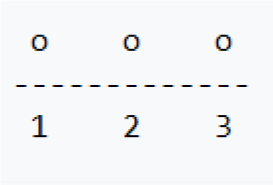
## Probability Mass Function (PMF)

Suppose the probability distribution of a discrete random variable *X* puts equal weights on 1, 2, and 3:

$$\boldsymbol{X = \begin{cases} 1 & \text{with probability } 1/3, \\ 2 & \text{with probability } 1/3, \\ 3 & \text{with probability } 1/3. \end{cases}}$$

The probability mass function of the random variable *X* may be depicted by the following bar graph:



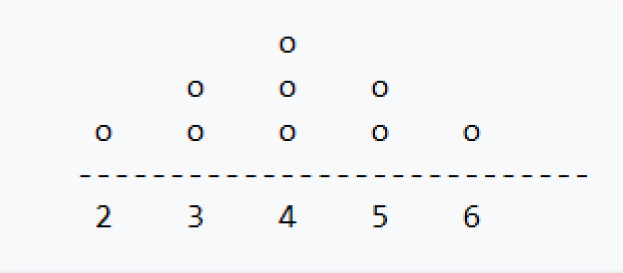
```
 o    o    o
-------------
 1    2    3
```

Clearly this looks nothing like the bell-shaped curve of the normal distribution. Contrast the above with the depictions below.

Now consider the sum of two independent copies of *X*:

$$\boldsymbol{\left\{\begin{matrix} 1+1 & = & 2 \\ 1+2 & = & 3 \\ 1+3 & = & 4 \\ 2+1 & = & 3 \\ 2+2 & = & 4 \\ 2+3 & = & 5 \\ 3+1 & = & 4 \\ 3+2 & = & 5 \\ 3+3 & = & 6 \end{matrix}\right\} = \left\{\begin{matrix} 2 & \text{with probability } 1/9 \\ 3 & \text{with probability } 2/9 \\ 4 & \text{with probability } 3/9 \\ 5 & \text{with probability } 2/9 \\ 6 & \text{with probability } 1/9 \end{matrix}\right\}}$$

The probability mass function of this sum may be depicted thus:



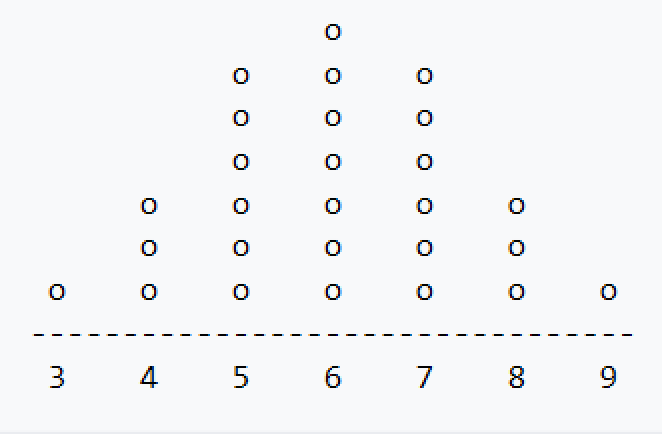
```
          o
     o    o    o
o    o    o    o    o
----------------------------
2    3    4    5    6
```

## Probability mass function of the sum of three terms

Now consider the sum of *three* independent copies of this random variable:

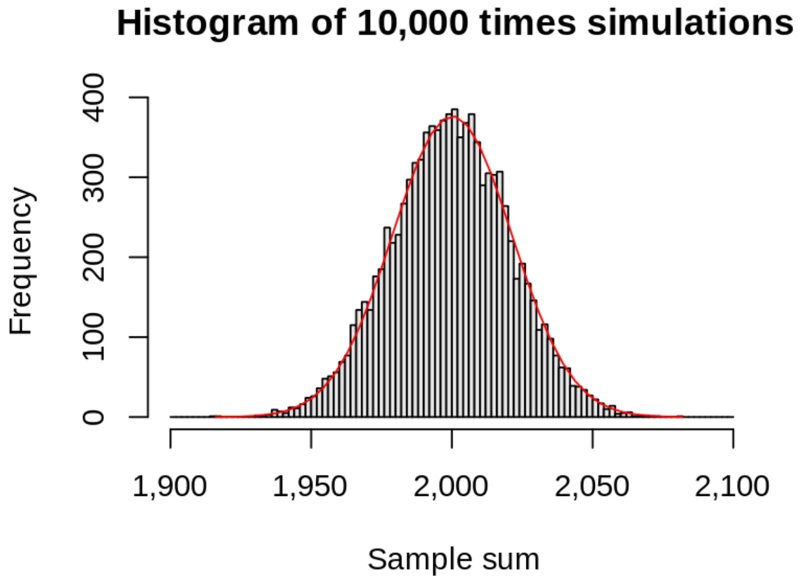
$$\left\{\begin{array}{lcl} 1+1+1 & = & 3 \\ 1+1+2 & = & 4 \\ 1+1+3 & = & 5 \\ 1+2+1 & = & 4 \\ 1+2+2 & = & 5 \\ 1+2+3 & = & 6 \\ 1+3+1 & = & 5 \\ 1+3+2 & = & 6 \\ 1+3+3 & = & 7 \\ 2+1+1 & = & 4 \\ 2+1+2 & = & 5 \\ 2+1+3 & = & 6 \\ 2+2+1 & = & 5 \\ 2+2+2 & = & 6 \\ 2+2+3 & = & 7 \\ 2+3+1 & = & 6 \\ 2+3+2 & = & 7 \\ 2+3+3 & = & 8 \\ 3+1+1 & = & 5 \\ 3+1+2 & = & 6 \\ 3+1+3 & = & 7 \\ 3+2+1 & = & 6 \\ 3+2+2 & = & 7 \\ 3+2+3 & = & 8 \\ 3+3+1 & = & 7 \\ 3+3+2 & = & 8 \\ 3+3+3 & = & 9 \end{array}\right\} = \left\{\begin{array}{ll} 3 & \text{with probability } 1/27 \\ 4 & \text{with probability } 3/27 \\ 5 & \text{with probability } 6/27 \\ 6 & \text{with probability } 7/27 \\ 7 & \text{with probability } 6/27 \\ 8 & \text{with probability } 3/27 \\ 9 & \text{with probability } 1/27 \end{array}\right\}$$

The probability mass function of this sum may be depicted thus:



```
               o
          o    o    o
          o    o    o
          o    o    o
     o    o    o    o    o
     o    o    o    o    o
o    o    o    o    o    o    o
-----------------------------------
3    4    5    6    7    8    9
```

## Probability mass function of the sum of 1,000 terms

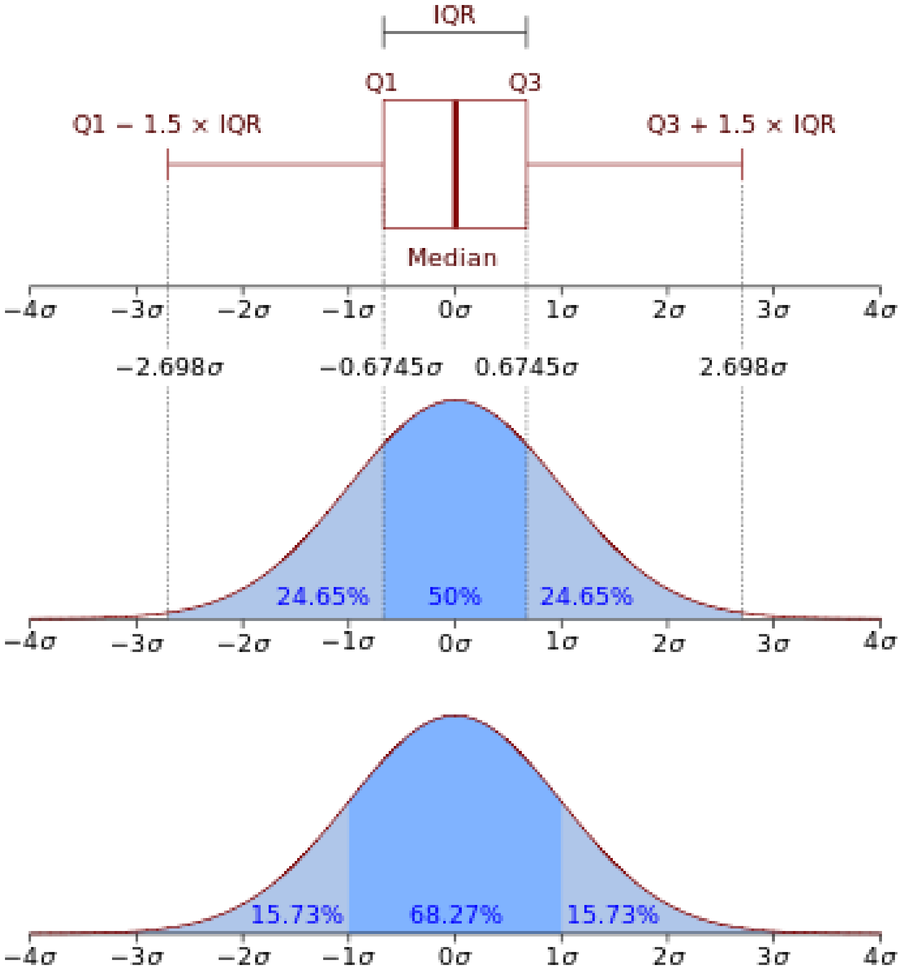


*A distribution is simply a collection of data, or scores, on a variable. Usually, these scores are arranged in order from smallest to largest and then they can be presented graphically.*

## Probability Density Function (PDF)

PDF, in statistics, a function whose integral is calculated to find probabilities associated with a continuous random variable.

Every random variable is associated with a probability density function (e.g., a variable with a normal distribution is described by a bell curve).



**Covariance** provides insight into how two variables are related to one another.

- covariance refers to the measure of how two random variables in a data set will change together.
- A positive covariance means that the two variables at hand are positively related, and they move in the same direction.
- A negative covariance means that the variables are inversely related, or that they move in opposite directions.

$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Covariance can tell how the stocks move together, but to determine the strength of the relationship,

The **correlation** should, therefore, be used in conjunction with the covariance, and is represented by this equation:

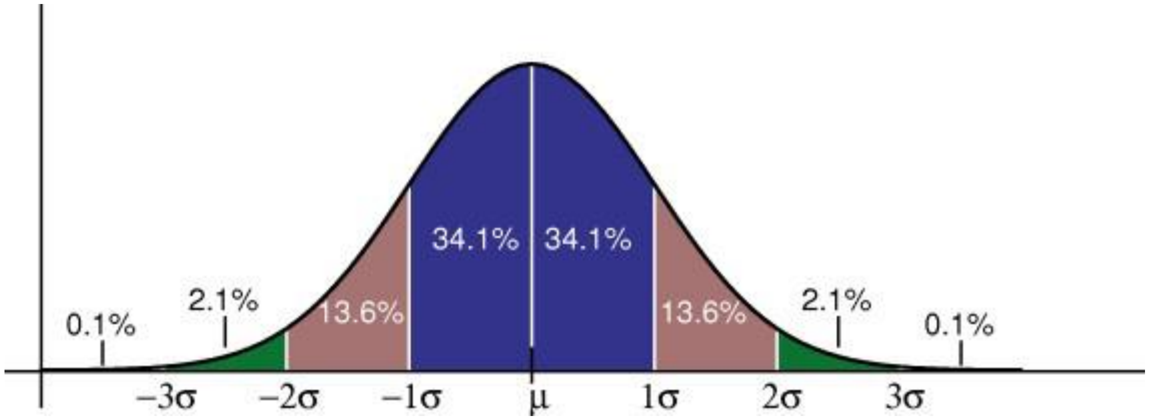
$$\text{Correlation} = \rho = \frac{cov\,(X, Y)}{\sigma_X \sigma_Y}$$

- The equation above reveals that the correlation between two variables is the covariance between both variables divided by the product of the standard deviation of the variables.
- While both measures reveal whether two variables are positively or inversely related, the correlation provides additional information by determining the degree to which both variables move together. The correlation will always have a measurement value between -1 and 1.
- If the correlation is 1, they move perfectly together, and if the correlation is -1, the stocks move perfectly in opposite directions. If the correlation is 0, neutral.

A **normal distribution**, sometimes called the bell curve, is a distribution that occurs naturally in many situations.

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

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



The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

## Z Score Formula:

The basic z score formula for a sample is:

**z = (x – μ) / σ**

For example, let's say you have a test score of 190. The test has a mean (μ) of 150 and a standard deviation (σ) of 25.

Assuming a normal distribution, your z score would be:

z = (x – μ) / σ

= 190 – 150 / 25 = 1.6.

The z score tells you how many standard deviations from the mean your score is. In this example, your score is 1.6 standard deviations *above* the mean.

$$z_i = \frac{x_i - \bar{x}}{s}$$

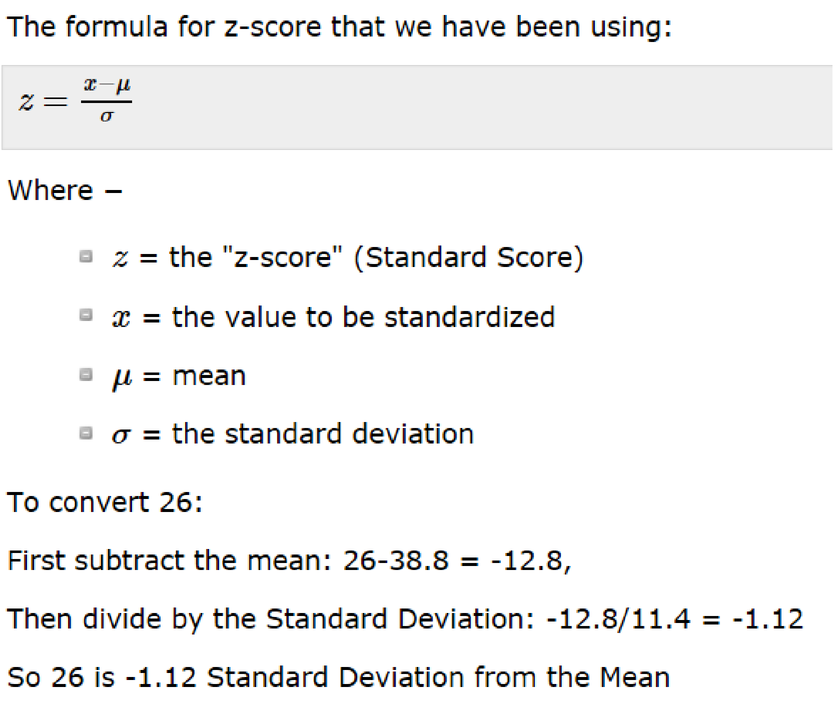
**Problem Statement:**

A survey of daily travel time had these results (in minutes):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | 33 | 65 | 28 | 34 | 55 | 25 | 44 | 50 | 36 | 26 | 37 | 43 | 62 | 35 | 38 | 45 | 32 | 28 | 34 |

The Mean is 38.8 minutes, and the Standard Deviation is 11.4 minutes. Convert the values to z - scores and prepare the Normal Distribution Graph.

**Solution:**



The formula for z-score that we have been using:

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

Where –

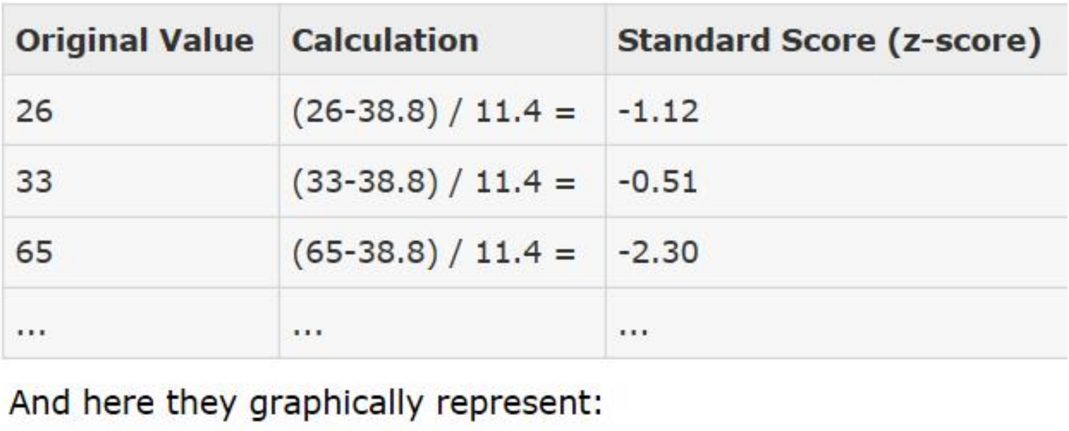
- $z$ = the "z-score" (Standard Score)
- $x$ = the value to be standardized
- $\mu$ = mean
- $\sigma$ = the standard deviation

To convert 26:

First subtract the mean: 26-38.8 = -12.8,

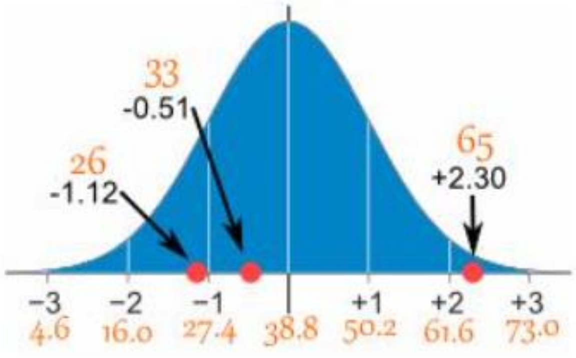
Then divide by the Standard Deviation: -12.8/11.4 = -1.12

So 26 is -1.12 Standard Deviation from the Mean

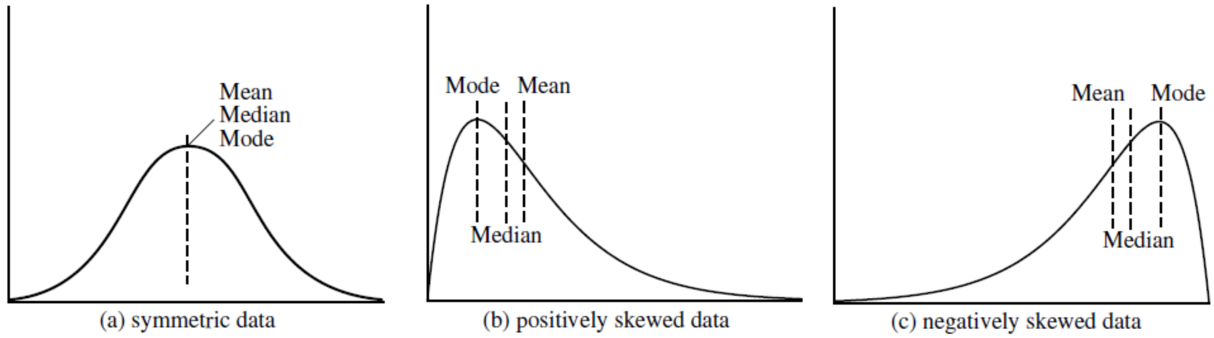


| Original Value | Calculation | Standard Score (z-score) |
|---|---|---|
| 26 | (26-38.8) / 11.4 = | -1.12 |
| 33 | (33-38.8) / 11.4 = | -0.51 |
| 65 | (65-38.8) / 11.4 = | -2.30 |
| ... | ... | ... |

And here they graphically represent:



Skewness



## Sampling methods

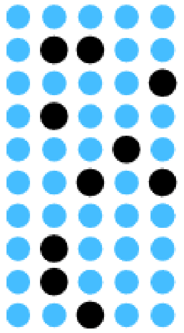
In a statistical study, sampling methods refer to how we select members from the population to be in the study.

If a sample isn't randomly selected, it will probably be biased in some way and the data may not be representative of the population.

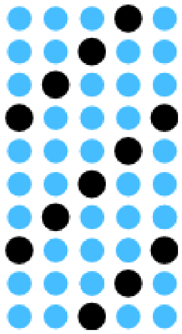
**Simple random sample:** Every member and set of members has an equal chance of being included in the sample.

Technology, random number generators, or some other sort of chance process is needed to get a simple random sample.

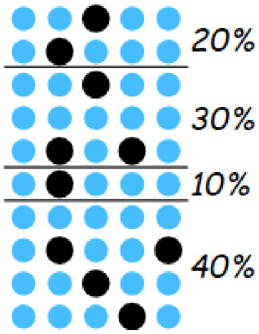
Random samples are usually representative since they don't favor certain members.



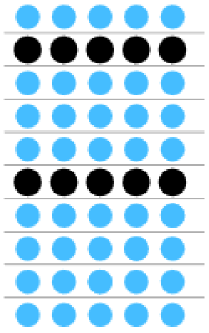
**Random Sample** (pick randomly from list)



**Systematic Sample** (such as every 4th)



**Stratified Sample** (randomly, but in ratio to group size)

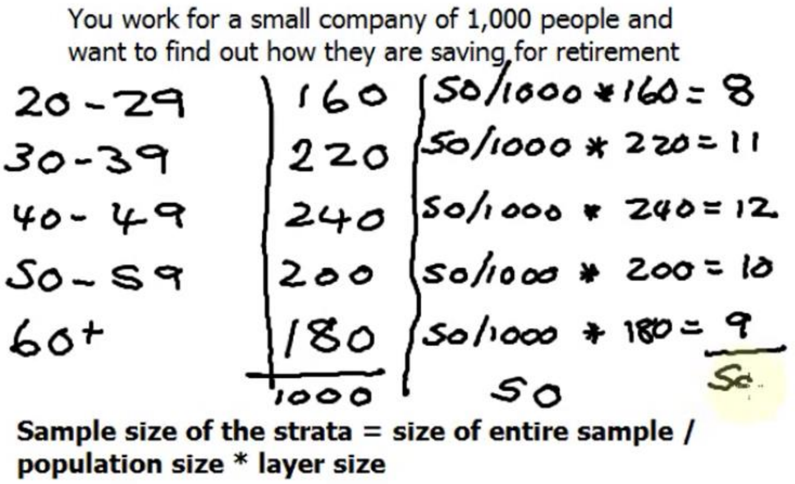


**Cluster Sample** (choose whole groups randomly)

**Stratified random sample:** The population is first split into groups.

The overall sample consists of some members from every group. The members from each group are chosen randomly.

A stratified sample guarantees that members from each group will be represented in the sample, so this sampling method is good when we want some members from every group.



You work for a small company of 1,000 people and want to find out how they are saving for retirement

| | | |
|---|---|---|
| 20-29 | 160 | 50/1000 * 160 = 8 |
| 30-39 | 220 | 50/1000 * 220 = 11 |
| 40-49 | 240 | 50/1000 * 240 = 12 |
| 50-59 | 200 | 50/1000 * 200 = 10 |
| 60+ | 180 | 50/1000 * 180 = 9 |
| | 1000 | 50    50 |

**Sample size of the strata = size of entire sample / population size * layer size**

**Cluster random sample:** The population is first split into groups. The overall sample consists of every member from some of the groups. The groups are selected at random.

A cluster sample gets every member from some of the groups, so it's good when each group reflects the population.

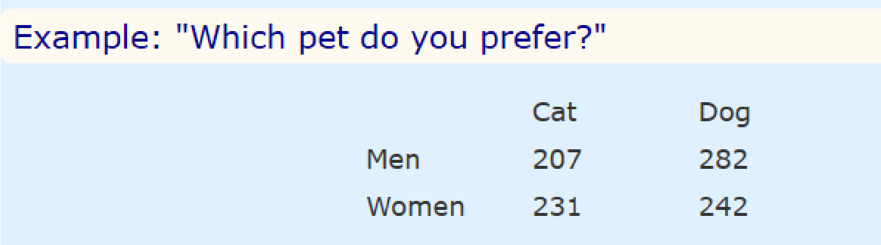
**Systematic random sample:** Members of the population are put in some order. A starting point is selected at random, and every nth, member is selected to be in the sample.

## Chi-Square Test

This is the formula for Chi-Square:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- O = the **Observed** (actual) value
- E = the **Expected** value



Example: "Which pet do you prefer?"

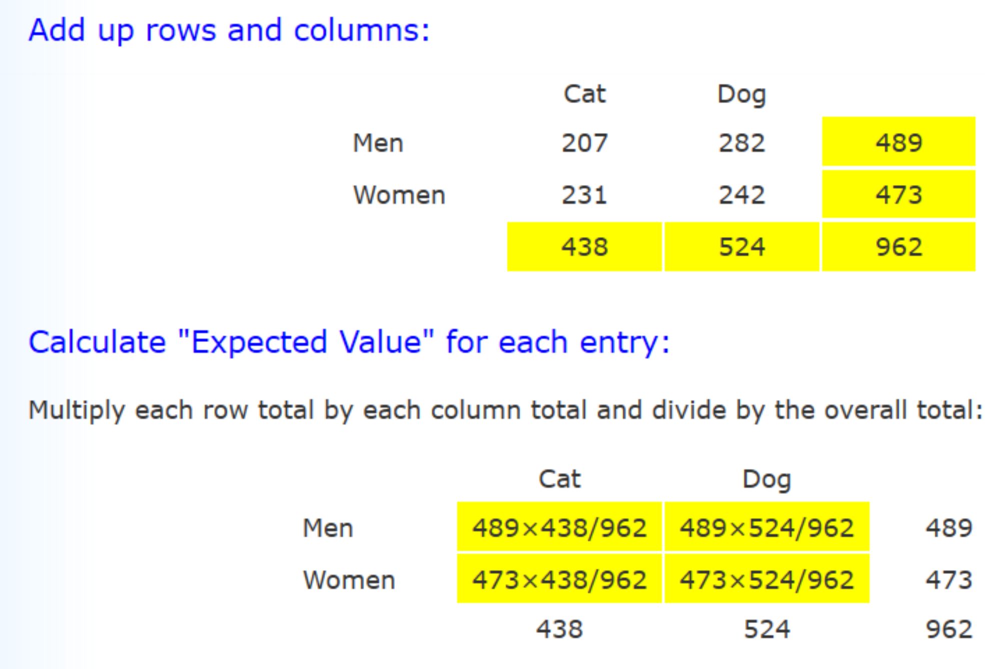
| | Cat | Dog |
|---|---|---|
| Men | 207 | 282 |
| Women | 231 | 242 |

This test only works for **categorical** data (data in categories), such as Gender {Men, Women} or color {Red, Yellow, Green, Blue} etc, but **not numerical** data such as height or weight.

**Hypothesis**: A statement that might be true, which can then be tested.

The two **hypotheses** are.

- Gender and preference for cats or dogs are **independent**.
- Gender and preference for cats or dogs are **not independent**.



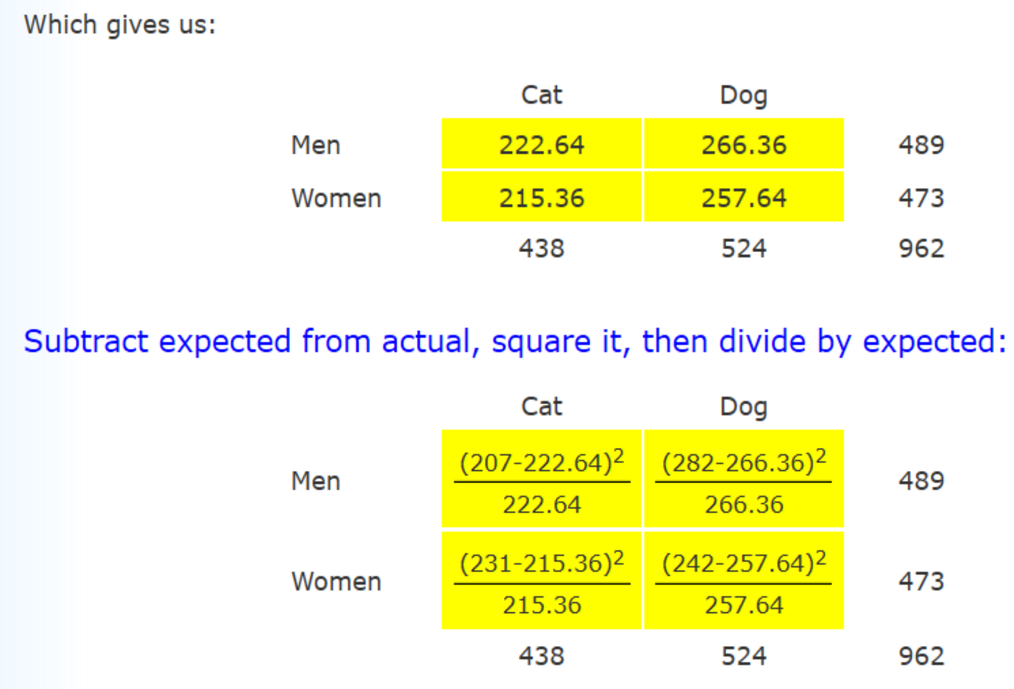
### Add up rows and columns:

| | Cat | Dog | |
|---|---|---|---|
| Men | 207 | 282 | 489 |
| Women | 231 | 242 | 473 |
| | 438 | 524 | 962 |

### Calculate "Expected Value" for each entry:

Multiply each row total by each column total and divide by the overall total:

| | Cat | Dog | |
|---|---|---|---|
| Men | 489×438/962 | 489×524/962 | 489 |
| Women | 473×438/962 | 473×524/962 | 473 |
| | 438 | 524 | 962 |

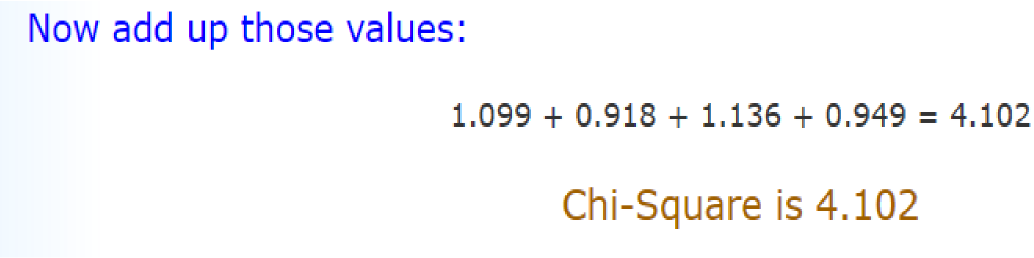


Which gives us:

| | Cat | Dog | |
|---|---|---|---|
| Men | 222.64 | 266.36 | 489 |
| Women | 215.36 | 257.64 | 473 |
| | 438 | 524 | 962 |

### Subtract expected from actual, square it, then divide by expected:

| | Cat | Dog | |
|---|---|---|---|
| Men | $\frac{(207-222.64)^2}{222.64}$ | $\frac{(282-266.36)^2}{266.36}$ | 489 |
| Women | $\frac{(231-215.36)^2}{215.36}$ | $\frac{(242-257.64)^2}{257.64}$ | 473 |
| | 438 | 524 | 962 |



### Now add up those values:

$$1.099 + 0.918 + 1.136 + 0.949 = 4.102$$

Chi-Square is 4.102

## Calculate Degrees of Freedom

Multiply (rows − 1) by (columns − 1)

Example: DF = (2 − 1)(2 − 1) = 1×1 = 1

## Result

The result is:

$$p = 0.04283$$

In this case $p < 0.05$, so this result is thought of as being "significant" meaning is that the variables are not independent.

In other words, because $0.043 < 0.05$ so Gender is linked to Pet Preference (Men and Women have different preferences for Cats and Dogs).

## Statistical Hypothesis Tests

Data must be interpreted in order to add meaning.

We can interpret data by assuming a specific structure our outcome and use statistical methods to confirm or reject the assumption.

The assumption is called a hypothesis and the statistical tests used for this purpose are called statistical hypothesis tests.

The assumption of a statistical test is called the null hypothesis, or hypothesis zero (H0 for short).

It is often called the default assumption, or the assumption that nothing has changed.

A violation of the test's assumption is often called the first hypothesis, hypothesis one, or H1 for short.

- **Hypothesis 0 (H0)**: Assumption of the test holds and is failed to be rejected.
- **Hypothesis 1 (H1)**: Assumption of the test does not hold and is rejected at some level of significance.

We can interpret the result of a statistical hypothesis test using a p-value.

The p-value is the probability of observing the data, given the null hypothesis is true.

A large probability means that the H0 or default assumption is likely.

A small value, such as below 5% (0.05) suggests that it is not likely and that we can reject H0 in favor of H1, or that something is likely to be different (e.g. a significant result).

A widely used statistical hypothesis test is the Student's t-test for comparing the mean values from two independent samples.

The default assumption is that there is no difference between the samples, whereas a rejection of this assumption suggests some significant difference. The tests assume that both samples were drawn from a Gaussian distribution and have the same variance.

The Student's t-Test is a statistical hypothesis test for testing whether two samples are expected to have been drawn from the same population.

The test works by checking the means from two samples to see if they are significantly different from each other. It does this by calculating the standard error in the difference between means, which can be interpreted to see how likely the difference is, if the two samples have the same mean (the null hypothesis).

The t statistic calculated by the test can be interpreted by comparing it to critical values from the t-distribution. The critical value can be calculated using the degrees of freedom and a significance level with the percent point function (PPF).

We can interpret the statistic value in a two-tailed test, meaning that if we reject the null hypothesis, it could be because the first mean is smaller or greater than the second mean. To do this, we can calculate the absolute value of the test statistic and compare it to the positive (right tailed) critical value, as follows:

- **If abs(t-statistic) <= critical value**: Accept null hypothesis that the means are equal.
- **If abs(t-statistic) > critical value**: Reject the null hypothesis that the means are equal.

We can also retrieve the cumulative probability of observing the absolute value of the t-statistic using the cumulative distribution function (CDF) of the t-distribution in order to calculate a p-value. The p-value can then be compared to a chosen significance level (alpha) such as 0.05 to determine if the null hypothesis can be rejected:

- **If p > alpha**: Accept null hypothesis that the means are equal.
- **If p <= alpha**: Reject null hypothesis that the means are equal.

There are two main versions of Student’s t-test:

- **Independent Samples**. The case where the two samples are unrelated.
- **Dependent Samples**. The case where the samples are related, such as repeated measures on the same population. Also called a paired test.

### Student’s t-Test for Independent Samples

The calculation of the t-statistic for two independent samples is as follows:

t = observed difference between sample means / standard error of the difference between the means
or
t = (mean(X1) - mean(X2)) / sed

Where *X1* and *X2* are the first and second data samples and *sed* is the standard error of the difference between the means.

The standard error of the difference between the means can be calculated as follows:

sed = sqrt(se1^2 + se2^2)

Where *se1* and *se2* are the standard errors for the first and second datasets.

The standard error of a sample can be calculated as:

```
se = std / sqrt(n)
```

These calculations make the following assumptions:

- The samples are drawn from a Gaussian distribution.
- The size of each sample is approximately equal.
- The samples have the same variance.

## Student’s t-Test for Dependent Samples

This is the case where we collect some observations on a sample from the population, then apply some treatment, and then collect observations from the same sample.

The result is two samples of the same size where the observations in each sample are related or paired.

The t-test for dependent samples is referred to as the paired Student’s t-test.

The main difference is in the calculation of the denominator.

```
t = (mean(X1) - mean(X2)) / sed
```

Where *X1* and *X2* are the first and second data samples and *sed* is the standard error of the difference between the means.

Here, *sed* is calculated as:

```
sed = sd / sqrt(n)
```

Where *sd* is the standard deviation of the difference between the dependent sample means and *n* is the total number of paired observations

The calculation of *sd* first requires the calculation of the sum of the squared differences between the samples:

```
d1 = sum (X1[i] - X2[i])^2 for i in n
```

It also requires the sum of the (non squared) differences between the samples:

```
d2 = sum (X1[i] - X2[i]) for i in n
```

We can then calculate sd as:

```
sd = sqrt((d1 - (d2**2 / n)) / (n - 1))
```

## Confidence Interval

A confidence interval is an interval that will contain a population parameter a specified proportion of the time. The confidence interval can take any number of probabilities, with the most common being 95% or 99%.

A confidence interval is how much uncertainty there is with any particular statistic. Confidence intervals are often used with a margin of error.

### Calculating the Confidence Interval

**Step 1**: find the number of observations **n**, calculate their mean $\boxed{X}$, and standard deviation **s**

example:

- Number of observations: **n = 40**
- Mean: $\boxed{\mathbf{X}}$ **= 175**
- Standard Deviation: **s = 20**

**Step 2**: decide what Confidence Interval we want: 95% or 99% are common choices. Then find the "Z" value for that Confidence Interval here:

| **Confidence Interval** | **Z** |
|---|---|
| 80% | 1.282 |
| 85% | 1.440 |
| 90% | 1.645 |
| 95% | 1.960 |
| 99% | 2.576 |
| 99.5% | 2.807 |
| 99.9% | 3.291 |

For 95% the Z value is **1.960**

**Step 3**: use that Z in this formula for the Confidence Interval

$$\overline{X} \ \pm \ Z\frac{s}{\sqrt{n}}$$

Where:

- $\boxed{\mathbf{X}}$ is the mean
- **Z** is the chosen Z-value from the table above
- **s** is the standard deviation
- **n** is the number of observations

And we have:

175 ± 1.960 × *20*√**40**

Which is:

**175cm ± 6.20cm**

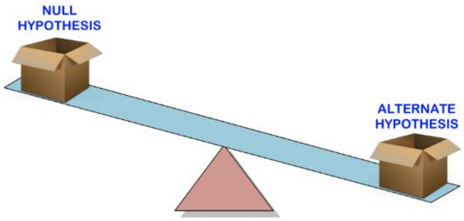
In other words: from 168.8cm to 181.2cm

The value after the ± is called the **margin of error**

The margin of error in our example is 6.20cm

## P Value

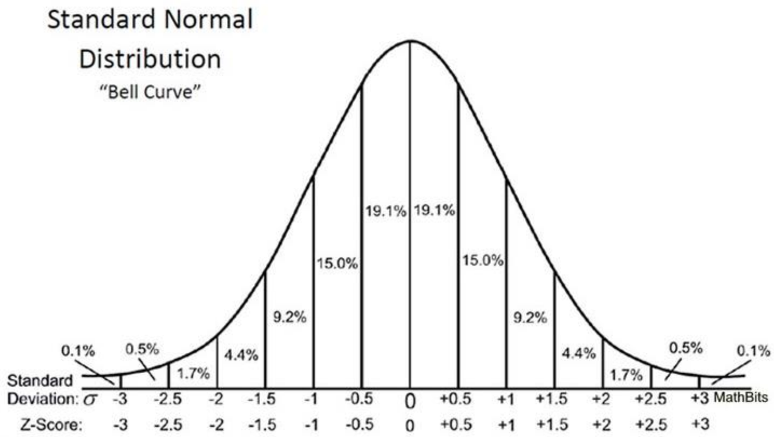
In statistical hypothesis testing, the ***p*-value** or **probability value** is, for a given statistical model, the probability that, when the null hypothesis is true, the statistical summary (such as the absolute value of the sample mean difference between two compared groups) would be greater than or equal to the actual observed results.



**Example:** Suppose a pizza place claims their delivery times are 30 minutes or less on average but you think it’s more than that. So you conduct a hypothesis test and randomly sample some delivery times to test the claim:

- **Null hypothesis** — The mean delivery time is 30 minutes or less
- **Alternative hypothesis** — The mean delivery time is greater than 30 minutes

If the p-value is lower than a predetermined **significance level** (*alpha*, *threshold*), then we reject the null hypothesis.



**Anova**

1. Calculate the sample means for each of our samples as well as the mean for all the sample data.
2. Calculate the sum of squares of error.
   - Here within each sample, we square the deviation of each data value from the sample mean. The sum of all the squared deviations is the sum of squares of error, abbreviated SSE.
3. Calculate the sum of squares of treatment. We square the deviation of each sample mean from the overall mean. The sum of all these squared deviations

is multiplied by one less than the number of samples we have. This number is the sum of squares of treatment, abbreviated SST.

4. Calculate the degrees of freedom. The overall number of degrees of freedom is one less than the total number of data points in our sample, or *n* - 1. The number of degrees of freedom of treatment is one less than the number of samples used, or *m* - 1. The number of degrees of freedom of error is the total number of data points, minus the number of samples, or *n* - *m*.
5. Calculate the mean square of error. This is denoted MSE = SSE/(*n* - *m*).

1. Calculate the mean square of treatment. This is denoted MST = SST/*m* - `1.
2. Calculate the *F* statistic. This is the ratio of the two mean squares that we calculated. So, *F* = MST/MSE.

Suppose we have **four independent populations** that satisfy the conditions for single factor ANOVA. We wish to test the null hypothesis $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. For purposes of this example, we will use a sample of size three from each of the populations being studied.

The data from our samples is:

- Sample from population #1: 12, 9, 12. This has a sample mean of 11.
- Sample from population #2: 7, 10, 13. This has a sample mean of 10.
- Sample from population #3: 5, 8, 11. This has a sample mean of 8.
- Sample from population #4: 5, 8, 8. This has a sample mean of 7.

The mean of all the data is 9.

**Sum of Squares of Error**

We now calculate the sum of the squared deviations from each sample mean. This is called the sum of squares of error.

- For the sample from population #1: $(12 – 11)^2 + (9– 11)^2 +(12 – 11)^2 = 6$
- For the sample from population #2: $(7 – 10)^2 + (10– 10)^2 +(13 – 10)^2 = 18$
- For the sample from population #3: $(5 – 8)^2 + (8 – 8)^2 +(11 – 8)^2 = 18$
- For the sample from population #4: $(5 – 7)^2 + (8 – 7)^2 +(8 – 7)^2 = 6$.

We then add all these sums of squared deviations and obtain 6 + 18 + 18 + 6 = 48.

**Sum of Squares of Treatment**

Here we look at the squared deviations of each sample mean from the overall mean, and multiply this number by one less than the number of populations:

$3[(11 – 9)^2 + (10 – 9)^2 +(8 – 9)^2 + (7 – 9)^2] = 3[4 + 1 + 1 + 4] = 30$.

**Degrees of Freedom**

There are 12 data values and four samples. Thus, the number of degrees of freedom of treatment is 4 – 1 = 3. The number of degrees of freedom of error is 12 – 4 = 8.

**Mean Squares**

We now divide our sum of squares by the appropriate number of degrees of freedom in order to obtain the mean squares.

- The mean square for treatment is 30 / 3 = 10.
- The mean square for error is 48 / 8 = 6.

**The F-statistic**

The final step of this is to divide the mean square for treatment by the mean square for error. This is the F-statistic from the data. Thus, for our example F = 10/6 = 5/3 = 1.667.