Mathematical Statistics

Anna Janicka

Lecture III, 10.03.2022

INTRODUCTION TO MATHEMATICAL STATISTICS

Plan for today

1. Introduction to Mathematical Statistics
   - the statistical model
2. Statistics and their distributions
   - the normal model

MATHEMATICAL STATISTICS

Assumptions

Empirical data reflect the functioning of a random mechanism

Therefore: we are dealing with random variables defined over some probabilistic space; the realizations of these random variables are the collected data.

Problem: we do not know the distribution of these random variables...

Difference between Probability Calculus and Mathematical Statistics

1. PC, example:
   - Phrasing: in a production process each produced unit may be defective. This happens with probability 10%. The defects of different units are independent.
   - Problems: What is the chance that in a batch of 50 items, exactly 6 will be defective? What is the average number of defective elements? What is the most probable number of defective elements?
   - Solution: we build a probabilistic model. Here: a Bernoulli Scheme with n=50, p=0,1.

     Alternatively, if we are interested in questions dealing with order (e.g. what is the chance that the first 5 items are defective?): a different model

Difference between Probability Calculus and Mathematical Statistics – cont.

2. MS, example:

- Formulation: An inspector verified a batch of 50 items, with the following results (1– item defective, 0 – OK):

  0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0

  0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1
- Problems: what is the probability that an item is defective (assessment)? Is the producer's declaration that defectiveness is equal to 10% credible?
- Solution: we build a statistical model, i.e. a probabilistic model with unknown distribution parameter(s).

Statistical Model

in PC: $(\Omega, \mathcal{F}, \mathcal{P})$

Statistical Model: $(\mathcal{X}, \mathcal{F}_{\mathcal{X}}, \mathcal{P})$

where:

$\mathcal{X}$ – the space of values for the observed random variable X (often n-dimensional, if we have an n-dimensional sample $X_1, \ldots, X_n$)

$\mathcal{F}_{\mathcal{X}}$ – σ-algebra on $\mathcal{X}$

$\mathcal{P}$ – a family of probability distributions P_θ, indexed by a parameter $\theta \in \Theta$

In a less formal setting we usually provide: **$\underline{\mathcal{X}, \mathcal{P}, \Theta}$**

Statistical model – example

$\mathcal{X} = \{0,1\}^n$ – sample space

Joint probability distribution:

$$P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

for $\theta \in [0,1]$

(we have n=50, $X_2 = X_{10} = X_{15} = X_{32} = X_{42} = X_{50} = 1$, other $X_i = 0$)

Statistical model – example cont.

Alternative formulation (if we only record the *number* of defective items in a sample):

$\mathcal{X} = \{0,1, 2, ..., n\}$ – sample space

Joint probability distribution:

$$P_\theta(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

for $\theta \in [0,1]$

(we have n=50 and X=6)

Statistical model – example cont. (2)
Possible questions

Based on the sample:

- What is the value of θ?
  - we are interested in a precise value
  - we are interested in an interval (*confidence*)

  → *estimation*

- Verification of the hypothesis that $\theta = 0.1$

  → *hypothesis testing*

- → *predictions*

Statistical Model: example 2
Growths on the market

An analyst studies the length of periods of growth on the stock market. She is interested in times of growth (until the first fall), in days. Assume the times of growth, $X_1, X_2, \ldots, X_n$ are a sample from an exponential distribution Exp(λ), where

$\lambda>0$ is an unknown parameter. We have:

$\mathcal{X}=(0,\infty)^n$ – sample space

Joint probability distribution:

$$P_\lambda(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) = \prod_{i=1}^{n}(1-e^{-\lambda x_i})$$

$$f_\lambda(x_1, x_2, \ldots, x_n) = \lambda^n e^{-\lambda \Sigma x_i}$$

for $\lambda > 0$

Statistical Model: example 3
Measurements with error

We repeat measuring μ, the results of measurements are independent random variables $X_1, X_2, \ldots, X_n$, (our machine is not perfect). Each measurement is normally distributed $N(\mu, \sigma^2)$, where μ, σ^2 – unknown parameters (so $\theta = (\mu, \sigma)$). Then:

$\mathcal{X} = \mathbb{R}^n$ – sample space

Joint probability distribution:

$$P_{\mu,\sigma}(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) = \prod_{i=1}^{n} \Phi\left(\frac{x_i - \mu}{\sigma}\right) \text{ or}$$

$$f_{\mu,\sigma}(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

for $\mu \in \mathbb{R}$, $\sigma > 0$

STATISTICS

(objects)

Statistics

Parameter estimation (both point and interval) as well as hypothesis testing are conducted based on *statistics*

Statistic = a function of observations, i.e. any random variable

$$T = T(X_1, X_2, \ldots, X_n)$$

The distribution of a statistic T depends on the distribution of X (and therefore on θ), but the statistic as such cannot depend on parameter θ, *e.g.*

$$\cancel{X_1 + X_2 - \theta}$$

Statistics – examples

$$T_1 = \sum_{i=1}^{n} X_i, \quad T_2 = \tfrac{1}{n}\sum_{i=1}^{n} X_i, \quad T_3 = \tfrac{1}{n}\sum_{i=1}^{n} X_i - 0.1$$

are statistics for a sample size of *n*;

$$T_1 = X, \quad T_2 = \frac{X}{n}, \quad T_3 = \frac{X}{n} - 0.1$$

are statistics for a single observation

The choice of a statistic depends on the question we want to answer.

Distribution of statistics

In many cases statistical models refer to a common set of assumptions $\rightarrow$ similar models are applied.

Similar questions are posed $\rightarrow$ similar statistics are calculated.

The most commonly used is the normal model

The normal model

$X_1, X_2, \ldots, X_n$ are a sample from N(μ, σ^2).

The most important statistics (*in general, not only for this model*):

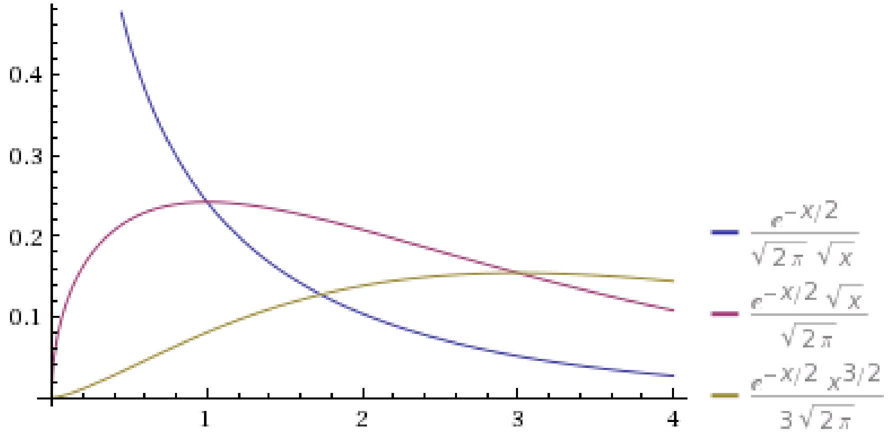
Mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

<u>sample</u> variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$

standard deviation: $S = \sqrt{S^2}$

what are their distributions?

Chi-square Distribution $\chi^2(n)$ – reminder



A special case of the gamma distribution.

The sum of squares of *n* IIN random variables (independent identically N(0,1) distributed) has a $\chi^2(n)$ distribution

$$\mathbb{E}X = n, \quad \mathrm{Var}X = 2n$$

The normal model – cont. (1)

Theorem: In the normal model, the $\bar{X}$ and S^2 statistics are independent random variables such that

$$\bar{X} \sim N(\mu, \sigma^2/n) \qquad \frac{(\bar{X}-\mu)}{\sigma}\sqrt{n} \sim N(0,1)$$

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$$

in particular:

$$E_{\mu,\sigma} S^2 = \sigma^2, \text{and } \mathrm{Var} S^2 = 2\sigma^4/(n-1)$$

The normal model – cont. (2)

In the normal model, *the variable*

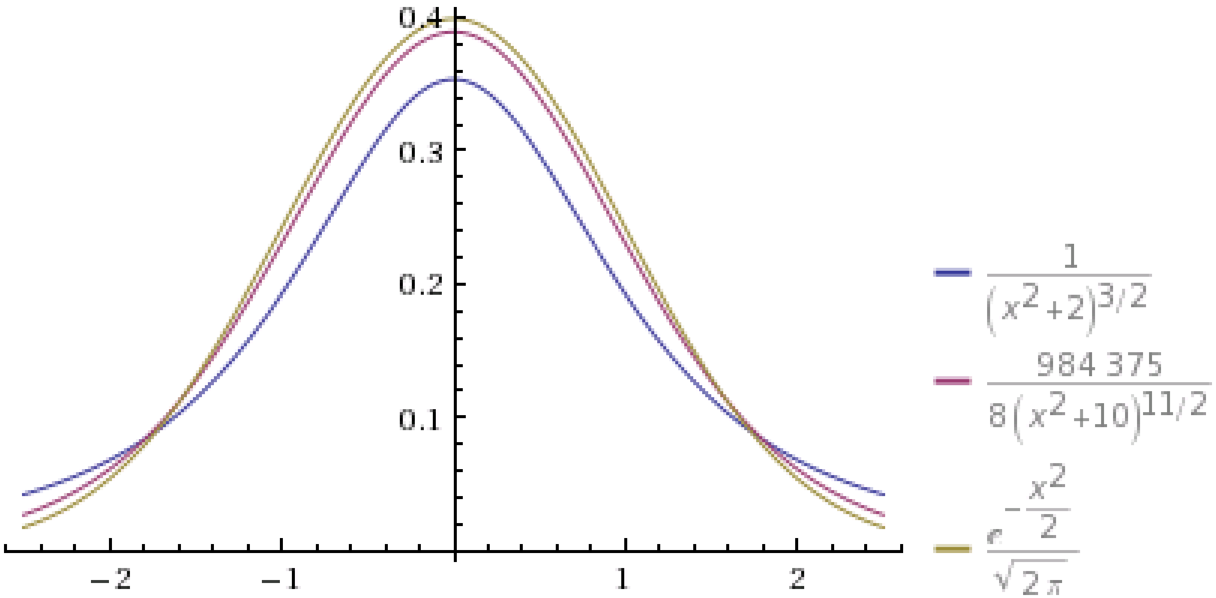
$$T = \sqrt{n}(\overline{X} - \mu) \Big/ S$$

has a t-Student distribution with n -1 degrees of freedom, T ~ t(n -1)

t-Student Distribution $t(n)$, n=1,2,…

defined as the distribution of the random variable

$\sqrt{n}X / \sqrt{Y}$ for independent X *and* Y, $X \sim N(0,1)$, $Y \sim \chi^2(n)$



$\mathbb{E}X = 0 \quad n > 1 \qquad \mathrm{Var}X = n/(n-2) \quad n > 2$