## Frequency distribution

* **Frequency** is how often something occurs.
* By counting frequencies we can make a **Frequency Distribution** table.
* <https://www.mathsisfun.com/data/frequency-distribution.html>

### Example: Newspapers

These are the numbers of newspapers sold at a local shop over the last 10 days:

22, 20, 18, 23, 20, 25, 22, 20, 18, 20

Let us count how many of each number there is:

|  |  |
| --- | --- |
| **Papers Sold** | **Frequency** |
| 18 | 2 |
| 19 | 0 |
| 20 | 4 |
| 21 | 0 |
| 22 | 2 |
| 23 | 1 |
| 24 | 0 |
| 25 | 1 |

## Frequency Distribution: values and their frequency (how often each value occurs).

## Chi-square

he **Chi-Square Test** gives a "p" value to help you decide!

### Example: "Which holiday do you prefer?"

|  |  |  |
| --- | --- | --- |
|  | Beach | Cruise |
| Men | 209 | 280 |
| Women | 225 | 248 |

### Does Gender affect Preferred Holiday?

If Gender (Man or Woman) **does** affect Preferred Holiday we say they are **dependent**.

By doing some special calculations (explained later), we come up with a "p" value:

p value is 0.132

Now, **p < 0.05** is the usual test for dependence. In this case **p is greater than 0.05**, so we believe the variables are **independent** (ie not linked together).

In other words Men and Women probably do **not** have a different preference for Beach Holidays or Cruises.

## Understanding "p" Value

"p" is the [probability](https://www.mathsisfun.com/data/probability.html) the variables are **independent**.

## Why p<0.05 ?

It is just a choice!**Using p<0.05 is common**, but we could have chosen p<0.01 to be even more sure that the groups behave differently, or any value really.

## Calculating P-Value

So how do we calculate this p-value? We use the Chi-Square Test!

## Chi-Square Test

Note:**Chi**Sounds like "Hi" but with a**K**, so say Chi-Square like "**Ki**square"

And Chi is the greek letter**Χ**, so we can also write it**Χ2**

Important points before we get started:

* 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.
* The numbers must be large enough. Each entry must be**5**or more. In our example we have values such as 209, 282, etc, so we are good to go.

### Our first step is to state our hypotheses:

**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**.

## Chi-Square Formula

This is the formula for Chi-Square:

chi square formula chisquare = sum (O-E)^2 / E

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

## Kurtosis / skew

# Skewed Data

Data can be "skewed", meaning it tends to have a **long tail** on one side or the other:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| data skewed left |  | data no skew |  | data skewed right |
| Negative Skew |  | No Skew |  | Positive Skew |

|  |  |  |
| --- | --- | --- |
|  |  | Negative Skew? Why is it called **negative** skew? Because the long "tail" is on the negative side of the peak.  People sometimes say it is "skewed to the left" (the long tail is on the left hand side)  The mean is also on the left of the peak. |

|  |  |
| --- | --- |
| The Normal Distribution has No Skew A [Normal Distribution](https://www.mathsisfun.com/data/standard-normal-distribution.html) is not skewed.  It is perfectly symmetrical.  And the Mean is exactly at the peak. |  |

## Central Tendency

When you have two or more numbers it is nice to find a value for the "center".

## Variability

# Variance

[more ...](https://www.mathsisfun.com/data/standard-deviation.html)

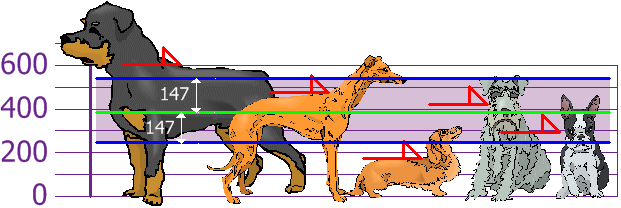
The Variance is a measure of how spread out numbers are  
  
It is the average of the squared differences from the Mean.

# Standard Deviation

[more ...](https://www.mathsisfun.com/data/standard-deviation.html)

The Standard Deviation is a measure of how spread out numbers are.  
  
It is the square root of the Variance,   
and the Variance is the average of the squared differences from the Mean.

*See:*[*Variance*](https://www.mathsisfun.com/definitions/variance.html)



Zscore

Here is the formula for z-score that we have been using:

|  |  |  |
| --- | --- | --- |
| z score = (x-mu)/sigma |  | * **z** is the "z-score" (Standard Score) * **x** is the value to be standardized * **μ** is the mean * **σ** is the standard deviation |

# Standard Deviation and Variance

*Deviation just means how far from the normal*

## Standard Deviation

The Standard Deviation is a measure of how spread out numbers are.

Its symbol is **σ** (the greek letter sigma)

The formula is easy: it is the**square root** of the **Variance.**So now you ask, "What is the Variance?"

## Variance

The Variance is defined as:

The average of the **squared** differences from the Mean.

To calculate the variance follow these steps:

* Work out the [Mean](http://www.mathsisfun.com/mean.html) (the simple average of the numbers)
* Then for each number: subtract the Mean and square the result (the *squared difference*).
* Then work out the average of those squared differences. ([Why Square?](http://www.mathsisfun.com/data/standard-deviation.html#WhySquare))

## Formulas

Here are the two formulas, explained at [Standard Deviation Formulas](http://www.mathsisfun.com/data/standard-deviation-formulas.html) if you want to know more:

|  |  |  |
| --- | --- | --- |
| The "**Population** Standard Deviation": |  | square root of [ (1/N) times Sigma i=1 to N of (xi - mu)^2 ] |
| The "**Sample** Standard Deviation**":** |  | square root of [ (1/(N-1)) times Sigma i=1 to N of (xi - xbar)^2 ] |

Looks complicated, but the important change is to   
divide by **N-1** (instead of **N**) when calculating a Sample Variance.

## <http://www.mathsisfun.com/data/standard-deviation.html#WhySquare>

## <https://www.mathsisfun.com/data/standard-normal-distribution.html>

## Standard Scores

The number of**standard deviations from the mean**is also called the "Standard Score", "sigma" or "z-score". Get used to those words!

### Example: In that same school one of your friends is 1.85m tall

|  |  |
| --- | --- |
| You can see on the bell curve that 1.85m is**3 standard deviations**from the mean of 1.4, so:  Your friend's height has a "z-score" of 3.0 | normal distribution 95% |

It is also possible to**calculate**how many standard deviations 1.85 is from the mean

How far is 1.85 from the mean?

It is 1.85 - 1.4 =**0.45m from the mean**

How many standard deviations is that?The standard deviation is 0.15m, so:

0.45m / 0.15m =**3 standard deviations**

So to convert a value to a Standard Score ("z-score"):

* first subtract the mean,
* then divide by the Standard Deviation

And doing that is called "Standardizing"