Statistics for animal experiments - experimental design

Geir Storvik

9 January 2012

Experimental design

- Aim: Collect data for testing scientific hypothesis
- Focus: How to collect data
- Huge field, large number of textbooks on this
  - Some parts: Repetition from introductory courses (STK1000)
  - Other parts: Require further study
  - Main message: Using time on designing experiment properly can save you a lot of trouble and time

Experimental design

- Aim: Collect data for testing scientific hypothesis
- Focus: How to collect data
- Huge field, large number of textbooks on this
  - Some parts: Repetition from introductory courses (STK1000)
  - Other parts: Require further study
  - Main message: Using time on designing experiment properly can save you a lot of trouble and time
- Simple situation: Comparing treatment against control
  - Specification of sample size
  - Selection of animals for experiment
  - Allocation of animals to different groups
- Require knowledge of statistical method for analyzing data

Outline

Use of data for answering biological questions

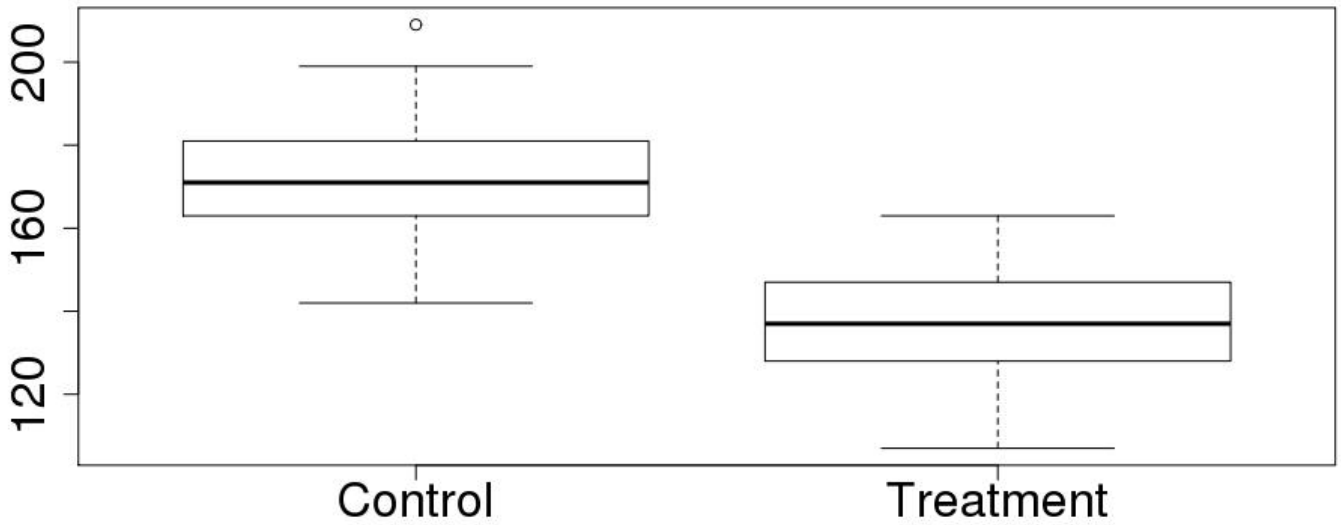
Example

- Are growth of rats reduced after a certain treatment?
- Empirical approach: Measure effects through experiments or field observations
- Data: Weight gain from first to 5th week for control/treatment group

Control	169	209	181	142	188	172	164	179	199	162	192	$\cdots$
Treatment	139	163	149	138	114	107	128	132	130	134	128	$\cdots$

(Gelfand et al. [1990])

Variability in observations



- Apparently differences
- Significant?
- Problem: Variability in data
- Need statistical methods to give a formal answer

Statistical hypothesis testing

1. Formulate biological question
2. Translate to a statistical hypothesis
   (Typically through specifying a mathematical model)
3. Select method for analysis of data
4. Design your experiment
5. Collect data
6. Analyze data
7. Write report/paper

Formulating hypotheses

Example: Are growth of rats reduced after a certain treatment?
Problem: Variability, not obvious that we get smaller growth for all rats after treatment.

Formulating hypotheses

Example: Are growth of rats reduced after a certain treatment?
Problem: Variability, not obvious that we get smaller growth for all rats after treatment.

Statistical approach

- Define a model including (unknown) parameters
- Formulate hypothesis through assumptions on parameters.
- Specify alternative hypothesis as what we want to show/prove
- Specify null hypothesis as the opposite of what we want to show/prove
- Analysis: Is there significant support in the data for rejecting null hypothesis

Formulating hypotheses

Example: Are growth of rats reduced after a certain treatment?
Problem: Variability, not obvious that we get smaller growth for all rats after treatment.

Statistical approach

- Define a model including (unknown) parameters
- Formulate hypothesis through assumptions on parameters.
- Specify alternative hypothesis as what we want to show/prove
- Specify null hypothesis as the opposite of what we want to show/prove
- Analysis: Is there significant support in the data for rejecting null hypothesis

Note: There might be that the collected data indicate that another method should be used. Some method is however needed in order to design your experiment.

Example: Growth of rats

Are growth of rats reduced after a certain treatment?

Example: Growth of rats

Are growth of rats reduced after a certain treatment?
Model

$$Y = \begin{cases} \mu_C + \varepsilon & \text{for control group} \\ \mu_T + \varepsilon & \text{for treatment group} \end{cases}$$

μ: Expected growth (population average), ε individual variability

Example: Growth of rats

Are growth of rats reduced after a certain treatment?
Model

$$Y = \begin{cases} \mu_C + \varepsilon & \text{for control group} \\ \mu_T + \varepsilon & \text{for treatment group} \end{cases}$$

μ: Expected growth (population average), ε individual variability
Define reduction in growth as $\mu_C > \mu_T$ or $\mu_C - \mu_T > 0$

$$H_0 : \mu_C - \mu_T = 0 \quad H_A : \mu_C - \mu_T > 0$$

T-test

Hypotheses

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_A : \mu_1 - \mu_2 > 0 \quad (\text{or } \mu_1 - \mu_2 \neq 0)$$

$\overline{Y}_i = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{i,j}$ estimates of $\mu_i, i = 1, 2$
Reasonable to look at $\overline{Y}_1 - \overline{Y}_2$

T-test

Hypotheses

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_A : \mu_1 - \mu_2 > 0 \quad (\text{or } \mu_1 - \mu_2 \neq 0)$$

$\overline{Y}_i = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{i,j}$ estimates of $\mu_i, i = 1, 2$
Reasonable to look at $\overline{Y}_1 - \overline{Y}_2$
Test statistic:

$$T = \frac{\overline{Y}_1 - \overline{Y}_2}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}}$$

Reject H_0 if T is large

T-test (cont)

Test statistic:

$$T = \frac{\overline{Y}_1 - \overline{Y}_2}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}}$$

$\hat{\sigma}\sqrt{1/n_1 + 1/n_2}$ is a scaling factor, taking into account the variability in $\overline{Y}_1 - \overline{Y}_2$.

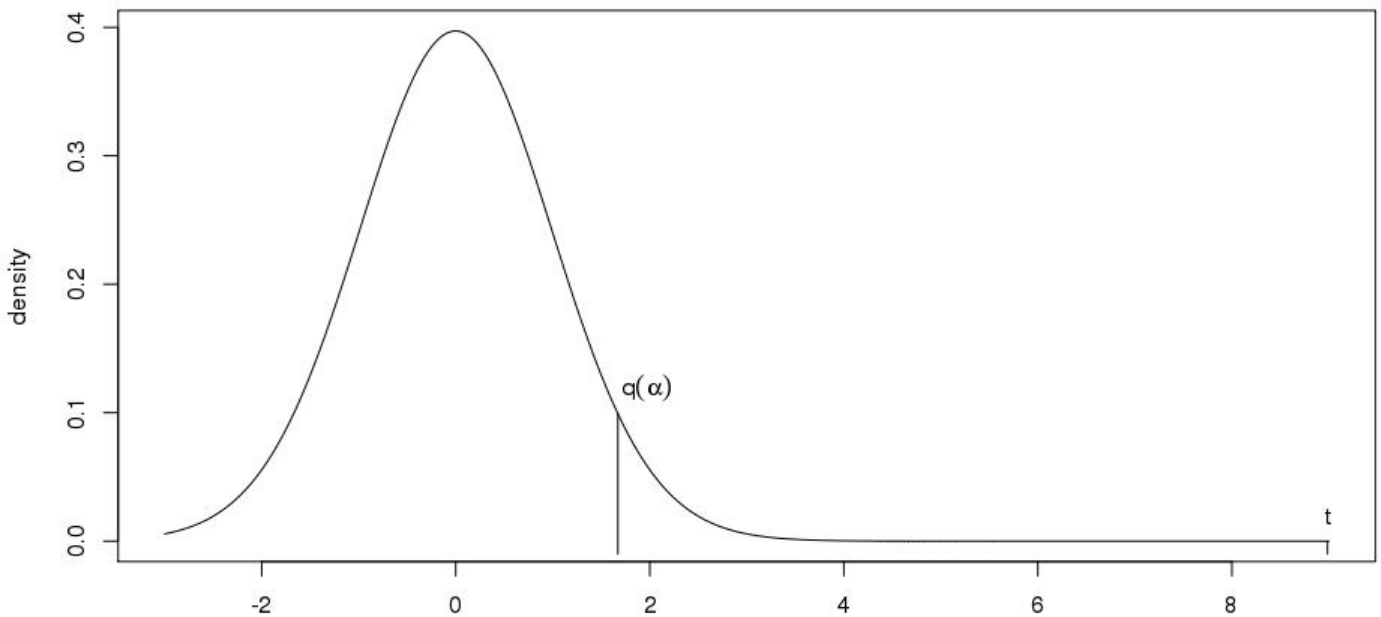
- $\hat{\sigma}$: Variability in individual observations
  Typically out of our control (but see later)
- $\sqrt{1/n_1 + 1/n_2}$: Reduction in variability due to many observations
  Specified through the experimental design!

T-test - How large is large?

- Under H_0, T is t-distribution with $n_1 + n_2 - 2$ degrees of freedom.
- Reject H_0 if T is larger than the upper α quantile

T-test - How large is large?

- Under H_0, T is t-distribution with $n_1 + n_2 - 2$ degrees of freedom.
- Reject H_0 if T is larger than the upper α quantile
- Example: $\overline{Y}_1 - \overline{Y}_2 = 35.97$, $T = 7.08$, $df = 58$, P-value: 7.0×10^{-13}



T-test - assumptions

T-test based on a number of assumptions on data $Y_{i,j}, i = 1, 2, j = 1, ..., n_i$

1. All observations are independent
   Can be made true by proper experimental design

T-test - assumptions

T-test based on a number of assumptions on data
$Y_{i,j}, i = 1, 2, j = 1, ..., n_i$

1. All observations are independent
2. All observations within a group are identically distributed
   Can be made by true proper experimental design

T-test - assumptions

T-test based on a number of assumptions on data $Y_{i,j}, i = 1, 2, j = 1, ..., n_i$

1. All observations are independent
2. All observations within a group are identically distributed
3. The variance σ^2 is the same in both groups
   - Alternative test when variances are different

T-test - assumptions

T-test based on a number of assumptions on data $Y_{i,j}, i = 1, 2, j = 1, ..., n_i$

1. All observations are independent
2. All observations within a group are identically distributed
3. The variance σ^2 is the same in both groups
4. $Y_{i,j}$ follows the normal distribution
   - (Actually enough that $\overline{Y}_i$ is normally distributed)
   - The central limit theorem:
     "The mean of independent observations will become normally distributed when n_i becomes large."
   - Beware of Outliers

Randomization

Concept:

- Draw randomly $n = n_1 + n_2$ individuals from the population of interest
- Draw randomly n_1 individuals from the n selected for the control group, the rest for the treatment group

Randomization

Concept:

- Draw randomly $n = n_1 + n_2$ individuals from the population of interest
- Draw randomly n_1 individuals from the n selected for the control group, the rest for the treatment group

Important aspects

- Make observations identically distributed and approximately independent

Randomization

Concept:

- Draw randomly $n = n_1 + n_2$ individuals from the population of interest
- Draw randomly n_1 individuals from the n selected for the control group, the rest for the treatment group

Important aspects

- Make observations identically distributed and approximately independent
- Avoid systematic differences between control group and treatment group
  (Accounting for variability in individuals related to factor that are not under our control)

Randomization

Concept:

- Draw randomly $n = n_1 + n_2$ individuals from the population of interest
- Draw randomly n_1 individuals from the n selected for the control group, the rest for the treatment group

Important aspects

- Make observations identically distributed and approximately independent
- Avoid systematic differences between control group and treatment group
  (Accounting for variability in individuals related to factor that are not under our control)
- Make individuals representative for the population of interest

Importance of randomization

Rats data

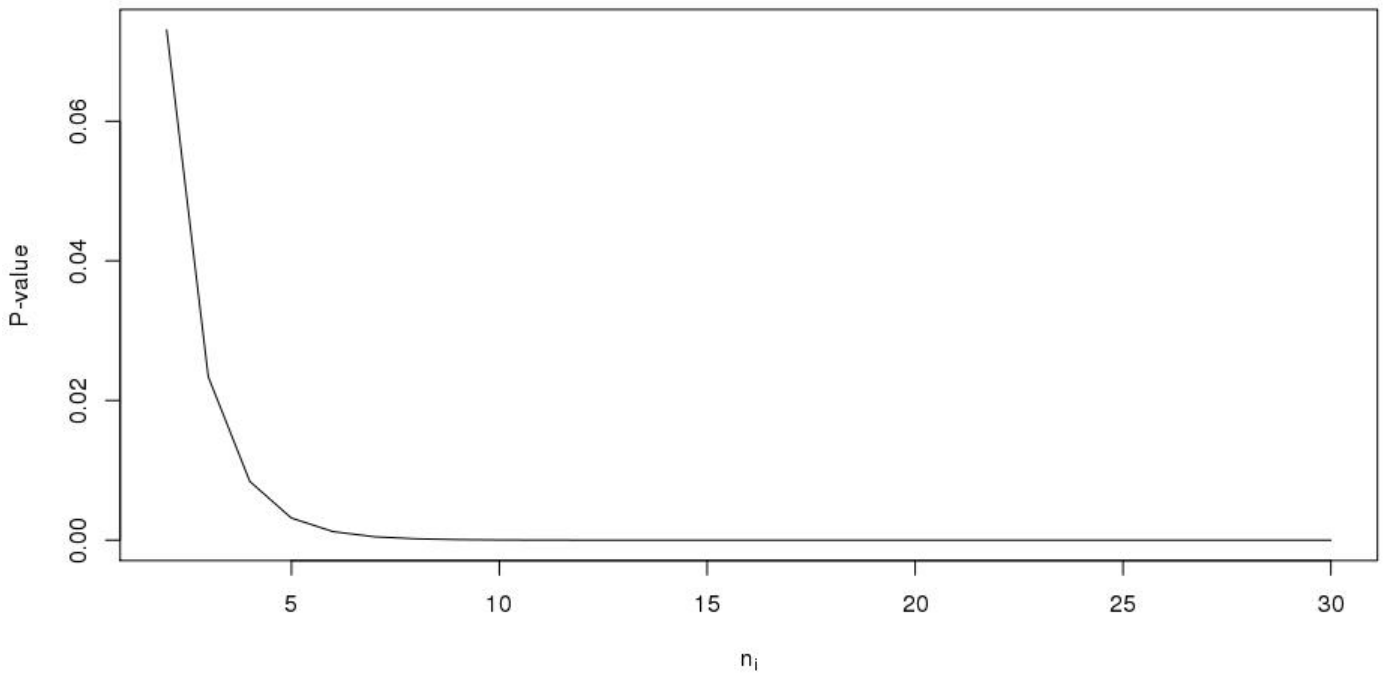
- Weight gain defined as increase from day 8 to day 36
- Also available weight at day 8
- Assume treatment given just after day 8
- Are there differences in control group and treatment group before treatment?

Control	151	145	147	155	135	159	141	159	$\cdots$
Treatment	114	140	133	132	119	155	117	129	$\cdots$

- T-test: $T = 2.43$, P-value=0.0182
- Give some indication of differences.
- Note: The actual treatment startet at day 1

How many samples do we need?

Example: $\overline{Y}_1 - \overline{Y}_2 = 35.97$, P-value: 7.01×10^{-13}.
Could we suffice with less observations?
Assume $\overline{Y}_1 - \overline{Y}_2 = 35.97$, but consider different values of n_i



- P-value increase with decreasing n_i
- Here: Very small P-values for much smaller n_i.

Sample size - how large should a study be?

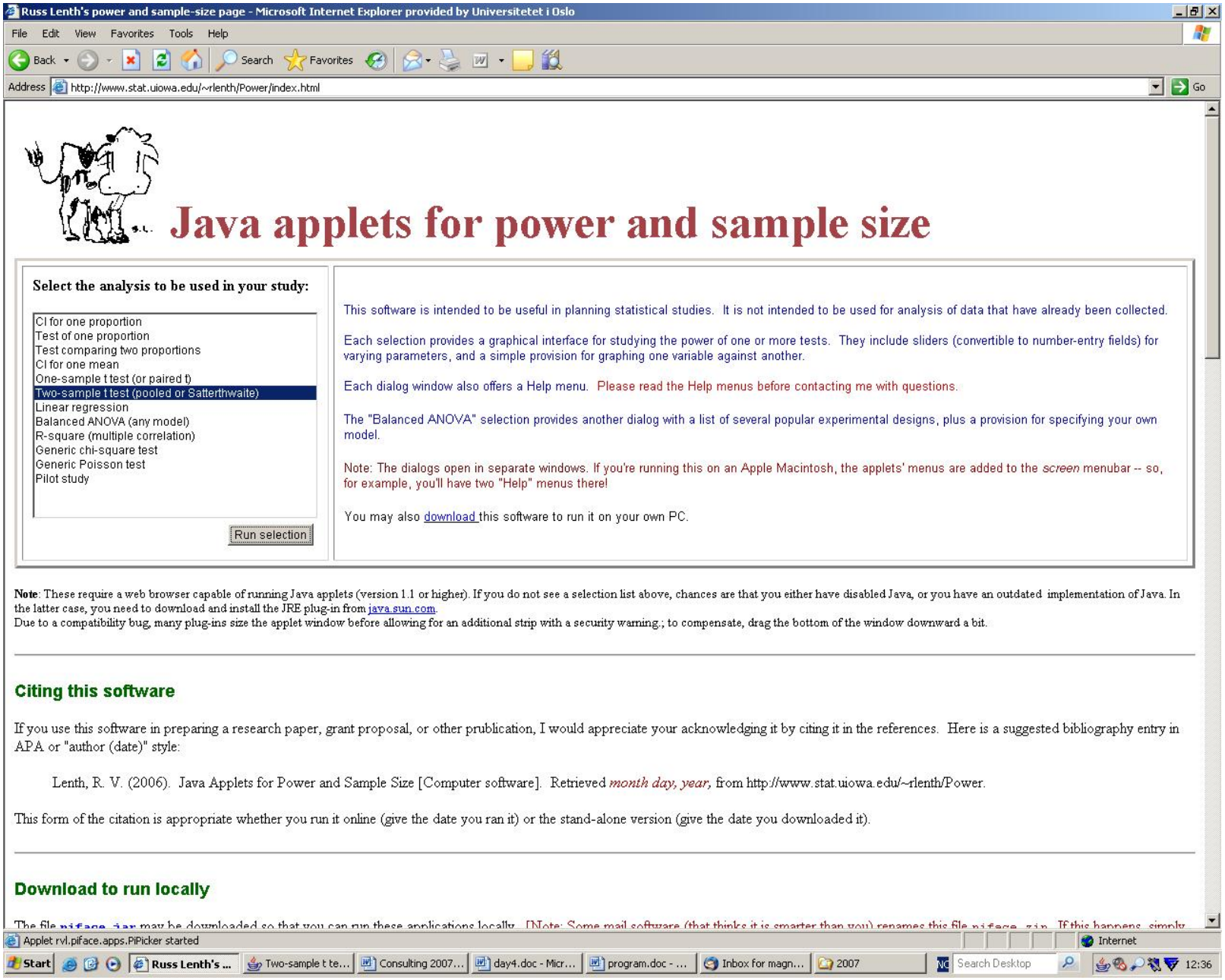
Possible decisions on H_0

	H_0 True	H_0 False
Not rejected	Right decision	Type II error
Rejected	Type I error	Right decision

Statistical tests:

- Make a limit, confidence level α, of probability for Type I error
- Want probability of Type II error small
  Want probability of rejecting H_0 when H_0 False high
  (high power, β)
- Specify sample size so large that the power is high enough

There exists software for performing this kind of calculations. There are also a number of "calculators" to be found on the Internet. We will be using one of them:
`http://www.stat.uiowa.edu/~rlenth/Power/index.html`

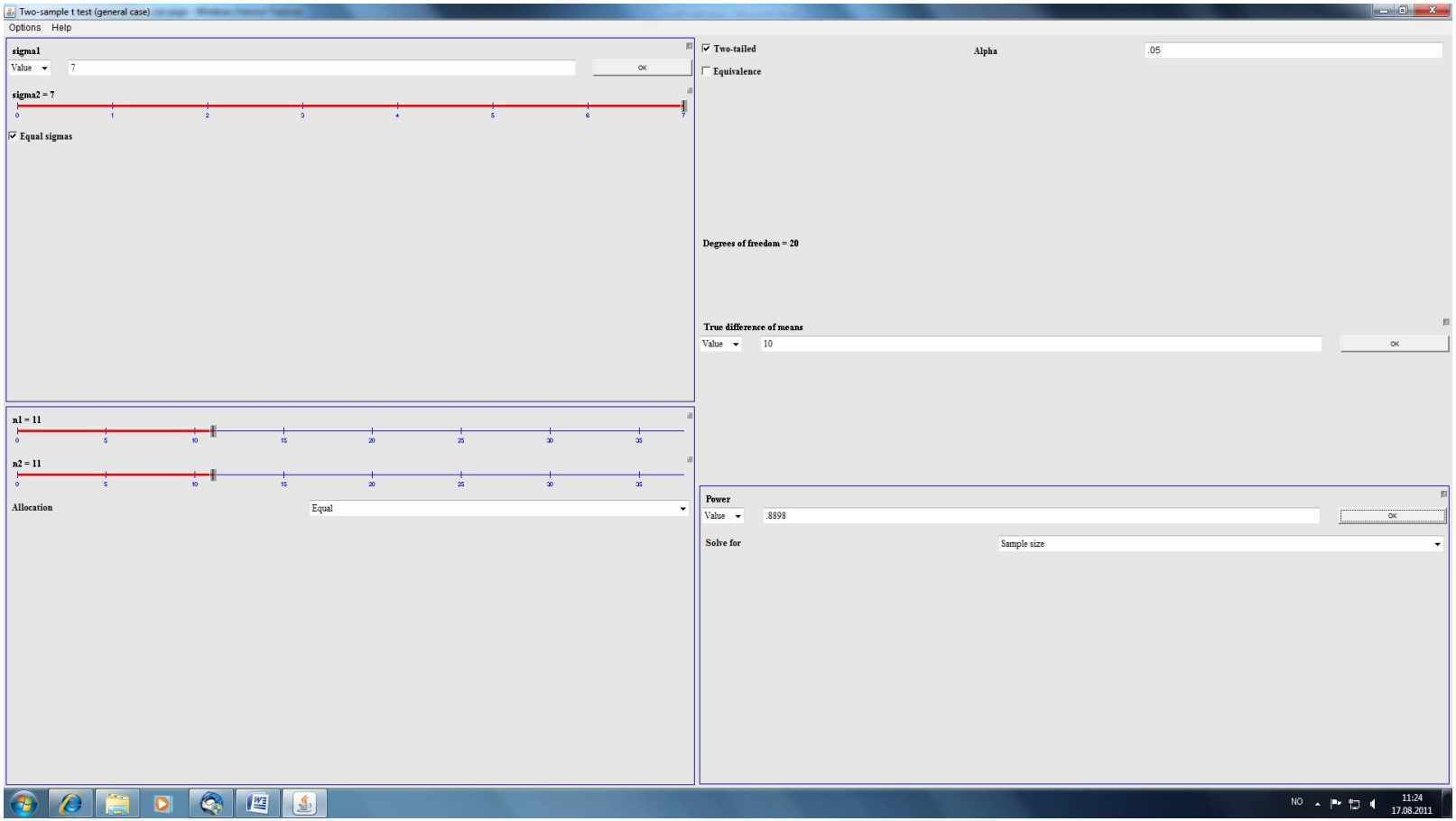

Russ Lenth's power and sample-size page - Microsoft Internet Explorer provided by Universitetet i Oslo
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.stat.uiowa.edu/~rlenth/Power/index.html
Go
Java applets for power and sample size
Select the analysis to be used in your study:
CI for one proportion
Test of one proportion
Test comparing two proportions
CI for one mean
One-sample t test (or paired t)
Two-sample t test (pooled or Satterthwaite)
Linear regression
Balanced ANOVA (any model)
R-square (multiple correlation)
Generic chi-square test
Generic Poisson test
Pilot study
Run selection
This software is intended to be useful in planning statistical studies. It is not intended to be used for analysis of data that have already been collected.
Each selection provides a graphical interface for studying the power of one or more tests. They include sliders (convertible to number-entry fields) for varying parameters, and a simple provision for graphing one variable against another.
Each dialog window also offers a Help menu. Please read the Help menus before contacting me with questions.
The "Balanced ANOVA" selection provides another dialog with a list of several popular experimental designs, plus a provision for specifying your own model.
Note: The dialogs open in separate windows. If you're running this on an Apple Macintosh, the applets' menus are added to the screen menubar -- so, for example, you'll have two "Help" menus there!
You may also download this software to run it on your own PC.
Note: These require a web browser capable of running Java applets (version 1.1 or higher). If you do not see a selection list above, chances are that you either have disabled Java, or you have an outdated implementation of Java. In the latter case, you need to download and install the JRE plug-in from java.sun.com.
Due to a compatibility bug, many plug-ins size the applet window before allowing for an additional strip with a security warning.; to compensate, drag the bottom of the window downward a bit.
Citing this software
If you use this software in preparing a research paper, grant proposal, or other prublication, I would appreciate your acknowledging it by citing it in the references. Here is a suggested bibliography entry in APA or "author (date)" style:
Lenth, R. V. (2006). Java Applets for Power and Sample Size [Computer software]. Retrieved month day, year, from http://www.stat.uiowa.edu/~rlenth/Power.
This form of the citation is appropriate whether you run it online (give the date you ran it) or the stand-alone version (give the date you downloaded it).
Download to run locally
Applet rvl.piface.apps.PiPicker started
Internet
Start
Russ Lenth's ...
Two-sample t te...
Consulting 2007...
day4.doc - Micr...
program.doc - ...
Inbox for magn...
2007
Search Desktop
12:36

T-test and sample size

Depend on

- Confidence level α (Typically 0.05 or 0.01)
- $\mu_1 - \mu_2$: Clinically / practically relevant effect / difference
- Variability in data (size of σ).
  - Guess from previous experiments
  - Perform a preliminary experiment
- Power, β, at relevant effect. Should be 80-95%!

Assume $\alpha = 0.05$, $\sigma = 15$, $\beta = 0.9$ and $\mu_1 - \mu_2 = 20$.
Gives $n_i = 10$


Two-sample t test (general case)
Options Help
sigma1
Value 7 OK
sigma2 = 7
Equal sigmas
Two-tailed
Equivalence
Alpha .05
Degrees of freedom = 20
True difference of means
Value 10 OK
n1 = 11
n2 = 11
Allocation Equal
Power
Value .8898 OK
Solve for Sample size
NO 11:24 17.08.2011

Ethics

- To run studies that are too small should be considered unethical.
  - Much resources are used without any reasonable chance of getting a good result.
- To run to studies that are too large
  - Can be to costly in time/money
  - Can be considered unethical as well

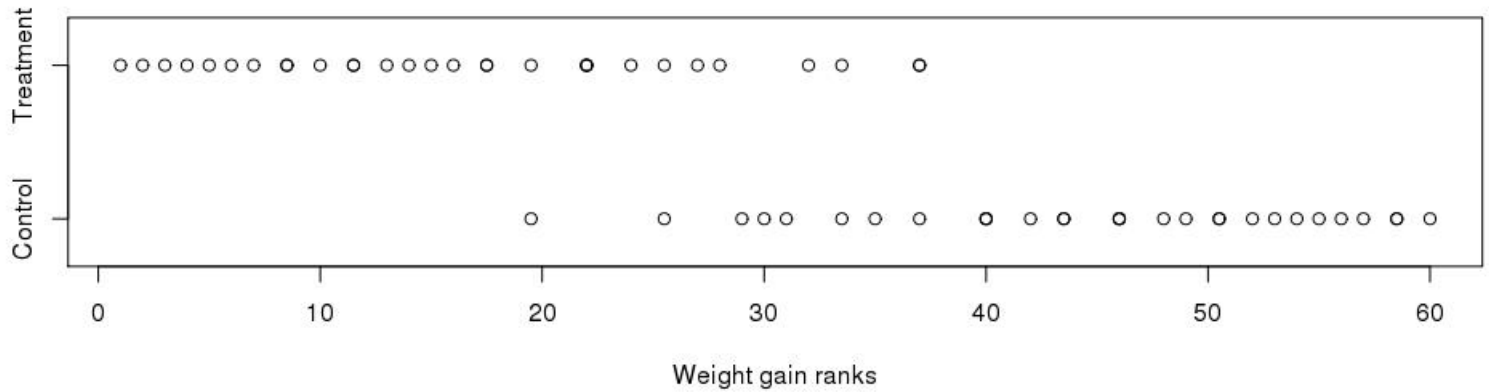
Mann-Whitney test

T-tests rely on normality
For small datasets

- Central limit theorem do not work properly
- Too few samples to check if normality is reasonable

Alternative to T-test: Mann-Whitney test
Main idea:
Rank the observations (independent of group). If the groups are similar, we will expect the observations to spread rather arbitrary. If we summarize the ranks for each group, we will then expect these sums to be similar. If the sum of ranks in the groups are very different, we will say that the groups differ.

Mann-Whitney on rats example



W = 864.5, p-value = 4.595e-10

No easy way to calculate sample size for such tests

Difference in power is surprisingly small [$<$ 5% compared to t-test, Quinn and Keough, 2002]

As a rule of thumb, for non-parametric tests, add 10% on sample size.

Mann-Whitney test versus T-test

Y_{1j} and Y_{2j} are distributed according to two different distributions

	Mann-Withney test	T-test
H_0	The two distributions are equal	$\mu_1 = \mu_2$

μ_i is the expectation in distribution i
Mann-Withney test

- Less assumptions on the shape of the distribution
- Strong assumptions on the similarity between the distributions
  In particular, the variance still needs to be the same

Quinn and Keough [2002]: Hard to recommend the rank-based tests except in situations where

- distributions are very weird and transformations do not help
- outlier are present

One-way anova

- Control/treatment: Factor with two levels
- What if several treatments are to be compared?
- Extension of t-test to one-way anova

$$Y_{i,j} = \mu_i + \varepsilon_{i,j}, i = 1, ..., I, j = 1, ..., J$$

- Can compute power/sample size also in this case
- More complicated, need
  - Variability within groups, $\sigma = SD(\varepsilon_{i,j})$
  - Power β
  - Variability between groups, $SD(\mu_i)$

    T-test: $\mu_2 - \mu_1$

Stratification

Randomization: Accounting for variability in individuals related to factor that are not under our control
Sometimes: Factors are under control
Rats example: Growth structure may depend on sex
Possible solutions

- Do separate analysis for each sex
  (Usually not possible when many levels on factor or many factors)
- Include sex as a factor in the model
- Stratification: Design your experiment such that the numbers/proportions from each sex are similar in the control group and the treatment group.
- Main effect: Reduce variability in individual observations
- Still important to do randomization within each strata!

Matched pairs

- Effect may vary with genetic variation
- Idea: Compare individuals with similar genetic properties
- Siblings/identical twins

$$Y_{i,j} = \mu_i + \eta_j + \varepsilon_{i,j}$$

η_j are variability in genetic properties (family factor)

$$\begin{aligned} D_j &= Y_{2,j} - Y_{1,j} = \mu_2 - \mu_1 + \varepsilon_{2,j} - \varepsilon_{1,j} \\ &= \tilde{\mu} + \tilde{\varepsilon}_j \\ H_0 &: \tilde{\mu} = 0 \quad \text{against} \quad H_A : \tilde{\mu} > 0 \end{aligned}$$

- Use randomization to allocate treatment group to the two siblings!

Factorial design

- Assume interest is in two types of treatments
  - Genetic manipulation against control
  - Two types of diets
- Possible approach
  1. Use one diet to test differences between genetic manipulation against control
  2. Use genetic manipulated individuals to test differences in diet types
- Better approach - Factorial design
  - Test both factors simultaneously
  - Gives more power
  - Allow for testing interactions

Other methods

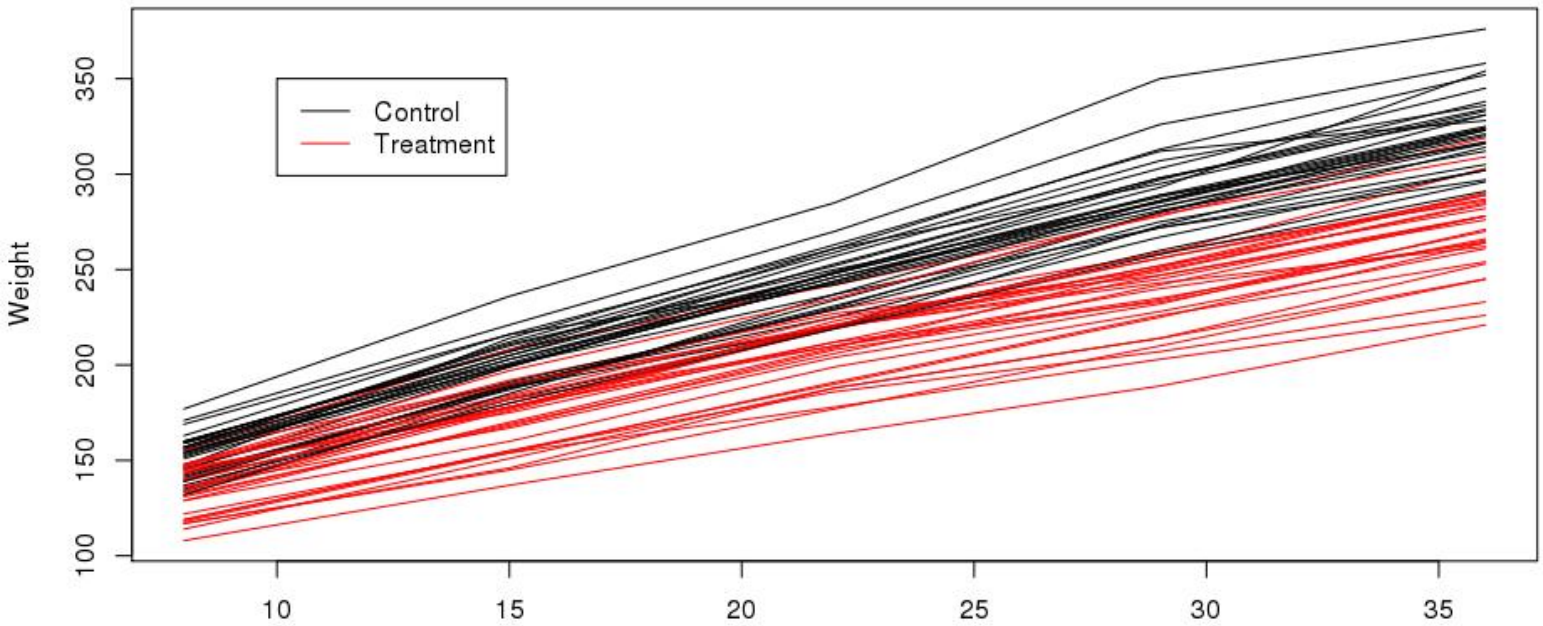
- T-test and one-way anova
- Non-parametric methods
- Anova with two or more factors - factorial designs
  - What to do when not all combinations are possible to measure
- Linear regression, response varying with numerical covariate
- Dependence between observations
- Other types of responses
  - Binary responses (Death/Survival) - Logistic regression
  - Count data (numbers in traps) - Poisson regression

Going further

- Possible courses
  - BIO2150 - Biostatistics and Study Design
  - BIO2150A - Biostatistics
  - STK4900/9900 - Statistical methods and applications
- Books
  - Quinn and Keough [2002] *Experimental design and data analysis for biologists*
  - Mead et al. [2003] *Statistical methods in agriculture and experimental biology*
  - Heath [1995] *An introduction to experimental design and statistics for biology*
  - Zuur et al. [2007] *Analysing ecological data*
  - Montgomery [2008] *Design and analysis of experiments*
  - Wu et al. [2000] *Experiments: planning, analysis, and parameter design optimization*

Full rats data

	ID	8 days	15 days	22 days	29 days	36 days	Gain
Control	1	151	199	246	283	320	169
	2	145	199	249	293	354	209
	3	147	214	263	312	328	181
	⋮						
Treatment	31	114	151	188	214	253	145
	32	140	189	229	258	303	145
	33	133	176	220	252	282	135
	⋮						



Linear regression model

Possible model for one rat:

$$Y_j = \alpha + \beta x_j + \varepsilon_j$$

where ε_j are independent and $N(0, \sigma^2)$
Assume interest in β
How many samples needed to detect a β significant from zero?
As for T-test, need to specify

- Confidence level α (Typically 0.05 or 0.01)
- Detectable β: The clinically meaningful value of the regression coefficient that you want to be able to detect
- Variability in data (size of σ)
- Power at relevant effect (80-95%)
- In addition: Variability in x_j through standard deviation of x's

Applet

Using the same applet as before with
$\alpha = 0.05, \beta = 1, \sigma = 16, sd(X) = 10$ and Power 0.9 gives $n = 28$
Question: Could we use 5 observations from 6 rats (total of $n = 30$ observations?)

Rats - using data from several individuals

Possible model

$$Y_{i,j} = \begin{cases} \alpha_C + \beta_C x_j + \varepsilon_{i,j} & \text{for control group} \\ \alpha_T + \beta_T x_j + \varepsilon_{i,j} & \text{for treatment group} \end{cases}$$

where $\varepsilon_{i,j}$ are independent and $N(0, \sigma^2)$
Quantity of interest: $\beta_C - \beta_T$

Rats - using data from several individuals

Possible model

$$Y_{i,j} = \begin{cases} \alpha_C + \beta_C x_j + \varepsilon_{i,j} & \text{for control group} \\ \alpha_T + \beta_T x_j + \varepsilon_{i,j} & \text{for treatment group} \end{cases}$$

where $\varepsilon_{i,j}$ are independent and $N(0, \sigma^2)$
Quantity of interest: $\beta_C - \beta_T$
Rats data using 3 individuals from each group

- $\hat{\beta}_C - \hat{\beta}_T = 1.3343$
- Testing $H_0 = \beta_C - \beta_T = 0$ gives P-value 0.0273
- Can we trust these results?
- Problem: Observations from same individual gives dependence
- Many different approaches for handling this
  - Correlation structure in $\varepsilon_{i,j}$'s
  - Including random effects accounting for individual variation
    Gave P-value 0.0095!

A.E. Gelfand, S.E. Hills, A. Racine-Poon, and A.F.M. Smith. Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, pages 972–985, 1990.

D. Heath. *An introduction to experimental design and statistics for biology*. CRC Press, 1995.

R. Mead, R.N. Curnow, and A.M. Hasted. *Statistical methods in agriculture and experimental biology*. CRC Press, 2003.

D.C. Montgomery. *Design and analysis of experiments*. John Wiley & Sons Inc, 2008.

G.P. Quinn and M.J. Keough. *Experimental design and data analysis for biologists*. Cambridge Univ Pr, 2002.

C.F.J. Wu, M. Hamada, and C.F. Wu. *Experiments: planning, analysis, and parameter design optimization*. Wiley New York, 2000.

A.F. Zuur, E.N. Ieno, and G.M. Smith. *Analysing ecological data*. Springer Verlag, 2007.