Drawing and testing assumptions

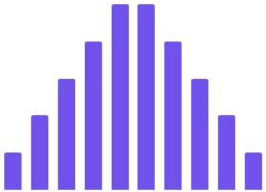
Parametric and non-parametric hypothesis tests

Statistical tests overview

> In hypothesis testing, **statistical tests** are used to decide whether we **reject or fail to reject the null hypothesis**. We use either **parametric** or **non-parametric tests**, depending on some factors of the underlying data.

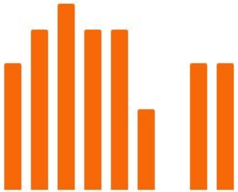
Parametric tests

- **Assume** that the data follow a **specific distribution**, such as a normal distribution.
- **More precise** estimates of the population parameters **when the assumptions are met**.
- When the data **violate the assumptions**, these tests may produce **inaccurate or unreliable** results.



Non-parametric tests

- **No assumptions** about the distribution of the data.
- **More robust to violations** of the assumptions but less powerful when the assumptions of the parametric test are met.
- Often preferred when the **sample size is small.**



Parametric tests overview

Parametric tests are a group of statistical tests that are used to **test hypotheses** on population parameters, such as the mean or variance, by making certain **assumptions about the underlying distribution of the data**.

T-test

Used to test hypotheses on the mean of a **single** population, or the **difference** between the **means** of two populations with **small sample sizes**.

Z-test

Similar to the t-test but the z-test assumes that the population **standard deviation is known** and the **sample is large**.

F-test

Used to test hypotheses on the **difference** between the **variances** of two or more populations with **large sample sizes**.

Analysis of Variance (ANOVA)

Used to test hypotheses on the **means** of three or more populations.

The t-test

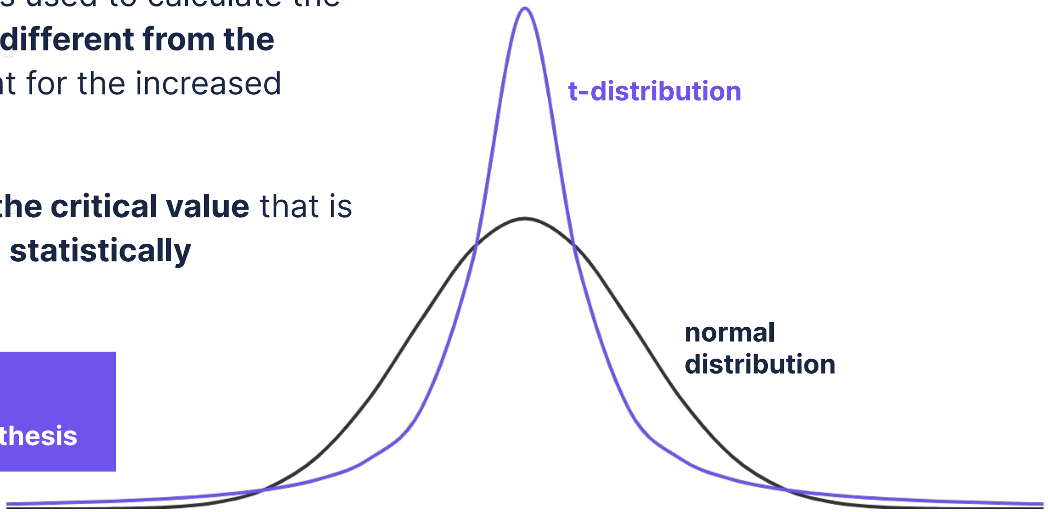
> The t-test is a parametric test based on the **t-distribution** and is used to test hypotheses on the **mean of a single population**, or the **difference between the means of two samples**, when the sample size is smaller.

The **t-distribution**, or Student's t-distribution, is used to calculate the probability of obtaining a **sample mean** that is **different from the population mean**. It has heavier tails to account for the increased variability in **smaller sample sizes**.

The **t-test** uses the t-distribution to **calculate the critical value** that is used to determine if the **difference** between is **statistically significant**.

|t-score| ≥ critical value → Reject the null hypothesis

|t-score| < critical value → Fail to reject the null hypothesis



One-sample t-test

To compare a **sample mean** with the **population mean**, we use a **one-sample t-test**.

The **test statistic *t* (t-score)** is:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$\bar{x}$ is the sample mean
s is the sample standard deviation
n is the sample size
μ is the population mean

We can use the **AVERAGE()**, **STDEV()**, and **COUNT()** Google Sheet functions to calculate the t-score.

Note: We also need to **determine the critical value** using the degrees of freedom and level of significance either from a statistical table, online calculator, or using Google Sheets.

Assumptions for one-sample t-test:

01. **Random sampling:** The data are collected using a random sampling method to ensure that the sample is representative of the population.
02. **Normality:** The distribution of the sample means is approximately normal.
03. **Independence:** The observations in the sample are independent of each other. In other words, the value of one observation is not related to the value of another observation.
04. **Homogeneity of variance (homoscedasticity):** The variance of the sample is approximately equal to the variance of the population.

Independent two-sample t-test

To compare the **means of two independent samples**, when there is no link between the two groups, we use an **independent two-sample t-test**.

The **t-score** is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$\bar{x}_1$ is sample one's mean
$\bar{x}_2$ is sample two's mean
s is pooled standard deviation
n_1 is sample one's size
n_2 is sample two's size

The pooled standard deviation is:

$$s = \sqrt{\frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}}$$

Assumptions for independent samples t-test:

01. **Normality:** The data in each group are normally distributed.
02. **Homoscedasticity:** The variance of the data in each group is equal.
03. **Independence:** The observations within each group are independent of each other, and the two groups are independent of each other.

Note: To find the p-value from the statistical table, we need to use the degrees of freedom as ($n_1 + n_2 - 2$) for the independent two-sample test.

Paired two-sample t-test

> To compare the **means of two related samples**, i.e. two sets from the same group, we use a **paired two-sample t-test**.

The **t-score** is:

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$$

$\bar{d}$ is the mean of the differences between the samples
s is the standard deviation of the differences
n is the sample size

Assumptions for paired samples t-test:

01. **Normality:** The differences between the paired observations are normally distributed.
02. **Independence:** The paired observations are independent of each other.

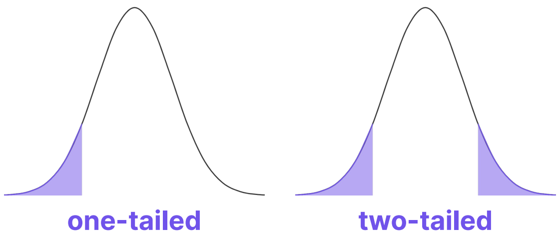
Note: To find the p-value from the statistical table, we need to use the degrees of freedom as (n - 1) for the paired two-sample test.

T-tests in Google Sheets

We can either use `AVERAGE()`, `STDEV()`, and `COUNT()` to calculate the t-score using Google Sheets and a p-value table to determine the p-value, or we can use the built-in function `T.TEST()` to calculate the p-value.

```
=T.TEST(range1, range2, tails, type)
```

- **`range1`** – The first sample of data or group of cells to consider for the t-test.
- **`range2`** – The second sample of data or group of cells to consider for the t-test.
- **`tails`** – Specify the number of distribution tails.
  - If **1**: uses a one-tailed distribution.
  - If **2**: uses a two-tailed distribution.



- **`type`** – Specifies the type of t-test.
  - If **1**: a paired test is performed.
  - If **2**: a two-sample equal variance (homoscedastic) test is performed.
  - If **3**: a two-sample unequal variance (heteroscedastic) test is performed.

If the populations being compared have equal variance, then we can use the two-sample equal variance test. If the variances differ, we use the two-sample unequal variance test.

Note: The `T.TEST()` function output is the probability associated with the t-test, i.e. the **p-value**.

The z-test

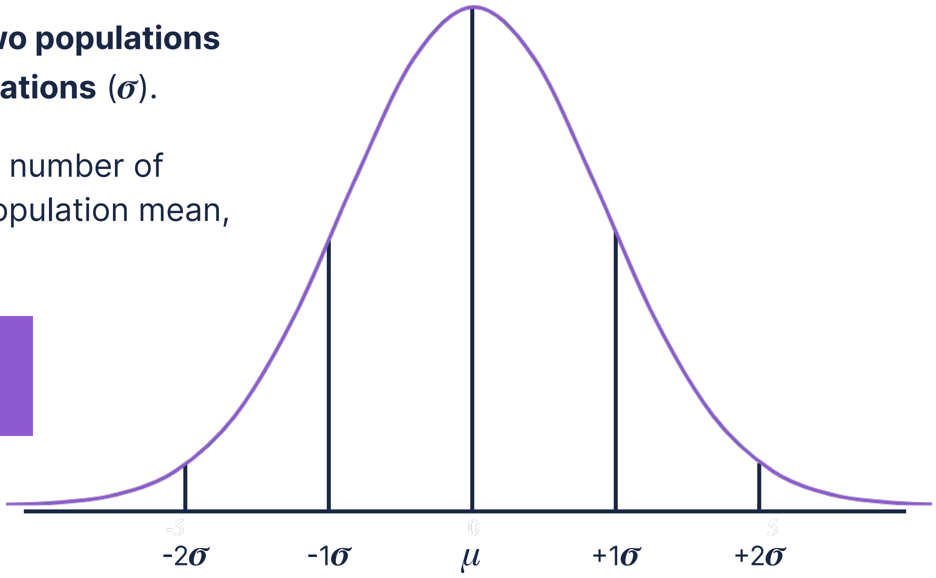
> The z-test is a parametric test based on the normal distribution (a.k.a. the z-distribution) and is similar to the t-test. However, the z-test is used when the **sample is large** and the **population standard deviation is known**.

In the **z-test**, the difference between the **means of two populations** is expressed in terms of the **number of standard deviations** (σ).

The **z-score** (test statistic z) therefore represents the number of standard deviations between the sample mean and population mean, assuming the null hypothesis is true.

|z-score| ≥ critical value → Reject the null hypothesis

|z-score| < critical value → Fail to reject the null hypothesis



One-sample z-test

To compare a **sample mean** with the **population mean** when the sample size is large and standard deviation known, we use a **one-sample z-test**.

The **test statistic *z*** is:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$\bar{x}$ is the sample mean
σ is the **population standard deviation**
n is the sample size
μ is the population mean

Note: The only difference between the test statistic *t* and *z* for a one-sample test is using the **sample** standard deviation for the *t* statistic and the **population** standard deviation for the *z* statistic.

Assumptions for one-sample z-test:

01. **Random sampling:** The sample is selected randomly from the population.
02. **Normal distribution:** The population from which the sample is drawn is normally distributed.
03. **Large sample size:** The sample size is sufficiently large, typically at least 30, so that the central limit theorem can be applied.
04. **Independence:** The observations in the sample are independent of each other.
05. **Known population standard deviation:** The standard deviation of the population is known.

Two-sample z-test

To compare the **means of two different samples** when the sample size is large and standard deviation known, we use a **two-sample z-test**.

The **test statistic *z*** for independent groups:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\bar{x}_1$ is sample one's mean
$\bar{x}_2$ is sample two's mean
σ_1 is population one's standard deviation
σ_2 is population two's standard deviation
n_1 is sample one's size
n_2 is sample two's size

Assumptions for two-sample z-test:

01. Normality: The data in each group are normally distributed.
02. Homoscedasticity: The variance of the data in each group is equal.
03. Independence: The two samples are independent of each other.
04. Known population standard deviations: The standard deviation of the populations are known.

The **test statistic *z*** for related groups:

$$z = \frac{\bar{d} - D}{\sqrt{\frac{\sigma^2}{n}}}$$

$\bar{d}$ is the mean of the differences between the samples
D is the hypothesised mean of the differences (usually equal to zero)
σ is the standard deviation of the differences
n is the sample size

Z-tests in Google Sheets

We can either use `AVERAGE()`, `STDEV()`, and `COUNT()` to calculate the z-score using Google Sheets and a statistical table to determine the p-value, or we can use the built-in function `Z.TEST()` to calculate the p-value.

```
=Z.TEST(data, value, [standard_deviation])
```

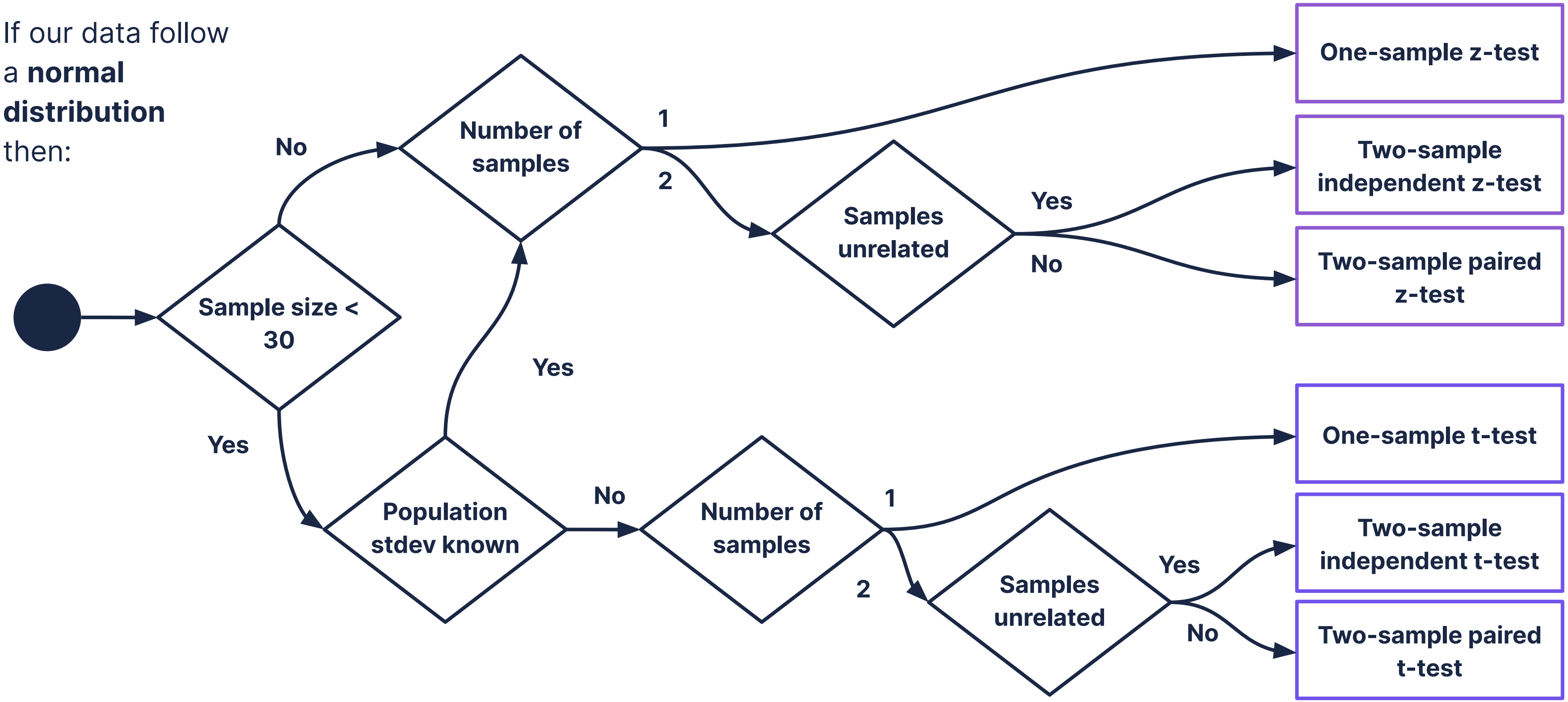
- **`data`** – The sample of data to consider for the z-test.
- **`value`** – The test statistic to use in the z-test, most often the population mean.
- **`[standard deviation]`** – The standard deviation to assume for the z-test, often the standard deviation of the population.

Note: The `Z.TEST()` function output is the probability associated with the z-test, i.e. the **p-value**.

It is important to note how the **Google Sheet** hypothesis testing **functions differ** based on the test we need to use, whether it is a one or two-sample test, and whether our samples are independent or paired.

How to choose a parametric test

If our data follow a **normal distribution** then:

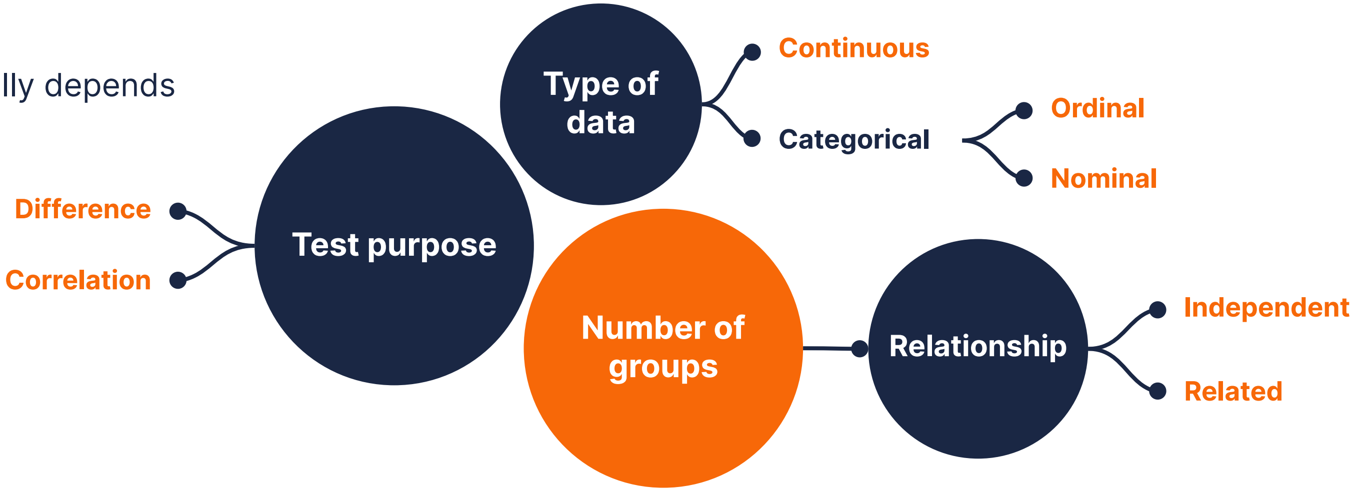


Non-parametric tests overview

Non-parametric tests are a group of statistical tests that are used to **test hypotheses** when the **underlying distribution** of the data is **unknown**.

Several non-parametric tests are available to use in hypothesis testing.

Which test to apply usually depends on:



Non-parametric tests overview

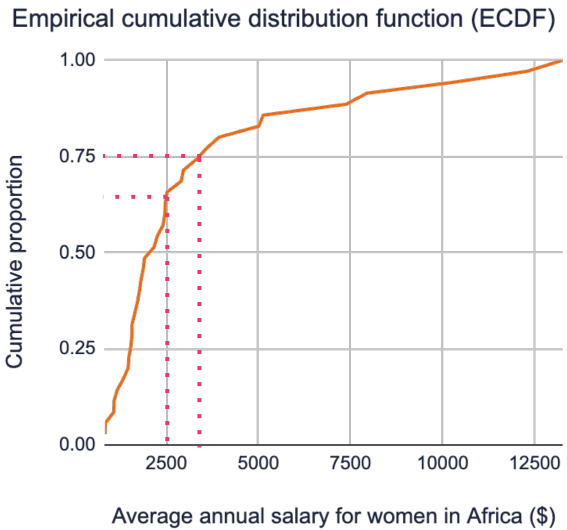
> Rather than relying on estimates of population parameters as parametric tests do, non-parametric tests are **based on ranks** or the **number of times certain events occur** in the data.

Some of the most commonly used non-parametric tests include:

- **Kolmogorov-Smirnov test:** Compares the distribution of a sample with a theoretical distribution.
- **Mann-Whitney U-test:** Determines if two independent groups come from populations with the same distribution.
- **Chi-square**: Tests for independence between categorical variables in a contingency table.
- **Spearman's rank correlation coefficient:** Measures the strength and direction of the relationship between two variables using ranks.
- **Wilcoxon signed rank test:** Determines if there is a significant difference between two paired samples using ranks.
- **Friedman test:** Tests for significant differences between three or more paired samples using ranks.
- **Kruskal-Wallis H test:** Tests for significant differences between three or more independent groups using ranks.

Kolmogorov-Smirnov and ECDF

> Kolmogorov-Smirnov (KS) is a **non-parametric** test based on the **empirical cumulative distribution function** (ECDF), which is a way to visually represent how data are **distributed.**



The ECDF **maps each observation** in a dataset **to the proportion of observations** that are less than or equal to it.

It is a step function that increases by 1/n at each point, where n is the sample size.

For example, considering the ECDF for the average annual salary for women in Africa, we see that more than 50% of women earn $2500 or less per year. We also see that 75% of women earn less than $3750 per year.

Kolmogorov-Smirnov test overview

Kolmogorov-Smirnov (KS) is used to test hypotheses on whether an **underlying distribution** observed in a sample is **similar to the hypothesised distribution**, or whether **two distributions are similar**.

Considering that KS helps us examine underlying distributions, it is a **useful tool to test for normality**, which is a prerequisite of parametric tests.

As with parametric tests, we need to state the **null** and **alternative hypotheses**:

- H_0 is that the sample is drawn from a population with a specific distribution, e.g. a normal distribution.
- H_A is that the sample is not drawn from a population with the specified distribution.

We will also need to specify the **level of significance** (α), calculate a **test statistic** (denoted **D** for Kolmogorov-Smirnov), and determine the **critical value** and **p-value**.

$|D| \geq$ critical value → Reject the null hypothesis

$|D| <$ critical value → Fail to reject the null hypothesis

p-value $\leq \alpha$ → Reject the null hypothesis

p-value $> \alpha$ → Fail to reject the null hypothesis

Kolmogorov-Smirnov test statistic

The **Kolmogorov-Smirnov test statistic** is:

$$D = \max_{1 \le i \le n} \left(\left| F(Y_i) - \frac{i-1}{n} \right|, \left| \frac{i}{n} - F(Y_i) \right| \right)$$

where

i is the index of the ordered sample $Y_1, Y_2, ..., Y_n$, i.e. the rank

n is the sample size

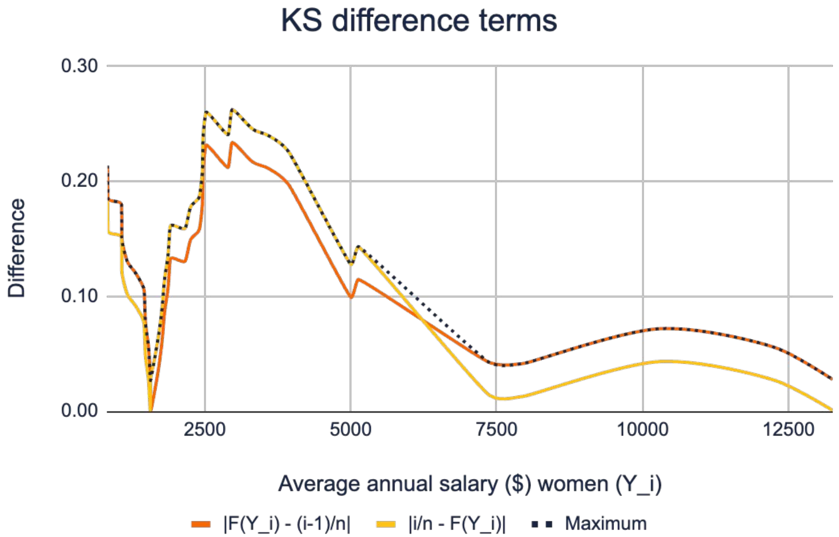
Y_i is the ith ordered value in the sample

$F(Y_i)$ is the hypothesised cumulative distribution function (CDF) evaluated at the ith ordered value of the sample data Y_i

Both **(i-1)/n** and **i/n** represent the empirical cumulative distribution function (**ECDF**)*.

The **test statistic D** is a single value which is the **maximum** across both difference terms, $|F(Y_i) - (i-1)/n|$ and $|i/n - F(Y_i)|$, for all sample values, Y_i.

We need to calculate both difference terms to ensure that the ECDF starts at 0 ((i-1)/n) and ends at 1 (i/n), i.e. the entire range.



*(i-1)/n represents the cumulative proportion observations that are expected to be strictly less than the ith ordered value, while i/n represents the proportion that is less than or equal to the ith.

Steps to the Kolmogorov-Smirnov test

The steps to performing KS:

01. State the **null** and **alternative** hypotheses:
  a. H_0 is that the sample is drawn from a population with a specific distribution, e.g. a normal distribution.
  b. H_A is that the sample is not drawn from a population with the specified distribution.

02. Specify the **level of significance** (α).

03. Calculate the **test statistic**, D, using the Kolmogorov-Smirnov test statistic formula.

04. Determine the **critical value** using the KS table, level of significance, and sample size.

05. Compare the test statistic (D) to the critical value.

Although the **p-value** for a KS test can be calculated using statistical software or a programming language like Python using built-in functions, it is much more involved and resource intensive in Google Sheets.

In theory, the p-value can be calculated using either the **exact** method, when $n \leq 35$, or the **approximate** (also known as asymptotic) method for larger sample sizes (n).