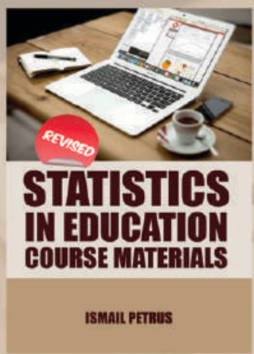


This course discusses some concepts and foundations of statistics necessary for undergraduate students to present and analyze the data in educational research. The course includes descriptive and inferential statistics, such as scales of measurement, measures of central tendency, variability and relative standing, correlation, linear regression, t-tests, one-way ANOVA, and some non-parametric tests. Upon the completion of the course, the students are expected to (a) have a solid understanding of statistical concepts and foundations, and (b) be able to do some statistical analyses in educational research.

Ismail Petrus is a lecturer of English Education at the Faculty of Teacher Training and Education,Sriwijaya University. He hold a Graduate Diploma of Arts in Interpreting/Tranlating from Deakin University, Melbourne, an M.A. in Linguistics from the University of Essex, UK, and a doctoral degree in English Education from Indonesia University of Education, Bandung, Indonesia. His research interest include linguistics, English language teaching and learning, and course desain.

Penerbit dan Percetakan
NoerFikri
Jl. Mayor Mahidin No. 142
Tlp./Fax. (0711) 366 625
E-mail : noerfikri@gmail.com
Palembang - Indonesia

ISBN 978-602-447-094-4
978-602-447-094-4

ISMAIL PETRUS

Statistics in Education Course Materials

REVISED

STATISTICS IN EDUCATION

COURSE MATERIALS

ISMAIL PETRUS

This course discusses some concepts and foundations of statistics necessary for undergraduate students to present and analyze the data in educational research. The course includes descriptive and inferential statistics, such as scales of measurement, measures of central tendency, variability and relative standing, correlation, linear regression, t-tests, one-way ANOVA, and some non-parametric tests. Upon the completion of the course, the students are expected to (a) have a solid understanding of statistical concepts and foundations, and (b) be able to do some statistical analyses in educational research.

Ismail Petrus is a lecturer of English Education at the Faculty of Teacher Training and Education,Sriwijaya University. He hold a Graduate Diploma of Arts in Interpreting/Tranlating from Deakin University, Melbourne, an M.A. in Linguistics from the University of Essex, UK, and a doctoral degree in English Education from Indonesia University of Education, Bandung, Indonesia. His research interest include linguistics, English language teaching and learning, and course desain.

Penerbit dan Percetakan
NoerFikri
Jl. Mayor Mahidin No. 142
Tlp./Fax. (0711) 366 625
E-mail : noerfikri@gmail.com
Palembang - Indonesia

ISBN 978-602-447-094-4
978-602-447-094-4

ISMAIL PETRUS

Statistics in Education Course Materials

REVISED

STATISTICS IN EDUCATION

COURSE MATERIALS

ISMAIL PETRUS

STATISTICS IN EDUCATION COURSE MATERIALS

REVISED

Ismail Petrus

Penerbit

NoerFikri Offset

Statistics in Education Course Materials Revised

Ismail Petrus

Penerbit

NoerFikri Offset

Statistics in Education Course Materials Revised

Penulis : Ismail Petrus
Layout : Nyimas Amrina Rosyada
Desain Cover : Sigit Dwi S

Hak Penerbit pada **NoerFikri Offset**, Palembang
Perpustakaan Nasional Katalog dalam Terbitan (KDT)
Anggota IKAPI (No. 012/SMS/13)

Dicetak oleh:
NoerFikri Offset
Jl. KH. Mayor Mahidin No. 142
Telp/Fax : 366 625
Palembang – Indonesia 30126
E-mail : noerfikri@gmail.com

Cetakan I: Januari 2019

ISBN : 978-602-447-094-4

OBJECTIVES OF THE COURSE

This course discusses some concepts and foundations of statistics necessary for undergraduate students to present and analyze the data in educational research. The course includes descriptive and inferential statistics, such as scales of measurement, measures of central tendency, variability and relative standing, correlation, linear regression, t-tests, one-way ANOVA, and some non-parametric tests. Upon the completion of the course, the students are expected to (a) have a solid understanding of statistical concepts and foundations, and (b) be able to do some statistical analyses in educational research.

TABLE OF CONTENTS

STATISTICS

Statistics is a branch of mathematics concerned with collection, classification, analysis, and interpretation of numerical facts, for drawing inferences on the basis of their quantifiable likelihood (probability) (http://www.businessdictionary.com/definition/statistics.html#ixzz3v geuhIAf). Statistics is a type of mathematical analysis involving the use of quantified representations, models and summaries for a given set of empirical data or real world observations (http://www.investopedia.com/terms/s/statistics.asp).

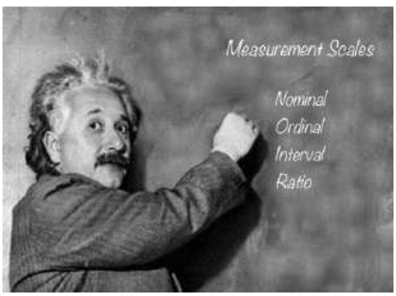
The science of statistics has evolved into two basic categories known as *descriptive* statistics and *inferential* statistics. The purpose of descriptive statistics is to summarize or display data so we can quickly obtain an overview. Inferential statistics allows us to make claims or conclusions about a population based on a sample of data from that population.

Data is the value assigned to an observation or a measurement and is the building blocks to statistical analysis. The plural form is data and the singular form is datum, referring to an individual observation or measurement. Data that describes a characteristic about a population is known as a **parameter**. Data that describes a characteristic about a sample is known as a **statistic. Information** is data that is transformed into useful facts that can be used for a specific purpose, such as making a decision. Primary data is data that you have collected for your own use. Secondary data is data collected by someone else that you are borrowing.

There are several ways to classify data: *quantitative-qualitative*, *discrete-continuous*, *nominal-ordinal-interval-ratio*. Quantitative data uses numerical values to describe something of interest. Qualitative data uses descriptive terms to measure or classify something of interest. One example of qualitative data is the name of a respondent in a survey and his or her level of education. Quantitative data can further be classified as discrete or continuous. Discrete data are quantitative data that are countable using a finite count, such as 1, 2, and so on. Continuous data are quantitative data that can take on any value within a range of values on a numerical scale in such a way that there are no gaps, jumps or other interruptions.

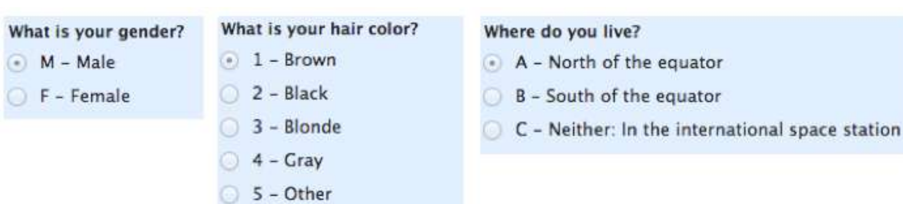
Scales of Measurement: Nominal, Ordinal, Interval and Ratio

There are four measurement scales (or types of data): nominal, ordinal, interval and ratio. (http://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/)



Nominal

Nominal scales are used for labeling variables, without any quantitative value. Nominal scales could simply be called labels. Here are some examples, below. Notice that all of these scales are mutually exclusive (no overlap) and none of them have any numerical significance. A good way to remember all of this is that nominal sounds a lot like name and nominal scales are kind of like names or labels.



Examples of Nominal Scales

Ordinal

With ordinal scales, the order of the values is what's important and significant, but the differences between each one is not really known. Take a look at the example below. In each case, we know that a #4 is better than a #3 or #2, but we don't know–and cannot quantify–how *much* better it is. For example, is the difference between "OK" and "Unhappy" the same as the difference between "Very Happy" and "Happy?" We can't say. Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc. With ordinal scales, it is the *order* that matters.

Advanced note: The best way to determine *central tendency* on a set of ordinal data is to use the mode or median; the mean cannot be defined from an ordinal set.

Example of Ordinal Scales

Interval

Interval scales are numeric scales in which we know not only the order, but also the exact differences between the values. The classic example of an interval scale is Celcius temperature because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees. Time is another good example of an interval scale in which the increments are known, consistent, and measurable. "Interval" itself means "space in between," which is the important thing to remember–interval scales not only tell us about order, but also about the value between each item.

Interval scales are nice because the realm of statistical analysis on these data sets opens up. For example, *central tendency* can be measured by mode, median, or mean; standard deviation can also be calculated.

Here's the problem with interval scales: they don't have a "true zero." For example, there is no such thing as "no temperature." Without a true zero, it is impossible to compute ratios. With interval data, we can add and subtract, but cannot multiply or divide. Confused? Ok, consider this: 10 degrees + 10 degrees = 20 degrees. No problem there. 20 degrees is not twice as hot as 10 degrees, however, because there is no such thing as "no temperature" when it comes to the Celsius scale. I hope that makes sense. Bottom line, interval scales are great, but we cannot calculate ratios, which brings us to our last measurement scale…

Example of Interval Scale

Ratio

Ratio scales are the ultimate nirvana when it comes to measurement scales because they tell us about the order, they tell us the exact value between units, and they also have an absolute zero–which allows for a wide range of both descriptive and inferential statistics to be applied. Everything above about interval data applies to ratio scales + ratio scales have a clear definition of zero. Good examples of ratio variables include height and weight.

Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

This Device Provides Two Examples of Ratio Scales (height and weight)

In summary, **nominal** variables are used to *name* or label a series of values. **Ordinal** scales provide good information about the *order* of choices, such as in a customer satisfaction survey. **Interval** scales give us the order of values + the ability to quantify *the difference between each one*. Finally, **ratio** scales give us the ultimate–order, interval values, plus the *ability to calculate ratios* since a "true zero" can be defined.

Summary of Data Types and Scale Measures

Provides:	Nominal	Ordinal	Interval	Ratio
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode, Median		✔	✔	✔
The "order" of values is known		✔	✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Exercises

A. For each item, classify it as *N* quantitative or *A* qualitative.

______ 1. Eye colors (blue, brown or so on) of babies
______ 2. Distances (in miles) traveled by students commuting to school
______ 3. Heights (in inches) of girls in a classroom
______ 4. Number of students in a classroom
______ 5. Number of teachers in favor of school uniforms
______ 6. Names of a group of students who took an exam
______ 7. Ages (in months) of children in a preschool.
______ 8. Ten-digit social security numbers of a group of citizens
______ 9. Foremost colors of flowers in a garden.
______ 10. Sex (male or female) of users of a website

B. For each item, classify it as *D* discrete or *C* continuous.

______ 1. Lengths (in inches) of newborn babies
______ 2. Distances (in miles) traveled by students commuting to school
______ 3. Heights (in inches) of girls in a classroom
______ 4. Number of students in a classroom
______ 5. Number of female teachers at a school
______ 6. Lengths (in meters) of broad jumps
______ 7. Weights (in pounds) of male police officers
______ 8. Number of correct answers on a quiz
______ 9. Number of heads in 100 tosses of a coin
______ 10. Daytime temperatures (in degrees Fahrenheit) over a 30-day period

C. Mark each item *N* nominal scale, *O* ordinal scale, *I* interval scale, or *R* ratio scale.

______ 1. Rankings of universities
______ 2. Customer satisfaction ratings of an internet store
______ 3. Raw scores on a standardized achievement test
______ 4. Sex (male or female) of users of a website

______ 5. Ranking of students in terms of amount of time on task during directed instruction
______ 6. IQ test score of 115
______ 7. Party affiliation (Republican, Democrat or so on) of voters
______ 8. Score of 635 on a college entrance exam
______ 9. Number of students whose ancestry is European, Asian, African
______ 10. Ten-digit social security numbers of a group of citizens
______ 11. Average monthly temperature in degrees Fahrenheit for a city throughout the year
______ 12. Average monthly rainfall in inches for the city of Wilmington throughout the year
______ 13. Education level of survey respondents
______ 14. Marital status of survey respondents
______ 15. Age of the respondents in the survey
______ 16. Gender of the respondents in the survey
______ 17. The year in which the respondent was born
______ 18. The uniform number of each member on a sports team
______ 19. A list of the graduating high school seniors by class rank
______ 20. Final exam scores for my statistics class on a scale of 0 to 100
______ 21. The state in which the respondents in a survey reside
______ 22. The voting intentions of the respondents in the survey classified as Republican, Democrat, or Undecided
______ 23. The race of the respondents in the survey classified as White, African American, Asian, or Other
______ 24. Performance rating of employees classified as Above Expectations, Meets Expectations, or Below Expectations

References

Donnelly, R.A. (2007). *The complete idiot's guide to statistics* (2nd Edition). New York: Alpha, a member of Penguin Group Inc.

Mueller, D. (1992). *An interactive guide to educational research: A modular approach*. Boston: Allyn and Bacon.

Descriptive Statistics: Measures of Central Tendency (Mean, Median, and Mode)

Measures of central tendency describe the center point of a data set with a single value. It's a valuable tool to help summarize many pieces of data with one number.

Mean

Mean or average is the most common measure of central tendency and is calculated by adding all the values in the data set and then dividing this result by the number of observations. The mathematical formula for the mean is as follows:

$$\bar{x} = \frac{\Sigma x}{N}$$

x stands for any of the numbers whose mean we are trying to compute
(Σ x) is the sum of all the x's

$\bar{x}$ (read as 'x-bar')

The mean can be calculated more conveniently from a frequency distribution table. The formula is as follows:

$$\bar{x} = \frac{\Sigma fx}{N}$$

x stands for any of the numbers whose mean we are trying to compute
(Σ fx) is the sum of all the x's multiplied by its frequency in the data

A *weighted mean* allows you to assign more weight to certain values and less weight to others. For example, let's say your grade this semester will be based on a combination of your final exam score, a homework score, and a final project, each weighted according to the following table.

Type	Score	Weight (Percent)
Exam	94	50
Project	89	35
Homework	83	15

We can calculate the final grade using the following formula for a weighted average. Note that here we are dividing by the sum of the weights rather than by the number of data values.

$$\bar{x} = \frac{\Sigma wx}{\Sigma w}$$

The symbol Σwx means "the sum of *w* times *x*." Each pair of *w* and *x* is first multiplied together, and these results are then summed. We can obtain the same result by plugging the numbers directly in to the formula for a weighted average:

$$\bar{x} = \frac{(0.50)(94) + (0.35)(89) + (0.15)(83)}{0.50 + 0.35 + 015} = 90.6$$

Median

Median is a measure of central tendency that represents the value in the data set for which half the observations are higher and half the observations are lower. When there is an even number of data points, the median will be the average of the two center points. For example,

3 4 4 4 5 6 7 7 9 17

Because we have an even number of data points (10), the median is the average of the two center points. In this case, that will be the values 5 and 6, resulting in a median of 5.5. Notice that there are four data values to the left (3, 4, 4, and 4) of these center points and four data values to the right (7, 7, 9, and 17).

3 4 4 4 5 6 7 7 9

In this instance, we only have one center point, which is the value 5. Therefore, the median for this data set is 5. Again, there are four data values to the left and right of this center point.

Mode

Mode is simply the observation in the data set that occurs the most frequently. There can be more than one mode of a data set if more than one value occurs the most frequent number of times. For example,

3 4 4 4 5 6 7 7 9 17

The mode is 4 because this value occurs three times in the data set.

Using Excel to Calculate Central Tendency

1. Using formulas
   Formulas→*fx* insert function→select a function
   Mean : =AVERAGE(B2:B11)
   Median : =MEDIAN(B2:B11)
   Mode : =MODE(B2:B11)
   Midpoint : =(MAX(B2:B11)+MIN(B2:B11))/2
2. Using Data Analysis
   Install Data Analysis
   Right click Office Button→Customize Quick Access Toolbar→Add-Ins→
   Analysis ToolPak→Go→Tick Analysis ToolPak→OK
   Excel will kindly calculate the mean, median, and mode for you all at once with a few mouse clicks.
   a. To begin, open a blank Excel worksheet.
   b. Click on the Tools menu at the top of the spreadsheet and select Data Analysis.
   c. Select Descriptive Statistics and click OK.

d. Select the Input Range, and select the Output Range. Then choose the Summary statistics check box and click OK.
e. Expand columns C and D slightly to see all the figures.

Exercises

1. Calculate the mean, median, and mode for the following data set: 20, 15, 24, 10, 8, 19, 24, 12, 21, 6.
2. Calculate the mean, median, and mode for the following data set: 84, 82, 90, 77, 75, 77, 82, 86, 82.
3. Calculate the mean, median, and mode for the following data set: 36, 27, 50, 42, 27, 36, 25, 40, 29, 15.
4. Calculate the mean, median, and mode for the following data set: 8, 11, 6, 2, 11, 6, 5, 6, 10.
5. A company counted the number of their employees in each of the following age classes. According to this distribution, what is the average age of the employees in the company?

Age Range	Number of Employees
20-24	8
25-29	37
30-34	25
34-39	48
40-44	27
45-49	10

6. Calculate the weighted mean of the following values with the corresponding weights.

Value	Weight
118	3
125	2
107	1

7. A company counted the number of employees at each level of years of service in the following table. What is the average number of years of service in this company?

Years of Service	Number of Employees
1	5
2	7
3	10
4	8
5	12
6	3

References

Butler, C. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell.

Donnelly, R.A. (2007). *The complete idiot's guide to statistics* (2nd Edition). New York: Alpha, a member of Penguin Group Inc.

Summary of the Properties of the Mean, Median, and Mode

The **mean** is

1. The balance point of a distribution;
2. The preferred measure for relatively symmetrical distributions and quantitative variables;
3. The measure with the best sampling stability;
4. Widely used in advanced statistical procedures;
5. Mathematically tractable;
6. The only measure whose value is dependent on the value of every core in the distribution;
7. More sensitive to extreme scores than the median and the mode and, hence, is not recommended for markedly skewed distributions;
8. Not appropriate for qualitative data; and
9. Not appropriate for open-ended distributions.

The **median** is

1. The point that divides the ordered scores into two samples of equal size;
2. Second to the mean in usefulness;
3. Widely used for markedly skewed distributions because it is sensitive only to the number rather than to the values of scores above and below it;

4. The most stable measure that can be used with open-ended distributions;
5. More subject to sampling fluctuation than the mean;
6. Less mathematically tractable than the mean; and
7. Less often used in advanced statistical procedures.

The **mode** is
1. The score that occurs most often and, therefore, the most typical value;
2. The only measure appropriate for unordered qualitative variables;
3. More appropriate than the mean or the median for quantitative variables that are inherently discrete;
4. The easiest measure to compute;
5. Much more subject to sampling fluctuation than the mean and the median;
6. Less mathematically tractable than the mean and the median;
7. Not necessarily existent, as when a distribution has two or more scores with the same maximum frequency; and
8. Rarely used in advanced statistical procedures.

The choice of an appropriate measure of central tendency depends on three considerations:
1. The level of measurement of the variables (nominal, ordinal, interval or ratio);
2. The shape of the frequency distribution;
3. What is to be done with the figure obtained.

The **mean** is really suitable only for ratio and interval data. For ordinal variables, where the data can be ranked but one cannot validly talk of equal differences between the values, the **median**, which is based on ranking, may be used. Where it is not even possible to rank the data, as in the case of a nominal variable, the **mode** may be the only measure available.

The median is also useful when a distribution contains one or more highly untypical values. Besides, if a distribution is strongly skewed, either positively or negatively, the median is a better indicator of central tendency than the mean. However, the mean has properties which make it particularly suitable for further statistical procedures.

Exercises

For the following sets of data, what measures of central tendency would you compute? Justify your choices.

a. 9, 6, 5, 7, 1, 6, 7, 8, 10, 6, 5, 4, 3, 6, 9, 7, 4, 5, 6, 8, 3, 2
b. 6, 5, 9, 6, 7, 5, 6, 8, 3, 4, 5, 7, 5, 4, 8, 5
c. 3, 5, 8, 5, 7, 9, 4, 2, 5, 6, 6, 23
d. 4, 3, 7, 5, 4, 2, 12, 6, 5, 4, 3, 3, 2, 7, 1, 6, 4, 5, 3, 5
e. Eye color: blue, brown, brown, blue, green, brown, gray, brown, blue
f. 7, 8, 6, 7, 8, 9, 1, 6, 5, 3, 7, 8, 7, 6, 7, 8, 5, 7
g. Family size: 4, 3, 5, 4, 1, 2, 4, 6, 5
h. Number of words in each chapter of a novel: 4841, 4724, 3420, 3861, 4017, 5129, 3450, 3481, 5001, 4482, 17807

References

Butler, C.S. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell Ltd.

Kirk, R.E. (2008). *Statistics: An introduction*. Fifth Edition. Belmont, C.A.: Thomson Wadsworth.

Descriptive Statistics: Measures of Dispersion (Range, Variance, Standard Deviation)

Measures of dispersion include range, variance, and standard deviation.

Range

Range is the simplest measure of dispersion and is calculated by finding the difference between the highest value and the lowest value in the data set. The range is a “quick-and-dirty” way to get a feel for the spread of the data set. However, the limitation is that it only relies on two data points to describe the variation in the sample. No other values between the highest and lowest points are part of the range calculation.

For example,

7 9 8 11 4 5

The range would be: 11 - 4 = 7

Variance

Variance is a measure of dispersion that describes the relative distance between the data points in the set and the mean of the data set. It summarizes the squared deviation of each data value from the mean. The formula is:

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{N-1}$$

s^2 = the variance of the sample
ξ = the sample mean
n = the size of the sample (the number of data values)
(x- $\bar{x}$) = the deviation from the mean for each value in the data set

Standard Deviation

Standard deviation is simply the square root of the variance. To calculate the standard deviation, you must first calculate the variance and then take the square root of the result. The standard deviation is actually a more useful measure than the variance because the standard deviation is in the units of the original data set.

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{N - 1}}$$

Using Excel to Calculate Dispersion

Using formulas

Formulas→*fx* insert function→select a function

Range : =MAX(B2:B11)-MIN(B2:B11)

Variance : =VAR(B2:B11)

Standard Deviation: =STDEV(B2:B11)

Using SPSS to Calculate Central Tendency and Dispersion

1. Analyze→Descriptive Statistics→Descriptives
2. Analyze→Descriptive Statistics→Explore

Exercises

1. Calculate the variance, standard deviation, and the range for the following data set: 20, 15, 24, 10, 8, 19, 24.
2. Calculate the variance, standard deviation, and the range for the following data set: 84, 82, 90, 77, 75, 77, 82, 86, 82.
3. Calculate the variance, standard deviation, and the range for the following data set: 36, 27, 50, 42, 27, 36, 25, 40.
4. A company counted the number of their employees in each of the age classes as follows. According to this distribution, what is the standard deviation for the age of the employees in the company?

Age Range	Number of Employees
20-24	8
25-29	37
30-34	25
34-39	48
40-44	27
45-49	10

5. A company counted the number of employees at each level of years of service in the table that follows. What is the standard deviation for the number of years of service in this company?

Years of Service	Number of Employees
1	5
2	7
3	10
4	8
5	12
6	3

References

Butler, C. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell.
Donnelly, R.A. (2007). *The complete idiot's guide to statistics* (2nd Edition). New York: Alpha, a member of Penguin Group Inc.

Mean Deviation

The mean of the distances of each value from their mean.

Yes, we use "**mean**" twice: Find the mean ... use it to work out distances ... then find the mean of those!

Three steps:

- 1. Find the mean of all values
- 2. Find the **distance** of each value from that mean (subtract the mean from each value, ignore minus signs)
- 3. Then find the **mean of those distances**

Like this:

Example: the Mean Deviation of 3, 6, 6, 7, 8, 11, 15, 16

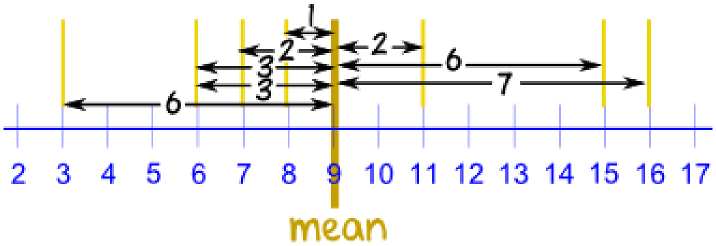
Step 1: Find the **mean**:

$$\text{Mean} = \frac{3+6+6+7+8+11+15+16}{8} = \frac{72}{8} = 9$$

Step 2: Find the **distance** of each value from that mean:

Value	Distance from 9
3	6
6	3
6	3
7	2
8	1
11	2
15	6
16	7

Which looks like this:



Step 3. Find the **mean of those distances**:

$$\text{Mean Deviation} = \frac{6+3+3+2+1+2+6+7}{8} = \frac{30}{8} = 3.75$$

So, the **mean = 9**, and the **mean deviation = 3.75**

It tells us how far, on average, all values are from the middle.

In that example the values are, on average, 3.75 away from the middle.

The formula is:

$$\text{Mean Deviation} = \frac{\Sigma|x - \mu|}{N}$$

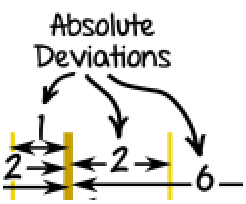
Let's learn more about those symbols!

Firstly:

- **μ** is the mean (in our example **μ = 9**)
- **x** is each value (such as 3 or 16)
- **N** is the number of values (in our example **N = 8**)

Absolute Deviation

Each distance we calculated is called an **Absolute Deviation**, because it is the Absolute Value of the deviation (how far from the mean).



To show "Absolute Value" we put "|" marks either side like this: **|-3| = 3**

For any value **x**:

Absolute Deviation = |x - μ|

From our example, the value **16** has **Absolute Deviation = |x - μ| = |16 - 9| = |7| = 7**

And now let's add them all up.

Sigma

The symbol for "Sum Up" is **Σ** (called Sigma Notation), so we have:

Sum of Absolute Deviations = Σ|x - μ|

Divide by how many values **N** and we have:

$$\text{Mean Deviation} = \frac{\Sigma|x - \mu|}{N}$$

Let's do our example again, using the proper symbols:

Example: the Mean Deviation of 3, 6, 6, 7, 8, 11, 15, 16

Step 1: Find the **mean**:

$$\mu = \frac{3+6+6+7+8+11+15+16}{8} = \frac{72}{8} = 9$$

Step 2: Find the **Absolute Deviations**:

x	$\|x - \mu\|$
3	6
6	3
6	3
7	2
8	1
11	2
15	6
16	7
	$\Sigma\|x - \mu\| = 30$

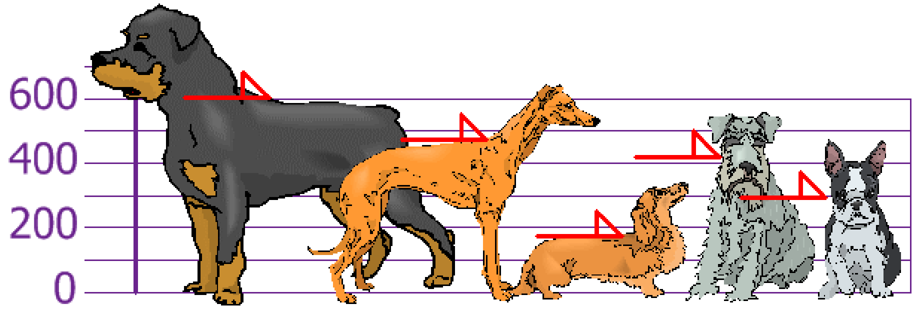
Step 3. Find the **Mean Deviation**:

$$\text{Mean Deviation} = \frac{\Sigma|x - \mu|}{N} = \frac{30}{8} = 3.75$$

Note: *the mean deviation is sometimes called the Mean Absolute Deviation (MAD) because it is the mean of the absolute deviations.* Mean Deviation tells us how far, on average, all values are from the middle.

Here is an example.

You and your friends have just measured the heights of your dogs (in millimeters):



The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Step 1: Find the **mean**:

$$\mu = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = \mathbf{394}$$

Step 2: Find the **Absolute Deviations**:

x	**|x - μ|**
600	206
470	76
170	224
430	36
300	94
	Σ|x - μ| = 636

Step 3. Find the **Mean Deviation**:

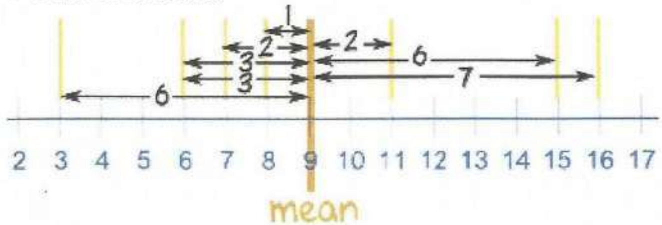
$$\text{Mean Deviation} = \frac{\Sigma|x - \mu|}{N} = \frac{636}{5} = 127.2$$

So, on average, the dogs' heights are **127.2 mm from the mean**. (Compare that with the Standard Deviation of **147 mm**)

A Useful Check

The deviations on **one side** of the mean should equal the deviations on the **other side**.
From our first example:

Example: 3, 6, 6, 7, 8, 11, 15, 16
The deviations are:



$$6 + 3 + 3 + 2 + 1 = 2 + 6 + 7$$
$$15 = 15$$

Likewise:

Example: Dogs

Deviations left of mean: 224 + 94 = **318**

Deviations right of mean: 206 + 76 + 36 = **318**

If they are not equal ... you may have made a mistake!

<u>Exercises</u>

1. Calculate the mean deviation for the numbers: 75, 83, 96, 100, 121 and 125.

2. Ten friends scored the following marks in their end-of-year math exam:
   23%, 37%, 45%, 49%, 56%, 63%, 63%, 70%, 72% and 82%.
   What was the mean deviation of their marks?

3. A booklet has 12 pages with the following numbers of words: 271, 354, 296, 301, 333, 326, 285, 298, 327, 316, 287 and 314
   What is the mean deviation of the number of words per page?

4. The Lakers scored the following numbers of goals in their last twenty matches:
   3, 0, 1, 5, 4, 3, 2, 6, 4, 2, 3, 3, 0, 7, 1, 1, 2, 3, 4, 6
   Calculate the Mean Deviation.

5. Ramiro did a survey of the number of pets owned by his classmates, with the following results:

Number of pets	Frequency
0	4
1	12
2	8
3	2
4	1
5	2
6	1

   What was the mean deviation of the number of pets?

6. The children in a class did a survey of the number of siblings (brothers and sisters) each of them had. The results are recorded in the following table:

Number of siblings	Frequency
0	3
1	6
2	8
3	5
4	4
5	2
6	1
7	0
8	0
9	1

Calculate the mean deviation of the number of siblings.

Coefficient of Variation

The coefficient of variation (CV) is the ratio of the standard deviation to the mean (average). For example, the expression "The standard deviation is 15% of the mean" is a coefficient of variation. The CV is particularly useful when you want to compare results from two different surveys or tests that have different measures or values. For example, if you are comparing the results from two tests that have different scoring mechanisms.

Coefficient of Variation is the percentage variation in mean, standard deviation being considered as the total variation in the mean. If we wish to compare the variability of two or more series, we can use the coefficient of variation. The series of data for which the coefficient of variation is large indicates that the group is more variable and it is less stable or less uniform. If a coefficient of variation is small it indicates that the group is less variable and it is more stable or more uniform.

The **formula for the coefficient of variation** is:
Coefficient of Variation = (Standard Deviation / Mean) * 100.
In symbols: CV = (SD/ $\bar{x}$) *100.
Multiplying the coefficient by 100 is an optional step to get a percentage, as opposed to a decimal.

Coefficient of Variation Example

A researcher is comparing two multiple-choice tests with different conditions. In the first test, a typical multiple-choice test is administered. In the second test, alternative choices (i.e. incorrect answers) are randomly assigned to test takers. The results from the two tests are:

	Regular Test	Randomized Answers
Mean	59.9	44.8
SD	10.2	12.7

Trying to compare the two test results is challenging. Comparing standard deviations doesn't really work, because the *means* are also different. Calculation using the formula CV=(SD/Mean)*100 helps to make sense of the data:

	Regular Test	Randomized Answers
Mean	59.9	44.8
SD	10.2	12.7
CV	17.03	28.35

Looking at the standard deviations of 10.2 and 12.7, you might think that the tests have similar results. However, when you adjust for the difference in the means, the results have more significance:

Regular test: CV = 17.03
Randomized answers: CV = 28.35

Note: The Coefficient of Variation should only be used to compare positive data on a ratio scale. The CV has little or no meaning for measurements on an interval scale. Examples of interval scales include temperatures in Celsius or Fahrenheit, while the Kelvin scale is a ratio scale that starts at zero and cannot, by definition, take on a negative value (0 degrees Kelvin is the absence of heat).

http://www.statisticshowto.com/how-to-find-a-coefficient-of-variation/

Exercises

1. Calculate the coefficient of variance of the marks secured by a student in the exam as given:
   82, 95, 75, 78, 87
2. The standard deviation and the mean of 16 values are 15.6 and 20.5. Find the coefficient of variation
3. A group of 80 candidates have their average height is 145.8 cm with coefficient of variance
   2.5%. What is the standard deviation of their height?
4. A company has two sections with 40 and 65 employees respectively. Their average weekly wages are $450 and $350. The standard deviations are 7 and 9. (i) Which section has a larger wage bill? (ii) Which section has larger variability in wages?
   Solution:
   (i) Wage bill for section A = 40 x 450 = 18000
   Wage bill for section B = 65 x 350 = 22750
   Section B is larger in wage bill.
   (ii) Coefficient of variation for Section A = 7/450 x 100 =1.56 %
   Coefficient of variation for Section B = 9/350 x 100 = 2.57%
   Section A is more consistent so there is greater variability in the wages of section B.

Index of Qualitative Variation

The index of qualitative variation (IQV) is a measure of variability for nominal variables, like race and ethnicity. It is based on the ratio of the total number of differences in the distribution to the maximum number of possible differences within the same distribution. The index can vary from 0.00 to 1.00. When all the cases in the distribution are in one category, there is no variation (or diversity) and IQV is 0.00. In contrast, when the cases in the distribution are distributed evenly across the categories, there is maximum variation (or diversity) and IQV is 1.00.

To calculate the IQV, use this formula:

$$\text{IQV} = \frac{K(100^2 - \sum Pct^2)}{100^2(K - 1)}$$

where:

K = the number of categories

N = the total number of cases in the distribution

ΣPct^2 = the sum of all squared percentages in the distribution

Suppose you live in Maine where the majority of residents are white and a small minority are Latino or Asian. Also suppose that your best friend lives in Hawaii where the numbers of white, Asian, and Hawaiian residents are about equal. The distributions for these two states are presented in following table. Which is more diverse? Clearly, Hawaii, where whites, Asians, and Hawaiian residents are more or less equally represented, is more diverse than Maine, where blacks and Latinos are but a small minority. You can also get a visual feel for the relative diversity in the two states by examining the two bar charts presented in the following figure.

Table: Top Three Racial/Ethnic Groups for Two States by Percentage

Racial/Ethnic Group	Maine	Hawaii
White	98.4	33.5
Latino	0.8	X
Asian	0.8	55.1
Native Hawaiian or Pacific Islander	X	11.4
Total	100.0	100.0

(Source: United States Bureau of the Census, *Statistical Abstract of the United States*, 2003, Tables 21 & 22)

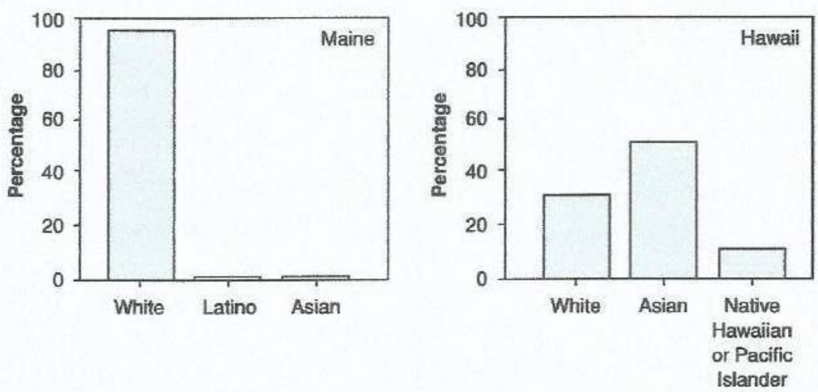


Figure: Top Three Racial/Ethnic Groups in Maine and Hawaii

Table: Squared Percentages for Three Racial/Ethnic Groups for Two States

Race/Ethnic Group	*Maine*		*Hawaii*	
	%	$(\%)^2$	%	$(\%)^2$
White	98.4	9,683	33.5	1122
Latino	0.8	.64	X	X
Asian	0.8	.64	55.1	3036
Native Hawaiian or Pacific Islander	X	X	11.4	130
Total	100.0	9684.3	100.0	4288.0

The IQV for Maine is

$$\text{IQV} = \frac{K(100^2 - \sum Pct^2)}{100^2(K-1)} = \frac{3(100^2 - 9{,}684.3)}{100^2(3-1)} = \frac{947.1}{20{,}000} = 0.05$$

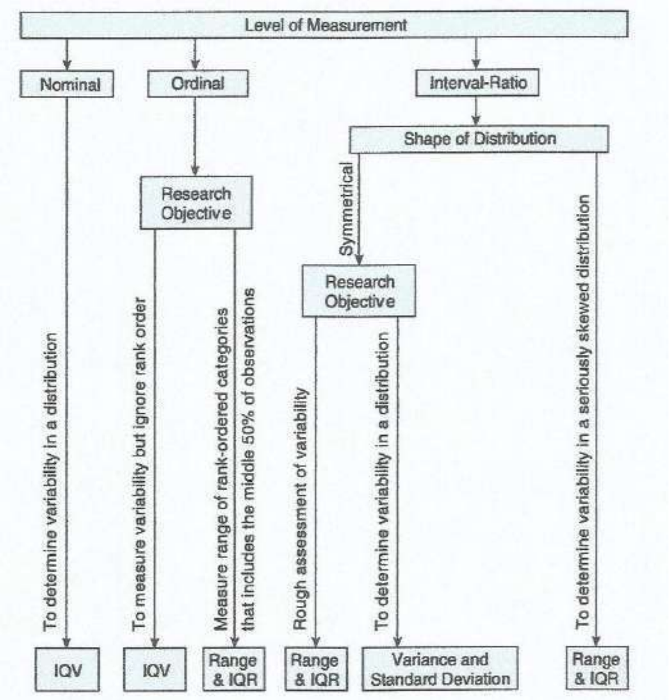
The IQV for Hawaii is

$$\text{IQV} = \frac{K(100^2 - \sum Pct^2)}{100^2(K-1)} = \frac{3(100^2 - 4{,}288)}{100^2(3-1)} = \frac{17{,}136}{20{,}000} = 0.86$$

In Hawaii, where IQV = 0.86, there is considerably more racial/ethnic variation than in Maine,
where IQV = 0.05.

The IQV can also be expressed as a percentage, rather than a proportion: simply multiply the IQV by 100. Expressed as a percentage, the IQV would reflect the percentage of racial/ethnic differences relative to the maximum possible differences in each distribution. Thus, an IQV of 0.05 indicates that the number of racial/ethnic differences in Maine is 5.0 percent (0.05 x l00) of the maximum possible differences. Similarly, for Hawaii, an IQV of 0.86 means that the number of racial/ethnic differences is 86.0 percent (0.86 x 100) of the maximum possible differences.

How to Choose a Measure of Variation



The choice of measure of variation for ordinal variables is more problematic. The IQV can be used to reflect variability in distributions of ordinal variables, but because it is not sensitive to the rank ordering of values implied in ordinal variables, it loses some information. Another possibility is to use the interquartile range. However, the interquartile range relies on distance between two scores to express variation, information that cannot be obtained from ordinal measured scores. The compromise is to use the interquartile range (reporting Ql and Q3) alongside the median, interpreting the interquartile range as the range of rank-ordered values that includes the middle 50 percent of fhe observations.

For interval-ratio variables, you can choose the variance (or standard deviation), range or interquartile range. Because the range, and to a lesser extent the interquartile range, is based on only two scores in the distribution (and therefore tends to be sensitive if either of the two points is extreme), the variance and/or standard deviation is usually preferred. However, if a distribution is extremely skewed so that the mean is no longer representative of the central tendency in the distribution, the range and the interquartile range can be used.

Exercises

1. Americans often think of themselves as quite diverse in their political opinions, within the continuum of liberal to conservative. The data from the 2002 General Social Survey show the diversity of political views. The frequency table displays respondents' self-rating of their political positions.

Political Views	Percent
Extremely liberal	3.5
Liberal	10.7
Slightly liberal	11.9
Moderate	39.2
Slightly conservative	15.7
Conservative	15.8
Extremely conservative	3.2
Total	100.0

a. How many categories (K) are we working with?
b. Calculate the sum of the squared percentages or ΣPct^2.
c. What is the IQV for this variable? Do you find it to be higher (closer to 1) or lower (closer to 0) than you might have expected for political views?

2. Using the information listed below, answer the following questions to get an idea about the educational attainment, by percentage, of the General Social Survey respondents in 2002.

Highest Educational Degree	Male	Female
Less than high school	17.0	13.6
High school graduate	55.8	55.8
Junior college	7.8	7.4
Bachelor's degree	9.1	15.7
Graduate degree	10.3	7.5
Total	100.0	100.0

a. What is the value of K?
b. Calculate the sum of the squared percentages or ΣPct^2 for males and females.
c. Calculate the IQV for males and females. Is there more diversity by degree for males or females?

3. The US Census Bureau annually estimates the percentage of Americans below the poverty level for various geographic areas. Use the information in the following table to characterize poverty in the southern versus the western portions of the United States. Compare the variability of the poverty rate of the states in the West with those in the South.

Percentage of Americans Below the Poverty Level: Southern and Western Regions, 1999

South	*Percent Below Poverty Level*	*West*	*Percent Below Poverty Level*
Alabama	16.1	Alaska	16.1
Florida	12.5	Arizona	13.9
Georgia	13.0	California	14.2
Kentucky	15.8	Idaho	11.8
Louisiana	19.6	Montana	14.6
Mississippi	10.5	Nevada	10.5
North Carolina	12.3	New Mexico	18.4
South Carolina	14.1	Oregon	11.6
Tennessee	13.5	Utah	9.4
		Washington	10.6

Source: U.S. Bureau of the Census, *Statistical Abstract of the United States*, 2003, Table 698.

References

Frankfort-Nachmias, C., & Leon-Guerrero, A. (2018). *Social statistics for a diverse society* (Eighth Edition). Thousand Oaks, California: SAGE Publications, Inc.

Descriptive Statistics: Measures of Relative Standing (Percentiles, z Scores, T scores, CEEB Scores)

Measures of relative standing include percentiles, z scores, T scores, CEEB scores.

Percentiles

Percentiles are like quartiles, except that percentiles divide the set of data into 100 equal parts while quartiles divide the set of data into 4 equal parts. Percentiles measure position from the bottom. Percentiles are most often used for determining the relative standing of an individual in a population or the rank position of the individual. Some of the most popular uses for percentiles are connected with test scores and graduation standings. Percentile ranks are an easy way to convey an individual's standing at graduation relative to other graduates. Unfortunately, there is no universally accepted definition of "percentile". Consider the following two slightly different definitions:

Definition 1: A **percentile** is a measure that tells us what percent of the total frequency scored at or below that measure. A *percentile rank* is the percentage of scores that fall at or below a given score.
Formula: To find the percentile rank of a score, *x*, out of a set of *n* scores, where *x* is included:

$$\frac{(B + 0.5E)}{n} \cdot 100 = \textit{percentile rank}$$

where *B* = number of scores below *x*
E = number of scores equal to *x*
n = number of scores

Example: If Jason graduated 25th out of a class of 150 students, then 125 students were ranked below Jason. Jason's percentile rank would be:

$$\frac{125 + 0.5(1)}{150} = \frac{125.5}{150} = .836 = 84^{th} \text{ percentile}$$

Jason's standing in the class at the 84th percentile is as higher or higher than 84% of the graduates. Good job, Jason!

Definition 2: A **percentile** is a measure that tells us what percent of the total frequency scored below that measure. A *percentile rank* is the percentage of scores that fall below a given score.
Formula: To find the percentile rank of a score, *x*, out of a set of *n* scores, where *x* is not included:

$$\frac{\text{number of scores below } x}{n} \cdot 100 = \text{percentile rank}$$

Example: If Jason graduated 25th out of a class of 150 students, then 125 students were ranked below Jason. Jason's percentile rank would be:

$$\frac{125}{150} = .83 = 83^{rd} \text{ percentile}$$

Jason's standing in the class at the 83rd percentile is higher than 83% of the graduates. Good job, Jason!

The slight difference in these two definitions can lead to significantly different answers when dealing with small amounts of data.

About Percentile Ranks:

- percentile rank is a number between 0 and 100 indicating the percent of cases falling at or below that score.
- percentile ranks are usually written to the nearest whole percent: 74.5% = 75% = 75^{th} percentile
- scores are divided into 100 equally sized groups
- scores are arranged in rank order from lowest to highest
- there is no 0 percentile rank - the lowest score is at the first percentile
- there is no 100th percentile - the highest score is at the 99th percentile.
- you cannot perform the same mathematical operations on percentiles that you can on raw scores. You cannot, for example, compute the mean of percentile scores, as the results may be misleading.

Consider:

1. Karl takes the big Earth Science test and his teacher tells him that he scored at the 92^{nd} percentile. Is Karl pleased with his performance on the test? **He should be. He scored as high or higher than 92% of the people taking the test.**

2. Sue takes the Chapter 4 math test. If Sue's score is the same as "the mean" score for the math test, she scored at the 50th percentile and she did "as well or better than" 50% of the students taking the test.

3. If Ty scores at the 75th percentile on the Social Studies test, he did "as well or better than" 75% of the students taking the test.

Examples: Finding Percentiles

1. The math test scores were: 50, 65, 70, 72, 72, 78, 80, 82, 84, 84, 85, 86, 88, 88, 90, 94, 96, 98, 98, 99. Find the percentile rank for a score of 84 on this test. Be sure the scores are ordered from smallest to largest. Locate the 84.

   Solution Using Formula:

   $$\frac{(B + 0.5E)}{n} \cdot 100 = percentile\ rank$$

   $$\frac{8 + 0.5(2)}{20} \cdot 100 = \frac{9}{20} \cdot 100 = 45^{th}\ percentile$$

   Solution Using Visualization:

   Since there are 2 values equal to 84, assign one to the group "above 84" and the other to the group "below 84".

   50, 65, 70, 72, 72, 78, 80, 82, 84, | 84, 85, 86, 88, 88, 90, 94, 96, 98, 98, 99

   $$\frac{number\ of\ scores\ below\ x}{n} \cdot 100 = percentile\ rank \qquad \frac{9}{20} \cdot 100 = 45^{th}\ percentile$$

   The score of 84 is at the 45th percentile for this test.

2. The math test scores were: 50, 65, 70, 72, 72, 78, 80, 82, 84, 84, 85, 86, 88, 88, 90, 94, 96, 98, 98, 99. Find the percentile rank for a score of 86 on this test. Be sure the scores are ordered from smallest to largest. Locate the 86.

Solution Using Formula:

$$\frac{(B+0.5E)}{n} \cdot 100 = percentile\ rank$$

$$\frac{11+0.5(1)}{20} \cdot 100 = \frac{11.5}{20} \cdot 100 = 58^{th}\ percentile$$

Solution Using Visualization:

Since there is only one value equal to 86, it will be counted as "half" of a data value for the group "above 86" as well as the group "below 86".

50, 65, 70, 72, 72, 78, 80, 82, 84, 84, 85, 8|6, 88, 88, 90, 94, 96, 98, 98, 99

$$\frac{number\ of\ scores\ below\ x}{n} \cdot 100 = percentile\ rank$$

$$\frac{11.5}{20} \cdot 100 = 57.5 = 58^{th}\ percentile$$

The score of 86 is at the 58th percentile for this test.

Calculating Percentiles

To calculate the *k*th percentile (where *k* is any number between one and one hundred), do the following steps:

1. Order all the numbers in the data set from the smallest to the largest.
2. Multiply *k* percent times the total number of numbers, *n*.

3a. If your result from Step 2 is a whole number, go to Step 4. If the result from Step 2 is not a whole number, round it up to the nearest whole number and go to Step 3b.

3b. Count the numbers in your data set from left to right (from the smallest to the largest number) until you reach the value indicated by Step 3a. The corresponding value in your data set is the *k*th percentile.

4. Count the numbers in your data set from left to right until you reach the one indicated by Step 2. The *k*th percentile is the average

of that corresponding value in your data set and the value that directly follows it.

For example, suppose you have 25 test scores, and in order from the lowest to the highest they look like this: 43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99. To find the 90*th* percentile for these ordered scores, start by multiplying 90% times the total number of scores, which gives 90% x 25 = 0.90 x 25 = 22.5. Rounding up to the nearest whole number, you get 23.

Counting from left to right (from the smallest to the largest number in the data set), you go until you find the 23rd number in the data set. That number is 98, and it's the 90th percentile for this data set.

Now say you want to find the 20th percentile. Start by taking 20% x 25 = 0.20 x 25 = 5; this is a whole number, so proceed from Step 3a to Step 4, which tells us the 20th percentile is the average of the 5^{th} and 6th numbers in the ordered data set (62 and 66). The 20th percentile then comes to (62+66)/2 = 64.
The median (the 50th percentile) for the test scores is the 13th score: 77.

z Scores

A z score (also known as a z value, standard score, and normal score) is used to describe a particular score in terms of where it fits into an overall group of scores. In other words, a z score is an ordinary score transformed so that it better describes the location of that score in a distribution. A z score has a mean of zero and a standard deviation of one. The formula is:

$$z = \frac{x - \bar{x}}{s}$$

where: x is any particular value of the variable
$\bar{x}$ is the mean
s is the standard deviation

T Scores

T scores are used to tell individuals how far their score is from the mean. T scores have a mean of 50 and a standard deviation of 10. Therefore, if a student's raw score was converted to a T score and their T score was 70 it would in turn mean that their score was 20 points above the mean. One advantage of using a T score over a z score is that T scores are relatively easy to explain to parents when reporting the student's assessment scores. The formula is T = 10z + 50.

CEEB Scores

College Entrance Examination Board (CEEB) scores are another variation of the z score that is often reported in the USA. To convert z scores to CEEB scores, multiply the z score by 100 and add 500, as follows: CEEB = 100z + 500.

Standard Nine (Stanine)

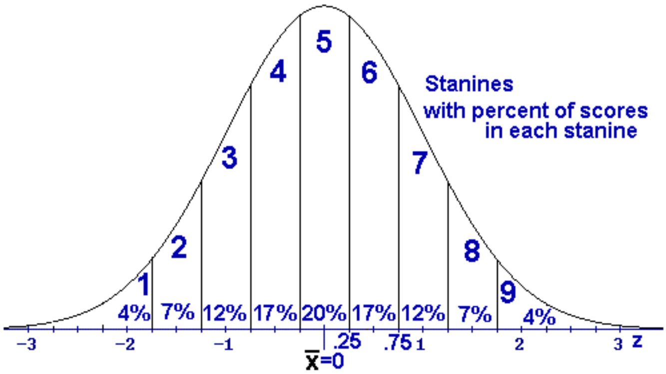
Standard Nine is a method of scaling test scores on a 9-point standard scale with a mean of five and a standard deviation of 2. It can be used to convert any test score to a single-digit score. Like z-scores and t-scores, stanines are a way to assign a number to a member of a group, relative to all members in that group. However, while z-scores and t-scores can be expressed with decimals like 1.2 or 3.25, stanines are always positive whole numbers from 0 to 9.

Stanines are also similar to normal distributions. You can think of these scores as a bell curve that has been sliced up into 9 pieces. These pieces are numbered 1 through 9, starting at the left hand section. However, where a standard normal distribution has a mean of 0 and a standard deviation of 1, stanines have a mean of 5 and a standard deviation of 2.

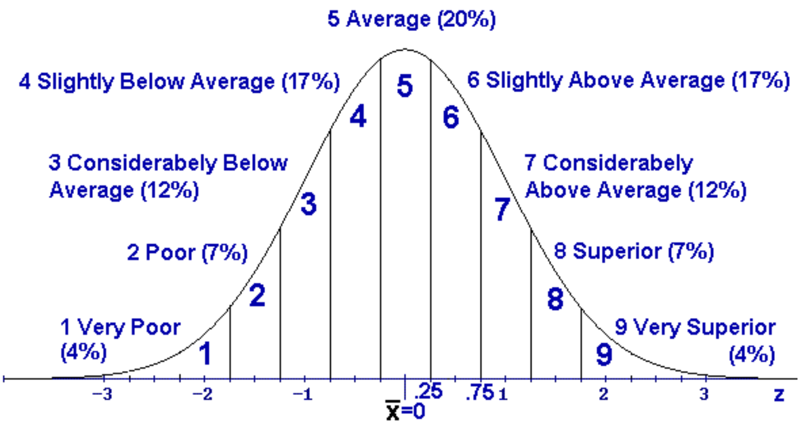
Test scores are scaled to stanine scores using the following algorithm:

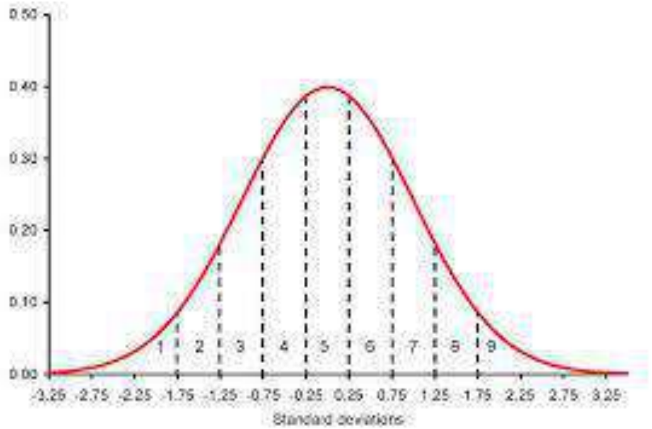
1. Rank results from lowest to highest
2. Give the lowest 4% a stanine of 1, the next 7% a stanine of 2, etc., according to the following table:

Stanine score	Percentage of scores
1	Bottom 4%
2	Next bottom 7%
3	Next bottom 12%
4	Next Bottom 17%
5	Middle 20%
6	Next top 17%
7	Next top 12%
8	Next top 7%
9	Top 4%

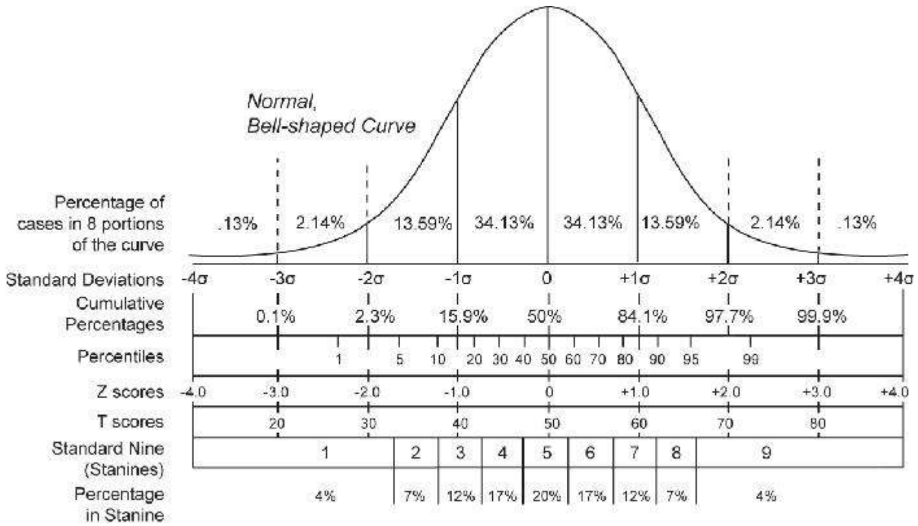


Stanines defined descriptively (with percent of scores):





www.mathnstuff.com/math/spoken/here/2class/90/stanine.htm
www.sciencing.com/calculate-stanine-scores-8459002.html
www.statisticshowto.com/stanine/



Using Excel to Calculate Relative Standing

Using formulas

Formulas→*fx* insert function→select a function

Percentile Rank : =PERCENTRANK($B$2:$B$11,B2)

75th Percentile : =PERCENTILE($B$2:$B$11,0.75)

z-score : =STANDARDIZE(B2,AVERAGE($B$2:$B$11),STDEV($B$2:$B$11))

Absolute value (=ABS), maintain the position ($B$10)

Finding stanine score from z-score: =MAX(1, MIN(9, ROUND((z-score+2)*2, 0)))

Using data analysis

Percentile Rank : Data Analysis→Rank and Percentile

Using SPSS to Calculate Relative Standing

Percentile Rank: Transform→Rank Cases→Rank Types→Fractional rank as %

(see Data View→PData)

(Display in output): Analyze→Reports→Case Summaries→Enter Variables:

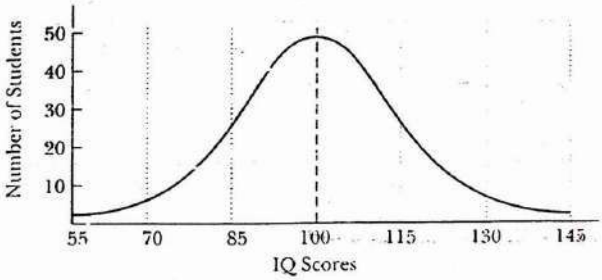
Data1 and PData1, check Display cases, Limit cases to first, Show only valid cases

Percentile:

a. Analyze → Descriptive Statistics → Frequencies → Statistics → Percentile (s):15Add
b. Analyze → Descriptive Statistics → Explore → Statistics → Percentiles z-score : Analyze → Descriptive Statistics → Descriptives → check Save standardized values as variables (see Data View → Z Data1) (Display in output): Analyze → Reports → Case Summaries → Enter Variables: Data1 and ZData1, check Display cases, Limit cases to first, Show only valid cases Sort data : Data → Sort Cases

APPLICATION EXERCISES

A. Look at the frequency polygon, and answer the questions that follow.



where: $\overline{X} = 100$
$S = 15$
$N = 947$

A1. What percentile score would an IQ score of 85 represent?

A2. About what percentage of students scored between 70 and 115?

A3. If Iliana had a score of 177 on this test, about how many standard deviations would she be above the mean? Does this mean that she is really intelligent?

A4. What would Iliana's z score be? T score? CEEB score?

B. In the table below, the raw score mean is 50, and the raw score standard deviation is 7. Fill in all the missing spaces by using the available information and what you now know about distributions and standardized scores.

Student	Raw score	z score	T score	CEEB score
A	64		70	
B	50			
C		–1		
D		–1.5		350
etc.				

C. Study the table below, and answer the questions that follow.

Test	Raw scores			Standardized scores	
	k^*	$\bar{X}$	S	$\bar{X}$	S
A	110	60	25	500	100
B	75	60	15	50	10
C	50	11	4	0	1

* Remember, k = number of items on the test

C1. Which test (A, B, or C) shows standardized scores that are probably:

a. z scores? ________

b. T scores? ________

c. CEEB scores? ________

C2. In raw scores, which test has:

a the largest standard deviation? ________

b. the lowest mean? ________

c. the largest number of items? ________

d. a negatively skewed distribution? ________

C3. In test C, a raw score of:

a. 11 equals what z score? ________

b. 7 equals what T score? ________

c. 19 equals what CEEB score? ________

[<u>Source</u>: Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, N.J.: Prentice Hall Regents, pp. 148-149]

References

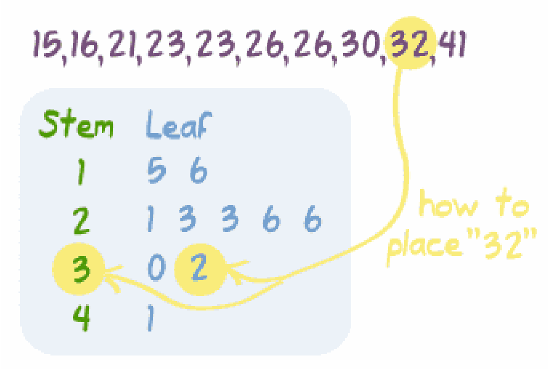
Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, N.J.: Prentice Hall Regents.

Rumsey, D.J. (2011). *Statistics for dummies*. 2nd Edition. Hoboken, N.J.: Wiley Publishing, Inc.

http://www.regentsprep.org/regents/math/algebra/ad6/quartiles.htm

Stem and Leaf Plots

A **Stem and Leaf Plot** is a special table where each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit). Like in this example:



The "stem" values are listed down, and the "leaf" values go right (or left) from the stem values. The "stem" is used to group the scores and each "leaf" shows the individual scores within each group.

Learning Objectives

1. Create and interpret basic stem and leaf displays
2. Create and interpret back-to-back stem and leaf displays
3. Judge whether a stem and leaf display is appropriate for a given data set

A stem and leaf display is a graphical method of displaying data. It is particularly useful when your data are not too numerous. In this section, we will explain how to construct and interpret this kind of graph.

As usual, an example will get us started. Consider Table 1 that shows the number of touchdown passes (TD passes) thrown by each of the 31 teams in the National Football League in the 2000 season.

Table 1. Number of touchdown passes.

37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6

A stem and leaf display of the data is shown in Figure 1. The left portion of Figure 1 contains the stems. They are the numbers 3, 2, 1, and 0, arranged as a column to the left of the bars. Think of these numbers as 10's digits. A stem of 3, for example, can be used to represent the 10's digit in any of the numbers from 30 to 39. The numbers to the right of the bar are leaves, and they represent the 1's digits. Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem.

```
3|2337
2|001112223889
1|2244456888899
0|69
```

Figure 1. Stem and leaf display of the number of touchdown passes.

To make this clear, let us examine Figure 1 more closely. In the top row, the four leaves to the right of stem 3 are 2, 3, 3, and 7. Combined with the stem, these leaves represent the numbers 32, 33, 33, and 37, which are the numbers of TD passes for the first four teams in Table 1. The next row has a stem of 2 and 12 leaves. Together, they represent 12 data points, namely, two occurrences of 20 TD passes, three occurrences of 21 TD passes, three occurrences of 22 TD passes, one occurrence of 23 TD passes, two occurrences of 28 TD passes, and one occurrence of 29 TD passes. We leave it to you to figure out what the third row represents. The fourth row has a stem of 0 and two leaves. It stands for the last two entries in Table 1, namely 9 TD passes and 6 TD passes. (The latter two numbers may be thought of as 09 and 06.)

One purpose of a stem and leaf display is to clarify the shape of the distribution. You can see many facts about TD passes more easily in Figure 1 than in Table 1. For example, by looking at the stems and the shape of the plot, you can tell that most of the teams had between 10 and 29 passing TDs, with a few having more and a few having less. The precise numbers of TD passes can be determined by examining the leaves.

We can make our figure even more revealing by splitting each stem into two parts. Figure 2 shows how to do this. The top row is reserved for numbers from 35 to 39 and holds only the 37 TD passes made by the first team in Table 1. The second row is reserved for the numbers from 30 to 34 and holds the 32, 33, and 33 TD passes made by the next three teams in the table. You can see for yourself what the other rows represent.

3	7
3	233
2	889
2	001112223
1	56888899
1	22444
0	69

Figure 2. Stem and leaf display with the stems split in two.

Figure 2 is more revealing than Figure 1 because the latter figure lumps too many values into a single row. Whether you should split stems in a display depends on the exact form of your data. If rows get too long with single stems, you might try splitting them into two or more parts.

There is a variation of stem and leaf displays that is useful for comparing distributions. The two distributions are placed back to back along a common column of stems. The result is a "back-to-back stem and leaf graph." Figure 3 shows such a graph. It compares the numbers of TD passes in the 1998 and 2000 seasons. The stems are in the middle, the leaves to the left are for the 1998 data, and the leaves to the right are for the 2000 data. For example, the second-to-last row shows that in 1998 there were teams with 11, 12, and 13 TD passes, and in 2000 there were two teams with 12 and three teams with 14 TD passes.

11	4	
	3	7
332	3	233
8865	2	889
44331110	2	001112223
987776665	1	56888899
321	1	22444
7	0	69

Figure 3. Back-to-back stem and leaf display. The left side shows the 1998 TD data and the right side shows the 2000 TD data.

Figure 3 helps us see that the two seasons were similar, but that only in 1998 did any teams throw more than 40 TD passes.

There are two things about the football data that make them easy to graph with stems and leaves. First, the data are limited to whole numbers that can be represented with a one-digit stem and a one-digit leaf. Second, all the numbers are positive. If the data include numbers with three or more digits, or contain decimals, they can be rounded to two-digit accuracy. Negative values are also easily handled. Let us look at another example.

Table 2 shows data from the case study Weapons and Aggression. Each value is the mean difference over a series of trials between the times it took an experimental subject to name aggressive words (like “punch”) under two conditions. In one condition, the words were preceded by a non-weapon word such as "bug." In the second condition, the same words were preceded by a weapon word such as "gun" or "knife." The issue addressed by the experiment was whether a preceding weapon word would speed up (or prime) pronunciation of the aggressive word compared to a non-weapon priming word. A positive difference implies greater priming of the aggressive word by the weapon word. Negative differences imply that the priming by the weapon word was less than for a neutral word.

Table 2. The effects of priming (thousandths of a second).

43.2, 42.9, 35.6, 25.6, 25.4, 23.6, 20.5, 19.9, 14.4, 12.7, 11.3, 10.2, 10.0, 9.1, 7.5, 5.4, 4.7, 3.8, 2.1, 1.2, -0.2, -6.3, -6.7, -8.8, -10.4, -10.5, -14.9, -14.9, -15.0, -18.5, -27.4

You see that the numbers range from 43.2 to -27.4. The first value indicates that one subject was 43.2 milliseconds faster pronouncing aggressive words when they were preceded by weapon words than when preceded by neutral words. The value -27.4 indicates that another subject was 27.4 milliseconds slower pronouncing aggressive words when they were preceded by weapon words.

The data are displayed with stems and leaves in Figure 4. Since stem and leaf displays can only portray two whole digits (one for the stem and one for the leaf), the numbers are first rounded. Thus, the value 43.2 is rounded to 43 and represented with a stem of 4 and a leaf of 3. Similarly, 42.9 is rounded to 43. To represent negative numbers, we simply use negative stems. For example, the bottom row of the figure represents the number -27. The second-to-last row represents the numbers -10, -10, -15, etc. Once again, we have rounded the original values from Table 2.

```
 4|33
 3|6
 2|00456
 1|00134
 0|1245589
-0|0679
-1|005559
-2|7
```

Figure 4. Stem and leaf display with negative numbers and rounding.

Observe that the figure contains a row headed by "0" and another headed by "-0." The stem of 0 is for numbers between 0 and 9, whereas the stem of –0 is for numbers between 0 and -9. For example, the fifth row of the table holds the numbers 1, 2, 4, 5, 5, 8, 9 and the sixth row holds 0, -6, -7, and -9. Values that are exactly 0 before rounding should be split as evenly as possible between the "0" and "-0" rows. In Table 2, none of the values are 0 before rounding. The "0" that appears in the "-0" row comes from the original value of -0.2 in the table.

Although stem and leaf displays are unwieldy for large data sets, they are often useful for data sets with up to 200 observations. Figure 5 portrays the distribution of populations of 185 US cities in 1998. To be included, a city had to have between 100,000 and 500,000 residents.

```
4|899
4|6
4|4455
4|333
4|01
3|99
3|677777
3|55
3|223
3|111
2|8899
2|666667
2|444455
2|22333
2|000000
1|8888888888889999999999
1|666666777777
1|44444444444455555555555
1|222222222222222222333333333
1|00000000000000111111111111111111111111111
```

Figure 5. Stem and leaf display of populations of 185 US cities with populations between 100,000 and 500,000 in 1998.

Since a stem and leaf plot shows only two-place accuracy, we had to round the numbers to the nearest 10,000. For example, the largest number (493,559) was rounded to 490,000 and then plotted with a stem of 4 and a leaf of 9. The fourth highest number (463,201) was rounded to 460,000 and plotted with a stem of 4 and a leaf of 6. Thus, the stems represent units of 100,000 and the leaves represent units of 10,000. Notice that each stem value is split into five parts: 0-1, 2-3, 4-5, 6-7, and 8-9.

Whether your data can be suitably represented by a stem and leaf graph depends on whether they can be rounded without loss of important information. Also, their extreme values must fit into two successive digits, as the data in Figure 5 fit into the 10,000 and 100,000 places (for leaves and stems, respectively). Deciding what kind of graph is best suited to displaying your data thus requires good judgment. Statistics is not just recipes!

http://onlinestatbook.com/2/graphing_distributions/stem.html

Using SPSS to Display Stem and Leaf Plots

Analyze → Descriptive Statistics → Explore → Plots → Descriptive → check Stem-and-Leaf

Exercises

Stem	Leaf
1	2 3
2	0 1 5 5 8
3	4 6 9
5	3 4 4 4 6
6	2 5 6 6 7 8
8	0 3 5
9	8

1. What is the range for the above stem and leaf plot?
2. For the above stem and leaf plot, what is the median value?
3. For the above stem and leaf plot, what is the mean value?
4. The highest value in the dataset is:

```
13|2
12|26
11|03457
10|011222445556788
 9|001112334445555666777789
 8|000011223333344444555666788999999
 7|000011111366777888
 6|00012223444455556677789999
 5|1355778899
 4|4789
 3|456689
 2|59
 1|9

Multiply stems by 10.0.
```

5. How many scores are between 40 and 50?

```
13|2
12|26
11|03457
10|011222445556788
 9|001112334445555666777789
 8|000011223333344444555666788999999
 7|000011111366777888
 6|00012223444455556677789999
 5|1355778899
 4|4789
 3|456689
 2|59
 1|9

Multiply stems by 10.0.
```

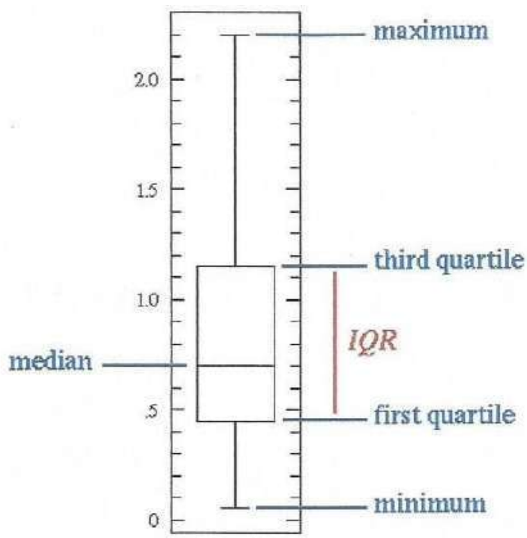
6. What are the smallest value and the largest value?

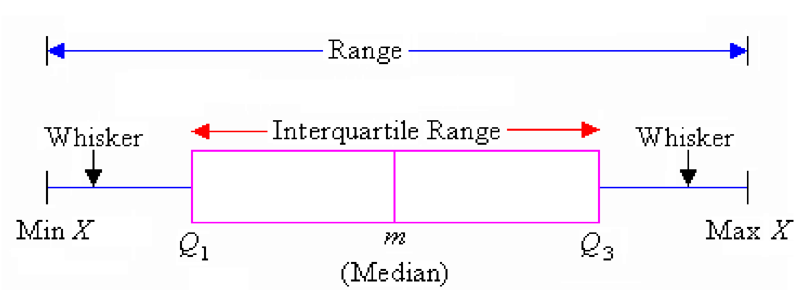
```
 6|0
 5|
 4|6
 3|2244
 2|1688
 1|23667
 0|457
-0|1234679
-1|011
-2|2568

Multiply stems by 1.0.
```

Box Plots

The box plot (a.k.a. box and whisker diagram) is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum. In the simplest box plot the central rectangle spans the first quartile to the third quartile (the *interquartile range* or *IQR*). A segment inside the rectangle shows the median and "whiskers" above and below the box show the locations of the minimum and maximum.

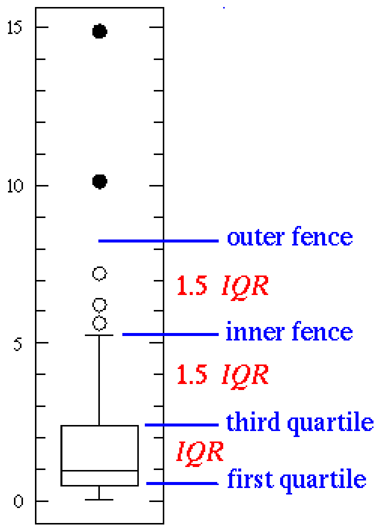




Not uncommonly real datasets will display surprisingly high maximums or surprisingly low minimums called *outliers*. John Tukey has provided a precise definition for two types of outliers:

- **Outliers** are either 3×*IQR* or more above the third quartile or 3×*IQR* or more below the first quartile.

- **Suspected outliers** are slightly more central versions of outliers: either 1.5×*IQR* or more above the third quartile or 1.5×*IQR* or more below the first quartile.



- **Draw a box-and-whisker plot for the following data set:**

  4.3, 5.1, 3.9, 4.5, 4.4, 4.9, 5.0, 4.7, 4.1, 4.6, 4.4, 4.3, 4.8, 4.4, 4.2, 4.5, 4.4

  My first step is to order the set. This gives me:

  3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

  The first number I need is the median of the entire set. Since there are seventeen values in this list, I need the ninth value:

  3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

  The median is $Q_2 = 4.4$.

  The next two numbers I need are the medians of the two halves. Since I used the "4.4" in the middle of the list, I can't re-use it, so my two remaining data sets are:

  3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4 and 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

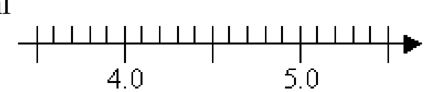
  The first half has eight values, so the median is the average of the middle two:

  $$Q_1 = (4.3 + 4.3)/2 = 4.3$$

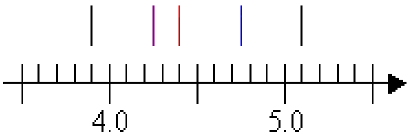
  The median of the second half is:

  $$Q_3 = (4.7 + 4.8)/2 = 4.75$$

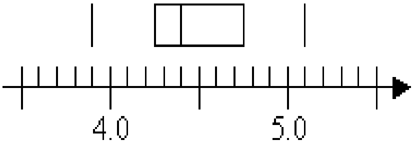
Since my list values have one decimal place and range from 3.9 to 5.1, I won't use a scale of, say, zero to ten, marked off by ones. Instead, I'll draw a number line from 3.5 to 5.5, and mark off by tenths.



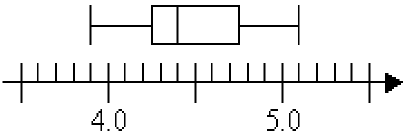
Now I'll mark off the minimum and maximum values, and Q_1, Q_2, and Q_3:



The "box" part of the plot goes from Q_1 to Q_3:



And then the "whiskers" are drawn to the endpoints:



By the way, box-and-whisker plots don't have to be drawn horizontally; they can be vertical, too.

Using Excel to Calculate Quartiles

Using formulas

Formulas→*fx* insert function→select a function

Quartile1 : =QUARTILE(B2:B11,1)

Quartile2 : =QUARTILE(B2:B11,2)

Quartile3 : =QUARTILE(B2:B11,3)

Using SPSS to Display Boxplots

Analyze → Descriptive Statistics → Explore → Plots → Boxplots → check Factor levels together

Exercises

Draw a boxplot for each data set of scores:

1. 76, 79, 76, 74, 75, 71, 85, 82, 82, 79, 81
2. 85, 78, 23, 33, 35, 75, 65, 68, 47, 41, 64, 48, 54, 53, 55, 58
3. 72, 48, 47, 35, 33, 36, 38, 34, 46, 43, 55, 59, 58, 66, 64, 11, 15, 23, 25, 25, 18, 20, 28, 26

http://www.physics.csbsju.edu/stats/box2.html

Standard Error

The standard deviation of the sampling distribution of a statistic. Standard error is a statistical term that measures the accuracy with which a sample represents a population. In statistics, a sample mean deviates from the actual mean of a population; this deviation is the standard error. The term "standard error" is used to refer to the standard deviation of various sample statistics such as the mean or median. For example, the "standard error of the mean" refers to the standard deviation of the distribution of sample means taken from a population.

The smaller the standard error, the more representative the sample will be of the overall population. The standard error is also inversely proportional to the sample size; the larger the sample size, the smaller the standard error because the statistic will approach the actual value.

If you measure a sample from a wider population, the average (or mean) of the sample will be an *approximation* of the population mean. But how accurate is this?

If you measure multiple samples, their means will not all be the same, and will be spread out in a distribution (although not as much as the population). Due to the central limit theorem, the means will be spread in an approximately normal, bell-shaped distribution.

The standard error, or *standard error of the mean*, of multiple samples is the standard deviation of the sample means, and thus gives a measure of their spread. Thus 68% of all sample means will be within one standard error of the population mean (and 95% within two standard errors).

What the standard error gives in particular is an indication of the likely accuracy of the sample mean as compared with the population mean. The smaller the standard error, the less the spread and the more likely it is that any sample mean is close to the population mean. A small standard error is thus a Good Thing.

When there are fewer samples, or even one, then the standard error, (typically denoted by SE or SEM) can be estimated as the standard deviation of the sample (a set of measures of x), divided by the square root of the sample size (n):

$SE = stdev(x_i) / sqrt(n)$

Example

This shows four samples of increasing size. Note how the standard error reduces with increasing sample size.

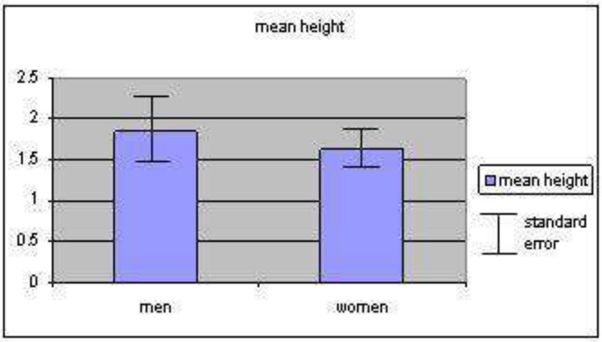
	Sample 1	Sample 2	Sample 3	Sample 4
	9	6	5	8
	2	6	3	1
	1	8	6	7
		8	4	1
		3	7	3
		8	2	3
			6	4
			9	7
			7	1
			1	8
			1	9
			7	9
				3
				1
				6
				8
				3
				4
Mean:	4.00	6.50	4.83	4.78
Std dev, s:	4.36	1.97	2.62	2.96
Sample size, n:	3	6	12	18
sqrt(n):	1.73	2.45	3.46	4.24
Standard error, s/sqrt(n):	2.52	0.81	0.76	0.70

Discussion

The standard error gives a measure of how well a sample represents the population. When the sample is representative, the standard error will be small. The division by the square root of the sample size is a reflection of the speed with which an increasing sample size gives an improved representation of the population, as in the example above.

An approximation of confidence intervals can be made using the mean +/- standard errors. Thus, in the above example, in Sample 4 there is a 95% chance that the population mean is within +/- 1.4 (=2*0.70) of the mean (4.78).

Graphs that show sample means may have the standard error highlighted by an 'I' bar (sometimes called an *error bar*) going up and down from the mean, thus indicating the spread, for example as below:



http://changingminds.org/explanations/research/statistics/standard_error.htm

Standard Error of the Mean

The standard error of the mean is the standard deviation of the sample mean estimate of a population mean. It is usually calculated by the sample estimate of the population standard deviation (sample standard deviation) divided by the square root of the sample size (assuming statistical independence of the values in the sample):

$$SEM = \frac{s}{\sqrt{n}}$$

Where:

SEM = standard error of the mean

s = sample standard deviation (see formula below)

n = size (number of observations) of the sample

The following is the sample standard deviation formula:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Where:
s = sample standard deviation
$x_1, ..., x_N$ = the sample data set
$\bar{x}$ = mean value of the sample data set
N = size of the sample data set

Using Excel to Calculate Standard Error
Standard error : =(STDEV(B2:B21))/(SQRT(COUNT(B2:B21)))

Distributions

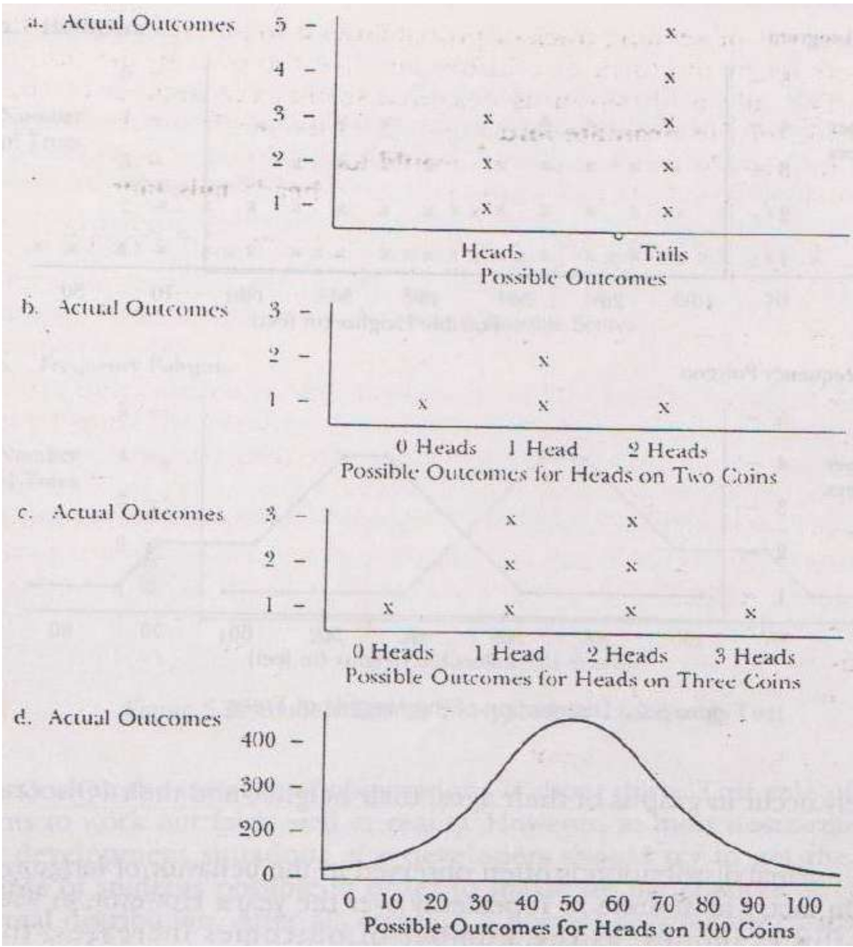


Figure: Histograms of Coin Flips

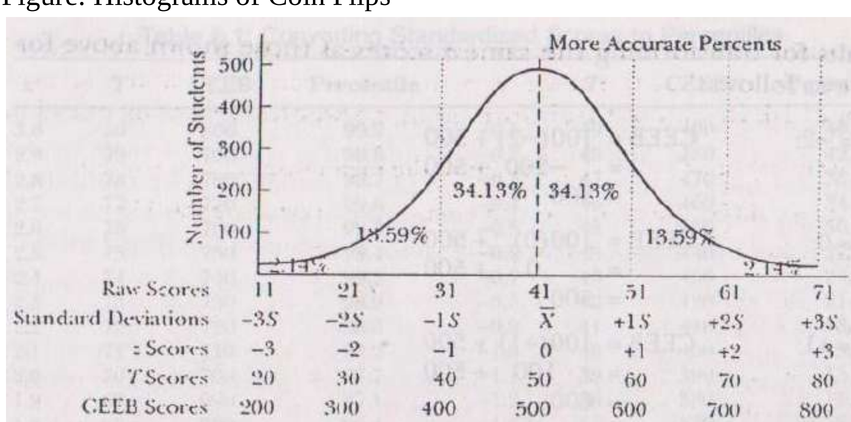


Figure: Comparison of Standard Score Distributions

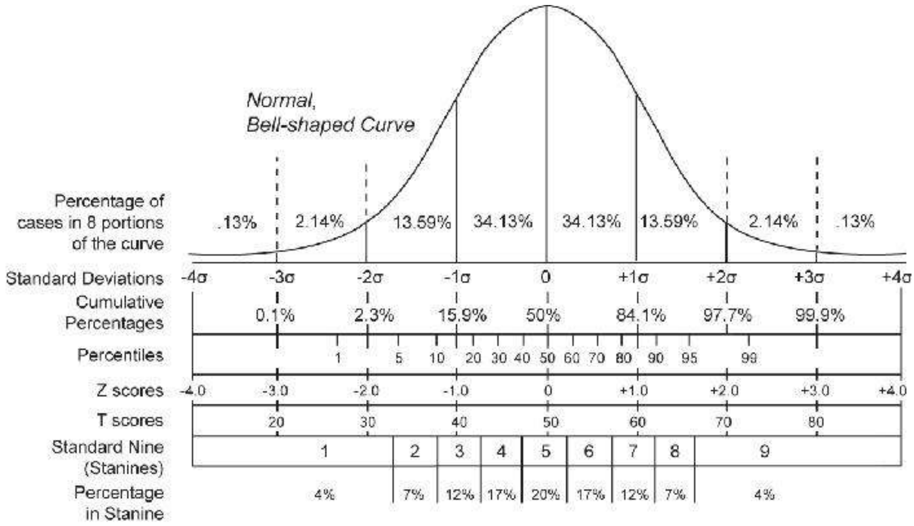
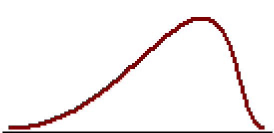
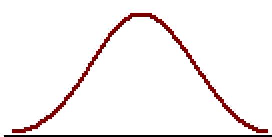


Figure: Comparison of Standard Score Distributions

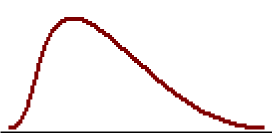
The coefficient of Skewness is a measure for the degree of symmetry in the variable distribution.



Negatively skewed distribution or Skewed to the left Skewness <0



Normal distribution Symmetrical Skewness = 0



Positively skewed distribution or Skewed to the right Skewness > 0

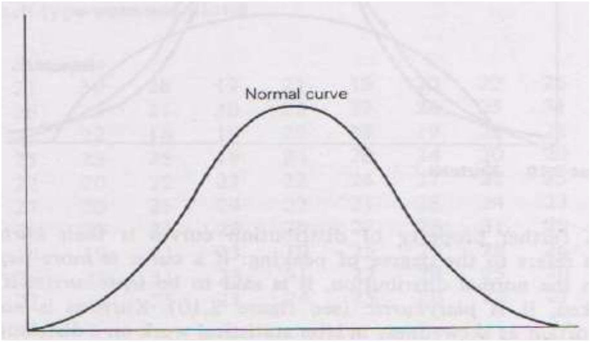


Figure: Normal distribution curve

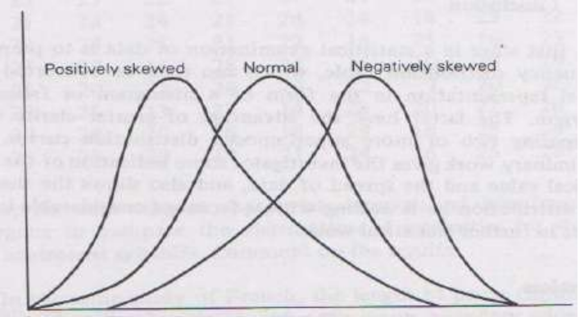


Figure: Skewed distribution

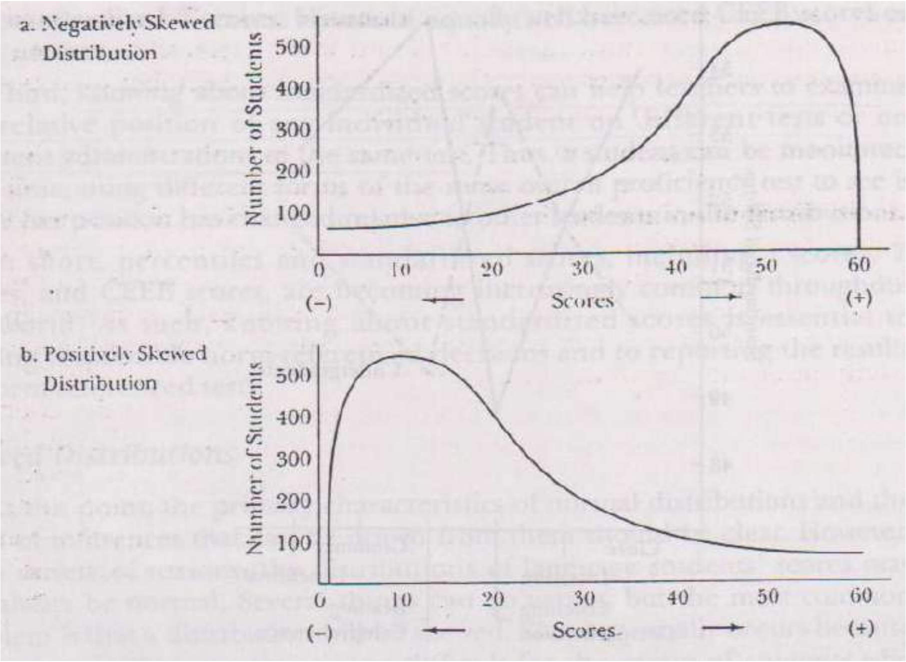


Figure: Skewed Distributions

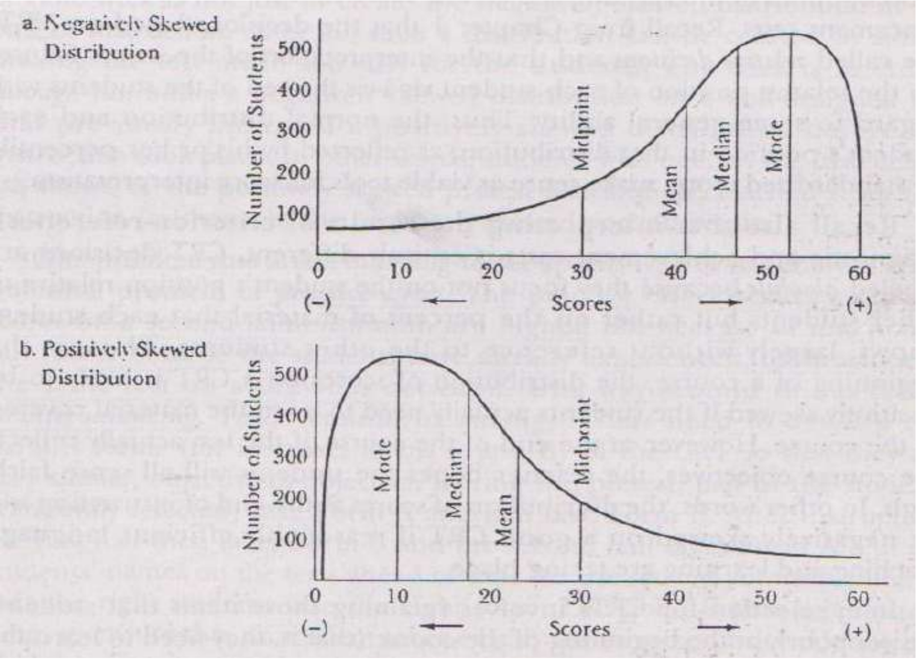


Figure: Indicators of Central Tendency in Skewed Distributions

Pearson's Index of Skewness

Pearson's index of skewness can be used to determine whether the data is symmetric or skewed. If the index is between -1 and 1, then the distribution is symmetric. If the index is less than -1 then it is skewed to the left and if it is more than 1, then it is skewed to the right.

$$\text{Pearson's Index of Skewness} = P = \frac{3(\bar{x} - \text{median})}{s}$$

Chebyshev's Rule

A rule that states the minimum amount of data that will like within *k* (*k*>1) standard deviations of the mean for any distribution of data. There will be at least 3/4 (75%) of the data within 2 standard deviations of the mean and at least 8/9 (89%) of the data within 3 standard deviations of the mean.

The rule states

At least $1-\frac{1}{k^2},\ (k>1)$ of the data will like within *k* standard deviations of the mean

To verify the rule

1. Rank the data from lowest to highest. This is not necessary, but it makes it easier.
2. Find the mean and standard deviation of the data.
3. Find the lower boundary by multiplying the standard deviation by *k* and subtracting from the mean.
4. Find the upper boundary by multiplying the standard deviation by *k* and adding it to the mean.
5. Count the number of values between these two boundaries.
6. Divide the number of values between the boundaries by the total number of values

Kurtosis

The coefficient of Kurtosis is a measure for the degree of peakedness/ flatness in the variable distribution.

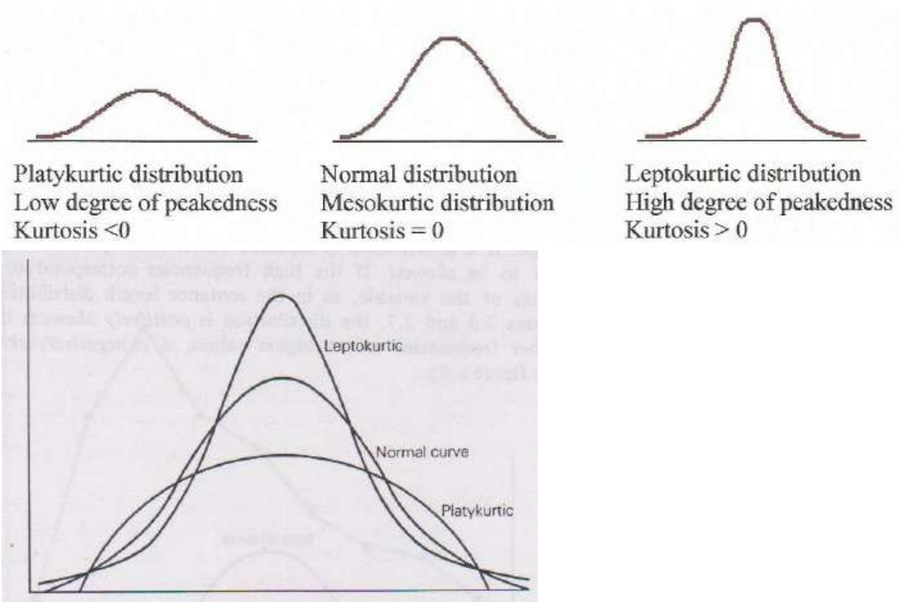


Figure: Kurtosis

Kurtosis is the degree of peakedness in a distribution, usually taken relative to a normal distribution. The peakedness property means that there is an excess frequency at the center of the distribution. The index of kurtosis is given by:

$$\delta = E\,\{(x\text{-}\mu)^4\} / \sigma^4 - 3$$

where $E\ \{\ (x\text{-}\mu)^4\ \}$ is the fourth moment and σ is the standard deviation of the distribution. Since the value of $E\ \{(x\text{-}\mu)^4\}$ is equal to 3 for a normal distribution, this index measures kurtosis relative to a normal distribution. Positive values of δ indicate longer thicker tails than a normal distribution, whereas the negative values of δ indicate shorter thinner tails. A distribution with positive kurtosis is called *Leptokurtic*, and a distribution with negative kurtosis is called *Platokurtic*. When $\delta = 0$, the distribution is called *Mesokurtic.*

Using Excel to Calculate Skewness and Kurtosis

Skewness : =SKEW(B2:B21)

Kurtosis : =KURT(B2:B21)

Using Excel to Display Histograms

a. Arrange the data in terms of Score Interval and Frequency.
b. Insert→Charts→Column
c. Layout→Data Labels, Trendline→More Trendline Options→Polynomial

Using SPSS to Display Histograms

Analyze→Descriptive Statistics→Explore→Plots→Descriptive→check Histogram

Double click the histogram to show the Chart Editor, then click Distribution Curve to show the normal distribution curve.

Exercise

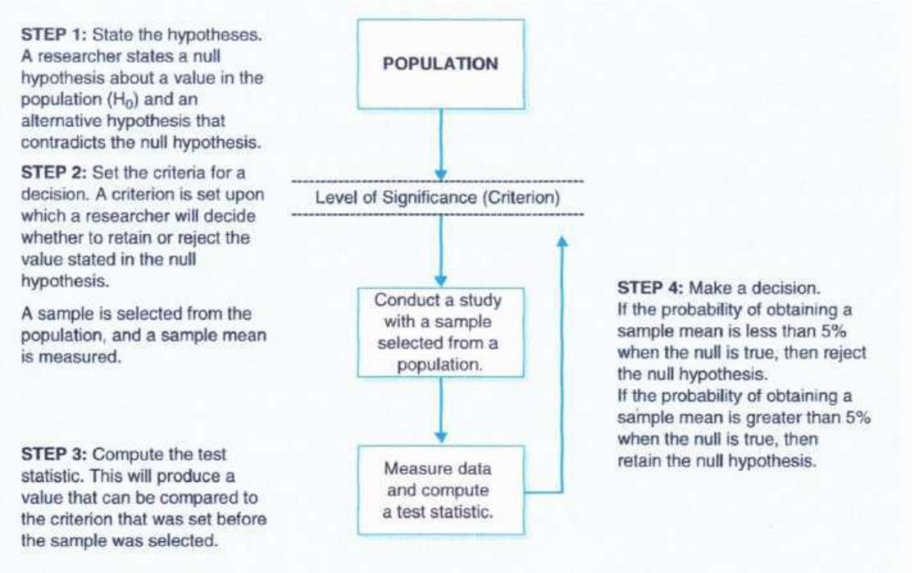
Draw the histogram for each data set.

1. 85, 78, 23, 33, 35, 75, 65, 68, 47, 41, 64, 48, 54, 53, 55, 58
2. 72, 48, 47, 35, 33, 36, 38, 34, 46, 43, 55, 59, 58, 66, 64, 11, 15, 23, 25, 25, 18, 20, 28, 26
3. 62, 72, 53, 44, 34, 22, 25, 35, 36, 44, 45, 55, 54, 55, 47, 57, 65, 75, 65, 68, 77, 78, 66, 16

Hypothesis Testing

Hypothesis testing, or **significance testing,** a method of testing a claim or hypothesis about a parameter in a population, using data measured in a sample. In this method, we test some hypothesis by determining the likelihood that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true.

The four steps of hypothesis testing are as follows:
– Step 1: State the hypotheses.
– Step 2: Set the criteria for a decision.
– Step 3: Compute the test statistic.
– Step 4: Make a decision.



The **null hypothesis (H_0),** stated as the **null,** is a statement about a population parameter, such as the population mean, that is assumed to be true. The null hypothesis is a starting point. We will test whether the value stated in the null hypothesis is likely to be true.

An **alternative hypothesis (H_1)** is a statement that directly contradicts a null hypothesis by stating that the actual value of a population parameter, such as the mean, is less than, greater than, or not equal to the value stated in the null hypothesis. The alternative hypothesis states what we think is wrong about the null hypothesis.

Level of significance refers to a criterion of judgment upon which a decision is made regarding the value stated in a null hypothesis. The criterion is based on the probability of obtaining a statistic measured in a sample if the value stated in the null hypothesis were true. In behavioral science, the criterion or level of significance is typically set at 5%. When the probability of obtaining a sample mean is less than 5% if the null hypothesis were true, then we reject the value stated in the null hypothesis. An **alpha (α) level** is the level of significance or criterion for a hypothesis test. It is the largest probability of committing a Type I error that we will allow and still decide to reject the null hypothesis.

The **test statistic** is a mathematical formula that allows researchers to determine the likelihood or probability of obtaining sample outcomes if the null hypothesis were true. The value of a test statistic can be used to make inferences concerning the value of population parameters stated in the null hypothesis.

A ***p*-value** is the probability of obtaining a sample outcome, given that the value stated in the null hypothesis is true. The *p*-value of a sample outcome is compared to the level of significance. When the *p*-value is less than 5% ($p < .05$), we reject the null hypothesis. We will refer to $p < .05$ as the criterion for deciding to reject the null hypothesis, although note that when $p = .05$, the decision is also to reject the null hypothesis. When the *p*-value is greater than 5% ($p > .05$), we retain the null hypothesis. The decision to reject or retain the null hypothesis is called **significance.** When the *p*- value is less than .05, we reach significance; the decision is to reject the null hypothesis. When the *p*-value is greater than .05, we fail to reach significance; the decision is to retain the null hypothesis. Researchers make decisions regarding the null hypothesis. The decision can be to retain the null ($p > .05$) or reject the null ($p < .05$).

Significance, or **statistical significance,** describes a decision made concerning a value stated in the null hypothesis. When a null hypothesis is rejected, a result is significant. When a null hypothesis is retained, a result is not significant.

The **degrees of freedom** (**df**) of an estimate is the number of independent pieces of information on which the estimate is based. The number of degrees of freedom generally refers to the number of independent observations in a sample minus the number of population parameters that must be estimated from sample data. Statisticians use the term "degrees of freedom" to describe the number of values in the final calculation of a statistic that are free to vary.

Critical values, which mark the cutoffs for the **rejection region,** can be identified for any level of significance. The value of the test statistic is compared to the critical values. When the value of a test statistic exceeds a critical value, we reject the null hypothesis; otherwise, we retain the null hypothesis.

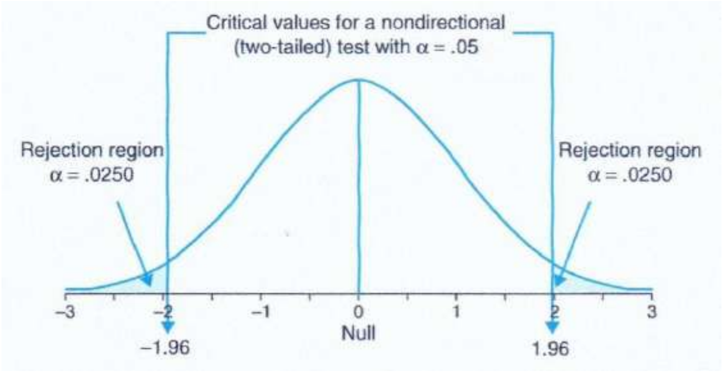
The **obtained value** is the value of a test statistic. This value is compared to the critical value(s) of a hypothesis test to make a decision. When the obtained value exceeds a critical value, we decide to reject the null hypothesis; otherwise, we retain the null hypothesis.
The **rejection region** is the region beyond a critical value in a hypothesis test. When the value of a test statistic is in the rejection region, we decide to reject the null hypothesis; otherwise, we retain the null hypothesis.

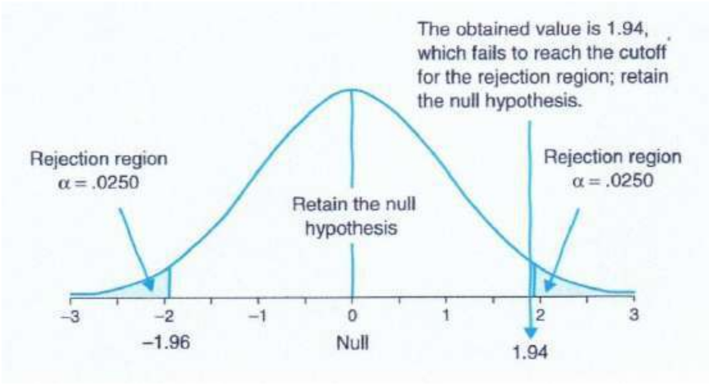
We can decide to retain or reject the null hypothesis, and this decision can be correct or incorrect. Two types of errors in hypothesis testing are called Type I and Type II errors. A **Type I error** is the probability of rejecting a null hypothesis that is actually true. The probability of this type of error is determined by the researcher and stated as the level of significance or alpha level for a hypothesis test. A **Type II error** is the probability of retaining a null hypothesis that is actually false.

Table: Four outcomes for making a decision. The decision can be either correct (correctly reject or retain null) or wrong (incorrectly reject or retain null).

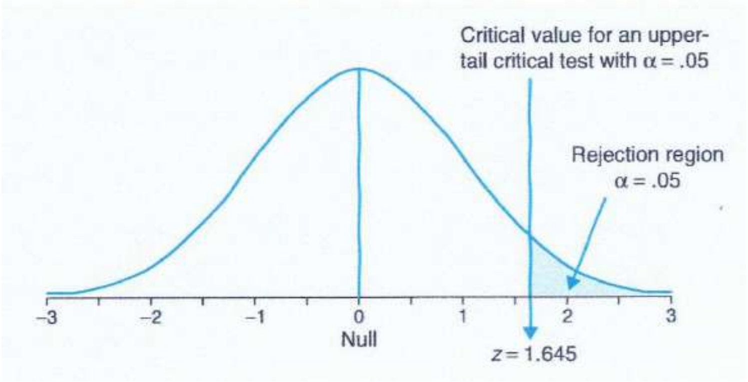
		Decision	
		Retain the null	**Reject the null**
Truth in the population	True	CORRECT $1-\alpha$	TYPE I ERROR α
	False	TYPE II ERROR β	CORRECT $1-\beta$ POWER

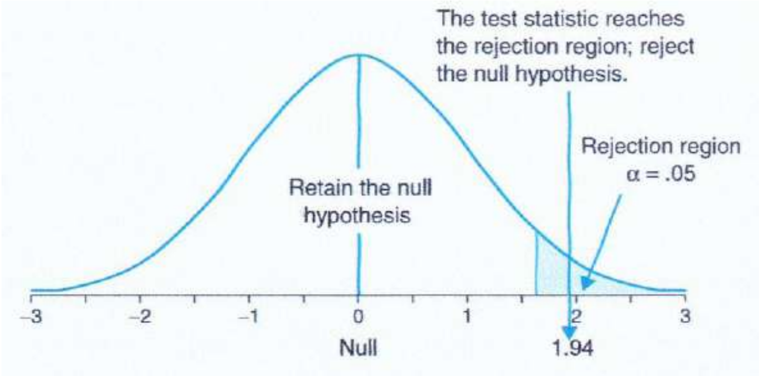
Nondirectional (two-tailed) tests are hypothesis tests where the alternative hypothesis is stated as *not equal to* (≠). So we are interested in any alternative from the null hypothesis.

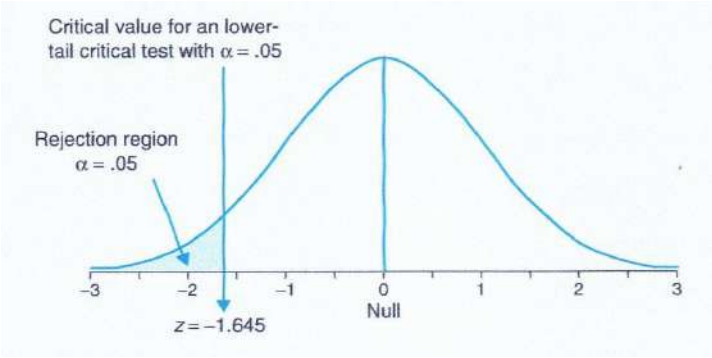




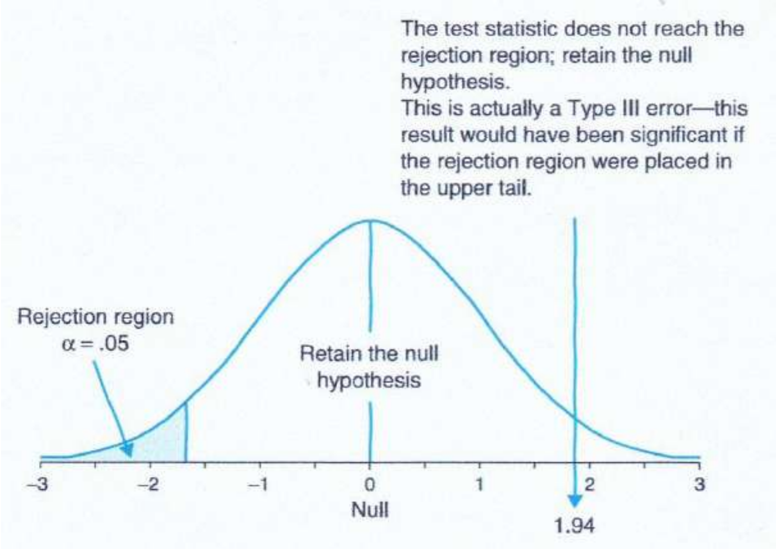
Directional (one-tailed) tests are hypothesis tests where the alternative hypothesis is stated as *greater than* (>) or *less than* (<) some value. So we are interested in a specific alternative from the null hypothesis.







A **Type III error** occurs for one-tailed tests where a result would have been significant in one tail, but the researcher retains the null hypothesis because the rejection region was placed in the wrong or opposite tail. The "wrong tail" refers to the opposite tail from where a difference was observed and would have otherwise been significant.



http://www.sagepub.com/sites/default/files/upm-binaries/40007_Chapter8.pdf

Examples

The four steps of hypothesis testing are as follows:

– Step 1: State the hypotheses.
– Step 2: Set the criteria for a decision.

(a) the critical-value approach

- If the obtained-value ≥ the critical value, H_0 is rejected. Therefore, H_1 is accepted.
- If the obtained-value < the critical value, H_0 is accepted.

(b) the ρ-value approach

- If the ρ-value ≤ the significance level (α = 0.05/0.01), H0 is rejected. Therefore, H1 is accepted
- If the ρ-value > the significance level (α = 0.05/0.01), H_0 is accepted.
-

– Step 3: Compute the test statistic.
– Step 4: Make a decision on the basis of the criteria.

NORMALITY

(1) Formulating the hypotheses

H_0 : The data set has the normal distribution.

H_1 : The data set does not have the normal distribution.

(2) Setting the criteria for a decision

(a) the critical-value approach

- If the obtained-value ≥ the critical value, H_0 is rejected. Therefore, H_1 is accepted.
- If the obtained-value < the critical value, H_0 is accepted.

(b) the ρ-value approach

- If the ρ-value ≤ the significance level (α = 0.05), H0 is rejected. Therefore, H1 is accepted
- If the ρ-value > the significance level (α = 0.05), H0 is accepted.

(3) Calculating the data by using, for example, Kolmogorov-Smirnov test or Shapiro-Wilk test.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Data1	.168	10	.200*	.942	10	.573

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

(4) Accepting/rejecting the null hypothesis on the basis of the results of the test:

(b) *the ρ-value approach*

Based on Kolmogorov-Smirnov test, the sig. value is 0.200. Because it is > 0.05, H_0 is accepted. It means that the dataset is considered normally distributed. Likewise, based on Shapiro-Wilk test, the sig. value is 0.573. Because it is > 0.05, H_0 is accepted. It means that the dataset is considered normally distributed.

HOMOGENEITY

(1) Formulating the hypotheses

H_0 : The two data sets have the same distribution.

H_1 : The two data sets do not have the same distribution.

(2) Setting the criteria for a decision

(3) Calculating the two data sets by using, for example, the *F*-test.

(a) For example: $F_{Stat} = 1.6350$. $F_{Critical\ 0.05,\ 39,\ 29} = 1.80$

(b) For example:

Test of Homogeneity of Variances

Data1

Levene Statistic	df1	df2	Sig.
.022	1	38	.882

(4) Accepting/rejecting the null hypothesis on the basis of the results of the test:

(a) *the critical-value approach*

F_{Stat} = 1.6350. At the significance level 0.10 in two-tailed testing, df in numerator=39 and df in denominator=29, $F_{Critical}$ = 1.80. F_{Stat} (1.6350) < F_{Crit} (1.80), therefore, H_0 is accepted. It means that the two data sets have *the same distribution*.

(b) *the p-value approach*

Based on the output of Levene's test, the sig. value is 0.882. Because it is > 0.05, H_0 is retained. It means that the datasets have the same variance or the same distribution.

LINEARITY

(1) Formulating the hypotheses

H_0 : The two data sets have the linear relationship.

H_1 : The two data sets do not have the linear relationship.

(2) Setting the criteria for a decision

(3) Calculating the two data sets by using, for example, ANOVA.

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Data1 * Data4	Between Groups	(Combined)	4242.950	16	265.184	.533	.826
		Linearity	3136.455	1	3136.455	6.307	.087
		Deviation from Linearity	1106.495	15	73.766	.148	.996
	Within Groups		1492.000	3	497.333		
	Total		5734.950	19			

(4) Accepting/rejecting the null hypothesis on the basis of the results of the test:

(b) *the ρ-value approach*

Based on the ANOVA Output Table, the significance value of Deviation from Linearity is 0.996. Because it is > 0.05, H_0 is retained. It means that the datasets have the linear relationship or the relationship between the variables is linear.

CORRELATION

(1) Formulating the hypotheses

H_0 : There is no significant correlation between *X* and *Y*.

H_1 : There is a significant correlation between *X* and *Y*.

H_0 : There is no significant *positive* correlation between *X* and *Y*.

H_1 : There is a significant *positive* correlation between *X* and *Y*.

H_0 : There is no significant *negative* correlation between *X* and *Y*.

H_1 : There is a significant *negative* correlation between *X* and *Y*.

(2) Setting the criteria for a decision

(3) Calculating the data by using, say, the Pearson product-moment correlation coefficient test. For example: the *r*-obtained = 0.809, the ρ-value = 0.005

		Grammar	Reading
Grammar	Pearson Correlation	1	.809**
	Sig. (2-tailed)		.005
	N	10	10
Reading	Pearson Correlation	.809**	1
	Sig. (2-tailed)	.005	
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

(4) Accepting/rejecting the null hypothesis on the basis of the results of the test:

(a) the critical-value approach

The *r*-obtained = 0.809

At the significance level 0.01 in two-tailed testing and n=10, the *r*-table = 0.765.

Because the *r*-obtained (0.809) > the r-table (0.765), H_0 is rejected. Therefore, H_1 is accepted. It means that there is a *significant correlation* between grammar and reading.

(b) the ρ-value approach

The ρ-value = 0.005

Because the ρ-value (0.005) < the significance level (0.01), H_0 is rejected.

Therefore, H_1 is accepted. There is a *significant correlation* between grammar and reading.

REGRESSION

(1) Formulating the hypotheses

H_0 : There is no significant influence of *X* on *Y*.
H_1 : There is a significant influence of *X* on *Y*.

H_0 : There is no significant *positive* influence of *X* on *Y*.
H_1 : There is a significant *positive* influence of *X* on *Y*.

H_0 : There is no significant *negative* influence of *X* on *Y*.
H_1 : There is a significant *negative* influence of *X* on *Y*.

(2) Setting the criteria for a decision

(3) Calculating the data by using, for example, the simple regression (ANOVA).

For example: the F-obtained = 28.223, the ρ-value = 0.000

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	378.829	1	378.829	28.223	.000[b]
	Residual	375.837	28	13.423		
	Total	754.667	29			

a. Dependent Variable: Writing

b. Predictors: (Constant), Vocabulary

(4) Accepting/rejecting the null hypothesis on the basis of the results of the test:

the ρ-value approach

The ρ-value = 0.000

Because the ρ-value (0.000) < the significant level (0.01), H_0 is rejected. Therefore, H_1 is accepted. It means that there is a *significant influence* of *X* on *Y*.

SIGNIFICANCE OF DIFFERENCES

(1) Formulating the hypotheses

Paired samples t-test

H_0 : There is no significant difference in vocabulary achievement before and after the treatment.

H_1 : There is a significant difference in vocabulary achievement before and after the treatment.

H_0 : There is no significant improvement in vocabulary achievement before and after the treatment.

H_1 : There is a significant improvement in vocabulary achievement before and after the treatment.

Independent-samples t-test

H_0 : There is no significant difference in vocabulary achievement between the students who are taught by using the A method and those who are not.

H_1 : There is a significant difference in vocabulary achievement between the students who are taught by using the A method and those who are not.

H_0 : The students who are taught by using the A method have the same vocabulary achievement as the students who are not taught by using the A method.
H_1 : The students who are taught by using the A method have a better vocabulary achievement than those who are not taught by using the A method.

(2) Setting the criteria for a decision

(3) Calculating the data by using, for example, the independent-samples t-test.
For example: the *t*-obtained = 0.912396327, the *ρ*-value = 0.373618943, df=18,
the t-table = 2.101

t-Test: Two-Sample Assuming Equal Variances		
	Group A	*Group B*
Mean	51.36363636	47.3333333
Variance	87.05454545	108.5
Observations	11	9
Pooled Variance	96.58585859	
Hypothesized Mean Difference	0	
df	18	
t Stat	0.912396327	
P(T<=t) one-tail	0.186809472	
t Critical one-tail	1.734063592	
P(T<=t) two-tail	0.373618943	
t Critical two-tail	2.100922037	

(4) Accepting/rejecting the null hypothesis on the basis of the results of the test:

(a) the critical-value approach

The *t*-obtained = 0.912396327
At the significance level 0.05 in two-tailed testing and df=18, the *t*-table = 2.101.

Because the *t*-obtained (0.912396327) < the *t*-table (2.101), H_0 is accepted.

It means that there is *no significant difference* in vocabulary achievement between the students who are taught by using the A method and those who are not.

(b) the ρ-value approach

The *ρ*-value = 0.373618943

Because the *ρ*-value (0.373618943) > the significance level (0.05), H_0 is accepted.

It means that there is *no significant difference* in vocabulary achievement between the students who are taught by using the A method and those who are not.

VALIDITY AND RELIABILITY

Validity and reliability of research instrument(s) are essential in research data collection. **Validity** is defined as the degree to which a research instrument measures what it claims or purports to be measuring. If a test claims to measure speaking proficiency, then the test should measure the ability to speak. An instrument is said to be valid if it can reveal the data of the variables to be studied.

One alternative to test validity is using Pearson Product-Moment Correlation Coefficient by correlating each item's scores with the total scores. There are 2 alternatives in making decision in the validity test

1. Seeing the value of significance
   a. If the significance value is < 0.05, the instrument is declared valid.
   b. If the significance value is > 0.05, the instrument is declared invalid.

2. Comparing the critical value of r-table
   a. If the obtained value is < the critical value of r-table, the instrument is declared invalid.
   b. If the obtained value is > the critical value of r-table, the instrument is declared valid.

For example

Correlations

Item_4	Item_5	Item_6	Item_7	Item_8	Item_9	Item_10	Score_Total
,349*	,289	,248	,373*	,241	,379*	,161	,613
,027	,070	,123	,018	,135	,016	,320	,000
40	40	40	40	40	40	40	40
,419**	,484**	,441**	,522**	,286	,369*	,176	,729**
,007	,002	,004	,001	,074	,019	,276	,000
40	40	40	40	40	40	40	40
,346*	,310	,332*	,239	,079	,640**	,578**	,704**
,029	,052	,036	,137	,629	,000	,000	,000
40	40	40	40	40	40	40	40
1	,571**	,351*	,318*	,308	,147	,201	,642**
	,000	,026	,046	,053	,367	,215	,000
40	40	40	40	40	40	40	40
,571**	1	,641**	,174	,240	,048	,151	,631**
,000		,000	,282	,136	,768	,351	,000
40	40	40	40	40	40	40	40

Based on the output above, the critical value of r-table at the significance level of 5% and N=40 is 0.312. Item 1 is valid because the significance value (0.000) is lower than 0.05 or because the obtained value r_{xy} (0.613) is higher than the critical value (0.312).

Reliability is defined as the extent to which the instrument's results can be considered consistent or stable. For example, if a teacher administers a test to her students on one occasion, she would like the scores to be very much the same if she were to administer the same test again tomorrow. The degree to which a test is consistent or reliable can be estimated by calculating a reliability coefficient. The reliability coefficient can go as low as 0 and as high as 1. One alternative to test the reliability of an instrument is using Cronbach's Alpha which was introduced by Cronbach in 1970. For research purposes, a rule of thumb is that reliability should be at least .70 and preferably higher (Wallen & Fraenkel, 2011, p.101)

For example

Reliability Statistics

Cronbach's Alpha	N of Items
,820	10

From the output of Reliability Statistics, the obtained value of Cronbach's Alpha is 0.820. Because the value is > 0.70, it can be concluded that the research instrument is reliable.

Using SPPS to test Validity and Reliability

Validity : Analyze→Correlate→Bivariate→Input all the data per item and total score per item→check Pearson in Correlation Coefficients, check Two-tailed in Test of Significance, check Flag significant correlations

Reliability: Analyze→Scale→Reliability Analysis→Input all the data per item→Model: Alpha

Exercise

A researcher has made a 10-item questionnaire with 5 alternative answers to the Likert scale: 5=very good, 4=good, 3=fair, 2=poor, 1=very poor. The questionnaire is given to twenty-five respondents. The following are the results of the questionnaire. Find out whether the questionnaire is valid and reliable.

No	Item Number									
	1	2	3	4	5	6	7	8	9	10
1	3	4	4	3	3	3	4	4	4	4
2	5	4	4	3	4	4	4	3	4	4
3	5	5	4	5	5	5	5	4	5	5
4	5	4	3	4	4	4	5	4	4	3
5	5	4	4	4	4	3	4	4	4	3
6	4	4	4	3	3	4	4	3	4	3
7	4	4	4	4	3	4	4	3	4	3
8	4	5	4	4	5	4	4	3	4	4
9	5	5	5	3	3	4	5	4	4	4
10	4	4	3	4	4	5	5	3	4	3
11	4	4	4	4	4	4	3	5	4	3
12	3	3	4	3	4	4	5	4	4	3
13	4	5	5	4	5	4	4	4	4	4
14	4	5	5	5	3	4	4	4	1	4
15	4	5	3	2	4	5	4	4	2	5
16	5	4	2	3	3	5	2	3	2	5
17	2	3	1	4	2	3	1	3	3	5
18	5	4	4	5	1	3	3	3	3	4
19	5	5	4	3	3	3	3	3	3	4
20	3	2	5	3	3	3	3	2	5	3
21	4	3	3	3	3	2	4	2	5	3
22	4	4	4	2	4	2	4	2	4	3
23	4	4	4	1	4	4	4	1	4	3
24	4	4	2	3	4	4	5	4	4	1
25	5	3	3	3	5	4	5	4	4	2

References

Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, N.J.: Prentice Hall Regents.

Cronbach, L.J. (1970). *Essentials of psychological testing* (3rd Edition). New York: Harper & Row.

Wallen, N.E., & Fraenkel, J.R. (2011). *Educational research: A guide to the process* (2nd Edition). Mahwah, N.J.: Lawrence Erlbaum.

www.spsstests.com

TESTING OF ASSUMPTIONS

In statistical analysis, all parametric tests assume some certain characteristics about the data, also known as assumptions. Violation of these assumptions changes the conclusions of the research and interpretation of the results. Therefore, all research must follow these assumptions for accurate interpretation Depending on the parametric analysis, the assumptions vary.

The following are the data assumptions commonly found in statistical research:

- **Assumption of normality:** Normality means that the distribution of the data is normally distributed with a symmetric bell shaped curve.
- **Assumption of homogeneity of variance:** Homogeneity of variance means that two datasets or more have equal variances or the same distribution.
- **Assumption of linearity**: Linearity means that the relationship between/among research variables is linear.

NORMALITY TEST

Many of the statistical procedures including correlation, regression, t-tests, and analysis of variance, namely parametric tests, are based on the assumption that the data set follows a normal distribution or a Gaussian distribution; that is, it is assumed that the populations from which the samples are taken are normally distributed. The assumption of normality is especially critical when constructing reference intervals for variables.

With large enough sample sizes (> 30 or 40), the violation of the normality assumption should not cause major problems; this implies that we can use parametric procedures even when the data are not normally distributed. If we have samples consisting of hundreds of observations, we can ignore the distribution of the data. According to the central limit theorem, (a) if the sample data are approximately normal then the sampling distribution too will be normal; (b) in large samples (> 30 or 40), the sampling distribution tends to be normal, regardless of the shape of the data; and (c) means of random samples from any distribution will themselves have normal distribution

We can look for normality visually by using normal plots or by significance tests; that is, comparing the sample distribution to a normal one. It is important to ascertain whether data show a serious deviation from normality.

Visual inspection of the distribution may be used for assessing normality, although this approach is usually unreliable and does not guarantee that the distribution is normal. However, when data are presented visually, readers of an article can judge the distribution assumption by themselves. The frequency distribution (histogram), stem-and-leaf plot, boxplot, P-P plot (probability-probability plot), and Q-Q plot (quantile-quantile plot) are used for checking normality visually. To test the assumption of normal distribution, skewness should be within the range ±2; kurtosis values should be within range of ±7

The normality tests are supplementary to the graphical assessment of normality. The main tests for the assessment of normality are Kolmogorov-Smirnov (K-S) test, Lilliefors corrected K-S test, Shapiro-Wilk test, Anderson-Darling test, Cramer-von Mises test, D'Agostino skewness test, Anscombe-Glynn kurtosis test, D'Agostino-Pearson omnibus test, and the Jarque-Bera test.

The K-S test is an empirical distribution function (EDF) in which the theoretical cumulative distribution function of the test distribution is contrasted with the EDF of the data. A limitation of the K-S test is its high sensitivity to extreme values; the Lilliefors correction renders this test less conservative. The Shapiro-Wilk test is based on the correlation between the data and the corresponding normal scores and provides better power than the K-S test even after the Lilliefors correction. Power is the most frequent measure of the value of a test for normality—the ability to detect whether a sample comes from a non-normal distribution. Some researchers recommend the Shapiro-Wilk test as the best choice for testing the normality of data.

Normality tests are associated to the null hypothesis that the population from which a sample is extracted follows a normal distribution. So when the **p-value** linked to a normality test is lower than the risk alpha, the corresponding distribution is significantly not-normal.

If the Asymp.Sig. value > 0.05, then the dataset is considered normally distributed.
If the Asymp.Sig. value < 0.05, then the dataset is considered not normally distributed.

For example

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Data1	.168	10	.200*	.942	10	.573

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

H_0 : The dataset has the normal distribution.
H_1 : The dataset has the normal distribution.
Based on Kolmogorov-Smirnov test, the sig. value is 0.200. Because it is > 0.05, H_0 is retained; therefore, the dataset is considered normally distributed. Likewise, based on Shapiro-Wilk test, the sig. value is 0.573. Because it is > 0.05, H_0 is retained; therefore, the dataset is considered normally distributed.

References

Ghasemi, A., & Zahediasl, S. (2012 Spring). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *Int J Endocrinol Metab*, *10*(2): 486–489. Doi: 10.5812/ijem.3505.
http://www.statisticssolutions.com/testing-of-assumptions/#

HOMOGENEITY TEST

In statistics, **homogeneity** and its opposite, **heterogeneity**, arise in describing the properties of a dataset, or several datasets. They relate to the validity of the often convenient assumption that the statistical properties of any one part of an overall dataset are the same as any other part.

Homogeneity can be studied to several degrees of complexity. For example, considerations of **homoscedasticity** examine how much the variability of data-values changes throughout a dataset. However, questions of homogeneity apply to all aspects of the statistical distributions, including the location parameter. Thus, a more detailed study would examine changes to the whole of the marginal distribution. An intermediate-level study might move from looking at the variability to studying changes in the skewness. In addition to these, questions of homogeneity apply also to the joint distributions.

The assumption of **homogeneity of variance** is an assumption of the independent samples t-test and ANOVA stating that all comparison groups have the same variance. The independent samples t-test and ANOVA utilize the *t* and *F* statistics respectively, which are generally robust to violations of the assumption as long as group sizes are equal. Equal group sizes may be defined by the ratio of the largest to smallest group being less than 1.5. If group sizes are vastly unequal and homogeneity of variance is violated, then the *F* statistic will be biased when large sample variances are associated with small group sizes. When this occurs, the significance level will be underestimated, which can cause the null hypothesis to be falsely rejected. On the other hand, the *F* statistic will be biased in the opposite direction if large variances are associated with large group sizes. This would mean that the significance level will be overestimated. This does not cause the same problems as falsely rejecting the null hypothesis, however, it can cause a decrease in the power of the test.

To test for homogeneity of variance, there are several statistical tests that can be used. These tests include Hartley's F_{max},, Cochran's test, Levene's test and Barlett's test. Several of these assessments have been found to be too sensitive to non-normality and are not frequently used. Of these tests, the most common assessment for homogeneity of variance is Levene's test. Levene's test uses an *F*-test to test the null hypothesis that the variance is equal across groups. A *p*-value less than .05 indicates a violation of the assumption. If a violation occurs, it is likely that conducting the non-parametric equivalent of the analysis is more appropriate.

If the Asymp.Sig. value > 0.05, then the datasets have the same distribution.

If the Asymp.Sig. value < 0.05, then the datasets do not have the same distribution.

For example

Test of Homogeneity of Variances

Data1

Levene Statistic	df1	df2	Sig.
.022	1	38	.882

H_0 : The two datasets have the same distribution.

H_1 : The two datasets do not have the same distribution.

Based on the output of Levene's Test, the sig. value is 0.882. Because it is > 0.05, H_0 is retained; therefore, the two datasets have the same variance or the same distribution.

http://www.statisticssolutions.com/the-assumption-of-homogeneity-of-variance/#

LINEARITY TEST

Linearity means that two variables, "x" and "y," are related by a mathematical equation "y = cx," where "c" is any constant number. The importance of testing for linearity lies in the fact that many statistical methods require an assumption of linearity of data (i.e. the data was sampled from a population that relates the variables of interest in a linear fashion).

Linearity test is intended to determine whether the relationship between/among research variables is linear or not. Linearity test is a requirement in the correlation and linear regression analysis.

If the sig.value of Deviation from Linearity > 0.05, then the relationship is considered linear.
If the sig.value of Deviation from Linearity < 0.05, then the relationship is not linear.

For example

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Data1 * Data4	Between Groups	(Combined)	4242.950	16	265.184	.533	.826
		Linearity	3136.455	1	3136.455	6.307	.087
		Deviation from Linearity	1106.495	15	73.766	.148	.996
	Within Groups		1492.000	3	497.333		
	Total		5734.950	19			

H_0 : The two datasets have the linear relationship.
H_1 : The two datasets do not have the linear relationship.
Based on the ANOVA Output Table, the significance value of Deviation from Linearity is 0.996. Because it is > 0.05, H_0 is retained; therefore, it can be concluded that the relationship between the variables is linear.

https://sciencing.com/test-linearity-spss-8600862.html
www.spsstests.com

Using SPSS to Test Normality, Homogeneity, and Linearity

Normality test :

a. Analyze → Descriptive Statistics → Explore → Plots → check Normality plots with tests
b. Analyze → Nonparametric Tests → Legacy Dialogs → 1-Sample K-S→ check Normal in Test Distribution

Homogeneity test: (Input the data) Variable View: 1. Data (Scale), 2. Value (Nominal): set the values

a. Analyze→ Descriptive Statistics→Explore→Dependent List (enter the data) and Factor List (enter the values)→Plots→check Power Estimation
b. Analyze→Compare Means→One way ANOVA→ Dependent List (enter the data) and Factor (enter the values)→Options→check Homogeneity of variance test

Linearity test : Analyze → Compare Means → Means → Options → check Test for linearity in Statistics for First Layer

Exercises

Data : 85, 78, 23, 33, 35, 75, 65, 68, 47, 41, 64, 48, 54, 53, 55, 58, 76, 25, 60, 76

Data2 : 72, 48, 47, 35, 33, 36, 38, 34, 46, 43, 55, 59, 58, 66, 64, 11, 15, 23, 25, 25, 18, 20, 28, 26

Data3 : 62, 72, 53, 44, 34, 22, 25, 35, 36, 44, 45, 55, 54, 55, 47, 57, 65, 75, 65, 68, 77, 78, 66, 16

Data4 : 50, 65, 70, 72, 86, 88, 88, 90, 94, 96, 98, 98, 99, 72, 78, 80, 82, 84, 84, 85

a. Does each dataset have the normal distribution?
b. Do Data1 and Data4 have the same distribution?
c. Do Data2 and Data3 have the same distribution?
d. Do Data2, Data3, and Data4 have the same distribution?
e. Is the relationship between Data1 (performance) and Data4 (competence) linear?
f. Is the relationship between Data2 (grammar) and Data3 (writing) linear?

CORRELATION

Correlation is the **go-togetherness** of two sets of scores. Correlation is a statistical technique that is used to measure and describe the relationship between two variables. Usually the two variables are simply observed as they exist naturally in the environment—there is no attempt to control or manipulate the variables.

A correlational study is trying to see:
(1) whether there is a relationship between the two variables,
(2) the **form** of the relationship (linear or non-linear/curvilinear),
(3) the **direction** of the relationship (positive or negative), and
(4) the **strength**/consistency of the relationship ($-1 \leq r \leq +1$).

The basic question being dealt with by **correlation** can be answered in one of three possible ways:

(1) In a **positive correlation,** the two variables tend to change in the same direction: As the value of the *X* variable increases from one individual to another, the *Y* variable also tends to increase; when the *X* variable decreases, the *Y* variable also decreases.
(2) In a **negative correlation,** the two variables tend to go in opposite directions: As the *X* variable increases, the *Y* variable decreases. That is, it is an inverse relationship.
(3) In a **little systematic tendency**, some of the high and low scores on the *X* variable are paired with high scores on the *Y* variable, whereas other high and low scores on the *X* variable are paired with low scores on the *Y* variable.

Scatter Plot (Scattergram)

A **scatter plot** (*scatter diagram* = *scattergram*) is a form of visual representation of a correlation, which graphs each ordered pair of (x,y). A scatter plot has a horizontal axis and a vertical axis. These axes are labeled to correspond to the two variables involved in the correlational analysis. The **abscissa** is marked off numerically so as to accommodate the obtained scores collected by the researcher on the variable represented by the horizontal axis; in a similar fashion, the **ordinate** is labeled so as to accommodate the obtained scores on the other variable.

A scatter plot reveals the relationship between two variables through the pattern that is formed by the full set of dots. Researchers often insert a straight line through the data points to help us see the nature of the relationship. If the line extends from lower-left to upper-right, we have a positive relationship. However, if the line extends from upper-left to lower-right, we have a negative relationships. If the line passing through the dots has a near zero tilt, this indicates that the two variables are not related very much at all. The strength of the relationship can be determined by asking how close the dots seem to lie from the line. If the dots are close to the line, the relationship is strong; if the dots are scattered both close to and far from the line, the relationship is best thought of as moderate, or even weak.

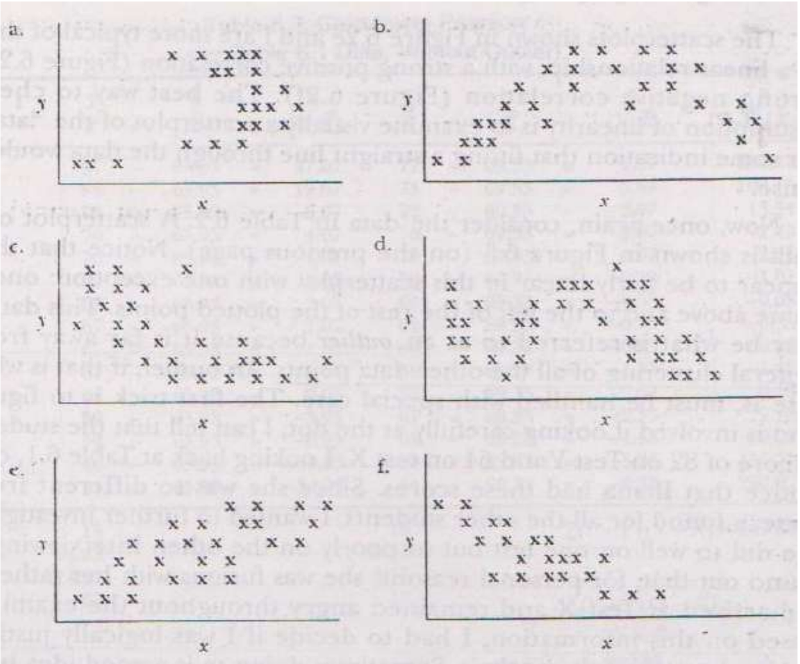


Figure: Curvilinear (a-d) and Linear (e-f) Scatterplots

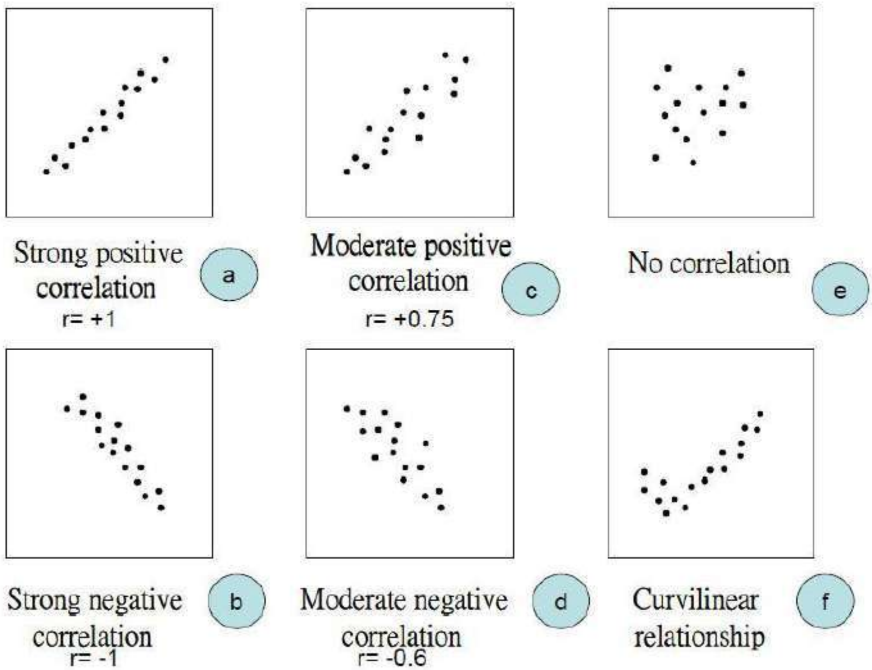


Figure: Positive and Negative Correlation

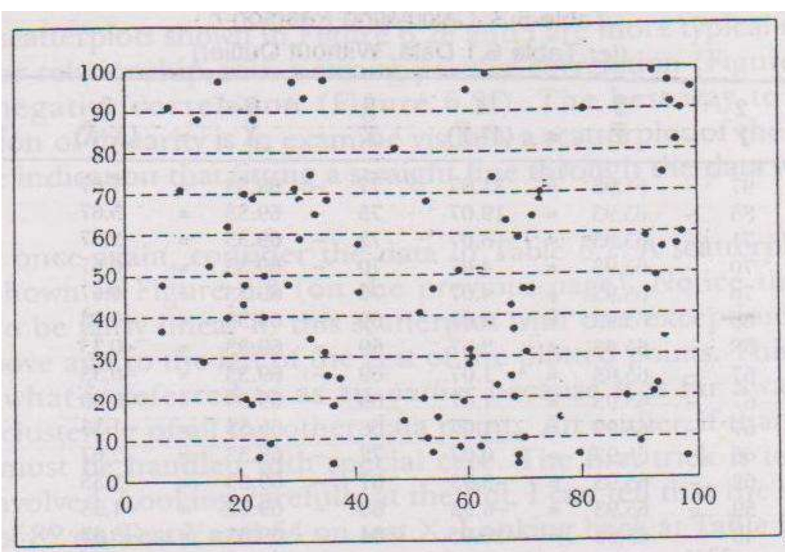
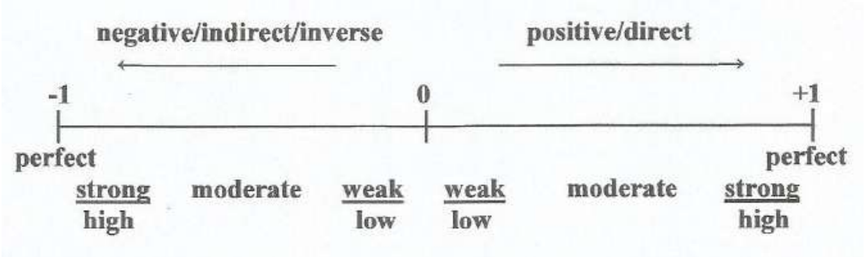


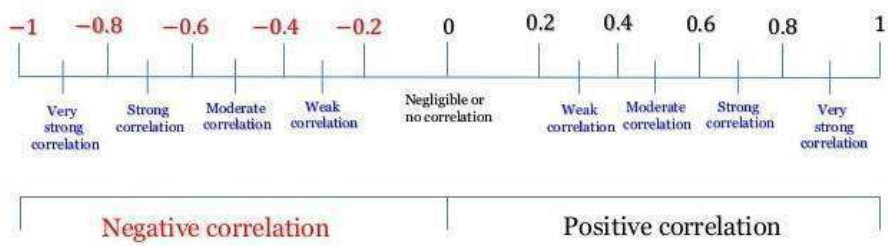
Figure: Scatterplot for Two Sets of Random Numbers

Correlation Coefficient

The degree to which two sets of scores co-vary or vary together is estimated statistically by calculating a **correlation coefficient**. Correlation coefficient**,** symbolized as *r*, is normally reported as a decimal number somewhere between -1.00 and +1.00. This correlational continuum helps you pin down the meaning of several adjectives that researchers use when talking about correlation coefficients or relationships: *direct, high, indirect, inverse, low, moderate, negative, perfect, positive, strong,* and *weak*. First, consider the two halves of the correlational continuum. Any ***r*** that falls on the right side represents a **positive correlation;** this indicates a **direct relationship** between the two measured variables. However, any result that ends up on the left side is a **negative correlation,** and this indicates an **indirect,** or **inverse, relationship**. If *r* lands on either end of our correlation continuum, the term **perfect** may be used to describe the obtained correlation. The term **high** comes into play when *r* assumes a value close to either end (thus implying a **strong** relationship); conversely, the term **low** is used when *r* lands close to the middle of the continuum (thus implying a **weak** relationship). Not surprisingly, any *r* that ends up in the middle area of the left or right sides of our continuum is called **moderate.**



Correlation Coefficient Interpretation Guideline:



For *interval* or *ratio* variables, the appropriate measure is **the Pearson product-moment correlation coefficient** (r). A measure of the correlation between two *ordinal* variables can be obtained by calculating **the Spearman rank-order correlation coefficient** (Spearman's rho or r_s or ρ), and use **Kendall's tau** (**tau-b)** for the data in the form of *ranks* with the issue of *ties*. For *nominal* variables, **the phi coefficient** (ϕ) or **Cramer's *V*** may be used. (see Butler, 1985, pp. 143-149). The appropriate statistic to apply when examining the relationship between a *nominal and* an *interval* scale is **the point-biserial correlation coefficient** (r_{pbi}).

Variables	Measures
Interval or Ratio variables	the Pearson product-moment correlation coefficient (r)
Ordinal variables	the Spearman rank-order correlation coefficient (r_s or ρ)
Ordinal variables with ties	Kendall's tau (tau-b)
Nominal variables	the phi coefficient (ϕ) or Cramer's V
Nominal and Interval variables	the point-biserial correlation coefficient (r_{pbi}).

Coefficient of Determination

To get a better feel for the strength of the relationship between two variables, many researchers square the value of the correlation coefficient. For example, if r turns out equal to .80, the researcher squares .80 and obtains .64. When r is squared like this, the resulting value is called the **coefficient of determination.** The coefficient of determination indicates the proportion of variability in one variable that is associated with (or explained by) variability in the other variable. The value of lies somewhere between 0 and 1.00, and researchers usually multiply by 100 so they can talk about the *percentage* of explained variability. Researchers sometimes refer to this percentage as the amount of variance in one variable that is accounted for by the other variable, or they sometimes say that this percentage indicates the amount of *shared variance.* The value indicates how much (proportionately speaking) variability in either variable is explained by the other variable. The implication of this is that the raw correlation coefficient (i.e., the value of r when not squared) exaggerates how strong the relationship really is between two variables. Note that *r* must be stronger than .70 for there to be at least 50 percent explained variability.

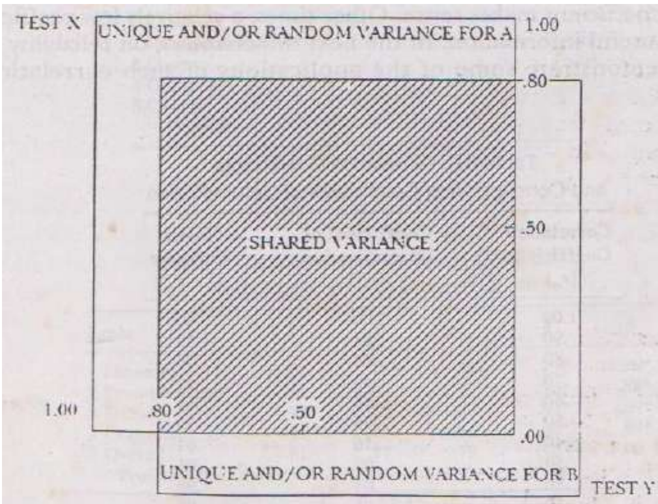


Figure: Overlapping Variance

The formula of the Pearson product-moment correlation coefficient is as follows.

$$r_{xy} = \frac{\sum(Y - \overline{Y})(X - \overline{X})}{N S_y S_x}$$

where r_{xy} = Pearson product-moment correlation coefficient
Y = each student's score on Test Y
$\overline{Y}$ = mean on Test Y
S_y = standard deviation on Test Y
X = each student's score on Test X
$\overline{X}$ = mean on Test X
S_x = standard deviation on Test X
N = the number of students who took the two tests

Table: Calculating a correlation coefficient:

Column 1 Students	2 Y	−	3 $\bar{Y}$	=	4 $(Y-\bar{Y})$	5 X	−	6 $\bar{X}$	=	7 $(X-\bar{X})$	8 $(Y-\bar{Y})(X-\bar{X})$
Robert	97	−	66.94	=	30.06	77	−	69.00	=	8.00	240.48
Millie	85	−	66.94	=	18.06	75	−	69.00	=	6.00	108.36
Iliana	82	−	66.94	=	15.06	64	−	69.00	=	−5.00	−75.30
Dean	71	−	66.94	=	4.06	72	−	69.00	=	3.00	12.18
Cuny	70	−	66.94	=	3.06	70	−	69.00	=	1.00	3.06
Bill	70	−	66.94	=	3.06	70	−	69.00	=	1.00	3.06
Corky	69	−	66.94	=	2.06	69	−	69.00	=	0.00	0.00
Randy	68	−	66.94	=	1.06	69	−	69.00	=	0.00	0.00
Monique	67	−	66.94	=	0.06	69	−	69.00	=	0.00	0.00
Wendy	67	−	66.94	=	0.06	69	−	69.00	=	0.00	0.00
Henk	67	−	66.94	=	0.06	68	−	69.00	=	−1.00	−0.06
Shenan	66	−	66.94	=	−0.94	72	−	69.00	=	3.00	−2.82
Jeanne	62	−	66.94	=	−4.94	67	−	69.00	=	−2.00	9.88
Elisabeth	59	−	66.94	=	−7.94	68	−	69.00	=	−1.00	7.94
Archie	40	−	66.94	=	−26.94	64	−	69.00	=	−5.00	134.70
Lindsey	31	−	66.94	=	−35.94	61	−	69.00	=	−8.00	287.52
N	16					16					$\Sigma(Y-\bar{Y})(X-\bar{X}) = 729.00$
Mean	66.94					69.00					
S	15.01					3.87					
Range	67					17					

$$r_{xy} = \frac{\Sigma(Y-\bar{Y})(X-\bar{X})}{N\,S_x\,S_y}$$

$$= \frac{729.00}{16(15.01)(3.87)}$$

$$= \frac{729.00}{929.42} = .7843601$$

$$= .78$$

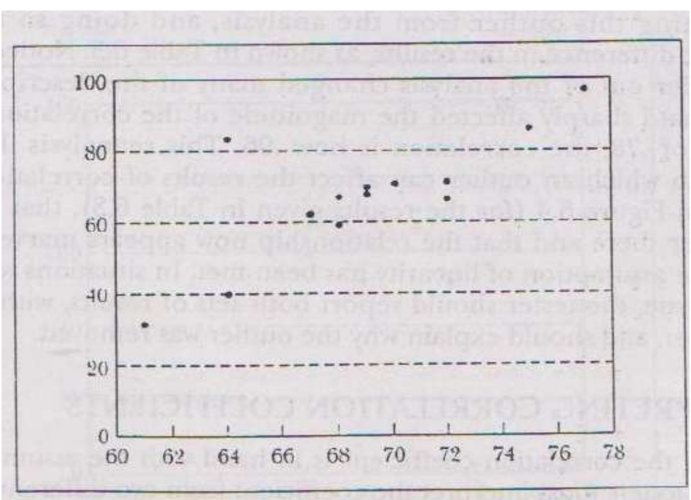


Figure: Scatterplot for the data above

The formula of the Spearman rank-order correlation coefficient is as follows.

$$\rho = 1 - \frac{6 \times \Sigma D^2}{N(N^2 - 1)}$$

where ρ = Spearman rank-order correlation coefficient

D = difference between ranks in each pair

N = number of students for whom you have pairs of ranks

Σ = sum

6 = a constant

Table: Calculating Spearman *rho*

Column 1	2		3		4	5	6
Student	Ranks on Test A		Ranks on Test B		D	D^2	Calculations
Robert	1	−	4	=	−3	9	
Millie	2	−	3	=	−1	1	
Iliana	3	−	2	=	1	1	$\rho = 1 - \frac{6 \times \Sigma D^2}{N(N^2 - 1)}$
Dean	4	−	1	=	3	9	
Cuny	5	−	5	=	0	0	$= 1 - \frac{6 \times 40}{9(81 - 1)}$
Bill	6	−	9	=	−3	9	
Corky	7	−	8	=	−1	1	$= 1 - \frac{240}{720}$
Randy	8	−	7	=	1	1	$= 1 - .33$
Monique	9	−	6	=	3	9	$= .67$*
*$p < .05$ (with $N = 9$)					$\Sigma D^2 = 40$		

The *phi* Coefficient

When the two variables are nominal, the question of whether they are correlated or not resolves itself into the question, 'If a particular item has property A, is there a high probability that it will also have property B?' A common application is to the statistics of testing in language teaching and other fields. Here the question at issue is, 'If a learner passes on test item 1, is he also likely to pass on item 2?' The general situation involved in such studies can be represented as a 2 x 2 table.

Variable 2	Variable 1 –	Variable 1 +
+	*A*	*B*
–	*C*	*D*

Here A, B, C and D are the frequencies of observations in the four cells of the table; for instance, A might be the number of learners who fail test item 1 but pass test item 2, B the number who pass both test items, and so on. The correlation between the two variables can be measured by the phi (ϕ) coefficient, defined as

$$\phi = \frac{BC - AD}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

It should be clear that this formula (the phi coefficient) is very closely related to the chi-square statistic. The two statistics are related as follows:

$$\phi = \sqrt{\chi^2/N} \quad \text{or} \quad \chi^2 = N\phi^2.$$

The significance of a phi coefficient can therefore be assessed by calculating $N\phi^2$ and then looking up critical values in the chi-square table with one degree of freedom (since there is a 2 x 2 table)

Table: Success and failure on two language test items

Test item 2	*Test item 1* Failure	*Test item 1* Success	*Total*
Success	49	21	70
Failure	14	16	30
Total	63	37	100

Suppose that we wish to know whether success on one item in a language test is significantly correlated with success on a second item. Imagine that we have tested 100 learners, with the results as shown in the table above. We calculate ϕ as

$$\phi = \frac{(21 \times 14) - (49 \times 16)}{\sqrt{70 \times 30 \times 63 \times 37}} = -0.22.$$

There is thus some degree of negative correlation between the two variables. To test its significance at the 5 per cent level, we calculate

$$\chi^2 = N\phi^2 = 100 \times (-0.22)^2 = 4.84.$$

The critical value of X^2 for one degree of freedom and the 5 per cent level is 3.84, so that the calculated value of ϕ is in fact significant at this level.

To calculate the r_{pbi} for each item, use the following formula:

$$r_{pbi} = \frac{\overline{X}_p - \overline{X}_q}{S_t}\sqrt{pq}$$

where: r_{pbi} = point-biserial correlation coefficient

$\overline{X}_p$ = mean on the whole test for those students who answered correctly (i.e., are coded as 1s)

$\overline{X}_q$ = mean on the whole test for those students who answered incorrectly (i.e., are coded as 0s)

S_t = standard deviation for whole test

Hypotheses

H_0: There is no significant correlation between *X* and *Y*.
H_a: There is a significant correlation between *X* and *Y*.

H_0: There is no significant positive correlation between *X* and *Y*.
H_a: There is a significant positive correlation between *X* and *Y*.

H_0: There is no significant negative correlation between *X* and *Y*.
H_a: There is a significant negative correlation between *X* and *Y*.

Using Excel to calculate correlation coefficients

Excel will do the work for you with the CORREL function that has the following characteristics:

CORREL (array1, array2)

where: array1 = the range of data for the first variable

array2 = the range of data for the second variable

Student	Grammar	Reading
1	65	70
2	76	66
3	56	45
4	80	79
5	35	56
6	70	80
7	90	87
8	54	65
9	87	78
10	45	55
	Correl	0.809157

Data analysis → Correlation

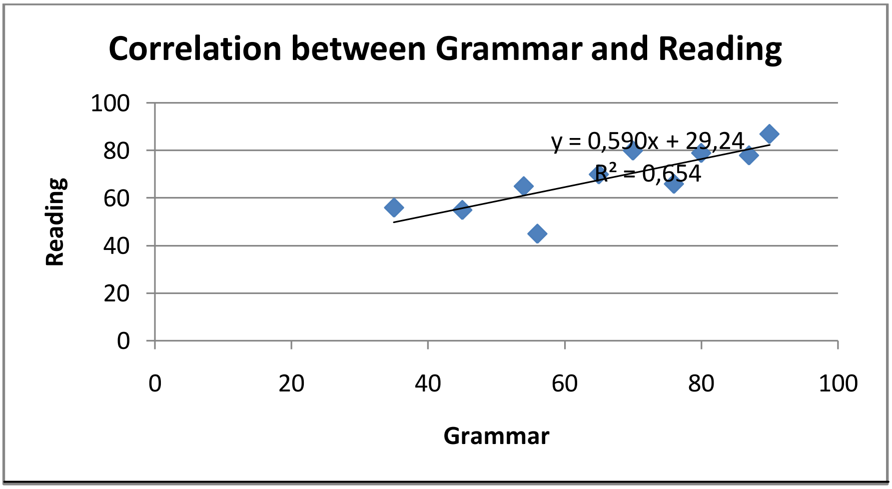
	Grammar	Reading
Grammar	1	
Reading	0.809157356	1

H_0: There is no significant correlation between grammar and reading.
H_a: There is a significant correlation between grammar and reading.

At the significance level 0.05 in two-tailed testing with N=10, the critical value r = 0.632 (see the table of Pearson Product-Moment Correlation Coefficient). The result of the Pearson Product-Moment Correlation shows that the r-obtained = 0.809. Since the obtained value (0.809) is higher than the critical value (0.632), H_0 is rejected and consequently H_a is accepted. It means that there is a significant correlation between grammar and reading.

Using Excel to display the scattergram

Insert → Scatter → Choose the simple one
To show the equation: Click the chart→Design→Chart Layouts → Choose fx chart



Using SPSS to calculate correlation coefficients

Analyze →Correlate →Bivariate

Correlations

		Grammar	Reading
Grammar	Pearson Correlation	1	.809**
	Sig. (2-tailed)		.005
	N	10	10
Reading	Pearson Correlation	.809**	1
	Sig. (2-tailed)	.005	
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

H_0: There is no significant correlation between grammar and reading.
H_a: There is a significant correlation between grammar and reading.

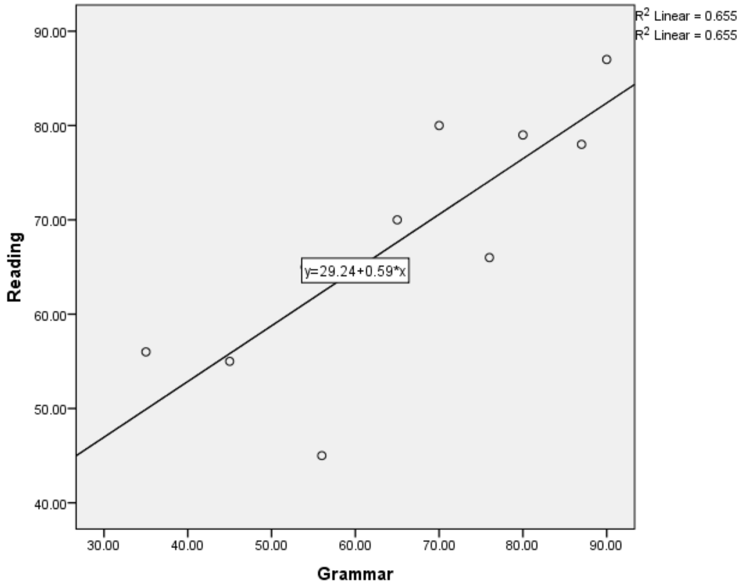
The results of the Pearson Product-Moment Correlation show that the r-obtained = 0.809 and the significance value (p-value) = 0.005. Since the p-value (0.005) is lower than 0.05, H_0 is rejected and consequently H_a is accepted. It means that there is a significant correlation between grammar and reading.

Using SPSS to display the scattergram

Graphs → Legacy Dialogs → Scatter/Dot → Click Simple Scatter → Define → Input the data to Y Axis and X Axis → OK

To show the equation: Double-click the chart → Click Add Fit Line at Total → Check Linear, check Attach label to line →Apply

Graph



Student	Grammar	Reading	Writing	Speaking
1	65	70	66	65
2	76	66	70	67
3	56	45	60	55
4	80	79	78	70
5	35	56	40	35
6	70	80	75	70
7	90	87	80	75
8	54	65	60	60
9	87	78	75	70
10	45	55	50	50

Excel

A Correlation Matrix

	Grammar	Reading	Writing	Speaking
Grammar	1			
Reading	0.809157356	1		
Writing	0.960578423	0.81519057	1	
Speaking	0.935143949	0.79802105	0.980517901	1

SPSS

A Correlation Matrix

Correlations

		Grammar	Reading	Writing	Speaking
Grammar	Pearson Correlation	1	.809**	.961**	.935**
	Sig. (2-tailed)		.005	.000	.000
	N	10	10	10	10
Reading	Pearson Correlation	.809**	1	.815**	.798**
	Sig. (2-tailed)	.005		.004	.006
	N	10	10	10	10
Writing	Pearson Correlation	.961**	.815**	1	.981**
	Sig. (2-tailed)	.000	.004		.000
	N	10	10	10	10
Speaking	Pearson Correlation	.935**	.798**	.981**	1
	Sig. (2-tailed)	.000	.006	.000	
	N	10	10	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

Exercises

Data1 : 85, 78, 23, 33, 35, 75, 65, 68, 47, 41, 64, 48, 54, 53, 55, 58, 76, 25, 60, 76

Data2 : 50, 65, 70, 72, 86, 88, 88, 90, 94, 96, 98, 98, 99, 72, 78, 80, 82, 84, 84, 85

Data3 : 72, 48, 47, 35, 33, 36, 38, 34, 46, 43, 55, 59, 58, 66, 64, 11, 15, 23, 25, 25, 18, 20, 28, 26

Data4 : 62, 72, 53, 44, 34, 22, 25, 35, 36, 44, 45, 55, 54, 55, 47, 57, 65, 75, 65, 68, 77, 78, 66, 16

a. Is there any significant correlation between Data1 and Data2?

b. Is there any significant correlation between Data3 and Data4?

REGRESSION

Regression analysis is used to predict and/or explain a variable with the use of one or more variables. It is assumed that a linear relationship exists between the dependent (to be predicted) variable (*y*) and one (or more) independent variable(s) (*x*). This means that the relationship can be visualized as a more or less straight line with a gradient that is either positive or negative. In a *simple* regression analysis, there is one independent and one dependent variable. The analysis becomes *multiple* when other *x* variables are added to measure their relative impact.

There are three different kinds of regression: **bivariate regression, multiple regression,** and **logistic regression.** Bivariate regression is similar to bivariate correlation because both are designed for situations in which there are just two variables. Multiple and logistic regression, however, are created for cases in which there are three or more variables.

Regression and correlation differ in three important respects: their purpose, the way variables are labeled, and the kinds of inferential tests applied to the data. Although regression and correlation are not the same, correlational concepts serve as some (but not all) of regression's building blocks. Correlation concentrates on the relationship, or link, that exists *between* variables; however, regression focuses on the variable(s) that exist on one or the other *ends* of the link. Depending on which end is focused on, regression tries to accomplish one or the other of two goals: **prediction** or **explanation.**

In some studies, regression is utilized to predict scores on one variable based on information regarding the other variable(s). For example, a college might use regression in an effort to predict how well applicants will handle its academic curriculum. An applicant's college grade point average (GPA) might be the main focus of the regression, with predictions made on the basis of available data on other variables (e.g., an entrance exam, the applicant's essay, and the recommendations written by high school teachers).

The second difference between regression and correlation concerns the labels attached to the variables. In a correlation analysis, variables A and B have no special names; they are simply the study's two variables. In a regression analysis, the two variables need to be identified as a **dependent variable** and an **independent variable.** This distinction is important because (1) the scatter diagram in bivariate regression is set up such that the vertical axis corresponds with the dependent variable whereas the horizontal axis represents the independent variable, and (2) the names of the two variables cannot be interchanged in verbal descriptions of the regression. For example, the regression of A on B is not the same as the regression of B on A.

The terms **criterion variable, outcome variable,** and **response variable** are synonymous with the term **dependent variable,** whereas the terms **predictor variable** or **explanatory variable** mean the same thing as **independent variable.** When the phrase "regression of __ on __" is used, the variable appearing in the first blank is the dependent variable whereas the variable appearing in the second blank is the independent variable.

The third difference between regression and correlation concerns the focus of inferential tests and confidence intervals. With correlation, there is just one thing that can be focused on: the sample correlation coefficient. With regression, however, inferences focus on the correlation coefficient, the regression coefficient(s), and the intercept.

BIVARIATE REGRESSION

The simplest kind of regression analysis is called **bivariate regression.** First, we must clarify the purpose of and the data needed for this kind of regression. Then, we consider scatter diagrams, lines of best fit, and prediction equations. Finally, we discuss inferential procedures associated with bivariate regression.

Bivariate regression involves just two variables. One of the variables serves as the dependent variable whereas the other functions as the independent variable. The purpose of this kind of regression can be either prediction or explanation; however, bivariate regression is used most frequently to see how well scores on the dependent variable can be predicted from data on the independent variable.

Scatter Diagram, Regression Line, and Regression Equation

The component parts and functioning of regression can best be understood by examining a scatter diagram. The technique of *simple regression* enables us to describe a straight line that best fits a series of ordered pairs (x,y). The equation for a straight line, known as a *linear equation,* takes the form:

$$\hat{y} = a + bx$$

where: $\hat{y}$ = the predicted value of y, given a value of x

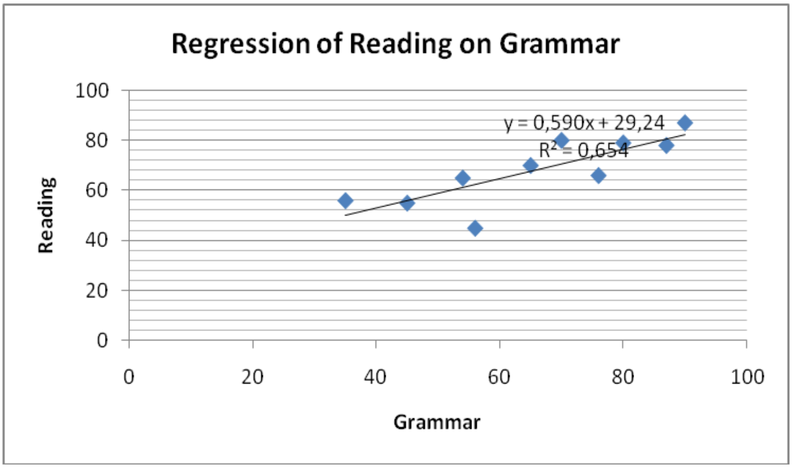
x = the independent variable

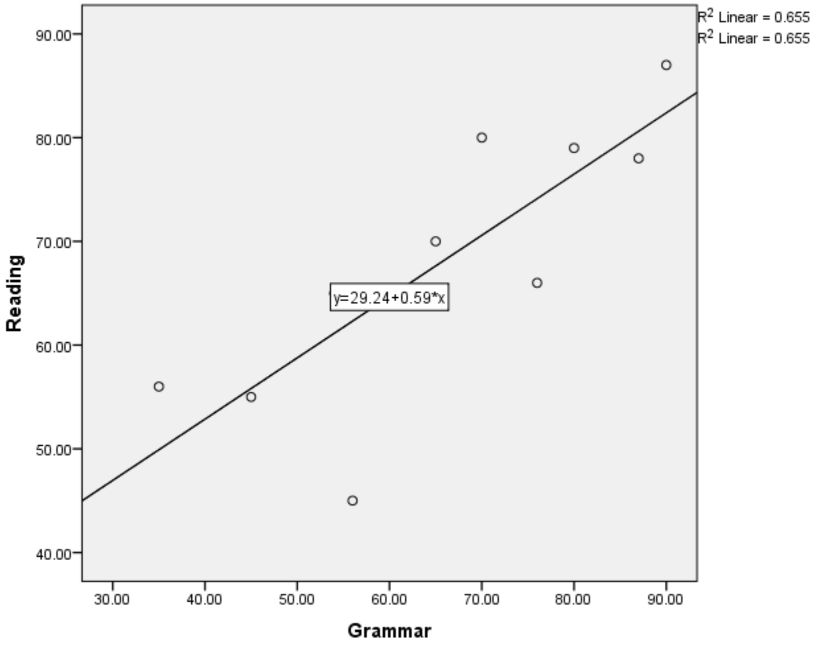
a = the y-intercept for the straight line or a *constant*

b = the slope of the straight line or the *regression coefficient*

The statistical technique for finding the best-fitting straight line for a set of data is called **simple linear regression analysis**, and the resulting slanted straight line passing through the data points of the scatter diagram is called the **regression line** or the **line of best fit,** and it functions as the tool to predict. The regression line is positioned so as to be as close as possible to all of the dots. A special formula determines the precise location of this line, and the formula is based on a statistical concept called ***least squares***.

The ***least squares principle*** simply means that when the squared vertical distances of the data points from the regression line are added together, they yield a smaller sum than would be the case for any other straight line that could be drawn through the scatter diagram's data points.





It should be noted that there are two kinds of regression equations that can be created in any bivariate regression analysis. One is called an **unstandardized regression equation**, and it has the form $\hat{y} = a + bx$. The other kind of regression equation (that can be generated using the same data) is called a **standardized regression equation** which has the form $\hat{z}_y = \beta z_x$. These two kinds of regression equations differ in three respects. First, a standardized regression equation involves *z*-scores on both the independent and the dependent variables, not raw scores. Second, the standardized regression equation does not have a *constant*. Finally, the symbol β (called a **beta weight**) is used in place of the regression coefficient, *b*. Finally, when one variable, *x*, is being used to predict a second variable, $\hat{y}$, the value of *beta* is equal to the Pearson correlation for *x* and *y*. Thus, the standardized form of the regression equation can also be written as $\hat{z}_y = r\, Z_x$.

Calculating and Interpreting a, b in Bivariate Regression

When used for predictive purposes, the regression equation has the form $\hat{y} = a + bx$. The two main ingredients in the equation are *a* and *b*.

The main component of the regression is *b*, the **regression coefficient**. When the regression line has been positioned within the data points of a scatter diagram, *b* simply indicates the **slope** of that line. *slope* means "rise over run." In other words, the value of *b* signifies how many predicted units of change (either up or down) in the dependent variable there are for any one unit increase in the independent variable. For example, a slope of 10/3 means as the *x*-value increases (moves right) by 3 units, the *y*-value moves up by 10 units on average. In a regression context, the slope is the heart and soul of the equation because it tells you how much you can expect *y* to change as *x* increases.

The formula for the *slope*, *b*, of the best-fitting line is $b = r\,(S_y/S_x)$
where: *r* is the correlation between *x* and *y*,
S_x and S_y are the standard deviations of the *x*-values and the *y*-values
(Think of S_y/S_x as the change in *y* over the change in *x* in units of *x* and *y*.)

The other component of the regression is *a*, the **constant** or the **intercept.** Simply stated, *a* indicates where the regression line in the scatter diagram would, if extended to the left, intersect the ordinate. It indicates, therefore, the value of $\hat{y}$ for the case where x = 0. For example, in the equation $\hat{y} = -6 + 2x$, the line crosses the y-axis at the point –6. The coordinates of this point are (0,–6).

The formula for the y-intercept, a, of the best-fitting line is $\boldsymbol{a} = \bar{y} - \boldsymbol{b}\,\bar{x}$
where: $\bar{x}$ and $\bar{y}$ are the means of the x-values and the y-values respectively,
b is the *slope*

Standard Error of Estimate

The **standard error of estimate (s_e)** gives a measure of the standard distance between the predicted ŷ values on the regression line and the actual *y* values in the data. The standard error of the estimate measures the amount of dispersion of the observed data around the regression line. Conceptually, the standard error of estimate is very much like a standard deviation: both provide a measure of standard distance. Also, the calculation of the standard error of estimate is very similar to the calculation of standard deviation.

To calculate the standard error of estimate, we first find the sum of squared deviations (*SS*). Each deviation measures the distance between the actual *y* value (from the data) and the predicted $\hat{y}$ value (from the regression line). This sum of squares is commonly called $SS_{residual}$ because it is based on the remaining distance between the actual *y* scores and the predicted values.

$$SS_{\text{residual}} = \Sigma\,(y - \hat{y})^2$$

The obtained *SS* value is then divided by its degrees of freedom to obtain a measure of variance. This procedure should be very familiar:

Variance = *SS*/*df*

The degrees of freedom for the standard error of estimate are ***df* = *n*-2**. The reason for having (*n-2*) degrees of freedom, rather than the customary (*n–1*), is that we now are measuring deviations from a line rather than deviations from a mean. To find the equation for the regression line, you must know the means for both the *x* and the *y* scores. Specifying these two means places two restrictions on the variability of the data, with the result that the scores have only ($n–2$) degrees of freedom.

The final step in the calculation of the standard error of estimate is to take the square root of the variance to obtain a measure of standard distance. The final equation is:

$$\textbf{standard error of estimate} = \sqrt{(SS_{residual}/df)} = \sqrt{\Sigma (y - \hat{y})^2 / (n\text{-}2)}$$

With a large correlation (near -1.00 or +1.00), the data points are close to the regression line, and the standard error of estimate is small. As a correlation gets smaller (near zero), the data points move away from the regression line, and the standard error of estimate gets larger. Because it is possible to have the same regression equation for several different sets of data, it is also important to consider r^2 and the standard error of estimate. The regression equation simply describes the best-fitting line and is used for making predictions. However, r^2 and the standard error of estimate indicate how accurate these predictions are.

Squaring the correlation provides a measure of the accuracy of prediction. The squared correlation, r^2, is called the coefficient of determination because it determines what proportion of the variability in y is predicted by the relationship with x. The coefficient of determination indicates the proportion of variability in the dependent variable that is explained by the independent variable. Because r^2 measures the predicted portion of the variability in the y scores, we can use the expression ($1\text{-}r^2$) to measure the unpredicted portion. Thus,

$$\textbf{predicted variability} = SS_{regression} = r^2 SS_y$$
$$\textbf{unpredicted variability} = SS_{residual} = (1 - r^2) SS_y$$

Using the equation to compute $SS_{residual}$, the standard error of estimate can be computed as

$$\textbf{standard error of estimate} = \sqrt{(SS_{residual}/df)} = \sqrt{(1 - r^2) SS_y / (n\text{-}2)}$$

Using Excel for Simple Regression

Data Analysis → Regression (*Input Y Range, Input X Range, Labels, Output Range*)
(*Input Ranges include Labels, grammar/reading*)

Student	Grammar	Reading
1	65	70
2	76	66
3	56	45
4	80	79
5	35	56
6	70	80
7	90	87
8	54	65
9	87	78
10	45	55

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.809157356
R Square	0.654735628
Adjusted R Square	0.611577581
Standard Error	8.270501073
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1037.690496	1037.690496	15.1706502	0.004577129
Residual	8	547.209504	68.401188		
Total	9	1584.9			

	Coefficients	*Standard Error*	*t Stat*	*p-value*	*Lower 95%*	*Upper 95%*
Intercept	29.24270735	10.31344324	2.835397128	0.021966988	5.459864605	53.0255501
Grammar	0.590536362	0.15161583	3.894951886	0.004577129	0.240909632	0.940163093

Adjusted r^2 = **0.612**. It means that grammar explained 61.2% of the variance in reading.

Because the value of sig. F (0.004577) is less than $\alpha = 0.05$, we can reject the null hypothesis and conclude that a significant relationship between the variables does exist.

The regression equation: $\hat{y} = a + bx$

$\hat{y}$ **= 29.2427** + **0.5905** x

Because the ρ-value for the independent variable Grammar is shown as **0.004577**, which is less than α = 0.05, we can reject the null hypothesis (that the independent variable provides no assistance in predicting scores on the dependent variable). [H_0: $\beta = 0$]

Using SPSS for Simple Regression

Analyze → Regression → Linear

Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.809[a]	.655	.612	8.271

a. Predictors: (Constant), Grammar

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1037.690	1	1037.690	15.171	.005[b]
	Residual	547.210	8	68.401		
	Total	1584.900	9			

a. Dependent Variable: Reading
b. Predictors: (Constant), Grammar

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	29.243	10.313		2.835	.022
	Grammar	.591	.152	.809	3.895	.005

a. Dependent Variable: Reading

Adjusted r^2 = **0.612**. It means that grammar explained 61.2% of the variance in reading.

Because the value of sig. F (0.005) is less than α = 0.05, we can reject the null hypothesis and conclude that a significant relationship between the variables does exist.

The unstandardized regression equation: $\hat{y} = a + bx$

$\hat{y}$ **= 29.243** + **0.591** x

The standardized regression equation : $\hat{Z}_y = \beta\ Z_x$

$\hat{Z}_y = 0.809\ Z_x$

Because the ρ-value for the independent variable Grammar is shown as **0.005**, which is less than α = 0.05, we can reject the null hypothesis (that the independent variable provides no assistance in predicting scores on the dependent variable). [H_0: $\beta = 0$]

Stepwise Linear Regression

Stepwise linear regression is a method of regressing multiple variables while simultaneously removing those that aren't important. Stepwise regression essentially does multiple regression a number of times, each time removing the weakest correlated variable. At the end you are left with the variables that explain the distribution best. The only requirements are that the data is normally distributed and that there is no correlation between the independent variables (known as collinearity).
(http://www.geog.leeds.ac.uk/courses/other/statistics/spss/stepwise/)

Stepwise regression is an automated tool used in the exploratory stages of model building to identify a useful subset of predictors. The process systematically adds the most significant variable or removes the least significant variable during each step.

Pitfalls of Stepwise Regression

There are some potential pitfalls to be aware of:

- If two independent variables are highly correlated, only one may end up in the model even though both may be important.
- Because the procedure fits many models, it could be selecting models that fit the data well due to chance alone.
- Stepwise regression may not always end with the model with the highest R^2 value possible for a given number of predictors.
- Automatic procedures cannot take into account special knowledge the analyst may have about the data. Therefore, the model selected may not be the most practical one.

- Graphing individual predictors against the response is often misleading because graphs do not account for other predictors in the model.

Using SPSS for Stepwise Regression

Analyze → Regression → Linear
Dependent → In
Independent → In → Next (in steps)
Method: Stepwise

No	Organization	Content	Grammar	Vocabulary	Mechanics	Writing (Total)
1	3	4	4	3	4	18
2	3	4	5	4	5	21
3	4	4	3	3	3	17
4	5	3	4	4	3	19
5	2	3	3	3	3	14
6	3	3	2	2	3	13
7	1	2	2	1	2	8
8	4	3	4	5	4	20
9	3	4	5	4	3	19
10	2	3	3	3	2	13

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.755[a]	.570	.517	2.834	.570	10.622	1	8	.012
2	.887[b]	.786	.725	2.137	.216	7.067	1	7	.033
3	.978[c]	.956	.935	1.042	.170	23.443	1	6	.003
4	.987[d]	.975	.955	.865	.019	3.702	1	5	.112
5	1.000[e]	1.000	1.000	.000	.025	.	1	4	.

a. Predictors: (Constant), Organization

b. Predictors: (Constant), Organization, Content

c. Predictors: (Constant), Organization, Content, Grammar

d. Predictors: (Constant), Organization, Content, Grammar, Vocabulary

e. Predictors: (Constant), Organization, Content, Grammar, Vocabulary, Mechanics

Exercise

The following are two sets of raw scores from a grammar test and a speaking test.

Grammar scores : 45 27 34 25 15 46 08 17 20 40
48 30 38 22 14 36 09 07 25 34
21 16 27 10 38 29 40 27 19 24

Speaking scores : 20 12 15 11 10 21 05 10 05 16
16 16 07 16 17 12 05 13 16
2 07 11 10 16 15 05 15 06 15

1. What is the regression line that best fits the data?
   (Find the regression equation for predicting *y* from *x*.)
2. Suppose a student has a grammar score of 32, what is the predicted score of speaking?
3. What percent of the variance in speaking is explained by the variable grammar?
4. Determine whether the relationship between the two variables is statistically significant or not.

References

Donnelly, R.A. (2007). *The complete idiot's guide to statistics*. (2^{nd} Edition). Indianapolis: Alpha Books, a member of Penguin Group, Inc.

Gravetter, F.J., and Wallnau, L.B. (2013). *Statistics for the behavioral sciences*. (9 Ed.). Belmont, C.A.: Wadsworth, Cengage Learning.

Grotenhuis, M., and Matthijssen, A. (2016). *Basic SPSS tutorial*. Thousand Oaks, California: SAGE Publications, Inc.

Huck, S.W. (2012). *Reading statistics and research*. (Sixth Edition). Boston: Pearson.

Pramesti, G. (2014). *Kupas tuntas data penelitian dengan SPSS 22*. Jakarta: Penerbit PT Elex Media Komputindo.

Rumsey, D. (2010). *Statistics essentials for dummies*. Hoboken, N.J.: Wiley Publishing, Inc.

Winarno, E., Zaki, A., & SmitDev Community. (2014). *Statistik dan analisisnya dengan Excel*. Jakarta: Penerbit PT Elex Media Komputindo.

T-TESTS

t-test is used to test the significance of differences with a single sample or two samples:

(1) hypothesis testing with one single sample
(2) hypothesis testing with two independent samples
(3) hypothesis testing with two dependent samples

Hypothesis testing with one single sample

In this hypothesis-testing situation, we begin with a population with an unknown mean (μ) and an unknown variance (σ), often a population that has received some treatment. The goal is to use a sample from the treated population (a treated sample) as the basis for determining whether the treatment has any effect. As always, the null hypothesis states that the treatment has no effect; specifically, H_0 states that the population mean is unchanged. Thus, the null hypothesis provides a specific value for the unknown population mean. The sample data provide a value for the sample mean. Finally, the variance and estimated standard error are computed from the sample data.

$$t = \frac{\textit{sample mean- population mean}}{\textit{(estimated) standard error}} = \frac{\bar{x} - \mu}{\sqrt{(s^2/n)}}$$

The *t* statistic forms a ratio. The numerator measures the actual difference between the sample data ($\bar{x} = M$) and the population hypothesis (μ). The estimated standard error in the denominator measures how much difference is reasonable to expect between a sample mean and the population mean. When the obtained difference between the data and the hypothesis (numerator) is much greater than expected (denominator), we obtain a large value for *t* (either large positive or large negative). In this case, we conclude that the data are not consistent with the hypothesis, and our decision is to "reject H_0." On the other hand, when the difference between the data and the hypothesis is small relative to the standard error, we obtain a *t* statistic near zero, and our decision is "fail to reject H_0."

One criticism of a hypothesis test is that it does not really evaluate the size of the treatment effect. Instead, a hypothesis test simply determines whether the treatment effect is greater than chance, where "chance" is measured by the standard error. In particular, it is possible for a very small treatment effect to be "statistically significant," especially when the sample size is very large. To correct for this problem, it is recommended that the results from a hypothesis test be accompanied by a report of effect size.

Measuring effect size is to determine how much of the variability in the scores is explained by the treatment effect. The concept behind this measure is that the treatment causes the scores to increase (or decrease), which means that the treatment is causing the scores to vary. If we can measure how much of the variability is explained by the treatment, we can obtain a measure of the size of the treatment effect. This value is called the ***percentage of variance accounted for by the treatment*** and is identified as r^2. It is computed using the obtained *t* value and the *df* value from the hypothesis test, exactly as was done for the single-sample *t* and for the independent-measures *t*.

$$r^2 = \frac{t^2}{t^2 + df}$$

The letter *r* is the traditional symbol used for a correlation. In the context of *t* statistics, the percentage of variance that we are calling r^2 is often identified by the Greek letter omega squared (ω^2).

<u>For example</u>:

A sample of $n = 16$ individuals is selected from a population with a mean of $\mu = 80$. A treatment is administered to the sample and, after treatment, the sample mean is found to be $M = 86$ with a standard deviation of $s = 8$.

a. Does the sample provide sufficient evidence to conclude that the treatment has a significant effect? Test with $\alpha = .05$.
b. Compute r^2 to measure the effect size.

Answers:

a. ***H_0*: $\bar{x}$ = *μ*** (There is no significant difference between the sample and population means)

   H_1*: $\bar{x}$ ≠ *μ (There is a significant difference between the sample and population means)

   The estimated standard error is $\sqrt{(s^2/n)} = \sqrt{(64/16)} = 2$ points and the data produce $t = 6/2 = 3.00$. With $df = 15$ in two-tailed testing and the significance level 0.05, the critical value of $t = 2.131$. Because the t-obtained (3) is higher than the critical value (2.131), H_0 is rejected and H_1 is accepted. It means that there is a significant difference between the sample and population means. (There is a significant treatment effect.)

b. For ***these*** data, $r^2 = 9/24 = 0.375$ or 37.5%.

Hypothesis testing with two samples

When a research study is designed to assess treatment effectiveness, probably the first statistic to come to mind is the *t*-test. This statistic is generally appropriate for two sample comparison designs (sometimes called two group comparison designs). Being a parametric statistical procedure, several assumptions have to be met before the *t*-test can be properly used. Another important consideration is whether the two-sample comparison is between independent or related samples. For example, a social psychologist may want to compare men and women in terms of their political attitudes, an educational psychologist may want to compare two methods for teaching mathematics, or a clinical psychologist may want to evaluate a therapy technique by comparing depression scores for patients before therapy with their scores after therapy. In each case, the research question concerns a mean difference between two sets of data.

There are two general research designs that can be used to obtain the two sets of data to be compared:

(a) The two sets of data could come from two completely separate groups of participants. For example, the study could involve a sample of men compared with a sample of women. The study could compare grades for one group of freshmen who are given laptop computers with grades for a second group who are not given computers. This research strategy, using completely separate groups, is called an ***independent-measures*** research design or a ***between-subjects*** design. These terms emphasize the fact that the design involves separate and independent samples and makes a comparison between two groups of individuals. With an independent-measures design, there is always a risk that the results are biased because the individuals in one sample are systematically different (smarter, faster, more extroverted, and so on) than the individuals in the other sample.

(b) The two sets of data could come from the same group of participants. For example, the researcher could obtain one set of scores by measuring depression for a sample of patients before they begin therapy and then obtain a second set of data by measuring the same individuals after 6 weeks of therapy. This strategy is called a ***repeated-measures design,*** or a ***within-subject design,*** in which the dependent variable is measured two or more times for each individual in a single sample. The same group of subjects is used in all of the treatment conditions. The main advantage of a repeated-measures study is that it uses exactly the same individuals in all treatment conditions. Thus, there is no risk that the participants in one treatment are substantially different from the participants in another.

Another strategy to produce correlated samples is *matching*. In a **matched-subjects design,** each individual in one sample is matched with an individual in the other sample. The matching is done so that the two individuals are equivalent (or nearly equivalent) with respect to a specific variable that the researcher would like to control. The goal of the matching process is to simulate a repeated-measures design as closely as possible. In a repeated-measures design, the matching is perfect because the same individual is used in both conditions. In a matched-subjects design, however, the best you can get is a degree of match that is limited to the variable(s) that are used for the matching process.

t-test is commonly used to determine whether the means within two groups deviate from each other. There are two *t*-test procedures for comparing the means of two groups.

(a) ***independent samples t-test***, for comparing two independent groups, and
(b) ***dependent samples t-test***, for comparing two dependent groups.

One advantage of the related *t*-test over the independent *t*-test is that statistical significance is attained, at a specified *p*-level, with a smaller difference between the two means (assuming other important attributes are equal). Should the independent t-test be considered inappropriate, an alternative nonparametric procedure is the **Wilcoxon Mann-Whitney test**. The null hypothesis is that the two samples have the same population distribution. An alternative nonparametric procedure to the repeated measures *t*-test is the **Wilcoxon Signed Ranks** test.

Hypothesis testing with two independent samples

Independent samples t-test is sometimes called **unpaired samples *t-test*, unmatched samples *t-test*,** or **uncorrelated samples *t-test***.

The goal of an independent-measures research study is to evaluate the mean difference between two populations (or between two treatment conditions). Because an independent-measures study involves two separate samples, we need some special notation to help specify which data go with which sample. This notation involves the use of subscripts, which are small numbers written beside a sample statistic. For example, the number of scores in the first sample would be identified by n_1; for the second sample, the number of scores is n_2. The sample means would be identified by *M*1 and *M*2. The sums of squares would be *SS*1 and *SS*2.

H_0: $\mu_1 = \mu_2$ (There is no significant difference between the two population means)

H_1*: *$\mu_1 \neq \mu_2$ (There is a significant difference between the two population means)

t =

$$\frac{\bar{x}_1 - \bar{x}_2}{\text{standard error of difference between means}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\Sigma x_1^2 - (\Sigma x_1)^2/N_1 + \Sigma x_2^2 - (\Sigma x_2)^2/N_2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Provided that the samples are large (from 30 upwards), z-test can be used:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\text{standard error of difference}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}$$

<u>For example</u>:

The following scores are obtained by two groups of subjects on a language proficiency test:

Group A : 41, 58, 62, 51, 48, 34, 64, 50, 53, 60, 44

Group B : 38, 40, 64, 47, 51, 49, 32, 44, 61

The researcher predicts that the means will differ significantly. Carry out a test to assess significance at the 5% level, giving reasons for selecting the test you decide to use.

<u>Answer</u>:

Since the samples are small, an *independent samples* t-test must be used.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\Sigma x_1^2 - (\Sigma x_1)^2/N_1 + \Sigma x_2^2 - (\Sigma x_2)^2/N_2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

H_0*: *$\mu_1 = \mu_2$ (There is no significant difference between the two population means)
H_1*: *$\mu_1 \neq \mu_2$ (There is a significant difference between the two population means)

The result of the calculation shows that t = 0.91.
The critical value of t for the 5% level and df = 11+9-2 = 18 in two-tailed testing is 2.101. Since the calculated value (0.91) is lower than the critical value (2.101), H_0 is retained. It means that there is no significant difference between the two population means. (The means ***do not differ significantly*** at the 5% level.)

Hypothesis testing with two dependent samples

Dependent samples t-test is sometimes called **paired samples *t-test*, matched samples *t-test* , dependent samples *t-test* ,** or **within samples *t-test***.

In a repeated-measures design or a matched-subjects design comparing two treatment conditions, the data consist of two sets of scores, which are grouped into sets of two, corresponding to the two scores obtained for each individual or each matched pair of subjects. Because the scores in one set are directly related, one-to-one, with the scores in the second set, the two research designs are statistically equivalent and share the common name *related-samples* designs (or *correlated-samples* designs).

The *t* statistic for a repeated-measures design is structurally similar to the other *t* statistics. It is essentially the same as the single-sample *t* statistic. The major distinction of the related-samples *t* is that it is based on *difference scores* rather than raw scores (*X* values). The sample data are represented by the sample mean of the difference scores (M_D)

H_0: $\mu_D = 0$ (There is no significant difference between the two sample means)
H_1: $\mu_D \neq 0$ (There is a significant difference between the two sample means)

$$t = \frac{\Sigma d}{\sqrt{\frac{N\Sigma d^2 - (\Sigma d)^2}{N-1}}}$$

<u>For example</u>:

The following show the scores of pretest and posttest on a reading test:

Pretest : 30, 41, 34, 28, 35, 39, 40, 29, 27, 33

Posttest : 27, 36, 35, 30, 38, 44, 46, 31, 33, 37

The researcher predicts that the means will differ significantly. Carry out a test to assess significance at the 5% level, giving reasons for selecting the test you decide to use.

<u>Answer</u>:

Since the samples are small, a *paired samples* t-test must be used.

$$t = \frac{\Sigma d}{\sqrt{\frac{N\Sigma d^2 - (\Sigma d)^2}{N-1}}}$$

H_0: $\mu_D = 0$ (There is no significant difference between the pretest and posttest means)

H_1: $\mu_D \neq 0$ (There is a significant difference between the pretest and posttest means)

The results of the calculation show that Σd = -21; Σd^2 = 165; t = -1.81. The critical value of t for the 5% level and df=9 in a non-directional test is 1.833.

Since the calculated value t (1.81) is lower than the critical value (1.833), H_0 is retained. It means that there is no significant difference between the pretest and posttest means. (The means ***are not significantly different*** at the 5% level.)

Using* Excel *to perform t-test

Critical t-scores can be generated by using Excel's TINV function, which has the following characteristics:

TINV(probability, deg-freedom)

Where:

Probability = the level of significance, α, for a two-tail test

Deg-freedom = the number of degrees of freedom

Cell A1 contains the Excel formula =TINV(0.05, 6) with the result being 2.447.

A one-tail test requires a slight modification. We need to multiply the probability in the TINV function by two because this parameter is based on a two-tail test.

Cell A1 contains the Excel formula =TINV(2*0.05, 9) with the result being 1.833, $\alpha = 0.05$ and $df = 9$

Microsoft Excel can help us with the *p*-value. Here is the format.

TDIST(test statistic, degrees of freedom, number of tails)

If we had a two-tailed test, with 11 degrees of freedom, and the test statistic calculated to be 2.038, we would find the *p*-value by typing the following: TDIST(2.038,11,2). The result is 0.0663, which is the *p*-value.

If we had a right-tailed test, with 13 degrees of freedom, and the test statistic calculated to be 2.405, we would find the *p*-value by typing the following: TDIST(2.405,13,1). The result is 0.0159, which is the *p*-value.

From the Tools menu, choose Data Analysis and select t-Test: Two-Sample Assuming Equal Variances.

For example, B1:B12 for Variable 1 Range and cells A1:A10 for Variable 2 Range. Set the Hypothesized Mean Difference to 0, Alpha to 0.05, and Output Range to cell D1.

t-Test: Two-Sample Assuming Equal Variances		
	Group A	*Group B*
Mean	51.36363636	47.3333333
Variance	87.05454545	108.5
Observations	11	9
Pooled Variance	96.58585859	
Hypothesized Mean Difference	0	
df	18	
t Stat	0.912396327	
P(T<=t) one-tail	0.186809472	
t Critical one-tail	1.734063592	
P(T<=t) two-tail	0.373618943	
t Critical two-tail	2.100922037	

From the Tools menu, choose Data Analysis and select t-Test: Paired Two Sample for Means.
For example, B1:B12 for Variable 1 Range and cells A1:A10 for Variable 2 Range. Set the Hypothesized Mean Difference to 0, Alpha to 0.05, and Output Range to cell D1.

t-Test: Paired Two Sample for Means		
	Pretest	*Posttest*
Mean	33.6	35.7
Variance	26.26666667	35.56666667
Observations	10	10
Pearson Correlation	0.791755889	
Hypothesized Mean Difference	0	
df	9	
t Stat	-1.811871522	
P(T<=t) one-tail	0.051713828	
t Critical one-tail	1.833112923	
P(T<=t) two-tail	0.103427656	
t Critical two-tail	2.262157158	

Using SPSS to perform t-test

Following are detailed instructions for using SPSS to perform **The Independent-Measures *t* Test**.

Data Analysis

1. Click **Analyze** on the tool bar, select **Compare Means,** and click on **Independent-Samples *t* Test.**
2. Highlight the column label for the set of scores (VAR0001) in the left box and click the arrow to move it into the **Test Variable(s)** box.
3. Highlight the label from the column containing the sample numbers (VAR0002) in the left box and click the arrow to move it into the **Group Variable** box.
4. Click on **Define Groups.**
5. Assuming that you used the numbers 1 and 2 to identify the two sets of scores, enter the values 1 and 2 into the appropriate group boxes.
6. Click **Continue.**
7. In addition to performing the hypothesis test, the program computes a confidence interval for the population mean difference. The confidence level is automatically set at 95% but you can select **Options** and change the percentage.
8. Click **OK.**

Following are detailed instructions for using SPSS to perform **The Repeated-Measures *t* Test**.

Data Entry

Enter the data into two columns (VAR0001 and VAR0002) in the data editor with the first score for each participant in the first column and the second score in the second column. The two scores for each participant must be in the same row.

Data Analysis

1. Click **Analyze** on the tool bar, select **Compare Means**, and click on **Paired- Samples t Test.**
2. One at a time, highlight the column labels for the two data columns and click the arrow to move them into the **Paired Variables** box.

3. In addition to performing the hypothesis test, the program computes a confidence interval for the population mean difference. The confidence level is automatically set at 95%, but you can select **Options** and change the percentage.
4. Click **OK.**

t-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Pretest	33.60	10	5.125	1.621
	Posttest	35.70	10	5.964	1.886

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Pretest - Posttest	-2.100	3.665	1.159	-4.722	.522	-1.812	9	.103

References

Donnelly, R.A. (2007). *The complete idiot's guide to statistics*. (2nd Edition). Indianapolis: Alpha Books, a member of Penguin Group, Inc.

Gravetter, F.J., and Wallnau, L.B. (2013). *Statistics for the behavioralsSciences*. (9 Ed.). Belmont, C.A.: Wadsworth, Cengage Learning.

Grotenhuis, M., and Matthijssen, A. (2016). *Basic SPSS tutorial*. Thousand Oaks, California: SAGE Publications.

Huck, S.W. (2012). *Reading statistics and research*. (Sixth Edition). Boston: Pearson.

Pramesti, G. (2014). *Kupas tuntas data penelitian dengan SPSS 22*. Jakarta: Penerbit PT Elex Media Komputindo.

Rumsey, D. (2010). *Statistics essentials for dummies*. Hoboken, N.J.: Wiley Publishing, Inc.

Winarno, E., Zaki, A., & SmitDev Community. (2014). *Statistik dan analisisnya dengan Excel*. Jakarta: Penerbit PT Elex Media Komputindo.

Exercises

1. The scores of a random sample of 8 students on a physics test are as follows: **60, 62, 67, 69, 70, 72, 75**, and **78**. Test to see if the sample mean is significantly different from 65 at the .05 level. Report the t and p values.

2. A principal wonders if her 5th graders score differently on a math test than 5th graders in the U.S. at large. She gets a random sample of 20 5th graders from her school. She knows that 5th graders in the U.S. at large have a mean of **88** on the test. Here are the scores of her sample of 20 5th graders: **75, 92, 85, 66, 93, 88, 75, 90, 90, 92, 84, 88, 67, 98, 99, 100, 79, 95, 88, 89.**

3. A researcher is interested in the number of days that successful athletes train to run a full marathon. She gets the workout logs for one year for all the athletes who completed a specific marathon in less than four hours. On the average, these athletes of the past trained for 305 days per year. She is in charge of the training for 20 athletes. After they run the marathon, she wonders whether their training differed in amount from the training of successful athletes in the past. Here are the number of days per year her 20 athletes trained: **230, 220, 214, 300, 310, 299, 200, 240, 305, 288, 297, 215, 199, 233, 255, 280, 303, 277, 280, 299**. Determine if the students' mean number of training days per year differs from the mean of 305 days in the past.

4. Participants threw darts at a target. In one condition, they used their preferred hand; in the other condition, they used their other hand. All subjects performed in both conditions (the order of conditions was counterbalanced). Their scores are shown below.

Preferred	Non-preferred
12	7
7	9
11	8
13	10

10	9

a. Which kind of t-test should be used?
b. Calculate the two-tailed t and p values using this t test.
c. Calculate the one-tailed t and p values using this t test.

5. The scores on a (hypothetical) vocabulary test of a group of 20 year olds and a group of 60 year olds are shown below. Test the mean difference for significance using the .05 level.

20 yr olds	60 yr olds
27	26
26	29
21	29
24	29
15	27
18	16
17	20
12	27
13	

6. The following are four sets of raw scores.
Data 1: 41, 58, 62, 51, 48, 34, 64, 50, 53, 60, 44, 69, 70, 59, 63, 57
Data 2: 38, 40, 64, 47, 51, 49, 32, 44, 61, 56, 60, 74, 76, 61, 63, 67
Data 3: 45, 58, 60, 51, 53, 59, 54, 40, 56, 56, 60, 71, 64, 58, 65
Data 4: 48, 58, 71, 56, 59, 62, 64, 62, 52, 69, 57, 66, 65, 60, 68

A teacher teaches vocabulary to her students by using two different media. Suppose that Data 1 and 2 are the results of the pretest and posttest in the control group, and Data 3 and 4 are the results of the pretest and posttest in the experimental group.
a. Is there any significant improvement in the control group?
b. Is there any significant improvement in the experimental group?
c. Do the students in the experimental group outperform those in the control group?

7. An investigator thinks that people under the age of forty have vocabularies that are different than those of people over sixty years of age. The investigator administers a vocabulary test to a group of 31 younger subjects and to a group of 31 older subjects. Higher scores reflect better performance. The mean score for younger subjects was 14.0 and the standard deviation of younger subject's scores was 5.0. The mean score for older subjects was 20.0 and the standard deviation of older subject's scores was 6.0. Does this experiment provide evidence for the investigator's theory?

   As part of your answer:

   a. Please provide, in words, the null and alternative hypotheses.
   b. Provide the decision rule for rejecting the null hypothesis, including the critical value(s) for the appropriate statistic.
   c. Using an alpha level of .05, test the null hypothesis. As part of this test, please compare the actual value for the appropriate statistic against the critical value(s) for the appropriate statistic.
   d. State your conclusion regarding the results from this test in language that a friend of yours with no knowledge of statistics could understand.

8. What is the difference in the averages between the two groups? Is this difference statistically significant?

   Table: Patients with high blood pressure

Group	Blood pressure
1=placebo	90
1	95
1	67
1	120
1	89
1	92
1	100
1	82
1	79
1	85
2=new drug	71
2	79
2	69
2	98
2	91
2	85
2	89
2	75
2	78
2	80

ONE-WAY ANOVA

Use ***analysis of variance*** (ANOVA) to compare the means of three or more populations.

A *factor* in ANOVA describes the cause of the variation in the data. When only one factor is being considered, the procedure is known as one-way ANOVA. A *level* in ANOVA describes the number of categories within the factor of interest.

The simplest type of ANOVA is known as ***completely randomized one-way ANOVA***, which involves an independent random selection of observations for each level of one factor.

To use one-way ANOVA, the following conditions must be present: (1) the populations of interest must be normally distributed, (2) the samples must be independent of each other, and (3) each population must have the same variance.

The hypotheses would look like the following:
H_0 : $\mu_1 = \mu_2 = \mu_3$ (All the population means are equal.)
H_1 : not all μ's are equal
If the null hypothesis is rejected, it means that a difference does exist. Analysis of variance could not compare population means to one another to determine which is greater. That task requires further analysis.

Partitioning the Sum of Squares

The hypothesis test for ANOVA compares two types of variations from the samples. The total variation in the data from the samples can be divided, or as statisticians like to say, "partitioned," into two parts. The first part is the variation within each sample, which is officially known as the ***sum of squares within*** (***SSW***) or the ***error sum of squares*** (***SSE***). This can be found using the following equation:

$$SSW = \sum_{i=1}^{k} (n_i - 1)\, s_i^2$$

where k = the number of samples (or levels).

For example: $s_1^2 = 1.01$ $s_2^2 = 1.70$ $s_3^2 = 0.96$
$n_1 = 6$ $n_2 = 6$ $n_3 = 6$

Note: ANOVA does *not* require that all the sample sizes are equal.

The sum of squares within can now be calculated as:
$SSW = (6 – 1)1.01 + (6 – 1)1.70 + (6 – 1)0.96 = 18.35$

The second partition is the variation among the samples, which is known as the ***sum of squares***
between (***SSB***). This can be found by:

$$SSB = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{\bar{x}})^2$$

where $\bar{\bar{x}}$ is the grand mean or the average value of all the observations.

For example: $\bar{x}_1 = 9.12$ $\bar{x}_2 = 10.92$ $\bar{x}_3 = 9.48$
Find the grand mean, using:

$$\bar{\bar{x}} = \frac{\Sigma x}{N}$$

where N = the total number of observations from all samples.

Calculate the sum of squares between:

$$SSB = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{\bar{x}})^2$$

$SSB = 6\,(9.12- 9.83)^2 + 6\,(10.92- 9.83)^2 + 6\,(9.48 – 9.83)^2 = 10.86$

Finally, the total variation of all the observations is known as the total sum of squares (SST) and can be found by:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{b} (x_{ij} - \bar{\bar{x}})^2$$

This total sum of squares calculation can be confirmed recognizing that: *SST* = *SSW* + *SSB*
SST = 18.35 + 10.86 = 29.21

Note that the variance of the original 18 observations, s^2, can be determined by:

$$s^2 = \frac{SST}{N-1} = \frac{29.21}{18-1} = 1.72$$

Determining the Calculated F-Statistic

To test the hypothesis for ANOVA, compare the calculated test statistic to a critical test statistic using the F-distribution. The calculated F-statistic can be found using the equation:

$$F = \frac{MSB}{MSW}$$

where *MSB* is the *mean square between,* found by:

$$MSB = \frac{SSB}{k - 1}$$

and *MSW* is the *mean square within,* found by:

$$MSW = \frac{SSW}{N - k}$$

For example:

$$MSB = \frac{SSB}{k\text{-}1} = \frac{10.86}{3\text{-}1} = 5.43$$

$$MSW = \frac{SSW}{N\text{-}k} = \frac{18.35}{18\text{-}3} = 1.22$$

$$F = \frac{MSB}{MSW} = \frac{5.43}{1.22} = 4.45$$

The mean square between (*MSB*) is a measure of variation between the sample means. The mean square within (*MSW*) is a measure of variation within each sample. A large *MSB* variation, relative to the *MSW* variation, indicates that the sample means are not very close to one another. This condition will result in a large value of *F*, the calculated F-statistic. The larger the value of *F*, the more likely it will exceed the critical F-statistic, leading us to conclude there is a difference between population means.

Determining the Critical F-Statistic

We use the F-distribution to determine the critical F-statistic, which is compared to the calculated F-statistic for the ANOVA hypothesis test. The critical F-statistic, $F_{\alpha,k\text{-}1,N\text{-}k}$, depends on two different degrees of freedom, which are determined by: $v_1 = k - 1$ and $v_2 = N - k$

For example:
$v_1 = 3 - 1 = 2$ and $v_2 = 18 - 3 = 15$
The critical F-statistic is read from the F-distribution table.
For $v_1 = 2$ and $v_2 = 15$, the critical F-statistic, $F_{.05,2,15} = 3.682$

Because F-calculated (4.45) > F-table (3.682), H_0 is rejected.
Therefore, H_1 is accepted.
It means that the population means are *not equal*.

Using Excel to Perform One-Way ANOVA

(1) Start by placing the data in Columns A, B, and C in a blank sheet.
(2) Go to the Tools menu and select Data Analysis.
(3) From the Data Analysis dialog box, select **Anova: Single Factor** and click OK.

For example:
Group 1: 10.2, 8.5, 8.4, 10.5, 9, 8.1
Group 2: 11.6, 12, 9.2, 10.3, 9.9, 12.5
Group 3: 8.1, 9, 10.7, 9.1, 10.5, 9.5

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Group 1	6	54.7	9.116666667	1.005666667
Group 2	6	65.5	10.91666667	1.701666667
Group 3	6	56.9	9.483333333	0.961666667

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	10.857778	2	5.428888889	4.438993368	0.03059861	3.682320344
Within Groups	18.345	15	1.223			
Total	29.202778	17				

Notice that the *p*-value = 0.0305 for this test, meaning that H_0 is rejected because this *p*-value ≤ α (0.05). H_1 is accepted. It means that not all the population means are equal.

Pairwise Comparisons

Once we have rejected H_0 using ANOVA, we can determine which of the sample means are different using the Scheffé test. This test compares each pair of sample means from the ANOVA procedure. For example, we would compare $\bar{x}_1$ versus $\bar{x}_2$, $\bar{x}_1$ versus $\bar{x}_3$, and $\bar{x}_2$ versus $\bar{x}_3$ to see whether any differences exist.

The Scheffé test, F_S, is as follows:

$$F_s = \frac{(\bar{X}_a - \bar{X}_b)^2}{\frac{SSW}{\sum^{k}(n_i - 1)}\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}$$

The critical value for the Scheffé test, F_{SC}, is determined by multiplying the critical F-statistic from the ANOVA test by (k – 1) as follows: $F_{SC} = (k-1)\, F_{\alpha,\, k-1,\, N-k}$

For example: $F_{.05,2,15} = 3.682$
$F_{SC} = (3–1)(3.682) = 7.364$

If $F_S \leq F_{SC}$, we conclude there is no difference between the sample means; otherwise, there is a difference.

For example:
Summary of Scheffé Test

Sample Pair	F_S	F_{SC}	**Conclusion**
x_1 and x_2	8.048	7.364	Difference
x_1 and x_3	0.323	7.364	No Difference
x_2 and x_3	5.142	7.364	No Difference

According to the results, the only statistically significant difference is between Groups 1 and 2.

Using SPSS to Perform One-Way ANOVA

(1) Determine the factor (for example, methods A, B, C) and the dependent variable (the test results)
(2) Click Analyze, Compare Means, One-Way ANOVA
(3) Click Post Hoc, click LSD (Scheffe), Continue
(4) Click Options, Descriptive, Homogeneity of variance test, Means Plot, Continue, OK

Method A: 10.2, 8.5, 8.4, 10.5, 9, 8.1
Method B: 11.6, 12, 9.2, 10.3, 9.9, 12.5
Method C: 8.1, 9, 10.7, 9.1, 10.5, 9.5

Oneway

Descriptives

Results

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
Method A	6	9.117	1.0028	.4094	8.064	10.169	8.1	10.5
Method B	6	10.917	1.3045	.5326	9.548	12.286	9.2	12.5
Method C	6	9.483	.9806	.4003	8.454	10.512	8.1	10.7
Total	18	9.839	1.3107	.3089	9.187	10.491	8.1	12.5

Test of Homogeneity of Variances

Results

Levene Statistic	df1	df2	Sig.
.989	2	15	.395

ANOVA

Results

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	10.858	2	5.429	4.439	.031
Within Groups	18.345	15	1.223		
Total	29.203	17			

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Results

	(I) Methods	(J) Methods	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Scheffe	Method A	Method B	-1.8000*	.6385	.041	-3.533	-.067
		Method C	-.3667	.6385	.849	-2.099	1.366
	Method B	Method A	1.8000*	.6385	.041	.067	3.533
		Method C	1.4333	.6385	.114	-.299	3.166
	Method C	Method A	.3667	.6385	.849	-1.366	2.099
		Method B	-1.4333	.6385	.114	-3.166	.299
LSD	Method A	Method B	-1.8000*	.6385	.013	-3.161	-.439
		Method C	-.3667	.6385	.574	-1.728	.994
	Method B	Method A	1.8000*	.6385	.013	.439	3.161
		Method C	1.4333*	.6385	.040	.072	2.794
	Method C	Method A	.3667	.6385	.574	-.994	1.728
		Method B	-1.4333*	.6385	.040	-2.794	-.072

*. The mean difference is significant at the 0.05 level.

Homogeneous Subsets

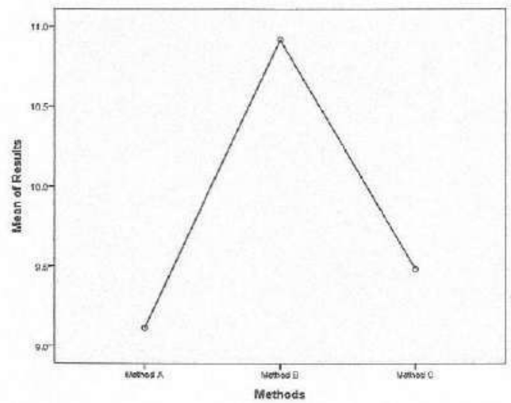
Results

	Methods	N	Subset for alpha = 0.05	
			1	2
Scheffe[a]	Method A	6	9.117	
	Method C	6	9.483	9.483
	Method B	6		10.917
	Sig.		.849	.114

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 6.000.

Means Plots



H_0 : $\mu_1 = \mu_2 = \mu_3$ (All the three means are equal.)
H_1 : not all μ's are equal (Not all the three means are equal.)
The results of one-way ANOVA show that the F-obtained = 4.439 and the p-value = 0.031. Since the p-value (0.031) is lower than 0.05, H_0 is rejected and H_1 is accepted. It means that not all the three means are equal.
The results of Scheffe's test show that there is a significant difference between Method A and Method B (p-value 0.041 < 0.05); there is no significant difference between Method A and Method C (p-value 0.849 > 0.05); there is no significant difference between Method B and Method C (p-value 0.114 > 0.05).
$\bar{x}_A$ = 9.117; $\bar{x}_B$ =10.917; therefore, Method B is significantly better than Method A.

References

Donnelly, R.A. (2007). *The complete idiot's guide to statistics*. (2nd Ed.). Indianapolis: Alpha Books, a member of Penguin Group, Inc.

Pramesti, G. (2014). *Kupas tuntas data penelitian dengan SPSS 22*. Jakarta: Penerbit PT Elex Media Komputindo.

Exercises

1. Not all hotdogs have the same calories. The following table contains calorie data on a random sample of Beef, Poultry, and Veggie dogs. (One extreme outlier for Veggie dogs was omitted from the data.) Does the mean calorie count differ depending on the type of hotdog? Run an ANOVA. State the value of the F-statistic, the degrees of freedom for F, the p-value of the test, and your conclusion.

   Table: Calorie content of hotdogs

Beef	Poultry	Veggie
110	60	40
110	60	45
130	60	45
130	70	45
140	70	50
150	70	50
160	80	55
160	90	57
170	90	60
170	100	60
175	100	70
180	100	80
180	110	80
180	110	81
190	110	90
190	120	95
190	120	100
200	130	100
210	140	110
230	150	

2. You are planning an experiment that will involve 4 equally sized groups, including a control group and 3 experimental groups. Each

group will contain 10 observations. Your expectation is that each of the 3 experimental treatments will have approximately the same effect, and that this effect will be small — roughly one-third a standard deviation improved performance over the control. Suppose that you run the above experiment, and obtain the data shown below. Perform a one-way ANOVA on the data.

No.	Control	Experiment 1	Experiment 2	Experiment 3
1	118	107	133	134
2	121	165	154	176
3	97	121	91	171
4	86	126	63	159
5	118	87	62	118
6	45	135	164	125
7	119	83	96	100
8	92	100	129	60
9	91	144	128	163
10	72	119	105	111

NON-PARAMETRIC TESTS

Four nonparametric or *distribution free* statistics that make use of rank scores are introduced: the Spearman Rank-Order Correlation Coefficient, the Mann-Whitney U-test, the Wilcoxon Signed-Ranks test, and the Sign test.

Nonparametric statistical tests are generally less powerful than parametric tests and are also less likely to mislead investigators because they are not dependent upon certain restrictive measurement and distributional assumptions. Nonparametric procedures are sometimes an effective way of dealing with non-normal or unknown distributions but are not always the answer to unequal variances. Nonparametric procedures are also well suited to small sample sizes, and rank tests are particularly helpful when outliers are present in a data set since ranks of raw scores are not affected by extreme values. Data may naturally form ranks, or ranks may be assigned on the basis of measurement (or combinations of different measures) or subjective judgment.

Spearman Rank-Order Correlation Coefficient

Correlations coefficients, *r*, computed from sample data provide an index of the strength of the relationship between two variables e.g., r_{xy} is the sample correlation between the variables *X* and *Y*. **Spearman rank-order correlation** is appropriate when variables are measured at an ordinal level, or when data is transformed to an ordinal scale. It is one of a number of alternative distribution-free correlation-type statistics.

Spearman rank-order correlation should be used when:
- the relationship between two variables is not linear (this can be checked by plotting the two variables);
- when measurement and distributional assumptions are not met (the variables are not interval or ratio measures and observations do not come from a bivariate normal distribution);
- when sample sizes are too small to establish an underlying distribution, or
- when the data naturally occur in the form of ranks.

Spearman rank-order correlation is equivalent to the Pearson Product Moment correlation (a parametric correlation procedure), performed on the ranks of the scores rather than on the raw scores themselves.

H_0: There is no significant correlation between the two variables. (H_0: $\rho_s = 0$)
H_1: There is a significant correlation between the two variables.

H_0: There is no significant *positive* correlation between the two variables.
H_1: There is a significant *positive* correlation between the two variables.

H_0: There is no significant *negative* correlation between the two variables.
H_1: There is a significant *negative* correlation between the two variables.

The formula of Spearman ρ coefficient is as follows:

$$\rho = 1 - \frac{6 \times \Sigma D^2}{N(N^2 - 1)}$$

where ρ = Spearman rank-order correlation coefficient
D = difference between ranks in each pair
N = number of students for whom you have pairs of ranks
Σ = sum
6 = a constant

Table: Calculating Spearman *rho*

Column 1	2		3		4	5	6
Student	Ranks on Test A		Ranks on Test B		D	D^2	Calculations
Robert	1	–	4	=	–3	9	
Millie	2	–	3	=	–1	1	
Iliana	3	–	2	=	1	1	$\rho = 1 - \frac{6 \times \Sigma D^2}{N(N^2-1)}$
Dean	4	–	1	=	3	9	
Cuny	5	–	5	=	0	0	$= 1 - \frac{6 \times 40}{9(81-1)}$
Bill	6	–	9	=	–3	9	
Corky	7	–	8	=	–1	1	$= 1 - \frac{240}{720}$
Randy	8	–	7	=	1	1	$= 1 - .33$
Monique	9	–	6	=	3	9	$= .67$

*$p < .05$ (with $N = 9$) $\Sigma D^2 = 40$

Using SPSS to Perform the Spearman Rank-Order Correlation Coefficient

Variable View: Measure → Ordinal
Analyze → Correlate → Bivariate → Check *Spearman* in Correlation Coefficients

Correlations

			TestX	TestY
Spearman's rho	TestX	Correlation Coefficient	1.000	.671**
		Sig. (2-tailed)	.	.004
		N	16	16
	TestY	Correlation Coefficient	.671**	1.000
		Sig. (2-tailed)	.004	.
		N	16	16

**. Correlation is significant at the 0.01 level (2-tailed).

H_0: There is no significant correlation between the two variables.
H_1: There is a significant correlation between the two variables.

The results of *Spearman rank-order correlation coefficient* show that the *rho*-obtained is 0.671 and the p-value = 0.004. At the significance level 0.05 in two-tailed testing with N=16, the *rho*-table is 0.503. Since the *rho*-obtained (0.671) is higher than the *rho*-table (0.503), H_0 is rejected and H_1 is accepted. It means that there is a significant correlation between the two variables.

The Mann-Whitney *U*-Test

The Mann-Whitney *U*-test is used to know if two independent sets of data show a significant overall difference in the magnitude of the variable, but *z*-test or *t*-test cannot be used because the assumptions relating to level of measurement, sample size, normality or equality of variance are not valid. The test assumes only an ordinal level of measurement, since it is based on the ranking of scores. The Mann-Whitney *U*-test can be used as a nonparametric counterpart of the *independent-samples* t-test.

H_0: There is no significant difference between the two sets of ratings.
H_1: There is a significant difference between the two sets of ratings.

$$U_1 = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1.$$

$$U_2 = N_1 N_2 + \frac{N_2(N_2+1)}{2} - R_2.$$

$$U_2 = N_1 N_2 - U_1.$$

U = the Mann-Whitney U-test, the smaller of U_1 and U_2
N_1= the number of scores in the smaller group
N_2= the number of scores in the larger group
R_1= the sum of ranks for the smaller group
R_2 = the sum of ranks for the larger group

First, rank the whole combined set of $N_1 + N_2$ scores from the lowest (rank 1) to the highest. If there is more than one occurrence of the same score (that is, tied ranks), each occurrence is given the mean of the ranks which would have been allocated if there had not been a tie. *If the calculated value is* ***smaller*** *than or equal to the critical value, the null hypothesis is rejected.*

For example: Two groups of students are asked to grade the text they have read on a scale of coherence from 0 (totally incoherent) to 10 (totally coherent). The researcher wishes to know whether there is any significant difference between the two sets of ratings at the 5 per cent level in a non-directional test.

Ratings and ranks for combined scores in two versions of a text

Original text ($N_2 = 9$)	Rank	Altered text (N_1=8)	Rank
7	9	7	9
8	13.5	4	1
6	5.5	5	3
9	16	6	5.5
10	17	8	13.5
7	9	5	3
7	9	5	3
8	13.5	7	9
8	13.5		
			$R_1 = 47$

$$U_1 = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1$$

$$= (8 \times 9) + \frac{8(8+1)}{2} - 47 = 61.$$

$U_2 = N_1N_2 - U_1$ = (8x9) –61 = 11

Since U_2 is the smaller of the two values, we have $U = U_2$ =11.
At the significance level 0.05 in two-tailed testing for N_1=8, N_2=9, the critical value of U is 15.
Since the calculated value (11) is less than the critical value (15), H_0 is rejected. It means that there is a significant difference between the two sets of ratings.

For larger samples (more than about 20), calculate a z value as follows:

$$z = \frac{U - N_1N_2/2}{\sqrt{\dfrac{N_1N_2(N_1+N_2+1)}{12}}}.$$

If the calculated value of z is greater than or equal to the critical value for the required significance level, the null hypothesis is rejected.

Suppose that we have calculated U for two samples, one of 30 and one of 35 scores, and have obtained a value of 618.

$$z = \frac{618 - (30 \times 35)/2}{\sqrt{\dfrac{30 \times 35 \times (30 + 35 + 1)}{12}}} = 1.22.$$

If the test is non-directional and at the 5 per cent level, the critical value of z is 1.96. Since the calculated value is smaller than the critical value, the null hypothesis cannot be rejected.

Using SPSS to Perform the Mann Whitney *U*-Test

Variable View: Measure → interval or ordinal

Define Groups

Analyze → Nonparametric Tests → Legacy Dialogs → 2 Independent Samples → Check *Mann- Whitney U* in Test Type

Mann-Whitney Test

Ranks

	Value12Rank	N	Mean Rank	Sum of Ranks
Group12Rank	Group1	9	11.78	106.00
	Group2	8	5.88	47.00
	Total	17		

Test Statistics[a]

	Group12Rank
Mann-Whitney U	11.000
Wilcoxon W	47.000
Z	-2.459
Asymp. Sig. (2-tailed)	.014
Exact Sig. [2*(1-tailed Sig.)]	.015[b]

a. Grouping Variable: Value12Rank

b. Not corrected for ties.

H_0: There is no significant difference between the two sets of ratings.

H_1: There is a significant difference between the two sets of ratings.

The results of *Mann-Whitney U-test* show that the *U*-obtained is 11 and the p-value of z-test = 0.014. At the significance level 0.05 in two-tailed testing for N_1=8, N_2=9, the critical value of U is 15. Since the U-calculated (11) is less than the U-table (15), H_0 is rejected and H_1 is accepted. It means that there is a significant difference between the two sets of ratings. Or since the p-value of z-test (0.014) is lower than 0.05, H_0 is rejected and H_1 is accepted. It means that there is a significant difference between the two sets of ratings.

The Wilcoxon Signed-Ranks Test

The Wilcoxon signed-ranks test can be used as a non-parametric counterpart of the paired-samples t-test. The data for the test consist of a number of pairs of scores, each derived from a single subject, or from a pair of matched subjects.

H_0: There is no significant difference between the two datasets.
H_1: There is a significant difference between the two datasets.

In order to calculate the test statistic, first find the difference between each pair of scores, subtracting consistently and recording the signs. Then, rank the absolute values of the differences, ignoring the signs. If two scores in a pair are the same (that is, if the difference is zero), that pair is ignored altogether. If two values of the difference are tied, they are given the mean of the ranks they would have had if they had been different in value. Each rank is now given the sign of the difference it corresponds to. The sum of the positive ranks is found, and also that of the negative ranks. The smaller of these two sums is the test statistic *W. The value of W must be* ***smaller*** *than or equal to the critical value if the null hypothesis is to be rejected*. It should be noted that in taking an appropriate value for the number of pairs of scores *N*, pairs which are tied, and so have been discarded, are not counted.

For example: A group of 10 students is asked to translate two passages of English into French. The following table shows the number of errors made by the students.

Differences and ranks for translation error scores

Student	Errors in Passage A	Errors in Passage B	A-B	Rank
1	8	10	-2	-4.5
2	7	6	+1	+2
3	4	4	0	-
4	2	5	-3	-7.5
5	4	7	-3	-7.5
6	10	11	-1	-2
7	17	15	+2	+4.5
8	3	6	-3	-7.5
9	2	3	-1	-2
10	11	14	-3	-7.5

The sum of the positive ranks is (+2) + (+4.5) = 6.5, while that of the negative ranks is 38.5. Take the smaller of these (6.5) as the value for *W*. At the significance level 0.05 in a non-directional test for N=9, the critical value is 5. Since the value of *W* (6.5) is larger than the critical value (5), the null hypothesis cannot be rejected. It means that there is no significant difference between the two data sets.

Where the number of pairs is greater than 20, a *z*-score can be calculated from the computed *W* value:

$$z = \frac{W - N(N+1)/4}{\sqrt{\dfrac{N(N+1)(2N+1)}{24}}}.$$

Suppose a value of *W*=209 has been calculated for a set of data with *N*=35.

$$z = \frac{209 - (35 \times 36)/4}{\sqrt{\dfrac{35 \times 36 \times \{(2 \times 35) + 1\}}{24}}} = -1.74.$$

Ignore the sign, which is normally negative. The values of 1.96 and 1.64 are required at the 5 per cent level for a non-directional or a directional test respectively. Therefore, if the test was non-directional,

H_0 could not be rejected. If the test was directional, H_0 could be rejected.

Using SPSS to Perform the Wilcoxon Signed-Ranks Test
Variable View: Measure→interval
Analyze→Nonparametric Tests→Legacy Dialogs→2 Related Samples→Check *Wilcoxon* in Test Type

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
PassageA - PassageB	Negative Ranks	7[a]	5.50	38.50
	Positive Ranks	2[b]	3.25	6.50
	Ties	1[c]		
	Total	10		

a. PassageA < PassageB
b. PassageA > PassageB
c. PassageA = PassageB

Test Statistics[a]

	PassageA - PassageB
Z	-1.921[b]
Asymp. Sig. (2-tailed)	.055

a. Wilcoxon Signed Ranks Test
b. Based on positive ranks.

H_0: There is no significant difference between the two datasets.
H_1: There is a significant difference between the two datasets.

The results of *Wilcoxon signed-ranks test* show that the sum of the positive ranks is 6.5, while that of the negative ranks is 38.5 (so the value for *W* is the smaller one = 6.5) and the p-value of z-test is 0.055. At the significance level 0.05 in a non-directional test for N=9, the critical value of *W* is 5. Since the value of *W* (6.5) is larger than the critical value (5), H_0 cannot be rejected. It means that there is no significant difference between the two datasets. Or since the p-value of z-test (0.055) is higher than 0.05, H_0 cannot be rejected. It means that there is no significant difference between the two datasets

The Sign Test

In many kinds of investigations of interest, only an ordinal level can be achieved. For example, the case where informants are asked to rate two sentences on a scale of acceptability or politeness. Such variables cannot be measured in units with equal intervals, and we cannot attach much importance to the magnitude of differences between ratings; we could, however, claim that a sentence rated 4 on an acceptability scale had been rated as more acceptable than a sentence given a rating of 3. We can take into account the direction of the differences between pairs, even though we cannot use their magnitude. The loss of information incurred in ignoring the magnitude of the differences means that the so-called sign test is less powerful than the Wilcoxon signed-ranked test, which does take the magnitude into account.

H_0: There is no significant difference between the two datasets.
H_1: There is a significant difference between the two datasets.

To carry out the sign test, first record the sign the difference for each pair of scores, subtracting consistently. Tied scores are dropped from the analysis, and the number of pairs (*N*) reduced accordingly. Now find the number of pairs with the less frequent sign, and call it *x*. *If the computed value of x is **smaller** than or equal to the critical value, the null hypothesis is rejected.*

For example: A group of subjects has been asked to rate a sentence on a scale of acceptability from 0 (totally unacceptable) to 5 (totally acceptable) for (a) informal spoken English and (b) formal written English. The researcher predicts that the sentence will be judged as more acceptable in informal spoken than in formal written English.

Sign of differences in acceptability scores

Student	Informal spoken	Formal written	Sign of (Informal-Formal)
1	5	5	0
2	4	2	+
3	5	3	+
4	4	4	0
5	3	1	+
6	2	3	−
7	4	3	+
8	5	1	+
9	4	2	+
10	2	3	−
11	4	2	+
12	4	3	+
13	5	3	+
14	3	5	−
15	3	0	+

We obtain the sign of each difference, and discount tied scores. We have 3 negative and 10 positive difference, so $x = 3$ and $N = 13$. At the significance level 0.05 in a directional test for N=13, the critical value of x is 3. Since the calculated value of x is equal to the critical value, the null hypothesis is rejected. It means that there is a significant difference between the two data sets.

If N is greater than about 25, we calculate a z value as follows:

$$z = \frac{N - 2x - 1}{\sqrt{N}}.$$

Suppose we have 100 pairs of (non-tied) scores and 42 of the differences are positive and 58 negative. Then, x=42.

$$z = \frac{100 - (2 \times 42) - 1}{\sqrt{100}} = 1.50.$$

This z value falls below the critical value for either a directional or a non-directional test at the 5 per cent level, so we should conclude that there is no significant difference between the two sets of scores.

Using SPSS to Perform the Sign Test

Variable View: Measure→interval or ordinal

Analyze→Nonparametric Tests→Legacy Dialogs→2 Related Samples→Check *Sign* in Test Type

Sign Test

Frequencies

		N
Informal - Formal	Negative Differences[a]	3
	Positive Differences[b]	10
	Ties[c]	2
	Total	15

a. Informal < Formal
b. Informal > Formal
c. Informal = Formal

Test Statistics[a]

	Informal - Formal
Exact Sig. (2-tailed)	.092[b]

a. Sign Test
b. Binomial distribution used.

H_0: There is no significant difference between the two datasets.
H_1: There is a significant difference between the two datasets.

The results of *sign test* show that there are 3 negative and 10 positive difference (so $x = 3$), and 2 ties (so $N = 15\text{-}2 = 13$), and the p-value (2-tailed) is 0.092.

At the significance level 0.05 in a *directional test* for N=13, the critical value of x is 3. Since the calculated value of x is equal to the critical value, H_0 is rejected and H_1 is accepted. It means that there is a significant difference between the two datasets.

At the significance level 0.05 in a *non-directional* test for N=13, the critical value of x is 2. Since the calculated value of x (3) is higher than the critical value (2), H_0 is retained. It means that there is no significant difference between the two datasets. Or since the p-value

(2-tailed) (0.092) is higher than 0.05, H_0 is retained. It means that there is no significant difference between the two datasets.

Deciding which test to use

The following flowchart shows the steps to be taken in deciding which of the three tests (the Mann-Whitney U-test, the Wilcoxon signed-ranks test, the sign test) should be used in a given case.

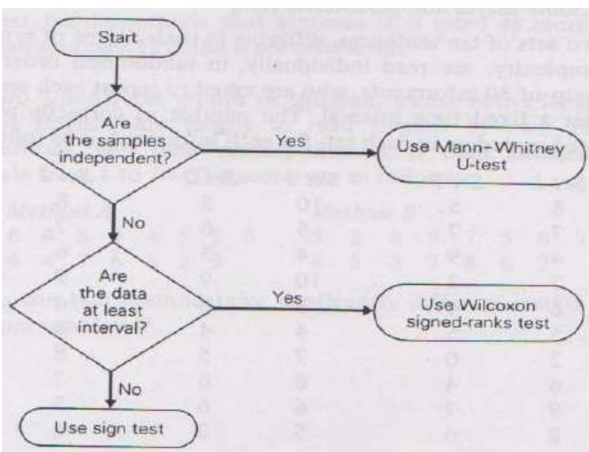


Figure: Deciding which test to use

References

Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, N.J.: Prentice Hall Regents.

Butler, C.S. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell Ltd.

Peers, I.S. (1996). *Statistical analysis for education and psychology researchers*. London: The Falmer Press.

Exercises

1. Calculate a Pearson *r* correlation coefficient and calculate a Spearman ρ correlation coefficient, and determine whether they are statistically significant.

Student	Test Y Scores	Test Z Scores	Test Y Ranks	Test Z Ranks
1	77	87	1	1
2	75	75	2	2
3	64	72	14.5	3
4	72	61	3.5	4
5	70	60	5.5	5.5
6	70	60	5.5	5.5
7	69	59	8.5	7
8	69	58	8.5	8
9	69	57	8.5	10
10	69	57	8.5	10
11	68	57	11.5	10
12	72	56	3.5	12
13	67	52	13	13
14	68	49	11.5	14
15	64	30	14.5	15
16	61	21	16	16

2. The following represent a teacher's assessment of reading skill for two groups of children.
   Group 1: 8, 6, 3, 5, 8, 7, 7, 6, 5, 6, 6, 7, 8
   Group 2: 4, 6, 3, 3, 7, 7, 5, 5, 4, 4, 6, 5, 6, 5, 4
   Test, at the 2.5 per cent significance level, the hypothesis that the first group has a higher level of reading skill than the second group.

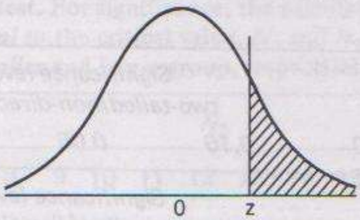
3. Two sets of 10 sentences are read individually to a group of 30 students. They are asked to repeat each sentence after a fixed time interval. The number of correctly remembered sentences in each set, for each informant, is as follows:
   Set 1: 8, 7, 4, 7, 6, 5, 2, 6, 9, 8, 10, 5, 4, 10, 8, 4, 7, 8, 6, 5, 8, 7, 6, 9, 4, 3, 8, 7, 5, 6
   Set 2: 5, 7, 9, 3, 4, 5, 0, 4, 7, 6, 5, 6, 3, 9, 10, 4, 5, 6, 6, 2, 7, 7, 9, 3, 6, 1, 7, 8, 1, 2
   Using a non-parametric test, test the hypothesis that the two sets of scores differ significantly at the 5 per cent level.

4. Fifteen informants are asked to rate two sentences on a scale from 1 (very impolite) to 5 (very polite), with the following results. Test the hypothesis that sentence 2 is rated as more polite than sentence 1, at the 5 per cent level.

Informant	Sentence 1	Sentence 2
1	1	3
2	2	2
3	1	4
4	2	3
5	3	1
6	2	4
7	1	1
8	2	3
9	3	5
10	1	3
11	2	3
12	1	4
13	2	1
14	2	4
15	1	3

Table: The Normal Distribution

The table gives the proportion of the total area under the curve which lies beyond any given z value (that is, the shaded area in the diagram). It is therefore appropriate for a one-tailed (directional) test. For a two-tailed (non-directional) test, the proportions must be doubled.



The figures down the left-hand side give values of z to the first decimal place, and those across the top give the second decimal place.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002

[Source: Butler, C.S. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell Ltd., p. 171]

Table: The F Distribution

The table gives the critical values of F for different numbers of degrees of freedom (df) in the numerator and in the denominator of the expression for F. For each entry, two values are given. The upper value is the critical value for the $p \leqslant 0.05$ level in a one-tailed/directional test, and for the $p \leqslant 0.10$ level in a two-tailed/non-directional test. The lower value is the critical value for the $p \leqslant 0.01$ level in a one-tailed/directional test and for the $p \leqslant 0.02$ level in a two-tailed/non-directional test.

Df in denominator	Df in numerator															
	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	250	252	254
	4 052	5 000	5 403	5 625	5 764	5 859	5 928	5 981	6 022	6 056	6 106	6 157	6 209	6 261	6 303	6 366
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.58	8.53
	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.5	26.4	26.1
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.70	5.63
	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.8	13.7	13.5
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.44	4.36
	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.38	9.24	9.02
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.75	3.67
	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.09	6.88
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.32	3.23
	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.86	5.65
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.02	2.93
	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.07	4.86
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.80	2.71
	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.52	4.31
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.64	2.54
	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.12	3.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.51	2.40
	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	3.94	3.81	3.60
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.40	2.30
	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70	3.57	3.36
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.31	2.21
	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.51	3.38	3.17
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.24	2.13
	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.35	3.22	3.00
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.18	2.07
	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	3.08	2.87
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19	2.12	2.01
	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.10	2.97	2.75
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15	2.08	1.96
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.00	2.87	2.65
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	2.04	1.92
	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.92	2.78	2.57

Table: The F Distribution (Continued)

Df in denominator	Df in numerator															
	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	∞
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07	2.00	1.88
	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.84	2.71	2.49
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.97	1.84
	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.64	2.42
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.92	1.84	1.71
	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.54	2.40	2.17
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.76	1.62
	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.25	2.01
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.96	1.88	1.79	1.70	1.56
	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.74	2.60	2.44	2.28	2.14	1.89
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74	1.66	1.51
	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.20	2.06	1.80
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	1.97	1.89	1.81	1.71	1.63	1.47
	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74	2.61	2.46	2.31	2.14	2.00	1.74
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.87	1.78	1.69	1.60	1.44
	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.42	2.27	2.10	1.95	1.68
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.56	1.39
	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.88	1.60
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.79	1.70	1.60	1.51	1.32
	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.42	2.27	2.12	1.94	1.79	1.49
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.77	1.68	1.57	1.48	1.28
	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.22	2.07	1.89	1.74	1.43
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55	1.46	1.25
	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.70	1.38
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.35	1.00
	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.70	1.52	1.00

[Source: Butler, C.S. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell Ltd., pp. 177-179]

Table: The t-Distribution

The table gives critical values of *t* for significance at various levels, in a two-tailed/non-directional or a one-tailed/directional test, for different numbers of degrees of freedom. These critical values are the values beyond which lies that proportion of the area under the curve which corresponds to the significance level.

	Significance level: two-tailed/non-directional				
	0.20	*0.10*	*0.05*	*0.02*	*0.01*
	Significance level: one-tailed/directional				
Degrees of freedom	*0.10*	*0.05*	*0.025*	*0.01*	*0.005*
1	3.078	6.314	12.71	31.82	63.66
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

[Source: Butler, C.S. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell Ltd., p. 172]

Table: The Pearson Product-Moment Correlation Coefficient

The table gives the critical values of the Pearson product–moment correlation coefficient, r, for different numbers of pairs of observations, N. For significance, the calculated value of r must be *greater than or equal to* the critical value.

	Significance level: two-tailed/non-directional			
	0.20	*0.10*	*0.05*	*0.01*
	Significance level: one-tailed/directional			
N	*0.10*	*0.05*	*0.025*	*0.005*
3	0.951	0.988	0.997	1.000
4	0.800	0.900	0.950	0.990
5	0.687	0.805	0.878	0.959
6	0.608	0.729	0.811	0.917
7	0.551	0.669	0.754	0.875
8	0.507	0.621	0.707	0.834
9	0.472	0.582	0.666	0.798
10	0.443	0.549	0.632	0.765
11	0.419	0.521	0.602	0.735
12	0.398	0.497	0.576	0.708
13	0.380	0.476	0.553	0.684
14	0.365	0.458	0.532	0.661
15	0.351	0.441	0.514	0.641
16	0.338	0.426	0.497	0.623
17	0.327	0.412	0.482	0.606
18	0.317	0.400	0.468	0.590
19	0.308	0.389	0.456	0.575
20	0.299	0.378	0.444	0.561
21	0.291	0.369	0.433	0.549
22	0.284	0.360	0.423	0.537
23	0.277	0.352	0.413	0.526
24	0.271	0.344	0.404	0.515
25	0.265	0.337	0.396	0.505
26	0.260	0.330	0.388	0.496
27	0.255	0.323	0.381	0.487
28	0.250	0.317	0.374	0.479
29	0.245	0.311	0.367	0.471
30	0.241	0.306	0.361	0.463
40	0.207	0.264	0.312	0.403
50	0.184	0.235	0.279	0.361
60	0.168	0.214	0.254	0.330
70	0.155	0.198	0.235	0.306
80	0.145	0.185	0.220	0.286
90	0.136	0.174	0.207	0.270
100	0.129	0.165	0.197	0.256
200	0.091	0.117	0.139	0.182

[Source: Butler, C.S. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell Ltd., p. 180]

Table: The Spearman Rank Correlation Coefficient

The table gives the critical values of the Spearman rank correlation coefficient, ρ, for different numbers of pairs of observations, N.

	Significance level: two-tailed/non-directional			
	0.20	*0.10*	*0.05*	*0.01*
	Significance level: one-tailed/directional			
N	*0.10*	*0.05*	*0.025*	*0.005*
5	0.800	0.900	1.000	–
6	0.657	0.829	0.886	1.000
7	0.571	0.714	0.786	0.929
8	0.524	0.643	0.738	0.881
9	0.483	0.600	0.700	0.833
10	0.455	0.564	0.648	0.794
11	0.427	0.536	0.618	0.755
12	0.406	0.503	0.587	0.727
13	0.385	0.484	0.560	0.703
14	0.367	0.464	0.538	0.679
15	0.354	0.446	0.521	0.654
16	0.341	0.429	0.503	0.635
17	0.328	0.414	0.488	0.618
18	0.317	0.401	0.472	0.600
19	0.309	0.391	0.460	0.584
20	0.299	0.380	0.447	0.570
21	0.292	0.370	0.436	0.556
22	0.284	0.361	0.425	0.544
23	0.278	0.353	0.416	0.532
24	0.271	0.344	0.407	0.521
25	0.265	0.337	0.398	0.511
26	0.259	0.331	0.390	0.501
27	0.255	0.324	0.383	0.492
28	0.250	0.318	0.375	0.483
29	0.245	0.312	0.368	0.475
30	0.240	0.306	0.362	0.467
35	0.222	0.283	0.335	0.433
40	0.207	0.264	0.313	0.405
45	0.194	0.248	0.294	0.382
50	0.184	0.235	0.279	0.363
55	0.175	0.224	0.266	0.346
60	0.168	0.214	0.255	0.331

[Source: Butler, C.S. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell Ltd., p. 181]

Table: The Mann-Whitney *U*-Test

The first table gives the critical values for significance at the $p \leqslant 0.05$ level in a two-tailed/non-directional test, and for the $p \leqslant 0.025$ level in a one-tailed/directional test. The second table gives the critical values for the $p \leqslant 0.01$ level in a two-tailed/non-directional test, and for the $p \leqslant 0.005$ level in a one-tailed/directional test. For significance, the calculated value of U must be *smaller than or equal to* the critical value. N_1 and N_2 are the number of observations in the smaller and larger group, respectively.

N_1 \ N_2	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$p \leqslant 0.05$ (two-tailed), $p \leqslant 0.025$ (one-tailed)																
5	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6		5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			8	10	12	14	16	18	20	22	24	26	28	30	32	34
8				13	15	17	19	22	24	26	29	31	34	36	38	41
9					17	20	23	26	28	31	34	37	39	42	45	48
10						23	26	29	33	36	39	42	45	48	52	55
11							30	33	37	40	44	47	51	55	58	62
12								37	41	45	49	53	57	61	65	69
13									45	50	54	59	63	67	72	76
14										55	59	64	69	74	78	83
15											64	70	75	80	85	90
16												75	81	86	92	98
17													87	93	99	105
18														99	106	112
19															113	119
20																127
$p \leqslant 0.01$ (two-tailed), $p \leqslant 0.005$ (one-tailed)																
5	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6		2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7			4	6	7	9	10	12	13	15	16	18	19	21	22	24
8				7	9	11	13	15	17	18	20	22	24	26	28	30
9					11	13	16	18	20	22	24	27	29	31	33	36
10						16	18	21	24	26	29	31	34	37	39	42
11							21	24	27	30	33	36	39	42	45	48
12								27	31	34	37	41	44	47	51	54
13									34	38	42	45	49	53	57	60
14										42	46	50	54	58	63	67
15											51	55	60	64	69	73
16												60	65	70	74	79
17													70	75	81	86
18														81	87	92
19															93	99
20																105

[Source: Butler, C.S. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell Ltd., p. 173]

Table: The Wilcoxon Signed-Ranks Test

The table gives critical values of W for different values of N (the number of non-tied pairs of scores). For significance, the calculated value must be *smaller than or equal to* the critical value.

	Significance level: two-tailed/non-directional	
	0.05	*0.01*
	Significance level: one-tailed/directional	
N	*0.025*	*0.005*
6	0	–
7	2	–
8	3	0
9	5	1
10	8	3
11	10	5
12	13	7
13	17	9
14	21	12
15	25	15
16	29	19
17	34	23
18	40	27
19	46	32
20	52	37
21	58	42
22	65	48
23	73	54
24	81	61
25	89	68

[Source: Butler, C.S. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell Ltd., p. 174]

Table: The Sign Test

The table gives critical values of *x* (the number of cases with the less frequent sign) for different values of *N* (the number of non-tied pairs of scores). For significance, the computed value of *x* must be *smaller than or equal to* the critical value.

	Significance level: two-tailed/non-directional		
	0.10	0.05	0.02
	Significance level: one-tailed/directional		
N	0.05	0.025	0.01
5	0	–	–
6	0	0	–
7	0	0	0
8	1	0	0
9	1	1	0
10	1	1	0
11	2	1	1
12	2	2	1
13	3	2	1
14	3	2	2
15	3	3	2
16	4	3	2
17	4	4	3
18	5	4	3
19	5	4	4
20	5	5	4
21	6	5	4
22	6	5	5
23	7	6	5
24	7	6	5
25	7	7	6

[Source: Butler, C.S. (1985). *Statistics in linguistics*. Oxford: Basil Blackwell Ltd., p. 175]

TEST SPECIFICATIONS

Course : Statistics in Education
Load : 3 credits
Lecturer : Ismail Petrus

Quiz	Objectives	Topics	Kinds of Test/Assignment	Time	Place
1	a. Understanding measures of central tendency, dispersion, and relative standing b. Displaying data by using a stem-and-leaf plot c. Understanding the normal distribution d. Calculating statistical values of data	• Central tendency • Dispersion • Relative standing • Stem-and-leaf plot • Normal distribution	○ Written test ○ Cognitive (application) ○ Skills of using Excel and/or SPSS	100 minutes	In the classroom
2	a. Understanding measures of central tendency, dispersion, and relative standing b. Understanding the normality, homogeneity, and linearity of the data c. Understanding correlation and linear regression d. Calculating statistical values of data and conducting statistical analyses	• Central tendency • Dispersion • Relative standing • Normal distribution, homogeneity, and linearity • Correlation • Linear regression	○ Written test ○ Cognitive (application) ○ Skills of using Excel and/or SPSS	100 minutes	In the classroom
3	a. Understanding correlation, linear regression, t-tests, one-way ANOVA, and non-parametric tests b. Conducting some statistical analyses by using appropriate statistical tests	• Normality and homogeneity • Correlation • Linear regression • t-tests • One-way ANOVA • Non-parametric test	○ Written test ○ Cognitive (application) ○ Skills of using Excel and/or SPSS	100 minutes	In the classroom

Quiz 1

Directions: *Answer all the questions.*

The following is the data set (Data A) of 40 values from a reading test.

30	***57***	***38***	***48***	***15***	***74***	***57***	***50***
65	***86***	***67***	***68***	***40***	***42***	***62***	***55***
56	***65***	***25***	***37***	***58***	***77***	***83***	***68***
91	***36***	***74***	***10***	***80***	***65***	***70***	***58***
43	***85***	***76***	***83***	***60***	***47***	***55***	***98***

A. Find the values of the following.

Mean	: ___________	The 33rd Percentile	: ___________
Median	: ___________	The 95th Percentile	: ___________
Mode	: ___________	Percentile Rank $_{(48)}$	: ___________
Midpoint	: ___________	Percentile Rank $_{(80)}$	: ___________
Range	: ___________	z-score $_{(67)}$	: ___________
Variance	: ___________	z-score $_{(91)}$	: ___________
Standard Deviation	: ___________	T-score $_{(50)}$	: ___________
Mean Deviation	: ___________	T-score $_{(77)}$	: ___________
Standard Error	: ___________	CEEB score $_{(36)}$	: ___________
Skewness	: ___________	IQR (Interquartile Range)	: ___________
Kurtosis	: ___________	CV (Coefficient of Variation)	: ___________

B. Construct a stem-and-leaf plot for the data set.

C. a. What range of values centered around the mean would represent 68 percent of the data points?

b. What is the minimum percent of values that would fall between 18 and 99?

D. Based on the IQR, determine whether there are any outliers in the data set.

E. Suppose there is another reading data set (Data B) with the standard deviation = 15.87 and the mean = 50.23, which data set is less variable?

Quiz 2

Directions: *Answer all the questions*.

The following are two sets of raw scores from a grammar test and a writing test.

Grammar scores:	45 27 34 26 15 47 06 18 28 40
	49 29 38 29 14 36 09 07 33 35
	24 16 36 15 37 50 42 05 35 44
Writing scores:	23 11 15 10 06 24 04 10 12 21
	25 14 20 13 05 17 05 04 14 15
	25 07 17 06 18 10 22 04 16 22

A. Find the values of the following.

Grammar Data		**Writing Data**	
Mean	: ________________	Mean	: ________________
Median	: ________________	Median	: ________________
Mode	: ________________	Mode	: ________________
Midpoint	: ________________	Midpoint	: ________________
Range	: ________________	Range	: ________________
Variance	: ________________	Variance	: ________________
Standard deviation:	________________	Standard deviation:	________________
Standard error :	________________	Standard error	: ________________
Skewness	: ________________	Skewness	: ________________
Kurtosis	: ________________	Kurtosis	: ________________
Percentile Rank $_{(38)}$:	________________	Percentile Rank $_{(7)}$:	________________
30^{th} Percentile:	________________	83^{rd} Percentile	: ________________
z-score $_{(15)}$:	________________	z-score $_{(21)}$	: ________________
T-score $_{(45)}$	: ________________	T-score $_{(10)}$	: ________________
IQR (Interquartile Range) :	________	IQR (Interquartile Range) :	__________
CV (Coefficient of Variation):	________	CV (Coefficient of Variation):	________

B. a. Determine whether each dataset follows the normal distribution.
   b. Determine whether the two datasets have the same distribution.
   c. Determine whether the two datasets have the linear relationship.

C. a. What is the correlation coefficient of the two data sets?
   b. What is the direction of the relationship? What does it mean?
   c. Determine whether the correlation is statistically significant or not.

d. Determine whether there are any outliers (on the basis of the scatter plot). If there are, which students have the scores that are considered as outliers?
e. Suppose a researcher eliminates the outlier(s) from the analysis, what would the coefficient be?

D. a. What is the regression equation that best fits the data?
b. Suppose a student has a grammar score of 43, what is the predicted score of writing?
c. What percent of the variance in writing is explained by the variable grammar?
d. Determine whether the relationship between the two variables is statistically significant or not.

Quiz 3

<u>Directions</u>: *Answer all the questions*.

Data1 : 63, 80, 85, 70, 76, 88, 73, 65, 87, 92, 85, 88, 60, 59, 80, 56, 79, 87, 76, 75, 69, 66, 80, 75, 90

Data2 : 70, 48, 34, 58, 84, 41, 38, 55, 63, 60, 74, 45, 48, 50, 58, 63, 56, 64, 28, 59, 79, 67, 56, 92, 55

Data3 : 68, 45, 61, 52, 68, 54, 30, 58, 52, 49, 80, 53, 51, 47, 45, 61, 50, 60, 40, 58, 85, 60, 58, 97, 75

Data4 : 45, 58, 60, 51, 53, 59, 56, 45, 56, 56, 60, 71, 64, 58, 65, 67, 78, 68, 56, 75

Data5 : 50, 60, 71, 56, 59, 62, 64, 62, 52, 69, 57, 66, 65, 60, 68, 68, 75, 83, 57, 85

A. A teacher of English wonders if her 12th graders score differently on an English test after a treatment than the 12th graders in the Province of South Sumatra. She gets a random sample of 25 students from her school. She knows that the provincial mean is 71.50 on the test. **Data1** shows the scores of her sample after the treatment. Determine also the effect size.

B. Twenty-five students learn both English vocabulary and listening. At the end of semester the students have the vocabulary and listening tests. **Data2** shows the results of vocabulary test and **Data3** the results of listening test.

   1. Does each data set have the normal distribution?
   2. Do the two data sets have the same distribution?
   3. Is there any significant correlation between the two data sets?
   4. Suppose that a student has a vocabulary score of 80, what is the predicted score of listening?
   5. What percent of the variance in listening is explained by the variable vocabulary?
   6. Is there any statistically significant influence of vocabulary on listening?

C. 1. A teacher teaches vocabulary to her students by using two different media. Suppose that **Data2** and **Data3** are the results of the pretest and posttest in the control group, and **Data4** and **Data5** are the results of the pretest and posttest in the experimental group, do the students in the experimental group significantly outperform those in the control group?

2. A teacher teaches 5 groups of students by using 5 different teaching strategies. Suppose that **Data1** shows the test results of the students who have been taught by using Strategy A, **Data2** Strategy B, **Data3** Strategy C, **Data4** Strategy D, and **Data5** Strategy E, which strategy is (the most) effective?

C. Two groups of students are taught English by two different methods, and their speaking fluency is then assessed on a scale from 1 to 10. The scores are as follows:

Method A: 6, 4, 8, 3, 4, 5, 5, 6, 4, 4, 7, 6, 6, 2, 3

Method B: 5, 5, 8, 9, 7, 5, 6, 7, 8, 3, 3, 7, 8, 6, 7

Do the two methods give significantly different results at the 5 per cent level?

Ismail Petrus is a lecturer of English Education at the Faculty of Teacher Training and Education, Sriwijaya University. He holds a Graduate Diploma of Arts in Interpreting/Translating from Deakin University, Melbourne, an M.A. in Linguistics from the University of Essex, UK, and a doctoral degree in English Education from Indonesia University of Education, Bandung, Indonesia. His research interests include linguistics, English language teaching and learning, and course design.

Penerbit dan Percetakan
NoerFikri
Jl. Mayor Mahidin No. 142
Tlp./Fax. 0711-366625
E-mail : noerfikri@gmail.com
Palembang - Indonesia

ISBN 978-602-447-094-4