# Machine Learning Methods

Siger

September 21, 2023

Si Dieu est infini, alors je suis une partie de Dieu sinon je serai sa limite...

# Contents

# Part I

# Collect and Pre-process Data

# Chapter 1

# Data cleaning

[1]

## 1.1 Data Quality

## 1.2 Validity

- **Data-Type Constraints**: for a given column a fixed data-type must be associated with.
- **Range Constraints**: only a range of values should be taken.
- **Mandatory Constraints**: some columns cannot be empty.
- **Unique Constraints**: across a given dataset a field or a combination of variables.
- **Foreign-key constraints**: a foreign key column cannot have a value that does not exist in the primary key.
- **Regular expression patterns**: text fields that have to follow a given alphanumerical pattern.
- **Cross-field validation**: consistency of values, for example considering a given man, his birth date have to be older than his death date.

## 1.3 Accuracy

The degree to which the data is close to the true value.

## 1.4 Completeness

The degree to which the all the required data is known.

## 1.5 Consistency

The degree to which the data is consistent, within the same data set or across multiple data sets.

## 1.6 Uniformity

The degree to which the data is specified using the same unit of measure.

## 1.7 The workflow

## 1.8 Inspection

Detect unexpected behavior in the data.

- **Data profiling**: summary statistics about the data, see ydata-profiling in Python.
- **Visualizations**: visualize the data using statistical metrics, see plotly
- **Software packages**: to note and check the constraints regarding the data see pydeequ

## 1.9 Cleaning

Fix or remove anomalies discovered in the above phase.

- **Irrelevant Data**: ask to the expert what can be the unnecessary columns, check them and remove them if they are not useful.
- **Duplicates**
- **Type conversion**: make sure the appropriate data type is associated with a given column.
- **Syntax errors**: white spaces, pad strings ...
- **Standardize**: same unit across the dataset, same pattern for text.
- **Scaling/Transformation**: in order to compare different scores for example.
- **Normalization**: useful for some statistical methods.
- **Missing values**:
  - Drop: only if the missing values in a column rarely and randomly occur.
  - Impute: many methods, *mean* is relevant when data is not skewed otherwise we should use *median*. A linear regression or a hot-deck (copying of values) approach can be taken as well, and more interestingly a *k-nearest* method approach.
  - Flag: let the missing value as it is.
- Outliers: Remove outliers only if they are harmful for the chosen model.
- In-record & cross-datasets errors: fix non-consistent situations like married kids, quantity being different of the one when we compute using other columns.

## 1.10 Verifying

Check correctness of the cleaning phase.

## 1.11 Reporting

Report about changes made, using one of the software summarising the data quality for example.

# Part II

# Statistics

# Chapter 2

# Fundamental probability concepts

## 2.1 Basic probability properties

### 2.1.1 What is a probability?

It is a mathematical measure of the uncertainty of a given event.

**Objectivist interpretation [3]** : assigns numbers describing some objective state, *Frequentist* interpretation claiming that the probability of a random event is quantified by the relative frequency in a given experiment.

**Subjectivist interpretation [3]** : assigns numbers quantifying the degree of belief that a given event occurs. *Bayesian* interpretation uses expert knowledge considered as subjective and represented by the prior, as well as experimental data represented by the likelihood. The normalized product of the 2 above quantity is the posterior probability distribution containing both expert knowledge and experimental data.

### 2.1.2 Properties

**Event and its opposite** $\mathbb{P}(\{A\}) + \mathbb{P}\left(\{\overline{A}\}\right) = 1$

**Not necessary mutually exclusive events** $\mathbb{P}(\{A \cup B\}) = \mathbb{P}(\{A\}) + \mathbb{P}(\{B\}) - \mathbb{P}(\{A \cap B\})$

**Independent events** $\mathbb{P}(\{A \cap B\}) = \mathbb{P}(\{A\}) \times \mathbb{P}(\{B\})$

**Conditional Probability** $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

**Law of Total Probability** $\begin{cases} (B_i)_{1 \leq i \leq n} : \text{ partition of a sample } \mathcal{S} \\ \forall i \in [\![1, n]\!],\ \mathbb{P}(\{B_i\}) \neq 0 \end{cases} \Rightarrow \mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i)\,\mathbb{P}(A|B_i)$

**Bayes' Theorem** Using Law of Total Probability:
$\begin{cases} (B_i)_{1 \leq i \leq n} : \text{ partition of a sample } \mathcal{S} \\ \forall i \in [\![1, n]\!],\ \mathbb{P}(\{B_i\}) \neq 0 \end{cases} \Rightarrow \mathbb{P}(B_i|A) = \dfrac{\mathbb{P}(B_i) \times \mathbb{P}(A|B_i)}{\sum_{k=1}^{n} \mathbb{P}(B_k)\,\mathbb{P}(A|B_k)}$

### 2.1.3 Moments

They are certain quantitative measures related to the shape of the function's graph. [2]

**$n^{th}$ moments of a random variable:** The $n^{th}$ moment about the origin of a random variable $X$ as denoted by $E(X^n)$, is defined to be:

$$\mathbb{E}(X^n) = \begin{cases} \sum_{x \in R_X} x^n f(x) \text{ if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x) dx \text{ if } X \text{ is continuous} \end{cases}$$

**Expected value:** The expected value of a random variable $X$ as denoted by $E(X)$, is defined to be:

$$\mathbb{E}(X) = \begin{cases} \sum_{x \in R_X} x f(x) \text{ if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x) dx \text{ if } X \text{ is continuous} \end{cases}$$

After normalized this moment by total mass we have the center of mass.

**Variance :** Let $X$ be a random variable with mean $\mu_X$. The variance of $X$ denoted by $\mathbb{V}(X)$ or $\sigma_X^2$ is defined by:

$$\mathbb{V}(X) = \mathbb{E}\left([X - \mu_X]^2\right)$$

After normalized this moment by total mass we have the moment of inertia.
If $X$ is a random variable with mean $\mu_X$ and variance $\sigma_X^2$ then:

$$\sigma_X^2 = \mathbb{E}(X^2) - \mu_X^2$$

And:

$$\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)$$

**Skewness and Kurtosis**

- Skewness: $\mathbb{E}\left(\left[\frac{X - \mu_X}{\sigma_X}\right]^3\right)$, indicates the direction (negative $\rightarrow$ left tail is longer, positive $\rightarrow$ right tail is longer) and relative magnitude of a distribution's deviation from the normal distribution.
- Kurtosis: $\mathbb{E}\left(\left[\frac{X - \mu_X}{\sigma_X}\right]^4\right)$, measures the outliers, data within one standard deviation will not contribute a lot to the kurtosis values conversely data exceeding one standard deviation will contribute a lot because of the fourth power.

### 2.1.4 Asymptotic properties

**Chebychev inequality** allows to find an estimate of the area between the values $\mu - k\sigma$ and $\mu + k\sigma$ for some given $k \neq 0$, showing that the area under $f(x)$ on the interval $[\mu - k\sigma, \mu + k\sigma]$ is at least $1 - k^2$. Let $X$ be a random variable with probability density function $f(x)$. If $\mu$ and $\sigma > 0$ are the mean and standard deviation of $X$ then:

$$\boxed{\mathbb{P}(\{|X - \mu| < k\sigma\}) \geq 1 - \frac{1}{k^2}}$$

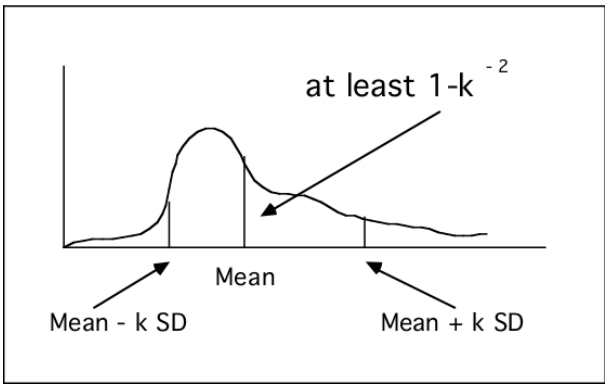


Figure 2.1: Illustration of Chebychev inequality

**Markov inequality**

$$X \neq \underline{0} \Rightarrow \mathbb{P}\left(\{X \geq t\}\right) \leq \frac{\mathbb{E}\left(X\right)}{t}$$

**Theorem weak law of large numbers:** Let $(X_i)_{1 \leq i \leq 1}\, n$: independent & identically distributed RV

$$\lim_{n \to \infty} \mathbb{P}\left(\left\{\left|\overline{S}_n - \mu\right| \geq \epsilon\right\}\right) = 0 \text{ with } \overline{S}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

**Convergence in probability** Suppose $(X_i)_{1 \leq i \leq 1}\, n$ is a sequence of random variables de- fined on a sample space S. The sequence "converges in probability" to the random variable X if, for any $\epsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}\left(\{|X_n - X| < \epsilon\}\right) = 1$$

**Convergence almost surely** Suppose the RV $X$ and $(X_i)_{1 \leq i \leq 1}\, n$ is a sequence of random variables de- fined on a sample space S. The sequence $X_n(\omega)$ "converges almost surely" to $X(\omega)$ if

$$\mathbb{P}\left(\left\{w \in S | \lim_{n \to \infty} X_n(\omega) = X(\omega)\right\}\right) = 1$$

**Properties**

- For a Bernoulli distribution, $\overline{S}_n$ converges in probability to $p$
- For a Normal distribution, $\overline{S}_n$ converges almost surely to $\mu$

### 2.1.5 Central Limit Theorem

The central limit therorem (Lindeberg-Levy Theorem) states that for any population distribution, the distribution of the standardized sample mean is approximately standard normal with better approximations obtained with the larger sample size.

$$\begin{cases} (X_i)_{1 \leq i \leq 1}\, n \hookrightarrow ?(\mu, \sigma^2) \\ n \to \infty \end{cases} \quad \Rightarrow \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \hookrightarrow \mathcal{N}(0,1)$$

### 2.1.6 Convergence in distribution

Consider $X$ with its cumulative density function $F$ and $(X_i)_{1 \leq i \leq 1}\, n$ with their cdf $(F_i)_{1 \leq i \leq n}$:

$$\lim_{n \to \infty} F_n(x) = F(x) \Rightarrow X_n \text{ "converges in distribution" to } X$$

### 2.1.7 Lévy Continuity Theorem

$$\begin{cases}(X_i)_{1\leq i\leq 1}\, n\text{RV}\\ (F_i)_{1\leq i\leq 1}\, n\text{distribution functions}\\ (M_{X_i})_{1\leq i\leq n}\, \text{moment generating function}\end{cases} \qquad \forall t\in[-h,h]\lim_{n\to\infty} M_{X_n}(t)=M_X(t)\Rightarrow \lim_{n\to\infty} F_n(x)=F(x)$$

## 2.2 Bivariate case

**Joint probability density function** Let $(X,Y):(\Omega_X,\Omega_Y)\to(R_X,R_Y)$ and $f:R_X\times R_Y\to\mathbb{R}$

$$\forall(x,y)\in R_X\times R_Y, f(x,y)=\mathbb{P}\left(\{X=x,Y=y\}\right)\Leftrightarrow$$
f is the joint probability density function for $X$ and $Y$

**Marginal probability density function** Let for all $(x,y)\in R_X\times R_Y$: $f(x,y)$ be the joint probability density of $X$ and $Y$

$$\begin{cases} f_1(x)=\displaystyle\int_{-\infty}^{\infty} f(x,y)dy \text{is the marginal probability density of } X\\ f_2(y)=\displaystyle\int_{-\infty}^{\infty} f(x,y)dx \text{is the marginal probability density of } Y\end{cases}$$

**Joint cumulative probability distribution function** Let $F:\mathbb{R}^2\to\mathbb{R}$

$$\forall(x,y)\in\mathbb{R}^2, F(x,y)=\mathbb{P}\left(\{X\leq x,Y\leq y\}\right)=\int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)dudv\Leftrightarrow$$
F is the joint cumulative probability density function for $X$ and $Y$

From the fundamental theorem of calculus: $f(x,y)=\dfrac{\partial^2 F(x,y)}{\partial x\partial y}$

**Conditional expectation** The conditional mean of $X$ given $Y=y$ is defined as:

$$\mathbb{E}\left(X|y\right)=\begin{cases}\displaystyle\sum_{x\in R_X} xg(x/y)\Leftarrow X \text{ discrete}\\ \displaystyle\int_{-\infty}^{\infty} xg(x/y)dx\Leftarrow X \text{ continuous}\end{cases}$$

Properties:

$$\begin{cases}\mathbb{E}_X\left(\mathbb{E}_{y|x}\left(Y|X\right)\right)=\mathbb{E}_y\left(Y\right)\\ \mathbb{E}\left(Y|\left\{X=x\right\}\right)=\mu_Y+\rho\dfrac{\sigma_Y}{\sigma_X}(x-\mu_X)\end{cases}$$

**Conditional Variance**

$$\begin{cases}\mathbb{V}\left(Y|x\right)=\mathbb{E}\left(Y^2|x\right)-\mathbb{E}\left(Y|x\right)^2\\ \mathbb{E}_x\left(\mathbb{V}\left(Y|X\right)=(1-\rho^2)\mathbb{V}\left(Y\right)\right)\end{cases}$$

## 2.3 Distribution function

### 2.3.1 Definition of probability density function (pdf):

Let $R_X$ be the space of the random variable $X$. The function: $f:R_X\to\mathbb{R}$ defined by:

$f(x)=\mathbb{P}\left(\{X=x\}\right)$ if $X$ is discrete.

$f(x)=\mathbb{P}\left(\{X\in A\}\right)=\displaystyle\int_A f(x)dx$ if $X$ is continuous, with $A$ a set of real numbers.

is called probability density function of $X$.

### 2.3.2 Definition of cumulative density function (cdf):

Let $R_X$ be the space of the random variable $X$. The function: $F : R_X \to \mathbb{R}$ defined by:

$$F(x) = \mathbb{P}(\{X \leq x\}) \text{ if } X \text{ is discrete.}$$
$$F(x) = \mathbb{P}(\{X \leq x\}) = \int_{-\infty}^{x} f(t)dt \text{ if } X \text{ is continuous, with } A \text{ a set of real numbers.}$$

### 2.3.3 Percentile for continuous random variables.

Let $p \in [0; 1]$, a $100p^{th}$ percentile of the distribution of a random variable $X$ is $q \in \mathbb{R}$ satisfying:

$$\mathbb{P}(\{X \leq q\}) \leq p$$
(Recall that the $F$ is a monotonically increasing function, then it has an inverse $F^{-1}$)
$$q = F^{-1}(p)$$

A $100p^{th}$ is a measure of location for the probability distribution in the sense that $q$ divides the distribution of the probability mass into 2 parts, one having probability mass $p$ and other having probability mass $1 - p$

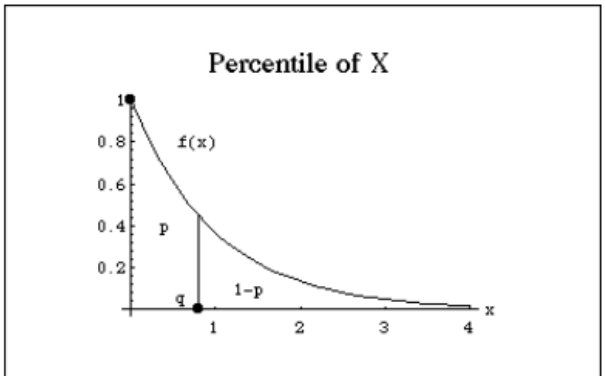


Figure 2.2: Percentile

The $50^{th}$ percentile of any distribution is called median of the distribution.

# Chapter 3

# Distributions

## 3.1 Discrete distributions with finite support

### 3.1.1 Bernoulli

### 3.1.2 Rademacher

### 3.1.3 Binomial

### 3.1.4 Beta-Binomial

### 3.1.5 Degenerate

### 3.1.6 Uniform

### 3.1.7 Hypergeometric

### 3.1.8 Negative Hypergeometric

### 3.1.9 Poisson Binomial

### 3.1.10 Fisher's noncentral hypergeometric

### 3.1.11 Benford's law

### 3.1.12 Zipf's law

### 3.1.13 Zipf-Mandelbrot law

# Chapter 4

# Bayesian approach

## 4.1 Components

### 4.1.1 Bayesian concept learning

Let be $\mathcal{D}$ the data, $h$ the hypothesis taken in account

### 4.1.2 Likelihood

$p(\mathcal{D}|h)$ the probability to get the observed data considering the hypothesis $h$.

### 4.1.3 Prior

$p(h)$ the probability of our hypothesis, many prior can be used, and this **subjective** aspect of Bayesian reasoning is a source of much controversy.

### 4.1.4 Posterior

The posterior is simply the likelihood times the prior, normalized.

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h) \times p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}, h')\, p(h')}$$

## 4.2 Summarizing posterior distributions

### 4.2.1 MAP (Maximum A Posteriori) estimation

Although most appropriate choice for:

$$\begin{cases} \text{Real valued quantity} & \rightarrow \textit{posterior median or mean} \\ \text{Discrete} & \rightarrow \textit{vector of posterior marginals} \end{cases}$$

The most popular choice is *posterior mode* aka MAP, because it reduces to optimization problems for which efficient algorithms often exist.
Some point to be aware about MAP:

- No measure of uncertainty
- Plugging in the MAP estimate can result in overfitting
- The mode is an untypical point, unlike the mean or median the mode is a point of measure 0, it does not take the volume of the space into account.
- MAP estimation is not invariant to reparameterization, for example passing from centimeters to inches can break things.)
  The MLE does not suffer from this since the likelihood is a function not a probability density

### 4.2.2 Credible intervals

With point estimates, we want a measure of confidence.

$$C_\alpha(\mathcal{D}) = (l, u) : \mathbb{P}(\{l \leq \theta \leq u | \mathcal{D}\}) \geq 1 - \alpha$$

In general, credible intervals are usually what people want to compute but confidence intervals are usually what they actually compute, because most people are taught frequentist statistics but not Bayesian statistics.
Sometimes with central intervals there might be points be outside the CI which have higher probability density.
More formally $p^*$ such that:

$$1 - \alpha = \int_{\theta : p(\theta|\mathcal{D}) > p^*} p(\theta|\mathcal{D}) d\theta$$

Then the HPD such that:

$$\mathcal{D} = \{\theta : p(\theta|\mathcal{D}) \geq p^*\}$$

## 4.3 Bayesian Model Selection

A more efficient approach than cross-validation, meaning fitting $k$ times each model, is to compute the posterior over models.

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m \in \mathcal{M}} p(m|\mathcal{D})}$$

From this we can compute the MAP model $\hat{m} = \underset{m}{\operatorname{argmax}}\, p(m|\mathcal{D})$

Then we have the marginal likelihood: $p(\mathcal{D}|\hat{m}) = \int p(\mathcal{D}|\hat{m})p(\theta|\hat{m})d\theta$

### 4.3.1 Baysian Occam's razor

In integrating out the parameters rather than maximizing them we are automatically protected from overfitting: model with more parameters do not necessarily have higher marginal likelihood.
A way to understand the Bayesian Occam's razor effect is to remember that probabilities must sum to one, meaning $\sum_{\mathcal{D}'} p(\mathcal{D}'|m) = 1$. Complex models, which can predict many things, must spread their probability mass thinly, and hence will not obtain as large a probability for any given data set as simpler models.

### 4.3.2 Computing the marginal likelihood (evidence)

For a fixed model we often write:

$$p(\boldsymbol{\theta}|\mathcal{D}, m) \propto p(\boldsymbol{\theta}|m)p(\mathcal{D}|\boldsymbol{\theta}, m)$$

This valid since $p(\mathcal{D}|m)$ is constant. However when comparing models we need to know how to compute the marginal likelihood, $p(\mathcal{D}|m)$. In general this can be quite hard, since we have to integrate over all possible parameter values, but when we have a conjugate prior, it is easy to compute.
Let $p(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{Z_0}$ be our prior, where $q(\boldsymbol{\theta})$ is an unnormalized distribution, and $Z_0$ is the normalization constant of the prior. Let $p(\mathcal{D}|\boldsymbol{\theta}) = \frac{q(\mathcal{D}|\boldsymbol{\theta})}{Z_l}$ be the likelihood, where $Z_l$ contains any constant factors in the likelihood. Finally let $p(\boldsymbol{\theta}|\mathcal{D}) = \frac{q(\boldsymbol{\theta}|\mathcal{D})}{Z_N}$ be our posterior where $q(\boldsymbol{\theta}|\mathcal{D}) = q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})$ is the

unnormalized posterior, and $Z_N$ is the normalization constant of the posterior.

We have: $\begin{cases} p(\boldsymbol{\theta}) = \dfrac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \\ \dfrac{q(\boldsymbol{\theta}|\mathcal{D})}{Z_N} = \dfrac{q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})}{Z_l Z_0 p(\mathcal{D})} \\ p(\mathcal{D}) = \dfrac{Z_N}{Z_0 Z_l} \end{cases}$

Simpler approach

- **BIC** In general $p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$ can be quite difficult to compute. A popular approximation is: $BIC \triangleq \log(p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})) - \dfrac{dof(\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})}{2}\log(N) \approx \log p(\mathcal{D})$
- **AIC**: $AIC(m, \mathcal{D}) \triangleq \log(p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{MLE})) - dof(m)$
  This is derived from Frequentist framework and cannot be interpreted as an approximation to the marginal likelihood. The penalty of AIC is less than BIC, it causes AIC pick more complex models. That can be better for predictive accuracy.
- Effect of the prior.
  If the prior is unknown, the correct Bayesian procedure is to put a prior on the prior. That is we should put a prior on the hyper-parameter $\alpha$ as well as the parameters $\boldsymbol{w}$. To compute the marginal likelihood we should integrate out all unknowns, we should compute: $\int\int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|\alpha, m)p(\alpha|m)d\boldsymbol{w}d\alpha$
  A computational shortcut is to optimize $\alpha$ rather than integrating it out. That is, we use $p(\mathcal{D}|m) \approx \int p(\mathcal{D}\boldsymbol{w})p(\boldsymbol{w}|\alpha, m)d\boldsymbol{w}$. where $\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, p(\mathcal{D}|\alpha, m) = \underset{\alpha}{\operatorname{argmax}} \int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|\hat{\alpha}, m)d\boldsymbol{w}$

### 4.3.3 Bayes Factors

When prior on models is uniform, then model selection is equivalent to picking the model with the highest marginal likelihood. Now suppose we just have two models we are considering, call them the null hypothesis, $M_0$ and the alternative hypothesis, $M_1$.

$$BF_{1,0} \triangleq \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} = \frac{\left.\frac{p(M_1|\mathcal{D})}{p(M_0|\mathcal{D})}\right\} \textit{Posterior odds}}{\left.\frac{p(M_1)}{p(M_0)}\right\} \textit{Prior odds}}$$

- *Posterior odds*: quantifies relative plausibility of the rival hypotheses **after** having seen the data.
- *Bayes Factor*, $BF_{1,0}$, quantifies the evidence provided by the data, this is like a likelihood ratio, except we integrate out the parameters, which allows us to compare models of different complexity.
- *Prior odds*: quantifies relative plausibility of the rival hypotheses **before** seeing the data.

| **Bayes Factor $BF(1,0)$** | **Interpretation** |
|---|---|
| $BF < \frac{1}{100}$ | Decisive evidence for $M_0$ |
| $BF < \frac{1}{10}$ | Strong evidence for $M_0$ |
| $\frac{1}{10} < BF < \frac{1}{3}$ | Modest evidence for $M_0$ |
| $\frac{1}{3} < BF < 1$ | Weak evidence for $M_0$ |
| $1 < BF < 3$ | Weak evidence for $M_1$ |
| $3 < BF < 10$ | Modest evidence for $M_1$ |
| $BF > 10$ | Strong evidence for $M_1$ |
| $BF > 100$ | Decisive evidence for $M_1$ |

### 4.3.4 Jeffreys-Lindley paradox

Problems can arise when we use improper priors (i.e. priors that do not integrate to 1) for model selection/ hypothesis testing, even though such priors may be acceptable for other purposes. In particular the Bayes Factor will always favor the simplest model since the probability of the observed data under a complex model with a very diffuse prior will be very small. Thus it is important to use proper priors when doing model selection.

## 4.4 Priors

The most controversial aspect of Bayesian statistics is its reliance on priors

### 4.4.1 Uninformative priors

If we do not have strong evidence on what $\theta$ should be, it is common to use an uninformative priors, to "let the data speak for itself".
One might think that the most uninformative prior would be the uniform distribution: $Beta(1,1)$, but the posterior would then be: $\mathbb{E}(\theta|\mathcal{D}) = \frac{N_1+1}{N_1+N_0+2}$, whereas the MLE is $\frac{N_1}{N_1+N_0}$.
As by decreasing the magnitude of the pseudo counts, we can lessen the impact of the prior, we can argue that the most non-informative prior is:

$$\lim_{\epsilon \to 0} Beta(\epsilon, \epsilon) = Beta(0,0)$$

Called the *Haldane prior*, it is an improper prior.
In general it is advisable to perform a some kind of sensitivity analysis, in which one checks how much one's conclusions or prediction change in response to change in the modelling assumptions which includes the choice of the prior and the likelihood as well. If the conclusion are relatively insensitive to the modelling assumption, one can have more confidence in the results.

### 4.4.2 Jeffreys priors

Harold Jeffreys designed a general purpose technique for creating non-informative priors. The key observation is that if $p(\phi)$ is non-informative then any re-parametrization of the prior, such as $\theta = h(\phi)$ for some function $h$ should also be non-informative.

- Start with a variable change: $p_\theta(\theta) = p_\phi(\phi)\left|\frac{d\phi}{d\theta}\right|$
- Consider the following constraint: $p_\phi(\phi) \propto \sqrt{\mathcal{I}(\phi)}$, where $\mathcal{I}(\phi)$ is the Fisher information. $\mathcal{I}(\phi) \triangleq -\mathbb{E}\left(2 \times \frac{d\log(p(X|\phi))}{d\phi}\right)$. This a measure of the curvature of the expected negative log likelihood and hence a measure of stability of the MLE.
- Now $\frac{d\log(p(x|\theta))}{d\theta} = \frac{d\log(p(X|\phi))}{d\phi}\frac{d\phi}{d\theta}$
- $\mathcal{I}(\theta) = \mathcal{I}(\phi)\left(\frac{d\phi}{d\theta}\right)^2$
- $\sqrt{\mathcal{I}(\theta)} = \sqrt{\mathcal{I}(\phi)}\left|\frac{d\phi}{d\theta}\right|$
- Finally $p_\theta(\theta) = p_\phi(\phi)\left|\frac{d\phi}{d\theta}\right| \propto \sqrt{\mathcal{I}(\phi)}\left|\frac{d\phi}{d\theta}\right| = \sqrt{\mathcal{I}(\theta)}$

### 4.4.3 Robust priors

To prevent an undue influence on the result, we build priors having heavy tails, which avoids forcing things to be too close to the prior mean.

### 4.4.4 Mixture of conjugate priors

Conjugate priors simplify the computation of robust priors, but are often not robust, and not flexible enough to encode our prior knowledge. However it turns out that a mixture of conjugate priors is also conjugate, and seem to be a good compromise.

## 4.5 Hierarchical and Empirical Bayes

### 4.5.1 Hierarchical Bayes

A key requirement for computing the posterior $p(\theta|\mathcal{D})$ is the specification of a prior $p(\theta|\eta)$ where $\eta$ are the hyper-parameters. A Bayesian approach is to put a prior on our priors. This is an example of a **hierarchical Bayesian Model**.

### 4.5.2 Empirical Bayes

In hierarchical Bayesian models, we need to compute the posterior on multiple levels of latent variables. For example, in a two-level model, we need to compute: $p(\eta, \theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta|\eta)p(\eta)$
We can approximate the posterior on the hyper-parameters with a point-estimate, $p(\eta|\mathcal{D}) \approx \delta_{\hat{\eta}}(\eta)$ where $\hat{\eta} = \text{argmax}_\eta\, p(\eta|\mathcal{D})$. Since $\eta$ is typically much smaller than $\theta$ in dimensionality, it is less prone to overfitting, so we can safely use a uniform prior on $\eta$. Then the estimate becomes:

$$\hat{\eta} = \operatorname*{argmax}_\eta p(\mathcal{D}|\eta) = \operatorname*{argmax}_\eta \int p(\mathcal{D}|\theta)p(\theta|\eta)d\theta$$

This overall approach is called **Empirical Bayes**
Empirical Bayes violates the principle that the prior should be chosen independently of the data. However, we can just view it as a computationally cheap approximation to inference in a hierarchical Bayesian model, just as we viewed MAP estimation as an approximation to inference in the one level model $\theta \to \mathcal{D}$. In fact, we can construct a hierarchy in which the more integrals one performs, the "more Bayesian" one becomes:

| **Method** | **Definition** |
|---|---|
| Maximum likelihood | $\hat{\theta} = \text{argmax}_\theta\, p(\mathcal{D}|\theta)$ |
| MAP estimation | $\hat{\theta} = \text{argmax}_\theta\, p(\mathcal{D}|\theta)p(\theta|\eta)$ |
| ML-II (Empirical Bayes) | $\hat{\eta} = \text{argmax}_\eta \int p(\mathcal{D}|\theta)p(\theta|\eta)d\theta = \text{argmax}_\eta\, p(\mathcal{D}|\eta)$ |
| MAP-II | $\hat{\eta} = \text{argmax}_\eta \int p(\mathcal{D}|\theta)p(\theta|\eta)p(\eta)d\theta = \text{argmax}_\eta\, p(\mathcal{D}|\eta)p(\eta)$ |
| Full Bayes | $p(\theta, \eta|\mathcal{D}) \approx p(\mathcal{D}|\theta)p(\theta|\eta)p(\eta)$ |

## 4.6 Bayesian Decision Theory

We can formalize any given statistical decision problem as a game against nature (as opposed to a game against other strategic players, which is the topic of game theory). In this game, nature picks a state or parameter or label, $y \in \mathcal{Y}$, unknown to us, and then generates an observation, $\boldsymbol{x} \in \mathcal{X}$ which we get to see. We then have to make a decision, that is, we have to choose an action $a$ from some **action space** $\mathcal{A}$. Finally we incur some **loss**, $L(y, a)$, which measures how compatible our action $a$ is with nature's hidden state $y$.
Our goal is to devise a decision procedure or policy, $\delta : \mathcal{X} \to \mathcal{A}$ which specifies the optimal action for each possible input which specifies the optimal action for each possible input, meaning the action that minimizes the expected loss:

$$\delta(\boldsymbol{x}) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}\left(L(y, a)\right)$$

In the Bayesian vision, the expected value of $y$ given the data we have seen so far, whereas in the frequentist vision the expected value refers to $x$ and $y$ that we expect to see in the future.
In the Bayesian vision the optimal action having observed $\boldsymbol{x}$ is defined as the action $a$ that minimizes the **posterior expected loss**:

$$\rho(a|\boldsymbol{x}) \triangleq \mathbb{E}_{p(y|x)}\left(L(y, a)\right) = \sum_y L(y, a)p(y|x)$$

Hence the Bayes estimator also called Bayes decision rule is given by:

$$\delta(\boldsymbol{x}) = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, \rho(\boldsymbol{a}|\boldsymbol{x})$$

### 4.6.1 Bayes estimators for common loss functions

- **MAP** estimate minimizes 0-1 loss: $L(y,a) = \mathbb{I}_{y \neq a} \begin{cases} 0 \text{ if } a = y \\ 1 \text{ else} \end{cases}$
- **Reject option**, in classification problems where $p(y|\boldsymbol{x})$ is very uncertain we may prefer to choose a reject action, in which we refuse to classify the example as any of the specified classes. Let choosing $a = C+1$ correspond to picking the reject action, and choosing $a \in \{1, ..., C\}$ correspond to picking one of the classes.
  $L(y = j, a = i) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, ..., C\} \\ \lambda_r & \text{if } i = C+1 \\ \lambda_s & \text{otherwise} \end{cases}$
  where $\lambda_r$ is the cost of the reject action, and $\lambda_s$ is the cost of a substitution error.
- **Squared Error ($l_2$)** for a continuous parameters. $L(y,a) = (y-a)^2$
- **Absolute Error ($l_1$)** more robust against outliers. $L(y,a) = |y-a|$. The optimal point is the median.
- **Supervised learning** considering a prediction function $\delta : \mathcal{X} \to \mathcal{Y}$ and some cost function $l(y, \delta(x))$. Then the loss incurred by taking action $\delta$ when the unknown state of nature is $\theta$ (the parameters of the data generating the mechanism). $L(\boldsymbol{\theta}, \delta) \triangleq \mathbb{E}_{(\boldsymbol{x},y)\ p(\boldsymbol{x},y|\boldsymbol{\theta})} \left(l(y, \delta(\boldsymbol{x}))\right) = \sum_{\boldsymbol{x}} \sum_{y} L\left(y, \delta(\boldsymbol{x}) p(\boldsymbol{x}, y|\boldsymbol{\theta})\right)$

### 4.6.2 Model evaluation metrics

- **False positive vs False negative trade-off** for binary decision problems three are 2 types of errors:
  1. false positive (false alarm) if $\hat{y} = 1 \wedge y = 0$
  2. false negative (missed detection) if $\hat{y} = 0 \wedge y = 1$

  We can consider the loss matrix:

| **Headers** | $\boldsymbol{y=1}$ | $\boldsymbol{y=0}$ |
|---|---|---|
| $\hat{\boldsymbol{y}} = \mathbf{1}$ | 0 | $L_{FP}$ |
| $\hat{\boldsymbol{y}} = \mathbf{0}$ | $L_{FN}$ | 0 |

  where $L_{FN}$ is the cost of a false negative and $L_{FP}$ the cost of a false positive.
- **ROC curves** From the below table

| **Headers** | | **Truth** | | **Count** |
|---|---|---|---|---|
| **Estimate** | 1 | $TP$ | $FP$ | $\hat{N}_+ = TP + FP$ |
| | 0 | $FN$ | $TN$ | $\hat{N}_- = FN + TN$ |
| **Count** | | $N_+ = TP + FN$ | $N_- = FP + TN$ | $N = N_+ + N_- = \hat{N}_+ + \hat{N}_-$ |

  we can generate the *confusion matrix* is the below table

| **Headers** | $\boldsymbol{y=1}$ | $\boldsymbol{y=0}$ |
|---|---|---|
| $\hat{\boldsymbol{y}} = \mathbf{1}$ | $\frac{TP}{N}$ (sensitivity/recall) | $\frac{FP}{N}$ (error type I/ false alarm) |
| $\hat{\boldsymbol{y}} = \mathbf{0}$ | $\frac{FN}{N}$ (error type II/ missed detection) | $\frac{TN}{N}$ (specificity) |

- **Precision recall curves** When trying to detect a rare event the number of negatives is very large, hence comparing *sensitivity* and *the error of type I* is not very informative. We would then like to use a measure that only talks about positives.

- **precision** $= \frac{TP}{\hat{N}_+}$
- **recall** $= \frac{TP}{N_+}$

A **precision recall curve** is a plot of *precision* vs *recall*.

- **F-scores** is the *harmonic mean of precision and recall*:
  $F_1 \triangleq \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$

# Chapter 5

# Frequentist approach

## 5.1 Sampling distribution

### 5.1.1 Sampling Distributions of an estimator

In frequentist statistic a parameter estimate $\hat{\theta}$ is computed by applying an estimator $\delta$ to some data $\mathcal{D}$, so $\hat{\theta} = \delta(\mathcal{D})$. The uncertainty in the parameter estimate can be measured by computing the *sampling distribution of the estimator*. Imagine sampling many different datasets $\mathcal{D}^{(s)}$ from some true model $p(\cdot|\theta^*)$ meaning $\mathcal{D}^{(s)} = \left\{x_i^{(s)} \hookrightarrow p(\cdot|\theta^*)\right\}_{1\leq i\leq N}$ for $1 \leq s \leq S$ and $\theta^*$ is the true parameter. Now apply the estimator $\hat{\theta}(\cdot)$ to each $\mathcal{D}^{(s)}$ to get a set of estimates $\{\hat{\theta}(\mathcal{D}^{(s)})\}_{1\leq s\leq S}$.
As we late $S \to \infty$, the distribution induced on $\hat{\theta}(\cdot)$ is the **sampling distribution of the estimator**.

### 5.1.2 Bootstrap

It is a simple *Monte Carlo* technique to approximate the sampling distribution. The idea is that if we knew the true parameters $\theta^*$, we could generate $S$ fake datasets of size $N$, from the true distribution. We could then compute our estimator from each sample, and use the empirical distribution of the resulting samples as our estimate of the sampling distribution.
Since $\theta$ is unknown, the idea of the **parametric bootstrap** is to generate the samples using $\hat{\theta}(\mathcal{D})$ instead. An alternative, called **non-parametric bootstrap** is to sample the $x_i^s$ (with replacement) from the original data $\mathcal{D}$ and then compute the induced distribution as before.

## 5.2 Fequentist decision theory

In Frequentist decision theory there is a loss function and a likelihood, but there is no prior and hence no posterior or posterior expected loss. Thus there is no automatic way of deriving an optimal estimator, unlike the Bayesian case.
Instead, we are free to choose any estimator or decision procedure $\delta : \mathcal{X} \to \mathcal{A}$ we want.
Having chosen an estimator, we define its expected loss or risk as follows

$$R(\theta^*, \delta) \triangleq \mathbb{E}_{p(\tilde{\mathcal{D}}|\theta^*)}\left(L(\theta^*, \delta(\tilde{\mathcal{D}}))\right) = \int L\left(\theta^*, \delta(\tilde{\mathcal{D}})\right) p(\tilde{\mathcal{D}}) d\tilde{\mathcal{D}}$$

where $\tilde{\mathcal{D}}$ is data sampled from 'nature's distribution' which is represented by parameter $\theta^*$. Whereas the Bayesian posterior expected loss:

$$p(a, \mathcal{D}, \pi) \triangleq \mathbb{E}_{p(\theta|\mathcal{D},\pi)}\left(L(\theta, a)\right) = \int_\Theta L(\theta, \boldsymbol{a}) p(\theta|\mathcal{D}, \pi) d\theta$$

We see that the Bayesian approach averages over $\theta$, which is unknown, and conditions on $\mathcal{D}$ which is known. Unlike the frequentist approach averages over $\tilde{\mathcal{D}}$, thus ignoring the observed data, and conditions on $\theta^*$ which is unknown.

### 5.2.1 Bayes risk

How to chose amongst the estimators? We need some way to convert $R(\theta^*, \delta)$ into single measure of quality, $R(\delta)$ which does not depend on knowing $\theta^*$. One approach is to put a prior on $\theta^*$ and then to define **Bayes risk** of an estimator as follows:

$$R_B(\delta) \triangleq \mathbb{E}_{p(\theta^*)}\left(R(\theta^*, \delta)\right) = \int R(\theta^*, \delta) p(\theta^*) d\theta^*$$

A **Bayes estimator** or **Bayes decision rule** is one which minimizes the expected risk: $\delta_B \triangleq \underset{\delta}{\operatorname{argmin}}\, R_B(\delta)$

**Connection Bayesian and Frequentist approaches to decision theory.**

- *Theorem 1* A Bayes estimator can be optained by minimizing the posterior expected loss for each $\boldsymbol{x}$
- *Theorem 2* Every admissible frequentist decision rule is a Bayes decision rule with respect to some possibly improper prior distribution.

**Minimax risk** Some frequentist statistic users avoid using Bayes risk since it requires the choice of a prior, although this is only in the evaluation of the estimator, not necessarily as part of its construction. An alternative approach is as follows:

1. Define the maximum risk of an estimator as:
   $R_{max}(\delta) \triangleq \max_{\theta^*} R(\boldsymbol{\theta}^*, \delta)$
2. A **minimax rule** is one which minimizes the maximum risk: $\delta_{MM} \triangleq \underset{\delta}{\operatorname{argmin}}\, R_{\max}(\delta)$

Minimax estimators have a certain appeal, however computing them can be hard and furthermore they are very pessimistic. In most statistical situations, excluding games theoretic ones, assuming nature is an adversary is not a reasonable assumption.

### 5.2.2 Admissible estimators

The basic problem with frequentis decision theory is that it relies on knowing the true distribution $p(\cdot|\theta^*)$ in order to evaluate the risk. However it might be the case that some estimators are worse than others regardless of the value of $\theta^*$.
In particular if for $\theta \in \Theta, R(\theta, \delta_1) \leq R(\theta, \delta_2)$ bayesthen we say that $\delta_1$ **dominates** $\delta_2$.
An estimator is said to be **admissible** if it is not strictly dominated by any other estimator.
**Admissibility is not enough**

## 5.3 Desirable properties of estimators

### 5.3.1 Consistent estimators

An estimator is said to be **consistent** if it eventually recovers the true parameters that generated the data as the sample size goes to infinity.

### 5.3.2 Unbiased estimator

The **bias** of an estimator is defined as

$$bias\left(\hat{\theta}(\cdot)\right) = \mathbb{E}_{p(\mathcal{D}|\theta^*)}\left(\hat{\theta}(\mathcal{D}) - \theta^*\right)$$

The estimator is unbiased when the bias is equal to 0.

### 5.3.3 Minimum variance estimators

A famous result called the **Cramerè-Rao lower bound** provides a lower bound on the variance of any unbiased estimator. More precisely: Let $(X_j)_{1\leq j\leq p} \hookrightarrow p(X|\theta_0)$ and $\hat{\theta}(\cdot)$ an unbiased estimator of $\theta^*$ Then, under various smoothness assumptions on $p(X|\theta_0)$ we have

$$\mathbb{V}(\hat{\theta}) \geq \frac{1}{nI(\theta^*)}$$

where $I(\theta^*)$ is the Fisher information matrix.

### 5.3.4 Bias-Variance Trade-off

As $MSE = variance + bias^2$
It might be wise to use a biased estimator, so long as it reduces our variance, assuming our goal is to minimize squared error.

## 5.4 Empirical Risk Minimization

### 5.4.1 Frequentist issue

Frequentist decision theory suffers from the fundamental problem that one cannot actually compute the risk function, since it relies on knowing the true data distribution. By contrast, the Bayesian posterior expected loss can always be computed since it conditions on the data rather that on $\theta^*$.
However there is one setting which avoids this problem, it is when the task is to predict observable quantities, as opposed to estimating hidden variables or parameters.
Instead of looking at loss functions of the form $L(\boldsymbol{\theta^*}, \delta(\mathcal{D}))$ let us look at loss functions of the form $L(y, \delta(\boldsymbol{x}))$.

Then the risk becomes: $R(p_*, \delta) \triangleq \mathbb{E}_{(\boldsymbol{x},y)\hookrightarrow p_*}\left(L(y, \delta(\boldsymbol{x}))\right) = \sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}} L(y, \delta(\boldsymbol{x}))p_*(\boldsymbol{x}, y)$ Where $p_*$ represents "nature's distribution", indeed this distribution is unknown, but a simple approach is to use the empirical distribution, derived from some training data to approximate $p_*(x, y) \approx p_{emp}(x, y) \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}(\boldsymbol{x})\delta_{y_i}(y)$ We define the empirical risk as follows:

$$R_{emp}(\mathcal{D}, \mathcal{D}) \triangleq R(p_{emp}, \delta) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \delta(x_i))$$

### 5.4.2 Regularized risk minimization

$$R'(\mathcal{D}, \delta) = R_{emp}(\mathcal{D}, \delta) + \lambda C(\delta)$$

where $C(\delta)$ measures the complexity of the prediction function $\delta(\boldsymbol{x})$ and $\lambda$ controls the strength of the complexity penalty. This approach is known as **regularized risk minimization**.

## 5.5 Components

### 5.5.1 Introduction

Avoid treating parameters as randome variables. The notion of variation across repeated trials forms the basis for modelling uncertainity.

### 5.5.2 Hypothesis Testing

A *frequentist* statistics, probabilities represent the frequencies at which particular events happen.

### 5.5.3 *p-value*

It is the heart of frequentist hypothesis testing, it tells us the probability of getting a particular test statistic $t$ as big as the one we have or bigger under the null hypothesis (that there is actually no effect). By convention we usually conclude an effect is *statistically significant* if the *p-value* is less than a threshold $\alpha$.

### 5.5.4 Confidence intervals

When we fit a model to our data we look for the *maximum of likelihood* parameters, meaning the parameters that are most consistent with our data. For each parameter we will able to construct 95% interval namely 95 of the 100 intervals generated will contain the true value of the parameter.
If $H_0 : \beta = 0$ is true, the probability of getting a 95% confidence interval that does not include 0 is less than 0.05. In other words, if the 95% confidence does not include 0, $p < 0.05$.

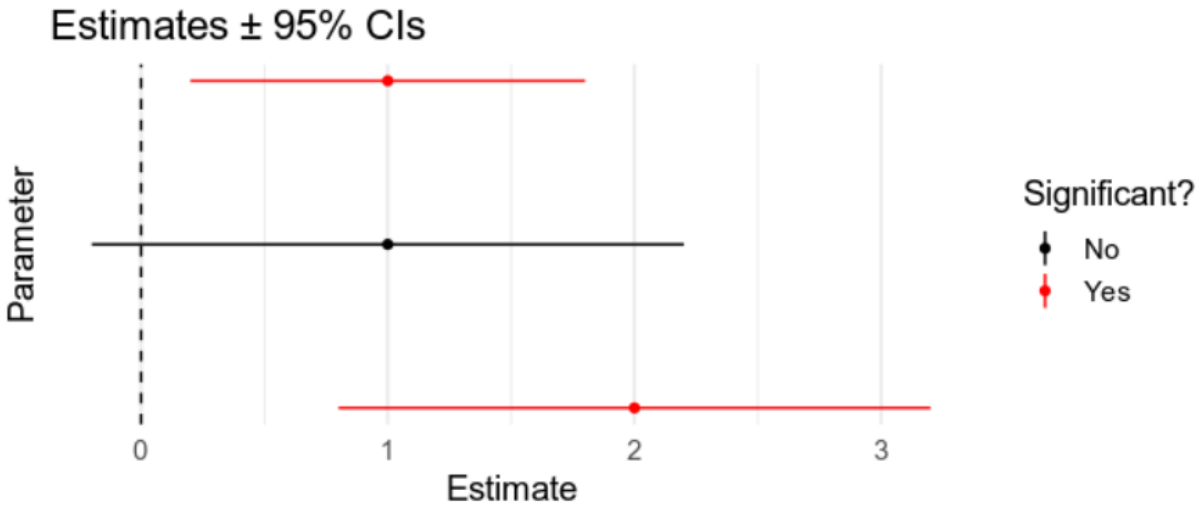


Figure 5.1: Confidence interval

### 5.5.5 Multiple comparisons

The more tests we run the more likely it is to we'll find at least one that is significant even though the null hypothesis is true. We can then apply a Bonferroni correction.
Let's say we are running $k$ tests, we can either adjust:

- the threshold $\alpha_{adj} = \frac{\alpha}{k}$ OR
- the *p-value* $p_{ajd} = k \times p$

## 5.6 Power Analysis

It has as general purpose to find the right sample number.

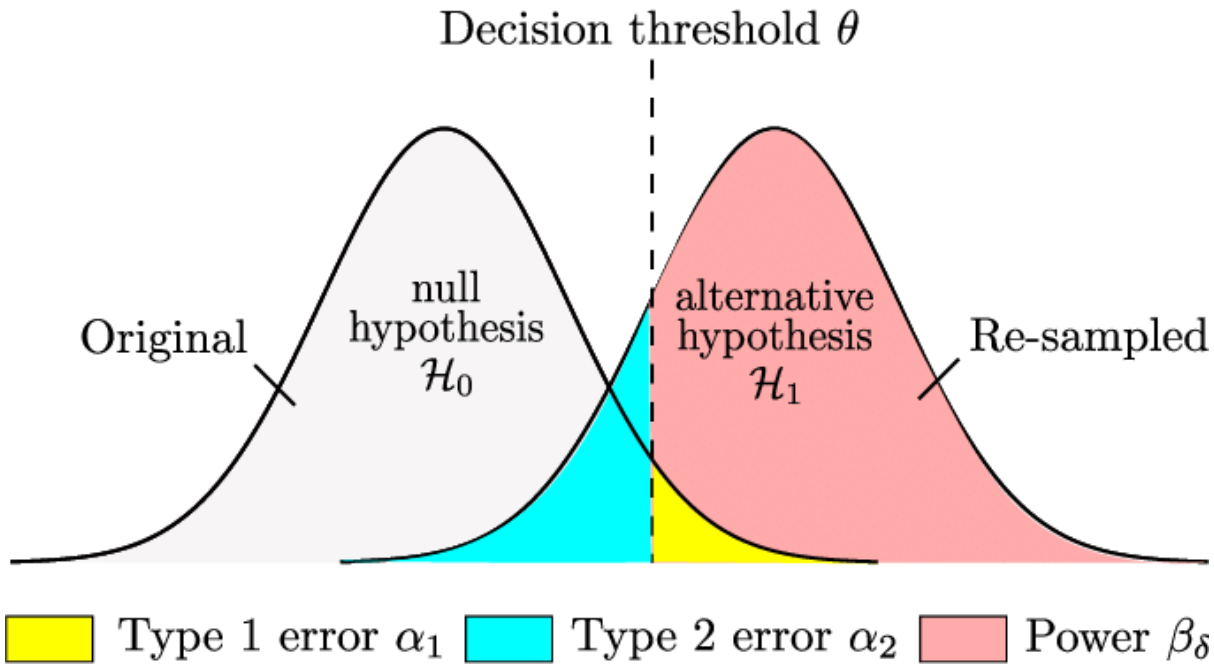


Figure 5.2: caption

| | $H_0$ ***is True*** | $H_1$ ***is True*** |
|---|---|---|
| **Do not reject** $H_0$ | Right decision | Type II Error ~ $\beta$ |
| **Reject** $H_0$ | Type I Error ~ $\alpha$ | Right decision |

### 5.6.1 Power of the test

Start by defining: $Power = 1 - \beta$, considering $H_1$ true it is the probability to correctly reject $H_0$

### 5.6.2 Significant threshold

Then propose $\alpha$, the probability to wrongly reject $H_0$. It will be the reference to which the *p-value* will be compared, the statistical test will be significant ($H_0$ rejected) if *p-value* $\leq \alpha$.

### 5.6.3 Effect size

It quantifies how meaningful the relationship between variables or the difference between group is, it indicates a practical significance.
While statistical significance (*p-value*) shows the existence of an effect, practical significance (*effect size*) shows if this effect is large enough to be meaningful in the real world.
There are dozens of measures for effect sizes, and the most common are *Cohen's d* and *Pearson's r*.

# Chapter 6

# Common statistical tests

## 6.1 Use of statistical tests

### 6.1.1 Terms

- *Paired* samples: one-to-one correspondence between data in the first and second set.
- *Matched* samples: every subject in one group with an equivalent in another.

### 6.1.2 Table of statistical hypothesis test

Statistical method table.

| | **Binomial/Discrete** | **Continuous, from Normal distribution** | **Continuous measurement (Score/Rank), from non-Normal distribution** |
|---|---|---|---|
| **Example of data sample** | List of patients recovering or not after a treatment | Reading of heart pressure from several patients | Ranking of several treatment efficiency |
| **Describe one data sample** | Proportions | Mean, Standard Deviation | Median |
| *Compare one data sample to a hypothetical distribution* | $\chi^2$ / *G-test* or Binomial test | 1-sample t-test | Sign test or Wilconox test |
| *Compare 2 paired samples* | Sign test | Paired t-test | Sign test or Wilconox test |
| *Compare 2 unpaired samples* | $\chi^2$ / *G-test* or Fisher's extract test | Unpaired t-test | Mann-Whitney test |
| *Compare 3 or more unmatched samples* | $\chi^2$ / *G-test* | 1-way ANOVA | Kruksal-Wallis test |
| *Compare 3 or more matched samples* | Cochrane Q test | Repeated-measures ANOVA | Friedman test |
| *Quantify association between 2 paired samples* | Contingency coefficients | Pearson correlation | Sperman correlation |

## 6.2 List of common statistical test

### 6.2.1 Binomial

To check if the deviations from a theoretically expected distribution of observations into 2 categories.

#### Assumptions

- Sample items are independent.

- Items are dichotomous and nominal.
- The sample size is significantly less than the population size
- The sample is a fiar representation of the population

**Frequentist** Let define a user-defined probability $p_0$, with $H_0 : p = p_0$ and $\begin{cases} H_1 : p \neq p_0 \text{: two-tailed test} \\ H_1 : p < p_0 \text{: left-tailed test} \\ H_1 : p > p_0 \text{: right-tailed test} \end{cases}$

**Bayesian** Define the prior distribution with a $Beta(a, b)$ distribution

*Return to the table.*

### 6.2.2 $\chi^2$ test

Either used to test *goodness-of-fit* and *independence* between 2 variables. It checks either if there is a significant difference between the expected and observed frequencies.

- *goodness-of-fit*: expected frequencies are computed with a theoretical relationship between observed frequencies
- *independence*: expected frequencies are computed with observed frequencies from the other sample

**Assumptions**

- simple random sample
- sample with a sufficiently large size is assumed, for small sample size see Cash test
- expected cell count has to be adequate, a rule of thumb is at least 5 for 2-by-2 table and 5 or more in 80% of cells in larger tables.
- Independence of the observations

**Frequentist** $\chi^2 = \sum_{i=1}^{n} \frac{\left(\frac{O_i}{N} - p_i\right)^2}{p_i}$ $\begin{cases} O_i \text{: number of observations of type } i \\ N \text{: total number of observations} \\ n \text{: number of cells in the table.} p_i \text{: expected proportions of the fraction of type} i \text{ in} \end{cases}$

**Bayesian** Does not exist, see *contingency table*

*Return to the table.*

### 6.2.3 Exact test of goodness-of-fit

Unlike the conventional statistical tests, there is no *test statistic*, we directly compute the *p-value* under the null hypothesis. The most common use are for dichotomous nominal variables or multinomial variables.

**Assumptions**

- Observations are independent.
- Small sample size $\lessapprox 1000$

**Frequentist** Let us define the list of, respectively, expected counts for each modality $i$, $(E_i)_{1 \leq i \leq m}$, and observed counts $(O_i)_{1 \leq i \leq m}$. Then $\begin{cases} \boldsymbol{H_0 : \forall i \in [\![1, m]\!],\ O_i = E_i} \\ H_1 : \exists i \in [\![1, m]\!],\ O_i \neq E_i \text{: two-tailed test} \end{cases}$

### 6.2.4 Fisher's exact test

To check the significance of the contingency between 2 kinds of classification of a given object, initially Fisher used this test to distinguish drink in which the tea has been put before the milk or vice-versa. For large sample use *G-test*

**Assumptions**

- In practice, small sample size $\lessapprox 1000$

**Frequentist** For example let's divide a population into male and female and for each persons indicating if this person is currently studying or not. We want to test if the proportion of studying students is higher among the women than among the men.

| | **Men** | **Women** | **Row Total** |
|---|---|---|---|
| **Studying** | $a$ | $b$ | $a+b$ |
| **Non-Studying** | $c$ | $d$ | $c+d$ |
| **Column Total** | $a+c$ | $b+d$ | $a+b+c+d=n$ |

The conditional on the margins of the table is distributed as $Hypergeometric(a+c, a+b, c+d)$ meaning $a+c$ draws from a population with $a+b$ success and $c+d$ failures. The probability of obtaining such set of values is given by

$$p = \frac{\binom{a+b}{a} \times \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b} \times \binom{c+d}{d}}{\binom{n}{b+d}}$$

**Bayesian** Does not exist, see *contingency table*

*Return to the table.*

### 6.2.5 G-test

It's a likelihood-ratio or a maximum likelihood statistical significance test. Either used to test *goodness-of-fit* and *independence* between 2 variables. It checks either if there is a significant difference between the expected and observed frequencies. This test tends to replace $\chi^2$*-test*

- *goodness-of-fit*: expected frequencies are computed with a theoretical relationship between observed frequencies
- *independence*: expected frequencies are computed with observed frequencies from the other sample
- *repeated tests*: first variable is analysed with a goodness-of-fit and the second one represents the repetition of the experiments multiple times. Thus it allows to assess the goodness-of-fit on a large sample instead of multiple lower samples. Expected frequencies is a theoretical relationship between the observed frequencies segmented in groups by the modalities of the second variable.

**Assumptions**

-
- Expected count must not be small in any modality.

**Strengths**

- Approximation to the theoretical $chi^2$ distribution is better attained with *G-test* than $\chi^2$ *test.*
- Cases where $O_i > 2 \times E_i$, *G-test* is always better than $\chi^2$ *test.*

**Weaknesses**

- in test of independence, for a small sample size use rather Fisher's extract test.

**Frequentist** We compare the observed counts in each modality with their expected counts. Let us define the list of, respectively, expected counts for each modality $i$, $(E_i)_{1\leq i\leq m}$, and observed counts $(O_i)_{1\leq i\leq m}$. Then $\begin{cases} \boldsymbol{H_0 : \forall i \in [\![1, m]\!],\ O_i = E_i} \\ H_1 : \exists i \in [\![1, m]\!],\ O_i \neq E_i \text{: two-tailed test} \end{cases}$

$$G = 2\sum_{i=1}^{m} O_i \times \ln\left(\frac{O_i}{E_i}\right)$$

$$\ln\left(\frac{L(\tilde{\theta}|x)}{L(\hat{\theta}|x)}\right) = \ln\left(\frac{\prod_{i=1}^{m}\tilde{\theta}^{x_i}}{\prod_{i=1}^{m}\hat{\theta}^{x_i}}\right) = \ln\left(\frac{\prod_{i=1}^{m}\left(\frac{x_i}{n}\right)^{x_i}}{\prod_{i=1}^{m}\left(\frac{e_i}{n}\right)^{x_i}}\right) = \ln\left(\prod_{i=1}^{m}\left(\frac{x_i}{e_i}\right)^{x_i}\right) = \sum_{i=1}^{m} x_i \ln\left(\frac{x_i}{e_i}\right)$$

Then we multiply by $-2$ to get *G-test* that is asymptotically equivalent to the *Pearson's* $\chi^2$ formula.

### 6.2.6 Cochran's Q test

It checks if $k$ treatments have identical effect, the response can take only 2 possible outcomes and a second variable segments the treatments.

| | **Treatment 1** | **Treatment 2** | $\cdots$ | **Treatment k** |
|---|---|---|---|---|
| *Block 1* | $x_{11}$ | $x_{12}$ | $\cdots$ | $x_{1k}$ |
| *Block 2* | $x_{21}$ | $x_{22}$ | $\cdots$ | $x_{2k}$ |
| *Block 3* | $x_{31}$ | $x_{32}$ | $\cdots$ | $x_{3k}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| *Block b* | $x_{b1}$ | $x_{b2}$ | $\cdots$ | $x_{bk}$ |

And $\forall (i, j) \in [\![1, b]\!] \times [\![1, k]\!], x_{ij} \in \{0, 1\}$

**Assumptions**

- The blocks are randomly selected from the population of all possible blocks.
- Outcome of the treatments are dichotomous, and should be coded in a standard way

**Frequentist** For example if $b$ respondents in a survey had each been asked $k$ *Yes/No* questions the *Q-test* could be use to test the null hypothesis that all questions were equally likely to elicit the answer "Yes".

We have $\begin{cases} H_0\text{: the treatments are equally effective} \\ H_a\text{: the treatments are } \textit{not} \text{ equally effective} \end{cases}$

$$T = k(k-1)\frac{\sum_{j=1}^{k}\left(x_{\cdot j} - \frac{N}{k}\right)^2}{\sum_{i=1}^{b} x_{i\cdot}\left(k - x_{i\cdot}\right)} \begin{cases} k\text{: number of treatments} \\ x_{\cdot j}\text{: column total for the } j\text{th treatment} \\ b\text{: number of blocks} \\ X_{i\cdot}\text{: row total for the } i\text{th block} \\ N\text{: grand total} \end{cases}$$

For significance level $\alpha$, the asymptotic critical region is $T > \chi^2_{1-\alpha,k-1}$ which is the $(a-\alpha)$ quantile of the $\chi^2$ distribution with $K-1$ degrees of freedom.

**Bayesian** Does not exist, see *contingency table*

### 6.2.7 Sign test

It is a statistical method to test for consistent differences between pairs of observations, such as the weight of subjects before and after treatment. For comparisons of paired observations $(x, y)$ the *sign-test* is most useful if comparison can only be expressed as $x > y,\ x = y,$ or $x < y$. If instead the differences can be expressed in numeric quantities it is worthy to use *t-test* or *Wilcoxon signed-rank test* will usually have greater power than the sign test to detect consistent differences.

**Frequentist** Let $p = \mathbb{P}(\{X > Y\})$, then $\begin{cases} H_0 : \ p = 0.5 \text{ meaning that given a random pair of measurements } (x_i, y_i) \text{ the} \\ H_a : \ p \neq 0.5 \end{cases}$

Pairs are omitted for which there is no differences so that there is a potential reduced sample of $m$ pairs. The statistics $W$ is defined as follow:

$$W = \mathbf{1}_{\{x_i > y_i\}} \hookrightarrow \mathcal{B}(m, 0.5)$$

**Assumptions** Let $\forall i \in [\![1, n]\!],\ Z_i = X_i - Y_i$

- $Z_i$ are assumed independent.
- Each $Z_i$ comes from the same continuous population.
- The values $X_i$ and $Y_i$ are ordered.

**Strengths**

- A fewer assumptions need to be made than for parametrical test

**Weaknesses**

- The power of test is lower than for a parametrical test

### 6.2.8 Contingency coefficients: Cramér's V

To quantify associations between 2 paired samples in a contingency table, it is based on $\chi^2$ and varies from 0 (no association) to 1 (complete association).

**Frequentist** Let a sample of size $n$ of the simultaneously distributed variable $A$ and $B$. $\forall (i, j) \in [\![1, r]\!] \times [\![1, c]\!],\ n_{ij} = Card(\{A_i, B_j\})$. Then $\chi^2 = \sum_{(i,j) \in [\![1,r]\!] \times [\![1,c]\!]} \frac{\left(n_{ij} - \frac{n_{i\cdot} \times n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} \times n_{\cdot j}}{n}}$

Finally the Cramér's V with bias correction is:

$$V = \sqrt{\frac{\max\left(0, \frac{\chi^2}{n} - \frac{(r-1)(c-1)}{n}\right)}{\min\left(r - \frac{(r-1)^2}{n-1} - 1, c - \frac{(c-1)^2}{n-1} - 1\right)}}$$

**Assumptions**

- The both variables have to be nominal.

**Strengths**

- Good analog of the $R^2$ for categorical variables.

**Weaknesses**

- Can tend to overestimate the strength of association.

### 6.2.9 Contingency table from a Bayesian perspective

To test the independence hypothesis between 2 variables.

**Bayesian** Let's consider 4 sampling plans, depending on which sampling plan is chosen the Bayes factor formula will change.

- *Poisson* sampling scheme: Each cell count is considered as random and so is the grand total, the cells are Poisson distributed. This design often occurs in purely observational work.
- *Joint multinomial* sampling scheme: same as above except that now, the grand total is fixed.
- *Independent multinomial* sampling scheme: either all row margins or all column margins are fixed, this scheme is frequently used in psychological studies.
- *Hypergeometric* sampling scheme: here both row margins and column margins are fixed. Practical use of this scheme is rare!

Bayes factors are often difficult to compute, as they are obtained by integrating out over the entire parameter space, a process that is non-trivial when the integrals are high-dimensional and intractable. Then we will use the 4 Bayes Factor developed by *Gunnel and Dickey in 1974*, because they only require computation of common functions such as gamma functions, for which numerical approximation are already available.
Here the logic: the Bayes Factor $BF_{01}^{i+1}$ computed at the observation $i+1$, contains the information up to the step $i$ with the extra information of the step $i+1$. We can then see $BF_{01}^{i+1}$ as the Bayes factor of the observation $i+1$ conditioned on the observation $i$.
Finally thanks to the successive conditionalization the Bayes Factor are easy to compute.

**Assumptions**

- We need to be consider data providing from one of the following sampling scheme: *Poisson*, *Joint multinomial*, *Independent multinomial* or *Hypergeometric*

**Strengths**

- Bayesian approach, then no issue to assess the significance
- Implemented in R

**Weaknesses**

- Restricted to the above sampling scheme today.

### 6.2.10 Wilconox test

Non parametric test, used to test the location of a population based on a data sample or to compare the locations of two populations using two matching samples.
It is a good alternative of the *t-test* when the mean is not of interest for the studied population.

**Frequentist** Let $Y$ and $X$ be 2 random variables, and $(x_i, y_i)_{1\leq i\leq n}$ a paired sample.

1. $\forall i \in [\![1, n]\!]\,,\ |x_i|$
2. Sort the $(|x_i|)_{1\leq i\leq n}$ and assign a rank $(R_i)_{1\leq i\leq n}$
3. The test statistic i $T = \sum_{i=1}^{n} sgn\,(X_i)\,R_i$
4. Produce a *p-value* by computing T to its distribution under the null hypothesis.

We will provide the logic for a one-sample test, the two-sample follows the same logic but with 2 variables.
Assume the data consists of independent and identically distributed (IID) samples from a distribution $F$ then consider 2 variables $(X_1, X_2) \hookrightarrow IID(F)$ Define $p_2 = \mathbb{P}\left(\left\{\frac{X_1+X_2}{2} > 0\right\}\right) = 1 - F^{(2)}(0)$ Then Wilcoxon signed-rank *sum* $\rightarrow H_0 :\ p_2 = \frac{1}{2}$ In restricting the distributions of interest we can reach more interpretable null and alternative hypotheses. On mildly restrictive is that $F^(2)$ has a unique median $\mu$. This median is called pseudo median of $F$ Then we have $H_0 :\ \mu = 0$

-

**Assumptions**

- Distribution $F$ is symmetric

**Strengths**

**Weaknesses** TO COMPLETE

### 6.2.11 Mann-Whitney test

Test for a randomly selected values $x$ and $y$ from 2 populations, $\mathbb{P}(\{x \leq y\}) = \mathbb{P}(\{x > y\})$

**Frequentist** Let $(n_1, n_2)\mathcal{N}_*^2$ and $(x_i)_{1 \leq i \leq n_1}$ and $(y_i)_{1 \leq i \leq n_2}$ both samples independent of each other.
Then $\begin{cases} U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \\ U_2 = n_1 n_2 + \frac{n_2(n_2+2)}{2} - R_2 \end{cases}$ $R_1, R_2$ being the sum of the ranks in groups 1 and 2.
Note that $AUC_1 = \dfrac{U_1}{n_1 n_2}$ meaning $U$-statistics is related to the area under the receiver operating characteristic.

**Assumptions**

- All observation from both groups are independent
- Values are at least ordinal
- $H_0$: the distribution of both population is identical
- $H_1$: the 2 distribution of population are different

**Strengths**

**Weaknesses**

### 6.2.12 Kruksal-Wallis test

Non-parametrical to test if samples originate from the same distribution.

**Frequentist** Let $N$ be the number of observations across all groups, $g$ number of groups, $n_i$ the number of observation in the group $i$, $r_{ij}$ the rank of observation $j$ from group $i$.
And $\begin{cases} \overline{r}_{i\cdot} = \dfrac{\sum_{j=1}^{n_i} r_{ij}}{n_i} \\ \overline{r} = \dfrac{N+1}{2} \end{cases}$

- Rank all data from all groups together
- $(N-1)\dfrac{\sum_{i=1}^{g} n_i \left(\overline{r}_{i\cdot} - \overline{r}\right)^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_j} n_i \left(r_{ij} - \overline{r}\right)^2}$ To actually check the stochastic differences.
- A correction can be brought for large number of ties.

**Assumptions**

- Independence
- All groups should have te same distributions

**Strengths**

- Non-parametrical test, no need of the normally distributed assumption.

**Weaknesses**

### 6.2.13 Friedman test

Non-parametric statistical test, analogu of the *repeated-measures ANOVA*. It use to detect differences in treatments across multiple test attempts.

**Frequentist**

- Consider a matrix of $n$ rows (the blocks) and $k$ columns (the treatments) and a single observation at the intersetion of each block and treatment. Then calculate the ranks.
- $\overline{r}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} r_{ij}$
- The test statistic is given by $Q = \frac{12n}{k(k+1)}\sum_{j=1}^{k}\left(\overline{r}_{\cdot j} - \frac{k+1}{2}\right)^2$ Note that the value of $Q$ does need to be adjusted for tied values in data.
- Finally when n or k is large ($n > 15 or k > 4$) the probability distribution of $Q$ can be approximated by a $\chi^2$ distribution.

**Assumptions**

- Independence

**Strengths**

**Weaknesses**

### 6.2.14 Sperman test

It assesses how well the relationship between 2 variables can be described using a monotonic function. Let $X, Y$ be 2 random variables, and $R$ the function transforming the realization of a random variable.

**Frequentist** $r_s = \rho_{R(X),R(Y)} = \frac{Cov\left(R(X), R(Y)\right)}{\sigma_{R(X)}\sigma_{R(Y)}} \begin{cases} \rho: \text{ Pearson correlation coefficient applied to the rank variables} \\ Cov\left(R(X), R(Y)\right) \end{cases}$

**Assumptions**

**Strengths**

**Weaknesses**

### 6.2.15 Pearson correlation coefficient

This coefficient is essentially a normalized measurement of the covariance such that the result has a value $-1$ and 1

**Frequentist** $\rho_{X,Y} = \frac{Cov\left(X, Y\right)}{\sigma_X \sigma_Y}$

**Assumptions**

**Strengths**

**Weaknesses**

- Not robust

### 6.2.16 Repeated-measures ANOVA

Used in repeated measure design, meaning when we measure multiple time the same variable taken on the same or matched subject either under different conditions or at different periods.

**Frequentist** Here $F = \frac{\frac{SS_{treatment}}{df_{treatment}}}{\frac{SS_{error}}{df_{error}}}$ In a *between-subjects* design there is a element of variance due to individual difference that is combined with the treatment and error term, meaning: $SS_{total} = SS_{treatment} + SS_{error}$ In a *repeated-measure* design it is possible to partition subject variability from the treatment and error term, meaning $SS_{total} = SS_{treatment(excludingindivdualdifferences)+SS_{subjects}+SS_{error}}$

**Assumptions**

- *Normality*: for each level of the within-subject factor, the dependent variable must have a normal distribution.
- *Sphericity*: difference scores computed between 2 levels of a within subject factor must have the same variance for the comparison of any 2 levels.
- *Randomness*: cases should be derived from a random sample.

**Strengths**

- ability to partition out variability due to individual differences.

**Weaknesses**

- Vulnerable to missing values, imputation, unequivalent time points between subjects and violation of Sphericity.

### 6.2.17 1-way ANOVA

Analysis of Variance describes the partition of the response variable sum of squares in a linear model into "explained" and "unexplained" components.

- Single categorical (or less common numerical) explanatory variable corresponds to One-Way ANOVA
- 2 factors to Two-Way ANOVA
- 3 factors to Three-Way ANOVA

The term "analysis of variance" is a bit of misnomer, we use variance-like quantities to study the equality or non-equality of population means, so we are analyzing means, not variances.

**Frequentist** examines equality of population means for a quantitative outcome and a single categorical explanatory variable with any number of levels.
The term "one-way" indicates that there is a single explanatory variable ("treatment") with 2 or more levels and only one level of treatment is applied at any time for a given subject.
And $H_0 : \forall (i,j) \in [\![1,k]\!]^2 \; \mu_i = \mu_j$

In ANOVA we work with variances and also "variance-like quantities" which are not really the variance of anything, but are still calculated as $\frac{SS}{df}$ all of these quantities are called "mean squares".

The deviation for subject $j$ of group $i$ in the figure above is mathematically equal to $Y_{ij} - \overline{Y}_i$ where $Y_{ij}$ is the observed value for subject $j$ of group $i$ and $\overline{Y}_i$ is the sample mean for group $i$.

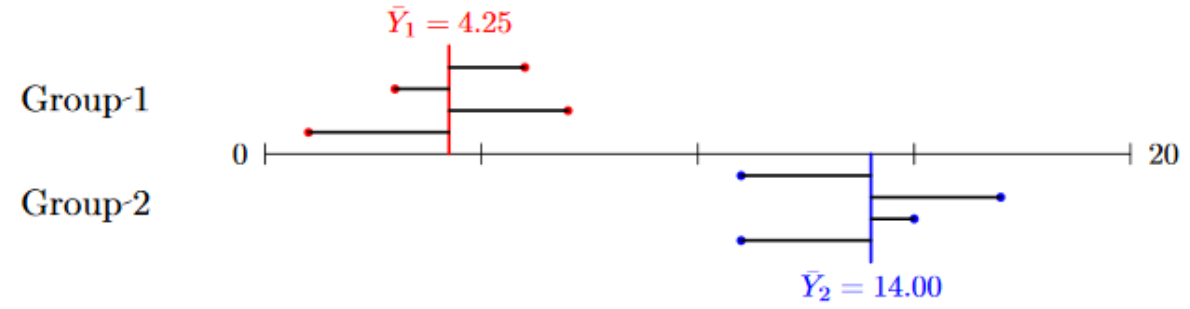


Figure 6.1: Deviations for within-group of squares

$$MS_{within} = \frac{SS_{within}}{df_{within}} \begin{cases} SS_{within} = \sum_{j=1}^{k} SS_j = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_{\bullet j}\right) \\ df_{within} = df_j = \sum_{j=1}^{k} (n_j - 1) = N - k \end{cases}$$

$MS_{within}$ is a good estimate of $\sigma^2$ from our model regardless of the truth of $H_0$. This is due to the way $SS_{within}$ is defined.

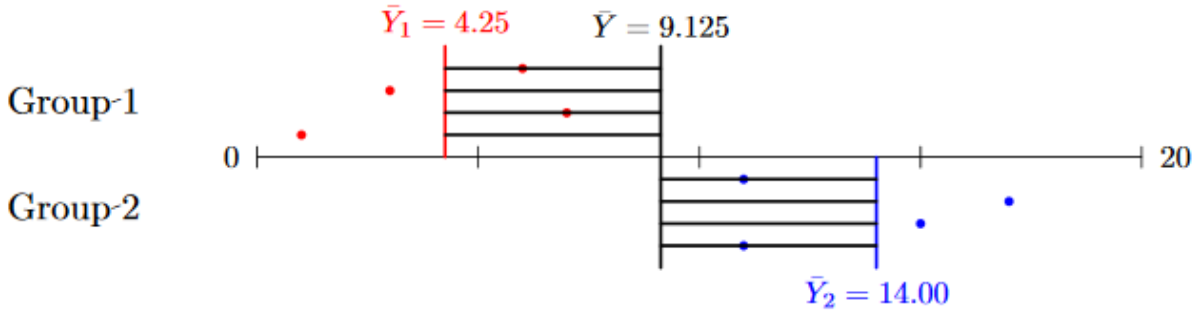


Figure 6.2: Deviations for betwwen-group sum of squares

$SS_{between}$ is the sum of the $N$ squared between-group deviations, where the deviation is the same for all subjects in the same group. The formula is :

$$MS_{Between} = \frac{SS_{Between}}{df_{between}} \begin{cases} SS_{between} = \sum_{j=1}^{k} n_j \left(\overline{Y}_{\bullet j} - \overline{Y}\right)^2 \\ df_{between} = k - 1 \end{cases}$$

Because of the way $SS_{between}$ is defined, $MS_{between}$ is a good estimate of $\sigma^2$ only if $H_0$ is true. Otherwise it tends to be larger.

The $F-statistic$ defined by $F = \dfrac{MS_{between}}{MS_{within}}$ tends to be larger if the alternative hypothesis is true than if the null hypothesis is true.

We can quantify "large" for the *F-statistic*, by comparing it to its null sampling distribution which is the specific *F-distribution* which has degrees of freedom matching the numerator and denominator of the *F-statistic*.

Concerning inferences to build the confidence interval we need the *standard error* (the standard deviation of the means) that is $\sqrt{\dfrac{MS_{within}}{n_i}}$

Numerically we have:
Given 2 samples with means $\mu_1$ and $\mu_2$, same variance $\sigma^2$ and $n = n_1 + n_2$ observations. Model:

$$\forall (j,i) \in [\![1,2]\!] \times [\![1,n_j]\!] \; y_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

$\alpha_j = \mu_j - \mu$ is called (treatment-) effect
Decomposition:

$$\begin{aligned} SS_{total} =& \sum_{j=1}^{n_1} (y_{1j} - \overline{y})^2 + \sum_{j=1}^{n_2} (y_{2j} - \overline{y})^2 \\ =& \sum_{j=1}^{n_1} (y_{1j} - \overline{y_1} + \overline{y_1} - \overline{y})^2 + \sum_{j=1}^{n_2} (y_{2j} - \overline{y_2} + \overline{y_2} - \overline{y})^2 \\ =& \underbrace{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}_{SS_{within}} + \underbrace{n_1(\overline{y}_1 - \overline{y}) + n_2(\overline{y}_2) - \overline{y}^2}_{SS_{between}} \end{aligned}$$

$SS_{between}$ corresponds to squared enumerator $(\overline{y}_1 - \overline{y}_2)^2$ of the statistic:

$$\begin{aligned} SS_{between} =& n_1(\overline{y}_1 - \overline{y})^2 + n_2(\overline{y}_2 - \overline{y})^2 \\ =& n_1 \left( \overline{y}_1 - \frac{n_1\overline{y}_1 + n_2\overline{y}_2}{n_1 + n_2} \right)^2 + n_2 \left( \overline{y}_2 - \frac{n_1\overline{y}_1 + n_2\overline{y}_2}{n_1 + n_2} \right)^2 \\ =& \frac{n_1 n_2}{n_1 + n_2} (\overline{y}_1 - \overline{y}_2)^2 \end{aligned}$$

$SS_{within}$ corresponds to denominator of *t-statistic*: $s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$
Pooled variance that is an estimate of the fixed common variance $\sigma^2$ underlying various populations that have different means. $\hat{\sigma} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$ Null hypothesis $H_0 : \mu_1 = \mu_2$ or $\alpha_1 = \alpha_2 = 0$ *F-test*
$(\overline{Y}_1 - \overline{Y}_2) \hookrightarrow \mathcal{N}\left(\mu_1 - \mu_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2\right)$
$\mathbb{E}\left(\left[\overline{Y}_1 - \overline{Y}_2\right]^2\right) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2 + (\mu_1 - \mu_2)^2$
$\mathbb{E}(MS_{between}) = \mathbb{E}\left(\frac{n_1 n_2}{n_1 + n_2}\left[\overline{Y}_1 - \overline{Y}_2\right]^2\right) = \sigma^2 + \frac{n_1 n_2}{n_1 + n_2}(\mu_1 - \mu_2)^2$
$\mathbb{E}(MS_{whithin}) = \sigma^2$
$F = \frac{MS_{between}}{MS_{within}}$ Here $F = t^2$

$$\begin{aligned} \text{Degrees of freedom} =& n - 1 \\ =& \underbrace{(n - m)}_{df_{within}} + \underbrace{(m - 1)}_{df_{between}} \end{aligned}$$

- $SS_{within}$ and $SS_{between}$ are independent
- under $H_0$ $\mathbb{E}(MS_{between}) = \mathbb{E}(MS_{within}) = \sigma^2$
- under $H_a$ $\mathbb{E}(MS_{between}) > \sigma^2$ and $\mathbb{E}(MS_{within}) = \sigma^2$

Hence

$$F = \frac{MS_{between}}{MS_{within}} \hookrightarrow F_{m-1,n-m}$$

In the case of 2 groups ("*t-test*") we received:

$$\overline{y}_1 - \overline{y}_2 \pm t_{n-2,1-\frac{\alpha}{2}} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

**Assumptions**

The statistical model for which one-way ANOVA is appropriate is that the

- (Quantitative) Outcomes for each group are normally distributed
- Outcome variances are all equal to ($\sigma^2$)
- The errors are assumed to be independent.

**Strengths**

**Weaknesses**

### 6.2.18 T-test

It is commonly used when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known.
When the scaling term is estimated from the data, under certain conditions, the test statistic follows a *Student's t-test*.
Most test statistics have the form $t = \frac{Z}{s}$, $Z$ may be sensitive to the alternative hypothesis, whereas $s$ isa scaling parameter allowing to determine the distribution $t$.

### 6.2.19 One-sample

**Frequentist**

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}, \begin{cases} \overline{x}\text{: sample mean} \\ s\text{: sample standard deviation} \\ n\text{: sample size} \end{cases}$$

By the central limit theorem, if the observations are independent and the second moment exist, then $t$ will approximately follow the distribution $\mathcal{N}(0, 1)$

**Assumptions** Although the parent population does not need to be normally distributed, the distribution of the population sample means $(\overline{x}_s)_{1 \leq s \leq S}$.

**Strengths**

**Weaknesses**

### 6.2.20 Slope of a regression line

Suppose one is fitting: $Y = \alpha + \beta x + \epsilon$, where $x$ is known and $\alpha$ and $\beta$ are unknown and finally $\epsilon \hookrightarrow \mathcal{N}(\prime, \infty)$. Symbols with hat will refer to estimators.

**Frequentist**

$$t_{score} = \frac{\hat{\beta} - \beta_0}{SE_{\hat{\beta}}} \hookrightarrow \mathcal{N}(0, 1)$$

**Assumptions**

**Strengths**

**Weaknesses**

### 6.2.21 Paired / Unpaired t-test

Depending on if the samples are paired or not and the differences between means and variances, the formula will be different. It is not worthful to dive in these formulas.

**Frequentist**

**Assumptions**

**Strengths**

**Weaknesses** Weimprove

# Chapter 7

# Data revovery

## 7.1 Sampling methods

### 7.1.1 Monte Carlo

## 7.2 Information theory

### 7.2.1 KL divergence

## 7.3 Key Mathematical functions

### 7.3.1 Softmax function

**Purpose** The softmax function takes as input a vector $z$ of $K$ real numbers, and normalizes it into a probability distribution consisting of $K$ probabilities proportional to the exponentials of the input numbers.

$$\sigma(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

**Interpretation** it is rather a smooth approximation of the *argmax* function, meaning the function returning the index of the maximum value of a given vector.

# Part III

# Classical Learning

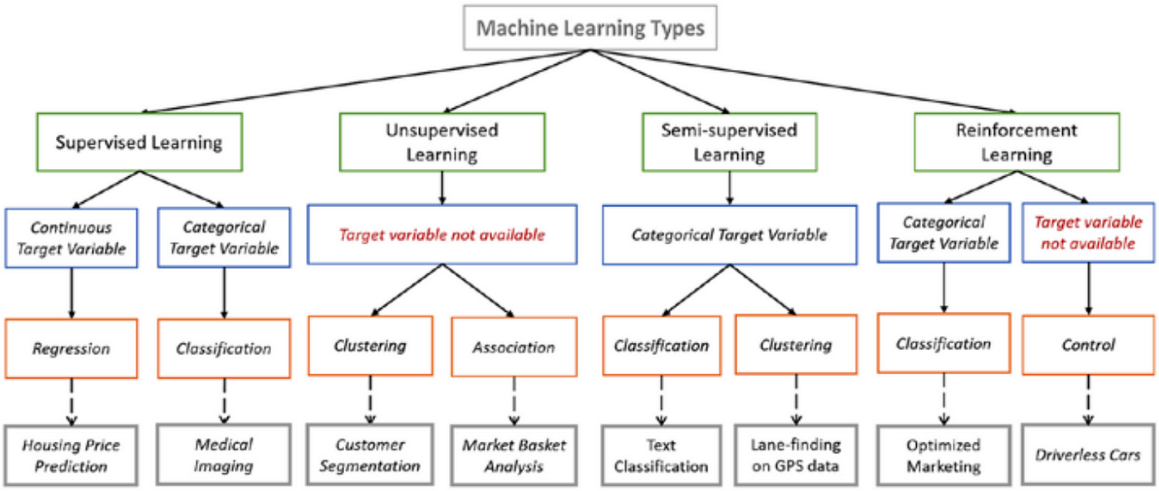


Figure 7.1: Types of machine learning methods

# Chapter 8

# Supervised Learning

## 8.1 Classification

### 8.1.1 Naive Bayes classifiers

**Purpose** Classifying vectors of discrete-valuated features $x \in \{i\}_{1\leq i\leq K}^{D}$, where $K$ is the number of values for each feature, and $D$ the number of features.

#### Assumptions

- Features are conditionally independent given the class label

**Theory** As a *generative* model, meaning of the form: $\mathbb{P}(y=c|\boldsymbol{x},\boldsymbol{\theta}) \propto \mathbb{P}(\boldsymbol{x}|y=c,\boldsymbol{\theta})\,\mathbb{P}(y=c|\boldsymbol{\theta})$. The key of such models is the possibility to specify a suitable form for the class-conditional density $\mathbb{P}(\boldsymbol{x}|y=c,\boldsymbol{\theta})$ which defineees what kind of data we expect to see in each class. And with the independence assumption we have:

$$\mathbb{P}(\boldsymbol{x}|y=c,\boldsymbol{\theta}) = \prod_{j=1}^{D}\mathbb{P}(x_j|y=c,\boldsymbol{\theta}_{jc})$$

with all $\mathbb{P}(x_j|y=c,\boldsymbol{\theta}_{jc})$ being able to follow a *normal*, *bernoulli* or *multinoulli* distribution.
Training a NBC consists in computing the MLE or the MAP estimate for the parameters.
For a single observation $\mathbb{P}(x_i,y_i|\boldsymbol{\theta}) = \mathbb{P}(y_i|\boldsymbol{\pi})\prod_j \mathbb{P}(x_{ij}|\boldsymbol{\theta}_j) = \prod_c \pi_c^{\mathbb{1}(y_i=c)} \prod_j\prod_c \mathbb{P}(x_{ij}|\theta_{jc})^{\mathbb{1}(y_i=c)}$

Hence the *log-likelihooh*: $\log(\mathcal{D}|\theta) = \sum_{c=1}^{C} N_c \log(\pi_c) + \sum_{j=1}^{D}\sum_{c=1}^{C}\sum_{i:y_i=c} \log\left(\mathbb{P}(x_{ij}|\boldsymbol{\theta}_{jc})\right)$
By optimizing the above equation we are able to find the $(\theta_{jc})_{1\leq j\leq D,\ 1\leq c\leq C}$ and we can then use them to predict the output of an observation $\boldsymbol{x}$ as: $\mathbb{P}(y=c|\boldsymbol{x},\mathcal{D}) \propto \mathbb{P}(y=c|\mathcal{D})\prod_{j=1}^{D}\mathbb{P}(x_j|y=c,\mathcal{D})$

#### Strengths

- Simple model, for $C$ classes and $D$ features, and hence relatively immune to overfitting

#### Weaknesses

- Unaccuracy because of the strong independence assumption

**Relationships with other methods** Logistic Regression: for discrete inputs *Naive Bayesian Classifiers* form a generative-discriminant pair with *Multinomial Logistic Regression*: each NBC can be considered a way of fitting a probability model that optimizes the joint likelihood $\mathbb{P}(C,\boldsymbol{x})$, while Multinomial Logistic Regression fits the same probability to optimize the conditional $\mathbb{P}(C|\boldsymbol{x})$

**Examples of application**

- Classifying documents using bag of words
- Determining the gender of a person, based on measured features

### 8.1.2 Linear/Quadratic Discriminant Analysis

It consists in defining the class conditional densities in a generative classifier: $\mathbb{P}(\boldsymbol{x}|y=c,\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x},\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$
As for a generative classifier we have the following equation:

$$\mathbb{P}(y=c|\boldsymbol{x},\boldsymbol{\theta}) = \frac{\overbrace{\mathbb{P}(\boldsymbol{x}|y=c,\boldsymbol{\theta})}^{\textit{class-conditional density}} \overbrace{\mathbb{P}(y=c|\boldsymbol{\theta})}^{\textit{class prior}}}{\sum_{c'} \mathbb{P}(y=c'|\boldsymbol{\theta})\,\mathbb{P}(\boldsymbol{x}|y=c',\boldsymbol{\theta})}$$

**Purpose of Quadratic Discriminant Analysis**

$$\mathbb{P}(y=c|\boldsymbol{x},\boldsymbol{\theta}) = \frac{|2\pi\Sigma_c|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}[\boldsymbol{x}-\boldsymbol{\mu}_c]^T\Sigma_c^{-1}[\boldsymbol{x}-\boldsymbol{\mu}_c]\right)\pi_c}{\sum_{c'}|2\pi\Sigma_{c'}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}[\boldsymbol{x}-\boldsymbol{\mu}_{c'}]^T\Sigma_{c'}^{-1}[\boldsymbol{x}-\boldsymbol{\mu}_{c'}]\right)\pi_{c'}}$$

The threshold of this results will be a quadratic function of $\boldsymbol{x}$.

**Purpose of Linear Discriminant Analysis** Same equation than above but this time, $\forall c \in [\![1,C]\!]\ \Sigma_c = \Sigma$, then quadratic term $\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}$ will cancel out from numerator and denominator. Then by considering the above cancellation and the fact that evidence is considered as a constant, we have:

$$\begin{aligned}\mathbb{P}(y=c|\boldsymbol{x},\boldsymbol{\theta}) &\propto \exp\left(\log(\pi_c) + \boldsymbol{\mu}_c^T\Sigma^{-1}\boldsymbol{x}\boldsymbol{\mu}_c\right)\\ &= \exp\left(\boldsymbol{\beta}_c^T\boldsymbol{x}+\gamma_c\right)\end{aligned}$$

Note also that we have exactly: $\mathbb{P}(y=c|\boldsymbol{x},\boldsymbol{\theta}) = \dfrac{e^{\boldsymbol{\beta}_c^T\boldsymbol{x}+\gamma_c}}{\sum_{c'} e^{\boldsymbol{\beta}_{c'}^T\boldsymbol{x}+\gamma_{c'}}} = S(\boldsymbol{\eta})_c$. With $\eta = (\boldsymbol{\beta}_c\boldsymbol{x}+\gamma_c)_{1\leq c\leq C}$ We recognize the *softmax* function.

**Assumptions**

- Independent variables are normal for each level of the grouping variable.
- Homoscedasticity for LDA: variances among group variables are the same across levels of predictors.
- Independence of the observations.

**Theory**

**Strengths**

**Weaknesses**

- Multicollinearity: predictive power can decrease with an increased correlation between predictor variables.

**Relationships with other methods**

**Examples of application**

### 8.1.3 Logistic Regression

**Purpose** With the generative approach we create a joint model of the form $\mathbb{P}(y, \boldsymbol{x})$, and then to condition on $\boldsymbol{x}$, thereby deriving $\mathbb{P}(y|bmx)$, it is the *generative* approach.
Alternatively, fitting directly a model of the form $\mathbb{P}(y|\boldsymbol{x})$ is a *discriminative* approach.

**Assumptions**

- Independence

**Theory** The data distribution is modelled by : $\boxed{\mathbb{P}(y|\boldsymbol{x}) = Bernoulli\left(y|\sigma\left(\boldsymbol{w}^T\boldsymbol{x}\right)\right)}$

With $\sigma$ being the *sigmoid* function, such that $\sigma = \begin{cases} \mathbb{R} \longrightarrow [0,1] \\ x \mapsto \dfrac{e^x}{1+e^x} \end{cases}$

**Maximum Likelihood Estimator**

$$\begin{aligned} NLL(\boldsymbol{w}) &= -\sum_{i=1}^{N} \log\left(\hat{y}_i^{\mathbb{1}_{\{y_i=1\}}}\left[1-\hat{y}_i^{\mathbb{1}_{\{y_i=0\}}}\right]\right) \\ &= -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + [1-y_i]\log(1-\hat{y}_i) \end{aligned}$$

This called *cross-entropy*

**Strengths**

**Weaknesses**

**Relationships with other methods**

**Examples of application**

### 8.1.4 Logistic Regression

**Purpose**

**Assumptions**

**Theory**

**Strengths**

**Weaknesses**

**Relationships with other methods**

**Examples of application**

### 8.1.5 Logistic Regression

**Purpose**

**Assumptions**

**Theory**

**Strengths**

**Weaknesses**

**Relationships with other methods**

**Examples of application**

### 8.1.6 Logistic Regression

**Purpose**

**Assumptions**

**Theory**

**Strengths**

**Weaknesses**

**Relationships with other methods**

**Examples of application**

## 8.2 Regression

### 8.2.1 Linear Regression

**Purpose**

**Assumptions**

**Theory**

**General** It is a model for which the data distribution (likelihood) is described by:

$$\mathbb{P}\left(y|\boldsymbol{x},\boldsymbol{\theta}\right)=\mathcal{N}\left(y|\boldsymbol{w}^T\phi(\boldsymbol{x}),\sigma^2\right)$$

with $\phi$ that can be a non-linear function, in this case we talk about *basis function expansion*.
To estimate the parameters we can use the *maximum likelihood estimation*: $\hat{\theta}\triangleq\operatorname{argmax}_\theta\log\left(\mathbb{P}\left(\mathcal{D}|\boldsymbol{\theta}\right)\right)$.
For computational purpose it is better to consider the minimization of the *Negative Log Likelihood* (NLL):

$$\begin{aligned}
NLL(\theta)&\triangleq-\log\left(p\left(\mathcal{D}|\theta\right)\right)\\
&=-\sum_{i=1}^{n}\log\left(\mathbb{P}\left(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}\right)\right)\\
&=-\sum_{i=1}^{n}\log\left(\left[\frac{1}{2\pi\sigma^2}\right]^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\left[y_i-\boldsymbol{w}^T\boldsymbol{x}_i\right]^2\right)\right)\\
&=\frac{n}{2}\log\left(2\pi\sigma^2\right)+\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\boldsymbol{w}^T\boldsymbol{x}_i\right)^2\\
&=\frac{n}{2}\log\left(2\pi\sigma^2\right)+\frac{1}{2\sigma^2}\;RSS(\boldsymbol{w})\\
&=\frac{n}{2}\log\left(2\pi\sigma^2\right)+\frac{1}{2\sigma^2}\left\|\epsilon\right\|_2^2
\end{aligned}$$

As the *MLE* for $\boldsymbol{w}$ is the one minimizing the *RSS* then this method is known as *least square*.

**Derivation of the MLE** it is better to use a matrix-vector representation. $NLL(\boldsymbol{w}) = \frac{1}{2}(y - \boldsymbol{X}\boldsymbol{w})^T (y - \boldsymbol{X}\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T \left(\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{w} - \boldsymbol{w}^T \left(\boldsymbol{X}^T y\right)$ Note that $\boldsymbol{X}^T\boldsymbol{X}$ is the *sum of squares matrix*. Then **gradient**, $g(\boldsymbol{w}) = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y}$ that we have to equate to zero to get $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$ to conclude that:

$$\boxed{\hat{\boldsymbol{w}}_{OLS} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}}$$

**Robust Linear Regression** It is very common to model the noise in regression models using a *Gaussian distribution*, meaning $\epsilon_i = y_i - \boldsymbol{w}^T\boldsymbol{x}_i \hookrightarrow \mathcal{N}, \sigma^{\in}$. One way to achieve *robustness* against *outliers* is to replace the Gaussian distribution for the response variable with a distribution having **heavy tails**.

| Likelihood | Prior | Name |
|---|---|---|
| Gaussian | Uniform | *Least Squares* |
| Gaussian | Gaussian | *Ridge* |
| Gaussian | Laplace | *Lasso* |

**Ridge** encourages parameters to be small by using a zero-mean Gaussian prior: $\mathbb{P}(\boldsymbol{w}) = \prod_{j=1}^{D} \mathcal{N}\left(\omega_j | 0, \tau^2\right)$, where $\frac{1}{\tau^2}$ controls the strength of the prior.

The corresponding *MAP* estimation problem becomes: $\text{argmax}_{\boldsymbol{w}} \sum_{i=1}^{n} \log\left(\mathcal{N}\left(y_i | \omega_0 + \boldsymbol{w}^T\boldsymbol{x}_i, \sigma^2\right)\right) + \sum_{j=1}^{D} \log\left(\mathcal{N}\left(\omega_j | 0, \tau^2\right)\right)$.

After some calculus and with where $\lambda \triangleq \frac{\sigma^2}{\tau^2}$ we deduce that:

$$\boxed{\hat{\boldsymbol{w}}_{Ridge} = \left(\lambda\boldsymbol{I}_D + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}\boldsymbol{y}}$$

Advantages of Ridge regression on OLS regression:

- $\left(\lambda\boldsymbol{I}_D + \boldsymbol{X}^T\boldsymbol{X}\right)$ is much better conditioned, and hence more likely to be invertible, than $\boldsymbol{X}^T\boldsymbol{X}$ at least for suitable large $\lambda$
- if we follow a *Singular Value Decomposition* $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ we find that $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}}_{Ridge} = \sum_{j=1}^{D} \boldsymbol{u}_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \boldsymbol{u}_j^T\boldsymbol{y}$ with $(\sigma_j)_{1 \leq j \leq D}$ the singular value of $\boldsymbol{X}$ whereas for OLS we have $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}}_{OLS} = \sum_{j=1}^{D} \boldsymbol{u}_j\boldsymbol{u}_j^T\boldsymbol{y}$. Meaning that with Ridge if $\sigma_j^2$ is small compared to $\lambda$ then direction $\boldsymbol{u}_j$ will not have much effect on the prediction. In term of predictive accuracy *Ridge* regression is more interesting than *PCA* regression.

**Strengths**

- Simple
- Customizable to achieve robustness

**Weaknesses**

- Not very powerful for non-linear data

**Relationships with other methods**

- Ridge Regression has similitude with PCA

**Examples of application**

# Part IV

# Deep Learning

# Part V

# Use-cases

# Bibliography

[1] Omar Elgabry. *The Ultimate Guide to Data Cleaning*. 2019. URL: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4.

[2] Wikipedia contributors. *Moments (Mathematics)*. [Online; accessed 21-August-2023]. 2023. URL: https://en.wikipedia.org/wiki/Moment_(mathematics).

[3] Wikipedia contributors. *Probability*. [Online; accessed 20-August-2023]. 2023. URL: https://en.wikipedia.org/wiki/Probability.