Bioinformatics and Biostatistics BB2440: Biostatistics
Lecture 4: Statistical Inference
Timo Koski

TK

11.09.2013

KTH Matematik

Outline of Lecture 4.

- Statistical inference: What is it?

Outline of Lecture 4.

- Statistical inference: What is it?
- PART I: Estimating a Population Proportion

KTH Matematik

Outline of Lecture 4.

- Statistical inference: What is it?
- PART I: Estimating a Population Proportion
  - The proportion estimate

KTH Matematik

Outline of Lecture 4.

- Statistical inference: What is it?
- PART I: Estimating a Population Proportion
  - The proportion estimate
  - Maximum likelihood estimate

KTH Matematik

Outline of Lecture 4.

- Statistical inference: What is it?
- PART I: Estimating a Population Proportion
  - The proportion estimate
  - Maximum likelihood estimate
  - Confidence interval (interval estimate), confidence level, critical value, standard error, margin of error

KTH Matematik

Outline of Lecture 4.

- Statistical inference: What is it?
- PART I: Estimating a Population Proportion
  - The proportion estimate
  - Maximum likelihood estimate
  - Confidence interval (interval estimate), confidence level, critical value, standard error, margin of error
- PART II: Estimating a Population Mean

KTH Matematik

Outline of Lecture 4.

- Statistical inference: What is it?
- PART I: Estimating a Population Proportion
  - The proportion estimate
  - Maximum likelihood estimate
  - Confidence interval (interval estimate), confidence level, critical value, standard error, margin of error
- PART II: Estimating a Population Mean
  - The estimate of a population mean, the model via Central limit theorem

KTH Matematik

Outline of Lecture 4.

- Statistical inference: What is it?
- PART I: Estimating a Population Proportion
  - The proportion estimate
  - Maximum likelihood estimate
  - Confidence interval (interval estimate), confidence level, critical value, standard error, margin of error
- PART II: Estimating a Population Mean
  - The estimate of a population mean, the model via Central limit theorem
  - Confidence interval (CI) (interval estimate), confidence level, critical value, standard error, margin of error

KTH Matematik

Outline of Lecture 4.

- Statistical inference: What is it?
- PART I: Estimating a Population Proportion
  - The proportion estimate
  - Maximum likelihood estimate
  - Confidence interval (interval estimate), confidence level, critical value, standard error, margin of error
- PART II: Estimating a Population Mean
  - The estimate of a population mean, the model via Central limit theorem
  - Confidence interval (CI) (interval estimate), confidence level, critical value, standard error, margin of error
    - CI for a population mean when standard deviation is known

KTH Matematik

Outline of Lecture 4.

- Statistical inference: What is it?
- PART I: Estimating a Population Proportion
  - The proportion estimate
  - Maximum likelihood estimate
  - Confidence interval (interval estimate), confidence level, critical value, standard error, margin of error
- PART II: Estimating a Population Mean
  - The estimate of a population mean, the model via Central limit theorem
  - Confidence interval (CI) (interval estimate), confidence level, critical value, standard error, margin of error
    - CI for a population mean when standard deviation is known
    - CI for a population mean when standard deviation is not known, t-distribution and critical values

KTH Matematik

Outline of Lecture 4.

- Statistical inference: What is it?
- PART I: Estimating a Population Proportion
  - The proportion estimate
  - Maximum likelihood estimate
  - Confidence interval (interval estimate), confidence level, critical value, standard error, margin of error
- PART II: Estimating a Population Mean
  - The estimate of a population mean, the model via Central limit theorem
  - Confidence interval (CI) (interval estimate), confidence level, critical value, standard error, margin of error
    - CI for a population mean when standard deviation is known
    - CI for a population mean when standard deviation is not known, t-distribution and critical values
  - A Word about Bootstrap

KTH Matematik

Statistical Inference

Statistical inference is the process of drawing conclusions from data that is subject to random variation, for example, observational errors or sampling variation

KTH Matematik

Statistical Inference

For the most part, statistical inference makes **statements about populations**, using sampled data drawn from the population of interest. Given a parameter or hypothesis related to the population about which one wishes to make inference, statistical inference most often uses:

- a **statistical model** (e.g. Signal + Noise), which describes the population that is supposed to generate the data

KTH Matematik

Statistical Inference

For the most part, statistical inference makes **statements about populations**, using sampled data drawn from the population of interest. Given a parameter or hypothesis related to the population about which one wishes to make inference, statistical inference most often uses:

- a **statistical model** (e.g. Signal + Noise), which describes the population that is supposed to generate the data
- Some common forms of a statistical inference are:

KTH Matematik

Statistical Inference

For the most part, statistical inference makes **statements about populations**, using sampled data drawn from the population of interest. Given a parameter or hypothesis related to the population about which one wishes to make inference, statistical inference most often uses:

- a **statistical model** (e.g. Signal + Noise), which describes the population that is supposed to generate the data
- Some common forms of a statistical inference are:
  - **an estimate**: a particular value that best approximates some parameter of interest

KTH VETENSKAP OCH KONST
KTH Matematik

Statistical Inference

For the most part, statistical inference makes **statements about populations**, using sampled data drawn from the population of interest. Given a parameter or hypothesis related to the population about which one wishes to make inference, statistical inference most often uses:

- a **statistical model** (e.g. Signal + Noise), which describes the population that is supposed to generate the data
- Some common forms of a statistical inference are:
  - **an estimate**: a particular value that best approximates some parameter of interest
  - **a confidence interval**: an interval constructed using a dataset drawn from a population so that, under repeated sampling of such datasets, such intervals would contain the true parameter value with the probability at the stated **confidence level**

KTH Matematik

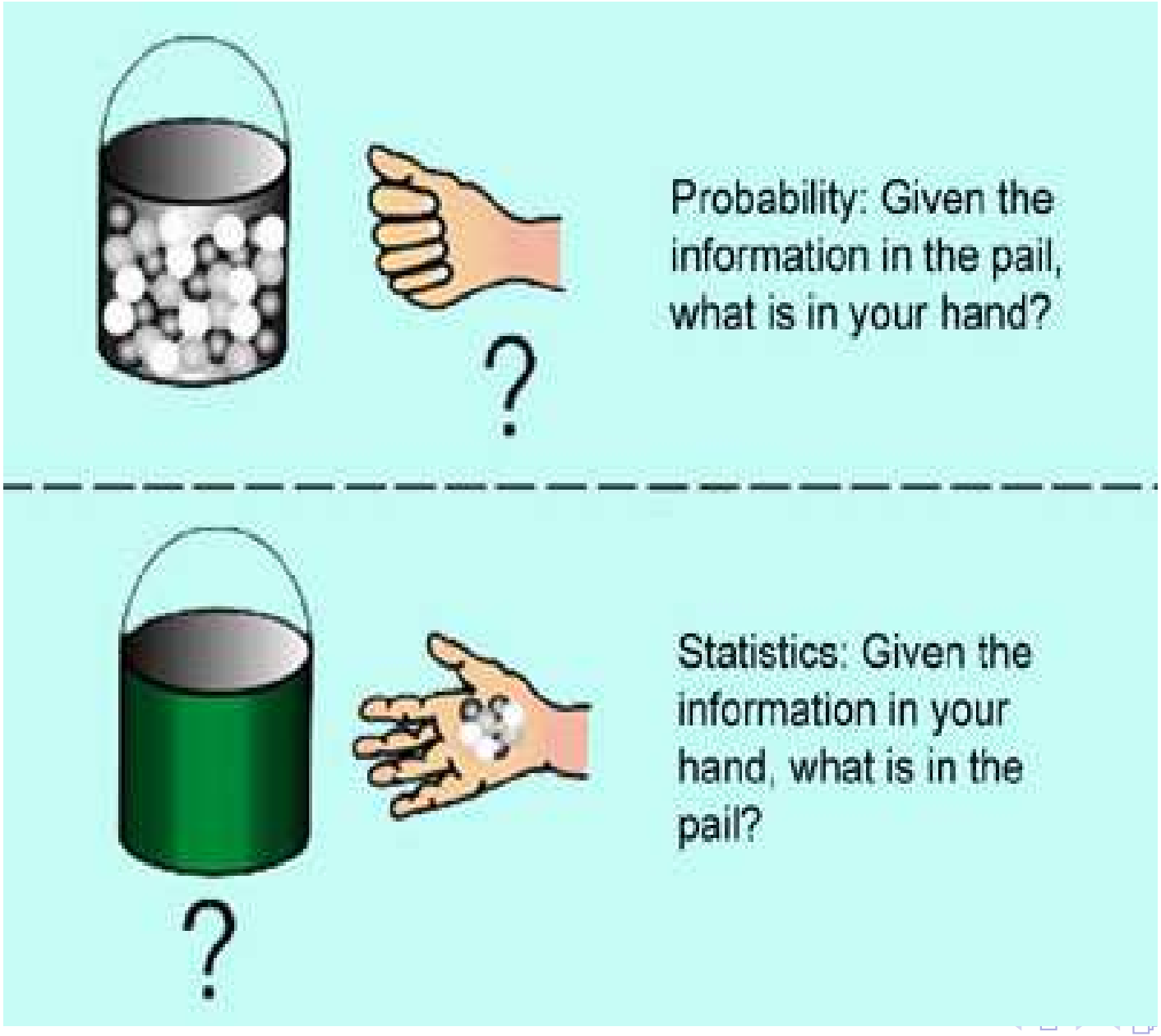
Statistical Inference

For the most part, statistical inference makes **statements about populations**, using sampled data drawn from the population of interest. Given a parameter or hypothesis related to the population about which one wishes to make inference, statistical inference most often uses:

- a **statistical model** (e.g. Signal + Noise), which describes the population that is supposed to generate the data
- Some common forms of a statistical inference are:
  - **an estimate**: a particular value that best approximates some parameter of interest
  - **a confidence interval**: an interval constructed using a dataset drawn from a population so that, under repeated sampling of such datasets, such intervals would contain the true parameter value with the probability at the stated **confidence level**
  - **rejection of a hypothesis** by a **statistical test**

KTH Matematik

Probability and Statistics



PART I: How to estimate a proportion ?

When Mendel conducted his famous genetics experiments with peas, one sample of offspring was obtained by crossing peas with green pods and peas with yellow pods. The offspring consisted of 580 peas. Among them 428 had green pods, and 152 had yellow pods. This is our data. Let us write

$$x = 152 \quad \text{yellow pods in one sample of offspring}$$

We want to infer, e.g., the **proportion of yellow pods that would be obtained in similar experiments**. We call this proportion the **population parameter** and denote it by p.
An estimate of p:

$$\widehat{p} = \frac{x}{428 + 152} \Rightarrow \widehat{p} = \frac{152}{580} = 26.2\%$$

KTH Matematik

PART I: A Quote

> Mendel . . . did something that, more than anything else, marks the birth of modern genetics, he **counted** the numbers of plants with each phenotype.

p. 25 in A.J.F. Griffiths, J.H. Miller, D.T. Suzuki, R.L. Lewontin, W.M. Gelbart: *An Introduction to Genetic Analysis*, W.H.Freeman and Company, New York, 1997.

KTH Matematik

How to estimate a proportion ? Questions

An estimate of p:

$$\widehat{p} = \frac{x}{428 + 152} \Rightarrow \widehat{p} = \frac{152}{580} = 26.2\%$$

Questions:

- What do we know about the accuracy of the estimate ?
- The theory of Mendel was that 25 % of the peas would have yellow pods. How do we explain the discrepancy to 26.2 % ?
- Is the discrepancy large enough to suggest that Mendel's 25 % was wrong?

KTH Matematik

Notations for Proportions

$$p = \textit{proportion in the entire population}$$

$$\widehat{p} = \frac{x}{n} = \textit{sample proportion of } x \textit{ successes in a sample of size } n$$

$$\widehat{q} = 1 - \widehat{p} = \textit{sample proportion of failures in a sample of size } n$$

KTH Matematik

Expand Your Mind

$p =$ *proportion in the entire population and a statistical model.*

Statistical Model !

We may for good reasons (recall the conditions for binomial distribution) consider the statistical model

$$X \in \text{Bin}(580, p)$$

and regard $x = 152$ as an outcome of X.

In words, we have now made the proportion in the entire population to a parameter of our statistical model.

Then

$$\widehat{p} = \frac{X}{580}$$

another random variable, the estimator of p.

KTH VETENSKAP OCH KONST
KTH Matematik

Statistical Model

$$X \in \text{Bin}(580, p)$$

We know that $E(X) = 580 \cdot p$, $V(X) = 580 \cdot p \cdot (1-p)$. This gives also

$$E(\widehat{p}) = E\left(\frac{X}{580}\right) = \frac{1}{580}E(X) = \frac{1}{580} \cdot 580 \cdot p = p.$$

$$V(\widehat{p}) = V\left(\frac{X}{580}\right) = \frac{1}{580^2}V(X) = \frac{1}{580^2} \cdot 580 \cdot p(1-p) = \frac{p(1-p)}{580}.$$

These expressions give us formulae for study of the accuracy of the estimate, but they depend on the unknown population proportion or the parameter p.

KTH Matematik

Accuracy

We insert the sample proportions of sucesses and failures

$$E(\widehat{p}) = p, \quad V(\widehat{p}) = \frac{p(1-p)}{580}.$$

$$D(\widehat{p}) = \sqrt{V(\widehat{p})} = \sqrt{\frac{p(1-p)}{580}}$$

and get the **standard error** (an estimate of $D(\widehat{p})$)

$$d(\widehat{p}) = \sqrt{\frac{0.262(1-0.262)}{580}} \approx 0.0183 = 1.83\%$$

KTH Matematik

A General Principle: Likelihood

We present next a general principle of estimation that explains the proportion estimation above. This is the method of maximum likelihood.

KTH Matematik

Maximum likelihood method for Bin(580, p)

Together with Mendel we observed $x = 152$ and modelled this by

$$X \in \text{Bin}(580, p)$$

We introduce the **likelihood function** for p

$$L(p) = \binom{580}{x} p^x (1-p)^{580-x}$$

This is just the propability for $Pr(X = x)$, if $X \in \text{Bin}(580, p)$, but we treat it now as a function of p.
We want to find the value of p that maximizes $L(p)$. We can understand this so that we find the value of p that maximizes the probability to observe $x = 152$.

KTH Matematik

Maximizing likelihood

We can maximize $L(p)$ by maximizing the logarithm of the likelihood function:

$$\ln L(p) = \ln \binom{580}{x} + x \ln p + (580 - x) \ln 1 - p$$

We differentiate $\ln L(p)$ w.r.t. p

$$\frac{d}{dp} \ln L(p) = x\frac{1}{p} - (580 - x)\frac{1}{1-p}$$

and solve $\frac{d}{dp} \ln L(p) = 0$ w.r.t. p and call the solution $\widehat{p}$.

$$x\frac{1}{p} - (580 - x)\frac{1}{1-p} = 0 \Leftrightarrow x\frac{1}{p} = (580 - x)\frac{1}{1-p}$$

$$\Leftrightarrow (1-p)x = p(580 - x) \Leftrightarrow x - px = p580 - px$$

and the maximum likelihood estimate is

$$\widehat{p} = \frac{x}{580} = \frac{152}{580} = 0.262.$$

KTH Matematik

Maximum likelihood

Hence the sample proportion of x successes in a sample of size n, or $\widehat{p} = \frac{x}{n}$, is the best estimate of p, the proportion in the entire population.

KTH Matematik

Confidence Interval

What do we know about the accuracy of the (best) estimate $\widehat{p} = \frac{x}{n}$? We introduce now a new topic to discuss this accuracy. This is called a **confidence interval**.

KTH Matematik

Confidence Interval CI

A **confidence interval** *(or* **interval estimate***) is an interval of values used to estimate a population parameter. A confidence interval is sometimes abbreviated as CI.*

KTH Matematik

Confidence Interval CI

A **confidence interval** *(or* **interval estimate***) is an interval of values used to estimate a population parameter. A confidence interval is sometimes abbreviated as CI.*

A confidence interval is associated with a confidence level.

A **confidence level** *is the probability* $1 - \alpha$ *(often expressed as the equivalent precentage value, e.g., 95 %) that is the proportion of times that the confidence interval actually does contain the population parameter, assuming the estimation process is repeated a large number of times.*

The confidence level is also called the **degree of confidence** or the **confidence coefficient**

KTH VETENSKAP OCH KONST
KTH Matematik

Confidence Interval CI

A **confidence level** *is the probability* $1 - \alpha$ *(often expressed as the equivalent precentage value, e.g., 95 %) that is the proportion of times that the confidence interval actually does contain the population parameter, assuming the estimation process is repeated a large number of times.*

The most common choices for the confidence level are 90% (with $\alpha = 0.10$), 95% (with $\alpha = 0.05$) and 99% (with $\alpha = 0.01$). The choice 95% is most common as it balances precision (the width of the interval) and reliability (expressed by the confidence level).

KTH Matematik

CI: an example

A **confidence interval** *(or* **interval estimate***) is an interval of values used to estimate a population parameter. A confidence interval is sometimes abbreviated as CI.*

An example of a confidence interval based on the sample data of 580 offspring peas with 26.2% of them having yellow pods:

Example

The 95% confidence interval estimate of the population parameter p is

$$0.226 < p < 0.298$$

KTH Matematik

CI: an example

Example

The 95% confidence interval estimate of the population parameter p is

$$0.226 < p < 0.298$$

How was this done ?? We shall now go through the whole procedure in several steps. Step one is the discussion of the critical value.

KTH Matematik

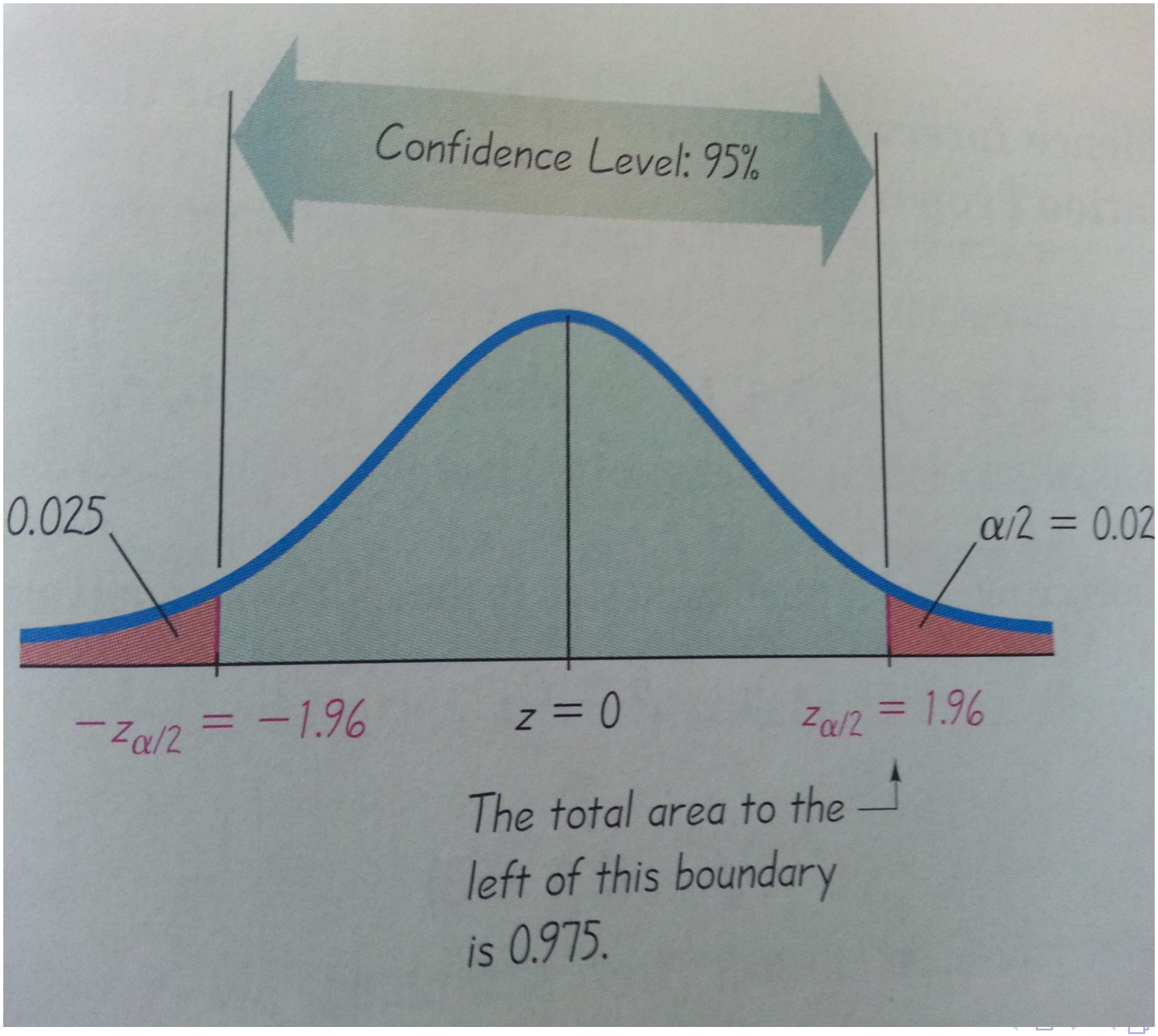
Step One: Critical value

This step is based on the fact that the distribution of the proportion estimator $\widehat{p} = \frac{X}{580}$ can be approximated by a normal distribution.

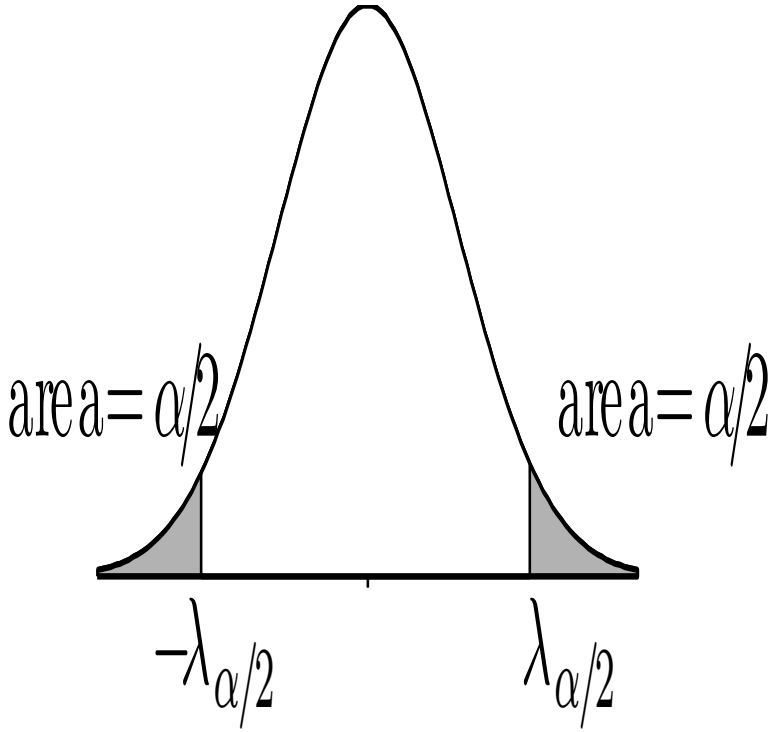
Notation for a critical value *Fix α. The critical value $z_{\alpha/2}$ is the positive z score that is at the vertical boundary separating an area of $\alpha/2$ in the right tail of the standard normal distribution. The value of $-z_{\alpha/2}$ is at the vertical boundary separating an area of $\alpha/2$ in the left tail of the standard normal distribution.*

KTH Matematik

Step One: Critical value



Step One: Critical value



Read $\lambda_{\alpha/2}$ as $z_{\alpha/2}$

Step One: Critical value

Since $z_{\alpha/2}$ separates an area of $\alpha/2$ in the right tail of the standard normal distribution, we find $\alpha/2$ mathematically speaking as the solution to the equation

$$\Phi\left(z_{\alpha/2}\right) = 1 - \alpha/2$$

Of course, this can be done only numerically. We enjoy the benefits of the work of past generations in the form of a table (There are, of course, computer and calculator routines, too).

KTH Matematik

Step One: Critical values

Critical values

$P(Z > z_\alpha) = \alpha$ där $Z \in N(0,1)$

α	z_α	α	z_α
0.10	1.2816	0.001	3.0902
0.05	1.6449	0.0005	3.2905
0.025	1.9600	0.0001	3.7190
0.010	2.3263	0.00005	3.8906
0.005	2.5758	0.00001	4.2649

KTH VETENSKAP OCH KONST
KTH Matematik

Step One: Critical values

We have fixed the 95% confidence level, where $\alpha = 0.05$. Thus $\alpha/2 = 0.025$ and the table gives us $z_{0.025} = 1.96$. This completes Step One.

KTH VETENSKAP OCH KONST
KTH Matematik

Step Two: Standard Error and the Margin of Error

In the preceding we saw the expression $d(\widehat{p}) = \sqrt{\frac{0.262(1-0.262)}{580}}$ *as an approximation of* $D(\widehat{p})$.

Step Two: Standard Error and the Margin of Error

In the preceding we saw the expression $d\left(\widehat{p}\right) = \sqrt{\frac{0.262(1-0.262)}{580}}$ *as an approximation of* $D\left(\widehat{p}\right)$.

Definition

Standard Error *The standard error* $d\left(\widehat{p}\right)$ *of the proportion estimator* $\widehat{p}$ $\left(\widehat{q} = 1 - \widehat{p}\right)$ *is*

$$d\left(\widehat{p}\right) \stackrel{\text{def}}{=} \sqrt{\frac{\widehat{p}\widehat{q}}{n}}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Step Two: Standard Error and the Margin of Error

Definition

Margin of Error *The margin of error E of the proportion estimator $\widehat{p}$ is*

$$E \stackrel{\text{def}}{=} z_{\alpha/2}\sqrt{\frac{\widehat{p}\widehat{q}}{n}}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Step Three: the confidence interval

Confidence Interval or the Interval Estimate for the population proportion p

$$\widehat{p} - E < p < \widehat{p} + E, \quad \text{where } E = z_{\alpha/2}\sqrt{\frac{\widehat{p}\widehat{q}}{n}}$$

Other equivalent expressions are

$$\widehat{p} \pm E$$

or

$$(\widehat{p} - E, \widehat{p} + E)$$

This completes the procedure of constructing the confidence interval for p.

KTH Matematik

The confidence interval for the proportion of yellow pods

Confidence Interval or the Interval Estimate for the population proportion p *We have* $\widehat{p} = 0.262$.

$$d\left(\widehat{p}\right) = \sqrt{\frac{0.262(1-0.262)}{580}} \approx 0.0183 = 1.83\%$$

$$z_{0.025} = 1.96$$

$$E = 1.96 \cdot 0.0183 = 0.035868$$

The CI is $0.262 - 0.035868 < p < 0.262 + 0.035868$, *i.e., (rounded off to three digits)*

$$0.226 < p < 0.298$$

KTH VETENSKAP OCH KONST
KTH Matematik

The statistical statement

The CI is

$$0.226 < p < 0.298$$

This CI is often reported with a statement such as this:

It is estimated that 26.2 % of the offspring peas will have yellow pods, with a margin of error of plus minus 3.6 percentage points.

People with knowldge in Swedish (and Sweden) know probably the statement " ... statistiska felmarginalen ", when media refers to opinion polls.

KTH Matematik

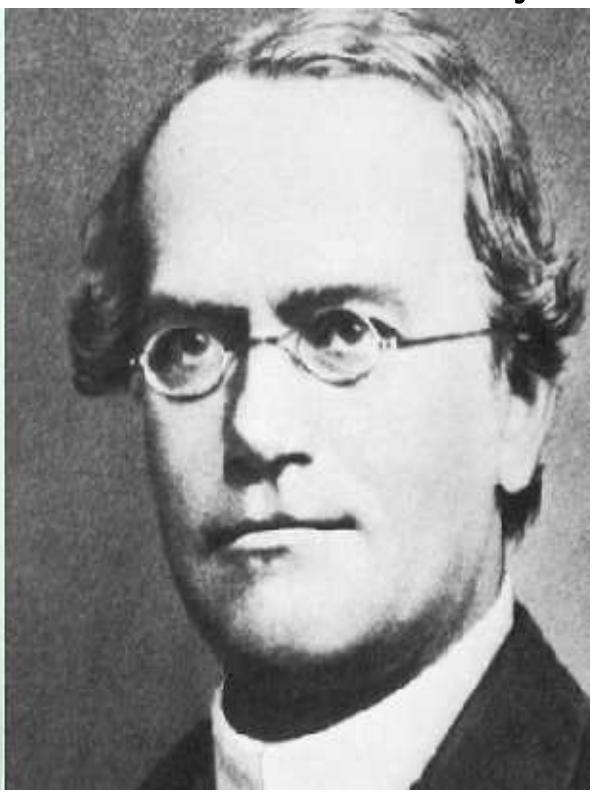
The statistical statement

People with knowledge in Swedish language (and Sweden) are probably familiar with the statement ” . . . statistiska felmarginalen ”, when media reports about opinion polls.

KTH Matematik

The statistical statement

We are now 95 % confident that the limits 22.6 % and 29.8 % contain the true percentage of offspring peas with yellow pods. The percentage of peas with yellow pods is likely to be any value between 22.6 % and 29.8 %. That interval includes 25%, so Mendel's expected value of 25% cannot be described as wrong. The results do not apper to provide significant evidence against the 25% rate claimed by Mendel.



KTH Matematik

Interpretation of the CI: Correct

We are 95 % confident that the limits 22.6 % and 29.8 % contain the true percentage of offspring peas with yellow pods.

This means that if we were to conduct many different experiments with 580 offspring peas and construct the corresponding CIs, 95 % of them would actually contain the value of the population proportion p. In this correct interpretation 95% refers to the success rate of the *process* being used to estimate the proportion and does not refer to the population proportion itself.

KTH Matematik

Interpretation of the CI: Wrong !

There is 95 % chance that the true value of p falls within the the limits 22.6 % and 29.8 %.

KTH VETENSKAP OCH KONST
KTH Matematik

Postscript: A Mathematical Aside

$$X \in \text{Bin}(n, p)$$

We know that $E(X) = n \cdot p$, $V(X) = n \cdot p \cdot (1-p)$. Then we know (Lecture 4.) that

$$\widehat{p} = \frac{X}{n} \quad \text{approximatively} \quad \sim N\left(p, \sqrt{\frac{p \cdot (1-p)}{n}}\right)$$

We form the population Z score

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}} \quad \text{approximatively} \sim N(0, 1)$$

Thus with the critical value $z_{\alpha/2}$

$$Pr\left(-z_{\alpha/2} \leq \frac{\widehat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

KTH VETENSKAP OCH KONST
KTH Matematik

Postscript: A Mathematical Aside

Then manipulation with inequalities gives

$$Pr\left(-z_{\alpha/2} \leq \frac{\widehat{p}-p}{\sqrt{\frac{p\cdot(1-p)}{n}}} \leq z_{\alpha/2}\right) =$$

$$Pr\left(-z_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} \leq \widehat{p}-p \leq z_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}\right) =$$

$$= Pr\left(-\widehat{p}-z_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} \leq -p \leq -\widehat{p}+z_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}\right)$$

$$= Pr\left(\widehat{p}-z_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} \leq p \leq \widehat{p}+z_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}\right)$$

KTH Matematik

Postscript: A Mathematical Aside

We have thus

$$Pr\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} \leq p \leq \widehat{p} + z_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}\right) \approx 1-\alpha$$

We replace $\frac{p\cdot(1-p)}{n}$ with $\frac{\widehat{p}\cdot\widehat{q}}{n}$ and thus get by definition of the margin of error E

$$Pr\left(\widehat{p} - E \leq p \leq \widehat{p} + E\right) \approx 1-\alpha$$

which is one of the expressions of our confidence statement in the preceding. End of postscript. □

PART II: Estimating a Population Mean

Now we consider the task of estimating a population mean. These are questions like

- What is the mean amount milk obtained from cows in Skåne in a year?

Here we think of proceeding by observing the amount of milk produced by cows in a randomly chosen subpopulation of farms in Skåne.

KTH VETENSKAP OCH KONST
KTH Matematik

Estimating a Population Mean: Statistical Model

We regard the estimation of mean as estimating the mean $E(x)$ of a statistical distribution, like

$$\mu \equiv E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

This is not the same as the sample mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

for data/samples $x_1, x_2, \ldots, x_n$.

We are going to take $\bar{x}$ as estimate of $E(X)$.

KTH VETENSKAP OCH KONST

KTH Matematik

Estimating a Population Mean: Statistical Rationale

We are going to take $\bar{x}$ as estimate of μ. Why ?

$X_1, X_2, \ldots, X_n$ are independent have the same distribution (=equally distributed) with mean μ och standard deviation σ. If n is very large we have

The Law of Large Numbers

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mu$$

where μ is the common mean of $X_1, X_2, \ldots, X_n, \ldots$

KTH VETENSKAP OCH KONST
KTH Matematik

Estimating a Population Mean: Statistical Rationale

We are going to take $\bar{x}$ as estimate of μ. Why ?

Consistency: by the Law of Large Numbers $\bar{x}$ is close to the true value μ !
There is no systematic error (a.k.a. bias) either, since $E(\bar{X}) = \mu$.

KTH VETENSKAP OCH KONST
KTH Matematik

Estimating a Population Mean: Statistical Rationale

$X_1, X_2, \ldots, X_n$ are all $N(\mu, \sigma)$, then $\bar{x}$ is the maximum likelihood estimate of μ, too. (Details omitted).

KTH Matematik

Estimating a Population Mean

We are going to follow the same general procedure as above, i.e,

- Find the critical value for chosen level of confidence α.
- Find the margin of error E
- Form the confidence interval (confidence level $1 - \alpha$)

$$\bar{x} - E < \mu < \bar{x} + E$$

KTH Matematik

Estimating a Population Mean: the critical value

When finding the critical value we have to consider two cases (a) and (b):

(a) $\sigma = \sqrt{V(X)} = \sqrt{\int_{-\infty}^{\infty}(x - E(X))^2 f(x)dx}$ *is* **known**

(b) σ *is* **not known**

KTH Matematik

Estimating a Population Mean (a) : the critical value when σ **known**

The critical value $z_{\alpha/2}$ is found as in PART I (estimation of population proportion).

KTH Matematik

Estimating a Population Mean (a)

This is justified by the central limit theorem applied to arithmetic mean

$$\bar{X} \text{ approximatively} \sim N\left(\mu, \sigma/\sqrt{n}\right)$$

Or, with with the cumulative density $\Phi(x)$

$$P(a < \bar{X} \leq b) \approx \Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right)$$

KTH Matematik

Z Score

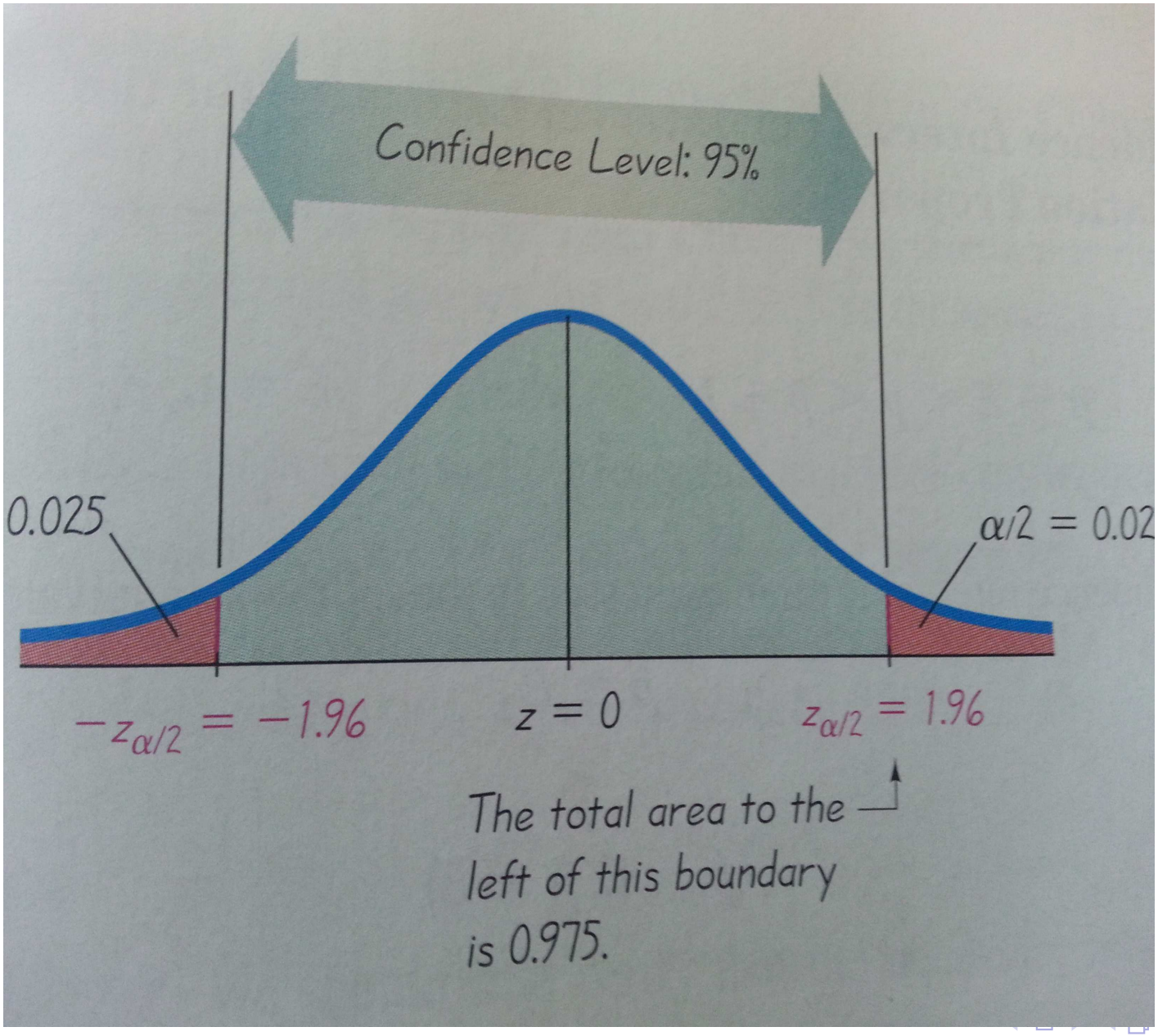
Therefore the population Z Score is standard normal

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

and we have for Z the same picture as above.

KTH VETENSKAP OCH KONST
KTH Matematik

Critical value again



Estimating a Population Mean (a): σ **known**

We can therefore to follow exactly the same procedure as above, i.e,

- Find the critical value $z_{\alpha/2}$ for chosen level of confidence α.
- Find the margin of error E
- Form the confidence interval (confidence level $1-\alpha$)

$$\bar{x} - E < \mu < \bar{x} + E$$

KTH Matematik

Estimating a Population Mean (a): the margin of error

When σ is **known**, the margin of error E is

$$E \stackrel{\text{def}}{=} z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

This reflects the fact that $\frac{\sigma}{\sqrt{n}}$ is standard deviation in $N\left(\mu, \sigma/\sqrt{n}\right)$.

KTH VETENSKAP OCH KONST
KTH Matematik

The confidence interval (a): σ is **known**

Confidence Interval or the Interval Estimate for the population mean μ

$$\bar{x} - E < \mu < \bar{x} + E, \quad \text{where } E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Other equivalent expressions are

$$\bar{x} \pm E$$

or

$$(\bar{x} - E, \bar{x} + E)$$

KTH Matematik

Interpretation of the CI: Correct

Suppose we have with $\alpha = 0.05$ and some data obtained by the procedure above $98.08 < \mu < 98.32$

We are 95 % confident that the interval with endpoints 98.08 and 98.32 contains the true value of μ.

KTH Matematik

Interpretation of the CI: Correct

This means that if we were to select many different samples of the same size and construct the corresponding CIs, 95 % of them would actually contain the value of the population proportion μ. In this correct interpretation 95% refers to the success rate of the *process* being used to estimate the mean.

Interpretation of the CI: Wrong !

Because μ is a fixed constant (but unknown) it would be wrong to say that " there is a 95% chance that μ will fall between 98.08 and 98.32 ". The CI does not describe the behaviour of individual sample values, so it would be wrong to say that " 95% of all data values fall between 98.08 and 98.32 ". The CI does not describe the behaviour of individual sample means, so it would also be wrong to say that " 95% of sample means fall between 98.08 and 98.32 ".

KTH Matematik

The confidence interval (b): σ **not known**

Since the standard deviation σ is not known we need first an estimate of σ. We already know

$$s = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

as such an estimate but we should in fact take (!)

$$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

KTH Matematik

t score

If $s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2}$, then we have the following:

The distribution of

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

is essentially a **t-distribution** *with* $n - 1$ *degrees of freedom .*

KTH VETENSKAP OCH KONST
KTH Matematik

t-distribution : Degreees of freedom ??

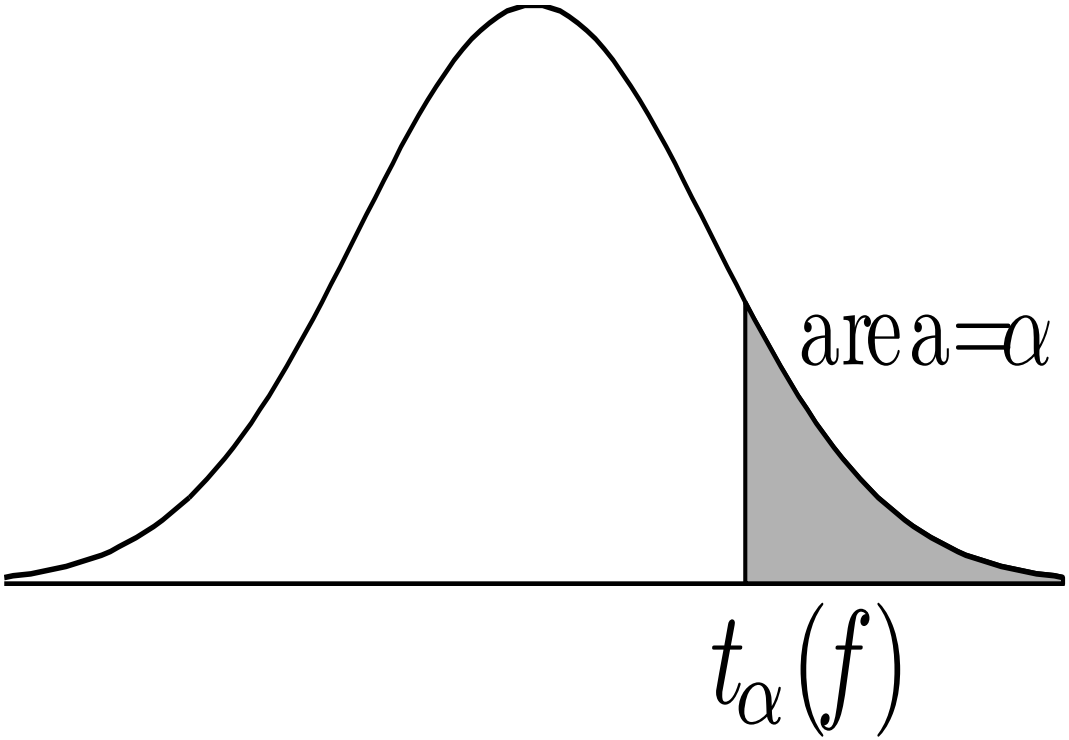
The number of **degrees of freedom** *is the number of sample values that can vary after certain restrictions have been imposed on all data values*

For example, if $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = 5$ is imposed, we can vary, e.g., $x_1, x_2, \ldots, x_n - 1$, but x_n must be chosen so that $\bar{x}$ is equal to 5. Hnece the degree of freedom is $n - 1$.

KTH VETENSKAP OCH KONST
KTH Matematik

Critical value (b): σ **not known**

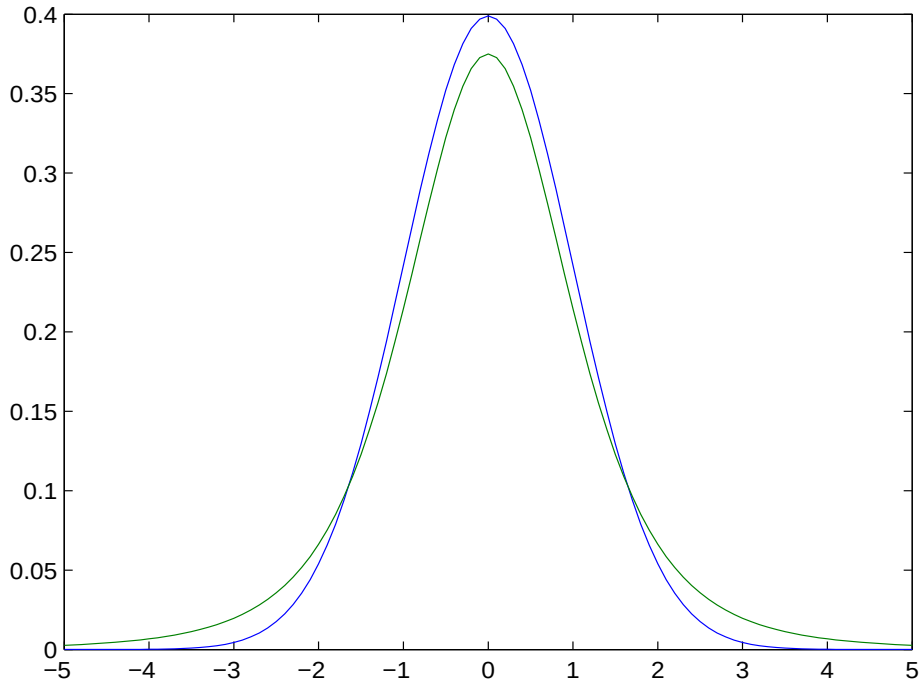
We use the t-distribution to find the critical value. These are now denoted by $t_{\alpha/2}$. The logic is the same as with $z_{\alpha/2}$, since t-distribution is symmetric as is $N(0,1)$.



KTH Matematik

t-distribution

The density curves for $N(0,1)$ (blue) and t with four degrees of freedom

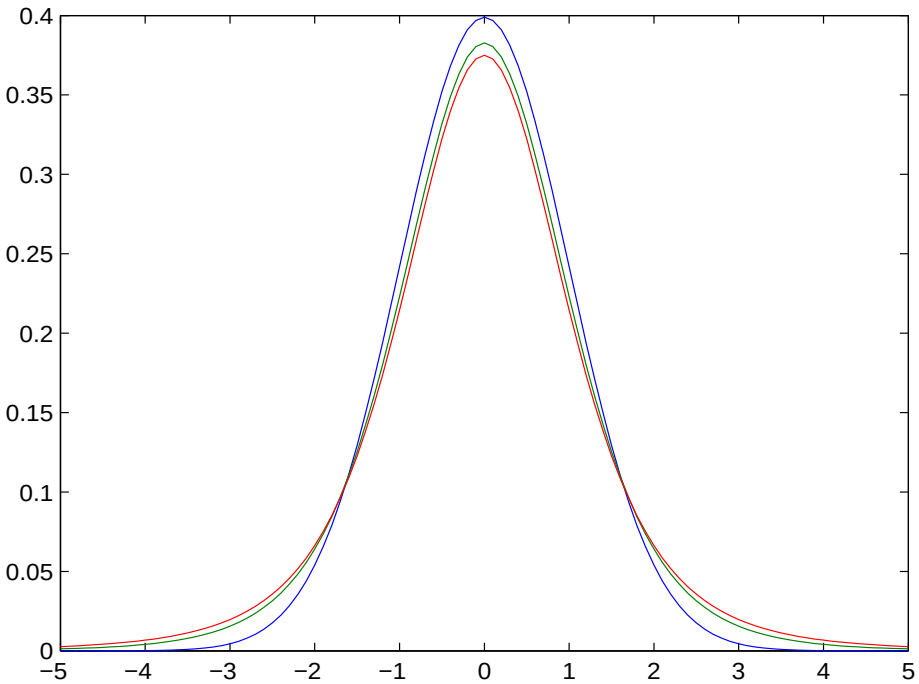


(green)

KTH Matematik

t-distribution

The density curves for $N(0,1)$ (blue) and t with four degrees of freedom (red) and t with six degrees of freedom (green)



KTH VETENSKAP OCH KONST
KTH Matematik

Critical value

Here we find t_α for various values of α and f =degrees of freedom.

f \ α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.08	6.31	12.71	31.82	63.66	318.31	636.62
2	1.89	2.92	4.30	6.96	9.92	22.33	31.60
3	1.64	2.35	3.18	4.54	5.84	10.21	12.92
4	1.53	2.13	2.78	3.75	4.60	7.17	8.61
5	1.48	2.02	2.57	3.36	4.03	5.89	6.87

KTH Matematik

For example $t_{0.05} = 2.02$ for 5 degrees of freedom. Norman & Steiner: Bare Essentials gives these values in TABLE C p. 361. Note their distinction of two- and one-sided tests. In other words, they give 2.025 in the two tailed test with 0.05.

KTH Matematik

Estimating a Population Mean: the margin of error (b)

When σ is **not known**, the margin of error E is

$$E \stackrel{\text{def}}{=} t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ has $n-1$ degrees of freedom.

KTH VETENSKAP OCH KONST
KTH Matematik

The confidence interval: σ not known

Confidence Interval or the Interval Estimate for the population mean μ σ not known

$$\bar{x} - E < \mu < \bar{x} + E, \quad \text{where } E = t_{\alpha/2}\frac{s}{\sqrt{n}}$$

or

$$\bar{x} \pm E$$

or

$$(\bar{x} - E, \bar{x} + E)$$

KTH VETENSKAP OCH KONST
KTH Matematik

t-distribution

The derivation of the t-distribution was first published in 1908 by William Sealy Gosset, while he worked at a Guinness Brewery in Dublin. He was prohibited from publishing under his own name, so the paper was written under the pseudonym Student.

KTH VETENSKAP OCH KONST
KTH Matematik

Bootstrap: dictionary

(1) Loop of leather or cloth sewn at the top rear, or sometimes on each side, of a boot to facilitate pulling it on. (2) a means of advancing oneself or accomplishing something relying entirely on one's efforts and resources.

KTH Matematik

Bootstrap & the Confidence Interval for population mean

We have samples $x_1, \ldots, x_n$. Then we pick at random with replacement a fictitious ' sample $x_1^*, \ldots, x_n^*$. This is called a bootstrap (re)sample. For example

$$x_1 = -0.2746, x_2 = -1.1730, x_3 = 1.4842, x_4 = 1.1454, x_5 = -1.6248,$$

$$x_6 = 0.9985, x_7 = 0.4571, x_8 = -1.2315, x_9 = 0.9868, x_{10} = -0.5941$$

so that a resample is

$$x^* = \begin{bmatrix} -1.6248 & -0.2746 & 0.9985 & 1.4842 & -1.1730 \\ -1.6248 & 0.9985 & 1.4842 & 0.4571 & -0.5941 \end{bmatrix};$$

KTH VETENSKAP OCH KONST
KTH Matematik

Bootstrap & the Confidence Interval for population mean

Then we repeat this, say thousand times, and compute for each of the thousand bootstrap samples x^* the values of the arithmetic mean $\bar{x}^*$ and $(s^2)^*$. Then we compute the empirical histogram of the thousand bootstrap z-scores

$$z^* = \frac{\bar{x} - \bar{x}^*}{s^*}$$

(same $\bar{x}$ computed from the original samples in every case). By a computer search we can find the quantiles and percentiles of the histogram.

KTH Matematik

Bootstrap & the Confidence Interval for population mean

Then the bootstrap confindence interval is

$$(\bar{x} - sz^*_{(1-\alpha)}, \bar{x} - sz^*_{(\alpha)})$$

where $z^*_{(1-\alpha)}$ denotes the $1-\alpha$ percentile[1] of the histogram of the bootstrapped z-scores z^*. s^2 is the variance of the original samples. This is useful and practical in such situations, where the assumptions (=asymptotic normal distribution) in PART II do not seem to hold.

[1] A **percentile** indicates the value below which a given percentage of observations in a group of observations fall.