



LUXEMBOURG
INSTITUTE
OF HEALTH

Multomics Data Science Group (MODAS)
Department of Cancer Research, LIH

Bioinformatics Platform (BIOINFO)
Department of Medical Informatics, LIH

BIOSTATISTICS for PhDs

Lecture 2

Testing Hypotheses

Peter Nazarov

19-02-2024

Email: petr.nazarov@lih.lu
Skype: pvn.public

<http://edu.modas.lu>

COURSE OVERVIEW

Outline *(to be updated during the course)*

◆ Lecture 1, 2024-02-05

- ◆ numerical measures (location/variability/association), parametric/nonparametric
- ◆ basic summary and visualization in R: barplot, boxplot, scatter plot
- ◆ z-score, detection of outliers
- ◆ continuous distributions (normal, Student, χ^2 , F), linkage to probability
- ◆ sampling distribution, methods for sampling



<https://cran.r-project.org/>



<https://posit.co/downloads/>

◆ Lecture 2, 2024-02-19

- ◆ interval estimations for mean and proportion
- ◆ hypotheses testing for mean(s), p-value, tails
- ◆ number of samples
- ◆ power of a test
- ◆ non-parametric tests
- ◆ multiple comparisons

◆ Lecture 3, 2024-03-04

- ◆ interval estimations and hypotheses for variance
- ◆ model fitting and test for independence
- ◆ linear models, ANOVA, posthoc analysis
- ◆ simple and multiple linear regression
- ◆ factors in linear regression
- ◆ logistic regression

◆ Lecture 4, 2024-03-18 *(please, propose!)*

- ◆ omics data analysis?
- ◆ survival analysis?
- ◆ clustering?
- ◆ more practical exercise?

Let's work at a comfortable speed!

Materials and other courses:

<http://edu.modas.lu>

INTERVAL ESTIMATES: Means and Proportions

Interval estimates, confidence intervals for means

Confidence intervals for proportions

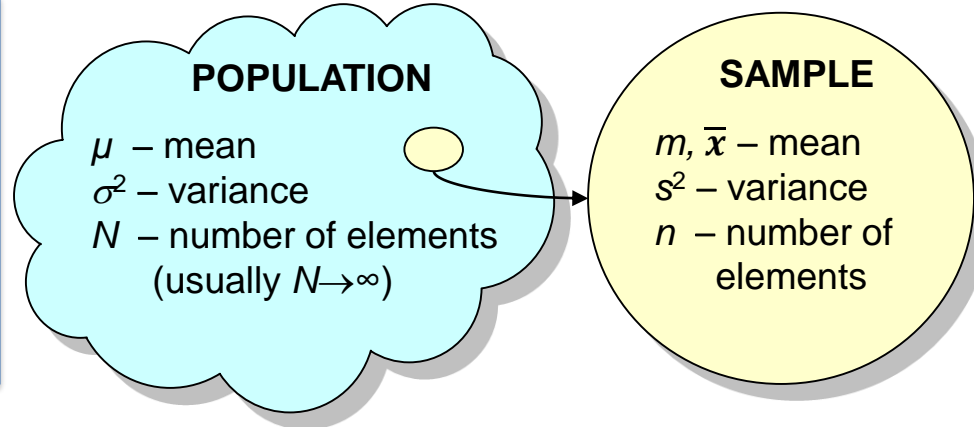
Interval estimation in the case of random functions

NUMERICAL MEASURES

Population and Sample

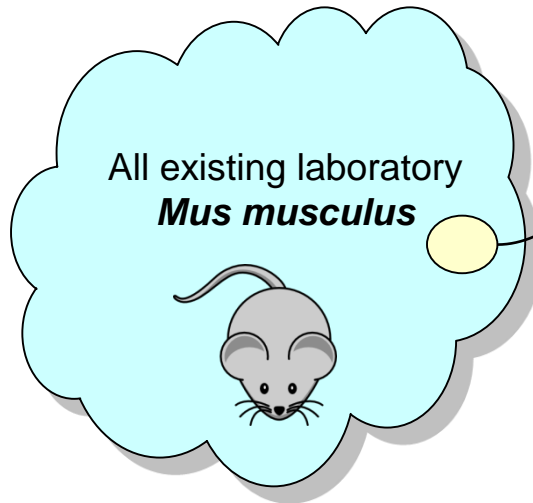
Population parameter

A numerical value used as a summary measure for a population of size N (e.g., the population mean μ , variance σ^2 , standard deviation σ)



Sample statistic (parameter)

A numerical value used as a summary measure for a sample of size n (e.g., the sample mean m , the sample variance s^2 , and the sample standard deviation s)



mice

790 mice from different strains

<http://phenome.jax.org>

ID	Strain	Sex	Starting age	Ending age	Starting weight	Ending weight	Weight change	Bleeding time	Ionized Ca in blood	Blood pH	Bone mineral density	Lean tissues weight	Fat weight
1	129S1/SvlmJ	f	66	116	19.3	20.5	1.062	64	1.2	7.24	0.0605	14.5	4.4
2	129S1/SvlmJ	f	66	116	19.1	20.8	1.089	78	1.15	7.27	0.0553	13.9	4.4
3	129S1/SvlmJ	f	66	108	17.9	19.8	1.106	90	1.16	7.26	0.0546	13.8	2.9
368	129S1/SvlmJ	f	72	114	18.3	21	1.148	65	1.26	7.22	0.0599	15.4	4.2
369	129S1/SvlmJ	f	72	115	20.2	21.9	1.084	55	1.23	7.3	0.0623	15.6	4.3
370	129S1/SvlmJ	f	72	116	18.8	22.1	1.176		1.21	7.28	0.0626	16.4	4.3
371	129S1/SvlmJ	f	72	119	19.4	21.3	1.098	49	1.24	7.24	0.0632	16.6	5.4
372	129S1/SvlmJ	f	72	122	18.3	20.1	1.098	73	1.17	7.19	0.0592	16	4.1
4	129S1/SvlmJ	f	66	109	17.2	18.9	1.099	41	1.25	7.29	0.0513	14	3.2
5	129S1/SvlmJ	f	66	112	19.7	21.3	1.081	129	1.14	7.22	0.0501	16.3	5.2
10	129S1/SvlmJ	m	66	112	24.3	24.7	1.016	119	1.13	7.24	0.0533	17.6	6.8
364	129S1/SvlmJ	m	72	114	25.3	27.2	1.075	64	1.25	7.27	0.0596	19.3	5.8
365	129S1/SvlmJ	m	72	115	21.4	23.9	1.117	48	1.25	7.28	0.0563	17.4	5.7
366	129S1/SvlmJ	m	72	118	24.5	26.3	1.073	59	1.25	7.26	0.0609	17.8	7.1
367	129S1/SvlmJ	m	72	122	24	26	1.083	69	1.29	7.26	0.0584	19.2	4.6
6	129S1/SvlmJ	m	66	116	21.6	23.3	1.079	78	1.15	7.27	0.0497	17.2	5.7
7	129S1/SvlmJ	m	66	107	22.7	26.5	1.167	90	1.18	7.28	0.0493	18.7	7
8	129S1/SvlmJ	m	66	108	25.4	27.4	1.079	35	1.24	7.26	0.0538	18.9	7.1
9	129S1/SvlmJ	m	66	109	24.4	27.5	1.127	43	1.29	7.29	0.0539	19.5	7.1

Load the data:

```
Mice = read.table("http://edu.modas.lu/data/txt/mice.txt", sep="\t", header=TRUE, stringsAsFactors = TRUE)
```

INTERVAL ESTIMATION

Definitions

Interval estimate

An estimate of a population parameter that provides an interval believed to contain the value of the parameter. The interval estimates have the form: point estimate \pm margin of error.

Margin of error

The \pm value added to and subtracted from a point estimate in order to develop an interval estimate of a population parameter.

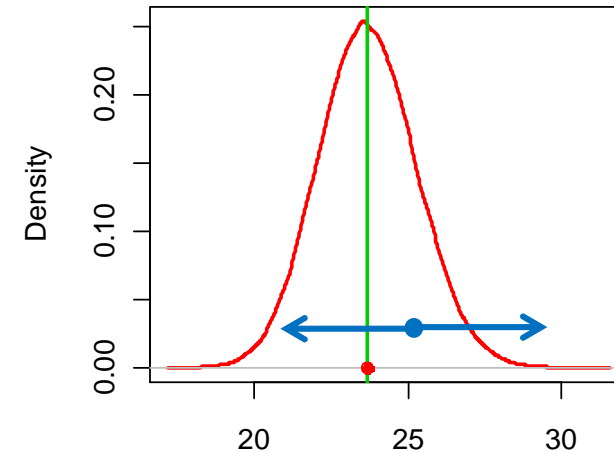
σ known

The condition existing when historical data or other information provides a good value for the population standard deviation prior to taking a sample. The interval estimation procedure uses this known value of σ in computing the margin of error.

σ unknown (usual situation)

The condition existing when no good basis exists for estimating the population standard deviation prior to taking the sample. The interval estimation procedure uses the sample standard deviation s in computing the margin of error.

Distribution of m

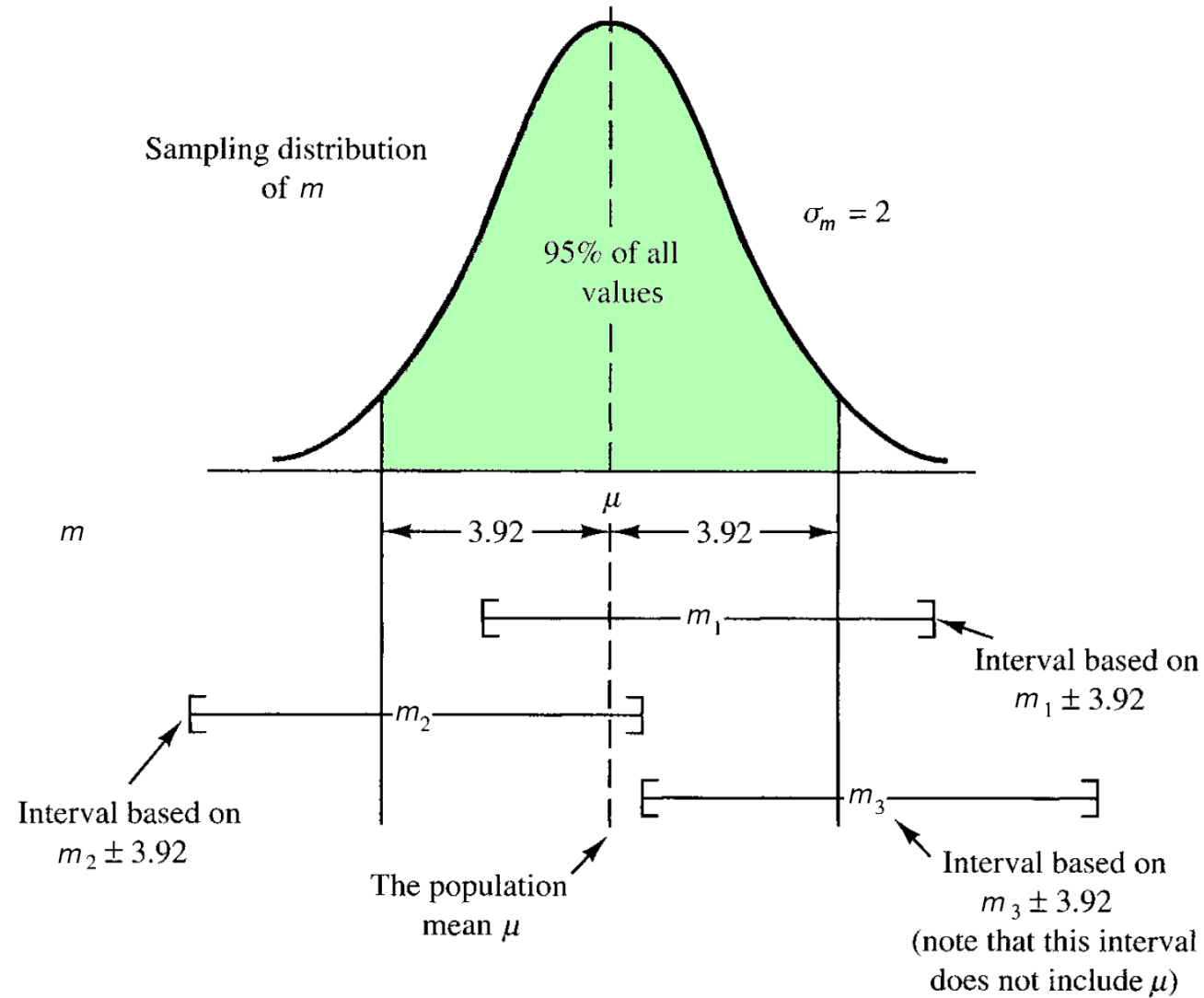


$$\mu = m \pm \text{margin of error}$$

We will consider theory for σ -known cases and then generalize to σ -unknown

INTERVAL ESTIMATION: σ KNOWN

Interval Estimation for the Mean



INTERVAL ESTIMATION: σ KNOWN

Confidence Interval

Confidence level

The confidence that is associated with an interval estimate. For example, if an interval estimation procedure provides intervals such that 95% of the intervals formed using the procedure will include the population parameter, the interval estimate is said to be constructed at the 95% confidence level. **Confidence = 1 - α**

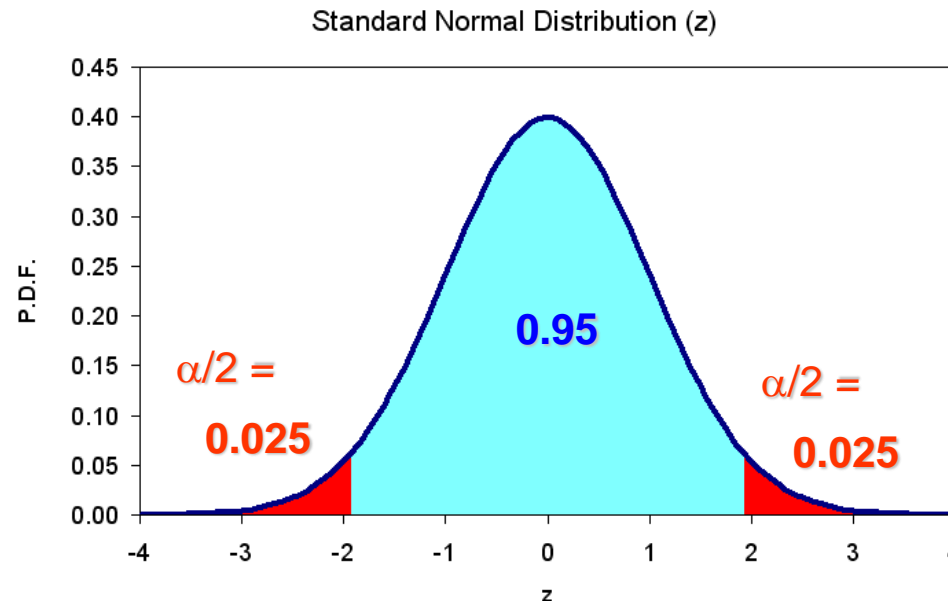
alpha = 1 - confidence

Confidence interval

Another name for an interval estimate.

$$\mu = m \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

For 95 % confidence $\alpha = 0.05$, which means that in each tail we have 0.025. Corresponding $z_{\alpha/2} = 1.96$



```
m = ... #mean
s = ... #st.dev
n = ... #number of observations
a = ... #alpha

me = -qnorm(0.025)*s/sqrt(n)

sprintf("mu= %g +/- %g",m,me)
```

◆ = CONFIDENCE.NORM(alpha, σ , n)

◆ = -NORM.S.INV(alpha/2)* σ /SQRT(n)

INTERVAL ESTIMATION: σ KNOWN

Example: Interval Estimation for the Mean

An engineer is testing a new measuring device. He tries to put 500 μl of water into tubes and then measure the resulting quantity. Based on 36 measurements she estimated the average volume of 498 μl . From technical documentation for the device she learnt that the standard deviation of the volume is around 5 μl . Calculate the 95% and 99% confidence intervals for the volume the researcher takes on average. Is the desired volume of 500 μl in the confidence intervals?

$$\mu = m \pm \text{merror} = m \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Marginal error (merror)

=CONFIDENCE.NORM(0.05,5,36)

=CONFIDENCE.NORM(0.01,5,36)

95% CI: $\mu = 498 \pm 1.63 = [496.4 \dots 499.6]$

99% CI: $\mu = 498 \pm 2.14 = [495.8 \dots 500.1]$

```

m = 498
s = 5
n = 36
a = 0.05
me = -qnorm(a/2)*s/sqrt(n)

sprintf("mu = %g +/- %g", m, me)
  
```


INTERVAL ESTIMATION: σ UNKNOWN

Population Mean: σ Unknown

Assume that we have a sample of 20 mice and would like to estimate an average size of a mice in population.

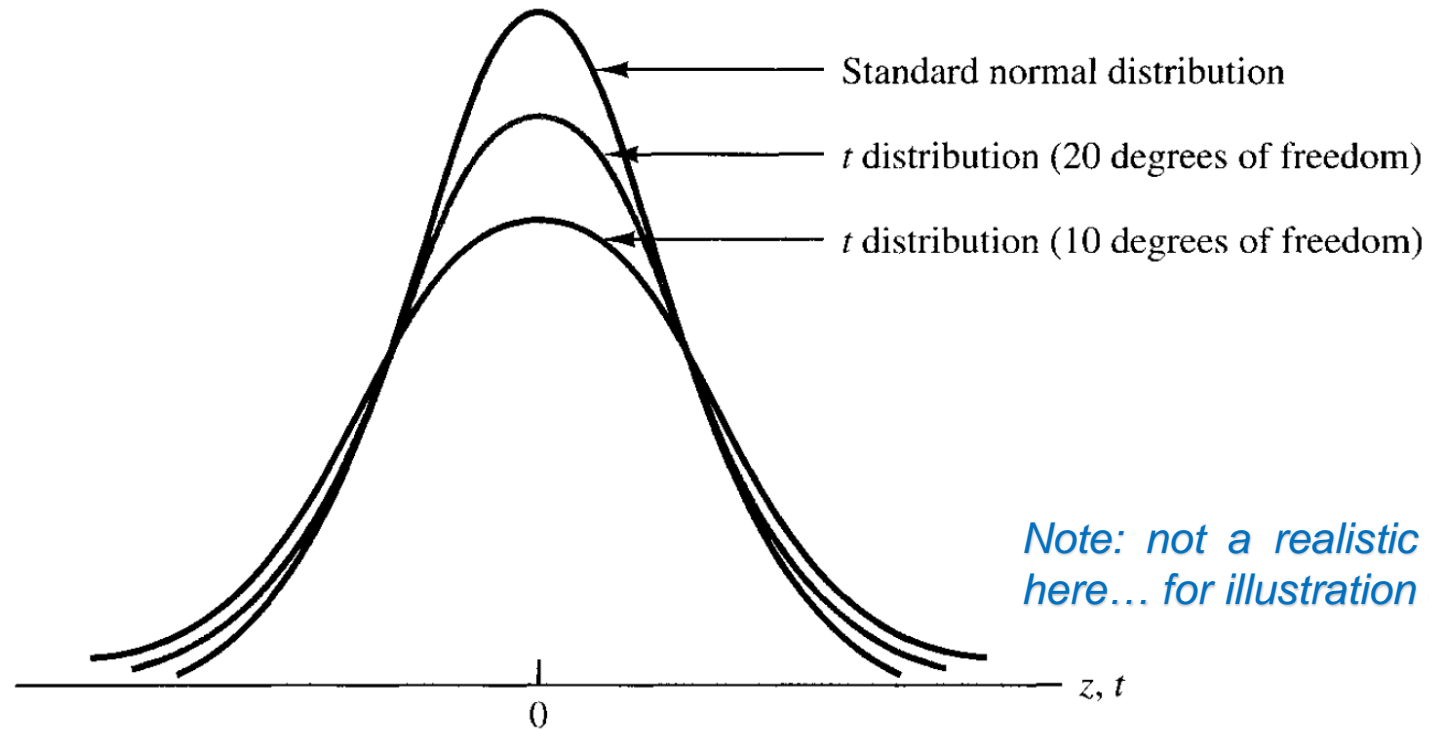
Weight
39.9
19.8
32.4
21
27.5
20.8
21.3
40
10.7
22.6
27
10.8
20.9
14.7
31.4
17.2
11.4
19.1
31.3
14.8

$$m = 22.73$$

$$s = 8.84$$

$$\sigma_m = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

As we replace $\sigma \rightarrow s$, we introduce an additional error and this change the distribution from z to t (Student)



Note: not a realistic scale here... for illustration only

INTERVAL ESTIMATION: σ UNKNOWN

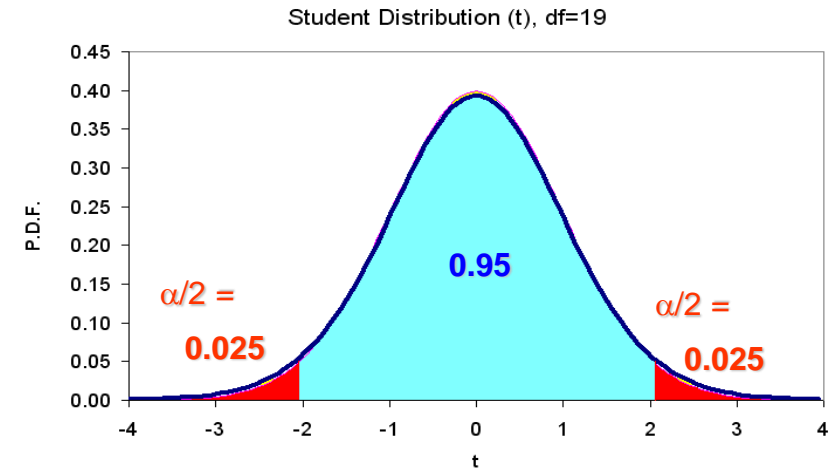
Interval Estimation for the Mean in Case of σ Unknown

Weight
39.9
19.8
32.4
21
27.5
20.8
21.3
40
10.7
22.6
27
10.8
20.9
14.7
31.4
17.2
11.4
19.1
31.3
14.8

$m = 22.73$
 $s = 8.84$

$s(m) = 1.98$
 $t = 2.09$
 $m.e. = 4.14$

$$\mu = m \pm t_{\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}}$$



```
m = ... #mean
s = ... #st.dev
n = ... #number of observations
a = ... #alpha

me = -qt(0.025,n-1)*s/sqrt(n)

sprintf("mu= %g +/- %g",m,me)
```

Variant 1 :

```
◆ =CONFIDENCE.T(alpha,s,n)
```

Variant 2 :

```
◆ =-T.INV(alpha/2,n-1)*s/SQRT(n)
```

Or simply:

```
t.test(x) # show
t.test(x)$conf.int # get
```

INTERVAL ESTIMATION

Population Proportion

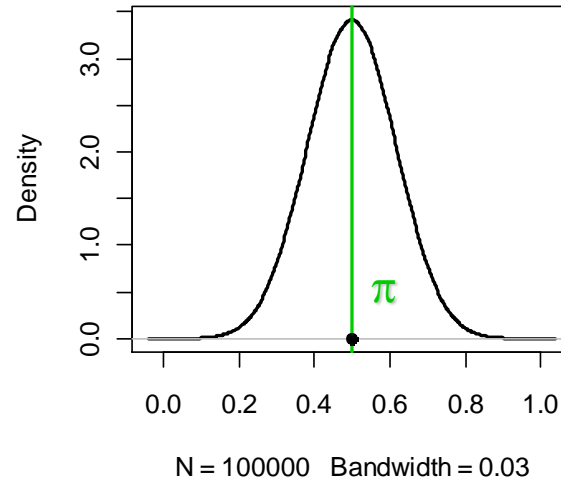
Standard error of proportion

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \rightarrow \sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

$$\pi = p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

if $np \geq 5$ and $n(1-p) \geq 5$

Sampling distribution
for proportion p



```
p = ... #proportion
n = ... #number of observations
a = ... #alpha

sp = sqrt(p*(1-p)/n)
me = -qnorm(a/2)*sp

sprintf("pi= %g +/- %g", p, me)
```

```
# simple way 1: p is known
prop.test(p*n, n)

# simple way 2:
# x - logical vector
prop.test(sum(x), length(x))
```

```
# exact
binom.test(p*n, n)
```

Variant 1:

- calculate n and p
- `=CONFIDENCE.NORM(alpha, SQRT(p*(1-p)), n)`

Variant 2:

- calculate n and p
- calculate st.error $\sigma_p = \text{SQRT}(p*(1-p)/n)$
- calculate $z_{\alpha/2}$ statistics `=-NORM.S.INV(alpha/2)`
- $= z_{\alpha/2} * \sigma_p$

$\alpha = 1 - \text{confidence}$

for 95% confidence $z_{0.025} = 1.96$

INTERVAL ESTIMATION

Population Proportion: Some Practical Aspects

$$\pi = p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

1. The normal distribution is applicable only when enough data points are observed. The rule of thumb is: $np \geq 5$ and $n(1-p) \geq 5$

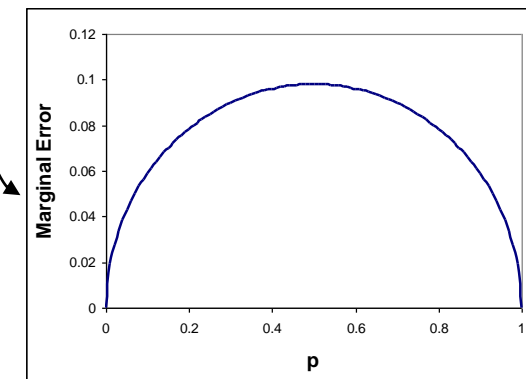
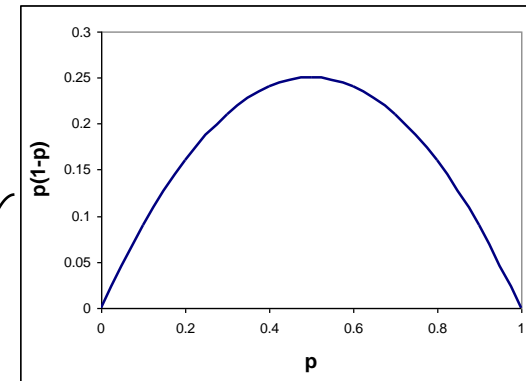
2. The maximal marginal error is observed when $p=0.5$

3. The estimation of the sample size can be obtained:

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{E^2}$$

$np \geq 5$ and $n(1-p) \geq 5$

where p is a best guess for π or the result of a preliminary study



INTERVAL ESTIMATION

Example: Population Proportion

pancreatitis

```
Pan =
read.table("http://edu.modas.lu/data/txt/
pancreatitis.txt", sep="\t", header=TRUE,
stringsAsFactors = TRUE)
```

Think whether you would like to use pooled groups (other, pancreatitis) or make independent analysis for each? Why?

n=	217
n(Never)=	56
p=	0.258065
st.error (st.dev.prop)=	0.029704
cut off for 95%, z(a/2)=	1.959964
Margin of Error=	0.058219

$$\pi = p \pm \text{error} = 25.8 \pm 5.8 \%$$

1. Define a 95% confidence interval for **never-smoking** proportion of people coming to a hospital.
2. How many patients you would need to have error (E) less than 1%

1

$$\pi = p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

2

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{E^2}$$

for 95% confidence $z_{0.025} = 1.96$

```
n = sum(Pan$Disease == "other")
x = sum(Pan$Smoking == "Never" &
Pan$Disease == "other")
p = x/n
a = 0.05

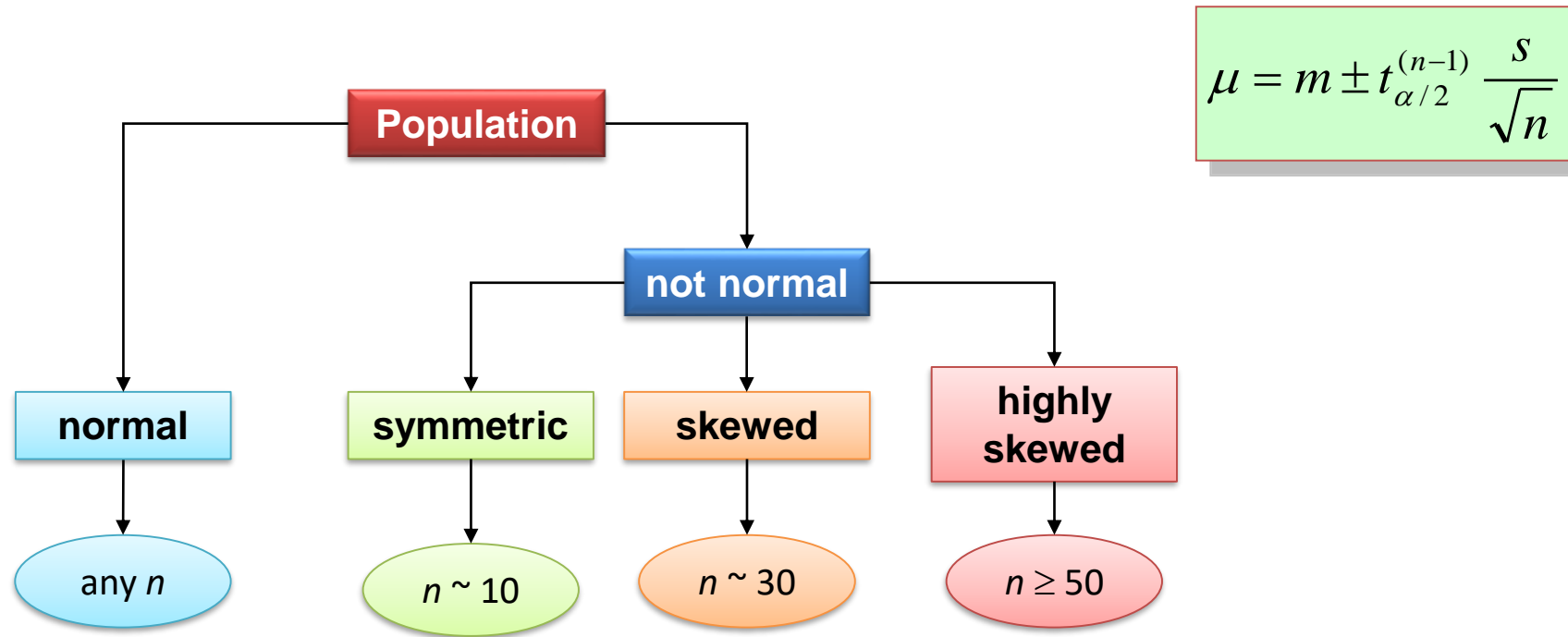
# method 1: manual calculation
sp = sqrt(p*(1-p)/n)
me = -qnorm(a/2)*sp
sprintf("pi= %g +/- %g",p,me)

# method 2: prop.test
prop.test(x,n,conf.level=1-a)$conf.int
```

```
E = 0.01
n1 = qnorm(a/2)^2 * p * (1-p) / E^2
n1
```

INTERVAL ESTIMATES

Population Mean: Practical Advice



$$\mu = m \pm t_{\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}}$$

Advice 2

if $n > 100$ you can, in principle, use z-statistics instead of t-statistics (error will be $< 1.5\%$)

INTERVAL ESTIMATES

Determining the Sample Size

Let's focus on another aspect: how to select a proper number of experiments.

$$\mu = m \pm E(n, \sigma)$$

$$E(n, \sigma) = E$$

$$n - ?$$

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

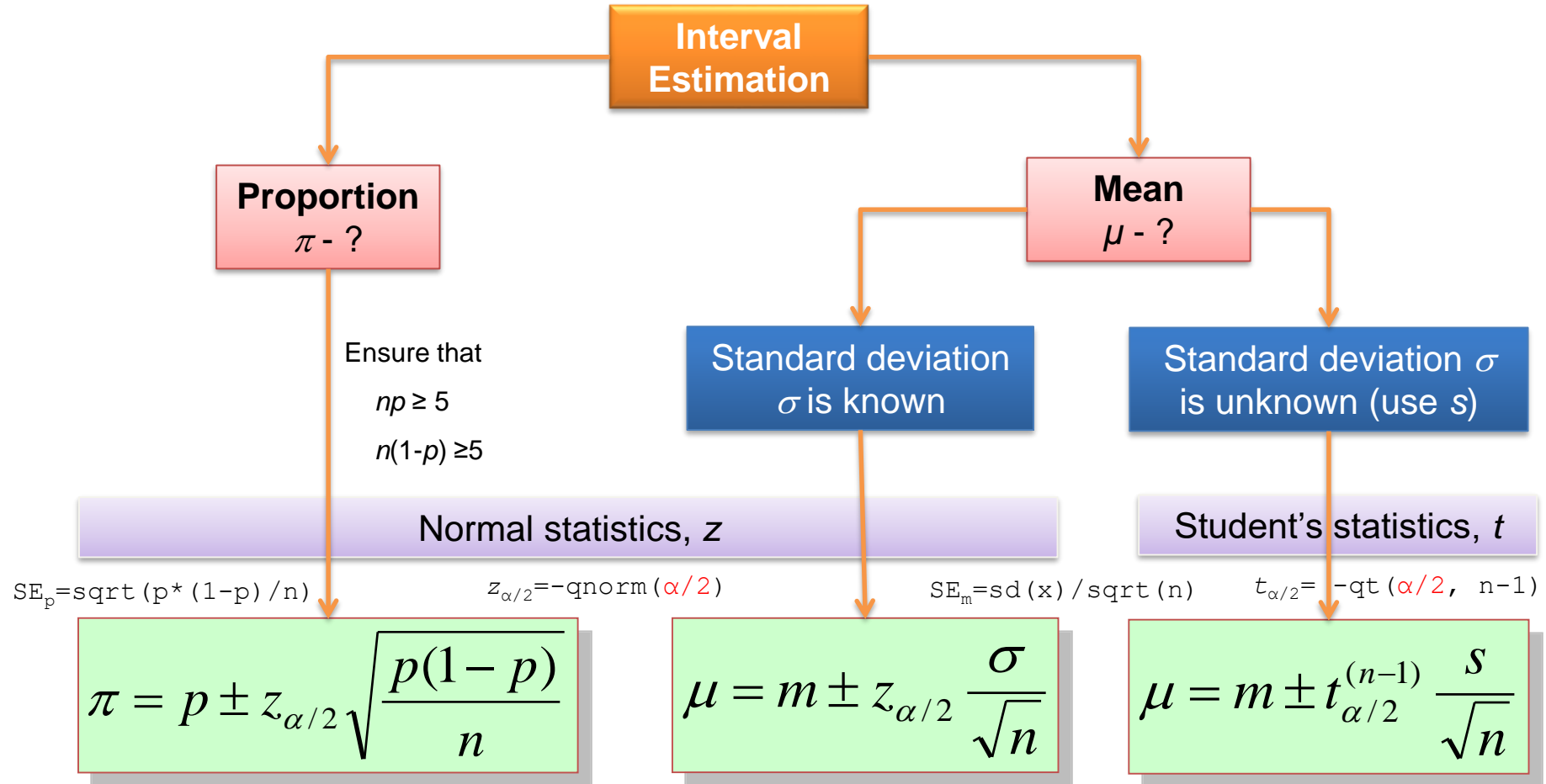
$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}$$

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}$$

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{E^2}$$

INTERVAL ESTIMATES

Summary



Alternatives:

= prop.test(c(x1, x2), c(n1, n2), conf.level=1-α)

= t.test(x, conf.level=1-α)

INTERVAL ESTIMATES FOR RANDOM FUNCTIONS

Sum and Square of Normal Variables

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.

$x \pm y \rightarrow \text{Normal distribution}$

$$E[x \pm y] = E[x] \pm E[y]$$

$$\sigma_{x \pm y}^2 = \sigma_x^2 + \sigma_y^2$$

Distribution of sum of squares on k standard normal random variables

The sum of squares of k standard normal random variables is a χ^2 with k degree of freedom.

if $x_1, \dots, x_k \rightarrow \text{Normal distribution}$

$$\sum_{i=1}^k x_i^2 \rightarrow \chi^2 \quad \text{with } d.f. = k$$

What to do in more complex situations?

$$\frac{x}{y} \rightarrow ?$$

$$\sqrt{x} \rightarrow ?$$

$$\log(|x|) \rightarrow ?$$

Try to solve analytically?

Simplest case. $E[x] = E[y] = 0$

Ratio distribution

From Wikipedia, the free encyclopedia

A **ratio distribution** (or *quotient distribution*) is a [probability distribution](#) constructed as the distribution of the [ratio of random variables](#) having two other known distributions. Given two random variables X and Y , the distribution of the random variable Z that is formed as the ratio

$$Z = X/Y$$

is a *ratio distribution*.

$$p_Z(z) = \frac{b(z) \cdot c(z)}{a^3(z)} \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y} \left[2\Phi\left(\frac{b(z)}{a(z)}\right) - 1 \right] + \frac{1}{a^2(z) \cdot \pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_y^2}{\sigma_y^2}\right)}$$

where

$$a(z) = \sqrt{\frac{1}{\sigma_x^2}z^2 + \frac{1}{\sigma_y^2}}$$

$$b(z) = \frac{\mu_x}{\sigma_x^2}z + \frac{\mu_y}{\sigma_y^2}$$

$$c(z) = e^{\frac{1}{2}\frac{b^2(z)}{a^2(z)} - \frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_y^2}{\sigma_y^2}\right)}$$

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

Practical Approach

Experimental values (x) and control (y) were measured for an experiment. 5 replicates were performed for each.

From previous experience, we know that the error between replicates is not too different from normal distribution.

Q: provide an interval estimation for the fold change x/y ($\alpha=0.05$)

(*)

#	Experiment	Control
1	215	83
2	253	75
3	198	62
4	225	91
5	240	70

Mean	226.2	76.2
StDev	21.39	11.26

Let us use a *numerical simulation...*

(*) this specific case can be solved in different ways, e.g. using log transformation:

$$\log(x/y) = \log(x) - \log(y)$$

Practical Approach

1. Calculate standard errors from s.t.d. $s_x \rightarrow se_x$

$$se_x = \frac{s_x}{\sqrt{n}}$$

Mean	226.2	76.2
StDev	21.39	11.26

Mean	226.2	76.2
StDev	9.57	5.03

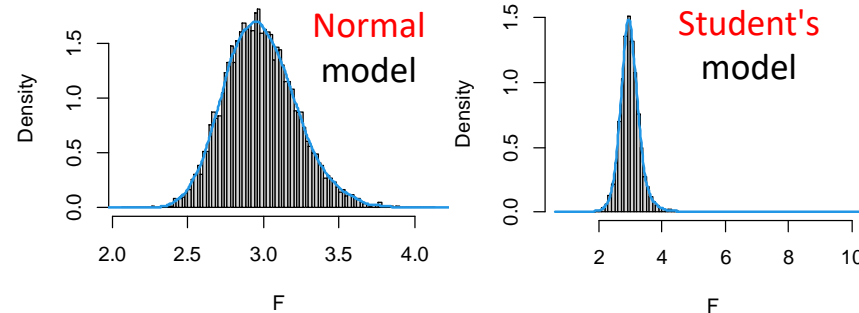
2. Generate 2 sets of 10k **normal** or **Student's** random variables M_x and M_y with means of x and y and standard deviations se_x , se_y corresponding to ones of experimental and control set (assume you perform 10000 series by $n=5$ experiments in each).

3. Build the target function: $f(x,y)$

4. Study the target function. Calculate summary, build histogram if necessary.

5. If you would like to have 95% interval, calculate 2.5% and 97.5% percentiles.

$$E[m_x/m_y] \in [2.39, 3.74]$$



for ratio $x,y > 0$, you can also use log:

6. If we need "prediction interval" instead of "confidence" – use x and y instead of m_x , m_y .

```
# enter data
x = c(215,253,198,225,240)
y = c(83,75,62,91,70)

# means and standard errors
mx = mean(x)
my = mean(y)
sex = sd(x)/sqrt(5)
sey = sd(y)/sqrt(5)

# simulation (can try rnorm)
Mx = mx + sex*rt(10000,5-1)
My = my + sey*rt(10000,5-1)
FXY = Mx/My

# visualization
hist(FXY, 50, freq = FALSE)
lines(density(FXY), lwd=2, col=4)

# confidence interval for means
quantile(Mx/My, c(0.025,0.975))
```

```
# use log-transformed data:
mx = mean(log2(x))
my = mean(log2(y))
sex = sd(log2(x))/sqrt(length(x))
sey = sd(log2(y))/sqrt(length(x))

m = mx - my
s = sqrt(sex^2 + sey^2)

me1 = m + qt(0.025,5-1)*s
me2 = m + qt(0.975,5-1)*s

2^m
2^me1
2^me2
```

HYPOTHESES about MEAN for ONE SAMPLE

Hypotheses

Confusion matrix: TP, FP, TN, FN and errors

Hypotheses about the mean of one sample

Hypotheses about the proportion of one sample

P-value

Power of the test

HYPOTHESES

Null and Alternative Hypotheses

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

$$H_0: \mu \leq \text{const}$$

$$H_a: \mu > \text{const}$$

$$H_0: \mu \geq \text{const}$$

$$H_a: \mu < \text{const}$$

$$H_0: \mu = \text{const}$$

$$H_a: \mu \neq \text{const}$$

Developing Null and Alternative Hypotheses: Example 1

Assume that the average survival time for glioblastoma patients (early state, age<50) is **18** months. You have developed a new treatment that should increase the survival time. Performing the clinical trial in order to determine the positive effect, you obtained an average survival of 20 months. You would like to ensure that this effect is real, so you perform the hypothesis testing. A research hypothesis should be stated as the alternative hypothesis as a general guideline. Hence, the appropriate null and alternative hypotheses for the study are

$$H_0: \mu \leq 18$$

$$H_a: \mu > 18$$

If the sample results indicate that H_0 cannot be rejected, researchers cannot conclude the new treatment is better. Perhaps more research and subsequent testing should be conducted. However, if the sample results indicate that H_0 can be rejected, researchers can make the inference that $H_a: \mu > 18$ is true. With this conclusion, the researchers gain the statistical support necessary to state that the new treatment increases survival time, and wide implementation of the treatment should be made.

Developing Null and Alternative Hypotheses: Example 2

Consider the situation of a manufacturer of soft drinks who states that it fills two-liter containers of its products with an average of at least **67.6** fluid ounces. A sample of two-liter containers will be selected, and the contents will be measured to test the manufacturer's claim. In this type of hypothesis testing situation, we generally assume that the manufacturer's claim is true unless the sample evidence is contradictory. Using this approach for the soft-drink example, we would state the null and alternative hypotheses as follows.

$$H_0: \mu \geq 67.6$$

$$H_a: \mu < 67.6$$

If the sample results indicate H_0 cannot be rejected, the manufacturer's claim will not be challenged. However, if the sample results indicate H_0 can be rejected, the inference will be made that $H_a: \mu < 67.6$ is true. With this conclusion, statistical evidence indicates that the manufacturer's claim is incorrect and that the soft-drink containers are being filled with a mean less than the claimed 67.6 ounces. Appropriate action against the manufacturer may be considered in a court.

Developing Null and Alternative Hypotheses: Example 3

For example, on the basis of a sample of parts from a shipment just received, a quality control inspector must decide whether to accept the shipment or to return the shipment to the supplier because it does not meet specifications. Assume that specifications for a particular part require a mean length of **2** inches per part. If the mean length is greater or less than the two-inch standard, the parts will cause quality problems in the assembly operation. In this case, the null and alternative hypotheses would be formulated as follows.

$$H_0: \mu = 2$$

$$H_a: \mu \neq 2$$

If the sample results indicate H_0 cannot be rejected, the quality control inspector will have no reason to doubt that the shipment meets specifications, and the shipment will be accepted. However, if the sample results indicate H_0 should be rejected, the conclusion will be that the parts do not meet specifications. In this case, the quality control inspector will have sufficient evidence to return the shipment to the supplier.

HYPOTHESES

Type I Error

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.

Conclusion

		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

poor sensitivity

**False Negative,
 β error**

**False Positive,
 α error**

poor specificity

Level of significance

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

HYPOTHESIS TESTING FOR MEAN

One-tailed Test

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

$$H_0: \mu \leq \mu_0$$

$$H_a: \mu > \mu_0$$

$$H_0: \mu \geq \mu_0$$

$$H_a: \mu < \mu_0$$



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.

$$\mu_0 = 3 \text{ lbm}$$

Suppose sample of $n=36$ coffee cans is selected. From the previous studies it's known that $\sigma = 0.18$ lbm

HYPOTHESIS TESTING FOR MEAN

One-tailed Test: Example

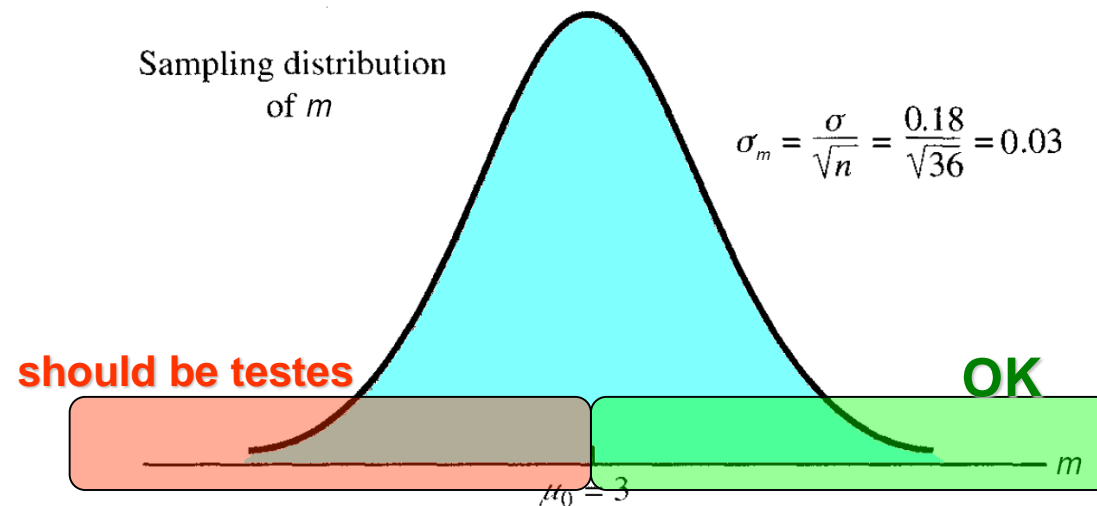
$$\mu_0 = 3 \text{ lbm}$$

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

$$H_0: \mu \geq 3 \quad \text{no action}$$

$$H_a: \mu < 3 \quad \text{legal action}$$

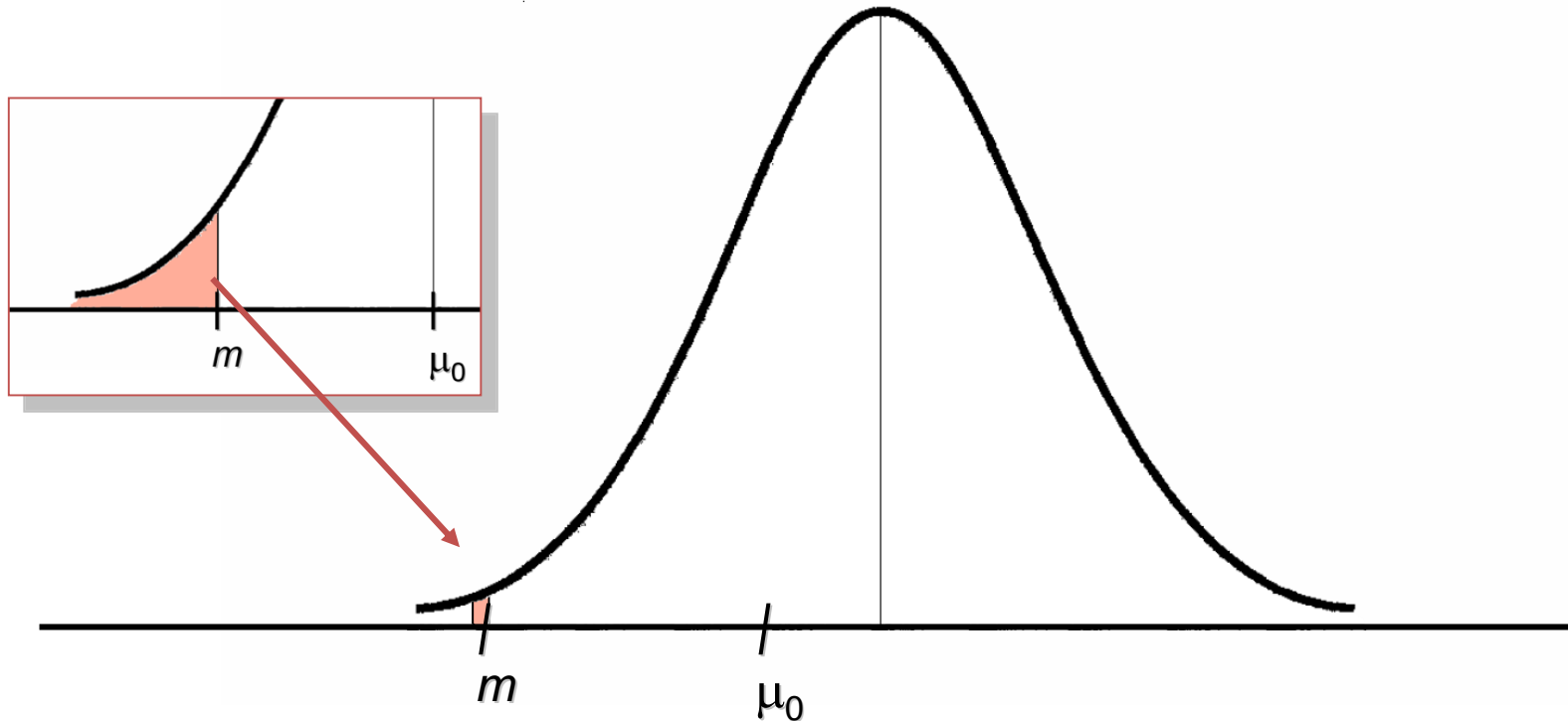
Let's say: **in the extreme case**, when $\mu=3$, we would like to be 99% **sure that we make no mistake**, when starting legal actions against Hilltop Coffee. It means that selected significance level is $\alpha = 0.01$



HYPOTHESIS TESTING FOR MEAN

Let's Try to Understand...

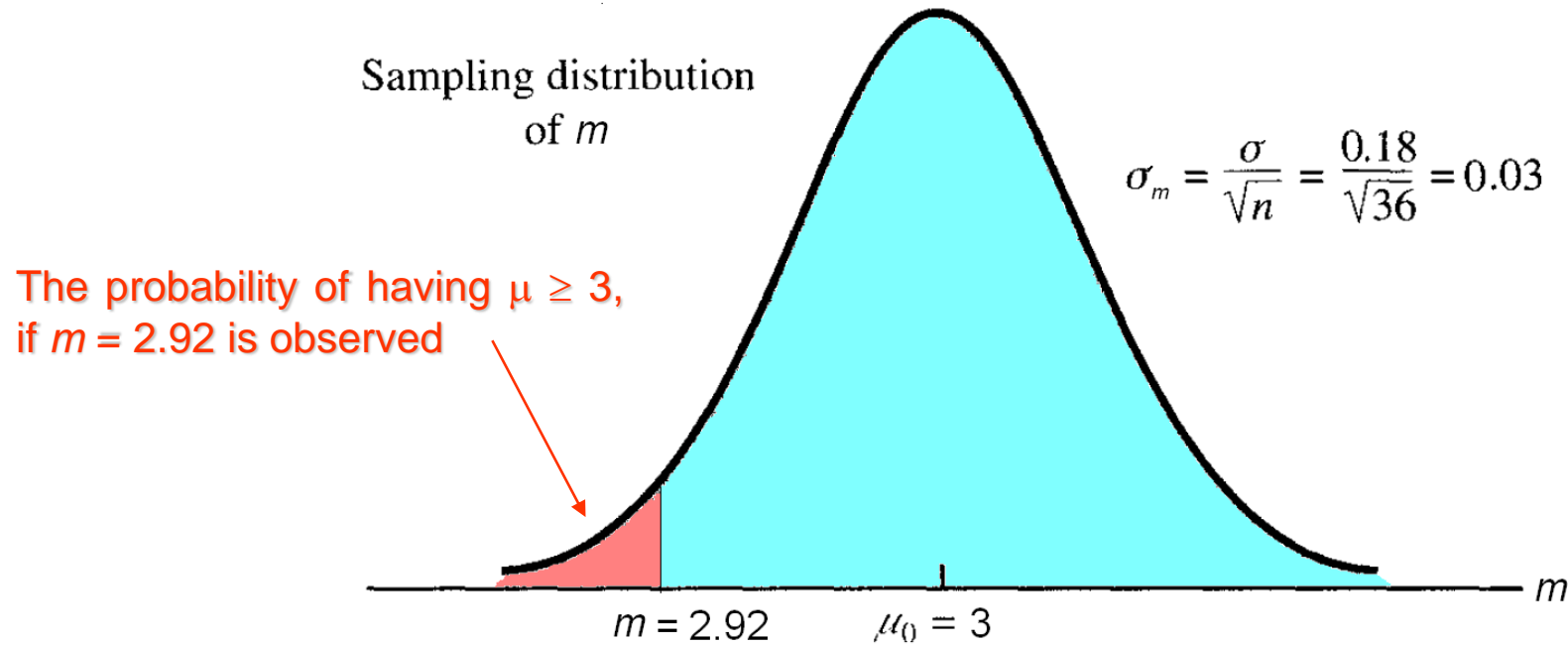
Let's find the probability of observation m for all possible $\mu \geq 3$. We start from an **extreme case** ($\mu=3$) and then probe all possible $\mu > 3$. See the behavior of the **small probability area** around measured m . What you will get if you **summarize its area for all possible $\mu \geq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



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

HYPOTHESIS TESTING FOR MEAN

p-value



p-value

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

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

...well, the statement, maybe, is not completely correct, but it helps get an idea.

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

```
# assumed population mean
mu0 = 3
# observed sample mean
m = 2.92
# known population st.dev.
sigma = 0.18
# sample size
n = 36
## standard error
se = sigma / sqrt(n)
## p-value
pnorm(m-mu0, mean = 0, sd = se)
```

HYPOTHESIS TESTING FOR THE MEAN

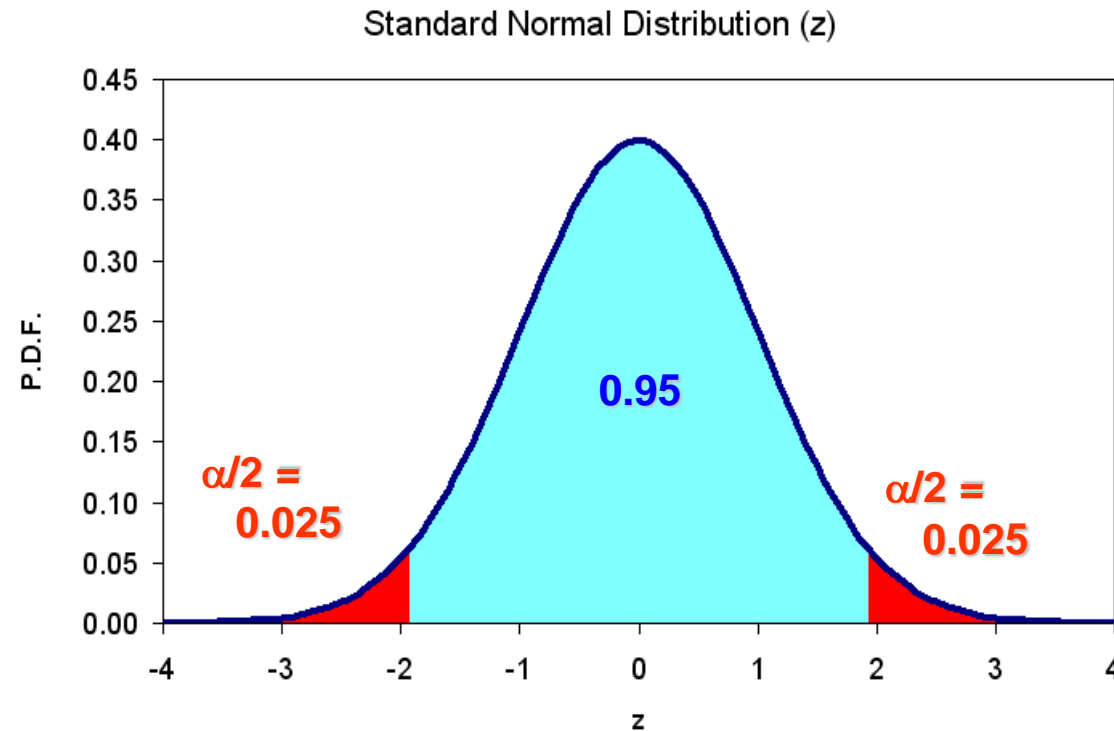
Two-tailed Test

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.

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$



HYPOTHESIS TESTING FOR THE MEAN

σ is Unknown (summary)

if σ is unknown:

$$\sigma \rightarrow S$$

$$Z \rightarrow t$$

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

Excel:

$m = \text{AVERAGE}(\dots)$

$n = \text{number of experiments}$

$\sigma = \text{population standard deviation}$

$\mu_0 = \text{population mean (constant)}$

$z = (m - \mu_0) / \sigma * \text{SQRT}(n)$

p-value = `T.DIST(-ABS(z), n-1, true)`

	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	$t = \frac{m - \mu_0}{s/\sqrt{n}}$	$t = \frac{m - \mu_0}{s/\sqrt{n}}$	$t = \frac{m - \mu_0}{s/\sqrt{n}}$
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 $t \geq t_{\alpha/2}$

$\times \leftarrow$ put in your data

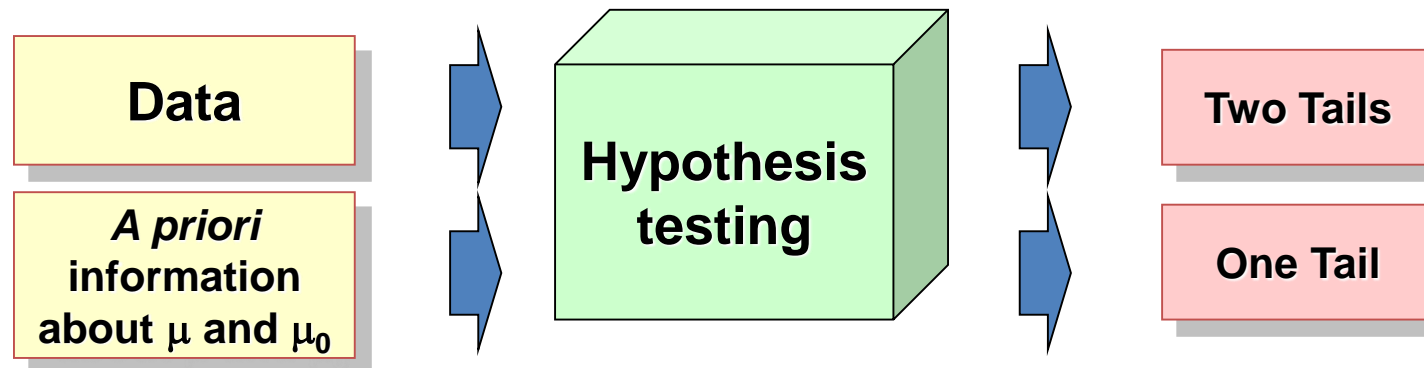
`t.test(x, mu=1,
alternative = "two.sided")`

other hypotheses "less", "greater"

HYPOTHESIS TESTING FOR THE MEAN

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.



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



HYPOTHESIS TESTING FOR THE MEAN

Example

The number of living cells in 5 wells under some conditions is given in the table, with an average value of **4705**. In a reference literature source authors claimed a mean quantity of **5000** living cells under the same conditions. Is our result significantly different?

Well	Cells
1	5128
2	4806
3	5037
4	4231
5	4322

One Tail

$H_0: \mu \geq 5000$
 $H_a: \mu < 5000$
 Let's use $\alpha=0.05$

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

Two Tails

$H_0: \mu = 5000$
 $H_a: \mu \neq 5000$
 Let's use $\alpha=0.05$

```
m = AVERAGE(A2:A6)
s = STDEV.S(A2:A6)
mu_0 = 5000
n = 5
t = (m - mu_0) / s * SQRT(n)
p-value1 = T.DIST(-ABS(t); n-1; true)
p-value2 = 2 * T.DIST(-ABS(t); n-1; true)
```

```
n          5
mean      4704.8
stdev     409.49
mu        5000
t         -1.612
p-value 2 t 0.1823
p-value 1 t 0.0911
```

```
x = c(5128, 4806, 5037, 4231, 4322)
pv1 = t.test(x, mu=5000, alternative="less")
pv2 = t.test(x, mu=5000, alternative="two.sided")
```

HYPOTHESIS TESTING FOR PROPORTION

Hypotheses for Proportions

Proportions

π – population proportion

p – experimental proportion

π_0 – tested proportion

For the proportions (in R):

- 1) if sample is large ($np > 5$ and $n(1-p) > 5$) you can use **prop.test()**
- 2) otherwise: **binom.test()**

n ← number of observations

p ← experimental proportion

π_0 ← tested proportion

τ ← number of tails

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0 : \pi \geq \pi_0$ $H_a : \pi < \pi_0$	$H_0 : \pi \leq \pi_0$ $H_a : \pi > \pi_0$	$H_0 : \pi = \pi_0$ $H_a : \pi \neq \pi_0$
Test Statistic	$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$	$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$	$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$
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 $z \leq -z_\alpha$	Reject H_0 if $z \geq z_\alpha$	Reject H_0 if $z \leq -z_{\alpha/2}$ or if $z \geq z_{\alpha/2}$

$$s_p = \text{SQRT}(\pi_0 * (1 - \pi_0) / n)$$

$$z = (p - \pi_0) / s_p$$

$$\text{pval} = \tau * \text{NORM.S.DIST}(-\text{ABS}(z), \text{TRUE})$$

```
prop.test(x, n, p, ...) # chi2 approx
binom.test(x, n, p, ...) # exact test
```

HYPOTHESES FOR ONE SAMPLE

Summary

p-value > α

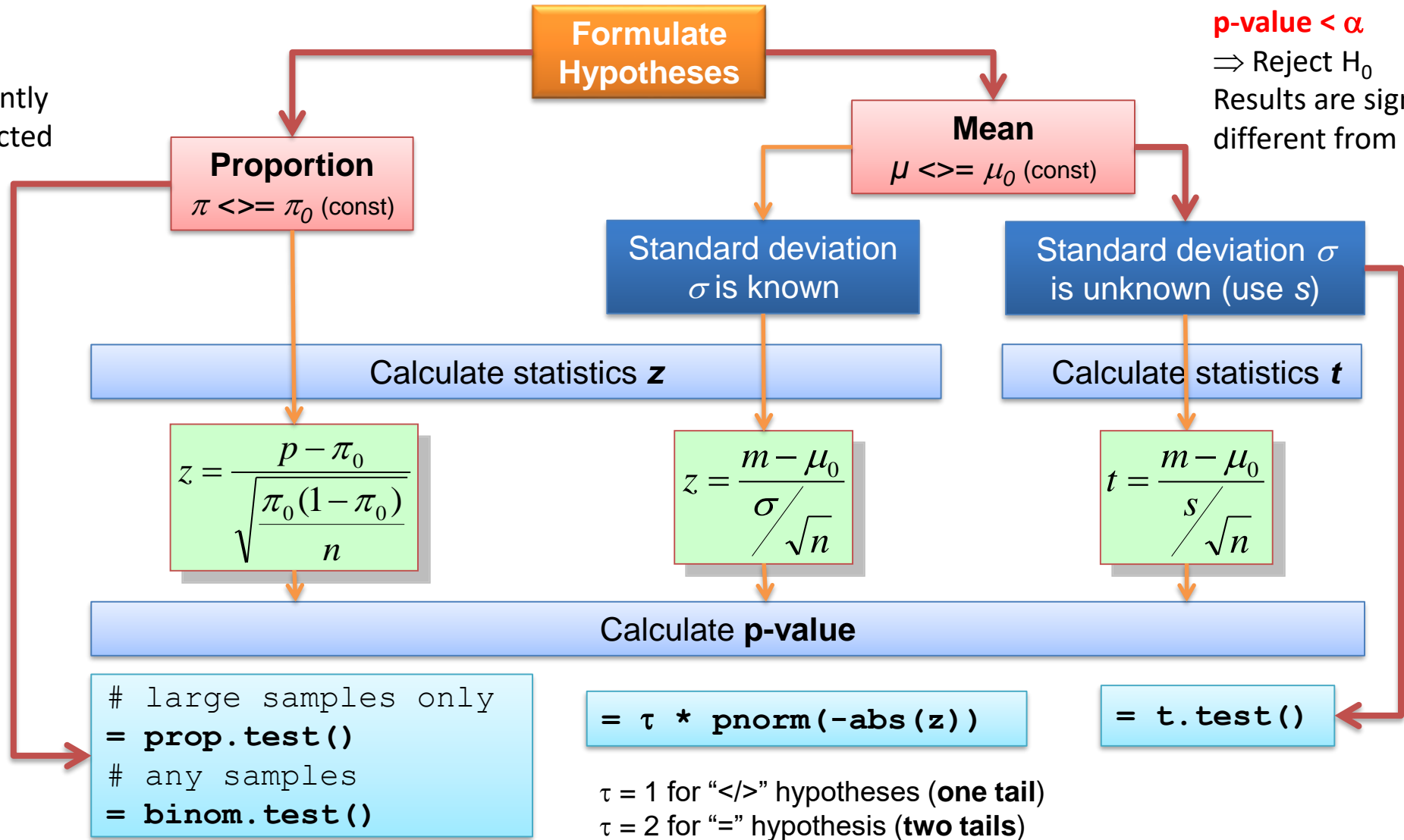
⇒ Accept H_0

Results are not significantly different from the expected

p-value < α

⇒ Reject H_0

Results are significantly different from the expected



HYPOTHESES

Power: how to control Type II Error

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

		Population Condition	
		H_0 True	H_a True
Conclusion	Accept H_0	Correct Conclusion	Type II Error <i>poor sensitivity</i> False Negative, β error
	Reject H_0	Type I Error False Positive, α error <i>poor specificity</i>	Correct Conclusion

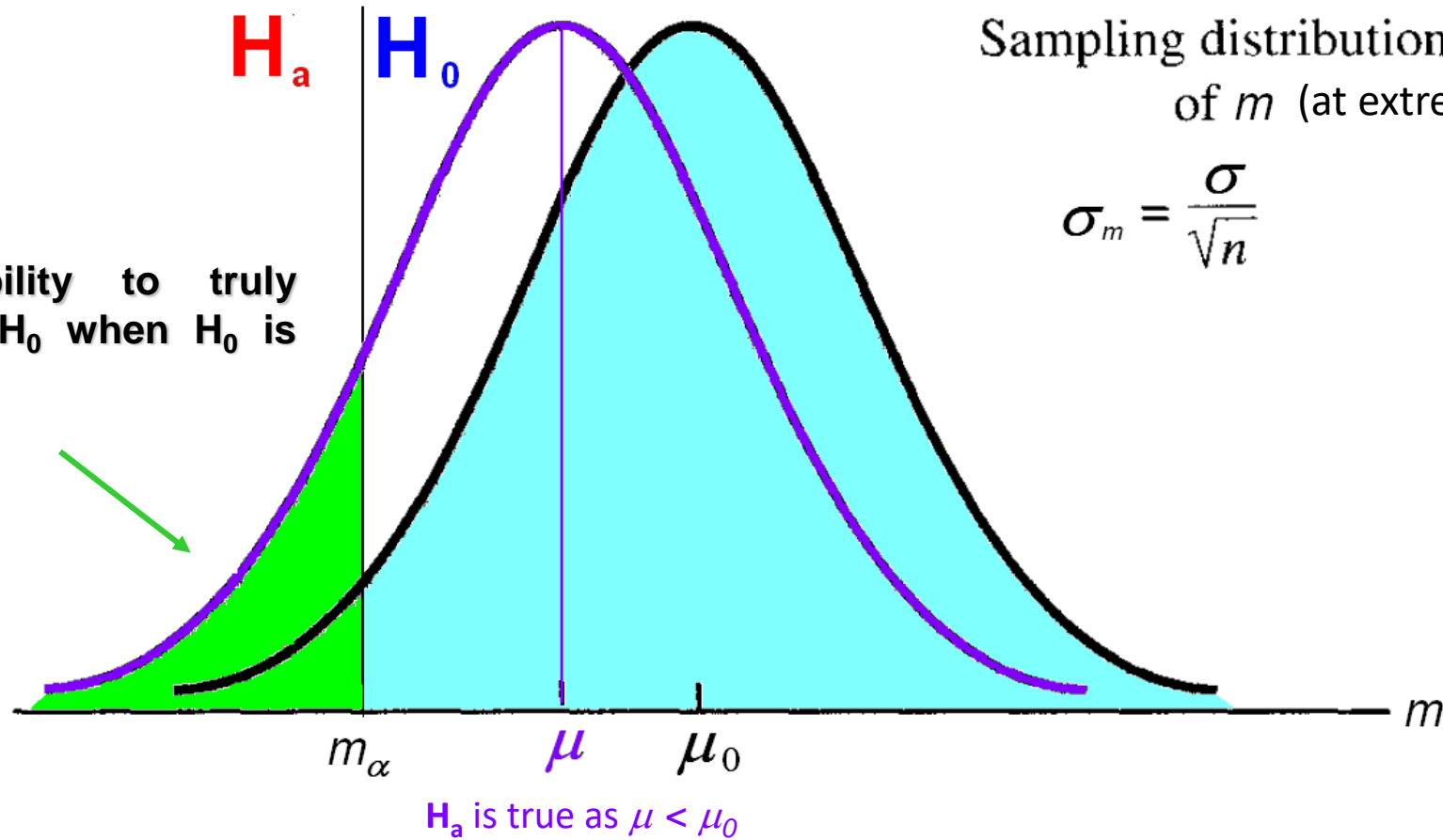
HYPOTHESES

Power Curve

$$H_0: \mu \geq \mu_0$$

$$H_a: \mu < \mu_0$$

Probability to truly reject H_0 when H_0 is false.



HYPOTHESES

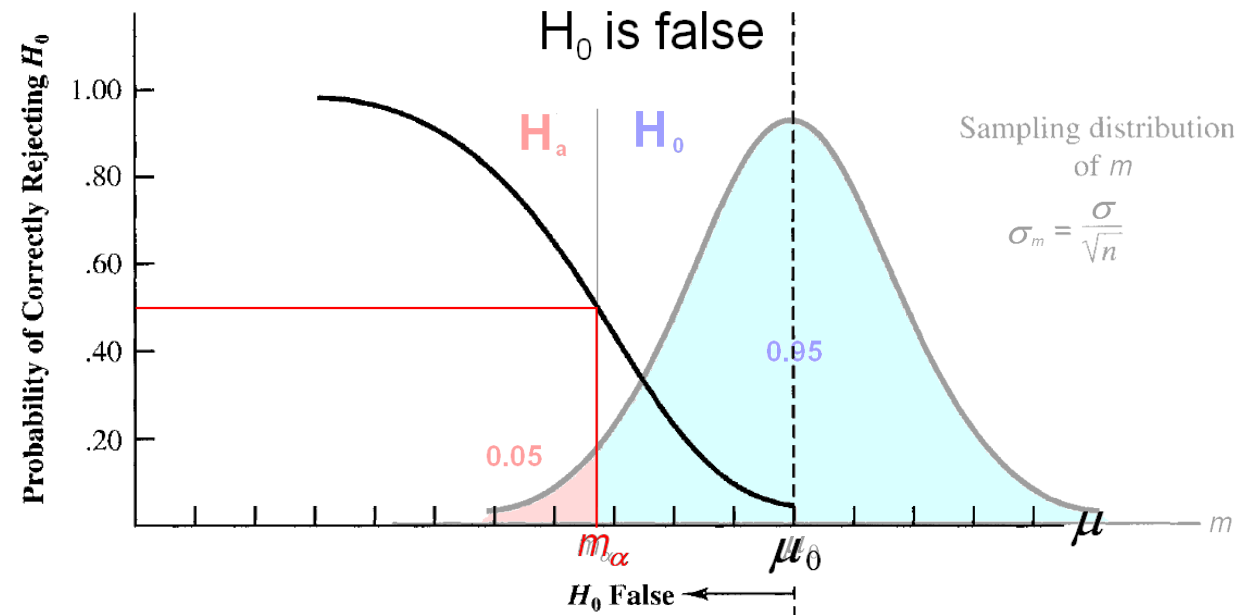
Power Curve

Power

The probability of correctly rejecting H_0 when it is false, **power** = $1 - \beta$

Power curve

A graph of the probability of rejecting H_0 for all possible values of the population parameter not satisfying the null hypothesis. The power curve provides the probability of correctly rejecting the null hypothesis



HYPOTHESES about MEANS for TWO SAMPLES

Independent and matched samples

Hypotheses about the means of two samples

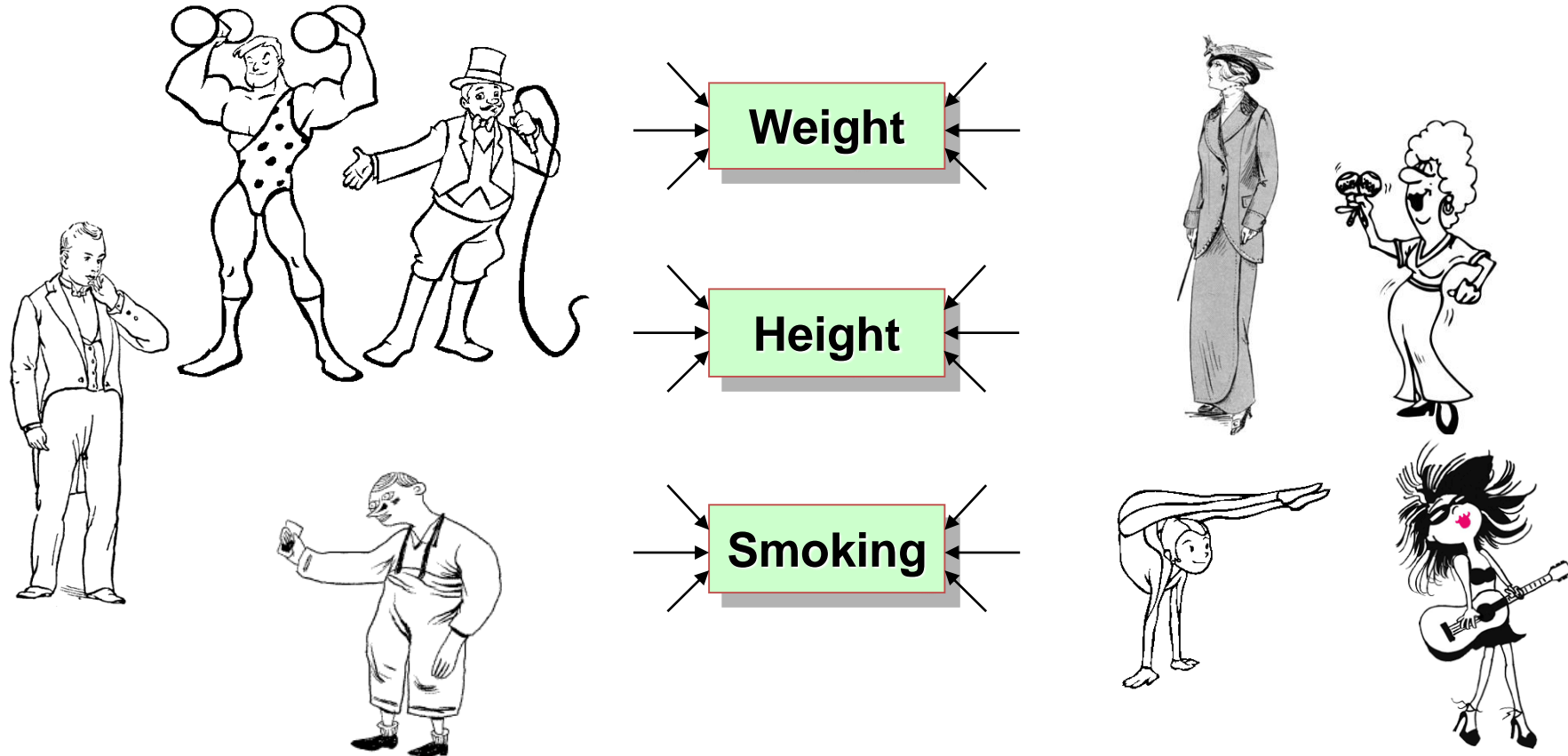
Hypotheses about the two proportions

TWO POPULATIONS

Independent Samples

Independent samples

Samples are 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.



TWO POPULATIONS

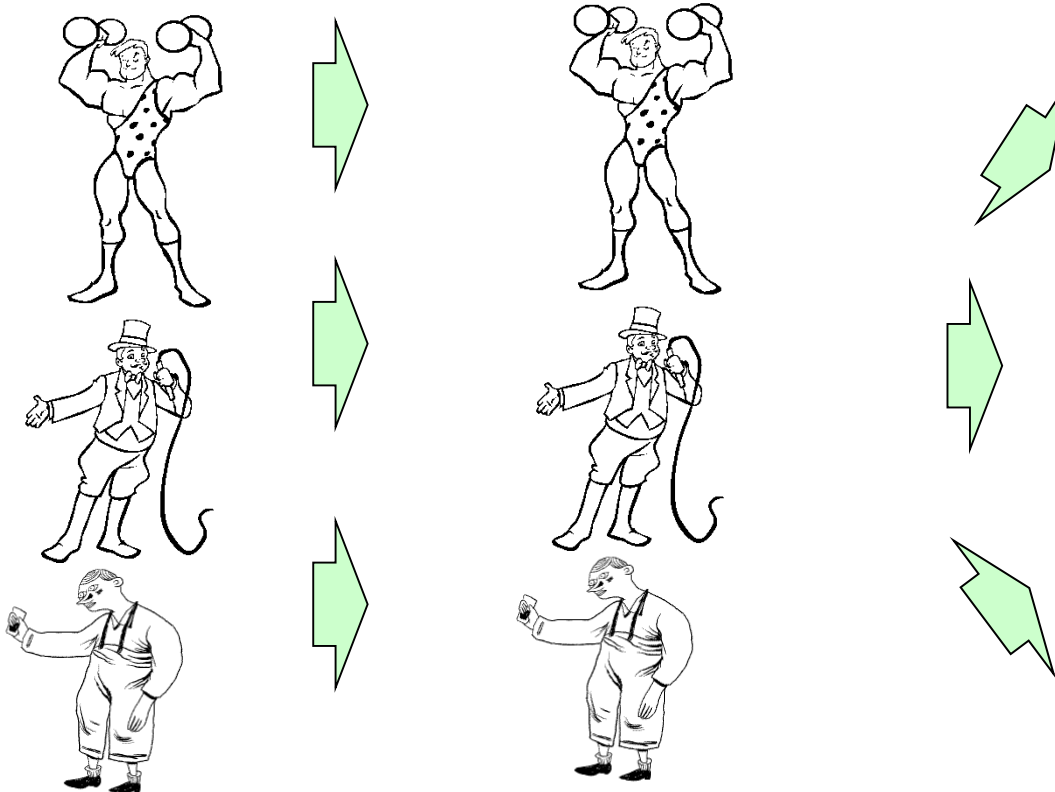
Matched (paired) Samples

Matched samples

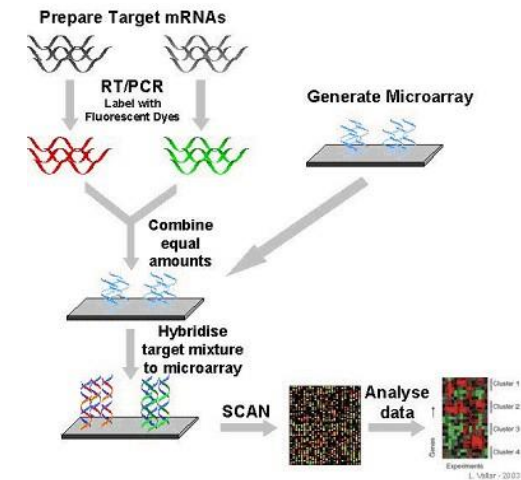
Samples, in which each data value of one sample is matched with a corresponding data value of the other sample.

Before treatment

After treatment



Analysis



MEANS OF TWO POPULATIONS

Independent Samples: Example

mice.xls



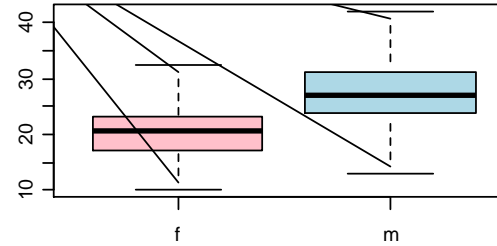
outliers
are
removed
from
boxplots

Q1: Is body weight significantly different for males and females?

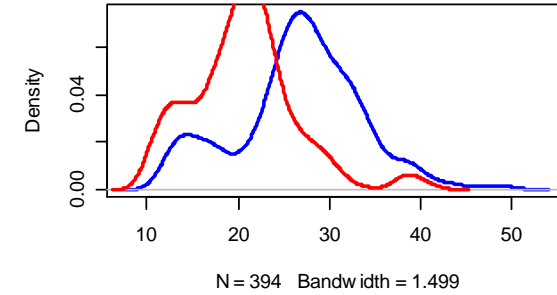
Q2: Is weight change significantly different for males and females?

Q3: Is bleeding time significantly different for males and females?

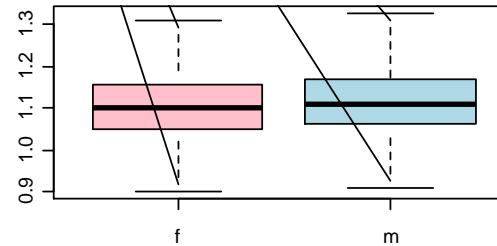
Final body weights (g)



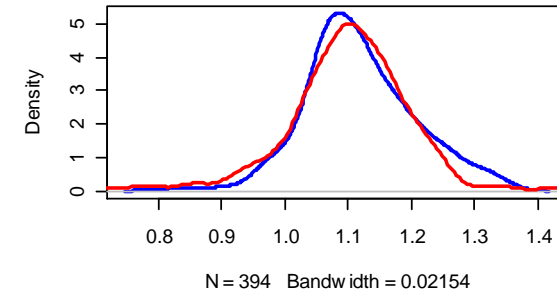
Body weight distributions



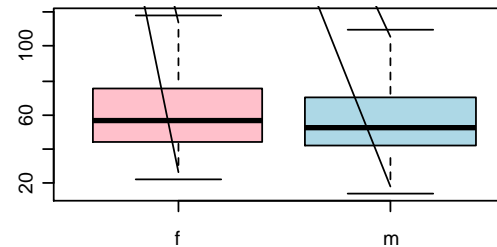
Weights change (g)



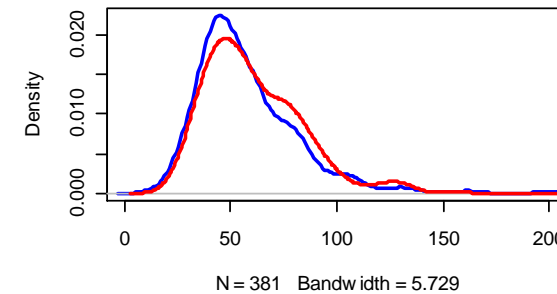
Distributions of weight change



Bleeding time (g)



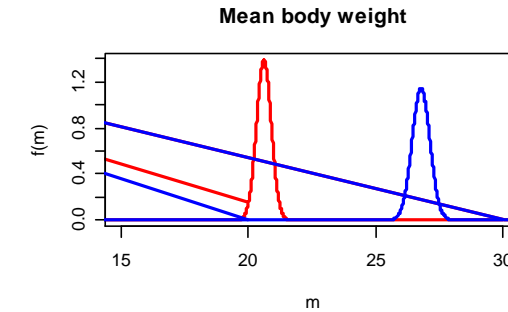
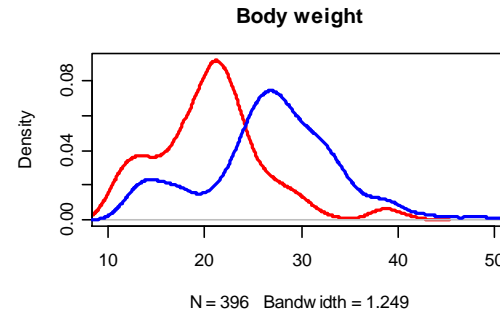
Distributions of bleeding times



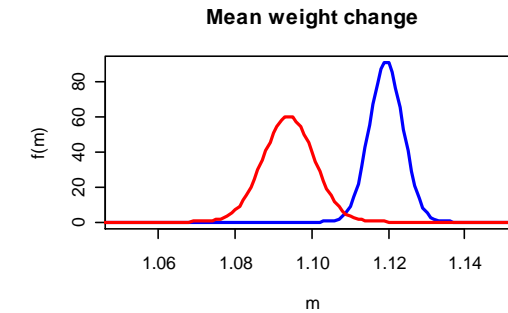
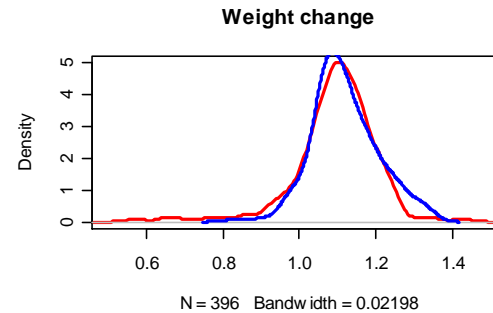
MEANS OF TWO POPULATIONS

Independent Samples: Example

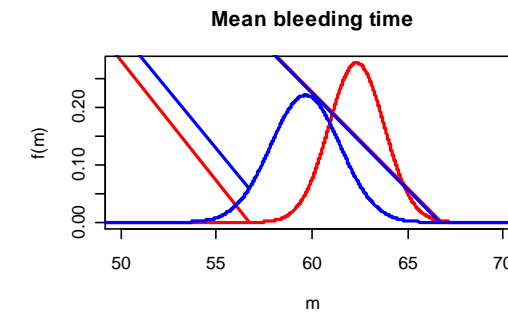
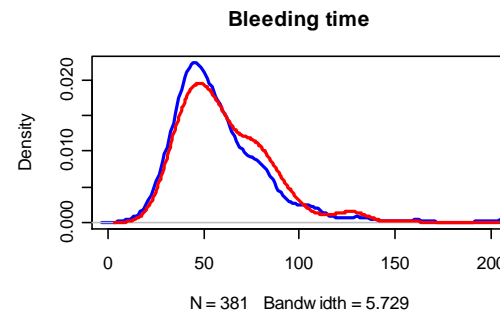
Q1: Is body weight for males and females significantly different?



Q2: Is weight change for males and females significantly different?



Q3: Is bleeding time significantly different for males and females?



HYPOTHESES

Theory

Two tail hypothesis

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

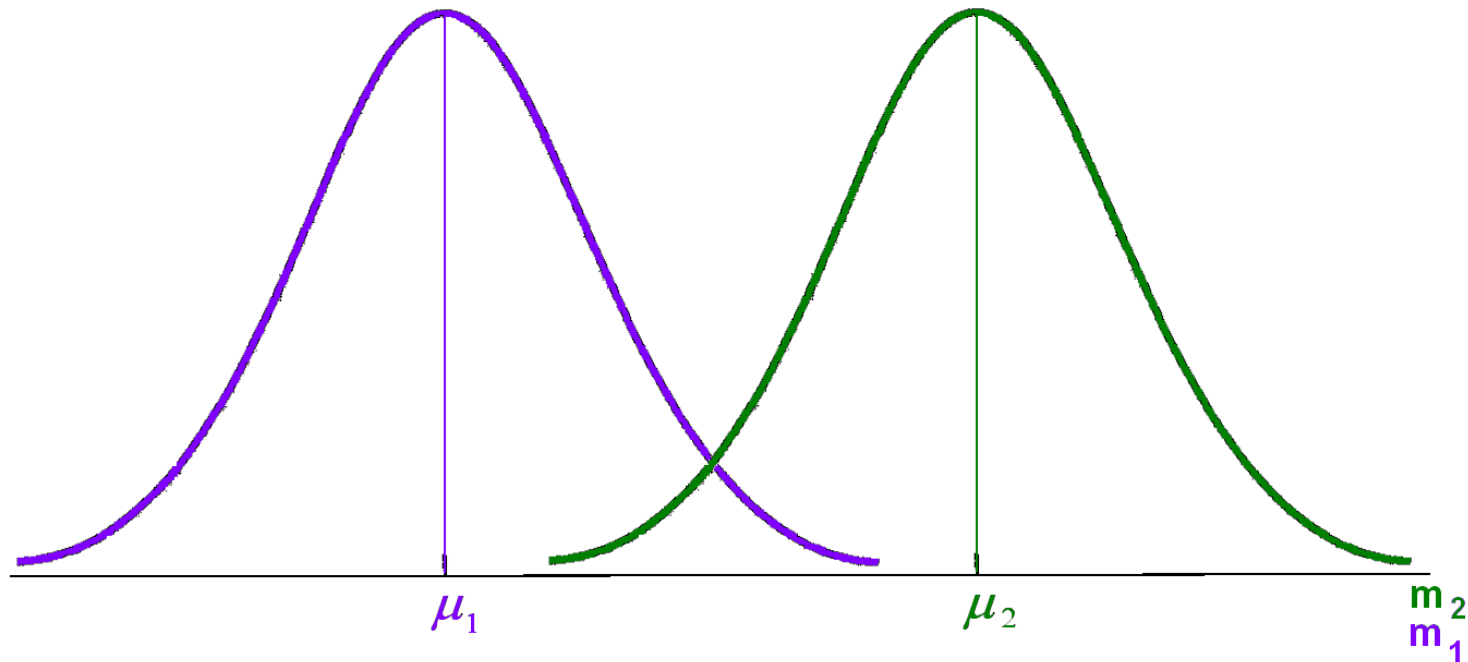
One tail hypothesis

$$H_0: \mu_1 \geq \mu_2$$

$$H_a: \mu_1 < \mu_2$$

$$H_0: \mu_1 \leq \mu_2$$

$$H_a: \mu_1 > \mu_2$$



COMPARING MEANS

Theory

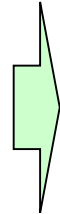
As we know how to work with standard hypotheses (comparison with constant μ_0), let us transform our hypothesis:

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$



$$H_0: \mu_2 - \mu_1 = 0$$

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

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	$\sigma_1^2 + \sigma_2^2$

COMPARING MEANS

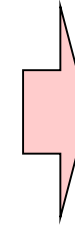
Theory

$$H_0: \mu_2 - \mu_1 = D_0$$

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

$$D_0 = \mu_2 - \mu_1$$

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



$$D_0 = m_2 - m_1$$

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

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 a t-test !!!

COMPARING MEANS

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:

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

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

$$(n_1 + n_2) / 2 < df < n_1 + n_2$$

2. Calculate the p-value

`pt(t, df)`

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

😊. Or simply do:

`t.test(x, y)`

`= T.TEST(data1, data2, 2, 3)`



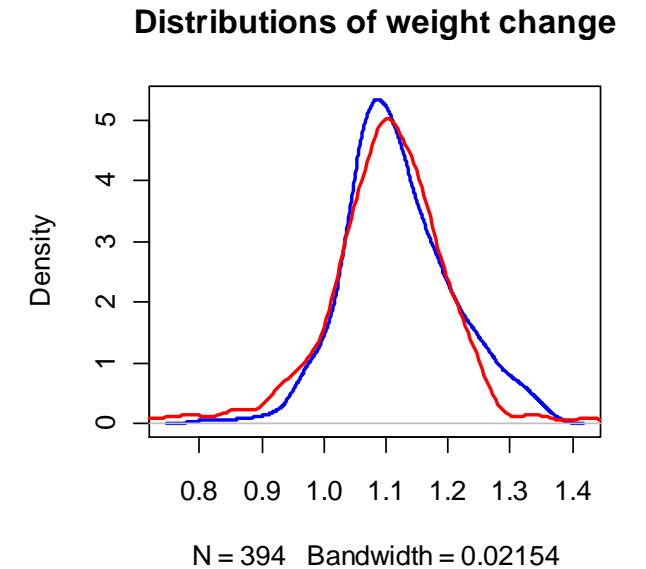
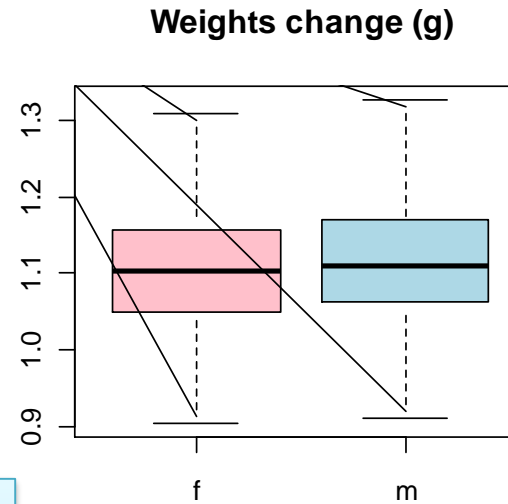
UNPAIRED T-TEST

Example

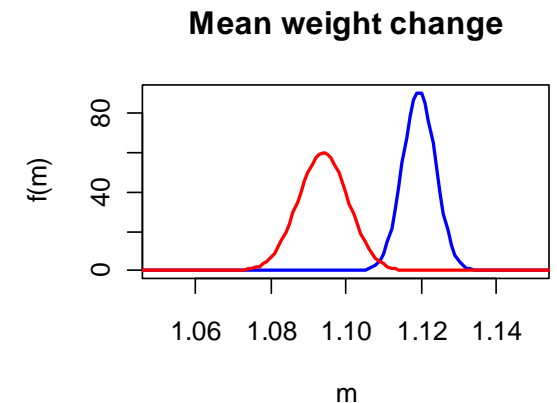
mice

Q2: Is the mean of weight change significantly different for males and females?

```
# slow way  
x = Mice$Weight.change[Mice$Sex == "f"]  
y = Mice$Weight.change[Mice$Sex == "m"]  
t.test(x,y)  
  
# fast way  
t.test(Mice$Weight.change ~ Mice$Sex)  
  
# get p-value  
t.test(...)$p.value
```



p-value = 0.0014



PAIRED T-TEST

Theory and Example

Paired t-test

In a paired t-test, instead of testing $H_0: \mu_2 - \mu_1 = 0$, use following steps:

1. Build a new random value $y = x_1 - x_2$ (subtract matched values).
2. Test whether one-sample mean $\mu_y = 0$

```
t.test(x, y,  
paired = TRUE)
```

bloodpressure

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

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.

Q: Does the treatment affect the systolic blood pressure?

Test	p-value
unpaired	0.414662
paired	0.014506

Unpaired test
= T.TEST (array1, array2, 2, 3)
Paired test
= T.TEST (array1, array2, 2, 1)

```
BP = read.table(  
"http://edu.modas.lu/data/txt/  
bloodpressure.txt",  
sep="\t", header=TRUE)  
  
t.test(BP$BP.before,  
BP$BP.after,  
paired = TRUE)
```

COMPARING PROPORTIONS

Theory

$H_0: \pi_1 = \pi_2$ $H_a: \pi_1 \neq \pi_2$	\Rightarrow	$H_0: \pi_1 - \pi_2 = 0$ $H_a: \pi_1 - \pi_2 \neq 0$	$\sigma_{p_1-p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
--	---------------	--	---

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

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

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

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

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

```
prop.test(x = c(x1, x2),
          n = c(n1, n2))
```


NON-PARAMETRIC TESTS

Non-parametric unpaired test

Mann-Whitney-Wilcoxon U-test

Non-parametric paired test

Wilcoxon Signed Rank Test

MEASURES

Parametric and Non-parametric Measures & Tests

Non-parametric measures

A statistical measure that does **not depend on particular data distribution**.

Non-parametric statistics / tests are usually performed on **ranks**.

Non-parametric procedures are more robust to outliers but less powerful than parametric ones.

parametric

Mean

Standard deviation

Pearson correlation

Non-paired (simple) t-test

Paired t-test

non-parametric

Median

MAD (median
absolute deviation)

Spearman correlation

Mann-Whitney test

Wilcoxon signed rank test

NON-PARAMETRIC TESTS

Mann-Whitney-Wilcoxon U-test

Non-paired (simple) t-test

Mann-Whitney test

`wilcox.test(x, y)`

Number of students on psychology and sociology

Psy	Soc
80	90
95	30
65	65
75	60
60	55
80	70
85	70
90	35
75	75
40	30

$$H_0: \mu_1 = \mu_2$$

t-Test: 0.06

$$H_a: \mu_1 \neq \mu_2$$

In Excel use for rank:

`RANK.AVG(x, TAB, 1)`

1. Create ranks for joint data and calculate sum for ranks for each dataset (R_1, R_2)
2. Calculate U1 and U2

$$U_i = n_1 n_2 + \frac{n_i(n_i + 1)}{2} - R_i$$

3. Take U = minimum (U1, U2)

4. Calculate z-statistics

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

5. Get p-value from z-stat by normal distribution

In Excel:

`2* NORM.S.DIST(z, TRUE)`

NON-PARAMETRIC TESTS

Mann-Whitney-Wilcoxon U-test



Psy	Soc	rank			Psy	Soc
80	90	15.5	18.5		Medians	77.5 62.5
95	30	20	1.5		SumRank	131.5 78.5 ← R1, R2
65	65	8.5	8.5		n=	10 10
75	60	13	6.5		U1, U2	23.5 76.5
60	55	6.5	5		U=	23.5
80	70	15.5	10.5		z=	-2.00321
85	70	17	10.5		p-val=	0.045155
90	35	18.5	3			
75	75	13	13			
40	30	4	1.5			

```
psy = c(80, 95, 65, 75, 60, 80, 85, 90, 75, 40)
soc = c(90, 30, 65, 60, 55, 70, 70, 35, 75, 30)
wilcox.test(psy, soc)
```

p-value = 0.04851

$$U_i = n_1 n_2 + \frac{n_i(n_i + 1)}{2} - R_i$$

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

NON-PARAMETRIC TESTS

Wilcoxon Signed Rank Test

Paired t-test

Wilcoxon signed rank test

before	after
12	412
42	312
31	63
462	632
1	0
25	20
63	124
754	5356
12	83
34	1245

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

paired t-Test p-value= **0.165**

Non significant... ???

Procedure for Wilcoxon (Siegel method):

1. Make difference between columns
2. Calculate rank for differences
3. Sum up ranks for positive and negative differences: $s+$ and $s-$. Now $T = \min(s+, s-)$
4. Calculate z-statistics and use standard z-test:

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

In Excel use for rank:

`RANK.AVG(diff,DIFF,1)`

`wilcox.test(x,y, paired=TRUE)`

NON-PARAMETRIC TESTS

Wilcoxon Signed Rank Test

Paired t-test

Wilcoxon signed rank test

before	after	diff	rank
12	412	400	8
42	312	270	7
31	63	32	3
462	632	170	6
1	0	-1	2
25	20	-5	1
63	124	61	4
754	5356	4602	10
12	83	71	5
34	1245	1211	9

s+ = SUMIF(diff;">0";rank)
s- = SUMIF(diff;"<0";rank)

s+=	52
s- =	3
T =	3
z =	-2.49727
pval(2t)=	0.012515

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

```
before = c(12, 42, 31,
462, 1, 25, 63, 754,
12, 34)
```

```
after = c(412, 312, 63,
632, 0, 20, 124, 5356,
83, 1245)
```

```
wilcox.test(before,
after,
paired = TRUE)
```

p-value = 0.009766

This is an approximate method. To increase power, use R

MULTIPLE TESTING

Correction for multiple testing

MULTIPLE TESTING

Correct and Wrong Conclusions

		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 Negative, β error (points to Type II Error)
 False Positive, α error (points to Type I Error)

Probability of an error in a multiple test:

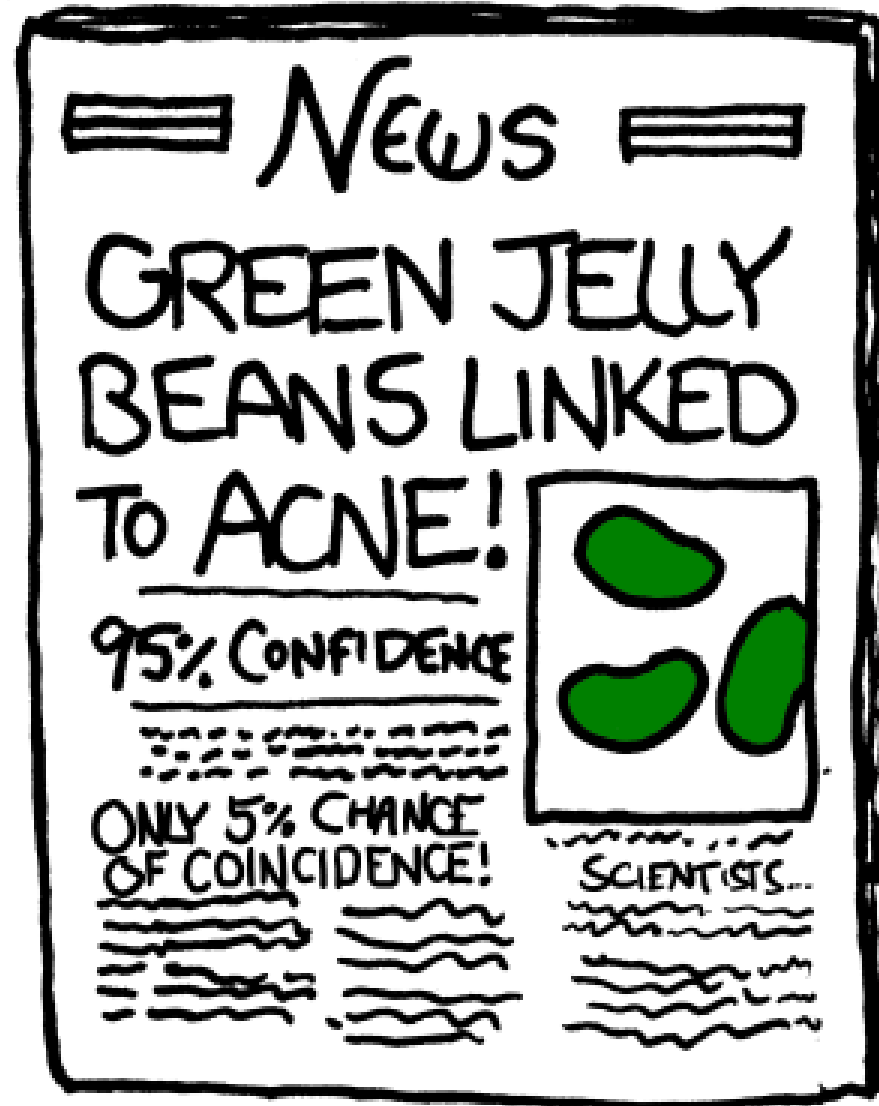
$$1 - (0.95)^{\text{number of comparisons}}$$

$$\text{n.o.c} = 10 \rightarrow p(\text{error}) = 0.4$$

$$\text{n.o.c} = 100 \rightarrow p(\text{error}) = 0.99$$

MULTIPLE TESTING

Example



<http://www.xkcd.com/882/>

MULTIPLE TESTING

False Discovery Rate

False discovery rate (FDR)

FDR control is a statistical method used in multiple-hypothesis testing to correct for multiple comparisons. FDR controls the expected proportion of incorrectly rejected null hypotheses (type I errors) in a list of rejected hypotheses.

		Population Condition		Total
		H ₀ is TRUE	H ₀ is FALSE	
Conclusion	Accept H ₀ (non-significant)	<i>U</i>	<i>T</i>	<i>m - R</i>
	Reject H ₀ (significant)	<i>V</i>	<i>S</i>	<i>R</i>
	Total	<i>m</i> ₀	<i>m - m</i> ₀	<i>m</i>

$$FDR = E\left(\frac{V}{V + S}\right)$$

Assume we need to perform $m = 100$ comparisons, and select maximum **FDR = $\alpha = 0.05$**

Independent tests

The **Simes procedure** ensures that its **expected value** $\mathbb{E} \left[\frac{V}{V + S} \right]$ is less than a given α (Benjamini and Hochberg

1995). This procedure is valid when the m tests are **independent**. Let $H_1 \dots H_m$ be the null hypotheses and $P_1 \dots P_m$ their corresponding **p-values**. Order these values in increasing order and denote them by

$P_{(1)} \dots P_{(m)}$. For a given α , find the largest k such that $P_{(k)} \leq \frac{k}{m} \alpha$.

Then reject (i.e. declare positive) all $H_{(i)}$ for $i = 1, \dots, k$.

Note that the mean α for these m tests is $\frac{\alpha(m+1)}{2m}$ which could be used as a rough FDR, or RFDR, " α adjusted

for m indep. tests." The RFDR calculation shown here provides a useful approximation and is not part of the Benjamini and Hochberg method; see AFDR below.

MULTIPLE EXPERIMENTS

FDR and FWER

Assume we need to perform $m = 100$ comparisons,
and select maximum **FDR = $\alpha = 0.05$**

k – is rank of p-value (order #)

$$FDR = E\left(\frac{V}{V+S}\right)$$

Expected value for $FDR < \alpha$ if

$$P_{(k)} \leq \frac{k}{m} \alpha$$

Benjamini-Hochberg (FDR)

$$\frac{mP_{(k)}}{k} \leq \alpha$$

```
p.adjust(pv,  
method="fdr")
```

Other Methods

Bonferroni – simple, but too stringent, not recommended

$$mP_{(k)} \leq \alpha$$

Holm – a more powerful and less stringent version of Bonferroni (ok)

$$(m - k + 1)P_{(k)} \leq \alpha$$

```
p.adjust(pv,  
method="holm")
```

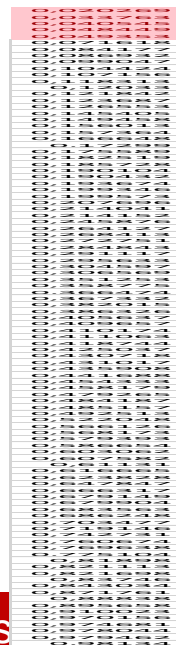
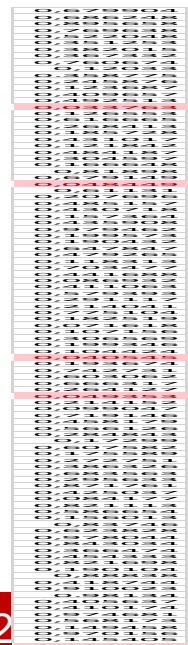

MULTIPLE EXPERIMENTS

Example: Random Data

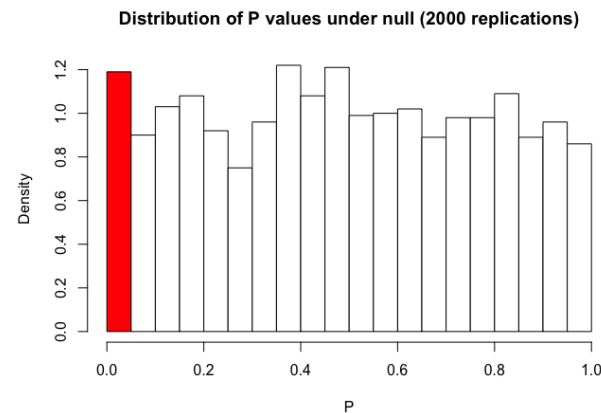
- ◆ Generate 6 columns of normal random variables (1000 points/candidates in each).
- ◆ Consider the first 3 columns as “treatment”, and the next 3 columns as “control”.
- ◆ Using t-test calculate p-values b/w “treatment” and “control” group. How many candidates have p-value < 0.05 ?
- ◆ Calculate FDR. How many candidates you have now?

Candidates.
5% are false

Same candidates.
Just sorted



Top 5%
selected
???



```
# create dataset
X = matrix(rnorm(1000*6),
           nrow=1000, ncol=6)

# test 1000 hypotheses
pv = 1

for (i in 1:nrow(X)) {
  res=t.test(X[i,1:3],
            X[i,4:6])

  pv[i]=res$p.value
}

# number of pv < 0.05
sum(pv<0.05)

# FDR adjustment
fdr = p.adjust(pv,"fdr")

sum(fdr<0.05)
```

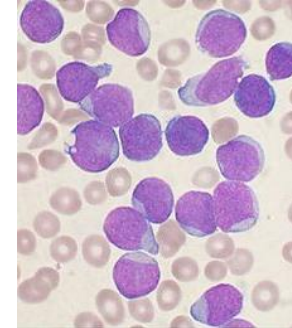
MULTIPLE EXPERIMENTS

Additional (optional) Task: Acute Lymphoblastic Leukemia

<http://edu.modas.lu/data>

all_data

Acute lymphoblastic leukemia (ALL), is a form of leukemia, or cancer of the white blood cells characterized by excess lymphoblasts.



all_data.xls contains the results of full-transcript profiling for ALL patients and healthy donors using Affymetrix microarrays. The data were downloaded from the ArrayExpress repository and normalized. The expression values in the table are in \log_2 scale.

Let us analyze these data:

- ◆ Calculate \log_2 -ratio (logFC) for each gene (simply subtract means)
- ◆ Calculate the p-value based on the t-test for each gene
- ◆ Perform the FDR-based **adjustment of the p-value**.

Calculate the number of up and down-regulated genes with $FDR < 0.01$

- ◆ How would you take into account logFC?
- ◆ Make volcano plot: $x = \log FC$, $y = -\log(\text{adj.pvalue})$

Example score:

$$\text{score} = -\log(\text{adj.p.value}) \cdot |\log FC|$$

FDR (adj. p-value) is a main measure. Other only help...

Thank you for your attention

