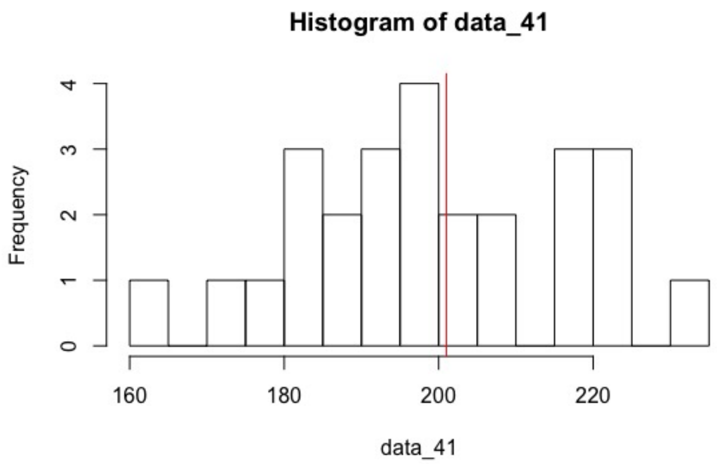
1. Exercise 4.1.2 on Pages 212–213

The weights of 26 professional baseball pitchers are given below, Suppose we assume that the weight of a professional baseball pitcher is normally distributed with mean μ and variance σ^2.

160	175	180	185	185	185	190	190	195	195	195	200	200
200	200	205	205	210	210	218	219	220	222	225	225	232

(a) Obtain a frequency distribution and a histogram or a stem-leaf plot of the data. Use 5-pound intervals. Based on the plot is a normal probability model credible?
Answer(a):



According to the plot, we can tell approximately, it is a normal distribution as it seems like a bell shape.
(b) Obtain the maximum likelihood estimate of μ, σ^2,σ and μ/σ. Locate your estimate of μ on your plot in part(a).
Answer(b):
Define, $\theta = \{\mu, \sigma\}$

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{26} f(x_i) \\ &= \prod_{i=1}^{26} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2} \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^{26}} e^{-\sum_{i=1}^{26}(x_i-\mu)^2/2\sigma^2} \end{aligned} \tag{1}$$

Therefore,

$$
\begin{aligned}
\ell(\theta) &= \log(L(\theta)) \\
&= -13\log(2\pi) - 13\log(\sigma^2) - \sum_{i=1}^{26}(x_i-\mu)^2/2\sigma^2
\end{aligned} \tag{2}
$$

Then, for mle of μ

$$
\begin{aligned}
\frac{\partial \ell(\mu)}{\partial \mu} &= 2\sum_{i=1}^{26}(x_i-\mu)/2\sigma^2 = 0 \\
\Rightarrow \hat{\mu} &= \frac{\sum_{i=1}^{n} x_i}{26} = 201
\end{aligned} \tag{3}
$$

for mle for σ^2, we have

$$
\begin{aligned}
\frac{\partial \ell(\sigma^2)}{\partial \sigma^2} &= -\frac{13}{\sigma^2} + \frac{\sum_{i=1}^{26}(x_i-\mu)^2}{2\sigma^4} = 0 \\
\Rightarrow \hat{\sigma^2} &= \frac{\sum_{i=1}^{26}(x_i-\mu)^2}{26}
\end{aligned} \tag{4}
$$

for mle for σ, we have

$$
\begin{aligned}
\frac{\partial \ell(\sigma)}{\partial \sigma} &= -\frac{13}{\sigma^2} \times 2\sigma + 2 \times \frac{\sum_{i=1}^{26}(x_i-\mu)^2}{2\sigma^3} = 0 \\
\Rightarrow \hat{\sigma} &= \sqrt{\frac{\sum_{i=1}^{26}(x_i-\mu)^2}{26}} = 17.14418
\end{aligned} \tag{5}
$$

The red line is the mean line in (a).

(c) Using the binomial model, obtain the maximum likelihood estimate of the proportion p of professional baseball pitchers who weigh over 215 pounds.

Answer(c):

For binomial distribution, assuming k is the number of players weigh higher than 215 pounds, the mle is

$$
L(p) = Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{6}
$$

Then, we take the derivative:

$$
l(p) = \log(p) = \log(\binom{n}{k}) + (k)\log(p) + (n-k)\log(1-p) \tag{7}
$$

Then,

$$
\frac{\partial l(p)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = 0 \Rightarrow \hat{p} = \frac{k}{n} = \frac{7}{26} \tag{8}
$$

Next,

$$
\begin{aligned}
Pr(X > 215) &= Pr(Z > \frac{215-201}{17.14418}) \\
&= 1 - 0.7929225 \\
&= 0.2070775
\end{aligned} \tag{9}
$$

(d) Determine the mle of p assuming that the weight of a professional baseball player follows the normal probability model $N(\mu, \sigma^2)$ with μ and σ known.

Answer(d):

$$np = \mu$$

Therefore,

$$\hat{p} = \frac{\mu}{n}$$

2. Exercise 4.2.1 on Page 219

Let the observed value of the mean $\bar{X}$ and of the sample variance of a random sample of size 20 from a distribution that is $N(\mu, \sigma^2)$ be 81.2 and 26.5, respectively. Find respectively 90%, 95% and 99% confidence intervals for μ. Note how the lengths of the confidence intervals increases as the confidence increases.
Answer:

$$t_{19,0.95} = 1.729133, \qquad t_{19,0.975} = 2.093024, \qquad t_{19,0.995} = 2.860935$$

Then, for the confidence interval, we have:

$$\begin{aligned}
90\% : \mu : \quad & c(\bar{X} - t_{19,0.95}\frac{S}{\sqrt{n}}, \bar{X} + t_{19,0.95}\frac{S}{\sqrt{n}}) = c(79.20962, 83.19038) \\
95\% : \mu : \quad & c(\bar{X} - t_{19,0.975}\frac{S}{\sqrt{n}}, \bar{X} + t_{19,0.975}\frac{S}{\sqrt{n}}) = c(78.79075, 83.60925) \\
99\% : \mu : \quad & c(\bar{X} - t_{19,0.995}\frac{S}{\sqrt{n}}, \bar{X} + t_{19,0.995}\frac{S}{\sqrt{n}}) = c(77.90682, 84.49318)
\end{aligned} \tag{10}$$

3. Exercise 4.2.5 on Page 220

In Exercise 4.1.2, the weights of 26 professional baseball pitchers were given. From the same data set, the weights of 33 professional baseball hitters (not pitchers) are given below. Assume that the data sets are independent of one another.

155	155	160	160	160	166	170	175	175	175	180	185	185
185	185	185	185	185	190	190	190	190	190	195	195	195
195	200	205	207	210	211	230						

Use expression (4.2.13) to find a 95% confidence interval for the difference in mean weights between the pitchers and the hitters. Which group (on the average) appears to be heavier? Why would this be so? (The sample means and variances for the weights of the pitchers and hitters are, receptively, Pitchers 201, 305.68 and Hitters 185.4,298.13)

Answer:

The confidence interval:

$$\begin{aligned}
&\left((\bar{X}_1-\bar{X}_2)-t_{n-2,0.975}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},(\bar{X}_1-\bar{X}_2)+t_{n-2,0.975}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)\\
&\Bigg(201-185.4-t_{57,0.975}\sqrt{\frac{25\times 305.68+32\times 298.13}{57}}\sqrt{\frac{1}{26}+\frac{1}{33}},\\
&201-185.4+t_{57,0.975}\sqrt{\frac{25\times 305.68+32\times 298.13}{57}}\sqrt{\frac{1}{26}+\frac{1}{33}}\Bigg)\\
&=(6.483065,24.71693)
\end{aligned} \tag{11}$$

where $n = n_1 + n_2$

$$Sp^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \tag{12}$$

4. Exercise 4.2.6 on Page 220

In the baseball data set discussed in the last exercise, it was found that out of 59 baseball players, 15 were left-handed. Is this odd, since the proportion of left-handed males in America is about 11%? Answer by using (4.2.7) to construct a 95% approximate confidence interval for p, the proportion of left-handed baseball players.

Answer:

(4.2.7) gives:

$$\begin{aligned}
&(\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})\\
&=c(0.1431301,0.3653444)
\end{aligned} \tag{13}$$

In this case

$$\hat{p}=\frac{15}{59} \qquad n=59 \qquad z_{\alpha/2}=1.959964$$

5. Exercise 4.2.11 on Page 221

Let $X_1, X_2, ..., X_n, X_{n+1}$ be a random sample of size $n+1$, $n > 1$, from a distribution that is $N(\mu, \sigma^2)$. Let $\bar{X} = \sum_{i=1}^n X_i/n$ and $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$. Find the constant c so that the statistic $c(\bar{X} - X_{n+1})/S$ has a t-distribution. If $n = 8$, determine k such that $P(\bar{X} - kS < X_9 < \bar{X} + kS) = 0.8$. The observed interval $(\bar{x} - ks, (\bar{x}) + ks)$ is often called an 80% prediction interval for X_9.

Answer:

Recall,

$$t=\frac{Z}{\sqrt{\chi^2(v-1)/(v-1)}}\sim t(d.f.=v-1)$$

and

$$X_{n+1} \sim N(\mu, \sigma^2)$$

and

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Then, the new R.V

$$X_{n+1} - \bar{X} \sim N(0, \sigma^2(1 + \frac{1}{n}))$$

$$\frac{X_{n+1} - \bar{X} - 0}{S} = \frac{X_{n+1} - \bar{X} - 0}{\sigma\sqrt{1+1/n}} \times \frac{\sqrt{1+1/n}}{S/\sigma} = \frac{X_{n+1} - \bar{X} - 0}{\sigma\sqrt{1+1/n}} \times \frac{\sqrt{1+1/n}}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}}$$

Which is

$$1/\sqrt{1+1/n} \times \frac{X_{n+1} - \bar{X} - 0}{S} = \frac{\frac{X_{n+1}-\bar{X}}{\sigma\sqrt{1+1/n}}}{\sqrt{\chi^2(n-1)/(n-1)}} = \frac{Z}{\sqrt{\chi^2(n-1)/(n-1)}} \sim t(n-1)$$

Thus, we know

$$c = 1/\sqrt{1 + \frac{1}{n}}$$

and, this is a t distribution, with

$$t \sim t(d.f. = n-1)$$

Next, from above conclusion, we have

k is

$$t(8-1, 0.90) * c = 1.414924 * \sqrt{9/8} = 1.500754$$

6. Exercise 4.2.27 on Pages 222–223

Let $X_1, X_2, ..., X_n$ and $Y_1, ..., Y_m$ be two independent random samples from the respective normal distribution $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, where the four parameters are unknown. To construct a confidence interval for ratio σ_1^2/σ_2^2, of the variances, form the quotient of the two independent χ^2 variables, each divided by its degree of freedom, namely,

$$F = \left(\frac{(m-1)S_2^2}{\sigma_2^2}/(m-1)\right) / \left(\frac{(n-1)S_1^2}{\sigma_1^2}/(n-1)\right) \tag{14}$$

where S_1^2 and S_2^2 are the respective sample variances.

(a) What kind of distribution does F have?

Answer(a):

As

$$\frac{(m-1)S_2^2}{\sigma_2^2}/(m-1) \sim \chi^2(m-1)$$

,

$$\frac{(n-1)S_1^2}{\sigma_1^2}/(n-1) \sim \chi^2(n-1)$$

Then,

$$F \sim F(m-1, n-1)$$

(b) From the appropriate table, a and b can be found so that $P(F < b) = 0.975$ and $P(a < F < b) = 0.95$

Answer(b):

$$b = F^*_{m-1,n-1,0.975}$$

and

$$a = F^*_{m-1,n-1,0.025}$$

(c) Rewrite the second probability statement as

$$P(a\frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < b\frac{S_1^2}{S_2^2}) = 0.95$$

The observed values, s_1^2 and s_2^2 can be inserted in these inequalities to provide a 95% confidence interval for σ_1^2/σ_2^2

Answer(c):

As this will be the same with (b) if we consider that

$$F = \frac{\sigma_1^2}{\sigma_2^2}\frac{S_2^2}{S_1^2}$$

Therefore, the same answer with (b)

$$b = F^*_{m-1,n-1,0.975} \qquad a = F^*_{m-1,n-1,0.025}$$

7. Exercise 4.4.3 on Page 236

Consider the sample of data:

13	5	202	15	99	4	67	83	36	11	301
23	213	40	66	106	78	69	166	84	64	

(a) Obtain the five-number summary of these data

Answer(a):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.0	23.0	67.0	83.1	99.0	301.0

(b) Determine if there are any outliers.

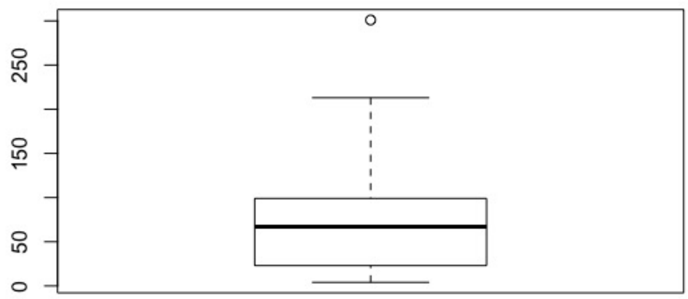
Answer(b):

$$c(q_1 - 1.5(q_3 - q_1), q_1 - 1.5(q_3 - q_1)) = c(-91, 213)$$

Therefore, 301 is the outlier.

(c) Boxplot the data. Comment on the plot.

Answer(c):



Through the boxplot, we can see the distribution is skewed to the right and one outlier. The median is rather low in the range.

8. Exercise 4.4.8 on Page 236

Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ denote the order statistics of a random sample of size 5 from a distribution having p.d.f. $f(x) = e^{-x}$,$0 < x < \infty$, zero elsewhere. Show that $Z_1 = Y_2$ and $Z_2 = Y_4 - Y_2$ are independent.

Hint: First find the joint p.d.f. of Y_2 and Y_4.

Answer: The p.d.f. for k-th largest number

$$f_X k; n(x) = \frac{n!}{(k-1)!(n-k)!}[F_X(x)]^{k-1}[1 - F_X(x)]^{n-k} f_X(x) \tag{15}$$

And the joint distribution of j-th largest and k-th largest distribution is

$$f_{X_{j;n}, X_{k;n}}(x, y) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!}[F_X(x)]^{j-1}[F_X(y) - F_X(x)]^{k-1-j}[1 - F_X(y)]^{n-k} f_X(x) f_X(y) \tag{16}$$

Therefore, using above formula, we can calculate Y_2, Y_4's joint distribution:

$$\begin{aligned} f_{Y_2,Y_4}(x,y) &= \frac{5!}{1!1!1!}(1-e^{-x})^1(e^{-x}-e^{-y})^1(e^{-y})^1e^{-x-y} \\ &= 120\big(e^{-2x-2y}-e^{-x-3y}-e^{-3x-2y}+e^{-2x-3y}\big) \end{aligned} \tag{17}$$

The corresponding transformation and Jacobian is:

$$\begin{gathered} Y_2 = Z_1 \qquad Y_4 = Z_1 + Z_2 \\ J = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = 1 \end{gathered} \tag{18}$$

$$\begin{aligned} f_{Z_1,Z_2}(z_1,z_2) &= 120[e^{-4z_1-2z_2}-e^{-4z_1-3z_2}-e^{-5z_1-2z_2}+e^{-5z_1-3z_2}] \\ f_{Z_2}(z_2) &= \int f_{Z_1,Z_2}(z_1,z_2)dz_1 \\ &= 120(e^{-2z_2}/4-e^{-3z_2}/4-e^{-2z_2}/5+e^{-3z_2}/5), \qquad z_2 \in c(0,\infty) \\ &= 120(e^{-2z_2}/20-e^{-3z_2}/20) \\ f_{Z_1}(z_1) &= 120[e^{-4z_1}/2-e^{-4z_1}/3-e^{-5z_1}/2+e^{-5z_1}/3] \\ &= 120[e^{-4z_1}/6-e^{-5z_1}/6], \qquad z_1 \in c(0,\infty) \end{aligned} \tag{19}$$

Therefore

$$f_{Z_1,Z_2}(z_1,z_2) = f_{Z_2}(z_2)f_{Z_1}(z_1)$$

9. Exercise 4.4.24 on Page 238

Let Y_n denote the n-th order statistic of a random sample of size n from a distribution of the continuous type. Find the smallest value of n for which the inequality $P(\xi_{0.9} < Y_n) \geq 0.75$ is true.

Answer:

$$p = 0.9$$

as given

$$\begin{aligned} P(Y_0 \leq \xi_{0.9} < Y_n) = \sum_{w=0}^{n-1}\binom{n}{w}p^w(1-p)^{n-w} &= 1-0.9^n > 0.75 \\ \Rightarrow \qquad n\log(0.9) &< \log(0.25) \\ n &> 13.15763 \end{aligned} \tag{20}$$