Statistik för bioteknik SF1911
Föreläsning 8: Modellbaserad data-analys

Timo Koski

TK

16.11.2017

KTH Matematik

Outline of Lecture

Population, data, models

Statistical models with unknown parameters

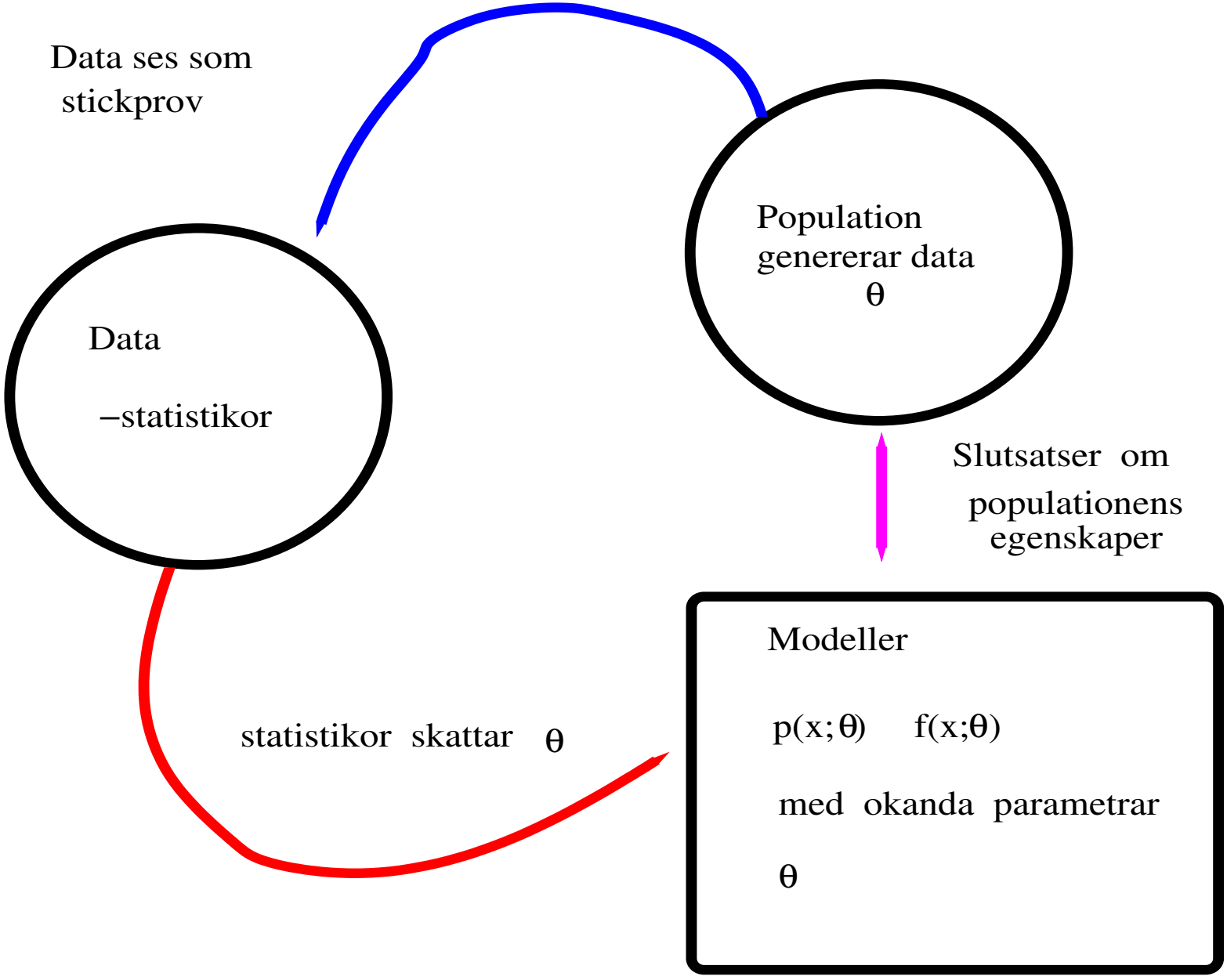
Estimation of a population proportion

Bias & standard error

Maximum likelihood (ML) estimate of p in $\mathcal{B}in(n, p)$.

Confidence interval (CI) for the ML -estimate of a population proportion.

- confidence degree, critical value, standard error



Population

Populations can be **existing** or **conceptual**.

Existing populations are well defined and could be identified (measured) directly.

Conceptual populations are non-existing, yet visualized or imaginable sets of measurements.

- ▶

Population, data, model: an example

Wolf, Yuri I and Grishin, Nick V and Koonin, Eugene V: *Estimating the number of protein folds and families from complete genome data*, **Journal of molecular biology**, 299, 4, pp. 897−905, 2000:

- There exists a **universal population of M protein (domain) families and N folds**. Each family belongs to exactly one fold; each fold includes at least one family.
- Both the database of protein structures and a genome are generated by a **random, independent sampling of families from the fold/family population**. In the sampling process, each of the M families has an equal probability of being drawn from universal population into the sample.
- In the fold/family population the distribution of the number of protein families in a fold is best approximated by a **logarithmic series distribution**.

Statistical model-based data-analysis

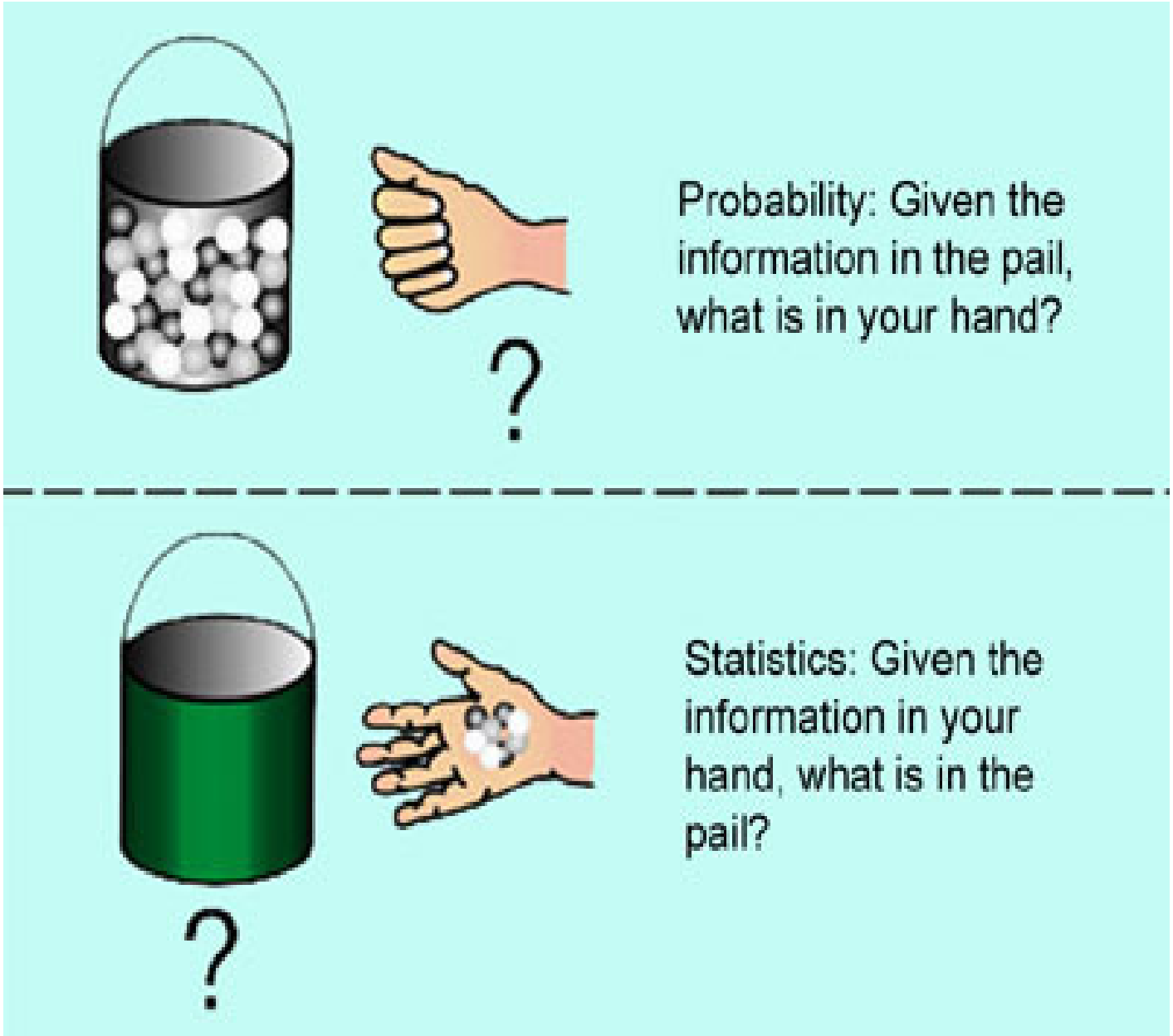
Statistical model-based data-analysis (a.k.a statistical inference) is the process of drawing conclusions about a population from data that is subject to random variation, for example, observational errors or sampling variation.

Statistical model-based data-analysis

For the most part, statistical model-based data-analysis a.k.a 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**, 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**
  - **testing a hypothesis about a population** by a **statistical test**

Probability and Statistics



How to estimate a population proportion ?

When Gregor Mendel conducted his famous genetics experiments with peas, one sample of offspring was obtained by crossing peas with green pods (=ärtbalja) 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 all similar experiments**. We call this proportion the **population proportion/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 VETENSKAP OCH KONST
KTH Matematik

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.

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?

Notations for Proportions

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

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

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

p = proportion in the entire population will be taken as an unknown parameter in a statistical model.

Statistical Model

Let us consider the following:

1. We have by crossed peas with green pods) and peas with yellow pods with the of 580 peas. We take the number like having been fixed in advance, at any rate it is not influenced by the number of yellow pods.
2. Each crossing/indidual experiment has one of two outcomes: yellow pod or green pod.
3. The probability of success = probability of yellow pod, is the unknown population proportion of yellow peas, and this proportion does not change from one crossing event to another, or we cannot think of any comparatively simple factor that could rapidly change this population proportion.
4. *We are interested in X = the number of yellow pods.*

These conditions imply or suggest the statistical model

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

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

Statistical Model !

In words, we have now made the proportion in the entire population to a parameter of our statistical model $X \sim \text{Bin}(580, p)$.

Then

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

another random variable, the estimator of p.

Mathemtical Properties of the Statistical Model: bias

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

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

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

$$E(\widehat{p}) = p$$

We say that $\widehat{p}$ is unbiased (väntevärdesriktig).

Statistical Model: bias

$$Var\left(\widehat{p}\right) = Var\left(\frac{X}{580}\right) = \frac{1}{580^2} Var\left(X\right)$$

$$= \frac{1}{580^2} \cdot 580 \cdot p(1-p) = \frac{p(1-p)}{580}.$$

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.

Accuracy

We insert the sample proportions of successes and failures

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

$$D\left(\widehat{p}\right) = \sqrt{Var\left(\widehat{p}\right)} = \sqrt{\frac{p(1-p)}{580}}$$

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

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

A General Principle: maximum likelihood

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

Maximum likelihood method for $\mathrm{Bin}(580, p)$

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

$$X \sim \mathrm{Bin}(580, p)$$

We introduce the **likelihood function** for p

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

This is just the propability for $P(X = x)$, if $X \sim \mathrm{Bin}(580, p)$, 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 $x = 152$.

Maximizing likelihood

We can maximize $L(p)$ by maximizing the natural 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

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

Maximum likelihood

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

Confidence Interval

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

Confidence Interval CI (Konfidensinterval KI)

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.*

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 (konfidensnivå).

A **confidence level** *is the probability* $1 - \alpha$ *(often expressed as the equivalent percentage 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** (konfidensgrad) or the **confidence coefficient**

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).

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

$$0.226 < p < 0.298$$

CI: an example

Example

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

$$0.226 < p < 0.298$$

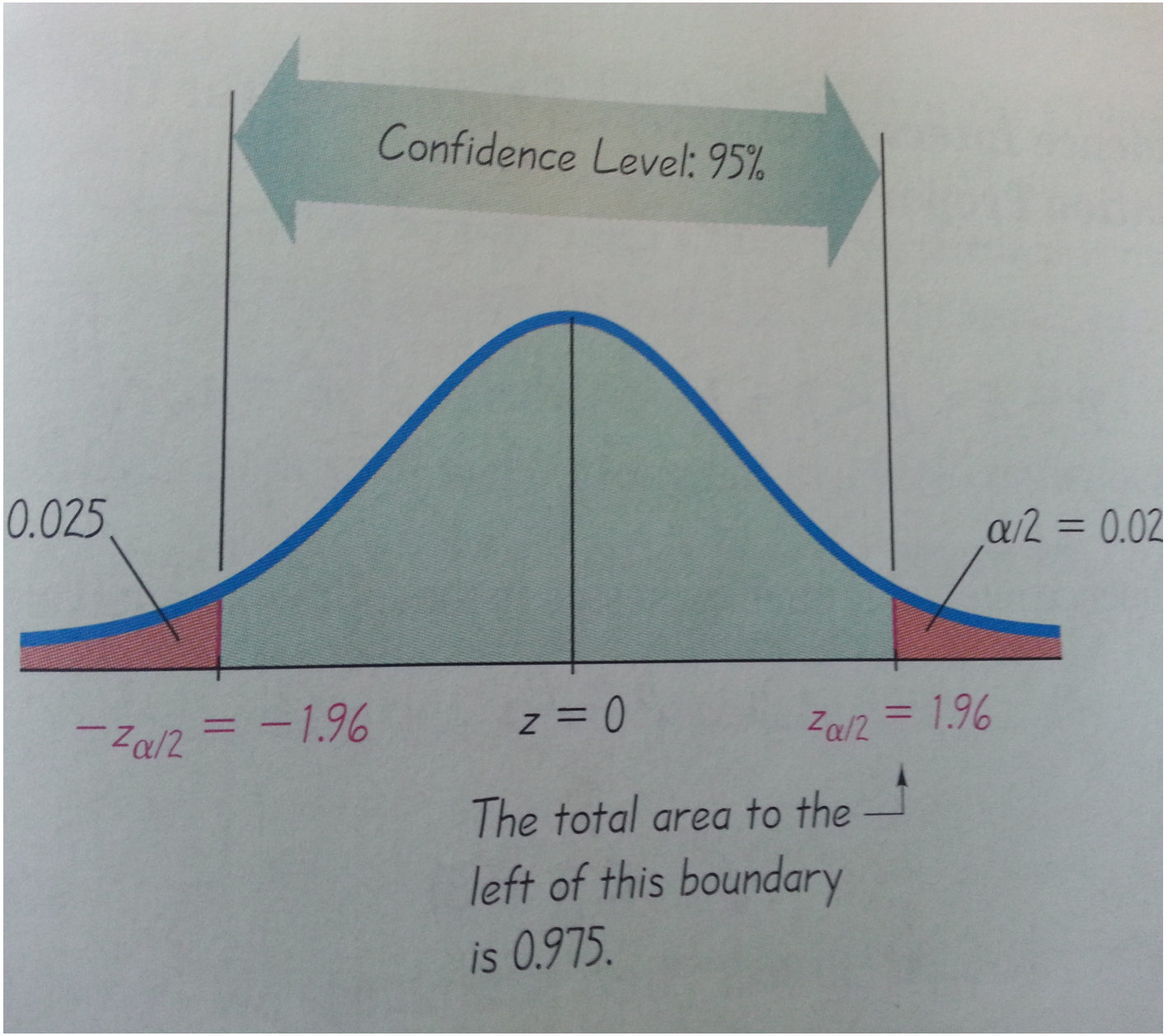
How was this done ?? We shall now go through the whole procedure in several steps. Step one is the discussion of the critical value.

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 $\lambda_{\alpha/2}$ 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 $-\lambda_{\alpha/2}$ is at the vertical boundary separating an area of $\alpha/2$ in the left tail of the standard normal distribution.*

Step One: Critical value Read $z_{\alpha/2}$ as $\lambda_{\alpha/2}$



Step One: Critical value

Since $\lambda_{\alpha/2}$ 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

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

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).

Step One: Critical values

Critical values

$P(Z > \lambda_\alpha) = \alpha$ där $Z \sim N(0,1)$

α	λ_α	α	λ_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

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 $\lambda_{0.025} = 1.96$. This completes Step One.

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* $D\left(\widehat{p}\right)$.

Step Two: Standard Error and the Margin of Error

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

Definition

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

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

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*

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

Step Three: the confidence interval

Confidence Interval or the Interval Estimate for the population proportion p

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

Other equivalent expressions are

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

or

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

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

KTH VETENSKAP OCH KONST
KTH Matematik

The confidence interval for the proportion of yellow pods

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

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

$$z_{0.025} = 1.96$$

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

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

$$0.226 < p < 0.298$$

The statistical statement

The CI is

$$0.226 < p < 0.298$$

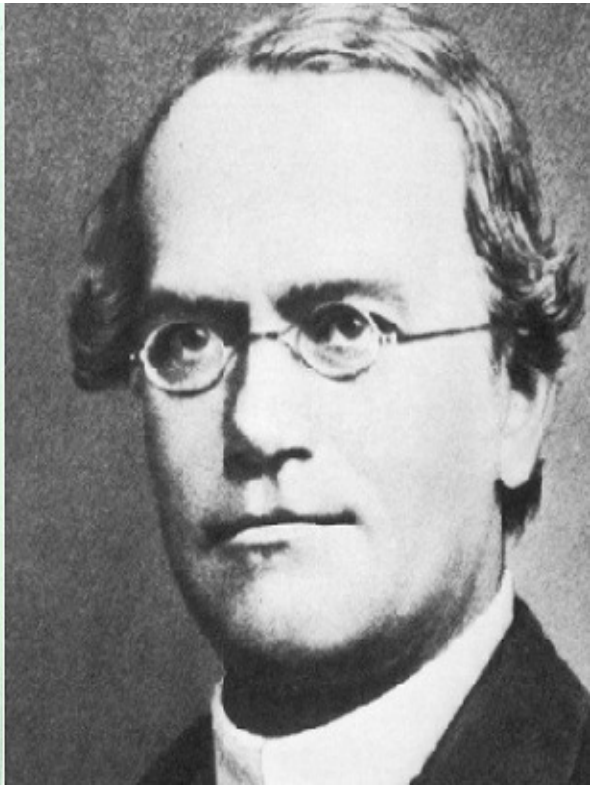
This CI is often reported with a statement such as this:

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.

" ... statistiska felmarginalen "

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.



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.

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 %.

Postscript: A Mathematical Aside

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

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

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

We form the population Z score

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

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

$$P\left(-\lambda_{\alpha/2} \leq \frac{\widehat{p}_{MLE} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}} \leq \lambda_{\alpha/2}\right) \approx 1 - \alpha$$

Postscript: A Mathematical Aside

Then manipulation with inequalities gives

$$P\left(-\lambda_{\alpha/2} \leq \frac{\widehat{p}_{MLE} - p}{\sqrt{\frac{p\cdot(1-p)}{n}}} \leq \lambda_{\alpha/2}\right) =$$

$$P\left(-\lambda_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} \leq \widehat{p}_{MLE} - p \leq \lambda_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}\right) =$$

$$= P\left(-\widehat{p}_{MLE} - \lambda_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} \leq -p \leq -\widehat{p}_{MLE} + \lambda_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}\right)$$

$$= P\left(\widehat{p}_{MLE} - \lambda_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} \leq p \leq \widehat{p}_{MLE} + \lambda_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}\right)$$

Postscript: A Mathematical Aside

We have thus

$$P\left(\widehat{p}_{MLE} - \lambda_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} \leq p \leq \widehat{p}_{MLE} + \lambda_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}\right) \approx 1 -$$

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

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

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