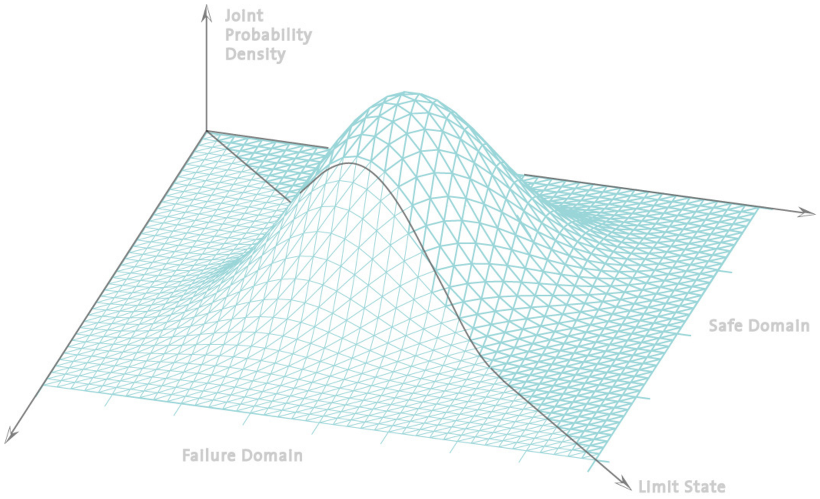


Exercise Tutorial 9

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber
Swiss Federal Institute of Technology Zurich
ETHZ

Exercise 7.3
(from the last tutorial: please correct the mistake, also in the solution given in the script)

c. Find an expression for the cumulative distribution function of the river's maximum discharge over the 20 year lifetime of an anticipated flood-control project. Assume that the individual annual maxima are independent random variables.

For independent random variables, the cumulative distribution function of the largest extreme in a period of nT is:

$$F^{\max}_{X,20T}(x) = \left\{ \; F^{\max}_{X,T}(x) \; \right\}^{20} = \left(e^{-e^{-\alpha(x-u)}} \right)^{20}$$

$$\xcancel{F^{\max}_{X,20T}(x) = e^{-e^{-20\alpha(x-u)}}} \quad \longrightarrow \quad F^{\max}_{X,20T}(x) = e^{-20e^{-\alpha(x-u)}}$$

$$1 - F^{\max}_{X,20T}(x) = 1 - F^{\max}_{X,20T}\left(15000\right) = 1 - e^{-20e^{-4.2756\bullet 10^{-4}(15000-8649.81)}}$$

$$= 1 - e^{-20e^{-2.7150}} = 1 - e^{-1.324} = 1 - 0.266 = 0.734$$

What is hypothesis testing?

An example:

1. We want to judge whether a coin is fair or not.
2. Therefore we toss the coin 20 times and count the number of heads/tails.
3. As a result, we obtain 16 heads.
4. We had already decided that we judge the coin to be fair if the number of heads is between 6 and 14 and otherwise we judge the coin to be not fair.
5. In accordance with the rule we had decided, we conclude that the coin is not fair.

What is important in this example are:

1. how many trials are necessary to judge the hypothesis?
2. how should we decide the operating rule?

Truth, hypothesis and judgment

Assume that the world consists of two sets, Θ_0 and Θ_1. Θ_0 and Θ_1 are complementary and one of them is the truth.

We want to judge in which set the truth belongs. Since we can never know the truth, what we can do is to believe one of them to be the truth by some artificial rule which we create by ourselves. So it is possible that we might make misjudgments!!

Hypothesis:

$$H_0 : \theta \in \Theta_0$$

$$H_1 : \theta \in \Theta_1$$

Operating rule:

$$\mathbf{x} \in C \Rightarrow reject\ H_0\ (= accept\ H_1)$$

$$\mathbf{x} \in \bar{C} \Rightarrow accept\ H_0$$

x is the value of sample statistic

C is the critical region

	Truth Judgment	Θ_0	Θ_1
Hypothesis	Accept H_0	Very good!	Type 2 error
	Accept H_1	Type 1 error	Very good!

Control the type 1 error

The probability that we make type 1 error is $P[\mathbf{X} \in C \mid \theta \in \Theta_0]$

We use the term "level of significance" α in order to state to which degree we allow the type 1 error as:

$$P[\mathbf{X} \in C \mid \theta \in \Theta_0] = \alpha$$

or equivalently

$$P[\mathbf{X} \in \bar{C} \mid \theta \in \Theta_0] = 1 - \alpha$$

Finally we can create the operating rule in accordance with the above equations.

We can use the knowledge we have gained about probability to create the rule!!!

Coin problem again!

We want to judge whether the coin is fair or not.

The world consists of

$\Theta_0 = \{p \mid p = 0.5\} = \{0.5\}$

$\Theta_1 = \{p \mid p \neq 0.5\} = \{$*the set of all possible values between 0 and 1 excluding 0.5*$\}$

Hypotheses are

H_0: $p = 0.5$

H_1: $p \neq 0.5$ *(p is the probability that the coin lands with the head up.)*

We toss the coin 20 times and count the number of heads.

Statistic utilized for this purpose : the number of heads = N

Note that N is a random variable **before** we begin tossing the coin.

We formulate a suitable operating rule as:

The null-hypothesis cannot be rejected at the α level of significance if

$$10 - k \leq N \leq 10 + k, \qquad k \text{ is an integer} > 0$$

We select a level of significance $\alpha = 0.05$, meaning that we accept the occurrence of a type 1 error with a probability of 5%

We thus need to judge

$$P[\mathbf{X} \in \bar{C} \mid p = 0.5] = 1 - \alpha$$

The operating rule hence becomes:

$$P[10 - k \le N \le 10 + k \mid p = 0.5] = 1 - \alpha = 1 - 0.05 = 0.95$$

The probability that $\underbrace{N \text{ is between } 10 - k \text{ and } 10 + k}_{\mathbf{X} \in \bar{C}}$ given $\underbrace{p = 0.5}_{\theta \in \Theta_0}$ is

$$P[10 - k \le N \le 10 + k \mid p = 0.5] = \sum_{i=10-k}^{10+k} \binom{20}{i} 0.5^{i} (1 - 0.5)^{20-i} = 0.95$$

By solving this equation we obtain $k = 4$.

Therefore the operating rule is obtained as:

“H_0 cannot be rejected if the realization of N is between 6 and 14, otherwise H_0 can be rejected, i.e. H_1 can be accepted”

Summary

We want to judge whether the coin is fair, $H_0 : p = 0.5$

We obtained 16 heads out of 20.

In accordance with the operating rule, we reject H_0 (= accept H_1).

We state this: "the hypothesis H_0 is rejected at the 5% level of significance."

As a conclusion, the coin is not fair.

General procedure for hypothesis testing

1. Specify what you want to judge as the null hypothesis: H_0 (complimentary set is H_1)
2. Determine the condition of sampling (what kind of and how many data?)
3. Create the operating rule (as a function of sample statistic)
4. Choose the level of significance α and evaluate the operating rule
5. Execute the sampling and obtain the result.
6. Judge the hypothesis H_0 depending on the result.

Exercise 8.2

In order to check the quality of the concrete production in a construction site, the compressive strength of the concrete produced were measured. Based on previous experience the compressive strength is assumed to follow the Normal distribution and its variance is assumed known and equal to 16.36 (MPa)^2. Acceptance criterion of the concrete production is that the mean of the population of the produced concrete is ~~equal to or more than 30 (MPa)~~ equal to 30 $\pm \Delta$ (MPa).

Make this correction in the exercise question; the solution is **correct**

15 samples have been tested and the results are shown in Table 8.2.1.

Sample Number	Compressive Strength [MPa]	Sample Number	Compressive Strength [MPa]
1	24.4	9	30.3
2	26.5	10	39.7
3	27.8	11	38.4
4	29.2	12	33.3
5	39.2	13	33.5
6	37.8	14	28.1
7	35.1	15	34.6
8	30.8		

Should the quality of the concrete production be accepted?
Test the hypothesis at significance levels of 10% and 1%.

1. Specify what you want to judge as the null hypothesis: H_0

H_0: The mean of the compressive strength μ is **equal to** or **more than** 30 [MPa]

H_1: The mean of the compressive strength μ is **less than** 30 [MPa]

2. Determine the condition of sampling (what kind of and how many data?)

15 concrete samples are taken from the construction site and their compressive strengths are measured .

Number of samples: N = 15

3. Create the operating rule (as a function of sample statistic)

$$30 - \Delta \leq \overline{X} \leq 30 + \Delta$$

where $\overline{X}$ is the sample mean value of the compressive strength

4. Choose the level of significance α and evaluate the operating rule

Level of significance α = 0.1 (10%)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{15} X_i$$

The sample statistic used is the sample mean, $\overline{X}$

is Normal distributed
with mean = 30
variance = 16.36/15

The operating rule is evaluated as

$$P(30-\Delta \le \overline{X} \le 30+\Delta) = 0.9 \Rightarrow P(\frac{(30-\Delta)-\mu}{\sigma/\sqrt{n}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{(30+\Delta)-\mu}{\sigma/\sqrt{n}}) = 0.9$$

$$P(\frac{-\Delta}{\sqrt{16.36/15}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{\Delta}{\sqrt{16.36/15}}) = 0.9 \Rightarrow \Phi(\frac{\Delta}{\sqrt{16.36/15}}) = 0.95$$

By solving this equation we obtain Δ = 1.72

4. Choose the level of significance α and evaluate the operating rule

Level of significance α = 0.1 (10%)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{15} X_i$$

is Normal distributed
with mean = 30
variance = 16.36/15

The sample statistic used is the sample mean, $\bar{X}$

The operating rule is evaluated as

$$P(30-\Delta \le \bar{X} \le 30+\Delta) = 0.9 \Rightarrow P\left(\frac{(30-\Delta)-\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{(30+\Delta)-\mu}{\sigma/\sqrt{n}}\right) = 0.9$$

$$P\left(\frac{-\Delta}{\sqrt{16.36/15}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{\Delta}{\sqrt{16.36/15}}\right) = 0.9 \Rightarrow \Phi\left(\frac{\Delta}{\sqrt{16.36/15}}\right) = 0.95$$

By solving this equation we obtain Δ = 1.72 MPa
The operating rule is hence obtained as:

"H_0 cannot be rejected at the 10% level of significance if the sample mean lies between **28.28** MPa and **31.72** MPa, otherwise H_0 can be rejected (and H_1 can be accepted)"

5. Execute the sampling and obtain the result.

The sample mean is

$$\begin{aligned}\overline{x} &= \frac{1}{15}(24.4+26.5+27.8+29.2+39.2+37.8+35.1+30.8+\\ &\qquad 30.3+39.7+38.4+33.3+33.5+28.1+34.6)\\ &= 32.58\ MPa\end{aligned}$$

6. Judge the hypothesis H_0 depending on the result.

The sample mean is 32.58 MPa

Since the sample mean lies **outside** the interval $\left[28.28MPa \le \bar{x} \le 31.72MPa\right]$,

the null hypothesis is rejected at the **10% level of significance**.

Following the same procedure, the interval for accepting the null hypothesis at the **1% level of significance** is obtained as $\left[27.31MPa \le \bar{X} \le 32.69MPa\right]$.

In this case , since the sample mean (32.58 MPa) is **within** this interval, the null hypothesis cannot be rejected at the **1% level of significance**.

Exercise 8.5
The weekly working hours in the building industry were decided to be reduced by 2 hours. On a construction site it is to be tested whether this new rule is applied or not since the labor union insists that the workers work the same hours as before the reduction. 9 workers were selected randomly and their weekly working hours were measured before (X) and (Y) after the reduction of the working hours It is assumed that X and Y are Normal distributed. The variance of the weekly working hours before and after the reduction is assumed to be $\sigma_X^2 = \sigma_Y^2 =$ 9.5 hours^2.

a) Can it be said, based on the data, that the mean of the working hours before the reduction is 40 hours per week at 5% significance level.

Number of workers	Working time Before reduction	Working time After reduction
1	38	38
2	41	39
3	40	41
4	42	39
5	43	40
6	40	40
7	39	39
8	37	38
9	43	40

1. Specify what you want to judge as the null hypothesis: H_0

H_0: The mean of the working hours per week before the reduction μ_X is **equal to** 40 hours per week.

H_1: The mean of the working hours per week before the reduction μ_X is **not equal to** 40 hours per week.

2. Determine the condition of sampling (what kind of and how many data?)

9 workers are selected randomly and their weekly working hours are measured .

Number of samples: N = 9

3. Create the operating rule (as a function of sample statistic)

$$40 - \Delta \leq \overline{X} \leq 40 + \Delta$$

where $\overline{X}$ is the sample mean value of the weekly working hours of the workers before the reduction

4. Choose the level of significance α and evaluate the operating rule

Level of significance α = 0.05 (5%)

The sample statistic used is the sample mean, $\overline{X}$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{15} X_i$$

is Normal distributed
with mean = 40
variance = 9.5/9

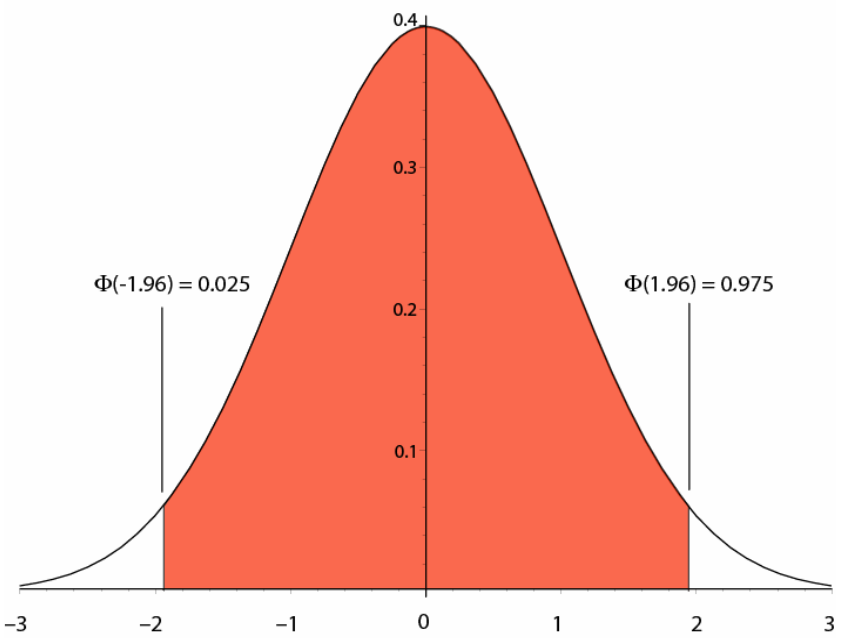
The operating rule is evaluated as

$$P(40-\Delta \le \overline{X} \le 40+\Delta) = 0.95 \Rightarrow P(\frac{(40-\Delta)-\mu}{\sigma/\sqrt{n}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{(40+\Delta)-\mu}{\sigma/\sqrt{n}}) = 0.95$$

$$P(\frac{-\Delta}{\sqrt{9.5/9}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{\Delta}{\sqrt{9.5/9}}) = 0.95$$

$$\Rightarrow \Phi(\frac{\Delta}{\sqrt{9.5/9}}) = 0.975$$

By solving this equation we obtain
Δ = 2.01 hours



4. Choose the level of significance α and evaluate the operating rule

Level of significance α = 0.05 (5%)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{15} X_i$$

The sample statistic used is the sample mean, $\overline{X}$

is Normal distributed
with mean = 40
variance = 9.5/9

The operating rule is evaluated as

$$P(40-\Delta \le \overline{X} \le 40+\Delta) = 0.95 \Rightarrow P(\frac{(40-\Delta)-\mu}{\sigma/\sqrt{n}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{(40+\Delta)-\mu}{\sigma/\sqrt{n}}) = 0.95$$

$$P(\frac{-\Delta}{\sqrt{9.5/9}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{\Delta}{\sqrt{9.5/9}}) = 0.95 \Rightarrow \Phi(\frac{\Delta}{\sqrt{9.5/9}}) = 0.975$$

By solving this equation we obtain Δ = 2.01 hours
The operating rule is hence obtained as:

"H_0 cannot be rejected at the 5% level of significance if the sample mean of the weekly working hours of the 9 workers lies between **37.99** hours and **42.01** hours, otherwise H_0 can be rejected (and H_1 can be accepted)"

5. Execute the sampling and obtain the result.

The sample mean is

$$\begin{aligned}\overline{x} &= \frac{1}{9}(39+41+40+42+43+40+39+37+43) \\ &= 40.33\ hours/week\end{aligned}$$

6. Judge the hypothesis H_0 depending on the result.

The sample mean is 40.33 hours/week

Since the sample mean lies **within** the interval $\left[37.99 hours \leq \bar{x} \leq 42.01 hours \right]$,

the null hypothesis cannot be rejected at the **5% level of significance**.

Exercise 8.5

The weekly working hours in the building industry were decided to be reduced by 2 hours. On a construction site it is to be tested whether this new rule is applied or not since the labor union insists that the workers work the same hours as before the reduction. 9 workers were selected randomly and their weekly working hours were measured before (X) and (Y) after the reduction of the working hours It is assumed that X and Y are Normal distributed. The variance of the weekly working hours before and after the reduction is assumed to be $\sigma_X^2 = \sigma_Y^2 =$ 9.5 hours^2.

b) Test the claim of the labor union at the 5% significance level.

Number	Working time	Working time
of workers	Before reduction	After reduction
1	38	38
2	41	39
3	40	41
4	42	39
5	43	40
6	40	40
7	39	39
8	37	38
9	43	40

1. Specify what you want to judge as the null hypothesis: H_0

The variables X and Y refer to the weekly working hours **before** and **after** the reduction of the working hours respectively.

The claim of the labor union is taken as the null hypothesis

$$H_0 : \mu_X = \mu_Y \Leftrightarrow \mu_X - \mu_Y = 0$$

$$H_1 : \mu_X \neq \mu_Y$$

2. Determine the condition of sampling (what kind of and how many data?)

9 workers are selected randomly and their weekly working hours **before** and **after** the reduction are measured .

Number of samples: $N = 9$

3. Create the operating rule (as a function of sample statistic)

$$\overline{X} = \frac{1}{k}\sum_{i=1}^{k} X_i$$

$$\overline{Y} = \frac{1}{l}\sum_{i=1}^{l} Y_i$$

Test two datasets in order to compare two populations.

We are interested in the difference μ$_X$ and μ$_Y$.

So we adopt $Z = \overline{X} - \overline{Y}$ as the sampling statistic for judgment.

Then Z has the following properties:

$$\mu_Z = \mu_{\overline{X}} - \mu_{\overline{Y}} = \mu_X - \mu_Y; \qquad \sigma_Z^2 = \sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2 = \frac{\sigma_X^2}{k} + \frac{\sigma_Y^2}{l}$$

According to the null hypothesis, $\mu_X = \mu_Y \Rightarrow \mu_Z = \mu_{\overline{X}} - \mu_{\overline{Y}} = \mu_X - \mu_Y = 0$

Also $\sigma_Z^2 = \sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2 = \frac{\sigma_X^2}{k} + \frac{\sigma_Y^2}{l} = \frac{9.5}{9} + \frac{9.5}{9} = 2.11$

What is more, Z is normally distributed.

3. Create the operating rule (as a function of sample statistic)

$$\overline{X} - \overline{Y} \leq \Delta$$

$$Z = \overline{X} - \overline{Y} \Rightarrow Z \leq \Delta$$

where $\overline{X}$ *and* $\overline{Y}$ are the sample mean values of the weekly working hours of the workers **before** and **after** the reduction respectively and Z is the difference between $\overline{X}$ *and* $\overline{Y}$

4. Choose the level of significance α and evaluate the operating rule

Level of significance α = 0.05 (5%)

The sample statistic is the difference between the sample means, $Z = \bar{X} - \bar{Y}$

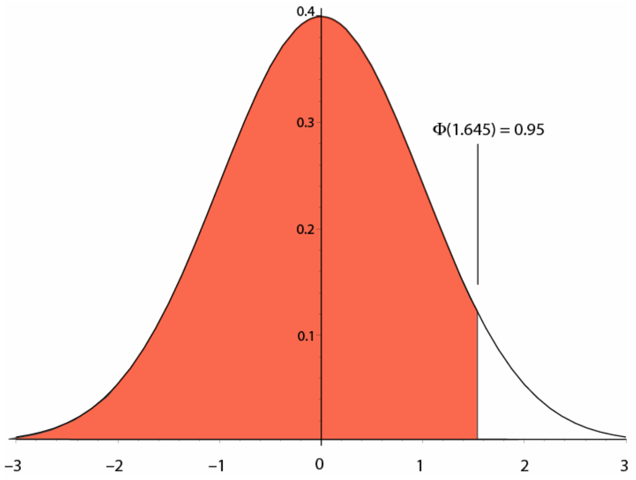
The operating rule is evaluated as

Z is Normal distributed
with mean = 0
variance = 2.11

$$P(Z \leq \Delta) = 1 - 0.05 = 0.95$$

$$P(\frac{Z-\mu_Z}{\sigma_Z} \leq \frac{\Delta-\mu_Z}{\sigma_Z}) = 0.95 \Rightarrow P(\frac{Z-0}{\sqrt{2.11}} \leq \frac{\Delta-0}{\sqrt{2.11}}) = 0.95$$

$$\Rightarrow \Phi(\frac{\Delta}{\sqrt{2.11}}) = 0.95$$



By solving this equation we obtain
Δ = 2.39 hours

4. Choose the level of significance α and evaluate the operating rule

Level of significance α = 0.05 (5%)

The sample statistic is the difference between the sample means, $Z = \overline{X} - \overline{Y}$

The operating rule is evaluated as

Z is Normal distributed
with mean = 0
variance = 2.11

$$P(Z \leq \Delta) = 1 - 0.05 = 0.95$$

$$P(\frac{Z-\mu_Z}{\sigma_Z} \leq \frac{\Delta-\mu_Z}{\sigma_Z}) = 0.95 \Rightarrow P(\frac{Z-0}{\sqrt{2.11}} \leq \frac{\Delta-0}{\sqrt{2.11}}) = 0.95$$

$$\Rightarrow \Phi(\frac{\Delta}{\sqrt{2.11}}) = 0.95$$

By solving this equation we obtain Δ = 2.39 hours
The operating rule is hence obtained as:

"H_0 cannot be rejected at the 5% level of significance if the difference in the sample mean of the weekly working hours before and after the reduction is smaller than or equal to **2.39** hours, otherwise H_0 can be rejected (and H_1 can be accepted)"

5. Execute the sampling and obtain the result.

The difference in the sample mean values is

$$\overline{x} = 40.33\ hours$$

$$\overline{y} = 39.33\ hours$$

$$z = \overline{x} - \overline{y} = 1\ hour$$

6. Judge the hypothesis H_0 depending on the result.

The difference in the sample mean is z = 1 hour

Since the sample mean lies **within** the interval $\left[z \leq 2.39\ hours\right]$,

the null hypothesis cannot be rejected at the **5% level of significance**.

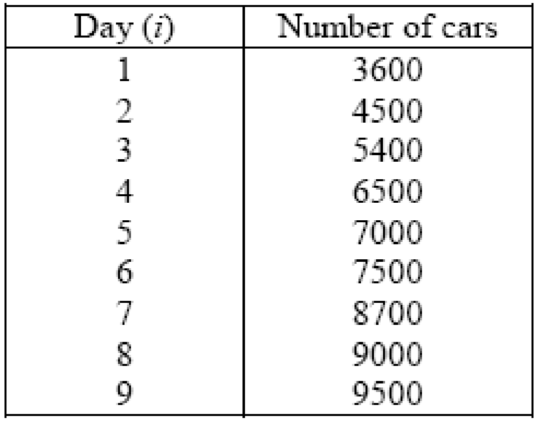
Exercise 8.6

Table 8.3 provides a number of data on the daily traffic flow in Rosengartenstrasse in Zürich.

a) Produce the probability paper for the triangular distribution given by

$$f_X(x) = \begin{cases} \dfrac{2}{10000^2}x & 0 \leq x \leq 10000 \\ 0 & otherwise \end{cases}$$

b) Check if the daily traffic flow is triangularly distributed with the help of the probability paper.



Day (i)	Number of cars
1	3600
2	4500
3	5400
4	6500
5	7000
6	7500
7	8700
8	9000
9	9500

What is a probability paper?

We are interested in whether the sample originates from the given distribution.

We then create the probability paper. If the sample comes from the distribution, the plotted points on the paper are on the straight line.

Probability paper, P-P plot Q-Q plot

All are useful to check the suitability of the modeling (assumed distribution)

But they differ. They are applied…

Before parameter estimation → Probability paper

- ✓ You do not need to estimate the parameters in advance.
- ✓ The suitability is checked whether or not the data on a straight line (not necessarily tangent of 45 degree)

After parameter estimation → P-P plot, Q-Q plot

- ✓ You need to estimate the parameters in advance.
- ✓ The suitability is checked whether or not the data on a straight line whose tangent is 45 degree.

The probability density function and the cumulative distribution functions are

$$f_X(x) = \begin{cases} \dfrac{2}{10000^2}x & 0 \leq x \leq 10000 \\ 0 & otherwise \end{cases}$$

$$F_X(x) = \begin{cases} 0 & 0 \leq x \\ \left(\dfrac{x}{10000}\right)^2 & 0 < x \leq 10000 \\ 1 & x > 10000 \end{cases}$$

Taking the square root of both sides of the first equation equation, a linear relationship between $F_x(x)$ and the square root of x is obtained:

$$F_X(x) = \left(\frac{x}{10000}\right)^2 \Leftrightarrow \sqrt{F_X(x)} = \frac{x}{10000}$$

For values of the cumulative distribution function in the interval $[0;1]$, the following table for the values of $\sqrt{F_X(x)}$ *and* $F_X(x)$ is obtained

$\sqrt{F_X(x)}$	$F_X(x)$
0	0
0.31	0.1
0.45	0.2
0.55	0.3
0.63	0.4
0.71	0.5
0.77	0.6
0.84	0.7
0.89	0.8
0.94	0.9
1.0	1.0

Empirical distribution function

We give the probability for i^{th} large sample

$$F_X(x_i) = \frac{i}{N+1}$$

where N is the total number of samples.

i	No. of cars	$F_X(x_i^o) = \frac{i}{N+1}$
1	3600	0.1
2	4500	0.2
3	5400	0.3
4	6500	0.4
5	7000	0.5
6	7500	0.6
7	8700	0.7
8	9000	0.8
9	9500	0.9

The probability paper is now created by rescaling the y-axis.

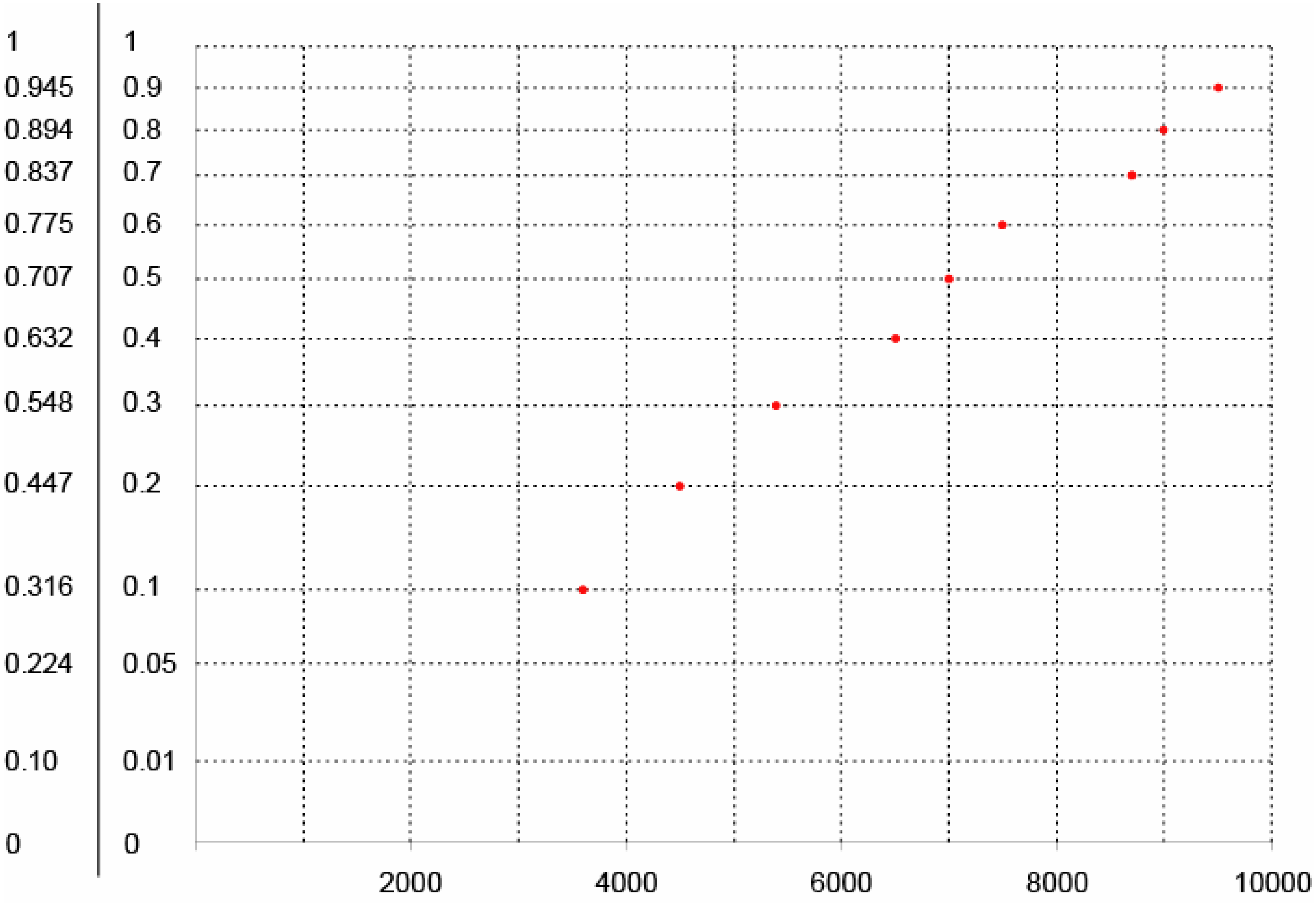
The data is plotted on the probability paper.

If the data fits on a straight line, it follows the triangular distribution.

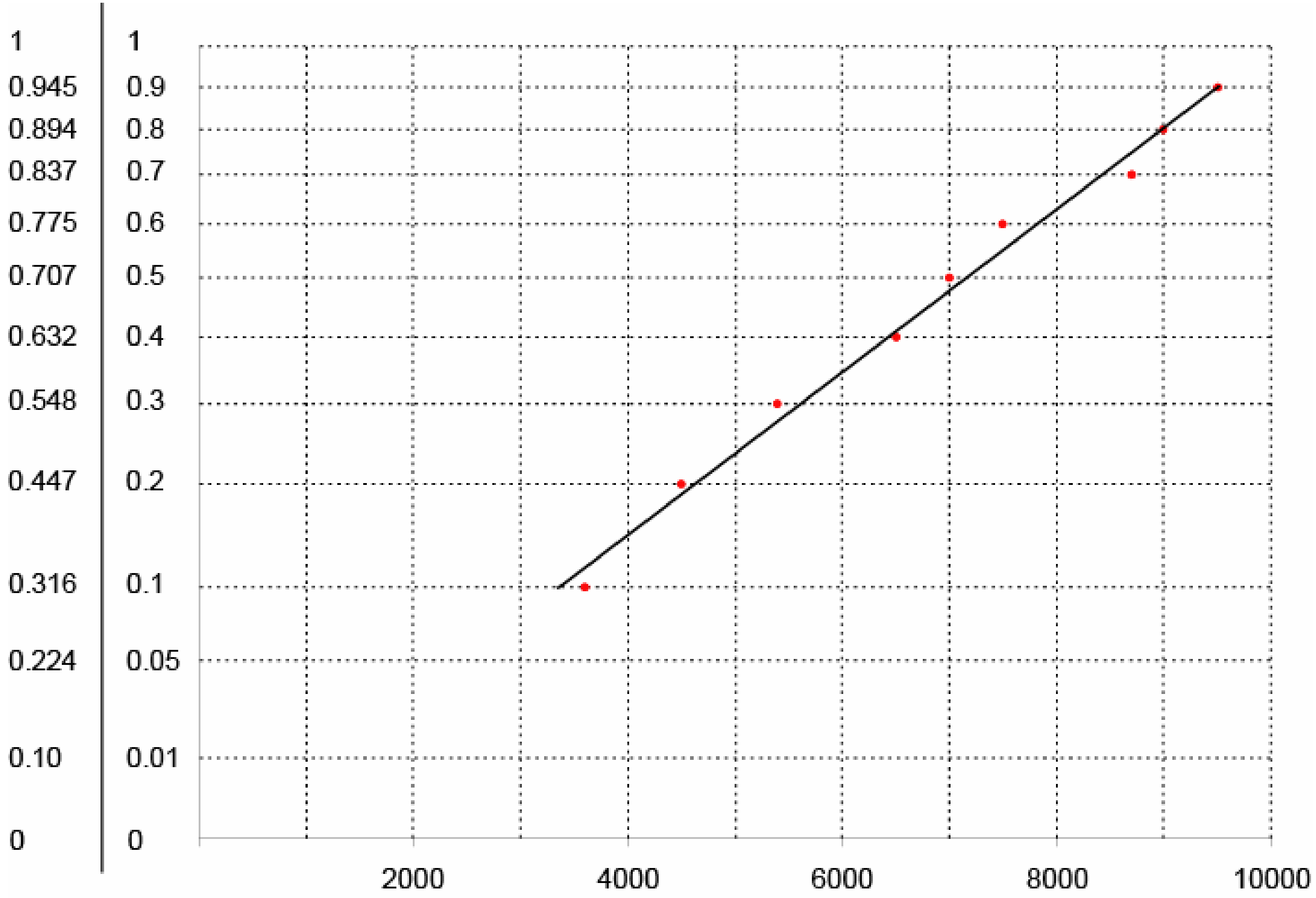
The cumulative distribution function used to plot the data is obtained from the following table:

i	No. of cars	$F_X(x_i^o) = \frac{i}{N+1}$
1	3600	0.1
2	4500	0.2
3	5400	0.3
4	6500	0.4
5	7000	0.5
6	7500	0.6
7	8700	0.7
8	9000	0.8
9	9500	0.9

Plotting the probability paper



Plotting the probability paper


1
0.945
0.894
0.837
0.775
0.707
0.632
0.548
0.447
0.316
0.224
0.10
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0.01
0
2000
4000
6000
8000
10000