1 Answers to the questions

Please, try to answer the questions first of all on the basis of the textbook. These answers should only acknowledge your answers if you think they are correct.

1.1 Variables and data

1. Data file is a set of measured data.

2. Plain form stores data in the order how they are measured. Alternative way is to store only different values and the numbers of their occurrence.

3. The ranks are 6.5, 4, 1, 6.5, 8, 5, 3, 2. For the same values, the average of their order is taken.

4. Level of data is given by expectation, mode or median, variability corresponds to variance, standard deviation.

5. Median is 4.

6. Average is 4.228, variance 5.80748.

7. Mode is 2.

8. General graph indicates values, histogram frequencies.

1.2 Probability and random variable

1. It is any action leading to some of correctly defined result. Repeating the action leads accidentally to different results.

2. Event is a set of results. It can be also specified verbally. E.g. at a dice, we can say “even number” and it determines the set $\{2, 4, 6\}$.

3. They are “Head” and “Tail”.

4. They are usually 0 and 1. They also can be 1 and 2 or any other two numbers.

5. They are that probability is (i) non-negative $(P(A) \geq 0)$, (ii) less or equal to 1 $(P(A) \leq 1)$, (iii) σ-additive (for events A, B such that $A \cap B = \emptyset$ it holds $P(A \cup B) = P(A) + P(B)$.

6. $P = \frac{m}{n}$, where m is number of all possible positive results and n is number of all possible results.

7. $P = \frac{M}{N}$, where M is number of all performed positive results and N is number of all performed results.

8. Classical definition takes into account possible results while statistical one speaks about performed results. The former is theoretical and its result is constant, the latter practical and its results a bit vary with a specific setof experiments.

9. Here we have used the statistical definition.

10. This relates to the classical definition.

11. It is denoted $P(A|B)$ and it is computed plainly as a probability of A but on the set of results which meet the condition. With a dice $P(even| < 5)$ is a probability of even on a set $\{1, 2, 3, 4\}$ which is $2/4 = 0.5$.
   The definition is
   $$P(A|B) = \frac{P(A, B)}{P(B)}$$

12. $A = "even" = (2, 4, 6)$, $B = " < 4" = \{1, 2, 3\}$. The set B has 3 elements, one of them is even: $P = 1/3$.

13. They are not independent as the probability depends on which ball was previously drown.

14. They are not. "different sides" has twice much possibilities.

15. $P(x|y) = P(x, y)/P(y) \underbrace{=}_{def..} P(x)$. Multiplication by $P(y)$ gives the result.

1.3 Description of random variable and vector

1. Random variable corresponds to random variable. However, values of random variable (which correspond to results of experiment) must be numeric. If results are not numeric (red, yellow, green), we must assign them numbers (e.g. 1,2,3).

2. Discrete and continuous.

3. Yes, it can.

4. There are two things: (i) common treatment of several random variables, (ii) mutual connection between random variables.

5. They are vectors of numbers.

6. $F_X(x) = P(X \leq x)$, X is random variable, x is a number.

7. Zero for $x \to -\infty$, one for $x \to \infty$ and non-decreasing on the whole support.

8. $P(x \in (a, b)) = F(b) - F(a)$.

9. Probability of each single number is zero for continuous random variable.

10. No, discrete ones are only piece-wise continuous.

11. It is a discrete function with values equal to individual values of random variable.

12. It is a derivative of the distribution function.

13. Non-negative values and sum (integral) equal to one.

14. (iii) is correct.

15. (iii) is correct.

16. $P(a,b) = \int_a^b f(x)\,dx$

17. $E[X] = \sum_{x_i \in X} x_i f(x_i)$

18. $E[X] = \int_{-\infty}^{\infty} x f(x)\,dx$

19. $\int_{-\infty}^{\zeta_\alpha} f(x)\,dx = \alpha$, $\int_{z_\alpha}^{\infty} f(x)\,dx = \alpha$

20. Median $\tilde{x} : \int_{-\infty}^{\tilde{x}} f(x) = 0.5.$

21. Mode $\hat{x} = \arg\max(f(x))$

22. $k = \frac{1}{5}$ (integral of $f(x)$ must be 1)

23. Distribution function is a cumulative sum of $f(x)$

24. (i) $f(x) = \begin{cases} 0.5 & \text{for } x = 0 \\ 0.5 & \text{for } x = 1 \end{cases}$, (ii) $f(x) = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases}$

25. The density function is a constant $\frac{1}{5}$ from $x = 0$ to $x = 5$. The distribution function is: zero on $x \in (-\infty, 0)$, $F(x) = \frac{1}{5}x$ for $x\,(0,5)$ and one for $x \in (0, \infty)$.

26. Joint is a product (they are independent).

27. Yes, they are. $f(x,y) = k \exp\left(x^2\right) \exp\left(2y^2\right) = f(x) f(y)$.

28. $P = 0.1 \cdot 0.2 = 0.02$. It is the volume of a prism with edges 0.1, 0.2 and 1.

29. $k = 1 - sum(\mathrm{of\,the\,rest}) = 0.1$

30. Marginal (sum over columns) 0.5, 0.2, 0.3; Conditional (joint columns divided by marginal) $\begin{matrix} \frac{2}{5} & \frac{1}{2} & \frac{1}{3} \\ \frac{3}{5} & \frac{1}{2} & \frac{2}{3} \end{matrix}$.

31. $f(x=1) = p^1 (1-p)^{1-1} = p$; Expected number is $p \times 100$ (p in one experiment, $p \times 100$ in 100 experiments).

1.4 Important distributions

1. Random variable with two outcomes 0 and 1, where 1 has probability p (that is constant). E.g. Selection of a product hat can be either good (1) or defective (0), if $p \times 100\%$ of products is good.

2. p is a probability of the result $x = 1$.

3. Binomial distribution $f(x = 2) = \binom{5}{2} 0.52^2 0.48^{5-2} = 0.299$.

4. We consider a Bernoulli trial with $P(x = 1) = p$. We perform n experiments a want to know with probability k of them will be 1. !! p must stay constant during the experiments) !!

5. It is he same table where the second row is divided by the sum of all its entries. 0.058, 0.224, 0.078, 0.340, 0.300.

6. Geometrical distribution $f(x = 3) = 0.5 \cdot (1 - 0.5)^2 = 0.125$.

7. It is the geometrical distribution with $p = 0.01$.

8. They are both the same $(-\infty, \infty)$.

9. $f(x) = N\left(10, \sigma^2\right)$, $f(y) = N\left(0, \sigma^2\right)$ where σ^2 is the variance of measurements.

10. For $X \sim U(a, b)$ it holds a) all values in the range (a, b) are equally probable and b) all values outside this interval are impossible.

11. $f(x)$ on $(3, 9)$ is $\frac{1}{6}$. So, $P(x \in (5, 7)) = \frac{1}{6}(7 - 5) = \frac{1}{3}$.

1.5 Regression analysis

1. $y = 4x - 3$ (line going through the points); $r = 0$ (residuals are zero)

2. The regression line is (use Statext) $y = 1.5x$. The residuals -0.5, 1, -0.5 $(= \cdots y - yp$, where yp is prediction - value on the line). Sum of squares is $0.25 + 1 + 0.25 = 1.5$.

3. It will be $0.21 \cdot 5000 - 100 = 950$.

4. It i the solution of he equation $1000 = 0.21x - 100$; $x = 4285.7$.

5. $\ln(y) = \ln(b_0) + b_1 x$

6. Quadratic: $y = b_0 + b_1 x + b_2 x^2$; cubic: $y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$.

7. It is $y = 2.1 + 0.6 \cdot 2 + 1.2 \cdot 2^2 + 0.1 \cdot 2^3 = 8.9$.

1.6 Population and data sample

1. Very roughly: population is big and constant, sample is relatively small and it changes when taking a new one.

2. Random sample is a vector of equally distributed and independent random variables and sample realization is a vector on measured numbers.

3. It is random variable (average of realization of sample is a number).

4. It is a number.

5. Population are speeds of all cars in the world. Random sample are speeds of a given number of cars that will potentially be measured (it is a definition of experiment - measure speeds of n cars), sample realization are really measured speeds.

6. Population are numbers $\{1, 2, 3, 4, 5, 6\}$ with equal probabilities. Random sample is a vector of 10 entries - each entry is prepared for a value obtained on the thrown dice. Sample realization is a vector with numbers that really were obtained on the dice.

7. Expectation is not mentioned so it must be estimated.

8. Parameter is an unknown constant, estimate is a number computed from measured data that is close to the true value of the parameter.

9. Statistics is a function of random sample which when a sample realization is inserted gives a value near the estimated value of the parameter.

10. They are sample average and sample variance.

11. $E\left[\bar{X}\right] = \mu$ and $D\left[\bar{X}\right] = \frac{\sigma^2}{n}$

12. It says that if we take very many sample realizations, from each we compute sample average and then we take average from these averages we get practically precise value of the population expectation.

13. It expresses the fact that the larger the sample is the higher is the precision of the sample average as an estimate of the expectation.

14. It is unbiased, consistent and the larger is the sample length the higher is its efficiency.

15. The one which is computed from larger sample.

1.7 Statistical inference

1. It is the sample average.
2. It is the sample variance.
3. It is the sample proportion.
4. It is the correlation coefficient.
5. It is normal with expectation μ and variance $\frac{\sigma^2}{n}$.
6. For computation of confidence interval we must use distribution of the statistics $f(\bar{x})$.
7. For both-sided interval we use $\frac{\alpha}{2}$ on both sides; for left-sided respectively right-sided interval we use α on the left respectively right side.
8. Zero H0 is currently valid, alternative HA rejects H0.
9. Region of acceptance is equal to the confidence interval, critical region is its complement.
10. T_t is the value of the statistics with the inserted sample realization.
11. HA for both-sided test says “is not equal to”; for left sided it says “is less than” and for right-sided “is greater than”.
12. Confidence interval is equal to the region of acceptance.
13. We reject H0.
14. We reject H0.

1.8 Validation in regression analysis

1. The closer the data points are to the regression curve, the better is the regression.
2. It tests if x and y are uncorrelated. If yes, the regression is not possible.
3. Yes, it is. (x and y are not uncorrelated)
4. Yes, it is.
5. Yes, it is.
6. Yes, they should.