Basics of Statistics

C - Probability and Statistics

2024

Contents

1. Statistical inference

- We must provide evidence based on data.

  For example, saying "my program works" requires evidence/data.

- It must be **reproducible**: only reproducible results might be of interest.

  For example, a **miraculous** cure will not be useful for future patients.

- It must be **transparent**

  to enable others to replicate the same results.

- **We infer** the characteristics of the **population** from a **random sample (RS)**.

  For example, I can infer the population-wide connection speed from a random sample of speeds.

1. Statistical inference. Risks

- The scientific and technical (statistical) method:
  - by **deduction** → data collection design (population → RS)
  - by **induction** → inferring (estimating) results (RS → population)
- Statistical inference defines and **quantifies the risks** of this process. [E.g., the mean connection speed of the entire population cannot be known unless data are available for the entire population, but statistics allows us to estimate and quantify the error from a specific **random sample**.]
- The **evidence** provided by **data** ends with **the analysis:** **e.g.,**
  - "My program works well"

    → estimating a measure (e.g., **average** performance) and its error.
  - "My program improves the results of..."

    → estimating performance improvement (e.g., **mean difference**) and its error.

1. Statistical inference. Types of variables

To analyse the relationship between variables, we must establish the role of each one:

- **Response *Y***. Measuring goal achievement – sometimes it can be an indirect measure.

  E.g., performance ***Y*** measured for a subject.

- **Decisions *X*.** We assign their values in experimental studies.

  They represent the potential to change the future: we want to measure the **effect** of ***X*** on ***Y***.

  An experimental design allows the **X** to be independent of other variables.

  E.g., a teaching method based on **printed lists** of exercises **(*X*=1)** compared with a method based on **e-status (*X*=2)**.

- **Co-variables *Z***. These represent the conditions observed in *real* data.

  We can use ***Z*** to reduce the uncertainty of ***Y*** (we will have to quantify its success).

  We can obtain ***Z*** in both experimental and observational studies.

  Z are usually interrelated (*colinear* or *non-orthogonal*).

  E.g., the marks of two previous subjects **(Z_1, Z_2)** usually have a certain relationship.

1. Statistical inference. Types of study

- **DO: Experimental studies**

We want **to change** the future ***Y*** through interventions in ***X*.**

In the analysis we estimate the **effects** of ***X*** on ***Y***.

E.g., To try to improve the marks ***Y***, we assign at random the students different work environments ***X***.

X represents an **assignable** and well-defined cause.

The key to intervening is to be **owners** of ***X***.

To guarantee independence from all *Z*, we assign ***X*** at random.

We assign respecting ethical and legal rights.

- **SEE: Observational studies**

They allow us **to predict** ***Y*** from the observed values ***Z***.

We will quantify the **capacity** of ***Z*** for **reducing** the **uncertainty** in the prediction of ***Y***.

E.g., we compare the prediction of ***Y*** according to $\boldsymbol{Z_1}$ or according to $\boldsymbol{Z_2}$, or depending on a certain **model m** of the two variables **m=f(**Z_1**,** Z_2**)**.

→ *The group* $\boldsymbol{Z_1}$ *reduces the uncertainty by 10%;* $\boldsymbol{Z_2}$ *by 20%; and the model* ***m****, with both, by 25%.*

We are not **owners** of the *Z* variables (the units already come with the *Z* value).

We can establish **relationships** between *Z* and *Y*, which we can use to **predict** the values of *Y* from *Z*.

But the covariates ***Z*** may be related (**collinear**), so their *effects* on **Y** may be **confounded**.

Establishing **causality** requires many premises (which are beyond an introductory course).

1. Statistical inference. Basic concepts

- **Parameter**: an indicator of the population that we wish to know or estimate. E.g., the expectation (μ) of the heights of FIB students.

- **Statistic**: any indicator that is obtained as a function of the data of a sample. E.g., the sum of the heights of the students in a sample.

- **Estimator**: a statistic of a sample used to know the value of a parameter of the population. E.g., the average height in a random sample of FIB students is an estimator of the expectation (μ) of the heights of FIB students.

Mean may mean ***expectation parameter*** regarding the centre of gravity of the population distribution, or ***statistical mean*** regarding the average of a series of values obtained from a sample.

2. Point estimation

- An estimator $\hat{\theta}$ of the unknown parameter θ from the sample $M(\omega_i)$ $(X_1,X_2,\ldots,X_n)$ (*a simple random sample defined in the appendix to Section B*) is a function of the RVs:

$$\hat{\theta} = f(X_1, X_2, \ldots, X_n)$$

- **Point estimation:** the value that the estimator $\hat{\theta}$ takes in a specific sample. E.g., $\bar{x} = \frac{\sum x_i}{n}$ is the sample mean and is a point estimate of μ.

Distinguish between the value $\bar{x}$ (small letter) of a specific simple random sample and the sample mean random variable $\bar{X}$ (capital letter).

- **Standard error:** the variability of the estimator. In the above case of MEAN, the **standard error of the mean** (or ***mean standard error***, or **SE**) is

$$se = \sqrt{V(\bar{X}_n)} = \sqrt{E[(\bar{X}_n - \mu)^2]} = \frac{\sigma}{\sqrt{n}}$$

Generally, the **σ** will be unknown and the standard error will have to be approximated using the co-rresponding estimator ($\hat{\sigma}$) with the sample data: $\widehat{se} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}}$ (with s the point estimator of σ).

Read the comments in point 5 of this section "Designs (how we obtain the data)" to use for Unit T

2. Point estimation. Cases

For the parameters we use letters of the Greek alphabet.

Parameter (θ) (**POPULATION**)	Estimator ($\hat{\theta}$) (**SAMPLE**)
μ (expectation, population mean)	**$\bar{x}$** (sample mean)
σ^2 (population variance) **σ** (population standard deviation)	**s^2** (sample variance) **s** (sample standard deviation)
π (probability)	**p** (proportion)

MEAN:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The sample mean is a **point estimate of the parameter μ** of central tendency.

STANDARD DEVIATION:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}}$$

The sample standard deviation is a **point estimate of the parameter σ** of dispersion.

PROPORTION:

$$p = \sum_{i,x_i=1} 1/n$$

The sample proportion is a **point estimate of the parameter π**.

We must take into account the properties of the estimators (see the Appendix, along with other potential estimators).

2. Estimators proprieties. Descriptive statistics

- Estimators proprieties:

  Parameters could have different **estimators** (approximations to the punctual unknown value) and have a **margin of error** depending on the sample

  So we could compare in base to some **proprieties**:

  - **Bias** that it's preferable to be near 0 in order to assure that the expected value will fit the real value
  - **Eficiency** that it's preferable to be high pointing out more precision and less dispersion
  - other proprieties like consistency, ...

  Annex C has more information about estimators proprieties

  At bibliography ("Estadística per a enginyers informatics") you could find more information about estimators proprieties in chapter 2)

- The previous punctual estimators are what is called **Descriptive Statistics** to summarize data numerically

  At the website of the subject you could find more information about Descriptive Statistics in R

  At the end of this unit you could find more information about Descriptive Statistics for unit T

3. Estimation using confidence intervals

- We know how to calculate an "interval" that contains $\bar{x}$ from μ. But the real problem is **to approximate μ from $\bar{x}$** (i.e., moving from an interval for the sample mean $\bar{x}$ to one for the population mean μ)

- From a probability 1-α between two (symmetric) values *a* and *b (with known σ)*:

$$P(a \le \bar{X}_n \le b) = 1 - \alpha \rightarrow P\left(\frac{a-\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \le \frac{b-\mu}{\sigma/\sqrt{n}}\right) = 1 - \alpha \rightarrow P\left(z_{\frac{\alpha}{2}} \le \frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \le z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

- we get the interval of the RV $\bar{X}_n$ with **probability** 1-α:

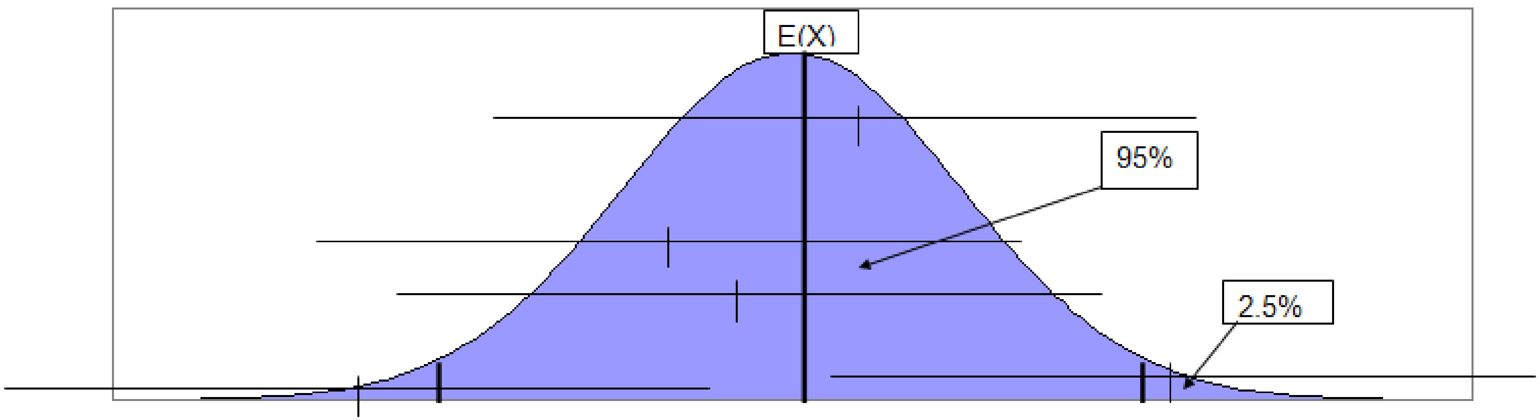
$$P\left(\mu + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \bar{X}_n \le \mu + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

- By rearranging, we get **the confidence interval (CI) 1-α of the parameter μ**:

$$P\left(\bar{X}_n + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

3. Estimation using confidence intervals

- $P\left(\bar{X}_n + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$ means that we can ensure that $E(X) = \mu$ will be in the calculated range (with a confidence of $1-\alpha$)
- If $1-\alpha$ is 95% (α =5%): **95% of the CIs will contain μ** (see a simulation in the Appendix)



- This procedure is correct $100\cdot(1-\alpha)$% of the time!
- We call **CI(μ, 1-α)** the **CONFIDENCE INTERVAL** $1-\alpha$ of μ

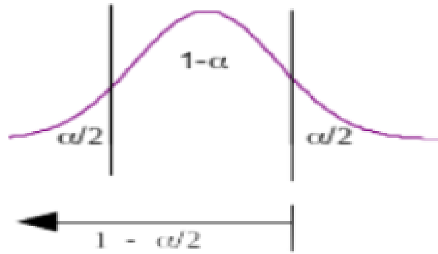
$$IC(\mu, 1-\alpha) = \bar{x} \mp z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

($z_{\alpha/2} = -z_{1-\alpha/2}$ bacause Z is simetric)

We will only observe one sample, and we will not know whether the found CI contains μ or not, but we do know that in the long run this procedure gives true values $100\cdot(1-\alpha)$% of the time

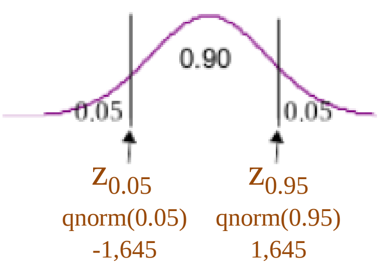
3.a. Confidence and risk

The calculation of a CI implies a confidence $1-\alpha$ (and therefore a risk α), which we can represent as

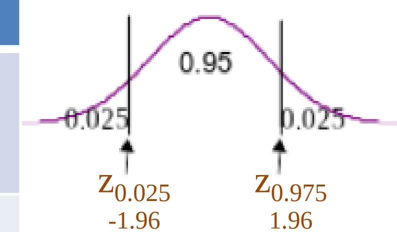


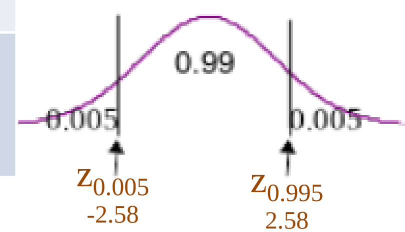
And we can relate the confidence value to the quantile that we need to build the CI

[E.g., the quantiles are indicated by a normal $Z(0,1)$, where we know that $z_{\alpha} = -z_{1-\alpha}$ or $z_{\alpha/2} = -z_{1-\alpha/2}$]



Confidence $1-\alpha$	Risk α	$\alpha/2$	$1 - \alpha/2$
0.95	0.05	0,025	0,975
0.90	0.10	0.05	0.95
0.99	0.01	0,005	0,995





3.b. Statistics for inference

- We will see statistics of two types:
  - Ratio of **"signal" or "information"** (difference between a value μ_0 of the parameter and the sample value) **to "noise" or "error"** (standard error, SE).

These statistics are modelled following the *Z* or Student *t** model (in some cases we evaluate the "*t*-ratio" that quantifies by how many times the signal is greater than the noise).

statistic $\hat{z} = \frac{(\bar{x}-\mu_0)}{\sigma/\sqrt{n}} = \frac{(\bar{x}-\mu_0)}{se}$ $\hat{z} \sim Z = N(0,1)$ (then the CI is $\mu \in \bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ or $\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot se$)

statistic $\hat{t} = \frac{(\bar{x}-\mu_0)}{S/\sqrt{n}} = \frac{(\bar{x}-\mu_0)}{\widehat{se}}$ $\hat{t} \sim$ T **Student** with degrees of freedom $\boldsymbol{v}$

  - **Ratio of variances**. These statistics are modelled following the F model.

statistic $\hat{F} = \frac{S_A^2}{S_B^2}$ $\hat{F} \sim$ F **Fisher–Snedecor** with degrees of freedom $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$

*Student $t_{\nu,\,1-\alpha/2}$; Fisher $F_{\nu1,\nu2,1-\alpha/2}$ and chi squared $\chi^2{}_{\nu,\,1-\alpha/2}$ are defined in Section B (Appendix). Those models are derived from the normal distribution, and they are **parameterised with** degrees of freedom (**v**), **depending on** the sizes (**n**) of the samples.*

3.c. Assumptions

- The fundamental assumption is to start from a **random sample**.

> We say that values come from independent and identically distributed (**IID**) random variables.

app app

- The premise of normality is necessary because CIs are based on the CLT theorem, which is based either on an original **normal** variable or a **"large" *n***.

> In small samples (n<30?), we will sustain the **premise of normality** with the **prior knowledge** of the response variable and with the graphic analysis **with R**.
>
> (see at the end of this Unit functions in R and Graphical Analysis of Normality in the Annex of Unit B)

4.a. Confidence interval for one parameter

Now we will see the **CI formulas** for 3 **single parameters**:

- The mean μ (with or without known population variance)

  E.g., the mean mark of a subject

- A proportion π

  E.g., the proportion of passes of a subject

- The variability σ^2

  E.g., the deviation from the mean mark of a subject

Confidence interval of μ (with known σ)

- The confidence interval 1−α of μ (with known σ) is calculated as

$$CI(\mu, 1-\alpha) = \bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

(R: $z_{1-\frac{\alpha}{2}}$ is `qnorm(1- α/2)`)

- Remember that we are using the CLT, which requires the random variable X to be normal or n to be "large". Therefore, the requirement for performing this calculation is either **X~N** *or* **n "large"**

- This CI can be obtained by setting apart the parameter μ from the statistic: $\hat{z} = \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}} = \frac{(\bar{x}-\mu)}{se}$ whose distribution we know to be $N(0,1)$
- Therefore the CI(μ,1-α) can be seen as $\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot se$.

When n increases, the CI accuracy increases (narrower range). If the confidence increases (decreasing the risk α of error), the accuracy of the CIs decreases (wider range).

To estimate μ, we need to know σ, which is an unrealistic situation because σ is usually an unknown parameter (we can also assume a reasonable value from prior knowledge).

Confidence interval of μ with unknown σ

The previous confidence interval 1−α of μ with unknown σ is calculated as

$$CI(\mu, 1-\alpha) = \overline{x} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

(R: $t_{n-1,1-\frac{\alpha}{2}}$ is `qt(1-α/2,n-1)`)

This CI is obtained by isolating the parameter μ from the statistic: $\hat{t} = \frac{(\bar{x}-\mu)}{s/\sqrt{n}} = \frac{(\bar{x}-\mu)}{se}$

When σ is unknown, we replace σ by *s;* and

the *Normal Z* by the ***Student t*** with *n-1 df*

being df=degrees of freedom

In this case the <u>initial variable *X* must be normal</u> (the premise of normality) because the definition of the ***Student t*** model is based on normal variables.

The situation of not knowing σ is more realistic and frequent: no value is assumed but it is approximated by its point estimate *s*.

t and *N*(0,1) are similar, increasingly so when *n* grows, $t_{n\text{-->}\infty}$ → *N*(0,1)

For small values of *n*, *t* has more variability, reflecting more uncertainty (as σ is approximated by *s*).

Therefore, the CI with unknown σ will be wider than the equivalent assuming the true value of σ.

Confidence interval of μ. Assumptions

To guarantee the confidence level of the CI, certain premises must be met.

The **fundamental** premise is that the origin of the sample must be **random**.

In addition:

- If sigma is known, one of the following two conditions is required:
  - **X~N** → since the linear combination of normals is also normal ($\bar{X} \sim$N)
  - **The sample is "large"** → by the CLT, $\bar{X} \sim$N
- If sigma is unknown, one of the following conditions is required:
  - **X~N** → $(\bar{x} - \mu)/\sqrt{s^2/n} \sim t_{n-1}$
  - **The sample is large ("large" n)** → by the CLT, $\bar{X} \sim$N

In larger samples, the variation of *s* will be smaller (*s* estimates σ well), and we can consider that $(\bar{x} - \mu)/\sqrt{s^2/n} \approx (\bar{x} - \mu)/\sqrt{\sigma^2/n} \sim \mathrm{N}(0,1)$.

In summary	...σ Is known	...σ is unknown
If X is normal and...	We use the Normal	We use the *Student t*
If X is not normal but *n* is "large" and...	We use the Normal	We use the Normal

Confidence interval of π

Let X ~ B(n,π) → E(X) = $\pi \cdot n$

V(X) = $\pi \cdot (1-\pi) \cdot n$

Then, P = X/n → E(P) = E(X/n) = E(X)/n = $\pi \cdot n$ / n = π

V(P) = V(X/n) = V(X)/n^2 = $\pi \cdot (1-\pi) \cdot n/n^2$ = $\pi \cdot (1-\pi)/n$

By using the convergence from B to N, $P \rightarrow N\left(\mu_P = \pi, \sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

So, the statistic $\hat{z} = \frac{(P-\pi)}{\sigma_P} = \frac{(P-\pi)}{se}$ is distributed as N(0,1) provided

n is "large" **and** π not extreme.

$$IC(\pi, 1-\alpha) = P \pm z_{1-\frac{\alpha}{2}}\ se = P \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

(R: $t_{n-1,1-\frac{\alpha}{2}}$ is `qt(1-α/2,n-1)`)

As a summary guide, check that

$\pi \cdot n \geq 5$ and **$(1-\pi) \cdot n \geq 5$**

The **paradox** that we need to know π to estimate the CI of π is usually solved in two ways:

a) by substituting $\hat{\pi}$ with P: $IC(\pi, 1-\alpha) = P \pm z_{1-\alpha/2} \cdot \sqrt{(P(1-P))/n}$

b) by obtaining the maximum of $\hat{\pi} \cdot (1-\hat{\pi})$, making $\hat{\pi}$ equal to 0.5: $IC(\pi, 1-\alpha) = P \pm z_{1-\alpha/2} \cdot \sqrt{(0.5(1-0.5))/n}$.

Confidence interval of σ^2

If $X_i \to N$ $\quad (n-1)\cdot\frac{s^2}{\sigma^2} = (n-1)\cdot\frac{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)/(n-1)}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 \sim \chi^2_{n-1}$

We can relate the variance ratio statistic (S2/σ2) to a χ^2

as the sum of squared normal variables is χ^2

(see models derived from the normal in the Appendix to Section B).

Therefore,

$$P\left(\chi^2_{n-1,\frac{\alpha}{2}} \le \frac{S^2\cdot(n-1)}{\sigma^2} \le \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(\frac{1}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \le \frac{\sigma^2}{S^2\cdot(n-1)} \le \frac{1}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) = 1-\alpha$$

$$P\left(\frac{S^2\cdot(n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \le \sigma^2 \le \frac{S^2\cdot(n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) = 1-\alpha$$

$$IC(\sigma^2, 1-\alpha) = \left[\frac{s^2(n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{s^2(n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}\right]$$

This is a CI for σ^2, not for σ!!

(R: $\chi^2_{n-1,1-\frac{\alpha}{2}}$ is `qchisq(1-α/2,n-1)` and $\chi^2_{n-1,\frac{\alpha}{2}}$ is `qchisq(α/2,n-1)`)

Since χ^2 is not symmetrical, it requires obtaining the upper and lower quantiles instead of doing $\pm$.

4.b. Confidence interval to compare 2 parameters

The **CI formulas** to…:

- Compare μ_1 and μ_2

  *E.g., the CI of the differential effect (μ_1-μ_2) comparing averages between two subjects**

  We must differentiate between

  - **paired samples**** (each case results in two measures, pairs of measures)

    (the same students in both subjects, $\mu_1-\mu_2 = \mu_{difference} = \mu_d$)
  - **independent samples** (each case is an independent measure)

    (different students in the two subjects)
- Compare π_1 and π_2

  *E.g., the CI of the differential effect (π_1-π_2) comparing averages between two subjects**
- Compare $\sigma^2{}_1$ and $\sigma^2{}_2$

  *E.g., the CI comparing deviations between two subjects**

* **The origin of the sample must be random**.

** If possible, a design with paired data will be more efficient (as we will see below).

CI of $\mu_1 - \mu_2$ (or of μ_D) in paired samples

If obtain a simple random **paired sample** of size *n,* and
we define **d** = $Y_1 - Y_2$ then $E(d) = \mu_d$ and $V(d) = \sigma_d^2$
and the *n* observed differences values have a mean $\bar{d}$ and deviation s_d.

- The statistic $\hat{t} = \frac{(\bar{d}-\mu_D)}{s_d/\sqrt{n}} = \frac{(\bar{d}-\mu_D)}{se}$ follows the t_{n-1} model

  where $(\bar{d} - \mu_D)$ is the **signal** and $se = s_d/\sqrt{n}$ is the standard **error**.

- The CI of the population mean difference with confidence 1-α is

$$IC(\mu_1 - \mu_2, 1-\alpha) = IC(\mu_d, 1-\alpha) = \bar{d} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}} = \bar{d} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot se$$

(R: $t_{n-1,1-\alpha/2}$ is `qt(1-α/2,n-1)`)

It may be of practical interest to evaluate the *t*-ratio: $t = \bar{d} \,/\, \frac{s_d}{\sqrt{n}} = \bar{d}/se$,
which says how many times the **signal** is greater than the **noise**.

CI of $(\mu_1 - \mu_2)$ independent samples

Be Y_1 with $E(Y_1) = \mu_1$, $V(Y_1) = \sigma_1^2$; and Y_2 with $E(Y_2) = \mu_2$, $V(Y_2) = \sigma_2^2$ with normal distributions (**σ_1 and σ_2 will be unknown values but must be assumed to be the same*)**, from which we obtain two **independent** simple random samples of size n_1 and n_2 with means $\bar{y}_1, \bar{y}_2$ and deviations s_1 and s_2 as estimators of the common parameter σ.

- The statistic $\hat{t} = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{se}$ follows the distribution $t_{n_1+n_2-2}$, with standard error

$\boldsymbol{se} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, where s is the root of the "pooled" variance $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$

- The CI of the difference with confidence 1-α is

$$IC(\mu_1 - \mu_2, 1-\alpha) = (\bar{y}_1 - \bar{y}_2) \pm t_{n_1+n_2-2,1-\frac{\alpha}{2}} \cdot se =$$

$$(\bar{y}_1 - \bar{y}_2) \pm t_{n_1+n_2-2,1-\frac{\alpha}{2}} \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(R: $t_{n_1+n_2-2,1-\frac{\alpha}{2}}$ is `qt(1-α/2,n1+n2-2)`)

The **fundamental** condition is random samples. We **know** it.

The other **two assumptions** (normality of Y_1, Y_2 and $\sigma_1^2=\sigma_2^2$) will be analysed graphically.

CI of $\pi_1 - \pi_2$

Let P_1 and P_2 be the sample proportions of two binomial populations with π_1, π_2, from which we obtain two independent simple random samples of size n_1 and n_2 .

- The statistic $\hat{z} = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{se}$ follows the distribution N(0,1) with standard error $se = \sqrt{P_1(1-P_1)/n_1 + P_2(1-P_2)/n_2}$.

- The CI of the difference with confidence 1-α is

$$IC(\pi_1 - \pi_2, 1-\alpha) = (P_1 - P_2) \pm z_{1-\frac{\alpha}{2}} \cdot se =$$

$$(P_1 - P_2) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{P_1(1-P_1)/n_1 + P_2\,(1-P_2)/n_2}$$

(R: $z_{1-\alpha/2}$ is `qnorm(1-α/2)`)

In this case, convergence requires “large” samples: usually $P \cdot n > 5$ and $(1-P) \cdot n > 5$.

CI of σ_1^2/σ_2^2

Let s_1 and s_2 be the sample deviations of two independent simple random samples of size n_1 and n_2 of two normal variables.

- The statistic $\hat{F} = \dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ follows the distribution $F_{(n_1-1,n_2-1)}$.
- The CI of the ratio of variances with confidence 1-α is
  (following the same reasoning as the CI of σ^2)

$$IC(\sigma^2{}_1/\sigma^2{}_2, 1-\alpha) = \left[\frac{s_1^2/s_2^2}{F_{(n_1-1,n_2-1),1-\frac{\alpha}{2}}}, \frac{s_1^2/s_2^2}{F_{(n_1-1,n_2-1),\frac{\alpha}{2}}}\right]$$

(R: $F_{(n_1-1,n_2-1),1-\frac{\alpha}{2}}$ is `qf(1-α/2,n₁-1,n₂-1)` and $F_{(n_1-1,n_2-1),\frac{\alpha}{2}}$ is `qf(α/2,n₁-1,n₂-1)`)

or (note the exchange of degrees of freedom of *F*)

$$IC(\sigma^2{}_1/\sigma^2{}_2, 1-\alpha) = \left[s_1^2/s_2^2 \; F_{(n_2-1,n_1-1),\frac{\alpha}{2}}, \; s_1^2/s_2^2 \; F_{(n_2-1,n_1-1),1-\frac{\alpha}{2}}\right]$$

4.c. A list of functions in R

Premise of normality (in Graphical Analysis of Normality in the Appendix to Section B):

```
qqnorm(X)
qqline(X)
```

app

CI of μ with known σ (for this function you need the BSDA library):

```
library(BSDA)
z.test(X,sigma.x= )        # for a sample when sigma is known
```

CI of μ (or μs) when σ (or σs) is unknown:

```
t.test(X)                              # for one sample
t.test(XY) or t.test(X,Y,paired=T)     # for paired samples
t.test(X,Y,var.equal=T)  # for two independent samples with equal variances
t-test(X,Y,var.equal=F)  # for two independent samples with different variances
```

CI of σs in two independent samples:

```
var.test(X,Y)
```

CI of π*:*

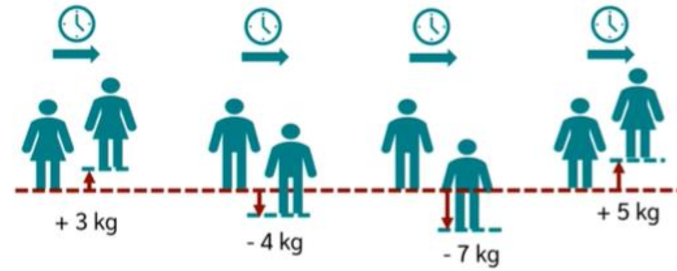
`prop.test` and **`binom.test`**

5. Designs (how we obtain the data)

Paired design:

One variable and two observations are taken from each unit (the two measures or responses),

Requirement: the first observation in a "pair" must not alter the state of the unit and therefore of the second observation.



Independent samples

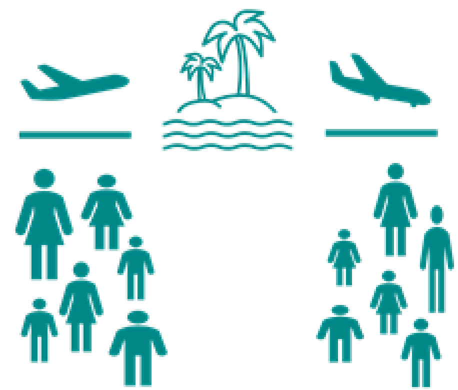
For each unit, one observation and two variables are taken:

(1) the outcome, the measure of the response and

(2) the category to compare.

Requirements:

- It must be possible to assign the category to the unit (it cannot be a condition, such as sex).
- In observational studies, when the group is not assignable, the samples are selected separately.



This is a simple approximation. The world of experiment design is much broader.

The key is **random**:

Collecting data arbitrarily does not guarantee a random sample: **willy-nilly ≠ at random.**

(1) **to plan** the random selection of the units to be measured; (2) to **carry out** the experiment correctly; (3) without **missing** values; and (4) to **document it** in a **reproducible** way.

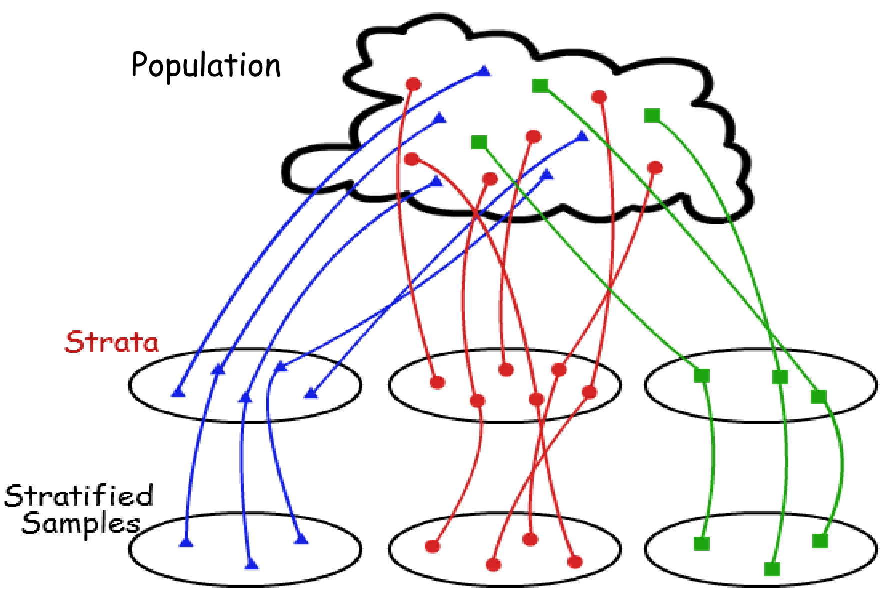
Designs

The chosen design conditions the subsequent statistical analysis.

If the collection is complex, the statistical model used is also complex:

- Cases with **nested** data (*clusters*): groups from the top level (e.g., *school*) are first randomly selected, then groups from the lower level (*class*), until the individual (*student*) is reached.
- Cases with **stratified** data: all the strata can be seen, but within each stratum the individuals are randomly selected.

Then, the group of students chosen is not strictly a random sample; they must be analysed with appropriate techniques (not explained at a first level statistical course).



It is unusual to have the listing of the complete population; nor to be able to access any unit under the same conditions (both are requirements for a simple random sample).

Usually, some units will be more “visible” than others and will have a higher probability of being chosen (e.g., results that can only be obtained in order).

6. Estimators and descriptive statistics

The above point estimators correspond to the functions of **descriptive statistics** for numerically summarising data (see more in the R section of the website).

The following table shows some (basic) functions in R for **descriptive statistics** in **univariate or bivariate numerical and categorical variables**:

	UNIVARIATE (numerical)	UNIVARIATE (categorical)	BIVARIATE
INDICATORS	length( ) * **mean( )** **var( )** **sd( )** **summary( )** **median( )**	**table( )**	**cov(,)** **cor(,)**
GRAPHICS	hist( ) boxplot( )	barplot(table( ))	plot(,)

* The sample size (*n*) is not an estimator, but we include it in the list for practicality.

(More graph functions in R: https://www.r-graph-gallery.com/)