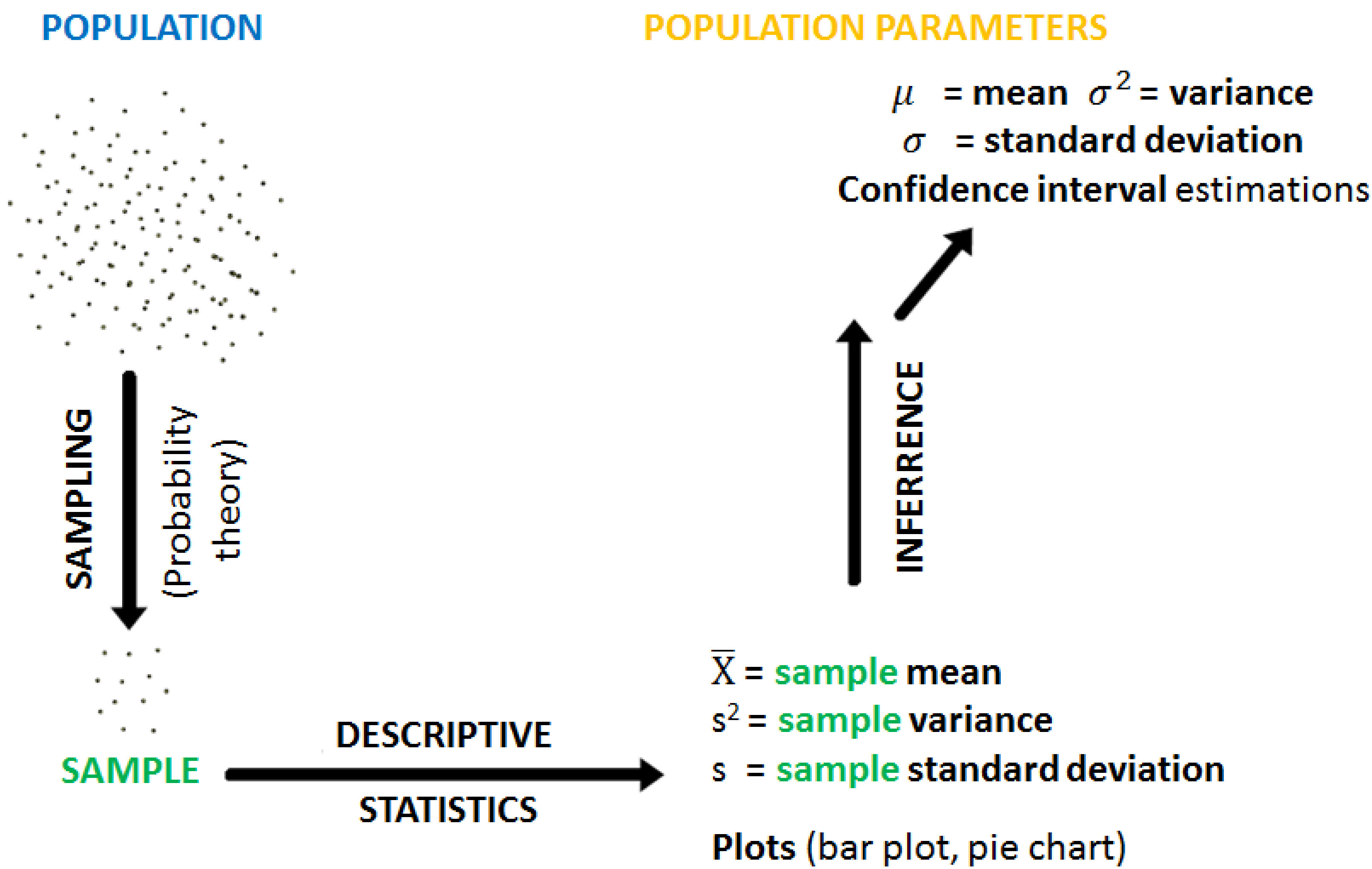
Analysing and presenting data: **practical hints**

Giorgio MATTEI

giorgio.mattei@centropiaggio.unipi.it

Course: Meccanica dei Tessuti Biologici
Date: 12 May 2016


POPULATION
POPULATION PARAMETERS
μ = mean σ^2 = variance
σ = standard deviation
Confidence interval estimations
SAMPLING
(Probability theory)
INFERENCE
SAMPLE
DESCRIPTIVE
STATISTICS
$\overline{X}$ = sample mean
s^2 = sample variance
s = sample standard deviation
Plots (bar plot, pie chart)

Basic probability theory

$$Pr\{A\} = P(A) = \lim_{n\to\infty} \frac{n_A}{n}$$

Event A probability

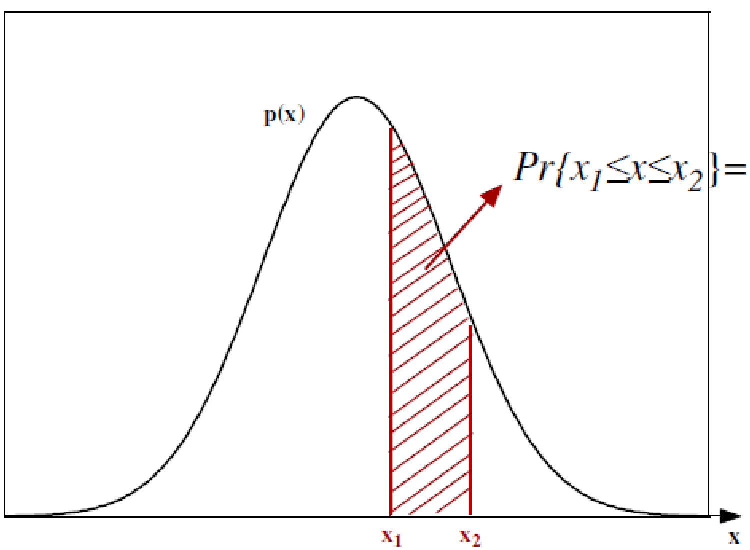
$$Pr\{S\} = P(S) = 1$$

Certain event probability

$$p(x) = \lim_{\Delta x\to 0} \frac{Pr\{x \le \bar{x} \le x + \Delta x\}}{\Delta x}$$

Probability density function (*pdf*) of $\boldsymbol{x}$

($\bar{x}$ is a **random variable** that assumes a given **value x** after the experiment)



$$Pr\{x_1 \le x \le x_2\} = \int_{x_1}^{x_2} p(x)\,dx$$

For $n \to \infty$ the relative frequency density approximates the *pdf*

Expectation operator and normal distribution

- **Mean** (μ) and **variance** (σ^2) for a random variable ($\bar{x}$) with a given *pdf* ($p(x)$) can be calculated through the **expectation operator**

$$\mu = \int x p(x) dx = E(\bar{x})$$

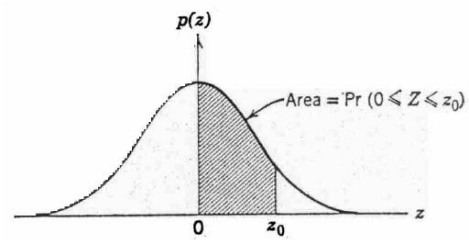
$$\sigma^2 = \int (x-\mu)^2 p(x) dx = E\{(x-\mu)^2\} = Var(\bar{x})$$

- $\bar{x}$ is **normal** with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$ if its *pdf* is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Standard normal variable ($\boldsymbol{\mu}$=0, $\boldsymbol{\sigma}^2$=1) and **variable standardisation**

- **Standardised normal probability density**

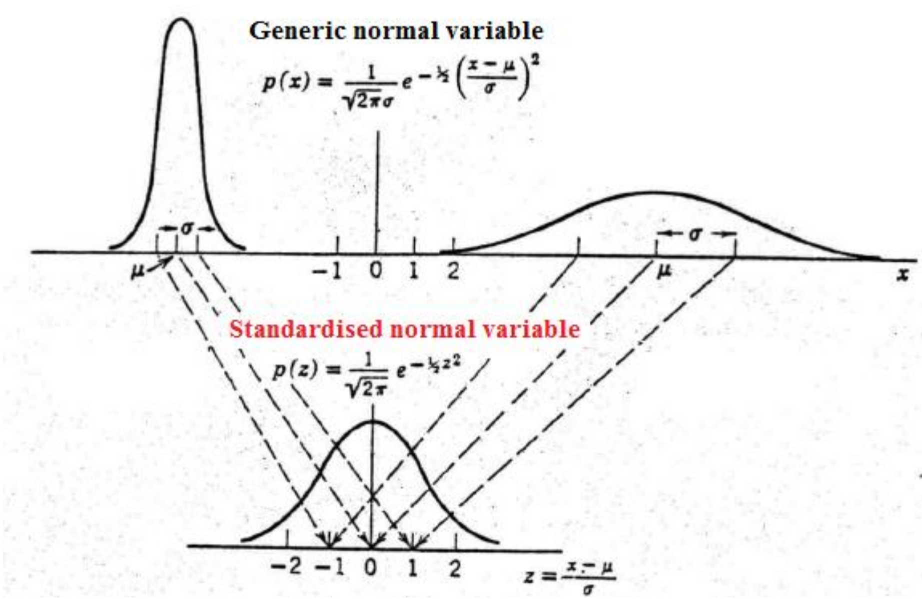


$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Pr {**-1.96** ≤ z ≤ **1.96**} = 0.95 = **95 %**

$\mathbf{z_{0.05}}$ = **1.96**

- **Generic normal variable standardisation** ($\bar{x} \rightarrow \bar{z}$)



$$\frac{\bar{x} - \mu}{\sigma} = \bar{z}$$

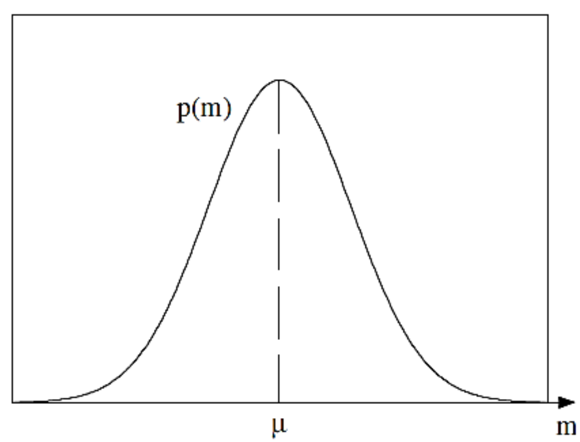
Inference

- **Population** parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$) are **constant** but **unknown**
- **Observed sample** parameters ($\overline{\boldsymbol{m}}$ and $\overline{\boldsymbol{s^2}}$) are **random variables** that may change with samples, according to a given ***pdf***
- **Population parameters** can be **inferred** from **observed samples** knowing the ***pdf*** of **the sample statistics**
- $\overline{\boldsymbol{m}}$ is an **un-biased estimator** of $\boldsymbol{\mu}$ (from probability theory)

$$E(\overline{m}) = \mu \Leftrightarrow \mu_{\overline{m}} = \mu$$

$$Var(\overline{m}) = \frac{\sigma^2}{n} \Leftrightarrow DS(\overline{m}) = \frac{\sigma}{\sqrt{n}}$$

$$\overline{m} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$



$$\boxed{\frac{\overline{m} - \mu}{\sigma / \sqrt{n}} = \overline{z}}$$

Standardised $\overline{\boldsymbol{m}}$

Confidence interval (CI) estimations

- In general $\boldsymbol{\mu} \neq \overline{\boldsymbol{m}}$, but $\boldsymbol{\mu} = \overline{\boldsymbol{m}} \pm \Delta$ and **↑CI → ↑**Δ
- 95% CI means that the error Δ is such that

$$Pr\{\overline{m} - \Delta \leq \mu \leq \overline{m} + \Delta\} = 95\% \longrightarrow Pr\{\mu - \Delta \leq \overline{m} \leq \mu + \Delta\} = 95\%$$

2 cases

Unknown $\boldsymbol{\mu}$
Known $\boldsymbol{\sigma^2}$

$$\frac{\overline{m} - \mu}{\sigma / \sqrt{n}} = \overline{z}$$

$\bar{z}$ statistic

Unknown $\boldsymbol{\mu}$
Unknown $\boldsymbol{\sigma^2}$

$$\bar{t} = \frac{\overline{m} - \mu}{\bar{s} / \sqrt{n}}$$

$\bar{t}$ statistic

Case A: unknown $\boldsymbol{\mu}$, known $\boldsymbol{\sigma^2}$
$\bar{z}$ *statistic*

$$\boxed{\frac{\overline{m}-\mu}{\sigma/\sqrt{n}}=\overline{z}}$$

$$Pr\{-z_0 \le \overline{z} \le +z_0\} = 95\%$$

From tables $\boldsymbol{z_{0.05}}$ **= 1.96**, hence:

$$Pr\left\{-1.96 \le \frac{\overline{m}-\mu}{\sigma/\sqrt{n}} \le +1.96\right\} = 95\%$$

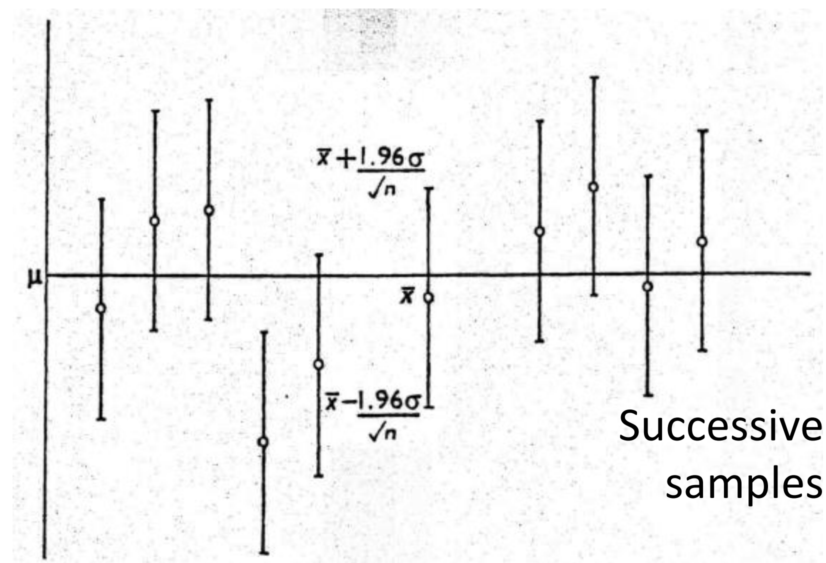
$$Pr\left\{\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \overline{m} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 95\%$$

$$Pr\left\{\overline{m} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{m} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 95\%$$

Thus **95% CI** is given by:

$$\mu = m \pm z_{0.05}\frac{\sigma}{\sqrt{n}} = m \pm 1.96\frac{\sigma}{\sqrt{n}}$$

Practical interpretation of **95% CI**



95% of CI include actual $\boldsymbol{\mu}$ (unknown)

Case B: unknown $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$

$\bar{t}$ *statistic* (i.e. use $\bar{\boldsymbol{s}}$ instead of $\boldsymbol{\sigma}$)

$$\bar{t} = \frac{\overline{m} - \mu}{\bar{s}/\sqrt{n}}$$

$$Pr\{-t_{v,0} \leq \bar{t} \leq +t_{v,0}\} = 95\%$$

$t_{v,0.05}$ from **tables** (**v = n-1**)

$$Pr\left\{-t_{v,0.05} \leq \frac{\overline{m} - \mu}{s/\sqrt{n}} \leq +t_{v,0.05}\right\} = 95\%$$

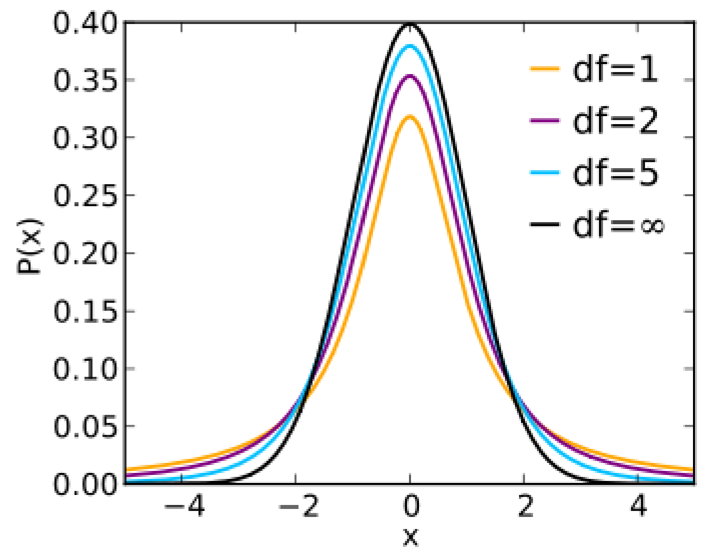
$$Pr\left\{\mu - t_{v,0.05}\frac{s}{\sqrt{n}} \leq \overline{m} \leq \mu + t_{v,0.05}\frac{s}{\sqrt{n}}\right\} = 95\%$$

$$Pr\left\{\overline{m} - t_{v,0.05}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{m} + t_{v,0.05}\frac{s}{\sqrt{n}}\right\} = 95\%$$

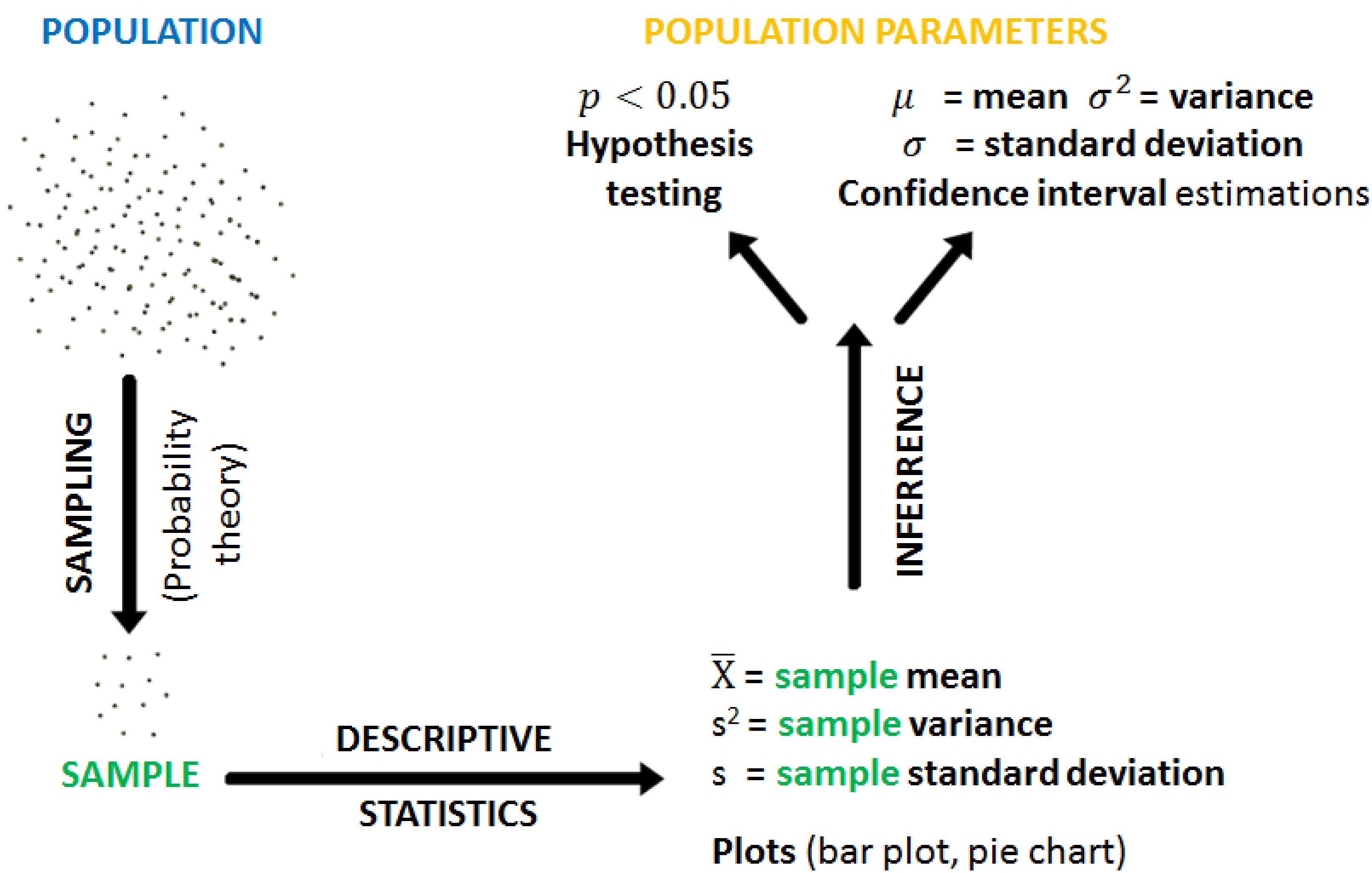
Thus **95% CI** is given by:

$$\mu = m \pm t_{v,0.05}\frac{s}{\sqrt{n}}$$

Student's *t*-distribution



$$t_{\alpha,\infty} = z_{\alpha}$$



POPULATION
POPULATION PARAMETERS
$p < 0.05$
Hypothesis testing
μ = mean σ^2 = variance
σ = standard deviation
Confidence interval estimations
SAMPLING
(Probability theory)
INFERRENCE
$\overline{X}$ = sample mean
s^2 = sample variance
s = sample standard deviation
SAMPLE
DESCRIPTIVE
STATISTICS
Plots (bar plot, pie chart)

Hypothesis testing

- **H_0 = null** hypothesis → the **sample belongs** to a **known population** (with known $\boldsymbol{\mu}$ and, eventually, $\boldsymbol{\sigma^2}$)

- **H_1 = alternative** hypothesis → the **2 treatments** are **different** each other

- **Hypothesis test** evaluates the **discrepancy** between the **sample** and the **H_0,** establishing whether it is statistically i) **significant** or ii) **not significant** for a **significance level α**

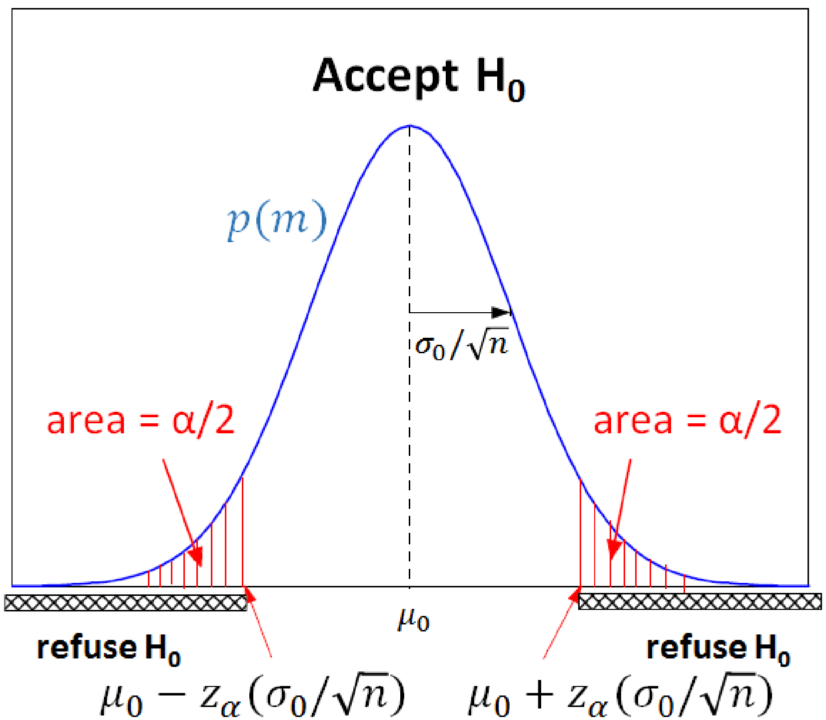
    i) H_0 is refused with a **significance level α**

    ii) H_0 cannot be **refused** with a **significance level α**

Case A: unknown $\boldsymbol{\mu}$, known $\boldsymbol{\sigma^2}$

$\bar{z}$ *statistic (**z-test**)*

- Mean survival time from the diagnosis of a given disease
  - **Population** = **38.3** ± **43.3** months ($\boldsymbol{\mu_0} \pm \boldsymbol{\sigma_0}$)
  - **100** patients treated with a **new technique** = **46.9** months ($\overline{\boldsymbol{m}}$) and $\boldsymbol{\sigma} = \boldsymbol{\sigma_0}$
- **$H_0 \rightarrow \mu = \mu_0$** *or* **$H_1 \rightarrow \mu \neq \mu_0$**



$$\bar{z} = \frac{\overline{m} - \mu_0}{\sigma_0/\sqrt{n}} = \frac{46.9 - 38.3}{43.3/\sqrt{100}} = \frac{8.6}{4.33} = 1.99$$

H_0 is refused with a **significance level α if $\bar{z}$ < - $z_{0.05}$ or $\bar{z}$ > $z_{0.05}$**

Since $\boldsymbol{z_{0.05} = 1.96}$ and $\boldsymbol{z_{0.01} = 2.58}$ **what can we say?**

CI estimations and hypothesis testing are equivalent

95% CI $m - 1.96\frac{\sigma}{\sqrt{n}} < \mu < m + 1.96\frac{\sigma}{\sqrt{n}}$ 46.9±1.96·4.33= 38.4 ÷ 55.4

$\overline{m}$ (38.3) **< *μ^-* → *refuse H_0***

99% CI $m \pm 2.58\frac{\sigma}{\sqrt{n}}$ = 46.9±2.58·4.33 =35.7±58.07

μ^-* < *$\overline{m}$ (38.3) **< *μ^+* → *H_0 cannot*** *be* ***refused***

A **confidence interval** can be considered as the **set of acceptable hypotheses** for a **certain level** of **significance**

Case b: unknown μ and σ^2
$\bar{t}$ *statistic* (***t-test***)

- Rat uterine weight
  - **Population** = **24** mg (μ_0)
  - n**=20** rats: [9, 14, 15, 15, 16, 18, 18, 19, 19, 20, 21, 22, 22, 24, 24, 26, 27, 29, 30, 32]
  - $\nu = n - 1 = 19$

- $H_0 \rightarrow \mu = \mu_0$ *or* $H_1 \rightarrow \mu \neq \mu_0$

$$\bar{t} = \frac{\overline{m} - \mu_0}{\overline{s}/\sqrt{n}} = \frac{21 - 24}{1.3219} = -2.27$$

Since $t_{19,\,0.05}$ **= 2.093** and $t_{19,\,0.02}$ **= 2.539** **what can we say?**

- **Equivalence** between ***t-test*** and **CI estimations**

$$m - t_{\nu,0.05}\frac{s}{\sqrt{n}} < \mu < m + t_{\nu,0.05}\frac{s}{\sqrt{n}}$$

95% CI $21 \pm 2.093 \cdot (1.3219) = 18.23 \div 23.77$

98% CI $21 \pm 2.539 \cdot (1.3219) = 17.64 \div 24.36$

Sample and **population** are **significantly different** with a **significance level** comprised between **2 %** and **5 %** (**0.02 <** p **< 0.05**; calculated *p-value* for $t_{19,\,p}$ **= 2.27** is p **= 0.035**)

MATLAB
z-test

H = 0, H_0 cannot be refused at **α**
H = 1, refuse H_0 at α

Confidence interval for the «true» value ***μ*** at a level **1 - α**

z-statistic value

significance level

[H,P,CI,ZVAL] = ZTEST(X,mean,sigma,alpha,tail)

p-value *(i.e. the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true)*

sample

population parameters

'both' → "$\overline{X}$ is not mean" (two-tailed test)
'right' → "$\overline{X}$ is greater than mean" (right-tailed test)
'left' → "$\overline{X}$ is less than mean" (left-tailed test)

MATLAB
z-test: example

```
>> X=[8.3 9.2 12.5 7.6 10.2 12.9 11.7 10.8 11.7 9.6];
>> sigma=2.1;
>> mean=12;
>> alpha=0.05;
>> [H,P,CI,ZVAL]=ztest(X,mean,sigma,alpha)
```

MATLAB
t-test

H = 0, H_0 cannot be refused at **α**
H = 1, refuse H_0 at α

Confidence interval for the «true» value ***μ*** at a level **1 - α**

Data **structure** containing **t-statistics value** and **number of DoF**

significance level

[H,P,CI,STATS] = TTEST(X,mean,alpha,tail)

p-value *(i.e. the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true)*

sample

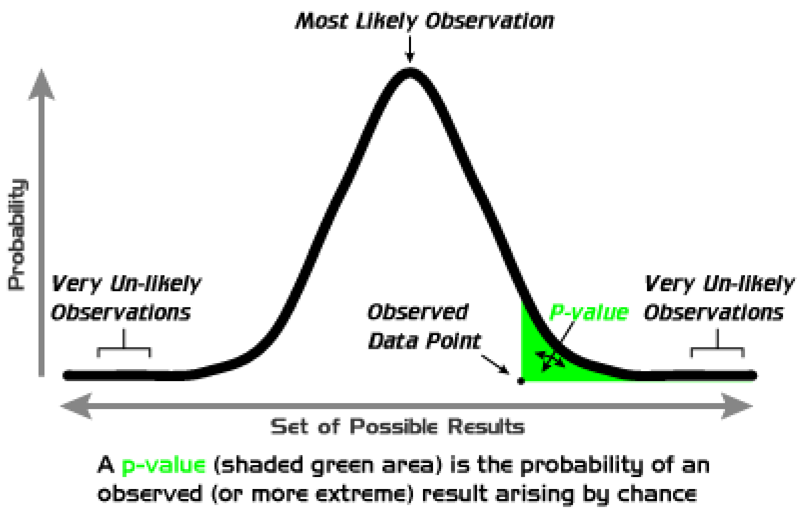
population mean

'both' → "$\overline{X}$ is not mean" (two-tailed test)
'right' → "$\overline{X}$ is greater than mean" (right-tailed test)
'left' → "$\overline{X}$ is less than mean" (left-tailed test)

MATLAB
t-test: example

```
>> X=[22.3 25.1 27 23.4 24.7 26.5 25.7 24.1 23.9 22.8];
>> mean=23;
>> alpha=0.05;
>> [H,P,CI,STAT]=ttest(X,mean,alpha)
```

Interpreting the *p*-value



In conclusion, the **smaller** the ***p*-value** the **more statistical evidence** exists to **support** the **alternative hypothesis** (H_1)