Statistical Methods for Data Science

Students cannot use teaching materials, smartphones or computers. They can only use a scientific calculator. Duration of the written exam is 2h.

Exercise 1 (6 points). Let X and Y be two independent random variables, where X has an N(2, 5) distribution and Y has an N(5, 9) distribution. Define Z = 3X−2Y +1.

a. Compute E[Z] and Var(Z).
b. What is the distribution of Z?
c. Compute P(Z ≤ 6).

Hint. For a random variable W ~ N(0, 1), we have P(W > 1) = 0.1587.

Solution. See full solution of Ex. 11.4 a,b,c at page 456 of [B1].

Exercise 2 (6 points). Environmentalists have taken 16 samples from the wastewater of a chemical plant and measured the concentration of a certain carcinogenic substance. They found $\bar{x}_{16}$ = 2.24 (ppm) and s^2_{16} = 1.12, and want to use these data in a lawsuit against the plant. It may be assumed that the data are a realization of a normal random sample.

a. Construct the 97.5% one-sided confidence interval that the environmentalists made to convince the judge that the concentration exceeds legal limits.
b. The plant management uses the same data to construct a 97.5% one-sided confidence interval to show that concentrations are not too high. Construct this interval as well.

Hint. $t_{15,0.025}$ = 2.131

Solution. See full solution of Ex. 24.6 a,b at page 470 of [B1].

Exercise 3 (6 points). One generates a number x from a uniform distribution on the interval [0, θ]. One decides to test H0 : θ = 2 against H1 : θ = 2 by rejecting H0 if x ≤ 0.1 or x ≥ 1.9.

a. Compute the probability of committing a type I error.
b. Compute the probability of committing a type II error if the true value of θ is 2.5.

Solution. The text is from Ex. 26.3 of [B1]. Here there are the solutions:

(a) A type I error is done when H0 is true but it is rejected. Hence it is the probability that x ≤ 0.1 or x ≥ 1.9. Given that the distribution is uniform this probability is $P_2(x \le 0.1)+P_2(x \ge 1.9)=(0.1)/2+(0.1)/2=0.1$ where P_2 is the probability for θ = 2.
(b) A type II error is done if the true value of θ is 2.5 but H0 is not rejected. Thus it is $P_{2.5}(0.1 \le x \le 1.9)$ = (1.9-0.1)/2.5 = 0.72, where $P_{2.5}$ is the probability for θ = 2.5.

Exercise 4 (6 points). Write R code to simulate 10 coin-tosses using:

a. function `sample()`
b. function `rbinom()`

Solution. For (a) see 3rd paragraph at page 56 of [B2]. For (b) see solution of Ex. 3.4 at page 339 of [B2].

Exercise 5 (6 points). Write an R function to compute p-value in 2-sided t-test without using the pre-defined built-in t.test().

Solution. (see also smdR6.R on the web page of the course)

```
pvalue = function(data, m, n)  # data = vector of n values, m = actual mean
{
  xbar <- mean(data) # sample mean
  sbar <- sd(data) # sample variance
  t0 <- sqrt(n)*(xbar-m)/sbar # studentized mean
  v = pt(t0, n-1) # P(t <= t0)
  p <- min(v, 1-v) # lower tail
  return (2*p) # 2-sided
}
```

References

[B1] F.M. Dekking C. Kraaikamp, H.P. Lopuha, L.E. Meester. A Modern Introduction to Probability and Statistics. Springer, 2005.

[B2] P. Dalgaard. Introductory Statistics with R. 2nd edition, Springer, 2008.