B34.UC2

Numerical Computation and Statistics in Engineering

Unit 1: Statistics - An Introduction

What Is Statistics?

Statistics is about collecting, presenting, and characterizing information to assist in data analysis and decision-making.

- *Descriptive statistics* is involved with the collection, presentation, and characterization of data sets to simply describe properties of a population.
- *Inferential statistics* aims to make inferences about a population based on information contained in a sample.

Here *population* is the set of data of our interest, and a *sample* is any selected subset of the population.

Typical areas of applications of statistics are business, science, politics, etc.

Examples

Example. A supermarket needs to know how many cashiers it needs so that on average 90% of their customers wait no longer than 2 minutes. □

Example. A bank wants to better serve their clients. It sends out a questionnaire to 2000 randomly selected clients with questions about their banking habits and their use of computers. Depending on the outcome of these questionnaires the bank has to decide whether to invest more into online banking or not. □

Example. A manufacturer of screws makes special screws for a customer. From every lot of 1000 screws the manufacturer wants to select screws randomly to check whether they match the customer's specifications or not. How many screws from each lot should be tested to be 98% confident that all screws in that lot meet the specifications? □

Examples

Example. In an experiment a scientist measures the speed of light. Even though theory tells her that there is only one actual value for the speed of light, she finds slightly different values in each trial due to external factors and inaccurate equipment. How should she estimate the speed of light given this set of data? □

Example. An airline wants to maximize profit and sell as many seats as possible. However, due to 'no-shows' seats often remain empty. Therefore many airlines overbook their planes. But now it can happen that passengers have to stay behind or have to be booked on planes of other companies.

How should the airline estimate the number of no-shows to plan the optimal number of reservations, that is, which number of reservations will maximize the number of filled seats, but minimizes the number of passengers with reservations who get denied. □

Types of Data

Quantitative data are data that represent an amount or a quantity. These can be discrete data or continuous data:

- number of books bought this month;
- height;
- weight, etc.

Qualitative data (also called categorical data) are data that have no quantitative interpretations:

- your favorite author;
- the grocery store where you do your weekly shopping;
- the names of the computer stores listed in the directory, etc.

Graphical Methods For Describing Data

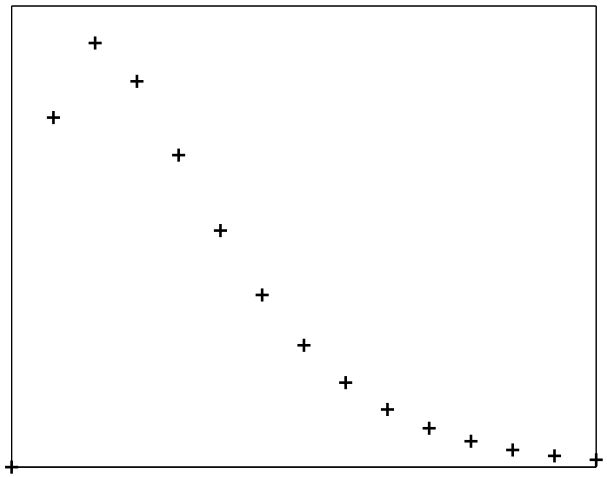
There are lots of different ways to represent quantitative data, for example tables, vertical or horizontal bar graphs, pie charts, scatter diagrams, etc. The following examples each show the money available for research at U.K. universities during the academic year 1998/99 (source: The Times Higher Education Supplement) in 1000 £:

Funding councils	1011835	Research councils	599606
Other UK government	316413	UK charities	429163
UK industry	221188	EU government	155435
Other	33598	Other EU	28218
Other overseas	91071		

Total: 2886527m £.

Numerical Values Associated to Samples

We consider a *relative frequency table* (also called the *relative frequency distribution*) of a sample.

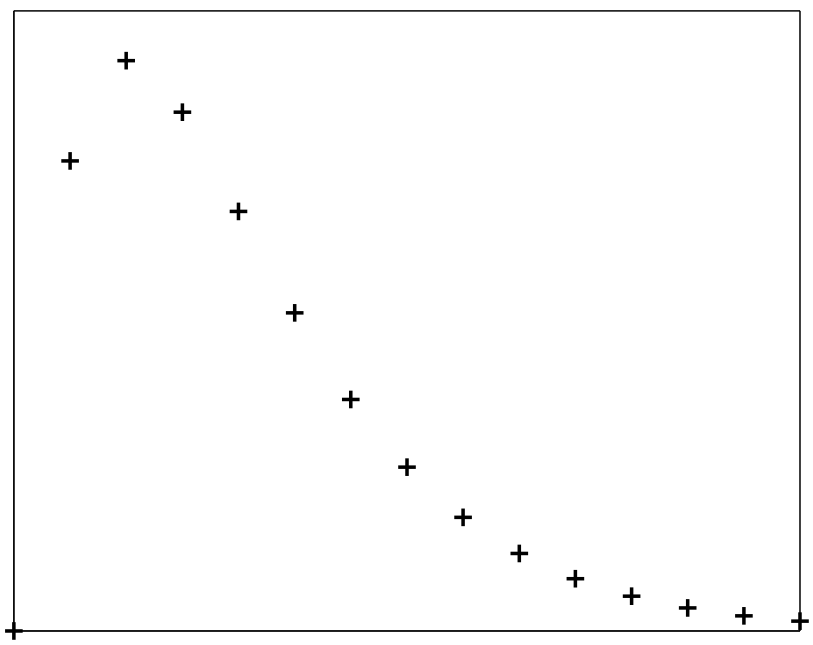


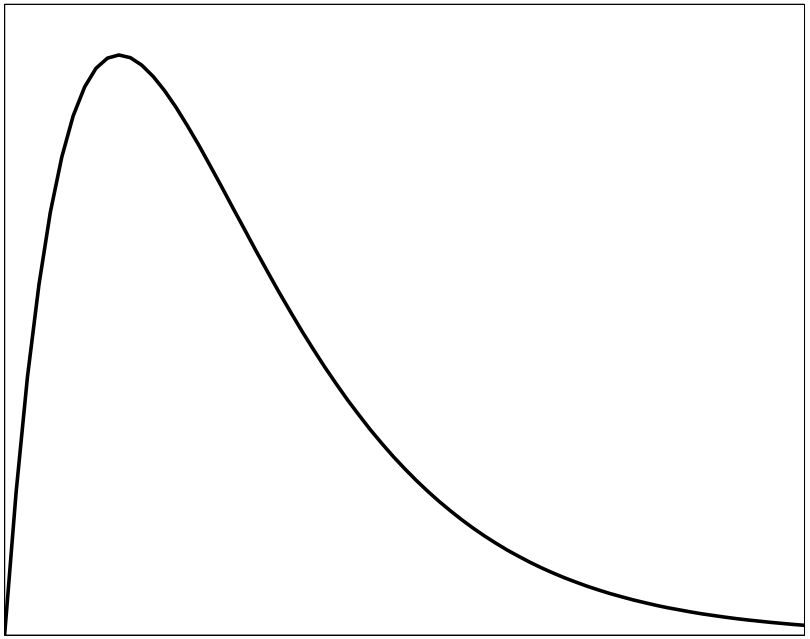
A variety of numerical values can be associated to a sample. These values aim

- to help locate the center of the relative frequency distribution of the data (arithmetic mean, median, mode); and
- to measure how the data is spread (range, variance, standard deviation).

Measures of Central Tendency

We consider a sample or an experiment where we made successively the observations $x_1, \ldots, x_n$. Arranged in a relative frequency table the data may look as on the left, or as sketched on the right:





Measures of Central Tendency

The *arithmetic mean* $\bar{x}$ of the sample $x_1, \ldots, x_n$ is the average

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

The *median* of the sample $x_1, \ldots, x_n$ is the middle number when the values are arranged in ascending or descending order.

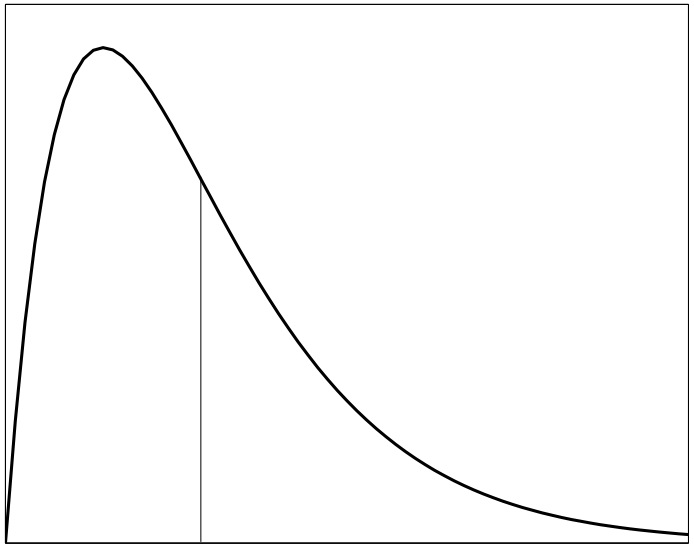
The *mode* of the sample $x_1, \ldots, x_n$ is the value which occurs with greatest frequency.

Example. Consider the measurement

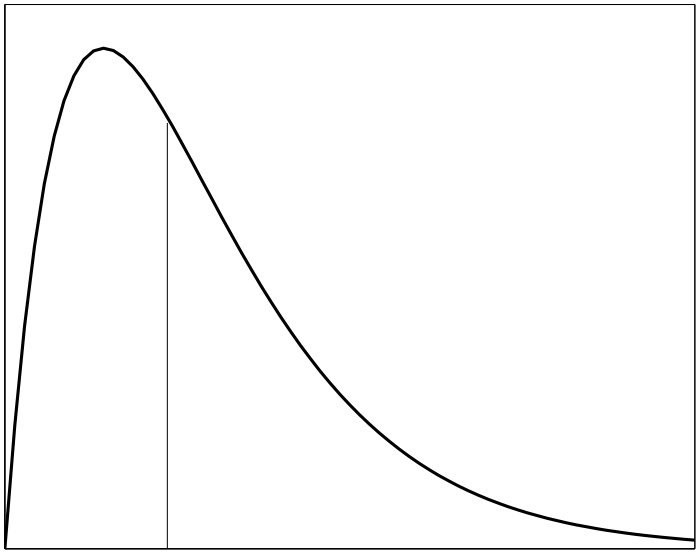
$$1, 2, 2, 2, 2, 3, 3, 4 .$$

The arithmetic means is $\bar{x} = \frac{19}{8}$, the median is $m = \frac{1}{2}(2 + 2) = 2$, and the mode is also 2. □

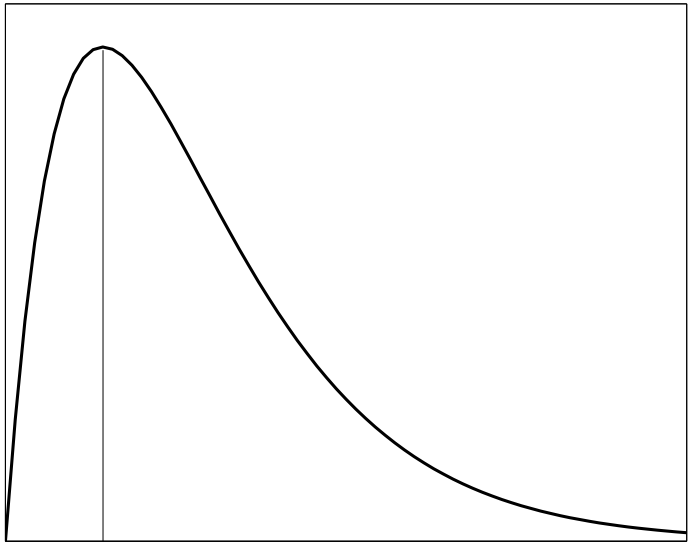
Measures of Central Tendency



mean (point of balance)



median



mode (peak point)

Measures of Variation

The *range* of a sample is the difference between the largest and smallest values in the sample.

The *variance* of the sample $x_1, \ldots, x_n$ is

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1} = \frac{\sum_i x_i^2 - \frac{1}{n}(\sum_i x_i)^2}{n-1}\,.$$

The *standard deviation* of a sample is the square root of the variance.

Example. With the data from the previous example continued the range is $4 - 1 = 3$, and the variance is

$$\begin{aligned} s^2 &= \frac{1}{7}(1 + 4 \cdot 2^2 + 2 \cdot 3^2 + 1 \cdot 4^2 - \frac{1}{8} \cdot 19^2) \\ &= \frac{1}{7}(49 - \frac{361}{8}) \\ &\approx 0.0554 \end{aligned}$$

□

The variance measures (almost) the arithmetic mean of the (square of

Measures of Variation

the) distance of the values in the sample $x_1, \ldots, x_n$ from the arithmetic mean $\bar{x}$.

Note that we have to square, that is, that we cannot consider $\frac{1}{n}\sum_i (x_i - \bar{x})$: A simple calculation shows that

$$\sum_i (x_i - \bar{x}) = \sum_i x_i - \sum_i \bar{x} = n\bar{x} - n\bar{x} = 0\,,$$

and $\frac{1}{n}\sum_i (x_i - \bar{x}) = 0$ does not contain any information.

The Role of the Standard Deviation

The variance of a sample is primarily of theoretical interest, but the standard deviation has a clear meaning:

Theorem. (Tchebysheff's Theorem) *For a sample of size n and $1 \leq k \leq n$, at least $1 - \frac{1}{k^2}$ of the observations lie within k standard deviations of their mean.*

The following empirical rule is also useful. It holds for mound-shaped (bell-shaped) frequency distributions of samples:

- Approximately 68% of the observations will lie within 1 standard deviation of their mean.
- Approximately 95% of the observations will lie within 2 standard deviations of their mean.
- Almost all observations will lie within 3 standard deviations of their mean.

Measures of Relative Standing

Some type of data (scores, health data) is often reported in a manner that describes their position *relative to other* data.

The *100pth percentile* of a data set is the value x of possible outcomes such that 100p% of the area of the relative frequency table lies left to the 100pth percentile.

The *lower quartile* Q_L for a data set is the 25th percentile, the *mid-quartile* m for a data set is the 50th percentile, and the *upper quartile* Q_U for a data set is the 75th percentile.

Example. If your mark in the statistics module is located on the 72th percentile then 72% of the students in your class had lower marks than you, and 28% had higher marks. □

An Example

The following table shows the results of an exam with 52 students:

0	0	0	0	0	0	40	41	42	50
50	52	60	60	60	61	62	63	63	63
63	65	66	67	68	70	70	71	71	72
72	73	74	75	75	77	80	80	80	80
80	81	81	81	81	82	87	87	88	90
94	95								

The 6 values 0 come from the fact that some students enrolled in the class, but didn't show up for the class or the exam. Such values are called *outliers*.

An Example

For this data set we get the following values:

$$\begin{aligned} \text{mean } \bar{x} &= 62.37 \\ \text{range} &= 95 \\ \text{median} &= 70 \\ \text{mode} &= 0 \\ \text{variance } s^2 &= 676.04 \\ \text{standard deviation } s &= 26.00 \end{aligned}$$

Out of the 52 data 40 lie in the interval $[\bar{x} - s, \bar{x} + s]$, which is 77%. Out of the data, 46 lie in the interval $[\bar{x} - 2s, \bar{x} + 2s]$, which is 88%. Note that the frequency distribution is *not* bell-shaped, but right-skewed.

An Example

As we have seen before, the first 6 values are clearly outliers. What would have been the results if we had removed these outliers?

$$\begin{aligned} \text{mean } \bar{x} &= 70.5 \\ \text{range} &= 95 \\ \text{median} &= 71.5 \\ \text{mode} &= 81 \\ \text{variance } s^2 &= 179.9 \\ \text{standard deviation } s &= 13.41 \end{aligned}$$

We can see that the mean and mode are not very robust to the outliers while the median is. Dealing with outliers is the branch of statistics called Robust Statitics and involves techniques such as RANSAC. This will not be developped in this course. Just remember that these techniques exist and that they normally give more robust but less precise estimations.

The Role of Statistics in Science

Experimental research, whether in engineering, life sciences, information science, business, or other sciences, involves experimental data. From this data the scientists derives properties of the whole population. Since the size of the whole population can be very large this is often the only possibility to gain insight into properties of the population.

However, this process of inference almost always involves an error. For example, a sample of 100 potential customers of a new product contains 25 people in favor of the new product, whereas a second sample of again 100 potential customers contains 32 people in favor of the new product. Hence, there is always *uncertainty* about the actual property.

Statistics provides scientific tools to enable such inferences with a probability of certainty, that is, provides methods to judge the reliability of such inferences. Statistics are also about predicting under uncertainty (Tracking of a missile) providing models of the underlying (random) processes encountered in physics, economics,...

Introduction to Probability Theory

Probability theory deals with the situation where the whole population is known. We calculate the likelihood that a particular sample is randomly selected from that population.

Probability theory plays some role in decision-making. If the introduction of the three previous new products of a company (say, new computer software) was a flop, would you invest in this company shortly before it launches its new product? If you are playing blackjack in a casino and the bank draws 3 blackjacks in a row, do you believe that the deck of cards is well-shuffled, or that the game is fair?

Objective Versus Subjective Probability

Very general, probability refers to the chance or likelihood that a particular event will occur.

We distinguish between *subjective* and *objective* probability. Examples of the former are the chances that Celtic will win the next Champions League, or that it will rain tomorrow. Examples of the latter are

- the probability that a thrown fair die will show 1 (which is $\frac{1}{6}$), and similarly for every other possible outcome;
- the probability that in a shuffled deck of 52 cards the top card is an ace (which is $\frac{1}{13}$).

Objective probability can be defined as

$$\frac{\text{number of outcomes}}{\text{total number of possible outcomes}}.$$

Objective Versus Subjective Probability

However care have to be taken when using this formula as:

- **It is only valid under the assumption that all outcomes are equally likely.**
- **This is the classical definition of probability where the probability of an event can be calculated a-priori. For experimental data, the number of possible outcome might not be known and the probability needs to be estimated from samples as we shall see later in the course. Classical probablities can then be used as working hypothesis.**

Example. Find the probability that 2 dices add to 7. Lets consider as the number of outcomes the 11 possible sums 2,3,4,...12. The probability of having 7 is then $P = \frac{1}{11}$ according to the definition. Any problem? This is wrong as the right answer is 1/6. Why? because the 11 sums are **not equally likely**. □

Events And Sample Space

Before we see some more examples let us introduce some new concepts. Each possible outcome of an experiment or an observation is called a *simple event*. An *event* is any possible outcome. The collection of all simple events is called the *sample space*.

Example. Let us consider again the example of throwing a die. A typical simple event is '1'. Another event is 'even' $= \{2, 4, 6\}$. The sample space consists of all possible outcomes, that is, of the set of numbers $\{1, 2, 3, 4, 5, 6\}$. □

Probability now refers to the probability of occurrence of an event. The notation

$$P(A)$$

denotes the probability that event A occurs.

Compound Events

The *union* of two events A and B, in symbols, $A \cup B$, is the event that either A or B occurs. The *intersection* of the two events A and B, in symbols $A \cap B$, is the event that both A and B occur.

Example. (Rolling a die.)
Let A be the event 'even' $= \{2, 4, 6\}$, and B the event 'divisible by 3' $= \{3, 6\}$. Then $A \cup B = \{2, 3, 4, 6\}$ and $A \cap B = \{6\}$. □

Example. (Tossing a fair coin twice.)
Let A be the event 'H in the first toss' and B be the event 'T in the second toss'. Then A and B is the event $\{\text{HT}\}$, whereas A or B is $\{\text{HH, HT, TT}\}$. □

It is important to note that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)\,.$$

Complementary Events

The *complement* of an event A, in symbols A^{C} or $\complement A$, is the union of all simple events not contained in A.

Example. (Rolling a die.)
If A is the event 'even', then A^{C} is the event 'odd'. Note that $P(A^{\mathrm{C}}) = 1 - P(A) = 1 - \frac{1}{2} = \frac{1}{2}$. □

It holds in general that

$$P(A^{\mathrm{C}}) = 1 - P(A)\,.$$

Summary of the Rules

- $0 \leq P(A) \leq 1$.
- $P(B_1 \cup B_2 \cup \cdots \cup B_k) = 1$,
  if the B_i are a collection of exhaustive events, that is, if at least one of the B_i must occur.
- $P(A) = \sum_{i=1}^{n} P(A \cap B_i)$,
  if the B_i are a collection of collectively exhaustive and mutually exclusive events, that is, if at least one of the B_i must occur, but not two of them can occur at the same time.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- $P(A^{\mathrm{C}}) = 1 - P(A)$.

Conditional Probability

Example. *(Tossing a die.)*
The probability $P(\text{'even'})$ is $\frac{1}{2}$. But suppose we know that the outcome was greater or equal than 4. Since this reduces the possible outcomes to $\{4, 5, 6\}$ it seems reasonable to bet with probability $P = \frac{2}{3}$ that the outcome is even, *since we know* that the outcome was either '4', '5', or '6'. □

The *conditional probability* that event A occurs given that B occurs is given by

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

This formula holds in general, but is easiest motivated through counting. Suppose that we have n simple events, out of which b are in B, and $c \leq b$ are in $A \cap B$. Then, by definition,

$$P(A \mid B) = \frac{c}{b} = \frac{\frac{c}{n}}{\frac{b}{n}} = \frac{P(A \cap B)}{P(B)}\,.$$

Conditional Probability

Example. A chip manufacturer sends large numbers of micro-chips to a customer. The customer makes random checks whether the chips meet his specifications. Suppose S is the event that a lot is shipped to the customer, and F the event that a lot contains faulty chips. Longterm inspections show the following table of probabilities:

$S \cap F^{\mathrm{C}}$	0.85	$S \cap F$	0.02
$S^{\mathrm{C}} \cap F^{\mathrm{C}}$	0.09	$S^{\mathrm{C}} \cap F$	0.04

Then the probability of a lot being send to the buyer is

$$P(S) = P(S \cap F) + P(S \cap F^{\mathrm{C}}) = 0.02 + 0.85 = 0.87\,,$$

and the conditional probability that a sent lot does *not* confirm to the customer's specifications is

$$P(F \mid S) = \frac{P(F \cap S)}{P(S)} = \frac{0.02}{0.087} \approx 0.023\,.$$

Independent Events

Two events A and B are said to be *independent* if the occurrence of B does not effect the occurrence of A, that is, if $P(A \mid B) = P(A)$. Otherwise we say that the events are *dependent*.

Example. *(Tossing a die.)*
We consider the events $A =$ 'even' and $B =$ 'less or equal to 3'. Then $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$, and $P(A \cap B) = \frac{1}{6}$. It follows that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3} \neq P(A)\,,$$

so that the events are dependent. □

Note that when $P(A \mid B) = P(A)$ then

$$P(B)P(A) = P(B)P(A \mid B) = P(A \cap B) = P(A)P(B \mid A)\,,$$

so that in this case we also have $P(B \mid A) = P(B)$.

Independent Events

Example. A manufacturer of hard drives offers a one year guaranty on his products. Analysis of customer complaints resulted in the following table:

	Reason for complaint		
	electrical failure	mechanical failure	Total
during the first year	31%	41%	72%
after one year	14%	14%	28%
	45%	55%	100%

Are the events $A =$ 'complaint during the first year' and $B =$ 'mechanical failure' dependent?

We know from the table that $P(A) = 0.72$, $P(B) = 0.55$, and $P(A \cap B) = 0.41$, thus $P(A \mid B) = \frac{0.41}{0.55} \approx 0.76$, and the events are *dependent*. □

Independent Events

Suppose that the events A and B are independent, then

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

and we get the following *multiplication rule for independent events*:

$$P(A \cap B) = P(A) \cdot P(B)\,.$$

Example. (Throwing a die twice.)
Let A be the event that we first observe 'H', and B be the event that we observe 'T' in the second throw.

Then $P(A) = \frac{2}{4} = \frac{1}{2} = P(B)$, and $P(A \cap B) = \frac{1}{4}$, so that the events are independent as expected. □

Counting

Many of our previous examples involved counting the outcomes relevant for some event and the total number of outcomes. For large sample spaces, however, it is not feasible to list all possible outcomes and we need rules for counting.

Rule1: Suppose we have k sets of elements, n_1 elements in the first set, n_2 elements in the second set, ..., n_k elements in the kth set. Suppose we want to sample k elements, *taking one element of each set*. Then there are

$$n_1 n_2 \cdots n_k$$

different possibilities.

Example. There are $2 \cdot 2 \cdots 2 = 2^{10} = 1024$ simple events (= possible outcomes) of tossing a coin 10 times. □

Counting

Rule 2: If n objects are given then they can be arranged in order in

$$n! = n(n-1)(n-2)\cdots 2\cdot 1$$

different ways. (By definition, $0! = 1$.)

The symbol ! is read *factorial*.

Example. There are $11\cdot 10\cdots 2\cdot 1 = 39916800$ possibilities that 11 students take seats on 11 chairs. □

Rule 3: Given a set of N elements we want to select $n \leq N$ elements of this set in order. Then there are

$$N(N-1)(N-2)\cdots(N-n+1) = \frac{N!}{(N-n)!}$$

possibilities.

Example. A sales agent has 10 different customers in Edinburgh. If she wants to visit 5 of them today, then there are $\frac{10!}{5!} = 10\cdot 9\cdot 8\cdot 7\cdot 6 = 30245$

different possible ways in which order she can visit 5 of her customers today. □

Rule 4: Given a set of N elements, we want to select $n \leq N$ elements of this set *without* regard of the order. Then there are

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

different possibilities.

Example. In the German lottery *6 aus 49* you mark 6 numbers out of the numbers $1, \ldots, 49$. There are

$$\binom{49}{6} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 13983816$$

possible ways to do so. □

Sampling With/Without Replacement

The previous rules contain indirectly also the following two rules of sampling of populations:

Rule 5: *(Random sampling with replacement.)* Given a set of N elements we want to select randomly n elements, returning each selected element back into the population. Then there are

$$N^n$$

possible outcomes.

Rule 6: *(Random sampling without replacement.)* Given a set of N elements we want to select randomly n elements, not returning the samples back into the population. Then there are

$$N(N-1)\cdots(N-n+1)$$

possible outcomes.

Example. There are $52 \cdot 51 \cdot 50 \cdot 49$ possibilities of picking 4 cards (in

order) out of a deck of 52 cards. □

Example. To draw a blackjack the dealer has to deal a card with value 10 ($4 \cdot 4$ possibilities) and an ace.

There are $52 \cdot 51$ different possibilities to draw 2 cards, and $16 \cdot 4 + 4 \cdot 16$ possibilities for drawing a blackjack. The probability of dealing a blackjack is thus $\frac{128}{2652} \approx 4.83\%$. □

Summary

- Statistics is about collecting, presenting and characterizing data and assists in data analysis and decision making.
- Statistics is usually about quantitative data. Often, such data is presented in diagrams.
- Basic analysis of data is about the central tendency of data (mean, median, mode), and about the variance of data (variance, standard deviation).
- Probability refers to the likelihood of a particular event.
- Probability theory is employed when the whole population is known. Often it involves counting the number of possible events.

B34.UC2

Numerical Computation and Statistics in Engineering

Unit 2: Probability Distributions

Random Variables

A *random variable* is a function taking numerical values which is defined over a sample space. Such a random variable is called *discrete* if it only takes countably many values.

Example. A quality control engineer checks randomly the content of bags, each containing 100 resistors. He selects 2 resistors and measures whether they match the specification (exact value plus or minus 10% tolerance). The number of resistors not matching the specification is a discrete random variable.

Another random variable would be the function taking values 0 and 1, for the outcomes that there are faulty resistors in the bag, or not. □

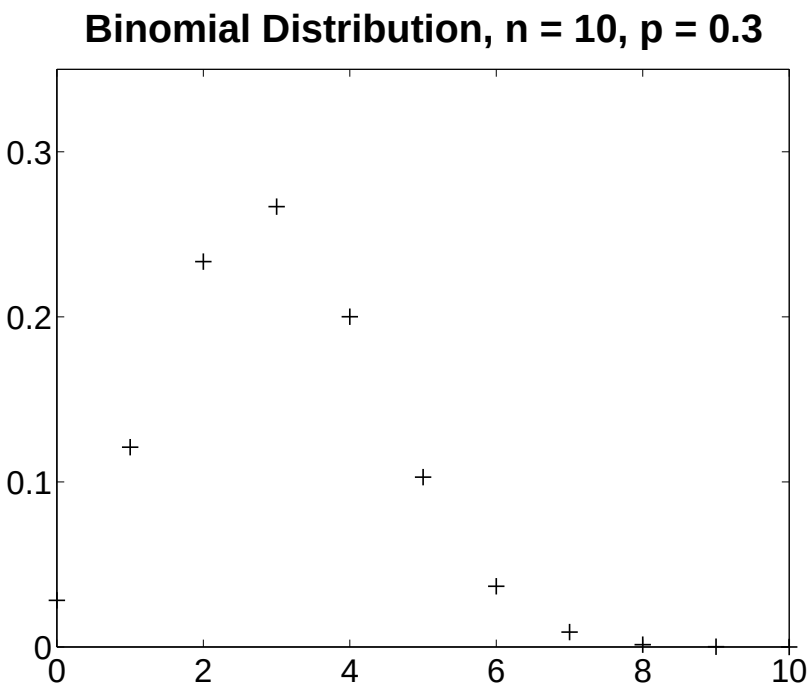
The *probability distribution* of a random variable is a table, graph, or formula that gives for each possible value of the random variable its probability. The requirements are that

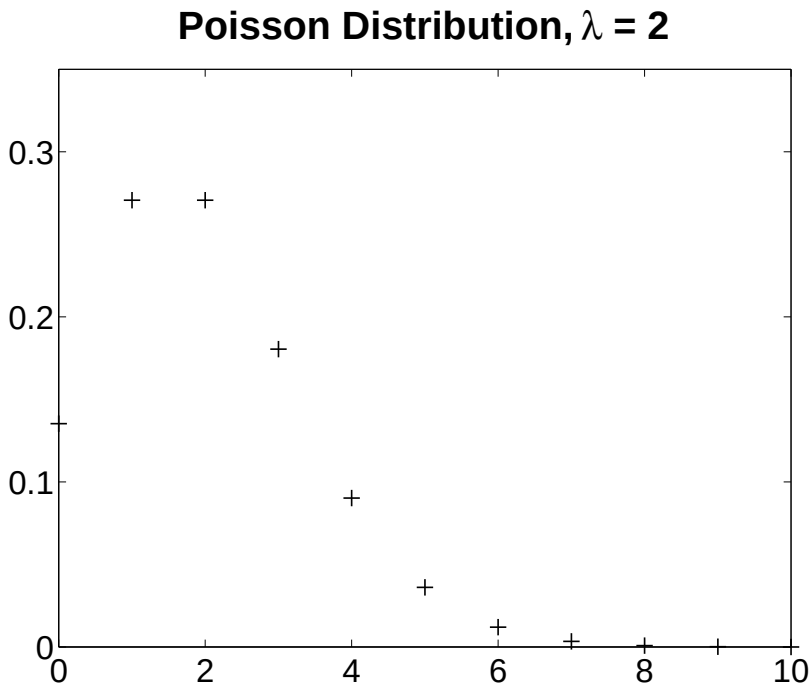
$$0 \leq p(x) \leq 1 \qquad \text{and} \qquad \sum_{\text{all } x} p(x) = 1.$$

The following two diagrams show examples of discrete probability distri-

Random Variables

butions:





Expected Value

For a discrete random variable x with probability distribution $p(x)$ the *expected value* (or *mean*) is defined as

$$\mu = E(x) = \sum_{\text{all } x} x \cdot p(x)\,.$$

Example. We consider throwing a fair die 6000 times. We expect roughly 1000 outcomes of each possible observations 1,. . . , 6. Thus the arithmetic mean of such an experiment will be approximately

$$1\frac{1000}{6000} + 2\frac{1000}{6000} + \cdots + 6\frac{1000}{6000} = 3.5\,.$$

The expected value is $\sum_{i=1}^{6} ip(i) = \frac{1}{6} \cdot 21 = 3.5$, as expected. □

Expected Value

Let x be any discrete random variable with probability distribution $p(x)$, and let g be any function of x. Then the *expected value of* $g(x)$ is defined as

$$E[g(x)] = \sum_{\text{all } x} g(x) \cdot p(x)\,.$$

The *variance* of a discrete random variable x with probability distribution $p(x)$ is defined as

$$\sigma^2 = E[(x-\mu)^2]\,,$$

the *standard deviation* is defined as $\sigma = \sqrt{E[(x-\mu)^2]}$.

Example. We return to the example of throwing a die. For the variance we find

$$\sigma^2 = (1-3.5)^2\frac{1}{6} + (2-3.5)^2\frac{1}{6} + \cdots + (6-3.5)^2\frac{1}{6} \approx 2.917\,.$$

For the standard deviation we find $\sigma \approx 1.708$. □

Properties of the Expected Value

Let x be a discrete random variable with probability distribution $p(x)$.

- $E(c) = c$, for every constant c;
- $E(cx) = cE(x)$, for every constant c;
- $E[g_1(x) + g_2(x)] = E[g_1(x)] + E[g_2(x)]$, for any two functions g_1, g_2 on x.

It follows the important formula that

$$\sigma^2 = E[x^2] - \mu^2 \, .$$

For the proof of this formula note that

$$\begin{aligned} \sigma^2 &= E[(x-\mu)^2] = E[x^2 - 2\mu x + \mu^2] \\ &= E[x^2] - 2\mu E[x] + \mu^2 E[1] \\ &= E[x^2] - 2\mu\mu + \mu^2 \, . \end{aligned}$$

The Binomial Probability Distribution

Example.

- Tossing a coin 10 times.
- Questioning 100 people on Princess Street in Edinburgh whether they know that Madonna's wedding takes place in a Scottish castle.
- Checking whether lots of transistors contain faulty transistors or not.

□

These experiments or observations are all examples of what is called a *binomial experiment* (the corresponding discrete random variable is called a *binomial random variable*).

The Binomial Probability Distribution

The examples have the following common characteristics:

- The experiment consists of n identical trials.
- In each trial there are exactly two possible outcomes (yes/no, pass/failure, or success/failure), denoted here 0 and 1 (for success).
- The probabilities for the outcomes 0 and 1 are the same in each trial (the trials are independent). These probabilities are usually denoted $p = P(\text{'1'})$ and $q = 1 - p = P(\text{'0'})$.
- The discrete (binomial) random variable is the number of successes (i.e., of 1's) in the n trials.

The Binomial Probability Distribution

The *binomial probability distribution* is given by the formula

$$p(x) = \binom{n}{x} p^x q^{n-x}\,, \qquad x \in \{0, \ldots, n\}\,,$$

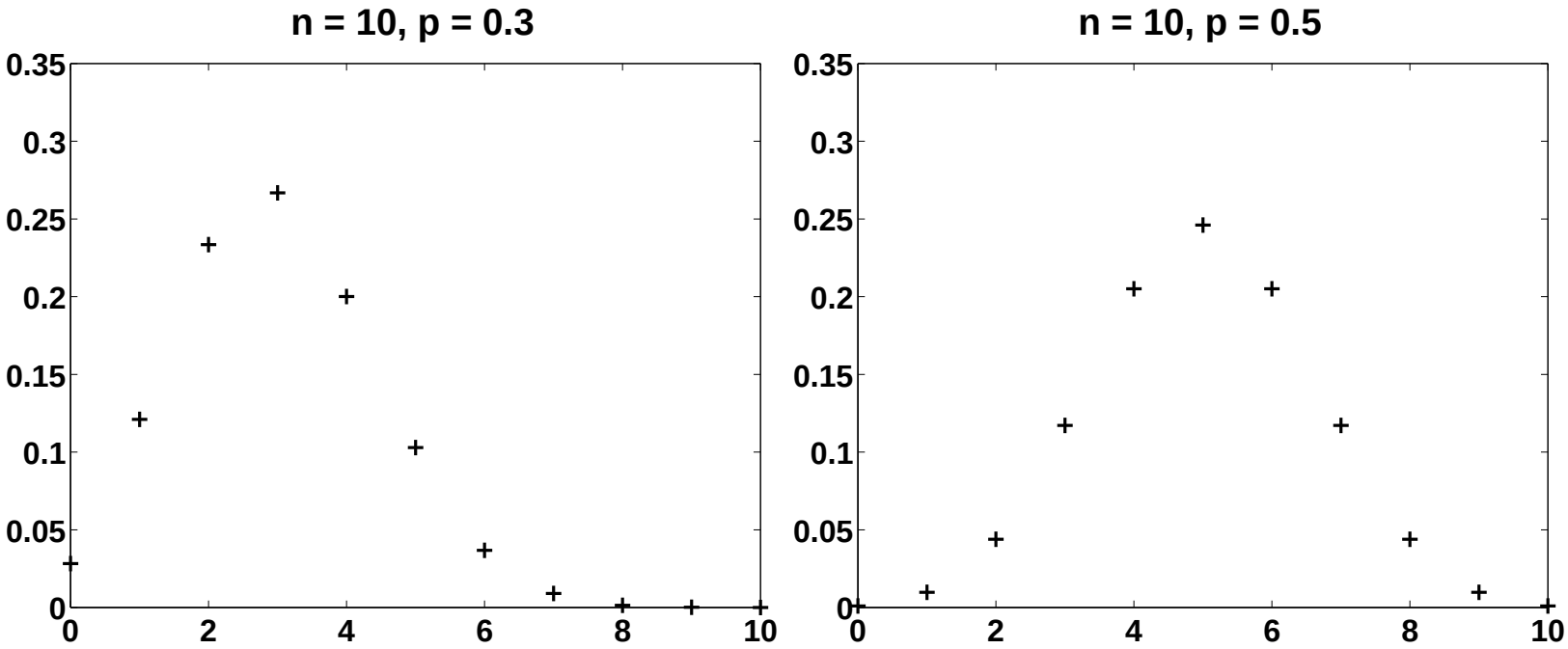
where

- p is the probability of a success in a single trial, and $q = 1 - p$;
- n is the number of trials; and
- x is the number of successes.

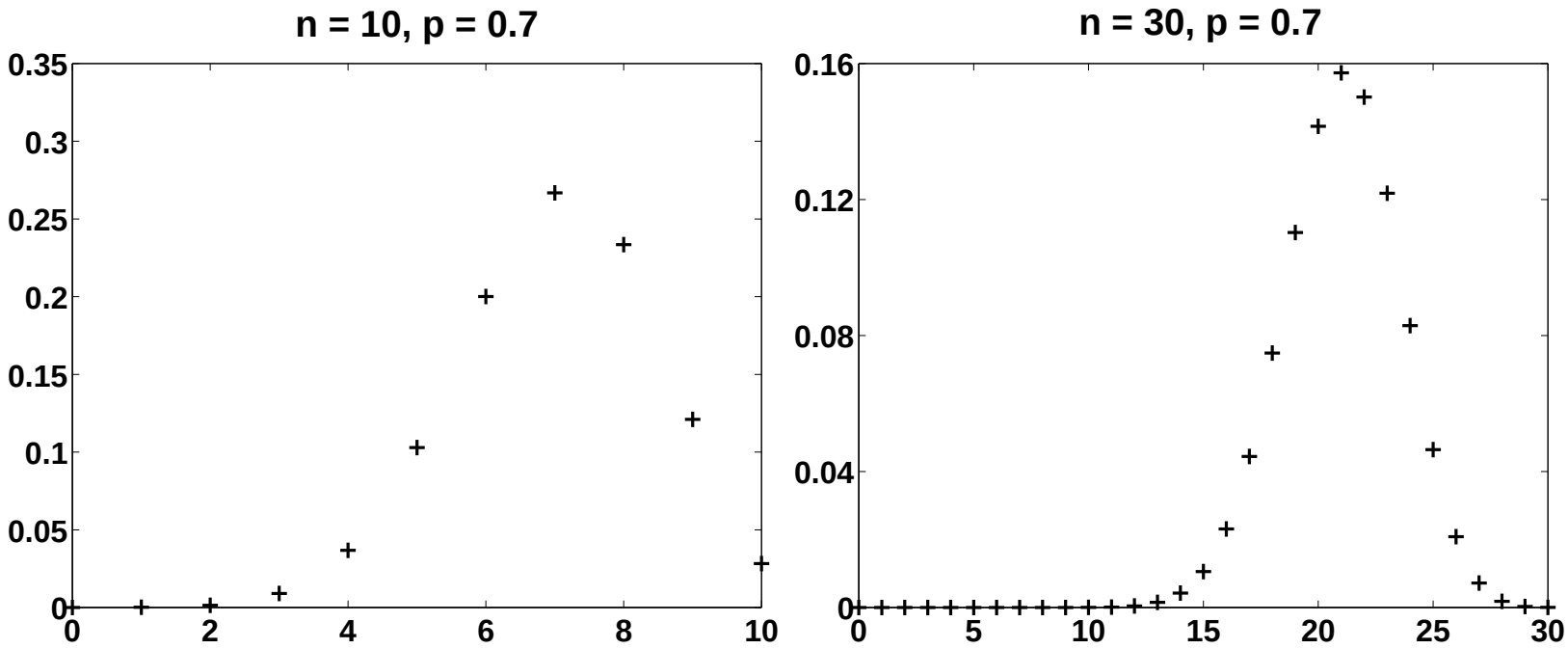
The expected value (mean) and standard deviation are given by

$$\mu = np \qquad \text{and} \qquad \sigma = \sqrt{npq}\,.$$

The Binomial Probability Distribution

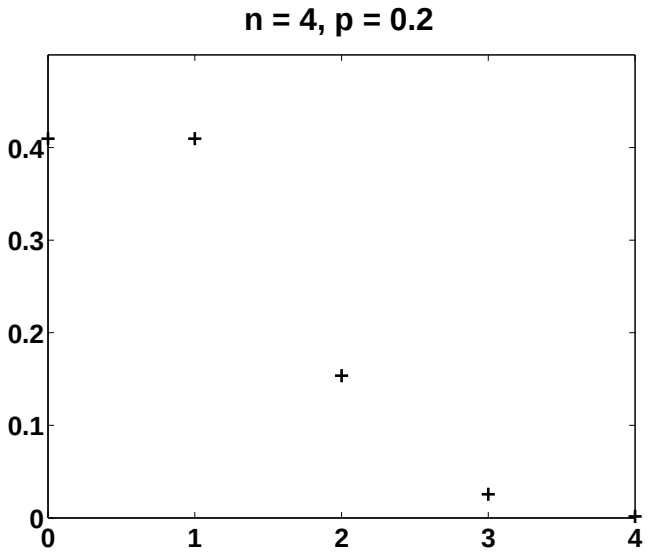


The Binomial Probability Distribution



The Binomial Probability Distribution

Example. Tests show that about 20% of all private wells in some specific region are contaminated. What are the probabilities that in a random sample of 4 wells exactly 2, fewer than 2, or at least 2 wells are contaminated?



Here $n = 4$, $p = 0.2$ (success for being contaminated). We find

$$
\begin{aligned}
P(\text{`}x = 2\text{'}) &= \binom{4}{2} 0.2^2 0.8^{4-2} = 0.1536\,, \\
P(\text{`}x < 2\text{'}) &= P(\text{`}x = 0\text{'}) + P(\text{`}x = 1\text{'}) = \binom{4}{0} 0.2^0 0.8^4 + \binom{4}{1} 0.2^1 0.8^3 = 0.8192\,, \\
P(\text{`}x \geq 2\text{'}) &= P(\text{`}x = 2\text{'}) + P(\text{`}x = 3\text{'}) + P(\text{`}x = 4\text{'}) \\
&= 0.1536 + \binom{4}{3} 0.2^3 0.8^1 + \binom{4}{0} 0.2^4 0.8^0 = 0.1808\,.
\end{aligned}
$$

The Geometric Probability Distribution

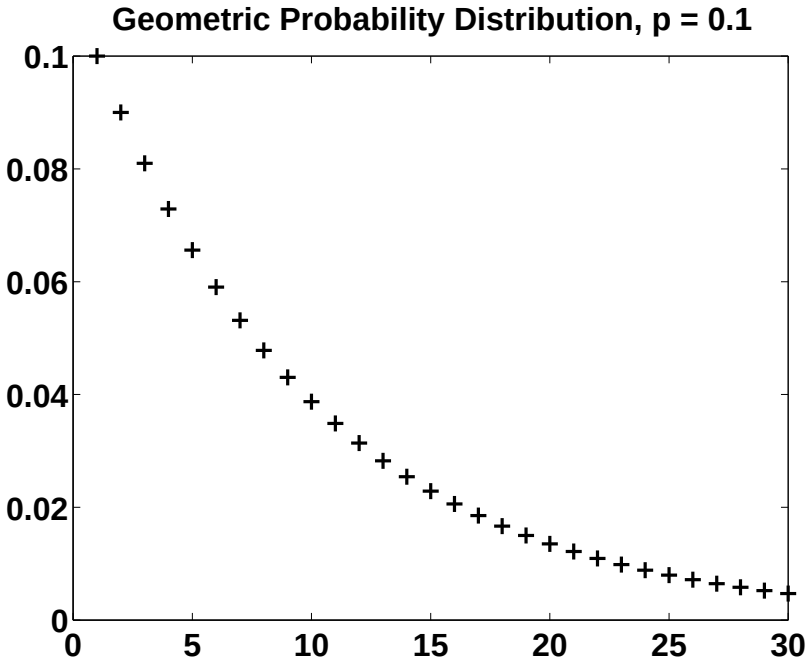
Example. Customers wait in line to be served at a wicket. Per time interval the probability that a customer is served is 10%. What is the probability that a customer has to wait 15 time intervals before being served? □

Such and similar events are modeled by the *geometric probability distribution*. Each time interval we have an 'independent experiment' which can succeed or fail with success probability p (as for the binomial probability distribution). To be successful in the xth try we need $x - 1$ failures (with probability $q = 1 - p$) and one success (with probability p).

The Geometric Probability Distribution

The data for the geometric probability distribution are

- $p(x) = pq^{x-1}, \qquad x = 1, 2, \ldots,$
  where x is the number of trials until the first success; and
- $\mu = \frac{1}{p}$, and $\sigma = \sqrt{\frac{q}{p^2}}$.



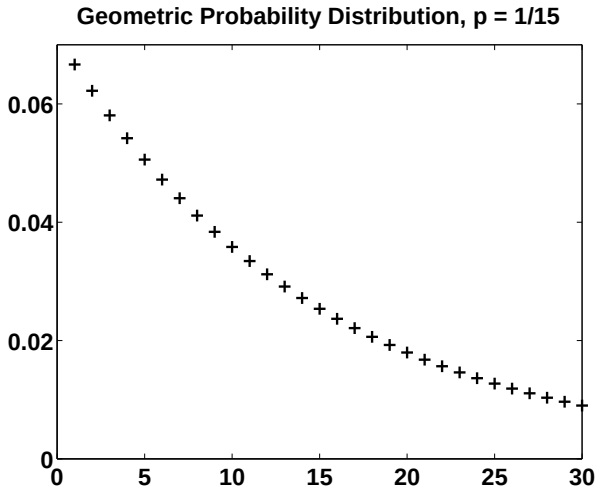
The Geometric Probability Distribution

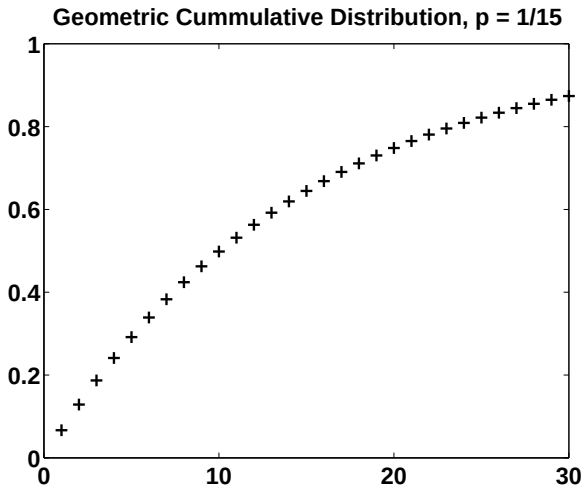
Example. The average life expectancy of a fuse is 15 months. What is the probability that the fuse will last exactly 20 months?

We have that $\mu = 15$ (months), or $p = \frac{1}{15}$, which is the probability that a fuse will break. For $x = 20$ we obtain

$$P(\text{`}x = 20\text{'}) = \frac{1}{15}(1 - \frac{1}{15})^{20-1},$$

which is approximately 0.018. For σ we find $\sqrt{210} = 14.49$. □





The Hypergeometric Distribution

The binomial and the geometric probability distribution are to be applied if, after observing a result, the sample is put back into the population. However, in practice, we often sample without replacement:

Example.

- If we test a bag of 1000 resistors whether they meet the specification we usually won't put back the tested items.
- Suppose people are randomly selected at Princess Street in Edinburgh to fill in a questionnaire about a new product. When people are approached they are usually first asked whether they have already taken part in this marketing research.
- A big manufacturing company maintains their machines on a regular basis. Suppose that on average 15% of the machines need repair. What is the probability that among the five machines inspected this week, one of them needs repair?

The Hypergeometric Distribution

- A box of 1000 fuses is tested one by one until the first defective fuse is found. Supposing that about 5% of the fuses are defective, what is the probability that a defective fuse is among the first 5 fuses tested?

□

Such and similar random variables have a *hypergeometric* probability distribution:

- The populations consists if N objects.
- The possible outcomes of the experiment are success or failure.
- Each sample of size n is equally likely to be drawn.

The Hypergeometric Distribution

The data for the hypergeometric probability distribution are

$$p(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \qquad 0, n-N+r \leq x \leq n, r,$$

where

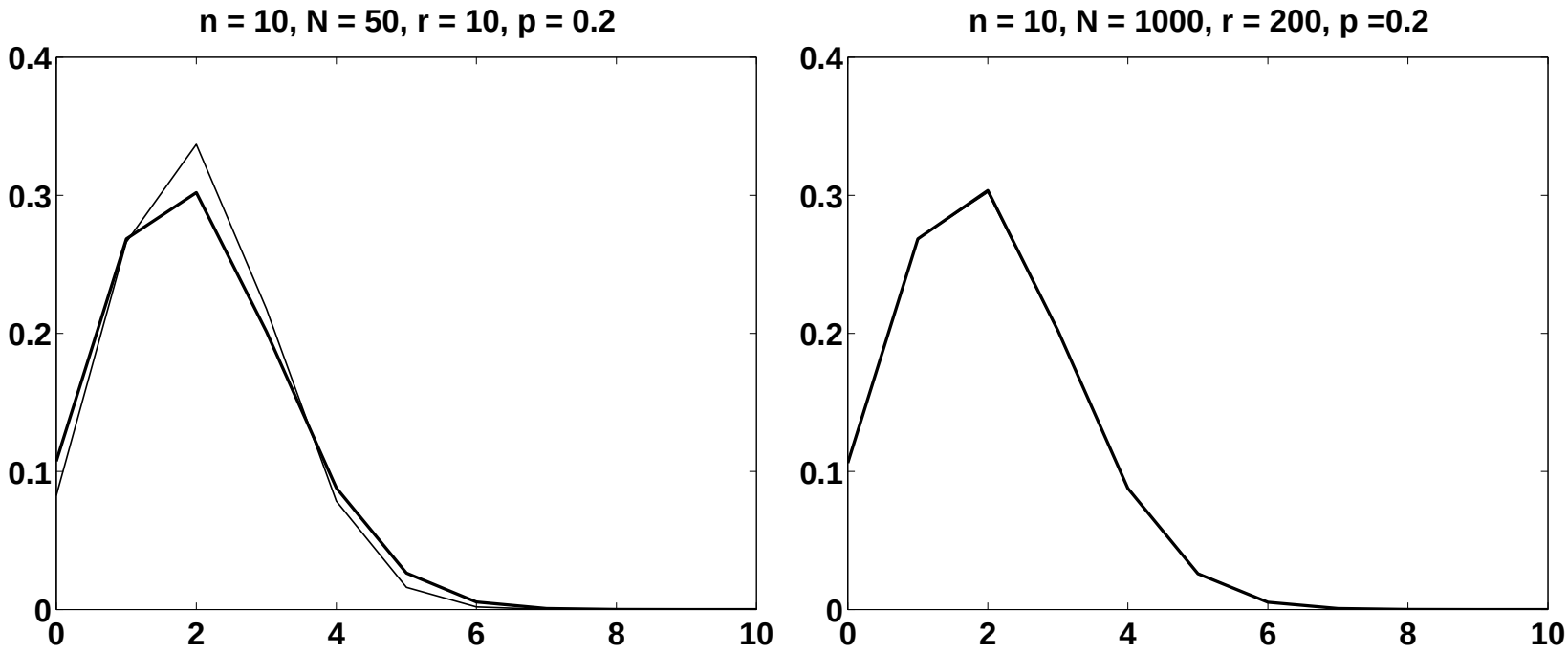
- N is the number of elements in the population;
- r is the number in the population for success;
- n is the number of elements drawn; and
- x is the number of successes in the n randomly drawn elements.

The mean and standard deviation are given by

$$\mu = n\frac{r}{N}, \qquad \text{and} \qquad \sigma = \sqrt{\frac{r(N-r)n(N-n)}{N^2(N-1)}}.$$

The Hypergeometric Distribution

If we write $p = \frac{r}{N}$ then $\mu = np$ and $\sigma = \sqrt{\frac{N-n}{N-1}np(1-p)}$. This shows that the binomial and the hypergeometric distributions have the same expected value, but different standard deviations. The correction factor $\frac{N-n}{N-1}$ is less than 1, but close to 1 if n is small relative to N.



The Hypergeometric Distribution

Example. A retailer sells computers. He buys lots of 10 motherboards from a manufacturer who sells them cheaply, but also offers low quality. Suppose the current lot contains one defective item. If the retailer usually tests 4 items per lot, what is the probability that the lot is accepted?

Here $N = 10$, $r = 1$, and $n = 4$, and we are looking for $P(\text{`}x = 0\text{'})$, which is

$$\begin{aligned} P(\text{`}x = 0\text{'}) &= \frac{\binom{1}{0}\binom{9}{4}}{\binom{10}{4}} = \frac{1 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{1 \cdot 2 \cdot 3 \cdot 4} \frac{1 \cdot 2 \cdot 3 \cdot 4}{10 \cdot 9 \cdot 8 \cdot 7} \\ &= \frac{6}{10}. \end{aligned}$$

We would use the same calculation if we would only know that *on average* 10% of the motherboards are faulty. □

The Hypergeometric Distribution

Example. We test lots of 100 fuses. On average 5% of the fuses are defective. If we test 4 fuses, what is the probability that we accept the current lot?

Again, the random variable is hypergeometric, and since $N = 100$ is large we can assume that there are 5 defective fuses in this lot. We find

$$\begin{aligned} P(\text{`}x = 0\text{'}) &= \frac{\binom{5}{0}\binom{95}{5}}{\binom{100}{5}} = \frac{\frac{5!}{0!5!}\frac{95!}{5!90!}}{\frac{100!}{5!95!}} \\ &= \frac{95 \cdot 94 \cdot 93 \cdot 92 \cdot 91}{100 \cdot 99 \cdot 98 \cdot 97 \cdot 96} \\ &\approx 0.7696\,. \end{aligned}$$

Later we will see how reliable this value is, as we don't know the exact number of faulty fuses in this lot. □

The Poisson Distribution

The *Poisson probability distribution* provides a model for the frequency of events, like the number people arriving at a counter, the number of plane crashes per month, or the number of micro-cracks in steel. (Micro-cracks in steel wheels of the German high-speed train ICE led to a disastrous rail accident in 1998.) The characteristics of a Poisson random variable are as follows:

- The experiment consists of counting events in a particular unit (time, area, volume, etc.).
- The probability that an event occurs in a given unit is the same for every unit.
- The number of events that occur in one unit is independent of the number of events that occur in other units.

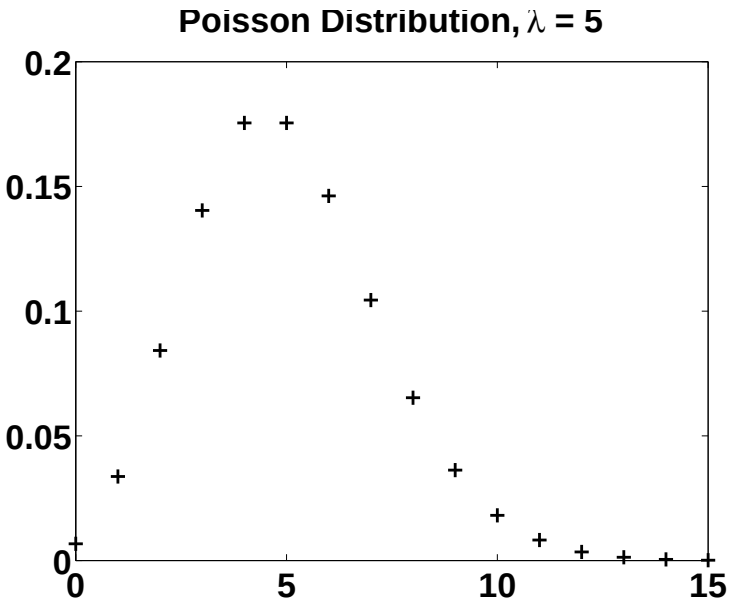
The Poisson Distribution

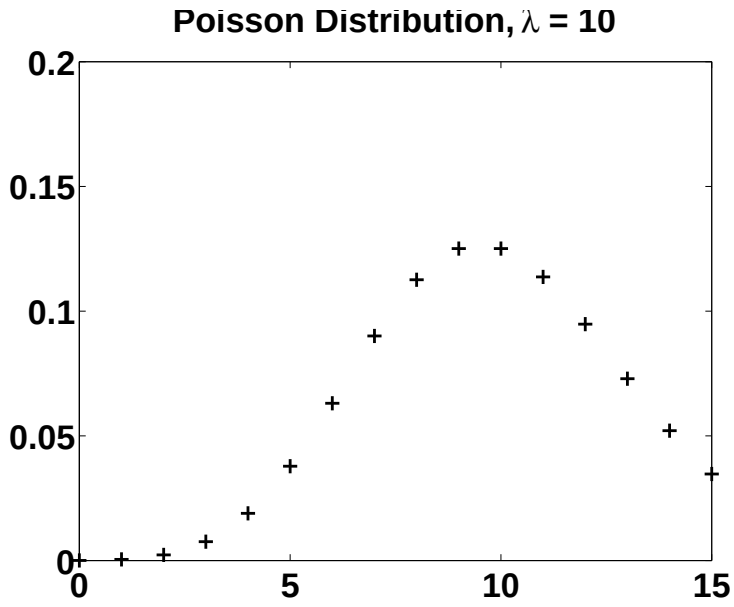
The Poisson probability distribution with mean λ is given by the formula

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}\,, \qquad (x = 0, 1, 2, \ldots)\,,$$

where e is the constant 2.71828. The expected value and standard deviation are

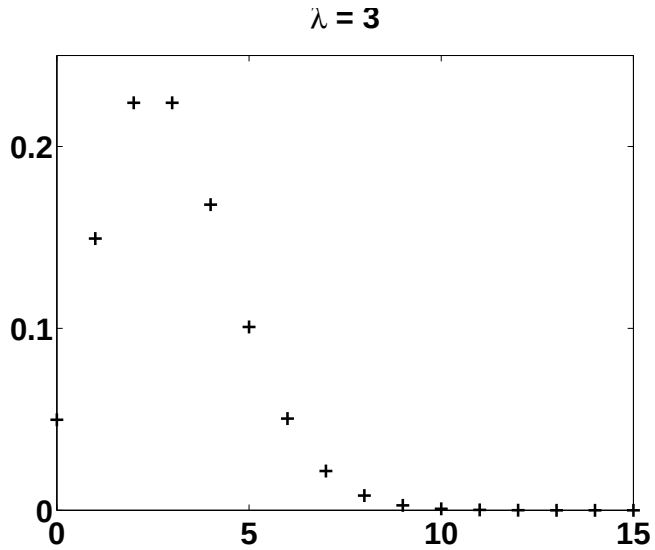
$$\mu = \lambda\,, \qquad \text{and} \qquad \sigma = \sqrt{\lambda}\,.$$





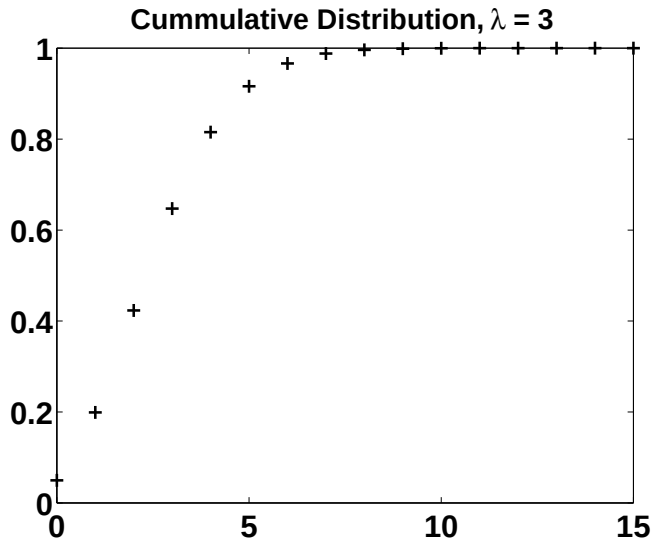
The Poisson Distribution

Example. Suppose customers arrive at a counter at an average rate of 6 per minute, and suppose that the random variable 'customer arrival' has a Poisson distribution. What is the probability that in a half-minute interval at most one new customer arrives?



Here $\lambda = \frac{6}{2} = 3$ customers per half-minute. So

$$\begin{aligned} P(\text{`}x \leq 1\text{'}) &= P(\text{`}x = 0\text{'}) + P(\text{`}x = 1\text{'}) \\ &= \frac{e^{-3}3^0}{0!} + \frac{e^{-3}3^1}{1!} \\ &= \frac{4}{e^3}\,, \end{aligned}$$



which equals approximately 0.199. □

As an example we will verify that the Poisson probability distribution

The Poisson Distribution

$p(x)$ really is a distribution, and that the mean is λ.

We note first that $0 \leq p(x)$ for all values of x. Also, since

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!},$$

$1 = e^{-\lambda}e^{\lambda} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = \sum_{x=0}^{\infty}\frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=0}^{\infty} p(x)$, which shows that $p(x) \leq 1$ and $\sum_{\text{all } x} p(x) = 1$.

For the mean we calculate

$$\begin{aligned} E(x) &= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} = 0 + \sum_{x=1}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \\ &= \sum_{x=1}^{\infty}\frac{\lambda e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \lambda\sum_{x=0}^{\infty}\frac{e^{-\lambda}\lambda^x}{x!} = \lambda\,. \end{aligned}$$

Continuous Random Variables

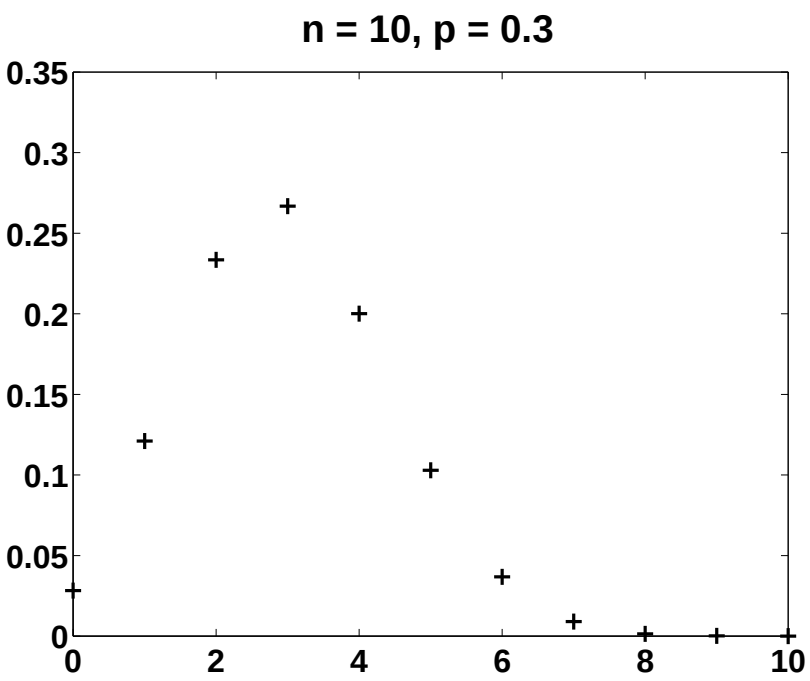
Many random variables arising in practice are not discrete. Examples are the strength of a beam, the height of a person, or the capacity of a conductor. Such random variables are called *continuous*.

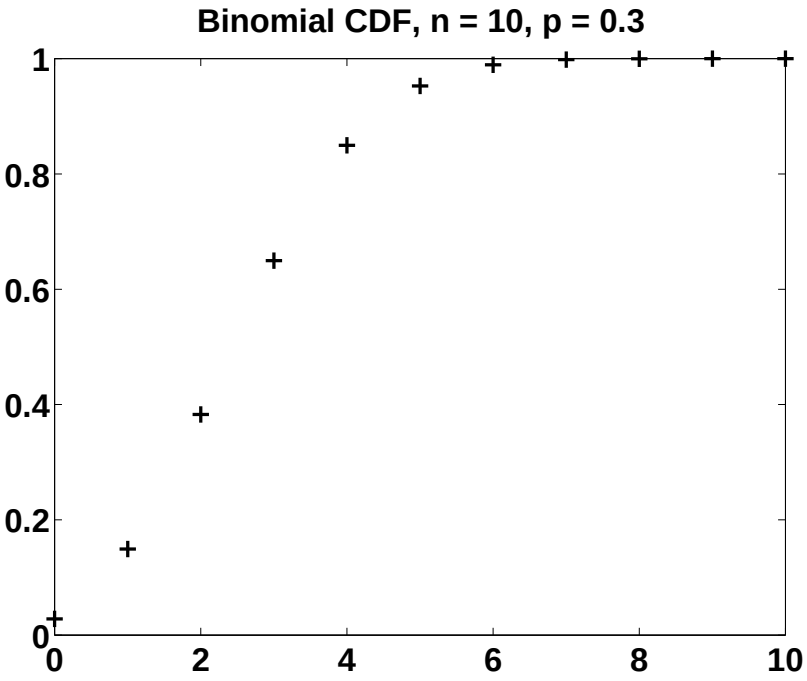
A practical problem arises, as it is *impossible* to assign finite amounts of probabilities to uncountably many values of the real line (or some interval) so that the values add up to 1. Thus, continuous probability distributions are usually based on *cumulative distribution functions*.

The cumulative distribution $F(x)$ of a random variable x is the function

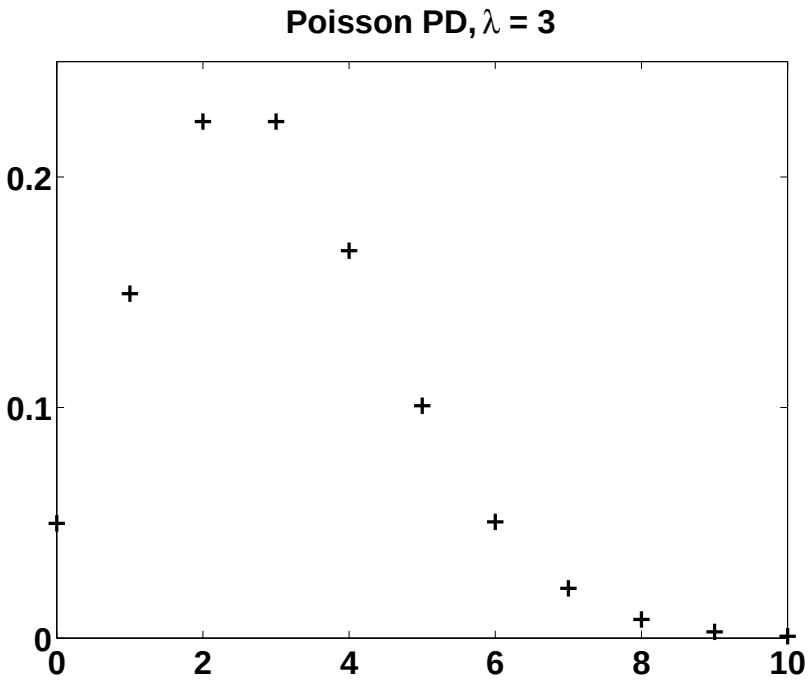
$$F(x_0) = P(\text{`}x \leq x_0\text{'}) \, .$$

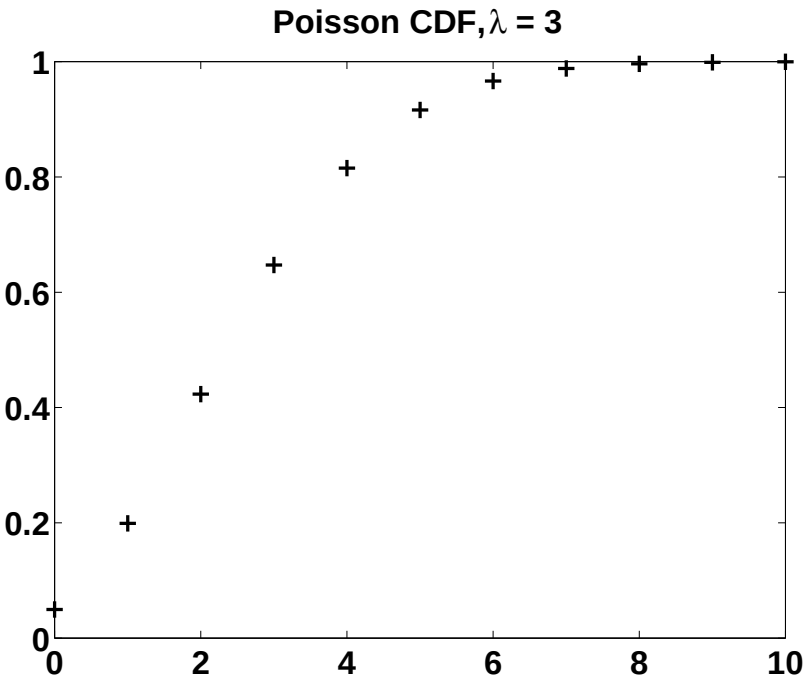
Continuous Random Variables





Continuous Random Variables





Density Functions

If F is the cumulative distribution of a continuous random variable x then the *density* function $\rho(x)$ for x is given by

$$\rho(x) = \frac{dF}{dx}$$

(provided that F is differentiable). It follows that

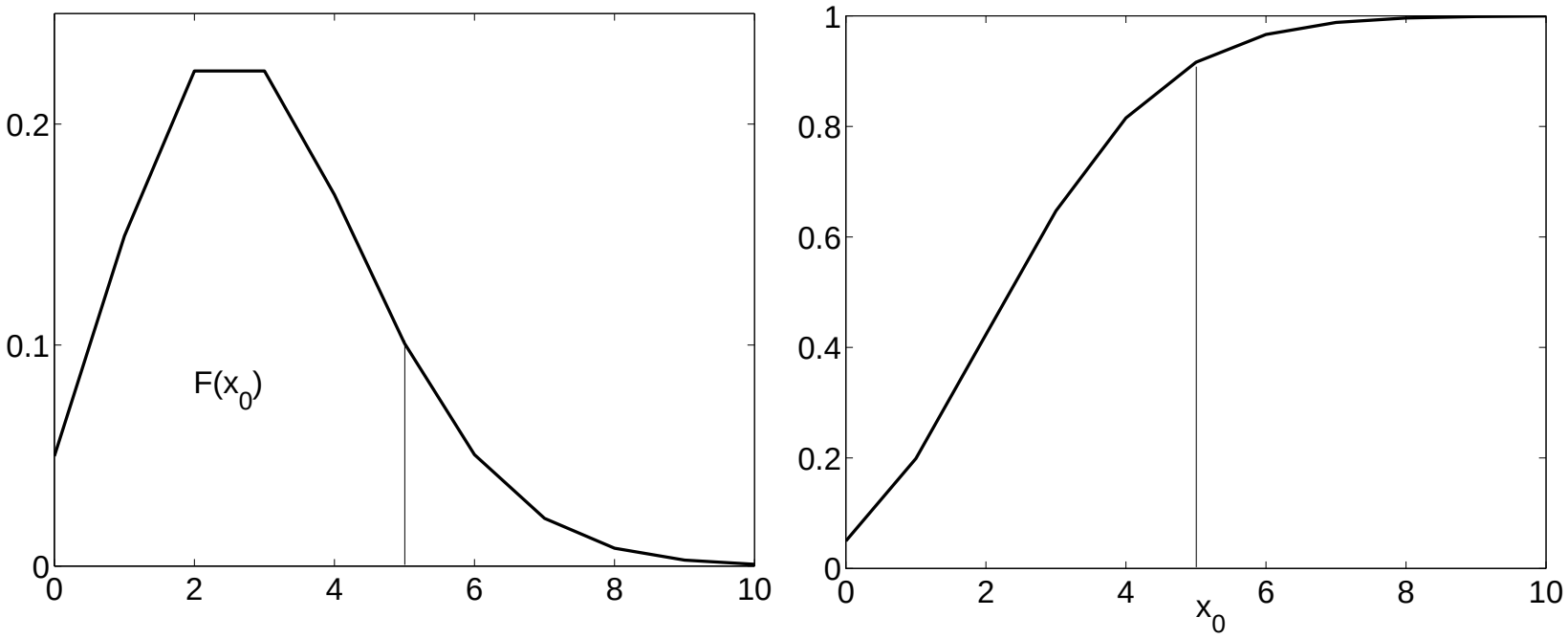
$$F(x) = \int_{-\infty}^{x} \rho(t)\, dt\,.$$

Moreover, the density function always satisfies the following two properties:

- $\rho(x) \geq 0$; and
- $\int_{-\infty}^{\infty} \rho(t)\, dt = 1.$

In particular, $P(a < x < b) = P(a \leq x \leq b) = \int_{a}^{b} \rho(t)\, dt.$

Density Functions



Expected Values

Let us recall from calculus that an integral is a limit process of a summation. Finding

$$F(x_0) = \int_{-\infty}^{x_0} \rho(x)\, dx$$

for a continuous random variable is analogous to finding

$$F(x_0) = \sum_{x \leq x_0} p(x)$$

for a discrete random variable. Thus, we define the *expected value* analogous to the discrete case.

The *expected value* of a continuous random variable x with density function $\rho(x)$ is given by

$$\mu = E(x) = \int_{-\infty}^{\infty} t\rho(t)\, dt\,.$$

Expected Values

If g is any function we define the *expected value of* $g(x)$ as

$$E[g(x)] = \int_{-\infty}^{\infty} g(t)\rho(t)\,dt\,,$$

provided that these integrals exist. The *standard deviation* is $\sigma = \sqrt{E[(x-\mu)^2]}$. Note that

- $E(c) = c$, for every constant c;
- $E(cx) = cE(x)$, for every constant c;
- $E[g_1(x) + g_2(x)] = E[g_1(x)] + E[g_2(x)]$, for any two functions g_1, g_2 on x.
- $\sigma^2 = E[x^2] - \mu^2$.

An Example

Example. We consider the density function

$$\rho = \begin{cases} \frac{1}{2}e^{-\frac{x}{2}} & \text{if } 0 \leq x < \infty, \\ 0 & \text{else.} \end{cases}$$

This density function is everywhere positive, and for $x \leq 0$, $F(x) = \int_{-\infty}^{x} \rho(t)\,dt = 0$, whereas for $x \geq 0$

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} \rho(t)\,dt = \int_{0}^{x} \frac{1}{2}e^{-\frac{t}{2}}\,dt \\ &= \left[-e^{-\frac{t}{2}}\right]_{0}^{x} = e^{0} - e^{-\frac{x}{2}} = 1 - e^{-\frac{x}{2}}\,. \end{aligned}$$

In particular,

$$\begin{aligned} \int_{-\infty}^{\infty} \rho(t)\,dt &= \lim_{x\to\infty} F(x) \\ &= 1 - \lim_{x\to\infty} e^{-\frac{x}{2}} = 1 - 0 = 1\,. \end{aligned}$$

An Example

For the expected value we find

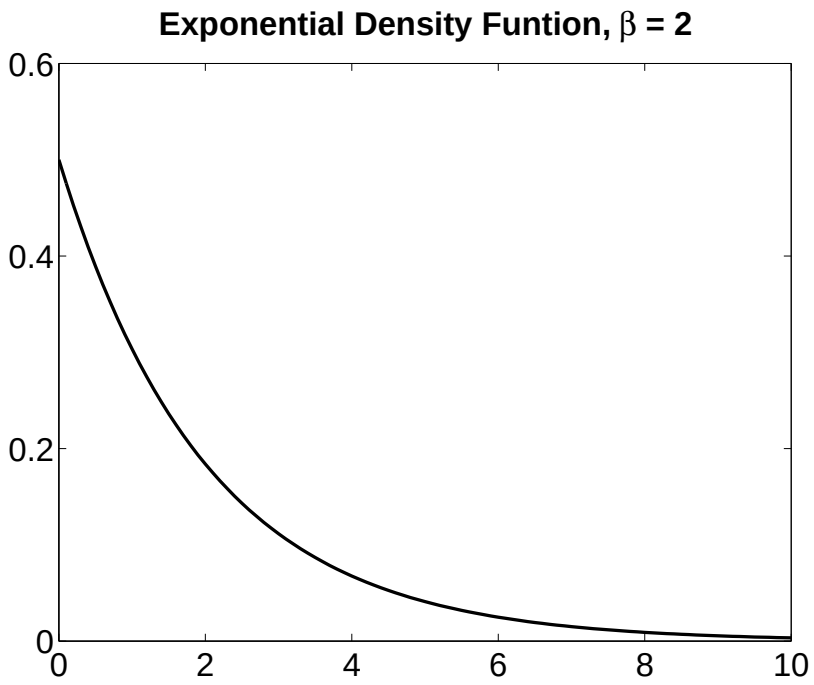
$$\begin{aligned}
\mu &= \int_{-\infty}^{\infty} t\rho(t)\,dt \;=\; \int_{-\infty}^{\infty} t\frac{1}{2}e^{-\frac{t}{2}}\,dt \\
&= \lim_{x\to\infty}\int_0^x t\frac{1}{2}e^{-\frac{t}{2}}\,dt \qquad\qquad \left|\int te^{at} = \frac{e^{at}}{a^2}(at-1)\right. \\
&= \frac{1}{2}\lim_{x\to\infty}\left[4e^{-\frac{t}{2}}(2\frac{t}{2}-1)\right]_0^x \\
&= \frac{1}{2}[0-4e^0(-0-1)] \;=\; 2\,.
\end{aligned}$$

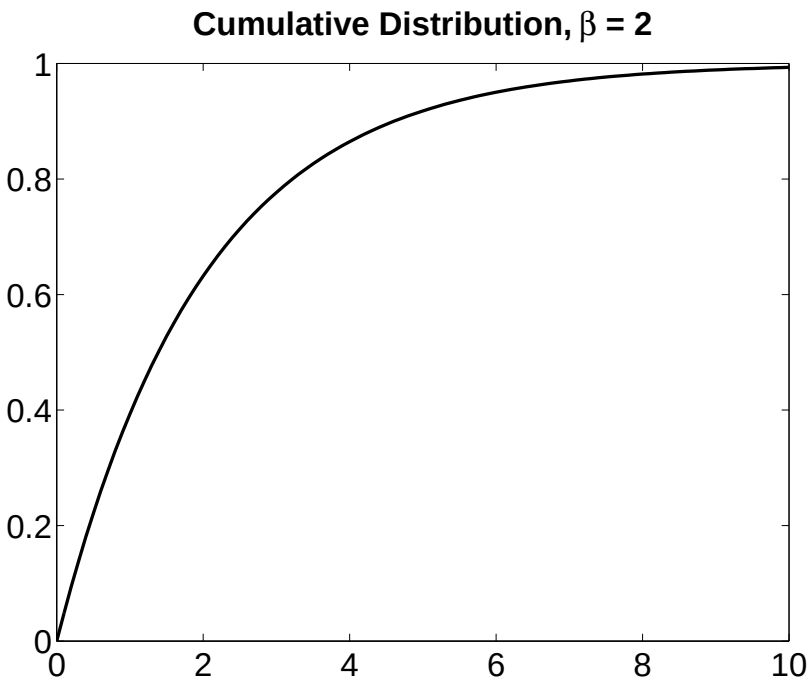
A similar calculation shows that $E(x^2) = 8$, so that $\sigma = \sqrt{E(x^2)-\mu^2} = \sqrt{8-4} = 2$. Finally, to do another calculation,

$$\begin{aligned}
P(\mu-\sigma<x<\mu+\sigma) &= \int_0^4 \frac{1}{2}e^{-\frac{t}{2}}\,dt \;=\; -e^{-\frac{t}{2}}\Big|_0^4 \\
&= -e^{-2}+e^0 \;=\; 1-\frac{1}{e^2} \;=\; 0.8647\,.
\end{aligned}$$

An Example

The empirical rule of Unit 1 suggested 68%. □





The Uniform Probability Distribution

If we select randomly a number in the interval $[a, b]$ then the corresponding random variable x is called a *uniform random variable*. Its density function is

$$\rho = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{else.} \end{cases}$$

For the mean and standard deviation one finds

$$\mu = \frac{a+b}{2} \qquad \text{and} \qquad \sigma = \frac{b-a}{2\sqrt{3}} = \frac{\sqrt{3}}{6}(b-a)\,.$$

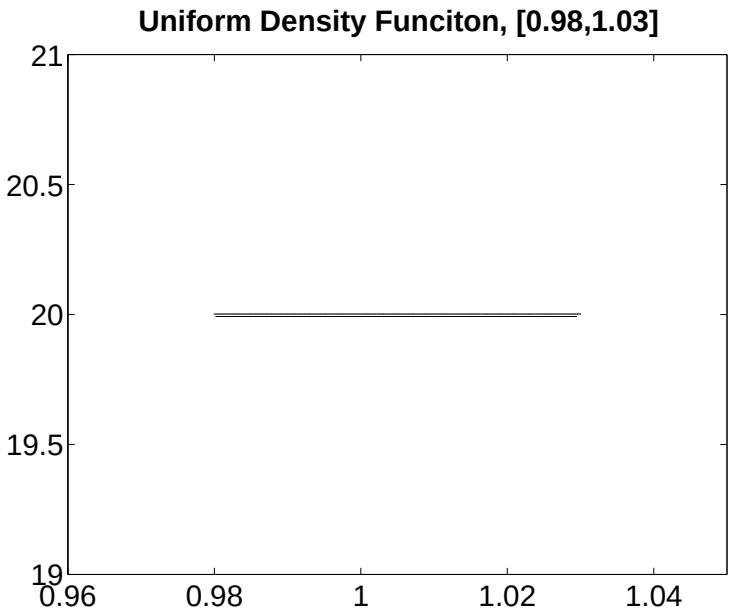
The Uniform Probability Distribution

Example. A manufacturer of wires believes that one of her machines makes wires with diameter uniformly distributed between 0.98 and 1.03 millimeters.

The mean of the thickness is $\frac{1.03+0.98}{2} = 1.005$ millimeters, and the standard deviation is $\sigma = \frac{\sqrt{3}}{6}(1.03 - 0.98) \approx 0.014$ millimeters.

The density function for this uniform random variable is $\rho = \frac{1}{.05} = 20$ for $0.98 \leq x \leq 1.03$, and 0 elsewhere. And, for example,

$$\begin{aligned} P(\text{`}x \leq 1.00\text{'}) &= \int_{0.98}^{1.00} 20\, dt \\ &= 20[1.00 - 0.98] \\ &= 0.4\,. \end{aligned}$$

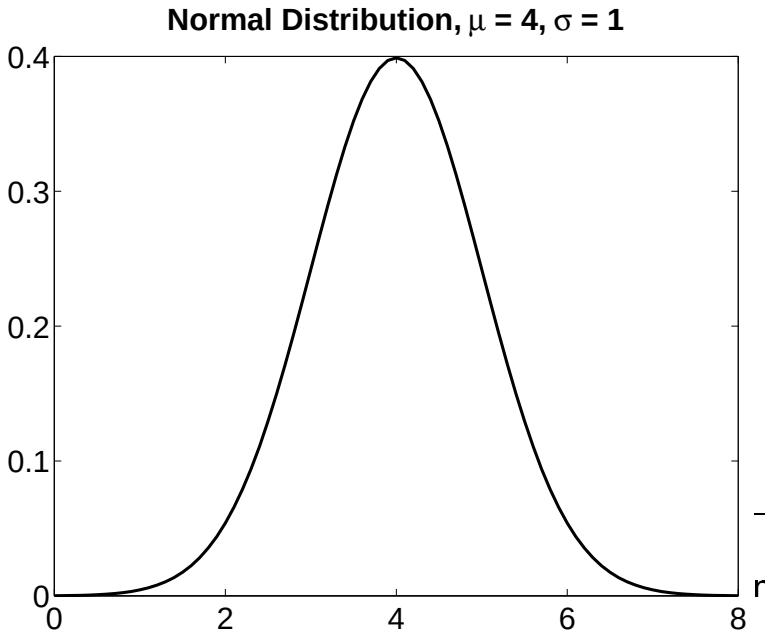


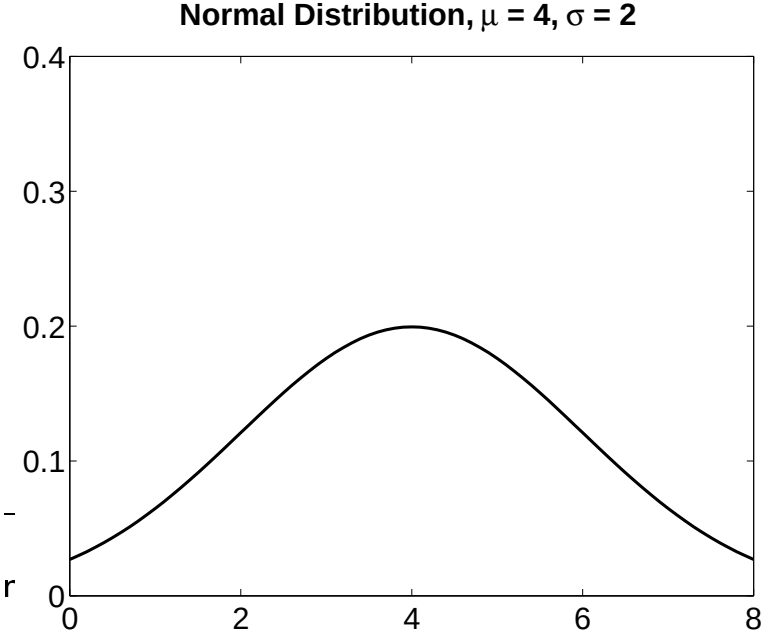
The Normal Probability Distribution

The *normal* probability distribution was suggested by C. F. Gauss as a model of the relative frequency distribution of errors (for example in measurements). The density function of this probability distribution is

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty\,,$$

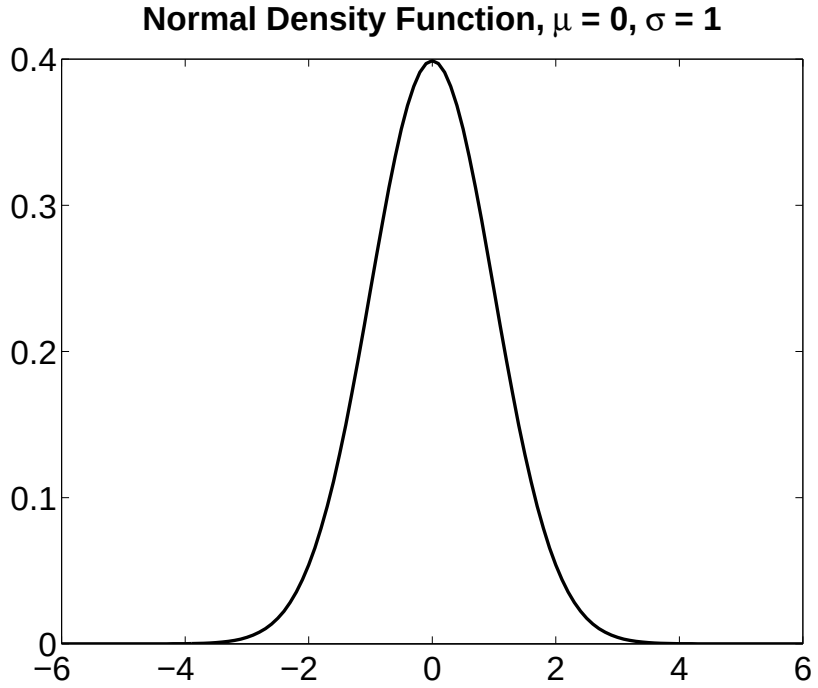
where μ and σ denote the mean and standard deviation, respectively (so these two values are parameters of the normal probability distribution).





The Normal Probability Distribution

The *standard normal random variable* has mean 0 and variance 1:



In practice it is enough to have tables for the standard normal probability distribution: Given a random variable x the variable $z = \frac{x-\mu}{\sigma}$ has mean 0 and standard deviation 1.

The Normal Probability Distribution

Example. Suppose a normally distributed random variable x has mean 10 and standard deviation 3. Find $P(\text{`}x \leq 11\text{'})$ using tables.

We set $z = \frac{x-10}{3}$, which has standard normal distribution. The x-value 11 corresponds to the z-value $\frac{11-10}{3} = \frac{1}{3}$. Then the table shows $P(\text{`}x \leq 11\text{'}) = P(\text{`}z \leq \frac{1}{3}\text{'}) = P(\text{`}z \leq 0\text{'}) + P(\text{`}0 \leq z \leq \frac{1}{3}\text{'}) \approx .5 + 0.1293 = 0.6293.$ □

Why is this justified? In the integral calculating $P(\text{`}x \leq 11\text{'})$ we substitute $z = \frac{x-\mu}{\sigma}$. Then $\frac{dz}{dx} = \frac{1}{\sigma}$, and

$$\begin{aligned} P(\text{`}x \leq 10\text{'}) &= \int_{-\infty}^{11} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\ &= \int_{-\infty}^{\frac{1}{3}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz \\ &= P(\text{`}z \leq \tfrac{1}{3}\text{'})\,. \end{aligned}$$

The Normal Probability Distribution

Example. An amplifier is built using two integrated circuits. Both have a life-length that is normally distributed, the first with mean 36000 hours and standard deviation 8000 hours, the second with mean 38000 hours and standard deviation 10000 hours. Which of the two integrated circuits is more likely to last at least 40.000 hours?

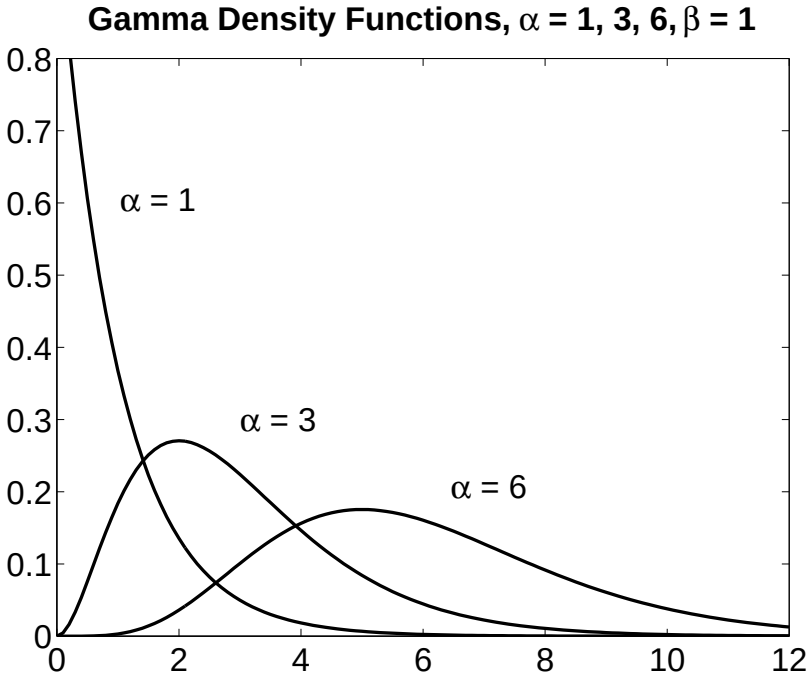
In both cases we ask for $P(\text{`}x \geq 40000\text{'}) = 1 - P(\text{`}x \leq 40000\text{'})$. The corresponding values for the standardized normal random variables z_1 and z_2 are $z_1 = \frac{1}{2}$, and $z_2 = \frac{1}{5}$. Thus

$$P(\text{`}x_1 \geq 40000\text{'}) = 1 - P(\text{`}z_1 < \tfrac{1}{2}\text{'}) \approx 1 - (0.5 + 0.1915) = 0.3085\,,$$

and similarly, $P(\text{`}x_2 \geq 40000\text{'}) = 1 - P(\text{`}z_2 < \frac{1}{5}\text{'}) \approx 1 - (0.5 + 0.0793) = 0.4207$. Thus, the second integrated circuit is more likely to last more than 40000 hours. □

The Gamma Distribution

Many continuous random variables can only take positive values, like height, thickness, life expectations of transistors, etc. Such random variables are often modeled by *gamma type random variables*. The corresponding density functions contain two parameters α, β. The first is known as the *shape* parameter, the second as the *scale* parameter.



The Gamma Distribution

The density function is given by

$$\rho(x) = \begin{cases} \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} & \text{if } 0 \leq x < \infty,\ \alpha, \beta > 0, \\ 0 & \text{else,} \end{cases}$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\,dt$. The mean and standard deviation are

$$\mu = \alpha\beta \qquad \text{and} \qquad \sigma = \sqrt{\alpha\beta^2}\,.$$

The gamma function plays an important role in mathematics. It holds that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, and $\Gamma(1) = 1$, so that for integer values of α, $\Gamma(\alpha) = \alpha!$. In general there is no closed form for the gamma function, and its values are approximated and taken from tables.

The Gamma Distribution

Example. A manufacturer of CPU's knows that the relative frequency complaints from customers (in weeks) about total failures is modeled by a gamma distribution with $\alpha = 2$ and $\beta = 4$. Exactly 12 weeks after the quality control department was restructured the next (first) major complaint arrives. Does this suggest that the restructuring resulted in an improvement of quality control?

We calculate $\mu = \alpha\beta = 8$ and $\sigma = 4\sqrt{2} \approx 5.657$. The value $x = 12$ lies well within one standard deviation from the (old) mean, so we would not consider it an exceptional value. Thus there is insufficient evidence to indicate an improvement in quality control given just this data. □

The Chi-Square Distribution

The χ^2 *(chi-square) probability distribution* plays an important role in statistics. The distribution is a special case of the gamma distribution for $\alpha = \frac{\nu}{2}$ and $\beta = 2$ (ν is called the *number of degrees of freedom*):

$$\rho(\chi^2) = c(\chi^2)^{\frac{\nu}{c}-1} e^{-\frac{\chi^2}{2}} ,$$

where $c(\chi^2) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}$. For mean and standard deviation one finds

$$\mu = \nu \qquad \text{and} \qquad \sigma = \sqrt{2\nu} .$$

The Exponential Density Function

The *exponential density function* is a gamma density function with $\alpha = 1$,

$$\rho(x) = \frac{e^{-\frac{x}{\beta}}}{\beta}, \qquad x \geq 0\,,$$

with mean $\mu = \beta$ and standard deviation $\sigma = \beta$. The corresponding random variable models for example the length of time *between* events (arrivals at a counter, requests to a CPU, etc) when the probability of an arrival in an interval is independent from arrivals in other intervals. This distribution also models the life expectancy of equipment or products, provided that the probability that the equipment will last t more time intervals is the same as for a new product (this holds for well-maintained equipment).

If the arrival of events follows a Poisson distribution with mean $\frac{1}{\beta}$ (arrivals per unit interval), then the time interval between two successive arrivals is modeled by the exponential distribution with mean β.

The Weibull Density Function

As the gamma probability distribution the *Weibull probability distribution* is often used to model length of life of products, equipment, or components. The density function is

$$\rho(x) = \begin{cases} \frac{\alpha}{\beta} x^{\alpha-1} e^{-\frac{x^\alpha}{\beta}} & \text{if } x \geq 0, \\ 0 & \text{else,} \end{cases}$$

with *shape parameter* α and *scale parameter* β. Moreover,

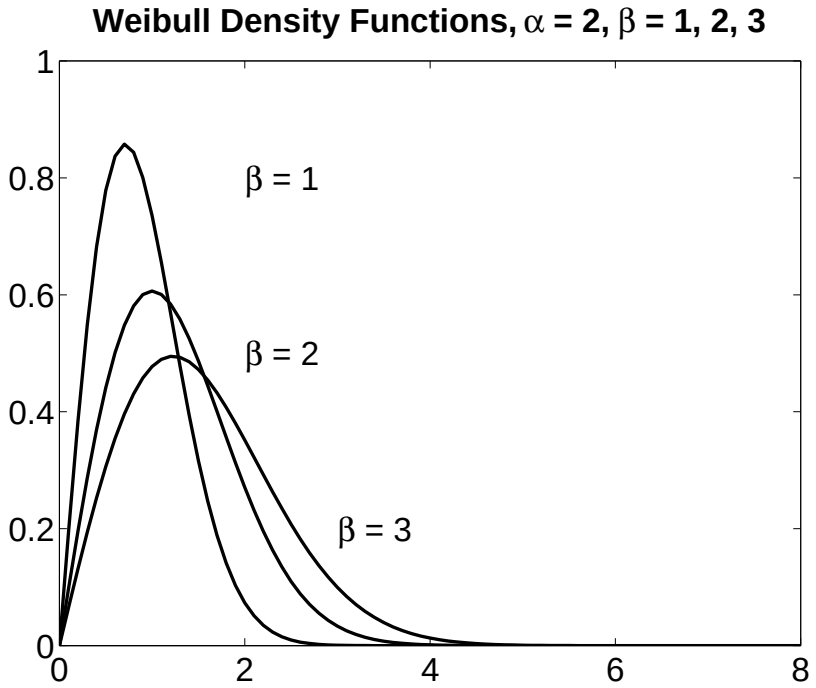
- $\mu = \beta^{\frac{1}{\alpha}} \Gamma(\frac{\alpha+1}{\alpha})$,
- $\sigma = \sqrt{\beta^{\frac{2}{\alpha}} [\Gamma(\frac{\alpha+2}{\alpha}) - \Gamma(\frac{\alpha+1}{\alpha})]}$.

For $\alpha = 1$ we get the exponential density function.

The Weibull Density Function

The Weibull cumulative distribution has a closed form; after substituting $y = x^\alpha$ and $dy = \alpha x^{\alpha-1}\,dx$ we find

$$\begin{aligned} F(x \le x_0) &= \int_0^{x_0} \frac{\alpha}{\beta} x^{\alpha-1} e^{-\frac{x^\alpha}{\beta}}\,dx \\ &= \int_0^{x_0^\alpha} \frac{1}{\beta} e^{-\frac{y}{\beta}}\,dy \\ &= -e^{-\frac{y}{\beta}}\Big|_0^{x_0^\alpha} \\ &= 1 - e^{-\frac{x_0^\alpha}{\beta}} \end{aligned}$$



The Weibull Density Function

Example. The length of life in years of a component in a camera is known to have a Weibull distribution with $\alpha = 2$ and $\beta = 100$. What is the probability that the component will last at least 6 years?

We are looking for $P(\text{`}x \geq 6\text{'})$ which is

$$\begin{aligned} P(\text{`}x \geq 6\text{'}) &= 1 - P(\text{`}x \leq 6\text{'}) \\ &= 1 - (1 - e^{-\frac{6^2}{100}}) \\ &= \frac{1}{e^{\frac{36}{100}}} \\ &\approx 0.698\,. \end{aligned}$$

□

Summary

- Random variables are functions assigning numerical values to each simple event of a sample space. We distinguish discrete and continuous random variables.
- The probability distribution of a discrete random variable is a function that gives for each event the probability that the event occurs.
- The expected value $E(x)$ is the mean, the standard deviation the square root of $E[(x - E(x))^2]$.
- Examples of discrete probability distribution are the binomial, geometric, hypergeometric and the Poisson distribution.
- For continuous random variables we have to give the cumulative probability distribution.

Summary

- The relative frequency distribution for a population with continuous random variable can be modeled using a density function $\rho(x)$ (usually a smooth curve) such that

$$\rho(x) \geq 0 \qquad \text{and} \qquad \int_{\infty}^{-\infty} \rho(x)\, dx = 1\,.$$

- Examples are the uniform distribution, normal distribution, gamma distribution, the exponential distribution and the Weibull distribution.

B34.UC2

Numerical Computation and Statistics in Engineering

Unit 3: Sampling

Introduction

In this unit we will mainly talk about *infinite* populations. However, most of the techniques and results will hold for (large) finite populations.

We are interested in taking random samples of a population to determine properties like the (unknown) mean or (unknown) standard deviation of the population. The process of taking random samples from a population is known as *sampling*.

In general, we want to have the following three properties of an estimator (say for the population mean) when sampling:

- unbiasedness,
- consistency, and
- efficiency.

Introduction

Unbiasedness refers to the fact that the average over *all* possible sample means (of a given size n) is equal to the population mean.

An estimator is *consistent* if, as the sample size increases, the difference between estimate and true population value (here mean) approaches zero. (For example, the formula for the standard deviation with $n-1$ in the denominator is unbiased and consistent, the one with n in the denominator is consistent, but biased.)

Efficiency, the last desirable property of an estimator, refers to the precision of the sample.

Basic Definitions

If $x_1, \ldots, x_n$ are *independent* and *identically distributed* random variables then they form a *random sample* from the population.

Example. A machine manufactures conductors. Each week *one* sample of fifty conductors is taken and the capacity is measured. If x_i is the average of the measures in week i, then the x_i form a random sample from the population. □

If $x_1, \ldots, x_n$ are a random sample then

$$\bar{x} = \frac{1}{n}\sum\nolimits_i x_i \qquad \text{and} \qquad s^2 = \frac{1}{n-1}\sum\nolimits_i (x_i - \bar{x})^2$$

are called the *sample mean* and *sample variance*. As before, the *sample standard deviation* is the square root of the sample variance.

The *standard error* of a statistics is the standard deviation of its sampling distribution.

Distribution of the Mean

If $x_1, \ldots, x_n$ are a random sample from an infinite population with mean μ and standard deviation σ, then

$$E(\bar{x}) = \mu \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\,.$$

This result says that if we sample then the expected value of these samples is the actual mean of the population, i.e., the correct value, and, the standard deviation of the sample decreases if the sample size increases: The more samples we take the more we can be assured that $\bar{x}$ is close to μ. Thus, the estimator $\bar{x}$ for the population mean is unbiased and consistent.

Together with Chebycheff's Theorem (see Unit 1) the previous result can be rephrased as follows:

> For every positive c, the probability that $\bar{x}$ will take a value in the interval $[\mu - c, \mu + c]$ is at least $1 - \frac{\sigma^2}{nc^2}$.

The Central Limit Theorem

Of more importance (both theoretically and practically) is the following version of the *Central Limit Theorem*:

> If n is sufficiently large (in practice $n \geq 30$) the random variable $\bar{x}$ can be approximated with a *normal* probability distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ (regardless of the actual shape of the sampled population (!)).

If the distribution of the population is symmetric then taking samples of size $n \geq 25$ is enough. If the distribution of the population is *normal* then $\bar{x}$ has normal distribution too, regardless of the size of n.

The Central Limit Theorem

Example. A coffee vending machine fills cups with coffee with mean 150 milliliter and standard deviation 15 milliliter. What is the probability that the average amount of coffee in a random sample of size 40 is at least 155 milliliters?

The distribution of $\bar{x}$ (the average amount in the sample of 40 cups of coffee) has sample mean $\mu_{\bar{x}} = 150$, and standard deviation $\sigma_{\bar{x}} = \frac{15}{\sqrt{40}}$, and this distribution is approximately normal. Using the standardized normal distribution $z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$ the corresponding z-value is $\frac{155-150}{\frac{15}{\sqrt{40}}} = \frac{\sqrt{40}}{3} \approx 2.10817$. Thus, tables show that

$$P(\bar{x} \geq 155) = P(z \geq 2.108) \approx 0.0175\,.$$

The table gives $P(0 \leq z \leq 2.10) = 0.4821$, and $P(0 \leq z \leq 2.11) = 0.4826$. With a linear approximation we find that $P(0 \leq z \leq 2.10817) \approx 0.4821 + 0.81(0.4826 - 0.4821) = 0.4825$. Thus, $P(z \geq 2.10817)$ is approximately $1 - P(z \leq 2.10817) \approx 1 - (.5 + 0.4825) = 0.0175$. $\square$

The Chi-Square Distribution

The importance of the chi-square distribution results from the following fact:

> If x has a standard normal distribution then x^2 has chi-square distribution (with $\nu = 1$ degree of freedom) with density function
>
> $$\rho(x) = \begin{cases} \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} x^{-\frac{1}{2}} e^{-\frac{x}{2}} & \text{if } x > 0, \\ 0 & \text{else,} \end{cases}$$
>
> with mean 1 and standard deviation $\sqrt{2}$.

In statistics we use the following more general fact:

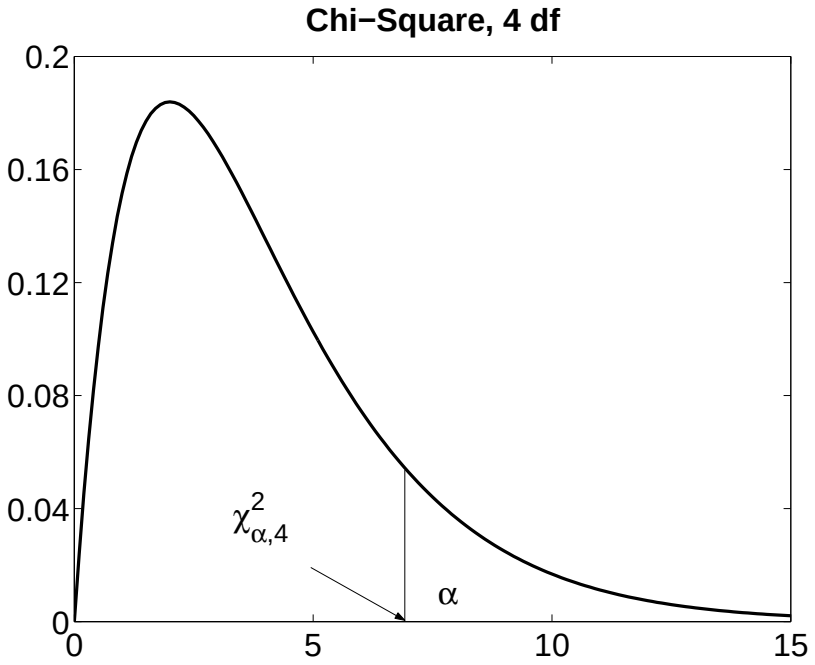
The Chi-Square Distribution

If $\bar{x}$ and s are the mean and standard deviation of a random sample of size n from a normal population with mean μ and standard deviation σ, then

- $\bar{x}$ and s^2 are independent, and
- the random variable $\frac{(n-1)s^2}{\sigma^2}$ has a chi-square distribution with $n-1$ degrees of freedom.

Tables of the chi-square distribution show, for given degree of freedom ν, the value of the random variable (here $\frac{(n-1)s^2}{\sigma^2}$, but often denoted $\chi^2_{\alpha,\nu}$), such that

$$P(\chi^2 \geq \chi^2_{\alpha,\nu}) \geq \alpha\,.$$



The Chi-Square Distribution

Example. Suppose the thickness of some semi-conductor part (with normal distribution) is critical, and suppose that the accepted variation around the mean is at most one standard deviation $\sigma = 0.6 \cdot 10^{-3}$ cm. Random samples of size 20 are taken each week to monitor the manufacturing process.

The machine is to be readjusted if the probability that s^2 will take a value greater than or equal to the observed value is 0.01 or less. What can we conclude if we find that in a sample $s = 0.84 \cdot 10^{-3}$ cm?

The machine is re-adjusted if

$$P(S^2 \geq (0.84 \cdot 10^{-3})^2) \leq 0.01\,.$$

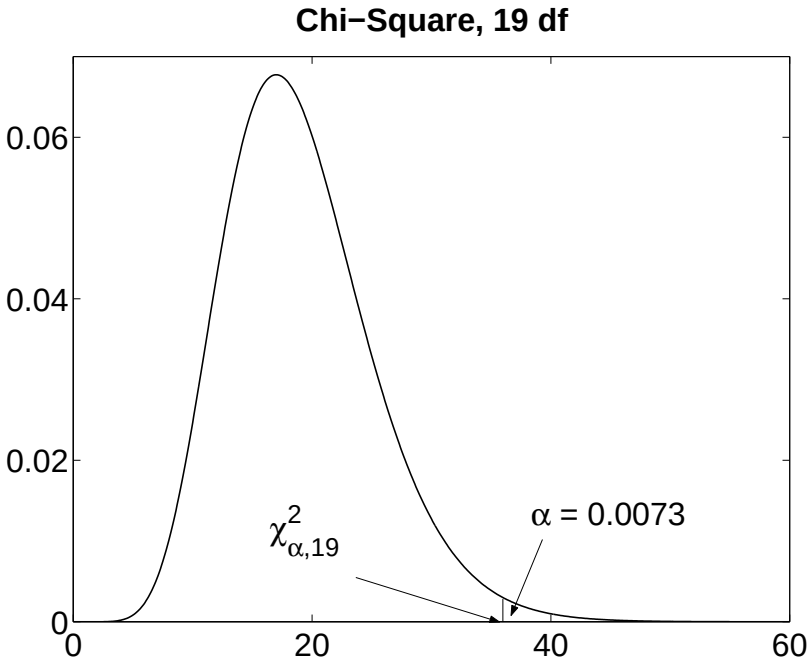
The probability on the left is that of $P(\frac{(n-1)S^2}{\sigma^2} \geq \frac{(n-1)s^2}{\sigma^2})$, which has a chi-square distribution with $20 - 1$ degrees of freedom.

The Chi-Square Distribution

For $n = 20$, $s = 0.84 \cdot 10^{-3}$ and $\sigma = 0.6 \cdot 10^{-3}$ we find $\frac{(n-1)s^2}{\sigma^2} = 37.24$, and

$$P(\frac{(n-1)S^2}{\sigma^2} \geq 37.24) \approx 1 - 0.9926 = 0.0073\,,$$

so that the machine has to be re-adjusted. (The value was calculated using MATLAB.) □



The Chi-Square Distribution

In practice we have to solve such a question using tables:

In our example we want $P(\frac{(n-1)S^2}{\sigma^2} \geq \frac{(n-1)s^2}{\sigma^2}) \leq 0.01$, for $\nu = 19$, $\alpha = 0.01$. From the table we find that his is the case if

$$\frac{(n-1)s^2}{\sigma^2} \geq 36.1908\,.$$

Since we found that $\frac{(n-1)s^2}{\sigma^2} = 37.24 \geq 36.1908$ the machine has to be re-adjusted.

The Student's t-Distribution

We already saw that for random samples from normal populations with mean μ and standard deviation σ that the random variable $\bar{x}$ has a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, that is, $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution.

We cannot apply this knowledge in practice since usually σ, the standard deviation of the population, is unknown.

Hence we replace σ by its estimation s which we get from the random sample. The probability distribution of the random variable

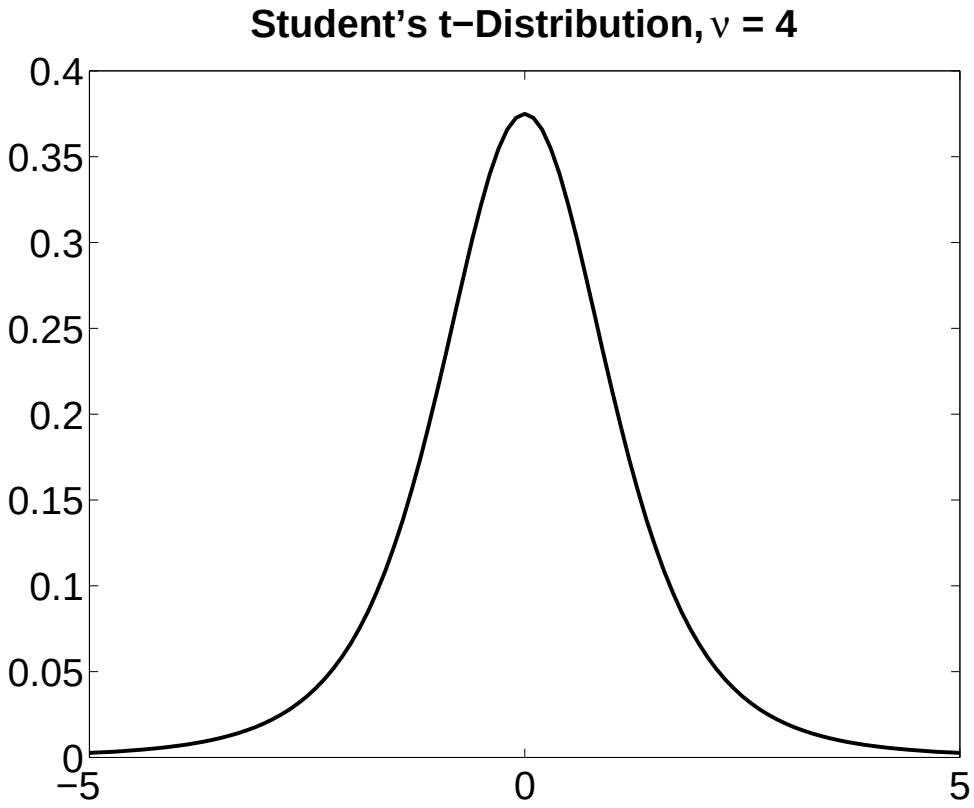
$$\frac{\bar{x}-\mu}{s/\sqrt{n}}$$

is the *t-distribution* with $\nu = n-1$ degrees of freedom, with density function

$$\rho(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \cdot \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \text{for } -\infty < t < \infty.$$

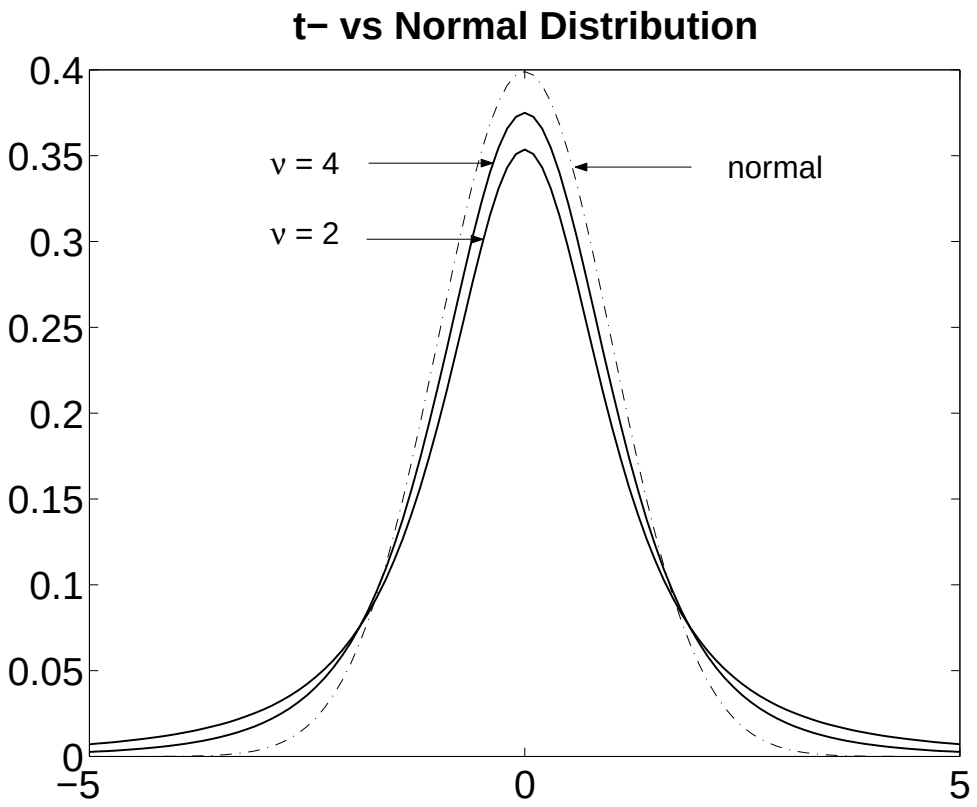
The Student's t-Distribution

W. S. Gossett discovered this distribution through his work at the Guinness brewery. At that time the brewery did not allow its staff to publish, so Gossett used the pseudonym Student, hence the name Student's t-distribution.



The Student's t-Distribution

The t-distribution is similar to the normal distribution, but is *wider* than the latter, i.e., has more tail. This is due to the fact that σ, the true population standard deviation, is only estimated. For larger ν the t-distribution becomes closer to the normal distribution.



The Student's t-Distribution

Example. The output of an old line printer is analyzed for a couple of days and it is found that the printer prints about 45 characters per second, with sample deviation 2 characters per second.

What is the probability that the sample mean of a random sample of 60 seconds will be between 44.5 and 45.3 characters per second?

Here $\bar{x} = 45$, $s = 2$, and $n = 60$. For $z = \frac{x-\bar{x}}{s/\sqrt{n}}$, which has a t-distribution, we find using MATLAB that

$$\begin{aligned} P(44.5 \leq x \leq 45.3) &= P(-1.9365 \leq z \leq 1.1619) \\ &= P(z \leq 1.1619) - P(z \leq -1.9365) \\ &= 0.8750 - 0.0288 \;=\; 0.8462. \end{aligned}$$

The Student's t-Distribution

Approximating the t-distribution by the normal distribution we find that

$$\begin{aligned} P(z \leq 1.1619) &= 0.5 + (0.3770 + .19(0.3790 - 0.3770)) \\ &= 0.8774\,, \end{aligned}$$

$$\begin{aligned} P(z \leq -1.9365) &= P(z \geq 1.9365) \\ &= 1 - P(z \leq 1.9365) \\ &= 1 - (0.5 + (0.4732 + 0.65(0.4732 - 0.4726))) \\ &= 0.0264\,, \end{aligned}$$

and thus $P(-1.9365 \leq z \leq 1.1619) = 0.851$. □

Sampling From Finite Populations

If $\bar{x}$ is the mean of a random variable of size n of a *finite* population of size N with mean μ and standard deviation σ then

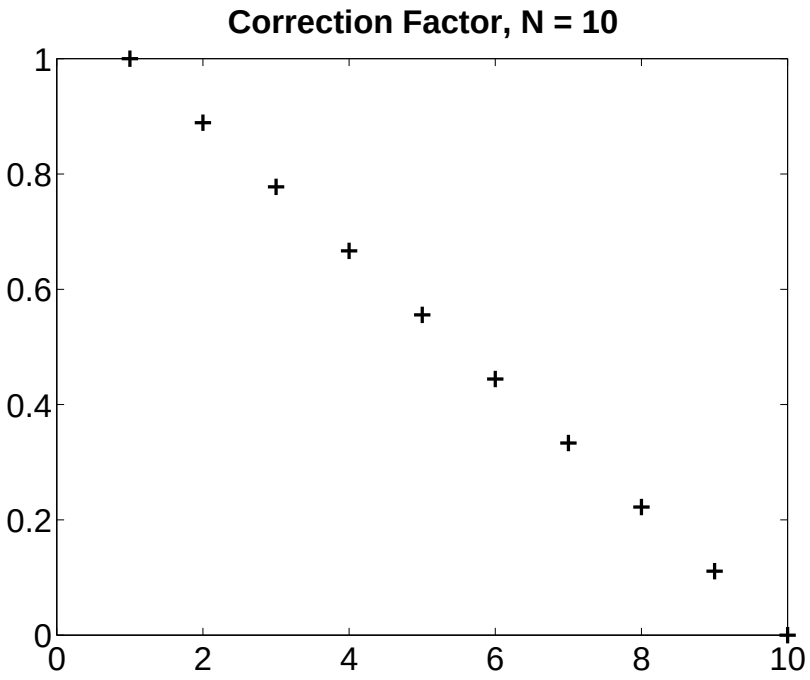
$$E(\bar{x}) = \mu \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\,.$$

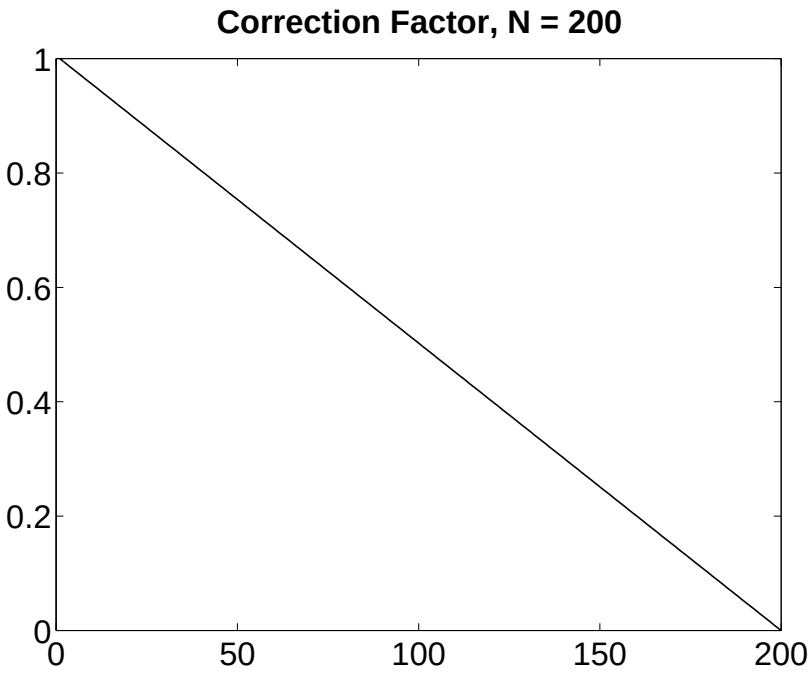
These formulae are similar to those for infinite populations, except for the *finite population correction factor*

$$\sqrt{\frac{N-n}{N-1}}\,.$$

If N is large relative to n then this correction factor is close to 1 and, indeed, the distribution of $\bar{x}$ is then approximated by the normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Sampling From Finite Populations





Normal as Approximation to the Binomial Distribution

Recall that for a binomial random variable the success probability was

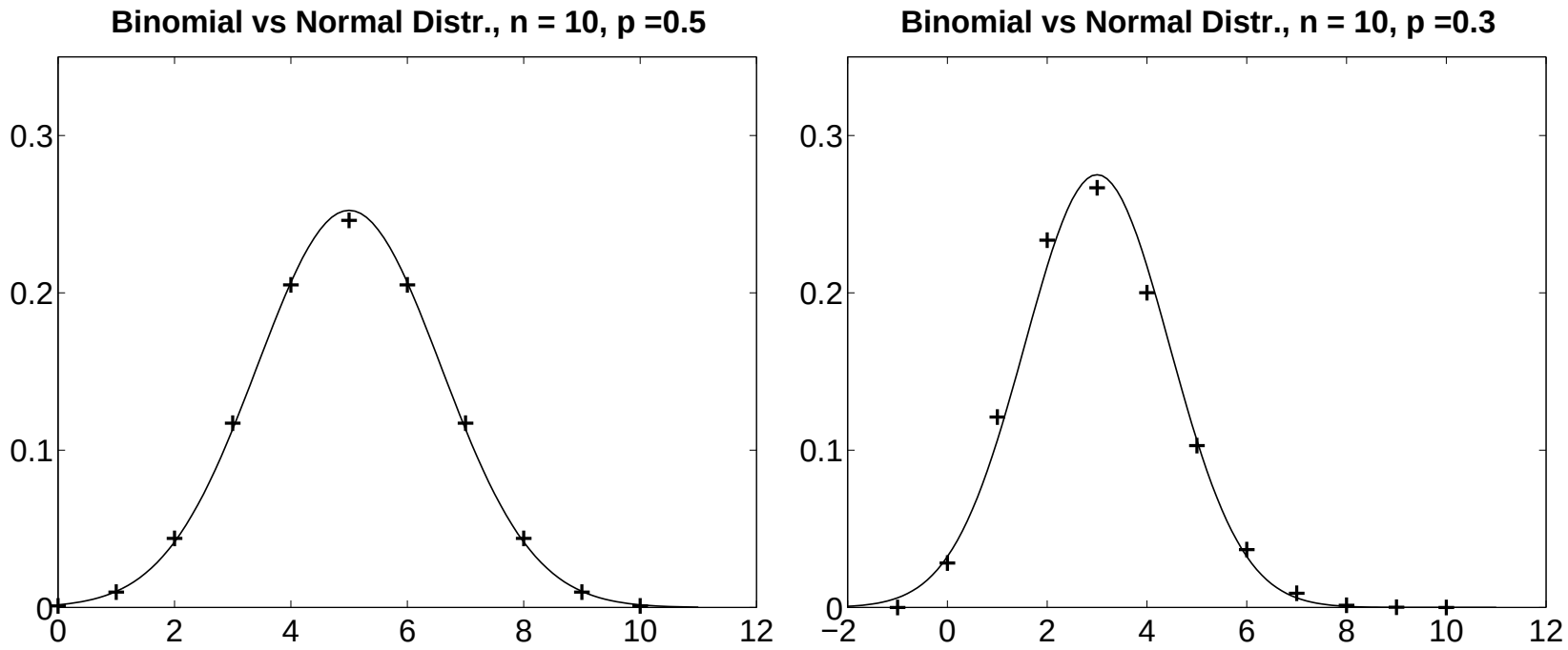
$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} ,$$

where n is the number of trials (or observations), and p is the success probability in one trial. We found that $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

The normal distribution approximates reasonably well the binomial distribution, even for small n ($n = 10$) when p is close to 0.5, and the distribution is symmetric around $\mu = np$. When p differs from 0.5 the binomial distribution is skewed, but the skewness disappears for large n. In general, the approximation is good for n large enough so that

$$0 \leq \mu - 2\sigma, \ \mu + 2\sigma \leq n \, .$$

Normal as Approximation to the Binomial Distribution



Note that, for example in the second case, $\mu = 3$ and $\sigma = 1.45$, so that $\mu - 2\sigma = 0.01 \geq 0$, and $\mu + 2\sigma = 5.9 \leq 10$.

Normal as Approximation to the Binomial Distribution

The following are known as *continuity correction* for the normal approximation: If x is a binomial random variable with parameters n and p, and if $z = \frac{x-\mu}{\sigma}$, then z has approximately standard normal distribution, and

- $P(x \leq a) \approx P(z \leq \frac{(a+0.5)-\mu}{\sigma})$,
- $P(x \geq a) \approx P(z \geq \frac{(a-0.5)-\mu}{\sigma})$,
- $P(a \leq x \leq b) \approx P(\frac{(a-0.5)-\mu}{\sigma} \leq z \leq \frac{(a+0.5)-\mu}{\sigma})$.

Normal as Approximation to the Binomial Distribution

Example. In quality control we randomly check 200 items if they meet the specifications. Suppose that the lot is accepted if the failure rate is below 6%. If, unknown to the quality control engineer, the failure rate is 8%, what is the probability that the lot is accepted?

In this example $n = 200$, $p = 0.08$, and we are looking for the probability that $P(x \leq 0.06 \cdot 200) = P(x \leq 12)$. Using the approximations above this is roughly

$$\begin{aligned} P(z \leq \frac{12.5 - 200 \cdot 0.08}{\sqrt{200 \cdot 0.08 \cdot 0.92}}) &= P(z \leq -0.9123) \\ &= P(z \geq 0.9123) \\ &= 0.5 - P(0 \leq z \leq 0.9123) \\ &\approx 0.5 - (0.3186 + 0.23(0.3213 - 0.3186)) \\ &= 0.180779\,. \end{aligned}$$

MATLAB gives as exact value 0.1821. □

Summary

- Sampling is about taking random sampling from (usually infinite) populations. Sampling is often used to calculate estimators for population parameters.
- The Central Limit Theorem states that for large sample size n the sample mean is approximately normally distributed with mean the true population mean, and standard deviation the true standard deviation of the population divided by the square root of the sample size.
- The random variable $\frac{(n-1)s^2}{\sigma^2}$, with s the sample error and σ the population standard deviation has a chi-square distribution with $(n-1)$ degrees of freedom.
- Usually, the true population standard deviation is not known. In this case the sample mean has a Student's t-distribution with mean the true population mean, and standard deviation the sample deviation divided by the square root of the sample size.

B34.UC2

Numerical Computation and Statistics in Engineering

Unit 4: Estimation

Estimators

Two different type of inferences (here for example about the mean) can be made from a sample:

- one can estimate the true mean of the population; or
- one can try to decide whether the true mean exceeds same value or lies within some interval.

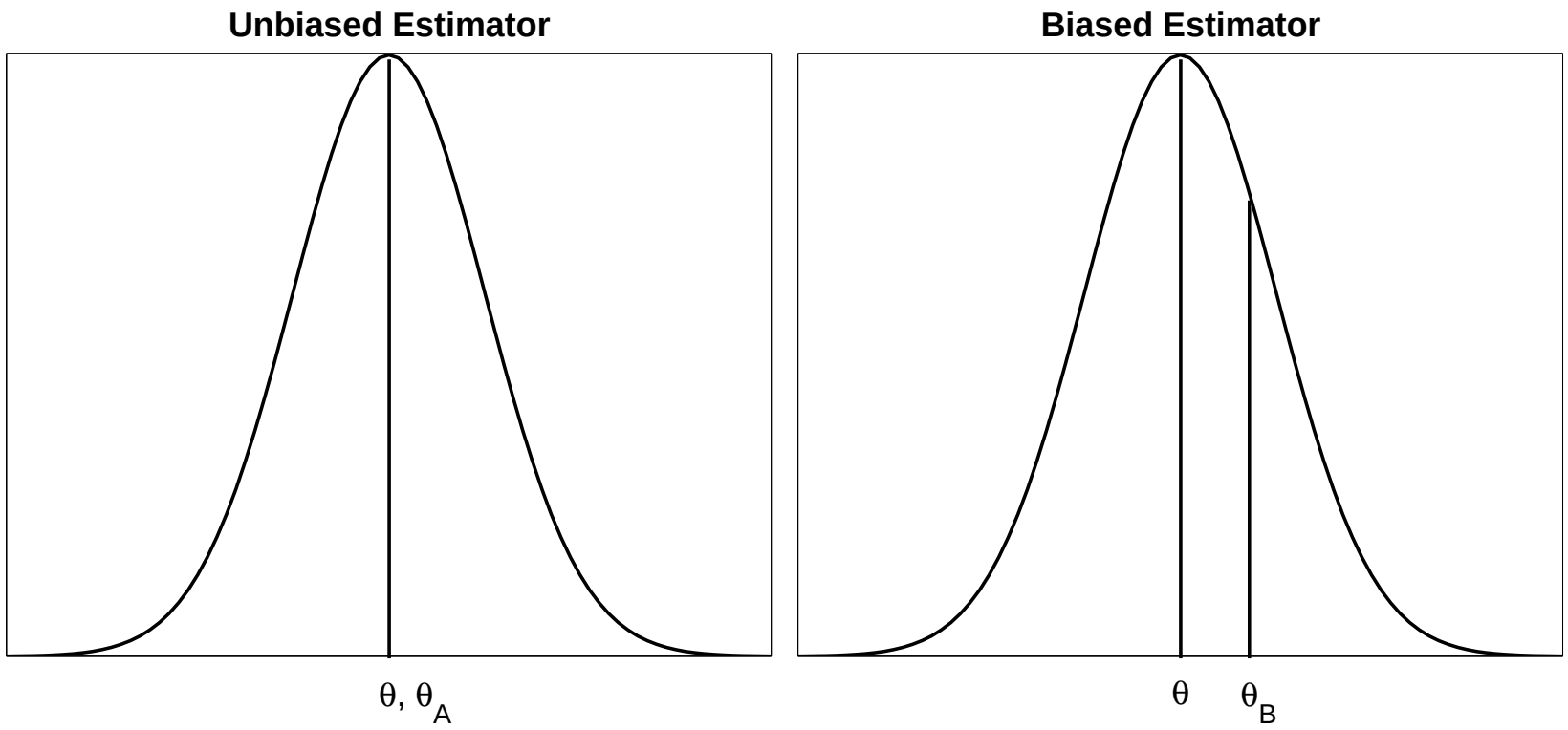
Suppose we want to estimate a population parameter θ (say mean, standard deviation, or $P(\text{`}x \leq d\text{'})$). A *point estimator* for theta is a rule that tells us how to compute from the sample data a single value $\hat{\theta}$ (also called a point estimator) that will serve as an estimator for θ.

An *interval estimator* is a rule computing an interval to estimate θ.

Example. If $x_1, \ldots, x_n$ is a random sample from a population then $\bar{x}$ is a point estimator for the true population mean, whereas $[\bar{x} - s, \bar{x} + s]$ is an interval estimator for the population mean. □

Bias

An estimator $\hat{\theta}$ is called *unbiased* if $E(\hat{\theta}) = \theta$. The *bias* of an estimator is $B = E(\hat{\theta}) - \theta$.



MVUE

In addition to unbiasedness we hope for a small standard deviation (or variance) of the probability distribution of $\hat{\theta}$. An *unbiased* estimator which has minimum variance among all unbiased estimators is called a *minimum variance unbiased estimator* (MVUE).

If such a MVUE does *not* exist one prefers the estimator which minimizes the *mean squared error*

$$E[(\theta - \hat{\theta})^2] \, .$$

Note that

$$\begin{aligned} E[(\theta - \hat{\theta})^2] &= E(\theta^2) - 2\theta E(\hat{\theta}) + E(\hat{\theta}^2) \\ &= \theta^2 - 2\theta E(\hat{\theta}) + \mathrm{var}_{\hat{\theta}} + E(\hat{\theta})^2 \\ &= B^2 + \mathrm{var}_{\hat{\theta}} \, . \end{aligned}$$

MVUE

In particular, if $B = 0$ then

- the mean squared error is equal to the variance of $\hat{\theta}$, and
- the estimator $\hat{\theta}$ that yields the smallest mean squared error is also a MVUE for θ.

Example. If x has binomial distribution with parameters n and p, then $\frac{x}{n}$ is an unbiased estimator for p.

Indeed, since $E(x) = np$ it follows that $E(\frac{x}{n}) = \frac{1}{n}E(x) = \frac{1}{n}np = p$. $\square$

MVUE

Example. If s^2 is the variance of a random sample from an *infinite* population then $E(s^2) = \sigma^2$, the true population variance, hence s^2 is an unbiased estimator for σ^2 (regardless of the nature of the sampled population). Here we use that $s^2 = \frac{1}{n-1}[\sum_i x_i^2 - \frac{1}{n}(\sum_i x_i)^2]$, and the fact that for any random variable, $E(y^2) = \sigma_y^2 + E(y)^2$.

$$
\begin{aligned}
E(s^2) &= \frac{1}{n-1}\left[\sum\nolimits_i E(x_i^2) - \frac{1}{n}E[\sum\nolimits_i x_i)^2]\right] \\
&= \frac{1}{n-1}\left[\sum\nolimits_i (\sigma^2 + \mu^2) - \frac{1}{n}(\sigma^2_{\sum_i x_i} + E(\sum\nolimits_i x_i)^2)\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 + n\mu^2 - \frac{1}{n}\cdot n\sigma^2 - \frac{1}{n}(n\mu)^2\right] \\
&= \frac{1}{n-1}[n\sigma^2 - \sigma^2] \;=\; \sigma^2\,.
\end{aligned}
$$

□

An Example

Consider the following three density functions:

$$\begin{aligned}
\rho(x) &= \frac{1}{2\pi\sigma^2} e^{-(x-\mu)^2/(2\sigma^2)} \qquad \text{for } -\infty < x < \infty, \\
\rho(x) &= \frac{1}{\pi(1+(x-\mu)^2)} \qquad \text{for } -\infty < x < \infty, \\
\rho(x) &= \frac{1}{2c} \qquad \text{for } -c \le x - \mu \le c, \text{ and } 0 \text{ else.}
\end{aligned}$$

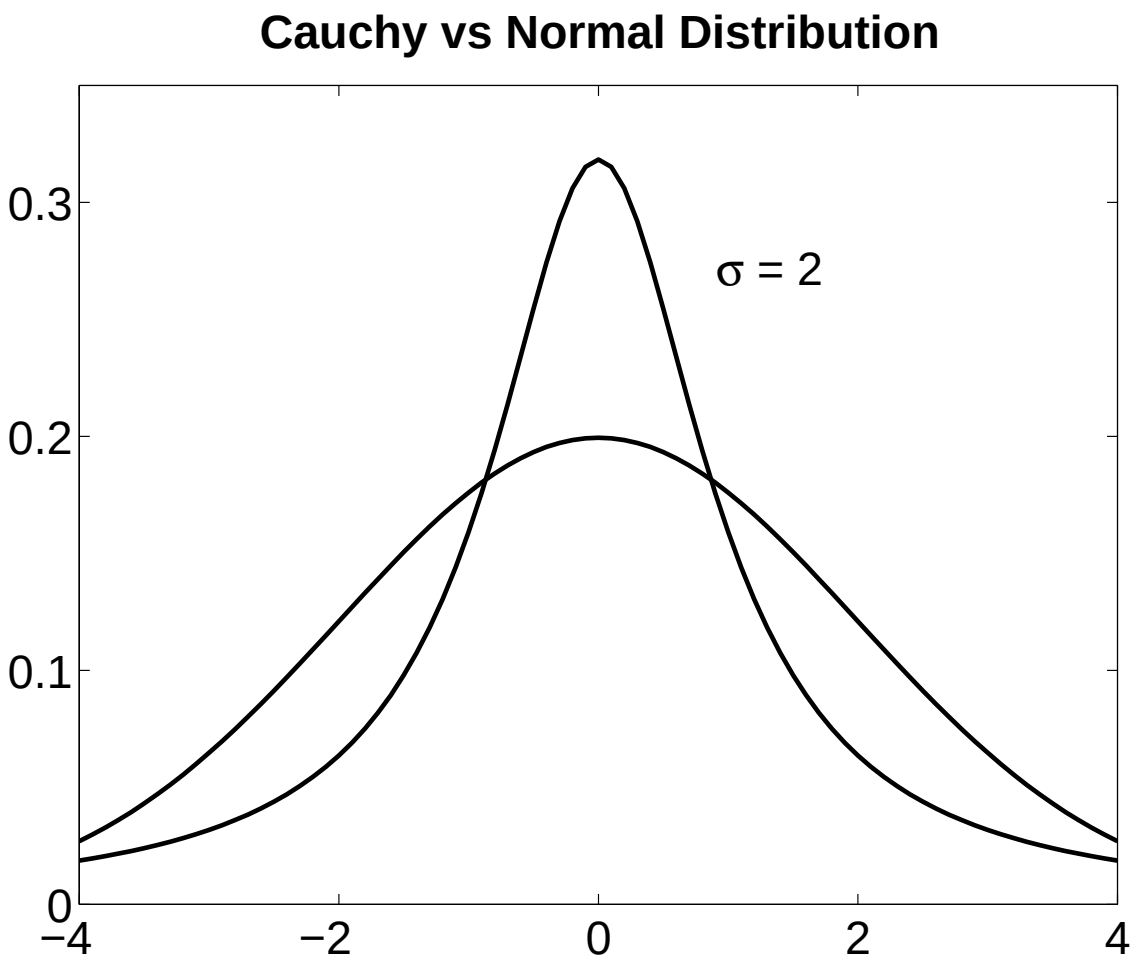
The first is the normal distribution, the second the Cauchy distribution, and the third the uniform distribution. All three have mean μ.

In theory we have at least three estimators for μ from a given sample, namely $\bar{x}$ (mean), $\tilde{x}$ (median), $\bar{x}_e$ (average between the two extreme observations).

An Example

- If the sample comes from a normal distribution, $\bar{x}$ is the best of the estimators as it is the MVUE.
- If the sample comes from a Cauchy distribution then $\bar{x}$ and $\bar{x}_e$ are bad estimators, whereas $\tilde{x}$ is quite good (the MVUE is not known). $\bar{x}$ is bad because it is sensitive to outliers, and the heavy tails of the Cauchy distribution will make such outliers very probable.
- If the distribution is uniform then $\bar{x}_e$ is the best estimator. x_e is sensitive to outliers, but the lack of tails makes such observations impossible.

An Example



Maximum Likelihood Estimators

Let $x_1, \ldots, x_n$ be a random sample. The *likelihood* of the sample is defined as

- $L = P(x_1, \ldots, x_n) = \prod_i P(x_i)$

  if the x_i are discrete random variables;

- $L = \rho(x_1, \ldots, x_n) = \prod_i \rho(x_i)$

  if the x_i are continuous random variables. (Note that $\rho(x_1, \ldots, x_n)$ is the density function of

$$F(x_1, \ldots, x_n) = P(t_1 \leq x_1, \ldots, t_n \leq x_n)\,.)$$

The *maximum likelihood estimator* for θ (or a list of parameters $\theta_1, \ldots, \theta_k$) is the estimator $\hat{\theta}$ (or $\hat{\theta}_1, \ldots, \hat{\theta}_n$) that *maximizes* L.

In practice one often maximizes the logarithm of $\rho(x_1, \ldots, x_n)$, which is easier to calculate and gives the same estimator since the logarithm function is strictly increasing.

Maximum Likelihood Estimators

Example. Let $x_1, \ldots, x_n$ be a random sample of n observations of a random variable x with exponential density function

$$\rho(x) = \begin{cases} \frac{e^{-\frac{x}{\beta}}}{\beta} & \text{if } 0 \leq x < \infty, \\ 0 & \text{else.} \end{cases}$$

What is the maximum likelihood estimator $\hat{\beta}$ for β?

The joint density function is $L(\beta) = \frac{1}{\beta^n} e^{\sum_i -x_i/\beta}$, and

$$\ln L = -n \ln \beta + \sum\nolimits_i -x_i/\beta$$

Setting $\frac{d \ln L}{d \beta}$ equal to 0 gives

$$\frac{\sum_i x_i}{\beta^2} - \frac{n}{\beta} = 0$$

or $\beta = \frac{1}{n} \sum_i x_i$. Thus $\hat{\beta} = \bar{x}$ is the maximum likelihood estimator for β.

□

Maximum Likelihood Estimators

Example. What is the maximum likelihood estimator of the success probability θ of a random sample from a population with binomial probability distribution?

Here $L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, and we maximize

$$\ln L(\theta) = \ln \binom{n}{x} + x \ln \theta + (n-x)\ln(1-\theta)\,.$$

Then

$$\frac{d \ln L}{d\theta} = 0 + \frac{x}{\theta} - \frac{n-x}{1-\theta}\,,$$

and thus $x - \theta x = \theta n - \theta x$, or $\theta = \frac{x}{n}$. □

Example. On 20 cold days a student gets his car started on the third, first, fifth, first, second, third, first, seventh, second, fourth, eighth, fourth, third, first, fifth, sixth, second, first, second, and sixth try. If the distribution of this random variable is modeled by a geometric probability distribution, what is the maximum likelihood estimator for θ?

Maximum Likelihood Estimators

The probability for success in the xth try is

$$\theta(1-\theta)^{x-1}$$

for $x = 1, 2, 3, \ldots$. Then $L(\theta) = \prod_i \theta(1-\theta)^{x_i-1} = \theta^n(1-\theta)^{(\sum_i x_i)-n}$, and $\ln L(\theta) = n \ln\theta + (\sum_i x_i - n)\ln(1-\theta)$, thus

$$\frac{d\,L}{d\,\theta} = \frac{n}{\theta} - \frac{\sum_i x_i - n}{1-\theta}\,.$$

The necessary condition for a maximum is thus

$$n - n\theta = \theta \sum_i x_i - n\theta\,,$$

or $\theta = \frac{n}{\sum_i x_i} = \bar{x}^{-1}$.

For our data, $n = 10$ and $\bar{x} = 3.35$, so that an estimator is given by 0.299. □

The Confidence Coefficient

We continue with interval estimators. The two numbers computed by an interval estimator are the endpoints of the *confidence interval*. The *confidence coefficient* for a confidence interval is the probability that the interval will contain the true (to be estimated) parameter.

As an example we consider the case when $\hat{\theta}$ is approximately normally distributed with mean $E(\hat{\theta}) = \theta$ and error (standard deviation) $\sigma_{\hat{\theta}}$. Then

$$z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

is approximately a standard random variable. We are looking for values z' such that

$$P(-z' \leq z \leq z') = 1 - \alpha\,,$$

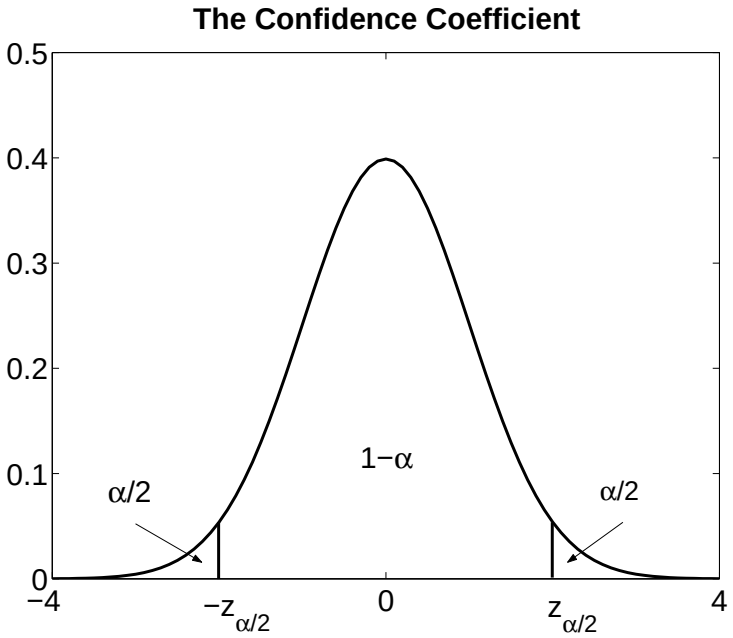
for $1 - \alpha$ the confidence coefficient of the interval $[-z', z']$. From the graph

The Confidence Coefficient

we see that $z' = z_{\alpha/2}$, which is the unique z' such that $P(z \leq z') = \alpha/2$. Substituting back the definition of z we find that for given confidence coefficient $1-\alpha$ the confidence interval for θ is

$$[\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}],$$

where $z_{\alpha/2}$ is the unique z' such that $P(z \leq z') = \alpha/2$ for the normally distributed random variable z with mean 0 and standard deviation 1.



Estimating the Mean

If the sampling size n is large ($n \geq 30$) then $\bar{x}$, the sampling mean, is approximately normally distributed with mean $E(\bar{x}) = \mu$, the true population mean, and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Thus $\bar{x}$ is an unbiased estimator for μ, and $\bar{x}$ is also the MVUE for μ. Since the distribution of $\bar{x}$ is approximately normal we can use the previous analysis to get the endpoints of the $(1-\alpha)100\%$ confidence interval for μ as

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}} = \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is the z-value that locates from $-\infty$ to $z_{\alpha/2}$ an area $\alpha/2$ under the standard normal density function.

Estimating the Mean

If the population is smaller, or if the value of σ has to be approximated by the sample deviation (sample error) s, then the t-distribution with $n-1$ degrees of freedom replaces the normal distribution so that the endpoints of the $(1-\alpha)100\%$ confidence interval for μ become

$$\bar{x} \pm t_{\alpha/2}\sigma_{\bar{x}} = \bar{x} \pm t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\,,$$

where $t_{\alpha/2}$ is the t-value that locates from $-\infty$ to $t_{\alpha/2}$ an area $\alpha/2$ under the density function of the t-distribution with $n-1$ degrees of freedom.

Estimating the Mean

Example. Time between server failures is recorded and for a sample of 20 failures the values $\bar{x} = 1500$ hours and $s = 210$ hours are computed. What is the 95% confidence interval for the mean based on this sample?

To apply the theory we have to assume a normal distribution for the time between server failures. Then $z = \frac{x-\bar{x}}{s/\sqrt{n}}$ has a t-distribution and for $\alpha = 0.05$ we find the endpoints of the confidence interval as

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 1500 \pm 2.093\frac{210}{\sqrt{20}} = 1500 \pm 98.282\,.$$

□

Estimating the Mean

Example. If a random sample of size 20 of a normal population with standard deviation 12.3 has mean 83.2, then we construct the confidence intervals with the following endpoints:

- 90%: $83.2 \pm z_{0.05}\frac{12.3}{\sqrt{20}} = 83.2 \pm 1.645\frac{12.3}{\sqrt{20}} = 83.2 \pm 4.524$
- 95%: $83.2 \pm z_{0.025}\frac{12.3}{\sqrt{20}} = 83.2 \pm 1.960\frac{12.3}{\sqrt{20}} = 83.2 \pm 5.390$
- 99%: $83.2 \pm z_{0.005}\frac{12.3}{\sqrt{20}} = 83.2 \pm 2.576\frac{12.3}{\sqrt{20}} = 83.2 \pm 7.084$

If the standard deviation of the population has to be estimated as well, and if 12.3 is an estimate based on the sample then the endpoints change as follows:

- 90%: $83.2 \pm t_{19,0.05}\frac{12.3}{\sqrt{20}} = 83.2 \pm 1.729\frac{12.3}{\sqrt{20}} = 83.2 \pm 4.755$
- 95%: $83.2 \pm t_{19,0.025}\frac{12.3}{\sqrt{20}} = 83.2 \pm 12.093\frac{12.3}{\sqrt{20}} = 83.2 \pm 5.756$
- 99%: $83.2 \pm t_{19,0.005}\frac{12.3}{\sqrt{20}} = 83.2 \pm 2.861\frac{12.3}{\sqrt{20}} = 83.2 \pm 7.868$

□

Estimating the Mean

Example. Readings from a machine show the following values:

11.3968	4.1666	0.8273	18.4765	7.8963
6.3828	6.4634	2.0181	5.2051	6.4615
16.7264	1.1679	3.2379	0.2825	3.0543
3.4829	0.6679	2.5763	6.3852	1.5892

What is the 95% confidence interval for the mean?

Here $\bar{x} = 5.4232$, and $s = 5.0388$, thus the 95% confidence interval has endpoints

$$\bar{x} \pm t_{19,0.025}\frac{s}{\sqrt{20}} = 5.4232 \pm 2.358\,.$$

(The numbers are random numbers for an exponential distribution with mean 4.) In theory our results do not apply since the sampling size is too small and the distribution is not bell-shaped! □

Estimating the Mean

Example. A company selling easy-to-assemble furniture wants to determine how long it takes to assemble a chest of drawers Bialitt. Using the following data (in minutes) gathered from 15 volunteers we construct a 95% confidence interval.

84.3487	59.6883	95.5066	98.7535	70.0706
116.8183	116.7833	92.2473	99.5458	96.4928
89.2658	107.5158	81.2337	136.6637	90.2721

Here $\bar{x} = 95.6804$ and $s = 19.1003$. Thus the confidence interval has endpoints

$$\bar{x} \pm t_{14,0.025} \frac{s}{\sqrt{15}} = 95.6804 \pm 10.5784\,.$$

(The data are random numbers from a normal distribution with mean 93 and standard deviation 20.) □

Estimating the Difference Between Means

If $\bar{x}_1$ and $\bar{x}_2$ are the values of the means and standard deviation of independent random samples of size n_1 and n_2 from normal populations with known standard deviations σ_1 and σ_2 respectively, then

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n_1 + n_2}}$$

are the endpoints of a $(1-\alpha)100\%$ confidence interval for the difference between the means $\mu_1 - \mu_2$. (By the central limit theorem this confidence interval can also be used for independent random samples from non-normal populations if $n_1, n_2 \geq 30$, or for even smaller samples when the density functions of the populations are known to be bell-shaped.)

Estimating the Difference Between Means

If σ_1 and σ_2 are to be estimated by the samples standard deviations then the endpoints of the $(1-\alpha)100\%$ confidence interval for the difference between the means are

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1+n_2-2} s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where

$$s_P = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}.$$

(s_P is called the pooled estimator for σ^2, and is unbiased.) Here the only assumption is that the two populations are normal.

Estimating the Difference Between Means

Example. Two machines make wires. 10 measurements taken in 1 minute intervals from both machines show the following diameters:

I:	1.0429	1.0627	0.9203	0.9280	1.0286
	0.9800	1.0345	1.0408	1.0356	1.0645
II:	1.0001	1.0157	0.9439	0.9794	0.9753
	0.9319	0.9877	0.9483	1.0225	0.9558

What is a 95% confidence interval for $\mu_1 - \mu_2$?

Estimating the Difference Between Means

Here $n_1 = n_2 = 10$, $\bar{x}_1 = 1.0138$, $s_1 = 0.0526$, $\bar{x}_2 = 0.9761$, and $s_2 = 0.0309$. Thus $\bar{x}_1 - \bar{x}_2 = 0.0377$ and $s_P^2 = \frac{9}{18}(s_1^2 + s_2^2) = 0.00186$.

The endpoints of the interval are thus

$$0.0377 \pm t_{18,0.025} s_P \sqrt{\frac{2}{10}} = 0.0377 \pm 2.101 \cdot 0.0136 = 0.0377 \pm 0.0287\,.$$

(The data are random numbers for normal distributions with mean and standard deviations 1.0, 0.05, and 0.98, 0.03 respectively.)

The analysis shows slightly more: Since we are 95% confident that the difference between the means is within the interval

$$[0.0090, 0.0664]$$

(which does *not* contain 0), we are also 95% confident that the means of the diameters of wires produced by the two machines differ. □

Estimating Proportions

We often try to estimate proportions, probabilities, or percentages such as faulty transistors, faulty lights, etc. If the sample size is large the corresponding random variable has approximately a binomial distribution (even though we often sample without replacement), and can be approximated by a normal distribution.

Thus, if θ denotes the true probability and $\hat{\theta} = \frac{x}{n}$ its estimate derived from a sample of size n then we can assert with $(1-\alpha)100\%$ confidence that the error we make is less than

$$z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\,.$$

Indeed, $z = \frac{x-\mu}{\sigma} = \frac{x-n\theta}{\sqrt{n\theta(1-\theta)}}$ is approximately a standard normal random variable and thus

$$P(-z_{\alpha/2} \leq \frac{x-n\theta}{\sqrt{n\theta(1-\theta)}} \leq z_{\alpha/2}) \leq 1-\alpha\,.$$

Estimating Proportions

Approximating θ by $\hat{\theta}$ under the radical and solving for θ gives

$$
\begin{aligned}
& P\left(-z_{\alpha/2}\sqrt{n\hat{\theta}(1-\hat{\theta})} \leq n\theta - x \leq z_{\alpha/2}\sqrt{n\hat{\theta}(1-\hat{\theta})}\right) \\
= \;& P\left(\frac{x}{n} - z_{\alpha/2}\frac{\sqrt{n\hat{\theta}(1-\hat{\theta})}}{n} \leq \theta \leq \frac{x}{n} + z_{\alpha/2}\frac{\sqrt{n\hat{\theta}(1-\hat{\theta})}}{n}\right) \\
= \;& P\left(\hat{\theta} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right) \\
\leq \;& 1-\alpha\,.
\end{aligned}
$$

Estimating Proportions

Example. A study is made to determine the proportion of people aged between 16 and 25 that use use the internet. If 316 out of 400 young people use the internet, what is a 95% confidence interval for $p = \frac{316}{400} = 0.79$?

Here $n = 400$, $\hat{\theta} = 0.79$, and $z_{0.025} = 1.960$. With 95% confidence the maximum error we make is

$$1.960\sqrt{\frac{0.79 \cdot 0.21}{400}} \approx 0.0399\,,$$

i.e., the interval is $[0.79 - 0.0399, 0.79 + 0.0399]$. □

Estimating Differences in Proportions

We often estimate differences between proportions (differences between males and females in favor of a certain candidate, difference between the percentage of faulty transistors manufactured by two machines, etc.).

If we have two samples x_1 and x_2 of size n_1 and n_2 respectively, then $\hat{\theta}_1 - \hat{\theta}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$ is an estimator for the difference between the two proportions. If both n_1 and n_2 are large then $\hat{\theta}_1 - \hat{\theta}_2$ is approximately normally distributed with mean $\theta_1 - \theta_2$ the difference between the true proportions, and variance

$$\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}\,.$$

Putting everything together and estimating θ_1 and θ_2 by $\hat{\theta}_1$ and $\hat{\theta}_2$ we find the endpoints of the $(1-\alpha)100\%$ confidence interval to be

$$(\hat{\theta}_1 - \hat{\theta}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}\,.$$

Estimating Differences in Proportions

Example. Voters are questioned after they went to the ballots. Out of 212 male voters 76 voted for candidate A, and out of 179 female voters 57 voted for the same candidate. What is a 99% confidence interval for the difference between the percentages of voters voting for candidate A?

Here $n_1 = 212$, $\hat{\theta}_1 = \frac{76}{212} = 0.3585$ and $n_2 = 179$, $\hat{\theta}_2 = \frac{57}{179} = 0.3184$. The endpoints of the interval are thus

$$
\begin{aligned}
& (\hat{\theta}_1 - \hat{\theta}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}} \\
&= (0.3585 - 0.3184) \pm 2.575 \sqrt{\frac{0.3585 \cdot 0.6415}{212} + \frac{0.3184 \cdot 0.6816}{179}} \\
&= 0.0401 \pm \sqrt{0.0011 + 0.0012} \\
&= 0.0401 \pm 0.0479 \,.
\end{aligned}
$$

The interval is thus $[-0.0078, 0.0880]$, which includes 0. This means that the case that there is *no* difference in voting habits among males and females is included. □

Summary

- Estimators are used to estimate population parameters from samples. We distinguish point estimators and interval estimators.
- In addition to unbiasedness we hope for a small standard deviation of the probability distribution of the estimator. An unbiased estimator with smallest variance among all unbiased estimators is called the minimum variance unbiased estimator (MVUE).
- Good estimators are often found the the method of maximum likelihood, which finds that parameter value which makes the observed sample most likely.
- The confidence coefficient for a confidence interval is the probability that the interval will contain the true population parameter.

B34.UC2

Numerical Computation and Statistics in Engineering

Unit 5: Hypothesis Testing

Hypothesis Testing

Statistical tests consist of the following elements:

- a *null hypothesis* H_0 about one or more population parameters;
- an *alternative hypothesis* H_1 (or H_a) that replaces H_0 if the test does not support H_0;
- the test statistics;
- acceptance and rejection regions indicating the values of the statistics that will lead to acceptance or rejection of H_0.

The term null hypothesis stems from the fact that we often test for 'something being equal to 0', for example $\mu - 4 = 0$ (i.e., the population mean equals 4), or $\mu_1 - \mu_2 = 0$ (i.e., the two populations have the same mean).

Errors

There are two types of errors we can make when testing a hypothesis:

		H_0 true	H_0 false
Decision:	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

Type I errors (rejecting H_0 while it is true) are usually denoted by the symbol α, type II errors (accepting H_0 while it is false) are denoted by the symbol β.

Example. A car retailer believes that more than 20% of his customers are willing to spend extra money for upgrading the stereo equipment of their new car. Before ordering new equipment the retailer wants to ask 10 of his customers whether they would buy the more expensive equipment.

Errors

Here we pick

$$H_0 : \ p = 0.2.$$

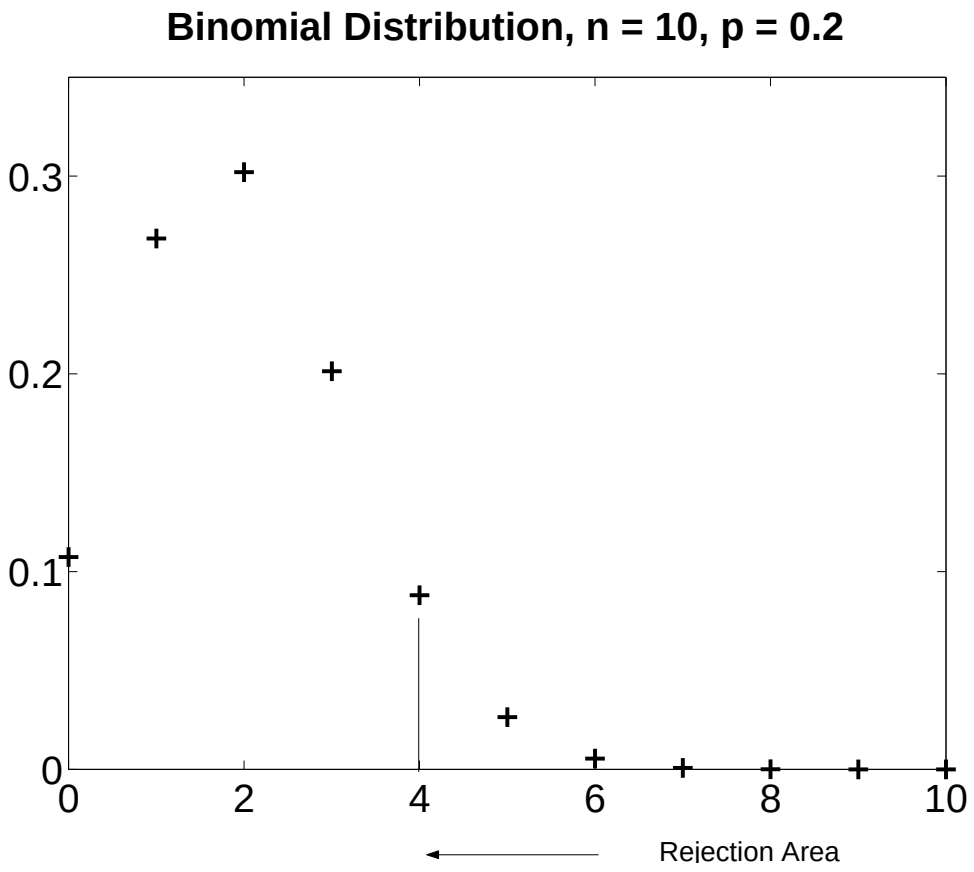
$$H_a : \ p > 0.2.$$

(We do not believe in $p < 0.2$.) The random variable x for this test is the number of people indicating that they would buy better stereo equipment for their cars. If $p = 0.2$ we expect $10 \cdot 0.2 = 2$ people to be in favor of the better product, thus, rejecting H_0 if $x \geq 4$ seems reasonable. For the type II error we find

$$\begin{aligned}
\alpha &= P(\text{reject } H_0 \text{ while it is true}) \\
&= P(p = 0.2 \text{ and } \ x \geq 4) \\
&= 1 - P(p = 0.2 \text{ and } \ x \leq 3) \\
&= 1 - \sum_{x=0}^{3} \binom{10}{x} p^x (1-p)^{10-x} \\
&\approx 0.121 \, .
\end{aligned}$$

Errors

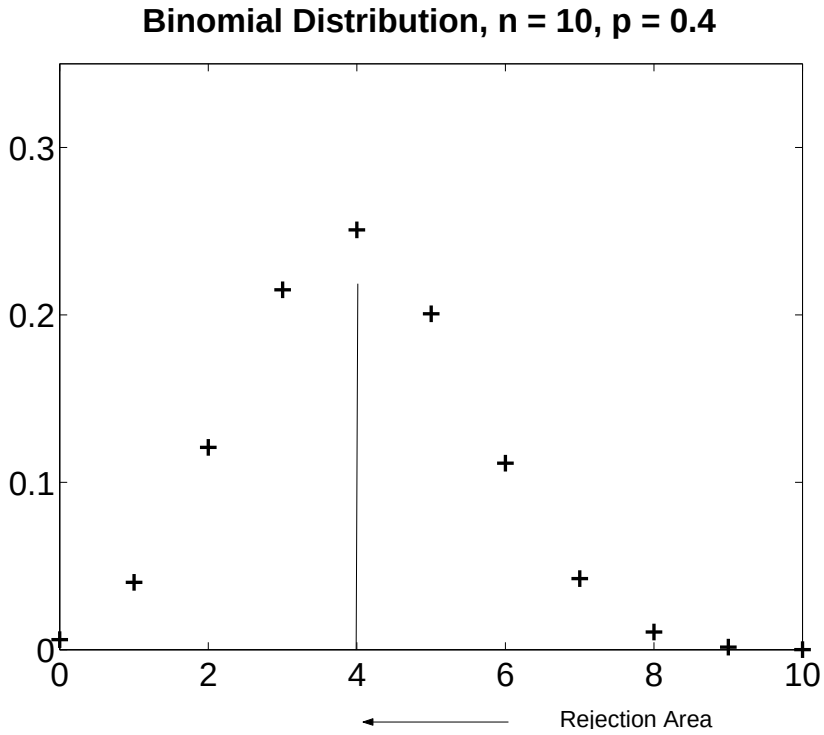
Questioning the customers the retailer finds that 4 out of 10 people are in favor of the better product, thus H_0 is rejected.



Errors

Suppose that the true parameter value is $p = 0.4$. Then

$$\begin{aligned}\beta &= P(\text{accept } H_0 \text{ while } p = 0.4) = P(x \leq 3 \text{ while } p = 0.4) \\ &= \sum_{x=0}^{3} \binom{10}{x} 0.4^x (1 - 0.4)^{10-x} \approx 0.3823\,.\end{aligned}$$



Using the Normal Distribution

In practice we can often use the normal distribution to find acceptance and rejection intervals. Suppose we want to test

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0$$

where θ is a parameter of a population (probability, mean, etc.). θ_0 is the value that we think θ has. We assume that the estimator $\hat{\theta}$ that we get from the sample has normal distribution with mean θ_0 and standard deviation $\sigma_{\hat{\theta}}$. Then

$$\text{statistics } z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

has a standard normal distribution. If the rejection region is

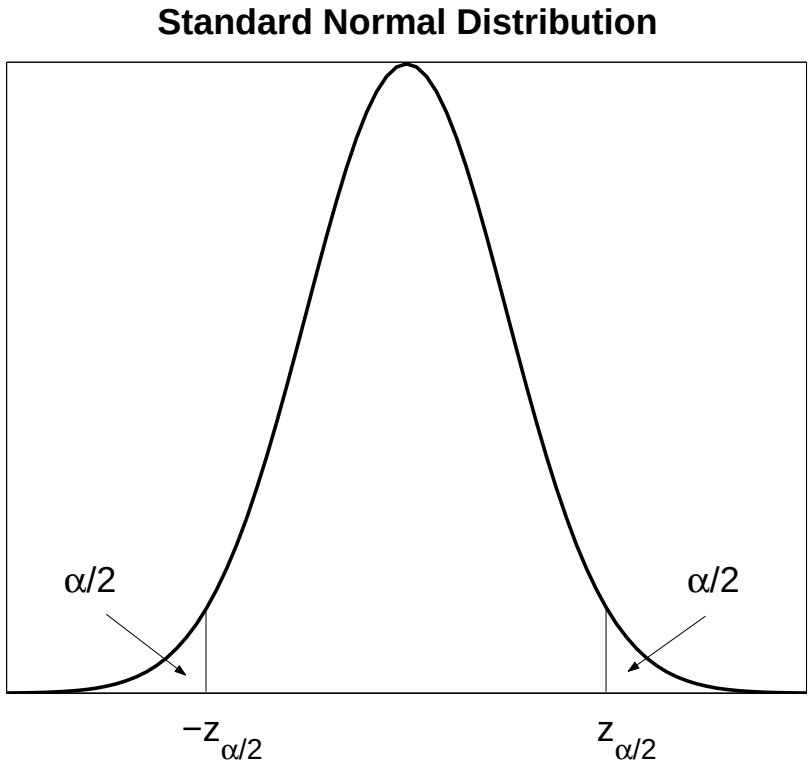
$$z < -z_{\alpha/2}, \qquad z_{\alpha/2} < z$$

then the type I error is α.

Using the Normal Distribution

Indeed,

$$\begin{aligned} & \text{type I error} \\ = \; & P(H_0 \text{ holds but is rejected}) \\ = \; & P(z \leq -z_{\alpha/2} \text{ or } z_{\alpha/2} \leq z) \\ = \; & \alpha\,. \end{aligned}$$



Such a test is called a *two-tailed test*.

Using the Normal Distribution

For a *one-tailed test* the data are

- $H_0 : \theta = \theta_0$;
- $H_1 : \theta > \theta_0$ $(\theta < \theta_0)$;
- Statistics: $z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$;
- Rejection region: $z > z_\alpha$ $(z < -z_\alpha)$;
- Type I error: α.

In practice, we are often given α in advance specifying the type I error probability that we are willing to accept, and we use this to find the acceptance and rejection interval.

Test About a Mean

One-tailed test	Two-tailed test
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu > \mu_0$	$H_1 : \mu \neq \mu_0$
(or $H_1 : \mu < \mu_0$)	
$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$	$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$
Rejection region:	Rejection region:
$z > z_\alpha$ (or $z < -z_\alpha$)	$z < -z_{\alpha/2}$ or $z_{\alpha/2} < z$

If the sample size is small ($n < 30$) or if $\sigma_{\bar{x}}$ has to be estimated by $s/\sqrt{n}$ then the normal distribution is replaced by the t-distribution with $n - 1$ degrees of freedom.

Test About a Mean

Example. We go back to the machines making wires with diameter approximately 1mm. Data taken from two machines showed the following values:

I:	1.0429	1.0627	0.9203	0.9280	1.0286
	0.9800	1.0345	1.0408	1.0356	1.0645
II:	1.0001	1.0157	0.9439	0.9794	0.9753
	0.9319	0.9877	0.9483	1.0225	0.9558

with $n_1 = n_2 = 10$, $\bar{x}_1 = 1.0138$, $s_1 = 0.0526$, $\bar{x}_2 = 0.9761$, and $s_2 = 0.0309$.

For

$$t_i = \frac{\bar{x}_i - 1.0}{s_i/\sqrt{10}}$$

we find $t_1 = 0.7877$ and $t_2 = -2.4459$. With $\alpha = 0.05$ we consider the following tests:

- $H_0 : \bar{x}_1 = 1.0$,

Test About a Mean

- $H_1 : \bar{x}_1 < 1.0$.

Since $t_1 \not> t_{9,0.05} = 1.8331$ the hypothesis H_0 is accepted.

- $H_0 : \bar{x}_1 = 1.0$,
- $H_1 : \bar{x}_1 \neq 1.0$.

Since $t_{9,0.025} = 2.2622$ and $-2.2622 < t_1 < 2.2622$ the hypothesis is accepted.

In both cases the decision is 'correct'; the data was random data with $\mu = 1.0$ and $\sigma = 0.05$.

Test About a Mean

For the second sample we consider

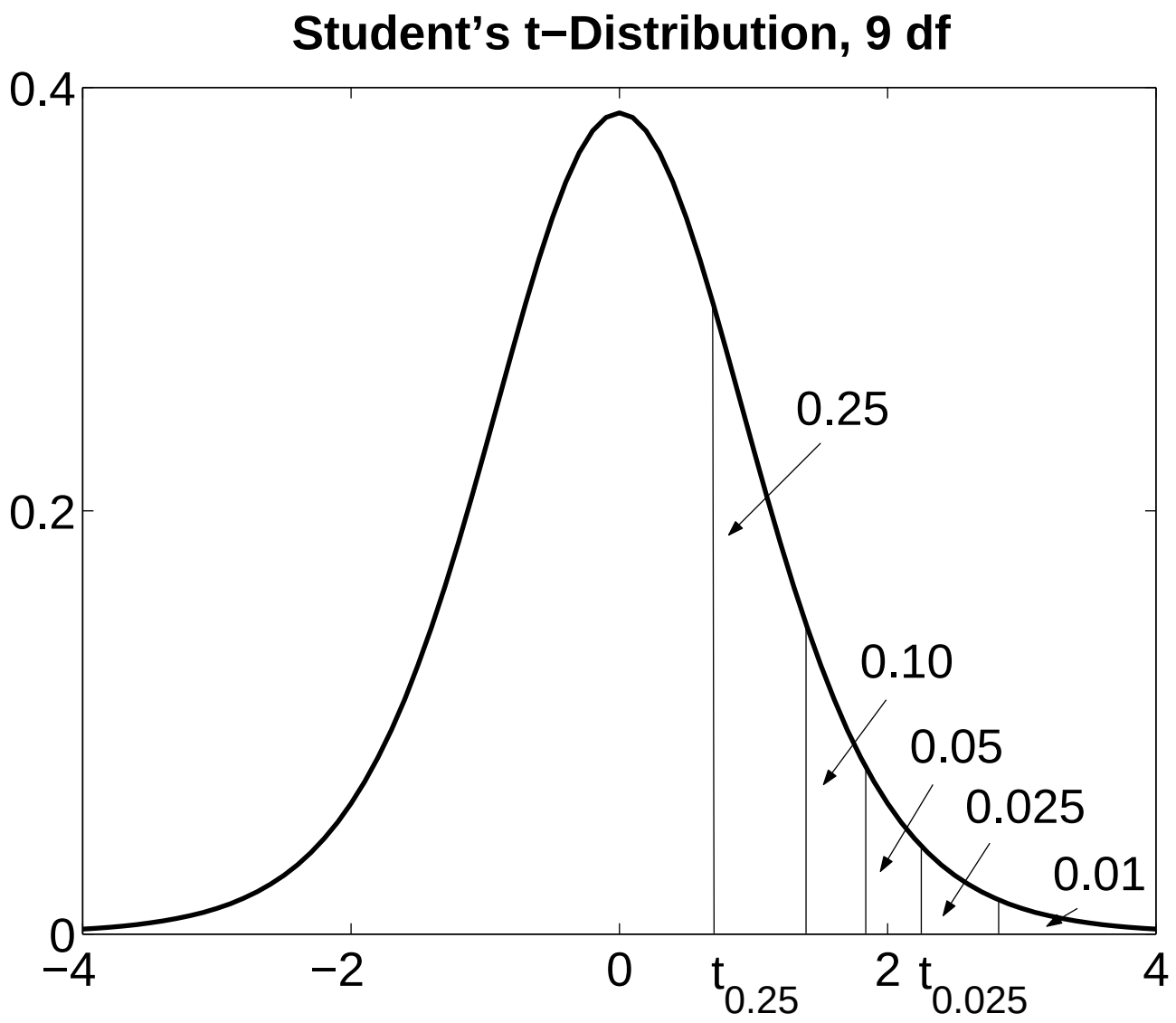
- $H_0 : \bar{x}_2 = 1.0$,
- $H_1 : \bar{x}_2 < 1.0$.

Since $t_2 = -2.4459 < t_{9,0.05} = -1.833$ the hypothesis is rejected. For the two-tailed test

- $H_0 : \bar{x}_2 = 1.0$,
- $H_1 : \bar{x}_2 \neq 1.0$

we see that $t_2 < -t_{9,0.025} = -2.262$, and we reject H_0 again. □

Test About a Mean



Testing For the Difference Between Means

One-tailed test	Two tailed test
$H_0 : \mu_1 - \mu_2 = d$	$H_0 : \mu_1 - \mu_2 = d$
$H_1 : \mu_1 - \mu_2 < d$	$H_1 : \mu_1 - \mu_2 \neq d$
(or $H_1 : \mu_1 - \mu_2 > d$)	
$z = \frac{(\bar{x}_1 - \bar{x}_2) - d}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$z = \frac{(\bar{x}_1 - \bar{x}_2) - d}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
Rejection region:	Rejection region:
$z > z_\alpha$ (or $z < z_\alpha$)	$z < -z_{\alpha/2}$ or $z_{\alpha/2} < z$

Testing For the Difference Between Means

If the sample sizes are small and σ_1 and σ_2 are unknown then the t-distribution with $n_1 + n_2 - 2$ degrees of freedom replaces the normal distribution, with

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d}{\sqrt{s_P^2(\frac{1}{n_1} + \frac{1}{n_2})}} \qquad \text{where} \qquad s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

provided that the two unknown variances are equal.

Example. The response times of two hard drives are tested. The values found are

Disk 1	Disk 2
$n_1 = 15$	$n_2 = 13$
$\bar{x}_1 = 16$	$\bar{x}_2 = 13$
$s_1 = 5$	$s_2 = 4$

What can be said about the difference between the mean response times?

Testing For the Difference Between Means

Here

$$H_0 : (\mu_1 - \mu_2) = 0, \qquad H_1 : (\mu_1 - \mu_2) \neq 0\,.$$

We calculate

$$\begin{aligned} s_P^2 &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \\ &= \frac{14 \cdot 5^2 + 12 \cdot 4^2}{15+13-2} \\ &= 20.8462 \end{aligned}$$

and

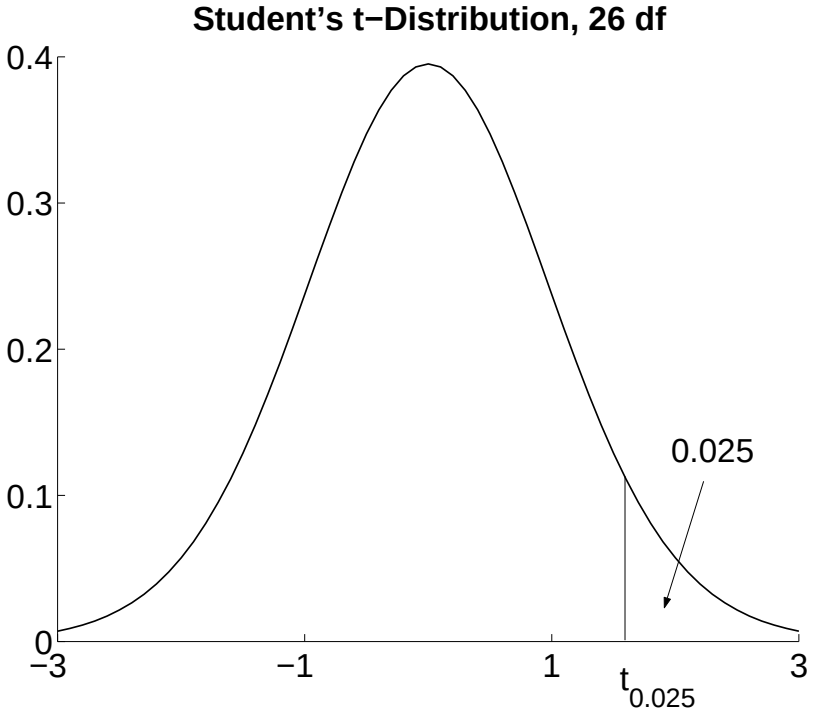
$$\begin{aligned} t &= \frac{(\bar{x}_1 - \bar{x}_2) - d}{\sqrt{s_P^2(\frac{1}{n_1} + \frac{1}{n_2})}} \\ &= \frac{16-13}{\sqrt{20.8462(\frac{1}{15} + \frac{1}{13})}} \\ &= 1.7340\,. \end{aligned}$$

Testing For the Difference Between Means

From the tables we know that for $n_1 + n_2 - 2 = 26$ and $\alpha = 0.1$ (for example) that

$$t_{0.05,26} = 1.706\,.$$

Since $t_{0.05,26} < t$ the hypothesis is rejected.



□

Testing For the Difference Between Means

Example. A study is done in the effectiveness of certain exercises to help weight loss. The following data are collected (in kg):

before	after	before	after
106	99	86	83
90	87	78	77
86	86	92	92
107	104	83	82
91	90	101	100
97	96	90	88
80	81	22	116
91	91	73	71

For the random variables before b and after a we find $\bar{b} = 92.0625$, $s_b = 12.3584$, $\bar{a} = 90.0625$, and $s_a = 11.0843$. Our hypotheses are

Testing For the Difference Between Means

- $H_0 : \bar{b} - \bar{a} = 0$,
- $H_1 : \bar{b} - \bar{a} > 0$,

and we will test at a 5% level of significance.

Then $t_{30,0.05} = 1.699$, and we reject H_0 if $t \geq 1.699$. To calculate further,

$$\begin{aligned} s_P^2 &= \frac{(n_b - 1)s_b^2 + (n_a - 1)s_a^2}{n_b + n_a - 2} \\ &= \frac{15}{30(s_b^2 + s_a^2)} \\ &= \frac{1}{2}(12.3584^2 + 11.0843^2) \\ &= 137.79562\,. \end{aligned}$$

Testing For the Difference Between Means

For t we find

$$\begin{aligned} t &= \frac{(\bar{b} - \bar{a}) - 0}{\sqrt{s_P^2(\frac{1}{n_b} + \frac{1}{n_a})}} \\ &= \frac{92.0625 - 90.0625}{\sqrt{137.79562 \cdot \frac{1}{5}}} \\ &= 0.38098\,. \end{aligned}$$

The hypothesis is accepted, and there is evidence that the exercises help reducing weight. □

Tests Concerning the Variance

Given a random sample from a normal population we will test the null hypothesis $\sigma^2 = \sigma_0^2$ against the alternatives $\sigma^2 \neq \sigma_0^2$ or $\sigma^2 < \sigma_0^2$ ($\sigma^2 > \sigma_0^2$). The random variable

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

has a χ^2 distribution with $n-1$ degrees of freedom. For a two-tailed test the null hypothesis is rejected if

$$\chi^2 \leq \chi^2_{1-\alpha/2,n-1} \qquad \text{or} \qquad \chi^2 \geq \chi^2_{\alpha/2,n-1}\,.$$

For a one-tailed test and the alternative hypothesis $\sigma^2 < \sigma_0^2$ we reject H_0 if $\chi^2 \leq \chi^2_{1-\alpha,n-1}$.

Example. Thickness of a semi-conductor part (in 10^{-5}m) is crucial in a production process. The machine manufacturing these semi-conductors needs to be readjusted if $\sigma^2 \leq 0.36$.

If in a sample of 20 measurements we find $s^2 = 0.74$, what can be said

at a $\alpha = 0.05$ level of significance?

We assume that thickness is normally distributed. Then

- H_0: $\sigma^2 = 0.36$,
- H_1: $\sigma^2 > 0.36$.

We reject the null hypothesis if $\chi^2 \geq \chi^2_{0.05,19} = 30.144$. With $s^2 = 0.74$, $\sigma_0^2 = 0.36$ and $n = 20$ we find

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{19 \cdot 0.74}{0.36} = 34.944\,,$$

and the machine needs to be readjusted.

Note that for $n = 20$, $\sigma_0^2 = 0.36$, and $\alpha = 0.05$ the machine needs readjustment if for a sample of 20 measurements the sample variation s^2 is greater or equal than 0.571. □

Testing for Proportions

Example. Suppose 5 out of 20 transistors are faulty. We test the hypothesis

- $H_0 : p = 0.5$,
- $H_1 : p \neq 0.5$,

at the 0.05 level of significance.

Instead of determining the rejection and acceptance interval we will find the smallest α which will reject H_0 (note for the calculation that the binomial distribution is symmetric):

$$\begin{aligned} \alpha/2 &= P(x \leq 5) \\ &= \sum_{x=0}^{5} \binom{20}{x} 0.5^x 0.5^{20-x} \\ &= 0.0207\,, \end{aligned}$$

so that $\alpha = 0.0414$. Since $\alpha < 0.05$ we will reject H_0. □

Testing for Proportions

If in a binomial test the size n is large we can use the normal distribution (with or without continuity correction) as an approximation for the random variable x. Then we get

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

$$z = \frac{\bar{x} - np_0}{\sqrt{np_0(1-p_0)}} \qquad \text{or } z = \frac{(\bar{x} \pm \frac{1}{2}) - np_0}{\sqrt{np_o(1-p_0)}}$$

Rejection region:

$$z < -z_{\alpha/2} \text{ or } z_{\alpha/2} < z$$

(If we use the correction factor we use a minus when x exceeds np_0, and a plus when x is less than np_0.)

For the on-tailed test

$$H_0 : p = p_0$$

$$H_1 : p > p_0$$

Testing for Proportions

we use the same statistics z as above with rejection interval $z \geq z_\alpha$.

Example. Suppose $p_0 = 0.2$ and we test

- $H_0 : p = 0.2$,
- $H_1 : p < 0.2$,

at the 0.01 level of significance. Then, using $z_{0.01} = 2.33$ we have the rejection region $z \leq -2.33$. If the test data are $n = 200$, $x = 22$, then

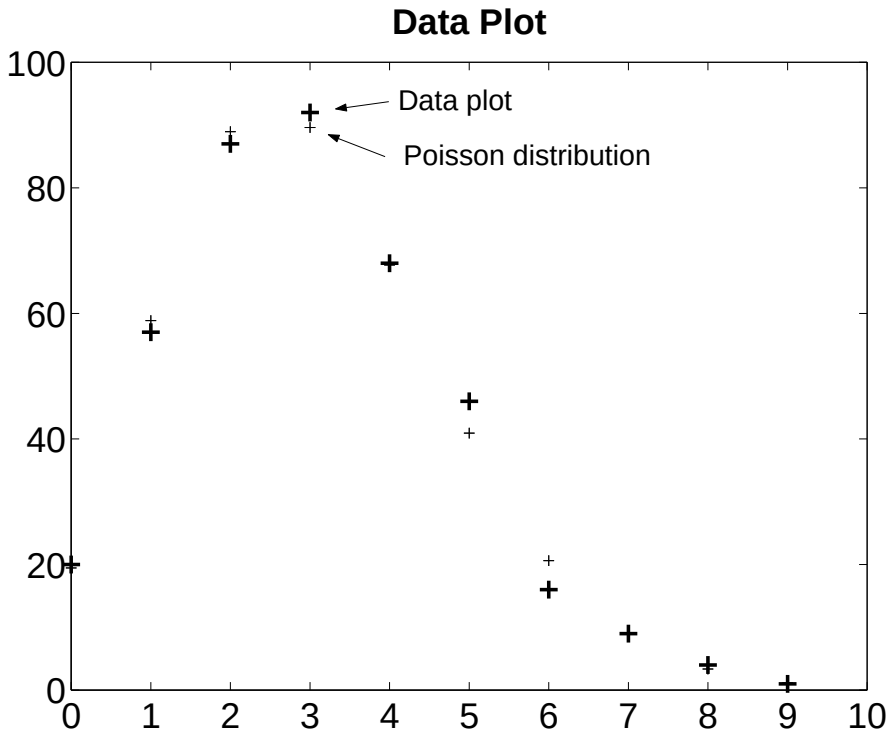
$$z = \frac{\bar{x} - np_0}{\sqrt{np_0(1-p_0)}} = \frac{22 - 200 \cdot 0.2}{\sqrt{200 \cdot 0.2 \cdot 0.8}} \approx -3.18\,,$$

and we reject the null hypothesis. □

Goodness-To-Fit

Goodness to fit tests are applied to test whether a set of data may be looked upon as a random sample from a population having a given distribution.

Suppose we have data from a Poisson distribution with $\lambda = 3$ (see next slide), which gives the following frequency diagram:



Goodness-To-Fit

x	Frequency f_i		Poisson $\lambda = 3$	Expected freq. e_i $(\lambda = 3.0225)$
0	20	0.0500	0.0498	19.4717
1	57	0.1425	0.1494	58.8534
2	87	0.2175	0.2240	88.9421
3	92	0.2300	0.2240	89.6092
4	68	0.1700	0.1680	67.7110
5	46	0.1150	0.1008	40.9313
6	16	0.0400	0.0504	20.6191
7	9	0.0225	0.0216	8.9030
8	4	0.0100	0.0081	3.3637
9	1	0.0025	0.0027	1.1296

For the expected frequency we first estimated λ using the third column

Goodness-To-Fit

as

$$\hat{\lambda} = 3.0225\,.$$

The random variable

$$\sum_{i=0}^{m} \frac{(f_i - e_i)^2}{e_i}$$

with m the number of different data (here 10) has a χ^2 distribution with $m - t - 1$ degrees of freedom, where t is the number of parameters estimated from the data (here 1).

Goodness-To-Fit

With the data above we want to test at a 0.05 level of significance whether the data are from a random variable having Poisson distribution. We set

- H_0: The data are from a Poisson random variable.
- H_1: The data are *not* from a Poisson random variable.

We reject H_0 if

$$\chi^2_{\alpha,m-t-1} \leq \chi^2 = \sum_{i=0}^{m} \frac{(f_i - e_i)^2}{e_i}$$

Here $m = 10$, $t = 1$, and $\chi^2_{0.05,10-1-1} = 15.507$. With our data, $\chi^2 = 1.9789$, and H_0 is accepted.

Summary

- Statistical tests often consist of a null hypothesis, and an alternative hypothesis. A type I error is made when the null hypothesis is true, but rejected. A type II error is made when the null hypothesis is false, but is accepted.
- We distinguish one-tailed and two-tailed tests.
- Statistical tests are based on sampling and confidence intervals. We thus use the normal distribution, the Student's t-distribution, and the chi-square distribution in standard tests.
- Goodness-To-Fit tests use the chi-square distribution to test whether a set of data fits a given distribution.

B34.UC2

Numerical Computation and Statistics in Engineering

Unit 6: Regression Analysis

Scatterplots

Scatterplots show the relationship between two quantitative variables. Examples are

- fuel consumption per speed;
- fuel consumption per weight;
- spending for leisure per income;
- etc.

Scatterplots

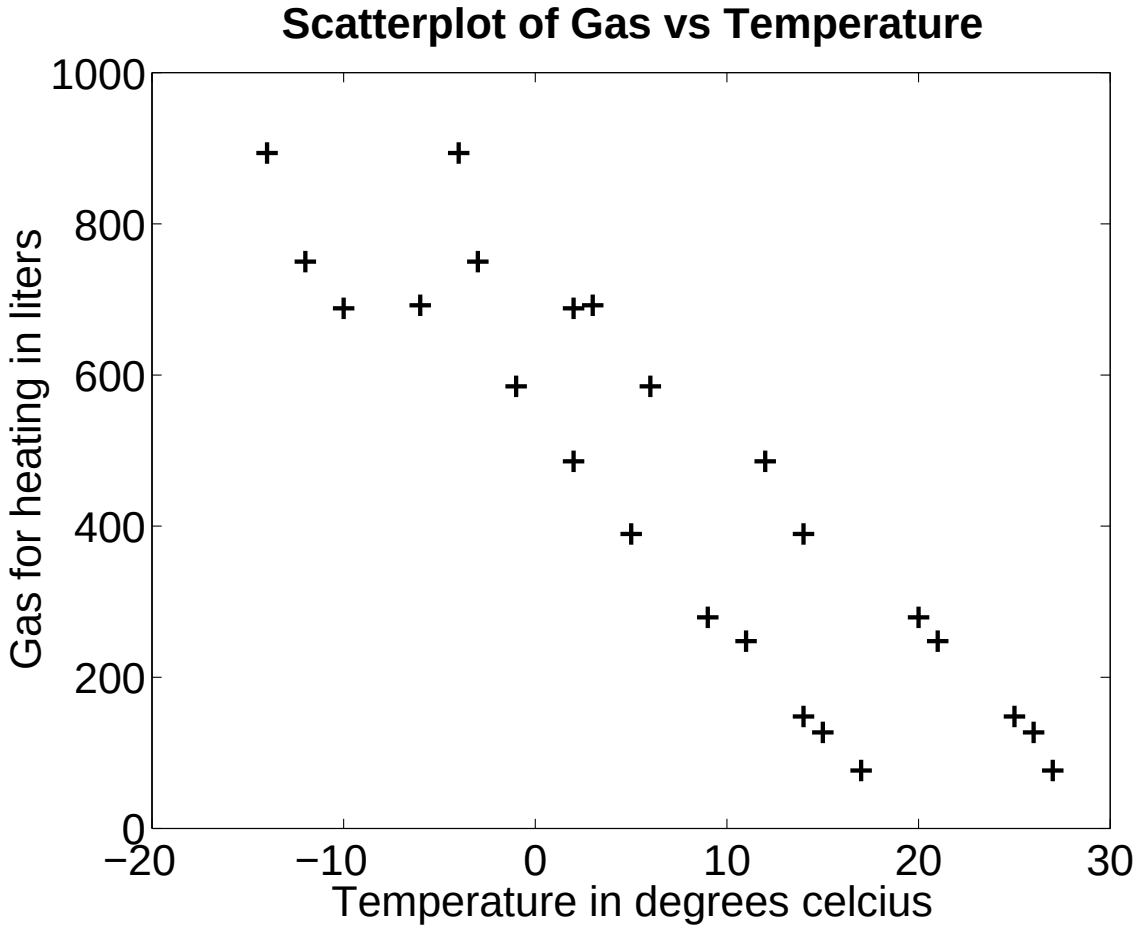
Example. We consider the following average high and low temperatures in Montreal (in degrees Celsius), and the measured gas consumption for heating (in liters) for a house:

	Jan.	Feb.	March	April	Mai	June
high	-4	-3	3	12	20	25
low	-14	-12	-6	2	9	14
gas	894	750	692	486	279	148
	July	Aug.	Sept.	Oct.	Nov.	Dec.
high	27	26	21	14	6	2
low	17	15	11	5	-1	-10
gas	77	127	248	390	584	688

The following is a data plot of the gas consumption against the low and

Scatterplots

hight temperatures. Both suggest a strong linear relationship.



□

Scatterplots

In general, interpreting a scatterplot we first look for an overall pattern, and describe form, direction, and strength.

Two variables are *positively associated* if above average values of one variable tend to result in above average values of the other. Two variables are *negatively associated* if above average values of one variable tend to result in below average values of the other.

Correlation

Correlation measures strength and direction of the *linear* relationship between two data sets. If the random variables are x and y, each consisting of n individual data, then the correlation r between x and y is defined as

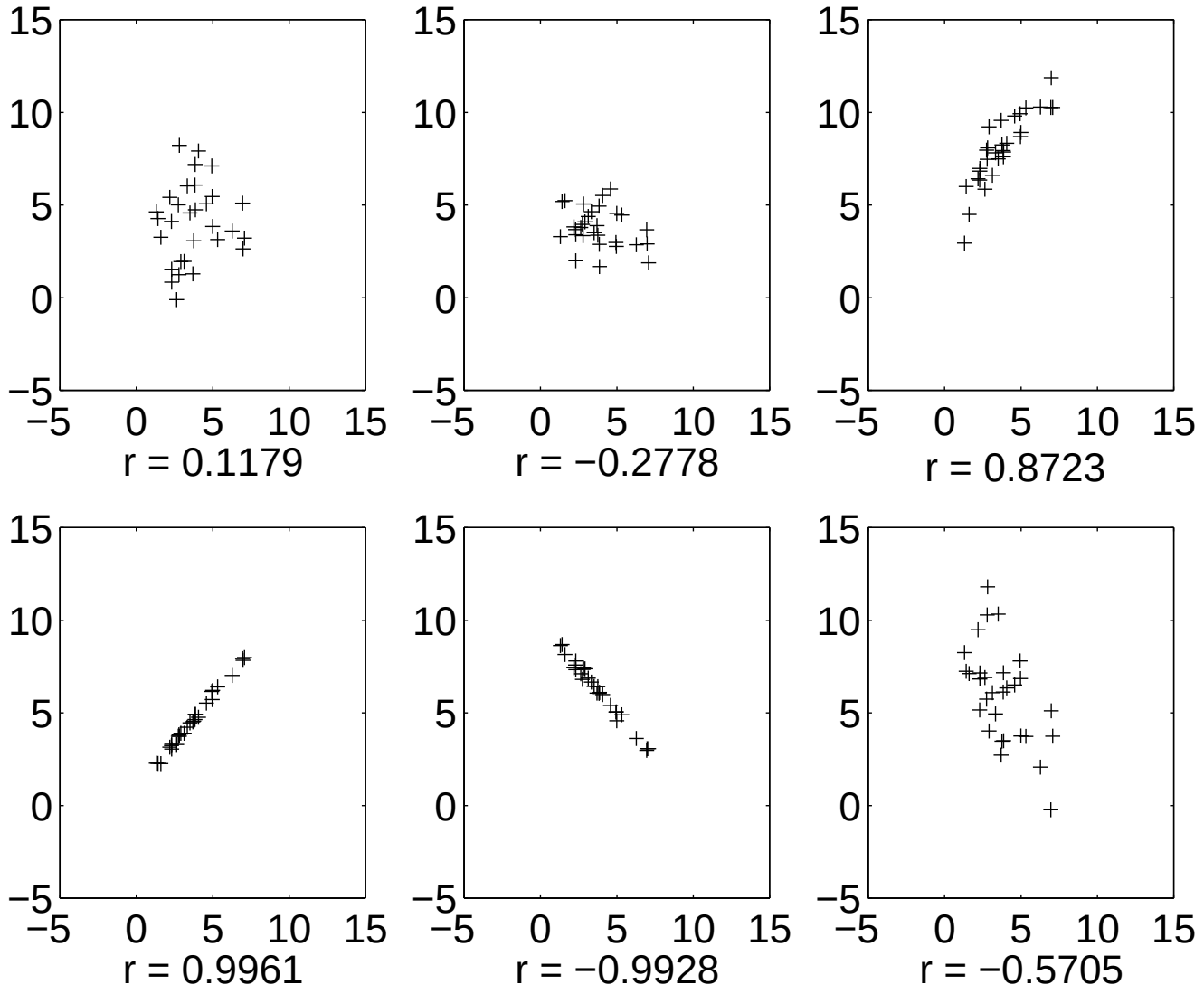
$$r = \frac{1}{n-1} \sum_i \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_x and s_y the sample errors. Correlation has the following properties:

- r does not have a dimension;
- the correlation is invariant under change of units of measurements;
- r is also independent from interchanging the role of x and y;
- r is *always* a number between -1 and 1.

Values close to -1 or 1 indicate strong linear relationships.

Correlation



Least Squares Regression Line

One of the variables is usually considered to be an *explanatory variable* (often denoted x), and the other a *response variable* (often denoted y).

A *regression line* is a line that describes how the response variable y changes when the explanatory variable x changes. Regression lines are used, among others, for prediction.

We use the notation

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$$

to denote a regression line for the data set x and y. Here $\hat{y}$ stands for the *predicted* value of the response variable (as opposed to the *observed* value). The quantity

$$\epsilon_i = \hat{y}_i - y_i = (\hat{\beta}_1 x_i + \hat{\beta}_0) - y_i$$

is called the error (or residual) of the observed data y_i, and is the vertical (!) distance between y_i and the regression line.

The *least squares regression line* is the regression line that minimizes the

Least Squares Regression Line

sum

$$\sum_i \epsilon_i^2 = \sum_i (\hat{y}_i - y_i)^2 \, .$$

The least squares regression line $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ has slope

$$\hat{\beta}_1 = r \frac{s_y}{s_x} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})(x_i - \bar{x})}$$

and intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \, .$$

- In least squares regression analysis the role of x and y are distinct. Changing the role of x and y gives *different* regression lines.
- A change of one standard deviation in x corresponds to a change of r standard deviation in y.
- Least square regression lines are sensitive to outliers.

How do we find the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the least squares regression line? We want to minimize $E(\hat{\beta}_0, \hat{\beta}_1) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (\hat{\beta}_0 +$

Least Squares Regression Line

$\hat{\beta}_1 x_i))^2$. Then

$$\begin{aligned} \frac{\partial E}{\partial \hat{\beta}_0} &= \sum\nolimits_i 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) \qquad \text{and} \\ \frac{\partial E}{\partial \hat{\beta}_1} &= \sum\nolimits_i 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i)\,. \end{aligned}$$

Setting both lines equal to 0 and simplifying gives

$$\begin{aligned} 0 &= \sum\nolimits_i (y_i - \hat{\beta}_1 x_i) - n\hat{\beta}_0 \\ &= -n\hat{\beta}_0 - \hat{\beta}_1 \sum\nolimits_i x_i + \sum\nolimits_i y_i \qquad \text{and} \\ 0 &= -\sum\nolimits_i x_i y_i + \hat{\beta}_0 \sum\nolimits_i x_i + \hat{\beta}_1 \sum\nolimits_i x_i^2\,. \end{aligned}$$

Solving this system of linear equations in $\hat{\beta}_0$ and $\hat{\beta}_1$ gives the formulae above.

Least Squares Regression Line

For our data relating the use of gas for heating we find the following parameters for the least squares regression lines:

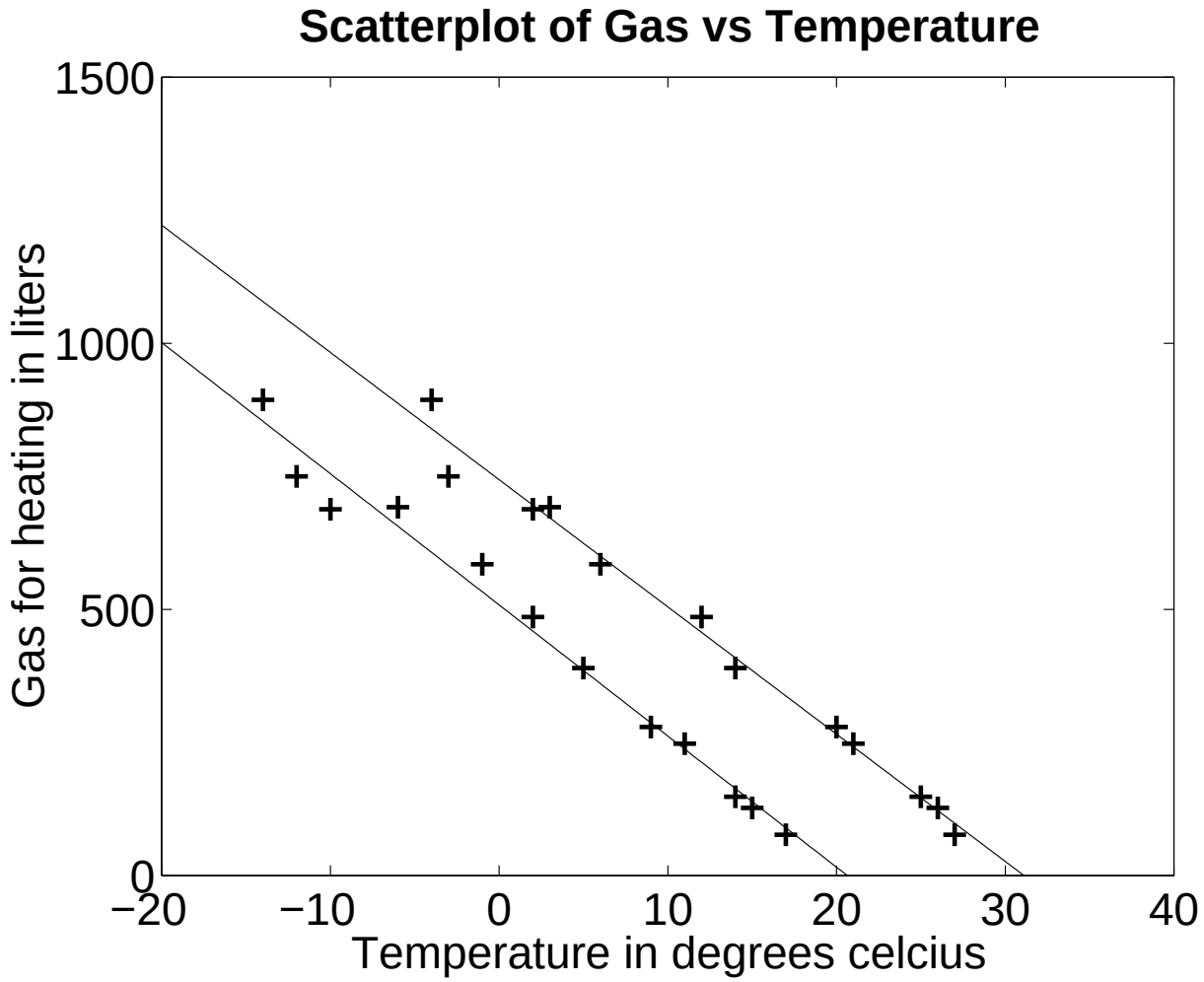
	r	s_x	s_y	$\hat{\beta}_1$	$\hat{\beta}_0$
high	-0.9940	11.4213	247.7005	-23.9072	743.8478
low	-0.9912	11.0495	247.7005	-24.6422	508.6054

The two equations for the regression lines are thus

$$\hat{y} = -23.9072x + 743.8478 \quad \text{and}$$
$$\hat{y} = -24.6422x + 508.6054\,.$$

Least Squares Regression Line

The following diagram shows the data and the regression lines:



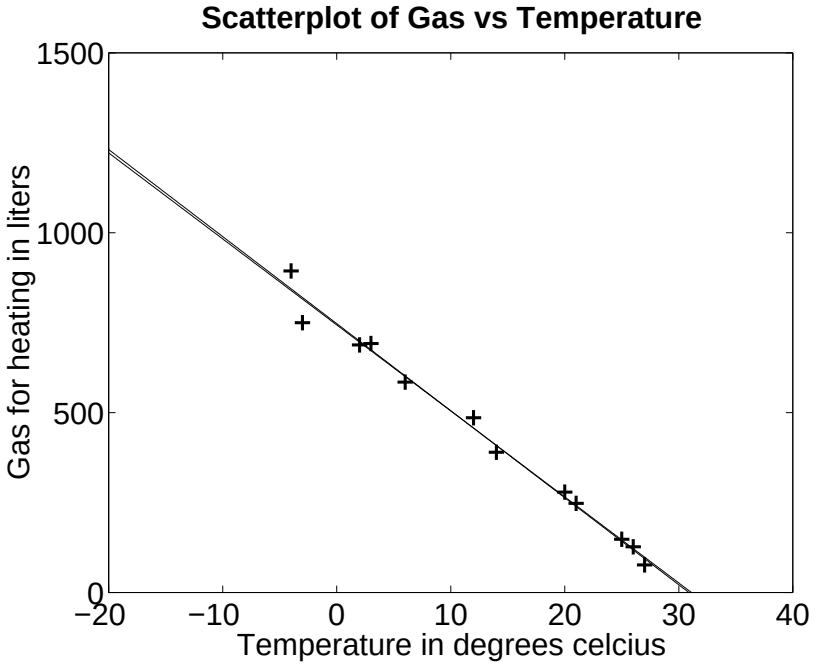
Least Squares Regression Line

If we interchange the role of x and y then we get the regression line

$$\hat{x} = \hat{\alpha}_1 y + \hat{\alpha}_0$$

with $\hat{\alpha}_1 = r\frac{s_x}{s_y}$ and $\hat{\alpha}_0 = \bar{x} - \hat{\alpha}_1 \bar{y}$. This yields the equation

$$\hat{x} = -0.0412y + 30.8903 \qquad \text{or} \qquad y = -24.1967\hat{x} + 747.4423\,.$$



Example

The following data represent the number of members in the EC Council of Ministers of current EC members and of potential EC members, and the populations of the member states:

	number	population (in 1,000,000)		number	population (in 1,000,000)
Germany	29	82.038	Portugal	12	9.980
Great-Britain	29	59.247	Sweden	10	8.854
France	29	58.966	Austria	10	8.082
Italy	29	57.610	Denmark	7	5.313
Spain	27	39.394	Finland	7	5.160
Netherlands	13	15.760	Irland	7	3.744
Grece	12	10.533	Luxembourg	4	0.429
Belgium	12	10.213			

The list of candidates is as follows, with the agreed number of seats

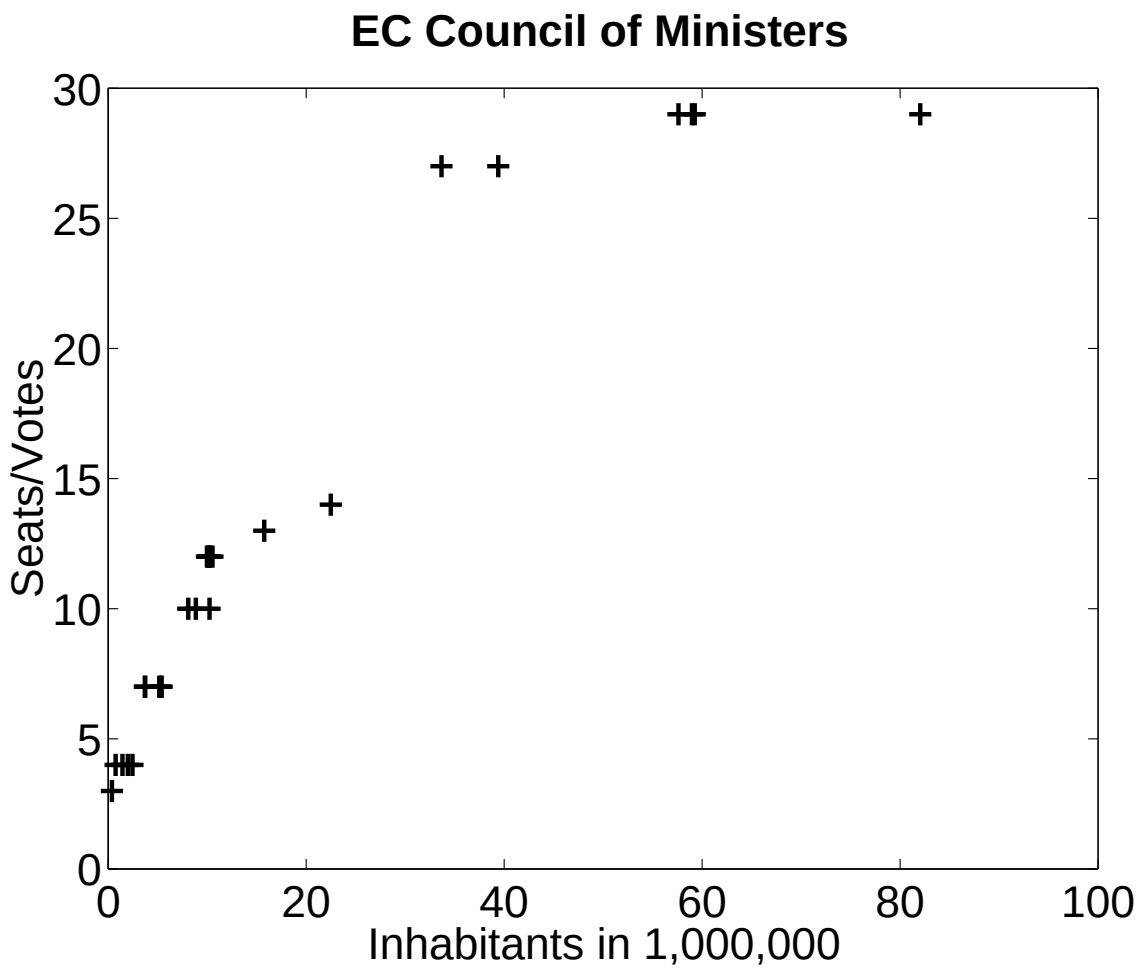
Example

according to the EC meeting in Nice end of 2000:

	number	population (in 1,000,000)		number	population (in 1,000,000)
Poland	27	33.667	Lithuania	7	3.701
Rumania	14	22.489	Letvia	4	2.439
Czech Republic	12	10.290	Slovenia	4	1.978
Hungaria	12	10.092	Estonia	4	1.446
Bulgaria	10	10.230	Cyprus	4	0.752
Slovakia	7	5.393	Malta	3	0.379

Example

The following diagram shows a scatterplot for these data:

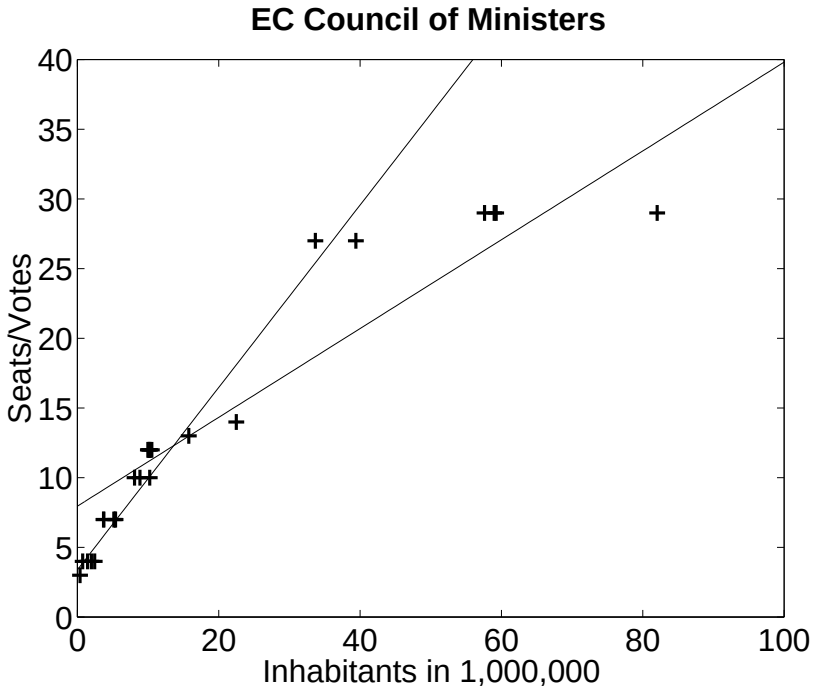


Example

For the regression lines we find

$$\hat{y} = 0.3186x + 7.9581 \qquad \text{and} \qquad \hat{y} = 0.6542x + 3.3924\,.$$

The correlation is 0.8930 for the first data set, and 0.9700 for the second.



The combined data set does suggest a logarithmic relationship.

Non-Linear Relationships

The following four data sets are from Frank J. Anscombe, Graphs in statistical analysis, *The American Statistician* 27: 17–21, 1973.

x_1	10	8	13	9	11	14	6	4	12	7	5
y_1	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
x_2	10	8	13	9	11	14	6	4	12	7	5
y_2	9.14	8.14	8.74	8.77	9.26	8.1	6.13	3.10	9.13	7.26	4.74
x_3	10	8	13	9	11	14	6	4	12	7	5
y_3	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
x_4	8	8	8	8	8	8	8	8	8	8	19
y_4	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

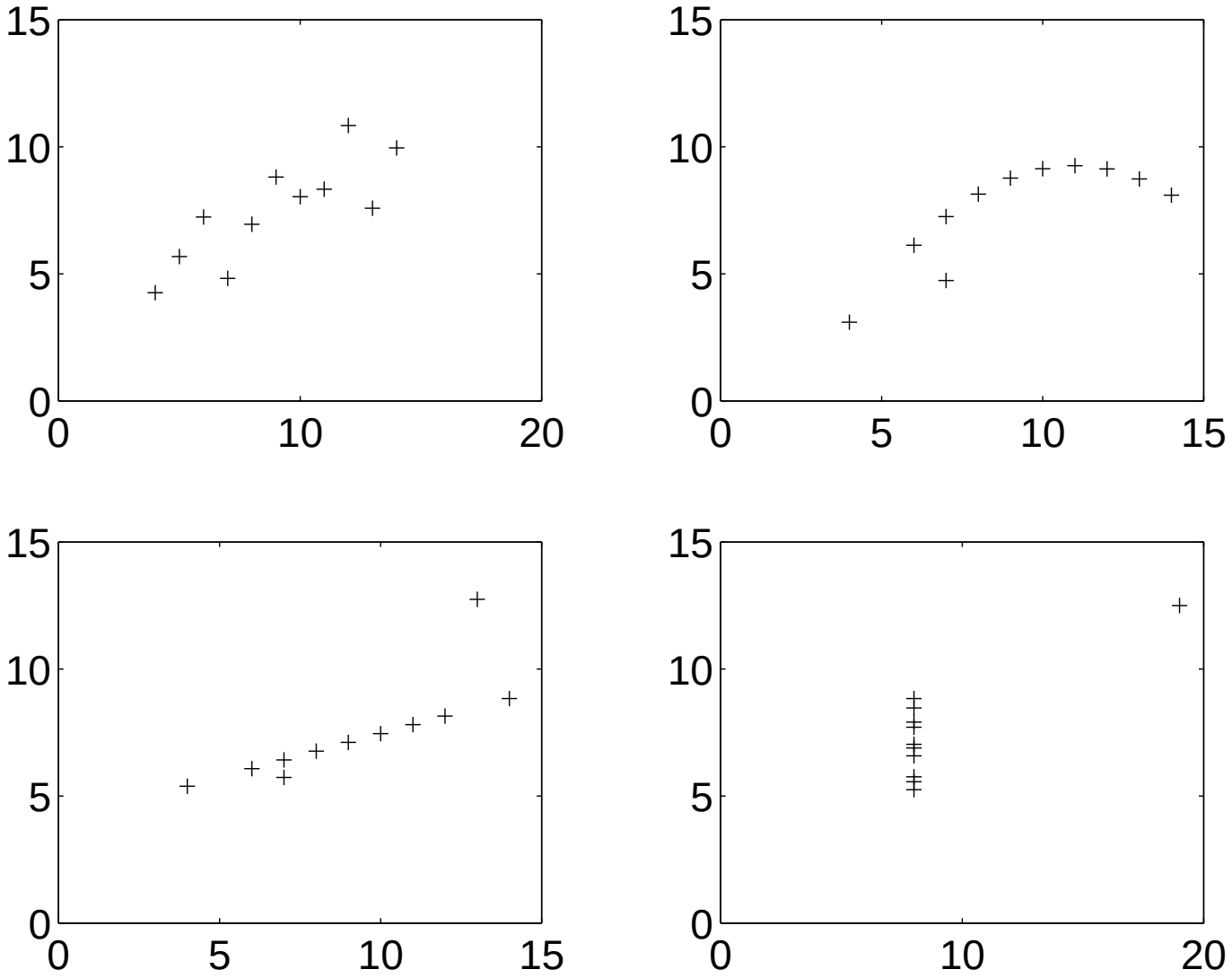
Non-Linear Relationships

The correlation and least squares lines are shown in the following list:

$$r_1 = 0.8164 \qquad \hat{y}_1 = 0.5001x_1 + 3.0001$$
$$r_2 = 0.8162 \qquad \hat{y}_2 = 0.5000x_2 + 3.0009$$
$$r_3 = 0.8163 \qquad \hat{y}_3 = 0.4997x_3 + 3.0025$$
$$r_4 = 0.8165 \qquad \hat{y}_4 = 0.4999x_4 + 3.0017$$

However, when plotting the data we realize that only the third and fourth represent strong linear relationships (both with one influential outlier). The first data set represents a moderate linear relationship, while the second represents a curved relationship.

Non-Linear Relationships



Confidence Intervals

We want to gain confidence in the least squares regression line. For this we have to make a couple of assumptions about ϵ:

- For the random variable ϵ we make the assumption that $E(\epsilon) = 0$.
- The standard deviation of ϵ is a constant σ.
- The distribution of ϵ is normal.
- Errors associated with different observations are independent.

Under the assumptions above let s^2 be

$$\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}.$$

Then $\frac{(n-2)s^2}{\sigma^2}$ has a chi-square distribution with $n-2$ degrees of freedom, and s^2 is an unbiased estimator for σ^2.

The standard deviation of ϵ can be interpreted as follows: We expect most observations y to lie within $2s$ of their least squares predicted value $\hat{y}$.

Confidence Intervals

If we make the four assumptions above then $\hat{\beta}_1$ has normal distribution with mean the true slope of the line and standard deviation

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}\,.$$

Here we use the notation

$$S_{uv} = \sum\nolimits_i (u_i - \bar{u})(v_i - \bar{v}) = \sum\nolimits_i u_i v_i - \frac{1}{n}\sum\nolimits_i u_i \sum\nolimits_i v_i$$

and note that

$$\sum (y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1 S_{xy}\,.$$

Confidence Intervals

We want to find confidence intervals for the slope. The random variable

$$\frac{\hat{\beta}_1 - \beta}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta}{s/\sqrt{S_{xx}}}$$

has a t-distribution with $n-2$ degrees of freedom (we cannot use the normal distribution since $\sigma_{\hat{\beta}_1}$ is estimated by $s_{\hat{\beta}_1}$). Thus, the endpoints of a $(1-\alpha)100\%$ confidence interval for the slope β_1 are

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} s_{\hat{\beta}_1} \qquad \text{where } s_{\hat{\beta}_1} = s/\sqrt{S_{xx}}\,.$$

Note that the endpoints of the interval are again of the form

$$\text{point estimator} \pm t_{n-2,\alpha/2} \text{ estimated standard error of the estimator}$$

Confidence Intervals

Example. The following data are given:

x_i	1	2	3	4	5	6
y_i	1	2	2	4	4	6

Here $\sum x_i = 21$, $\sum y_i = 19$, $\sum x_i^2 = 91$, $\sum y_i^2 = 77$ and $\sum x_i y_i = 83$.
It follows that

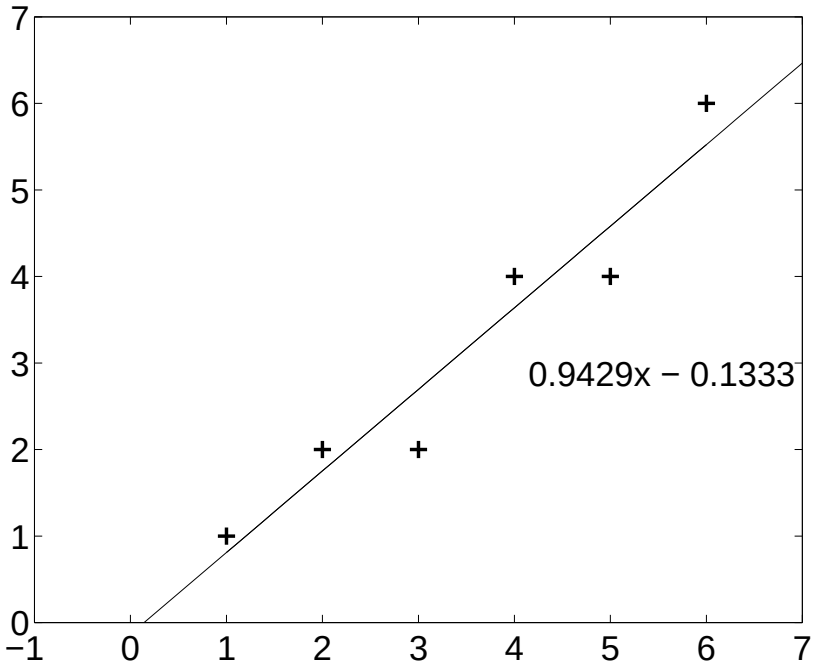
$$\begin{aligned} S_{xx} &= \sum x_i^2 - \frac{1}{6}(\sum x_i)^2 = 91 - \frac{21^2}{6} \\ &= 17.5\,, \quad \text{and} \\ S_{xy} &= \sum x_i y_i - \frac{1}{6}(\sum x_i)(\sum y_i) = 83 - \frac{21 \cdot 19}{6} \\ &= 16.5\,. \end{aligned}$$

Confidence Intervals

Thus we calculate for slope and intercept of the least squares regression line

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.9429\,,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{6} \cdot 19 - 0.9429 \cdot \frac{21}{6} = -0.13\,.$$



Confidence Intervals

An estimator for the variance s^2 of the error $\epsilon = y - \hat{y}$ is

$$\frac{\sum (y_i - \hat{y}_i)^2}{n-2} = S_{yy} - \hat{\beta}_1 S_{xy}\,.$$

With the data above we find that $S_{yy} = 77 - \frac{1}{6}(19)^2 = 16.8333$ so that

$$s^2 = 1.2755\,.$$

The endpoints of the 95% confidence interval for $\hat{\beta}_1$ are

$$\hat{\beta}_1 \pm t_{4,0.025} s_{\hat{\beta}_1} = 0.9429 \pm 2.776 \cdot \frac{\sqrt{1.2755}}{\sqrt{17.5}} = 0.9429 \pm 0.7494\,.$$

□

Multiple Linear Regression

By way of example we consider multiple linear regression. We consider the following table of apartment blocks sold in a big city:

#appartments	#floors	price (in 1,000,000)
60	10	78.2
40	5	45.4
80	10	100.0
30	6	35.7
60	3	80.5
40	6	42.8
90	12	120.4
80	7	90.5

We want to find the equation of a plane (!) allowing to predict the price z of an apartment block with x apartments and y floors using the

Multiple Linear Regression

method of least squares.

The equation of the plane is

$$\hat{z} = \hat{\alpha}x + \hat{\beta}y + \hat{\gamma}\,.$$

For observed data z_i we have the error

$$\epsilon_i = z_i - \hat{z}_i = z_i - \hat{\alpha}x_i - \hat{\beta}y_i - \hat{\gamma}\,,$$

and we try to minimize

$$E(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = \sum\nolimits_i \epsilon_i^2 = \sum\nolimits_i (z_i - \hat{\alpha}x_i - \hat{\beta}y_i - \hat{\gamma})^2\,.$$

Multiple Linear Regression

Then

$$\begin{aligned}
\frac{\partial E}{\partial \hat{\alpha}} &= \sum\nolimits_i 2(z_i - \hat{\alpha}x_i - \hat{\beta}y_i - \hat{\gamma})(-x_i) \\
\frac{\partial E}{\partial \hat{\beta}} &= \sum\nolimits_i 2(z_i - \hat{\alpha}x_i - \hat{\beta}y_i - \hat{\gamma})(-y_i) \\
\frac{\partial E}{\partial \hat{\gamma}} &= \sum\nolimits_i 2(z_i - \hat{\alpha}x_i - \hat{\beta}y_i - \hat{\gamma})(-1)
\end{aligned}$$

Thus we have to solve the system of linear equations

$$\begin{aligned}
0 &= -\sum z_i x_i + \hat{\alpha}\sum x_i^2 + \hat{\beta}\sum x_i y_i + \hat{\gamma}\sum x_i \\
0 &= -\sum z_i y_i + \hat{\alpha}\sum x_i y_i + \hat{\beta}\sum y_i^2 + \hat{\gamma}\sum y_i \\
0 &= -\sum z_i + \hat{\alpha}\sum x_i + \hat{\beta}\sum y_i + n\hat{\gamma}
\end{aligned}$$

Multiple Linear Regression

With the data above we find

$$\begin{array}{lllllllll} \sum x_i^2 & = & 33200 & \sum x_i y_i & = & 3840 & \sum x_i & = & 480 \\ \sum y_i^2 & = & 499 & \sum x_i z_i & = & 40197 & \sum y_i & = & 59 \\ \sum z_i^2 & = & 50499.4 & \sum y_i z_i & = & 4799.8 & \sum z_i & = & 593.5 \end{array}$$

Thus we have to solve the system of linear equations

$$\begin{pmatrix} 33200 & 3840 & 480 \\ 3840 & 499 & 59 \\ 480 & 59 & 8 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} 40197.0 \\ 4799.8 \\ 593.5 \end{pmatrix}.$$

Linear algebra (or MATLAB) tells us that the solution is

$$\hat{\alpha} = 1.3067, \quad \hat{\beta} = 0.4813, \quad \hat{\gamma} = -7.7611\,.$$

We can use the equation of the plane

$$\hat{y} = 1.3068x + 0.4813y - 7.7611$$

to predict the price of an apartment block with 50 apartments and 5 floors, which will have a predicted price of

$$\hat{z} = 1.3068 \cdot 50 + 0.4813 \cdot 5 - 7.7611 \approx 59.98$$

million.

The Previous Example in MATLAB

```
>> x = [60 40 80 30 60 40 90 80];
>> y = [10 5 10 6 3 6 12 7];
>> z = [78.2 45.4 100.0 35.7 80.5 42.8 120.4 90.5];
>> Sxx = sum(x.*x);
>> Syy = sum(y.*y);
>> Szz = sum(z.*z);
ans =
     1.0e+004 *
           3.2200     0.0499     5.0449
>> Sxy = sum(x.*y);
>> Sxz = sum(x.*z);
>> Syz = sum(z.*y);
>> [Sxx Syy Szz]
```

The Previous Example in MATLAB

```
>> [Sxy Sxz Syz]
ans =
     1.0e+004 *
           0.3840     4.0197     0.4800
>> [sum(x) sum(y) sum(z)]
ans =
     480.0000     59.0000     593.5000
>> A=[Sxx Sxy sum(x); Sxy Syy sum(y); sum(x) sum(y) 8]
A =
       32200        3840         480
        3840         499          59
         480          59           8
>> b=[Sxz; Syz; sum(z)]
```

The Previous Example in MATLAB

```
b =
  1.0e+004 *

    4.0197
    0.4800
    0.0594
>> inv(A)
ans =
    0.0005   -0.0024   -0.0127
   -0.0024    0.0267   -0.0556
   -0.0127   -0.0556    1.2996
>> (inv(A)*b)'
ans =
    1.3067     0.4813     -7.7611
```

Example

Example. A study is made into the response time to 911 calls. It is measured the time (in minutes) it takes the ambulance to arrive at the scene against the distance (in kilometers) between the station and the scene. The following data are collected:

dist. x	3.4	1.8	4.6	2.3	3.1
time y	2.5	1.8	3.0	2.6	2.9
dist. x	5.2	0.6	2.9	2.7	4.0
time y	3.9	1.6	2.3	2.0	3.4
dist. x	2.3	1.0	6.3	4.5	3.5
time y	2.5	1.8	2.6	2.8	2.8

Example

To find the least squares regression line $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ we first calculate

$$\begin{aligned} S_{xx} &= 33.5573 \\ S_{yy} &= 5.3933 \\ S_{xy} &= 10.0367 \end{aligned}$$

and thus

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{10.0367}{33.5573} = 0.2991\,, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 2.5667 - 0.2991 \cdot 3.2133 = 1.6056\,. \end{aligned}$$

and the least squares regression line is

$$\hat{y} = 0.2991x + 1.6056\,.$$

Next we look at the probability distribution of the random error compo-

Example

nent ϵ. For s^2 we find

$$s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{2.3915}{13} = 0.1840\,.$$

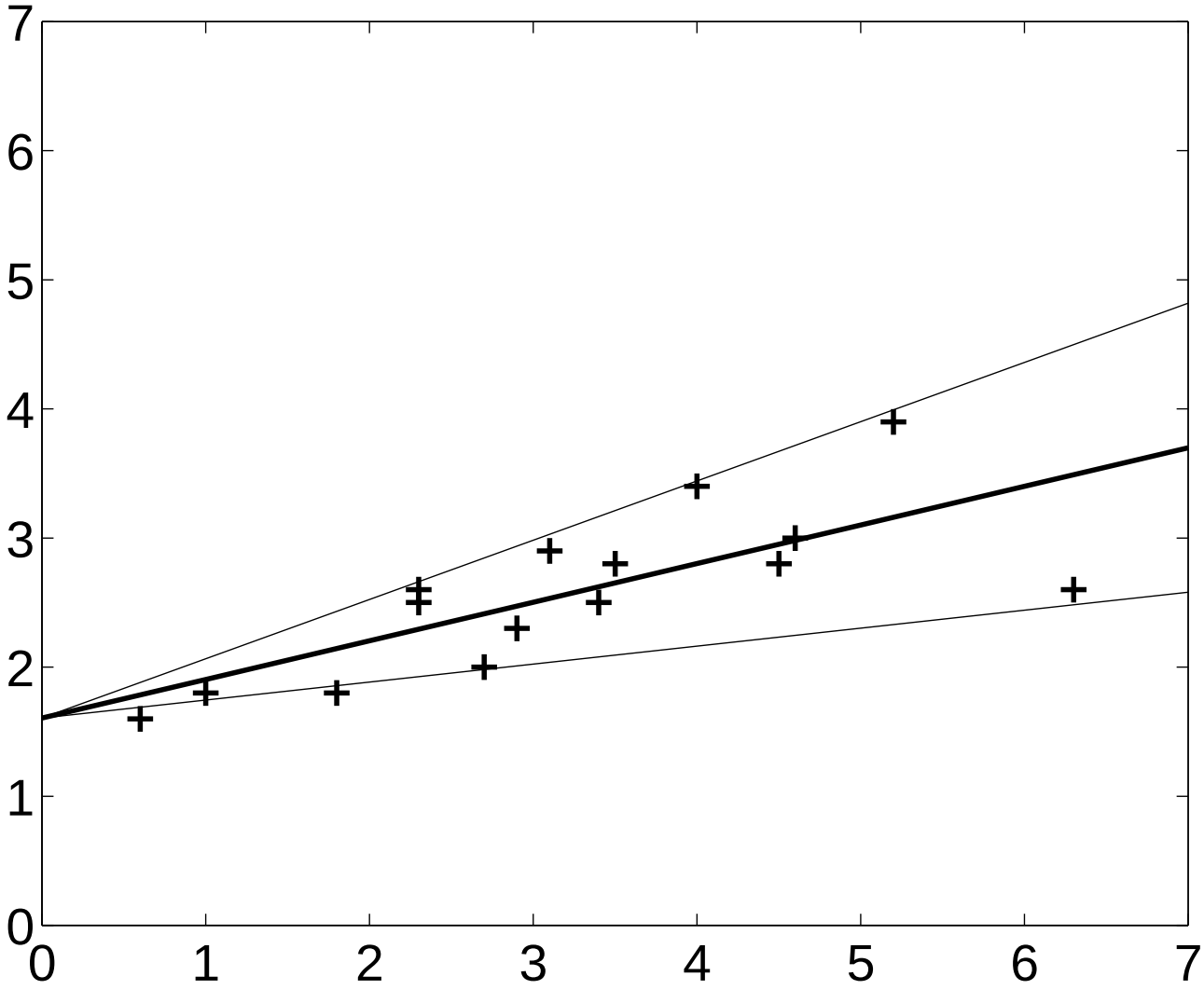
Thus $s = 0.4289$.

The endpoints of the 95% confidence interval for $\hat{\beta}_1$ are

$$\hat{\beta}_1 \pm t_{13,0.025} \frac{s}{\sqrt{S_{xx}}} = 0.2991 \pm 2.160 \frac{0.4289}{\sqrt{33.5573}}$$

which gives the interval $[0.1392, 0.4590]$. The following diagram shows the data set, the least squares regression line, and the two boundary lines for the interval estimating $\hat{\beta}_1$:

Example



□

Summary

- Regression analysis aims to find relationships between variables where there is an estimated dependency.
- Correlation measures strength and direction of a linear relationship between two data sets. However, correlation is sensitive to outliers.
- The least squares regression line is the line that fits best best two data sets. This line minimizes the vertical distance between the observed values and the predicted values on the line.
- Multiple linear relationships and non-linear relationships can be tackled with similar methods.