Probability and Statistics course

Introduction

Gabor Vigh

TTK Department of Probability Theory and Statistics
ELTE

November 1, 2025

Week 8

Retake of midterm and final exam

15th of Dec 8:00 - 9:30: same room - 0-804 Lóczy Lajos class room

Materials

Baron, M. (2014).
Probability and statistics for computer scientists.
CRC Press.

Agenda for rest of the semester

[Baron, 2014]

1. Descriptive statistics (3rd of Nov)
2. The basics of statistics: estimation (maximum likelihood, moments) (10th of Nov)
3. Confidence intervals (17th of Nov)
4. Parametric and non-parametric probes (24th of Nov)
5. Linear regression (1st of Dec)
6. Final Exam (8th of Dec)

Introduction to Statistics

1. So far we focused on finding probabilities, expectations, and other characteristics for a variety of situations
2. Ultimately, we needed to know the distribution and the parameters of the (fitted) distribution had to be reported to us explicitly, or they had to follow directly from the problem. This, however, is rarely the case in practice.
3. Then, how can one apply the knowledge that we learned compute probabilities?
4. Answer is simple: we need to collect data. A properly collected sample of data can provide rather sufficient information about parameters of the observed system. In the next sections and chapters, we learn how to use this sample
   - to visualize data, understand the patterns, and make quick statements about the system's behavior;
   - to characterize this behavior in simple terms and quantities;
   - to estimate the distribution parameters;
   - to assess reliability of our estimates;
   - to test statements about parameters and the entire system;
   - to understand relations among variables;
   - to fit suitable models and use them to make forecasts.

Descriptive Statistics

Descriptive statistics deals with the observed data, not taking into account the randomness. In **mathematical statistics** our starting point is the **sample**, which is a sequence $X_1, \ldots, X_n$ of independent, identically distributed random variables. We do not know the distribution of the sample elements. In many cases it is determined by a (real) parameter ϑ. Some statistics of the sample:

- Sample Mean or Average: $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- Empirical Standard Deviation: $S_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2}$

  (The square root of the average squared deviation from the mean)
- Corrected Empirical Standard Deviation: $S_n^* = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}$

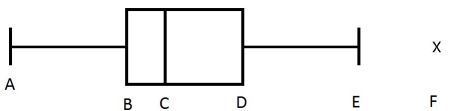
  (Often called the Sample Standard Deviation)
- Coefficient of Variation (or Relative Standard Deviation): $V = \frac{S_n}{\overline{X}} = \frac{S_n}{\overline{X}} \cdot 100\%$ (The average deviation from the mean expressed as a percentage)

Descriptive statistics

- k-th Empirical Moment ($k \geq 1, k \in \mathbb{Z}$): $m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$
- Empirical Mode: the most frequently occurring value
- Ordered Sample: $X_1^* \leq ... \leq X_n^*$ - sample elements in non-decreasing order
- Empirical Median: $X^*_{\frac{n+1}{2}}$, if n is odd, and $\frac{X^*_{\frac{n}{2}} + X^*_{\frac{n}{2}+1}}{2}$, if n is even.
- Range: $R = X_n^* - X_1^*$ (largest $-$ smallest sample element)
- Range: $R = X_n^* - X_1^*$ (largest $-$ smallest sample element)
- z-Quantile: $q_z = \inf\{x : F(x) \geq z\}$. If F is invertible, $q_z = F^{-1}(z)$.
- Empirical z-Quantile (Interpolation Method): First determine the rank: $(n+1)z = e + t$ (e: integer, t: fractional part), then $q_z = X_e^* + t(X_{e+1}^* - X_e^*)$.
- Quartiles (Special quantiles):
  - Lower (or first) quartile: $Q_1 = q_{\frac{1}{4}}$
  - Median: $Q_2 = q_{\frac{1}{2}}$
  - Upper (or third) quartile: $Q_3 = q_{\frac{3}{4}}$
- Interquartile Range: $IQR = q_{\frac{3}{4}} - q_{\frac{1}{4}} = Q_3 - Q_1$

Descriptive statistics

An important diagram: Boxplot, where the letters mean the following:



- $A = \max\{x_1^*, Q_1 - 1,5 \cdot IQR\}$; $B = Q_1$; $C = Me$; $D = Q_3$;
- $E = \min\{x_n^*, Q_3 + 1,5 \cdot IQR\}$; F: outliers that fall outside A or E.

Glivenko-Cantelli

Empirical Distribution Function (EDF): $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i < x)$, where

$I(X_i < x) = \begin{cases} 1 & \text{if } X_i < x \\ 0 & \text{if } X_i \geq x \end{cases}$ is the indicator function.

Theorem (Glivenko-Cantelli)

The maximum deviation between the Empirical Distribution Function ($F_n(x)$) and the Theoretical Distribution Function ($F(x)$) converges to 0 with probability 1. This means that for a sufficiently large sample, the value of $F_n(x)$ is arbitrarily close to the value of $F(x)$ for all x and remains in its vicinity as n increases.

Unbiased Estimator

The statistics $T(X_1, \ldots, X_n)$ is an **unbiased estimate** of function $g(\vartheta)$ of the unknown parameter ϑ if

$$\mathbb{E}_\vartheta(T(X_1, \ldots, X_n)) = g(\vartheta)$$

holds for all possible values ϑ. That is, the expectation of our estimate is equal to the quantity that we want to estimate, for all possible values of the unknown parameter.
Example:

$$\mathbb{E}(\overline{X}) = \mathbb{E}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \mathbb{E}(X_1),$$

that is, the mean is an unbiased estimate of the expectation.
Definition of the empirical standard deviation (`sd` in R):

$$s_n^* = \sqrt{\frac{1}{n-1}\left(\sum_{j=1}^{n}(X_j - \overline{X})^2\right)} = \sqrt{\frac{n}{n-1}\left(\left(\frac{1}{n}\sum_{j=1}^{n}X_j^2\right) - \overline{X}^2\right)}$$

One can prove that

$$\mathbb{E}(s_n^{*2}) = Var(X_1),$$

that is, the empirical variance is an unbiased estimator of the variance (however, in general, empirical standard deviation is not an unbiased estimator of the standard deviation).

Consitent Estimator

The sequence of estimators (T_n) is **consistent** for the function $g(\vartheta)$ of the unknown parameter ϑ if

$$T_n(X_1, \ldots, X_n) \to g(\vartheta)$$

holds with probability 1 for all possible values of ϑ. That is, the sequence of estimators converges with probability 1 to the value that we want to estimate, for all possible values of the unknown parameter. For example, by the strong law of large numbers, we have that

$$\frac{X_1 + \ldots + X_n}{n} \to \mathbb{E}(X_1)$$

holds with probability 1 if $X_1, X_2, \ldots$ are independent, identically distributed, and have finite expectation. That is, the mean is a consistent estimator of the expectation.
One can prove that

$$s_n^* \to s.d.(X_1) \quad \text{and} \quad s_n^{*2} \to Var(X_1)$$

holds with probability 1. Hence the empirical standard deviation is consistent for the standard deviation, and empirical variance is consistent for the variance.

Examples

1. Let $X_1, \ldots, X_n$ be independent, identically distributed random variables with mean m. Our goal is to estimate the unknown parameter m. Consider the following statistics and determine which of them are unbiased! If an estimator is not unbiased, how could we make it unbiased?
   $$T_1(\mathbf{X}) = X_8, \qquad T_2(\mathbf{X}) = \frac{X_9 + X_{19}}{9}, \qquad T_3(\mathbf{X}) = \overline{X}$$
2. Give an unbiased estimator for the parameter ϑ of the independent, identically $E[0, \vartheta]$ distributed sample $X_1, \ldots, X_n$, using the sample mean!
3. Let $X_1, X_2, \ldots, X_n$ be a sample from exponential distribution with unknown parameter $\lambda > 0$.
   (a) Give an unbiased estimate for $e^{-3\lambda}$ and then for $\frac{1}{\lambda}$.
   (b) Give a consistent estimate for λ and λ^2.
   (c) Randomize a sample of size $n = 10000$ with some arbitrarily chosen parameters λ, and calculate the estimates. Compare the values with the quantities that we wanted to estimate.