Statistics: home-test to exam

1 Data

1. The data have been measured

$$\{4,3,2,3,4,2,3,3,1,3,4,3,5,2,4,4,3,2,1,5,3,2,3,5,2,1,3,2,3,2\}$$

a) Write them in the form of different values and frequencies.

b) Compute: average, variance, mode and median

[

1	2	3	4	5
3	8	11	5	3

, $\bar{x} = 2.6$, $s^2 = 1.223$, $\hat{x} = 3$, $\tilde{x} = 3$]

2. For data

$$\{50,13,21,77,49,4,22,15,31,63\}$$

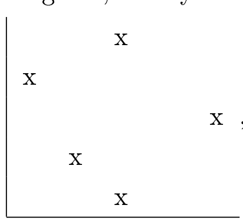
determine ranks.

[sorted: 4, 13, 15, 21, 22, 31, 49, 50, 63, 77, ranks: 8, 2, 4, 10, 7, 1, 53, 6, 9]

3. Draw scatter plot for the data

x	5	3	2	3	1
y	3	1	2	5	4

and guess, if they are suitable for liner regression.

[ , no linear regression]

2 Probability

1. What is the sample space for experiment "flipping two coins" with head = 1 and tail = 0 if a) we distinguish the coins and b) we do not distinguish the coins and c) determine probabilities of the results.

[$\{[0,0],[0,1],[1,0],[1,1]\}$; $\{[0,0],[0,1],[1,1]\}$; $\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}$; $\frac{1}{4},\frac{1}{2},\frac{1}{4}$]

2. Determine conditional probability $f(A|B)$ for the experiment throwing a dice, if A is "even number" and B is a) "greater than 3", b) "less then 4", c) "less or equal to 4". Explain the difference.

[$\frac{2}{3}$, $\frac{1}{3}$, $\frac{1}{2}$]

3. Probability $P(A|B)$ is equal to 0.8 and $P(B)$ is 0.3. What is the probability $P(A,B)$?

   [0.24]

3 Random variable

1. The random variable X has distribution function

$$F(x) = k \cdot x, \ x \in (0,10)$$

   a) Determine the constant k.
   b) Calculate the probability $P(x \leq 3)$.

   [$k = \frac{1}{10}$, $P = \frac{3}{10}$

2. Given the density function

$$f(x) = \frac{1}{\delta} \exp\left(-\frac{x}{\delta}\right), \ \ x \in (0,\infty)$$

   and δ is non-negative constant. Determine:
   a) Distribution function,
   b) expectation
   c) median,
   d) probability $P(x > 5)$.

   [$F = 1 - \exp\left(-\frac{x}{\delta}\right)$, $E[X] = \delta$, median $= -\delta \exp(0.5)$, $P = 1 - \exp\left(-\frac{5}{\delta}\right)$]

3. For discrete random variables $X \in \{1,2\}$ and $Y \in \{1,2\}$ for which

$$P(X = 1, Y = 1) = 0.3, \ P(X = 1, Y = 2) = 0.1$$

$$P(X = 2, Y = 1) = 0.2, \ P(X = 2, Y = 2) = 0.4$$

   determine $f(x|y)$.

   [f(x|y) $= \begin{bmatrix} 0.6 & 0.2 \\ 0.4 & 0.8 \end{bmatrix}$]

4 Distributions

1. Write probability function of random variable X with Poisson distribution with parameter λ and for $\lambda = 4$ determine probability $P(X > 3)$.

   [$f(x) = \exp(-\lambda) \frac{\lambda^x}{x!}$; $P = 1 - 0.433 = 0.567$]

2. Write density function of standard normal distribution and determine its expectation, variance and standard deviation.

   [$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$, $\mu = 0$, $\sigma^2 = 1$]

5 Statistical inference

1. Write formulas for expectation of sample average and variance. Explain their meaning.

   [$E\left[\bar{X}\right] = \mu$, $D\left[\bar{X}\right] = \frac{\sigma^2}{n}$] for n growing grows accuracy of estimation]

2. A population has normal distribution with expectation 120 and variance 16. a) Write distribution of sample average with sample length equal to 100.

   b) What can you say about this distribution if the distribution of the population is not normal?

   [$f(\bar{x}) = \frac{1}{\sqrt{2\pi \cdot 0.16}} \exp\left(-\frac{1}{2}\frac{(x-120)^2}{0.16}\right)$, it will be normal as $n = 100$ is sufficiently large]

3. Write statistics for point estimate of exponential distribution.

   [$\bar{\delta} = \bar{x}$ or $\bar{a} = \frac{1}{\bar{x}}$]

4. a) Write definition of unbiased point estimate.

   b) Show that sample average is unbiased point estimate of expectation.

   [$E[T] = \theta$, $E\left[\bar{X}\right] = \mu$]

6 Confidence intervals

1. Flour is sold in kilogram packs. Inspectors checked their weight each day for one year. From the measurements they calculated average 1.05 kg and variance 0.03 kg^2. Determine the weight interval in which a newly purchased package of flour will lie with a probability of 10%.

   [$I = (1.035,\ 1.065)$]

2. From normal distribution the following set of data has been measured

   $$\{5.2, 7.7, 3.6, 6.8, 7.2, 4.3, 5.1, 7.8\}$$

   Determine confidence interval for 5%

   [$I = (4.61,\ 7.31)$]

7 Statistical testing

1. In two classes, A and B, a mathematics test was written. Tests were scored (more points, better score). The scores obtained are in the following table

A	8	9	8	8	9	5	10	9	7	6	6	9
B	5	9	10	10	10	8	9	8	8	9		

At the level 5% test zero hypothesis H0: The results in math are better in the class A.

[left, pv=0.126, do not reject]

2. One class took an English test at half-term and the same pupils took a similar test at the end of the year. The scores (more is better) for the students who were just shy are in the table below

Pupil	1	2	3	4	5	6	7	8	9	10	11	12
Half-term	6	9	8	7	9	5	4	6	5	7	7	6
End	6	8	9	9	7	6	4	5	6	8	6	7

At the level 5% test zero hypothesis H0: Pupils improved towards the end of the year. Test it under assumption:
a) The data come from the normal distribution.
b) Normal distribution cannot be assumed.
[a) left, pv=0.319; b) pv=0.323]

3. One class took an English test and then the students took an intensive course in English. After that, a new test was written by the same pupils. In the tests, the pupils either succeeded (1) or failed (0). The results are in the table below

Pupil	1	2	3	4	5	6	7	8	9	10	11	12
1st test	1	1	0	1	1	0	0	1	0	1	1	0
2nd test	1	1	1	0	0	0	1	1	0	1	1	1

At the level 5% test zero hypothesis H0: Intensive course had an effect.
[make contingency table, pv=0.655]

4. In four periods: in winter - 2 weeks, in spring - 5 weeks, in summer - 3 weeks and in autumn - 6 weeks, traffic violations were recorded at a certain location. The measured values are shown in the table below

winter	spring	summer	autumn
8	35	9	31

At the level 5% test zero hypothesis H0: The violations occur uniformly.
[E =10.375, 25.9375, 15.5625, 31.125; pv=0.09]

5. Five racing motorcycles were tested on the track. Each motorcycle was driven around the track five times and timed. The results are in the table

bike	1	2	3	4	5
time 1	128	112	135	108	122
time 2	122	120	124	109	120
time 3	125	111	131	112	125
time 4	131	115	130	105	121
time 5	124	118	137	119	117

At the level 10% test H0: the performance of the racing motorcycles is the same:
a) assuming normality,
b) not assuming normality.
[anova, pv=0.973; kruskal-wallis, pv=0.986]

6. The profits of a particular firm in a given years were

year	2005	2010	2012	2015	2018	2021	2022	2023
profit [thous. $]	510	540	520	560	570	560	590	630

a) Perform (*i*) linear, (*ii*) exponential, (*iii*) quadratic regression and write p-values.
b) Decide which regression is best.
c) Determine trend of the profits (rising or falling).

[pv1 = 0.004, pv2 = 0.0032, pv3 = 0.0166; second is best; rising $b_1 > 0$]