Statistics for Mathematics (Stat4Math)

Sara van de Geer

Spring 2021

Contents

Chapter 1

Introduction

With todays large amounts of data, statistics is more relevant than ever. Understanding what statistical algorithms (or machine learning algorithms) do, and how to interpret their outcome is essential in scientific research, and also in daily life. In this lecture we treat some classical concepts and methods from statistics, with an outlook to more modern mathematical statistics in Chapter 12 (new developments are coming in very fast). The mathematical theory relies on various branches of mathematics: probability theory, (functional, numerical) analysis, optimization, geometry, topology, algebra, ... Moreover, mathematical statistics has its own mathematics. With the present lecture notes we will not be able to treat all this. There will be very few formal theorems with formal proofs. The idea is rather to get a first glimpse of the statistical philosophy. We present approaches to statistical problems that intuitively should "make sense", but most of the time we do not formally prove any optimality properties. The latter is the main theme of the lecture *Fundamentals of Mathematical Statistics*.

An overview of standard distributions is given in Appendix A.

1.1 Notation

In the lectures on probability theory, we have seen random variables X with distribution P. Formally, one starts with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and a random variable X is defined as a measurable mapping $X: \Omega \mapsto \mathcal{X}$ where $\mathcal{X} = \mathbb{R}^k$ or more generally some measurable space. The distribution P of X is given by

$$P(A) := \mathbb{P}(\omega \in \Omega : \ X(\omega) \in A) = \mathbb{P}(X \in A)$$

for measurable sets $A \subset \mathcal{X}$. Shorthand notation: $X \sim P$. In what follows, we will implicitly assume measurability without stating this explicitly. Moreover, we sometimes apply the abuse of notation

$$P(A) := P(X \in A).$$

In statistics, the distribution P is unknown and we aim at estimating it from data. For example, X could be the yearly expenditure of a person living in Switzerland. We do not know the distribution of X because we did not ask everybody what his/her expenditures were. We can estimate the distribution by asking n persons their expenditures. The data then consists of their answers $X_1, \ldots, X_n$. As another example, in classification one observes $X = (Y, Z)$ where $Y \in \{0, 1\}$ is a label and Z are features. This could be for instance $X =$ a painting, $Z =$ colours used, composition (coded in an suitable way), abstraction level (coded in a suitable way), etc., and $Y = 1$ if the picture is a Picasso and $Y = 0$ otherwise. We do not know the distribution of Y given Z, i.e., the probability of a Picasso given the features Z of the painting. We also do not know the distribution of the features. The data may be n paintings where we know the features and whether or not it concerns a Picasso. We aim at learning from the data (so-called supervised learning) how to recognize a Picasso up to a small probability of making a mistake.

In most of the theory in these lecture notes, the data (observations) are assumed to be independent identically distributed[1](which we abbreviate to i.i.d.) random variables $X_1, \ldots, X_n$ each having the same distribution P on $\mathcal{X}$.

We call $\mathcal{X}$ the observation space (typically (a subset of) Euclidean space). The sample is $\mathbb{X} = \mathbb{X}_n := (X_1, \ldots, X_n) \in \mathcal{X}^n$ and n the sample size. We say that $X_1, \ldots, X_n$ are i.i.d. copies of a random variable $X \in \mathcal{X}$.

1.2 Statistical models

Definition 1.2.1 *A statistical model*[2] *says that $X \sim P \in \mathcal{P} := \{P_\theta : \ \theta \in \Theta\}$. The set Θ is called the parameter space.*

Notation If $X \in \mathbb{R}^k$ has distribution P_θ its expectation depends on θ. We (often) write the expectation with a subscript: $E_\theta X$.

Example 1.2.1 *.The normal distribution is commonly used to model "measurement error". If $X \in \mathbb{R}$ and its mean $\mu := EX$ exists can write*

$$X = \mu + \epsilon,$$

[1] Thus for all $A_1, \ldots, A_n$ measurable subsets of $\mathcal{X}$,

$$\mathbb{P}(\omega : X_1(\omega) \in A_1, \ldots, X_n(\omega) \in A_n) = \prod_{i=1}^n \mathbb{P}(\omega : \ X_i(\omega) \in A_i).$$

[2] Formally, one calls $\{(\Omega, \mathcal{F}, \mathbb{P}_\theta) : \ \theta \in \Theta\}$ a statistical experiment. The observations are $X_i : \ \Omega \to \mathcal{X}$, $i = 1, \ldots, n$, and in the i.i.d. case, for all $A_1, \ldots, A_n$ measurable subsets of $\mathcal{X}$,

$$\mathbb{P}_\theta(\omega : \ X_1(\omega) \in A_1, \ldots, X_n(\omega) \in A_n) = \prod_{i=1}^n \mathbb{P}_\theta(\omega : X_i(\omega) \in A_i).$$

where $\epsilon = X - \mu$ can be seen as measurement error or noise. If we assume $\mu \in \mathbb{R}$ to be unknown and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ where the variance $\sigma^2 > 0$ is also unknown, the statistical model is

$$\mathcal{P} = \Big\{ P_\theta \text{ is the normal distribution with mean } \mu \text{ and variance } \sigma^2, \\ \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+ \Big\}.$$

A statistical model is typically an idealization of the real world situation. For instance in Example 1.2.1 above, the assumption of a normal distribution is perhaps for its ease in computations, or inspired by the central limit theorem. The model is at best only an approximation of the truth. In these lecture notes, we will however assume throughout that the model is correct, unless otherwise stated (as we will in Chapters 11 and 12). If the model is not correct we call it misspecified. There exists a large body of statistical methods that are robust against model misspecification. It is a topic in itself and closely related to the theory for adversarial learning developed in e.g. the computer science literature.

Note that if we know nothing about the distribution P we have

$$\mathcal{P} = \{ \text{ all distributions on } \mathcal{X} \}.$$

Then we may take $\Theta = \mathcal{P}$. In other words, the parameter space Θ may be finite-dimensional such as is the case in Example 1.2.1, but it can also be a rather abstract space such as the space of all distributions. Of course if $\mathcal{X}$ is finite, say $|\mathcal{X}| = q$, then the space of all distributions on $\mathcal{X}$ is finite-dimensional, in the sense that it can be described by q (in fact $q - 1$) Euclidean parameters (the probabilities $P(X = x)$, $x \in \mathcal{X}$). But otherwise, the class $\mathcal{P} := \{\text{all distributions on } \mathcal{X}\}$ cannot be described by finitely many Euclidean parameters. We call a model with parameter space Θ that cannot be described by finitely many Euclidean parameters nonparametric.

1.3 Parameter of interest and estimators

Let $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ be a statistical model.

Definition 1.3.1 *A parameter of interest is $\gamma := Q(P)$ where Q is a given map $Q : \mathcal{P} \to \Gamma$ with domain some given space Γ (typically $\Gamma = \mathbb{R}$ or some subset thereof). We then write $g(\theta) := Q(P_\theta)$ where $g : \Theta \mapsto \Gamma$.*

In Example 1.2.1: $X \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$, the parameter μ (the signal) is typically the parameter of interest. It is the quantity we observe with measurement error ϵ (the noise). The variance σ^2 is then called a nuisance parameter.

Example 1.3.1 *Let $X = (Y, Z) \in \{0, 1\} \times \mathbb{R}$ where Z = pitch of voice and Y = gender ($Y = 0$ is male and $Y = 1$ is female). Suppose the probability that*

$Y = 1$ given the pitch of voice $Z = z$ is strictly increasing in z. A parameter of interest could then be the value $\gamma \in \mathbb{R}$ for which given $Z = z$:

$$\begin{cases} Y = 0 \text{ (male) is more likely} & \text{if } z < \gamma \\ Y = 1 \text{ (female) is more likely} & \text{if } z > \gamma \end{cases}.$$

(This example corresponds mathematically to Example 3.2.5).

We consider a sample $\mathbb{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$.

Definition 1.3.2 *An estimator (or statistic) T of a parameter of interest $\gamma \in \Gamma$ is a given (measurable) map $T: \mathcal{X}^n \to \Gamma$. We then also call $T(X_1, \ldots, X_n)$ an estimator (or statistic).*

Remark 1.3.1 *With some abuse of notation, we write shorthand*

$$T = T(\mathbb{X}) = T(X_1, \ldots, X_n).$$

That is we do not make a the distinction in notation between the map T and its evaluation at $\mathbb{X} = (X_1, \ldots, X_n)$. It should then be clear from the context what is meant. For example, we write $\mathbb{E}_\theta T =: \mathbb{E}_\theta T(\mathbb{X})$.

Remark 1.3.2 *Often we denote estimators with a "hat", e.g. $\hat{\gamma} = \hat{\gamma}(\mathbb{X})$ as estimator of γ.*

We present ways to construct estimators in the coming chapters. It depends on your creativity, your computational limits and the model assumptions you are prepared to make. Estimators should preferably "make sense": For example, having a law of large numbers in mind they should be close to what one is trying to estimate when the sample size n is large. What estimator would you use for γ in Example 1.3.1 about pitch of voice? Well, for $\{X_i = (Y_i, Z_i)\}_{i=1}^n$, being the data, a reasonable estimate of γ could be the value $\hat{\gamma}$ that makes the smallest number of errors in the sample, i.e.

$$|\{i: \ Y_i = 0, \ Z_i > \hat{\gamma}\} \cup \{Y_i = 1, \ Z_i < \hat{\gamma}\}|$$

$$\in \min_z |\{i: \ Y_i = 0, \ Z_i > z\} \cup \{Y_i = 1, \ Z_i < z\}|.$$

1.4 The law of large numbers as source of inspiration

The law of large numbers is an important result for developing statistical theory, and we use the abbreviation LLN. We recall that for $X_1, \ldots, X_n$ i.i.d. copies of $X \in \mathbb{R}$ where $E|X| < \infty$, the LLN says that the sample average $\bar{X} := \sum_{i=1}^n X_i/n$ is for n large close to the theoretical mean $\mu := EX$. More precisely, $\bar{X} = \bar{X}_n$ converges to μ in probability as $n \to \infty$. One has in fact convergence almost surely if $X_1, \ldots, X_n$ are the first n of an infinite sequence. Thus it makes sense to estimate μ by $\bar{X}$. Similarly, for a given function $g: \mathcal{X} \to \mathbb{R}$, inspired by the LLN, an estimator of $Eg(X)$ is $\sum_{i=1}^n g(X_i)/n$ and for a given (continuous) function $h: \mathbb{R} \to \mathbb{R}$ an estimator of $h(\mu)$ is $h(\bar{X})$, etc.

For example $\sigma^2 = EX^2 - \mu^2 = E(X-\mu)^2$ by definition, so the LLN leads to the estimator

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

An estimator $\hat{\gamma} = \hat{\gamma}_n$ of γ is called consistent if $\hat{\gamma}_n$ converges to γ in probability. In this lecture, we judge estimators that are consistent as “making sense”, but we will also see situations where the estimator makes sense, but proving its consistency is beyond the scope of this lecture and usually requires some additional conditions (see *Fundamentals of Mathematical Statistics* for consistency proofs).

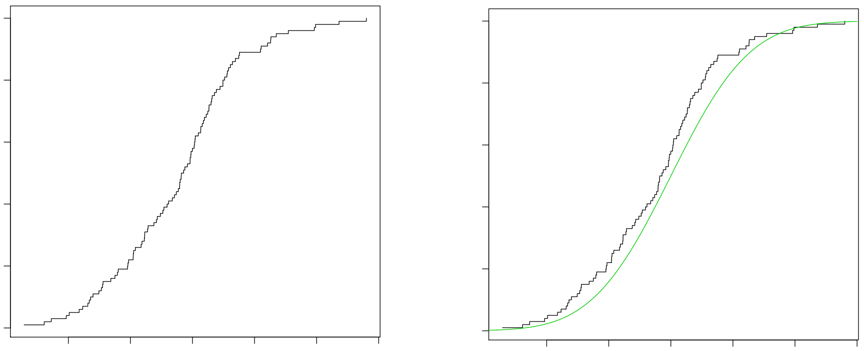
Let

$$F(x) := P(X \leq x),\ x \in \mathbb{R}$$

be the cumulative distribution function (CDF) of X. Again, inspired by the LLN, an estimator of F is

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{l}_{\{X_i \leq x\}},\ x \in \mathbb{R}.$$

The function $\hat{F}_n$ is called the empirical distribution function.



(a) $\hat{F}_n$ when $X \sim \mathcal{N}(0,1)$ and $n = 100$ (b) The theoretical $F (= \Phi)$ in green

Figure 1.1: A realization of the empirical distribution function $\hat{F}_n$ in 1.1a and comparison with the theoretical distribution function F in 1.1b.

One can use similar inspirations when $\mathcal{X}$ is not the real line ($\mathcal{X} = \mathbb{R}^k$ for example, or even a more abstract space). We will encounter these later.

1.5 Mean square error

Ideally, we aim at estimators of a parameter of interest that are in some sense "good". Then one needs to make precise what is meant by "good". For real-valued parameters of interest $\gamma \in \mathbb{R}$ the mean square error is a popular criterion for accessing the performance of an estimator $T \in \mathbb{R}$. Another criterion is unbiasedness.

Definition 1.5.1 *The mean square error (MSE) of an estimator $T \in \mathbb{R}$ of $\gamma := g(\theta) \in \mathbb{R}$ is*

$$\text{MSE}_\theta(T) = E_\theta(T - g(\theta))^2.$$

The bias of T is

$$\text{bias}_\theta(T) = E_\theta T - g(\theta).$$

The estimator T is called unbiased if

$$\text{bias}_\theta(T) = 0, \ \forall \ \theta \in \Theta.$$

A little warning may be in place: MSE (and bias) may be difficult to compute exactly. For instance, in Example 1.3.1 about pitch of voice, the MSE of $\hat{\gamma}$ has to the best of our knowledge never been considered exactly. This warning indicates that we can only handle "toy" examples. Note moreover that the MSE depends the underlying unknown distribution, and hence is typically unknown. A further warning is that unbiased estimators often do not exist and if they do they cannot stand non-linear transformations! That nevertheless MSE and bias remain important throughout the statistical literature comes from the fact that there is much theory on approximate MSE and bias (e.g. using "asymptotics" beyond LLN's).

The following lemma presents the famous bias-variance decomposition for MSE. If you like, it is Pythagoras' rule in abstract terms.

Lemma 1.5.1

$$\text{MSE}_\theta(T) = \text{bias}^2_\theta(T) + \text{Var}_\theta(T).$$

Proof. Write $q(\theta) := E_\theta(T)$. Then

$$\begin{aligned}
\text{MSE}_\theta(T) &= E_\theta\biggl(T - q(\theta) + q(\theta) - g(\theta)\biggr)^2 \\
&= E_\theta\biggl(T - q(\theta)\biggr)^2 + \biggl(q(\theta) - g(\theta)\biggr)^2 \\
&+ 2\biggl(q(\theta) - g(\theta)\biggr)\underbrace{E_\theta\biggl(T - q(\theta)\biggr)}_{=0} \\
&= \text{Var}_\theta(T) + \text{bias}^2_\theta(T).
\end{aligned}$$

□

Example 1.5.1 *Let $X_1, \ldots, X_n$ be i.i.d. copies of $X \in \mathbb{R}$ where $EX =: \mu$ and $\mathrm{Var}(X) =: \sigma^2$. Then the sample average $\bar{X} = \sum_{i=1}^n X_i/n$ is an unbiased estimator of μ.*

For $n \geq 2$ the sample variance $S^2 := \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ is an unbiased estimator of σ^2. To see this, note that

$$\sum_{i=1}^n (X_i - \mu - (\bar{X} - \mu))^2 = \sum_{i=1}^n (X_i - \mu)^2 - (\bar{X} - \mu)^2$$

(the latter being again Pythagoras' rule) and

$$\mathbb{E}(X_i - \mu)^2 = \sigma^2, \ \mathbb{E}(\bar{X} - \mu)^2 = \sigma^2/n.$$

Thus $\mathbb{E}S^2 = \sigma^2$. The estimator

$$\hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

that was inspired by the LLN is not unbiased. If one wants to compare the mean square error of S^2 and $\hat{\sigma}^2$ one needs to calculate the variance of $\sum_{i=1}^n (X_i - \bar{X})^2$ which requires additional distributional assumptions. It turns out that when $X_i \sim \mathcal{N}(\mu, \sigma^2)$ ($\forall$ i), then actually $\hat{\sigma}^2$ wins from S^2 in terms of MSE: $\hat{\sigma}^2$ is biased but has smaller variance! Of course for $n \to \infty$ the difference in MSE's disappears.

We further observe that S is generally **not** *an unbiased estimator of σ: by Jensen's inequality*

$$ES \leq \sqrt{(ES^2)} = \sigma$$

with equality only in the degenerate case where $\mathrm{var}(S) = 0$. Nevertheless, albeit biased, S remains a "reasonable" estimator by the LLN. Thus, non-linear transformations generally ruin unbiasedness and one often need not to be too upset about that.

1.6 The central limit theorem with estimated variance

Let for $n \geq 1$, $X_1, \ldots, X_n$ be i.i.d. copies of $X \in \mathbb{R}$ where $EX =: \mu$ and $\mathrm{var}(X) =: \sigma^2 < \infty$. Let $\bar{X}_n := \sum_{i=1}^n X_i/n$ be the average of $X_1, \ldots, X_n$. By the central limit theorem (which we abbreviate to CLT)

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq z \right) = \Phi(z) \ \forall z \in \mathbb{R}$$

where Φ is the standard normal distribution function. This result is frequently applied in statistics to construct approximate confidence intervals for the unknown μ when the data are $X_1, \ldots, X_n$ (as we will do in the next section).

However, in most statistical situations the variance σ^2 is unknown as well. One can show ("Slutsky's Theorem", see *Fundamental of Mathematical Statistics*) that for $\hat{\sigma}_n^2 > 0$ a sequence of random variables which converges in probability to σ^2, the CLT still holds with σ^2 replace by $\hat{\sigma}_n^2$:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}_n} \le z \right) = \Phi(z) \ \forall z \in \mathbb{R}.$$

1.7 An example using the central limit theorem

The example in this section serves as a look ahead: more theory is to follow in Chapter 8. It will illustrate that there a several ways to tackle statistical problems (for example several ways to estimate a parameter of interest). Then one would like to know what the best approach is. And the answer is: it depends!

We illustrate the use of the CLT here for the case where $X \sim \text{Poisson}(\lambda)$, with $\lambda > 0$ an unknown parameter (see Appendix A for a definition of the Poisson and other distributions). Suppose we observe $X_1, \ldots, X_n$, i.i.d. Poisson(λ)-distributed random variables. We estimate $E_\lambda X = \lambda$ by the sample average $\hat{\lambda} := \bar{X}$ (omitting the subscript n). It holds that $\mathbb{E}_\lambda \bar{X} = \lambda$ for all $\lambda > 0$ so $\bar{X}$ is unbiased. Moreover $\text{Var}_\lambda(\bar{X}) = \lambda/n$. By the CLT, $\bar{X}$ is approximately $\mathcal{N}(\lambda, \lambda/n)$-distributed for n large. Thus for all $z \ge 0$

$$\mathbb{P}_\lambda\Bigg(|\bar{X} - \lambda| \le z\sqrt{\lambda/n} \Bigg) \approx \Phi(z) - \Phi(-z) = 2\Phi(z) - 1.$$

Now we choose $z = 1.96$ which gives $2\Phi(z) - 1 = 2\Phi(1.96) - 1 = 2(0.975) - 1 = 0.95$. To clean up the formula's[3], we replace $1.96 \approx 2$ by 2.

Moreover we let

$$\begin{aligned} I_1(\bar{X}) &:= \Bigg\{ \lambda > 0 : \ |\bar{X} - \lambda| \le 2\sqrt{\lambda/n} \Bigg\} \\ &= \Bigg\{ \lambda > 0 : \bar{X} + \frac{2}{n} - 2\sqrt{\frac{\bar{X} + 1/n}{n}} \le \lambda \le \bar{X} + \frac{2}{n} + 2\sqrt{\frac{\bar{X} + 1/n}{n}} \Bigg\} \end{aligned}$$

where the second equality follows from some computations. Then we have by the CLT

$$\mathbb{P}_\lambda\Bigg(\lambda \in I_1(\bar{X}) \Bigg) \approx 0.95.$$

We call $I_1(\bar{X})$ an approximate 95% confidence interval for λ.

An alternative way to use the CLT to build a confidence interval for λ is based on an estimate of the variance of $\bar{X}$:

$$\widehat{\text{Var}}_\lambda(\bar{X}) := \hat{\lambda}/n = \bar{X}/n.$$

[3] Rule of thumb: an approximate 95 % confidence interval for $\gamma \in \mathbb{R}$ is $\hat{\gamma} \pm 2\times$ the (estimated) standard deviation of $\hat{\gamma}$, provided $\hat{\gamma} - \gamma$ is approximately a centered normally distributed random variable.

As stated in Section 1.6, the CLT still holds with this estimated variance

$$\mathbb{P}_\lambda\left(|\bar{X}-\lambda| \le z\sqrt{\bar{X}/n}\right) \approx 2\Phi(z)-1.$$

Let now

$$\begin{aligned} I_2(\bar{X}) &:= \left\{\lambda>0:\ |\bar{X}-\lambda| \le 2\sqrt{\bar{X}/n}\right\} \\ &= \left\{\lambda>0:\ \bar{X}-2\sqrt{\bar{X}/n} \le \lambda \le \bar{X}+2\sqrt{\bar{X}/n}\right\}. \end{aligned}$$

Then (using again $2 \approx 1.96$), $I_2(\bar{X})$ is approximately a 95% confidence interval for λ:

$$\mathbb{P}_\lambda\left(\lambda \in I_2(\bar{X})\right) \approx .95.$$

The two intervals $I_1(\bar{X})$ and $I_2(\bar{X})$ are for n large approximately equal (the first one is slightly more conservative).

One may also use $S^2 = \sum_{i=1}^n (X_i-\bar{X})^2/(n-1)$ as estimator of the variance of $\bar{X}$ and use this estimator in the CLT for $\bar{X}$. This would give a third confidence interval $I_3(\bar{X}, S^2)$.

Since $\text{Var}_\lambda(X) = \lambda$ we see that S^2 is also an alternative estimator of λ. One may ask which one of the two estimators, $\bar{X}$ or S^2, is "better". One may want to compare them by calculating the MSE's of the two estimators (calculation the MSE of S^2 is not an easy exercise). One could try to apply a CLT for S^2 instead of $\bar{X}$ (this is indeed possible) and base a confidence interval for λ on that. Then one needs to estimate the (asymptotic) variance of S^2 (which is possible too).

Statistical theory says that the differences between the confidence intervals $I_1(\bar{X})$, $I_2(\bar{X})$ and $I_3(\bar{X}, S^2)$ vanish as $n \to \infty$, provided that the Poisson model is correct. If the model may be wrong, $I_3(\bar{X}, S^2)$ is a safer (more conservative) choice than $I_1(\bar{X})$ and $I_2(\bar{X})$. The fourth approach where the confidence interval is based on S^2 instead of $\bar{X}$ is asymptotically valid under the assumption that the model is correct, but it is more conservative than $I_1(\bar{X})$, $I_2(\bar{X})$ and $I_3(\bar{X}, S^2)$.

If the sample size is small, one may prefer to construct exact confidence intervals for λ instead of approximate ones. This is possible too, see Chapter 9.

Numerical example

In a numerical example (and in real life), one sees the "realizations" of the random variables involved. These realizations are denoted with lower case letters. A realization of an estimator is called an estimate.

This is from Example 10.19 in DasGupta [2011]. Let the data be

x_i	# days
0	100
1	60
2	32
3	8
≥ 4	0

Thus $n = 200$ and the observed value for $\bar{X}$ is $\bar{x} = .74$. Then an approximate 95% confidence interval for λ is

$$I_2(\bar{x}) = \bar{x} \pm 2\sqrt{\bar{x}/n} = [0.62, 0.84].$$

Let $\gamma := g(\lambda) := P_\lambda(X \geq 4)$ be the parameter of interest. Then

$$\hat{\gamma} = \widehat{g(\lambda)} := g(\hat{\lambda}) = g(\bar{x}) = .00697,$$

and, since $\lambda \mapsto g(\lambda)$ is a monotone function, an approximate 95% confidence interval for γ is

$$g\left(\bar{x} \pm 2\sqrt{\bar{x}/n}\right) = [0.0038, 0.01].$$

Suppose now we estimate the variance $\text{Var}_\lambda(X)$ of X by the sample variance S^2.

$x_i - \bar{x}$	$(x_i - \bar{x})^2$	# days
-.74	.5476	100
.26	.0676	60
1.26	1.5876	32
2.26	5.1076	8

We find that the observed value of S^2 is $s^2 := \sum_{i=1}^n (x_i - \bar{x})^2/(n-1) = .7561$. Since $\text{Var}_\lambda(X) = \lambda$, both $\bar{x} = .74$ and $s^2 = .7561$ are unbiased estimates of λ. The fact that these values are not very different can be seen as an indication that the Poisson model is appropriate.

Invoking s^2 to construct another approximate 95% confidence interval for λ yields

$$I_3(\bar{x}, s^2) = \bar{x} \pm 2\sqrt{s/n} = [0.62, 0.86].$$

Chapter 2

The method of moments

The method of moments is a procedure for constructing an estimator of the parameter describing the distribution, when this parameter is finite-dimensional, say of dimension d.

Let $X \in \mathbb{R}$ and let the data $X_1, \ldots, X_n$ be i.i.d. copies of X.

Definition 2.0.1 *For $k \in \mathbb{N}$ the k-th moment of X is*

$$\mu_k := EX^k$$

(if the expectation exists).

Definition 2.0.2 *The k-th sample moment (or empirical moment) is*

$$\hat{\mu}_k := \frac{1}{n} \sum_{k=1}^{n} X_i^k, \ k \in \mathbb{N}.$$

Note By the LLN $\hat{\mu}_k \approx \mu_k$ for n large (provided the moment exists).

2.1 Definition of the method of moments estimator

Suppose that X has distribution P_θ, where $\theta \in \Theta \subset \mathbb{R}^d$. Then the moments of X also depend on θ:

$$\mu_k = \mu_k(\theta) = E_\theta X.$$

Definition 2.1.1 *The methods of moments estimator $\hat{\theta}$ is a solution of*

$$\mu_k(\vartheta)_{\vartheta = \hat{\theta}} = \hat{\mu}_k, \ k = 1, \ldots, d.$$

(assuming a solution exists).

Thus with the method of moments, one creates d equations with d unknowns and tries to solve these. These d equations are based on the sample moments. The parameter θ is a solution of the d equations with the sample moments replaced by the theoretical moments. Since the sample moments are close to to the theoretical moments by the law of large numbers, the estimator $\hat{\theta}$ "makes sense": if the inverse map of $\vartheta \mapsto \{\mu_k(\vartheta)\}_{k=1}^d$ is continuous, then $\hat{\theta}$ will be close θ.

2.2 Examples

Example 2.2.1 *Let the data $X_1, \ldots, X_n$ be i.i.d. copies of $X \sim \mathcal{N}(\mu, \sigma^2)$, where both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown. Then the methods of moments estimator is*

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

Example 2.2.2 *Let $X \sim \mathrm{Gamma}(\alpha, \lambda)$ (see Appendix A):*

$$E_\theta X = \alpha/\lambda,\ \mathrm{Var}_\theta(X) = \alpha/\lambda^2.$$

Then $E_\theta X^2 = \alpha(\alpha+1)/\lambda^2$. So the methods of moments estimator $(\hat{\alpha}, \hat{\lambda})$ solve the two equations

$$\hat{\mu}_1 = \hat{\alpha}/\hat{\lambda},\ \hat{\mu}_2 - \hat{\mu}_1^2 = \hat{\alpha}/\hat{\lambda}^2.$$

It follows that

$$\hat{\lambda} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2},\ \hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}.$$

Example 2.2.3 *Let the data $X_1, \ldots, X_n$ be i.i.d. copies of X where X has Lebesgue density*

$$p_\theta(x) = \frac{1 + \theta x}{2},\ -1 \le x \le 1,\ -1 \le \theta \le 1.$$

Then

$$E_\theta(X) = \frac{\theta}{3}.$$

The methods of moments estimator is thus $\hat{\theta} = 3\bar{X}$.

Example 2.2.4 *(Gaussian mixture) Let X have density*

$$p_\theta(x) := \pi_1 \frac{1}{\tau_1}\phi\biggl(\frac{x - \nu_1}{\tau_1}\biggr) + (1 - \pi_1)\frac{1}{\tau_2}\phi\biggl(\frac{x - \nu_2}{\tau_2}\biggr)$$

where ϕ is the standard normal density. To simplify, we assume that $\pi_1 = \frac{1}{2}$, $\nu_1 = 0$ and $\tau_1 = 1$ are given. We write $\nu := \nu_2$ and $\tau := \tau_2$. The unknown parameter is $\theta = (\nu, \tau)$. We have

$$EX = \frac{1}{2}\nu,\ EX^2 = \frac{1}{2} + \frac{1}{2}(\nu^2 + \tau^2).$$

So the method of moments estimator $(\hat{\nu}, \hat{\tau})$ solve

$$\frac{1}{2}\hat{\nu} = \hat{\mu}_1, \ \frac{1}{2} + \frac{1}{2}(\hat{\nu}^2 + \hat{\tau}^2) = \hat{\mu}_2.$$

Hence

$$\hat{\nu} = 2\hat{\mu}_1, \ \hat{\tau}^2 = 2\hat{\mu}_2 - 4\hat{\mu}_1^2 - 1.$$

2.3 Plug in method

The method of moments is inspired by the LLN, but the LLN can also be a source of inspiration for further constructions. The idea is to mimic the unknown theoretical parameter of interest by its empirical counterpart. We present two examples.

Example 2.3.1 *Let $(X,Y) \in \mathbb{R}^2$. The best linear predictor of Y given X is defined as $\alpha + \beta X$ where*

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} := \arg\min \left\{ E\left(Y - (a+bX) \right)^2 : \ \begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^2 \right\}.$$

*Here "*arg*" stands for "argument", i.e. the location of (in this case) the minimum. By direct calculations one sees that*

$$\alpha = EY - \beta EX, \ \beta = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}.$$

Let now $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. copies of (X,Y). Then, the LLN leads to the estimators

$$\hat{\alpha} := \bar{Y} - \hat{\beta}\bar{X}, \ \hat{\beta} := \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}.$$

The estimator $(\hat{\alpha}, \hat{\beta})^\top$ is called the least squares estimator. Note that

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \arg\min \left\{ \frac{1}{n} \sum_{i=1}^n \left(Y_i - (a + bX_i) \right)^2 : \begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^2 \right\},$$

see also Example 11.1.1.

Example 2.3.2 *Let $X \in \mathbb{R}$ have CDF F. Assume the median $m := F^{-1}(\frac{1}{2})$ exists. Let $\hat{F}_n$ be the empirical distribution function (see Section1.4). We can estimate m by a solution $\hat{m}$ of $\hat{F}_n(\hat{m}) \approx \frac{1}{2}$. The sample median is*

$$\hat{m} := \begin{cases} X_{(\frac{n+1}{2})}, & n \text{ odd} \\ \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}, & n \text{ even} \end{cases}.$$

Here $X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics.

Chapter 3

Maximum likelihood

Maximum likelihood is another way to construct estimators. Let $X \in \mathcal{X}$ and suppose that $X \sim P_\theta \in \mathcal{P} := \{P_\vartheta : \vartheta \in \Theta\}$. In order to be able to define the maximum likelihood estimator of θ we need to assume that the family $\mathcal{P}$ is dominated by some sigma-finite measure ν. We then call, for $\vartheta \in \Theta$

$$p_\vartheta := \frac{dP_\vartheta}{d\nu}$$

the density of P_ϑ (with respect to ν). Typically, we consider one of the two cases:

- The space $\mathcal{X}$ is finite or countably infinite. Then we can take ν as the counting measure, and for $x \in \mathcal{X}$

  $$p_\vartheta(x) = P_\vartheta(\{x\}),$$

  which we write with some abuse of notation as $p_\vartheta(x) = P_\vartheta(X = x)$.

- The space $\mathcal{X}$ is a subset of $\mathbb{R}^k$ and ν is Lebesgue measure. Then P_ϑ is absolutely continuous, and p_ϑ is the Lebesgue density of P_ϑ. In that case, for F_ϑ the CDF of P_ϑ, we can take for ν-almost all $x = (\xi_1, \ldots, \xi_k)$

  $$p_\vartheta(x) = \frac{\partial^k}{\partial \xi_1 \cdots \partial \xi_k} F(\xi_1, \ldots, \xi_k).$$

3.1 Definition of the maximum likelihood estimator

Let $\mathbb{X} = (X_1, \ldots, X_n)$ be a sample of size n of i.i.d. copies of X. We use the notation: for a real-valued function f on some domain $\mathcal{Z}$: $\arg\max_{z \in \mathcal{Z}} f(z) :=$ the location of the maximum of f.

Definition 3.1.1 *The likelihood function is*

$$L_{\mathbb{X}} : \Theta \to \mathbb{R},$$

with

$$L_{\mathbb{X}}(\vartheta) := \prod_{i=1}^n p_\vartheta(X_i), \ \vartheta \in \Theta.$$

The *maximum likelihood estimator (MLE)* *of* θ *is*

$$\hat{\theta} = \hat{\theta}_{\text{MLE}} := \arg\max_{\vartheta \in \Theta} L_{\mathbb{X}}(\vartheta)$$

(assuming the maximum exists).

One may ask why maximum likelihood "makes sense". Is the MLE $\hat{\theta}$ close to θ when n is large? The answer is: yes, under certain conditions it is. In the background there is again the LLN which indicates maximum likelihood is potentially a good idea. We will not give the theory here, see Remark 3.1.2 for a first hint and see for example the lecture *Fundamentals of Mathematical Statistics.*

Maximum likelihood also intuitively "makes sense". Here is an example. Suppose you throw a coin $n = 10$ times. The probability θ of heads is unknown, but suppose we know that either $\theta = 1/2$ or $\theta = 1/4$, i.e. $\Theta = \{1/2, 1/4\}$. Now after throwing the coin, one finds 7 heads. What would you then be your estimate of θ? I would say $\hat{\theta} = 1/2$ because we found many heads, which makes the value $\vartheta = 1/2$ more likely than the value $\vartheta = 1/4$:

$$\mathbb{P}_{\vartheta=1/2}(7 \text{ heads}) = \binom{10}{7}\left(\frac{1}{2}\right)^{10} = 0.117,$$

and

$$\mathbb{P}_{\vartheta=1/4}(7 \text{ heads}) = \binom{10}{7}\left(\frac{1}{4}\right)^{7}\left(\frac{3}{4}\right)^{4} = 0.016.$$

In other words, $L_{\mathbb{X}=7}(1/2) = 0.117$, $L_{\mathbb{X}=7}(1/4) = 0.016$.

One may note that the likelihood function is nothing else then the density of $\mathbb{X}$, which is $\prod_{i=1}^n p_\vartheta(x_i)$, evaluated at $(x_1, \ldots, x_n)$ being the sample $\mathbb{X} = (X_1, \ldots, X_n)$. The difference between the concept likelihood and the concept density is that the likelihood function considers $\prod_{i=1}^n p_\vartheta(x_i)$ as function of the parameter ϑ, whereas the density considers $\prod_{i=1}^n p_\vartheta(x_i)$ as function of $(x_1, \ldots, x_n)$.

Remark 3.1.1 *Since* $z \mapsto \log z$, $z > 0$ *is a monotone transformation, one may also maximize the log-likelihood* $\log L_{\mathbb{X}}$.

$$\hat{\theta} = \hat{\theta}_{\text{MLE}} = \arg\max_{\vartheta \in \Theta} \log L_{\mathbb{X}}(\vartheta) = \arg\max_{\vartheta \in \Theta} \sum_{i=1}^n \log p_\vartheta(X_i).$$

If $\Theta \subset \mathbb{R}^d$ *is finite-dimensional, the MLE can often (not always!) be obtained by setting the derivative of the log-likelihood to zero:*

$$\sum_{i=1}^n s_{\hat{\theta}}(X_i) = 0,$$

where

$$s_\vartheta(\cdot) := \frac{\partial}{\partial\vartheta}\log p_\vartheta(\cdot).$$

Remark 3.1.2 *(LLN as source of inspiration) One can show that*

$$\theta = \arg\max_{\vartheta\in\Theta} E_\theta \log p_\vartheta(X),$$

and also, when $\Theta \subset \mathbb{R}^d$ and under regularity conditions,

$$E_\theta s_\theta(X) = 0,\ s_\vartheta := \frac{\partial}{\partial\vartheta}\log p_\vartheta.$$

3.2 Examples

Example 3.2.1 *Let the data be $X_1, \ldots, X_n$ be i.i.d. copies of $X \sim \mathcal{N}(\mu, \sigma^2)$, where both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown, i.e. $\theta = (\mu, \sigma^2)$. Writing $\vartheta := (\tilde{\mu}, \tilde{\sigma}^2)$ the log-likelihood is*

$$L_{\mathbb{X}}(\vartheta) = \sum_{i=1}^n \log p_\vartheta(X_i) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tilde{\sigma}^2 - \frac{\sum_{i=1}^n (X_i - \tilde{\mu})^2}{2\tilde{\sigma}^2}.$$

Taking derivatives w.r.t. $\tilde{\mu}$ gives

$$\frac{\sum_{i=1}^n (X_i - \hat{\mu}_{\text{MLE}})}{\hat{\sigma}^2_{\text{MLE}}} = 0,$$

so that $\hat{\mu}_{\text{MLE}} = \bar{X}$. As

$$\bar{X} = \arg\min_{\tilde{\mu}} \sum_{i=1}^n (X_i - \tilde{\mu})^2,$$

it is also called the least squares estimator (LSE) of μ.

Inserting $\hat{\mu}_{\text{MLE}} = \bar{X}$ and differentiating w.r.t. $\tilde{\sigma}^2$ gives

$$-\frac{n}{2\hat{\sigma}^2_{\text{MLE}}} + \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{2\hat{\sigma}^4_{\text{MLE}}} = 0$$

so $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. Thus, in this case the MLE equals the method of moments estimator (see 2.2.1).

Example 3.2.2 *Let the data $X_1, \ldots, X_n$ be i.i.d. copies of $X \sim \text{Laplace}(\mu, \sigma^2)$, where both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown, i.e. $\theta = (\mu, \sigma^2)$. The (Lebesgue) density of X is*

$$p_\theta(x) = \frac{1}{2\sigma}\exp\left[-\frac{|x-\mu|}{\sigma}\right],\ x \in \mathbb{R}.$$

The log-likelihood based on the sample $\mathbb{X} = (X_1, \ldots, X_n)$ is

$$L_{\mathbb{X}}(\vartheta) = \sum_{i=1}^n \log p_\vartheta(X_i) = -n\log 2 - n\log\tilde{\sigma} - \frac{\sum_{i=1}^n |X_i - \tilde{\mu}|}{\tilde{\sigma}},\ \vartheta = (\tilde{\mu}, \tilde{\sigma}).$$

It follows that

$$\hat{\mu}_{\text{MLE}} = \arg\min_{\tilde{\mu}} \sum_{i=1}^{n} |X_i - \tilde{\mu}|.$$

For n even the minimizer is not unique. We take the sample median

$$\hat{\mu}_{\text{MLE}} = \hat{m} := \begin{cases} X_{(\frac{n+1}{2})} & n \text{ odd} \\ \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2} & n \text{ even} \end{cases}$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics (see also Section 1.4). The sample median is often called the least absolute deviations (LAD) estimator of μ.

What is still left to do in this example is to calculate the MLE of σ. By differentiating the log-likelihood w.r.t. $\tilde{\sigma}$ one gets

$$-\frac{n}{\hat{\sigma}_{\text{MLE}}} + \frac{\sum_{i=1}^{n} |X_i - \hat{m}|}{\hat{\sigma}^2_{\text{MLE}}} = 0,$$

which gives $\hat{\sigma}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \hat{m}|$.

Let us briefly present an alternative view how LLN can make sense out of the estimator $\hat{m} \approx \arg\min_{\tilde{\mu}} \sum_{i=1}^{n} |X_i - \tilde{\mu}|/n$, even when the data are not Laplacian. One may verify that

$$E|X - \tilde{\mu}| = 2 \int_{x > \tilde{\mu}} (1 - F(x)) dx + \tilde{\mu} - EX,$$

where F is the CDF of X. One can find

$$\arg\min_{\tilde{\mu}} E|X - \tilde{\mu}|$$

by setting the derivative of $E|X - \tilde{\mu}|$ to zero

$$-2(1 - F(\tilde{\mu}))|_{\tilde{\mu} = \arg\min} + 1 = 0.$$

In other words

$$\arg\min_{\tilde{\mu}} E|X - \tilde{\mu}| = F^{-1}(\tfrac{1}{2}),$$

is the theoretical median (provided it exists).

Remark 3.2.1 *Estimating the mean EX by the LSE $\bar{X}$ remains a valid procedure also for non-Gaussian data. Similarly, the LAD estimator $\hat{m}$ remains a valid estimator of the median $F^{-1}(\frac{1}{2})$ also when the data are not Laplacian.*

Example 3.2.3 *Let the data be $X \sim \text{Binomial}(n, \theta)$, where the success probability $0 < \theta < 1$ is unknown. Then for $x \in \{0, 1, \ldots, n\}$*

$$p_{\vartheta}(x) = P_{\vartheta}(X = x) = \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x},$$

and

$$\log L_X(\vartheta) = \log p_\vartheta(X) = \log \binom{n}{X} + X \log \vartheta + (n - X) \log(1 - \vartheta).$$

We have

$$\frac{d}{d\vartheta} \log p_\vartheta(X) = \frac{X}{\vartheta} - \frac{n - X}{1 - \vartheta}.$$

Setting this to zero gives

$$\frac{X}{\hat{\theta}_{\rm MLE}} - \frac{n - X}{1 - \hat{\theta}_{\rm MLE}} = 0,$$

giving

$$\hat{\theta}_{\rm MLE} = \frac{X}{n}.$$

Example 3.2.4 *Let the data $X_1, \ldots, X_n$ be i.i.d. copies of $X \in \{1, \ldots, q\}$. For example, X represents a "class label". The probability of a particular label is unknown:*

$$P_\theta(X = j) := \theta_j,\ j = 1, \ldots, q,$$

where

$$\theta \in \Theta = \Big\{ \vartheta \in \mathbb{R}^q :\ \vartheta_j \ge 0\ \forall\ j,\ \sum_{j=1}^q \vartheta_j = 1 \Big\}.$$

We may write

$$\log p_\vartheta(x) = \sum_{j=1}^q {\rm l}_{\{x=j\}} \log \vartheta_j.$$

Hence the log-likelihood based on $\mathbb{X} = (X_1, \ldots, X_n)$ is

$$L_{\mathbb{X}}(\vartheta) = \sum_{i=1}^n \log p_\vartheta(X_i) = \sum_{i=1}^n \sum_{j=1}^q {\rm l}_{\{X_i=j\}} \log \vartheta_j = \sum_{j=1}^q N_j \log \vartheta_j,$$

where $N_j := \sum_{i=1}^n {\rm l}_{\{X_i=j\}} = \#\{X_i = j\}$ counts the number of observations with the label j ($j = 1, \ldots, q$). To find the maximum of the log-likelihood under the restriction that $\sum_{j=1}^q \vartheta_j = 1$ we use a Lagrange multiplier λ: we maximize

$$\sum_{j=1}^q N_j \log \vartheta_j + \lambda \Big(1 - \sum_{j=1}^q \vartheta_j \Big).$$

Differentiating and setting to zero gives for the MLE $\hat{\theta}$

$$\frac{\partial}{\partial \vartheta_j} \Big\{ \sum_{j=1}^q N_j \log \vartheta_j + \lambda \Big(1 - \sum_{j=1}^q \vartheta_j \Big) \Big\} \Big|_{\hat{\theta}} = \frac{N_j}{\hat{\theta}_j} - \lambda = 0.$$

Thus

$$\hat{\theta}_j = \frac{N_j}{\lambda},\ j = 1, \ldots, q.$$

The restriction now gives

$$1 = \sum_{j=1}^{q} \frac{N_j}{\lambda},$$

and since $\sum_{j=1}^{q} N_j = n$ *we obtain* $\lambda = n$. *The MLE is therefore*

$$\hat{\theta}_j = \frac{N_j}{n},\ j = 1, \ldots, q.$$

Example 3.2.5 *This example concerns a case where the parameter space is infinite-dimensional. We present it to illustrate that maximum likelihood can also be used when the parameter is non-Euclidean (see e.g. Groeneboom and Wellner [1992] for more about nonparametric maximum likelihood and in particular the problem described here). This example is not part of the exam.*

Let Z *be the arrival time of (slow) mail. The arrival time* Z *is never observed exactly. You check your (physical) mailbox every day at a random time* T. *Then either the mail arrived:* $Y = 1$, *or it did not:* $Y = 0$. *Aim is now to estimate the distribution of* Z. *The problem is called "interval censored". Let* F *be the CDF of* Z. *We have* $P(Y = 1|T = t) = F(t)$ *and* $P(Y = 0|T = t) = 1 - F(t)$. *Thus the density (with dominating measure the distribution of* T*) is*

$$p_F(y, t) = F^y(t)(1 - F(t))^{1-y}$$

and so

$$\log p_F(y, t) = y \log F(t) + (1 - y) \log(1 - F(t)).$$

Having checked the mailbox for n *days, the data are i.i.d. copies* $\mathbb{X} = \{Y_i, T_i\}_{i=1}^n$, *of* $X = (Y, T)$. *The log-likelihood is*

$$L_{\mathbb{X}}(\tilde{F}) = \sum_{i=1}^{n} \biggl(Y_i \log \tilde{F}(T_i) + (1 - Y_i) \log(1 - \tilde{F}(T_i)) \biggr),$$

where the parameter $\tilde{F}$ *ranges over the parameter space* $\mathcal{F}$ *of all CDF's. The ("nonparametric") MLE is*

$$\hat{F}_{\text{MLE}} := \arg\max_{\tilde{F} \in \mathcal{F}} L_{\mathbb{X}}(\tilde{F}).$$

Questions are now: does it exist, how to compute it, what are its properties? (This example is closely related to Example 1.3.1 when the parameter of interest is $F^{-1}(1/2)$.*)*

Chapter 4

Hypothesis testing

In this chapter, we denote the data by X, i.e. we replace $\mathbb{X}$ by X (and $\mathbb{P}$ by P). This makes the notation less Baroque.

Let $X \in \mathcal{X}$, $X \sim P_\theta$, $\theta \in \Theta$. We consider two hypotheses about the parameter θ: for $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$
$H_0 : \ \theta \in \Theta_0$ the null hypothesis,
$H_1 : \ \theta \in \Theta_1$ the alternative hypothesis.

Example Let $X \sim \text{Binomial}(n, \theta)$ and
$H_0 : \ \theta = \frac{1}{2}$,
$H_1 : \ \theta = \frac{3}{4}$.
Suppose we observe the value $X = 14$. We have
$P_{H_0}(X = 14) = .074$,
$P_{H_1}(X = 14) = .112$.
We see that the likelihood $P_{H_1}(X = 14)$ is larger than the likelihood $P_{H_0}(X = 14)$. The value $\theta = \frac{3}{4}$ is the maximum likelihood estimate over $\{\frac{1}{2}, \frac{3}{4}\}$. The likelihood ratio is

$$\frac{P_{H_1}(X = 14)}{P_{H_0}(X = 14)} = 1.51.$$

Is this large enough to reject H_0 in favour of H_1?

To answer the question in the above example, we need to agree on a criterion for evaluating whether or not rejecting the null hypothesis is a good decision. The point of view one uses in statistical hypothesis testing is that the null hypothesis H_0 represents a situation where "everything is as usual", or "no evidence found". For example[1], if it concerns the decision of putting someone in prison (for murder) or not, it makes sense to choose
H_0 : the person is innocent,
H_1 : the person is guilty,
when convicting an innocent person is an error considered worse than not to convict a guilty person. The Bayesian approach is to put a prior on H_0 and H_1

[1] The use of statistics in the court room is under debate. We only use this illustration to explain the idea of hypothesis testing more vividly.

(see Chapter 10), i.e. in the above example, a prior belief whether a person is a murderer. In the frequentist approach, no prior is used.

4.1 Definition of a test

We can make two errors: rejecting H_0 (accepting H_1) when H_0 is true (error first kind) and not rejecting H_0 when H_1 is true (error second kind). It is (generally) not possible to keep **both** errors under control. The idea is now to keep the probability of the error of first kind below a (small) prescribed value α.

	H_0	H_1
$\phi = 1$	error first kind	probability = power
$\phi = 0$		error second kind

Definition 4.1.1 *A statistical test*[2] *at given level* α *(*$0 < \alpha < 1$*) is a (measurable) map* $\phi : \mathcal{X} \to \{0, 1\}$ *such that*

$$\phi(X) = \begin{cases} 1 & \text{means } H_0 \text{ is rejected} \\ 0 & \text{means } H_0 \ \text{ is not rejected} \end{cases},$$

and such that

$$P_{\theta_0}(\phi(X) = 1) \le \alpha \ \forall \ \theta_0 \in \Theta_0.$$

The power of the test at $\theta_1 \in \Theta_1$ *is* $P_{\theta_1}(\phi(X) = 1)$*.*

Thus, in a loose notation, $\phi = 1$ is in favour of H_1 and $\phi = 0$ is in favour of H_0. I.e. the decision is

$$H_\phi = \begin{cases} H_1, & \phi = 1 \\ H_0, & \phi = 0 \end{cases}.$$

A statistical test is often based on a real-valued test statistic, say $T = T(X)$, such that $\phi(X) = 1$ iff $T(X) > c$, where c is called the critical value of the test.

Once the null hypothesis is rejected, this can be reason for further research. It may also be a reason for publication of the findings, and then the results should be reproducible. If the null hypothesis is not rejected, one says that the result is not significant. One can see this too as an interesting result that might be publishable. However, one should be careful here, as it could just be due to a lack of power of the test. For example, if pharmaceutic industry wants to show that the effect of a new (cheaper to produce) drug is not significantly different from the existing drug (bio-equivalence), it could stir towards a non-significant effect by basing the test on very little test persons.

[2] We extend this to "randomized" tests $\phi : \mathcal{X} \to [0, 1]$ in the next section

Example $X \sim \text{Binomial}(n, \theta)$, with $n = 20$.
H_0: $\theta \leq \frac{1}{2}$,
H_1: $\theta > \frac{1}{2}$.
We choose $\alpha = .05$. Let

$$\phi(X) := \begin{cases} 1 & X > c \\ 0 & X \leq c \end{cases},$$

where we now need to choose the "critical value" c is such a way that

$$P_{\theta_0}(X > c) \leq \alpha \ \forall \ \theta_0 \leq \tfrac{1}{2}.$$

Consider the map

$$\vartheta \mapsto P_\vartheta(X > c) = \sum_{x=c+1}^{n} \binom{n}{x} \vartheta^x (1-\vartheta)^{n-x}.$$

It is increasing in ϑ so that

$$\max_{\theta_0 \leq \frac{1}{2}} P_{\theta_0}(X > c) = P_{\theta_0 = \frac{1}{2}}(X > c) = \sum_{x=c+1}^{n} \binom{n}{x} \frac{1}{2^n}.$$

It holds that

$$\underbrace{P_{\theta_0=\frac{1}{2}}(X > 15)}_{=0.0207} < \underbrace{\alpha}_{=0.05} < \underbrace{P_{\theta_0=\frac{1}{2}}(X > 14)}_{=0.0577}.$$

We choose the critical value c as small as possible: $c = 15$.

4.2 Definition of a randomized test

We have seen in the previous example that for discrete distributions, it is not always possible to make the error of first kind exactly equal to α. In a sense, a part of α is then left unused (comparable with a knapsack that is not completely filled, but all items that are not in the knapsack are too large to put in). By applying a randomized test, this problem is overcome (comparable to cutting a too large item so that it fits in the knapsack).

Definition 4.2.1 *A randomized statistical test at given level α ($0 < \alpha < 1$) is a (measurable) map $\phi: \ \mathcal{X} \to [0, 1]$ such that*

$$\phi(X) = \begin{cases} 1 & \text{means } H_0 \text{ is rejected} \\ q \in (0,1) & \text{means } H_0 \text{ is rejected with probability } q \\ 0 & \text{means } H_0 \ \text{ is not rejected} \end{cases},$$

and such that

$$E_{\theta_0}\phi(X) \leq \alpha \ \forall \ \theta_0 \in \Theta_0.$$

The power of the test at $\theta_1 \in \Theta_1$ is $E_{\theta_1}\phi(X)$.

In other words, when $\phi(X) \in (0,1)$ one throws a coin with probability $\phi(X)$ of success (heads) and rejects when it is heads. One may object that this is something one will never do in practice. Yet, it does make sense. Imagine a lab that carries out experiments every day, and based on these data, tests hypotheses every day. Then the same outcome on different days can sometimes mean rejection, sometimes not. In other words, one does not always stay on the conservative side as this would lead to a decrease of power. On the other hand, a judge who considers two cases with the same evidence, will not put one person in jail and the other person not. I.e. there may be ethical reasons not to randomize.

Example $X \sim \text{Binomial}(n, \theta)$, with $n = 20$.
H_0: $\theta \le \frac{1}{2}$,
H_1: $\theta > \frac{1}{2}$.
We choose $\alpha = .05$. We have

$$P_{\theta_0=\frac{1}{2}}(X > 15) < \alpha < P_{\theta_0=\frac{1}{2}}(X > 14)$$

so we can write

$$\alpha = P_{\theta_0=\frac{1}{2}}(X > 15) + \gamma P_{\theta_0=\frac{1}{2}}(X = 15)$$

where

$$q = \frac{\alpha - P_{\theta_0=\frac{1}{2}}(X > 15)}{P_{\theta_0=\frac{1}{2}}(X = 15)} = 0.79.$$

Thus a test at level α is

$$\phi(X) = \begin{cases} 1 & X > 15 \\ .79 & X = 15 \\ 0 & X < 15 \end{cases} .$$

Suppose we observe $X = 14$. Then H_0 cannot be rejected.

4.3 Simple hypothesis versus simple alternative

The simple hypothesis versus simple alternative problem is
H_0 : $\theta = \theta_0$,
H_1 : $\theta = \theta_1$.
The term "simple" refers to the fact that there is only one parameter under H_0 and only one parameter under H_1. In other words, the distributions under H_0 and H_1 are known. This is maybe an exceptional situation, but it is good to start with something simple.

There are only two distributions in the class $\mathcal{P}$, that is $\mathcal{P} = \{P_{\theta_0}, P_{\theta_1}\}$. Then there is always a dominating measure (for example $\nu = P_{\theta_0} + P_{\theta_1}$). Let $p_0(\cdot) := p_{\theta_0}(\cdot)$ be the density under H_0 and $p_1(\cdot) := p_{\theta_1}$ be the density under H_1. This could be the probability mass function in the discrete case, or the Lebesgue density in the absolutely continuous case.

Definition 4.3.1 *A Neyman-Pearson test is of the form*

$$\phi_{\rm NP}(X) := \begin{cases} 1 & \frac{p_1(X)}{p_0(X)} > c_0 \\ q & \frac{p_1(X)}{p_0(X)} = c_0 \\ 0 & \frac{p_1(X)}{p_0(X)} < c_0 \end{cases}$$

where $c_0 \ge 0$ and $q \in [0,1]$ are given constants.

Note that a Neyman-Pearson test "makes sense": if p_1 is much larger than p_0 it means that θ_1 is more likely than θ_0.

Lemma 4.3.1 *(Neyman-Pearson Lemma) Let $\alpha \in (0,1)$ be a given level. Choose c_0 and q in such a way that*

$$E_{\theta_0}\phi_{\rm NP}(X) = \alpha.$$

Then for all (randomized) tests $\tilde{\phi}$ with $E_{\theta_0}\tilde{\phi}(X) \le \alpha$ it holds that

$$E_{\theta_1}\tilde{\phi}(X) \le E_{\theta_1}\phi_{\rm NP}(X).$$

In other words, $\phi_{\rm NP}$ has maximal power among all tests with level α.

Proof for the discrete case. We have

$$E_{\theta_1}\biggl(\tilde{\phi}(X) - \phi_{\rm NP}(X)\biggr) = \sum_x \biggl(\tilde{\phi}(x) - \phi_{\rm NP}(x)\biggr) p_1(x)$$

$$= \sum_{p_1/p_0 > c_0} \underbrace{(\tilde{\phi} - \phi_{\rm NP})}_{\le 0} p_1 + \sum_{p_1/p_0 = c_0} (\tilde{\phi} - \phi_{\rm NP}) p_1 + \sum_{p_1/p_0 < c_0} \underbrace{(\tilde{\phi} - \phi_{\rm NP})}_{\ge 0} p_1$$

$$\le c_0 \sum_{p_1/p_0 > c_0} (\tilde{\phi} - \phi_{\rm NP}) p_0 + c_0 \sum_{p_1/p_0 = c_0} (\tilde{\phi} - \phi_{\rm NP}) p_0 + c_0 \sum_{p_1/p_0 < c_0} (\tilde{\phi} - \phi_{\rm NP}) p_0$$

$$= c_0 E_{\theta_0}\biggl(\tilde{\phi}(X) - \phi_{\rm NP}(X)\biggr) = c_0\biggl(E_{\theta_0}\tilde{\phi}(X) - \alpha\biggr) \le 0.$$

□

4.4 Examples

Example 4.4.1 *Consider $X \sim \text{Binomial}(n, \theta)$ and*
$H_0: \ \theta = \theta_0$,
$H_1: \ \theta = \theta_1$,
where $\theta_1 > \theta_0$. Then

$$\frac{p_1(x)}{p_0(x)} = \left[\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right]^x \left(\frac{1-\theta_1}{1-\theta_0}\right) > c_0$$

$$\Leftrightarrow$$

$$\underbrace{x \log\left[\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right]}_{>0 \text{ as } \theta_1>\theta_0} + n\log\left(\frac{1-\theta_1}{1-\theta_0}\right) > \log c_0$$

$$\Leftrightarrow$$

$$x > \frac{\log c_0 - n\log\left(\frac{1-\theta_1}{1-\theta_0}\right)}{\log\left[\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right]} := c.$$

A Neyman-Pearson test is thus

$$\phi_{\mathrm{NP}}(X) = \begin{cases} 1 & X > c \\ q & X = c \\ 0 & X < c \end{cases}.$$

If we choose the critical value c in such a way that

$$\underbrace{P_{\theta_0}(X > c)}_{=\sum_{x>c}\binom{n}{x}\theta_0^x(1-\theta_0)^{n-x}} \quad \le \alpha \le \quad \underbrace{P_{\theta_0}(X > c-1)}_{=\sum_{x>c-1}\binom{n}{x}\theta_0^x(1-\theta_0)^{n-x}}$$

and then

$$q = \frac{\alpha - P_{\theta_0}(X > c)}{P_{\theta_0}(X = c)},$$

then $E_{\theta_0}\phi_{\mathrm{NP}}(X)) = \alpha$ and ϕ_{NP} is most powerful among all tests with level α. Note that c and q do not depend on θ_1: the test only depends on the sign of $\theta_1 - \theta_0$.

Example 4.4.2 *In this example, we have a sample $\mathbb{X} = (X_1, \ldots, X_n)$ of i.i.d. $\mathcal{N}(\mu, \sigma_0^2)$ -distributed random variables where μ is unknown and σ_0^2 is known. Write the density of $(X_1, \ldots, X_n)$ as*

$$\mathbf{p}_\mu(x_1, \ldots, x_n) := \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left[-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma_0^2}\right].$$

Then

$$\begin{aligned} \frac{\mathbf{p}_{\mu_1}(x_1, \ldots, x_n)}{\mathbf{p}_{\mu_0}(x_1, \ldots, x_n)} &= \exp\left[-\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^n (x_i - \mu_1)^2 - \sum_{i=1}^n (x_i - \mu_0)^2\right)\right] \\ &= \exp\left[\frac{1}{2\sigma_0^2}\left(-2\sum_{i=1}^n (x_i - \mu_0) + n(\mu_1 - \mu_0)^2\right)\right] \\ &= \exp\left[\frac{1}{\sigma_0^2}\left(n\bar{x} - n\mu_0 - n(\mu_1 - \mu_0)^2/2\right)\right] \end{aligned}$$

It follows that

$$\frac{\mathbf{p}_{\mu_1}(\mathbb{X})}{\mathbf{p}_{\mu_0}(\mathbb{X})} > c_0 \Leftrightarrow \begin{cases} \bar{X} > c & \text{if } \mu_1 > \mu_0 \\ \bar{X} < c & \text{if } \mu_1 < \mu_0 \end{cases}.$$

To test $H_0: \ \mu = \mu_0$ we consider 3 alternative hypotheses.

$\boxed{\textit{Right sided}}$

$H_1: \ \mu = \mu_1 > \mu_0$. Then $\phi_{\mathrm{NP}}(\mathbb{X}) = \mathrm{l}_{\{\bar{X} > c\}}$ where the critical value c is such that $\mathbb{E}_{\mu_0}\phi_{\mathrm{NP}}(\mathbb{X}) = \alpha$. We have

$$\mathbb{E}_{\mu_0}\phi_{\mathrm{NP}}(\mathbb{X}) = \mathbb{P}_{\mu_0}(\bar{X} > c) = \mathbb{P}_{\mu_0}\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma_0} > \frac{\sqrt{n}(c - \mu_0)}{\sigma_0}\right) = \alpha$$

for

$$\frac{\sqrt{n}(c - \mu_0)}{\sigma_0} = \Phi^{-1}(1 - \alpha).$$

Thus

$$c = \mu_0 + \Phi^{-1}(1 - \alpha)\sigma_0/\sqrt{n}.$$

For example for $\alpha = .05$ it holds that $\Phi^{-1}(1 - \alpha) = 1.65$.

$\boxed{\textit{Left sided}}$

$H_1: \ \mu = \mu_1 < \mu_0$. Reject H_0 if

$$\bar{X} < \mu_0 - \Phi^{-1}(1 - \alpha)\sigma_0/\sqrt{n}.$$

$\boxed{\textit{Two sided}}$

$H_1: \ \mu \neq \mu_0$. The Neyman-Pearson Lemma cannot be used. It can be shown (see e.g. Fundamentals of Mathematical Statistics) that the following test is in some sense optimal (it has largest power among all tests of level α for which the power is larger than the probability of rejecting under H_0) : reject H_0 if

$$\bar{X} > \mu_0 + \Phi^{-1}(1 - \tfrac{\alpha}{2})\sigma_0/\sqrt{n} \text{ or } \bar{X} < \mu_0 - \Phi^{-1}(1 - \tfrac{\alpha}{2})\sigma_0/\sqrt{n}.$$

For example for $\alpha = .05$ it holds that $\Phi^{-1}(1 - \frac{\alpha}{2}) = 1.96$. If one agrees that $1.96 \approx 2$ we see the rule of thumb: reject H_0 if the difference between $\bar{X}$ and μ_0 is more than twice the standard deviation of $\bar{X}$, i.e. if $|\bar{X} - \mu_0| > 2\sigma_0/\sqrt{n}$.

Chapter 5

One-sample tests

In this chapter we consider a sample of real-valued observations from a CDF F with unknown "location parameter" (mean or median in this chapter), and aim at testing whether the location parameter is below (or above, or equal) to a given value.

One may think for example of having a group of n test persons, which have been given a drug to reduce blood pressure, and one observes the difference between blood pressure at the beginning of the test period and the end of the test period. Then one may ask: did the mean (or median) blood pressure decrease?

We will present two tests: Student's test and the sign test. The first is based on the assumption that the data are normally distributed, and the second only assumes the median exists and continuity near the median.

In the previous chapter, we showed that the Neyman-Pearson test is most powerful for the "simple null-hypothesis versus simple alternative" problem. In this and the next chapter, we are dealing with composite hypotheses. For such problems, the theory on optimal (most powerful) tests is more involved (see *Fundamentals of Mathematical Statistics*). Here, we only present the tests, but do not explain why they are a good idea. But they do "make sense".

5.1 The Student distribution

The Student distribution (or t-distribution) is symmetric around 0. We will encounter below the Student distribution with $n-1$ "degrees of freedom", the t_{n-1}-distribution. The Lebesgue density of the t_{n-1}-distribution is

$$f_{n-1}(t) = \frac{\Gamma(\frac{n}{2})}{\sqrt{(n-1)\pi}\Gamma(\frac{n-1}{2})}\left(1+\frac{t^2}{n-1}\right)^{-n/2}, \ t \in \mathbb{R}.$$

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Let $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$ be the sample average. Then $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. If we subtract the mean and divide by the standard deviation, this is called standardization. So the standardized form of the sample average is

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}.$$

It has the standard normal distribution. Now if σ^2 is unknown one may want to replace σ in the standardization by the sample standard deviation $S := \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}$. This is called studentization.

Theorem 5.1.1 *Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$-distributed. The studentized sample average*

$$\frac{\sqrt{n}(\bar{X}-\mu)}{S}$$

has a Student distribution with $n-1$ degrees of freedom (t_{n-1}-distribution).

Proof. We first show that, for all i, $X_i - \bar{X}$ and $\bar{X}$ are independent. This follows from

$$\begin{aligned} \mathrm{Cov}(X_i - \bar{X}, \bar{X}) &= \mathrm{Cov}(X_i, \bar{X}) - \underbrace{\mathrm{Cov}(\bar{X}, \bar{X})}_{=\mathrm{Var}(\bar{X})} \\ &= \frac{1}{n}\sum_{j=1}^n \mathrm{Cov}(X_i, X_j) - \frac{\sigma^2}{n} = 0. \end{aligned}$$

The independence now follows from the fact that for multivariate normal random variables, zero covariance implies independence.
Thus S^2 and $\bar{X}$ are also independent. Moreover

$$\sum_{i=1}^n \frac{(X_i-\mu)^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i-\bar{X})^2}{\sigma^2} + \frac{n(\bar{X}-\mu)^2}{\sigma^2}.$$

By the definition of the χ^2-distribution (see Appendix A), the left hand side has a χ^2_n-distribution. Moreover $\frac{n(\bar{X}-\mu)^2}{\sigma^2}$ has a χ^2_1-distribution. Since moreover $\sum_{i=1}^n \frac{(X_i-\bar{X})^2}{\sigma^2}$ is independent of $\frac{n(\bar{X}-\mu)^2}{\sigma^2}$ it must have a χ^2_{n-1}-distribution. The result now follows from the definition of the Student distribution (see Appendix A). □

5.2 The Student test

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ We consider the same testing problem as in Example 4.4.2, but now for the case both μ and σ^2 unknown.

Let $c(n-1, \alpha)$ be the $(1-\alpha)$-quantile of the t_{n-1}-distribution. One can show that $\forall\ \alpha \in (0,1)$

$$c(n-1,\alpha)\begin{cases} > \ \Phi^{-1}(1-\alpha) & \forall\ n \\ \to \Phi^{-1}(1-\alpha) & n \to \infty \end{cases}.$$

The latter in fact follows from the consistency of S^2 as estimator of σ^2, i.e., $S^2 \to \sigma^2$ (in probability) as $n \to \infty$.

We know from Theorem 5.1.1 that

$$\mathbb{P}_\mu\biggl(\frac{\sqrt{n}(\bar{X}-\mu)}{S} > c(n-1,\alpha) \biggr) = \alpha,$$

$$\mathbb{P}_\mu\biggl(\frac{\sqrt{n}(\bar{X}-\mu)}{S} < -c(n-1,\alpha) \biggr) = \alpha,$$

and

$$\mathbb{P}_\mu\biggl(\frac{\sqrt{n}|\bar{X}-\mu|}{S} > c(n-1,\tfrac{\alpha}{2}) \biggr) = \alpha.$$

The first will be applied in the right sided test, the second in the left sided test, and the third in the two sided test.

$\boxed{\text{Right sided}}$
$H_0: \ \mu \le \mu_0$,
$H_1: \ \mu > \mu_0$.
Reject H_0 if

$$\bar{X} > \mu_0 + c(n-1,\alpha)S/\sqrt{n}.$$

Then

$$\max_{\mu \le \mu_0} \mathbb{P}_\mu(H_0 \text{ rejected}) = \mathbb{P}_{\mu_0}(H_0 \text{ rejected}) = \alpha.$$

$\boxed{\text{Left sided}}$
$H_0: \ \mu \ge \mu_0$,
$H_1: \ \mu < \mu_0$.
Reject H_0 if

$$\bar{X} < \mu_0 - c(n-1,\alpha)S/\sqrt{n}.$$

$\boxed{\text{Two sided}}$
$H_0: \ \mu = \mu_0$,
$H_1: \ \mu \ne \mu_0$.
Reject H_0 if

$$\bar{X} > \mu_0 + c(n-1,\tfrac{\alpha}{2})S/\sqrt{n} \text{ or } \bar{X} < \mu_0 - c(n-1,\tfrac{\alpha}{2})S/\sqrt{n}.$$

Numerical example:

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
4.5	0	0
4	-.5	.25
3.5	-1	1
6	1.5	2.25
5	.5	.25
4	-.5	.25

We have $n = 6$, $\bar{x} = 4.5$, $\sum(x_i - \bar{x})^2 = 4$, $s^2 = .8$ and $s/\sqrt{n} = .365$. With $\alpha = .05$ the $(1 - \frac{\alpha}{2})$-quantile of the t_5-distribution is $c(5, 0.025) = 2.571$. Thus $c(5, 0.025)s/\sqrt{n} = .939$.
For example
$H_0 : \ \mu = 5.1$
is rejected when $|\bar{x} - 5.1| > .939$. Thus $H_0 : \ \mu = 5.1$ is not rejected as

$$|\bar{x} - 5.1| = .6 < .939.$$

The values for μ which are not rejected are all μ such that $|\bar{x} - \mu| \le .939$, that is all $\mu \in [3.561, 5.439]$. We call $[3.561, 5.439]$ a 95% confidence interval for μ (see Chapter 8).

5.3 Sign test

Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables with common CDF F. We assume $m := F^{-1}(\frac{1}{2})$ exists, and that F is continuous near m. Consider the testing problem
$H_0 : \ m = m_0$,
$H_1 : \ m \neq m_0$.
As test statistic we take

$$T := \#\{X_i > m_0\}$$

and as (non-randomized) test

$$\phi(T) := \begin{cases} 1 & |T - \frac{n}{2}| > c \\ 0 & |T - \frac{n}{2}| \le c \end{cases}$$

where c is such that

$$\underbrace{\mathbb{P}_{H_0}\left(\left|T - \frac{n}{2}\right| > c\right)}_{=\sum_{|k-\frac{n}{2}|>c}\binom{n}{k}2^{-n}=:1-G_Z(c)} \quad \le \alpha$$

and c is as small as possible. One calls $1 - G_Z(Z-)$ where $Z := |T - \frac{n}{2}|$ the p-value: see the next section for its definition. We reject H_0 if the p-value is at most α. We can write for $\tilde{c} < n/2$,

$$\phi(T) := \begin{cases} 1 & T \le \tilde{c} \text{ or } T \ge n - \tilde{c} \\ 0 & \text{else} \end{cases},$$

where

$$\underbrace{\mathbb{P}_{H_0}(T \le \tilde{c}) + \mathbb{P}_{H_0}(T \ge n - \tilde{c})}_{=2\sum_{k \le \tilde{c}}\binom{n}{k}2^{-n}} \le \alpha.$$

Numerical example continued
The normal distribution is symmetric around μ so the median m is equal to μ.

We test
$H_0: \ \mu = 5.1 \quad ,$
$H_1: \ \mu \neq 5.1 \quad .$
We have

$$\begin{aligned} 1 - G_Z(2) &= \mathbb{P}_{H_0}(T \le 0 \text{ or } T \ge 6) = \mathbb{P}_{H_0}(T = 0) + \mathbb{P}_{H_0}(T = 6) \\ &= \frac{2}{64} = .03125 < .05 \end{aligned}$$

so we can take $\tilde{c} = 0$.[1] The observed value of T is $T = 1$. Therefore we cannot reject H_0. Since $n = 6$ we have $Z := |T - \frac{n}{2}| = 2$. In the next section one can find the general definition of a p-value. In this example, the p-value is

$$1 - G_Z(2-) = 1 - G_Z(1) = \frac{14}{64} = .21875 > .05.$$

5.4 Definition of p-value

Definition 5.4.1 *Let Z be a test statistic such that large values of Z are evidence against $H_0 : \theta = \theta_0$. We reject H_0 when $Z \ge c$ where the critical value c is chosen such that the probability of rejection when the null hypothesis is true is at most α:*

$$1 - G_Z(c-) \le \alpha$$

with $1 - G_Z(c-) := \mathbb{P}_{H_0}(Z \ge c)$. The p-value is then $1 - G_Z(Z-)$.

Note $1 - G_Z$ is a decreasing function, so

$$Z \ge c \Rightarrow 1 - G_Z(Z-) \le 1 - G_Z(c-) \le \alpha.$$

Thus if the p-value is at most α we reject H_0.

Note In the two-sided case, one typically starts with a test statistic T such that large values of $Z := |T|$ are evidence against $H_0 : \theta = \theta_0$. If we reject H_0 for $|T| \ge c$ and T has CDF G_T under the null hypothesis, then since $\mathbb{P}_{H_0}(|T| \ge c) = 1 - G_T(c-) + G_T(-c)$ the p-value is $1 - G_T(|T|-) + G_T(-|T|)$. If in addition G_T is continuous and symmetric the p-value becomes $2(1 - G_T(|T|))$. Thus then we reject H_0 if $(1 - G_T(|T|) \le \alpha/2$.

[1] A randomized test at level $\alpha = .05$ is

$$\tilde{\phi}(T) = \begin{cases} 1 & T = 0 \text{ or } T = 6 \\ \frac{1}{10} & T = 1 \text{ or } T = 5 \\ 0 & \text{else} \end{cases} .$$

Indeed

$$\mathbb{E}_{H_0}\tilde{\phi}(T) = \mathbb{P}_{H_0}(T = 0 \text{ or } T = 6) + \frac{1}{10}\mathbb{P}_{H_0}(T = 1 \text{ or } T = 5) = .05.$$

Chapter 6

Two-sample tests

Suppose we carry out an experiment with a treatment group and a control group. The data then consists of two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$. We assume that $X_1, \ldots, X_n$ are i.i.d. real-valued random variable with CDF F and $Y_1, \ldots, Y_m$ are i.i.d. real-valued random variables with CDF G. We moreover assume that the two samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ are independent. Our goal is to test whether F and G are equal. As in Chapter 5, one can build a test under the assumption that the data follow a normal distribution. This leads to Student's test. If one is not ready to assume normality, one can try to build a test assuming no extra conditions, except maybe continuity. This leads to Wilcoxon's test.

When one constructs a test, the general idea is to try to find a real-valued test statistic such that its extreme values, say large values, are evidence against the null-hypothesis. Then for such extreme values the null is rejected. But what is extreme? For that one needs to know what the distribution of the statistic is under the null hypothesis: the null-distribution. Then, given a level $\alpha \in (0, 1)$, one takes the critical value c such that when the null hypothesis is true, the probability that the statistic is larger than c is at most α. One rejects H_0 if the statistic is larger than the critical value c.

6.1 The two-sample student test

Model:

$$\underbrace{X_1, \ldots, X_n}_{\sim\mathcal{N}(\mu_1,\sigma^2)}, \underbrace{Y_1, \ldots, Y_m}_{\sim\mathcal{N}(\mu_2,\sigma^2)} \text{ independent}$$

We want to test
$H_0: \ \mu_1 = \mu_2$
$H_1: \ \mu_1 \neq \mu_2$.

Note that we assume that the observations in both samples have the same variance σ^2. If the variance of the observations in one sample may be different from those in the other sample, and these two variances are unknown, the problem is known as the Behrens-Fisher problem, and there is no test statistic with a simple null-distribution and good power. Assuming equal variance is mathematically convenient but is perhaps not realistic.

If $\mu_1 = \mu_2$ then for n large $\bar{X} \approx \bar{Y}$. Therefore it makes sense to reject H_0 if $|\bar{X} - \bar{Y}| > c$ where the critical value c is to be chosen in such a way that

$$\mathbb{P}_{H_0}\left(|\bar{X} - \bar{Y}| > c\right) = \alpha$$

where $0 < \alpha < 1$ is a given level. So we need to find the distribution of $\bar{X} - \bar{Y}$ under H_0. It holds that

$$\bar{X} \sim \mathcal{N}\biggl(\mu_1, \frac{\sigma^2}{n}\biggr), \ \bar{Y} \sim \mathcal{N}\biggl(\mu_2, \frac{\sigma^2}{m}\biggr).$$

Moreover

$$\mathbb{E}(\bar{X} - \bar{Y}) = \mu_1 - \mu_2,$$

and since $\bar{X}$ and $\bar{Y}$ are independent

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2\biggl(\frac{n+m}{nm}\biggr).$$

Thus

$$\bar{X} - \bar{Y} \sim \mathcal{N}\biggl(\mu_1 - \mu_2, \sigma^2\biggl(\frac{n+m}{nm}\biggr)\biggr).$$

Standardizing gives

$$\sqrt{\frac{nm}{n+m}}\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma} \sim \mathcal{N}(0,1).$$

We consider two cases.

$\boxed{\sigma^2 = \sigma_0^2 \text{ known:}}$ Then we can take as test statistic

$$T_0 := \sqrt{\frac{nm}{n+m}}\frac{\bar{X} - \bar{Y}}{\sigma_0}.$$

Under H_0 the statistic T_0 has a standard normal distribution. We reject H_0 when $|T_0| > \Phi^{-1}(1 - \frac{\alpha}{2})$. Then

$$\mathbb{P}_{H_0}(H_0 \text{ rejected}) = \mathbb{P}_{H_0}\biggl(|T_0| > \Phi^{-1}(1 - \tfrac{\alpha}{2})\biggr) = \alpha.$$

In other words the critical value is $c = \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{\frac{n+m}{nm}}\sigma_0$. (With the "common" choice $\alpha = .05$ it holds that $c = (1.96)\sqrt{\frac{n+m}{nm}}\sigma_0$, i.e., roughly twice the standard deviation of $\bar{X} - \bar{Y}$).

$\boxed{\sigma^2 \text{ unknown:}}$ To estimate the standard deviation of $\bar{X} - \bar{Y}$ we need an estimator of σ^2. A good choice turns out to be the "pooled sample" variance

$$\tilde{S}^2 := \frac{1}{n+m-2}\left\{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2\right\},$$

which is unbiased. Standardizing with the estimated standard deviation gives the statistic

$$T := \sqrt{\frac{nm}{n+m}}\frac{\bar{X} - \bar{Y}}{\tilde{S}}.$$

But because $\tilde{S}$ is random T is no longer normally distributed. This is not really a problem, as long as its distribution under H_0 does not depend on unknown parameters. It is now not difficult to show that under H_0, T has a Student distribution with $n+m-2$ degrees of freedom, the t_{n+m-2}-distribution[1]. Therefore, with $c(n+m-2, \frac{\alpha}{2})$ the $(1-\frac{\alpha}{2})$-quantile of the t_{n+m-2}-distribution, we reject H_0 if $|T| > c(n+m-2, \frac{\alpha}{2})$ or equivalently if $|\bar{X} - \bar{Y}| > \tilde{c}$ where the critical value $\tilde{c}$ is $\tilde{c} = c(n+m-2, \frac{\alpha}{2})\sqrt{\frac{n+m}{nm}}\tilde{S}$.

6.2 Two-sample Wilcoxon test, or Mann-Whitney U test

Model:

$$\underbrace{X_1, \ldots, X_n}_{\sim F}, \; \underbrace{Y_1, \ldots, Y_m}_{\sim G} \text{ independent}$$

where F and G are two unknown continuous distributions.

We want to test
$H_0 : \; F = G$,
$H_1 : \; F \neq G$.

We construct a test statistic as follows. Let $N := n + m$ be the pooled sample size and $(Z_1, \ldots, Z_N) := (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ be the pooled sample. In the pooled sample, let $Z_{(1)} < \cdots < Z_{(N)}$ be the order statistics. Let $R_i := \text{rank}(X_i)$ in the pooled sample (i.e. $Z_{(R_i)} = X_i$), $i = 1, \ldots, n$, and $R_{n+j} := \text{rank}(Y_j)$ in the pooled sample, $j = 1, \ldots, m$. If $F = G$ then $(R_1, \ldots, R_n, R_{n+1}, \ldots, R_N)$ is a random permutation of the numbers $\{1, \ldots, N\}$. This means that under H_0 the ranks $R_1, \ldots, R_n$ have the same distribution as a random sample without replacement of size n from an urn with N balls numbered from 1 to N. The Mann-Whitney U statistic is

$$U := \sum_{i=1}^{n} R_i.$$

[1] As in the one sample case, $\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2$ has a χ^2_{n-1}-distribution. Similarly, $\sum_{i=1}^{n}(Y_j - \bar{Y})^2/\sigma^2$ has a χ^2_{m-1}-distribution. The two sums-of-squares are independent and independent of $\bar{X}$ and $\bar{Y}$.

The Wilcoxon test statistic is

$$W := \#\{X_i > Y_j\}.$$

One may verify that U and W are equivalent:

$$U = W + \frac{n(n+1)}{2}.$$

numerical example

z	rank
$x_1 = 36$	8
$x_2 = 9$	4
$x_3 = 7$	2
$x_4 = 100$	9
$x_5 = 3$	1
$y_1 = 5$	3
$y_2 = 37$	7
$y_3 = 11$	5
$y_4 = 12$	6

Table 6.1: $n = 5$, $m = 4$, $E_{H_0}(U) = 25$, $u = 24$, $w = 9$

Lemma 6.2.1
i) $\mathbb{E}_{H_0}(U) = \frac{n(N+1)}{2}$
ii) $\mathrm{Var}_{H_0}(U) = \frac{nm(N+1)}{12}$.

Proof.
i) For all i

$$\mathbb{P}_{H_0}(R_i = k) = \frac{1}{N},\ k = 1, \ldots N.$$

Hence

$$\mathbb{E}_{H_0} R_i = \sum_{k=1}^{N} k \frac{1}{N} = \frac{N+1}{2}$$

and so

$$\mathbb{E}_{H_0}(U) = \frac{n(N+1)}{2}.$$

ii) For all i

$$\mathbb{E}_{H_0} R_i^2 = \sum_{k=1}^{N} k^2 \frac{1}{N} = \frac{(N+1)(2N+1)}{6}$$

so

$$\mathrm{Var}_{H_0}(R_i) = \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} = \frac{N^2-1}{12} =: \sigma^2.$$

Further for $i \neq j$

$$\mathbb{E}_{H_0} R_i R_j = \sum_{k \neq l} kl \frac{1}{N(N-1)}$$

$$= \frac{N(N+1)^2}{4(N-1)} - \frac{(N+1)(2N+1)}{6(N-1)} = \frac{(N+1)(3N^2-N-2)}{12(N-1)}.$$

Thus

$$\mathrm{Cov}_{H_0}(R_i, R_j) = \frac{(N+1)(3N^2-N-2)}{12(N-1)} - \frac{(N+1)^2}{4} = -\frac{\sigma^2}{N-1}.$$

It follows that

$$\mathrm{Var}_{H_0}\left(\sum_{i=1}^n R_i\right) = n\sigma^2 - n(n-1)\frac{\sigma^2}{N-1} = n\sigma^2 \frac{N-n}{N-1}.$$

□

Corollary 6.2.1 $\mathbb{E}_{H_0}(W) = \frac{nm}{2}$, $\mathrm{Var}_{H_0}(W) = \frac{nm(N+1)}{12}$.

Standardizing under H_0:

$$T := \frac{U - \mathbb{E}_{H_0}(U)}{\sqrt{\mathrm{Var}_{H_0}(U)}} = \frac{W - \mathbb{E}_{H_0}(W)}{\sqrt{\mathrm{Var}_{H_0}(W)}}.$$

For n and m large, T has under H_0 approximately a $\mathcal{N}(0,1)$-distribution. (No proof: this does not follow from the "usual" CLT.)

Numerical example continued

$$|T| = \frac{|24-25|}{\sqrt{\frac{20\times 8}{12}}} = \sqrt{\frac{3}{7}} = .655.$$

The approximate p-value (see Section 5.4 for its definition) is $2(1-\Phi(.655)) = .513$.

Chapter 7

Goodness-of-fit tests

In this chapter, we study the construction of tests when the hypothesis is that the data follow some given distribution (simple hypothesis), or a distribution in some given parametric family (composite hypothesis). For example, one may want to test whether the digits of π are uniformly distributed, or whether the square-root function follows Benford's law. For such questions one may use the χ^2-test. Perhaps one wants to test whether waiting times in queueing theory follow an exponential distribution. Then the distribution of the data is continuous one may invoke binning and again use a χ^2-test, or refrain from binning and use a Kolmogorov-Smirnov test.

7.1 Kolmogorov-Smirnov tests

Model: $X_1, \ldots, X_n$ i.i.d. with CDF F on $\mathbb{R}$.

$H_0: \ F = F_0$.

Recall the empirical distribution function

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^n \mathrm{l}_{\{X_i \le x\}}, \ x \in \mathbb{R}.$$

Kolmogov-Smirnov tests are based on a comparison of $\hat{F}_n$ with F_0. The test statistic is

$$T_\infty := \sup_x |\hat{F}_n(x) - F_0(x)|,$$

or its variants

$$T_p := \int |\hat{F}_n(x) - F_0(x)|^p dF_0(x), \ 1 \le p < \infty.$$

An approximation of the distribution of T_p ($1 \le p \le \infty$) under the null hypothesis follows from probability theory (not treated here). One may also simulate the null-distribution.

7.2 The χ^2-test: simple hypothesis

Let $X \in \{1, \ldots, q\}$ represent a class label. Write

$$P_\theta(X = j) := \theta_j,$$

where

$$\theta \in \Theta := \{\vartheta = (\vartheta_1, \ldots, \vartheta_q) : \ \vartheta_j \geq 0 \ \forall \ j, \ \sum_{j=1}^q \vartheta_j = 1\}.$$

Suppose we want to test
$H_0 : \ \theta = \theta_0$.
The data consist of i.i.d. copies $X_1, \ldots, X_n$ of X. The maximum likelihood estimator of θ is

$$\hat{\theta}_j = \frac{N_j}{n}, \ N_j := \#\{X_i = j\}, \ j = 1, \ldots, q$$

(see Example 3.2.4). The idea is now to reject H_0 if $\hat{\theta}$ is very different from the hypothesized θ_0. One may use for instance the Euclidean distance between $\hat{\theta}$ and θ_0 as a test statistic. One may however want to take into account the different variances of the estimators of the components. A test statistic that does so is the so-called χ^2 test statistic

$$\chi^2 := n \sum_{j=1}^q \frac{(\hat{\theta}_j - \theta_{0,j})^2}{\theta_{0,j}} = \sum_{j=1}^q \frac{(N_j - n\theta_{0,j})^2}{n\theta_{0,j}}.$$

Theorem 7.2.1 *For n large, $\mathbb{P}_{H_0}(\chi^2 \leq t) \approx G(t)$ for all t, where G is the CDF of a $\chi^2(q-1)$-distribution.*

No proof. (See *Fundamentals of Mathematical Statistics.*)

Special case: $q = 2$. Then $X := N_1 \sim$ Binomial(n, p) where $p := \theta_1$, and $N_2 = n - X$, $\theta_2 = 1 - p$. So

$$\chi^2 = \frac{(X - np)^2}{np} + \frac{(n - X - n(1-p))^2}{n(1-p)} = \frac{(X - np)^2}{np(1-p)}.$$

By the CLT

$$\frac{X - np}{\sqrt{np(1-p)}}$$

is approximately $\mathcal{N}(0, 1)$-distributed, and so its square

$$\frac{(X - np)^2}{np(1-p)}$$

is approximately $\chi^2(1)$-distributed (by the definition of the χ^2-distribution).

7.3 The χ^2-test: composite hypothesis

The random variable $X \in \{1, \ldots, q\}$ again represent a class label and

$$P_\theta(X = j) := \theta_j, \ j = 1, \ldots, q.$$

Suppose we want to test $m < q - 1$ restrictions
$H_0 : \ R_k(\theta) = 0, \ k = 1, \ldots, m$. Let

$$\hat{\theta}_0 := \arg \max_{\vartheta \in \Theta: \ R_k(\vartheta)=0, \ k=1,\ldots,m} \sum_{j=1}^{q} N_j \log \vartheta_j$$

be the maximum likelihood estimator under the m restrictions. Define the test statistic

$$\chi^2 := \sum_{j=1}^{q} \frac{(N_j - n\hat{\theta}_{0,j})^2}{n\hat{\theta}_{0,j}}.$$

Under some regularity conditions (see *Fundamentals of Mathematical Statistics*), the distribution of χ^2 under H_0 is approximately $\chi^2(m)$. Thus we reject H_0 when $\chi^2 > G^{-1}(1-\alpha)$ where G is the CDF of the $\chi^2(m)$-distribution. Then

$$\mathbb{P}_{H_0}(H_0 \text{ rejected}) \approx \alpha.$$

Note A special case is the simple hypothesis $H_0 : \ \theta = \theta_0$. This corresponds to $m = q - 1$ restrictions.

7.4 Contingency tables

This section treats a special case of the previous section.

Let $X := (Y, Z) \in \{(k, l) : \ k = 1, \ldots, p, \ l = 1, \ldots, q\}$ and

$$P_\theta\biggl(X = (k, l)\biggr) := \theta_{k,l}$$

where

$$\begin{aligned} \theta \ &\in \ \Theta \\ &= \ \biggl\{\vartheta = \{\vartheta_{k,l} : \ k = 1, \ldots, p, \ l = 1, \ldots, q\}, \ \vartheta_{k,l} \geq 0 \ \forall \ k, l \ \sum_{k=1}^{p} \sum_{l=1}^{q} \vartheta_{k,l} = 1\biggr\}. \end{aligned}$$

We aim at testing whether Y and Z are independent. Define the marginals

$$\eta_k := \sum_{l=1}^{q} \theta_{k,l} \ (k = 1, \ldots, p), \quad \xi_l := \sum_{k=1}^{p} \theta_{k,l} \ (l = 1, \ldots, q).$$

The null hypothesis is $H_0 : \ \theta_{k,l} = \eta_k \xi_l, \ \forall \ k, l$.

The data are i.i.d. copies $\{X_i = (Y_i, Z_i) : \; i = 1, \ldots, n\}$ of $X = (Y, Z)$. The maximum likelihood estimator is as before (see Example 3.2.4)

$$\hat{\theta}_{k,l} = \frac{N_{k,l}}{n}, \; k = 1, \ldots, p, \; l = 1, \ldots, q,$$

where $N_{k,l} = \#\{(Y_i, Z_i) = (k, l)\}$, $k = 1, \ldots, p$, $l = 1, \ldots, q$.

Write

$$N_{k,+} := \sum_{l=1}^{q} N_{k,l} \; (k = 1, \ldots, p), \quad N_{+,l} := \sum_{k=1}^{p} N_{k,l} \; (l = 1, \ldots, q).$$

Lemma 7.4.1 *The maximum likelihood under the restrictions of H_0 is*

$$\hat{\eta}_k = \frac{N_{k,+}}{n} \; (k = 1, \ldots, p), \; \hat{\xi}_l = \frac{N_{+,l}}{n} \; (l = 1, \ldots, q).$$

Proof. The log-likelihood is

$$\sum_{k=1}^{p} \sum_{l=1}^{q} N_{k,l} \log \vartheta_{k,l}.$$

We now have the restriction $\vartheta_{k,l} = \tilde{\eta}_k \tilde{\xi}_l$ for some non-negative $\tilde{\eta}_k$, $\tilde{\xi}_l$, with $\sum_{k=1}^{p} \tilde{\eta}_k = 1$ and $\sum_{l=1}^{q} \tilde{\xi}_l = 1$. The restricted log-likelihood is therefore

$$\begin{aligned} & \sum_{k=1}^{p} \sum_{l=1}^{q} N_{k,l} \log(\tilde{\eta}_k \tilde{\xi}_l) \\ = \; & \sum_{k=1}^{p} \sum_{l=1}^{q} N_{k,l} \log \tilde{\eta}_k + \sum_{k=1}^{p} \sum_{l=1}^{q} N_{k,l} \log \tilde{\xi}_l \\ = \; & \sum_{k=1}^{p} N_{k,+} \log \tilde{\eta}_k + \sum_{l=1}^{q} N_{+,l} \log \tilde{\xi}_l. \end{aligned}$$

The two terms can now be maximized separately, as done in Example 3.2.4 (where we used a Lagrange multiplier). □

It follows that

$$\chi^2 = \sum_{k=1}^{p} \sum_{l=1}^{q} \frac{(N_{k,l} - N_{k,+} N_{+,l}/n)^2}{N_{k,+} N_{+,l}/n}.$$

The original number of free parameters is

$$pq - 1.$$

The number of free parameters under H_0 is

$$p - 1 + q - 1.$$

The number of restrictions is therefore

$$m = \Big(pq - 1 \Big) - \Big(p - 1 + q - 1 \Big) = (p-1)(q-1).$$

So χ^2 is approximately $\chi^2((p-1)(q-1))$-distributed under H_0.

7.5 Special case: (2×2)-table

$N_{1,1}$	$N_{1,2}$	$N_{1,+}$
$N_{2,1}$	$N_{2,2}$	N_{2+}
$N_{+,1}$	$N_{+,2}$	n

or, using different symbols,

A	B	R
C	D	S
P	Q	n

Then

$$\chi^2 = \frac{n(AD - BC)^2}{PQRS}.$$

It has approximately a $\chi^2(1)$-distribution under H_0.

Numerical example

	left-handed	right-handed	All
Arts	16	40	56
Science	25	35	60
All	41	75	116

Table 7.1: Rows: Beverage, Columns: Personality

In the above example

$$\chi^2 = 116 \times \frac{(16 \times 35 - 40 \times 25)^2}{41 \times 75 \times 60 \times 56} = 2.174.$$

Remark Let $X \sim \text{Binomial}(n_1, p_1)$ and $Y \sim \text{Binomial}(n_2, p_2)$ be independent and suppose we want to test
$H_0: \ p_1 = p_2 =: p$ where $0 < p < 1$ is an unknown common value.
An estimator of p_1 is $\hat{p}_1 = X/n_1$ and an estimator of p_2 is $\hat{p}_2 = Y/n_2$. Thus it makes sense to reject H_0 if $|\hat{p}_1 - \hat{p}_2|^2$ is large.

X	Y	$X + Y$
$n_1 - X$	$n_2 - Y$	$n - (X + Y)$
n_1	n_2	$n := n_1 + n_2$

We have

$$\text{Var}_{H_0}(\hat{p}_1 - \hat{p}_2) = p(1-p)\frac{n}{n_1 n_2},$$

and we can estimate this by

$$\widehat{\text{Var}}_{H_0}(\hat{p}_1 - \hat{p}_2) := \hat{p}(1-\hat{p})\frac{n}{n_1 n_2},$$

where $\hat{p} = (X + Y)/n$. The H_0-standardized test statistic is now

$$T := \frac{|\hat{p}_1 - \hat{p}_2|^2}{\hat{p}(1-\hat{p})\frac{n}{n_1 n_2}} = \frac{n(AD - BC)^2}{PQRS} = \chi^2$$

as before.

Chapter 8

Confidence intervals

Let $\hat{\gamma} \in \mathbb{R}$ be an estimator of $\gamma \in \mathbb{R}$. As a rule of thumb, $\hat{\gamma}\pm$ twice the (estimated) standard deviation of $\hat{\gamma}$ is approximately a 95% confidence interval for γ. This is true if $\hat{\gamma} - \gamma$ is approximately normally distributed with mean zero, and if you are okay with the approximation $\Phi^{-1}(1 - \alpha/2) = 1.96 \approx 2$ for $\alpha = 0.05$. We remark here that many estimators are indeed approximately normally distributed. This is for instance the case for method of moment estimators provided certain differentiability conditions hold. It is also true for maximum likelihood estimators of a finite-dimensional parameter, assuming (rather involved) regularity conditions. We refer to *Fundamentals of Mathematical Statistics*.

If the sample size is not very large, one replaces $\Phi^{-1}(1 - \alpha/2)$ by a larger value so that the confidence interval becomes wider (and one thus is more conservative). The most popular choice is replacing the $(1 - \alpha/2)$-quantile $\Phi^{-1}(1 - \alpha/2)$ of the standard normal distribution by the $(1 - \alpha/2)$-quantile of the t_{n-1} distribution. This leads to an exact 95 % confidence interval for the mean μ using the estimator $\hat{\mu} := \bar{X}$ if the data are i.i.d. normally distributed. If the data are not normally distributed, one can often also find exact confidence intervals (instead of approximate ones based on some CLT), but this requires some inventivity.

Numerical example:

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
4.5	0	0
4	-.5	.25
3.5	-1	1
6	1.5	2.25
5	.5	.25
4	-.5	.25

We have $n = 6$, $\bar{x} = 4.5$, $s^2 = .8$ and $s/\sqrt{n} = .365$. With $\alpha = .05$ the $(1 - \frac{\alpha}{2})$-quantile of the t_5-distribution is $c(5, 0.025) = 2.571$ which is substantially larger than $\Phi^{-1}(1-\alpha/2) = 1.96$. Thus $c(5, 0.025)s/\sqrt{n} = .939$. Assuming i.i.d. Gaussian data the interval

$$\bar{x} \pm c(5, 0.025)s/\sqrt{n} = 4.5 \pm .939 = [3.561, 5.439]$$

is an exact 95% confidence interval for μ. If the Gaussian assumption does not hold, it is an approximate 95 % confidence interval provided the common variance σ^2 of the observations is finite.

8.1 Definition of a confidence interval

Consider an $X \in \mathcal{X}$ with distribution P_θ depending on $\theta \in \Theta$. Let $\gamma := g(\theta) \in \mathbb{R}$ be a parameter of interest. Write $\Gamma := \{g(\theta) : \ \theta \in \Theta\} \subset \mathbb{R}$. Let $\mathbb{X} := (X_1, \ldots, X_n)$ be a sample from P_θ

Recall that a real-valued statistic is a measurable map $\mathbb{X} \to R$.

Definition 8.1.1 *Let $\underline{T} = \underline{T}(\mathbb{X}) \in \mathbb{R}$ and $\bar{T} = \bar{T}(\mathbb{X}) \in \mathbb{R}$ be two statistics with $\underline{T} \leq \bar{T}$. One calls $[\underline{T}, \bar{T}]$ a $\underline{(1-\alpha)\text{-confidence interval}}$ for $g(\theta)$ if*

$$\mathbb{P}_\theta\biggl(\underline{T} \leq g(\theta) \leq \bar{T}\biggr) \geq 1-\alpha, \ \forall \ \theta \in \Theta.$$

8.2 Exact confidence intervals when the data are Gaussian

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$.

$\boxed{\text{Confidence interval for } \mu,\ \sigma^2 =: \sigma_0^2 \text{ known}}$

Then

$$\biggl[\bar{X} - \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n}, \bar{X} + \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n}\biggr]$$

is a $(1-\alpha)$-confidence interval for μ:

$$\begin{aligned} & \mathbb{P}_\mu\biggl(\bar{X} - \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n} \leq \mu \leq \bar{X} + \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n}\biggr) \\ = \ & \mathbb{P}_\mu\biggl(\mu - \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n} \leq \bar{X} \leq \mu + \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n}\biggr) \\ = \ & \mathbb{P}_\mu\biggl(\frac{\sqrt{n}|\bar{X}-\mu|}{\sigma_0} \leq \Phi^{-1}(1-\tfrac{\alpha}{2})\biggr) = 1-\alpha. \end{aligned}$$

$\boxed{\text{Confidence interval for } \mu,\ \sigma^2 \text{ unknown}}$

Then

$$\biggl[\bar{X} - c(n-1, \tfrac{\alpha}{2})S/\sqrt{n}, \bar{X} + c(n-1, \tfrac{\alpha}{2})S/\sqrt{n}\biggr],$$

is a $(1-\alpha)$-confidence interval for μ. Here

$$S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$$

is the sample variance and $c(n-1, \frac{\alpha}{2})$ the $(1-\frac{\alpha}{2})$-quantile of the Student distribution with $n-1$ degrees of freedom.

Confidence interval for σ^2, $\mu = \mu_0$ known

Then

$$\left[\frac{n\hat{\sigma}^2}{G_n^{-1}(1-\frac{\alpha}{2})}, \frac{n\hat{\sigma}^2}{G_n^{-1}(\frac{\alpha}{2})}\right]$$

is a $(1-\alpha)$-confidence interval for σ^2. Here

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2$$

and G_n is the CDF of the $\chi^2(n)$-distribution. Indeed, since $n\hat{\sigma}^2/\sigma^2 \sim \chi^2(n)$,

$$\begin{aligned} & \mathbb{P}_{\sigma^2}\left(\frac{n\hat{\sigma}^2}{G_n^{-1}(1-\frac{\alpha}{2})} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{G_n^{-1}(\frac{\alpha}{2})}\right) \\ = \ & \mathbb{P}_{\sigma}\left(G_n^{-1}(\tfrac{\alpha}{2}) \le \frac{n\hat{\sigma}^2}{\sigma^2} \le G_n^{-1}(1-\tfrac{\alpha}{2})\right) = 1-\alpha. \end{aligned}$$

Confidence interval for σ^2, μ unknown

Then

$$\left[\frac{(n-1)S^2}{G_{n-1}^{-1}(1-\frac{\alpha}{2})}, \frac{(n-1)S^2}{G_{n-1}^{-1}(\frac{\alpha}{2})}\right]$$

is a $(1-\alpha)$-confidence interval for σ^2. Here

$$S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$$

and G_{n-1} is the CDF of the $\chi^2(n-1)$-distribution. A one-sided confidence interval for σ^2 (right-sided) is

$$\left[0, \frac{(n-1)S^2}{G_{n-1}^{-1}(\alpha)}\right],$$

since

$$\mathbb{P}_{\mu,\sigma^2}\left(\sigma^2 \le \frac{(n-1)S^2}{G_{n-1}^{-1}(\alpha)}\right) = \mathbb{P}_{\mu,\sigma^2}\left(\frac{(n-1)S^2}{\sigma^2} \ge G_{n-1}^{-1}(\alpha)\right) = 1-\alpha.$$

Numerical example continued

The sample size is $n = 6$. We take $\alpha = .05$. Then $G_{n-1}^{-1}(1-\frac{\alpha}{2}) = 12.83$ and

$G_{n-1}^{-1}(\frac{\alpha}{2}) = .83$. The sample variance is $s^2 = .8$. So a 95% confidence interval for σ^2 is

$$.312 \le \sigma^2 \le 4.18$$

and so a 95% confidence interval for σ is

$$.56 = \sqrt{.312} \le \sigma \le \sqrt{4.18} = 2.19.$$

If one is interested in a upper bound for σ^2 we use that $G_{n-1}^{-1}(\alpha) = 1.145$. So a one-sided 95% confidence interval for σ^2 is

$$\sigma^2 \le 3.491$$

and a one-sided 95% confidence interval for σ is

$$\sigma \le \sqrt{3.491} = 1.868.$$

8.3 Approximate confidence interval when the data are Poisson

We revisit the example of Section 1.7. Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ)-distributed. Then $\mathbf{X} := \sum_{i=1}^n X_i$ has a Poisson($n\lambda$)-distribution. By the duality between tests and confidence sets as explained in Chapter 9.2 ahead, one can construct an exact $(1-\alpha)$-confidence set for λ by testing for all λ_0 the hypothesis $H_0 : \lambda = \lambda_0$ at level α and taking the confidence set as those values λ_0 that are not rejected. Yes, this is possible, but not easy and does not give us explicit expressions. Therefore, let us use the central limit theorem instead.

We have $\mathbf{X} \sim$ Poisson(λ_n) with $\lambda_n := n\lambda$. In other words, we can reduce to situation to one where we have one observation $\mathbf{X}$ from a Poisson distribution with parameter λ_n.
We take $\alpha = .05$ and for simplicity replace $\Phi^{-1}(1 - \frac{\alpha}{2}) = 1.96$ by 2.

Approximate confidence interval for λ_n using the CLT

For λ_n large, $(\mathbf{X} - \lambda_n)/\sqrt{\lambda_n}$ is approximately $\mathcal{N}(0,1)$ distributed. Hence

$$\mathbb{P}_{\lambda_n}\left(\frac{|\mathbf{X} - \lambda_n|}{\sqrt{\lambda_n}} \le 2 \right) \approx .95.$$

Rewrite this to

$$\mathbb{P}_{\lambda_n}\left(\lambda_n \in \left[\mathbf{X} + 2 - 2\sqrt{\mathbf{X}+1}, \mathbf{X} + 2 + 2\sqrt{\mathbf{X}+1} \right] \right) \approx .95.$$

So

$$\left[\mathbf{X} + 2 - 2\sqrt{\mathbf{X}+1}, X + 2 + 2\sqrt{\mathbf{X}+1} \right]$$

is an approximate 95% confidence interval for λ_n.

Approximate confidence interval for λ_n using the CLT and estimated variance

We can estimate the variance by

$$\widehat{\mathrm{Var}}(\mathbf{X}) := \mathbf{X}.$$

For λ_n large $(\mathbf{X}-\lambda)/\sqrt{\mathbf{X}}$ is approximately $\mathcal{N}(0,1)$-distributed (see e.g. *Fundamentals of Mathematical Statistics*). An approximate 95% confidence interval based on this is

$$\left[\mathbf{X}-2\sqrt{\mathbf{X}}, \mathbf{X}+2\sqrt{\mathbf{X}}\right].$$

Chapter 9

Duality between confidence sets and tests

In this chapter we replace $\mathbb{X} \in \mathcal{X}^n$ by $X \in \mathcal{X}$ and $\mathbb{P}$ by P to make the notation less Baroque. We assume $P = P_\theta$ with $\theta \in \Theta$ and consider a paramater of interest $\gamma = g(\theta) \in \Gamma = \{g(\vartheta) : \ \vartheta \in \Theta\}$ which is possibly not real-valued.

9.1 Definition of a confidence set

Let I be a mapping

$$I : \ \mathcal{X} \to \{\text{subsets of } \Gamma\}$$

such that $\{x \in \mathcal{X} : \ \gamma \in I(x)\}$ is measurable for all $\gamma \in \Gamma$.

Definition 9.1.1 *One calls $I(X)$ a $(1-\alpha)$-confidence set for $g(\theta)$ if*

$$P_\theta\biggl(g(\theta) \in I(X) \biggr) \ge 1 - \alpha, \ \forall \ \theta \in \Theta.$$

9.2 The duality theorem

Consider some set $C \subset \mathcal{X} \times \Gamma$ and let for $\gamma \in \Gamma$

$$J_\gamma := \{x \in \mathcal{X} : \ (x, \gamma) \in C\} \subset \mathcal{X},$$

and for $x \in \mathcal{X}$

$$I(x) := \{\gamma \in \Gamma : \ (x, \gamma) \in C\} \subset \Gamma.$$

We assume that J_γ is measurable for all $\gamma \in \Gamma$.

Theorem 9.2.1 *(duality theorem)*
The set $I(X)$ is a $(1-\alpha)$-confidence set
$\Leftrightarrow$
For all $\gamma_0 \in \Gamma$, $\phi_{\gamma_0}(X) := 1_{J^c_{\gamma_0}}(X)$ is a level α test for $H_0 : \ \gamma = \gamma_0$.

Proof.

$$\begin{aligned} P_\theta\Big(\phi_\gamma(X)=1\Big) &= P_\theta\Big(X\notin J_\gamma\Big) \\ &= P_\theta\Big((X,\gamma)\notin C\Big) \\ &= 1-P_\theta\Big((X,\gamma)\in C\Big) \\ &= 1-P_\theta\Big(\gamma\in I(X)\Big). \end{aligned}$$

□

Example 9.2.1 *In this example we return to the notation $\mathbb{X}=(X_1,\ldots,X_n)$. Let $X_1,\ldots,X_n$ be i.i.d. $\mathcal{N}(\mu,\sigma^2)$ with $\sigma^2=:\sigma_0^2$ known. We let $\gamma:=\mu$. Then we may take*

$$I(\mathbb{X})=\Big[\bar{X}-\Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n},\bar{X}+\Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n}\Big],$$

and then

$$J_\mu=\Big[\mu-\Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n},\mu+\Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n}\Big].$$

9.3 Confidence intervals when X is binomial

Consider $X\sim\text{Binomial}(n,\theta)$ with $0\le\theta\le 1$ unknown. We present three ways for the construction of confidence intervals for θ.
$\boxed{\text{Exact confidence interval using the Duality Theorem}}$
For the hypothesis
$H_0:\ \theta=\theta_0\ \ ,$
we use the test

$$\phi(X,\theta_0):=\begin{cases}1 & X>\bar{c}(\theta_0)\ \text{or}X<\underline{c}(\theta_0)\\ 0 & \text{else}\end{cases},$$

where $\underline{c}(\theta_0)\le\bar{c}(\theta_0)$ (both in $\{0,\ldots,n\}$) are determined by

$$\underbrace{P_{\theta_0}\Big(X>\bar{c}(\theta_0)\Big)}_{=\sum_{k>\bar{c}(\theta_0)}\binom{n}{k}\theta_0^k(1-\theta_0)^{n-k}}\le\frac{\alpha}{2}\le P_{\theta_0}\Big(X>\bar{c}(\theta_0)-1\Big)$$

$$P_{\theta_0}\Big(X<\underline{c}(\theta_0)\Big)\le\frac{\alpha}{2}\le P_{\theta_0}\Big(X<\underline{c}(\theta_0)+1\Big).$$

So

$$J_{\theta_0}=\{x\in\{0,\ldots,n\}:\underline{c}(\theta_0)\le x\le\bar{c}(\theta_0)\}$$

and

$$C = \{(x,\theta) \in \{0,\ldots,n\} \times [0,1] : \ \underline{c}(\theta) \le x \le \bar{c}(\theta)\},$$
$$I(x) = \{\theta \in [0,1] : \ \underline{c}(\theta) \le x \le \bar{c}(\theta)\}.$$

We let for $x \in \{0,\ldots,n-1\}$, $\bar{\theta}(x)$ be defined by

$$\sum_{k<x} \binom{n}{k} \bar{\theta}(x)^k (1-\bar{\theta}(x))^{n-k} = \frac{\alpha}{2}$$

and for $x \in \{1,\ldots,n\}$, $\underline{\theta}(x)$ be defined by

$$\sum_{k>x} \binom{n}{k} \underline{\theta}(x)^k (1-\underline{\theta}(x))^{n-k} = \frac{\alpha}{2}$$

and further take $\bar{\theta}(n) = 1$ and $\underline{\theta}(0) = 0$. Then $[\underline{\theta}(X), \bar{\theta}(X)]$ is an exact $(1-\alpha)$-confidence interval for θ.

Approximate confidence interval using the CLT

We reject

$H_0 : \ \theta = \theta_0$,

when

$$\frac{|X - n\theta_0|}{\sqrt{n\theta_0(1-\theta_0)}} > \underbrace{\Phi^{-1}(1-\tfrac{\alpha}{2})}_{:=z}.$$

So

$$\begin{aligned} I(X) &= \left\{\theta : \ \frac{|X-n\theta|}{\sqrt{n\theta(1-\theta)}} > z\right\} \\ &= \left\{\theta \in \frac{X + \frac{z^2}{2}}{n+z^2} \pm \frac{\sqrt{\frac{z^2 X(n-X)}{n} + \frac{z^4}{4}}}{n+z^2}\right\}, \end{aligned}$$

where the second equality follows after some calculations. Approximate confidence interval using the CLT a

By the CLT

$$\frac{X - n\theta}{\sqrt{\mathrm{Var}_\theta(X)}}$$

is approximately $\mathcal{N}(0,1)$-distributed. We have $\mathrm{Var}_\theta(X) = n\theta(1-\theta)$ which can be estimated by

$$\widehat{\mathrm{Var}}_\theta(X) := n\hat{\theta}(1-\hat{\theta}).$$

Then

$$\frac{X - n\theta}{\sqrt{\widehat{\mathrm{Var}}_\theta(X)}}$$

is still approximately $\mathcal{N}(0,1)$-distributed (see Section 1.6). We can then take

$$I(X) := \left\{\theta \in \frac{X}{n} \pm z\sqrt{\frac{X}{n}\left(1-\frac{X}{n}\right)}/\sqrt{n}\right\}$$

$$= \left\{ \theta \in \frac{X}{n} \pm \frac{\sqrt{\frac{z^2 X(n-X)}{n}}}{n} \right\}.$$

Numerical example

Let $n = 38$ and suppose we observe $X = 20$. Then, using the third method above, an approximate 95% confidence interval for θ (and using $\Phi^{-1}(.975) \approx 2$) is

$$\frac{20}{38} \pm 2\sqrt{\frac{20 \times 18}{38^3}} = .526 \pm .162.$$

Chapter 10

Bayesian statistics

In this chapter we again replace $\mathbb{X}$ by X, etc, to avoid a too Baroque notation. Thus $X \in \mathcal{X}$ represents the data. We assume X has distribution P_θ, with $\theta \in \Theta$ an unknown parameter. In frequentist statistics one assumes the unknown θ to be fixed (nonrandom). In Bayesian statistics on assumes θ to be random.

For example, suppose you visit your doctor. You ask your doctor: what is the probability θ that someone like me has the disease? He might say: some studies indicate θ is about $1/2$, others experts find it is almost zero, but there are also reports which point in the direction to it being close to one. And he continues: I would personally say this probability θ can be anything between 0 and 1, each value is equally likely. Then the doctor seems to model a probability as a random variable, assigning uniform weights to all possible values. The doctor has a uniform prior for θ. Now the doctor carries out some tests on you. Given the outcome X of these tests, you ask the doctor: what is now the probability that I have the disease? With the data in hand, the doctor updates beliefs for your case to posterior beliefs, and will hopefully share these to you.

Suppose $\mathcal{P} := \{P_\theta : \ \theta \in \Theta\}$ is dominated by a sigma-finite measure ν. Before, we wrote for $\theta \in \Theta$ the densities as

$$p_\theta(x) = \frac{dP_\theta}{d\nu}(x), \ x \in \mathcal{X}.$$

In the Bayesian notation

$$p_\theta(x) = p(x|\theta), \ x \in \mathcal{X},$$

is the density given the parameter value is θ. To make this work we suppose Θ is measurable space.

10.1 Prior, marginal and posterior

Definition 10.1.1 *Let Π be a given probability distribution on Θ, the prior. For a dominating measure μ the prior density of θ is*

$$w(\vartheta) := \frac{d\Pi}{d\mu}(\vartheta), \ \vartheta \in \Theta.$$

Remark 10.1.1

- *If Θ is countable we let $w(\cdot)$ be the probability mass function of θ.*
- *If $\Theta = \mathbb{R}$ and if Π is absolutely continuous, we let $w(\cdot)$ be the Lebesgue density of θ.*
- *In both discrete and absolutely continuous case we call $w(\cdot)$ a density. Other cases will not be considered in this lecture.*

Definition 10.1.2 *The marginal density of X is*

$$p(x) = \int p(x|\vartheta)w(\vartheta)d\mu(\vartheta) = \begin{cases} \sum_\vartheta p(x|\vartheta)w(\vartheta) & \theta \text{ discrete} \\ \int_\vartheta p(x|\vartheta)w(\vartheta)d\vartheta & \theta \text{ abs. continuous} \end{cases}, \ x \in \mathcal{X}.$$

Definition 10.1.3 *For $p(x) > 0$ the posterior density of θ given $X = x$ is*

$$w(\vartheta|x) := \frac{p(x|\vartheta)w(\vartheta)}{p(x)}.$$

The posterior density is thus given by Bayes' rule.

10.2 The maximum a posteriori estimator

With the Bayesian approach, the data X lead to a posterior distribution for θ. But one might also want a point estimator of θ, some value as representative of the parameter. This could be the mean or the median of the posterior distribution (when $\Theta \subset \mathbb{R}$). Another representative is the most likely value for θ given the data X.

Definition 10.2.1 *The maximum a posteriori (MAP) estimator is*

$$\hat{\theta}_{\text{MAP}} := \hat{\theta}_{\text{MAP}}(X) := \arg\max_{\vartheta \in \Theta} w(\vartheta|X),$$

provided the maximum exists.

Note To find $\hat{\theta}_{\text{MAP}}$ you do not need to calculate the marginal distribution $p(\cdot)$:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\vartheta \in \Theta} p(X|\vartheta)w(\vartheta).$$

We apply the proportional symbol $\propto$: for real-valued functions f and g with domain Θ, we write $f(\vartheta) \propto g(\vartheta) \neq 0$ if $f(\vartheta)/g(\vartheta)$ does not depend on $\vartheta \in \Theta$. So $w(\vartheta|x) \propto p(x|\vartheta)w(\vartheta)$.

Note We may also write

$$\hat{\theta}_{\rm MAP} = \arg\max_{\vartheta\in\Theta}\biggl\{\log p(X|\vartheta) + \log w(\vartheta)\biggr\}.$$

In other words, the MAP maximizes the log-likelihood $\log L_X(\vartheta)$ penalized with a "regularization term" $\log w(\vartheta)$.

Example 10.2.1 *Let, given $\theta \in \mathbb{R}$, $\mathbb{X} = (X_1, \ldots, X_n)$ be an i.i.d. sample of the $\mathcal{N}(\theta, 1)$-distribution. Suppose the prior on θ is the $\mathcal{N}(0, 1/\lambda^2)$-distribution, where $\lambda > 0$ is given. Then*

$$\hat{\theta}_{\rm MAP} = \arg\max_{\vartheta\in\mathbb{R}}\biggl\{-\frac{1}{2}\sum_{i=1}^n (X_i - \vartheta)^2 - \frac{1}{2}\lambda^2\vartheta^2\biggr\} = \frac{\bar{X}}{1+\lambda^2/n}.$$

We see that the MAP is a shrinked version of the MLE $\bar{X}$. This makes sense, because the $\mathcal{N}(0, 1/\lambda^2)$-prior has a preference for values of θ near zero, and this is reflected in the MAP by the shrinkage of $\bar{X}$ to zero.

10.3 Bayes' decision in classification

Consider two given densities $p_0(x)$ and $p_1(x)$, $x \in \mathcal{X}$. Given an observation X, we want to classify it as coming from distribution P_0 (with density p_0) or P_1 (with density p_1). Let the prior be

$$w(\vartheta) = \begin{cases} w_0, & \vartheta = 0 \\ w_1, & \vartheta = 1 \end{cases},$$

for given $0 < w_0 < 1$ and $w_1 = 1 - w_0$. Then the MAP estimator is

$$\hat{\theta}_{\rm MAP} = \begin{cases} 1 & \frac{p_1(X)}{p_0(X)} > \frac{w_0}{w_1} \\ q & \frac{p_1(X)}{p_0(X)} = \frac{w_0}{w_1} \\ 0 & \frac{p_1(X)}{p_0(X)} < \frac{w_0}{w_1} \end{cases}$$

where $q \in \{0, 1\}$ is arbitrary. Here, use that

$$w(\vartheta|x) = \begin{cases} p_0(x)w_0/p(x), & \vartheta = 0 \\ p_1(x)w_1/p(x), & \vartheta = 1 \end{cases}.$$

Note that

$$p(x) = w_0 p_0(x) + w_1 p_1(x),\ x \in \mathcal{X},$$

is a mixture of p_0 and p_1.

The estimator $\hat{\theta}_{\text{MAP}}$ is called Bayes' decision, which we write as ϕ_{Bayes}. Note that ϕ_{Bayes} is of the same form as a Neyman-Pearson test defined in Chapter 4.

Now let us reformulate the classification problem. We simply use the notation Y instead of θ. Let $Y \in \{0, 1\}$ be a label and $X \in \mathcal{X}$ be the features. Write

$$\eta(x) = P(Y = 1|X = x), \ x \in \mathcal{X}.$$

Then Bayes' decision is is choosing the most likely label: given $X = x$, predict $Y = 1$ if $\eta(x) > 1/2$ and predict $Y = 0$ if $\eta(x) < 1/2$ and randomize if $\eta(x) = 1/2$. In other words,

$$\phi_{\text{Bayes}}(X) = \mathrm{l}\{\eta(X) > 1/2\} + q\mathrm{l}\{\eta = 1/2\}.$$

One can present the situation in terms of decision theory. There are two possible actions $a = 0$ (classify as coming from p_0) and $a = 1$ (classify as coming from p_1). The action space is thus $\mathcal{A} := \{0, 1\}$. We define the loss function as the event of making a mistake:

$$L(y, a) := \mathrm{l}_{\{y \neq a\}}, \quad (y, a) \in \{0, 1\}^2.$$

This means one unit loss for taking a wrong action. We call a function $\phi : \mathcal{X} \to \{0, 1\}$ a decision and define its risk as

$$R(y, \phi) := E[L(y, \phi(X))|Y = y].$$

Thus

$$R(y, \phi) = \begin{cases} P_0(\phi(X) = 1), & y = 0 \\ P_1(\phi(X) = 0), & y = 1 \end{cases} .$$

We then define the Bayes risk of ϕ as the average risk over Y, where $P(Y = 1) = w_1$ and $P(Y = 0) = w_0$

$$r_w(\phi) = w_0 P_0(\phi(X) = 1) + w_1 P_1(\phi(X) = 0) = P(\phi(X) \neq Y).$$

Bayes' decision is the minimizer of the Bayes risk

$$\phi_{\text{Bayes}} = \arg \min_{\phi:\ \mathcal{X} \to \{0,1\}} r_w(\phi).$$

Remark 10.3.1 *This remark makes a link to machine learning. You may find it useful to see this connection, but it is not exam material for this course.*

In the supervised learning setup for classification, one again has a label $Y \in \{0, 1\}$ and features $X \in \mathcal{X}$. The distribution of (X, Y) is (in part) unknown. One either starts with a model for $\eta(x) := P(Y = 1|X = x)$, or for the two densities $p_1(x) = p(x|Y = 1)$ and $p_0(x) = p(x|Y = 0)$. For instance, in logistic regression, one uses given feature mappings $\psi_j : \mathcal{X} \to \mathbb{R}$, $j = 1, \ldots, d$, and takes

$$\eta(x) = \frac{1}{1 + \exp[\sum_{j=1}^{d} \theta_j \psi_j(x)]},$$

where $\theta = (\theta_1, \ldots, \theta_d)^\top \in \mathbb{R}^d$ is an known parameter. In linear discriminant analysis one has $\mathcal{X} = \mathbb{R}^d$ (possibly after a feature mapping) and models p_1 as the $\mathcal{N}(\mu_1, \Sigma)$ density and p_0 as the $\mathcal{N}(\mu_0, \Sigma)$ density with $\mu_1 \in \mathbb{R}^d$ and $\mu_0 \in \mathbb{R}^d$ unknown means and $\Sigma \in \mathbb{R}^{d\times d}$ an unknown covariance matrix. In either case, based on data $\{X_i, Y_i\}_{i=1}^n$ one estimates the unknown parameters to obtain an estimator $\hat{\eta}(\cdot)$ of $\eta(\cdot)$ and applies the estimated Bayes' decision

$$\hat{\phi}_{\text{Bayes}}(X) = \mathrm{l}\{\hat{\eta}(X) > 1/2\} + q\mathrm{l}\{\hat{\eta} = 1/2\}.$$

10.4 Bayesian inference for the binomial distribution

Let $X|\theta \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(r, s)$. Then the prior mean is $E\theta = \frac{r}{r+s}$. The posterior density is

$$\begin{aligned} w(\vartheta|x) &\propto p(x|\vartheta)w(\vartheta) \propto \vartheta^x(1-\vartheta)^{n-x}\vartheta^{s-1}(1-\vartheta)^{r-1} \\ &= \vartheta^{x+s-1}(1-\vartheta)^{n-x+r-1}. \end{aligned}$$

So $\theta|X = x \sim \text{Beta}(x + r, n - x - s)$ and the posterior mean is

$$E(\theta|X) = \frac{X + r}{n + r + s}.$$

The MAP estimator is

$$\hat{\theta}_{\text{MAP}} = \frac{X + r - 1}{n + s + r - 2}.$$

If for example one starts with the uniform distribution as prior, one finds as posterior

$$w(\vartheta|X) = (n+1)\binom{n}{X}\vartheta^X(1-\vartheta)^{n-X}$$

and $\hat{\theta}_{\text{MAP}}$ is equal to the maximum likelihood estimator $\hat{\theta}_{\text{MLE}} = X/n$. This follows more generally from Definition 10.2.1: if Π is the uniform[1] distribution on Θ, its density is constant over Θ so it plays no role in the maximization, and therefore that $\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}}$

10.5 Bayesian inference for the normal distribution

We revisit Example 10.2.1. Let $X|\theta \sim \mathcal{N}(\theta, \sigma^2)$ were $\theta \in \mathbb{R}$ and where σ^2 is known. Suppose $\theta \sim \mathcal{N}(0, \tau^2)$ for some given $\tau^2 > 0$. Then the posterior is

$$\theta|X \sim \mathcal{N}\left(\frac{\tau^2}{\tau^2 + \sigma^2}X, \frac{\tau^2\sigma^2}{\tau^2 + \sigma^2}\right).$$

We see that the posterior mean is

$$E(\theta|X) = \frac{\tau^2}{\tau^2 + \sigma^2}X.$$

In this case this is also the MAP estimator.

[1] If $|\Theta|$ is infinite there is no uniform distribution on Θ.

Chapter 11

The linear model

The linear model will be defined in Definition 11.2.1. Models are only approximations. In this chapter we allow for a misspecified linear model and study "linear approximations". Recall Example 2.3.1, where we defined the best linear prediction of $\mathbf{Y} \in \mathbb{R}$ given $\mathbf{X} \in \mathbb{R}$ (in that example $(\mathbf{X}, \mathbf{Y})$ was called (X, Y), but in this chapter X and Y will have another meaning). Let $x_i := \{x_{i,j}\}_{j=1}^d \in \mathbb{R}^d$ be fixed and $Y_i \in \mathbb{R}$ be random, $i = 1, \ldots, n$. These data could be based on i.i.d. copies $\{(X_i, Y_i)\}_{i=1}^n$of a pair $(\mathbf{X}, \mathbf{Y})$, with $\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{Y} \in \mathbb{R}$, and we condition on the realizations x_i of X_i, $i = 1, \ldots, n$. We call $\{x_i\}$ (or $\{X_i\}$) the co-variables. Note that given these co-variables $X_i = x_i$, $i = 1, \ldots, n$, the random variables $Y_1, \ldots, Y_n$ remain independent but are (possibly) no longer identically distributed.

We aim at estimating the best linear approximation (defined formally in Definition 11.2.2 below) of Y_i given $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, by minimizing

$$\sum_{i=1}^n \biggl(Y_i - a - \sum_{j=1}^d x_{i,j} b_j \biggr)^2 .$$

over $a \in \mathbb{R}$ and $b = (b_1, \ldots, b_d)^\top \in \mathbb{R}^d$.

To simplify the expressions, we rename the quantities involved as follows. Define for all i, $x_{i,0} := 1$ and define $b_0 := a$. Then for all i we have $a + \sum_{j=1}^p x_{i,j} b_j = \sum_{j=0}^d x_{i,j} b_j$.

Then we minimize

$$\sum_{i=1}^n \biggl(Y_i - \sum_{j=0}^d x_{i,j} b_j \biggr)^2 .$$

over $b = (b_0, b_1, \ldots, b_d)^T \in \mathbb{R}^{d+1}$.

11.1 Definition of the least squares estimator

We let $p := d + 1$ and

$$X := \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,d} \\ 1 & x_{2,1} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,d} \end{pmatrix} \in \mathbb{R}^{n \times p},\ Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

Thus, in this chapter $X \in \mathbb{R}^{n \times p}$ is a given matrix with (non-random) entries $\{x_{i,j}\}$.

One calls X the design matrix or input matrix. We assume it to be non-random, which is called the case of fixed design. Moreover, $Y \in \mathbb{R}^n$ is the vector of responses or output vector.

We will assume throughout this chapter:

Condition 11.1.1 *The design matrix X has rank p.*

Let us denote the Euclidean norm of a vector $v \in \mathbb{R}^n$ by

$$\|v\|_2 := \sqrt{\sum_{i=1}^n v_i^2}.$$

Then

$$\sum_{i=1}^n \biggl(Y_i - \sum_{j=0}^d x_{i,j} b_j \biggr)^2 = \|Y - Xb\|_2^2,\ b \in \mathbb{R}^p.$$

Definition 11.1.1 *One calls*

$$\hat{\beta} := \arg\min_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2$$

the least squares estimator (LSE).

The distance between Y and the space $\{Xb :\ b \in \mathbb{R}^p\}$ spanned by the columns of X is minimized by projecting Y on this space. In fact, one has

Lemma 11.1.1 *Suppose X has rank p. Then*

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

Proof. It holds that

$$\frac{1}{2} \frac{\partial}{\partial b} \|Y - Xb\|_2^2 = -X^\top (Y - Xb).$$

It follows that $\hat{\beta}$ is a solution of the so-called normal equations

$$X^\top (Y - X\hat{\beta}) = 0$$

or

$$X^\top Y = X^\top X\hat{\beta}.$$

As X has rank p, the matrix $X^\top X$ has an inverse $(X^\top X)^{-1}$ and we get

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

□

The projection of Y on $\{Xb : \ b \in \mathbb{R}^p\}$ is

$$\underbrace{X(X^\top X)^{-1}X^\top}_{\text{projection}} Y.$$

Recall that a projection is a linear map of the form $PP^\top$ such that $P^\top P = I$. We can write $X(X^\top X)^{-1}X^\top := PP^\top$, where P is an orthonormal basis for the column space of X.[1]

Example 11.1.1 *(Example with $d = 1$)*
For $d = 1$

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Then

$$X^\top X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix},$$

$$(X^\top X)^{-1} = \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.$$

Moreover

$$X^\top Y = \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix}.$$

We now obtain (with $\hat{\alpha} := \hat{\beta}_0$, $\hat{\beta} := \hat{\beta}_1$)

$$\begin{aligned}
\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} &= (X^T X)^{-1} X^T Y \\
&= \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix} \\
&= \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \sum_{i=1}^n x_i^2 \bar{Y} - \bar{x}\sum_{i=1}^n x_i Y_i \\ -n\bar{x}\bar{Y} + \sum_{i=1}^n x_i Y_i \end{pmatrix} \\
&= \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \sum_{i=1}^n (x_i - \bar{x})^2 - \bar{x}(\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}) \\ \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} \end{pmatrix}.
\end{aligned}$$

[1] Write the singular value decomposition of X as $X = P\phi Q^\top$, where $\phi = \text{diag}(\phi_1, \ldots, \phi_p)$ contains the singular values of $X$and where $P^\top P = I$ and $Q^\top Q = I$.

Here we used that $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2$. *We can moreover write*

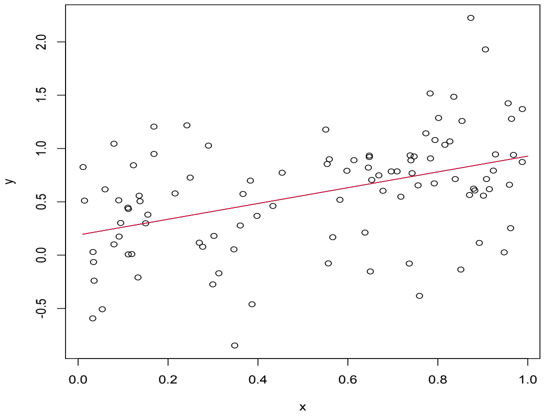
$$\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}).$$

Thus

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \bar{Y} - \hat{\beta}\bar{x} \\ \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{pmatrix}.$$

These expressions coincide with what we derived as method of moments estimators, see Example 2.3.1.

Note that if we assume that $\bar{x} = 0$ *the calculations become much simpler. If* $\bar{x} = 0$, *the matrix* $X^\top X$ *becomes a diagonal matrix, and we find* $\hat{\alpha} = \bar{Y}$ *and* $\hat{\beta} = \sum_{i=1}^n Y_i x_i / \sum_{i=1}^n x_i^2$.



Simulated data with $Y = .3 + .6 \times x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \frac{1}{4})$, $\hat{\alpha} = .19$, $\hat{\beta} = .740$

11.2 Theoretical properties of the least squares estimator

We define the mean vector $f = \mathbb{E}Y \in \mathbb{R}^n$. (Recall that X is fixed. One may see f_i as the mean of Y_i given x_i, $i = 1, \ldots, n$.) We call f the signal. The noise is defined as $\epsilon = Y - f$. This gives the signal+noise model

$$Y = f + \epsilon.$$

Definition 11.2.1 *The linear model (or linear regression model) is*

$$f = X\beta$$

where $\beta \in \mathbb{R}^p$ *is an unknown parameter.*

Thus in the linear regression model

$$Y = X\beta + \epsilon,$$

where $\beta \in \mathbb{R}^p$ is unknown and ϵ is an unobservable noise vector with independent, mean zero entries.

If the linear model is true, the LSE $\hat{\beta}$ is an estimator of β. We however allow for a misspecified model. Then $\hat{\beta}$ is an estimator of β^* given in the following definition.

Definition 11.2.2 *Let $\beta^* := (X^\top X)^{-1} X^\top f$. We call $X\beta^*$ the best linear approximation of the vector f.*

Thus $X\beta^*$ is the projection of f on the space spanned by the rows of X.

Lemma 11.2.1 *Suppose $\mathbb{E}\epsilon\epsilon^\top = \sigma^2 I$. Then*
i) $\mathbb{E}\hat{\beta} = \beta^*$, $\text{Cov}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$,
ii) $\mathbb{E}\|X(\hat{\beta} - \beta^*)\|_2^2 = \sigma^2 p$,
iii) $\mathbb{E}\|X\hat{\beta} - f\|_2^2 = \underbrace{\|X\beta^* - f\|_2^2}_{\substack{\text{approximation}\\ \text{error}}} + \underbrace{\sigma^2 p}_{\substack{\text{estimation}\\ \text{error}}}$.

Proof.
i) By straightforward computation

$$\hat{\beta} - \beta^* = \underbrace{(X^\top X)^{-1} X^\top}_{:=B} \epsilon.$$

We therefore have

$$\mathbb{E}(\hat{\beta} - \beta^*) = B\mathbb{E}\epsilon = 0,$$

and the covariance matrix of $\hat{\beta}$ is

$$\text{Cov}(\hat{\beta}) = \text{Cov}(B\epsilon) = B \underbrace{\text{Cov}(\epsilon)}_{=\sigma^2 I} B^\top$$

$$= \sigma^2 BB^\top = \sigma^2 (X^\top X)^{-1}.$$

ii) Define the projection $PP^\top := X(X^\top X)^{-1} X^\top$. Then

$$\|X(\hat{\beta} - \beta^*)\|_2^2 = \|PP^\top \epsilon\|_2^2 := \sum_{j=1}^p V_j^2,$$

where $V := P^\top \epsilon$,

$$\mathbb{E}V = P^\top \mathbb{E}\epsilon = 0,$$

and

$$\text{Cov}(V) = P^\top \text{Cov}(\epsilon) P = \sigma^2 I.$$

It follows that

$$\mathbb{E}\sum_{j=1}^p V_j^2 = \sum_{j=1}^p \mathbb{E}V_j^2 = \sigma^2 p.$$

iii) It holds by Pythagoras' rule for all b

$$\|Xb - f\|_2^2 = \|X(b - \beta^*)\|_2^2 + \|X\beta^* - f\|_2^2$$

since $X\beta^* - f$ is orthogonal to X. □

Lemma 11.2.2 *Suppose $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Then we have*
i) $\hat{\beta} - \beta^* \sim \mathcal{N}(0, \sigma^2 (X^\top X)^{-1})$,
ii) $\frac{\|X(\hat{\beta}-\beta^*)\|_2^2}{\sigma^2} \sim \chi^2(p)$.

Proof.
i) Since $\hat{\beta}$ is a linear function of the multivariate normal noise vector ϵ, the least squares estimator $\hat{\beta}$ is also multivariate normal. The result follows from Lemma 11.2.1.
ii) Define the projection $PP^\top := X(X^\top X)^{-1} X^\top$. Then

$$\|X(\hat{\beta} - \beta^*)\|_2^2 = \|PP^\top \epsilon\|_2^2 := \sum_{j=1}^{p} V_j^2 .$$

Now $V := P^\top \epsilon$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. □

Remark 11.2.1 *More generally, under appropriate conditions, many estimators are approximately normally distributed (for example the sample median) and many test statistics have approximately a χ^2 null-distribution (for example the χ^2 goodness-of-fit statistic). This phenomenon occurs because many models can in a certain sense be approximated by the linear model and many minus log-likelihoods resemble the least squares loss function (applying a two-term Taylor expansion). Understanding the linear model is a first step towards understanding a wide range of more complicated models.*

Corollary 11.2.1 *Suppose the linear model is well-specified:*

$$Y = X\beta + \epsilon$$

Assume moreover that $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. If $\sigma^2 := \sigma_0^2$ is known, a test for
$H_0 : \ \beta = \beta_0 \ \ ,$
is:
reject H_0 when $\|X(\hat{\beta} - \beta^0)\|_2^2 / \sigma_0^2 > G_p^{-1}(1-\alpha)$,
where G_p is the CDF of a $\chi^2(p)$-distributed random variable.

Remark 11.2.2 *When σ^2 is unknown one may estimate it using the estimator*

$$\hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|_2^2}{n-p},$$

where $\hat{\epsilon} := Y - X\hat{\beta}$ is the vector of residuals. Under the assumptions of the previous corollary (but now with possibly unknown σ^2) the test statistic $\|X(\hat{\beta} - \beta^0)\|_2^2 / p / \hat{\sigma}^2$ has under H_0 a so-called F-distribution with p and $n-p$ degrees of freedom.

Chapter 12

High-dimensional statistics

Let $X_1, \ldots, X_n$ be independent observations with distribution depending on some parameter $\theta \in \Theta \subset \mathbb{R}^p$. Thus, the number of parameters is p and the number of observations is n. In high-dimensional statistics, p is "large", possibly $p \gg n$. We consider here a prototype example, namely the (approximate) linear model.

12.1 Definition of ridge estimator and Lasso

As in the previous chapter, the data are $(x_1, Y_1), \ldots, (x_n, Y_n)$ with co-variable $x_i = \{x_{i,j}\}_{j=0}^d \in \mathbb{R}^p$ a given p-dimensional vector with $x_{i,0} = 1$ and $Y_i \in \mathbb{R}$ a random response ($i = 1, \ldots, n$). One wants to find a good linear approximation using the least squares loss function

$$b \mapsto \sum_{i=1}^n \biggl(Y_i - \sum_{j=0}^d x_{i,j} b_j \biggr)^2 .$$

Define as in Chapter 11 $p := d + 1$ and

$$X := \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,d} \\ 1 & x_{2,1} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,d} \end{pmatrix} \in \mathbb{R}^{n \times p},\ Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} .$$

Then

$$\sum_{i=1}^n \biggl(Y_i - \sum_{j=0}^d x_{i,j} b_j \biggr)^2 = \| Y - Xb \|_2^2,\ b \in \mathbb{R}^p.$$

The difference with the previous chapter is now that we consider the high-dimensional situation where p is "large". This covers the case $p \geq n$ or even

$$\boxed{p \gg n}.$$

If $p \geq n$ the matrix X has rank at most n. If its rank is equal to n, then minimizing $\|Y - Xb\|_2^2$ over all $b \in \mathbb{R}^p$ gives a "perfect" solution $\hat{\beta}_{\text{LSE}}$ with $X\hat{\beta}_{\text{LSE}} = Y$. This solution just interpolates the data. It is of no use when design is fixed[1]: we say that it overfits the data.

To avoid overfitting one may use a penalization term that penalizes a too good fit. In general, the constant term b_0 is not penalized. Recall $b = \{b_j\}_{j=0}^d$. Let us define $b_{-0} = \{b_j\}_{j=1}^d$.

Definition 12.1.1 *The ridge regression estimator is*

$$\hat{\beta}_{\text{ridge}} := \arg\min_{b \in \mathbb{R}^p} \left\{ \|Y - Xb\|_2^2 + \lambda^2 \|b_{-0}\|_2^2 \right\},$$

where $\lambda > 0$ is a regularization parameter.

Definition 12.1.2 *The Lasso (Tibshirani [1996]: least absolute shrinkage and selection operator) is*

$$\hat{\beta}_{\text{Lasso}} := \arg\min_{b \in \mathbb{R}^p} \left\{ \|Y - Xb\|_2^2 + 2\lambda \|b_{-0}\|_1 \right\},$$

where $\lambda > 0$ is a regularization parameter and $\|b_{-0}\|_1 := \sum_{j=1}^d |b_j|$ is the ℓ_1-norm of b_{-0} .

Remark 12.1.1 *Suppose the linear model is correct: $Y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. The ridge regression estimator is the MAP estimator using as prior $\beta_1, \ldots, \beta_d$ i.i.d. $\sim \mathcal{N}(0, \tau^2)$. The Lasso estimator is the MAP using as prior $\beta_1, \ldots, \beta_d$ i.i.d. $\sim \text{Laplace}(0, \tau^2)$. The tuning parameter is then in both cases $\lambda^2 = \sigma^2/\tau^2$.*

Both ridge estimator and Lasso are biased. As λ increases the bias increases, but the variance decreases.

The regularization parameter λ is for example chosen by using "cross validation" or (information) theoretic or Bayesian arguments.

12.2 Theory for ridge estimator and Lasso

Let is write

$$\mathbf{x}_0 := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$$

and

$$X = \begin{pmatrix} \mathbf{x}_0 & \mathbf{x}_1 & \cdots & \mathbf{x}_d \end{pmatrix}.$$

In this section, we assume

[1] In recent work, data interpolation has been rehabilitated as it can be useful when the design is random.

Condition 12.2.1 *For $j = 1, \ldots, n$, the entries of $\mathbf{x}_j$ add up to zero.*

This condition can be made without loss of generality. It means that $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are orthogonal to $\mathbf{x}_0$: $\mathbf{x}_j^\top \mathbf{x}_0 = 0$. (Compare with Example 11.1.1.) Define

$$X_{-0} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_d \end{pmatrix} \in \mathbb{R}^{n \times d}.$$

Lemma 12.2.1 *Let $\hat\beta = (\hat\beta_0, \hat\beta_1 \ldots, \hat\beta_d$ be either the ridge estimator or the Lasso. Then $\hat\beta_0 = \bar Y$. Moreover,*

$$\hat\beta_{-0} = \arg \min_{b_{-0} \in \mathbb{R}^d} \left\{ \|Y - X_{-0} b_{-0}\|_2^2 / n + \text{pen}(b_{-0}) \right\}$$

where

$$\text{pen}(b_{-0}) = \begin{cases} \lambda^2 \|b_{-0}\|_2, & \text{if } \hat\beta \text{ is the ridge estimator} \\ 2\lambda \|b_{-0}\|_1, & \text{if } \hat\beta \text{ is the Lasso} \end{cases}.$$

Proof. It holds that

$$\|Y - Xb\|_2^2 = \|Y - \mathbf{x}_0 b_0\|_2^2 + \|Y - X_{-0} b_{-0}\|_2^2 - \|Y\|_2^2,$$

where we applied Pythagoras' rule. □

Lemma 12.2.2 *For $\hat\beta = \hat\beta_{\text{ridge}}$ we have*

$$\hat\beta_{-0} = (X_{-0}^\top X_{-0} + \lambda^2 I)^{-1} X_{-0}^\top Y.$$

Proof. We apply Lemma 12.2.1. We have

$$\begin{aligned} & \frac{1}{2} \frac{\partial}{\partial b_{-0}} \left\{ \|Y - X_{-0} b_{-0}\|_2^2 + \lambda^2 \|b_{-0}\|_2^2 \right\} \\ = \quad & -X_{-0}^\top (Y - X_{-0} b_{-0}) + \lambda^2 b_{-0} \\ = \quad & -X_{-0}^\top Y + \left(X_{-0}^\top X_{-0} + \lambda^2 I \right) b_{-0}. \end{aligned}$$

The estimator $\hat\beta_{-0}$ puts this to zero. □

For the Lasso estimator there is no explicit expression in general. We therefore only consider the special case of orthogonal design and that all columns in X_{-0} have the same length.

Lemma 12.2.3 *Suppose X is a fixed design matrix and $X_{-0}^\top X_{-0} = nI$ (thus $p \le n$ necessarily). Define $Z := X_{-0}^\top Y$. Then for $\hat\beta = \hat\beta_{\text{Lasso}}$, and for $j = 1, \ldots, d$,*

$$\hat\beta_j = \begin{cases} (Z_j - \lambda)/n & Z_j \ge \lambda \\ 0 & |Z_j| \le \lambda \\ (Z_j + \lambda)/n & Z_j \le -\lambda \end{cases}.$$

Proof. We apply Lemma 12.2.1. We can write

$$\begin{aligned} \|Y - X_{-0}b_{-0}\|_2^2 &= \|Y\|_2^2 - 2b_{-0}^T X_{-0}^\top Y + nb_{-0}^\top X_{-0}^\top X_{-0} b_{-0} \\ &= -2b_{-0}^\top Z + nb_{-0}^\top b_{-0}. \end{aligned}$$

Thus for each $j \in \{1, \ldots, d\}$ we minimize

$$-2b_j Z_j + nb_j^2 + 2\lambda |b_j|.$$

If $\hat{\beta}_j > 0$ it must be a solution of putting the derivative of the above expression to zero:

$$-Z_j + n\hat{\beta}_j + \lambda = 0,$$

or

$$\hat{\beta}_j = (Z_j - \lambda)/n.$$

Similarly, if $\hat{\beta}_j < 0$ we must have

$$-Z_j + n\hat{\beta}_j - \lambda = 0.$$

Otherwise $\hat{\beta}_j = 0$. □

From Lemma 12.2.2 we conclude that as λ grows the ridge estimator shrinks the coefficients towards zero. They will however not be set exactly to zero. From Lemma 12.2.3, the coefficients of the Lasso estimator shrink to zero as well and some - or even many - are set exactly to zero. It can be shown that this remains true when the design is not orthogonal. The ridge estimator can be useful if p is moderately large. For very large p the Lasso is to be preferred. The idea is that one should not try to estimate a signal when it is below the noise level. Instead, then one should simply put it to zero.

Some notation
∘ For a vector $z \in \mathbb{R}^d$ we let $\|z\|_\infty := \max_{1 \le j \le d} |z_j|$ be its ℓ_∞-norm.
∘ For a subset $S \subset \{1, \ldots, d\}$ with cardinality $s := |S|$ we let $b_S := \{b_j\}_{j \in S}$ and $X_S := \{\mathbf{x}_j\}_{j \in S}$.
∘ We let $f_{-0} = X_{-0}\beta_{-0}^*$ be the projection of f on the linear space spanned by the columns of X_{-0}.

In the next theorem we again assume orthogonal design.

Theorem 12.2.1 *Consider again orthogonal design with $X_{-0}^\top X_{-0} = nI$. Fix some level $\alpha \in (0,1)$ and suppose that for some λ_α it holds that $\mathbb{P}(\|X_{-0}^\top \epsilon\|_\infty > \lambda_\alpha) \le \alpha$. Let $\hat{\beta} = \hat{\beta}_{\text{Lasso}}$. Then for $\lambda > \lambda_\alpha$ we have with probability at least $1 - \alpha$*

$$\|X_{-0}\hat{\beta}_{-0} - f_{-0}\|_2^2 \le \min_S \Bigg\{ \underbrace{\|X_S \beta_S^* - f_{-0}\|_2^2}_{\substack{\text{approximation} \\ \text{error}}} + \underbrace{(\lambda + \lambda_\alpha)^2 s}_{\substack{\text{estimation} \\ \text{error}}} \Bigg\}.$$

Proof. On the set where $\|X_{-0}^\top \epsilon\|_\infty \le \lambda_\alpha$ we have for $j \in \{1, \ldots, d\}$
- $n|\beta_j^*| > \lambda + \lambda_\alpha \Rightarrow n|\hat{\beta}_j - \beta_j^*| \le \lambda + \lambda_\alpha$,

- $n|\beta^*_j| \leq \lambda + \lambda_\alpha \Rightarrow |\hat{\beta}_j - \beta^*_j| \leq |\beta^*_j|$.
So with probability at least $(1-\alpha)$,

$$\begin{aligned} \|X_{-0}\hat{\beta}_{-0} - f_{-0}\|_2^2 &\leq \sum_{n|\beta^*_j| \leq \lambda + \lambda_\alpha} n\beta_j^{*2} + (\lambda+\lambda_\alpha)^2 \Big(\#\{j : \ n|\beta^*_j| > \lambda + \lambda_\alpha\} \Big) \\ &= \min_S \Big\{ \|X_S\beta^*_S - f_{-0}\|_2^2 + (\lambda+\lambda_\alpha)^2 s \Big\}. \end{aligned}$$

□

Corollary 12.2.1 *Suppose that* β^*_{-0} *has* $s_* := \#\{j \in \{1, \ldots, d\} : \ \ \beta^*_j \neq 0\}$ *non-zero components. Then under the conditions of the above theorem, with probability at least* $1-\alpha$

$$\|X_{-0}(\hat{\beta}_{-0} - \beta^*_{-0})\|_2^2 \leq (\lambda+\lambda_\alpha)^2 s_*.$$

The above corollary tells us that the Lasso estimator adapts to favourable situations where β^* has many zeroes (i.e. where β^*_{-0} is sparse).

To complete the story, we need to study a bound for λ_α. It turns out that for many types of error distributions, one can take λ_α of order $\sqrt{\log p}$.

Remark 12.2.1 *The value* $\alpha = \frac{1}{2}$ *in Theorem 12.2.1 thus gives a bound for the median of* $\|X\hat{\beta}_{-0} - f_{-0}\|_2^2$*. In the case of Gaussian errors one may use "concentration of measure" to deduce that* $\|X\hat{\beta}_{-0} - f_{-0}\|_2^2$ *is "concentrated" around its median.*

Appendix A

Standard distributions

Standard discrete distributions

1. Bernoulli distribution with success parameter $p \in (0,1)$. $X \in \{0,1\}$ and

$$P(X = 1) = p, \quad EX = p, \quad \mathrm{Var}(X) = p(1-p).$$

2. Binomial distribution with n trials and success parameter $p \in (0,1)$. $X \in \{0, 1, \ldots, n\}$

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots n,$$

$$EX = np, \quad \mathrm{Var}(X) = np(1-p).$$

3. Poisson distribution with parameter $\lambda > 0$. $X \in \{0, 1, \ldots\}$

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \ldots,$$

$$EX = \lambda, \quad \mathrm{Var}(X) = \lambda.$$

Standard continuous distributions

4. Gaussian distribution with mean μ and variance σ^2. $X \in \mathbb{R}$,

$$f_X(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \ \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad x \in \mathbb{R}.$$

Denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$.

$$EX = \mu, \ \operatorname{var}(X) = \sigma^2.$$

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Leftrightarrow \quad Z := \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

$\mathcal{N}(0,1)$ is called the standard normal (or Gaussian).

5. The standard normal distribution function.

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2} \, dz, \quad x \in \mathbb{R}.$$

Let Φ^{-1} be its inverse function. Then,

$$\Phi^{-1}(0.9) = 1.28, \quad \Phi^{-1}(0.95) = 1.64, \quad \Phi^{-1}(0.975) = 1.96.$$

6. Exponential distribution with parameter $\lambda > 0$. $X \in \mathbb{R}_+ := [0, \infty)$,

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

$$EX = 1/\lambda, \quad \operatorname{Var}(X) = 1/\lambda^2.$$

7. Gamma distribution with parameters α, λ. $X \in \mathbb{R}_+ := [0, \infty)$,

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \ x^{\alpha-1} \ e^{-\lambda x}, \quad x \geq 0.$$

Here $\Gamma(\alpha)$ is the Gamma function and for integer values $\Gamma(m) = (m-1)!$.

$$EX = \alpha/\lambda, \quad \operatorname{Var}(X) = \alpha/\lambda^2.$$

8. Beta distribution with parameters r, s. $X \in [0,1]$,

$$f_X(x) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \ x^{r-1} \ (1-x)^{s-1}, \quad x \in [0,1].$$

$$EX = \frac{r}{r+s}, \quad \operatorname{Var}(X) = \frac{rs}{(r+s)^2 \ (1+r+s)}.$$

9. Chi-Square (χ^2) distribution.

   The χ^2 distribution with m degrees of freedom is the Gamma distribution with parameters $(m/2, 1/2)$. Denoted by $\chi^2(m)$. In particular,

$$X \sim \mathcal{N}(0,1) \quad \Rightarrow \quad X^2 \sim \chi^2(1),$$

$$X_j \sim \mathcal{N}(0,1),\ j = 1, \ldots, m,\ \text{i.i.d.} \quad \Rightarrow \quad \sum_{j=1}^{m} X_j^2 \sim \chi^2(m),$$

10. Student distribution.

    If $Z \sim \mathcal{N}(0,1)$, $Y \sim \chi^2(m)$, $Z \perp Y$, then,

$$T := \frac{Z}{\sqrt{Y/m}},$$

   has a student distribution with m degrees of freedom.

   Its density is given by

$$f_T(t) = \frac{\Gamma((m+1)/2)}{\sqrt{m\pi}\ \Gamma(m/2)} \left(1 + \frac{t^2}{m}\right)^{-(m+1)/2}, \quad t \in \mathbb{R}.$$

11. Studentizing. Let $\{X_i\}_{i=1}^n$ be i.i.d. with $\mathcal{N}(\mu, \sigma^2)$ distribution. Let $\overline{X}_n := \sum_{i=1}^n X_i/n$ and set

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

   Then, $\overline{X}_n$ and S_n^2 are independent and

$$\frac{\sqrt{n}\,\left[\overline{X}_n - \mu\right]}{S_n}$$

   has a Student distribution with $n-1$ degrees of freedom.

Bibliography

Anirban DasGupta. *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*. Springer Science & Business Media, 2011.

Piet Groeneboom and Jon A. Wellner. *Information Bounds and Nonparametric Maximum Likelihood Estimation*, volume 19. Springer Science & Business Media, 1992.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.