The z-score, combines a raw score with the mean and standard deviation of a distribution in a way that allows you to know relative standing of the raw score in the distribution.

A z-score description works regardless of the kind of scores you using or the shape of the distribution.

Suppose Student 1 says she got a 95 on a statistics exam. What does that tell you about his statistical ability?

A score gets its meaning from its relation to the mean and the variability of other scores in the distribution.

A z-score is a mathematical way to modify an individual raw score so that the result conveys the score’s relationship to the mean and standard deviation of its fellow scores.

The formula is \( z = \frac{X - \mu}{\sigma} \)

Remember that \( X - \mu \) is an acquaintance of yours, the deviation score. A z-score describes the relation of \( X \) to \( \mu \) with respect to the variability of the distribution.

For instance, if you know that a score \( (X) \) is 5 units from the mean \( (X - \bar{X} = 5) \), you know only that the score is better than average, but you have no idea how far above average it is.
If the distribution has a range of 10 units and $X = 50$, then an $X$ of 55 is a very high score.

On the other hand, if the distribution has a range of 100 units, an $X$ of 55 is barely above average.

To know the score’s position in a distribution, the variability of the distribution must be taken into account (divide $X - \overline{X}$ by a unit that measures the variability, the standard deviation).

The z-score is sometimes referred to as a standard score because it is a deviation score expressed in standard deviation units.

Any distribution of raw scores can be converted into a distribution scores.

For each of raw score, there is one z score.
Positive $z$ scores represent raw scores that are greater than the mean; negative $z$ scores go with raw scores that are less than the mean.

In both cases, the absolute value of the $z$ score tells the number of standard deviations the score is from the mean.

Converting a raw score to a $z$ score gives you a number that indicates the raw score’s relative position in the distribution.

If two raw scores are converted to $z$ scores, you will know their positions relative to each other as well as to the distribution.

---

$Z$ scores are also used to compare two scores from different distributions, even when the scores are measuring different things.

Let say, one of your lecturer returned tests with a $z$-score rather than a percentage score. This $z$ score was the key to figuring out your grade. A $z$ score of +1.50 or higher was an A, and -1.50 or lower was an F.

<table>
<thead>
<tr>
<th>Student</th>
<th>Raw Score</th>
<th>$Z$-score</th>
<th>Raw Score</th>
<th>$Z$-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>76</td>
<td>2.20</td>
<td>76</td>
<td>-1.67</td>
</tr>
<tr>
<td>S2</td>
<td>51</td>
<td>.00</td>
<td>86</td>
<td>.00</td>
</tr>
<tr>
<td>S3</td>
<td>58</td>
<td>+0.40</td>
<td>82</td>
<td>-0.67</td>
</tr>
<tr>
<td>S4</td>
<td>58</td>
<td>+0.40</td>
<td>90</td>
<td>+0.67</td>
</tr>
<tr>
<td>Mean, $X = 54$</td>
<td>Mean, $X = 86$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma = 10$</td>
<td>$\sigma = 6$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Z scores give you a way to compare raw scores. The basis of the comparison is the distribution itself rather than some external standard (such as grading scale of 90-80-70-60 percent for As, Bs, and so on).

Z is used as both descriptive statistic and an inferential statistic. As a descriptive statistic, its range is limited.

For a distribution of 100 or so scores, the z scores might range from approximately -3 to +3.

The formula to determine a raw score from a z score is $X = \mu + z\sigma$.

For example, what is raw score value corresponds to a z-score of $z = -2.00$ if $\mu = 60$ and $\sigma = 5$?

From the formula,

$X = \mu + z\sigma$

$= 60 + (-2)(5)$

$= 60 - 10 = 50$

For many distributions, especially when $N$ is small, the range is more narrow.

As an inferential statistic, z value are not limited to ±3. The value of z depends heavily on how different the two populations actually are.

Thus, values much greater than 3 occur when z is used as an inferential statistic.
For a normal distribution with particular values of \( \mu \) and \( \sigma \), the probability falling within \( z \) standard deviations of the mean depends on the value of \( z \).

This is the area under the bell-shaped curve between \( \mu - z\sigma \) and \( \mu + z\sigma \).

For every normal distribution, this probability is .68 for \( z = 1 \), .95 for \( z = 2 \), and nearly 1.0 for \( z = 3 \).

For instance, adult female height has the normal distribution with mean, \( \mu = 65.0 \) and standard deviation, \( \sigma = 3.5 \) inches.

The probability equals .95 that a randomly selected female has height between \( \mu - 2\sigma = 65.0 - 2(3.5) = 58 \) inches and \( \mu + 2\sigma = 65.0 + 2(3.5) = 72 \) inches.

Adult male height has a normal distribution with mean, \( \mu = 70 \) and standard deviation, \( \sigma = 4.0 \) inches, so the probability equals .95 that a randomly selected male has height between \( \mu - 2\sigma = 70 - 2(4.0) = 62 \) inches and \( \mu + 2\sigma = 70 + 2(4.0) = 78 \) inches.
A continuous variable graphically described by a certain bell-shaped curve is said to have the normal probability distribution.

A normal distribution is a bell shaped, symmetrical, theoretical distribution based on a mathematical formula rather than on any empirical distributions.

Each normal distribution has two parameters – the mean, μ and the standard deviation, σ.

The mean, the median, and the mode are the same score – the score on the X axis where the curve peaks.

If a line is drawn from the peak to the mean score on the X axis, the area under the curve to the left of the line is half the total area – 50 percent – leaving half the area to the right of the line.

The tails of the curve are asymptotic to the X axis – it is convenient to think of (and to draw) the curve as extending from -3σ to +3σ.
The theoretical normal distribution is used to determine the probability of an event.

**Example of finding the proportion of a population that has scores of a particular size or greater.**

Suppose you are faced with finding out what proportion of the population has an IQ of 120 or higher.

1. Begin by sketching a normal curve.
2. Note on the baseline the positions of IQs of 100 and 120.
3. Calculate the z score, $z = \frac{X - \mu}{\sigma} = \frac{120 - 100}{15} = \frac{20}{15} = 1.33$

4. From the Table, the proportion beyond $z = 1.33$ is 0.918 or 9.18 percent of the population to have an IQ of 120 or higher.
Example of finding the score that separates the population into two proportions.

What IQ score is required to be in the top 10 percent of the population?

1. Begin by sketching a normal curve.
2. Note on the baseline the positions of IQs of 100.
3. Separate the top 10 percent portion with a vertical line.
4. From the Table, look in the column area beyond z for .1000. You have a choice between .0985 and .1003 (and .1003 (z = 1.28) closer to the desired .1000).
5. Calculate the score,

\[ X = \mu + z\sigma = 100 + (1.28)(15) = 119 \]

Example of finding the proportion of a population between two scores.

What proportion of the population falls in the range what proportion of the population has an IQ in the range from 90 to 110.

1. We need z scores that correspond to the IQ scores of 90 and 110.

\[ z = \frac{90 - 100}{15} = -\frac{10}{15} = -0.67 \]

\[ z = \frac{110 - 100}{15} = \frac{10}{15} = 0.67 \]

The proportion of the distribution between the mean and z = 0.67 is 0.2486, and, of course, the same proportion is between the mean and z = -0.67. Therefore, (2)(0.2486) = 0.4972 or 49.72%.
Example of finding the extreme scores in a population.

What two heart rates (beats per minute) separate the middle 95 percent of the population from the extreme 5 percent?

(According to study by Milnor, the mean heart rate for human is 71 beats per minute and the s.d. is 9 beats per minute).

1. Use the formula $X = \mu + (z)(\sigma)$. Substituting the values given, plus the z score you looked up, you get:

<table>
<thead>
<tr>
<th>Upper score (2.5%)</th>
<th>Lower score (2.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 71 + (1.96)(9)$</td>
<td>$X = 71 - (1.96)(9)$</td>
</tr>
<tr>
<td>$= 71 + 17.6$</td>
<td>$= 71 - 17.6$</td>
</tr>
<tr>
<td>$= 88.6$</td>
<td>$= 53.4$</td>
</tr>
</tbody>
</table>

Thus, 95 percent of the population is expected to have pulse rates between 53 and 89, which leaves 5 percent of the population above 89 and below 53.

If we have a set of measurement scores on different measures using z-scores we can tell how the scores are placed in their distribution.

Knowing the mean and standard deviation of a distribution allows us to compare and determine the percentile rank of that distribution score.

Z-scores act like a ruler; placed data on a standard scale.

Z-scores express the distance of a single score from the mean in standard deviation units.

If data are normally distributed, z-scores to identify or cut off specific percentages/area of distribution.
By now, you should be able to describe a set of data by using a graph, a few choice words, and numbers such as mean, and a standard deviation.

The other statistical technique is inferential statistics.

With inferential statistics, you use measurements from a sample to reach conclusion about a larger, unmeasured population.

Inferential statistics is a method that takes chance factors into account when samples are used to reach conclusions about populations.

Inferential statistics was developed as a way to measure and reduce the effects of chance that come with sampling.

Inferential procedures are typically built around the concept of probability - the relationships between samples and populations are usually defined in terms of probability.
The probability of a particular outcome is the proportion (or fraction) of times that outcome would occur in a long run of repeated observations.

For instance, if the possible outcomes are identified as A, B, C, D, and so on, then

\[
\text{Probability of A} = \frac{\text{number of outcomes classified as A}}{\text{total number of possible outcomes}}
\]

For example, with coin flipping, the probability of head faces upwards is \( \frac{1}{2} \) or .50 = 50%.

A head is a success and it can occur in only one way.

The total number of possible outcomes is two (head and tail).

The probability values range from .00 (there is no possibility that an event will occur) to 1.00 (the event is certain to happen) – as a proportion or a number between 0 and 100 when expressed as percentage.
Similarly, if the weather forecaster says that the probability of rain today is 70% - this means that in long series of days with atmospheric conditions like those today, rain occurs on 70% of the days.

Notation - the probability, $p$, followed by the specific outcome in parentheses such as the probability of selecting spade is $p$ (spade) = \( \frac{13}{52} = \frac{1}{4} = 0.25 = 25\% \).

A variable can take at least two different values. The probability distribution of the variable lists the possible outcomes together with their probabilities.
The central limit theorem always applies when the sample size is adequate and we know $\sigma$.

Usually, we often do not know $\sigma$ and we don’t have population data to calculate it.

What to do if we don’t have $\sigma$?

W. S. Gosset (1876 – 1937) solution was to work out the mathematics of distribution based on $\bar{s}$.

Gosset found that the distribution depended on the sample size, with a different distribution for each $N$.

These distributions make up a family of curves that have come to be called the $t$ distribution.

The different curves that make up the $t$ distribution are distinguished from one another by their degree of freedom.

The $t$ distribution is bell shaped, like the normal curve, but less peaked and with fatter tails.
Degree of freedom (df) range from 1 to \( \infty \).

Knowing the df for your data tells you which \( t \) distribution to use.

From Figure 1, their df are 2, 9, and \( \infty \).

We can see that as the df increase, less and less of the curve is in the tails.

Note that the smaller the df, the larger the value needed for the rejection (critical region).
Look at Table D. The first column shows df, ranging from 1 to ∞.

Across the top are three rows; the row we use depends on the kind of problem we are using the \( t \) distribution for.

Each column is associated with a percent or probability.

The body of the table contains \( t \) values.

Remember, that hypothesis testing is one of the most important concepts in Statistics.

This is how we decide if:

a. Effect actually occurred
b. Treatment have effects.
c. Groups differ from each other.
d. One variable predicts another.

There are TWO possibilities:
NOTHING happened (Null hypothesis) or
SOMETHING happened (Alternative hypothesis).
When we test a statistical hypothesis, we are trying to see if something happened and are comparing against the possibility that nothing happened.

For example:
We were told that the average person sleeps 8 hours a day. We want to know whether university students sleep less. We survey 10 students to see how much they sleep. Our data as follows:

<table>
<thead>
<tr>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

The statistics we need to decide whether to reject or fail to reject the Null hypothesis is the ONE-SAMPLE t TEST.

This statistics compares the sample mean to the population mean and gets an estimate of the probability that the sample mean is different by chance.

The widely accepted solution to the problem of where to break the continuum is use the .05 mark (or 95% sure – degree of certainty).

0.01 – 99 percent sure; 0.001 – 99.9 percent sure.
Hypothesis testing with two samples is similar to hypothesis testing with one sample.

The same hypothesis-testing reasoning will be used, but you will have data from **two samples or two group experiment**.

In a well-designed, well-executed experiment, all imaginable results are included in this statement: Either Treatment A has an effect or it does not have an effect.

Let’s making a tentative assumption that Treatment A does not have an effect.

Gather data.

Using a sampling distribution based on the assumption that Treatment A has no effect, find the probability of the data obtained.

If the probability is low, abandon your tentative assumption and conclude that Treatment A has an effect.

If the probability is not low, you are back where you began: The data have not given you evidence against either of the two logical possibilities.
In the language of experiment:

Begin with two logical possibilities, the null hypothesis, $H_0$, and alternative hypothesis, $H_1$.

$H_0$: Treatment A does not have an effect; that is, the mean of the population scores of those who receive Treatment A is equal to the mean of the population of scores of those who do not receive Treatment A.

The difference between population means is zero. In statistical language:

$H_0: \mu_A - \mu_{\text{no A}} = 0 \quad \text{or} \quad H_0: \mu_A = \mu_{\text{no A}}$

$H_1$: Treatment A does have an effect; that is, the mean of the population of scores of those who receive Treatment A is not equal to the mean of the population of scores of those who do not receive Treatment A.

The alternative hypothesis, which should be chosen before the data are gathered, can be two-tailed or one-tailed.

An alternative hypothesis consists of one of the three $H_1$’s that follow.

Two-tailed alternative: $H_1: \mu_A \neq \mu_{\text{no A}}$
In the example of the sample experiment, this hypothesis says that Treatment A has an effect, but it does not indicate whether the treatment improves or disrupts performance on task Q.

A two-tailed test allows either conclusion.

One-tailed alternative:  
\[ H_1: \mu_A > \mu_{\text{no } A} \]  
\[ H_1: \mu_A < \mu_{\text{no } A} \]

The first two-tailed alternative hypothesis allows you to conclude that Treatment A increases scores.

However, no outcome of the experiment can lead to the conclusion that Treatment A produces lower scores than no treatment.

The second alternative hypothesis permits a conclusion that Treatment A reduces scores, but not that it increases scores.

2. Tentatively assume that Treatment A has no effect (that is, assume \( H_0 \)). If \( H_0 \) is true, the two samples will be alike except for the usual chance variation in samples.

3. Decide on a level \( \alpha \) level. (Usually, \( \alpha = .05 \).)

4. Choose an appropriate inferential statistical test.

This test will have
The *t* test is a traditional way to analyze the results from an experiment that is designed to compare two treatments.

How you are going to analyze an experiment that has more than two treatments?

The technique is called the analysis of variance (ANOVA) – also a hypothesis testing technique, and a new effect size will be explained.

ANOVA is based on the very versatile concept of assigning the variability in the data to various sources.

The simplest ANOVA design is called one-way ANOVA – the one in one-way means that one independent variable is being analyzed.

One-way ANOVA is used to find out if there are many differences among three or more populations means.

For examples:
Four groups learned a task. A different schedule of reinforcement was used for each group.
Afterwards, response persistence was measured.
A common first reaction to the task of determining if there is a difference among three or more population means is to run \( t \) tests on all possible pairs of sample means.

For the three populations (example earlier), three \( t \) tests would be required (\( \bar{X}_1 \) vs \( \bar{X}_2 \), \( \bar{X}_1 \) vs \( \bar{X}_3 \), and \( \bar{X}_2 \) vs \( \bar{X}_3 \)).

For four populations, six tests will needed.

This multiple \( t \) test approach will not work and doing a lot of \( t \) tests – is tedious.

Here is the reasoning: Suppose you have 15 samples all drawn from the same population.

These 15 sample means will vary from one another as a result of chance factors.

Now, let say we run every possible \( t \) test (all 105 of them), retaining or rejecting each null hypothesis at the .05 level.

How many times would we reject the null hypothesis??? – about 5 times (100 \( t \) tests will produce about five Type I errors).
We can protect ourselves from disaster if we use the statistical technique that keeps the overall risk of a Type I error at one acceptable level (such as .05 or .01).

Sir Ronald A. Fisher (1890 – 1962), a biologist and statistician, developed a technique called ANOVA and its sampling distribution, the $F$ distribution.

The question that ANOVA addresses is whether the populations that the samples come from have the same $\mu$.

If the answer is no, then at least one of the populations has a different $\mu$.

The $F$ distribution, like the $t$ distribution, is a sampling distribution.

The $F$ value that is calculated from the data is a ratio; one estimate of the population variance is divided by a second estimate.

The $F$ ratio can be described

$$F = \frac{\text{variation between treatment means}}{\text{variation within treatments}}$$
The F distribution is a sampling distribution that shows the probability of various ratios of variances when the null hypothesis is true.

Each variance has a number of degrees of freedom associated with it – Table F gives us critical values of F for α levels of .05 and .01.

There are number of similarities between the F distribution and the t distribution.

Like t, the F distribution is also a family of curves, with each member of the family characterized by different degrees of freedom.

A particular F distribution depends on the degrees of freedom in both numerator estimate of $\sigma^2$ and the denominator estimate of $\sigma^2$.

As the number of degrees of freedom approaches infinity, the F curve approaches a normal distribution, which is true for t as well.

The mathematical relationship between t and F is $t^2 = F$ (for an F with 1 df in the numerator).
Sum of squares (SS).
As you recall, the numerator of the basic formula for the standard deviation \( \Sigma(X - \bar{X})^2 \) is the sum of squares.
So, \( SS = \Sigma(X - \bar{X})^2 \).
A more descriptive name for sum of squares is sum of the squared deviations.

Mean square (MS) is the ANOVA term for variance \( s^2 \). The mean square is a sum of squares (SS) divided by its degrees of freedom (df).

Grand mean is the mean of all scores; it is computed without regard for the fact that the scores come from different groups (samples).

tot (a subscript) after a symbol means that the symbol stands for all such numbers in the experiment; \( \Sigma X_{tot} \) is the sum of all X scores.

t (a subscript) after symbol means – applies to the treatment group; for example, \( \Sigma(\Sigma X_t)^2 \) tells us to sum the scores in each group, square each sum, and then sum these squared value.
\( N_t \) is the number of scores in one treatment group.
\( K \) is the number of treatments in the experiment. This is the same as the number of levels of the independent variable.
It is time to do the calculations.

Table 1 shows fictional data from an experiment in which there were three treatments for patients with psychological disorders.

<table>
<thead>
<tr>
<th>Drug A</th>
<th>Drug B</th>
<th>Drug C</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>$\bar{x}_A$ = 7.25</td>
<td>$\bar{x}_B$ = 6.75</td>
<td>$\bar{x}_C$ = 2.25</td>
</tr>
<tr>
<td>$\sum x_{A}$ = 21</td>
<td>$\sum x_{B}$ = 19</td>
<td>$\sum x_{C}$ = 9</td>
</tr>
</tbody>
</table>

In this experiment (and in all one-way ANOVA), the null hypothesis is that the populations are identical. In terms of treatment means in the drug experiment, $H_0: \mu_A = \mu_B = \mu_C$. 

The I.V. was drug therapy. Three levels of the I.V. were used - Drug A, Drug B, and Drug C.

The D.V. was the number of psychotic episodes each patient had during the therapy period.
In this particular experiment, the researcher believed that Drug C would be better than Drug A and Drug B, which were in common use – (in this case better means fewer psychotic episodes).

The $H_1$ in ANOVA is that one or more populations are different from the others. (No greater or less than is specified.)

The next step in an ANOVA is to calculate the sum of squares, which is one measures of variability.

\[ SS_{\text{tot}} = \sum X^2 - \left( \frac{\sum X}{N} \right)^2 = 427.7 - \frac{(40)^2}{12} = 84.917 \]

\[ SS_{\text{between}} = \sum \left[ \frac{(\sum X)^2}{N_t} \right] - \frac{1}{N} \left( \sum X \right)^2 \]
\[ = \frac{332 + 237 + 497}{4} - \frac{1}{12} \left( \frac{40}{2} \right)^2 \]
\[ = 111 + 2.7875 + 20.375 - 332.083 = 60.667 \]

\[ SS_{\text{within}} = \sum \left[ X^2 - \left( \frac{\sum X}{N_t} \right)^2 \right] \]
\[ = (319 - \frac{29^2}{4}) + (191 - \frac{29^2}{4}) + (27 - \frac{9^2}{4}) \]
\[ = 8.750 + 8.750 + 6.750 = 24.250 \]

\[ SS_{\text{tot}} = SS_{\text{between}} + SS_{\text{within}} = 60.667 + 24.250 = 84.917 \]
After you find SS, the next step is to find the mean squares.

A mean square is simply a sum of squares divided by its degrees of freedom.

It is an estimate of the population variance, $\sigma^2$, when the null hypothesis is true.

Each sum of squares has a particular number of degrees of freedom associated with it.

In one-way classification, the df are $df_{total}$, $df_{treat}$, and $df_{error}$.

The relationship among degrees of freedom is the same as that among sums of squares:

$$df_{total} = df_{treat} + df_{error}$$

where:

$$df_{total} = N_{total} - 1$$
$$df_{treat} = K - 1$$
$$df_{error} = N_{total} - K$$

Mean squares can be found using the following formulas:

$$MS_{treat} = \frac{SS_{treat}}{df_{treat}}$$
The next two steps in an ANOVA are to calculate an F value and to interpret it using the F distribution.

An F test consists of dividing $MS_{treat}$ by $MS_{error}$ to obtain an F value.

$$F = \frac{MS_{treat}}{MS_{error}}$$

Every F value has two degrees of freedom associated with it.

The first is the df associated with $MS_{treat}$ (the numerator) and the second is the df associated with $MS_{error}$ (the denominator).

For the data in Table 1,

$$F = \frac{MS_{treat}}{MS_{error}} = \frac{30.333}{2.694} = 11.26; \text{ df } = 2, 9$$

Table 2 shows summary of ANOVA for data in Table 1

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>60.667</td>
<td>2</td>
<td>30.333</td>
<td>11.26</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Error</td>
<td>24.250</td>
<td>9</td>
<td>2.694</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>84.917</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(a) a test statistic that can be calculated from the data,
(b) a sampling distribution of the test statistic that shows its distribution when H₀ is true, and
(c) a critical value for the α level.

5. Calculate a test statistic using the sample data.

6. Compare the test statistic to the critical value from the sampling distribution. If the test statistic’s probability is less than α, reject H₀.

If larger, retain H₀. (For the t distribution, if the data-based t value is greater than the critical value in the table, reject H₀. If the data-based t value is less than the critical value, retain H₀.)

7. Write a conclusion that uses the terms of the experiment.

Your conclusion should describe how the levels of the independent variable differ on the dependent variable in the populations the samples came from.

In your use of the t distribution, you have determining degrees of freedom by rule of thumb techniques:

\[ N - 1 \] when you determine whether a sample mean \( \bar{X} \) comes from a population with a mean \( \mu_0 \), and \[ N - 2 \] when you determine whether a correlation coefficient \( r \) is significantly different from .00.
The ‘freedom’ refers to the freedom of a number to have any possible value.

If you are asked to pick two numbers and there is no restrictions, both numbers are free to vary (take any value) and you have two degrees of freedom.

If a restriction is imposed – say, \( \Sigma X = 0 \) – then one degree of freedom is lost because of that restriction; that is, when you now pick the two numbers, only one of them is free to vary.

For example, if you choose 3 for the first number, the second number must be -3.

In similar way, if you are to pick five numbers with a restriction that \( \Sigma X = 0 \), you have four degrees of freedom.

When you found \( s_{\bar{X}} \), as required in the formula for the one-sample t test, used

\[
s_{\bar{X}} = \frac{s}{\sqrt{N}} = \sqrt{\frac{\Sigma (X - \bar{X})^2}{N-1}}
\]

The restriction that is built is that \( \Sigma (X - \bar{X}) \) is always zero and, in order to meet that requirement, one of the X’s is determined.

The degree of freedom for \( s_{\bar{X}} \) is \( N - 1 \).
An experiment with two groups can be either a correlated-samples design or an independent-samples design.

You must decide what kind of design you have before you analyze the data.

These different design require different t test formulas.

The difference can be tell by knowing whether scores in one group are paired with scores in a second group.

The three types of correlated-samples designs are natural pairs, matched pairs, and repeated measures.

In independent-samples design, a t test is used to decide whether two populations have the same mean.

The null hypothesis is

\[ H_0: \mu_1 = \mu_2 \]

where the subscripts 1 and 2 are assigned arbitrarily to the two populations.
If this null hypothesis is true, any difference between the two sample means is due to chance.

The task is to establish an $\alpha$ level, calculate a $t$ test value, and compare that value with a critical value of $t$ in Table D.

If the $t$ value calculated from the data is greater than the critical value (that is, less probable than $\alpha$), reject $H_0$ and conclude that the two samples came from populations with different means.

If the data-based $t$ value is not large as critical value, retain $H_0$.

For an independent-samples design, the formula for $t$ test

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

The term $s_{\bar{x}_1 - \bar{x}_2}$ is the standard error of a difference.
The $t$ Test for Independent Sample Designs

When a $t$ test is used to decide whether two populations have the same mean, the null hypothesis is

$$H_0: \mu_1 = \mu_2;$$

where the subscripts 1 and 2 are assigned arbitrarily to the two populations.

The task is to establish the alpha level, calculate a $t$ test value, and compare the value with a critical value of $t$ in Table D.

The formula for the $t$ test is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

where; $s_{\overline{X}_1 - \overline{X}_2}$ is the standard error of difference (the s.d. of a sampling distribution of differences between means).

Use this formula when the two samples have an equal number of scores ($N_1 = N_2$).

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{\sum X_1^2}{N_1} + \sum X_2^2 - \frac{(\sum X_1)^2}{N_1}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$
If $N_1 = N_2$, use this formula:

$$\sqrt{\frac{\sum x_1^2 - \frac{(\sum x_1)^2}{n_1} + \sum x_2^2 - \frac{(\sum x_2)^2}{n_2}}{n_1 (n_2 - 1)}}$$

The following example shows the use of an independent-samples design to evaluate the effect of a drug on learning a complex problem solving task.

The drug group (7 monkeys) received pills, while the other group (6 monkeys) was given an inert substance (a placebo).

Pills and training on the task continued for six days.

The total number of errors that each monkey made was the dependent variable.

$H_0$ is the drug has no effect on errors.

A two-tailed test is called for because the investigator is interested in any effect the drug has, either to enhance or to inhibit performance.

The number of errors each monkey made as follows:

<table>
<thead>
<tr>
<th>Drug Group, $X_1$</th>
<th>Placebo Group, $X_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>37</td>
</tr>
<tr>
<td>52</td>
<td>57</td>
</tr>
<tr>
<td>26</td>
<td>68</td>
</tr>
<tr>
<td>47</td>
<td>74</td>
</tr>
<tr>
<td>41</td>
<td>49</td>
</tr>
<tr>
<td>37</td>
<td>57</td>
</tr>
<tr>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>
Applying the $t$ test (because the $N$’s are unequal for the two samples, the $N1\#N2$ formula is used).

\[
t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sum x_1^2 - (\sum x_1)^2}{N_1} + \frac{\sum x_2^2 - (\sum x_2)^2}{N_2}}} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}
\]

\[
= \frac{-1.92}{\sqrt{\frac{17}{10}}}
\approx -2.99
\]

The $t$ value for these data is -2.99, a negative value.

(For a two-tailed test, always use the absolute value of $t$).

To evaluate a data-based $t$ value of 2.99 with 11 $df$, you need a critical value from Table D.

The critical value in the column for a two-tailed test with $\alpha = .05$ is 2.201.

That is $t.05(11) = 2.201$. Thus the null hypothesis can be rejected.
Because 2.99 is also greater than critical value at the .02 level (2.718), the results would usually be reported as “significant at the .02 level.”

The final step is to interpret the results.

“The drug group average fewer errors than the placebo group (37.71 vs 57.33). Thus, the drug facilitated learning (p < .02).”

Exercise.

A psychologist doing research on how divorce influence a children behavior. The hypothesis that he developed was “Children from divorced parents have an aggressive behavior.” To test this hypothesis, the researcher observed 17 children from divorced parents and 15 children from happy family (live together.)

<table>
<thead>
<tr>
<th>Divorced Family</th>
<th>Happy Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 8</td>
<td>N = 6</td>
</tr>
<tr>
<td>$\sum X = 320$</td>
<td>$\sum X = 210$</td>
</tr>
<tr>
<td>$\sum X^2 = 15800$</td>
<td>$\sum X^2 = 6690$</td>
</tr>
<tr>
<td>$\bar{X}_1 = 40$</td>
<td>$\bar{X}_2 = 37$</td>
</tr>
</tbody>
</table>
The t Test For Correlated Sample Designs (Natural Pairs, Matched Pairs, Or Repeated Measures).

Fortunately, the actual arithmetic of calculating a t test value is the same for all three.

The formula has a familiar theme: a difference between means divided by the standard error of a difference.

The standard error of a difference between means of correlated samples is symbolized $s_D$.

The formula is

$$t = \frac{\bar{D}}{s_D}$$

where

$$s_D = \sqrt{\frac{\sum D^2 - (\sum D)^2}{N}}^{\frac{1}{N-1}}$$

$df = N - 1$, where $N$ is number of pairs

The algebraically equivalent formula for $s_D$ does not require you to calculate $r$ as follows:

$$s_D = \frac{\bar{D}}{\sqrt{N}}$$

where

$\bar{D} = \frac{\sum D}{N}$

$N = \text{number of pairs}$
Exercise
Suppose you are interested in the effects of interracial contact on racial attitudes. You administer the test one Monday morning to a biracial group of 14 12 year-old girls who do not know one another but who have signed up for a week-long camp. The campers then spend the next week taking nature walks, playing ball, eating lunch, swimming, making things, and doing the kinds of things that camp directors dream up to keep 12 year-old girls busy. On Saturday morning the girls again are given the racial attitude test. Thus, the data consist of 14 pairs of before-and-after scores.

The null hypothesis is that the mean of the population of ‘after’ scores is equal to the mean of the population ‘before’ scores or, in terms of the specific experiment, that a week of interracial contact has no effect on racial attitudes.

<table>
<thead>
<tr>
<th></th>
<th>before</th>
<th>after</th>
<th>Δ</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kyle</td>
<td>54</td>
<td>38</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>Marsha</td>
<td>22</td>
<td>19</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>Didi</td>
<td>25</td>
<td>36</td>
<td>-11</td>
<td>121</td>
</tr>
<tr>
<td>Max</td>
<td>31</td>
<td>40</td>
<td>-9</td>
<td>81</td>
</tr>
<tr>
<td>Shana</td>
<td>27</td>
<td>36</td>
<td>-9</td>
<td>81</td>
</tr>
<tr>
<td>Navia</td>
<td>32</td>
<td>31</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Simon</td>
<td>38</td>
<td>43</td>
<td>-5</td>
<td>26</td>
</tr>
<tr>
<td>Martin</td>
<td>57</td>
<td>56</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Ward</td>
<td>30</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nate</td>
<td>24</td>
<td>31</td>
<td>-7</td>
<td>25</td>
</tr>
<tr>
<td>Pauline</td>
<td>16</td>
<td>34</td>
<td>-18</td>
<td>724</td>
</tr>
<tr>
<td>Cheryl</td>
<td>24</td>
<td>31</td>
<td>-7</td>
<td>49</td>
</tr>
<tr>
<td>Susan</td>
<td>26</td>
<td>30</td>
<td>-10</td>
<td>100</td>
</tr>
<tr>
<td>Anne</td>
<td>29</td>
<td>37</td>
<td>-8</td>
<td>44</td>
</tr>
</tbody>
</table>
Using the sums of the \( D \) and \( D^2 \) column in the last table, find \( s_{\bar{D}}^2 \):

\[
\bar{D}^2 = \frac{\sum D^2 - (\sum D)^2}{N - 1} = \sqrt{897 - (\frac{283.6}{14})^2} = \sqrt{12.95} \approx 3.6
\]

Thus,

\[
\bar{D}^2 = \frac{283.6}{N} = \frac{283.6}{14} = 20.26
\]

\[
t = \frac{\bar{D} - \bar{Y}}{s_{\bar{D}}} = \frac{283.6 - 14.14}{1.83} = \frac{-3.78}{1.83} \approx -2.06
\]

\( df = N - 1 = 14 - 1 = 13 \)

Because \( t.01(13 \text{ df}) = 3.012 \), a \( t \) value of 3.78 is significant beyond the .01 level; that is, \( p < .01 \).

Because the ‘after’ mean is greater than the ‘before’ mean, conclude that racial attitudes were significantly more positive after camp than before.
In an experiment, the researcher sets $\alpha$, gathers data, and uses a sampling distribution to find the probability ($p$) of such data.

This $p$ value is correct only when the null hypothesis is true.

If $p < \alpha$, reject $H_0$. Thus, if $\alpha = .05$ and $p \leq .05$, reject $H_0$.

If $\alpha = .05$ and $p > .05$, retain $H_0$. Of course, if $p = .03$, or .01, or .001, reject $H_0$.

If the probability is .051 or greater, retain $H_0$.

When $H_0$ is rejected, the difference is described as statistically significant.

When $H_0$ is retained - the difference is not significant.

The rejection region is the area of the sampling distribution that includes all the differences that have a probability equal to or less than $\alpha$ - any event in the rejection region leads to rejection the null hypothesis.
Most statistical tables do not provide us with probability figures for every outcome of a statistical test.

What they provide are statistical test values that correspond to commonly chosen $\alpha$ values. These specific test values are called critical values.

If we have a $df$ that is not in the table, it is conventional to use the next smaller $df$ or to interpolate a $t$ value that corresponds to the exact $df$.

The $t$ values in the table separate the rejection region from the rest of the sampling distribution. Data produced $t$ values that are equal to or greater than the critical value fall in the rejection region, and the null hypothesis is rejected.

$t_{.05} (14 \ df) = 2.145$.

This expression indicates the sampling distribution that was used ($t$), $\alpha$ level (.05), $df$ (14), and critical value from the table (2.145, for a two-tailed test).
The formula for the one-sample t test is:

\[ t = \frac{\bar{X} - \mu_0}{s_\bar{X}}; \quad df = N - 1 \]

Where \( \bar{X} \) is the mean of the sample
\( \mu_0 \) is the hypothesized mean of the population
\( s_\bar{X} \) is the standard error of the mean

Table 1 shows the two ways to make mistake.

<table>
<thead>
<tr>
<th>( H_0 ) true</th>
<th>( H_0 ) false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject Ho</td>
<td>Type I error</td>
</tr>
<tr>
<td>Retain Ho</td>
<td>Correct decision</td>
</tr>
</tbody>
</table>
Hypothesis testing answers the question of whether the population a sample is drawn from is different from the population specified by the null hypothesis.

If the answer is YES, another question would ask, “How big is the difference?”

The effect size index, $d$, will determine just how different two sample means were.

The formula for $d$ that corresponds to a one-sample t test is:

$$d = \frac{\bar{X} - \mu_0}{\sigma}$$

Where $\bar{X}$ is the mean of the sample
$\mu_0$ is the mean specified by the null hypothesis
$\sigma$ is the standard deviation of the null hypothesis population

Cohen guidelines that help evaluate the absolute size of $d$:
small effect, $d = .20$;
medium effect, $d = .50$; and
large effect, $d = .80$. 