University of BRISTOL

WEEK 6

Statistical inference for multiple populations

Felipe Campelo

In this lecture...

In this lecture, we will…

- Explore how inference extends from one population to several - comparing means across two or more groups.
- See how A/B testing and multi-model evaluation fit naturally into the statistical testing framework.
- Learn the logic behind Welch’s t-test (for comparing two independent samples) and and one-way ANOVA (for multiple groups).
- Understand the assumptions of those statistical tests and learn how to verify those.

Part I

Comparison of two means

Motivation

The concepts of comparison between two populations based on information obtained from samples follow the same principles we for testing hypotheses about a single population.

Inferences using two samples frequently arise when comparing the effect of a technique (or *treatment*) against a control group - a *placebo* group (when testing the efficacy of a therapy), a baseline model (when testing the effect of a modelling approach), etc;

Comparisons of two populations also emerge quite frequently in *competitive testing*: given two methods / tools to solve a problem, is one better than the other?

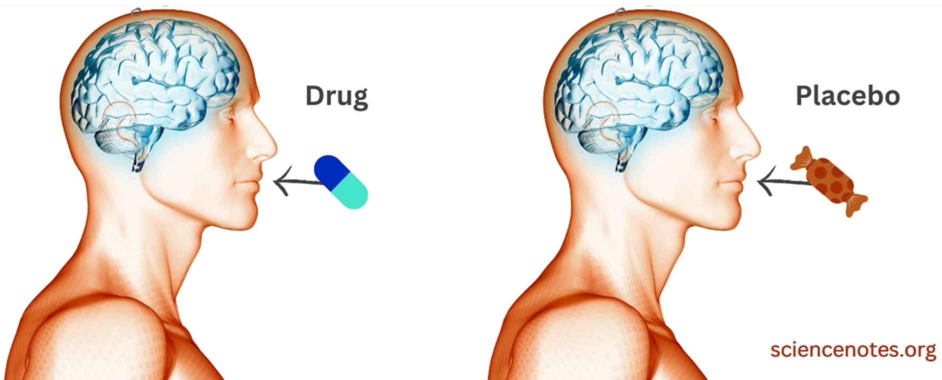


Image source: https://sciencenotes.org/placebo-effect-what-it-is-and-how-it-works/

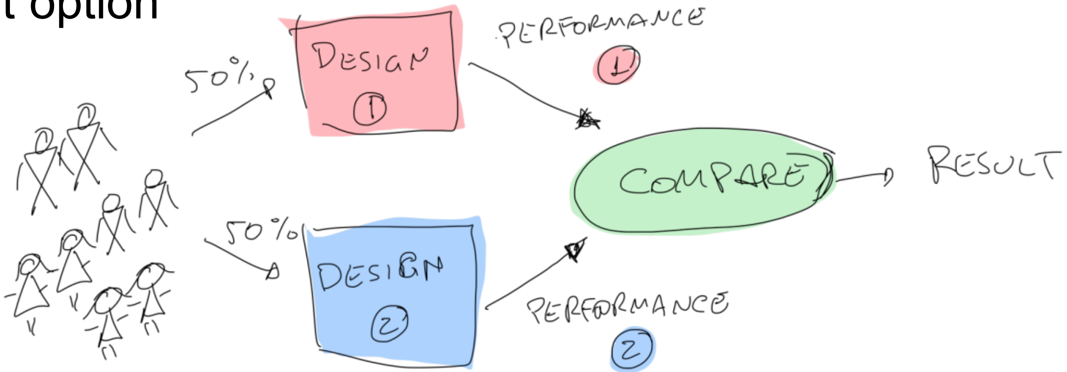
Motivating example: A/B testing

Assume that you are a data scientist at a tech company (e.g., an e-commerce site or streaming platform). The product team is launching a **new service,** and they want to know which (if any) of two *competing layouts* results in the highest average user engagement time, at a 99% confidence level.

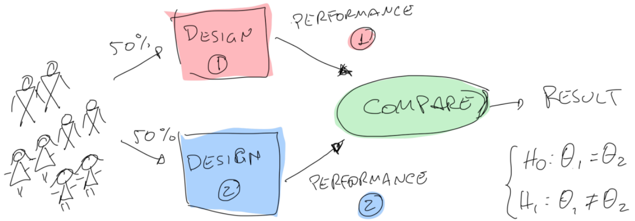
Users are ***randomly assigned*** to one of two versions of the service:

- **Group 1**: sees the first layout option
- **Group 2**: sees the second layout option

After one week, the company measures the **average session duration (in minutes)** for each group.



Some definitions



We can now introduce some standard nomenclature normally used in comparative testing.

The variable that is manipulated in a *designed experiment* (i.e., the one for which we're interested in checking the effect) is commonly called an *experimental factor,* or just a *factor.* In this example, the factor is "layout version".

The set of discrete values that a factor can take in a given experiment are the *factor levels*, or just *levels*. Sometimes these are also called *treatments*. In this example, the levels of the factor "layout version" are "layout 1" and "layout 2".

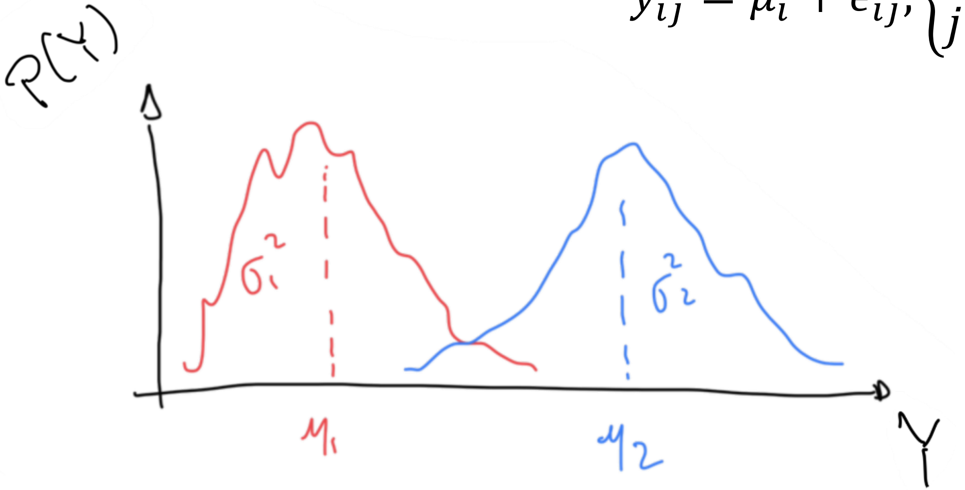
In this initial simple comparative experiment, we commonly have **one** factor with **two** levels. We'll later see how to expand this to more that two levels.

Comparative testing – some definitions

Assume that, at the end of the experiment, you obtained two samples, Y_1 and Y_2, with respective sample sizes N_A, N_B. Assume that the observations in these samples are *independent.*

We can model the jth observation of the ith level as a simple model:

$$y_{ij} = \mu_i + \epsilon_{ij}, \begin{cases} i \in \{1,2\} \\ j \in \{1, \dots, N_i\} \end{cases}$$

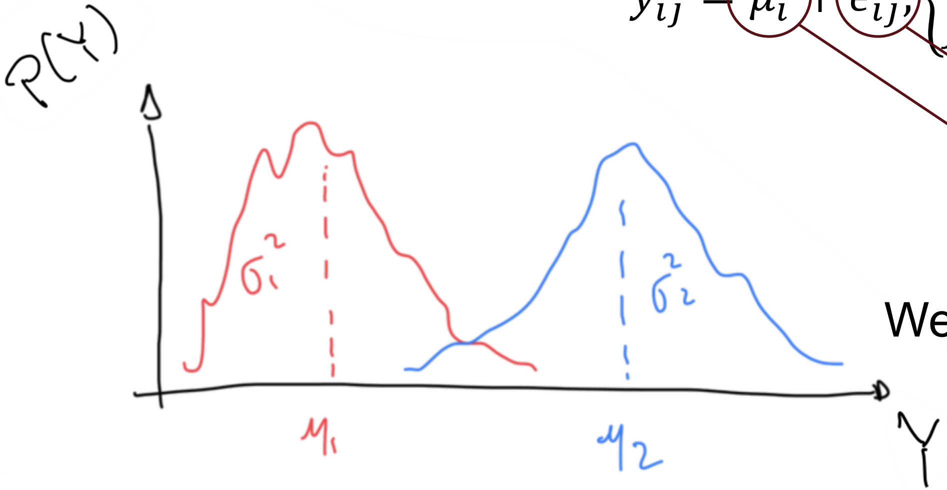


Comparative testing – some definitions

Assume that, at the end of the experiment, you obtained two samples, Y_1 and Y_2, with respective sample sizes N_A, N_B. Assume that the observations in these samples are *independent.*

We can model the jth observation of the ith level as a simple model:

$$y_{ij} = \underbrace{\mu_i}_{\textit{level means}} + \underbrace{\epsilon_{ij}}_{\textit{residuals}}, \begin{cases} i \in \{1,2\} \\ j \in \{1, \dots, N_i\} \end{cases}$$



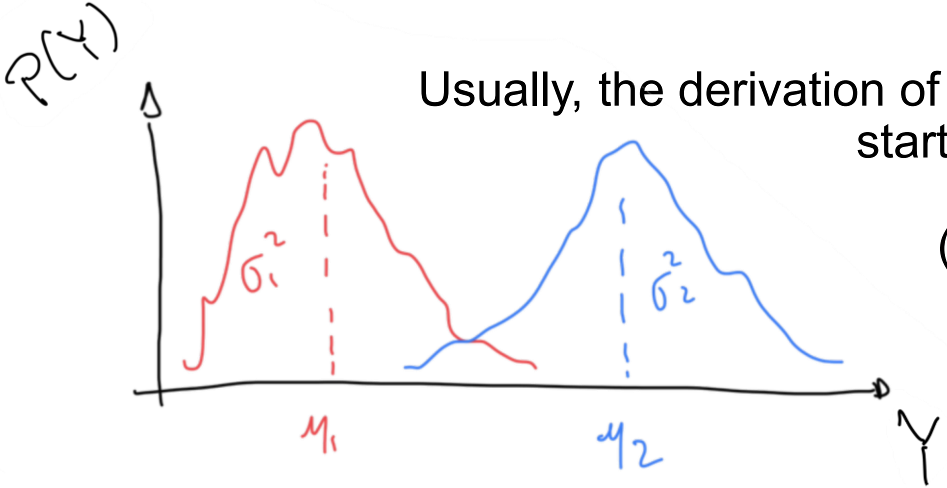
We are interested in this case in performing inference about the differences of the means.

Test hypotheses

We are interested in this case in performing inference about the differences of the means.

We can write:

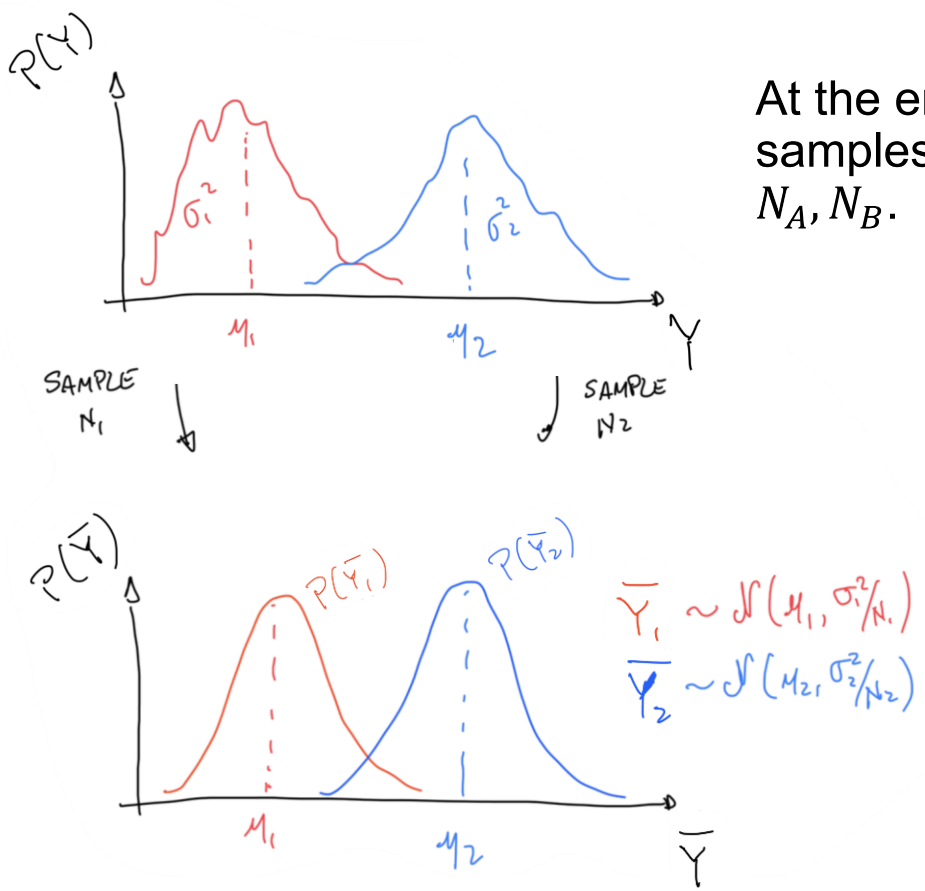
$$\begin{cases} H_0: \mu_1 - \mu_2 = \Delta\mu_0 \\ H_1: \mu_1 - \mu_2 \neq \Delta\mu_0 \end{cases} \quad \text{or, } \textit{if } \Delta\mu_0 = 0, \quad \begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$$



Usually, the derivation of the *parametric* test for these hypotheses starts from strong distributional assumptions: normality of the two populations (or, equivalently, of the residuals for each level) and equality of variances.

But we'll go straight to the more general case!

Test derivation



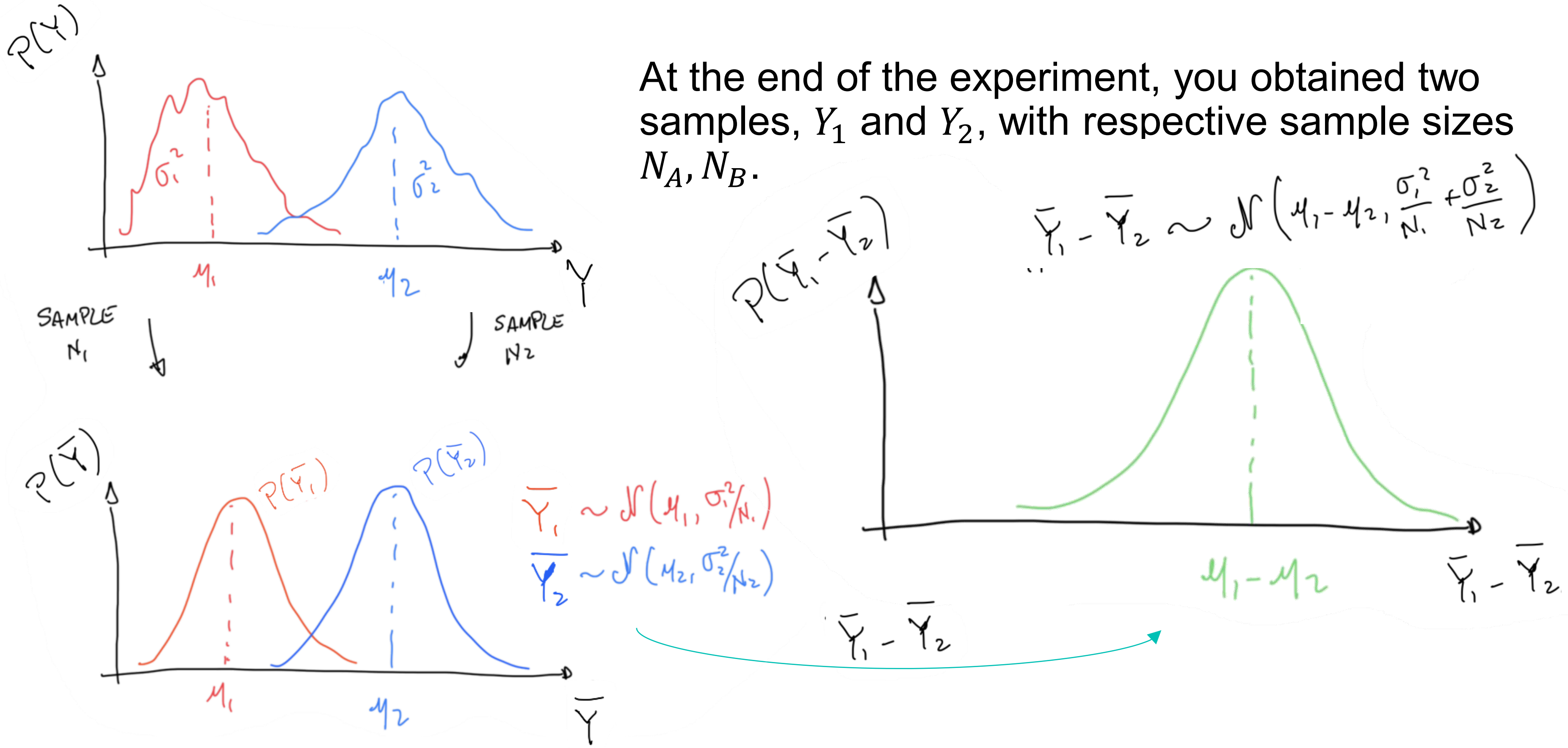
$$\bar{Y}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2/N_1)$$

$$\bar{Y}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2/N_2)$$

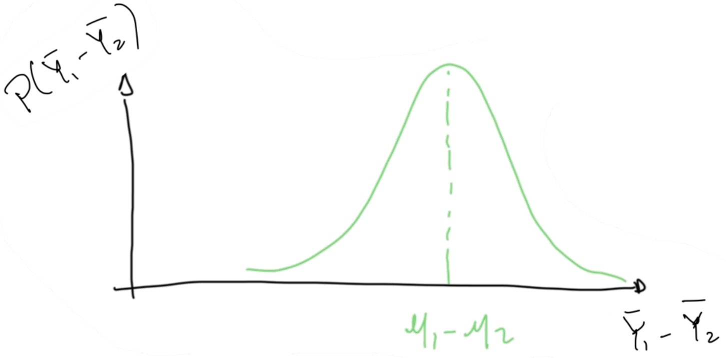
At the end of the experiment, you obtained two samples, Y_1 and Y_2, with respective sample sizes N_A, N_B.

Test derivation

At the end of the experiment, you obtained two samples, Y_1 and Y_2, with respective sample sizes N_A, N_B.


$P(Y)$
σ_1^2
σ_2^2
μ_1
μ_2
Y
SAMPLE N_1
SAMPLE N_2
$P(\bar{Y})$
$P(\bar{Y}_1)$
$P(\bar{Y}_2)$
$\bar{Y}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2/N_1)$
$\bar{Y}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2/N_2)$
μ_1
μ_2
$\bar{Y}$
$\bar{Y}_1 - \bar{Y}_2$
$\therefore \bar{Y}_1 - \bar{Y}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}\right)$
$P(\bar{Y}_1 - \bar{Y}_2)$
$\mu_1 - \mu_2$
$\bar{Y}_1 - \bar{Y}_2$

Test derivation



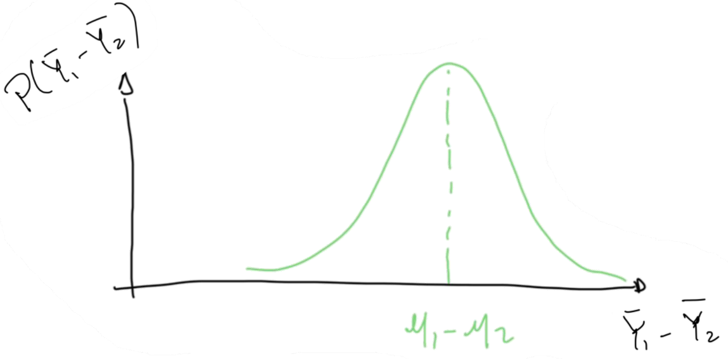
$$\bar{Y}_1 - \bar{Y}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}\right)$$

At the end of the experiment, you obtained two samples, X_1 and X_2, with respective sample sizes N_A, N_B.

You can calculate sample means and variances, $\bar{X}_1, \bar{X}_2, s_1^2, s_2^2$, from the samples. With that, we can derive the test exactly as we did for the one-sample test.

We can calculate a t-statistic as before: by centring and scaling the *sampling distribution of the difference of means:*

Test derivation



$$\bar{Y}_1-\bar{Y}_2 \sim \mathcal{N}\left(\mu_1-\mu_2, \frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}\right)$$

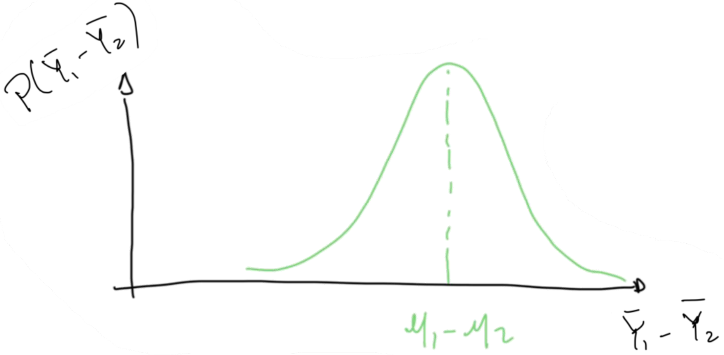
At the end of the experiment, you obtained two samples, X_1 and X_2, with respective sample sizes N_A, N_B.

You can calculate sample means and variances, $\bar{X}_1, \bar{X}_2, s_1^2, s_2^2$, from the samples. With that, we can derive the test exactly as we did for the one-sample test.

We can calculate a t-statistic as before: by centring and scaling the *sampling distribution of the difference of means:*

$$t=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}}$$

Test derivation



$$\bar{Y}_1 - \bar{Y}_2 \sim \mathcal{N}\left(\boxed{\mu_1 - \mu_2}, \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}\right)$$

At the end of the experiment, you obtained two samples, X_1 and X_2, with respective sample sizes N_A, N_B.

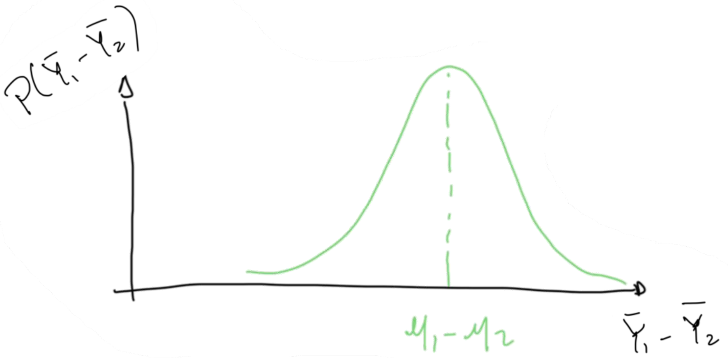
You can calculate sample means and variances, $\bar{X}_1, \bar{X}_2, s_1^2, s_2^2$, from the samples. With that, we can derive the test exactly as we did for the one-sample test.

We can calculate a t-statistic as before: by centring and scaling the *sampling distribution of the difference of means:*

Actual diff of means

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \boxed{(\mu_1 - \mu_2)}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Test derivation



$$\bar{Y}_1 - \bar{Y}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \boxed{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}\right)$$

At the end of the experiment, you obtained two samples, X_1 and X_2, with respective sample sizes N_A, N_B.

You can calculate sample means and variances, $\bar{X}_1, \bar{X}_2, s_1^2, s_2^2$, from the samples. With that, we can derive the test exactly as we did for the one-sample test.

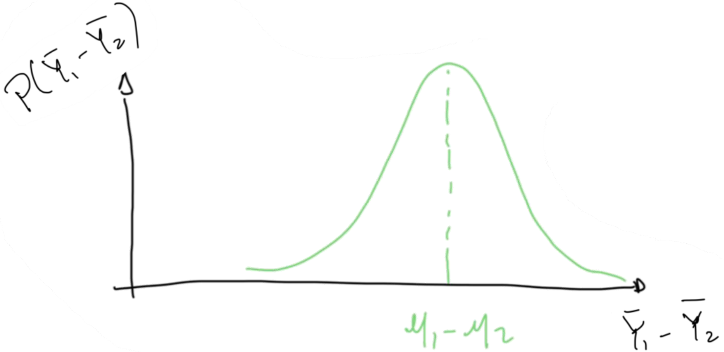
We can calculate a t-statistic as before: by centring and scaling the *sampling distribution of the difference of means:*

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\boxed{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}}$$

Estimated standard error

Test derivation

$$\begin{cases} H_0: \mu_1 - \mu_2 = \Delta\mu_0 \\ H_1: \mu_1 - \mu_2 \neq \Delta\mu_0 \end{cases}$$

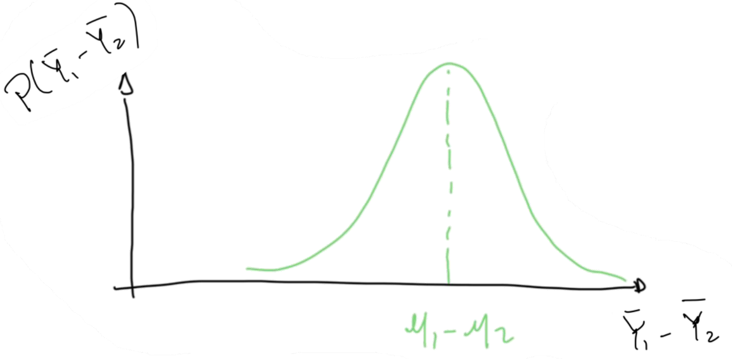


As before, we don't know the actual difference $\mu_1 - \mu_2$ - but, if H_0 is true, we do!

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - \cancel{(\mu_1 - \mu_2)}}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} \quad = \Delta\mu_0, \text{ if } H_0 \text{ is true}$$

Test derivation

$$\begin{cases} H_0: \mu_1 - \mu_2 = \Delta\mu_0 \\ H_1: \mu_1 - \mu_2 \neq \Delta\mu_0 \end{cases}$$



$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - \cancel{(\mu_1 - \mu_2)}}{\sqrt{\underbrace{s_1^2/N_1}_{se_1^2} + \underbrace{s_2^2/N_2}_{se_2^2}}}$$

$= \Delta\mu_0$, if H_0 is true

As before, we don't know the actual difference $\mu_1 - \mu_2$ - but, if H_0 is true, we do!

So, we have that, under H_0, the test statistic

$$t_0 = \frac{(\bar{Y}_1 - \bar{Y}_2) - \Delta\mu_0}{\sqrt{se_1^2 + se_2^2}}$$

is distributed as a t-distribution $t^{(\nu)}$, with ν degrees of freedom,

$$\nu = \frac{\left(se_1^2 + se_2^2\right)^2}{se_1^2/(N_1 - 1) + se_2^2/(N_2 - 1)}$$

Welch's t-test

$$\begin{cases} H_0\colon \mu_1 - \mu_2 = \Delta\mu_0 \\ H_1\colon \mu_1 - \mu_2 \neq \Delta\mu_0 \end{cases}$$

The test we just built is known as ***Welch's t-test***, and it represents a robust test for the difference of the *means of two independent populations.*

$$t_0 = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta\mu_0}{\sqrt{se_1^2 + se_2^2}} \sim t^{(\nu)} \; under\ H_0$$

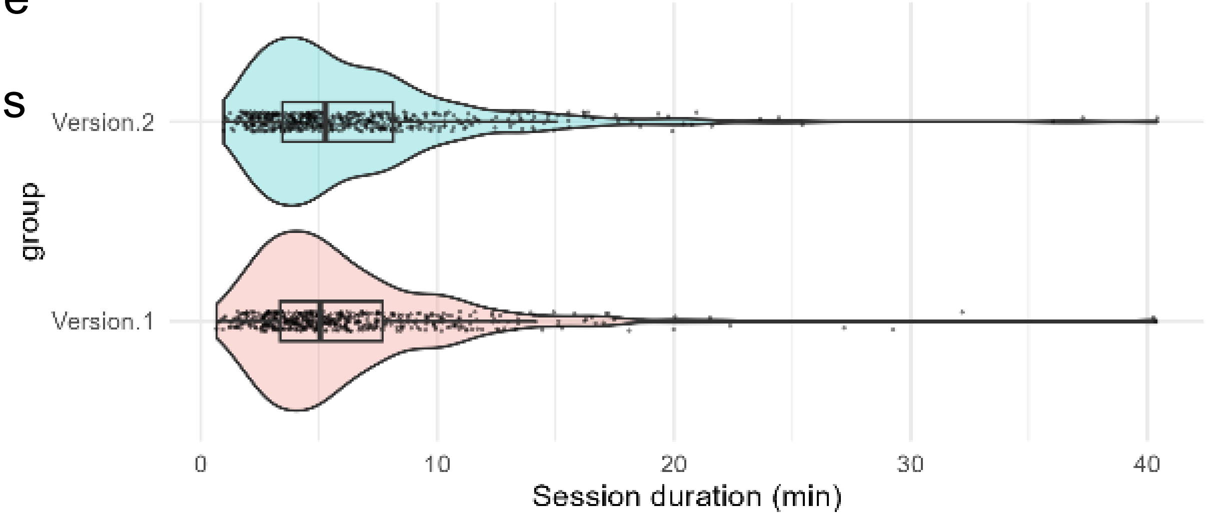
The rejection criteria are the same ones used for the simple t-test we discussed in the previous lecture - i.e., compare the obtained value of t_0 vs. the relevant quantiles of the reference t-distribution. Other aspects such as the directionality of H_1, the calculation o p-values etc., are also treated in a similar manner as the simple t-test.

There are a couple of special cases (e.g., if variances are known, or if they are unknown but $\sigma_1^2 \approx \sigma_2^2$), but in general this is the standard parametric test for the difference of means of two independent populations.

Example – A/B test

Going back to the A/B test example, we start by checking the distribution of our observations (in this case, the average session duration for each user, under either proposed version of the layout).

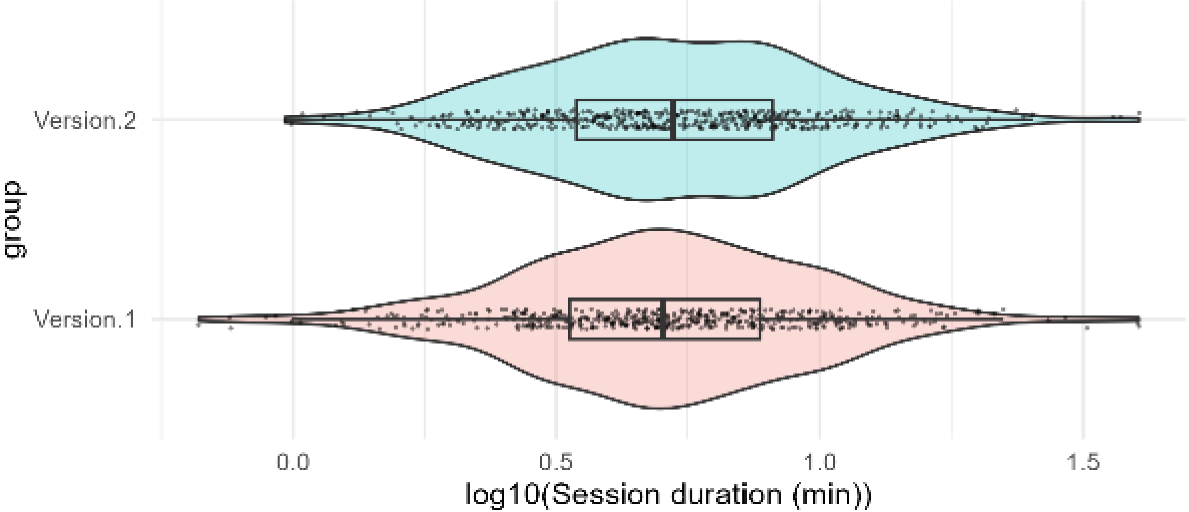
As is often the case for duration-type data, we see a heavily right-skewed data distribution. We know that the Welch t-test is robust to deviations from normality (particularly with large sample sizes), but maybe there is a different way of looking at this data?



Example – A/B test

When you have strictly-positive right-skewed observations, an often-useful *data transformation* is to take the logarithm of the values.

For this example, we can see that this stabilises the data into a more bell-shaped distribution. We can perform inference on this data with the Welch t-test, then interpret our conclusions either in the log-transformed scale or transform back to the original scale.



Example – A/B test

```
> t.test(x1, x2, conf.level = 0.99)
        Welch Two Sample t-test
data:  x1 and x2
t = -1.3022, df = 997.8, p-value = 0.1931
alternative hypothesis: true difference in
means is not equal to 0
99 percent confidence interval:
 -0.06800223  0.02239137
sample estimates:
mean of x mean of y
0.7069365 0.7297419
```

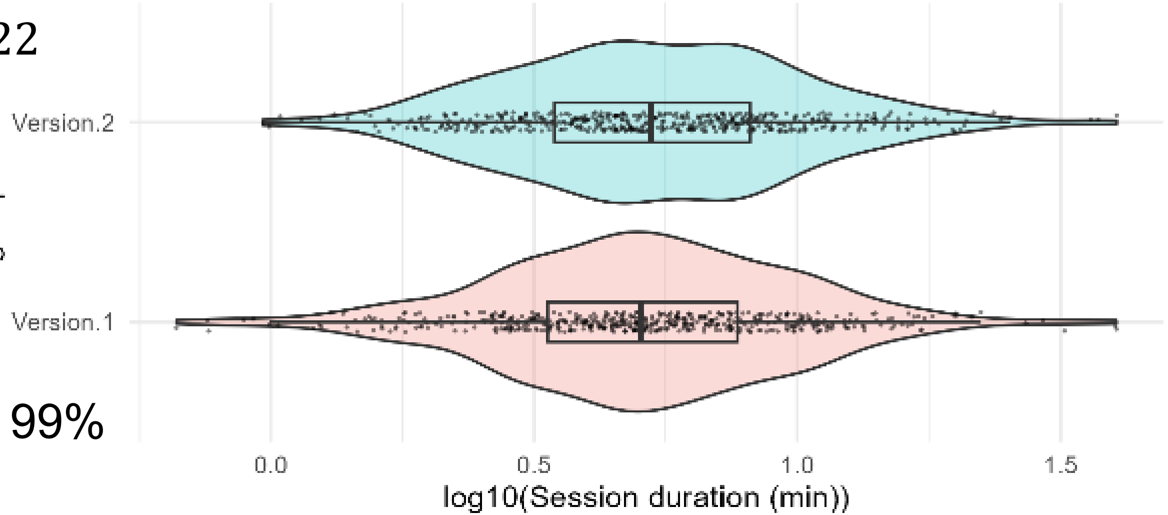
$\bar{x}_1 = 0.706$
$\bar{x}_2 = 0.730$
$s_1 = 0.279$
$s_2 = 0.275$
$N_1 = N_2 = 500$

$$t_0 = \frac{\bar{x}_1 = \bar{x}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} = -1.3022$$

P-value:

```
> 2*pt(-1.3022, df = 997.8)
[1] 0.1931485
```

i.e., we cannot reject H_0 at the 99% confidence level.



Test assumptions

Welch’s t-test is based on two mains assumptions:

- *Independence*. This commonly refers to the observations within a sample being representative of the population to which you want the inference to generalise, without co-dependencies (within-sample independence) as well as the absence of unmodelled dependencies between the two populations (between-samples independence).

  Violations of the within-sample independence usually occur when repeated measurements are not consolidated (e.g., if you have many observations coming from the same user – commonly prevented through *aggregation*) or when there is some level of information being communicated between observations (e.g., if some observations are influenced by others – commonly prevented by setting up data collection to minimise this sort of communication).

Test assumptions

Welch’s t-test is based on two mains assumptions:

- *Normality*. As was the case of the one-sample t-test, there is the assumption that the sampling distribution of the difference of means is well-approximated by a normal distribution. This is guaranteed if the data-generating distributions are normal, or if the sample sizes N_1, N_2 are large enough for the sampling distribution to converge to an approximately normal shape (lower N required for bell-shaped or symmetric distributions, larger N needed for heavily asymmetric / heavy-tailed ones).

  Large sample sizes and reasonably bell-shaped data distributions are good indications that this assumption is satisfied.
  If those characteristics are not present, visual inspection of the bootstrapped sampling distribution of the differences of means is a good alternative.

Part II

Comparisons of multiple means

Example

A fintech lender wants to choose which of **four candidate credit-scoring models** to deploy.

Instead of running a lot of pairwise A/B tests, the product team runs a multi-arm randomised experiment: each new loan applicant is randomly assigned to one of the five models, which decides whether to approve the loan and at what rate.

After some time, the company measures profit per customer (interest earned minus defaults and servicing costs) for every applicant. Assume that each arm of this experiment included 1000 customers, followed for 90 days.

Assume a desired significance level $\alpha = 0.05$.

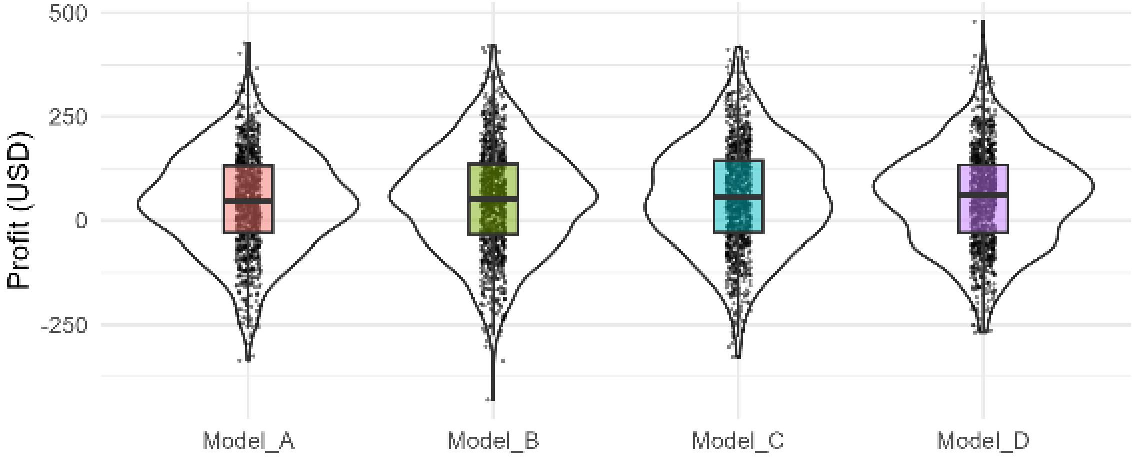
Motivating example

We have (using the same nomenclature discussed in part I):

- One **experimental factor** (model) with $a = 4$ **levels** (model alternatives)
- One **response variable** (profit)
- $n = 1000$ observations in each level $i = 1, \ldots, 4$.

There is a lot of overlap between the samples? Are any differences of means significant?

```
> ggplot(X, aes(x = model, y = profit_usd)) +
   geom_violin(alpha = 0) +
   geom_jitter(alpha = .5, stroke=0,
               size = .75, width = .05) +
   geom_boxplot(aes(fill = model),
                alpha = 0.5, width = .2,
                show.legend = FALSE,
                outlier.shape = NA) +
   theme_minimal() +
   xlab("") + ylab("Profit (USD)")
```



Modelling the data

We can model this data using a similar model to the one we used fort the t-test for two independent samples,

$$y_{ij} = \mu_i + \epsilon_{ij}, \begin{cases} i \in \{1,2,3,4,5\} \\ j \in \{1, \dots, N\} \end{cases}$$

We can split the μ_i terms into a *grand mean*, μ, and *level effect* terms, τ_i:

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}$$

In the derivation of the statistical test for the existence of differences in the group means, we will initially consider a few assumptions about the *residuals*:

$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ *(Residuals are identically and independently distributed as a normal variable with zero mean and constant variance)*

Modelling the data

We can model this data using a similar model to the one we used fort the t-test for two independent samples,

$$y_{ij} = \mu_i + \epsilon_{ij}, \begin{cases} i \in \{1,2,3,4,5\} \\ j \in \{1, \dots, N\} \end{cases}$$

We can split the μ_i terms into a *grand mean*, μ, and *level effect* terms, τ_i:

$$y_{ij} = \overbrace{\mu_i + \epsilon_{ij}}^{\text{“Means model”}} = \overbrace{\mu + \tau_i + \epsilon_{ij}}^{\text{“Effects model”}}$$

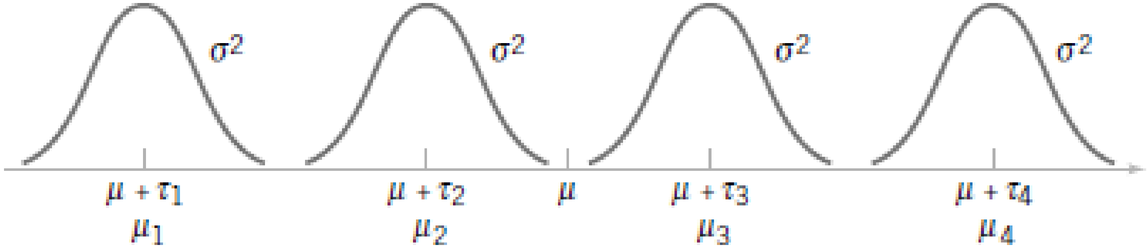
In the derivation of the statistical test for the existence of differences in the group means, we will initially consider a few assumptions about the *residuals*:

$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ *(Residuals are identically and independently distributed as a normal variable with zero mean and constant variance)*

Modelling the data

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}$$
$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

Under this modelling, the data is expected to be distributed like this:



$$\sum_{\forall i} \tau_i = 0$$

To test for differences in the mean values of each population, we can describe the hypotheses as:

$$\begin{cases} H_0: \mu_i = \mu_j, \forall i, j \\ H_1: \mu_i \neq \mu_j, \text{for any } i, j \end{cases}$$

or, equivalently,

$$\begin{cases} H_0: \tau_i = 0, \forall i \\ H_1: \tau_i \neq 0, \text{for any } i \end{cases}$$

If data collection is performed in a randomised manner under constant experimental conditions, we have what’s called a *completely randomized design*

Modelling the data

Based on the effects model, $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, we can represent the total variability in the data as a quantity known as *total sum of squares*,

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{..}\right)^2$$

Sample mean of all observations

$$\bar{y}_{..} = \frac{1}{an} \sum_{i,j} y_{ij}$$

Modelling the data

Based on the effects model, $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, we can represent the total variability in the data as a quantity known as *total sum of squares*,

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2$$

Sample mean of all observations

$$\bar{y}_{..} = \frac{1}{an}\sum_{i,j} y_{ij}$$

Sample mean of group i

$$\bar{y}_{i.} = \frac{1}{n}\sum_{j} y_{ij}$$

With some relatively simple algebra, we can separate SS_T into two components, a between-groups variability and a within-groups variability,

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2 = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

Between-groups variability, SS_{Groups}

Within-groups variability, SS_E

Modelling the data

From the decomposition of the variances, we can derive two *variance estimators*, one biased and one unbiased:

$$MS_{levels} = \frac{n}{a-1}\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

$$MS_E = \frac{1}{a(n-1)}\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2$$

$$\mathrm{E}[\mathrm{MS_E}] = \sigma^2$$

$$\mathrm{E}[\mathrm{MS_{levels}}] = \sigma^2 + \frac{n}{a-1}\sum\nolimits_i \tau_i^2$$

Derivation of the test

$$\mathrm{E}[\mathrm{MS_E}] = \sigma^2$$

$$\mathrm{E}[\mathrm{MS_{levels}}] = \sigma^2 + \frac{n}{a-1}\sum_i \tau_i^2$$

Recall the hat the null hypothesis of no difference of the means can be expressed in terms of the level effects,

$$\begin{cases} H_0\colon \tau_i = 0, \forall i \\ H_1\colon \tau_i \neq 0, \text{for any } i \end{cases}$$

This means that, *under the null hypothesis*, $\mathrm{E}[\mathrm{MS_{levels}}] = \mathrm{E}[\mathrm{MS_E}] = \sigma^2$ and the test statistic

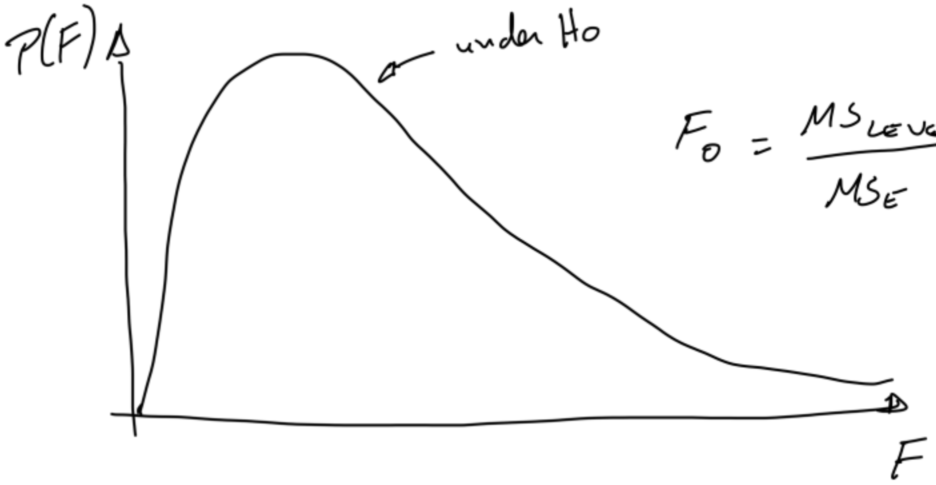
$$F_0 = \frac{MS_{levels}}{MS_E}$$

is a random variable that follows an F-distribution with $(a-1)$ numerator degrees-of-freedoms and $a(n-1)$ denominator degrees-of-freedom.

Derivation of the test

$$E[\mathrm{MS_E}] = \sigma^2$$

$$E[\mathrm{MS_{levels}}] = \sigma^2 + \frac{n}{a-1}\sum_i \tau_i^2$$



$$F_0 = \frac{MS_{LEVELS}}{MS_E}$$

If Ho is false,
$E[MS_{Levels}] > E[MS_E]$

$$\begin{cases} H_0\colon \tau_i = 0, \forall i \\ H_1\colon \tau_i \neq 0, \text{for any } i \end{cases}$$

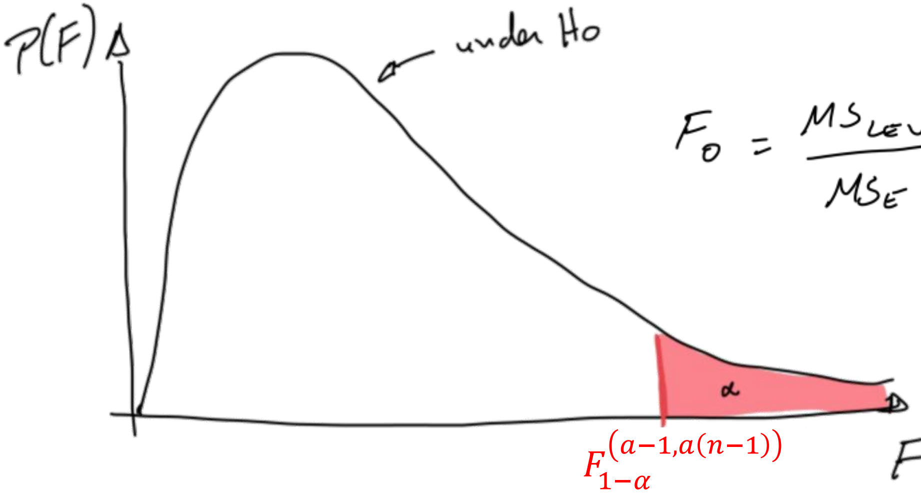
$$F_0 = \frac{MS_{levels}}{MS_E} \sim F^{(a-1, a(n-1))}$$

(under H_0)

Derivation of the test

$$\mathrm{E}[\mathrm{MS_E}] = \sigma^2$$

$$\mathrm{E}[\mathrm{MS_{levels}}] = \sigma^2 + \frac{n}{a-1}\sum_i \tau_i^2$$



$$F_0 = \frac{MS_{LEVELS}}{MS_E}$$

If Ho is false,
E[MS

$$\begin{cases} H_0: \tau_i = 0, \forall i \\ H_1: \tau_i \neq 0, \text{for any } i \end{cases}$$

$$F_0 = \frac{MS_{levels}}{MS_E} \sim F^{(a-1,a(n-1))}$$

(under H_0)

We reject H_0 at the $1-\alpha$ confidence level if $F_0 > F_{1-\alpha}^{(a-1,a(n-1))}$.

Analysis of variance

$$\begin{cases} H_0: \tau_i = 0, \forall i \\ H_1: \tau_i \neq 0, \text{for any } i \end{cases}$$

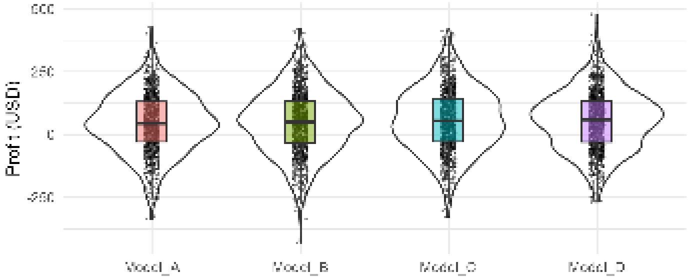
This test to detect the existence of differences in the means of multiple populations is called *analysis of variance* (ANOVA for short).

$$F_0 = \frac{MS_{levels}}{MS_E} \sim F^{(a-1, a(n-1))}$$

$(under\ H_0)$

More specifically, this is known as a **one-way ANOVA** (to reflect the fact that there is only one experimental factor being considered) using the **fixed effects model** (to reflect the fact that the levels of the factor are fixed, and not a sample from a population of possible levels).

Example: credit scoring models



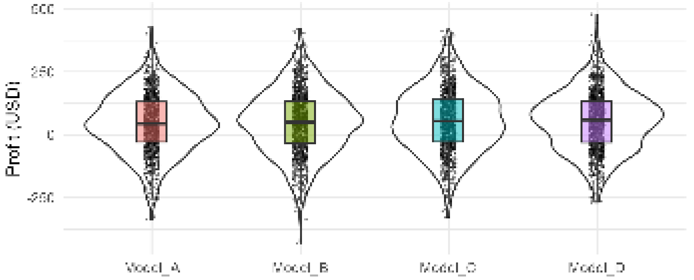
Recall our motivating example: comparing the mean 90-day profit of four credit-scoring models.

Although mathematically more complex than the t-tests, we can run this analysis in R quite easily:

```
> myaov <- aov(profit_usd ~ model, data = X)
> summary(myaov)

              Df   Sum Sq Mean Sq F value Pr(>F)
model          3    92431   30810   2.055  0.104
Residuals   3996 59902573   14991
```

Example: credit scoring models



Recall our motivating example: comparing the mean 90-day profit of four credit-scoring models.

Although mathematically more complex than the t-tests, we can run this analysis in R quite easily:

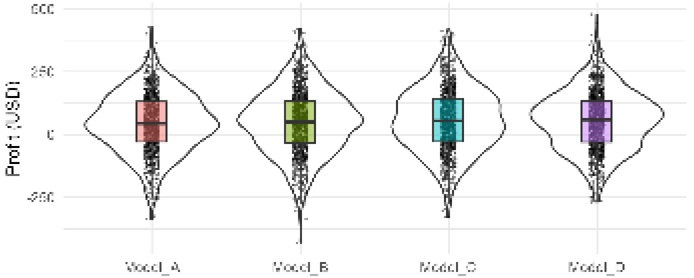
```
> myaov <- aov(profit_usd ~ model, data = X)
> summary(myaov)

              Df    Sum Sq Mean Sq F value Pr(>F)
model          3     92431   30810   2.055  0.104
Residuals   3996  59902573   14991
```

Components of variability

Experimental factor, residuals.

Example: credit scoring models



Recall our motivating example: comparing the mean 90-day profit of four credit-scoring models.

Although mathematically more complex than the t-tests, we can run this analysis in R quite easily:

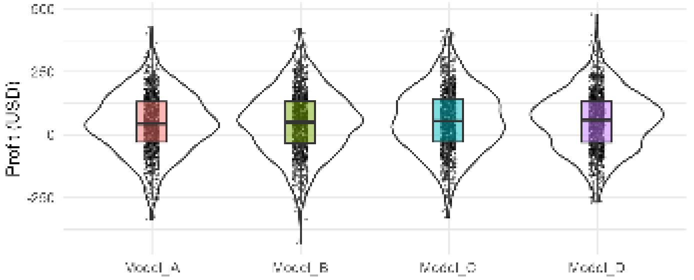
```
> myaov <- aov(profit_usd ~ model, data = X)
> summary(myaov)

              Df   Sum Sq Mean Sq F value Pr(>F)
model          3    92431   30810   2.055  0.104
Residuals   3996 59902573   14991
```

Degrees of freedom

Always check that the DoF for the factor levels equals $a - 1$. Function *aov()* expects the factor variable to be a *character* or *factor* variable – if it is expressed as a number, it needs to be converted, otherwise the function fits a linear regression instead, which gives a DoF of 1 instead of $a - 1$.

Example: credit scoring models



Recall our motivating example: comparing the mean 90-day profit of four credit-scoring models.

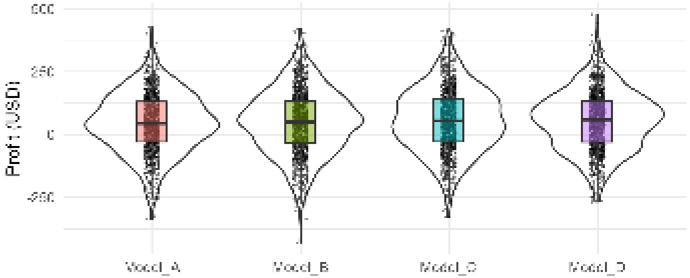
Although mathematically more complex than the t-tests, we can run this analysis in R quite easily:

```
> myaov <- aov(profit_usd ~ model, data = X)
> summary(myaov)

               Df   Sum Sq Mean Sq F value Pr(>F)
model           3    92431   30810   2.055  0.104
Residuals    3996 59902573   14991
```

Sum of squares, Mean squares (MS_{levels}, MS_E)
Reported, but commonly not used for the final analysis.

Example: credit scoring models



Recall our motivating example: comparing the mean 90-day profit of four credit-scoring models.

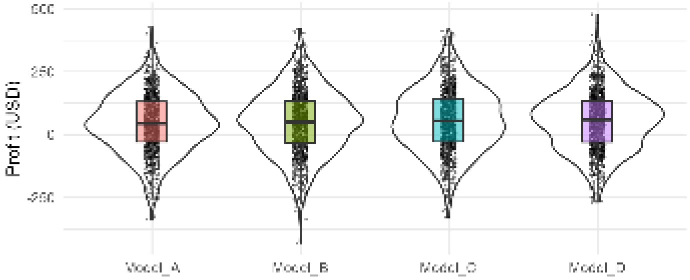
Although mathematically more complex than the t-tests, we can run this analysis in R quite easily:

```
> myaov <- aov(profit_usd ~ model, data = X)
> summary(myaov)

              Df   Sum Sq Mean Sq F value Pr(>F)
model          3    92431   30810   2.055  0.104
Residuals   3996 59902573   14991
```

Computed value of F_0

Example: credit scoring models



Recall our motivating example: comparing the mean 90-day profit of four credit-scoring models.

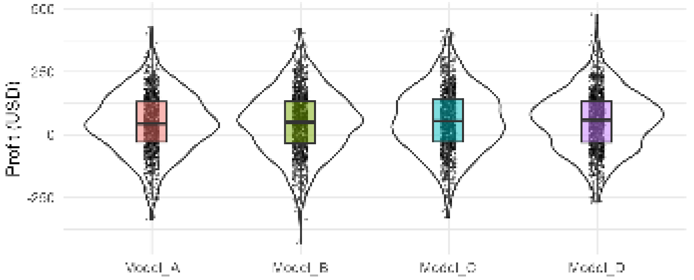
Although mathematically more complex than the t-tests, we can run this analysis in R quite easily:

```
> myaov <- aov(profit_usd ~ model, data = X)
> summary(myaov)

              Df   Sum Sq Mean Sq F value Pr(>F)
model          3    92431   30810   2.055  0.104
Residuals   3996 59902573   14991
```

P-value

Example: credit scoring models



Recall our motivating example: comparing the mean 90-day profit of four credit-scoring models.

Although mathematically more complex than the t-tests, we can run this analysis in R quite easily:

```
> myaov <- aov(profit_usd ~ model, data = X)
> summary(myaov)

              Df    Sum Sq Mean Sq F value Pr(>F)
model          3     92431   30810   2.055  0.104
Residuals   3996  59902573   14991
```

In this example, the different scoring models do not present statistically significant differences in terms of mean profit.

We’ll discuss how to proceed if the results are statistically significant in this week’s lab. For now, we’ll finish with a brief discussion of the test assumptions.

ANOVA assumptions

The assumptions of the ANOVA are encapsulated in the expected distribution of the model residuals. Recall the data model used in the derivation of the test:

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}$$
$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

*Residuals are **identically** and **independently** distributed as a **normal** variable with zero mean and **constant variance***

ANOVA assumptions

The assumptions of the ANOVA are encapsulated in the expected distribution of the model residuals. Recall the data model used in the derivation of the test:

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}$$
$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

This can be summarised into three *statistical assumptions*:

- **Normality** of model residuals.
- **Equality of variances** of model residuals across factor levels.
- **Independence** of model residuals.

It’s easy to extract the residuals directly from the model object in R:

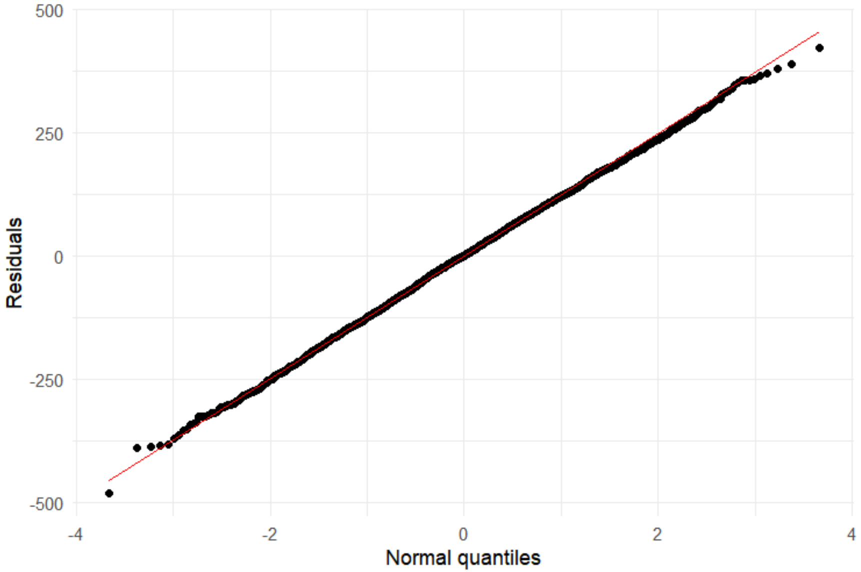
```
myaov$residuals
```

ANOVA assumptions

Normality of model residuals.

This is usually well-assessed by a normal QQ-plot.

If the sample size is reasonably **small**, a normality test such as the Shapiro-Wilk test (R function *shapiro.test()*) can also be used.



```
> ggplot(data.frame(x = myaov$residuals),
             aes(sample = x)) +
   geom_qq() + geom_qq_line(col = "red") +
   theme_minimal() +
   xlab("Normal quantiles") + ylab("Residuals")
```

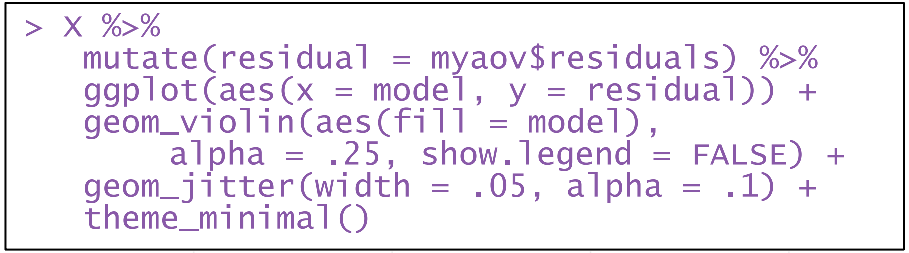
ANOVA is reasonably robust to moderate deviations from normality, as long as the other assumptions are met.

ANOVA assumptions

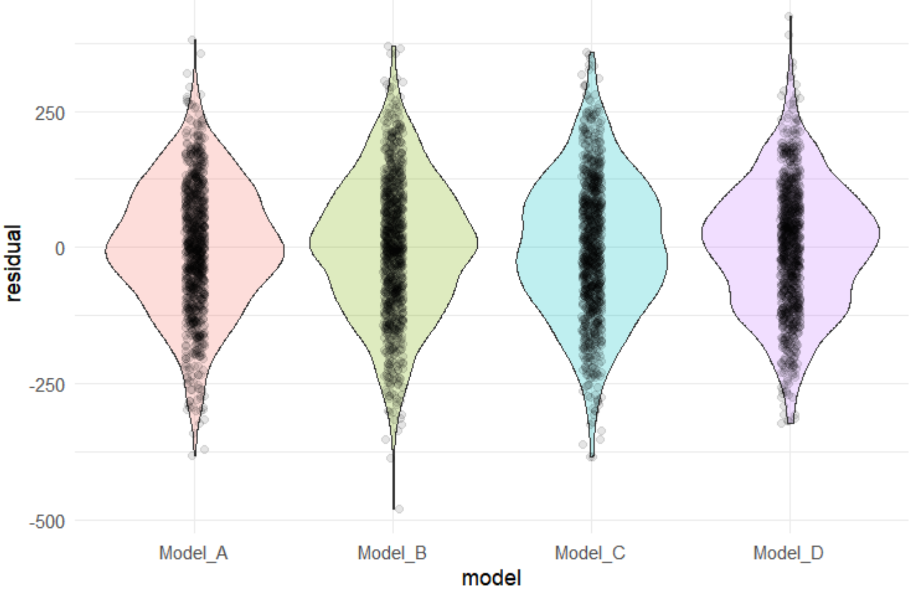
Equality of variances across factor levels (a.k.a. *homoscedasticity).*

This is usually well-assessed by a plot of *level means x residuals*.
It can also be assessed using
a test called *Fligner-Kileen* test
(R function *fligner.test()*)

ANOVA is also reasonably robust to moderate deviations from this assumption.



```
> X %>%
   mutate(residual = myaov$residuals) %>%
   ggplot(aes(x = model, y = residual)) +
   geom_violin(aes(fill = model),
        alpha = .25, show.legend = FALSE) +
   geom_jitter(width = .05, alpha = .1) +
   theme_minimal()
```

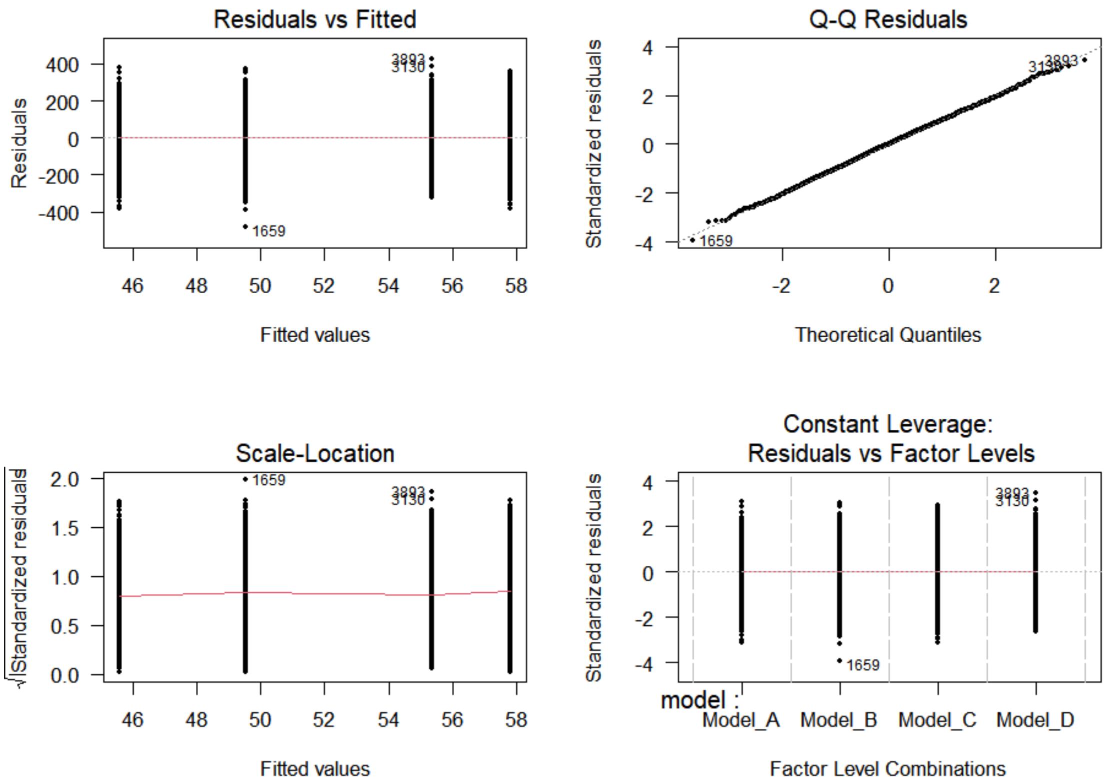


ANOVA assumptions

```
> par(mfrow = c(2,2))
> plot(myaov, pch = 16, cex = .5, las = 1)
> par(mfrow = c(1,1))
```

A practical way to visualise ANOVA residuals is to plot the object returned by AOV directly.

This provides diagnostic plots that, although less visually appealing than ggplot2, provide the necessary information to diagnose the assumptions of normality and homogeneity of variances.



ANOVA assumptions

Independence of model residuals.

The assumption of independence is the most important (not only for ANOVA, but for most if not all statistical tests), and the most difficult to test generally.

Ideally, this should be guaranteed to the best of the data scientist’s knowledge, with support from domain experts:

- On the *experimental design* and *data collection* stages: ensuring that the data collection prevents the emergence of ordering effects, location effects, or other data dependencies that could contaminate the data.

- On the *analysis* stage: making sure that what is considered an *observation* during the analysis really represents an independent measurement from the population of interest (consolidate repeated measurements or other repeated measurements with potential co-dependencies).

ANOVA assumptions

Independence of model residuals.

Statistical tests can be quite sensitive to violations of the independence assumption. *Randomization* (of the data collection ordering, positional allocation etc) and attention to potentially influential covariates can help avoiding violations of this assumption.

To test for *serial correlations* (which is only one type of independence violation) we can use the Durbin-Watson test (R function *dwtest*() from package *lmtest*), but that only makes sense if the data is presented to the DW test ordered by an unmodelled and possibly influential variable (such as the order of data collection).

In this lecture we discussed…

- A standard test for comparing the means of two populations (with no expectation of known or equal variances).
- The one-way ANOVA for comparing the means of multiple populations (under an equal-variances assumption)
- The underlying statistical assumptions of these tests and how to verify them.

In this week’s lab, we’ll revisit these concepts and expand them to include nonparametric (bootstrap) alternatives, post-ANOVA pairwise tests, and how to deal with matched-samples scenarios.

Further reading

DC Montgomery, GC Runger, *Applied statistics and probability for engineers, 5th ed.*
[Chapters 10.1, 10.2, 13.1 and 13.2]

- *Read the chapters and try to solve the worked examples by yourself*.