1 Basic Statistical Concepts

Consider a poll with two answers, A and B, regarding political parties. Let:

- N: total number of voters,
- M: number of voters supporting A,
- n: size of the poll,
- $X_1, X_2, \dots, X_n$: responses,
- Each $X_i \in \{0, 1\}$ if $X_i = 1$ supports A.

Additionally, assume:

- We select n individuals from N at random and record their truthful reply,
- Every person asked replies (no selection bias),
- People can be asked repeatedly.

The aim of the poll is to estimate the fraction of party A supporters, say θ.

Definition 1 (Estimator)**.** *An intuitive estimator is:*

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

This estimator will be analyzed in the following sections to determine whether it is unbiased, consistent, and optimal.

2 Statistical Models

Let $(X, \mathcal{F})$ be a measurable space, i.e., a set X with a sigma-algebra $\mathcal{F}$, in which our statistical observations take values.

Definition 2 (Statistical Model)**.** *Let $(X, \mathcal{F})$ be some sample space. We call the parameter space Θ. A statistical model is a family of probability measures $\{P_\theta\}_{\theta \in \Theta}$.*

Remark 1. *Often $(X, \mathcal{F})$ is a product space. For example, if $X_i \in \{0, 1\}$, each P_θ is a product distribution, i.e., $X_1, X_2, \dots, X_n$ are independent and identically distributed (iid). Then we say $\{P_\theta : \theta \in \Theta\}$ is an iid statistical model.*

Remark 2. *If every person could only be asked once, we would have P_θ as a hypergeometric distribution, which converges to the Bernoulli model as $N, M \to \infty$.*

3 Parameter Estimation

Assume $(\Omega, \mathcal{F}, P_\theta)$ is the setting of parametric statistics. Assume Θ is measurable.

Definition 3 (Estimator)**.** *An estimator for θ is any measurable function $\hat{\theta} : X \to \Theta$, i.e., any function that, based on some data X, outputs a guess $\hat{\theta}(X)$ for θ.*

4 Unbiased and Consistent Estimators

4.1 Unbiased Estimator

Definition 4 (Unbiased Estimator)**.** *Let $(\Omega, \mathcal{F}, P_\theta)$ be a measurable space. An estimator $\hat{\theta}$ is called unbiased if:*

$$\mathbb{E}[\hat{\theta}] = \theta \quad \forall \theta \in \Theta$$

where $\mathbb{E}_{P_\theta}$ denotes expectation under the law P_θ. In more explicit terms, unbiasedness means no systematic error.

Proof. For the Bernoulli model, we compute:

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

Thus, $\hat{\theta}_n$ is an unbiased estimator of θ. □

4.2 Consistent Estimator

Definition 5 (Consistent Estimator)**.** *Let $\{P_{\theta,n} : n \geq 1\}$ be a sequence of statistical models on the same parameter space. Let $\hat{\theta}_n$ be a sequence of estimators. The sequence $\hat{\theta}_n$ is called consistent if for every $\theta \in \Theta$:*

$$\hat{\theta}_n \to \theta \quad \text{in probability as } n \to \infty$$

or equivalently:

$$P_\theta\left(\lim_{n\to\infty}\hat{\theta}_n = \theta\right) = 1$$

Proof. For the Bernoulli model:

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

We know $\mathbb{E}[\hat{\theta}_n] = \theta$ and $\mathrm{Var}(\hat{\theta}_n) = \frac{\theta(1-\theta)}{n}$. Using Chebyshev's inequality, for any $\epsilon > 0$:

$$P\left(|\hat{\theta}_n - \theta| > \epsilon\right) \leq \frac{\mathrm{Var}(\hat{\theta}_n)}{\epsilon^2} = \frac{\theta(1-\theta)}{n\epsilon^2}$$

As $n \to \infty$, this probability tends to 0, proving that $\hat{\theta}_n$ is consistent. □

5 Maximum Likelihood Estimation (MLE)

Definition 6 (Maximum Likelihood Estimator)**.** *The maximum likelihood estimator (MLE) is the parameter that maximizes the likelihood function:*

$$L(\theta) = \prod_{i=1}^{n} P_\theta(X_i)$$

5.1 Proof: MLE for Bernoulli Model

Proof. For the Bernoulli model, $P_\theta(X_i) = \theta^{X_i}(1-\theta)^{1-X_i}$, so the likelihood function is:

$$L(\theta) = \prod_{i=1}^{n}\theta^{X_i}(1-\theta)^{1-X_i} = \theta^{\sum X_i}(1-\theta)^{n-\sum X_i}$$

Taking the logarithm:

$$\log L(\theta) = \sum X_i \log\theta + (n - \sum X_i)\log(1-\theta)$$

Setting the derivative with respect to θ equal to 0 gives:

$$\frac{d}{d\theta}\log L(\theta) = \frac{\sum X_i}{\theta} - \frac{n - \sum X_i}{1-\theta} = 0$$

Solving for θ, we get:

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

which is the MLE. □

6 Bayesian Methods

Definition 7 (Posterior Distribution in Bayesian Inference)**.** *In Bayesian statistics, a key element is the prior distribution, denoted by $\pi(\theta)$, which reflects our beliefs about the parameter θ before observing data. The posterior distribution is given by:*

$$\pi(\theta|X) \propto P_\theta(X)\pi(\theta)$$

6.1 Example: Posterior for Bernoulli Model

Example 1. *Suppose we have a Beta prior for θ, $\pi(\theta) \sim Beta(\alpha, \beta)$, and observe $X_1, \dots, X_n$ as Bernoulli trials. The likelihood is:*

$$P(X|\theta) = \theta^{\sum X_i}(1-\theta)^{n-\sum X_i}$$

The posterior is proportional to the product of the prior and likelihood:

$$\pi(\theta|X) \propto \theta^{\sum X_i+\alpha-1}(1-\theta)^{n-\sum X_i+\beta-1}$$

Thus, $\pi(\theta|X) \sim Beta(\sum X_i + \alpha, n - \sum X_i + \beta)$.

Notes on Bayes and Posterior

Posterior = prior × likelihood

Normalizing Constant

$$\int \text{Posterior}\,\text{dx} = 1$$

So,

$$\int \text{Posterior}\,\text{dx} = 1$$

Prior → Posterior via Bayes.

Let $\mathcal{F}_0$ be a σ-algebra on Ω and suppose $(\Omega, \mathcal{F}_0, P_\theta)$ is a dominated statistical model with densities $p(x|\theta)$. Assume

$$x, \theta \in \Omega \quad \Rightarrow \quad p(x|\theta)$$

is jointly measurable with respect to $\mathcal{F}_0 \times \mathcal{F}_1$.

Let π be a prior distribution on Ω with density $\pi(\theta)$ with respect to measure ν. Define posterior density

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)\,d\theta}$$

The corresponding probability measure is called the **posterior distribution**.

Think of $p(x|\theta)$ as a Lebesgue measure. Let ν be a Lebesgue density.

Exception: If $\Omega = \{0, 1\}$, then we take ν to be the counting measure.

From the posterior, we can derive several estimators. For example, $E[\theta|X = x]$ is convex:

$$\int \theta p(x|\theta)\,d\theta = E[\theta|X = x]$$

Example: Binomial model $X|\theta \sim \text{Binomial}(n, \theta)$ with prior $\theta \sim \text{Unif}(0, 1)$.

For a uniform prior, we know the MAP and MLE.

Posterior mean:

$$\theta_{\text{MAP}} = \frac{k+1}{n+2}$$

In the case of coin flips, $X \sim \text{Binomial}(n, \theta)$, where k is the number of heads, we conclude $\theta|X \sim \text{Beta}(k + 1, n - k + 1)$.

$$\theta|X \sim \text{Beta}(k+1, n-k+1)$$

Conjugate Bayes Models: Let $P_\theta \in \mathcal{P}$ be a statistical model. Then some family of priors is called **conjugate** if

$$P_\theta \in \mathcal{P} \Rightarrow \theta|X \in \mathcal{P}$$

for all $X \in \mathcal{X}$, where $\mathcal{X}$ is the sample space.

$$\theta|X \sim \text{Beta}(a, b), \quad X \sim \text{Bernoulli}(p)$$

Loss Functions and Risk

Loss Function: A function $L : \Theta \times \mathcal{X} \to [0, \infty)$ is a basis function if for every $\theta \in \Theta$, $L(\theta, \cdot)$ is measurable.

Given an estimator δ, the expected loss is

$$R(\theta, \delta) = E_\theta[L(\theta, \delta)]$$

Mean Squared Error (MSE):

$$L(x, y) = (x - y)^2 \Rightarrow R(\theta, \delta) = E_\theta[(\delta - \theta)^2]$$

Bias-Variance Decomposition:

$$L(x, y) = (x - y)^2$$

Proof: Let $\delta(x) = E[\theta|X = x]$.

$$R(\theta, \delta) = E_\theta[(\delta(X) - \theta)^2]$$

Bias-variance decomposition:

$$E[(\delta(X) - \theta)^2] = \text{Var}(\delta(X)) + (\text{Bias})^2$$

Minimax and Bayes Risk

Minimax Risk: Given an estimator δ in a model $P_\theta \in \mathcal{P}$, the maximal risk of it is

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

The minimax of a model P_θ is given as $\inf_\delta \sup_\theta R(\theta, \delta)$, where the inf is over all estimators.

An estimator is called minimax if

$$\sup_\theta R(\theta, \delta) = \inf_\delta \sup_\theta R(\theta, \delta)$$

Bayes Risk: Given an estimator δ and prior π on Θ, the Bayes risk of δ is defined as

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

The posterior risk of an estimator $\delta(X)$ is defined by

$$R(\delta|X = x) = E[L(\theta, \delta(X))|X = x]$$

Suppose δ^* is an estimator that minimizes the posterior risk, $\delta^*(x) = E[\theta|X = x]$. Then it also minimizes the Bayes risk.

If $L(x, y) = (x - y)^2$, the Bayes optimal estimator $\delta(x)$ is the posterior mean.

We want to construct $C(x)$ s.t. $P_\theta(\theta \in C(x)) \geq 1 - \alpha$, $\forall \theta \in [0, 1]$

$$x^{(1)} \qquad (\quad) \quad C(x^{(1)})$$

$$x^{(k)} \qquad (\quad) \quad C(x^{(k)})$$
$$\theta \to \quad \to \quad \to \quad \text{contains true param } 3/4 \text{ times}$$

Example cont.:

Best guess: $C(x) = \left[\frac{\bar{X}_n - a}{n}, \frac{\bar{X}_n + b}{n}\right]$

$$P_\theta^n(\theta \in C(x)) = P_\theta^n\left(\frac{\bar{X}_n}{n} - \theta \in [-b, a]\right)$$
$$= F_\theta^n(a) - F_\theta^n(-b) + \rho_n$$

where $F_\theta^n : \mathbb{R} \to [0,1], F_\theta^n(t) = P_\theta^n\left(\frac{\bar{X}_n - \theta}{n} \le t\right)$ is the CDF of $\frac{\bar{X}_n - \theta}{n}$ under P_θ and $\rho_n = P_\theta^n\left(\frac{\bar{X}_n}{n} - \theta = -b\right)$.

How to choose a and b:

$$\text{CDF} \quad \text{CDF} \quad \leftarrow \quad -b \quad a \to t$$

We'd like to choose $a = (F_\theta^n)^{-1}\left(1 - \frac{\alpha}{2}\right)$ and $b = (F_\theta^n)^{-1}\left(\frac{\alpha}{2}\right)$, where

$$(F_\theta^n)^{-1}(p) := \inf\{t \in \mathbb{R} : F_\theta^n(t) \ge x\} \quad \text{(Quantile Function)}$$

Let's use a normal approximation, for $\sigma^2 = \theta(1-\theta)$:

$$\sqrt{n}\left(\frac{\bar{X}_n}{n} - \theta\right) = \frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - \theta}{\sigma} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{[CLT]}$$

$X_k \sim \text{Ber}(\theta)$

Then it follows that

$$F_\theta^n(a_n) = P_\theta^n\left(\frac{\bar{X}_n}{n} - \theta \le a_n\right)$$
$$= P_\theta^n\left(\frac{\sqrt{n}}{\sigma}\left(\frac{\bar{X}_n - \theta}{n}\right) \le \sqrt{n} a_n\right)$$
$$= \Phi\left(\frac{\sqrt{n}}{\sigma} a_n\right),$$

where the convergence is valid if $a_n := \text{const.}\frac{1}{\sqrt{n}}$.

Now, let us choose

$$a := \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$$

where $z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ is the $1 - \frac{\alpha}{2}$ quantile of $\mathcal{N}(0,1)$ and $b = a$. Then

$$C(x) = \left[\frac{\bar{X}_n}{n} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \frac{\bar{X}_n}{n} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right]$$

It follows

$$P_\theta^n(\theta \in C(x)) = F_\theta^n(a_n) - F_\theta^n(b) + \rho_n = 1 - \frac{\alpha}{2} + o(1) + o(1)$$
$$= 1 - \alpha + o(1) \text{ as } n \to \infty$$

$$\Rightarrow \text{ Asymptotically valid confidence set}$$

One more problem: σ depends on θ

- Upper bound: $\sup_{\theta \in [0,1]} \theta(1-\theta) = \frac{1}{4}$ (maximized at $\theta = \frac{1}{2}$)
- Empirical Variance: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \frac{1}{n}\sum_{i=1}^n X_i)^2$

$$\frac{\hat{\sigma}^2}{\sigma^2} \xrightarrow{P_\theta} 1$$

Slutsky's Theorem:

$$X_n \xrightarrow{d} X, \quad Y_n \xrightarrow{d} \text{const.} \Rightarrow X_n Y_n \xrightarrow{d} CX$$

Exercise: Use this to deduce that $a_n = \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$ is also valid

Remark:

Hypothesis Testing

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

- A statistical test is a measurable function of the data $\varphi : (\mathcal{X}, \mathcal{F}) \to [0, 1]$
- If $\forall x \in \mathcal{X}, \varphi(x) \in \{0, 1\}$, then φ is a non-randomized test
- Else φ is randomized

Definitions:

- $H_0 : \theta \in \Theta_0$ is called the null hypothesis
- $H_1 : \theta \in \Theta_1$ is called the alternative hypothesis
- The map $\theta \to \beta_\varphi(\theta) = P_\theta[\varphi = 1]$ is called the power function of a test φ

1 $\beta_\varphi(\theta)$ 0 Θ_0 Θ_1 Θ

- For $\theta \in \Theta_0$, $\beta_\varphi(\theta)$ is the type-I-error under θ [Wrongly rejecting the null]
- For $\theta \in \Theta_1$, $1 - \beta_\varphi(\theta)$ is the type-II-error

Note:

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

Definition: [Level]

$\varphi : \mathcal{X} \to [0, 1]$ has level $\alpha \in [0, 1]$ if

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

Definition: [Uniformly most powerful test]

Given a level $\alpha \in (0, 1)$, $\varphi : \mathcal{X} \to [0, 1]$ is called UMP if for every other test φ' of level α and all $\theta \in \Theta_1$,

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

1 α 0 $\beta_\varphi(\theta)$ $\beta_{\varphi'}(\theta)$ Θ_0 Θ_1

Remark:

In general, it is very hard to find UMP tests. But: for simple hypotheses, i.e. $\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$, it is possible. Here, likelihood ratio tests are UMP.

Theorem: [Neyman-Pearson Lemma]

Let $\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$ be simple:

1. **Existence:** There exists a test φ and a constant $k \in [0, \infty)$, s.t. $P_{\theta_0}(\varphi = 1) = \alpha$, of the form

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

   Here $p_{\theta_1}, p_{\theta_0}$ are densities w.r.t. some dominated measure μ, e.g. $\mu = p_{\theta_0} + p_{\theta_1}$. Finite Θ implies measure is always dominated (likelihood always exists).

2. **Sufficiency:** If φ satisfies $P_{\theta_0}(\varphi = 1) = \alpha$ and $(*)$ 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 $(*)$, and it also satisfies $P_{\theta_0}(\varphi_k = 1) = \alpha$, or else it must satisfy $P_{\theta_1}(\varphi_k = 1) = 1$.

Proof:

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

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

   Then define also $\alpha(t) = 1 - F_0(t) = P_{\theta_0}(r(x) > t)$

   - α is right-continuous:

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

   - α is non-increasing
   - α has left limits

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

   α is cadlag:

   - Continuous from the right
   - Limit from the left

$$\text{There exists some } k \in [0, \infty) \text{ s.t. } \alpha \le \alpha(k^-) \quad \text{and} \quad \alpha \ge \alpha(k)$$

   We define our test

$$\varphi(x) = \begin{cases} 1 & \text{if } r(x) > k \\ \gamma & \text{if } r(x) = k \quad [\text{reject null w.p. } \gamma] \\ 0 & \text{if } r(x) < k \end{cases}$$

   We set

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

   The level of φ is

$$\begin{aligned} E_{\theta_0}[\varphi(x)] &= P_{\theta_0}(\varphi(x) = 1) \\ &= P_{\theta_0}(r(x) > k) + P_{\theta_0}(r(x) = k) \cdot \gamma \\ &= \alpha(k) + \left[\alpha(k^-) - \alpha(k)\right] \cdot \frac{\alpha - \alpha(k)}{\alpha(k^-) - \alpha(k)} = \alpha \end{aligned}$$

   (randomizing the test)

Lecture 6

Neyman-Pearson

Power of a test:

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

Likelihood ratio test:

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

LR test

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

for some $k \in [0, \infty)$, $\gamma \in [0, 1]$.

Note: LR tests are UMP for simple hypothesis testing:

- Given some α, if LR satisfies $E_{\theta_0}[\varphi] = \alpha$, it represents a Type I error.
- φ minimizes the Type II error
$$E_{\theta_1}[\varphi] \geq E_{\theta_1}[\varphi'] \quad \forall \varphi'$$

Cont. of proof (part of UMP)

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

Goal: $E_{\theta_1}[\varphi] \geq E_{\theta_1}[\varphi']$. 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: $p \geq 0$.

Observe:

- If $p_{\theta_1}(x) - kp_{\theta_0}(x) > 0 \Rightarrow \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k \Rightarrow \varphi(x) = 1$.
- If $p_{\theta_1}(x) - kp_{\theta_0}(x) < 0 \Rightarrow \varphi(x) = 0$.
- If $p_{\theta_1}(x) - kp_{\theta_0}(x) = 0 \Rightarrow$ integrand $= 0$.

$$\Rightarrow p = 0$$

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

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

Part (3) UMP $\Rightarrow$ (LR): Take φ^* a UMP test, $E_{\theta_0}[\varphi^*] = \alpha$, and let φ be the LR test with $E_{\theta_0}[\varphi] = \alpha$ with (*).

Goal: $\varphi = \varphi^*$ a.e. except on $\{r(x) = k\}$.

Define

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

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

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

Like before, we have

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

Thus if $\mu(\tilde{x}) > 0$,

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

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

But also

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

$$\Rightarrow \text{Cannot be } \varphi^* \text{ is UMP.}$$

Example (Gaussian Location Model)

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

$$H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1, \quad \mu_0 < \mu_1$$

Then:

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

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

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

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

To determine K_α:

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

$$\Rightarrow \mathbb{L} = P_{H_0}\left(\bar{X}_n \geq K_\alpha\right) = 1 - P_{H_0}\left(\bar{X}_n < K_\alpha\right)$$

$$= 1 - \Phi\left(\frac{\sqrt{n}}{\sigma}(K_\alpha - \mu_0)\right) \quad \text{(CDF for } \mathcal{N}(0,1))$$

$$\Rightarrow \text{solving for } K_\alpha \text{ gives } K_\alpha = \mu_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha),$$

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

Corollary

Consider simple hypothesis testing. Let φ be UMP, for level α. Then,

$$\alpha = E_{H_0}[\varphi_0] = E_{\theta_0}[\varphi_0] \leq E_{\theta_1}[\varphi]$$

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

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

Also since $\varphi_0 \in \{\varphi, \beta\}$ we conclude that $p_{\theta_1} = Kp_{\theta_0}$ a.s.

But

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

Correspondence theorem

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

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

$$\text{If } \Pr_{\theta}(\phi_\theta = 1) = \alpha$$

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

(i) Let $C = C(X)$ be a level-α confidence set, then

$$\phi_{\theta_0}(x) = 1\{\theta_0 \notin C(x)\}$$

is a level-α test of $\theta = \theta_0$ vs. $\theta \neq \theta_0$.

(ii) Suppose $\{\phi_{\theta_0} : \theta_0 \in \Theta\}$ 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) $\Pr_{\theta_0}(\phi_{\theta_0} = 1) = \Pr_{\theta_0}(\theta_0 \notin C(X)) = \alpha$

(ii) $\Pr_\theta(\theta \notin C(X)) = \Pr_\theta(\theta \notin \{\tilde{\theta} \in \Theta : \phi_{\tilde{\theta}}(X) = 0\}) = \Pr_\theta(\phi_\theta(X) = 1) \leq \alpha$

UMPT Tests in Models with Monotone Likelihoods
Proposition: Let $\Theta \subseteq \mathbb{R}$. Consider testing $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$, for some $\theta_0 \in \mathbb{R}$. Assume there exists some test statistic $T : X \to \mathbb{R}$ and a function $h : \mathbb{R} \times \Theta \times \Theta$ such that

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

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

The simplest model for the relationship between Y_i and X_i assumes a linear relationship:

$$Y_i = aX_i + b + \varepsilon_i$$

for $i = 1, \dots, n$, where ε_i is centered, i.e., $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. Suppose $\varepsilon \sim N(0, \sigma^2)$ with σ known.

The statistical model is given by

$$(\mathbb{R}, B(\mathbb{R}), (\bigotimes_{i=1}^{n} N(ax_i + b, \sigma^2))_{(a,b)\in\mathbb{R}^2})$$

The likelihood within the statistical model is

$$L((a,b)|y) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - ax_i - b)^2\right)$$

The MLE satisfies the optimization problem

$$(\hat{a}, \hat{b}) = \arg\min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n}(y_i - (ax_i + b))^2$$

Provided that $x_i \neq x_j$ for $i \neq j$, the least squares problem has a solution with minimum given by (Gauss, 1801):

$$(\hat{a}, \hat{b}) = \left(\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \bar{y} - \hat{a}\bar{x}\right)$$

Definition 8 (Linear Model)**.** *A random vector $Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$ stems from a linear model if there exists a parameter vector $\beta \in \mathbb{R}^p$, a matrix $X \in \mathbb{R}^{n\times p}$, and a random vector $\varepsilon \in \mathbb{R}^n$ such that*

$$Y = X\beta + \varepsilon$$

1. *A linear model is called regular if*
   (a) *$p \leq n$ (parameter size is smaller than sample size),*
   (b) *X has full rank. $rank(X) = p \leq n$ (design with full rank)*
   (c) *$E(\varepsilon) = 0$ (noise is controlled)*
   (d) *The covariance matrix is postive definate, $\Sigma = (Cov(\varepsilon_i, \varepsilon_j))_{i,j\in[n]}$*
2. *A linear model is called ordinary if $\Sigma = \sigma^2 E_n$ (and is usually the noise is Gaussian)*

Remark 3. 1. *There are several synonyms*
   (a) *Y a dependent variable, responce, regressand*
   (b) *X, a independent variable, predictor, design matrix, regressor*
   (c) *ε Error, perturbation, reression function*

2. *The matrix Σ is symmetric and diagonalizable, i.e. $\Sigma = UDU^T$ for some diagonal matrix, $D = diag(\lambda_1, \dots, \lambda_n) \in \mathbb{R}^{n\times n}$*
3. *Positive semi-definate, i.e. $\lambda_i \geq 0$*

$$\langle \Sigma u, u\rangle = \langle E[(\varepsilon - E[\varepsilon])(\varepsilon - E[\varepsilon])^T]u, u\rangle$$

$$= E[(\varepsilon - E[\varepsilon])^2] \geq 0, u \in \mathbb{R}^n$$

   item If Σ is positve definate ($\lambda_i > 0$) for $i = 1, \dots, n$, then there exists the inverse $\Sigma^{-1} = UD^{-1}U^T$ and $\Sigma^{-1/2} = UD^{-1/2}U^T$.

4. *If X is not deterministic, we speak of random design.*

In the regular linear model, $\hat{\beta}$ is called weighted least squares estimate, (LSE). if

$$||\sigma^{-1/2}(Y - X\hat{\beta})||^2 = \inf_{\beta \in \mathbb{R}^n} ||\sigma^{-1/2}(Y - X\beta)||^2 = \inf_{\beta \in \mathbb{R}^n} ||\sigma^{-1/2}Y - X_\Sigma \beta||^2$$

where $X_\Sigma = \Sigma^{-1/2}X$. $X_\Sigma \hat{\beta}$ is the point within the subspace,

$$U = \{X_\Sigma \beta \mid \beta \in \mathbb{R}^n\} \subseteq \mathbb{R}^n$$

with the smallest distance to the vector $\Sigma^{-1/2}Y$. Thus, $X_\Sigma \hat{\beta} = \Pi_U(\Sigma^{-1/2}Y)$ where Π_U is the orthogonal projection onto U. $\Pi_U u = u$ for all $u \in U$ $\langle \Pi_U v - v, u \rangle = 0$ for all $u \in U$ and $r \in \mathbb{R}^n$. Provided that $(X_\Sigma^T X_\Sigma)^{-1}$ exists, we can confirm by direct computation that the projection satisfies

$$\Pi_U = X_\Sigma (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T$$

For $u = X_\Sigma \beta$ we have,

$$X_\Sigma (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T X_\Sigma \beta = X_\Sigma \beta = u$$

By symmetry,

$$\langle \Pi_U v - v, u \rangle = \langle v, \Pi_U u \rangle - \langle v, u \rangle = \langle v, u \rangle - \langle v, u \rangle = 0$$

for all $u \in U$.

Lemma 1. *Representation for the LSE Consider a regular linear model, then the LSE exists uniquely, and is given by*

$$\hat{\beta} = (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y = X_\Sigma^+ \Sigma^{-1/2} Y$$

Proof. $\ker(X_\Sigma^T X_\Sigma)$ is invertible. Suppose that $X_\Sigma^T X_\Sigma v = 0$ ($v \in \ker(X_\Sigma^T X_\Sigma)$)

$$0 = v^T X_\Sigma^T X_\Sigma v = (X_\Sigma^T v)^T X_\Sigma v = \langle X_\Sigma v, X_\Sigma v \rangle = ||X_\Sigma v||^2 = ||\Sigma^{-1/2} X v||^2 \implies ||Xv||^2 = 0 \implies v = 0$$

So then

$$\begin{aligned} X_\Sigma \hat{\beta} &= \Pi_u \Sigma^{-1/2} Y = X_\Sigma (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y \\ X_\Sigma^T X_\Sigma \hat{\beta} &= X_\Sigma^T X_\Sigma (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y \\ \implies \hat{\beta} &= (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y \end{aligned}$$

□

Remark 4. *1. If $p > n$, then $(X_\Sigma^T X_\Sigma)^{-1}$ does not exist and the LSE is not unique.*

$$\left\{ \beta \cdot ||\Sigma^{-1/2} Y - X_\Sigma \beta||^2 = 0 \right\}$$

is a $p - n$ dim subspace and each solution interpolates the data

Theorem 1. *Optimality of the LSE, Gauss-Markov Theorem Consider an ordinary linear model for $\sigma > 0$, then*

1. *The least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$ is linear and the unbiased parameter for the parameter β.*
2. *For the desired parameter $\alpha = \langle \beta, v \rangle$ for $v \in \mathbb{R}$, the estimator $\hat{\alpha} = \langle \hat{\beta}, v \rangle$ is the best linear unbiased estimator (BLUE), meaning that $\hat{\alpha}$ has the optimal value within the class of linear unbiased estimators for α*
3. *$\hat{\sigma}^2 = \frac{||Y - X\hat{\beta}||^2}{n-p}$ is an unbiased estimator of σ^2*

Proof.

$$\hat{\beta}(y + \tilde{y}) = \hat{\beta}(y) + \hat{\beta}(\tilde{y}) \text{for } y, \tilde{y} \in \mathbb{R}^n$$

$$\begin{aligned} E[\hat{\beta}] &= (X^T X)^{-1} X^T E[Y] && (1) \\ &= (X^T X)^{-1} X^T E[X\beta + \varepsilon] && (2) \\ &= (X^T X)^{-1} (X^T X) \beta && (3) \\ &= \beta && (4) \end{aligned}$$

Suppose that $\tilde{\alpha}$ is some other linear unbiased estimator of α. Since the estimator is linear, there exists some element w such that $\tilde{\alpha} = \langle y, w \rangle$

$$\langle \beta, v \rangle = \alpha = E[\tilde{\alpha}] = E[\langle y, w \rangle] = \langle X\beta, w \rangle = \langle \beta, X^T w \rangle$$

This implies that $v = X^T w$, therefore we have,

$$\begin{aligned}
\text{Var} &= \text{Var}(\langle x\beta, w \rangle + \langle \varepsilon, w \rangle) &(5)\\
&= \text{Var}(\langle \varepsilon, w \rangle) + E\left[\left(\sum_{i=1}^{n} \varepsilon w\right)^2\right] &(6)\\
&= \sigma^2 \sum_{i=1}^{p} w_i^2 = \sigma^2 ||w||^2 &(7)
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\alpha}) &= E[\langle \hat{\beta} - \beta, v \rangle^2] &(8)\\
&= E[\langle (X^T X)^{-1} X^T \beta + (X^T X)^{-1} X^T \varepsilon - \beta, v \rangle^2] &(9)\\
&= E[\langle (X^T X)^{-1} X^T \varepsilon, v \rangle^2] &(10)\\
&= \sigma^2 ||X(X^T X)^{-1} v||^2 = \sigma^2 ||X(X^T X)^{-1} X^T w||^2 &(11)\\
&= \sigma^2 ||\Pi_u w||^2 &(12)
\end{aligned}$$

Thus, $\text{Var}(\hat{\alpha}) \leq \text{Var}\tilde{\alpha}$ □

7 Lecture 8

Recall linear model

$$Y = X\beta + \varepsilon$$

where $\text{cov}(\varepsilon) = \Sigma$.

OLD: $\hat{\beta} = (X_\Sigma^T X_\Sigma)^{-1} X_\varepsilon^T \Sigma^{-1/2} Y$.

$X\hat{\beta}$ = Projection of $\Sigma^{-1/2} Y$ onto span $\{X_{\varepsilon,1}, \ldots, X_{\varepsilon,p}\}$

Theorem 2 (Gauss-Markov)**.** *1. $\hat{\beta}_{OLS}$ is the best linear unbiased est (BLUE)*

2. $\alpha_i = \langle \beta, v \rangle$ is BLUE.

3. $\hat{\sigma}^2 = \frac{||Y - X\hat{\beta}||^2}{n-p}$ is unbiased est for $\sigma^2 > 0$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}^T + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$ Where our data is $(Y_i, X_i)_{i=1}^n \in (\mathbb{R} \times \mathbb{R}^p)^{\otimes p}$

Remark 5. *Is this an iid model? Depends!*

1. *Typically ε_i are iid.*
2. *If X_i are random then "random design".*
3. *If X_i are iid, then linear model is iid model.*
4. *If X_i are deterministic, then not iid model.*

$\beta \mapsto ||Y - X\hat{\beta}||$.

Proof. This is a continuation of point 3 in our theorem above.

We already introdiced $\Pi_U = X(X^T X)^{-1} X^T$ projection onto col space U of X. Thus $I_n - \Pi_U$ is another projection operator, onto $U^\perp$ (othrogonal complement),

$$U^\perp = \{z \in \mathbb{R}^n \mid \langle z, X_k \rangle \forall k = 1, \ldots, p\}.$$

Choose a basis $e_1, \dots e_{n-p}$, orthonormal, of $U^\perp$, then

$$(I_n - \Pi_U)z = \Pi_{U^\perp} z = \sum_{n=1}^{n-p} \langle z, e_k \rangle e_k.$$

$$\begin{aligned}
||Y - X\hat{\beta}|| &= ||Y - \underbrace{X(X^TX)^{-1}X^T}_{\Pi_U} Y||^2 &(13)\\
&= ||(I_n - \Pi_n)Y||^2 &(14)\\
&= ||(I_n - \Pi_n)(X\beta + \varepsilon)||^2 &(15)\\
&= ||(I_n - \Pi_n)\varepsilon||^2 &(16)\\
&= \sum_{i=1}^{n-p} \langle \varepsilon, e_i \rangle^2 &(17)\\
& &(18)
\end{aligned}$$

Hence,

$$E[||Y - X\hat{\beta}||^2] = \sum_{i=1}^{n-p} E[\langle \varepsilon, e_i \rangle^2] = n - p \implies E[\hat{\sigma}] = n - p$$

□

Remark 6. *Recall the $N(\mu, \sigma^2)$ model, where the MLE is*

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2.$$

The unbiased estimator for σ^2 was $\frac{1}{n-1} \sum_{i=1}^{n} (X - i - \hat{\mu})^2$. This is related to the $n - p$ factor in point 3.

Remark 7. *1. If linearity is dropped, there exists better estimators than $\hat{\beta}_{OLS}$. For example a connstant estimator, $\hat{\beta} = \beta^*$*

2. The MSE of $\hat{\beta}_{OLS}$ is

$$E[||\hat{\beta}_{OLS} - \beta||^2] = E\left[\sum_{i=1}^{p} \langle \hat{\beta}_{OLS} - \beta, \underbrace{e_i}_{ONB of \mathbb{R}^n} \rangle^2\right] = \sum_{i=1}^{n} Var_\beta(\langle \hat{\beta}_{OLS}, e_i \rangle) = \sum_{i=1}^{n} \sigma^2 ||X(X^TX)^{-1} e_k||^2$$

We say X satisfies orthogonal design if

$$X^TX = nI_p$$

"The different covariants are uncorrelated.=" $(X^TX)_{ij} = \langle X_i, X_j \rangle = n\delta_{ij}$ For orthogonal design,

$$E_\beta[||\hat{\beta}_{OLS} - \beta||^2] = \frac{1}{n^2} \sigma^2 \sum_{i=1}^{n} \underbrace{||xe_i||^2}_{n} = \frac{\sigma^2 P}{n}.$$

and this is equal to noise level times the number of parameters, divided by the number of data points.

Theorem 3 (Bayes in Linear Models)**.** *Consider a linear model $Y = X\beta + \varepsilon$, and $\varepsilon \sim N(0, \sigma^2 I_n)$ with $\sigma > 0$ known and $\beta \sim N(m, M)$ where $m \in \mathbb{R}^p, M \in \mathbb{R}^{p \times p}$ positive semi definate. Then, the posterior $\Pi(\beta | Y_i X)$ is given by*

$$\Pi(\beta | Y, X) = N(\mu_{post}, \Sigma_{post}) \text{ for}$$

$$\mu_{post} = \sigma_{post}(\sigma^{-2} X^T y + M^{-1} m) \quad \Sigma_{post} = (\sigma^{-2} X^T X + M^{-2})^{-1}$$

Remark 8. *Σ_{post} independent of Y. For "$M^{-2} \to 0$", then "$\mu_{post} \to \hat{beta}_{OLS}$"*

Proof.

$$L(X,Y,\beta)\pi(\beta) \propto \exp\left(-\frac{1}{2\sigma^2}||Y-X\beta|| - \frac{1}{2}(\beta-m)^T M^{-1}(\beta-m)\right)$$

We want this to be proportional to $\exp\left(-\frac{1}{2}(\beta-\mu_{\text{post}})^T\sigma_{\text{post}}^{-1}(\beta-\mu_{\text{post}})\right)$.

Now,

$$\exp\left(-\frac{1}{2}(\beta-\mu_{\text{post}})^T\sigma_{\text{post}}^{-1}(\beta-\mu_{\text{post}})\right) \propto \exp\left(-\frac{1}{\sigma^2}\beta^T X^T X\beta - \frac{1}{2}\beta^T M^{-1}\beta + \frac{1}{\sigma^2}\beta^T X^T Y + \beta^T M^{-1}m\right)$$

and this is equal to

$$\exp\left(-\frac{1}{2}\beta^T\left(\frac{1}{\sigma^2}X^TX + M^{-1}\right)\beta + \beta^T(\frac{1}{\sigma^2}X^TY + M^{-1}m\right)$$

and this is

$$\propto \exp\left(-\frac{1}{2}(\beta-\mu_{\text{post}})^T\sigma_{\text{post}}^{-1}(\beta-\mu_{\text{post}})\right)$$

□

Corollary 1. *For $\ell = ||\cdot||^2$, the Bayes estimator is $\hat{\beta}_\Pi = \mu_{post}$*

Proposition 1. *Consider the previous setting (from the theorem), with $m = 0$, and $M = \tau^2 I_p$ (centered, isotropic, normal prior). The, $\mu_{post} = \hat{\beta}_\Pi$ minimizes*

$$\beta \mapsto ||Y-X\beta||^2_{\mathbb{R}^n} + \underbrace{\frac{\sigma^2}{\tau^2}||\beta||^2_{\mathbb{R}^p}}_{\text{"penalty" or "regularization"}}$$

Proof.

$$\nabla_\beta J(\beta) = -ZX^T\underbrace{(Y-X\beta)}_{\mathbb{R}^n} + \frac{2\sigma^2}{\tau^2}\beta = 0 \in \mathbb{R}^p$$

This happens if and only if

$$Z(X^TX + \frac{\sigma^2}{\tau^2})\beta - 2X^TY = 0 \iff \hat{\beta} = (X^TX + \frac{\sigma^2}{\tau^2}I)^{-1}X^TY.$$

Thus,

$$\mu_{\text{post}} = \Sigma_{\text{post}}^{-1}\left(\sigma^{-2}X^TY + M^{-1},\right) \tag{19}$$

$$= \left(\sigma^{-2}X^TX + \frac{1}{\tau^2}I_p\right)^{-1}\sigma^{-2}X^TY \tag{20}$$

$$= (X^TX + \frac{\sigma^2}{\tau^2}I_p)^{-1}X^TY. \tag{21}$$

□

Remark 9. *$\hat{\beta}$ is defined even if rank$(X) < p$, in particular, even for $n < p$.*

Definition 9. *$\hat{\beta}$ given by*

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} ||Y-\beta||^2_{\mathbb{R}^n} + \lambda||\beta||^2_{\mathbb{R}^p}$$

is called a ridge regression estimator. Have $\lambda > 0$ is called regularization parameter. We sometimes write $\hat{\beta}_{ridge}$. $\hat{\beta}_{ridge}$ is always uniquely defined.

Proposition 2 (MSE of $\hat{\beta}_{\text{ridge}}$)**.** *Consider a linear model with $\varepsilon \sim N(0,\sigma I_n)$, $\sigma > 0$ known, and $X^TX = nI_p$ with design. Let $\alpha = \langle\beta, v\rangle$, for $v \in \mathbb{R}$ and $\hat{\alpha}_{ridge} = \langle\hat{\beta}_{ridge}, v\rangle$. Then,*

1. $$E_\beta[|\hat{\alpha}_{ridge} - \alpha|^2] = (1+\frac{n}{\lambda})\langle\beta,v\rangle^2 + \frac{\sigma^2||v||^2}{n}\left(1+\frac{\lambda}{n}\right)$$

2. $$E_\beta[|\hat{\beta}_{ridge} - \beta|^2] = \left(1+\frac{n}{\lambda}\right)^{-2}||\beta||^2 + \underbrace{\frac{p\sigma^2}{n}}_{\text{risk of } \hat{\beta}_{OLS}}\left(1+\frac{\lambda}{n}\right)^{-2}$$