CENTRE DE RECHERCHE PUBLIC

Microarray Center

UNIVERSITÉ DU LUXEMBOURG

the Luxembourg Bioinformatics Network

STATISTICAL DATA ANALYSIS IN EXCEL

Part 2

Practical Aspects

Dr. Petr Nazarov

petr.nazarov@crp-sante.lu

14-06-2010

- **Hypotheses** *(theoretical)*
- **Unpaired t-test**
- **Paired t-test**
- **ANOVA**
- **Linear regression**
- **Multiple comparison**
- **Empirical confidence interval calculation**
- ***Goodness of fit and inependence*** *(optional)*

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

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: $\mu \leq$ 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: $\mu_1 \neq \mu_2$

crp SANTÉ
CENTRE DE RECHERCHE PUBLIC

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

H_0: $\mu \geq 3$

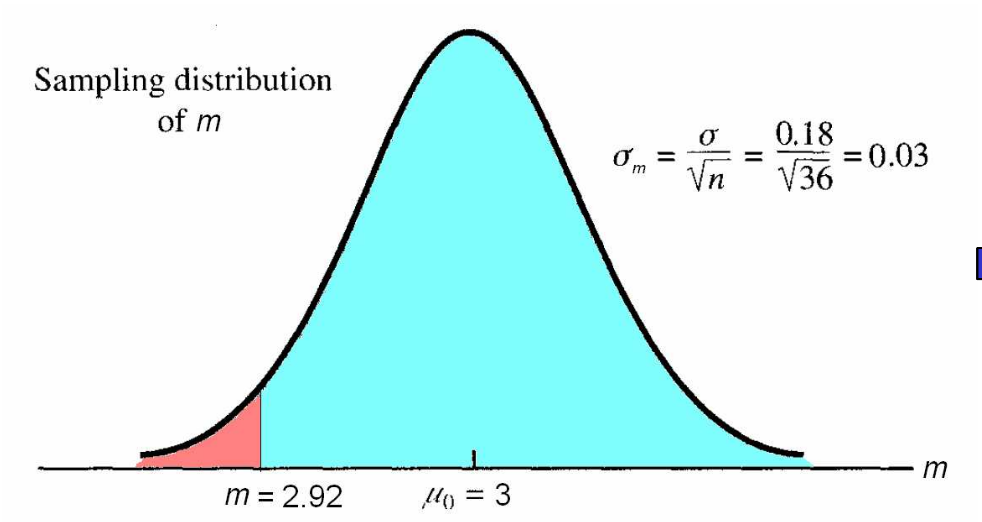
H_a: $\mu < 3$

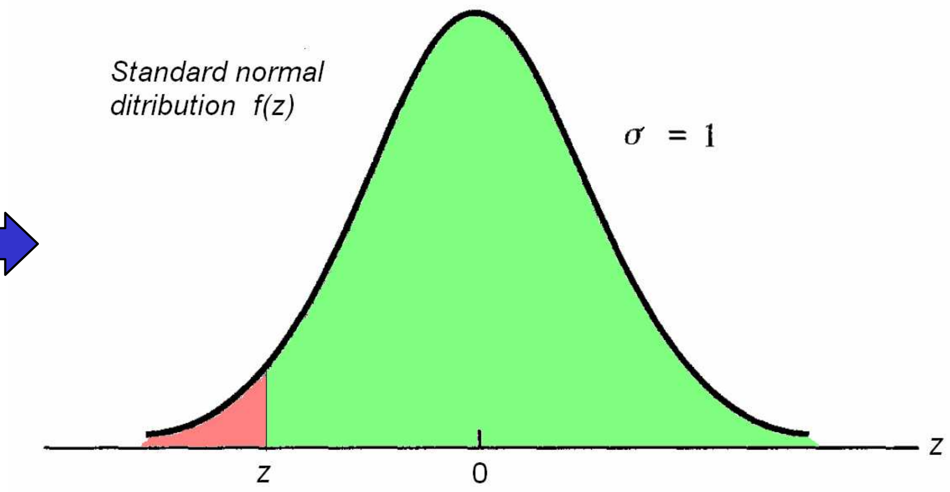
Assume that we have obtained experimentally m=2.92. Is it significant?

Step 1. Introduce the test statistics

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

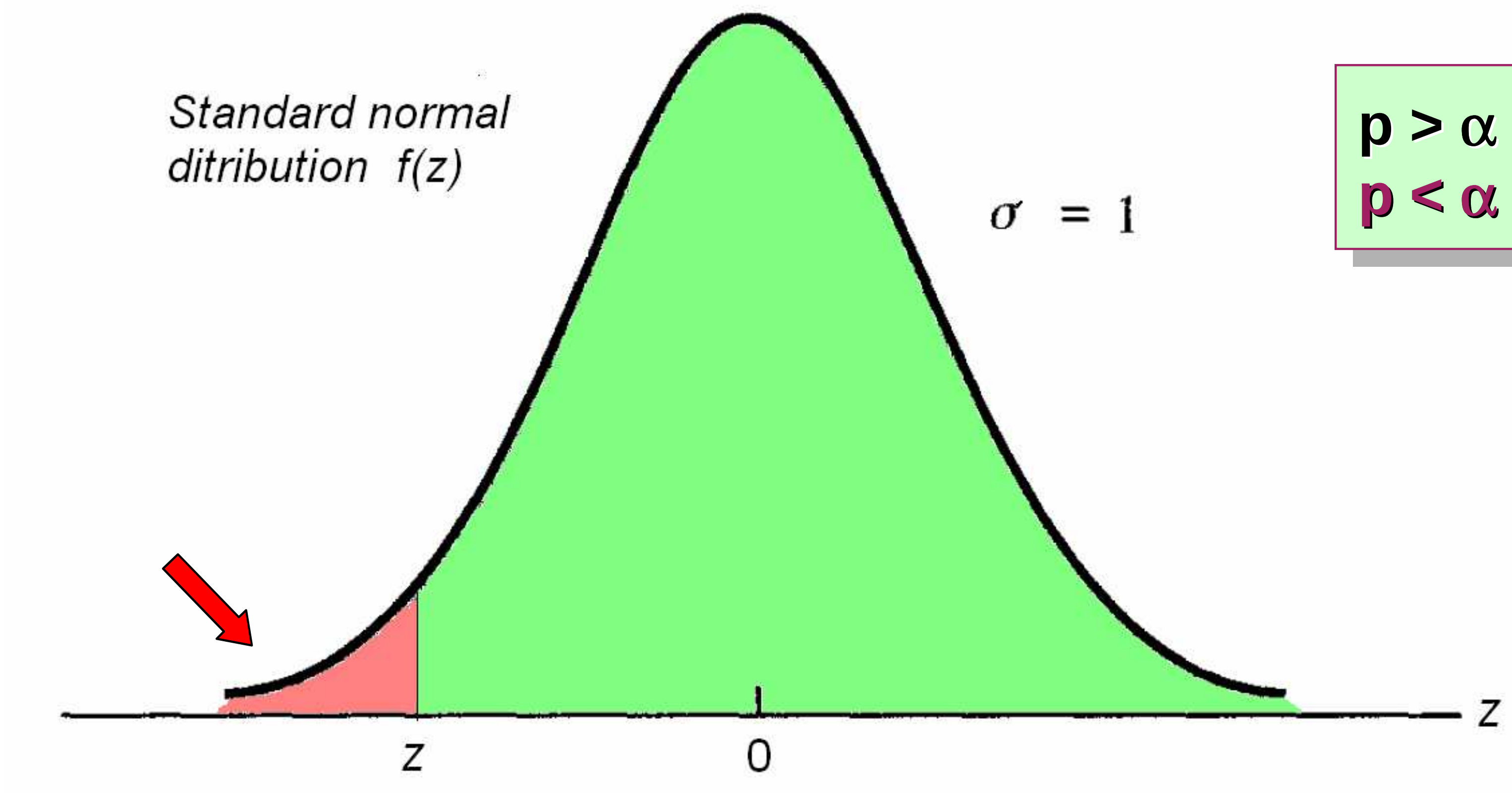
$$z = \frac{m - \mu_0}{\sigma / \sqrt{n}}$$





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



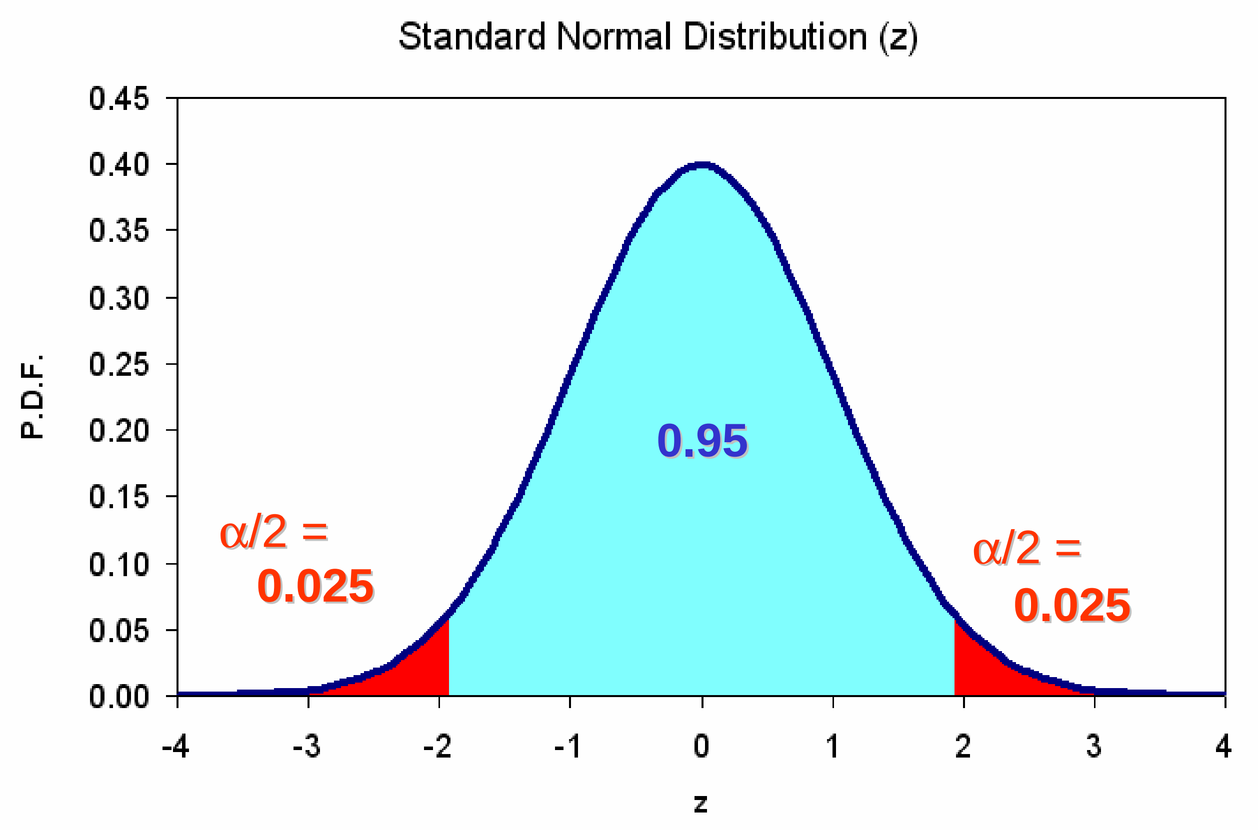
$p > \alpha \Rightarrow H_0$
$p < \alpha \Rightarrow H_a$

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: $\mu \neq \mu_0$



if σ in unknown:

σ → *s*

z → *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 $t \geq t_{\alpha/2}$

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.

Data

A priori information about μ and μ_0

Hypothesis testing

Two Tails

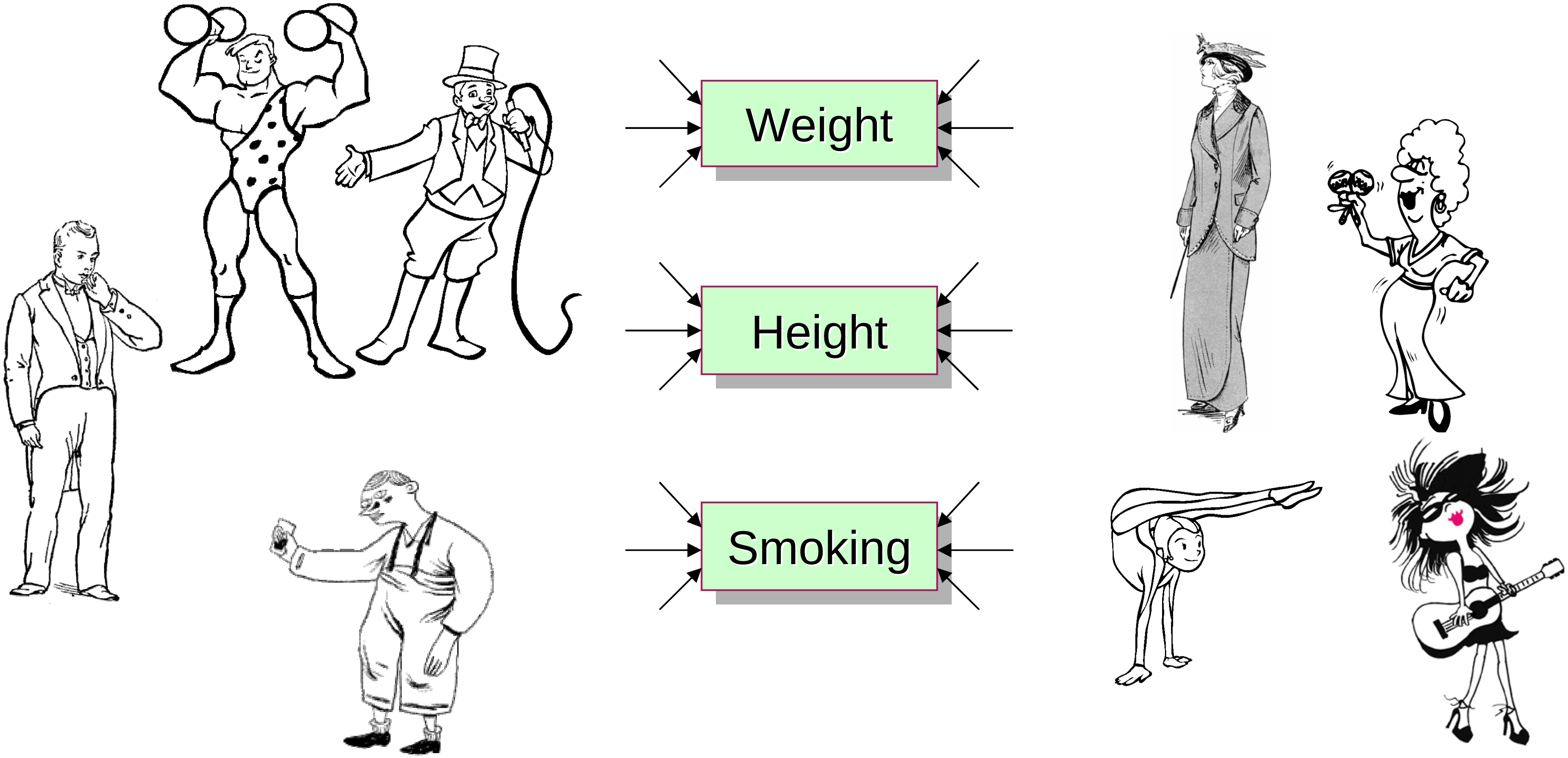
One Tail

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

Unpaired t-test

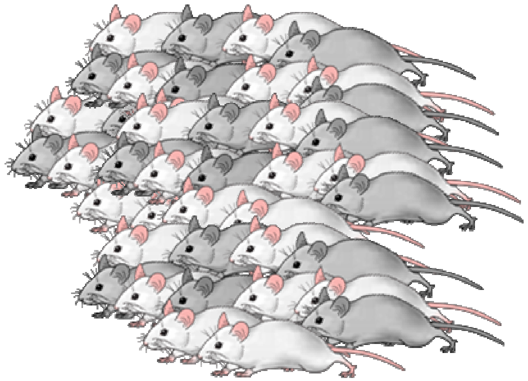
Independent samples

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

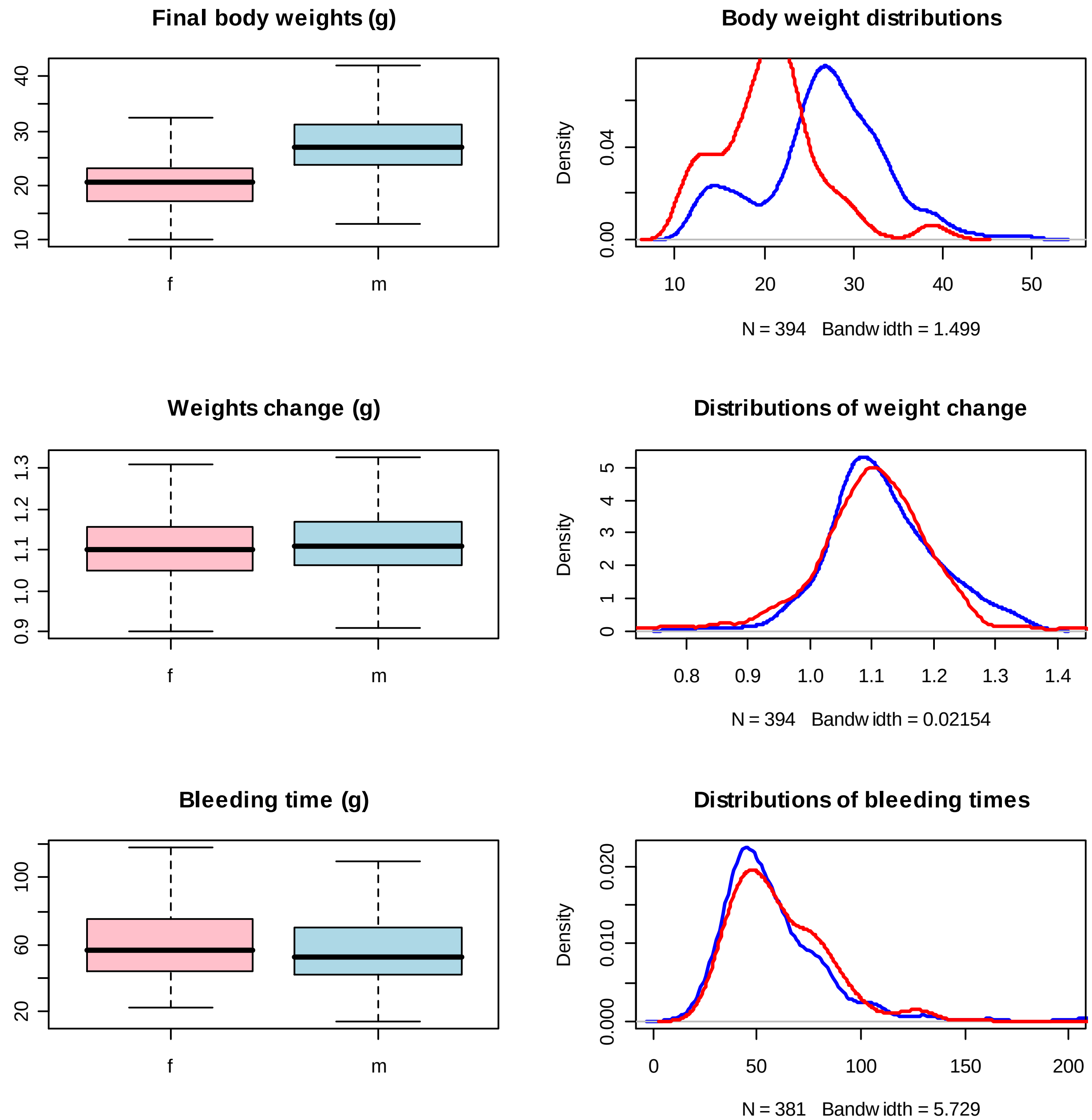
mice.xls



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?



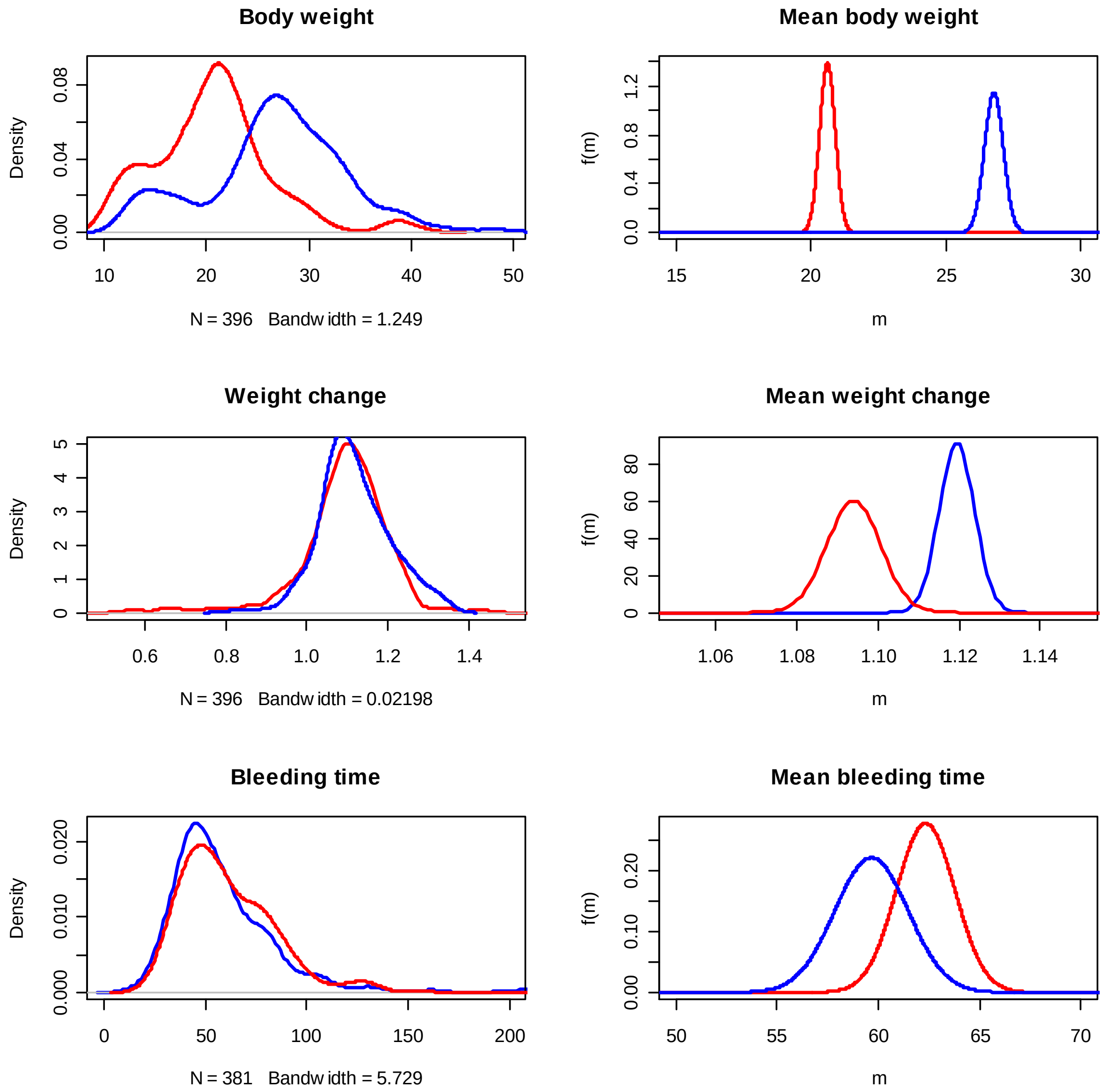
UNPAIRED t-TEST

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?



mice.xls

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

```
= TTEST (array1, array2, 2, 3)
```

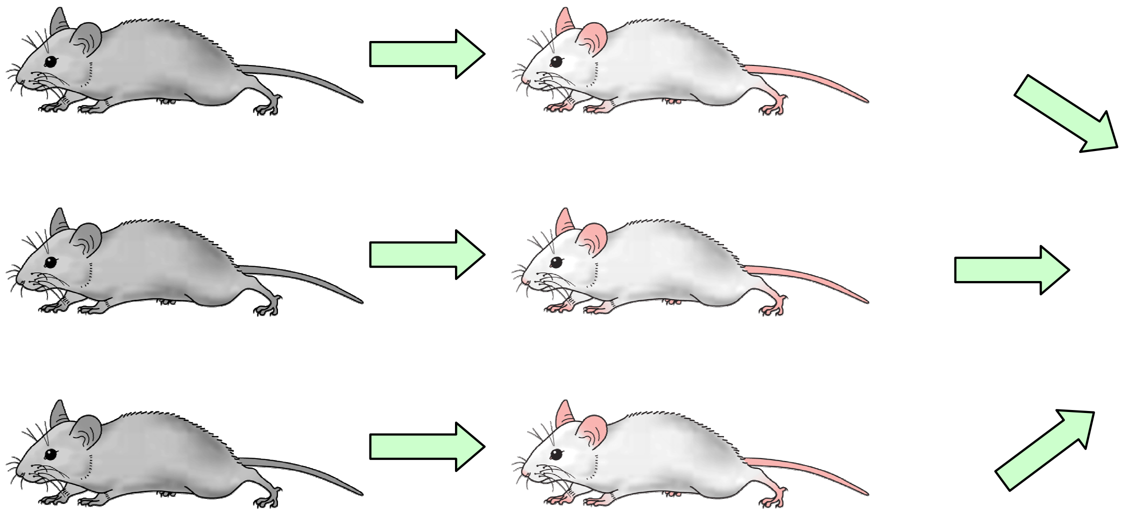
parameter	pval	female	male
Starting age	0.165799	65.90	66.52
Ending age	0.223033	113.91	114.61
Starting weight	5.48E-34	18.91	23.86
Ending weight	8.98E-38	20.62	26.78
Weight change	0.001405	1.09	1.12
Bleeding time	0.248716	62.34	59.67
Ionized Ca in blood	0.271336	1.23	1.24
Blood pH	0.009593	7.21	7.19
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

CENTRE DE RECHERCHE PUBLIC

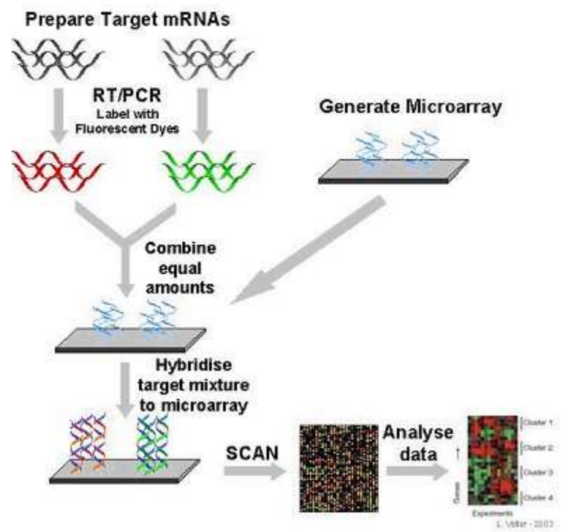
Paired t-test

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

- `= TTEST (array1, array2, 2, 3)`

Paired test

- `= TTEST (array1, array2, 2, 1)`

Test	p-value
unpaired	0.414662
paired	0.014506

ANOVA

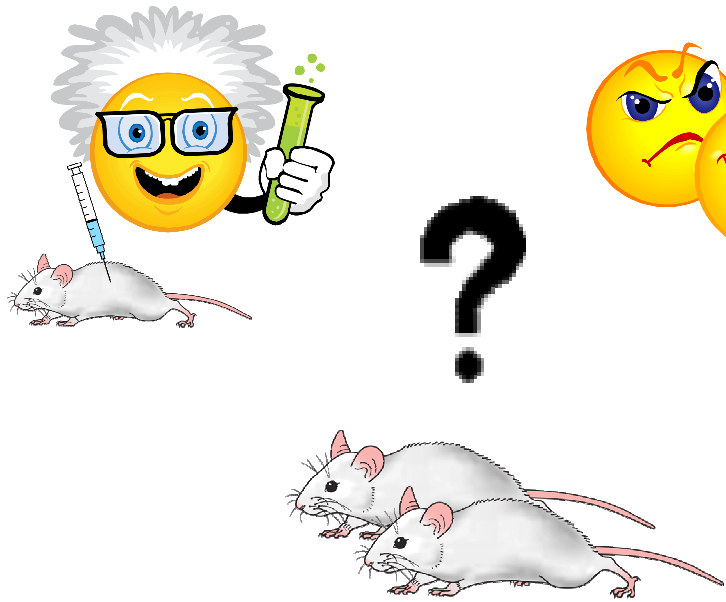
Means for more than 2 populations
We have measurements for 5 conditions. Are the means for these conditions equal?

Validation of the effects
We assume that we have several factors affecting our data. Which factors are more significant? Which can be neglected?

If we would use pairwise comparisons, what will be the probability of getting error?

Number of comparisons: $C_2^5 = \frac{5!}{2!3!} = 10$

Probability of an error: $1 - (0.95)^{10} = 0.4$



http://easylink.playstream.com/affymetrix/ambsymposium/partek_08.wvx

As part of a long-term study of individuals 65 years of age or older, sociologists and physicians at the Wentworth Medical Center in upstate New York investigated the relationship between geographic location and depression. A sample of 60 individuals, all in reasonably good health, was selected; 20 individuals were residents of Florida, 20 were residents of New York, and 20 were residents of North Carolina. Each of the individuals sampled was given a standardized test to measure depression. The data collected follow; higher test scores indicate higher levels of depression.

Q: Is the depression level same in all 3 locations?

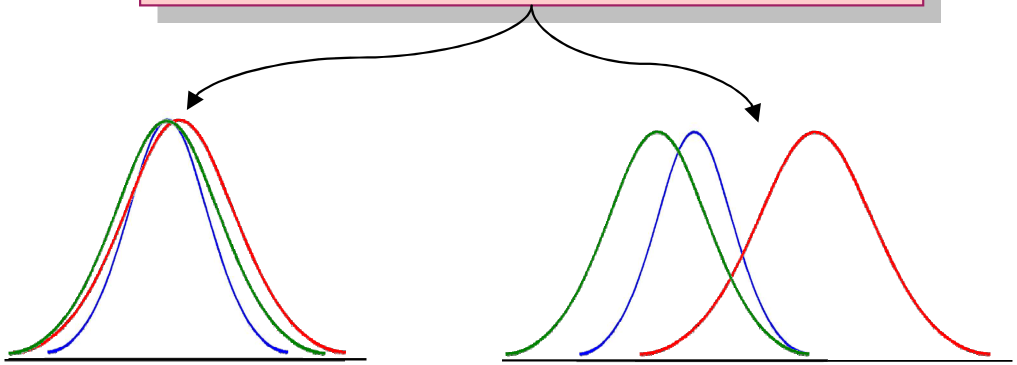
`depression.xls`

1. Good health respondents

Florida	New York	N. Carolina
3	8	10
7	11	7
7	9	3
3	7	5
8	8	11
8	7	8
...	...	...

H_0: $\mu_1 = \mu_2 = \mu_3$

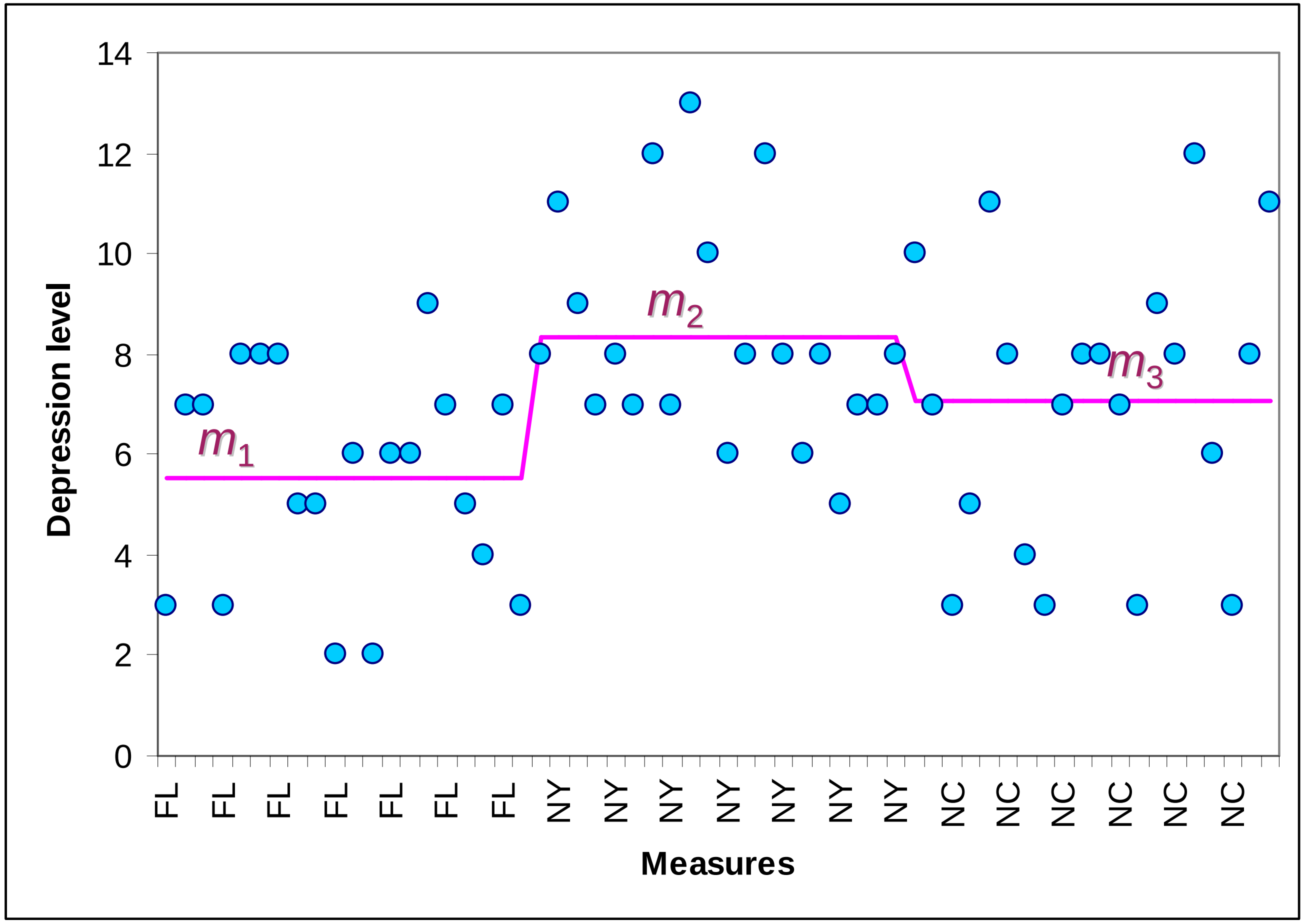
H_a: not all 3 means are equal



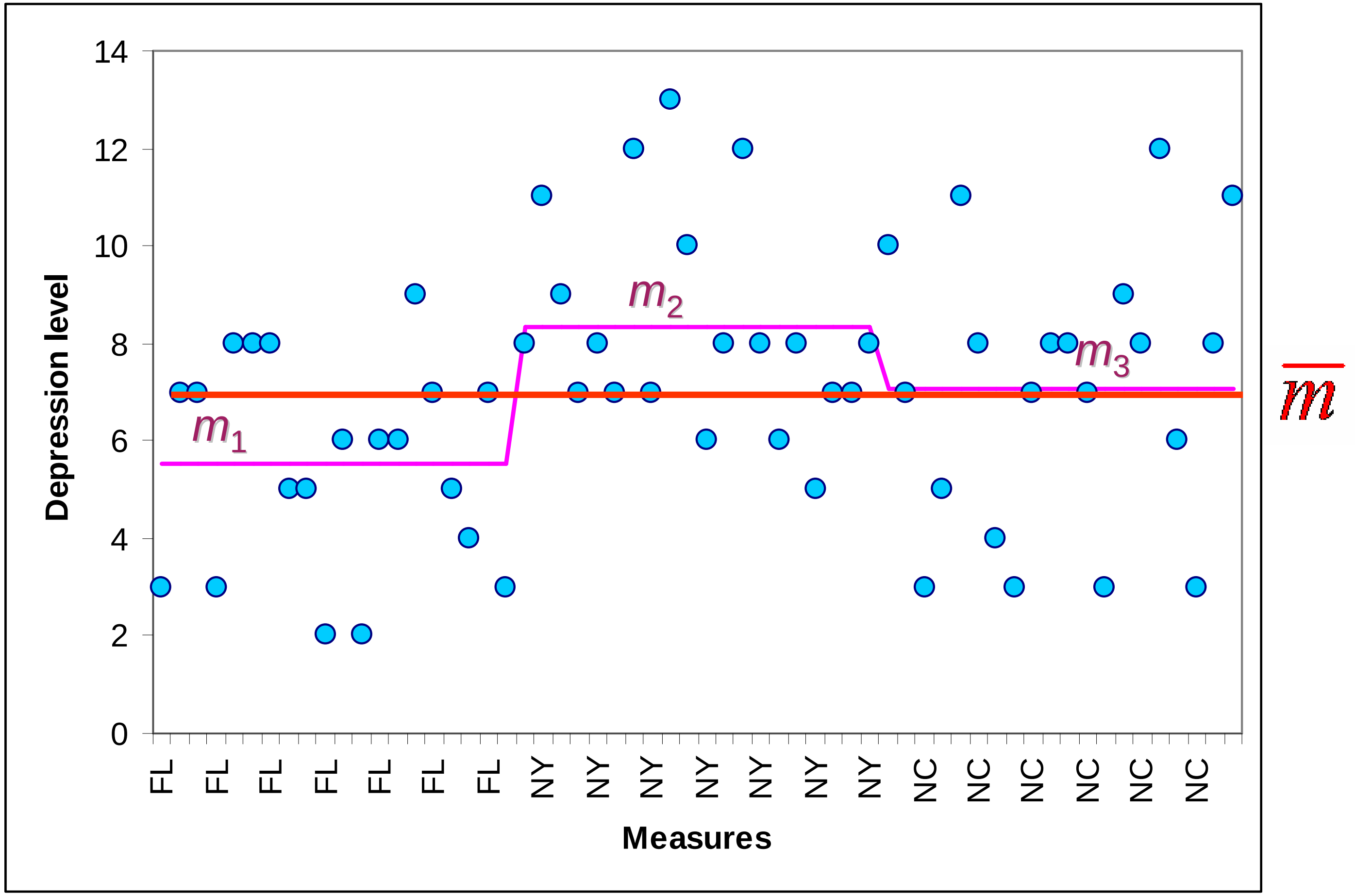
Meaning

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: not all 3 means are equal



Example



$$SST = SSTR + SSE$$

ANOVA table

A table used to summarize the analysis of variance computations and results. It contains columns showing the source of variation, the sum of squares, the degrees of freedom, the mean square, and the *F* value(s).

In Excel use:

- Tools → Data Analysis → ANOVA Single Factor

depression.xls

Let's perform for dataset 1: "good health"

SSTR

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	78.53333	2	39.26667	6.773188	0.002296	3.158843
Within Groups	330.45	57	5.797368			
Total	408.9833	59				

SSE

Factor
Another word for the independent variable of interest.

Factorial experiment
An experimental design that allows statistical conclusions about two or more factors.

Treatments
Different levels of a factor.

depression.xls

Factor 1: Health → good health, bad health

Factor 2: Location → Florida, New York, North Carolina

$$\text{Depression} = \mu + \text{Health} + \text{Location} + \text{Health}\times\text{Location} + \varepsilon$$

Interaction
The effect produced when the levels of one factor interact with the levels of another factor in influencing the response variable.

2-factor ANOVA with *r* Replicates: Example

depression.xls

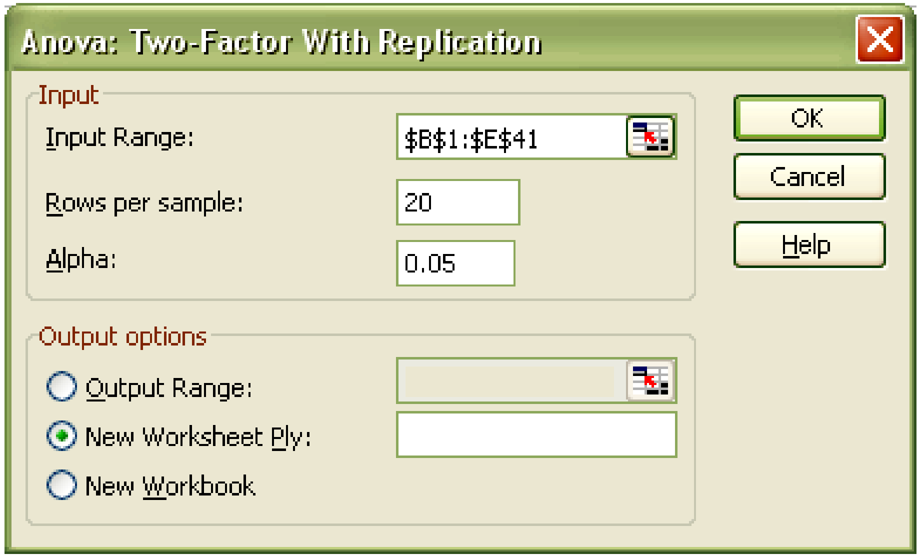
Factor 1: Health

Factor 2: Location

1. Reorder the data into format understandable for Excel

	Florida	New York	North Carolina
Good health	3	8	10
	7	11	7
	7	9	3
	3	7	5
	...	...	...
	7	7	8
	3	8	11
bad health	**13**	**14**	**10**
	12	**9**	**12**
	17	**15**	**15**
	17	**12**	**18**
	...	...	...
	11	**13**	**13**
	17	**11**	**11**

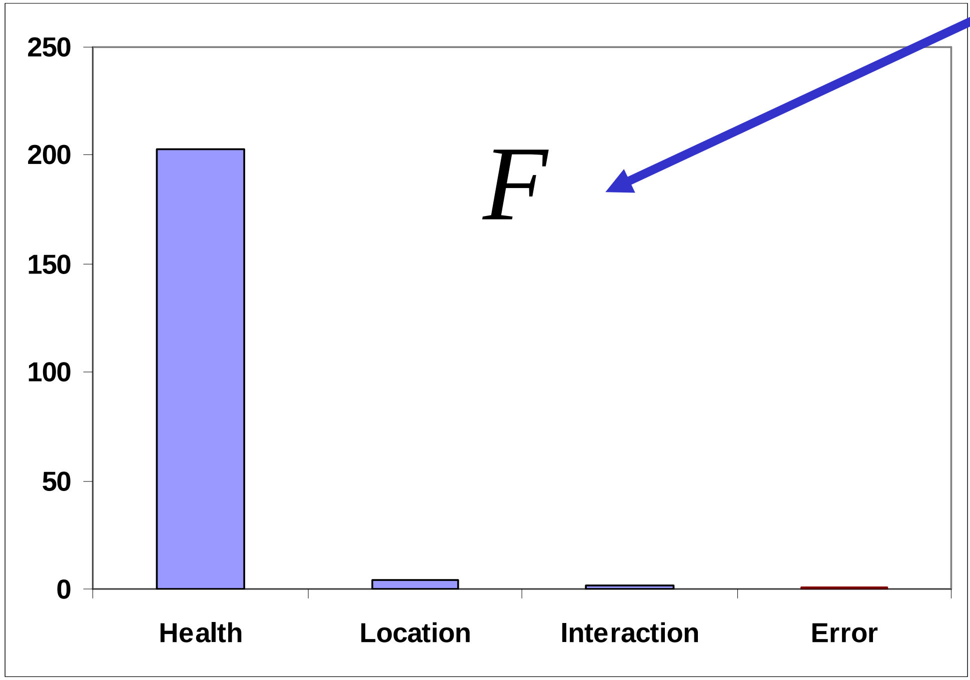
2. Use Tools → Data Analysis → ANOVA: Two-factor with replicates

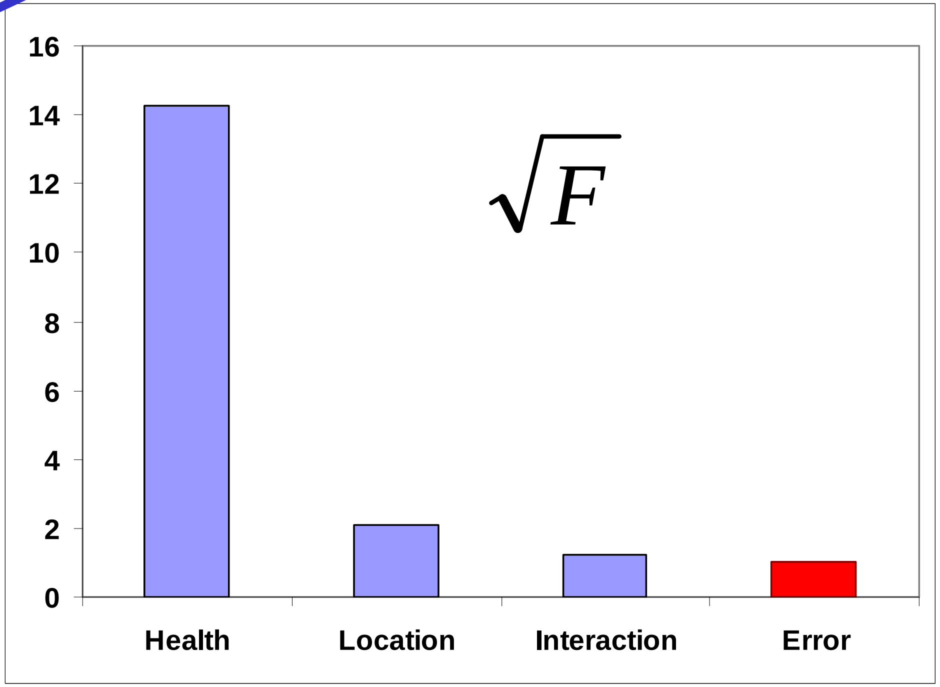


2-factor ANOVA with *r* Replicates: Example

ANOVA

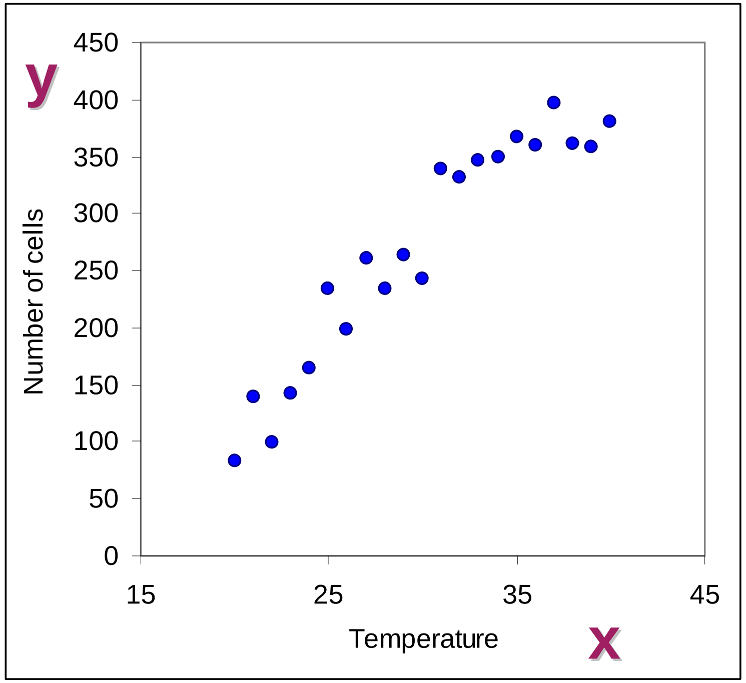
	Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Health	Sample	1748.033	1	1748.033	203.094	4.4E-27	3.92433
Location	Columns	73.85	2	36.925	4.290104	0.015981	3.075853
Interaction	Interaction	26.11667	2	13.05833	1.517173	0.223726	3.075853
Error	Within	981.2	114	8.607018			
	Total	2829.2	119				





Linear Regression

Temperature	Cell Number
20	83
21	139
22	99
23	143
24	164
25	233
26	198
27	261
28	235
29	264
30	243
31	339
32	331
33	346
34	350
35	368
36	360
37	397
38	361
39	358
40	381



Cells are grown under different temperature conditions from 20° to 40°. A researched would like to find a dependency between T and cell number.

Dependent variable

The variable that is being predicted or explained. It is denoted by ***y*****.**

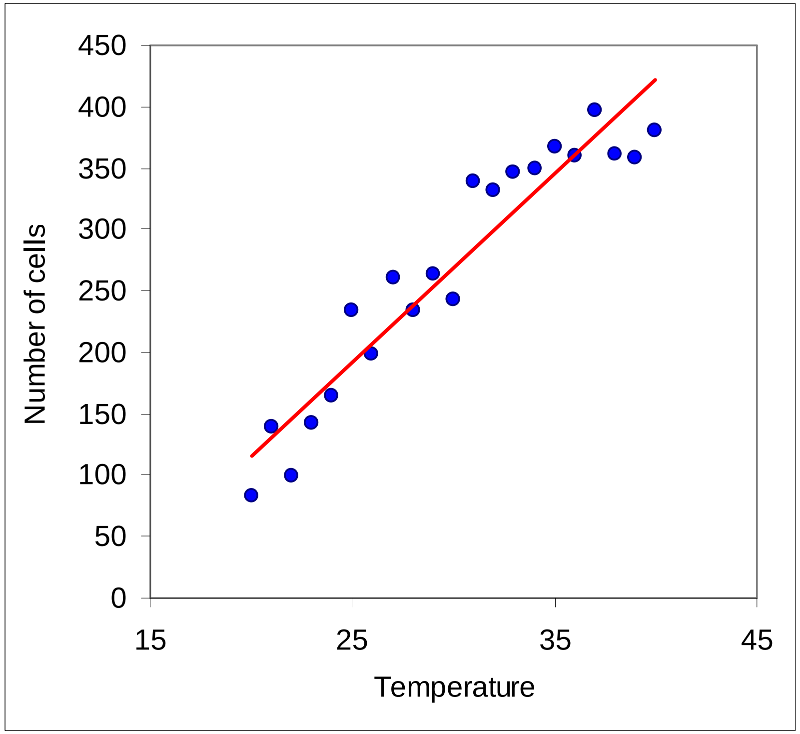
Independent variable

The variable that is doing the predicting or explaining. It is denoted by ***x***.

Simple linear regression

Regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line.

- Building a *regression* means finding and tuning the model to explain the behaviour of the data



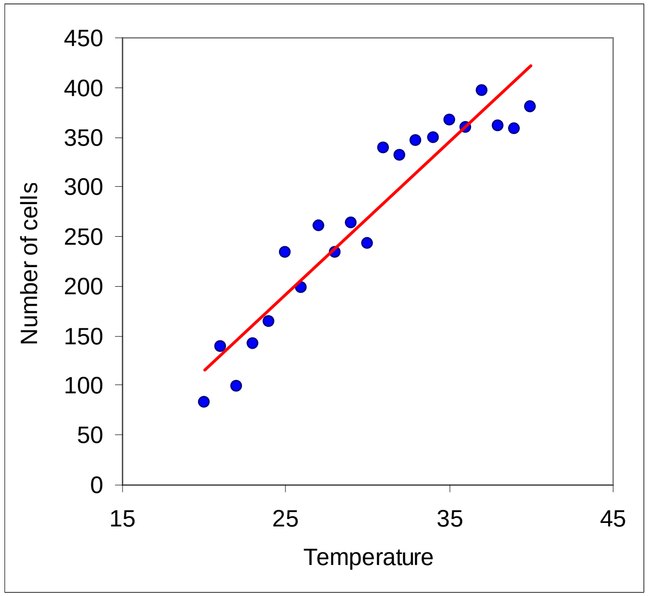
Regression model

The equation describing how y is related to x and an error term; in simple linear regression, the regression model is $y = \beta_0 + \beta_1 x + \varepsilon$

Regression equation

The equation that describes how the mean or expected value of the dependent variable is related to the independent variable; in simple linear regression,

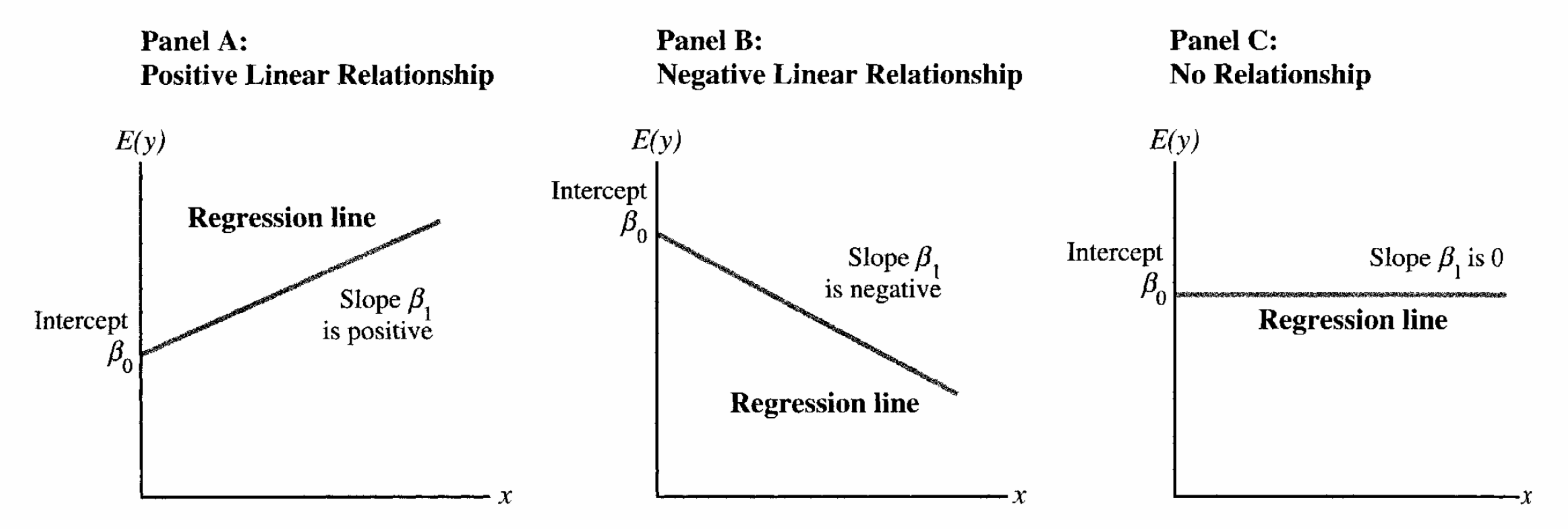
$E(y) = \beta_0 + \beta_1 x$



- Model for a simple linear regression:

$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$

$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$



Estimated regression equation
The estimate of the regression equation developed from sample data by using the least squares method. For simple linear regression, the estimated regression equation is ***y* = *b*$_0$ + *b*$_1$*x***

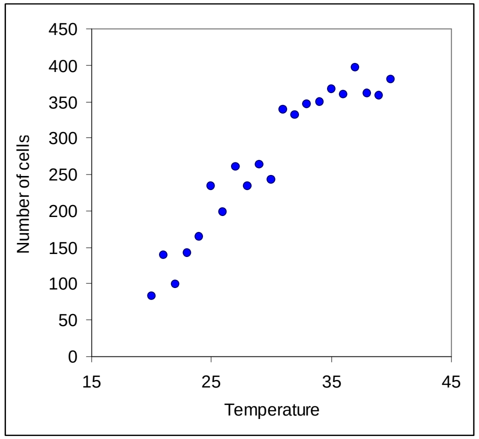
$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$

$$\hat{y}(x) = b_1 x + b_0$$

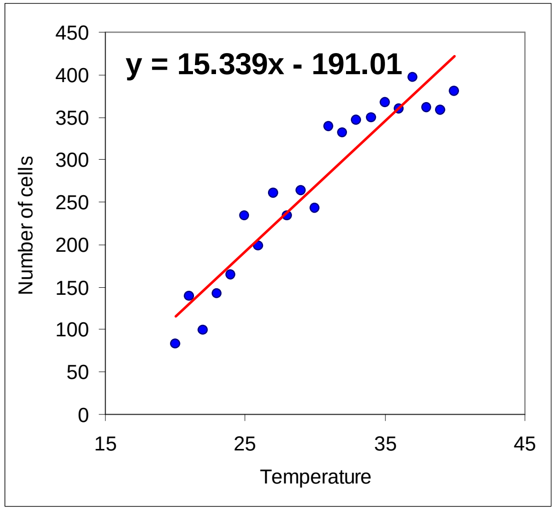
$$E\left[y(x)\right] = b_1 x + b_0$$

cells.xls

1. Make a scatter plot for the data.



2. Right click to "Add Trendline". Show equation.



Coefficient of Determination

Sum squares due to **error**

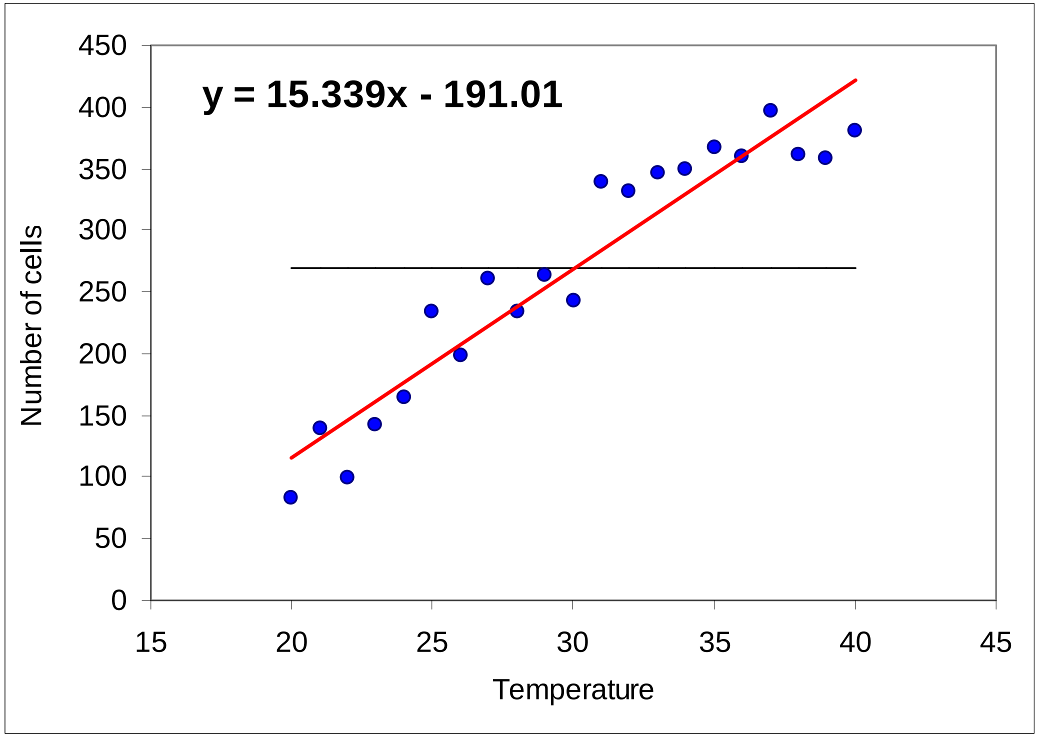
$$SSE = \sum \left(y_i - \hat{y}_i \right)^2$$

Sum squares total

$$SST = \sum \left(y_i - \bar{y} \right)^2$$

Sum squares due to regression

$$SSR = \sum \left(\hat{y}_i - \bar{y} \right)^2$$



The Main Equation

$$\boxed{SST = SSR + SSE}$$

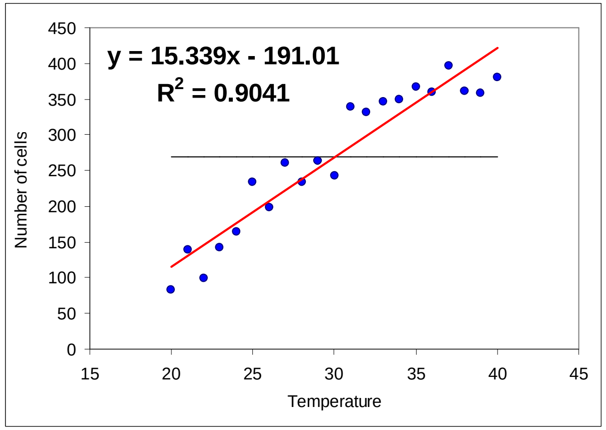
Coefficient of Determination

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SST = \sum (y_i - \bar{y})^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

$$SST = SSR + SSE$$



Coefficient of determination

A measure of the goodness of fit of the estimated regression equation. It can be interpreted as the proportion of the variability in the dependent variable *y* that is explained by the estimated regression equation.

$$R^2 = \frac{SSR}{SST}$$

Correlation coefficient

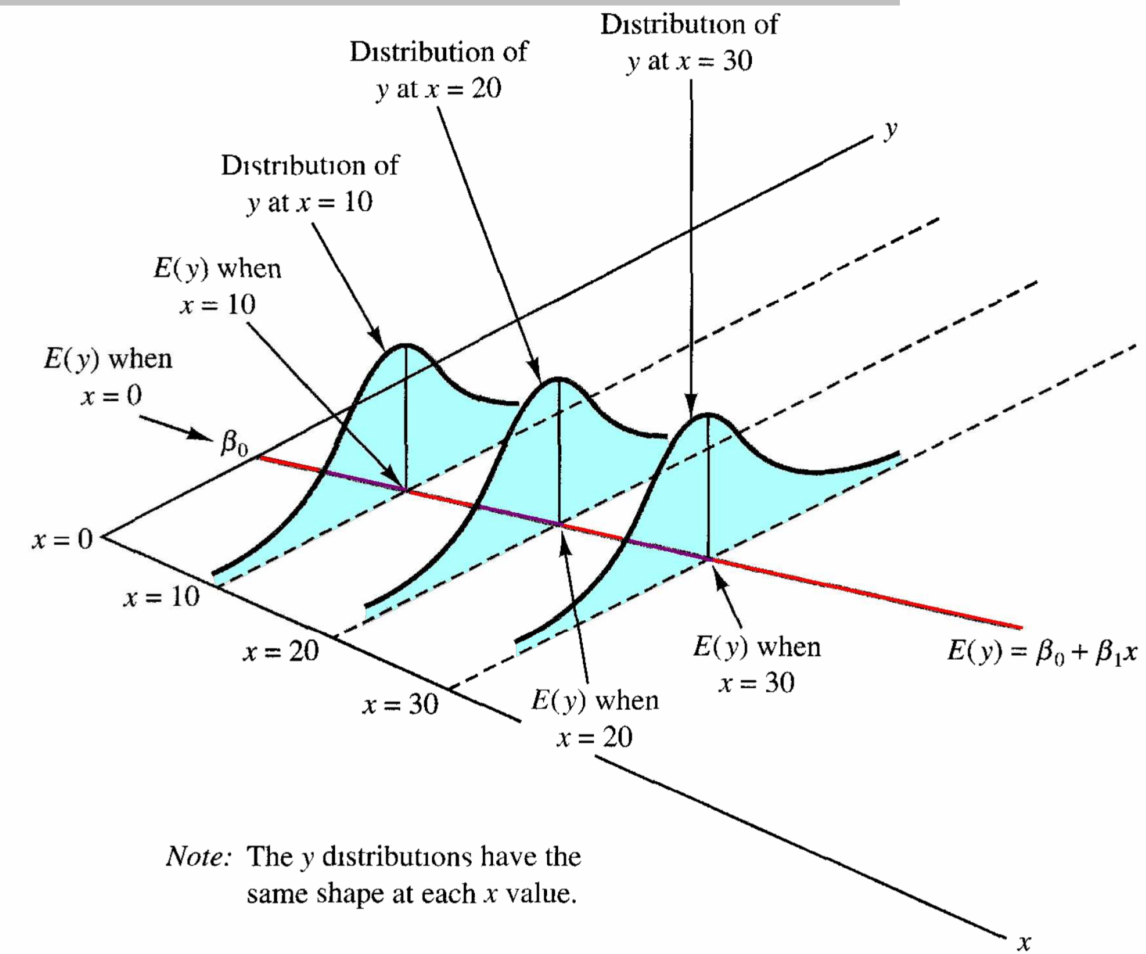
A measure of the strength of the linear relationship between two variables (previously discussed in Lecture 1).

$$r = \mathrm{sign}(b_1)\sqrt{R^2}$$

Assumptions for Simple Linear Regression

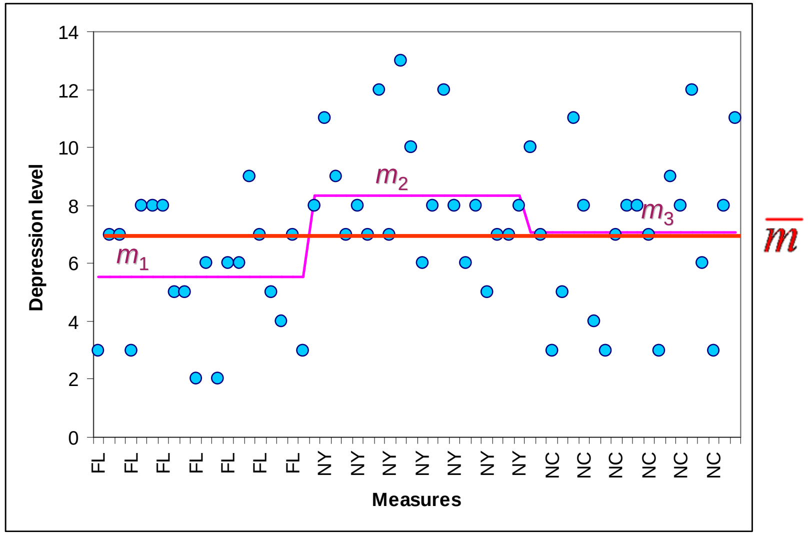
1. The error term ε is a random variable with 0 mean, i.e. E[ε]=0
2. The variance of ε, denoted by σ^2, is the same for all values of *x*
3. The values of ε are independent
3. The term ε is a normally distributed variable

$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$

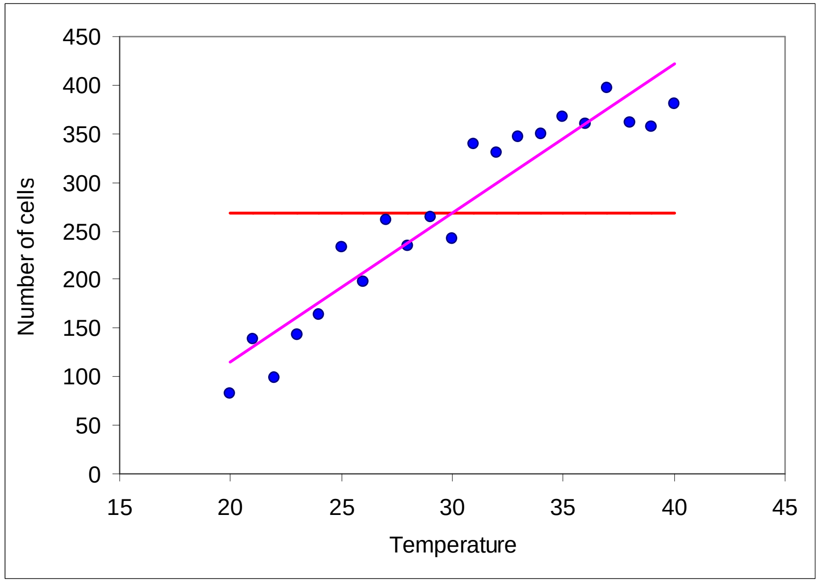


Note: The y distributions have the same shape at each x value.

ANOVA and Regression: Testing for Significance



$$SST = SSTR + SSE$$



$$SST = SSR + SSE$$

H_0: $\beta_1 = 0$ *insignificant*

H_a: $\beta_1 \neq 0$

Example

cells.xls

1. Calculate manually b_1 and b_0

Intercept	b0=	-191.008119
Slope	b1=	15.3385723

In Excel use the function:

- `= INTERCEPT(y,x)`
- `= SLOPE(y,x)`

2. Let's do it automatically Tools → Data Analysis → Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950842308
R Square	0.904101095
Adjusted R Square	0.899053784
Standard Error	31.80180903
Observations	21

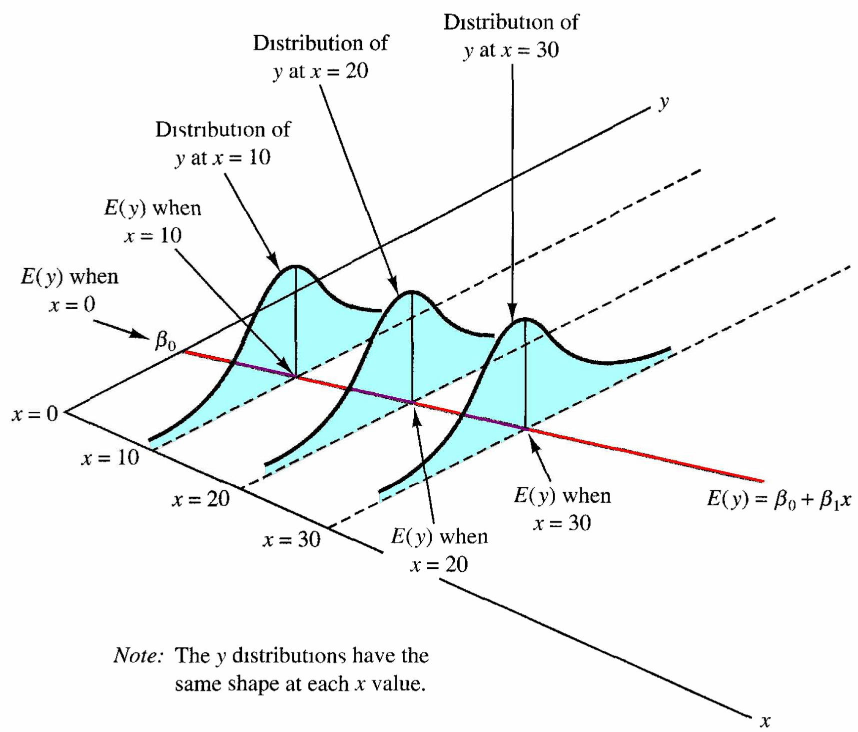
ANOVA

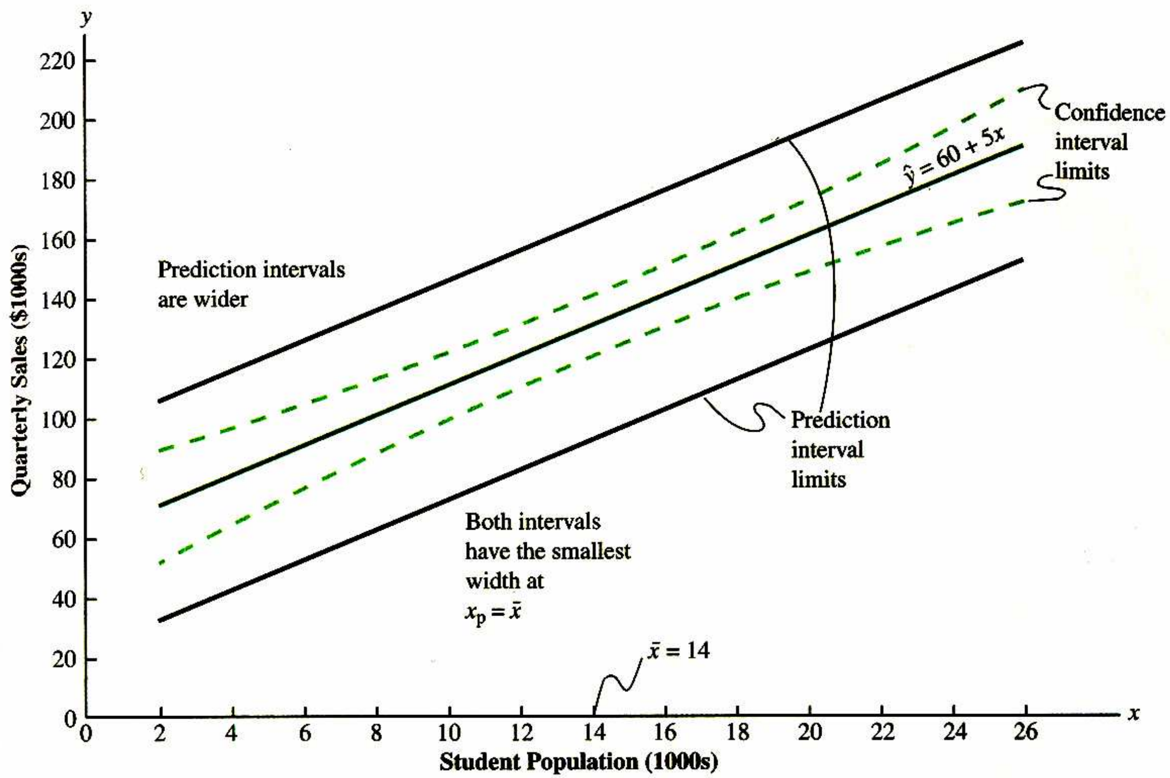
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	181159.2853	181159.3	179.1253	4.01609E-11
Residual	19	19215.7461	1011.355		
Total	20	200375.0314			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-191.0081194	35.07510626	-5.445689	2.97E-05	-264.4211603	-117.5950784	-264.4211603	-117.5950784
X Variable 1	15.33857226	1.146057646	13.38377	4.02E-11	12.93984605	17.73729848	12.93984605	17.73729848

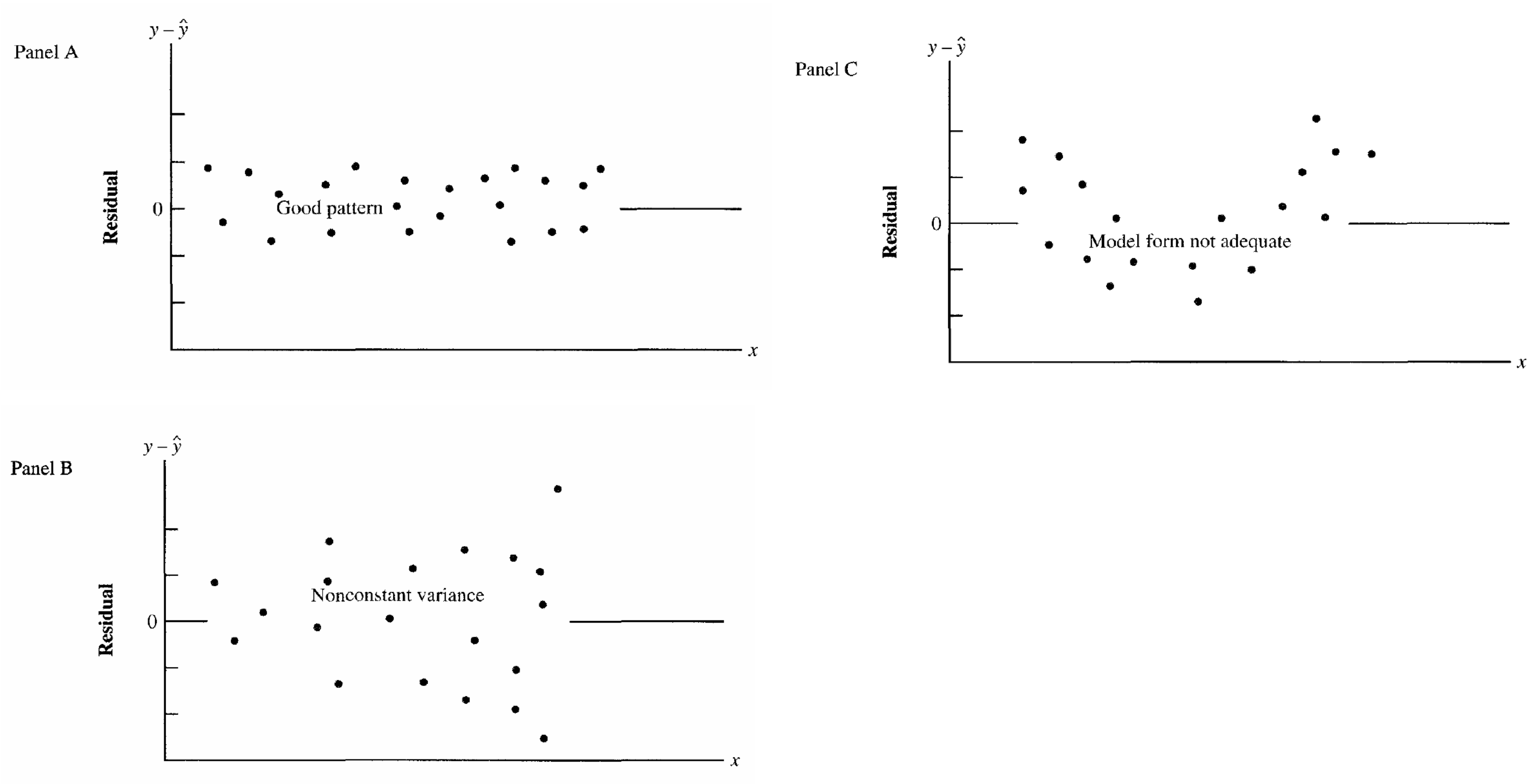
Confidence interval
The interval estimate of the mean value of y for a given value of x.

Prediction interval
The interval estimate of an individual value of y for a given value of x.





Residuals



Panel A
$y - \hat{y}$
Residual
0
Good pattern
x
Panel C
$y - \hat{y}$
Residual
0
Model form not adequate
x
Panel B
$y - \hat{y}$
Residual
0
Nonconstant variance
x

Correction for Multiple Comparison

Please download the data from edu.sablab.net/data/xls

`all_data.xls`

MULTIPLE EXPERIMENTS

Correct Results and Errors

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

Probability of an error in a multiple test:

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

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

		Population Condition		
		H_0 is TRUE	H_0 is FALSE	Total
Conclusion	Accept H_0 (non-significant)	***U***	***T***	$m - R$
	Reject H_0 (significant)	***V***	***S***	R
	Total	m_0	$m - m_0$	$\boldsymbol{m}$

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

Assume we need to perform k = 100 comparisons, and select maximum FDR = α = 0.05

Independent tests

The Simes procedure ensures that its expected value $\mathrm{E}\left[\frac{V}{V+S}\right]$ is less than a given α (Benjamini and Hochberg 1995). This procedure is valid when the m tests are independent. Let $H_1 \ldots H_m$ be the null hypotheses and $P_1 \ldots P_m$ their corresponding p-values. Order these values in increasing order and denote them by $P_{(1)} \ldots P_{(m)}$. 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 $i = 1, \ldots, k$.

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.

Assume we need to perform k = 100 comparisons, and select maximum FDR = α = 0.05

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

Expected value for FDR < α if

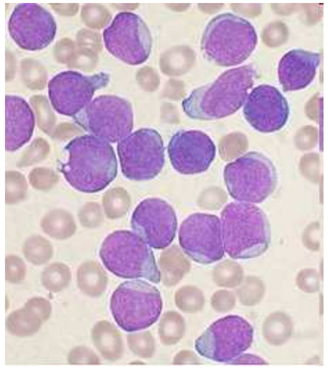
$$P_{(k)} \leq \frac{k}{m}\alpha$$

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

Example: Acute Lymphoblastic Leukemia

all_data.xls

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



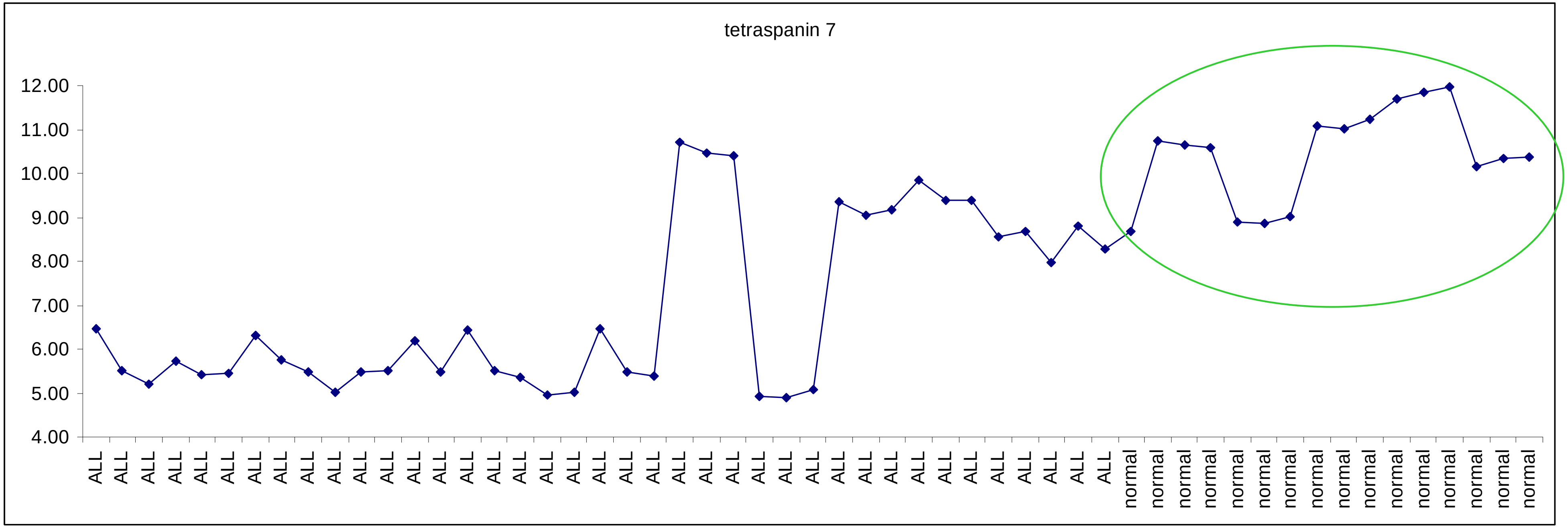
all_data.xls contains the results of full-trancript profiling for ALL patients and healthy donors using Affymetrix microarrays. The data were downloaded from ArrayExpress repository and normalized. The expression values in the table are in $\log_2$ scale.

Let us analyze these data:

- Calculate log-ratio (logFC) for each gene
- Calculate the p-value based on 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?

Example score: $$score = -\log(adj.p.value) \cdot |logFC|$$



look for "tetraspanin 7" + leukemia in google ☺

Results are never perfect…

Empirical Interval Estimation 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.

$$x \pm y \rightarrow Normal\ distribution$$

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

$$\sigma^2_{x \pm y} = \sigma^2_x + \sigma^2_y$$

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.

$$if \quad x_1, \ldots, x_k \rightarrow Normal\ distribution$$

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

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

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

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

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

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

INTERVAL ESTIMATIONS FOR RANDOM FUNCTIONS

Practical Approach

Two rates where measured for a PCR experiment: experimental value (X) and control (Y). 5 replicates where performed for each.

From previous experience we know that the error between replicates is normally distributed.

Q1: provide an interval estimation for the fold change X/Y (α=0.05)

Q2: provide an interval estimation for the log fold change $\log_2(X/Y)$

#	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

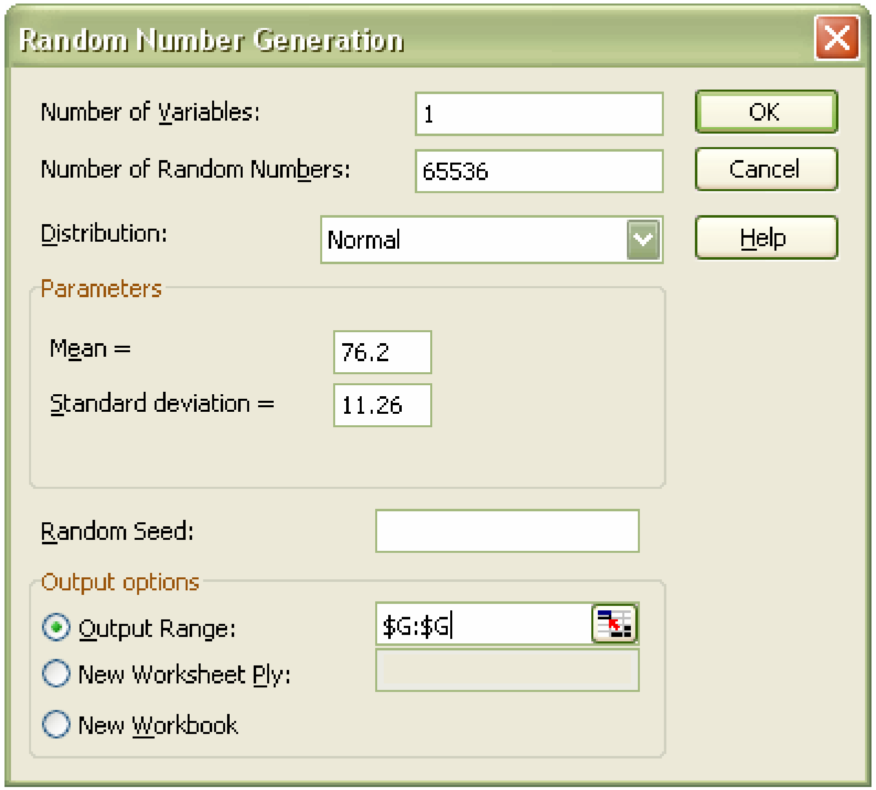
Let us use a ***numerical simulation…***

1. Generate 2 sets of 65536 normal random variable with means and standard deviations corresponding to ones of experimental and control set.

Mean	226.2	76.2
StDev	21.39	11.26

In Excel go: Tools → Data Analysis:

- **Random Number Generation**



If you do not have Data Analysis tool – approximate normal distribution by sum of uniform:

$$N(x, m_x, \sigma_x) = m_x + \sigma_x \left(\sum_{i=1}^{12} U(x_i) - 6 \right)$$

- **= RAND()** ← $U(x)$

INTERVAL ESTIMATIONS FOR RANDOM FUNCTIONS

Practical Approach

1. Generate 2 sets of 65536 normal random variable with means and standard deviations corresponding to ones of experimental and control set.

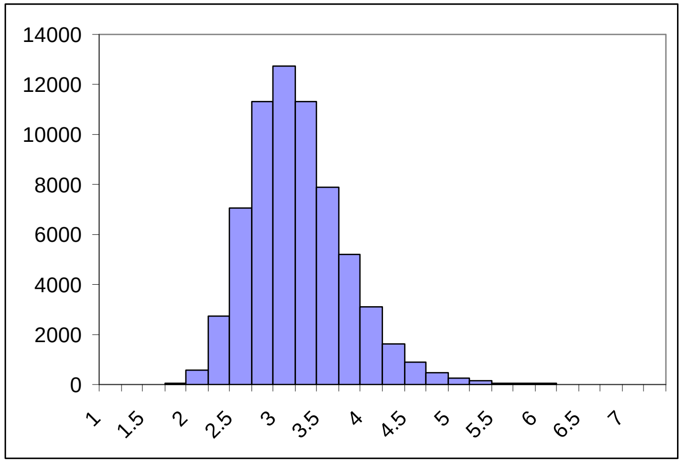
Mean	226.2	76.2
StDev	21.39	11.26

sim.m	226.088799	76.2823
sim.s	21.379652	11.2885

2. Build the target function. For Q1 build X/Y

X/Y.m	3.03289298
X/Y.s	0.566865
min	-8.14098141
max	7.72162205

3. Study the target function. Calculate summary, build histogram.



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

In Excel use function

- **=PERCENTILE(data,0.025)**

$X/Y \in [\ 2.13,\ \ 4.33\ ]$

What was a “mistake” in the previous case?

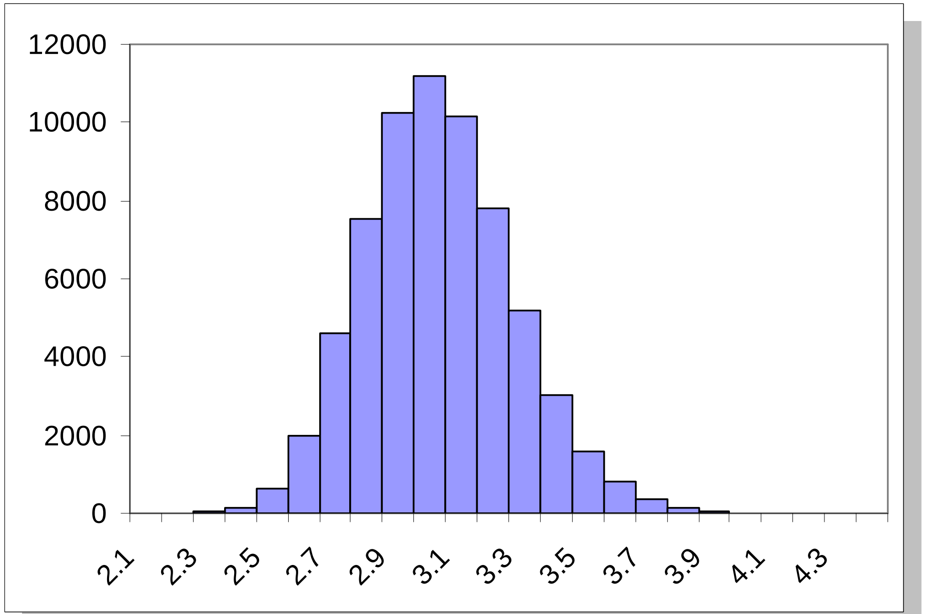
$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

There we spoke about **prediction interval** of X/Y. Now let’s produce the **interval estimation for mean X/Y**

Mean	226.2	76.2
StDev	**9.57**	**5.03**

X/Y.m	2.98047943
X/Y.s	0.23616818
min	2.01556098
max	4.31131109

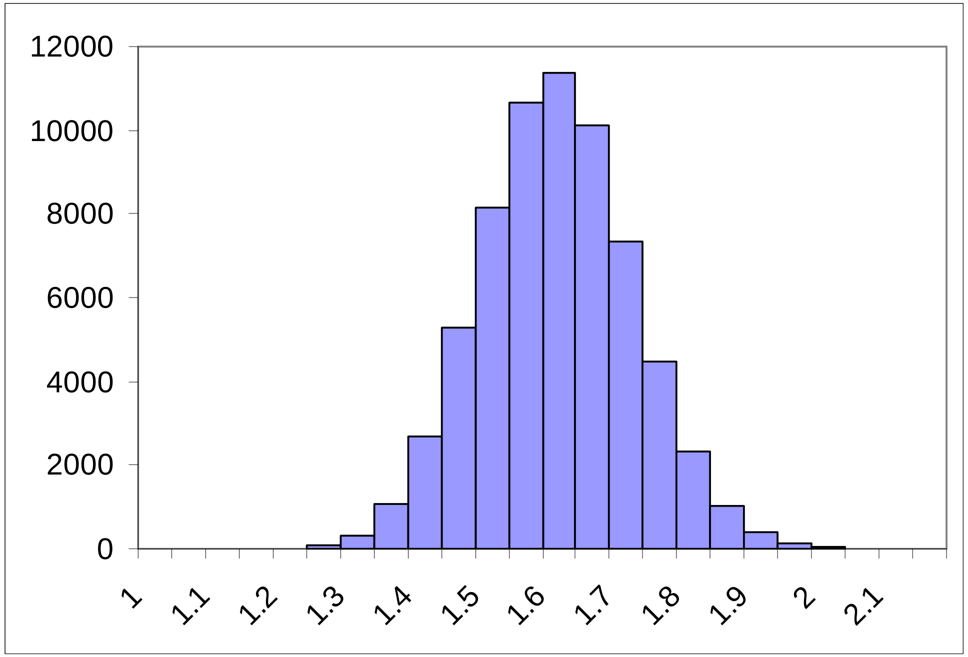
E[X/Y] ∈ [2.55, 3.48]



Q2: provide an interval estimation for the log fold change log2(X/Y)

Mean	1.571052
Standard Devi	0.113705

E[log(X/Y)] ∈ [1.35, 1.80]



	Simulation	Normal
2.50%	1.3546	1.3482
97.50%	1.7998	1.7939

Goodness of Fit and Independence

Multinomial population
A population in which each element is assigned to one and only one of several categories. The multinomial distribution extends the binomial distribution from two to three or more outcomes.

Contingency table = Crosstabulation
Contingency tables or crosstabulations are used to record, summarize and analyze the relationship between two or more categorical (usually) variables.

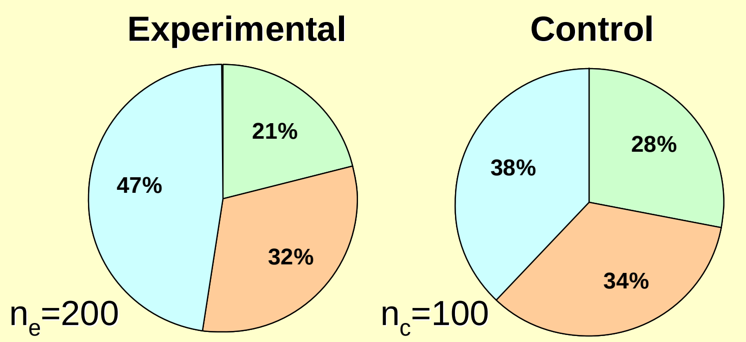
The new treatment for a disease is tested on 200 patients. The outcomes are classified as:

- **A** – patient is **completely treated**
- **B** – disease transforms into a **chronic form**
- **C** – treatment is **unsuccessful** ☹

In parallel the 100 patients treated with standard methods are observed

Category	Experimental	Control
A	94	38
B	42	28
C	64	34
Sum	200	100

◆ The proportions for 3 "classes" of patients with and without treatment are:



Are the proportions *significantly different* in control and experimental groups?

Goodness of fit test
A statistical test conducted to determine whether to reject a hypothesized probability distribution for a population.

Model – our assumption concerning the distribution, which we would like to test.

Observed frequency – frequency distribution for experimentally observed data, f_i

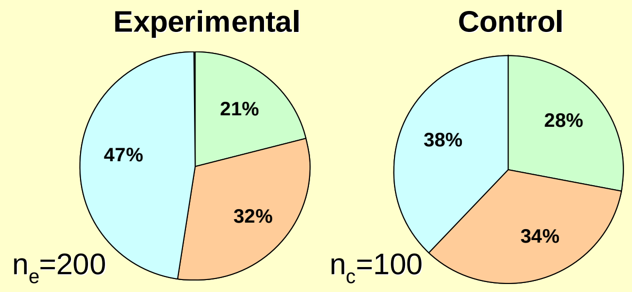
Expected frequency – frequency distribution, which we would expect from our **model**, e_i

Hypotheses for the test:

H_0: the population follows a multinomial distribution with the probabilities, specified by **model**

H_a: the population does not follow … **model**

◆ The proportions for 3 “classes” of patients with and without treatment are:



Are the proportions ***significantly different*** in control and experimental groups?

Test statistics for goodness of fit

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

χ^2 has ***k*–1** degree of freedom

At least 5 expected must be in each category!

TEST OF GOODNESS OF FIT

Example

The new treatment for a disease is tested on 200 patients. The outcomes are classified as:

- **A** – patient is **completely treated**
- **B** – disease transforms into a **chronic form**
- **C** – treatment is **unsuccessful** ☹

In parallel the 100 patients treated with standard methods are observed

Category	Experimental	Control
A	94	38
B	42	28
C	64	34
Sum	200	100

1. Select the model and calculate expected frequencies

Let's use control group (classical treatment) as a model, then:

Category	Control frequencies	Model for control	Expected freq., e	Experimental freq., f
A	38	0.38	76	94
B	28	0.28	56	42
C	34	0.34	68	64
Sum	100	1	**200**	200

2. Compare expected frequencies with the experimental ones and build χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

Category	**(f-e)2/e**
A	4.263
B	3.500
C	0.235
Chi2	**7.998**

3. Calculate p-value for χ^2 with d.f. = k–1

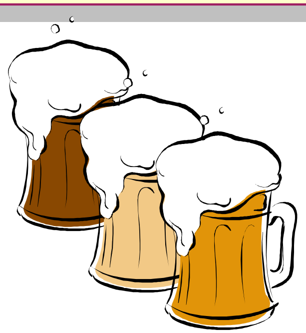
- `= CHIDIST(`χ^2`,d.f.)`
- `= CHITEST(`f, e`)`

p-value = 0.018, reject H$_0$

Goodness of Fit for Independence Test: Example

Alber's Brewery manufactures and distributes three types of beer: **white**, **regular**, and **dark**. In an analysis of the market segments for the three beers, the firm's market research group raised the question of whether preferences for the three beers differ among **male** and **female** beer drinkers. If beer preference is independent of the gender of the beer drinker, one advertising campaign will be initiated for all of Alber's beers. However, if beer preference depends on the gender of the beer drinker, the firm will tailor its promotions to different target markets.

beer.xls



H_0: Beer preference is **independent** of the gender of the beer drinker

H_a: Beer preference is **not independent** of the gender of the beer drinker

sex\beer	White	Regular	Dark	Total
Male	20	40	20	**80**
Female	30	30	10	**70**
Total	**50**	**70**	**30**	**150**

Goodness of Fit for Independence Test: Example

1. Build model assuming independence

sex\beer	White	Regular	Dark	Total
Male	20	40	20	**80**
Female	30	30	10	**70**
Total	**50**	**70**	**30**	**150**

	White	Regular	Dark	Total
Model	0.3333	0.4667	0.2000	1

2. Transfer the model into expected frequencies, multiplying model value by number in group

sex\beer	White	Regular	Dark	Total
Male	26.67	37.33	16.00	**80**
Female	23.33	32.67	14.00	**70**
Total	**50**	**70**	**30**	**150**

$$e_{ij} = \frac{(Row\ i\ Total)(Column\ \ j\ Total)}{Sample\ Size}$$

3. Build χ^2 statistics

$$\chi^2 = \sum_i^n \sum_j^m \frac{\left(f_{ij} - e_{ij}\right)^2}{e_{ij}}$$

χ^2 =6.122

χ^2 distribution with d.f.=(n – 1)(m – 1), provided that the expected frequencies are 5 or more for all categories.

4. Calculate p-value

p-value = 0.047, reject H_0

TEST FOR CONTINUOUS DISTRIBUTIONS

Test for Normality: Example

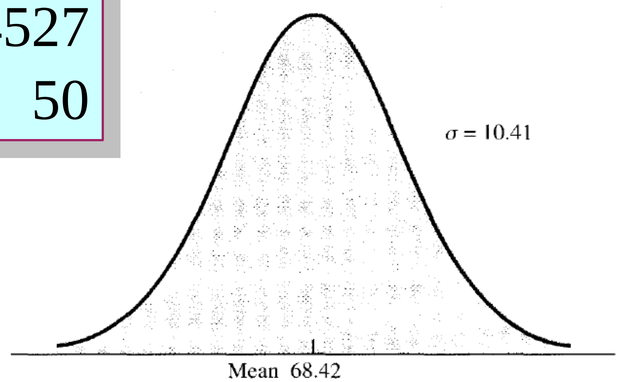
Chemline hires approximately 400 new employees annually for its four plants. The personnel director asks whether a normal distribution applies for the population of aptitude test scores. If such a distribution can be used, the distribution would be helpful in evaluating specific test scores; that is, scores in the upper 20%, lower 40%, and so on, could be identified quickly. Hence, we want to test the null hypothesis that the population of test scores has a normal distribution. The study will be based on 50 results.

chemline.xls

Mean	68.42
Standard Deviation	10.4141
Sample Variance	108.4527
Count	50

Aptitude test scores

71	86	56	61	65
60	63	76	69	56
55	79	56	74	93
82	80	90	80	73
85	62	64	54	54
65	54	63	73	58
77	56	65	76	64
61	84	70	53	79
79	61	62	61	65
66	70	68	76	71



H_0: The population of test scores **has a normal distribution** with mean **68.42** and standard deviation **10.41**

H_a: the population **does not have** a mentioned distribution

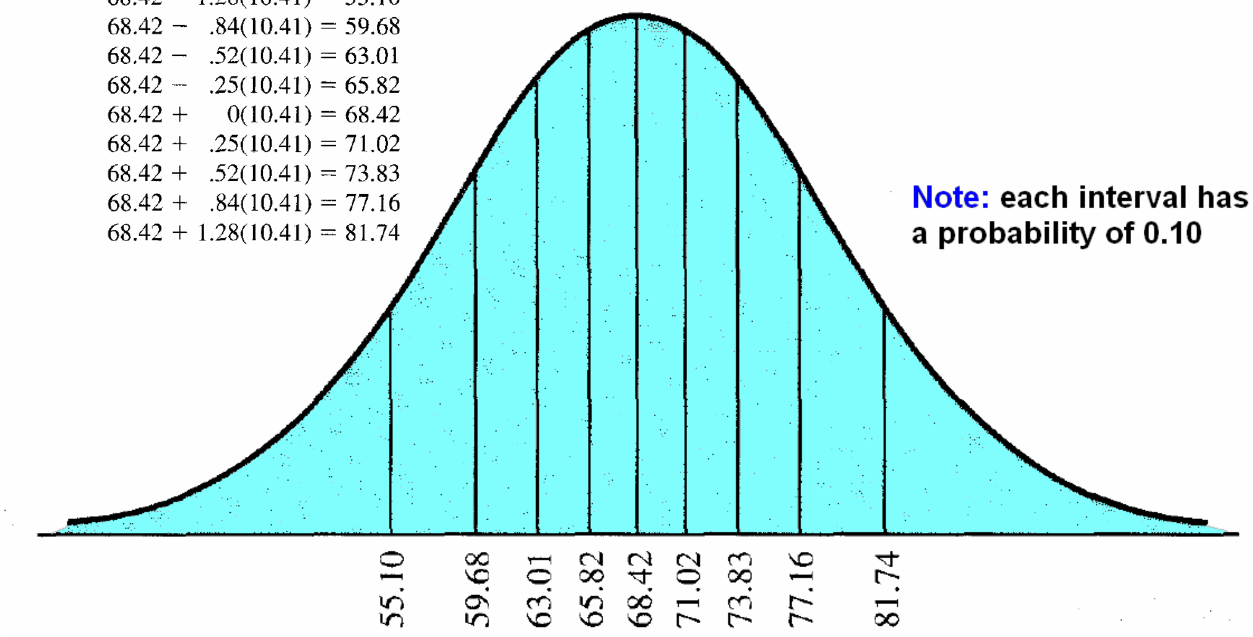
Test for Normality: Example

chemline.xls

Mean	68.42
Standard Deviation	10.4141
Sample Variance	108.4527
Count	50

Bin	*Observed frequency*	*Expected frequency*
55.1	5	5
59.68	5	5
63.01	9	5
65.82	6	5
68.42	2	5
71.02	5	5
73.83	2	5
77.16	5	5
81.74	5	5
More	6	5
Total	**50**	**50**

Lower 10%:	68.42 − 1.28(10.41) = 55.10
Lower 20%:	68.42 − .84(10.41) = 59.68
Lower 30%:	68.42 − .52(10.41) = 63.01
Lower 40%:	68.42 − .25(10.41) = 65.82
Mid-score:	68.42 + 0(10.41) = 68.42
Upper 40%:	68.42 + .25(10.41) = 71.02
Upper 30%:	68.42 + .52(10.41) = 73.83
Upper 20%:	68.42 + .84(10.41) = 77.16
Upper 10%:	68.42 + 1.28(10.41) = 81.74



Note: each interval has a probability of 0.10

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

χ^2 distribution with d.f.= $n - p - 1$,

where p – number of estimated parameters

$p = 2$ includes mean and variance

d.f. = 10 – 2 – 1

$\chi^2 = 7.2$

p-value = 0.41, cannot reject H_0

Thank you for your attention