Analysing and presenting data: **practical hints**

Giorgio MATTEI

giorgio.mattei@centropiaggio.unipi.it

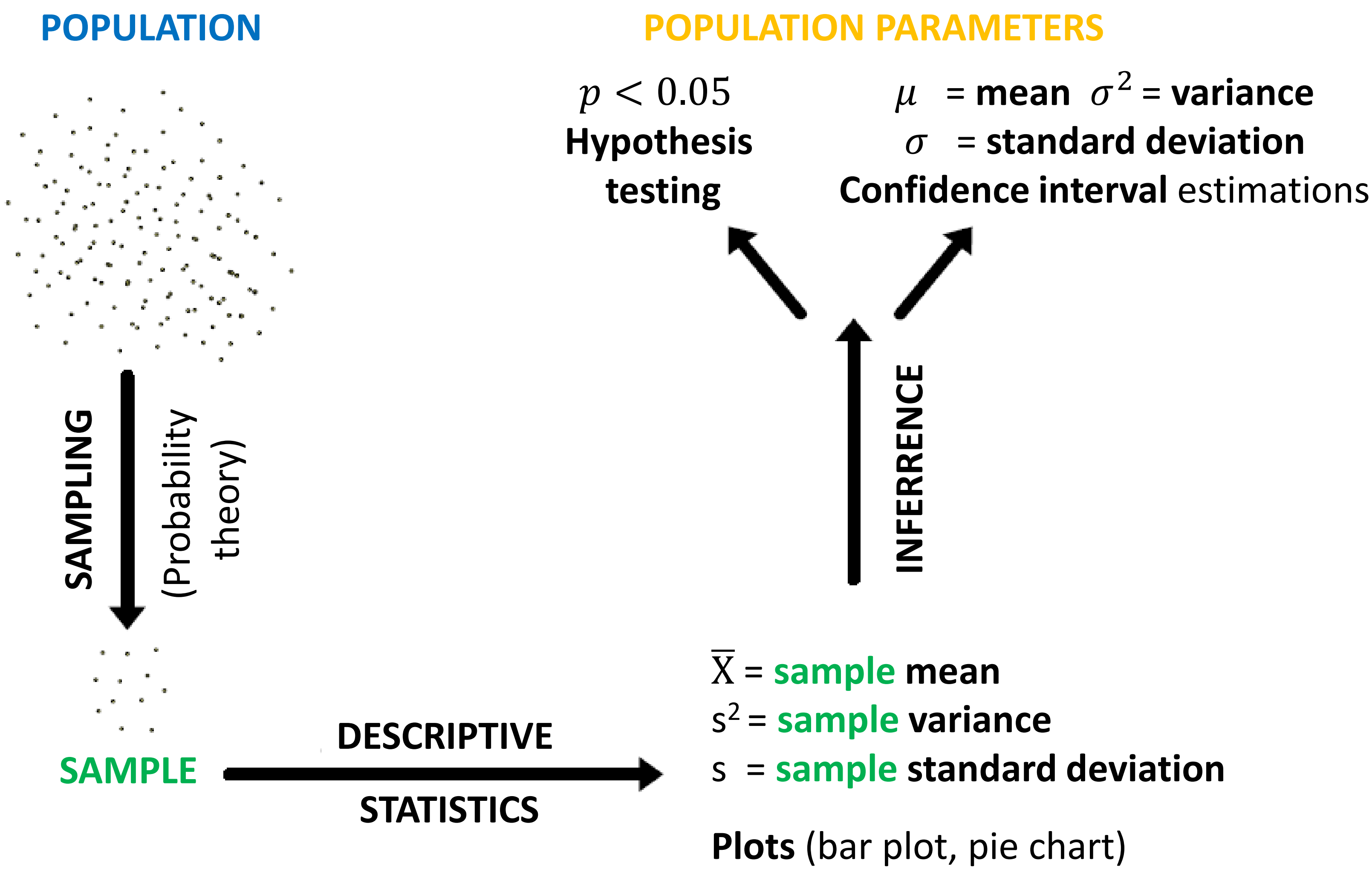
Course: Fenomeni di trasporto biologico
Date: 06 Oct 2014

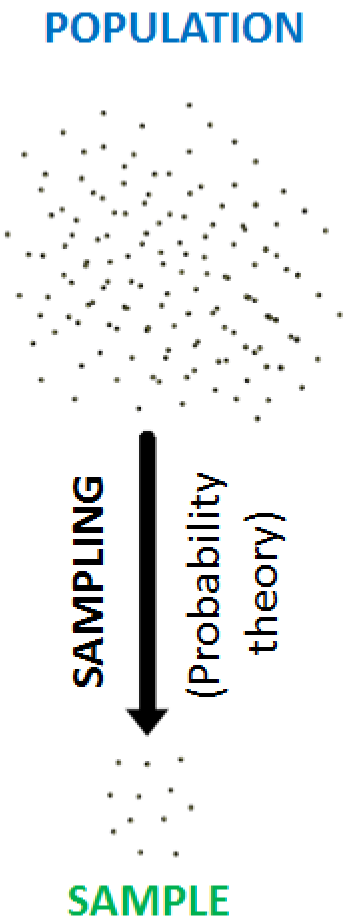
What is **statistics**?

Statistics is the study of the **collection, organization, analysis, interpretation, and presentation** of data. It deals with all aspects of this, including the **planning of data collection** in terms of the **design of surveys and experiments**. [*Wikipedia*]

- In general, **the population is too large** to be studied in its entirety → a **sample of *n* individuals** is extracted from the same population as a representative to study its properties

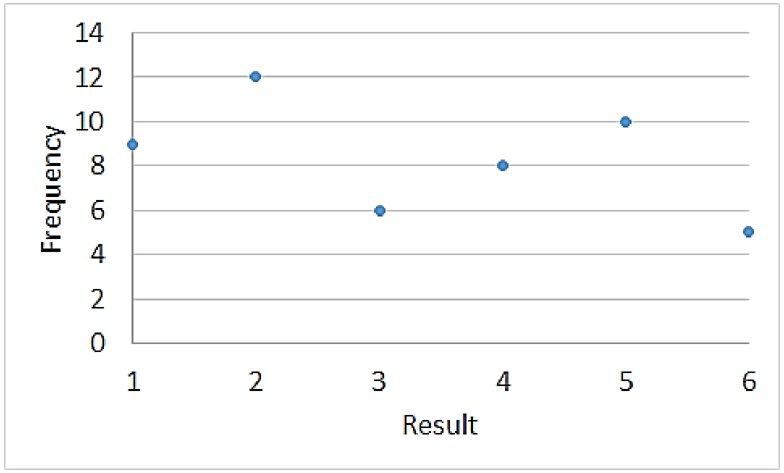
The statistical process

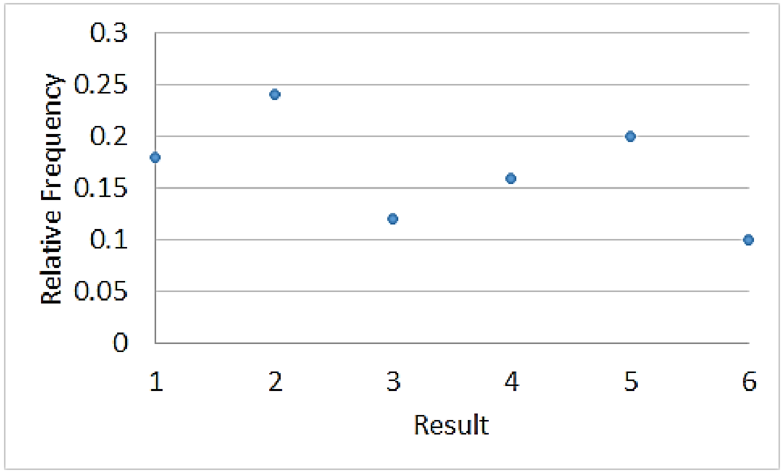



POPULATION
SAMPLING
(Probability theory)
SAMPLE

Tables and frequency graphs
Discrete domain: dice throw

Result	Frequency (n)	Relative frequency (n/N)
1	9	0.18
2	12	0.24
3	6	0.12
4	8	0.16
5	10	0.2
6	5	0.1
TOTAL	**50**	**1**





Tables and frequency graphs
Continuous domain: human height

Interval	Central value	Frequency	Relative frequency
141.5-148.5	145	2	0.01
148.5-155.5	152	7	0.035
155.5-162.5	159	22	0.11
162.5-169.5	166	13	0.065
169.5-176.5	173	44	0.22
176.5-183.5	180	36	0.18
183.5-190.5	187	32	0.16
190.5-197.5	194	13	0.065
197.5-204.5	201	21	0.105
204.5-211.5	208	10	0.05

Need to group data defining **histogram bins**

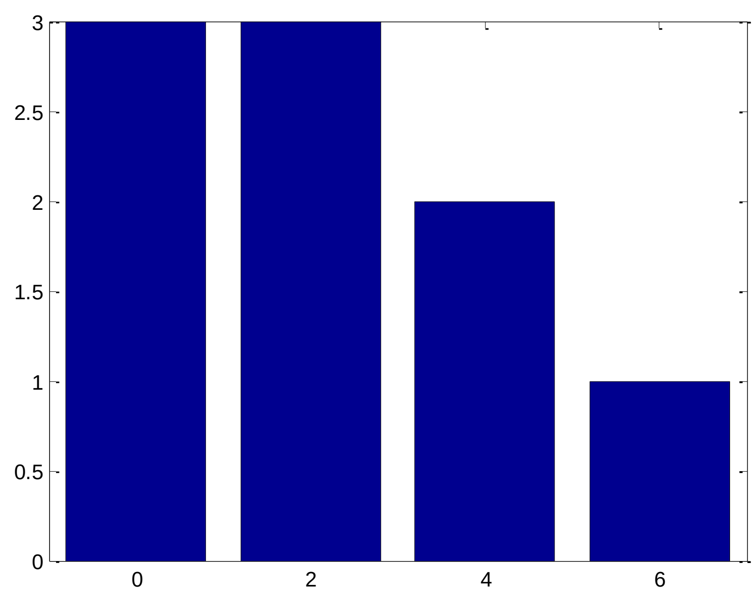
There is no best/optimal number of bins and **different bin sizes can reveal different features of the data**

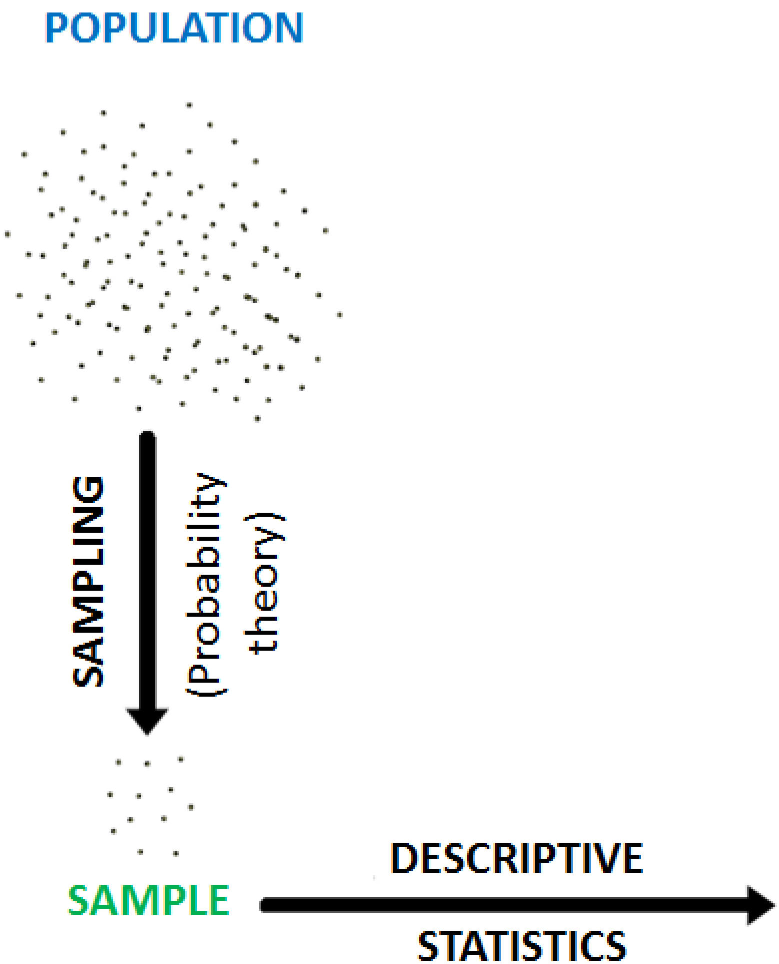
- ✓ Methods for determining optimal number of bins generally make strong assumptions about the shape of the distribution
- ✓ **Appropriate bin widths should be experimentally determined depending on the actual data distribution and the goals of the analysis**
- ✓ However there are various useful guidelines and rules of thumb

MATLAB
Frequency graphs

- **stem(X,Y)** *discrete variables*
- **bar(X,Y)** *continuous variables*
  - **f=histc(X, edges)** *number of elements between edges*

```
>> X=[0.5 1 1.2 2.1 3 3.2 4.6 5 6];
>> edges=[0 2 4 6];
>> f=histc(X,edges)

f =
    3    3    2    1

>> bar(edges,f)
```





$\overline{X}$ = sample **mean**

s^2 = sample **variance**

s = sample **standard deviation**

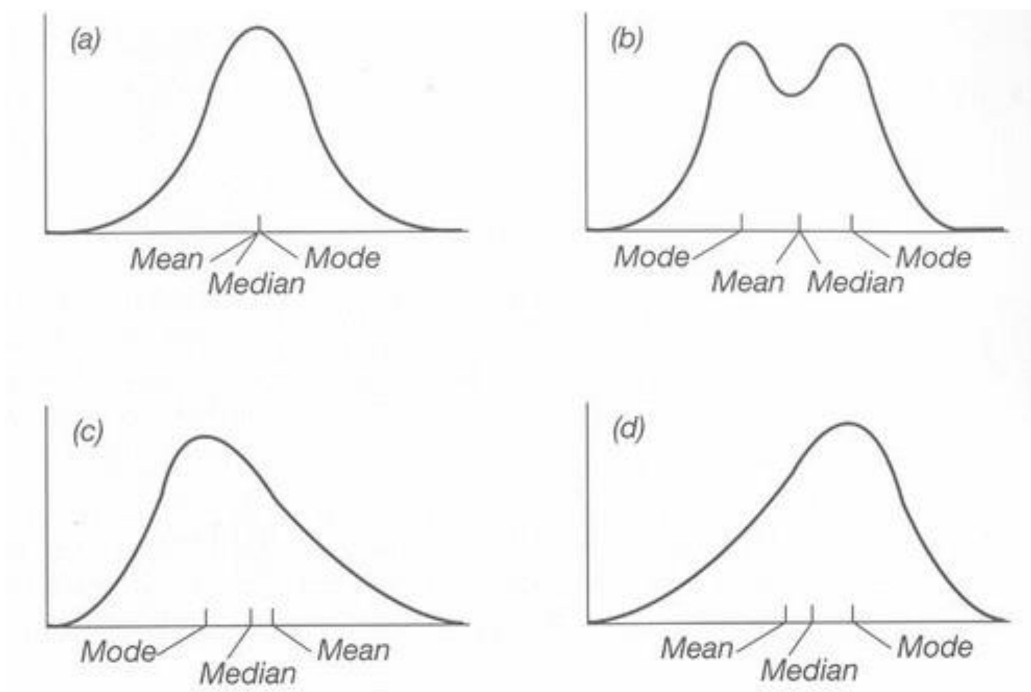
Plots (bar plot, pie chart)

Position (or **central tendency**) *mode, median* and *mean*

- **Mode**: the value(s) that occurs most often
- **Median**: the middle value of a data set arranged in ascending order
- **Arithmetic mean**: sum of all of the data values divided by their number

Simmetric (unimodal)

Simmetric (bimodal)



Positively skewed (unimodal)

Negatively skewed (unimodal)

Mean (m) calculation

What we know?

Case A: values (x_i) of each of the n observations

$$m = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Case B: x_i are not known: n data grouped in k intervals

$$m \cong \frac{1}{n} \cdot \sum_{i=1}^{k} f_i x_i = \sum_{i=1}^{k} x_i \left(\frac{f_i}{n} \right)$$

where f_i is the number of observation within the interval centred on the value x_i

Dispersion (or **scatter**) *variance* and *standard deviation*

- **The measure of scatter should be**
  - **proportional to the scatter of the data** (small when the data are clustered together, and large when the data are widely scattered)
  - **independent of the number of values in the data set** (otherwise, simply by taking more measurements the value would increase even if the scatter of the measurements was not increasing).
  - **independent of the mean** (since now we are only interested in the spread of the data, not its central tendency)
- Both the **variance** and the **standard deviation meet these three criteria** for **normally-distributed** data sets

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - m)^2$$

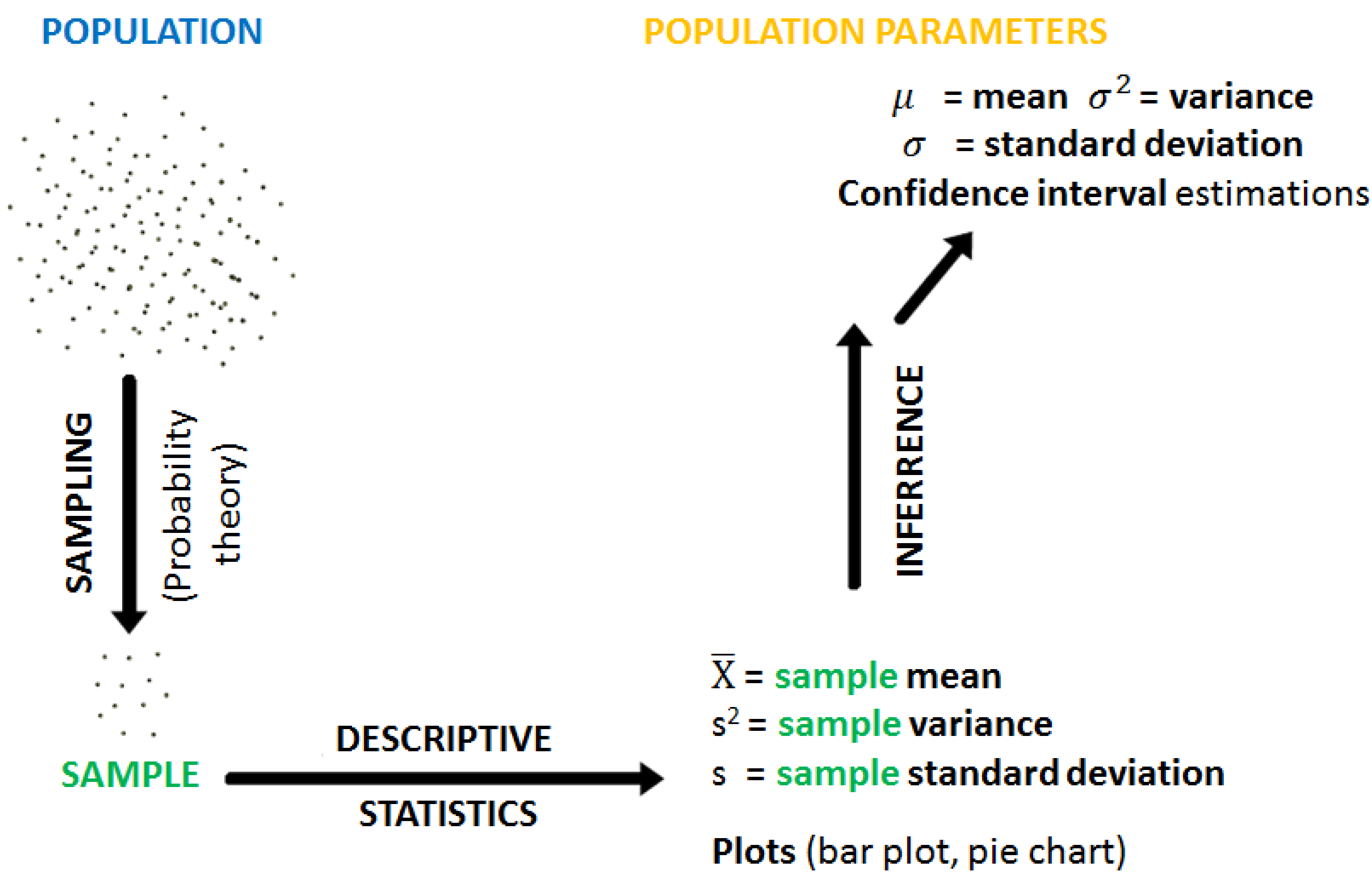
Variance

$$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - m)^2}$$

Standard deviation

MATLAB
Position and dispersion

- **mode(X)**
- **median(X)**
- **mean(X)**
- **var(X)**
- **std(X)**
  - Note that **std(X) = sqrt(var(X))**


POPULATION
POPULATION PARAMETERS
μ = mean σ² = variance
σ = standard deviation
Confidence interval estimations
SAMPLING
(Probability theory)
INFERENCE
SAMPLE
DESCRIPTIVE
STATISTICS
X̄ = sample mean
s² = 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

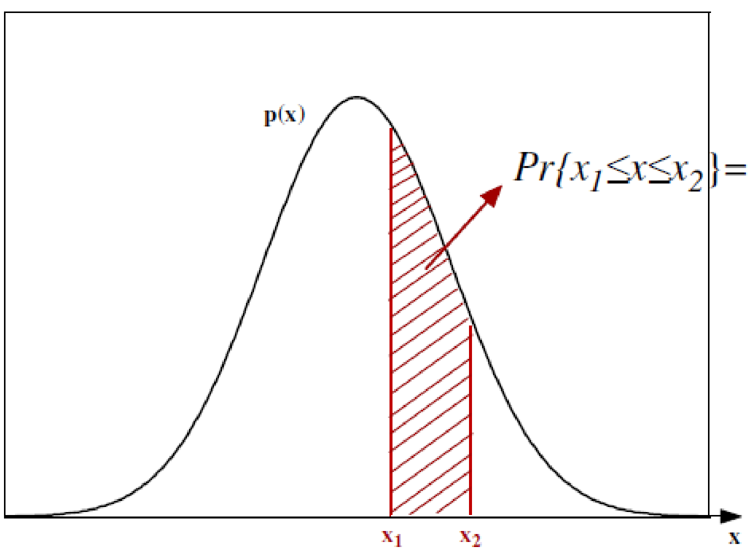
$$Pr\{S\} = P(S) = 1$$

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 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**

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

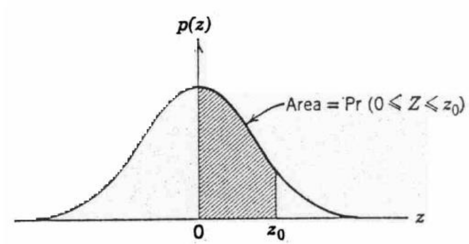
$$\sigma^2 = \int (x-\mu)^2 p(x) dx = E\{(x-\mu)^2\} = Var(\bar{x})$$

- $\bar{x}$ 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 (μ=0, σ^2=1) and variable standardisation

- **Standardised normal probability density**

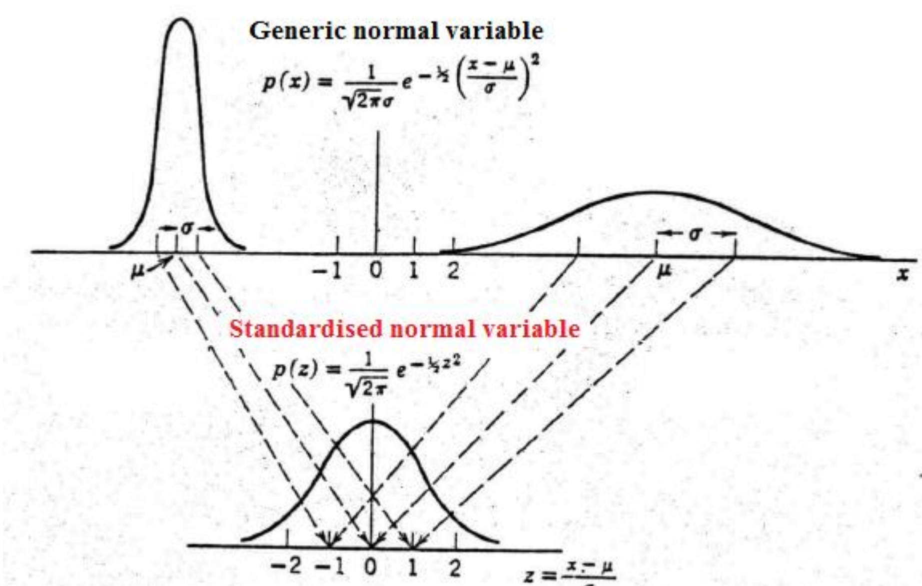


$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

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

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

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



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

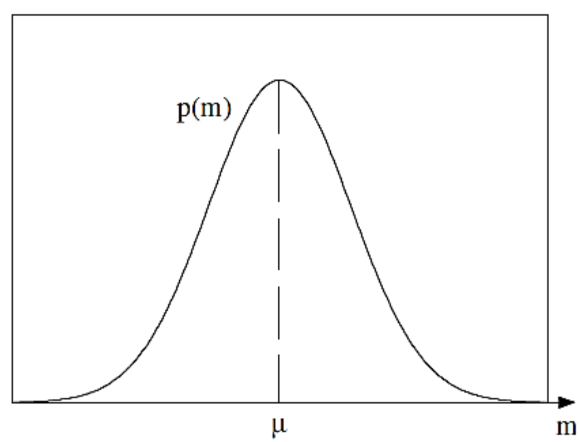
Inference

- **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 $\boldsymbol{\sigma^2}$	Unknown $\boldsymbol{\mu}$ Unknown $\boldsymbol{\sigma^2}$
$\dfrac{\overline{m} - \mu}{\sigma / \sqrt{n}} = \bar{z}$	$\bar{t} = \dfrac{\overline{m} - \mu}{\bar{s} / \sqrt{n}}$
$\bar{z}$ statistic	$\bar{t}$ 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:

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

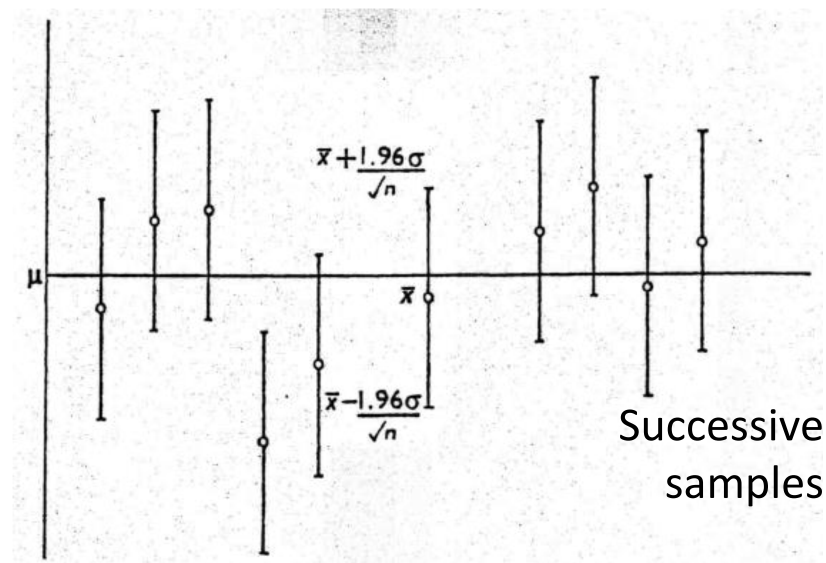
$$Pr\left\{\mu - 1.96\,\frac{\sigma}{\sqrt{n}} \le \overline{m} \le \mu + 1.96\,\frac{\sigma}{\sqrt{n}}\right\} = 95\%$$

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

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 $\boldsymbol{\sigma}$)

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

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

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

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

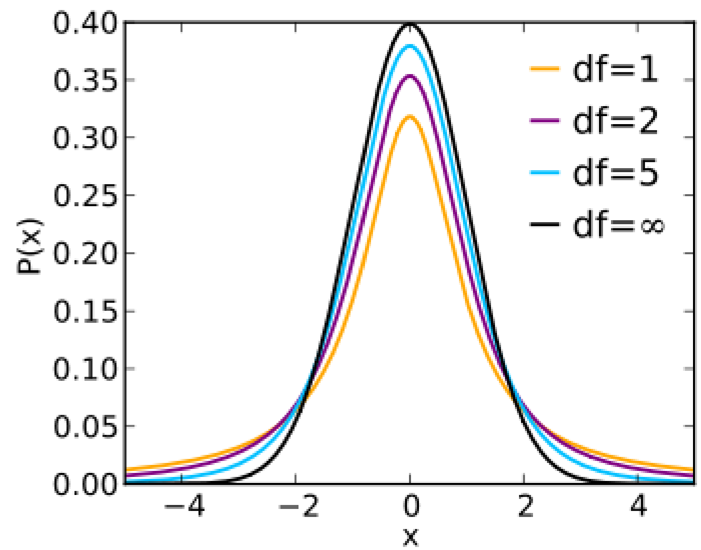
$$Pr\left\{\mu - t_{v,0.05}\frac{s}{\sqrt{n}} \leq \overline{m} \leq \mu + t_{v,0.05}\frac{s}{\sqrt{n}}\right\} = 95\%$$

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

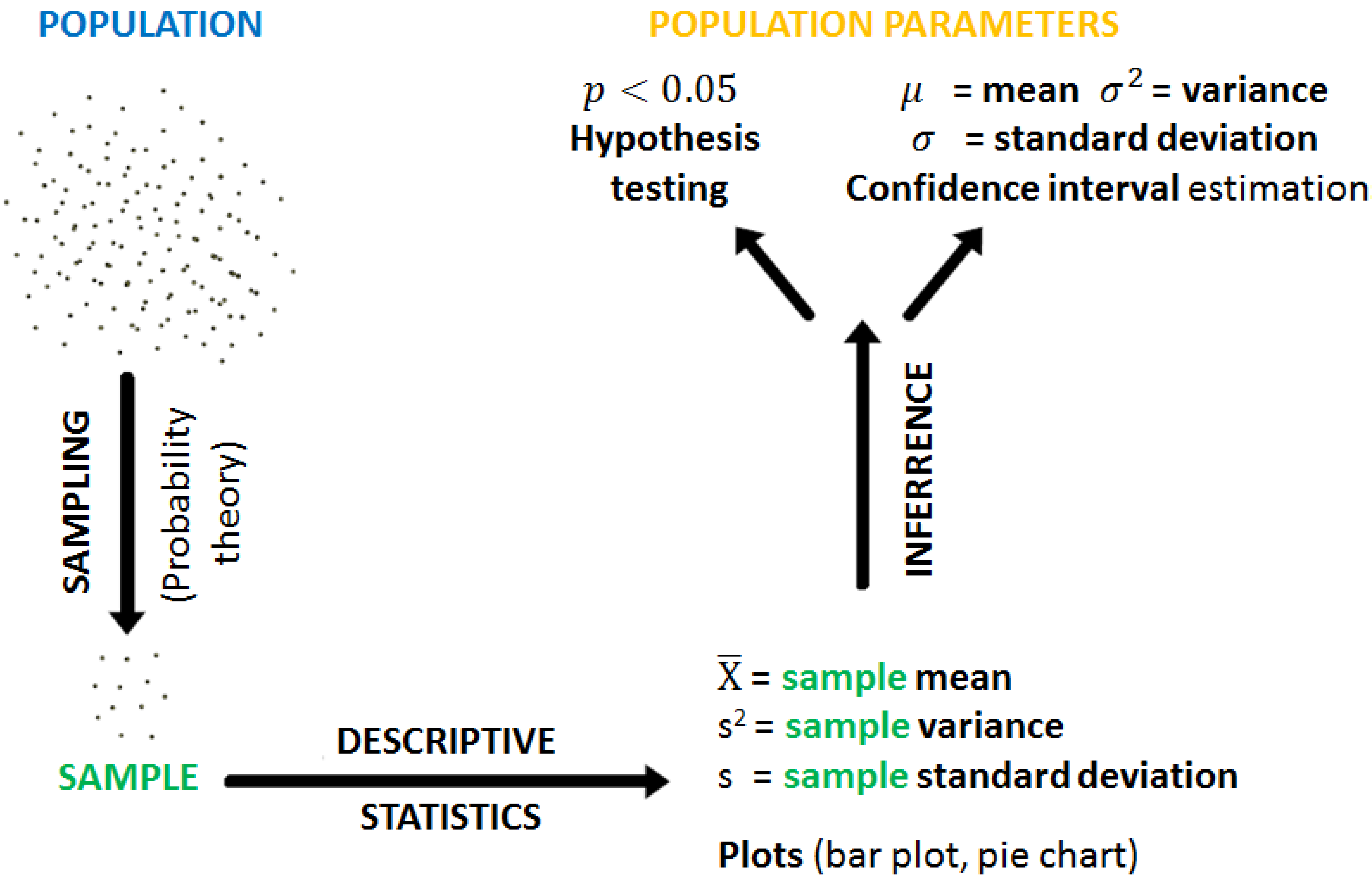
Thus **95% CI** is given by:

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

Student's *t*-distribution



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



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

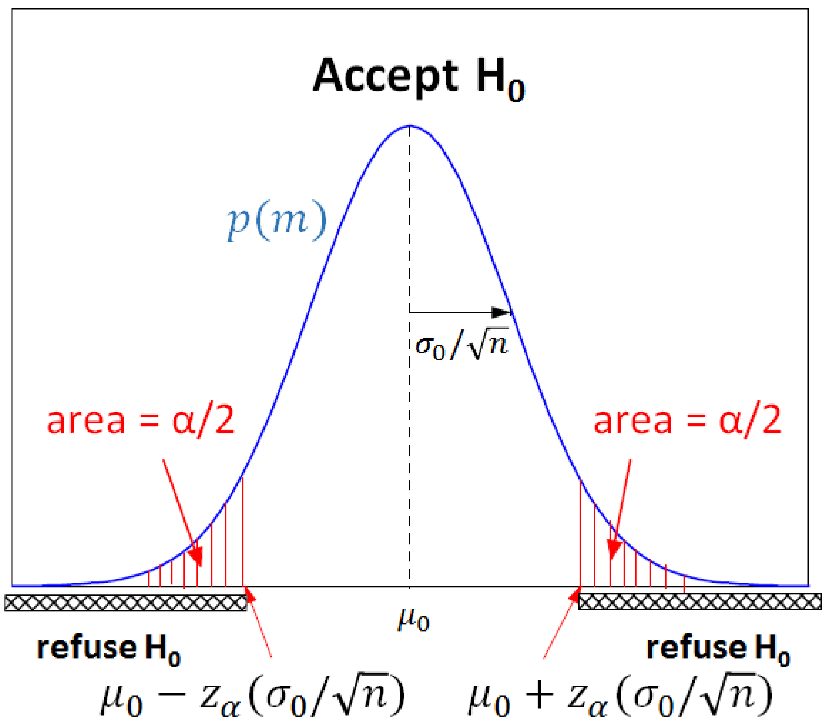
- **Hypothesis test** evaluates the **discrepancy** between the **sample** and the **H_0**, establishing whether it is statistically i) **significant** or ii) **not significant** for a **significance level α**

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

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

Case A: unknown μ_0, known σ_0
$\bar{z}$ *statistic (**z-test**)*

- Mean survival time from the diagnosis of a given disease
  - **Population** = **38.3** ± **43.3** months ($\mu_0 \pm \sigma_0$)
  - **100** patients treated with a **new technique** = **46.9** months ($\bar{m}$)
- **H$_0$ → $\bar{m} = \mu_0$** and $\bar{s} = \sigma_0$ and **H$_1$ → $\bar{m} \neq \mu_0$**



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

H$_0$ is refused with a **significance level α if $\bar{z}$ < - $z_{0.05}$ or $\bar{z}$ > $z_{0.05}$**

Since ***$z_{0.05}$ = 1.96*** and ***$z_{0.01}$ = 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

$\bar{m}$ (38.3) $< \mu^-$ → ***refuse*** H_0

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

$\mu^- < \bar{m}$ (38.3) $< \mu^+$ → 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_0, unknown σ_0
$\bar{t}$ *statistic (**t-test**)*

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

- **H_0 → $\bar{m} = \mu_0$** ~~and $\bar{s} = \sigma_0$~~

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

Since $\boldsymbol{t_{19,\,0.05}}$ **= 2.093** and $\boldsymbol{t_{19,\,0.02}}$ **= 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 $\boldsymbol{t_{19,\,p}}$ **= 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)

sample

population parameters

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)*

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

H = 1

P = 0.0196

CI = 9.1484   11.7516

ZVAL = -2.3341
```

What happens using $\boldsymbol{\alpha} = 0.01$?

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 $\boldsymbol{\mu}$ 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

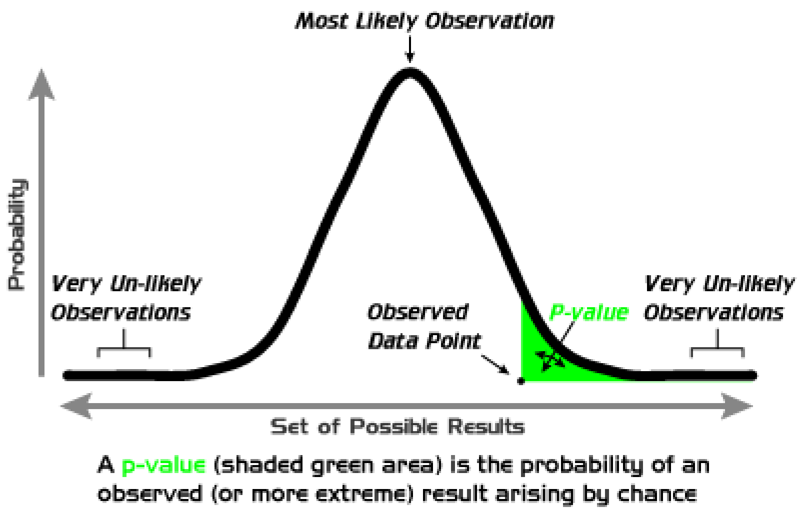
'both' → "$\overline{X}$ is not mean" (two-tailed test)
'right' → "$\overline{X}$ is greater than mean" (right-tailed test)
'left' → "$\overline{X}$ is less than mean" (left-tailed test)

MATLAB
t-test: example

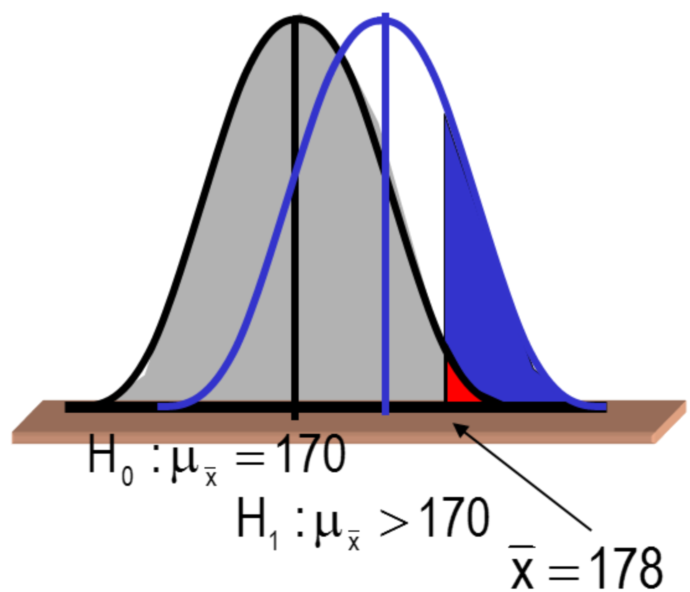
```
>> X=[22.3 25.1 27 23.4 24.7 26.5 25.7 24.1 23.9 22.8];
>> mean=23;
>> alpha=0.05;
>> [H,P,CI,STAT]=ttest(X,mean,alpha)

H = 1

P = 0.0114

CI = 23.4437   25.6563

STAT =   tstat: 3.1694
         df: 9
         sd: 1.5465
```

What happens using **α** = 0.01?

Interpreting the *p*-value



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



Equal or different?

*The case of **two samples***

Independent two-sample *t-test*

__Equal__ sample sizes (__n__), __equal__ variances ($S_{X_1X_2}$)

The ***t* statistic** to test whether the **means** of **group 1** ($\boldsymbol{\overline{X_1}}$) and **group 2** ($\boldsymbol{\overline{X_2}}$) **are different** can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

$$S_{X_1X_2} = \sqrt{\frac{1}{2}(S_{X_1}^2 + S_{X_2}^2)}$$

«pooled» standard deviation

t-test DoFs = 2n - 2

H_0 is refused with a **significance level α if**

$t < -\,t_{DoF,\alpha}$ or $t > t_{DoF,\alpha}$

Independent two-sample *t-test*

Unequal sample sizes (*n*$_1$ and *n*$_2$), *equal* variances ($S_{X_1X_2}$)

The ***t* statistic** to test whether the **means** of **group 1** ($\overline{X_1}$) and **group 2** ($\overline{X_2}$) **are different** can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad S_{X_1X_2} = \sqrt{\frac{(n_1 - 1)S^2_{X_1} + (n_2 - 1)S^2_{X_2}}{n_1 + n_2 - 2}}$$

«pooled» standard deviation

t-test DoFs = $n_1 + n_2 - 2$

H_0 is refused with a **significance level α if**
$t < -t_{DoF,\alpha}$ or $t > t_{DoF,\alpha}$

Independent two-sample *t-test*

Unequal sample sizes (***n₁*** *and* ***n₂***)*,* ***unequal*** *variances* ($S_{X_1X_2}$)

The ***t* statistic** to test whether the **means** of **group 1** ($\overline{X_1}$) and **group 2** ($\overline{X_2}$) **are different** can be calculated as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

«unpooled» standard deviation

$$t\text{-}test \text{ DoFs} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

Welch–Satterthwaite equation

H_0 is refused with a **significance level α if**
$t < -t_{DoF,\alpha}$ or $t > t_{DoF,\alpha}$

Independent two-sample *t-test* (*unequal sample sizes* and *equal variances*): an *example*

- Two groups of 10 *Dapnia magna* eggs, randomly extracted from the same clone, were reared in two different concentrations of hexavalent chromium
- After a month survived individuals were measured: 7 in group A and 8 in group B

	A	B
	2,7	2,2
	2,8	2,1
	2,9	2,2
	2,5	2,3
	2,6	2,1
	2,7	2,2
	2,8	2,3
		2,6
Mean	2.714	2.250

$$s_p^2 = \frac{0{,}10825 + 0{,}18000}{6 + 7} = 0{,}022173$$ «pooled» variance

$$t_{13} = \frac{2{,}714 - 2{,}250}{\sqrt{0{,}022173 \cdot \left(\frac{1}{7} + \frac{1}{8}\right)}} = 6{,}02$$ **t** with 13 DoF

Since $\boldsymbol{t_{13,\,0.05}}$ **= 2.160 what can we say?**

MATLAB

Independent two-sample t-test

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

Confidence interval for the «true» **difference** of **population means**

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

significance level

$$[H,P,CI,STATS] = TTEST2(\underbrace{X,Y}_{\textbf{samples}},alpha,tail,vartype)$$

p-value *(i.e. the probability of observing the given result, or one more extreme, by chance if the null hypothesis is true)*

'equal' or **'unequal'**

'both' → “means are not equal" (two-tailed test)
'right' → "$\overline{X}$ is greater than $\overline{Y}$" (right-tailed test)
'left' → "$\overline{X}$ is less than $\overline{Y}$" (left-tailed test)

MATLAB

Ind. 2-sample t-test: an example

```
>> X=[2.7 2.8 2.9 2.5 2.6 2.7 2.8]';
>> Y=[2.2 2.1 2.2 2.3 2.1 2.2 2.3 2.6]';
>> [H,P,CI,STATS] = ttest2(X,Y,0.05,'both','equal')

H =  1
P =  4.2957e-05
CI = 0.2977    0.6309
STATS =
    tstat: 6.0211
       df: 13
       sd: 0.1490
```

Dependent two-sample *t-test*

one sample tested twice or *two "paired" samples*

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}}$$

- ✓ Calculate the differences between all ***n*** pairs (X_D), then substitute their average ($\overline{X_D}$) and standard deviation (s_D) in the equation above to test if the average of the differences is significantly different from $\boldsymbol{\mu_0}$ ($\boldsymbol{\mu_0}$ **= 0** under $\mathbf{H_0}$**, DoFs =** ***n*** **- 1**)
- ✓ The "pairs" can be either one person's pre-test and post-test scores (repeated measures) or persons matched into meaningful groups (e.g. same age)

Example of repeated measures

Number	Name	Test 1	Test 2
1	Mike	35%	67%
2	Melanie	50%	46%
3	Melissa	90%	86%
4	Mitchell	78%	91%

Example of matched pairs

Pair	Name	Age	Test
1	John	35	250
1	Jane	36	340
2	Jimmy	22	460
2	Jessy	21	200

Dependent two-sample *t-test*: an example

Student	Pre-module score	Post-module score	Difference
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

$$t = \frac{2.05}{0.634} = 3.231 \qquad \text{on 19 df}$$

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

MATLAB

Dependent two-sample t-test

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

Confidence interval for the «true» **difference** of **population means**

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

significance level

$$\underbrace{[H,P,CI,STATS]}\ = \mathrm{TTEST}(\underbrace{X,Y}_{\textbf{samples}},alpha,tail)$$

p-value *(i.e. the probability of observing the given result, or one more extreme, by chance if the null hypothesis is true)*

'both' → “means are not equal" (two-tailed test)
'right' → "$\overline{X}$ is greater than $\overline{Y}$" (right-tailed test)
'left' → "$\overline{X}$ is less than $\overline{Y}$" (left-tailed test)

MATLAB
Dep. 2-sample t-test: an example

```
>> X=[22 25 17 24 16 29 20 23 19 20 15 15 18 26 18 24 18 25 19 16]';
>> Y=[18 21 16 22 19 24 17 21 23 18 14 16 16 19 18 20 12 22 15 17]';
>> [H,P,CI,STATS] = ttest(X,Y,0.05,'both')

H =  1
P =  0.0044
CI = 0.7221    3.3779
STATS =
    tstat: 3.2313
       df: 19
       sd: 2.8373
```