Part V:

Statistics

Literatur: Wassermann: all of statistics

Point estimates,
bias & variance,
consistency

Standard setup in parametric statistics

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

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

$\theta = (\mu, \sigma)$

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

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

one particular parameter

space of all possible parameters

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

Conventions

Parameter space Θ ("capital theta")

True (unknown) parameter θ ("lower case theta")

P_θ, E_θ, ... refers to the probability, expectation under the distribution f_θ

Estimates typically get a "hat": $\hat{\theta}$, $\hat{\mu}$, ...

Point estimation

The goal of point estimation is to estimate θ.

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

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

Bias of an estimator

Def The bias of such an estimator is defined as

$$\text{bias}(\hat{\theta}_n) := \underbrace{E_\theta}_{} (\underbrace{\hat{\theta}_n}_{\text{estimate}}) - \underbrace{\theta}_{\text{true para}}$$

expectation wrt the distribution f_θ (the true one!)

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

An estimate is unbiased if its bias is zero.

Variance and standard error

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

Example

$$X_1, \dots, X_n \sim \overset{\{0,1\}}{\overset{\shortparallel}{\text{Bernoulli}}}(p), \quad \text{parameter } \overbrace{p}^{\theta} \in \overbrace{[0,1]}^{\Theta},$$

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

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

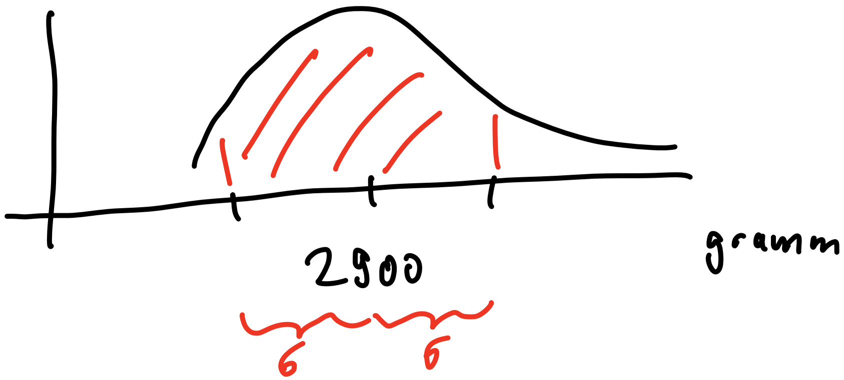
The standard error of this estimate is

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

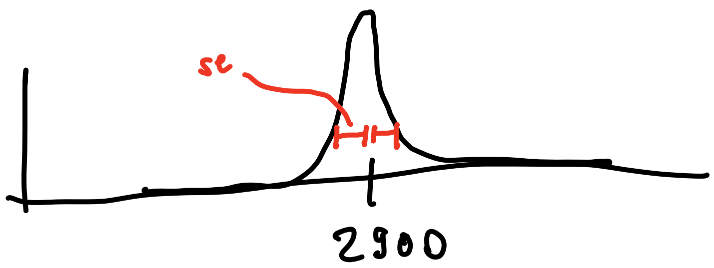
We can for example estimate it by

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

Example: weight of baby



Distribution of individual data pts: $\mu = 2900$, $\sigma = 500$



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

Mean squared error

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

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

deterministic

Bias-Variance decomposition

<u>Theorem</u> : bias-variance-decomposition

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

Proof

$$E_\theta\left((\hat\theta_n - \theta)^2\right) =$$

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

$$= E_\theta\left(\overbrace{(\hat\theta_n - E\hat\theta_n)^2}^{a^2}\right) + \underbrace{2E_\theta\left(\overbrace{(\hat\theta_n - E\hat\theta_n)}^{a}\underbrace{\overbrace{(E\hat\theta_n - \theta)}^{b}}_{\text{deterministic}}\right)}_{2(E\hat\theta_n - \theta)\cdot \underbrace{E_\theta(\hat\theta_n - E\hat\theta_n)}_{= E_\theta(\hat\theta_n) - E_\theta\hat\theta_n = 0} \;=\; 0} + E_\theta\left(\overbrace{(E\hat\theta_n - \theta)^2}^{b^2}\right)$$

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

$\blacksquare$

Example

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

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.

$$\hat{\sigma}_1^{\,2} := \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 \qquad \text{"first estimate"}$$

$$\hat{\sigma}_2^{\,2} := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 \qquad \text{"second estimate"}$$

$$E(\hat{\sigma}_1^2) = \frac{n-1}{n}\sigma^2$$ so the bias is $\frac{1}{n}\sigma^2$

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

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

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

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

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

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

Consistent estimator

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

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

as $n \to \infty$

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

Confidence sets

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 $a_n = a(X_1, \dots, X_n)$, $b_n = b(X_1, \dots, X_n)$ are functions of the sample $X_1, \dots, X_n$ such that

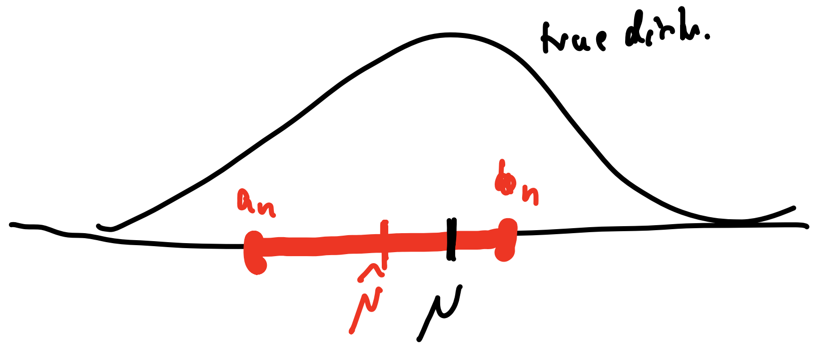
$$P_{\theta}(\underbrace{\theta}_{\text{deterministic}} \in \underbrace{C_n}_{\text{random}}) \geq \underbrace{1-\alpha} \quad \text{for all } \theta \in \Theta.$$

true (unknown) parameter

$(1-\alpha)$ is called the coverage of the confidence interval.

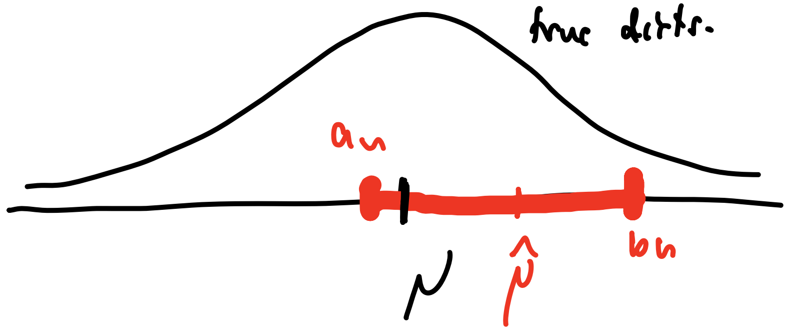
Illustration

First experiment
$x_1, \dots, x_n \rightsquigarrow \hat{\mu}$



First confidence set:

second experiment
$x_1, \dots x_n \rightsquigarrow \hat{\mu}$



second confidence set

... 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, define

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

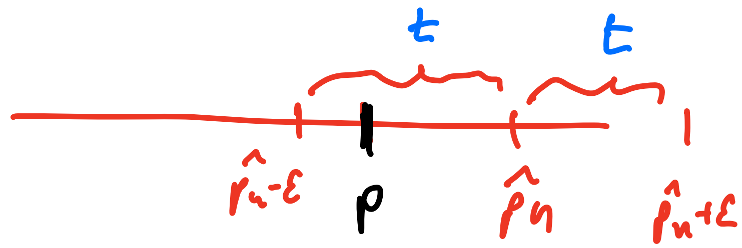
Proposition: $C_n := \left(\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n \right)$ is a CI with coverage $1-\alpha$.

Proof (example)

Proof: By Hoeffding inequality, for any t we have

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

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



and solve for t:

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

Chose $\varepsilon_n = t$. $\blacksquare$

Maximum likelihood estimator

Motivating example

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

adjacency matrices of graphs

Observe k random walks from the graph of length 10.

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

Likelihood

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

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

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

notation!

Maximum likelihood

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

$$\hat{\theta} := \underset{\theta \in \Theta}{\arg\max}\ P(x_1, \dots, x_n \mid \theta) \overset{\text{ind.}}{=} \underset{\theta}{\arg\max} \prod_{i=1}^{n} P(x_i \mid \theta)$$

This is equivalent to the problem

$$\hat{\theta} = \arg\max \log\left(\prod_{i=1}^{n} P(x_i \mid \theta) \right) = \arg\max \sum_{i=1}^{n} \underbrace{\log \underbrace{P(x_i \mid \theta)}_{\in [0,1]}}_{< 0}$$

which is equivalent to minimizing the negative log-likelihood:

$$\hat{\theta} = \underset{\theta}{\arg\min} \sum_{i=1}^{n} \underbrace{- \log P(x_i \mid \theta)}_{> 0}$$

= maximum-likelihood estimator **MLE**

Solving max. likelihood problems

Sometimes this optimization problem is easy:

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

Example for an analytic solution

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

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

$Var(X) = \lambda$.

Observe $x_1, \dots, x_n \sim \text{Poisson}(\lambda)$

Want to construct the ML-estimator for λ.

Compute the likelihood:

$$\mathcal{L}(\lambda) = P(x_1, \dots, x_n \mid \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Example (continued)

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

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

Now want to optimize for λ. Take the derivative (wrt λ):

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

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

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

MLE properties

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:

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

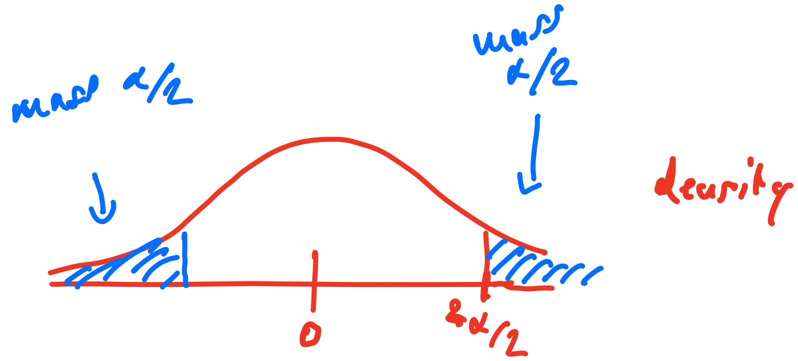
$$\frac{\hat{\theta}_{MLE} - \theta}{\hat{se}} \xrightarrow{\text{in distr.}} N(0,1)$$

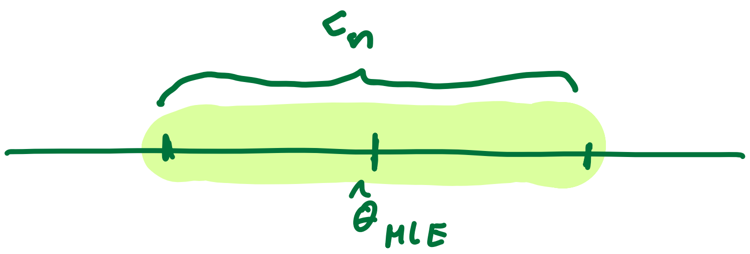
(5) This can be used to construct (approximate) confidence intervals:

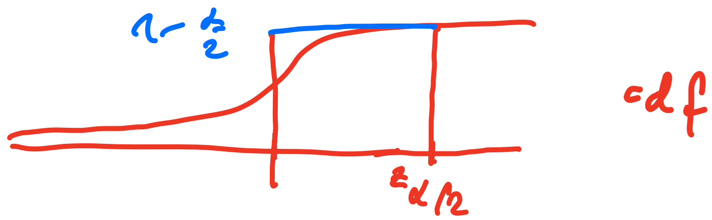
$$C_n := \left(\hat{\theta}_{MLE} - \underbrace{z_{\alpha/2} \, \hat{se}}_{-\varepsilon} \, , \; \hat{\theta}_{MLE} + \underbrace{z_{\alpha/2} \cdot \hat{se}}_{+\varepsilon} \right)$$

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

Φ: cdf of $N(0,1)$







C_n is an "approximate CI" in the sense that

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

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?

If yes we would like to call the statistics $T(X_1, \dots, X_n)$

"sufficient".

Sufficiency

Which properties would we need to assert sufficiency?

- when we observe two samples $x_1, \dots, x_n$ and $x_1', \dots, x_n'$, and $T(x_1, \dots, x_n) = T(x_1', \dots, x')$, then we would infer the same θ.
- when we know $T(x_1, \dots, x_n)$, then we would need some way to estimate θ just based on θ.

Formal definition is technical, skipped.

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 $\mathcal{F}$:

$$\theta \neq \theta' \Rightarrow f_\theta \neq f_{\theta'}$$

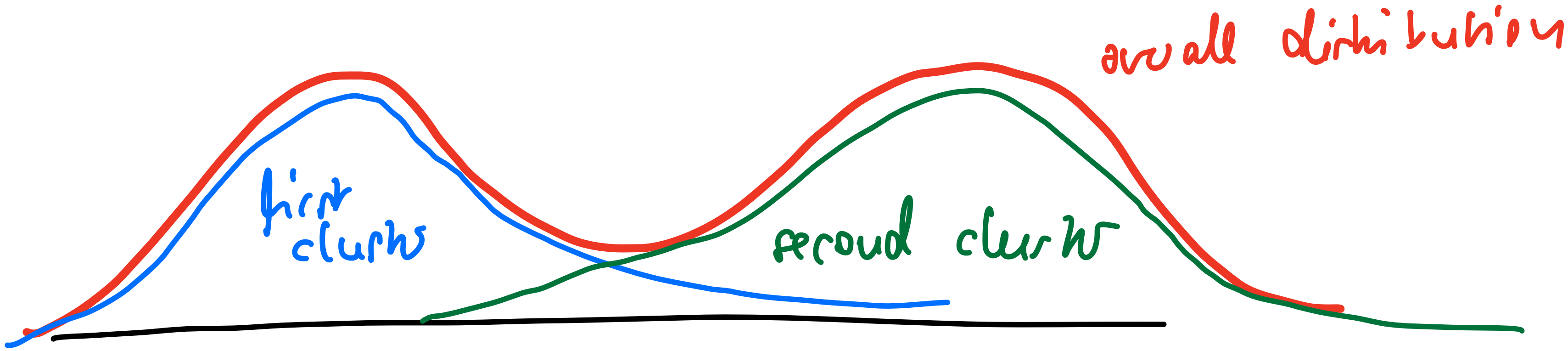
(Identifiability is a property of the model itself, not of the data)

Example (identifiability)

The problem of clustering consists in finding "groups" in data.
From a statistical point of view, this is equivalent to the problem of decomposing the underlying distribution as a mixture, say

$$p = \alpha_1 p_1 + \alpha_2 p_2$$

mixing weights



Example (identifiability)

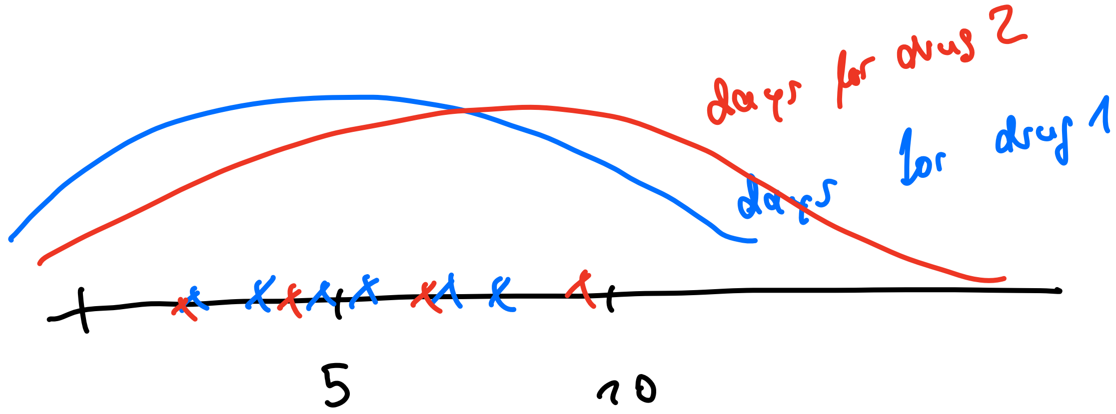
Depending on our statistical model, such patterns might be identifiable or not. Often, parametric models are identifiable, for example mixtures of Gaussians.

Non-parametric models, for example implicit in kernel algorithms, are often not identifiable.

Hypothesis testing

Motivation

Example: Two drugs D_1, D_2, we measure number of days to recovery for both drugs: $x_1, \dots, x_n$ treated with D_1
$x'_1, \dots, x'_n$ D_2

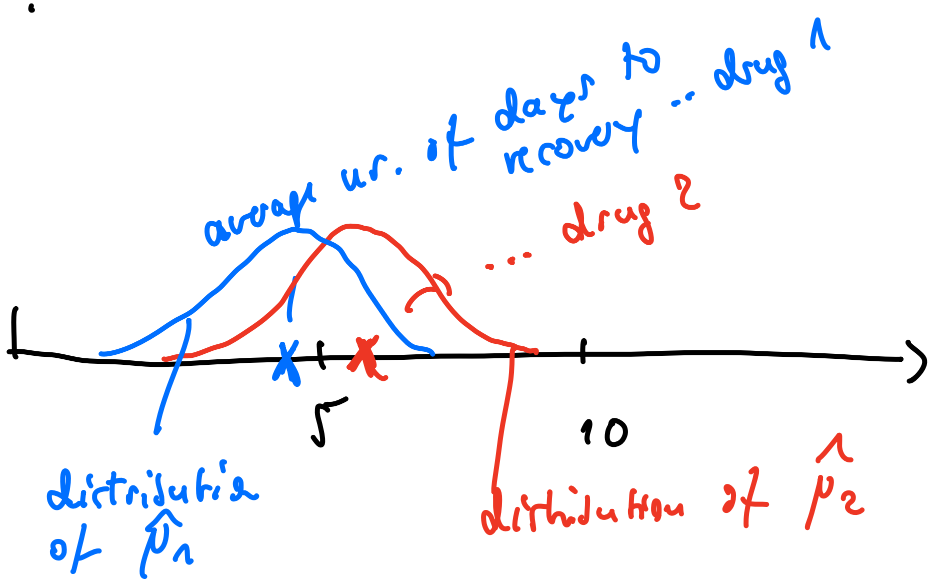


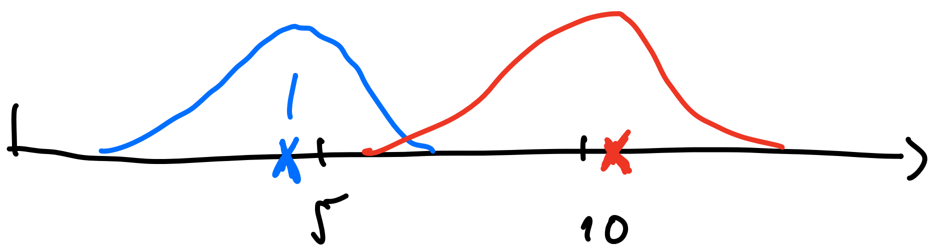
Question: is Drug 1 better than Drug 2 ?

General idea

... will be: consider the distributions of the estimates $\hat{\mu}_1$, $\hat{\mu}_2$.

If they are "far apart", we would say that they are different...





But how do we know what "far apart" is?

__Example__ Want to test whether a coin is fair.

__Null hypothesis:__ H_0 : coin is fair

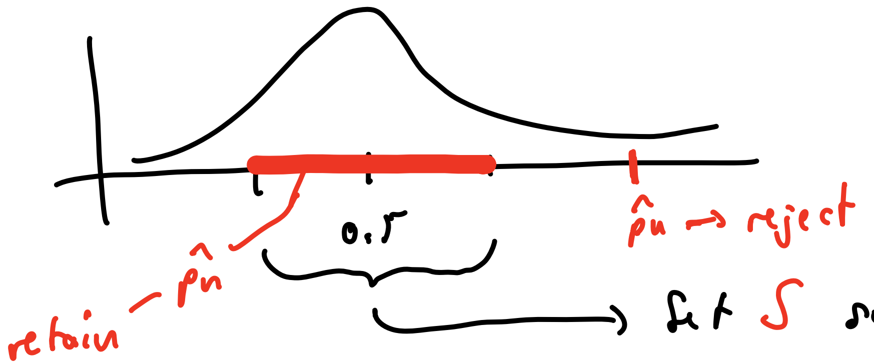
__Alternative hypothesis:__ H_1 : coin is unfair

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:



$\longrightarrow$ 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

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

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 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	☺

Roles of H_0 vs H_1

⚠ The role of the two hypotheses is not "symmetric".

It is like a legal trial: we believe H_0 unless there is strong evidence against it.

So if a test does not reject H_0, it does not imply that H_0 is correct.

Power of a test, β

Def The **power function** of a test with rejection region R is the function

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

- If $\theta_0 \in \Theta_0$ then $T(x)$ should not end up in R.
  For such θ_0, $\beta(\theta_0) = P(\text{Type I error})$.
  Ideally, $\beta(\theta_0)$ should be small.
- If $\theta_1 \in \Theta_1$ then we hope that $T(x) \in R$. So
  $\beta(\theta_1) = 1 - P(\text{Type II error})$. Here, $\beta(\theta_1)$ should be large.

Level of a test

Def 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: $\alpha \mathrel{\hat{=}}$ type I error)

Power of a test

We always specify the power of a test against a <u>fixed</u> alternative parameter $\theta_1 \in \Theta_1$.

The power of test against alternative θ_1 is given as

$$1 - \beta(\theta_1)$$

(Intuition: $\beta \mathrel{\hat{=}} 1 - P(\text{type-II-error})$)

Standard approach for testing

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.

Uniformly most powerful test

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

$H_0: \theta \in \Theta_0$ vs $H_1: \theta \notin \Theta_0$.

A test in $\mathcal{T}$ with power function $\beta(\theta)$ is

uniformly most powerful (UMP) if

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

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.

Example: summer or winter?

Assume we want to test whether it currently is summer or winter. Say, we typically believe it is summer, unless evidence is against it:

$$H_0 = \{\text{summer}\}$$
$$H_1 = \{\text{winter}\}$$

Now we construct several tests based on the average temperature of the current day (1 measurement)

Example: summer or winter?

Test 1: we look at the temperature outside. Whenever it is below 5° degrees, we reject H_0 and accept H_1.

Type-I-error: how often does it get below 5° in summer?
Say only in 5% of all summer days.
P(Test rejects H_0 | H_0 is true)

Type-II-error: how often does it get higher than 5° in winter? Say, on 50% of the days.

P(Test rejects H_1 | H_1 = true) = 50% !

Not so great.

Example: summer or winter?

Test 2: So maybe 0 is a bad threshold.

Reject H_0 if $T \leq 0°$

- Reduces type-I-error, but still increases type-II-error

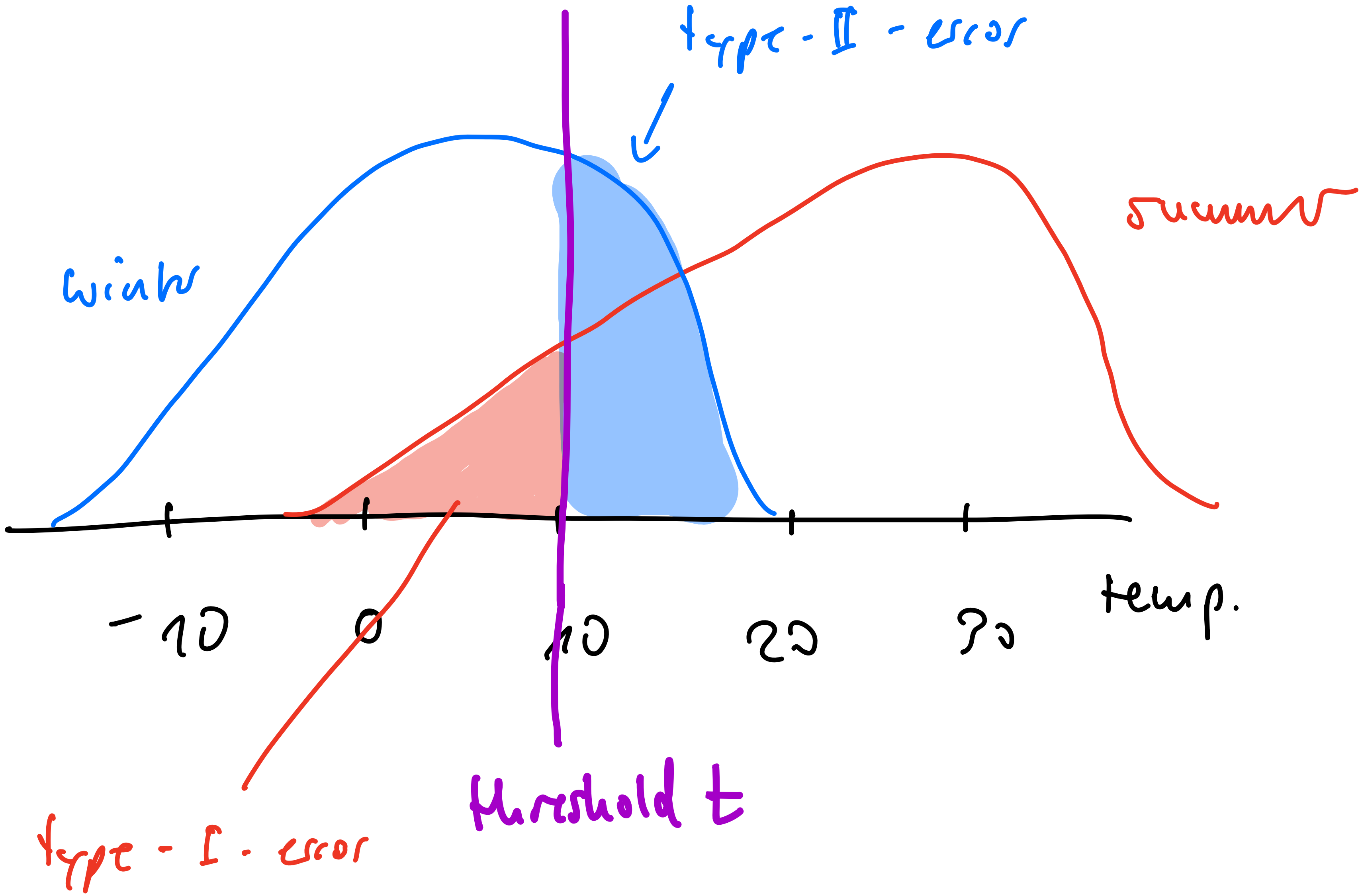
Test 3: Reject if $T \leq +10°$.

- increases type-I-error
- Reduces type-II-error.

Cannot really construct a good test, why?

- Temperature distributions are overlapping
- just 1 measurement

temp.



type - II - error
winter
summer
-10
0
10
20
30
temp.
threshold t
type - I - error

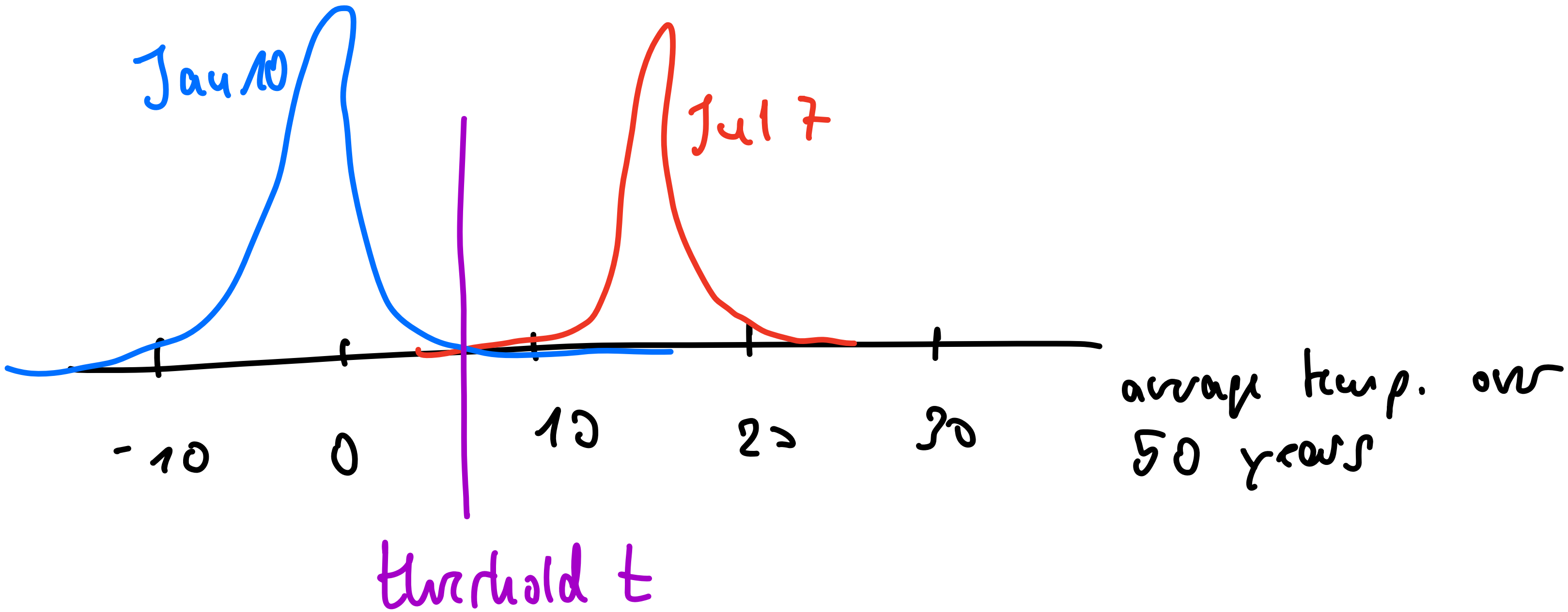
Example: summer or winter?

Often, to increase the power of a test we need more sample pts.

Need to be independent!

Say, I fix a date, July 7, and average over the temperatures on this day over the last 50 years.

And same for winter, Jan 10.



Example: summer or winter?

Test 4 Look outside the window. If somebody with skis is walking by, reject H_0 (summer) and accept H_1 (winter).

Type-I-error: say, 1% (perhaps somebody just bought skis in summer)

Type-II-error: 99% as well! (even if it is winter, don't see skiers in front of my window)

$\rightsquigarrow$ low power!

Binomial example

We throw 5 coins with $P(\text{"1"}) = \theta \in [0,1]$.

Want to test $H_0: \theta \leq \frac{1}{2}$ vs $H_1: \theta > \frac{1}{2}$

Test 1 : reject H_0 if we observe 5 times 1 : 1 1 1 1 1

Power function: $\beta(\theta) = P_\theta(\text{reject}) = \theta^5$

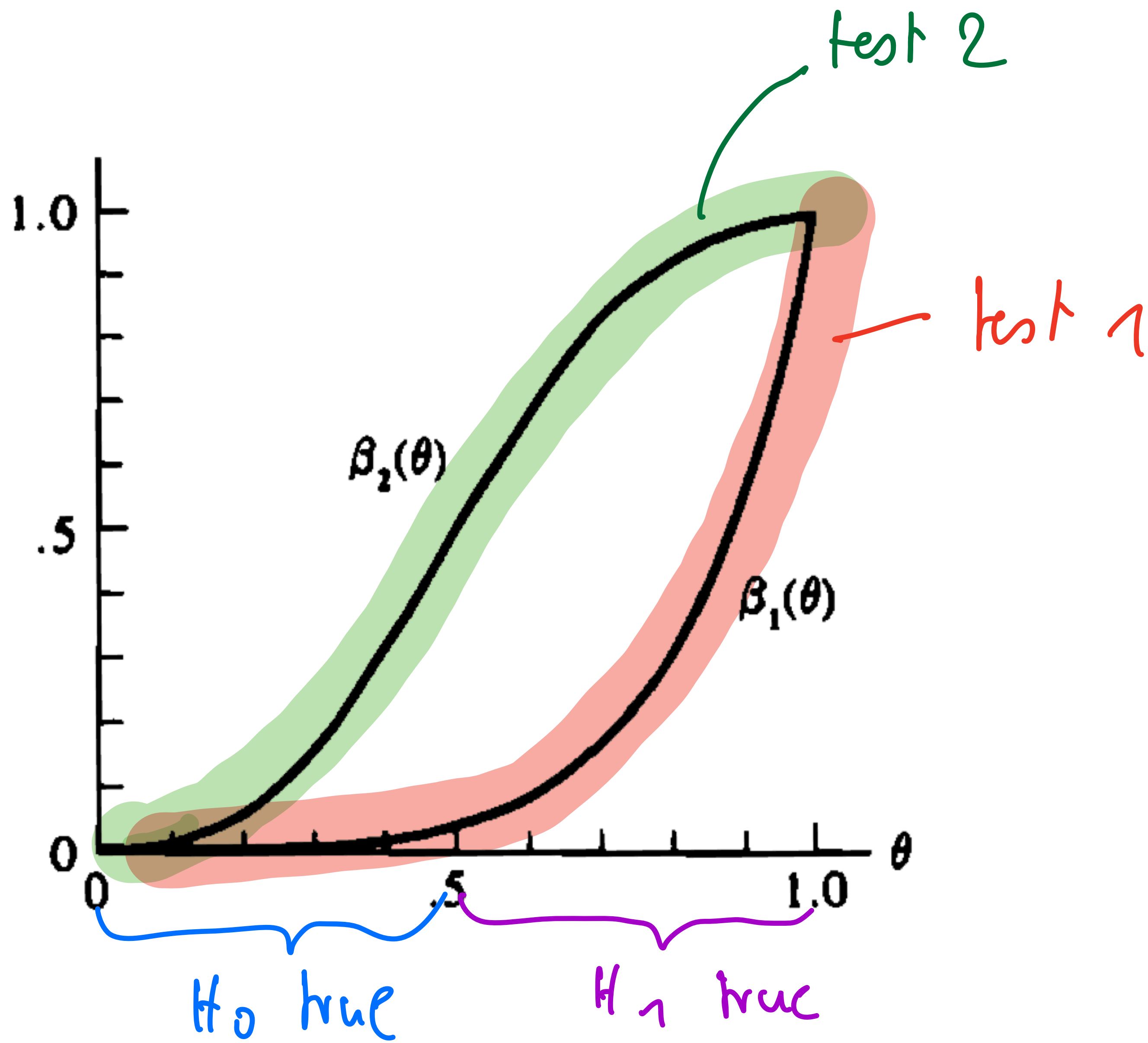
Test 2 : reject H_0 if we observe at least 3 times 1 : 0 0 1 1 1 or
0 1 0 1 1 or ...

Power fct : $\beta(\theta) = P_\theta(3,4,5 \text{ times } 1) =$

$$= \binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta) + \theta^5$$

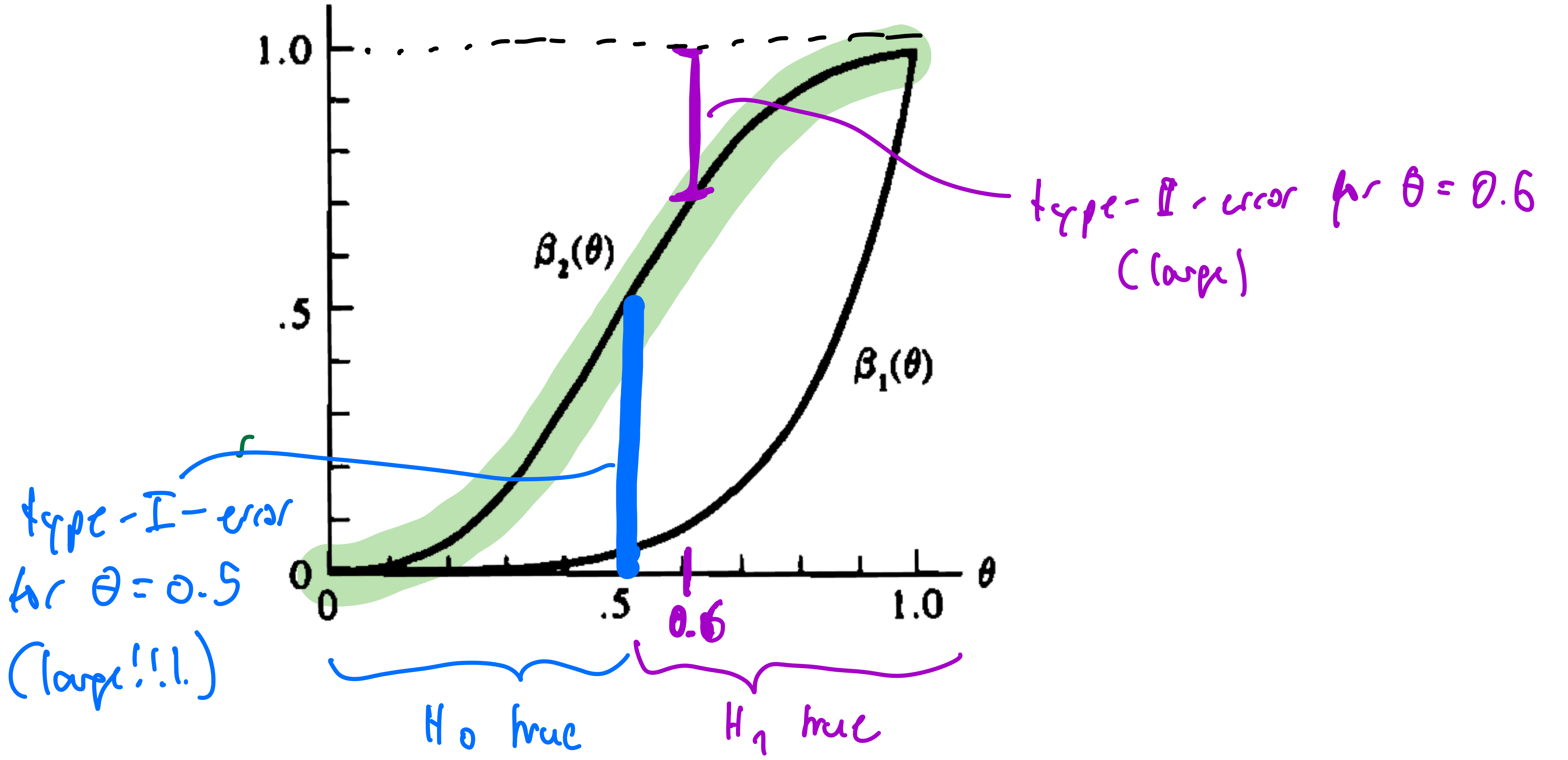
Binomial example

Power functions of both tests: $\beta(\theta) = P_\theta(\text{reject } H_0)$



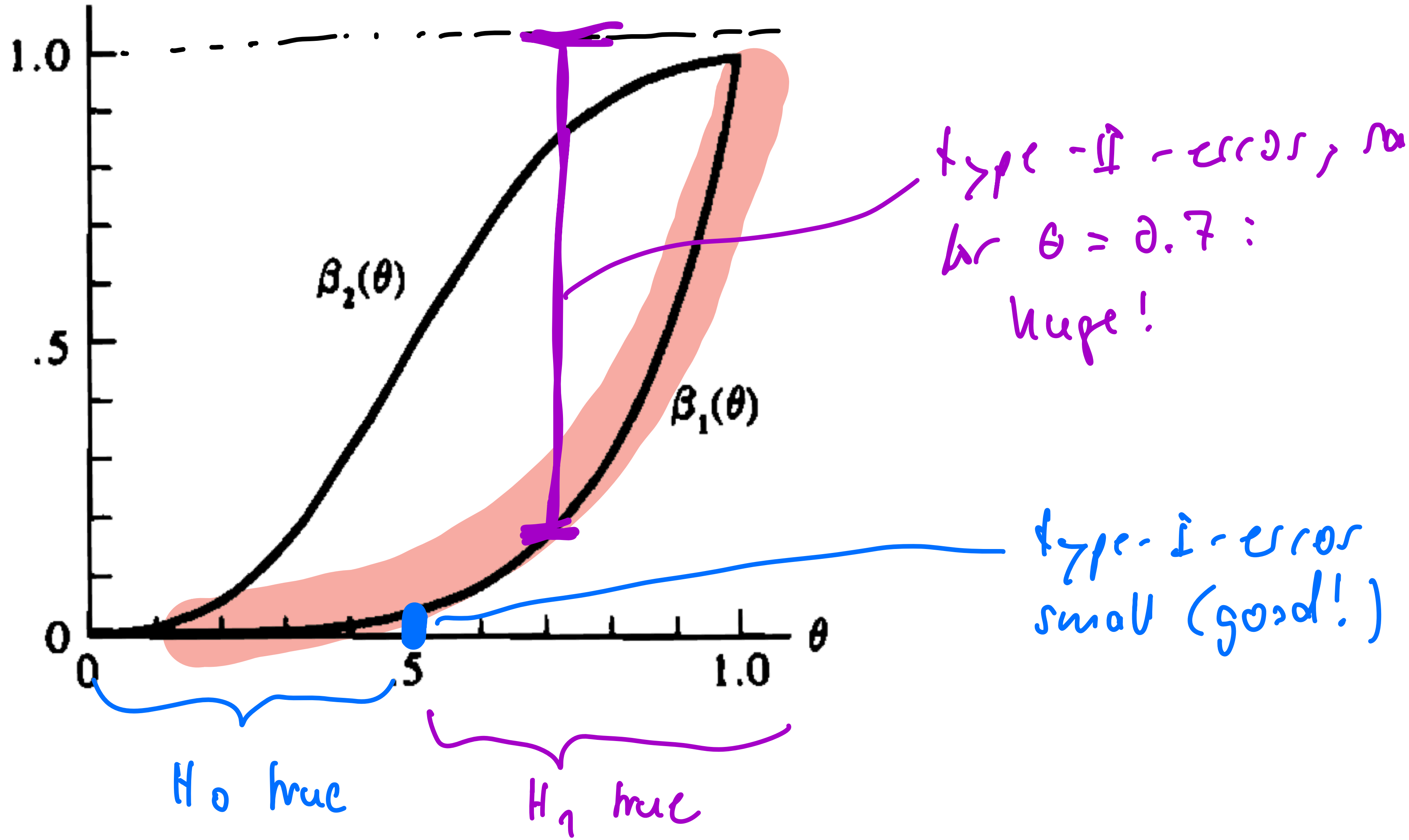
Binomial example

Power functions of both tests : $\beta(\theta) = P_\theta$ (reject H_0)



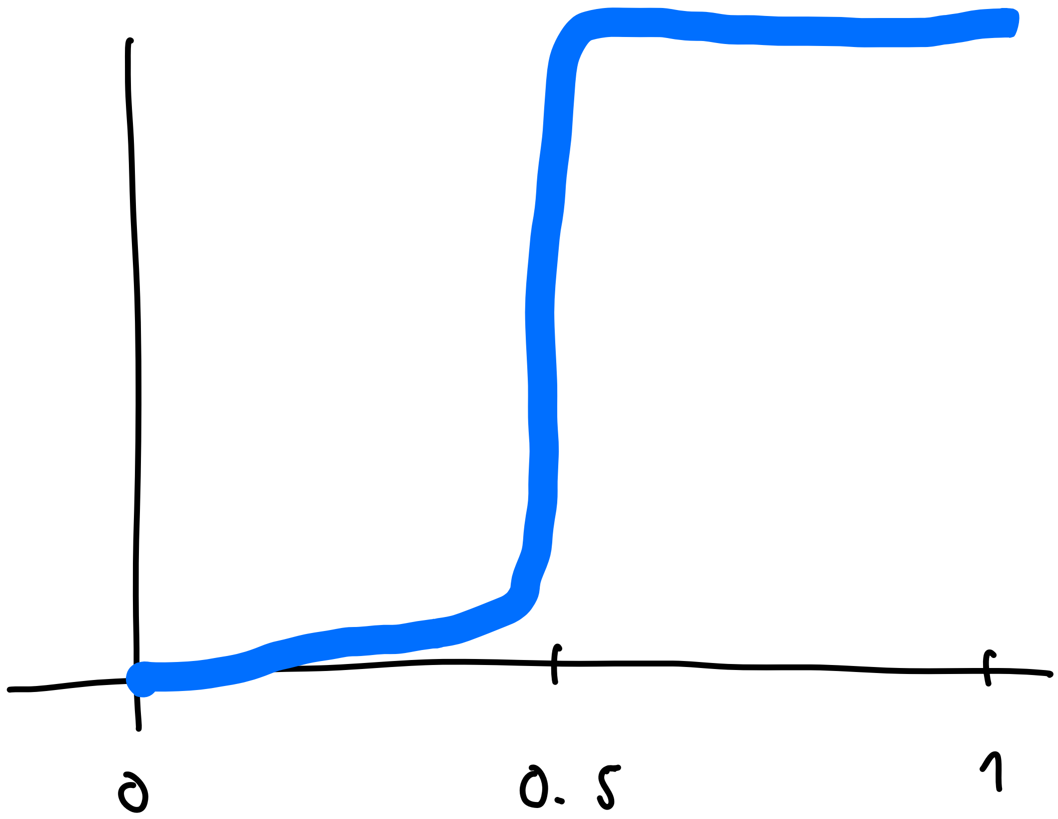
Binomial example

Power functions of both tests



Binomial example

Would like to have such a power fct:

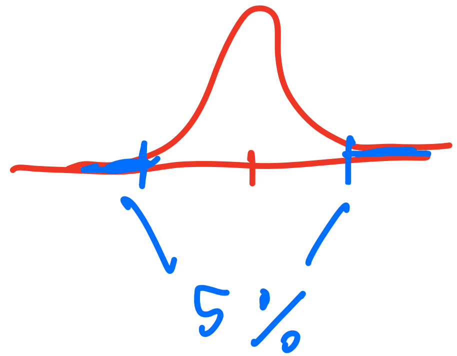


How could we achieve this? Increase the sample size...

z - Test (Wald test)

If the distribution of the test statistic is asymptotically normal:

$$\frac{\hat{\theta} - \theta_0}{\hat{se}} \rightsquigarrow N(0,1)$$



$H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$

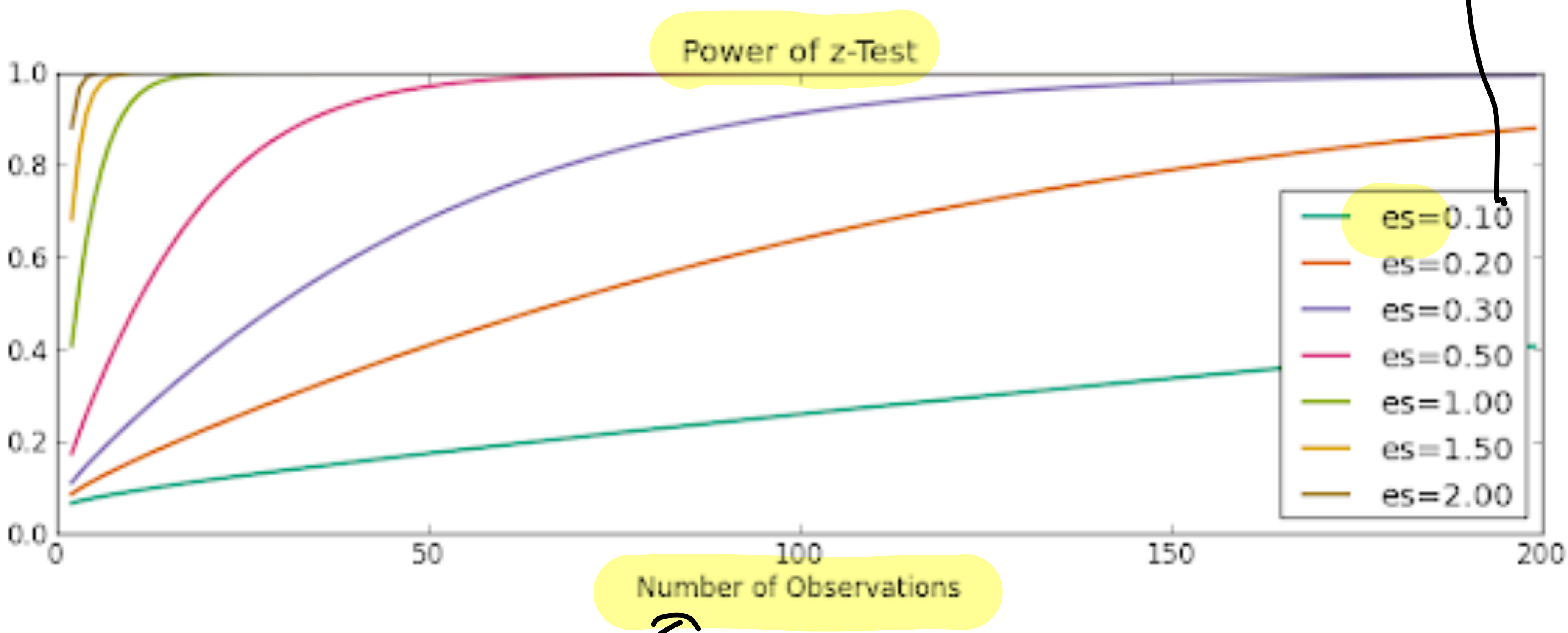
Test: reject when $\left| \frac{\hat{\theta} - \theta_0}{\hat{se}} \right| > z_{\alpha/2}$ where

$z_{\alpha/2}$ is the real nr. s.t. the normal cdf attains value $\frac{\alpha}{2}$.

z - Test (Wald test)

Power of test increases with sample size:

effect size $|\theta - \theta_0|$

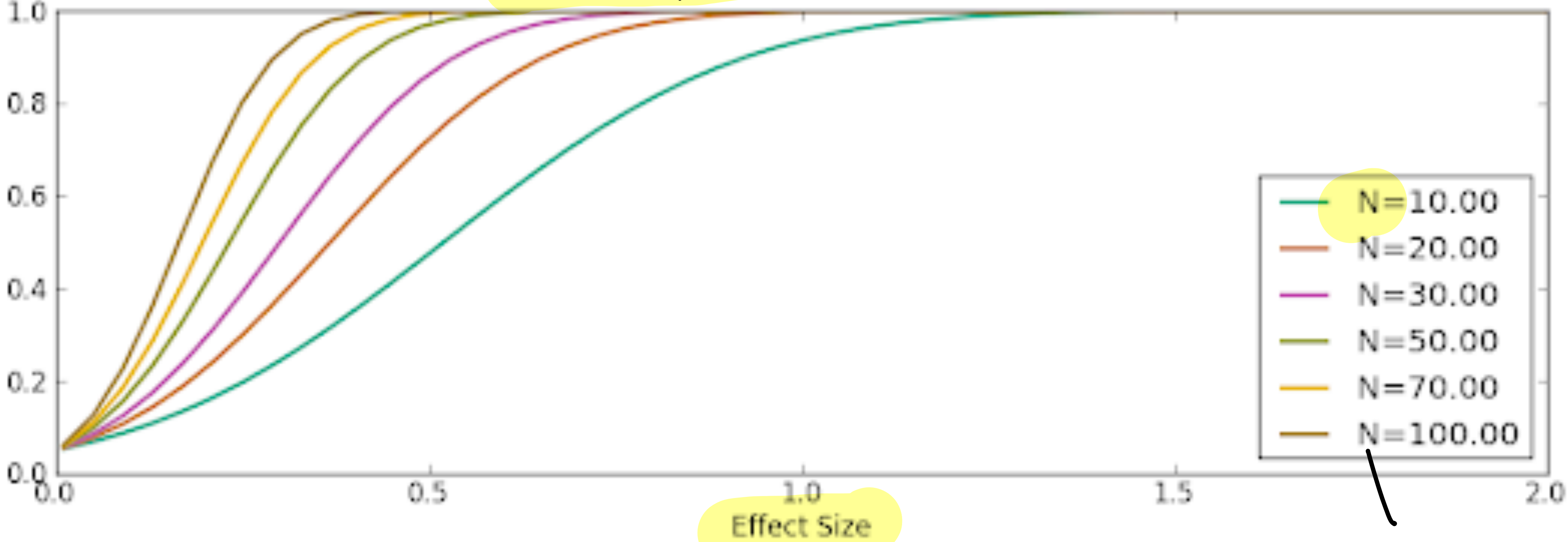


sample size

z - Test (Wald test)

Power of test increases with Effect size $|\theta - \theta_0|$:

Power of z-test



sample size

Neyman - Pearson - lemma
&
likelihood ratio tests

Neyman - Pearson

<u>Theorem</u> Suppose we test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

Consider

$$T = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{\prod_{i=1}^{n} f(x_i \mid \theta_1)}{\prod_{i=1}^{n} f(x_i \mid \theta_0)}$$

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

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

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

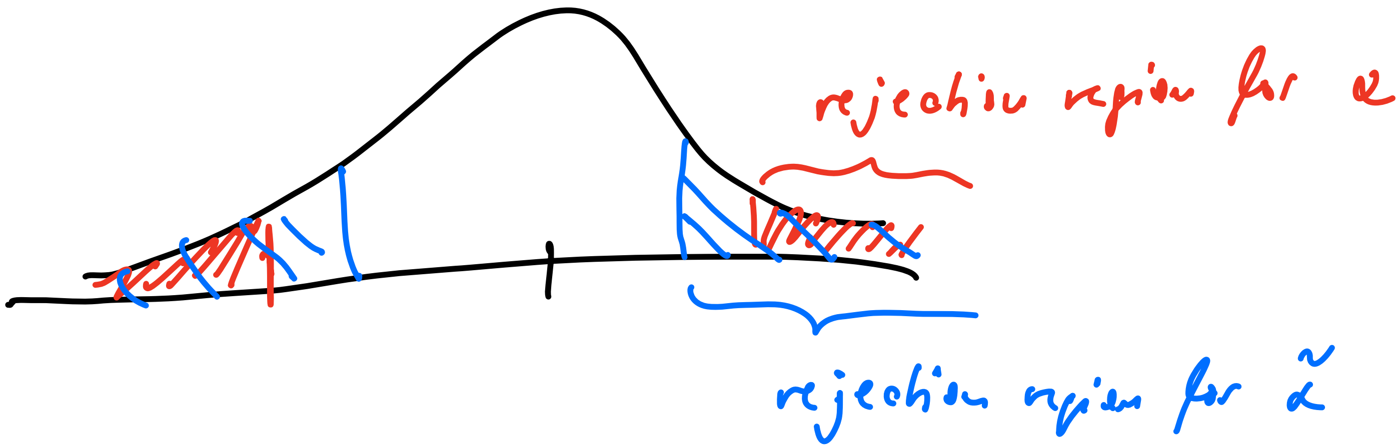
p-values

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 $\alpha < \tilde{\alpha} \Rightarrow R_\alpha \subset R_{\tilde{\alpha}}$)



p-value

<u>Def</u> The p-value is defined as

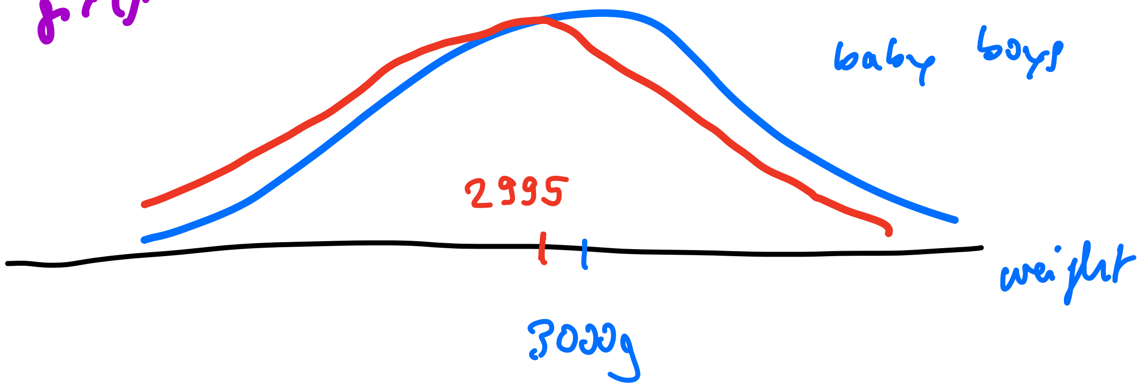
$$p = \inf \{ \alpha \mid T(X_1, \dots, X_n) \in R_\alpha \}$$

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

Example baby boys and girls

sample distribution

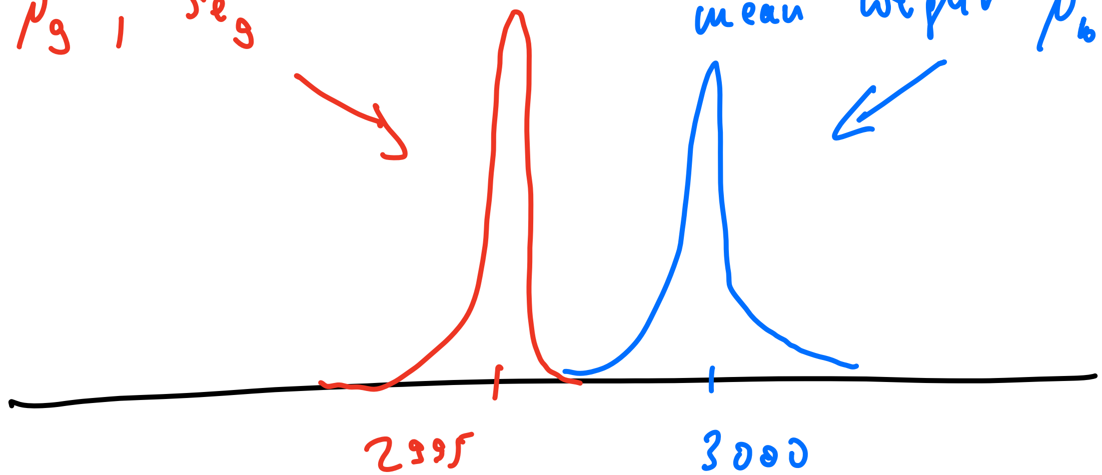


Sample many baby girls, many baby boys

$\hat{\mu}_g$, $\hat{se}_g$

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

distribution of test statistics



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

Statistically significant

Convention: $p \leq 0.05$ → "significant"

$p \leq 0.01$ → "highly significant"

⚠ Even though an effect might be "statistically significant", it might not be "scientifically significant".
For example, the effect size might be tiny

⚠

- p-value does <u>not</u> give the probability that the null-hypothesis is true (Θ is not a rv in the frequentist setup)

- We can have small p-values, yet the "effect size" can be tiny.

- A large p-value is not necessarily strong evidence for H_0. It could just mean that our test has little power.

Neyman - Pearson - lemma

&

likelihood ratio tests

Neyman-Pearson

<u>Theorem</u> Suppose we test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

Consider

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

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

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

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

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

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

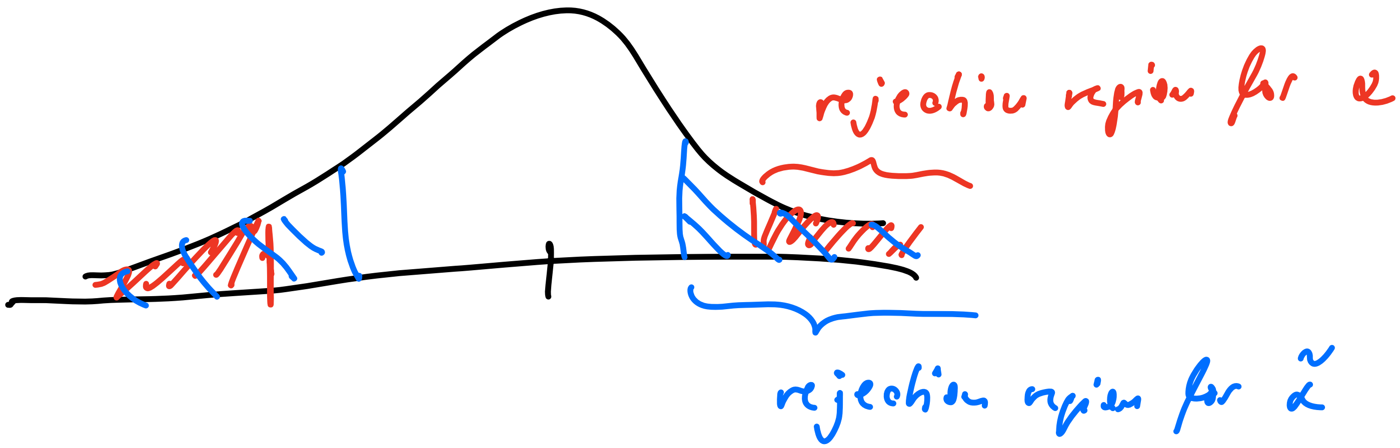
p-values

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 $\alpha < \tilde{\alpha} \Rightarrow R_\alpha \subset R_{\tilde{\alpha}}$)



p-value

<u>Def</u> The **p-value** is defined as

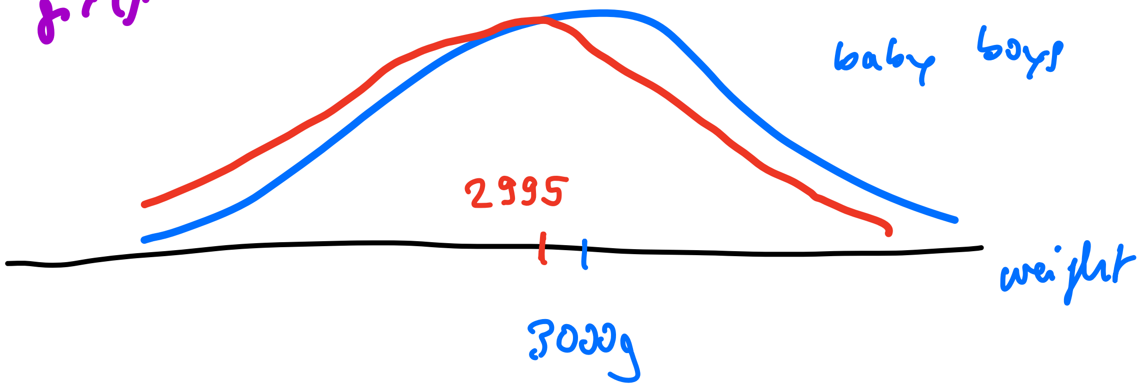
$$p = \inf \{ \alpha \mid T(x_1, \dots, x_n) \in R_\alpha \}$$

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 distribution

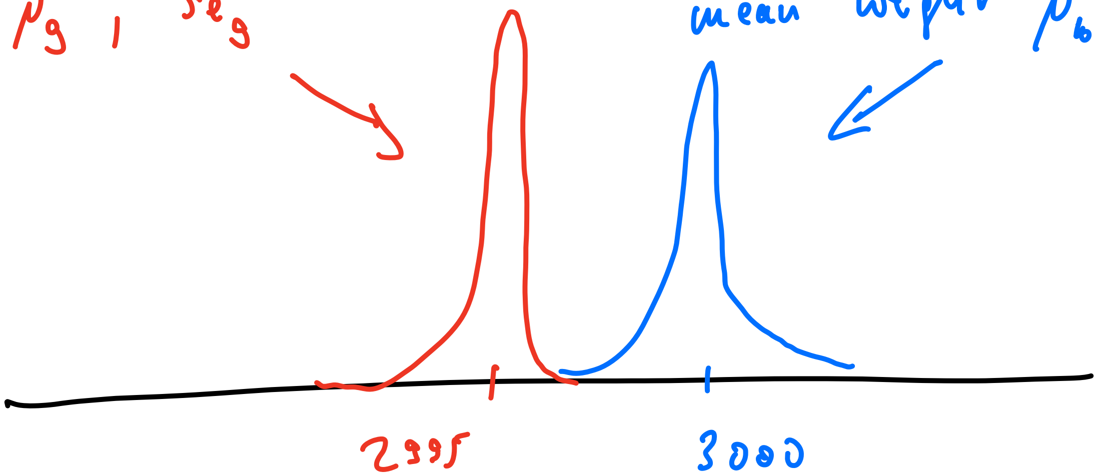


Sample many baby girls, many baby boys

$\hat{\mu}_g$, $\hat{se}_g$

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

distribution of test statistics



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

Multiple testing

Motivation

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 <u>or</u> t_2 error <u>or</u> .. <u>or</u> t_m makes error$)$

$= 1 - P($no error in t_1 <u>and</u> no error in t_2 <u>and</u> ...$) =$ assume independence

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

$m = 1 \Rightarrow * = 0.05$

$m = 10 \Rightarrow * = 0.40$

$m = 50 \Rightarrow * = 0.92$

Many "wrong" tests!

Family-wise error rate (FWER)

<u>Definition:</u> 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:

$$\text{FWER} = P(t_1 \text{ makes type-I-error } \underline{\text{or}}$$
$$t_2 \ \dots$$
$$\dots$$
$$t_m \text{ makes type-I-error}).$$

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:

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

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

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

Bonferroni, discussion

Bonferroni controls the FWER.

Advantage: simple, correct

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

the test barely discovers anything!

False discovery rate, FDR

Def Assume we have a family of m tests. We call

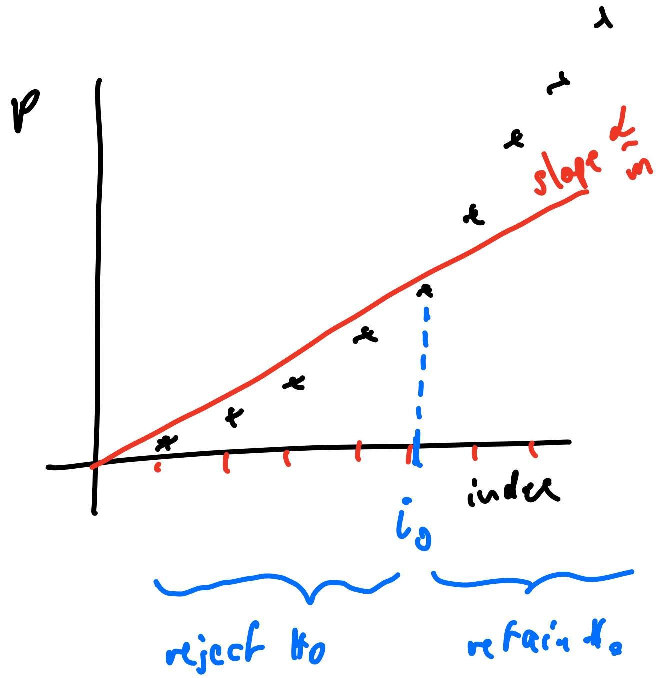
$$E\left(\frac{\#\text{ false rejections}}{\#\text{ all rejections}}\right) =: \text{FDR}$$

the false discovery rate.

Benjamini/Hochberg: Controlling FDR

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.



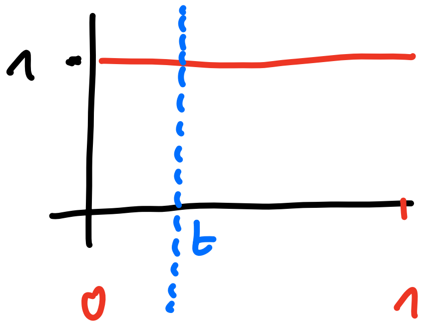
Theorem (Benjamini–Hochberg)

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 $\mathrm{FDR} \leq \alpha$.

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

Intuition

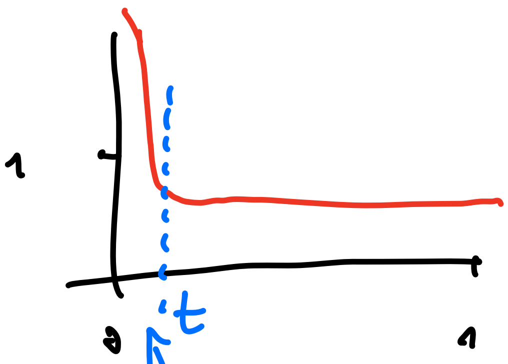
- Under the null hypothesis, the p-values always have a uniform distribution on [0, 1].



density of p-values under H_0

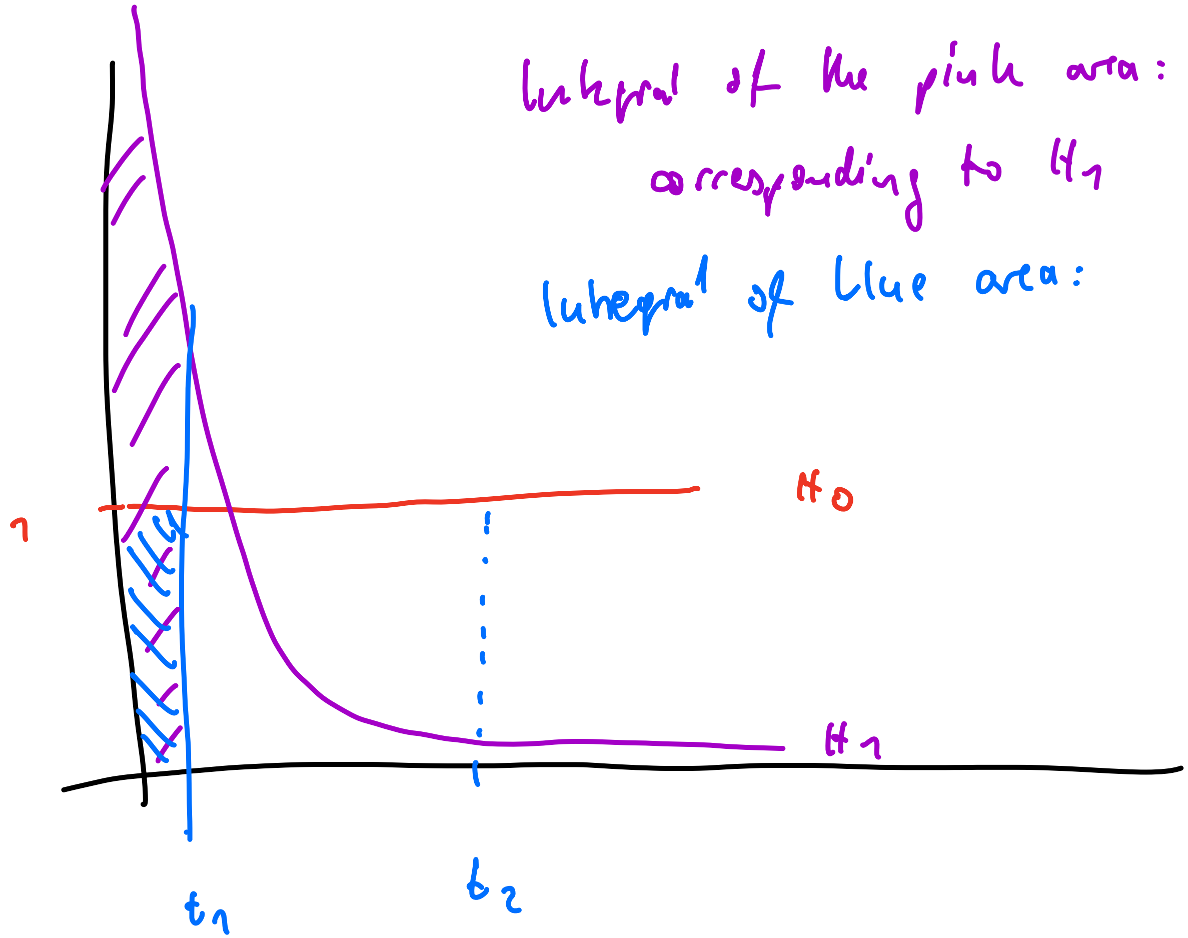
density of p-values under H_1

If we have some H_0 and some H_1 being true ~~an~~ then the dens.h would maybe look like that:



here we have (hopefully) many of the H_1 s but we also have some H_0 s.

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

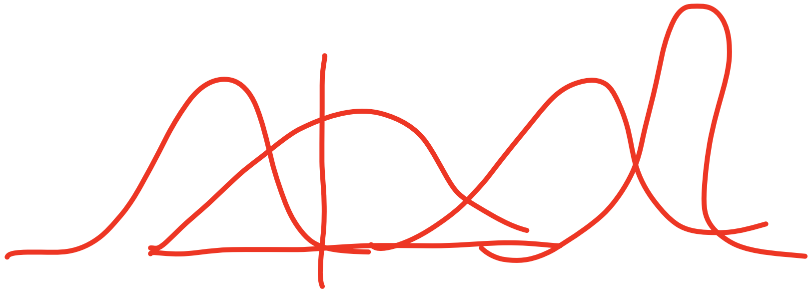
- BH tends to have more power than Bonferroni
- BH controls FDR, not FWER (overall type-I-error)!
- BH works best in sparse regime where only few tests reject the null
- BH gives guarantees on FDR, but in general does not minimize it.
- When all the H_0 are true, BH $\approx$ Bonferroni.

Non-parametric tests

Non-parametric tests

Standard (parametric scenario):

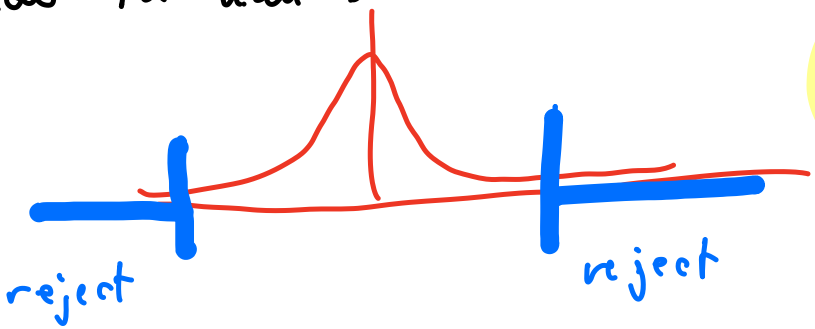
- Statistical model $\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \}$



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.

Goodness-of-fit test (gof)

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

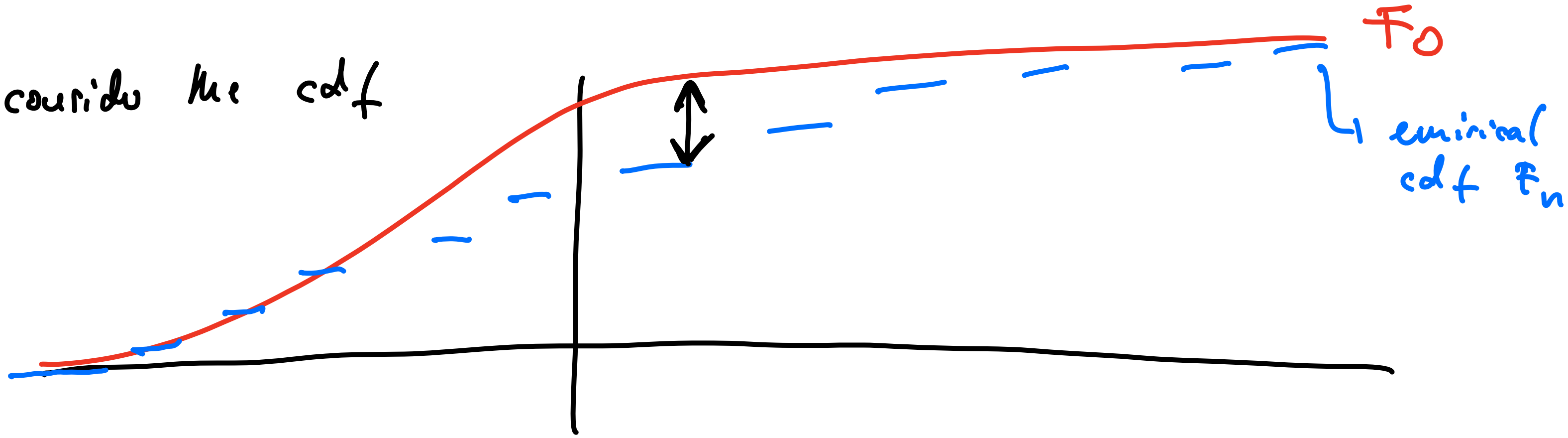
H_0: $F(x) = F_0(x)$

true distribution that generated the data

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

Kolmogorov-Smirnov test for g.o.f

We consider the cdf



F_0 = cdf of the given distribution

F_n = cdf of the data

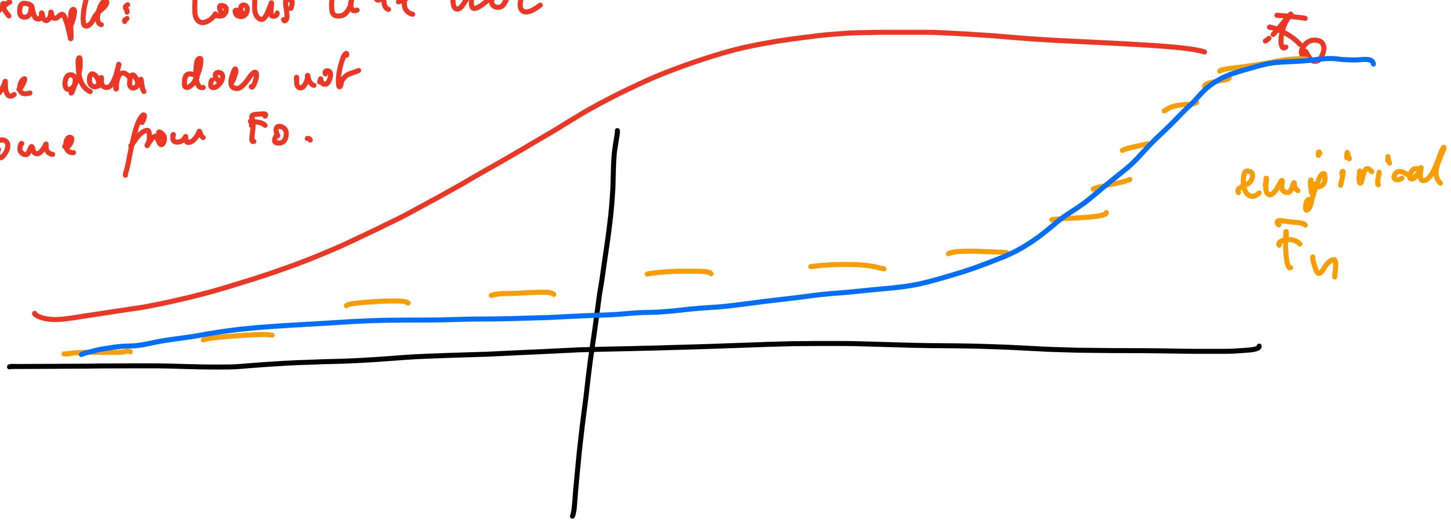
$$D_n := \sup_{x \in \mathbb{R}} \left| F_n(x) - F_0(x) \right|$$

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.

Example: Looks like here the data does not come from F_0.



two sample test

Two sample test: $X_1, \dots, X_n \sim F_1$ a first sample distributed according to F_1, $Y_1, \dots, Y_m \sim F_2$ a second sample distributed acc. to F_2.

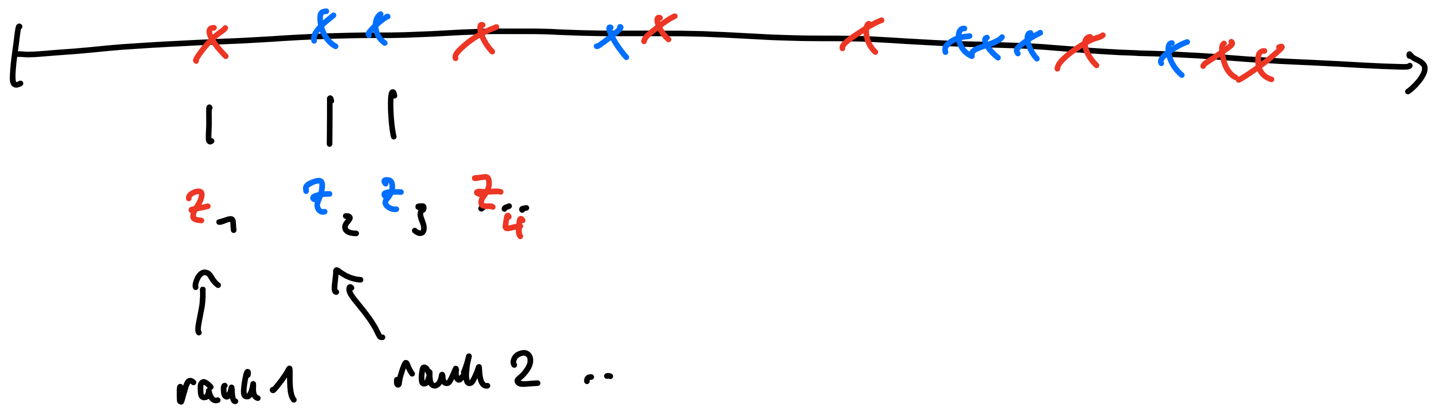
Question: $F_1 = F_2$?

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

Wilcoxon - Mann - Whitney test (based on ranks)

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

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



- Compute the rank sums for both groups:

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

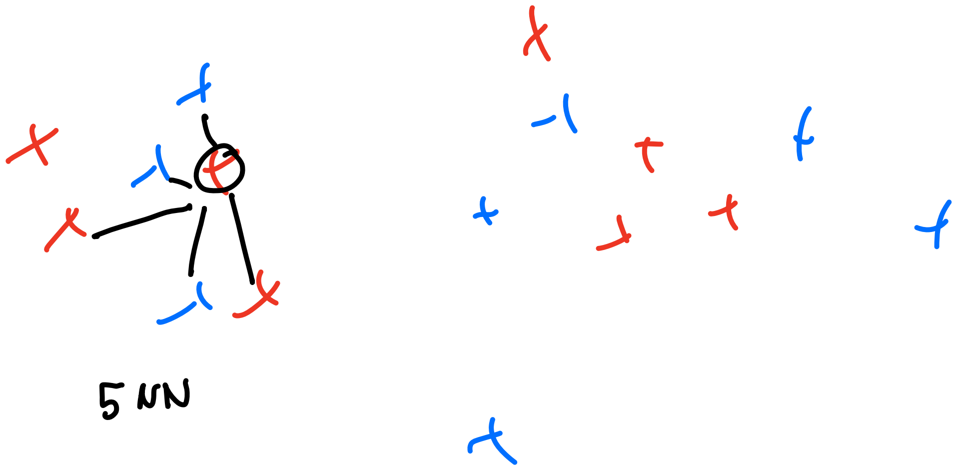
$$W_{blue} = \sum_{i \in \text{blue pop.}} \text{rank}(Y_i)$$

- If $|W_{red} - W_{blue}|$ is small, we retain H_0,

  if $|W_{red} - W_{blue}|$ is large, reject H_0.

Extension to a multivariate setting using k nearest neighbors

- Two samples, we pool them:



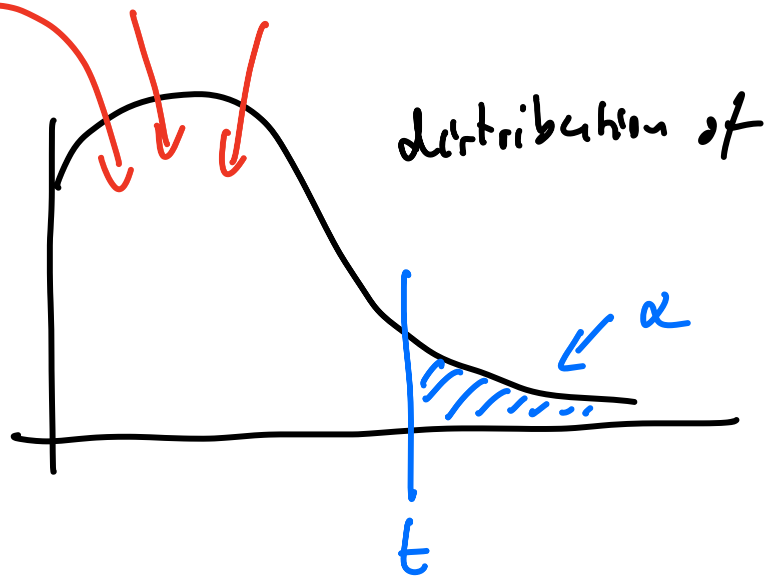
- For each point, we look at the colors of the k nearest neighbors:

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

Permutation (randomization) tests

- Sample $X_1, \dots, X_n$ group A $\rightsquigarrow$ mean $\bar{X}$
  $Y_1, \dots, Y_n$ group B mean $\bar{Y}$?
- Compute observed statistics T_{observed} = mean (red) − mean (blue)
- Pool the sample
- For $k = 1, \dots, 10^3$: shuffle the group membership ("colors")
- Compute the difference T_k = mean (red) − mean (blue)

2) $T_1, T_2, \dots, T_{1000}$



- Find α-quantile to determine rejection threshold.

- Check whether the observed T_{observed} on the true data is $\leq t$.

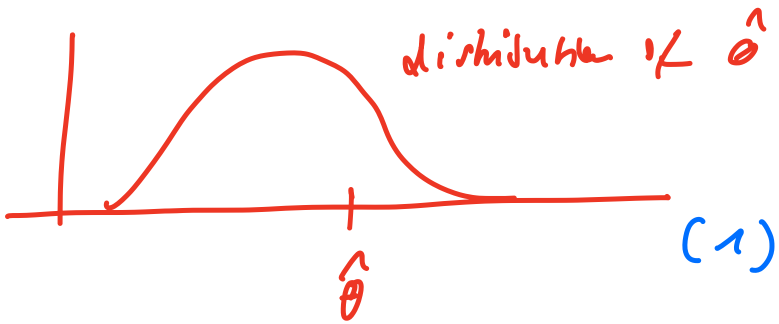
Bootrhcp fesrp

Motivation

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.

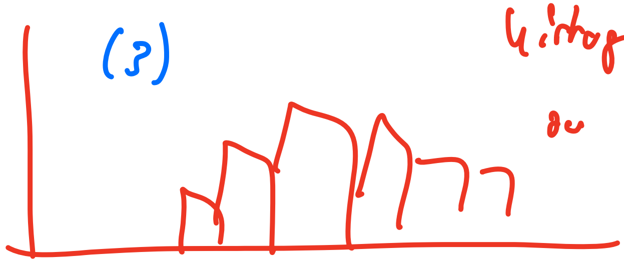
(2)

Then we could maybe build a hist on this.

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

(3)



histogram (of θ^*) based on resampled data

"Hope: histogram (3) of $\hat{\theta}^*$ 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 pseudo code

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

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

- Sample $x_1^*, \dots, x_n^*$ 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 $\widehat{se}$ of the original estimate $\hat{\theta}$ by the standard dev. of the bootstrap replicates:

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

Does it always work?

Consistency result for bootstrap

<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.

Example where it goes wrong

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

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

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

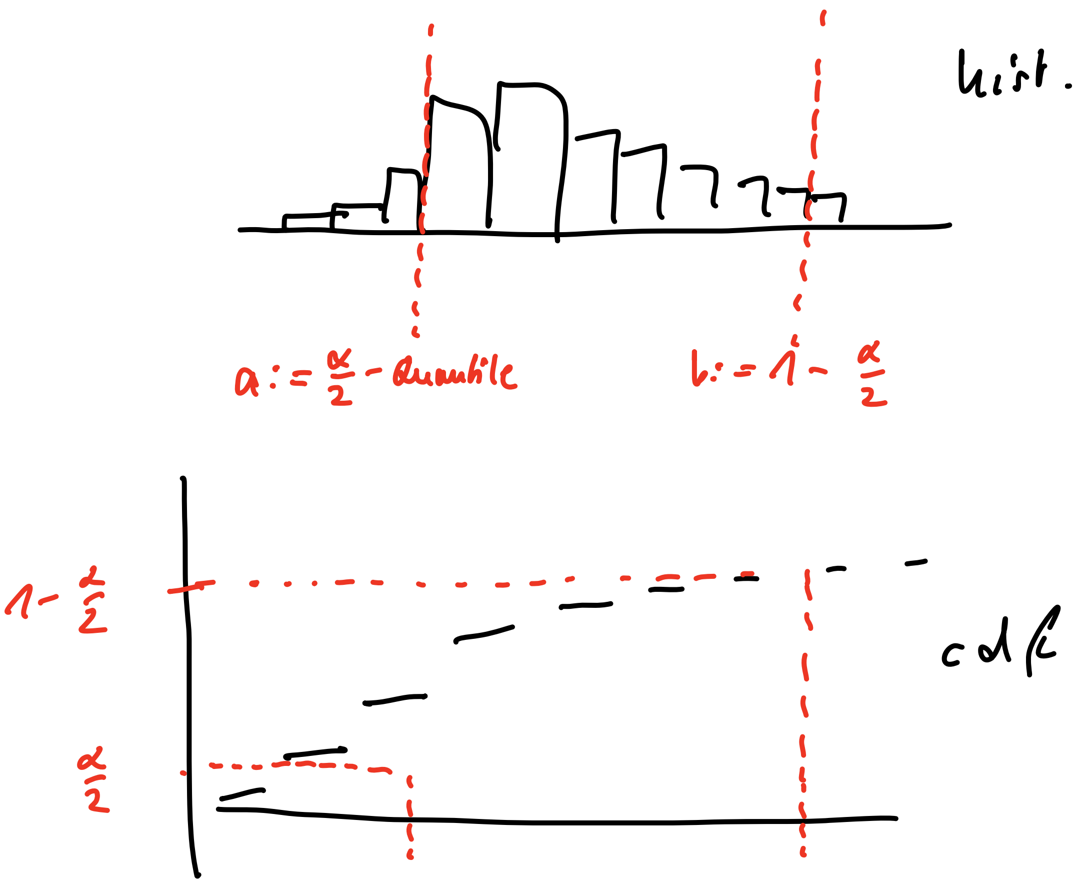
Estimating the $\hat{se}$ by bootstrap is going to fail.

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

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

Bayesian statistics

Frequentist vs. Bayesian statistics

Frequentist statistics:

- probability = limiting frequency
- parameters θ are constants, we cannot assign probabilities to them
- statistics behaves 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: the model

Assume a statistical model $\{f_\theta \mid \theta \in \Theta\}$, as in frequentist approach.

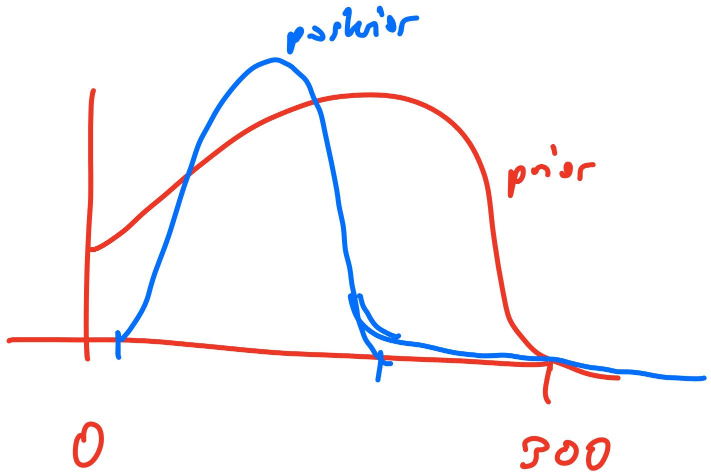
It encodes our prior assumptions on the data-generating process in general.

θ unknown, want to estimate it.

Bayesian approach: prior distribution

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

the parameters, not the data



Bayesian statistics: the likelihood

Observe data $X_1, \dots, X_n$ iid from some of the f_θ (θ unknown).

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

density data parameter

(In frequentist world, we could now use MLE to select the para that maximizes the likelihood)

Bayesian statistics: posterior

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

$$\overbrace{f(\theta | x_1 \dots x_n)}^{\text{posterior}} = \frac{\overbrace{f(x_1 \dots x_n | \theta)}^{\text{likelihood}} \cdot \overbrace{f(\theta)}^{\text{prior}}}{\underbrace{\int f(x_1 \dots x_n | \theta) f(\theta)\, d\theta}_{\text{normalizing constant}}}$$

normalizing constant
(does not depend on θ any more)

The posterior is a distribution.

Statistics derived from posterior

- 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

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

Discussion

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

High-dimensional probability and statistics

Literature:

Books: Wainwright: High-dimensional statistics.

Vershynin: High-dim probability

Papers: • PCA in high dimensions (Johnstone, Paul, 2018)

• Practice, theory, and theorems for random matrix theory in ML. (Mahoney, 2022)

High-dimensional statistics is different!

General setting:

- high-dim data (images, feature-based, genetic data, ...). d is easily 100 or 1000.

- In some applications, one also has very few data points (eg, medical case: $n = 20$ patients, $d = 1000$ genetic markers)

- ML models tend to have many parameters, m.

Classic vs. high-dim

"Classic" statistics: d fixed, $n \to \infty$: LLN, CLT, ...

"Modern regime": $d \to \infty$, $n \to \infty$

Different ways to model this, the question is always how to "couple" d and n and m

E.g, $d/n \to$ const.

Things behave very different and unexpected in this regime.

Concentration phenomena

in high-dim spaces

Norms of high-dim vectors are concentrated

Consider $X_1, \dots, X_d$ independent, $\mathbb{E}(X_i) = 0$, $\mathrm{Var}(X_i) = 1$, $X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$

Expected norm of X:

$$\mathbb{E}(\|X\|^2) = \mathbb{E}\left(\sum_{i=1}^{d} X_i^2\right) = \sum_{i=1}^{d} \underbrace{\mathbb{E}X_i^2}_{1} = d$$

Concentration of the norm:

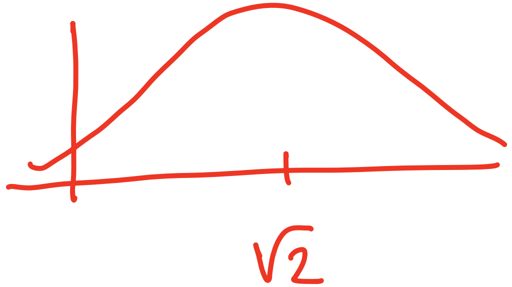
By Bernstein inequality one can prove:

$$P\left(\left|\|X\| - \sqrt{d}\right| > t\right) \leq 2\exp(-ct^2)$$

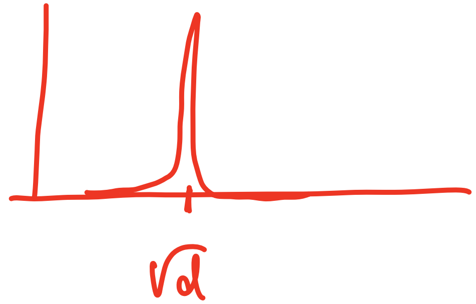
Norms of high-dim vectors are concentrated

Illustration:

2-dim



d-dim

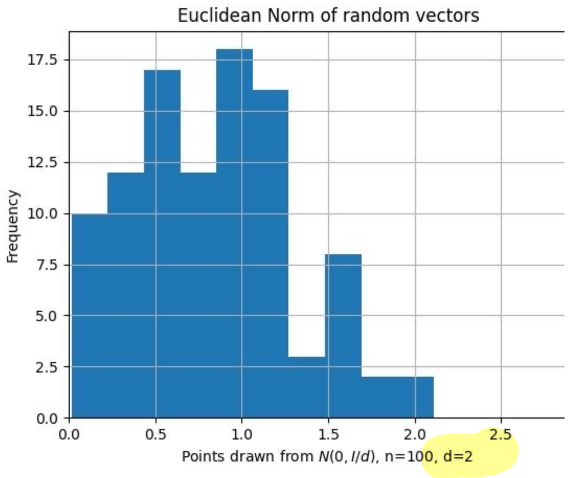


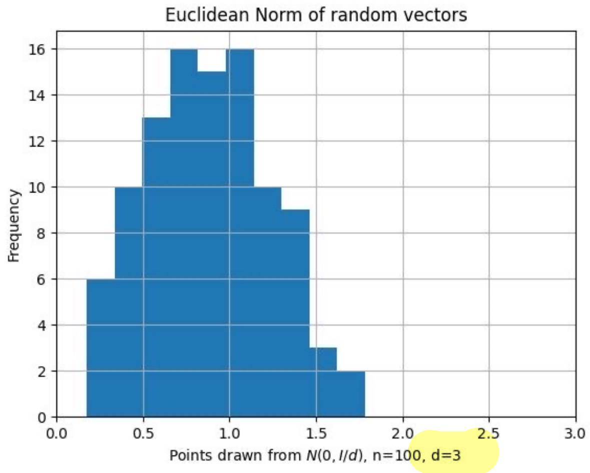
$X \sim N(0, I_2)$

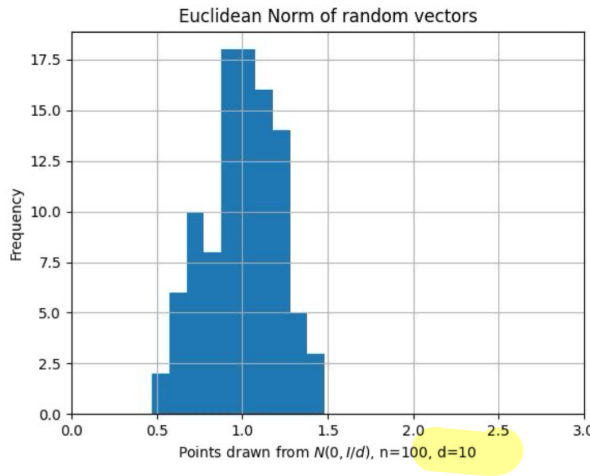
$X \sim N(0, I_d)$

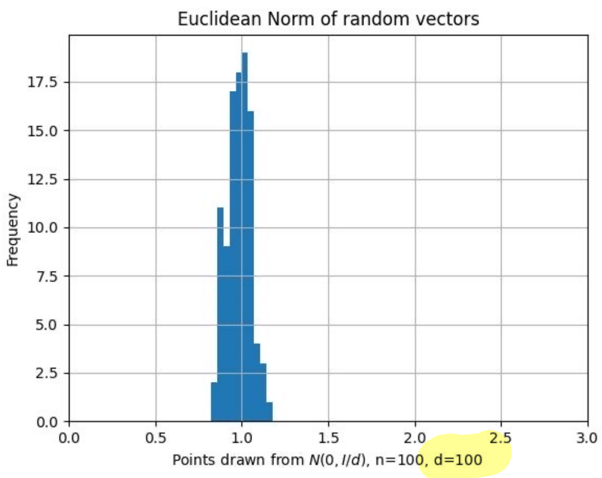
Norms of high-dim vectors are concentrated

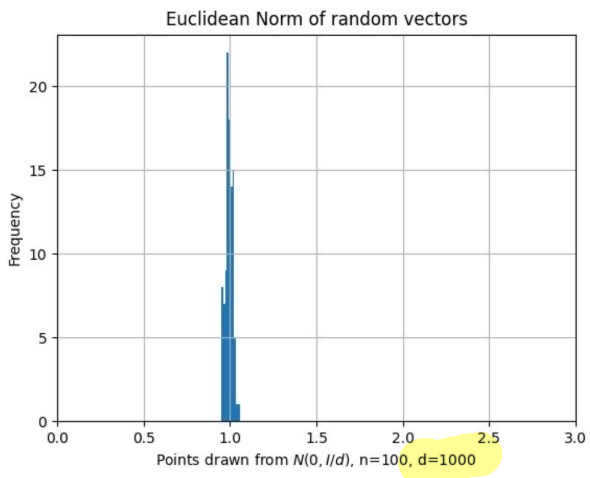
Simulations: points drawn from d-dim normal distr. $N(0, \frac{1}{d} \cdot I_d)$
in increasing dimensions ($d = 2, 3, 10, 100, 1000$)





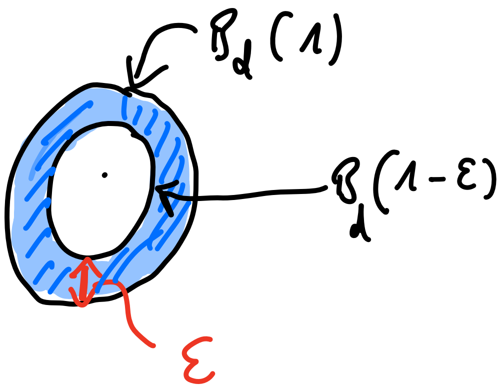






Volume of balls is concentrated along "surface"

Concentration can be seen from the point of view of sampling (previous slide), but also from the point of view of geometry:



$$\mathrm{vol}\left(B_d(1) \setminus B_d(1-\varepsilon) \right) =$$

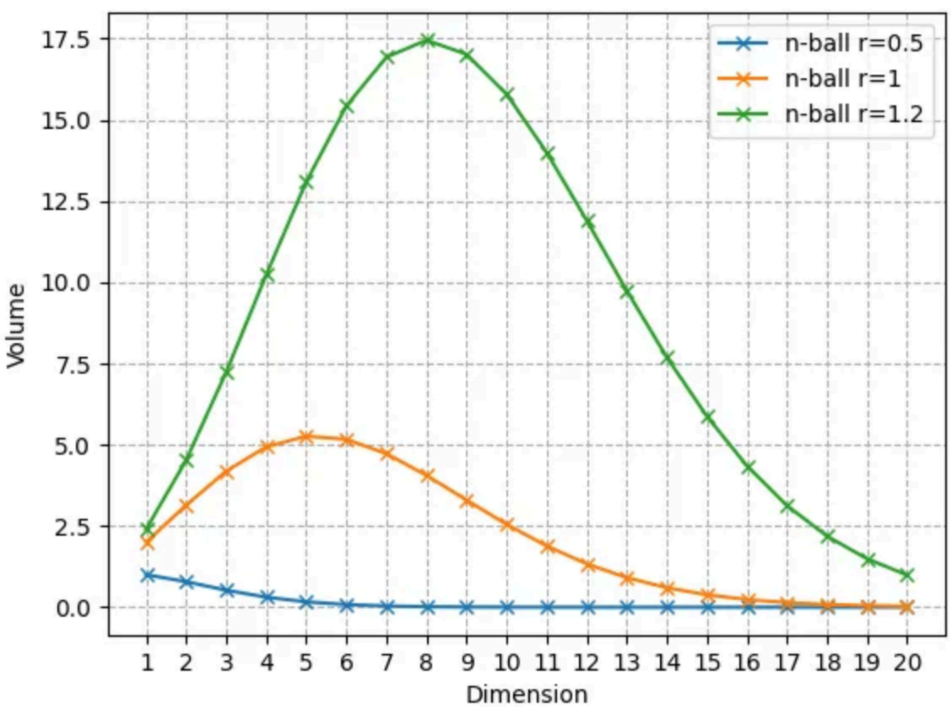
$$= 1^d \cdot \mathrm{vol}\left(B_d(1)\right) - (1-\varepsilon)^d \, \mathrm{vol}\left(B_d(1)\right)$$

$$= \underbrace{\left(1 - (1-\varepsilon)^d\right)}_{\longrightarrow 1 \text{ as } d \to \infty} \mathrm{vol}\, B_d(1)$$

Volume of unit balls gets tiny

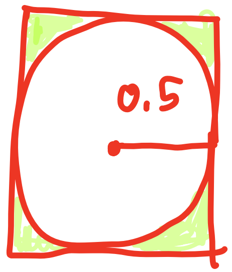
The Euclidean volume of the unit ball in $\mathbb{R}^d$ goes to 0 as $d \to \infty$.

Figure credit: Maxim Wolf, Medium 2024



Volume of d-dim unit ball

2-dim



$\mathrm{vol}_2(\text{square}) = 1$

$\mathrm{vol}_2(\text{ball}) = \pi/4$

≈ 0.75

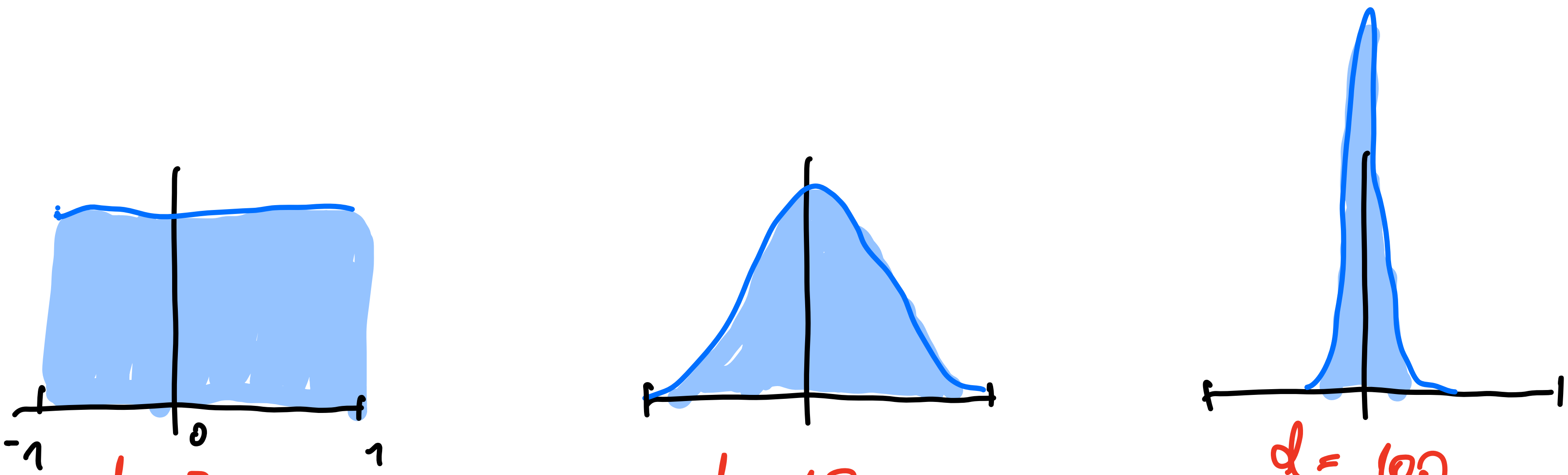
as d increases, "corners" get more volume. vol (ball) gets smaller

Random vectors on the sphere are nearly orthogonal

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix}$$ drawn uniformly from sphere

$$E(|\langle x, y \rangle|) = 1/d \quad (\approx 0 \text{ if } d \text{ is large!})$$

Distribution of $\langle x, y \rangle$ for two independent random vectors on the sphere:

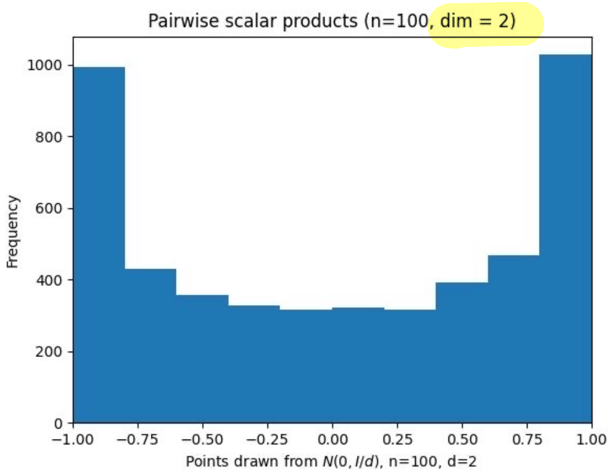


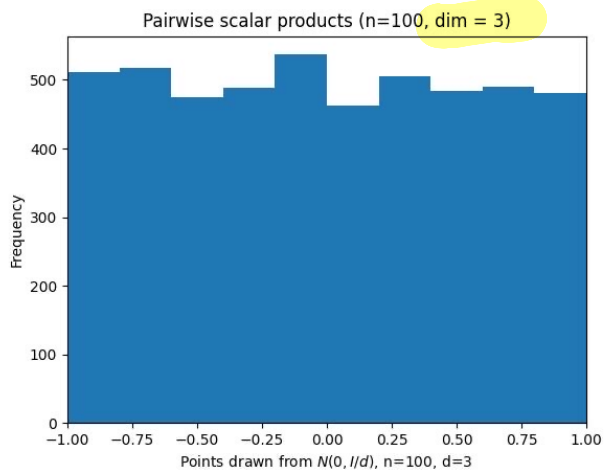
$d = 3$ $\qquad$ $d = 10$ $\qquad$ $d = 100$

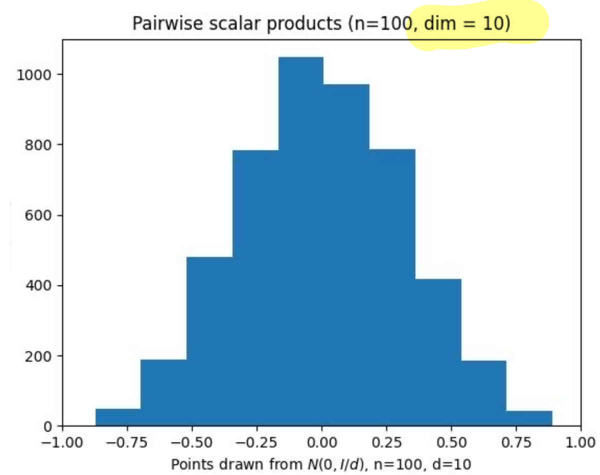
$\leadsto$ PCA gets problematic, see later

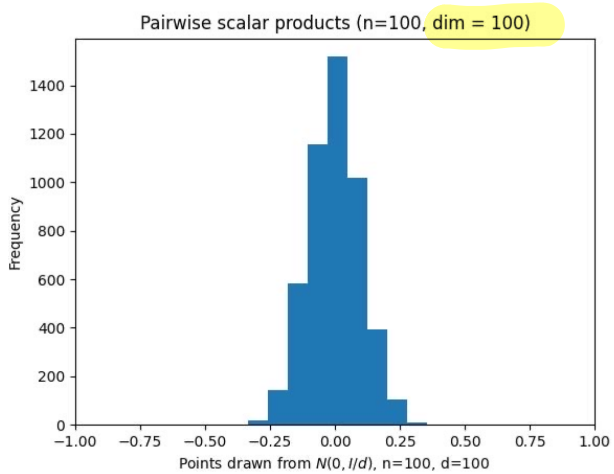
Random vectors on the sphere are nearly orthogonal

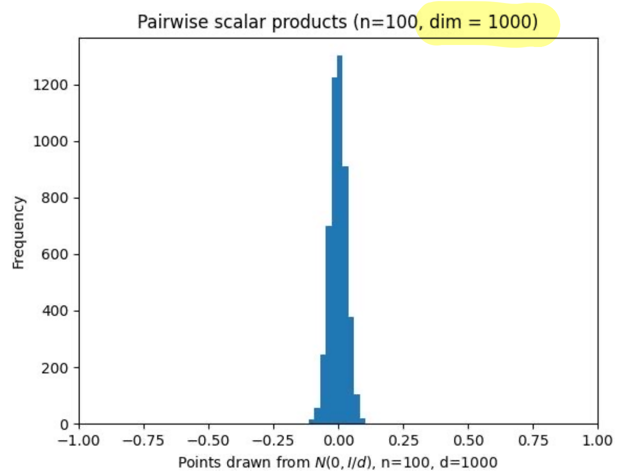
Simulations: points drawn from d-dim normal distn. $N(0, \frac{1}{d} \cdot I_d)$
in increasing dimensions ($d = 2, 3, 10, 100, 1000$)











Let's look at distances in

high-dim spaces

If we are lucky ... pairwise distances are still meaningfull

⇝ Johnson-Lindenstrauss

Can work "as if we were in a low dimension":

Theorem of Johnson-Lindenstrauss:
Consider n points in $\mathbb{R}^d$, fix $\varepsilon > 0$. One can find a linear map $f: \mathbb{R}^d \to \mathbb{R}^k$ with $k \approx \log n / \varepsilon^2$ such that

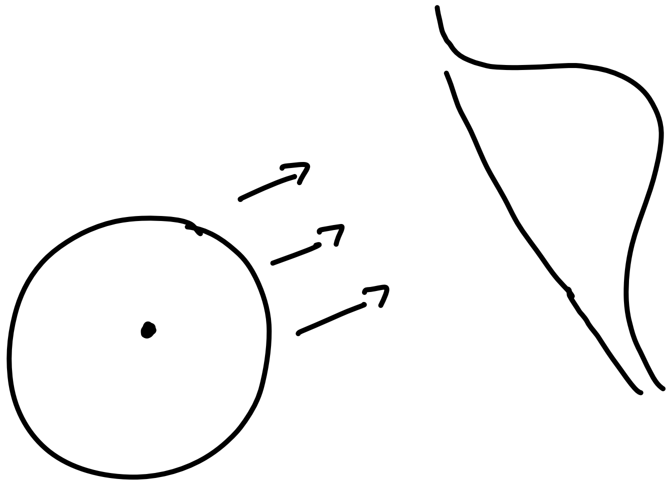
the distances do not change by "more than ε":

$$(1-\varepsilon) \| x_i - x_j \| \leq \| f(x_i) - f(x_j) \| \leq (1+\varepsilon) \| x_i - x_j \|$$

⇝ "random projections"

But often we are not lucky: random projections look like (simple) Gaussians

$X \sim \text{Unif}(S_{d-1}\sqrt{d})$, Project X on a fixed vector v:



Then $\langle X, v\rangle \longrightarrow N(0,1)$ in distribution

In practice, this often also holds for non-uniform data.

We don't see any class or cluster structure!

Distances in high-dim spaces are dominated by noise

$X \sim N(\mu_1, \sigma_1^2 I_d)$, $Y \sim N(\mu_2, \sigma_2^2 I_d)$

$$E(\|X-Y\|^2) = E\left(\sum_{i=1}^{d} |X_i - Y_i|^2\right)$$

$$= \sum_{i=1}^{d} \left(\mathrm{var}(X_i - Y_i) + \left(E(X_i - Y_i)\right)^2\right)$$

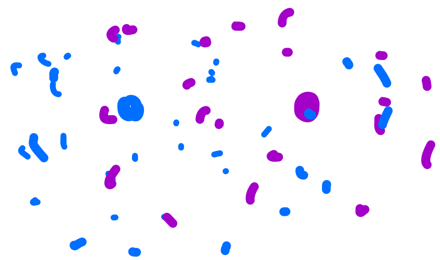
$$= \underbrace{d \cdot \left(\sigma_1^2 + \sigma_2^2\right)}_{\text{noise}} + \underbrace{\|\mu_1 - \mu_2\|^2}_{\text{signal}}$$

classification is difficult!

2-dim

$\mathrm{dist}(X,Y) = \|\mu_1 - \mu_2\| +$ small noise

d-dim



noise >> signal

Within-class-distances << between-class-distances !?

Mixture of Gaussians with different variances:

$$x, x' \sim \mathcal{N}(\mu_1, \sigma_1^2 I_d), \quad y, y' \sim \mathcal{N}(\mu_2, \sigma_2^2 I_d)$$

Within-class distance:

$$E(\|x - x'\|^2) = 2d\sigma_1^2 \quad \text{(class 1)}$$

$$2d\sigma_2^2 \quad \text{(class 2)}$$

Between-class-distances:

$$E(\|x - y\|^2) = \|\mu_1 - \mu_2\|^2 + d(\sigma_1^2 + \sigma_2^2)$$

If $\sigma_1 << \sigma_2$, then $\underbrace{d \cdot (\sigma_1^2 + \sigma_2^2)}_{\text{between}} << \underbrace{d \cdot 2\sigma_2^2}_{\text{within-class-2}}$

clustering is difficult!

ML consequence:

- different classes overlap heavily and are hard to distinguish based on individual distances

- Might hope for assumptions that tell us that actually, our space does not have such a high dimension:
  - sparsity (→ book by Wainwright)
  - manifold assumption

Let's take a look at matrices...

Estimating covariance matrices in high dim

Assume we have n data points in d dimensions $\rightsquigarrow$ $n \cdot d$ "measurements"

Covariance matrix has d^2 entries... obvious trouble if $n \ll d$.

The default estimate for the covariance matrix, the sample cov. matrix

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i'$$ might not converge to the true

underlying cov. matrix.

If n is low and d is high, (d/n large), the entries of the sample covariance matrix are unreliable and do not converge to the true values. If $n \ll d$, there is no way to fix that!!!

Estimating covariance matrices in high dim

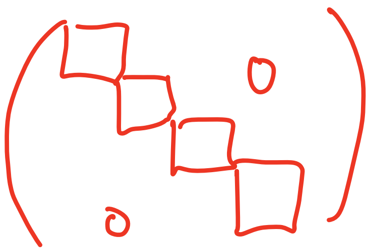
Alternatively, if $d \gg n$, one needs to make structural assumptions, for example:

- Spiked covariance model where the "signal" just lives in few data dimensions

- Sparsity assumptions $\quad \text{cov} = \begin{pmatrix} 1 & 1 & & 1 \\ 1 & & 1 & 1 \end{pmatrix}$

- Block-matrix assumptions

...

Cov. = 

Eigenvalues in high dimension

Consider a $n \times d$ -matrix with iid random entries (eg, Gaussian $N(0,1)$):
n points, d dimensions each.

The true covariance matrix is the identity.

Can we use the sample cov. matrix to estimate the eigenvalues of the true cov. matrix? (d numbers to estimate) ?

Different things happen, depending on the ratio d/n :

Eigenvalues in high dimension

If $d \ll n$, then eigenvalues concentrate:

Concentration theorem (Vershynin book Th. 4.7.1)

X subgaussian vector in $\mathbb{R}^d$, denote by Σ the true covariance matrix, and by Σ_n the empirical covariance matrix based on n sample points. Then

$$\mathbb{E}\left(\|\Sigma_n - \Sigma\|\right) \leq \text{const.} \left(\underbrace{\sqrt{\frac{d}{n}} + \frac{d}{n}}\right) \|\Sigma\| .$$

subgaussian:
$P(|X| > t) \leq 2\exp(-t^2 \cdot \text{const})$

if $d/n \to$ const, this term does not vanish any more!

ℓ_2-operator norm ($\approx$ distance between largest eigenvalues)

Eigenvalues in high dimension

If $d/n \to$ const., then the distribution of eigenvalues is known:

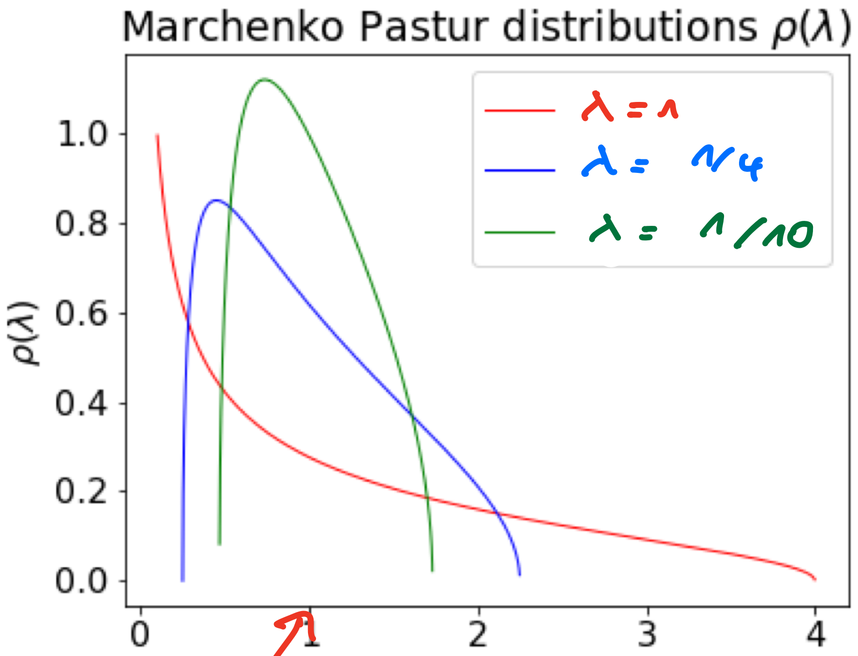
Consider random vectors whose entries are iid with variance 1.

The theorem of Marchenko-Pastur characterizes the distribution of eigenvalues of the sample cov. matrix as $d/n \to \lambda$:
(It has a closed-form expression that I did not want to put on the slide).

General behavior:

The larger λ, the more the eigenvalues "spread", see next slide:

$$\lambda = \frac{d}{n}$$



true eig is 1

True cov. matrix is I_d, so all true eigs are 1.

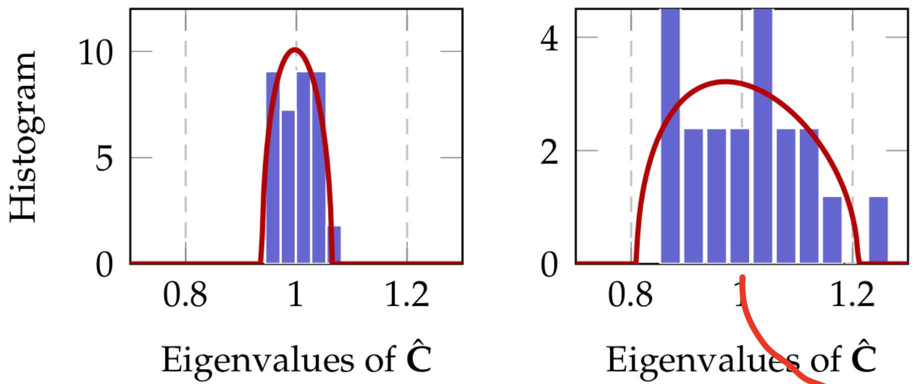


Figure by M. Mahoney: empirical eigs vs. limit distribution for a matrix with standard normal entries

$d = 20$
$n = 1000$

$d = 20$
$n = 100$

true eigenvalue is 1

Eigenvectors (PCA) in high dim

PCA on high-dim data is problematic, to say the least.

Even the leading eigenvector is typically inconsistent!!!

Why should we care?

Weights in a DNN can behave funny as well if you try to analyze their statistics

(see for example papers by M. Mahoney)

Outlook: double descent

Finally, let's look at "model size" (in terms of the number of parameters) as well.

Relation between model size in ML and overfitting:

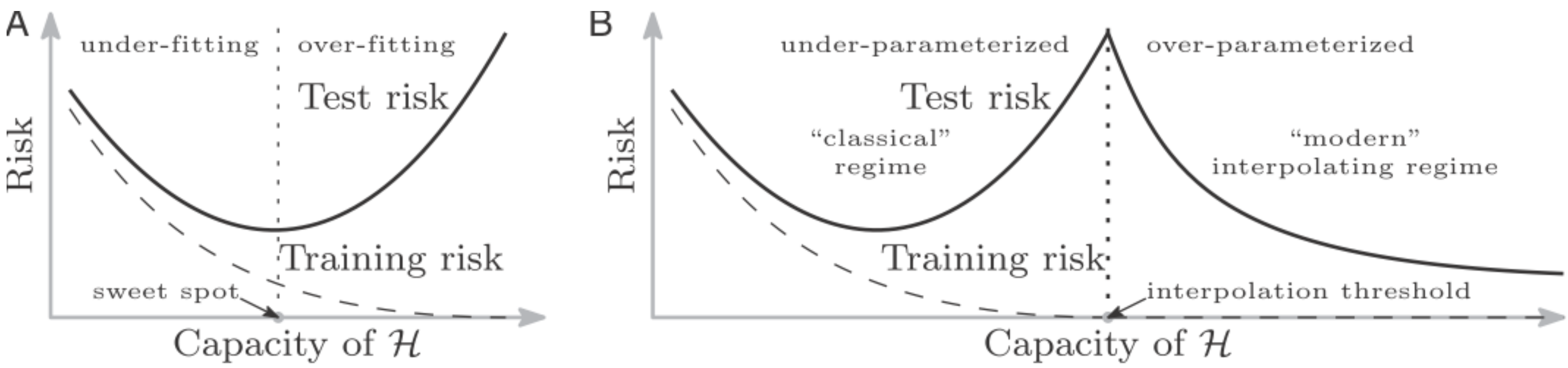


Figure credit: Mikhail Belkin

Long debate ... see the Statistical ML lecture next term ...