12. KEYS TO SOLUTION

12.1. Lecture 1

1. Exploratory data analysis is often a first step in revealing information hidden in large amount variables and their variants.
2. The base kinds of variables are a quantitative variable (nominal, ordinal) and a qualitative variable (discrete, continuous).
3. The frequency table is concerning with absolute and relative frequencies (for a qualitative variable).
4. The outliers are the variable values which significantly differ from other values.
5. b)
6. qualitative variable – bar chart, pie chart
   quantitative variable – box plot, stem and leaf plot
7. b) 14 thousand, d) cca (9;19)

12.2. Lecture 2

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2. $P(A \cap B) = P(A) \cdot P(B)$
3. Two events are independent if intersection probability of these two events is equal to a product of individual event probabilities.

12.3. Lecture 3+4

1. $F(x) = \sum_{x_i < x} P(X = x_i)$
2. $F(x) = \int_{-\infty}^{x} f(t)dt$ for $-\infty < x < \infty$
3. 50% quantile is called a median

   A mode is a value in which the discrete RV comes with the biggest probability.
4. The conditional distribution is the distribution of one variable at a fixed value of the other jointly distributed random variable.
5. $X_1 \ldots X_n$ are mutually independent $\Leftrightarrow$ $F(x_1,\ldots,x_n) = F_1(x_1) \ldots F_n(x_n)$.
6. The correlation coefficient measures the strength of the relation between two random variables.

12.4. Lecture 5

1. discrete distribution – binomial, geometric, negative binomial
   continuous distribution – poisson, exponential, Weibull, Gamma
2. A sequence of Bernoulli trials is defined as a sequence of random events which are mutually independent and which have only two possible outcomes and the probability

of event occur p is constant in any trial. On the basis of these trials expectations we can define the following random variables: binomial, geometric and negative binomial.
mean of the binomial random variable: EX=np

3. A Gamma distribution describes a time to k-th event occurrence in a Poisson process
4. β=2

12.5. Lecture 6

1. X ... RV with $N(\mu,\sigma^2)$ => $Z = \frac{X-\mu}{\sigma}$ ... N(0,1)
2. Chebyshev's inequality puts an upper bound on the probability that an observation should be far from its mean.
3. The law of large number is a theorem about convergence of means in the sequence of the random variables.
4. Chí-square distribution is a distribution derived from sum of squared standard normal random variables.

12.6. Lecture 7

1. Inferential statistics or statistical induction comprises the use of statistics to make inferences concerning some unknown aspect of a population.
2. A random sample is a set of items that have been drawn from a population in such a way that each time an item was selected, every item in the population had an equal opportunity to appear in the sample.

12.7. Lecture 8

1. The p-value calculation depends on defined null hypothesis:

a) $H0: \mu < \mu_0$ => p-value=F(x_{obs})

b) $H0: \mu > \mu_0$ => p-value=1-F(x_{obs})

c) $H0: \mu = \mu_0$ => p-value=2{F(x_{obs}), 1-F(x_{obs})}

2. It is a hypothesis that is accepted in case the rejection of null hypothesis.
3. $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\cdot(1-\hat{p})}\cdot\left(\frac{1}{n_1}-\frac{1}{n_2}\right)}$ ... N(0,1) , where $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$

12.8. Lecture 9

1. An $\hat{\theta}$ estimation is consistent if

   a) $\hat{\theta}$ is asymptotically unbiassed, $E\hat{\theta} \to \theta$

   b) $\lim_{n \to \infty} D\hat{\theta} = 0$

2. $P(T_D(\overline{X}) \le \theta \le T_H(\overline{X})) \ge 1-\alpha$

12.9. Lecture 10

1. $F = \frac{S_B^2}{S_W^2}$ ... F-distribution with *(k*-1*)* and *(N-k)* degrees of freedom
2. ANOVA table
3. The post hoc analysis is a second step of ANOVA and consists of comparing the means of all pairs of groups of purpose to choose homogenous groups.

12.10. Lecture 11

1. y=a+b*x, where y is a dependent variable and x is an independent variable. The values a (intercept) and b (slope) are estimates of regression line parameters
2. x and y are independent variables, in the case that p-value > 0,05
3. Coefficient of determination predicate about suitability of used model