STK4900/9900 - Lecture 2

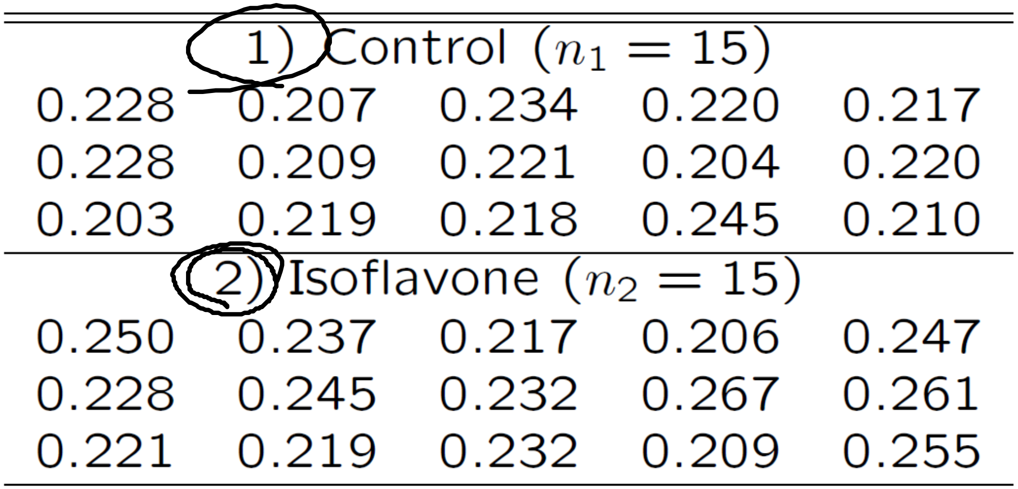
Program

1. Comparing two or more groups ✓
2. One-way analysis of variance (ANOVA) ✓
3. Multiple testing and FDR
4. Covariance and correlation
5. Simple linear regression

- Section 13.4.1
- Section 2.4
- Sections 3.1.4, 3.2 (not 3.2.2), 3.3
- Supplementary material on FDR, covariance, correlation and one-way ANOVA

Comparing two groups

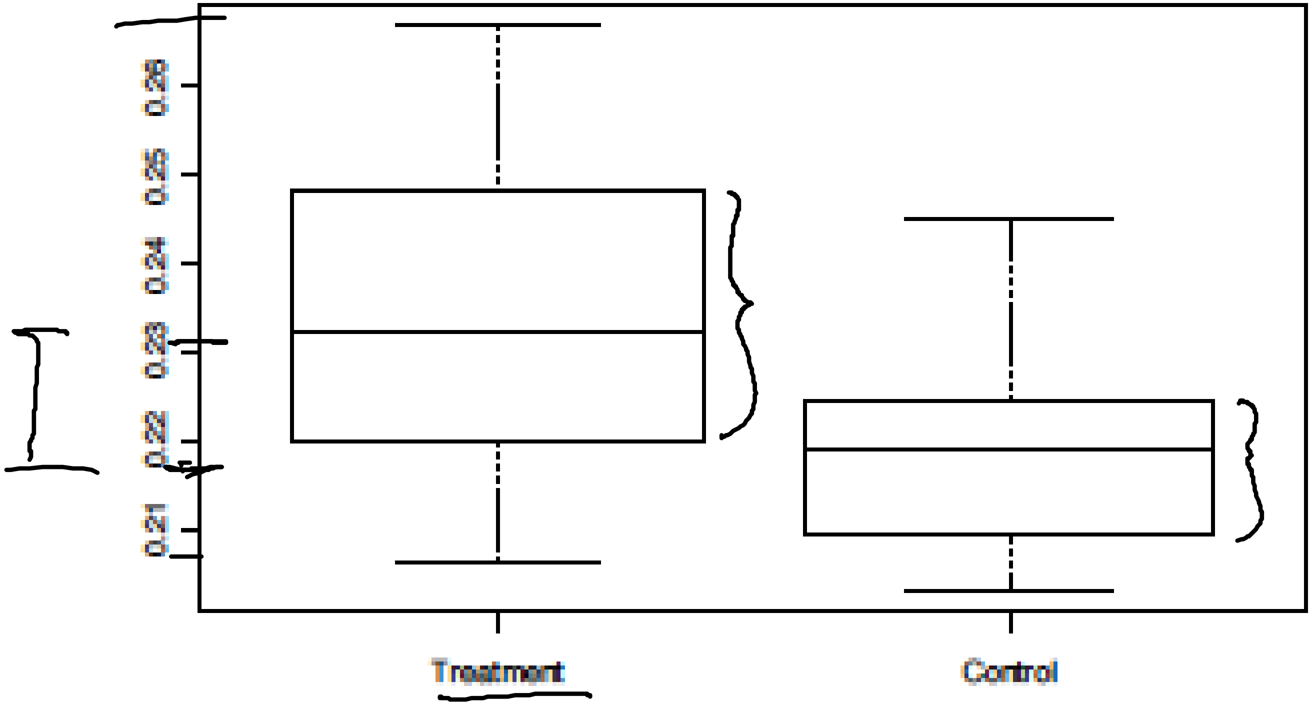
In Lecture 1 we considered an example where we measured bone mineral density (in g/cm^2) for rats given isoflavone and for rats in a control group:



1) Control ($n_1 = 15$)				
0.228	0.207	0.234	0.220	0.217
0.228	0.209	0.221	0.204	0.220
0.203	0.219	0.218	0.245	0.210
2) Isoflavone ($n_2 = 15$)				
0.250	0.237	0.217	0.206	0.247
0.228	0.245	0.232	0.267	0.261
0.221	0.219	0.232	0.209	0.255

Question: Does isoflavone have an effect on bone mineral density?

A boxplot gives a graphical comparison of the two groups:



We would like to determine a confidence interval for the treatment effect and test if the difference is statistically significant (cf. next slide)

R-commands:

```
cont=c(0.228, 0.207, 0.234, 0.220, 0.217, 0.228, 0.209, 0.221, 0.204, 0.220,
       0.203, 0.219, 0.218, 0.245, 0.210)
treat=c(0.250, 0.237, 0.217, 0.206, 0.247, 0.228, 0.245, 0.232, 0.267, 0.261,
       0.221, 0.219, 0.232, 0.209, 0.255)
boxplot(treat, cont,names=c("Treatment","Control"))
t.test(treat, cont , var.equal=TRUE)
```

$\bar{x}_2-\bar{x}_1 \pm c\,se(\bar{x}_2-\bar{x}_1)$

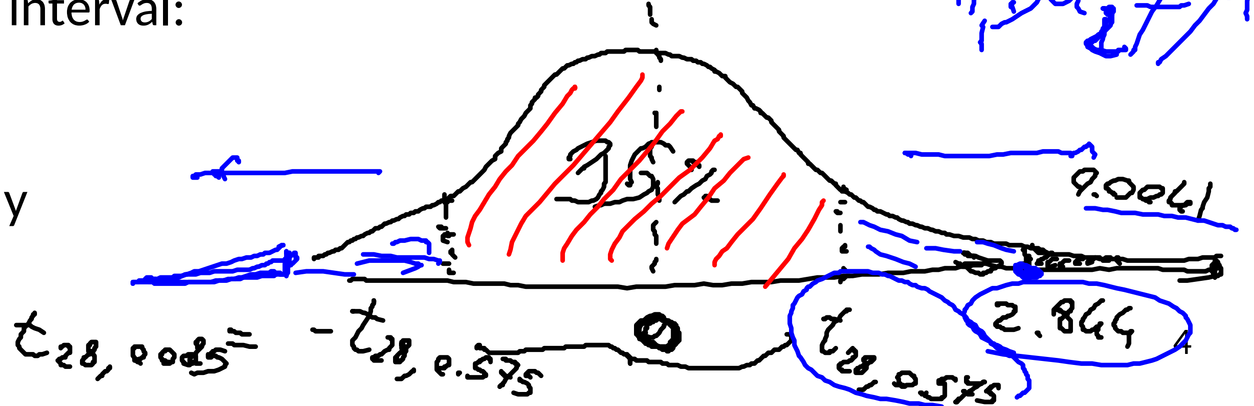
R-output (slightly edited)

$$t=\frac{\bar{x}_2-\bar{x}_1}{se(\bar{x}_2-\bar{x}_1)} \sim t_{n_1+n_2-2}$$

```
Two Sample t-test

data:  treat and cont
t = 2.844,  df = 28,  p-value = 0.0082
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0045   0.0279
sample estimates:
mean of x      mean of y
0.2351         0.2189
```

$\boxed{\mu_2=\mu_1}$

$H_1: \mu_2 \neq \mu_1$



Suppose that the data for the two groups are random samples from $\underline{N(\mu_1, \sigma^2)}$ and $\underline{N(\mu_2, \sigma^2)}$, respectively

Consider testing the null hypothesis $\underline{H_0: \mu_1 = \mu_2}$ versus the alternative $\underline{H_A: \mu_1 \neq \mu_2}$

Test statistic:

$$\underline{t = \frac{\bar{x}_2 - \bar{x}_1}{se(\bar{x}_2 - \bar{x}_1)}}$$

where

$$\underline{se(\bar{x}_2 - \bar{x}_1) = s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

with

$$s_p = \sqrt{\frac{n_1 - 1}{n_1 + n_2 - 2} s_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} s_2^2}$$

We reject $\underline{H_0}$ for large values of $\underline{|t|}$

P-value (two-sided) : $P = \underline{2\,P(T > |t|)},$

where T is t-distributed with $n_1 + n_2 - 2$ df.

Comparing more than two groups: one-way ANOVA

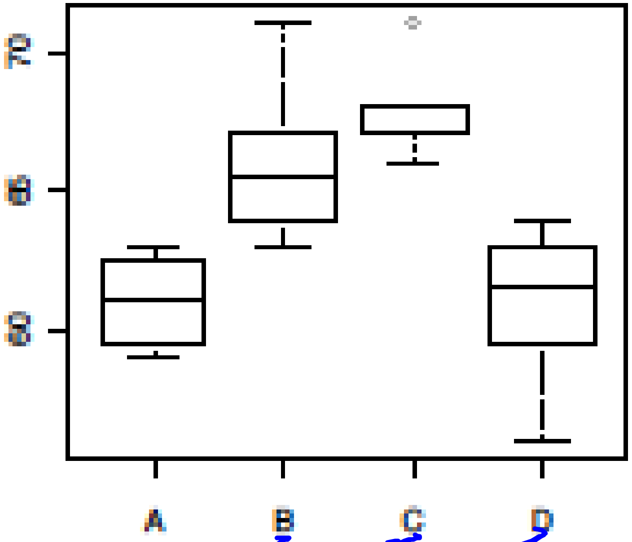
In an experiment 24 rats were randomly allocated to four different diets, and the blood coagulation time (in seconds) was measured for each animal

$N(\mu_A, \sigma_A^2)$

?

$\sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$

Diets (treatment)			
A	B	C	D
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
	65	68	63
	66	68	64
			63
			59



A vs B
A vs C
A vs D
B vs C
B vs D
C vs D

Question: Does diet have an effect on coagulation time?

We may compare two and two diets, using two sample procedures

We would, however, also like to have an overall test

$k = 4$

In general we have observations from *K* groups:

x_{ik} =observation number i in group k

$(i = 1, \ldots, n_k \qquad k = 1, \ldots, K)$

1, 2, 3, 4 :-> 8 A, B, C, D

We assume that all observations are independent and that the observations from group k are a random sample from $N(\mu_k, \sigma^2)$

Notation:

$\bar{X}$

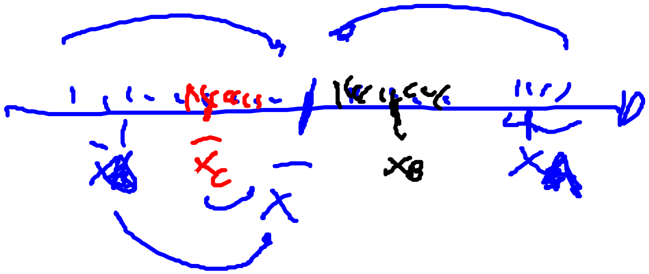
Total number of observations: $n = \sum_k n_k$

Mean in group k: $\bar{x}_k = \frac{1}{n_k} \sum_i x_{ik}$

Overall mean: $\bar{x} = \frac{1}{n} \sum_{i,k} x_{ik} = \frac{1}{n} \sum_k n_k \, \bar{x}_k$

$= \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} x_{ik}$

We want to test the null hypothesis $H_0 : \mu_1 = \ldots\ldots = \mu_K$ versus the alternative that *not* all the μ_k are equal



Introduce the sums of squares:

$$TSS = \sum_{i,k} (x_{ik} - \bar{x})^2$$

(total sum of squares)

$$MSS = \sum_{k} n_k (\bar{x}_k - \bar{x})^2$$

(model sum of squares)

$$RSS = \sum_{i,k} (x_{ik} - \bar{x}_k)^2$$

(residual sum of squares)

Important decomposition:

$$TSS = MSS + RSS$$

Unbiased estimator of σ^2 :

$$s^2 = RSS/(n - K)$$

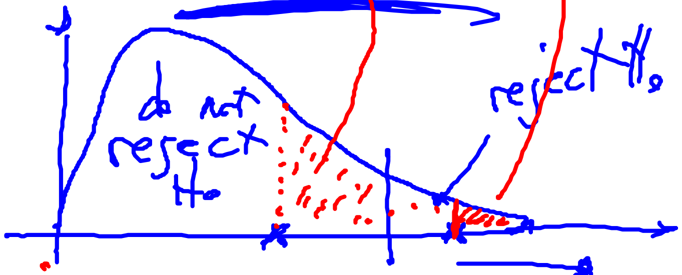
Under the null hypothesis σ^2 may also be estimated by :

$$MSS/(K - 1)$$

However, when the null hypothesis does not hold, the latter estimate tends to be larger than σ^2

We reject the null hypothesis for large values of the test statistic

$$F = \frac{MSS/(K-1)}{RSS/(n-K)}$$



The test statistic is F-distributed with *K* – 1 and *n* – *K* degrees of freedom under the null hypothesis

This result is used to compute the P-value

The result may be summarized in an ANOVA table:

Source	df	Sum of squares	Mean sum of squares	F statistic	P-value
Model	$K-1$	MSS	$MSS/(K-1)$	$F=\dfrac{MSS/(K-1)}{RSS/(n-K)}$	→ P
Residual	$n-K$	RSS	$RSS/(n-K)$		
Total	$n-1$	TSS			

The P-value is found by:

$$P = P(\overbrace{F > \text{observed value of } F}^{F^{df}})$$

where F is F-distributed with $K-1$ and $n-K$ degrees of freedom

In Lecture 3 we will see how one-way ANOVA is a special case of multiple linear regression

R commands for coagulation times:

```
rats=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/data/
      rats.txt",header=T)
rats$diet=factor(rats$diet)     # defines diet to be a categorical variable
aov.rats=aov(time~diet,data=rats)
summary(aov.rats)
```

$K=4$

$n=24$

$H_0: \mu_A = \mu_B = \mu_C = \mu_D$

H_A: at least one different

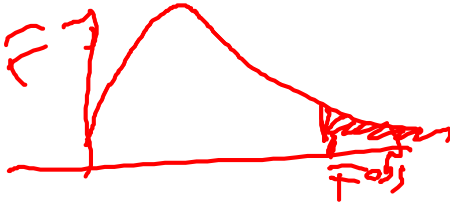
R output (edited):

```
            Df     Sum Sq      Mean Sq   F value     Pr(>F)
diet         3        228         76.0      13.6    4.7e-05
Residuals   20        112          5.6
```

total 23

Relation to two-sample t-test (two-sided)

Consider the situation with two groups, i.e. $K = 2$



Will test the null hypothesis $H_0: \mu_1 = \mu_2$ versus the alternative hypothesis $H_A: \mu_1 \neq \mu_2$

t-test statistic:

$$t = \frac{\bar{x}_2 - \bar{x}_1}{se(\bar{x}_2 - \bar{x}_1)} \longleftrightarrow F = \frac{MSS/(K-1)}{RSS(n-K)}$$

We reject H$_0$ for large values of $|t|$

$K = 2$

We may show that

$$t^2 = \frac{MSS/(2-1)}{RSS/(n-2)} = F$$

The usual (two-sided) t-test for two samples is a special case of the F-test in one-way ANOVA

R-commands for bone density example:

```
bonedensity=read.table("http://www.uio.no/studier/emner/matnat/math/
                STK4900/data/bonedensity.txt",header=T)
aov.density=aov(density~group,data=bonedensity)
summary(aov.density)
```

R-output (edited)

```
            Df     Sum Sq       Mean Sq      F value        Pr(>F)
group        1     0.00197      0.00197       8.09          0.0082
Residuals   28     0.00681      0.000243
```

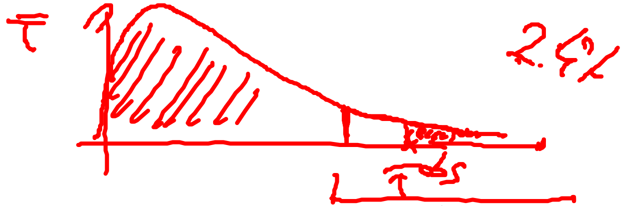
Note that $t^2 = 2.844^2 = 8.09 = F$

10 test

p-values vs 0.05

pvalues vs 0.005 ← $\frac{0.05}{10}$

Multiple testing



In Lecture 1, we performed a hypothesis test and calculated a P-value (using a t-test).

Now in Lecture 2 we have discussed one-way ANOVA for the null hypothesis:

$$H_0 : \mu_1 = \ldots\ldots = \mu_K$$

We could also be interested in testing pair-wise differences in mean between category levels: $H_{0jk} : \mu_j = \mu_k$

Assume all $H_{0jk} : \mu_j = \mu_k$ are true and are tested with a significance level α.

Note: This will consist of m=K (K-1)/2 different tests, i.e. *multiple tests*.

Then the overall probability of rejecting one or more null hypotheses (falsely) will be greater than α, *but* less than m α.

Thus: With a initial level α'= α/m we can ensure an overall level of α.

Such a procedure is called a Bonferroni correction. Although appealing Bonferroni corrections can be seriously conservative.

23'000

Multiple testing, cont.

Often, we perform a very large number test at the same time.

For example, in genomics, maybe m=10000 tests are performed simultaneously. For each test, we have a probability α of erroneously rejecting H_0, resulting in a false discovery ("Type I error").

With $\alpha = 0.05$, and 10000 independent tests, we expect 500 false discoveries. Even for small m, the probability of at least one false discovery is large. With f.ex. m=10 independent tests, we get

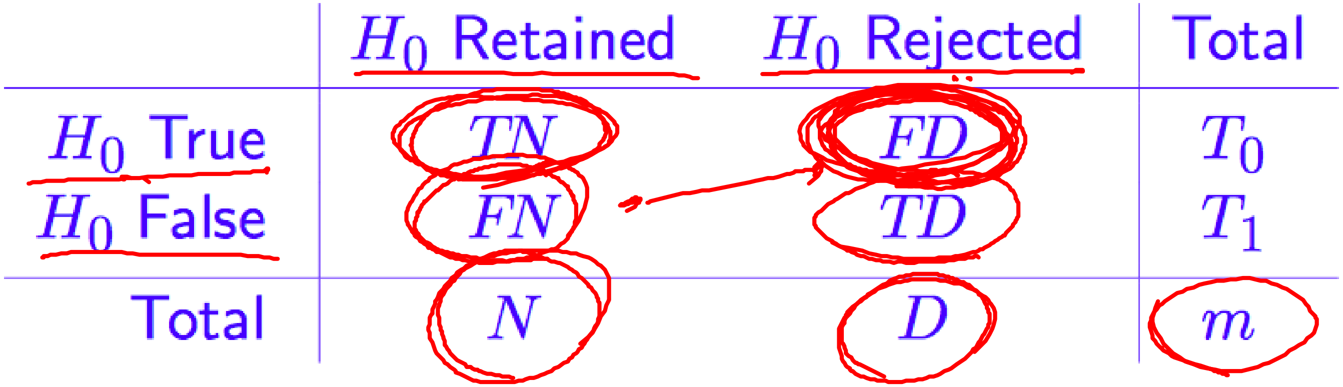
P(at least one false discovery among 10 tests) = 1 – P(no false discoveries)

$$= 1 - (1-\alpha)^{10} = 1 - (1-0.05)^{10} = 0.4$$

Multiple testing setting

- We perform m simultaneous tests with a common procedure.

- For a given procedure, classify the results as:



	H_0 Retained	H_0 Rejected	Total
H_0 True	TN	FD	T_0
H_0 False	FN	TD	T_1
Total	N	D	m

- TN = # True Non-discoveries, FN = # False Non-discoveries, FD = # False Discoveries, TD = # True Discoveries.

- Only N, D, m are observed, FD (for instance) is not known

How to choose a threshold?

- Control Per-Comparison Type I Error (PCER)
  - a.k.a. "uncorrected testing," many type I errors
  - Gives $\mathbb{P}\{FD_i > 0\} \leq \alpha$ marginally for all $1 \leq i \leq m$

- Control Familywise Type I Error (FWER)
  - e.g.: Bonferroni: use per-comparison significance level α/m
  - Guarantees $\mathbb{P}\{FD > 0\} \leq \alpha$
- Control False Discovery Rate (FDR)
  - first defined by Benjamini & Hochberg (BH, 1995, 2000)
  - Guarantees FDR $\equiv \mathbb{E}\left(\frac{FD}{D}\right) \leq \alpha$

Borrowed from C.R. Genovese

We use the term *raw P-values* for the original P-values P_1, P_2, ..., P_m, and produce *adjusted P-values* P_1^{adj}, P_2^{adj}, ..., P_m^{adj} based on the type of control above.

Bonferroni adjustment (simplest to understand, but conservative)

All hypotheses with raw P-values < α/m are rejected. Guarantees a probability of any FD below α (as pointed out above).

Adjusted P-values will be $P_i^{adj} = \min(mP_i, 1)$, $i = 1, 2, ..., m$

In R: Let P be a vector of raw P-values.

```
> p.adjust(P, method="...")
```

returns a vector of adjusted P-values. Choices of methods for p.adjust can for instance be "bonferroni” or "BH” for Benjamini-Hochberg controlling the FDR.

$$\frac{0.05}{10'000}$$

0 0 000005

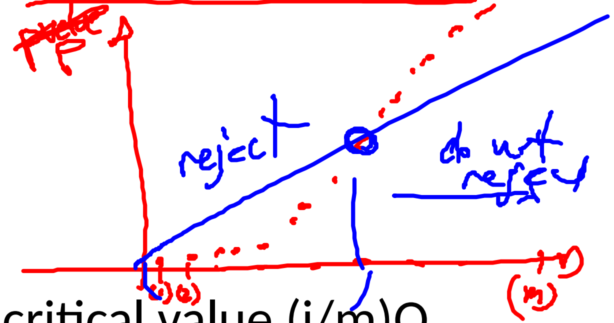
↑

FDR adjustment

Bonferroni, controls the overall probability of having at least one false discovery. Bonferroni is very strict, and may rule out discoveries of interest as false.

FDR, on the other hand, controls the expected proportion of false discoveries relative to the total number of discoveries, and tolerates some false discoveries.

With an FDR of f.ex. 10 % (0.10), on average 10% of the discoveries will represent false discoveries. Dropping the mathematics behind, the Benjamini-Hochberg procedure can be summarized as:



- Choose a false discovery rate Q (f.ex. 10% or 20%)
- Sort the raw P-values, giving $P_{(1)}$, $P_{(2)}$, …, $P_{(m)}$
- Compare each $P_{(i)}$-value to its Benjamini-Hochberg critical value (i/m)Q
- The largest $P_{(i)}$-value that has $P_{(i)} < (i/m)Q$ is significant, and *all* of the P-values smaller than it are also significant.

8

$\frac{i}{m} \cdot Q$

25 =

$\frac{1}{\cancel{25}_{5}} \cdot \frac{\cancel{15}^{3}}{100}$

2

3

0.006

0.012

0.018

0.

The BH adjusted P-value is the raw P-value times m/i. If the adjusted P-value is smaller than the false discovery rate Q, the test is significant.

Example

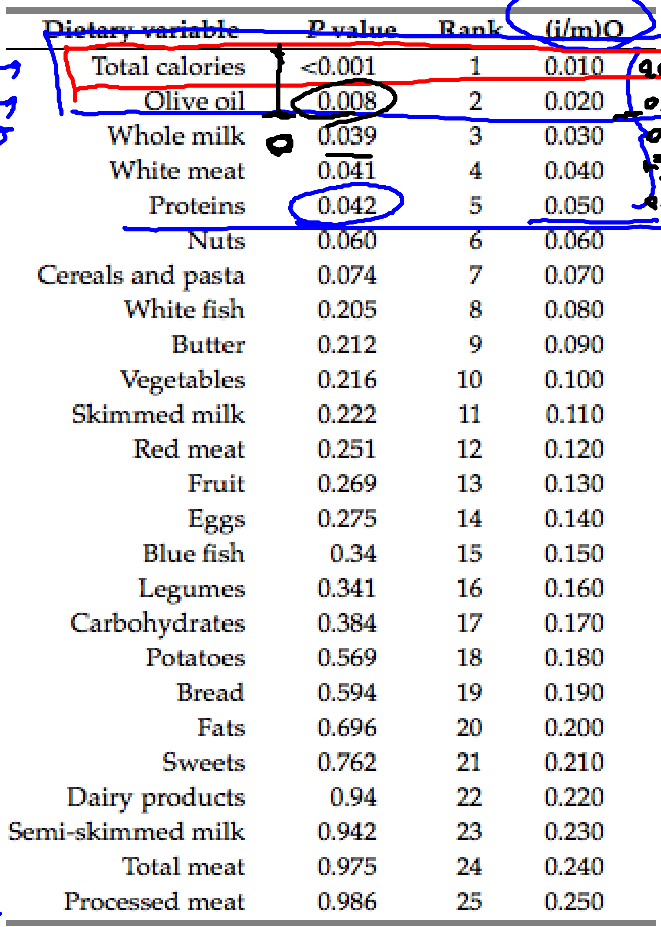
Garcia-Arenzana et al.(2014) Associations between dietary variables and breast cancer risk

m=25 tests, giving raw P-values in column 2

FDR-corrected, with Q=0.25 (!large!), we see from column 4, that Proteins and the other variables above are significant.

FDR-corrected with Q=0.15 gives Olive Oil and Total calories as significant (check!)

Using Bonferroni-correction, only the variables with raw P-value < 0.05/25 = 0.002 are significant, that is only Total calories



Dietary variable	P value	Rank	(i/m)Q	
Total calories	<0.001	1	0.010	0.006
Olive oil	0.008	2	0.020	0.012
Whole milk	0.039	3	0.030	0.018
White meat	0.041	4	0.040	.024
Proteins	0.042	5	0.050	0.030
Nuts	0.060	6	0.060	
Cereals and pasta	0.074	7	0.070	
White fish	0.205	8	0.080	
Butter	0.212	9	0.090	
Vegetables	0.216	10	0.100	
Skimmed milk	0.222	11	0.110	
Red meat	0.251	12	0.120	
Fruit	0.269	13	0.130	
Eggs	0.275	14	0.140	
Blue fish	0.34	15	0.150	
Legumes	0.341	16	0.160	
Carbohydrates	0.384	17	0.170	
Potatoes	0.569	18	0.180	
Bread	0.594	19	0.190	
Fats	0.696	20	0.200	
Sweets	0.762	21	0.210	
Dairy products	0.94	22	0.220	
Semi-skimmed milk	0.942	23	0.230	
Total meat	0.975	24	0.240	
Processed meat	0.986	25	0.250	

Two numerical variables

For one-way ANOVA we study how a numerical variable (e.g. blood coagulation time) depends on a categorical variable (e.g. diet)

Often we want to study the relation between two numerical variables

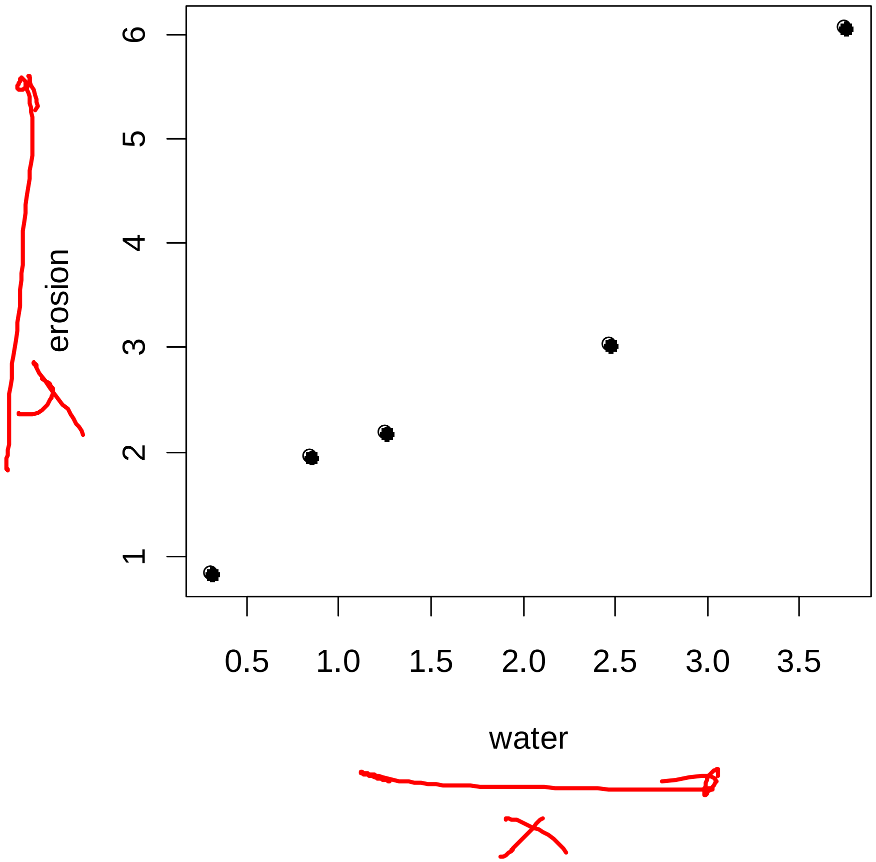
Example A: When water flows across a field, some of the soil will be washed away (eroded). An experiment has been performed in order to investigate how the amount of water affects the amount of soil that is eroded.

Amount of water (l/s)	0.31	0.85	1.26	2.47	3.75
Erosion (kg)	0.82	1.95	2.18	3.02	6.07

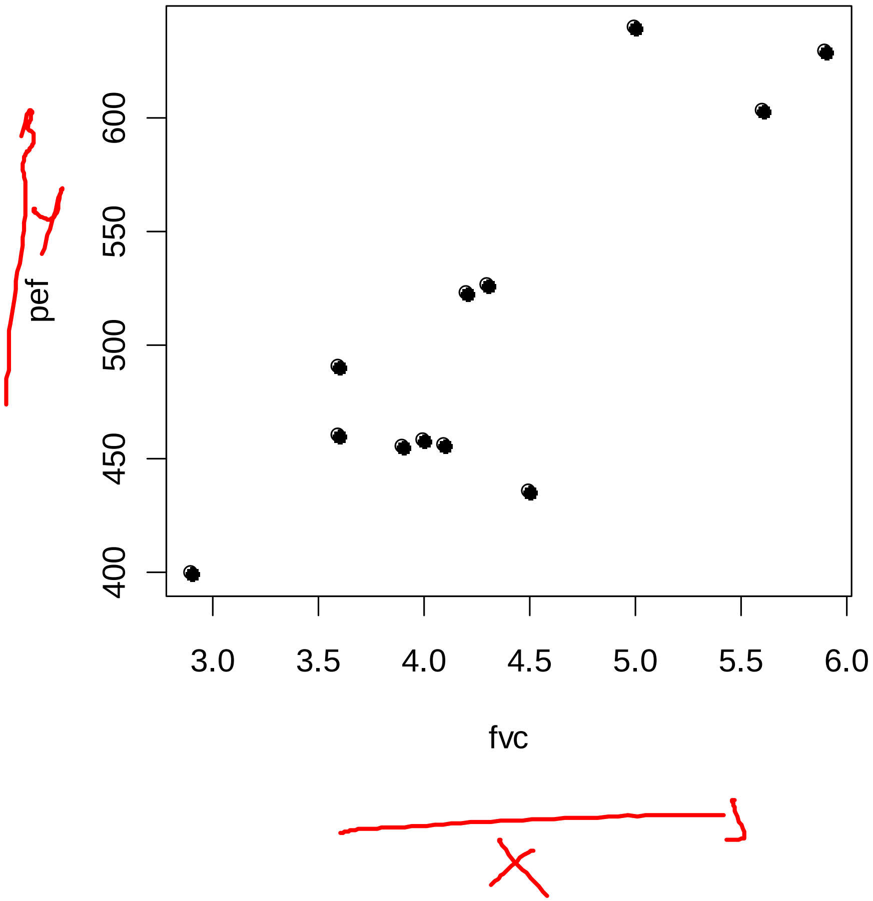
Example B: Forced vital capacity (FVC) and peak expiratory flow (PEF) have been measured for 12 adults (in liter and liter per minute, respectively). What is the relation between these two measures of lung function?

Person	1	2	3	4	5	6
FVC	3.9	5.6	4.1	4.2	4.0	3.6
PEF	455	603	456	523	458	460
Person	7	8	9	10	11	12
FVC	5.9	4.5	3.6	5.0	2.9	4.3
PEF	629	435	490	640	399	526

Example A



Example B



1, 2, ..., n

We will consider two situations:

1. The data (x_1,y_1) , … , (x_n,y_n) are considered as independent replications of a pair of random variables $(X\ ,Y\ )$
2. The data are described by a linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \qquad i = 1,\ldots.,n$$

Here y_1 , … , y_n are the outcomes that are considered to be realizations of random variables, while x_1 , … , x_n are considered to be fixed (i.e. non-random) and the ε_i's are random errors (noise)

Situation 1 occurs for observational studies (like Example B), while situation 2 occurs for planned experiments, where the values of the x_i's are under the control of the experimenter (like Example A)

In situation 1 we will often *condition* on the observed values of the x_i's, and analyze the data as if they are from situation 2

We start out by considering situation 1

Bivariate distributions

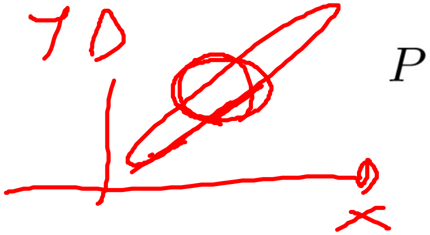
$X \sim f_X$

$Y \sim f_Y$

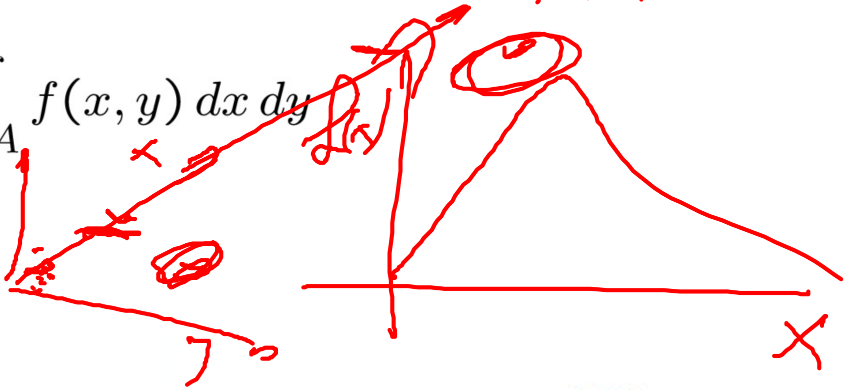
We describe the joint distribution of a pair of random variables (X, Y) through their *bivariate probability density,* $f(x,y)$

$X, Y \sim f$

This is defined so that



$$P(\,(X,Y)\in A\,) = \int_A f(x,y)\,dx\,dy$$



The bivariate normal distribution depends on the parameters:

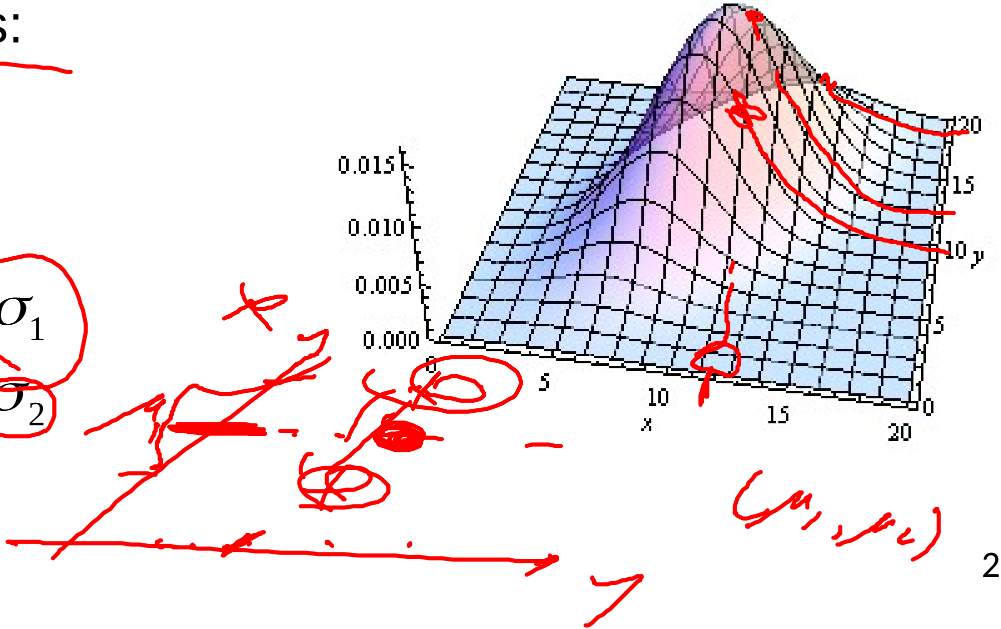
Mean of X : μ_1

Mean of Y : μ_2

Standard deviation of X : σ_1

Standard deviation of Y : σ_2

Correlation : ρ



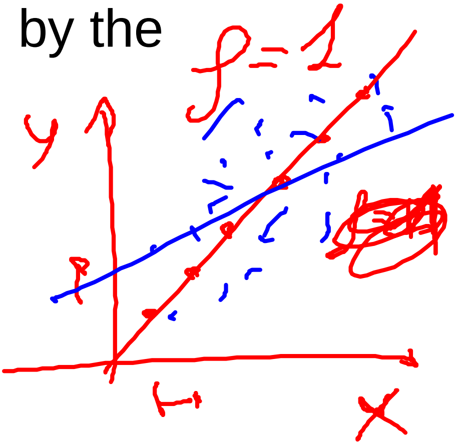
(μ_1, μ_2)

Covariance and correlation

ρ = 0.6

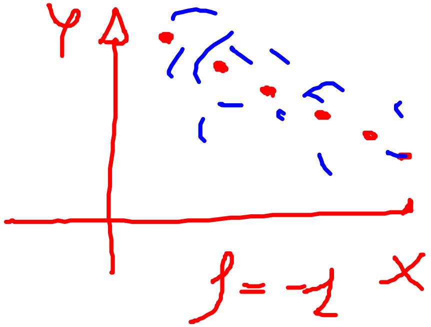
The dependence between *X and Y* may be summarized by the *covariance:*

$$\mathrm{Cov}(X,Y) = E[(X-\mu_1)(Y-\mu_2)]$$



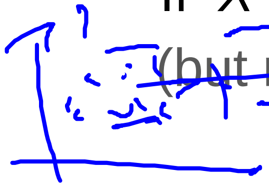
or by the *correlation coefficient:*

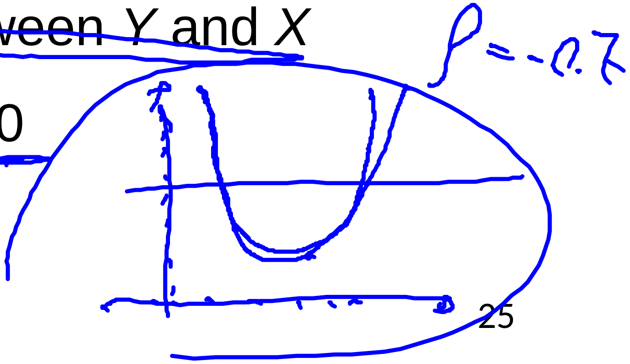
$$\rho = \mathrm{corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{sd}(X)\,\mathrm{sd}(Y)}$$



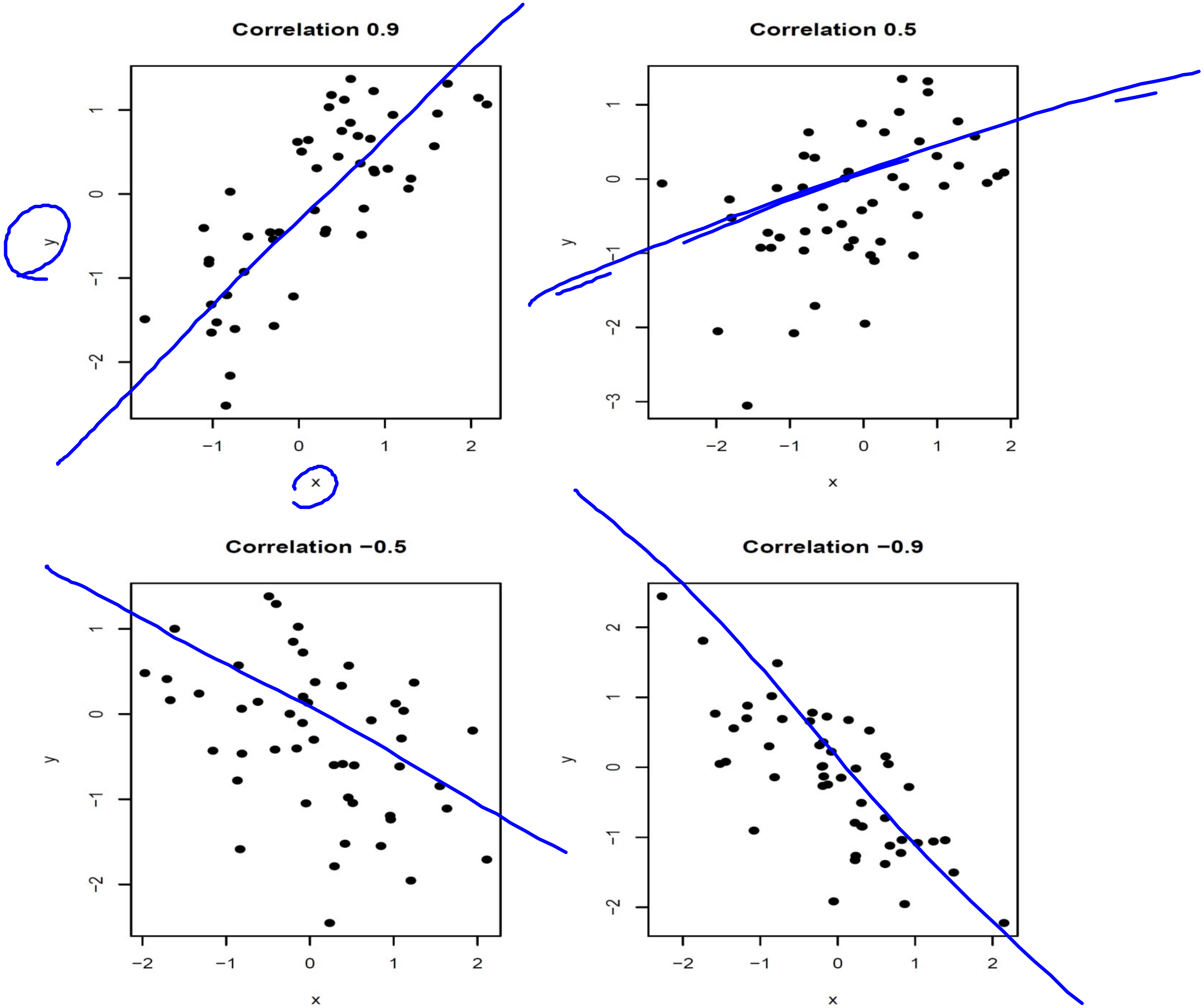
Important properties of the correlation coefficient:

- corr(*X,Y*) takes values between -1 and 1
- corr(*X,Y*) describes the *linear* relationship between *Y* and *X*
- If *X* and *Y* are independent, then corr(X,Y)=0 (but not necessarily the other way around)

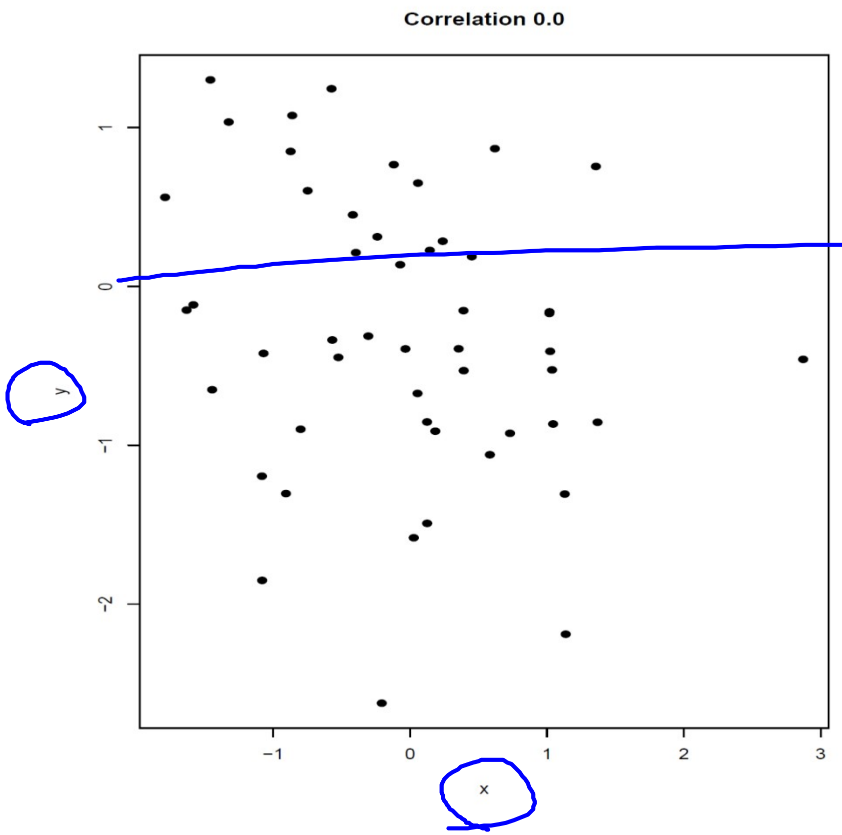


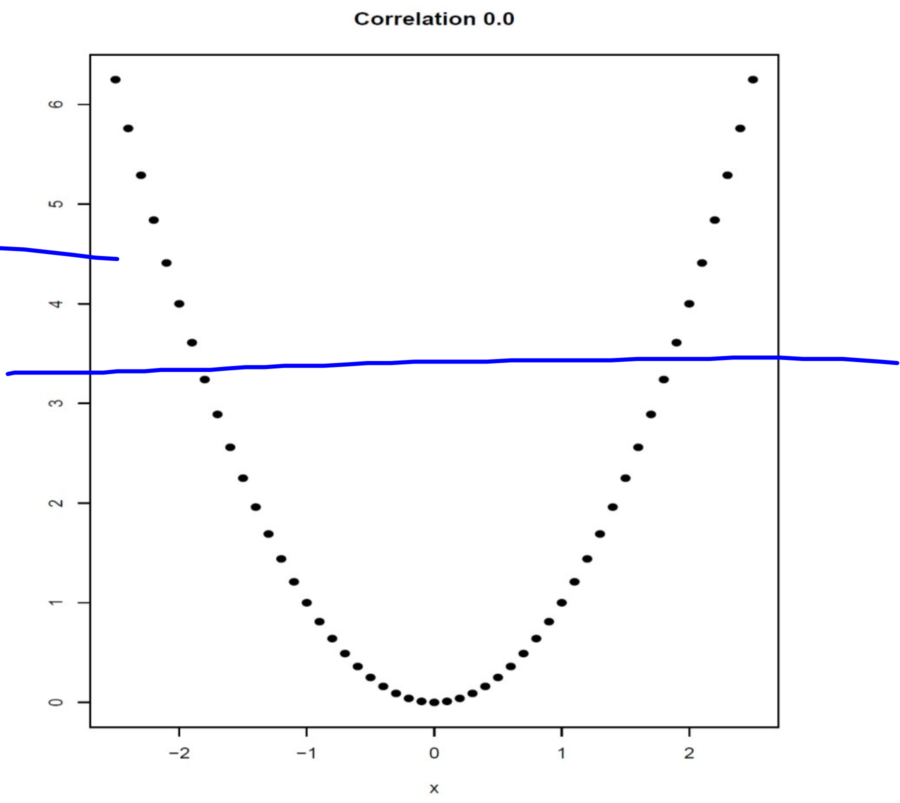


Examples of correlated data:



Examples of uncorrelated data:


Correlation 0.0


Correlation 0.0

$y = x^2$

Empirical correlation

ρ r

The empirical correlation coefficient is an estimator of the theoretical correlation coefficient, and it takes the form

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})/(n-1)}{s_x \cdot s_y}$$

Here s_x and s_y are the empirical standard deviations of the x_i's and the y_i's

r is called the *Pearson correlation coefficient*

The properties of the Pearson correlation coefficient are similar to those of the theoretical correlation coefficient

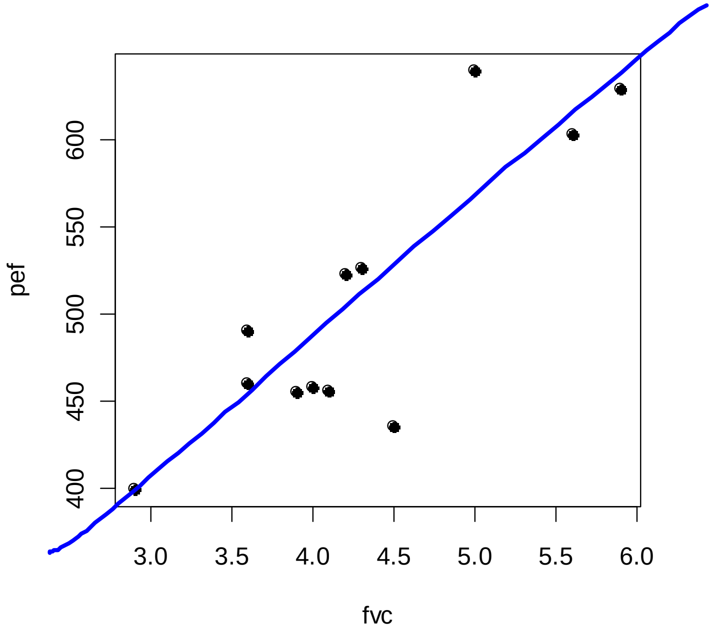
$-1 \le r \le 1$

linear dependency

$X \perp\!\!\!\perp Y \Rightarrow r = 0$ but $r = 0 \not\Rightarrow X \perp\!\!\!\perp Y$

Consider the example with measures of lung function:

Person	1	2	3	4	5	6	7	8	9	10	11	12
FVC	3.9	5.6	4.1	4.2	4.0	3.6	5.9	4.5	3.6	5.0	2.9	4.3
PEF	455	603	456	523	458	460	629	435	490	640	399	526



R-commands and results:

```
fvc=c(3.9,5.6,4.1,4.2,4.0,3.6,5.9,4.5,3.6,5.0,2.9,4.3)
pef=c(455,603,456,523,458,460,629,435,490,640,399,526)
cov(fvc,pef)
cov(fvc,pef)/(sd(fvc)*sd(pef))
0.856

cor(fvc,pef)
0.856
```

Test and confidence interval for correlation

We assume that (x_1,y_1) , … , (x_n,y_n) are a random sample from a bivariate normal distribution

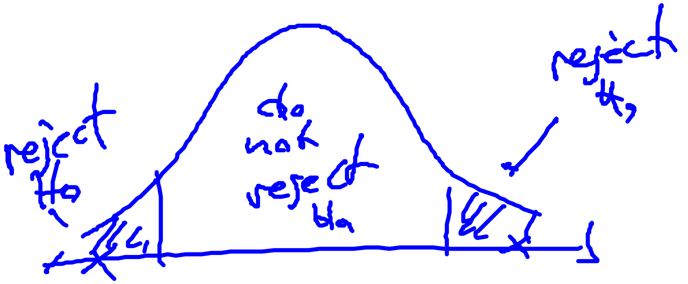
$X, Y \sim N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}; \Sigma\right)$

Consider testing the null hypothesis $H_0 : \rho = 0$ versus the alternative $H_0 : \rho \neq 0$

Test statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

t_{n-2}



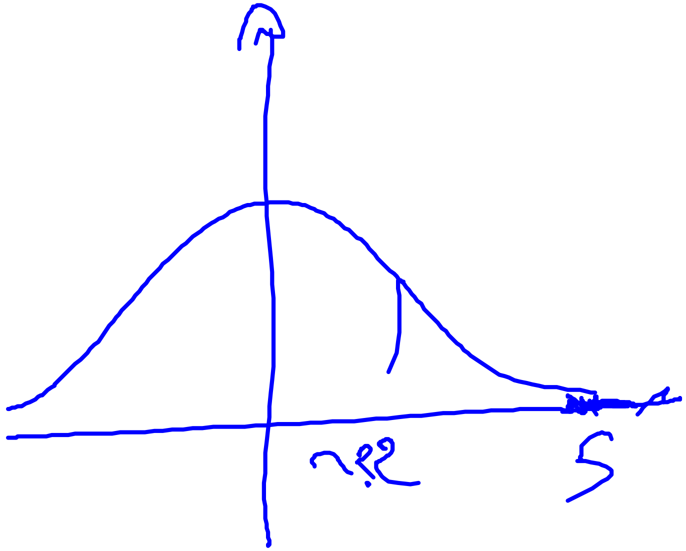
We reject H$_0$ for large values of $|t|$

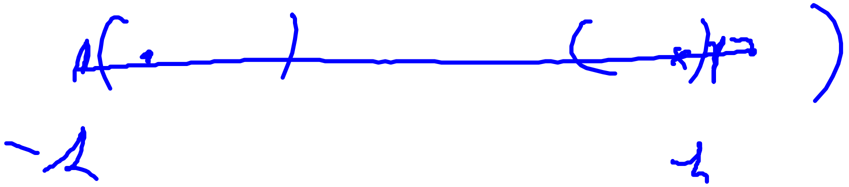
Under H$_0$ the test statistic is t-distributed with n – 2 df

It is more complicated to describe how one may obtain a confidence interval for r (but one is obtained by the R code on the following slide)

R-command and results:

```
cor.test(fvc,pef)

        Pearson's product-moment correlation

data:  fvc and pef
t = 5.23, df = 10, p-value = 0.00038
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.554     0.959
sample estimates:
      cor
0.856
```





Note that the confidence interval is not symmetric

Spearman (rank) correlation

The Pearson correlation is sensitive to outliers in the data, and measures degree of linear relation.

An alternative correlation measure is the Spearman correlation:

x1, x2, x3, x4, x5
y1, y2, y3, y4, y5

The smallest x_i is replaced by rank $r_i = 1$,

the second smallest x_i is replaced by rank $r_i = 2$, and so on to

the largest x_i which is replaced by rank $r_i = n$.

X3 X1 X2 X5 X4
y2 y1 y3 y5 y4

Similarly, the y_i are replaced by ranks s_i .

x 2 3 1 5 5
y 1 2 3 4 3

The Spearman correlation is then simply the Pearson correlation of the ranks (r_1,s_1) , … , (r_n,s_n).

5 4 1 3 2
5 4 1 3 2

1 2 3 4 5
1 2 3 4 3

In R:

```
> cor(fvc, pef, method="spearman")
[1] 0.669
```

5 4 1 3 2
1 5 4 2 3

Simple linear regression

We have data $(x_1, y_1), \ldots, (x_n, y_n)$

Here:

y_i = outcome
(or response)
(or dependent variable)

$y = f(x) + \varepsilon$

x_i = predictor
(or covariate)
(or explanatory variable)
(or independent variable)

Model:

$$y_i = E(y_i \mid x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where the x_i's are considered to be fixed quantities, and the ε_i's are independent error terms ("noise") that are assumed to be $N(0, \sigma_\varepsilon^2)$-distributed

Consider the erosion example:

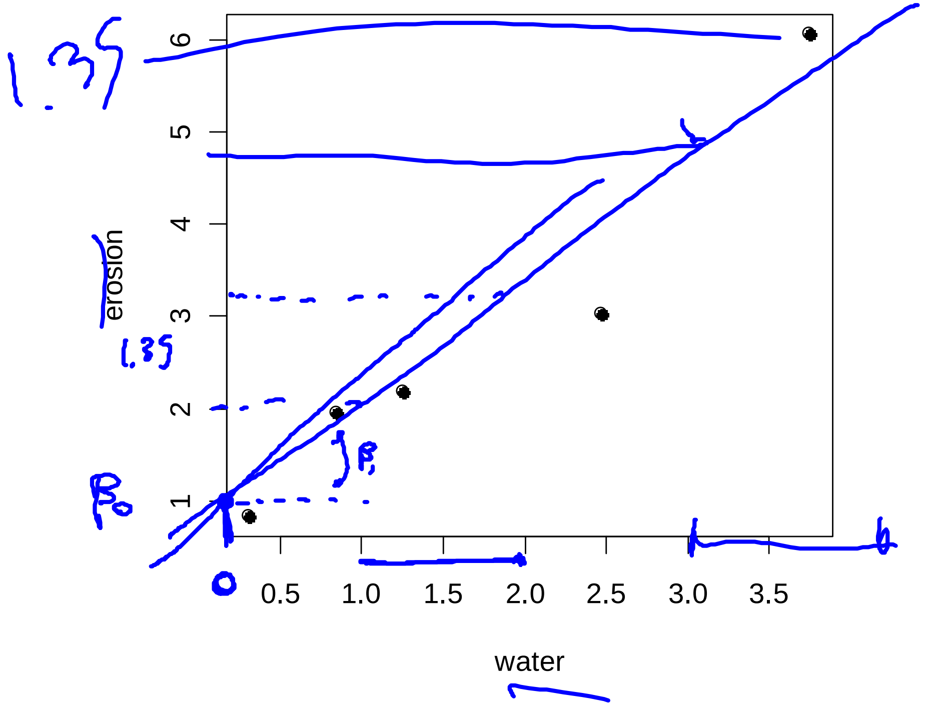
Amount of water (l/s)	0.31	0.85	1.26	2.47	3.75
Erosion (kg)	0.82	1.95	2.18	3.02	6.07

Response = erosion

Predictor = amount of water

Model:

$$\underline{\text{erosion}} = \boxed{\beta_0} + \boxed{\beta_1}\,\underline{\text{water}} + \underline{\varepsilon}$$



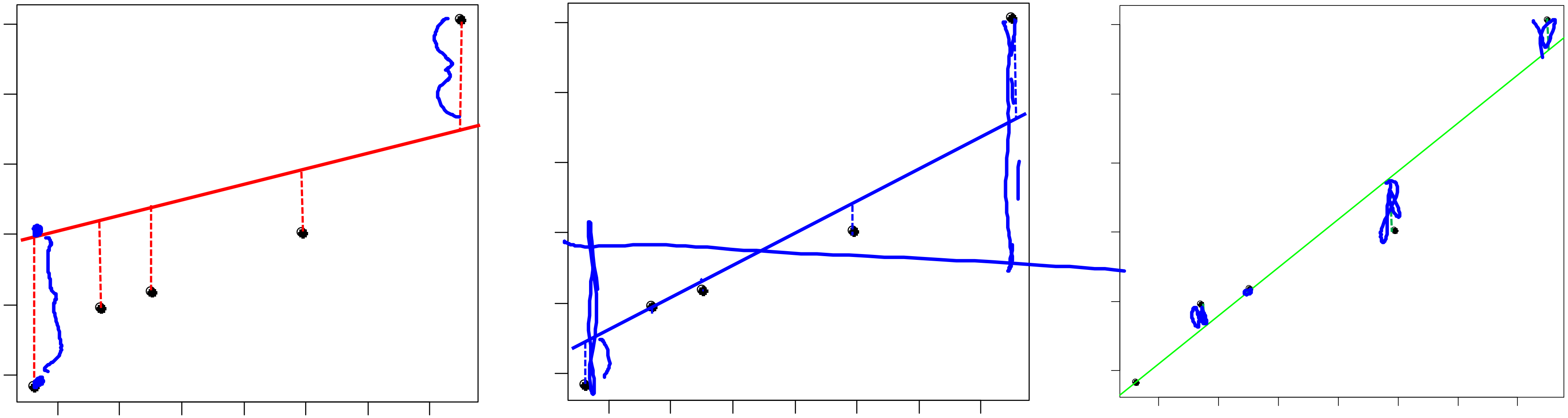
Least squares

We estimate the regression coefficients using the method of least squares, i.e. the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained as the values of b_0 and b_1 that minimize the sum of squares

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

$$\sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \underline{y_i} \right)^2$$

Illustration:



R-commands:

```
water=c(0.31,0.85,1.26,2.47,3.75)
erosion=c(0.82,1.95,2.18,3.02,6.07)
fit=lm(erosion~water)
summary(fit)
plot(water,erosion,pch=19)
abline(fit)
```

$fit \overset{=}{<-} lm(y \sim x)$

$erosion = 0.406 + 1.390\, water + \varepsilon$

$t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} \sim t_{n-2}$

R-output (edited)

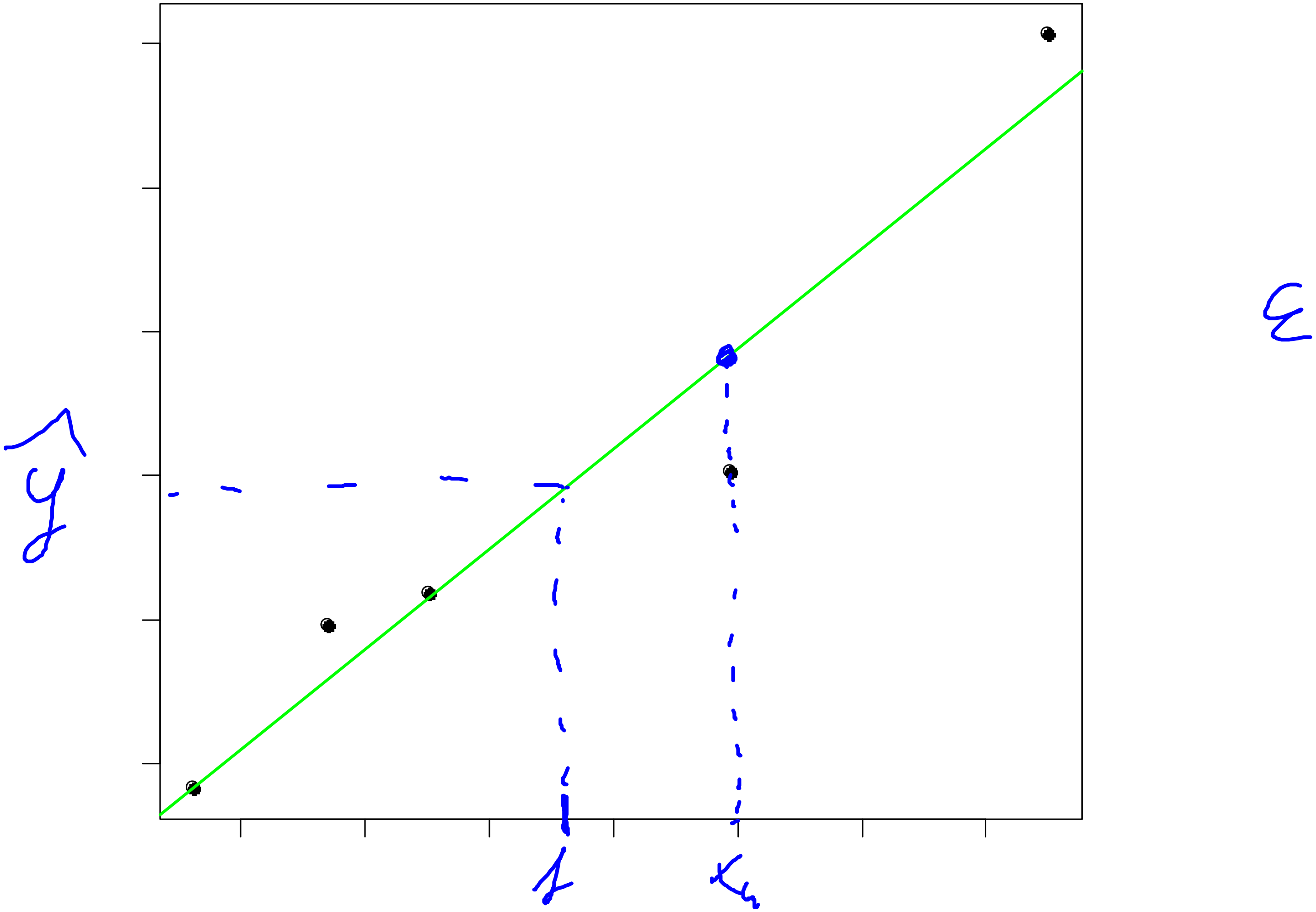
```
Coefficients:
                    Estimate    Std. Error     t value       Pr(>|t|)
(Intercept)         0.406       0.445          0.912         0.429
water               1.390       0.210          6.630         0.007

Residual standard error: 0.580 on 3 degrees of freedom
Multiple R-squared: 0.936     Adjusted R-squared: 0.915
F-statistic: 44.0 on 1 and 3 DF,  p-value: 0.007
```

$\hat{\beta}_0$

$\hat{\beta}_1$

$H_0\ \beta_1 = 0$

"Estimate" denotes the least squares estimates (the meaning of the other parts of the output will be made clear in the



Fitted regression line: $\text{erosion} = 0.406 + 1.390 \times \text{water}$

Fitted values and residuals

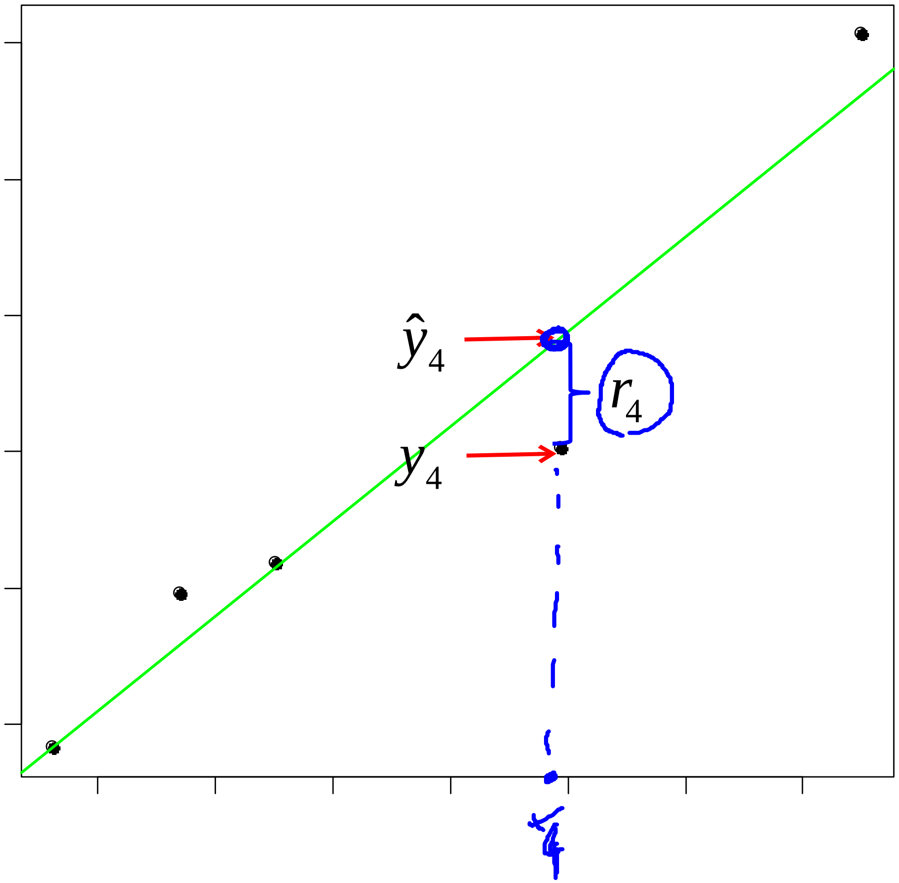
Fitted values:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Residuals:

$$r_i = y_i - \hat{y}_i$$

The residuals are estimates of the unobserved ε_i's



Sums of squares

In a similar manner as for one-way ANOVA, we have the sums of squares:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ (total sum of squares)

$$MSS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ (model sum of squares)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ (residual sum of squares)

Decomposition:

$$TSS = \underbrace{MSS}_{\beta_0+\beta_1 x} + \underbrace{RSS}_{\varepsilon}$$

Standard errors

$\hat{\sigma}_{\varepsilon}$

Unbiased estimator of σ^2_ε :

β_0 β_1

$$\mathrm{V\hat{a}r}(\varepsilon) = s^2_{y|x} = RSS/(n-2)$$

$s_{y|x}$ is the "residual standard error" in the R output

The variance of $\hat{\beta}_1$ is estimated by :

$$\mathrm{V\hat{a}r}(\hat{\beta}_1) = \frac{s^2_{y|x}}{(n-1)\,s^2_x}$$

where $s^2_x = \sum_{i=1}^{n} (x_i - \overline{x})^2 / (n-1)$ is the sample variance of the x_i's

Standard error: $se(\hat{\beta}_1) = \sqrt{\mathrm{V\hat{a}r}(\hat{\beta}_1)}$

Similar formulas hold for the variance and standard error of $\hat{\beta}_0$

The standard errors are denoted "Std. Error" in the R output

Hypothesis tests

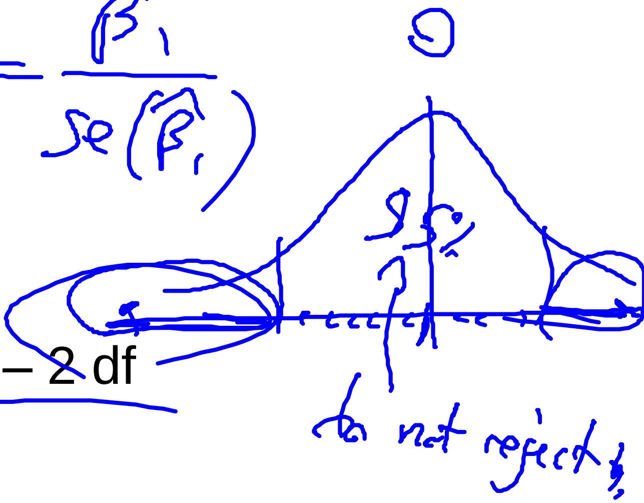
Consider testing the null hypothesis $H_0 : \beta_1 = 0$ versus the alternative $H_A : \beta_1 \neq 0$

Test statistic:

$$t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$$

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

We reject H$_0$ for large values of $|t|$



Under H$_0$ the test statistic is t-distributed with $n - 2$ df

P-value (two-sided) : $P = 2\,P(T > |t|)$,

where T is t-distributed with $n - 2$ df.

Testing the null hypothesis $H_0 : \beta_0 = 0$ is performed similarly (but is usually not of much interest)

t-statistics and P-values are given in the R output as "t value" and "Pr(>|t|)"

Confidence intervals

95% confidence interval for β_1 :

$$\hat{\beta}_1 \pm c \cdot se(\hat{\beta}_1)$$

$t_{n-2;0.975}$

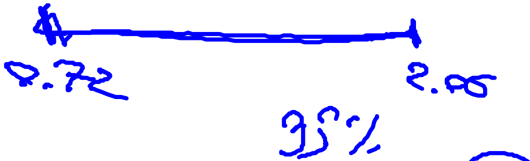
$-t_{n-2;0.975} = t_{n-2;0.025}$

where *c* is the upper 97.5% percentile in the t-distribution with $n-2$ df

95% confidence interval in the erosion example:

$$1.39 \pm 3.18 \cdot 0.210$$

i.e. from 0.72 to 2.06



Note that the confidence interval does not contain 0 if and only if the P-value for the test is less than 5%

Correlation and regression

The least squares estimate for the slope is given by:

$$\underline{\hat{\beta}_1} = \boxed{r}\,\frac{s_y}{s_x}$$

where

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/(n-1)}{\underline{s_x} \cdot s_y}$$

is the Pearson correlation coefficient (and s_x and s_y are the empirical standard deviations of the x_i's and the y_i's)

Further the test for $\underline{H_0 : \beta_1 = 0}$ in a linear regression model (slide 40) is numerically equivalent to the test for $\underline{H_0 : \rho = 0}$ for bivariate data (slide 29)

Coefficient of determination

The coefficient of determination is given by

TSS

$$\boxed{R^2} = \boxed{\frac{MSS}{TSS}} = 1 - \boxed{\frac{RSS}{TSS}}$$

$0 \leq R^2 \leq 1$

This may be interpreted as the proportion of the total variability in the outcomes (TSS) that is accounted for by the model (MSS)

R^2 is given as " Multiple R-squared" in the R output

For the simple linear regression model R^2 is just the square of the Pearson correlation coefficient:

$$\boxed{R^2 = r^2}$$