Traditional, frequentist statistics

Point estimation, bias, variance

We assume that data is generated by a particular family of distributions, for example

$$\mathcal{F} = \{ N(\mu, \sigma^2) \mid \underbrace{\mu \in \mathbb{R}, \sigma^2 > 0}_{\Theta} \}.$$

More generally, this family is typically denoted as follows:

$$\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \}$$

one particular parameter

space of all possible parameters

θ one particular parameter

Θ space of parameters

(parametric statistics!)

The family $\mathcal{F}$ is called the statistical model.

We are given a sample $X_1, \dots, X_n \sim f_\theta$ (typically, iid) but the true, underlying θ is unknown.

The goal of point estimation is to estimate θ.

Convention: Para space Θ, true parameter θ, P_θ, E_θ refers to the probability and expectation under the distribution (density) f_θ, $\hat{\theta}$ estimated parameter

Def Given a statistical model $\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \}$, and a sample $X_1, \dots, X_n \sim F \in \mathcal{F}$. A point estimator $\hat{\theta}_n$ of parameter θ is a function

$$\hat{\theta}_n := g(X_1, \dots, X_n)$$

Def The bias of such an estimator is defined as

$$\text{bias}(\hat{\theta}_n) := \underbrace{E_\theta}_{\text{expectation wrt the distribution } f_\theta \text{ (the true one!)}} (\underbrace{\hat{\theta}_n}_{\text{estimate}}) - \underbrace{\theta}_{\text{true para}}$$

repeat the procedure very often (infinitely often) and average over the estimate $\hat{\theta}_n$.

An estimate is unbiased if its bias is zero.

Def The variance of an estimator is defined as $\text{Var}_\theta(\hat{\theta}_n)$. The corresponding standard deviation is called the standard error se. Typically, se is unknown, but it can be estimated: $\widehat{se}$.

Example: $X_1, \dots, X_n \sim \underset{\{0,1\}}{\text{Bernoulli}}(p)$, parameter $\underbrace{p}_{\theta} \in \underbrace{[0,1]}_{\Theta}$,

$\hat{p}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$ an estimate of p.

$$E_p(\hat{p}_n) = E_p\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \overbrace{E_p(X_i)}^{p} = p.$$

Thus, $\hat{p}_n$ is unbiased because

$$E_p(\hat{p}_n) - p = p - p = 0.$$

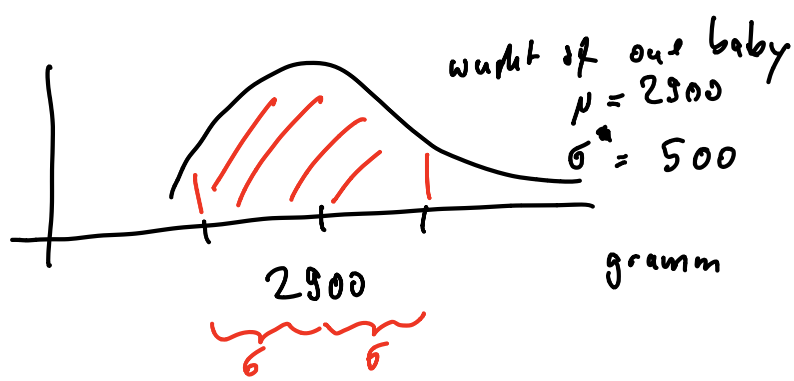
The standard error of this estimate is

$$se = \sqrt{\mathrm{Var}_p(\hat{p}_n)} = \sqrt{\frac{1}{n}\mathrm{Var}_p(X_1)} = \sqrt{\frac{p(1-p)}{n}}$$

We can for example estimate it by

$$\hat{se} = \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}.$$

Example: weight of baby



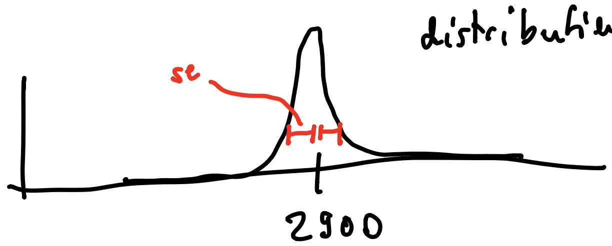
Estimate mean weight of ~~babys~~. Measure weight of 100 ~~babys~~ babies.

$\rightsquigarrow$ 2950 g

Do it a second time $\rightsquigarrow$ 2890 g

$\vdots$

distribution of the estimate $\hat{p}_n$



Def The mean squared error (MSE) of an estimate is the quantity

$$MSE(\hat{\theta}, \theta) = E_\theta\left((\hat{\theta}_n - \underbrace{\theta}_{\text{deterministic}})^2 \right)$$

Theorem : bias-variance-decomposition

$$\underbrace{MSE(\hat{\theta}_n, \theta)}_{\text{how good is our estimate}} = \text{bias}^2(\hat{\theta}_n) + Var_\theta(\hat{\theta}_n)$$

Proof $E_\theta\left((\hat{\theta}_n - \theta)^2 \right) =$

$$= E_\theta\left((\overbrace{\hat{\theta}_n - E\hat{\theta}_n}^{a} + \overbrace{E\hat{\theta}_n - \theta}^{b})^2 \right)$$

$$= E_\theta\left((\hat{\theta}_n - E\hat{\theta}_n)^2 \right) + \underbrace{2E_\theta\left((\hat{\theta}_n - E\hat{\theta}_n)(\underbrace{E\hat{\theta}_n - \theta}_{\text{deterministic}}) \right)}_{2(E\hat{\theta}_n - \theta)\cdot \underbrace{E_\theta(\hat{\theta}_n - E\hat{\theta}_n)}_{= E_\theta(\hat{\theta}_n) - E_\theta\hat{\theta}_n = 0} = 0} + E\left((E\hat{\theta}_n - \theta)^2 \right)$$

$$= \underbrace{E_\theta\left((\hat{\theta}_n - E\hat{\theta}_n)^2 \right)}_{Var(\hat{\theta}_n)} + \underbrace{\cancel{E}\left((\underbrace{E\hat{\theta}_n - \theta}_{\text{deterministic}})^2 \right)}_{= (E\hat{\theta}_n - \theta)^2 = (\text{bias}(\hat{\theta}_n))^2}$$

$\blacksquare$

<u>Example:</u> Model $\mathcal{F} = \{ N(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma > 0 \}$

Sample: $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ with unknown μ, σ^2, iid

$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an unbiased estimate of μ.

$$\hat{\sigma}_1^2 := \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2$$

$$\hat{\sigma}_2^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu})^2$$

Let's compute:

$$E(\hat{\sigma}_1^2) = \overbrace{\frac{n-1}{n}}^{!} \overset{\text{true para } \sigma}{\sigma^2} \quad \text{so the bias is } \frac{1}{n}\sigma^2$$

$$E(\hat{\sigma}_2^2) = \sigma^2 \quad \text{unbiased!}$$

$$\text{Var}(\hat{\sigma}_1^2) = \frac{2(n-1)\sigma^4}{n^2}$$

$$\text{Var}(\hat{\sigma}_2^2) = \frac{2\sigma^4}{n-1}$$

$$MSE(\hat{\sigma}_1^2) = \text{bias}^2 + \text{var} = \dots = \left(\frac{2n-1}{n^2}\right) \sigma^4$$

$$MSE(\hat{\sigma}_2^2) = \dots = \frac{2}{n-1} \sigma^4$$

$$\Rightarrow MSE(\hat{\sigma}_1^2) < MSE(\hat{\sigma}_2^2)$$

Def A point estimator $\hat{\theta}_n$ of θ is <u>consistent</u> (strongly consistent) if

$$\hat{\theta}_n \to \theta \text{ in probability (a.s.)}$$

as $n \to \infty$

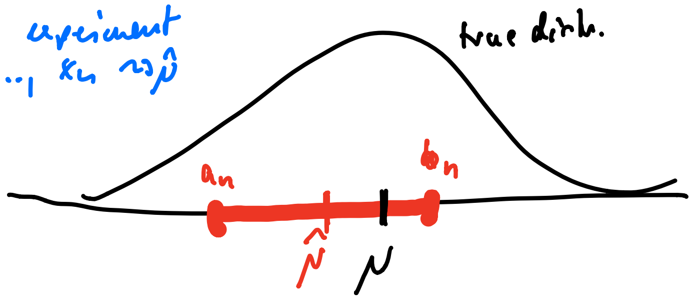
Theorem If an estimate satisfies bias $\to 0$ and se $\to 0$ as $n \to \infty$, the estimate is consistent.

Confidence sets

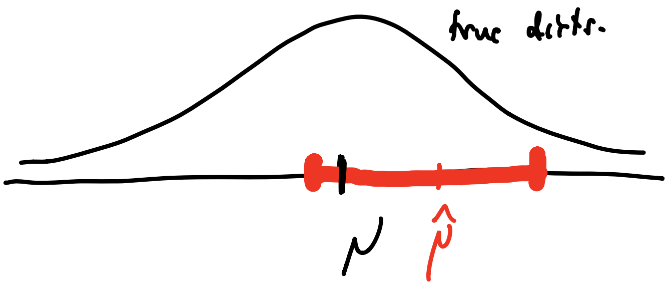
Def A $(1-\alpha)$-confidence interval for a parameter $\theta \in \mathbb{R}$ is an interval $C_n = (a_n, b_n)$ where $a_n = a(X_1, \dots, X_n)$, $b_n = b(X_1, \dots, n)$ are functions of the sample $X_1, \dots, X_n$ such that

$$P_\theta(\underbrace{\theta}_{\text{deterministic}} \in \underbrace{C_n}_{\text{random}}) \geq 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

First experiment
$X_1, \dots, X_n \rightsquigarrow \hat{\mu}$



Second experiment
$X_1, \dots, X_n \rightsquigarrow \hat{\mu}$



... in $(1-\alpha)$ of the repetitions, the true μ is inside the red interval.

Example: Coin flips, with $P(X=1) = p$, $P(X=0) = 1-p$, $p \in [0,1]$ unknown. Want to estimate it.

$\rightsquigarrow$ Observe $X_1, \dots, X_n \sim f_p$

$\hat{p}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$. Choose a confidence level α,

now want to define $C_n = (a_n, b_n)$. To this end,

$$\varepsilon_n^2 := \frac{\log(2/\alpha)}{2n}.$$

Then the interval

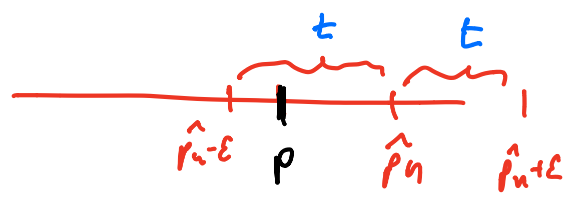
$$C_n := \left(\hat{p}_n - \varepsilon_n, \ \hat{p}_n + \varepsilon_n \right)$$

is a CI with coverage $1-\alpha$.

Proof:

By Hoeffding inequality, for any t we have

$$P(|\hat{p}_n - p| > t) \leq \underbrace{2\exp(-2nt^2)}_{\alpha}$$



$$\alpha = 2\exp(-2nt^2)$$

$$\log\left(\frac{\alpha}{2}\right) = -2nt^2 \Rightarrow t^2 = \frac{-\log(\alpha/2)}{2n} = \frac{\log(2/\alpha)}{2n}$$

Chose $\varepsilon_n = t$. ∎

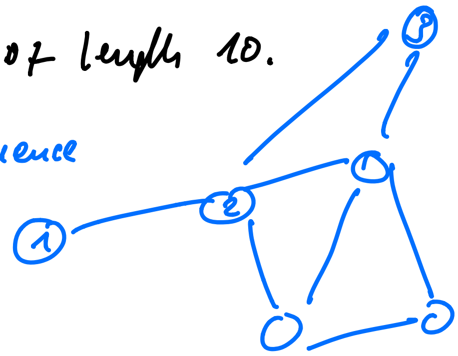
Maximum likelihood estimator

Example: $\mathcal{F} = \{ A \mid A \text{ symmetric}, \ a_{ij} \in \{0,1\}^{n \times n} \}$

adjacency matrices of graphs

Observe k random walks from the graph of length 10.

one random walk produces a sequence $x_1, x_2, \dots, x_{10}$ of vertices.



Goal: reconstruct (estimate) A

Idea: among all adjacency matrices $A \in \mathcal{F}$, select the one that has the highest likelihood to have produced the random walks you have observed.

$\rightsquigarrow$ Maximum likelihood approach

More formally: Parametric family $\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \}$,

observe iid points $x_1, \dots, x_n \sim f_\theta \in \mathcal{F}$.

The likelihood of the data given a parameter θ_0 is

$$P_{\theta_0}(x_1, \dots, x_n) = P(x_1, \dots, x_n \mid \theta_0)$$

$$= \prod_{i=1}^{n} P(x_i \mid \theta_0)$$

To estimate the true parameter θ, we now select $\hat{\theta}$ such that this likelihood is maximized:

$$\hat{\theta} := \underset{\theta \in \Theta}{\operatorname{argmax}} \; P(X_1, \dots, X_n \mid \theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_i P(X_i \mid \theta)$$

This is equivalent to the problem

$$\hat{\theta} = \operatorname{argmax} \log \left(\prod_{i=1}^{n} P(X_i \mid \theta) \right)$$

$$= \operatorname{argmax} \sum_{i=1}^{n} \underbrace{\log \underbrace{P(X_i \mid \theta)}_{\in [0,1]}}_{< 0}$$

which is equivalent to minimizing the negative log-likelihood:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \underbrace{- \log P(X_i \mid \theta)}_{> 0}$$

This is the maximum likelihood approach.

Sometimes this optimization problem is easy:

- it might be able to solve it analytically (rare)
- if you are lucky it is convex
- Most typically, it is not convex.

Example (analytic solution)

Model: $X \sim \text{Poisson}(\lambda)$, this means that

$$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{it has } E(X) = \lambda, \; Var(X) = \lambda.$$

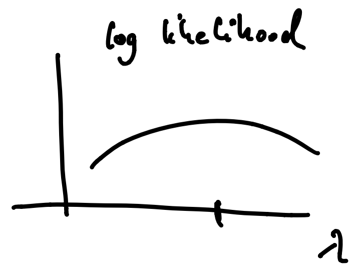
Observe $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$

Want to construct the ML-estimator.

$$\mathcal{L}(\lambda) = P(x_1, \ldots, x_n \mid \lambda) = \prod \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

$$\log(\ldots) = \sum_{i=1}^n \log\left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right)$$

$$= \underbrace{\sum_{i=1}^n x_i \log \lambda - \lambda - \log(x_i!)}_{f(\lambda)}$$



Now want to optimize for λ. Take the derivative

typo fixed (removed paranthesis)

$$f'(\lambda) = \sum_{i=1}^n \frac{x_i}{\lambda} - n = \frac{1}{\lambda}\left(\sum_{i=1}^n x_i\right) - n \overset{!}{=} 0$$

$$\Rightarrow \lambda = \frac{1}{n} \sum_{i=1}^n x_i$$

So $\hat{\lambda} := \frac{1}{n} \sum_{i=1}^n x_i$ is the ML estimate of λ.

□ Example

From the theory side, MLE often (but not always) has nice properties:

(1) If the model $\mathcal{F}$ consists of "nice" functions, then the MLE based on an iid sample is consistent.

(2) If $\mathcal{F}$ consists of "nice" functions, the MLE estimate $\hat{\theta}_{MLE}$ is asymptotically normal:

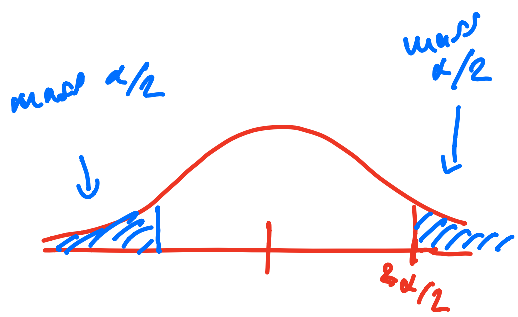
$$\frac{\hat{\theta}_{MLE} - \theta}{se} \xrightarrow{\text{in distr.}} N(0,1) \qquad \text{and}$$

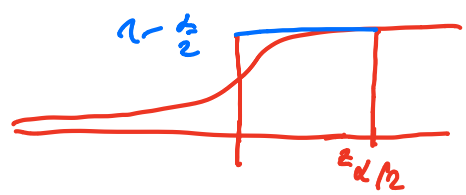
$$\frac{\hat{\theta}_{MLE} - \theta}{\hat{se}} \xrightarrow{\text{in distr.}} N(0,1)$$

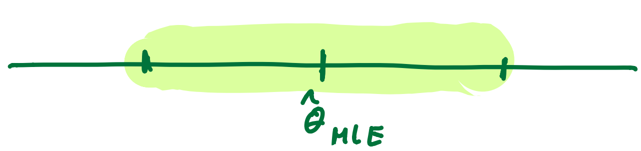
(3) This can be used to construct confidence intervals:

$$C_n := \Big(\hat{\theta}_{MLE} - \underbrace{z_{\alpha/2} \, \hat{se}}_{-\varepsilon} \, , \ \hat{\theta}_{MLE} + \underbrace{z_{\alpha/2} \cdot \hat{se}}_{+\varepsilon} \Big)$$

where $z_{\alpha/2} := \underset{\text{cdf of } N(0,1)}{\Phi}^{-1}\left(1 - \frac{\alpha}{2}\right)$







C_n is an approximate $\underline{CI}$ in the sense that

$$P_\theta(\theta \in C_n) \longrightarrow 1-\alpha \quad \text{as } n \to \infty.$$

Sufficiency, identifiability

Sufficiency

Intuition: given sample $X_1, \dots, X_n \sim f_\theta \in \mathcal{F}$

we typically convert the (large) sample to a statistic

$T(X_1, \dots, X_n)$ (in the extreme case, one number).

Question: can we recover the true parameter θ from this statistic?

- Intuition: when we observe two samples $X_1, \dots, X_n$, $X_1', \dots, X_n'$ and $T(X_1, \dots, X_n) = T(X_1', \dots, X_n')$, then we want to infer the same θ.

- When we know $T(X_1, \dots, X_n)$, then we can calculate the likelihood of the data.

Formal definition is technical.

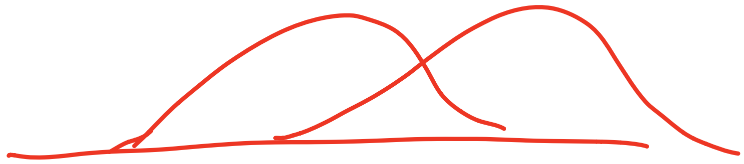
Identifiability

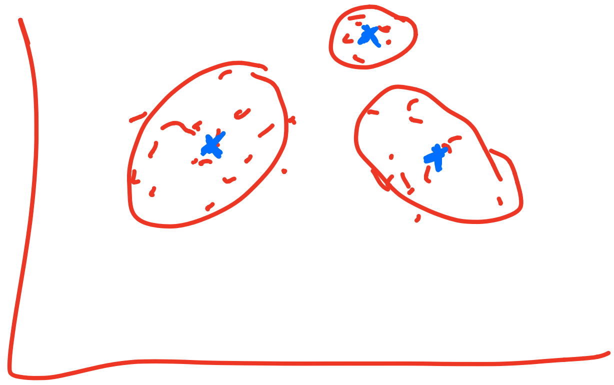
Sometimes families of distributions can be described in different ways with different sets of parameters.

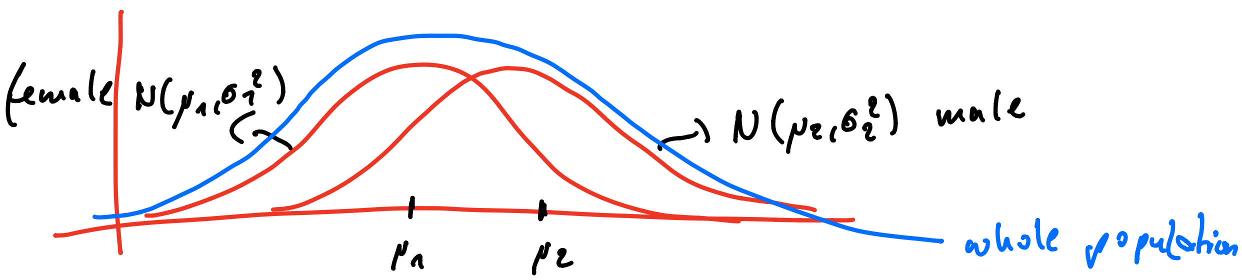
__Def__ A parameter θ for a family $\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \}$ is identifiable if distinct values of θ correspond to distinct pdfs in $\mathcal{F}$:
$$\theta \neq \theta' \Rightarrow f_\theta \neq f_{\theta'}$$

Example: Mixture distributions

$$\mathcal{F} = \left\{ \sum \alpha_i \, N(\mu_i, \sigma_i^2) \right\} \quad \text{with} \quad \sum \alpha_i = 1$$







$$\hat{=} \quad 0.5\, N(\mu_1, \sigma_1^2) + 0.5\, N(\mu_2, \sigma_2^2)$$

You observe samples from the whole population.

→ one way to model the data is in terms of the mixture male/female as above

→ another way to model the data is in terms of the mixture "glasses" / "no glasses"

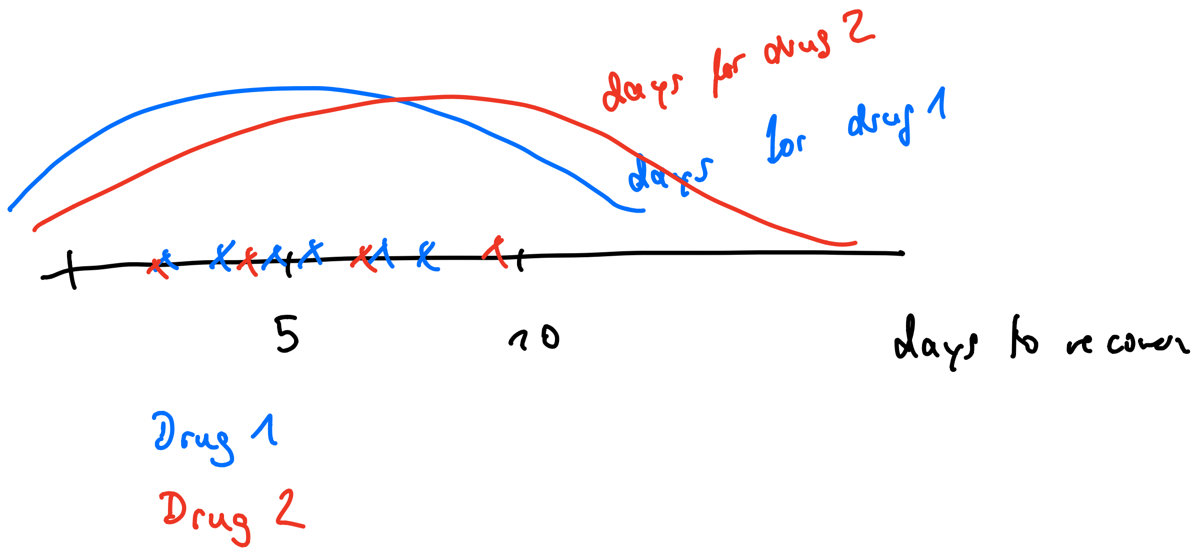
$$0.3\, N(\mu_3, \sigma_3^2) + 0.7\, N(\mu_4, \sigma_3^2)$$

Not identifiable!

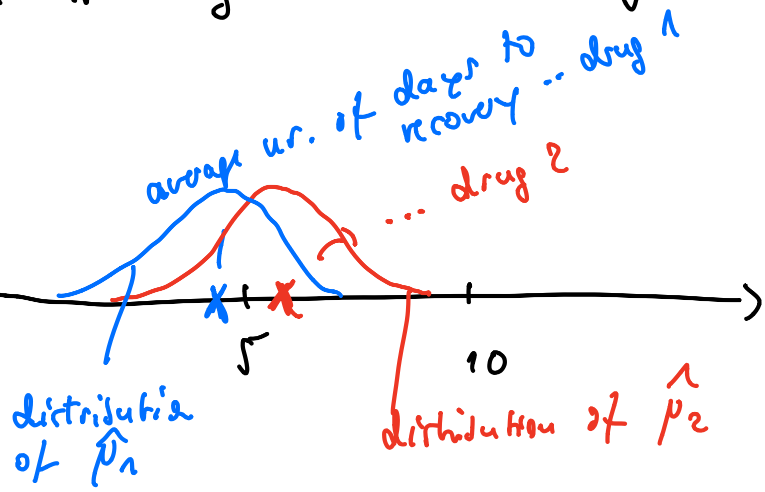
Hypothesis testing

Example: Two drugs D_1, D_2, we measure number of days to recovery

$x_1, \dots, x_n$ treated with D_1 ~



Question: is Drug 1 better than Drug 2?



Example: Want to test whether a coin is fair.

Null hypothesis: H_0 : coin is fair

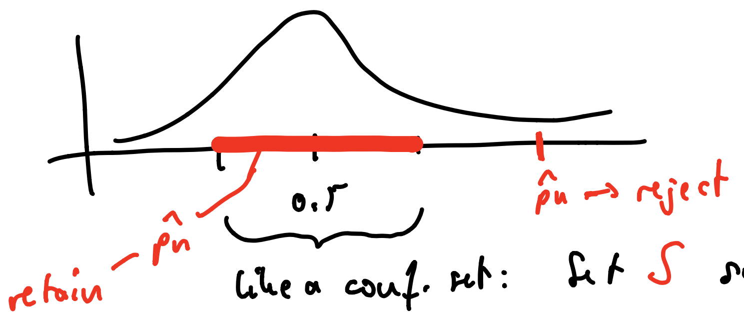
Alternative hypothesis: H_1 : coin is unfair

Sample many coin flips and estimate $\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$.

We want to reject H_0 if $\hat{p}_n$ is "far away" from 0.5.

Question: "far away"?

Look at the distribution of $\hat{p}$ under the null hypothesis:



Like a conf. set: Set S such that

$$P_p(\hat{p}_n \in S) \geq 95\%$$

More formal setup

Statistical model $\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \}$. Assume that

$\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$.

Want to test

$$\underbrace{H_0 : \theta \in \Theta_0}_{\text{null hyp.}} \quad \text{against} \quad \underbrace{H_1 : \theta \in \Theta_1}_{\text{alternative hyp.}}.$$

Sample data from the unknown f_θ, compute a test statistic $T(X_1, \dots, X_n)$. Now we construct a rejection region R_n such that

$$T(X_1, \dots, X_n) \in R_n \Rightarrow \text{reject } H_0$$

$$T(X_1, \dots, X_n) \notin R_n \Rightarrow \text{retain } H_0$$

Typical hypotheses are the of the form

- $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$
- $H_0 : \theta < \theta_0$ vs $H_1 : \theta \geq \theta_0$

Two types of error can occur:

	Test retains H_0	Test rejects H_0
H_0 true	☺	Type I error
H_1 true	Type II error	☺

<u>Def</u> The <u>power function</u> of a test with rejection region R is the function

$$\beta(\theta) := P_\theta(T(X) \in R)$$

- If $\theta \in \Theta_0$ then $T(X)$ should not end up in R. For such θ, $\beta(\theta) = P(\text{Type I error})$. Ideally, $\beta(\theta)$ should be small.
- If $\theta \in \Theta_1$ then we hope that $T(X) \in R$. So
$$\beta(\theta) = 1 - P(\text{Type II error}).$$
Ideally, $\beta(\theta)$ is large.

<u>Def</u> We say that a test is of level α if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$$

Intuition: worst case guarantee, no matter which $\theta \in \Theta_0$ we pick, the type I error is not larger than α.

(Intuition to remember: $\alpha \hat{=}$ type I error)

Standard procedure: We fix the level α of a test in advance,

for example 0.05 or 0.01.

Then we can also look at the type II error. For example among several tests of level α, you might now choose the one that has the smallest type-II-error.

Notation used often in literature:

$\alpha = P(\text{type I error})$ $\qquad \alpha$ = level of the test

$\beta = P(\text{type II error})$ $\qquad 1-\beta$ = power of a test

Remark: the power of a test is typically evaluated when we test against a concrete hypothesis $\theta \in \Theta_1$. We say "the power of the test against alternative $\theta_1 \in \Theta_1$".

Def Let $\mathcal{T}$ be a set of tests of level α for testing

$$H_0: \theta \in \Theta_0 \quad \text{vs} \quad H_1: \theta \notin \Theta_0 .$$

A test in $\mathcal{T}$ with power function $\beta(\theta)$ is uniformly most powerful (UMP) if

$$\beta(\theta) \geq \beta'(\theta) \qquad \text{for all } \theta \in \Theta_0^C$$

and for all β' that are power functions for other tests in $\mathcal{T}$.

Remark: In practice it is often impossible to find an UMP test.

Neyman-Pearson-Lemma and likelihood ratio tests

Theorem Suppose we test $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$.

Consider

$$T = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{\prod_{i=1}^{n} f(x_i \mid \theta_1)}{\prod_{i=1}^{n} f(x_i \mid \theta_0)} \quad \Bigg\} \text{ likelihood ratio.}$$

Assume we reject H_0 if $T > k$ (for some k).

If we choose k such that $P_{\theta_0}(T > k) = \alpha$,

then this is the most powerful level-α-test.

More general likelihood-ratio-test:

Parameter space Θ, $\Theta_0 \subset \Theta$, $\Theta_1 = \Theta_0^c$. Then we

consider the test statistic

$$\tilde{T} = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta_1} \mathcal{L}(\theta)} \qquad \text{or even simpler}$$

$$T = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)}$$

and we determine a parameter λ such that the rejection region is of the form $R = \{ T \leq \lambda \}$.

In practice the difficulties are

- compute the suprema (in practice)
- fix R, fix λ (in theory)

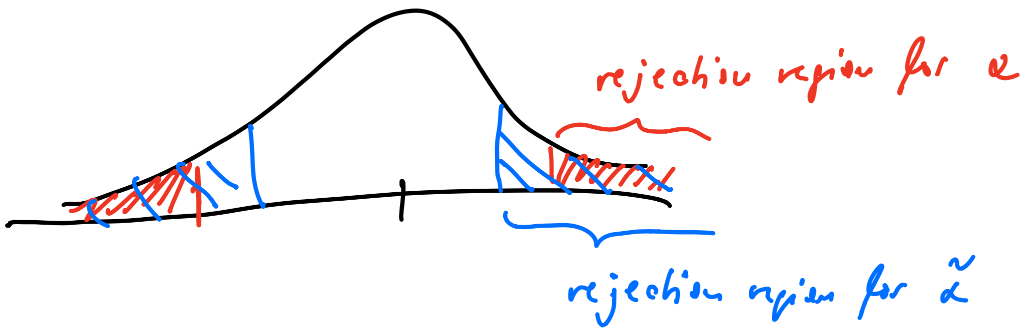
p-values

Consider a test at level α, and denote its rejection region as R_α.

Recall: $\alpha = P(\text{Type-I-error})$.

The smaller α, the more difficult does it get to reject H_0.

(we often even have that $\alpha < \tilde{\alpha} \Rightarrow R_\alpha \subset R_{\tilde{\alpha}}$)



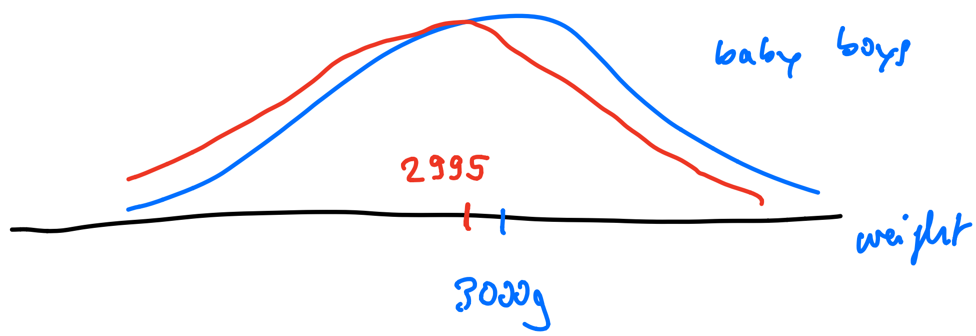
Def The p-value is defined as

$$p = \inf \{ \alpha \mid T(X_1, \dots, X_n) \in R_\alpha \}$$

i.e. the smallest α for which the level-α-test would reject the null hypothesis.

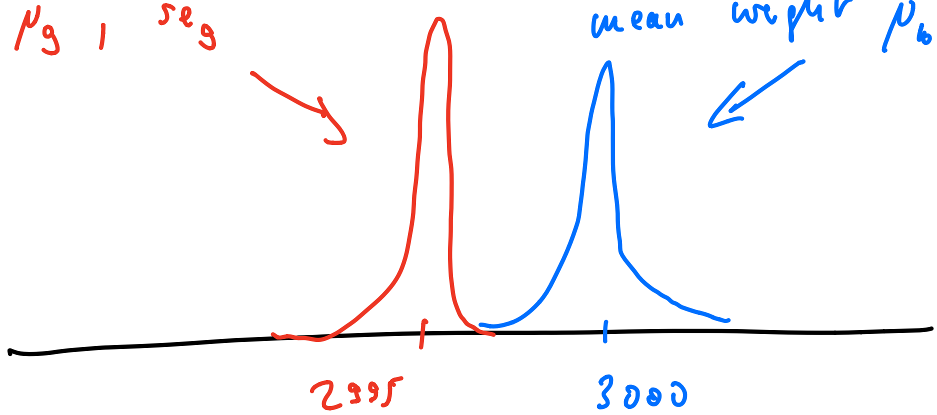
Intuition: smaller p-values are "better", more evidence for rejecting the null (less error).

Example baby boys and girls



Sample many baby girls, many baby boys

$\hat{\mu}_g$, $\hat{se}_g$ mean weight $\hat{\mu}_b$, & $\hat{se}_b$



For n large test will find a statistically significant difference. ~ small p

Multiple testing

Example: gene expression data

	patients with cancer ($n = 20$)	control group ($n = 20$)
gene 1		
gene 2		
⋮		
gene 17	0.5 0.2 0.9 0.8 0.5	0.01 0.05 0.1 0.02
⋮		
gene 105		
gene 1000 $=: m$		

α % of the test will "ring a bell"

Assume we run, for each gene, a test of level α

$P($ Test i makes type-I-error $) = 5\%$

Now we have m tests.

$P($ at least one of the tests makes a type-I-error $) =$

$= P(t_1$ makes error $\underline{\text{or}}$ t_2 error $\underline{\text{or}}$ .. $\underline{\text{or}}$ t_m makes error $)$

$= 1 - P($ no error in t_1 $\underline{\text{and}}$ no error in t_2 $\underline{\text{and}}$ ... $) \overset{\text{assume independence}}{=}$

$$= 1 - \prod_{i=1}^{m} P(\text{no error in } t_i) = \underbrace{1 - (0.95)^m}_{*} \xrightarrow[m \to \infty]{} 1$$

$$\left.\begin{array}{lll} m=1 & \Rightarrow & * = 0.05 \\ m=10 & \Rightarrow & * = 0.40 \\ m=50 & \Rightarrow & * = 0.92 \end{array}\right\} = \text{FWER (see below)}$$

Bonferroni: controlling FWER

Definition: Consider a family of m tests. The family-wise error rate (FWER) is the probability that at least one type-I-error occurs in the family:

$$\begin{aligned} \text{FWER} = P(\, & t_1 \text{ makes type-I-error } \underline{\text{or}} \\ & t_2 \; \ldots \\ & \ldots \\ & t_m \text{ makes type-I-error }). \end{aligned}$$

Bonferroni correction:

Assume we run m tests, and we want to achieve a FWER α (e.g. $\alpha = 0.05$). Then we run the individual tests with level $\frac{\alpha}{m} =: \alpha_{single}$.

Then we have:

$$\mathrm{FWER} = P(\text{at least one type-I-error}) =$$

$$= P(t_1 \text{ error } \underline{\text{or}}\ t_2 \ \dots) \leq$$

$$\leq \sum_{i=1}^{m} \underbrace{P(t_i \text{ makes error})}_{\alpha_{\text{single}}} = m \cdot \alpha_{\text{single}} = m \cdot \frac{\alpha}{m} = \alpha.$$

Advantage: simple, correct

Disadvantage: too conservative, low power (high type-II-error)

the test barely discovers anything!

Benjamini/Hochberg: Controling FDR

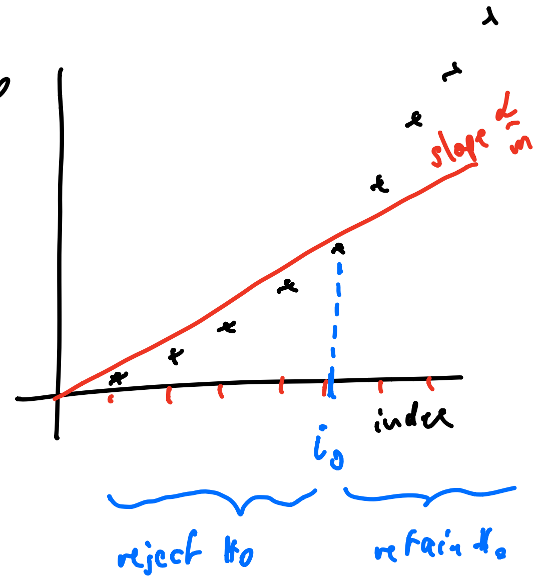
<u>Def</u> Assume we have a family of m tests. We call

$$E\left(\frac{\#\text{ false rejections}}{\#\text{ all rejections}} \right) =: \underline{\mathrm{FDR}}$$

the <u>false discovery rate</u>.

Benjamini/Hochberg (1998) approach:

- Fix FDR α in advance.
- Run the m individual tests and evaluate their p-values.
- Sort p-values increasingly: $p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq \dots \leq p_{(m)}$
- Define thresholds $l_i := i \cdot \frac{\alpha}{m}$
- Find the largest index i_0 such that $p_{(i_0)} \leq l_{i_0}$.
  (below the red line)
- Reject the hypotheses for $i = 1, \dots, i_0$, retain all the others.



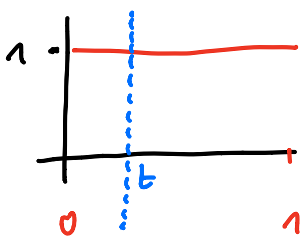
Theorem: If the Benjamini-Hochberg procedure is applied (and the tests are independent), then regardless of how many null hypotheses are true and regardless of the distribution of p-values when the null is false, we obtain

$$FDR \leq \alpha.$$

(Remark: similar approach also works without independence assumption, many modifications exist.)

Intuition:

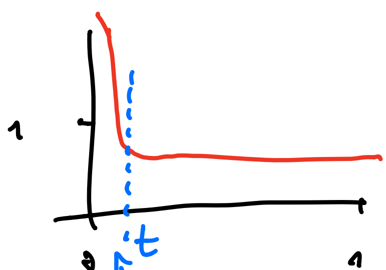
- Under the null hypothesis, the p-values always have a uniform distribution on [0, 1].



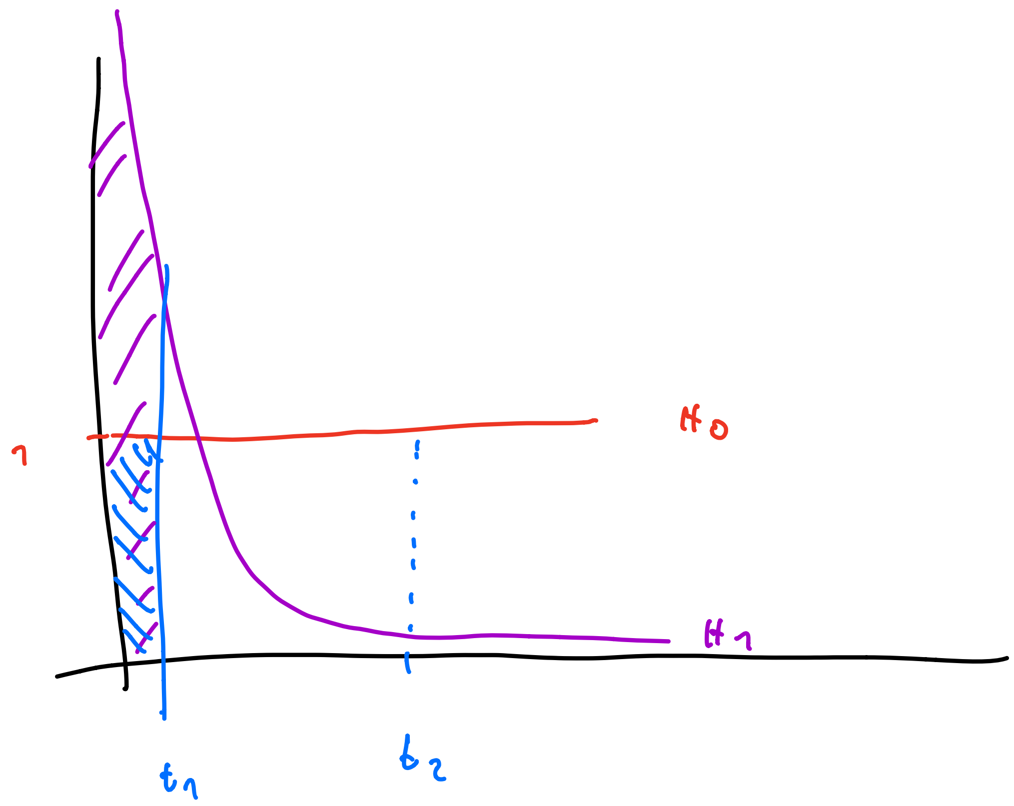
density of p-values under H_0

density of p-values under H_1

If we have some H_0 and some H_1 being true ~~an~~ then the density would maybe look like this:



Goal: set threshold t such that FDR satisfies what we want.



Integral of the pink area: Expected number of pvalues corresponding to H_1 that are below t_1

Integral of blue area: ... H_0

By moving t from 0 to 1 we control the FDR:

For t_1, the FDR is small

t_2 large

General Remarks:

- BH tends to have more power than Bonferroni
- BH controls FDR, not FWER (overall type-I-error)!

- BH works best in sparse regime where only few tests reject the null

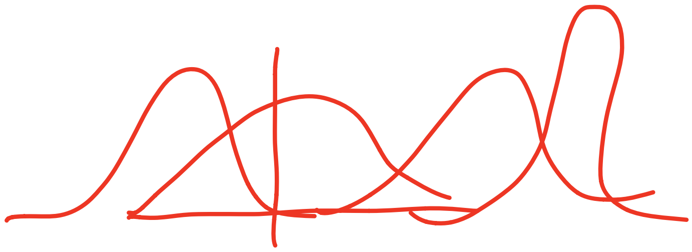
- BH gives guarantees on FDR, but in general does not minimize it.

- When all the H_0 are true, BH $\approx$ Bonferroni.

Non-parametric tests

Standard (parametric scenario):

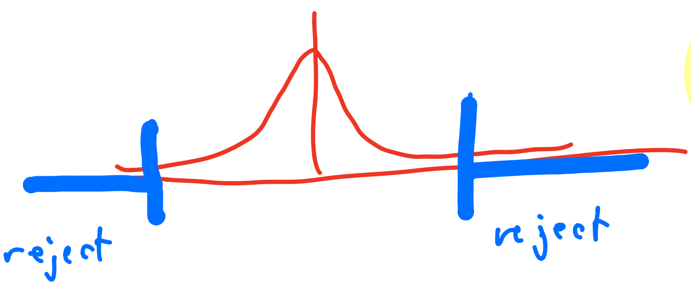
- Statistical model $\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \}$



distribution of the samples

- Observe data, compute a test statistics T_n, for example the mean $\bar{X}$

- Need to know the distribution of the test statistics T_n under the null distribution:



distribution of T_n under the null hyp.

- Construct rejection threshold / region: if the observed T_n is in this region, reject the null hyp.

Kolmogorov–Smirnov test for goodness-of-fit

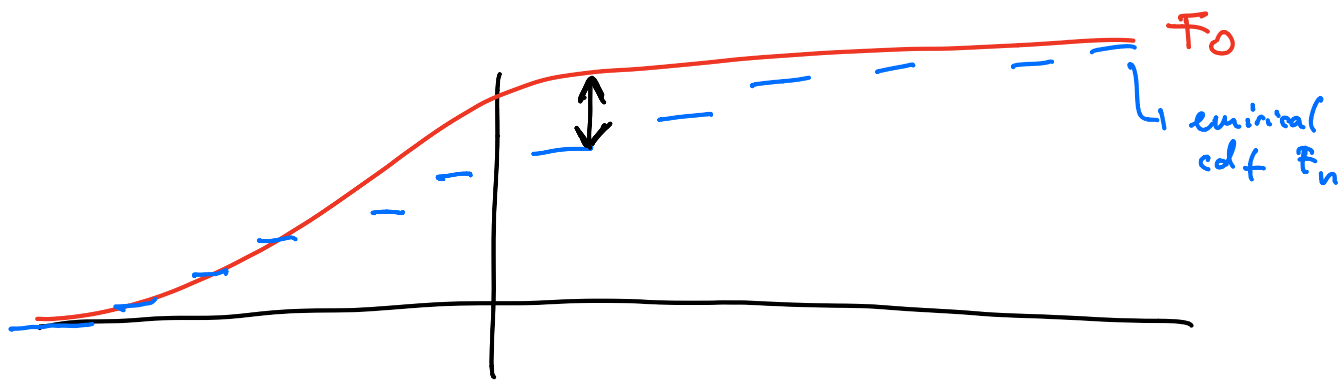
Goodness-of-fit tests: Goal is to test whether a data set comes from a particular distribution F_0

H_0: $F(x) = F_0(x)$

true distribution that generated the data

H_1: $F(x) \neq F_0(x)$

Kolmogorov-Smirnov: We consider the cdf



F_0 = cdf of the given distribution

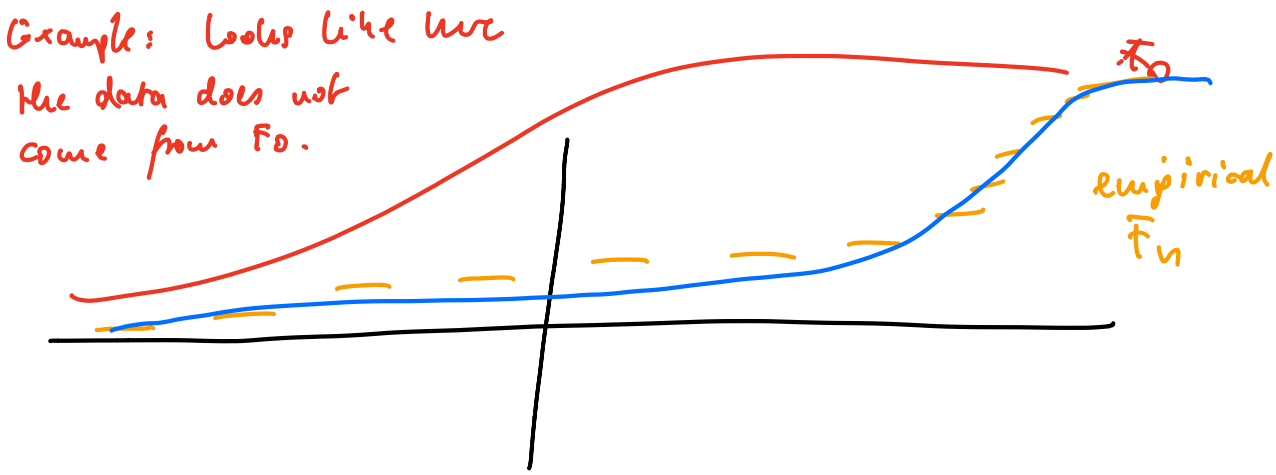
F_n = cdf of the data

$$D_n := \sup_{x \in \mathbb{R}} | F_n(x) - F_0(x) |$$

By the Glivenko–Cantelli theorem we know that under the null hypothesis, $F_n \to F_0$ uniformly a.s.

It is possible to compute the distribution of D_n, and it is independent of F_0, it just depends on n.

From this we can compute rejection thresholds and design a test.



Wilcoxon - Mann - Whitney test

(two sample test based on ranks)

<u>Two sample test</u>: $X_1, \dots, X_n \sim F_1$ a first sample distributed according to F_1,

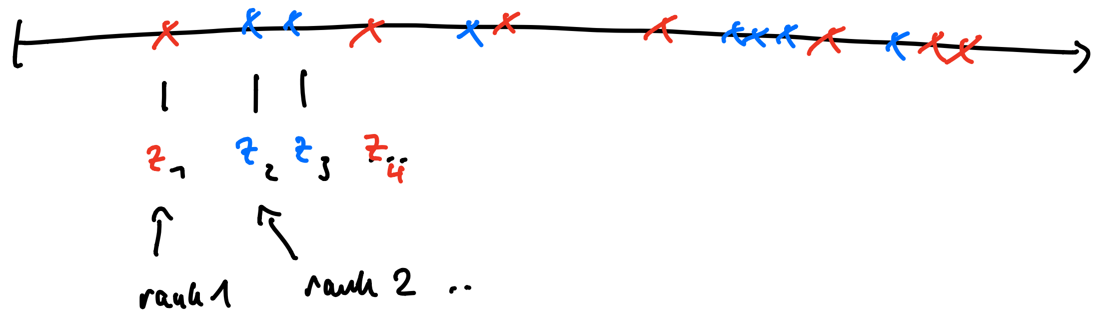
$Y_1, \dots, Y_m \sim F_2$ a second sample distributed acc. to F_2.

Question: $F_1 = F_2$?

$H_0: F_1 = F_2$, $H_1: F_1 \neq F_2$

Test: • "Pool the sample": $\underbrace{x_1, \dots, x_n}$, $\underbrace{y_1, \dots, y_m} \in \mathbb{R}$

• Sort the pooled sample in increasing order and retrieve the rank of all points $\rightsquigarrow$ rank(x_i)
rank(y_i)



• Compute the rank sums for both groups:

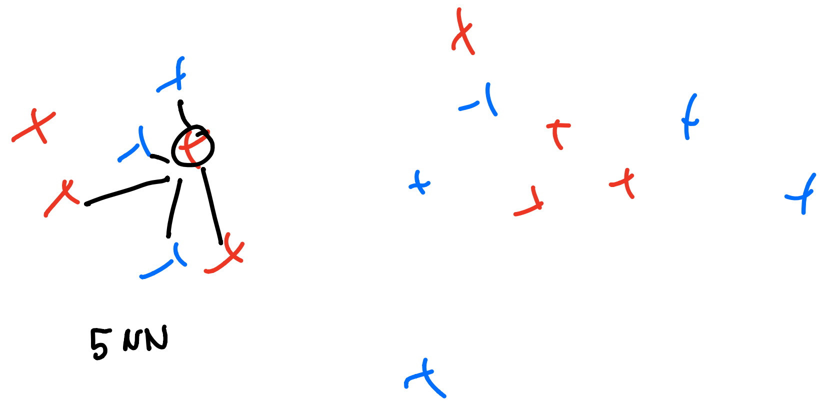
red group: $W_{red} = \sum_{i \in \text{red population}} \text{rank}(x_i)$

$W_{blue} = \sum_{i \in \text{blue pop.}} \text{rank}(y_i)$

• If $|W_{red} - W_{blue}|$ is small, we retain H_0,
if large, reject H_0.

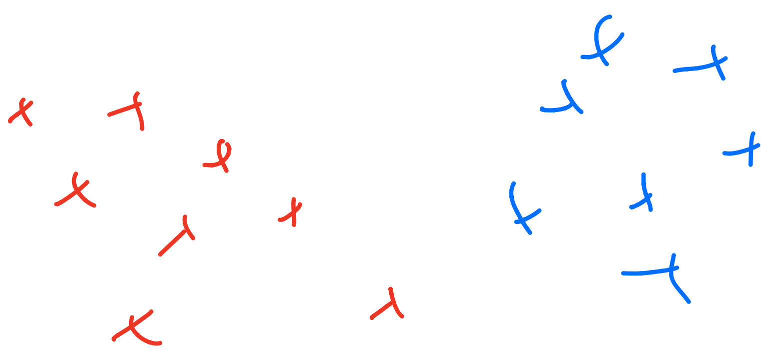
Extension to a multivariate setting using k nearest neighbors

- Two samples, we pool them:

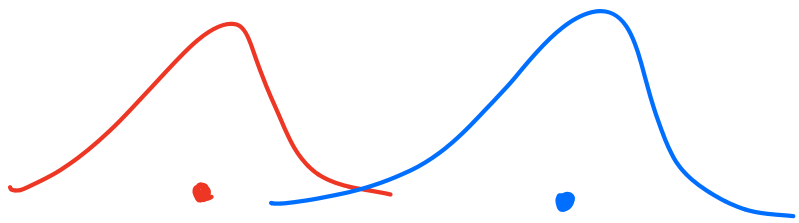


- For each point, we look at the colors of the k nearest neighbors:

- Under the null hypothesis we expect that the number of red neighbors $\approx$ number of blue neighbors.



Permutation (randomization) tests

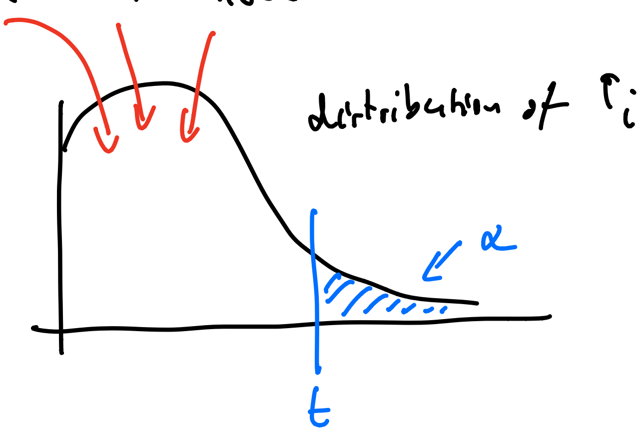


- Sample $X_1, \ldots, X_n$ group A $\rightsquigarrow$ mean $\bar{X}$

  $Y_1, \ldots, Y_n$ group B mean $\bar{Y}$

  ?
- Compute observed statistics $T_{observed}$ = mean (red) − mean (blue)
- Pool the sample
- For $k = 1, \ldots, 10^3$: shuffle the group membership ("colors")
- Compute the difference T_k = mean (red) − mean (blue)

$\rightsquigarrow$ $T_1, T_2, \ldots, T_{1000}$



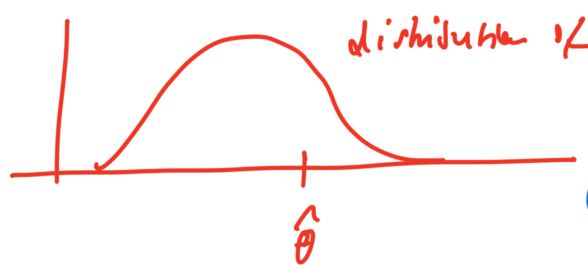
- Find α-quantile to determine rejection threshold.
- Check whether the observed $T_{observed}$ on the true data is $\lesssim t$.

Bootstrap

Motivation: $X_1, \dots, X_n \sim F$, no knowledge on F, want to estimate a parameter $\theta = t(F)$. You generate an estimate $\hat{\theta}$ based on $X_1, \dots, X_n$, want to know how reliable $\hat{\theta}$ is.

The first thing to look at is the standard error se,

- If we have assumptions on F, we can analytically compute the distribution of $\hat{\theta}$, the $\hat{se}$, ... (this is rare!)



(1)

- We could also try to obtain many samples
  - $X_1^{(1)}, \dots, X_n^{(1)}$
  - $X_1^{(2)} \dots X_n^{(2)}$

    $\vdots$

    $X_1^{(m)} \dots X_n^{(m)}$

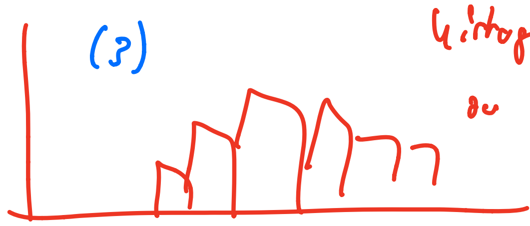
and then estimate the distribution of $\hat{\theta}$:

Problem: need too many samples.

(2)

Idea of the bootstrap:

- Given the sample $x_1, \dots, x_n$ $\rightsquigarrow$ estimate $\hat{\theta}_{orig}$
- Draw a subsample of $x_1 \dots x_n$, compute $\hat{\theta}^*$, repeat very often



histogram based (of θ^*) on resampled data

"Hope: histogram of $\hat{\theta}^*$ (3) is "close" to histogram of $\hat{\theta}$ (2), which is close to (1)

Example: estimate the standard error of an estimate $\hat{\theta}$

Algorithm in pseudocode

Input: $x_1, \dots, x_n$ — number of original sample points

For $b = 1, \dots, B$ — number of bootstrap replications

- Sample $x_1^*, \dots, \underline{\underline{x_n}}^*$ uniformly with replacement from $x_1, \dots, x_n$
- Estimate the parameter $\hat{\theta}_b^*$

gives us $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$

Estimate the standard error $\hat{se}$ of the original estimate $\hat{\theta}$ by the standard dev. of the bootstrap replicates:

$$\hat{se}_B := \left(\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \underbrace{\left(\frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^* \right)}_{\text{mean of replicates}} \right)^2 \right)^{1/2}$$

Does it always work?

<u>Theorem</u> (Consistency of the estimate of the standard error)

- Assume that $X_1, \dots, X_n \sim F$, iid, and $E(\|X_1\|^2) < \infty$.
- Let $\hat{\theta}_n = g(X_1, \dots, X_n)$ be the parameter that we estimate. Assume that g is continuously differentiable in a neighborhood of $\mu = EX_1$, with a non-zero gradient.

Then the bootstrap estimate of the standard error is strongly consistent.

<u>Example where it goes wrong:</u>

$X_1, \dots, X_n \sim \text{Uniform}[0, \theta]$, where $\theta \in [0, 1]$ unknown.

Want to estimate θ. The ML estimate of θ is simply the largest number we observe:

$$\hat{\theta} = \max_{i=1 \dots n} X_i .$$

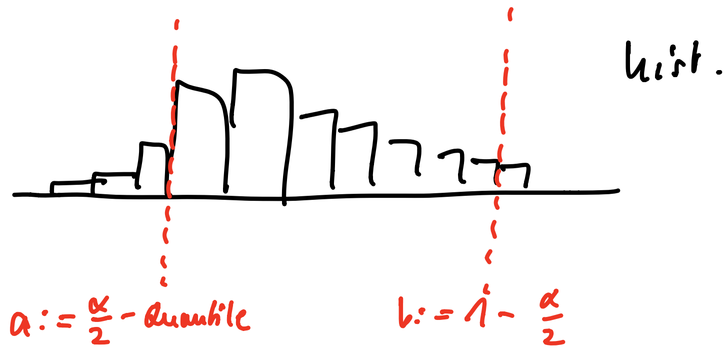
Estimating the $\hat{se}$ by bootstrap is going to fail.

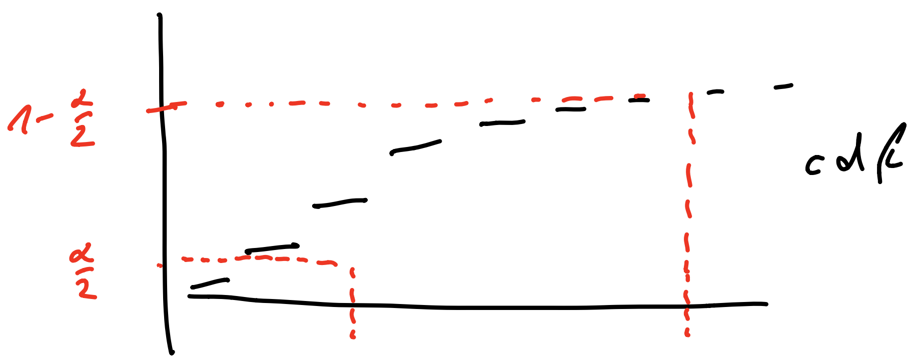
Estimating tails or extreme values by bootstrap is problematic.

Confidence sets by bootstrap

Bootstrap-percentile-method:

- Given sample $X_1, \dots, X_n$, estimate $\hat{\theta}$
- Generate bootstrap replicates $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$
- look at the histogram of the $\hat{\theta}_b^*$





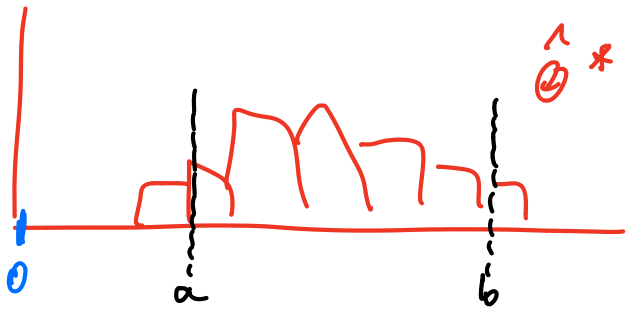
- $CI = [a, b]$

  It has coverage $1 - \alpha$ because

  $$P_\theta(\hat{\theta} \in CI) \geq 1 - \alpha$$

  (approximately, because n, B finite)

Subsequently you can construct bootstrap tests in the obvious way



$H_0: \hat{\theta} = 0$ vs $H_1: \hat{\theta} \neq 0$

Bayesian statistics

Frequentist statistics:

- probability = limiting frequency
- parameters θ are constants, we cannot assign probabilities to them
- statistics behave well when repeated often

Bayesian statistics

- probability = degree of belief
- parameters do have probabilities
- have a prior belief about the world, update it based on observed data.

Bayesian statistics:

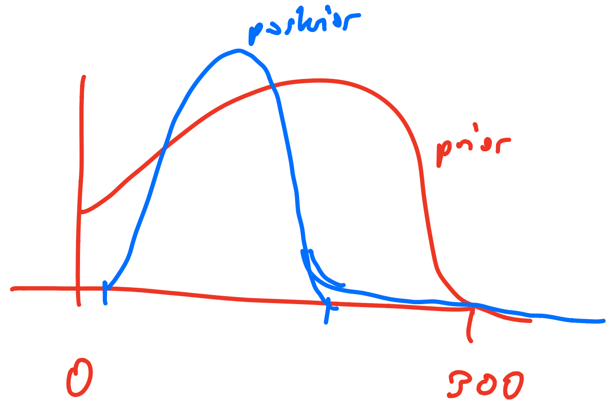
- Assume a statistical model $\{f_\theta \mid \theta \in \Theta\}$,

  we call $f(x \mid \theta)$ the likelihood of the data given the parameter θ

  density data parameter

- Goal: Investigate θ

- We assume that we have a prior belief about the parameters θ: $f(\theta)$ prior distribution



- Observe data $X_1, \dots, X_n$ iid

- Now we update our belief, we compute the posterior using Bayes rule: $f(\theta | x_1 \dots x_n)$

$$\overbrace{f(\theta | x_1 \dots x_n)}^{\text{posterior}} = \frac{\overbrace{f(x_1 \dots x_n | \theta)}^{\text{likelihood}} \cdot \overbrace{f(\theta)}^{\text{prior}}}{\underbrace{\int f(x_1 \dots x_n | \theta)\, f(\theta)\, d\theta}_{\substack{\text{normalizing constant} \\ \text{(does not depend on } \theta \text{ any more)}}}}$$

The posterior is a distribution.

- Now you can make statements based on the posterior.
  - If you want to return one "best guess" for θ, you could use
    - max of posterior (MAP)
    - mean of posterior

- You can construct confidence intervals:

  find a, b such that

  $$P(\theta \in [a, b]) = 95\%.$$

Advantages:

- easy to interpret
- natural way to incorporate prior knowledge

Disadvantages

- analytic solutions are rare, typically you have to solve computationally hard problems
- need to choose a prior