22.4UC2

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

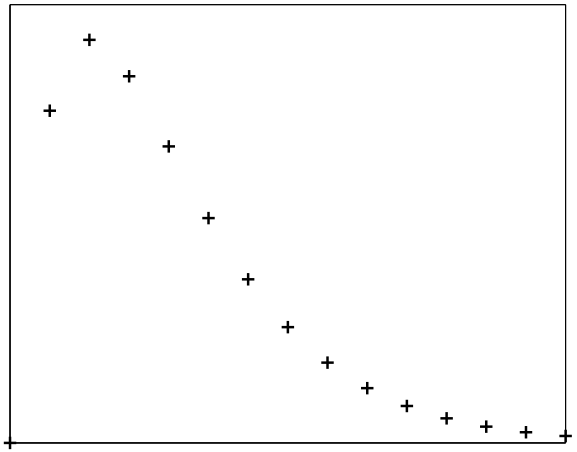
There are lots of different ways to represent quantitative data, for example tables, vertical or horizontal bar graphs, pie charts, scatter diagrams, etc. The following examples each show the money available for research at U.K. universities during the academic year 1998/99 (source: The Times Higher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

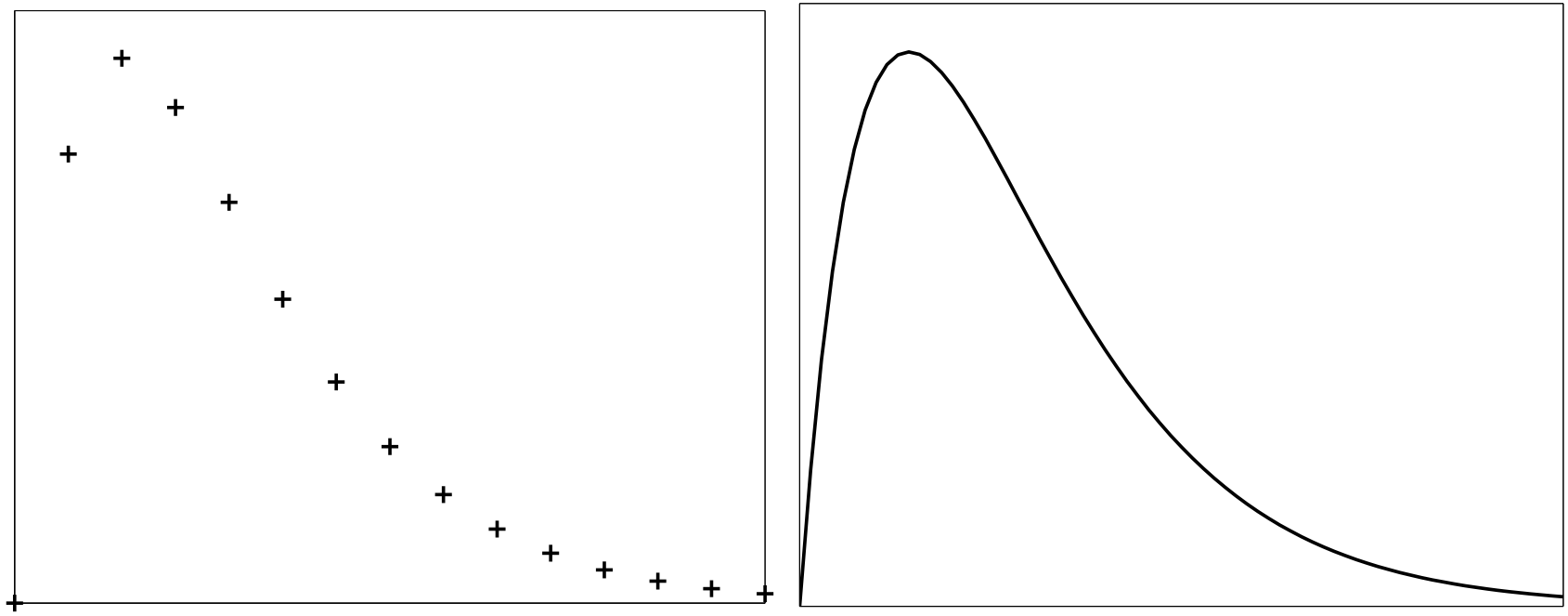


A variety of numerical values can be associated to a sample. These values aim

- 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.

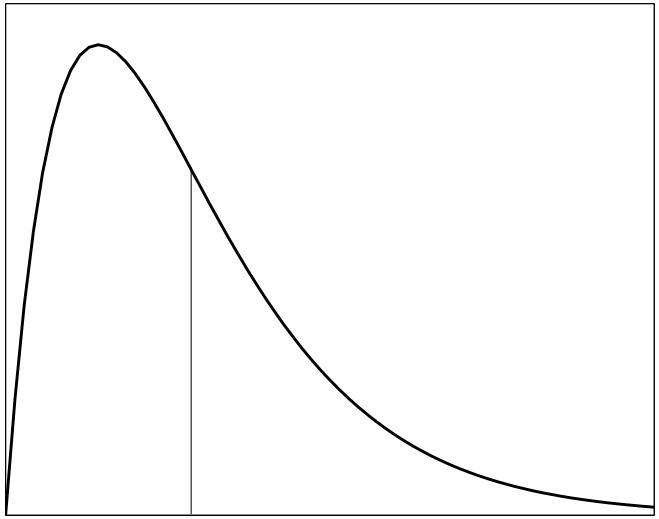
The *mode* of the sample $x_1, \ldots, x_n$ is the value which occurs with greatest frequency.

Example. Consider the measurement

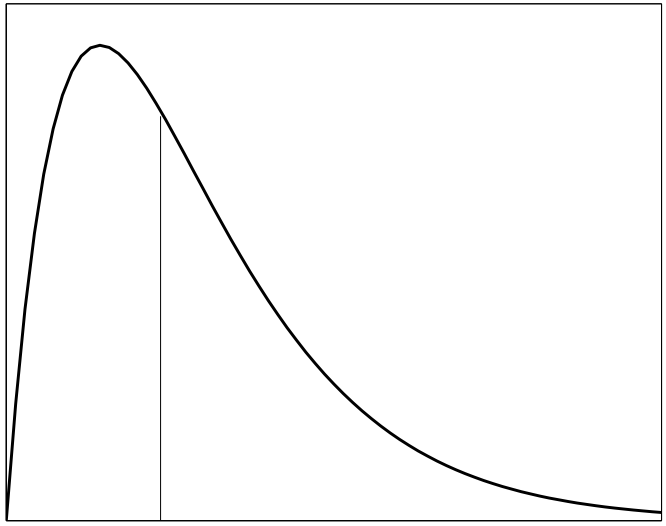
$$1, 2, 2, 2, 2, 3, 3, 4 .$$

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. □

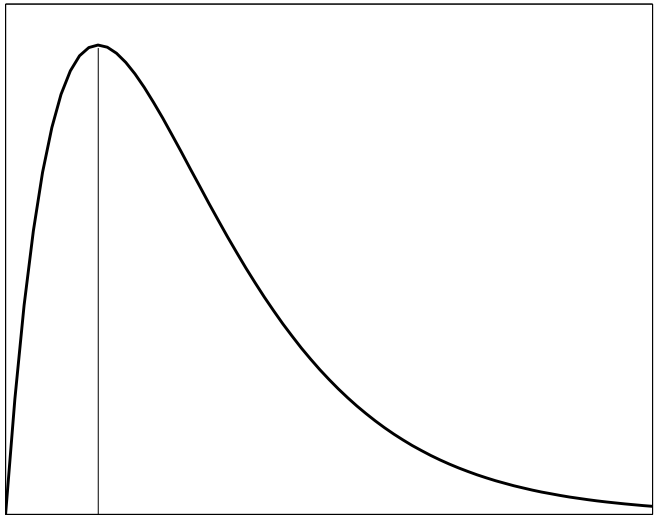
Measures of Central Tendency



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

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

□

Measures of Variation

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

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 (here, being in favor of the new product) of the total population.

Statistics provides scientific tools to enable such inferences with a probability of certainty, that is, provides methods to judge the reliability of such inferences.

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 Bayern-München 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

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

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 {HT}, whereas A or B is {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.

Sampling With/Without Replacement

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.