Teaching Mathematics and Statistics in Sciences - IPA HU-SRB/0901/221/088
Interesting Mathematical Problems in Sciences and Everyday Life - 2011

Probability and statistical models appearing in teaching mathematics to science teachers and university students

Danijela Rajter-Ćirić
Department of Mathematics and Informatics, University of Novi Sad, Serbia
Trg Dositeja Obradovića 4, Novi Sad, Serbia
danijela.rajter-ciric@dmi.uns.ac.rs
http://www.dmi.uns.ac.rs

Probability theory, stochastic analysis and statistics are the topics one cannot avoid if wants to understand a lot of phenomena in the nature and society. Therefore, not only mathematicians are interested in using the methods of these three, but also the people who work in physics, chemistry, economy, pharmacy, agriculture, psychology, etc. If one takes, for example, the stochastic analysis, then it is really a challenge to teach the complicated theory on a level which can be understandable for someone who has some knowledge in mathematics (although mathematics is not her or his main field of interest) but who, at the end, should use the methods of the stochastic analysis in the work. For instance, the theory of the Markov processes, Poisson processes or Brownian motion are very important in many fields of applications but the corresponding mathematical theory is not easy to explain. Here, in the first part of the chapter, we will introduce the powerful methods of the probability theory and stochastic analysis. We will mostly concentrate to the probability theory since it is the basics for the stochastic analysis and it is simpler and easier to understand. The second part of the chapter is devoted to statistics. When it comes to teaching statistics, the situation is a little bit easier since one can teach statistics in many different levels. No matter which level one chooses, the main goal is to teach the students to understand the problem and to choose the statistical tool correctly in order to solve the problem.

1. Introduction

Any realistic model of a real-world phenomenon must take into account the possibility of randomness. Probability is a part of our everyday-lives. A basic understanding of probability makes it possible to understand everything from batting averages to the weather report or someone's chances of being struck by lightning! Probability is an important topic in mathematics because the probability of certain events happening - or not happening - can be important to us in the real world.

On the other hand, statistics is a set of techniques and procedures that are used to analyze what is happening in the world around us. Today we live in the Information Age and much of this information was determined mathematically by using statistics. When used correctly, statistics tell us any trends in what happened in the past and can also be very useful in predicting what may or may not happen in the future.

2. Basics of the probability theory

Probability is simply the chance that something might happen. When we calculate the probability of an event we look at chances of getting what we want versus all the possible things that can happen. The probability of an event that you know for sure will happen is 100% or 1 while the probability of an event that will never happen is 0% or just plain 0. What about other events that we're not so sure about? The probability of these events can be given as a percent or as an odds ratio. The probability of all the events that are possible must add up to 100%.

2.1. Sample space and events

Suppose that we will perform an experiment whose outcome is not predictable in advance and let us suppose that the set of all possible outcomes is known. This set is called the *sample space* and here will be denoted by Ω. For instance, if the experiment consists of rolling a die, the sample space is

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

where the outcome i stands for the case when the number i appear on the die, $i = 1, 2, 3, 4, 5, 6$.

An event A is a subset of the sample space Ω. In the example with rolling a die, the event $A = \{2, 4, 6\}$ is the event that an even number appears on the roll. Specially, $\emptyset$ refers to the event consisting of no outcomes.

For any event A we define the *complement* of A as $\overline{A} = \Omega \setminus A$. It will occur if and only if A does not occur. In the example with rolling a die, if A is the event that an even number appears on the roll then $\overline{A} = \{1, 3, 5\}$.

Since events are sets, all usual set operations are allowed such as, union, intersection, taking a subset,... For instance, for any two events A and B, we may define the intersection of A and B, denoted by $A \cap B$, or simply by AB. It is, of course, the event which will occur if both A and B occur. If $AB = \emptyset$, then A and B are said to be *mutually exclusive*. The union of events A and B, denoted by $A \cup B$, will occur if either A or B occurs. If A and B are mutually exclusive events, that the union of A and B will be denoted by $A + B$. We also define unions and intersections of more than two events in a similar manner.

2.2. Probabilities of events

Suppose that, for each event A of the sample space Ω, a number $P(A)$ is defined and satisfies the following conditions:

(i) $P(\Omega) = 1$.

(ii) $0 \leq P(A) \leq 1$.

(iii) For any family of mutually exclusive events $A_1, A_2, \ldots$ (that means that $A_i A_j = \emptyset$ when $i \neq j$), the following holds:

$$P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

The number $P(A)$ is called the probability of the event A.

Here we will not discuss in detail the conditions under which the probability of an event exists (in order to do this, one should introduce the notion of σ-algebra), we will simply suppose that we are dealing with events for which it is possible to calculate their probabilities.

In the example with rolling a die, if we suppose that all six numbers are equally likely to appear, then we have

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = 1/6.$$

From the property (iii) follows that the probability of getting an even number is

$$P(\{2,4,6\}) = P(\{2\}) + P(\{4\}) + P(\{6\} = 1/2.$$

In the sequel we give some of the properties of the probability P.

(1) $P(\emptyset) = 0$.

Proof

Since $\Omega = \Omega + \emptyset + \emptyset + \dots$, one has that $P(\Omega) = P(\Omega) + P(\emptyset) + P(\emptyset) + \dots$ i.e. $P(\emptyset) = 0$. □

(2) $P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$, for any mutually exclusive events $A_1, \dots A_n$.

(3) If $A \subseteq B$, then $P(A) \leq P(B)$.

Proof

If $A \subseteq B$, then $B = A + \overline{A}B$. Therefore $P(B) = P(A) + P(\overline{A}B)$ implying that $P(A) \leq P(B)$. □

(4) $P(\overline{A}) = 1 - P(A)$.

Proof

Since the events A and $\overline{A}$ are mutually exclusive such that $A + \overline{A} = \Omega$, we have (see property (2) from above):

$$1 = P(\Omega) = P(A + \overline{A}) = P(A) + P(\overline{A}),$$

i.e. $P(\overline{A}) = 1 - P(A)$. □

One can easily prove the following

(5) $P(A \cup B) = P(A) + P(B) - P(AB)$.

In case when A and B are mutually exclusive, one has $P(AB) = \emptyset$ and therefore

$$P(A \cup B) = P(A + B) = P(A) + P(B).$$

2.3. Conditional probability

Suppose that we toss two dice and that all 36 possible outcomes are equally likely to occur (all of them have probability 1/36). Suppose that we observe that number three appears on the first die. Given this information, what is the probability that the sum of the two dice equals seven?

We think in the following way: Knowing that the first die is three, there can be six possible outcomes of our experiment, (3,1), (3,2), (3,3), (3,4), (3,5) and (3,6). Each of these outcomes originally had the same probability of occurring and they should still have equal probabilities (our knowledge about the outcome on the first die does not change this property). Therefore, given the first die is three, the (conditional) probabilities of all outcomes (3,1), (3,2), (3,3), (3,4), (3,5) and (3,6) are the same, 1/6, while the (conditional) probabilities of other 30 points in the sample space is 0. Only one of the outcomes (3,1), (3,2), (3,3), (3,4), (3,5) and (3,6) describes the property that the sum of dice equals seven, this is the outcome (3,4) and it has the probability 1/6.

Denote by A the event that the sum of the dice is seven and by B the event the first die is three. Then the probability that we have just obtained is called the *conditional* probability that A occurs given that B has occurred and it is denoted by $P(A|B)$.

A formula for $P(A|B)$ which is general, i.e. valid for all events A and B, is obtained in the same manner: If the event B occurs, than (since we want that A occurs) it is necessary that AB occurs. But, knowing that B has occurred, B becomes our new sample space and hence the probability that AB occurs will be equal to the probability that AB occurs relative to the probability of B. That is,

$$P(A|B) = \frac{P(AB)}{P(B)}.$$

Note that $P(A|B)$ is well defined only if $P(B) > 0$. If both $P(A) > 0$ and $P(B) > 0$, then from

$$P(A|B) = \frac{P(AB)}{P(B)} \text{ and } P(B|A) = \frac{P(AB)}{P(A)}$$

one obtains

$$P(AB) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A).$$

Example 1 Elena can either take course in computers or in statistics. If she takes the computer course, the probability that she will receive the best grade is 1/2, and if she takes the statistics course she will obtain the best grade with probability 1/3. Elena will decide which to choose by flipping a fair coin. What is the probability that Elena will get the best grade in statistics?

Solution: If we let B be the event that Elena takes the statistics and denote by A the event that she receives the best grade in whatever course she takes, then the desired probability is $P(AB)$ and it calculated as

$$P(AB) = P(B)P(A|B) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}.$$

The previous formula can be generalized (by induction) to

$$P(A_1A_2 \dots A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1A_2) \cdots P(A_n|A_1A_2 \cdots A_{n-1}). \tag{1}$$

2.4. Independent events

Intuitively, two events A and B are independent if the realization of either of this two has no impact on the probability that the other one occurs:

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B).$$

Based on this, one obtains the formal definition: Two events A and B are *independent* if

$$P(AB) = P(A) \cdot P(B).$$

One can extend the definition of independence to more than two events. The events $A_1, A_2, \dots, A_n$ are said to be (mutually) independent if elements of any subset $A_{i_1}, A_{i_2}, \dots, A_{i_k}$, $k \leq n$, of those events satisfy

$$P(A_{i_1}A_{i_2} \dots A_{i_k}) = P(A_{i_1})P(A_{i_2}) \dots P(A_{i_k}).$$

If the events $A_1, A_2, \dots, A_n$ in relation (1) are independent, then

$$P(A_1A_2 \dots A_n) = P(A_1) \cdot P(A_2) \cdots P(A_n).$$

Example 2 Suppose we toss two fair coins. What is the probability that on both coins heads appear?

Solution: To solve the problem, denote by A_1 and A_2 the events that head appears on the first and the second coin, respectively. Than, the desired probability is $P(A_2A_2)$ and, since A_1 and A_2 are independent, can be calculated as

$$P(A_1A_2) = P(A_1)P(A_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

2.5. Bayes' formula

Let A and B be events. Then

$$A = A\Omega = A(B + \overline{B}) = AB + A\overline{B},$$

where Ω is the sample space.

The previous formula states that if a point is in A, then it is either in both A and B or in A but not in B. Then we have

$$P(A)=P(AB)+P(A\overline{B})=P(B)P(A|B)+P(\overline{B})P(A|\overline{B}).$$

Now we have

$$P(B|A)=\frac{P(BA)}{P(A)}=\frac{P(B)P(A|B)}{P(A)}=\frac{P(B)P(A|B)}{P(B)P(A|B)+P(\overline{B})P(A|\overline{B})}.$$

The previous formulas can be generalized: Suppose that $H_1, H_2, \ldots, H_n$ are mutually exclusive events satisfying $\sum\limits_{i=1}^{n} H_i=\Omega$ and $P(H_i)>0$, $i=1,\ldots,n$. (In other words, exactly one of the events $H_1, H_2, \ldots, H_n$ will occur.)

Consider an event A such that $P(A)>0$. Knowing that $A=AH_1+\cdots+AH_n$ one obtains

$$P(A)=\sum_{i=1}^{n} P(AH_i)=\sum_{i=1}^{n} P(H_i)P(A|H_i),$$

and

$$P(H_i|A)=\frac{P(H_i)P(A|H_i)}{\sum_{i=1}^{n} P(H_i)P(A|H_i)},\quad i=1,\ldots,n.$$

The last equation is known as Bayes' formula.

Example 3 Consider two urns. The first urn contains two white and seven black balls and the second one contains five white and six black balls. We select the urn by chance and then draw a ball from the urn. What is the probability that, if a white ball is drawn, it is taken from the first urn?

Solution: Let W be the event that a white ball is drawn. Let H_1 be the event that the first urn is selected and let H_2 be the event that the second urn is selected. The desired probability $P(H_1|W)$ can be calculated as

$$P(H_1|W)=\frac{P(H_1W)}{P(W)}=\frac{P(H_1)P(W|H_1)}{P(W)}=\frac{P(H_1)P(W|H_1)}{P(H_1)P(W|H_1)+P(H_2)P(W|H_2)}.$$

By assumption, we select the urn by chance and therefore $P(H_1)=P(H_2)=1/2$. The probability that a white ball is drawn if the first urn is selected is $P(W|H_1)=2/9$ since there are two white balls out of nine balls in the first urn. Similarly, $P(W|H_2)=5/11$. Now, it is easy to calculate $P(H_1|W)=22/67$.

3. Random variables

It is very often that in performing some experiment we are more interested in some functions of outcome then in outcome itself. For instance, in tossing dice we are interested in the sum of two dice and rarely in the actual outcome. These real valued functions defined on the sample space are *random variables*. We may assign probabilities to certain sets of possible values of the random variable since those values are determined by the outcomes of the experiment.

For instance, suppose that the experiment consists of tossing two fair coins and denote by H the outcome when head appears on whatever coins and by T the outcome when tail appears. If X denote the number of heads appearing, then X is a random variable taking one of the values 0,1,2 with probabilities

$$\begin{aligned}
&P\{X=0\}=P\{(T,T)\}=1/4,\\
&P\{X=1\}=P\{(H,T)\}+P\{(T,H)\}=2/4,\\
&P\{X=2\}=P\{(H,H)\}=1/4.
\end{aligned}$$

Of course, $P\{X=0\}+P\{X=1\}+P\{X=2\}=1$.

Note that the random variable X given above has finite number of values. In case when random variable takes finite or countable number of possible values it is called *discrete*.

The *distribution function* F_X of the random variable X is defined for any real number b, $-\infty < b < \infty$, by

$$F_X(b) = P\{X < b\}.$$

It is nondecreasing and continuous from left function, such that

$$\lim_{b\to-\infty} F_X(b) = F_X(-\infty) = 0, \quad \lim_{b\to\infty} F_X(b) = F_X(\infty) = 1.$$

All probability questions about X can be answered in terms of distribution function. For instance,

$$P(a \leq X < b) = F_X(b) - F_X(a), \quad \text{for all } a, b \in \mathbb{R},\ a < b.$$

3.1. Discrete random variables

A random variable that can take at most a countable number of possible values is said to be discrete. For a discrete random variable X define

$$p(b) = P\{X = b\}$$

and it is called the *probability mass function* of X. If X must take one of the values x_1, x_2,..., then $p(x_i) > 0$, $i = 1, 2, \ldots$ and $p(x) = 0$, for all other values of x. This can be written as

$$X : \left(\begin{array}{ccccc} x_1 & x_2 & \ldots & x_n & \ldots \\ p(x_1) & p(x_2) & \ldots & p(x_n) & \ldots \end{array} \right).$$

It is obvious that

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

since X must take one of the values x_i. The distribution function F of X is

$$F_X(b) = \sum_{all\ x_i < b} p(x_i).$$

Let X be a discrete random variable with probability mass function

$$p(0) = \frac{1}{2}, \quad p(1) = \frac{3}{8}, \quad p(2) = \frac{1}{8}.$$

Then, the distribution function of X is given by

$$F_X(b) = \left\{ \begin{array}{ll} 0, & b \leq 0 \\ 1/2, & 0 < b \leq 1 \\ 7/8, & 1 < b \leq 2 \\ 1, & b > 2 \end{array} \right. .$$

3.1.1. The binomial random variable

Suppose that experiment consists of n independent trials during which an event A occurs with the probability p (hence, it does not occur with probability $q = 1 - p$). If X represents the number of occurring of event A in the n trials, then X is a *binomial* random variable with parameters n and p (we write $X : \mathcal{B}(n;p)$), where $n \in \mathbb{N}$ and $0 < p < 1$. Then

$$X : \left(\begin{array}{cccccc} 0 & 1 & \ldots & k & \ldots & n \\ p(0) & p(1) & \ldots & p(k) & \ldots & p(n) \end{array} \right),$$

where

$$p(i) = \left(\begin{array}{c} n \\ i \end{array} \right) p^i q^{n-i}, \quad i = 0, 1, \ldots, n.$$

Note that,

$$\sum_{i=0}^{n} p(i) = \sum_{i=0}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) p^i q^{n-i} = (p + (1-p))^n = 1.$$

Example 4 Suppose that four fair coins are flipped. If the outcomes are assume to be independent, what is the probability that three heads and one tail are obtained?

Solution: In order to find the solution, we denote by X the number of heads that appear. In that case, X is a binomial random variable with parameters $n = 4$ and $p = 1/2$. Hence,

$$P\{X = 3\} = p(3) = \left(\begin{array}{c} 4 \\ 3 \end{array} \right) \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 = \frac{3}{8}.$$

Example 5 It is known that any item produced by a certain machine will be defective with probability 0.1, independently of any other item. What is the probability that in a sample of three items, at most one will be defective?

Solution: If X is the number of defective items in the sample, then X is a binomial random variable with parameters $n = 3$ and $p = 0.1$. Therefore,

$$P\{X = 0\} + P\{X = 1\} = \left(\begin{array}{c} 3 \\ 0 \end{array} \right) (0.1)^0 (0.9)^3 + \left(\begin{array}{c} 3 \\ 1 \end{array} \right) (0.1)^1 (0.9)^2 = 0.972.$$

3.1.2. The Poisson Random variable

The random variable X is said to be *Poisson* random variable with parameter $\lambda > 0$ (we write $X : \mathcal{P}(\lambda)$), if

$$X : \left(\begin{array}{cccccc} 0 & 1 & \dots & n & \dots \\ p(0) & p(1) & \dots & p(n) & \dots \end{array} \right),$$

where

$$p(i) = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0, 1, \dots, n.$$

Then,

$$\sum_{i=0}^{\infty} p(i) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} e^{\lambda} = 1.$$

One of important properties of Poisson random variable is that it can be used to approximate a binomial random variable when the binomial parameter n is large and p is small. Namely, if $X : \mathcal{B}(n; p)$ and $n \to \infty$ and $p \to 0$, but $np \to const$, then

$$p(k) \to \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \dots,$$

where $p(k)$ is probability mass function for binomial random variable.

Example 6 Suppose that the number of typographical errors on a single page of some book has a Poisson distribution with parameter $\lambda = 1$. What is the probability that there is at least one error on a certain page?

Solution: $\quad P\{X \geq 1\} = 1 - P\{X = 0\} = 1 - e^{-1} \approx 0.633.$

Example 7 If the number of accidents occurring on a highway each day is a Poisson random variable with parameter $\lambda = 3$, what is the probability that no accident occur on a certain day?

Solution: $\quad P\{X = 0\} = e^{-3} \approx 0.05.$

3.1.3. The geometric random variable

Suppose that experiment consists of n independent trials during which an event A occurs with the probability p. If X is the number of trials until the first appearance of A, then X is said to be the *geometric* random variable with parameter $0 < p < 1$ (we write $X : \mathcal{G}(p)$), and

$$X : \begin{pmatrix} 1 & 2 & \dots & n & \dots \\ p(1) & p(2) & \dots & p(n) & \dots \end{pmatrix},$$

where

$$p(i) = p\, q^{i-1}, \quad i = 1, 2, \dots, \quad q = 1 - p.$$

Again, we have

$$\sum_{i=1}^{\infty} p(i) = p \sum_{i=1}^{\infty} q^{i-1} = p \sum_{i=0}^{\infty} q^{i} = p\, \frac{1}{1-q} = \frac{p}{p} = 1.$$

Example 8 Luka goes target shooting and he hits the target with the probability 0.7. If he shoots until he misses the target, what is the probability that he shoots exactly three times?

Solution: If X is the number of shooting, then X is a geometric random variable with $p = 0.3$. Then,

$$P\{X = 3\} = (0.7)^2 \cdot 0.3 = 0.147.$$

3.2. Continuous random variables

As mentioned above, continuous random variables have uncountable set of possible values. More precisely, X is said to be continuous if there exists a nonnegative function $\varphi_X : \mathbb{R} \mapsto \mathbb{R}^+$, such that

$$P\{X \in B\} = \int_B \varphi_X(t)\, dt, \quad \text{for any set } B \text{ of real numbers.}$$

The function φ_X is called the *probability density function* of random variable X, or just the density function of X.

By choosing $B = (-\infty, \infty)$ one conclude that φ_X must satisfy

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} \varphi_X(t)\, dt.$$

By choosing $B = [a, b]$ one obtains

$$P\{a < X < b\} = P\{X \in [a, b]\} = \int_a^b \varphi_X(t)\, dt.$$

If $a = b$ it follows that

$$P\{X = a\} = \int_0^0 \varphi_X(t)\, dt = 0.$$

By choosing $B = (-\infty, b)$, one obtains

$$F_X(b) = \int_{-\infty}^{b} \varphi_X(t)\, dt, \quad \text{for any } b \in \mathbb{R},$$

where F_X is the distribution function of random variable X. Differentiating both sides gives

$$\frac{d}{db} F_X(b) = \varphi_X(b).$$

3.2.1. The uniform random variable

A continuous random variable X is said to be *uniformly distributed* over the interval (a,b), $a,b\in\mathbb{R}$ and $a<b$, (we write $X:\mathcal{U}(a,b)$), if its density function is

$$\varphi_X(x)=\begin{cases}\dfrac{1}{b-a}, & x\in(a,b)\\ 0, & x\notin(a,b)\end{cases}.$$

The distribution function of $X:\mathcal{U}(a,b)$ is

$$F_X(x)=\begin{cases}0, & x\le a\\ \dfrac{x-a}{b-a}, & a<x\le b\\ 1, & x>b\end{cases}.$$

For example, if X is the number randomly selected from the interval (a,b), then $X:\mathcal{U}(a,b)$.

Example 9 We select a number from the interval (0,10) by chance. What is the probability that it is less then 3?

Solution: If X is the number we select, then it is uniformly distributed over the interval (0,10). Therefore,

$$P\{X<3\}=\int_0^3\frac{dx}{10}=\frac{3}{10}.$$

3.2.2. The exponential random variable

A continuous random variable is said to be exponentially distributed with parameter $\lambda>0$, (we write $X:\mathcal{E}(\lambda)$), if its density function is

$$\varphi_X(x)=\begin{cases}\lambda\, e^{-\lambda x}, & x>0\\ 0, & x\le 0\end{cases}.$$

The distribution function of $X:\mathcal{E}(\lambda)$ is

$$F_X(x)=\begin{cases}0, & x\le 0\\ 1-e^{-\lambda x}, & x>0\end{cases}.$$

Example 10 Suppose that the amount of the time (in minutes) one spends in a bank is exponentially distributed with $\lambda=1/10$. What is the probability that a customer will spend more than 15 minutes in the bank?

Solution: If X represents the amount of time that the customer spends in the bank, then the probability is just

$$P\{X>15\}=e^{-15\lambda}=e^{-3/2}\approx 0.22.$$

3.2.3. The normal random variable

A continuous random variable is said to be normally distributed with parameters $m\in\mathbb{R}$ and $\sigma^2>0$, (we write $X:\mathcal{N}(m;\sigma^2)$), if its density function is

$$\varphi_X(x)=\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{(x-m)^2}{2\sigma^2}},\ x\in\mathbb{R}.$$

The distribution function of $X:\mathcal{N}(m;\sigma^2)$ is

$$F_X(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(t-m)^2}{2\sigma^2}}\,dt,\ x\in\mathbb{R}.$$

The important property is that, if X is normally distributed with parameters m and σ^2, then $Y = aX + b$ is also normally distributed but with parameters $am + b$ and $a^2\sigma^2 > 0$.

In case when $m = 0$ and $\sigma^2 = 1$ one obtains so-call standard normal distribution, $\mathcal{N}(0; 1)$ which has many applications, (for instance, in statistics).

4. Jointly distributed random variables

4.1. Joint distribution function

We are often interested in probability statements concerning two or more random variables. To deal with such probabilities, we define, for any two random variables X and Y, the *joint cumulative probability distribution function* of X and Y by

$$F_{XY}(a, b) = P\{X < a,\ Y < b\}, \quad -\infty < a, b < \infty.$$

The distribution of X is obtained from the joint distribution of X and Y as follows:

$$F_X(a) = P\{X < a\} = P\{x < a,\ Y < \infty\} = F_{XY}(a, \infty).$$

Similarly, $F_Y(b) = F_{XY}(\infty, b)$.

4.2. Independent random variables

The random variables X and Y are said to be *independent* if, for all a, b,

$$P\{x < a,\ Y < b\} = P\{x < a\} \cdot P\{Y < b\}.$$

In other word, X and Y are independent if, for all a and b, the events $\{X < a\}$ and $\{X < b\}$ are independent. In the terms of distribution functions, X and Y are independent if

$$F_{XY}(a, b) = F_X(a) \cdot F_Y(b), \qquad \text{for all } a, b.$$

5. Expectation and variance

Sometimes, it is very convenient to know some numerical characteristics of random variables. We study some of them in the sections that follow.

5.1. Definitions and basic properties

In probability theory, the expectation (or expected value, or mathematical expectation, or mean, or the first moment) of a random variable is the weighted average of all possible values that this random variable can take on. The expectation of the random variable X, $E(X)$, is number defined by

- $E(X) = \sum_i x_i p(x_i)$, in case X is discrete with the probability mass function $P\{X = x_i\} = p(x_i)$;
- $E(X) = \int_{-\infty}^{\infty} x\ \varphi_X(x)\ dx$, in case X is continuous with the density function $\varphi_X(x)$,

assuming that the sum and the integral above are absolutely convergent.

Some properties of expectation are

1. If c is a constant, then $E(c) = c$ and $E(cX) = cE(X)$.
2. Let Y be a function of random variable X, that is $Y = g(X)$. Then
   - $E(Y) = E(g(X)) = \sum_i g(x_i)p(x_i)$, in case X is discrete with the probability mass function $p(x_i)$.
   - $E(Y) = E(g(X)) = \int_{-\infty}^{\infty} g(x)\ \varphi_X(x)\ dx$, in case X is continuous with the probability density function $\varphi_X(x)$,

   assuming that the sum and the integral above are absolutely convergent.
3. For any X and Y, $E(X+Y) = E(X) + E(Y)$, assuming that $E(X)$ and $E(Y)$ exist.
4. If X and Y are independent, $E(XY) = E(X)E(Y)$, assuming that $E(X)$ and $E(Y)$ exist.
5. $E(X - E(X)) = 0$, assuming that $E(X)$ exists.

The variance of the random variable X is defined as

$$D(X) = E\left((X - E(X))^2\right).$$

One easily calculates

$$\begin{aligned} D(X) &= E\left((X - E(X))^2\right) = E\left((X^2 - 2E(X)X + E^2(X))\right) \\ &= E(X^2) - 2E(X)E(X) + E^2(X) = E(X^2) - E^2(X). \quad \square \end{aligned}$$

From the definition of the variance it is obvious that $D(X) \geq 0$. One can easily show that $D(X) = 0$ if and only if X is a constant and that $D(cX) = c^2 D(X)$ and $D(X + c) = D(X)$, $c =$ const. If X and Y are independent, then $D(X + Y) = D(X) + D(Y)$.

Since the variance is always nonnegative, one can define $\sigma(X) = \sqrt{D(X)}$ and it is called the *standard deviation* of the random variable X.

For any random variable X we can define

$$X^* = \frac{X - E(X)}{\sqrt{D(X)}}.$$

One can easily show

$$E(X^*) = 0 \text{ and } D(X^*) = 1.$$

5.2. Expectation and variance of some random variables

5.2.1. Binomial distribution $\mathcal{B}(n;p)$

Random variable $X : \mathcal{B}(n;p)$ we write as

$$X = X_1 + \cdots + X_n$$

where $X_1, \ldots, X_n$ are independent random variables:

$$X_i : \left(\begin{array}{cc} 0 & 1 \\ p & q \end{array}\right), \quad i = 1, \ldots, n,$$

where $q = 1 - p$. It is obvious that

$$E(X_i) = p \quad \text{and} \quad D(X_i) = pq, \text{ for all } i = 1, \ldots, n.$$

Then

$$\begin{aligned} E(X) = E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = np, \\ D(X) = D(X_1 + \cdots + X_n) = D(X_1) + \cdots + D(X_n) = npq. \end{aligned}$$

5.2.2. Poisson distribution $\mathcal{P}(\lambda)$

The expectation of the random variable $X:\mathcal{P}(\lambda)$ is

$$E(X)=\sum_{i=0}^{\infty} i\,\frac{\lambda^i}{i!}\,e^{-\lambda}=\lambda e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^{i-1}}{(i-1)!}=\lambda,$$

and the variance is

$$\begin{aligned}D(X) &= E(X^2)-E^2(X)=\sum_{i=0}^{\infty} i^2\,\frac{\lambda^i}{i!}\,e^{-\lambda}-\lambda^2\\ &= \sum_{i=1}^{\infty}(i-1+1)\,\frac{\lambda^i}{(i-1)!}\,e^{-\lambda}-\lambda^2\\ &= e^{-\lambda}\left(\lambda^2\sum_{i=2}^{\infty}\frac{\lambda^{i-2}}{(i-2)!}+\lambda\sum_{i=1}^{\infty}\frac{\lambda^{i-1}}{(i-1)!}\right)-\lambda^2\\ &= e^{-\lambda}(\lambda^2e^{\lambda}+\lambda e^{\lambda})-\lambda^2=\lambda.\end{aligned}$$

5.2.3. Uniform distribution $\mathcal{U}(a,b)$

The expectation of the random variable $X:\mathcal{U}(a,b)$ is

$$E(X)=\int_a^b x\,\frac{1}{b-a}\,dx=\frac{a+b}{2},$$

and the variance is

$$\begin{aligned}D(X) &= E(X^2)-E^2(X)\\ &= \int_a^b x^2\,\frac{1}{b-a}\,dx-\left(\frac{a+b}{2}\right)^2\\ &= \frac{(b-a)^2}{12}.\end{aligned}$$

5.2.4. Exponential distribution $\mathcal{E}(\lambda)$

The expectation of the random variable $X:\mathcal{E}(\lambda)$ is

$$E(X)=\int_0^{\infty} x\,\lambda\,e^{-\lambda x}\,dx=\frac{1}{\lambda^2}\int_0^{\infty} te^{-t}\,dt=\frac{1}{\lambda},$$

where we have used the change of variables $ax=t$ and the gamma function

$$\int_0^{\infty} t^{n-1}e^{-t}\,dt=\Gamma(n)=(n-1)!.$$

The variance of the random variable $X:\mathcal{E}(\lambda)$ is

$$\begin{aligned}D(X) &= E(X^2)-E^2(X)\\ &= \int_0^{\infty} x^2\,\lambda\,e^{-\lambda x}\,dx-\frac{1}{\lambda^2}\\ &= \frac{1}{\lambda^2}\,\Gamma(3)-\frac{1}{\lambda^2}=\frac{1}{\lambda^2}.\end{aligned}$$

6. Stochastic processes

A *stochastic process* $\{X(t),\ t \in T\}$ is a collection of random variables. That means that, for each $t \in T$, $X(t)$ is a random variable. The index t is often interpreted as time. Therefore, we refer to $X(t)$ as the *state* of the process at time t. For example, $X(t)$ might be the number of customers in the supermarket at time t.

The set T is called the *index* set of the process. If T is countable set then the stochastic process is said to be a *discrete-time* process. On the other hand, if T is an interval of the real line, the stochastic process is said to be a *continuous-time* process. Usually, by $\{X_n,\ n = 0, 1, \dots\}$ we denote a discrete-time process indexed by nonnegative integers, while $\{X(t),\ t \geq 0\}$ usually denotes a continuous-time stochastic process indexed by nonnegative real numbers.

The *state space* of a stochastic process is the set of all possible values that the random variables $X(t)$ can assume. Thus, a stochastic process is a family of random variables that describes the evolution through the time of some (physical) process.

6.1. Poisson process

A stochastic process $\{N(t),\ t \geq 0\}$ is said to be a counting process if $N(t)$ is the total number of events that occur by time t.

The counting process $\{N(t),\ t \geq 0\}$ is said to be a Poisson process having rate λ, $\lambda > 0$, if

(i) N(0)=0.

(ii) The process has independent increments (the number of events that occur in disjoint time intervals are independent).

(iii) The number of events in any interval of length t is Poisson distributed with mean λt. That is, for all $s, t \geq 0$,

$$P\{N(t+s) - N(s) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad n = 0, 1, 2, \dots$$

Consider a Poisson process and denote the time of the first event by T_1. Further, for $n > 1$, let T_n denote the elapsed time between the $(n-1)$st and nth event.

The sequence $\{T_n,\ n = 1, 2, \dots\}$ is called the sequence of interarrival times.

$T_n,\ n = 1, 2, \dots$ are independent identically distributed exponential random variables having mean $1/\lambda$.

The arrival (waiting) time of the nth event is $S_n = T_1 + \cdots + T_n$ and it has the probability density function

$$\varphi_{S_n}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}, \quad t \geq 0.$$

The random variable having the probability density function above is said to be Gamma distributed with parameters n and λ.

Example 11 Suppose that people immigrate into a territory at a Poisson rate $\lambda = 1$ per day.

a) What is the expected time until the tenth immigrant arrives?

a) What is the probability that the elapsed time between the tenth and the eleventh arrival exceeds two days?

Solution:

a) $E(S_{10}) = 10/\lambda = 10$ days.

a) $P\{T_{11} > 2\} = e^{-2\lambda} = e^{-2} \approx 0.133$.

7. Introduction to Statistics

Statistics refers to a range of techniques and procedures for analyzing data, interpreting data, displaying data and making decisions based on data. It includes descriptive statistics (the study of methods and tools for collecting data, and mathematical models to describe and interpret data) and inferential statistics (the systems and techniques for making probability-based decisions and accurate predictions based on incomplete (sample) data).

One important use of descriptive statistics is to summarize a collection of data in a clear and understandable way. How might data collected with some purpose be summarized? There are two basic methods: numerical and graphical. By using the numerical approach one might compute some statistics such as the mean and standard deviation. These statistics convey information about the average or the spread of the data in comparison to the average. Using the graphical approach one might create a stem and leaf display and a box plot. These plots contain detailed information about the distribution of data. Graphical methods are better than numerical methods for identifying patterns in the data. On the other hand, numerical approaches are more precise and objective. Therefore, it is wise to use them both.

Inferential statistics are used to draw inferences about a population from a sample. There are two main methods used in inferential statistics: estimation and hypothesis testing. In estimation, the sample is used to estimate a parameter and a confidence interval about the estimate is constructed.

Introducing basic statistical notions we start with a *population*. A population consists of an entire set of objects, observations, or scores that have something in common. For example, a population might be defined as all females between the ages of 15 and 19 (the high school female population). A *sample* is a subset of a population. Since it is usually impractical to test every member of a population, a sample from the population is typically the best approach available.

7.1. Measures of central tendency

Measures of central tendency are measures of the location of the middle or the center of a distribution. The term ”central tendency” can refer to a wide variety of measures. The mean is the most commonly used measure of central tendency. The following measures of central tendency are discussed in this text: mean, median and mode.

7.1.1. Arithmetic mean

The arithmetic mean is simply the arithmetic average of a group of numbers or data set. It is calculated by adding up all of the values in a data set and dividing by the number of values in that data set: For example, the mean of the set of data $\{1, 2, 3, 4, 5\}$ is:

$$\frac{1+2+3+4+5}{5} = 3.$$

In general, the formula for calculating the arithmetic mean of population is

$$m = \frac{\sum_{i=1}^{N} x_i}{N},$$

where m is the population mean, x_i is the i-th score in the population and N is the number of scores (population size). If the scores are from a sample, then the symbol $\bar{x}_n$ refers to the mean and n refers to the sample size. The formula for $\bar{x}_n$ is the same as the formula for m:

$$\bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}.$$

7.1.2. Median

The median is the ”middle value” in a set. That is, the median is the number in the center of a data set that has been ordered sequentially. Half the scores are above the median and half are below the median. The median is less sensitive to extreme scores than the mean and this makes it a better measure than the mean for highly skewed distributions.

For example, let's look at the data set: $\{9, 14, 87, 3, 70, 99, 1\}$. What is its median?

First, we sort our data set sequentially: $\{1, 3, 9, 14, 70, 87, 99\}$. Next, we determine the total number of points in our data set (in this case, 7.) Finally, we determine the central position of or data set (in this case, the 4th position), and the number in the central position is our median. Thus, in our set $\{1, 3, 9, 14, 70, 87, 99\}$, number 14 is the median.

An easy way to determine the central position or positions for any ordered set is to take the total number of points, add 1, and then divide by 2. If the number we get is a whole number, then that is the central position. If the number we get is a fraction, we take the two whole numbers on either side.

Because our data set had an odd number of points, determining the central position was easy - it will have the same number of points before it as after it. But what if our data set has an even number of points?

Let's take the same data set, but add a new number to it: $\{1, 3, 9, 14, 67, 70, 99, 100\}$. What is the median of this set?

When you have an even number of points, you must determine the two central positions of the data set. So for a set of 8 numbers, we get

$$\frac{8+1}{2} = \frac{9}{2} = 4,5,$$

which has 4 and 5 on either side.

Looking at our data set, we see that the 4th and 5th numbers are 14 and 70. Now, we take the mean of these two to determine the median:

$$\frac{14+70}{2} = \frac{84}{2} = 42.$$

7.1.3. Mode

The mode is the most common or "most frequent" value in a set. In the set $\{1, 2, 3, 4, 4, 4, 5, 6, 7, 8, 8, 9\}$, the mode would be 4 as it occurs a total of three times in the set, more frequently than any other value in the set. A data set can have more than one mode: for example, in the set $\{1, 2, 2, 3, 3\}$, both 2 and 3 are modes. If all points in a data set occur with equal frequency, it is equally accurate to describe the data set as having many modes or no mode.

The advantage of the mode as a measure of central tendency is that its meaning is obvious. Further, it is the only measure of central tendency that can be used with so-called nominal data. A set of data is said to be nominal if one can assign to its values/observations a code in the form of a number where the numbers are simply labels. One can count but not order or measure nominal data.

The mode is greatly subject to sample fluctuations and is therefore not recommended to be used as the only measure of central tendency. A further disadvantage of the mode is that many distributions have more than one mode. These distributions are called "multi modal."

The relationship of the mean, median and mode to each other can provide some information about the relative shape of the data distribution. If the mean, median, and mode are approximately equal to each other, the distribution can be assumed to be approximately symmetrical. If

$$\text{mean} > \text{median} > \text{mode},$$

the distribution will be skewed to the right or positively skewed. If

$$\text{mean} < \text{median} < \text{mode},$$

the distribution will be skewed to the left or negatively skewed. In a normal distribution, the mean, median and mode are identical.

7.2. Variance and standard deviation

The variance and the standard deviation are measures of how spread out a distribution is. In other words, they are measures of variability.

The variance is computed as the average squared deviation of each number from its mean. For example, for the numbers 1, 2, 3, 4, and 5, the mean is 3 and the variance is:

$$\sigma^2 = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5} = 2.$$

The formula for the variance in a population is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2,$$

where m is the mean of the population, x_i is the i-th score in the population and N is the population size.

When the variance is computed in a sample, then the symbol $\bar{s}_n^2$ refers to the variance and n refers to the sample size. The formula for $\bar{s}_n^2$ is the same as the formula for σ^2:

$$\bar{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2,$$

where $\bar{x}_n$ is the mean of the sample.

One can show that $\bar{s}_n^2$ is a biased estimate of σ^2. (A statistic is biased if, in the long run, it consistently over or underestimates the parameter it is estimating i.e. if its expected value is not equal to the parameter.)

The most common formula for computing variance in a sample is:

$$\hat{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2,$$

which gives an unbiased estimate of σ^2. Since samples are usually used to estimate parameters, $\hat{s}_n^2$ is the most commonly used estimation of variance. Calculating the variance is an important part of many statistical applications and analysis.

Standard deviation is simply the square root of the variance. It is an extremely useful measure of spread. An important attribute of the standard deviation as a measure of spread is that if the mean and standard deviation of a normal distribution are known, it is possible to compute the percentile rank associated with any given score. In a normal distribution, about 68% of the scores are within one standard deviation of the mean and about 95% of the scores are within two standard deviations of the mean.

8. Estimating parameters

8.1. Method of moments

The k-th moment of the random variable X, if exists, is simply $E(X^k)$.

The method of moments is a method of estimation of population parameters such as mean, variance, median, etc. (which need not be moments), by equating sample moments with unobservable population moments and then solving those equations for the quantities to be estimated.

Example 12 Let X be an exponentially distributed random variable, $X : \mathcal{E}(\lambda)$. Based on a sample having size n, estimate the parameter λ by using the method of moments.

Solution: Since X is $\mathcal{E}(\lambda)$-distributed, we know that $E(X) = \dfrac{1}{\lambda}$. Since the first moment is

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

we take

$$\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}_n},$$

for the estimation of the parameter λ.

8.2. Maximum likelihood estimation

Maximum likelihood estimation (MLE) is a popular statistical method used for fitting a statistical model to data, and providing estimates for the model's parameters.

In general, for a fixed set of data and underlying probability model, the method of maximum likelihood selects values of the model parameters that produce the distribution most likely to have resulted in the observed data (i.e. the parameters that maximize the likelihood function). Maximum likelihood estimation gives a unified approach to estimation, which is well-defined in the case of the normal distribution and many other problems. However, in some complex problems, difficulties do occur: in such problems the maximum-likelihood estimators may be unsuitable or may not even exist.

Let θ be an unknown parameter appearing in distribution of X. We want to estimate θ by using (MLE), based on a sample having size n.

First, we construct the likelihood function $L = L(x_1, x_2, \ldots, x_n, \theta)$ in the following way:

- If X is a discrete random variable with the probability mass function
$$P\{X = x_i\} = p(x_i, \theta),$$
then the likelihood function is
$$L = L(x_1, x_2, \ldots, x_n, \theta) = p(x_1, \theta)\ p(x_2, \theta) \cdots p(x_n, \theta).$$
- If X is a continuous random variable with the density function $\varphi(x, \theta)$, then the likelihood function is
$$L = L(x_1, x_2, \ldots, x_n, \theta) = \varphi(x_1, \theta)\ \varphi(x_2, \theta) \cdots \varphi(x_n, \theta).$$

The estimation of θ is

$$\hat{\theta}_{MLE} = \max_{\theta} L = \max_{\theta} L(x_1, x_2, \ldots, x_n, \theta).$$

If L does not reach its maximum with respect to θ, then the estimation obtained by (MLE) does nor exist.

Since the likelihood function is a product, we seek its maximum by solving the equation

$$\frac{\partial \ln L}{\partial \theta} = 0,$$

instead of solving

$$\frac{\partial L}{\partial \theta} = 0,$$

if, of course, differentiation makes sense.

Since the logarithm function is monotone, function $\ln L$ reaches its maximum in the same points as L.

The function $\ln L$ is often called log-likelihood function.

Example Let X is exponentially distributed with parameter λ. Based on a sample having size n estimate the parameter λ by using (MLE).

Solution: Since $X : \mathcal{E}(\lambda)$, the likelihood function is

$$L(x_1, x_2, \ldots, x_n, \lambda) = \lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} x_i\right\},$$

and

$$\ln L = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Therefore

$$\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i.$$

By solving the equation

$$\frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0$$

we obtain that

$$\hat{\lambda}_{MLE} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}_n},$$

is an estimation of parameter λ.

Note that we obtained the same estimation of the parameter λ as in case when we used the method of moments.

8.3. Confidence intervals

Let's say that the population we are investigated is specified and it consists of all high school students. We are interested in their knowledge in mathematics. of course, we can not test all students attending high school, so the next step is to take a sample from the population. In our example, suppose that 10 students are drawn and each student's knowledge in mathematics is tested.

The way to estimate the mean of all high school students, m, is to compute the mean of the 10 students in the sample. Indeed, the sample mean is an unbiased estimate of m. But it will certainly not be a perfect estimate.

For the estimate of m to be of value, one must have some idea of how precise it is. That is, how close to m is the estimate likely to be?

An excellent way to specify the precision is to construct a so-called confidence interval. A confidence interval is a range of values computed in such a way that it contains the estimated parameter a high probability. The 95% confidence interval is constructed so that 95% of such intervals will contain the parameter. Similarly, 99% of 99% confidence intervals contain the parameter. The wider the interval, the more confident you are that it contains the parameter.

8.3.1. Confidence interval for m when standard deviation is known

Here we want to compute the confidence interval for the mean of a normally-distributed variable for which the population standard deviation is known. of course, in practice, the standard deviation of population is rarely known. Therefore, the next step would be to compute a confidence interval when the standard deviation has to be estimated and this would be easier to do after considering this case.

To construct a confidence interval for m we need to know the sample mean $\bar{x}_n$, the value of z (which depends on the level of confidence), the standard deviation σ and the sample size n. The confidence interval has $\bar{x}_n$ for its center and its formula is:

$$\bar{x}_n - z\,\frac{\sigma}{\sqrt{n}} \le m \le \bar{x}_n + z\,\frac{\sigma}{\sqrt{n}}.$$

Assume that our 10 high school students are tested in mathematics and their scores (out of 1000) are 320, 380, 400, 420, 500, 520, 600, 660, 720, and 780. We suppose that the standard deviation of mathematics test scores in a high school system is known to be 100. We wish to compute a 95% confidence interval for the mean m from a random sample of these 10 scores.

First, note that $n = 10$ and $\sigma = 100$ are known. One can easily compute $\bar{x}_n = 530$. It remains to determine the value of z for the 95% confidence interval. It is the number of standard deviations one must go from the mean (in both directions) to contain 0.95 of the scores. The value $z = 1.96$ can be found by using a so-called z table.

All the components of the confidence interval are now known and we can easily compute the 95% confidence interval for m:

$$(530 - 1.96\,\frac{100}{\sqrt{10}},\ 530 + 1.96\,\frac{100}{\sqrt{10}}) = (468.02,\ 591.98).$$

If a larger sample size had been used, the range of scores would have been smaller.

The computation of the 99% confidence interval is exactly the same except that $z = 2.58$. The 99% confidence interval is: $(448.54,\ 611.46)$. As it must be, the 99% confidence interval is wider than the 95% confidence interval.

8.3.2. Confidence interval for m when standard deviation is estimated

It is very rare that a researcher wishing to estimate the mean of a population already knows its standard deviation. Therefore, the construction of a confidence interval in most of cases involves the estimation of both m and σ.

Whenever the standard deviation is estimated, the t distribution rather than the normal distribution (z distribution) should be used.

To construct a confidence interval for m when standard deviation is estimated, we need to know the sample mean $\bar{x}_n$, the value of t (which, same as z, depends on the level of confidence), the standard deviation of a sample $\bar{s}_n$ and the sample size n. The confidence interval again has $\bar{x}_n$ for its center and its formula is:

$$\bar{x}_n - t\,\frac{\hat{s}_n}{\sqrt{n}} \leq m \leq \bar{x}_n + t\,\frac{\hat{s}_n}{\sqrt{n}}.$$

As in the case for the value of z which we can found in z table, the value of t can be determined by using a so-called t table. It is actually the quantile of order $(1+\gamma)/2$ of t distribution, where γ is the confidence level, for example, $\gamma = 0.95$.

9. Hypothesis testing

A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

There are two types of statistical hypotheses: null hypothesis and alternative hypothesis. The null hypothesis, denoted by H_0, is usually the hypothesis that sample observations result purely from chance. The alternative hypothesis, denoted by H_1, is the hypothesis that sample observations are influenced by some non-random cause.

Now, suppose that we want to test some hypothesis about a population parameter. We start with the null hypothesis. Then, we make a sample from the population and collect data. The next step is to determine the viability of the null hypothesis in light of the data. If the data are very different from what would be expected under the assumption that the null hypothesis is true, then the null hypothesis is rejected. If the data are not greatly at variance with what would be expected under the assumption that the null hypothesis is true, then the null hypothesis is not rejected.

The null hypothesis is often the reverse of what the experimenter actually believes; it is put forward to allow the data to contradict it.

9.1. Testing of m, $H_0(m = m_0)$, when standard deviation is known

Let the population variable X be normally $\mathcal{N}(m; \sigma^2)$-distributed, where m is unknown parameter, and suppose we know σ. Denote by $X_1, \ldots X_n$ the sample and by $x_1, \ldots, x_n$ its realizations.

Suppose that $m = m_0$.

We consider the deviation of the sample arithmetic mean with respect to m_0 and calculate, by using the z table, the probability

$$\begin{aligned} P(|\bar{X}_n - m_0| \geq |\bar{x}_n - m_0|) &= P\left\{\left|\frac{\bar{X}_n - m_0}{\sigma}\sqrt{n}\right| \geq \left|\frac{\bar{x}_n - m_0}{\sigma}\sqrt{n}\right|\right\} \\ &= 2\left(1 - \Phi\left(\frac{\bar{x}_n - m_0}{\sigma}\sqrt{n}\right)\right) \\ &= \alpha^*. \end{aligned}$$

If $\alpha^* < \alpha$, where α is the significance level, we reject $H_0(m = m_0)$, otherwise we do not reject it.

9.2. Testing of m, $H_0(m = m_0)$, when standard deviation is estimated

Let the population variable X be normally $\mathcal{N}(m; \sigma^2)$-distributed, where both m and σ are unknown. As always, denote by $X_1, \ldots X_n$ the sample and by $x_1, \ldots, x_n$ its realizations.

Suppose that $m = m_0$.

We start with

$$\frac{\bar{X}_n - m_0}{\bar{S}_n}\sqrt{n-1},$$

which is t_{n-1}-distributed. Then we calculate, by using the t table, the probability

$$\alpha^* = P\left\{\left|\frac{\bar{X}_n - m_0}{\bar{S}_n}\sqrt{n-1}\right| \geq \left|\frac{\bar{x}_n - m_0}{\bar{S}_n}\sqrt{n-1}\right|\right\}.$$

If $\alpha^* < \alpha$, where α is the significance level, we reject $H_0(m = m_0)$, otherwise we do not reject it.

Example 13 Let X be normally $\mathcal{N}(m; \sigma^2)$-distributed. Based on the sample having size 5, it is obtained $\bar{x}_5 = 10$ and $\hat{s}_n^2 = 0.09$. Test the hypothesis $H_0(m = 9.9)$ for $\alpha = 0.1$.

Solution: We calculate

$$\frac{\bar{x}_n - m_0}{\bar{S}_n}\sqrt{n-1} = \frac{10 - 9.9}{0.3}\sqrt{4} \approx 0.67.$$

Now, by using the t table we determine $\alpha^* = 0.5$. Since $\alpha^* \geq \alpha$, hypothesis $H_0(m = 9.9)$ can not be rejected.

9.3. Pearson's chi-square test

Statistical procedures whose results are evaluated by reference to the chi-square distribution are all named chi-square tests. Pearson's chi-square test is the best-known of several chi-square tests. It tests a null hypothesis stating that the frequency distribution (a frequency distribution is a tabulation of the values that one or more variables take in a sample) of certain events observed in a sample is consistent with a particular theoretical distribution. The events considered must be mutually exclusive and have total probability 1.

Pearson's chi-square is used to assess two types of comparison: tests of goodness of fit and tests of independence. A test of goodness of fit establishes whether or not an observed frequency distribution differs from a theoretical distribution. A test of independence assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other.

9.3.1. Test for fit of a distribution

We start with the null hypothesis saying that X has the distribution function:

$$F_X(x) = F_0(x).$$

The procedure goes as follows.

- We divide the set of real numbers in k disjunct intervals
$$I_1, I_2, \dots, I_k, \;\; I_m = [a_{m-1}, a_m), \;\; m = 1, 2, \dots, k.$$
- If the distribution F_0 has some unknown parameters, we estimate them (by using method of moments, MLE,...). We denote the number of unknown parameters in F_0 by s.
- We calculate so-called theoretical probabilities:
$$p_m = P\{X \in I_m\} = F_0(a_m) - F_0(a_{m-1}), \;\; m = 1, 2, \dots, k.$$
- We set
$$Z = \sum_{m=1}^{k} \frac{(N_m - np_m)^2}{np_m}.$$
Random variable N_m, $m = 1, 2, \dots, k$ are the number of values from the sample belonging to the interval I_m.
- Based on the sample we calculate the value of Z:
$$z = \sum_{m=1}^{k} \frac{(n_m - np_m)^2}{np_m},$$
where n_m are actual values of random variables N_m obtained from the sample.
- For given α, in the table for χ^2-distribution, we find the value $\chi^2_{\alpha,k-1-s}$, so that
$$P\left\{z \ge \chi^2_{\alpha,k-1-s}\right\} = \alpha.$$
- Finally, we compare the values of $\chi^2_{\alpha,k-1-s}$ and z in order to decide whether to accept the null hypothesis or not. If $\chi^2_{\alpha,k-1-s} > z$ we accept the null hypothesis $F_X(x) = F_0(x)$, otherwise we reject it.

Example 14 The die is rolled for 120 times. The result of the experiment is given in the table:

the number on the die	1	2	3	4	5	6
frequence	20	14	23	12	26	25

Taking $\alpha = 0.1$ test the hypothesis that the die is correct (the probabilities of showing any of six numbers are equal).

Solution: We test the hypothesis that X, representing the number appearing on the die after the rolling, has the distribution

$$X : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}.$$

The sample size is $n = 120$. We divide the set of real numbers in intervals

$$I_1 = (-\infty, 1.5)\,,\; I_2 = [1.5, 2.5)\,,\; I_3 = [2.5, 3.5)\,,$$
$$I_4 = [3.5, 4.5)\,,\; I_5 = [4.5, 5.5)\,,\; I_6 = [6.5, +\infty)\,,$$

having frequencies

$$n_1 = 20,\ n_2 = 14,\ n_3 = 23,\ n_4 = 12,\ n_5 = 26,\ n_6 = 25.$$

Now, we have to find the theoretical probabilities

$$p_m = P_{H_0}\{X \in I_m\},\quad m = 1,2,3,4,5,6.$$

We have

$$p_1 = P_{H_0}\{X \in (-\infty, 1.5]\} = P_{H_0}\{X = 1\} = \frac{1}{6},$$
$$p_2 = P_{H_0}\{X \in (1.5, 2.5]\} = P_{H_0}\{X = 2\} = \frac{1}{6}.$$

Analogously, we obtain

$$p_3 = p_4 = p_5 = p_6 = \frac{1}{6}.$$

We find the value

$$\chi^2_{6-1} = \chi^2_5 = \sum_{m=1}^{6} \frac{(N_m - n \cdot p_m)^2}{n \cdot p_m}.$$

From the sample we obtain

$$\chi^2_5 = \frac{\left(20 - 120 \cdot \frac{1}{6}\right)^2}{120 \cdot \frac{1}{6}} + \frac{\left(14 - 120 \cdot \frac{1}{6}\right)^2}{120 \cdot \frac{1}{6}} + \frac{\left(23 - 120 \cdot \frac{1}{6}\right)^2}{120 \cdot \frac{1}{6}} +$$

$$+ \frac{\left(12 - 120 \cdot \frac{1}{6}\right)^2}{120 \cdot \frac{1}{6}} + \frac{\left(26 - 120 \cdot \frac{1}{6}\right)^2}{120 \cdot \frac{1}{6}} + \frac{\left(25 - 120 \cdot \frac{1}{6}\right)^2}{120 \cdot \frac{1}{6}} = 8.5.$$

Finally, from the table for χ^2-distribution we determine $\chi^2_{5;0.1} = 9.236$. Since $9.236 > 8.5$, we accept the null hypothesis.

References

[1] Z.A.Ivković, Mathematical Statistics, Naučna knjiga, Beograd, 1980. (In Serbian)

[2] D. Rajter-Ćirić, Probability, University of Novi Sad, 2009. (In Serbian)

[3] D. Rajter-Ćirić, Stochastic Analysis, notes for students. (In Serbian)

[4] S. Ross, Introduction to Probability Models, Academic Press, 2003.