MAT 2379, Introduction to biostatistics

Assignment 3 - solutions

Part I: Solve the following problems using a calculator permitted by the Faculty of Science (TI30, TI34, Casio fx-260 and Casio fx-300), and the table for the cumulative distribution of the standard normal.

[10] **Question 1:** Problem 7.4 from the textbook.

Let x be the weight of the horns of a white rhino (in kg).

(a) We computed the following sums: $\sum_{i=1}^{28} x_i = 35.9$ and $\sum_{i=1}^{28} x_i^2 = 79.03$. The mean and standard deviation are respectively

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{35.9}{28} = 1.9944,$$

and

$$s = \sqrt{\frac{(\sum_{i=1}^{28} x_i^2) - (\sum_{i=1}^{28} x_i)^2/n}{n-1}} = \sqrt{\frac{(79.03) - (35.9)^2/18}{18-1}} = 0.6611.$$

(b) We will need to get the order statistics, that is arrange the 18 values from smallest to largest:

0.9	1.1	1.3	1.3	1.5	1.6	1.6	1.8	2.0
2.1	2.1	2.2	2.3	2.4	2.7	2.9	3.0	3.1

The rank of the median is $(n+1)\,50\% = 9.5$. Thus,

$$\text{median} = \tilde{x} = (1-0.5)y_9 + 0.5\,y_{10} = 0.5\,(2.0) + 0.5\,(2.1) = 2.05.$$

The rank of q_1 is $(n+1)\,25\% = 4.75$. Thus,

$$q_1 = (1-0.75)y_4 + 0.75\,y_5 = 0.25\,(1.3) + 0.75\,(1.5) = 1.450.$$

The rank of q_3 is $(n+1)\,75\% = 14.25$. Thus,

$$q_3 = (1-0.25)y_{14} + 0.25\,y_{15} = 0.75\,(2.4) + 0.25\,(2.7) = 2.475.$$

The interquartile range is $\text{IQR} = q_3 - q_1 = 2.475 - 1.450 = 1.025$.

(c) We compute the fences:

$$\text{lower fence} = q_1 - 1.5\,\text{IQR} = 1.450 - 1.5\,(1.025) = -0.0875,$$

$$\text{upper fence} = q_3 + 1.5\,\text{IQR} = 2.475 + 1.5\,(1.025) = 4.0125.$$

All values are within the fences. So there are no outliers.

[3] **Question 2:** Problem 7.10 from the textbook.

Let $y = \ln(c)$, where c is the concentration c at $t = 450$ seconds. Since the intial concentration is $C_0 = 0.3$, then at $t = 45$ seconds, we

$$y = \ln(C_0) - k\,t = \ln(0.3) - 450\,k \quad \Rightarrow \quad k = -(1/450)\,y + \ln(0.3)/450.$$

(a) The geometric mean of the concentration c at $t = 450$ seconds is $0.22 = e^{\bar{y}}$. Thus, $\bar{y} = \ln(0.22)$. This implies that

$$\bar{k} = -\frac{1}{450}\,\bar{y} + \frac{\ln(0.3)}{450} = -\frac{1}{450}\,\ln(0.22) + \frac{\ln(0.3)}{450} = 0.000689.$$

(b) The geometric standard deviation of the concentration c at $t = 450$ seconds is $1.17 = e^{s_y}$. Thus, $s_y = \ln(1.17)$. This implies that

$$s_k = |-(1/450)|\,s_y = \ln(1.17)/450 = 0.0003489.$$

[6] **Question 3:**

(a) Let X be the height of a plant in cm. X has a normal distribution with mean $\mu = 145$ and variance $\sigma^2 = 22^2$. We want

$$\begin{aligned} P(135 < \bar{X} < 155) &= \Phi\left(\frac{155-145}{22}\right) - \Phi\left(\frac{135-145}{22}\right) \\ &= \Phi(0.45) - \Phi(-0.45) = 0.6736 - 0.3264 = 0.3472 = 34.72\%. \end{aligned}$$

(b) Let $\overline{X}$ be the average height of the 16 selected plants in cm. $\bar{X}$ has a normal distribution with mean $\mu = 145$ and variance $\sigma^2/n = 22^2/16$. We want

$$\begin{aligned} P(135 < \bar{X} < 155) &= \Phi\left(\frac{155-145}{22/\sqrt{16}}\right) - \Phi\left(\frac{135-145}{22/\sqrt{16}}\right) \\ &= \Phi(1.82) - \Phi(-1.82) = 0.9656 - 0.0344 = 0.9312. \end{aligned}$$

Question 4: Problem 14.23 (a), (b), and (c) from the textbook.

(a) The mean is $\bar{x} = \sum_{i=1}^{n} x_i/x = 14,971/50 = 299.42$ and the standard deviation is

$$s = \sqrt{\frac{(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2/n}{n-1}} = \sqrt{\frac{(4,486,567) - (14,971)^2/50}{50-1}} = 8.9786.$$

(b) The rank of the median is $(n+1)\,50\% = 25.5$. Thus,

$$\text{median} = \tilde{x} = (1-0.5)\,y_{25} + 0.5\,y_{26} = 0.5(297) + 0.5(298) = 297.5.$$

The rank of q_1 is $(n+1)\,25\% = 12.75\%$. Thus,

$$q_1 = (1-0.75)\,y_{12} + 0.75\,y_{13} = 0.25\,(293) + 0.75\,(293) = 293.$$

The rank of q_2 is $(n+1)\,75\% = 38.25\%$. Thus,

$$q_3 = (1-0.25)\,y_{38} + 0.25\,y_{39} = 0.75\,(307) + 0.25\,(307) = 307.$$

(c) The interquartile range is $\mathrm{IQR} = q_3 - q_1 = 307 - 293 = 14$. The fences are

$$\text{lower fence} = q_1 - 1.5\,\mathrm{IQR} = 293 - 1.5\,(14) = 272,$$

$$\text{upper fence} = q_3 + 1.5\,\mathrm{IQR} = 307 + 1.5\,(14) = 328.$$

All values are within the fences. So there are no outliers.

Part (II) Use `R` for your computations to the following problems. Please attach the `R` commands, output and graphs that you used to answer the question. The `R` output **alone** is not an answer to the question. Please write a sentence or two to properly answer each question.

[4] **Question 5:** Assume that the distribution of the duration of human pregnancies can be approximated with a normal distribution with a mean of 266 days and a standard deviation of 16 days.

Let X be the duration of a human pregnancy in days. X has a normal distribution with $\mu = 266$ and $\sigma = 16$.

(a) We want $P(260 < X < 280) = F_X(280) - F_X(260) = 45.54\%$.

(b) We want x such that $0.1 = P(X > x)$. Thus, $0.9 = P(X \le x)$. Which means $x = F_X^{-1}(0.9) = 286.5048$ days.

(c) We select 500 pregnant women at random. Let N be the number of pregnancies in the sample with a duration between 260 and 280 days. Compute

$$P(200 \le N \le 300) \quad \text{and} \quad P(N = 265).$$

N has a binomial distribution with $n = 500$ and $p = P(260 < X < 280) = 0.4554$. We want

$$P(200 \le N \le 300) = F_N(300) - F_N(199) = 0.9945$$

and $P(N = 265) = 0.00013$.

(d) Let $\bar{X}$ be the average duration of 10 pregnancies in days. $\bar{X}$ has a normal distribution with $\mu = 266$ and standard deviation $\sigma/\sqrt{n} = 16/\sqrt{10}$. We want $P(\bar{X} < 260) = F_{\bar{X}}(260) = 0.11784$.

(e) Let $\bar{X}$ be the average duration of 60 pregnancies in days. $\bar{X}$ has a normal distribution with $\mu = 266$ and standard deviation $\sigma/\sqrt{n} = 16/\sqrt{60}$. We want $P(\bar{X} < 260) = F_{\bar{X}}(260) = 0.00184$.

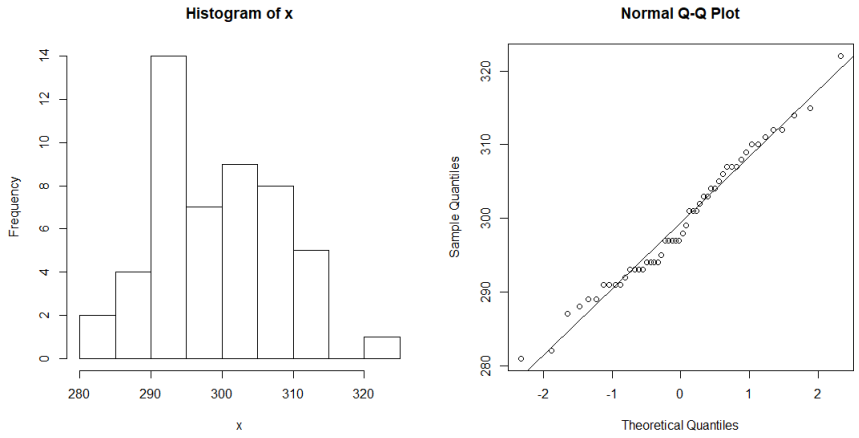
(f) Since n is large in (e), then this probability is probably approximately correct. However, the probability in (d) might not be a good approximation since n is small.

With `R`:

```
> ## (a)
> pnorm(280,266,16)-pnorm(260,266,16)
[1] 0.4553828
> ## (b)
> qnorm(0.9,266,16)
[1] 286.5048
> ## (c)
> pbinom(300,500,.4554)-pbinom(199,500,.4554)
[1] 0.9944893
> dbinom(265,500,.4554)
[1] 0.0001346836
> ## (d)
> pnorm(260,266,16/sqrt(10))
[1] 0.11784
> ## (e)
> pnorm(260,266,16/sqrt(60))
[1] 0.001837806
```

[4] **Question 6:** Problem 14.23 (c) and (d) from the textbook.

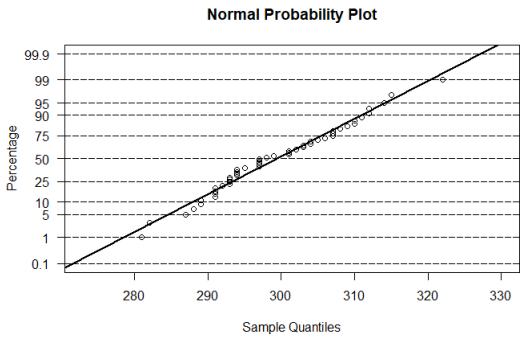
We assigned to the data to the numeric vector `x` with `R`. Using the command `hist(x)` and the commands `qqnorm(x)` and `abline(mean(x),sd(x))`, we produced a histogram and a QQ-plot for the daily water consumption, respectively. Here are the two plots.



(c) The distribution of the water usage is slightly skewed to the right or even describing it as approximately symmetric would be acceptable.

(d) There is a linear tendency in the QQ-plot with slight deviations in the extremes. It is reasonable to assume that the daily consumption of water is normally distributed.

Comment: Producing a normal probability plot instead of the QQ-plot would also be acceptable. Using the command `ppnorm(x)`, where `pporm` is from the file `plots.r`, gives the following normal probability plot.



Marking Scheme:

Question1: (a) 1pt for $\overline{x}$ and 1pt for s. Lose 1/2 pt if they gave s^2 instead of s.
(b) for each of median, q_1 and q_3. 1pt for getting the rank, and 1pt for getting the correct value.
(c) 1/2pt for lower fence; 1/2pt for upper fence. 1 pt for concluding that there are no outliers

Question2: 1 pt for writing k as a linear transformation of $\ln(C) = y$. 1/2pt for $\bar{y} = \ln(0.22)$ and 1/2 pt for $s_y = \ln(1.17)$. 1/2pt for writing $\bar{k}$ as a function of $\bar{y}$ and 1/2pt for writing s_k as a function of s_y.

Question3: (a) 1/2 for defining X and 1/2 for identifying that it has a certain normal distribution. 1pt for properly standardizing, 1pt for final answer;
(b) (a) 1/2 for defining $\bar{X}$ and 1/2 for identify that it has a certain normal distribution. 1pt for properly standardizing, 1pt for final answer;

Question5: (d) 1pt for noticing that $\sigma_{\overline{X}} = \sigma^2/n$. 1pt for the correct final answer. (f) 1 pt for noticing that in (d) n is small, so the approximation might not be good; 1pt for noticing that in (e) n is large, so the approximation most likely will be good.

Question6: 1pt for providing the histogram, 1pt for providing the qq-plot or the normal probability plot. 1 pt for identifying a linear tendency in the qq-plot. 1 pt to conclude that it is reasonable to assume that the water usage is normally distributed.

[/27]