UNIT 1: Advanced Statistical Inference

Classes: Sept 9-Oct 21. **Exam:** Nov 7

Syllabus

1. Preliminaries and Inferential questions
2. The frequentist approach: point estimation, some finite sample properties
3. The frequentist approach: Hypothesis testing and permutational approach
4. Interval estimation
5. Several Appendices

Unit 1: Main and supplementary references

1. Wood (2015). *Core Statistics*. Chapter 2.
2. Olive (2014). *Statistical Theory and Inference*. Sections 5.1, 6.1, 6.2, 7.1, 7.2, 7.3
3. Held and Sabanés (2014). Applied Statistical Inference. Section 2.5. Chapter 3
4. Cox (2006). *Principles of Statistical Inference*. Cambridge, Chapters 1, 2, 3, Section 8.3
5. Casella-Berger (2002). *Statistical Inference*. Sections 6.1, 6.2, 7.2.2, 7.3.1, 7.3.2, 7.3.3, 8.3.1, 8.3.2, 9.2.1, 9.2.2, 9.3.1
6. Gómez y Delicado (2001). *Inferencia y Decisión*. Sections 3,2, 3.3, 5.1, 6.1 (not necessarily in full or in this order)

Outline

Inferential Aims

READ: Chapter 1 in Cox. Principles of Statistical Inference. Cambridge, 2006

Statistical inference is the science of analyzing and interpreting data. It provides essential tools for processing information, summarizing the amount of knowledge gained and quantifying the remaining uncertainty. (Held and Sabanés)

- Statistics aims to **extract information from data**
- Two main difficulties:
  - Not easy to **INFER** what we want
  - Random variability in data

Inferential Aims: Some hints

- Start with a subject-matter question
- Data **HAS TO BE AVAILABLE** to address the question
- Check data quality, do simple tabulations and graphs
- Specify a statistical model to describe how “our” data might be generated
- Use the statistical model to **go in the reverse direction**: infer the values of the model **consistent with the observed data**

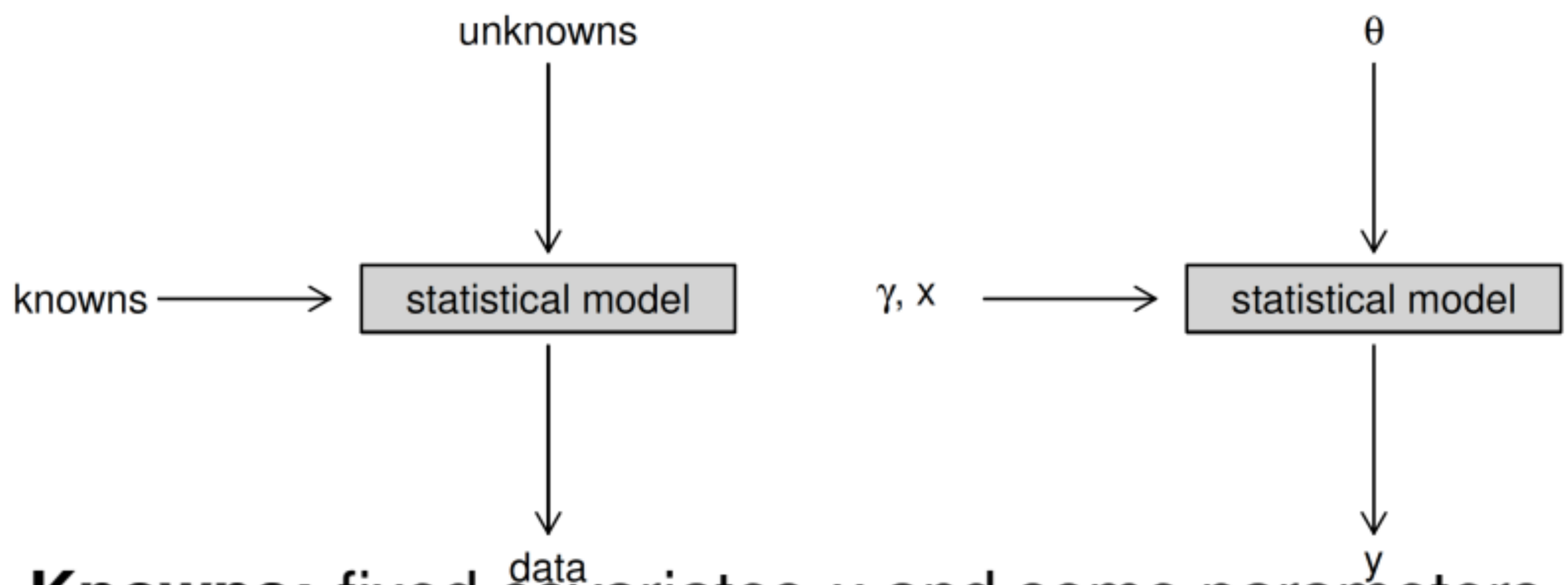
Formulation of the statistical model

The choice of the model is crucial to fruitful applications

Small notation:

- y: random vector containing the **observed data**. Y: corresponding random variable vector
- θ: vector of **parameters of unknown value.**
- **Density function** $f_Y(y;\theta)$ describes the family of models.
  It could also be used
  - cumulative distribution
  - survival function
- **Statistical model:** a "recipe" by which y might have been generated from Y given appropriate values of θ

Statistical model



- **Knowns:** fixed covariates *x* and some parameters γ
- **Unknowns:** Parameter vector θ and nuisance parameters λ needed to specify the model.
- **Data:** observations *y* from *Y*

Statistical methods aim to reverse the direction of the vertical arrows: infer unknown θ from observed data *y*.

Statistical models must allow random data to be simulated in a way stochastically similar to *y*

Parametric statistical models

General notation: $\psi, \theta, \lambda, \Omega_\psi$, for instance:

$$\psi = (\theta, \lambda)$$
$$\Omega_\psi \text{ full parameter space}$$
$$d = dim(\Omega_\psi)$$

- **Fully Parametric models**: $\psi = (\theta, \lambda) \in \Omega_\psi \in \mathcal{R}^d$. For instance, $Y \sim N(\mu, \sigma^2)$, $\psi = (\mu, \sigma^2) \in \mathcal{R}^2$.

  Parameter values which make the data appear relatively probable according to the model are more likely to be correct than parameter values which make the data appear relatively improbable according to the model. (Wood).

  Often use Maximum Likelihood Estimation (MLE) to estimate ψ

Non-parametric and semi-parametric statistical models

General notation: $\psi, \theta, \lambda, \Omega_\psi$, for instance: $\psi = (\theta, \lambda)$

Ω_ψ full parameter space

$d = dim(\Omega_\psi)$

- **Non-parametric models**: Dimension of Ω_ψ unspecified (hypothetically $= \infty$).
  Not restricted to a family of distributions (not specified in terms of ψ)) and solely based on the data.
  For instance, a histogram is an example of a nonparametric estimate of a probability distribution
  Convenient for Goodness-of-ft checking
- **Semi-parametric problems**: $\psi = (\theta, \lambda) \in \mathcal{R}^m x \mathcal{R}^\infty$ (λ is unspecified). Combines parametric and nonparametric parts.
  For instance, Cox proportional hazards model in survival

Statistical Inference purpose: Answering questions (I)

Theory of inference takes the family of statistical models as given (with θ parameter of interest and λ nuisance parameter) and answers questions about the model in the light of the data y.

Basic questions:

1. What values of θ are most consistent with y? **Answers via Point Estimation:** Providing reasonable parameter value guesses (numerical approximations) for θ from the observed data y
2. What ranges of θ are consistent with y? **Answers via Interval Estimation:** Providing intervals or set of values within which θ is likely to lie.

Statistical Inference purpose is answering questions (II)

Data y
Statistical model $\psi = (\theta, \lambda)$: θ parameter of interest, λ nuisance parameter

3. Is some defined restriction on θ consistent with y?
   Answers via Hypothesis Testing: Assessing the consistency of the data with a particular parameter value θ_0.
4. Is the model consistent with the data for any values of θ at all?
   Answers via Model checking: Showing that the model is wrong in some serious and detectable way.

Statistical Inference purpose is answering questions (III)

5. What if the target of the study is the value of an unobserved random variable?
   Answers via Prediction: Predicting unobserved random variables from the same random system that generated the data using the estimated statistical model.
   SUPP READING: *To Explain or to Predict?* **by G. Shmueli. Statistical Science (2010) Vol 25, 289–310**

6. What if you want to reach a decision?
   Answers via Decision Analysis: Data are analysed with a view to reaching a decision. A set of decisions (rules) has to be established as well as the consequences of each one

Example 1.1: DMBA Data Case Study

$Y =$ times (in days) from injection with carcinogenic DMBA* to death due to vaginal cancer in two groups of rats depending on the diet before the injection. [*DMBA=(7,12- dimethylbenz[a]anthracene)]

Group 1	143	164	188	188	190	192	206
	209	213	216	220	227	230	234
	246	265	304	216+	244+		
Group 2	142	156	163	198	205	232	232
	233	233	233	233	239	240	261
	280	280	296	296	323	204+	344+

Table: Times in days to death from vaginal cancer. Times with the symbol $+$ indicate rats that have been moved away from the study (Source: Kalbfleisch and Prentice, 2002)

Question of interest: Does the time until death from vaginal cancer differs by diet?

Example 1.1. Goodness of fit for both groups

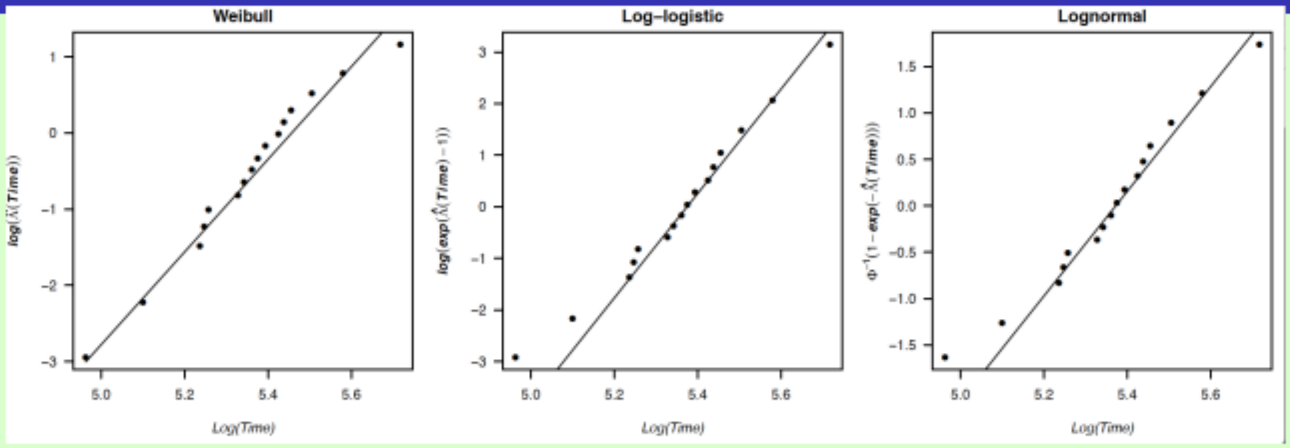


Figure: Group 1: GOF Weibull, Log-logistic, Lognormal

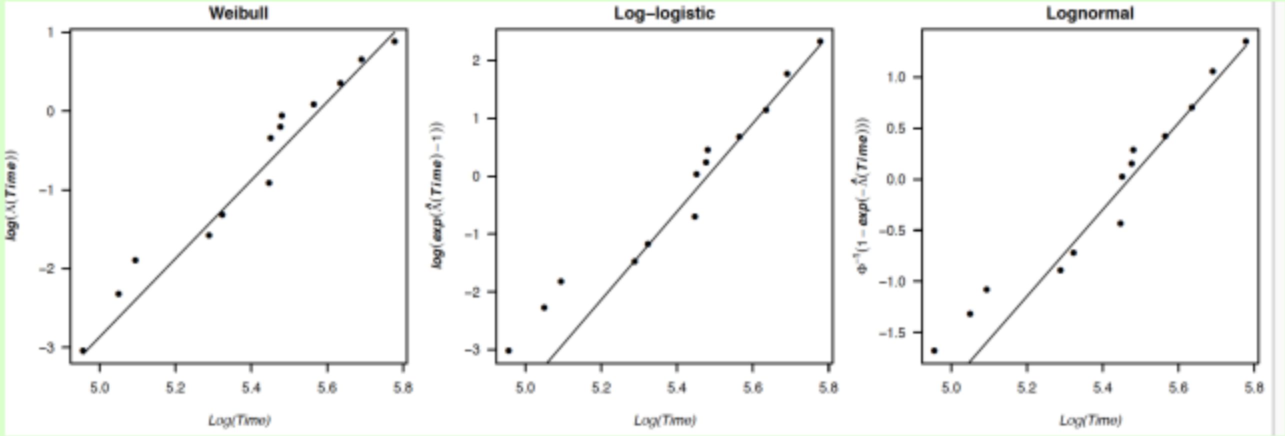


Figure: Group 2: GOF Weibull, Log-logistic, Lognormal

No fit is very convincing for both groups. Would a shifted distribution provide a better fit?

GOF for both groups shifted 100 days

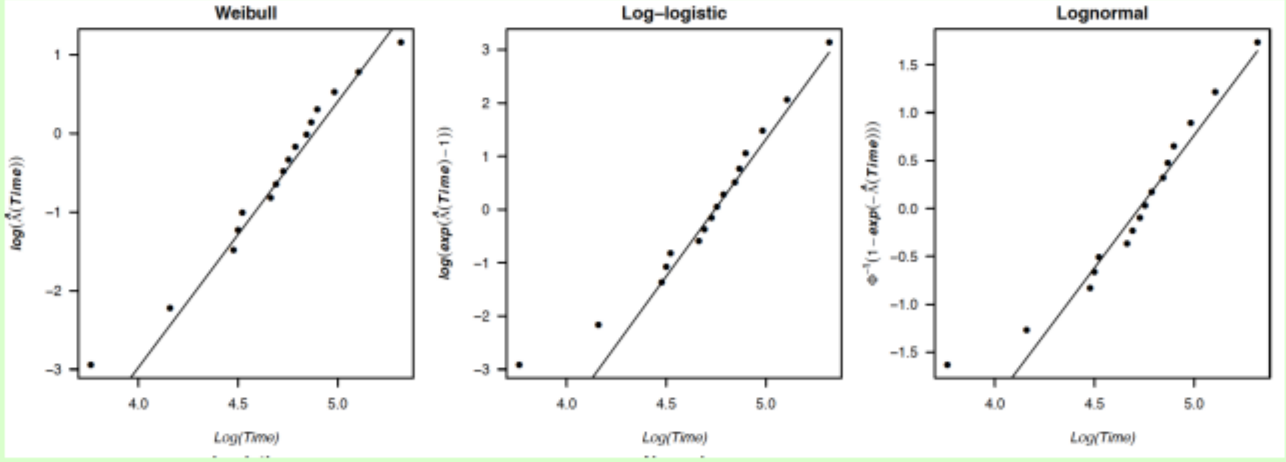


Figure: Group 1: GOF Weibull, Log-logistic, Lognormal

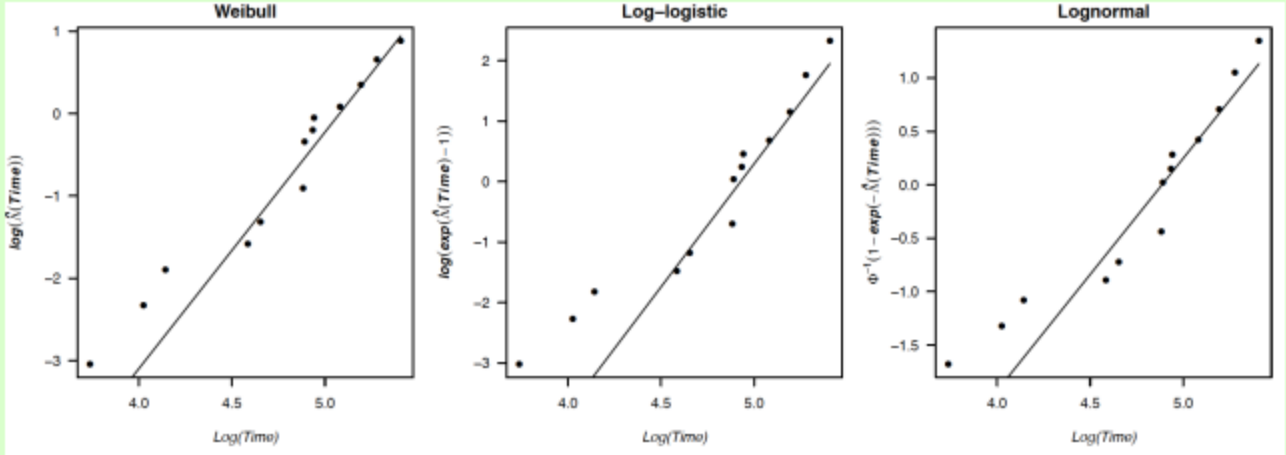


Figure: Group 2: GOF Weibull, Log-logistic, Lognormal

Shifted Weibull distribution seems to provide a better fit and Weibull law is a good interpretative option

GofCens R library

Graphical tools and goodness-of-fit tests for complete data and right-censored data can be found in the GofCens R library `https://cran.r-project.org/web/packages/GofCens/index.html`. It covers

1. Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling tests based on the empirical distribution function for complete data and their extensions for right-censored data.
2. Generalized χ^2 (Chi-squared)-type tests based on the squared difference between observed and expected counts using random cells with right-censored data.
3. A series of graphical tools such as probability or cumulative hazard plots to guide the decision about the parametric model that best fits the data.

Example 1.1: Weibull model

Advanced Statistical Inference

Guadalupe Gómez (UPC) & Alex Sánchez (UB)

Preliminaries

The Frequentist approach

Reasonable model for Y: Weibull's distribution

$$f_\theta(y) = \begin{cases} \alpha\beta y^{\beta-1} e^{-\alpha y^\beta} & \text{si } y \geq 0 \\ 0 & \text{si } y < 0 \end{cases}$$

Associated family of distributions:

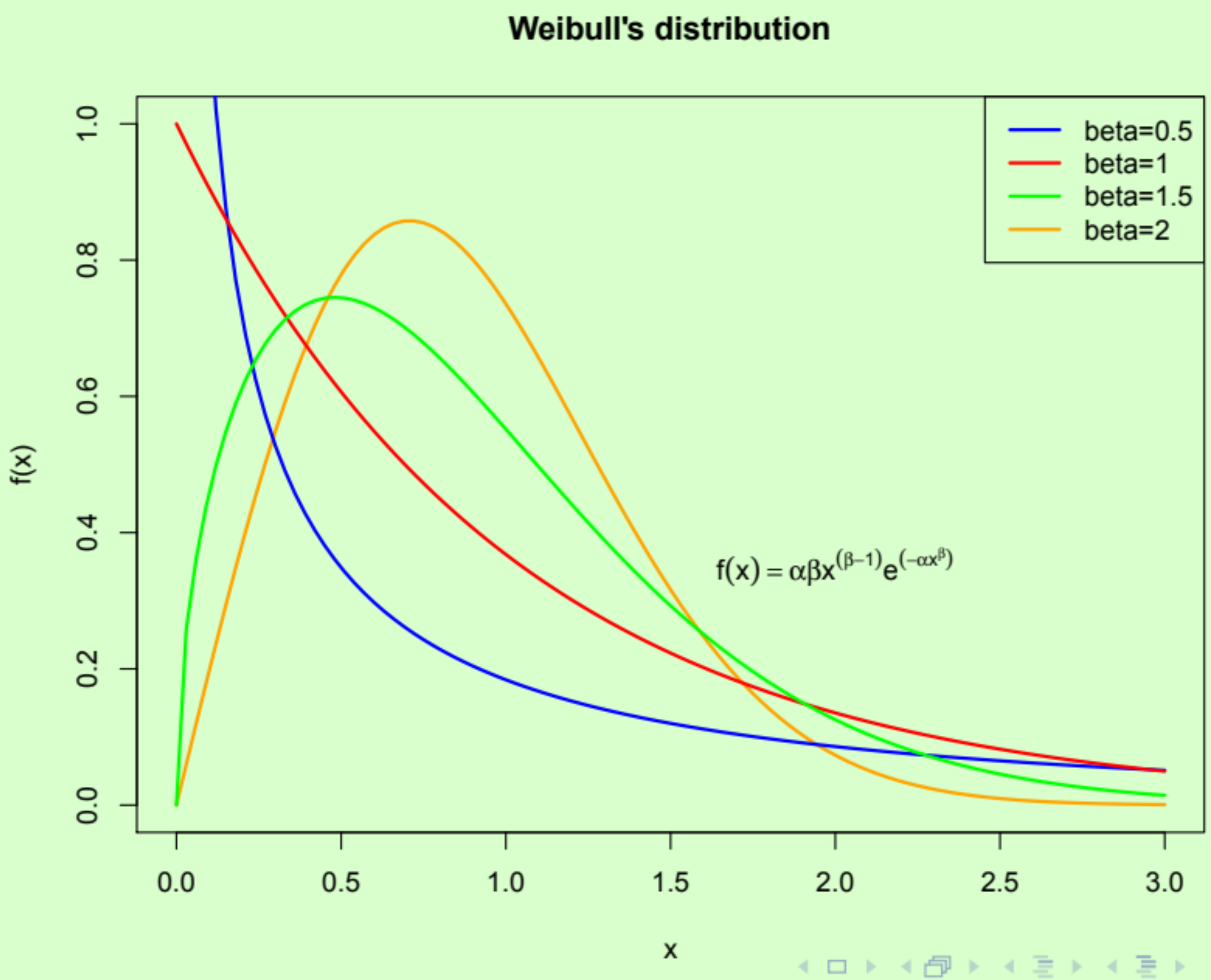
$$\mathcal{F} = \{f_\theta(\cdot) : \theta = (\alpha, \beta) \in (0, \infty) \times (0, \infty)\}$$

Joint density based on sample $(y_1, \cdots, y_n)$:

$$\begin{aligned} f_\theta(y_1, \cdots, y_n) &= \prod_{i=1}^n f_\theta(y_i) = \prod_{i=1}^n \{\alpha\beta y_i^{\beta-1} e^{-\alpha y_i^\beta}\} \\ &= \alpha^n \beta^n (\prod_{i=1}^n y_i)^{\beta-1} e^{-\alpha \sum_{i=1}^n y_i^\beta} \end{aligned}$$

Exponential law: $\beta = 1 \Rightarrow f_\theta(y) = \alpha e^{-\alpha y} 1_{(0,+\infty)}(y)$

Example 1.1: Weibull density plots



Example 1.1: Shifted Weibull's density

Goodness-of-fit plots indicate that the **shifted Weibull model is a better option to analyze DMBA data.**

Extended model for Y: shifted Weibull's density for

$\forall y \geq G; \theta' = (\alpha, \beta, G)$

$$f_{\theta'}(y) = \alpha\beta(y - G)^{\beta-1} e^{-\alpha(y-G)^{\beta}}$$

Maximum Likelihood estimators (MLE):

- Common value of $\hat{G} = 100$ and $\hat{\beta} = 3$ for both groups.
- The scale parameter, α, is different for each group:
  $\hat{\alpha}_1 = 0.0076$ and $\hat{\theta}'_1 = (0.0076, 3, 100)$
  $\hat{\alpha}_2 = 0.0062$ and $\hat{\theta}'_2 = (0.0062, 3, 100)$

How can we use and interpret these findings?

Example 1.1: Shifted Weibull's survival

By means of the survival function: $\forall y \geq G$

$$S_{\theta'}(y) = \text{Prob}\{Y > y\} = 1 - F_{\theta'}(y) = e^{-\alpha(y-G)^{\beta}}$$

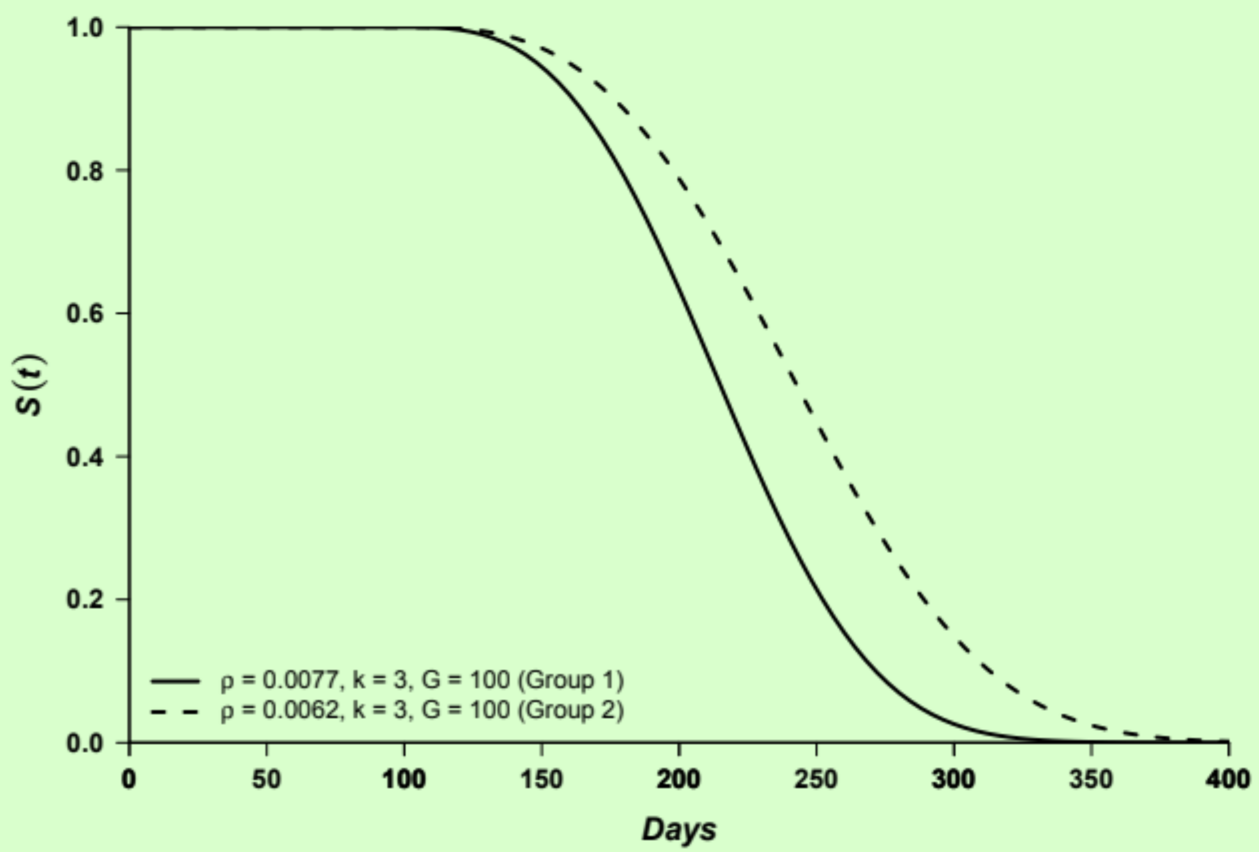


Figure: Estimated survival function in both groups of rats.

Example 1.1: Shifted Weibull's survival

Based on the estimates, the survival after 250 days is:

- $\hat{S}_{\hat{\theta}'_1}(250) = \exp\{-(0,0077(250-100))^3\} = 0.2288$
- $\hat{S}_{\hat{\theta}'_2}(250) = \exp\{-(0,0062(250-100))^3\} = 0.4423$

The estimates show a better survival for rats on Group 2 (Diet 2).
For instance, the Probability of surviving 250 days on Diet 1 is 23% while on Diet 2 is 44%.

These point estimates have to be accompanied by their standard errors or confidence intervals. Later in the course, we will work on them.

Statistical schools. Two broad approaches

- Frequentist
- Bayesian (covered in 2nd sem.)

Both approaches have variants. Essentially

- Frequentist based on *sampling theory*. Takes θ as a fixed unknown constant
- Bayesian uses *inverse probability*. Treats θ as having a probability distribution
- Differences between both approaches, not as huge as they seem. From a practical perspective have much in common
- If properly applied, they usually produce results that differ by less than the analyzed models are likely to differ from reality

SUPP READING: R.A. Fisher in the 21st century (Efron)

Main leaders of the Frequentist approach (I)

- **Karl Pearson** (1857–1936). English biostatistician, eugenicist, and mathematician. Credited with establishing the discipline of mathematical statistics
- **Ronald Fisher** (1890–1962). English great statistician, evolutionary biologist, geneticist and eugenicist. He layed the foundations for modern statistical science (Fisher's information, ANOVA, etc.).
- **Egon Pearson** (1895–1980), son of K. Pearson and, like his father, a leading British statistician. Known for development of Neyman–Pearson lemma of statistical hypothesis testing.

Main leaders of the Frequentist approach (II)

- **Jerzy Neyman** (1894–1981). Polish mathematician and statistician who first introduced the modern concept of a confidence interval into statistical hypothesis testing and, with Egon Pearson, revised Ronald Fisher's null hypothesis testing.
- **Abraham Wald** (1902–1950). Jewish Hungarian mathematician who contributed to decision theory, geometry and econometrics, and founded the field of sequential analysis. One of his well-known statistical works was written during World War II on how to minimize the damage to bomber aircraft and took into account the survivorship bias in his calculations.

But important to know that ...

Fisher, K.Pearson and Galton: Eugenics

Their work was closely integrated into **eugenics** (means literally “well-born”).
They were racist who believed in the sterilization of those they considered inferior.
Pearson himself spoke approvingly of Nazi plans during his retirement speech in 1934

SUPP READINGS (see links in Atenea):

- *Award “retired” over R.A.Fisher’s links to eugenics*. Significance Vol 17 (4), page 2
- *RA Fisher and the science of hatred.* Science & Tech (28 July 2020).

Is Statistics Racist? Need to think about

Tools of statistics are not employed in an abstract mathematical space, but, rather, their purpose is to tell us something about the real world.
In order to do this, we need to make assumptions to connect the mathematics to reality.

These assumptions range from the way in which we describe the world in mathematical terms, the types of questions that we ask and the methods we use to answer them, and the way in which we then interpret the real world meaning of our statistical results.

There is plenty of opportunity to "embed bias in these assumptions"
D. Cleather: `https://medium.com/swlh/is-statistics-racist-59cd4ddb5fa9`

Outline

Estimation

Given a model and some data:

- Think about the meaning of the parameters
- It is possible to get reasonable "guesses" from the data
- If this process can be written down as a mathematical "recipe", we call the "guess" an **estimate**
- We can study its properties under **data replication** to get an idea of its uncertainty
- General approaches:
  - Method of moments (develop in full in Unit 2)
  - Method of least squares (assumed known)
  - Maximum Likelihood estimation: MLE (Only likelihood function and related functions. MLE postponed to Unit 3)
  - Bayesian Inference (Bayesian Analysis course in Spring semester)

Definition of Statistic

- We start with samples $(Y_1, \cdots, Y_n)$ coming from a population defined by $Y \sim f(\mathbf{y}, \theta)$.

Definition

Any measurable function of the sample $(Y_1, \cdots, Y_n)$ that does not depend on θ is a statistic.

- If θ is known, the probability characteristics of the population Y are known.
- θ **is usually unknown**. The goal is to obtain good approximations of θ based on the observed values $(y_1, \cdots, y_n)$.

Estimands, Estimators and Estimates

Definition

Estimand: The quantity of interest that needs to be estimated.

For instance, θ
$\tau(\theta) = \log(\theta)$, $\tau(\theta) = 1/\theta$.
If $\theta = (\mu, \sigma^2)$, $\tau(\theta) = \frac{\sigma}{\mu}$ the coefficient of variation.

Definition

Estimator: The method or rule used to obtain an approximation of the estimand.

For instance MLE, method of moments, etc

Definition

Estimate: The specific value obtained from a given method and dataset $(y_1, \cdots, y_n)$.

Point Estimators and Point Estimates

Definition

A point estimator $\hat{\theta}$ of θ is any measurable function of the sample $(Y_1, \cdots, Y_n)$ that does not depend on θ. Any statistic is a point estimator.

For instance, $\hat{\theta} = \hat{\theta}(Y_1, \cdots, Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is a random variable.

Definition

A point estimate $\hat{\theta}$ of θ is the actual result of the estimation process, often presented as a numerical value, and based on the observed values $(y_1, \cdots, y_n)$.

For instance, $\hat{\theta} = \hat{\theta}(y_1, \cdots, y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i$ is a real value.

Likelihood function

Parameter values that make the observed data appear relatively probable are more ***likely*** *to be correct than parameter values that make the observed data appear relative improbable.*

Definition

$f(\mathbf{y}|\theta)$: joint pdf (pmf) of a sample $(Y_1, \cdots, Y_n)$.

Given the observations $\mathbf{y} = (y_1, \cdots, y_n)$, the likelihood function for θ is a function of θ defined as

$$L(\theta|\mathbf{y}) = f(\mathbf{y}|\theta) = \prod_{i=1}^{n} f_Y(y_i|\theta).$$

Example 1.2. Joint density function for $Y \sim$ Poisson (λ)

The joint density for $\mathbf{y} = (y_1, \cdots, y_n)$ is:

$$f(\mathbf{y}|\lambda) = \prod_{i=1}^{n} f_Y(y_i|\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{y_i}}{y_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} = e^{-n\lambda} \frac{\lambda^{n\bar{y}_n}}{\prod_{i=1}^{n} y_i!}$$

That is, $f(\mathbf{y}|\lambda)$ varies with y for given λ.

Suppose $n = 12$ and $\lambda = 2$

$$f(\mathbf{y}|\lambda = 2) = e^{-24} \frac{2^{\sum_{i=1}^{12} y_i}}{\prod_{i=1}^{12} y_i!} = e^{-24} \frac{2^{12\bar{y}_{12}}}{\prod_{i=1}^{12} y_i!}$$

Remark: $f(\mathbf{y}|\lambda = 2)$ takes different values for different $(y_1, \cdots, y_{12})$ where $y_j \in \{0, 1, \cdots, 12\}$. In fact, it only depends on two functions of $(y_1, \cdots, y_{12})$: $\sum_{i=1}^{12} y_i$ and $\prod_{i=1}^{12} y_i!$

Example 1.2. Likelihood function for $Y \sim$ Poisson (λ)

The likelihood function for λ

$$L(\lambda|\mathbf{y}) = e^{-n\lambda}\frac{\lambda^{n\bar{y}_n}}{\prod_{i=1}^{n} y_i!}$$

In this case $L(\lambda|\mathbf{y})$ varies with λ for given y.

Suppose $n = 12$ and $\prod_{i=1}^{12} y_i! = 41472$ and $\sum_{i=1}^{12} y_i = 1.83$

$$L(\lambda|\mathbf{y}) = e^{-12\lambda}\lambda^{1.33}/41472$$

Remark: $L(\lambda|\mathbf{y})$ takes different values for different λ where $\lambda \in (0, \mathcal{R}]$. In fact, it only depends on two functions of λ: $e^{-n\lambda}$ and $\lambda^{n\bar{y}_n}$.
In many occasions, we only need to work with the part of $L(\theta|\mathbf{y})$ depending on λ, for instance with $L(\lambda|\mathbf{y}) = e^{-n\lambda}\lambda^{n\bar{y}_n}$ because $L(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)$

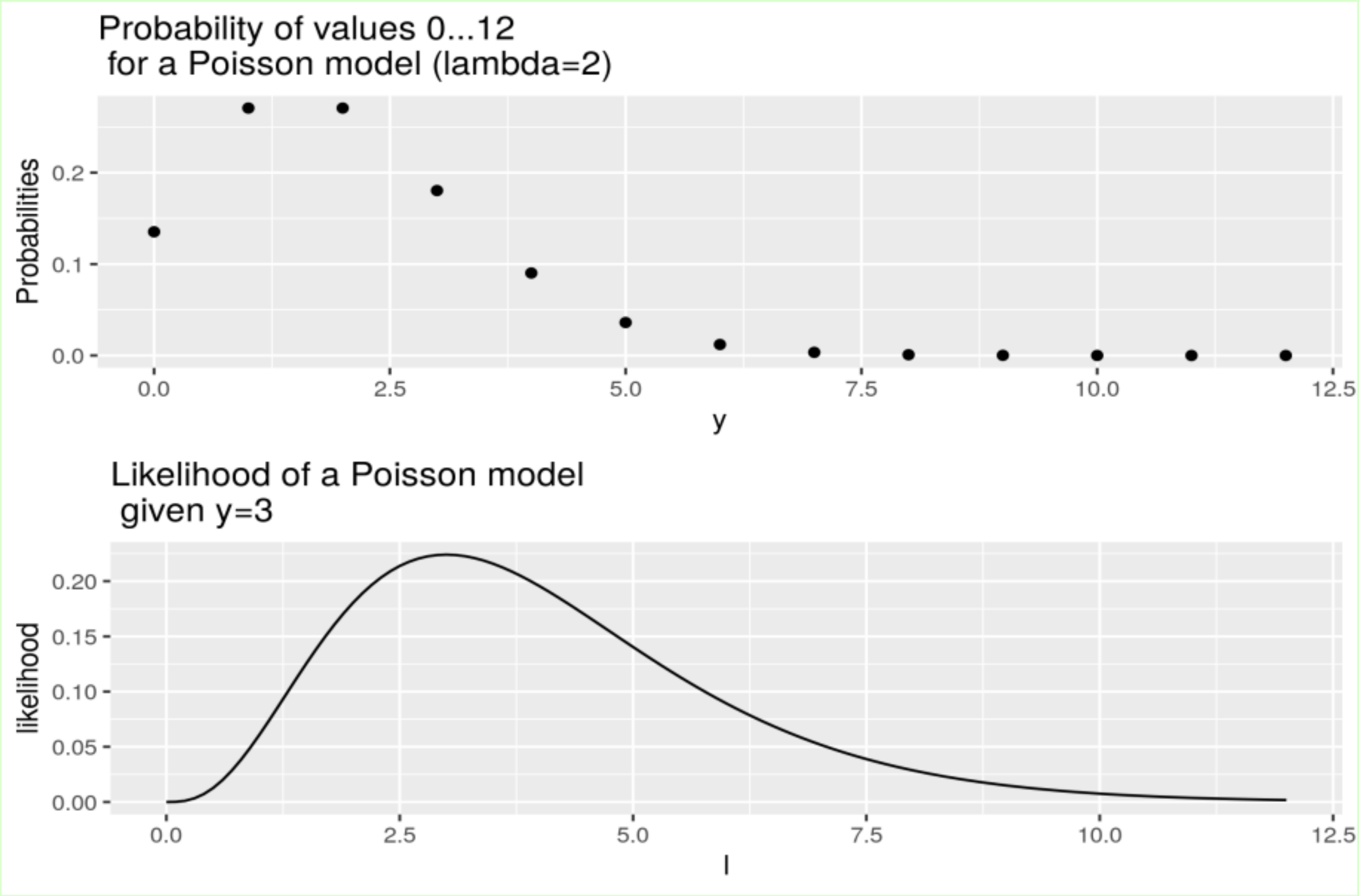
Example 1.2. Probability/density versus likelihood ($n = 1$)

- **Probability (joint density**) describes a function of the outcome $Y = y$ given a fixed parameter value. For example, if $Y \sim \text{Poisson}(\lambda)$ with $\lambda = 2$, *what is the probability of* $Y = 3$*?*
  $P(Y = 3|\lambda = 2)) = e^{-2}\frac{2^3}{3!} = 0.1804$
  If an experiment with a Poisson(2) is repeated many, many times, 18% *of the cases we will observe* $y = 3$.

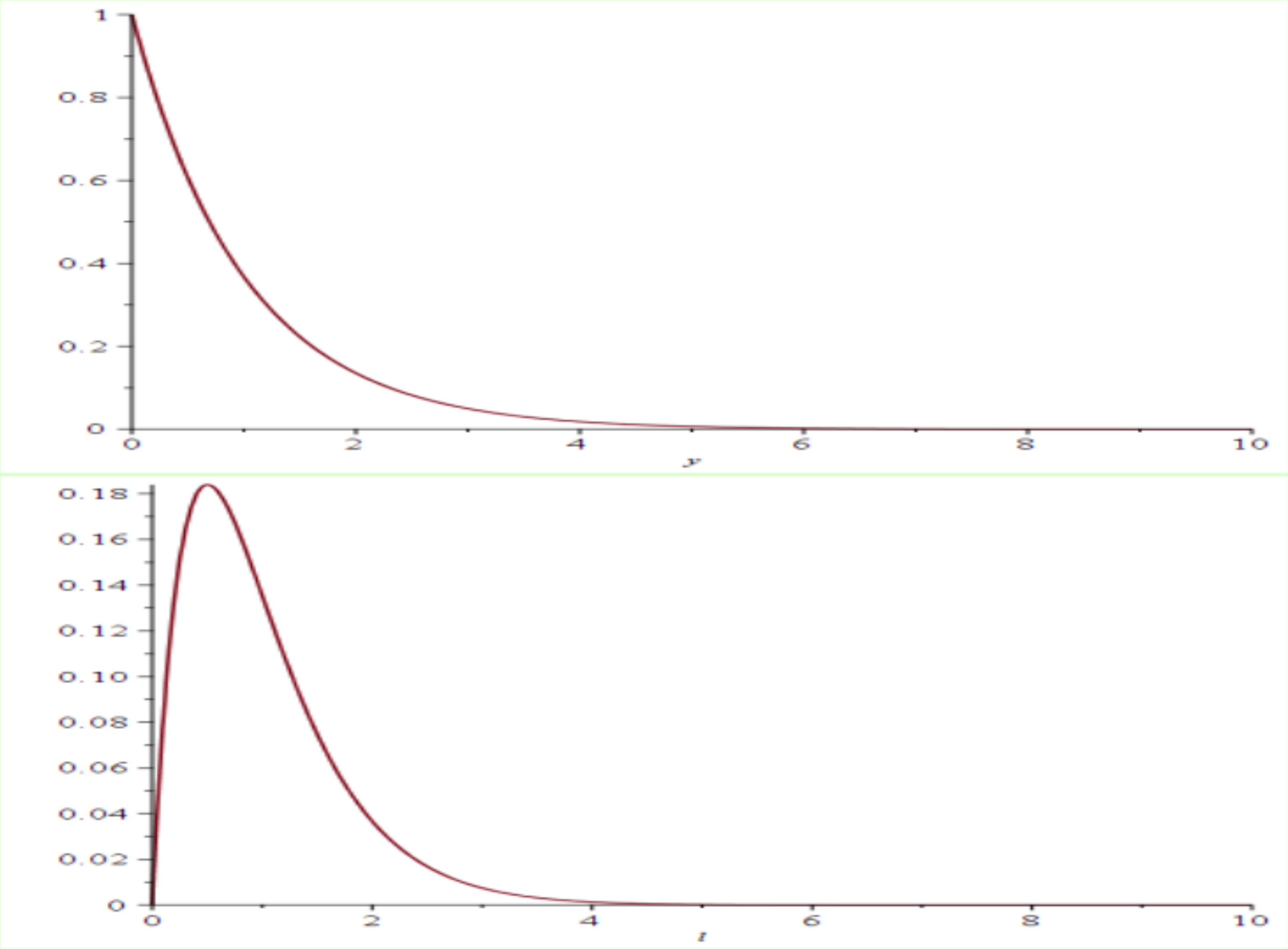
Example 1.2. Probability/density versus likelihood ($n = 1$)

- **Probability (joint density)** describes a function of the outcome $Y = y$ given a fixed parameter value. For example, if $Y \sim \text{Poisson}(\lambda)$ with $\lambda = 2$, *what is the probability of $Y = 3$?*
  $P(Y = 3|\lambda = 2)) = e^{-2}\frac{2^3}{3!} = 0.1804$
  If an experiment with a Poisson(2) is repeated many, many times, 18% *of the cases we will observe $y = 3$.*
- **Likelihood** describes a function of a parameter given an outcome.
  If we have observed $y = 3$, *how likely is that $\lambda = 2$?*
  $L(\lambda = 2|Y = 3) = e^{-2}\frac{2^3}{3!} = 0.1804$
  Given this observation $y = 3$, the "chances" that Y comes from a Poisson(2) is only 18%.
  0.1804 is the same in both cases but refers to 2 different concepts

Example 1.2. Probability vs Likelihood using R (Poisson)



Example 1.3. Density vs Likelihood using Maple (Exponential)



Top: Density of exponential for $\lambda = 1$
Bottom: Likelihood of exponential if observed $y = 2$

Maximum likelihood estimators

For each sample $\mathbf{y} \in \mathcal{Y}$, the *maximum likelihood estimator* $\hat{\theta}$ of θ is the value of Θ that maximizes the likelihood $L(\cdot|\mathbf{y})$:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|\mathbf{y}).$$

- Intuitively, $\hat{\theta}$ is the value of the parameter that makes most plausible the observed sample.
- Maximum likelihood estimators are usually good estimators and, in general, have optimality properties.
- What do we mean by a **good estimator** and by **good properties**? Need to consider repeated estimation under repeated replications of the data-generating process (frequentistic approach)

Postponed to UNIT 3

Bias and Variance

If we replicate the random data and repeat the estimation process it would result in a different **estimate** $\hat{\theta}$ for each replicate.
These values are observations of a random vector, the **estimator** of θ, also denoted by $\hat{\theta}$, or by $\hat{\theta}_n$ if we need to keep the sample size in mind.

Two properties are desirable

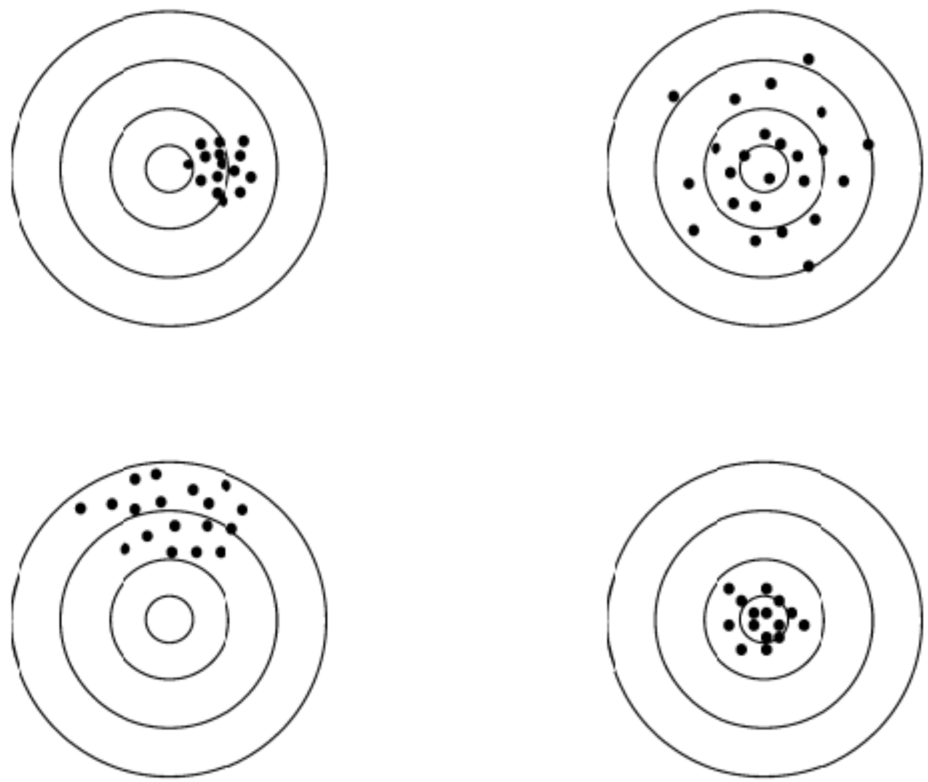
- **Small Bias** of $\hat{\theta}$ for θ implying the estimator is right on average.
  Biais $B_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$ is a systematic deviation from θ.
  $\hat{\theta}$ is UNBIASED if $B_\theta(\hat{\theta}) = 0$.
- **Low Variance** of $\hat{\theta}$ for θ implying that any individual estimate is quite precise
  $V_\theta(\hat{\theta}) = Var_\theta(\hat{\theta}) = E_\theta(\hat{\theta} - E_\theta(\hat{\theta}))^2$

MSE: Mean squared error (a tradeoff)

Mean squared error (MSE) of an estimator $\hat{\theta}$ of a parameter θ is $MSE_\theta(\hat{\theta}) = E_\theta\left((\hat{\theta}-\theta)^2\right) = V_\theta(\hat{\theta}) + \left(B_\theta(\hat{\theta})\right)^2$

✍ Deduce the formula

(a) (b)



(c) (d)

Example 1.4: Normal distribution

$X_1, \ldots, X_n$ s.r.s (simple random sample) of $X \sim N(\mu, \sigma^2)$, both $\theta = (\mu, \sigma^2)$ unknown

- $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$
- $E(\overline{X}_n) = \mu$
- $MSE_\theta\left(\overline{X}_n\right) = V_\theta(\overline{X}_n) = \frac{\sigma^2}{n}$

- $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$
- $E(S_n^2) = \sigma^2$

$$MSE_{\sigma^2}(S_n^2) = E_\theta\left((S_n^2 - \sigma^2)^2\right) = V_\theta(S_n^2) = 2\frac{\sigma^4}{n-1}.$$

✍ Deduce $V_\theta(S_n^2) = 2\frac{\sigma^4}{n-1}$

Example 1.4: Normal distribution

$X_1, \ldots, X_n$ s.r.s of $X \sim N(\mu, \sigma^2)$, both $\theta = (\mu, \sigma^2)$ unknown

✍ Compute the maximum likelihood estimator $\hat{\sigma}_n^2$ of σ^2

- $\hat{\sigma}_n^2 = \frac{n-1}{n} S_n^2$ is the maximum likelihood estimator of σ^2 and has lower mean squared error than S_n^2.

$$E(\hat{\sigma}_n^2) = \frac{n-1}{n}\sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2 \Longrightarrow B_\theta(\hat{\sigma}_n^2) = -\frac{1}{n}\sigma^2;$$

$$V(\hat{\sigma}_n^2) = \left(\frac{n-1}{n}\right)^2 V(S_n^2) = \frac{2(n-1)}{n^2}\sigma^4.$$

$$MSE_{\sigma_n^2}(\hat{\sigma}_n^2) = \frac{2(n-1)}{n^2}\sigma^4 + \frac{1}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4 < \frac{2}{n-1}\sigma^4.$$

The MLE $\hat{\sigma}_n^2$ is often preferred, although it has small bias, because its optimal asymptotic properties.

Search of the Best Unbiased Estimators

Multiple estimators $\hat{\theta}_n$ can be obtained for the same parameter θ.
It seems clear that MSE could be a **good criterion** to choose between two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$.

However, it is difficult to prove **general results** about minimum MSE estimators.
Our search of the minimum variance estimator will be restricted to **the class of unbiased estimators of θ.**

Change of notation!!! Estimators will be denoted by W (or W_n) instead of $\hat{\theta}_n$

Class of unbiased estimators of $\tau(\theta)$

For a given reparametrization $\tau(\theta)$, let $\mathcal{C}_\tau$ be the class of unbiased estimators of $\tau(\theta)$, that is, $\mathcal{C}_\tau = \{W : E_\theta(W) = \tau(\theta)\}$.

For instance, if $\theta = (\mu, \sigma^2)$, we might be interested in the coefficient of variation and take $\tau(\theta) = \frac{\mu}{\sigma}$.

In this case

$$MSE_{\tau(\theta)}(W) = E_\theta\left((W - \tau(\theta))^2\right) = V_\theta(W)$$

We will search the minimum variance estimator within the class $\mathcal{C}_\tau$ of unbiased estimators of $\tau(\theta)$.

Best unbiased estimators: Definition

GOAL: Estimation of $\tau(\theta)$

Definition

An estimator W^* (if needed W_n^*) is the BEST UNBIASED ESTIMATOR OF $\tau(\theta)$ or UMVUE (UNIFORM MINIMUM VARIANCE UNBIASED ESTIMATOR OF $\tau(\theta)$), if:

- $E_\theta(W^*) = \tau(\theta)$ for all $\theta \in \Theta$
- for any other estimator W, such that $E_\theta(W) = \tau(\theta)$ for all $\theta \in \Theta$, we have that $V_\theta(W^*) \leq V_\theta(W)$, for all $\theta \in \Theta$.

Example 1.2: $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$

- Compare $\overline{X}_n$ and S_n^2, both unbiased estimators of λ.
- .
$$V(S_n^2) = \frac{\lambda}{n} + \frac{2\lambda^2}{n-1} > \frac{\lambda}{n} = V(\overline{X}_n) \text{ if } n \geq 2.$$
S_n^2 cannot be the UMVUE for λ.
- The class of estimators $W_a = a\overline{X}_n + (1-a)S_n^2$, with $a \in \mathbf{R}$, only contains unbiased estimators of λ
$$E_\lambda(W_a) = a\lambda + (1-a)\lambda = \lambda,$$
- We can find a^* such that W_{a^*} is the best unbiased estimator of λ among those who have the form W_a.
- We cannot find the UMVUE by inspection

We introduce the concept of sufficiency and completeness. Rao-Blackwell and Lehman-Scheffé Theorems set the ground to obtain UMVUE estimators

Sufficient statistics

READ 4.1, 4.2 in Olive, 6.1, 6.2 in Casella and Berger, 2.5 in Held and Sabanés

A statistic T is sufficient for a parameter θ if it captures all the information that the sample contains about θ.
That is:

1. Any additional information (apart from the value of the statistic T) that the sample can contribute, does not provide relevant information about θ.
2. The information that provides a sufficient statistic about the parameter is all the information that the complete sample would provide.
3. The main advantage and goal of a sufficient statistic is the **reduction of the initial dimension**

Sufficiency principle (1)

If T is a sufficient statistic for θ, then any inference about θ should depend on the sample $\mathbf{X} = (X_1, \ldots, X_n)$ ***only via the value of*** $T(\mathbf{X})$**.**

EXAMPLE

- $X \sim N(\mu, \sigma^2)$
- Sample $\mathbf{X} = (X_1, \ldots, X_n)$
- Estimation of μ
  - We only have to collect and keep $T_1(\mathbf{X}) = \sum_{i=1}^n X_i$.
- Estimation of σ^2
  - We only have to collect and keep $T_2(\mathbf{X}) = \sum_{i=1}^n X_i^2$.

Sufficiency principle (2)

If we have two samples $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ such that $T(\mathbf{x}) = T(\mathbf{y})$, inference about θ should be the same whether $\mathbf{x}$ or $\mathbf{y}$ has been observed.

EXAMPLE

- $X \sim N(\mu, \sigma^2)$
- $T(\mathbf{X}) = \sum_{i=1}^n X_i$
- Samples $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ such that $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$, hence $T(\mathbf{x}) = \sum_{i=1}^n x_i = \sum_{i=1}^n y_i = T(\mathbf{y})$
- Estimation of μ: $\hat{\mu}(\mathbf{x}) = \frac{\sum_{i=1}^n X_i}{n}$
  - Since $T(\mathbf{x}) = T(\mathbf{y})$, $\hat{\mu}(\mathbf{x}) = \hat{\mu}(\mathbf{y})$, and the estimated value for μ would be in both cases $\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{i=1}^n y_i}{n}$.

We will show $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is sufficient for μ.

Definition Sufficient statistic

Definition

A statistic $T(\mathbf{X})$ is *sufficient* for θ if

- the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$, does not depend on θ.

  In other words, for a suitable function $g(t)$ that does not depend on θ
- If the law of $\mathbf{X}$ is discrete: $P[\mathbf{X} = \mathbf{x} | T = t; \theta] = g(t)$
- If the law of $\mathbf{X}$ is continuous: $f_{\mathbf{X}}(\mathbf{x} | T = t; \theta) = g(t)$

The above definition is operative, we will be able to use it to check sufficiency.

Example 1.5: Proportion p of faulty pieces in a production process from n pieces at random

Simple random sample $X_1, X_2, \ldots, X_n$ where

$$X_i = \begin{cases} 1 = \text{faulty} & \text{with probability } p \\ 0 = \text{correct} & \text{with probability } 1 - p \end{cases}$$

We will prove that to estimate p we only need $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$, that is, the number of 1 and 0 in the sample.

- A statistic that would take into account the position of the 1 and 0 in the sample would be **useless**.
- A statistic that will not consider every value, as for example $T(\mathbf{X}) = X_1$, would be **less appropriate**.
- All the samples of size n with the same number of 1 have the **same probability.**

Example 1.5: Proportion p of faulty pieces I

- The probability function of a sample $(x_1, x_2, \ldots, x_n)$ is:
$$f(x_1, x_2, \ldots, x_n; p) = p^t(1-p)^{n-t}$$
where $t = \sum_{i=1}^n x_i$, $x_i \in \{0, 1\}$, $i = 1, 2, \ldots, n$.
- Note that $f(x_1, x_2, \ldots, x_n; p)$ only depends on the number of 1 and 0 and not on the order in which they appear
- To estimate p, the statistic $T(\mathbf{X}) = \sum_{i=1}^n X_i$ **contains the same information** on p that the whole sample $X_1, X_2, \ldots, X_n$
- Differences between using $T(\mathbf{X})$ or $X_1, X_2, \ldots, X_n$:
  - Going from $(X_1, X_2, \ldots, X_n)$ to $\sum_{i=1}^n X_i$ there is a **data reduction** which does not involve any loss of information.
  - Diff. samples might give rise to the **same value of T.**

Example 1.5: Proportion p of faulty pieces II

- Compute, for all $t = 0, 1, \ldots, n$, the conditional probability of $\mathbf{X}$ given $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$:

$$P[\mathbf{X} = \mathbf{x} | T = t; p] = \frac{P[\mathbf{X} = \mathbf{x}, T = t; p]}{P(T = t; p)}$$

$$= \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}}$$

Given $(x_1, x_2, \ldots, x_n) \in \{0, 1\}^n$ and $t \in \{0, 1, \ldots, n\}$:

$$P[\mathbf{X} = \mathbf{x} \mid T = t; p] = \begin{cases} 0 & \text{if } t \neq \sum_{i=1}^{n} x_i \\ \frac{1}{\binom{n}{t}} & \text{if } t = \sum_{i=1}^{n} x_i \end{cases}$$

Example 1.5: Proportion p of faulty pieces III

- $P[\mathbf{X} = \mathbf{x}; p]$ depends on p which is the parameter we want to estimate and can be decomposed into two factors:

$$P(\mathbf{X} = \mathbf{x}; p) = P[\mathbf{X} = \mathbf{x} \mid T = t; p] \cdot P(T = t; p)$$

  - $P(T = t; p)$ depends on p
  - the conditional probability $P[\mathbf{X} = \mathbf{x} \mid T = t; p]$ *does not depend* on p.

- This idea will be formalized in the **factorization theorem.**

Factorization theorem

Advanced Statistical Inference

Guadalupe Gómez (UPC) & Alex Sánchez (UB)

Preliminaries

The Frequentist approach

Point estimation. Main results

Properties of point estimators

Theorem (Factorization theorem)

Let $f_{\mathbf{X}}(\mathbf{x}|\theta)$ the joint density function of $\mathbf{X} = (X_1, \ldots, X_n)$. The statistic $T(\mathbf{X})$ is sufficient for θ if and only if exist functions:

- $g(t; \theta)$
- $h(\mathbf{x})$

such that for all $\mathbf{x} \in \mathcal{X}$ and all $\theta \in \Theta$ the function $f(\mathbf{x}; \theta)$ can be factorized as:
$f_{\mathbf{X}}(\mathbf{x}; \theta) = g(T(\mathbf{x}); \theta)h(\mathbf{x})$.

✍ Try to prove it yourself for the discrete case
See a general proof in Casella and Berger

Example 1.5

$$P(\mathbf{X} = \mathbf{x}; p) = P(T = t; p) \cdot P[\mathbf{X} = \mathbf{x} \mid T = t; p]$$

We have:

- $\theta = p$
- $f_{\mathbf{X}}(\mathbf{x}; \theta) = P_p(\mathbf{X} = \mathbf{x}; p)$
- $T_1(\mathbf{X}) = \sum_{i=1}^n X_i$
- $g(T_1(\mathbf{x}); \theta) = P(T = t; p)$
- $h(\mathbf{x}) = P[\mathbf{X} = \mathbf{x} \mid T_1 = t; p] = 1/\binom{n}{t}$

Hence, $T_1(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for p:
$f_{\mathbf{X}}(\mathbf{x}; \theta) = P(\mathbf{X} = \mathbf{x}; p) = g(T_1(\mathbf{x}); \theta)h(\mathbf{x})$.

✍ Is $T_2(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i$ a sufficient statistic as well? **Not unicity of the sufficient statistics**

✍ Practice with the distributions from the exponential family to find out $T(\mathbf{X})$ in each case.

Rao-Blackwell's theorem

Rao-Blackwell's and Lehmann-Scheffé's theorems set the ground to obtain UMVUE estimators

Theorem (Rao-Blackwell theorem)

- $X_1, \ldots, X_n \sim X$ *(random sample) with density (or mass probability)* $f_{\mathbf{X}}(x|\theta)$.
- $T(\mathbf{X})$: *sufficient statistic for* θ
- $W(\mathbf{X})$: *unbiased estimator of* $\tau(\theta)$.
- *Define* $W_T = E_\theta(W|T)$.

Then:

i. W_T *is a statistic, function of* $\mathbf{X}$ *uniquely via* $T(\mathbf{X})$,

ii. W_T *is unbiased, that is,* $E_\theta(W_T) = \tau(\theta)$,

iii. $V_\theta(W_T) \leq V_\theta(W)$ *for all* $\theta \in \Theta$.

✍ Review Conditional Expected random variables

Exemples: $x_1, \dots, x_n$

1) $X_1, \dots, X_n \sim Poiss(\lambda)$

Volem aplicar teorema Rao...

$E(X_1) = \dots = E(X_n) = \lambda$

$T_1(X_1, \dots, X_n) = \underset{\substack{\| \\ \hat{\lambda}_1}}{X_1}$ unove per λ

$T_2(X_1, \dots, X_n) = \sum X_i$ sufficient statistic per λ

Objectiu: $T_3(X_1, \dots, X_n) = E(X_1 | T_2)$
$= \hat{\lambda}_2(X_1, \dots, X_n)$
$\downarrow$
Quin és el valor?

$E(X_1|T_2) = \dots = E(X_n|T_2)$ per Poiss

$$\left.\begin{aligned} \sum^n E(X_i|T_2) &= n\,E(X_1|T_2) \\ \| \\ E\left(\sum_{i=1}^n X_i | T_2\right) &= T_2 = \sum X_i \end{aligned}\right\} \Rightarrow \underset{\substack{\| \\ \hat{\lambda}_2}}{E(X_1|T_2)} = \frac{\sum^n X_i}{n} = \bar{X}$$

Notes: volem estimar λ
↳ mirem els sst (sufficient statistics) de poisson.

Rao Blackwell diu que $E(X_1|T_2)$ és un sst amb menys variància que $\hat{\lambda}_1$

Exemple: Rao-Blackwell

$\theta = e^{-\lambda}$; $P(X_1=k) = e^{-\lambda}\frac{\lambda^k}{k!}$

$$P(X_1=0) = e^{-\lambda}$$

$T_1(X_1,\dots,X_n) = \mathbb{1}\{X_1=0\}$ unb for θ

$T_2(X_1,\dots,X_n) = \sum^n X_i$ s St.

$T_3(X_1,\dots,X_n) = E(T_1|T_2)$ ← Objectiu.

$$E(T_1|T_2) = E(\mathbb{1}(X_1=0)|T_2) = P(X_1=0|\sum X_i)$$

Assumim que $\sum_{i=1}^n X_i = t$, un valor. Ara volem trobar T_2 amb t donada.

$$P(X_1=0|\sum X_i=t) = \frac{P(X_1=0,\sum X_i=t)^{(2)}}{P(\sum^n X_i=t)_{(1)}} = (*)$$

(1) = Sabem que $\sum X_i \sim P(n\lambda)$ ergo

$$P(\sum X_i=t) = e^{-n\lambda}\frac{(n\lambda)^t}{t!}$$

(2) $P(X_1=0,\sum X_i=t) = P(X_1=0,\sum_{i=2}^n X_i=t) =$

Sense X_1 ⇒ són independents

$$= P(X_1=0)P(\sum_2^n X_i=t) = e^{-\lambda}\frac{[e^{-(n-1)\lambda}]^t}{t!}$$

(*)

$$= \frac{e^{-\lambda}e^{-(n-1)\lambda}\frac{((n-1)\lambda)^t}{t!}}{e^{-n\lambda}\frac{(n\lambda)^t}{t!}} = \left(\frac{n-1}{n}\right)^t$$

Nota: si al final depèn de λ, la has liat, només ha de dependre de les dades

Ergo:

$$T_3(X_1,\dots,X_n) = \left(\frac{n-1}{n}\right)^{T_2} = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^n x_i}$$

Exemple Rao-Blackwell

$X_1,\dots,X_n \sim \text{Unif}[0,\theta]$ Estima θ

Intuició $\begin{cases} \to \max \\ \to E(\text{Unif})? \end{cases}$

1. Unbiased estimator:

$T_1(X_1,\dots,X_n) = 2X_1$ Unb θ

$E(X_i) = \frac{\theta}{2}$

2. Sufficient statistic

$T_2(X_1,\dots,X_n) = \max X_i$ s St of θ

$T_3(X_1,\dots,X_n) = E(2X_1 / T_2)$?

$E(X_1 | \max X_i) = (\max X_i)\frac{1}{n} + \max X_i \cdot \frac{n-1}{n} =$

↳ Dos casos: $X_1 = \max X_n \to \frac{1}{n}$

$X_2 \neq \max X_n \to \frac{n-1}{n}$

$\frac{\max X_i}{n}\left(1+\frac{1}{2}(n-1)\right)$

$\frac{\max X_i}{n} \frac{2+n-1}{2} = \frac{\max X_i}{n} \cdot \frac{n+1}{2}$

$T_3(X_1,\dots,X_n) = \left(\frac{n+1}{n}\right)\max X_i$

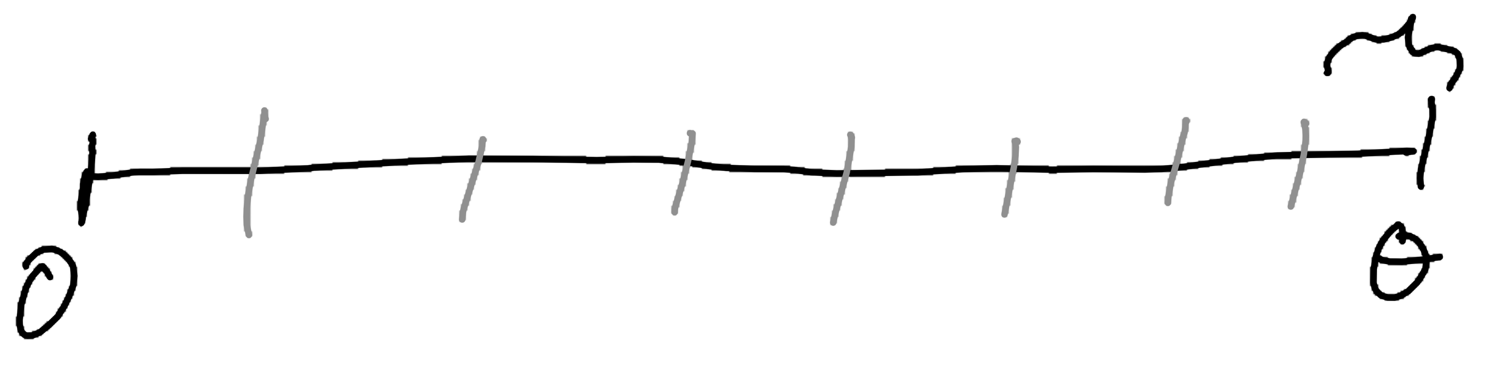
$\max X_i$ → sempre està esbiaixat! El que hem fet és desesbiaixar-lo en certa manera.

Idea: trobar la densitat $\max X_i$ i fer el

$E(\max X_i) = \frac{n}{n+1}\theta$

↓

Aquest és un 1



Proof of Rao-Blackwell theorem

i. Since T is sufficient for θ, the distribution of $\mathbf{X}|T$ does not depend on $\theta \implies$ the conditional distribution of $W(\mathbf{X})|T(\mathbf{X})$, does not depend on $\theta \implies E(W|T)$ does not depend on θ and is uniquely function of T.

ii. From the Law of iterated expectation,

$$E_\theta(W_T) = E_\theta(E(W|T)) = E_\theta(W) = \tau(\theta).$$

iii. From the Law of iterated variance,

$$V_\theta(W) = V_\theta(E(W|T)) + E_\theta(V_\theta(W|T)) =$$

$$= V_\theta(W_T) + E_\theta(V_\theta(W|T)) \geq V_\theta(W_T) \implies$$

$V_\theta(W_T) \leq V_\theta(W)$ for all $\theta \in \Theta$.

✍ Review the Law of iterated variance

Rao-Blackwellization: Blackboard exercises

$X_1, \cdots, X_n$ sample from X

1. $X \sim Poisson(\lambda)$. Estimate λ
2. $X \sim Poisson(\lambda)$. Estimate $e^{-\lambda}$
3. $X \sim Unif[0, \theta]$. Estimate θ

Example 1.6: $X \sim B(k, \theta)$, k known and θ unknown. WORK ON YOUR OWN

Based on sample $X_1, \cdots, X_n \sim B(k, \theta)$. Estimate $\tau(\theta) = P_\theta(X = 1) = k\theta(1-\theta)^{k-1}$

- $T(\mathbf{X}) = \sum_{i=1}^n X_i \sim B(nk, \theta)$ sufficient and complete for θ.
- The estimator $W(\mathbf{X}) = I\{X_1 = 1\}$ is unbiased for $\tau(\theta)$, that is, $E(W(\mathbf{X})) = E(I\{X_1 = 1\}) = P_\theta(X = 1) = \tau(\theta)$
- The **UMVUE** for $\tau(\theta) = P_\theta(X = 1) = k\theta(1-\theta)^{k-1}$ is:

$$W_T(\mathbf{X}) = E\left[I\{X_1 = 1\} | T(\mathbf{X}) = \sum_{i=1}^n X_i \right] = \phi\left(\sum_{i=1}^n X_i \right)$$

Practicality issues from Rao-Blackwell's and Lehmann-Scheffé's theorem

Whenever you can, you should restrict the search to those **unbiased estimators that are function of a sufficient and complete statistic** T.

Rao-Blackwell's and Lehmann-Scheffé's method to get the UMVUE consists on:

- Find an unbiased estimator W for the parameter of interest, **the simplest you might think**
- Build another unbiased estimator conditioning the previous one to a sufficient and complete statistic, $W_T(\mathbf{X}) = E_\theta(W|T)$

The new unbiased estimator W_T:

- is a sufficient statistic
- $V(W_T) \leq V(W)$, in fact W_T is the UMVUE
- is unique if $V(W_T) < \infty$

Fisher's information: Definitions

- LOG-LIKELIHOOD FUNCTION: $l(\theta|\mathbf{x}) = \log L(\theta|\mathbf{x})$
- SCORE FUNCTION, $S_{\mathbf{X}}(\theta)$ or $S(\theta|\mathbf{X})$, is the first derivative of $l(\theta|\mathbf{x})$: $S_{\mathbf{X}}(\theta) = S(\theta|\mathbf{X}) = \frac{\partial}{\partial\theta} l(\theta|\mathbf{X})$
- FISHER'S INFORMATION that $\mathbf{X}$ carries about θ:

$$i_{\mathbf{X}}(\theta) = -\frac{\partial^2}{\partial\theta^2} l(\theta|x) = -\frac{\partial}{\partial\theta} S(\theta|\mathbf{X})$$

- OBSERVED FISHER'S INFORMATION that $\mathbf{X}$ carries about θ: $i_{\mathbf{X}}(\hat{\theta}_{MLE}(\mathbf{x}))$ where θ has been replaced by $\hat{\theta}_{MLE}$
- EXPECTED FISHER'S INFORMATION that $\mathbf{X}$ carries about θ: $I_{\mathbf{X}}(\theta) = E_\theta \left[i_{\mathbf{X}}(\theta) \right]$

Remark: Use upper case $I_{\mathbf{X}}(\theta)$ for expected Fisher's information

Ex 1.2: Likelihood related functions for Poisson

- LIKELIHOOD FUNCTION for λ is:

$$L(\lambda|\mathbf{y}) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{y_i}}{y_i!} = e^{-n\lambda} \frac{\lambda^{n\bar{y}_n}}{\prod_{i=1}^{n} y_i!}$$

- LOG-LIKELIHOOD FUNCTION for λ is:

$$\log L(\lambda|\mathbf{y}) = -n\lambda + n\bar{y}_n \log\lambda - \log\left(\prod_{i=1}^{n} y_i!\right)$$

- SCORE FUNCTION for λ is: $S(\lambda|\mathbf{y}) = \frac{\partial}{\partial\lambda} l(\lambda|\mathbf{y}) = -n + \frac{n\bar{y}_n}{\lambda}$
- FISHER'S INFORMATION for λ is:

$$i_{\mathbf{Y}}(\lambda) = -\frac{\partial^2}{\partial\lambda^2} l(\lambda|\mathbf{y}) = -\frac{\partial}{\partial\lambda} S_{\mathbf{Y}}(\lambda) = +\frac{n\bar{y}_n}{\lambda^2}$$

- EXPECTED FISHER'S INFORMATION for λ is:

$$I_{\mathbf{Y}}(\lambda) = E\left\{\frac{n\bar{y}_n}{\lambda^2}\right\} = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$$

Example 1.2: Likelihood functions for Poisson: Working with data

Suppose $n = 12$, $\prod_{i=1}^{12} y_i! = 41472$, $\sum_{i=1}^{12} y_i = 1.83$

$$
\begin{aligned}
L(\lambda|(y_1,\cdots,y_{12})) &= e^{-12\lambda}\lambda^{1.83}/41472 \\
l(\lambda|(y_1,\cdots,y_{12})) &= -12\lambda + 1.83\log(\lambda) - \log(41472) \\
S(\lambda|(y_1,\cdots,y_{12})) &= -n + \frac{n\bar{y}_{12}}{\lambda} = -12 + \frac{1.83}{\lambda} \\
i(\lambda|(y_1,\cdots,y_{12})) &= +\frac{n\bar{y}_{12}}{\lambda^2} = \frac{1.83}{\lambda^2} > 0
\end{aligned}
$$

$$
S(\lambda|(y_1,\cdots,y_{12})) = 0 \Longrightarrow \hat{\lambda}_{MLE} = \bar{y}_{12} = 0.1525
$$

$$
\begin{aligned}
L(\hat{\lambda}_{MLE}|(y_1,\cdots,y_{12})) &= 0,000001238425632 \\
l(\hat{\lambda}_{MLE}|(y_1,\cdots,y_{12})) &= -6,907130068 \\
S(\hat{\lambda}_{MLE}|(y_1,\cdots,y_{12})) &= 0 \\
i(\hat{\lambda}_{MLE}|(y_1,\cdots,y_{12})) &= 78,68852459
\end{aligned}
$$

Fisher's regularity conditions

Fisher's information concept and its properties depend on some regularity conditions on the behaviour of the density that we summarize next:

$\mathbf{X} = (X_1, \ldots, X_n) \sim X$ with density $f(\mathbf{x}|\theta)$, $\theta \in \Theta \subseteq \mathbf{R}$.

1. The parameter space Θ is an open interval
2. The support of $f(\mathbf{x}|\theta)$ does not depend on θ
3. Identifiability: $f(\mathbf{x}|\theta_1) \neq f(\mathbf{x}|\theta_2)$ whenever $\theta_1 \neq \theta_2$
4. $L(\theta)$ is twice continuously differentiable wrt θ
5. $\int f(\mathbf{x}|\theta)dx$ can be twice differentiated under $\int$
6. H1: $\int$ and $\frac{d}{d\theta}$ can be exchanged
7. H2: $\int$ and $\frac{d^2}{d\theta^2}$ can be exchanged

See Appendix on Estimation for more information on these conditions

Properties of Score function and Expected Fisher's information

Given a sample $\mathbf{X} = (X_1, \ldots, X_n) \sim X$ with joint density $f(\mathbf{x}|\theta)$. Recall the score and Fisher's information functions:

$$\begin{aligned} S_{\mathbf{X}}(\theta) &= S(\theta|\mathbf{X}) = \frac{\partial}{\partial\theta} l(\theta|\mathbf{X}) = \frac{\partial}{\partial\theta} \log f(\mathbf{x}|\theta) \\ I_{\mathbf{X}}(\theta) &= -E\left\{\frac{\partial}{\partial\theta} S(\theta|\mathbf{X})\right\} = -E\left\{\frac{\partial^2}{\partial\theta^2} \log f(\mathbf{x}|\theta)\right\} \end{aligned}$$

1. Under H1: The score function has mean 0. $E_\theta\{S_{\mathbf{X}}(\theta)\} = 0$
2. Under H1 and H2: $I_{\mathbf{X}}(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta} \log f(X|\theta)\right)^2\right]$
3. Under H1 and H2: The variance of the score function is equal to the expected Fisher information. $I_{\mathbf{X}}(\theta) = V_\theta\{S_{\mathbf{X}}(\theta)\}$
4. Under H1: $I_{\mathbf{X}}(\theta) = nI_{X_1}(\theta)$

Fisher's information **Matrix** for $\theta \in \Theta \subseteq \mathbf{R}^k$ in UNIT 3

Proof of 1: $E_\theta \left\{ S_{\mathbf{X}}(\theta) \right\} = 0$

Advanced Statistical Inference

Guadalupe Gómez (UPC) & Alex Sánchez (UB)

Preliminaries

The Frequentist approach

Point estimation. Main results

Properties of point estimators

Appendix

Under H1:

$$
\begin{aligned}
E(S_{\mathbf{X}}(\theta)) &= E\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) = E\left(\frac{\frac{\partial}{\partial\theta}f(\mathbf{X}|\theta)}{f(\mathbf{x}|\theta)}\right) \\
&= \int\cdots\int \frac{\frac{\partial}{\partial\theta}f(\mathbf{X}|\theta)}{\cancel{f(\mathbf{x}|\theta)}}\cancel{f(\mathbf{x}|\theta)}dx_1\cdots dx_n \quad \text{“}\theta \text{ és la “truth”} \\
&= \int\cdots\int \frac{\partial}{\partial\theta}f(\mathbf{X}|\theta)dx_1\cdots dx_n \\
&\overset{H1}{=} \frac{\partial}{\partial\theta}\int\cdots\int f(\mathbf{X}|\theta)dx_1\cdots dx_n \overset{f \text{ is density}}{=} \frac{\partial}{\partial\theta}1 = 0
\end{aligned}
$$

The score function at the truth θ is on average 0.

Score és VA amb mitjana 0.

Proof of 2: $I_{\mathbf{X}}(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]$

$$I_{\mathbf{X}}(\theta) = -E\left\{\frac{\partial}{\partial\theta}S(\theta|\mathbf{X})\right\} = -E\left\{\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{x}|\theta)\right\}$$

Under H1 and H2:

-> regla cadena. $(uv)' = u'v + v'u$

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{x}|\theta) &= \frac{\partial}{\partial\theta}\left[\frac{1}{f(\mathbf{x}|\theta)}\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta)\right] \\
&= \frac{-1}{f^2(\mathbf{x}|\theta)}\left(\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta)\right)^2 + \frac{1}{f(\mathbf{x}|\theta)}\frac{\partial^2}{\partial\theta^2}f(\mathbf{x}|\theta) \\
&= -\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2 + \frac{1}{f(\mathbf{x}|\theta)}\frac{\partial^2}{\partial\theta^2}f(\mathbf{x}|\theta) = A + B
\end{aligned}$$

mateix d'a posterior

$$* \; E(B) = E_\theta\left[\frac{1}{f(\mathbf{X}|\theta)}\frac{\partial^2}{\partial\theta^2}f(\mathbf{X}|\theta)\right] = \int \frac{\partial^2}{\partial\theta^2}f(\mathbf{x}|\theta)dx \overset{H2}{=} 0$$

$$I_{\mathbf{X}}(\theta) = -E\left\{\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{x}|\theta)\right\} = -E(A) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right]$$

Proof of 3: $I_{\mathbf{X}}(\theta) = V_\theta\left\{S_{\mathbf{X}}(\theta)\right\}$

(trivial, diu la Lupe)

Under H1 and H2.
Since $E_\theta\left\{S_{\mathbf{X}}(\theta)\right\} = 0$,

$$
\begin{aligned}
V_\theta\left\{S_{\mathbf{X}}(\theta)\right\} &= E_\theta\left\{\left(S_{\mathbf{X}}(\theta)\right)^2\right\} \\
&= E_\theta\left\{\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right\} \underset{*}{=} I_{\mathbf{X}}(\theta)
\end{aligned}
$$

* Diapo anterior, just demostrat.

Proof of 4: $I_{\mathbf{X}}(\theta) = nI_{X_1}(\theta)$

Under H1 and H2 and taking into account that
$f(\mathbf{X}|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$

pf 3

$$
\begin{aligned}
I_{\mathbf{X}}(\theta) &= V_\theta \left\{ S_{\mathbf{X}}(\theta) \right\} \\
&= V_\theta \left(\frac{\partial}{\partial\theta} \log f(\mathbf{X}|\theta) \right) = V_\theta \left(\frac{\partial}{\partial\theta} \log \prod_{i=1}^{n} f(X_i|\theta) \right) \\
&= V_\theta \left(\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta) \right) \\
&= nV_\theta \left(\frac{\partial}{\partial\theta} \log f(X_1|\theta) \right) = nI_{X_1}(\theta).
\end{aligned}
$$

iid i independent

Because for every $i = 1, \cdots, n$, the random variables
$\frac{\partial}{\partial\theta} \log f(X_i|\theta)$ are iid.

Example 1.2: $X_1, \ldots, X_n$ s.r.s of $X \sim \text{Poisson}(\lambda)$

New random variables (upper case in X_i) are defined:

- $V_1 = f(X_i|\lambda) = e^{-\lambda}\frac{\lambda^{X_i}}{X_i!}$ $\rightarrow$ $\log(V_1) = V_2$
- $V_2 = l(\lambda|X_i) = -\lambda + X_i \log(\lambda) - \log(X_i!)$ $\rightarrow$ $\frac{\partial}{\partial \lambda} V_2$
- $V_3 = \underline{S(\lambda|X_i)} = -1 + \frac{X_i}{\lambda}$
- $V_4 = i_{\mathbf{X}}(\lambda) \overset{??}{=} \frac{X_i}{\lambda^2}$ $\rightarrow$ expected fishers information.

and it makes sense to take $E(V_4) = I_{X_i}(\lambda)$.
Hence the expected Fisher's information for λ is

$$I_{X_i}(\lambda) = E_\lambda\left\{\frac{X_i}{\lambda^2}\right\} = \frac{E_\lambda(X_i)}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda} = \frac{1}{Var(x)}$$

Note that
$\mathrm{Var}(V_3) = \mathrm{Var}(S(\lambda|X_i)) = \mathrm{Var}(-1 + \frac{X_i}{\lambda}) = \frac{\mathrm{Var}_\lambda(X_i)}{\lambda^2} = \frac{1}{\lambda}$ and it holds that $\mathrm{Var}(S(\lambda|X_i)) = I_{X_i}(\lambda)$

Expected Fisher's information of a one-to-one transformation $\psi = \tau(\theta)$ Canvi de variable a Fisher

Let θ be a scalar parameter and $\psi = \tau(\theta)$ a one- to-one transformation of θ (hence $\tau^{-1}(\psi) = \theta$)

The Expected Fisher information $I_X(\psi)$ of ψ is given by:

$$I_X(\psi) = I_X(\theta)\left\{\frac{d\tau(\theta)}{d\theta}\right\}^{-2}$$

Depèn de $\tau(\theta)$

Indeed, in general (**Work on your own**) Lo que farien (Regla cadena!!)

$$(h(\tau(\theta)))' = (h(\psi))'\frac{d\tau(\theta)}{d\theta} \implies (h(\psi))' = (h(\tau(\theta)))'\left(\frac{d\tau(\theta)}{d\theta}\right)^{-1}$$

Hence,

$$S_X(\psi) = \frac{\partial}{\partial\psi}\{\log f(\mathbf{X}|\psi)\} = \frac{\partial}{\partial\theta}\{\log f(\mathbf{X}|\theta)\}\left(\frac{d\tau(\theta)}{d\theta}\right)^{-1} \implies$$

$$I_X(\psi) = V_\psi\{S_X(\psi)\} = V_\theta\{S_X(\theta)\}\left\{\frac{d\tau(\theta)}{d\theta}\right\}^{-2}$$

Cas Poisson anterior

$$e^{-\lambda} = P(X=0)$$
$$\parallel$$
$$\psi \Rightarrow -\log(\psi) = \lambda$$

$$e^{-\lambda}\frac{\lambda^x}{x!}$$
$$\Uparrow$$
$$\psi\frac{(-\log\psi)^x}{x!}$$

Cramér-Rao's bound for θ

$\mathbf{X} = (X_1, \ldots, X_n) \sim X$ with df $f(\mathbf{x}|\theta)$, $\theta \in \Theta \subseteq \mathbf{R}$ holding H1.

Is dades

Let $W(\mathbf{X})$ be unbiased estimator of θ, $E_\theta(W(\mathbf{X})) = \theta$

Theorem

The inequality of Cramér-Rao states that

la variància més petita que pots tenir.

$$V_\theta(W(\mathbf{X})) \geq \frac{1}{I_{\mathbf{X}}(\theta)}$$

and $1/I_{\mathbf{X}}(\theta)$ *is known as the* CRAMER RAO'S BOUND.

Equivalently
$$I_{\mathbf{X}}(\theta) \geq \frac{1}{V_\theta(W(\mathbf{X}))}$$

Recall that $I_{\mathbf{X}}(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta} \log f(\mathbf{X}|\theta)\right)^2\right] = nI_{X_1}(\theta)$

Si el teu $W(x)$ compleix la cota $\Rightarrow$

$W(x)$ és UMVUE

General Cramér-Rao's bound for $\tau(\theta)$

L'anterior però en general.

Interest relies on $\tau(\theta)$ and we assume $\tau(\theta)$ is differentiable at θ. (Surt aplicant el canvi de variable!)

Let $W(\mathbf{X})$ be unbiased estimator of $\tau(\theta)$, $E_\theta(W(\mathbf{X})) = \tau(\theta)$

Theorem

The inequality of Cramér-Rao states that

$$V_\theta(W(\mathbf{X})) \geq \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{I_{\mathbf{X}}(\theta)}$$

and $\frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{I_{\mathbf{X}}(\theta)} = \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{nI_{X_1}(\theta)}$ *is the* CRAMER RAO'S BOUND.

Proof Cramér-Rao's inequality:

For every rv X and Y use Cauchy-Schwarz's inequality:

$$V(X) \geq \frac{(\mathrm{Cov}(X,Y))^2}{V(Y)}$$

- Consider taking $X = W(\mathbf{X})$ and
- $Y = \frac{\partial}{\partial\theta}(\log f(\mathbf{X}|\theta)) = S(\theta|\mathbf{x})$,
- and proving that $V(Y) = I_{\mathbf{X}}(\theta)$ and → visit proofs.

$\underbrace{(\mathrm{Cov}(X,Y))^2 = \left(\frac{d}{d\theta}\tau(\theta)\right)^2}_{\text{hem de demostrar-ho}}$ to get

$$V_\theta(W(\mathbf{X})) \geq \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{I_{\mathbf{X}}(\theta)}$$

[A.] Since $E_\theta\left[S(\theta|\mathbf{x})\right] = 0$ we have already proved that

$$V(Y) = V(S(\theta|\mathbf{x})) = I_{\mathbf{X}}(\theta)$$

Perque Cauchy Swartz és certa?

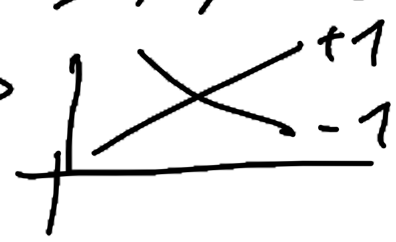
$\hookrightarrow \mathrm{Corr}(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$

$\mathrm{Corr}^2(X,Y) = \dfrac{\mathrm{Cov}^2(X,Y)}{\mathrm{Var}(X)\,\mathrm{Var}(Y)} \leq 1 \Rightarrow$ Cauchy Swartz!

$\mathrm{Corr}(X,Y) = \pm 1 \Rightarrow X, Y$ seen linealment dependents.

$\Updownarrow$

$Y = \alpha + \beta X$



Proof Cramér-Rao's inequality:

[B.]
$\mathrm{Cov}\left(W(\mathbf{X}), S(\theta|\mathbf{X})\right) = \mathrm{Cov}\left(W(\mathbf{X}), \frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) = \frac{d}{d\theta}\tau(\theta)$.
Indeed:

$$
\begin{aligned}
& \mathrm{Cov}\left(W(\mathbf{X}), \frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) \\
= \; & E_\theta\left[W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right] = E_\theta\left[W(\mathbf{X})\frac{\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta)}{f(\mathbf{x}|\theta)}\right] \\
= \; & \int\cdots\int W(\mathbf{x})\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta)dx_1\ldots dx_n \\
\overset{H2, h(\mathbf{x})=W(\mathbf{x})}{=} \; & \frac{d}{d\theta}\underbrace{\int\cdots\int W(\mathbf{x})f(\mathbf{x}|\theta)dx_1\ldots dx_n}_{=E_\theta(W(\mathbf{X}))=\tau(\theta)} = \frac{d}{d\theta}\tau(\theta).
\end{aligned}
$$

$$\mathrm{Cov}(X,Y) = E\left[(X-E(X))(Y-E(Y))\right] = E(XY) - \underset{\downarrow \atop 0}{E(Y)}E(X)$$

Efficiency

Definition

An estimator $W(\mathbf{X})$ is **Efficient** if it is unbiased and reaches Cramér Rao's bound, that is:

1. $E_\theta(W(\mathbf{X})) = \tau(\theta) \leftarrow$ unbiased
2. $\mathrm{Var}_\theta(W(\mathbf{X})) = \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{I_{\mathbf{X}}(\theta)}$ CRAMÉR RAO'S BOUND.

Definition

The **Efficiency, e,** of an unbiased estimator $W(\mathbf{X})$ of $\tau(\theta)$ is the ratio between the Cramer Rao's bound and its variance:

$$e(W(\mathbf{X})) = \frac{\frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{I_{\mathbf{X}}(\theta)}}{V_\theta(W(\mathbf{X}))} \begin{array}{l} \to \text{Cramer Rao} \\ \to \text{Var.} \end{array}$$

Consell: reproduir exemple 1.2 amb totes les famílies

Example 1.2: $X_1, \ldots, X_n \sim X \sim \text{Poisson}(\lambda)$

$E(X_1) = \lambda$, $\tau(\lambda) = \lambda$, $\tau'(\lambda) = 1$. $I_\lambda(x) = \frac{1}{\lambda} \cdots \frac{1}{\lambda} = \frac{n}{\lambda}$

$$I_\lambda(\mathbf{X}) = E_\lambda\left[\left(\frac{\partial}{\partial\lambda}\log\prod_{i=1}^{n} f(X_i|\lambda)\right)^2\right] = -nE_\lambda\left[\frac{\partial^2}{\partial\lambda^2}\log f(X_1|\lambda)\right]$$

$$= -nE_\lambda\left[\frac{\partial^2}{\partial\lambda^2}\log\left(\frac{e^{-\lambda}\lambda^{X_1}}{X_1!}\right)\right] = -nE_\lambda\left[-\frac{X_1}{\lambda^2}\right] = \frac{n}{\lambda}.$$

For any unbiased estimator of λ, W,

$$V_\lambda(W) \geq \frac{1}{n/\lambda} = \frac{\lambda}{n} = \text{CR bound}$$

Since $V_\lambda(\overline{X}) = \frac{\lambda}{n} \implies \hat{\lambda}_n = \overline{X}$ achieves the CR lower bound, so it is efficient (and UMVUE).

Other properties of Cramér-Rao theorem

1. $W(\mathbf{X})$ unbiased estimator of $\tau(\theta)$. $W(\mathbf{X})$ is efficient for $\tau(\theta) \iff \exists a(\theta)$ such that, $\forall \theta$,

$$a(\theta)(W(\mathbf{x}) - \tau(\theta)) = \frac{\partial}{\partial\theta} \log L(\theta|\mathbf{x}) = S(\theta|\mathbf{x})$$

To prove it, we need to establish a connection between the estimator $W(\mathbf{X})$ such that $E[W(\mathbf{X})] = \tau(\theta)$, the Cramér-Rao (CR) bound, and the score function $S(\theta|\mathbf{X}) = \frac{\partial}{\partial\theta} \log L(\theta|\mathbf{X})$.

($\Longrightarrow$) Assume $W(\mathbf{X})$ is Efficient

If $W(\mathbf{X})$ it is unbiased and efficient for $\tau(\theta) \implies W(\mathbf{X})$ reaches the CR bound, that is,

$$\mathrm{Var}(W(\mathbf{X})) = \frac{\left(\frac{d\tau(\theta)}{d\theta}\right)^2}{I_{\mathbf{X}}(\theta)}$$

Proof 1. ($\Longrightarrow$) Assume $W(\mathbf{X})$ is Efficient

Recall that we have proved that if $W(\mathbf{X})$ achieves the CR bound, equality in Cramér-Rao inequality holds , which is achieved if and only if:

$$Cov(W(\mathbf{X}), S(\theta|\mathbf{X})) = \frac{d\tau(\theta)}{d\theta}.$$

Furthermore, we have,

$$\begin{aligned} Corr(W(\mathbf{X}), S(\theta|\mathbf{X})) &= \frac{Cov(W(\mathbf{X}), S(\theta|\mathbf{X}))}{\sqrt{Var(W(\mathbf{X}))V(S(\theta|\mathbf{X}))}} \\ &= \frac{\frac{d\tau(\theta)}{d\theta}}{\sqrt{Var(W(\mathbf{X}))I_{\mathbf{X}}(\theta)}} = \pm 1 \end{aligned}$$

Recall that $Corr(X, Y) = \pm 1 \iff \exists a$ such that

$$a(X - E(X)) = Y - E(Y)$$

Hence we conclude that $\exists a(\theta)$ such that $a(\theta)(W(\mathbf{X}) - \tau(\theta)) = S(\theta|\mathbf{X})$.

Proof 1. ($\Longleftarrow$) Assume the Existence of $a(\theta)$

Suppose there exists a function $a(\theta)$ such that:

$$a(\theta)(W(\mathbf{X}) - \tau(\theta)) = S(\theta|\mathbf{X}).$$

Taking expectations on both sides yields:

$$E[a(\theta)(W(\mathbf{X}) - \tau(\theta))] = E[S(\theta|\mathbf{X})] = 0 \implies E[W(\mathbf{X})] = \tau(\theta)$$

Taking variances on both sides yields:

$$Var(W(\mathbf{X})) = \frac{E[S^2(\theta|\mathbf{X})]}{a(\theta)^2} = \frac{I_{\mathbf{X}}(\theta)}{a(\theta)^2}.$$

To achieve efficiency, we need:

$$Var(W(\mathbf{X})) = \frac{\left(\frac{d\tau(\theta)}{d\theta}\right)^2}{I_{\mathbf{X}}(\theta)}$$

which can be satisfied by choosing $a(\theta)^2 = \frac{(I_{\mathbf{X}}(\theta))^2}{\left(\frac{d\tau(\theta)}{d\theta}\right)^2}$

Other properties of Cramér-Rao theorem

2. $W(\mathbf{X})$ unbiased estimator of $\tau(\theta)$. If $\log L(\theta|\mathbf{x})$ is an absolutely continuous function of $\theta \in \Theta$, the equality

$$a(\theta)(W(\mathbf{x}) - \tau(\theta)) = \frac{\partial}{\partial\theta} \log L(\theta|\mathbf{x}) = S(\theta|\mathbf{x})$$

holds $\Longleftrightarrow$ the distribution of X belongs to the **exponential family.** That is, if $\exists$ $h(\mathbf{x})$, $c(\theta)$ and $k(\theta)$ s.t.

$$L(\theta|\mathbf{x}) = h(\mathbf{x})c(\theta)\exp(W(\mathbf{x})k(\theta)),$$

The score function relates efficiency and maximum likelihood: Indeed, to obtain the maximum likelihood estimator of θ we solve the equation $S(\theta|\mathbf{x}) = 0$.
Remark: The proof is in the appendix. A more technical derivation can be found in the paper of Joshi (1976) in Atenea.

Example 1.7: No guaranty that CR bound is reached for any unbiased estimator

$\mathbf{X} = (X_1, \ldots, X_n) \sim N(\mu, \sigma^2)$ with (μ, σ^2) unknown. Let $W(\mathbf{X})$ an unbiased estimator of σ^2. Applying last result (slide 82) and developing the likelihood of a $N(\mu, \sigma^2)$

$$\frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2 | \mathbf{x}) = \frac{n}{2\sigma^4} \left(\frac{\sum_{i=1}^n (x_i - \mu)^2}{n} - \sigma^2 \right),$$

hence $W(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$ would be the unbiased estimator reaching the Cramér-Rao bound ($= \frac{1}{nI(\sigma^2)} = \frac{2\sigma^4}{n}$.) This estimator only makes sense if μ is known.

Note that $S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ is the UMVUE for σ^2 (by Lehmann-Scheffe) but does not reach Cramér-Rao bound $V(S^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$

Outline

APPENDIX ON ESTIMATION

- The log-likelihood function
- The family of binomial distributions (with n known) is complete
- Proof Lehmann-Scheffé's theorem
- Fisher's regularity conditions
- Extra comments on the Score function
- Proofs of some properties related to Cramér-Rao's bound

The log-likelihood function

- The MLE of θ is obtained as:
  $\hat{\theta}_{MLE}(\mathbf{y}) = \arg\max_{\theta \in \Theta} L(\theta|\mathbf{y}) = \arg\max_{\theta \in \Theta} f_{\mathbf{Y}}(\mathbf{y}|\theta)$
- Due to the form of many likelihoods it is usually simpler to maximise the log-likelihood

$$l(\theta|\mathbf{y}) = \log L(\theta|\mathbf{y}),$$

which yields the same result due to the monotonicity of the $\log$ function.

$$\hat{\theta}_{MLE} = \hat{\theta}_{MLE}(\mathbf{y}) = \arg\max_{\theta \in \Theta} L(\theta|\mathbf{y}) = \arg\max_{\theta \in \Theta} l(\theta|\mathbf{y}).$$

Indeed, the log-likelihood and its first and second derivatives play an important role in statistical inference.

Example 1.3: MLE for Exponential

$$Y \sim \exp(\lambda) \Longrightarrow f(y|\lambda) = \lambda e^{-\lambda y} I_{[0,\infty)}(y),\ \lambda > 0.$$

- Suppose that $n = 1$ and $y_1 = 2$ is observed.
- $L(\lambda|y_1 = 2) = \lambda e^{-2\lambda}$ and we look for its maximum for values $\lambda > 0$.
- $$L'(\lambda|2) = e^{-2\lambda}(1 - 2\lambda);\ L'(\lambda|2) = 0 \Longrightarrow \lambda = \frac{1}{2}.$$

  Since $L(\lambda|2) \geq 0$ and $\lim_{\lambda \longrightarrow 0} L(\lambda|2) = \lim_{\lambda \longrightarrow \infty} L(\lambda|2) = 0$ it follows that the critical point of $L(\lambda|2)$ is a maximum. So,

$$\hat{\lambda}_{MLE}(y_1 = 2) = \frac{1}{2}.$$

Complete statistics

A distribution family is complete if the only unbiased estimators of 0 are those identical to 0.

Definition

$f_T(t|\theta)$ density function of a statistic T.
The distribution family $\{f_T(t|\theta) : \theta \in \Theta\}$ is **complete** if for any real-valued function $g(t)$ such that $E_\theta(g(T)) = 0$ for all θ then $P_\theta(g(T) = 0) = 1$ for all θ.

In that case we will say that T is a **complete statistic**.

Example: The family of binomial distributions (with n known) is complete

- $T \sim B(n, p)$, $0 < p < 1$ and n known
- g function such that $E(g(T)) = 0$ for all $p \in (0, 1)$.

Then:

$$
\begin{aligned}
0 = E(g(T)) &= \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} \\
&= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p} \right)^t .
\end{aligned}
$$

- This is an identically zero polynomial of degree n in $p/(1-p)$ and $p/(1-p) \in (0, \infty) \Rightarrow$
- $g(t)\binom{n}{t} = 0$ for all $t \in 0, 1, \ldots, n \Rightarrow g(t) = 0$ for all $t \in 0, 1, \ldots, n \Rightarrow$
- $P_p(g(T) = 0) = 1$, for all $p \Rightarrow$
- The distribution family of T is complete.

Proof Lehmann-Scheffé's theorem (self work):

- As we already know, $W_T(\mathbf{X})$ is unbiased.
- Whatever unbiased estimator of $\tau(\theta)$ is used, the same estimator $W_T(\mathbf{X})$ is obtained: Indeed:
  - W' be such that $E_\theta(W') = \tau(\theta)$: $W'_T = E_\theta(W'|T)$;

$$g(T) = E_\theta(W|T) - E_\theta(W'|T).$$

  - $E_\theta(g(T)) = \tau(\theta) - \tau(\theta) = 0 \implies P_\theta(g(T) = 0) = 1$ because T is complete $\implies P_\theta(W_T = W'_T) = 1$.
- $W_T(\mathbf{X})$ estimator is the UMVUE. Indeed: Let W' be such that $E_\theta(W') = \tau(\theta)$. By Rao-Blackwell theorem,

$$V_\theta(E(W'|T)) \leq V_\theta(W') \quad \text{for all } \theta,$$

but we have just seen that $W_T = E(W'|T)$, then $V_\theta(W_T) \leq V_\theta(W')$ for all θ, hence W_T is UMVUE.
- The uniqueness of W_T is guaranteed by previous theorem: *Unicity of UMVUE if it exists.*

Fisher's regularity conditions (I)

Fisher's information concept and its properties depend on some regularity conditions on the behaviour of the density that we summarize next (see the Estimation Appendix for these conditions stated in a rigorous manner)

$\mathbf{X} = (X_1, \ldots, X_n) \sim X$ with density f. $f(\mathbf{x}|\theta)$, $\theta \in \Theta \subseteq \mathbf{R}$.

1. The parameter space Θ is an open interval, θ must not be at the boundary of the parameter space Θ
2. The support of $f(\mathbf{x}|\theta)$ does not depend on θ
3. Identifiability: $f(\mathbf{x}|\theta_1) \neq f(\mathbf{x}|\theta_2)$ whenever $\theta_1 \neq \theta_2$
4. The likelihood $L(\theta) = f(\mathbf{x}|\theta)$ is twice continuously differentiable with respect to θ

Fisher's regularity conditions (II)

$\mathbf{X} = (X_1, \ldots, X_n) \sim X$ with density f. $f(\mathbf{x}|\theta)$, $\theta \in \Theta \subseteq \mathbf{R}$.

5. the integral $\int f(\mathbf{x}|\theta)dx$ can be twice differentiated under the integral sign and in particular:
   - H1: For any $h(\mathbf{x})$ such that $E_\theta|h(\mathbf{X})| < \infty$: $\int$ and ∂ can be exchanged:

$$\frac{d}{d\theta}\int..\int h(\mathbf{x})f(\mathbf{x}|\theta)dx_1..dx_n = \int..\int h(\mathbf{x})\left[\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta)\right]dx_1..dx_n.$$

H2: For any $h(\mathbf{x})$ such that $E_\theta|h(\mathbf{X})| < \infty$: $\int$ and ∂^2 can be exchanged:

$$\frac{d^2}{d\theta^2}\int..\int h(\mathbf{x})f(\mathbf{x}|\theta)dx_1..dx_n = \int..\int h(\mathbf{x})\left[\frac{\partial^2}{\partial\theta^2}f(\mathbf{x}|\theta)\right]dx_1..dx_n.$$

Extra comments on the Score function

$$S_{\mathbf{X}}(\theta) = \frac{\partial}{\partial\theta} l(\theta|\mathbf{X}) = \frac{\partial}{\partial\theta} \log L(\theta|\mathbf{X}) = \frac{\partial}{\partial\theta} \log f(\mathbf{X}|\theta)$$

The score function at the true θ is on average 0. Hence, the MLE which has a score function value of 0 is on average equal to the true θ, but "only" true asymptotically.

$I(\theta|\mathbf{X})$ is the average negative curvature of $l(\theta)$ at the true θ. If $l(\theta)$ is steep and has a lot of curvature, then the information with respect to θ is large, and $S_{\mathbf{X}}(\theta)$ at the true θ will vary a lot.
Conversely, if $l(\theta)$ is flat, then $S_{\mathbf{X}}(\theta)$ will not vary much at true θ (does not have much information w.r.t. θ.

Some intuition behind $a(\theta)(W(\mathbf{X}) - \tau(\theta)) = S(\theta|\mathbf{X})$

The reason this is true can be intuitively understood as follows:
- The score function $S(\theta|\mathbf{X})$ is the "sensitivity" of the log-likelihood with respect to θ. If the scaled deviation $a(\theta)(W(\mathbf{X}) - \tau(\theta))$ is proportional to the score function, then it means the deviation of $W(\mathbf{X})$ from $\tau(\theta)$ is aligned with the sensitivity of the likelihood function. This alignment implies that $W(\mathbf{X})$ is "as good as possible" in estimating $\tau(\theta)$, achieving the lowest possible variance given by the CR bound.

In summary, the equality $E[(W(\mathbf{X}) - \tau(\theta))S(\theta|\mathbf{X})] = \frac{d\tau(\theta)}{d\theta}$ holds if and only if the condition $a(\theta)(W(\mathbf{X}) - \tau(\theta)) = S(\theta|\mathbf{X})$ is satisfied. This condition guarantees that $W(\mathbf{X})$ reaches the CR bound.

Alternative Proof of ($\Longrightarrow$) Assume $W(\mathbf{X})$ is Efficient

The CR theorem is based on

$$\left(\mathrm{Corr}(W(\mathbf{X}), \frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X}))\right)^2 = (\mathrm{Corr}(W(\mathbf{X}), S(\theta|\mathbf{X})))^2 \leq 1,$$

with equality if and only if exist $a(\theta)$ and $b(\theta)$ such that

$$\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X}) = a(\theta)W(\mathbf{X}) + b(\theta), \text{ with probability 1.}$$

Since

$$0 = E\left(\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X})\right) = a(\theta)E(W(\mathbf{X}))+b(\theta) = a\tau(\theta)+b(\theta),$$

$b(\theta) = -a(\theta)\tau(\theta)$. Then,

$$\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X}) = a(\theta)(W(\mathbf{X}) - \tau(\theta)), \text{ with probability 1.}$$

Proof of: $a(\theta)(W(\mathbf{x}) - \tau(\theta)) = \frac{\partial}{\partial\theta} \log L(\theta|\mathbf{x})$ holds $\iff X \sim$ exponential family

$$\frac{\partial}{\partial\theta} \log L(\theta|\mathbf{X}) = a(\theta)(W(\mathbf{X}) - \tau(\theta)) \implies$$

the solution to the differential equation is given by:

$$L(\theta|\mathbf{x}) = h(\mathbf{x})c(\theta)\exp(W(\mathbf{x})k(\theta))$$

if exist functions $h(\mathbf{x})$, $c(\theta)$ and $k(\theta)$.
Note that $c(\theta) = \exp(\int a(\theta)\tau(\theta)d\theta)$, $k(\theta) = \exp(\int a(\theta)d\theta)$.

Hence, whenever we have a sample from the exponential family, the sufficient and complete statistic $W(\mathbf{X})$ (unbiased of $E(W(\mathbf{X})) = \tau(\theta)$) reaches the Cramér-Rao bound, hence is efficient.
If the family is not an exponential family, the CR bound may not be a lower bound on the variance of unbiased estimators of $\tau(\theta)$.

Outline

Hypothesis testing. Introduction

Recall the question whether some defined restriction on θ is consistent with the data

Definition

A STATISTICAL HYPOTHESIS is an statement about the unknown values of the parameters of the population distribution or about the distribution itself.

READING:
The Fisher Neyman-Pearson Theories of Testing Hypothesis: One theory or two? by E. Lehmann

Olive Chapter 7 (Sections 7.1 to 7.4) for a concise exposition and many other examples

Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing by J.D. Perezgonzalez

Paper xulo per mirar

This paper presents a tutorial for data testing procedures, often referred to as **hypothesis testing theories**.

1. The first procedure introduced is Fisher's approach to data testing (1925) known as **tests of significance**.
2. The second is Neyman-Pearson's approach (1928) known as **tests of acceptance**.
3. The final procedure is the **incongruent combination** of the previous two theories into the current approach known as **null hypothesis significance testing (NHST)** (2004) and **will not be covered** because is wrong and should not be used.

↳ LMAO, buscar perquè

FISHER'S APPROACH TO DATA TESTING

Fisher's approach to data testing (1925) known as **tests of significance** is a tool for identifying research results of interest.

↳ no clarificar/decidir

Fisher's approach is eminently inferential and some of the steps can be set up a posteriori, once the research data are ready to be analyzed.

↳ Ta bé, perquè no estan fent Neymann-Pearson. Tú bé amb Fisher.

Some steps such as the setting of hypotheses and levels of significance can be worked out a priori.

Fisher's approach can be summarized into five steps:

Fisher's steps 1 and 2: Test and H_0

Quina pregunta contestem.

1. **Select an appropriate test T** for the research goal of interest and consider how the variables have been measured.

2. **Set up the null hypothesis H_0.**

   Es el que volem refutar, el que passa si no fas res

   com estava

The statistical distribution of T under H_0 is established theoretically and represents the random variability theoretically expected.

Example 1.9: Fisher's steps 1 and 2

Advanced Statistical Inference

Guadalupe Gómez (UPC) & Alex Sánchez (UB)

Preliminaries

The Frequentist approach

Appendix

Hypothesis testing

Tests of significance: Fisher's approach

Tests of acceptance: Neymann-Pearson's approach

For example, how to assess differences in the scores reported to a questionnaire in 2 different groups?

- $X_1, \cdots, X_{n_1} \sim N(\mu_1, \sigma)$ random sample scores group 1
- $Y_1, \cdots, Y_{n_2} \sim N(\mu_2, \sigma)$ random sample scores group 2
- (μ_1, μ_2, σ^2) unknown.
- $H_0: \ \mu_1 = \mu_2$ → no hi ha diferencia entre X, Y
- Consider the statistic $T_{n_1+n_2-2}$ to test $H_0: \ \mu_1 = \mu_2$

↳ $T_{n_1+n_2-2}$ és esta distic

$$T_{n_1+n_2-2} = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t$$

pooled estimator of the variance

with $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ unbiased estimator of σ^2.

Some "quantities" such as the variance or the degrees of freedom are estimated from the sample.

$$(n_1-1)S_1^2 + (n_2-1)S_2^2 = \cancel{(n_1-1)}\frac{\sum(x_i-\overline{x}_{n_1})^2}{\cancel{n_1-1}} +$$

$$\cancel{(n_2-1)}\frac{\sum Y_i - \overline{Y}_{n_2}}{\cancel{n_2-1}}$$

$$\frac{\sum(x_i-\overline{x}_{n})^2 + \sum(y_i-\overline{y}_{n_2})^2}{n_1-1+n_2-1}$$ és estimador

Fisher's step 3: *p*-value →també és VA

3. **Calculate the theoretical probability of the results under** H_0. Once the corresponding theoretical distribution is established, the probability (p-value) of any datum under H_0 is established
The p-value comprises the probability of the observed results and also of any other more extreme results
↳ a la dreta del p-valor

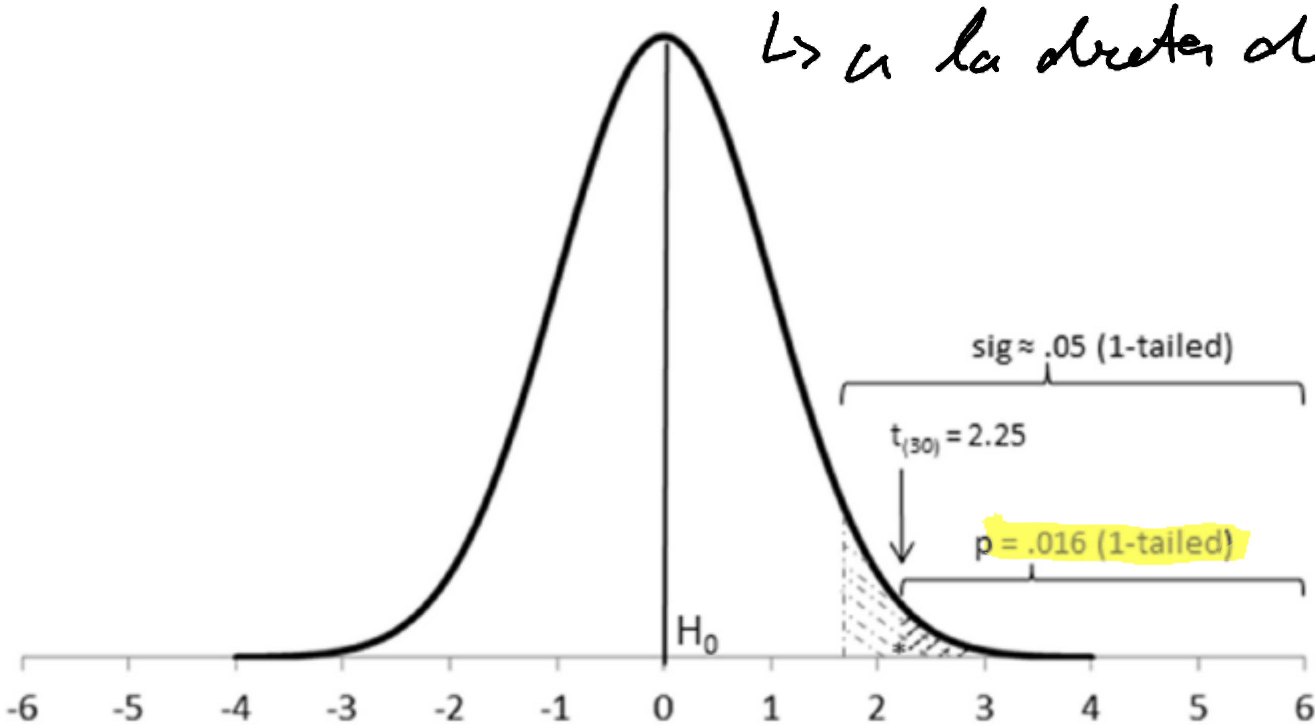


Figure: Location *t*-value and *p*-value on a t_{30}

Fisher's step 4: results of interest

4. **Assess the statistical significance of the results.**

- A research's result with a low p-value may be taken as evidence against H_0.
- How small these results ought to be in order to be considered statistically significant? Depends on the researcher question and may vary from research to research. Reporting exact p-values is very informative.
- What p-value is sufficiently small as to warrant rejection of H_0? Fisher, in the context of agricultural and biological experiments, proposed 0.01 or 0.05. In other setups, it might need to be much smaller,

Fisher's step 4: Level of significance

The assessment of research results is bound to a given level of significance, by comparing whether the research p-value, p, is smaller than such level of significance or not.

- If $p \leq$ level of significance, the result is considered statistically significant.
- If $p >$ level of significance, the result is considered statistically non-significant.

Level of significance: theoretical p-value used as a point of reference to help identify statistically significant results.

Remark: No need to set up a level of significance to be used in all occasions.
No need to be rigid: 0.049 and 0.051 have the same statistical significance.

Fisher's step 5: Interpretation and Conclusions

5. **Assess the statistical significance of the results.**
   A significant result is literally interpreted as a dual statement:
   - Either a rare result that occurs only with probability p (or lower) just happened, or ↳ Molta sort, hi ha chance.
   - H_0 does not explain the research results satisfactorily.

   Common interpretations:
   - "The null hypothesis did not seem to explain the research results well, thus we inferred that other processes exist that account for the results".
   - "The research results were statistically significant, thus we inferred that the treatment used accounted for such difference." → Fenòmens no extrems

   Remark: Non-significant results might be ignored but can provide useful information: they can show expected results in the expected direction and provide a hint about their magnitude.

Highlights of Fisher's approach

amb significance levels changing between H_0

- **Flexibility:** Most of the work is done a posteriori. Any number of H_0 might be tested. In this case, correction of the level of significance is needed.
- **Better suited for ad-hoc research projects and for exploratory research.**

→ Disadvantages

- **No power analysis:** Fisher talked about sensitiveness (similar concept) and how this could be increased by increasing sample size. There is no mathematical procedure for controlling sensitiveness predictably.
- **No alternative hypothesis.** There is no point in rejecting H_0 without an alternative explanation available. Fisher only considered alternative hypotheses as the negation of H_0. → No hi ha alternativa

Controversies about the *p*–value

- The *p*–value has become one of the most used (and abused) tool in applied statistics.
- This has entailed multiple wrong conceptions about its meaning and its interpretation, such as:
  - *Confusion with the null hypothesis probability* or
  - *Used to argue in favor of the null hypotheses when it has large values*

→ P(value) no és la probabilitat de que Ho sigui certa

All started because p-values were not used correctly and *the editors of Basic and Applied Social Psychology (BASP) announced (in 2015) that the journal would no longer publish papers containing P values because the statistics were too often used to support lower-quality research.* 1

Important reading (and watching) about the p-value controversy

- 2 ASA statement on p-values (2016): (llegir)
  https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf
- 3 Sir David Cox. "In gentle praise of Significance Tests" (2019). `https://www.youtube.com/watch?v=txLj_P9UlCQ`
- Follow Kareem Carr's thread (2020) on `https://twitter.com/kareem_carr/status/1312783404975493122?s=09`
- * 5-year Review (2024): P-Value Statements and Their Unintended(?) Consequences: The June 2019 ASA President's Corner (b)
  `https://errorstatistics.com/2024/05/10/5-year-review-p-value-statements-and-their-`
  This blog is recommended

Neymann-Pearson's (NP) approach to data testing

Jerzy Neyman and Egon Sharpe Pearson tried to improve Fisher's procedure and ended up developing an alternative approach to data testing.
Main conceptual innovation: explicit **alternative hypothesis** when testing research data.

NP approach to data testing can be considered as **tests of acceptance** and is the approach to follow for confirmatory analysis (in clinical trials, for instance) and to make decisions.

NP approach is more mathematical than Fisher's and does much of its work **a priori, at the planning stage of the research project**. It is summarized in the following eight main steps:

Neymann-Pearson's (NP) steps

A PRIORI STEPS for a research question of interest

1. Set up the expected effect size in the population
2. Set up the main hypothesis H_0 and the alternative hypothesis H_1. Perezgonzalez paper refers to H_M (main) and not H_0 (null) and H_A instead of H_1
3. Control type I error, α
4. Control type II error, β
5. Select an optimal test T
6. Calculate sample size required for optimal power, $1-\beta$
7. Calculate the critical value t_α of the test T

A POSTERIORI STEPS

8. Calculate the observed test value $T = t$
9. Decide in favor of either the main (null) or the alternative hypothesis.

The steps here do not follow those of Perezgonzalez's paper

Diferencies amb Fisher
-> 2 hipo
-> No p-value ⇓
Critical value per comparar amb l'estadistic.

-> Podem trobar el MILLOR test.

Step 1: Expected effect size in the population

The two groups under scrutiny differ by some degree: the effect size

- the smaller the effect size, the more difficult to appreciate such differences
- the larger the effect size, the easier to appreciate such differences.

Example 1.9

Assessing differences in the scores reported to a questionnaire in 2 different groups.
Assuming $X_1, \cdots, X_{n_1}$ and $Y_1, \cdots, Y_{n_2}$ random samples from $X \sim N(\mu_1, \sigma)$ and $Y \sim N(\mu_2, \sigma)$.
The effect size might be given by $\mu_1 - \mu_2$.

Step 2: Set up the hypothesis H_0 and H_1

NP's approach considers two competing hypotheses.
BUT, it only tests data under one of them.
The hypothesis **you do not want to reject too often** is the one tested.
The alternative hypothesis represents a second population that sits alongside the population of the main hypothesis (H_0) on the same continuum of values.

H_0 is often written such as $H_0 : \mu_1 - \mu_2 = 0$
but can also be written incorporating the minimum expected effect (MES) size $H_0 : \mu_1 - \mu_2 = 0 \pm MES$

- Values within such minimum threshold are considered reasonably probable under H_0,
- values outside are more likely under H_1.

Step 2: Simple and Composite Hypothesis

Advanced Statistical Inference

Guadalupe Gómez (UPC) & Alex Sánchez (UB)

Preliminaries

The Frequentist approach

Appendix

Hypothesis testing

Tests of significance: Fisher's approach

Tests of acceptance: Neymann-Pearson's approach

Definition

A statistical hypothesis that completely specifies the population distribution is called a *simple hypothesis*, otherwise is called a *composite hypothesis*.

For example, if $f \in \{f_\theta : \theta \in \Theta \subseteq \mathbf{R}\}$, one might have:

- A simple hypothesis such as: $H : \theta = \theta_0$. $\theta \in \Theta$
- A composite hypothesis such as: $H : \theta > \theta_0$. $\Theta_0 \subset \Theta$

In parametric tests, where $\Theta \subseteq \mathbf{R}$, we usually distinguish:

- **One-sided tests:**
  1. $H_0 : \theta = \theta_0 \quad H_1 : \theta > \theta_0$ if $\Theta = [0, \infty)$
  2. $H_0 : \theta \leq \theta_0 \quad H_1 : \theta > \theta_0$ if $\Theta = \mathbf{R}$
  3. $H_0 : \theta \geq \theta_0 \quad H_1 : \theta < \theta_0$ if $\Theta = \mathbf{R}$
- **Two-sided tests:** $H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$

La null hipòtesi NO S'ACCEPTA, només es demostra que no és rebutja o si és rebutja

La que s'accepta com a certa sempre és H_1.

Recorda: que no puguis provar una hipòtesi no en demostra una altra (a.k.a no rebutjar $H_0 \not\Rightarrow H_0$ és certa)

Example 1.10. Z Test

$X_1, \ldots, X_n$ simple random sample of $X \sim N(\mu, \sigma^2)$ with σ^2 known. Let $\overline{X}_n$ be the sample mean.

$H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$

You all might know that, under H_0,

$$Z = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

And for any particular value $\mu_1 > \mu_0$ under H_1

$$Z = \frac{\overline{X}_n - \mu_1}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Step 3: Type I error and α

- **A Type I error is made every time the main hypothesis H_0 is wrongly rejected**, or equivalently, every time H_1 is wrongly accepted.

  Because the hypothesis under test is your main hypothesis, you want to minimize this error as much as possible.
- α **is the probability of committing a Type I error** in the long run, or the probability level at which the main hypothesis will be rejected in favor of the alternative hypothesis.

When setting H_0 you want to control the **Type I error**.

Neyman and Pearson often worked with convenient levels such as $\alpha = 0.05, 0.01$ but smaller values are convenient sometime.

Step 3: Critical region and Critical value

The α level draws a critical region, or rejection region, on the probability distribution of the test under H_0.

Any research value that:

- falls outside this critical region will be taken as reasonably likely under H_0
- falls within the critical region will be taken as most likely under H_1

The α level helps identify the location of the **critical value** of such test.
The **critical value** corresponds to the boundary for deciding between hypotheses.

Step 3: Critical or rejection region R

Definition

- The CRITICAL REGION (OR REJECTION REGION) R associated to a test is a subset of the sample space.
- **CRITERION:** if the observed sample falls within the critical region we reject the null hypothesis in favour of the alternative
- The complement of R is known as the ACCEPTANCE REGION A.
- The rejection region is specified in terms of a TEST STATISTIC, for instance Z, Student's t with $n-1$ degrees of freedom (df), T_{n-1}, χ^2 Chi-square with ν df, χ^2_ν, etc.
- Since the rejection region depends on the level of the test, α, we will often use the notation R_α.

W és regió crítica si

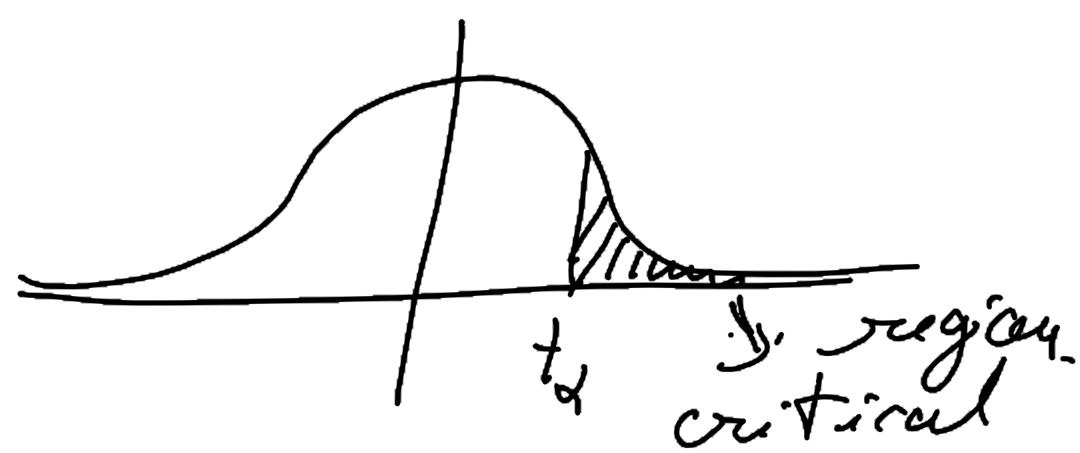
$W = \{x \mid P(x \in R \mid H_0) \leq \alpha\}$ -> is small

$x \in W \Rightarrow$ Reject H_0

$x \notin W \Rightarrow$ Do not reject H_0

$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

$[z > t_\alpha] \equiv x \in W$



Example 1.10. Z Test: $Z = \sqrt{n}(\overline{X}_n - \mu_0)/\sigma \sim N(0,1)$ under H_0

$X_1, \ldots, X_n$ srs of $X \sim N(\mu, \sigma^2)$ with σ^2 known. Let $\overline{X}_n$ be the sample mean.

$H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$

We know that:
$Z = \sqrt{n}(\overline{X}_n - \mu_0)/\sigma \sim N(0,1)$ under H_0.
We will see later, using the Neymann-Pearson lemma, that the rejection region for the most powerful test is given by

$$R_\alpha = \{\mathbf{x} : \overline{x}_n \geq B = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\} = \{\mathbf{x} : \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma} \geq z_\alpha\}$$

It holds
$\alpha = P(R_\alpha | H_0) = P(\overline{X}_n \geq B | H_0) = P(Z = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha)$
having an α level test as we required.

Step 4: Type II error and β

- **A Type II error is made every time the main hypothesis H_0 is wrongly accepted**, thus, every time H_1 is wrongly rejected.
- **β is the probability of committing a Type II error in the long run.**

Making a Type II error is less critical than making a Type I error, but still you want to minimize the probability of making this error once you have decided which α level to use. Hence, you want to make β as small as possible, although not smaller than α.
When setting H_1 you want to control for the **Type II error**.

Neyman and Pearson proposed $\alpha < \beta \leq 0.20$

L'error tipus II és incontrolable, no com el tipus I que es pot fixar, perquè?

Step 4: Errors and Decisions

$H_0 : \theta \in \Theta_0$ *versus* $H_1 : \theta \in \Theta_1, \quad \Theta_0, \Theta_1 \subseteq \mathbf{R}$

R is the Critical or Rejection Region.

The probabilities of making an **error** are:

1. $\alpha = P(\mathbf{X} \in R | H_0 \text{ true}) = \sup_{\theta \in \Theta_0} P_{\theta_0} (\mathbf{X} \in R)$
2. $\beta = P(\mathbf{X} \notin R | H_0 \text{ false}) = \sup_{\theta \in \Theta_1} P_{\theta_1} (\mathbf{X} \notin R)$

The probabilities of making **right decisions** are:

1. $P(\mathbf{X} \notin R | H_0 \text{ true}) = 1 - \alpha$
2. $P(\mathbf{X} \in R | H_0 \text{ false}) = 1 - \beta$

Bàsicament, estem dient casos respecte si x està o no a la regió crítica.

Step 4: Errors and Decisions in a 2×2 table

$H_0 : \theta \in \Theta_0$ *versus* $H_1 : \theta \in \Theta_1, \quad \Theta_0, \Theta_1 \subseteq \mathbf{R}$

R is the Critical or Rejection Region

L'error menys greu

DECISION

	$\mathbf{X} \notin R \rightarrow$ Do not reject H_0	$\mathbf{X} \in R \rightarrow$ Reject H_0
H_0 true	Right decision TP	TYPE I Error TN $\sup_{\theta \in \Theta_0} P_{\theta_0}(\mathbf{X} \in R) = \alpha$
H_0 false	TYPE II Error FP $\sup_{\theta \in \Theta_1} P_{\theta_1}(\mathbf{X} \notin R) = \beta$	Right decision FN

Step 5: Select an optimal test

Neyman-Pearson introduces **the power of a test** (see mathematical definition below) as the probability of taking the right decision when rejecting the null hypothesis.

The optimal test is **the most powerful test** for your research project and you are better off choosing this one.

In general terms: (menos un)

- parametric tests are more powerful than non-parametric tests, (familia)
- one-tailed tests are more powerful than two-tailed tests
- increasing sample size increases power

Difícil de

Fácil de demo

Step 5: Power function, size and level of a test

Power: probability of correctly rejecting the main hypothesis in favor of the alternative hypothesis (i.e.,of correctly accepting H_1).

Definition

Let $X \in \{f_\theta : \theta \in \Theta\}$ and the test hypotheses

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1 \;, \; \Theta_0 \cup \Theta_1 = \Theta, \; \Theta_0 \cap \Theta_1 = \emptyset.$$

The **power function** $\eta(\theta)$ of a test with rejection region R is defined as the function:

$$\eta(\theta) = P_\theta(\mathbf{X} \in R) = \begin{cases} \text{error prob. type I} & \text{if } \theta \in \Theta_0 \\ 1 - \text{error prob. type II} & \text{if } \theta \in \Theta_1 \end{cases}$$

which, for $0 \leq \alpha \leq 1$, has **size** α if $\sup_{\theta \in \Theta_0} \eta(\theta) = \alpha$.
If $\sup_{\theta \in \Theta_0} \eta(\theta) \leq \alpha$ we refer to an α **level** test.

Power of the test: probabilitat de rebutjar

rebutjem

H_0 $\quad \eta(\theta) = P_\theta(x \in W)$

$\theta \in \Theta_0 \to$ rebutjem incorrectament $\to$

$\theta \in \Theta_1 \to$ rebutjem correctament.

$\to$ Prob(type 1 error)

$\to$ 1 - Prob(type 2 error)

Step 5: From "power of a test" to "power function"

- For single hypotheses tests the *power*, $1 - \beta$, is described by a single value because the alternative hypothesis consists of only one value (say θ_1).
- For composite hypotheses tests the *power*, $1 - \beta(\theta)$, is described as *a function of* θ.

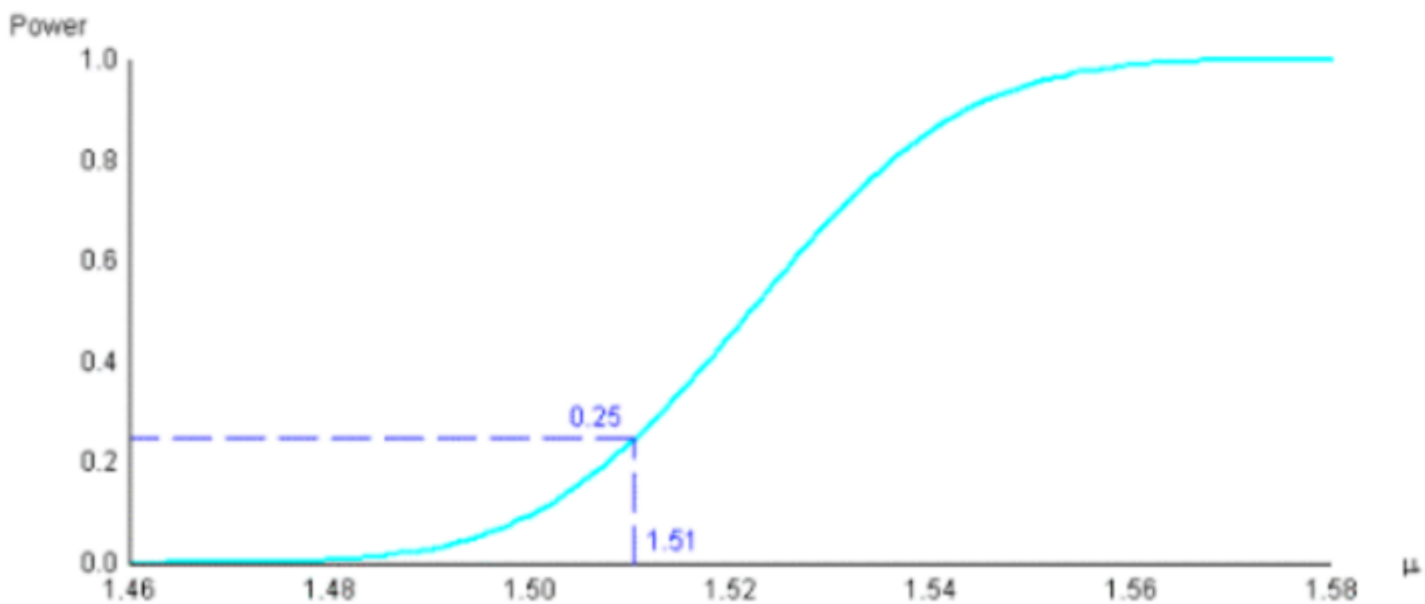
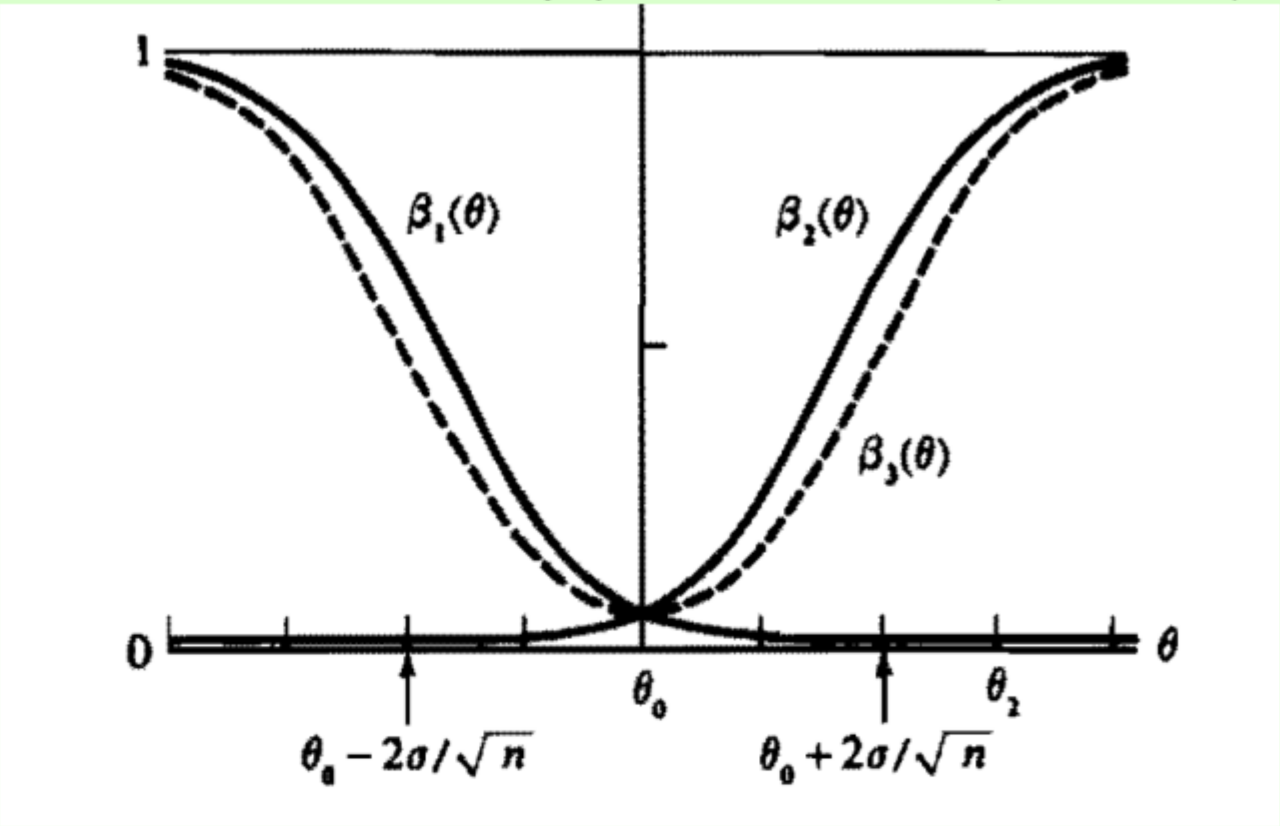


Figure: Power function for $H_0 : \mu \leq 1.5$ vs $H_2 : \mu > 1.5$. Reject H_0 if $\overline{X}_n > 1.5 + \sigma z_\alpha / \sqrt{n}$. For instance $\eta(1.51) = 0.25$

Example 1.10. $H_0 : \mu \geq \mu_0$ (in 1), $H_0 : \mu \leq \mu_0$ (in 2) and $H_0 : \mu = \mu_0$ (in 3)

1. $H_1 : \mu < \mu_0$. Power $= \eta_1(\mu)$. Reject H_0 if $\overline{X}_n < \mu_0 - \sigma z_\alpha/\sqrt{n}$
2. $H_2 : \mu > \mu_0$. Power $= \eta_2(\mu)$. Reject H_0 if $\overline{X}_n > \mu_0 + \sigma z_\alpha/\sqrt{n}$
3. $H_3 : \mu \neq \mu_0$. Power $= \eta_3(\mu)$. Reject H_0 if $|\overline{X}_n - \mu_0| > \sigma z_\alpha/\sqrt{n}$



Casella Berger

Figure: Note that in this plot θ is μ and β is η

✍ Deduce the rejection regions for cases 1 and 3

Step 6: Calculate the sample size n required for optimal power $1 - \beta$

NP's approach works a priori in order to ensure that the **research to be done has good power.**

Given the "operating characteristics" α and $1 - \beta$ for a given effect size, the sample size is calculated.

This computed sample size ensures the a priori power and hence sets the maximum allowed probability of committing a type II error.

It is often taken $0.80 \leq 1 - \beta < 1 - \alpha$

Example 1.10. Needed sample size for Z test

Advanced Statistical Inference

Guadalupe Gómez (UPC) & Alex Sánchez (UB)

Preliminaries

The Frequentist approach

Appendix

Hypothesis testing

Tests of significance: Fisher's approach

Tests of acceptance: Neymann-Pearson's approach

For operating characteristics α (significance level) and $1-\beta(\mu_1)$ (power for $\mu = \mu_1$) the required sample size n is given by:

1. $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$:

$$n = (z_\alpha - z_\eta)^2 \frac{\sigma^2}{(\mu_o - \mu_1)^2}$$

2. $H_0 : \mu = \mu_0$ versus $H_1 : \mu < \mu_0$:

$$n = (z_\alpha - z_\eta)^2 \frac{\sigma^2}{(\mu_o - \mu_1)^2}$$

3. $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$:

$$n = (z_{\alpha/2} - z_\eta)^2 \frac{\sigma^2}{(\mu_o - \mu_1)^2}$$

Add: compute sample size needed to reach this power.

✍ Work these formulae on your own

Step 7: Calculate the critical value of the test

The test T, the significance level α and the sample size n can be used for calculating the critical value of the test; that is, the value to be used as the **cut-off point** for deciding between hypotheses.

In example 1.10 with $H_0 : \mu = \mu_0$ versus $H_1 : \mu < \mu_0$.

The test rejects H_0 if $\overline{X}_n < \mu_0 - \sigma z_\alpha / \sqrt{n} = c_\alpha$.

This value c_α is the critical value of the test

Steps 8 and 9: A POSTERIORI STEPS

Step 8: Calculate the test value for the research question of interest obtained from previous steps

Step 9: Decide in favor of either the main or the alternative hypothesis.

- If the observed result falls within the rejection region, **reject H_0 and accept H_1**
- If the observed result falls outside the rejection region and the test has good power **accept H_0**
- If the observed result falls outside the rejection region and the test has low power **conclude nothing**

In example 1.10 with $H_0 : \mu = \mu_0$ versus $H_1 : \mu < \mu_0$.
The value c_α is the cut-off point: If the sample mean is lower than c_α the null hypothesis should be rejected in favor of the alternative $H_1 : \mu < \mu_0$.

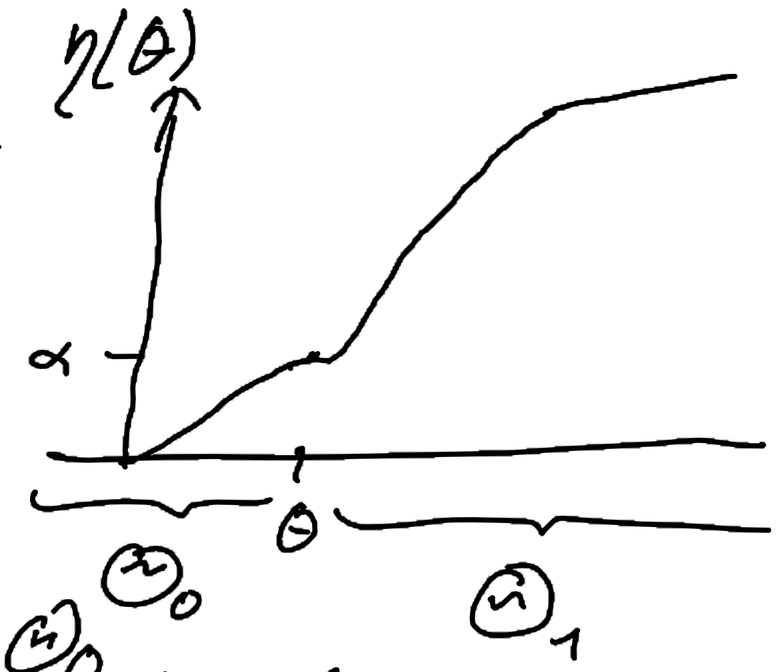
Highlights of Neyman-Pearson's approach

- **More powerful:** NP's approach more powerful than Fisher's for testing data in the long run
- **Better suited for repeated sampling projects** such as industrial quality control or large scale diagnostic testing
- **Deductive:** The approach is deductive and rather mechanical once the a priori steps have been setup
- **Less Flexible than Fisher's approach.** Because most of the work is done a priori, this approach is less flexible for accommodating tests not known to before hand and for doing exploratory research
- **Unique**. Expected size and β taken into account for designing research with good power

Si $H_0: \theta \in \Theta_0$ vs $H_1: \theta \in \Theta_1$

$\Theta_0 \cap \Theta_1 \neq \emptyset$

$\Theta_0 \cup \Theta_1 = \Theta$, podem dir



$\eta(\theta_*) = P(\text{Rej } H_0 | \theta = \theta_*)$

ja que $P(R\ H_0 | H_0 \text{ true}) \Rightarrow \theta_* \in \Theta_0$

$P(R\ H_0 | H_1 \text{ true}) \Rightarrow \theta_* \in \Theta_1$

↳ El poder ens hauria de sortir que és menor que α quan $\theta \in H_0$, que és

A strategy for testing hypotheses

- It is not possible to build tests that *simultaneously* control type I and type II errors.
- Neymann-Pearson's approach suggests to look for the most powerful test between those having a predefined significance level α, or equivalently those tests having level α:

power function

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

Definition

A test that minimizes $\beta = P_\theta(\mathbf{X} \in \overline{R} | H_1)$ (or equivalently maximizes the power $1 - \beta$) between those who have size α is known as the *most powerful test of size* α.

$P(\text{Rej } H_0 | H_0 \text{ true}) = \alpha \overset{*}{=} \alpha(\theta_0)$

$P(\text{Rej } H_1 | H_1 \text{ true}) = \beta$

$P(\text{Accep } H_0 | H_1 \text{ true})$

$P(\text{Rej } H_0 | H_1 \text{ true}) = 1 - \beta \overset{*}{=} \eta(\theta_1)$

* Si $H_0: \theta = \theta_0$ i $H_1: \theta = \theta_1$ → seguiria d'alt

Uniformly Most Powerful Tests

Definition

Consider all level α tests of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$.

A **uniformly most powerful (UMP) level α test** with power function $\eta_{UMP}(\theta)$ is such that:

- $\sup_{\theta \in \Theta_0} \eta_{UMP}(\theta) \leq \alpha$ and
- for any other test of H_0 versus H_1 with power function η^* and level α, we have that $\eta_{UMP}(\theta) \geq \eta^*(\theta)$, for all $\theta \in \Theta_1$.

The following theorems, derived by Neyman and Pearson, can be used to find UMP tests of level α

La gràcia és que no sabem que Θ_1 / Θ_0 els valors que prenen i el f^t va igual.

Neyman-Pearson's lemma (NPL) for simple hypothesis

Theorem

- *Consider testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$*
- *$X_1, \ldots, X_n$ s.r.s of X with joint density function $f(\mathbf{x}|\theta_i)$ and corresponding likelihood $L(\theta_i|\mathbf{x})$ $(i = 0, 1)$*

The most powerful test of size α for H_0 and H_1 is the test whose critical region is defined by a constant $A > 0$ such that:

$$R_\alpha = \left\{ \mathbf{x} \in \mathcal{X}^n : \frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} \geq A \right\}$$

The rejection region of this most powerful (MP) test is called the Optimal Critical region (OCR).

↳ Reject if ratio is bigger que A.

Intuició:

$$\frac{L(\theta_1|x)}{L(\theta_0|x)} \geq A \rightarrow$$ és més probable estar a l'alternativa que la Nul·l

Example 1.10. Z Test: $Z = \sqrt{n}(\overline{X}_n - \mu_0)/\sigma \sim N(0,1)$ under H_0

$X_1, \ldots, X_n$ srs of $X \sim N(\mu, \sigma^2)$ with σ^2 known.
$H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ (with $\mu_1 > \mu_0$)

The likelihood ratio is as:

$$
\begin{aligned}
\frac{L(\mu_1|\mathbf{x})}{L(\mu_0|\mathbf{x})} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\Sigma_{i=1}^n (x_i-\mu_1)^2\right\}}{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\Sigma_{i=1}^n (x_i-\mu_0)^2\right\}} \\
&= \exp\left\{\frac{1}{2\sigma^2} n\left(2\overline{x}_n(\mu_1-\mu_0) + (\mu_0^2-\mu_1^2)\right)\right\} \\
&= \exp\left\{\frac{n}{2\sigma^2}\left(2\overline{x}_n(\mu_1-\mu_0)\right)\right\} \exp\left\{\frac{n}{2\sigma^2}(\mu_0^2-\mu_1^2)\right\} \\
&= \exp\left\{\frac{n\overline{x}_n(\mu_1-\mu_0)}{\sigma^2}\right\} K
\end{aligned}
$$

no té X, exp és constant per la regió crítica

where $K = \exp\left\{\frac{n}{2\sigma^2}(\mu_0^2-\mu_1^2)\right\} > 0$.

Example 1.10. Z Test: Rejection Region

The likelihood ratio is an increasing function of the sufficient statistic $\overline{x}$ because $\mu_1 - \mu_0 > 0$.

By NP the following rejection region yields the UMP test:

$$
\begin{aligned}
R_\alpha &= \left\{ \mathbf{x} \in \mathcal{X}^n : \frac{L(\mu_1|\mathbf{x})}{L(\mu_0|\mathbf{x})} \geq A \right\} \\
&= \{\mathbf{x} : \underline{\exp\left\{ \frac{n\overline{x}_n(\mu_1 - \mu_0)}{\sigma^2} \right\}} K \geq A\} \\
&= \{\mathbf{x} : \overline{x}_n \geq B\}
\end{aligned}
$$

això és creixent ja que $\mu_1 > \mu_0$

where $B = \frac{\sigma^2 \log(A/K)}{n(\mu_1 - \mu_0)}$ and we have used that $\mu_1 > \mu_0$.

si haguéssim estat $\mu_1 < \mu_0$, $\Rightarrow \{x : \overline{x}_n \leq B\}$

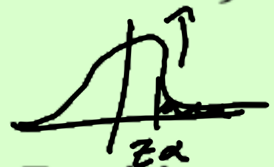
Example 1.10: α level test

A α level test is such that: (hem d'imposar el valor d'α encara)

$\alpha = P(R_\alpha|H_0) = P(\overline{X}_n \geq B|H_0) = P(\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq \frac{B-\mu_0}{\sigma/\sqrt{n}}|H_0).$

Since $\frac{\overline{X}_n-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ under $H_0 \implies B = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$. and

$R_\alpha = \{\mathbf{x} : \overline{x}_n \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\} = \{\mathbf{x} : \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma} \geq z_\alpha\}$ Hem de trobar



Exercise:

Test $H_0 : \mu = \mu_0 = 2$ versus $H_1 : \mu = \mu_1 = 5$ with $\alpha = 0.05$ from a sample of size $n = 10$, if $\sigma = 3$.

We have just seen that the rejection region yielding the UMP test is given by

$$R_{0.05} = \{\mathbf{x} : \overline{x}_{10} \geq 2 + 1.64\frac{3}{\sqrt{10}}\} = \{\mathbf{x} : \overline{x}_{10} \geq 3.55\}$$

Hence, if the sample mean based on the 10 data points is larger than 3.55 you can conclude that $H_1 : \mu = \mu_1 = 5$.

$R_\alpha \to X$ únic, perquè depen de les dades

$P(R_\alpha|H_0) \to X$, ja que ha de ser VA

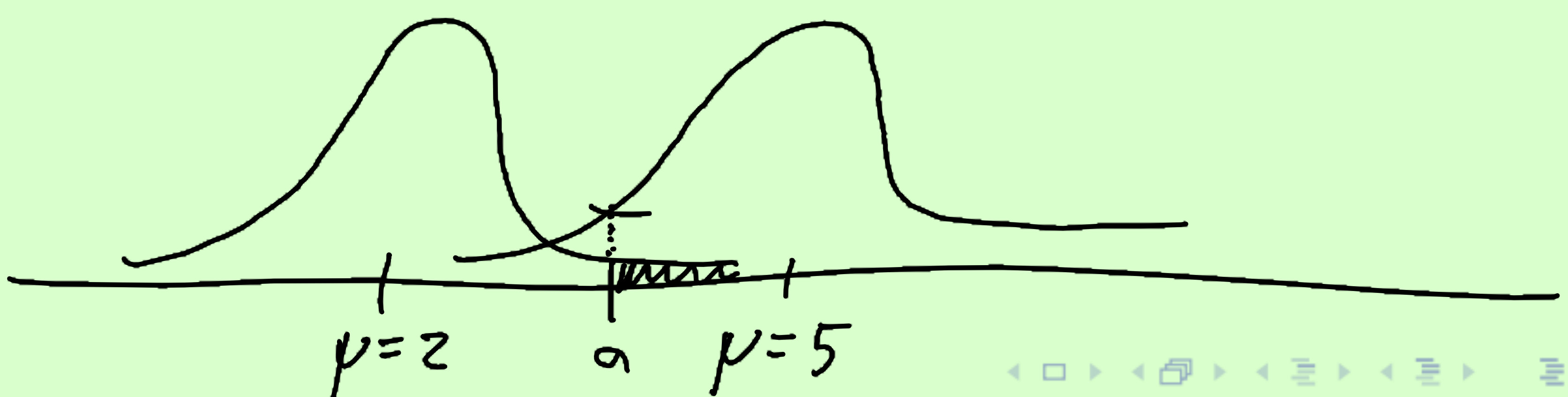
Example 1.10: Power

Exercise:

In this case, the power is

$$\eta = P\{\mathbf{x} : \overline{x}_{10} \geq 3.55 | \mu_1 = 5\} = P\{\mathbf{x} : \frac{\overline{x}_{10} - 5}{\sigma/\sqrt{n}} \geq \frac{3.55 - 5}{3/\sqrt{10}} | \mu_1 =$$

$$\eta = P\{Z \geq -1.52\} = 0.93$$



Example 1.10. Required sample size for specific data

If you want to calculate the required sample size for testing $H_0 : \mu = \mu_0 = 2$ versus $H_1 : \mu = \mu_1 = 5$ with significance level $\alpha = 0.05$ and power $\eta = 0.80$ assuming the data come from a normal distribution with $\sigma = 3$ the formula we have seen before is: → intentar deduir la fòrmula.

1. variability
2. the effect size
3. distància a-valor.

$$n = (z_\alpha - z_\eta)^2 \frac{\sigma^2}{(\mu_o - \mu_1)^2}$$

hence

$$n = (1.645 - (-0.842))^2 \frac{3^2}{(2-5)^2} = 2.487$$

We only need a sample of size 3.

Manera de calcular la sample size a priori per trobar la fòrmula.

Neyman-Pearson Lemma (NPL) for simple versus composite hypotheses

Theorem (NPL for composite hypothesis) Nivell 2

- $H_0 : \theta = \theta_0 \quad H_1 : \theta \in \Theta_1 = \Theta - \{\theta_0\}$.
- *For every $\theta_1 \in \Theta_1$, take the rejection regions of the UMP level α test to test*
  $H_0 : \theta = \theta_0$ *versus* $H_1 : \theta = \theta_1$ *provided by the Neyman-Pearson Lemma, that is,*

  no ha de dependre de Θ_1.

$$R(\theta_1) = \left\{ \mathbf{x} : \frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} \geq A(\theta_1) \right\} .$$

- *If those regions do not depend on θ_1, that is, if $R(\theta_1) = R$ for all $\theta_1 \in \Theta_1$, then the statistical test with rejection region R is the Uniformly Most Powerful (UMP) test of size α for these hypotheses.*

Example 1.10. Rejection region for $H_0 : \mu = 2$ vs $H_1 : \mu > 2$

Still assuming that $\sigma = 3$ and $n = 10$, suppose you want to test $H_0 : \mu = 2$ versus $H_1 : \mu > 2$ with significance level $\alpha = 0.05$.
The rejection region yielding the UMP test for H_0 versus H_1 was given by

$$R_{0.05}(\mu_1 = 5) = \{\mathbf{x} : \overline{x}_{10} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\} = \{\mathbf{x} : \overline{x}_{10} \geq 3.55\}$$

Observe that $R_{0.05}(\mu_1 = 5) = R_{0.05}(\mu_1)$ for any $\mu_1 > 2$ does not depend on the specific value $\mu = \mu_1 = 5$; in the previous developments we only took into account that $\mu_1 > \mu_0$.
Hence, by NPL Theorem, the statistical test with rejection region $R = \{\mathbf{x} : \overline{x}_{10} \geq 3.55\}$ is the UMP test of size 0.05 for $H_0 : \mu = 2$ vs $H_1 : \mu > 2$.
If the sample mean based on the 10 data points is larger than 3.55 you can conclude that $H_1 : \mu > 2$.

One-Sided UMP Tests via the Neyman-Pearson Lemma

Theorem (NPL for one-sided test) Nivell 3!

AIM: *To test* $H_0 : \theta \leq \theta_0$ *versus* $H_1 : \theta > \theta_0$

Consider the UMP test for $H_0^* : \theta = \theta_0$ *vs.* $H_1^* : \theta = \theta_1$ *(*$\theta_1 \in \Theta_1 = (\theta_0, \infty), \theta_1 > \theta_0$*), rejecting* $H_0^* : \theta = \theta_0$ *if* $\frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} > A(\theta_1)$ *for some* $A(\theta_1) \geq 0$.

If $A(\theta_1) = A$ ***for all*** $\theta_1 \in \Theta_1$*, that is, it does not depend on the value* $\theta_1 \in \Theta_1$ *and* $\sup_{\theta \in \Theta_0} \eta(\theta) = \eta(\theta_0)$*, the same test is also the UMP level* α *test for*

$H_0 : \theta \leq \theta_0$ *versus* $H_1 : \theta > \theta_0$

Analogously for $H_0 : \theta \geq \theta_0$ *versus* $H_1 : \theta < \theta_0$ with $\Theta_0 = [\theta_0, \infty)$ and $\Theta_1 = (-\infty, \theta_0)$

Example 1.10: UMP test for $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$.

$X_1, \ldots, X_n \sim X \sim N(\mu, \sigma^2)$, σ^2 known. Consider the test $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$.
We have seen that

$$R_\alpha(\mu_1) = \{\mathbf{x} : \overline{x_n} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} = R_\alpha\}$$

is the same for all $\mu_1 \in \Theta_1 = (\mu_0, \infty)$.
Hence, this test is also UMP of size α to test $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. The power function of this test is an increasing function of μ:

* Programar en R com a operació

$$\begin{aligned}\eta(\mu) &= P(\overline{X}_n \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} | \mu) = P\left(Z \geq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right) \\ &= 1 - P\left(Z \leq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right) = 1 - \phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right)\end{aligned}$$

Neyman-Pearson and sufficient statistic

Corollary

Consider the hypotheses: $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. ***If*** *there exists a test procedure based on a sufficient statistic T, with density $g(t|\theta)$, with rejection region R such that:*

1. *The test has level α,*
2. *There exists $\theta_0 \in \Theta_0$ such that $P_{\theta_0}(T \in R) = \alpha$,*
3. *Given $\theta_0 \in \Theta_0$, if for every $\theta_1 \in \Theta_1$ there exists $A(\theta_1) \geq 0$ such that for $t = T(\mathbf{x})$:*

$$\frac{g(t|\theta_1)}{g(t|\theta_0)} > A(\theta_1) \Longrightarrow t \in R, \text{ and, } \frac{g(t|\theta_1)}{g(t|\theta_0)} < A(\theta_1) \Longrightarrow t \in \overline{R}$$

then *this test is UMP of level α to test H_0 vs H_1.*

Corolari: si tens famílies exponencials, agafes el SSt i mires si rebutja o no.

(ull amb les discretes, que $P_{\theta_0}(T \in R) = \alpha$

One-Sided UMP Tests for Exponential Families

The last corollary can be applied to densities from the exponential family. Aquest és el test més freqüent.

- $X_1, \ldots, X_n$ sample with a joint density $f(\mathbf{x}|\theta)$
- $f(\mathbf{x}|\theta)$ from a one-parameter exponential family: $f(\mathbf{x}|\theta) = h(\mathbf{x})c(\theta)\exp\{w(\theta)T(\mathbf{x})\}$
- $w(\theta)$ **is increasing**
- $T(\mathbf{x})$ is the complete sufficient statistic

1. For $\theta_1 > \theta_0$.
   The test that rejects H_0 if $T(\mathbf{x}) > A$ where A is such that $\text{Prob}_{\theta_0}\{T(\mathbf{X}) > A\} = \alpha$ is the UMP level α test in the following 3 situations:
   1. $H_0 : \theta = \theta_0$ *versus* $H_1 : \theta = \theta_1$
   2. $H_0 : \theta = \theta_0$ *versus* $H_1 : \theta > \theta_0$
   3. $H_0 : \theta \le \theta_0$ *versus* $H_1 : \theta > \theta_0$

One-Sided UMP Tests for Exponential Families

Analogously

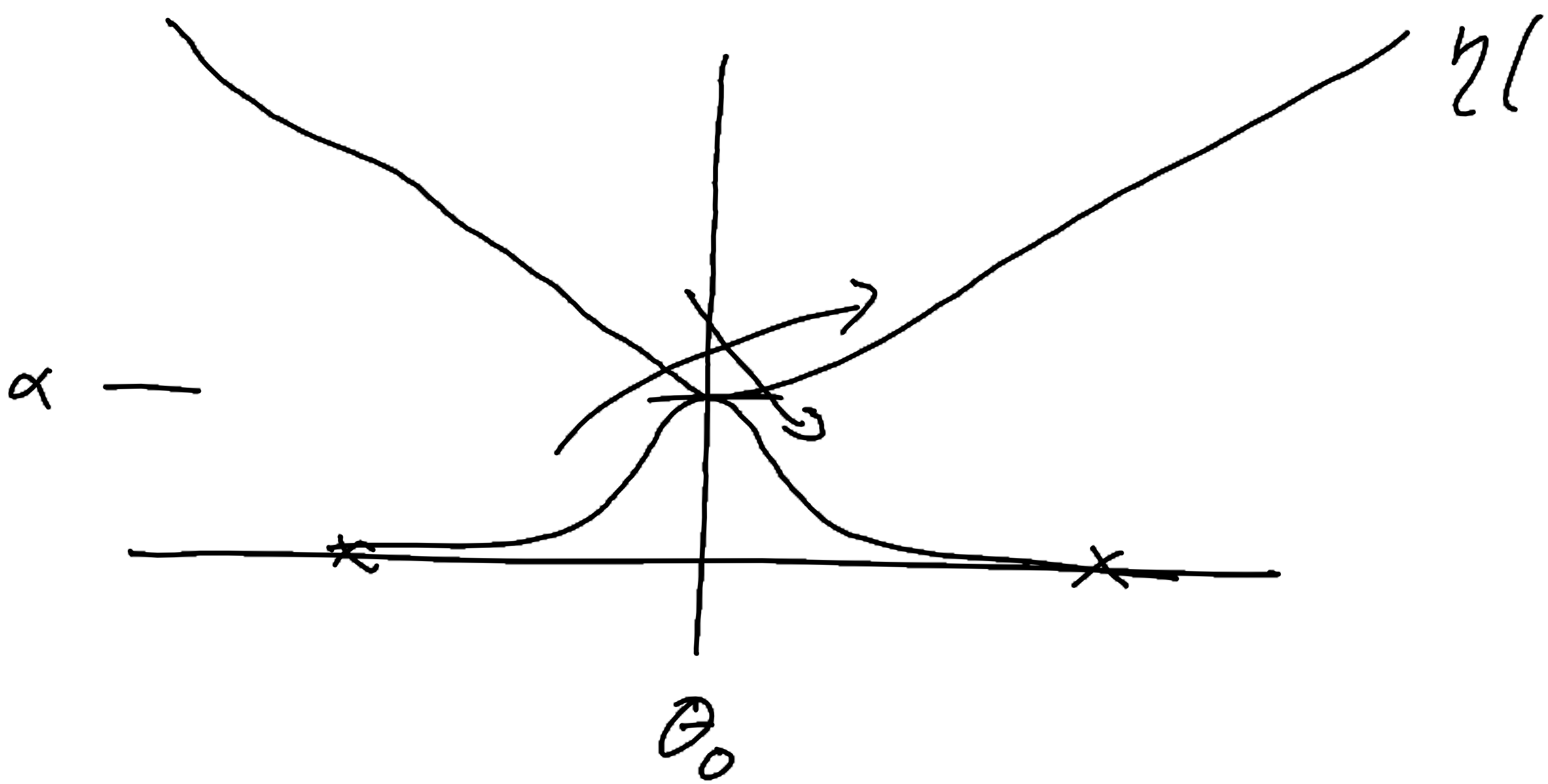
2. For $\theta_1 < \theta_0$.
   The test that rejects H_0 if $T(\mathbf{x}) < A$ where A is such that $\text{Prob}_{\theta_0}\{T(\mathbf{X}) < A\} = \alpha$ is the UMP level α test in the following 3 situations:
   1. $H_0 : \theta = \theta_0$ *versus* $H_1 : \theta = \theta_1$
   2. $H_0 : \theta = \theta_0$ *versus* $H_1 : \theta < \theta_0$
   3. $H_0 : \theta \geq \theta_0$ *versus* $H_1 : \theta < \theta_0$

✍ The proof follows from Neyman-Pearson and sufficient statistic corollary
✍ Why do you need $w(\theta)$ to be increasing in both results
✍ Apply this result to derive UMP tests for samples from $N(\mu, \sigma^2)$ with σ^2 known, Poisson, Geometric, Exponential

Unbiased tests

A hypothesis test for $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ of size α and with power function $\eta(\theta)$ is an **unbiased test** if $\eta(\theta) \geq \alpha$ for all $\theta \in \Theta_1$.

Unbiasedness is a reasonable requirement (recall Step 4 in NP theory):
The probability of (correctly) accepting H_1 when H_0 is false is always bigger than the probability of (erroneously) rejecting H_0 when it is true.



Example 1.10: $\mathbf{X} \sim N(\mu, \sigma^2)$, $H_0 : \mu = \mu_0$ vs $H_3 : \mu \neq \mu_0$.

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with σ^2 known.
Test: $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

The critical region of the UMP test for H_0 vs. $H_1' : \mu < \mu_0$ is $C_1 = \{\mathbf{x} : \overline{x}_n \leq A_1\}$.
The critical region of the UMP test H_0 vs. $H_1'' : \mu > \mu_0$, is $C_2 = \{\mathbf{x} : \overline{x}_n \geq A_2\}$.

To test H_0 vs H_1 we suggest a critical region given by:

$$C = \{\mathbf{x} : \overline{x}_n \leq A_1 \text{ or } \overline{x}_n \geq A_2\}, \quad \text{such that}$$

$$P(\overline{X}_n \leq A_1 | \mu = \mu_0) + P(\overline{X}_n \geq A_2 | \mu = \mu_0) = \alpha.$$

For example we could take

$$P(\overline{X}_n \leq A_1 | \mu = \mu_0) = P(\overline{X}_n \geq A_2 | \mu = \mu_0) = \frac{\alpha}{2},$$

and then, $A_1 = \mu_0 - z_{\alpha/2}\sigma/\sqrt{n}$ and $A_2 = \mu_0 + z_{\alpha/2}\sigma/\sqrt{n}$.

Example 1.10

The test that rejects $H_0 : \mu = \mu_0$ if $|\overline{X}_n - \mu_0| \geq z_{\alpha/2}\sigma/\sqrt{n}$, when the family of probabilities is normal and σ^2 is known, is unbiased and is UMP in the class of unbiased tests.

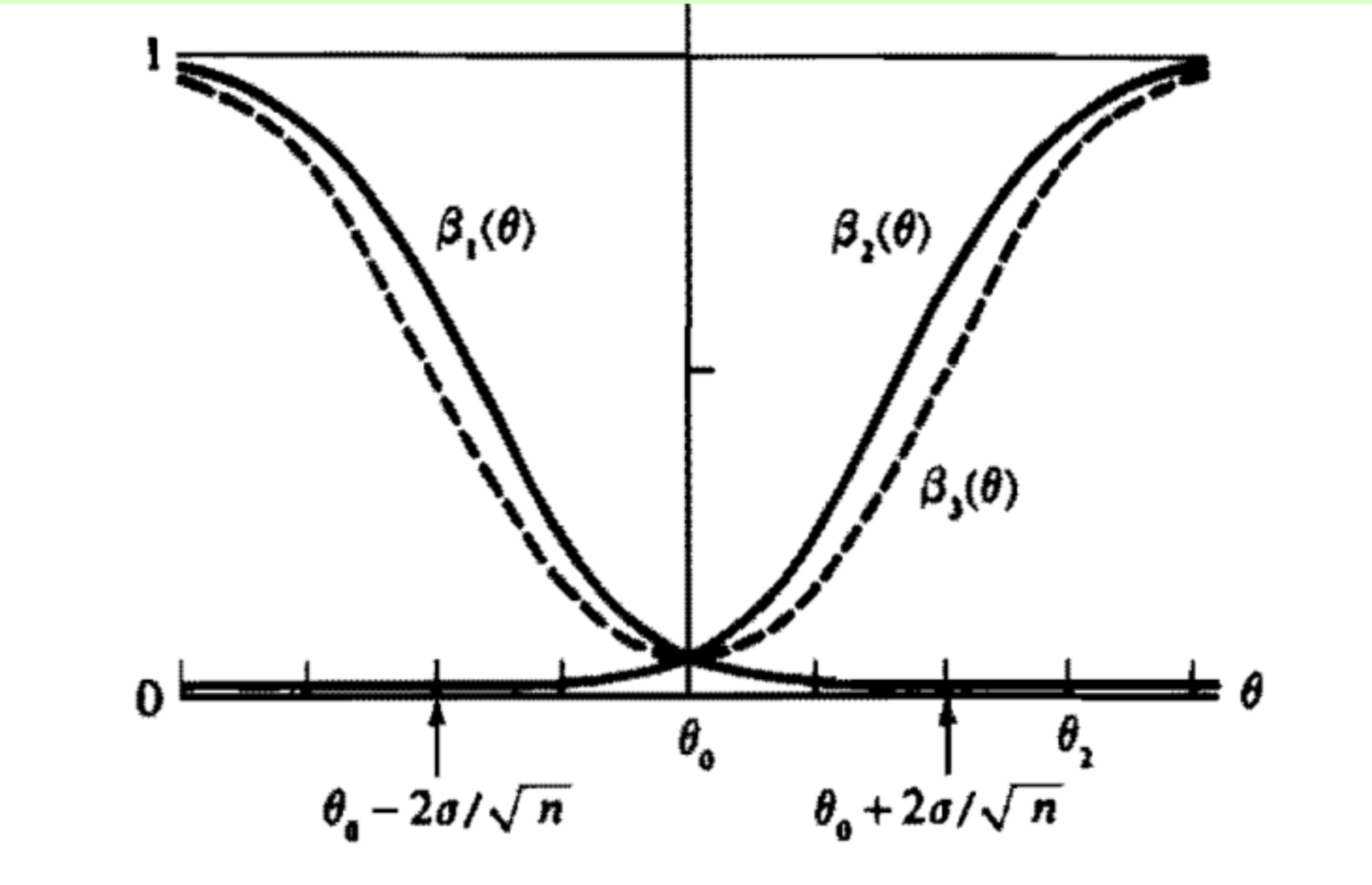


Figure: $\beta_3(\theta)$ corresponds to the power of this test and $\theta = \mu$

Searching for UMP tests

- Up to here, the search for UMP tests has been based on the Neymann-Pearson lemma that provides an optimal solution under certain conditions:
  - If the two hypotheses H_0 and H_1 are simple ones.
  - With one-sided composite hypothesis if the α level can be computed at the boundary of the H_0 parameter space and the optimal critical region does not depend on the value of the parameter of the H_1 parameter space.
- When none of these conditions holds, the existence of UMP tests is not guaranteed.
- **What do we do?** Go for the second best: **Likelihood Ratio Tests**. Explain in Unit 3

Permutation tests

Permutation tests, exact tests or randomization tests are nonparametric tests, light on assumptions, widely applicable, and very intuitive.

- Most of the theory until now for testing a null hypothesis, H_0, needs the sampling distribution of the corresponding test statistic under H_0. When this is not feasible a permutation test gives a simple way to compute the sampling distribution, under the strong null hypothesis of no effect on the outcome.
- To estimate the sampling distribution we generate samples under H_0 shuffling the data set we have collected (resampling the observed data). The ranking of the real test statistic among the shuffled test statistics gives the p-value.

Following presentation by Jared Wilber

We follow now Jared Wilber's presentation on the permutation tests by means of an Alpaca Shepherd (pastor de ovejas de alpaca) story.
The test statistic here used is the difference of means.

Let's watch it now!!
`https://www.jwilber.me/permutationtest/`

Permutation algorithm summary

1. Determine the initial test-statistic T you want to use
2. Calculate the initial test-statistic for the observed data, called it t_{obs}
3. Run simulations resampling the labels of the observed data and calculate the test-statistics for each new dataset, $t_1, t_2, \cdots, t_m$, for m shuffled samples
4. Build the approximate sampling distribution of the test-statistic from the values $t_1, t_2, \cdots, t_m$ obtained in step 3. It can be done empirically and it can also be approximated by a normal distribution.
5. Calculate the p-value as the $\mathrm{Prob}\{T > t_{obs}\}$.

An important limitation of this approach is that it assumes that observations are exchangeable under H_0, and as a consequence that both groups have equal variances, as it does the two-sample Student t test.

Other permutation test. Self readings

- Permutation tests by Ken Rice and Thomas Lumley with R commands and ouputs. See `https://faculty.washington.edu/kenrice/sisg/SISG-08-06.pdf`
- Permutation tests by Thomas J. Leeper in html also with R commands `https://thomasleeper.com/Rcourse/Tutorials/permutationtests.html`

Outline

Appendix A

Next slides provide extra information and details of all the concepts developed previously. All of them are relevant, but some of them are required for you to learn, for instance:

- Further discussion of Fisher and Neyman-Pearson
- Neyman-Pearson minimum effect size
- Proof NPL for composite hypothesis
- Monotone likelihood ratio property
- Karlin-Rubin theorem

Appendix B

Common tests are also summarized in this appendix.

Most of the common tests, we assume you are familiar with,

- Z-test for one sample and for difference of means
- t-test for one sample and for difference of means
- χ^2-test and F-test

APPENDIX A

A bit more on *Fisher, Neyman-Pearson or NHST?*

SUMMARY of ABSTRACT of *Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing* by J.D. Perezgonzalez.

Despite frequent calls for the overhaul of **null hypothesis significance testing (NHST)**, this controversial procedure **remains ubiquitous** in behavioral, social and biomedical teaching and research.
This paper presents a tutorial for the teaching of data testing procedures, often referred to as **test of hypothesis testing theories**. The first procedure introduced is Fisher's approach to data testing (1925) known as **tests of significance**. The second is Neyman-Pearson's approach (1928) known as **tests of acceptance**. The final procedure is the **incongruent combination** of the previous two theories into the current approach known as **NSHT** (2004) and **will not be covered**.

NOT COVERED: Bayes's hypotheses testing (Lindley,1965) and Wald's (1950) decision theory

Fisher's step 4: Extra reading

- **Directional and non-directional hypotheses.**
- **One-tailed and two-tailed tests.**
- **Correction of level of significance for multiple tests**.
  If multiple tests are performed, the probability of finding statistical significant results which are due to mere chance variation increases.
  To keep such probability at acceptable levels overall, the level of significance may be corrected downwards
- **Bonferroni's correction** reduces the level of significance proportionally to the number of tests carried out. **Remark:** Is too conservative.
  Other methods:
  - Familywise error rate metods
  - False discovery rate
  - Resampling methods (bootstrap)
  - Permutaiton test

Kareem Carr (PhD student at Harvard, MA (Math), MSc(Biostats) on twitter (self reading)

- Don't know what a P-VALUE is?
- Don't know why P-VALUES work?
- Don't know why sometimes P-VALUES don't work?

DEFINITION OF A P-VALUE. Assume your theory is false. The P-VALUE is the probability of getting an outcome as extreme or even more extreme than what you got in your experiment.

THE LOGIC OF THE P-VALUE. Assume my theory is false. The probability of getting extreme results should be very small but I got an extreme result in my experiment. Therefore, I conclude that this is strong evidence that my theory is true. That's the logic of the p-value.

THE P-VALUE IS REASONABLE IN THEORY BUT TRICKY IN PRACTICE. Follow the thread on `https://twitter.com/kareem_carr/status/1312783404975493122?s=09`

NP Step 1: MES: Minimum effect size

We work with distributions of samples, which have narrower standard errors than population distributions.

MES represents that part of the main hypothesis that is not going to be rejected by the test. MES captures values of no research interest which you want to leave under H_0

The minimum effect size might be given, for instance, by $\mu_1 - \mu_2 = d = 0.8$.

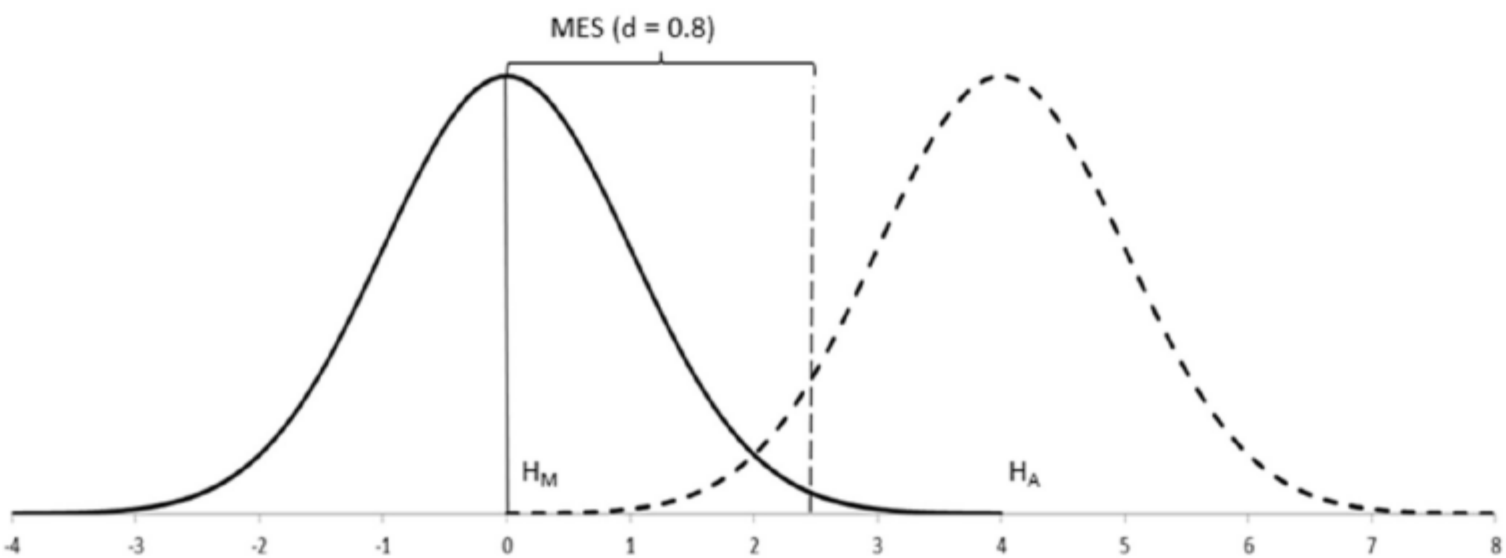


Figure: Sampling distr. ($N = 50$ each) of two Normal distr.

NP Step 3: Rejection region, α, β

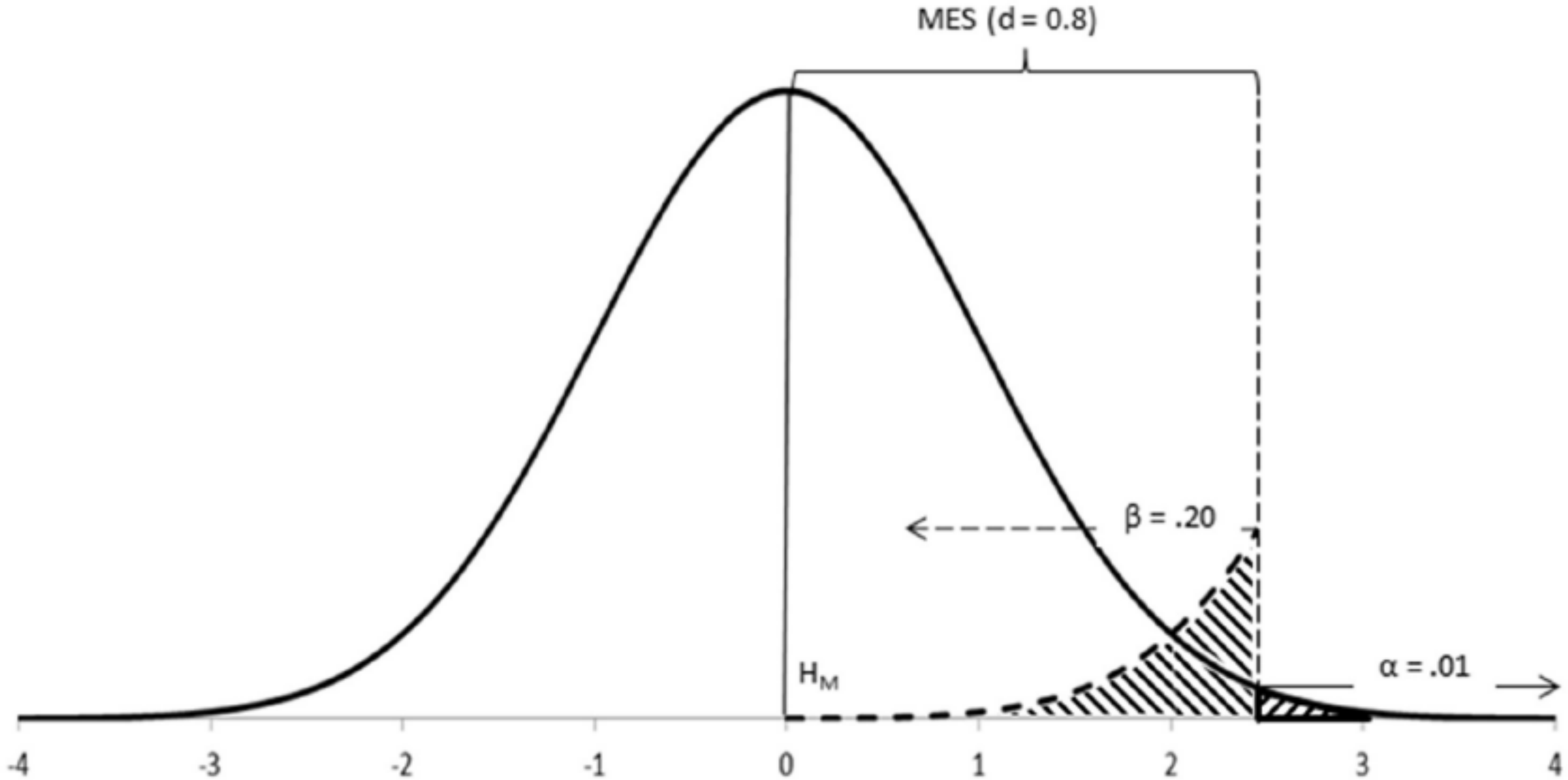


Figure: NP's approach tests data under H_0 using the rejection region delimited by α. H_1 contributes MES and β. Differences of research interest will be $\geq$ MES and will fall within this rejection region.

Example 1.10. p-value (self-work)

The p–value for an observed sample $\mathbf{x}_{obs}$ is:

$$p(\mathbf{x}_{obs}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x}_{obs})).$$

where $W(\mathbf{X})$ is generic for a test based on $\mathbf{X}$
In the example of the Z test having observed $\overline{x}_n$:

$$p(\mathbf{x}) = P(\overline{X}_n \geq \overline{x}_n | \mu = \mu_0) = P\left(Z \geq \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}\right)$$

where Φ is the distribution of $Z \sim N(0,1)$.
For $\sigma = 1$, $\mu_0 = 5$, $n = 4$ and $\overline{x} = 5.2$ we have

$$p(\mathbf{x}) = P\left(\frac{\overline{X}_4 - 5}{1/\sqrt{4}} \geq \frac{5.2 - 5}{1/\sqrt{4}}\right) = P(Z \geq 0.4) = 0.3446;$$

not sufficient evidence against H_0 to reject $H_0 : \mu = \mu_0$.

Proof NPL for composite hypothesis (self work)

- Let $\eta(\theta)$ be the power function of the test with critical region C.
- Because of the definition of C this test has level α.
- Let η^* be the power function of any other test of level α.
- For any $\theta_1 \in \Theta_1$, from the Neyman-Pearson's Lemma, the test with critical region C is the best to test

$$H_0 : \theta = \theta_0 \quad H_1 : \theta = \theta_1$$

and then, $\eta(\theta_1) \geq \eta^*(\theta_1)$.
- As this happens for all $\theta_1 \in \Theta_1$ thus,

$$\eta(\theta_1) \geq \eta^*(\theta_1)$$

for all $\theta_1 \in \Theta_1$, so that the test given by C is UMP level α.

Searching for UMP tests

The problem of finding UMP tests for more general situations than those provided by the NP lemmas can be faced in different forms:

- By restricting the probability families considered:
  - Restrict to families with *monotone likelihood ratio* (MLR).
  - Restrict to the exponential family of distributions (which also has MLR).
- By restricting the class of test that are built:
  - Considering only *unbiased tests*
  - Considering only *locally most powerful* tests (we are not covering this)

Monotone likelihood ratio property

Definition

- Let T be a univariate random variable T.
- A familiy of probability density functions (pdfs) or probability mass functions (pmfs) $\{g(t|\theta) : \theta \in \Theta\}$ is said to have *monotone likelihood ratio* (MLR) if, for every $\theta_2 > \theta_1$ the ratio

$$g(t|\theta_2)/g(t|\theta_1)$$

is a nondecreasing function of t for the values of t such that

$$g(t|\theta_2) > 0 \text{ and } g(t|\theta_1) > 0.$$

Examples of Monotone likelihood ratio (MLR)

- The Poisson distribution has MLR.
- The exponential familiy has MLR.
- The Cauchy distribution has not MLR.

✍ Prove these results

The Karlin–Rubin theorem

Theorem (Karlin-Rubin)

Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.
Suppose that T is a sufficient statistic for θ and the family of pdfs or pmfs $\{g(t|\theta) : \theta \in \Theta\}$ of T has MLR.
Then for any t_0 the test that rejects H_0 if and only if $T > t_0$ is a UMP level α test where $\alpha = P_{\theta_0}(T > t_0)$.

- The Karlin-Rubin Theorem (1956) states that if we set $\alpha = P(t > t_0)$ and reject H_0 for an observed $t > t_0$ (t_0 a known constant), then this test is the most powerful relative to any other possible test of H_0 with this α level.

APPENDIX B

Common tests

Most of the common tests, we assume you are familiar with:

- Z-test for one sample and for difference of means
- t-test for one sample and for difference of means
- χ^2-test
- F-test

are derived from the Neyman-Pearson lemma or as likelihood ratio tests (in Unit 3) .

In general,

- one-sided UMP Tests are derived following Neyman-Pearson lemma,
- while two-sided unbiased UMP tests rely on the Likelihood ratio approach.

YOU NEED TO KNOW ALL THESE TESTS

Z test

Let $X_1, X_2, \ldots, X_n$ be a random sample from $X \sim N(\mu, \sigma)$, and assume σ^2 is known.
Consider $H_0: \ \mu = \mu_0$ versus $H_1: \ \mu \neq \mu_0$

The test that rejects H_0 if $|Z| > z_\alpha$, where z_α is the $1 - \alpha$ percentile of a $N(0, 1)$ and

$$Z = \frac{\overline{X}_n - \mu_0}{\sqrt{\sigma^2 / n}} \sim N(0, 1)$$

is an unbiased UMP test of size α.

✍ Apply the likelihood ratio approach to derive this result, after Unit 3

t test

Let $X_1, X_2, \ldots, X_n$ be a random sample from $X \sim N(\mu, \sigma)$, and assume σ^2 is unknown.
Consider $H_0: \ \mu = \mu_0$ versus $H_1: \ \mu \neq \mu_0$
Define

$$T = \frac{\overline{X}_n - \mu_0}{S/\sqrt{n}}$$

with $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-1}$ unbiased estimator of 'σ^2.
Under $H_0: \ \mu = \mu_0$ we have $T \sim t_{n-1}$

The test that rejects H_0 if $|T| > t(\alpha, n-1)$, where $t(\alpha, n-1)$ is the $1-\alpha$ percentile of a Student t with $n-1$ degrees of freedom is an unbiased UMP test of size α.

✍ Apply the likelihood ratio approach to derive this result, after Unit 3

χ^2 test for one-sample variance

Let $X_1, X_2, \ldots, X_n$ be a random sample from $X \sim N(\mu, \sigma)$, and assume σ^2 is unknown.
Consider $H_0: \ \sigma^2 = \sigma_0^2$ versus $H_1: \ \sigma^2 \neq \sigma_0^2$

Define for $S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{n-1}$

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

Under $H_0: \ \sigma^2 = \sigma_0^2$ we have $\chi^2 \sim \chi^2(n-1)$

The test that rejects H_0 if

$$\chi^2 > \chi_{(n-1),\alpha/2} \text{ or } \chi^2 < \chi_{(n-1),1-\alpha/2},$$

where $\chi_{(n-1),\alpha/2}$ is the $1-\alpha$ percentile of a $\chi^2(n-1)$ with $n-1$ degrees of freedom is an unbiased UMP test of size α.

✍ Apply the likelihood ratio approach to derive this result,

Test F for two-sample variances

Let **X** and **Y** two independent random samples, sizes n_1 and n_2, from $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, respectively.
Consider $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$

Define $S_1^2 = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X}_{n_1})^2}{n_1 - 1}$ and $S_2^2 = \frac{\sum_{i=1}^{n_2}(Y_i - \overline{Y}_{n_2})^2}{n_2 - 1}$.

$$\text{Define} \quad F = \frac{S_1^2}{S_2^2}$$

Under $H_0 : \sigma_1^2 = \sigma_2^2$ we have $F \sim F(n_1 - 1, n_2 - 1)$

The test that rejects H_0 if $F > F_{n_1, n_2, \alpha/2}$ o $F < F_{n_1, n_2, 1-\alpha/2}$, where $F_{n_1, n_2, \alpha/2}$ is the $1 - \alpha$ percentile of a Fisher F with n_1, n_2 degrees of freedom is an unbiased UMP test of size α.

✍ Apply the likelihood ratio approach to derive this result, after Unit 3.

Test Z for difference of means

Let **X** and **Y** two independent random samples, sizes n_1 and n_2, from $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, respectively, with known σ_1^2, σ_2^2.
Consider $H_0: \ \mu_1 = \mu_2$ versus $H_1: \ \mu_1 \neq \mu_2$.
Define

$$Z = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

Under $H_0: \ \mu_1 = \mu_2$ we have $Z \sim N(0, 1)$.

The test that rejects H_0 if $|Z| > z_\alpha$ is an unbiased UMP test of size α.

Test t for difference of means

Let **X** and **Y** two independent random samples, sizes n_1 and n_2, from $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, respectively, with unknown σ_1^2, σ_2^2 but $\sigma_1^2 = \sigma_2^2$.
Consider $H_0 : \; \mu_1 = \mu_2$ versus $H_1 : \; \mu_1 \neq \mu_2$.
Define

$$T = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\dfrac{(n_1 S_1^2 + n_2 S_2^2)\,(n_1 + n_2)}{(n_1 + n_2 - 2)(n_1\, n_2)}}}$$

Under $H_0 : \; \mu_1 = \mu_2$ we have $T \sim t(n_1 + n_2 - 2)$

The test that rejects H_0 if $|T| > t(\alpha, n_1 + n_2 - 2)$ is an unbiased UMP test of size α.
Remark: If $\sigma_1^2 \neq \sigma_2^2$ this test is not the right solution. The different solutions are known as the Behrens-Fisher problem

Outline

Interval estimators

Definition

An **interval estimator** $[L(\mathbf{X}), U(\mathbf{X})]$ of a parameter $\theta \in \Theta \subseteq \mathbf{R}$ is formed by any pair of real functions $L(\mathbf{x})$ and $U(\mathbf{x})$ defined on the sample space $\mathcal{X}$ such that $L(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}$.

If the value $\mathbf{X} = \mathbf{x}$ is observed, we can calculate an **interval estimate** of θ, $[L(\mathbf{x}), U(\mathbf{x})]$ allowing to infer that "$L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$" .

Example 1.11

Let X_1, X_2, X_3, X_4 be a sample of size 4 of $X \sim N(\mu, 1)$. An interval estimator of μ is $[\overline{X}_4 - 1, \overline{X}_4 + 1]$. For every observed sample x_1, x_2, x_3, x_4, the interval estimate of μ is $[\overline{x}_4 - 1, \overline{x}_4 + 1]$.

Precision versus Confidence

If we estimate θ based on an interval, the inference is **less precise** than if we estimate it with a point estimator:

- We NOW state that $\theta \in [L(\mathbf{x}), U(\mathbf{x})]$, while before, we gave a specific value $\hat{\theta}$ as its estimation.
- We lose precision. Which benefit do we have?
- We gain **CONFIDENCE**

Example 1.11: $X_1, X_2, X_3, X_4 \sim N(\mu, 1)$.

If we estimate μ by $\overline{X}_4$, we have that $P_\mu(\overline{X}_4 = \mu) = 0$, because $\overline{X}_4 \sim N(\mu, 1/4)$. However,

$$P_\mu(\mu \in [\overline{X}_4 - 1, \overline{X} + 1]) = P_\mu(\overline{X}_4 - 1 \leq \mu \leq \overline{X}_4 + 1) =$$

$$P_\mu(-1 \leq \overline{X}_4 - \mu \leq 1) = P_\mu\left(-2 \leq \frac{\overline{X}_4 - \mu}{1/\sqrt{4}} \leq 2\right) = 0.954.$$

where the probability is taken w.r.t. the **true value of μ.**

Thus, we have over a 95% chance of **covering the unknown parameter** with our interval estimator.
The purpose of using an interval estimator rather than a point estimator is to have some guarantee of capturing the value of the parameter of interest.

Coverage probability

For an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ the **Coverage probability of** $[L(\mathbf{X}), U(\mathbf{X})]$ is the probability (empirical proportion) that the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ contains the true value of the parameter of interest θ , that is

$$\text{coverage}(\theta) = P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})]).$$

Important to keep in mind that the interval is the random quantity, not the parameter and these probability statements refer to $\mathbf{X}$, not θ.

The coverage probability can be a constant function of θ (often for continuous distributions) or can change with θ (for discrete distributions).

Confidence interval

For an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ the **Confidence level** of $[L(\mathbf{X}), U(\mathbf{X})]$ is the infimum of the coverage probabilities, that is

$$\text{confidence level} = \inf_{\theta \in \Theta} P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})]).$$

An interval estimator together with a confidence level is known as a **confidence interval**. We can use the notation $CI_{1-\alpha}(\theta)$ when referring to a $(1-\alpha)$ confidence interval for θ.

Since we do not know the true value of θ, we can only guarantee a coverage probability equal to the infimum, the confidence level. If the coverage probability is a constant function of θ, the infimum coincides with all the values. In other cases, however, the coverage probability can be a fairly variable function of θ.

Example 1.11: $X_1, X_2, X_3, X_4 \sim N(\mu, 1)$

We have seen that

$$P_\mu(\mu \in [\overline{X}_4 - 1, \overline{X}_4 + 1]) = P_\mu\left(-2 \leq \frac{\overline{X}_4 - \mu}{1/\sqrt{4}} \leq 2\right) = 0.954.$$

Note that this probability **does not depend on the μ value** and hence the confidence level of $[\overline{X}_4 - 1, \overline{X}_4 + 1]$ is

$$\inf_{\mu \in R} P_\mu(\mu \in [\overline{X}_4 - 1, \overline{X}_4 + 1]) = 0.954$$

and $[\overline{X}_4 - 1, \overline{X}_4 + 1]$ is a 95.4% confidence interval for μ.

Example 1.11: $X_1, X_2, X_3, X_4 \sim N(\mu, 1)$

If the true value of μ is μ_0, then $\overline{X}_4 \sim N(\mu_0, 1/4)$. We could compute the probability for other values μ_1 of μ:

$$
\begin{aligned}
& P_{\mu_1}(\mu_0 \in [\overline{X}_4 - 1, \overline{X}_4 + 1]) = P_{\mu_1}(\overline{X}_4 - 1 \leq \mu_0 \leq \overline{X}_4 + 1) \\
= \quad & P_{\mu_1}(\mu_0 - 1 \leq \overline{X}_4 \leq 1 + \mu_0) \\
= \quad & P_{\mu_1}\left(\mu_0 - \mu_1 - 1 \leq \overline{X}_4 - \mu_1 \leq \mu_0 - \mu_1 + 1\right) =
\end{aligned}
$$

$$
P_{\mu_1}\left(2(\mu_0 - \mu_1 - 1) \leq \frac{\overline{X}_4 - \mu_1}{1/\sqrt{4}} \leq 2(\mu_0 - \mu_1 + 1)\right) =
$$

$$
\phi(2(\mu_0 - \mu_1 + 1)) - \phi(2(\mu_0 - \mu_1 - 1))
$$

because under μ_1, $\frac{\overline{X}_4 - \mu_1}{1/\sqrt{4}} \sim N(0, 1)$

What a confidence interval **is not**

Notice that in expressions such as $P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$, the value of the parameter θ is fixed and the extremes of the interval are **random variables**:

$$P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})]) = P_\theta(\{L(\mathbf{X}) \leq \theta\} \cap \{U(\mathbf{X}) \geq \theta\}).$$

Thus, a confidence interval does not represent the probability that a parameter is between its lower and upper limits.
This can be reinforced with a frequentist vision of the CI concept, and it is relativized with a Bayesian approach.

Example 1.12. $X_1, \cdots, X_n \sim U(0,\theta)$ and $Y_n = \max\{X_1, \cdots, X_n\}$

For a, b, c, d specified constants, consider the following two interval estimators: **1:** $[aY_n, bY_n]$, $1 \leq a < b$ and **2:** $[Y_n + c, Y_n + d]$, $0 \leq c < d$

1. The coverage probability is independent of θ. $\left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n$ is the confidence coefficient of $[aY, bY]$.

$$P_\theta(\theta \in [aY, bY]) = P_\theta\left(\frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\right) = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n$$

2. The coverage probability depends on θ.

$$\begin{aligned} P_\theta(\theta \in [Y+c, Y+d]) &= P_\theta\left(1 - \frac{d}{\theta} \leq \frac{Y}{\theta} \leq 1 - \frac{c}{\theta}\right) \\ &= \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n \end{aligned}$$

For any c and d, the confidence coefficient of this interval is 0 because $\lim_{\theta\to\infty} \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n = 0$

CI for transformations of one parameter

Let $[L(\mathbf{x}), U(\mathbf{x})]$ be a $(1-\alpha)$ confidence interval for θ.
If we want to build an interval for an invertible transformation $\tau(\theta)$ of the parameter θ:

- for increasing $\tau(\theta)$ the interval

$$[\tau(L(\mathbf{x})), \tau(U(\mathbf{x}))]$$

  is a $(1-\alpha)$ confidence interval for $\tau(\theta)$.

- for decreasing $\tau(\theta)$ the interval

$$[\tau(U(\mathbf{x})), \tau(L(\mathbf{x}))]$$

  is a $(1-\alpha)$ confidence interval for $\tau(\theta)$.

Methods for building interval estimators

There exists multiple methods to build interval estimators of a parameter $\theta \in \Theta \subseteq \mathbf{R}$.

1. Inverting a Test Statistic. There is a very strong correspondence between hypothesis testing and interval estimation. In general, every confidence set corresponds to a test and vice versa.
2. Pivotal Quantities
3. Bayesian Intervals (not covered)
4. Likelihood Intervals (not covered)
5. Asymptotic Intervals (not covered)
6. Bootstrap Intervals (not covered)

Inverting a Test Statistic

Theorem

Testing $H_0 : \theta = \theta_0$ for $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of an α level test. For each $\mathbf{x} \in \mathcal{X}$ define a set $CI(\mathbf{x}) \subseteq \Theta$ as

$$CI(\mathbf{x}) = \{\theta_0 \in \Theta : \mathbf{x} \in A(\theta_0)\} \subseteq \Theta$$

Then the random set $CI(\mathbf{X})$ is a $1 - \alpha$ confidence set for θ.

Conversely, let $CI(\mathbf{X})$ be a $1 - \alpha$ confidence set for θ. For any $\theta_0 \in \Theta$ define

$$A(\theta_0) = \{\mathbf{x} \in \mathcal{X} : \theta_0 \in C(\mathbf{x})\} \subset \mathcal{X}$$

Then $A(\theta_0)$ is the acceptance region of α a level test of $H_0 : \theta = \theta_0$.

Correspondence between CI and HT

The correspondence between testing and interval estimation is because both procedures look for consistency between sample statistics and population parameters, but from a slightly different perspective.

- In a hypothesis test the parameters are fixed and one looks for sample values in agreement with the parameters
- In interval estimation, we fix the observed sample and look for those parameter values that make this sample valid.

Example 1.10: CI for the mean from inverting a test statistic

Let $X_1, \ldots, X_n$ be a simple random sample of $X \sim N(\mu, \sigma^2)$ with σ known. Consider testing

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0 \ .$$

For a fixed α level, the unbiased (see appendix) uniformly most powerful test rejects H_0 if

$$|\overline{x}_n - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}$$

that is, the region of the sample space $\mathcal{X}$ where H_0 is accepted is the set of $\mathbf{x}$ such that

$$\overline{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \overline{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Example 1.10:

Given that the test has α size, we have that $P(\text{accept } H_0 | \mu = \mu_0) = 1 - \alpha$. So, for all μ_0

$$P\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \middle| \mu = \mu_0\right) = 1 - \alpha,$$

hence

$$P_\mu\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

for all μ, then $[\overline{X}_n - z_{\alpha/2}\sigma/\sqrt{n}, \overline{X}_n + z_{\alpha/2}\sigma/\sqrt{n}]$ is an interval estimator for μ with $1 - \alpha$ confidence level.

Pivotal quantities

Perhaps one of the most elegant methods of building set estimators and calculating coverage probabilities is the use of *pivotal quantities*.

Definition

Let $\mathbf{X} = (X_1, \ldots, X_n)$ a s.r.s of $X \sim F(x; \theta)$.

A function $Q(\mathbf{X}, \theta)$ of the sample and the parameter is a *pivotal quantity* if the distribution of $Q(\mathbf{X}, \theta)$ does not depend on the parameter θ, that is, $Q(\mathbf{X}, \theta)$ has the same distribution for all values of θ.

Pivotal quantities

Given a pivotal quantity $Q(\mathbf{X}, \theta)$, for any set A of the image space of Q, we have that $P_\theta(Q(\mathbf{X}, \theta) \in A)$ does not depend on θ. So, if we chose a set A_α such that

$$P_\theta(Q(\mathbf{X}, \theta) \in A) = 1 - \alpha, \text{ for all } \theta,$$

and we observe the sample $\mathbf{X} = \mathbf{x}$, then the set

$$CI(\mathbf{x}) = \{\theta : Q(\mathbf{x}, \theta) \in A\}$$

is a $1 - \alpha$ confidence set for θ.
If $\theta \in \mathbf{R}$, the use of pivotal quantities does not guarantee, in general, that the confidence set is an interval.

Pivotal quantities in location and scale families

If X has distribution belonging to a location and scale family then it is possible to define different pivotal quantities. Between them, we will emphasize the following:

Location:	$f_\mu(x) = f(x-\mu)$	$Q(\mathbf{x},\mu) = \overline{X} - \mu$
Scale:	$f_\sigma(x) = (1/\sigma) f(x/\sigma)$	$Q(\mathbf{x},\sigma) = \overline{X}/\sigma$
Loc. and Scale:	$f_{\mu,\sigma}(x) = (1/\sigma) f((x-\mu)/\sigma)$	$Q_1(\mathbf{x},\mu,\sigma) = (\overline{X}-\mu)/\sigma$, $Q_2(\mathbf{x},\mu,\sigma) = (\overline{X}-\mu)/S$

✍ Prove that they are pivotal quantities

Construction of CI from pivotal quantities

In practice, the way we build a confidence interval from a pivotal quantity is the following:

- We will assume that $Q(\mathbf{x}, \theta) \in \mathbf{R}$ and $\theta \in \mathbf{R}$. For a given value α, we look for numbers a and b such that

$$P_\theta(a \leq Q(\mathbf{X}, \theta) \leq b) = 1 - \alpha.$$

  Notice that a and b do not depend on θ because Q is a pivotal quantity, and the choice of a and b will not be unique, in general.

- For each θ_0, the set

$$A(\theta_0) = \{\mathbf{x} : a \leq Q(\mathbf{x}, \theta) \leq b\}$$

  is the acceptance region for a α level test of $H_0 : \theta = \theta_0$ based on the statistic $T(\mathbf{X}) = Q(\mathbf{X}, \theta_0)$.

Construction of CI from pivotal quantities

- Inverting this test, we obtain the $1-\alpha$ confidence set for θ:
$$C(\mathbf{x}) = \{\theta : a \leq Q(\mathbf{x}, \theta) \leq b\}.$$
- If $g_{\mathbf{x}}(\theta) = Q(\mathbf{x}, \theta)$ is a monotone function of θ for each fixed $\mathbf{x}$, then $C(\mathbf{x})$ will be an interval.
  - If $g_{\mathbf{x}}(\theta)$ is increasing, then $C(\mathbf{x}) = [L(\mathbf{x}, a), U(\mathbf{x}, b)]$, while
  - if $g_{\mathbf{x}}(\theta)$ is decreasing, then $C(\mathbf{x}) = [L(\mathbf{x}, b), U(\mathbf{x}, a)]$.
- If $g_{\mathbf{x}}(\theta)$ is invertible, then
$$C(\mathbf{x}) = [\min\{g_{\mathbf{x}}^{-1}(a), g_{\mathbf{x}}^{-1}(b)\}, \max\{g_{\mathbf{x}}^{-1}(a), g_{\mathbf{x}}^{-1}(b)\}].$$

Example 1.10: CI for the mean and variance of the normal distribution

If $X \sim N(\mu, \sigma)$ with σ known, then the distribution of X is from a location family and therefore, $Q(\mathbf{X}, \mu) = (\overline{X}_n - \mu)/\sigma$ is a pivotal quantity. Furthermore,

$$Z = \sqrt{n}Q(\mathbf{X}, \mu) = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim N(0, 1).$$

If σ is unknown, a pivotal quantity is $Q(\mathbf{X}, \mu) = (\overline{X}_n - \mu)/S$. Moreover,

$$t = \sqrt{n}Q(\mathbf{X}, \mu) = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S} \sim t_{n-1}.$$