Applied Statistics

for Computer Science BSc, Exam

Probability Theory and Mathematical Statistics

for Computer Science Engineering BSc, Term grade

István Fazekas

University of Debrecen

2020/21 fall

This work was supported by the construction EFOP-3.4.3-16-2016-00021. The project was supported by the European Union, co-financed by the European Social Fund.

Main topics

1. Probability theory
2. Statistics

Mathematical tools: combinatorics, calculus

Computer tool: Matlab

Book:

Yates, Goodman:

Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers

Lecture 10

Basic notions of statistics

The scope of statistics

Statistical methods are used everywhere
Industry: quality control,...
Agriculture: estimating crops,...
Science: processing data of experiments,...
Medicine: spread of certain desease, efficiency of certain therapy,...
Government: data of a country, population, ...
Statistical offices
Eurostat: https://ec.europa.eu/eurostat
Hungarian Central Statistical Office, http://www.ksh.hu/?lang=en
Inflation, GDP, population, unemployment rate, industrial production

Methods of statistics

Mathematical tools: probability theory, calculus, discrete mathematics, numerical methods,...
Software tools: data base, statistical program packages, data mining, machine learning, data science,...
Statistical companies, statistical packages:
SAS (Statistical Analysis Software),
SPSS (Statistical Product and Service Solutions, IBM),
R (The R Project for Statistical Computing, free),
Matlab (integrated software tool for analysing data),...

Statistical sample

Probability Theory $\leftrightarrow$ Statistics
The mathematical base (models) of statistics are given by probability theory.
But in probability theory we suppose, that we know the distributions, the parameters,...
In statistics we have data only and we should find the unknown distributions and parameters form the data.
Statistical sample: data, measurements, observations,...

Statistical sample...

Definition.

$$X_1, X_2, \ldots, X_n$$

is called sample if they are independent, identically distributed random variables.
Let Ω be the background probability space. Then for $\omega \in \Omega$ the numbers

$$x_1 = X_1(\omega),\ x_2 = X_2(\omega), \ldots, x_n = X_n(\omega)$$

are called sample realization.
Sample: theory
Sample realization: practice
Example. We weight 10 tablets of chocolate. The results are

$$x_1 = 100.1,\ x_2 = 99.7,\ x_3 = 99.5, \ldots,\ x_{10} = 99.9$$

It is a sample realization.

Ordered statistics

Let $\omega \in \Omega$ be fixed.
Denote by

$$X_1^*(\omega) \leq X_2^*(\omega) \leq \cdots \leq X_n^*(\omega)$$

the increasing ordering of the sample realization
$X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$
The random variables

$$X_1^* \leq X_2^* \leq \cdots \leq X_n^*$$

are called order statistics (ordered sample).

Empirical distribution function (empirical CDF)

Let $X_1^*, X_2^*, \ldots, X_n^*$ the order statistics. Then the function

$$F_n^*(x) = \begin{cases} 0, & \text{if } x \leq X_1^*, \\ \frac{k}{n}, & \text{if } X_k^* < x \leq X_{k+1}^*, \ k = 1, \ldots, n-1, \\ 1, & \text{if } x > X_n^* \end{cases}$$

is called the empirical distribution function.
It is a step function jumping $\frac{1}{n}$ at any value of the sample.
F_n^* is a random function.

Empirical distribution function (empirical CDF)...

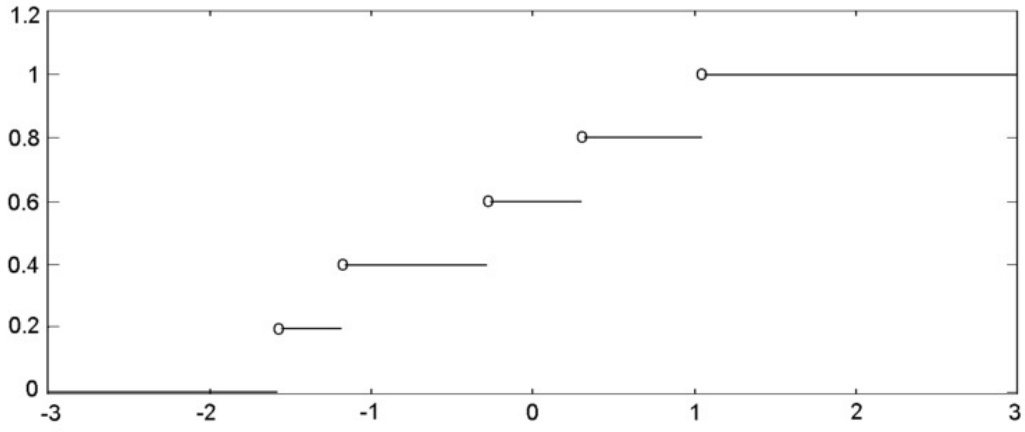

Figure: The empirical CDF of a 5-element sample

Empirical distribution function (empirical CDF)...

Theorem.
Let $X_1, X_2, \ldots, X_n$ be a sample from a population having CDF F.
Let $F_n^*(x)$ be the empirical CDF.
Then for any fixed $x \in \mathbb{R}$ we have

- a) $nF_n^*(x)$ has binomial distribution with parameters n and $p = F(x)$;
- b) the expectation of $F_n^*(x)$ is $F(x)$;
- c) the variance of $F_n^*(x)$ converges to 0 as $n \to \infty$;
- d) $F_n^*(x) \to F(x)$ in probability (stochastically) as $n \to \infty$.

Empirical distribution function (empirical CDF)...

Proof.

a) The event $\{nF_n^*(x) = k\}$ occurs if precisely k out of the n sample values are less than x. The events $\{X_k < x\}$, $k = 1, \ldots, n$, are independent and they have probability $p = F(x)$, so we obtain

$$P(nF_n^*(x) = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

for an fixed x.

b) The expectation of the binomial distribution with parameters n and p is np, its variance is $np(1-p)$. Therefore

$$\mathbb{E}F_n^*(x) = p = F(x),$$

$$\mathrm{Var}F_n^*(x) = \frac{p(1-p)}{n} = \frac{F(x)(1-F(x))}{n}.$$

Empirical distribution function (empirical CDF)...

Proof...

c)

$$\mathrm{Var}F_n^*(x) = \frac{p(1-p)}{n} = \frac{F(x)(1-F(x))}{n} \to 0$$

as $n \to \infty$.

d) For $\varepsilon > 0$, by Chebyshev's inequality,

$$P(|F_n^*(x) - F(x)| > \varepsilon) \leq \frac{1}{n}\frac{F(x)(1-F(x))}{\varepsilon^2} \to 0$$

as $n \to \infty$.

Empirical distribution function (empirical CDF)...

Theorem. For $x \in \mathbb{R}$

$$\lim_{n\to\infty} F_n^*(x) = F(x) \quad \text{almost surely,}$$

$$\lim_{n\to\infty} F_n^*(x+0) = F(x+0) \quad \text{almost surely.}$$

In the above theorem we have convergence almost surely which is stronger than stochastic convergence in the preceding theorem.

Empirical distribution function (empirical CDF)...

Proof. The random variables $I_{(-\infty,x)}(X_1), \ldots, I_{(-\infty,x)}(X_n)$ are independent and identically distributed having Bernoulli distribution. (Here $I_{(-\infty,x)}(X_1) = 1$ if $X_1 \in (-\infty, x)$ and it is 0 otherwise.) So

$$P(I_{(-\infty,x)}(X_i) = 1) = P(X_i < x) = F(x) = p,$$

$$P(I_{(-\infty,x)}(X_i) = 0) = 1 - F(x)$$

so $\mathbb{E}I_{(-\infty,x)}(X_i) = p$. Therefore, by the strong law of large numbers,

$$F_n^*(x) = \frac{1}{n}(I_{(-\infty,x)}(X_1) + \cdots + I_{(-\infty,x)}(X_n)) \to p = F(x)$$

almost surely.

The fundamental theorem of mathematical statistics (Glivenko and Cantelli)

The convergence in the preceding theorem is uniform, that is

Theorem.

$$\sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)| \to 0, \quad \text{if } n \to \infty,$$

is satisfied almost surely (that is with probability 1).

Exercise for the Glivenko-Cantelli theorem

Visualize the approximation of F by F_n^*. Generate random samples of small, moderate and large sizes from the standard normal population. Draw the graphs of F and F_n^*.

Solution.

```
x=-6:0.01:6; y=normcdf(x,0,1);
for n=[50 500 5000]
figure plot(x,y,'r-'); hold on;
z=normrnd(0,1,[1,n]); z=sort(z);
for i=1:n-1
plot([z(i) z(i+1)],[i/n i/n],'b-','LineWidth',2); hold on
end
end
```

Exercise for the Glivenko-Cantelli theorem...

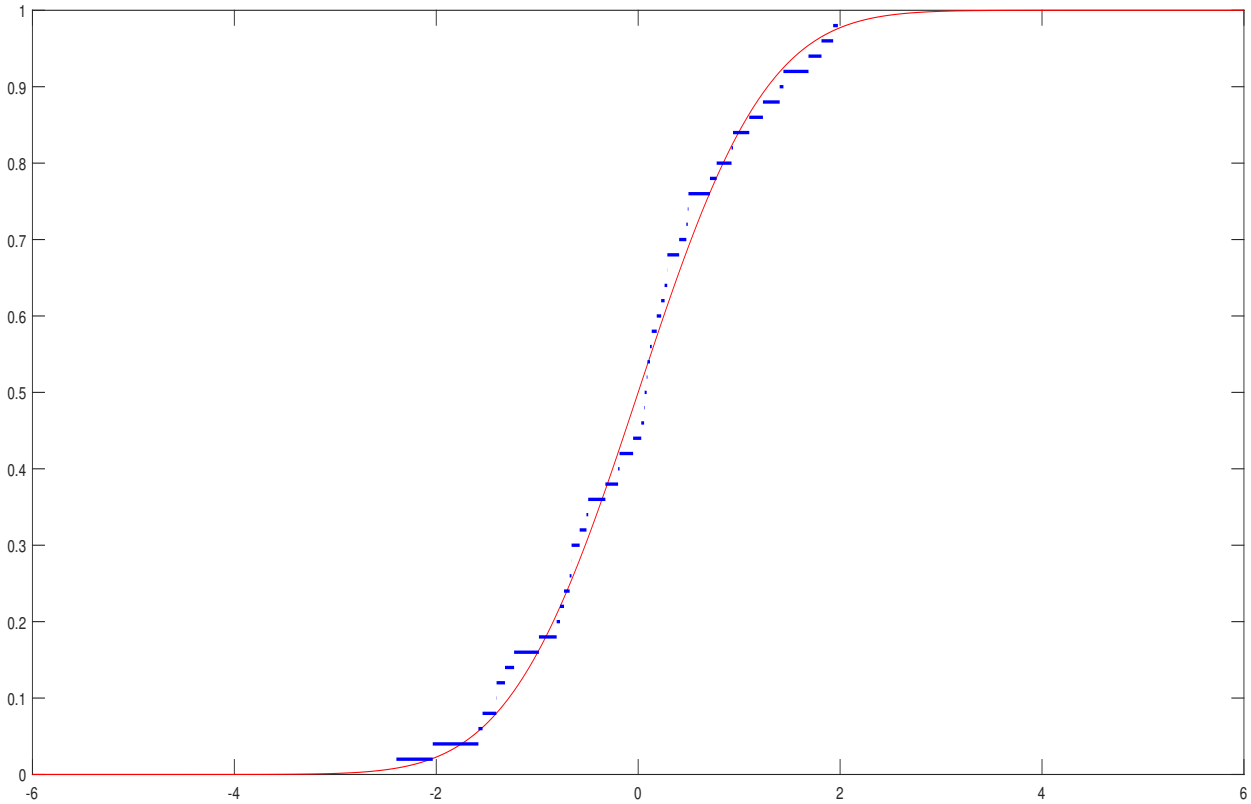

Figure: The standard normal CDF (red) and the empirical CDF (blue) for n=50

Exercise for the Glivenko-Cantelli theorem...

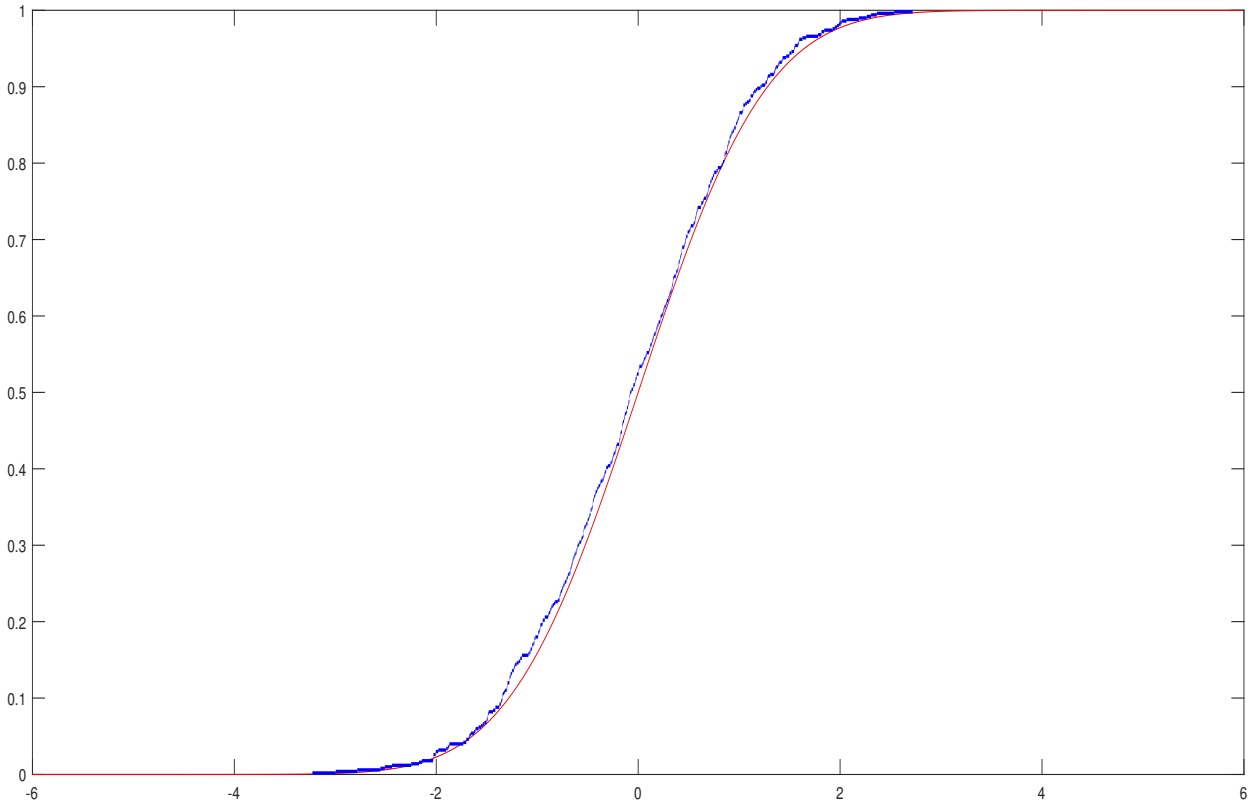

Figure: The standard normal CDF (red) and the empirical CDF (blue) for n=500

Exercise for the Glivenko-Cantelli theorem...

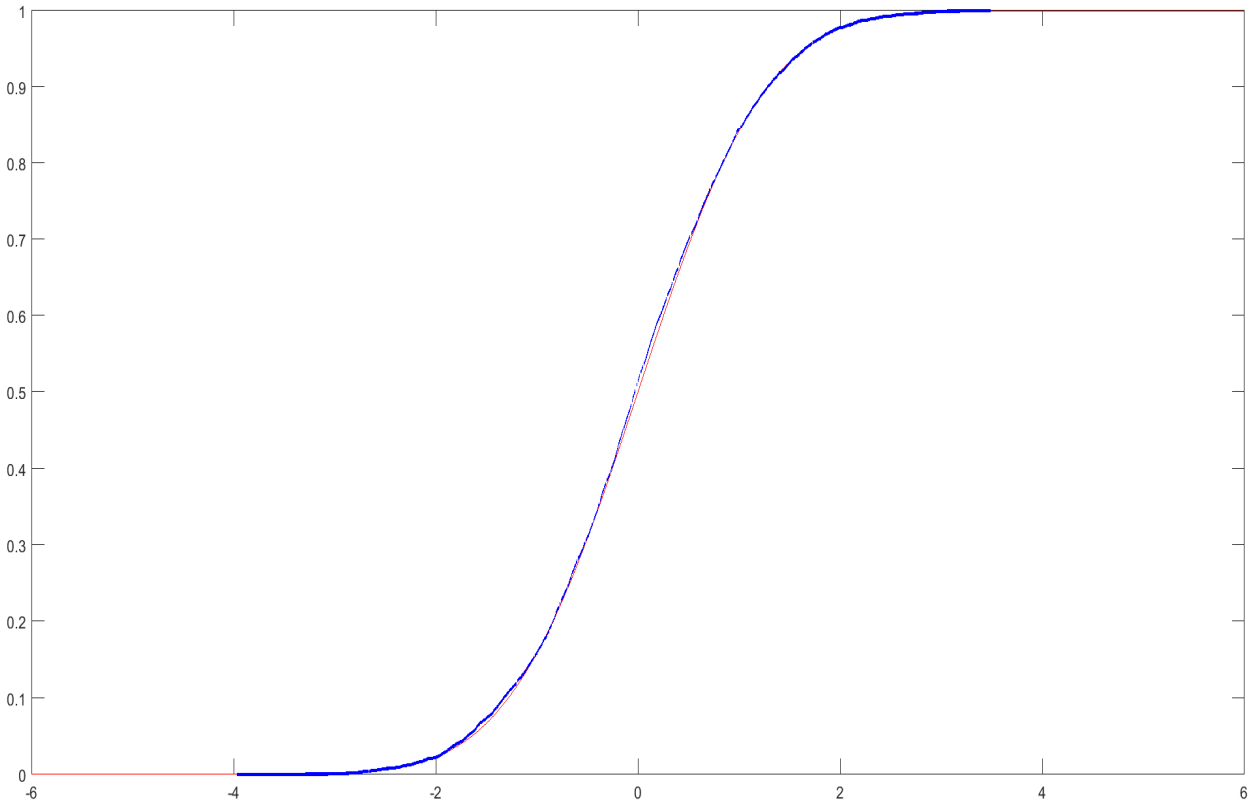

Figure: The standard normal CDF (red) and the empirical CDF (blue) for n=5000

Histograms

A histogram is column diagram approximating the distribution of numerical data. It was introduced by K. Pearson.
Let $X_1, X_2, \ldots, X_n$ be a sample.
Divide the range of the sample into subintervals with the points $y_0 < y_1 < \cdots < y_r$.
(The sample values should be in the interval (y_0, y_r).)
The subintervals $[y_{i-1}, y_i)$ are called bins.
Let ν_i be the number of sample values falling into the bin $[y_{i-1}, y_i)$, $i = 1, \ldots, r$.
Then a rectangle is erected over the bin with area proportional to the frequency ν_i, $i = 1, \ldots, r$.
So we obtain a histogram.

Frequency histograms and density histograms

If the cumulated area of the rectangles is n, then we obtain the frequency histogram. In this case the histogram is the real function f_n for which

$$f_n(x) = \begin{cases} \frac{\nu_i}{(y_i - y_{i-1})}, & \text{if } x \in [y_{i-1}, y_i), i = 1, \ldots, r, \\ 0, & \text{if } x \notin [y_0, y_r). \end{cases}$$

If the cumulated area of the rectangles is 1, then we obtain the density histogram. In this case the height of the ith rectangle is $\frac{\nu_i}{n(y_i - y_{i-1})}$.
The density histogram is an approximation of the PDF.

Histograms. Examples

Exercise. Using the random number generator, generate a sample of size 200 from the standard normal population. Construct the density histogram with 4 bins. Then use 7 bins, finally 18 bins.

A bad histogram

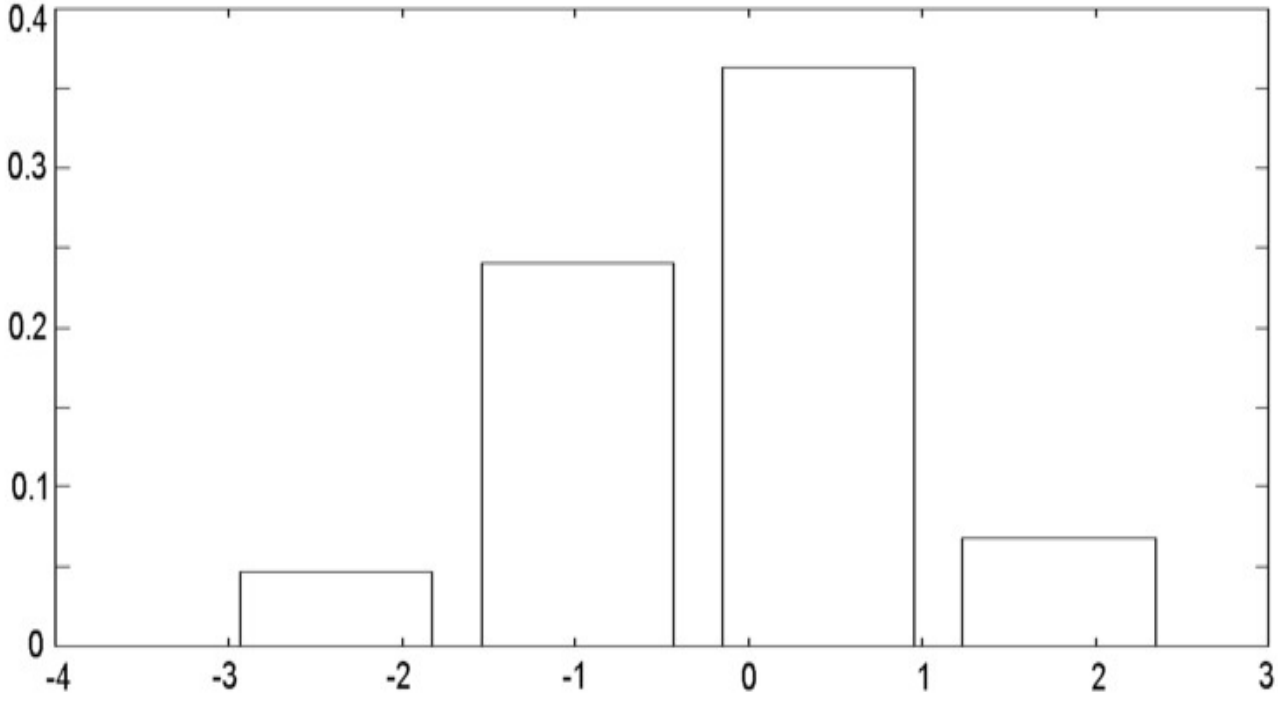

Figure: A histogram with 4 bins. Too few bins.

A good histogram

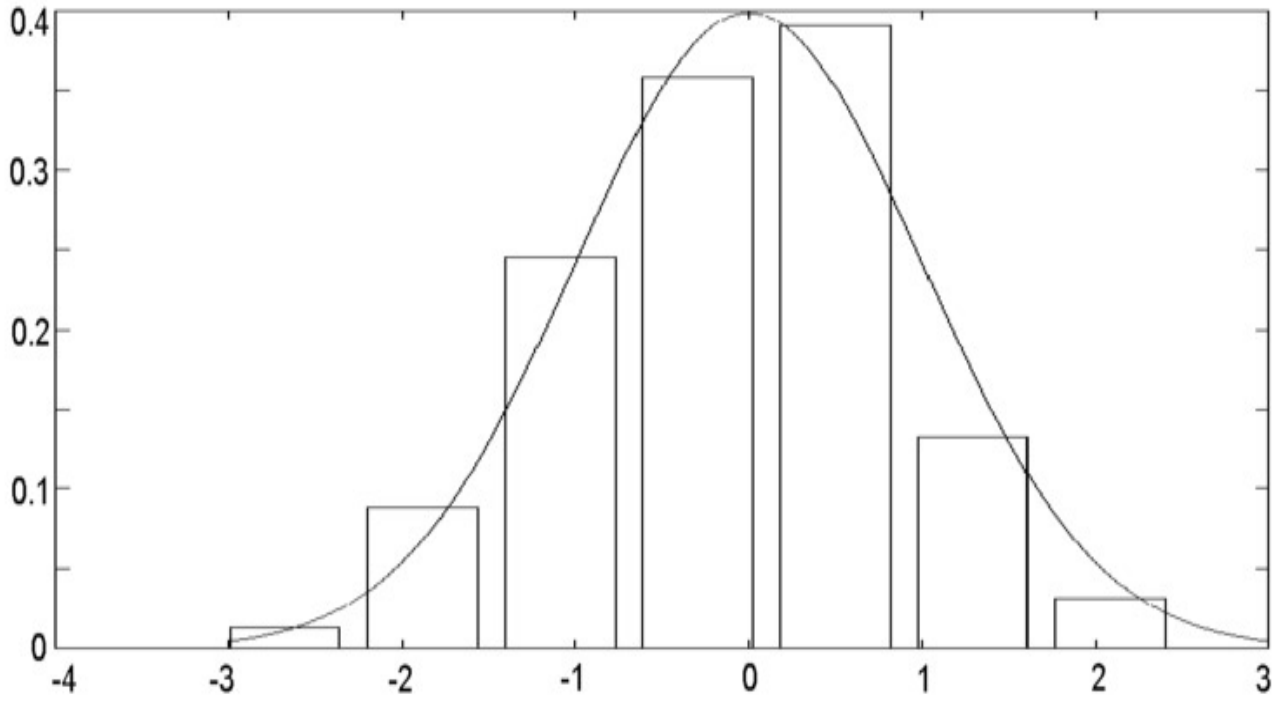

Figure: A histogram with 7 bins and the corresponding theoretical PDF

A bad histogram

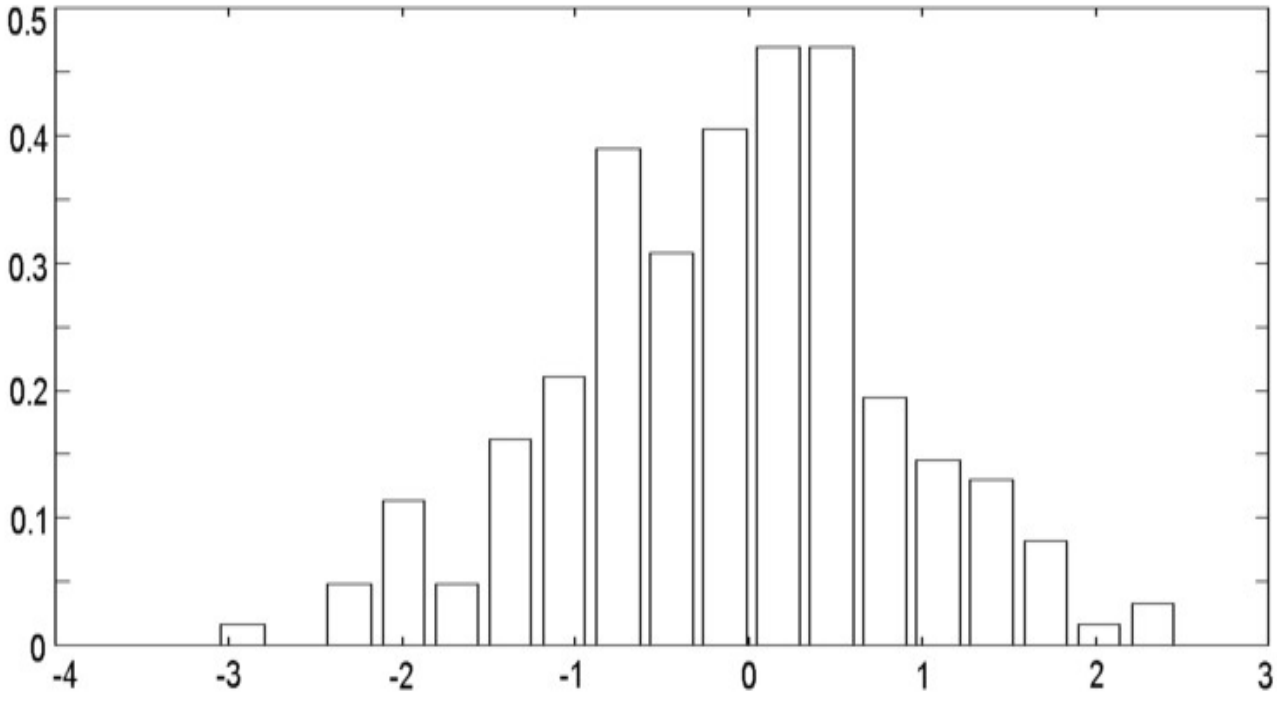

Figure: A histogram with 18 bins. Too many bins.

The empirical mean (sample average)

Let $X_1, X_2, \ldots, X_n$ be sample for X. Then

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

is called the empirical mean.
Assume that X has finite expectation $m = \mathbb{E}X$.
In statistics the value of m is unknown. We can estimate it by $\overline{X}$.
The expectation and the variance of $\overline{X}$ are

$$\mathbb{E}\overline{X} = \frac{1}{n}\sum_{1}^{n} \mathbb{E}X_i = m,$$

$$\mathrm{Var}(\overline{X}) = \frac{1}{n^2}\sum_{1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n},$$

where $\sigma^2 = \mathrm{Var}X$ is the theoretical variance.

The empirical mean (sample average)...

Theorem. Assume that $\mathbb{E}|X| < \infty$. Then the empirical mean is an unbiased and (strongly) consistent estimator of the theoretical expectation.

Proof.

Unbiased:

$$\mathbb{E}\overline{X} = m.$$

Consistent:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \to m$$

stochastically, by the WLLN.

Strong consistent:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \to m$$

almost surely, by the SLLN.

The empirical variance

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is called empirical variance. It is an estimator of the theoretical variance.
A better estimator of the variance is the corrected empirical variance, that is

$$s_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Steiner's formula will help us to calculate and to study s_n^2 and s_n^{*2}.

Steiner's formula

Theorem. For any real number a we have

$$ns_n^2 = (n-1)s_n^{*2} = \sum_{i=1}^{n}(X_i - a)^2 - n(\overline{X} - a)^2.$$

Proof.

$$\begin{aligned} ns_n^2 &= \sum_{i=1}^{n}\left[(X_i - a) - (\overline{X} - a)\right]^2 \\ &= \sum_{i=1}^{n}(X_i - a)^2 - 2\sum_{i=1}^{n}(X_i - a)(\overline{X} - a) + \sum_{i=1}^{n}(\overline{X} - a)^2 \\ &= \sum_{i=1}^{n}(X_i - a)^2 - n(\overline{X} - a)^2. \end{aligned}$$

The expectation of s_n^{*2}

Inserting $a = m$ into Steiner's formula, we obtain

$$\begin{aligned}
\mathbb{E}s_n^{*2} &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - m)^2 - \frac{n}{n-1}(\overline{X} - m)^2\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i - m)^2 - \frac{n}{n-1}\mathbb{E}(\overline{X} - m)^2 \\
&= \frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\frac{\sigma^2}{n} = \sigma^2.
\end{aligned}$$

The corrected empirical variance

Theorem. Assume that $\mathbb{E}X^2 < \infty$. Then the corrected empirical variance is an unbiased and (strongly) consistent estimator of the theoretical variance.

Proof. Unbiased:

$$\mathbb{E}s_n^{*2} = \sigma^2.$$

Consistent: Inserting $a = 0$ into Steiner's formula, we obtain

$$s_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i)^2 - \frac{n}{n-1}(\overline{X})^2 \to \mathbb{E}X^2 - m^2 = \sigma^2$$

stochastically, by the WLLN.

Strong consistent:

$$s_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i)^2 - \frac{n}{n-1}(\overline{X})^2 \to \mathbb{E}X^2 - m^2 = \sigma^2$$

almost surely, by the SLLN.

Sample quantiles

The zth percentile or $z\%$ quantile is the the smallest sample element x_k such that $z\%$ of the whole sample is smaller than or equal to x_k.

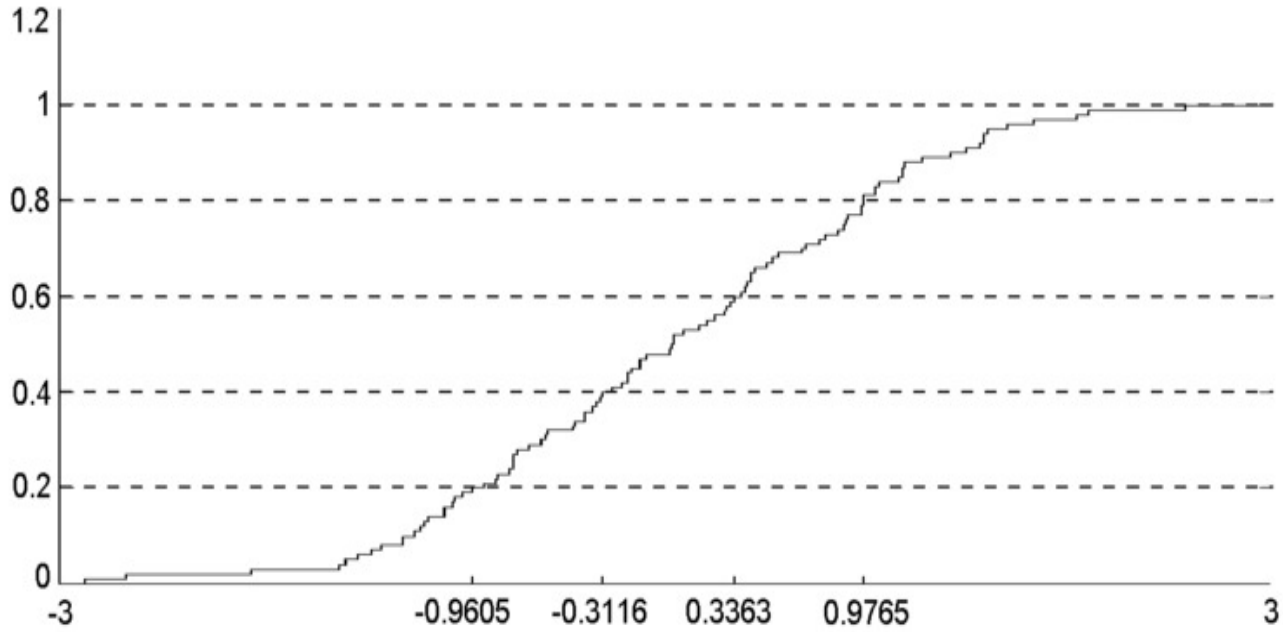

Figure: The 20th, 40th, 60th and 80th sample percentiles of a sample from normal population

Boxplot

A useful method of descriptive statistics.

A boxplot displays the dataset using the minimum, the maximum, the sample median, and the first and third quartiles.

Order the sample.

Q_2 = the median (50th percentile) is the middle value of the dataset.

Q_1 = first (lower) quartile (25th percentile) is the middle value of the lower half of the dataset.

Q_3 = third (upper) quartile (75th percentile) is the middle value of the upper half of the dataset.

The median and the two quartiles are visualized by a box.

Interquartile range is the distance between the upper and lower quartiles, it is denoted by h.

A sample value is an outlier if it is either between $Q_3 + 1.5h$ and $Q_3 + 3h$ or between $Q_1 - 1.5h$ and $Q_1 - 3h$. Outliers are usually denoted by $+$.

Boxplot...

A data value is extreme, if it is outside the interval $[Q_1 - 3h, Q_3 + 3h]$. Extreme values are usually denoted by $*$.
The minimum is the lowest data point excluding outliers and extremes.
The maximum is the largest data point excluding outliers and extremes.
The minimum and the maximum values are visualised by whiskers.

Exercise. Generate samples from uniform, normal, Cauchy, and Poisson distributions. Construct the boxplots.
Solution. n=100;
x=normrnd(0,1,n,1); y=3*rand(n,1)-1.5; z=0.1*trnd(1,n,1);
v=poissrnd(1,n,1);
boxplot([y x z v]);

Boxplots

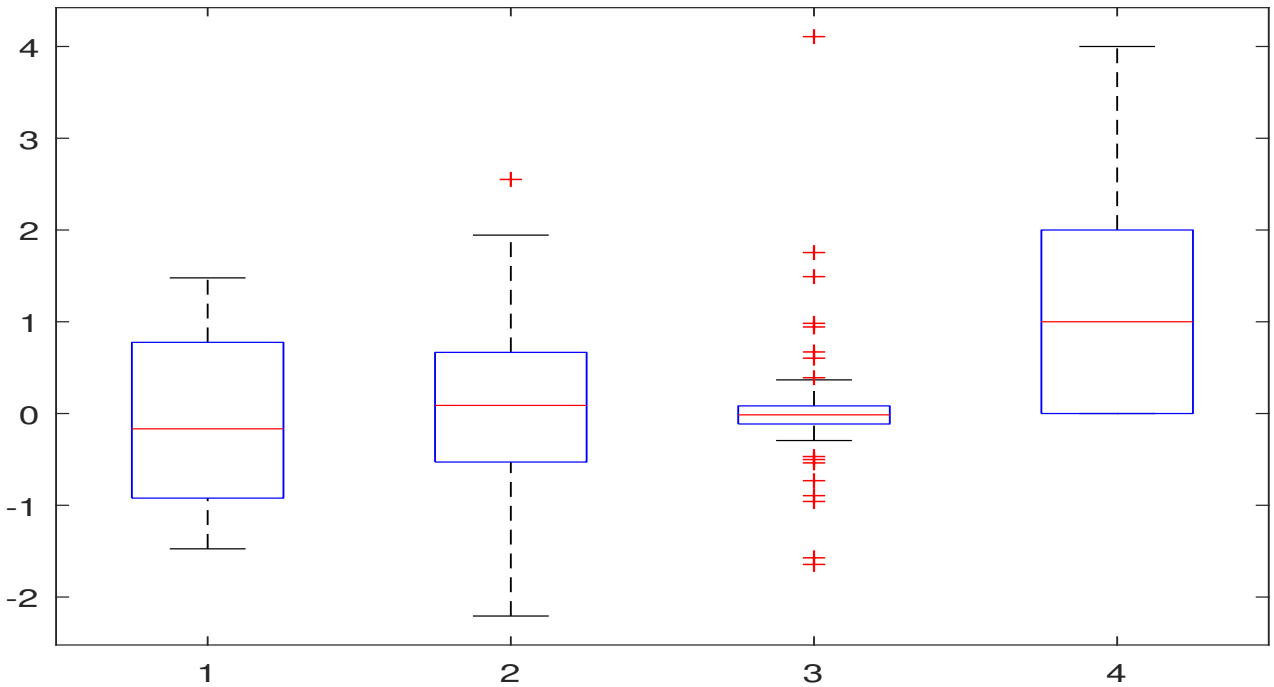

Figure: Boxplots for samples from uniform, normal, Cauchy, and Poisson distributions.

Boxplots...

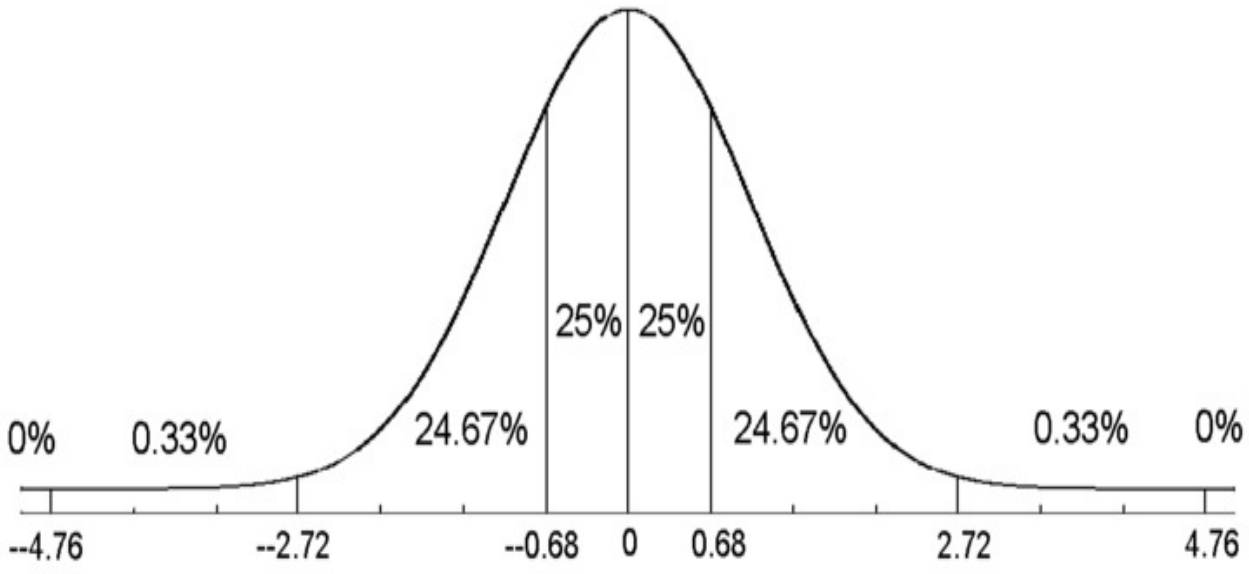

Figure: The regions of the outliers and extreme values for standard normal distribution.