Analysis of Variance (ANOVA)

Sébastien Biass

sebastien.biasse@unige.ch

Earth Sciences, University of Geneva 🇨🇭

Stéphane Guerrier

Stephane.Guerrier@unige.ch

Earth Sciences & Pharmaceutical Sciences, University of Geneva 🇨🇭

Dominique Couturier

Cancer Research UK, University of Cambridge 🇬🇧

Yuming Zhang

T.H. Chan School of Public Health, Harvard University 🇺🇸

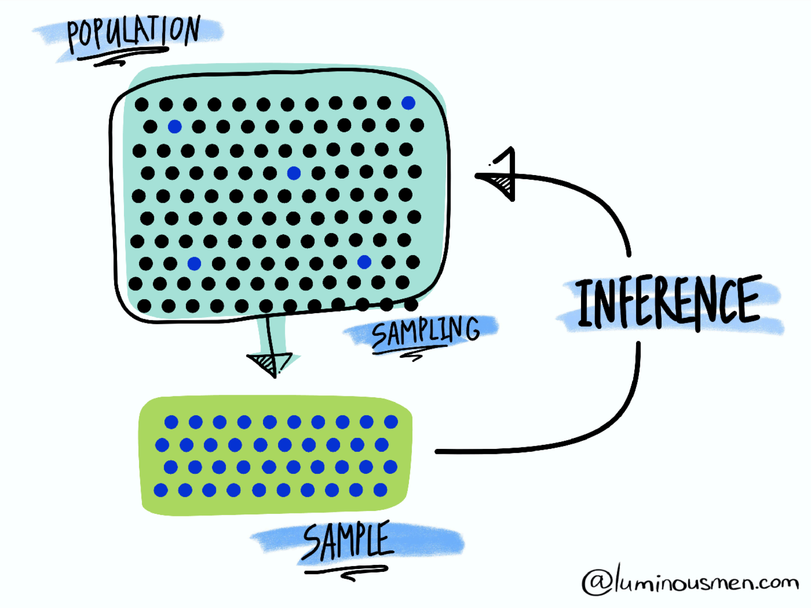
This document was prepared with the help of Lionel Voirol and Filippo Salmaso

How does it work?

Statistical methods are based on several fundamental concepts, the most central of which is to consider the information available (in the form of data) resulting from a random process.

As such, the data represent a random sample of a totally or conceptually accessible population.

Then, statistical inference allows to infer the properties of a population based on the observed sample. This includes deriving estimates and testing hypotheses.



Source: luminousmen

Hypothesis testing

- In general (scientific) hypotheses can be translated into a set of (non-overlapping idealized) statistical hypotheses:

$$H_0 : \theta \in \Theta_0 \quad \text{and} \quad H_a : \theta \notin \Theta_0$$

- In a hypothesis test, the statement being tested is called the null hypothesis H_0. A hypothesis test is designed to assess the strength of the evidence against the null hypothesis.
- The alternative hypothesis H_a is the statement we hope or suspect to be true instead of H_0.
- Each hypothesis excludes the other, so that one can exclude one in favor of the other using the data.
- Example: a drug represses the progression of cancer

$$H_0 : \mu_{\text{drug}} = \mu_{\text{control}} \quad \text{and} \quad H_a : \mu_{\text{drug}} < \mu_{\text{control}}.$$

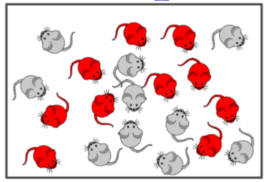
Hypothesis testing

Statistical Hypotheses

H_0: $\mu_{Drug} = \mu_{Control}$

H_a: $\mu_{Drug} < \mu_{Control}$

Sample



Data: Tumour size at day 42

$(x_{D,1}; x_{D,2}; \ldots; x_{D,n_D})$

$(x_{C,1}; x_{C,2}; \ldots; x_{C,n_C})$

Point estimation

$\widehat{\mu}_{Drug} - \widehat{\mu}_{Control}$

Inference: Under H0

$$T_{obs} = \frac{\widehat{\mu}_{Drug} - \widehat{\mu}_{Control}}{s_p \sqrt{\frac{1}{n_D} + \frac{1}{n_C}}} \sim St_{n_D + n_C - 2}$$

Idea:

Drug represses the progression of cancer

Hypothesis testing

Outcome	H_0 is true	H_0 is false
Can't reject H_0	✅ Correct decision (prob=$1-\alpha$)	⚠️ Type II error (prob=$1-\beta$)
Reject H_0	⚠️ Type I error (prob=α)	✅ Correct decision (prob=β)

- The type I error corresponds to the probability of rejecting H_0 when H_0 is true (also called false positive). The type II error corresponds to the probability of not rejecting H_0 when H_a is true (also called false negative).

- A test is of significance level α when the probability of making a type I error equals α. Usually we consider $\alpha = 5\%$, however, this can vary depending on the context.

- A test is of power β when the probability to make a type II error is $1 - \beta$. In other words, the power of a test is its probability of rejecting H_0 when H_0 is false (or the probability of accepting H_a when H_a is true).

What are p-values?

In statistics, the **p-value** is defined as the probability of observing a test statistic that is at least as extreme as actually observed, assuming that the null hypothesis H_0 is true.

Informally, a p-value can be understood as a measure of plausibility of the null hypothesis given the data. A small p-value indicates strong evidence against H_0.

When the p-value is small enough (i.e., smaller than the significance level α), the test based on the null and alternative hypotheses is considered significant, meaning we reject the null hypothesis in favor of the alternative. This is generally what we want because it "verifies" our (research) hypothesis.

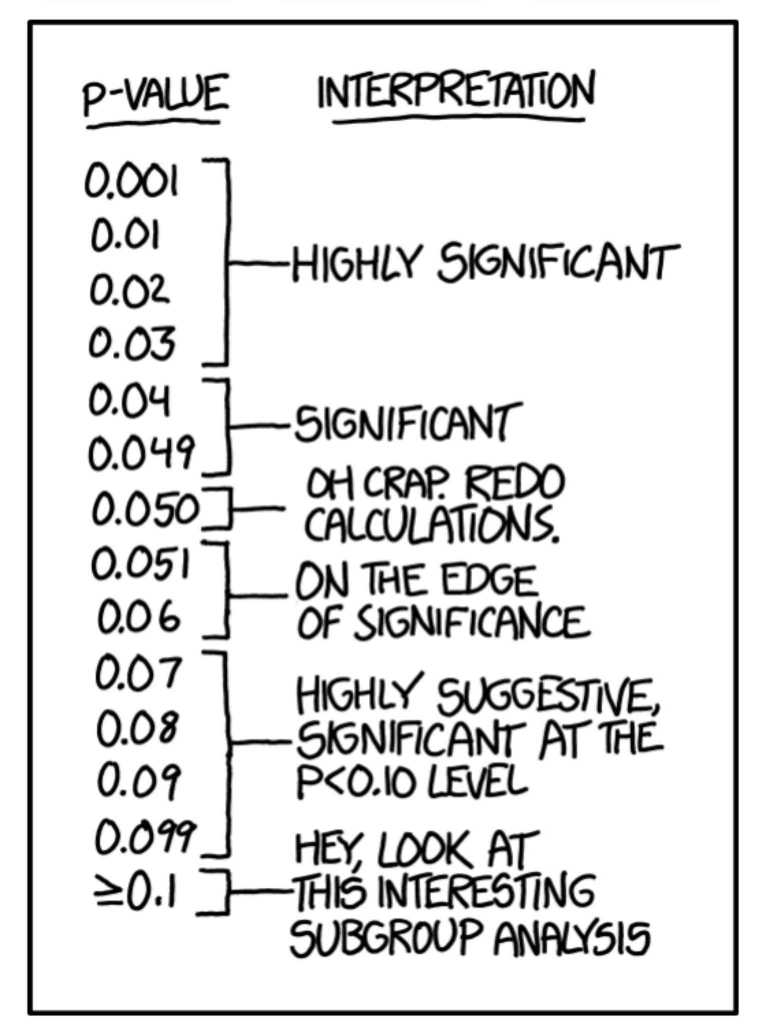
When the p-value is not small enough, with the available data, we cannot reject the null hypothesis, so nothing can be concluded. 🤔

The obtained p-value summarizes the incompatibility between the data and the model constructed under the set of assumptions.

"Absence of evidence is not evidence of absence." 👋

👋 From the British Medical Journal.

How to understand p-values?



👋 If you want to know more have a look here.

This document was prepared with the help of Lionel Voirol and Filippo Salmaso

P-values may be controversial

P-values have been misused many times because understanding what they mean is not intuitive!

👋 If you want to know more have a look here.

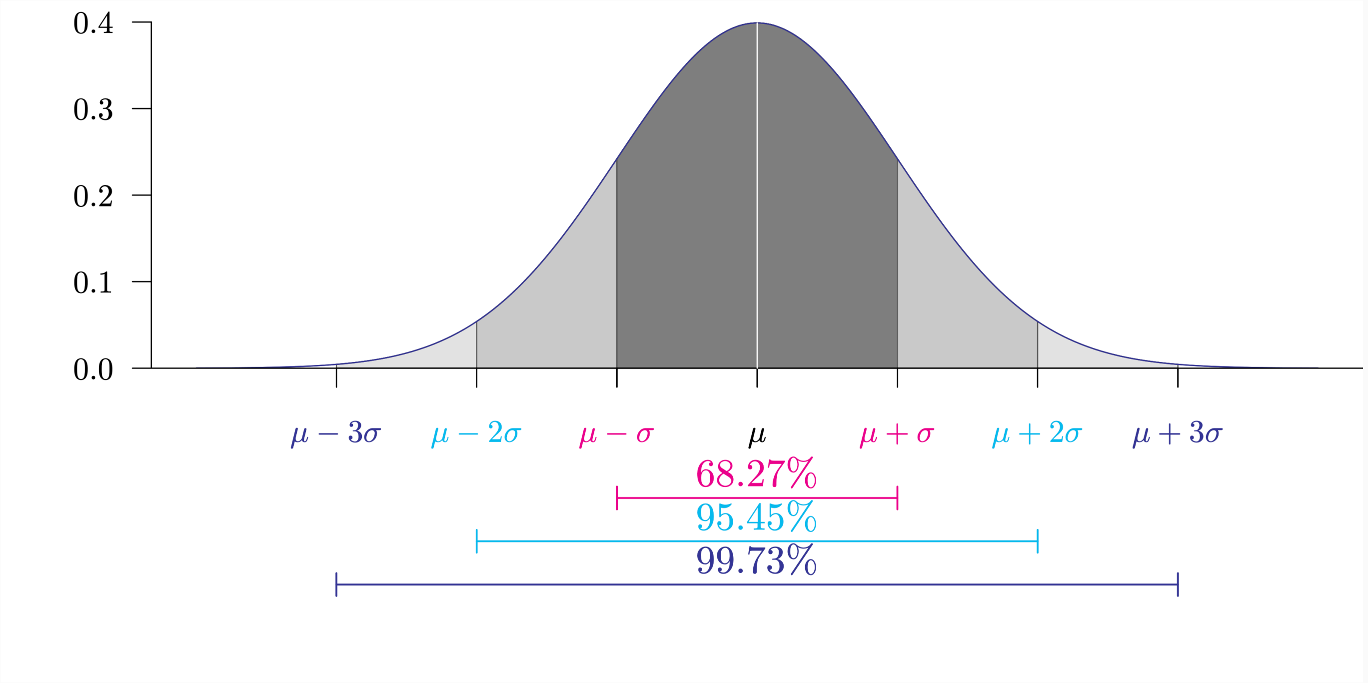
Quick review: Normal distribution

$$Y \sim N(\mu, \sigma^2), \qquad f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

$$\mathbb{E}[Y] = \mu, \qquad \text{Var}[Y] = \sigma^2,$$

$$Z = \frac{Y-\mu}{\sigma} \sim N(0,1), \qquad f_Z(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}}.$$

Probability density function of a normal distribution:



This document was prepared with the help of Lionel Voirol and Filippo Salmaso

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 $g = A, B$, $i = 1, \ldots, n_g$, $\varepsilon_{i(g)} \overset{iid}{\sim} N(0, \sigma^2)$ 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:

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

Test statistic's distribution under H_0:

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

Discussion — Student's t-test

Python function:

```
from scipy import stats

stats.ttest_ind(a = ..., b = ..., alternative = ..., equal_var = 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**.

This document was prepared with the help of Lionel Voirol and Filippo Salmaso

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 $g = A, B$, $i = 1, \ldots, n_g$, $\varepsilon_{i(g)} \overset{iid}{\sim} N(0, \sigma_g^2)$ 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:

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

Test statistic's distribution under H_0:

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

Discussion — Welch's t-test

Python function:

```
from scipy import stats

stats.ttest_ind(a = ..., b = ..., alternative = ..., equal_var = False)
```

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 loses 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 $g = A, B$, $i = 1, \ldots, n_g$, $\varepsilon_{i(g)} \overset{iid}{\sim} N(0, \sigma^2)$ and $\sum n_g \delta_g = 0$.

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

Hypotheses:

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

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

Python function:

```
from scipy import stats

stats.mannwhitneyu(a = ..., b = ...)
```

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 **symmetric distributions** and **equality in variances** .

Then, we have $\theta_A = \mu_A$ and $\theta_B = \mu_B$.

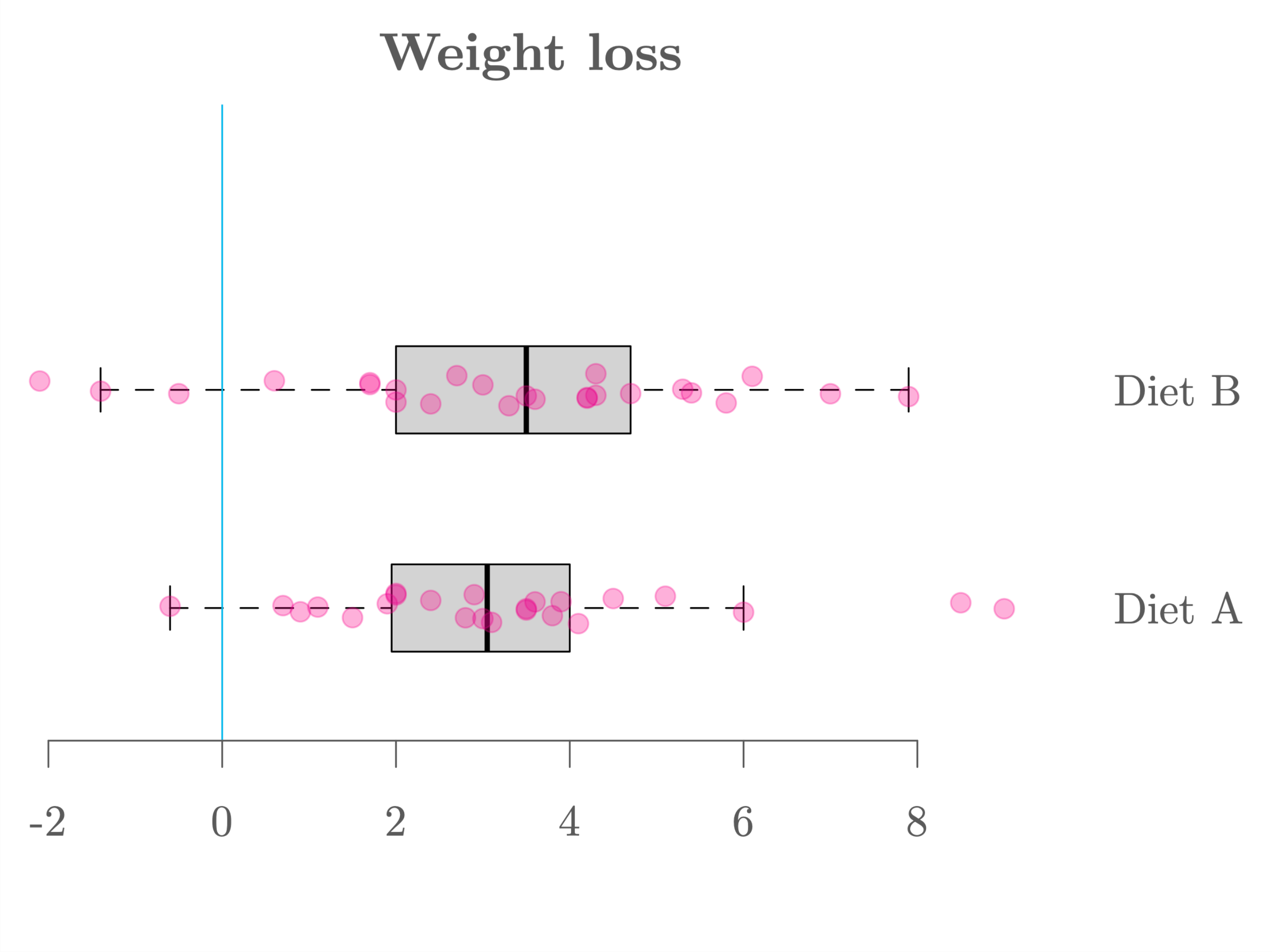
Compared to the t-tests, the Mann–Whitney–Wilcoxon test is less powerful if their requirements (Gaussian and possibly same variances) are met.

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.

This document was prepared with the help of Lionel Voirol and Filippo Salmaso

Comparing diets A and B

Graph



This document was prepared with the help of Lionel Voirol and Filippo Salmaso

Import

```
# Import data
import pandas as pd
diet = pd.read_csv("https://raw.githubusercontent.com/ELSTE-Master/Data-Science/main/Data/diet.
n = 5
print(diet.head(n)) #print the first n rows of the dataset
```

```
   id  gender  age  height diet.type  initial.weight  final.weight
0   1  Female   22     159         A              58          54.2
1   2  Female   46     192         A              60          54.0
2   3  Female   55     170         A              64          63.3
3   4  Female   33     171         A              64          61.1
4   5  Female   50     170         A              65          62.2
```

```

# Compute weight loss
diet["weight.loss"] = diet["initial.weight"] - diet["final.weight"]

# Select diet
posA = diet["diet.type"] == "A"
posB = diet["diet.type"] == "B"

# Variable of interest
dietA = diet["weight.loss"][diet["diet.type"]=="A"]
dietB = diet["weight.loss"][diet["diet.type"]=="B"]
```

Student

```python
from scipy.stats import ttest_ind

t_stat, p_val = ttest_ind(dietA, dietB, alternative="two-sided", equal_var=True)

print("t-statistic:", round(t_stat, 4))
```

```
t-statistic: 0.0475
```

```python
print("p-value:", round(p_val, 4))
```

```
p-value: 0.9623
```

Welch

```
from scipy import stats

# Welch's t-test (equal_var=False makes it Welch)
t_stat, p_value = stats.ttest_ind(dietA, dietB, equal_var=False)

print("t-statistic:", round(t_stat, 4))
```

```
t-statistic: 0.0476
```

```
print("p-value:", round(p_value, 4))
```

```
p-value: 0.9622
```

Wilcox

```python
from scipy import stats

# Wilcoxon rank-sum test (Mann–Whitney U)
stat, p_value = stats.mannwhitneyu(dietA, dietB)

print("Wilcoxon statistic:", round(stat, 4))
```

```
Wilcoxon statistic: 277.0
```

```python
print("p-value:", round(p_value, 4))
```

```
p-value: 0.6526
```

Results

- Step 1: Define hypotheses $H_0 : \mu_A = \mu_B, \quad H_a : \mu_A \neq \mu_B$

- Step 2: Define α We consider $\alpha = 5\%$

- Step 3: Compute p-value
  Welch's t-test appears suitable in this case, and therefore, we obtain $\text{p-value} = 96.22\%$ (see Python output tab for details).

- Step 4: Conclusion
  We have $\text{p-value} > \alpha$ so we fail to reject the null hypothesis at the significance level of 5%.

Problems with multiple samples

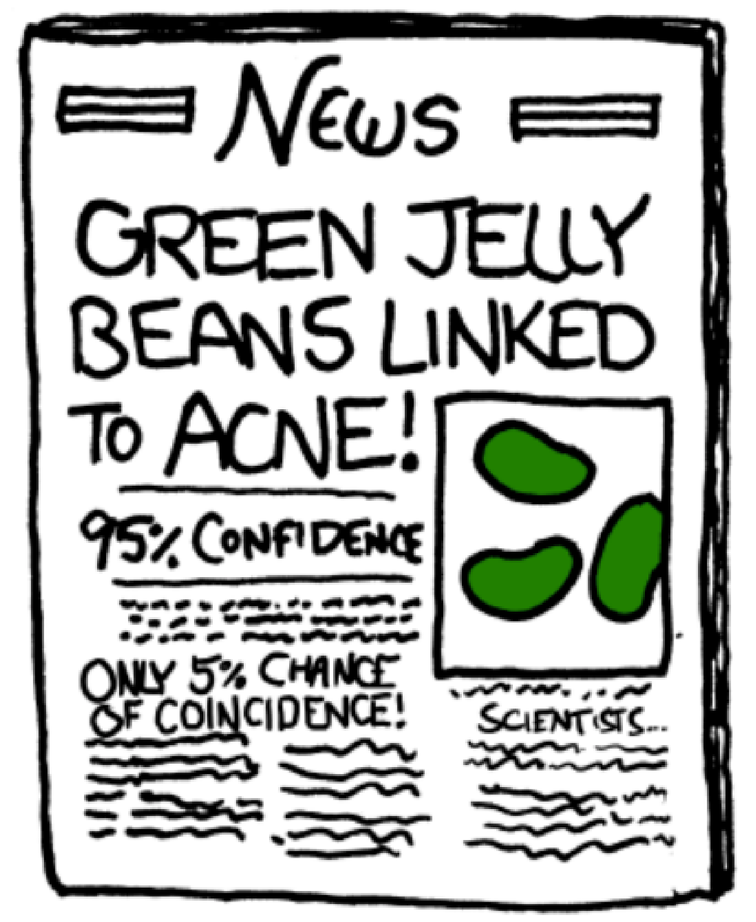
In practice, we often even encounter situations where we need to compare the means of more than 2 groups. For example, we want to compare the weight loss efficacy of several diets, say diets **A**, **B**, **C**. Your theory could, for example, be the following: $0 < \mu_A = \mu_B < \mu_C$. A possible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

1. Show that μ_C is greater than μ_A and μ_B (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 (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 (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.

This document was prepared with the help of Lionel Voirol and Filippo Salmaso

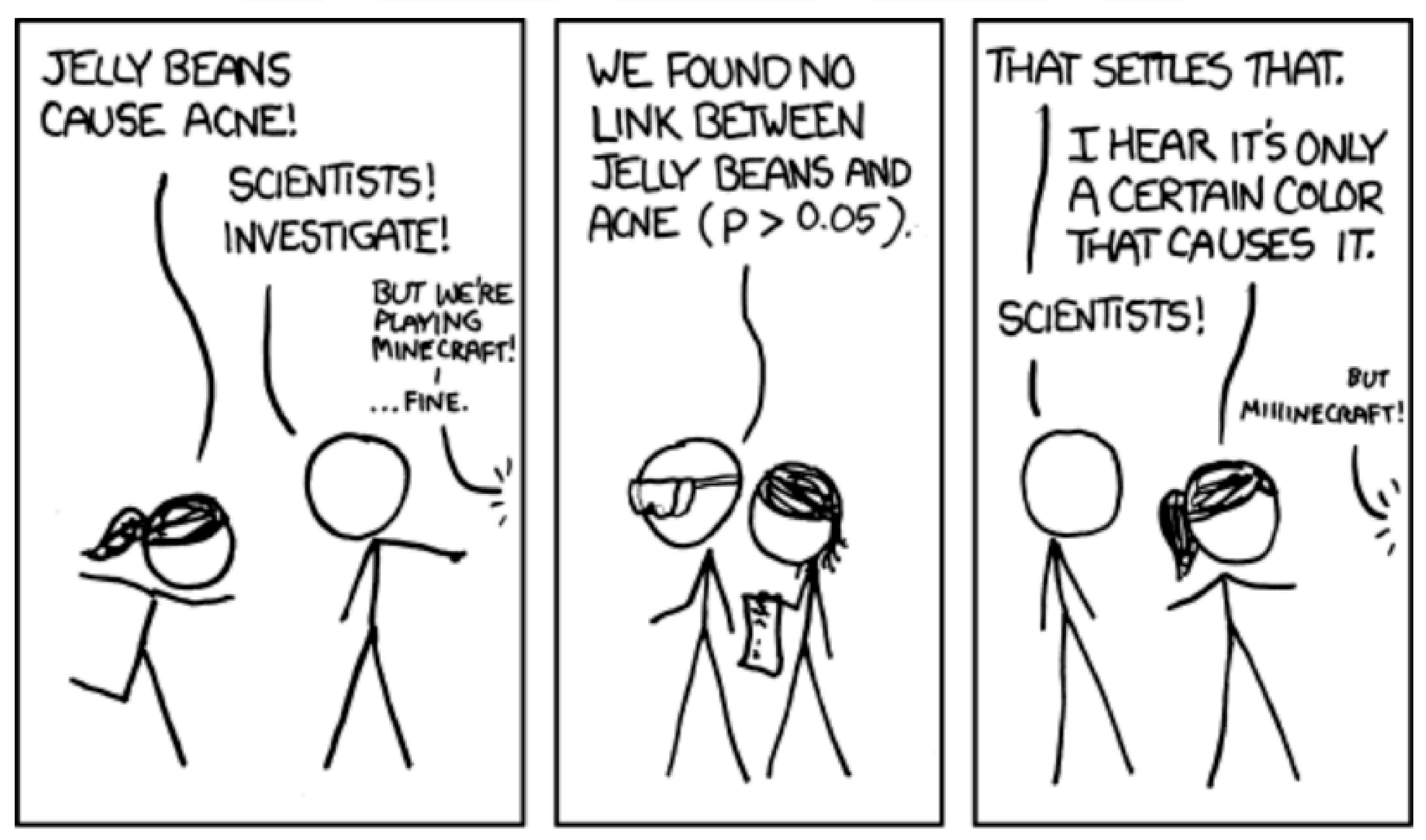
Is there a problem in doing many tests?

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



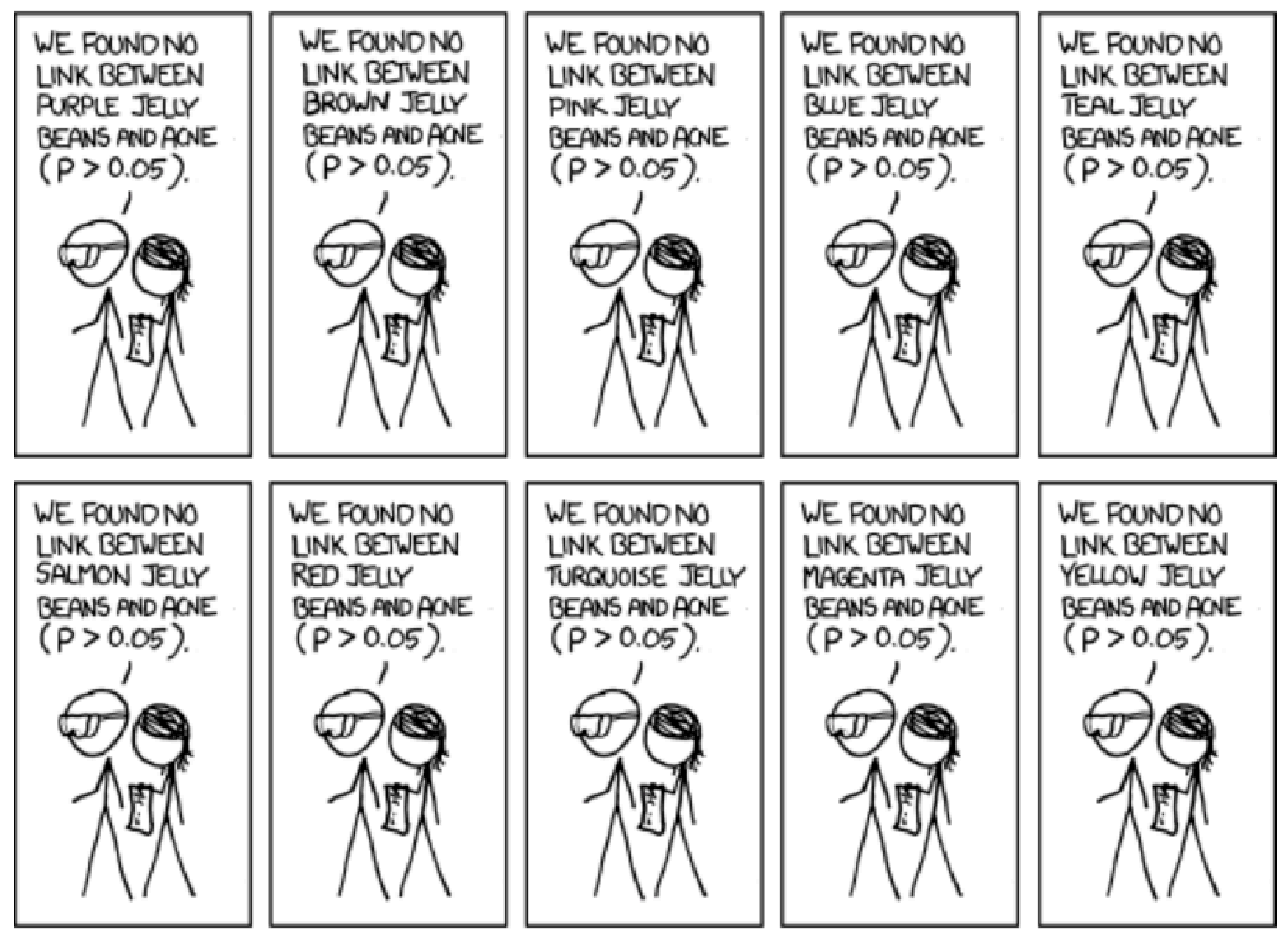
Source: xkcd

Are jelly beans causing acne?



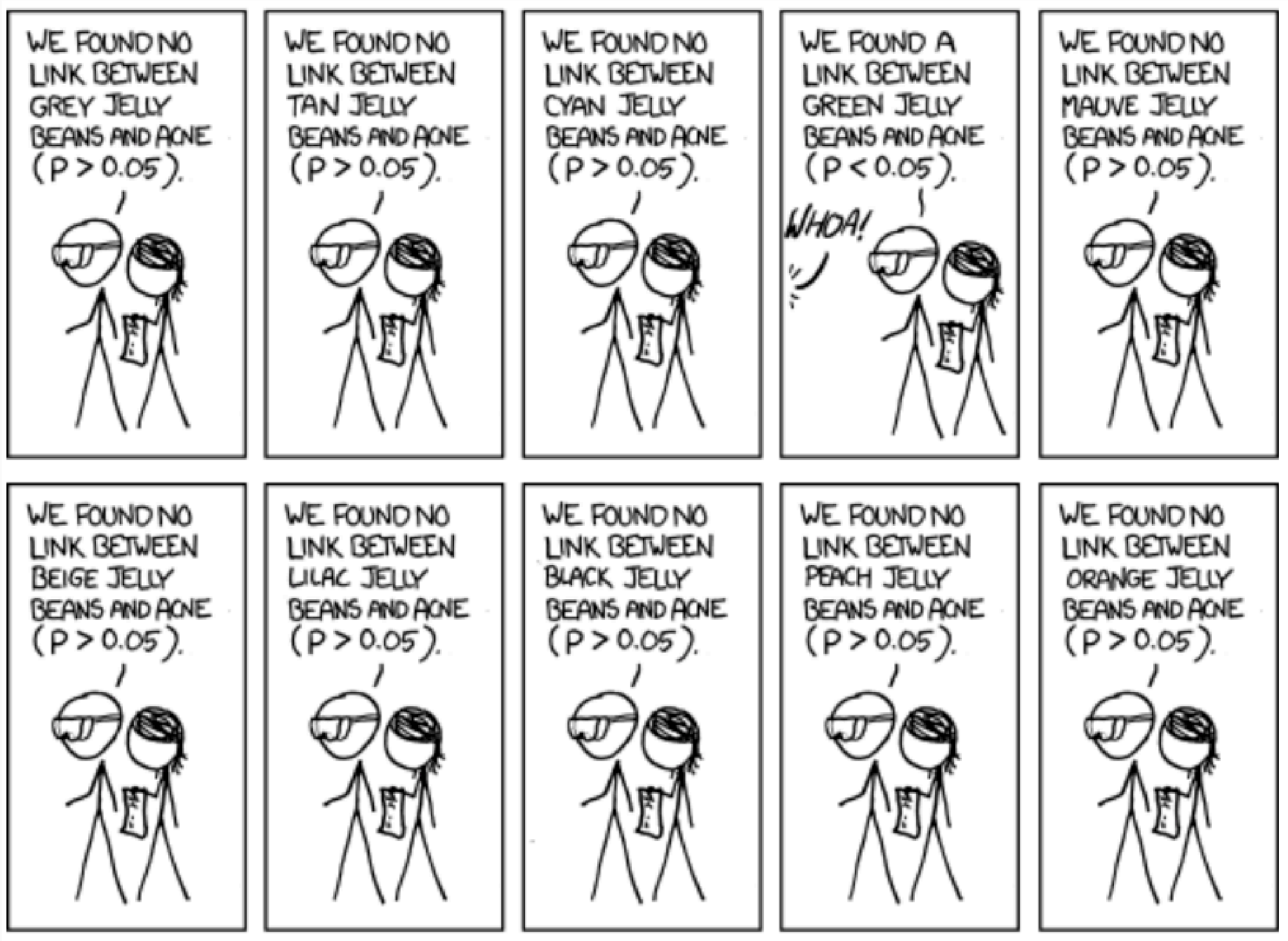
Source: xkcd

Maybe a specific color?



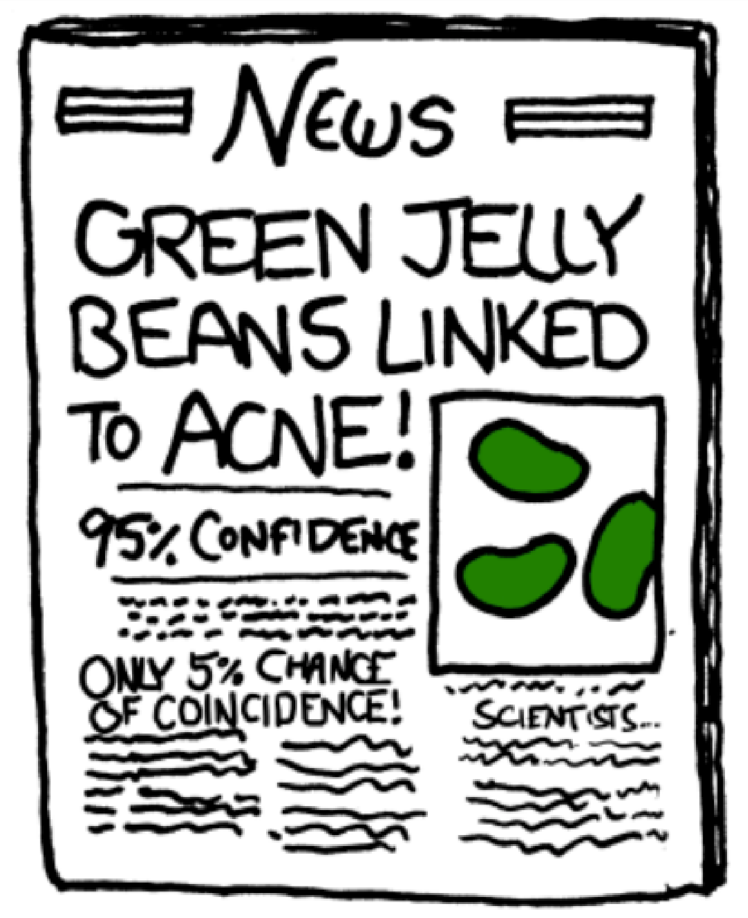
Source: xkcd

Maybe a specific color?



Source: xkcd

And finally...



Source: xkcd

👋 If you want to know more about this comic, 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

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

- 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 $g = 1, \dots, G$, $i = 1, \dots, n_g$, $\varepsilon_{i(g)} \overset{iid}{\sim} N(0, \sigma^2)$ 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:

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

Test statistic's distribution under H_0:

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

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

Discussion - Fisher's one-way ANOVA

Python function:

```
from scipy import stats

stats.f_oneway(group_A, group_B, group_C)
```

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

Hypotheses:

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

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

Python function:

```
from statsmodels.stats.oneway import anova_oneway

anova_oneway(data, groups=groups, use_var='unequal', welch_correction=True)
```

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 loses 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 $g = 1, \ldots, G$, $i = 1, \ldots, n_g$, $\varepsilon_{i(g)} \overset{iid}{\sim} N(0, \sigma^2)$ and $\sum n_g \delta_g = 0$.

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

Hypotheses:

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

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

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

Python function:

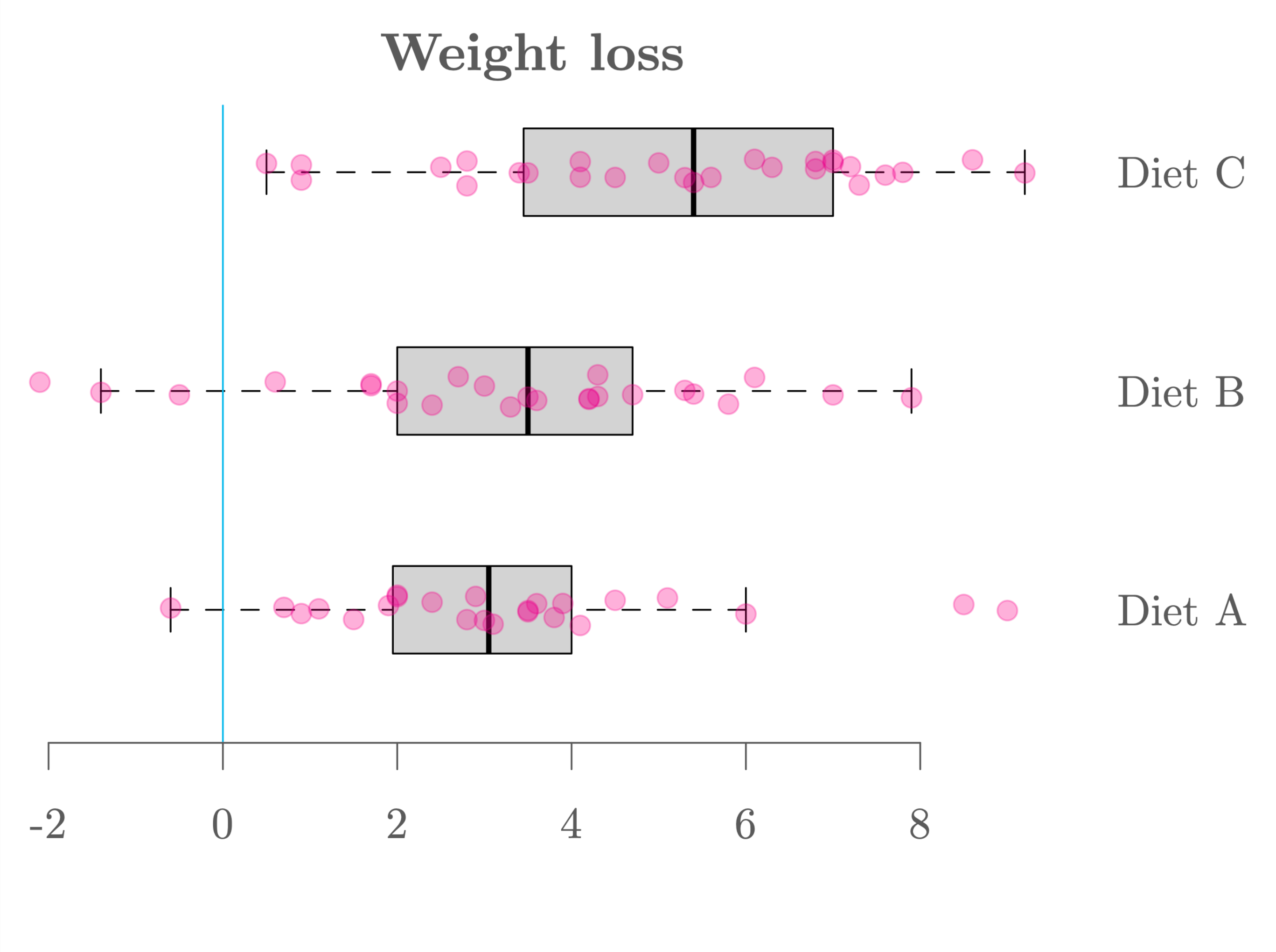
```
from scipy import stats

stats.kruskal(group_A, group_B, group_C)
```

- 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 ANOVAs, 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



Exercise: Comparing diets A, B and C

Import

```
# Import data
import pandas as pd
diet = pd.read_csv("https://raw.githubusercontent.com/ELSTE-Master/Data-Science/main/Data/
diet["weight.loss"] = diet["initial.weight"] - diet["final.weight"]

# Variable of interest
dietA = diet["weight.loss"][diet["diet.type"]=="A"]
dietB = diet["weight.loss"][diet["diet.type"]=="B"]
dietC = diet["weight.loss"][diet["diet.type"]=="C"]

# Create data frame
dat = pd.DataFrame({
    "response": list(dietA) + list(dietB) + list(dietC),
    "groups": (["A"] * len(dietA)) + (["B"] * len(dietB)) + (["C"] * len(dietC))
})
```

This document was prepared with the help of Lionel Voirol and Filippo Salmaso