Introduction to Data Analytics

Part I: Introduction to Statistical Inference

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 🇨🇭

R and RStudio

In this class, we will use the statistical software **R** together with the integrated development environment **RStudio**, which can be downloaded with the following:

- Latest version of R:
  https://cran.r-project.org/
- Latest version of RStudio:
  https://www.rstudio.com/

Note: You cannot use RStudio without having installed R on your computer.


Source: allisonhorst

R package idarps

We have developed an `R` package for this class which is called `idarps` (which stands for Introduction to Data Analytics with R for Pharmaceutical Sciences) and that you can install by running the following code in your `R` console:

```
install.packages("idarps")
```

Note: Find the GitHub repository of the package here.

What is statistics?

Statistics is a science that uses mathematics and computer science to deal with the collection, analysis, interpretation, and presentation of masses of numerical data. Informally, it is the science of learning from data.
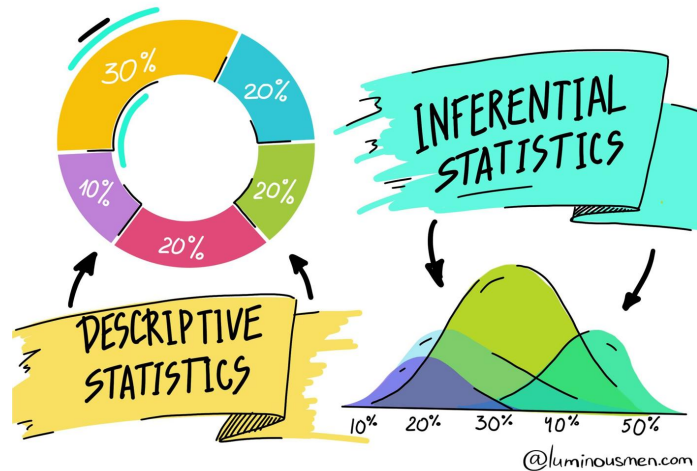
Statistics is a crucial part of our life. However, statistical methods are often consciously (or not) misused. This can lead to contradictory studies and conclusions (as seen during the current COVID-19 pandemic).

Source: luminousmen

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Source: Atoz Markets

How can statistics be useful?

Statistics can be used (among others) to

1. Visualize data (e.g. propagation of COVID-19 in different countries).
2. Understand and interpret data (e.g. main causes of cancer).
3. Assess the validity of a hypothesis (e.g. is a drug working?).
4. Make predictions (e.g. predicting unemployment or risk indices).

Learning more about statistics allows to

1. Better understand arguments based on data.
2. Be able to apply critical thinking about statistics used as evidence.
3. Understand how statistical associations are used to evaluate claims (hypotheses) and assess causal connections.

Understanding and knowing how to interpret statistical analyses is therefore becoming an increasingly vital skill.

How to test a (scientific) hypothesis?

"In god we trust, all others must bring data." 👋

- To assess the validity of a (scientific) hypothesis, the scientific community (generally) agrees on a specific procedure.
- These hypotheses can be nearly anything, such as:
    1. Coffee consumption increases blood pressure.
    2. Republican politicians are bad/good for the American Economy.
    3. A glass of red wine is as good as an hour at the gym.
- This procedure involves the design of an experiment and then the collection of data to compute a metric, called **p-value**, which evaluates the adequacy between the data and your original hypothesis.
- There is generally a specific threshold (typically 5%), and if the p-value falls below this threshold we can claim that we have statistically significant result(s) validating our hypothesis.

👋 From W. Edwards Deming.

Statistics vs Truth 🤥

- Statistically significant results are not necessarily the truth, as there isn't a threshold (e.g. 5%) that separates real results from the false ones.
- This procedure simply provides us with one piece of a puzzle that should be considered in the context of other evidence.

Everything we eat both causes and prevents cancer
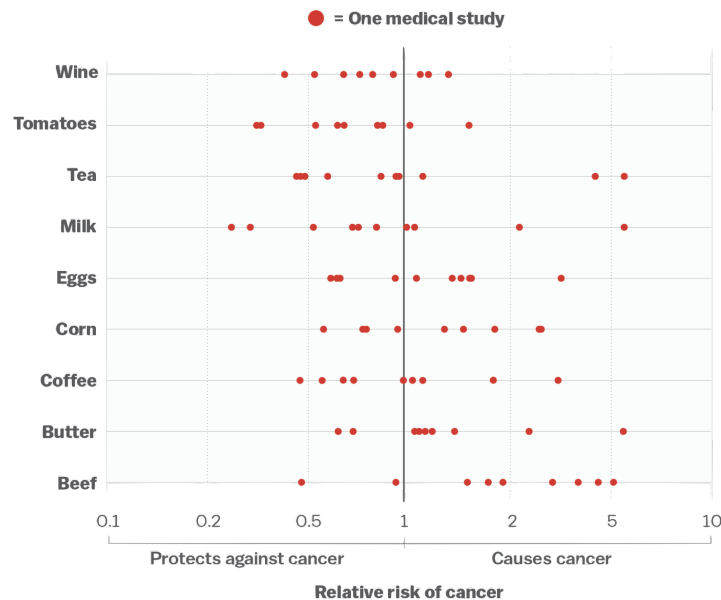
● = One medical study

SOURCE: Schoenfeld and Ioannidis, *American Journal of Clinical Nutrition*

Vox

👋 Read the original article: "*This is why you shouldn't believe that exciting new medical study*" here.

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.
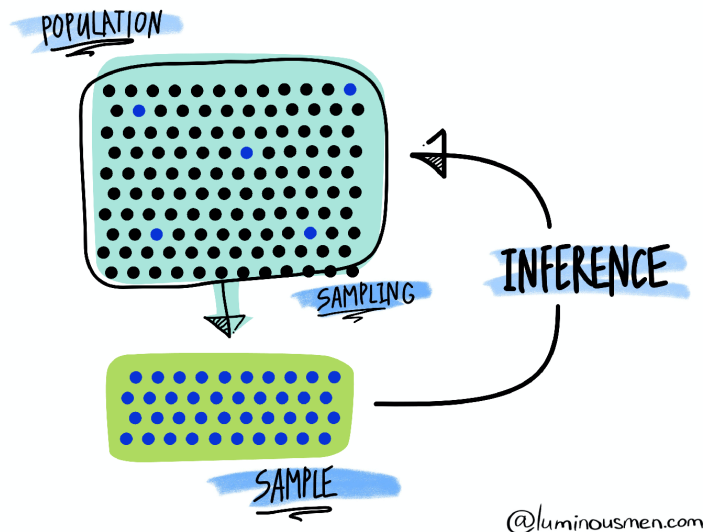
Hypothesis testing

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

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

- 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

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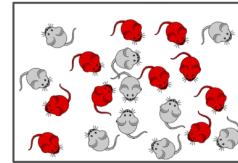
Hypothesis testing

Statistical Hypotheses
H_0: $\mu_{Drug} = \mu_{Control}$
H_a: $\mu_{Drug} < \mu_{Control}$

Sample

Idea:
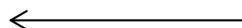Drug represses the progression
of cancer

Data: Tumour size at day 42
$$(x_{D,1}; x_{D,2}; ...; x_{D,n_D})$$
$$(x_{C,1}; x_{C,2}; ...; x_{C,n_C})$$

Inference: Under H0
$$T_{obs} = \frac{\widehat{\mu}_{Drug} - \widehat{\mu}_{Control}}{s_p \sqrt{\frac{1}{n_D} + \frac{1}{n_C}}} \sim St_{n_D + n_C - 2}$$

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

Hypothesis testing

	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 $= \alpha$)	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?

- The **p-value** is defined as the probability of observing a test statistic that is at least as extreme as actually observed, assuming that H_0 is true.
- Informally, a p-value can be understood as a measure of plausibility of the null hypothesis given the data. Small p-value indicates strong evidence against H_0.
- When the p-value is small enough (i.e. smaller than the significance level α), one says that the test based on the null and alternative hypotheses is significant or that the null hypothesis is rejected 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 somehow 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.
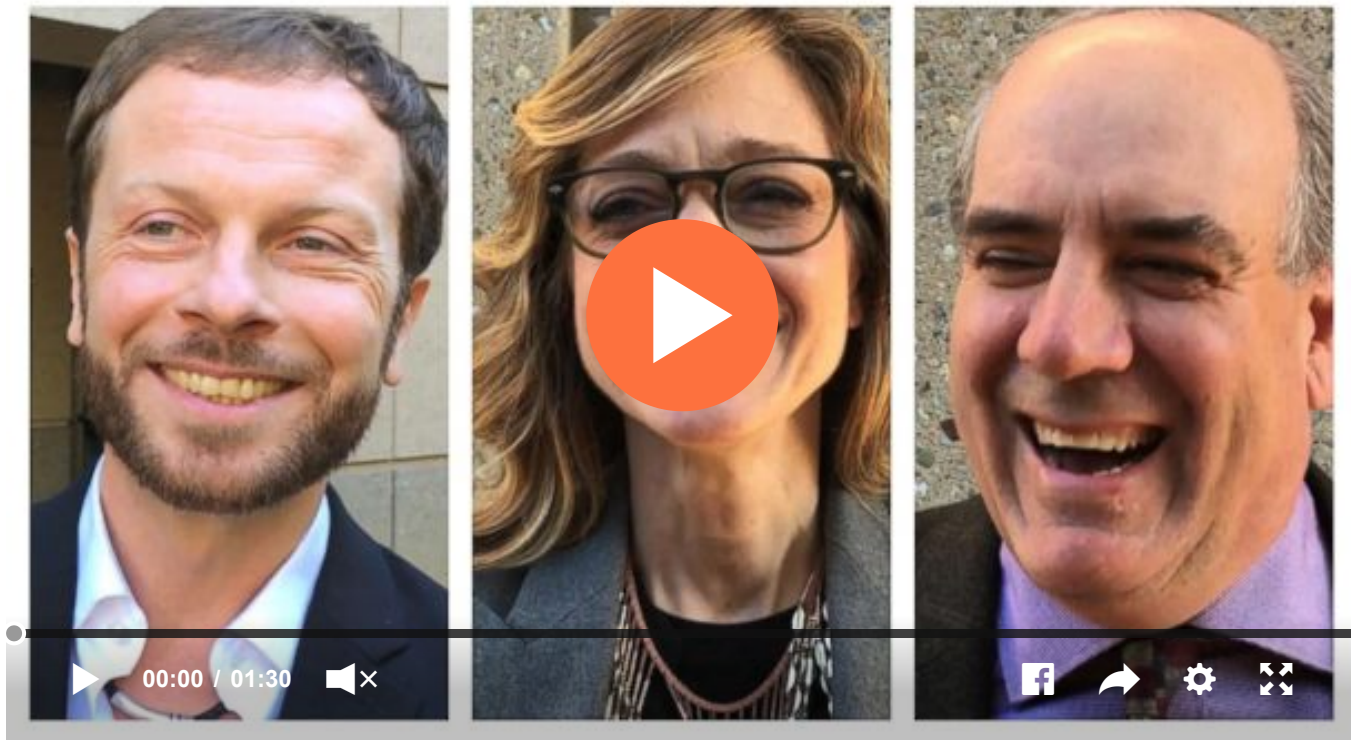
How to understand p-values?

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.
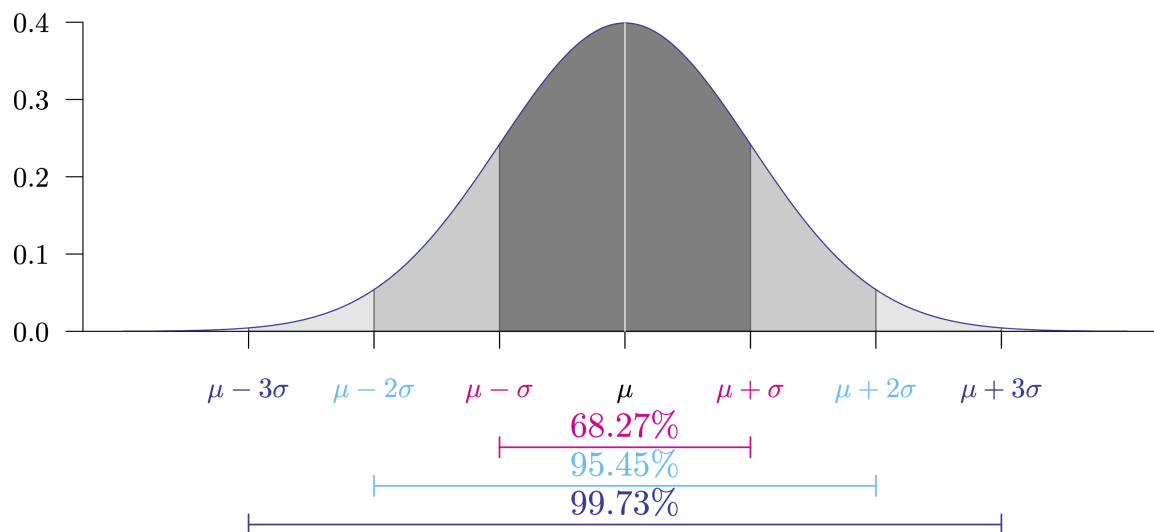
Quick review: Normal distribution

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

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

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

Probability density function of a normal distribution:
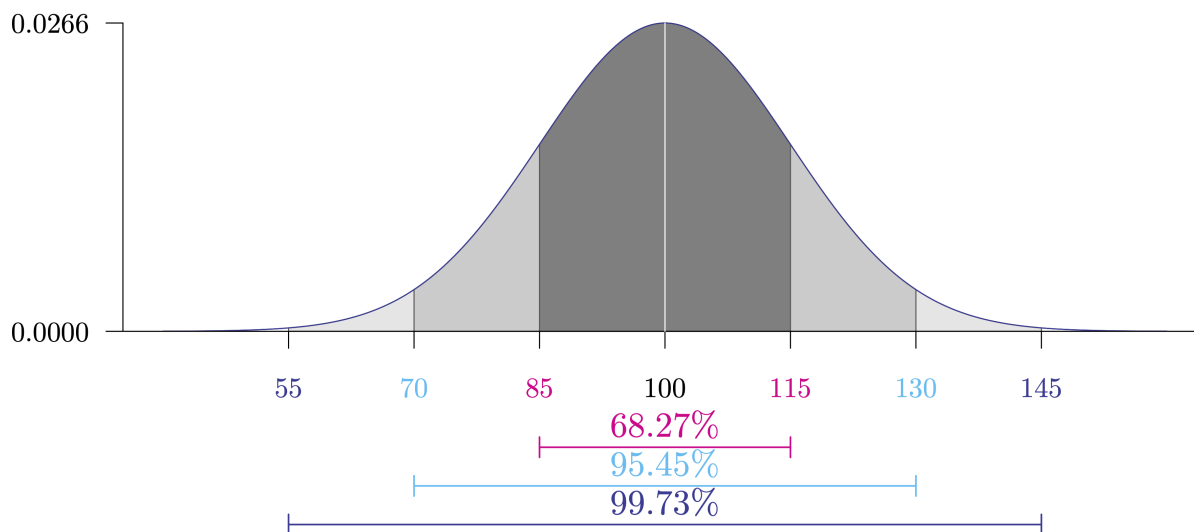
$$Y \sim \mathcal{N}(\mu, \sigma^2), \quad f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

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

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

Suitable modeling for a lot of phenomena: IQ $\sim \mathcal{N}(100, 15^2)$

Normal distribution as an approximation?

The **Central Limit Theorem(s) (CLT)** states (very informally): *the sampling distribution of the average of independent (or "not too strongly dependent") random variables (whose distributions are "not too different" nor "too extreme") tends to a normal distribution as the sample size gets larger.*

As an example, one of the simplest version of the CLT (known as the *Lindeberg–Lévy CLT*) states: *Suppose that $\{X_1, \ldots, X_n\}$ is a sequence of iid random variables such that $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$, then as n approaches infinity, the random variables $\sqrt{n}(\bar{X}_n - \mu)$* converge in distribution *to a normal $\mathcal{N}(0, \sigma^2)$.*

This result can be extended (under some conditions) to dependent (i.e. X_i and X_j are not independent for $i \neq j$) and/or non identically distributed (i.e. X_i and X_j don't have the same distribution for $i \neq j$) data.

Normal distribution as an approximation?

Loosely speaking, we can translate the results of CLTs as

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \overset{\cdot}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

where $\overset{\cdot}{\sim}$ corresponds "approximately distributed as". This result is essential in statistics and a vast majority of inferential methods are based on some versions of the CLT.

Informally, it means that when a measurement can be thought of as the average (or the sum) of **numerous factors**, its distribution tends to go to a normal distribution. For example, the height of adults can be thought of as the sum of their genetic information, diet, life style, …

Central Limit Theorem - Example

People have used the Galton Board as a practical device to demonstrate the CLT, in particular that with sufficient sample size the binomial distribution approximates a normal distribution.

Normal distribution in nature?

> Amends might be made in the interest of the new generation of students by printing in leaded type in future editions of existing text-books and in all new text-books:
>
> *Normality is a myth; there never was, and never will be, a normal distribution.* 😟

👋 *Testing for normality*, R.C. Geary, Biometrika, 1947

> All models are wrong, but some are useful. 🧐

👋 *Science and Statistics*, G. Box, Journal of the American Statistical Association, 1976

So how does it work?

Suppose that we conduct an experiment on n patients where a new diet is tested. We measure X_i as the weight loss after 4 months. Our hope is to show that the diet allows to significantly reduce the weight of the participants.

Suppose that the (possibly dependent) data are such that $X_i \sim F_i$, $i = 1, \ldots, n$ and $\mathbb{E}[X_i] = \mu$. This is a rather plausible assumption (why? 🤔). To verify our hypothesis (i.e. diet reduces weight) we consider:

$$H_0 : \mu = 0 \quad \text{and} \quad H_a : \mu > 0.$$

This implies that we are considering the following model:

$$X_i = \mu + \varepsilon_i,$$

where $\varepsilon_i = X_i - \mu$ can be understood as "residuals".

So how does it work?

Then, by the CLT we have

$$T = \frac{\sqrt{n}\left(\bar{X}_n - \mu_{H_0}\right)}{S} = \frac{\sqrt{n}\bar{X}_n}{S} \underset{H_0}{\sim} G \to \mathcal{N}(0,1),$$
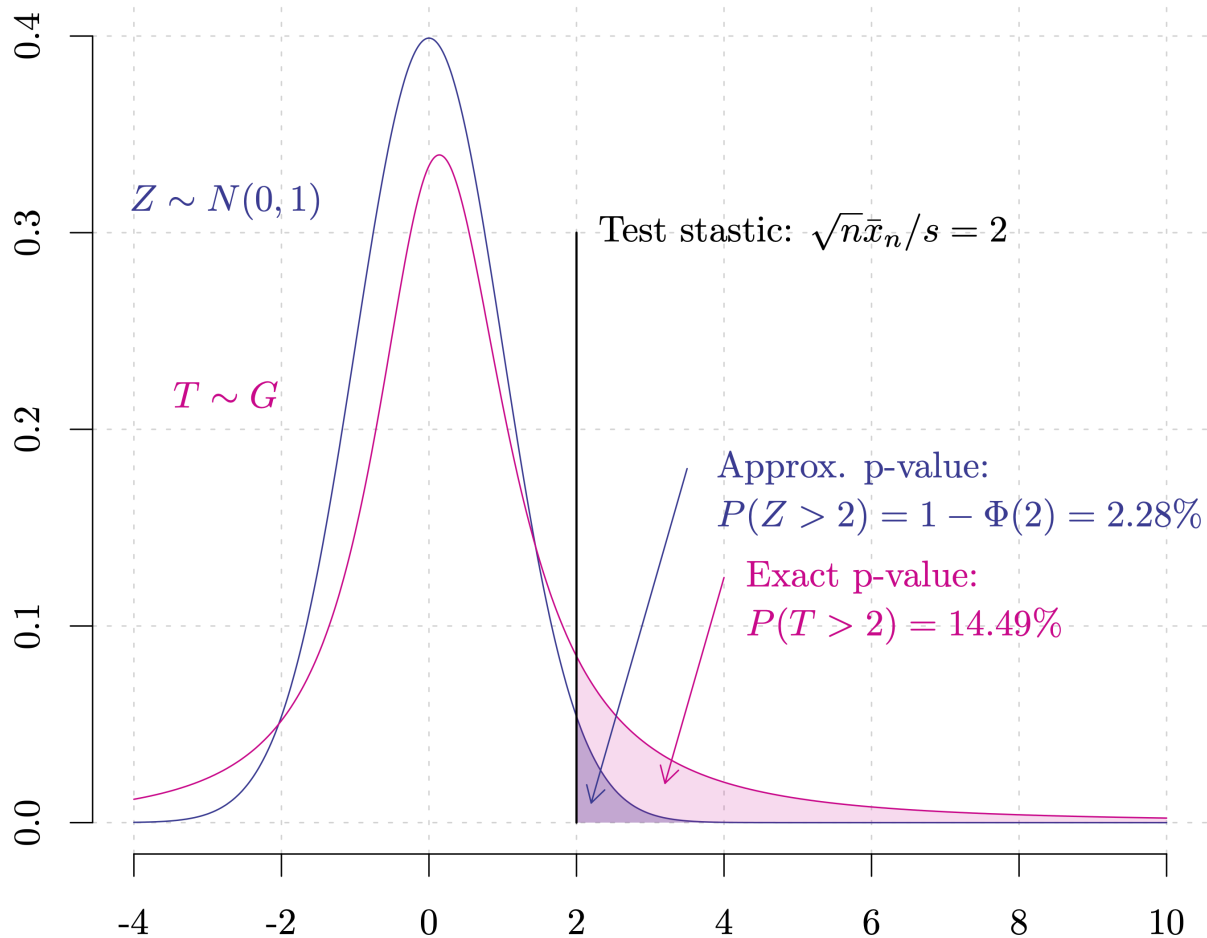
where $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$ and $\underset{H_0}{\sim}$ corresponds "distributed under

H_0 as". In the above formula T is a random variable but we can compute its **realization** based on our sample, i.e. $\sqrt{n}\bar{x}_n/s$.

Let $Z \sim \mathcal{N}(0,1)$, then using the definition of the p-value👋, we have

$$\text{p-value} = \Pr\left(T > \frac{\sqrt{n}\bar{x}_n}{s}\right) \overset{CLT}{\approx} \Pr\left(Z > \frac{\sqrt{n}\bar{x}_n}{s}\right) = 1 - \Phi\left(\frac{\sqrt{n}\bar{x}_n}{s}\right).$$

👋 Reminder: 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 is true.
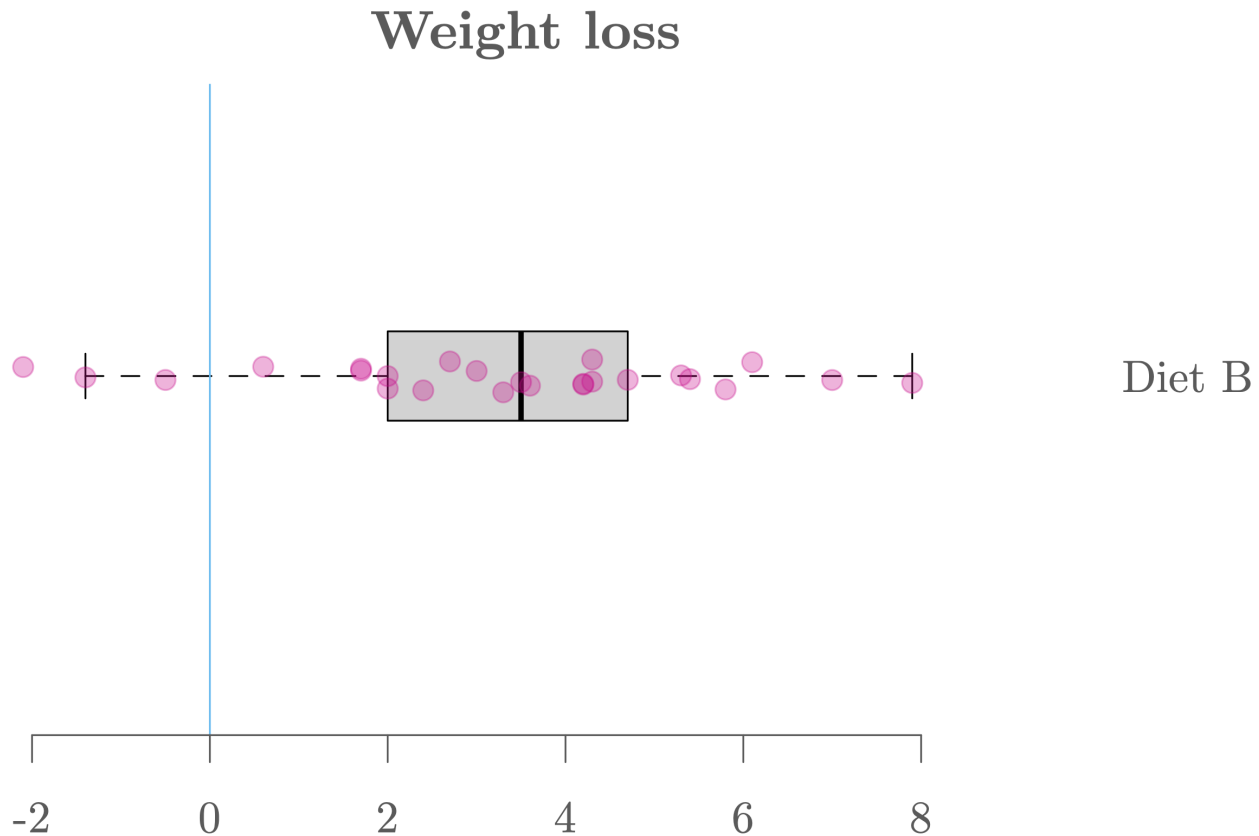
So how does it work?

$Z \sim N(0,1)$

$T \sim G$

Test stastic: $\sqrt{n}\bar{x}_n/s = 2$

Approx. p-value:
$P(Z > 2) = 1 - \Phi(2) = 2.28\%$

Exact p-value:
$P(T > 2) = 14.49\%$

Diet Example

Weight loss

Diet B

One-sample Student's t-test

Before we assume that the (possibly correlated) data is such that $X_i \sim F_i$, $i = 1, \ldots, n$ and $\mathbb{E}[X_i] = \mu$. However, in the very special case where

$$X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2),$$

which corresponds to the following model:

$$X_i = \mu + \varepsilon_i,$$

with $\varepsilon_i = X_i - \mu \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, we have

$$T = \frac{\sqrt{n}\left(\bar{X}_n - \mu_{H_0}\right)}{S} = \frac{\sqrt{n}\bar{X}_n}{S} \underset{H_0}{\sim} \text{Student}(n - 1) \to \mathcal{N}(0, 1).$$

Unlike our previous result, T follows exactly a $\text{Student}(n - 1)$ distribution for all n. 📝 $\text{Student}(n) \to \mathcal{N}(0, 1)$ as $n \to \infty$.
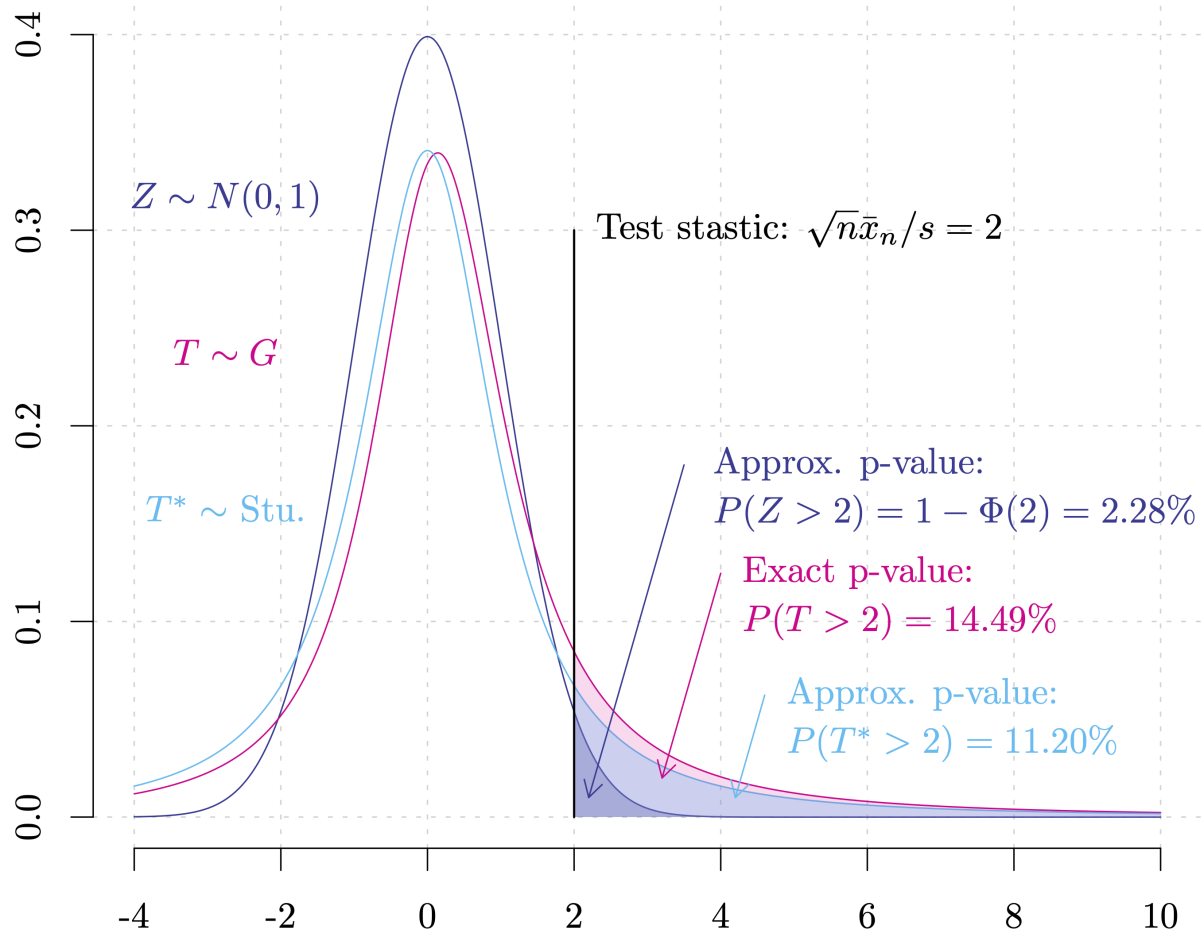
Remarks

- **Why is it called Student?** The Student's t distributions were discovered in 1908 by William S. Gosset, who is a statistician employed by the Guinness brewing company. Gosset devised the t-test as an economical way to monitor the quality of stout 🍻. The company forbade its scientists from publishing their findings, so Gosset published his statistical work under the pen name "Student".

- The t-test (similarly to the test previously discussed) can be used to test if the population mean is greater, smaller, or different than a hypothesized value, i.e.

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

- The t-test accounts for the uncertainty of sample variance and we have:

$$\text{p-values based on } \mathcal{N}(0, 1) < \text{p-values based on } \mathrm{Student}(n).$$

$Z \sim N(0,1)$

$T \sim G$

$T^* \sim$ Stu.

Test stastic: $\sqrt{n}\bar{x}_n/s = 2$

Approx. p-value:
$P(Z > 2) = 1 - \Phi(2) = 2.28\%$

Exact p-value:
$P(T > 2) = 14.49\%$

Approx. p-value:
$P(T^* > 2) = 11.20\%$

R syntax for the t-test

In `R`, we can use the function `t.test( ... )` to compute the p-value for the one-sample Student's t-test. For more information, have a look at `?t.test`. Here are some examples with different alternative hypotheses:

$$H_0 : \mu = 5 \quad \text{and} \quad H_a : \mu > 5$$

```
t.test(data, alternative = "greater", mu = 5)
```

$$H_0 : \mu = 0 \quad \text{and} \quad H_a : \mu < 0$$

```
t.test(data, alternative = "less", mu = 0)
t.test(data, alternative = "less")
```

$$H_0 : \mu = 0 \quad \text{and} \quad H_a : \mu \neq 0$$

```
t.test(data, alternative = "two.sided", mu = 0)
t.test(data)
```

Diet Example (with t-test)

Results   R output

1. Define hypotheses: $H_0 : \mu = 0$ and $H_a : \mu > 0$.
2. Define α: We consider $\alpha = 5\%$.
3. Compute p-value: p-value = 3.697×10^{-5} % (see R output tab for details).
4. Conclusion: We have p-value < α and so we can reject the null hypothesis at the significance level of 5% and conclude that the diet significantly reduces the weight of the participants.

Limitations of the one-sample t-test

The reliability of the t-test strongly relies on:

1. The absence of outliers;
2. For moderate and small sample, the sample distribution is at least approximately normal with no strong skewness (i.e. heavy tails). ✋

⚠️ Therefore, before proceeding to any inference, we should check the data preliminarily using boxplot or histogram or QQ plot 👋 to see if a t-test can be used.

In the diet example, these assumptions appear plausible. When this is not the case, the Wilcoxon signed rank test can be used as an alternative that is less sensitive to outliers and doesn't assume that the data are normally distributed. This test is, however, less powerful than the t-test (when its assumptions are satisfied).

✋ If you want to know more here is an interesting reference. 👋 Check out QQ plot.

One-sample Wilcoxon signed rank test

- The rank is the position index of each observation when we order them from smallest to largest, starting with rank 1 for the smallest observation.
- The Wilcoxon signed rank test only uses the ranks (i.e. the ordering) of the observations, and makes no use of their actual numerical values. Therefore, it is a nonparametric test.
- The Wilcoxon signed rank test depends on the Wilcoxon signed rank statistic, which is the sum of the ranks of only the positive values (or only the negative values). Its distribution under the null can be obtained by different methods (e.g. exact, asymptotic normal, ...) but the details are beyond the scope of this class.
- Unlike the t-test whose hypotheses are on the population mean, the Wilcoxon signed rank test states the hypotheses in terms of population median. When the distribution is symmetric, the mean and the median of the population have the same value.

One-sample Wilcoxon signed rank test

- Under the assumption of a symmetric distribution, the assumed model for this method is given by $X_i = \mu + \varepsilon_i$, where $\varepsilon_i = X_i - \mu \overset{iid}{\sim} (0, \sigma^2)$.

- The Wilcoxon signed rank test considers the same alternative hypothesis as the t-test and its use in `R` is based on the function `wilcox.test( ... )`, which has a similar syntax as the function `t.test( ... )`.

- In general, assessing whether the data is normally distributed is very difficult. (In fact, there is no satisfactory statistical method for it.)

- A reasonable strategy is to always use the one-sample Wilcoxon signed rank test. Compared to the t-test, we may lose a bit of power using the Wilcoxon signed rank test if all the conditions are indeed satisfied (although the difference is generally very small), but if this is not the case the results will be far more reliable.

Diet Example (with Wilcoxon test)

Results   R output

1. Define hypotheses: $H_0 : \mu = 0$ and $H_a : \mu > 0$.
2. Define α: We consider $\alpha = 5\%$.
3. Compute p-value: p-value = 2.71×10^{-3} % (see R output tab for details).
4. Conclusion: We have p-value < α and so we can reject the null hypothesis at the significance level of 5% and conclude that the diet significantly reduces the weight of the participants.