Statistics for EES
Parameter estimation — Frequentistic and Bayesian approaches

Dirk Metzler

July 24, 2017

1 Confidence intervals for expectation values

1.1 Example: Carapace length of the black squat lobster

Example: black squat lobster

Galathea squamifera

Carapace length:

What is the mean Carapace length of female black squat lobsters?

Estimate mean carapace length from sample

How precise is this estimation?

Aim: find a *confidence interval*, i.e. an interval that has a high probability to contain the true mean carapace length.

Galathea: Carapace lengths in a sample

females: $\overline{x} = 3.23$ mm $sd(x) = 0.9$ mm $n = 29$ $sem(x) = \frac{sd(x)}{\sqrt{n}} = \frac{0.9}{\sqrt{29}} = 0.17$ $\qquad (= sd(\overline{x}))$

We know the following rules of thumb:

- 2/3 rule: the interval
$$\left[\overline{x} - sem(x), \overline{x} + sem(x)\right]$$
contains the true mean value with a probability of $\approx 2/3$.

- 95% rule of thumb: The interval
$$\left[\overline{x} - 2 * sem(x), \overline{x} + 2 * sem(x)\right]$$
(more precisely
$$\left[\overline{x} - 1.96 * sem(x), \overline{x} + 1.96 * sem(x)\right])$$
has a probability of $\approx 95\%$ to contain the true mean.

More precisely: let $t_{0.025}$ `<- qt(0.025,length(x)-1)` be the 2.5% quantile of the t distribution with $n-1$ degrees of freedom. Then, the probability of the interval

$$\left[\overline{x} - t_{0.025} * sem(x), \overline{x} + t_{0.025} * sem(x)\right]$$

to contain the true mean is 95%.

With the values $\overline{x} = 3.23$, $t_{0.025} = 2.05$ and $\text{sem}(x) = 0.17$ in

$$\left[\overline{x} - t_{0.025} * sem(x), \overline{x} + t_{0.025} * sem(x)\right]$$

we obtain the 95% confidence interval

$$\left[2.88, 3.58\right]$$

for the true mean, i.e. the error probability is 5%.

Remark

For large sample sizes n or, equivalently, many degrees of freedom $n-1$, we can approximate the t distribution by the normal distribution and use

$$\left[\overline{x} - 1.96 \cdot sem(x), \overline{x} + 1.96 \cdot sem(x)\right]$$

as an approximate 95% confidence interval.

1.2 Theory

Confidence interval for the true mean

Aim: Find a confidence interval for the true mean with error risk α

The confidence interval for the true mean value μ with confidence level q is an interval

$$\big[A(x), B(x)\big]$$

that is estimated from the data $x = (x_1, \ldots, x_n)$ and has the following property.

$$\Pr\big(\mu \in \big[A(x), B(x)\big]\big) \geq q$$

Among all valid confidence intervals we prefer those that tend to be as small as possible.

confidence interval for the true mean

Solution: We already know that the t statistic

$$t := \frac{\overline{x} - \mu}{\mathrm{sem}(x)}$$

is approximately t distributed with $n-1$ degrees of freedom (if n is not too small).

Let t_α `<- -qt(`α`,length(x)-1)` be the α-Quantile of Student's t distribution with n degrees of freedom. Then,

$$\big[\overline{x} - t_\alpha * sem(x), \overline{x} + t_\alpha * sem(x)\big]$$

is a confidence interval with confidence level $1 - 2 \cdot \alpha$.

Substantiation:

$$\begin{aligned}
&\Pr\big(\mu \in \big[\overline{x} - t_\alpha * \mathrm{sem}(x), \overline{x} + t_\alpha * \mathrm{sem}(x)\big]\big) \\
&\quad = \Pr\big(\mu - \overline{x} \in \big[-t_\alpha * \mathrm{sem}(x), t_\alpha * \mathrm{sem}(x)\big]\big) \\
&\quad = \Pr\big(\frac{\mu - \overline{x}}{\mathrm{sem}(x)} \in \big[-t_\alpha, t_\alpha\big]\big) \\
&\quad = \Pr\big(\big|\frac{\mu - \overline{x}}{\mathrm{sem}(x)}\big| \leq t_\alpha\big) \\
&\quad = \Pr\big(|t| \leq t_\alpha\big) \\
&\quad = 1 - 2 \cdot \alpha
\end{aligned}$$

t_α is chosen such that the last equation is fulfilled.

confidence intervals in general

Let θ be a parameter of the underlying distribution.

The confidence interval for the parameter θ with confidence level q (or e.g. "95% confidence interval" if $q = 0.95$) is an interval

$$\big[A(x), B(x)\big]$$

that is estimated from the data $x = (x_1, \ldots, x_n)$ and fulfills

$$\Pr\left(\theta \in \big[A(x), B(x)\big]\right) \geq q$$

2 Confidence intervals for proportions

2.1 Example: sex ratio in porcelain crabs

Family: *Porcellanidae*

23 females and 30 males were caught on 21.Feb.1992 in the Helgoländer Tiefe Rinne (*Pisidiae longicornis*), i.e. the proportion of males in the sample was $30/53 = 0.57$.

What does this tell about the proportion of males in the entire sample?

Wanted: 95% confidence interval for the proportion of males in the population (0.57±??).

2.2 Theory

We observe X males in a sample of size n and aim to estimate the proportion p of males in the entire population.

An obvious estimator is the relative frequency $\widehat{p} := \frac{X}{n}$ in the sample.

How reliable is this estimation?

Find an interval $[\widehat{p}_\ell, \widehat{p}_u]$ that depends on the data and has the property

$$\Pr_p\Big([\widehat{p}_\ell, \widehat{p}_u] \text{ covers } p\Big) \geq q$$

for *any choice* of p. We prefer methods that give us short intervals fulfilling these requirements.

General solution:

For a binomially distributed number K with known total number n we want a 95% confidence interval for the proportion parameter p p. We observe a value of k for K.[0.5cm]

Remember:

$$\begin{aligned} \Pr(K = k) &= \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \\ \mathbb{E}K &= n \cdot p \\ \mathsf{var}(k) &= n \cdot p \cdot (1-p) \end{aligned}$$

A simple solution is Wald's confidence interval:

$$\left[\widehat{p} - 1.96 \cdot \sqrt{\widehat{p} \cdot (1-\widehat{p})/n}\ ,\ \widehat{p} + 1.96 \cdot \sqrt{\widehat{p} \cdot (1-\widehat{p})/n}\right]$$

Wald's confidence interval is based on the following considerations:

$$\begin{aligned} \mathsf{var}(\widehat{p}) &= \mathsf{var}(K/n) = \mathsf{var}(K)/n^2 \\ &= n \cdot p \cdot (1-p)/n^2 \approx \widehat{p} \cdot (1-\widehat{p})/n \end{aligned}$$

We approximate the distribution of $\widehat{p}$ by the normal distribution with mean $\mu = p$ and variance $\sigma^2 = \widehat{p} \cdot (1-\widehat{p})/n$.

The difference between the value a normally distributed random variable and its mean is smaller than $1.96 \cdot \sigma$ in 95% of the cases.

2.3 Example: Mallards

image (c) Andreas Trepte

Anas platyrhynchos Mallard (in German: Stockente)

Foxes hunt Mallards. Male mallards have noticeable colors and are thus easier to descry. Does this bias the sex ratio in mallards?

Data: Sample of size $n = 2200$. Relative frequency of males: 0.564.

References

[Smi68] Johnson, Sargeant (1977) Impact of red fox predation on the sex ratio of prairie mallards *United States fish & wild life service*

- The normal-distribution approximation used in the Wald confidence interval is only valid if n is large and p is neither close to 0 nor to 1. A rule of thumb is that the variance $n \cdot p \cdot (1-p)$ should be ≥ 9.
- The idea of confidence intervals comes from the *frequentistic* approach of statistics. If we repeat an experiment many times and compute the q-confidence intervals for each repetition, approx. $q \cdot 100\%$ of the confidence intervals will contain the true value. This must be true, no matter what the true parameter values are.

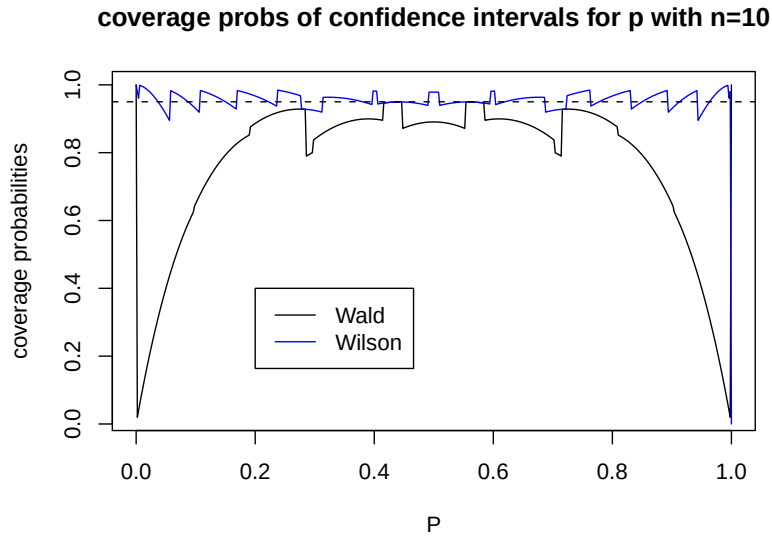
There are also other methods for getting confidence intervals for the proportion variable p of a binomially distributed random variable. Some are available with the R command binconf in the R package Hmisc and with the R command binom.confint from the package binom.

An example is Wilson's method, the default of the R command binconf. We do not treat the theoretical backgrounds here but compare its results to the Wald method.

Remember: Confidence intervals are random because they depend on the data.

Theoretically, a method for generating 95% confidence intervals should have a probability of approximately 95% or slightly more to output an interval that covers (i.e. includes) the true value. This should be true for all possible true parameter values.

For the case that a proportion is to be estimated we can compute this *coverage probability*. We do this for $n = 10$, and $n = 100$, combined with values of p between 0 and 1.



We see that the coverage probabilities of the Wald confidence interval is too low if p is

close to 0 or 1.

Reason: If $p = 0.1$, then $K = 0$ is quite probable. In this case we estimate $\widehat{p} = K/n = 0/n = 0$ and $\mathsf{var}(\widehat{p}) \approx \widehat{p} \cdot (1 - \widehat{p})/n = 0$. This leads to a Wald confidence interval of $[0, 0]$ that does not contain the true value.

A simple trick to solve this problem is to compute the confidence interval as if the $K + 1$ was observed instead of K (to avoid $\widehat{p} = 0$ in the case of $K = 0$) and as if the total number was $n + 2$ instead of n (to avoid $\widehat{p} = 1$ in the case of $K = n$).

The "k+1, n+2" trick

cf p. 121 in

References

[KW08] G"otz Kersting, Anton Wakolbinger (2008) *Elementare Stochastik*, Birkh"auser, Basel.

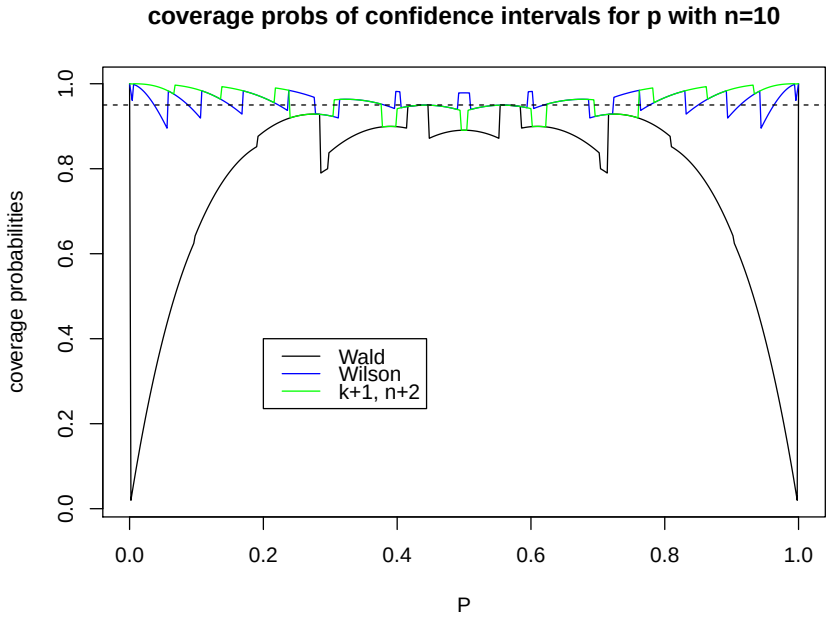
If k successes in n trials are observed, estimate the success probability by

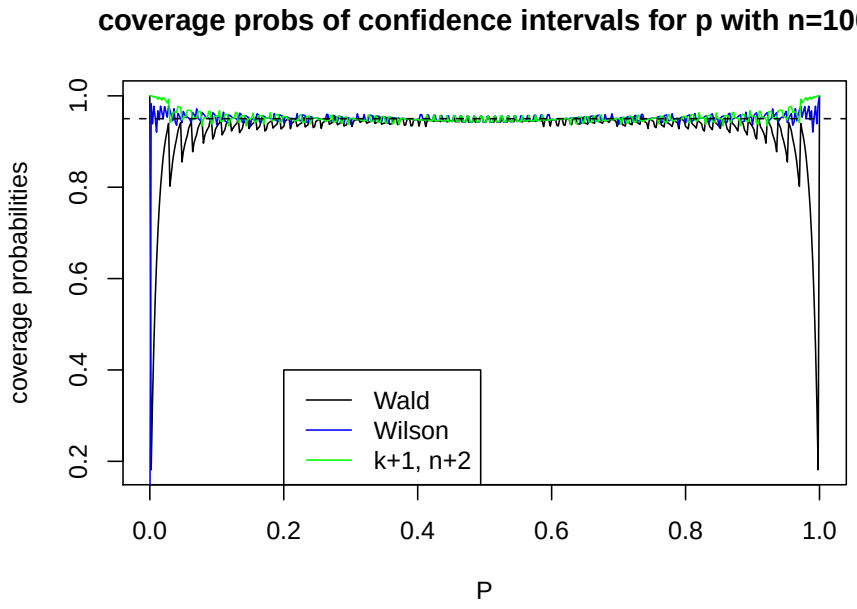
$$\widetilde{p} = (k + 1)/(n + 2)$$

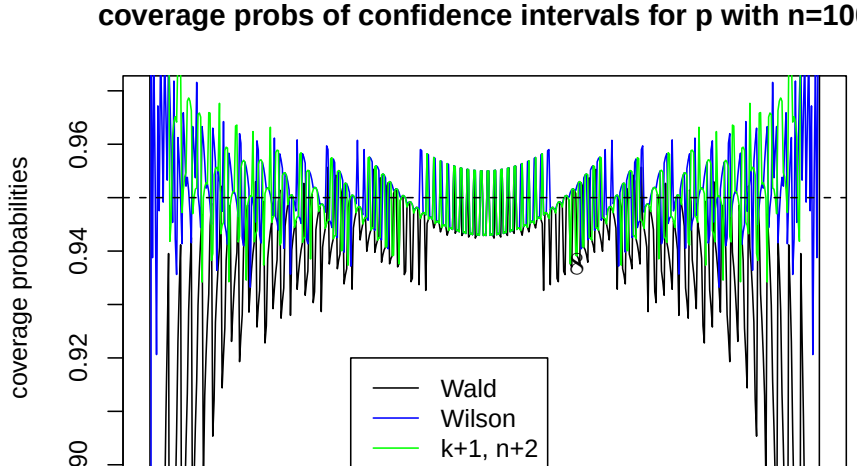
we use this $\widetilde{p}$ instead of $\hat{p}$ to compute the Wald confidence interval

$$\left[\widetilde{p} - 1.96 \cdot \sqrt{\widetilde{p} \cdot (1 - \widetilde{p})/n}\ ,\ \widetilde{p} + 1.96 \cdot \sqrt{\widetilde{p} \cdot (1 - \widetilde{p})/n}\right]$$

This works astonishingly well, even if the true p is close to 0 or 1.

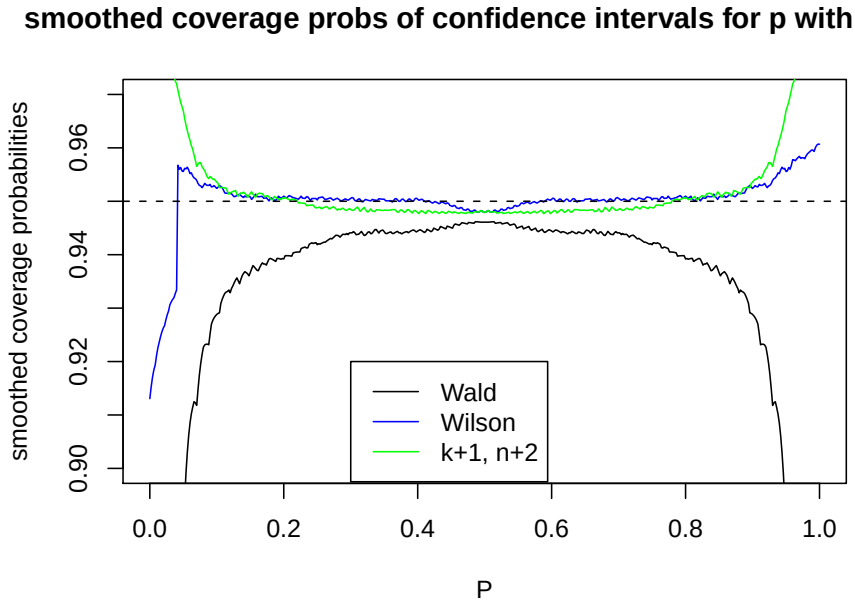






there are some values of p for which the coverage probability is smaller than 95%. Slightly different true values of p may lead to coverage probabilities higher than 95%. [1cm]

To obtain a clearer impression we smooth the curves by taking means over small ranges of p.



Both for $n = 10$ and for $n = 100$, the Wilson method and the "k+1, n+2"-Wald method give more reliable confidence intervals than the simple Wald method, especially for p close to 0 or 1.

We will revisit the "k+1, n+2" trick in the context of Bayesian statistics.

3 Frequentistic Statistics

3.1 Foundations of frequentistic statistics

Principles of frequentistic statistics

- Parameters are unknown but not random.
- Data depend on parameter values and on random (according to model assumptions).
- frequentistic interpretation of probability: If an event has a probability of p, this means that on the long run it will take place in a proportion of p of all cases (assuming independent repetitions).
- If we perform a test with significance level α, on the long run we falsely reject the null-hypotheses in a proportion of α of the cases where the null-hypothesis is actually fulfilled.

- On the long run, 95% of my 95% confidence intervals will contain the true value.

3.2 Duality of tests and confidence intervals

```
> X
 [1] 4.111007 5.023229 5.489230 4.456054 4.343212
 [5] 5.431928 3.944405 3.471677 4.337888 5.412292
> n <- length(X)
> m <- mean(X)
> sem <- sd(X)/sqrt(n)
> t <- -qt(0.025,n-1)
> konf <- c(m-t*sem,m+t*sem)
> konf
[1] 4.100824 5.103360
```

$[4.100824, 5.103360]$

```
> t.test(X,mu=4)

One Sample t-test

data:  X
t = 2.7172, df = 9, p-value = 0.02372
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 4.100824 5.103360
sample estimates:
mean of x
 4.602092
```

Notice: R t-tests output confidence intervals!

$[4.100824, 5.103360]$

```
> t.test(X,mu=4.1)

One Sample t-test

data:  X
t = 2.2659, df = 9, p-value = 0.0497
alternative hypothesis: true mean is not equal to 4.1
95 percent confidence interval:
 4.100824 5.103360
sample estimates:
mean of x
 4.602092
```

Notice: R t-tests output confidence intervals!
$[4.100824, 5.103360]$

```
> t.test(X,mu=4.1009)

One Sample t-test

data:  X
t = 2.2618, df = 9, p-value = 0.05003
alternative hypothesis: true mean is not equal to 4.1009
95 percent confidence interval:
 4.100824 5.103360
sample estimates:
mean of x
 4.602092
```

Notice: R t-tests output confidence intervals!
$[4.100824, 5.103360]$

```
> t.test(X,mu=5.1)

One Sample t-test

data:  X
t = -2.247, df = 9, p-value = 0.05125
alternative hypothesis: true mean is not equal to 5.1
95 percent confidence interval:
 4.100824 5.103360
sample estimates:
mean of x
 4.602092
```

Notice: R t-tests output confidence intervals!
$[4.100824, 5.103360]$

```
> t.test(X,mu=5.1034)

One Sample t-test

data:  X
t = -2.2623, df = 9, p-value = 0.04999
alternative hypothesis: true mean is not equal to 5.1034
95 percent confidence interval:
 4.100824 5.103360
sample estimates:
```

```
mean of x
 4.602092
```

Notice: R t-tests output confidence intervals!

Duality Tests $\leftrightarrow$ Confidence Intervals

If $[a, b]$ is a $(1-\alpha)$-confidence interval for a parameter θ, there is a corresponding test with confidence interval α that rejects $\theta = x$ if and only if $x \notin [a, b]$.[0.5cm]

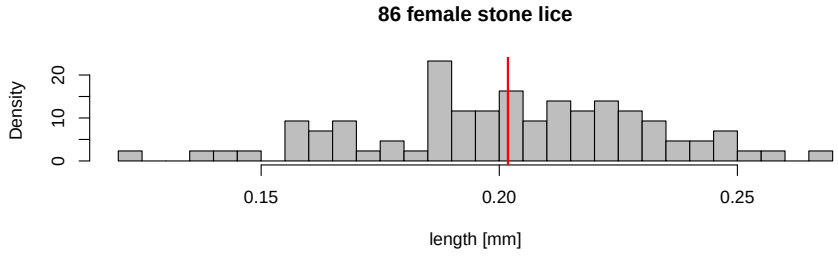
If T_x is a one- or two-sided test with null-hypothesis $\theta = x$ and significance level α, the set of all values x, for which the test would *not* reject the null hypothesis $\theta = x$ form a $(1-\alpha)$ confidence interval for θ.

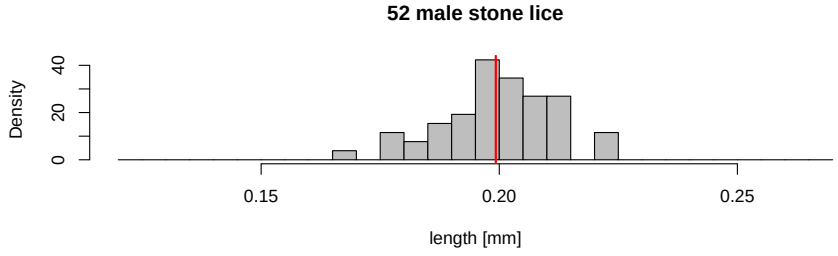
Confidence ranges are especially helpful if the test does *not* indicate significance.

Example: Is there a sex-specific differentiation for body length in stone lice *Petrophaga lorioti* (in German: Steinlaus)?

Data: Lengths of 86 female (F) and 52 male (M) Stone lice.

Small sample size because the stone louse is an endangered species! http://www.youtube.com/watch?v= http://en.wikipedia.org/wiki/Stone_louse






```
> t.test(F,M)

Welch Two Sample t-test

data:  F and M
t = 0.7173, df = 122.625, p-value = 0.4746
alternative hypothesis: true difference in means is
```

```
                                        not equal to 0
95 percent confidence interval:
 -0.004477856  0.009567353
sample estimates:
mean of x mean of y
0.2018155 0.1992707
```

How should we report the result of this test?

- There is no difference in length between female and male stone lice. ~~There is no difference in length between female and male stone lice.~~
- On average, male and female stone lice have the same length.~~On average, male and female stone lice have the same length.~~
- The data do not show a significant difference in length between male and female stone lice. The data do not show a significant difference in length between male and female stone lice.
- A 95% confidence range of the difference in length between male and female stone lice is [-0.0045,0.0096] A 95% confidence range of the difference in length between male and female stone lice is [-0.0045,0.0096]✓

3.3 Maximum-Likelihood (ML) Estimator

- Even if it is preferable to give confidence intervals for estimated parameters, there is sometimes the desire to output just one estimation value. The preferred method in frequentistic statistic to estimate a parameter is *Maximum-Likelihood (ML)* estimation.
- It does not make sense to ask for the "most probable" value of a parameter because (from the perspective of frequentistic statistics) parameters are not random and thus do not have a probability.
- Instead, we search for the parameter value for which the data have the highest probability. The *Likelihood* $L_D(x)$ of a value x for a parameter θ is the probability $\Pr_x(D)$ of the observed data, assuming that $\theta = x$:

$$L_D(x) := \Pr_{\theta=x}(D)$$

- The *Likelihood* of a value x for a parameter θ is the probability of the observed data D, assuming $\theta = x$:

$$L_D(x) := \Pr_{\theta=x}(D)$$

- The *Maximum-Likelihood Estimator* (ML estimator) assigns to each dataset D the parameter value $\widehat{\theta}$ that maximizes the likelihood function L_D:

$$\widehat{\theta} = \arg\max_x L_D(x)$$

Example: A DNA strand of length 100 bp shows 7 differences between human and chimpanzee. What is the probability p to observe a difference in the neighboring position 101?

Obvious estimator $\tilde{p} = 7/100$

ML estimator: Model the number K of the mutations as binomially distributed with $n = 100$ and unknown p. It follows

$$L(p) = \Pr_p(K = 7) = \binom{100}{7} p^7 \cdot (1-p)^{93}$$

and

$$\begin{aligned} \widehat{p} &= \arg\max_p \binom{100}{7} p^7 \cdot (1-p)^{93} = \arg\max_p p^7 \cdot (1-p)^{93} \\ &= \arg\max_p \log\left(p^7 \cdot (1-p)^{93}\right) \end{aligned}$$

Wanted: the p that maximizes

$$f(p) := \log\left(p^7 \cdot (1-p)^{93}\right) = 7 \cdot \log(p) + 93 \cdot \log(1-p).$$

A common approach to find the maximum is to find the root of the derivative

$$0 = f'(p) = 7 \cdot \frac{1}{p} + 93 \frac{1}{1-p} \cdot (-1)$$

(remember that $\log'(x) = 1/x$.) Solving the equation for p, we get:

$$\widehat{p} = 7/100$$

Thus, we have found a theoretic reasoning for the obvious estimator $\tilde{p}$.

In many situations, the ML estimator is *consistent*, i.e. if sufficiently many independent data are available and the model assumptions are fulfilled, the ML estimator will find the true value (or approximate it with arbitrary precision).

For small data sets, other estimators may be preferable.

Example: If $X_1, \ldots, X_n$ is a sample from a normal distribution, then $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ is the ML estimator for the variance σ^2. Usually, the bias-corrected estimator $\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is preferred.

4 Bayesian Statistics

Principles of Bayesian Statistics

- Also parameter are considered to be random.

- The *priori (probability) distribution* of a parameter reflects how probable the different possible values are assumed to be before looking at the data.
- With the Bayes-Formula we obtain the *posterior (probability) distribution*, the conditional probability distribution of θ given the data D.

$$\Pr(\theta_0|D) \;=\; \frac{\Pr(D|\theta_0)\cdot\Pr(\theta_0)}{\Pr(D)} \;=\; \frac{\Pr(D|\theta_0)\cdot\Pr(\theta_0)}{\sum_\theta \Pr(D|\theta)\Pr(\theta)}$$

  This only works if the prior probabilities $\Pr(\theta)$ are defined. $\Pr(D|\theta_0)$ is just the Likelihood $L_D(\theta)$ from frequentistics statistics. Usually, we have to deal with continuous parameter spaces. Then, we have to replace the prior and posterior probabilities by prior and posterior densities and the sums by integrals.

- If we can compute or simulate posterior probabilities for parameters we can analyze which parameters come into question.
- Instead of the ML estimator, Baysian statistics uses the expectation value of the posterior probability or the parameter value with the highest posterior probability (density). [MAP=maximum a-posteriori].
- The analogs to the confidence intervals of frequentistic statistics in Bayesian statistics are the credibility ranges. A 95% credibility range is a parameter range that contains the parameter value with a probability of 95% according to the posterior distribution.

Example: $n = 20$ **experiments,** $K = 3$ **successes,** $p =?$

K is binomially distributed with $n = 20$. We observe $K = 3$. The ML estimator is $\widehat{p} = 3/20$.

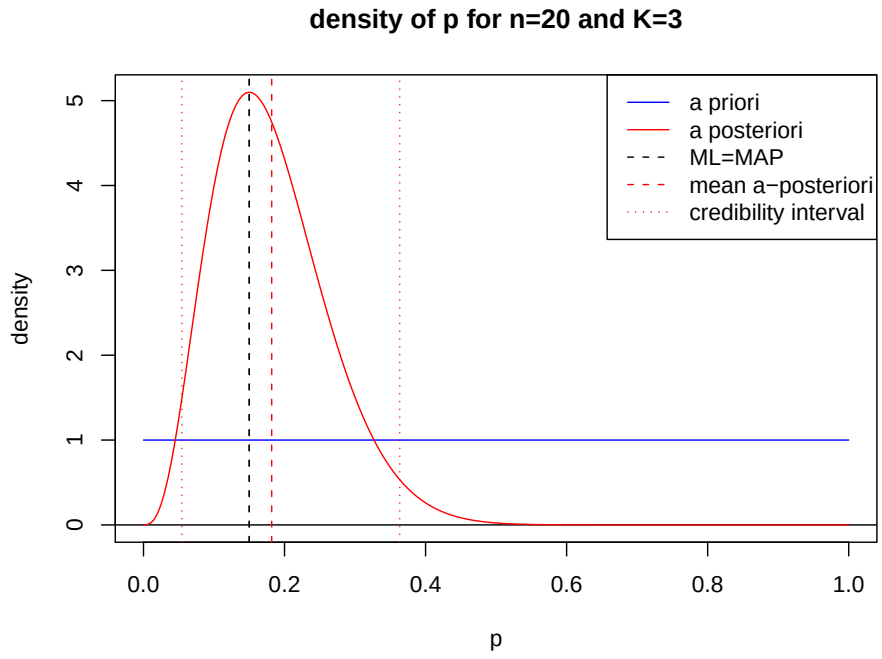
What is the posterior distribution for p?

It is only defined if we first define a prior distribution for p. We just take the uniform distribution on $[0, 1]$ as a prior.

The posterior distribution is then the Beta$(1 + K,1 + n - K)$ distribution, cf. p. 106 in

References

[KW08] G. Kersting, A. Wakolbinger (2008) *Elementare Stochastik*, Birkh"auser, Basel.



- The ML estimator and the MAP estimator coincide because we used a uniform prior distribution.
- The expectation value of the posterior distribution Beta$(1+K, 1+n-K)$ is

$$\mathbb{E}(p|K) = \frac{K+1}{n+2}.$$

  We have already seen this estimator from the "$k+1$, $n+2$" trick as $\widetilde{p}$. Thus, we have a Bayesian justification for this estimator.

- Interval estimators:

Wald confidence interval:	[0, 0.306]
"$k+1$, $n+1$" confidence interval:	[0.013, 0.351]
Wilson confidence interval:	[0.052, 0.360]
credibility range:	[0.054, 0.363]

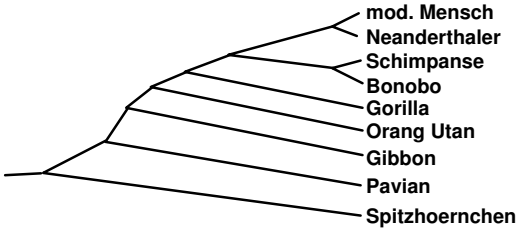
Frequentists vs. Bayesians

- For quite some time frequentists and Bayesians fought about the "right" interpretation of statistics.
- The strongest point of criticism on Bayesian Methods: The choice of a prior distribution is subjective.
- Nowadays, most statisticians use frequentistic and Bayesian methods according to the requirements.

- The choice of a prior distribution is still a critical point. Using a uniform distribution does not always solve this problem.

Example: Phylogenetic tree reconstruction

```
Bonobo       ATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGAAGCAAATTTAAGTGCCACCCAAGTATTGGCTCA...
Schimpanse   ATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGAAGCAAATTTAAGTACCACCTAAGTACTGGCTCA...
Gibbon       TATTCTCATGTGGAAGCCATTTTGGGTACAACCCCAGTACTAACCCACTTCTCCACAACTCTATGTACTT...
Gorilla      ATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGAGACAAATTTGGGTACCACCCAAGTATTGGCTAA...
mod. Mensch  ATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCA...
Neanderth    CCAAGTATTGACTCACCCATCAACAACCGCCATGTATTTCGTACATTACTGCCAGCCACCATGAATATTG...
Pavian       TATTTTATGTTGTACAAGCCCCACAGTACAACCTTAGCACTAGCTAACTTTTAATGCCACTATGTAATTC...
Oran Utan    TTCTTTCATGGGGGACCAGATTTGGGTGCCACCCCAGTACTGACCCATTTCTAACGGCCTATGTATTTCG...
Spitzhrn     CGTGCATTAATGCTTTACCACATTAATATATGGTACAGTACATAACTGTATATAAGTACATAGTACATTT...
```



- Parameter values are not always numbers.
- In phylogeny estimation, the tree is the parameter to be estimated.
- ML programs like PHYLIP/dnaml search the ML phylogeny, i.e. the tree for which the sequence data are most probable.
- Bayesian programs MrBayes or BEAST first generate many trees according to the posterior distribution (given the sequence data) and summarize then which assertions (e.g. "human, chimpanzee and bonobo form a monophyletic group") is fulfilled for which proportion of the sampled trees.
- More about this in the next semester.