Statistics 2020: Lectures 13/14 (Extra Lectures not on Quiz/Homeworks)

Richard Veale

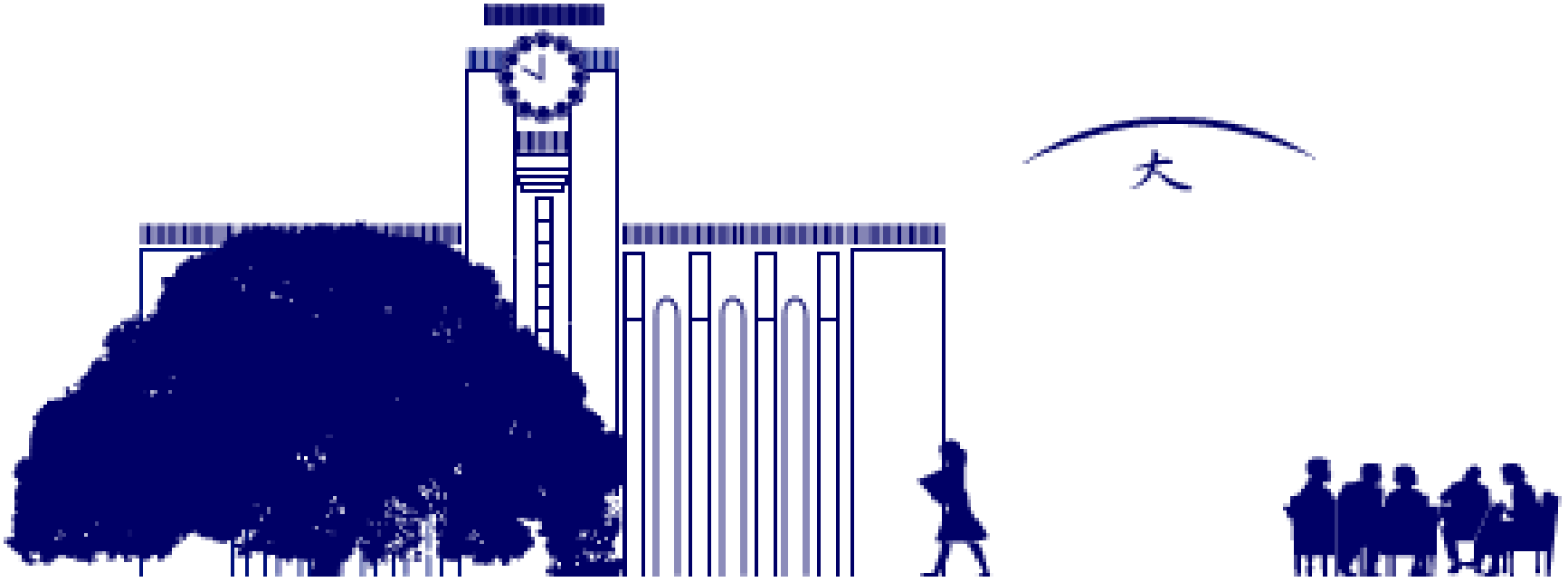
https://youtu.be/JxyFF3g7GxI

Lecture Video above (Youtube)

Introductory Statistics

Class 13: Correct usage of statistics

Richard Veale
Graduate School of Medicine
Center of Medical Education



Today's aims

- To understand how Pearson correlations can be misleading and to learn about alternative approaches.

- To learn that statistical p-values are often misunderstood.

- To understand some ways in which researchers fool themselves and others with statistics and how to avoid them.

Today's topics

1) Some comments on correlations (causality, selection bias, outliers)

2) Non-parametric correlation (Spearman's ρ)

3) The meaning of p-values

4) "Sins" when using statistics: p-value hacking/fishing

Correlation is not causality

Relationships that show a significant correlation, but is there causality?

Amount of sold ice cream ↑

and

Deaths by drowning ↑

Number of police officers ↑

and

Number of crimes ↑

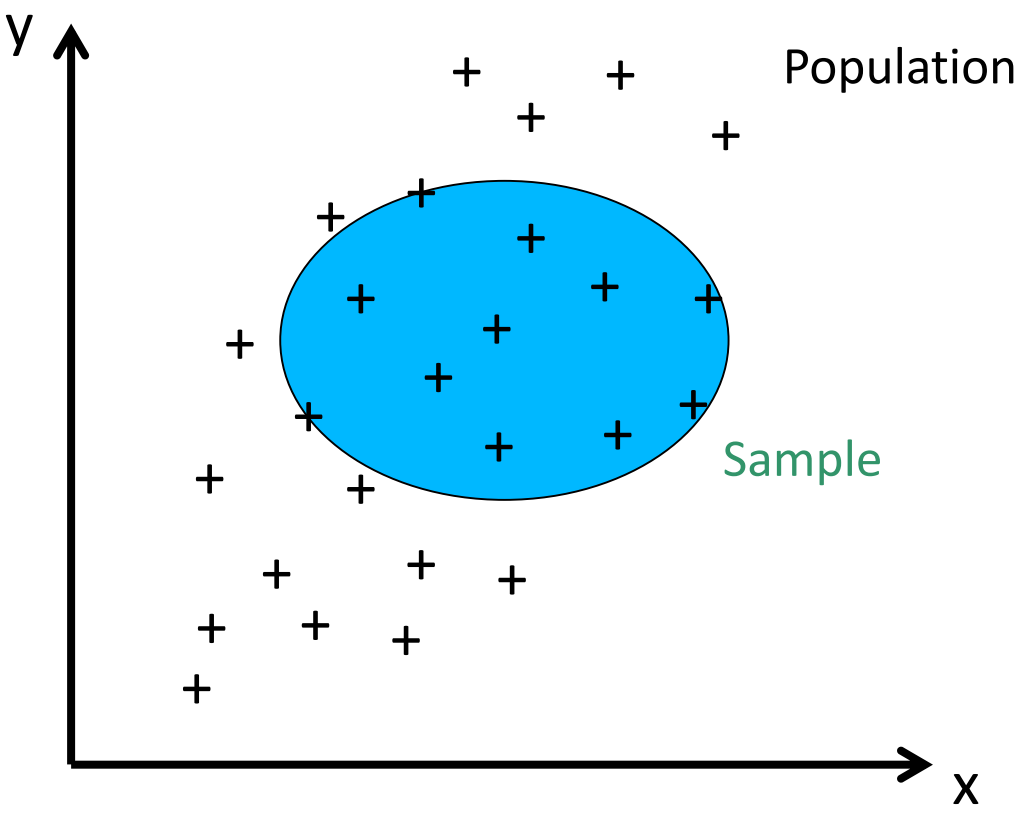
Number of storks sighted ↑

and

Population of Oldenburg, Germany ↑

Selection bias

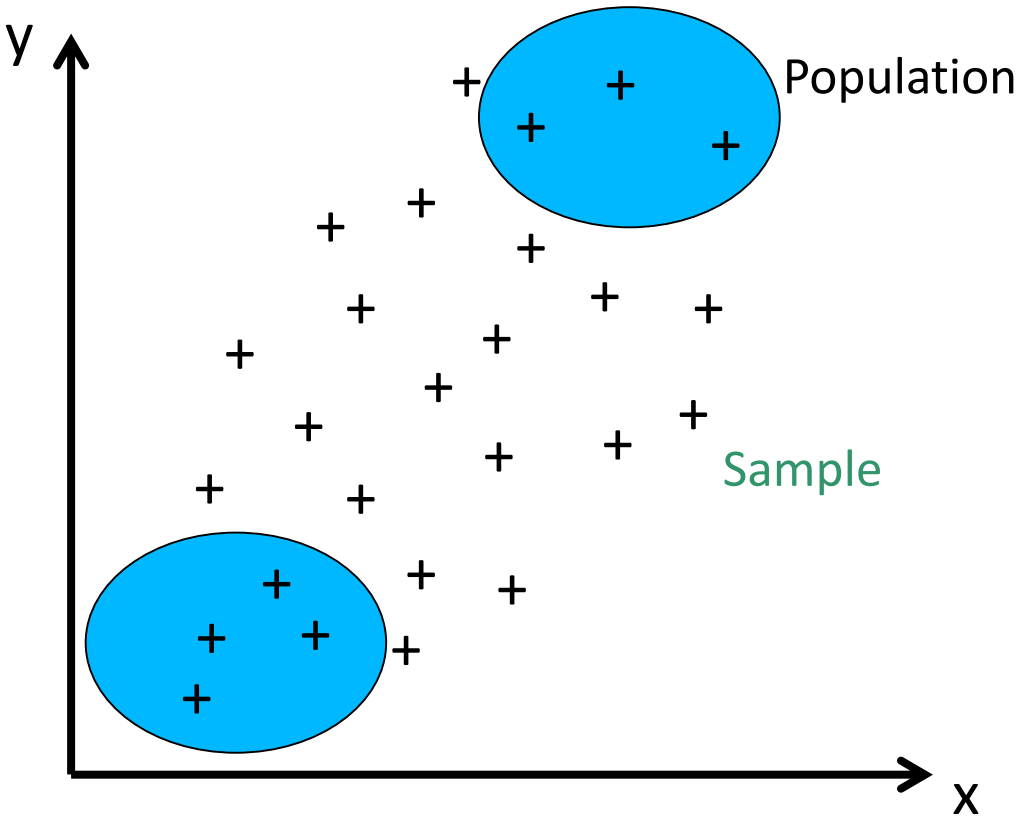
When sampling to calculate a correlation, the sample should not have restricted variance:



This sample would underestimate the correlation.

Selection bias

When sampling to calculate a correlation, the sample should not have restricted variance:

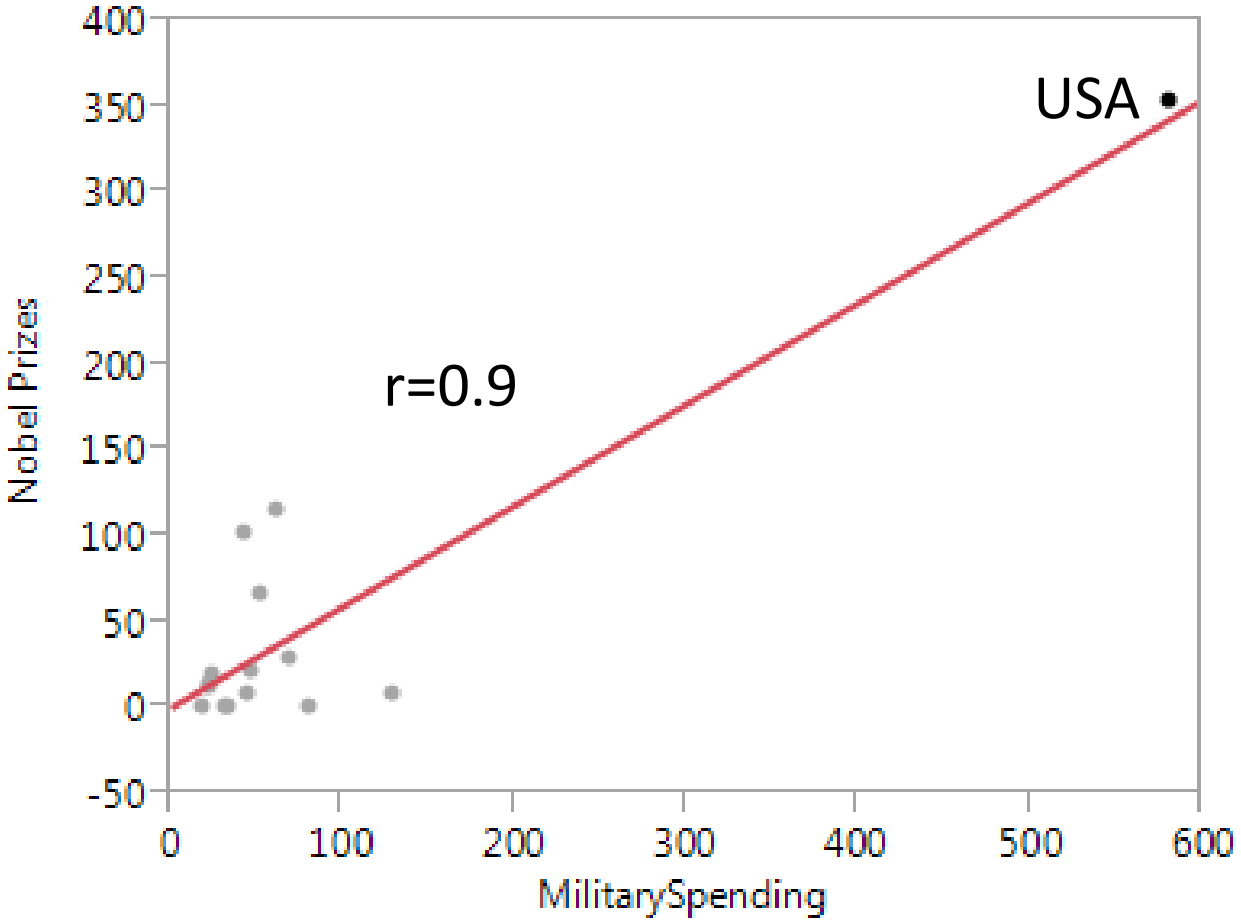


This sample might <u>overestimate</u> the correlation.

Outliers

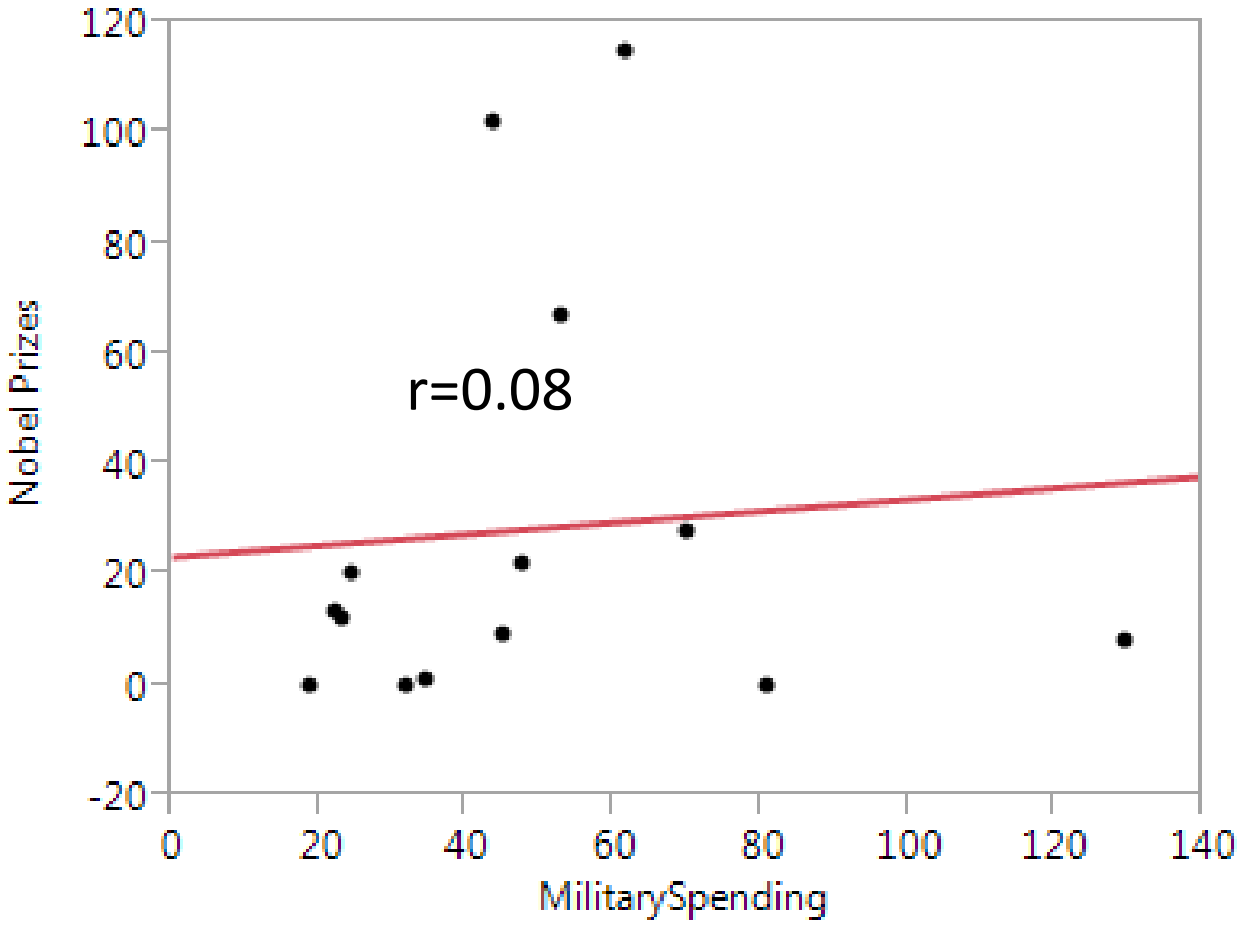
Some values have extreme influence on a correlation:

For example, countries by Nobel prizes and Military Spending (billion $/year)



Outliers

Some values have extreme influence on a correlation:



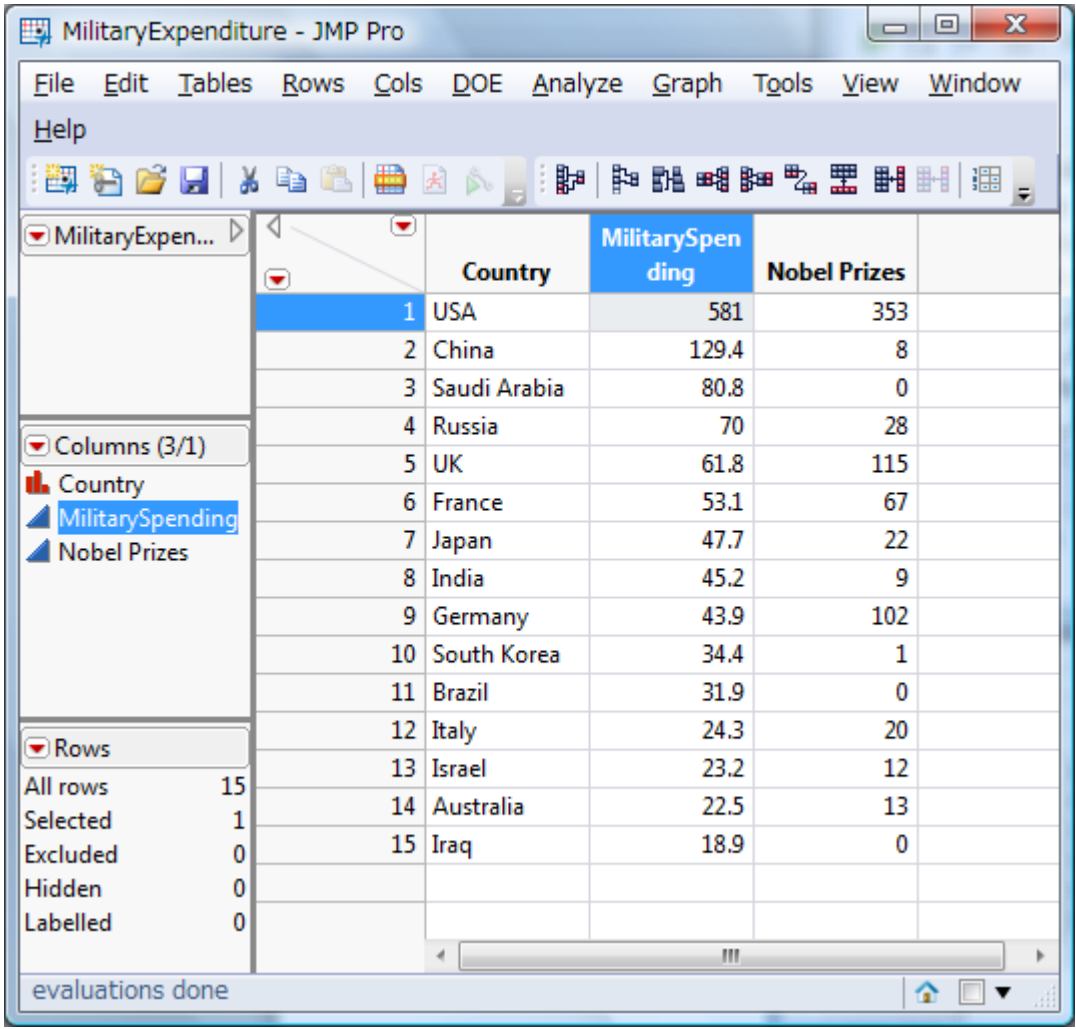
USA excluded

Spearman’s rank correlation

Spearman’s rank correlation is a non-parametric correlation:
(it is less sensitive to outliers and can handle ordinal data)

1) First column is ranked Military spending:

   USA->1
   China->2
   Saudi Arabia->3
   Russia->4
   etc.

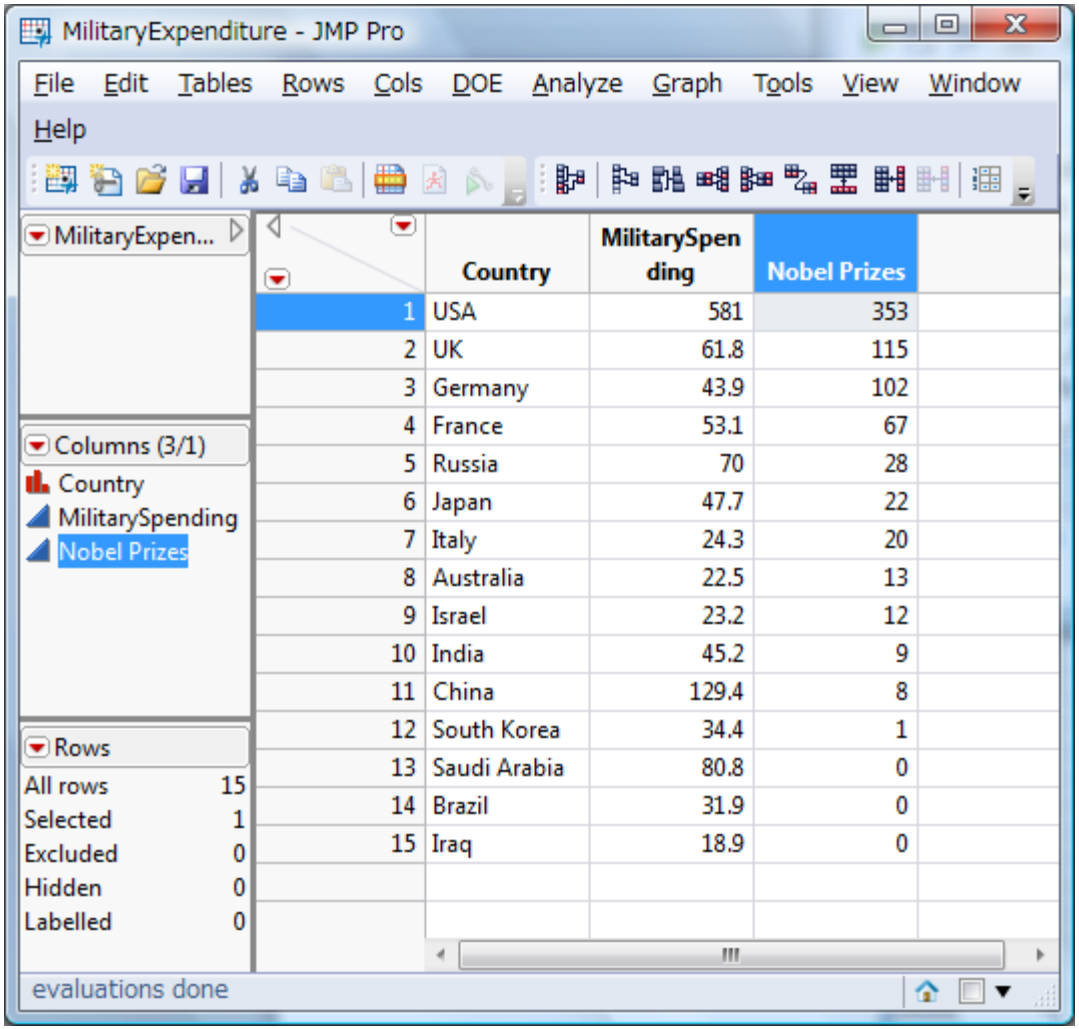


Spearman's rank correlation

2) Second column is ranked Nobel Prizes :

USA->1
UK->2
Germany->3
France->4
Russia->5
Japan->6
etc

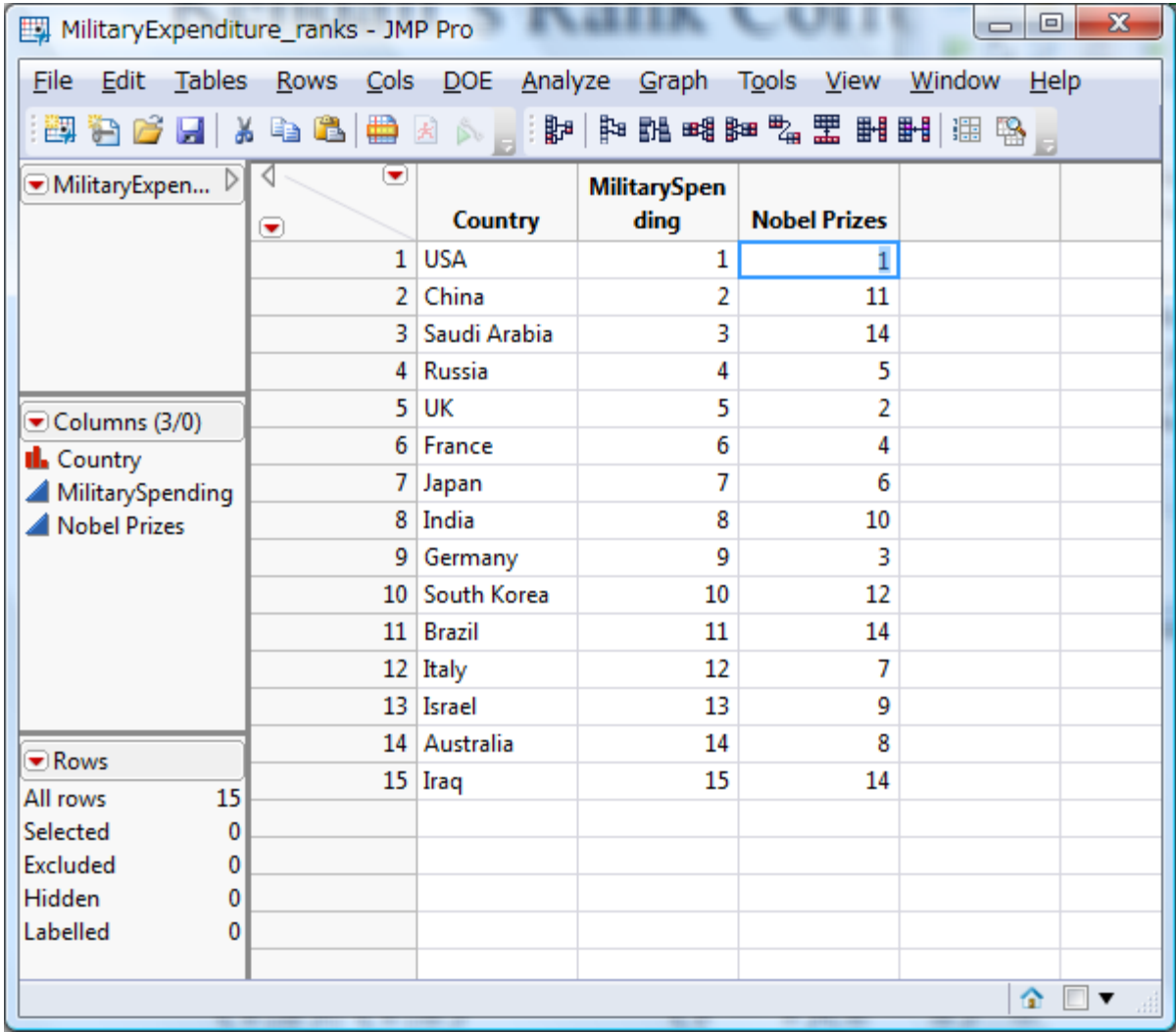
Ties (0):
Saudi Arabia->14
Brazil->14
Iraq->14



	Country	MilitarySpending	Nobel Prizes
1	USA	581	353
2	UK	61.8	115
3	Germany	43.9	102
4	France	53.1	67
5	Russia	70	28
6	Japan	47.7	22
7	Italy	24.3	20
8	Australia	22.5	13
9	Israel	23.2	12
10	India	45.2	9
11	China	129.4	8
12	South Korea	34.4	1
13	Saudi Arabia	80.8	0
14	Brazil	31.9	0
15	Iraq	18.9	0

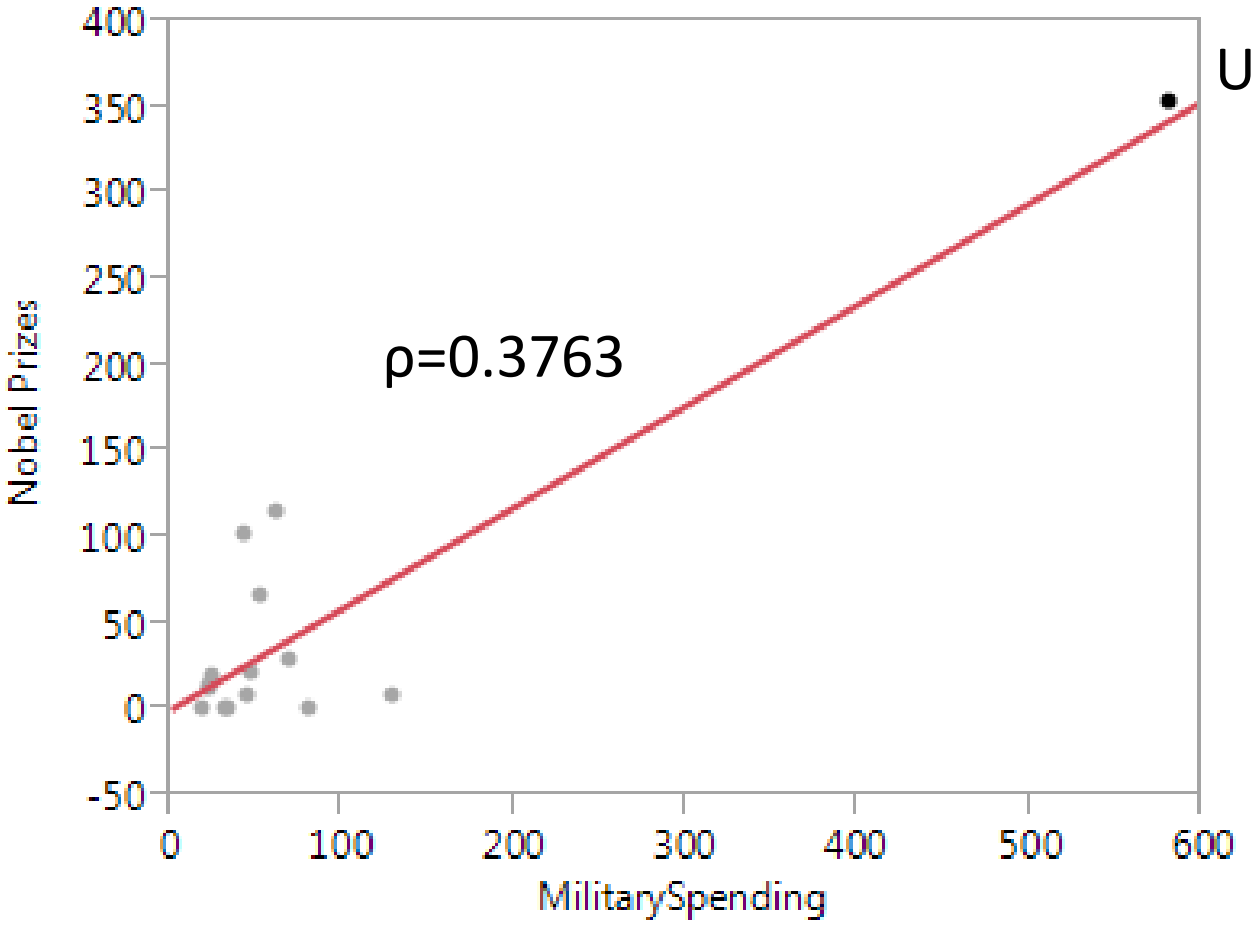
Spearman's rank correlation

3) Then, we compute Pearson's product-moment correlation on the ranks.



	Country	MilitarySpending	Nobel Prizes
1	USA	1	1
2	China	2	11
3	Saudi Arabia	3	14
4	Russia	4	5
5	UK	5	2
6	France	6	4
7	Japan	7	6
8	India	8	10
9	Germany	9	3
10	South Korea	10	12
11	Brazil	11	14
12	Italy	12	7
13	Israel	13	9
14	Australia	14	8
15	Iraq	15	14

Spearman's rank correlation

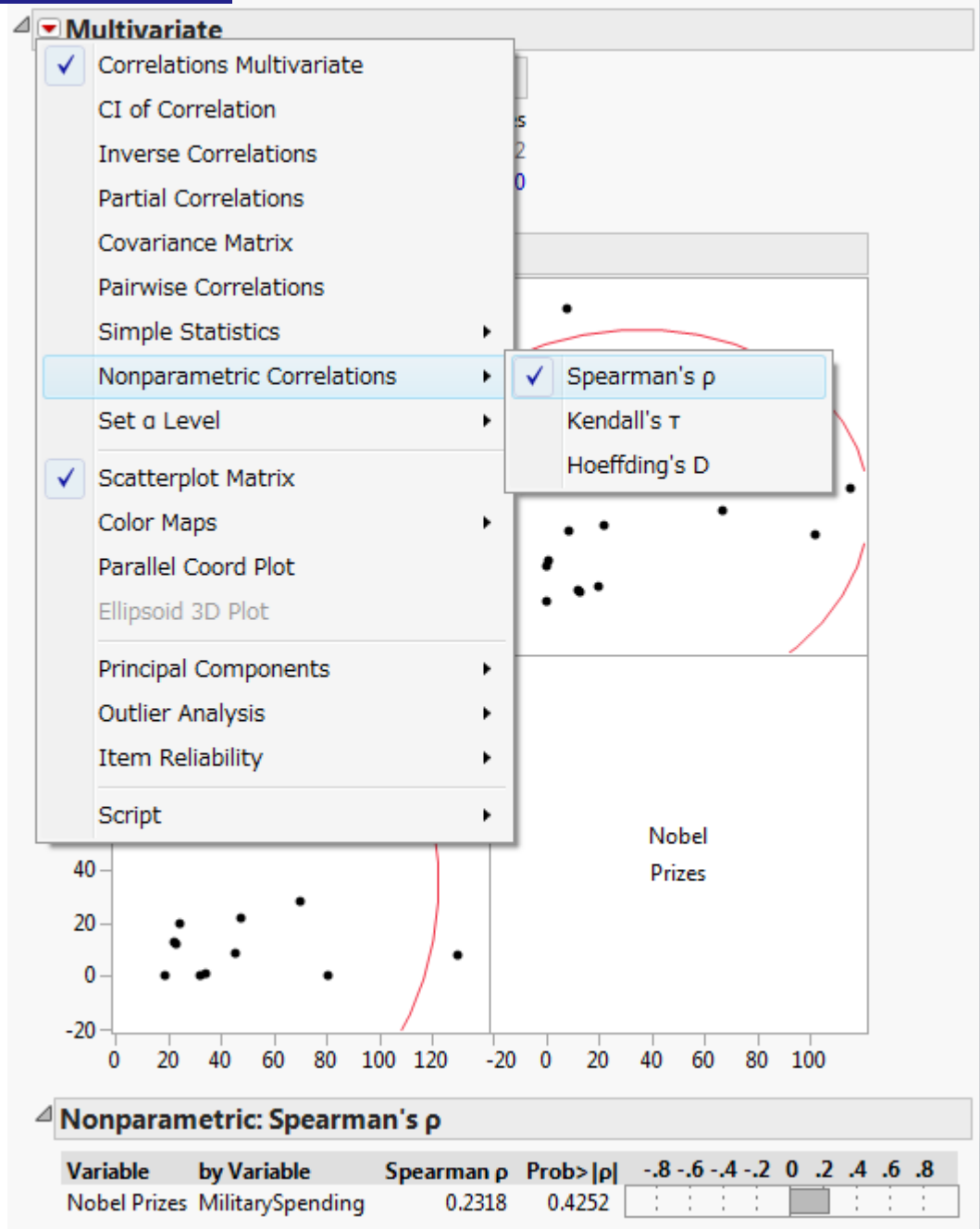


without USA:
ρ=0.2318

Thus, the effect of an extreme case is usually smaller when using Spearman's rank correlation. It can also be used for ordinal data.

Spearman's rank correlation

In JMP:
Analysis->
Multivariate Methods->
Nonparametric
Correlations->
Spearman's ρ



p-value

What is a p-value?

"Informally, the p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value."

ASA Statement on statistical significance and p-values, The American Statistician, 2016

p-value: principles

1. P-values can indicate how compatible the data are with a specified statistical model.

This specified statistical model is commonly the null hypothesis, e.g., no difference between two groups. However, many assumptions go into how this null hypothesis is modeled.

(Bold font: ASA statement; Non-bold: explanation/comment by C.F.Altmann)
ASA Statement on statistical significance and p-values, The American Statistician, 2016

p-value: principles

2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

Probability that the studied hypothesis H_0 is true:

$P(H_0|D)$ D: observed data

P-value: $P(D|H_0)$

ASA Statement on statistical significance and p-values, The American Statistician, 2016

p-value: principles

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

The widely used threshold of $\alpha=0.05$ is arbitrary.

ASA Statement on statistical significance and p-values, The American Statistician, 2016

p-value: principles

4. Proper inference requires full reporting and transparency.

When conducting statistical tests, as much information as possible should be given: which test is used and why and which assumptions are made.

Furthermore, if multiple analyses were conducted, it is not acceptable to report only those that resulted in significant p-values. All analysis steps and decisions have to be reported.

ASA Statement on statistical significance and p-values, The American Statistician, 2016

p-value: principles

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Small p-values do not mean that an effect is important (such as a strong effect of BMI on blood pressure). Large p-values do not mean that there is no effect, this could be a result of a study with small sample size.

Thus it is important not only to report the p-value, but also an estimate for the effect (if applicable).

ASA Statement on statistical significance and p-values, The American Statistician, 2016

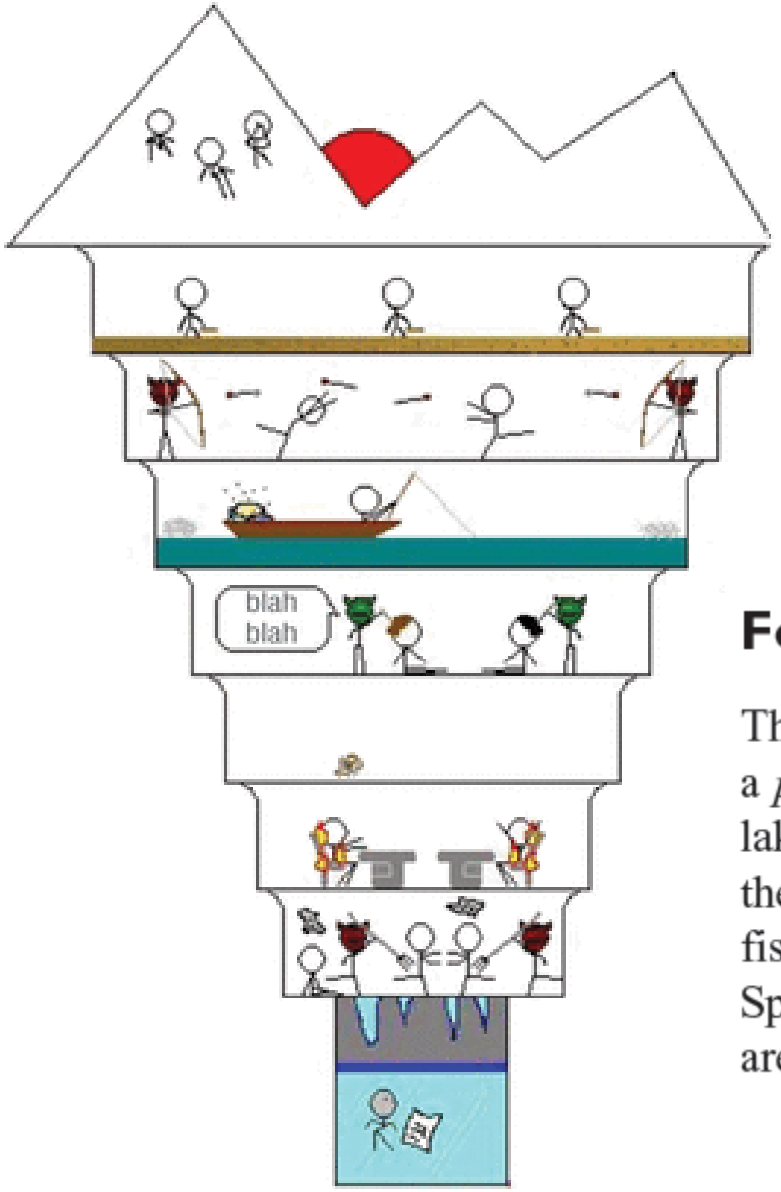
p-value: principles

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

So, context information should be provided and if possible the alternative hypothesis be developed.

ASA Statement on statistical significance and p-values, The American Statistician, 2016

Sin: p-value fishing



I Limbo

II Overselling

III Post-Hoc Storytelling

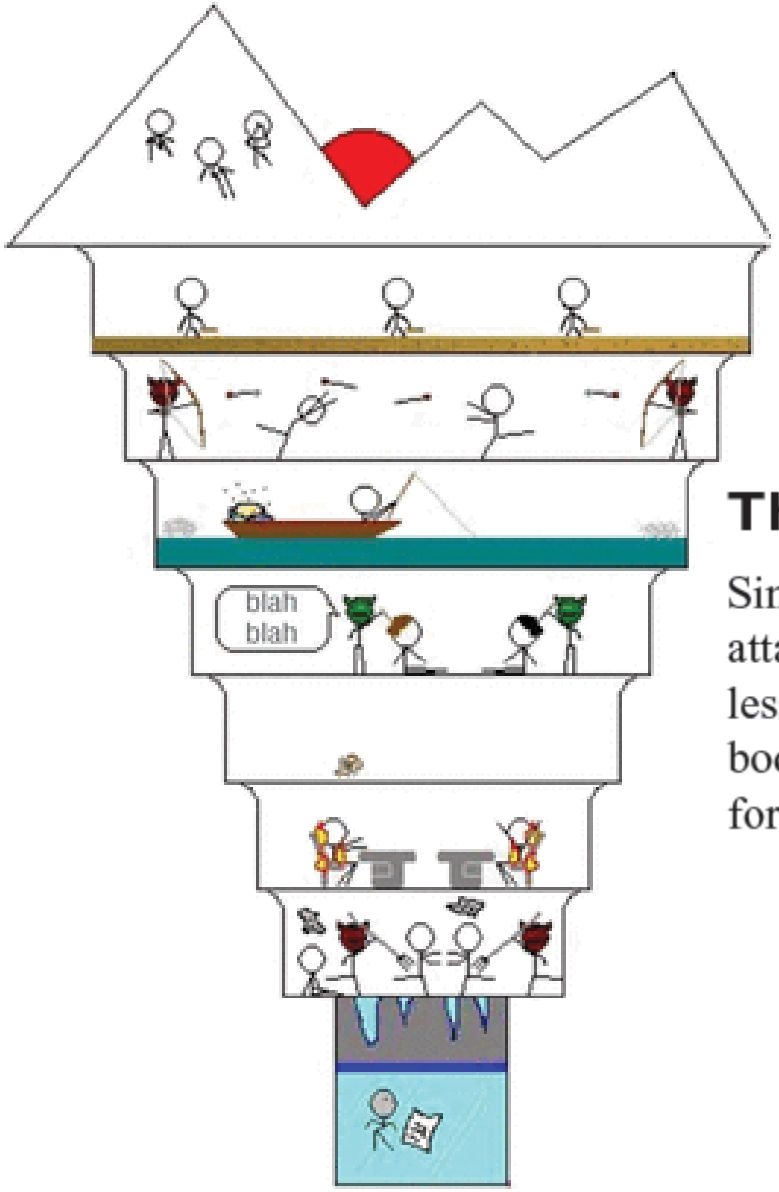
IV P-Value Fishing

Fourth Circle: *p* Value Fishing

Those who tried every statistical test in the book until they got a *p* value less than .05 find themselves here, in an enormous lake of murky water. Sinners sit on boats and must fish for their food. Fortunately, they have a huge selection of different fishing rods and nets (brand names include Bayes, Student, Spearman, and many more). Unfortunately, only one in 20 fish are edible, so the sinners in this circle are constantly hungry.

Neurosceptic, Perspectives on Psychological Science, 2012

Sin: post-hoc storytelling



I Limbo

II Overselling

III Post-Hoc Storytelling

Third Circle: Post-Hoc Storytelling

Sinners condemned to this circle must constantly dodge the attacks of demons armed with bows and arrows, firing more or less at random. Every time someone is hit in some part of their body, a demon proceeds to explain at length that it was aiming for that exact spot all along.

VIII Partial Publication

IX Inventing Data

Neurosceptic, Perspectives on Psychological Science, 2012

Summary

- Outliers can heavily influence a correlation: in such cases data trimming or non-parametric correlation (on ranked data) might be appropriate.

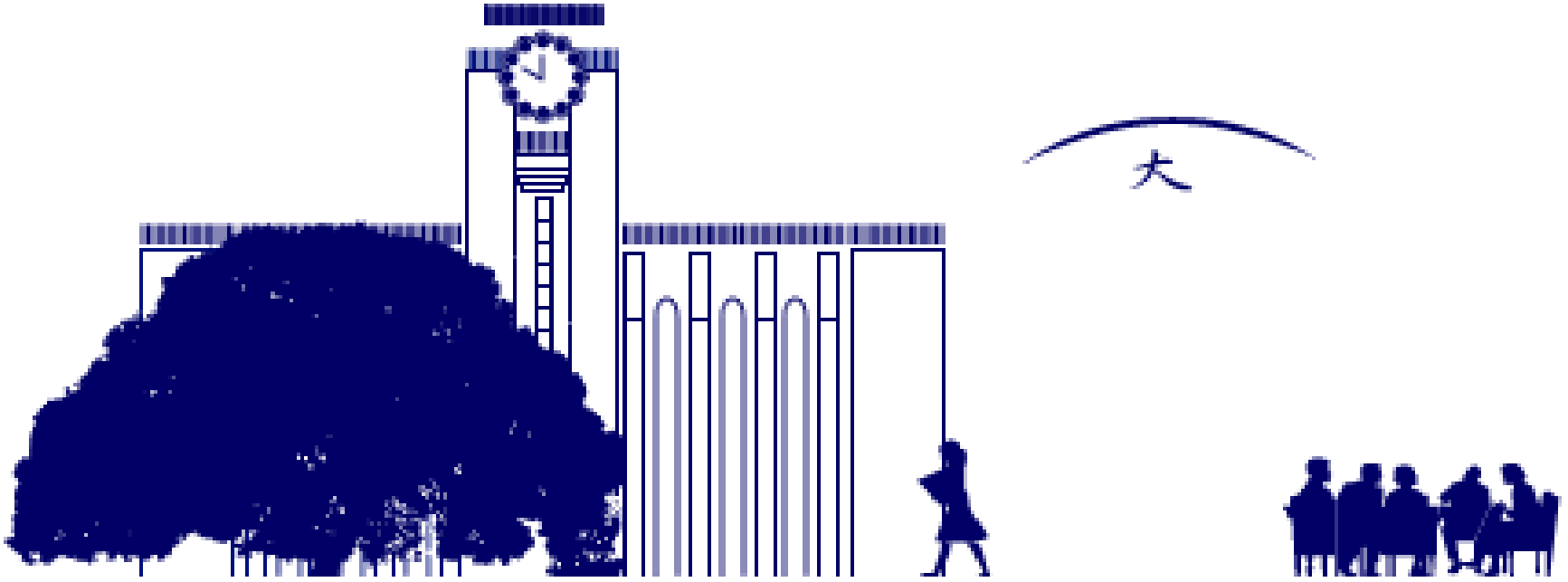
- Statistical p-values are often “misused”: it is important to know what they are.

- Researchers can easily delude themselves and others when using statistics: beware!

Introductory Statistics

Class 14: Further Studies

Richard Veale
Graduate School of Medicine
Center of Medical Education



Today's topics

Further studies

1) Analysis of Variance (ANOVA)

2) Cluster analysis

3) Testing if a variable conforms to a probability distribution

4) Bayes' theorem

5) Course evaluation

ANOVA - example

Let's say we have three groups of depressive patients.
The three groups are treated with placebo, a new antidepressant on a single dose, or the antidepressant with a double dose.

Pre-tests have shown similar levels of depression in the three groups, after 6 weeks they are retested with a depression inventory. This questionnaire asks about wellbeing, suicidal tendencies, loss of appetite, etc.:

Scoring:
0-8: no depression
9-13: minimal depression
14-19: mild depression
20-28: moderate depression
>29: severe depression

ANOVA - example

Placebo	Single Dose	Double Dose
18	19	16
22	16	13
25	16	12
19	15	12
22	17	14
19	16	16
21		13
		13
		14

Group means $\bar{y}_1 = 20.86$ $\bar{y}_2 = 16.5$ $\bar{y}_3 = 13.67$

Grand mean: $\bar{y} = 16.72$

Are these group means significantly different from each other?

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

H_a: alternative hypothesis: the means are not equal

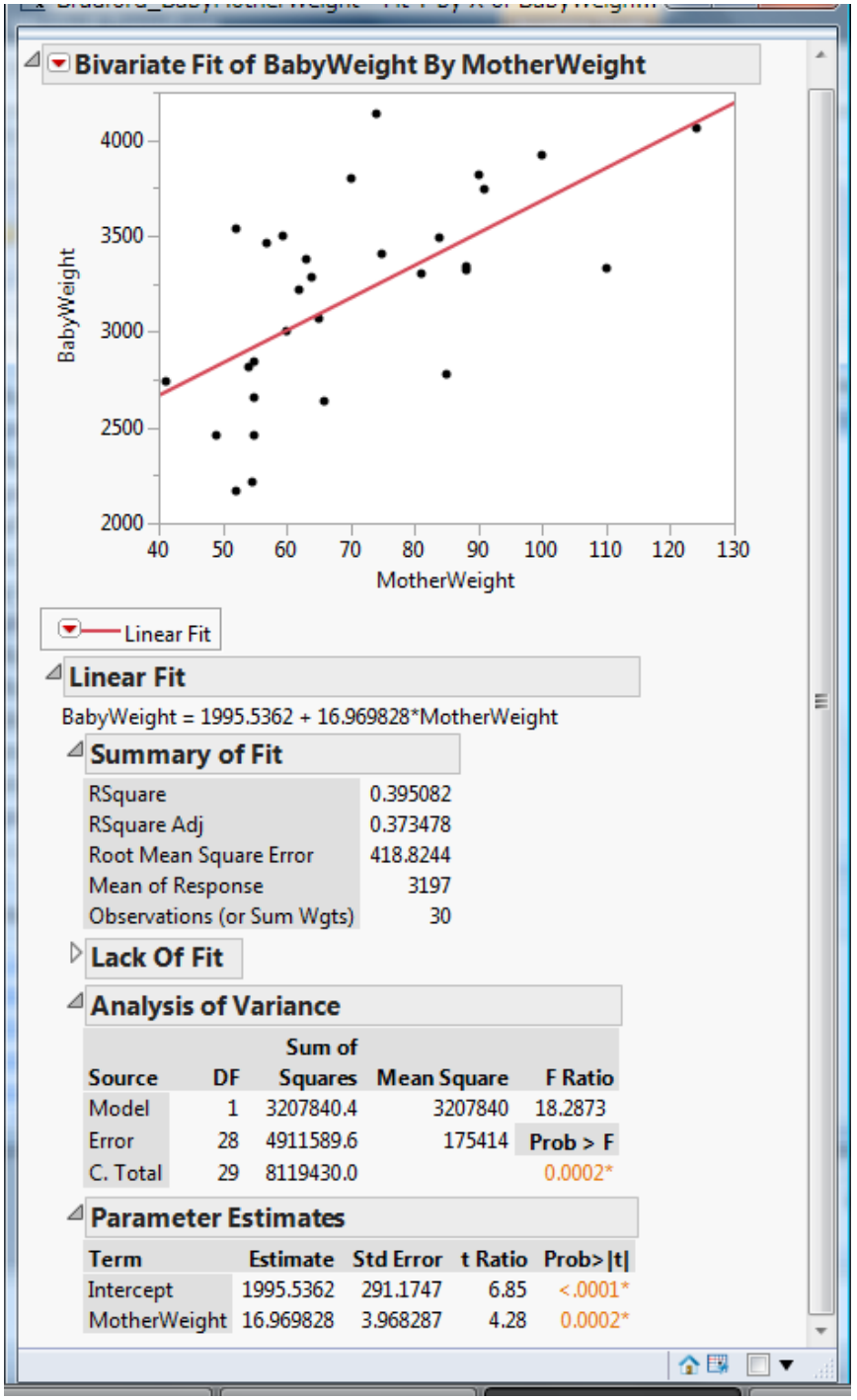
Linear regression: F-Test

Significance test for the whole linear regression model:

$$F = \frac{MS_M}{MS_R} = \frac{SS_M/df_M}{SS_R/df_R}$$

df_M: model degrees of freedom = number of x-variables

df_R: residual degrees of freedom = number of observations – number of estimated parameters (a,b)

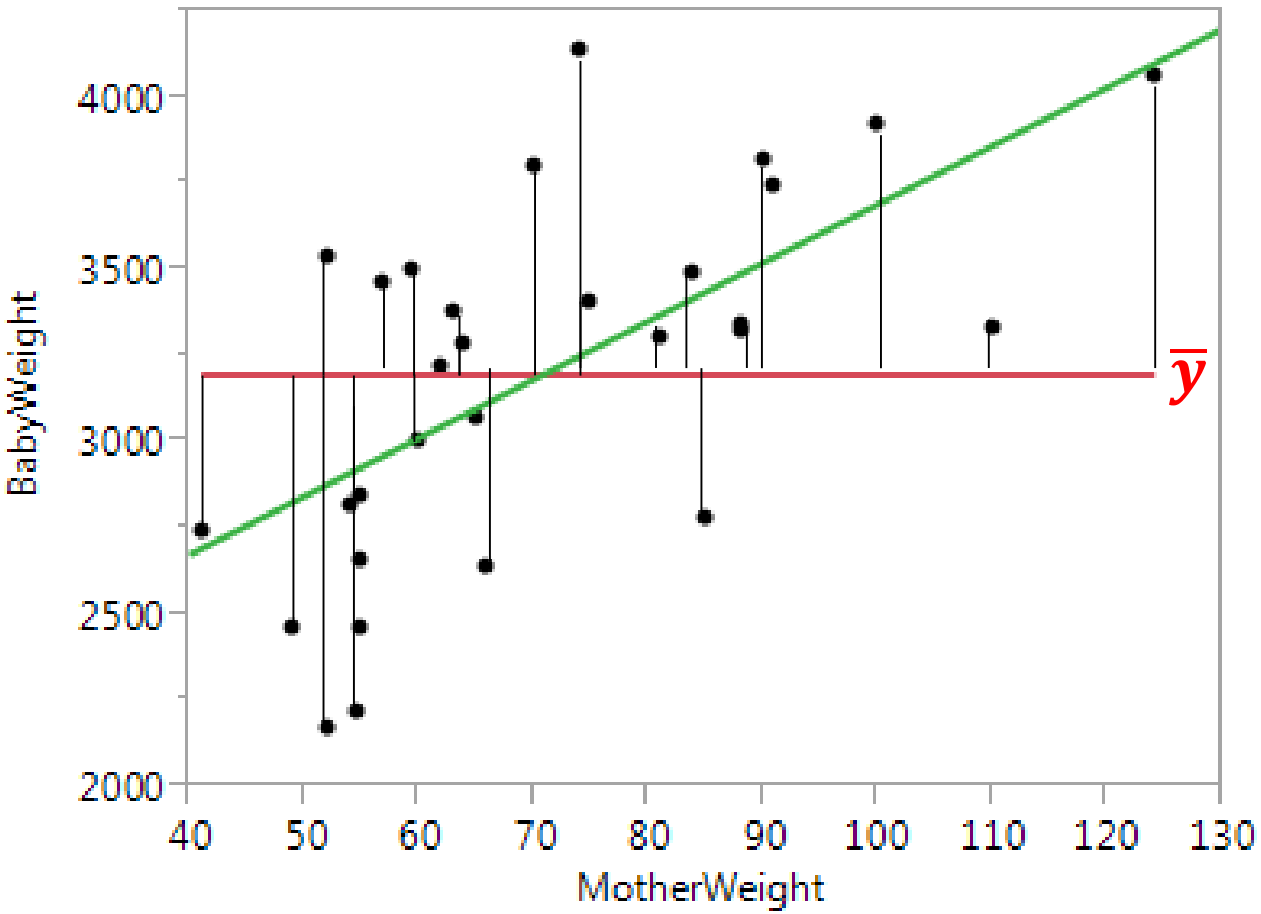


Linear Regression – testing the model

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Total sum of squares

$a + bx_i$



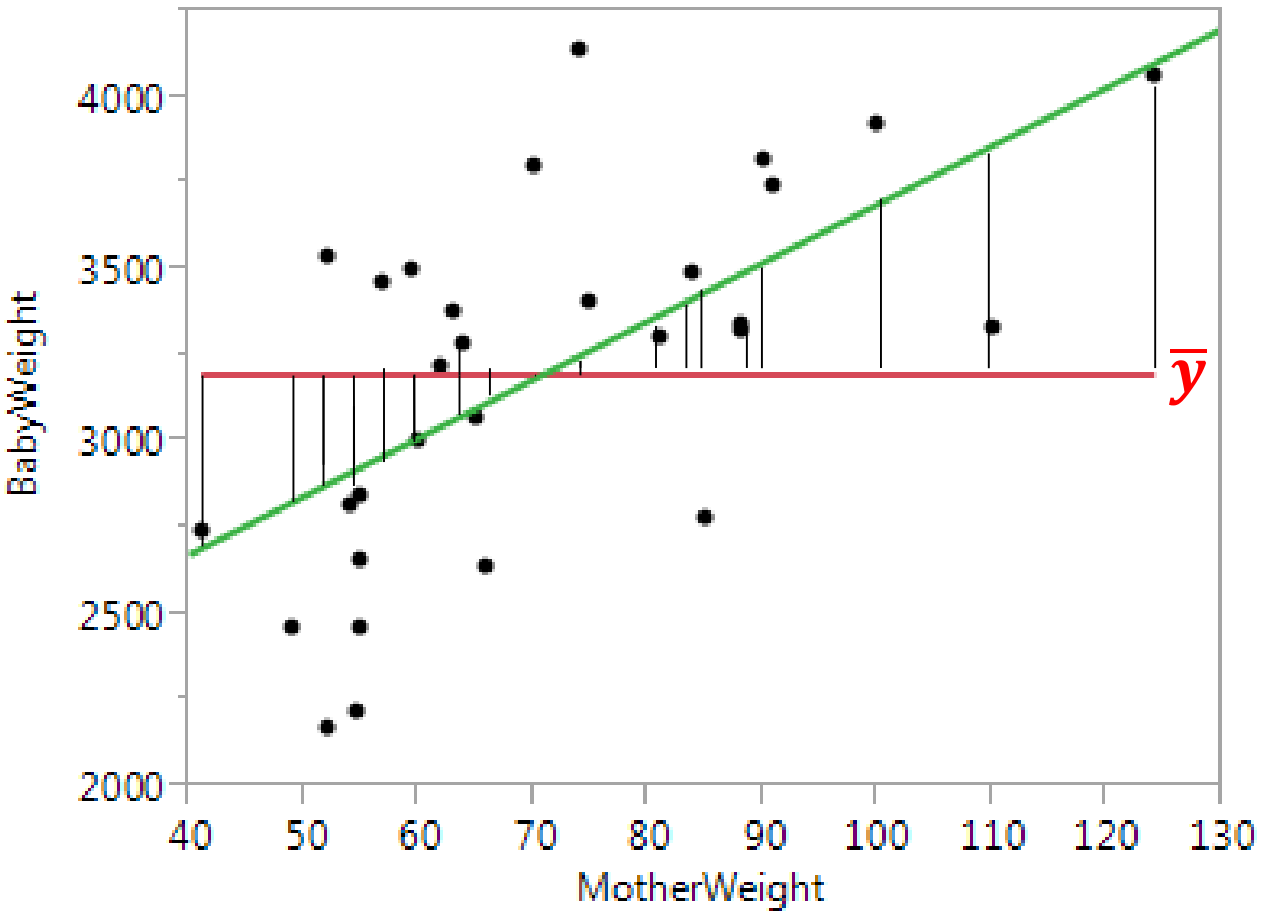
Linear Regression – testing the model

$$SS_M = \sum_{i=1}^{n} (a + bx_i - \bar{y})^2$$

Model sum of squares

$a + bx_i$

$\mathrm{df_M}$ = number of x-variables (here: 1)



Linear Regression – testing the model

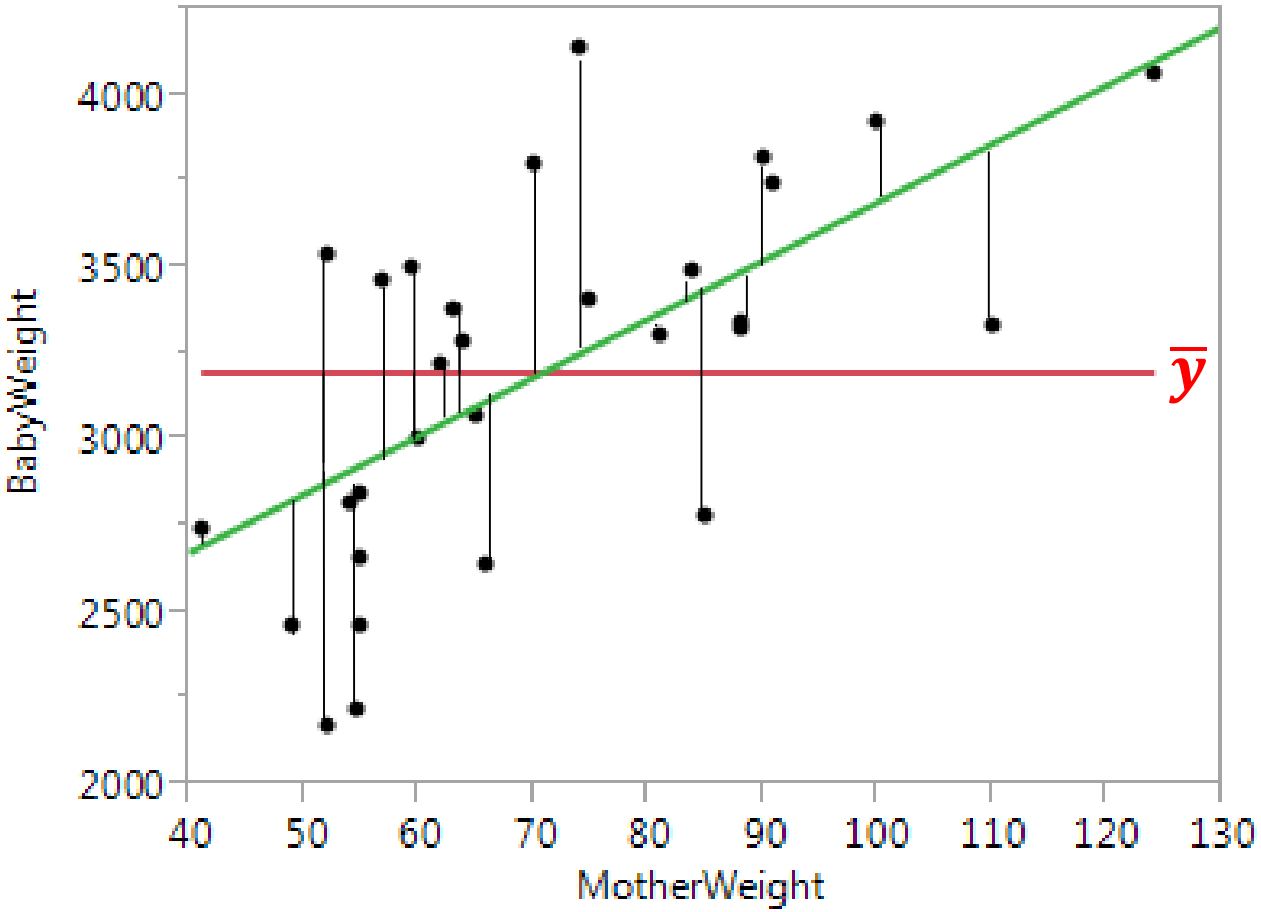
$$SS_R = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Residual sum of squares (error)

$$SS_T = SS_M + SS_R$$

$a + bx_i$

df_R: = number of observations – number of estimated parameters (a,b) (here: 30-2=28)

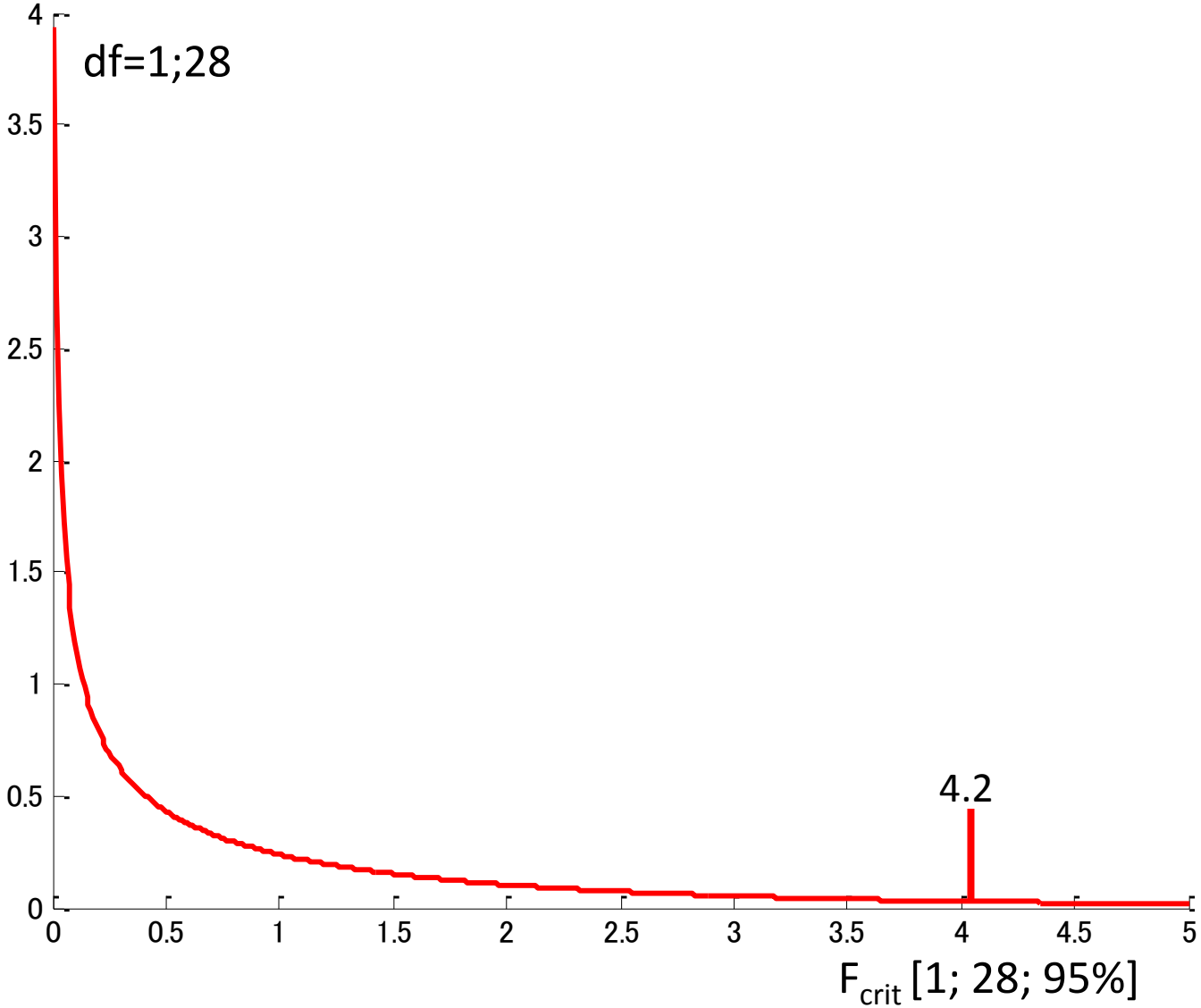


F-Test

Under the Null hypothesis (that b=0):

$P(F < F_{crit}) = \alpha = 0.05$

$$F = \frac{MS_M}{MS_R} = \frac{SS_M/df_M}{SS_R/df_R}$$



ANOVA - model

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

H_a: alternative hypothesis: the means are not equal

$$y_i = b_o + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + \varepsilon_i$$

Effect coding:

$x_{1i} = 1$ when patient i belongs to group 1, else $x_{1i} = 0$

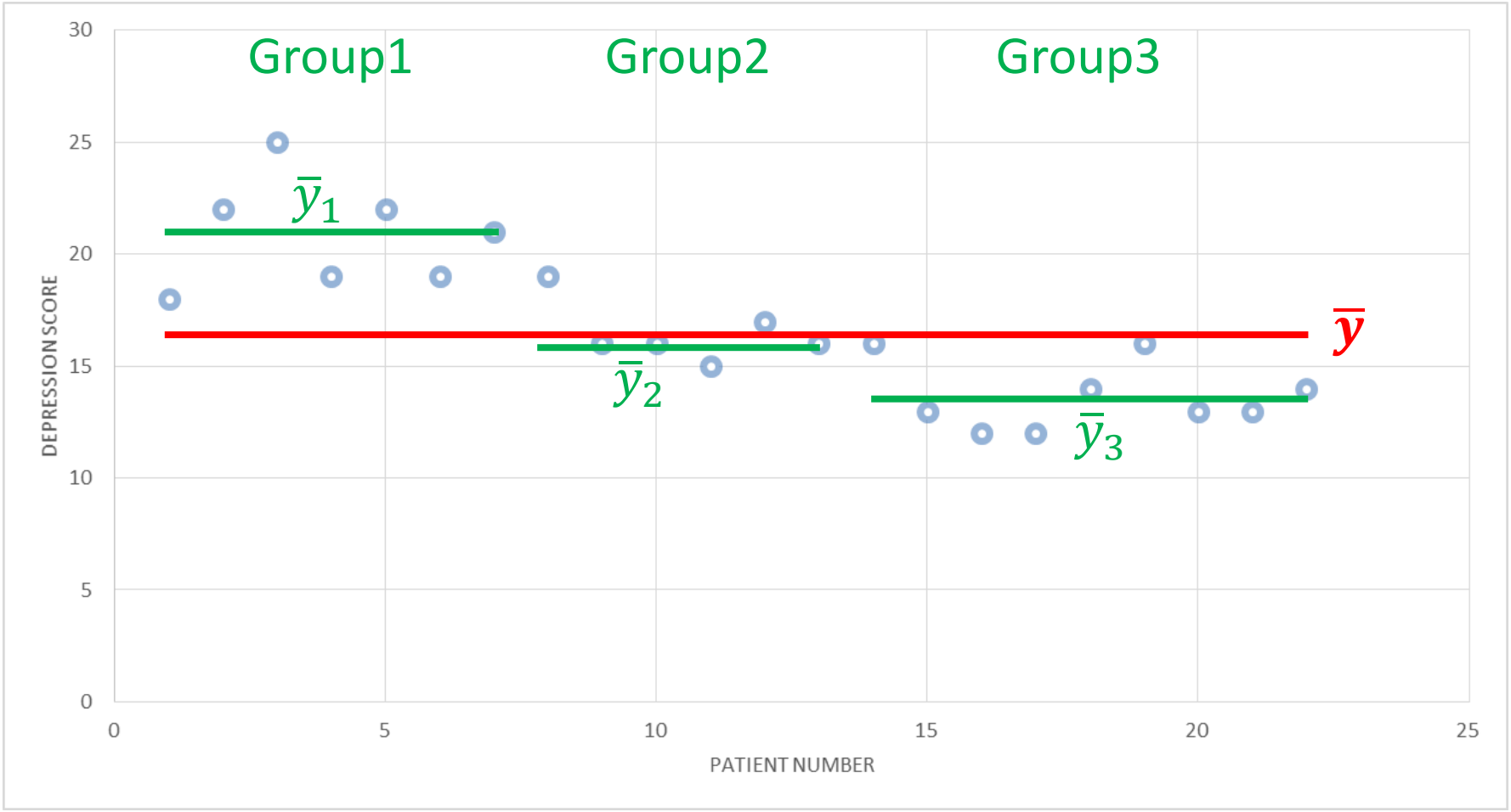
$x_{2i} = 1$ when patient i belongs to group 2, else $x_{2i} = 0$

$x_{3i} = 1$ when patient i belongs to group 3, else $x_{3i} = 0$

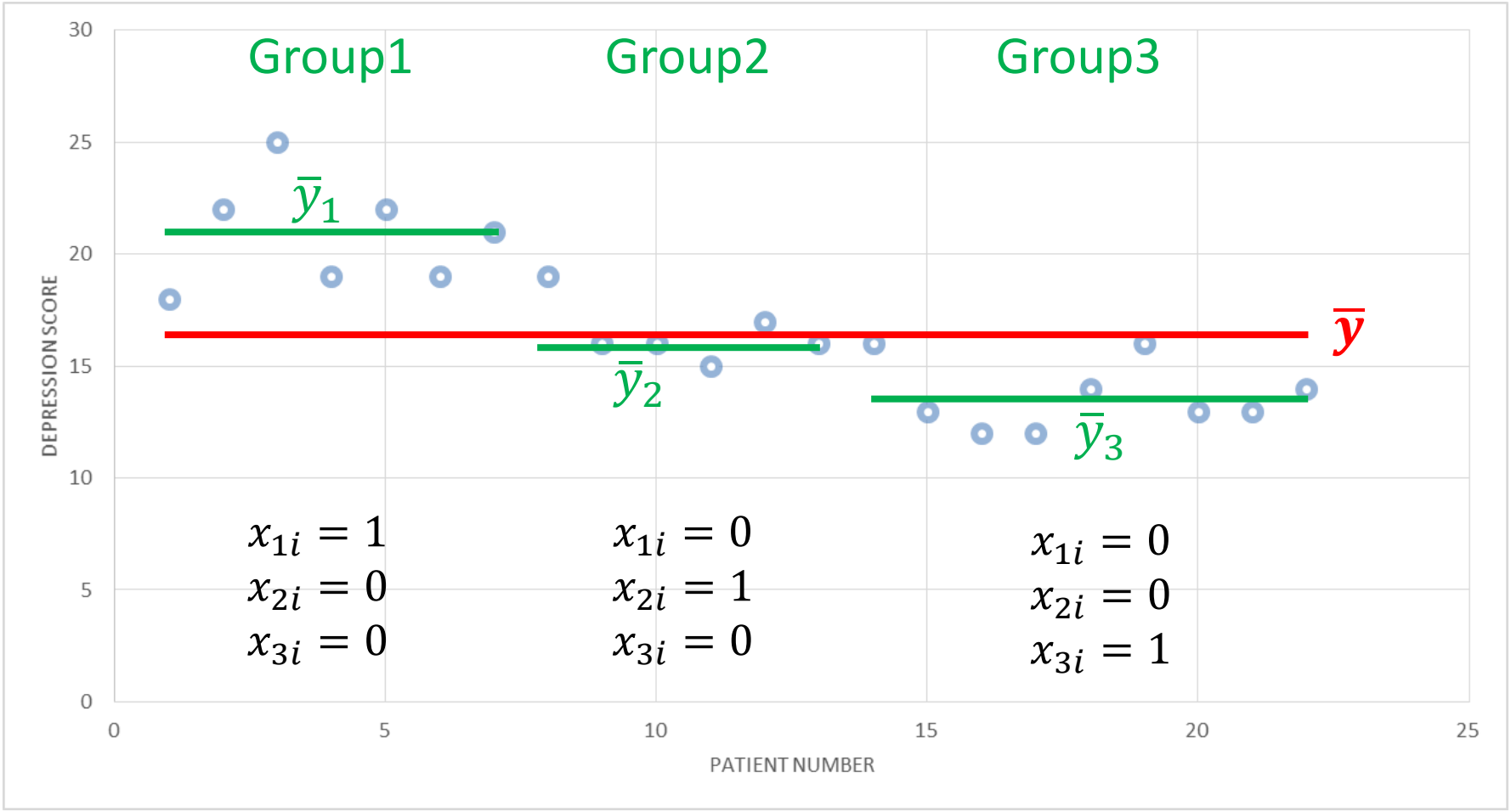
$$b_o = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$b_1 = \bar{y}_1 - \bar{y}$$
$$b_2 = \bar{y}_2 - \bar{y}$$
$$b_3 = \bar{y}_3 - \bar{y}$$

ANOVA – effect coding

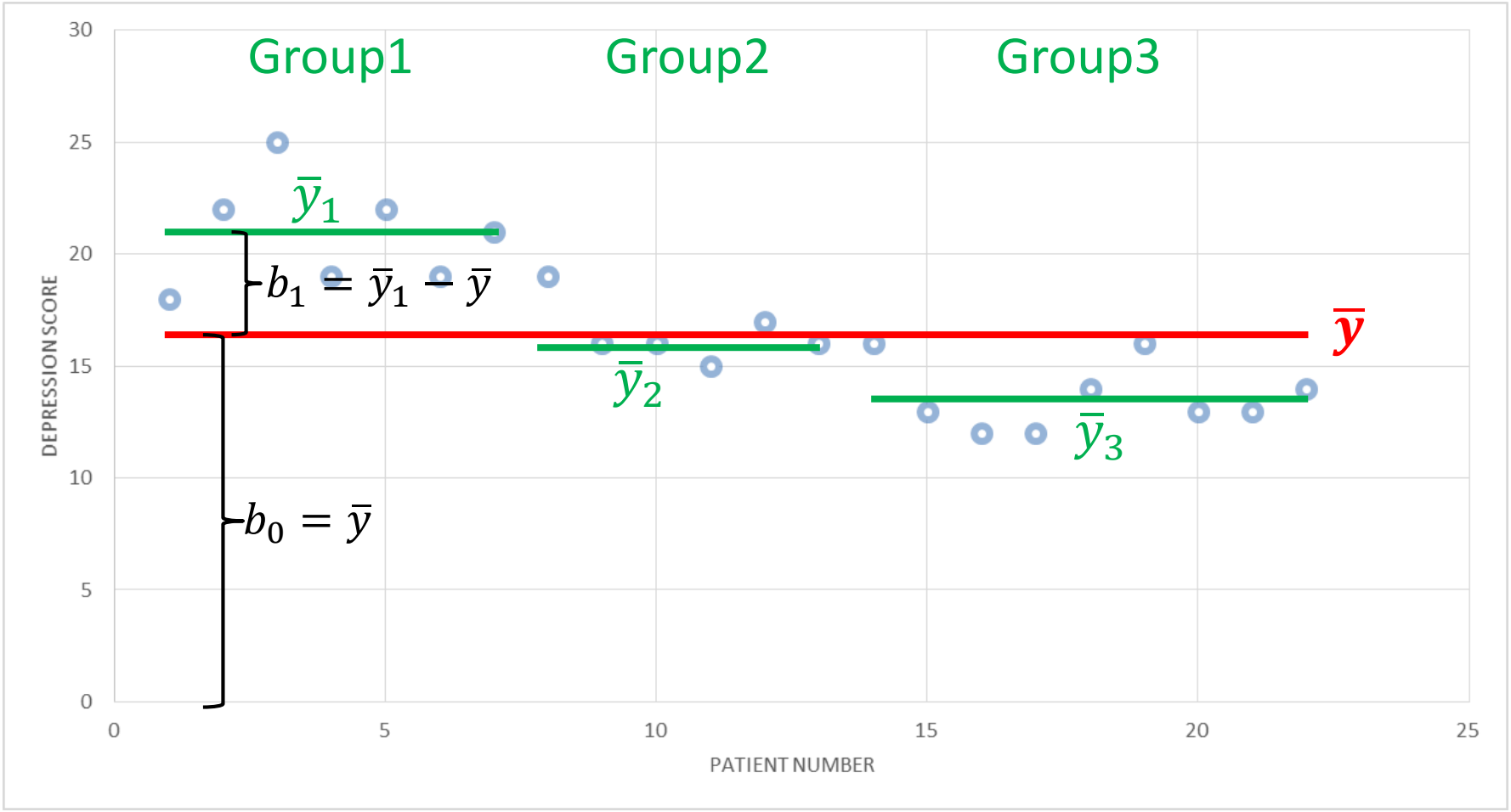


ANOVA – effect coding

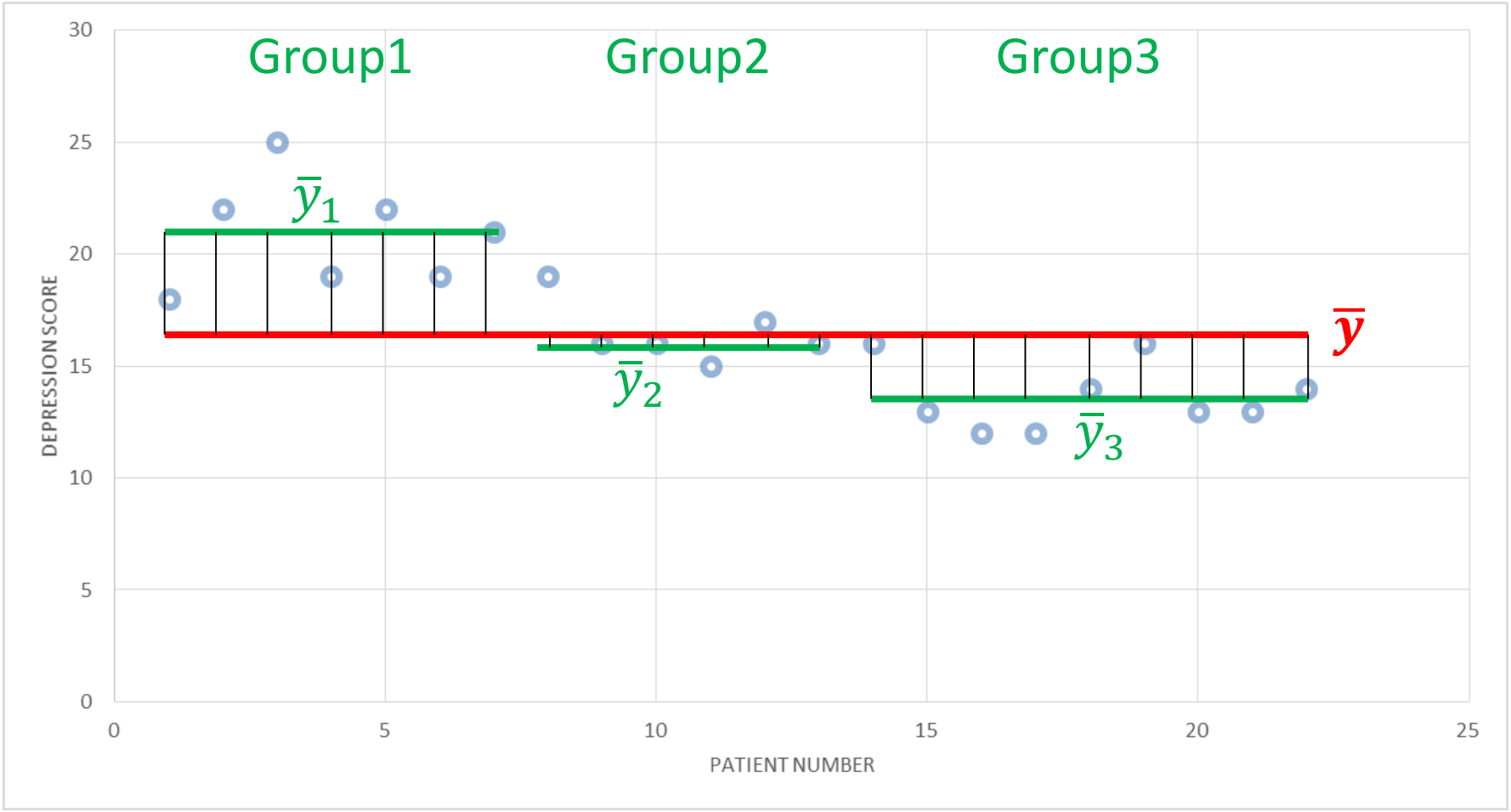


Model for group 1: $y_i = b_o + b_1 + \varepsilon_i = \bar{y} + \bar{y}_1 - \bar{y} + \varepsilon_i = \bar{y}_1 + \varepsilon_i$

ANOVA – effect coding



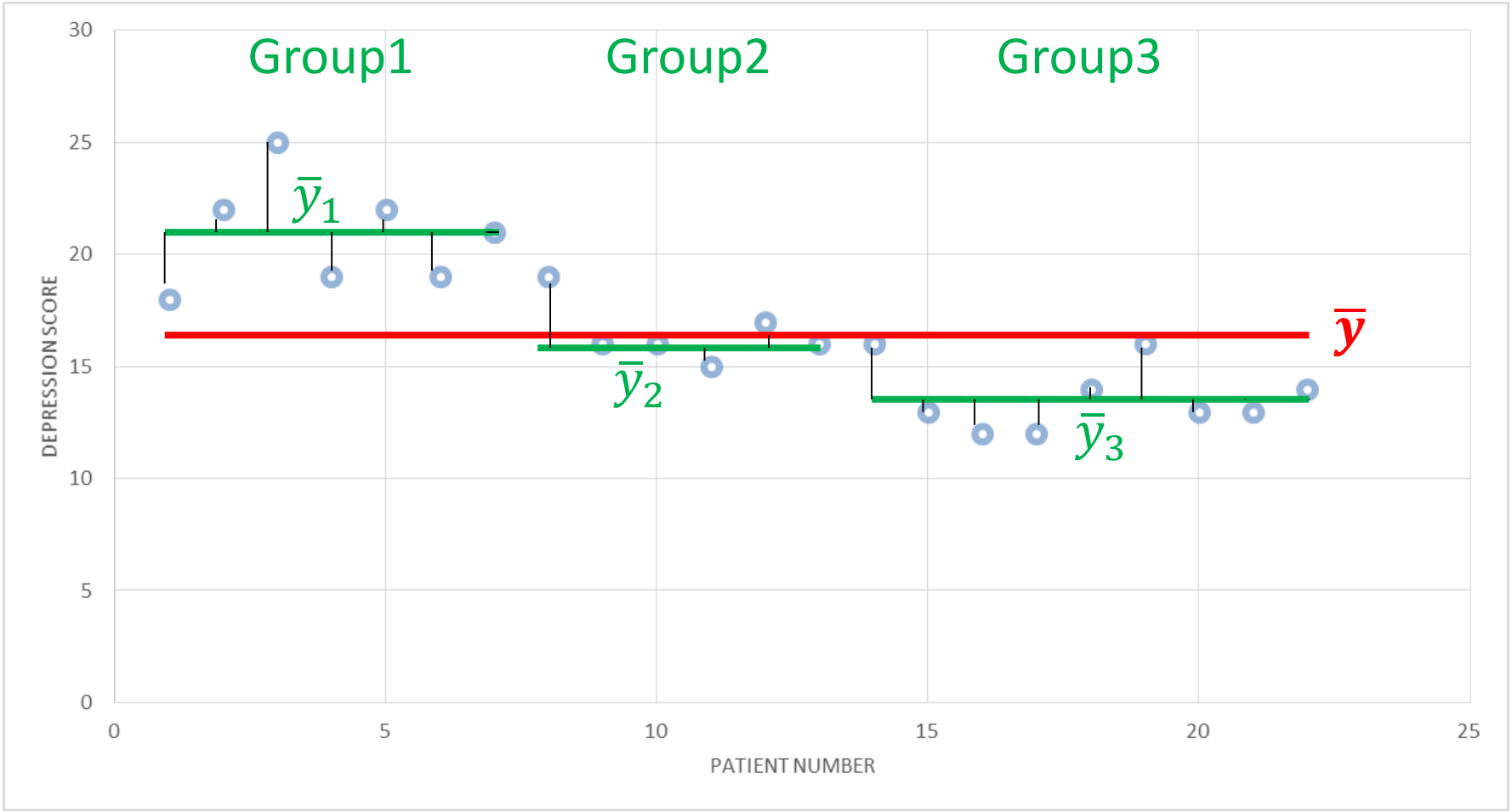
ANOVA – testing the model



$$SS_M = \sum_{i=1}^{n}(b_o + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} - \bar{y})^2 = \sum_{i=1}^{n}(b_1x_{1i} + b_2x_{2i} + b_3x_{3i})^2$$

Model sum of squares

ANOVA – testing the model



$$SS_R = \sum_{i=1}^{n} (y_i - b_o + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i})^2$$

Residual sum of squares (error)

ANOVA – testing the model

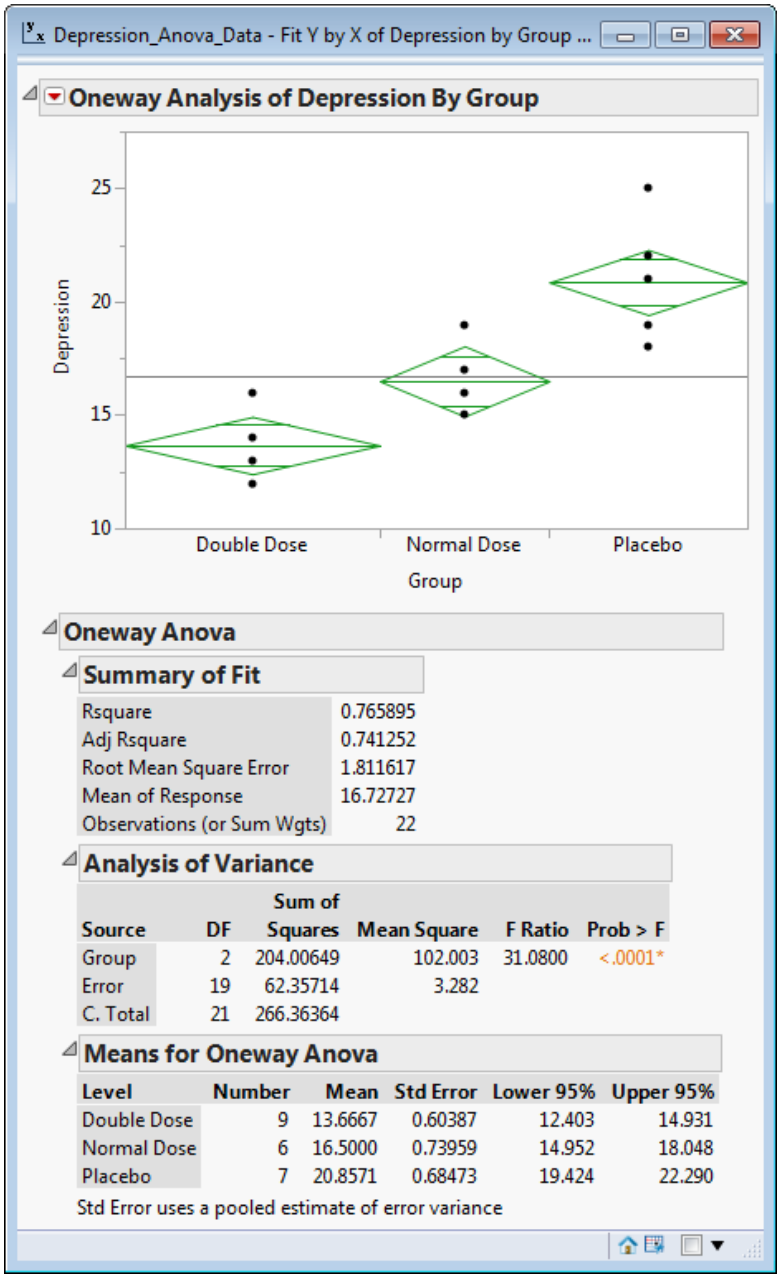
Analysis of variance: F-test:

$$F = \frac{MS_M}{MS_R} = \frac{SS_M/df_M}{SS_R/df_R}$$

df_M: model degrees of freedom = number of groups - 1

df_R: residual degrees of freedom = number of observations - number of groups

Compare with the critical F-value derived from the F[df_M; df_R]-distribution.



ANOVA – requirements

Using the F-test to test for significance assumes the following:

- the distribution within groups is normal.

- variances in each group are similar (homogeneity of variance).

- observations should be independent.

ANOVA – single t-tests

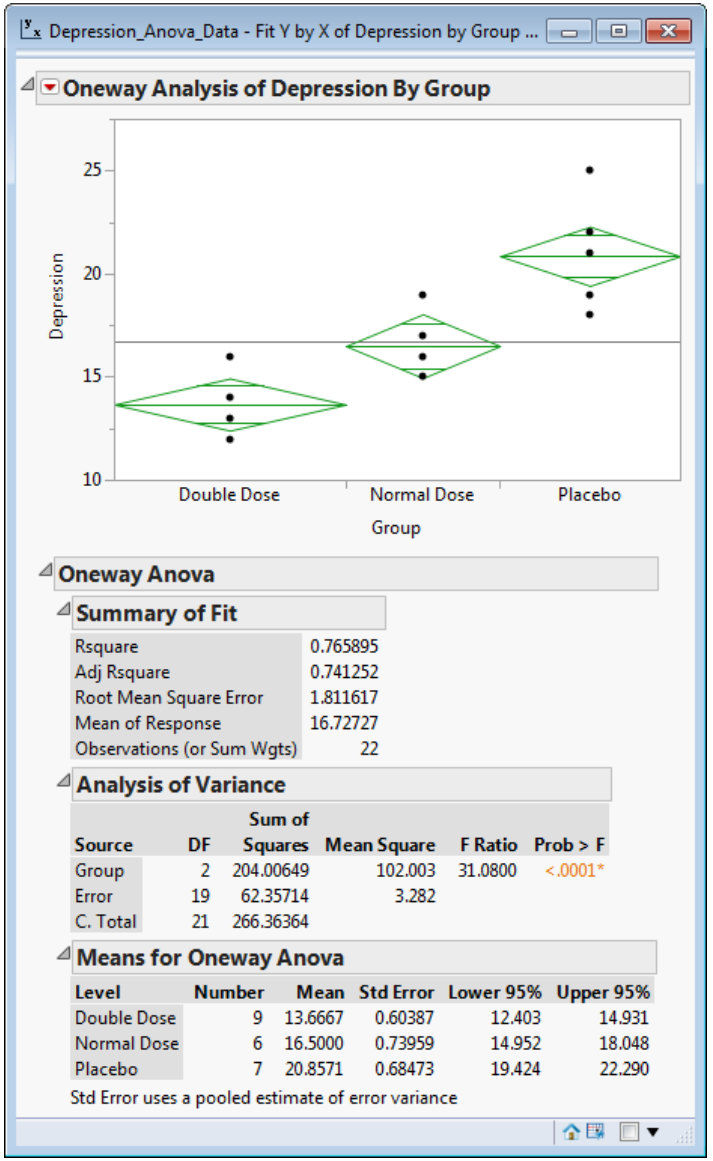
H_0: null hypothesis: $\mu_1 = \mu_2 = \mu_3$

H_a: alternative hypothesis: the means are not equal

"There was a significant effect of treatment group (placebo, single dose, double dose) on depression inventory scores, as tested by one-way independent analysis of variance, F[2,19] = 31.08, p<0.0001."

However, this does not tell us which groups were significantly different from each other (e.g., placebo and single dose).

-> pair-wise t-tests (with correction for multiple comparisons)



ANOVA – alternative methods

If observations are not independent

-> repeated measurements ANOVA

If observations have non-normal distributions

-> non-parametric approaches (e.g., Kruskall-Wallis-test)

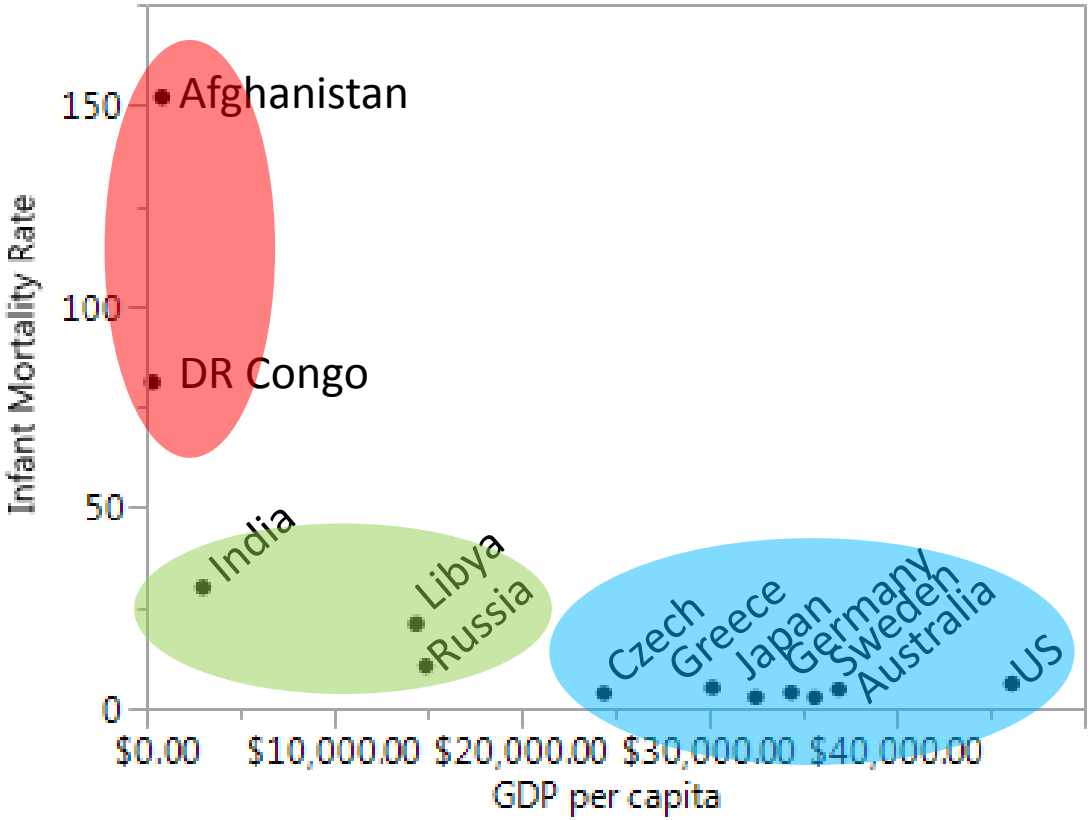
If there is more than one independent variable (factor)

-> factorial ANOVA (e.g., two-way or three-way ANOVA)

Cluster analysis

Let's say you have n entities that are characterized by k variables, can we find a way to put them into meaningful categories/clusters?

For example: different countries characterized by infant mortality and GDP/capita.



Cluster analysis

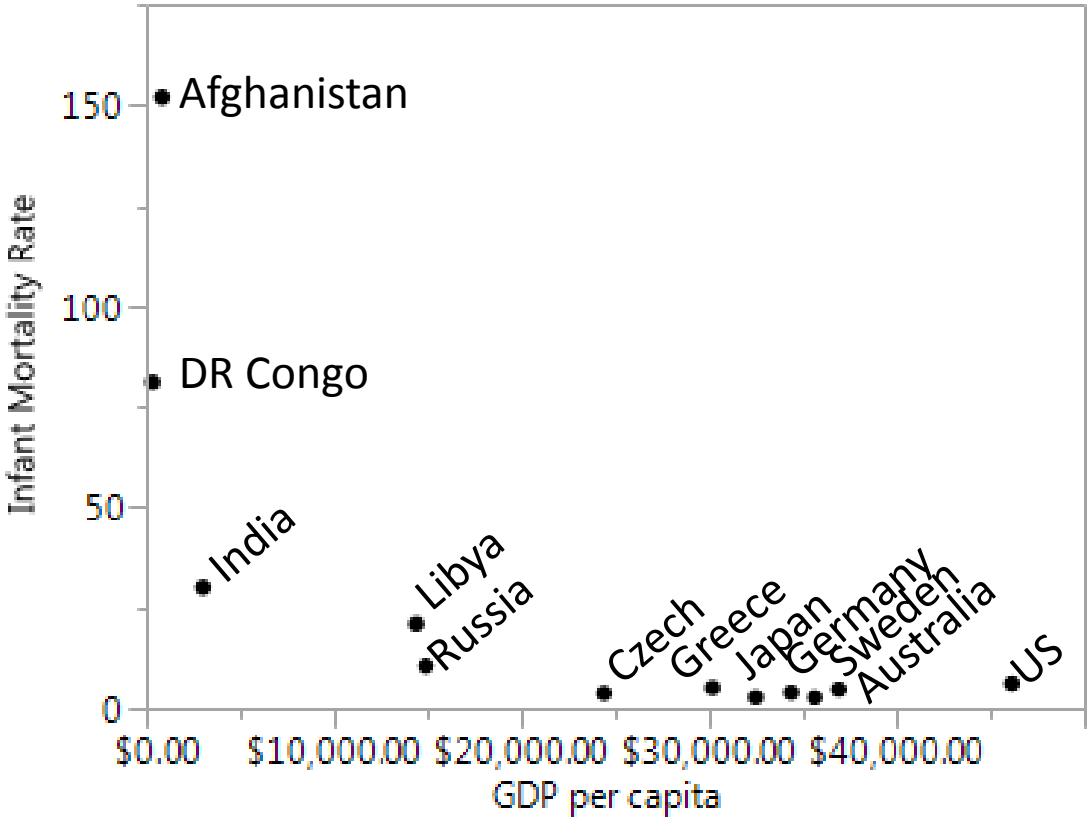
How can we define clusters with similar entities?

First, define distance, e.g.: $d_{ij} = \sqrt{\sum_{l=1}^{k}(x_{il} - x_{jl})^2}$

x_{i1}: e.g., GDP of country i
x_{j2}: e.g., IMR of country j

Variables are often standardized with a z-transform:

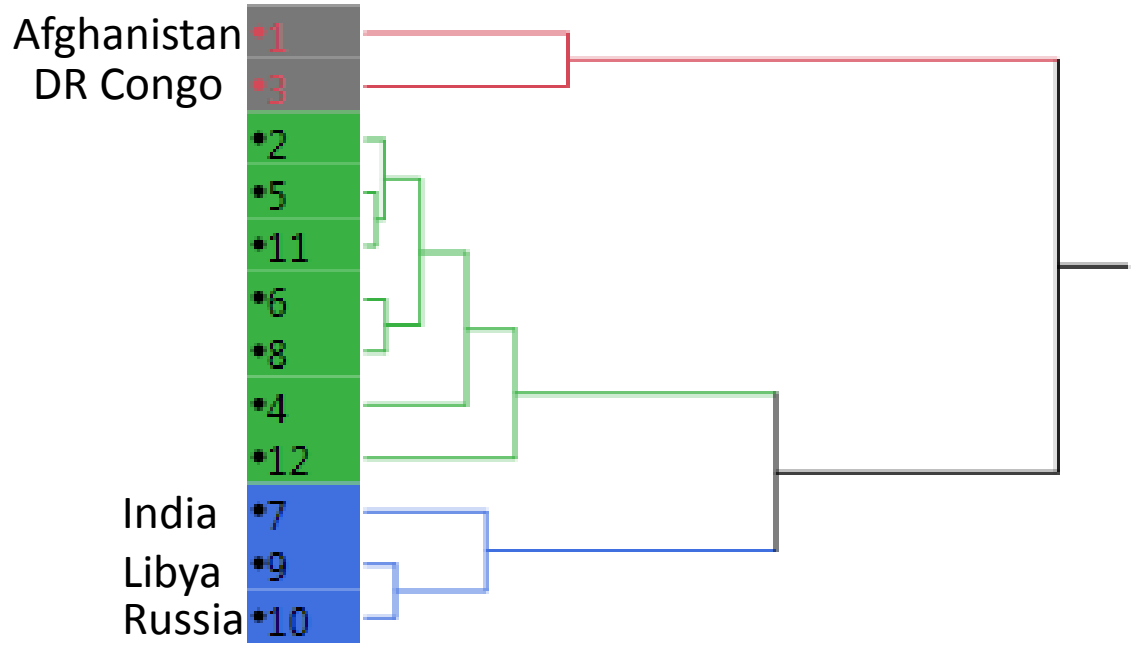
$$z_{il} = \frac{x_{il} - \overline{x_{.l}}}{s(x_{.l})}$$



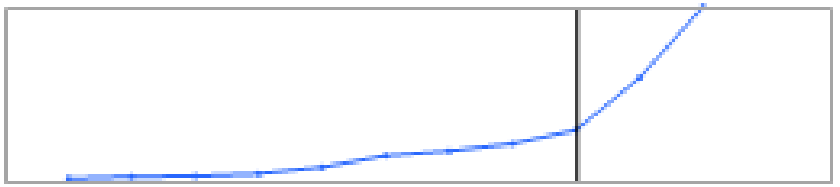
Cluster analysis

A common method (Ward method) for cluster analysis starts with clusters that contain only 1 element. Successively, elements are joined into clusters so that the residual sum of squares (computed like in ANOVA) increases the least.

Dendrogram:



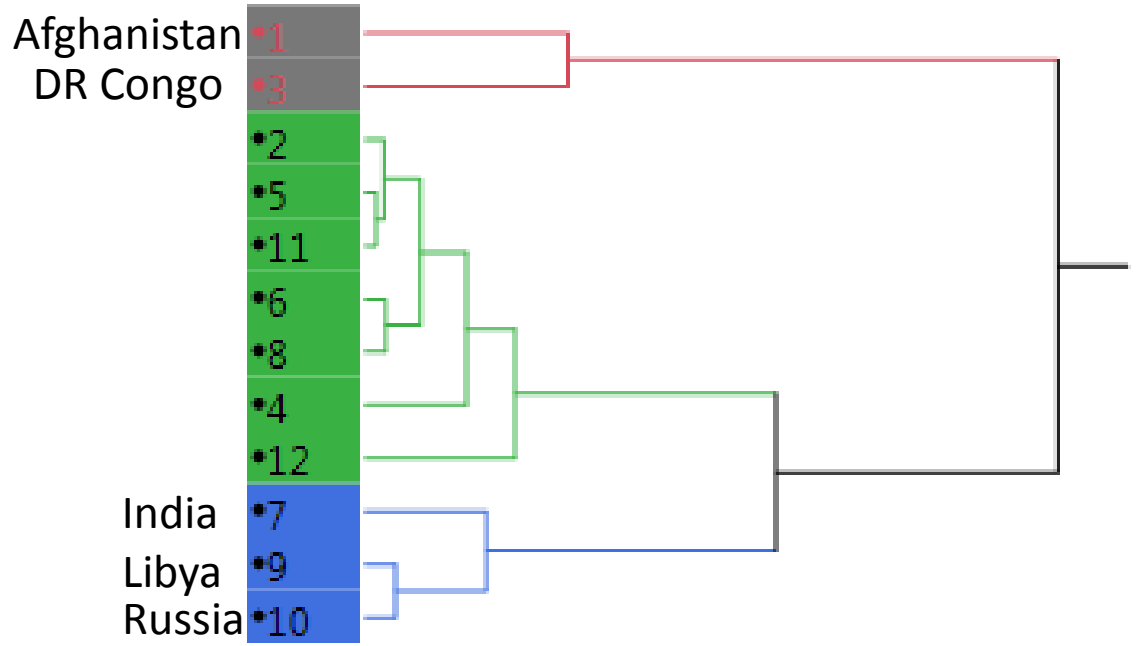
Distance graph: shows the increase in the residual sum of squares with each fusion step.



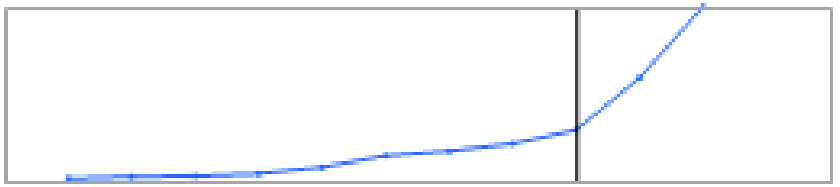
Cluster analysis

The optimal number of clusters can be estimated from the distance graph. Rule of thumb: take the value before the residual sum of squares increase sharply ('elbow'). Other approaches: e.g., Calinski and Harabasz, 1974.

Dendrogram:



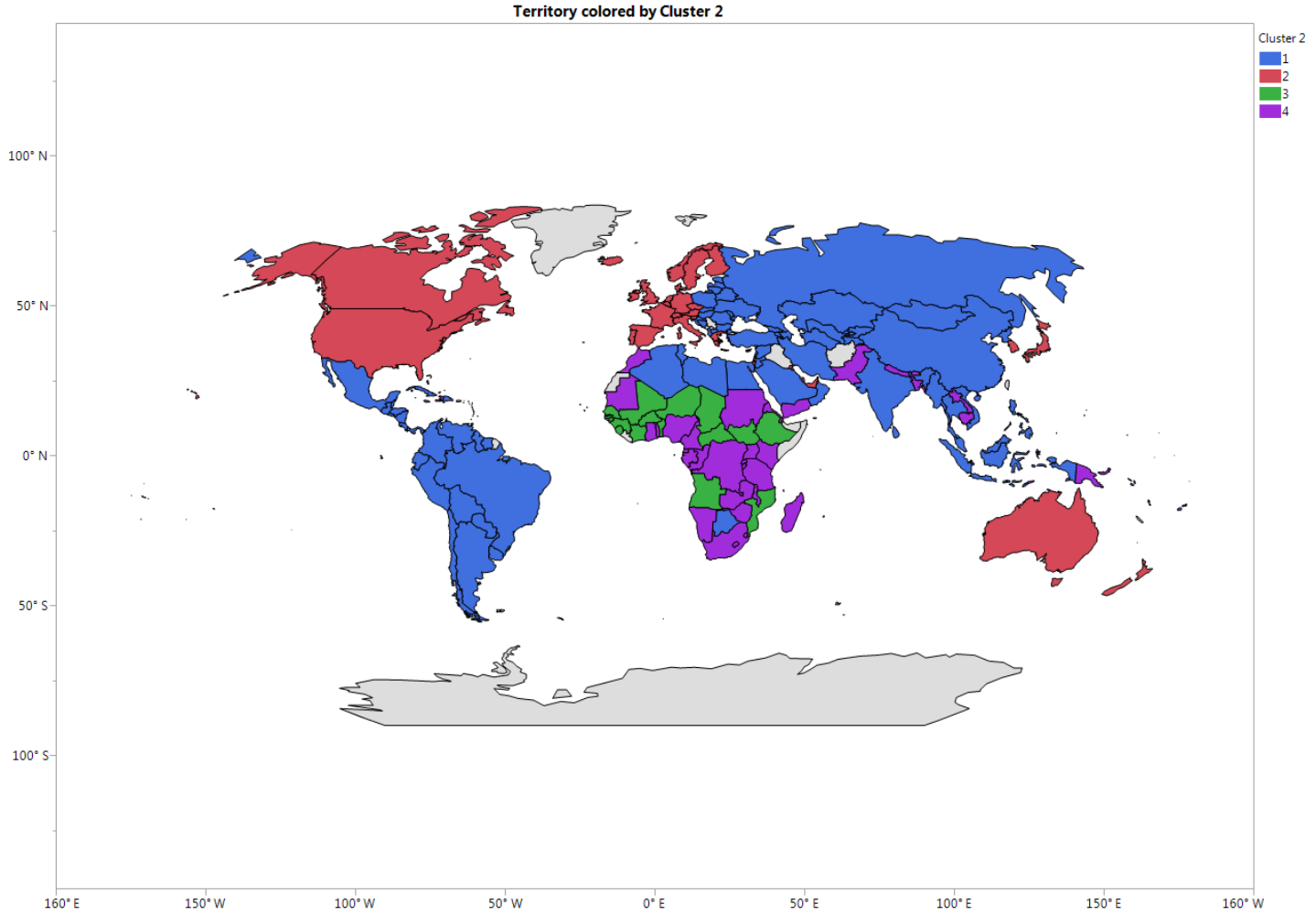
Distance graph: shows the increase in the residual sum of squares with each fusion step.



Cluster analysis

JMP: Analyze->Clustering

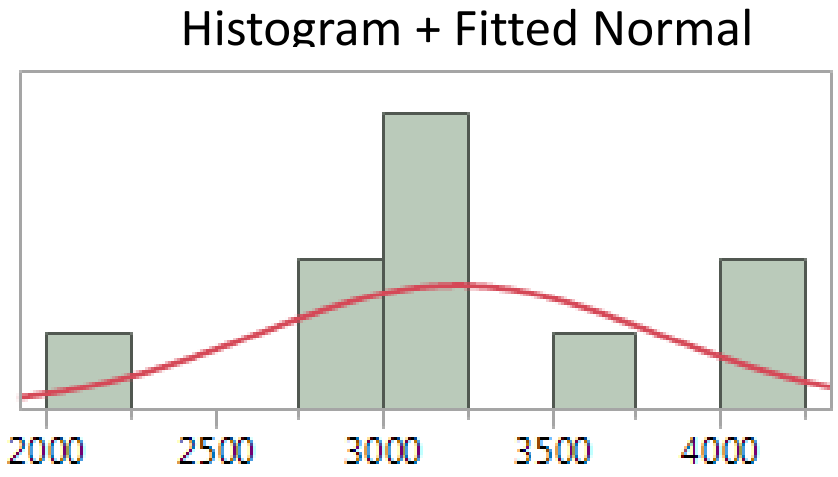
Visualization with Graph Builder (Cluster numbers can be imported into the data table with ‘Save Clusters’). Here, the world’s nations were clustered by infant mortality rate, life expectancy, literacy rate, and GDP per capita.



Test for normality

Many parametric statistical tests require a normal distribution of the investigated variable. Can we test this?

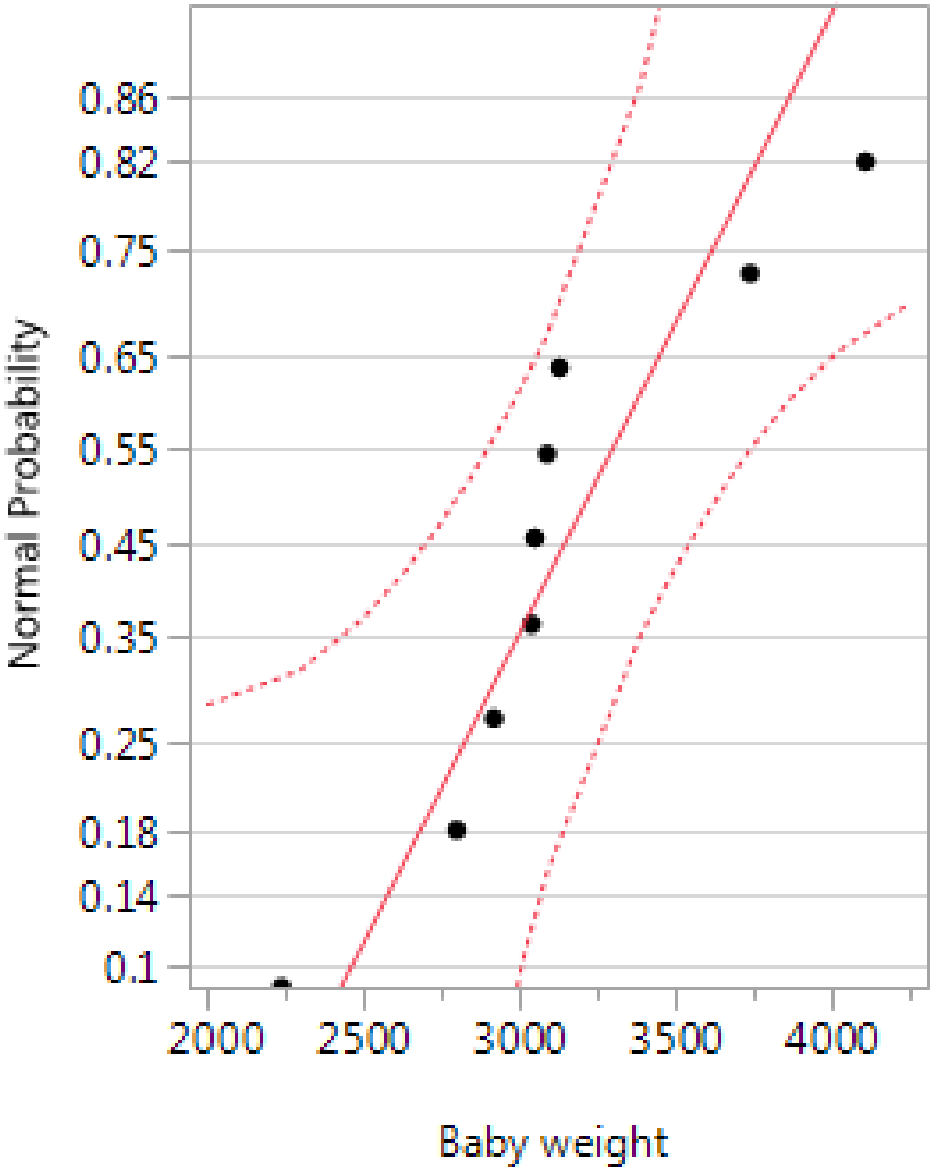
For example Bradford's baby weights (Assignment 1):
Smoked during pregnancy:
2240 3050 4110 3740 3040 2920 2800 3090 4110 3130



$\bar{x}_1 = 3223$
$s_1 = 593$

Test for normality

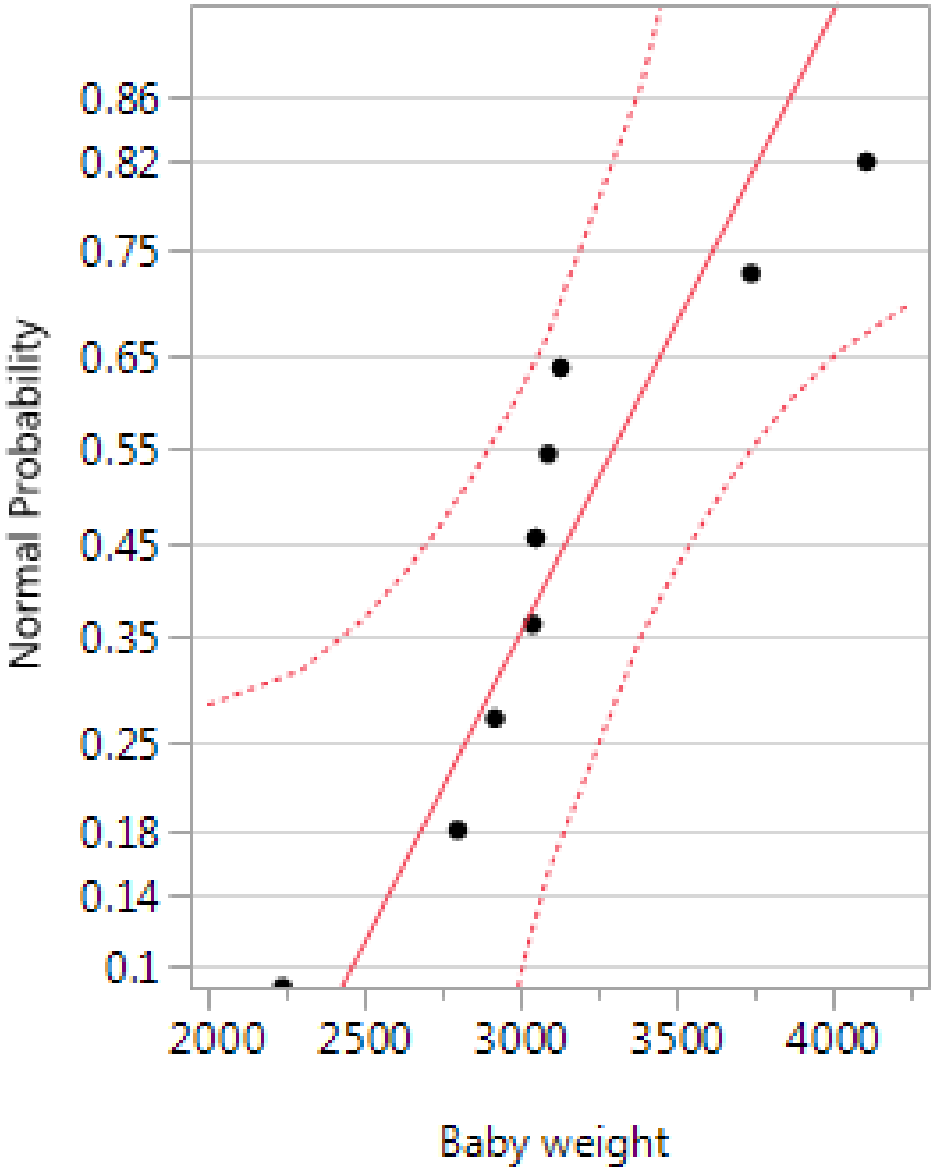
Normal quantile plot or Q-Q plot:



Normal quantile plot:
Plots the observed quantiles versus the expected quantiles.

"Expected": under the assumption the data is normally distributed.

Test for normality: Shapiro-Wilk test

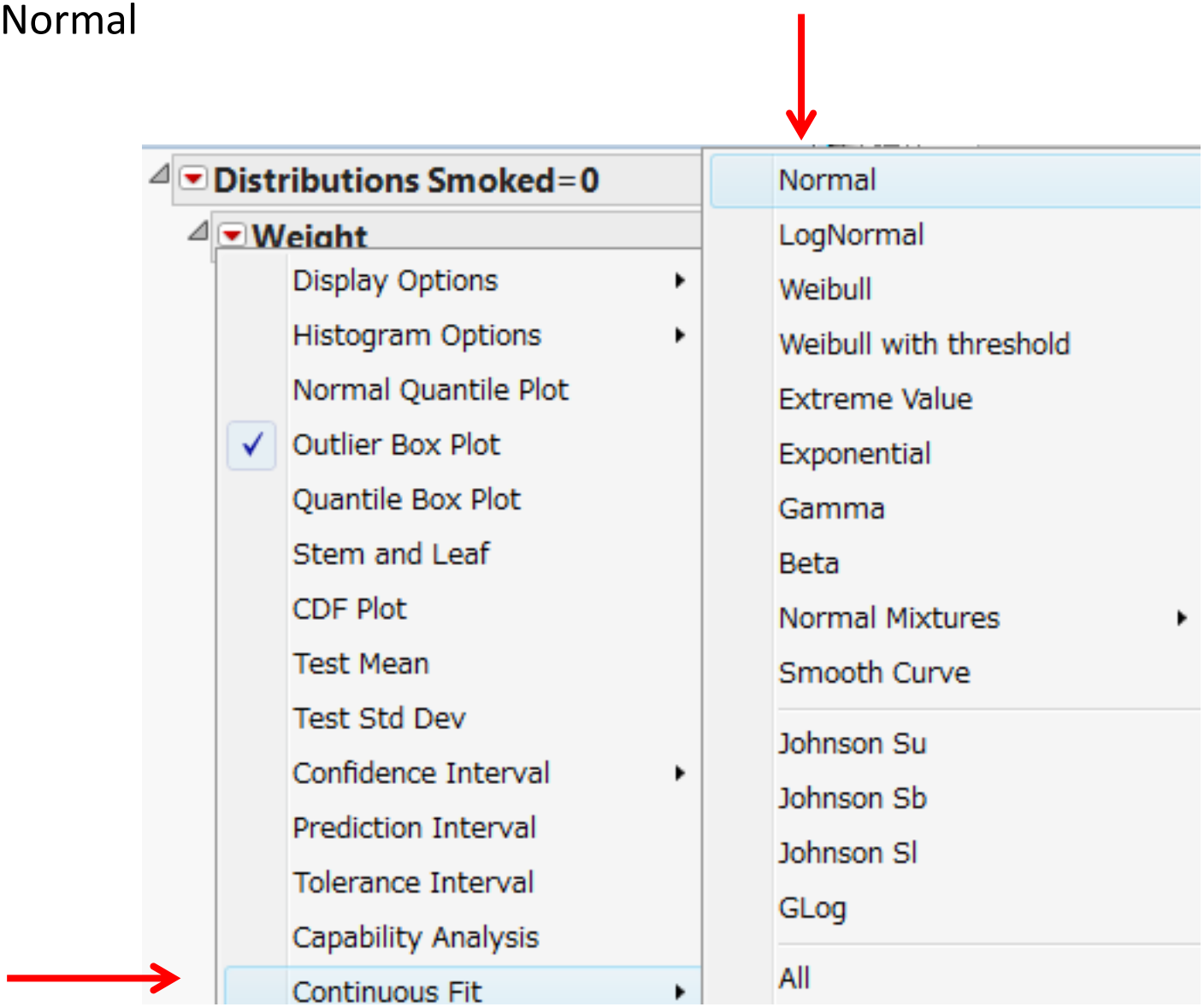


W ≈ 0.90; p = 0.24 -> We cannot reject that the data is normally distributed.

W is similar to a correlation coefficient with values close to 1 indicating high correspondence of the distribution with a normal distribution.

Test for normality: Shapiro-Wilk test

In JMP: Analyze->Distribution
Continuous Fit->Normal



Test for normality: Shapiro-Wilk test

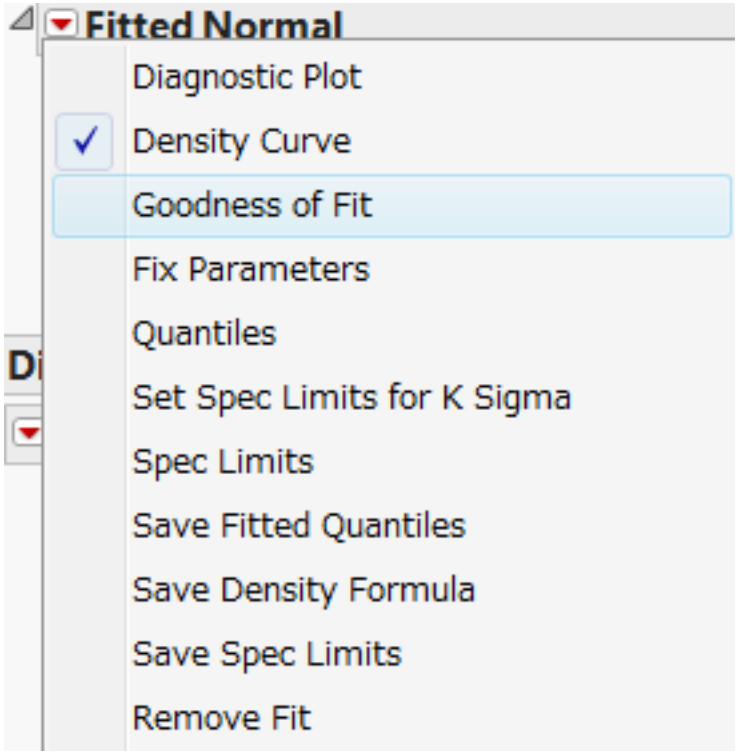
Fitted Normal->Goodness of Fit

Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	3223	2798.6078	3647.3922
Dispersion	σ	593.25936	408.06467	1083.0603

-2log(Likelihood) = 155.091404073508



Goodness-of-Fit Test

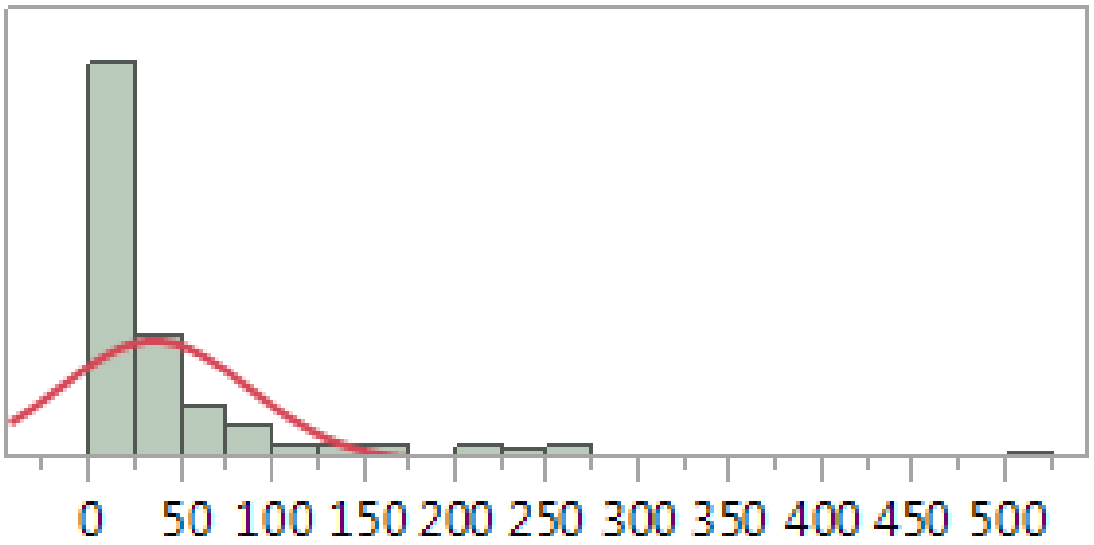
Shapiro-Wilk W Test

W	Prob<W
0.904026	0.2424

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

Test for normality: Shapiro-Wilk test

How about Titanic's fare prizes: normally distributed?



Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.527825	<.0001*

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

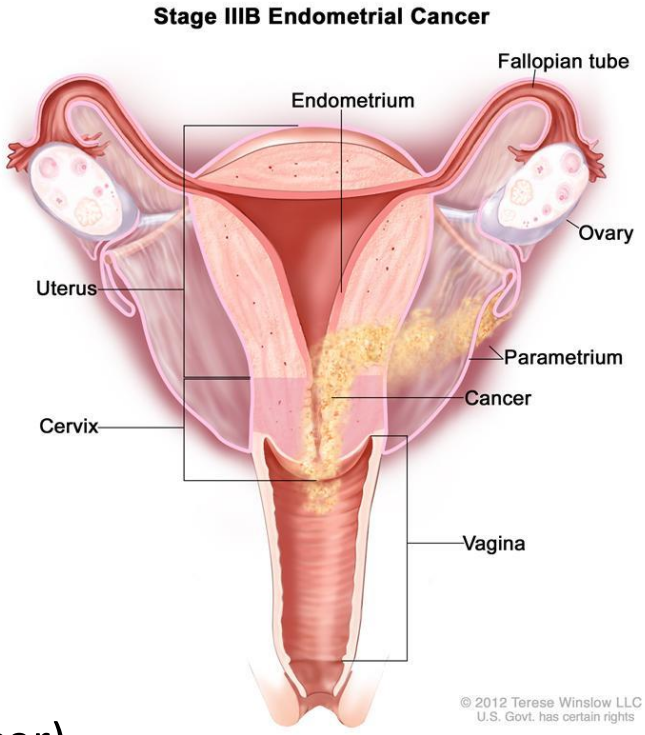
p<0.05 !

This distribution is skewed and we can reject the hypothesis that it is normal.

Test for normality: Shapiro-Wilk test

Clear deviation from normal distribution (red line)!

Question



You do a pap smear test to detect cervical cancer: Positive!

What is the probability that you really have cervical cancer?

Sensitivity:
(probability that the test is positive when having cancer)
$P(T+|C+) = 0.84$

Specificity:
(probability that the test is negative when having no cancer)
$P(T-|C-) = 0.90$

Prevalence: $P(C+) = 0.00008$

C+: having cervical cancer
C-: not having cervical cancer
T+: test indicates cervical cancer
T-: test does not indicate cervical cancer

Bayes' theorem

We need Bayes' theorem to solve this:

$$P(C+|T+) = \frac{P(C+ \cap T+)}{P(T+)} = \frac{P(T+|C+) \cdot P(C+)}{P(T+|C+) \cdot P(C+) + P(T+|C-) \cdot P(C-)}$$

$$= \frac{0.84 \cdot 0.00008}{0.84 \cdot 0.00008 + 0.1 \cdot 0.99992} \approx 0.0672\%$$

Bayes' theorem

Posterior probability

Likelihood

Prior probability

$$P(C+|T+) = \frac{P(T+|C+) \cdot P(C+)}{P(T+)}$$

Probability of Evidence

Before the test, you might belief that your chance for cervical cancer is 0.008% (prior probability). After you got evidence from the test, you update your belief to 0.0672% (posterior probability).

Summary

Categorical Variable x Categorical Variable

Data visualization and summary:

- 2x2 contingency table
- Odds-ratio, Risk-ratio

Statistical tests:

- χ^2-test
- Fisher's exact test (small samples)
- z-test for proportions

Summary

Categorical Variable x Metric Variable

Data visualization and summary:

- Box plots (quantiles)
- Bar graph
- Mean, Variance

Statistical tests:

- t-test
- Mann-Whitney U test (non-parametric)
- Analysis of Variance: for more than 2 groups

Summary

Metric Variable x Metric Variable

Data visualization and summary:

- Scatter plot
- Pearson's correlation coefficient
- Spearman's rank correlation (non-parametric)
- linear regression
- multiple linear regression

Statistical tests:

- testing significance of r (against t-distribution)
- testing significance of regression parameters (t)
- testing whole regression model (F)