Inductive Statistics

Inductive Statistics: Main Tasks

- **Parameter Estimation**
  Given an assumption about the type of distribution of the underlying random variable the parameter(s) of the distribution function is estimated.
- **Hypothesis Testing**
  A hypothesis about the data generating process is tested by means of the data.
  - *Parameter Test*
    Test whether a parameter can have certain values.
  - *Goodness-of-Fit Test*
    Test whether a distribution assumption fits the data.
  - *Dependence Test*
    Test whether two attributes are dependent.
- **Model Selection**
  Among different models that can be used to explain the data the best fitting is selected, taking the complexity of the model into account.

Inductive Statistics: Random Samples

- In inductive statistics probability theory is applied to make inferences about the process that generated the data. This presupposes that the sample is the result of a random experiment, a so-called **random sample**.
- The random variable yielding the sample value x_i is denoted X_i.
  x_i is called a **instantiation** of the random variable X_i.
- A random sample $x = (x_1, \ldots, x_n)$ is an instantiation of the **random vector** $X = (X_1, \ldots, X_n)$.
- A random sample is called **independent** if the random variables $X_1, \ldots, X_n$ are (stochastically) independent, i. e. if

$$\forall c_1, \ldots, c_n \in \mathbb{R}: \quad P\left(\bigwedge_{i=1}^{n} X_i \leq c_i\right) = \prod_{i=1}^{n} P(X_i \leq c_i).$$

- An independent random sample is called **simple** if the random variables $X_1, \ldots, X_n$ have the same distribution function.

Parameter Estimation

Given:

- A data set and
- a family of parameterized distributions functions of the same type, e.g.
  - the family of binomial distributions $b_X(x; p, n)$ with the parameters p, $0 \leq p \leq 1$, and $n \in \mathbb{N}$, where n is the sample size,
  - the family of normal distributions $N_X(x; \mu, \sigma^2)$ with the parameters μ (expected value) and σ^2 (variance).

Assumption:

- The process that generated the data can be described well by an element of the given family of distribution functions.

Desired:

- The element of the given family of distribution functions (determined by its parameters) that is the best model for the data.

Parameter Estimation

- Methods that yield an estimate for a parameter are called **estimators**.
- Estimators are **statistics**, i.e. functions of the values in a sample.

  As a consequence they are functions of (instantiations of) random variables and thus (instantiations of) random variables themselves.

  Therefore we can use all of probability theory to analyze estimators.
- There are two types of parameter estimation:
  - **Point Estimators**
    Point estimators determine the best value of a parameter
    w.r.t. the data and certain quality criteria.
  - **Interval Estimators**
    Interval estimators yield a region, a so-called **confidence interval**,
    in which the true value of the parameter lies with high certainty.

Point Estimation

Not all statistics, that is, not all functions of the sample values are reasonable and useful estimator. Desirable properties are:

- **Consistency**
  With growing data volume the estimated value should get closer and closer to the true value, at least with higher and higher probability.
  Formally: If T is an estimator for the parameter θ, it should be

$$\forall \varepsilon > 0 : \quad \lim_{n \to \infty} P(|T - \theta| < \varepsilon) = 1,$$

  where n is the sample size.

- **Unbiasedness**
  An estimator should not tend to over- or underestimate the parameter.
  Rather it should yield, on average, the correct value.
  Formally this means

$$E(T) = \theta.$$

- **Efficiency**
  The estimation should be as precise as possible, that is, the deviation from the true value should be as small as possible. Formally: If T and U are two estimators for the same parameter θ, then T is called *more efficient* than U if

$$D^2(T) < D^2(U).$$

- **Sufficiency**
  An estimator should exploit all information about the parameter contained in the data. More precisely: two samples that yield the same estimate should have the same probability (otherwise there is unused information).
  Formally: an estimator T for a parameter θ is called sufficient iff for all samples $x = (x_1, \ldots, x_n)$ with $T(x) = t$ the expression

$$\frac{f_{X_1}(x_1; \theta) \cdots f_{X_n}(x_n; \theta)}{f_T(t; \theta)}$$

  is independent of θ.

Point Estimation: Example

Given: a family of **uniform distributions** on the interval $[0, \theta]$, i. e.

$$f_X(x; \theta) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \leq x \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Desired: an estimate for the unknown parameter θ.

- We will now consider two estimators for the parameter θ and compare their properties.
  - $T = \max\{X_1, \ldots, X_n\}$
  - $U = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}$
- **General approach:**
  - Find the probability density function of the estimator.
  - Check the desirable properties by exploiting this density function.

Point Estimation: Example

To analyze the estimator $T = \max\{X_1, \ldots, X_n\}$, we compute its density function:

$$\begin{aligned} f_T(t;\theta) &= \frac{\mathrm{d}}{\mathrm{d}t} F_T(t;\theta) = \frac{\mathrm{d}}{\mathrm{d}t} P(T \leq t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} P(\max\{X_1, \ldots, X_n\} \leq t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} P\Big(\textstyle\bigwedge_{i=1}^{n} X_i \leq t\Big) = \frac{\mathrm{d}}{\mathrm{d}t} \displaystyle\prod_{i=1}^{n} P(X_i \leq t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} (F_X(t;\theta))^n = n \cdot (F_X(t;\theta))^{n-1} f_X(t,\theta) \end{aligned}$$

where

$$F_X(x;\theta) = \int_{-\infty}^{x} f_X(x;\theta)\mathrm{d}x = \begin{cases} 0, & \text{if } x \leq 0, \\ \frac{x}{\theta}, & \text{if } 0 \leq x \leq \theta, \\ 1, & \text{if } x \geq \theta. \end{cases}$$

Therefore it is

$$f_T(t;\theta) = \frac{n \cdot t^{n-1}}{\theta^n} \quad \text{for} \quad 0 \leq t \leq \theta, \quad \text{and 0 otherwise.}$$

Point Estimation: Example

- The estimator $T = \max\{X_1, \ldots, X_n\}$ is **consistent**:

$$\begin{aligned}
\lim_{n\to\infty} P(|T-\theta| < \epsilon) &= \lim_{n\to\infty} P(T > \theta - \epsilon) \\
&= \lim_{n\to\infty} \int_{\theta-\epsilon}^{\theta} \frac{n \cdot t^{n-1}}{\theta^n} \mathrm{d}t \;=\; \lim_{n\to\infty} \left[\frac{t^n}{\theta^n}\right]_{\theta-\epsilon}^{\theta} \\
&= \lim_{n\to\infty} \left(\frac{\theta^n}{\theta^n} - \frac{(\theta-\epsilon)^n}{\theta^n}\right) \\
&= \lim_{n\to\infty} \left(1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n\right) \;=\; 1
\end{aligned}$$

- It is **not unbiased**:

$$\begin{aligned}
E(T) &= \int_{-\infty}^{\infty} t \cdot f_T(t;\theta) \mathrm{d}t \;=\; \int_0^{\theta} t \cdot \frac{n \cdot t^{n-1}}{\theta^n} \mathrm{d}t \\
&= \left[\frac{n \cdot t^{n+1}}{(n+1)\theta^n}\right]_0^{\theta} \;=\; \frac{n}{n+1}\theta \;<\; \theta \quad \text{for } n < \infty.
\end{aligned}$$

Point Estimation: Example

- The estimator $U = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}$ has the density function

$$f_U(u;\theta) = \frac{n^{n+1}}{(n+1)^n} \frac{u^{n-1}}{\theta^n} \qquad \text{for } 0 \leq t \leq \frac{n+1}{n}\theta, \text{ and 0 otherwise.}$$

- The estimator U is **consistent** (without formal proof).

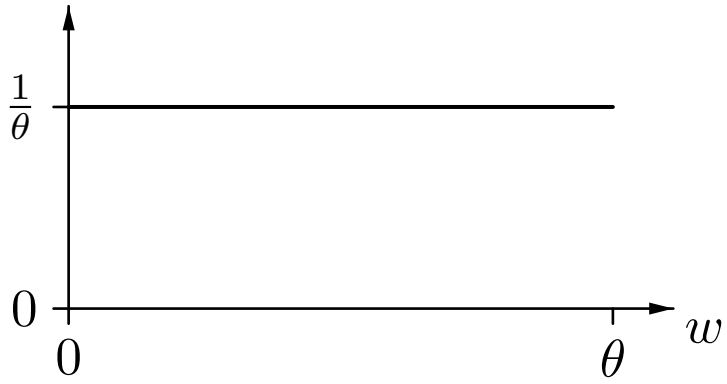
- It is **unbiased**:

$$\begin{aligned}
E(U) &= \int_{-\infty}^{\infty} u \cdot f_U(u;\theta) \mathrm{d}u \\
&= \int_0^{\frac{n+1}{n}\theta} u \cdot \frac{n^{n+1}}{(n+1)^n} \frac{u^{n-1}}{\theta^n} \mathrm{d}u \\
&= \frac{n^{n+1}}{(n+1)^n \theta^n} \left[\frac{u^{n+1}}{n+1} \right]_0^{\frac{n+1}{n}\theta} \\
&= \frac{n^{n+1}}{(n+1)^n \theta^n} \cdot \frac{1}{n+1} \left(\frac{n+1}{n}\theta \right)^{n+1} = \theta
\end{aligned}$$

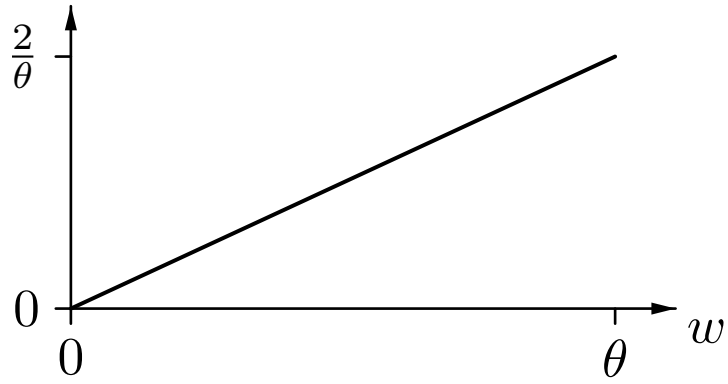
Densities of Estimators

What does the density of the estimator $W = \max\{X_1, \ldots, X_n\}$ look like? (w. r. t. n)

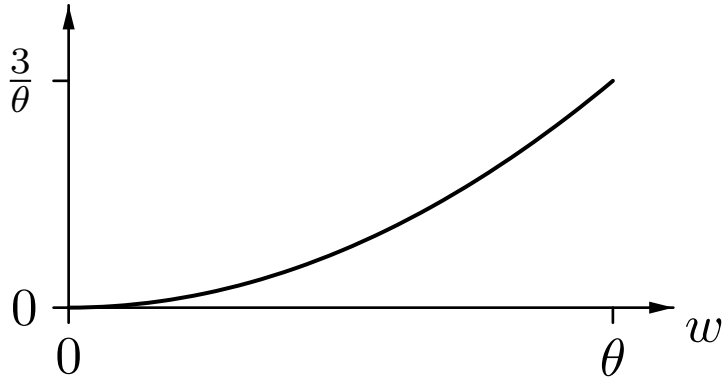
$f_W(w; \theta, 1)$



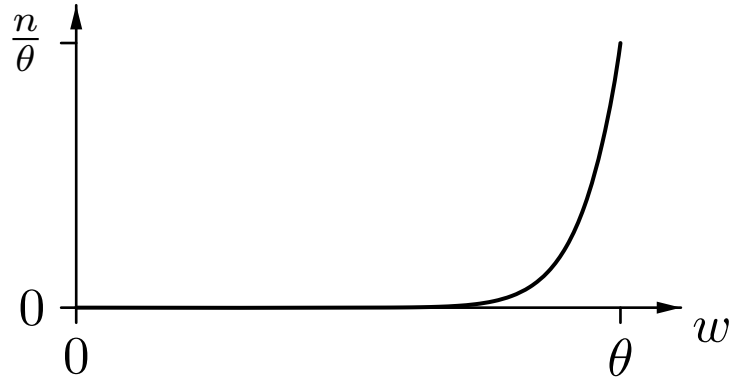
$f_W(w; \theta, 2)$



$f_W(w; \theta, 3)$



$f_W(w; \theta, n)$



Note the different scales for the y-axes!

Point Estimation: Example

Given: a family of **normal distributions** $N_X(x; \mu, \sigma^2)$

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Desired: estimates for the unknown parameters μ and σ^2.

- The median and the arithmetic mean of the sample
  are both consistent and unbiased estimators for the parameter μ.

  The median is less efficient than the arithmetic mean.

- The function $V^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a consistent, but **biased** estimator for the parameter σ^2 (it tends to underestimate the variance).

  The function $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, however, is a consistent and **unbiased** estimator for σ^2 (this explains the definition of the empirical variance).

Point Estimation: Example

Given: a family of **polynomial distributions**
(synonym: multinomial distribution)

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k;\theta_1,\ldots,\theta_k,n) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i},$$

(n is the sample size, the x_i are the frequencies of the different values a_i, $i = 1,\ldots,k$, and the θ_i are the probabilities with which the values a_i occur.)

Desired: estimates for the unknown parameters $\theta_1,\ldots,\theta_k$

- The relative frequencies $R_i = \frac{X_i}{n}$ of the different values a_i, $i = 1,\ldots,k$, are
  - consistent,
  - unbiased,
  - most efficient, and
  - sufficient estimators for the θ_i.

Polynomial Distribution: Example

- Consider the random experiment of picking a person out of a population (with replacement, technically) and determine her eye color.
- $k = 3$: $a_1 \mathrel{\widehat{=}} \mathsf{blue}$, $a_2 \mathrel{\widehat{=}} \mathsf{green}$, $a_3 \mathrel{\widehat{=}} \mathsf{brown}$
- $\theta_1 = 0.3, \quad \theta_2 = 0.3, \quad \theta_3 = 0.4$

The probability of finding 2 persons with blue eyes, 4 persons with green eyes and 4 persons with brown eyes (in a sample of size 10) is:

$$f_{X_1,X_2,X_3}(2, 4, 4;\, \theta_1, \theta_2, \theta_3, 10) = \frac{10!}{2!4!4!} \cdot 0.3^2 \cdot 0.3^4 \cdot 0.4^4 \approx 0.0588$$

Note:

- $\sum_{i=1}^{k} x_i = n \quad \text{and} \quad \sum_{i=1}^{k} \theta_i = 1$

How Can We Find Estimators?

- Up to now we analyzed given estimators,
  now we consider the question how to find them.

- There are three main approaches to find estimators:

  - **Method of Moments**
    Derive an estimator for a parameter from the moments of a distribution and its generator function.
    (We do not consider this method here.)

  - **Maximum Likelihood Estimation**
    Choose the (set of) parameter value(s) that makes the sample most likely.

  - **Maximum A-posteriori Estimation**
    Choose a prior distribution on the range of parameter values, apply Bayes' rule to compute the posterior probability from the sample, and choose the (set of) parameter value(s) that maximizes this probability.

Maximum Likelihood Estimation

- General idea: **Choose the (set of) parameter value(s) that makes the sample most likely.**
- If the parameter value(s) were known, it would be possible to compute the probability of the sample. With unknown parameter value(s), however, it is still possible to state this probability as a function of the parameter(s).
- Formally this can be described as choosing the value θ that maximizes

$$L(D; \theta) = f(D \mid \theta),$$

  where D are the sample data and L is called the **Likelihood Function**.
- Technically the estimator is determined by
  - setting up the likelihood function,
  - forming its partial derivative(s) w.r.t. the parameter(s), and
  - setting these derivatives equal to zero (necessary condition for a maximum).

Brief Excursion: Function Optimization

Task: Find values $\vec{x} = (x_1, \ldots, x_m)$ such that $f(\vec{x}) = f(x_1, \ldots, x_m)$ is optimal.

Often feasible approach:

- A necessary condition for a (local) optimum (maximum or minimum) is that the partial derivatives w.r.t. the parameters vanish (Pierre Fermat).
- Therefore: (Try to) solve the equation system that results from setting all partial derivatives w.r.t. the parameters equal to zero.

Example task: Minimize $f(x, y) = x^2 + y^2 + xy - 4x - 5y$.

Solution procedure:

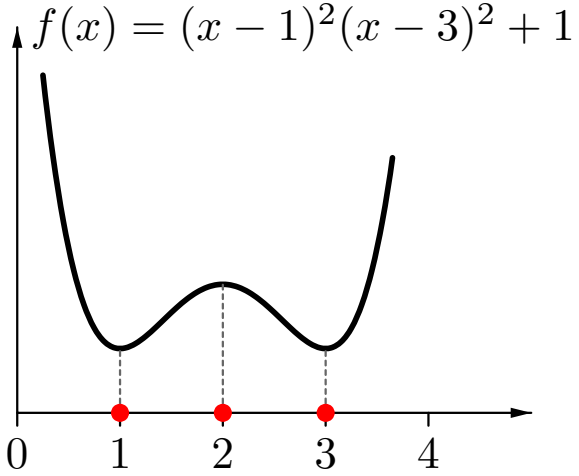
1. Take the partial derivatives of the objective function and set them to zero:

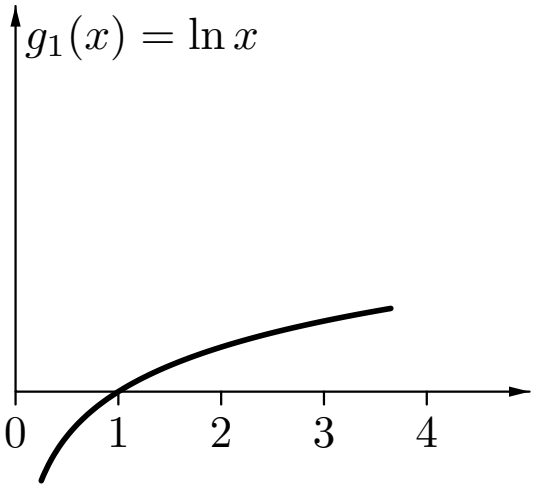
$$\frac{\partial f}{\partial x} = 2x + y - 4 = 0, \qquad \frac{\partial f}{\partial y} = 2y + x - 5 = 0.$$

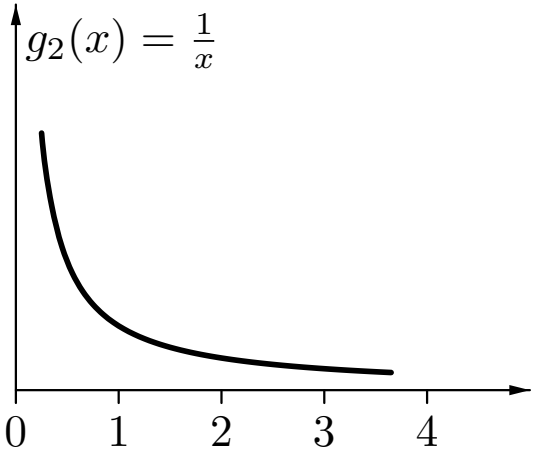
2. Solve the resulting (here: linear) equation system: $x = 1, \quad y = 2.$

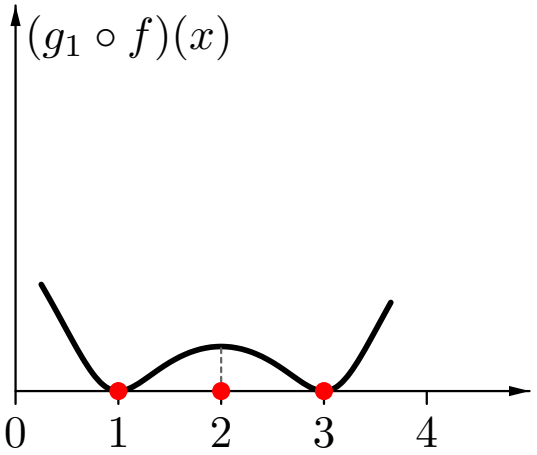
Optima of a Function

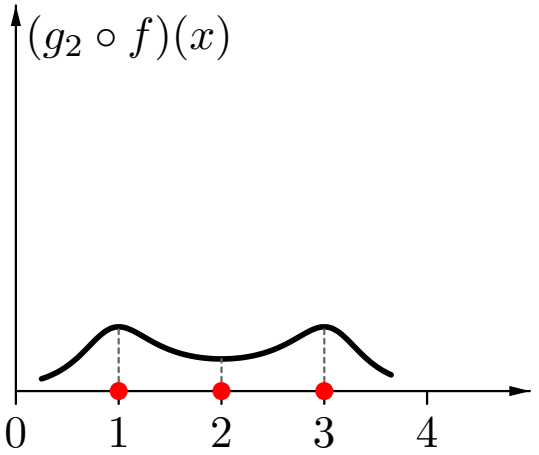
- The **locations** of the optima of a function f do not change if f is composed with a **strictly monotonic** (increasing or decreasing) function g.











Maximum Likelihood Estimation: Example

Given: a family of **normal distributions** $N_X(x; \mu, \sigma^2)$

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Desired: estimators for the unknown parameters μ and σ^2.

The **Likelihood Function**, which describes the probability of the data, is

$$L(x_1, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

To simplify the technical task of forming the partial derivatives,
we consider the natural logarithm of the likelihood function, i. e.

$$\ln L(x_1, \ldots, x_n; \mu, \sigma^2) = -n \ln\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Maximum Likelihood Estimation: Example

- Estimator for the **expected value** μ:

$$\frac{\partial}{\partial \mu} \ln L(x_1, \ldots, x_n; \mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \stackrel{!}{=} 0$$

$$\Rightarrow \qquad \sum_{i=1}^{n} (x_i - \mu) = \left(\sum_{i=1}^{n} x_i \right) - n\mu \stackrel{!}{=} 0 \qquad \Rightarrow \qquad \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Estimator for the **variance** σ^2:

$$\frac{\partial}{\partial \sigma^2} \ln L(x_1, \ldots, x_n; \mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 \stackrel{!}{=} 0$$

$$\Rightarrow \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \frac{1}{n^2} \left(\sum_{i=1}^{n} x_i \right)^2 \qquad \text{(biased!)}$$

Maximum A-posteriori Estimation: Motivation

Consider the following three situations:

- A drunkard claims to be able to predict the side on which a thrown coin will land (head or tails). On ten trials he always states the correct side beforehand.
- A tea lover claims that she is able to taste whether the tea or the milk was poured into the cup first. On ten trials she always identifies the correct order.
- An expert of classical music claims to be able to recognize from a single sheet of music whether the composer was Mozart or somebody else. On ten trials he is indeed correct every time.

Maximum likelihood estimation treats all situations alike, because formally the samples are the same. However, this is implausible:

- We do not believe the drunkard at all, despite the sample data.
- We highly doubt the tea drinker, but tend to consider the data as evidence.
- We tend to believe the music expert easily.

Maximum A-posteriori Estimation

- Background knowledge about the plausible values can be incorporated by
  - using a **prior distribution** on the domain of the parameter and
  - adapting this distribution with **Bayes' rule** and the data.
- Formally maximum a-posteriori estimation is defined as follows:
  find the parameter value θ that maximizes

$$f(\theta \mid D) = \frac{f(D \mid \theta) f(\theta)}{f(D)} = \frac{f(D \mid \theta) f(\theta)}{\int_{-\infty}^{\infty} f(D \mid \theta') f(\theta') \mathrm{d}\theta'}$$

- As a comparison: maximum likelihood estimation maximizes

$$f(D \mid \theta)$$

- Note that $f(D)$ need not be computed: It is the same for all parameter values and since we are only interested in the value θ that maximizes $f(\theta \mid D)$ and not the *value of* $f(\theta \mid D)$, we can treat it as a normalization constant.

Maximum A-posteriori Estimation: Example

Given: a family of **binomial distributions**

$$f_X(x; \theta, n) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

Desired: an estimator for the unknown parameter θ.

a) **Uniform prior**: $\quad f(\theta) = 1, \quad 0 \leq \theta \leq 1.$

$$f(\theta \mid D) = \gamma \binom{n}{x} \theta^x (1-\theta)^{n-x} \cdot 1 \qquad \Rightarrow \qquad \hat{\theta} = \frac{x}{n}$$

b) **Tendency towards $\frac{1}{2}$**: $\quad f(\theta) = 6\theta(1-\theta), \quad 0 \leq \theta \leq 1.$

$$f(\theta \mid D) = \gamma \binom{n}{x} \theta^x (1-\theta)^{n-x} \cdot \theta(1-\theta) = \gamma \binom{n}{x} \theta^{x+1} (1-\theta)^{n-x+1}$$

$$\Rightarrow \qquad \hat{\theta} = \frac{x+1}{n+2}$$

Excursion: Dirichlet's Integral

- For computing the normalization factors of the probability density functions that occur with polynomial distributions, **Dirichlet's Integral** is helpful:

$$\int_{\theta_1} \dots \int_{\theta_k} \prod_{i=1}^{k} \theta_i^{x_i} \, \mathrm{d}\theta_1 \dots \mathrm{d}\theta_k \;=\; \frac{\prod_{i=1}^{k} \Gamma(x_i + 1)}{\Gamma(n + k)}, \qquad \text{where} \quad n = \sum_{i=1}^{k} x_i$$

  and the Γ-function is the so-called **generalized factorial**:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \mathrm{d}t, \qquad x > 0,$$

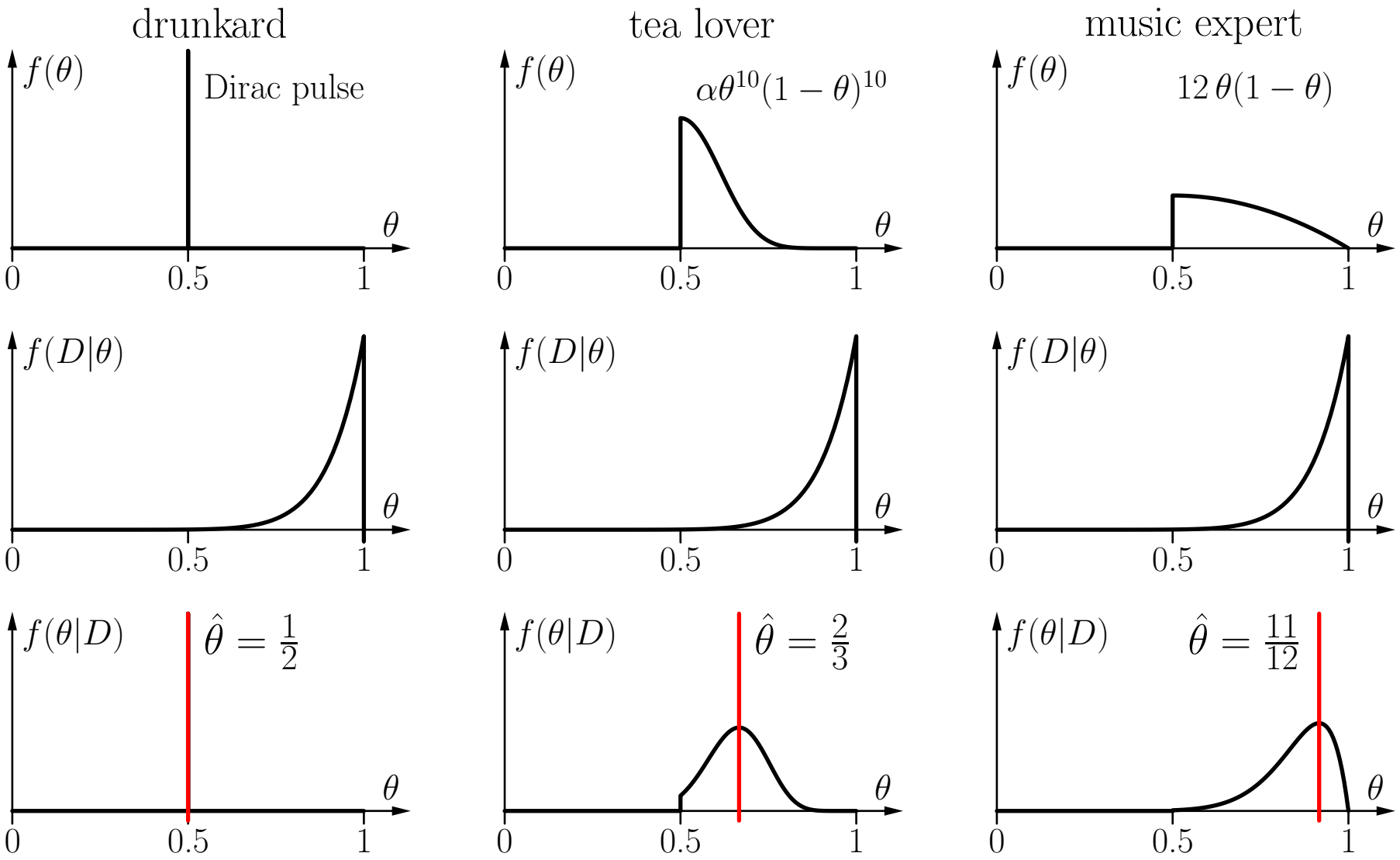
  which satisfies

$$\Gamma(x+1) = x \cdot \Gamma(x), \qquad \Gamma(\tfrac{1}{2}) = \sqrt{\pi}, \qquad \Gamma(1) = 1.$$

- **Example:** the normalization factor α for the binomial distribution prior $f(\theta) = \alpha\, \theta^2 (1-\theta)^3$ is

$$\alpha \;=\; \frac{1}{\int_\theta \theta^2 (1-\theta)^3 \mathrm{d}\theta} \;=\; \frac{\Gamma(5+2)}{\Gamma(2+1)\,\Gamma(3+1)} \;=\; \frac{6!}{2!\,3!} \;=\; \frac{720}{12} \;=\; 60.$$

Maximum A-posteriori Estimation: Example



Interval Estimation

- In general the estimated value of a parameter will differ from the true value.
- It is desirable to be able to make an assertion about the possible deviations.
- The simplest possibility is to state not only a point estimate,
  but also the standard deviation of the estimator:

$$t \pm D(T) = t \pm \sqrt{D^2(T)}.$$

- A better possibility is to find intervals that contain the true value with high probability. Formally they can be defined as follows:
  Let $A = g_A(X_1, \ldots, X_n)$ and $B = g_B(X_1, \ldots, X_n)$ be two statistics, such that

$$P(A < \theta < B) = 1 - \alpha, \qquad P(\theta \leq A) = \frac{\alpha}{2}, \qquad P(\theta \geq B) = \frac{\alpha}{2}.$$

  Then the random interval $[A, B]$ (or an instantiation $[a, b]$ of this interval) is called $(1-\alpha)\cdot 100\%$ **confidence interval** for θ. The value $1-\alpha$ is called **confidence level**.

Interval Estimation

- This definition of a confidence interval is not specific enough:
  A and B are not uniquely determined.
- Common solution: Start from a point estimator T for the unknown parameter θ and define A and B as functions of T:

$$A = h_A(T) \quad \text{and} \quad B = h_B(T).$$

- Instead of $A \leq \theta \leq B$ consider the corresponding event w.r.t. the estimator T, that is, $A^* \leq T \leq B^*$.
- Determine $A = h_A(T)$ and $B = h_B(T)$ from the inverse functions $A^* = h_A^{-1}(\theta)$ and $B^* = h_B^{-1}(\theta)$.

$$\begin{array}{lrcl}
\text{Procedure:} & P(A^* < T < B^*) & = & 1 - \alpha \\
\Rightarrow & P(h_A^{-1}(\theta) < T < h_B^{-1}(\theta)) & = & 1 - \alpha \\
\Rightarrow & P(h_A(T) < \theta < h_B(T)) & = & 1 - \alpha \\
\Rightarrow & P(A < \theta < B) & = & 1 - \alpha.
\end{array}$$

Interval Estimation: Example

Given: a family of **uniform distributions** on the interval $[0, \theta]$, i.e.

$$f_X(x;\theta) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \leq x \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Desired: a confidence interval for the unknown parameter θ.

- Start from the unbiased point estimator $U = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}$:

$$P(U \leq B^*) = \int_0^{B^*} f_U(u;\theta)\mathrm{d}u = \frac{\alpha}{2}$$

$$P(U \geq A^*) = \int_{A^*}^{\frac{n+1}{n}\theta} f_U(u;\theta)\mathrm{d}u = \frac{\alpha}{2}$$

- From the study of point estimators we know

$$f_U(u;\theta) = \frac{n^{n+1}}{(n+1)^n} \frac{u^{n-1}}{\theta^n}.$$

Interval Estimation: Example

- Solving the integrals gives us

$$B^* = \sqrt[n]{\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta \qquad \text{and} \qquad A^* = \sqrt[n]{1-\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta,$$

that is,

$$P\left(\sqrt[n]{\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta < U < \sqrt[n]{1-\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta\right) = 1-\alpha.$$

- Computing the inverse functions leads to

$$P\left(\frac{U}{\sqrt[n]{1-\frac{\alpha}{2}}\frac{n+1}{n}} < \theta < \frac{U}{\sqrt[n]{\frac{\alpha}{2}}\frac{n+1}{n}}\right) = 1-\alpha,$$

that is,

$$A = \frac{U}{\sqrt[n]{1-\frac{\alpha}{2}}\frac{n+1}{n}} \qquad \text{and} \qquad B = \frac{U}{\sqrt[n]{\frac{\alpha}{2}}\frac{n+1}{n}}.$$

Interval Estimation: Common Misconceptions

- Since A and B are functions of random variables $\vec{X} = (X_1, \ldots, X_n)$ (modelling the underlying random sampling), they are random variables themselves and thus a statement like
$$P(A(\vec{X} < \theta < B(\vec{X})) = 1 - \alpha$$
makes sense.

- If, however, applied to a specific random sample $\vec{x}$ the interval borders $a = A(\vec{x})$ and $b = B(\vec{x})$ become fixed and are not random anymore.

- A probability statement about $a < \theta < b$ would be nonsensical because either $\theta \in [a, b]$ or $\theta \notin [a, b]$.

- Therefore it is incorrect to say:
"The true parameter θ lies with $(1 - \alpha) \cdot 100\%$ probability within the confidence interval."

- Correct: **"This confidence interval has been generated by a procedure which returns for $(1 - \alpha) \cdot 100\%$ of all possible samples $\vec{x}$ an interval that contains the true parameter θ."**

Interval Estimation: Common Misconceptions

Relation to sample size n and confidence level α.

- Width of a confidence interval can be considered a measure of imprecision or inaccuracy, i. e., the smaller the interval the more accurate the estimation. (although the real parameter may not be within the interval at all, of course).
- Increasing n yields a smaller interval.
- Increasing α yields a smaller interval. (Often misunderstood!)
- Example: random variable X with binomial distribution: $b_X(x; p, n)$
- Let x be the number of positive outcomes in the sample of size n.
  The $(1 - \alpha) \cdot 100\%$ confidence interval for p reads:

$$\left[r - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{r(1-r)}{n-1}}, \quad r + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{r(1-r)}{n-1}} \right] \quad \text{with} \quad r = \frac{x}{n}$$

  (with $z_a = \Phi^{-1}(a)$ and Φ being the standard normal distribution function)

Hypothesis Testing

- A **hypothesis test** is a statistical procedure with which a decision is made between two contrary hypothesis about the process that generated the data.
- The two hypotheses can refer to
  - the value of a parameter **(Parameter Test)**,
  - a distribution assumption **(Goodness-of-Fit Test)**,
  - the dependence of two attributes **(Dependence Test)**.
- One of the two hypothesis is preferred, that is, in case of doubt the decision is made in its favor. (One says that it gets the "benefit of the doubt".)
- The preferred hypothesis is called the **Null Hypothesis** H_0,
  the other hypothesis is called the **Alternative Hypothesis** H_a.
- Intuitively: the null hypothesis H_0 is put on trial.
  Only if the evidence is strong enough, it is convicted (i.e. rejected).
  If there is doubt, however, it is acquitted (i.e. accepted).

Hypothesis Testing

- The test decision is based on a **test statistic**,
  that is, a function of the sample values.
- The null hypothesis is rejected if the value of the test statistic
  lies inside the so-called **critical region** C.
- Developing a hypothesis test consists in finding the critical region
  for a given test statistic and significance level (see below).
- The test decision may be wrong. There are two possible types of errors:

  Type 1: The null hypothesis H_0 is rejected, even though it is correct.
  Type 2: The null hypothesis H_0 is accepted, even though it is false.
- Type 1 errors are considered to be more severe,
  since the null hypothesis gets "the benefit of the doubt".
- Therefore it is tried to restrict the probability of a type 1 error to a certain maximum α. This maximum value α is called **significance level**.

Example: Outlier Detection - Single Attributes

Numerical attributes:

- Outliers in boxplots.
  Problems: Asymmetric distribution, large data sets
- Statistical tests, for example *Grubb's test*:

Define the statistic

$G = \frac{\max\{\|x_i - \overline{x}\| : 1 \leq i \leq n}{s}$, where $x_1, \ldots, x_n$ is the sample, $\overline{x}$ its mean value and s its empirical standard deviation. For a given significance level α, the null hypothesis that the sample coming from a *normal distribution* does not contain outliers is rejected if

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{1-\alpha/(2n),n-2}}{n-2+t^2_{1-\alpha/(2n),n-2}}}$$

where $t_{1-\alpha/(2n),n-2}$ denotes the $(1 - \alpha/(2n))$-quantile of the t-distribution with $n-$ degrees of freedom.

Parameter Test

- In a parameter test the contrary hypotheses refer to the value of a parameter, for example (one-sided test):

$$H_0: \quad \theta \geq \theta_0, \qquad\qquad H_a: \quad \theta < \theta_0.$$

- For such a test usually a point estimator T is chosen as the test statistic.
- The null hypothesis H_0 is rejected if the value t of the point estimator does not exceed a certain value c, the so-called **critical value** (i.e. $C = (-\infty, c]$).
- Formally the critical value c is determined as follows: We consider

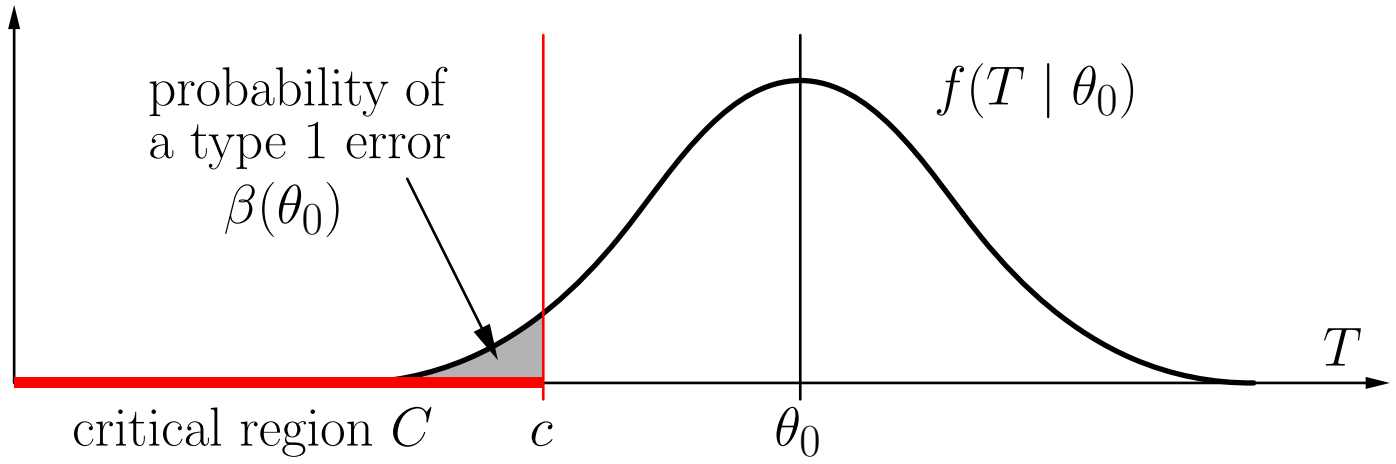
$$\beta(\theta) = P_\theta(H_0 \text{ is rejected}) = P_\theta(T \in C),$$

  the so-called **power** β of the test.
- The power must not exceed the significance level α for values θ satisfying H_0:

$$\max_{\theta:\theta \text{ satisfies } H_0} \beta(\theta) \leq \alpha. \qquad\qquad \text{(here:} \qquad \beta(\theta_0) \leq \alpha)$$

Parameter Test: Intuition

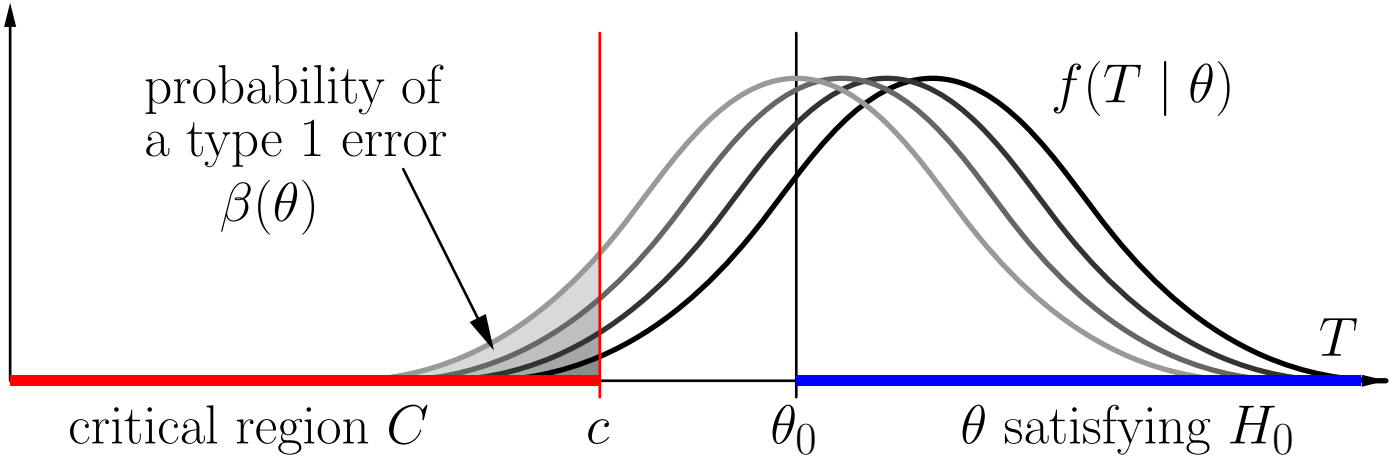
- The probability of a type 1 error is the area under the estimator's probability density function $f(T \mid \theta_0)$ to the left of the critical value c. (Note: This example illustrates $H_0 : \theta \geq \theta_0$ and $H_a : \theta < \theta_0$.)



- Obviously the probability of a type 1 error depends on the location of the critical value c: higher values mean a higher error probability.
- Idea: Choose the location of the cricital value so that the maximal probability of a type 1 error equals α, the chosen significance level.

Parameter Test: Intuition

- What is so special about θ_0 that we use $f(T \mid \theta_0)$?



- In principle, all θ satisfying H_0 have to be considered,
  that is, all density functions $f(T \mid \theta)$ with $\theta \geq \theta_0$.

- Among these values θ, the one with the highest probability of a type 1 error
  (i.e., the one with the highest power $\beta(\theta)$) determines the critical value.

  Intuitively: we consider the **worst possible case**.

Parameter Test: Example

- We consider a one-sided test of the expected value μ of a normal distribution $N(\mu, \sigma^2)$ with known variance σ^2, i.e., we consider the hypotheses

$$H_0: \quad \mu \geq \mu_0, \qquad\qquad H_a: \quad \mu < \mu_0.$$

- As a test statistic we use the standard point estimator for the expected value

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

  This point estimator has the probability density

$$f_{\bar{X}}(x) = N\left(x; \mu, \frac{\sigma^2}{n}\right).$$

- Therefore it is (with the $N(0,1)$-distributed random variable Z)

$$\alpha = \beta(\mu_0) = P_{\mu_0}(\bar{X} \leq c) = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = P\left(Z \leq \frac{c - \mu_0}{\sigma/\sqrt{n}}\right).$$

Parameter Test: Example

- We have as a result that
$$\alpha = \Phi\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right),$$
where Φ is the distribution function of the standard normal distribution.
- The distribution function Φ is tabulated, because it cannot be represented in closed form. From such a table we retrieve the value z_α satisfying $\alpha = \Phi(z_\alpha)$.
- Then the critical value is
$$c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$
(Note that the value of z_α is negative due to the usually small value of α. Typical values are $\alpha = 0.1$, $\alpha = 0.05$ or $\alpha = 0.01$.)
- H_0 is rejected if the value $\bar{x}$ of the point estimator $\bar{X}$ does not exceed c, otherwise it is accepted.

Parameter Test: Example

- Let $\sigma = 5.4$, $n = 25$ and $\bar{x} = 128$. We choose $\mu_0 = 130$ and $\alpha = 0.05$.
- From a standard normal distribution table we retrieve $z_{0.05} \approx -1.645$ and get

$$c_{0.05} \approx 130 - 1.645\frac{5.4}{\sqrt{25}} \approx 128.22.$$

  Since $\bar{x} = 128 < 128.22 = c$, we reject the null hypothesis H_0.
- If, however, we had chosen $\alpha = 0.01$, it would have been (with $z_{0.01} \approx -2.326$):

$$c_{0.01} \approx 130 - 2.326\frac{5.4}{\sqrt{25}} \approx 127.49$$

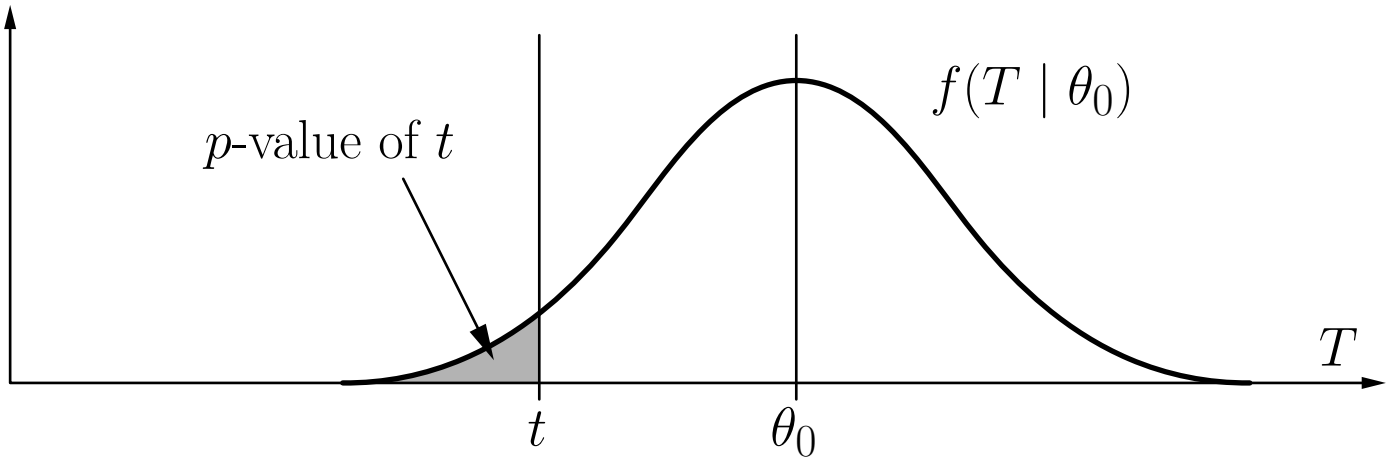
  Since $\bar{x} = 128 > 127.49 = c$, we would have accepted the null hypothesis H_0.
- Instead of fixing a significance level α one may state the so-called **p-value**

$$p = \Phi\left(\frac{128 - 130}{5.4/\sqrt{25}}\right) \approx 0.032.$$

  For $\alpha \geq p = 0.032$ the null hypothesis is rejected, for $\alpha < p = 0.032$ accepted.

Parameter Test: p-value

- Let t be the value of the test statistic T
  that has been computed from a given data set.
  (Note: This example illustrates $H_0 : \theta \geq \theta_0$ and $H_a : \theta < \theta_0$.)



- The **p-value** is the probability that a value of t or less
  can be observed for the chosen test statistic T.

- The p-value is a **lower limit for the significance level** α
  that may have been chosen if we wanted to reject the null hypothesis H_0.

Parameter Test: p-value

Attention: p-values are often misused or misinterpreted!

- A low p-value does **not** mean that the result is very reliable!

  All that matters for the test is whether the computed p-value
  is **below the chosen significance level or not**.

  (A low p-value could just be a chance event, an accident!)

- The significance level may **not** be chosen **after** computing the p-value,
  since we tend to choose lower significance levels if we know that they are met.

  Doing so would undermine the reliability of the procedure!

- Stating p-values is only a convenient way of avoiding a fixed significance level.
  (Since significance levels are a matter of choice and thus user-dependent.)

  However: A significance level must still be chosen
  before a reported p-value is looked at.

Relevance of the Type-2 Error

- Reminder: There are two possible types of errors:

  Type 1: The null hypothesis H_0 is rejected, even though it is correct.

  Type 2: The null hypothesis H_0 is accepted, even though it is false.

- Type-1 errors are considered to be more severe,
  since the null hypothesis gets "the benefit of the doubt".

- However, **type-2 errors should not be neglected** completely:

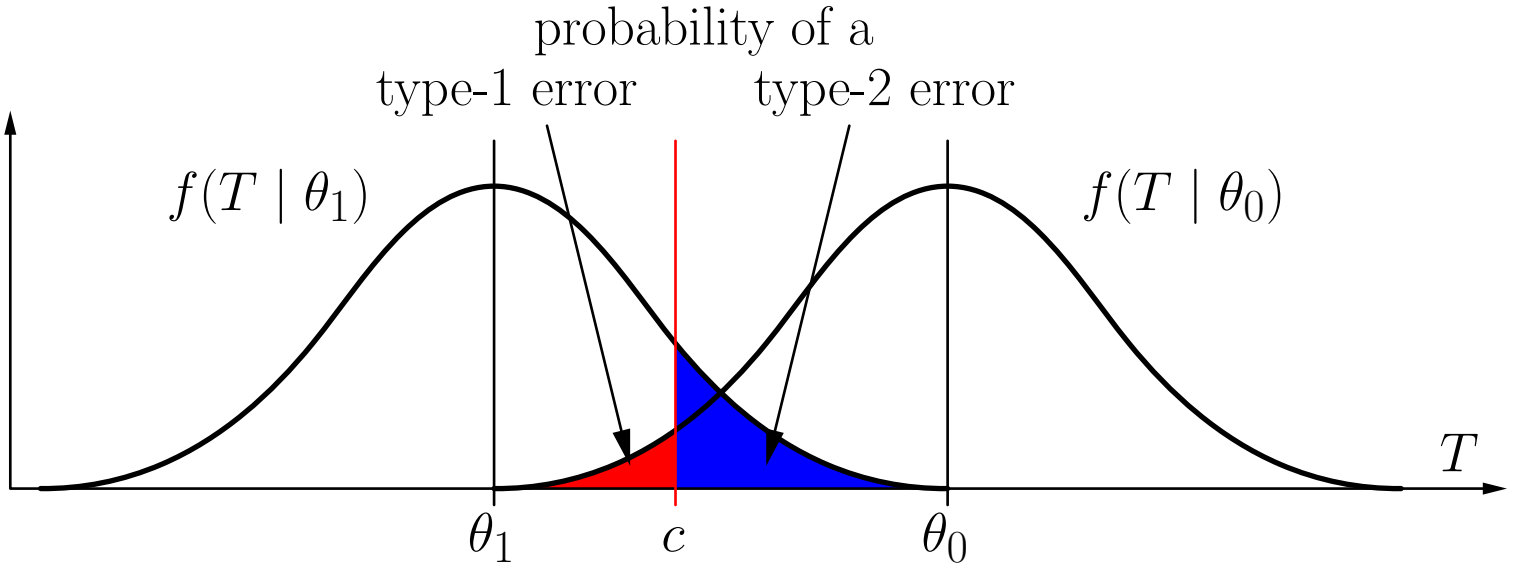
  - It is always possible to achieve a vanishing probability of a type-1 error:
    Simply accept the null hypothesis in all instances, regardless of the data.
  - Unfortunately such an approach maximizes the type-2 error.

- Generally, **type-1 and type-2 errors are complementary quantities**:

  The lower we require the type-1 error to be (the lower the significance level),
  the higher will be the probability of a type-2 error.

Relationship between Type-1 and Type-2 Error

- Suppose there are only two possible parameter values θ_0 and θ_1 with $\theta_1 < \theta_0$. (That is, we have $H_0 : \theta = \theta_0$ and $H_a : \theta = \theta_1$.)



- Lowering the significance level α moves the critical value c to the left: lower type-1 error (red), but higher type-2 error (blue).
- Increasing the significance level α moves the critical value c to the right: higher type-1 error (red), but lower type-2 error (blue).