Advanced Statistics

Janette Walde

janette.walde@uibk.ac.at

Department of Statistics
University of Innsbruck

Contents

"*We are pattern-seeking story-telling animals.*" (Edward Leamer)

"Statistics does not hand truth to the user on a silver platter. However, statistics confines arbitrariness and provides comprehensible conclusions."

"*Es gibt keine Tatsachen, es gibt nur Interpretationen.*" (Friedrich Nietzsche)

Preliminary comments

1. You will learn to apply statistical tools correctly, interpret the findings appropriately and get an idea about the possibilities of analyzing research questions employing statistics.
2. It is not possible and not worthwhile to learn all statistical methods in such a course. However, this course is successful if it enables you to improve your knowledge in statistical methods on your own. Therefore this course gives you profound knowledge about some statistical analyzing tools and shows you the correct application of them.

Preliminary comments

3. Although knowing the most sophisticated analyzing instruments one may be confronted with limits in getting results or finding appropriate interpretations or applying tools in the given framework. This has to be accepted ("*If we torture the data long enough, they will confess.*").
4. Be aware: Never confuse statistical significance with biological significance.

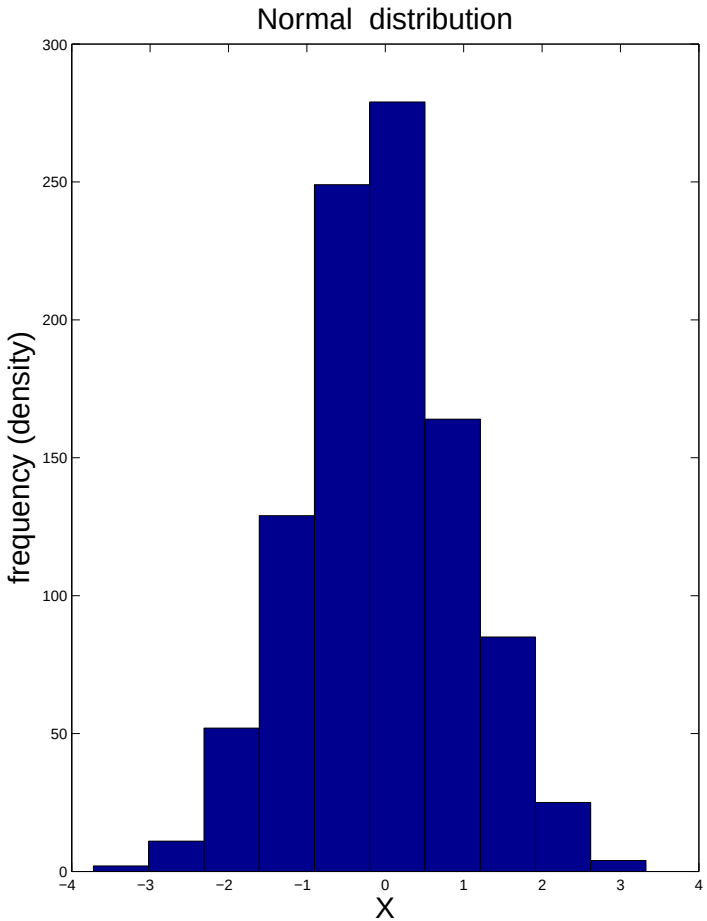
Scales of measurement

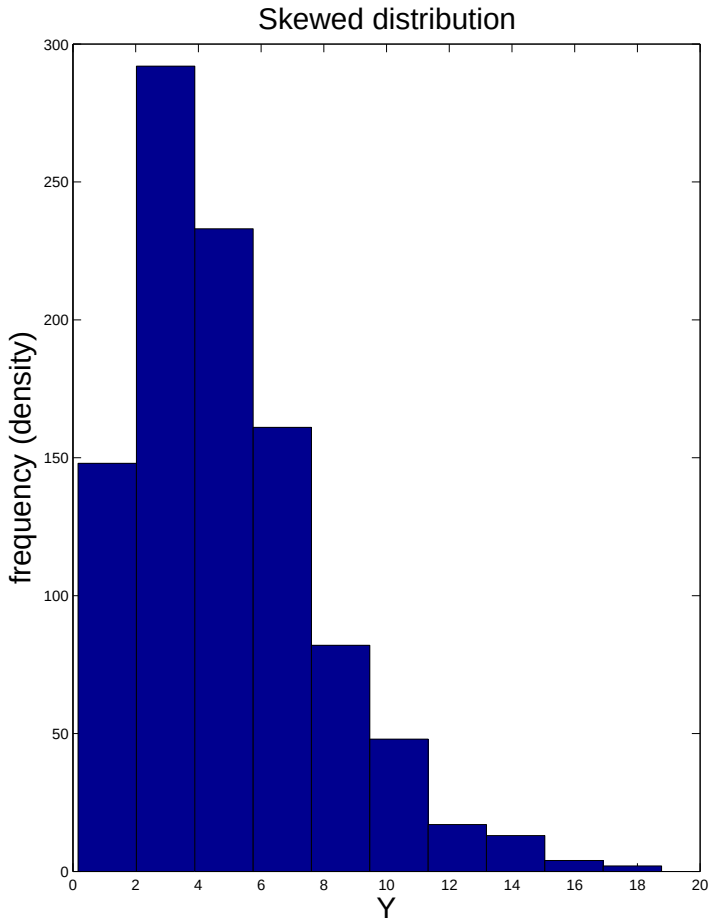
1. Nominal Scale. Nominal data are attributes like sex or species, and represent measurement at its weakest level. We can determine if one object is different from another, and the only formal property of nominal scale data is *equivalence*.
2. Ranking Scale. Some biological variables cannot be measured on a numerical scale, but individuals can be ranked in relation to one another. Two formal properties occur in ranking data: *equivalence* and *greater than*.

Scales of measurement

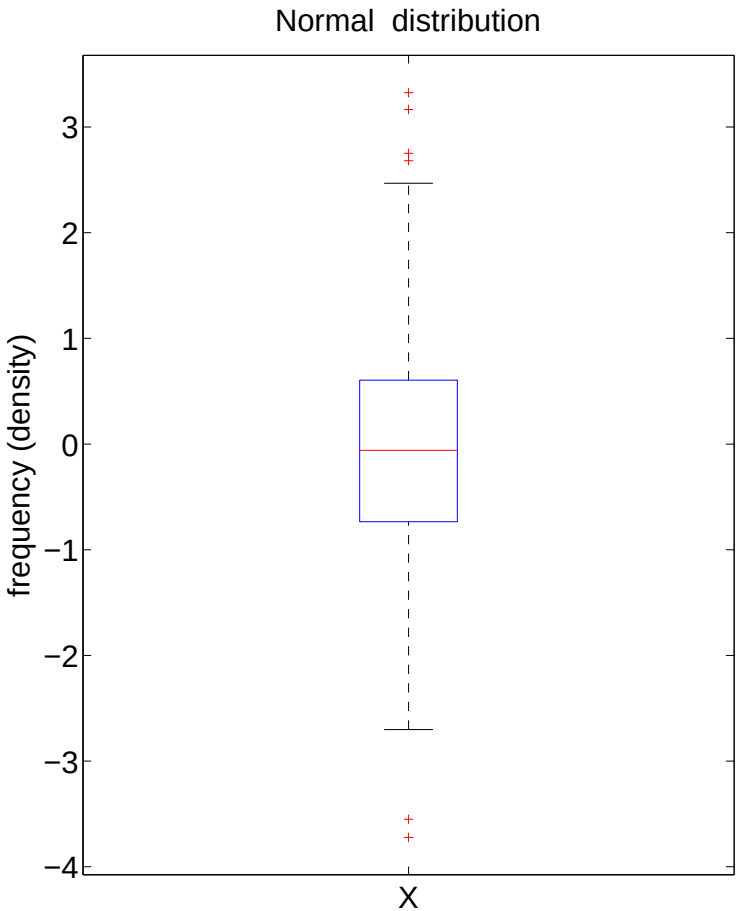
3. Interval and Ratio Scales. Interval and ratio scales have all the characteristics of the ranking scale, but we know the distances between the classes. If we have a true zero point, we have a ratio scale of measurement.

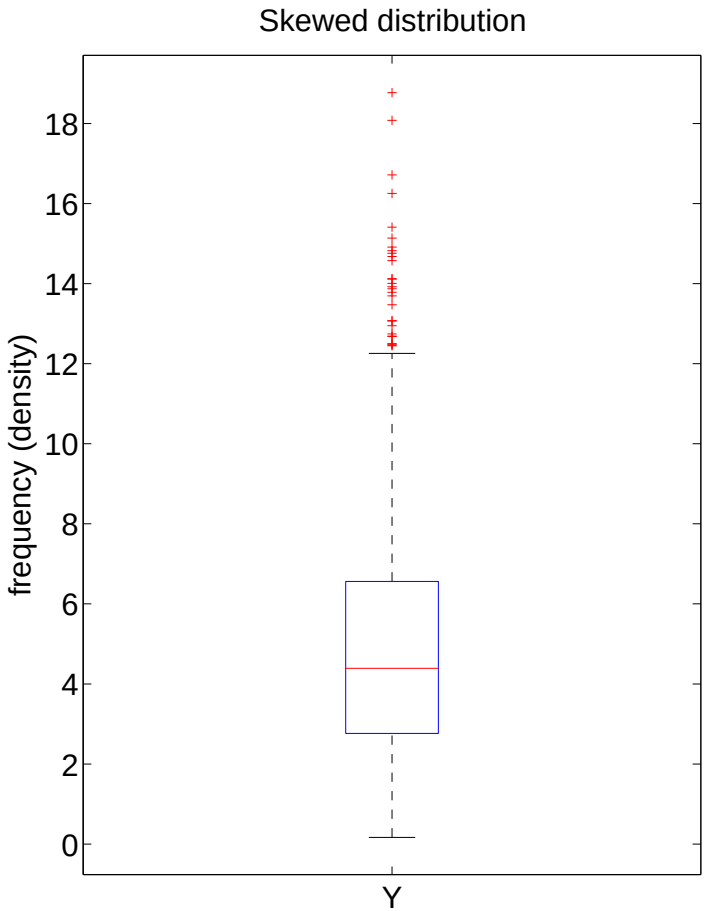
Histogram


Normal distribution
frequency (density)
X
300
250
200
150
100
50
0
−4
−3
−2
−1
0
1
2
3
4


Skewed distribution
frequency (density)
Y
300
250
200
150
100
50
0
0
2
4
6
8
10
12
14
16
18
20

Box Plot

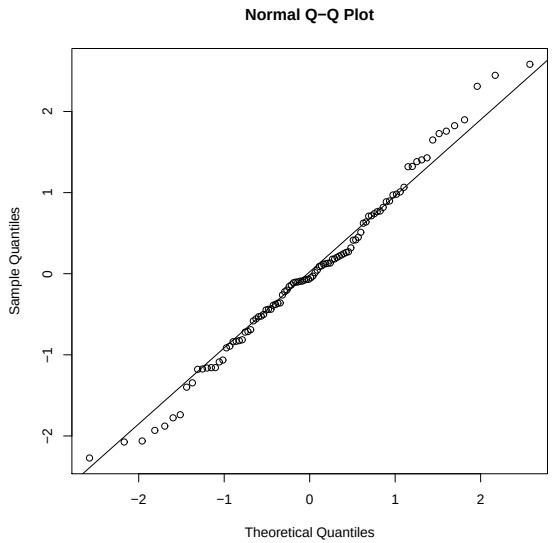

Normal distribution
frequency (density)
3
2
1
0
−1
−2
−3
−4
X

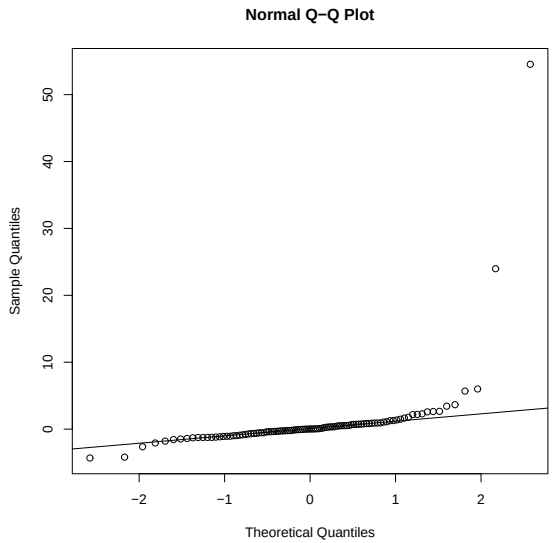

Skewed distribution
frequency (density)
18
16
14
12
10
8
6
4
2
0
Y

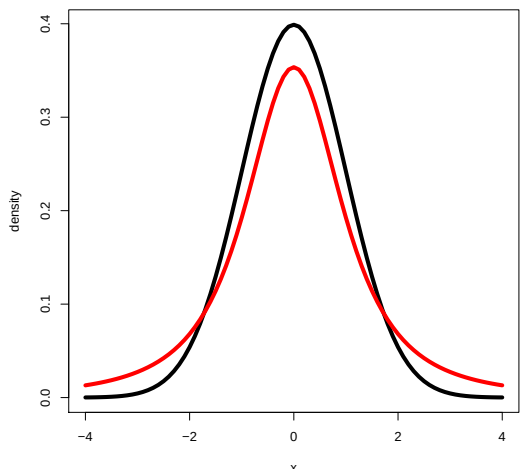
Q-Q Plot

- Many statistical methods make some assumptions about the distribution of the data (e.g. normality).
- The quantile-quantile plot provides a way to visually investigate such an assumption.
- The QQ-plot shows the theoretical quantiles versus the empirical quantiles. If the distribution assumed (theoretical one) is indeed the correct one, we should observe a straight line.

Q-Q Plot


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
0
10
20
30
40
50


density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4

Summary Statistic

- Mean, median
- Percentiles, inter quartile range
- Minimum, maximum, range
- Standard deviation, variance
- Coefficient of variation
- Median absolute deviation, mean absolute deviation

Fundamental concepts

Populations must be defined at the start of any study and this definition should include the spatial and temporal limits to the inference. The formal statistical inference is restricted to these limits.

Possibility of drawing samples randomly.

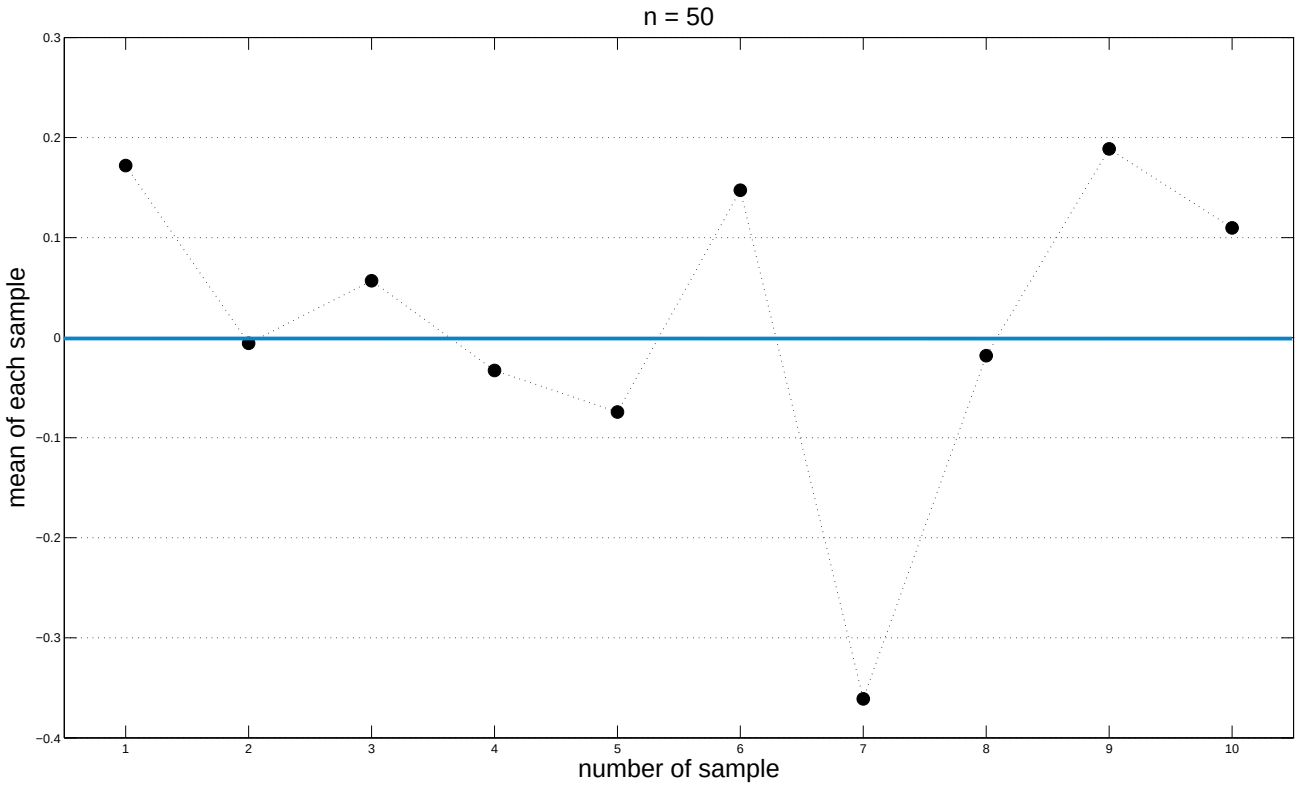
Population parameters are considered to be fixed but unknown values (in contrast to the Bayesian approach).

Characteristics of an estimator

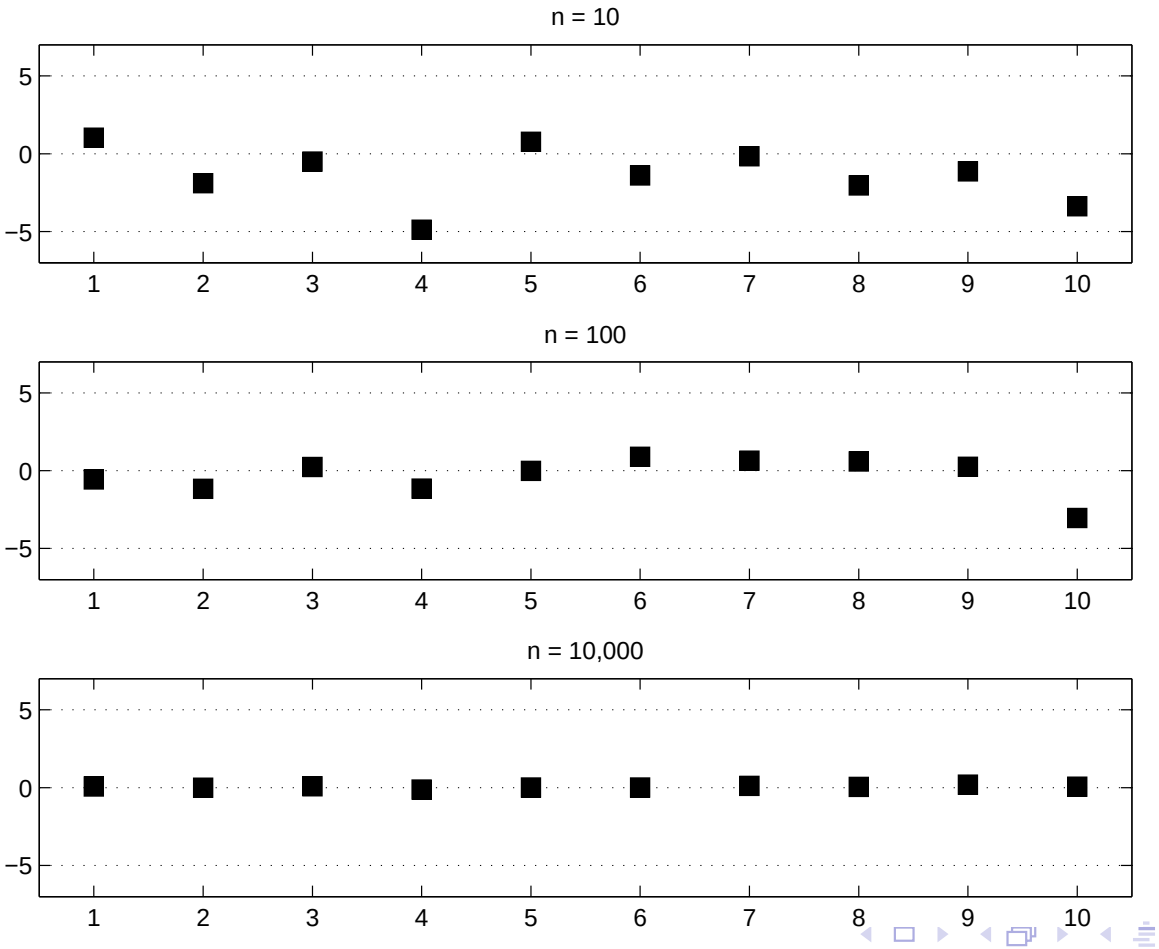
A good estimator of a population parameter should have the following characteristics:

- The estimator should be unbiased, meaning that the expected value of the sample statistic (the mean of its probability distribution) should equal the parameter.
- It should be consistent so as the sample size increases then the estimator will get closer to the population parameter.
- It should be efficient, meaning it has the lowest variance among all competing estimators.

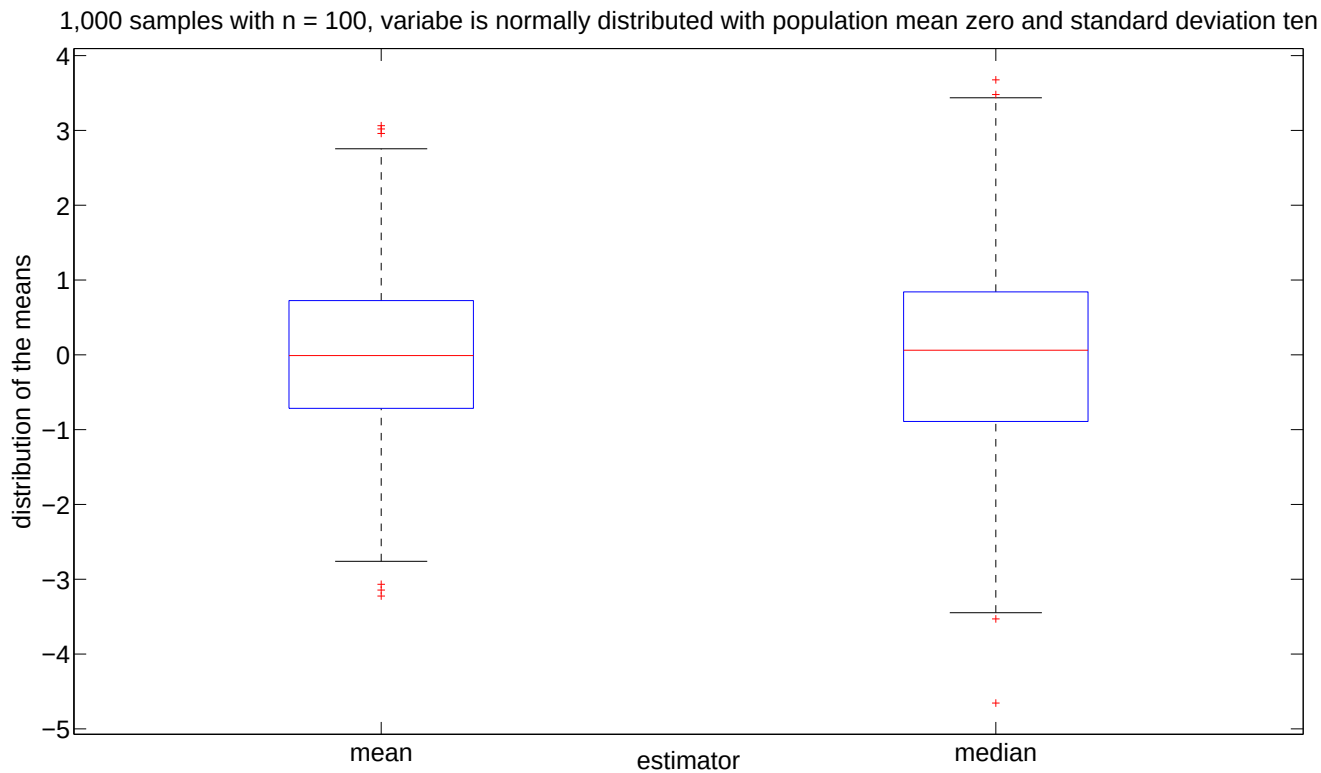
Unbiasedness of sample mean as estimator for the population mean



Consistency of the sample mean as estimator for the population mean



Efficiency of the sample mean and of the median as an estimator for the population central tendency

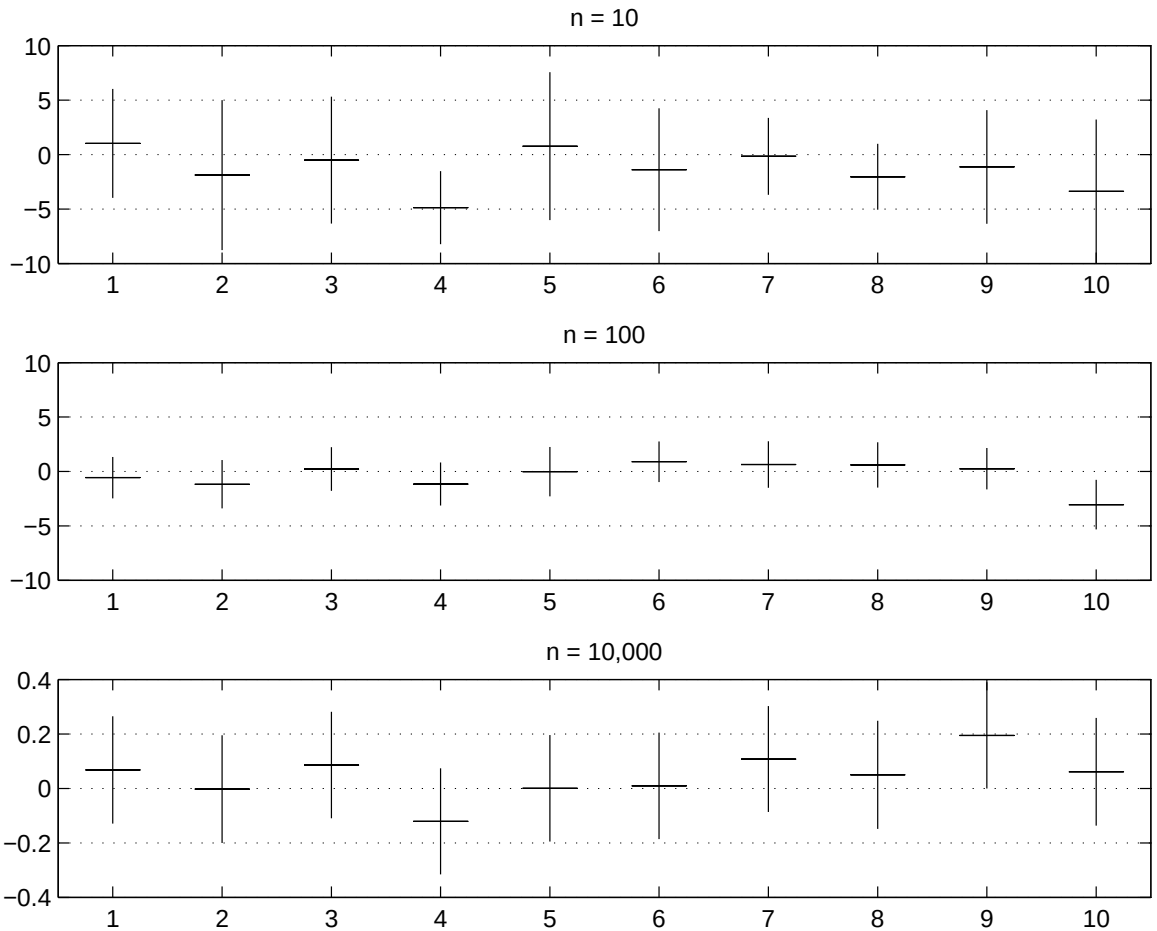


Confidence interval for the population mean

Consider a population of N observations of the variable X. We take a random sample of n observations $\{x_1, x_2, ..., x_n\}$ from the population.

- Median versus sample mean ($\bar{x}$).
- Having an estimate of a parameter is only the first step in estimation. We also need to know how precise our estimate is: Standard error. Standard error of the mean: $se_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}$
- Confidence interval for the population mean: $CI_{(1-\alpha)} : [\bar{x} - t_{df=n-1,1-\alpha} se_{\bar{x}}; \bar{x} + t_{df=n-1,1-\alpha} se_{\bar{x}}]$

95% confidence interval for the population mean



Statistical tests and scientific hypotheses

A statistical test is a confrontation of the real world (observations) to a theory (model) with the aim of falsifying the model.

Model: $H_0 : \mu = 0$ and $H_a : \mu \neq 0$
Real world: $\bar{x}, s$

Statistical tests and scientific hypotheses

As such the statistical test (as a scientific method) fits directly into the philosophy of science described by the English philosopher Karl Popper (1902–1994) (see e.g. The Logic of Scientific Discovery, 1972). Basically the philosophy says that 1) theories can not be empirically verified but only falsified and 2) scientific progress happens by having a theory until it is falsified. That is, if we observe a phenomenon (data) which under the model (theory) is very unlikely, then we reject the model (theory).

Statistical tests and scientific hypotheses

"*No amount of experimentation can ever prove me right; a single experiment can prove me wrong.*" (Albert Einstein)

In other words, experiments can mainly be used for falsifying a scientific hypothesis – never for proving it! When we have a scientific theory, we conduct an experiment in order to falsify it. Therefore, the strong conclusion arising from an experiment is when a hypothesis is rejected. Accepting (more precisely – not rejecting) a hypothesis is not a very strong conclusion (maybe acceptance is simply due to that the experiment is too small).

Example

Suppose we have a coin, and that our hypothesis is that the coin is fair, i.e. that P(head) = P(tail) = 1/2. Suppose we toss a coin $n = 25$ times and observe 21 heads. The probability of actually observing these data under the model is P(21 heads, 4 tails) = 0.0004. It is a very unlikely (but possible) event to see such data if the model is true. In this falsification process we employ the interpretation principle of statistics:

Unlikely events do not occur...

Statistical tests and scientific hypotheses

If we do not employ this principle we can never say anything at all on the basis of statistics (observations): An opponent can always claim that the present observations just are "an unfortunate outcome" which - no matter how unlikely they are - are possible.

Statistical tests and scientific hypotheses

In practice the statistical interpretation principle needs more structure:

- In a large sample space, all possible outcomes will have a very small probability, so it will be unlikely to have the data one has.
- In addition there is also the question about how small a probability is needed in order to classify data as being unlikely.
- Concepts of *p*-value and significance level α.

Two Types of Errors

Recall that the following four outcomes are possible when conducting a test:

Reality	Our Decision	
	H_0	H_a
H_0	$\surd$ (Prob $= 1 - \alpha$)	Type I Error Prob $= \alpha$
H_a	Type II Error Prob $= \beta$	$\surd$ (Prob $= 1 - \beta$)

The significance level α of any fixed level test is the probability of a Type I error.

Acceptable levels of errors

- Type I error (α)
  - Typically $\alpha = 0.05$ (This convention is due to R.A. Fisher)
  - For more stringent tests $\alpha = 0.01$ or $\alpha = 0.001$
  - Exploratory or preliminary experiments $\alpha = 0.10$
- Type II error (β)
  - Typically 0.20
  - Often unspecified and much less than 0.20
- Statistical power=$(1 - \beta)$

The power of a statistical test

The power of a significance test measures its ability to detect an alternative hypothesis.
The power against a specific alternative is calculated as the probability that the test will reject H_0 when that specific alternative is true.

Example: Computing statistical power

Does exercise make strong bones?

Can a 6-month exercise program increase the total body bone mineral content (TBBMC) of young women? A team of researchers is planning a study to examine this question. Based on the results of a previous study, they are willing to assume that $\sigma = 2$ for the percent change in TBBMC over the 6-month period. A change in TBBMC of 1% would be considered important, and the researcher would like to have a reasonable chance of detecting a change this large or larger. Are 25 subjects a large enough sample for this project?

Example (cont.)

1. State the hypotheses: let μ denote the mean percent change:

$$H_0 : \mu = 0$$
$$H_a : \mu > 0$$

2. Calculate the rejection region: The z test rejects H_0 at the $\alpha = 0.05$ level whenever:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x}}{2/\sqrt{25}} \geq 1.645$$

That is we reject H_0 when $\bar{x} \geq 0.658$.

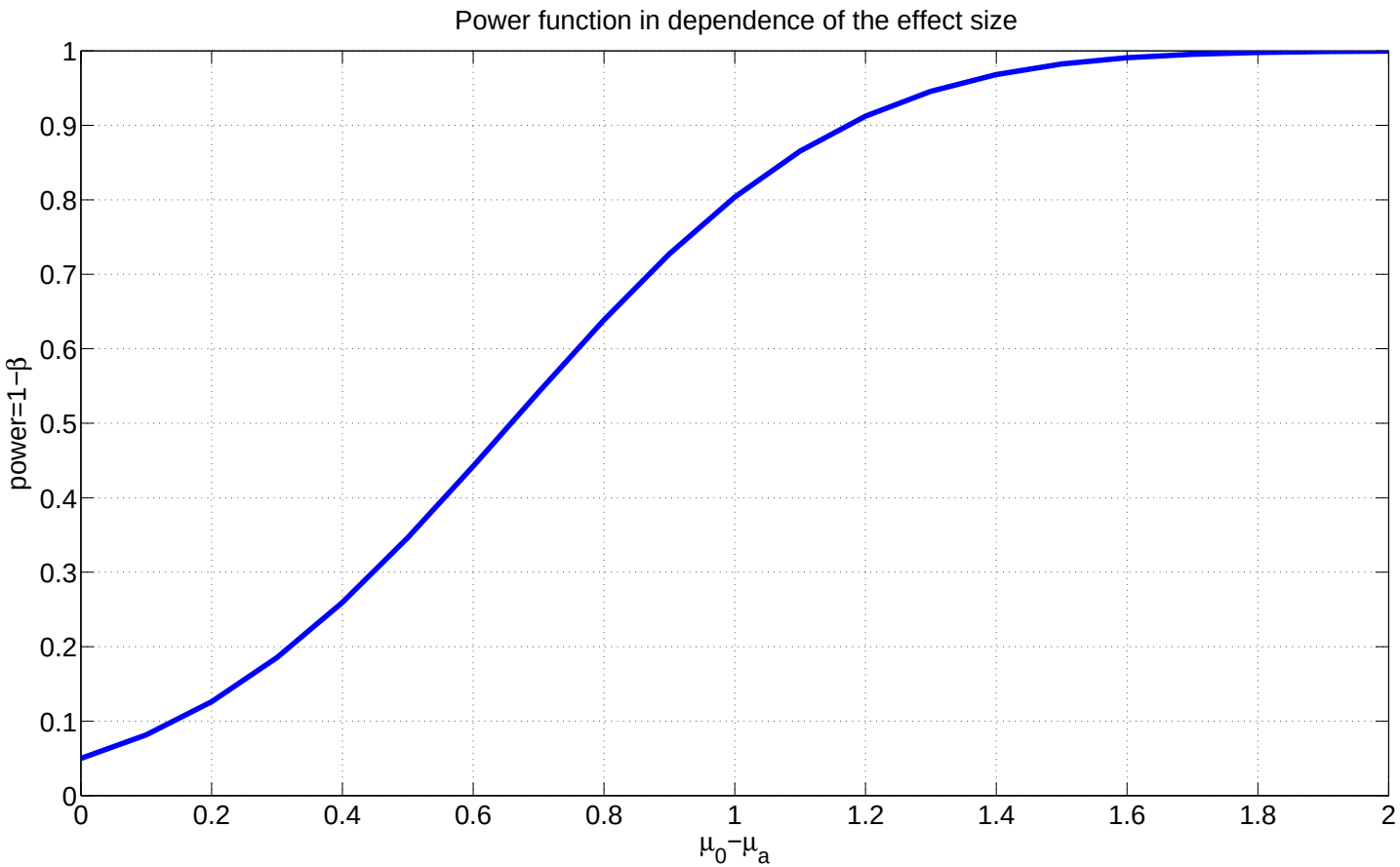
Example (cont.)

3. Compute the power at a specific alternative: The power of the test at alternative $\mu = 1$ is

$$P(\bar{x} \geq 0.658 | \mu_a = 1) = 0.8$$

   Plot graph.
4. Statistical power is the probability of rejecting H_0 given population effect size (ES), α and sample size (n). This calculation also requires knowledge of the sampling distribution of the test statistic under the alternative hypothesis: Power curve.

Example (cont.)



Ways to increase power

- Increase α. A 5% test of significance will have a greater chance of rejecting the alternative than a 1% test because the strength of evidence required for rejection is less.
- Consider a particular alternative that is farther away from μ_0. Values of μ that are in H_a but lie close to the hypothesized value μ_0 are harder to detect than values of μ that are far from μ_0.
- Increase the sample size. More data will provide more information about $\bar{x}$ so we have a better chance of distinguishing values of μ.

Ways to increase power

- Decrease σ. This has the same effect as increasing the sample size: it provides more information about μ. Improving the measurement process and restricting attention to a subpopulation are two common ways to decrease σ.

How many samples are needed to achieve a power of 0.8 in a *t*-test?

Effect size index for the *t*-test for a difference between two independent means.

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

where d is the effect size index, μ_1 and μ_2 are means, σ is the common standard deviation of the means.

Effect size indices are available for many statistical tests.

How many samples are needed to achieve a power of 0.8 in a t-test?

Effect Size	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
Large effect ($d = 0.8$)	20	26	38
Medium effect ($d = 0.5$)	50	64	95
Small effect ($d = 0.2$)	310	393	586

Source: Cohen (1992), p. 158.

Recommendation: Use estimates of statistical power as a guide to planning experiments (a priori power analysis).

Is lack of statistical power a widespread problem?

"We estimated the statistical power of the first and last statistical test presented in 697 papers from 10 behavioral journals ... On average statistical power was 13-16% to detect a small effect and 40-47% to detect a medium effect. This is far lower than the general recommendation of a power of 80%. By this criterion, only 2-3%, 13-21%, and 37-50% of the tests examined had a requisite power to detect a small, medium, or large effect, respectively."

Jennions, M.D., and A.P. Moeller 2003. Behavioral Ecology 14, 438-455.

Further readings

Cohen, J. 1992. A power primer. Psychological Bulletin 112: 155-159.

Jennions, M.D., and A.P. Moeller 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. Behavioral Ecology 14: 438-455.

Hoenig, J.M., and D.M. Heisey 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. American Statistician 55: 19-24.

Why multivariate analysis?

	Male	Female
Accept	35	20
Refuse entry	45	40
Total	80	60

- Example: 44% of male applicants are admitted by a university, but only 33% of female applicants.
- Does this mean there is unfair discrimination?
- University investigates and breaks down figures for Engineering and English programmes.

Simpson's Paradox

Engineering	Male	Female
Accept	30	10
Refuse entry	30	10
Total	60	20

English	Male	Female
Accept	5	10
Refuse entry	15	30
Total	20	40

- No relationship between sex and acceptance for either programme. So no evidence of discrimination. Why?
- More females apply for the English programme, but it is hard to get into. More males applied to Engineering, which has a higher acceptance rate than English. Must look deeper than single cross-tab to find this out!