Humboldt-Universität zu Berlin

Methods of Statistics (M25)

Lecturer: Prof. Dr. Sven Wang
WS 24/25

Last Update: Tuesday 5th November, 2024

Contents

1 Intro and Disclaimer

These lecture notes are based on the material presented by Professor Wang during class and are written by students. They may contain errors or omissions. Please refer to the in-person lectures and the literature on Moodle for accurate and authoritative information. If you find an error or want to help, please send an email to:

said.kassner@student.hu-berlin.de

stephensonmonroe@gmail.com

salihiad@hu-berlin.de

2 Basic Statistical Concepts

Lecture 1

Here is the literature that the class is based on.

Literature:

- WS 19/20 R. Altmeyer *"Gliederung Methoden der Statistik"*
- L. Wasserman, *All of Statistics*
- M. Trabs, K. Krenz, M. Jirak and M. Reiss. *Methoden der Statistik.*
- Hastie, Tibshirani, et al., *Elements of Statistical Learning*

Let's start with a (simplest possible) example:

Example 1 (Polling). Consider a poll with two answers A and B (representing political parties).

- N = total number of votes
- M = total number of votes supporting party A

Poll Definitions:

- n = size of the poll
- $x = (x_1, ..., x_n)$ = responses, where:

$$x_i = \begin{cases} 0 & \text{if the i-th person supports B} \\ 1 & \text{if the i-th person supports A} \end{cases}$$

Additional Assumptions:

- n-times, we select a person randomly from the set $\{1, ..., N\}$, and record their (truthful) response.
- Every asked person responds (i.e., no selection bias).
- People can be asked repeatedly.

Aim of the Poll: The aim of the poll is to estimate the fraction of party A supporters. This can be written as:

$$\theta = \frac{M}{N} \in [0,1]$$

An intuitive estimate of θ is:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Question: Is this a good (or best possible) estimator? What properties does it have?

To answer this, we formalize some statistical notions.

Definition 1 (Sample space). A *sample space* is a measurable space $(\mathcal{X}, \mathcal{F})$, i.e., a set $\mathcal{X}$ with a σ-algebra $\mathcal{F}$, in which our statistical observations take values.

Definition 2 (Statistical model). Let $(\mathcal{X}, \mathcal{F})$ be some sample space and let Θ be a set, which we call the *parameter space.* A *statistical model* is a family of probability measures $\{P_\theta : \theta \in \Theta\}$ on $(\mathcal{X}, \mathcal{F})$.

Remark 1. Often, $(\mathcal{X}, \mathcal{F})$ is a "product space." For example, in Example **??**, $\mathcal{X} = \{0,1\}^n$, and each P_θ is a product distribution, i.e., $x_1, \ldots, x_n$ are independent, identically distributed. Then we say $\{P_\theta : \theta \in \Theta\}$ is an *iid statistical model.*

*Remark 2 (Back to Example **??**).* Here:

- $\mathcal{X} = \{0,1\}^n$
- $\Theta = [0,1]$
- $\mathcal{F} = \mathcal{P}(\{0,1\}^n)$
- $P_\theta = (\text{Bernoulli}(\theta))^{\otimes n}$

Remark 3. If every person could only be asked once, we would have $P_\theta = \text{Hypergeometric}(N, M, n)$, which "converges" to the Bernoulli model as $N, M \to \infty$. We might have to discretize Θ and take $\theta = \frac{M}{N}$. (Exercise: Think about it!)

3 Parameter Estimation

Assume that $\Theta \subseteq \mathbb{R}^p$, for $p \geq 1$. This is the setting of parametric statistics. [Assume Θ is measurable.]

Definition 3 (Estimator). An estimator for $\theta \in \Theta$ is any measurable function:

$$\hat{\theta} : (\mathcal{X}, \mathcal{F}) \to \Theta.$$

Any function that, based on some data $x \in \mathcal{X}$, outputs a guess / estimate $\hat{\theta}(x) \in \Theta$.

Lecture 2

Last time: Statistical model = family of probability measures on $(\mathcal{X}, \mathcal{F})$ indexed by $\theta \in \Theta$.

Sample space: $(\mathcal{X}, \mathcal{F})$

Estimator: = measurable function $(\mathcal{X}, \mathcal{F}) \to \Theta$

Now, what are some desirable properties we would like to have?

Definition 4 (Unbiased estimator). Let $\Theta = \mathbb{R}^p$ (measurable), $p \geq 1$. An estimator $\hat{\theta}$ is unbiased if

$$\mathbb{E}_\theta[\hat{\theta}] = \mathbb{E}_{\mathbb{P}_\theta}[\hat{\theta}] = \theta, \text{ for all } \theta \in \Theta.$$

Where $\mathbb{E}_\theta[\cdot] = \mathbb{E}_{\mathbb{P}_\theta}[\cdot]$ denotes expectation under the law $\mathbb{P}_\theta$.

In more explicit terms:

$$\mathbb{E}_{x \sim \mathbb{P}_\theta}[\hat{\theta}(x)] = \theta \quad \forall \theta$$

Remark 4 (Unbiasedness). Unbiasedness means ”no systematic errors.” However, we'd also like a ”good” $\hat{\theta}$ to be concentrated around the data-generating parameter.

Definition 5 (Consistent estimator). Let $(\mathbb{P}_\theta^n)_{\theta \in \Theta}$ be a sequence of statistical models ($n \geq 1$), on the same parameter space Θ not depending on $n \geq 1$.

Let $\hat{\theta}_n$ be a sequence of estimators. Then $\hat{\theta}_n$ is called consistent if for every $\theta \in \Theta$,

$$\hat{\theta}_n \xrightarrow{\mathbb{P}_\theta^n} \theta$$

or explicitly, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}_\theta^n(|\hat{\theta}_n - \theta| > \epsilon) = 0.$$

Back to Example ??:

- $X_i = \{0,1\}^n$
- $\Theta = [0,1]$
- $\mathbb{P}_\theta^n = \text{Bernoulli}(\theta)^{\otimes n}$
- $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$

Unbiasedness:

Let $\theta \in \Theta$, then

$$\mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta[X_i] = \frac{1}{n}\sum_{i=1}^n \theta = \theta.$$

Thus, $\mathbb{E}_\theta[\hat{\theta}_n] = \theta$.

Consistency:

- We could use the Weak Law of Large Numbers (WLLN).
- Alternatively,

  $$\text{Var}_\theta(\hat{\theta}_n) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}_\theta(X_i) = \frac{1}{n^2}\sum_{i=1}^n \theta(1-\theta) = \frac{\theta(1-\theta)}{n}$$

  which tends to zero as $n \to \infty$.

It follows: For every $\epsilon > 0$,

$$\mathbb{P}_\theta^n(|\hat{\theta}_n - \theta| > \epsilon) = \mathbb{P}_\theta^n(|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]| > \epsilon)$$

By Markov's inequality:

$$\mathbb{P}_\theta^n(|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]| > \epsilon) \leq \frac{\mathbb{E}_\theta\left[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2\right]}{\epsilon^2} = \frac{\text{Var}_\theta(\hat{\theta}_n)}{\epsilon^2} = \frac{\theta(1-\theta)}{n\epsilon^2}$$

which tends to zero as $n \to \infty$. Thus,

$$(\hat{\theta}_n : n \geq 1) \text{ is consistent.} \quad \square$$

3.1 Maximum Likelihood Principle

Is there another way to motivate $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$? Yes, it turns out it is the maximum likelihood estimator, i.e.,

MLE = ”parameter which assigns the highest probability to the observed data.”

In our example, each $\mathbb{P}_\theta^n$ has a probability density (likelihood)

$$\mathbb{P}_\theta^n(x) = \prod_{i=1}^n \mathbb{P}_\theta(x_i) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$$
$$= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Fixing $x \in \{0,1\}^n$ and maximizing in $\theta \in [0,1]$ gives the following:

- If $\sum_{i=1}^n x_i = 0$, then $\hat{\theta}_n = 0$ is the maximizer.
- If $\sum_{i=1}^n x_i = n$, then $\hat{\theta}_n = 1$ is the maximizer.
- If $\hat{\theta}_n \in \{1,\dots,n-1\}$, then writing $S_n = \sum_{i=1}^n x_i$ gives:

$$\frac{\partial}{\partial\theta}\mathbb{P}_\theta^n(x) = S_n\theta^{S_n-1}(1-\theta)^{n-S_n-1} - (n-S_n)\theta^{S_n}(1-\theta)^{n-S_n-1} = 0$$
$$\Leftrightarrow S_n(1-\theta) - \theta(n-S_n)$$
$$\Leftrightarrow \theta = \frac{S_n}{n}. \quad \square$$

Definition 6 (Dominated statistical model & MLE). A model $(\mathbb{P}_\theta^n)_{\theta\in\Theta}$ is called dominated if there exists a measure μ on $(\mathcal{X},\mathcal{F})$ such that for every $\theta \in \Theta$, $\mathbb{P}_\theta \ll \mu$ or equivalently (by Radon-Nikodym), for all $\theta \in \Theta$, there is a probability density $\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}$ of $\mathbb{P}_\theta$ with respect to μ.

The MLE is defined as any $\hat{\theta} \in \Theta$ that maximizes the function

$$\theta \mapsto \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}(x) = \mathbb{P}_\theta(x).$$

Remark 5 (Caveats).

- MLE might not be unique.
- MLE might not exist.
- It’s not always clear that some selection $\hat{\theta}(x) \in \arg\max_\theta \mathbb{P}_\theta(x)$ is a measurable function of $x \in \mathcal{X}$. However, there are measurable selection theorems that permit a measurable choice of $\hat{\theta}$ under very general conditions.

Remark 6. In all the models we study, we will work with the Lebesgue measure (for continuous data) or the counting measure (for discrete data).

Example 2 (Normal model). Consider random samples $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ for some unknown $\mu \in \mathbb{R}$, $\sigma^2 > 0$, and let λ denote the Lebesgue measure on $\mathbb{R}^n$.

$$\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty).$$

Then the likelihood is:

$$\begin{aligned} L(\mu, \sigma^2 \mid x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{X_i - \mu}{\sigma}\right)^2\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\|X - \mu \cdot 1_n\|^2\right), \end{aligned}$$

where by 1_n we denote the vector of ones of dimension n.

Here, the MLE is given as:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{[Sample mean]}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 \quad \text{[Sample variance]}.$$

$$\mathbb{E}_\theta[\hat{\mu}] = \mu, \quad \mathbb{E}_\theta[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2, \quad \text{so } \hat{\sigma}^2 \text{ is biased.}$$

⇒ MLE is not always unbiased, and not always a ”good” method.

3.2 Bayesian method

Motivation In Bayesian statistics, a key element is the prior distribution, which we denote by π, reflecting our ”beliefs” about the parameter $\theta \in \Theta$ before observing data (π is a probability measure on Θ).

A prior π, together with a model $(P_\theta : \theta \in \Theta)$, gives rise to a joint probability distribution for the pair $(\theta, x) \in \Theta \times \mathcal{X}$.

The Bayesian approach bases statistical inference on the posterior distribution of θ conditioned on x.

Joint probability:

$$(\theta, x) \mapsto \pi(\theta) P_\theta(x)$$

conditional distribution of $x \mid \theta$

Posterior:

$$\pi(\theta \mid x) = \frac{\pi(\theta)P_\theta(x)}{\int_\Theta \pi(\theta)P_\theta(x)\,d\theta}$$

Remark 7. Bayesian methods automatically generate "error bars" because the posterior is not an estimator but a whole probability distribution.

Lecture 3

Definition 7 (Prior, Posterior, Bayes' Rule). Let $\mathcal{F}_\Theta$ be a σ-algebra on Θ, and suppose

$$\{P_\theta : \theta \in \Theta\}$$

is a dominated statistical model with densities $p_\theta(x)$, and assume that

$$(\theta, x) \mapsto p_\theta(x)$$

is "jointly measurable" (i.e., w.r.t. $\sigma(\mathcal{F}_\Theta \times \mathcal{F})$).

Let Π be a prior distribution on Θ, with density $\pi(\theta)$ w.r.t. measure $\nu(\cdot)$. Then, define the posterior density

$$\pi(\theta \mid x) := \frac{p_\theta(x)\pi(\theta)}{\int_\Theta p_{\tilde{\theta}}(x)\,d\Pi(\tilde{\theta})}.$$

The corresponding probability measure $\Pi(\cdot \mid x)$ is called the posterior distribution:

$$\Pi(B \mid x) = \int_B \pi(\theta \mid x)\,d\nu(\theta), \quad B \in \mathcal{F}_\Theta.$$

$$= \frac{\int_B p_\theta(x)\,\pi(\theta)\,d\nu(\theta)}{\int_\Theta p_{\tilde{\theta}}(x)\,d\Pi(\tilde{\theta})},$$

$$= \frac{\int_B p_\theta(x)\,d\Pi(\theta)}{\int_\Theta p_{\tilde{\theta}}(x)\,d\Pi(\tilde{\theta})},$$

Remark 8. Think of $\Theta \subseteq \mathbb{R}^p$, $\nu(\cdot)$ as a Lebesgue measure, $\pi(\cdot)$ as a Lebesgue density.

Exception: $\Theta = \{0, 1\}$ in hypothesis testing. Then, we'd take $\nu(\cdot)$ to be the counting measure.

From the posterior, we can derive several estimators:

- **Maximum-a-posterior (MAP) estimator:**

$$\hat{\theta}_{\text{MAP}}(x) = \operatorname*{argmax}_{\theta \in \Theta} \pi(\theta \mid x).$$

- **Posterior mean:** Say $\Theta \subseteq \mathbb{R}^p$ convex

$$\hat{\theta}(x) = \int_{\Theta} \theta \, \pi(\theta \mid x) \, d\nu(\theta) \in \mathbb{R}^p.$$

Back to Example 1.1: *Binomial model:* $\mathcal{X} = \{0, 1, \ldots, n\}$, $p_\theta = \mathrm{Bin}(n, \theta)$, $\theta \in \Theta = [0, 1]$.

Prior (uniform): $\Pi = \mathrm{Unif}(0, 1)$.

We know:

$$\hat{\theta}_{\mathrm{MAP}} = \hat{\theta}_{\mathrm{MLE}} \quad \text{(for the uniform prior)},$$

$$\hat{\theta}_{\mathrm{MAP}} = \frac{X}{n}.$$

- **Posterior mean**:

$$\pi(\theta|x) = \frac{p_\theta(x)}{\int p_{\tilde{\theta}}(x) d\tilde{\theta}} \propto \binom{n}{k} \theta^x (1-\theta)^{n-x}.$$

- **Binomial distribution**:

$$\mathrm{Bin}(n, p)(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

  where $k \in \{0, \ldots, n\}$ is the number of successes, and p is the probability of success, and n would be interpreted as the ”number of coin flips”.

$$\pi(\theta|x) \propto \theta^x (1-\theta)^{n-x}.$$

and

$$\int_0^1 \pi(\theta \mid x) d\nu = 1$$

We conclude that $\pi(\theta|x)$ is a **Beta-distribution** on $[0, 1]$,

$$\mathrm{Beta}(x+1, n-x+1).$$

The mean is given by:

$$\hat{\theta} = \frac{x+1}{n+2}.$$

Remark 9 (Beta distribution). The Beta distribution is defined as:

$$\mathrm{Beta}(a, b), \quad a, b \geq 0.$$

The probability density function of the Beta distribution is given by:

$$P_{\mathrm{Beta}(a,b)}(x) = x^a(1-x)^b.$$

Definition 8 (Conjugate Bayesian models). Let $(P_\theta : \theta \in \Theta)$ be a statistical model. Then, some family $\mathcal{D}$ of p.m.s on Θ is called *conjugate* if

$$\Pi \in \mathcal{D} \implies \Pi(\cdot|x) \in \mathcal{D} \quad \text{for all } x \in \mathcal{X}.$$

Examples:

- $(\mathrm{Bin}(n,\theta)) : \theta \in [0,1], \quad \mathcal{D} = \mathrm{Beta}(a,b),\ a,b \geq 0.$
- $(\mathcal{N}(\mu,\sigma^2)) : \mu \in \mathbb{R}, \quad \mathcal{D} = \left\{\mathcal{N}\left(\mu, n^2\right), \mu \in \mathbb{R}, n^2 > 0\right\},\ \sigma^2$ known.

4 Decision Theory

Here suppose that $\Theta \subseteq \mathbb{R}^p$.

Definition 9 (Loss function). A function $\ell : \Theta \times \mathbb{R}^p \to [0, \infty)$ is a *loss function* if for every $\theta \in \Theta$, $\ell(\theta, \cdot)$ is measurable. Given some estimator $\hat{\theta}$, the expected loss is

$$\mathcal{R}(\theta, \hat{\theta}) = \mathbb{E}_\theta \left[\ell(\theta, \hat{\theta}) \right].$$

Example: Take $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_{\mathbb{R}^p}^2$. Then,

$$\mathcal{R}(\theta, \hat{\theta}) = \mathbb{E}_\theta \left[\|\theta - \hat{\theta}\|_{\mathbb{R}^p}^2 \right]$$

is the mean squared error (MSE).

Proposition 1 (Bias-Variance Decomposition). *Let $\hat{\theta} \in L^2(\mathbb{P}_\theta)$. Then it holds that:*

$$R(\hat{\theta}, \theta) = (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2 + \mathrm{Var}_\theta(\hat{\theta}).$$

Proof. We have

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}] + \mathbb{E}_\theta[\hat{\theta}] - \theta)^2]$$

Expanding the squared term:

$$= \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2] + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2 + 2\mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])(\mathbb{E}_\theta[\hat{\theta}] - \theta)].$$

Since $\mathbb{E}_\theta[\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]] = 0$, the last term vanishes, leaving us with:

$$R(\hat{\theta}, \theta) = \mathrm{Var}_\theta(\hat{\theta}) + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2.$$

Definition 10 (Minimax Risk). Given an estimator $\hat{\theta}$ in a model $(\mathbb{P}_\theta : \theta \in \Theta)$, the maximal risk of $\hat{\theta}$ is

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}).$$

The minimax risk of a model $(\mathbb{P}_\theta : \theta \in \Theta)$ is given as

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta}),$$

where the infimum is taken over all estimators. An estimator is called minimax if

$$\sup_{\theta\in\Theta} R(\theta,\hat{\theta}) = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} R(\theta,\hat{\theta}).$$

Definition 11 (Bayes Risk). Given a prior π on Θ, the π-Bayes risk of a decision rule δ for the loss function L is defined as

$$R_\Pi(\delta) = \mathbb{E}_\Pi[R(\delta,\theta)] = \int_\Theta R(\delta,\theta)\pi(\theta)d\theta = \int_\Theta \int_{\mathcal{X}} L(\delta(x),\theta)\pi(\theta)p_\theta(x)dxd\theta.$$

A Π-Bayes decision rule $\hat{\theta}_\Pi$ is any decision rule that minimizes $R_\Pi(\hat{\theta})$.

SW: ℓ instead of L below, $p_\theta(x)$ instaed of $f(x,\theta)$ Note: Has been corrected.

Definition 12 (Posterior Risk). For a Bayesian model, the posterior risk R_Π is defined as the average loss under the posterior distribution for some observation $x \in \mathcal{X}$:

$$R_{\Pi(\cdot|x)}(\delta) = \mathbb{E}_\Pi[\ell(\delta(x),\theta)|x].$$

Here, the notation $\mathbb{E}_\Pi[\cdot|x]$ stands for the expectation under the posterior distribution.

Proposition 2 (Bayes Risk and Posterior Risk). *An estimator δ that minimizes the Π-posterior risk R_Π also minimizes the π-Bayes risk R_π.*

Proof. The π-Bayes risk can be rewritten as

$$\begin{aligned}
R_\pi(\delta) &= \int_\Theta \mathbb{E}_\theta[\ell(\delta(X),\theta)]\pi(\theta)d\theta \\
&= \int_\Theta \int_{\mathcal{X}} \ell(\delta(x),\theta)\pi(\theta)p_\theta(x)dxd\theta \\
&= \int_{\mathcal{X}} \int_\Theta \ell(\delta(x),\theta)\frac{p_\theta(x)\pi(\theta)}{\int_\Theta p_{\theta'}(x)\pi(\theta')d\theta'} \times \underbrace{\int_\Theta p_{\theta'}(x)\pi(\theta')d\theta'}_{=:n(x)\geq 0} dxd\theta \\
&= \int_{\mathcal{X}} \mathbb{E}_\Pi[\ell(\delta(x),\theta)|x]n(x)dx.
\end{aligned}$$

[Notation $n(x)$ motivated by the word ‘normalising constant’].

Let δ_Π be a decision rule that minimizes the posterior risk, i.e., such that for all $x \in \mathcal{X}$,

$$\mathbb{E}_\Pi[\ell(\delta_\Pi(x),\theta)|x] \leq \mathbb{E}_\Pi[\ell(\delta(x),\theta)|x].$$

Multiplying by $m(x) \geq 0$ and integrating on both sides over $\mathcal{X}$ yields the desired result.□

Example 3. For the quadratic risk with the squared-loss, the posterior risk is minimized by taking $\delta(X) = \mathbb{E}_\Pi[\theta|X]$, by minimizing the quadratic function in δ. Other losses will give other ways to minimize the posterior risk, and other Bayes decision rules.

Proposition 3. *Let $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model and let $\hat{\theta}$ be an estimator. Then we have*

$$\sup_{\theta\in\Theta} R(\hat{\theta}, \theta) = \sup_{\Pi} \int_\Theta R(\hat{\theta}, \theta)\, \Pi(d\theta),$$

where the supremum is taken over all prior distributions Π*.*

Proof. Obviously, we have

$$\int_\Theta R(\hat{\theta}, \theta)\, \Pi(d\theta) \leq \sup_{\theta\in\Theta} R(\hat{\theta}, \theta).$$

On the other hand, by using the prior distributions δ_θ (Dirac measure on $\theta \in \Theta$), we obtain

$$\sup_{\Pi} \int_\Theta R(\hat{\theta}, \theta)\, \Pi(d\theta) \geq \int_\Theta R(\hat{\theta}, \theta)\, \delta_\theta(d\theta) = R(\hat{\theta}, \theta).$$

[Note: In the following we use the notation δ for decision rules while on the blackboard we used $\hat{\theta}$ or $\tilde{\theta}$. If you want to adjust this please contact me so that I can give you access.]

Proposition 4. *Let π be a prior on Θ such that*

$$R_\pi(\delta_\pi) = \sup_{\theta\in\Theta} R(\delta_\pi, \theta),$$

where δ_π is a π-Bayes rule. Then it holds that

1. *The rule δ_π is minimax.*
2. *If δ_π is unique Bayes, then it is unique minimax.*

Proof. Let δ be any decision rule. Then

$$\sup_{\theta\in\Theta} R(\delta, \theta) \geq \mathbb{E}_\pi[R(\delta, \theta)],$$

$$\int_\Theta R(\delta, \theta)\pi(\theta)\, d\theta \geq \mathbb{E}_\pi[R(\delta, \theta)],$$

$$\int_\Theta R(\delta, \theta)\pi(\theta)\, d\theta = R_\pi(\delta_\pi) = \sup_{\theta\in\Theta} R(\delta_\pi, \theta).$$

Taking the infimum over δ gives the result.

2. If δ_π is unique Bayes, the second inequality is strict for any $\delta' \neq \delta_\pi$. □

Corollary 1. *If a (unique) Bayes rule δ_π has constant risk in θ, then it is (unique) minimax.*

Proof. If a Bayes rule δ_π has constant risk, then

$$R_\pi(\delta_\pi) = \mathbb{E}_\pi[R(\delta_\pi, \theta)] = \sup_{\theta \in \Theta} R(\delta_\pi, \theta),$$

where $R(\delta_\pi, \theta)$ is constant in θ. Uniqueness of the Bayes rule implies uniqueness of the minimax rule, as in part 2 of the former proposition. □

Example 4. Hence, if the maximal risk of a Bayes rule δ_π equals the Bayes risk, then π is least favorable, and the corresponding Bayes rule is minimax.

- In a $\mathrm{Bin}(n, \theta)$ model, let $\pi_{a,b}$ be a $\mathrm{Beta}(a, b)$ prior on $\theta \in [0, 1]$. Then the unique Bayes rule for $\pi_{a,b}$ over the quadratic risk is the posterior mean $\delta_{a,b} = \bar{\theta}_{a,b}$. Trying to solve the equation
$$R(\delta_{a,b}, \theta) = \text{const.} \quad \forall \theta \in [0, 1]$$
we can find a prior π_{a^*,b^*} and a corresponding Bayes rule $\delta_{\pi_{a^*,b^*}}$ of constant risk. It is therefore unique minimax, and different from the MLE (see Examples sheet).
- In a $\mathcal{N}(\theta, 1)$ model, $\bar{X}_n$ is minimax, as proved later.

4.1 Another optimality concept: Admissibility

Definition 13. A decision rule δ is *inadmissible* if there exists δ' such that

$$R(\delta', \theta) \leq R(\delta, \theta) \quad \forall \theta \in \Theta \quad \text{and} \quad R(\delta', \theta) < R(\delta, \theta) \quad \text{for some } \theta \in \Theta.$$

Remark 10.

- The intuition is that there is no reason to choose an inadmissible estimator or decision rule: it would always be better to choose another estimator that dominates it.
- Admissibility is not the only criterion to evaluate an estimator: In most cases, a constant estimator will be admissible for the quadratic risk, but it is often not reasonable.

Proposition 5.

1. *A unique Bayes rule is admissible.*
2. *If δ is admissible and has constant risk, then it is minimax.*

Proof may be done in the Examples sheet.

Definition 14. For a vector $X \in \mathbb{R}^p$, the *James–Stein estimator* is defined as

$$\delta^{JS}(X) = \left(1 - \frac{p-2}{\|X\|^2}\right) X.$$

In a Gaussian model $X \sim \mathcal{N}(\theta, I_p)$ for $\theta \in \mathbb{R}^p$ (with a single observation, to simplify notation), the risk of the MLE is given by

$$R(\hat{\theta}_{\mathrm{MLE}}, \theta) = \mathbb{E}_\theta[\|X - \theta\|^2] = \sum_{j=1}^{p} \mathbb{E}_\theta[(X_j - \theta_j)^2] = p.$$

For $X \sim \mathcal{N}(\theta, I_p)$ with $p \geq 3$, the risk of the James–Stein estimator satisfies for all $\theta \in \mathbb{R}^p$

$$R(\delta^{JS}, \theta) < p.$$

5 Confidence Sets

Definition 15 (Confidence Set). Let $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model. For a given $\alpha \in [0,1]$, consider sets $C_{1-\alpha}(x) \subseteq \Theta$ for each $x \in \mathcal{X}$. Then $C_{1-\alpha}(x)$ is called a random confidence set at level $1-\alpha$ (or with coverage probability $1-\alpha$) if

$$\forall\theta \in \Theta : \mathbb{P}_\theta(\theta \in C_{1-\alpha}) = \mathbb{P}_\theta(\{x \in \mathcal{X} : \theta \in C_{1-\alpha}(x)\}) \geq 1-\alpha.$$

Note: The following example was only started in Lecture 4 and may be fully covered in Lecture 5.

Example 5. Consider the statistical model $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), (\mathbb{P}_p)_{p\in[0,1]})$ with $\mathbb{P}_p = \mathrm{Ber}(p)^{\otimes n}$ and independent observations $X_k \sim \mathrm{Ber}(p)$, $k \in \{0,\dots,n\}$. We are looking for a confidence interval $C_{1-\alpha}$ around $\hat{p} = \overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$, i.e.,

$$C_{1-\alpha} = [\overline{X}_n - a, \overline{X}_n + b] \quad \Big(C_{1-\alpha}(x) = [\overline{X}_n(x) - a(x), \overline{X}_n(x) + b(x)]\Big)$$

should satisfy (where a and b might be random) that

$$1-\alpha \leq \mathbb{P}_p\left(p \in C_{1-\alpha}\right) = \mathbb{P}_p\left(\overline{X}_n - a \leq p \leq \overline{X}_n + b\right) = \mathbb{P}_p\left(-b \leq \overline{X}_n - p \leq a\right).$$

Let $t \mapsto F_p^n(t) := \mathbb{P}_p\left(\overline{X}_n - p \leq t\right)$ be the distribution function. Then,

$$\begin{aligned}\mathbb{P}_p\left(-b \leq \overline{X}_n - p \leq a\right) &= \mathbb{P}_p\left(\overline{X}_n - p \leq a\right) - \mathbb{P}_p\left(\overline{X}_n - p < -b\right)\\ &= F_p^n(a) - F_p^n(-b) + R_n,\end{aligned}$$

where $R_n = \mathbb{P}_p\left(\overline{X}_n - p = -b\right)$. Choose a, b as quantiles of $\mathbb{P}_p^n$, i.e., $a = \left(F_p^n\right)^{-1}(1-\alpha/2)$ and $-b = \left(F_p^n\right)^{-1}(\alpha/2)$ (with quantile function $t \mapsto \left(F_p^n\right)^{-1}(t) := \inf\{t \in \mathbb{R} : F_p^n(t) \geq t\}$).

However, F_p^n and thus a, b are unknown. Consider two possibilities:

Normal Approximation. It holds that $\mathbb{E}_p^n[X_k] = p$, $\sigma := \mathrm{Var}_p^n(X_k) = p(1-p)$. By the central limit theorem, we have

$$\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - p\right) = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{X_k - p}{\sigma} \xrightarrow{d} N(0,1), \quad n \to \infty.$$

For $Z \sim N(0,1)$, it holds that

$$F_p^n(a) = \mathbb{P}_p^n\left(\overline{X}_n - p \leq a\right) = \mathbb{P}_p^n\left(\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - p\right) \leq \frac{\sqrt{n}}{\sigma}a\right) \approx \mathbb{P}(|Z| \leq \frac{\sqrt{n}}{\sigma}a)$$

$$= \Phi\left(\frac{\sqrt{n}}{\sigma}a\right) = \Phi(z_\beta)$$

for $a := \frac{\sigma}{\sqrt{n}}z_\beta$ (where z_β is the β-quantile of the $N(0,1)$-distribution, i.e., $\Phi(z_\beta) = \beta$). In particular, $R_n = o(1)$ (i.e., $R_n \to 0$ as $n \to \infty$). For $a = b$ (since the $N(0,1)$-distribution is symmetric) and because $\Phi(-x) = 1 - \Phi(x)$, it follows that

$$\mathbb{P}_p^n(p \in C_{1-\alpha}) = F_p^n(a) - F_p^n(-a) + R_n \approx \Phi(z_\beta) - (1 - \Phi(z_\beta)) + o(1) = 2\Phi(z_\beta) - 1 + o(1).$$

For $\beta = 1 - \alpha/2$, $C_{1-\alpha}$ is an asymptotically correct confidence interval. However, p and therefore σ and a are unknown. Solutions:

- Estimate $\sigma = p(1-p) \leq 1/4$ to widen the confidence interval.
- For the empirical variance $\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2$, we have $\hat{\sigma}^2 \to \sigma^2$ almost surely (by the law of large numbers). Using Slutsky's lemma (Lemma: For random variables $(X_n, Y_n)_{n\geq 1}$ with $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} c \in \mathbb{R}$ (where c is deterministic), it holds that $X_n + Y_n \xrightarrow{d} X + c$ and $X_n \cdot Y_n \xrightarrow{d} c \cdot X$), it follows that

$$\frac{\sqrt{n}}{\hat{\sigma}}\left(\overline{X}_n - p\right) = \frac{\sigma}{\hat{\sigma}} \cdot \frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - p\right) \xrightarrow{d} N(0,1), \quad n \to \infty.$$

From this, we derive $a = \frac{\hat{\sigma}}{\sqrt{n}}z_{1-\alpha/2}$ (randomly chosen).

$$\mathbb{P}_p^n(p \in C_{1-\alpha}) = \mathbb{P}_p^n\left(|\overline{X}_n - p| \leq a\right) = \mathbb{P}_p^n\left(\left|\frac{\sqrt{n}}{\hat{\sigma}}(\overline{X}_n - p)\right| \leq z_{1-\alpha/2}\right)$$

$$\approx \mathbb{P}(|Z| \leq z_{1-\alpha/2}) = 2\Phi(z_{1-\alpha/2}) - 1 = 1 - \alpha.$$

5.1 Hypothesis Testing

5.1.1 Basic Definitions

Let $(P_\theta : \theta \in \Theta)$ be a statistical model, and let $\Theta = \Theta_0 \cup \Theta_1$ be a partition:

- A **statistical test** is a measurable function of the data $\varphi : (\mathcal{X}, \mathcal{F}) \to [0,1]$.
- If $\varphi(x) \in \{0,1\}$ for all $x \in \mathcal{X}$, then φ is a **non-randomized test**; otherwise, it is **randomized**.
- $H_0 : \theta \in \Theta_0$ is the **null hypothesis**.
- $H_1 : \theta \in \Theta_1$ is the **alternative hypothesis**.
- The map $\theta \to \beta_\varphi(\theta) = P_\theta(\varphi = 1)$ is called the **power function** of a test φ.

5.1.2 Type I and Type II Errors

- For $\theta \in \Theta_0$, $\beta_\varphi(\theta)$ represents the **Type I error** (wrongly rejecting the null).
- For $\theta \in \Theta_1$, $1 - \beta_\varphi(\theta)$ represents the **Type II error** (failing to reject the alternative when it is true).

$$1 \qquad \beta_\varphi(\theta) \quad 0 \qquad \Theta_0 \qquad \Theta_1 \qquad \Theta$$

Note:

$$1 - P_\theta(\varphi = 1) = P_\theta(\varphi = 0) = P_\theta \text{ (wrongly accepting the null)}$$

5.1.3 Level and Uniformly Most Powerful Tests

Definition 16 (Level). A test $\varphi : \mathcal{X} \to [0,1]$ has **level** $\alpha \in [0,1]$ if

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

Definition 17 (Uniformly Most Powerful Test). Given a level $\alpha \in (0,1)$, $\varphi : \mathcal{X} \to [0,1]$ is called **uniformly most powerful (UMP)** if, for every other test φ' of level α and all $\theta \in \Theta_1$,

$$\beta_\varphi(\theta) \geq \beta_{\varphi'}(\theta).$$

5.1.4 The Neyman-Pearson Lemma

The Neyman-Pearson Lemma provides a basis for constructing the most powerful tests for simple hypotheses:

Theorem 1 (Neyman-Pearson Lemma). *Let $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ be simple hypotheses:*

1. ***Existence:*** *There exists a test φ and a constant $k \in [0, \infty)$ such that $P_{\theta_0}(\varphi = 1) = \alpha$, with*

$$\varphi(x) = \begin{cases} 1, & \textit{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k \\ 0, & \textit{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < k \end{cases}$$

   Here, p_{θ_1} and p_{θ_0} are densities with respect to some dominated measure μ.

2. ***Sufficiency:*** *If φ satisfies $P_{\theta_0}(\varphi = 1) = \alpha$ and the above form, then φ is a UMP level α test.*
3. ***Necessity:*** *If φ_k is UMP for level α, then it must be of the form shown above.*

5.1.5 Proof of the Neyman-Pearson Lemma

1. Define the likelihood ratio $r(x) = \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \in [0, \infty)$. Let F_0 be the CDF of $r(x)$ under P_{θ_0}.

$$F_0(t) = P_{\theta_0}(r(x) \leq t).$$

   Define $\alpha(t) = 1 - F_0(t) = P_{\theta_0}(r(x) > t)$ and note:

   - α is right-continuous:

$$\lim_{\epsilon \to 0} \alpha(t + \epsilon) = P_{\theta_0}(r(x) > t).$$

   - α is non-increasing.
   - α has left limits.

   α **is cadlag:** It is continuous from the right and has a left limit.

$$\text{There exists } k \in [0, \infty) \text{ such that } \alpha \leq \alpha(k^-) \quad \text{and} \quad \alpha \geq \alpha(k).$$

   We define the test

$$\varphi(x) = \begin{cases} 1 & \text{if } r(x) > k, \\ \gamma & \text{if } r(x) = k, \\ 0 & \text{if } r(x) < k. \end{cases}$$

Set $\gamma = \frac{\alpha - \alpha(k)}{\alpha(k^-) - \alpha(k)}$.

The level of φ is

$$E_{\theta_0}[\varphi(x)] = P_{\theta_0}(\varphi(x) = 1).$$

$$= P_{\theta_0}(r(x) > k) + P_{\theta_0}(r(x) = k) \cdot \gamma = \alpha.$$

6 Lecture 6: Neyman-Pearson Lemma and Likelihood Ratio Tests

Neyman-Pearson Lemma

Power of a Test

The **power** of a test is defined as:

$$E_{\theta_1}[\varphi] = P_{\theta_1}(\varphi = 1)$$

Likelihood Ratio Test

The **likelihood ratio** is given by:

$$\Lambda(x) = \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = r(x)$$

Likelihood Ratio (LR) Test

The LR test is defined as:

$$\varphi(x) = \begin{cases} 1 & \text{if } r(x) > k, \\ \gamma & \text{if } r(x) = k, \\ 0 & \text{if } r(x) < k, \end{cases}$$

where $k \in [0, \infty)$ and $\gamma \in [0, 1]$.

Note: LR tests are Uniformly Most Powerful (UMP) for simple hypothesis testing:

- Given a significance level α, if the LR test satisfies $E_{\theta_0}[\varphi] = \alpha$, it controls the Type I error.
- The LR test minimizes the Type II error:

$$E_{\theta_1}[\varphi] \geq E_{\theta_1}[\varphi'] \quad \forall \varphi'$$

Continuation of Proof (Part of UMP)

Let φ' be another level α test such that $E_{\theta_0}[\varphi'] \leq \alpha$.

Goal: Show that $E_{\theta_1}[\varphi] \geq E_{\theta_1}[\varphi']$.

Let μ be the dominating measure. Consider:

$$\int (\varphi(x) - \varphi'(x))\,(p_{\theta_1}(x) - kp_{\theta_0}(x))\,d\mu(x) = 0$$

Claim: $p \geq 0$.

Observation:

- If $p_{\theta_1}(x) - kp_{\theta_0}(x) > 0$, then $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k$, implying $\varphi(x) = 1$.
- If $p_{\theta_1}(x) - kp_{\theta_0}(x) < 0$, then $\varphi(x) = 0$.
- If $p_{\theta_1}(x) - kp_{\theta_0}(x) = 0$, then the integrand is 0.

Thus, $p = 0$, leading to:

$$\int (\varphi - \varphi')p_{\theta_1}\,d\mu = \int (\varphi - \varphi')p_{\theta_0}\,d\mu = k\,[E_{\theta_0}[\varphi] - E_{\theta_0}[\varphi']] \geq 0$$

Therefore,

$$E_{\theta_1}[\varphi] \geq E_{\theta_1}[\varphi']$$

Part (3) UMP $\Rightarrow$ LR

Assume φ^* is a UMP test with $E_{\theta_0}[\varphi^*] = \alpha$. Let φ be the LR test satisfying $E_{\theta_0}[\varphi] = \alpha$.

Goal: Show that $\varphi = \varphi^*$ almost everywhere except on $\{r(x) = k\}$.

Define the sets:

$$x^+ = \{x : \varphi(x) > \varphi^*(x)\}, \quad x^- = \{x : \varphi(x) < \varphi^*(x)\}, \quad x^0 = \{x : \varphi(x) = \varphi^*(x)\}$$

$$\tilde{x} = (x^+ \cup x^-) \cap \{x : p_{\theta_1}(x) \neq kp_{\theta_0}(x)\}$$

It suffices to show $\mu(\tilde{x}) = 0$.

On $\tilde{x}$:

$$(\varphi - \varphi^*)(p_{\theta_1} - kp_{\theta_0}) > 0$$

If $\mu(\tilde{x}) > 0$, then:

$$\int_{\mathcal{X}} (\varphi - \varphi^*)(p_{\theta_1} - kp_{\theta_0})\,d\mu \geq 0$$

$$\int_{\tilde{x}} (\varphi - \varphi^*)(p_{\theta_1} - kp_{\theta_0})\,d\mu \geq 0$$

However,

$$E_{\theta_1}[\varphi] - E_{\theta_1}[\varphi^*] > k\,[E_{\theta_0}[\varphi] - E_{\theta_0}[\varphi^*]] \geq 0$$

This leads to a contradiction, implying $\mu(\tilde{x}) = 0$ and thus $\varphi = \varphi^*$ almost everywhere.

Example: Gaussian Location Model

Consider:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

Testing:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1, \quad \mu_0 < \mu_1$$

The likelihood ratio is:

$$\frac{p_1(X_1, \ldots, X_n)}{p_0(X_1, \ldots, X_n)} = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_1)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right)$$

Simplifying:

$$= \exp\left(-\frac{n}{2\sigma^2}(\mu_1^2 - \mu_0^2) - \frac{2(\mu_1 - \mu_0)}{\sigma^2}\sum_{i=1}^{n} X_i\right) \geq K_\alpha$$

This implies:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geq K_\alpha, \quad \text{for some } K_\alpha \in \mathbb{R}$$

To determine K_α:

$$\bar{X}_n := \frac{1}{n}\sum X_i \overset{H_0}{\sim} \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right)$$

Thus:

$$P_{H_0}\left(\bar{X}_n \geq K_\alpha\right) = 1 - \Phi\left(\frac{\sqrt{n}}{\sigma}(K_\alpha - \mu_0)\right)$$

Solving for K_α:

$$K_\alpha = \mu_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \alpha)$$

Therefore, the LR test is:

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if } \bar{X}_n \geq \mu_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \alpha), \\ 0 & \text{otherwise.} \end{cases}$$

Corollary

Consider simple hypothesis testing. Let φ be a UMP test at level α. Then:

$$\alpha = E_{H_0}[\varphi] \leq E_{\theta_1}[\varphi]$$

Suppose $E_{\theta_1}[\varphi] = E_{\theta_1}[\varphi_0]$. Then φ_0 is also UMP, implying φ_0 is an LR test:

$$\varphi_0 = \begin{cases} 1 & \text{if } \frac{p_{\theta_1}}{p_{\theta_0}} \geq K \quad \text{a.s., for some } K, \\ 0 & \text{otherwise.} \end{cases}$$

Since $\varphi_0 \in \{\varphi, \beta\}$, it follows that $p_{\theta_1} = K p_{\theta_0}$ almost surely.

Moreover:

$$\int p_{\theta_0}\, d\mu = K \int p_{\theta_0}\, d\mu = 1 \quad \Rightarrow \quad K = 1$$

Correspondence Theorem

Statement: There is a correspondence between tests and confidence regions.

$$\text{Tests} \quad \longleftrightarrow \quad \text{Confidence regions } C(x)$$

with

$$\Pr_\theta(\theta \in C(x)) \geq 1 - \alpha$$

and

$$\Pr_\theta(\phi_\theta(x) = 1) = \alpha$$

Theorem: Let $\{P_\theta : \theta \in \Theta\}$ be a statistical model and $\alpha \in (0,1)$.

(i) If $C = C(X)$ is a level-α confidence set, then

$$\phi_{\theta_0}(x) = \mathbb{I}\{\theta_0 \notin C(x)\}$$

is a level-α test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

(ii) If $\{\phi_\theta : \theta \in \Theta\}$ is a family of level-α tests, then

$$C(X) = \{\theta \in \Theta : \phi_\theta(X) = 0\}$$

is a $1 - \alpha$ confidence set.

Proof:

(i)

$$\Pr_{\theta_0}(\phi_{\theta_0} = 1) = \Pr_{\theta_0}(\theta_0 \notin C(X)) \leq \alpha$$

(ii)

$$\Pr_{\theta}(\theta \notin C(X)) = \Pr_{\theta}(\phi_\theta(X) = 1) \leq \alpha$$

UMPT Tests in Models with Monotone Likelihoods

Proposition: Let $\Theta \subseteq \mathbb{R}$. Consider testing:

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0,$$

for some $\theta_0 \in \mathbb{R}$.

Assume there exists a test statistic $T : X \to \mathbb{R}$ and a function $h : \mathbb{R} \times \Theta \times \Theta \to \mathbb{R}$ such that:

$$\frac{P_\theta(X)}{P_{\tilde{\theta}}(X)} = h(T(X), \theta, \tilde{\theta})$$

and for all $\theta \geq \tilde{\theta}$, the function $t \mapsto h(t, \theta, \tilde{\theta})$ is monotone increasing.

Conclusion: LR tests are also UMP for level α. Specifically, the LR test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ for any $\theta_1 > \theta_0$ will be UMP.