LUXEMBOURG INSTITUTE OF HEALTH

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

Bioinformatics Platform (BIOINFO)
Department of Medical Informatics, LIH

BIOSTATISTICS for PhDs

Lecture 3

Linear Models

Peter Nazarov

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

2024-03-04

http://edu.modas.lu

LUXEMBOURG INSTITUTE OF HEALTH

COURSE OVERVIEW

Outline

- **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
- **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

https://cran.r-project.org/ https://posit.co/downloads/

- **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
- **Lecture 4, 2024-04-08** *(please, propose!)*
  - factors in linear regression
  - logistic regression
  - omics data analysis?
  - survival analysis?
  - clustering?
  - more practical exercise?

Let's work at a comfortable speed!

Materials and other courses:

http://edu.modas.lu

LUXEMBOURG INSTITUTE OF HEALTH

HYPOTHESES FOR VARIANCE

Confidence intervals for variance

Hypotheses for variance

Goodness of fit, test for independence

ANalysis Of VAriance (ANOVA)

Linear regression

Logistic regression

INTERVAL ESTIMATION FOR VARIANCE

Variance Sampling Distribution

Variance
A measure of variability based on the squared deviations of the data values about the mean.

population

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

sample

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

The interval estimation for variance is build using the following measure:

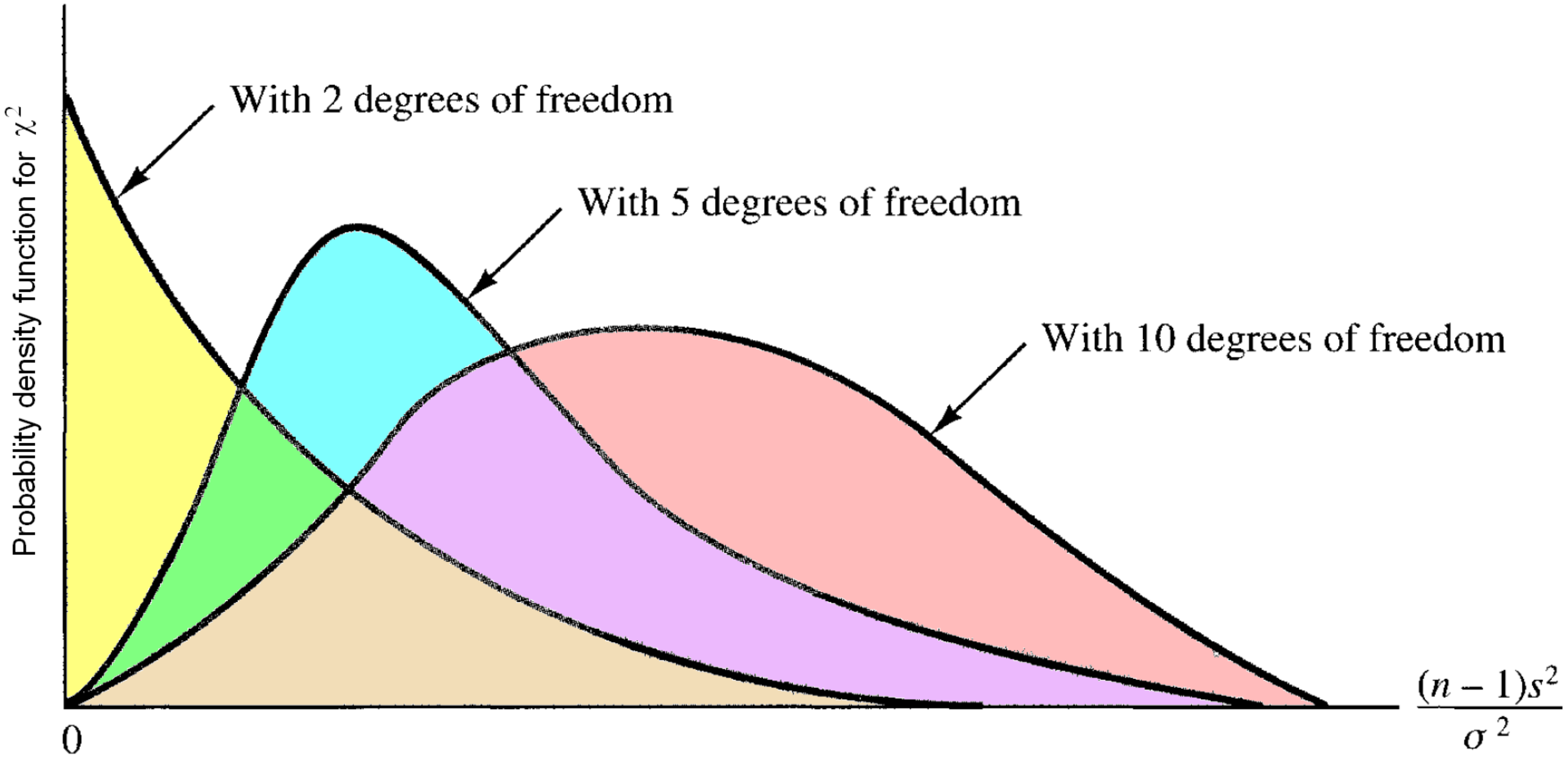
$$(n-1)\frac{s^2}{\sigma^2}$$

Sampling distribution of $(n\text{-}1)s^2/\sigma^2$
Whenever a simple random sample of size n is selected from a normal population, the sampling distribution of **$(n\text{-}1)s^2/\sigma^2$** has a **chi-square distribution** (χ^2) with n-1 degrees of freedom.

$$(n-1)\frac{s^2}{\sigma^2} = \chi^2_{df=n-1}$$

INTERVAL ESTIMATION FOR VARIANCE

χ^2 Distribution

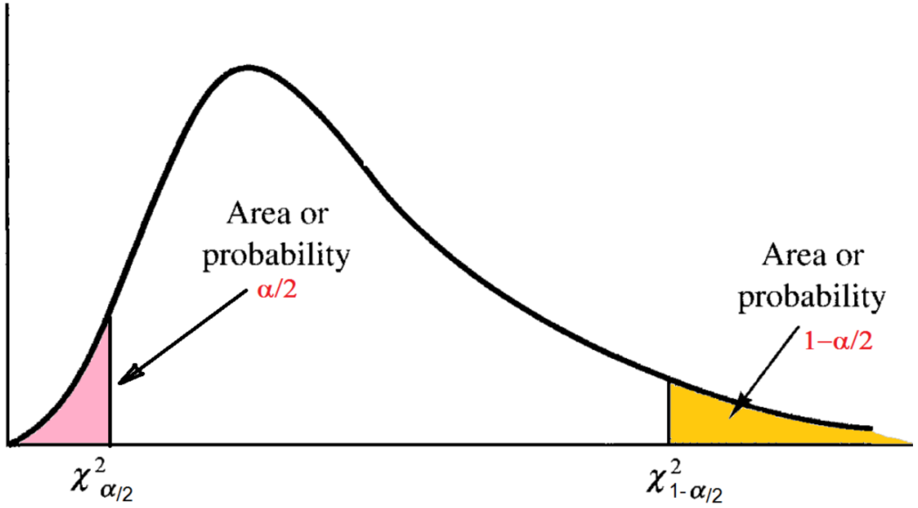


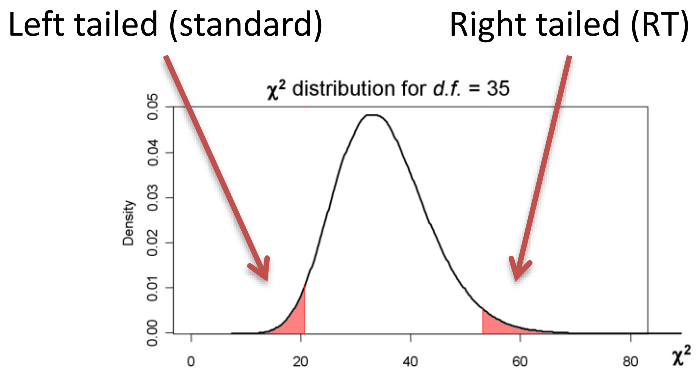
χ^2 distribution works only for sampling from normal population

$$\chi^2_{df=k} = \sum_{i=1}^{k} x_i^2 \quad where\ x_i - normal$$

INTERVAL ESTIMATION FOR VARIANCE

χ^2 Probabilities in Table and Excel






```
= CHISQ.DIST(χ², n-1, true)
= CHISQ.DIST.RT(χ², n-1)
= CHISQ.INV(α/2, n-1)
= CHISQ.INV.RT(α/2, n-1)
```

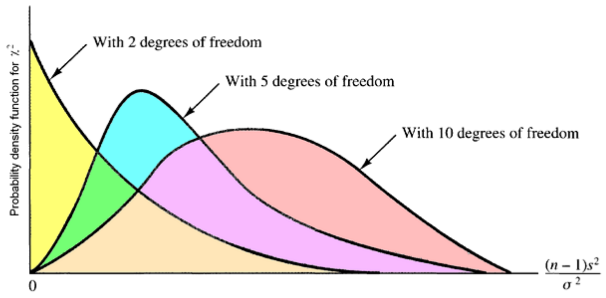
```
pchisq(x = χ², df = n-1)
qchisq(p = α/2, df = n-1)
```

Degrees of Freedom	Area in Upper Tail .99	.975	.95	.90	.10	.05	.025	.01
1	.000	.001	.004	.016	2.706	3.841	5.024	6.635
2	.020	.051	.103	.211	4.605	5.991	7.378	9.210
3	.115	.216	.352	.584	6.251	7.815	9.348	11.345
4	.297	.484	.711	1.064	7.779	9.488	11.143	13.277
5	.554	.831	1.145	1.610	9.236	11.070	12.832	15.086
6	.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191

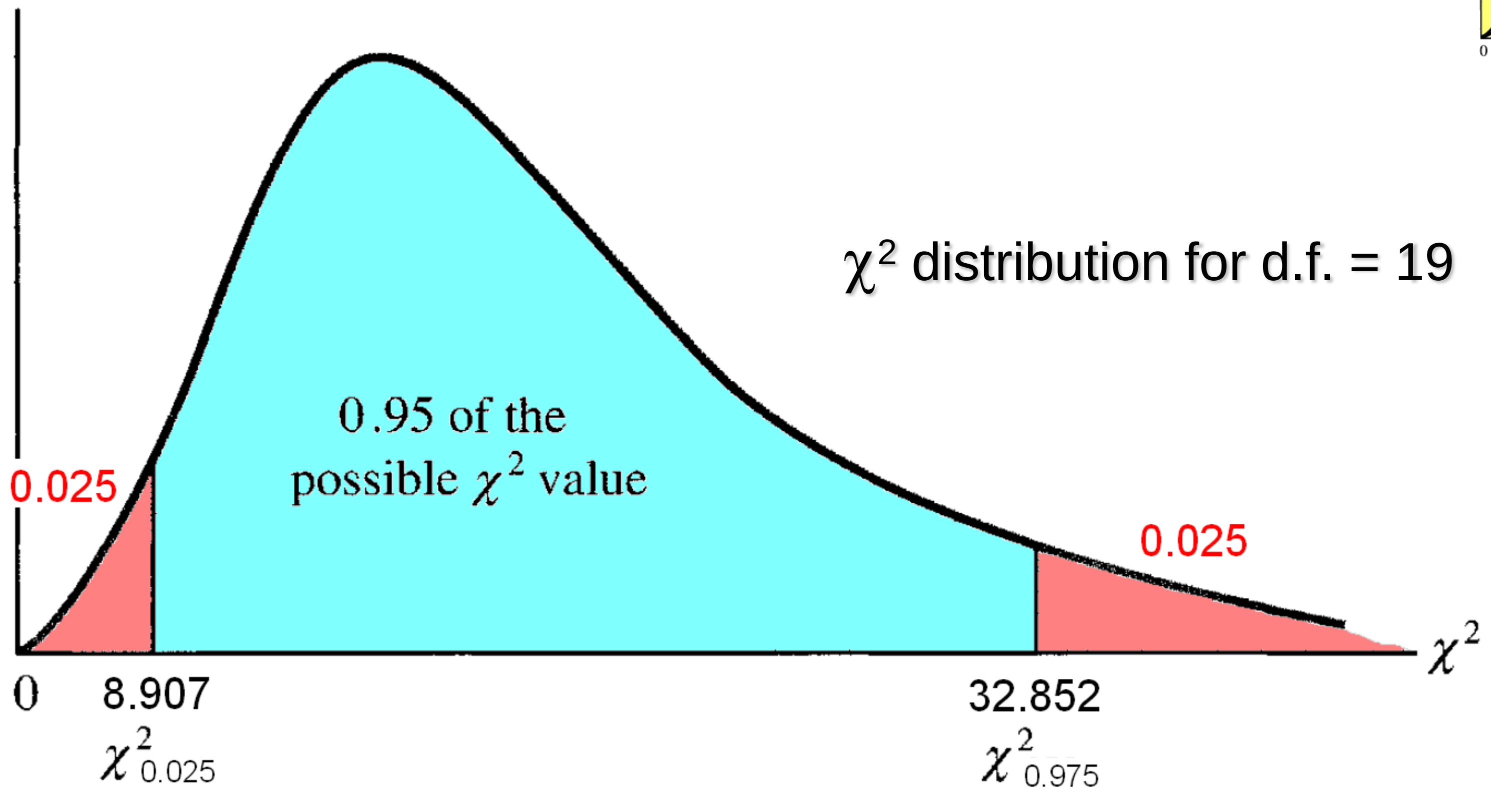
INTERVAL ESTIMATION FOR VARIANCE

χ^2 Distribution for Interval Estimation

$$\chi^2 = (n-1)\frac{s^2}{\sigma^2}$$



χ^2 distribution for d.f. = 19



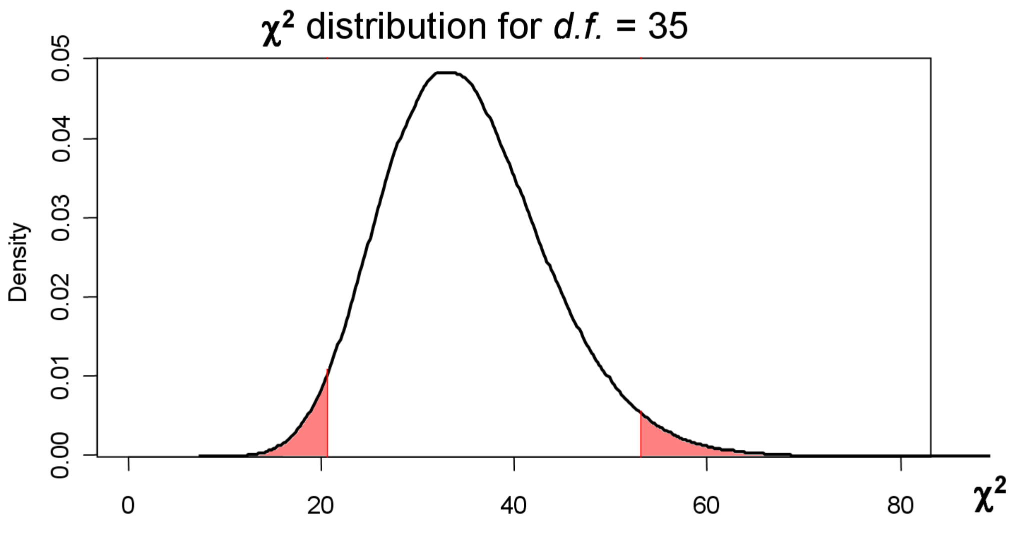

```
qchisq(0.025,19)
qchisq(0.975,19)
```

INTERVAL ESTIMATION FOR VARIANCE

Interval Estimation

$$\chi^2_{\alpha/2} \le (n-1)\frac{s^2}{\sigma^2} \le \chi^2_{1-\alpha/2}$$

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$$



Suppose sample of $n = 36$ coffee cans is selected and $m = 2.92$ and $s = 0.18$ lbm is observed. Provide 95% confidence interval for the standard deviation

$$\frac{(36-1)0.18^2}{53.203} \le \sigma^2 \le \frac{(36-1)0.18^2}{20.569}$$

```
= CHISQ.INV(α/2, n-1)
= CHISQ.INV.RT(α/2, n-1)
```

```
qchisq(0.025,36-1)
qchisq(1-0.025,36-1)
```

$$0.0213 \le \sigma^2 \le 0.0551$$

$$0.146 \le \sigma \le 0.235$$

INTERVAL ESTIMATION FOR VARIANCE

Hypotheses about Population Variance

H_0: $\sigma^2 \le$ const
H_a: $\sigma^2 >$ const

H_0: $\sigma^2 \ge$ const
H_a: $\sigma^2 <$ const

H_0: $\sigma^2 =$ const
H_a: $\sigma^2 \ne$ const

	Lower Tail Test	**Upper Tail Test**	**Two-Tailed Test**
Hypotheses	$H_0: \sigma^2 \ge \sigma_0^2$ $H_a: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 \le \sigma_0^2$ $H_a: \sigma^2 > \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_a: \sigma^2 \ne \sigma_0^2$
Test Statistic	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$
Rejection Rule: p-Value Approach	Reject H_0 if p-value $\le \alpha$	Reject H_0 if p-value $\le \alpha$	Reject H_0 if p-value $\le \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $\chi^2 \le \chi^2_{(1-\alpha)}$	Reject H_0 if $\chi^2 \ge \chi^2_{\alpha}$	Reject H_0 if $\chi^2 \le \chi^2_{(1-\alpha/2)}$ or if $\chi^2 \ge \chi^2_{\alpha/2}$

VARIANCES OF TWO POPULATIONS

Sampling Distribution

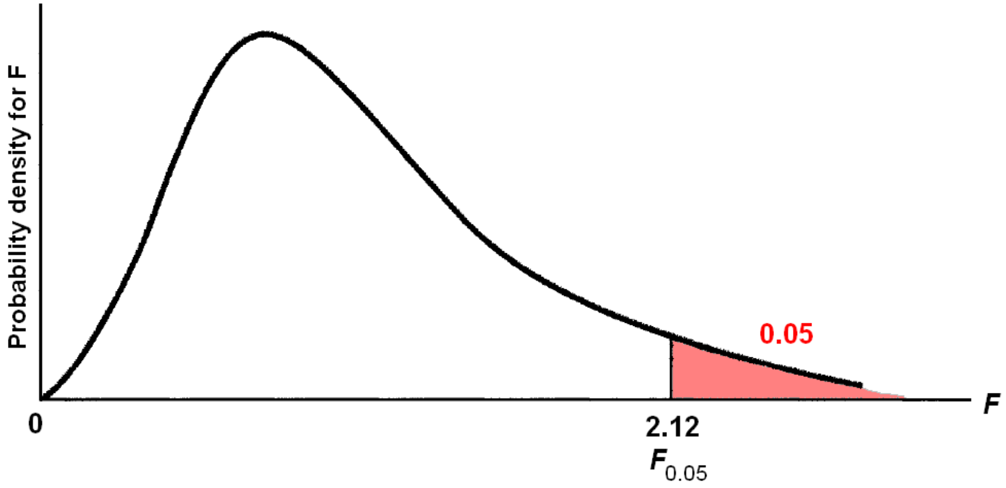
In many statistical applications we need a comparison between variances of two populations. In fact well-known ANOVA-method is base on this comparison.

The statistics is build for the following measure:

$$F = \frac{s_1^2}{s_2^2}$$

Sampling distribution of s_1^2/s_2^2 when $\sigma_1^2 = \sigma_2^2$
Whenever a independent simple random samples of size n_1 and n_2 are selected from two normal populations with equal variances, the sampling of s_1^2/s_2^2 has **F-distribution** with n_1-1 degree of freedom for numerator and n_2-1 for denominator.

F-distribution for 20 d.f. in numerator and 20 d.f. in denominator



Distributions

```
= F.DIST(x, df1,
         df2,TRUE)
= F.INV(p, df1,
        df2,TRUE)
```

```
pf(x,df1,df2,…)

qf(p,df1,df2,…)
```

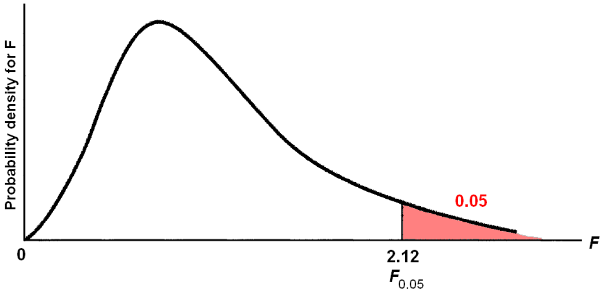
Tests

```
= F.TEST(data1,data2)
```

```
var.test(data1,data2)
```

VARIANCES OF TWO POPULATIONS

Hypotheses about Variances of Two Populations



H_0: $\sigma_1^2 \leq \sigma_2^2$
H_a: $\sigma_1^2 > \sigma_2^2$

H_0: $\sigma_1^2 = \sigma_2^2$
H_a: $\sigma_1^2 \neq \sigma_2^2$

	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_a : \sigma_1^2 > \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_a : \sigma_1^2 \neq \sigma_2^2$ *Note: Population 1 has the lager sample variance*
Test Statistic	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{s_1^2}{s_2^2}$
Rejection Rule: p-Value Approach	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 $F \geq F_\alpha$	Reject H_0 if $F \geq F_\alpha$

Tests

```
= F.TEST(data1,data2)
```

```
var.test(data1,data2)
```

VARIANCES OF TWO POPULATIONS

Example

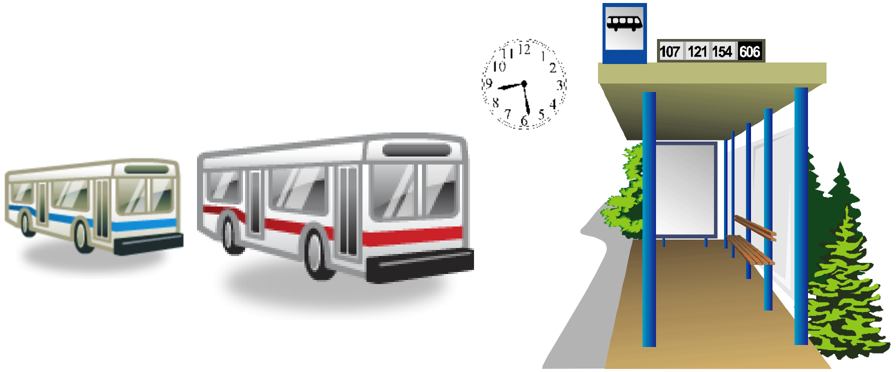
schoolbus

#	Milbank	Gulf Park
1	35.9	21.6
2	29.9	20.5
3	31.2	23.3
4	16.2	18.8
5	19.0	17.2
6	15.9	7.7
7	18.8	18.6
8	22.2	18.7
9	19.9	20.4
10	16.4	22.4
11	5.0	23.1
12	25.4	19.8
13	14.7	26.0
14	22.7	17.1
15	18.0	27.9
16	28.1	20.8
17	12.1	
18	21.4	
19	13.4	
20	22.9	
21	21.0	
22	10.1	
23	23.0	
24	19.4	
25	15.2	
26	28.2	

Dullus County Schools is renewing its school bus service contract for the coming year and must select one of two bus companies, the Milbank Company or the Gulf Park Company. We will use the variance of the arrival or pickup/delivery times as a primary measure of the quality of the bus service. Low variance values indicate the more consistent and higher-quality service. If the variances of arrival times associated with the two services are equal Dullus School administrators will select the company offering the better financial terms However, if the sample data on bus arrival times for the two companies indicate a significant difference between the variances, the administrators may want to give special consideration to the company with the better or lower variance service. The appropriate hypotheses follow

$$H_0\colon \sigma_1^2 = \sigma_2^2$$
$$H_a\colon \sigma_1^2 \neq \sigma_2^2$$

If H_0 can be rejected, the conclusion of unequal service quality is appropriate. We will use a level of significance of $\alpha = .10$ to conduct the hypothesis test.



VARIANCES OF TWO POPULATIONS

Example

schoolbus

#	Milbank	Gulf Park
1	35.9	21.6
2	29.9	20.5
3	31.2	23.3
4	16.2	18.8
5	19.0	17.2
6	15.9	7.7
7	18.8	18.6
8	22.2	18.7
9	19.9	20.4
10	16.4	22.4
11	5.0	23.1
12	25.4	19.8
13	14.7	26.0
14	22.7	17.1
15	18.0	27.9
16	28.1	20.8
17	12.1	
18	21.4	
19	13.4	
20	22.9	
21	21.0	
22	10.1	
23	23.0	
24	19.4	
25	15.2	
26	28.2	

1. Let us start from estimation of the variances for 2 data sets

Milbank: $s_1^2 = 48$, $n_1 = 26$

Gulf Park: $s_2^2 = 20$, $n_2 = 16$

interval estimation (optionally)

Milbank: $\sigma_1^2 \approx 48$ $(29.5 \div 91.5)$

Gulf Park: $\sigma_2^2 \approx 20$ $(10.9 \div 47.9)$

2. Let us calculate the *F*-statistics

$$F = \frac{s_1^2}{s_2^2} = \frac{48}{20} = 2.40$$

3. … and p-value = 0.08

p-value = 0.08 < α = 0.1

In Excel use <u>one</u> of the functions:

- = **2*F.DIST.RT**(F,n_1-1,n_2-1)
- = **F.TEST**(data1,data2)

In R use <u>one</u> of solutions:

```
2*(1-pf(2.4,25,15))
var.test(data1,data2)
```

HYPOTHESES FOR VARIANCE

Confidence intervals for variance

Hypotheses for variance

Goodness of fit, test for independence

ANalysis Of VAriance (ANOVA)

Linear regression

Logistic regression

TEST OF GOODNESS OF FIT

Multinomial Population

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.

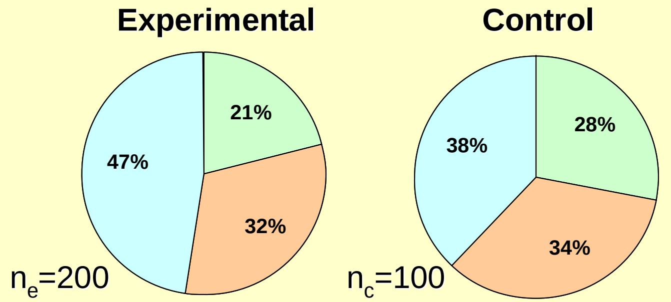
The new treatment for a disease is tested on 200 patients. The outcomes are classified as:

- **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?

TEST OF GOODNESS OF FIT

Goodness of Fit

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

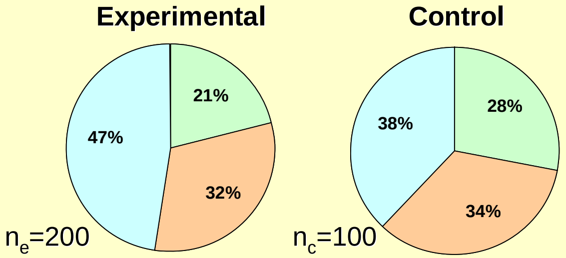
Expected frequency – frequency distribution, which we would expect from our **model**, e_i

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

```
# input data
Tab = cbind(c(94,42,64),
            c(38,28,34))
colnames(Tab) =
          c("exp","ctrl")

rownames(Tab) =
          c("A","B","C")

# control defines Model
mod=Tab[,2]/sum(Tab[,2])

# test Model for 'exp'
chisq.test(Tab[,1],p=mod)
```

1. Select the model and calculate expected frequencies

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

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	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

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

Here k=3 => df=2

= **CHISQ.DIST.RT**(χ^2,d.f.)

p-value = 0.018, reject H_0

TEST OF INDEPENDENCE

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

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**

TEST OF INDEPENDENCE

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**

Model	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

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

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

= **CHISQ.DIST.RT**(χ^2,d.f.)

p-value = 0.047, reject H_0

```
# input data
Tab = rbind(c(20,40,20),
            c(30,30,10))
colnames(Tab) = c("white",
          "regular","dark")
rownames(Tab) =
          c("male","female")
Tab

# it is simple:
chisq.test(Tab)
```

TEST FOR CONTINUOUS DISTRIBUTIONS

Test for Normality: Example

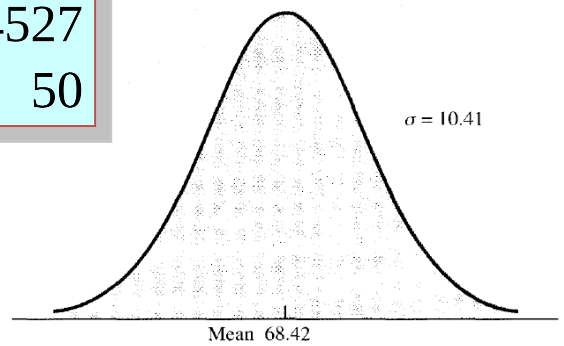
Chemline hires approximately 400 new employees annually for its four plants. The personnel director asks whether a normal distribution applies for the population of aptitude test scores. If such a distribution can be used, the distribution would be helpful in evaluating specific test scores; that is, 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

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 CONTINUOUS DISTRIBUTIONS

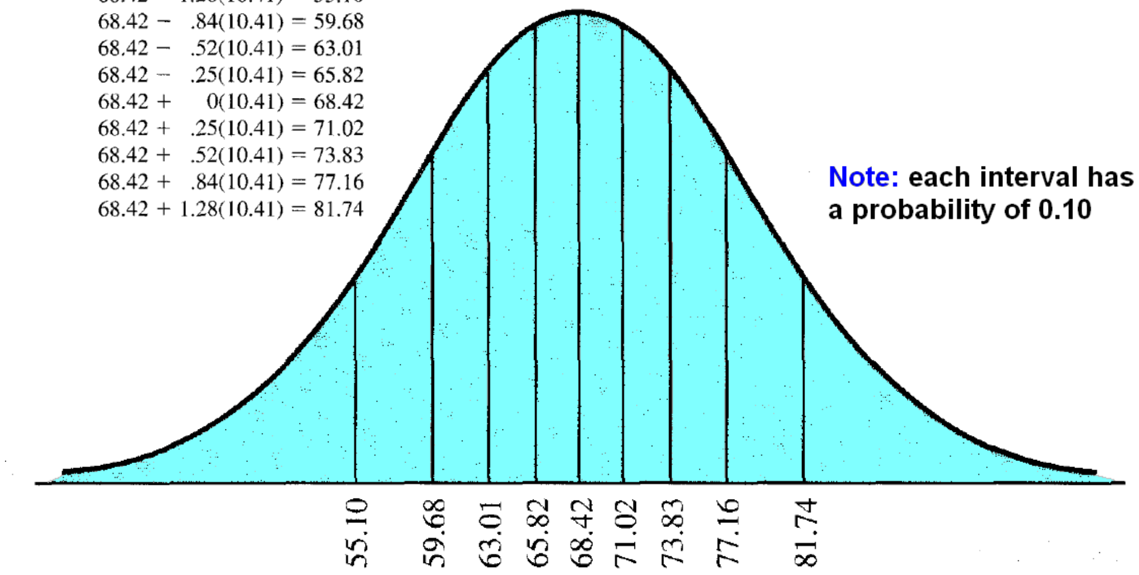
Test for Normality: Example

chemline

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.= $k - p - 1$,

where p – number of estimated parameters, k – number of bins

p = 2 includes mean and variance
d.f. = 10 – 2 – 1
$\chi^2 = 7.2$

p-value = 0.41, cannot reject H_0

More precise: $\chi^2 = 6.4$ ☺

R: more advanced

```
#input data
x = scan(
"http://edu.modas.lu/data/txt/chemline.txt", skip=1)

#Shapiro-Wilk
shapiro.test(x)

#Kolmogorov-Smirnov
ks.test(x,"pnorm",
        mean=mean(x),
        sd=sd(x))

#Jarque-Bera
library(tseries)
jarque.bera.test(x)
```

https://datasharkie.com/how-to-test-for-normality-in-r/

HYPOTHESES FOR VARIANCE

Confidence intervals for variance

Hypotheses for variance

Goodness of fit, test for independence

ANalysis Of VAriance (ANOVA)

Linear regression

Logistic regression

INTRODUCTION TO ANOVA

Why ANOVA?

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

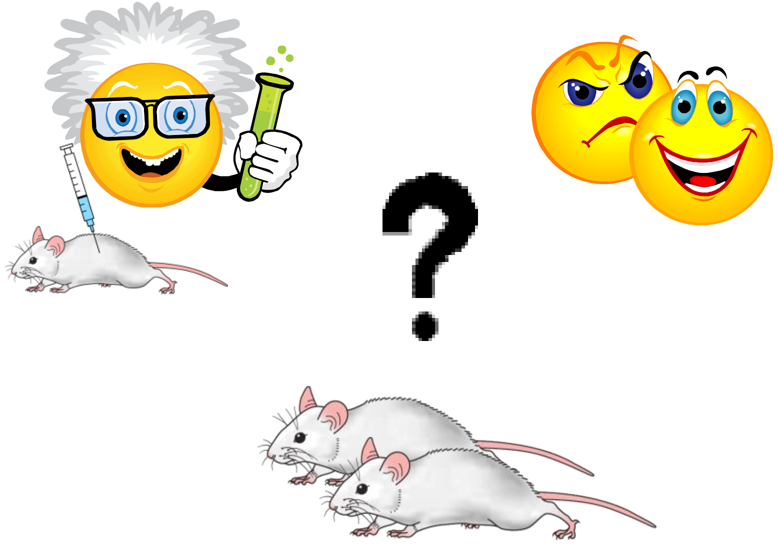
Validation of the effects
We assume that we have several factors affecting our data. Which factors are most significant? Which can be neglected?

ANOVA
example from Partek™

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: $1-(0.95)^{10} = 0.4$



INTRODUCTION TO ANOVA

Example

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?

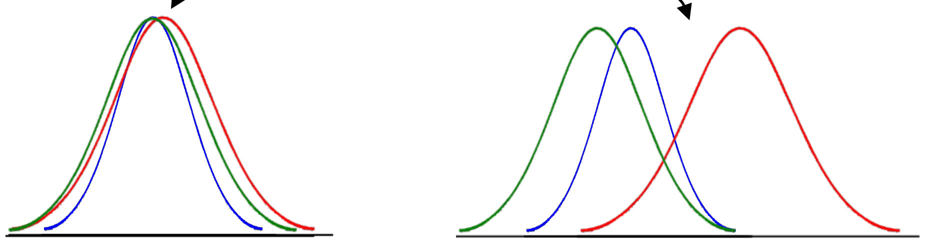
depression

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



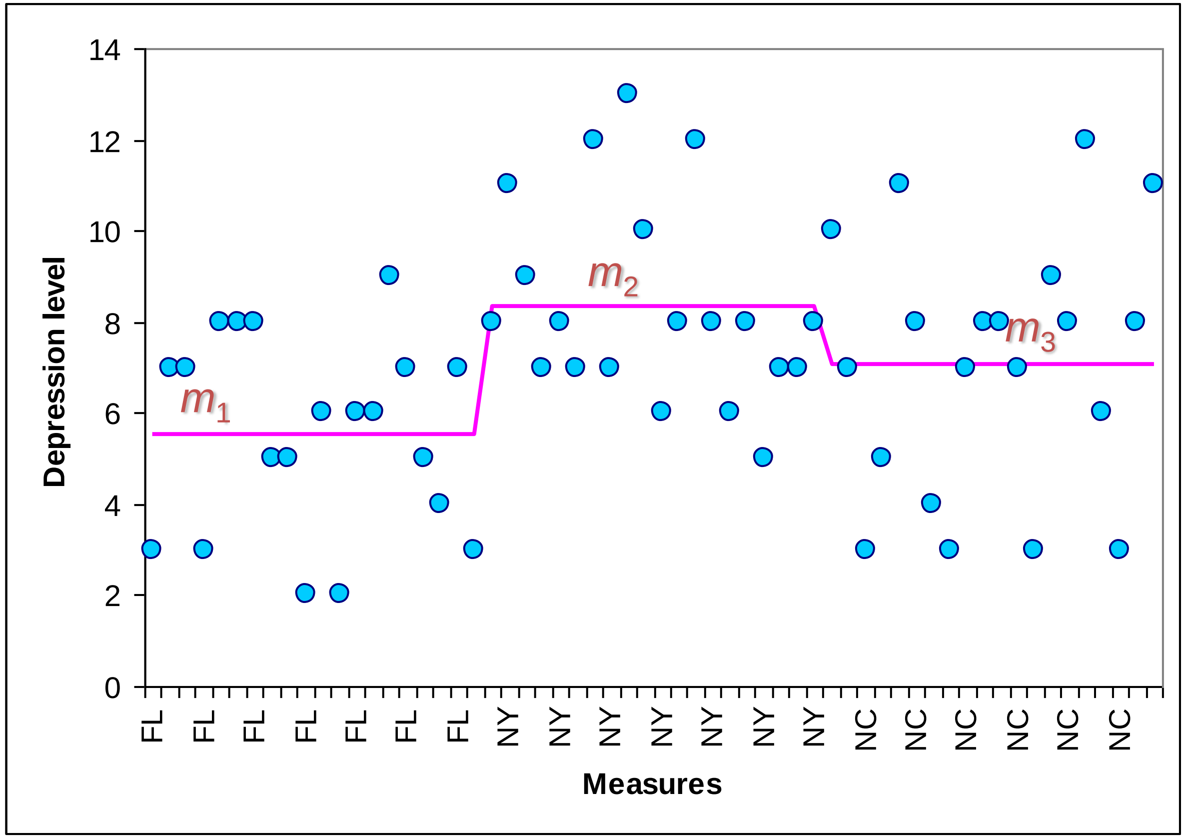
LUXEMBOURG INSTITUTE OF HEALTH

INTRODUCTION TO ANOVA

Meaning

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

H_a: not all 3 means are equal

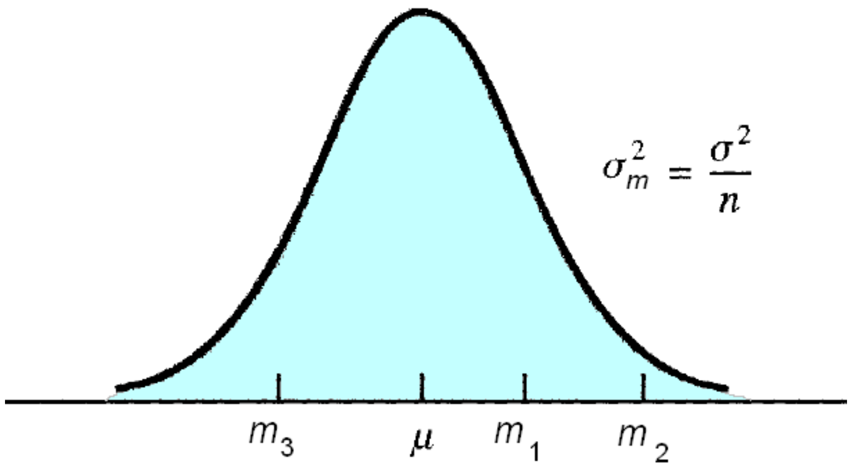


INTRODUCTION TO ANOVA

Assumptions for ANOVA, ANOVA in R

Assumptions for Analysis of Variance

1. For each population, the response variable is **normally distributed**
2. The variance of the respond variable, denoted as **σ^2 is the same** for all of the populations.
3. The observations must be **independent.**




```
# build the model
model = aov(x ~ fact1 + …, data)
```

```
# summary (anova table)
summary(model)
anova(model)
```

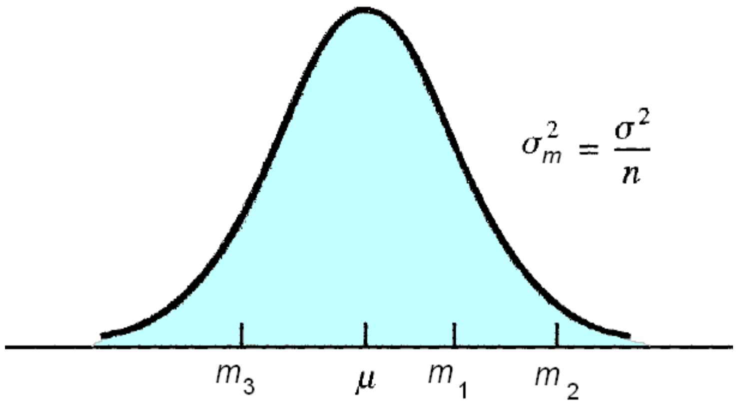
```
# posthoc
TukeyHSD(model)
```

```
# check for normality
shapiro.test( residuals(model) )
```

INTRODUCTION TO ANOVA

Some Calculations

Parameter	Florida	New York	N. Carolina
m=	5.55	8.35	7.05
overall mean=	6.98333		
var=	4.5763	4.7658	8.0500



Let's estimate the variance of sampling distribution. If H_0 is true, then all m_i belong to the same distribution

$$\sigma_m^2 = \frac{\sum_{i=1}^{k}(m_i - \overline{m})^2}{k-1} = \frac{(5.55-6.98)^2 + (8.35-6.98)^2 + (7.05-6.98)^2}{3-1} = 1.96$$

$$\boxed{\sigma^2 = n\sigma_m^2 = 20 \times 1.96 = 39.27}$$

– this is called between-treatment estimate, works only at H_0

At the same time, we can estimate the variance just by averaging out variances for each populations:

– this is called within-treatment estimate

$$\sigma^2 = \frac{\sum_{i=1}^{k}\sigma_i^{\,2}}{k} = \frac{4.58+4.77+8.05}{3} = 5.8$$

Does between-treatment estimate and within-treatment estimate give variances of the same "population"?

SINGLE-FACTOR ANOVA

Theory

H$_0$: $\mu_1 = \mu_2 = \ldots = \mu_k$

H$_a$: **not all *k* means are equal**

Means for treatments

$$m_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

Variances treatments

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - m_j)^2}{n_j - 1}$$

Total mean

$$\overline{m} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T}$$

$$n_T = n_1 + n_2 + \cdots + n_k$$

due to treatment

Sum squares

$$SSTR = \sum_{j=1}^{k} n_j (m_j - \overline{m})^2$$

Mean squares, $\sigma_{beetween}^2$

$$MSTR = \frac{SSTR}{k-1}$$

due to error

Sum squares

$$SSE = \sum_{j=1}^{k} (n_j - 1) s_j^2$$

Mean squares, σ_{within}^2

$$MSE = \frac{SSE}{n_T - k}$$

Test of variance equality

$$F = \frac{MSTR}{MSE}$$

p-value for the treatment effect

$$p - value$$

SINGLE-FACTOR ANOVA

The Main Equation

Total sum squares

$$SST=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{m}\right)^2$$

SS due to treatment

$$SSTR=\sum_{j=1}^{k}n_j\left(m_j-\overline{m}\right)^2$$

$$SST = SSTR + SSE$$

SS due to error

$$SSE=\sum_{j=1}^{k}\left(n_j-1\right)s_j^2$$

Total variability of the data include variability due to treatment and variability due to error

$$d.f.(SST)=d.f.(SSTR)+d.f.(SSE)$$
$$n_T-1=\left(k-1\right)+\left(n_T-k\right)$$

Partitioning
The process of allocating the total sum of squares and degrees of freedom to the various components.

SINGLE-FACTOR ANOVA

Example

Sum squares **total, SST**

distances from ● to –

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \overline{m})^2$$

Sum squares due to **error, SSE**

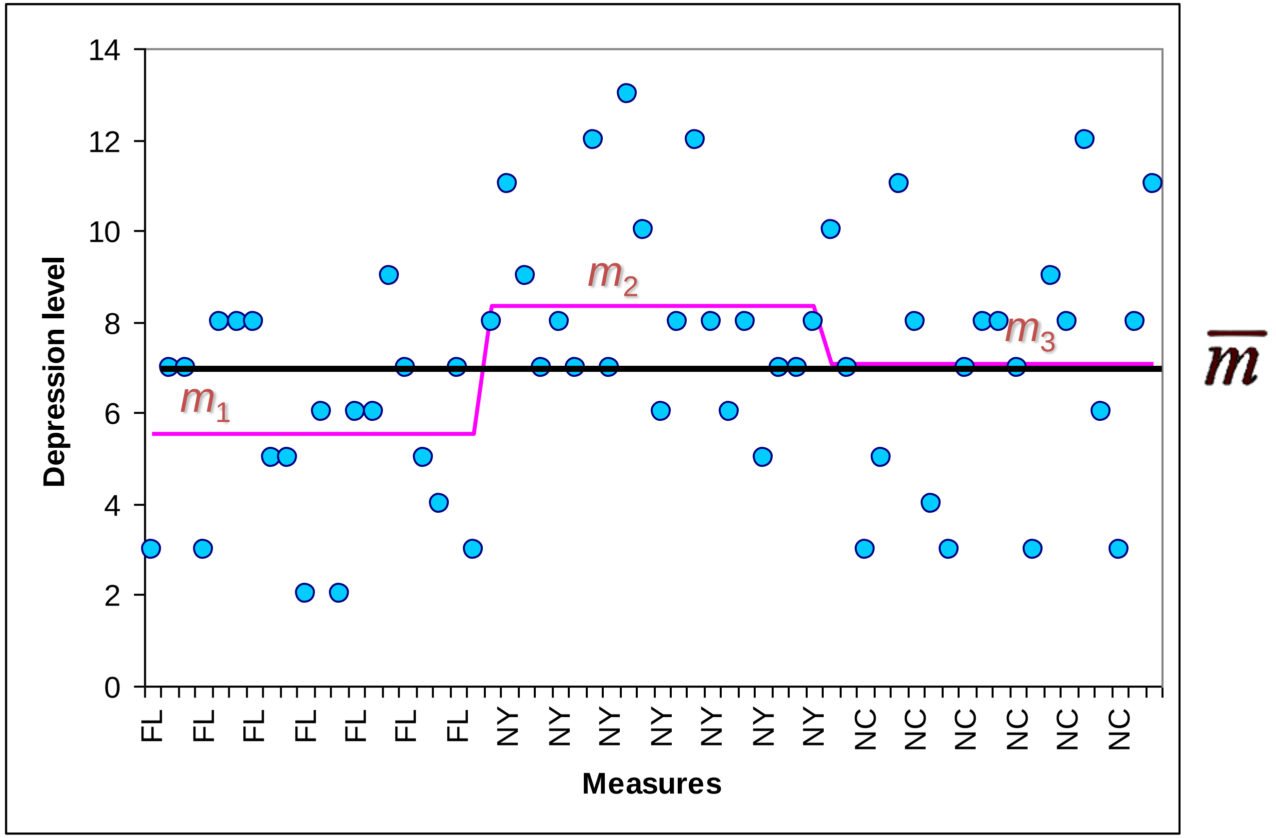
distances from ● to –

$$SSE = \sum_{j=1}^{k} (n_j - 1) s_j^2$$

Sum squares due to **treatment, SSTR**

distances from – to –

$$SSTR = \sum_{j=1}^{k} n_j (m_j - \overline{m})^2$$



$$SST = SSTR + SSE$$

SINGLE-FACTOR ANOVA

Example: ANOVA in R

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:

- Data → Data Analysis → ANOVA Single Factor

depression2

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

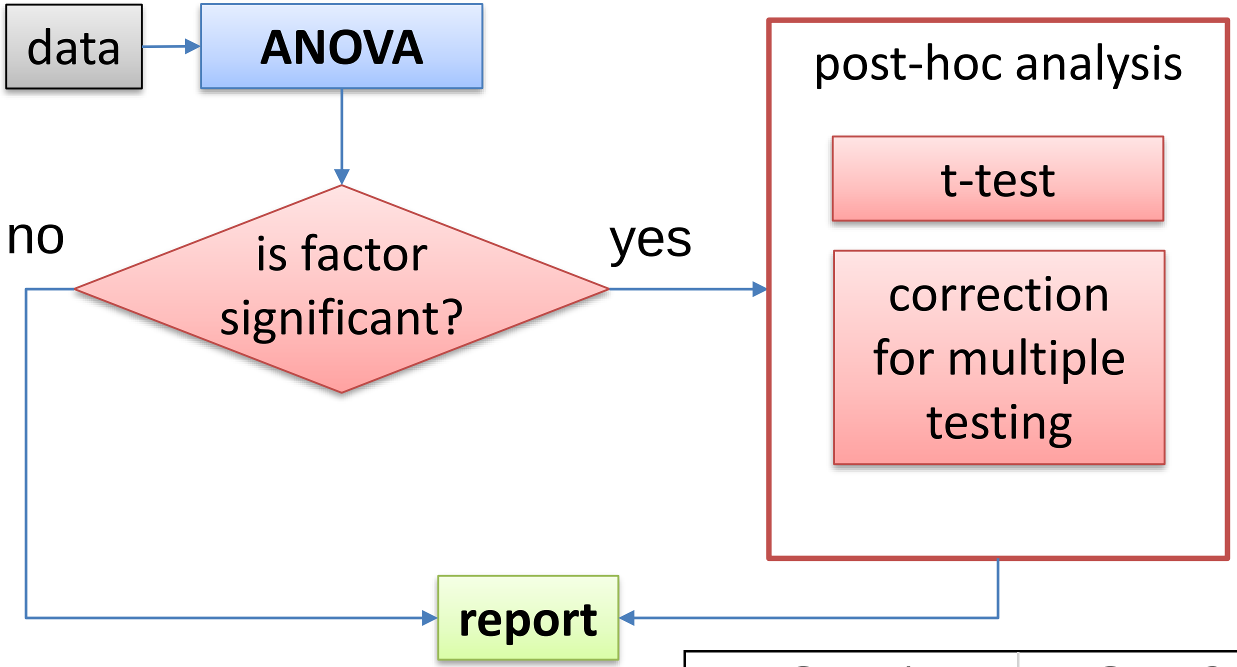
```
# read dataset

Dep = read.table(
"http://edu.modas.lu/data/
txt/depression2.txt",
  header=T,
  sep="\t",
  as.is=FALSE)

str(Dep)

# consider only healthy

DepGH = Dep[Dep$Health ==
             "good",]

# build 1-way ANOVA model

res1 = aov(Depression ~
            Location, DepGH)

summary(res1)
```

SINGLE-FACTOR ANOVA

Post-hoc Analysis

Post-hoc analysis
allows for additional exploration of significant differences in the data, when significant effect of the factor was already confirmed (for example, by ANOVA).

```
# build 1-way ANOVA model
res1 = aov(Depression ~
           Location, DepGH)
summary(res1)

# add post-hoc analysis
TukeyHSD(res1)
```



Calculate rank (*k*) by

= RANK.AVG(...)

$$FDR = \frac{pval * nComparisons}{k}$$

Group1	Group2	p-value	k	FDR
Florida	New York	0.00021	1	0.00063
Florida	North Carolina	0.0667	2	0.10005
New York	North Carolina	0.11264	3	0.11264

If you can – use **Tukey Honest Significant Differences**

if not – just do FDR-adjustment

SINGLE-FACTOR ANOVA

Non-parametric (Kruskal-Wallis)

Kruskal-Wallis rank sum test
is a non-parametric version of 1-way ANOVA (ANOVA on ranks).

```
# non-parametric
kruskal.test(DepGH)
# posthoc 1
pairwise.wilcox.test(DepGH$Depression,
DepGH$Location, p.adjust.method = "bonf")
# posthoc 2
#install.packages("dunn.test")
library(dunn.test)
dunn.test(DepGH$Depression, DepGH$Location)
```

MULTI-FACTOR ANOVA

Factors and Treatments

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
```

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.

```r
# read dataset

Dep = read.table(
"http://edu.modas.lu/data/
txt/depression2.txt",
  header=T,
  sep="\t",
  as.is=FALSE)

str(Dep)

# build 2-way ANOVA model

res2 = aov( Depression ~
  Health + Location+
  Health*Location, Dep)

summary(res2)

# post-hoc

TukeyHSD(res2)
```

MULTI-FACTOR ANOVA

2-factor ANOVA with *r* Replicates

Replications
The number of times each experimental condition is repeated in an experiment.

a = number of levels of factor A
b = number of levels of factor B
r = number of replications
n_T = total number of observations taken in the experiment; $n_T = abr$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Factor A	SSA	$a - 1$	$\text{MSA} = \dfrac{\text{SSA}}{a - 1}$	$\dfrac{\text{MSA}}{\text{MSE}}$
Factor B	SSB	$b - 1$	$\text{MSB} = \dfrac{\text{SSB}}{b - 1}$	$\dfrac{\text{MSB}}{\text{MSE}}$
Interaction	SSAB	$(a - 1)(b - 1)$	$\text{MSAB} = \dfrac{\text{SSAB}}{(a - 1)(b - 1)}$	$\dfrac{\text{MSAB}}{\text{MSE}}$
Error	SSE	$ab(r - 1)$	$\text{MSE} = \dfrac{\text{SSE}}{ab(r - 1)}$	
Total	SST	$n_T - 1$		

MULTI-FACTOR ANOVA

Example

```
                 Df Sum Sq Mean Sq F value Pr(>F)
Health            1 1748.0  1748.0 203.094 <2e-16 ***
Location          2   73.9    36.9   4.290  0.016 *
Health:Location   2   26.1    13.1   1.517  0.224
Residuals       114  981.2     8.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# check normality
shapiro.test( residuals(model) )
```

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Depression ~ Health + Location + Health * Location,
          data = Dep)

$Health
              diff       lwr       upr p adj
good-bad -7.633333 -8.694414 -6.572252     0

$Location
                            diff        lwr       upr     p adj
New York-Florida           1.850  0.2921599 3.4078401 0.0155179
North Carolina-Florida     0.475 -1.0828401 2.0328401 0.7497611
North Carolina-New York -1.375 -2.9328401 0.1828401 0.0951631

$`Health:Location`
                                        diff         lwr       upr     p adj
good:Florida-bad:Florida               -8.95 -11.6393115 -6.260689 0.0000000
bad:New York-bad:Florida                0.90  -1.7893115  3.589311 0.9264595
good:New York-bad:Florida              -6.15  -8.8393115 -3.460689 0.0000000
bad:North Carolina-bad:Florida         -0.55  -3.2393115  2.139311 0.9913348
good:North Carolina-bad:Florida        -7.45 -10.1393115 -4.760689 0.0000000
bad:New York-good:Florida               9.85   7.1606885 12.539311 0.0000000
good:New York-good:Florida              2.80   0.1106885  5.489311 0.0361494
bad:North Carolina-good:Florida         8.40   5.7106885 11.089311 0.0000000
good:North Carolina-good:Florida        1.50  -1.1893115  4.189311 0.5892328
good:New York-bad:New York             -7.05  -9.7393115 -4.360689 0.0000000
bad:North Carolina-bad:New York        -1.45  -4.1393115  1.239311 0.6244461
good:North Carolina-bad:New York       -8.35 -11.0393115 -5.660689 0.0000000
bad:North Carolina-good:New York        5.60   2.9106885  8.289311 0.0000003
good:North Carolina-good:New York      -1.30  -3.9893115  1.389311 0.7262066
good:North Carolina-bad:North Carolina -6.90  -9.5893115 -4.210689 0.0000000
```

MULTI-FACTOR ANOVA

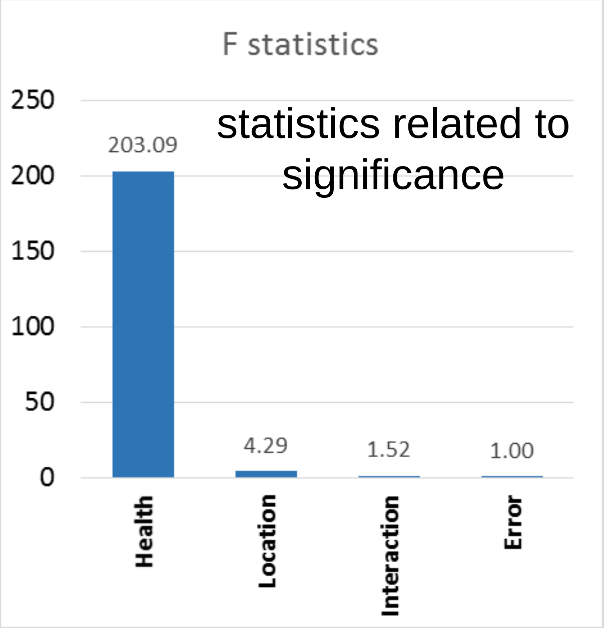
Example & Effect size

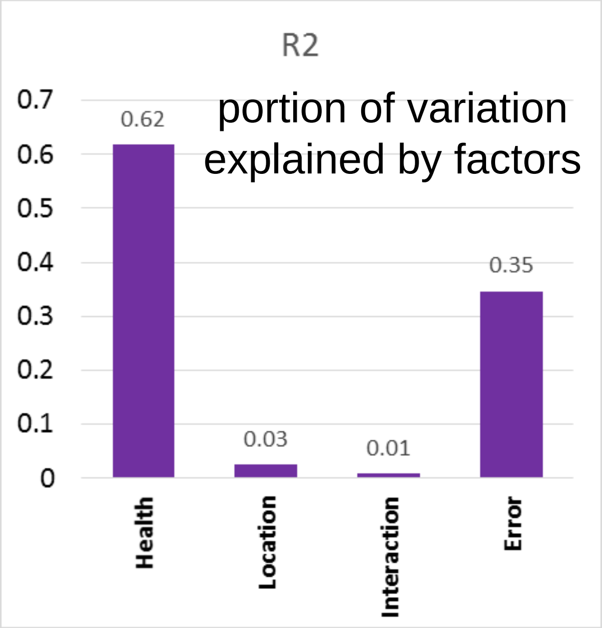
ANOVA

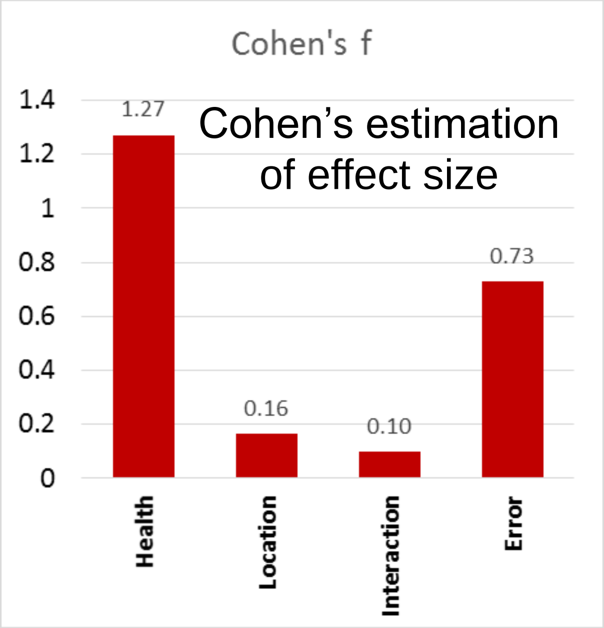
	Source of Variation	SS	df	MS	F	P-value	F crit
Health	Sample	1748.033	1	1748.033	203.094	4.4E-27	3.92433
Location	Columns	73.85	2	36.925	4.290104	0.015981	3.075853
Interaction	Interaction	26.11667	2	13.05833	1.517173	0.223726	3.075853
Error	Within	981.2	114	8.607018			
	Total	2829.2	119				

η^2 or $R^2 = SSx / SST$

$f = \mathrm{sqrt}(R^2 / (1-R^2))$







MULTI-FACTOR ANOVA

Example 2

salaries

Salary/week	Occupation	Gender
872	Financial Manager	Male
859	Financial Manager	Male
1028	Financial Manager	Male
1117	Financial Manager	Male
1019	Financial Manager	Male
519	Financial Manager	Female
702	Financial Manager	Female
805	Financial Manager	Female
558	Financial Manager	Female
591	Financial Manager	Female

Q: Which factors have significant effect on the salary

```
# read dataset
Sal = read.table(
"http://edu.modas.lu/data/txt/salaries.txt",
  header=T,sep="\t",as.is=FALSE)
str(Sal)
# build 2-way ANOVA model
mod = aov(Salary.week ~
  Occupation + Gender + Occupation*Gender, Sal)
summary(mod)
# post-hoc
TukeyHSD(mod)
```

Sourceof Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	221880	1	221880	21.254	0.000112	4.25968
Columns	276560	2	138280	13.246	0.000133	3.40283
Interaction	115440	2	57720	5.5289	0.010595	3.40283
Within	250552	24	10439.7			
Итого	864432	29				

MULTI-FACTOR ANOVA

Example 2

```
                   Df Sum Sq  Mean Sq F value   Pr(>F)
Occupation          2 276560  138280  13.246 0.000133 ***
Gender              1 221880  221880  21.254 0.000112 ***
Occupation:Gender   2 115440   57720   5.529 0.010595 *
Residuals          24 250552   10440
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

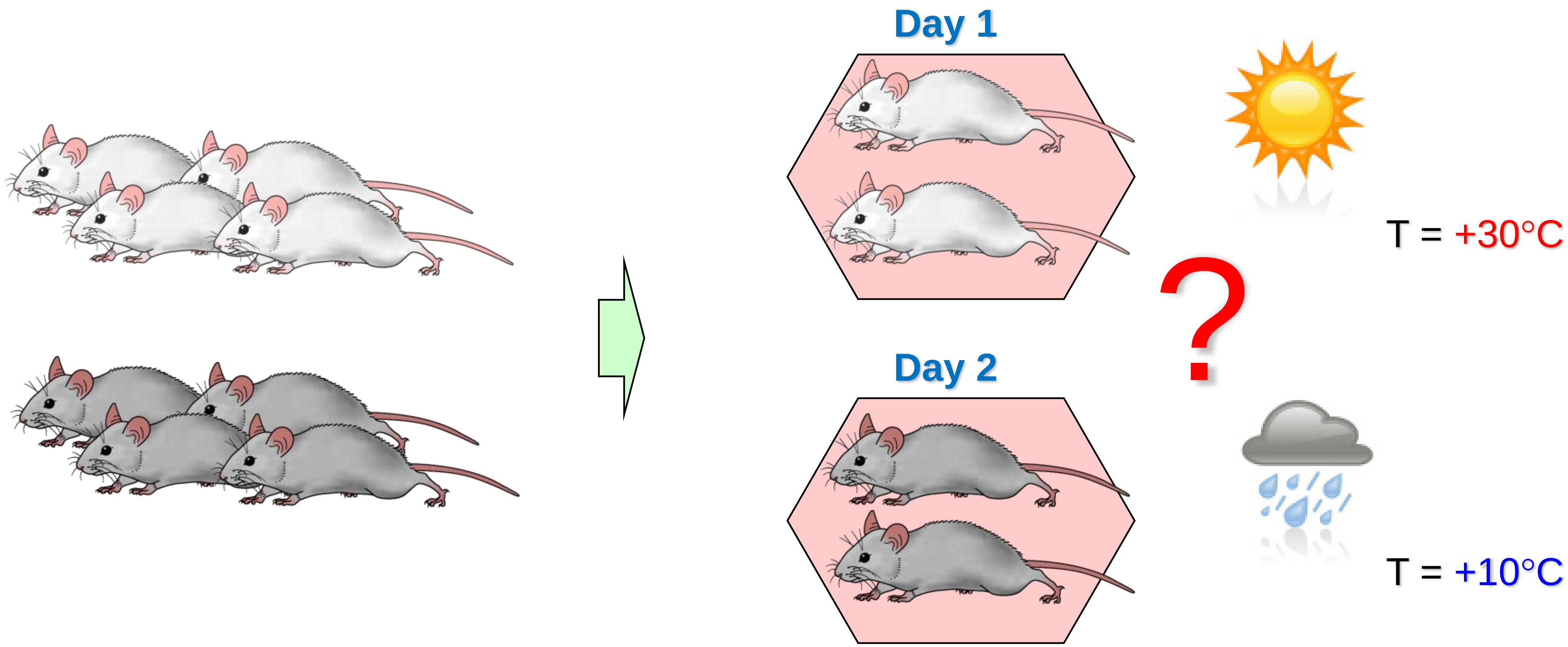
```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Salary.week ~ Occupation + Gender + Occupation * Gender, data = Sal)

$Occupation
                                      diff        lwr      upr     p adj
Financial Manager-Computer Programmer   38 -76.11081 152.1108 0.6874260
Pharmacist-Computer Programmer         220 105.88919 334.1108 0.0001903
Pharmacist-Financial Manager           182  67.88919 296.1108 0.0015387

$Gender
            diff      lwr      upr     p adj
Male-Female  172 94.99818 249.0018 0.0001119

$`Occupation:Gender`
                                                    diff         lwr       upr     p adj
Financial Manager:Female-Computer Programmer:Female -106 -305.80351  93.80351 0.5814961
Pharmacist:Female-Computer Programmer:Female         190   -9.80351 389.80351 0.0689592
Computer Programmer:Male-Computer Programmer:Female   56 -143.80351 255.80351 0.9508750
Financial Manager:Male-Computer Programmer:Female    238   38.19649 437.80351 0.0131635
Pharmacist:Male-Computer Programmer:Female           306  106.19649 505.80351 0.0010255
Pharmacist:Female-Financial Manager:Female           296   96.19649 495.80351 0.0015025
Computer Programmer:Male-Financial Manager:Female    162  -37.80351 361.80351 0.1616324
Financial Manager:Male-Financial Manager:Female      344  144.19649 543.80351 0.0002396
Pharmacist:Male-Financial Manager:Female             412  212.19649 611.80351 0.0000185
Computer Programmer:Male-Pharmacist:Female          -134 -333.80351  65.80351 0.3334443
Financial Manager:Male-Pharmacist:Female              48 -151.80351 247.80351 0.9743050
Pharmacist:Male-Pharmacist:Female                    116  -83.80351 315.80351 0.4872344
Financial Manager:Male-Computer Programmer:Male      182  -17.80351 381.80351 0.0889147
Pharmacist:Male-Computer Programmer:Male             250   50.19649 449.80351 0.0084855
Pharmacist:Male-Financial Manager:Male                68 -131.80351 267.80351 0.8950589
```

EXPERIMENTAL DESIGN

Experiments

Aware of Batch Effect !

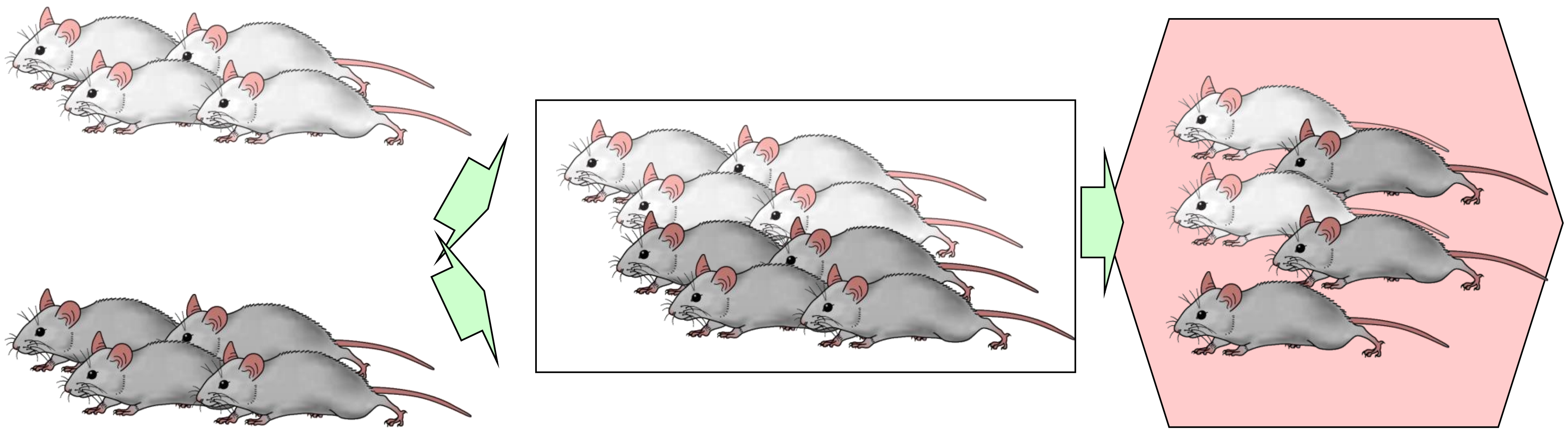
When designing your experiment always remember about various factors which can effect your data: batch effect, personal effect, lab effect...



EXPERIMENTAL DESIGN

Experiments

Completely randomized design
An experimental design in which the treatments are randomly assigned to the experimental units.



We can nicely randomize:

Day effect

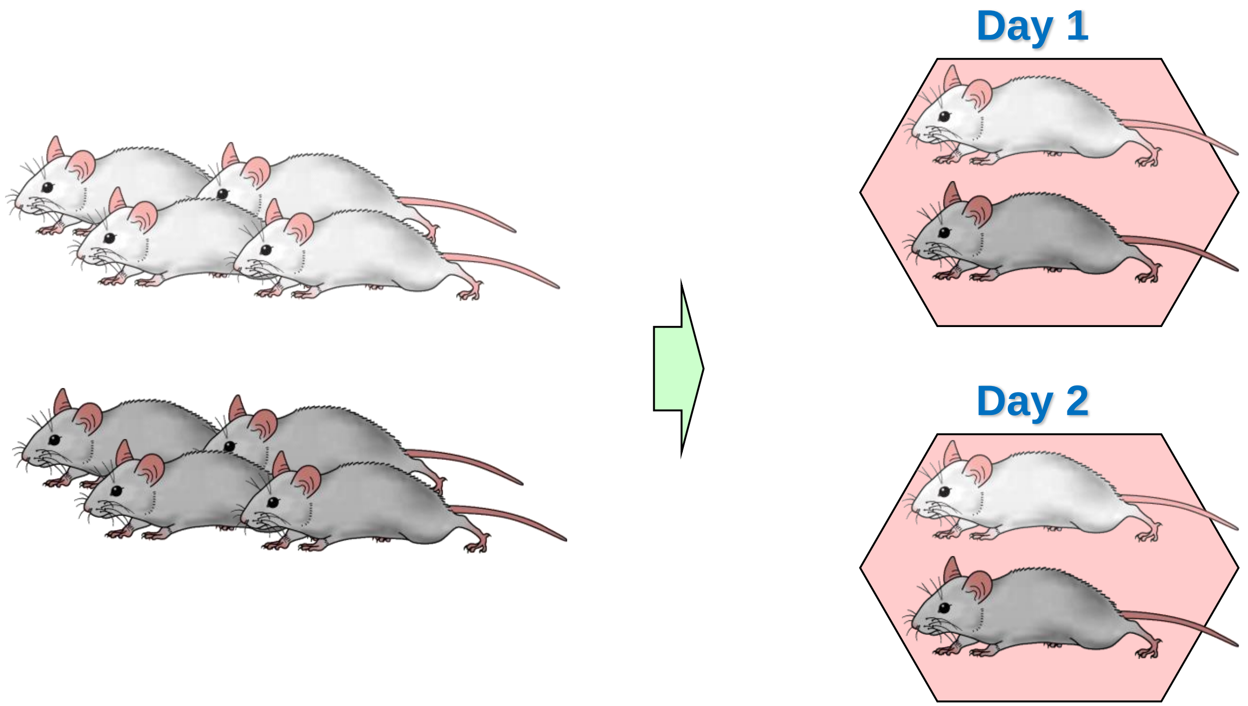
Batch effect

EXPERIMENTAL DESIGN

Experiments

Blocking

The process of using the same or similar experimental units for all treatments. The purpose of blocking is to remove a source of variation from the error term and hence provide a more powerful test for a difference in population or treatment means.



EXPERIMENTAL DESIGN

Experiments

A good suggestion… ☺

Block what you can block, **randomize** what you cannot, and try to **avoid** unnecessary factors

ANOVA

Task

mice

Q: Does mouse strain affect the weight (e.g. Starting weight)? Show the effects of **sex** and **strain** using ANOVA

		129S1/SvImJ	A/J	AKR/J	BALB/cByJ	BTBR_T+_	BUB/BnJ	C3H/HeJ
1	Female	20.5	23.2	24.6	22.8	28	27.1	21.4
2		20.8	22.4	26	23.5	25.8	24.1	28.2
3		19.8	22.7	31	23.8	26	25.9	23.5
4		21	21.4	25.7	22.7	26.5	25.9	23.9
5		21.9	22.6	23.7	19.7	26.3	26	22.8
6		22.1	20	21.1	26.2	27	27.1	18.4
7		21.3	21.8	23.7	24.1	26	26.2	21.8
8		20.1	20.8	24.5	23.5	28.8	27.5	25
9		18.9	19.5	32.3	23.8	28	30.2	20.1
10	Male	24.7	25.8	42.8	29.3	34.1	36.2	31.2
11		27.2	27.7	32.6	32.2	33	36.9	28.2
12		23.9	29.9	34.8	29.7	38.7	34.4	26.7
13		26.3	24.8	32.8	30	39	34.3	29.3
14		26	22.9	34.8	27	31	31.7	33.1
15		23.3	24.5	32.8	30	32	33	28.2
16		26.5	24.6	33.6	33.1	33.7	33.2	31.2
17		27.4	21.6	30.7	30.6	33.1	34	27.7
18		27.5	26.9	36.5	28.7	32.5	31	27.5

HYPOTHESES FOR VARIANCE

Confidence intervals for variance

Hypotheses for variance

Goodness of fit, test for independence

ANalysis Of VAriance (ANOVA)

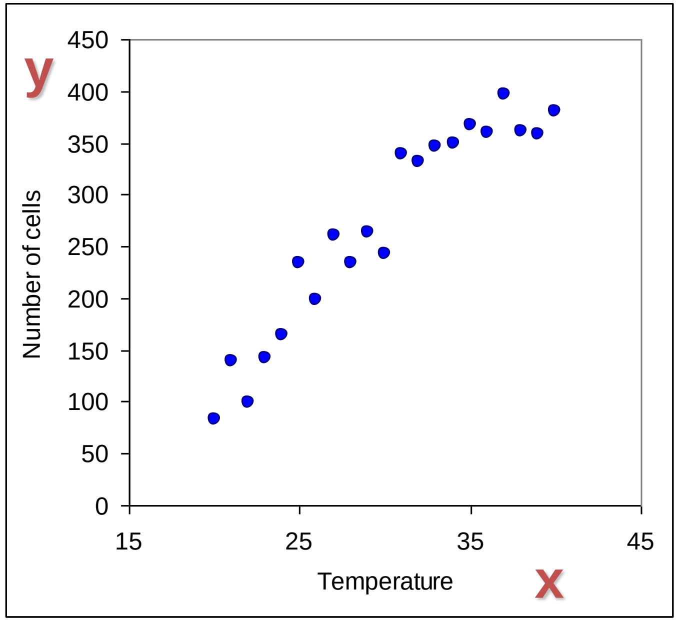
Linear regression

Logistic regression

SIMPLE LINEAR REGRESSION

Example

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.

cells

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

str(Cells)

plot(Cells, pch=19)
```

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

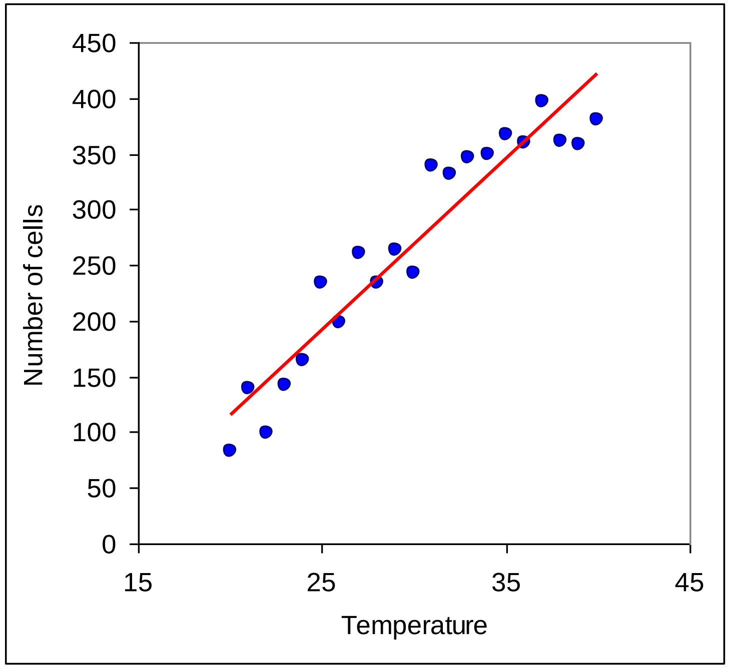
Independent variable
The variable that is doing the predicting or explaining. It is denoted by *x*.

SIMPLE LINEAR REGRESSION

Regression Model and Regression Line

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



SIMPLE LINEAR REGRESSION

Regression Model and Regression Line

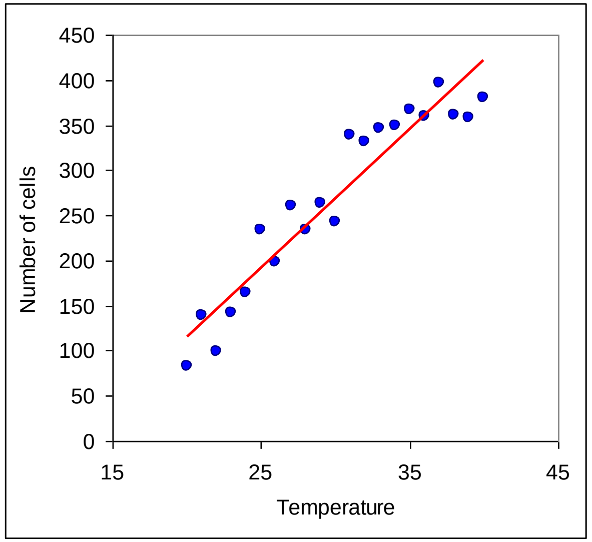
Regression model

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,

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



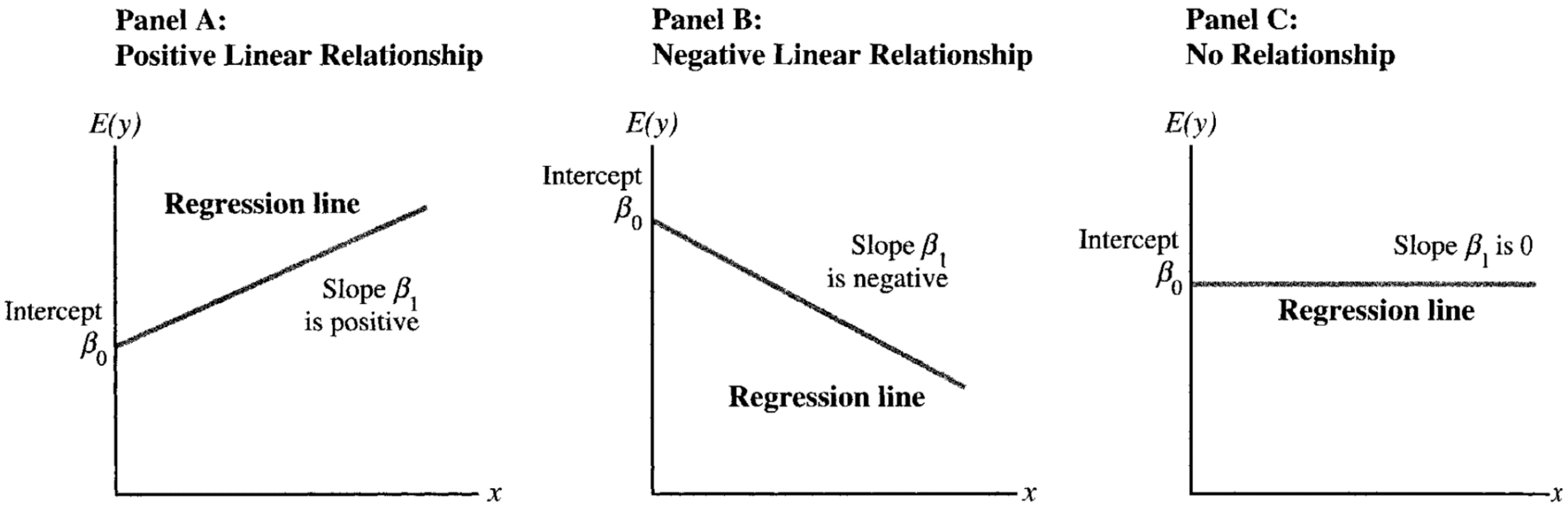
- Model for a simple linear regression:

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

SIMPLE LINEAR REGRESSION

Regression Model and Regression Line

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



SIMPLE LINEAR REGRESSION

Estimated regression equation

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 $y = b_0 + b_1 x$

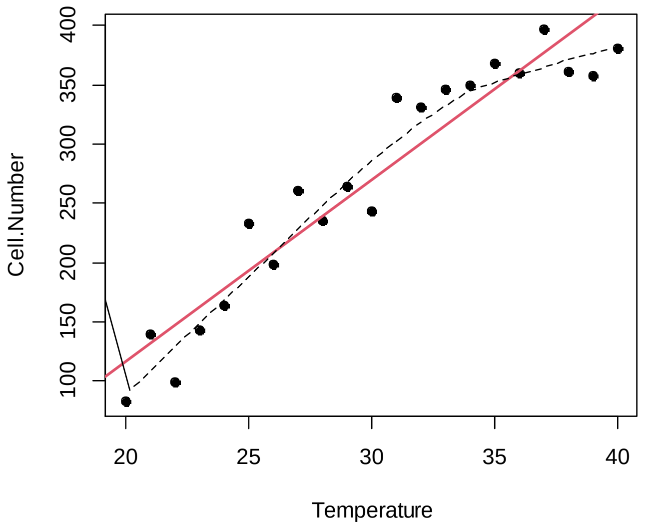
$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$

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

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

`cells`

```
plot(Cells, pch=19)
abline(lm(Cell.Number ~ Temperature, Cells),col=2, lwd=2)

# add smooth curve (loess/lowess) (just fun)
lines(lowess(Cells$Temperature, Cells$Cell.Number),lty=2)
```



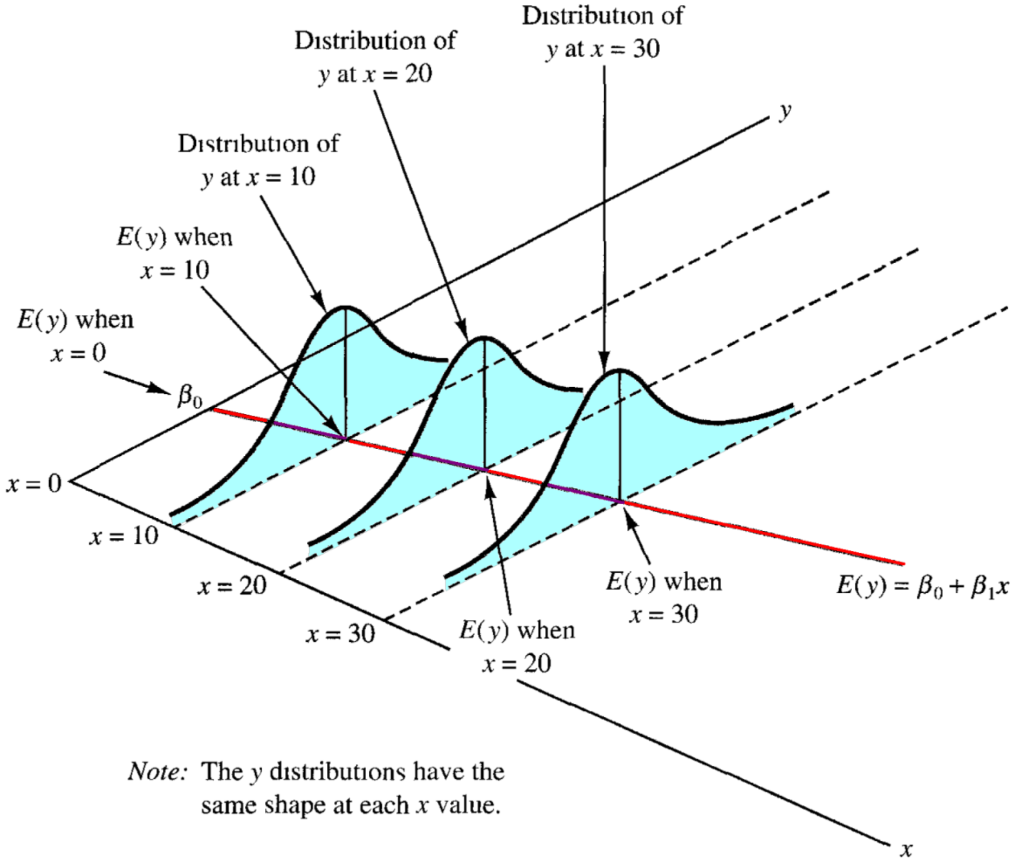
LINEAR REGRESSION

Assumptions

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.

SIMPLE LINEAR REGRESSION

Exact calculation for the simplest case

Least squares method

A procedure used to develop the estimated regression equation.

The objective is to minimize $\sum (y_i - \hat{y}_i)^2$

y_i = observed value of the dependent variable for the ith observation
$\hat{y}_i$ = estimated value of the dependent variable for the ith observation

Slope: $$b_1 = \frac{\sum (x_i - m_x)(y_i - m_y)}{(x_1 - m_x)^2}$$

Intersect: $$b_0 = m_y - b_1 m_x$$

SIMPLE LINEAR REGRESSION

The Main Equation

Sum squares due to error

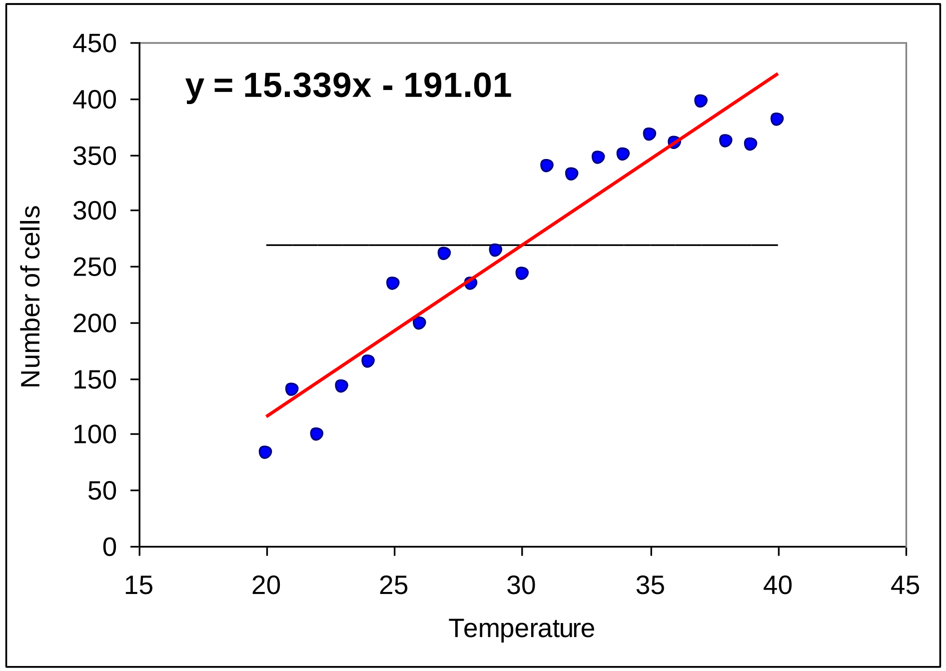
distances from ● to – $$SSE = \sum (y_i - \hat{y}_i)^2$$

Sum squares total

distances from ● to – $$SST = \sum (y_i - m_y)^2$$

Sum squares due to regression

distances from – to – $$SSR = \sum (\hat{y}_i - m_y)^2$$

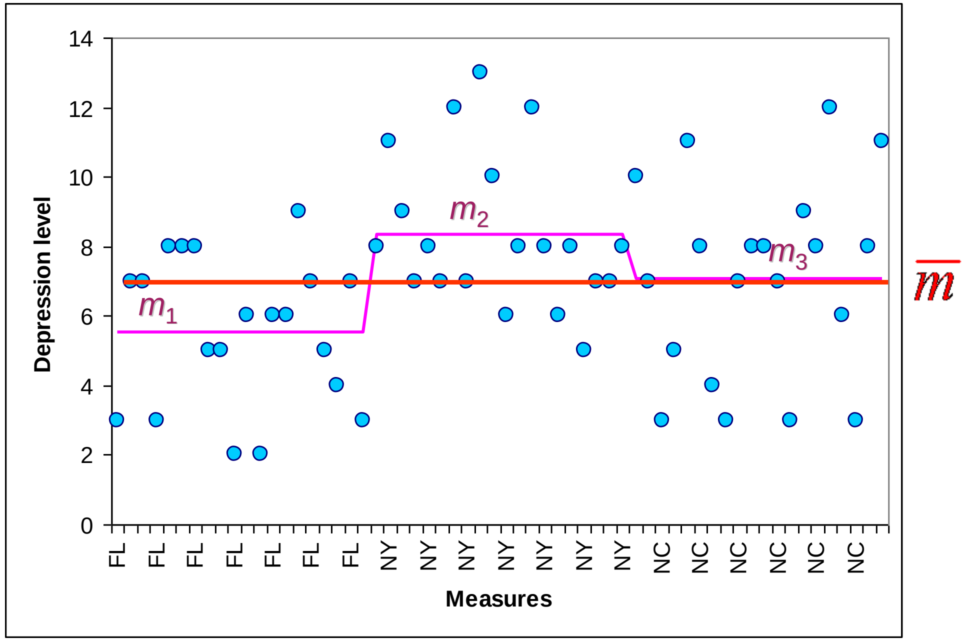


The Main Equation

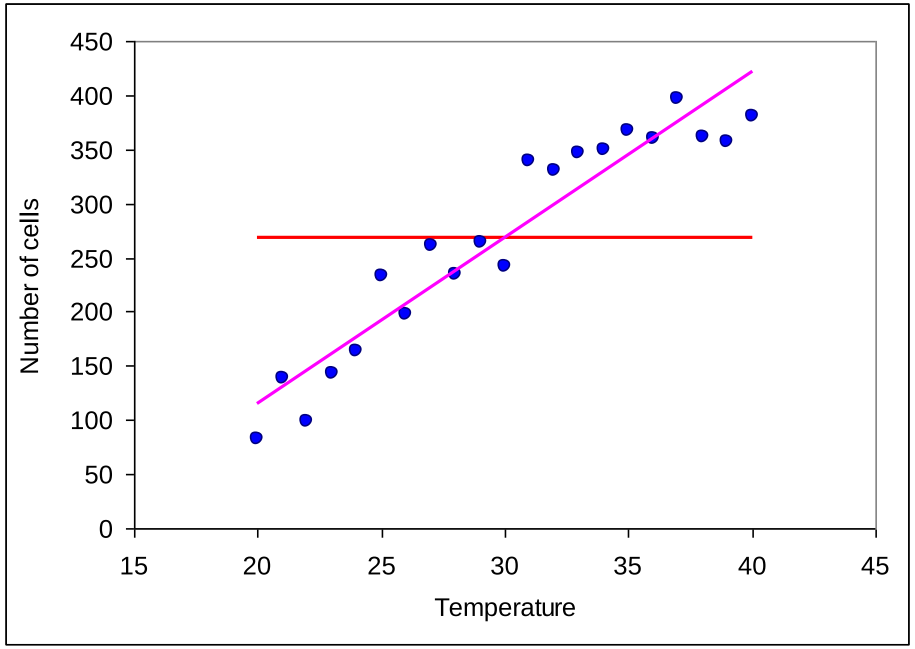
$$\boxed{SST = SSR + SSE}$$

SIMPLE LINEAR REGRESSION

ANOVA and Regression



$$SST = SSTR + SSE$$



$$SST = SSR + SSE$$

SIMPLE LINEAR REGRESSION

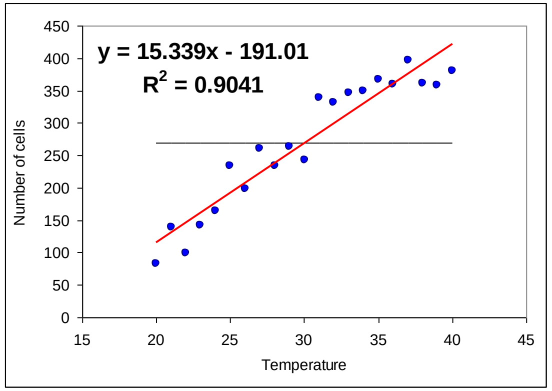
Coefficient of Determination

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

$$SST = \sum (y_i - m_y)^2$$

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

$$SST = SSR + SSE$$



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.

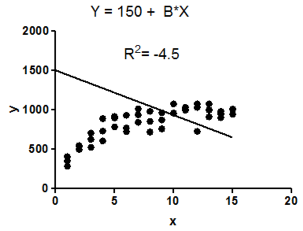
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Correlation coefficient

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

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

NOTE: There is a non-obvious case when $R^2 < 0$.
It means that the model is worse than the mean value

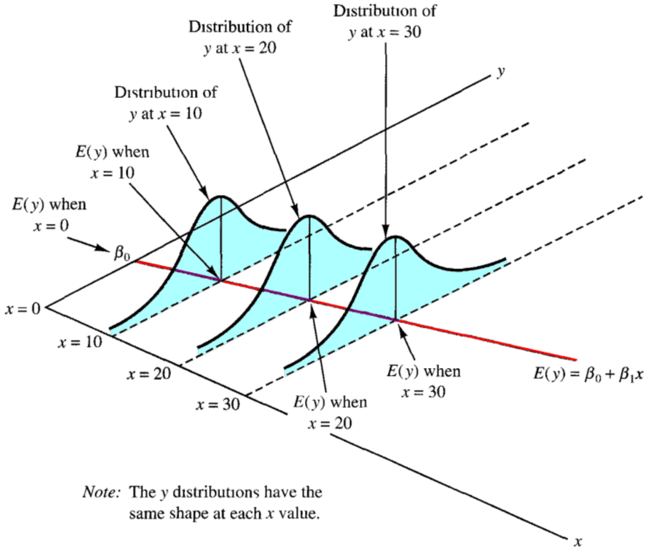


TESTING FOR SIGNIFICANCE

Estimation of σ^2

***i*-th residual**
The difference between the observed value of the dependent variable and the value predicted using the estimated regression equation; for the *i*-th observation the *i*-th residual is: $y_i - \hat{y}_i$

Mean square error
The unbiased estimate of the variance of the error term σ^2. It is denoted by MSE or s^2.
Standard error of the estimate: the square root of the mean square error, denoted by *s*. It is the estimate of σ, the standard deviation of the error term ε.



$$s^2 = MSE = \frac{SSE}{n-2}$$

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

TESTING FOR SIGNIFICANCE

Sampling Distribution for b_1

If assumptions for ε are fulfilled, then the sampling distribution for b_1 is as follows:

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

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

Expected value $E[b_1] = \beta_1$

St.deviatiation $\sigma_{b_1} = \dfrac{\sigma}{\sqrt{\sum (x_i - m_x)^2}}$ = Standard Error

Distribution: ***normal***

Interval Estimation for β_1

$$\beta_1 = b_1 \pm t_{\alpha/2}^{(n-2)} \frac{\sigma}{\sqrt{\sum (x_i - m_x)^2}}$$

$$\beta_1 = b_1 \pm t_{\alpha/2}^{(n-2)} SE$$

TESTING FOR SIGNIFICANCE

2 Ways to Test for Significance

H_0: $\beta_1 = 0$ *insignificant*

H_a: $\beta_1 \neq 0$

1. Build a t-test statistics.

$$t = \frac{b_1}{\sigma_{b_1}} = \frac{b_1}{s}\sqrt{\sum (x_i - m_x)^2}$$

2. Calculate p-value for *t*

p-value approach: Reject H_0 if p-value $\leq \alpha$

Critical value approach: Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

1. Build a F-test statistics.

$$F = \frac{MSR}{MSE}$$

$$\text{MSR} = \frac{\text{SSR}}{\text{Number of independent variables}}$$

2. Calculate a p-value

REGRESSION ANALYSIS

Example

cells

In R you should run the complete analysis:

```
model=lm(Cell.Number~Temperature, data=Cells)
```

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

```
# Regression table
summary(model)

# ANOVA table
anova(model)

# intercept/slope
model$coefficients
```

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.95091908
R Square	0.9042471
Adjusted R Square	0.89920747
Standard Error	31.7623796
Observations	21

ANOVA

	df	SS	MS	F	Significance F
Regression	1	181015.1117	181015.11	179.4274	3.95809E-11
Residual	19	19168.12641	1008.8488		
Total	20	200183.2381			

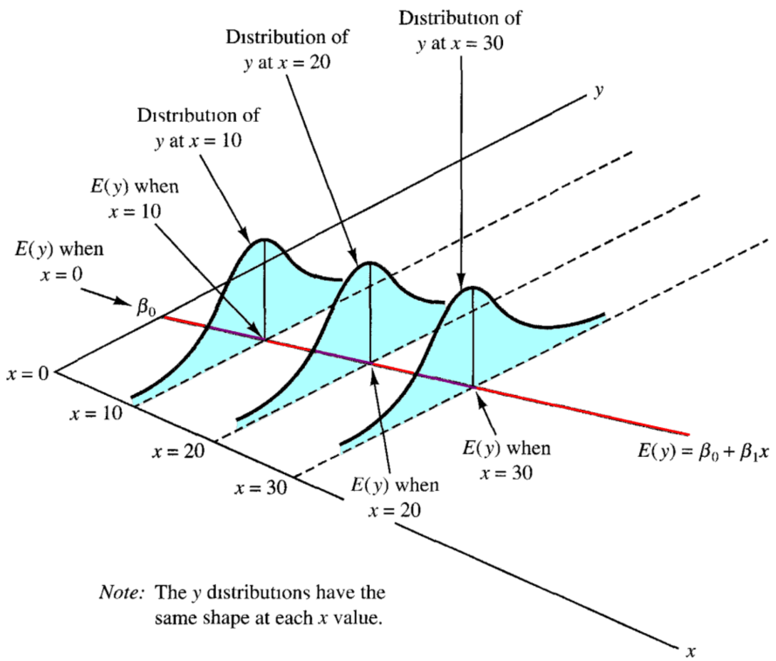
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-190.783550	35.031618	-5.446039	2.96E-05	-264.10557	-117.46153	-264.10557	-117.46153
Temperature	15.332468	1.144637	13.395051	3.96E-11	12.93671537	17.7282197	12.93671537	17.7282197

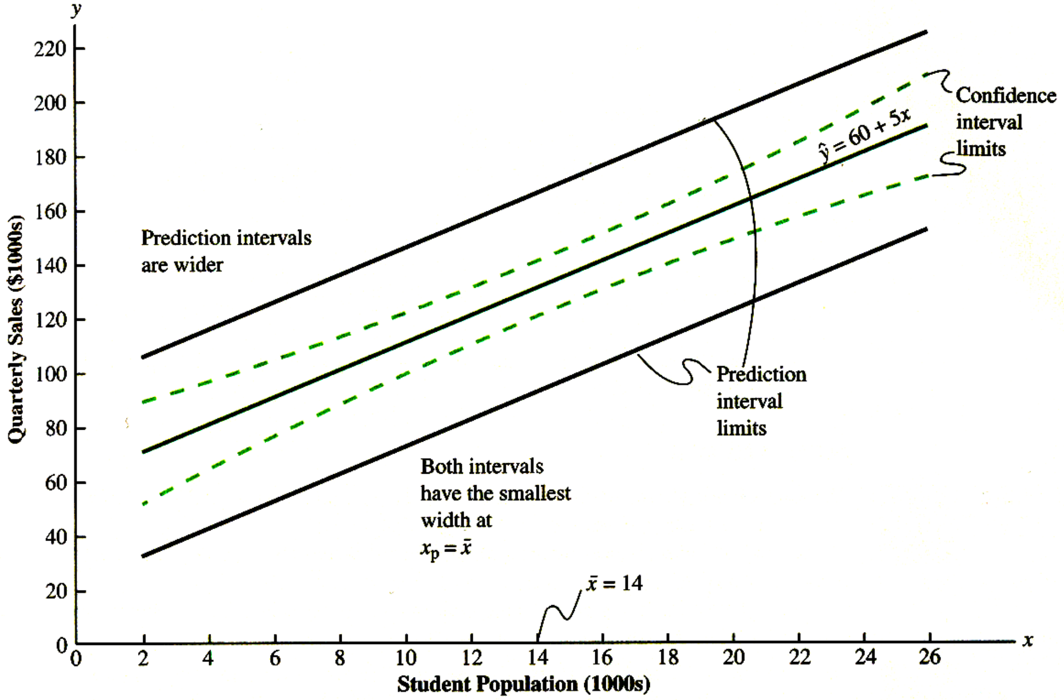
REGRESSION ANALYSIS

Confidence and Prediction

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.



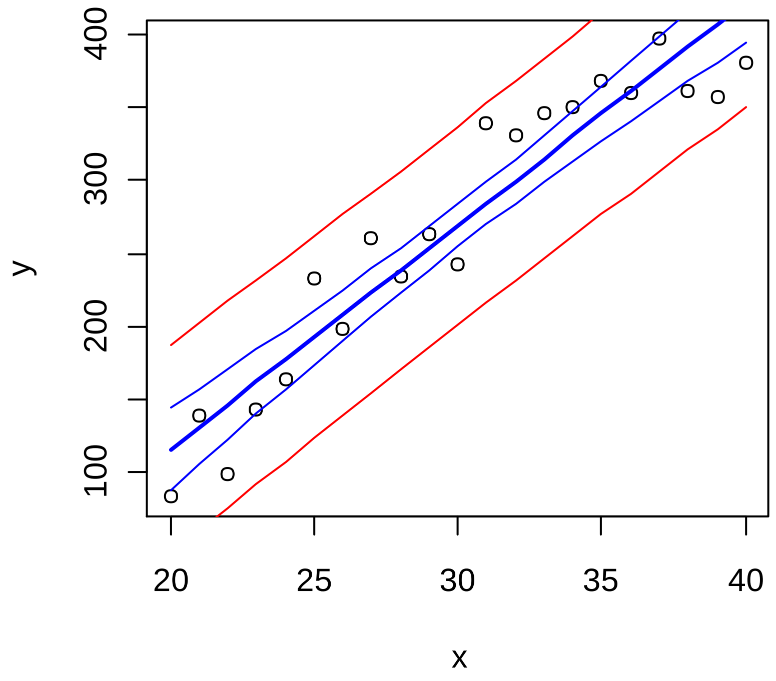


REGRESSION ANALYSIS

Example

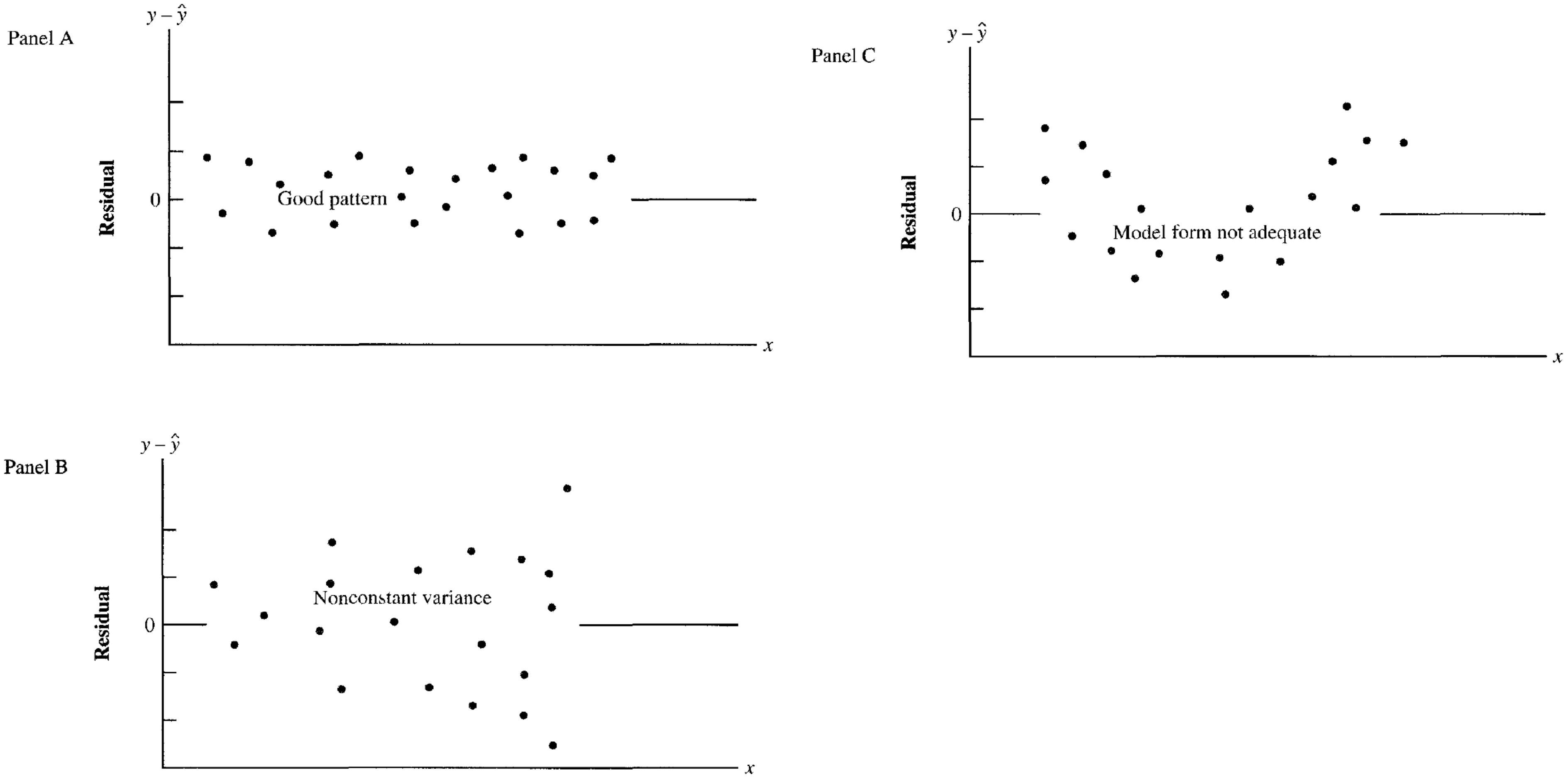
cells

```
x = data$Temperature
y = data$Cell.Number
res = lm(y~x)
res
summary(res)

# draw the data
x11()
plot(x,y)
# draw the regression and its confidence (95%)
lines(x, predict(res,int = "confidence")[,1],col=4,lwd=2)
lines(x, predict(res,int = "confidence")[,2],col=4)
lines(x, predict(res,int = "confidence")[,3],col=4)
# draw the prediction for the values (95%)
lines(x, predict(res,int = "pred")[,2],col=2)
lines(x, predict(res,int = "pred")[,3],col=2)
```



REGRESSION ANALYSIS

Residuals



REGRESSION ANALYSIS

Task

rana

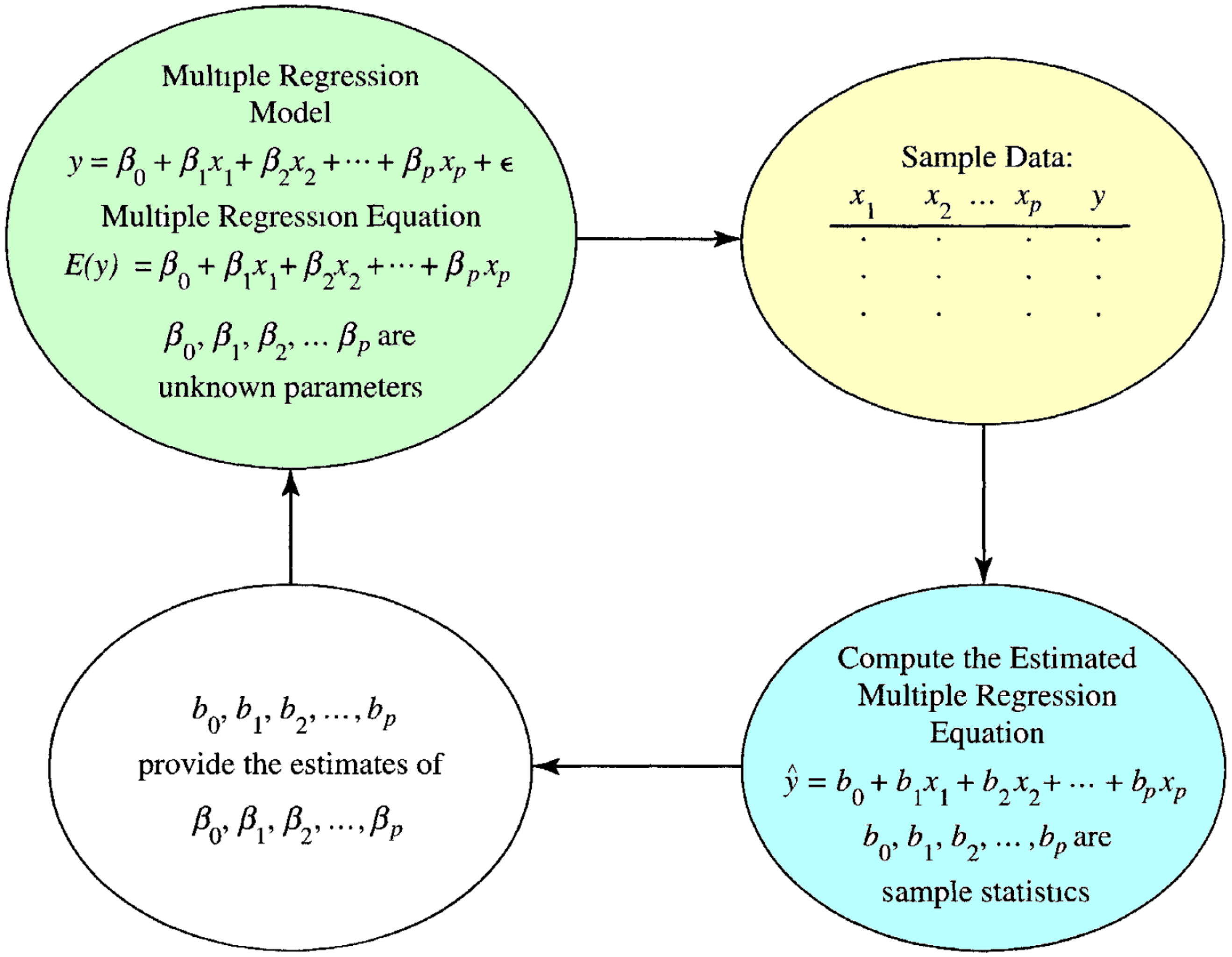
A biology student wishes to determine the relationship between temperature and heart rate in leopard frog, *Rana pipiens*. He manipulates the temperature in 2° increment ranging from 2 to 18°C and records the heart rate at each interval. His data are presented in table rana.txt

1) Build the model and provide the p-value for linear dependency

2) Provide interval estimation for the slope of the dependency

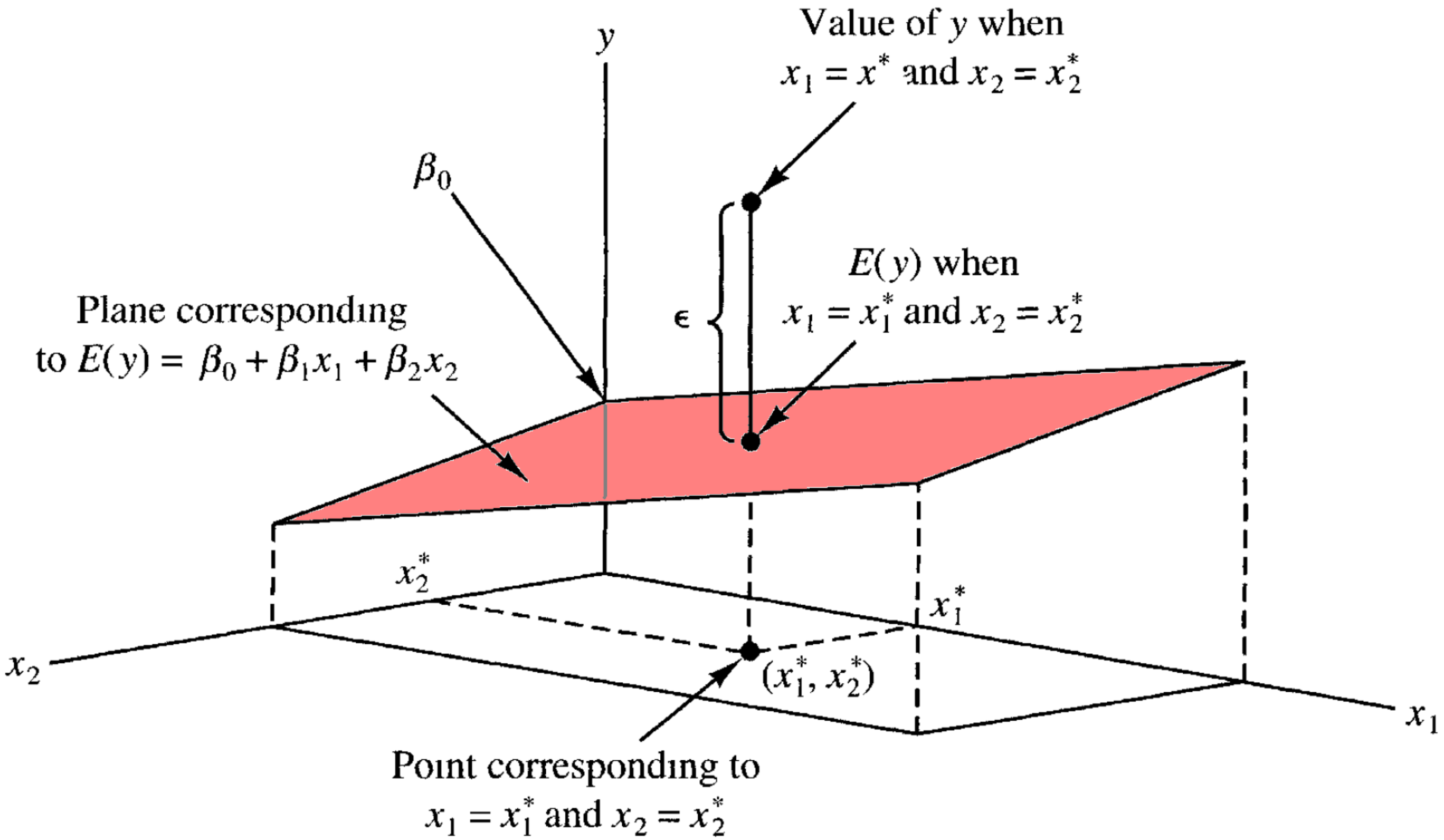
3) Estimate 95% prediction interval for heart rate at 15°

Multiple Regression



REGRESSION ANALYSIS

Multiple Regression



MULTIPLE REGRESSION

Example: swiss dataset

swiss

Often one variable is not enough, and we need several independent variables to predict dependent one. Let's consider R internal swiss dataset: standardized fertility measure and socio-economic indicators for 47 French-speaking provinces of Switzerland at about 1888. See **?swiss**

```
## 'data.frame':    47 obs. of  6 variables:
##  $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
##  $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
##  $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
##  $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
##  $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
##  $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

```
#install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)
chart.Correlation(swiss)
modAll = lm(Fertility ~ . , data = swiss)
summary(modAll)

plot(swiss$Fertility, predict(modAll,swiss),xlab="Real
Fertility",ylab="Predicted Fertility",pch=19)
abline(a=0,b=1,col=2,lty=2)
```

Check further analysis in the HTML…

MULTIPLE REGRESSION

Recommendations

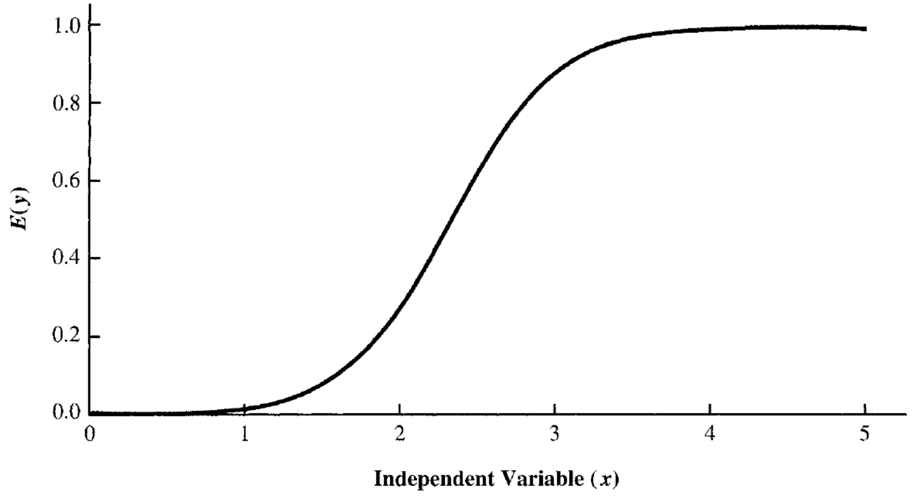
swiss

- Check whether your linear model is adequate (visualize residual, draw **lowess** curve)
- Check the significance of the variables
- Check and try to avoid correlated variables
- If you need to choose optimal variables:
  - maximize R^2
  - minimize information criteria: BIC and AIC
- Add / remove variable and compare models using likelihood ratio or chi2 test.
  - anova(modAll, modSig)

REGRESSION ANALYSIS

Logistic Regression

FIGURE 15.12 LOGISTIC REGRESSION EQUATION FOR $\beta_0 = -7$ AND $\beta_1 = 3$



$$E(y) = P\left(y = 1 \mid x_1, x_2, \ldots, x_p\right) = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right)}$$

Example:

in **R**: glm(..., family="binomial")

```
Mice = read.table(
"http://edu.modas.lu/data/txt/mice.txt",
header=T,sep="\t",as.is=FALSE)
str(Mice)
## let's remove animals with NA values
ikeep = apply(is.na(Mice),1,sum) == 0

model = glm( Sex ~ Blood.pH +
Bone.mineral.density + Lean.tissues.weight
+ Ending.weight,
          data = Mice[ikeep,],
          family = "binomial")

summary(model)
```

http://edu.modas.lu/modas_pm/part2.html

To be continued in Lecture 4…

QUESTIONS ?

Thank you for your attention