MAR GREGORIOS COLLEGE OF ARTS & SCIENCE

Block No.8, College Road, Mogappair West, Chennai – 37

Affiliated to the University of Madras
Approved by the Government of Tamil Nadu
An ISO 9001:2015 Certified Institution

DEPARTMENT OF COMPUTER SCIENCE

SUBJECT NAME: ALLIED STATISTICS II

SUBJECT CODE: SBA0D

SEMESTER: III

PREPARED BY: PROF. D. SELVARAJ

SYLLABUS

Allied - Paper II - Statistical Methods and Their Applications II

Note:

The emphasis is solely upon the applicational understanding and practice of statistical methods, with specific reference to problems in physical, natural, and earth sciences.

UNIT - 1:

Bivariate frequency table and its uses - scatter diagram - Regression lines - linear prediction - Rank correlation coefficient - curve fitting by the method of least squares.

UNIT - 2:

Standard distributions - Binomial, Poisson and Normal distributions - Fitting of distributions.

UNIT - 3:

Concept of sampling distributions - standard error - Tests of significance based on t, Chi-square arid F - distributions with respect to mean, variance and correlation coefficient. Theory of attributes and tests of independence in contingency table.

UNIT - 4:

Sampling from finite population - Simple random sampling, Stratified and systematic random sampling procedures - Estimation mean and total and their standard errors. Concepts of sampling and non-sampling errors.

UNIT - 5:

Principle of scientific experiments - Randomization, replication, and local control Analysis of variance - One way and two way classification Analysis of CRD and RBD - Latin square designs. Concepts of factorial experiments (without confounding).

Books for Study and References:

Mode, E.B.: Elements of Statistics - Prentice Hall
Wilks, S.S.: Elementary Statistical Analysis -Oxford and IBH
Snedecor, G.W., & Cochran, W.G.: Statistical Methods, Oxford and IBH
Simpson and Kafka: Basic Statistics
Burr, I.W.: Applied Statistical Methods, Academic Press.
Croxton, FE. and Cowden, D.J.: Applied General Statistics, Prentice Hall
Ostleo, B.: Statistics in Research, Oxford & IBH.

ALLIED STATISTICS II

UNIT I

Bivariate Frequency Distributions

It is known that the frequency distribution of a single variable is called univariate distribution. When a data set consists of a large mass of observations, they may be summarized by using a two-way table. A two-way table is associated with two variables, say X and Y. For each variable, a number of classes can be defined keeping in view the same considerations as in the univariate case. When there are m classes for X and n classes for Y, there will be m × n cells in the two-way table. The classes of one variable may be arranged horizontally, and the classes of another variable may be arranged vertically in the two way table. By going through the pairs of values of X and Y, we can find the frequency for each cell. The whole set of cell frequencies will then define a bivariate frequency distribution. In other words, a bivariate frequency distribution is the frequency distribution of two variables.

Table 3.17 shows the frequency distribution of two variables, namely, age and marks obtained by 50 students in an intelligent test. Classes defined for marks are arranged horizontally (rows) and the classes defined for age are arranged vertically (columns). Each cell shows the frequency of the corresponding row and column values. For instance, there are 5 students whose age fall in the class 20 – 22 years and their marks lie in the group 30 – 40.

Table 3.17

Bivariate Frequency Distribution of Age and Marks

Marks	Age in Years				Total
	16 – 18	18 - 20	20 - 22	22 – 24	
10 – 20	2	1	1	-	4
20 – 30	3	2	3	1	9
30 – 40	3	3	5	6	17
40 – 50	2	2	3	4	11
50 – 60	-	1	2	2	5
60 – 70	-	1	2	1	4
Total	10	10	16	14	50

Stem and Leaf Plot (Stem and Leaf Diagram)

The *stem* and *leaf* plot is another method of organizing data and is a combination of sorting and graphing. It is an alternative to a tally chart or a grouped frequency distribution. It retains the original data without loss of information. This is also a type of bar chart, in which the numbers themselves would form the bars.

Stem and *leaf* plot is a type of data representation for numbers, usually like a table with two columns. Generally, *stem* is the label for left digit (leading digit) and leaf is the label for the right digit (trailing digit) of a number.

For example, the leaf corresponding to the value 63 is 3. The digit to the left of the leaf is called the stem. Here the stem of 63 is 6. Similarly for the number 265, the leaf is 5 and the stem is 26.

The elements of data 252, 255, 260, 262, 276, 276, 276, 283, 289, 298 are expressed in Stem and leaf plot as follows:

Actual data	Stem (Leading digits)	Leaf (Trailing digits)
252, 255	25	2 5
260, 262	26	0 2
276, 276, 276	27	6 6 6
283, 289	28	9
298	29	8

From the Stem and Leaf plot, we find easily the smallest number is 252 and the largest number is 298.

Also, in the class 270 – 280 we find 3 items are included and that group has the highest frequency.

The procedure for plotting a Stem and Leaf diagram is illustrated through an example given below:

Example 3.17

Construct a Stem and Leaf plot for the given data.

1.13, 0.72, 0.91, 1.44, 1.03, 0.88, 0.99, 0.73, 0.91, 0.98, 1.21, 0.79, 1.14, 1.19, 1.08, 0.94, 1.06, 1.11, 1.01

Solution:

Step 1: Arrange the data in the ascending order of magnitude:

0.72, 0.73, 0.79, 0.88, 0.91, 0.94, 0.98, 0.99, 1.01,

1.03, 1.06, 1.08, 1.11, 1.13, 1.14, 1.19, 1.21, 1.39, 1.44

Step 2: Separate the data according to the first digit as shown

0.72, 0.73, 0.79

0.88

0.91, 0.91, 0.94, 0.98, 0.99

1.01, 1.03, 1.06, 1.08

1.11, 1.13, 1.14, 1.19

1.21

1.39

1.44

Step 3: Now construct the stem and leaf plot for the above data.

Stem (Leading digits)

Stem (Leading digits)

0.7

0.8

0.9

1.0

1.1

1.2

1.3

1.4

Leaves (Trailing digits)

2 3 9

8

1 1 4 8 9

1 3 6 8

1 3 4 9

1

9

4

Using a Stem and Leaf plot, finding the Mean, Median, Mode and Range

-

We know how to create a stem and leaf plot. From this display, let us look at how we can use it to analyze data and draw conclusions. First, let us recall some statistical terms already we used in the earlier classes.

· The mean is the data value which gives the sum of all the data values, divided by the number of data values.

· The median is the data value in the middle when the data is ordered from the smallest to the largest.

· The mode is the data value that occurs most often. On a stem and leaf plot, the mode is the repeated leaf.

· The range is the difference between the highest and the least data value.

Example 3.18

Determine the mean, median, mode and the range on the stem and leaf plot given below:

Stem	Leaf
25	2 5
26	0 2
27	6 6 6
28	3 9
29	8

Solution:

From the display, combine the stem with each of its leaves. The values are in the order from the smallest to the largest on the plot. Therefore, keep them in order and list the data values as follows:

252, 255, 260, 262, 276, 276, 276, 283, 289, 298

To determine the mean, add all the data values and then divide the sum by the number of data values.

(252 + 255 + 260 + 262 + 276 + 276 + 276 + 283 + 289 + 298) ÷ 10

=2727 ÷ 10 = 272.7

Mean = 272.7.

The data is already arranged in ascending order. Therefore, identify the number in the middle position of the data. In this case, two data values share the middle position. To find the median, find the mean of these two middle data values.

The two middle numbers are 276 and 27 6.

The median is (276 +276) ÷ 2 = 276.

The mode is the data value that occurs more frequently. Looking at the stem and leaf plot, we can see the data value 276 appears thrice.

Therefore the mode is 276.

Recall that the range is the difference of the greatest and least values. On the stem and leaf plot, the greatest value is the last value and the smallest value is the first value.

The range is 298 – 252 = 46.

Scatter Diagram

The scatter diagram graphs pairs of numerical data, with one variable on each axis, to look for a relationship between them. If the variables are correlated, the points will fall along a line or curve. The better the correlation, the tighter the points will hug the line. This cause analysis tool is considered one of the seven basic quality tools.

- When to use a scatter diagram
- Scatter diagram procedure
- Scatter diagram example
- Scatter diagram considerations
- Scatter diagram resources

WHEN TO USE A SCATTER DIAGRAM

- When you have paired numerical data
- When your dependent variable may have multiple values for each value of your independent variable
- When trying to determine whether the two variables are related, such as:
  - When trying to identify potential root causes of problems
  - After brainstorming causes and effects using a fishbone diagram to determine objectively whether a particular cause and effect are related
  - When determining whether two effects that appear to be related both occur with the same cause
  - When testing for autocorrelation before constructing a control chart

SCATTER DIAGRAM PROCEDURE

1. Collect pairs of data where a relationship is suspected.

2. Draw a graph with the independent variable on the horizontal axis and the dependent variable on the vertical axis. For each pair of data, put a dot or a symbol where the x-axis value intersects the y-axis value. (If two dots fall together, put them side by side, touching, so that you can see both.)
3. Look at the pattern of points to see if a relationship is obvious. If the data clearly form a line or a curve, you may stop because variables are correlated. You may wish to use regression or correlation analysis now. Otherwise, complete steps 4 through 7.
4. Divide points on the graph into four quadrants. If there are X points on the graph:

   - Count X/2 points from top to bottom and draw a horizontal line.
   - Count X/2 points from left to right and draw a vertical line.
   - If number of points is odd, draw the line through the middle point.

5. Count the points in each quadrant. Do not count points on a line.
6. Add the diagonally opposite quadrants. Find the smaller sum and the total of points in all quadrants.

   A = points in upper left + points in lower right

   B = points in upper right + points in lower left

   Q = the smaller of A and B

   N = A + B
7. Look up the limit for N on the trend test table.

   - If Q is less than the limit, the two variables are related.
   - If Q is greater than or equal to the limit, the pattern could have occurred from random chance.

Table 5.18 Trend test table.

N	Limit	*N*	Limit
1–8	0	51–53	18
9–11	1	54–55	19
12–14	2	56–57	20
15–16	3	58–60	21
17–19	4	61–62	22
20–22	5	63–64	23
23–24	6	65–66	24
25–27	7	67–69	25
28–29	8	70–71	26
30–32	9	72–73	27
33–34	10	74–76	28
35–36	11	77–78	29
37–39	12	79–80	30
40–41	13	81–82	31
42–43	14	83–85	32
44–46	15	86–87	33
47–48	16	88–89	34
49–50	17	90	35

SCATTER DIAGRAM EXAMPLE

The ZZ-400 manufacturing team suspects a relationship between product purity (percent purity) and the amount of iron (measured in parts per million or ppm). Purity and iron are plotted against each other as a scatter diagram, as shown in the figure below.

There are 24 data points. Median lines are drawn so that 12 points fall on each side for both percent purity and ppm iron.

To test for a relationship, they calculate:

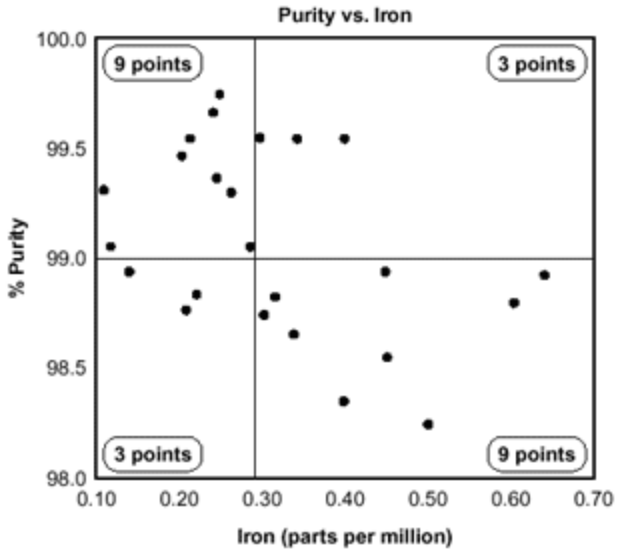
A = points in upper left + points in lower right = 9 + 9 = 18

B = points in upper right + points in lower left = 3 + 3 = 6

Q = the smaller of A and B = the smaller of 18 and 6 = 6

N = A + B = 18 + 6 = 24

Then they look up the limit for N on the trend test table. For N = 24, the limit is 6. Q is equal to the limit. Therefore, the pattern could have occurred from random chance, and no relationship is demonstrated.



Scatter Diagram Example

Additional Scatter Diagram Examples

Below are some examples of situations in which might you use a scatter diagram:

- Variable A is the temperature of a reaction after 15 minutes. Variable B measures the color of the product. You suspect higher temperature makes the product darker. Plot temperature and color on a scatter diagram.
- Variable A is the number of employees trained on new software, and variable B is the number of calls to the computer help line. You suspect that more training reduces the number of calls. Plot number of people trained versus number of calls.
- To test for autocorrelation of a measurement being monitored on a control chart, plot this pair of variables: Variable A is the measurement at a given time. Variable B is the same measurement, but at the previous time. If the scatter diagram shows correlation, do another diagram where variable B is the measurement two times previously. Keep increasing the separation between the two times until the scatter diagram shows no correlation.

SCATTER DIAGRAM CONSIDERATIONS

- Even if the scatter diagram shows a relationship, do not assume that one variable caused the other. Both may be influenced by a third variable.

- When the data are plotted, the more the diagram resembles a straight line, the stronger the relationship.
- If a line is not clear, statistics (*N* and Q) determine whether there is reasonable certainty that a relationship exists. If the statistics say that no relationship exists, the pattern could have occurred by random chance.
- If the scatter diagram shows no relationship between the variables, consider whether the data might be stratified.
- If the diagram shows no relationship, consider whether the independent (x-axis) variable has been varied widely. Sometimes a relationship is not apparent because the data do not cover a wide enough range

Regression line -Linear Prediction

The **Regression** Line is the line that best fits the data, such that the overall distance from the line to the points (variable values) plotted on a graph is the smallest. In other words, a line used to minimize the squared deviations of predictions is called as the **regression** line

The Linear Regression Equation

Linear regression is a way to model the relationship between two variables. You might also recognize the equation as the **slope formula**. The equation has the form Y= a + bX, where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

The first step in finding a linear regression equation is to determine if there is a relationship between the two variables. This is often a judgment call for the researcher. You'll also need a list of your data in x-y format (i.e. two columns of data—independent and dependent variables).

Warnings:

1. Just because two variables are related, it does not mean that one *causes* the other. For example, although there is a relationship between high GRE scores and better performance in grad school, it doesn't mean that high GRE scores **cause** good grad school performance.

2. If you attempt to try and find a linear regression equation for a set of data (especially through an automated program like Excel or a TI-83), you *will* find one, but it does not necessarily mean the equation is a good fit for your data. One technique is to make a scatter plot first, to see if the data roughly fits a line *before* you try to find a linear regression equation.

How to Find a Linear Regression Equation: Steps

Step 1: *Make a chart of your data, filling in the columns in the same way as you would fill in the chart if you were finding the Pearson's Correlation Coefficient*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

From the above table, Σx = 247, Σy = 486, Σxy = 20485, Σx2 = 11409, Σy2 = 40022. n is the sample size (6, in our case).

Step 2: Use the following equations to find a and b.

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

a = **65.1416**

b = **.385225**

Click here if you want easy, step-by-step instructions for solving this formula.

Find a:

- $((486 \times 11{,}409) – ((247 \times 20{,}485)) / 6\ (11{,}409) – 247^2)$
- 484979 / 7445
- =**65.14**

Find b:

- $(6(20{,}485) – (247 \times 486)) / (6\ (11409) – 247^2)$
- $(122{,}910 – 120{,}042) / 68{,}454 – 247^2$
- 2,868 / 7,445
- = **.385225**

Step 3: *Insert the values into the equation.*

y’ = a + bx

y’ = 65.14 + .385225x

Rank Correlation Coefficient

In statistics, Spearman's rank correlation coefficient or Spearman's ρ, named after Charles Spearman and often denoted by the Greek letter \rho or as r_{s}, is a nonparametric measure of rank correlation. It assesses how well the relationship between two variables can be described using a monotonic function.

What values can the Spearman correlation coefficient, r_s, take?

The Spearman correlation coefficient, r_s, can take values from +1 to -1. A r_s of +1 indicates a perfect association of ranks, a r_s of zero indicates no association between ranks and a r_s of -1 indicates a perfect negative association of ranks. The closer r_s is to zero, the weaker the association between the ranks.

An example of calculating Spearman's correlation

To calculate a Spearman rank-order correlation on data without any ties we will use the following data:

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

We then complete the following table:

English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d
56	66	9	4	5
75	70	3	2	1
45	40	10	10	0
71	60	4	7	3
62	65	6	5	1
64	56	5	9	4
58	59	8	8	0
80	77	1	1	0
76	67	2	3	1
61	63	7	6	1

Where d = difference between ranks and d^2 = difference squared.

We then calculate the following:

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

We then substitute this into the main equation with the other information as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2-1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

as $n = 10$. Hence, we have a ρ (or r_s) of 0.67. This indicates a strong positive relationship between the ranks individuals obtained in the maths and English exam. That is, the higher you ranked in maths, the higher you ranked in English also, and vice versa.

curve fitting by the method of least squares.

The given example explains how to find the equation of a straight line or a least square line by using the method of least square, which is very useful in statistics as well as in mathematics.

Example:

Fit a least square line for the following data. Also find the trend values and show that ∑(Y–Yˆ)=0∑(Y–Y^)=0.

TXX	**1**	**2**	**3**	**4**	**5**

YY	**2**	**5**	**3**	**8**	**7**

Solution:

XX	YY	XYXY	X2X2	Yˆ=1.1+1.3XY^=1.1+1.3X	Y–YˆY–Y^
1	**2**	**2**	**1**	**2.4**	**-0.4**
2	**5**	**10**	**4**	**3.7**	**+1.3**
3	**3**	**9**	**9**	**5.0**	**-2**
4	**8**	**32**	**16**	**6.3**	**1.7**
5	**7**	**35**	**25**	**7.6**	**-0.6**
∑X=15∑X=15	∑Y=25∑Y=25	∑XY=88∑XY=88	∑X2=55∑X2=55	**Trend Values**	∑(Y–Yˆ)=0∑(Y–Y^)=0

The equation of least square line Y=a+bXY=a+bX

Normal equation for '*a*' ∑Y=na+b∑X 25=5a+15b∑Y=na+b∑X 25=5a+15b —- (1)

Normal equation for '*b*' ∑XY=a∑X+b∑X2 88=15a+55b∑XY=a∑X+b∑X2 88=15a+55b —- (2)

Eliminate aa from equation (1) and (2), multiply equation (2) by **3** and subtract from equation (2). Thus we get the values of aa and bb.

Here a=1.1a=1.1 and b=1.3b=1.3, the equation of least square line becomes Y=1.1+1.3XY=1.1+1.3X

Curve Fitting

Curve fitting is the process of introducing mathematical relationships between dependent and independent variables in the form of an equation for a given set of data.

Method of Least Squares

The method of least squares helps us to find the values of unknowns aa and bb in such a way that the following two conditions are satisfied:

- The sum of the residual (deviations) of observed values of YY and corresponding expected (estimated) values of YY will be zero. $\sum(Y-Y\hat{}) = 0 \sum(Y-Y\hat{}) = 0$.
- The sum of the squares of the residual (deviations) of observed values of YY and corresponding expected values ($Y\hat{}Y\hat{}$) should be at least $\sum(Y-Y\hat{})2 \sum(Y-Y\hat{})2$.

Fitting of a Straight Line

A straight line can be fitted to the given data by the method of least squares. The equation of a straight line or least square line is $Y=a+bXY=a+bX$, where aa and bb are constants or unknowns. To compute the values of these constants we need as many equations as the number of constants in the equation. These equations are called normal equations. In a straight line there are two constants aa and bb so we require two normal equations.

Normal Equation for '*a*' $\sum Y=na+b\sum X \sum Y=na+b\sum X$

Normal Equation for '*b*' $\sum XY=a\sum X+b\sum X2 \sum XY=a\sum X+b\sum X2$

The direct formula of finding aa and bb is written as

$b=\sum XY-(\sum X)(\sum Y)n\sum X2-(\sum X)2n$,$a=\overline{Y}-bX$

UNIT II

Binomial Distribution :

A**binomial distribution** can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The**binomial** is a type of **distribution** that has two possible outcomes (the prefix "bi" means two, or twice).

A **binomial** experiment is one that possesses the following properties:

1. The experiment consists of *n* repeated trials;
2. Each trial results in an outcome that may be classified as a **success** or a **failure** (hence the name, **binomial**);

3. The probability of a success, denoted by p, remains constant from trial to trial and repeated trials are independent.

The number of successes X in n trials of a binomial experiment is called a **binomial random variable**.

The probability distribution of the random variable X is called a **binomial distribution**, and is given by the formula:

\displaystyle{P}{\left({X}\right)}={{C}_{{x}}^{{n}}}{p}^{x}{q}^{{{n}-{x}}}$P(X)=Cxnpxqn−x$

where

n = the number of trials

x = 0, 1, 2, ... n

p = the probability of success in a single trial

q = the probability of failure in a single trial

(i.e. $q = 1 - p$)

\displaystyle{{C}_{{x}}^{{n}}}Cxn is a combination

For example, many experiments share the common element that their outcomes can be classified

into one of two events, e.g. a coin can come up heads or tails; a child can be male or female; a

person can die or not die; a person can be employed or unemployed. These outcomes are often

labeled as “success” or “failure.” Note that there is no connotation of “goodness” here - for

example, when looking at births, the statistician might label the birth of a boy as a “success” and

the birth of a girl as a “failure,” but the parents wouldn’t necessarily see things that way. The

usual notation is

p = probability of success,

q = probability of failure = 1 - p.

Note that p + q = 1. In statistical terms, A Bernoulli trial is each repetition of an experiment

involving only 2 outcomes.

We are often interested in the result of independent, repeated bernoulli trials, i.e. the number of

successes in repeated trials.

1. independent - the result of one trial does not affect the result of another trial.

2. repeated - conditions are the same for each trial, i.e. p and q remain constant

across trials. Hayes refers to this as a stationary process. If p and q can change from trial to trial,

the process is nonstationary. The term identically distributed is also often used.

B. A binomial distribution gives us the probabilities associated with independent, repeated

Bernoulli trials. In a binomial distribution the probabilities of interest are those of receiving

a certain number of successes, r, in n independent trials each having only two possible

outcomes and the same probability, p, of success. So, for example, using a binomial

distribution, we can determine the probability of getting 4 heads in 10 coin tosses.

How does the binomial distribution do this? Basically, a two part process is involved. First, we

have to determine the probability of one possible way the event can occur, and then determine

the number of different ways the event can occur. That is,

P(Event) = (Number of ways event can occur) * P(One occurrence).

Suppose, for example, we want to find the probability of getting 4 heads in 10 tosses. In this

case, we'll call getting a heads a "success." Also, in this case, n = 10, the number of successes is

r = 4, and the number of failures (tails) is n – r = 10 – 4 = 6. One way this can occur is if the first

4 tosses are heads and the last 6 are tails, i.e.

Approximately 30 percent of obese patients develop diabetes. If a physician sees 10 patients who are obese:

a) What is the probability that half of them. will develop diabetes?

b) What is the probability that none will develop diabetes?

c) How many would you expect to develop diabetes?

Probabilities Knowing Sample Size:

When we have a sample of individuals, and we know the probability of success, we identify a probability distribution, since it is intrinsic that the probabilities for each individual are independent.

Answer and Explanation:

We have a binomial probability distribution with parameters:

Number of trials:n=10Probability of success:p=0.3Probability of failure:q=0.7Number of trials:n=10Probability of success:p=0.3Probability of failure:q=0.7
Binomial formula:

P(x)=n!(n−x)!·x!·px·qn−x or P(X=x)=(nx)·px·(1−p)n−xforx=0,1,2,…,nP(x)=n!(n−x)!·x!·px·qn−x or P(X=x)=(nx)·px·(1−p)n−xforx=0,1,2,…,n

a)

Required:

P(5)=P(x=5)=?P(5)=P(x=5)=?
Finding the corresponding probability we have:

P(5)=10!(10−5)!·5!·0.35·0.710−5P(5)=252·0.35·0.75P(5)=252·0.00243·0.16807P(5)=0.1029P(5)=10.29%P(5)=10!(10−5)!·5!·0.35·0.710−5P(5)=252·0.35·0.75P(5)=252·0.00243·0.16807P(5)=0.1029P(5)=10.29%

b)

P(0)=10!(10−0)!·0!·0.30·0.710−0P(0)=1·0.30·0.710P(0)=1·1·0.028247525P(0)=0.0282P(0)=2.82%P(0)=10!(10−0)!·0!·0.30·0.710−0P(0)=1·0.30·0.710P(0)=1·1·0.028247525P(0)=0.0282P(0)=2.82%

c)

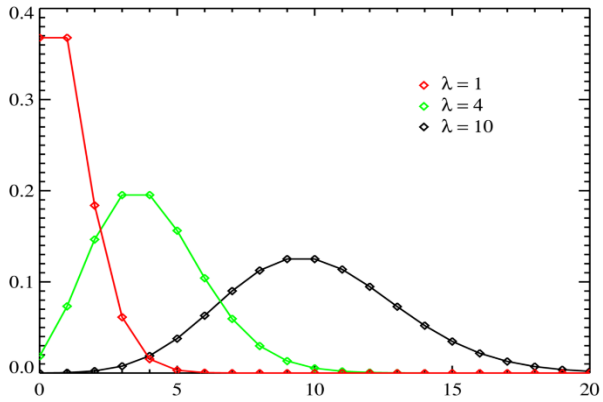
We go to find the expected value:

$E(x)=n\cdot pE(x)=10E(x)=3E(x)=n\cdot pE(x)=10E(x)=3$

We expect three out of ten patients to develop diabetes.

Poisson Distribution

A Poisson distribution is a tool that helps to predict the probability of certain events from happening when you know how often the event has occurred. It gives us the **probability of a given number of events happening in a fixed interval of time**.



Poisson distributions, valid only for integers on the horizontal axis. λ (also written as μ) is the expected number of event occurrences.

Practical Uses of the Poisson Distribution

A textbook store rents an average of 200 books every Saturday night. Using this data, you can **predict the probability that more books will sell** (perhaps 300 or 400) on the following Saturday nights. Another example is the number of diners in a certain restaurant every day. If the average number of diners for seven days is 500, you can predict the probability of a certain day having more customers.

Because of this application, Poisson distributions are used by businessmen to make **forecasts** about the number of customers or sales on certain days or seasons of the year. In business, overstocking will sometimes mean losses if the goods are not sold. Likewise, having too few stocks would still mean a lost business opportunity because you were not able to maximize your sales due to a shortage of stock. By using this tool,

businessmen are able to estimate the time when demand is unusually higher, so they can purchase more stock. Hotels and restaurants could prepare for an influx of customers, they could hire extra temporary workers in advance, purchase more supplies, or make contingency plans just in case they cannot accommodate their guests coming to the area. With the Poisson distribution, companies can adjust supply to demand in order to keep their business earning good profit. In addition, waste of resources is prevented.

Calculating the Poisson Distribution

> *The Poisson Distribution pmf is:* $P(x; \mu) = (e^{-\mu} * \mu^{x}) / x!$

Where:

- The symbol "!" is a factorial.
- μ (the expected number of occurrences) is sometimes written as λ. Sometimes called the **event rate** or rate parameter.

Example question

The average number of major storms in your city is 2 per year. What is the probability that exactly 3 storms will hit your city next year?

Step 1: Figure out the components you need to put into the equation.

- μ = 2 (average number of storms per year, historically)
- x = 3 (the number of storms we think might hit next year)
- e = 2.71828 (e is Euler's number, a constant)

Step 2: Plug the values from Step 1 into the Poisson distribution formula:

- $P(x; \mu) = (e^{-\mu}) (\mu^{x}) / x!$
- $= (2.71828^{-2}) (2^{3}) / 3!$
- = (0.13534) (8) / 6
- = 0.180

The probability of 3 storms happening next year is 0.180, or 18%

Poisson distribution vs. Binomial

The above example was over-simplified to show you how to work through a problem. However, it can be challenging to figure out if you should use a binomial distributionor a

Poisson distribution. If you aren't given a specific guideline from your instructor, use the following general guideline.

- If your question has an **average probability** of an event happening per unit (i.e. per unit of time, cycle, event) **and** you want to find probability of a certain number of events happening in a period of time (or number of events), then use the Poisson Distribution.
- If you are given an **exact probability** and you want to find the probability of the event happening a certain number out times out of x (i.e. 10 times out of 100, or 99 times out of 1000), use the Binomial Distribution formula.

Mean and Variance of Poisson Distribution

If μ is the average number of successes occurring in a given time interval or region in the Poisson distribution, then the mean and the variance of the Poisson distribution are both equal to μ.

$$E(X) = \mu$$

and

$$V(X) = \sigma^2 = \mu$$

Note: In a Poisson distribution, only **one** parameter, μ is needed to determine the probability of an event

Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

Solution: This is a Poisson experiment in which we know the following:

- μ = 5; since 5 lions are seen per safari, on average.
- x = 0, 1, 2, or 3; since we want to find the likelihood that tourists will see fewer than 4 lions; that is, we want the probability that they will see 0, 1, 2, or 3 lions.
- e = 2.71828; since *e* is a constant equal to approximately 2.71828.

To solve this problem, we need to find the probability that tourists will see 0, 1, 2, or 3 lions. Thus, we need to calculate the sum of four probabilities: P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5). To compute this sum, we use the Poisson formula:

$P(x \leq 3, 5) = P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)$

$P(x \leq 3, 5) = [(e^{-5})(5^0) / 0!] + [(e^{-5})(5^1) / 1!] + [(e^{-5})(5^2) / 2!] + [(e^{-5})(5^3) / 3!]$

$P(x \leq 3, 5) = [(0.006738)(1) / 1] + [(0.006738)(5) / 1] + [(0.006738)(25) / 2] + [(0.006738)(125) / 6]$

$P(x \leq 3, 5) = [0.0067] + [0.03369] + [0.084224] + [0.140375]$

$P(x \leq 3, 5) = 0.2650$

Thus, the probability of seeing at no more than 3 lions is 0.2650

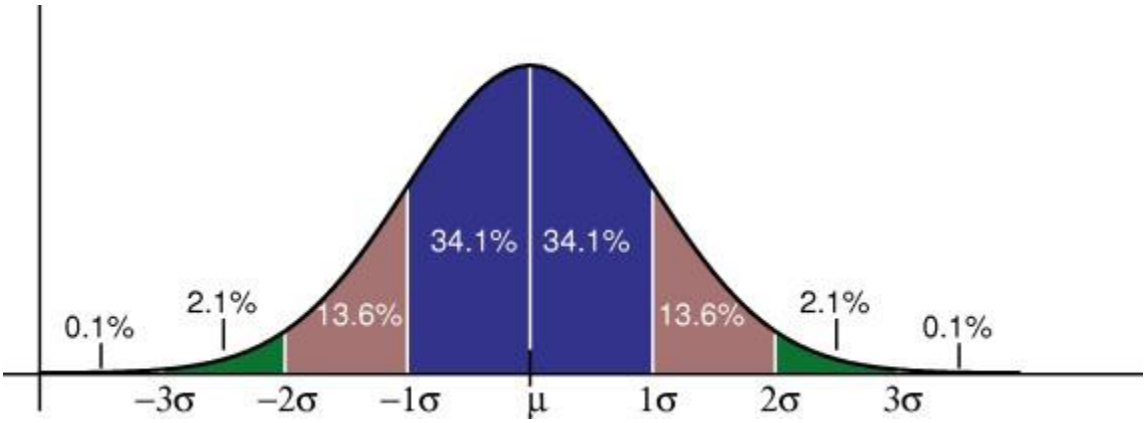
Normal Distribution

A normal distribution, sometimes called the bell curve, is a distribution that occurs naturally in many situations. For example, the bell curve is seen in tests like the SAT and GRE. The bulk of students will score the average (C), while smaller numbers of students will score a B or D. An even smaller percentage of students score an F or an A. This creates a distribution that resembles a bell (hence the nickname). The bell curve is symmetrical. Half of the data will fall to the left of the mean; half will fall to the right.
Many groups follow this type of pattern. That's why it's widely used in business, statistics and in government bodies like the FDA:

- Heights of people.
- Measurement errors.
- Blood pressure.
- Points on a test.
- IQ scores.
- Salaries.

The empirical rule tells you what percentage of your data falls within a certain number of standard deviations from the mean:

- 68% of the data falls within one standard deviation of the mean.
- 95% of the data falls within two standard deviations of the mean.
- 99.7% of the data falls within three standard deviations of the mean.



The standard deviation controls the spread of the distribution. A smaller standard deviation indicates that the data is tightly clustered around the mean; the normal distribution will be taller. A larger standard deviation indicates that the data is spread out around the mean; the normal distribution will be flatter and wider.

Fitting of Binomial Distribution :

When a Binomial distribution is to be fitted to an observed data the following procedure is adopted:-

(i) Find mean $\overline{x} = \frac{\sum fx}{\sum f} = np$

(ii) Find $p = \frac{\overline{x}}{n}$

(iii) Find $q = 1 - p$

(iv) Write the probability mass function : $P(x) = nC_x p^x q^{n-x}$ $x = 0, 1, 2, \ldots n$

(v) Put $x = 0$; find $P(0) = nC_0 p^0 q^{n-0} = q^n$

(vi) Find the expected frequency of $X = 0$ i.e., $F(0) = N \times P(0)$, where $N = \sum f_i$

(vii) The other expected frequencies are obtained by using the recurrence formula

$F(x+1) = \frac{n-x}{x+1} \times \frac{p}{q} \times F(x)$

Example 10.34

A set of three similar coins are tossed 100 times with the following results

Number of heads	0	1	2	3
Frequency	36	40	22	2

Fit a binomial distribution and estimate the expected frequencies.

Solution :

x	f	fx
0	36	0
1	40	40
2	22	44
3	2	44
Total	100	90

(i) Mean $\bar{x} = \frac{\sum fx}{\sum f} = \frac{90}{100} = 0.9$

(ii) $p = \frac{\bar{x}}{n} = \frac{0.9}{3} = 0.3$

(iii) $q = 1 - p = 1 - 0.3 = 0.7$

(iv) $P(x) = nC_x p^x q^{n-x} = 3C_x \; 0.3_x \; 0.7^{3-x}$

(v) $P(0) = 3C_0 \; 0.3^0 \; (0.7)^{3-0} = 0.7^3 = 0.343$

(vi) $F(0) = N \times P(0) = 100 \times 0.343 = 34.3$

(vii) $F(x+1) = \frac{n-x}{x+1} \times \frac{p}{q} \times F(x)$

$\therefore \quad F(1) = F(0+1) = \frac{3-0}{0+1} \times \frac{0.3}{0.7} \times 34.3 = 44.247$

$F(2) = F(1+1) = \frac{3-1}{1+1} \times \frac{0.3}{0.7} \times 44.247 = 19.03$

$F(3) = F(2+1) = \frac{3-2}{2+1} \times \frac{0.3}{0.7} \times 19.03 = 2.727$

Solution :

(i) The fitted binomial distribution is

$P(X = x) = 3C_x\, 0.3^x\, 0.7^{(3-x)}, \quad x = 0,1,2,3$

(ii) The expected frequencies are :

x	0	1	2	3	Total
Observed frequencies (O_i)	36	40	22	2	100
Expected Frequencies (E_i)	34	44	19	3	100

Fitting of Poisson Distribution

When a Poisson distribution is to be fitted to an observed data the following procedure is adopted:

(i) Find the mean: $\overline{x} = \frac{\Sigma fx}{\Sigma f}$

(ii) Poisson parameter = $\lambda = \overline{x}$

(iii) Probability mass function is: $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $x = 0, 1, 2....$

(iv) Put X = 0 and find P(0) = $\frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}$

(v) Expected frequency for $x = 0$ is $F(0) = N \times P(0)$, $\Sigma f_i = N$

(vi) Other expected frequencies can be found using

$F(x + 1) = \frac{\lambda}{x+1} \times F(x)$ for $x = 0, 1, 2....$

Example 10.35

The following mistakes per page were observed in a book

Number of Mistakes (per page)	0	1	2	3	4
Number of pages	211	90	19	5	0

Fit a Poisson distribution and estimate the expected frequencies.

Solution:

x	f	fx
0	211	0
1	90	90
2	19	38
3	5	15
4	0	0
Total	325	143

(i) Mean $\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{143}{325} = 0.44$

(ii) $\lambda = \bar{x} = \quad = 0.44$

(iii) $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0.44}(0.44)^x}{x!}$

(iv) $P(0) \quad = \frac{e^{0.44} \times 0.44^0}{0!} = e^{-0.44} \quad = 0.6440$ (from the Poisson table)

(v) $F(0) \;=\; N \times P(0) \;= 325 \times 0.6440 = 209.43$

(vi) $F(x+1) = \frac{\lambda}{x+1} F(x)$

$F(1) \;=\; F(0+1) \;= \frac{0.44}{0+1} \times 209.43 = 92.15$

$F(2) \;=\; F(1+1) \;= \frac{0.44}{1+1} \times 92.15 = 20.27$

$F(3) \;=\; F(2+1) \;= \frac{0.44}{2+1} \times 2.972$

$F(4) \;=\; F(3+1) \;= \frac{0.44}{3+1} \times 2.97 = 0.33$

Result:

(1). Fitted Poisson distribution is $P(X = x) = \frac{e^{-0.44}0.44^x}{x!}$, $x = 0, 1, 2....$

(2). Expected frequencies are given below :

x	0	1	2	3	4	Total
Observed frequencies (O_i)	211	90	19	5	0	325
Expected Frequencies (E_i)	210	92	20	3	0	325

Fitting of Normal Distribution

In fitting a Normal distribution to the observed data, given in class intervals, we follow the following procedure:-

(i) Calculate μ and σ of the distribution

(ii) Find x_i the lower class boundary

(iii) Find $z_i = \frac{x_i - \mu}{\sigma}$

(iv) Find Z_i (Zi) $= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{z^2}{2}} dz$

(v) Find $\Delta \phi$ $(Z_i) = \phi$ $(Z_{i+1}) - \phi(Z_i)$

(vi) Find expected frequency $E_i = = N \Delta \phi(Z_i)$

Example 10.36

Find expected frequencies for the following data, if its calculated mean and standard deviation are 79.945 and 5.545.

Class	60-65	65-70	70-75	75-80	80-85	85-90	90-95	95-10
Frequency	3	21	150	335	326	135	26	4

Solution:

Given μ= 79.945, σ = 5.545, and N = 1000

Hence the equation of Normal curve fitted to the data is

$$f(x) = \frac{1}{5.545\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-79.945}{5.545}\right)^2}$$

Theoretical Normal frequencies can be obtained as follows:

UNIT III

t-test

The t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. A t-test is the most commonly applied when the test

Class	Lower Class boundry (X_i)	$z_i = \frac{X_i - \mu}{\sigma}$	$\phi(Z_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{z^2}{2}} dz$	$= \phi(Z_{i+1}) - \phi(Z_i)$	Expected frequencies $N \Delta \phi(Z_i)$
Below 60	- ∞	- ∞	0	0.001	*1*
60 - 65	60	-3.59693	0.001	0.0026	2.6 ≈ 3
65 - 70	65	-2.69522	0.0036	0.0331	33.1 ≈ 33
70 - 75	70	-1.79351	0.0367	0.15	150
75 - 80	75	-0.89179	0.1867	0.3173	317.3 ≈ 317
80 - 85	80	0.009919	0.504	0.3146	314.6 ≈ 315
85 - 90	85	0.911632	0.8186	0.1463	146.3 ≈ 146
90 - 95	90	1.813345	0.9649	0.0318	31.8 ≈ 32
95 – 100	95	2.715059	0.9967	0.0032	3.2 ≈ 3
100 and Over	100	3.616772	0.9999		

statistic would follow a normal distribution if the value of a scaling term in the test statistic were known.

What is a Paired T Test (Paired Samples T Test / Dependent Samples T Test)?

A paired t test (also called a **correlated pairs t-test**, a **paired samples t test** or **dependent samples t test**) is where you run a t test on dependent samples. Dependent samples are essentially connected — they are tests on the same person or thing. For example:

- Knee MRI costs at two different hospitals,
- Two tests on the same person before and after training,
- Two blood pressure measurements on the same person using different equipment.

When to Choose a Paired T Test / Paired Samples T Test / Dependent Samples T Test

Choose the paired t-test if you have two measurements on the same item, person or thing. You should also choose this test if you have two items that are being measured with a unique condition. For example, you might be measuring car safety performance in Vehicle Research and Testing and subject the cars to a series of crash tests. Although the manufacturers are different, you might be subjecting them to the same conditions.

With a "regular" two sample t test, you're comparing the means for two different samples. For example, you might test two different groups of customer service associates on a business-related test or testing students from two universities on their English skills. If you take a random sample each group separately and they have different conditions, your samples are independent and you should run an independent samples t test (also called between-samples and unpaired-samples).

The null hypothesis for the independent samples t-test is $\mu_1 = \mu_2$. In other words, it assumes the means are equal. With the paired t test, the null hypothesis is that the *pairwise difference* between the two tests is equal (H_0: $\mu_d = 0$). The difference between the two tests is very subtle; which one you choose is based on your data collection method.

Paired Samples T Test By hand

Example question: Calculate a paired t test by hand for the following data:

Subject #	Score 1	Score 2
1	3	20
2	3	13
3	3	13
4	12	20
5	15	29
6	16	32
7	17	23
8	19	20
9	23	25
10	24	15
11	32	30

Step 1: Subtract each Y score from each X score.

Subject #	Score 1	Score 2	X-Y
1	3	20	-17
2	3	13	-10
3	3	13	-10
4	12	20	-8
5	15	29	-14
6	16	32	-16
7	17	23	-6
8	19	20	-1
9	23	25	-2
10	24	15	9
11	32	30	2

Step 2: Add up all of the values from Step 1. Set this number aside for a moment.

Subject #	Score 1	Score 2	X-Y
1	3	20	-17
2	3	13	-10
3	3	13	-10
4	12	20	-8
5	15	29	-14
6	16	32	-16
7	17	23	-6
8	19	20	-1
9	23	25	-2
10	24	15	9
11	32	30	2
		SUM:	**-73**

Step 3: Square the differences from Step 1.

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	

Step 4: Add up all of the squared differences from Step 3.

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	**1131**

Step 5: Use the following formula to calculate the t-score:

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

- ΣD: Sum of the differences (Sum of X-Y from Step 2)
- ΣD^2: Sum of the squared differences (from Step 4)
- $(\Sigma D)^2$: Sum of the differences (from Step 2), squared.

If you're unfamiliar with Σ you may want to read about summation notation first.

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{(-73)^2}{11}\right)}{(11\text{-}1)(11)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{5329}{11}\right)}{110}}}$$

$$t = -\ 2.74$$

Step 6: Subtract 1 from the sample size to get the degrees of freedom. We have 11 items, so 11-1 = 10.

Step 7: Find the p-value in the t-table, using the degrees of freedom in Step 6. If you don't have a specified alpha level, use 0.05 (5%). For this example problem, with df = 10, the t-value is 2.228.

Step 8: Compare your t-table value from Step 7 (2.228) to your calculated t-value (-2.74). The calculated t-value is greater than the table value at an alpha level of .05. The p-value is less than the alpha level: p <.05. We can reject the null hypothesis that there is no difference between means.

Extra Problems - t tests

1. A manufacturer of running shoes knows that the average lifetime for a particular model of shoes is 15 months. Someone in the research and development division of the shoe company claims to have developed a longer lasting product. This new product was worn by 30 individuals and lasted on

average for 17 months. The variability of the original shoe is estimated based on the standard deviation of the new group which is 5.5 months. Is the designer's claim of a better shoe supported by the trial results? Please base your decision on a two tailed testing using a level of significance of $p < .05$.

2. Average heart rate for Americans is 72 beats/minute. A group of 25 individuals participated in an aerobics fitness program to lower their heart rate. After six months the group was evaluated to identify is the program had significantly slowed their heart. The mean heart rate for the group was 69 beats/minute with a standard deviation of 6.5. Was the aerobics program effective in lowering heart rate?

3. A research team wants to investigate the usefulness of relaxation training for reducing levels of anxiety in individuals experiencing stress. They identify 30 people at random from a group of 100 who have "high stress" jobs. The 30 people are divided into two groups. One group acts as the control group - they receive no training. The second group of 15 receive the relaxation training. The subjects in each group are then given an anxiety inventory. The summarized results appear below where higher scores indicate greater anxiety.

Control Relaxation

$X = 30$ $X = 26$

$S = 6.63$ $S = 6.20$

$n = 15$ $n = 15$

4. A colleague of the investigators is problem 3 repeats the experiment but matches the samples on the dimensions of sex and job type. the raw data appear below. Evaluate her experiment using the criteria of $p < .05$. Assume it is a two tailed test.

Pairs: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Contr: 38 40 35 36 35 32 31 30 28 26 24 21 18 34 22

Relax: 35 32 30 34 30 32 28 27 22 22 18 17 17 25 21

Answer Key – t test extra problems

1. This is a problem that requires a t test for single samples. You have been given the fact that the population mean is 15. The sample mean is 17 with a standard dev. of 5.5.

Step One: Generally you should start by computing the st. dev. but that has been done so you can move on to computing the st. error. In this case st. error = 1.00

Step Two: Compute t using the sample mean, pop mean and st. error. In this case, t = 2.00

Step Three: Evaluate. The crit. value of t for a two tailed test is 2.045, for a one tailed test is 1.699. So, if you wrote a two tailed test you must accept the null. If you wrote a one tailed test you must reject the null and accept the alternative.

2. The pop. mean is given as 72 beats per minutes. The sample of 25 has an average of 69 with a standard dev. of 6.5.

Step One: Again you need to solve for st. error. St. error = 1.30

Step Two: Solve for t test for single samples t = -2.31

Step Three: Evaluate. The critical value is 2.064. The computed value exceeds this value so there is a significant effect of the ind. var. of fitness.

3. You have no info about the population and there are two samples so this calls for a t test for independent samples. You can use the shortened version of the t formula if you want since the size of each sample is the same. Remember that the value that is given is the st. dev. of each sample and that the formula (either one) requires the variance so that the first thing to do is to square each of the st.dev.s.

The numerator of the t formula is 4. The denominator is 2.34. The overall t value is 1.71. The critical value at df = 28 is 2.048 so that this outcome is not

statistically significant.

4. This last problem is a t test for matched samples. In order to solve this you must first find D - the difference between the control subject and the relaxation subject in each matched pair. The sum of D= 60, the mean value of D = 4 and the sum of D squared is 332. The st. dev. of D = 2.56 and the st. error equals .66. This makes

t = 6.06. When you evaluate this the critical value at df = 15-1 or 14 is 2.145. The computed t value exceeds this and so it is a significant outcome. The relaxation

group is significantly different than the control group.

F Test

An **F-test** is any statistical **test** in which the **test statistic** has an **F**-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled..

An "F Test" is a catch-all term for **any test that uses the F-distribution**. In most cases, when people talk about the F-Test, what they are actually talking about is The *F-Test to Compare Two Variances.* However, the f-statistic is used in a variety of tests.

If you're running an F Test, you should use Excel, SPSS, Minitab or some other kind of technology to run the test. Why? Calculating the F test by hand, including variances, is tedious and time-consuming. Therefore you'll probably make some errors along the way.

If you're running an F Test using technology (for example, an F Test two sample for variances in Excel), the only steps you really need to do are Step 1 and 4 (dealing with the null hypothesis). Technology will calculate Steps 2 and 3 for you.

1. State the null hypothesis and the alternate hypothesis.
2. Calculate the F value. The F Value is calculated using the formula $F = (SSE_1 – SSE_2 / m) / SSE_2 / n-k$, where SSE = residual sum of squares, m = number of restrictions and k = number of independent variables.
3. Find the F Statistic (the critical value for this test). The F statistic formula is: **F Statistic = variance of the group means / mean of the within group variances.** You can find the F Statistic in the F-Table.
4. Support or Reject the Null Hypothesis.

Back to Top

F Test to Compare Two Variances

A **Statistical F Test** uses an F Statistic to compare two variances, s_1 and s_2, by dividing them. The result is always a positive number (because variances are always positive). The equation for comparing two variances with the f-test is:

$F = s^2_1 / s^2_2$

If the variances are equal, the ratio of the variances will equal 1. For example, if you had two data sets with a sample 1 (variance of 10) and a sample 2 (variance of 10), the ratio would be 10/10 = 1.

You **always** test that the population variances are equal when running an F Test. In other words, you always assume that the variances are equal to 1. Therefore, your null hypothesis will always be that *the variances are equal*.

Assumptions

Several **assumptions** are made for the test. Your population **must be approximately normally distributed** (i.e. fit the shape of a bell curve) in order to use the test. Plus, the samples must be independent events. In addition, you'll want to bear in mind a few important points:

- The larger variance should always go in the numerator (the top number) to force the test into a right-tailed test. Right-tailed tests are easier to calculate.
- For two-tailed tests, divide alpha by 2 before finding the right critical value.
- If you are given standard deviations, they must be squared to get the variances.
- If your degrees of freedom aren't listed in the F Table, use the larger critical value. This helps to avoid the possibility of Type I errors.

Back to Top

F Test to compare two variances by hand: Steps

Need help with a specific question? Check out our tutoring page!

Warning: F tests can get really tedious to calculate by hand, especially if you have to calculate the variances. You're much better off using technology (like Excel — see below).

These are the general steps to follow. Scroll down for a specific example (watch the video underneath the steps).

Step 1: *If you are given standard deviations, go to Step 2. If you are given variances to compare, go to Step 3.*

Step 2: *Square both standard deviations to get the variances.* For example, if $\sigma_1 = 9.6$ and $\sigma_2 = 10.9$, then the variances (s_1 and s_2) would be $9.6^2 =$ **92.16** and $10.9^2 =$ **118.81**.

Step 3: *Take the largest variance, and divide it by the smallest variance to get the f-value.* For example, if your two variances were $s_1 = 2.5$ and $s_2 = 9.4$, divide 9.4 / 2.5 = **3.76**. Why? Placing the largest variance on top will force the F-test into a right tailed test, which is much easier to calculate than a left-tailed test.

Step 4: Find your degrees of freedom. Degrees of freedom is your sample size minus 1. As you have two samples (variance 1 and variance 2), you'll have two degrees of freedom: one for the numerator and one for the denominator.

Step 5: *Look at the f-value you calculated in Step 3 in the f-table.* Note that there are several tables, so you'll need to locate the right table for your alpha level. Unsure how to read an f-table? Read What is an f-table?.

Step 6: Compare your calculated value (Step 3) with the table f-value in Step 5. If the f-table value is smaller than the calculated value, you can reject the null hypothesis.

That's it

The difference between running a one or two tailed F test is that the alpha level needs to be halved for two tailed F tests. For example, instead of working at $\alpha = 0.05$, you use $\alpha = 0.025$; Instead of working at $\alpha = 0.01$, you use $\alpha = 0.005$.

With a two tailed F test, you just want to know if the variances are not equal to each other. In notation:

$H_a = \sigma^2 1 \neq \sigma^2 2$

Example problem: Conduct a two tailed F Test on the following samples: Sample 1: Variance = 109.63, sample size = 41. Sample 2: Variance = 65.99, sample size = 21.

Step 1: Write your hypothesis statements:

H_o: No difference in variances.

H_a: Difference in variances.

Step 2: Calculate your F critical value. Put the highest variance as the numerator and the lowest variance as the denominator:

F Statistic = variance 1/ variance 2 = 109.63 / 65.99 = 1.66

Step 3: Calculate the degrees of freedom: The degrees of freedom in the table will be the sample size -1, so: Sample 1 has 40 df (the numerator). Sample 2 has 20 df (the denominator).

Step 4: Choose an alpha level. No alpha was stated in the question, so use 0.05 (the standard "go to" in statistics). This needs to be halved for the two-tailed test, so use 0.025.

Step 5: Find the critical F Value using the F Table. There are several tables, so make sure you look in the alpha = .025 table. Critical F (40,20) at alpha (0.025) = 2.287.

/	df_1=1	2	24	30	40	60	120	∞
df_2=1	647.7890	799.5000	997.2492	1001.414	1005.598	1009.800	1014.020	1018.258
2	38.5063	39.0000	39.4562	39.465	39.473	39.481	39.490	39.498
3	17.4434	16.0441	14.1241	14.081	14.037	13.992	13.947	13.902
4	12.2179	10.6491	8.5109	8.461	8.411	8.360	8.309	8.257
5	10.0070	8.4336	6.3780	6.227	6.175	6.123	6.069	6.015
16	6.1151	4.6867	2.6252	2.568	2.509	2.447	2.383	2.316
17	6.0420	4.6189	2.5598	2.502	2.442	2.380	2.315	2.247
18	5.9781	4.5597	2.5027	2.445	2.384	2.321	2.256	2.187
19	5.9216	4.5075	2.4523	2.394	2.333	2.270	2.203	2.133
20	5.8715	4.4613	2.4076	2.349	2.287	2.223	2.156	2.085

Step 6: Compare your calculated value (Step 2) to your table value (Step 5). If your calculated value is higher than the table value, you can reject the null hypothesis: F calculated value: 1.66

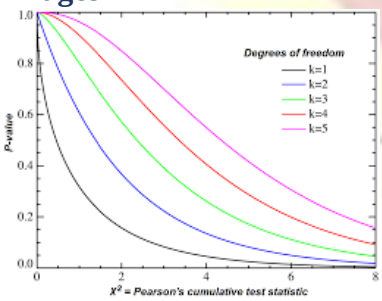
F value from table: 2.287.

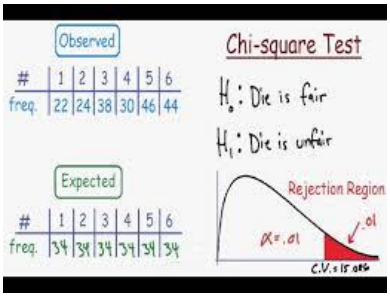
1.66 < 2 .287.

So we cannot reject the null hypothesis.

Chi-squared teMain results

Images





Marital Status by Education | n = 300

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher	Total
Never married	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
Total	39	90	84	54	33	300

What is a Chi Square Test?

There are two types of chi-square tests. Both use the chi-square statistic and distribution for different purposes:

- A chi-square goodness of fit test determines if sample data matches a population. For more details on this type, see: Goodness of Fit Test.
- A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.
    - A very small chi square test statistic means that your observed data fits your expected data extremely well. In other words, there is a relationship.
    - A very large chi square test statistic means that the data does not fit very well. In other words, there isn't a relationship.

The formula for the chi-square statistic used in the chi square test is:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The chi-square formula.

The subscript "c" is the degrees of freedom. "O" is your observed value and E is your expected value. It's very rare that you'll want to actually *use* this formula to find a critical chi-square value by hand. The summation symbol means that you'll have to perform a calculation for every single data item in your data set. As you can probably imagine, the calculations can get very, very, lengthy and tedious. Instead, you'll probably want to use technology:

Uses

The chi-squared distribution has many uses in statistics, including:

- Confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation.
- Independence of two criteria of classification of qualitative variables.
- Relationships between categorical variables (contingency tables).
- Sample variance study when the underlying distribution is normal.
- Tests of deviations of differences between expected and observed frequencies (one-way tables).
- The chi-square test (a goodness of fit test).

Example question: 256 visual artists were surveyed to find out their zodiac sign. The results were: Aries (29), Taurus (24), Gemini (22), Cancer (19), Leo (21), Virgo (18), Libra (19), Scorpio (20), Sagittarius (23), Capricorn (18), Aquarius (20), Pisces (23). Test the hypothesis that zodiac signs are evenly distributed across visual artists.

Step 1: **Make a table** with columns for "Categories," "Observed," "Expected," "Residual (Obs-Exp)", "$(\text{Obs-Exp})^2$" and "Component $(\text{Obs-Exp})^2$ / Exp." Don't worry what these mean right now; We'll cover that in the following steps.

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp

Step 2: **Fill in your categories**. Categories should be given to you in the question. There are 12 zodiac signs, so:

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries					
Taurus					
Gemini					
Cancer					
Leo					
Virgo					
Libra					
Scorpio					
Sagittarius					
Capricorn					
Aquarius					
Pisces					

Step 3: **Write your counts**. Counts are the number of each items in each category in column 2. You're given the counts in the question:

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29				
Taurus	24				
Gemini	22				
Cancer	19				
Leo	21				
Virgo	18				
Libra	19				
Scorpio	20				
Sagittarius	23				
Capricorn	18				
Aquarius	20				
Pisces	23				

Step 4: **Calculate your expected value** for column 3. In this question, we would expect the 12 zodiac signs to be evenly distributed for all 256 people, so 256/12=21.333. Write this in column

3.

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333			
Taurus	24	21.333			
Gemini	22	21.333			
Cancer	19	21.333			
Leo	21	21.333			
Virgo	18	21.333			
Libra	19	21.333			
Scorpio	20	21.333			
Sagittarius	23	21.333			
Capricorn	18	21.333			
Aquarius	20	21.333			
Pisces	23	21.333			

Step 5: **Subtract the expected value (Step 4) from the Observed value (Step 3)** and place the result in the "Residual" column. For example, the first row is Aries: 29-21.333=7.667.

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667		
Taurus	24	21.333	2.667		
Gemini	22	21.333	0.667		
Cancer	19	21.333	-2.333		
Leo	21	21.333	-0.333		
Virgo	18	21.333	-3.333		
Libra	19	21.333	-2.333		
Scorpio	20	21.333	-1.333		
Sagittarius	23	21.333	1.667		
Capricorn	18	21.333	-3.333		
Aquarius	20	21.333	-1.333		
Pisces	23	21.333	1.667		

Step 6: **Square your results from Step 5** and place the amounts in the $(Obs-Exp)^2$ column.

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	
Taurus	24	21.333	2.667	7.112889	
Gemini	22	21.333	0.667	0.44889	
Cancer	19	21.333	-2.333	5.442889	
Leo	21	21.333	-0.333	0.110889	
Virgo	18	21.333	-3.333	11.108889	
Libra	19	21.333	-2.333	5.442889	
Scorpio	20	21.333	-1.333	1.776889	
Sagittarius	23	21.333	1.667	2.778889	
Capricorn	18	21.333	-3.333	11.108889	
Aquarius	20	21.333	-1.333	1.776889	
Pisces	23	21.333	1.667	2.778889	

Step 7: **Divide the amounts in Step 6 by the expected value (Step 4)** and place those results in the final column.

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	2.755490976
Taurus	24	21.333	2.667	7.112889	0.333421882
Gemini	22	21.333	0.667	0.44889	0.021042048
Cancer	19	21.333	-2.333	5.442889	0.255139408
Leo	21	21.333	-0.333	0.110889	0.005198003
Virgo	18	21.333	-3.333	11.108889	0.520737308
Libra	19	21.333	-2.333	5.442889	0.255139408
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricorn	18	21.333	-3.333	11.108889	0.520737308
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457

Step 8: Add up (sum) all the values in the last column.

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	2.755490976
Taurus	24	21.333	2.667	7.112889	0.333421882
Gemini	22	21.333	0.667	0.44889	0.021042048
Cancer	19	21.333	-2.333	5.442889	0.255139408
Leo	21	21.333	-0.333	0.110889	0.005198003
Virgo	18	21.333	-3.333	11.108889	0.520737308
Libra	19	21.333	-2.333	5.442889	0.255139408
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricorn	18	21.333	-3.333	11.108889	0.520737308
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457
					5.094017203

This is the chi-square statistic: 5.094.

Theory of attributes and tests of independence in contingency table

In statistics, a **contingency table** (also known as a cross tabulation or crosstab) is a type of **table** in a matrix format that displays the (multivariate) frequency distribution of the variables. They are heavily used in survey research, business intelligence, engineering, and scientific research.

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

A contingency table, sometimes called a two-way frequency table, is a tabular mechanism with at least two rows and two columns used in statistics to present categorical data in terms of frequency counts. More precisely, an $r \times c$ contingency table shows the observed frequency of

two variables, the observed frequencies of which are arranged into r rows and c columns. The intersection of a row and a column of a contingency table is called a cell.

gender	cup	cone	sundae	sandwich	other
male	592	300	204	24	80
female	410	335	180	20	55

For example, the above contingency table has two rows and five columns (not counting header rows/columns) and shows the results of a random sample of 2200 adults classified by two variables, namely gender and favorite way to eat ice cream (Larson and Farber 2014). One benefit of having data presented in a contingency table is that it allows one to more easily perform basic probability calculations, a feat made easier still by augmenting a summary row and column to the table.

gender	cup	cone	sundae	sandwich	other	total
male	592	300	204	24	80	1200
female	410	335	180	20	55	1000
total	1002	635	384	44	135	2200

The above table is an extended version of the first table obtained by adding a summary row and column. These summaries allow easier computation of several different probability-related quantities. For example, there's a $1002/2200 \approx 45.54\ \%$ probability that the person sampled prefers their ice cream in a cup, while the probability that a random participant is female is $1000/2200 \approx 45.45\ \%$. What's more, computing conditional probabilities is made easier using contingency tables, e.g., the probability that a person prefers ice cream sandwiches given that the person is male is $24/1200 = 2\ \%$, while the conditional probability that a person is male given that ice cream sandwiches are preferred is $24/44 \approx 54.54\ \%$.

Other common statistical analyses can be performed on data given in contingency table form. For example, one useful value to know is the so-called expected frequency $E_{c,r}$ of the cell at the intersection of column c and row r, the formula for which is given by

$$E_{c,r} = \frac{(\text{sum of row } r)\cdot(\text{sum of column } c)}{\text{sample size}}. \quad (1)$$

Computing $E_{1,1}$ says that the value one would expect at cell $(1, 1)$--i.e., the expected number of men who prefer to eat ice cream from a cup--is approximately

$$E_{1,1} = \frac{1200\cdot 1002}{2200} \approx 546.54, \quad (2)$$

whereby one may deduce that there are somehow "more than expected" of that particular demographic included in the given sample. Note, too, that knowing $E_{1,1}$ automatically gives, e.g., $E_{2,1}$, without repeated application of ():

$$E_{2,1} = (\text{total people who prefer cups}) - E_{1,1} \approx 1002 - 546.54 = 455.46. \quad (3)$$

One of the major benefits of computing expected frequencies is the ability to test whether the two variables being examined--in this case, gender and favorite way to eat ice cream--are actually independent as they've been assumed throughout. This is done by computing, for each cell (c, r), the expected frequency $E = E_{c,r}$, comparing it to the observed frequency $O = O_{c,r}$, and then performing a chi-squared test.

Another common test associated to contingency tables is so-called homogeneity of proportions test which is a form of chi-squared test used to determine whether several proportions are equal when samples are taken from different populations (Larson and Farber 2014). Worth noting is that both of the above-mentioned instances of chi-squared testing requires a randomly-selected sampling of observed frequencies, each of whose expected frequency is at least 5. These tests play important roles throughout various branches of statistics.

UNIT IV

Simple random sampling:

In statistics, a simple random sample is a subset of individuals chosen from a larger set in which each individual is chosen randomly and entirely by chance

When to use simple random sampling

Simple random sampling is used to make statistical inferences about a population. It helps ensure high internal validity: randomization is the best method to reduce the impact of potential confounding variables.

In addition, with a large enough sample size, a simple random sample has high external validity: it represents the characteristics of the larger population.

However, simple random sampling can be challenging to implement in practice. To use this method, there are some prerequisites:

- You have a complete list of every member of the population.
- You can contact or access each member of the population if they are selected.
- You have the time and resources to collect data from the necessary sample size.

Simple random sampling works best if you have a lot of time and resources to conduct your study, or if you are studying a limited population that can easily be sampled.

In some cases, it might be more appropriate to use a different type of probability sampling:

- **Systematic sampling** involves choosing your sample based on a regular interval, rather than a fully random selection. It can also be used when you don't have a complete list of the population.
- **Stratified sampling** is appropriate when you want to ensure that specific characteristics are proportionally represented in the sample. You split your population into strata (for example, divided by gender or race), and then randomly select from each of these subgroups.

- **Cluster sampling** is appropriate when you are unable to sample from the entire population. You divide the sample into clusters that approximately reflect the whole population, and then choose your sample from a random selection of these clusters.

How to perform simple random sampling

There are 4 key steps to select a simple random sample.

Step 1: Define the population

Start by deciding on the population that you want to study.

It's important to ensure that you have access to every individual member of the population, so that you can collect data from all those who are selected for the sample.

Example: PopulationIn the American Community Survey, the population is all 128 million households who live in the United States (including households made up of citizens and non-citizens alike).

Step 2: Decide on the sample size

Next, you need to decide how large your sample size will be. Although larger samples provide more statistical certainty, they also cost more and require far more work.

There are several potential ways to decide upon the size of your sample, but one of the simplest involves using a formula with your desired confidence interval and confidence level, estimated size of the population you are working with, and the standard deviation of whatever you want to measure in your population.

The most common confidence interval and levels used are 0.05 and 0.95, respectively. Since you may not know the standard deviation of the population you are studying, you should choose a number high enough to account for a variety of possibilities (such as 0.5).

You can then use a sample size calculator to estimate the necessary sample size.

Example: Sample sizeThe ACS follows 3.5 million households each year. This is a small fraction of the overall population of 128 million households, but it is a large enough sample size to gather detailed data on all geographical regions and demographic groups in the United States, including those usually underrepresented in surveys.

Step 3: Randomly select your sample

This can be done in one of two ways: the lottery or random number method.

In the **lottery method**, you choose the sample at random by "drawing from a hat" or by using a computer program that will simulate the same action.

In the **random number method**, you assign every individual a number. By using a random number generator or random number tables, you then randomly pick a subset of the population. You can also use the random number function (RAND) in Microsoft Excel to generate random numbers.

Example: Random selectionThe Census Bureau randomly selects addresses of 295,000 households monthly (or 3.5 million per year). Each address has approximately a 1-in-480 chance of being selected.

Step 4: Collect data from your sample

Finally, you should collect data from your sample.

To ensure the validity of your findings, you need to make sure every individual selected actually participates in your study. If some drop out or do not participate for reasons associated with the question that you're studying, this could bias your findings.

For example, if young participants are systematically less likely to participate in your study, your findings might not be valid due to the underrepresentation of this group.

Example: Data collectionThe Census Bureau first sends a letter to ask the respondents to fill the survey out online. If occupants of an address do not respond, the Bureau calls the home telephone number. If all else fails, a representative visits the address in person.

Through this variety of methods, the officials collecting data for the ACS manage to receive responses from 95% of those randomly selected, a high response rate that supports the validity of their results.

Stratified sampling

In statistics, stratified sampling is a method of sampling from a population which can be partitioned into subpopulations. In statistical surveys, when subpopulations within an overall population vary, it could be advantageous to sample each subpopulation independently.

When to use stratified sampling

To use stratified sampling, you need to be able to divide your population into mutually exclusive and exhaustive subgroups. That means every member of the population can be clearly classified into exactly one subgroup.

Stratified sampling is the best choice among the probability sampling methods when you believe that subgroups will have different mean values for the variable(s) you're studying. It has several potential advantages:

- Ensuring the diversity of your sample

A stratified sample includes subjects from every subgroup, ensuring that it reflects the diversity of your population. It is theoretically possible (albeit unlikely) that this would not happen when using other sampling methods such as simple random sampling.

- Ensuring similar variance

If you want the data collected from each subgroup to have a similar level of variance, you need a similar sample size for each subgroup.

With other methods of sampling, you might end up with a low sample size for certain subgroups because they're less common in the overall population.

- Lowering the overall variance in the population

Although your overall population can be quite heterogeneous, it may be more homogenous within certain subgroups.

For example, if you are studying how a new schooling program affects the test scores of children, both their original scores and any change in scores will most likely be highly correlated with family income. The scores are likely to be grouped by family income category.

In this case, stratified sampling allows for more precise measures of the variables you wish to study, with lower variance within each subgroup and therefore for the population as a whole.

- Allowing for a variety of data collection methods

Sometimes you may need to use different methods to collect data from different subgroups.

For example, in order to lower the cost and difficulty of your study, you may want to sample urban subjects by going door-to-door, but rural subjects using mail.

Research exampleYou are interested in how having a doctoral degree affects the wage gap between men and women among graduates of a certain university.

Because only a small proportion of this university's graduates have obtained a doctoral degree, using a simple random sample would likely give you a sample size too small to properly compare the differences between men and women with a doctoral degree versus those without one.

Therefore, you decide to use a stratified sample, relying on a list provided by the university of all its graduates within the last ten years.

Step 1: Define your population and subgroups

Like other methods of probability sampling, you should begin by clearly defining the population from which your sample will be taken.

Choosing characteristics for stratification

You must also choose the characteristic that you will use to divide your groups. This choice is very important: since each member of the population can only be placed in only one subgroup, the classification of each subject to each subgroup should be clear and obvious.

Stratifying by multiple characteristics

You can choose to stratify by multiple different characteristics at once, so long as you can clearly match every subject to exactly one subgroup. In this case, to get the total number of subgroups, you multiply the numbers of strata for each characteristic.

For instance, if you were stratifying by both race and gender, using four groups for the former and two for the latter, you would have 2 x 4 = 8 groups in total.

ExampleYour population is all graduates of the university within the last ten years. You will stratify by both gender and degree received.

Cluster sampling

In cluster sampling, researchers divide a population into smaller groups known as clusters. They then randomly select among these clusters to form a sample.

Cluster sampling is a method of probability sampling that is often used to study large populations, particularly those that are widely geographically dispersed. Researchers usually use pre-existing units such as schools or cities as their clusters.

How to cluster sample

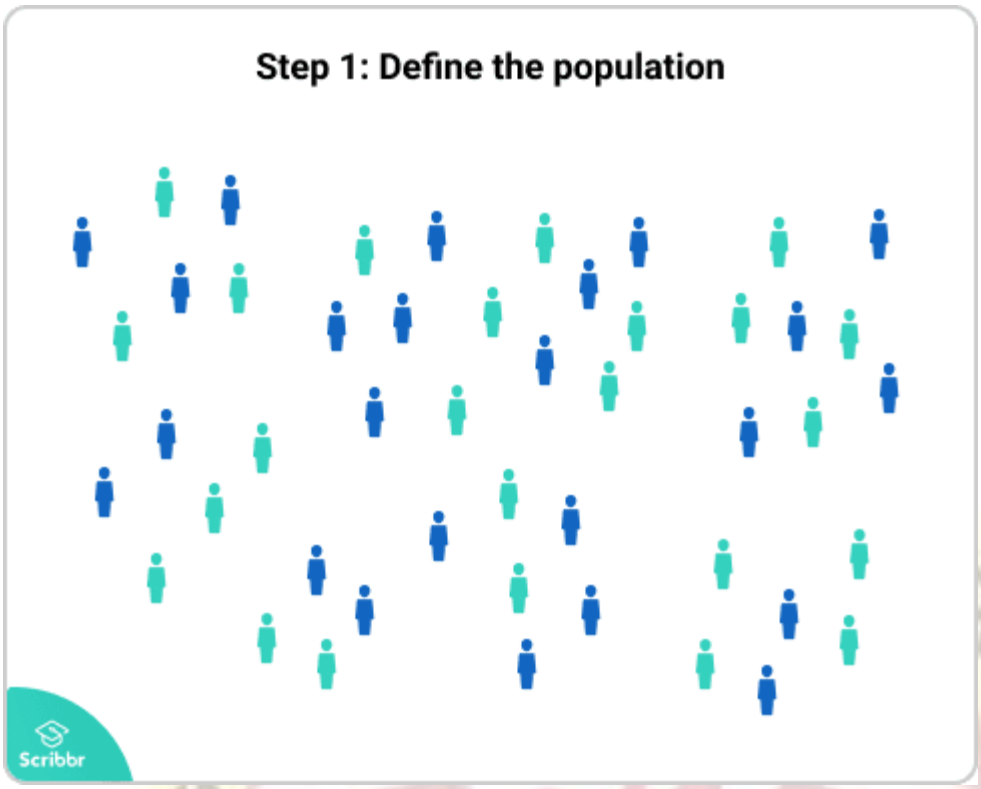
The simplest form of cluster sampling is **single-stage cluster sampling**. It involves 4 key steps.

Research exampleYou are interested in the average reading level of all the seventh-graders in your city.

It would be very difficult to obtain a list of all seventh-graders and collect data from a random sample spread across the city. However, you can easily obtain a list of all schools and collect data from a subset of these. You thus decide to use the cluster sampling method.

Step 1: Define your population

As with other forms of sampling, you must first begin by clearly defining the population you wish to study.



PopulationIn your reading program study, your population is all the seventh-graders in your city.

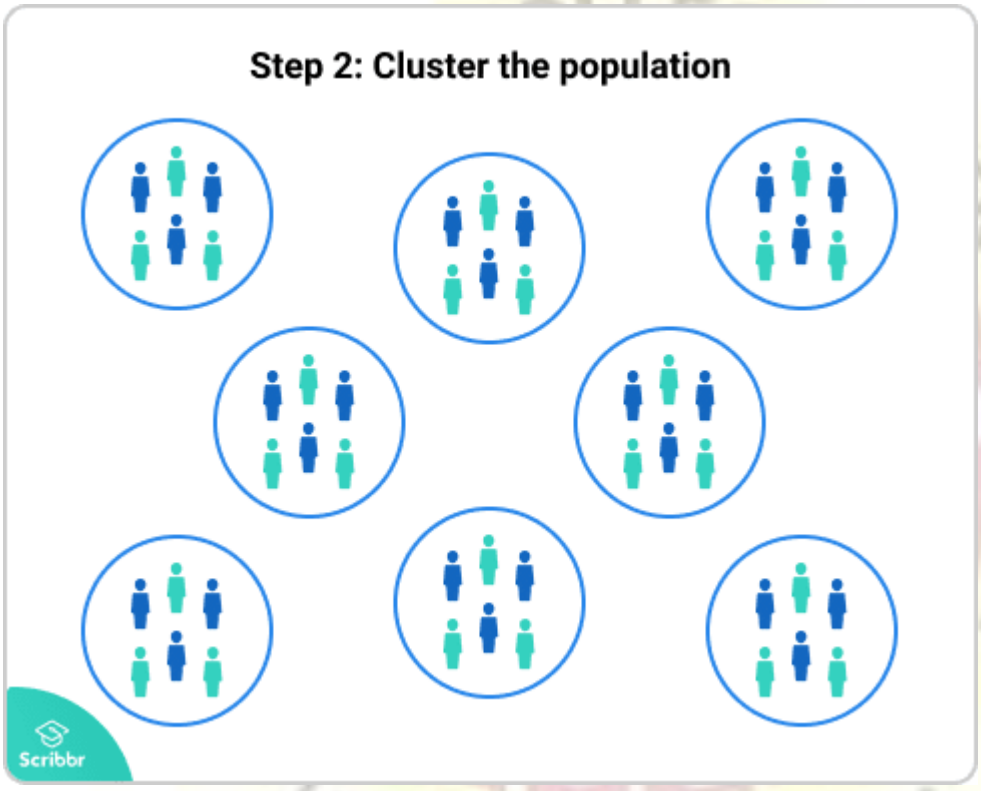
Step 2: Divide your sample into clusters

This is the most important part of the process. The quality of your clusters and how well they represent the larger population determines the validity of your results. Ideally, you would like for your clusters to meet the following criteria:

- Each cluster's population should be as diverse as possible. You want every potential characteristic of the entire population to be represented in each cluster.
- Each cluster should have a similar distribution of characteristics as the distribution of the population as a whole.
- Taken together, the clusters should cover the entire population.
- There not be any overlap between clusters (i.e. the same people or units do not appear in more than one cluster).

Ideally, each cluster should be a mini-representation of the entire population. However, in practice, clusters often do not perfectly represent the population's characteristics, which is why this method provides less statistical certainty than simple random sampling.

Because clusters are usually naturally occurring groups, such as schools, cities, or households, they are often more homogenous than the population as a whole. You should be aware of this when performing your study, as it might affect its validity.

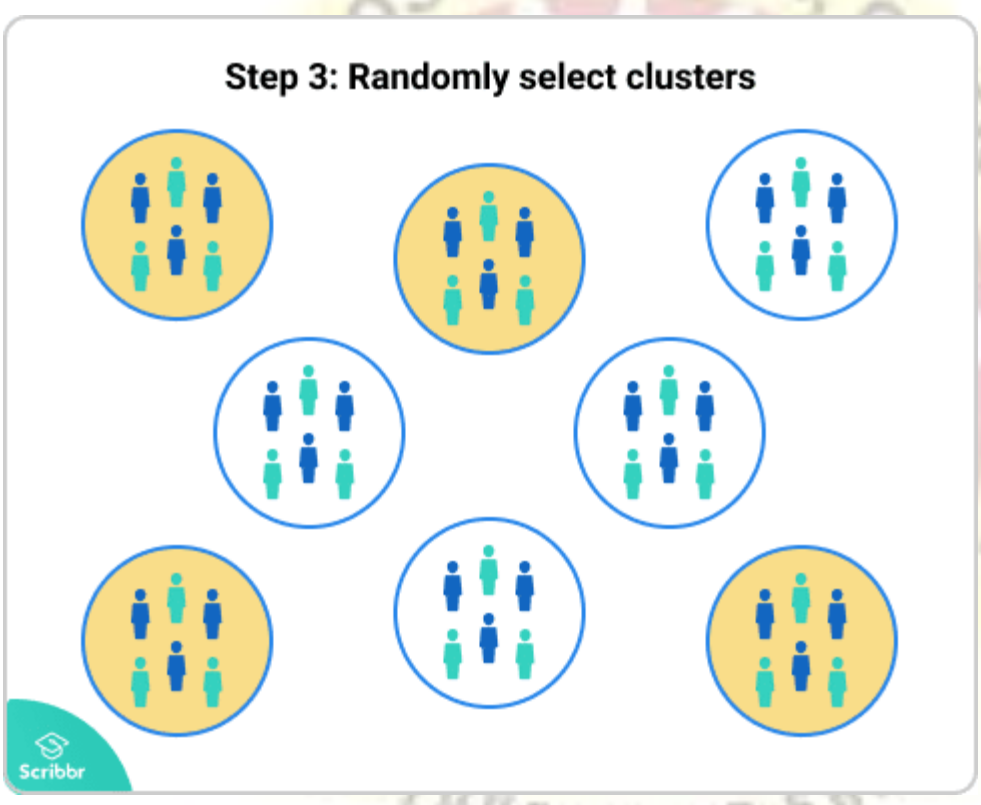


ClustersYou cluster the seventh-graders by the school they attend. To cover the whole population, you need to include every school in the city. There is no overlap because each student attends only one school.

Step 3: Randomly select clusters to use as your sample

If each cluster is itself a mini-representation of the larger population, randomly selecting and sampling from the clusters allows you to imitate simple random sampling, which in turn supports the validity of your results.

Conversely, if the clusters are not representative, then random sampling will allow you to gather data on a diverse array of clusters, which should still provide you with an overview of the population as a whole.



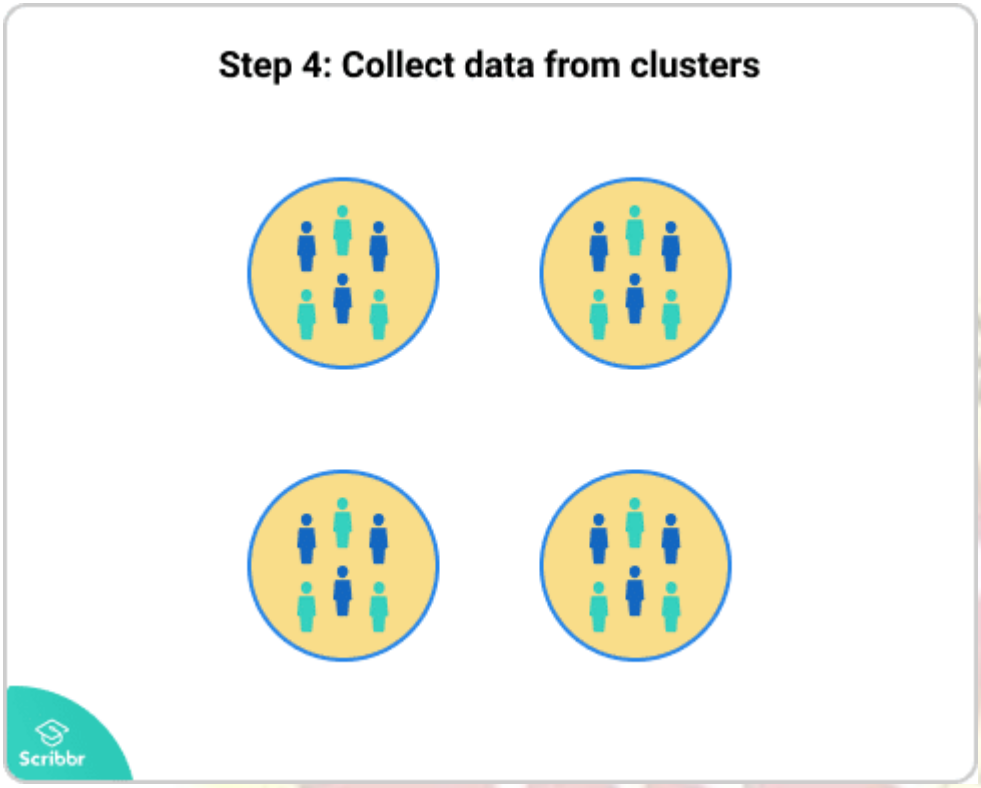
SampleYou assign a number to each school and use a random number generator to select a random sample.

You choose the number of clusters based on how large you want your sample size to be. This in turn is based on the estimated size of the entire seventh-grade population, your desired confidence interval and confidence level, and your best guess of the standard deviation (a measure of how spread apart the values in a population are) of the reading levels of the seventh-graders.

You then use a sample size calculator to estimate the required sample size.

Step 4: Collect data from the sample

You then conduct your study and collect data from every unit in the selected clusters.

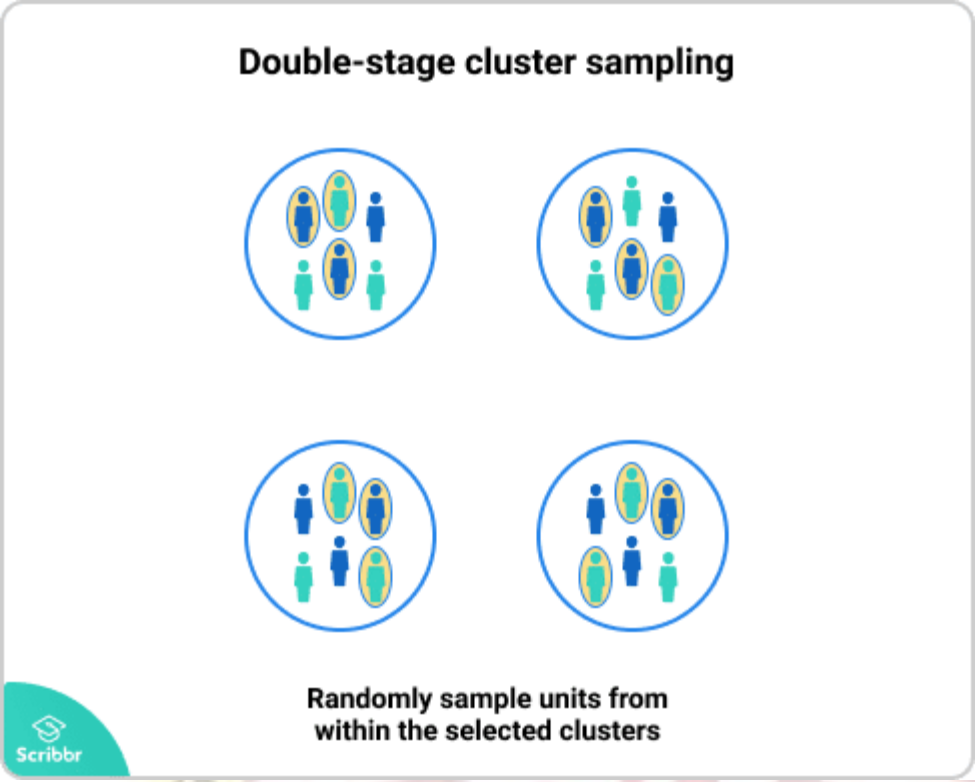


Data collectionYou test the reading levels of every seventh-grader in the schools that were randomly selected for your sample.

Multi-stage cluster sampling

In **multi-stage clustering**, rather than collect data from every single unit in the selected clusters, you randomly select individual units from within the cluster to use as your sample.

You can then collect data from each of these individual units – this is known as **double-stage sampling**.



You can also continue this procedure, taking progressively smaller and smaller random samples, which is usually called **multi-stage sampling**.

You should use this method when it is infeasible or too expensive to test the entire cluster.

Example: Multistage samplingInstead of collecting data from every seventh-grader in the selected schools, you narrow down your sample in two additional stages:

1. From each school, you randomly select a sample of seventh-grade classes.
2. From within those classes, you randomly select a sample of students.

The resulting sample is much smaller and therefore easier to collect data from.

Estimation mean and total and their standard errors.

Procedure:

Step 1: Calculate the mean (Total of all samples divided by the number of samples).

Step 2: Calculate each measurement's deviation from the mean (Mean minus the individual measurement).

Step 3: Square each deviation from mean. Squared negatives become positive.

Step 4: Sum the squared deviations (Add up the numbers from step 3).

Step 5: Divide that sum from step 4 by one less than the sample size (n-1, that is, the number of

measurements minus one)

Step 6: Take the square root of the number in step 5. That gives you the "standard deviation (S.D.)."

Step 7: Divide the standard deviation by the square root of the sample size (n). That gives you the

"standard error".

Step 8: Subtract the standard error from the mean and record that number. Then add the standard

error to the mean and record that number. You have plotted mean± 1 standard error (S. E.),

the distance from 1 standard error below the mean to 1 standard error above the mean

Example:

5 Divide by number of measurements-1. (m-i)2 / (n-1) = 272.70 / 4 = 68.175

6 Standard deviation = square root of (m-i)2/n-1 = 68.175 = 8.257

7 Standard error = Standard deviation/ n = 8.257/2.236 = 3.69

8 m ± 1SE = 162 ± 3.7 or 159cm to 166cm for the men (162.4 - 3.7 to 162.4 + 3.7).

Concepts of sampling and non-sampling errors.

Meaning **Sampling error** is a type of **error**, occurs due to the **sample** selected does **not** perfectly represents the population. An **error** occurs due to sources other than **sampling**, while conducting survey activities is known as **non sampling error**.

Non-sampling error is the error that arises in a

data collection process as a result of factors other

than taking a sample.

Non-sampling errors have the potential to cause

bias in polls, surveys or samples.

There are many different types of non-sampling

errors and the names used to describe them are

not consistent.

This may be due to poor sampling method,

measurement errors, and behavioural effect.

Comparison Chart

BASIS FOR COMPARISON	SAMPLING ERROR	NON-SAMPLING ERROR

BASIS FOR COMPARISON	SAMPLING ERROR	NON-SAMPLING ERROR
Meaning	Sampling error is a type of error, occurs due to the sample selected does not perfectly represents the population of interest.	An error occurs due to sources other than sampling, while conducting survey activities is known as non sampling error.
H	Deviation between sample mean and population mean	Deficiency and analysis of data
Type	Random	Random or Non-random
Occurs	Only when sample is selected.	Both in sample and census.
Sample size	Possibility of error reduced with the increase in sample size.	It has nothing to do with the sample size.

UNIT V

Randomization: A method based on chance alone by which study participants are assigned to a treatment group. **Randomization** minimizes the differences among groups by equally distributing people with particular characteristics among all the trial arms. The researchers do not know which treatment is better.

Replication

Replication, with **randomization**, will provide a basis for estimating the error variance. ... **Local control**, like **replication** is yet another device to reduce or **control** the variation due to extraneous factors and increase the precision of the experiment.

ANOVA one way

In statistics, one-way analysis of variance is a technique that can be used to compare means of two or more samples. This technique can be used only for numerical response data, the "Y", usually one variable, and numerical or categorical input data, the "X", always one variable, hence "one-way".

Oneway ANOVA practice problems

Problem 1

Using the following data, perform a oneway analysis of variance using α=.05α=.05. Write up the results in APA format.

[Group15145334567][Group22343234345][Group35676748756][Group15145334567][Group22343234345][Group35676748756]

Solution

Sample means ($\bar{x}\bar{x}$) for the groups: = 48.2, 35.4, 69.8

Intermediate steps in calculating the group variances:

[[1]]

	value	mean	deviations	sq deviations
1	51	48.2	2.8	7.84
2	45	48.2	-3.2	10.24
3	33	48.2	-15.2	231.04
4	45	48.2	-3.2	10.24
5	67	48.2	18.8	353.44

[[2]]

	value	mean	deviations	sq deviations
1	23	35.4	-12.4	153.76
2	43	35.4	7.6	57.76
3	23	35.4	-12.4	153.76
4	43	35.4	7.6	57.76
5	45	35.4	9.6	92.16

[[3]]

	value	mean	deviations	sq deviations
1	56	69.8	-13.8	190.44
2	76	69.8	6.2	38.44
3	74	69.8	4.2	17.64
4	87	69.8	17.2	295.84

5	56	69.8	-13.8	190.44

Sum of squared deviations from the mean (SS) for the groups:

[1] 612.8 515.2 732.8

Var1=612.85−1=153.2Var1=612.85−1=153.2

Var2=515.25−1=128.8Var2=515.25−1=128.8

Var3=732.85−1=183.2Var3=732.85−1=183.2

MSerror=153.2+128.8+183.23=155.07MSerror=153.2+128.8+183.23=155.07 *Note: this is just the average within-group variance; it is not sensitive to group mean differences!*

Calculating the remaining *error* (or *within*) terms for the ANOVA table:

dferror=15−3=12dferror=15−3=12

SSerror=(155.07)(15−3)=1860.8SSerror=(155.07)(15−3)=1860.8

Intermediate steps in calculating the variance of the sample means:

Grand mean ($\bar{x}_{grand}$) = 48.2+35.4+69.83=51.1348.2+35.4+69.83=51.13

group	mean	grand mean	deviations	sq deviations
	48.2	51.13	-2.93	8.58
	35.4	51.13	-15.73	247.43
	69.8	51.13	18.67	348.57

Sum of squares (SSmeans)=604.58(SSmeans)=604.58

Varmeans=604.583−1=302.29Varmeans=604.583−1=302.29

MSbetween=(302.29)(5)=1511.45MSbetween=(302.29)(5)=1511.45 *Note: This method of estimating the variance IS sensitive to group mean differences!*

Calculating the remaining *between* (or *group*) terms of the ANOVA table:

dfgroups=3−1=2dfgroups=3−1=2

SSgroup=(1511.45)(3−1)=3022.9SSgroup=(1511.45)(3−1)=3022.9

Test statistic and critical value

F=1511.45155.07=9.75F=1511.45155.07=9.75

Fcritical(2,12)=3.89Fcritical(2,12)=3.89

Decision: reject H0 Decision: reject H0

ANOVA table

source	SS	df	MS	F
group	3022.9	2	1511.45	9.75
error	1860.8	12	155.07	
total	4883.7			

Effect size

η2=3022.94883.7=0.62η2=3022.94883.7=0.62

APA writeup

F(2, 12)=9.75, *p* <0.05, η2η2=0.62.

Problem 2

Using the following summary data, perform a oneway analysis of variance using α=.01α=.01.

⎡⎣⎢⎢⎢⎢n303030mean50.2645.3253.67sd10.4512.7611.47⎤⎦⎥⎥⎥⎥[nmeansd3050.2610.453045.3212.763053.6711.47]

Solution

Var1=10.452=109.2Var1=10.452=109.2
Var2=12.762=162.82Var2=12.762=162.82
Var3=11.472=131.56Var3=11.472=131.56
MSerror=109.2+162.82+131.563=134.53MSerror=109.2+162.82+131.563=134.53 *Note: this is just the average within-group variance; it is not sensitive to group mean differences!*
Calculating the remaining *error* (or *within*) terms for the ANOVA table:

dferror=90−3=87dferror=90−3=87

SSerror=(134.53)(90−3)=11703.82SSerror=(134.53)(90−3)=11703.82

Intermediate steps in calculating the variance of the sample means:

Grand mean ($\bar{x}_{grand}$) = 50.26+45.32+53.673=49.7550.26+45.32+53.673=49.75

group mean	grand mean	deviations	sq deviations

50.26	49.75	0.51	0.26
45.32	49.75	-4.43	19.62
53.67	49.75	3.92	15.37

Sum of squares (SSmeans)=35.25(SSmeans)=35.25

Varmeans=35.253−1=17.62Varmeans=35.253−1=17.62

MSbetween=(17.62)(30)=528.75MSbetween=(17.62)(30)=528.75 *Note: This method of estimating the variance IS sensitive to group mean differences!*

Calculating the remaining *between* (or *group*) terms of the ANOVA table:

dfgroups=3−1=2dfgroups=3−1=2

SSgroup=(528.75)(3−1)=1057.5SSgroup=(528.75)(3−1)=1057.5

Test statistic and critical value

F=528.75134.53=3.93F=528.75134.53=3.93

Fcritical(2,87)=4.86Fcritical(2,87)=4.86

Decision: fail to reject H0 Decision: fail to reject H0

ANOVA table

source	**SS**	**df**	**MS**	**F**
group	1057.5	2	528.75	3.93
error	11703.82	87	134.53	
total	12761.32			

Effect size

η2=1057.512761.32=0.08η2=1057.512761.32=0.08

APA writeup

F(2, 87)=3.93, *p* >=0.01, η2η2=0.08.

Two way Anova

In statistics, the two-way analysis of variance is an extension of the one-way ANOVA that examines the influence of two different categorical independent variables on one continuous dependent variable.

When to use a two-way ANOVA

You can use a two-way ANOVA when you have collected data on a quantitative dependent variable at multiple levels of two categorical independent variables.

A **quantitative variable** represents amounts or counts of things. It can be divided to find a group mean.

Bushels per acre is a quantitative variable because it represents the amount of crop produced. It can be divided to find the average bushels per acre.

A **categorical variable** represents types or categories of things. A level is an individual category within the categorical variable.

Fertilizer types 1, 2, and 3 are levels within the categorical variable **fertilizer type**. Planting densities 1 and 2 are levels within the categorical variable **planting density**.

You should have enough observations in your data set to be able to find the mean of the quantitative dependent variable at each combination of levels of the independent variables.

Both of your independent variables should be categorical. If one of your independent variables is categorical and one is quantitative, use an ANCOVA instead.

How does the ANOVA test work?

ANOVA tests for significance using the F-test for statistical significance. The F-test is a groupwise comparison test, which means it compares the variance in each group mean to the overall variance in the dependent variable.

If the variance within groups is smaller than the variance between groups, the F-test will find a higher F-value, and therefore a higher likelihood that the difference observed is real and not due to chance.

A two-way ANOVA with interaction tests three null hypotheses at the same time:

- There is no difference in group means at any level of the first independent variable.
- There is no difference in group means at any level of the second independent variable.
- The effect of one independent variable does not depend on the effect of the other independent variable (a.k.a. no interaction effect).

A two-way ANOVA without interaction (a.k.a. an additive two-way ANOVA) only tests the first two of these hypotheses.

Two-way ANOVA hypothesesIn our crop yield experiment, we can test three hypotheses using two-way ANOVA:

There is no difference in average yield for any fertilizer type.	There is a difference in average yield by fertilizer type.
There is no difference in average yield at either planting density.	There is a difference in average yield by planting density.
The effect of one independent variable on average yield does not depend on the effect of the other independent variable (a.k.a. no interaction effect).	There is an interaction effect between planting density and fertilizer type on average yield.

Completely randomized design

A completely randomized **design** (**CRD**) is one where the treatments are assigned completely at random so that each **experimental** unit has the same chance of receiving any one treatment. For the **CRD**, any difference among **experimental** units receiving the same treatment is considered as**experimental** error.

Randomized block design

A **randomized block design** is an experimental**design** where the experimental units are in groups called **blocks**. The treatments are randomly allocated to the experimental units inside each**block**. When all treatments appear at least once in each **block**, we have a **completely randomized block design**.

Latin square design

The **Latin square design** applies when there are repeated exposures/treatments and two other factors. ... Agricultural**examples** often reflect geographical**designs** where rows and columns are literally two dimensions of a grid in a field. Rows and columns can be any two sources of variation in an experiment.

*****