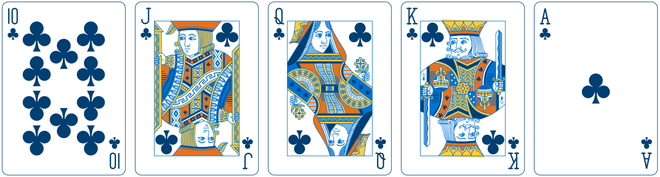
2P7: Probability & Statistics

Decision, Estimation and Hypothesis Testing

Thierry Savin

Lent 2024



the *royal flush*, the best possible hand in poker, has a probability 0.000154%

Introduction

Course's contents

UNIVERSITY OF CAMBRIDGE
Department of Engineering

1. Probability Fundamentals
2. Discrete Probability Distributions
3. Continuous Random Variables
4. Manipulating and Combining Distributions
5. Decision, Estimation and Hypothesis Testing

Introduction

This lecture's contents

Introduction

Decision and estimation theory

Hypothesis testing

Course survey

Introduction

Decision, Estimation and Hypothesis Testing

In the last lectures, we have seen:

- that discrete random variables are described by their probability mass function
- that continuous random variables are described by their probability density function
- how to manipulate random variables and their distributions.
- the central limit theorem.

In this lecture, we will

- learn how to make decisions or estimates on random variables using their measurements
- explain how to assess the statistical significance of an experiment using hypothesis testing.

Unfortunately, we will only scratch the surface of these vast subjects. . .

Decision and estimation theory

Problem statement & definitions

- Suppose that the PDF (or PMF) of a random variable X is a function $f_{\mathrm{X}}(x;\theta)$ (or $P_{\mathrm{X}}(x;\theta)$) that depends on a parameter θ.
- We wish to find the value of θ from observations of X:
  - If θ can adopt a continuous set of values, we want to best estimate its value given the observations.
  - If θ can only take a value from a discrete set, we want to decide which value is the correct one given the observations.
- We define an observation as a measurement of X. We usually have n observations $\boldsymbol{x} = [x_1, \ldots, x_n]^{\mathsf{T}}$.
- We define the sample as the set of random variables underlying the observations $\mathbf{X} = [\mathrm{X}_1, \ldots, \mathrm{X}_n]^{\mathsf{T}}$.
  Usually, these are independent and identically distributed (i.i.d.) with $f_{\mathrm{X}_i}(x;\theta) = f_{\mathrm{X}}(x;\theta)$ (or $P_{\mathrm{X}_i}(x;\theta) = P_{\mathrm{X}}(x;\theta)$) for all $i \in \{1, \ldots, n\}$.
- The estimate (or decision) for θ is a function of the observations, $\hat{\theta}(\boldsymbol{x})$.
- The estimator (or decision rule) is the corresponding *random variable* $\hat{\Theta} = \hat{\theta}(\mathbf{X})$.

The goal is to find an expression for $\hat{\theta}$.

Decision and estimation theory

Bayesian approach

In Bayesian statistics:

- The unknown parameter θ is viewed as the value of a random variable Θ;
- The distribution of the sample is then interpreted as the conditional distribution $f_{\mathbf{X}|\Theta}(\boldsymbol{x}|\theta)$ (or $P_{\mathbf{X}|\Theta}(\boldsymbol{x}|\theta)$);
- The *prior* information is used to assign *somehow* a PDF $f_{\Theta}(\theta)$ (or PMF $P_{\Theta}(\theta)$) to the random variable Θ.
- The problem of estimating (or deciding) the unknown parameter θ is thus changed to the problem of predicting the value θ of the random variable Θ.
- We define:
  - the prior function $f_{\Theta}(\theta)$ or $P_{\Theta}(\theta)$ (prior to the measurements)
  - the likelihood function $f_{\mathbf{X}|\Theta}(\boldsymbol{x}|\theta)$ or $P_{\mathbf{X}|\Theta}(\boldsymbol{x}|\theta)$
  - the posterior function $f_{\Theta|\mathbf{X}}(\theta|\boldsymbol{x})$ or $P_{\Theta|\mathbf{X}}(\theta|\boldsymbol{x})$ (after the measurements)

Decision and estimation theory

Common estimators

- Maximum likelihood estimator (ML)

$$\hat{\theta}_{\mathsf{ML}}(\boldsymbol{x}) = \arg\max_{\theta} f_{\mathbf{X}|\Theta}(\boldsymbol{x}|\theta)$$

- Maximum a posteriori estimator (MAP)

$$\hat{\theta}_{\mathsf{MAP}}(\boldsymbol{x}) = \arg\max_{\theta} f_{\Theta|\mathbf{X}}(\theta|\boldsymbol{x})$$

- Minimum mean squared error (MMSE)

$$\hat{\theta}_{\mathsf{MMSE}}(\boldsymbol{x}) = \mathbb{E}[\Theta|\mathbf{X} = \boldsymbol{x}] = \int \theta \, f_{\Theta|\mathbf{X}}(\theta|\boldsymbol{x}) \mathrm{d}\theta$$

  minimises the mean squared error $\mathbb{E}[(\theta - \Theta)^2|\mathbf{X} = \boldsymbol{x}]$.

Note that

- The posterior function is obtained from Bayes' rule:

$$f_{\Theta|\mathbf{X}}(\theta|\boldsymbol{x}) = \frac{f_{\mathbf{X}|\Theta}(\boldsymbol{x}|\theta) f_{\Theta}(\theta) \quad \longleftarrow \text{only this needs to be maximised for } \hat{\theta}_{\mathsf{MAP}}}{f_{\mathbf{X}}(\boldsymbol{x}) \quad \longleftarrow \text{doesn't depend on } \theta}$$

- If $f_{\Theta}(\theta) =$ constant (i.e. Θ uniformly distributed), then $\hat{\theta}_{\mathsf{ML}} = \hat{\theta}_{\mathsf{MAP}}$.

Decision and estimation theory

Examples

We draw n observations $\{x_1, \ldots, x_n\}$ of the random variable $\mathrm{X} \sim \mathcal{N}(\theta, 1)$ from an i.i.d. sample $\{\mathrm{X}_1, \ldots, \mathrm{X}_n\}$. What is the ML estimator of θ?

The likelihood function is[1]:

$$f_{\mathbf{X}|\Theta}(\boldsymbol{x}|\theta) = f_{\mathrm{X}_1|\Theta}(x_1|\theta) \times f_{\mathrm{X}_2|\Theta}(x_2|\theta) \times \cdots \times f_{\mathrm{X}_n|\Theta}(x_n|\theta) = \frac{e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}}{(2\pi)^{n/2}}$$

Maximising $f_{\mathbf{X}|\Theta}(\boldsymbol{x}|\theta)$ is minimising $\sum_{i=1}^{n}(x_i - \theta)^2$:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\sum_{i=1}^{n}(x_i - \theta)^2 = -2\sum_{i=1}^{n}(x_i - \theta) = 2n\theta - 2\sum_{i=1}^{n} x_i = 0$$

is solved for $\hat{\theta}_{\mathrm{ML}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{\boldsymbol{x}}$, the sample mean. □

[1] Why can we write $f_{\mathrm{X}_1 \ldots \mathrm{X}_n|\Theta} = f_{\mathrm{X}_1|\Theta} \times f_{\mathrm{X}_2|\Theta} \times \ldots \times f_{\mathrm{X}_n|\Theta}$? We haven't seen conditional independence, and it may look a bit peculiar. First note, that $f_{\mathrm{X}_1 \ldots \mathrm{X}_n\Theta} = f_{\mathrm{X}_2 \ldots \mathrm{X}_n|\mathrm{X}_1\Theta} \times f_{\mathrm{X}_1\Theta}$. Since $\mathrm{X}_2, \ldots, \mathrm{X}_n$ are *conditionally independent* of X_1, $f_{\mathrm{X}_2 \ldots \mathrm{X}_n|\mathrm{X}_1\Theta} = f_{\mathrm{X}_2 \ldots \mathrm{X}_n|\Theta} = f_{\mathrm{X}_2 \ldots \mathrm{X}_n\Theta}/f_{\Theta}$ and $f_{\mathrm{X}_1 \ldots \mathrm{X}_n\Theta} = f_{\mathrm{X}_2 \ldots \mathrm{X}_n\Theta} \times f_{\mathrm{X}_1\Theta}/f_{\Theta} = f_{\mathrm{X}_2 \ldots \mathrm{X}_n\Theta} \times f_{\mathrm{X}_1|\Theta}$. So we've proved $f_{\mathrm{X}_1 \ldots \mathrm{X}_n\Theta} = f_{\mathrm{X}_1|\Theta} \times f_{\mathrm{X}_2 \ldots \mathrm{X}_n\Theta}$. The same way, $f_{\mathrm{X}_2 \ldots \mathrm{X}_n\Theta} = f_{\mathrm{X}_2|\Theta} \times f_{\mathrm{X}_3 \ldots \mathrm{X}_n\Theta}$, and so on. We get $f_{\mathrm{X}_1 \ldots \mathrm{X}_n\Theta} = f_{\mathrm{X}_1|\Theta} \times f_{\mathrm{X}_2|\Theta} \times \ldots \times f_{\mathrm{X}_{n-1}|\Theta} \times f_{\mathrm{X}_n\Theta}$ and dividing by f_{Θ} on both sides gives the result.

Decision and estimation theory

Examples

We draw n observations $\{x_1, \dots, x_n\}$ of the random variable $\mathrm{X} \sim \mathcal{N}(\theta, 1)$ from an i.i.d. sample $\{\mathrm{X}_1, \dots, \mathrm{X}_n\}$, with the *belief* that $\Theta \sim \mathcal{N}(\vartheta, \sigma^2)$. What is the MAP estimator of θ?

The posterior function is:

$$f_{\Theta|\mathbf{X}}(\theta|\boldsymbol{x}) \propto f_{\mathbf{X}|\Theta}(\boldsymbol{x}|\theta) f_{\Theta}(\theta) \propto \frac{e^{-\frac{1}{2}\left[\left(\frac{\theta-\vartheta}{\sigma}\right)^2 + \sum_{i=1}^n (x_i-\theta)^2\right]}}{(2\pi)^{n/2}}$$

Maximising $f_{\Theta|\mathbf{X}}(\theta|\boldsymbol{x})$ is minimising $\left(\frac{\theta-\vartheta}{\sigma}\right)^2 + \sum_{i=1}^n (x_i - \theta)^2$:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\left(\frac{\theta-\vartheta}{\sigma}\right)^2 + \sum_{i=1}^n (x_i - \theta)^2\right] = 2\frac{\theta-\vartheta}{\sigma^2} + 2n(\theta - \bar{\boldsymbol{x}}) = 0$$

is solved for $\hat{\theta}_{\mathrm{MAP}}(\boldsymbol{x}) = \dfrac{\vartheta + n\sigma^2 \bar{\boldsymbol{x}}}{1 + n\sigma^2}$. □

- We verify $\hat{\theta}_{\mathrm{MAP}}(\boldsymbol{x}) \xrightarrow{\sigma\to\infty} \hat{\theta}_{\mathrm{ML}}(\boldsymbol{x}) = \bar{\boldsymbol{x}}$ as the prior is "flattened" (hence uniform) when $\sigma \to \infty$.
- Here the posterior is Gaussian, for which maximum and mean coincide, and $\hat{\theta}_{\mathrm{MMSE}}(\boldsymbol{x}) = \hat{\theta}_{\mathrm{MAP}}(\boldsymbol{x})$.

We draw n observations $\{x_1, \ldots, x_n\}$ of the random variable $\mathrm{X} \sim \mathrm{U}(0, \theta)$ from an i.i.d. sample $\{\mathrm{X}_1, \ldots, \mathrm{X}_n\}$. That is, the common PDF is $f_{\mathrm{X}_i}(x_i) = \begin{cases} \theta^{-1} & \text{if } 0 \leq x_i \leq \theta, \\ 0 & \text{otherwise;} \end{cases}$, for all $i = 1 \ldots n$.

What is the ML estimator of θ?

The likelihood function in this case is given by

$$\begin{aligned} f_{\mathrm{X}|\Theta}(\boldsymbol{x}|\theta) &= \frac{1}{\theta^n} \quad 0 \leq x_i \leq \theta \text{ for } i = 1 \ldots n \\ &= \frac{1}{\theta^n} \quad 0 \leq \max\{x_1, \ldots, x_n\} \leq \theta \end{aligned}$$

It is maximised by the minimum value of θ and since $\theta \geq \max\{x_1, \ldots, x_n\}$, we get $\hat{\theta}_{\mathrm{ML}}(\boldsymbol{x}) = \max\{x_1, \ldots, x_n\}$. □

Hypothesis testing

Principle

Hypothesis testing is designed to tell us if the observations indicate that something "unusual" or "interesting" happened with an experiment.

- We are concerned with the statistical significance of an effect in our data, against the hypothesis that no such effect is present. The latter hypothesis is called the null hypothesis and is written $\mathcal{H}_0$.
- In other words, a result has statistical significance when it is very unlikely to have occurred given the null hypothesis.
- More precisely:
  - We define the p-value of a result as the probability p of obtaining a result *at least as extreme*, given $\mathcal{H}_0$ is true.
  - Choose a significance level, denoted by α (typically 1% or 5%). If $p \leq \alpha$, the result is unlikely to happen under $\mathcal{H}_0$: we reject $\mathcal{H}_0$ and the result is said to be statistically significant.

This is probabilistic *reductio ad absurdum*.

Hypothesis testing

The null hypothesis $\mathcal{H}_0$ - Examples

The null hypothesis $\mathcal{H}_0$ is typically a statement of "no effect" or "no difference".

Question	Null hypothesis $\mathcal{H}_0$
Are boys taller than girls at age eight?	"they are the same average height"
Do teens use restaurant locator apps more than adults?	"they use these apps the same amount"
Does eating an apple a day reduce visits to the doctor?	"apples do not reduce doctor visits"
Are small states more densely populated than large states?	"small states have the same population density as large states"
Does the size of a state affect population density?	"all states have the same population density"
Do large dogs prefer large food kibbles?	"large dogs have no preference for kibble size"
Do cats prefer fish or milk?	"cats have no preference; they like them the same"

Hypothesis testing

Example

Can people, in a blind test, taste the difference between single malt and blended whisky? We get 10 people to blind test, each given a randomly selected drink:

- We observe that 7 people correctly identify their drink, and 3 respond incorrectly. What do we conclude?

Null hypothesis $\mathcal{H}_0$: "people can't taste the difference" (i.e. they give a random response).

Under $\mathcal{H}_0$, the number X of correct people is binomial:

$$\mathrm{X}|\mathcal{H}_0 \sim \mathrm{B}(10, {}^1\!/_2)\,.$$

Assuming the null hypothesis is true, what is the probability of the observed outcome, or something more extreme?

$$\sum_{k=7}^{10} {}^{10}\mathrm{C}_k \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{10-k} \approx 0.172 = 17.2\%$$

So, under the hypothesis that people respond randomly, this outcome or something more extreme would happen in about 17% of cases. This doesn't constitute strong evidence against the hypothesis.

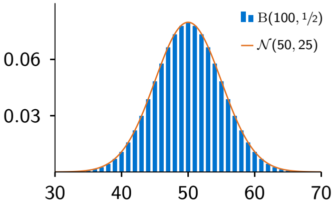
Hypothesis testing

Example

- We repeat the experiment with 100 people, and get 70 correct, 30 incorrect. What do we now conclude?

Under $\mathcal{H}_0$, the probability of the observed outcome, or something more extreme, is now $\sum_{k=70}^{100} {}^{100}\mathrm{C}_k \left(\frac{1}{2}\right)^k \left(1-\frac{1}{2}\right)^{100-k}$.

To calculate this, remember that for n independent Bernoulli trials $\{\mathrm{X}_i \sim \mathrm{Ber}(p)\}_{i=1\ldots n}$ we have $\sum_{i=1}^{n} \mathrm{X}_i \sim \mathrm{B}(n,p)$. Invoking the central limit theorem, $\mathrm{B}(n,p) \overset{n \gg 1}{\approx} \mathcal{N}(np, np(1-p))$ and $\mathrm{B}(100, {}^1\!/\!_2) \approx \mathcal{N}(50, 25)$. Hence the p-value:



$$\sum_{k=70}^{100} {}^{100}\mathrm{C}_k \left(\frac{1}{2}\right)^k \left(1-\frac{1}{2}\right)^{100-k} \approx 1-\Phi\left(\frac{70-50}{\sqrt{25}}\right) = 1-\Phi(4) < 10^{-4} \qquad \square$$

Under $\mathcal{H}_0$, this outcome is exceedingly unlikely and we can thus reject $\mathcal{H}_0$ (i.e. people can actually taste the difference).

Hypothesis testing

One-sided vs two-sided

We want to establish whether a coin is fair. Out of 10 tosses, we get 7 heads and 3 tails. What is the evidence against the null hypothesis that the coin is fair?

Under $\mathcal{H}_0$, the probability of the observed outcome, or something more extreme, is now:

$$\sum_{k=7}^{10} {}^{10}\mathrm{C}_k \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{10-k} + \sum_{k=0}^{3} {}^{10}\mathrm{C}_k \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{10-k} = 34.4\%$$

where we account both tails of the distribution. Observing 3 heads (or less) is also as extreme (or more) as the observed outcome.

- Note how the coin example is different from the whisky tasting example. The tasting example is asymmetric: we test whether people are better than chance at distinguishing two types of whisky, not whether their ability is different from chance.
- Figuring out whether one-sided or two-sided tests are appropriate may require some attention...

Hypothesis testing

Remarks about the *p*-value

- The *p*-value is the probability of observing an effect "at least as extreme as" the one in your sample data, under the assumption that no such effect is present in the population your observations are drawn from.

  Why "a result at least as extreme"? Using only the particular observed outcome may lead to small probabilities.

  Consider the whisky blind test with 500 people being correct out of 1000. Then $\mathbb{P}[500|\mathcal{H}_0] = \frac{{}^{1000}\mathrm{C}_{500}}{2^{1000}} = 0.025 < 5\%$, and $\mathcal{H}_0$ would be rejected...

- Hypothesis testing is used *a lot*. One of the most common misuses of the *p*-value is to assign it to the probability of the null hypothesis. This is wrong, as we know:

  $$\mathbb{P}[\text{results}|\mathcal{H}_0] \neq \mathbb{P}[\mathcal{H}_0|\text{results}] \neq \mathbb{P}[\mathcal{H}_0]$$

  This mistake has led to some dramatic miscarriages of justice...

- The alternative, sometimes written $\mathcal{H}_1 = \mathcal{H}_0^{\complement}$ does not carry a lot of information.

You can attempt all problems of Examples Paper 6