


Data Analytics for Pharmaceutical Sciences

Part II: Analysis of Variance (ANOVA)

Stéphane Guerrier, Data Analytics Lab, University of Geneva 

Dominique-L. Couturier, Cancer Research UK, University of Cambridge 

Yuming Zhang, Data Analytics Lab, University of Geneva 



License: CC BY NC SA 4.0

Two-sample location tests

In practice, we often encounter problems where our goal is to compare the means (or locations) of two samples. For example,

1. A scientist is interested in comparing the vaccine efficacy of the Pfizer-BioNTech and the Moderna vaccine.
2. A bank wants to know which of two proposed plans will most increase the use of its credit cards.
3. A psychologist wants to compare male and female college students' impression on a selected webpage.

We will discuss three two-sample location tests:


1. Two independent sample Student's t-test
2. Two independent sample Welch's t-test
3. Two independent sample Mann-Whitney-Wilcoxon test

Two independent sample Student's t-test

This test considers the following assumed model for group **A** and **B**

$$X_{i(g)} = \mu_g + \varepsilon_{i(g)} = \mu + \delta_g + \varepsilon_{i(g)},$$

where $g = A, B, i = 1, \dots, n_g, \varepsilon_{i(g)} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and $\sum n_g \delta_g = 0$.

 n_A = sample size of group **A**, $\mu_A = \mu + \delta_A$ = population mean of group **A**, n_B and $\mu_B = \mu + \delta_B$ are similarly defined for group **B**.

Hypotheses:

$$H_0 : \mu_A - \mu_B = \mu_0 \quad \text{and} \quad H_a : \mu_A - \mu_B \left[> \text{ or } < \text{ or } \neq \right] \mu_0.$$

Test statistic's distribution under H_0 :

$$T = \frac{(\bar{X}_A - \bar{X}_B) - \mu_0}{s_p \sqrt{n_A^{-1} + n_B^{-1}}} \underset{H_0}{\sim} \text{Student}(n_A + n_B - 2).$$

Discussion - Student's t-test

- R function:

```
t.test(x = ... , y = ... , alternative = ... , var.equal = TRUE).
```


- This test strongly relies on the **assumed absence of outliers**. If outliers appear to be present the Mann-Whitney-Wilcoxon test (see later) is (probably) a better option.
- For moderate and small sample sizes, the sample distribution should be at least **approximately normal** with no strong skewness to ensure the reliability of the test.
- In practice, the assumption of equal variance is hard to verify so **we recommend to avoid this test in practice**.

Two independent sample Welch's t-test

This test considers the following assumed model for group **A** and **B**

$$X_{i(g)} = \mu_g + \varepsilon_{i(g)} = \mu + \delta_g + \varepsilon_{i(g)},$$

where $g = A, B, i = 1, \dots, n_g, \varepsilon_{i(g)} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_g^2)$ and $\sum n_g \delta_g = 0$.

 n_A = sample size of group **A**, $\mu_A = \mu + \delta_A$ = population mean of group **A**, n_B and $\mu_B = \mu + \delta_B$ are similarly defined for group **B**.

Hypotheses:

$$H_0 : \mu_A - \mu_B = \mu_0 \quad \text{and} \quad H_a : \mu_A - \mu_B \left[> \text{ or } < \text{ or } \neq \right] \mu_0.$$

Test statistic's distribution under H_0 :

$$T = \frac{(\bar{X}_A - \bar{X}_B) - \mu_0}{\sqrt{s_A^2/n_A + s_B^2/n_B}} \underset{H_0}{\sim} \text{Student}(df).$$

Discussion - Welch's t-test

- R function:

```
t.test(x = ... , y = ... , alternative = ... ).
```


- This test strongly relies on the **assumed absence of outliers**. If outliers appear to be present the Mann-Whitney-Wilcoxon test (see later) is (probably) a better option.
- For moderate and small sample sizes, the sample distribution should be at least **approximately normal** with no strong skewness to ensure the reliability of the test.
- This test does not require the variances of the two groups to be equal. If the variances of the two groups are the same (which is rather unlikely in practice), the Welch's t-test losses a little bit of power compared to the Student's t-test.
- The computation of ***df*** (i.e. the degrees of freedom of the distribution under the null) is beyond the scope of this class.

Mann-Whitney-Wilcoxon test

This test considers the following assumed model for group **A** and **B**

$$X_{i(g)} = \theta_g + \varepsilon_{i(g)} = \theta + \delta_g + \varepsilon_{i(g)},$$

where $g = A, B, i = 1, \dots, n_g, \varepsilon_{i(g)} \stackrel{iid}{\sim} (0, \sigma^2)$ and $\sum n_g \delta_g = 0$.

 n_A = sample size of group **A**, $\theta_A = \theta + \delta_A$ = population location of group **A**, n_B and $\theta_B = \theta + \delta_B$ are similarly defined for group **B**.

Hypotheses: $H_0 : \theta_A - \theta_B = \theta_0$ and $H_a : \theta_A - \theta_B [> \text{ or } < \text{ or } \neq] \theta_0$.

Test statistic's distribution under H_0 :

$$Z = \frac{\sum_{i=1}^{n_B} R_{i(g)} - [n_B(n_A + n_B + 1)/2]}{\sqrt{n_A n_B (n_A + n_B + 1)/12}},$$

where $R_{i(g)}$ denotes the global rank of the i -th observation of group g .

Discussion - Mann-Whitney-Wilcoxon test

- R function: `wilcox.test(x = ... , y = ... , alternative = ...)`.
- This test is "robust" in the sense that (unlike the t-tests) it is not overly affected by outliers.
- For the Mann-Whitney-Wilcoxon test to be comparable to the t-tests (i.e. testing for the mean) we need to assume: (1) The distributions are symmetric, (2) the variances are the same. Then, we have $\theta_A = \mu_A$ and $\theta_B = \mu_B$.
- Compared to the t-tests, the Mann-Whitney-Wilcoxon test is less powerful if their requirements (Gaussian and possibly same variances) are met.
- The distribution of this method under the null is complicated and can be obtained by different methods (e.g. exact, asymptotic normal, ...). The details are beyond the scope of this class.

Comparing diets A and B

Graph

Import

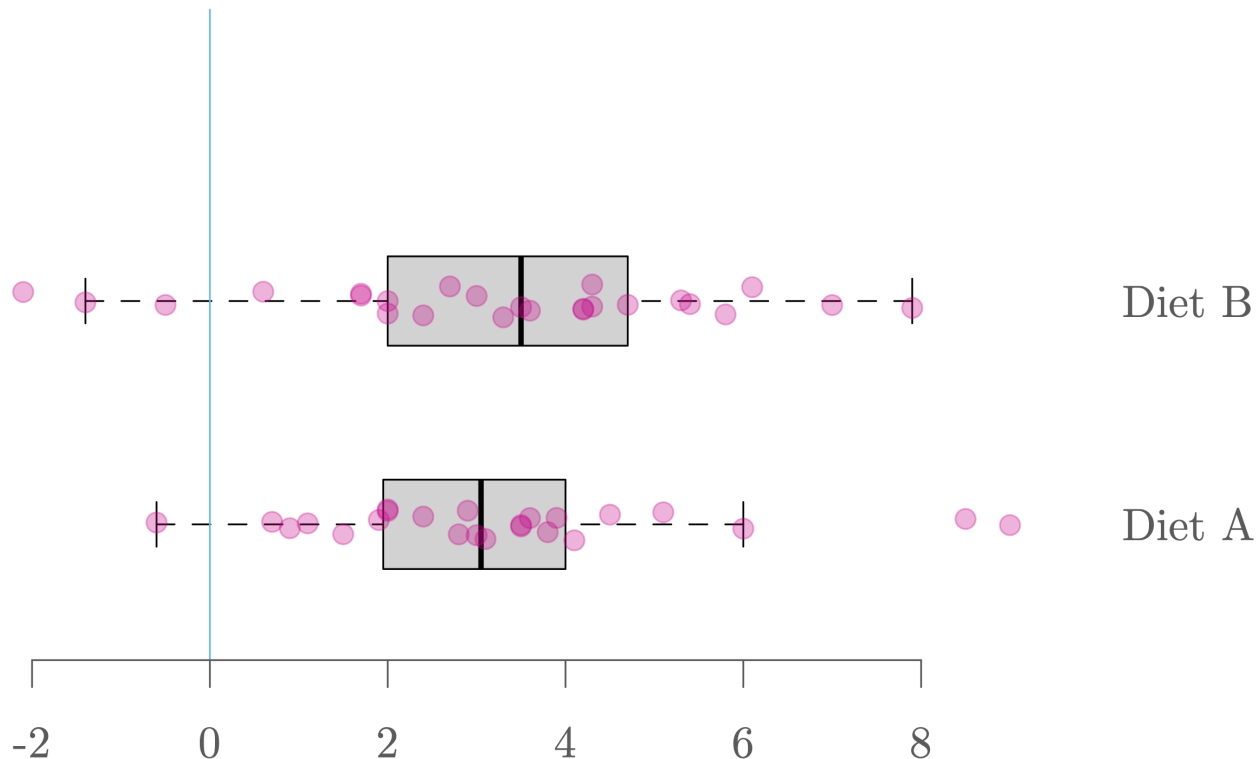
Student

Welch

Wilcox


Results

Weight loss



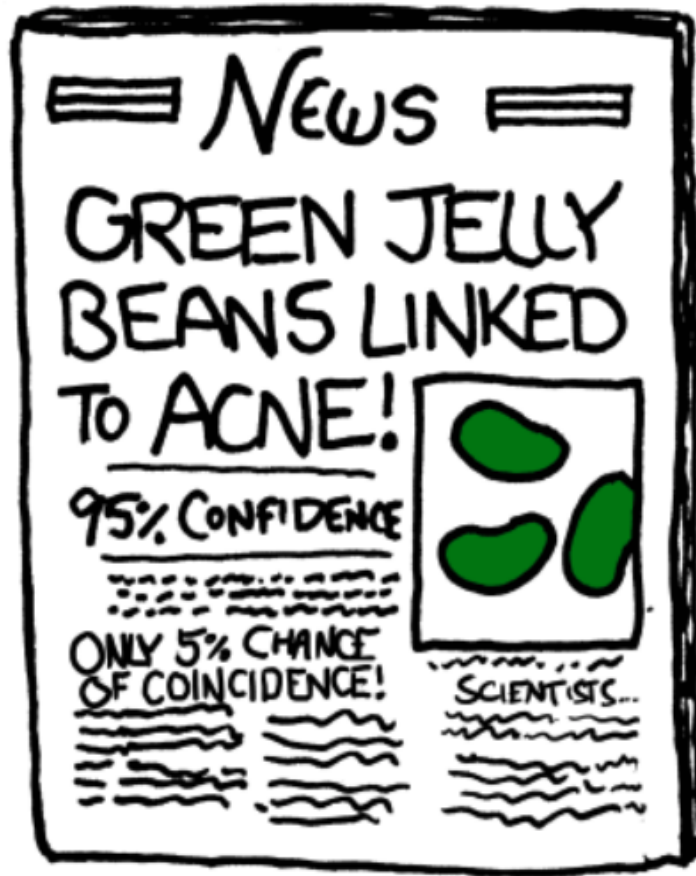
Problems with multiple samples

In practice, we often even encounter situations where we need to **compare the means of more than 2 groups**. For example, we want to compare the weight loss efficacy of several diets, say diets **A**, **B**, **C**. Your theory could, for example, be the following: $0 < \mu_A = \mu_B < \mu_C$. A possible approach to evaluate its validity:

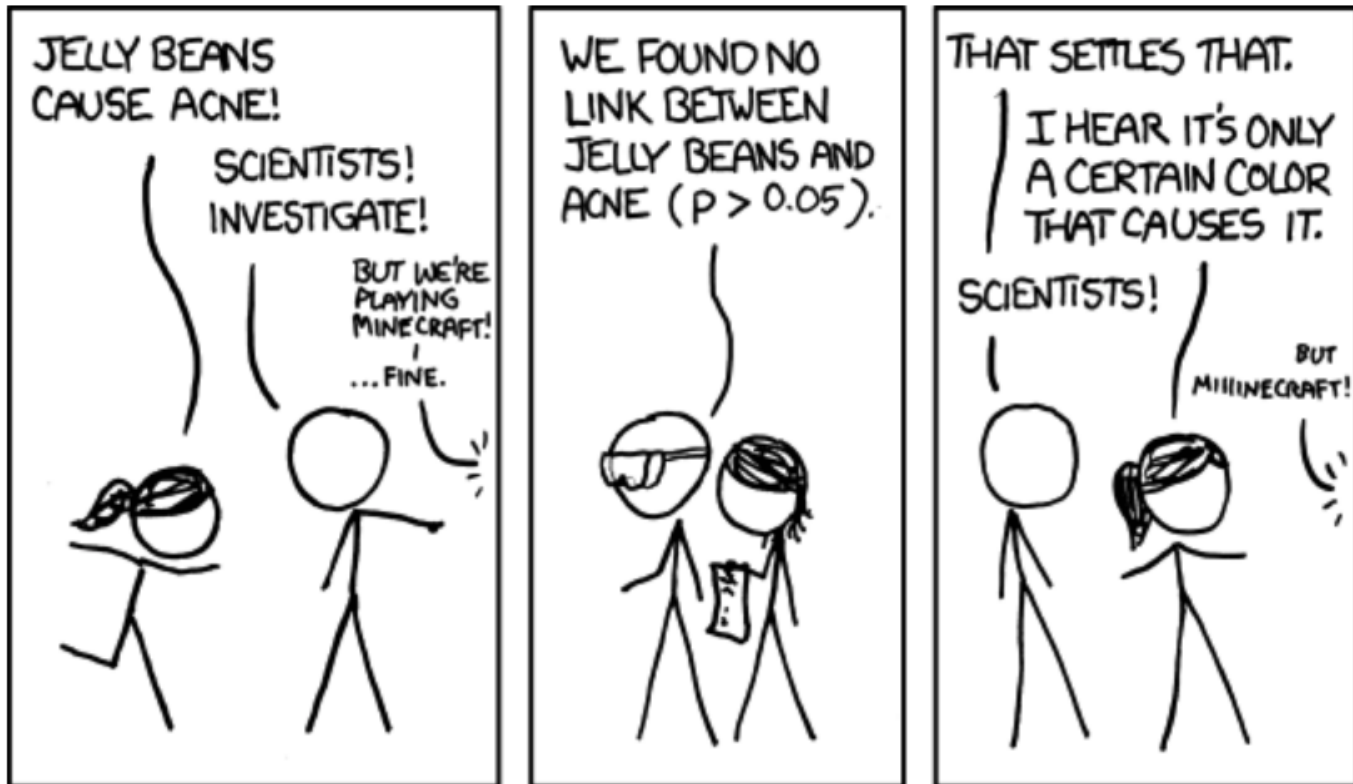
1. Show that μ_C is greater than μ_A and μ_B (i.e. Test 1: $H_0: \mu_A = \mu_C$, $H_a: \mu_A < \mu_C$; Test 2: $H_0: \mu_B = \mu_C$, $H_a: \mu_B < \mu_C$). Here we hope to reject H_0 in both cases.
2. Show that μ_A and μ_B are greater than 0 (i.e. Test 3: $H_0: \mu_A = 0$, $H_a: \mu_A > 0$; Test 4: $H_0: \mu_B = 0$, $H_a: \mu_B > 0$). Here we also hope to reject H_0 in both cases.
3. Compare μ_A and μ_B (i.e. Test 5: $H_0: \mu_A = \mu_B$, $H_a: \mu_A \neq \mu_B$). Here we hope not to reject H_0 .  This does not imply that $\mu_A = \mu_B$ is true but at least the result would not contradict our theory.

Is there a problem in doing many tests?

Are jelly beans causing acne? Maybe... but why only green ones? 🤔

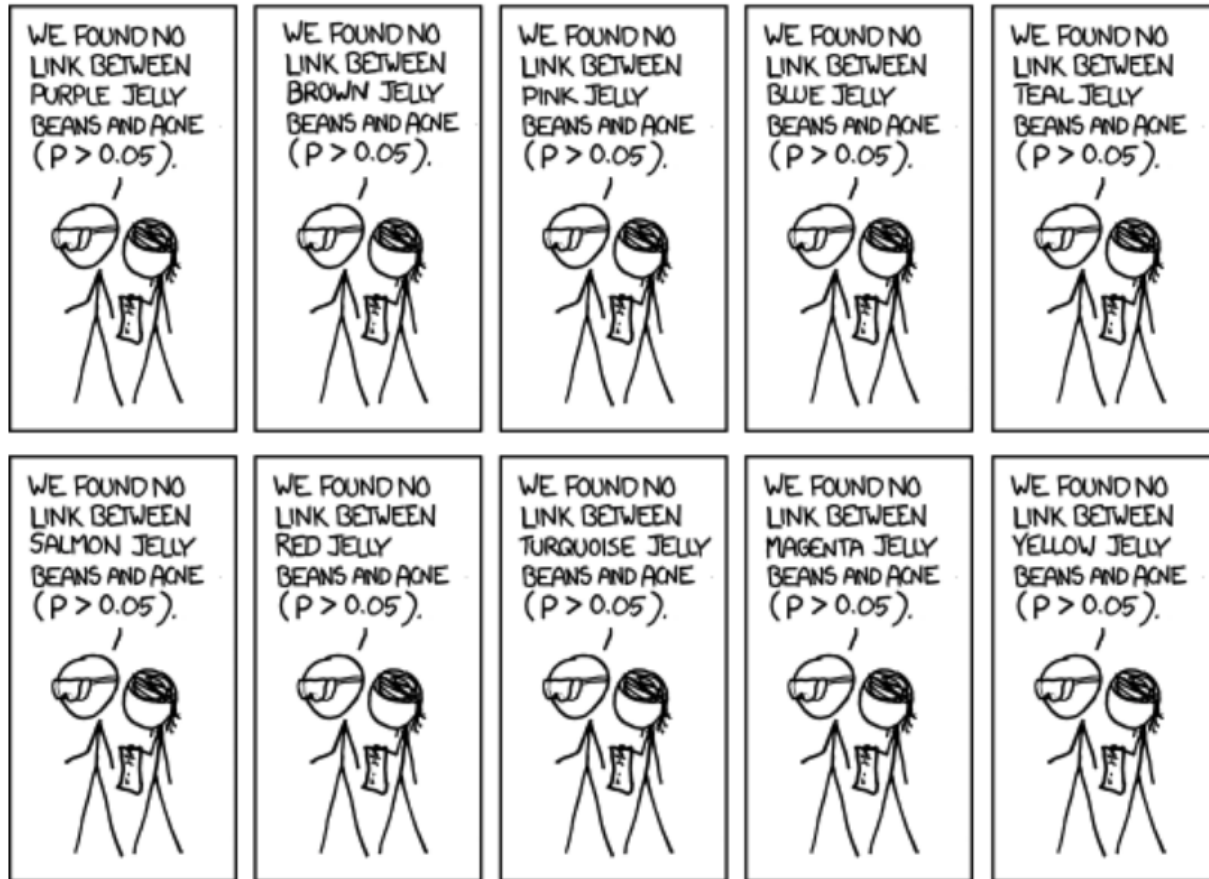


Are jelly beans causing acne?



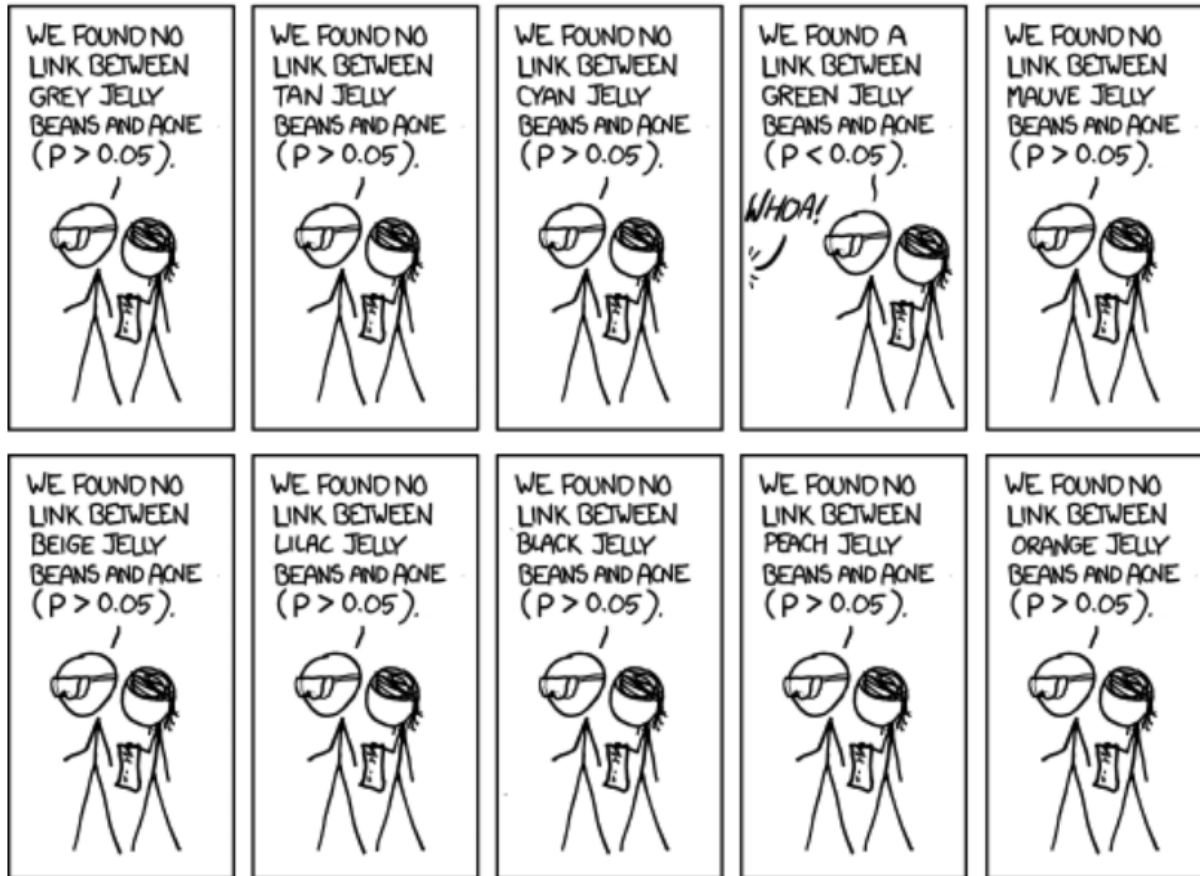
Source: [xkcd](#)

Maybe a specific color?



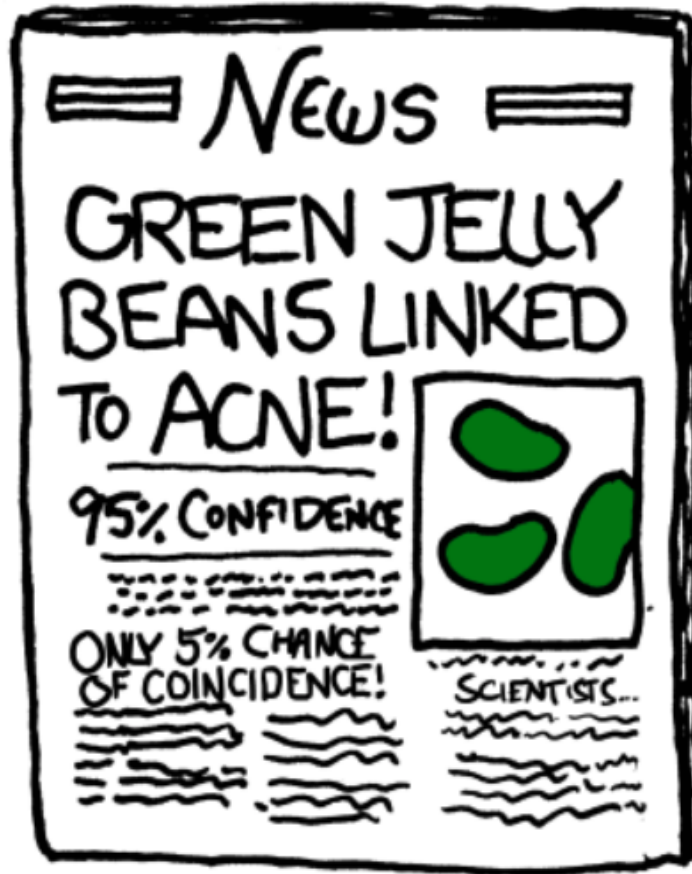
Source: [xkcd](#)

Maybe a specific color?



Source: [xkcd](#)

And finally...



Source: [xkcd](#)



If you want to know more about this comics have a look [here](#).

Multiple testing can be dangerous!

- Remember that a p-value is **random** as its value depends on the data.
- If multiple hypotheses are tested, the chance of observing a rare event increases, and therefore, the chance to incorrectly reject a null hypothesis (i.e. making a Type I error) increases.
- For example, if we consider k (independent) tests (whose null hypotheses are all correct), we have

$$\begin{aligned}\alpha_k &= \Pr(\text{reject } H_0 \text{ at least once}) \\ &= 1 - \Pr(\text{not reject } H_0 \text{ test 1}) \times \dots \times \Pr(\text{not reject } H_0 \text{ test } k) \\ &= 1 - (1 - \alpha) \times \dots \times (1 - \alpha) = 1 - (1 - \alpha)^k\end{aligned}$$

- Therefore, α_k increases rapidly with k (e.g. $\alpha_1 = 0.05$, $\alpha_2 \approx 0.098$, $\alpha_{10} \approx 0.4013$, $\alpha_{100} \approx 0.9941$).
- Hence **performing multiple tests, with the same or different data, is dangerous** ⚠️ (but very common! 😞) as it can lead to significant results, when actually there are none!

Possible solutions

Suppose that we are interested in making k tests and that we want the probability of rejecting the null at least once (assuming the null hypotheses to be correct for all tests) α_k to be equal to α (typically 5%). Instead of using α for the individual tests we will use α_c (i.e. a corrected α). Then, for k (potentially dependent) tests we have

$$\begin{aligned}\alpha_k &= \alpha = \Pr(\text{reject } H_0 \text{ at least once}) \\ &= \Pr(\text{reject } H_0 \text{ test 1 OR } \dots \text{ OR reject } H_0 \text{ test } k) \\ &\leq \sum_{i=1}^k \Pr(\text{reject } H_0 \text{ test } i) = \alpha_c \times k.\end{aligned}$$

Solving for α_c we obtain: $\alpha_c = \alpha/k$, which is called **Bonferroni correction**. By making use of the **Boole's inequality**, this approach does not require any assumptions about dependence among the tests or about how many of the null hypotheses are true.

Possible solutions

The Bonferroni correction can be conservative if there are a large number of tests, as it comes at the cost of reducing the power of the individual tests (e.g. if $\alpha = 5\%$ and $k = 20$, we get $\alpha_c = 0.05/20 = 0.25\%$). There exists a (slightly) "tighter" bound for α_k , which is given by

$$\alpha_k = \Pr(\text{reject } H_0 \text{ at least once}) \leq 1 - (1 - \alpha_c)^k.$$

Solving for α_c we obtain: $\alpha_c = 1 - (1 - \alpha)^{1/k}$, which is called **Dunn-Šidák correction**. This correction is (slightly) less stringent than the Bonferroni correction (since $1 - (1 - \alpha)^{1/k} > \alpha/k$ for $k \geq 2$).

There exist many other alternative methods for multiple testing corrections. It is important to mention that when k is large (say > 100) the Bonferroni and Dunn-Šidák corrections become inapplicable and methods based on the idea of **False Discovery Rate** should be preferred. However, these recent methods are beyond the scope of this class.

Multiple-sample location tests

To compare several means of different populations, a standard approach is to start our analysis by using the **multiple-sample location tests**. More precisely, we proceed our analysis with the following steps:

- **Step 1:** We first perform the multiple-sample location tests, where the null hypothesis states that all the locations are the same. If we cannot reject the null hypothesis, we stop our analysis here. Otherwise, we move on to Step 2.
- **Step 2:** We compare the groups mutually (using α_c) with two-sample location tests in order to verify our hypothesis.

We will discuss three **multiple-sample location tests**:

1. Fisher's one-way ANalysis Of VAriance (ANOVA)
2. Welch's one-way ANOVA
3. Kruskal-Wallis test

Fisher's one-way ANOVA

This test considers the following assumed model for G groups

$$X_{i(g)} = \mu_g + \varepsilon_{i(g)} = \mu + \delta_g + \varepsilon_{i(g)},$$

where $g = 1, \dots, G$, $i = 1, \dots, n_g$, $\varepsilon_{i(g)} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and $\sum n_g \delta_g = 0$.



n_i = sample size of group i , $\mu_i = \mu + \delta_i$ = population mean of group i , $i = 1, \dots, G$.

Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_G \quad \text{and} \quad H_a : \mu_i \neq \mu_j \text{ for at least one pair of } (i, j).$$

Test statistic's distribution under H_0 :

$$F = \frac{N s_{\bar{X}}^2}{s_p^2} \underset{H_0}{\sim} \text{Fisher}(G-1, N-G), \text{ where } s_{\bar{X}}^2 = \frac{1}{G-1} \sum_{g=1}^G \frac{n_g}{N} (\bar{X}_g - \bar{X})^2,$$

$$s_p^2 = \frac{1}{N-G} \sum_{g=1}^G (n_g - 1) s_g^2, \quad N = \sum_{g=1}^G n_g, \text{ and } \bar{X} = \frac{1}{N} \sum_{g=1}^G n_g \bar{X}_g.$$

Discussion - Fisher's one-way ANOVA

- R function:

```
aov(response ~ groups, data = mydata).
```

- This test strongly relies on the **assumed absence of outliers**. If outliers appear to be present the Kruskal-Wallis test (see later) is (probably) a better option.
- For moderate and small sample sizes, the sample distribution should be at least **approximately normal** with no strong skewness to ensure the reliability of the test.
- In practice, the assumption of equal variance is hard to verify so **we recommend to avoid this test in practice**.

Welch's one-way ANOVA

This test considers the following assumed model for G groups

$$X_{i(g)} = \mu_g + \varepsilon_{i(g)} = \mu + \delta_g + \varepsilon_{i(g)},$$

where $g = 1, \dots, G$, $i = 1, \dots, n_g$, $\varepsilon_{i(g)} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_g^2)$ and $\sum n_g \delta_g = 0$.

Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_G \quad \text{and} \quad H_a : \mu_i \neq \mu_j \text{ for at least one pair of } (i, j).$$

Test statistic's distribution under H_0 :

$$F^* = \frac{s_{\bar{X}}^{*2}}{1 + \frac{2(G-2)}{3\Delta}} \underset{H_0}{\sim} \text{Fisher}(G-1, \Delta),$$

where $s_{\bar{X}}^{*2} = \frac{1}{G-1} \sum_{g=1}^G w_g (\bar{X}_g - \bar{X}^*)^2$, $\Delta = [\frac{3}{G^2-1} \sum_{g=1}^G \frac{1}{n_g} (1 - \frac{w_g}{\sum_{g=1}^G w_g})]^{-1}$, $w_g = \frac{n_g}{s_g^2}$, and

$$\bar{X}^* = \sum_{g=1}^G \frac{w_g \bar{X}_g}{\sum_{g=1}^G w_g}.$$

Discussion - Welch's one-way ANOVA

- R function:

```
oneway.test(response ~ groups, data = mydata).
```


- This test strongly relies on the **assumed absence of outliers**. If outliers appear to be present the Kruskal-Wallis test (see later) is (probably) a better option.
- For moderate and small sample sizes, the sample distribution should be at least **approximately normal** with no strong skewness to ensure the reliability of the test.
- This test does not require the variances of the groups to be equal. If the variances of all the groups are the same (which is rather unlikely in practice), the Welch's one-way ANOVA losses a little bit of power compared to the Fisher's one-way ANOVA.

Kruskal-Wallis test

This test considers the following assumed model for G groups

$$X_{i(g)} = \theta_g + \varepsilon_{i(g)} = \theta + \delta_g + \varepsilon_{i(g)},$$

where $g = 1, \dots, G$, $i = 1, \dots, n_g$, $\varepsilon_{i(g)} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and $\sum n_g \delta_g = 0$.

 n_i = sample size of group i , $\theta_i = \theta + \delta_i$ = population location of group i ,
 $i = 1, \dots, G$.

Hypotheses:

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_G \quad \text{and} \quad H_a : \theta_i \neq \theta_j \text{ for at least one pair of } (i, j).$$

Test statistic's distribution under H_0 : $H = \frac{\frac{12}{N(N+1)} \sum_{g=1}^G \frac{\bar{R}_g}{n_g} - 3(N+1)}{1 - \frac{\sum_{v=1}^V t_v^3 - t_v}{N^3 - N}} \underset{H_0}{\sim} \chi(G-1)$, where

$\bar{R}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} R_{i(g)}$ with $R_{i(g)}$ denoting the global rank of the i^{th} observation of group g ,
 V is the number of different values/levels in X and t_v denotes the number of times a given value/level occurred in X .

Discussion - Kruskal-Wallis test

- R function:

```
kruskal.test(response ~ groups, data = mydata).
```

- This test is "robust" in the sense that (unlike the one-way ANOVA) it is not overly affected by outliers.
- For the Kruskal-Wallis test to be comparable to the one-way ANOVAs (i.e. testing for the mean) we need to assume: (1) The distributions are symmetric, (2) the variances are the same. Then, we have
 $\theta_i = \mu_i, i = 1, \dots, G.$
- Compared to the one-way ANOVA, the Kruskal-Wallis test is less powerful if their requirements (Gaussian and possibly same variances) are met.

Exercise: Comparing diets A, B and C

Graph

Import

