

Statistics

Contents

1	Probability	2
2	Population and Sample	8
3	Descriptive Statistics	10
4	Central Tendency and Variability: Questions	16
5	Normal Distributions	19
6	Hypothesis Testing	22
7	Confidence Intervals	23
8	Confidence Intervals: Questions	26
9	z and Confidence: More Questions	29
10	z and t	31
11	ANOVA	34
12	Correlation	41
13	Regression	43
14	Chi-Square	45

1 Probability

If 20% of the bolts produced by a machine are defective, determine the probability that out of 4 bolts chosen at random:

- a) exactly 1 bolt will be defective.
- b) exactly 0 bolt will be defective.
- c) less than 2 bolts will be defective (0 or 1 bolt).

d = defective
n = not defective
p = probability
o = outcome

$$p(d) = .2$$
$$p(n) = 1 - .2 = .8$$

Two possible values (**d** or **n**) and four observations (4 bolts chosen), so there are $2^4 = 16$ possible outcomes. “**dddd**” means 1st bolt **d** AND 2nd **d** AND 3rd **d** AND 4th **d** (all four bolts are defective). You multiply probabilities when you have “AND” situations. The 16 possible outcomes and their probabilities (how likely each outcome will happen) are:

o =	p =
dddd	$.2 \times .2 \times .2 \times .2 = 0.0016$
dddn	$.2 \times .2 \times .2 \times .8 = 0.0064$
ddnd	$.2 \times .2 \times .8 \times .2 = 0.0064$
ddnn	$.2 \times .2 \times .8 \times .8 = 0.0256$
dndd	$.2 \times .8 \times .2 \times .2 = 0.0064$
dndn	$.2 \times .8 \times .2 \times .8 = 0.0256$
dnnd	$.2 \times .8 \times .8 \times .2 = 0.0256$
dnnn	$.2 \times .8 \times .8 \times .8 = 0.1024$
nddd	$.8 \times .2 \times .2 \times .2 = 0.0064$
nddn	$.8 \times .2 \times .2 \times .8 = 0.0256$
ndnd	$.8 \times .2 \times .8 \times .2 = 0.0256$
ndnn	$.8 \times .2 \times .8 \times .8 = 0.1024$
nndd	$.8 \times .8 \times .2 \times .2 = 0.0256$
nndn	$.8 \times .8 \times .2 \times .8 = 0.1024$
nnnd	$.8 \times .8 \times .8 \times .2 = 0.1024$
nnnn	$.8 \times .8 \times .8 \times .8 = 0.4096$

Answers to the questions:

- a) Find the probability that exactly one bolt will be defective. It does not specify which bolt, so you have to add the probabilities of all the outcomes that contain exactly one **d**. Exactly one **d** means **nnnd OR nndn OR ndnn OR dnnn**. When you have “OR” situations, you add probabilities. Remember disjoint events? You cannot have **nnnd AND nndn** at the same time.

$$\begin{aligned} p(1d) &= p(nnnd) + p(nndn) + p(ndnn) + p(dnnn) \\ &= 0.1024 + 0.1024 + 0.1024 + 0.1024 \\ &= 4 \times 0.1024 = 0.4096 \end{aligned}$$

b) $p(0d) = p(nnnn) = 0.4096$

c) $p(0d \text{ OR } 1d) = p(1d) + p(0d) = 0.4096 + 0.4096 = 0.8192$

$$\begin{aligned} p(0d) &= p(nnnn) \\ &= 0.4096 \end{aligned}$$

$$\begin{aligned} p(1d) &= p(nnnd) + p(nndn) + p(ndnn) + p(dnnn) \\ &= 0.1024 + 0.1024 + 0.1024 + 0.1024 \\ &= 0.4096 \end{aligned}$$

$$\begin{aligned} p(2d) &= p(ddnn) + p(dndn) + p(dnnd) + p(nddn) + p(ndnd) + p(nndd) \\ &= 0.0256 + 0.0256 + 0.0256 + 0.0256 + 0.0256 + 0.0256 \\ &= 0.1536 \end{aligned}$$

$$\begin{aligned} p(3d) &= p(dddn) + p(ddnd) + p(dndd) + p(nddd) \\ &= 0.0064 + 0.0064 + 0.0064 + 0.0064 \\ &= 0.0256 \end{aligned}$$

$$\begin{aligned} p(4d) &= p(dddd) \\ &= 0.0016 \end{aligned}$$

You use these combined numbers when you make a histogram (see next page). The probability of having one **d** is 0.4096 (not just 0.1024, which is one specific case of **1d**). They should add up to 1.

$$\begin{aligned} p(0d) + p(1d) + p(2d) + p(3d) + p(4d) \\ &= 0.4096 + 0.4096 + 0.1536 + 0.0256 + 0.0016 \\ &= 1 \end{aligned}$$

Say you randomly sample 4 bolts and record the number of defective bolts in your sample. You repeat this procedure 100 times (i.e., obtain 100 samples). Out of your 100 samples, the number of samples you expect to have one **d** = $p(1d) \times 100 = 0.4096 \times 100 = 40.96 \approx 41$ times.

To plot a histogram, you use the combined numbers. For example, there are four ways you can get one defective bolt: nnnd, nndn, ndnn, and dnnn. You need to add the probability associated with each of these outcomes to obtain the probability of having exactly one defective bolt out of four bolts chosen. That is, $p(1d) = p(nnnd) + p(nndn) + p(ndnn) + p(dnnn)$.

$$\begin{aligned} p(0d) &= p(nnnn) \\ &= 0.4096 \end{aligned}$$

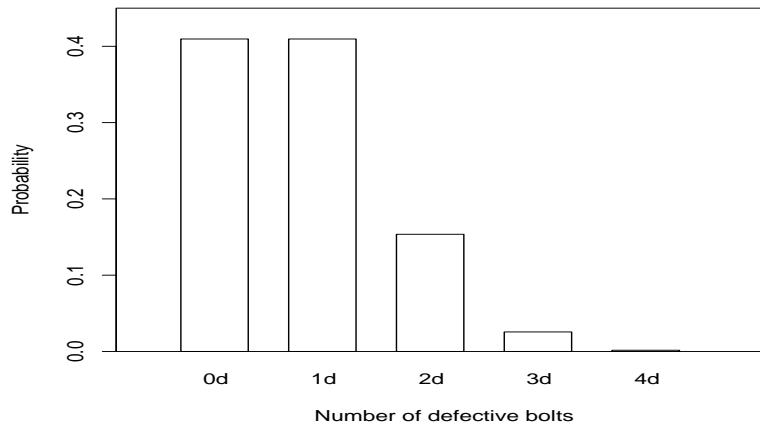
$$\begin{aligned} p(1d) &= p(nnnd) + p(nndn) + p(ndnn) + p(dnnn) \\ &= 0.1024 + 0.1024 + 0.1024 + 0.1024 \\ &= 0.4096 \end{aligned}$$

$$\begin{aligned} p(2d) &= p(ddnn) + p(dndn) + p(dnnd) + p(nddn) + p(ndnd) + p(nndd) \\ &= 0.0256 + 0.0256 + 0.0256 + 0.0256 + 0.0256 + 0.0256 \\ &= 0.1536 \end{aligned}$$

$$\begin{aligned} p(3d) &= p(dddn) + p(ddnd) + p(dndd) + p(nddd) \\ &= 0.0064 + 0.0064 + 0.0064 + 0.0064 \\ &= 0.0256 \end{aligned}$$

$$\begin{aligned} p(4d) &= p(dddd) \\ &= 0.0016 \end{aligned}$$

In histograms, the x-axis will represent the outcome you are interested in, such as the number of heads you will get when you flip a fair coin four times. In this example, the x-axis represents the number of defective bolts in your sample of four bolts. The y-axis might represent the frequency or the probability of obtaining each outcome you are interested in. Here, the y-axis is the probability of having 0d, 1d, 2d, 3d, and 4d.



The probability that a child will be a boy is 50%. Find the probability that in a family of five children there will be:

- a) two boys or more.
- b) not more than one boy.
- c) exactly 3 boys.

b = boy
g = girl
o = outcome
p = probability

$$p(b) = .5$$
$$p(g) = 1 - .5 = .5$$

The number of possible values = 2 (b or g) and the number of observations = 5 (5 children). So there are $2^5 = 32$ possible outcomes (see next page).

Answers to the questions:

- a) You find all the outcomes that contain two or more **b** and add the probabilities. There are 26 outcomes with two or more **b** (see next page). Because all the outcomes are equally likely (probability is the same for each outcome), you can simply do (probability) \times (number of observations).

$$p(\text{two or more b}) = 0.03125 \times 26 = 0.8125$$

- b) You find the outcomes with 0 or 1 **b**. There are 6 such outcomes (see next page).

$$p(\text{not more than one b}) = 0.03125 \times 6 = 0.1875$$

- c) You find the outcomes that contain exactly 3 **b**. There are 10 of them (see next page).

$$p(3b) = 0.03125 \times 10 = 0.3125$$

32 possible outcomes.

g = girl, **b** = boy, **o** = outcome, and **p** = probability.

o =	p =
ggggg	0.03125 ($.5 \times .5 \times .5 \times .5 \times .5$ or $1/32$)
ggggb	0.03125
gggbg	0.03125
gggbb	0.03125
ggbgg	0.03125
ggbgb	0.03125
ggbbg	0.03125
ggbbb	0.03125
gbggg	0.03125
gbggb	0.03125
gbgbg	0.03125
gbgbb	0.03125
gbbgg	0.03125
gbbgb	0.03125
gbbbg	0.03125
gbbbb	0.03125
bgggg	0.03125
bgggb	0.03125
bggbg	0.03125
bggbb	0.03125
bgbgg	0.03125
bgbgb	0.03125
bgbbg	0.03125
bgbbb	0.03125
bbggg	0.03125
bbggb	0.03125
bbgbg	0.03125
bbgbb	0.03125
bbbgb	0.03125
bbbgb	0.03125
bbbgb	0.03125
bbbbg	0.03125
bbbbg	0.03125

Probability distributions

A frequency distribution is simply an indication of the different scores in a distribution along with the frequency with which each occurs. There is usually an interval for which we get the best overall picture of the distribution of scores.

We take a frequency distribution and instead of plotting the absolute number of scores in each interval, we plot the percentage. This percentage distribution is a relative frequency distribution. If we keep plotting the percentage (or proportion) of scores rather than the absolute number of scores and make the interval small, we will have a smooth, continuous curve (instead of discrete bars as in histograms) and the area under the curve will always be 1.



Now we have a probability distribution and the y -value at a particular x -value is a probability density. We can no longer talk about the area represented at a particular x -value because the interval is infinitely small (there is no area at one particular x -value). Even though the area under the curve for an exact value is 0, the area under a section of the curve (such as $x = 0$ to $x = 1$) still has meaning. It will be the probability that a score falls between $x = 0$ and $x = 1$.

When we get to the point of having a probability distribution, we are talking about a theoretical set of scores. The set of scores is population or the set of all possible observations. A subset of population is a sample. We usually obtain data for samples and try to infer something about the population.

2 Population and Sample

Population → parameter

- **Population** refers to the “parent” distribution that contains every possible observation.
 - Often times, it is impossible to obtain population data.
 - If we do have population data, we don’t need to do inferential statistics (descriptive statistics such as finding mean and standard deviation will be enough).
- μ (mu) refers to the mean of the population – true, actual, theoretical mean.
- σ (sigma) refers to the standard deviation of the population – true, actual, theoretical standard deviation. Standard deviation refers to how much individual scores vary around the mean.

Sample → statistics

- **Sample** refers to a subset of population.
 - This is what we usually obtain doing experiments.
 - We will infer what the population will look like from what we know about the sample (inferential statistics).
- \bar{x} (x-bar) refers to the mean of the sample (mean of your sample data).
- s_x refers to the standard deviation of the sample (standard deviation of your sample data).

Sampling distribution of the mean (more on this later)

- Sampling distribution of the mean.
 - We take a sample of size n from the population and find \bar{x} .
 - Repeat lots of times and each time you plot \bar{x} .
 - We'll have a distribution of \bar{x} .
 - This is the sampling distribution of the mean.
- When n (sample size, not the number of samples) is big, *the sampling distribution of the mean will look like a normal distribution, even when the "parent" distribution is not normally distributed* (this will occur more rapidly with larger samples).
- It's good if we can repeat sampling numerous times, but we usually don't (maybe replicates several times). Still, we can estimate the mean of \bar{x} s from an \bar{x} .
 - We can obtain the standard deviation associated with \bar{x} if we know σ .
 - This tells us how much \bar{x} will vary if we resample again and again from the same population and obtain \bar{x} .
 - This is the standard deviation of the sampling distribution of the mean ($s_{\bar{x}}$). $s_{\bar{x}}$ tells you how much each \bar{x} varies around the mean of the sampling distribution of the \bar{x} .
- Even though we don't resample, we pretend that we do. We then have lots of \bar{x} s and can find the mean of \bar{x} s. So each \bar{x} is like an individual score in the sampling distribution of the mean. The standard deviation of the sampling distribution of the mean, therefore, measures how much individual score (\bar{x}) varies around the mean (mean of \bar{x} s). Again, $s_{\bar{x}}$ measures how much \bar{x} s vary from one sample to another. Notice this is the concept behind confidence interval (discussed later). We use $s_{\bar{x}}$ as an error (deviation) measurement (margin of error) in finding confidence interval.
- When σ is known, $s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Notice as n gets bigger, $s_{\bar{x}}$ gets smaller.
- If we take infinitely many repeated samples, the sampling distribution of the mean will have the same mean as the "parent" population.

Using \bar{x} , we can estimate the mean of \bar{x} s. (premise 1)

The mean of \bar{x} s becomes the same as μ with big n . (premise 2)

\bar{x} therefore is a "good" estimator of μ with big n . (conclusion)

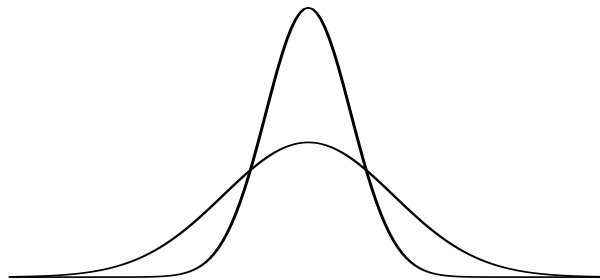
3 Descriptive Statistics

Central tendency (where is the group centered?)

- is a single number that represents the overall group magnitude.
- makes sense when the distribution has one peak (not a good measure for multi-peaked distributions).
- can be measured by:
 - Mean = sum of all scores / number of scores
 - Median = middle score
 - Mode = most frequent score

Variability (how consistent or scattered the data are – spread)

- Central tendency alone leaves out a lot.
- How representative of the distribution is the central tendency?
 - Scores can be tightly clustered around the center.
 - Scores can be widely scattered.
 - Yet the same center (mean, median, mode).
 - So we need variability to represent the data.
- Ways of describing the variability
 - Quartiles (using median)
 - Variance and standard deviation (using mean)



Mean

- is the most useful measure of central tendency for statistics.

- $$= \frac{\sum_{i=1}^n x_i}{n}$$

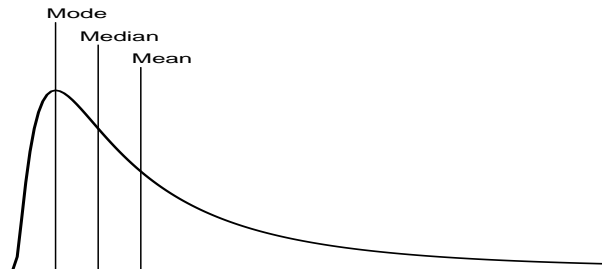
where n is the number of observations and x represents the data.

For example, x could be (0, 1, 2, 3, 4, 5, 6, 7, 8, 9).

Then, mean = $\frac{(0+1+2+3+4+5+6+7+8+9)}{10} = \frac{45}{10} = 4.5$

- is the score, around which the deviations score to 0.
 - A deviation score is how far away one number is from the mean.
 - Using the numbers from the above example, the deviations are:
(0 – 4.5, 1 – 4.5, 2 – 4.5, ..., 7 – 4.5, 8 – 4.5, 9 – 4.5)
= (-4.5, -3.5, -2.5, -1.5, -0.5, 0.5, 1.5, 2.5, 3.5, 4.5)
 - Sum of all the deviation scores equals 0.
 $\sum(-4.5, -3.5, -2.5, -1.5, -0.5, 0.5, 1.5, 2.5, 3.5, 4.5) = 0$
- is the score that makes the sum of the squared deviations minimum.
- is not resistant to outliers – the extreme values in the tail of a skewed distribution pull the mean into the tail.

Positive skew



A distribution is skewed if one of its tails is longer than the other. The distribution shown above has a positive skew; it has a long tail in the positive direction (“skewed to the right”). As an example, family income is a positively skewed distribution. A few very wealthy families will skew the distribution to the right and thus raise the mean. However, a few very wealthy people will have little effect on the median. Thus, the median is a preferred measure of central tendency for family income. But in general, the mean is the most useful measure of central tendency in statistics – the mean is the number which has the smallest squared distance from all other numbers in the distribution.

Median

- is the observation at the 50th percentile (the middle score).
- Half of the observations are above the median, half are below.
- Finding the median:
 1. First rearrange the scores from lowest to highest.
 2. When n is odd, median is the middle observation.
 3. When n is even, median is the mean of the two observations around the middle.
- is resistant to outliers (unaffected by a single outlier).
- is resistant to skew.

Median values are often used for skewed distributions (median housing prices).

- minimizes the sum of the absolute deviations around itself.
- Examples:

$x = (3, 5, 7, 6, 2, 1, 4, 8, 0, 9)$

– First put numbers in order: $(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)$.

– Even number (10) of observations, so $\frac{(4+5)}{2} = 4.5$.

– The median is 4.5.

$x = (3, 5, 7, 6, 2, 1, 4, 8, 0)$

– $(0, 1, 2, 3, 4, 5, 6, 7, 8)$

– Odd number (9) of observations, so find the middle score.

– The median is 4.

Mode

- is the most frequent value.
- is not often used in statistics.
- is used when you want to guess a number in your data and you want to be exactly right the greatest proportion of time.
- Examples:

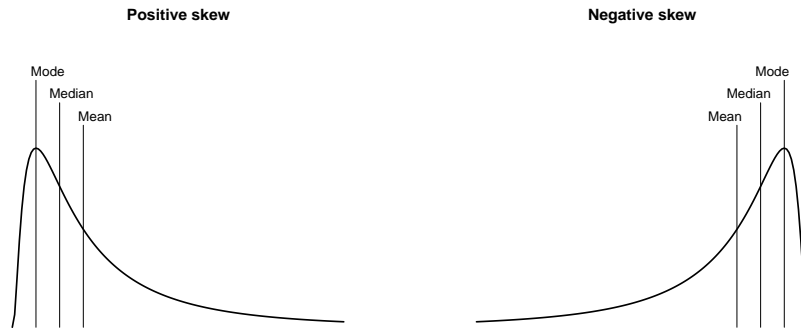
$x = (0, 1, 2, 3, 4, 4, 4, 5, 6, 7, 8, 9)$

The mode is 4.

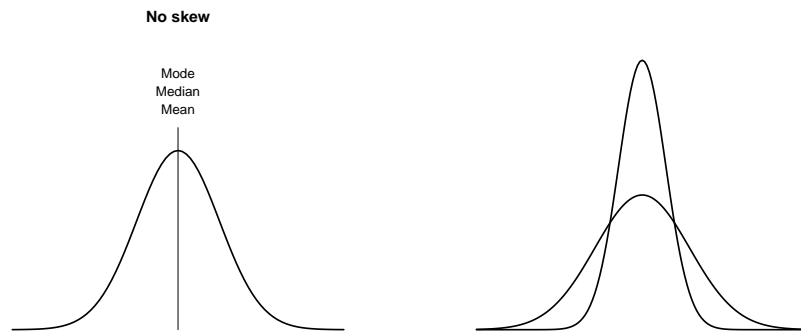
$x = (0, 1, 2, 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 9)$

Two modes, 2 and 5. Bimodal.

Central tendency alone leaves out a lot. Given a mean of scores, you cannot tell how the scores are distributed. The distribution of the scores might be positively skewed like the left distribution shown below. The distribution might have a negative skew (a long tail in the negative direction – “skewed to the left”) like the distribution on the right.



The distribution may be symmetric and have no skew like the distribution shown below. Still, scores can be tightly clustered or widely scattered.



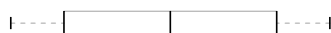
We need variability to represent the data. **Quartiles** are measures of variability using the median and show us potential outliers and how skewed the distribution is. **Variance** and **standard deviation** use the mean and thus are not resistant to outliers but they are most common measures of variability because most of the distributions we deal with are symmetric and have no skew. This is because the sampling distribution will tend toward normality (no skew) even if the parent distribution is not normally distributed (more on this later).

Quartiles

- Measure of variability using the median.
- A percentile is the observation such that $p\%$ of the observations are below that observation.
- The 25th percentile is often called the first quartile ($q1$).
 $q1$ is the median of the observations from the smallest to the median.
- The median is the 50th percentile ($q2$).
- The 75th percentile is the third quartile ($q3$).
 $q3$ is the median of the observations from the median to the largest.
- The interquartile range is the size of the interval between $q1$ and $q3$.
- The five number summary
 - A distribution can be summarized by five numbers.
 - minimum, $q1$, median ($q2$), $q3$, maximum
 - The five number summary can be graphed as a box-plot.

Examples:

Scores = (2, 3, 4, 5, 2, 3, 4, 5, 2, 3, 4, 5, 2, 3, 4, 5, 2, 3, 4, 5)
 [min, $q1$, $q2$, $q3$, max] = [2, 2.75, 3.5, 4.25, 5]



Scores = (1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7)
 [min, $q1$, $q2$, $q3$, max] = [1, 1.75, 3, 5, 7]



Scores = (4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 5, 5, 5, 5, 10)
 [min, $q1$, $q2$, $q3$, max] = [3, 3.75, 4, 4.25, 10] or
 [min, $q1$, $q2$, $q3$, max] = [3, 3.75, 4, 4.25, 5] + outlier 10



Standard deviation

- The most common measure of the variability using the mean.
- Not the most resistant measure because the mean is used to calculate standard deviation.
- How do you find the standard deviation?

An example: Scores = (0, 1, 2, 3, 4)

1. Find the mean.

$$\text{Mean} = 2$$

2. Square the deviation scores and make them all positive.

$$\text{Deviations}^2 = (-2, -1, 0, 1, 2)^2 = (4, 1, 0, 1, 4)$$

3. Sum all of the squared deviations and divide by the number of observations minus 1 (degrees of freedom, $n - 1$).

$$\text{Variance} = \frac{\sum(4,1,0,1,4)}{5-1} = 2.5$$

Variance = mean squared deviations

$$\text{Variance } (s_x^2) = \frac{\sum(\text{each score} - \text{mean})^2}{n-1} = \frac{\sum(x-\bar{x})^2}{n-1}$$

- We use $n - 1$ instead of n because we are using the mean and only $n - 1$ scores are truly random or free to vary.
- Variance is good, but its scale is the square of the values rather than the values themselves. So, we take the square root of the variance.

4. Take the square root of the variance.

$$\text{Standard deviation} = \sqrt{2.5} = 1.58$$

Standard deviation = square root of variance

$$\text{Standard deviation } (s_x) = \sqrt{s_x^2} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

We will think of a score's position in a group as how many standard deviations above or below the mean the score falls. This is called a **standard score** or a **z score** (more on this later).

$$z = \frac{\text{score} - \text{mean}}{\text{standard deviation}}$$

When we say a particular value has a z score of 1, we are saying that it is one standard deviation above the mean. Many groups of scores tend to follow a pattern known as the **normal distribution**, in which there will be about **68% of the scores within one standard deviation of the mean** (i.e., between $z = -1$ and $z = 1$). One nice thing about transforming the data to z scores is that **the mean is always 0 and the standard deviation is always 1 when we use z scores**.

4 Central Tendency and Variability: Questions

In a survey, 13 people gave the age of their oldest child. The ages mentioned were: 6, 10, 3, 19, 3, 10, 13, 10, 2, 11, 12, 3, 8.

1. What is the mean age?
2. What is the median?
3. What is the mode?
4. What is the variance?
5. What is the standard deviation?
6. $\sum(x + 1) = ?$ $x = (6, 10, 3, 19, 3, 10, 13, 10, 2, 11, 12, 3, 8)$
7. $\sum(3x) = ?$
8. $(\sum x)^2 = ?$
9. Express the first three ages (6, 10, and 3) as z scores.
10. If you change all the scores to z score, what would the new mean and standard deviation be?
11. What would the mean and standard deviation be if you multiplied the raw data by 4.5 and subtracted 2 from the product, then change all of the transformed data to z scores?

Central Tendency and Variability: Answers

$$x = (6, 10, 3, 19, 3, 10, 13, 10, 2, 11, 12, 3, 8)$$

$$\begin{aligned} 1. \text{ Mean} &= \frac{\sum x}{n} \\ &= \frac{6+10+3+19+3+10+13+10+2+11+12+3+8}{13} \\ &= 8.461538 \end{aligned}$$

$$\mathbf{Mean} = 8.46$$

If you had to guess the age of the child in the sample and always guessed 8.46, you would be wrong on every guess (because there is no 8.46) but your mistake in guessing would be as small as possible.

$$\sum(x - \text{mean}) = 0$$

2. Put the values in order.

$$2, 3, 3, 3, 6, 8, 10, 10, 10, 11, 12, 13, 19$$

$$\mathbf{Median} = \text{middle most value} = 10$$

Use the median when outliers could affect the descriptive value of the typical score.

3. Mode is the most frequent value.

$$\mathbf{Modes} = 3 \text{ and } 10$$

The distribution has two modes (bimodal).

Use the mode when you want to have the best chance of being exactly correct.

4. Variance = $\frac{\text{sum of squared deviations}}{\text{degrees of freedom}}$

$$x = (6, 10, 3, 19, 3, 10, 13, 10, 2, 11, 12, 3, 8)$$

$$\text{Mean} = 8.46$$

Deviation = $x - \text{mean} =$

$$(-2.46, 1.54, -5.46, 10.54, -5.46, 1.54, 4.54, 1.54, -6.46, 2.54, 3.54, -5.46, -0.46)$$

Squared deviation = $(x - \text{mean})^2 =$

$$(6.0516, 2.3716, 29.8116, 111.0916, 29.8116, 2.3716, 20.6116, 2.3716, 41.7316, 6.4516, 12.5316, 29.8116, 0.2116)$$

Sum of squared deviations = $\sum(x - \text{mean})^2 =$

$$6.0516 + 2.3716 + 29.8116 + 111.0916 + 29.8116 + 2.3716 + 20.6116 + 2.3716 + 41.7316 + 6.4516 + 12.5316 + 29.8116 + 0.2116 = 295.2308$$

Variance = $\frac{\text{sum of squared deviations}}{\text{degrees of freedom}}$

$$= \frac{\sum(x - \text{mean})^2}{n-1} = \frac{295.2308}{12} = 24.60257$$

Variance = 24.60

5. Standard deviation = $\sqrt{\text{variance}} = \sqrt{24.60} = 4.959839$

Standard deviation = 4.96

6. $\sum(x + 1) = (6 + 1) + (10 + 1) + \dots + (3 + 1) + (8 + 1) = 123$

7. $\sum 3x = (6 \times 3) + (10 \times 3) + \dots + (3 \times 3) + (8 \times 3) = 330$

Also $3(\sum x) = 3(110) = 330$

8. $(\sum x)^2 = (110)^2 = 12100$

9. $z = \frac{x - \text{mean}}{\text{standard deviation}}$

6: $z = \frac{6 - 8.46}{4.96} = -0.4959677 = -0.50$

10: $z = \frac{10 - 8.46}{4.96} = 0.3104839 = 0.31$

3: $z = \frac{3 - 8.46}{4.96} = -1.100806 = -1.10$

10. New mean = 0 and new standard deviation = 1. The purpose of transforming the data to z scores is to have **standardized values with a mean of 0 and a standard deviation of 1.**

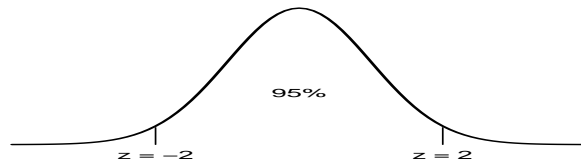
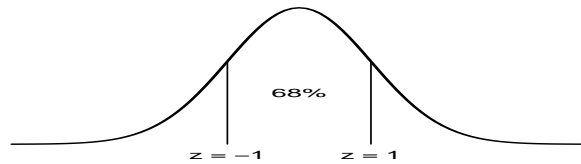
11. The mean will still be 0 and the standard deviation will still be 1. Nice.

5 Normal Distributions

There are many theoretical probability distributions. One is of overwhelming importance and is called normal distributions. The distribution of many different quantities in the real world tend to follow the normal distribution. It is a bell-shaped, symmetrical distribution we are familiar with.

There are infinitely many normal distributions defined by different values of the two parameters determining the exact function. The two parameters are the mean and the standard deviation. We can take scores in any normal distribution and express each in standard deviation units above or below the mean. The resultant numbers will have a mean of 0 and a standard deviation of 1. This normalized (standardized) score is a z score. The distribution of z scores is called standardized normal distribution.

When we say a particular value has a z score of 1, we are saying that it is one standard deviation above the mean. From this we can figure out where in the distribution it rests (what proportion of the scores are above or below it). As shown in the standardized normal distribution below, there will be about 68.27% of scores within one standard deviation of the mean (between $z = -1$ and $z = 1$). About 95.45% of scores will be within two standard deviations of the mean. Three standard deviations will have 99.73% of scores.



Sampling distribution

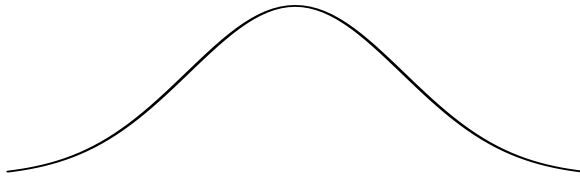
We choose a random sample of n values from a normal distribution. We calculate the mean and record it. We choose another sample of n values, record the mean along with the first. If we continue this process indefinitely, we will have a large set of means of samples of n . In theory, the process could be continued forever and we could end up with a population of possible means of samples of n drawn from some “parent” population.

This population of means is called the sampling distribution of the mean. When we randomly sample n values from a normal parent, we are more likely to sample values that are around the mean, especially when we sample a large number of values. This is because there are more values around the mean than away from the mean. Then, the mean of the sampling distribution of the mean is an unbiased, consistent, and sufficient estimator of the mean of the parent population. In addition, because we are more likely to sample values around the mean, our sample will have less variability than the parent population. Again, this is especially true with big n .

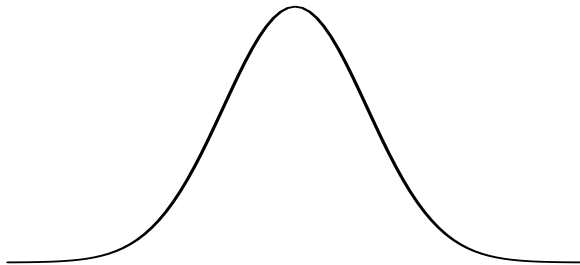


Standard deviation of the sampling distribution of the mean, $s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the parent population. Notice, if we sample one value from the parent population, the standard deviation of the sampling distribution of the mean is the same as the standard deviation of the parent population ($\frac{\sigma}{\sqrt{1}} = \sigma$). But with a large sample size, the standard deviation of the sampling distribution of the mean becomes smaller than the standard deviation of the parent population.

Parent distribution



Sampling distribution of the mean



Central limit theorem

One theorem that serves as a basis for nearly all of the tests we use is the central limit theorem. It says if we take infinitely many repeated samples of size n from a normally distributed population (“parent” population) and form a new distribution of sample \bar{x} s (sampling distribution of the mean), then the distribution of the \bar{x} s will have the same mean as the “parent” population and will have a standard deviation of $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the “parent” population. This distribution of the means is called the sampling distribution of the mean and its standard deviation is called the standard deviation of the sampling mean.

Central limit theorem also says that even if the parent distribution is not normally distributed, the sampling distribution will tend toward normality. This will occur more rapidly with larger sample sizes.

6 Hypothesis Testing

Is the result statistically significant? Do we reject the null? To make our decision, we need to figure out how likely, or unlikely, the result (our observation) is ASSUMING OUR NULL HYPOTHESIS IS RIGHT.

Null hypothesis is like a model under which we can calculate the sample distribution of statistics and thus the probability of results. We think of the null model as being a hypothesis that groups don't differ. Or we can have something like the population mean is equal to 7.

How much emphasis should we put on failing to reject the null hypothesis? If we fail to reject the null hypothesis, we cannot claim that the means of the groups are the same (hence fail to reject rather than accept the null). We could only say that we do not have sufficient evidence to claim that the groups are different. Clearly, if we have a great deal of variance, whether because of natural variability or a lack of experimental control, the groups could be quite different and we would still fail to detect it. We may have an inadequate n for the design. If we have only modest power, we clearly cannot put a very strong claim on the truth of the null hypothesis. Power, by the way, is the probability of rejecting the null hypothesis when it is false.

Are two groups different? Suppose we find that the means for the two groups are different (when we do an experiment, we usually do not get the same mean for each group). We want to know whether the observed difference is large enough that the two groups indeed differ or the observed difference between groups is simply due to chance. We need to figure out how likely, or unlikely, the result (the observed difference) is assuming that the two groups, in fact, do not differ (the null hypothesis).

Say we find that the probability of finding the result, assuming the groups do not differ, is .02 ($p = .02$). It means that in cases where the null hypothesis is right, we will incorrectly reject it 2% of the time in the long run. In other words, our result can happen simply by chance 2% of the time assuming that the two groups do not differ.

Is 2% rare enough to conclude that the null hypothesis is wrong? That's something we decide. We decide that 5% is unlikely enough to reject the null hypothesis ($\alpha = .05$). Then we conclude that our result ($p = .02$) is statistically significant. In other words, the observed difference or effect (our result) is so large that it is unlikely, or $< .05$, to occur by chance if indeed the groups do not differ (assuming the null hypothesis is right). We then reject the null hypothesis.

Again, how unlikely the result has to be before we reject the null hypothesis is something we choose. It is usually .05 or .01 (α level). If instead of .05 we decide that .01 is unlikely enough to reject the null, then our result is no longer statistically significant because $p = .02$ is not unlikely enough.

7 Confidence Intervals

Generally, we want to know the population mean not just sample.

- The population mean is usually unknown, however.
- How can we make a guess about what the population mean actually is?
- Collect a sample. The mean of the sample is an estimate of the population mean.

How is the sample mean related to the population mean?

- Sampling distribution of the mean is a distribution of means of samples of a particular size.
- We take a lot of samples of size n , record the means each time, and make a distribution of the means.
- Now we find the mean of this sampling distribution of the means.
- This mean of the sampling distribution of the means will be approximately the same as the population mean according to the central limit theorem.

How closely sample mean will approximate the population mean will depend on the sample size. according to the central limit theorem:

- The larger the sample size, the more accurate the estimate of the true/population mean, and
- For a large sample size, the sampling distribution of the mean is normally distributed even if the population distribution is not normally distributed.

Where is the population mean likely to be?

- You can construct an interval that is likely to contain the population mean.
- This interval is called a **confidence interval**.
- You can build an interval that is as wide as the confidence you want to have in it.

Building confidence intervals (a procedure to use when σ is known).

You want to find a confidence interval of the population mean (μ) using the sample mean (\bar{x}). You are guessing where the μ is likely to be located.

- a) Use sample mean (\bar{x}).
- b) Use standard deviation of the sampling distribution of the mean ($s_{\bar{x}}$). So you have to do $\frac{\sigma}{\sqrt{n}}$ unless the given standard deviation is the standard deviation of the sampling distribution of the mean.
- c) Use the z score associated with the level of confidence you are interested in.
- d) Confidence interval of the population mean (μ) is:

$$\bar{x} - (z \times s_{\bar{x}}) < \mu < \bar{x} + (z \times s_{\bar{x}})$$

$-(z \times s_{\bar{x}})$ and $+(z \times s_{\bar{x}})$ are margins of error.

Say you want to find a 95% confidence interval. Then when you calculate a z score, you want it to fall within 95% area under the z curve. You don't want your z score to fall under 2.5% areas in the two extremes. So you want your z score to be anywhere between -1.96 and 1.96 (see z table). So for 95% confidence interval, you use 1.96 as your z value.

Confidence intervals: An example

Sample mean (\bar{x}) is 100

Standard deviation (σ) is 10

Sample size (n) is 25

95% confidence interval of the population mean (μ)

- a) $\bar{x} = 100$
- b) $s_{\bar{x}} = \frac{10}{\sqrt{25}} = 2$
- c) $z(95\%) = 1.96$ (2.5% in each tail)
- d) $100 - (1.96 \times 2) < \mu < 100 + (1.96 \times 2)$
 $= 96.08 < \mu < 103.92$

You have 95% level of confidence that the mean will fall between 96.08 and 103.92 if you resample again and again and again. Because the mean of the sampling distribution of the mean is the best estimate of the mean of the population mean, we can say that we are estimating the confidence interval of the population mean (or true/actual mean).

Confidence intervals: Why?

When we want to analyze individual scores, $z = \frac{\text{score} - \text{mean}}{\text{standard deviation}}$ or $z = \frac{x - \mu}{\sigma}$. We are standardizing individual scores or estimating how far individual scores are from the mean. When we build confidence intervals, we are doing the same thing except that the individual score is \bar{x} and the mean is the mean of \bar{x} s. We are using the sampling distribution of the mean. So, we use $s_{\bar{x}}$ rather than σ .

$$z = \frac{\bar{x} - \text{mean of } \bar{x}s}{s_{\bar{x}}}$$

According to the central limit theorem, with big n , mean of $\bar{x}s = \mu$. then,

$$z = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

μ is the unknown, and we want to estimate using \bar{x} . If we want to find 95% confidence interval, we want our z score to be between -1.96 and 1.96 .

$$\begin{array}{l|l} & -1.96 < z < 1.96 \\ \times s_{\bar{x}} & -1.96 < \frac{\bar{x} - \mu}{s_{\bar{x}}} < 1.96 \\ -\bar{x} & -1.96 \times s_{\bar{x}} < \bar{x} - \mu < 1.96 \times s_{\bar{x}} \\ \times (-1) & -\bar{x} - (1.96 \times s_{\bar{x}}) < -\mu < -\bar{x} + (1.96 \times s_{\bar{x}}) \\ \text{notice } < \text{ changes to } > & \bar{x} + (1.96 \times s_{\bar{x}}) > \mu > \bar{x} - (1.96 \times s_{\bar{x}}) \\ \text{notice } > \text{ changes to } < & \bar{x} - (1.96 \times s_{\bar{x}}) < \mu < \bar{x} + (1.96 \times s_{\bar{x}}) \end{array}$$

We know everything else but population mean. We just want to know what value (population mean) will produce the z score that falls within the confidence interval we choose. Using the numbers from the previous example:

$$\begin{array}{l|l} \times 2 & -1.96 < \frac{100 - ?}{2} < 1.96 \\ -100 & -1.96 \times 2 < 100 - ? < 1.96 \times 2 \\ \times (-1) & -3.92 - 100 < -? < 3.92 - 100 \\ \text{notice } < \text{ changes to } > & 3.92 + 100 > ? > -3.92 + 100 \\ \text{notice } > \text{ changes to } < & 100 - 3.92 < ? < 100 + 3.92 \\ & 96.08 < ? < 103.92 \end{array}$$

Or you can just solve two equations for $?$:

$$\begin{array}{ll} -1.96 = \frac{100 - ?}{2} & 1.96 = \frac{100 - ?}{2} \\ 100 - ? = -3.92 & 100 - ? = 3.92 \\ -? = -103.92 & -? = -96.08 \\ ? = 103.92 & ? = 96.08 \end{array}$$

and make the interval $96.08 < ? < 103.92$.

8 Confidence Intervals: Questions

1. Suppose you are scouting for potential olympic long jumpers. You observe 5000 sixth graders in the standing broad jump. The distribution of the sample looks well-behaved and you find:

Mean = 6.53 feet

Standard deviation = 1.14

You pick 6 numbers from your 5000 numbers and want to find whether or not there is a potential candidate for long jump. Find the z -scores for jumps of:

6.21 feet.

6.53 feet.

4.38 feet.

7.21 feet.

9.77 feet.

3.11 feet.

2. $\bar{x} = 100$, $n = 100$, and $s_{\bar{x}} = 10$.

Find the 95% confidence interval of the μ .

Find the 99% confidence interval of the μ .

3. $\bar{x} = 15$, $n = 100$, and $\sigma = 5$.

Find the 95% confidence interval of the μ .

Find the 99% confidence interval of the μ .

4. $\bar{x} = 32$, $n = 100$, and $s_{\bar{x}} = 5$.

Find the value of σ .

Find the 95% confidence interval of the μ .

Find the 99% confidence interval of the μ .

5. The standard deviation for the graduation age is 5.27. In the spring of 1990, a random sample of 75 college graduates showed a mean age of 23.1. Find the 95% confidence interval for the mean age at graduation in 1990.

Confidence Intervals: Answers

1. The mean = 6.53 is μ and the standard deviation = 1.14 is σ . This could get confusing. We obtained a sample of 6 jumps from a population of 5000 in this question. We want to know how far away individual score (as opposed to the mean or \bar{x} of the sample of 6) is from the μ , so we use σ (rather than $s_{\bar{x}}$). The formula is $\frac{x-\mu}{\sigma}$.

$$6.21 \rightarrow \frac{6.21-6.53}{1.14} = -0.2807018$$

$$6.53 \rightarrow \frac{6.53-6.53}{1.14} = 0$$

$$4.38 \rightarrow \frac{4.38-6.53}{1.14} = -1.885965$$

$$7.21 \rightarrow \frac{7.21-6.53}{1.14} = 0.5964912$$

$$9.77 \rightarrow \frac{9.77-6.53}{1.14} = 2.842105$$

$$3.11 \rightarrow \frac{3.11-6.53}{1.14} = -3$$

Remember, when you use z scores, **the mean is always 0 and the standard deviation is always 1**. z score of 0 means that the score is 0 standard deviation away from the mean (same as the mean). z score of 2 means the score is 2 standard deviation above the mean, and z score of -2 is 2 standard deviation below the mean. We know that for a normal distribution, about 95% of observations fall within 2 standard deviation away from (both below and above) the mean. Then, the probability of observing a score that is 2 standard deviation away from the mean is less than 2.5%. $z = 2.84$ (jump = 9.77 feet), then, is a rare event. Her jump is way above the mean. $z = -3$ (jump = 3.11 feet) is also a rare event. His jump is extremely below the mean. Look at the z table and find out the probability of having -3 or lower z score.

2. Since we know $s_{\bar{x}}$, we don't need to use n .

95% confidence interval : $\bar{x} = 100$, $s_{\bar{x}} = 10$, $z = 1.96$ (from the table).

$$100 - (1.96 \times 10) < \mu < 100 + (1.96 \times 10) \\ 80.4 < \mu < 119.6$$

99% confidence interval : $\bar{x} = 100$, $s_{\bar{x}} = 10$, $z = 2.58$ (from the table).

$$100 - (2.58 \times 10) < \mu < 100 + (2.58 \times 10) \\ 74.2 < \mu < 125.8$$

When calculating the 95% confidence interval, we have to split the 5% "error" into two tails of 2.5% each. 1.96 is the z score at which the upper 2.5% of the curve is cut off (or 97.5% are below 1.96). Notice 99% confidence interval is wider. Why? You want to be 99% sure that the mean will fall within this interval if you resample again. If you have a wider interval, you are more likely to be right.

3. We don't know $s_{\bar{x}}$, so we need to calculate $s_{\bar{x}}$.

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{10} = 0.5.$$

95% confidence interval : $\bar{x} = 15$, $s_{\bar{x}} = 0.5$, $z = 1.96$ (from the table).

$$\begin{aligned} 15 - (1.96 \times 0.5) &< \mu < 15 + (1.96 \times 0.5) \\ 14.02 &< \mu < 15.98 \end{aligned}$$

99% confidence interval : $\bar{x} = 15$, $s_{\bar{x}} = 0.5$, $z = 2.58$.

$$\begin{aligned} 15 - (2.58 \times 0.5) &< \mu < 15 + (2.58 \times 0.5) \\ 13.71 &< \mu < 16.29 \end{aligned}$$

4. $\bar{x} = 32$, $n = 100$, and $s_{\bar{x}} = 5$.

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ then } \sigma = s_{\bar{x}}\sqrt{n} = 5\sqrt{100} = 50$$

95% confidence interval : $\bar{x} = 32$, $s_{\bar{x}} = 5$, $z = 1.96$.

$$\begin{aligned} 32 - (1.96 \times 5) &< \mu < 32 + (1.96 \times 5) \\ 22.2 &< \mu < 41.8 \end{aligned}$$

99% confidence interval : $\bar{x} = 32$, $s_{\bar{x}} = 5$, $z = 2.58$.

$$\begin{aligned} 32 - (2.58 \times 5) &< \mu < 32 + (2.58 \times 5) \\ 19.1 &< \mu < 44.9 \end{aligned}$$

5. $n = 75$, $\bar{x} = 23.1$, $\sigma = 5.27$.

$$s_{\bar{x}} = \frac{5.27}{\sqrt{75}} = 0.6085272$$

95% confidence interval : $\bar{x} = 23.1$, $s_{\bar{x}} = 0.61$, $z = 1.96$.

$$\begin{aligned} 23.1 - (1.96 \times 0.61) &< \mu < 23.1 + (1.96 \times 0.61) \\ 21.9 &< \mu < 24.3 \end{aligned}$$

9 z and Confidence: More Questions

1. Jane needs to score in the top 2.5% percent of an aptitude test to qualify for a job. The mean of test is 75 and standard deviation is 10. How high of a score does she need to get?
2. The standard deviation for the number of students attending a class is 5. Over five classes, 25, 35, 30, 26, and 20 students attend the class. What is the mean number of students attending for these five classes? What is the standard deviation associated with this mean?
3. Mia needs to figure out the mean number of hours children spend watching TV per week. She needs the standard deviation associated with the mean to be less than or equal to 2. The standard deviation for how many hours children watch TV is 6 hours per week. How many children does she need to survey?
4. Dave runs an experiment with 1241 subjects. Kate runs the same experiment (on the same population) with 476 subjects. What is Kate's standard deviation of the mean divided by Dave's standard deviation of the mean?
5. The weight of a seal has a standard deviation of 20 pounds. An animal trainer at Sea World is interested in how heavy the seals are at the park (on average). Eight seals at the park are weighed and weigh 200, 220, 300, 175, 218, 315, 180, and 200 pounds. What is the mean weight of the seals? What is the 95% confidence interval for the above mean? What is the 99% confidence interval for the above mean?

z and confidence: Answers

1. Top 2.5%, so $z = 1.96$ or higher. An individual score vs. the mean, so use the standard deviation as it is. No need to obtain the estimate ($s_{\bar{x}}$).

$$\frac{x-75}{10} \geq 1.96 \quad \text{We want to know what score } x \text{ Jane needs.}$$

$$x - 75 \geq 1.96 \times 10$$

$$x \geq 19.6 + 75 = 94.6$$

Jane needs to score 94.6 or higher.

2. $\bar{x} = \frac{25+35+30+26+20}{5} = 27.2$
 $s_{\bar{x}} = \frac{5}{\sqrt{5}} = 2.236068$

The mean would stay within 2.236068 away from 27.2 if repeated.

3. Mia needs to survey 9 or more kids. Why? Because $\frac{6}{\sqrt{n}} = 2$ or less. \sqrt{n} has to be 3 or larger. n has to be 9 or more.

$$\frac{\sigma}{\sqrt{n}} \leq 2$$

$$\frac{6}{\sqrt{n}} \leq 2$$

$$6 \leq 2\sqrt{n}$$

$$3 \leq \sqrt{n}$$

$$9 \leq n$$

4. Suppose $s = 10$ for both because the same population is used. They cancel out anyway, so you don't need to know.

$$\frac{\frac{10}{\sqrt{476}}}{\frac{10}{\sqrt{1241}}} = \frac{10}{\sqrt{476}} \frac{\sqrt{1241}}{10} = \frac{\sqrt{1241}}{\sqrt{476}} = 1.614665$$

or you can just do $\sqrt{\frac{1241}{476}} = 1.614665$

5. $\bar{x} = \frac{200+220+300+175+218+315+180+200}{8} = 226$

We are estimating the mean instead of the individual scores. So we use $s_{\bar{x}}$ instead of σ .

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{8}} = \frac{20}{2.828427} = 7.071068$$

95% confidence interval:

$$226 - (1.96 \times 7.071068) < \text{mean} < 226 + (1.96 \times 7.071068)$$

$$212.1407 < \text{mean} < 239.8593$$

99% confidence interval:

$$226 - (2.58 \times 7.071068) < \text{mean} < 226 + (2.58 \times 7.071068)$$

$$207.7566 < \text{mean} < 244.2434$$

When to use what:

Continuous data

Two groups $\rightarrow z$ and t

More than two groups \rightarrow ANOVA

Categorical data (when data are not continuous)

Chi-square (χ^2) test

10 z and t

- σ known

- Comparing a single score to the population mean (x vs. μ)

$$z = \frac{x - \mu}{\sigma}$$

- Comparing a sample mean to a population (\bar{x} vs. μ)

$$z = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{where } s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- σ unknown

- Comparing a single score to the sample mean (x vs. \bar{x})

$$t = \frac{x - \bar{x}}{s_x}$$

used when calculating r (correlation).

- Comparing a sample mean to a population (\bar{x} vs. μ)

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

where $s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$ where $s_x = \sqrt{\frac{df s_x^2}{df}} = \sqrt{\frac{\text{sum squared deviations}}{df}}$

A different t distribution for each $df = n - 1$ (degrees of freedom)

- Comparing two sample means (\bar{x}_1 vs. \bar{x}_2)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{p}}}$$

where $s_{\bar{p}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where $s_p = \sqrt{\frac{(df_1)(s_{x1})^2 + (df_2)(s_{x2})^2}{df_1 + df_2}}$

s_p is the pooled standard deviation for the two groups.

df_1 is the df and s_1^2 is the variance for group 1,

df_2 is df and s_2^2 is the variance for group 2,

n_1 is the number of subjects in group 1,

and n_2 is the number of subjects in group 2.

$s_{\bar{p}}$ looks complicated but is calculated the same way as $s_{\bar{x}}$.

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = s_x \frac{1}{\sqrt{n}} = s_x \sqrt{\frac{1}{n}} \quad s_{\bar{p}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Comparing two sample means: An example

An American and an Englishman get in an argument over who is more boring: Canadians or Belgians (subsequently a Belgian and a Canadian get in an argument over the relative intelligence levels of the English and the Americans, but that is a different story). To resolve the debate, the American and the Englishman collect some data on Canadians and Belgians using the North Atlantic Boredom Scale (NABS). A high rating indicates a boring person. The following data are obtained:

Canadians: 25, 5, 24, 14
Belgians: 31, 41, 22, 42

- a. State the null and alternative hypothesis.

Null: Canadians and Belgians do not differ in their level of boringness.
Alternative: Canadians and Belgians differ in their level of boringness.

- b. Do Canadians and Belgians differ significantly in their level of boringness (use α at .05)?

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{p}}} \quad \text{where } s_{\bar{p}} = \sqrt{\frac{(df_1)(s_{x1})^2 + (df_2)(s_{x2})^2}{df_1 + df_2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\bar{x}_1 = \frac{25+5+24+14}{4} = 17 \quad \bar{x}_2 = \frac{31+41+22+42}{4} = 34$$

$$n_1 = 4 \quad n_2 = 4 \quad df_1 = 4 - 1 = 3 \quad df_2 = 4 - 1 = 3$$

$$(df_1)(s_{x1})^2 = 3 \frac{(25-17)^2 + (5-17)^2 + (24-17)^2 + (14-17)^2}{3} = 266$$

$$(df_2)(s_{x2})^2 = 3 \frac{(31-34)^2 + (41-34)^2 + (22-34)^2 + (42-34)^2}{3} = 266$$

$$s_{\bar{p}} = \sqrt{\frac{266+266}{3+3}} \sqrt{\frac{1}{4} + \frac{1}{4}} = \sqrt{88.66667} \sqrt{0.7071068} = 6.658328$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{p}}} = \frac{17-34}{6.658328} = -2.553194$$

The critical value for t with 6 ($df_1 + df_2$) df at .05 α (upper tail .025 + lower tail .025) is 2.447. We observed $t(6) = 2.553$. We reject the null hypothesis that Canadians and Belgians do not differ in their level of boringness. They differ significantly at α .05 level.

z and t

- σ known

- Comparing a single score to the population mean (x vs. μ)

$$z = \frac{x - \mu}{\sigma}$$

- Comparing a sample mean to a population (\bar{x} vs. μ)

$$z = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{where } s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

n = the number of scores

$$\bar{x} = \frac{\sum(\text{each score})}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

- σ unknown

- Comparing a single score to the sample mean (x vs. \bar{x})

$$t = \frac{x - \bar{x}}{s_x}$$

used when calculating r (correlation).

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$$

- Comparing a sample mean to a population (\bar{x} vs. μ)

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

where $s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$

A different t distribution for each $df = n - 1$ (degrees of freedom)

- Comparing two sample means (\bar{x}_1 vs. \bar{x}_2)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{p}}}$$

where $s_{\bar{p}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

where $s_p = \sqrt{\frac{(df_1)(s_{x1})^2 + (df_2)(s_{x2})^2}{df_1 + df_2}}$

11 ANOVA

ANOVA: Analysis of Variance

- We usually think of ANOVA as a way to detect significant variation of means in three or more groups, but there is no reason we cannot apply it to two groups.
- We are going to calculate two variance estimates. One is the averaged or pooled within group variance estimate. The other is the variance estimate based on the variability among group means (between group estimate).
- We take the ratio of the two variances (between/within) and refer this value to the F -distribution with the appropriate degrees of freedom to find out if there is significantly more scatter among the means (between) than we would expect on the basis of the scatter of individual scores within the group.
- The null hypothesis of the analysis is that the variance of group means around the grand mean (between group variance) is due to chance ($\mu_1 = \mu_2 = \mu_3 \dots$).

An example of ANOVA

Group 1: 11, 25, 22, 17, 20

Group 2: 24, 34, 33, 25, 29

Group 3: 8, 28, 22, 17, 15

The null hypothesis is $\mu_1 = \mu_2 = \mu_3$.

The alternative is they are not all equal.

The results of ANOVA computations are normally presented in a table.

Source	SS	df	MS	F	p
Between					
Within					
Total					

SS : sum of squared deviations

df : degrees of freedom

MS : mean squared deviations or variance ($\frac{SS}{df}$)

F : F -value ($\frac{MS_{between}}{MS_{within}}$)

p : p -value

Group 1: 11, 25, 22, 17, 20

Group 2: 24, 34, 33, 25, 29

Group 3: 8, 28, 22, 17, 15

We want to figure out $MS_{between}$ (between group variance) and MS_{within} (within group variance) to calculate F -ratio:

$$F\text{-ratio} = \frac{MS_{between}}{MS_{within}}$$

- Between: between group (“treatment”) variance. Treatments raise or lower the group means but they do not affect within group variance. So, the variance of the group means around the overall (“grand”) mean measures treatment effects.
- Within: within group (“error”) variance. Random errors raise or lower individual group member scores but do not change the group mean (because random errors tend to cancel out). So, the variance of scores in a group around the group mean measures random error. Treatments do not affect within group variance because everybody in the group gets the same treatment.

So, we are trying to find out if the treatment effect (between group variance) is larger than we would expect on the basis of random errors (within group variance).

Source	SS	df	MS	F	p
Between					
Within					
Total					

SS : sum of squared deviations

df : degrees of freedom

MS : mean squared deviations or variance ($\frac{SS}{df}$)

F : F -value ($\frac{MS_{between}}{MS_{within}}$)

p : p -value

Group 1: 11, 25, 22, 17, 20

Group 2: 24, 34, 33, 25, 29

Group 3: 8, 28, 22, 17, 15

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between					
Within					
Total					

SS_{within} – Individual differences or random errors

$$SS_{within} = SS_1 + SS_2 + SS_3$$

$$SS_1 = \sum (x - \bar{x})^2 \text{ for group 1}$$

$$SS_2 = \sum (x - \bar{x})^2 \text{ for group 2}$$

$$SS_3 = \sum (x - \bar{x})^2 \text{ for group 3}$$

Notice $\frac{SS}{df}$ = variance. If you know the variance and it's *df*, you can obtain *SS* ($SS = df \times \text{variance}$).

$$\bar{x}_1 = (11 + 25 + 22 + 17 + 20)/5 = 19$$

$$\bar{x}_2 = (24 + 34 + 33 + 25 + 29)/5 = 29$$

$$\bar{x}_3 = (8 + 28 + 22 + 17 + 15)/5 = 18$$

$$SS_1 = (11 - 19)^2 + (25 - 19)^2 + (22 - 19)^2 + (17 - 19)^2 + (20 - 19)^2 = 114$$

$$SS_2 = (24 - 29)^2 + (34 - 29)^2 + (33 - 29)^2 + (25 - 29)^2 + (29 - 29)^2 = 82$$

$$SS_3 = (8 - 18)^2 + (28 - 18)^2 + (22 - 18)^2 + (17 - 18)^2 + (15 - 18)^2 = 226$$

$$SS_{within} = 114 + 82 + 226 = 422$$

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between					
Within	422				
Total					

Group 1: 11, 25, 22, 17, 20

Group 2: 24, 34, 33, 25, 29

Group 3: 8, 28, 22, 17, 15

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between					
Within	422				
Total					

SS_{between} – Treatment effect

$$SS_{\text{between}} = n_1(\bar{x}_1 - gm)^2 + n_2(\bar{x}_2 - gm)^2 + n_3(\bar{x}_3 - gm)^2$$

n_1 is the number of observations in group 1,

\bar{x}_1 is the mean for group 1,

and gm is the grand mean (the mean for all the observations).

- We are assuming that group 1: 19, 19, 19, 19, 19 where 19 is the mean for group 1, group 2: 29, 29, 29, 29, 29, and group 3: 18, 18, 18, 18, 18. We are assuming that there are no variations within groups.

- Then we do $\sum(x - gm)^2$ using these assumed numbers (19, 19, 19, 19, 19, 29, 29, 29, 29, 29, 18, 18, 18, 18, 18). This is exactly the same as doing $SS_{\text{between}} = n_1(\bar{x}_1 - gm)^2 + n_2(\bar{x}_2 - gm)^2 + n_3(\bar{x}_3 - gm)^2$

$$\bar{x}_1 = (11 + 25 + 22 + 17 + 20)/5 = 19$$

$$\bar{x}_2 = (24 + 34 + 33 + 25 + 29)/5 = 29$$

$$\bar{x}_3 = (8 + 28 + 22 + 17 + 15)/5 = 18$$

$$gm = (11 + 25 + 22 + 17 + 20 + 24 + 34 + 33 + 25 + 29 + 8 + 28 + 22 + 17 + 15)/15 = 22$$

$$SS_{\text{between}} = 5(19 - 22)^2 + 5(29 - 22)^2 + 5(18 - 22)^2 = 370$$

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between	370				
Within	422				
Total					

Group 1: 11, 25, 22, 17, 20

Group 2: 24, 34, 33, 25, 29

Group 3: 8, 28, 22, 17, 15

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between	370				
Within	422				
Total					

SS_{total}

Were we to throw all 15 numbers (11, 25, 22, 17, 20, 24, 34, 33, 25, 29, 8, 28, 22, 17, 15) into a big group and find $\sum(x - gm)^2$, that is SS_{total} and would be the same as the sum of the $SS_{between}$ and SS_{within} .

$$SS_{total} = (11 - 22)^2 + (25 - 22)^2 + (22 - 22)^2 + (17 - 22)^2 + (20 - 22)^2 + (24 - 22)^2 + (34 - 22)^2 + (33 - 22)^2 + (25 - 22)^2 + (29 - 22)^2 + (8 - 22)^2 + (28 - 22)^2 + (22 - 22)^2 + (17 - 22)^2 + (15 - 22)^2 = 792$$

$$SS_{total} = ss_{between} + ss_{within} = 370 + 422 = 792$$

This is a good way to double check you calculated $SS_{between}$ and SS_{within} correctly.

source	<i>ss</i>	<i>df</i>	<i>ms</i>	<i>F</i>	<i>p</i>
between	370				
within	422				
total	792				

df – degrees of freedom

We have 3 groups and an n of 5 in each group. The between group estimate will have $3 - 1 = 2$ degrees of freedom because there are three group means contributing to the estimate. The averaged or pooled within group variance estimate will have a total of 12 degrees of freedom. Why? The variance estimate calculated around each group mean will have $5 - 1 = 4$ *dfs* so the total across the three groups is $4 + 4 + 4 = 4 \times 3 = 12$. Were we to throw all $5 \times 3 = 15$ numbers into a big group and get a variance estimate, it would have $15 - 1 = 14$ *dfs*, which is the same as the sum of the within *dfs* and between *dfs* ($12 + 2 = 14$).

$$df_{between} = \text{number of groups} - 1$$

$$df_{within} = \sum(\text{number of subjects in each group} - 1)$$

$$df_{total} = \text{number of total subjects} - 1 = df_{between} + df_{within}$$

So far we have:

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between	370	2			
Within	422	12			
Total	792	14			

MS – mean squared deviations

$$MS = \frac{SS}{df} \text{ (= variance)}$$

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{370}{2} = 185 \quad MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{422}{12} = 35.17$$

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between	370	2	185		
Within	422	12	35.17		
Total	792	14			

F – *F*-value

$$F(df_{between}, df_{within}) = \frac{ms_{between}}{ms_{within}}$$

$$F(2, 12) = \frac{185}{35.17} = 5.26$$

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between	370	2	185	5.26	
Within	422	12	35.17		
Total	792	14			

p – *p*-value.

You find *p* using the *F*-value and the two *dfs* (*df*_{between} and *df*_{within}) used to calculate *F*.

F(2, 12) at the .05 level is 3.89.

F(2, 12) = 5.26 will have *p*-value less than .05.

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between	370	2	185	5.26	<.05
Within	422	12	35.17		
Total	792	14			

This is our final ANOVA table:

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between	370	2	185	5.26	<.05
Within	422	12	35.17		
Total	792	14			

ANOVA is always one-tail. The probability of obtaining our result assuming that the between group variance (treatment effect) is due to chance is less than 5%. So, we reject the null hypothesis. There is significantly more scatter among the group means than we would expect on the basis of random errors.

You can recreate someone's anova if you know the mean, the standard deviation, and the number of observations in each group. For example, say you are given the followings:

- Group 1: $\bar{x} = 19, s = 5.34, n = 5$
- Group 2: $\bar{x} = 29, s = 4.53, n = 5$
- Group 3: $\bar{x} = 18, s = 7.52, n = 5$

Try to do ANOVA on the above data. You should get roughly the same result as the one we just did. "Roughly" because of round-off errors.

Once you find the variance for each group (variance = s^2), you can variance \times *df* and obtain *SS* for each group. Then you add each *SS* to get *SS_{within}*.

$$SS_{within} = SS_1 + SS_2 + SS_3$$

You find the grand mean. You can simply $\frac{19+29+18}{3}$ because *n* is the same in each group. Or you can $\frac{(19 \times 5) + (29 \times 5) + (18 \times 5)}{5+5+5}$ and this will work even when *n* is different in each group. Once you calculate the grand mean, you can find *SS_{between}*.

$$SS_{between} = n_1(\bar{x}_1 - gm)^2 + n_2(\bar{x}_2 - gm)^2 + n_3(\bar{x}_3 - gm)^2$$

Usually, the mean, standard error, and *n* can be found in a research paper. Then, you can recreate their ANOVA.

$$\begin{aligned} \text{Standard error of the mean} &= \frac{\text{standard deviation}}{\sqrt{n}} \\ \text{Standard deviation} &= \text{standard error} \times \sqrt{n} \end{aligned}$$

12 Correlation

- Correlation is a measure of a linear association between two variables.
- Correlation is a standardized measurement that generates an easily interpretable number ranging from -1 to 1 .
- Correlation is always a measure of the linear relationship between two variables. We can calculate a correlation when our data have no linear association, but the correlation will not be very accurate. So, we need to be careful.

r – Pearson’s correlation coefficient

$$r = \frac{\sum(\frac{x-\bar{x}}{s_x} \frac{y-\bar{y}}{s_y})}{df}$$

s_x : standard deviation of variable x

s_y : standard deviation of variable y

df : number of xy pairs - 1

An example:

Subject	1	2	3	4	5
How many hotdogs eaten	1	3	2	5	4
Stomach discomfort	1	27	8	125	64

What is the correlation (r) between the number of hotdogs eaten and stomach discomfort?

Number of hotdogs eaten could be variable x

Stomach discomfort could be variable y

$$r = \frac{\sum(\frac{x-\bar{x}}{s_x} \frac{y-\bar{y}}{s_y})}{df}$$

Subject	1	2	3	4	5
How many hotdogs eaten	1	3	2	5	4
Stomach discomfort	1	27	8	125	64

Standardize each value in x (number of hotdogs eaten) and y (stomach discomfort):

$$\bar{x} = \frac{1+3+2+5+4}{5} = 3$$

$$s_x^2 = \frac{ss_x}{df_x} = \frac{(1-3)^2+(3-3)^2+(2-3)^2+(5-3)^2+(4-3)^2}{5-1} = \frac{10}{4} = 2.5$$

$$s_x = \sqrt{2.5} = 1.58$$

$$\bar{y} = \frac{1+27+8+125+64}{5} = \frac{225}{5} = 45$$

$$s_y^2 = \frac{ss_y}{df_y} = \frac{(1-45)^2+(27-45)^2+(8-45)^2+(125-45)^2+(64-45)^2}{5-1} = \frac{10390}{4} = 2597.5$$

$$s_y = \sqrt{2597.5} = 50.97$$

Subject	1	2	3	4	5
$\frac{x-\bar{x}}{s_x}$	$\frac{1-3}{1.58}$	$\frac{3-3}{1.58}$	$\frac{2-3}{1.58}$	$\frac{5-3}{1.58}$	$\frac{4-3}{1.58}$
$\frac{y-\bar{y}}{s_y}$	$\frac{1-45}{50.97}$	$\frac{27-45}{50.97}$	$\frac{8-45}{50.97}$	$\frac{125-45}{50.97}$	$\frac{64-45}{50.97}$

Calculate $\frac{x-\bar{x}}{s_x} \frac{y-\bar{y}}{s_y}$:

Subject	1	2	3	4	5
$\frac{x-\bar{x}}{s_x}$	-1.26	0	-0.63	1.26	0.63
$\frac{y-\bar{y}}{s_y}$	-0.86	-0.35	-0.73	1.57	0.37
$\frac{x-\bar{x}}{s_x} \frac{y-\bar{y}}{s_y}$	1.09	0	0.46	1.99	0.24

$$r = \frac{\sum(\frac{x-\bar{x}}{s_x} \frac{y-\bar{y}}{s_y})}{df} = \frac{1.09+0+0.46+1.99+0.24}{5-1} = \frac{3.78}{4} = 0.95$$

A strong positive correlation ($r = 0.95$) between the number of hotdogs eaten and stomach discomfort.

13 Regression

- Linear regression gives us a rule for using one variable to predict another.
- The rule for making predictions can be represented by a line on a graph (regression equation/line).
- The regression equation is called the “line of best fit” because it minimizes the sum of squared deviations between predicted and actual values.

Regression equation

$$y = a + bx$$

y = predicted value

a = the y intercept (the place where the regression line crosses the y axis)

b = the slope of the regression line

x = the predictor value

The slope $b = r(\frac{s_y}{s_x})$. So, we need to find r first. Then, use \bar{x} and \bar{y} to find the intercept a ($a = y - bx$, so $a = \bar{y} - r(\frac{s_y}{s_x})\bar{x}$).

An example:

Subject	1	2	3	4	5
How many hotdogs eaten	1	3	2	5	4
Stomach discomfort	1	27	8	125	64

Find the regression line.

Number of hotdogs eaten could be variable x

Stomach discomfort could be variable y

$$r = \frac{\sum(\frac{x-\bar{x}}{s_x} \frac{y-\bar{y}}{s_y})}{df}$$

Subject	1	2	3	4	5
How many hotdogs eaten	1	3	2	5	4
Stomach discomfort	1	27	8	125	64

Find r :

$$\bar{x} = \frac{1+3+2+5+4}{5} = 3$$

$$s_x = \sqrt{\frac{(1-3)^2+(3-3)^2+(2-3)^2+(5-3)^2+(4-3)^2}{5-1}} = 1.58$$

$$\bar{y} = \frac{1+27+8+125+64}{5} = \frac{225}{5} = 45$$

$$s_y = \sqrt{\frac{(1-45)^2+(27-45)^2+(8-45)^2+(125-45)^2+(64-45)^2}{5-1}} = 50.97$$

$$r = \frac{\sum \left(\frac{x-\bar{x}}{s_x} \frac{y-\bar{y}}{s_y} \right)}{df}$$

$$= \frac{\left(\frac{1-3}{1.58} \right) \left(\frac{1-45}{50.97} \right) + \left(\frac{3-3}{1.58} \right) \left(\frac{27-45}{50.97} \right) + \left(\frac{2-3}{1.58} \right) \left(\frac{8-45}{50.97} \right) + \left(\frac{5-3}{1.58} \right) \left(\frac{125-45}{50.97} \right) + \left(\frac{4-3}{1.58} \right) \left(\frac{64-45}{50.97} \right)}{5-1} = 0.95$$

Find the slope b :

$$b = r \left(\frac{s_y}{s_x} \right) = 0.95 \left(\frac{50.97}{1.58} \right) = 30.65$$

Find the intercept a using \bar{x} and \bar{y} :

$$y = a + bx$$

$$a + bx = y$$

$$a = y - bx$$

The two points we know for x and y are \bar{x} and \bar{y} .

$$a = \bar{y} - b\bar{x} = 45 - (30.65 \times 3) = 45 - 91.95 = -46.95$$

The regression equation is:

$$y = -46.95 + 30.65x$$

14 Chi-Square

When you have count (categorical) data (when data are not continuous), you do a chi-square (χ^2) test. You calculate the number of observations expected in each cell according to the null hypothesis, and compare how well the actual data corresponds to the null hypothesis. χ^2 is always positive \rightarrow it is one-tailed. The critical value for χ^2 is different for different degrees of freedom.

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Degrees of freedom:

One-way χ^2 : $df = n_c - 1$

Two-way χ^2 : $df = (n_c - 1)(n_r - 1)$

where n_c = number of columns and n_r = number of rows

Expected cell count (based on null hypothesis):

One-way χ^2 : total \times probability

Two-way χ^2 : column total $\times \frac{\text{row total}}{\text{grand total}}$ (or row total $\times \frac{\text{column total}}{\text{grand total}}$)

In two-way χ^2 , the null hypothesis is always that there is no relationship between the rows and the columns.

Do not use when expected count is less than 5.

One-way chi-square

In Monty Hall's game show there are three doors A, B, and C. The null hypothesis is that contestants will choose each door with equal frequency. What is the expected cell count if there are 60 contestants?

$$60 \times \frac{1}{3} = 20$$

A	B	C
20	20	20

Observed data:

A	B	C
10	10	40

Do we reject the null hypothesis?

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \\ &= \frac{(10-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(40-20)^2}{20} = 5 + 5 + 20 = 30\end{aligned}$$

$$df = n_c - 1 = 3 - 1 = 2$$

The critical value for χ^2 with 2 df at .05 α is 5.99. The observed χ^2 is 30. The observed counts are significantly different from the expected counts. We reject the null hypothesis that contestants will choose each door with equal frequency.

Two-way chi-square (two independent variables)

Observed data:

	Female	Male
Popularity	14	31
Fitness	7	18
Self-confidence	21	5
Entertainment	25	13

Expected counts:

	Female	Male	Total
Popularity	$67 \times \frac{45}{134}$	$67 \times \frac{45}{134}$	45
Fitness	$67 \times \frac{25}{134}$	$67 \times \frac{25}{134}$	25
Self-confidence	$67 \times \frac{26}{134}$	$67 \times \frac{26}{134}$	26
Entertainment	$67 \times \frac{38}{134}$	$67 \times \frac{38}{134}$	38
Total	67	67	134

	Female	Male	Total
Popularity	22.5	22.5	45
Fitness	12.5	12.5	25
Self-confidence	13	13	26
Entertainment	19	19	38
Total	67	67	134

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} = \frac{(14-22.5)^2}{22.5} + \frac{(7-12.5)^2}{12.5} \\ &+ \frac{(21-13)^2}{13} + \frac{(25-19)^2}{19} + \frac{(31-22.5)^2}{22.5} + \frac{(18-12.5)^2}{12.5} + \frac{(5-13)^2}{13} + \frac{(13-19)^2}{19} \\ &= 3.21 + 2.42 + 4.92 + 1.89 + 3.21 + 2.42 + 4.92 + 1.89 = 24.88 \end{aligned}$$

$$df = (n_c - 1)(n_r - 1) = (2 - 1)(4 - 1) = 3$$

The critical χ^2 value for 3 df at .05 α level is 7.81. The observed χ^2 is 27.88. We reject the null hypothesis that there is no relationship between the rows and the columns.