Summary

▼ Hypothesis testing

▼ testing statistical hypothesis

▼ Setup

Let $X \sim F$ be a r.v. with $F \in \mathcal{F}$ unknown ($\mathcal{F}$ space of all possible cdfs) . Then we define two sets $H_0, H_1 \subset \mathcal{F}$ and test if:

$$F \in H_0 \text{ or } F \in H_1$$

Hypothesis (H_0, H_1) could be

- simple
  - $|H_0| = 1$
- composite
  - $|H_0| > 1$

▼ Range

The distributions in H_0 and H_1 imply a **range** of outcomes $R(H_0), R(H_1)$. The range is all possible outcomes under the distributions of H_1 and H_0.

▼ Critical region

A **critical region** is a set of outcomes C_r such that iff $x \in C_r$ we reject H_0.

The acceptance region is the **compliment of the critical region $C_a = C_r^c$.**

We therefore partitioning the sample space into two sets

▼ test statistic

Sometimes it is hard to find C_r for a given signficance level α (more on this in type one and type two error section) We therefore look at the **test-statistic**

Let $t : \mathbb{R}^k \to \mathbb{R}$. Then we call the test stastitic $t(\vec{X})$ and $C_r = \{x : t(x) \in C_r^T\}$.

- type one and type two error

  The type one error is the false rejection of H_0:

  $$\alpha = P(\text{type one error }) = P(X \in C_r | F \in H_0)$$

  The type two error is the false acceptance of H_0:

  $$\beta = P(\text{type two error}) = P(X \in C_a | F \in H_1)$$

Test decision	True probability distribution: H is true	True probability distribution: H is not true
Rejection of H	Type I Error	correct decision
Acceptance of H	correct decision	Type II Error

- properties
  - Parametric vs non-parametric

    Parametric: $H_0 = \{f(x;\theta) | \theta \in \Theta\}$ characterized by a set of **parameters**

    Non-parametric: everything else

  - power function
    - Parametric definition

      The powerfunction is given by the probability that an outcome in C_r is realized:

      $$\pi(\theta) = P(X \in C_r | \theta)$$

      If $f(x;\theta) \in H_0$ then $\pi(\theta) = P(\text{type one error})$ and if $f(x;\theta) \in H_1$ we can say $\pi(\theta) = P(\text{correct decision})$.

      The **ideal power function is:**

      $$\pi_o(\theta) = \begin{cases} 1 \text{ if } f(x;\theta) \in H_1 \\ 0 \text{ else} \end{cases}$$

- Inadmissability

  A set C_r is **inadmissable iff** there exists an alternative critical region C_r^* such that the power-function of this alternative is better than the on by C_r.

  We than say, that the test with C_r is **dominated** by C_r^*.

- Size of a test

  We call the size α of a test:

$$\alpha = \sup_{\theta \in H_0} \pi(\theta) = \sup_{\theta \in H_0} P(X \in C_r | \theta)$$

- Uniformly most powerful tests

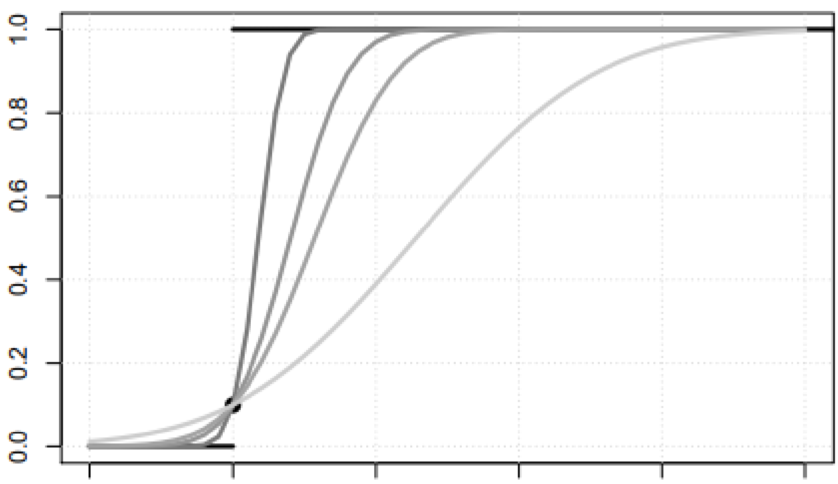
  Basically the best C_r possible.

- Consistency

  Let C_{rn} be the critical region induced by the sample $(X_1, ..., X_n)$. Let the significance level be fixed at $\alpha > 0$. Then the test is called **consistent** iff:

$$\pi_{C_{rn}}(\theta) \longrightarrow 1 \quad \forall \theta \in H_1$$

  Plotting the sequence of induced power-functions could look smth like:



- asymptotic distributions
- test statistics
  - Likelihood ratio

    We want to compare, how likely $\theta \in H_0$ is compared to $\theta \in H_0 \cup H_1$. Therefore we construct the test statistic:

    $$\lambda(x) = \frac{\sup\limits_{\theta \in H_0} \mathcal{L}(\theta; x)}{\sup\limits_{\theta \in H_0 \cup H_1} \mathcal{L}(\theta; x)}$$

    Where $\mathcal{L}$ is the generalized Likelyhood function. The critical region is defined by a $c \in \mathbb{R}$ where the test rejects H_0 iff $\lambda(x) \leq c$.

    - asymptotic null distribution

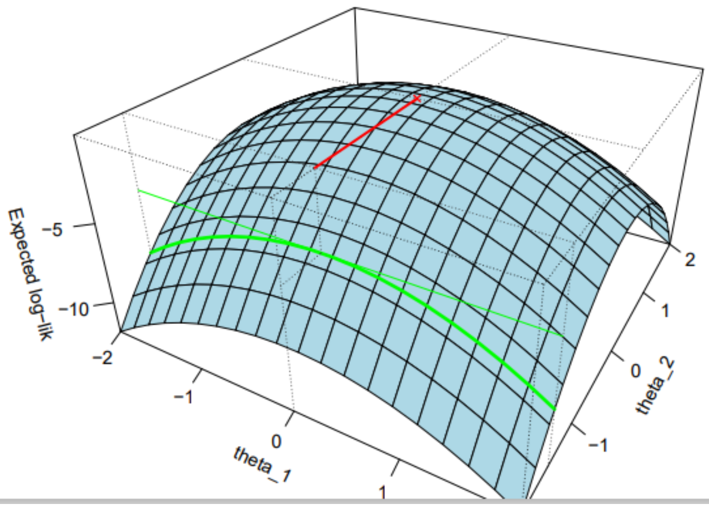
      If we have a restriction of $H_0 : R(\theta) = R$ and $H_1 : R(\theta) \neq R$. Where $R(\theta)$ is a smooth q-dimensional vector function, then under H_0 it holds that:

      $$-2 \log \lambda(X) \xrightarrow{d} \mathcal{X}_q^2$$

  - Wald

    This test looks at the distance between the restricted and unrestricted estimate.

Wald's idea



This is basically a t-test. This is why it is not further discussed in the lecture.

- Lagrange multiplier (or score test)

  Looks at the gradient of the restricted maximum of the log likelyhood.

  Looking at the restricted maximization problem:

  $$\max_{\theta,\lambda}\{\ln \mathcal{L}(\theta; x) - \lambda'[R(\theta) - R]\}$$

  If the restriction $R(\theta) - R$ is binding, the λ' should be non zero. So only if the restriction is keeping you away from the maximum, you get Lagrange coefficients not equal to zero

  - what if the restricted maximum is a local maximum?

    This why we look at different tests for log likelyhoodss

- multiple testing

- Confidence Intervalls
  - basic setup

    Let there be a “data dependent interval $[t_1(X), t_2(X)]$ such that $P(\theta \in [t_1(X), t_2(X)])$ becomes maximal and $t_2(X) - t_1(X)$ becomes minimal.

    We can generalize this idea to confidence sets $C(X)$. So choose $C(X)$ such that $P(\theta \in C(X))$ (this metric is called the **coverage probability**) becomes maximal and the volume of $C(X)$ becomes minimal

  - confidence level

    The confidence level is the smallest possible coverage probability:

    $$CL_C = \inf_{\theta \in \Theta} P_\theta(\theta \in C(X))$$

  - Pivotal variables

    Choose a transformation $T(X_1, ..., X_n, \theta)$, that is not dependent on θ. Then do the following steps:

    1. Find $a, b \in \mathbb{R}$ such that: $P(a \leq T \leq b) = 1 - \alpha$ for all $\theta \in \Theta$.
    2. “Invert” T such that $\theta = t^*(X, T)$.

3. Construct $C(X)$ with $t^*(X, a)$ and $t^*(X, b)$.

We covered an example with the empirical variance with iid sample $(X_1, ..., X_n)$ with $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then it holds that:

$$S_n^2 := \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \implies nS_n^2/\sigma^2 \sim \mathcal{X}_{n-1}^2$$

But then it holds that:

$$P\left(\frac{nS_n^2}{\mathcal{X}_{n-1;\alpha}^2} \leq \sigma^2 \leq \frac{nS_n^2}{\mathcal{X}_{n-1;1-\alpha}^2} \right) = 1 - \alpha$$

But because this is the confidence level, we at the same time form a critical region for a test of size α with the compliment of this interval.

- Basics of decision theory
  - Decision space

    Set of possible decisions D.

    If $D = \Omega$ then we are looking at the forecast.

    If $D = \Theta$ we are looking at (point) estimation
  - Loss function

    Let $X = (X_1, ..., X_d)$ be a random vector with $X \sim F(x|\theta)$ a **parametric** cdf.

    We call the function $L : \Theta \times D \to \mathbb{R}$ a loss function. We would like to chose $d \in D$ such that: $L(\theta, d)$ gets minimal.
  - Utility function

    If we have benifits, that arise from our choice of $d \in D$ we can capture them in a **utility-function** $U : \Theta \times D \to \mathbb{R}$**.**

    Then choose d such that:

    $$\min_d U(\theta, d)$$
  - Risk

Let $L : \Theta \times D \to \mathbb{R}$ be a loss function. And $\delta(x)$ be a deterministic decision rule. Then the **risk** is defined by:

$$R(\theta, \delta(x)) = \mathbb{E}(L(\theta, \delta(X)))$$

The expected loss

- Examples

  Solutions for loss functions:

$$\begin{aligned} L(\theta, d) = (\theta - d)^2 &\implies d(x) = \mathbb{E}(\theta|x) \\ L(\theta, d) = |\theta - d| &\implies d(x) = m(x|\theta) \end{aligned}$$

- Optimization techniques
  - Bayes

    We can give a prior distribution of $\theta \in \Theta$. The Bayes-risk is then defined by:

$$B(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)d\theta$$

    The Bayes-decision rule is now the one with the lowest Bayes-Risk:

$$B(\pi, \delta_B) = \arg\min_{\delta \in D} B(\pi, \delta)$$

    - non informative prior

      If you have no information on the prior, every θ is equally likely:

$$\pi(\theta) = c \in \mathbb{R} \quad \forall \theta \in \Theta$$

    - Non-randomized Bayes

      There is no randomized prior, that improves over a non-randomized prior. So all Bayes-Risks are based on non-random priors.

    - Admissibility

      If continuity on Θ for the prior $\pi(\theta)$ is given and the support $\mathcal{S}_\pi = \Theta$ is the whole parameter space, the resulting rule δ_B is admissible

      - Shape of risk-points

The set of admissible decisions $\delta \in \mathcal{D}$ have risk-points $\mathcal{R}$ with **convex** shape

- positive Bayes-rule

  If Θ is finite with $|\Theta| = k$ and $\pi(\theta) > 0 \forall \theta \in \Theta$. Then the resulting rule δ_B is called **positive Bayes-Rule**

  Every positive Bayes-Rule is positive

- MiniMax
  - Definition with Bayes-Risk

    Let Θ^* be the set of all prior distributions then the minimax rule δ_M is:

    $$\sup_{\pi \in \Theta^*} B(\pi, \delta_M) = \inf_{\delta \in \mathcal{D}} \sup_{\pi \in \Theta^*} B(\pi, \delta)$$

    While $\mathcal{D}$ is a class of decisions.

    We call this the **upper value $\bar{V}$:**

    $$\bar{V} := \inf_{\delta \in \mathcal{D}} \sup_{\pi \in \Theta^*} B(\pi, \delta)$$

    The **lower value $\underline{\mathrm{V}}$ is:**

    $$\underline{\mathrm{V}} := \sup_{\pi \in \Theta^*} \inf_{\delta \in \mathcal{D}} B(\pi, \delta)$$

    It comes natural that a given rule δ_M is a minimax rule iff:

    $$\sup_{\pi \in \Theta^*} B(\pi, \delta_M) = \bar{V}$$

    - least favourable distribution

      The least favourable distribution τ_0 is a l.f.d. iff:

      $$\inf_{\delta \in \mathcal{D}} B(\tau_0, \delta) = \underline{\mathrm{V}}$$

    - value of the game

If $\underline{V} = \bar{V} = V$ then V is called the value of the game.

Also the minimax rule

▼ How to find the least favourable distribution

Some times boundary priors are least favourable, because the risk increases at the boundaries. Sometimes there exists a sequence of constants c_m and of priors $\pi_m(\theta)$ such that:

$$c_m \pi_m(\theta) \underset{n\to\infty}{\longrightarrow} \tau_0(\theta)$$

▼ Another way of defining a minimax rule

A minimax rule δ_0 fulfills the property, with π_m being an approximation of the least favourable rule. :

$$R(\theta, \delta_0) \leq \lim_{m\to\infty} B(\pi_m, \delta_m) \quad \forall \theta \in \Theta$$

This means, regardless of the parameter θ, the risk is always lower than baysian with the worst prior.

▼ Equalizer Rules

An equalizer rule $\delta \in D$ yields a constant risk over all $\theta \in \Theta$:

$$R(\theta, \delta) = c \in \mathbb{R} \quad \forall \theta \in \Theta$$

Every admissible equalizer rule is also a minimax rule.

▼ Example

Let $X_i \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $L(\theta, d) = (d-\theta)^2$. Then $d = \bar{X}$ is an equalizer rule for $n = 1$:

$$\begin{aligned} R(\theta, \bar{X}) &= \mathbb{E}((\bar{X} - \theta)^2) \\ &= Var(\bar{X}) \\ &= \sigma^2/n \\ &= \sigma^2 \end{aligned}$$

So regardless of the prior $\pi(\theta)$ the Bayes-risk is always σ^2.

- Shrinkage methods

  We have seen that $d = X$ is a valid minimax rule for $X \sim \mathcal{N}(\theta, \sigma^2)$. But what about the multidimensional case?

  If we have $X \sim \mathcal{N}(\theta, I_p)$ does this still hold?

  In that case we would get:

  $$R(\theta, X) = p$$

  But for $p > 2$ we get an **R-better** rule by way of **shrinkage**.

  This **James Stein** estimator shrinks the estimation:

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

  It turns out that:

  $$R(\theta, \delta_{JS}) = p - (p-2)^2 \mathbb{E}\left(\frac{1}{\|X\|^2}\right) < \mathbb{E}(\theta, X)$$

  With shrinkage, we introduce a bias and reduce the variance. But the reduction in variance is always be greater than the increase in bias.

  - Optimal shrinkage estimator

    The optimal shrinkage b^* requires knowledge of $\|\theta\|$. We can estimate $\|\theta\|$ by $\|X\|^2 - p$. But this would lead to a bias in the estimator. There is nothing on $R(\theta, b^* X)$

  If we have a more general setup like $X \sim \mathcal{N}(\theta, \Sigma)$ we can use:

  $$\delta_{JS} = \left(1 - \frac{\tilde{p} - 2}{X' \Sigma^{-1} X}\right)$$

  With $\tilde{p} = tr(\Sigma)/\lambda_{max}(\Sigma)$ with λ_{max} being the largest eigenvalue of Σ.

  Why an estimation-problem with one observation?

  Consider this linear regression problem:

  $$Y = X\beta + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2), X \in \mathbb{R}^n$$

Then it holds that:

$$Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

Now we don't know β. Let's estimate it using OLS:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Since X is deterministic, we are looking at a **deterministic and linear** transformation of Y and therefore know the distribution of $\hat{\beta}$:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$$

But get this: $X\hat{\beta} = X(X'X)^{-1}X'Y$ so we know the distribution of that too:

$$X\hat{\beta} \sim \mathcal{N}(X\beta, \sigma^2 X(X'X)^{-1}X')$$

We are now looking at the same problem stated above. We have one realization of $X\hat{\beta}$ and are interested in $\theta = X\beta$. We therefore know that $X\hat{\beta}$ is NOT an admissible decision rule for $X\beta$. We can now apply shrinkage and get an better estimation of $X\beta$. The James Stein estimator is now:

$$\delta_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\hat{\beta}'(X'X)\hat{\beta}}\right) X\hat{\beta}$$

- Hypothesis Testing
  - randomized decisions

    Consider the set up of $\Theta = \{\theta_1, \theta_2\}$ and $D = \{d_0, d_1\}$ (classic test setup) then we can randomize the decision rule by:

    $$\delta(x) = (1 - \phi(x), \phi(x))$$

    Where simply:

    $$\phi(x) = P(\delta = d_1 | X = x)$$

Under $L(\theta, \delta) = 1_{\{\delta \neq \theta\}}$ we have:

$$R(\theta_0, \delta) = P_{\theta_0}(\delta(x) = d_1)$$
$$R(\theta_1, \delta) = P_{\theta_1}(\delta(x) = d_0)$$

- size and power of the test

  We can say that:

$$\alpha(\phi) = P_{\theta_0}(\delta(X) = d_1) = R(\theta_0, \delta)$$
$$\beta(\phi) = P_{\theta_1}(\delta(X) = d_1) = 1 - R(\theta_1, \delta)$$

  This gives us the risk points:

$$\begin{aligned} \mathcal{R} &= \{R(\theta_0, \delta), R(\theta_1, \delta), \forall \delta \in \mathcal{D}\} \\ &= \{\alpha(\phi), 1 - \beta(\phi), \forall \phi \in \mathcal{D}\} \end{aligned}$$

- uniformly most powerful test

  The best test test of level α_0 minimizes the type II error while keeping the probability of a type I error under α_0.

$$\mathbb{E}_{\theta_0}(\phi(X)) = \alpha(\phi) \leq \alpha_0 \tag{1}$$
$$\mathbb{E}_{\theta_1}(1 - \phi(X)) \leq \mathbb{E}_{\theta_1}(1 - \phi'(X)) \forall \phi' \text{ satisfying } (1) \tag{2}$$

- Neyman-Pearson Lemma

  Consider the previous set up and the corresponding densities $f(x|\theta_0) = f_0(x)$ and $f(x|\theta_1) = f_1(x)$. With:

$$f_i : \mathbb{R}^n \to \mathbb{R} \quad i \in \{0, 1\}$$

  We consider the tests with the mapping $\phi(x)$ of the form:

$$\phi(x) = \begin{cases} 1 & f_1(x) > k f_0(x) \\ \gamma(x) & f_1(x) = k f_0(x) \\ 0 & f_1(x) < k f_0(x) \end{cases}$$

With $\gamma : \mathbb{R}^n \to [0, 1]$ and $k \in [0, \infty]$.

The Neyman-Pearson Lemma states, there exists a most powerful test in this class for every level α_0.

▼ Uniformly most powerful onesided testing