Statistics

1. | Point estimation

Statistical models

Definition 1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space[1], Θ be a set, $n \in \mathbb{N}$ and $x_1, \ldots, x_n$ be a collection of data that we may assume that they are the outcomes of a random vector $\mathbf{X}_n = (X_1, \ldots, X_n)$ defined on $(\Omega, \mathcal{A}, \mathbb{P})$. Suppose, moreover, that the outcomes of $\mathbf{X}_n$ are in a set $\mathcal{X} \subseteq \mathbb{R}^n$, the law $\mathbf{X}_n$ is one in the set $\mathcal{P} = \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\}$ and $\mathcal{F}$ is a σ-algebra over $\mathcal{X}$[2]. We define a *statistical model* as the triplet $(\mathcal{X}, \mathcal{F}, \mathcal{P})$[3]. The set $\mathcal{X}$ is called *sample space*, and the set Θ, *parameter space*. The random vector $\mathbf{X}_n$ is called *random sample*. If, moreover, $X_1, \ldots, X_n$ are i.i.d. random variables, $\mathbf{X}_n$ is called a *simple random sample*. The value $(x_1, \ldots, x_n) \in \mathcal{X}$ is called a *realization* of $(X_1, \ldots, X_n)$.

Definition 2. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model. We say $\mathcal{P} = \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\}$ is *identifiable* if the function

$$\begin{aligned} \Theta &\longrightarrow \mathcal{P} \\ \theta &\longmapsto \mathbb{P}_\theta^{\mathbf{X}_n} \end{aligned}$$

is injective[4].

Definition 3. A statistical model $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ is said to be *parametric* if $\Theta \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$[5].

Statistics and estimators

Definition 4 (Statistic). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model. We define a *statistic* $\mathbf{T}$ as a Borel measurable function. That is, $\mathbf{T}$ can be written as $\mathbf{T} = \mathbf{h}(X_1, \ldots, X_n)$, where $\mathbf{h} : \mathcal{X} \to \mathbb{R}^m$ is a Borel measurable function. Hence, $\mathbf{T}$ is a random vector. The value m is the *dimension* of the statistic.

Definition 5. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model. We define the *sample mean* as the statistic:

$$T(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i =: \overline{X}_n$$

Given a realization $(x_1, \ldots, x_n) \in \mathcal{X}$, we will denote $\overline{x}_n := \overline{X}_n(x_1, \ldots, x_n)$[6].

Definition 6. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model. We define the *sample variance* as the statistic:

$$T(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^n {(X_i - \overline{X}_n)}^2 =: {S_n}^2$$

We define the *corrected sample variance* as the statistic:

$$T(X_1, \ldots, X_n) = \frac{1}{n-1} \sum_{i=1}^n {(X_i - \overline{X}_n)}^2 =: {\tilde{S}_n}^2$$

Given a realization $(x_1, \ldots, x_n) \in \mathcal{X}$, we will denote ${s_n}^2 := {S_n}^2(x_1, \ldots, x_n)$ and ${\tilde{s}_n}^2 := {\tilde{S}_n}^2(x_1, \ldots, x_n)$[7].

Proposition 7. Let $X_1, \ldots, X_n$ be random variables. Then:

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

Definition 8. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model, $\theta \in \Theta$ and $g : \Theta \to \Theta$ be a function. An *estimator* of $g(\theta)$ is a statistic $\hat{\theta}$ whose outcomes are in Θ and does not depend on any unknown parameter. It is used to give an estimation of the (supposedly unknown) parameter $g(\theta)$.

Properties of estimators

Definition 9 (Bias). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model, $g : \Theta \to \Theta$ be a function and $\hat{\theta}$ be an integrable estimator of $g(\theta) \in \Theta$. We define the *bias* of $\hat{\theta}$ with respect to θ as:

$$\operatorname{bias}(\hat{\theta}) := \mathbb{E}(\hat{\theta}) - g(\theta)$$

We say that $\hat{\theta}$ is an *unbiased estimator* of $g(\theta)$ if $\operatorname{bias}(\hat{\theta}) = 0\ \forall \theta \in \Theta$. Otherwise we say that it is a *biased estimator* of $g(\theta)$.

Proposition 10. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model, $g : \Theta \to \Theta$ be a function and $\hat{\theta}$ be an integrable estimator of $g(\theta) \in \Theta$. Suppose that $\operatorname{bias}(\hat{\theta}) = cg(\theta)$ for some $c \in \mathbb{R}$. Then, $\frac{\hat{\theta}}{c+1}$ is an unbiased estimator for $g(\theta)$

Proposition 11. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model such that $X_1, \ldots, X_n$ are square-integrable[8] i.i.d. random variables with mean μ and variance σ^2. Then:

$$\mathbb{E}(\overline{X}_n) = \mu \quad \text{and} \quad \operatorname{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

Hence, the estimator $\overline{X}_n$ of μ is unbiased.

[1] From now on we will assume that the random variables are defined always in the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$, so we will omit to say that.

[2] That is, $\mathcal{P}$ denotes a family of probability distributions of $\mathbf{X}_n$ in $(\mathcal{X}, \mathcal{F})$, indexed by $\theta \in \Theta$. Note that we denote that distribution of $\mathbf{X}_n$ by $\mathbb{P}^{\mathbf{X}_n}$ to distinguish it from the probability distribution $\mathbb{P}_{\mathbf{X}_n}$ in $(\Omega, \mathcal{A}, \mathbb{P})$.

[3] Often we will take $\mathcal{F} = \mathcal{B}(\mathcal{X})$.

[4] From now on, we will suppose that all the sets $\mathcal{P}$ are always identifiable.

[5] There are cases where Θ is not a subset of $\mathbb{R}^d$. For example, we could have $\Theta = \{f : \mathbb{R} \to \mathbb{R}_{\geq 0} : \int_{-\infty}^{+\infty} f(x)\, dx = 1\}$.

[6] Some times, and if the context is clear, we will denote $\overline{x}_n$ simply as $\overline{x}$.

[7] Some times, and if the context is clear, we will denote ${s_n}^2$ and ${\tilde{s}_n}^2$ simply as s^2 and $\tilde{s}^2$, respectively.

[8] That is, with finite 2nd moments.

Proposition 12. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model such that $X_1, \ldots, X_n$ are square-integrable i.i.d. random variables with mean μ and variance σ^2. Then:

$$\mathbb{E}({S_n}^2) = \frac{n-1}{n}\sigma^2 \quad \text{and} \quad \mathbb{E}({\tilde{S}_n}^2) = \sigma^2$$

Hence, the estimator ${\tilde{S}_n}^2$ of σ^2 is unbiased whereas the estimator ${S_n}^2$ of σ^2 is biased.

Definition 13. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model, $g : \Theta \to \Theta$ be a function and $\hat{\theta}$ be a square-integrable integrable estimator of $g(\theta) \in \Theta$. The *mean squared error* (*MSE*) of $\hat{\theta}$ is the function:

$$\mathrm{MSE}(\hat{\theta}) := \mathbb{E}\left(\left(\hat{\theta} - g(\theta)\right)^2\right)$$

Proposition 14. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model, $g : \Theta \to \Theta$ be a function and $\hat{\theta}$ be a square-integrable integrable estimator of $g(\theta) \in \Theta$. Then:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \left(\mathrm{bias}(\hat{\theta})\right)^2$$

Definition 15. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model, $g : \Theta \to \Theta$ be a function and $\hat{\theta}$, $\tilde{\theta}$ be estimators of $g(\theta) \in \Theta$. We say that $\hat{\theta}$ is *more efficient than* $\tilde{\theta}$ if

$$\mathrm{Var}(\hat{\theta}) < \mathrm{Var}(\tilde{\theta}) \quad \forall \theta \in \Theta$$

Definition 16. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model and $\hat{\theta}$ be a square integrable estimator of $\theta \in \Theta$. We say that $\hat{\theta}$ is a *minimum-variance unbiased estimator* (*MVUE*) if it is an unbiased estimator that has lower variance than any other unbiased estimator $\forall \theta \in \Theta$.

Proposition 17. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model. Then, the MVUE is unique almost surely.

Sufficient statistics

Definition 18. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model and $\mathbf{T}$ be a statistic. We say that $\mathbf{T}$ is *sufficient* for $\theta \in \Theta$ if the joint conditional distribution of $(X_1, \ldots, X_n)$ given $\mathbf{T}(X_1, \ldots, X_n) = \mathbf{t}$ does not depend on θ.

Theorem 19 (Fisher-Neyman factorization theorem). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model and T be a statistic. Then, T is sufficient if and only if $\forall \mathbf{x}_n \in \mathcal{X}$ we have:

1. For the discrete case:

$$p_{\mathbf{X}_n}(\mathbf{x}_n; \theta) = g(T(\mathbf{x}_n); \theta) h(\mathbf{x}_n)$$

2. For the continuous case:

$$f_{\mathbf{X}_n}(\mathbf{x}_n; \theta) = g(T(\mathbf{x}_n); \theta) h(\mathbf{x}_n)$$

for certain functions g and h. Here we have denoted by $p_{\mathbf{X}_n}(\mathbf{x}_n; \theta)$ the joint pmf of $\mathbf{X}_n$ (in the discrete case) and by $f_{\mathbf{X}_n}(\mathbf{x}_n; \theta)$ the joint pdf of $\mathbf{X}_n$ (in the continuous case).

Asymptotic properties

Definition 20. For each $n \in \mathbb{N}$, let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model with $X_1, \ldots, X_n$ being i.i.d., $g : \Theta \to \Theta$ be a function and $\hat{\theta}_n$ be an estimator of $g(\theta) \in \Theta$. We say that the sequence $(\hat{\theta}_n)$ is a *weakly consistent estimator* of $g(\theta)$ if $\hat{\theta}_n \xrightarrow{\mathbb{P}} g(\theta)$.

Definition 21. For each $n \in \mathbb{N}$, let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model with $X_1, \ldots, X_n$ being i.i.d., $g : \Theta \to \Theta$ be a function and $\hat{\theta}_n$ be an estimator of $g(\theta) \in \Theta$. We say that the sequence $(\hat{\theta}_n)$ is a *strongly consistent estimator* of $g(\theta)$ if $\hat{\theta}_n \xrightarrow{\text{a.s.}} g(\theta)$.

Definition 22. For each $n \in \mathbb{N}$, let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model with $X_1, \ldots, X_n$ being i.i.d., $g : \Theta \to \Theta$ be a function and $\hat{\theta}_n$ be an estimator of $g(\theta) \in \Theta$. We say that the sequence $(\hat{\theta}_n)$ is a *consistent estimator in* L^2 of $g(\theta)$ if

$$\lim_{n \to \infty} \mathbb{E}\left(\left(\hat{\theta}_n - g(\theta)\right)^2\right) = \lim_{n \to \infty} \mathrm{MSE}(\hat{\theta}_n) = 0$$

Proposition 23. For each $n \in \mathbb{N}$, let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model with $X_1, \ldots, X_n$ being i.i.d., $g : \Theta \to \Theta$ be a function and $\hat{\theta}_n$ be a consistent estimator in L^2 of $g(\theta) \in \Theta$. Then, $\hat{\theta}_n$ is a weakly consistent estimator of $g(\theta)$.

Definition 24. For each $n \in \mathbb{N}$, let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model with $X_1, \ldots, X_n$ being i.i.d., $g : \Theta \to \Theta$ be a function and $\hat{\theta}_n$ be an estimator of $g(\theta) \in \Theta$. We say that the sequence $(\hat{\theta}_n)$ is an *asymptotically unbiased estimator* of $g(\theta)$ if

$$\mathbb{E}(\hat{\theta}_n) \longrightarrow g(\theta)$$

Definition 25. For each $n \in \mathbb{N}$, let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a parametric statistical model with $X_1, \ldots, X_n$ being i.i.d. whose variance is σ^2, $g : \Theta \to \Theta$ be a function and $\hat{\theta}_n$ be an estimator of $g(\theta) \in \Theta$. We say that the sequence $(\hat{\theta}_n)$ is *an asymptotically normal estimator* of $g(\theta)$ with asymptotically variance $\frac{\sigma^2}{n}$ if

$$\sqrt{n}(\hat{\theta}_n - g(\theta)) \xrightarrow{\text{d}} N(0, \sigma^2) \qquad \forall \theta \in \Theta$$

In that case, we denote it by $\hat{\theta}_n \overset{\text{a}}{\sim} N\left(g(\theta), \frac{\sigma^2}{n}\right)$.

Methods of estimation

Definition 26 (Method of moments). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_{\boldsymbol{\theta}}^{\mathbf{X}_n} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\})$ be a parametric statistical model such that $X_1, \ldots, X_n$ are i.i.d. random variables, and μ_k be k-th moment of each of them. Suppose $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. Then, given a realization $\mathbf{x}_n = (x_1, \ldots, x_n) \in \mathcal{X}$ of $\mathbf{X}_n$, an estimator $\tilde{\boldsymbol{\theta}}(\mathbf{x}_n) = (\tilde{\theta}_1(\mathbf{x}_n), \ldots, \tilde{\theta}_d(\mathbf{x}_n))$ of $\boldsymbol{\theta}$ is given by the solution of the following system:

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i = \mu_1(\theta_1,\dots,\theta_d) \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} {x_i}^2 = \mu_2(\theta_1,\dots,\theta_d) \\ \qquad\vdots \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} {x_i}^d = \mu_d(\theta_1,\dots,\theta_d) \end{cases}$$

Proposition 27. The estimators obtained by the method of moments are strongly consistent and consistent in L^2.

Definition 28 (Likelihood). Let $(\mathcal{X},\mathcal{F},\{\mathbb{P}_\theta^{\mathbf{X}_n}:\theta\in\Theta\subseteq\mathbb{R}\})$ be a parametric statistical model, $\mathbf{x}_n\in\mathcal{X}$ be a realization of $\mathbf{X}_n$.

1. For the discrete case, let $p_{\mathbf{X}_n}(\mathbf{x}_n;\theta)$ be the pmf of $\mathbb{P}_\theta^{\mathbf{X}_n}$. In this case, we define the *likelihood function* as the function:
$$\begin{aligned} L(\cdot;\mathbf{x}_n):\Theta &\longrightarrow \mathbb{R} \\ \theta &\longmapsto p_{\mathbf{X}_n}(\mathbf{x}_n;\theta) \end{aligned}$$
2. For the continuous case, let $f_{\mathbf{X}_n}(\mathbf{x}_n;\theta)$ be the pdf of $\mathbb{P}_\theta^{\mathbf{X}_n}$. In this case, we define the *likelihood function* as the function:
$$\begin{aligned} L(\cdot;\mathbf{x}_n):\Theta &\longrightarrow \mathbb{R} \\ \theta &\longmapsto f_{\mathbf{X}_n}(\mathbf{x}_n;\theta) \end{aligned}$$

Definition 29 (Maximum likelihood method). Let $(\mathcal{X},\mathcal{F},\{\mathbb{P}_\theta^{\mathbf{X}_n}:\theta\in\Theta\})$ be a statistical model and $\mathbf{x}_n\in\mathcal{X}$ be a realization of $\mathbf{X}_n$. A *maximum likelihood estimator* (*MLE*) of $\theta\in\Theta$ is the estimator $\hat{\theta}$ such that:

$$L(\hat{\theta};\mathbf{x}_n)=\max\{L(\theta;\mathbf{x}_n):\theta\in\Theta\}$$ [9]

Definition 30. Let $(\mathcal{X},\mathcal{F},\{\mathbb{P}_{\boldsymbol{\theta}}^{\mathbf{X}_n}:\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^d\})$ be a parametric statistical model, $\mathbf{x}_n\in\mathcal{X}$ be a realization of $\mathbf{X}_n$. We define the *log-likelihood function* as:

$$\ell(\boldsymbol{\theta};\mathbf{x}_n):=\ln L(\boldsymbol{\theta};\mathbf{x}_n)$$

We define the *score function* as:

$$\mathbf{S}(\boldsymbol{\theta};\mathbf{x}_n):=\frac{\partial\ell}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{x}_n)$$

Proposition 31. Let $(\mathcal{X},\mathcal{F},\{\mathbb{P}_{\boldsymbol{\theta}}^{\mathbf{X}_n}:\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^d\})$ be a parametric statistical model, $\mathbf{x}_n\in\mathcal{X}$ be a realization of $\mathbf{X}_n$ Then, a MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is the one that satisfies:

$$\frac{\partial L}{\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}};\mathbf{x}_n)=\mathbf{0}$$

Or equivalently, $\frac{\partial\ell}{\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}};\mathbf{x}_n)=\mathbf{0}$.

Proposition 32 (Invariance of the MLE). Let $(\mathcal{X},\mathcal{F},\{\mathbb{P}_\theta^{\mathbf{X}_n}:\theta\in\Theta\})$ be a statistical model and $g:\Theta\to\Theta$ be a measurable function. Suppose $\hat{\theta}$ is a MLE of θ. Then, $g(\hat{\theta})$ is a MLE of $g(\theta)$.

Regular statistical models

Definition 33. A statistical model $(\mathcal{X},\mathcal{F},\{\mathbb{P}_\theta^{\mathbf{X}_n}:\theta\in\Theta\})$ is said to be *regular* if it satisfies the following conditions:

1. Θ is open.
2. The support of $\mathbb{P}_\theta^{\mathbf{X}_n}$ does not depend on θ.
3. The function $L(\theta;\mathbf{x}_n)$ is two times differentiable with respect to θ $\forall\mathbf{x}_n\in\mathcal{X}$ (except in a set of probability zero) and moreover:
   i) For the discrete case:
   $$\frac{\partial^2}{\partial\theta^2}\sum_{\mathbf{x}_n\in\mathcal{X}}L(\theta;\mathbf{x}_n)=\sum_{\mathbf{x}_n\in\mathcal{X}}\frac{\partial^2 L}{\partial\theta^2}(\theta;\mathbf{x}_n)$$
   ii) For the continuous case:
   $$\frac{\partial^2}{\partial\theta^2}\int_{\mathcal{X}}L(\theta;\mathbf{x}_n)\,\mathrm{d}\mathbf{x}_n=\int_{\mathcal{X}}\frac{\partial^2 L}{\partial\theta^2}(\theta;\mathbf{x}_n)\,\mathrm{d}\mathbf{x}_n$$
4. For all $\theta\in\Theta$, we have:
$$0<\int_{\mathcal{X}}\left(\frac{\partial^2\ell}{\partial\theta^2}(\theta;\mathbf{x}_n)\right)^2 f_{\mathbf{X}_n}(\mathbf{x}_n;\boldsymbol{\theta})\,\mathrm{d}\mathbf{x}_n<\infty$$

Definition 34. Let $(\mathcal{X},\mathcal{F},\{\mathbb{P}_{\boldsymbol{\theta}}^{\mathbf{X}_n}:\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^d\})$ be a regular parametric statistical model, $\mathbf{x}_n\in\mathcal{X}$ be a realization of $\mathbf{X}_n$. We define the *observed information* of the model as:

$$\mathbf{J}(\boldsymbol{\theta};\mathbf{x}_n)=-\frac{\partial^2\ell}{\partial\boldsymbol{\theta}^2}(\boldsymbol{\theta};\mathbf{x}_n)$$

We define the *Fisher information* of the model as:

$$\mathbf{I}(\boldsymbol{\theta})=\mathbb{E}(\mathbf{J}(\boldsymbol{\theta};\mathbf{X}_n))=-\mathbb{E}\left(\frac{\partial^2\ell}{\partial\boldsymbol{\theta}^2}(\boldsymbol{\theta};\mathbf{X}_n)\right)$$ [10]

Proposition 35. Let $(\mathcal{X},\mathcal{F},\{\mathbb{P}_{\boldsymbol{\theta}}^{\mathbf{X}_n}:\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^d\})$ be a regular parametric statistical model, $\mathbf{x}_n\in\mathcal{X}$ be a realization of $\mathbf{X}_n$. Then, $\mathbb{E}(\mathbf{S}(\boldsymbol{\theta};\mathbf{X}_n))=0$ and

$$\mathbf{I}(\boldsymbol{\theta})=\mathrm{Var}(\mathbf{S}(\boldsymbol{\theta};\mathbf{X}_n))=\mathbb{E}\left[\left(\frac{\partial\ell}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta};\mathbf{X}_n)\right)^2\right]$$

for all $\boldsymbol{\theta}\in\Theta$.

Proposition 36. Let $(\mathcal{X},\mathcal{F},\{\mathbb{P}_{\boldsymbol{\theta}}^{X_1}:\boldsymbol{\theta}\in\Theta\})$ be a regular parametric statistical model of one observation $x_1\in\mathcal{X}$. Then, the model corresponding to n i.i.d. observations $x_1,\dots,x_n$ is regular and

$$\mathbf{I}(\boldsymbol{\theta})=n\mathbf{I}_1(\boldsymbol{\theta})$$

where $\mathbf{I}_1(\boldsymbol{\theta})$ denotes the Fisher information in the model with one observation.

Definition 37. Let $(\mathcal{X},\mathcal{F},\{\mathbb{P}_\theta^{\mathbf{X}_n}:\theta\in\Theta\})$ be a statistical model and T be a statistic. We say that T is *regular* if

[9] Note that sometimes this estimator is not unique or may not even exist.

[10] Since generally $\mathbf{J}(\boldsymbol{\theta};\mathbf{X}_n)$ will be a matrix, the expectation of $\mathbf{J}(\boldsymbol{\theta};\mathbf{X}_n)$ is taken component by component.

1. for the discrete case:
$$\frac{\partial}{\partial \theta}\sum_{\mathbf{x}_n\in\mathcal{X}} T(\mathbf{x}_n)L(\theta;\mathbf{x}_n) = \sum_{\mathbf{x}_n\in\mathcal{X}} T(\mathbf{x}_n)\frac{\partial L}{\partial \theta}(\theta;\mathbf{x}_n)$$
2. for the continuous case:
$$\frac{\partial}{\partial \theta}\int_{\mathcal{X}} T(\mathbf{x}_n)L(\theta;\mathbf{x}_n)\,\mathrm{d}\mathbf{x}_n = \int_{\mathcal{X}} T(\mathbf{x}_n)\frac{\partial L}{\partial \theta}(\theta;\mathbf{x}_n)\,\mathrm{d}\mathbf{x}_n$$

for all $\theta \in \Theta$.

Theorem 38 (Cramér-Rao bound). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a regular parametric statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$, $g : \Theta \to \Theta$ be a differentiable function and $\hat{\theta}$ be a regular estimator of $g(\theta) \in \Theta$. Then:
$$\mathrm{Var}(\hat{\theta}) \geq \frac{g'(\theta)^2}{I(\theta)}\left[1 + \left(\mathrm{bias}'(\hat{\theta})\right)^2\right]$$
Moreover if the estimator $\hat{\theta}$ is unbiased we have:
$$\mathrm{Var}(\hat{\theta}) \geq \frac{g'(\theta)^2}{I(\theta)}$$

Definition 39. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a regular statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$, $g : \Theta \to \Theta$ be a differentiable function and $\hat{\theta}$ be a regular and unbiased estimator of $g(\theta) \in \Theta$. We say that $\hat{\theta}$ is an *efficient estimator* of $g(\theta)$ if
$$\mathrm{Var}(\hat{\theta}) = \frac{g'(\theta)^2}{I(\theta)}$$
We say that $\hat{\theta}$ is an *asymptotic efficient estimator* of $g(\theta)$ if the asymptotic variance of $\hat{\theta}$ is $\frac{g'(\theta)^2}{I(\theta)}$.

Proposition 40. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a regular statistical model, $g : \Theta \to \Theta$ be a function and $\hat{\theta}$ be a regular, unbiased and efficient estimator of $g(\theta) \in \Theta$. Then, $\hat{\theta}$ is a MVUE in the class of regular estimators.

Theorem 41. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a regular statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$ and $\hat{\theta}$ be a MLE of $\theta \in \Theta$. Suppose that $\frac{\partial^2 \ell}{\partial \theta^2}$ is a continuous function of θ and that
$$\left|\frac{\partial^2 \ell}{\partial \theta^2}(\tilde{\theta};\mathbf{x}_n)\right| < h(\mathbf{x}_n;\theta)$$
for all $\tilde{\theta}$ in a neighbourhood of θ with $\int_{\mathcal{X}} h(\mathbf{x}_n;\theta)L(\theta;\mathbf{x}_n)\,\mathrm{d}\mathbf{x}_n < \infty$. Then:
$$\hat{\theta} \xrightarrow{\mathrm{d}} N\left(\theta, \frac{1}{I(\theta)}\right)$$
Thus, $\hat{\theta}$ is an asymptotically efficient estimator of θ. Hence, an asymptotic confidence interval for θ of confidence $1-\alpha$ is:
$$\theta \in \left(\hat{\theta} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{I(\hat{\theta})}}, \hat{\theta} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{I(\hat{\theta})}}\right)$$
where $z_{1-\frac{\alpha}{2}}$ denote the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution (see Theorem 45).

Theorem 42 (Delta method). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a statistical model, $g : \Theta \to \Theta$ be a two-times differentiable function such that $g'(\theta) \neq 0$ and $\hat{\theta}$ be an estimator of $\theta \in \Theta$. Then:
$$g(\hat{\theta}) \xrightarrow{\mathrm{d}} N\left(g(\theta), g'(\theta)^2\mathrm{Var}(\hat{\theta})\right)$$

Order statistics

Definition 43. Let $X_1, \ldots, X_n$ be random variables. We define the k-th *order statistic*, denoted by $X_{(k)}$ of the sample $X_1, \ldots, X_n$ as the k-th smallest value of it. In particular:
$$X_{(1)} := \min\{X_1, \ldots, X_n\} \quad X_{(n)} := \max\{X_1, \ldots, X_n\}$$
The sample $X_{(1)}, \ldots, X_{(n)}$ is usually called *order statistics*.

2. | Distributions relating $N(\mu, \sigma^2)$

Standard normal distribution

Definition 44. We denote by $\Phi(t)$ the cdf of a standard normal distribution $N(0,1)$.

Definition 45 (Quantile). We define quantile function $Q_X(p)$ of a distribution of a random variable X as the inverse function of the cdf. That is, $Q_X(p)$ satisfies:
$$\mathbb{P}(X \leq Q_X(p)) = p$$
In particular, we denote the quantile of a standard normal distribution as $z_p := Q_X(p) = \Phi^{-1}(p)$.

Multivariate normal distribution

Definition 46. Let $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{\Sigma} \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive-definite matrix and $\mathbf{X}$ be a random vector. We say that $\mathbf{X}$ has *multivariate normal distribution*, and we denote it by $\mathbf{X} \sim N(\mathbf{b}, \mathbf{\Sigma})$ if has density function:
$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{n}{2}}(\det \mathbf{\Sigma})^{-\frac{1}{2}}\mathrm{e}^{-\frac{(\mathbf{x}-\mathbf{b})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{b})}{2}}$$
The vector $\mathbf{b}$ is called *mean vector* and the matrix $\mathbf{\Sigma}$, *covariance matrix*.

Proposition 47. Let $\mathbf{\Sigma} \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive-definite matrix. Then, $\exists \mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$ such that $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$[11].

Proposition 48. Let $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^{\mathrm{T}} \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive-definite matrix with $\mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$ and $\mathbf{X}$, $\mathbf{Z}$ be random vectors.

- If $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_n)$, then $\mathbf{A}\mathbf{Z} + \mathbf{b} \sim N(\mathbf{b}, \mathbf{\Sigma})$.
- If $\mathbf{X} \sim N(\mathbf{b}, \mathbf{I}_n)$, then $\mathbf{A}^{-1}(\mathbf{X} - \mathbf{b}) \sim N(\mathbf{0}, \mathbf{\Sigma})$.

[11] When $\mathbf{\Sigma}$ is the covariance matrix, the matrix $\mathbf{A}$ such that $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$ plays the role of the *multivariate standard deviation*.

Proposition 49. Let $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{\Sigma} \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive-definite matrix and $\mathbf{X} = (X_1, \dots, X_n) \sim N(\mathbf{b}, \mathbf{\Sigma})$. Then, $\mathbb{E}(\mathbf{X}) = \mathbf{b}$ and $\mathrm{Var}(\mathbf{X}) = \mathbf{\Sigma}$ and moreover:

$$\mathbf{b} = (\mathbb{E}(X_1), \dots, \mathbb{E}(X_2))^{\mathrm{T}}$$

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Var}(X_n) \end{pmatrix}$$

Proposition 50. Let $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$, $\mathbf{\Sigma} \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive-definite matrix, $\mathbf{B} \in \mathrm{GL}_n(\mathbb{R})$, $\mathbf{X} \sim N(\mathbf{b}, \mathbf{\Sigma})$ and $\mathbf{Y} := \mathbf{B}\mathbf{X} + \mathbf{c}$. Then:

$$\mathbf{Y} \sim N(\mathbf{B}\mathbf{b} + \mathbf{c}, \mathbf{B}\mathbf{\Sigma}\mathbf{B}^{\mathrm{T}})$$

χ^2-distribution

Proposition 51. Let $n \in \mathbb{N}$ and $X_1, \dots, X_n$ be independent random variables such that $X_i \sim \mathrm{Gamma}(\alpha_i, \beta)$ for $i = 1, \dots, n$. Then:

$$\sum_{i=1}^n X_i \sim \mathrm{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right)$$

Corollary 52. Let $n \in \mathbb{N}$ and $Z_1, \dots, Z_n$ be i.i.d. random variable with standard normal distribution. Then:

$${Z_1}^2 + \cdots + {Z_n}^2 \sim \mathrm{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$$

Definition 53. We define the *chi-squared distribution with n degrees of freedom*, denoted as ${\chi_n}^2$, as the distribution

$${\chi_n}^2 := \mathrm{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$$

which is the distribution of ${Z_1}^2 + \cdots + {Z_n}^2$, where $Z_1, \dots, Z_n \sim N(0,1)$ are i.i.d. random variables. Its pdf is:

$$f_{{\chi_n}^2}(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} \mathrm{e}^{-\frac{x}{2}} \mathbf{1}_{(0,\infty)}(x)$$

We will denote by ${\chi_{n;p}}^2 := Q_{{\chi_n}^2}(p)$ the quantile of the ${\chi_n}^2$.

Proposition 54. Let $X \sim {\chi_a}^2$ and $Y \sim {\chi_b}^2$ be i.i.d. random variables. Then:

$$X + Y \sim {\chi_{a+b}}^2$$

Proposition 55. Let $X \sim \mathrm{Gamma}(\alpha, \beta)$ and $c \in \mathbb{R}_{>0}$. Then, $cX \sim \mathrm{Gamma}(\alpha, \beta/c)$. In particular, if $X \sim \mathrm{Gamma}(n, 1)$, then $2X \sim {\chi_{2n}}^2$.

Student's t-distribution

Definition 56. Let $n \in \mathbb{N}$ and $Z \sim N(0,1)$ and $Y \sim {\chi_n}^2$ be independent random variables. We define the *Student's t-distribution with n degrees of freedom* as the distribution of:

$$\frac{Z}{\sqrt{Y/n}}$$

We will denote by $t_{n;p} := Q_{t_n}(p)$ the quantile of the t_n.

Proposition 57. Let $n \in \mathbb{N}$. Then, the pdf of t_n is:

$$f_{t_n}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$ [12]

Fisher's theorem

Theorem 58 (Fisher's theorem). Let $(\mathcal{X}, \mathcal{F}, \{X_1, \dots, X_n \sim N(\mu, \sigma^2) \text{ i.i.d.} : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{\geq 0}\})$ be a parametric statistical model. Then:

1. $\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
2. ${\tilde{S}_n}^2 \sim \frac{\sigma^2}{n-1} {\chi_{n-1}}^2$
3. $\overline{X}_n$ and ${\tilde{S}_n}^2$ are independent.

Corollary 59. Let $n \in \mathbb{N}$ and $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ be i.i.d. random variables. Then:

$$\frac{\overline{X}_n - \mu}{\frac{\tilde{S}_n}{\sqrt{n}}} \sim t_{n-1}$$

Corollary 60. Let $n \in \mathbb{N}$ and $X \sim t_n$ be a random variable. Then:

$$X \xrightarrow{\mathrm{d}} N(0,1)$$

Hence, $N(0,1) = t_\infty$.

Corollary 61. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a parametric statistical model and suppose $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ are i.i.d. random variables. Then, the estimators $\overline{X}_n$ of μ and ${\tilde{S}_n}^2$ of σ^2 are unbiased and consistent.

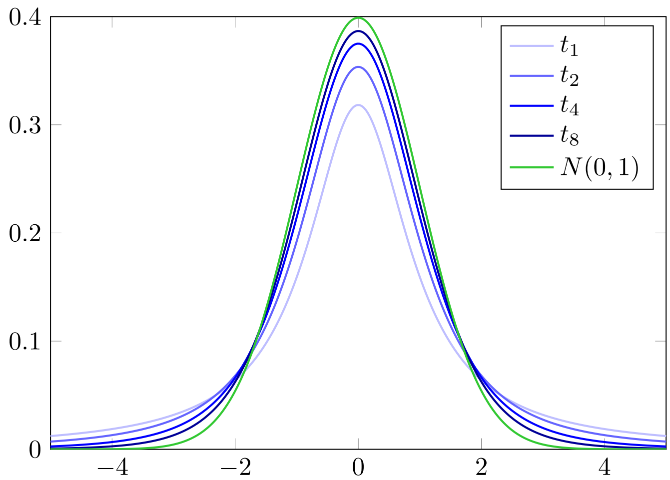

Figure 1: Probability density function of 4 Student's t-distribution together with a standard normal $N(0,1) = t_\infty$.

[12] It makes sense if we replace the value $n \in \mathbb{N}$ for a value $\nu \in \mathbb{R}_{>0}$. However the original definition of t_n from the ${\chi_n}^2$ fails.

3. | Confidence intervals

Confidence regions

Definition 62. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_{\boldsymbol{\theta}}^{\mathbf{X}_n} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\})$ be a parametric statistical model and $\mathbf{g} : \Theta \to \mathbb{R}^m$ be a function with $m \leq d$. A *confidence region* for $\mathbf{g}(\boldsymbol{\theta})$ with *confidence level* $\gamma \in [0, 1]$ is a random region $C(\mathbf{X}_n)$ such that:

$$\mathbb{P}(\mathbf{g}(\boldsymbol{\theta}) \in C(\mathbf{X}_n)) \geq \gamma \quad \forall \theta \in \Theta$$

If $d = 1$, we talk about *confidence intervals*. The value $\alpha := 1 - \gamma$ is called *significance level*.

Definition 63. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_{\boldsymbol{\theta}}^{\mathbf{X}_n} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\})$ be a parametric statistical model and $\mathbf{g} : \Theta \to \mathbb{R}^m$ be a function with $m \leq d$. A *pivot* for $\mathbf{g}(\boldsymbol{\theta})$ is a measurable function

$$\begin{aligned} \pi : \mathcal{X} \times \mathbf{g}(\Theta) &\longrightarrow \mathbb{R}^m \\ (\mathbf{x}_n, \mathbf{g}(\boldsymbol{\theta})) &\longmapsto \boldsymbol{\pi}(\mathbf{x}_n, \mathbf{g}(\boldsymbol{\theta})) \end{aligned}$$

such that the distribution of $\boldsymbol{\pi}(\mathbf{x}_n, \mathbf{g}(\boldsymbol{\theta}))$ does not depend on $\boldsymbol{\theta}$.

Proposition 64. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_{\boldsymbol{\theta}}^{\mathbf{X}_n} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\})$ be a parametric statistical model, $\gamma \in [0, 1]$, $\mathbf{g} : \Theta \to \mathbb{R}^m$ be a function with $m \leq d$, $\boldsymbol{\pi}(\mathbf{x}_n, \mathbf{g}(\boldsymbol{\theta}))$ be a pivot for $\mathbf{g}(\boldsymbol{\theta})$ and $B \in \mathcal{B}(\mathbb{R}^m)$ such that:

$$\mathbb{P}(\boldsymbol{\pi}(\mathbf{X}_n, \mathbf{g}(\boldsymbol{\theta})) \in B) \geq \gamma \quad \forall \boldsymbol{\theta} \in \Theta$$

Then:

$$C(\mathbf{X}) = \{\mathbf{g}(\boldsymbol{\theta}) : \boldsymbol{\pi}(\mathbf{X}_n, \mathbf{g}(\boldsymbol{\theta})) \in B\} \subseteq \mathbf{g}(\boldsymbol{\Theta})$$

is a confidence region with confidence level γ.

Confidence intervals for the relative frequency

Proposition 65. Let $(\mathcal{X}, \mathcal{F}, \{X_1, \ldots, X_n \sim \mathrm{Ber}(p) \text{ i.i.d.} : p \in (0, 1)\})$ be a parametric statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $(X_1, \ldots, X_n)$ and $\alpha \in [0, 1]$. Let $\hat{p} = \overline{x}_n$. Then, an asymptotic confidence interval for p of confidence level $1 - \alpha$ is:

$$p \in \left(\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Confidence intervals for $N(\mu, \sigma^2)$

Proposition 66 (Interval for μ with σ known). Let $\sigma \in \mathbb{R}_{\geq 0}$ be a known parameter, $(\mathcal{X}, \mathcal{F}, \{X_1, \ldots, X_n \sim N(\mu, \sigma^2) \text{ i.i.d.} : \mu \in \mathbb{R}\})$ be a parametric statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $(X_1, \ldots, X_n)$ and $\alpha \in [0, 1]$. Then, a confidence interval for μ of confidence level $1 - \alpha$ is:

$$\mu \in \left(\overline{x}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

Proposition 67 (Intervals for μ and σ^2). Let $(\mathcal{X}, \mathcal{F}, \{X_1, \ldots, X_n \sim N(\mu, \sigma^2) \text{ i.i.d.} : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{\geq 0}\})$ be a parametric statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $(X_1, \ldots, X_n)$ and $\alpha \in [0, 1]$. Then, a confidence interval for μ of confidence level $1 - \alpha$ is:

$$\mu \in \left(\overline{x}_n - t_{n-1;1-\frac{\alpha}{2}}\frac{\tilde{s}_n}{\sqrt{n}}, \overline{x}_n + t_{n-1;1-\frac{\alpha}{2}}\frac{\tilde{s}_n}{\sqrt{n}}\right)$$

A confidence interval for σ^2 of confidence level $1 - \alpha$ is:

$$\sigma^2 \in \left(\frac{(n-1){\tilde{s}_n}^2}{\chi_{n;1-\frac{\alpha}{2}}}, \frac{(n-1){\tilde{s}_n}^2}{\chi_{n;\frac{\alpha}{2}}}\right)$$

Confidence intervals for two-samples problems

Proposition 68 (Independent samples with known variances). Let $\sigma_x, \sigma_y \in \mathbb{R}_{\geq 0}$ be known parameters, $(\mathcal{X}, \mathcal{F}, \{X_1, \ldots, X_{n_x} \sim N(\mu_x, {\sigma_x}^2) \text{ i.i.d.}, Y_1, \ldots, Y_{n_y} \sim N(\mu_y, {\sigma_y}^2) \text{ i.i.d.} : (\mu_x, \mu_y,) \in \mathbb{R}^2\})$ be a parametric statistical model such that each X_i is independent of Y_j $\forall (i, j) \in \{1, \ldots, n_x\} \times \{1, \ldots, n_y\}$, $\mathbf{x}_{n_x} \in \mathcal{X}$ be a realization of $(X_1, \ldots, X_{n_x})$, $\mathbf{y}_{n_y} \in \mathcal{X}$ be a realization of $(Y_1, \ldots, Y_{n_y})$ and $\alpha \in [0, 1]$. Then, an asymptotic confidence interval for $\mu_x - \mu_y$ of confidence level $1 - \alpha$ is:

$$\mu_x - \mu_y \in \left(\overline{x}_{n_x} - \overline{y}_{n_y} - z_{1-\frac{\alpha}{2}}s, \overline{x}_{n_x} - \overline{y}_{n_y} + z_{1-\frac{\alpha}{2}}s\right)$$

where $s = \sqrt{\frac{{\sigma_x}^2}{n_x} + \frac{{\sigma_y}^2}{n_y}}$.

Proposition 69 (Independent samples with unknown equal variances). Let $(\mathcal{X}, \mathcal{F}, \{X_1, \ldots, X_{n_x} \sim N(\mu_x, \sigma^2) \text{ i.i.d.}, Y_1, \ldots, Y_{n_y} \sim N(\mu_y, \sigma^2) \text{ i.i.d.} : (\mu_x, \mu_y,) \in \mathbb{R}^2 \times {\mathbb{R}_{\geq 0}}^2\})$ be a parametric statistical model such that each X_i is independent of Y_j $\forall (i, j) \in \{1, \ldots, n_x\} \times \{1, \ldots, n_y\}$, $(x_1, \ldots, x_n) \in \mathcal{X}$ be a realization of $(X_1, \ldots, X_n)$, $(y_1, \ldots, y_n) \in \mathcal{X}$ be a realization of $(Y_1, \ldots, Y_n)$ and $\alpha \in [0, 1]$. Let ${\tilde{s}_{n_x}}^2 := \frac{1}{n_x - 1}\sum_{i=1}^n (x_i - \overline{x})^2$ and ${\tilde{s}_{n_y}}^2 := \frac{1}{n_y - 1}\sum_{i=1}^n (y_i - \overline{y})^2$. Then, an asymptotic confidence interval for $\mu_x - \mu_y$ of confidence level $1 - \alpha$ is:

$$\mu_x - \mu_y \in \left(\overline{x}_{n_x} - \overline{y}_{n_y} - t_{\nu;1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}, \right.$$
$$\left. \overline{x}_{n_x} - \overline{y}_{n_y} + t_{\nu;1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}\right)$$

where ${s_p}^2 = \frac{(n_x - 1){\tilde{s}_{n_x}}^2 + (n_y - 1){\tilde{s}_{n_y}}^2}{n_x + n_y - 2}$ and $\nu = n_x + n_y - 2$.

Proposition 70 (Independent samples with unknown variances). Let $(\mathcal{X}, \mathcal{F}, \{X_1, \ldots, X_{n_x} \sim N(\mu_x, {\sigma_x}^2) \text{ i.i.d.}, Y_1, \ldots, Y_{n_y} \sim N(\mu_y, {\sigma_y}^2) \text{ i.i.d.} : (\mu_x, \mu_y,) \in \mathbb{R}^2 \times {\mathbb{R}_{\geq 0}}^2\})$ be a parametric statistical model such that each X_i is independent of Y_j $\forall (i, j) \in \{1, \ldots, n_x\} \times \{1, \ldots, n_y\}$, $\mathbf{x}_{n_x} \in \mathcal{X}$ be a realization of $(X_1, \ldots, X_{n_x})$, $\mathbf{y}_{n_y} \in \mathcal{X}$ be a realization of $(Y_1, \ldots, Y_{n_y})$ and $\alpha \in [0, 1]$. Let ${\tilde{s}_{n_x}}^2 := \frac{1}{n_x - 1}\sum_{i=1}^n (x_i - \overline{x})^2$ and ${\tilde{s}_{n_y}}^2 := \frac{1}{n_y - 1}\sum_{i=1}^n (y_i - \overline{y})^2$. Then, an asymptotic confidence interval for $\mu_x - \mu_y$ of confidence level $1 - \alpha$ is:

$$\mu_x - \mu_y \in \left(\overline{x}_{n_x} - \overline{y}_{n_y} - t_{\nu;1-\frac{\alpha}{2}}\sqrt{\frac{{\tilde{s}_{n_x}}^2}{n_x} + \frac{{\tilde{s}_{n_y}}^2}{n_y}}, \right.$$
$$\left. \overline{x}_{n_x} - \overline{y}_{n_y} + t_{\nu;1-\frac{\alpha}{2}}\sqrt{\frac{{\tilde{s}_{n_x}}^2}{n_x} + \frac{{\tilde{s}_{n_y}}^2}{n_y}}\right)$$

where

$$\nu = \frac{\left(\frac{\tilde{s}_{n_x}{}^2}{n_x} + \frac{\tilde{s}_{n_y}{}^2}{n_y}\right)^2}{\frac{\left(\frac{\tilde{s}_{n_x}{}^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{\tilde{s}_{n_y}{}^2}{n_y}\right)^2}{n_y - 1}}$$

Proposition 71 (Related samples with unknown variances). Let $(\mathcal{X}, \mathcal{F}, \{X_1, \ldots, X_n \sim N(\mu_x, {\sigma_x}^2)$ i.i.d., $Y_1, \ldots, Y_n \sim N(\mu_y, {\sigma_y}^2)$ i.i.d. $: (\mu_x, {\sigma_x}^2, \mu_y, {\sigma_y}^2) \in \mathbb{R}^2 \times {\mathbb{R}_{\geq 0}}^2\})$ be a parametric statistical model such that each $W_i := X_i - Y_i \sim N(\mu_x - \mu_y, {\sigma_x}^2 - {\sigma_y}^2)$ are i.i.d., $(x_1, \ldots, x_n) \in \mathcal{X}$ be a realization of $(X_1, \ldots, X_n)$, $(y_1, \ldots, y_n) \in \mathcal{X}$ be a realization of $(Y_1, \ldots, Y_n)$ and $\alpha \in [0, 1]$. Then, we can proceed as if we only had the sample $(W_1, \ldots, W_n)$. In particular, a confidence interval for $\mu_x - \mu_y$ of confidence level $1 - \alpha$ is:

$$\mu_x - \mu_y \in \left(\overline{x}_n - \overline{y}_n - t_{n-1;1-\frac{\alpha}{2}} \frac{\hat{s}_n}{\sqrt{n}}, \right. \\ \left. \overline{x}_n - \overline{y}_n + t_{n-1;1-\frac{\alpha}{2}} \frac{\hat{s}_n}{\sqrt{n}}\right)$$

where $\hat{s}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - y_i - (\overline{x} - \overline{y}))^2$.

4. | Hypothesis testing

Hypothesis test

Definition 72. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model and $\Theta_0, \Theta_1 \subset \Theta$ be disjoint subsets. Our goal is to know whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$ (even if it isn't neither of them) and we will use a sample $\mathbf{x}_n \in \mathcal{X}$ to conclude our objective. We define the following two propositions which we will call *hypothesis*:

$$\mathcal{H}_0 : \theta \in \Theta_0 \qquad \mathcal{H}_1 : \theta \in \Theta_1$$

$\mathcal{H}_0$ is called *null hypothesis* and $\mathcal{H}_1$ is called *alternative hypothesis*. We say that the hypothesis $\mathcal{H}_i$ is *simple* if $\Theta_i = \{\theta_0\}$ for some $\theta_0 \in \Theta$. Otherwise we say that the hypothesis $\mathcal{H}_i$ is compound.

Definition 73 (Hypothesis test). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model. A *hypothesis test* is a function

$$\begin{aligned} \delta : \mathcal{X} &\longrightarrow \{\mathcal{H}_0, \mathcal{H}_1\} \\ \mathbf{x}_n &\longmapsto \delta(\mathbf{x}_n) \end{aligned}$$

The set $A_0 := \delta^{-1}(\mathcal{H}_0) \subseteq \mathcal{X}$, which is the set of samples that will lead us to accept[13] $\mathcal{H}_0$, is called *acceptation region*. The set $A_1 := \delta^{-1}(\mathcal{H}_1) \subseteq \mathcal{X}$, which is the set of samples that will lead us to accept $\mathcal{H}_1$ (and therefore reject $\mathcal{H}_0$), is called *critical region*[14].

Definition 74. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model and $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ be a hypothesis test. An *error of type I* is the rejection of $\mathcal{H}_0$ when it is true. An *error of type II* is the acceptation of $\mathcal{H}_0$ when it is false. We define the probabilities α and β as:

$$\begin{aligned} \alpha &:= \mathbb{P}(\text{Error of type I}) = \mathbb{P}(\text{Reject } \mathcal{H}_0 \mid \mathcal{H}_0 \text{ is true}) \\ \beta &:= \mathbb{P}(\text{Error of type II}) = \mathbb{P}(\text{Accept } \mathcal{H}_0 \mid \mathcal{H}_0 \text{ is false}) \end{aligned}$$

More precisely, if $\mathbf{x}_n \in \mathcal{X}$ is a realization of $\mathbf{X}_n$, then:

$$\alpha := \sup\{\mathbb{P}(\mathbf{x}_n \in A_1 \mid \theta) : \theta \in \Theta_0\}$$

The value $1 - \beta$ is called *power* of the test and the value α, *size* of the test. Moreover, we say that the test has *significance level* $\alpha \in [0, 1]$ if its size is less than or equal to α. In many cases the size of the test and the significance level are equal, hence the use of the same letter[15][16].

Definition 75. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model, $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ be a hypothesis test and $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$. We define the *power function* as:

$$\Pi(\theta) = \mathbb{P}(\text{Reject } \mathcal{H}_0) = \mathbb{P}(\mathbf{x}_n \in A_1)$$

Proposition 76. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$ and $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ be a hypothesis test. Then:

$$\Pi(\theta) = \begin{cases} \alpha & \text{if } \theta \in \Theta_0 \\ 1 - \beta & \text{if } \theta \in \Theta_1 \end{cases}$$

Test statistic and p-value

Definition 77. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$ and $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ be a hypothesis test. A statistic T used to decide whether or not reject the null hypothesis is called a *test statistic*.

Definition 78. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$, $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ be a hypothesis test such that $\Theta_0 = \{\theta_0\}$, and T be a test statistic. Suppose that we have observed the value $t := T(\mathbf{x}_n)$. We define the *p-value* as the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.

Definition 79. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$, $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ be a hypothesis test such that $\Theta_0 = \{\theta_0\}$. We say that the test is a

- *one-sided right tail test* if $\Theta_1 = (\theta_0, \infty)$.
- *one-sided left tail test* if $\Theta_1 = (-\infty, \theta_0)$.
- *two-sided test* if $\Theta_1 = \mathbb{R} \setminus \{\theta_0\}$.

[13] Some authors prefer to say that they don't reject $\mathcal{H}_0$ instead of saying that they accept $\mathcal{H}_0$.

[14] In order to denote these concepts more compactly, we will write $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ to denote the hypothesis test whose acceptation and critical regions are A_0 and A_1, respectively.

[15] In particular, for simple hypothesis they are the same thing.

[16] In practice, we fix a significance level $\alpha \in [0, 1]$ small enough (≈ 0.05 but may vary depending on the problem) with which we accept making mistakes and from here we try to minimize the β (or maximize the power $1 - \beta$). Moreover, having fixed α, we obtain a confidence level $1 - \alpha$. And if we impose $\mathbb{P}(\mathbf{x}_n \in A_0 \mid \mathcal{H}_0) = 1 - \alpha$, we are able to determine A_0.

Proposition 80. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta \subseteq \mathbb{R}\})$ be a statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$, $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ be a hypothesis test such that $\Theta_0 = \{\theta_0\}$, and T be a test statistic. Suppose that we have observed the value $t := T(\mathbf{x}_n)$ and let p be the p-value of the test. Then:

1. One-sided right tail test:
$$p = \mathbb{P}(T \geq t \mid \mathcal{H}_0)$$
2. One-sided left tail test:
$$p = \mathbb{P}(T \leq t \mid \mathcal{H}_0)$$
3. Two-sided test:
$$p = 2\min\{\mathbb{P}(T \geq t \mid \mathcal{H}_0), \mathbb{P}(T \leq t \mid \mathcal{H}_0)\}$$ [17]

And given a significance level $\alpha \in (0, 1)$ we will reject $\mathcal{H}_0$ if $p < \alpha$ and accept $\mathcal{H}_0$ if $p \geq \alpha$.

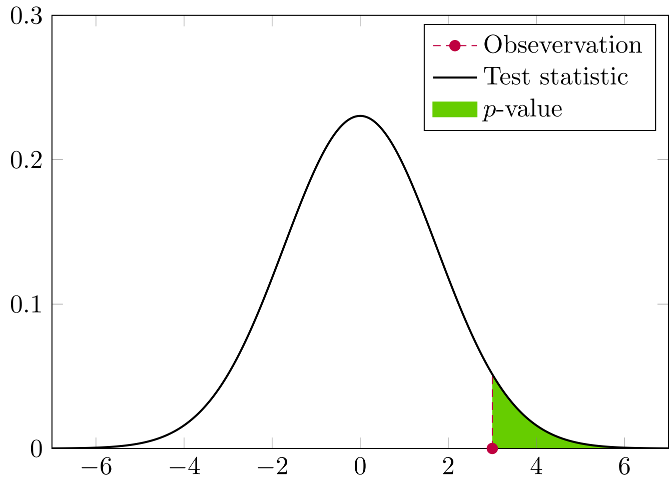

Figure 2: Probability density function of a test statistic (assuming the null hypothesis) together with an observed value and the p-value of the one-sided right tail test.

Neymann-Pearson test

Definition 81. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model. We say that a test $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ of significance level α is a *uniformly most powerful* (*UMP*) *test* if it has the greatest power among all the tests with significance level α.

Lemma 82. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model. We say that a test $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ of significance level α. If A_1 does not depend on the parameter $\theta \in \Theta_1$, then δ is a UMP test.

Definition 83 (Neymann-Pearson test). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\mathbf{X}_n} : \theta \in \Theta\})$ be a statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$ and $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ be a hypothesis test such that $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. We say that δ is a *Neyman-Pearson test* of significance level $\alpha \in [0, 1]$ if $\exists C > 0$ such that:
$$\left\{\mathbf{x}_n \in \mathcal{X} : \frac{L(\theta_0; \mathbf{x}_n)}{L(\theta_1; \mathbf{x}_n)} > C\right\} = A_0$$
$$\left\{\mathbf{x}_n \in \mathcal{X} : \frac{L(\theta_0; \mathbf{x}_n)}{L(\theta_1; \mathbf{x}_n)} \leq C\right\} = A_1$$
and $\mathbb{P}(\mathbf{x}_n \in A_1 \mid \mathcal{H}_0) = \alpha$.

Lemma 84 (Neyman-Pearson lemma). Any Neyman-Pearson test is a UMP test.

Theorem 85. Any UMP test is a Neyman-Pearson test.

Likelihood-ratio test

Definition 86 (Likelihood-ratio test). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_{\boldsymbol{\theta}}^{\mathbf{X}_n} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\})$ be a statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$, $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ be a hypothesis test of compound hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} \in \Theta_0$ and $\mathcal{H}_1 : \boldsymbol{\theta} \in \Theta_1$. Then, the *likelihood ratio test* (*LRT*) is given by the critical region:
$$\left\{\mathbf{x}_n \in \mathcal{X} : \frac{\sup\{L(\boldsymbol{\theta}; \mathbf{x}_n) : \boldsymbol{\theta} \in \Theta_0\}}{\sup\{L(\boldsymbol{\theta}; \mathbf{x}_n) : \boldsymbol{\theta} \in \Theta\}} \leq C\right\} = A_1$$
for some constant $C > 0$[18].

Proposition 87. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_{\boldsymbol{\theta}}^{\mathbf{X}_n} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\})$ be a statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $\mathbf{X}_n$, $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ be a hypothesis test of simple hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$. Then, the LRT is a Neyman-Pearson test.

Theorem 88 (Asymptotic behaviour of the LRT). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_{\boldsymbol{\theta}}^{\mathbf{X}_n} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\})$ be a parametric regular statistical model and consider the test $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ of compound hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} \in \Theta_0$ and $\mathcal{H}_1 : \boldsymbol{\theta} \in \Theta_1$. Let
$$\Lambda(\mathbf{x}_n) := -2\ln\frac{\sup\{L(\boldsymbol{\theta}; \mathbf{x}_n) : \boldsymbol{\theta} \in \Theta_0\}}{\sup\{L(\boldsymbol{\theta}; \mathbf{x}_n) : \boldsymbol{\theta} \in \Theta\}}$$
which is called *LRT test statistic*. Then if the model is regular, we have:
$$\Lambda(\mathbf{x}_n) \xrightarrow{\text{d}} {\chi_r}^2$$
where $r = \dim\Theta - \dim\Theta_0$.

Definition 89 (Goodness of fit). Suppose we have a random variable X whose outcomes are $x_1, \ldots, x_n$ and that we classify these outcomes in k classes. Thus, we obtain a table of the form:

Class	a_1	$\cdots$	a_j	$\cdots$	a_k	**Total**
Frequency	n_1	$\cdots$	n_j	$\cdots$	n_k	n

We want a test for:
$$\begin{cases} \mathcal{H}_0 : & X \sim f_{\boldsymbol{\theta}} \\ \mathcal{H}_1 : & P(X \in a_i) = \frac{n_i}{n} \quad \forall i \end{cases}$$
If π_i denotes the probability of being in the cell i under $\mathcal{H}_0$, we can approximate π_i by $\hat{\pi}_i = \mathbb{P}(X \in a_i \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}})$,

[17] If the statistic T is symmetric with respect to the origin, then $p = \mathbb{P}(|T| \geq |t| \mid \mathcal{H}_0)$.

[18] Note that $L(\hat{\boldsymbol{\theta}}, \mathbf{x}_n) = \sup\{L(\boldsymbol{\theta}; \mathbf{x}_n) : \boldsymbol{\theta} \in \Theta\}$, where $\hat{\boldsymbol{\theta}}$ is the MLE. Similarly $L(\hat{\boldsymbol{\theta}}_0, \mathbf{x}_n) = \sup\{L(\boldsymbol{\theta}; \mathbf{x}_n) : \boldsymbol{\theta} \in \Theta\}$, where $\hat{\boldsymbol{\theta}}_0$ is the MLE restricted to Θ_0.

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, for $i = 1, \dots, k-1$ and let $\hat{\pi}_k := 1 - \sum_{i=1}^{k-1} \hat{\pi}_i$. Hence since the distribution of the data in the table follows a multinomial distribution, we have[19]:

$$\Lambda = 2 \sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n\hat{\pi}_i}\right) \xrightarrow{\text{d}} {\chi_{k-2}}^2$$

Definition 90 (Test of homogenity). Consider r i.i.d. random variables $X_1, \dots, X_r$ whose outcomes can be classified in the classes $a_1, \dots, a_s$ each with probability $\mathbb{P}(X_i = a_j) = p_{ij}\ \forall i, j$. Suppose we have n_{ij} observations of the variable X_i taking the value a_j and denote $n_{i\cdot} := \sum_{j=1}^{s} n_{ij}$, $n_{\cdot j} := \sum_{i=1}^{r} n_{ij}$ and $n := \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$. That is, we have the following table:

	a_1	$\cdots$	a_j	$\cdots$	a_s	**Total**
X_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$n_{r\cdot}$
Total	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot s}$	n

Table 1

We want a test for:

$$\begin{cases} \mathcal{H}_0: & p_j := p_{1j} = \cdots = p_{rj}\ \forall j \\ \mathcal{H}_1: & \text{otherwise} \end{cases}$$

Again, the distribution of the data in the table follows a multinomial distribution, so under $\mathcal{H}_0$ we get the following MLEs (with the constraint that $\sum_{j=1}^{s} p_j = 1$):

$$\hat{p}_j = \frac{n_{\cdot j}}{n} \qquad \forall j$$

And in general, using the constraint $\sum_{i=1}^{r} \sum_{j=1}^{s} p_{ij} = 1$, we have:

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \qquad \forall i, j$$

Finally we have:

$$\Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} \log\left(\frac{n_{ij} n}{n_{i\cdot} n_{\cdot j}}\right) \xrightarrow{\text{d}} {\chi_{(r-1)(s-1)}}^2$$

Definition 91 (Test of independence). Consider r i.i.d. random variables $X_1, \dots, X_r$ whose outcomes can be classified in the classes $a_1, \dots, a_s$ each with probability $\mathbb{P}(X_i = a_j) = p_{ij}\ \forall i, j$. Suppose we have n_{ij} observations of the variable X_i taking the value a_j and denote $n_{i\cdot} := \sum_{j=1}^{s} n_{ij}$, $n_{\cdot j} := \sum_{i=1}^{r} n_{ij}$ and $n := \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$. That is, we have again the Table 1. We want a test for:

$$\begin{cases} \mathcal{H}_0: & p_{ij} = \theta_i \phi_j\ \forall i, j \\ \mathcal{H}_1: & \text{otherwise} \end{cases}$$

Again, the distribution of the data in the table follows a multinomial distribution, so under $\mathcal{H}_0$ we get the following MLEs for θ_i and ϕ_j (with the constraints that $\sum_{i=1}^{r} \theta_i = \sum_{j=1}^{s} \phi_j = 1$):

$$\hat{\theta}_i = \frac{n_{i\cdot}}{n} \quad \hat{\phi}_j = \frac{n_{\cdot j}}{n} \qquad \forall i, j$$

And in general, using the constraint $\sum_{i=1}^{r} \sum_{j=1}^{s} p_{ij} = 1$, we have:

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \qquad \forall i, j$$

Finally we have:

$$\Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} \log\left(\frac{n_{ij} n}{n_{i\cdot} n_{\cdot j}}\right) \xrightarrow{\text{d}} {\chi_{(r-1)(s-1)}}^2$$

t-test

Definition 92 (t-test). Let $(\mathcal{X}, \mathcal{F}, \{X_1, \dots, X_n \sim N(\mu, \sigma^2) \text{ i.i.d. } : \mu \in \mathbb{R}\})$ be a statistical model, $\mathbf{x}_n \in \mathcal{X}$ be a realization of $(X_1, \dots, X_n)$. The *t-test* is the test $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$, where $\mathcal{H}_0 : \mu = \mu_0$ for some $\mu_0 \in \mathbb{R}$ and $\mathcal{H}_1$ can be either of $\{\mu > \mu_0, \mu < \mu_0, \mu \neq \mu_0\}$. In this case the test statistic that is taken is:

$$\frac{\overline{X}_n - \mu}{\frac{\tilde{s}_n}{\sqrt{n}}} \sim t_{n-1}$$

Wald and score tests

Definition 93 (Wald test). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_{\theta}^{\mathbf{X}_n} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\})$ be a parametric regular statistical model and consider the test $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ of simple hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. The *Wald test* is the test whose statistic is:

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{\text{a}}{\sim} {\chi_d}^2$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ and $d = \dim \Theta$. If $\mathcal{H}_0 : \boldsymbol{\theta} \in \Theta_0$, we shall replace $\boldsymbol{\theta}_0$ by the MLE under $\mathcal{H}_0$, $\hat{\boldsymbol{\theta}}_{\mathbf{0}}$, in the test statistic. For the 1-dimensional case, we have:

$$I(\hat{\theta})(\hat{\theta} - \theta_0)^2 \overset{\text{a}}{\sim} {\chi_1}^2$$

Corollary 94. Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_{\theta}^{\mathbf{X}_n} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\})$ be a parametric regular statistical model and consider the test $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ of simple hypothesis $\mathcal{H}_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{r}$ and $\mathcal{H}_1 : \mathbf{R}\boldsymbol{\theta} \neq \mathbf{r}$, where $\mathbf{R} \in \mathcal{M}_{k \times d}(\mathbb{R})$, $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\mathbf{r} \in \mathbb{R}^k$. The test statistic of Wald test is:

$$(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r})^{\mathrm{T}} \left[\mathbf{R}\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1} \mathbf{R}^{\mathrm{T}}\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}) \overset{\text{a}}{\sim} {\chi_k}^2$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$. The matrix $\mathbf{R}$ is called *contrast matrix*.

Definition 95 (Score test). Let $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_{\theta}^{\mathbf{X}_n} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\})$ be a parametric regular statistical model and consider the test $\delta : \mathcal{X} = A_1 \sqcup A_2 \to \{\mathcal{H}_0, \mathcal{H}_1\}$ of simple hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. The *score test* is the test whose statistic is:

$$\mathbf{S}(\boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{I}^{-1}(\boldsymbol{\theta}_0) \mathbf{S}(\boldsymbol{\theta}_0) \overset{\text{a}}{\sim} {\chi_d}^2$$

[19] In order to have the expected asymptotic behaviour we need to check that the expectations of each $\hat{\pi}_i$ are greater than or equal to 5 (heuristic criterion), i.e. $n_i \hat{\pi}_i \geq 5\ \forall i$. If this is not the case, we should reduce the number of classes by groupping some of them together.

where $d = \dim \Theta$. If $\mathcal{H}_0 : \boldsymbol{\theta} \in \Theta_0$, we shall replace $\boldsymbol{\theta}_0$ by the MLE under $\mathcal{H}_0$, $\hat{\boldsymbol{\theta}}_0$, in the test statistic. For the 1-dimensional case, we have:

$$\frac{S(\theta_0)^2}{I(\theta_0)} \overset{a}{\sim} {\chi_1}^2$$

5. | Bootstrapping

Parametric and non-parametric bootstrap

Definition 96 (Non-parametric bootstrap). Consider a sample $x_1, \ldots, x_n$ from i.i.d. random variables $X_1, \ldots, X_n \sim F$ of an unknown distribution F. Our goal is to get an estimator of a parameter $\theta = T(X_1, \ldots, X_n)$. To do so, we define the *empirical distribution* F_n as follows: if $X \sim F_n$ then:

$$\mathbb{P}(X = x_i) = \frac{1}{n} \quad \forall i \in \{1, \ldots, n\}$$

First we compute $\hat{\theta} := T(x_1, \ldots, x_n)$ from the data given. Now we generate B samples of data $\{x_1^b, \ldots, x_n^b\}$, $b = 1, \ldots, B$, taken from the distribution F_n (with replacement) and for each sample we compute the respective estimator $\hat{\theta}_b = T(x_1^b, \ldots, x_n^b)$. This gives us the *bootstrap distribution* of $\hat{\theta}$ and so the bias and variance of $\hat{\theta}$ is:

$$\text{bias}_B(\hat{\theta}) = \overline{\theta}_B - \hat{\theta} \quad \text{Var}_B(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b - \overline{\theta}_B\right)^2$$

where $\overline{\theta}_B := \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b$. And hence, the bootstrap estimate of the standard error is:

$$\tilde{s}_B(\hat{\theta}) = \sqrt{\frac{\sum_{b=1}^{B} \left(\hat{\theta}_b - \overline{\theta}_B\right)^2}{B-1}}$$

If we want an unbiased estimator $\hat{\theta}$ of θ we can replace $\hat{\theta}$ by $2\hat{\theta} - \overline{\theta}_B$.

Definition 97 (Parametric bootstrap). Consider a sample $x_1, \ldots, x_n$ from i.i.d. random variables $X_1, \ldots, X_n \sim F_\theta$ and we are interested in estimating $\theta = T(X_1, \ldots, X_n)$. First we compute $\hat{\theta} := T(x_1, \ldots, x_n)$ from the data given. Now we generate B samples of data $\{x_1^b, \ldots, x_n^b\}$, $b = 1, \ldots, B$, taken from the distribution $F_{\hat{\theta}}$ (with replacement) and for each sample we compute the respective estimator $\hat{\theta}_b = T(x_1^b, \ldots, x_n^b)$. This gives us the *parametric bootstrap distribution* of $\hat{\theta}$[20].

Bootstrap confidence intervals

Definition 98 (Normal confidence interval). Consider a sample $x_1, \ldots, x_n$ from i.i.d. random variables $X_1, \ldots, X_n \sim F$ of an unknown distribution F. The *normal confidence interval* for $\theta = T(X_1, \ldots, X_n)$ of level α is:

$$\theta \in (\hat{\theta} - z_{1-\alpha/2} \tilde{s}_B(\hat{\theta}), \hat{\theta} + z_{1-\alpha/2} \tilde{s}_B(\hat{\theta}))$$

To use a bias-corrected bootstrap estimator, replace $\hat{\theta}$ by $2\hat{\theta} - \overline{\theta}_B$.

Definition 99 (Basic bootstrap confidence interval). Consider a sample $x_1, \ldots, x_n$ from i.i.d. random variables $X_1, \ldots, X_n \sim F$ of an unknown distribution F. The *basic bootstrap confidence interval* for $\theta = T(X_1, \ldots, X_n)$ of level α is:

$$\theta \in (2\hat{\theta} - \hat{\theta}_{1-\alpha/2}, 2\hat{\theta} - \hat{\theta}_{\alpha/2})$$

where $\hat{\theta}_\alpha$ is the sample quantiles of the bootstrap relicates.

Definition 100 (Bootstrap-t confidence interval). Consider a sample $x_1, \ldots, x_n$ from i.i.d. random variables $X_1, \ldots, X_n \sim F$ of an unknown distribution F. The *bootstrap-t confidence interval* for $\theta = T(X_1, \ldots, X_n)$ of level α is:

$$\theta \in (\hat{\theta} - t_{1-\alpha/2} \tilde{s}_B(\hat{\theta}), \hat{\theta} + t_{1-\alpha/2} \tilde{s}_B(\hat{\theta}))$$

To use a bias-corrected bootstrap estimator, replace $\hat{\theta}$ by $2\hat{\theta} - \overline{\theta}_B$.

Definition 101 (Percentile confidence interval). Consider a sample $x_1, \ldots, x_n$ from i.i.d. random variables $X_1, \ldots, X_n \sim F$ of an unknown distribution F. Once we have computed B estimates of $\hat{\theta}$ we order them:

$$\hat{\theta}_{(1)} \leq \cdots \leq \hat{\theta}_{(B)}$$

The *percentile confidence interval* for $\theta = T(X_1, \ldots, X_n)$ of level α is:

$$\theta \in (\hat{\theta}_{\alpha/2}, \hat{\theta}_{1-\alpha/2})$$

6. | Bayesian inference

Prior and posterior distributions

Definition 102. Let $\mathbf{X}$ be a random vector with pdf $f(\mathbf{x} \mid \theta)$. As always we would like to estimate the unknown parameter $\theta \in \Theta$. Bayesian approach to statistical inference treats the parameter θ as a random variable with an appropriate *prior distribution* (or simply *prior*) $f(\theta)$.

Definition 103. Let $\mathbf{X}$ be a random vector with pdf $f(\mathbf{x} \mid \theta)$, $f(\theta)$ be the prior of θ and $\mathbf{x}$ be a realization of $\mathbf{X}$. We define the *posterior distribution* (or simply *posterior*) of θ as the pdf $f(\theta \mid \mathbf{x})$ given by Bayes' theorem:

$$f(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta) f(\theta)}{f(\mathbf{x})} = \frac{f(\mathbf{x} \mid \theta) f(\theta)}{\int f(\mathbf{x} \mid \theta) f(\theta) \, d\theta}$$ [21]

If the prior and the posterior are of the same distribution type, prior and observation model are called *conjugate*.

Definition 104 (Bayesian point estimates). Let $\mathbf{X}$ be a random vector with pdf $f(\mathbf{x} \mid \theta)$, $f(\theta)$ be the prior of $\theta \in \Theta$ and $\mathbf{x}$ be a realization of $\mathbf{X}$.

- The *posterior mean* $\mathbb{E}(\theta \mid \mathbf{x})$ is:

$$\mathbb{E}(\theta \mid \mathbf{x}) = \int_\Theta \theta f(\theta \mid \mathbf{x}) \, d\theta$$

[20] Note that in order for bootstrapping to work we need some regular conditions of the function T (Hadamard differentiability). Statistics like minimum or maximum doesn't satisfy these restrictions.

[21] For practical purposes it is sometimes sufficient to study only $f(\mathbf{x} \mid \theta) f(\theta)$, since $f(\theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta) f(\theta)$.

- The *posterior mode* $\text{Mod}(\theta \mid \mathbf{x})$ is:
$$\text{Mod}(\theta \mid \mathbf{x}) = \arg\max\{f(\theta \mid \mathbf{x}) : \theta \in \Theta\}$$
- The *posterior median* $\text{Med}(\theta \mid \mathbf{x})$ is any number a such that:
$$\int_{-\infty}^{a} f(\theta \mid \mathbf{x})\,\mathrm{d}\theta \quad \text{and} \quad \int_{a}^{+\infty} f(\theta \mid \mathbf{x})\,\mathrm{d}\theta$$

Choice of the prior

Definition 105. Let $\theta \in \Theta$ be the parameter of interest of our model. A prior distribution with pdf $f(\theta)$ is called *flat* if
$$f(\theta) \propto \text{const.} \qquad \theta \in \Theta$$

Definition 106. Let $\theta \in \Theta$ be the parameter of interest of our model. A prior distribution with pdf $f(\theta) \geq 0$ is called *improper* if
$$\int_{\Theta} f(\theta)\,\mathrm{d}\theta = \infty \quad \text{or} \quad \sum_{\theta \in \Theta} f(\theta) = \infty$$
for continuous or discrete parameters θ, respectively.

Definition 107 (Jeffrey's prior). Let $\mathbf{X}$ be a random vector with pdf $f(\mathbf{x} \mid \theta)$ where $\theta \in \Theta$ is the parameter of interest. *Jeffrey's prior* is defined as:
$$f(\theta) \propto \sqrt{I(\theta)}$$
If $\boldsymbol{\theta}$ is a vector valued parameter, Jeffrey's prior is defined as:
$$f(\boldsymbol{\theta}) \propto \sqrt{\det \mathbf{I}(\boldsymbol{\theta})}$$

Proposition 108. Let $\mathbf{X}$ be a random vector with pdf $f(\mathbf{x} \mid \theta)$ where $\theta \in \Theta$ is the parameter of interest and $\eta = h(\theta)$ where h is an injective function. If the prior $f_\theta(\theta)$ of θ is flat, then:
$$f_\eta(\eta) \propto \left|\frac{\mathrm{d}h^{-1}(\eta)}{\mathrm{d}\eta}\right|$$
where $f_\eta(\eta)$ is the prior of η. And so, $f_\eta(\eta)$ is flat if and only if h is a linear transformation.

Proposition 109. Let $\mathbf{X}$ be a random vector with pdf $f(\mathbf{x} \mid \theta)$ where $\theta \in \Theta$ is the parameter of interest and $\eta = h(\theta)$ where h is an injective function. Suppose $f_\theta(\theta)$ is the prior of θ and $f_\eta(\eta)$ is the prior of η. If $f_\theta(\theta) \propto \sqrt{I(\theta)}$, then $f_\eta(\eta) \propto \sqrt{I(\eta)}$

Properties of Bayesian point and interval estimates

Definition 110. Let $\theta \in \Theta$ be the parameter of interest of our model. A *loss function* $\ell(\hat{\theta}, \theta) \in \mathbb{R}$ quantifies the loss encountered when estimating the true parameter θ by $\hat{\theta}$. Commonly used loss functions are the following ones:

- *Quadratic loss function*: $\ell(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$
- *Linear loss function*: $\ell(\hat{\theta}, \theta) = \left|\hat{\theta} - \theta\right|$
- *Zero-one loss function*:
$$\ell_\varepsilon(\hat{\theta}, \theta) = \begin{cases} 0 & \text{if } \hat{\theta} = \theta \\ 1 & \text{if } \hat{\theta} \neq \theta \end{cases}$$

Definition 111. Let $\mathbf{X}$ be a random vector with pdf $f(\mathbf{x} \mid \theta)$, $f(\theta)$ be the prior of $\theta \in \Theta$ and $\mathbf{x}$ be a realization of $\mathbf{X}$. A *Bayes estimate* of θ with respect to a loss function $\ell(\hat{\theta}, \theta)$ minimizes the expected loss with respect to the posterior distribution $f(\theta \mid \mathbf{x})$, i.e. it minimizes:
$$\mathbb{E}(\ell(\hat{\theta}, \theta) \mid \mathbf{x}) = \int_{\Theta} \ell(\hat{\theta}, \theta) f(\theta \mid \mathbf{x})\,\mathrm{d}\theta$$

Proposition 112.

1. The posterior mean is the Bayes estimate with respect to quadratic loss.
2. The posterior median is the Bayes estimate with respect to linear loss.
3. The posterior mode is the Bayes estimate with respect to zero-one loss.

Theorem 113. Let $\mathbf{X}$ be a random vector (of length n) with pdf $f(\mathbf{x} \mid \theta)$ where $\theta \in \Theta$ is the parameter of interest which has a prior $f(\theta)$. Suppose the MLE of θ is $\hat{\theta}_n$. We define:
$$m_0 := \arg\max\{f(\theta) : \theta \in \Theta\}, \quad J_0 = -\left.\frac{\partial^2(\log f(\theta))}{\partial \theta^2}\right|_{\theta = m_0}$$
Then:
$$\theta \mid \mathbf{X} \overset{a}{\sim} N(m_n, {J_n}^{-1})$$
where:
$$J_n = J_0 + J(\hat{\theta}_n) \quad \text{and} \quad m_n = \frac{J_0 m_0 + J(\hat{\theta}_n)\hat{\theta}_n}{J_n}$$
If n is large enough, we will have:
$$\theta \mid \mathbf{X} \overset{a}{\sim} N\left(\hat{\theta}_n, I(\hat{\theta}_n)^{-1}\right)$$

Definition 114. Let $\mathbf{X}$ be a random vector with pdf $f(\mathbf{x} \mid \theta)$ where $\theta \in \Theta$ is the parameter of interest. A subset $C \subseteq \Theta$ is called a *credible region* for θ of confidence γ if:
$$\int_C f(\theta \mid \mathbf{x})\,\mathrm{d}\theta = \gamma$$
If C is a real interval, C is also called *credible interval*. If the parameter is discrete, we will define a credible region of confidence γ as:
$$\sum_{\theta \in C \cap \Theta} f(\theta \mid \mathbf{x}) \geq \gamma$$

Definition 115. Let $\mathbf{X}$ be a random vector with pdf $f(\mathbf{x} \mid \theta)$ where $\theta \in \Theta$ is the parameter of interest. A γ credible region C is called a *highest posterior density* (HPC) *region* if $f(\theta \mid \mathbf{x}) \geq f(\tilde{\theta} \mid \mathbf{x})$ $\forall \theta \in C$ and all $\tilde{\theta} \notin C$.

Proposition 116. Let $\mathbf{X}$ be a random vector with pdf $f(\mathbf{x} \mid \theta)$ where $\theta \in \Theta$ is the parameter of interest whose posterior is $f(\theta \mid \mathbf{x})$. Then, among all γ credible region C, the HPC minimizes the expected loss for the function $\ell(C, \theta) = |C| - \mathbf{1}_C(\theta)$, where $|C|$ is the size of the region.

7. | Analyzing data

Comparizing distributions

Definition 117 (Q-Q plots). Consider a sample $x_1, \dots, x_n$ of data and the ordered sample $x_{(1)}, \dots, x_{(n)}$. We would like to know whether the data come from a distribution F or not. To do conclude something, we define:

$$y_{(k)} = y_k := F^{-1}\left(\frac{k}{n+1}\right) \quad k = 1, \dots, n$$

Then, we plot the pairs $(y_{(k)}, x_{(k)})$ (or equivalently the pairs $(F(y_{(k)}), F(x_{(k)}))$). The more similar is the plot to a line, the more possible is for the data to come from the distribution F. This plot is called *Quantile-Quantile plot* (or *Q-Q plots*), $y_{(k)}$ are called the *theoretical quantiles* and $x_{(k)}$ the *sample quantiles*, that is, the quantile function of the discrete distribution induced by the sample (assigning probability $1/n$ to each data of the sample)[22].

Proposition 118 (Normal Q-Q plots). Consider a sample $x_1, \dots, x_n$ of data and the ordered sample $x_{(1)}, \dots, x_{(n)}$. We would like to know whether the data come from a normal distribution $N(\mu, \sigma^2)$ or not. We define

$$z_{(k)} = z_{\frac{k}{n+1}} = \Phi^{-1}\left(\frac{k}{n+1}\right) \quad k = 1, \dots, n$$

If the data are reasonably normal, the Q-Q plot of the pairs $(z_{(k)}, x_{(k)})$ is approximately a line of slope σ and y-intercept μ.

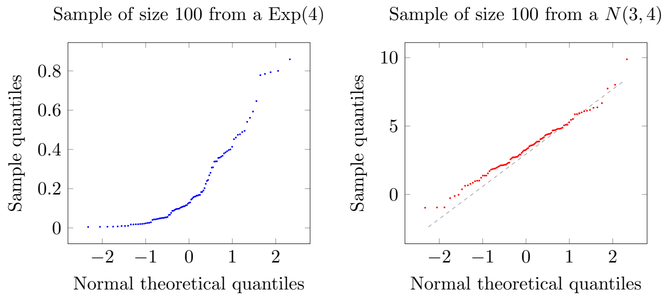

Figure 3: Two normal Q-Q plots of samples of size 100. On the right-hand side the sample comes from an exponential distribution and so we can see that the data doesn't fit quite well in a line. On the left-hand side the data come from a normal distribution and so the data is approximately fitted in a line whose slope is $2.369 \approx \sqrt{4}$ and whose y-intersect is $2.956 \approx 3$.

[22]Sometimes the theoretical quantiles taken are $F^{-1}\left(\frac{k-0.5}{n}\right)$ instead of $F^{-1}\left(\frac{k}{n+1}\right)$.