Analysis and Interpretation of Data 1
aka
Introduction to Probability & Statistics

University of Ostrava

Version of August 21, 2018

Literature

- F.M. Dekking, C. Kraaikamp, H.P. Lopuhaä, and L.E. Meester, *A Modern Introduction to Probability and Statistics*, Springer, 2005
  (referred as DEKKING ET AL.)
- L. Gonick and W. Smith, *The Cartoon Guide to Statistics*, HarperPerennial, 1993
- R.L. Graham, D.E. Knuth, and O. Patashnik, *Concrete Mathematics*, 2nd ed., Addison-Wesley, 1994
  (referred as GRAHAM–KNUTH–PATASHNIK)
- F. Mosteller, *Fifty Challenging Problems in Probability with Solutions*, Dover, 1987
- J.A. Rice, *Mathematical Statistics and Data Analysis*, 3rd ed., Thomson Brooks/Cole, 2007
  (referred as RICE)
- N.Ya. Vilenkin, *Combinatorics*, Academic Press, 1971
  (referred as VILENKIN)

(All images are courtesy of Wikipedia)

1.
A subject of probability and statistics

What is probability and statistics?

Probability theory is a branch of mathematics which tries to argue rigorously about random and uncertain things.

Though it cannot be said anything definite about outcomes of a single random event, when considering a number of such events in their totality, certain patterns emerge; these patterns are amenable to a rigorous mathematical study.

Statistics (also called sometimes "data science") deals with collection, analysis, interpretation, presentation, visualization, and organization of various ("real-world") data. Mathematical foundations of statistics are based on probability theory.

Examples of applications of statistics

- Establishing links between genotype of an individual and his risks to die from a certain disease (cancer, cardiovascular, etc.).
- The reason of the space shuttle "Challenger" disaster.
- Detection of election frauds.
- Revelation that the "most cited papers" are not read by those who cite them.

For details and more examples, see DEKKING ET AL., pp. 1–11.

2.
Basic combinatorics: combinations, variations, permutations

Permutations

Definition

A *permutation* of a set is a bijection of the set to itself.

In other words, a permutation is a way of arranging elements of a set into some order.

Theorem 1

The number of permutations of a set of n elements is equal to $n!$.

Try to prove this! (**Hint**: use induction).

Variations

Definition

A *variation without repetition* is a way to choose k elements out of n elements, taking into account the order of elements.

Theorem 2

The number of variations without repetition is equal to $\frac{n!}{(n-k)!}$.

Definition

A *variation with repetition* is a way to choose k elements out of n elements, taking into account the order of elements, and with possible repetitions of elements.

Theorem 3

The number of variations with repetition is equal to n^k.

For proofs and examples, see Vilenkin, pp. 3–6,18–19.

Combinations

Definition

A *combination* is a way to choose k elements out of n elements, without taking into account the order of elements, and not allowing repetitions.

Theorem 4

The number of combinations is equal to $\binom{n}{k}$.

Proof. This is the same as doing variation without repetitions, but without accounting for different permutations of elements, i.e. the number in Theorem 2 should be divided by the number of permutations of k elements, which, according to Theorem 1, is equal to $k!$: $\frac{n!}{k!(n-k)!}$.

Example of combinations

In many card games, each player gets 6 cards out of the standard deck of 52 cards. There are

$$\binom{52}{6} = \frac{52!}{46!\,6!} = 20,358,520$$

possibilities for a 6-card hand.

For more examples, see Vilenkin, pp. 28–32, and Rice, p. 10,12–13.

Binomial coefficients

The binomial coefficients $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ occur in the (well known from the high school) binomial formula:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

and can be arranged into the *Pascal triangle*:

$$\begin{array}{lccccccccc}
n = 0: & & & & & 1 & & & & \\
n = 1: & & & & 1 & & 1 & & & \\
n = 2: & & & 1 & & 2 & & 1 & & \\
n = 3: & & 1 & & 3 & & 3 & & 1 & \\
n = 4: & 1 & & 4 & & 6 & & 4 & & 1
\end{array}$$

Binomial coefficients (cont.)

The main properties of binomial coefficients are:

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$
$$\binom{n}{k} = \binom{n}{n-k}$$

For more properties, see Graham–Knuth–Patashnik, pp. 153–196, and Vilenkin, pp. 34–42,61–63.

A particular case of the binomial formula is

$$2^n = \sum_{k=0}^{n} \binom{n}{k}.$$

A combinatorial meaning of this formula: all possible ways to choose elements out of n elements, i.e., the number of n-length binary sequences (like $0110\ldots$, etc.)

Inclusion-exclusion principle

In counting, the following *inclusion-exclusion principle* is often used.

Theorem

The number of elements in the union of sets $A_1, \dots, A_n$ is determined by the formula

$$
\begin{aligned}
|A_1 \cup \cdots \cup A_n| = & \\
& |A_1| + \cdots + |A_n| \\
& -|A_1 \cap A_2| - |A_1 \cap A_3| - \cdots - |A_{n-1} \cap A_n| \\
& +|A_1 \cap A_2 \cap A_3| + \cdots + |A_{n-2} \cap A_{n-1} \cap A_n| \\
& - \cdots + \\
& +(-1)^{n+1}|A_1 \cap A_2 \cap \cdots \cap A_n|.
\end{aligned}
$$

For a proof and examples, see Vilenkin, pp. 12–17.

3.
Sample space, events, probability function

Sample space, events

A *sample space* (also called *probability space*) is a set whose elements represent the possible outcomes of the event we are interested in.

An *event* is a subset of the sample space.

An *elementary event* is a subset of the sample space consisting of one element.

The set-theoretic operations on events correspond to their logical combinations: the intersection $A \cap B$ of events A and B occurs when both A and B occur; the union $A \cup B$ occurs when either A or B occurs; the complement $\Omega \setminus A$, where Ω is the whole sample space, occurs when A does not occur.

Examples of sample spaces and events

- When tossing a coin, the sample space is the 2-element set $S = \{\text{head}, \text{tail}\}$.
- When tossing a dice, the sample space is the 6-element set $\{1, 2, 3, 4, 5, 6\}$.
- If we are throwing an (idealized) dart (i.e., a point) at an (idealized) dartboard (say, a circle with radius 1 with the center at the origin), the sample space is

  $$\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}.$$

  The perfect hit, $\{(0, 0)\}$, is an elementary event, while hitting, say, the right half of the dartboard:

  $$\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1, x \geq 0\}$$

  will constitute an event.

Exercise

Give more examples.

More examples: tossing two coins

When tossing two coins simultaneously, the sample space is the Cartesian product $S \times S$, i.e. the set

$$\{(\text{head}, \text{head}), (\text{head}, \text{tail}), (\text{tail}, \text{head}), (\text{tail}, \text{tail})\}.$$

The event of having head first is

$$A = \{(\text{head}, \text{head}), (\text{head}, \text{tail})\},$$

and the event of having head second is

$$B = \{(\text{head}, \text{head}), (\text{tail}, \text{head})\}.$$

Their intersection is an elementary event having both heads, $\{(\text{head}, \text{head})\}$, and their union is an event of having at least one head:

$$A \cup B = \{(\text{head}, \text{head}), (\text{tail}, \text{head}), (\text{head}, \text{tail})\}.$$

The complement of A is an event of having tails first:

$$S \setminus A = \{(\text{tail}, \text{head}), (\text{tail}, \text{tail})\}.$$

More examples: birthday dates

We may sample birthday dates in a certain group of people. The sample space in this case is the subset of the cartesian product

$$\{1, 2, \dots, 31\} \times \{\text{Jan}, \text{Feb}, \dots, \text{Dec}\}$$

(subset, as certain pairs, like February 30 and June 31, are excluded). Now we may be interested, for example, in people born at the specific date, so the one-element sets $\{(1, \text{Jan})\}$ and $\{(10, \text{Mar})\}$ will constitute elementary events, while all birthdays occurring in February:

$$\{1, 2, \dots, 29\} \times \{\text{Feb}\},$$

or all birthdays occurring at the end of the month:

$$\{(30, \text{Jan}), (28, \text{Feb}), (29, \text{Feb}), \dots, (31, \text{Dec})\}$$

will constitute (just) events.

Probability function

Probability is a numerical expression of how likely an event occurs. If all outcomes in the sample space Ω occur equally likely, then the probability $P(A)$ of an event A is equal to $\frac{|A|}{|S|}$, where $|X|$ is the cardinality (number of elements) of the set X. In particular, the probability of an elementary event is equal to $\frac{1}{|S|}$.
More formally:

Definition

A *probability function* P on a sample space Ω is a function from the set of all possible events, i.e. the powerset $\mathscr{P}(\Omega)$, to the interval $[0,1]$, such that $P(\Omega) = 1$, and

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

if A and B do not occur simultaneously, i.e. $A \cap B = \varnothing$.

The probability of an event can be computed by summing up probabilities of all outcomes (elementary events) comprising it.

Examples

- In the example with two coins tossing, we have

$$P(A) = P(B) = P(S \setminus A) = \frac{2}{4} = \frac{1}{2}$$
$$P(A \cap B) = \frac{1}{4}$$
$$P(A \cup B) = \frac{3}{4}$$

- If we are throwing a crooked dice, where 6 can occur with the probability $\frac{1}{5} = 0.2$ (instead of the fair $\frac{1}{6}$), and the rest of points, from 1 till 5, can occur with the equal probability 0.16, the probability to get an even number of points is equal to

$$P(\{2,4,6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = 2 \times 0.16 + 0.2 = 0.52$$

For more examples, see Graham–Knuth–Patashnik, pp. 382–383, and Rice, pp. 6–7,10–11.

Probability function (cont.)

Lemma 1

If P is a probability function, then $P(\varnothing) = 0$.

Proof. Follows from $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ for $A_1 \cap A_2 = \varnothing$ (take $A_2 = \varnothing$).

Lemma 2

If P is a probability function on a sample space Ω, then for any $A \subseteq \Omega$, $P(\Omega \backslash A) = 1 - P(A)$.

Proof. Follows from Lemma 1, and from $P(\Omega) = 1$ (take $A_1 = A$, $A_2 = S \backslash A$).

Probability function on infinite sample spaces

Still, these notions of event and probability are not entirely satisfactory, as we can run into problems with infinite sets. When the sample space Ω is infinite, the appropriate notion of event appears to be not an arbitrary subset of Ω, but an element of a σ-algebra, i.e. a set of subsets of S closed with respect to complements, and countable unions and intersections. Then the probability function P is defined as a measure on the σ-algebra, normalized by the condition $P(\Omega) = 1$.

Examples

- We are tossing an unbiased coin till the fist head. Our sample space is $\Omega = \{1, 2, 3, \dots, n, \dots\}$, where n signifies that the first head occurs at nth toss. Then $P(n) = \frac{1}{2^n}$, and
$$P(1) + P(2) + P(3) + \cdots = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 1,$$
as expected.
- A general way to give examples of events and probability functions: define a (finite or countable) set Ω, and define $P(a) \in [0, 1]$ for each $a \in \Omega$ such that $\sum_{a \in \Omega} P(a) = 1$.
- In the throwing darts example, the probability to hit any measurable subset A of our idealized dartboard is equal to $\frac{\mu(A)}{\pi}$. For example, the probability of the perfect hit is zero (as the measure of a set consisting of a single point is zero), while the probability to hit the right half of the dartboard is $\frac{1}{2}$.

The bottom line: to compute probabilities, we have to count, be it counting of discrete sets, like in combinatorics, or counting of areas of geometric figures, like in mathematical analysis.

Application of inclusion-exclusion principle

For arbitrary events $A_1, A_2, \ldots, A_n$, not necessary disjoint, we have

$$\begin{aligned} P(A_1 \cup \cdots \cup A_n) = & \\ & P(A_1) + \cdots + P(A_n) \\ & -P(A_1 \cap A_2) - P(A_1 \cap A_3) - \cdots - P(A_{n-1} \cap A_n) \\ & +P(A_1 \cap A_2 \cap A_3) + \cdots + P(A_{n-2} \cap A_{n-1} \cap A_n) \\ & - \cdots + \\ & +(-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n), \end{aligned}$$

what is exactly the inclusion-exclusion principle.

4.

Conditional probability, Bayes' formula, independent events

Conditional probability

Definition

A *conditional probability*, denoted by $P(A|B)$, is a probability of an event A assuming that another event B has occurred. It is defined as

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

(Of course, we assume here that $P(B) > 0$).

Theorem

Let Ω be a sample space, and B an event. Then $Q : \mathscr{P}(\Omega) \to [0,1]$ defined as $Q(A) = P(A|B)$, is a probability function on Ω.

Example

Assuming a non-leap year, let A be an event that a person has a birthday at the first day of a month, and B an event that a person has a birthday at an odd-numbered day at summer. Then $A \cap B$ is an event that a person has a birthday at the first day of a summer month,

$$P(A \cap B) = \frac{3}{365},$$
$$P(B) = \frac{15+16+16}{30+31+31} = \frac{47}{92},$$
$$P(A|B) = \frac{\frac{3}{365}}{\frac{47}{92}} = \frac{276}{17155} \approx 0.016.$$

Compare this with the value of unconditional probability

$$P(A) = \frac{12}{365} \approx 0.033.$$

Bayes' formula

Having two events A and B with nonzero probabilities, along with the conditional probability $P(A|B)$, we may consider the conditional probability

$$P(B|A) = \frac{P(A \cap B)}{P(A)},$$

what implies

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

This is known as *Bayes' formula*.

Generalization of Bayes' formula

Theorem

Let Ω be a sample space, $B_1, \dots, B_n$ events such that $B_1 \cup \dots \cup B_n = \Omega$, and B_i's are pairwise disjoint. Then

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + \dots + P(A|B_n)P(B_n)}.$$

Proof. 1st way: by induction. 2nd way: using (several times) additivity of the conditional probability, and Bayes' formula.

Independent events

Definition

Two events A and B are called *independent*, if one of the following six equivalent condition holds:

(i) $P(A|B) = P(A)$

(ii) $P(B|A) = P(B)$

(iii) $P(A \cap B) = P(A)P(B)$

and the same conditions (i)-(iii) with A and B being replaced by $\overline{A}$ and $\overline{B}$ (complements), respectively.

The equivalence follows from the definition of conditional probability, and Bayes' formula.

Several independent events

Generalization of the definition from the previous slide to the case of several events is not as straightforward as one might think at the first glance:

Definition 1

Events $A_1, \ldots, A_n$ are called *independent*, if

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \ldots P(A_{i_k})$$

for any $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$.

Definition 2

Events $A_1, \ldots A_n$ are called independent, if

$$P(B_1 \cap \cdots \cap B_n) = P(B_1) \ldots P(B_n)$$

where each B_i's is either A_i or $\overline{A}_i$.

Exercises

Exercise 1

Prove that these two definitions are equivalent.

Exercise 2

When events A and $\overline{A}$ are independent?

(Answer: if and only if $P(A) = 0$ or 1).

Example

For simplicity of calculation in this example assume that every month of a year has 30 days. For example, an event of having a birthday specified in terms of the day of the month (e.g., at the 10th day of the month, at odd days, from 10th till 15th day, etc.) is independent from the event of having birthday specified in terms of the month (e.g., at January, at spring, at the last 3 months of the year, etc.). On the other hand, the events of having birthday at summer, and at the odd-numbered months are not independent (intuitively this is clear, but check it numerically!)

For more examples, see Dekking et al., pp. 26–29 and Rice, pp. 16,24–26.

5. Discrete random variable, distribution function

Discrete random variable

In some situations, we may be interested not in the sample space itself, but only in some of its features. This leads us to the notion of a random variable.

Definition

A *discrete random variable* is a function on the sample space Ω with values in $\mathbb{R}$, accepting finite or countable number of different values.

Of course, if the sample space if finite, then any random variable defined on it is discrete.

Example

When throwing pair of dices, we may be interested not in the exact outcome, but merely in the sum of two throws, or in the maximum of two throws.

Mass and distribution functions

Definition

Let $X : \Omega \to \mathbb{R}$ be a discrete random variable defined on a sample space Ω. The *mass function* of X is the function $f_X : \mathbb{R} \to [0, 1]$ defined for any $x \in \mathbb{R}$ as

$$f_X(x) = P(X = x).$$

Definition

The *distribution function* of X is the function $F_X : \mathbb{R} \to [0, 1]$ defined for any $x \in \mathbb{R}$ as

$$F_X(x) = P(X \leq x).$$

Notational warning

Formally, the right-hand sides of the two formulas from the previous slide had to be written as

$$P(\{s \in \Omega \mid X(s) = x\}),$$

and

$$P(\{s \in \Omega \mid X(s) \leq x\}),$$

respectively, but here and in similar situations below, we use the universally accepted shorthands.

Properties of distribution functions

Theorem

For any distribution function F_X of a discrete random variable X, the following holds:

1. F_X is non-decreasing.
2. F_X is piecewise constant (i.e., has “jumps” only in a finite or countable number of points).
3. $\lim_{x\to-\infty} F_X(x) = 0$, $\lim_{x\to+\infty} F_X(x) = 1$.

Proof. By definition, the mass function attains possibly non-zero values in the finite or countable number of points (the values of the discrete random variable X), and is zero elsewhere. Thus we have

$$F_X(x) = \sum_{t\leq x} f_X(t)$$

for any $x \in \mathbb{R}$.

Example

In the tossing coin example, let us assign numerical values of 0 and 1 to tail and head respectively, and on the sample space Ω of outcomes of tossing 3 coins simultaneously, consider the random variable X equal to the sum of all 3 outcomes, so the possible values of X are $0, 1, 2, 3$. Let us compute the corresponding mass and distribution functions.

$$
\begin{aligned}
f_X(0) &= P(X=0) = P(\{(0,0,0)\}) = \tfrac{1}{8};\\
f_X(1) &= P(X=1) = P(\{(1,0,0),(0,1,0),(0,0,1)\}) = \tfrac{3}{8};\\
f_X(2) &= P(X=2) = P(\{(1,1,0),(1,0,1),(0,1,1)\}) = \tfrac{3}{8};\\
f_X(3) &= P(X=3) = P(\{(1,1,1)\}) = \tfrac{1}{8};
\end{aligned}
$$

$$
\begin{aligned}
F_X(0) &= f_X(0) = \tfrac{1}{8};\\
F_X(1) &= f_X(0) + f_X(1) = \tfrac{1}{8} + \tfrac{3}{8} = \tfrac{1}{2};\\
F_X(2) &= f_X(0) + f_X(1) + f_X(2)\\
&= \tfrac{1}{8} + \tfrac{3}{8} + \tfrac{3}{8} = \tfrac{7}{8};\\
F_X(3) &= f_X(0) + f_X(1) + f_X(2) + f_X(3)\\
&= \tfrac{1}{8} + \tfrac{3}{8} + \tfrac{3}{8} + \tfrac{1}{8} = 1.
\end{aligned}
$$

Example (cont.)

At any other point, the value of f_X is zero, and

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0; \\ F_X([x]) & \text{if } 0 \leq x \leq 3; \\ 1 & \text{if } x > 3. \end{cases}$$

(Here $[x]$ denotes the integer part of x).

For other examples, see Dekking et al., pp. 43–44.

6.
Continuous random variable, density, distribution function

Continuous random variable, density function

If a real-valued function defined on the sample space Ω attains not a discrete, but a continuous range of values, we arrive at the notion of a continuous random variable. Formally:

Definition

A random variable $X : \Omega \to \mathbb{R}$ is *continuous*, if

$$P(a \le X \le b) = \int_a^b f_X(t)\,\mathrm{d}\,t$$

for some function $f_X : \mathbb{R} \to \mathbb{R}$, and any $a, b \in \mathbb{R}$, $a \le b$. The function f_X is called the *density function* of X.

Example

Picking a point at a circle of radius R, the random variable $X(r)$ is the probability that the point will lie in a circle with radius r, $0 \le r \le R$.

Continuous random variables (cont.)

In the real world, we are dealing with discrete random variables, even with a particular case of them which involves only finite number of possible values. Continuous random variables are very useful mathematical abstractions helping to capture important properties of the discrete case when the number of possible values is becoming huge. This explains a big similarity between discrete and continuous random variables: as a rule of thumb, any formula, result, or reasoning involving the discrete case can be turned into the continuous one, by replacing summation by integration.

We can operate with random variables defined on the same sample space, both discrete and continuous, the same way as we operate with functions: we can add them, multiply them, apply other functions to them, etc.

Properties of density functions

Density is a continuous analog of the mass function of a discrete random variable.

Warning

Density is not probability!

Theorem

For any density function f_X of a continuous random variable X, the following holds:

1. f_X attains only non-negative values.
2. $\int_{-\infty}^{\infty} f_X(t)\,\mathrm{d}\,t = 1$.

Distribution function

Definition

The distribution function $F_X : \mathbb{R} \to [0,1]$ of a continuous random variable X is defined the same way as for a discrete one:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}\,t$$

for any $x \in \mathbb{R}$.

7.

Numerical characteristics of a random variable: expectation, quantile, median, standard deviation

Expectation

Random variables may contain a huge amount of data in a very complicated form, so sometimes one wants to summarize that or another property of a random variable by a single number.

The *expected value*, or *mean*, of a random variable X, denoted by $E[X]$, is its average value, or, in other words, the center of the corresponding distribution function. More formally:

Definition

For a discrete random variable X attaining values $x_1, x_2, \ldots$, the *expected value* is the weighted mean of the values, with weights being the respective probabilities:

$$E[X] = \sum_{i=1,2,\ldots} f_X(x_i) x_i.$$

Expectation (cont.)

Warning

Note that, generally, we are dealing here with an infinite sum, which may not exist. However, it does exist in most of the important cases occurring on practice. Of course, if the random variable X attains only finite number of values, the sum is finite and thus exists always.

Example

The expected value of the random variable equal to the number of points got in one throw of a dice is equal to

$$\frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = 3.5.$$

For example of a random variable with non-existing expectation, see DEKKING ET AL., p. 92.

Expectation (cont.)

Definition

The *expected value* of a continuous random variable X is defined as

$$E[X] = \int_{-\infty}^{\infty} t f_X(t) \, \mathrm{d}\, t.$$

Theorem

Let X be a random variable, and $g : \mathbb{R} \to \mathbb{R}$ a real function. If X is discrete, taking values $a_1, a_2, \ldots,$ then

$$E[g(X)] = \sum_i g(a_i) P(X = a_i).$$

If X is continuous, with density function f, then

$$E[g(X)] = \int_{-\infty}^{+\infty} g(t) f(t) \, \mathrm{d}\, t.$$

Expectation (cont.)

The formulas in the theorem from the previous slide are called the change-of-variable formulas. An important particular case is:

Corollary

For any random variable X (discrete or continuous), and any $a, b \in \mathbb{R}$,

$$E[aX + b] = aE[X] + b.$$

Quantiles, median

Definition

The p-th *quantile* of a random variable X, where p is a number between 0 and 1, is the smallest number q_p such that

$$P(X \leq q_p) = p.$$

Sometimes mean is not an adequate characteristic of a random variable. For example, the mean of the yearly income per household in a given country would exhibit values much higher then "expected", due to a relatively small number of embarrassingly wealthy individuals. In such cases, a more adequate representation of a "mean" value would be given by *median* which is defined as the 0.5th quantile. Informally, the median is the value which "sits in the middle", and it is much less sensitive than mean to extreme values in the data.
Another frequently used in practice quantiles are *quartiles*, which are defined as 0.25th, 0.50th, and 0.75th quantiles.

Variance and standard deviation

Definition

The *variance* of a random variable X, denoted by $Var(X)$, is the number $E[(X - E[X])^2]$.

Variance signify how “spread”, around the mean, is the random variable.

Theorem 1

For any random variable X, $Var(X) = E[X^2] - E[X]^2$.

Proof goes separately for discrete and continuous distributions.

Theorem 2

For any random variable X, and any $a, b \in \mathbb{R}$, $Var(aX + b) = a^2 Var(X)$.

Proof uses definition of variance and Theorem 1.

Definition

The *standard deviation* of a random variable X (both discrete and continuous), denoted by $\sigma(X)$, is defined as $\sigma(X) = \sqrt{Var(X)}$.

Example

If X is a discrete random variable attaining, with the equal probability $\frac{1}{n}$, a finite number of n distinct values $x_1, \ldots, x_n$, then

$$E[X] = \frac{x_1 + \cdots + x_n}{n}$$

and

$$\sigma(X) = \sqrt{\frac{(x_1 - E[X])^2 + \cdots + (x_n - E[X])^2}{n}}.$$

The latter formula explains why indeed the standard deviation is a good measure of how spread the data is: the more the values x_i stay away from their mean $E[X]$, the bigger $\sigma(X)$ would be.

For more examples, see Graham–Knuth–Patashnik, pp. 387–394.

8. Basic types of discrete distributions: uniform, binomial, Poisson, hypergeometric

Uniform distribution

Some types of distributions are of utmost importance, as they appear often on practice, and provide a convenient material for building effective statistical models.

Perhaps, the simplest possible distribution is a uniform one.

Definition

A discrete random variable is distributed *uniformly*, if its mass function attains the same value at the finite number of n points.

In what follows, we assume the number n to be fixed.

Lemma

The mass function of a uniform distribution is of the form

$$f_n(k) = \frac{1}{n},$$

where $k = 1, 2, \ldots, n$.

Uniform distribution (cont.)

Lemma

The expected value and the standard deviation of an uniformly distributed random variable are equal to $\frac{n+1}{2}$ and $\sqrt{\frac{n^2-1}{12}}$, respectively.

Example

Our favorite random variable examples of throwing a single (fair) dice, or tossing a single (fair) coin are uniformly distributed.

Binomial distribution

Definition

The *binomial distribution* with parameters $n = 1, 2, \dots$ and p, where $0 \leq p \leq 1$, is the discrete distribution of the number of successes in a sequence of n experiments with a binary outcome (success/failure), each of which yields success with probability p.

Lemma

The mass function of the binomial distribution has the form

$$f_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where $k = 0, 1, 2, \dots, n$.

Example

Suppose that a biased coin comes up heads with probability 0.3. Then the probability to have 4 heads after 6 tosses is equal to

$$f_{6,0.3}(4) = \binom{6}{4} \times 0.3^4 \times (1-0.3)^{6-4} \approx 0.0595$$

Binomial distribution (cont.)

Lemma

The expected value of a binomially distributed random variable is equal to

$$\sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np,$$

and the standard deviation is equal to

$$\sqrt{np(1-p)}.$$

Poisson distribution

The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

Definition

The *Poisson distribution* is defined as a discrete distribution with parameter $\mu > 0$ (estimated number of events) whose mass function has the form

$$f_\mu(k) = \frac{\mu^k}{k!} e^{-\mu},$$

where $k = 0, 1, 2, \ldots$.

Poisson distribition (cont.)

Lemma

The expected value of a random variable whose distribution function is Poisson, is equal to

$$\sum_{k=0}^{\infty} k \frac{\mu^k}{k!} e^{-\mu} = \mu,$$

and the standard deviation is equal to $\sqrt{\mu}$.

Hypergeometric distribution

Definition

The *hypergeometric distribution* with parameters N, K, n, where N is a non-negative integer, and K and n are integers ranging from 0 till N, describes the number of successes in n binary (success/failure) draws, without replacement, from a finite set of N elements, that contains exactly K successes.

Lemma

The mass function of the hypergeometric distribution has the form

$$f_{N,K,n}(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where $k = 0, 1, 2, \ldots, \min(n, K)$.

Hypergeometric distribution (cont.)

Lemma

The expected value of the hypergeometric distribution is equal to

$$\frac{nK}{N},$$

and the standard deviation is equal to

$$\frac{1}{N}\sqrt{\frac{nK(N-K)(N-n)}{N-1}}.$$

9.
Basic types of continuous distributions: uniform, normal, exponential

Continuous uniform distribution

Again, among continuous distributions the continuous uniform distribution has the most simple form: its density function is a constant within a given range. More precisely:

Definition

The density function of the *continuous uniform distribution* on the interval $[a, b]$ is defined as

$$f_{a,b}(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

Lemma

The expected value and the standard deviation of an uniformly distributed continuous random variable are equal to $\frac{a+b}{2}$ and $\frac{b-a}{2\sqrt{3}}$, respectively.

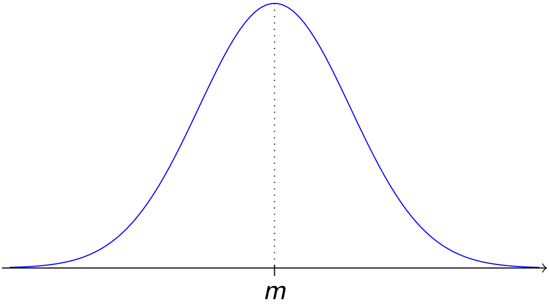
Normal distribution

Definition

The *normal distribution* with parameters m and σ is a continuous distribution with the density function of the form

$$f_{m,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}.$$

The graph of this density function has the famous "bell-shaped" form, with the maximum around $x = m$:



Normal distribution (cont.)

Normal distributions are sometimes called *Gaussian distributions*, in honor of Carl Friedrich Gauss (1777–1855):

Normal distribution (cont.)

The normal distribution is, perhaps, the single most important distribution, due to the Central Limit Theorem, one of the cornerstones results in probability and statistics. Roughly, this theorem says that, under certain natural conditions, the average of a large number of identically distributed random variables is distributed normally, no matter what the initial distribution was. This is the reason why normally distributed random variables appear so often on practice.

Lemma

The expected value of a normally distributed random variable is equal to

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} te^{-\frac{1}{2}(\frac{t-m}{\sigma})^2}\,\mathrm{d}\,t = m,$$

and the standard deviation is equal to σ.

Exponential distribution

The exponential distribution describes the time between events in a process in which events occur continuously and independently at a constant average rate $\lambda > 0$. More formally:

Definition

The *exponential distribution* is the continuous distribution with the density function of the form

$$f_\lambda(x) = \begin{cases} 0 & \text{if } x < 0, \\ \lambda e^{-\lambda x} & \text{if } x \geq 0. \end{cases}$$

Lemma

The expected value of an exponentially distributed random variable is equal to

$$\lambda \int_0^\infty t e^{-\lambda t} \, \mathrm{d}\, t = \frac{1}{\lambda},$$

and the standard deviation is equal to $\frac{1}{\lambda}$ too.

10.
Sum of random variables, covariance, correlation

Sum of random variables, covariance, correlation

Theorem

For any random variables $X_1, \dots, X_n$, and any $a_1, \dots, a_n, b \in \mathbb{R}$,

$$E[a_1 X_1 + \cdots + a_n X_n + b] = a_1 E[X_1] + \cdots + a_n E[X_n] + b.$$

In particular, $E[X + Y] = E[X] + E[Y]$.

As an application of Theorem, one may derive the formula $E[X] = pn$ for a binomial distribution with parameters p, n, without evaluating of the corresponding combinatorial sums: indeed, a binomially distributed random variable X can be represented as $X = X_1 + \cdots + X_n$, where each X_i is a random variable taking the value 1 with probability p, and 0 with probability $1 - p$. Thus $E[X_i] = p$, and

$$E[X] = p + \cdots + p \,(n \text{ times}).$$

Covariance

Theorem

For any two random variables X, Y,

$$Var(X+Y) = Var(X) + Var(Y) + 2E[(X-E[X])(Y-E[Y])].$$

The "extra" term (up to factor 2) is what is called covariance, and express the way X and Y "influence" each other

Definition

The *covariance* of two random variables X, Y, denoted by $Cov(X,Y)$, is defined as

$$E[(X-E[X])(Y-E[Y])].$$

Theorem

For any two random variables X, Y,

$$Cov(X,Y) = E[XY] - E[X]E[Y].$$

Dependence vs. correlation

Definition

Two random variables X, Y are called *uncorrelated*, if $Cov(X, Y) = 0$.

Theorem

If X and Y are independent random variables, then $E[XY] = E[X]E[Y]$.

Corollary

If two random variables are independent, then they are uncorrelated.

Warning

The opposite is not true!

For an example, see DEKKING ET AL., pp. 141–142.

Correlation

Definition

The *correlation* between two random variables X, Y, denoted by $Cor(X, Y)$, is defined as

$$\frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

Theorem

For any two random variables X, Y:

1. $Cor(X, X) = 1$
2. $Cor(X, Y) = Cor(Y, X)$
3. $-1 \leq Cor(X, Y) \leq 1$

11.
Data analysis. Statistical models

Linear regression

Suppose we have two numerical datasets, $x_1, \dots, x_n$, and $y_1, \dots, y_n$, and we want to find how y_i depends on x_i. This is the task for *statistical models*. In the simplest case, we assume that the relationship is linear, modulo errors of measurement.

Definition

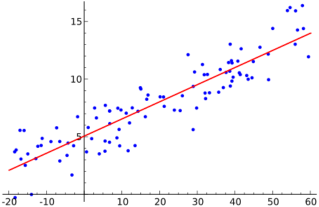
In a *linear regression* model for a bivariate dataset $(x_1, y_1), \dots, (x_n, y_n)$, we assume that $x_1, \dots, x_n$ are nonrandom, and that $y_1, \dots, y_n$ are realizations of random variables $Y_1, \dots, Y_n$ satisfying

$$Y_i = \alpha + \beta x_i + U_i$$

for $i = 1, \dots, n$, where $U_1, \dots, U_n$ are independent random variables with $E[U_i] = 0$ and $Var(U_i) = \sigma^2$.

Linear regression (cont.)

The line $y = \alpha + \beta x$ is called the *regression line*.



For example, see DEKKING ET AL., p. 258.

The regression line could be found with the *method of least squares* (first developed by Gauss): which line satisfies the condition that the sum of squares of *residuals*, i.e. the differences $\alpha + \beta x_i - y_i$, is minimal? This boils done to the standard problem from analysis of minimization of the real function in two variables α, β. See DEKKING ET AL., pp. 329–331 for details and examples.

In more complicated statistical models, nonlinear functions may be used.

12.
Hypothesis testing. Null and alternative hypotheses

Hypotheses testing. Null and alternative hypotheses

The more sophisticated methods of choosing a suitable statistical distribution, and estimation of the distribution parameters are usually performed in the framework of *hypothesis testing*. Hypothesis testing is also used for establishing a relationship (or lack thereof) between two datasets, or, more generally, in deriving any kind of statistical observation about one or more datasets.

Usually, this is done by specifying two rival and mutually exclusive hypotheses, the *null* and *alternative hypotheses*, and their subsequent comparison by certain statistical procedures.

Null and alternative hypotheses (cont.)

There is no rule of thumb how null and alternative hypotheses should be formed. However, the usual statistical practice stipulates that the null hypothesis states that the phenomenon being studied produces no effect or makes no difference. The null hypothesis is also usually the hypothesis one wants to reject, or “nullify”. For example, when investigating relationship between two datasets, the null hypothesis should state that there is no relationship at all, while the alternative hypothesis should indicate the existence of such relationship. Or, say, when investigating the impact of smoking on lung cancer, the null hypothesis would state that smoking does not have any impact.

See DEKKING ET AL., pp. 373–374 for more examples.

Type I and type II errors. *p*-values

A *type I error* is the (incorrect) rejection of a true null hypothesis. The probability of type I error is called the *significance level* of a test. A *type II error* is the (incorrect) acceptance of a false null hypothesis. The probability of not making a type II error, i.e. the (correct) rejection of a false null hypothesis, is called the *power* of a test. The probabilities of making type I and type II errors are traded off against each other: for any given sample set, the effort to reduce one type of error generally results in increasing the other type of error. For a given test, the only way to reduce both error rates is to increase the sample size, and this may not be feasible.

One of the most used test statistics in hypothesis testing is *p-value*, which is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis H_0 is true. What is "more extreme" and how it is measured, depends on the context.

p-values (cont.)

For a “double-tailed” event, the p-value of a random variable X assuming values “more extreme” than x (the observed value), might be defined as

$$2 \cdot \min\{\, P(X \geq x \mid H_0),\ P(X \leq x \mid H_0) \,\},$$

while for “left-tailed” events the same value might be defined as

$$P(X \leq x \mid H_0),$$

and similarly for “right-tailed” ones.

The p-value measures statistical significance of the test, but it should not be confused with the probability of the hypothesis being true, the probability of observing the given data, etc. p-values are often misused and misinterpreted.

Examples

In the flipping coin example, suppose that the null hypothesis specifies that the coin is fair. In a double-tailed model, the alternative hypothesis would be that the coin is biased either way, while in an one-tailed model the alternative hypothesis says that the coin is biased towards, say, heads. Suppose that one gets 5 heads in a row in one experiment. In the one-tailed model this is the most extreme possible value, with a *p*-value equal to

$$\left(\frac{1}{2}\right)^5 = \frac{1}{32} \approx 0.03\,.$$

In the double-tailed model, the corresponding *p*-value would be twice as that:

$$2 \cdot \left(\frac{1}{2}\right)^5 = \frac{1}{16} \approx 0.06\,.$$

One frequently sets 0.05 as the threshold for the *p*-value of the test to be statistically significant. Under this assumption, we should reject the null hypothesis in the first case, while we cannot do that in the second one.

Hypotheses testing (cont.)

An essentially equivalent procedure, but not using the concept of a *p*-value, would run as follows:

1. Choose a test statistics (for example, just the number of heads in the flipping coin example);
2. Derive the distribution of the test statistics under the null hypothesis (the binomial distribution in our example);
3. Select the significance level of the test (the common values are 0.05 and 0.01);
4. Determine the *critical* (or *rejection*) *region* – the values of the test statistics for which the null hypothesis is rejected;
5. Perform the test, derive from it the empirical value of test statistics, and see whether it falls into the critical region or not.

Confidence interval

There is a direct relationship between the critical region and the confidence interval. The *confidence interval* of a certain statistical parameter is the interval, calculated from the sample, that contains the specified value of the parameter with the specified probability. A typical situation when this notion occurs naturally is estimation of the average of the mean of identically distributed random variables. For example, if a certain random variable is normally distributed with the same mean μ and standard deviation σ, then it is known that the average $\overline{x}$ of n observations is normally distributed around μ with standard deviation $\frac{\sigma}{\sqrt{n}}$. A 95% confidence interval for μ is determined then as

$$\overline{x} + N_{0.025}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + N_{0.975}\frac{\sigma}{\sqrt{n}},$$

where $N_{0.975} \approx 1.96$ and $N_{0.025} = -N_{0.975}$ are the 97.5% and 2.5% quantiles in the standard (i.e. with parameters $\mu = 0$ and $\sigma = 1$) normal distribution.

Confidence interval (cont.)

Now, suppose that for some parameter θ and its value θ_0, we test the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$ (one-tailed test). Then we reject H_0 in favor of H_1 (i.e., θ is *not* in the critical region) at the significance level α if and only if θ_0 is not in the $100(1 - \alpha)\%$ one-tailed confidence interval for θ. A similar statement is true in the case of a double-tailed test.

The End