MAT 2379, Introduction to biostatistics

Solution to Assignment 3

Problem 7.4

[2] a) The mean is $\bar{x} = \sum x_i/n = 35.9/18 = 1.99$ and the standard deviation is

$$s = \sqrt{\frac{\sum x_i^2 - (\sum x_i)^2/n}{n-1}} = \sqrt{\frac{79.03 - (35.9)^2/18}{18-1}} = 0.661.$$

[4] b) We arrange the data in increasing order:

$$y_1 = 0.9 \quad y_2 = 1.1 \quad y_3 = 1.3 \quad y_4 = 1.3 \quad y_5 = 1.5 \quad y_6 = 1.6 \quad y_7 = 1.6 \quad y_8 = 1.8 \quad y_9 = 2.0$$

$$y_{10} = 2.1 \quad y_{11} = 2.1 \quad y_{12} = 2.2 \quad y_{13} = 2.3 \quad y_{14} = 2.4 \quad y_{15} = 2.7 \quad y_{16} = 2.9 \quad y_{17} = 3.0 \quad y_{18} = 3.1$$

Since $n = 18$ is odd, the median is

$$\tilde{x} = \frac{y_9 + y_{10}}{2} = \frac{2.0 + 2.1}{2} = 2.05$$

To compute the first quartile, we note that $(n+1)/4 = 4.75$, and hence

$$q_1 = (0.25)y_4 + (0.75)y_5 = (0.25)(1.3) + (0.75)(1.5) = 1.45$$

To compute the third quartile, we note that $3(n+1)/4 = 14.25$, and hence

$$q_3 = (0.75)y_{14} + (0.25)y_{15} = (0.75)(2.4) + (0.25)(2.7) = 2.475$$

Hence $IQR = q_3 - q_1 = 2.475 - 1.45 = 1.025$.

[3] c) To find if there are outliers, we need to compute the two fences:

$$\text{Fence1} = q_1 - (1.5)IQR = 1.45 - (1.5)(1.025) = -0.0875$$

$$\text{Fence2} = q_3 + (1.5)IQR = 2.475 + (1.5)(1.025) = 4.0125$$

There are no values outside the two fences. Hence, there are no outliers.

[4] Problem 7.10

Let $y = \ln C$ be the log-concentration at time $t = 450$, so $y = \ln C_0 - k\,(450)$. Thus,

$$k = a\,y + b, \quad \text{where} \quad a = -\frac{1}{450} \quad \text{and} \quad b = \frac{\ln(C_0)}{450} = \frac{\ln(0.3)}{450}.$$

Therefore, $\bar{k} = a\,\bar{y} + b$ and $s_k = |a|\,s_y$. **(1 point each for expressing $\bar{k}$ and s_k in terms of $\bar{y}$ and s_y).**

The geometric mean concentration at time $t = 450$ seconds of 0.22 mol/L. This means, $g = e^{\bar{y}} = 0.22$, which implies $\bar{y} = \ln(0.22) = -1.514128$.
The geometric standard deviation of the concentration at time $t = 450$ seconds is 1.17. This means, $e^{s_y} = 1.17$, which implies $s_y = \ln(1.17) = 0.1570037$.
(1 point each for getting $\bar{y}$ and s_y from $e^{\bar{y}}$ and e^{s_y}.)

Therefore,

$$\bar{k} = a\,\bar{y} + b = -\frac{1}{450}(-1.514128) + \frac{\ln(0.3)}{450} = 0.000689233$$

and

$$s_k = |a|\,s_y = \frac{1}{450}(0.1570037) = 0.0003489.$$

[6] Problem 7.14

a) Group 2
b) Group 1
c) Group 1 (Note: We generally use the IQR as the preferred measure of dispersion in the boxplot.)
d) Group 2 (Note: The range is the difference between the largest and the smallest value.)
e) Both groups 2 and 3 have a median survival time near 2.25 years.
f) Groups 2 and 3 have boxes of similar sizes. Hence, the values within these two groups are similarly dispersed.

Problem 7.16

[2] a) $\mu_{\bar{X}} = E[\overline{X}] = \mu = 3.2$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 0.17/\sqrt{33} = 0.02959$.

[3] b) By the Central Limit Theorem, $\overline{X} \sim N(3.2, 0.02959)$ approximately. Thus,
$P(\overline{X} > 3.3) = 1 - P(Z < (3.3 - 3.2)/0.02959) \approx 1 - \Phi(3.38) = 0.0004$.

Additional question:

(a) For the mean of Y:

$$\mu_Y = E[Y] = E[2X_1 - 3X_2 + 4X_3] = 2E[X_1] - 3E[X_2] + 4E[X_3] = 4 - 12 + 24 = 16.$$

For the variance of Y:

$$\sigma_Y^2 = V[Y] = V[2X_1 - 3X_2 + 4X_3] = 4V[X_1] + 9V[X_2] + 16V[X_3] = 12 + 45 + 112 = 169.$$

(b) $Y \sim N(\mu_Y = 32, \sigma_Y^2 = 169)$.

(c) $P(Y > 58) = 1 - P(Y \leq 58) = 1 - P\left(\frac{Y-\mu_Y}{\sigma_Y} \leq \frac{58-\mu_Y}{\sigma_Y}\right) = 1 - P\left(Z \leq \frac{58-16}{13}\right) = 1 - \Phi(3.23) = 1 - 0.9994 = 0.0006$.

Part (II)

Problem 1

We assigned the data with `R` to the number vector `x`.
(a) The mean is $\bar{x} = 50.23333$ and the standard deviation is $s = 31.61504$.
With `R`:

```
> mean(x)
[1] 50.23333
> sd(x)
[1] 31.61504
```

(b) The 5-number summary is

$$\text{min}= 10.9,\ q_1 = 19.7,\ \tilde{x} = 47.4,\ q_3 = 85.8,\ \text{max}= 98.8.$$

With `R`:

```
> quantile(x,type=6)
  0%  25%  50%  75% 100%
10.9 19.7 47.4 85.8 98.8
```

[/24]