Microarray Center

STATISTICAL DATA ANALYSIS IN EXCEL

Lecture 3

Testing of Hypotheses for Means

dr. Petr Nazarov

petr.nazarov@crp-sante.lu

31-10-2011

- Hypotheses for means
- Unpaired t-test
- Paired t-test

`http://edu.sablab.net/data`

Hypotheses for Mean

Here we continue the discussion of statistical inference by showing how hypothesis testing can be used to determine whether a statement about the value of a population parameter should or should not be rejected.

In hypothesis testing we begin by making a tentative assumption about a population parameter, i.e. by formulation of a null hypothesis.

Null hypothesis
The hypothesis tentatively assumed true in the hypothesis testing procedure, H_0

Alternative hypothesis
The hypothesis concluded to be true if the null hypothesis is rejected, H_a

H_0: $\mu \leq$ const
H_a: $\mu >$ const

H_0: $\mu \geq$ const
H_a: $\mu <$ const

H_0: $\mu =$ const
H_a: $\mu \neq$ const

H_0: $\mu_1 \leq \mu_2$
H_a: $\mu_1 > \mu_2$

H_0: $\mu_1 \geq \mu_2$
H_a: $\mu_1 < \mu_2$

H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

Type I error
The error of rejecting H_0 when it is true.

Type II error
The error of accepting H_0 when it is false.

Level of significance
The probability of making a Type I error when the null hypothesis is true as an equality

poor sensitivity

False Negative, β error

		Population Condition	
		H_0 True	H_a True
Conclusion	Accept H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

False Positive, α error

poor specificity

One-tailed test

A hypothesis test in which rejection of the null hypothesis occurs for values of the test statistic in one tail of its sampling distribution

H_0: $\mu \le \mu_0$ H_0: $\mu \ge \mu_0$

H_a: $\mu > \mu_0$ H_a: $\mu < \mu_0$

A Trade Commission (TC) periodically conducts statistical studies designed to test the claims that manufacturers make about their products. For example, the label on a large can of Hilltop Coffee states that the can contains 3 pounds of coffee. The TC knows that Hilltop's production process cannot place exactly 3 pounds of coffee in each can, even if the mean filling weight for the population of all cans filled is 3 pounds per can. However, as long as the population mean filling weight is at least 3 pounds per can, the rights of consumers will be protected. Thus, the TC interprets the label information on a large can of coffee as a claim by Hilltop that the population mean filling weight is at least 3 pounds per can. We will show how the TC can check Hilltop's claim by conducting a lower tail hypothesis test.

μ_0 = 3 lbm   Suppose sample of n=36 coffee cans is selected. From the previous studies it's known that σ = 0.18 lbm

One-tailed Test: Example

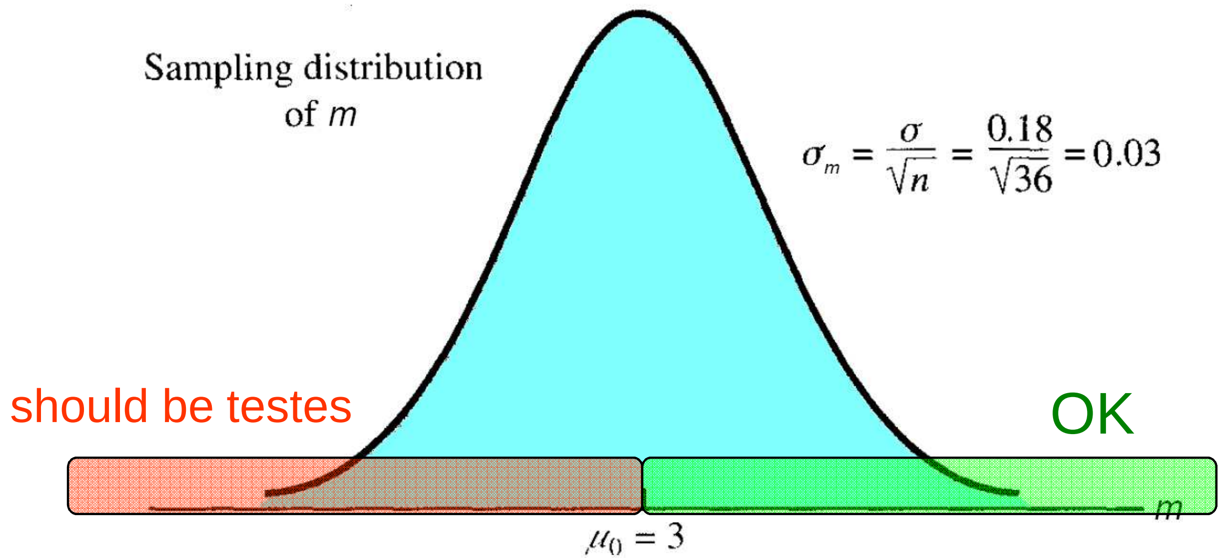
μ_0 = 3 lbm

Suppose sample of n = 36 coffee cans is selected and m = 2.92 is observed. From the previous studies it's known that σ = 0.18 lbm

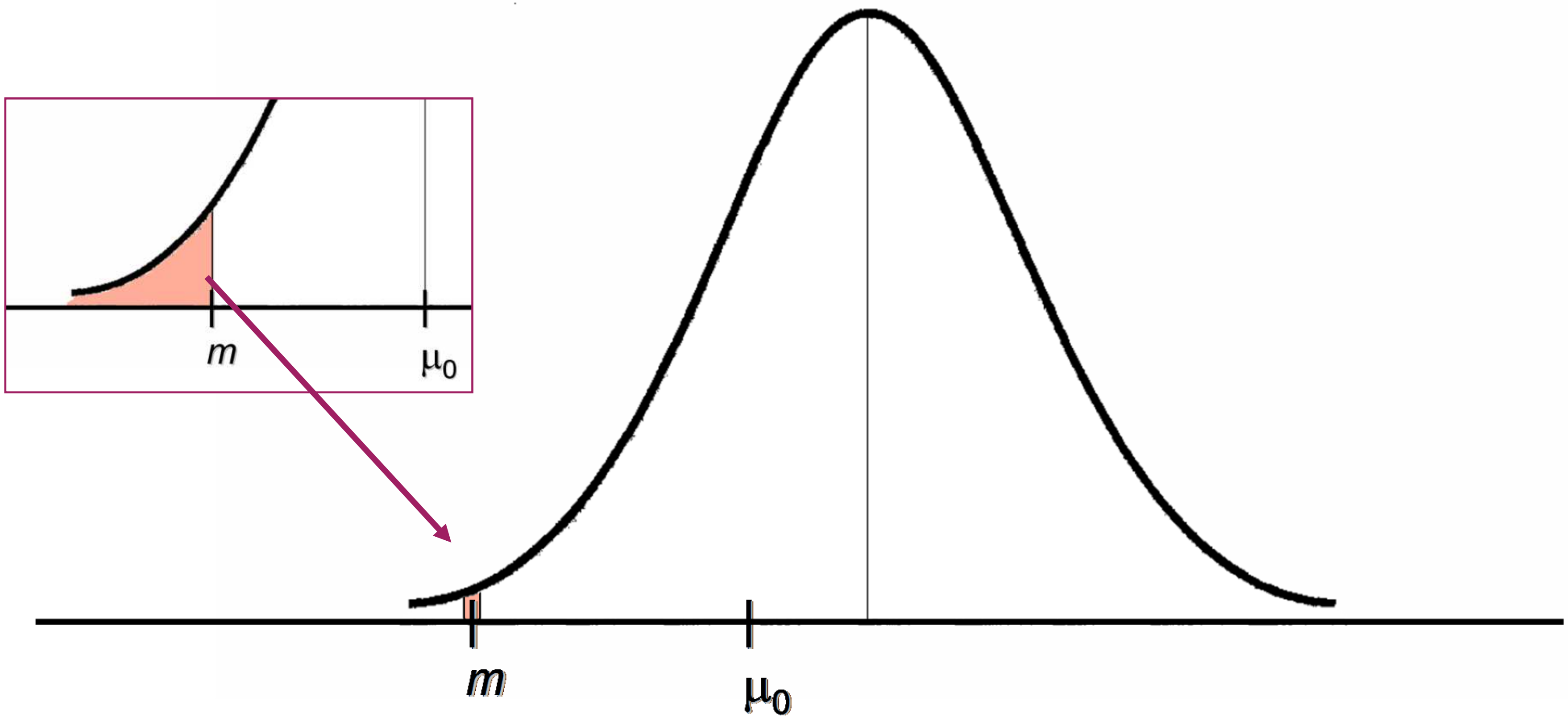
H_0: $\mu \geq 3$ no action

H_a: $\mu < 3$ legal action

Let's say: in the extreme case, when μ=3, we would like to be 99% sure that we make no mistake, when starting legal actions against Hilltop Coffee. It means that selected significance level is $\alpha = 0.01$



Let's find the probability of observation ***m*** for all possible **$\mu \geq 3$**. We start from an extreme case ($\mu=3$) and then probe all possible $\mu>3$. See the behavior of the small probability area around measured *m*. What you will get if you summarize its area for all possible $\mu \geq 3$?

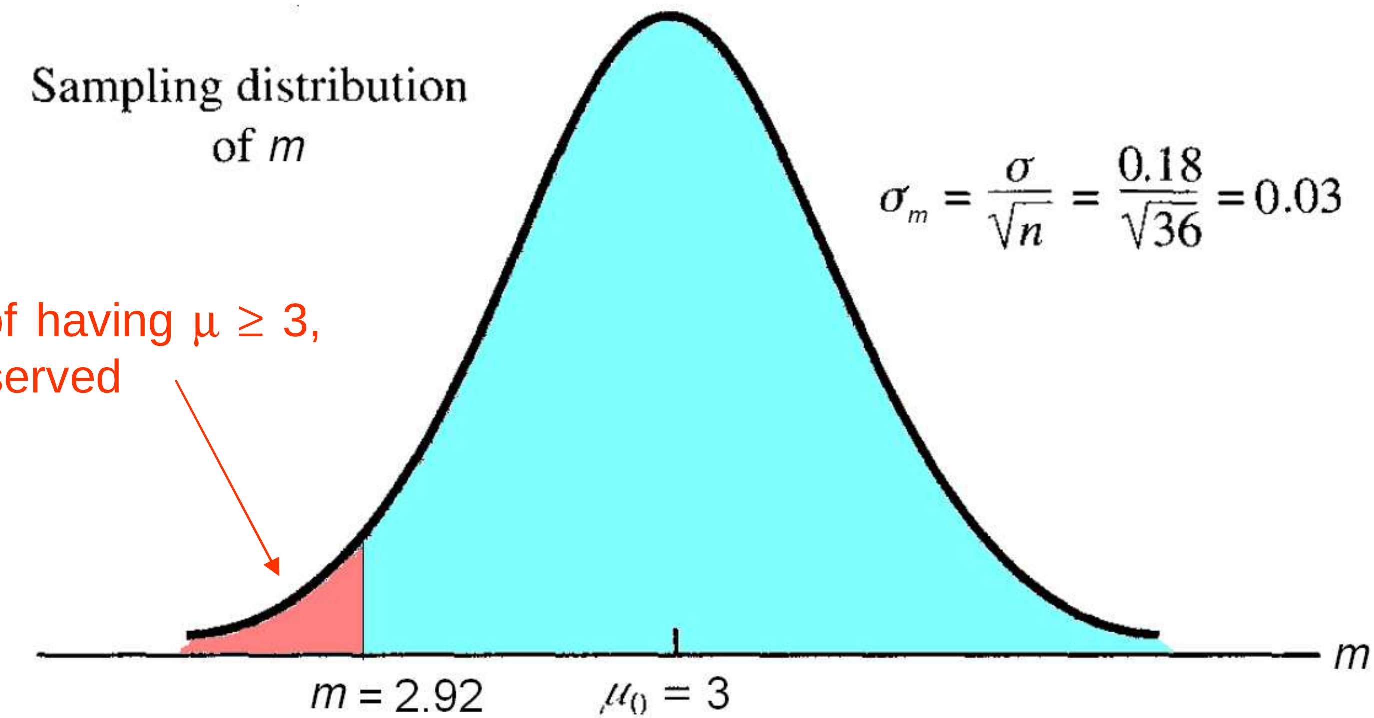


P(*m*) for all possible $\mu \geq \mu_0$ is equal to $P(\overline{x}<m)$ for an extreme case of $\mu=\mu_0$

HYPOTHESIS TESTING FOR THE MEAN

Let’s Try to Understand…

The probability of having $\mu \geq 3$,
if m = 2.92 is observed



In other words, red area characterizes the probability of the null hypothesis.

To be completely correct, the **red area** gives us a **probability of making an error** when rejecting the null hypothesis, or the **p-value.**

if σ in unknown:
$\sigma \to s$
$z \to t$

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Test Statistic	$t = \dfrac{m - \mu_0}{s/\sqrt{n}}$	$t = \dfrac{m - \mu_0}{s/\sqrt{n}}$	$t = \dfrac{m - \mu_0}{s/\sqrt{n}}$
Rejection Rule: p-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $t \leq -t_\alpha$	Reject H_0 if $t \geq t_\alpha$	Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

H_0: $\mu \geq 3$

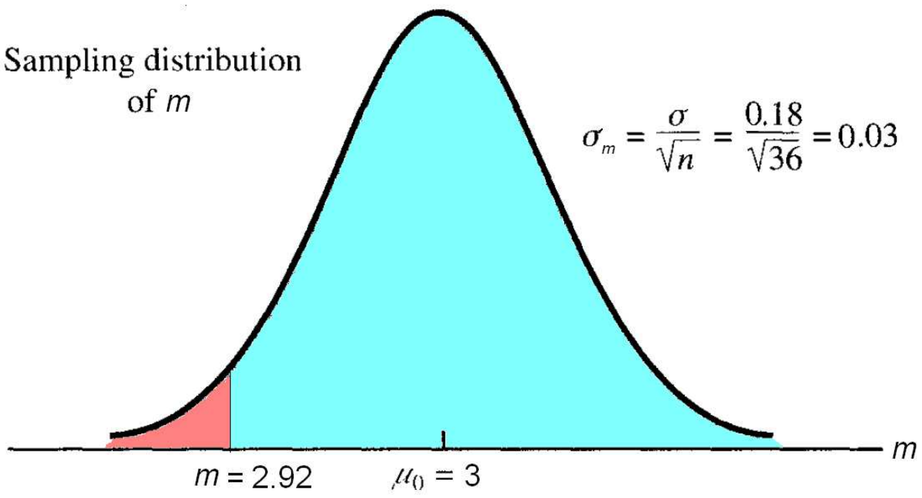
H_a: $\mu < 3$

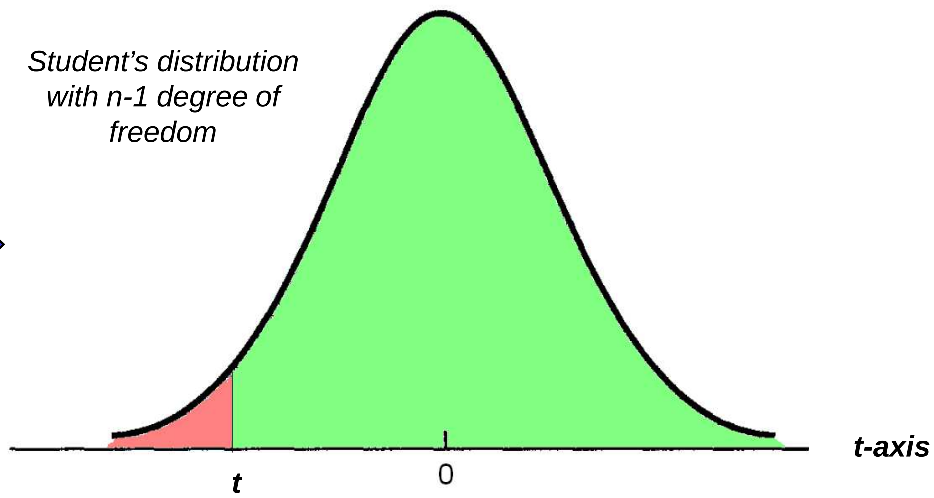
Assume that we have obtained experimentally m=2.92. Is it significant?

Step 1. Introduce the test statistics

Test statistic
A statistic whose value helps determine whether a null hypothesis can be rejected

$$t = \frac{m - \mu_0}{s/\sqrt{n}} = \frac{m - \mu_0}{s}\sqrt{n}$$

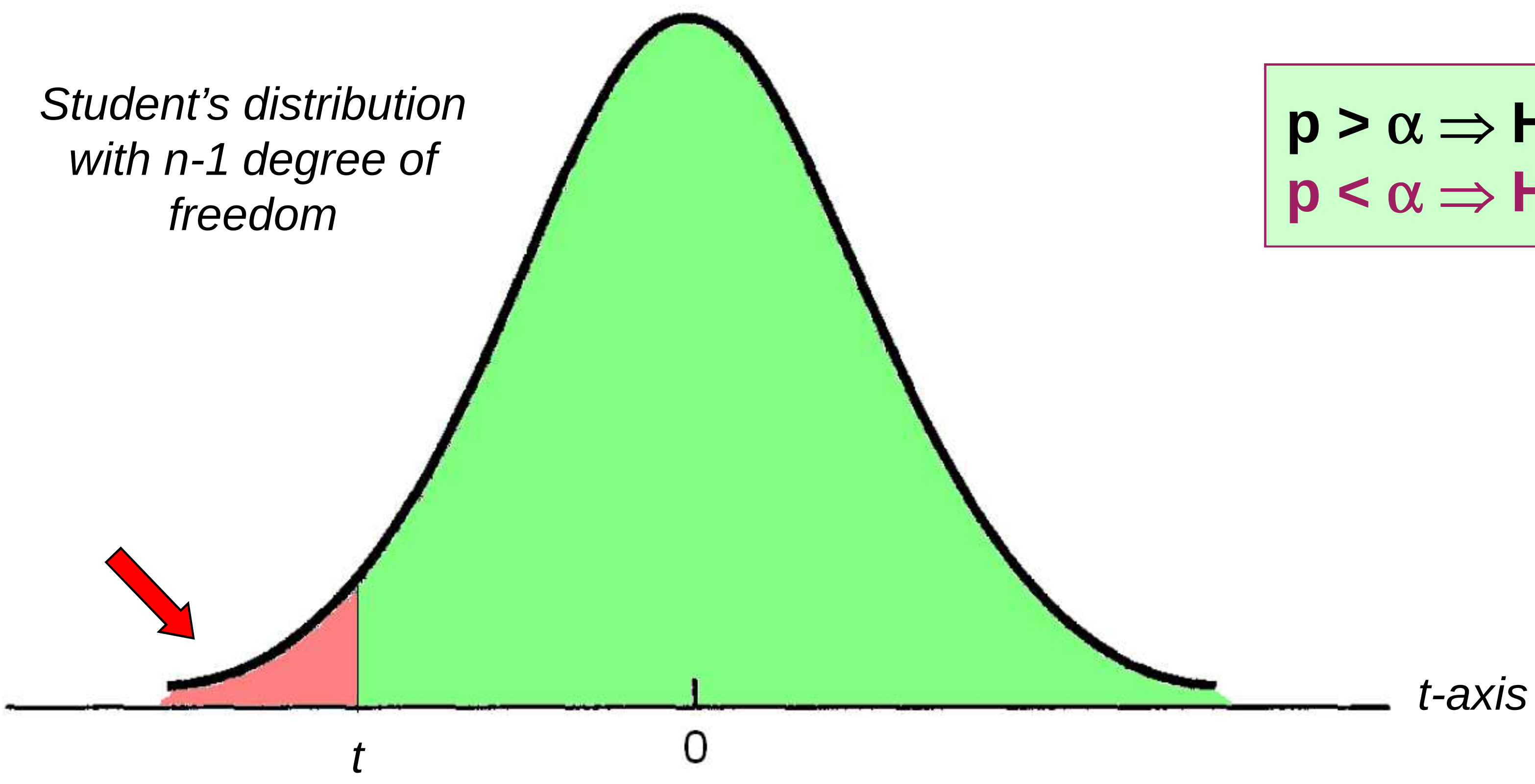




Step 2. Calculate p-value and compare it with α

p-value

A probability, computed using the test statistic, that measures the support (or lack of support) provided by the sample for the null hypothesis. It is a probability of making error of type I



$p > \alpha \Rightarrow H_0$

$p < \alpha \Rightarrow H_a$

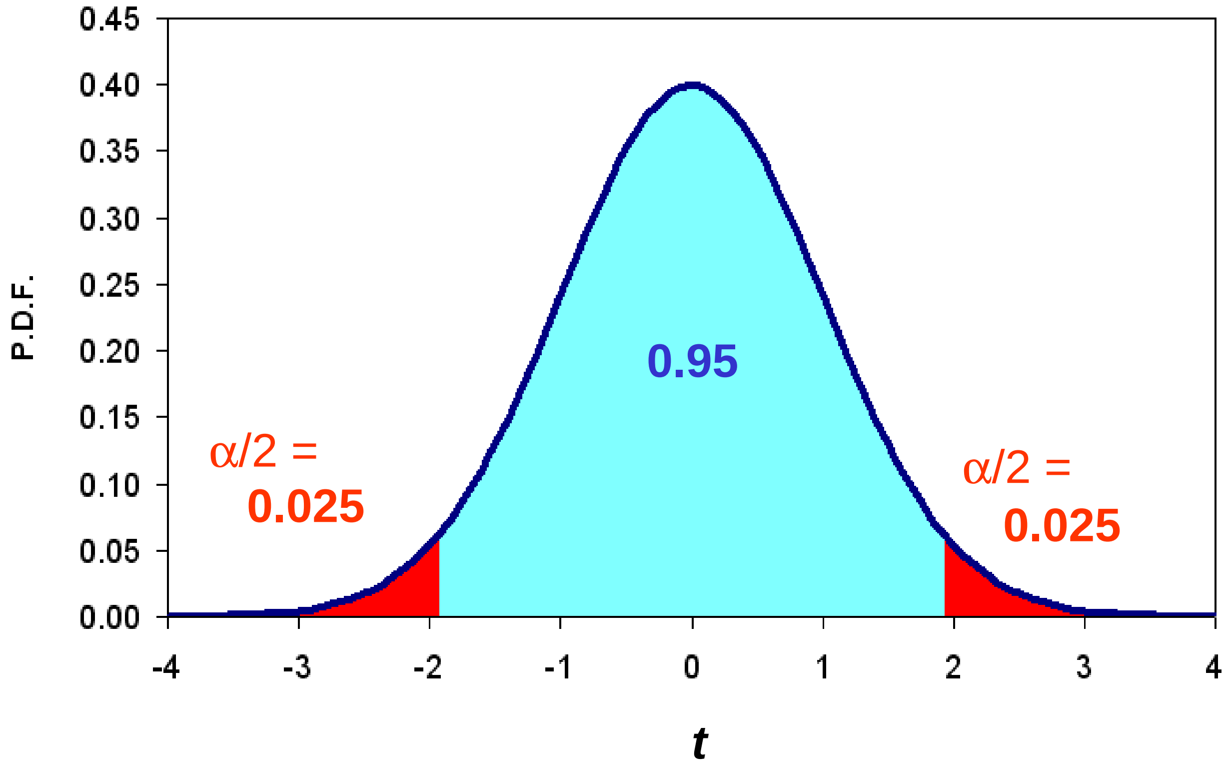
Two-tailed test

A hypothesis test in which rejection of the null hypothesis occurs for values of the test statistic in either tail of its sampling distribution.

$H_0: \mu = \mu_0$

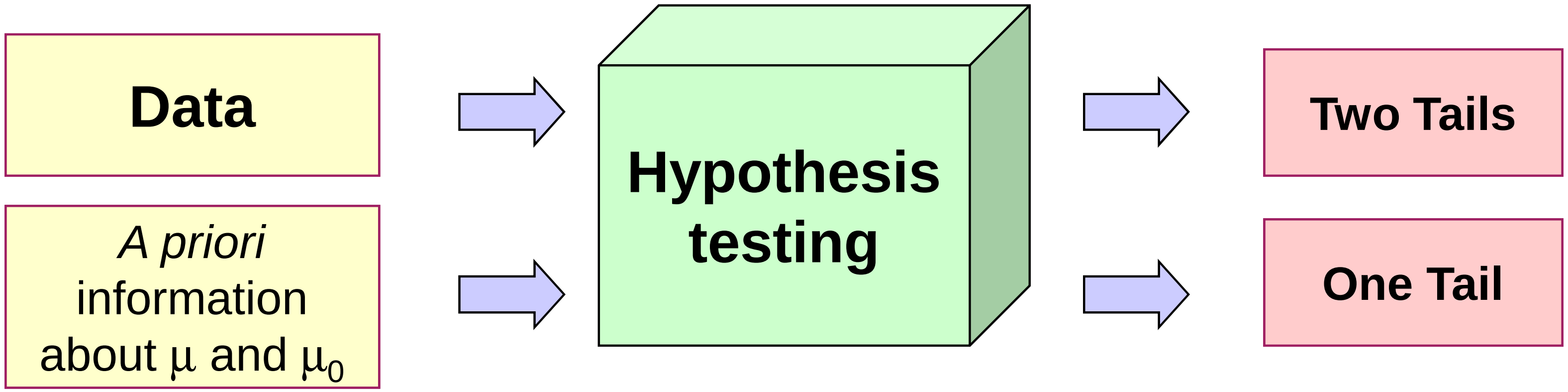
$H_a: \mu \neq \mu_0$

Student's distribution with n-1 degree of freedom



One Tail Test vs. Two Tail Test

There is a raging controversy (for about the last hundred years) on whether or not it is ever appropriate to use a one-tailed test. The rationale is that if you already know the direction of the difference, why bother doing any statistical tests. While it is **generally safest to use a two-tailed tests**, there are situations where a one-tailed test seems more appropriate. The bottom line is that **it is the choice of the researcher** whether to use one-tailed or two-tailed research questions.



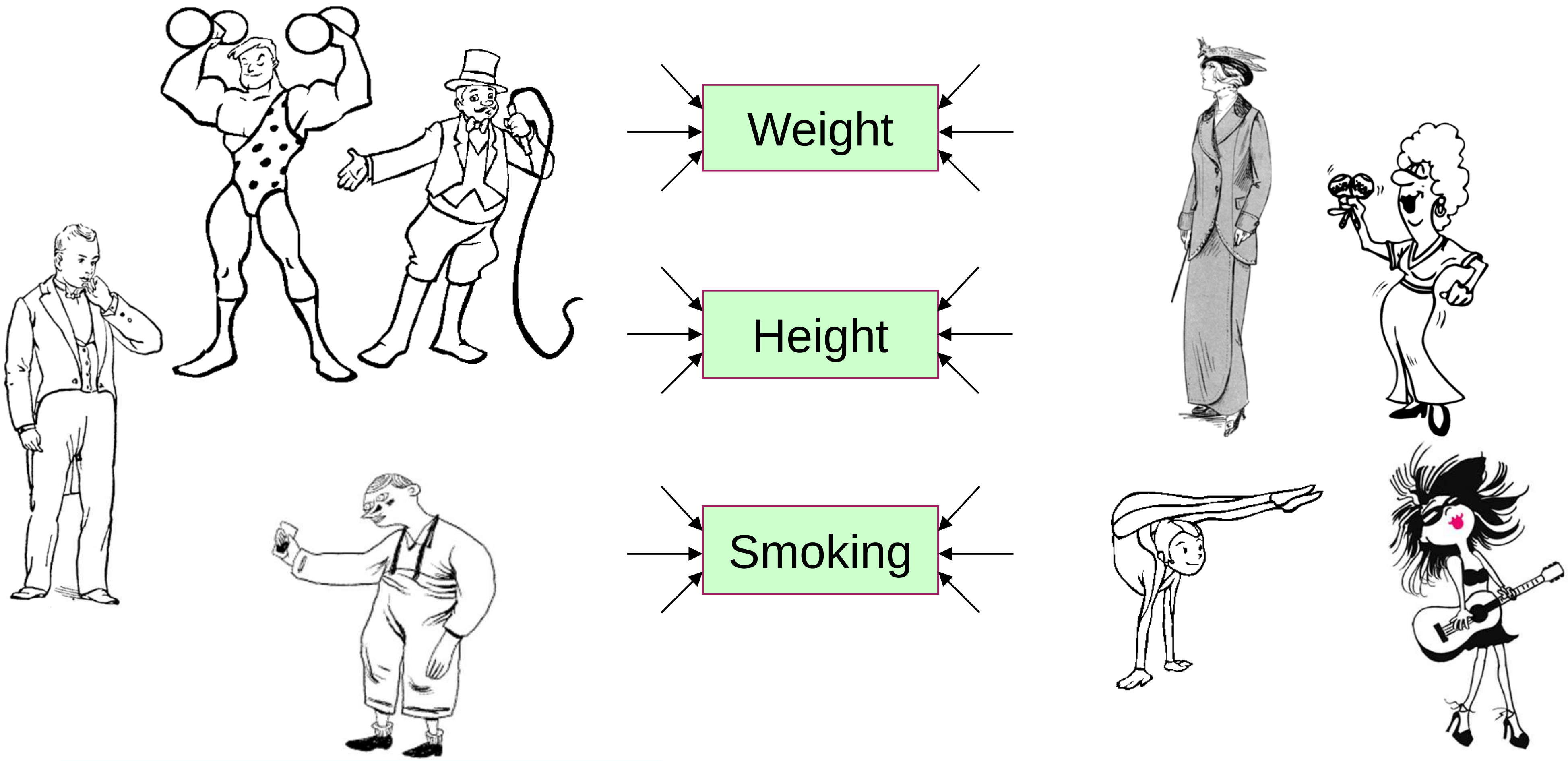
$$2 \times \text{p-value}_{(1\ \text{tail})} = \text{p-value}_{(2\ \text{tails})}$$

Unpaired t-test

UNPAIRED t-TEST

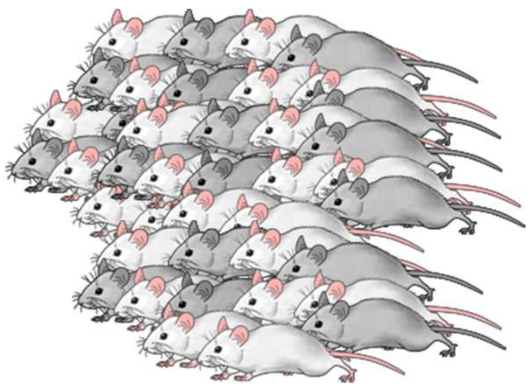
Independent Samples

Independent samples
Samples selected from two populations in such a way that the elements making up one sample are chosen independently of the elements making up the other sample.



Example

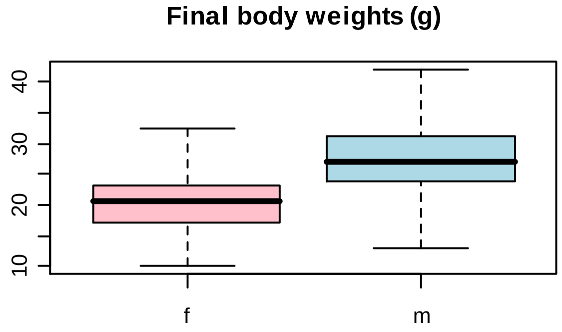
`mice.xls`

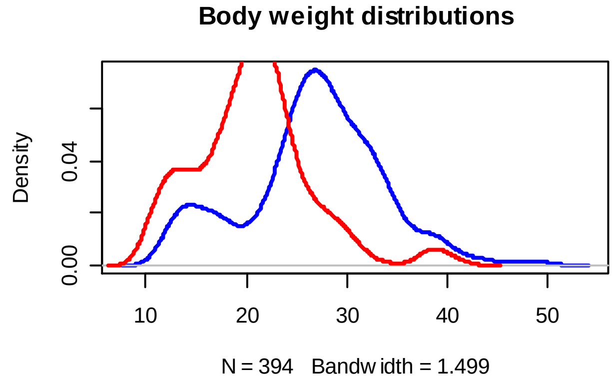


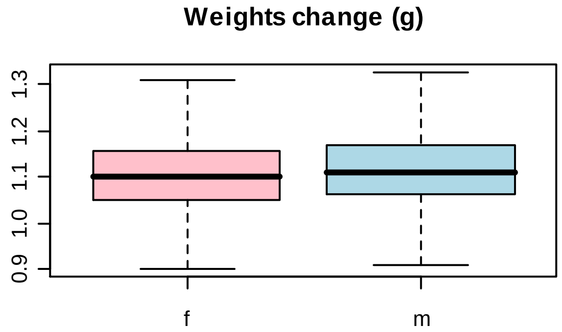
Q1: Is **body weight** for male and female significantly different?

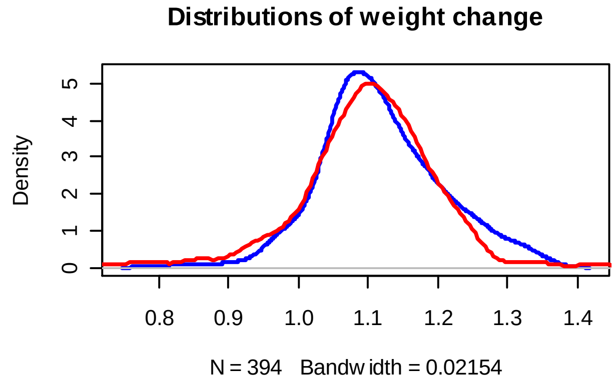
Q2: Is **weight change** for male and female significantly different?

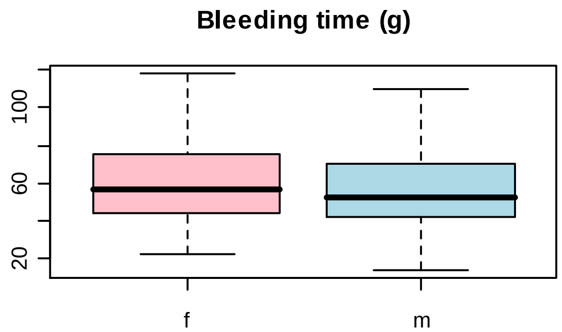
Q3: Is **bleeding time** for male and female significantly different?

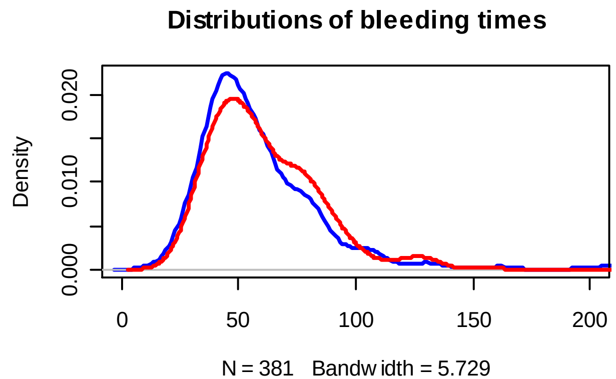








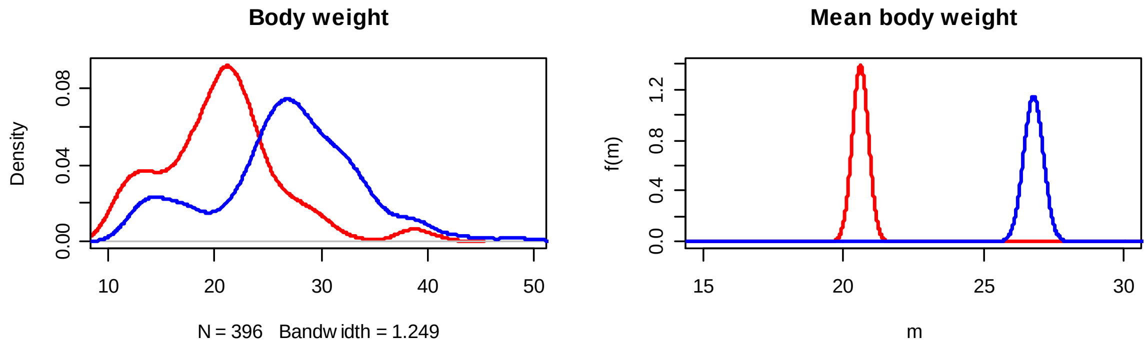




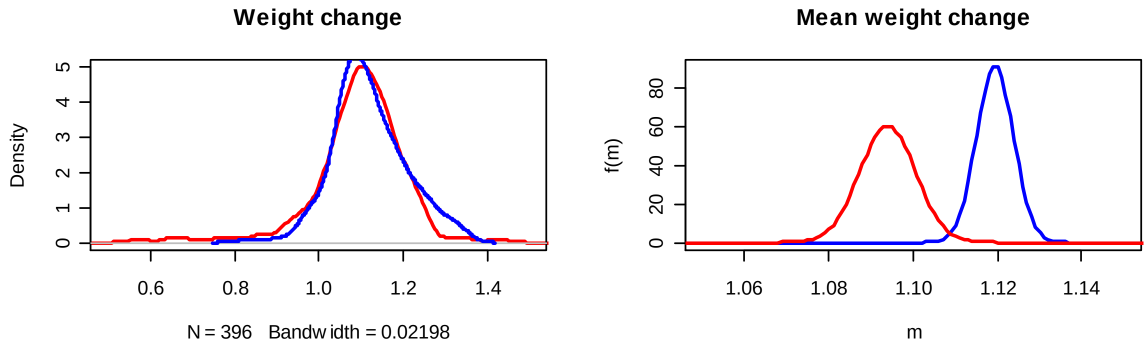
UNPAIRED t-TEST

Example

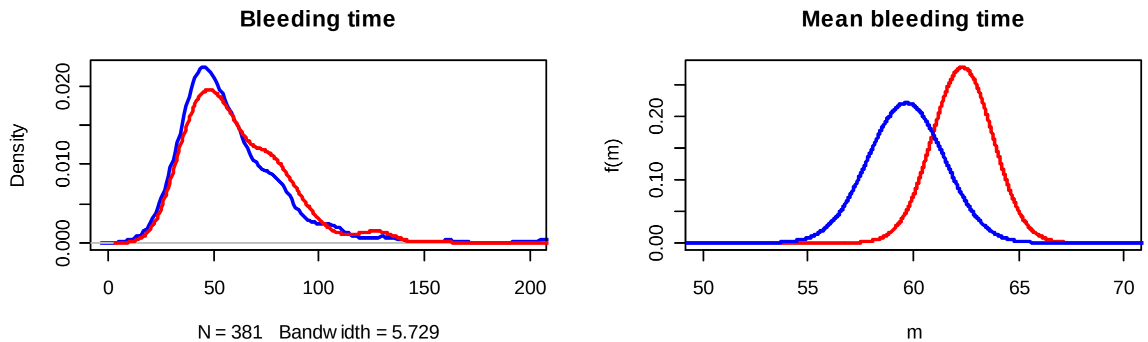
Q1: Is **body weight** for male and female significantly different?



Q2: Is **weight change** for male and female significantly different?



Q3: Is **bleeding time** for male and female significantly different?



HYPOTHESIS ABOUT MEANS OF 2 POPULATIONS

Theory

Two tail hypothesis

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

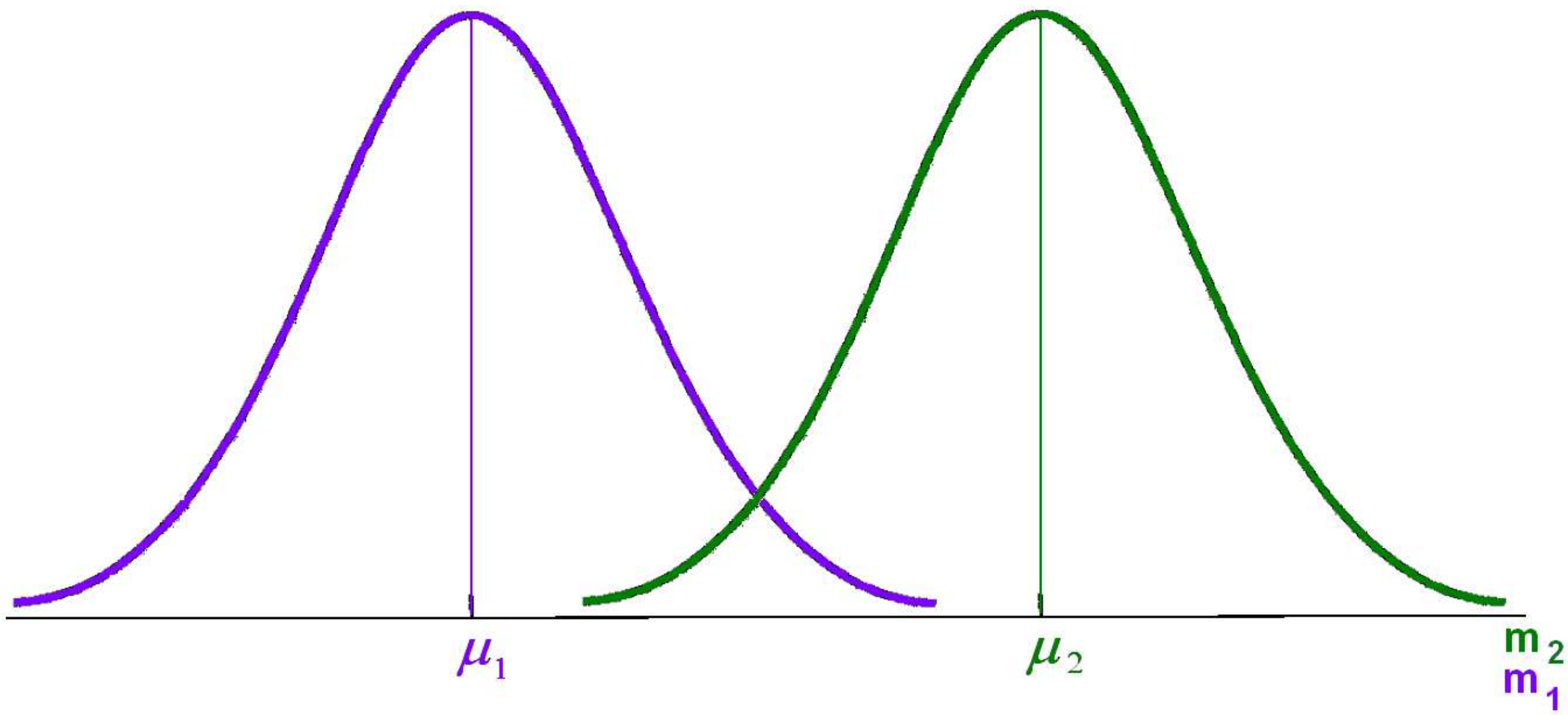
One tail hypothesis

H_0: $\mu_1 \geq \mu_2$

H_a: $\mu_1 < \mu_2$

H_0: $\mu_1 \leq \mu_2$

H_a: $\mu_1 > \mu_2$



H_0: $\mu = \mu_0$

H_a: $\mu \neq \mu_0$

As we know how to work with standard hypotheses (comparison with constant μ_0), let us transform our hypothesis:

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

$\Rightarrow$

H_0: $\mu_2 - \mu_1 = 0$

H_a: $\mu_2 - \mu_1 \neq 0$

To use it, we need to know what is the distribution of **D = $m_2 - m_1$**

Distribution of sum or difference of 2 normal random variables

The sum/difference of 2 (or more) normal random variables is a normal random variable with mean equal to sum/difference of the means and variance equal to **SUM** of the variances of the compounds.

Variables	m_1	m_2	$m_2 - m_1$
Means	μ_1	μ_2	$\mu_2 - \mu_1$
Variances	σ_1^2	σ_2^2	$\sigma_1^2 + \sigma_2^2$

H_0: $\mu_2 - \mu_1 = D_0$

H_a: $\mu_2 - \mu_1 \neq D_0$

$$D_0 = \mu_2 - \mu_1$$

$$\sigma_{m_2 - m_1} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$D_0 = m_2 - m_1$$

$$s_{m_2 - m_1} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Statistics to be used for hypothesis testing:

if σ is known: z-statistics

$$z = \frac{m_2 - m_1 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

if σ is unknown: t-statistics

$$t = \frac{m_2 - m_1 - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This is what we call t-test !!!

HYPOTHESIS ABOUT MEANS OF 2 POPULATIONS

Unpaired t-test: Algorithm

H_0: $\mu_2 - \mu_1 = D_0$

H_a: $\mu_2 - \mu_1 \neq D_0$

$$D_0 = m_2 - m_1$$

Usually $D_0 = 0$

$$s_{m_2 - m_1} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

1. Build the statistics to be used for hypothesis testing:

$$t = \frac{m_2 - m_1 - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t-distribution has following degrees of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$df = (n-1)\frac{\left(s_1^2 + s_2^2\right)^2}{\left(s_1^4 + s_2^4\right)}$$

$(n_1+n_2)/2 < \mathbf{df} < n_1+n_2$

2. Calculate the p-value

- `= T.DIST(ABS(t),df,2)`

☺. Or simply do:

- `= T.TEST (array1, array2, 2, 3)`
- `= TTEST (…)` *(Office 2007,2003)*

mice.xls

Using the t-test define which parameter in the table is sex-dependent

- **= T.TEST (array1, array2, 2, 3)**

parameter	pval	female	male
Starting age	0.165799	65.90	66.52
Ending age	0.223033	113.91	114.61
Starting weight	5.48E-34	18.91	23.86
Ending weight	8.98E-38	20.62	26.78
Weight change	0.001405	1.09	1.12
Bleeding time	0.248716	62.34	59.67
Ionized Ca in blood	0.271336	1.23	1.24
Blood pH	0.009593	7.21	7.19
Bone mineral density	2.41E-05	0.05	0.05
Lean tissues weight	4.66E-33	15.32	19.21
Fat weight	2.28E-21	4.85	7.30

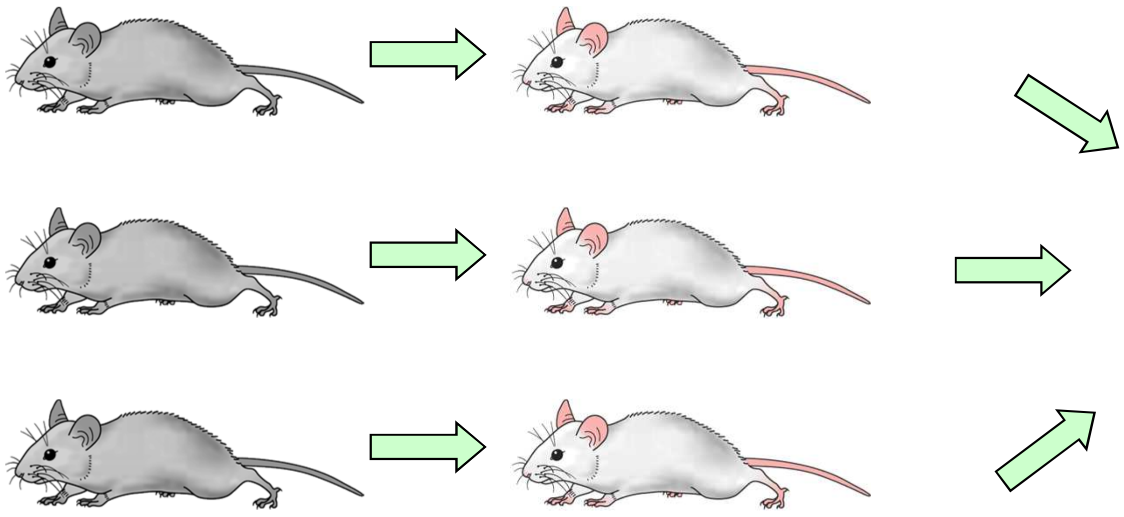
Paired t-test

HYPOTHESIS ABOUT MEANS OF 2 POPULATIONS

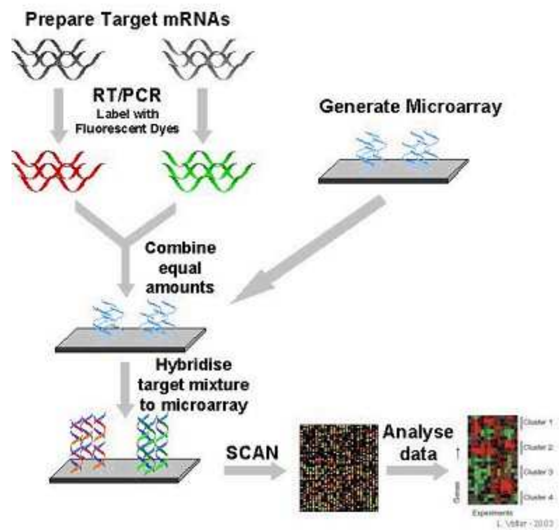
Dependent Samples

Matched samples
Samples in which each data value of one sample is matched with a corresponding data value of the other sample.

Before treatment | After treatment



Transcriptomic Analysis



Paired t-test: Task

bloodpressure.xls

The systolic blood pressures of n=12 women between the ages of 20 and 35 were measured before and after usage of a newly developed oral contraceptive.

Systolic blood pressure (mmHg)

Subject	BP before	BP after
1	122	127
2	126	128
3	132	140
4	120	119
5	142	145
6	130	130
7	142	148
8	137	135
9	128	129
10	132	137
11	128	128
12	129	133

Q: Does the treatment affect the systolic blood pressure?

Unpaired test

- `= T.TEST (array1, array2, 2, 3)`

Paired test

- `= T.TEST (array1, array2, 2, 1)`

Test	p-value
unpaired	0.414662
paired	0.014506

Hypotheses for 2 Proportions

HYPOTHESIS ABOUT PROPORTIONS OF 2 POPULATIONS

Theory

H_0: $\pi_1 = \pi_2$

H_a: $\pi_1 \neq \pi_2$

H_0: $\pi_1 - \pi_2 = 0$

H_a: $\pi_1 - \pi_2 \neq 0$

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Pooled estimator of π

An estimator of a population proportion obtained by computing a weighted average of the point estimators obtained from two independent samples.

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$\sigma_{p_1 - p_2} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

= 2*(1-NORM.S.DIST(ABS(z),TRUE))

HYPOTHESIS ABOUT PROPORTIONS OF 2 POPULATIONS

Example

SWR/J	MA/MyJ
f	f
f	f
f	f
f	f
f	f
f	f
f	f
f	f
f	m
f	m
m	m
m	m
m	m
m	m
m	m
m	m
m	m
m	m
m	m
	m
	m
	m
	m

mice.xls

Q: Is the male proportion significantly different in these mouse strains (0.47 and 0.65)?

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

```
= 2*(1-NORM.S.DIST(ABS(z),TRUE))
```

	SWR/J	MA/MyJ	*pooled*
count male	9	15	*24*
n	19	23	*42*
p	0.474	0.652	*0.571*
z	-1.16		
p-val	**0.244658997**		

p-value = 0.24

Thank you for your attention

to be continued…