Lecture 8

Sample Mean and Variance, Histogram, Empirical Distribution Function

Sample Mean and Variance

Consider a random sample $\boldsymbol{X} = (x_1, x_2, \ldots x_n)$, where n is the sample size (number of elements in $\boldsymbol{X}$).

Then, the *sample mean* $\bar{x}$ is given by

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

Sample Mean and Variance

The sample mean is an unbiased estimator of the expected value of a random variable the sample is generated from:

$$\bar{x} \equiv \hat{E}[X].$$

The sample mean is the sample statistic, and is itself a random variable.

Sample Mean and Variance

In many practical situations, the true variance of a population is not known *a priori* and must be computed somehow.

When dealing with extremely large populations, it is not possible to count every object in the population, so the computation must be performed on a sample of the population.

Sample Mean and Variance

Sample variance can also be applied to the estimation of the variance of a continuous distribution from a sample of that distribution.

Sample Mean and Variance

Consider a random sample $\boldsymbol{X} = (x_1, x_2, \ldots x_n)$ of size n. Then, we can define the sample variance as

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

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

Sample Mean and Variance

However, $\widehat{\sigma_X}^2$ gives an estimate of the population variance that is biased by a factor of $\frac{n-1}{n}$.

For this reason, $\widehat{\sigma_X}^2$ is referred to as the ***biased sample variance***. Correcting for this bias yields the *unbiased sample variance*:

$$s^2 = \frac{n}{n-1}\widehat{\sigma_X}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

Sample Mean and Variance

While s^2 is an unbiased estimator for the variance, s is still a biased estimator for the standard deviation, though markedly less biased than the uncorrected sample standard deviation $\widehat{\sigma_X}$.

This estimator is commonly used and generally known simply as the "sample standard deviation". The bias may still be large for small samples (n less than 10).

As sample size increases, the amount of bias decreases.

Sample Mean and Variance

When using Mathcad, we can obtain the sample mean by use of $mean(X)$ function.

The biased and unbiased variances are obtained by $var(X)$ and $Var(X)$, respectively (mind the letter case!).

The uncorrected and corrected standard deviation are obtained by $stdev(X)$ and $Stdev(X)$.

Sample Mean and Variance

We can also obtain sample raw moments of the *k*th order:

$$\hat{\mu}'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \,,$$

and sample central moments of the *k*th order:

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k \,,$$

Sample Mean and Variance

… as well as sample standardized moments of the *k*th order:

$$\hat{\nu}_k = \frac{\hat{\mu}_k}{s^k},$$

where s is corrected sample standard deviation.

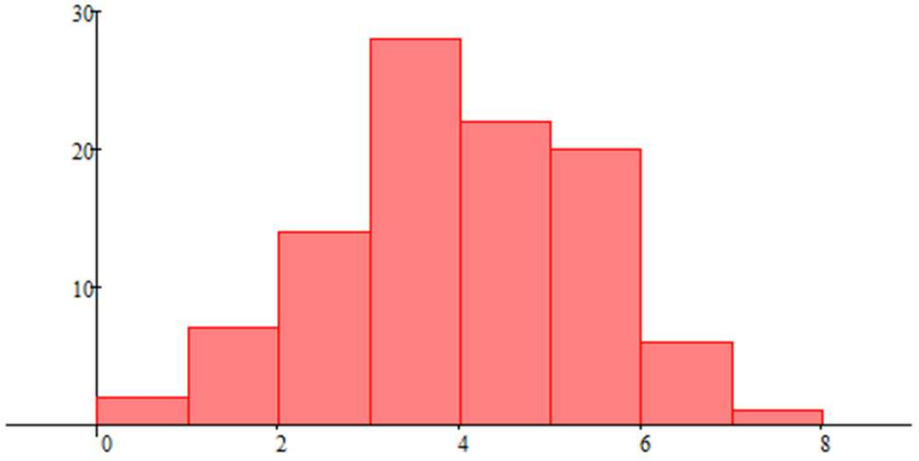
Histogram

A histogram is an approximate representation of the distribution of numerical or categorical data.

To construct a histogram, the first step is to "bin" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval.

Histogram

The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent, and are often (but not required to be) of equal size.



Histogram

If the bins are of equal size, a rectangle is erected over the bin with height proportional to the frequency—the number of cases in each bin.

A histogram may also be normalized to display "relative" frequencies. It then shows the proportion of cases that fall into each of several categories, with the sum of the heights equaling 1.

Histogram

Histograms give a rough sense of the density of the underlying distribution of the data, and often are used for density estimation.

The total area of a histogram used for probability density is always normalized to 1.

Histogram

In a more general mathematical sense, a histogram is a function m_i that counts the number of observations that fall into each of the disjoint categories (bins), whereas the graph of a histogram is merely one way to represent a histogram.

Histogram

Thus, if we let n be the total number of observations (sample size) and k be the total number of bins, the histogram m_i meets the following conditions:

$$n = \sum_{i=1}^{k} m_i \, .$$

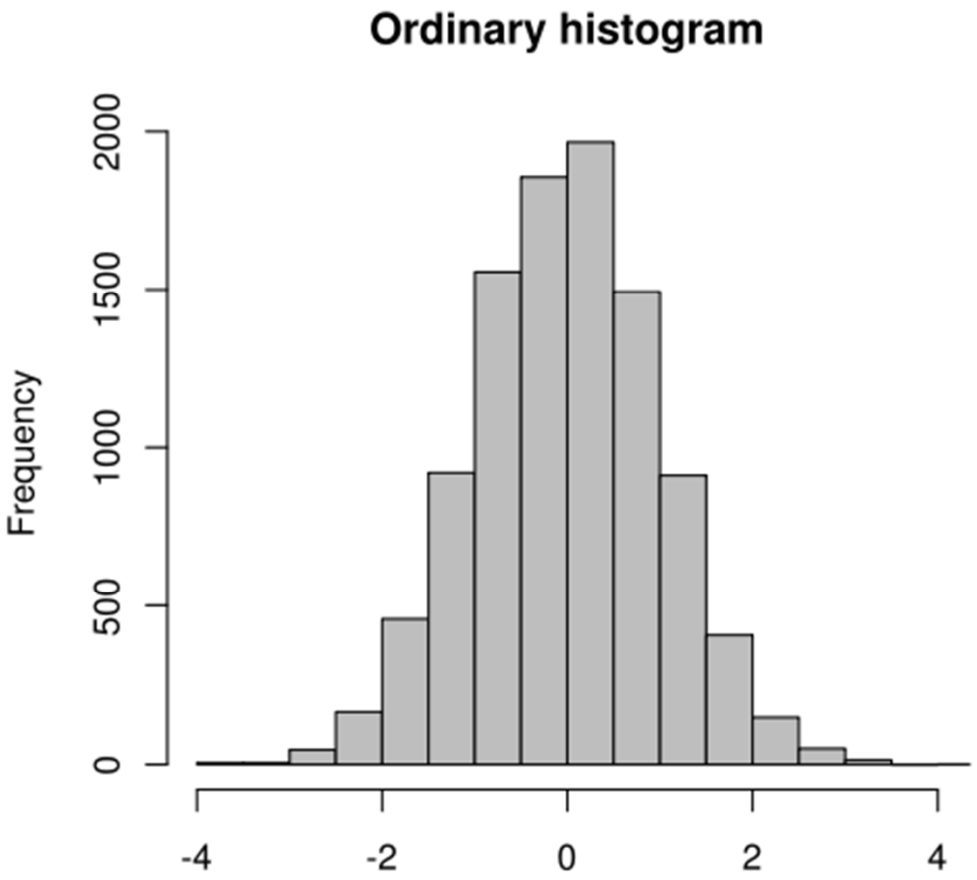
Histogram

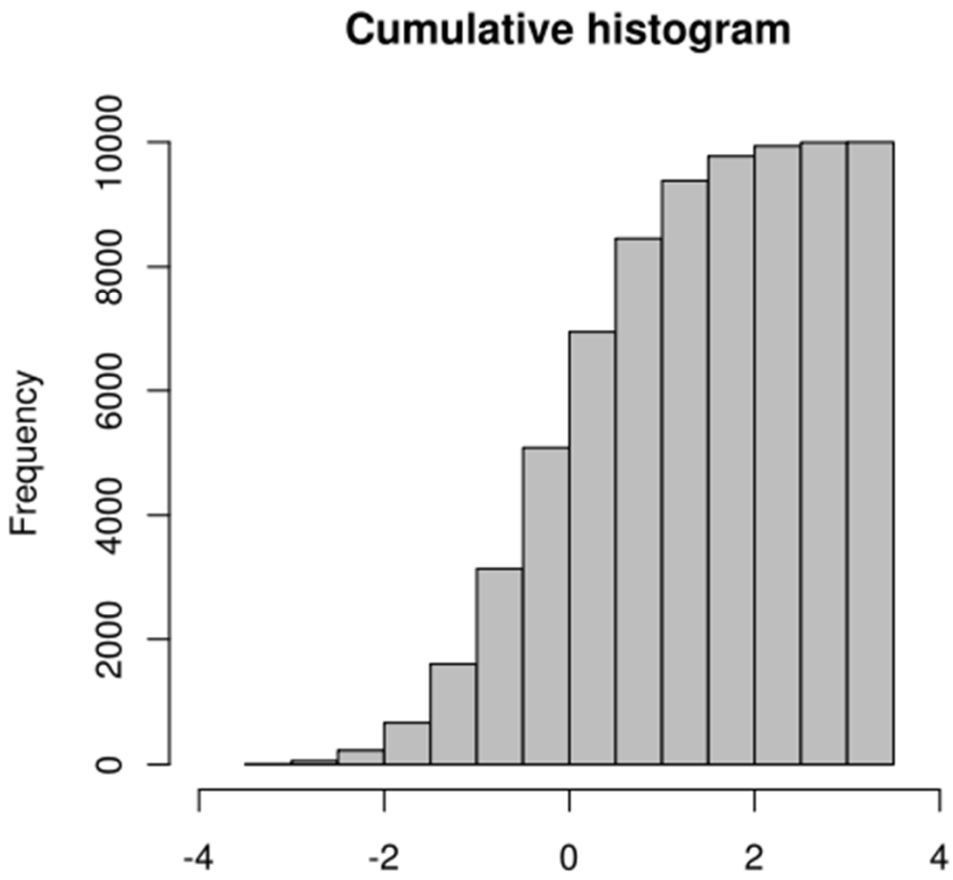
A cumulative histogram is a mapping that counts the cumulative number of observations in all of the bins up to the specified bin.

That is, the cumulative histogram M_i of a histogram m_j is defined as:

$$M_i = \sum_{j=1}^{i} m_j \, .$$

Histogram


Ordinary histogram
Frequency
0
500
1000
1500
2000
-4
-2
0
2
4


Cumulative histogram
Frequency
0
2000
4000
6000
8000
10000
-4
-2
0
2
4

Histogram

There is no "best" number of bins, and different bin sizes can reveal different features of the data.

Using wider bins where the density of the underlying data points is low reduces noise due to sampling randomness; using narrower bins where the density is high gives greater precision to the density estimation. Thus varying the bin-width within a histogram can be beneficial.

Nonetheless, equal-width bins are widely used.

Histogram

Some theoreticians have attempted to determine an optimal number of bins, but these methods generally make strong assumptions about the shape of the distribution.

Depending on the actual data distribution and the goals of the analysis, different bin widths may be appropriate, so experimentation is usually needed to determine an appropriate width.

There are, however, various useful guidelines and rules of thumb.

Histogram

The number of bins k can be assigned directly or can be calculated from a suggested bin width h as:

$$k = \left\lceil \frac{max(\boldsymbol{X}) - min(\boldsymbol{X})}{h} \right\rceil,$$

where braces indicate the ceiling function.

Square-root choice

$$k = \lceil \sqrt{n} \rceil,$$

where n is the sample size.

Sturges' formula

Sturges' formula implicitly assumes an approximately normal distribution.

$$k = 1 + \lceil \log_2 n \rceil$$

The Sturges' formula can perform poorly if $n < 30$, because the number of bins will be small—less than seven—and unlikely to show trends in the data well. It may also perform poorly if the data are not normally distributed.

Rice Rule

$$k = \lceil 2\sqrt[3]{n} \rceil$$

The Rice Rule is presented as a simple alternative to Sturges' formula.

Doane's formula

Doane's formula is a modification of Sturges' formula which attempts to improve its performance with non-normal data.

$$k = 1 + \left\lfloor \log_2 n + \log_2 \left(1 + \frac{|\hat{\nu}_3|}{\sigma_\nu} \right) \right\rfloor$$

where $\hat{\nu}_3$ is sample's third standardized moment and

$$\sigma_\nu = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}.$$

Scott's normal reference rule

$$h = \frac{3.49s}{\sqrt[3]{n}}$$

Scott's normal reference rule is optimal for random samples of normally distributed data.

After determining h the number of bins is obtained by

$$k = \left\lceil \frac{max(\boldsymbol{X}) - min(\boldsymbol{X})}{h} \right\rceil$$

Freedman-Diaconis rule

$$h = \frac{2 \cdot IQR(\boldsymbol{X})}{\sqrt[3]{n}},$$

where $IQR(\boldsymbol{X})$ is an interquartile range given by the difference between third and first sample quartiles:

$$IQR(\boldsymbol{X}) = Q_3 - Q_1.$$

When using Mathcad, we can obtain the interquartile range by the following expression:

$$IQR := precentile(X, 0.75) - percentile(X, 0.25)$$

The ceiling function can be implemented as $ceil(X)$.

Histogram

In order to get the ordinary histogram for a sample $\boldsymbol{X}$ in Mathcad we use the function $histogram(k, X)$, specifying the number of bins k and the dataset X.

The $histogram$ function results in a $(k \times 2)$ matrix, with first column containing the coordinates for the bins' midpoints and second – the number of elements of the dataset falling into each bin.

Empirical Distribution Function

Another way to graphically represent a dataset is to plot the data in a cumulative manner. This can be done using the empirical cumulative distribution function (EDF, or ECDF) of the data.

It is denoted by F_n and is defined at a point x as the proportion of elements in the dataset that are less than or equal to x:

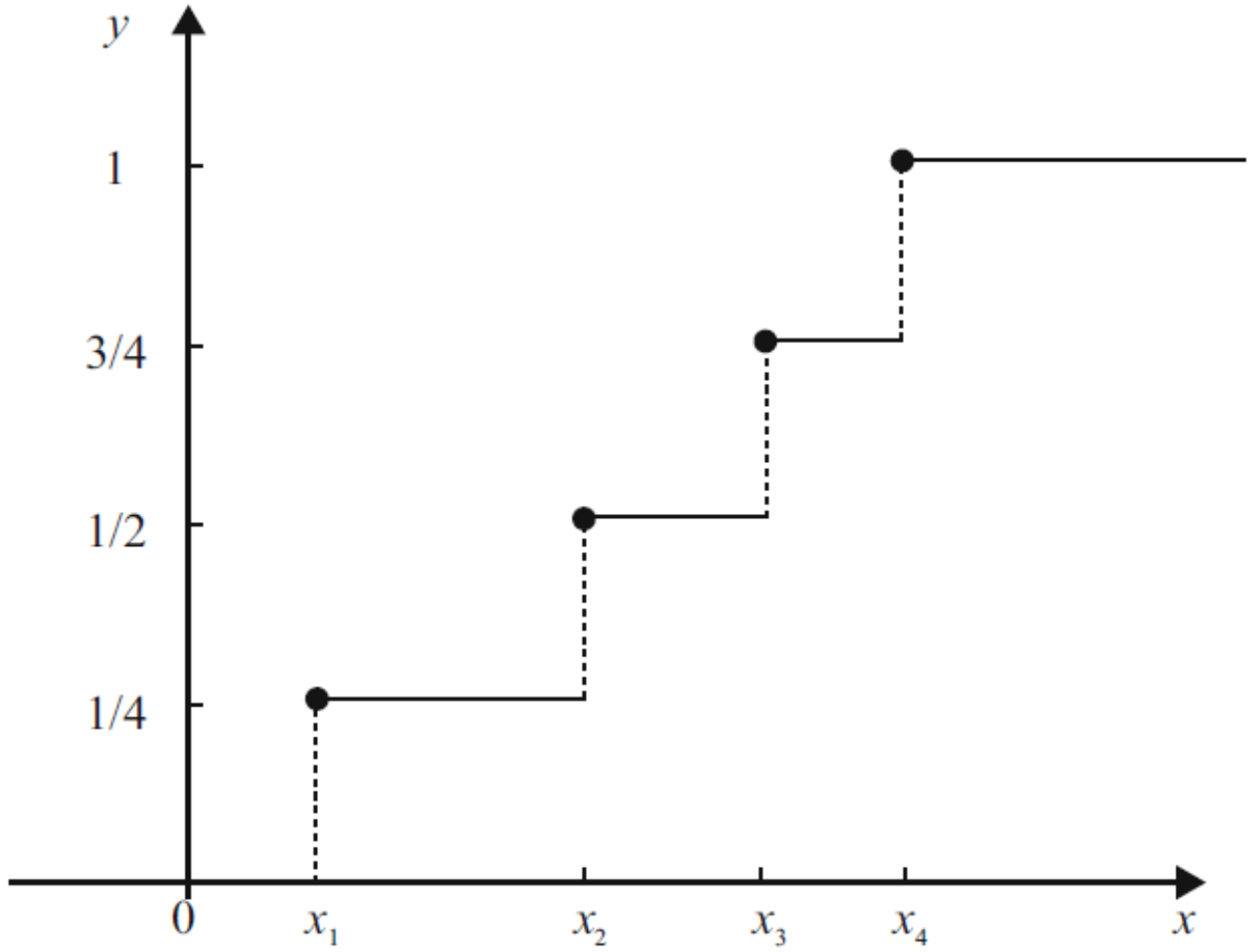
$$F_n(x) = \frac{number\ of\ elements\ in\ the\ dataset\ \leq x}{n}$$

Empirical Distribution Function

In general, the graph of $F_n(x)$ has the form of a staircase, with $F_n(x) = 0$ for all x smaller than the minimum of the dataset and $F_n(x) = 1$ for all x greater than the maximum of the dataset.

Between the minimum and maximum, $F_n(x)$ has a jump of size $1/n$ at each element of the dataset and is constant between successive elements.

Empirical Distribution Function



Empirical Distribution Function

The EDF $F_n(x)$ is an estimator for the *cdf* of the random variable, $F(x)$. To indicate this fact, sometimes it is denoted by $\hat{F}(x)$.

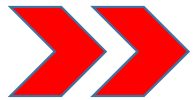
Preliminaries for the Laboratory Work 2

1. Generate two datasets*:

- for normally distributed r.v.;
- for exponentially distributed r.v.

2. Save the datasets into separate Excell files.

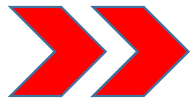
* The distributions' parameters and samples' sizes are listed below, according to your variant.



Preliminaries for the Laboratory Work 2

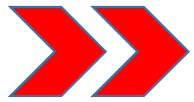
3. For each dataset:

a) determine minimal and maximal elements;
b) determine sample mean, sample variance (biased and unbiased), sample standard deviation (biased and unbiased);
c) calculate the number of bins recommended by formulae mentioned above;
d) plot histograms for the distinct number of bins;
e) select a histogram, normalize and plot it;



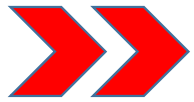
Preliminaries for the Laboratory Work 2

f) plot the *pdf* of the corresponding r.v. over the normalized histogram;
g) write a user-defined Mathcad function that takes in the result of $histogram$ function as an input and returns the cumulative histogram;
h) plot the normalized cumulative histogram;
i) plot the *cdf* of the corresponding r.v. over the normalized cumulative histogram;



Preliminaries for the Laboratory Work 2

j) write a user-defined Mathcad function that takes in the dataset as an input and returns the EDF (*extra credit*);
k) plot the EDF;
l) plot the *cdf* of the corresponding r.v. over the EDF.



Preliminaries for the Laboratory Work 2

Variant	Normal Distribution			Exponential Distribution	
	n	μ	σ	n	λ
A	80	3	2.5	140	0.125
B	200	10	5	70	0.04

Laboratory Work 2 Example

X is the sample ($n = 100$) generated from $N(1,1)$ distribution, ($\mu = 1, \sigma = 1$).

$$n := 100 \qquad \mu := 1 \qquad \sigma := 1$$

$$X := \mathrm{norm}(n, \mu, \sigma)$$

Laboratory Work 2 Example

$$\min(X) = -1.844 \qquad \max(X) = 3.517$$

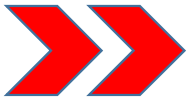
$$x_{mean} := \text{mean}(X) = 0.946$$ sample mean value

Biased variance and standard deviation

$$\text{var}(X) = 0.964 \qquad \text{stdev}(X) = 0.982$$

Unbiased variance and standard deviation

$$\text{Var}(X) = 0.974 \qquad \text{Stdev}(X) = 0.987$$



Laboratory Work 2 Example

Square root

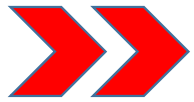
$$k_{sqr} := \text{ceil}\left(\sqrt{n}\right) = 10$$

Sturges' formula

$$k_{sturges} := 1 + \text{ceil}(\log(n, 2)) = 8$$

Rice rule

$$k_{rice} := \text{ceil}\left(2 \cdot \sqrt[3]{n}\right) = 10$$



Laboratory Work 2 Example

Doane

$$\nu_3 := \frac{\frac{1}{n}\cdot\sum_{i=0}^{99}\left(X_i - x_{mean}\right)^3}{Stdev(X)} = -0.105 \qquad \sigma_\nu := \sqrt{\frac{6\cdot(n-2)}{(n+1)\cdot(n+3)}} = 0.238$$

$$k_{doane} := 1 + ceil\left(\log(n,2) + \log\left(1 + \frac{|\nu_3|}{\sigma_\nu},2\right)\right) = 9$$

Laboratory Work 2 Example

Scott's

$$h_{scott} := \frac{3.49 \cdot \mathrm{Stdev}(X)}{\sqrt[3]{n}} = 0.742$$

$$k_{scott} := \mathrm{ceil}\left(\frac{\max(X) - \min(X)}{h_{scott}}\right) = 8$$

Laboratory Work 2 Example

Freedman - Diaconis

$$IQR := \text{percentile}(X, 0.75) - \text{percentile}(X, 0.25) = 1.31$$

$$h_{fd} := \frac{2 \cdot IQR}{\sqrt[3]{n}} = 0.565$$

$$k_{fd} := \text{ceil}\left(\frac{\max(X) - \min(X)}{h_{fd}}\right) = 10$$

Laboratory Work 2 Example

According to the formulae, we have obtained three distinct number of bins: 8, 9, and 10. Now, let's calculate the width of bins for these values:

$$h_8 := \frac{\max(X) - \min(X)}{8} = 0.67 \qquad h_9 := \frac{\max(X) - \min(X)}{9} = 0.596$$

$$h_{10} := \frac{\max(X) - \min(X)}{10} = 0.536$$

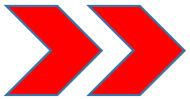
Laboratory Work 2 Example

$$H8 := \text{histogram}(8, X) = \begin{pmatrix} -1.509 & 2 \\ -0.838 & 6 \\ -0.168 & 11 \\ 0.502 & 27 \\ 1.172 & 24 \\ 1.842 & 19 \\ 2.512 & 9 \\ 3.182 & 2 \end{pmatrix}$$

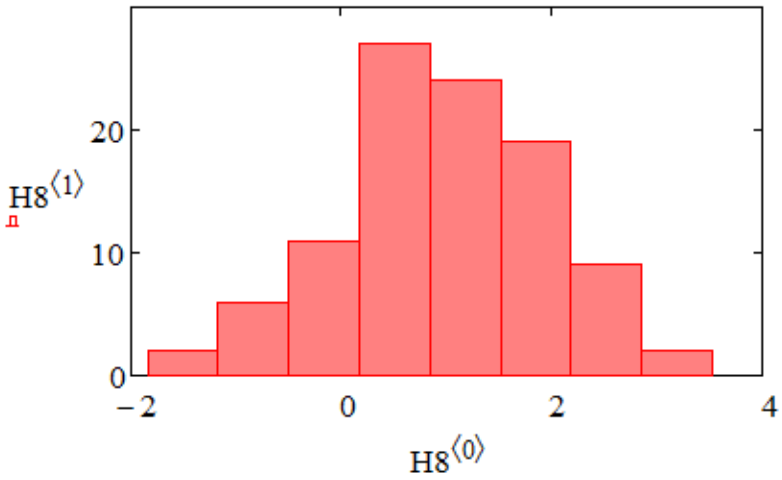
$$H9 := \text{histogram}(9, X) = \begin{pmatrix} -1.546 & 2 \\ -0.95 & 6 \\ -0.355 & 5 \\ 0.241 & 22 \\ 0.837 & 24 \\ 1.432 & 19 \\ 2.028 & 17 \\ 2.623 & 3 \\ 3.219 & 2 \end{pmatrix}$$

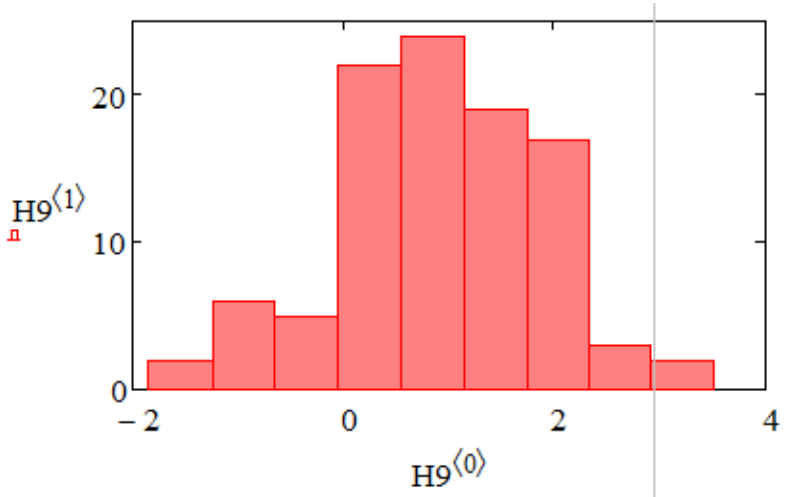
$H10 := \text{histogram}(10, X) =$

	0	1
0	-1.576	2
1	-1.039	3
2	-0.503	5
3	0.033	15
4	0.569	21
5	1.105	18
6	1.641	18
7	2.177	13
8	2.713	3
9	3.249	2

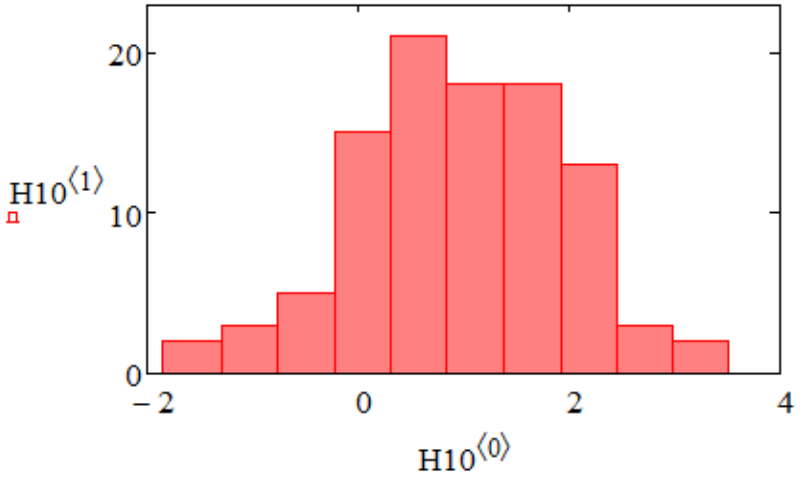


Laboratory Work 2 Example



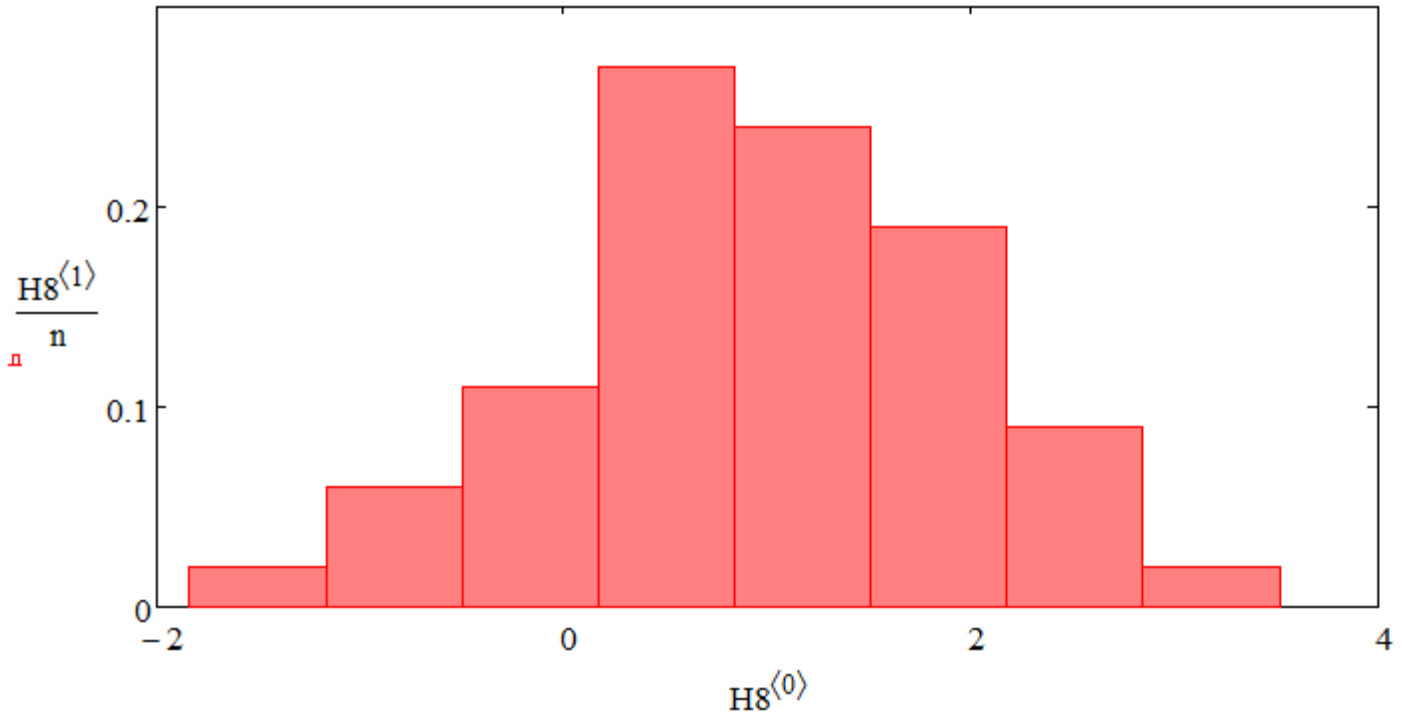


We've plotted 3 histograms.
Let's select one of them, say H8, and continue with it.



Laboratory Work 2 Example

By dividing the values in the second column of the matrix H8 we obtain the normalized histogram:



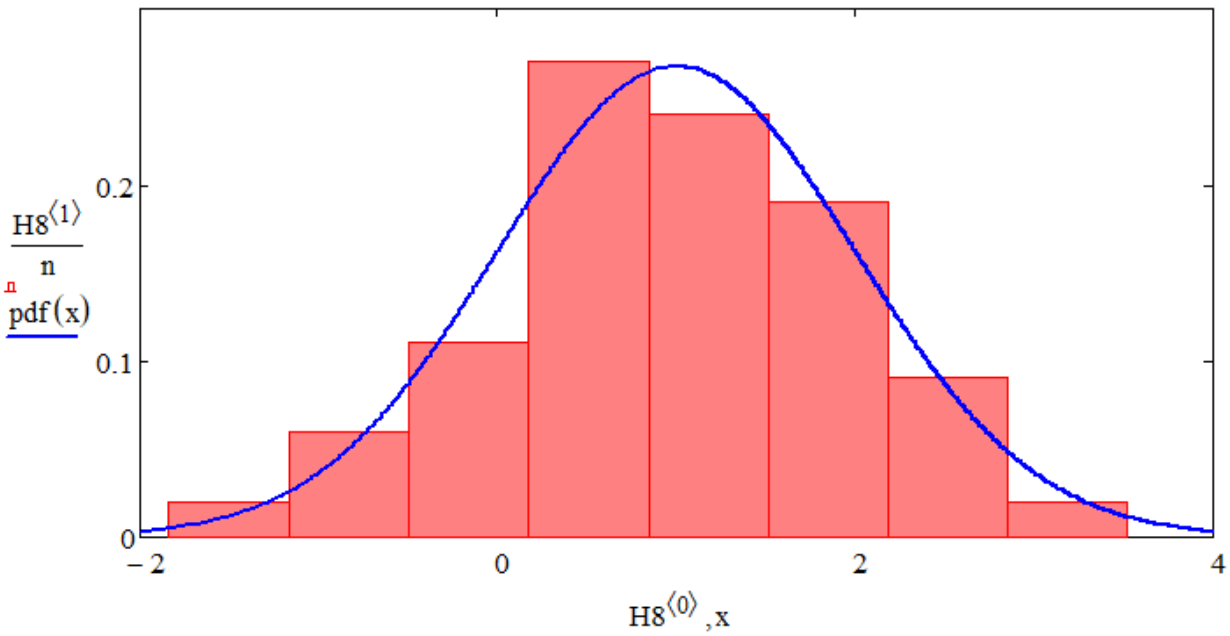
$$\frac{H8^{\langle 1 \rangle}}{n} = \begin{pmatrix} 0.02 \\ 0.06 \\ 0.11 \\ 0.27 \\ 0.24 \\ 0.19 \\ 0.09 \\ 0.02 \end{pmatrix}$$

Laboratory Work 2 Example

To observe how well the histogram approximates the *pdf* of random variable, we need to factor in the width of the bin, when plotting the *pdf*:

$x := -2, -1.999 .. 4$

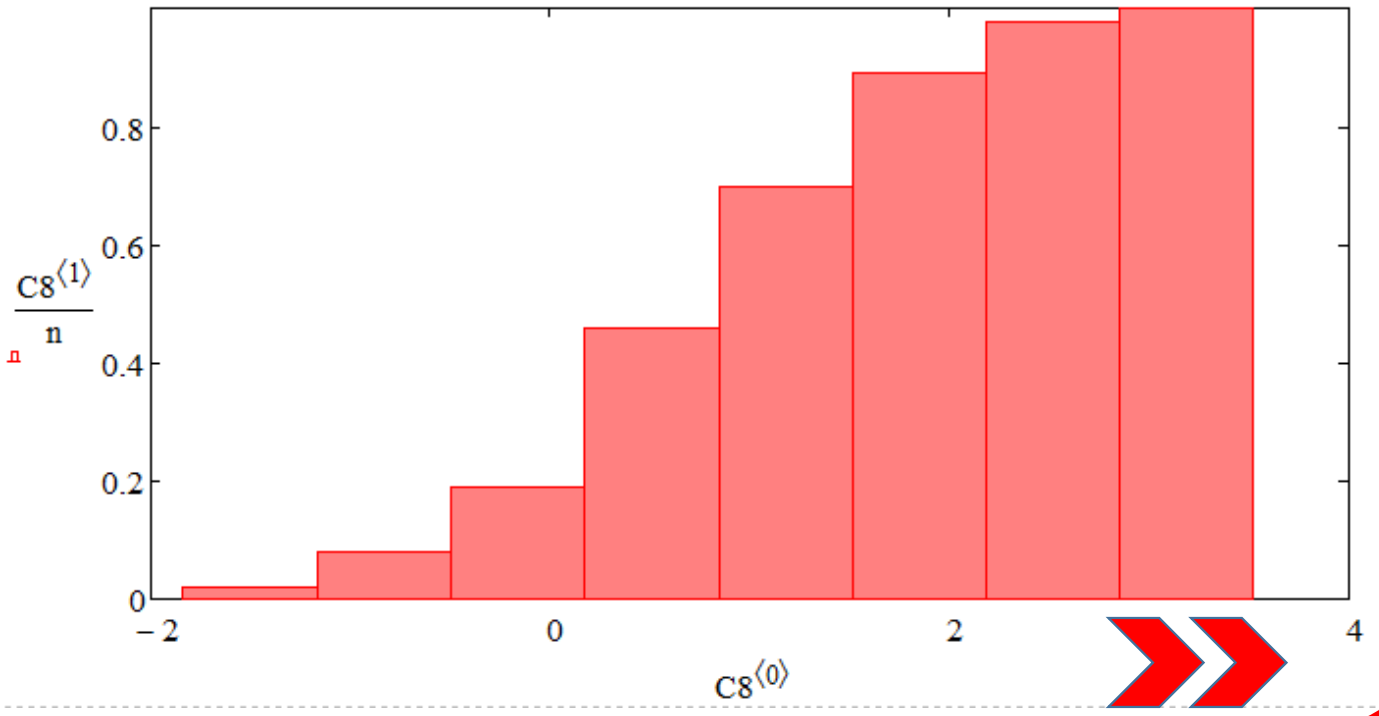
$pdf(x) := h_8 \cdot dnorm(x, \mu, \sigma)$



Laboratory Work 2 Example

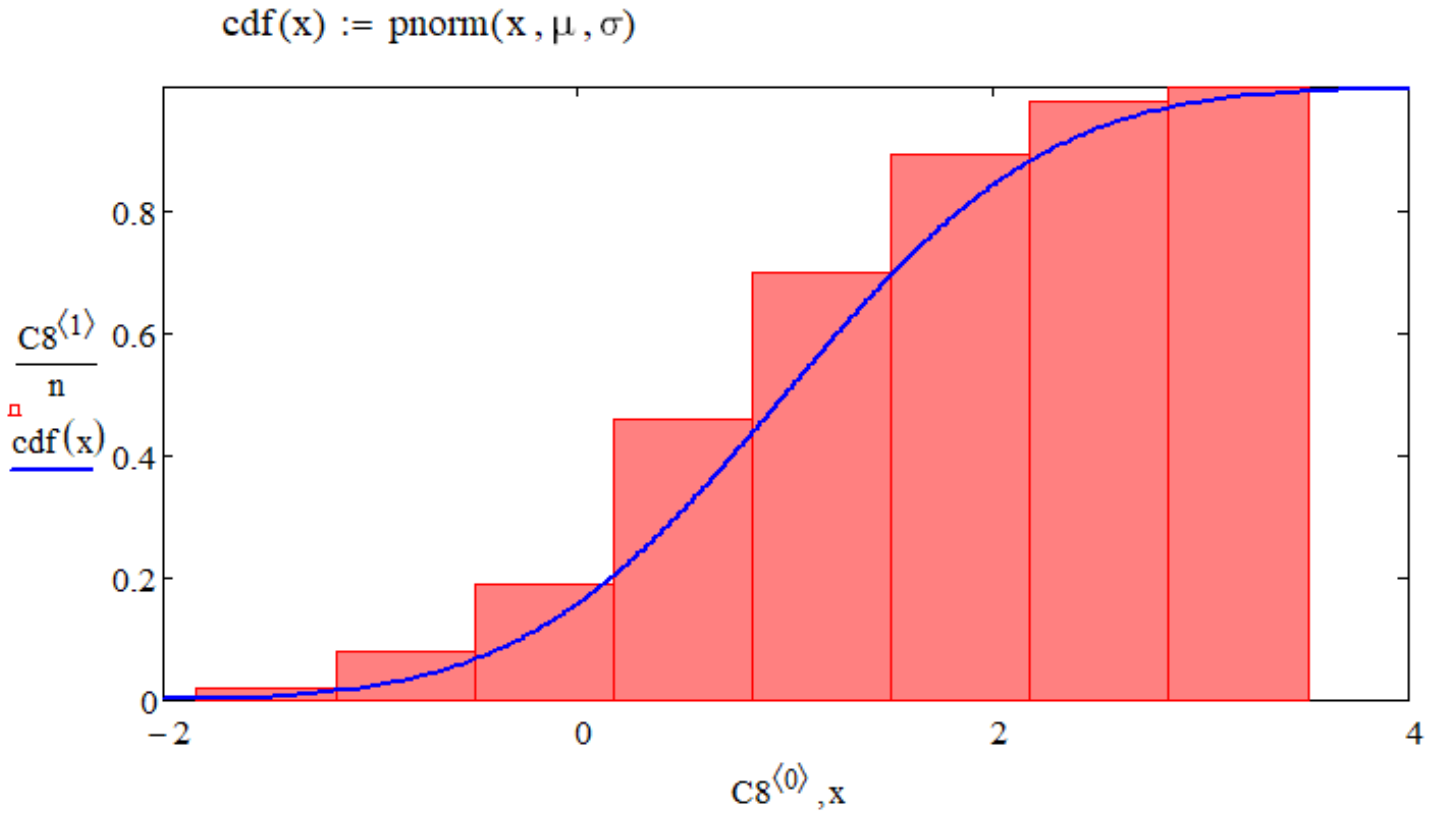
You should write the program for the cumulative histogram yourself, but the result should look something like this:

$$C8 := \text{histocum}(H8) = \begin{pmatrix} -1.509 & 2 \\ -0.838 & 8 \\ -0.168 & 19 \\ 0.502 & 46 \\ 1.172 & 70 \\ 1.842 & 89 \\ 2.512 & 98 \\ 3.182 & 100 \end{pmatrix}$$



Laboratory Work 2 Example

Comparing the normalized cumulative histogram and the cdf, we get



Textbook Assignment

F.M. Dekking et al. *A Modern Introduction to…*

- Chapters 15 & 16.