Traditional, frequentist statistics

Point estimation, bias, variance

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

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

More generally, this family is typically denoted as follows:

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

(Parametric statistics!)

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

We are given a sample \( X_1, \ldots, X_n \sim f_{\theta} \) (typically, iid) but the true, underlying \( \theta \) is unknown.

The goal of point estimation is to estimate \( \theta \).

Convention: Para space \( \Theta \), true parameter \( \theta \),

\( \hat{\theta} \in \Theta \) refers to the probability and expectation under the distribution (density) \( f_{\theta} \), \( \hat{\theta} \) estimated parameter
Given a statistical model \( \mathcal{F} = \{ f(\theta) \mid \theta \in \Theta \} \), and a sample \( x_1, \ldots, x_n \sim F \in \mathcal{F} \). A point estimator \( \hat{\theta}_n \) of parameter \( \Theta \) is a function
\[
\hat{\theta}_n := g(x_1, \ldots, x_n)
\]

**Def.** The bias of such an estimator is defined as
\[
\text{bias}(\hat{\theta}_n) := E_{\theta}(\hat{\theta}_n) - \theta
\]

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 \( \text{se} \). Typically, \( \text{se} \) is unknown, but it can be estimated: \( \hat{\text{se}} \).

**Example:** \( X_1, \ldots, X_n \sim \text{Bernoulli}(\theta) \) for parameter \( \theta \in [0,1] \),

\[
\hat{\theta}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{an estimate of } \theta.
\]
\[ E_p (\hat{\mu}_n) = E_p \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n} \sum_{i=1}^{n} E_p (X_i) = \mu. \]

Thus, \( \hat{\mu}_n \) is unbiased because
\[ E_p (\hat{\mu}_n) = \mu - \mu = 0. \]

The standard error of this estimate is
\[ se = \sqrt{\operatorname{Var}_p (\hat{\mu}_n)} = \sqrt{\frac{1}{n} \operatorname{Var}_p (\hat{\mu}_n)} = \sqrt{\frac{\sigma_p (1-p)}{n}}. \]

We can for example estimate it by
\[ \widehat{se} = \sqrt{\frac{\hat{\mu}_n (1-\hat{\mu}_n)}{n}}. \]

**Example:** weight of baby

![Distribution of baby weights](image)

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

\[ \bar{x} = 2950 \text{ g} \]

Do it a second time \( \approx 2890 \text{ g} \)

Distribution of the estimate \( \hat{\mu}_n \)
The mean squared error (MSE) of an estimate is the quantity

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

**Theorem:** bias-variance decomposition

\[
\text{MSE}(\hat{\theta}_n, \theta) = \text{bias}^2(\hat{\theta}_n) + \text{Var}_{\theta}(\hat{\theta}_n)
\]

**Proof:**

\[
E_{\theta} \left( (\hat{\theta}_n - \theta)^2 \right) =
\]

\[
= E_{\theta} \left( \left( \hat{\theta}_n - E\hat{\theta}_n + E\hat{\theta}_n - \theta \right)^2 \right)
\]

\[
= E_{\theta} \left( (\hat{\theta}_n - E\hat{\theta}_n)^2 \right) + 2E_{\theta} \left( (\hat{\theta}_n - E\hat{\theta}_n)(E\hat{\theta}_n - \theta) \right) + E \left( (E\hat{\theta}_n - \theta)^2 \right)
\]

\[
= E_{\theta} \left( (\hat{\theta}_n - E\hat{\theta}_n)^2 \right) + 2E_{\theta} \left( (\hat{\theta}_n - E\hat{\theta}_n)(E\hat{\theta}_n - \theta) \right) + E \left( (E\hat{\theta}_n - \theta)^2 \right)
\]

\[
= E_{\theta} \left( (\hat{\theta}_n - E\hat{\theta}_n)^2 \right) + 2E_{\theta} \left( (\hat{\theta}_n - E\hat{\theta}_n)(E\hat{\theta}_n - \theta) \right)
\]

\[
\overset{\text{deterministic}}{=} E_{\theta} (\hat{\theta}_n - \theta) \cdot E_{\theta} (\hat{\theta}_n - E\hat{\theta}_n)
\]

\[
= E_{\theta} (\hat{\theta}_n) - E_{\theta} E\hat{\theta}_n = 0
\]

\[
= 0
\]

\[
= E_{\theta} \left( (\hat{\theta}_n - E\hat{\theta}_n)^2 \right) + E \left( (E\hat{\theta}_n - \theta)^2 \right)
\]

\[
\overset{\text{deterministic}}{=} E_{\theta} (\hat{\theta}_n - \theta)^2
\]

\[
= \left( \text{bias} (\hat{\theta}_n) \right)^2
\]
Example: \( \mathcal{F} = \{ N(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma > 0 \} \)

Sample: \( X_1, \ldots, X_n \sim N(\mu, \sigma^2) \) with unknown \( \mu, \sigma^2 \), iid

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{is an unbiased estimate of } \mu.
\]

\[
\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2
\]

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

Let's compute: true para \( \sigma \)

\[
E(\hat{\sigma}_n^2) = \frac{n-1}{n} \sigma^2 \quad \text{so the bias is } \frac{1}{n} \sigma^2
\]

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

\[
\text{Var}(\hat{\sigma}_n^2) = \frac{2(n-1)}{n} \sigma^4
\]

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

\[
\text{MSE}(\hat{\sigma}_n^2) = \text{bias}^2 + \text{var} = \ldots = \left( \frac{2(n-1)}{n^2} \right) \sigma^4
\]

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

\[\Rightarrow \text{MSE} (\hat{\sigma}_n^2) < \text{MSE}(\hat{\sigma}_2^2)\]
**Def.** A point estimator $\hat{\theta}_n$ of $\theta$ is **consistent** (strongly consistent) if

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

as $n \to \infty$

**Theorem.** If an estimator satisfies $\text{bias} \to 0$ and $\text{se} \to 0$ as $n \to \infty$, then the estimator is consistent.
Def. A \((1-\alpha)\)-confidence interval for a parameter \(\theta \in \mathbb{R}\) is an interval \(c_n = (a_n, b_n)\) where 
a
\[ a_n = a(x_1, \ldots, x_n), \quad b_n = b(x_1, \ldots, x_n) \]
are functions of the sample \(x_1, \ldots, x_n\) such that 
\[ P_{\theta}\left(\theta \in c_n\right) \geq 1 - \alpha \] for all \(\theta \in \Theta\).

First experiment \(x_1, \ldots, x_n \sim \theta\)

Second experiment \(x_1, \ldots, x_n \sim \theta\)

... in \((1-\alpha)\) of the repetitions, the true \(\mu\) is inside the red interval.
Example: Coin flips, with \( P(X = 1) = \rho \), \( P(X = 0) = 1 - \rho \), \( \rho \in [0, 1] \) unknown. Want to estimate it.

\[ \hat{\rho}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \]

Choose a confidence level \( \alpha \), now want to define \( c_n = (a_n, b_n) \). To this end,

\[ \epsilon_n^2 := \frac{\log \left( \frac{2}{\alpha} \right)}{2n} \]

Then the interval

\[ c_n := (\hat{\rho}_n - \epsilon_n, \hat{\rho}_n + \epsilon_n) \]

is a CI with coverage \( 1 - \alpha \).

Proof:

By Hoeffding inequality, for any \( t \) we have

\[ P \left( \left| \hat{\rho}_n - \rho \right| > t \right) \leq 2 \exp(-2n t^2) \]

\[ \alpha = 2 \exp(-2n t^2) \]

\[ \log \left( \frac{\alpha}{2} \right) = -2n t^2 \Rightarrow t^2 = \frac{-\log \left( \frac{\alpha}{2} \right)}{2n} = \frac{\log \left( \frac{2}{\alpha} \right)}{2n} \]

Choose \( \epsilon_n = t \).
Example

$$\mathcal{F} = \left\{ A \mid A \text{ symmetric, } a_{ij} \in \{0, 1\} \right\}$$

adjacency matrices of graphs

Observe 2 random walks from the graph for length 10.

one random walk produces a sequence

$$x_1, x_2, \ldots, x_{10}$$

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 walk you have observed.

\(\Rightarrow\) Maximum likelihood approach

More formally: Parametric family \( \mathcal{F} = \{ f_\theta \mid \theta \in \Theta \} \), observe iid points $$X_1, \ldots, X_n \sim f_\theta \in \mathcal{F}$$.

The likelihood of the data given a parameter $$\Theta$$ is

$$p_{\theta_0}(x_1, \ldots, x_n) = p(x_1, \ldots, x_n \mid \theta_0)$$

$$= \prod_{i=1}^{n} p(x_i \mid \theta_0)$$
To estimate the true parameter $\theta$, we now select $\hat{\theta}$ such that their likelihood is maximized:

$$\hat{\theta} := \arg\max_{\theta} P(X_1, \ldots, X_n | \theta)$$

This is equivalent to the problem

$$\hat{\theta} = \arg\max \prod_{i=1}^n P(X_i | \theta)$$

which is equivalent to minimizing the negative log-likelihood:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^n -\log P(X_i | \theta)$$

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

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

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

Want to construct the MLE estimator.

$$\mathcal{L}(\lambda) = P(X_1, \ldots, X_n | \lambda) = \prod \frac{\lambda^{X_i} e^{-\lambda}}{X_i!},$$

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

$$= \sum_{i=n}^{n} X_i \log \lambda - n - \log(X_i !)$$

Now want to optimize for $\lambda$. Take the derivative

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

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

So $\hat{\lambda} = \frac{1}{n} \sum_{i=n}^{n} X_i$ is the MLE estimator of $\lambda$. 

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

1. If the model $F$ consists of "nice" functions, then the MLE based on an iid sample is **consistent**.

2. If $F$ consists of "nice" functions, the MLE estimate $\hat{\theta}_{\text{MLE}}$ is **asymptotically normal**:

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

3. This can be used to construct confidence intervals:

$$c_n := \left( \hat{\theta}_{\text{MLE}} - \frac{z_{\alpha/2} \cdot \text{se}}{\sqrt{n}}, \hat{\theta}_{\text{MLE}} + \frac{z_{\alpha/2} \cdot \text{se}}{\sqrt{n}} \right)$$

where $z_{\alpha/2} := \Phi^{-1}(1 - \frac{\alpha}{2})$
$c_n$ is an approximate CI in the sense that

$$P_{\theta} (\theta \in c_n) \rightarrow 1 - \alpha \text{ as } n \rightarrow \infty.$$
Sufficiency, identifiability

Intuition: given sample $x_1, ..., x_n \sim f_\theta$ and $\theta \in \Theta$.

- We typically count the $(x_1, ..., x_n)$ sample as a statistic $T(x_1, ..., x_n)$ (in the extreme case, one number).

- Question: can we recover the true parameter $\theta$ from this statistic?

- Intuition: when we observe two samples $x_1, ..., x_n$, $x_1', ..., x_n'$ and $T(x_1, ..., x_n) = T(x_1', ..., x_n')$, then we want to infer the same $\theta$.

- When we know $T(x_1, ..., x_n)$, 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.

**Definition:** A parameter $\theta$ for a family $\mathcal{F} = \{f_\theta | \theta \in \Theta\}$ is identifiable if distinct values of $\theta$ correspond to distinct pdfs in $\mathcal{F}$: $\Theta \neq \Theta' \implies f_\theta \neq f_{\theta'}$. 
Example: Mixture distribution

\[ f = \left\{ \sum x_i \ N(\mu_i, \sigma_i^2) \right\}^2 \] with \[ \sum x_i = 1 \]

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 "glucose"/"no glucose"

\[ 0.3 \ N(\mu_3, \sigma_3^2) + 0.7 \ N(\mu_4, \sigma_4^2) \]
Hypothesis Testing

Example: Two drugs $D_1, D_2$, we measure number of days to recovery $x_1, \ldots, x_n$ treated with $D_1$.

![Graph showing comparison of days to recovery for drugs $D_1$ and $D_2$.]

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 $\hat{p}_n = \frac{\sum x_i}{n}$. 
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:

\[
\hat{p}_n \sim \text{Beta}(n, n)
\]

Imagine a confidence set: Let $S$ such that

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

More formal setup

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

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

Want to test

$H_0: \theta \in \Theta_0$ against $H_1: \theta \in \Theta_1$.

Null Hyp.

Alternative Hyp.

Sample data from the unknown $f_\theta$, compute a test statistic $T(x_1, \ldots, x_n)$. Now we construct a rejection region $R_n$ such that

$T(x_1, \ldots, x_n) \in R_n \Rightarrow$ reject $H_0$

$T(x_1, \ldots, x_n) \in R_n \Rightarrow$ 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:

<table>
<thead>
<tr>
<th>$H_0$ true</th>
<th>$H_1$ true</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test rejects $H_0$</td>
<td>Type II error</td>
</tr>
<tr>
<td>Test rejects $H_0$</td>
<td>Type I error</td>
</tr>
</tbody>
</table>

**Def** The power function 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 $\theta$, $\beta(\theta) = P(\text{Type I error})$. 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.

**Def** We say that a test is of level $\alpha$ if

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

Intuition: worst case guarantees no mistake when $\theta \in \Theta_0$. In other words, the type I error is not larger than $\alpha$. 
(Intuition to remember: \( \alpha \leq \text{type I error} \))

**Standard procedure:** We fix the level \( \alpha \) 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 \( \alpha \), you might want choose the one that has the smallest type II error.

**Notation used often in literature:**

\[
\begin{align*}
\alpha &= P(\text{type I error}) \\
\beta &= P(\text{type II error}) \\
\alpha - \beta &= \text{power of a test}
\end{align*}
\]

**Remark:** The power of a test is typically evaluated when we test against a concrete hypothesis \( \theta \in \Theta_i \). We say “the power of the test against alternative \( \theta \in \Theta_n \).”

**Def:** Let \( T \) be a set of tests of level \( \alpha \) for testing

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

A test in \( T \) 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 \( \beta' \) that are power functions for other tests in \( T \).

**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

\[
T = \frac{L(\theta_1)}{L(\theta_0)} = \left( \frac{\prod_{i=1}^{n} f(x_i | \theta_1)}{\prod_{i=1}^{n} f(x_i | \theta_0)} \right)
\]

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-\( \alpha \)-test.

**Hare general likelihood-ratio test:**

Parameter space \( \Theta \), \( \Theta_0 \subset \Theta \), \( \Theta_1 = \Theta_0^c \). Then we consider the test statistic

\[
\tilde{T} = \sup_{\theta \in \Theta_0} \frac{L(\theta)}{\sup_{\theta \in \Theta_1} L(\theta)}
\]
or even simpler

\[
T = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta_1} L(\theta)}
\]
and we determine a parameter \( \lambda \) such that the rejection region is of the form \( R = \{ \Gamma \leq \lambda \} \).

In practice the difficulties are
- compute the suprema \((\text{in practice})\)
- fix \( R \), fix \( \lambda \) \((\text{in theory})\)
Consider a test at level $\alpha$, and denote its rejection region as $R_\alpha$.
Recall: $\alpha = P(\text{Type-I-error})$.
The smaller $\alpha$, the more difficult does it get to reject $H_0$.
(we often even have that $\alpha < \bar{\alpha} \Rightarrow R_\alpha \subset R_{\bar{\alpha}}$)

Def. The $p$-value is defined as

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

i.e. the smallest $\alpha$ for which the level- $\alpha$-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

\[ \hat{\mu}_g, \hat{\mu}_b, \text{ and } \text{se}_g, \text{ se}_b \]

For a large test will find a statistically significant difference. In small \( p \)
**Multiple testing**

Example: gene expression data

<table>
<thead>
<tr>
<th>patients with cancer (n = 20)</th>
<th>Control group (n &gt; 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1</td>
<td>0.5 0.2 0.9 0.8 0.5</td>
</tr>
<tr>
<td>gene 2</td>
<td>0.04 0.05 0.1 0.02</td>
</tr>
</tbody>
</table>

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

\[
P(\text{test i makes type I error}) = 5\% 
\]

Now we have m tests.

\[
P(\text{at least one of the tests makes a type I error}) = 
\]

\[
= P(\text{test i makes error or test } b_2 \text{ makes error or \ldots or test } b_m \text{ makes error}) 
\]

\[
= 1 - P(\text{no error in test } t_1 \text{ and no error in test } b_2 \text{ and \ldots}) 
\]

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

\[
\text{FWER} = P(\text{at least one } b_i \text{ makes type-I-error})
\]

Bonferroni correction:

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

Then we have:

\[
m = 1 \Rightarrow \alpha = 0.05 \]
\[
m = 10 \Rightarrow \alpha = 0.005 \]
\[
m = 50 \Rightarrow \alpha = 0.001 \]

= FWER

(see below)
\[ \text{FWER} = P(\text{at least one type-I error}) = \]
\[ = P(\text{t_1 error or t_2 ...}) \leq \]
\[ \leq \sum_{i=1}^{m} P(\text{t_i makes error}) = m \cdot \alpha_{\text{single}} = m \cdot \frac{\alpha}{m} = \alpha. \]

Advantage: simple, correct

Disadvantage: too conservative, low power (high type-II error) the test barely discover anything!

Benjamini/Hochberg: Controlling 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 (1995) approach:

- Fix FDR $\alpha$ in advance.
- Run the $m$ individual tests and evaluate their $p$-values.
- Sort $p$-values increasingly: $p(1) \leq p(2) \leq p(3) \leq \ldots \leq p(m)$.
- 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 $i = 1, \ldots, i_0$, 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 $\text{FDR} \leq \alpha$.

Remark: Similar approach also works without independence assumption.
Inuition:

- Under the null hypothesis, the p-values always have a uniform distribution on $[0, 1]$. Ideally, it would look like this:

\[ I \]
\[ 0 \quad 1 \]

Density of p-values under $H_0$.

If we have some $H_0$ and some $H_1$, being true and false, the density would maybe look like this:

\[ I \]
\[ 0 \quad 1 \]

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

Goal: set threshold to such that FDR satisfies what we want.
Integral of the pink area: Expected number of p-values corresponding to $H_1$ and one below $t_1$

Integral of blue area: ... $H_0$

By moving $t$ from 0 to $b_1$ we control the FDR.
For $b_1$, the FDR is small

$$b_2$$ large

General Remarks:
- $B_H$ tends to learn when your learn Foulon
- $B_H$ controls FDR, not FWER (small type-I error)
BT works best in sparse regime where all few tests reject the null.

- BT gives guarantee on FDR, but in general does not minimize it.

- When all the $H_0$ are true, $BT \approx$ Bonferroni.
Non-parametric tests

Standard (parametric scenario):

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

  ![Distribution of the samples](image)

- Observe data, compute a test statistics, 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 hypothesis](image)

- Construct rejection threshold/region: if the observed \( T_n \) is in this region, reject the null hypothesis.
Kolmogorov-Smirnov Test for Goodness-of-Fit

Goodness-of-Fit Tests: Goal is to test whether a data set comes from a particular distribution $F_0$

- $H_0$: $F(x) = F_0(x)$
- $H_1$: $F(x) \neq F_0(x)$

Kolmogorov-Smirnov: We consider the cdf $F_0$

$F_0 = \text{cdf of the true distribution}$

$F_n = \text{cdf of the data}$

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

By the Glivenko-Cantelli theorem we know that under the null hypothesis, $F_n \rightarrow F_0$ uniformly, a.s.
It is possible to compute the distribution of $D_n$, and if it independent of $F_0$, it just depends on $u$. From this we can compute rejection thresholds and design a test.

Example: looks like the data does not come from $F_0$.

**Wilcoxon - Mann-Whitney test**

(two sample test based on ranks)

Two sample test: $x_1, \ldots, x_n \sim F_1$, a first sample distributed according to $F_1$;

$y_1, \ldots, y_m \sim F_2$, a second sample distributed acc. to $F_2$.

Question: $F_1 = F_2$?

$H_0: F_1 = F_2 \quad | \quad H_1: F_1 \neq F_2$
Test: "Pool the sample": $x_1, \ldots, x_n, y_1, \ldots, y_m \lor R$

- Sort the pooled sample in increasing order and return the rank of all points $\Rightarrow \text{rank}(x_i)$ $\Rightarrow \text{rank}(y_i)$

\[ \begin{align*}
&\text{rank} 1 \quad \text{rank} 2 \quad \ldots \\
&z \quad p \quad \varepsilon \quad y
\end{align*} \]

- Compute the rank sums for both groups:

\[
\text{red group: } W_{\text{red}} = \sum_{i \in \text{red population}} \text{rank}(x_i) \\
W_{\text{blue}} = \sum_{i \in \text{blue pop.}} \text{rank}(y_i)
\]

- If $|W_{\text{red}} - W_{\text{blue}}|$ is small, we retain $H_0$; if large, reject $H_0$. 
Extension to a multivariate setting using k nearest neighbors

1. Two samples, we pool them:

![Diagram showing two sets of data points with a circle indicating a point and its nearest neighbors]

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

2. Under the null hypothesis, we expect that the number of red neighbors is equal to the number of blue neighbors.
**Permutation (randomization) tests**

- Sample $X_1, \ldots, X_n$ \textbf{group A} mean $\bar{X}$
- $X_1, \ldots, X_n$ \textbf{group B} mean $\bar{Y}$
- Compute observed statistic $T_{\text{observed}} = \text{mean (red)} - \text{mean (blue)}$
- Pool the sample
- For $k = 1, \ldots, 10^3$: shuffle the gray membership ("colors")
- Compute the difference $T_k = \text{mean (red)} - \text{mean (blue)}$

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

- Find $\alpha$-quantile to determine rejection threshold.
- Check whether the observed $T_{\text{observed}}$ on the true data is $\leq \alpha$. 

\[ T_n, T_2, \ldots, T_{1000} \]

Distribution of $T_i$
Bootstrap

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

The first thing to look at is the standard error \( \text{se} \).

If we have assumptions on \( F \), we can analytically compute the distribution of \( \hat{\theta} \), the \( \text{se} \), ...

(That is rare!)

\begin{equation}
\text{Distribution of } \hat{\theta}
\end{equation}

We could also try to obtain many samples

- \( X_1^{(1)}, \ldots, X_n^{(1)} \)
- \( X_1^{(2)}, \ldots, X_n^{(2)} \)
- \( \vdots \)
- \( X_1^{(m)}, \ldots, X_n^{(m)} \)

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

Problem: need too many samples.
Idea of the bootstrap:

1. Given the sample $X_1, \ldots, X_n$ estimate $\hat{\Theta}_{\text{orig}}$.
2. Draw a subsample of $X_1, \ldots, X_n$, compute $\hat{\Theta}^*$, repeat many often.

\[ \text{histogram of } \hat{\Theta}^* \text{ is "close" to histogram of } \hat{\Theta} \text{ (2), which is close to (1)} \]

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

Algorithm in pseudo code

- Input: $X_1, \ldots, X_n$
- For $b = 1, \ldots, B$
  - Sample $X_1^*, \ldots, X_n^*$ uniformly with replacement from $X_1, \ldots, X_n$
  - Estimate the parameter $\hat{\Theta}^*_b$

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

\[
\hat{\sigma} = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left( \hat{\Theta}^*_b - \left( \frac{1}{B} \sum_{b=1}^{B} \hat{\Theta}^*_b \right) \right)^2}
\]
Does it always work?

**Theorem (Consistency of the estimate of the standard error)**

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

**Example where it goes wrong:**

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

Want to estimate $\theta$. The ML estimate of $\theta$ is simply the largest number we observe:

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

Estimating the SE by bootstrap is going to fail.

Estimating tails or extreme values by bootstrap is problematic.
Confidence sets by Bootstrap

Bootstrap - percentile - method:
1. Given sample \( x_1, \ldots, x_n \), estimate \( \hat{\theta} \).
2. Generate bootstrap replicates \( \hat{\theta}_1^*, \ldots, \hat{\theta}_B^* \).
3. Look at the histogram of the \( \hat{\theta}_b^* \).

\[ a := \frac{\alpha}{2} - \text{percentile} \quad b := 1 - \frac{\alpha}{2} \]

- \( CI = [a, b] \)
- It has coverage \( 1 - \alpha \) because
  \[ p_\theta ( \hat{\theta} \in CI) \geq 1 - \alpha \]

(approximate \( \gamma \),
because \( n, \theta \) finite)
Subsequently you can construct bootstrap tests in the obvious way.

\[ H_0: \hat{\theta} = \theta \quad \text{vs.} \quad H_1: \theta \neq 0 \]
Bayesian statistics

Frequentist statistics:

- Probability = limiting frequency
- Parameters $\theta$ 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 $\{ f_{\theta}(x) \}$, we call $f(x|\theta)$ the likelihood of the data given the parameter $\theta$.
- Goal: investigate $\theta$. 
We assume that we have a prior belief about the parameters \( \Theta \): \( f(\Theta) \) prior distribution.

\[
f(\Theta | x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n | \Theta) \cdot f(\Theta)}{\int f(x_1, \ldots, x_n | \Theta) \cdot f(\Theta) \, d\Theta}
\]

The posterior is a distribution.

Now we update our belief; we compute the posterior using Bayes rule: \( f(\Theta | x_1, \ldots, x_n) \).

- Observe data \( x_1, \ldots, x_n \) iid.
- Now we update our belief; we compute the posterior using Bayes rule: \( f(\Theta | x_1, \ldots, x_n) \).
- Now you can make statements based on the posterior.
  - If you want to make the "best guess" for \( \Theta \), 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