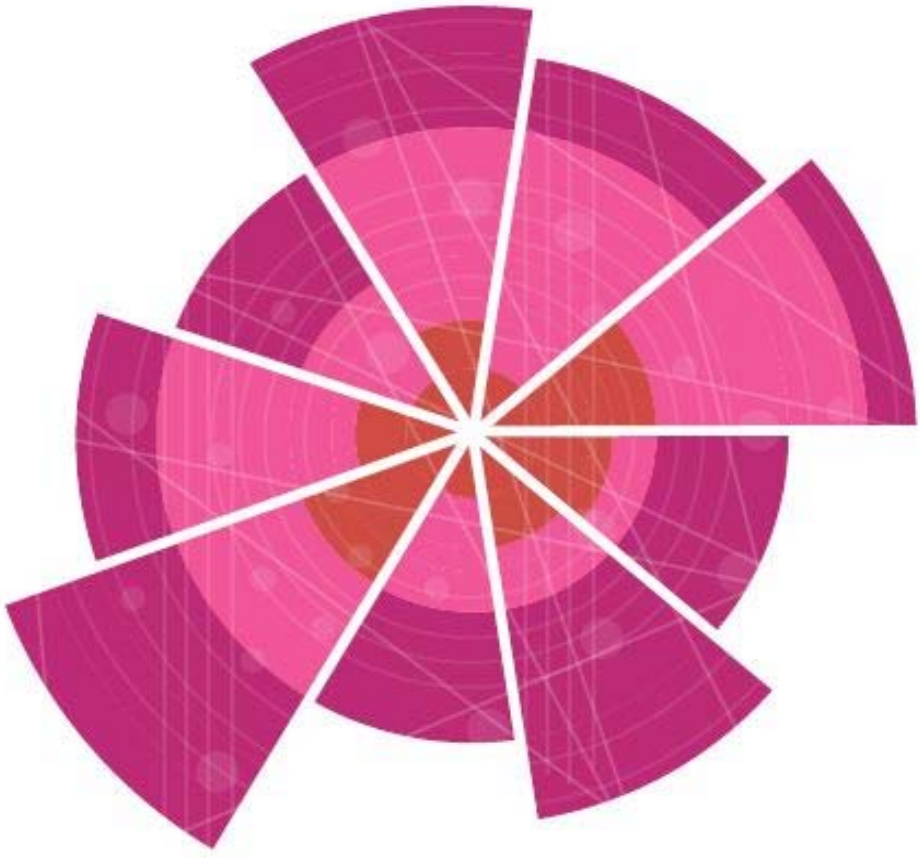
STAT 221/231
Problem Solutions
Winter 2024 Edition

Department of Statistics and Actuarial Science
University of Waterloo



STAT 221/231 PROBLEM SOLUTIONS

Department of Statistics and Actuarial Science, University of Waterloo

Winter 2024 Edition

Contents

SOLUTIONS TO CHAPTER 1 PROBLEMS

1.1 (a)

$$\sum_{i=1}^{n}(y_i-\bar{y})=\sum_{i=1}^{n}y_i-\bar{y}\left(\sum_{i=1}^{n}1\right)=\sum_{i=1}^{n}y_i-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)(n)=\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}y_i=0$$

(b) The sample mean of the transformed data set is

$$\bar{u}=\frac{1}{n}\sum_{i=1}^{n}(a+by_i)=\frac{1}{n}\left(na+b\sum_{i=1}^{n}y_i\right)=a+b\bar{y}$$

and the sample median of the transformed data set is

$$\hat{m}_u=a+b\hat{m}$$

(c) There is no general result for the sample mean but if all $y_i \geq 0$ and n is an odd number then the new sample median is $(\hat{m})^2$.

(d) The sample mean of the augmented data set is

$$\bar{y}_a(y_0)=\frac{1}{n+1}\left(y_0+\sum_{i=1}^{n}y_i\right)=\frac{n\bar{y}+y_0}{n+1}$$

which depends on y_0. Since

$$\lim_{y_0\to\infty}\frac{n\bar{y}+y_0}{n+1}=\infty$$

and

$$\lim_{y_0\to-\infty}\frac{n\bar{y}+y_0}{n+1}=-\infty$$

this means that an additional very large positive observation or very large negative observation has a large effect on the sample mean.

(e) **Case 1**: If n is odd then the sample median of the original data set is

$$\hat{m}=\hat{m}(y_1,y_2,\ldots,y_n)=y_{\left(\frac{n+1}{2}\right)}$$

The augmented data set will have an even number of observations.

If $y_0 \geq y_{\left(\frac{n+1}{2}+1\right)}$ then the sample median of the augmented data set is

$$\hat{m}_a = \frac{1}{2}\left[y_{\left(\frac{n+1}{2}\right)} + y_{\left(\frac{n+1}{2}+1\right)}\right]$$

which does not depend on the value of y_0. If $y_{\left(\frac{n+1}{2}\right)}$ and $y_{\left(\frac{n+1}{2}+1\right)}$ are close in value then the sample median will change by very little.

If $y_0 \leq y_{\left(\frac{n+1}{2}-1\right)}$ then the sample median of the augmented data set is

$$\hat{m}_a = \frac{1}{2}\left[y_{\left(\frac{n+1}{2}-1\right)} + y_{\left(\frac{n+1}{2}\right)}\right]$$

which does not depend on the value of y_0. If $y_{\left(\frac{n+1}{2}-1\right)}$ and $y_{\left(\frac{n+1}{2}\right)}$ are close in value then the sample median will change by very little.

If $y_{\left(\frac{n+1}{2}-1\right)} < y_0 < y_{\left(\frac{n+1}{2}+1\right)}$ then the sample median of the augmented data set is

$$\hat{m}_a = \frac{1}{2}\left[y_0 + y_{\left(\frac{n+1}{2}\right)}\right]$$

which does depend on the value of y_0. If y_0 and $y_{\left(\frac{n+1}{2}\right)}$ are close in value then the sample median will change by very little.

Case 2: If n is even then sample median of the original data set is

$$\hat{m} = \hat{m}\left(y_1, y_2, \ldots, y_n\right) = \frac{1}{2}\left[y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}+1\right)}\right]$$

The augmented data set will have an odd number of observations.

If $y_0 \geq y_{\left(\frac{n}{2}+1\right)}$ then the sample median of the augmented data set is

$$\hat{m}_a = y_{\left(\frac{n}{2}+1\right)}$$

which does not depend on the value of y_0. If $y_{\left(\frac{n}{2}\right)}$ and $y_{\left(\frac{n}{2}+1\right)}$ are close in value then the sample median will change by very little.

If $y_0 \leq y_{\left(\frac{n}{2}\right)}$ then the sample median of the augmented data set is

$$\hat{m}_a = y_{\left(\frac{n}{2}\right)}$$

which does not depend on the value of y_0. If $y_{\left(\frac{n}{2}\right)}$ and $y_{\left(\frac{n}{2}+1\right)}$ are close in value then the sample median will change by very little.

If $y_{\left(\frac{n}{2}\right)} < y_0 < y_{\left(\frac{n}{2}+1\right)}$ then the sample median of the augmented data set is

$$\hat{m}_a = y_0$$

which does depend on the value of y_0. If $y_{\left(\frac{n}{2}\right)}$ and $y_{\left(\frac{n}{2}+1\right)}$ are close in value then the sample median will change by very little.

(f) Unlike the sample mean, the sample median is not affected by outliers (very large positive y_0 or very large negative y_0) so it is a more robust numerical summary of location. For example, in many countries there are usually a few people with very large incomes. The mean income is affected by these few very large incomes so reporting the mean income rather than the median income would give the false impression that people are doing well in general with respect to income.

(g)
$$\frac{d}{d\mu}V(\mu) = -2\sum_{i=1}^{n}(y_i - \mu) = -2n(\bar{y} - \mu) = 0 \text{ if } \mu = \bar{y}$$
and by the First Derivative Test, $V(\mu)$ is minimized at $\mu = \bar{y}$.

1.2 (a)
$$\begin{aligned}
S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y}) = \sum_{i=1}^{n} y_i(y_i - \bar{y}) - \bar{y}\sum_{i=1}^{n}(y_i - \bar{y}) \\
&= \sum_{i=1}^{n} y_i(y_i - \bar{y}) - 0 \text{ since } \sum_{i=1}^{n}(y_i - \bar{y}) = 0 \\
&= \sum_{i=1}^{n} y_i(y_i - \bar{y}) = \sum_{i=1}^{n} y_i^2 - \bar{y}\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\sum_{i=1}^{n} y_i \\
&= \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \\
&= \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 \qquad (1.1)
\end{aligned}$$

(b) Let s_u^2 be the sample variance of the transformed data set $\{u_1, u_2, \ldots, u_n\}$. Then
$$\begin{aligned}
s_u^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})^2 = \frac{1}{n-1}\sum_{i=1}^{n}[a + by_i - (a + b\bar{y})]^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(by_i - b\bar{y})^2 = \frac{b^2}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= b^2 s^2
\end{aligned}$$

The sample standard deviation of the transformed data set is
$$s_u = |b|\, s$$

The IQR of the transformed data set is
$$IQR(u_1, u_2, \ldots, u_n) = |b|\, IQR(y_1, y_2, \ldots, y_n)$$

The range of the transformed data set is
$$range(u_1, u_2, \ldots, u_n) = |b|\left(y_{(n)} - y_{(1)}\right)$$

Note that $|b|$ is necessary since b can be a negative number and all the summaries of variability are positive.

(c) We first note that if we rearrange

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\left(\bar{y}\right)^2\right]$$

we obtain

$$\sum_{i=1}^{n} y_i^2 = (n-1)\, s^2 + n\left(\bar{y}\right)^2 \tag{1.2}$$

Therefore the sample variance of the augmented data set $\{y_0, y_1, y_2, \ldots, y_n\}$ is given by

$$\begin{aligned}
s_a^2\left(y_0\right) &= \frac{1}{n}\left\{\sum_{i=0}^{n} y_i^2 - (n+1)\left[\bar{y}_a\left(y_0\right)\right]^2\right\} \quad \text{using (1.1)} \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} y_i^2 + y_0^2 - (n+1)\frac{\left(n\bar{y}+y_0\right)^2}{(n+1)^2}\right] \quad \text{since } \bar{y}_a\left(y_0\right) = \frac{n\bar{y}+y_0}{n+1} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 + \frac{y_0^2}{n} - \frac{n^2\left(\bar{y}\right)^2 + 2n\bar{y}y_0 + y_0^2}{n\left(n+1\right)} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 + \frac{(n+1)\, y_0^2 - n^2\left(\bar{y}\right)^2 - 2n\bar{y}y_0 - y_0^2}{n\left(n+1\right)} \\
&= \frac{1}{n}\left[(n-1)\, s^2 + n\left(\bar{y}\right)^2\right] + \frac{ny_0^2 - n^2\left(\bar{y}\right)^2 - 2n\bar{y}y_0}{n\left(n+1\right)} \quad \text{using (1.2)} \\
&= \frac{(n-1)\, s^2}{n} + \left(\bar{y}\right)^2 + \frac{y_0^2 - n\left(\bar{y}\right)^2 - 2\bar{y}y_0}{(n+1)} \\
&= \frac{(n-1)\, s^2}{n} + \frac{(n+1)\left(\bar{y}\right)^2 - n\left(\bar{y}\right)^2}{(n+1)} + \frac{y_0^2 - 2\bar{y}y_0}{(n+1)} \\
&= \frac{(n-1)\, s^2}{n} + \frac{\left(\bar{y}\right)^2}{(n+1)} + \frac{y_0\left(y_0 - 2\bar{y}\right)}{(n+1)}
\end{aligned}$$

Therefore

$$\begin{aligned}
\lim_{y_0 \to \pm\infty} s_a\left(y_0\right) &= \lim_{y_0 \to \pm\infty}\left[\frac{(n-1)\, s^2}{n} + \frac{\left(\bar{y}\right)^2}{(n+1)} + \frac{y_0\left(y_0 - 2\bar{y}\right)}{(n+1)}\right]^{1/2} \\
&= \left[\frac{(n-1)\, s^2}{n} + \frac{\left(\bar{y}\right)^2}{(n+1)} + \frac{1}{(n+1)}\lim_{y_0 \to \pm\infty} y_0\left(y_0 - 2\bar{y}\right)\right]^{1/2} \\
&= \infty
\end{aligned}$$

This means that an additional very large positive observation or very large negative observation has a large effect on the sample standard deviation. The sample standard deviation is not a robust measure of variability.

(d) Suppose the size of the data set n is reasonably large. Using an argument similar to the argument given in Problem 1 (e) for the sample median, it can be shown that y_0 has a small effect on the IQR of the augmented data set. In particular,

if $y_0 \geq q(0.75)$ or $y_0 \leq q(0.25)$, then y_0 has a small effect on the IQR of the augmented data set. Therefore as $|y_0|$ increases there will be little effect on the IQR of the augmented data set. The IQR is not affected by outliers and is a robust measure of variability.

(e) If $y_{(1)} \leq y_0 \leq y_{(n)}$ then the range of the augmented data set is the same as the original data set.

If $y_0 \leq y_{(1)}$ then the range of the augmented data set is $y_{(n)} - y_0$.

If $y_0 \geq y_{(n)}$ then the range of the augmented data set is $y_0 - y_{(1)}$.

In both cases the range increases as $|y_0|$ increases. The range is not a robust measure of variability.

1.3 Since

$$\begin{aligned}
\frac{\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})^2\right]^{3/2}} &= \frac{\frac{1}{n}\sum_{i=1}^{n}(by_i-b\bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(by_i-b\bar{y})^2\right]^{3/2}} \\
&= \frac{b^3}{(b^2)^{3/2}}\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{3/2}} \\
&= \left[\frac{b}{|b|}\right]^3 g_1
\end{aligned}$$

Therefore $g_1(u_1, u_2, \ldots, u_n) = g_1$ if $b > 0$ and $g_1(u_1, u_2, \ldots, u_n) = -g_1$ if $b < 0$. In summary the magnitude of the sample skewness remains unchanged but the sample skewness changes sign if $b < 0$.

Since

$$\begin{aligned}
\frac{\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})^2\right]^{2}} &= \frac{\frac{1}{n}\sum_{i=1}^{n}(by_i-b\bar{y})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(by_i-b\bar{y})^2\right]^{2}} \\
&= \frac{b^4}{(b^2)^{2}}\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{2}} = g_2
\end{aligned}$$

therefore the sample kurtosis is the same for both data sets.

1.4 For the revenues:
sample mean $= (-7)(2500) + 1000 = -16500$
sample standard deviation $= |-7|(5500) = 38500$

sample median $= (-7)(2600) + 1000 = -17200$
sample skewness $= (-1)(1.2) = -1.2$
sample kurtosis $= 3.9$
range $= (7)(7500) = 52500$

1.5 (a) The relative frequency histogram of the piston diameters is given in Figure 1.1.

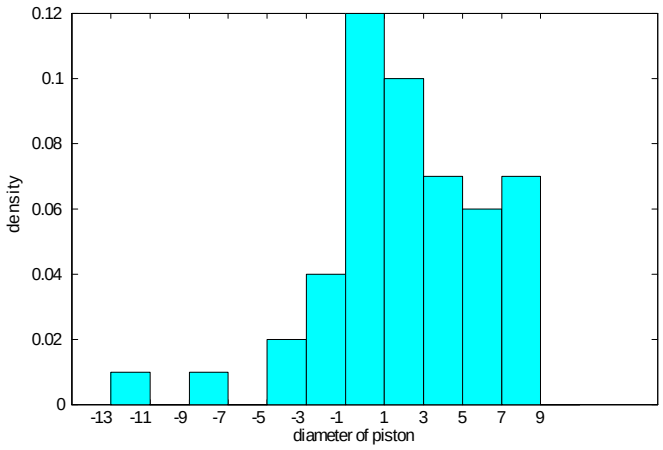


Figure 1.1: Histogram of Piston Diameters

(b) $\bar{y} = 100.7/50 = 2.014$, $\hat{m} = q(0.5) = \frac{1}{2}\left(y_{(25)} + y_{(26)}\right) = \frac{1}{2}(2.1 + 2.5) = 2.3$

(c) $s^2 = \frac{1}{49}\left[1110.79 - 50(2.014)^2\right] = 18.5302, \quad s = 4.3047$
$q(0.25) = \frac{1}{2}\left(y_{(12)} + y_{(13)}\right) = \frac{1}{2}[-0.7 + (-0.6)] = -0.65$
$q(0.75) = \frac{1}{2}\left(y_{(38)} + y_{(39)}\right) = \frac{1}{2}[5.1 + 5.4] = 5.25$
$IQR = 5.25 - (-0.65) = 5.9$

(d) The five number summary is: -12.8, -0.65, 2.3, 5.25, 8.9

(e) $Ppk = 0.6184$

(f) If $\bar{y} \approx \pm 10$ then $Ppk \approx 0$. Values of $\bar{y}$ less than -10 or bigger than $+10$ indicate that performance is poor. If $\bar{y} \approx 0$ then $Ppk \approx 10/3s$. Recall that for Normal data we would expect approximately 99% of the observed data to lie between $\mu - 3\sigma \approx \bar{y} - 3s$ and $\mu + 3\sigma \approx \bar{y} + 3s$. Therefore if $\bar{y} \approx 0$ and $3s \approx 10$ or $10/3s \approx 1$ then this indicates that performance is good. Therefore $Ppk \approx 10/3s = 1$ indicates good performance.

(g) Let $Y \sim G(2.014, 4.3047)$ then

$$\begin{aligned} P(\text{diameters out of specification}) &= 1 - P(-10 < Y < 10) \\ &= 0.03408 \quad \text{using the } N(0,1) \text{ table} \\ &= 0.034412 \quad \text{using R} \end{aligned}$$

1.7 The data from smallest to largest are:
1.1 3.9 4.3 4.5 5.2 6.3 7.2 7.6 8.5 14.0

$$\begin{aligned} q(0.25) &= \frac{1}{2}\left(y_{(2)} + y_{(3)}\right) = \frac{1}{2}(3.9 + 4.3) = 4.1 \\ q(0.5) &= \frac{1}{2}\left(y_{(5)} + y_{(6)}\right) = \frac{1}{2}(5.2 + 6.3) = 5.75 \\ q(0.75) &= \frac{1}{2}\left(y_{(8)} + y_{(9)}\right) = \frac{1}{2}(7.6 + 8.5) = 8.05 \\ IQR &= q(0.75) - q(0.25) = 8.05 - 4.1 = 3.95 \\ q(0.25) - 1.5 \times IQR &= 4.1 - 1.5(3.95) = -1.825 \\ q(0.75) + 1.5 \times IQR &= 8.05 + 1.5(3.95) = 13.975 \end{aligned}$$

The top of the box is at 8.05, the bottom is at 4.1 and the line inside the box is at 5.75. The upper whisker is at 8.5 and the lower whisker is at 1.1. There is one outlier at 14.0.

The empirical cumulative distribution function is a step function which jumps a height of 0.1 at each of the points: 1.1 3.9 4.3 4.5 5.2 6.3 7.2 7.6 8.5 14.0

1.8 (a)

	$y_{(1)}$	$q(0.25)$	$q(0.5)$	$q(0.75)$	$y_{(n)}$	IQR	range
Dataset 1	0.1	1.3	2.8	5.5	13.6	4.2	13.5
Dataset 2	0.2	2.1	4.8	6.7	9.5	4.6	9.3

(b) Since Dataset 2 has more jumps of a smaller size than Dataset 1, therefore Dataset 2 has more observations.

(c) From the shapes of the empirical cumulative distribution functions we can see that the relative frequency histogram of Dataset 1 would not be symmetric but would have a long right tail and the relative frequency histogram of Dataset 2 would be reasonably symmetric.

(d) Based on (c), the sample skewness for Dataset 1 would be positive and the sample skewness for Dataset 2 would be close to zero.

(e) Seventy-five percent of the observations for Dataset 1 are in the interval $[0.1, 5.5]$ while 75 percent of the observations for Dataset 2 are in the interval $[0.2, 6.7]$ These intervals are reasonably similar in width. Dataset 1 which has a long right tail has 25 percent of its observations in the interval $[5.5, 13.6]$ while Dataset 2 has 25 percent of its observations in the interval $[6.7, 9.5]$ which is a much narrower interval. This information indicates that Dataset 1 will have a larger sample standard deviation than Dataset 2.

Alternately we could note that the shape of the empirical cumulative distribution function for Dataset 1 is similar to the shape of the cumulative distribution function of an Exponential(θ) random variable which has standard deviation θ. The median of an Exponential(θ) random variable is $m = \theta \log(2)$ which can be rearranged as $\theta = m/\log(2)$. Since the sample median for Dataset 1 is $\hat{m} = 2.8$

then an estimate of the standard deviation θ would be $2.8/\log(2) = 4.04$.

The shape of the empirical cumulative distribution function for Dataset 2 is similar to the shape of the cumulative distribution function of a Uniform(a, b) random variable which has standard deviation $\sqrt{(b-a)^2/12}$. Since the range for Dataset 2 is 9.3 then an estimate of the sample standard deviation for these data would be $\sqrt{(9.3)^2/12} = 2.7$ which also indicates the sample standard deviation of Dataset 1 would be larger than the sample standard deviation of Dataset 2.

(f) $\hat{F}_1(9)$ is approximately equal to 0.83 and $\hat{F}_2(3)$ is approximately equal to 0.35.

1.9 (a)

	$y_{(1)}$	$q(0.25)$	$q(0.5)$	$q(0.75)$	$y_{(n)}$	IQR	range
Dataset 1	0.0	1.4	3.0	4.5	6.0	3.1	6.0
Dataset 2	0.7	2.1	2.9	3.7	5.8	1.6	5.1
Dataset 3	0.0	0.3	0.8	1.4	5.9	1.1	5.9

(b) From the shapes of the boxplots we can see that the relative frequency histograms of Dataset 1 and Datatset 2 would both be reasonably symmetric, and the relative frequency histogram of Dataset 3 would not be symmetric but would have a long right tail.

(c) Based on (b), the sample skewness for Dataset 1 and Dataset 2 would be close to zero and the sample skewness for Dataset 3 would be positive.

(d) The shape of the boxplot for Dataset 3 indicates that its relative frequency histogram would not be symmetric but would have a long right tail. Therefore the relative frequency histogram of Dataset 3 would not be bell-shaped or uniform. The sample skewness for these data would be positive.

For Dataset 1 the center line in the box, which corresponds to the sample median, divides both the box and the whiskers approximately in half which indicates that the relative frequency histogram would be reasonably symmetric. The shape of the boxplot for Dataset 1 also indicates that approximately 25 percent of the observations lie in 4 intervals of approximately equal width which would suggest that the shape of the relative frequency histogram would be reasonably uniform.

For Dataset 2 the center line in the box, which corresponds to the sample median, divides both the box and the whiskers approximately in half which indicates that the relative frequency histogram would be reasonably symmetric. The distance from the sample median to the whiskers is approximately 2.5 times the distance from the sample median to the edge of the box which indicates that the relative frequency histogram for Dataset 2 would be reasonably bell-shaped.

Therefore the relative frequency histogram for Dataset 2 would look most bell-shaped while the relative frequency histogram for Dataset 1 would look more uniform.

(e) The sample kurtosis of Dataset 1 would be less than 3 because the distribution of the dataset is reasonably uniform.

(f) The sample kurtosis of Dataset 2 would be approximately equal to 3 because the distribution of the dataset is reasonably bell-shaped.

(g) For a data set which is reasonably symmetric, the sample median and sample mean will be approximately equal. Therefore we can compare the variability in Dataset 1 and Dataset 2 by looking at the distribution of the points about the sample median. Since the width of the box and the distance between whiskers for Dataset 1 are larger than for Dataset 2, the sample standard deviation for Dataset 1 will be larger than the sample deviation for Dataset 2.

Alternately, the shape of the boxplot for Dataset 1 suggested a Uniform distribution. Since the range for Dataset 1 is 6.0 then an estimate of the sample standard deviation would be $\sqrt{(6)^2/12} \approx 1.7$. The shape of the boxplot for Dataset 2 suggested a bell-shaped or Gaussian distribution. For Gaussian data we expect the IQR to be approximately equal to 1.35σ. Since the IQR for Dataset 2 is 1.6 then an estimate of the standard deviation would be $1.6/1.35 \approx 1.2$ and therefore the sample standard deviation for Dataset 1 will be larger than the sample deviation for Dataset 2.

1.11 (a)

Sex	$y_{(1)}$	$q\,(0.25)$	$q\,(0.5)$	$q\,(0.75)$	IQR	range
Female	71.00	85.75	89.75	93.12	7.37	31.5
Male	78.00	87.50	92.00	96.00	8.5	27.00

Sex	$\bar{y}$	s	g_1	g_2
Female	89.24	6.548	-0.444	3.730
Male	92.06	6.696	-0.088	2.434

(b) The relative frequency histograms are given in Figures 1.2 and 1.3.

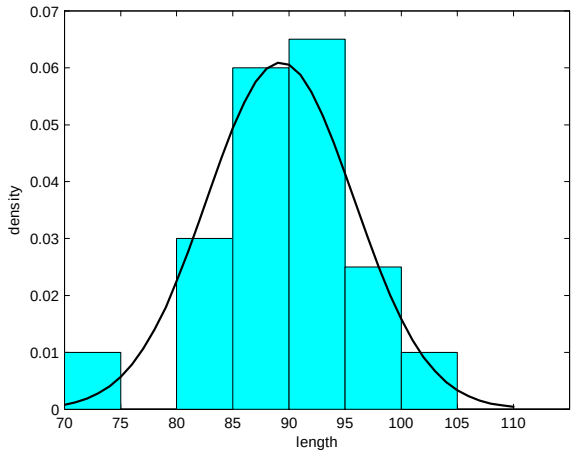


Figure 1.2: Relative Frequency Histogram for Lengths of Female Coyotes

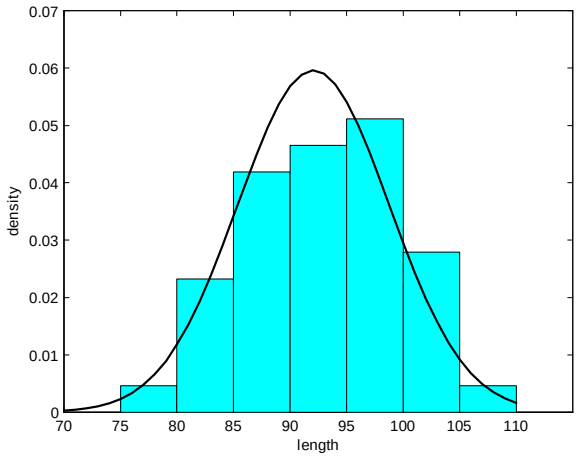


Figure 1.3: Relative Frequency Histogram for Lengths of Male Coyotes

(c) The dataset for female lengths consists of 40 observations. The sample mean and sample median are close in value, the sample skewness is negative but close to zero, and the sample kurtosis is slightly larger than 3. The IQR is equal to 7.37 which is reasonably close to $1.35s = 1.35\,(6.548) = 8.84$. These results are what we would expect for a Gaussian data set of this reasonably small sample size. In Figure 1.2 the superimposed Gaussian probability density function fits the data reasonably well for a data set of size 40. Based on the comparisons of the observed and expected summaries and the fact that this is a reasonably small data set, we would conclude that the Gaussian model fits these data reasonably well.

The dataset for male lengths consists of 43 observations. The sample mean and sample median are close in value, the sample skewness is negative but very close to zero, and the sample kurtosis is slightly smaller than 3. The IQR is equal to 8.5 which is reasonably close to $1.35s = 1.35\,(6.696) = 9.04$. These results are what we would expect for a Gaussian data set of this reasonably small sample size. In Figure 1.3 the superimposed Gaussian probability density function fits the data reasonably well for a data set of size 43. Based on the comparisons of the observed and expected summaries and the fact that this is a reasonably small data set, we would conclude that the Gaussian model fits these data reasonably well.

(d) The boxplots are shown in Figure 1.4.The shape of each boxplot suggests that the distribution of the corresponding data set is reasonably symmetric and similar to a boxplot that we would see for Gaussian data. For the data set of female lengths we notice two outliers, that is, observations which are much smaller than the other observations in the data set. The data set for male lengths has no

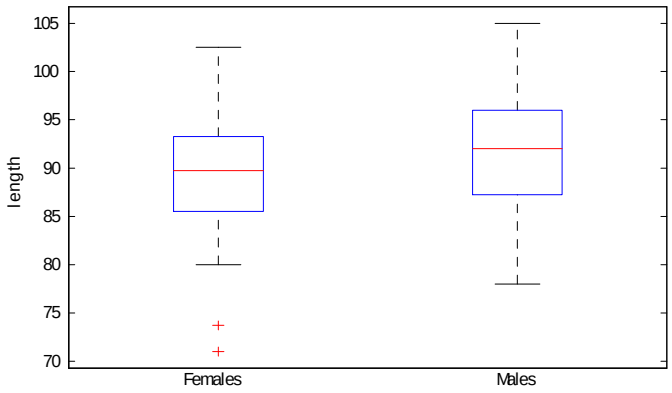


Figure 1.4: Boxplots for lengths of female and male coyotes

outliers. The center of the box for male lengths is taller than the center of the box for female lengths which is consist with the sample median for male lengths being larger than the sample median for female lengths. In general the boxplots look similar in shape with the boxplots for males being shifted up which is consistent with male coyotes generally being larger than female coyotes.

(e) We note that the empirical cumulative distribution function for the male coyotes is to the right of the empirical cumulative distribution function for the female coyotes which would indicate that male lengths of coyotes are generally larger than females as you might expect. We also notice that the shapes of both empirical cumulative distribution functions are similar in shape and the shape is similar to what we would see for Gaussian data.

1.12 The identity

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2$$

follows from Problem 2(a) by replacing the y_i's with the x_i's.

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i (y_i - \bar{y}) - \bar{x} \sum_{i=1}^{n} (y_i - \bar{y}) = \sum_{i=1}^{n} x_i (y_i - \bar{y}) - 0 \\ &= \sum_{i=1}^{n} x_i (y_i - \bar{y}) \\ S_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x}) y_i - \bar{y} \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x}) y_i - 0 \\ &= \sum_{i=1}^{n} (x_i - \bar{x}) y_i \end{aligned}$$

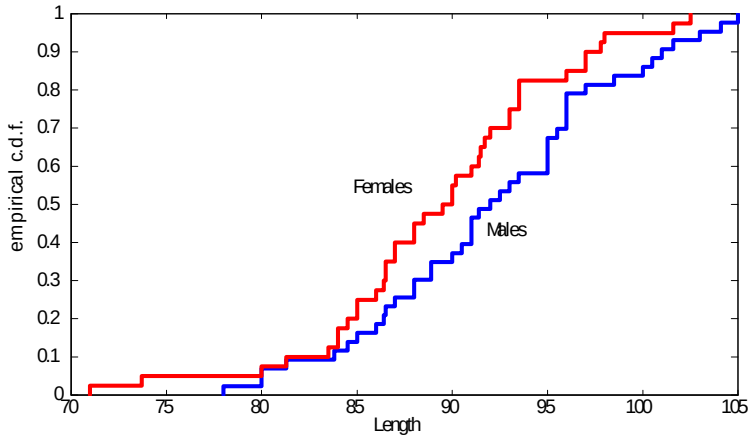


Figure 1.5: Empirical c.d.f.'s for lengths of female and male coyotes

1.13 (a) The two variates are Value (x) and Gross (y), where Value is the average amount the actor's movies have made (in millions of U.S. dollars), and Gross is the amount of the highest grossing movie in which the actor played as a major character (in millions of U.S. dollars). Since the goal is to study the effect of an actor's value (x) on the amount grossed in a movie (y), we choose x as the explanatory variate and y as the response variate.

(b) A scatterplot of the data is given in Figure 1.6.

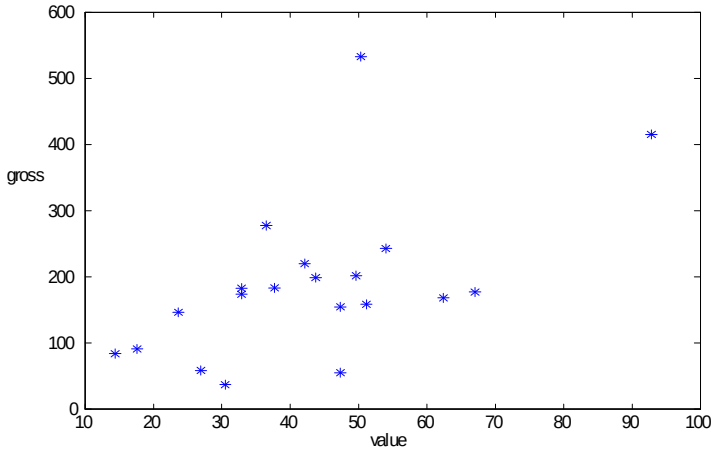


Figure 1.6: Scatterplot of gross versus value

(c) The sample correlation is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{184540.93 - \frac{1}{20}(860.6)(3759.5)}{\left[43315.04 - \frac{1}{20}(860.6)^2\right]^{1/2}\left[971560.19 - \frac{1}{20}(3759.5)^2\right]^{1/2}}$$
$$= 0.558$$

There is a moderately strong positive linear relationship between x and y.

(d) In this example we do not have enough evidence to conclude that a causal relationship exists. Another plausible explanation for the observed data is that there is a third variate such as "the talent of the actor" that affects both the Value (x) and Gross(y) (of course it is very difficult to measure the variate "talent"). Consequently, x and y are expected to be positively correlated, and this is what we observe in this data set.

1.14 (a) The sample correlations between location 1 with locations $2, 3, 4, 5$ are all greater than 0.95 whereas the sample correlations between location 1 and locations $6, 7, 8, 9$ are smaller in value. You would expect the thicknesses at locations which are adjacent to be more highly correlated.

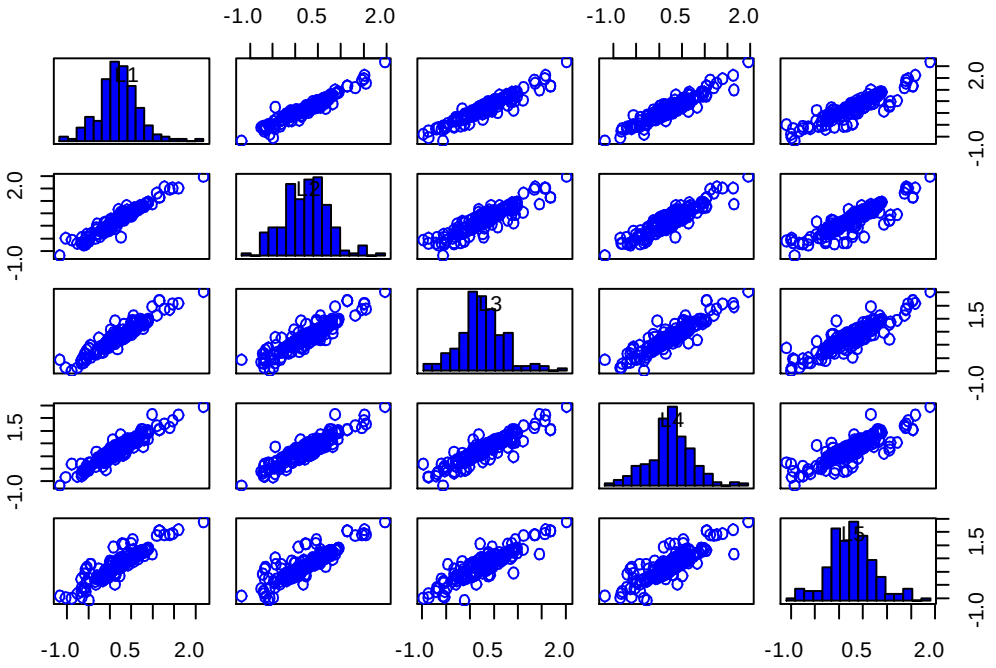


Figure 1.7: Scatterplots of wafer thicknesses for locations 1-5

(b) The variability in the points for locations 1-5 is smaller than for locations 6-9.

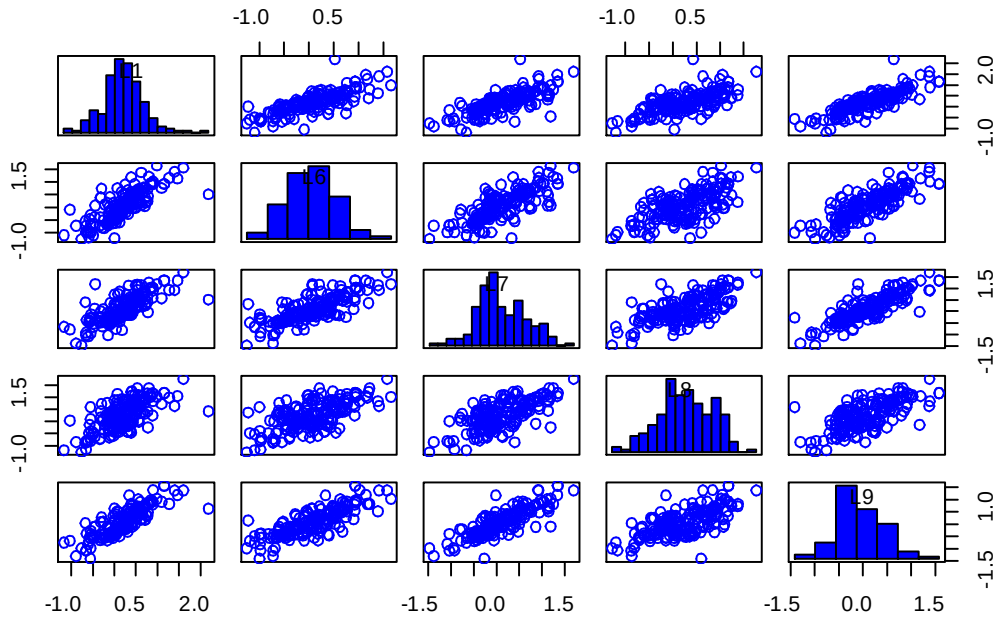


Figure 1.8: Scatterplot of wafer thicknesses for locations 1, 6-9

1.15 (a) The two-way table is:

	Cold	No Cold	Total
Vitamin C	20	80	100
Placebo	30	70	100
Total	50	150	200

(b) The relative risk of a cold in the vitamin C group as compared to the placebo group is

$$\frac{20/(20+80)}{30/(30+70)} = \frac{2}{3}$$

(c) The group taking Vitamin C are only two-thirds as likely to catch a cold as compared to the placebo group which might suggest that taking Vitamin C is associated with fewer colds. (More on this in Chapter 7.)

1.16 (a) For a Binomial model the assumptions are a sequence of independent trials, two outcomes on each trial (S = Success, F = Failure), and $P(S) = \theta$ is the same on each trial.

In this context the trials are people and the two possible outcomes are the person has the blood type A (Success) and the person does not have blood type A (Failure).

The assumption of independent trials may not be a valid. For example, if the population of people consists of families who are highly related then the people would not be independent with respect to the event of interest.

The people would be chosen without replacement from the population in which case the trials would not be independent. However if the sample drawn at random from a large population is small then the trials would be very close to being independent. (Recall the Binomial approximation to the Hypergeometric.)

(b) Since $Y \sim \text{Binomial}(n, \theta)$, the probability function for the random variable $Y =$ the number of people with blood type A is given by

$$P(Y = y) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad \text{for } y = 0, 1, \ldots, n, \quad 0 \leq \theta \leq 1$$

The mean of Y is $E(Y) = n\theta$ and the variance of Y is $Var(Y) = n\theta(1-\theta)$.

(c)

$$P(Y = 20) = \binom{50}{20} \theta^{20} (1-\theta)^{30} \quad \text{for } 0 \leq \theta \leq 1$$

(d) A reasonable estimate of θ is given by the proportion of observed successes in $n = 50$ trials which is $20/50 = 0.4$. An estimate of the probability that in a sample of $n = 10$ there will be at least one person with blood type A is given by

$$\begin{aligned} & 1 - \binom{10}{0} (0.4)^0 (0.6)^{10} = 1 - (0.6)^{10} \\ = \; & 0.9940 \end{aligned}$$

(e) If y successes are observed in n Bernoulli trials then a reasonable estimate of θ is given by the (sample) proportion of successes, that is, a reasonable estimate of θ is y/n.

(f)

$$\begin{aligned} E\left(\frac{Y}{n}\right) &= \frac{1}{n} E(Y) = \frac{1}{n}(n\theta) = \theta \\ Var\left(\frac{Y}{n}\right) &= \left(\frac{1}{n}\right)^2 Var(Y) = \left(\frac{1}{n}\right)^2 [n\theta(1-\theta)] = \frac{\theta(1-\theta)}{n} \to 0 \quad \text{as } n \to \infty \end{aligned}$$

For large values of n, Y/n should be close to θ. By the Central Limit Theorem

$$\begin{aligned} & P\left(\frac{Y}{n}-1.96\sqrt{\frac{\theta\left(1-\theta\right)}{n}}\leq\theta\leq\frac{Y}{n}+1.96\sqrt{\frac{\theta\left(1-\theta\right)}{n}}\right) \\ = & P\left(\left|\frac{\frac{Y}{n}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right|\leq 1.96\right)\approx P\left(\left|Z\right|\leq 1.96\right) \quad \text{where } Z\sim N\left(0,1\right) \\ = & 2P\left(Z\leq 1.96\right)-1=2\left(0.975\right)-1=0.95 \end{aligned}$$

(g) Since there are now four possible outcomes on each independent trial the joint distribution of Y_1 = no. of A types, Y_2 = no. of B types, Y_3 = no. of AB types, Y_4 = no. of O types is given by the Multinomial distribution.

$$P\left(Y_1=y_1,Y_2=y_2,Y_3=y_3,Y_4=y_4\right)=\frac{n!}{y_1!y_2!y_3!y_4!}\theta_1^{y_1}\theta_2^{y_2}\theta_3^{y_3}\theta_4^{y_4}$$

$$\begin{aligned} \text{for } \ y_i &= 0,1,\ldots,n; \quad i=1,2,3,4 \quad \sum_{i=1}^{4}y_i=n \\ \text{and } 0 &< \theta_i<1; \quad i=1,2,3,4 \quad \sum_{i=1}^{4}\theta_i=1 \end{aligned}$$

(h) Since we observe outcome A, y_1 times in a sample of n people a reasonable estimate of θ_1 = proportion of type A in the large population is given by the sample proportion y_1/n. Similarly a reasonable estimate of θ_i is y_i/n for $i=2,3,4$.

1.17 (a) Since $Y\sim$ Poisson(θ) the probability density function of Y is

$$f\left(y\right)=P\left(Y=y\right)=\frac{\theta^y e^{-\theta}}{y!} \quad \text{for } y=0,1,2,\ldots \text{ and } \theta\geq 0$$

for which $E\left(Y\right)=\theta$ and $Var\left(Y\right)=\theta$.

(b) Independence: the number of occurrences in non-overlapping intervals are independent. This assumption seems reasonable since there is no obvious way in which the number of website visits in one time interval would affect the number of calls in another interval.

Individuality: for sufficiently short time periods of length Δt, the probability of 2 or more events occurring in the interval is close to zero, that is, events occur singly not in clusters. A situation in which this assumption might not be reasonable is if the website was selling tickets to a concert for a popular singer and you were looking at the number of visits in the first few minutes of the tickets going on sale. Visits to the website may then occur in clusters due to the high demand.

Homogeneity or Uniformity: events occur at a uniform or homogeneous rate θ over time so that the probability of one occurrence in an interval $(t, t+\Delta t)$ is approximately $\theta\Delta t$ for small Δt for any value of t. The homogeneity assumption may be an issue since the rate of visits may vary with the time of day.

(c)

(i) Let Y_i = no. of visits in the i'th second, $i = 1, 2, \ldots, 10$. Then

$$\begin{aligned} & P(Y_1 = 1, Y_2 = 4, Y_3 = 5, Y_4 = 1, Y_5 = 0, \\ Y_6 &= 2, Y_7 = 5, Y_8 = 4, Y_9 = 3, Y_{10} = 2) \\ &= P(Y_1 = 1) P(Y_2 = 4) P(Y_3 = 5) P(Y_4 = 1) P(Y_5 = 0) \\ & \quad P(Y_6 = 2) P(Y_7 = 5) P(Y_8 = 4) P(Y_9 = 3) P(Y_{10} = 2) \\ &= \left(\frac{\theta^1 e^{-\theta}}{1!}\right)\left(\frac{\theta^4 e^{-\theta}}{4!}\right)\left(\frac{\theta^5 e^{-\theta}}{5!}\right)\left(\frac{\theta^1 e^{-\theta}}{1!}\right)\left(\frac{\theta^0 e^{-\theta}}{0!}\right) \\ & \quad \left(\frac{\theta^2 e^{-\theta}}{2!}\right)\left(\frac{\theta^5 e^{-\theta}}{5!}\right)\left(\frac{\theta^4 e^{-\theta}}{4!}\right)\left(\frac{\theta^3 e^{-\theta}}{3!}\right)\left(\frac{\theta^2 e^{-\theta}}{2!}\right) \\ &= \frac{\theta^{27} e^{-10\theta}}{1!4!5!1!0!2!5!4!3!2!} \quad \text{for } \theta \geq 0 \end{aligned}$$

(ii) A reasonable estimate of the mean θ is the sample mean $\bar{y} = 27/10 = 2.7$.

(iii) An estimate of the probability that there is at least one visit to the website in a one second interval is

$$1 - \frac{(2.7)^0 e^{-2.7}}{0!} = 1 - e^{-2.7} = 0.9328$$

(d)

(i)

$$\begin{aligned} E(\bar{Y}) &= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \theta = \frac{1}{n}(n\theta) = \theta \\ \text{and } Var(\bar{Y}) &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var(Y_i) \quad \text{since } Y_i's \text{ are independent r.v.'s} \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \theta = \left(\frac{1}{n}\right)^2 (n\theta) = \frac{\theta}{n} \to 0 \quad \text{as} \quad n \to \infty \end{aligned}$$

For large values of n, the sample mean $\bar{Y}$ should be close to the mean θ.

(ii) By the Central Limit Theorem

$$\begin{aligned} & P\left(\bar{Y} - 1.96\sqrt{\theta/n} \leq \theta \leq \bar{Y} + 1.96\sqrt{\theta/n}\right) \\ &= P\left(\left|\frac{\bar{Y} - \theta}{\sqrt{\theta/n}}\right| \leq 1.96\right) \approx P(|Z| \leq 1.96) \quad \text{where } Z \sim N(0,1) \\ &= 2P(Z \leq 1.96) - 1 = 2(0.975) - 1 = 0.95 \end{aligned}$$

1.18 (a) Since $Y \sim G(\mu, \sigma)$ the probability density function of Y is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \quad \text{for } y \in \Re,\ \mu \in \Re,\ \sigma > 0$$

for which $E(Y) = \mu$ and $Var(Y) = \sigma^2$.

(b) To find the pth quantile of the $G(\mu, \sigma)$ distribution we need to find $Q(p)$ such that $P[Y \leq Q(p)] = p$ or equivalently $P\left(Z \leq \frac{Q(p)-\mu}{\sigma}\right) = p$. Since the cumulative distribution function of a Gaussian random variable can only be written as an integral, let Φ^{-1} be the inverse cumulative distribution function of a $G(0,1)$ random variable. Then $\frac{Q(p)-\mu}{\sigma} = \Phi^{-1}(p)$ or $Q(p) = \mu + \sigma\Phi^{-1}(p)$ as required.

By symmetry of the $G(0,1)$ distribution we know that $\Phi^{-1}(0.5) = 0$ and therefore the median $= Q(0.5) = \mu + \sigma(0) = \mu$.

The IQR of the $G(\mu, \sigma)$ distribution is equal to

$$\begin{aligned} Q(0.75) - Q(0.25) &= \sigma\left[\Phi^{-1}(0.75) - \Phi^{-1}(0.25)\right] \\ &= 0.6744898 - (-0.6744898) \\ &= 1.34898\sigma \approx 1.35\sigma \end{aligned}$$

(c)

(i) A reasonable estimate of the mean μ is the sample mean

$$\bar{y} = \frac{1916}{16} = 119.75$$

(ii) A reasonable estimate of the variance σ^2 is the sample variance

$$s^2 = \frac{1}{15}\left[231618 - 16(119.75)^2\right] = 145.1\dot{3}$$

(iii) An estimate of the probability that a randomly chosen UWaterloo Math student will have an IQ greater than 120 is given by

$$\begin{aligned} P(Y \geq 120) \quad \text{where } Y &\sim N\left(119.75, 145.1\dot{3}\right) \\ &= P\left(Z \geq \frac{120 - 119.75}{\sqrt{145.1\dot{3}}}\right) \quad \text{where } Z \sim N(0,1) \\ &= P(Z \geq 0.0208) \approx P(Z \geq 0.02) = 1 - 0.50798 \\ &= 0.49202 \end{aligned}$$

(d)

(i) The distribution of a linear combination of Gaussian (Normal) random variables has a Gaussian (Normal) distribution. Since

$$\begin{aligned} E\left(\bar{Y}\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\left(n\mu\right) = \mu \\ \text{and } Var\left(\bar{Y}\right) &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var\left(Y_i\right) \quad \text{since } Y_i's \text{ are independent r.v.'s} \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n}\sigma^2 = \left(\frac{1}{n}\right)^2\left(n\sigma^2\right) = \frac{\sigma^2}{n} \end{aligned}$$

therefore $\bar{Y} \sim G\left(\mu, \sigma/\sqrt{n}\right)$.

$$Var\left(\bar{Y}\right) = \frac{\sigma^2}{n} \to 0 \quad \text{as } n \to \infty$$

For large values of n, the sample mean $\bar{Y}$ should be close to the mean μ.

(ii)

$$\begin{aligned} & P\left(\bar{Y} - 1.96\sigma/\sqrt{n} \le \mu \le \bar{Y} + 1.96\sigma/\sqrt{n}\right) \\ = \; & P\left(\left|\bar{Y} - \mu\right| \le 1.96\sigma/\sqrt{n}\right) = P\left(|Z| \le 1.96\right) \quad \text{where } Z \sim G\left(0,1\right) \\ = \; & 2P\left(Z \le 1.96\right) - 1 \\ = \; & 2\left(0.975\right) - 1 = 0.95 \end{aligned}$$

(iii) We want $P\left(\left|\bar{Y} - \mu\right| \le 1.0\right) \ge 0.95$ where $\bar{Y} \sim G\left(\mu, 12/\sqrt{n}\right)$ or

$$\begin{aligned} P\left(\left|\bar{Y} - \mu\right| \le 1.0\right) &= P\left(\frac{\left|\bar{Y} - \mu\right|}{12/\sqrt{n}} \le \frac{1.0}{12/\sqrt{n}}\right) \\ &= P\left(|Z| \le \frac{\sqrt{n}}{12}\right) \ge 0.95 \quad \text{where } Z \sim G\left(0,1\right) \end{aligned}$$

Since $P\left(|Z| \le 1.96\right) = 0.95$ we want $\sqrt{n}/12 \ge 1.96$ or
$n \ge \left(1.96\right)^2\left(144\right) = 553.2$. Therefore $n = 554$.

1.19 (a) If $Y \sim$ Exponential(θ) then the memoryless property is

$$P\left(Y > c + b | Y > b\right) = P\left(Y > c\right)$$

This implies that if a battery has lasted b units of time then the probability the battery will last an additional c units of time does not depend on b but only depends on c. In other words the battery is not deteriorating over time.

(b) Since $Y \sim$ Exponential(θ) the probability density function of Y is

$$f\left(y\right) = \frac{1}{\theta}e^{-y/\theta} \quad \text{for } \; y > 0 \text{ and } \theta > 0$$

for which $E\left(Y\right) = \theta$ and $Var\left(Y\right) = \theta^2$.

(c) To find the pth quantile of the Exponential(θ) distribution we need to find $Q(p)$ such that $P[Y \leq Q(p)] = p$ or equivalently $1 - e^{-Q(p)/\theta} = p$. Solving for $Q(p)$ we obtain $Q(p) = -\theta \log(1-p)$.

The median of the Exponential(θ) distribution is $m = Q(0.5) = -\theta \log(0.5) = \theta \log 2$.

The IQR of the Exponential(θ) distribution is equal to

$$\begin{aligned} Q(0.75) - Q(0.25) &= -\theta \log(1-0.75) - [-\theta \log(1-0.25)] \\ &= \theta [\log(0.75) - \log(0.25)] \\ &= \theta \log\left(\frac{0.75}{0.25}\right) \\ &= \theta \log(3) \end{aligned}$$

(d)

(i) A reasonable estimate of the mean θ is the sample mean $\bar{y} = \frac{7442.8}{20} = 372.14$.

(ii) An estimate of

$$P(Y > 100) = \int_{100}^{\infty} \frac{1}{\theta} e^{-y/\theta} dy = e^{-100/\theta}$$

is $e^{-100/372.14} = 0.7644$

(e)

(i)

$$\begin{aligned} E(\bar{Y}) &= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \theta = \frac{1}{n}(n\theta) = \theta \\ \text{and } Var(\bar{Y}) &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var(Y_i) \quad \text{since } Y_i's \text{ are independent r.v.'s} \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \theta^2 = \left(\frac{1}{n}\right)^2 (n\theta^2) = \frac{\theta^2}{n} \to 0 \quad \text{as } n \to \infty \end{aligned}$$

For large values of n, the sample mean $\bar{Y}$ should be close to the mean θ.

(ii) By the Central Limit Theorem

$$\begin{aligned} & P\left(\bar{Y} - 1.6449\theta/\sqrt{n} \leq \theta \leq \bar{Y} + 1.6449\theta/\sqrt{n}\right) \\ = \; & P\left(\left|\frac{\bar{Y} - \theta}{\theta/\sqrt{n}}\right| \leq 1.6449\right) \approx P(|Z| \leq 1.6449) \quad \text{where } Z \sim N(0,1) \\ = \; & 2P(Z \leq 1.6449) - 1 = 2(0.95) - 1 = 0.9 \end{aligned}$$

1.20 (a) $E\left(Y_i^2\right) = Var\left(Y_i\right) + \left[E\left(Y_i\right)\right]^2 = \sigma^2 + \mu^2$.

(b)

$$E(\bar{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}\left(n\mu\right) = \mu$$

Since $Y_1, Y_2, \ldots, Y_n$ are independent random variables

$$\begin{aligned} Var(\bar{Y}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var\left(Y_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \sigma^2 \\ &= \left(\frac{1}{n}\right)^2 \left(n\sigma^2\right) = \frac{\sigma^2}{n} \\ E\left[(\bar{Y})^2\right] &= \left[E\left(\bar{Y}\right)\right]^2 + Var\left(\bar{Y}\right) = \mu^2 + \frac{\sigma^2}{n} \end{aligned}$$

(c)

$$\begin{aligned} E\left(S^2\right) &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} E\left(Y_i^2\right) - nE\left[\left(\bar{Y}\right)^2\right]\right\} \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\mu^2 + \sigma^2\right) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right] \\ &= \frac{1}{n-1}\left[n\left(\mu^2 + \sigma^2\right) - n\mu^2 - \sigma^2\right] \\ &= \frac{1}{n-1}\left[\left(n-1\right)\sigma^2\right] = \sigma^2 \end{aligned}$$

1.24 (a) This is an experimental study because the researchers controlled the treatments that the premature babies received.

(b) The researchers are interested in the population of premature babies in New York state at the time of the study. The units are premature babies.

(c) One variate is heart rate. This is a discrete variate.
Another variate is respiratory rate. This is a discrete variate.
Another variate is oxygen saturation. This is a continuous variate.
Another variate is sucking pattern. This is an ordinal variate.
Another variate is activity level. This is a categorical variate.
Another variate is treatment type. This is a categorical variate.
Another variate is the hospital the baby was in. This is a categorical variate.

(d) Here are several examples of attributes. There are other possible attributes:
The differences in mean or average heart rate between the four treatments.
The differences in mean or average respiratory rate between the four treatments.
The differences in mean or average oxygen saturation between the four treatments.
The differences in the distribution of sucking behaviours between the four treatments.
The differences in the distribution of activity levels between the four treatments.

SOLUTIONS TO CHAPTER 2 PROBLEMS

2.1 (a)

$$\begin{aligned} G(\theta) &= \theta^a (1-\theta)^b \quad \text{for } 0<\theta<1 \\ g(\theta) &= \log G(\theta) = a\log\theta + b\log(1-\theta) \quad \text{for } 0<\theta<1 \\ g'(\theta) &= \frac{a}{\theta} - \frac{b}{1-\theta} = \frac{a(1-\theta)-b\theta}{\theta(1-\theta)} = \frac{a-(a+b)\theta}{\theta(1-\theta)} \\ g'(\theta) &= 0 \text{ if } \theta = \frac{a}{a+b} \end{aligned}$$

Since $g'(\theta) > 0$ for $0 < \theta < \frac{a}{a+b}$ and $g'(\theta) < 0$ for $1 > \theta > \frac{a}{a+b}$ then by the First Derivative Test $g(\theta)$ has a maximum value at $\theta = \frac{a}{a+b}$.

(b)

$$\begin{aligned} G(\theta) &= \theta^{-a} e^{-b/\theta} \quad \text{for } \theta > 0 \\ g(\theta) &= \log G(\theta) = -a\log\theta - \frac{b}{\theta} \quad \text{for } \theta > 0 \\ g'(\theta) &= \frac{-a}{\theta} + \frac{b}{\theta^2} = \frac{-a\theta+b}{\theta^2} \\ g'(\theta) &= 0 \text{ if } \theta = \frac{b}{a} \end{aligned}$$

Since $g'(\theta) > 0$ for $0 < \theta < \frac{b}{a}$ and $g'(\theta) < 0$ for $\theta > \frac{b}{a}$ then by the First Derivative Test $g(\theta)$ has a maximum value at $\theta = \frac{b}{a}$.

(c)

$$\begin{aligned} G(\theta) &= \theta^a e^{-b\theta}, \quad \theta > 0 \\ g(\theta) &= \log G(\theta) = a\log\theta - b\theta, \quad \theta > 0 \\ g'(\theta) &= \frac{a}{\theta} - b = \frac{a-b\theta}{\theta} \\ g'(\theta) &= 0 \text{ if } \theta = \frac{a}{b} \end{aligned}$$

Since $g'(\theta) > 0$ for $0 < \theta < \frac{a}{b}$ and $g'(\theta) < 0$ for $\theta > \frac{a}{b}$ then by the First Derivative Test $g(\theta)$ has a maximum value at $\theta = \frac{a}{b}$.

(d)

$$\begin{aligned} G(\theta) &= e^{-a(\theta-b)^2} \quad \text{for } \theta \in \Re \\ g(\theta) &= \log G(\theta) = -a(\theta-b)^2 \quad \text{for } \theta \in \Re \\ g'(\theta) &= -2a(\theta-b) \\ g'(\theta) &= 0 \text{ if } \theta = b \end{aligned}$$

Since $g'(\theta) > 0$ for $\theta < b$ and $g'(\theta) < 0$ for $\theta > b$ then by the First Derivative Test $g(\theta)$ has a maximum value at $\theta = b$.

2.2 If $y = 0$ then

$$L(\theta) = P(Y = 0; \theta) = \binom{n}{0}\theta^0(1-\theta)^n = (1-\theta)^n \quad \text{for } 0 \le \theta \le 1$$

$L(\theta)$ is a decreasing function for $\theta \in [0,1]$ and its maximum value on the interval $[0,1]$ occurs at the endpoint $\theta = 0$ and so $\hat{\theta} = 0 = \frac{0}{n}$.

If $y = n$ then

$$L(\theta) = P(Y = n; \theta) = \binom{n}{n}\theta^n(1-\theta)^0 = \theta^n \quad \text{for } 0 \le \theta \le 1$$

$L(\theta)$ is an increasing function for $\theta \in [0,1]$ and its maximum value on the interval $[0,1]$ occurs at the endpoint $\theta = 1$ and so $\hat{\theta} = 1 = \frac{n}{n}$.

In both cases $\hat{\theta} = \frac{y}{n}$.

2.3 (a) The probability of the observed results for Experiment 1 is

$$\begin{aligned} & P(\text{total number of individuals examined } = 100; \theta) \\ = \ & \binom{99}{9}\theta^{10}(1-\theta)^{90} \quad \text{for } 0 \le \theta \le 1 \end{aligned}$$

The probability of the observed results for Experiment 2 is

$$\begin{aligned} & P(10 \text{ individuals with blood type B }; \theta) \\ = \ & \binom{100}{10}\theta^{10}(1-\theta)^{90} \quad \text{for } 0 \le \theta \le 1 \end{aligned}$$

The likelihood function in both cases simplifies to

$$L(\theta) = \theta^{10}(1-\theta)^{90} \quad \text{for } 0 \le \theta \le 1$$

if we ignore constants with respect to θ. The log likelihood function is

$$l(\theta) = 10\log\theta + 90\log(1-\theta) \quad \text{for } 0 < \theta < 1$$

Now

$$\begin{aligned} l'(\theta) &= \frac{10}{\theta} - \frac{90}{1-\theta} \\ &= \frac{10-100\theta}{\theta(1-\theta)} = 0 \quad \text{if } \theta = \frac{10}{100} = 0.1 \end{aligned}$$

and the maximum likelihood estimate of θ is $\hat{\theta} = 0.1$.

(b) Let Y = the number of donors with blood type B. Then $Y \sim \text{Binomial}(n, 0.1)$ and $E(Y) = 0.1n$ and $Var(Y) = 0.1(0.9)n = 0.09n$. We went to find n such that $P(Y \geq 10) \geq 0.90$. By the Normal approximation to the Binomial we have

$$P(Y \geq 10) \approx P\left(Z \geq \frac{9.5 - 0.1n}{\sqrt{0.09n}}\right) \quad \text{where } Z \sim N(0,1)$$

Since $P(Z \geq -1.2816) = 0.90$ we solve

$$\frac{9.5 - 0.1n}{\sqrt{0.09n}} = -1.2816$$

or

$$n^2 - 204.78n + 9025 = 0$$

which gives $n = 140.6$. Since

$$\begin{aligned} P(Y \geq 10) &= \texttt{1-pbinom(9,139,0.1)} = 0.8981225 \\ P(Y \geq 10) &= \texttt{1-pbinom(9,140,0.1)} = 0.9027362 \\ P(Y \geq 10) &= \texttt{1-pbinom(9,141,0.1)} = 0.9071738 \end{aligned}$$

we can see that $n = 140$ is the smallest value of n such that $P(Y \geq 10) \geq 0.90$.

2.4 (a) The likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \theta^{y_i - 1}(1-\theta) = \theta^{\sum_{i=1}^{n} y_i - n}(1-\theta)^n \\ &= \theta^{n(\bar{y}-1)}(1-\theta)^n \quad \text{for } 0 \leq \theta < 1 \end{aligned}$$

if $\bar{y} > 1$. The log likelihood is

$$l(\theta) = n(\bar{y}-1)\log\theta + n\log(1-\theta) \quad \text{for } 0 < \theta < 1$$

Solving

$$l'(\theta) = \frac{n(\bar{y}-1)}{\theta} - \frac{n}{1-\theta} = \frac{n(\bar{y}-1)(1-\theta) - n\theta}{\theta(1-\theta)} = 0$$

gives the maximum likelihood estimate

$$\hat{\theta} = \frac{\bar{y}-1}{\bar{y}}$$

If $\bar{y} = 1$ then $L(\theta) = (1-\theta)^n$ which is a decreasing function for $0 < \theta < 1$. The maximum value of $L(\theta) = (1-\theta)^n$ on the interval $[0, 1)$ occurs at the endpoint $\theta = 0$ and so $\hat{\theta} = 0$. Therefore

$$\hat{\theta} = \frac{\bar{y} - 1}{\bar{y}}$$

holds for all values of $\bar{y}$.

(b) The relative likelihood function is

$$\begin{aligned} R(\theta) &= \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^{n(\bar{y}-1)}(1-\theta)^n}{\hat{\theta}^{n(\bar{y}-1)}\left(1-\hat{\theta}\right)^n} \\ &= \left[\left(\frac{\theta}{\hat{\theta}}\right)^{(\bar{y}-1)}\left(\frac{1-\theta}{1-\hat{\theta}}\right)\right]^n \quad \text{for } 0 \le \theta < 1 \text{ and } \bar{y} > 1 \end{aligned}$$

If $n = 200$ and $\sum_{i=1}^{200} y_i = 400$ then $\bar{y} = \frac{400}{200} = 2$, $\hat{\theta} = \frac{2-1}{2} = 0.5$, and

$$R(\theta) = \left[\left(\frac{\theta}{0.5}\right)\left(\frac{1-\theta}{0.5}\right)\right]^{200} = [4\theta(1-\theta)]^{200} \quad \text{for } 0 \le \theta < 1$$

(c) A graph of $R(\theta)$ is given in Figure 2.1.

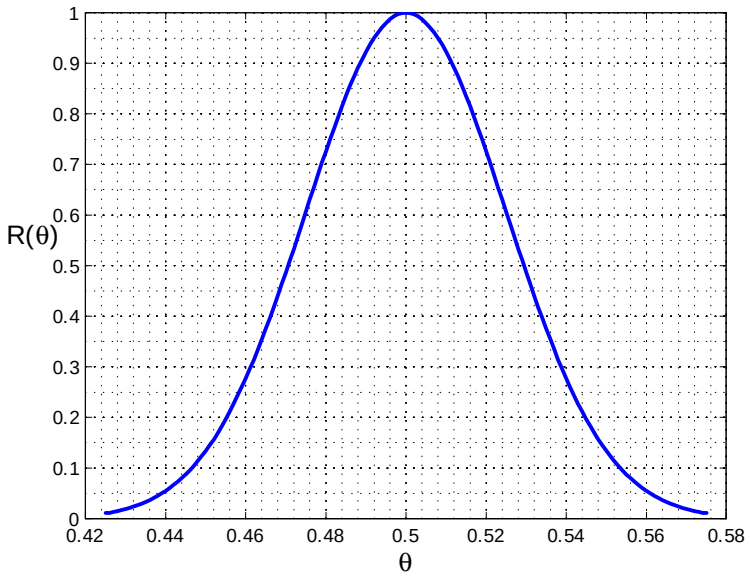


Figure 2.1: Relative likelihood function for fracture data

(d) Since $p = P(Y = 1; \theta) = (1-\theta)$ then by the invariance property of maximum likelihood estimates the maximum likelihood estimate of p based on the data in (c) is $\hat{p} = (1-\hat{\theta}) = 1 - 0.5 = 0.5$.

2.5 (a) Since $t = 1$, the likelihood function is

$$L(\theta) = \prod_{i=1}^{10} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \left(\prod_{i=1}^{10} y_i! \right)^{-1} \theta^{41} e^{-10\theta} \quad \text{for } \theta \geq 0$$

or more simply (ignoring constants with respect to θ)

$$L(\theta) = \theta^{41} e^{-10\theta} \quad \text{for } \theta > 0$$

The log likelihood function is

$$l\left(\theta\right) = 41 \log \theta - 10\theta \quad \text{for } \theta > 0$$

Solving

$$l'\left(\theta\right) = \frac{41}{\theta} - 10 = 0$$

gives the maximum likelihood estimate $\hat{\theta} = 4.1$.

(b) Since

$$p = P\left(\text{no transactions in a two minute interval }; \theta\right) = \frac{\left(2\theta\right)^0 e^{-2\theta}}{0!} = e^{-2\theta}$$

then by the invariance property of maximum likelihood estimates the maximum likelihood estimate of p is $\hat{p} = e^{-2\hat{\theta}} = 0.000275$.

2.6 (a) The joint probability density function of the observations $y_1, y_2, ..., y_n$ is given by

$$\begin{aligned} \prod_{i=1}^{n} f(y_i; \theta) &= \prod_{i=1}^{n} \frac{2y_i}{\theta} e^{-y_i^2/\theta} \\ &= 2^n (\prod_{i=1}^{n} y_i) \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta} \sum_{i=1}^{n} y_i^2 \right) \quad \text{for } \ \theta > 0 \end{aligned}$$

The likelihood function (ignoring constants with respect to θ) is

$$L\left(\theta\right) = \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta} \sum_{i=1}^{n} y_i^2 \right) \quad \text{for } \theta > 0$$

and the log likelihood is

$$l(\theta) = -n \log(\theta) - \frac{1}{\theta} \sum_{i=1}^{n} y_i^2 \quad \text{for } \theta > 0$$

Solving

$$l'(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} y_i^2 = \frac{1}{\theta^2} \left(\sum_{i=1}^{n} y_i^2 - n\theta \right) = 0$$

gives the maximum likelihood estimate

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i^2$$

(b) The relative likelihood function is

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\frac{1}{\theta^n}\exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} y_i^2\right)}{\frac{1}{\hat{\theta}^n}\exp\left(-\frac{1}{\hat{\theta}}\sum_{i=1}^{n} y_i^2\right)} \quad \text{for } \theta > 0$$

But

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$$

so

$$\sum_{i=1}^{n} y_i^2 = n\hat{\theta}$$

Therefore

$$\begin{aligned} R(\theta) &= \frac{\frac{1}{\theta^n}\exp\left(-\frac{n\hat{\theta}}{\theta}\right)}{\frac{1}{\hat{\theta}^n}\exp(-n)} \\ &= \left(\frac{\hat{\theta}}{\theta}\right)^n e^{n\left(1-\hat{\theta}/\theta\right)} \quad \text{for } \theta > 0 \end{aligned}$$

(c) Graphs of $R(\theta)$ for $n = 20$, $\hat{\theta} = 3.6$ (solid line) and $n = 60$, $\hat{\theta} = 3.6$ (dotted line) are given in Figure 2.2.

(d) Both relative likelihood functions have a maximum value of 1 which occurs at the maximum likelihood estimate $\theta = \hat{\theta}$. The relative likelihood function for $n = 20$ is more asymmetric and skewed to the right while the relative likelihood function for $n = 60$ is more symmetric about the maximum likelihood estimate $\theta = \hat{\theta}$ and rather bell-shaped. The relative likelihood function for $n = 20$ is more spread out as compared to the relative likelihood function for $n = 60$ or equivalently the relative likelihood function for $n = 60$ is more concentrated about the maximum likelihood estimate $\hat{\theta}$.

(e) Since

$$\begin{aligned} P(Y > 1;\theta) &= 1 - \int_0^1 \frac{2y}{\theta} e^{-y^2/\theta} \\ &= 1 + \left(e^{-y^2/\theta}\Big|_0^1\right) \\ &= e^{-1/\theta} \end{aligned}$$

therefore by the invariance property of maximum likelihood estimates the maximum likelihood estimate of $P(Y > 1;\theta)$ is

$$e^{-1/\hat{\theta}} = e^{-1/3.6} = 0.7574651$$

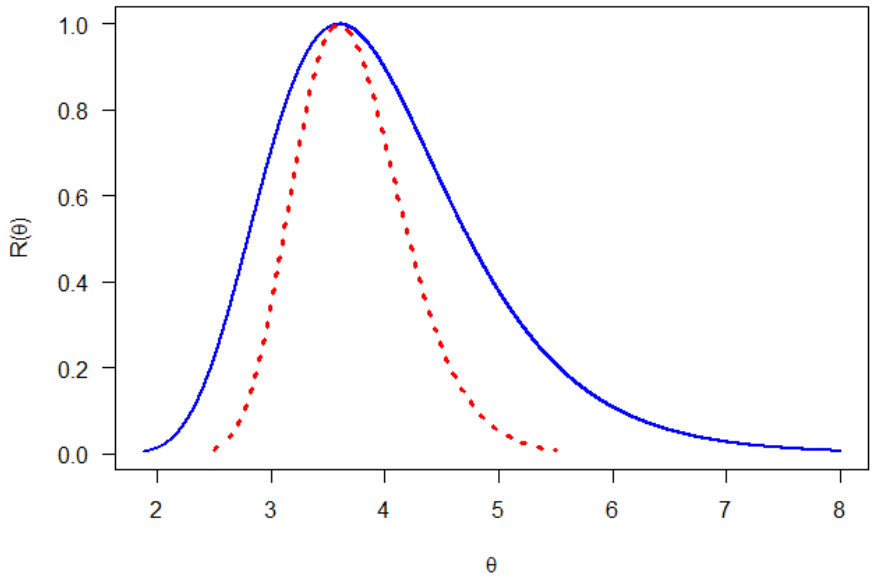


Figure 2.2: Relative likelihood functions for Problem 6

2.7 (a) From Example 2.3.2 the joint likelihood function of μ and σ ignoring constants can be written as

$$L(\mu,\sigma) = \sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]\exp\left[-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right] \quad \text{for } \mu\in\Re \text{ and } \sigma>0$$

If σ is known then the likelihood function of μ ignoring constants with respect to μ is

$$L(\mu) = \exp\left[-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right] \quad \text{for } \mu\in\Re$$

The log likelihood function is

$$l(\mu) = -\frac{n}{2\sigma^2}(\bar{y}-\mu)^2 \quad \text{for } \mu\in\Re$$

and

$$l'(\mu) = \frac{n}{\sigma^2}(\bar{y}-\mu) = 0 \quad \text{if } \mu=\bar{y}$$

and therefore the maximum likelihood estimate of μ is $\hat{\mu}=\bar{y}$ which does not depend on σ.

(b) If μ is known then the likelihood function of σ is

$$\begin{aligned} L(\sigma) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \\ &= (2\pi)^{-n/2}\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \quad \text{for } \sigma>0 \end{aligned}$$

or more simply (ignoring constants with respect to σ)

$$L(\sigma) = \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \quad \text{for } \sigma > 0$$

The log likelihood function is

$$l(\sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \quad \text{for } \sigma > 0$$

and

$$l'(\sigma) = \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\mu)^2 = \frac{1}{\sigma^3}\left[-n\sigma^2 + \sum_{i=1}^{n}(y_i-\mu)^2\right] = 0$$

if

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2$$

Therefore the maximum likelihood estimate of σ is

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2}$$

which does depend on μ.

2.8 (a) The likelihood function

$$L(\theta) = \prod_{i=1}^{n}(\theta+1)y_i^{\theta} = (\theta+1)^n\left(\prod_{i=1}^{n}y_i\right)^{\theta} \quad \text{for } \theta > -1$$

The log likelihood function is

$$l(\theta) = n\log(\theta+1) + \theta\sum_{i=1}^{n}\log(y_i) \quad \text{for } \theta > -1$$

Solving

$$\frac{d}{d\theta}l(\theta) = \frac{n}{1+\theta} + \sum_{i=1}^{n}\log(y_i) = 0$$

gives

$$\hat{\theta} = \frac{n}{-\sum_{i=1}^{n}\log(y_i)} - 1$$

(b) The log relative likelihood function is

$$r(\theta) = l(\theta) - l(\hat{\theta}) = n\log\left(\frac{\theta+1}{\hat{\theta}+1}\right) + (\theta-\hat{\theta})\sum_{i=1}^{n}\log(y_i) \quad \text{for } \theta > -1$$

Since

$$\hat{\theta} = \frac{n}{-\sum_{i=1}^{n}\log(y_i)} - 1$$

therefore

$$\sum_{i=1}^{n} \log(y_i) = -\frac{n}{\hat{\theta}+1}$$

and

$$r(\theta) = n \log\left(\frac{\theta+1}{\hat{\theta}+1}\right) + n\left(\frac{\hat{\theta}-\theta}{\hat{\theta}+1}\right) \quad \text{for } \theta > -1$$

(c) Graphs of $r(\theta)$ for $n = 15$, $\hat{\theta} = -0.5652$ (solid line) and $n = 45$, $\hat{\theta} = -0.5652$ (dotted line) are given in Figure 2.3.

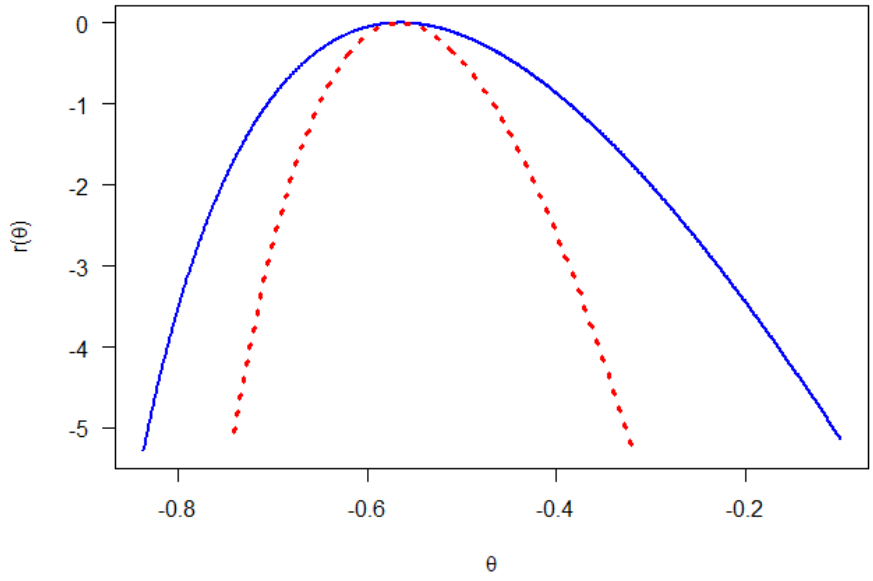


Figure 2.3: Log relative likelihood functions for Problem 8

(d) Both log relative likelihood functions have a maximum value of 0 which occurs at the maximum likelihood estimate $\theta = \hat{\theta}$. Both log relative likelihood functions are concave down. The log relative likelihood function for $n = 45$ is more symmetric about the maximum likelihood estimate $\theta = \hat{\theta}$ and more quadratic in shape than the log relative likelihood function for $n = 15$. The log relative likelihood function for $n = 15$ is more spread out as compared to the relative likelihood function for $n = 45$ or equivalently the log relative likelihood function for $n = 45$ is more concentrated about the maximum likelihood estimate $\hat{\theta}$.

2.9 (a) The joint probability density function of the observations $y_1, y_2, ..., y_n$ is given by

$$\begin{aligned}\prod_{i=1}^{n} f(y_i;\theta) &= \prod_{i=1}^{n} \frac{\theta}{y_i^{\theta+1}} = \frac{\theta^n}{\prod_{i=1}^{n} y_i^{\theta+1}} = \frac{\theta^n}{\left(\prod_{i=1}^{n} y_i\right)^{\theta+1}} \\ &= \frac{\theta^n}{\left(\prod_{i=1}^{n} y_i\right)\left(\prod_{i=1}^{n} y_i\right)^{\theta}} \quad \text{for } \theta > 1\end{aligned}$$

The likelihood function (ignoring constants with respect to θ) is

$$L(\theta) = \theta^n \left(\prod_{i=1}^{n} y_i\right)^{-\theta} \quad \text{for } \theta > 1$$

The log likelihood function is

$$l(\theta) = n\log(\theta) - \theta \sum_{i=1}^{n} \log(y_i) \quad \text{for } \theta > 1$$

Solving

$$\frac{d}{d\theta} l(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} \log(y_i) = 0$$

gives

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log(y_i)}$$

(b) The relative likelihood function is

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^n \left(\prod_{i=1}^{n} y_i\right)^{-\theta}}{\hat{\theta}^n \left(\prod_{i=1}^{n} y_i\right)^{-\hat{\theta}}} = \left(\frac{\theta}{\hat{\theta}}\right)^n \left(\prod_{i=1}^{n} y_i\right)^{\hat{\theta}-\theta} \quad \text{for } \theta > 1$$

2.10 (a)

$$\begin{aligned}P(MM) &= P(FF) = P(FF|\text{pair is identical})\, P(\text{pair is identical}) \\ &\quad + P(FF|\text{pair is not identical})\, P(\text{pair is not identical}) \\ &= \left(\frac{1}{2}\right)\alpha + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)(1-\alpha) = \frac{1+\alpha}{4} \\ P(MF) &= 1 - P(MM) - P(FF) = 1 - \left(\frac{1+\alpha}{4}\right) - \left(\frac{1+\alpha}{4}\right) = \frac{1-\alpha}{2}\end{aligned}$$

where M = male, F = female and α = probability the pair is identical.

(b)

$$L(\alpha) = \frac{n!}{n_1!n_2!n_3!}\left(\frac{1+\alpha}{4}\right)^{n_1}\left(\frac{1+\alpha}{4}\right)^{n_2}\left(\frac{1-\alpha}{2}\right)^{n_3} \quad \text{where } n = n_1 + n_2 + n_3$$

or more simply (ignoring constants with respect to α)

$$L(\alpha) = (1+\alpha)^{n_1+n_2}(1-\alpha)^{n_3} \quad \text{for } 0 \leq \alpha \leq 1$$

Maximizing $L(\alpha)$ gives $\hat{\alpha} = (n_1 + n_2 - n_3)/n$. For $n_1 = 16$, $n_2 = 16$ and $n_3 = 18$, $\hat{\alpha} = 0.28$.

2.11 (a) The parameter θ represents the mean number of points scored in a game by Wayne when he played for the Edmonton Oilers.

(b) Independence: If Wayne's performance in a game affected how well he played in the next game, the assumption of independence (number of points scored in non-overlapping games are independent) would not hold.
Individuality: This assumption seems reasonably since, if we took a game of sufficiently small "size" then the probability that Wayne scored 2 or more points would be close to zero (events occur singly not in clusters).
Homogeneity or Uniformity: The homogeneity assumption would implied that Wayne was a consistent player during the 11 years he played in Edmonton. Given the type of player Wayne was, this assumption seems reasonable.

(c) The sample mean is

$$\begin{aligned} & \frac{1}{696}\sum_{y=0}^{8} y f_y \\ = \; & \frac{1}{696}\left[0(69) + 1(155) + 2(171) + 3(143) + 4(79) + 5(57) + 6(14) + 7(6) + 8(2)\right] \\ = \; & \frac{1669}{696} \approx 2.3980 \end{aligned}$$

Since

$$\begin{aligned} & \sum_{y=0}^{8} y^2 f_y \\ = \; & (0)^2(69) + (1)^2(155) + (2)^2(171) + (3)^2(143) \\ & + (4)^2(79) + (5)^2(57) + (6)^2(14) + (7)^2(6) + (8)^2(2) \\ = \; & 5741 \end{aligned}$$

therefore the sample variance is

$$\frac{1}{695}\left[5741 - \frac{(1669)^2}{696}\right] = 2.501809$$

The sample mean and sample variance are reasonably close in value for these data.

(d) The likelihood function based on the Poisson model and the frequency table is

$$\begin{aligned} L(\theta) &= \frac{696!}{69!155!171!143!79!57!14!6!2!0!} \\ &\quad \times \left(\frac{\theta^0 e^{-\theta}}{0!}\right)^{69} \left(\frac{\theta^1 e^{-\theta}}{1!}\right)^{155} \left(\frac{\theta^2 e^{-\theta}}{2!}\right)^{171} \left(\frac{\theta^3 e^{-\theta}}{3!}\right)^{143} \left(\frac{\theta^4 e^{-\theta}}{4!}\right)^{79} \\ &\quad \times \left(\frac{\theta^5 e^{-\theta}}{5!}\right)^{57} \left(\frac{\theta^6 e^{-\theta}}{6!}\right)^{14} \left(\frac{\theta^7 e^{-\theta}}{7!}\right)^{6} \left(\frac{\theta^8 e^{-\theta}}{8!}\right)^{2} \left(\sum_{y=9}^{\infty} \frac{\theta^y e^{-\theta}}{y!}\right)^{0} \end{aligned}$$

or more simply (ignoring constants with respect to θ)

$$\begin{aligned} L(\theta) &= \theta^{0(69)+1(155)+2(171)+3(143)+4(79)+5(57)+6(14)+7(6)+8(2)} \\ &\quad \times e^{-(69+155+171+143+79+57+14+6+2)\theta} \\ &= \theta^{1669} e^{-696\theta} \quad \text{for } \theta \geq 0 \end{aligned}$$

The log likelihood function is

$$l(\theta) = 1669 \log \theta - 696\theta \quad \text{for } \theta > 0$$

and

$$l'(\theta) = \frac{1669}{\theta} - 696 = \frac{1669 - 696\theta}{\theta} = 0 \quad \text{if } \theta = \frac{1669}{696} \approx 2.3980$$

The maximum likelihood estimate of θ is $\hat{\theta} = 1669/696$ which is the sample mean.

(e) The expected frequencies are calculated using

$$e_y = 696 \frac{\left(\frac{1669}{696}\right)^y e^{-1669/696}}{y!} \quad \text{for } y = 0, 1, \ldots, 8$$

and are given in the table below rounded to 2 decimal places:

Number of Points in a Game: y	Observed Number of Games with y points: f_y	Expected Number of Games with y points: e_y
0	69	63.27
1	155	151.71
2	171	181.90
3	143	145.40
4	79	87.17
5	57	41.81
6	14	16.71
7	6	5.72
8	2	1.72
≥ 9	0	0.60
Total	696	696.01

(f) There is quite good agreement between the observed and expected frequencies. Also the sample mean and sample variance are close in value which is what we expect for Poisson data. The Poisson model fits these data well. Recall the homogeneity assumption for the Poisson process. Since a Poisson model fits the data well this suggests that Wayne was a very consistent player when he played with the Edmonton Oilers.

2.12 (a) The parameter θ represents the mean number of points scored in a game by Wayne when he played for the Edmonton Oilers.

(b) See the answer for part (b) for Problem 11.

(c) The sample mean is

$$\begin{aligned} & \frac{1}{783}\sum_{y=0}^{8} y f_y \\ = \; & \frac{1}{783}\left[0\,(219)+1\,(259)+2\,(185)+3\,(90)+4\,(24)+5\,(4)+6\,(2)\right] \\ = \; & \frac{1027}{783} \approx 1.311622 \end{aligned}$$

Since

$$\begin{aligned} & \sum_{y=0}^{8} y^2 f_y \\ = \; & (0)^2\,(219)+(1)^2\,(259)+(2)^2\,(185)+(3)^2\,(90)+(4)^2\,(24)+(5)^2\,(4)+(6)^2\,(2) \\ = \; & 2365 \end{aligned}$$

therefore the sample variance is

$$\frac{1}{782}\left[2365-\frac{(1027)^2}{783}\right] = 1.301745$$

The sample mean and sample variance are very close in value for these data.

(d) The maximum likelihood estimate is equal to the sample mean so $\hat{\theta} = 1027/783 \approx 1.311622$.

(e) The expected frequencies are calculated using

$$e_y = 783\frac{\left(\frac{1027}{783}\right)^y e^{-1027/783}}{y!} \quad \text{for } y = 0, 1, \ldots, 6$$

and are given in the table below rounded to 2 decimal places:

Number of Points in a Game: y	Observed Number of Games with y points: f_y	Expected Number of Games with y points: e_y
0	219	210.93
1	259	276.66
2	185	181.43
3	90	79.32
4	24	26.01
5	4	6.82
6	2	1.82
≥ 7	0	0.33
Total	783	783

(f) There is quite good agreement between the observed and expected frequencies. Also the sample mean and sample variance are close in value which is what we expect for Poisson data. The Poisson model fits these data well.

2.13 (a) Since $P(Y = 1; \theta) = \theta$, the parameter θ represents the probability a randomly chosen family has one child.

(b) Let $F_y =$ the number of families with y children. The probability of observing the data

y	0	1	$\cdots$	$y_{\max}$	$> y_{\max}$	Total
f_y	f_0	f_1	$\cdots$	$f_{\max}$	0	n

is

$$\frac{n!}{f_0!f_1!\cdots f_{\max}!0!}\left(\frac{1-2\theta}{1-\theta}\right)^{f_0}(\theta)^{f_1}\left(\theta^2\right)^{f_2}\cdots\left(\theta^{y_{\max}}\right)^{f_{\max}}\left(\sum_{y=y_{\max}+1}^{\infty}\theta^y\right)^0$$

$$\frac{n!}{f_0!f_1!\cdots f_{\max}!}\left(\frac{1-2\theta}{1-\theta}\right)^{f_0}\prod_{y=1}^{y_{\max}}\theta^{yf_y} \quad \text{for } 0 \leq \theta \leq 0.5$$

If we ignore constants with respect to θ, the likelihood function is

$$L(\theta) = \left(\frac{1-2\theta}{1-\theta}\right)^{f_0}\prod_{y=1}^{y_{\max}}\theta^{yf_y}$$

and the log likelihood is

$$l(\theta) = f_0 \log\left(\frac{1-2\theta}{1-\theta}\right) + \left(\sum_{y=1}^{y_{\max}} yf_y\right)\log\theta \quad \text{for } 0 < \theta < 0.5$$

$$= f_0 \log(1-2\theta) - f_0 \log(1-\theta) + T\log\theta \quad \text{where } T = \sum_{y=1}^{y_{\max}} yf_y$$

Now

$$\begin{aligned} l'(\theta) &= \frac{-2f_0}{1-2\theta} + \frac{f_0}{1-\theta} + \frac{1}{\theta}T \\ &= \frac{1}{\theta(1-\theta)(1-2\theta)}\left[2T\theta^2 - (f_0+3T)\theta + T\right] \end{aligned}$$

and $l'(\theta) = 0$ if

$$\theta = \frac{(f_0+3T) \pm [(f_0+3T)^2 - 8T^2]^{1/2}}{4T}$$

and since

$$\frac{(f_0+3T) + [(f_0+3T)^2 - 8T^2]^{1/2}}{4T} \geq \frac{f_0+3T}{4T} \geq \frac{3}{4} > 0.5$$

therefore

$$\hat{\theta} = \frac{(f_0+3T) - [(f_0+3T)^2 - 8T^2]^{1/2}}{4T}$$

(c) For $y = 1, 2, ...,$ the probability that a randomly selected family has y children is θ^y. Suppose for simplicity there are N families in the population where N is very large. Then the number of families that have y children is $N \times$ (probability a family has y children) $= N\theta^y$ for $y = 1, 2, \ldots$ and there is a total of $yN\theta^y$ children in families of y children and a total of $\sum_{y=1}^{\infty} yN\theta^y$ children altogether. Therefore the probability a randomly chosen child is in a family of x children is

$$\frac{xN\theta^x}{\sum_{x=1}^{\infty} xN\theta^x} = cx\theta^x \quad \text{for } x = 1, 2, \ldots$$

Since

$$\sum_{x=1}^{\infty} cx\theta^x = 1$$

and

$$\sum_{x=1}^{\infty} x\theta^x = \frac{\theta}{(1-\theta)^2}$$

we obtain $c = (1-\theta)^2/\theta$ and

$$P(X = x; \theta) = \frac{(1-\theta)^2}{\theta} x\theta^x = x(1-\theta)^2\theta^{x-1} \quad \text{for } x = 1, 2, \ldots \text{ and } 0 < \theta \leq \frac{1}{2}$$

(d) The probability of observing the given data for model (c) is

$$\frac{33!}{22!7!3!1!}\left[(1-\theta)^2\right]^{22}\left[2(1-\theta)^2\theta\right]^7\left[3(1-\theta)^2\theta^2\right]^3\left[4(1-\theta)^2\theta^3\right] \quad \text{for } 0 < \theta \leq \frac{1}{2}$$

The likelihood function is

$$\begin{aligned} L(\theta) &= (1-\theta)^{2(22+7+3+1)}\theta^{7+2(3)+3} \\ &= \theta^{16}(1-\theta)^{66} \quad \text{for } 0 < \theta \leq \frac{1}{2} \end{aligned}$$

which is maximized for $\theta = 16/(16+66) = 16/82 = 8/41 = 0.1951$.

Since the probability a family has no children is

$$P(Y=0;\theta) = \frac{1-2\theta}{1-\theta} = g(\theta)$$

then by the Invariance Property of maximum likelihood estimates the maximum likelihood of $g(\theta)$ is

$$g(\hat{\theta}) = \frac{1-2\hat{\theta}}{1-\hat{\theta}} = \frac{1-2(0.1951)}{1-0.1951} = 0.7576$$

(e) For these data $f_0 = 0$, $T = 49$. and $l'(\theta) = 49/\theta = 0$ has no solution. Since $l'(\theta) = 49/\theta > 0$ for all $0 < \theta \leq 0.5$, therefore $l(\theta)$ is an increasing function on this interval. Thus the maximum value of $l(\theta)$ occurs at the endpoint $\theta = 0.5$ and therefore $\hat{\theta} = 0.5$.

2.14 (a) Let Y_i = the number of particles emitted in time interval i of length t_i, $i = 1, 2, \ldots, n$. We assume that the Y_i's are independent random variables. The likelihood function is the probability of observing the data $y_1, y_2, \ldots, y_n$ which is

$$\begin{aligned} L(\theta) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n; \theta) \\ &= \prod_{i=1}^{n} P(Y_i = y_i; \theta) = \prod_{i=1}^{n} \frac{(\theta t_i)^{y_i} e^{-\theta t_i}}{y_i!} \\ &= \prod_{i=1}^{n} \frac{(t_i)^{y_i}}{y_i!} \prod_{i=1}^{n} \theta^{y_i} e^{-\theta t_i} \end{aligned}$$

or more simply (ignoring constants with respect to θ)

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i} e^{-\theta t_i} = \theta^{n\bar{y}} e^{-\theta n\bar{t}} \quad \text{for } \theta \geq 0$$

where $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$. The log likelihood function is

$$l(\theta) = (n\bar{y}) \log\theta - (n\bar{t})\,\theta \quad \text{for } \theta > 0$$

and

$$l'(\theta) = \frac{n\bar{y}}{\theta} - n\bar{t} = \frac{n}{\theta}(\bar{y} - \bar{t}\theta) = 0$$

if

$$\theta = \frac{\bar{y}}{\bar{t}}$$

so $\hat{\theta} = \frac{\bar{y}}{\bar{t}}$ is the maximum likelihood estimate of θ.

(b) Let X = number of intervals of length t with no particles emitted. Then $X \sim \text{Binomial}(n, p)$ where

$$p = P(X = 0; \theta) = \frac{(\theta t)^0 e^{-\theta t}}{0!} = e^{-\theta t}$$

Suppose that x intervals were observed with no particles. Since $X \sim \text{Binomial}(n, p)$ the maximum likelihood estimate of p is $\hat{p} = \frac{x}{n}$. Since $p = e^{-\theta t}$ implies $\theta = \frac{-(\log p)}{t}$ then by the invariance property of maximum likelihood estimates $\hat{\theta} = \frac{-(\log \hat{p})}{t}$.

2.16 (a)

$y_{(1)}$	$q(0.25)$	$q(0.5)$	$q(0.75)$	$y_{(n)}$	IQR	range	$\bar{y}$	s	g_1	g_2
3	16.375	19.5	22	30	5.625	27	19.13	4.4498	−0.50	4.32

(b) The proportion of observations in the interval $[\bar{y} - s, \bar{y} + s] = [14.68, 23.58]$ is $71/100 = 0.71$. If $Y \sim G(\mu, \sigma)$ then

$$\begin{aligned} P(Y \in [\mu - \sigma, \mu + \sigma]) &= P(|Y - \mu| \leq \sigma) = P\left(\frac{|Y - \mu|}{\sigma} \leq 1\right) \\ &= P(|Z| \leq 1) = 2P(Z \leq 1) - 1 \quad \text{where } Z \sim N(0, 1) \\ &= 2(0.84134) - 1 = 0.68268 \\ &\approx 0.68 \end{aligned}$$

The observed proportion of observations in the interval which is 0.71 is slightly higher than what would be expected for Gaussian data (0.68).

(c) A boxplot and qqplot of the data are given in Figures 2.4 and 2.5.

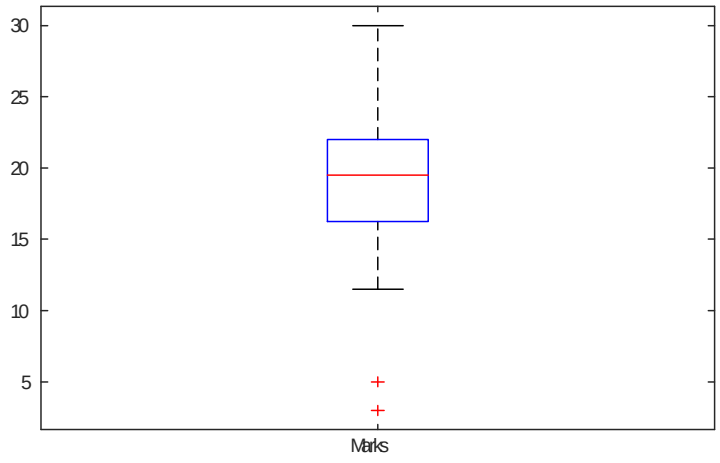


Figure 2.4: Boxplot for tutorial test 1 data

(d) For Gaussian data we expect the sample skewness to be close to zero and the sample mean and sample median to be approximately equal. For these data

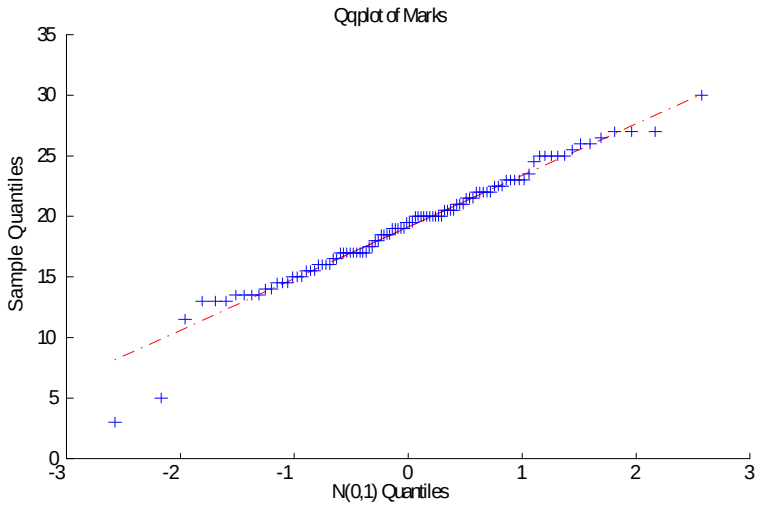


Figure 2.5: **Qqplot for tutorial test data**

the sample skewness $= -0.50$ and the sample median $= 19.5 >$ sample mean $= 19.14$. Both of these results indicate that the data are not symmetric but **slightly** skewed to the left. This is also evident in the boxplot in which neither the box nor the whiskers are divided approximately in half by the sample median.

For Gaussian data we expect the sample kurtosis to be close to 3. The sample kurtosis for these data equals 4.32 which indicates that there are more observations in the tails then would be expected for Gaussian data.

In the list of observations as well as the boxplot we observe two extreme observations, 3 and 5, which are also evident in the qqplot (see lower left hand corner of graph). These extremes have a large influence on the sample mean as well as on the sample skewness and sample kurtosis. If the sample mean, sample median, sample skewness and sample kurtosis were recalculated with these observations removed then the values of these numerical summaries would be more in agreement with what we expect to see for Gaussian data.

For Gaussian data we expect the points to be scattered about a straight line although the points at both ends may lie further from the straight line since the quantiles of the Gaussian distribution change more rapidly in both tails of the distribution. Except for the outliers, the points in this qqplot lie are scattered about a straight line.

For these data the $IQR = 5.75$ is close in value to $1.349s = 1.349\,(4.4498) = 6.00$ which is what we expect for Gaussian data.

The proportion of observations in the interval $[\bar{y} - s, \bar{y} + s]$ is slightly higher than we would expect for Gaussian data. This also agrees with the sample kurtosis value of 4.3 being larger than 3.

Overall, except for the two outliers, a Gaussian model fits these data well. It would be a good idea to do any formal analyses of the data with and without the outliers to determine the effect of these outliers on the conclusions of the analyses. Note also that the variate we are modeling here is a discrete variate since there were only a finite number of possible marks out of 30 that were assigned in this test. Therefore the model is only approximate.

2.17 (a)

$y_{(1)}$	$q(0.25)$	$q(0.5)$	$q(0.75)$	$y_{(n)}$	IQR	range	$\bar{y}$	s	g_1	g_2
142	156	160	164	178	8	36	159.77	6.03	0.13	3.16

(b) The number of observations in the interval $[\bar{y}-s, \bar{y}+s] = [153.75, 165.80]$ is 244 or 69.5% and the number of observations in the interval $[\bar{y}-2s, \bar{y}+2s] = [147.72, 171.83]$ is 334 or 95.2%.
If $Y \sim G(\mu, \sigma)$ then $P(Y \in [\mu-\sigma, \mu+\sigma]) = 0.68268$ and $P(Y \in [\mu-2\sigma, \mu+2\sigma]) = 0.9545$.
The observed and expected proportions and very close to what one would expect if the data were Normally distributed.

(c) The frequency histogram and superimposed Gaussian probability density function are given in the top left graph in Figure 2.6.

(d) The empirical cumulative distribution function and superimposed Gaussian cumulative distribution function are given in the top right graph in Figure 2.6.

(e) The boxplot is given in the bottom left graph in Figure 2.6. The qqplot is given in the bottom right graph in Figure 2.6. The "steplike" behaviour of the plot is due to the rounding of the data to the nearest centimeter.

(f) Note that the sample size of this data set is reasonably large ($n = 351$). The sample skewness for these data is 0.13 while for Gaussian data we expect a sample skewness close to 0. The sample kurtosis for these data is 3.16 while for Gaussian data we expect a sample kurtosis close to 3. Both the sample skewness and the sample kurtosis are reasonably close to what we expect for Gaussian data.
For Gaussian data we expect the IQR to be close to $1.349\sigma \approx 1.349s = 8.13$ which is very close to $IQR = 8$.
All the numerical summaries indicate good agreement with the model. The relative frequency histogram has the shape of a Gaussian probability density function. The empirical cumulative distribution function and the Gaussian cumulative distribution function also have similar shapes. The boxplot is consistent with Gaussian data. The points in the qqplot are scattered along a straight line with more variability at both ends which is what we expect for Gaussian data.
The agreement between the observed and expected summaries for this data set indicated above is what we expect for a data set of this size. A Gaussian model seems very reasonable for these data.

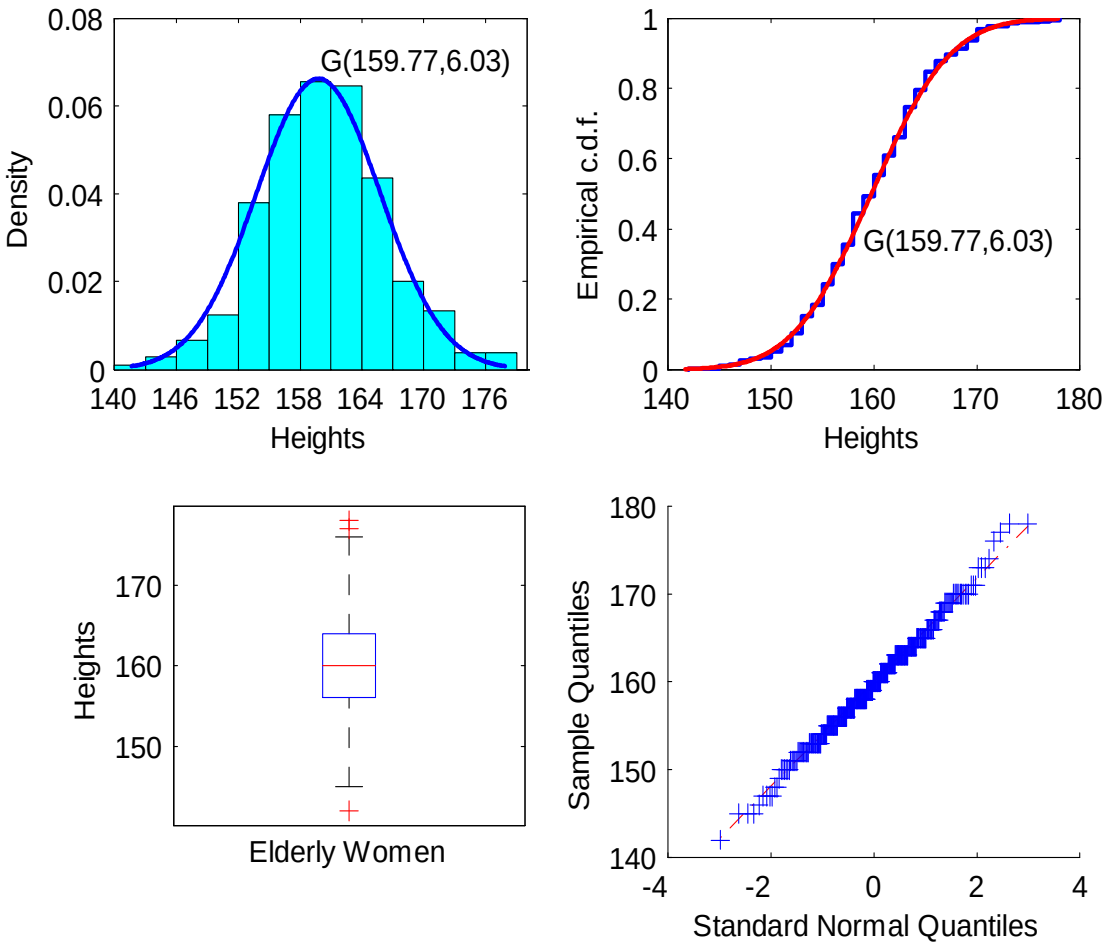


Figure 2.6: Plots for Heights of Elderly Women

2.18 (a) $\hat{\mu} = 1.744$, $\hat{\sigma} = 0.0664$ (M) $\hat{\mu} = 1.618$, $\hat{\sigma} = 0.0636$ (F)

(b) 1.659 and 1.829 (M) 1.536 and 1.670 (F)

(c) 0.098 (M) and 0.0004 (F)

(d) $11/50 = 0.073$ (M) 0 (F)

2.19 See Figure 2.7. Note that the qqplot for the $\log y_i$'s is far more linear than for the y_i's indicating that the Normal model is more reasonable for the transformed data.

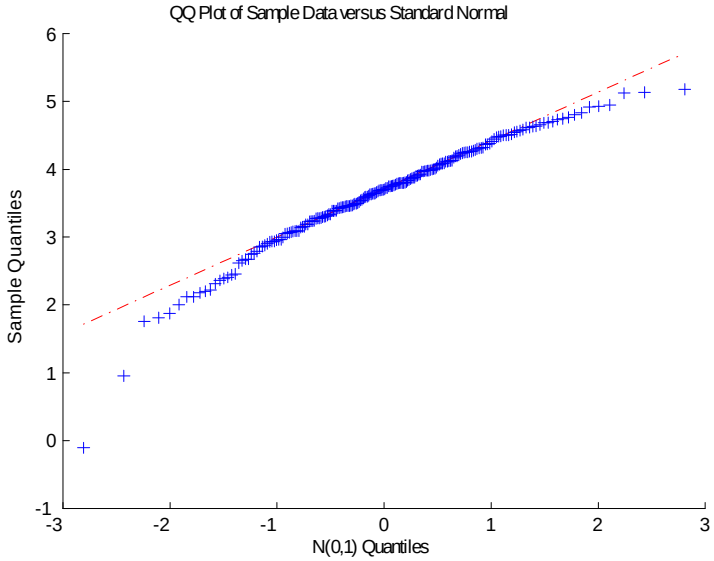


Figure 2.7: Qqplot of log brake pad lifetimes

2.20 (a) If they are independent $P(S \text{ and } H) = P(S)P(H) = \alpha\beta$. The others are similar.

(b) The Multinomial probability function evaluated at the observed values is

$$L(\alpha, \beta) = \frac{100!}{20!15!22!43!} (\alpha\beta)^{20} [\alpha(1-\beta)]^{15} [(1-\alpha)\beta]^{22} [(1-\alpha)(1-\beta)]^{43}$$

or more simply (ignoring constants with respect to α and β)

$$L(\alpha, \beta) = \alpha^{35}(1-\alpha)^{65}\beta^{42}(1-\beta)^{58} \quad \text{for } 0 \le \alpha \le 1,\ 0 \le \beta \le 1$$

The log likelihood is

$$l(\alpha, \beta) = 35\log(\alpha) + 65\log(1-\alpha) + 42\log(\beta) + 58\log(1-\beta) \quad \text{for } 0 < \alpha < 1,\ 0 < \beta < 1$$

Setting the derivatives to zero gives the maximum likelihood estimates $\hat{\alpha} = 0.35$ and $\hat{\beta} = 0.42$.

(c) The expected frequencies are

$$100\hat{\alpha}\hat{\beta},\ 100\hat{\alpha}\left(1-\hat{\beta}\right),\ 100\left(1-\hat{\alpha}\right)\hat{\beta},\ 100\left(1-\hat{\alpha}\right)\left(1-\hat{\beta}\right)$$

or 14.7, 20.3, 27.3, 37.7 which can be compared with $20, 15, 22, 43$. The observed and expected frequencies do not appear to be very close. In Chapter 7 we will see how to construct a formal test of the model.

2.21 See Figures 2.8 and 2.9.

For the data sets of size $n = 100$ the points lie closely to a straight line. For the data

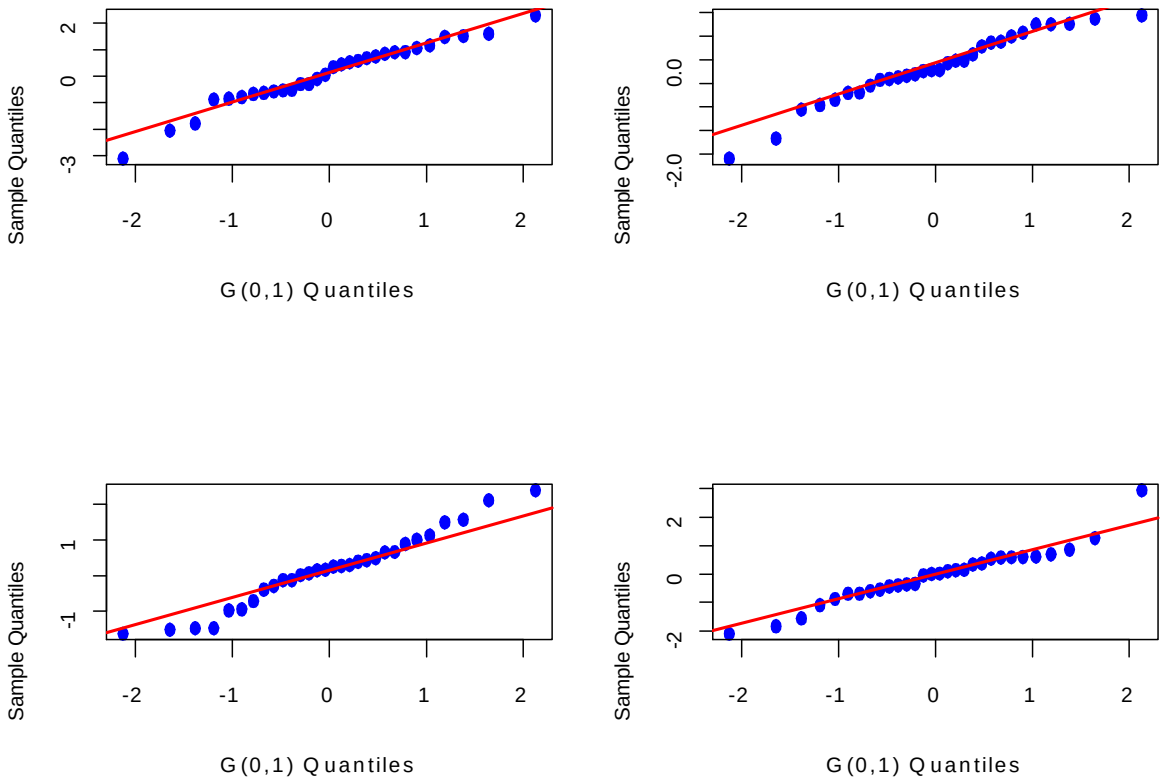


Figure 2.8: Qqplots for Gaussian data for sample size $n = 30$

sets of size $n = 30$ the points lie roughly along a straight line but not as closely to a straight line as for $n = 100$. This means that for smaller data sets it would be more difficult to decide on the basis of just a qqplot how reasonable the Gaussian model is. Note also that, even for datasets of size $n = 100$, some unusual behaviour can be observed.

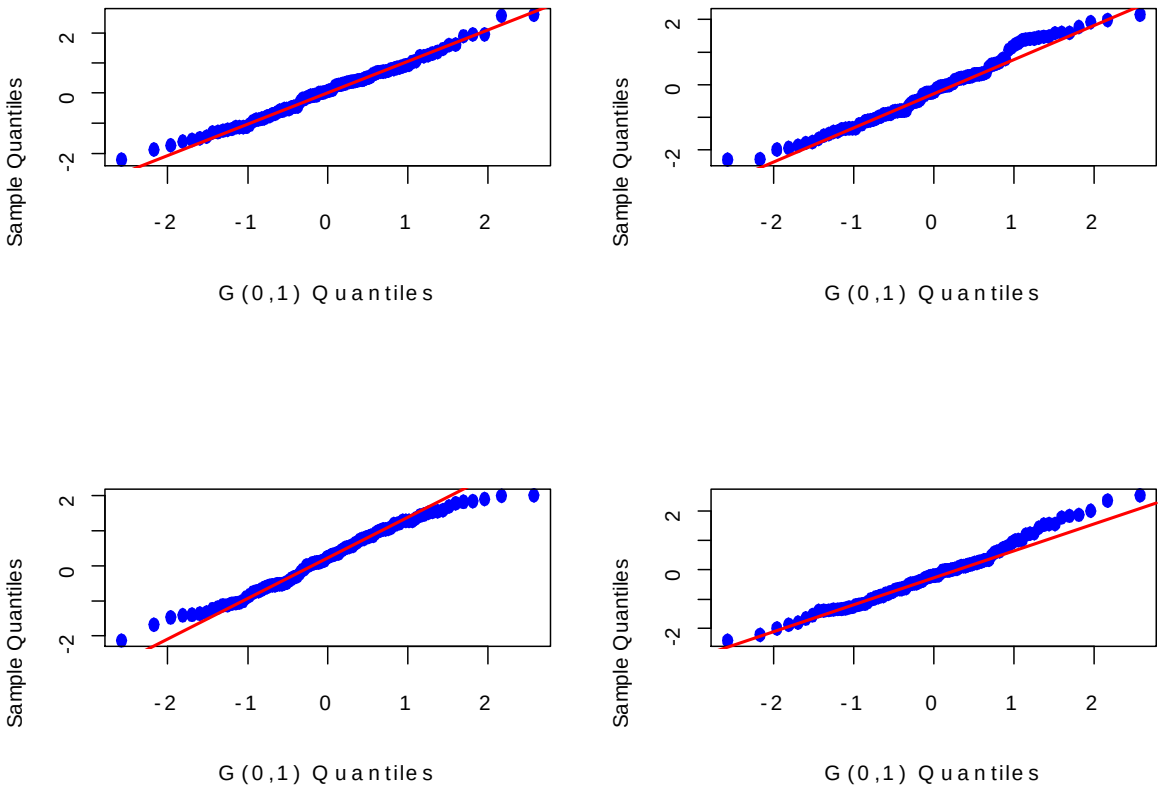


Figure 2.9: Qqplots for Gaussian data for sample size $n = 100$

2.22 See Figures 2.10 and 2.11.

Note that the plots are grouped into four just to save space but it is best to view these graphs as separate plots in order to obtain the information provided by the plot.

Qqplot 1 looks S-shaped which indicates the distribution of the data is symmetric and the sample skewness would be close to 0. The symmetry of the qqplot indicates that the distribution of the data would be symmetric about the sample median which is close to 0.45. This S-shape also indicates that the sample kurtosis would be less than 3. A symmetric model with lighter tails than the Gaussian distribution (e.g. Uniform) would provide a better fit to the data.

Qqplot 2 looks somewhat S-shaped which indicates the distribution of the data is symmetric and the sample skewness would be close to 0. This S-shape also indicates that the sample kurtosis would be less than 3. The sample median is approximately 0.3. There are many observations with values close to 1 and many observations with

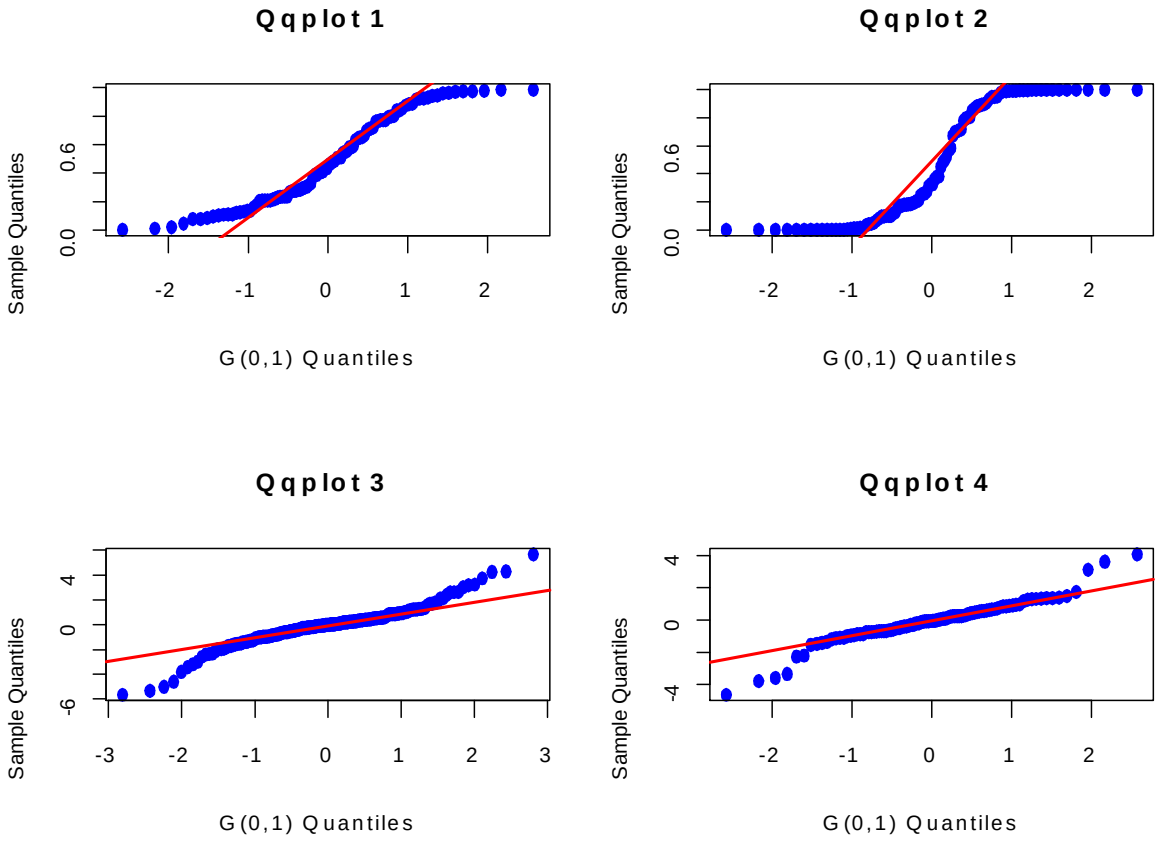


Figure 2.10: QQplots for Problem 21

values close to 0. The other observations are more uniformly distributed between 0 and 1. A symmetric model with more observations at both ends of the interval $[0, 1]$ and fewer observations in the middle (e.g. a U shaped probability density function on the interval $[0, 1]$) would provide a better fit to the data.

Qqplot 3 looks like an S-shape which had been turned upside down which indicates the underlying distribution is symmetric with skewness close to 0. The symmetry of the qqplot indicates that the distribution of the data would be symmetric about the sample median which is close to 0. The underlying distribution has more observations in the tails than is expected with Gaussian data so the sample kurtosis will be greater than 3.

Qqplot 4 also looks like an S-shape which had been turned upside down which indicates the underlying distribution is symmetric with sample skewness close to 0. The symmetry of the qqplot indicates that the distribution of the data would be symmetric about the sample median which is close to 0. The underlying distribution has more

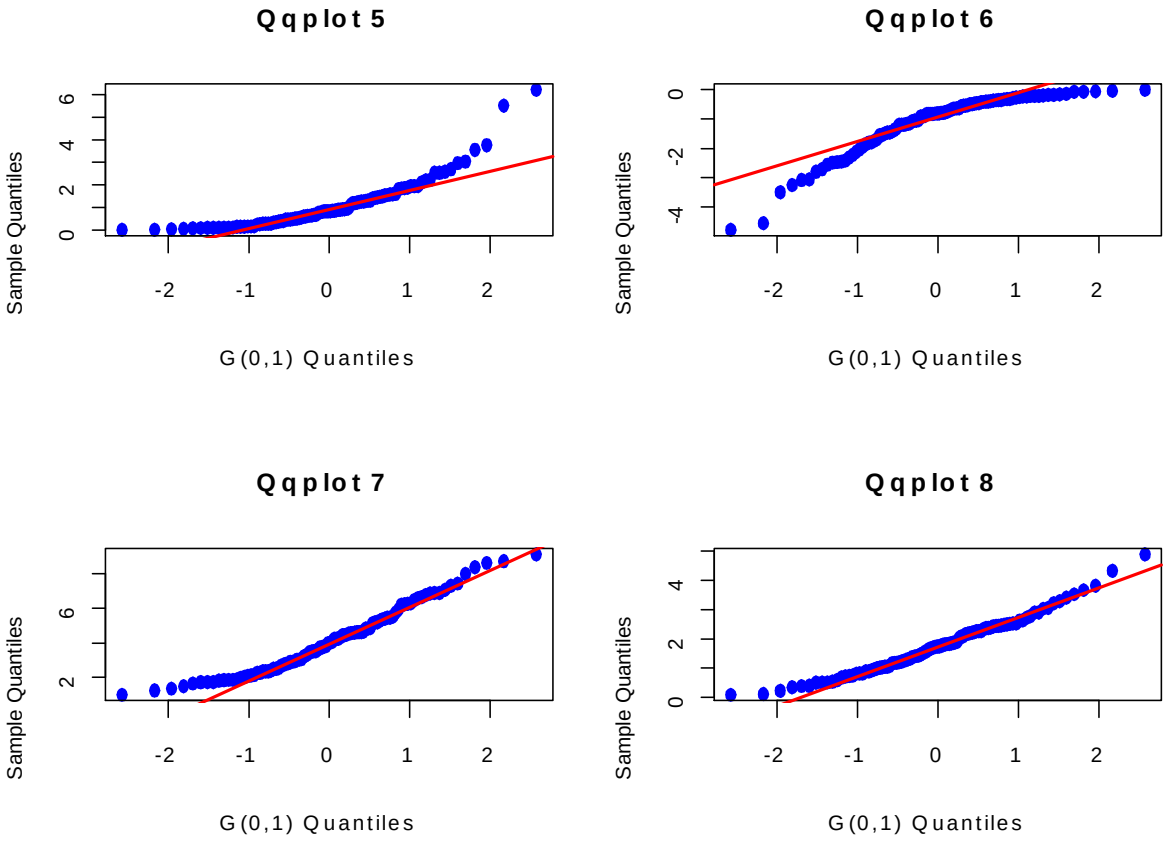


Figure 2.11: Qqplots for Problem 21

observations in the tails than is expected with Gaussian data so the sample kurtosis will be greater than 3.

Qqplot 5 is very U-shaped which indicates the distribution of the data is not symmetric but has a long right tail and the sample skewness would be positive. A non-symmetric model with a long right tail (e.g. Exponential) would provide a better fit to the data.

Qqplot 6 is shaped like an upside down U which indicates the distribution of the data is not symmetric but has a long left tail and the sample skewness would be negative. A non-symmetric model with a long left tail would provide a better fit to the data.

Qqplot 7 is slightly U-shaped which indicates the underlying distribution is not symmetric and has a slightly longer right tail so the sample skewness is positive.

Qqplot 8 is slightly U-shaped which indicates the underlying distribution is not symmetric and has a slightly longer right tail so the sample skewness is positive.

2.23 (a) The median of the $G(0,1)$ distribution is $m = 0$. Reading from the qqplot the sample quantile on the $y-$axis which corresponds to 0 on the $x-$axis is approximately equal to 1.0 so the sample median for these data is approximately 1.0.

(b) To determine $q(0.25)$ for the these data we note that $P(Z \leq -0.6745) = 0.25$ if $Z \sim N(0,1)$. Reading from the qqplot the sample quantile on the $y-$axis which corresponds to -0.67 on the $x-$axis is approximately equal to 0.4 so $q(0.25)$ is approximately 0.4. To determine $q(0.75)$ for the these data we note that $P(Z \leq 0.6745) = 0.75$ if $Z \sim N(0,1)$. Reading from the qqplot the sample quantile on the $y-$axis which corresponds to 0.67 on the $x-$axis is approximately equal to 1.5 so $q(0.75)$ is approximately 1.5. The IQR for these data is approximately $1.5 - 0.4 = 1.1$.

(c) The range of the data can be determined approximately by looking at the height of the minimum observation which is approximately 0 and the height of the maximum observation which is approximately 2 so the range is approximately $2 - 0 = 2$.

(d) The frequency histogram of the data would be approximately symmetric about the sample mean.

(e) The frequency histogram would most resemble a Uniform probability density function.

2.24

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(y_i;\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} \quad \text{if } \theta \geq y_i \;\; i = 1,2,\ldots,n \\ &= \frac{1}{\theta^n} \quad \text{if } \theta \geq y_{(n)} = \max(y_1, y_2, \ldots, y_n) \end{aligned}$$

where θ^{-n} is a decreasing function of θ. Note also that $L(\theta) = 0$ for $0 < \theta < y_{(n)}$. Therefore the maximum value of $L(\theta)$ occurs at $\theta = y_{(n)}$ and therefore the maximum likelihood estimate of θ is $\hat{\theta} = y_{(n)}$.

2.25 (a)

$$P(Y > c;\theta) = \int_c^{\infty} \frac{1}{\theta} e^{-y/\theta} dy = e^{-c/\theta}$$

(b) For the i'th piece that failed at time $y_i < c$, the contribution to the likelihood is $\frac{1}{\theta} e^{-y_i/\theta}$. For those pieces that survive past time c, the contribution to the likelihood is the probability of the event, $P(Y > c;\theta) = e^{-c/\theta}$. Therefore the likelihood is

$$L(\theta) = \left(\prod_{i=1}^{k} \frac{1}{\theta} e^{-y_i/\theta} \right) \left(e^{-c/\theta} \right)^{n-k} \quad \text{for } \theta > 0$$

and the log relative likelihood is

$$l(\theta) = -k\log(\theta) - \frac{1}{\theta}\sum_{i=1}^{k} y_i - (n-k)\frac{c}{\theta} \quad \text{for } \theta > 0$$

Solving $l'(\theta) = 0$ we obtain the maximum likelihood estimate,

$$\hat{\theta} = \frac{1}{k}\left[\sum_{i=1}^{k} y_i + (n-k)c\right]$$

(c) When $k = 0$ and $c > 0$ the maximum likelihood estimator is $\hat{\theta} = \infty$. In this case there are no failures in the time interval $[0, c]$ and this is more likely to happen when $E(Y) = \theta$ is very large.

2.26 (a) If there is adequate mixing of the tagged animals, the number of tagged animals caught in the second round is a random sample selected without replacement so follows a hypergeometric distribution (see the STAT 230 Course Notes).

(b)
$$\frac{L(N+1)}{L(N)} = \frac{(N+1-k)(N+1-n)}{(N+1-k-n+y)(N+1)}$$
and $L(N)$ reaches its maximum within an integer of kn/y.

(c) The model requires sufficient mixing between captures that the second stage is a random sample. If they are herd animals this model will not fit well.

2.27 The likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{(\alpha + \beta x_i)^{y_i}}{y_i!} e^{-(\alpha+\beta x_i)}$$

or more simply (ignoring constants with respect to α and β)

$$L(\alpha, \beta) = \prod_{i=1}^{n} (\alpha + \beta x_i)^{y_i}\, e^{-(\alpha+\beta x_i)}$$

The log likelihood is

$$l(\alpha, \beta) = \sum_{i=1}^{n} \left[y_i(\alpha + \beta x_i) - e^{(\alpha+\beta x_i)}\right]$$

To maximize we set the partial derivatives equal to zero and solve

$$\frac{\partial}{\partial \alpha} l(\alpha, \beta) = \sum_{i=1}^{n} \left[y_i - e^{(\alpha+\beta x_i)}\right] = 0$$
$$\frac{\partial}{\partial \beta} l(\alpha, \beta) = \sum_{i=1}^{n} x_i \left[y_i - e^{(\alpha+\beta x_i)}\right] = 0$$

For a given set of data we can solve this system of equations numerically but not explicitly.

SOLUTIONS TO CHAPTER 3 PROBLEMS

3.1 (a) This study would best be described as a sample survey since the population of interest (university students in Ontario in 2019) is finite. The purpose of the study was to learn about the attributes of this population and the researchers did not attempt to change or control any of the variates for the sampled units.

(b) The target population is the set of all university students in Ontario at the time of the study (or in 2019).

(c) The study population is the set of university students living in the Kitchener Waterloo region in September 2019.

(d) The sampling protocol consisted of taking a random sample of 250 students attending a specific Laurier-Waterloo football game in September 2019. The sample consists of the 250 students and the sample size is 250.

(e) One variate is whether the student is male or female. The other variate is whether the student agrees or disagrees with the statement "I have significant trouble paying my bills." Both variates are categorical variates.

(f) One attribute is the proportion of male students who agree with the statement "I have significant trouble paying my bills." Another attribute is the proportion of female students who agree with the statement

(g) There may be systematic differences between KW university students and the population of Ontario university students, for example, university students in Toronto and Thunder Bay may have different financial worries then KW university students. This could be a possible source of study error since this could result in the proportion of students who agree with the statement in the target population (Ontario university students) to be different than the proportion of students in the study population (KW university students).

(h) There may be systematic differences between KW university students and students who attend a particular Laurier-Waterloo football game, for example, university students at a football game may have different financial worries since they are attending a football game and not working at a part-time job. This could

be a possible source of sample error since this could result in the proportion of students who agree with the statement in the study population (KW university students) to be different than the proportion of students in the sample (students at the football game).

(i) An estimate of the proportion of males in the study population who agree with the statement "I have significant trouble paying my bills." based on the sample is 68/145 or approximately 47%. An estimate of the proportion of females in the study population who agree with the statement "I have significant trouble paying my bills." based on the sample is 42/105 or 40%.

(j) The most serious limitation is that the study population only consisted of KW university students which could lead to study error. Another serious limitation is that the sample only consisted of students who attended a particular university football game which could lead to sample error.

3.2 (a) This study would best be described as an experimental study because the researchers determined, using randomization, which patient was in the aspirin group and which patient was in the standard care alone group.

(b) The Problem was to determine if the aspirin treatment reduced 28−day mortality from Covid-19 as compared to standard care alone treatment.

(c) This is a causative Problem because this was an experimental study in which the researchers wanted to determine if the aspirin treatment caused a reduction in the 28−day mortality from Covid-19 as compared to standard care alone treatment.

(d) The target population for this study could reasonably described as people at least 18 years of age who were hospitalized for Covid-19 between November 2020 and March 2021 in all hospitals in the United Kingdom. (Other answers are possible. The target population could be increased to include adults in other countries which have a similar health care system to the British health care system. All adults in the world would not be a suitable targer population.)

(e) The study population for this study consists of people at least 18 years of age who were hospitalized for Covid-19 between November 2020 and March 2021 at the 177 hospitals in the United Kingdom.

(f) Between November 2020 and March 2021, doctors at 177 hospitals in the United Kingdom who had adult patients who tested positive for COVID-19 and needed to be hospitalized, were encouraged to recruit these patients to the study. Patient consent was required to be in the study. The sample consisted of the $7351 + 7541 = 14892$ patients who were enroled in the study. The sample size was 14892.

(g) One important variate is which treatment the patient received, aspirin or standard care alone. This variate is categorical. Another important variate is

whether or not the patient died of Covid-19 within 28 days. This variate is categorical.

(h) An important attribute of interest is the proportion of patients in the aspirin group who died of Covid-19 within 28 days. Another important attribute of interest is the proportion of patients in the standard care alone group who died of Covid-19 within 28 days.

(i) Here are other variates mentioned in the article:
time to discharge from hospital - this is a continuous variate
whether a patient was on invasive mechanical ventilation at randomisation - this is a categorical variate
time to invasive mechanical ventilation for patients not on invasive mechanical ventalation at randomisation - this is a continuous variate
time to death for patients not on invasive mechanical ventalation at randomisation - this is a continuous variate
identity of the recruiting physician for the patient - this is a categorical variate
patient's age at recruitment - this is a continuous variate
patient's sex - this is a categorical variate
Covid-19 onset date - this is a discrete variate
Covid-19 severity - this is a ordinal variate

(j) Attributes for the variates in (i):
time to discharge from hospital - mean (or median) time to discharge
whether a patient was on invasive mechanical ventilation at randomisation - proportion of patiens on invasive mechanical ventilation at time of recruitment
time to invasive mechanical ventilation for patients not on invasive mechanical ventalation at randomisation - mean (or median) time to invasive mechanical ventilation
time to death for patients not on invasive mechanical ventalation at randomisation - mean (or median) time to death
identity of the recruiting physician for the patient - proportion of patients recruited by a particular physician
patient's age at recruitment - mean (or median) age
patient's sex - proportion of patients who are females
Covid-19 onset date - run chart of onset dates
Covid-19 severity - proportion of patients with a certain level of severity

(k) Only 177 hospitals in the UK were involved in the study. Suppose the hospitals in the study were systematically different in the care offered to COVID-19 patients compared to other hospitals in the UK. This could be a possible source of study error since this could result in the proportion of patients dying of Covid-19 within 28 days in the target population being different than the proportion of patients dying of Covid-19 within 28 days in the study population.

(l) Doctors could decide which patients to recruit. Patients needed to give their consent, that is, they could decide whether to be part of the study or not. Therefore the sample may not be a representative sample from the study population. For example, doctors may not have referred their patients who were extremely ill. This may result in the proportion who die within 28 days being lower in the sample as compared to the proportion who die within 28 days in the study population.

(m) If the hospital staff member who entered the data, incorrectly entered which treatment the patient received (or whether or not they died within 28 days) then this is an example of measurement error.

(n) It is important for the researchers to randomly assign the participants to the two different groups. If the researchers were allowed to choose the groups they might inadvertently assign the patients who are more ill to the treatment they believe is better. A difference in deaths rates between the groups could not then be attributed to just the treatment.

(o) The control group which is the group that received the standard care acts as a baseline. It allows the researchers to determine the effectiveness of the aspirin treatment over and above the standard care.

(p) An estimate, based on the sample, of the proportion of adults in the study population who would die within 28 days if they received standard care only is 17%. An estimate, based on the sample, of the proportion of adults in the study population who would die within 28 days if they received standard care plus aspirin is 17%.

(q) The limitation is that the study was conducted in the UK which has a universal healthcare system and therefore the conclusions of the is study might not hold in other countries with different healthcare systems.

3.3 (a) This study would best be described as an experimental study because the researchers determined, using randomization, which participants were assigned to which training groups.

(b) The Problem is to compare generalized cognitive improvement among people who engage in brain training and those who do not, as well as to compare generalized cognitive improvement between groups who engage in different types of brain training activity.

(c) This is a causative problem as the researchers are interested in whether brain training causes an improvement in generalized cognitive improvement.

(d) A suitable target population for this study would be healthy adults aged $18-60$ living in the United Kingdom at the time of the study or healthy adults aged $18-60$ living in the European Union at the time of the study.

(e) A suitable study population for this study would be viewers of the television programme 'Bang Goes the Theory' aged 18-60 at the time of the study.

(f) The sampling protocol consisted of inviting adults (aged 18+) viewers of the BBC popular science programme 'Bang Goes The Theory' to participate in a six-week online study of brain training. The sample consisted of the participants who registered for the study, completed both benchmarking assessments as well as at least two full training sessions during the six-week period. The sample size was $11,430$.

(g) A variate which was recorded is the group a participant was assigned to (experimental group 1 or 2, or the control group) which is a categorical variate.
In the study, scores for four different tests (reasoning, verbal short-term memory, spatial working memory and paired-associates learning) were acquired at baseline and after six weeks. A variate would be associated with each of the four different tests and each time point (baseline and six weeks) for a total of eight different variates. Each of these scores would be a discrete variate.
Scores were also recorded for the six training tasks the participants were asked to complete three times a week over the six week period. A discrete variate would be associated with the score for each of the $6 \times 3 \times 6 = 108$ different tasks.

(h) The following are just four examples of attributes. There are many other examples.
One attribute is the difference between the mean score on the baseline test for reasoning and the mean score after six weeks of training tasks of the type used for group 1.
Another attribute is the difference between the mean score on the baseline test for verbal short-term memory and the mean score after six weeks of training tasks of the type used for group 2.
Another attribute is the difference between the mean score on the baseline test for spatial working memory and the mean score after six weeks of training tasks of the type used for the control group.
Another attribute is the standard deviation in scores on the baseline test for reasoning.

(i) Consider the attribute which is the mean score on the benchmark test for reasoning. Since the study population consisted of viewers of the television programme, the fact that they watched this programme could mean that their reasoning ability is different than the adults in the target population. This could be a possible source of study error since this could result in the average score on the benchmark test for reasoning for the viewers of the television programme to be different than the average score on the benchmark test for reasoning for adults in the target population.

(j) Consider the attribute which is the difference between the mean score on the

baseline test for verbal short-term memory and the mean score after six weeks of training tasks of the type used for group 2. The adults who volunteered to participate in the study were motivated to complete all the tests and training, they may be systematically different than the adults in the study population in terms of their ability to improve their short term memory. This could be a possible source of study error since this could result in the mean score on the baseline test for verbal short-term memory and the mean score after six weeks of training tasks of the type used for group 2 for the participants to be different from the mean score on the baseline test for verbal short-term memory and the mean score after six weeks of training tasks of the type used for group 2 for adults in the study population.

(k) A possible source of measurement error is that participants may cheat on the assessments (for example, by asking a friend for help). This could mean that their scores are not an accurate reflection of what the researchers wish to measure, which is their scores without any outside help.

(l) Randomization ensures that any difference in cognitive improvement is due to the group assignment, and not due to other potential confounders (e.g., if individuals could choose which type of brain training exercise to do, we may find individuals who prefer one type of exercise may be more/less likely to then benefit from that exercise).

(m) The control group allows the researchers to compare those who engaged in brain training exercises with those who did not.

(n) The most important limitation is that the participants in this study were volunteers who were keen enough to do all the testing and training exercises. Therefore the conclusions of this study may not apply to the study population.
Another important limitation is that the study population consisted of adults who watched the television programme. Therefore the conclusions of this study may not apply to the target population.

3.4 (a) This study would best be described as an experimental study since the researchers are in control of which schools received the regular curriculum and which schools are using the JUMP program.

(b) The Problem is to compare the performance in math of students at Ontario schools using the current provincial curriculum as compared to the performance in math of students at Ontario schools using the JUMP math program.

(c) This is a causative problem since the researcher are interested in whether the JUMP program causes better student performance in math.

(d) The target population is all elementary students in Ontario public schools at the time of the study or all elementary students in Ontario public schools at the time of the study and into the future.

(e) The study population is all Ontario elementary students in Grades 2 and 5 in public schools at the time the study.

(f) The sampling protocol is to select the schools in one school board in Ontario and conduct the study using the students in those schools who were in Grades 2 and 5 at the time the study began. The researchers did not indicate how this school board was chosen.

(g) One variate is whether a student receives the standard curriculum or the standard curriculum plus the JUMP program which is a categorical study.
Another variate is the score on a classroom test used to study the impact of the two programs on student learning. This variate is a discrete variate since scores only take on a finite number of countable values.
Another variate is the score on a lab test used to study the impact of the two programs on student learning. This is a discrete variate since scores only take on a finite number of countable values.

(h) For the variate which indicates what program the student received, an attribute would be the proportion of student who receive the standard circulum.
For the variate which is the score on a classroom test, an attribute would be the mean score on the test or the difference between the mean score on a classroom test and the mean score on the same classroom test after using the JUMP program.
For the variate which is the score on a lab test, an attribute would be the mean score on the test or the difference between the mean score on a lab test and the mean score on the same lab test after using the JUMP program.

(i) A possible source of study error is that the ability of students in Grades 2 and 5 to learn math skills might be different than students in other grades and therefore there might be a systematic difference in the mean scores on the classroom tests for the different grades.

(j) A possible source of sample error is that the schools in the chosen school board may not be representative of all the elementary schools in Ontario. For example, the schools in the chosen board may have larger class sizes compared to other schools. Student in larger classes may not receive as much help to improve their math skills as students in smaller classes. Another example is that the chosen school board might be in a low income area of a city. Students from low income families may respond differently to changes in the Math curriculum as compared to students from middle class families. These sources of sample may cause a systematic difference in mean scores on a classroom test between the sample and study population.

(k) The most serious limitation to this study is that only schools in one school board in Toronto will be used in the sample so the conclusions may be subject to both study error and sample error as described above.

3.5 (a) This study would best be described as an experimental study because the researchers controlled, using randomization, which students were assigned to the racing-type game and which students were assigned to the game of solitaire.

(b) The Problem is to determine whether playing racing games makes players more likely to take risks in a simulated driving test.

(c) This is a causative type Problem because the researchers were interested in whether playing racing games as compared to playing a game like solitaire caused players to take more risks in the driving test.

(d) A suitable target population for this study is young adults living in China at the time of the study OR students attending university in China at the time of the study.

(e) A suitable study population for this study is students attending Xi'an Jiaotong University at the time of the study.

(f) From the article it appears that the researchers recruited volunteers for the study. The article does not indicate how these volunteers were obtained.

(g) One variate is whether the student played the racing-type driving game or the game of solitaire. This is a categorical variate.
A different variate would be associated with each of the 24 "risky" videotaped road-traffic situations on a computer screen. Each variate would be how long, in seconds, the student waited to hit the "stop" key in the Vienna Risk-Taking Test for a given road-traffic situation viewed. Each of these 24 variates is a continuous variate.

(h) An attribute can be associated with each of the 24 variates given in (g). Each attribute would be the mean difference in the time to hit the "stop" key in the Vienna Risk-Taking Test between adults who play racing games compared to adults who play neutral games for each of the 24 road-traffic situations.

(i) Young adults living in China (the target population) might be systematically different than the students attending university in China at the time of the study (the study population) since students who attend university are more educated and more intelligent. This could be a possible source of study error since this could result in differences in the means for the different road-traffic situations between the target population and the study population.

(j) The sample consisted of volunteers. The sample was not a random sample of students from the Xi'an Jiaotong University. Students who volunteer for such studies may be higher risk takers on average as compared to non-volunteers who might be more conservative. This could be a possible source of study error since this could result in differences in the means for the different road-traffic situations between the study population and the sample.

(k) An estimate based on the data of the mean difference in the time to hit the "stop" key in the Vienna Risk-Taking Test between young adults who play racing games compared to young adults who play neutral games in the study population is $12 - 10 = 2$ seconds.

3.6 (a) This study would best be described as an observational study because the researchers did not attempt to change or control any of the variates for the sampled units and the population/process of interest is not finite.

(b) The Problem is to examine the association between coffee consumption and the risk of mortality (death) in a middle-aged Mediterranean cohort.

(c) This problem would best be described as a descriptive problem. It does appear that the researchers were interested in a causative problem since they wanted to know how coffee consumption was related to mortality. However since this is an observational study in which coffee consumption was not controlled by the researchers, there could be many other explanations for the association which was observed. Note that the title of the article "**Higher coffee consumption associated with lower risk of early death**" makes it clear that the researchers knew that a causal relationship could not be concluded in such a study.

(d) Since this study recruited new participants to the study every year since 1999 it would be best to define a target process. A suitable target process for this study is the set of adults living in a Mediterranean country when the study began and into the future.

(e) A suitable study process for this study is the set of Spanish university graduates when the study began and into the future.

(f) Another variate was whether the person had died during the follow-up period. This is a categorical variate.
Another variate is the number of cups of coffee consumed per day. It is not clear how this variate was collected. If the question was how many cups of coffee do you drink a day on average then this would be a discrete variate. If the question (for example) was how many cups of coffee do you drink a day on average: 0, $1-2$, $3-4$, >4, then this would be a categorical variate.

(g) The attributes are the mortality rates among middle-aged people living in the Midterranean for the different levels of coffee consumption.

(h) There were only Spanish university graduates in the study process. University graduates may have different coffee consumption habits on average and they may also have better overall health habits. This might be a possible source of error since it could result in a lower mortality for people in the study process as compared to the mortality for people in the target process which included people who are not university graduates.

(i) Consider the variate which is the number of cups of coffee consumed per day. This variate seems only to have been measured on entering the study. If the person suddenly gave up drinking coffee or started drinking coffee soon after entry into the study then coffee consumption would not be accurately measured. It would also be easy for a person to over or understate the number of cups they drink per day.
Consider the variate which is whether the person died during the follow-up period. Information on mortality was obtained from study participants and their families, postal authorities, and the National Death Index. It is possible that whether or not a person has died was not recorded correctly.

(j) The purpose of the study was to examine the association between coffee consumption and the risk of mortality (death) in a middle-aged Mediterranean cohort. However since there were only Spanish university graduates in the study the conclusions are limited to this study process. Also coffee consumption was only measured at entry into the study and did not take into account that the participants could have changed their coffee drinking habits after entry into the study.

(k) It is not a good idea to make a decision about whether or not to start drinking four cups of coffee a day based on just one study particularly when the study is observational. The conclusion was "Our findings suggest that drinking four cups of coffee each day can be part of a healthy diet in healthy people." It is not clear from this study that drinking coffee is actually improving a person's health. As well we are all living in North America, not the Mediterranean, so the conclusion does not really apply to us.

3.7 (a) The purpose of the study was to study whether live music, played or sung, was beneficial to premature babies in terms of helping to slow their heartbeats, calm their breathing, improve their sucking behaviors (important for feeding), aid their sleep and promote their states of quiet alertness.

(b) This is a causative Problem because the researchers were interested in whether live music could cause beneficial results to premature babies.

(c) A suitable target population for this study would be premature infants in New York state (or the United States).

(d) A suitable study population for this study could be premature infants who satisfied the criteria (aged ≤ 32 weeks with respiratory distress syndrome, clinical sepsis, and/or SGA) between January 2011 and December 2012 in the 11 hospitals who were given approval by their review boards.

(e) There is no information in the article on how the babies were selected from the 11 hospitals. Presumably the parents of the babies needed to give their consent for their baby to be in the study. It is unclear whether the babies in the sample

were in fact all the babies whose parents gave their consent in the 11 approved hospitals during the two year study period.

(f) The sample consists of the 272 premature babies who participated in the study. The sample size is 272.

(g) The study population only consisted of premature infants from 11 hospitals in New York state at the time of the study. It might be that premature infants at other hospitals in New York state (or the United States) are systematically different from the premature infants at these 11 hospitals with respect to the attributes discussed in the solution to Chapter 1, Problem 26.

(h) No information was given on how these premature babies were selected from the study population. It could be that some parents of premature babies refused to participate in the study. The premature babies of the parents who refused to participate may be systematically different from the premature babies of the parents who agreed to participate with respect to the attributes of interest.

(i) Respiratory rate (number of breaths per minute) and heart rate (heartbeats per minute) would presumably be measured by an experienced nurse or technician. The skill and reliability of the nurse or technician are possible sources of measurement error. An unskilled nurse or technician could take measurements which are not the true values.

Oxygen saturation would probably be measured using a blood sample and the value determined in a lab. The skill of the lab technicians and the reliability of the equipment used are possible sources of measurement error. For example, a device for determining oxygen concentration may be badly calibrated and may always underdetermine the oxygen saturation.

Sucking pattern (active/medium/slow/none) would need to be determined by a technician trained to be able to detect the differences between these levels. The skill of the technician is a possible source of measurement error.

(j) The most serious limitation is that this study only involved the 11 hospitals whose boards gave approval to participate in the study. It is possible that these 11 hospitals are systematically different with respect to the attributes of interest as compared to the other hospitals in New York state. It would be wrong to conclude that the results of this study apply to all premature babies in New York state or the United States.

SOLUTIONS TO CHAPTER 4 PROBLEMS

4.1 (b) For samples of size $n = 30$ the histogram of simulated means should be centred very close to $\mu = 2.326$, the variability of the sample means should be smaller compared to the variability for samples of size $n = 15$ since $sd\left(\bar{Y}\right) \approx \sigma/\sqrt{n}$ and the histogram should be more symmetric. The estimate of $P\left(\left|\bar{Y} - 2.326\right| \leq 0.5\right)$ should increase since $P\left(\left|\bar{Y} - 2.326\right| \leq 0.5\right)$ increases as n increases.

(c) Since $E\left(\bar{Y}\right) = \mu$, the mean of the original population will affect the location of the center of the histogram of simulated means.
Since $sd\left(\bar{Y}\right) \approx \sigma/\sqrt{n}$, the standard deviation of the original population will affect the spread of the histogram. Larger values of σ will result in histograms with more spread.
From the Central Limit Theorem we know that if the original population is very symmetric then the distribution of $\bar{Y}$ approaches a Normal distribution more rapidly as n increases as compared to the case in which the original distribution is very non-symmetric. Therefore if the original distribution is reasonably symmetric then the histogram will be very symmetric even if the sample size n is not large. If the original distribution is not symmetric then you would not expect the histogram to be reasonably symmetric unless n is large.

(d) Since $sd\left(\bar{Y}\right) \approx \sigma/\sqrt{n}$, the spread of the histogram will be affected by the sample size n. Larger sample sizes will result in histograms which are more concentrated around the mean μ which intuitively makes sense.

4.4 From Chapter 2, Problem 4 we have

$$\begin{aligned} R(\theta) &= \left[\left(\frac{\theta}{0.5}\right)\left(\frac{1-\theta}{0.5}\right)\right]^{200} \\ &= [4\theta(1-\theta)]^{200} \quad \text{for} \quad 0 \leq \theta \leq 1 \end{aligned}$$

The 15% likelihood interval is $[0.45, 0.55]$ which can be obtained from the graph of $R(\theta)$ given in Figure 4.1 or by using the R commands

```
uniroot(function(x)((4*x*(1-x))^200-0.15),lower=0.42,upper=0.46)$root
uniroot(function(x)((4*x*(1-x))^200-0.15),lower=0.52,upper=0.56)$root
```

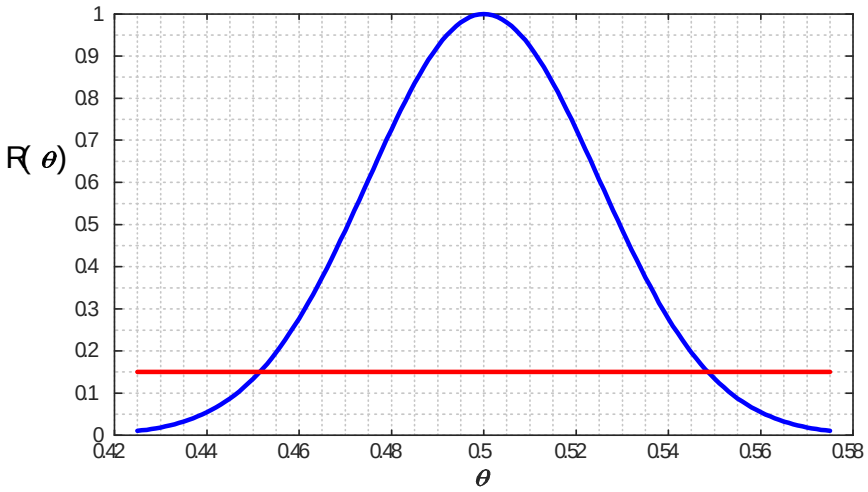


Figure 4.1: Relative likelihood function for fracture data

4.5 From Chapter 2, Problem 6 we have

$$R(\theta) = \left[\frac{\hat{\theta}}{\theta} e^{(1-\hat{\theta}/\theta)}\right]^n \quad \text{for } \theta > 0$$

For $n = 20$ and $\hat{\theta} = 3.6$ the 15% likelihood interval is $[2.40, 5.76]$. For $n = 60$ and $\hat{\theta} = 3.6$ the 15% likelihood interval is $[2.83, 4.68]$. These intervals can be obtained approximately from the graphs of $R(\theta)$ given in Figure 4.2 or by using the R commands

```
RLF<-function(x,that,n) {exp(n*log(that/x)+n*(1-that/x))}
uniroot(function(x) (RLF(x,3.6,20)-0.15),lower=1,upper=3)$root
uniroot(function(x) (RLF(x,3.6,20)-0.15),lower=5,upper=6)$root
uniroot(function(x) (RLF(x,3.6,60)-0.15),lower=2,upper=3)$root
uniroot(function(x) (RLF(x,3.6,60)-0.15),lower=4,upper=5)$root
```

The width of the likelihood interval for $n = 60$ is narrower than the interval for $n = 20$. This is expected since, as the same size increases, we would expect to have more information about the unknown parameter θ. More information would generally result in a narrower interval of plausible values for θ.

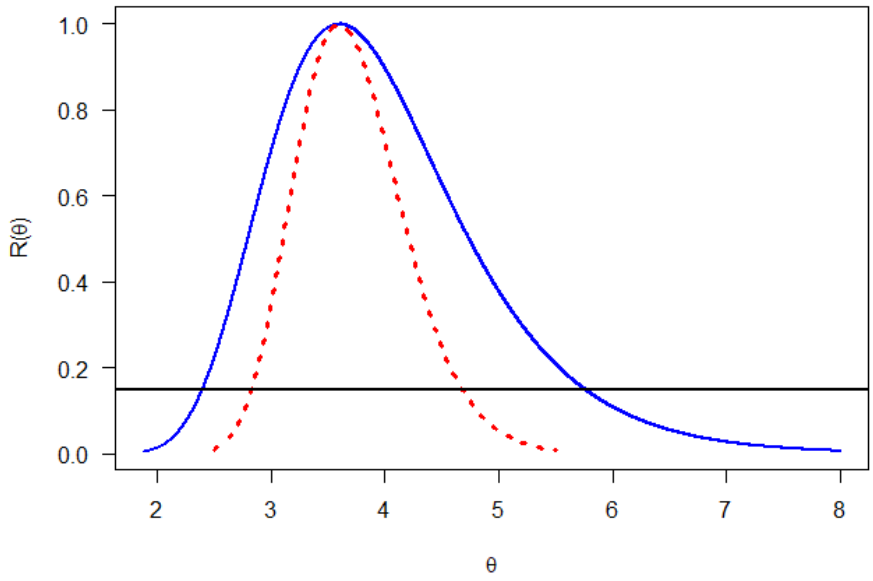


Figure 4.2: Relative likelihood functions for Problem 5

4.6 From Chapter 2, Problem 8 we have

$$r(\theta) = n \log\left(\frac{\theta+1}{\hat{\theta}+1}\right) + n\left(\frac{\hat{\theta}-\theta}{\hat{\theta}+1}\right) \quad \text{for } \theta > -1$$

For $n = 15$ and $\hat{\theta} = -0.5652$ the 15% likelihood interval is $[-0.75, -0.31]$. For $n = 45$ and $\hat{\theta} = -0.5652$ the 15% likelihood interval is $[-0.68, -0.43]$. These intervals can be obtained approximately from the graphs of $r(\theta)$ given in Figure 4.3 or by using the R commands

```
rlf<-function(x,that,n) {n*log((x+1)/(that+1))+n*(that-x)/(that+1)}
lg15<-log(0.15)
uniroot(function(x) (rlf(x,-0.5652,15)-lg15),lower=-0.8,upper=-0.6)$root
uniroot(function(x) (rlf(x,-0.5652,15)-lg15),lower=-0.6,upper=-0.2)$root
uniroot(function(x) (rlf(x,-0.5652,45)-lg15),lower=-0.8,upper=-0.6)$root
uniroot(function(x) (rlf(x,-0.5652,45)-lg15),lower=-0.6,upper=-0.2)$root
```

The width of the likelihood interval for $n = 45$ is narrower than the interval for $n = 15$. This is expected since, as the same size increases, we would expect to have more information about the unknown parameter θ. More information would generally result in a narrower interval of plausible values for θ.

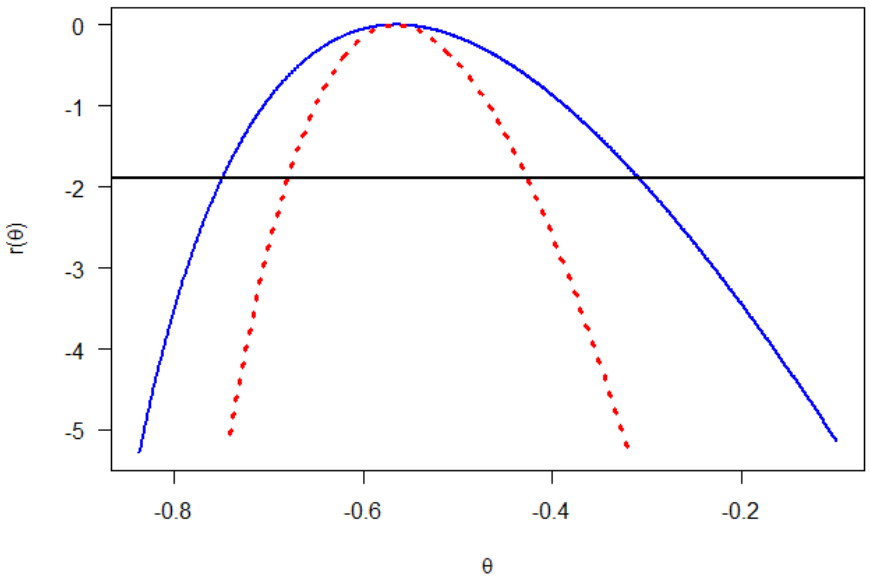


Figure 4.3: Log relative likelihood functions for Problem 6

4.7 (a) For the data $n_1 = 16$, $n_2 = 16$ and $n_3 = 18$, $\hat{\alpha} = 0.28$ and

$$R(\alpha) = \frac{(1+\alpha)^{32}(1-\alpha)^{18}}{(1+0.28)^{32}(1-0.28)^{18}} \quad \text{for } 0 \leq \alpha \leq < 1$$

Looking at Figure 4.4 we can see that $R(0) > 0.1$ and since $\alpha > 0$ the lower endpoint of the 10% likelihood interval is 0. We can also see that $R(\alpha) = 0.1$ corresponds to α between 0.5 to 0.6. We use the R command

```
uniroot(function(x)(((1+x)/1.28)^32*((1-x)/0.72)^18-0.1),
lower=0.5,upper=0.6)$root
```

to obtain the answer 0.55. Therefore the 10% likelihood interval is $[0, 0.55]$. Since the 10% likelihood interval is very wide compared to the set of possible values, α is not very well determined by these data.

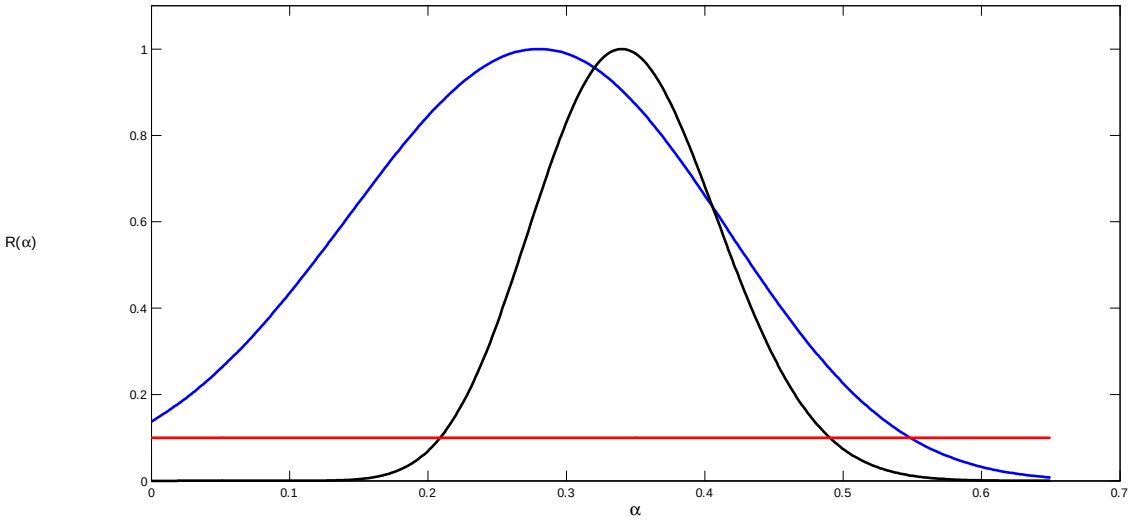


Figure 4.4: Relative likelihood functions for Twin Data

(b) For the data for which 17 identical pairs were found, $\hat{\alpha} = 17/50 = 0.34$ and the relative likelihood function is

$$R(\alpha) = \frac{\alpha^{17}(1-\alpha)^{33}}{(0.34)^{17}(1-0.34)^{33}} \quad \text{for } 0 \leq \alpha \leq 1$$

We use the R commands

```
uniroot(function(x)((x/0.34)^17*((1-x)/0.66)^33-0.1),
lower=0,upper=0.3)$root
uniroot(function(x)((x/0.34)^17*((1-x)/0.66)^33-0.1),
lower=0.4,upper=0.6)$root
```

to obtain the 10% likelihood interval $[0.21, 0.49]$. This interval is much narrower

than the interval in (a) which indicates that α is more accurately determined by the second data set.

4.8 From Chapter 2, Problem 12 (d) we have

$$L(\theta) = \theta^{16}(1-\theta)^{66} \quad \text{for } 0 \leq \theta \leq \frac{1}{2}$$

$$\hat{\theta} = \frac{16}{82} = \frac{8}{41}$$

and

$$R(\theta) = \frac{\theta^{16}(1-\theta)^{66}}{(8/41)^{16}(33/41)^{66}} \quad \text{for } 0 \leq \theta \leq \frac{1}{2}$$

A graph of $R(\theta)$ is given in Figure 4.5.

The 15% likelihood interval is $[0.12, 0.29]$ which can be obtained from the graph of $R(\theta)$ or by using the R commands

```
uniroot(function(x)((41*x/8)^16*(41*(1-x)/33)^66-0.15),
lower=0.1,upper=0.15)$root
uniroot(function(x)((41*x/8)^16*(41*(1-x)/33)^66-0.15),
lower=0.2,upper=0.3)$root
```

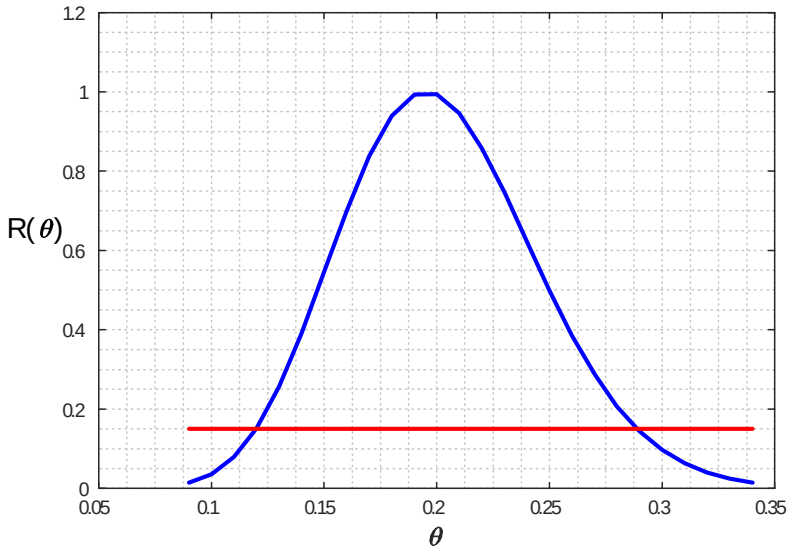


Figure 4.5: Relative likelihood function for size of family data

4.9 (a) The probability a group tests negative is $p = (1-\theta)^k$. The probability that y out of n groups test negative is

$$\binom{n}{y} p^y (1-p)^{n-y} \quad \text{for } y = 0, 1, ..., n$$

We are assuming that the nk people represent independent trials and that θ does not vary across subpopulations of the population of interest.

(b) Since $L(p) = p^y(1-p)^{n-y}$ is the usual Binomial likelihood we know $\hat{p} = y/n$. Solving $p = (1-\theta)^k$ for θ we obtain $\theta = 1 - p^{1/k}$. Therefore by the Invariance Property of maximum likelihood estimates, the maximum likelihood estimate of θ is

$$\hat{\theta} = 1 - (\hat{p})^{1/k} = 1 - (y/n)^{1/k}$$

(c) For $n = 100$, $k = 10$ and $y = 89$ we have $\hat{p} = 89/100 = 0.89$ and $\hat{\theta} = 1 - (89/100)^{1/10} = 0.0116$.
A 10% likelihood interval for p is found by using the R commands

```
uniroot(function(x)((x/0.89)^89*((1-x)/0.11)^11-0.1),
lower=0.8,upper=0.9)$root
uniroot(function(x)((x/0.89)^89*((1-x)/0.11)^11-0.1),
lower=0.9,upper=0.99)$root
```

which gives the interval $[0.8113, 0.9451]$ for p.
The 10% likelihood interval for θ is

$$\begin{aligned} & \left[1 - (0.9451)^{1/10}, 1 - (0.8113)^{1/10}\right] \\ = \; & [0.0056,\ 0.0207] \end{aligned}$$

4.10 (a) From Example 2.2.2 the likelihood function for Poisson data is

$$L(\theta) = \theta^{n\bar{y}} e^{-n\theta} \quad \text{for } \theta > 0$$

with corresponding maximum likelihood estimate $\hat{\theta} = \bar{y}$. For Company A, $n = 12$ and $\hat{\theta} = 20$ and the relative likelihood function is

$$R(\theta) = \frac{\theta^{n\bar{y}} e^{-n\theta}}{\bar{y}^{n\bar{y}} e^{-n\bar{y}}} \quad \text{for } \theta > 0$$

See Figure 4.6 for a graph of the relative likelihood function (graph on the right).

(b) For Company B, $n = 12$ and $\hat{\theta} = 11.67$. See Figure 4.6 for a graph of the relative likelihood function (graph on the left).

(c) The 15% likelihood interval for Company A is: $[17.59,\ 22.62]$ and the 15% likelihood interval for Company B is: $[9.84,\ 13.71]$. It is clear from these approximate 95% confidence intervals that the mean number of service calls for Company A is much larger than for Company B which implies the decision to go with Company B is a good one.

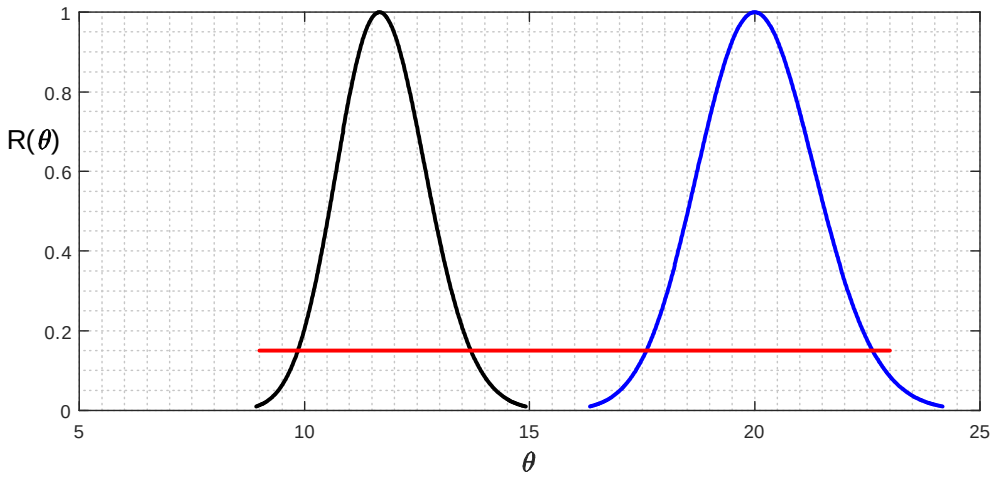


Figure 4.6: Relative Likelihood Functions for Company A and Company B Photocopiers

(d) The assumptions of the Poisson process (individuality, independence and homogeneity) would need to hold.

4.11 (a) If $n = 1000$ and $\theta = 0.5$ then

$$\begin{aligned} P\left(-0.03 \leq \tilde{\theta} - \theta \leq 0.03\right) &= P\left(\frac{-0.03}{\sqrt{\frac{(0.5)(0.5)}{1000}}} \leq \frac{\frac{Y}{1000} - 0.5}{\sqrt{\frac{(0.5)(0.5)}{1000}}} \leq \frac{0.03}{\sqrt{\frac{(0.5)(0.5)}{1000}}}\right) \\ &= P\left(-1.897367 \leq Z \leq 1.897367\right) \\ &= 2P\left(Z \leq 1.897367\right) - 1 \quad \text{where } Z \sim N(0,1) \\ &= 2 * \text{pnorm}\left(1.897367\right) - 1 = 0.9422205 \quad \text{using R} \end{aligned}$$

(b)

$$\begin{aligned} P\left(-0.03 \leq \tilde{\theta} - \theta \leq 0.03\right) &= P\left(-0.03 \leq \frac{Y}{n} - 0.5 \leq 0.03\right) \\ &= P\left(\frac{-0.03}{\sqrt{\frac{(0.5)(0.5)}{n}}} \leq \frac{\frac{Y}{n} - 0.5}{\sqrt{\frac{(0.5)(0.5)}{n}}} \leq \frac{0.03}{\sqrt{\frac{(0.5)(0.5)}{n}}}\right) \\ &\approx P\left(-0.06\sqrt{n} \leq Z \leq 0.06\sqrt{n}\right) \end{aligned}$$

where $Z \sim N(0,1)$. Since $P\left(-1.96 \leq Z \leq 1.96\right) = 0.95$, we need $0.06\sqrt{n} \geq 1.96$ or $n \geq (1.96/0.06)^2 = 1067.1$. Therefore n should be at least 1068.

(c)

$$\begin{aligned} P\left(-0.03 \leq \tilde{\theta} - \theta \leq 0.03\right) &= P\left(-0.03 \leq \frac{Y}{n} - \theta \leq 0.03\right) \\ &= P\left(\frac{-0.03}{\sqrt{\frac{\theta(1-\theta)}{n}}} \leq \frac{\frac{Y}{n} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \leq \frac{0.03}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right) \\ &\approx P\left(-\frac{0.03\sqrt{n}}{\sqrt{\theta\left(1-\theta\right)}} \leq Z \leq \frac{0.03\sqrt{n}}{\sqrt{\theta\left(1-\theta\right)}}\right) \end{aligned}$$

where $Z \sim N(0,1)$. Since $P\left(-1.96 \leq Z \leq 1.96\right) = 0.95$, we need

$$\frac{0.03\sqrt{n}}{\sqrt{\theta\left(1-\theta\right)}} \geq 1.96$$

or

$$n \geq \left(\frac{1.96}{0.03}\right)^2 \theta\left(1-\theta\right)$$

Since θ is unknown we take $\theta = 0.5$ so the inequality is true for all $0 < \theta < 1$. Thus

$$n \geq \left(\frac{1.96}{0.03}\right)^2 (0.5)^2 = 1067.1$$

and n should be at least 1068.

4.13 (a) Suppose the experiment which was used to estimate μ was conducted a large number of times and each time a 95% confidence interval for μ was constructed using the observed data. Then, approximately 95% of these constructed intervals would contain the true, but unknown value of μ. Since we only have one interval $[42.8, 47.8]$, we do not know whether it contains the true value of μ or not. We can only say that we are 95% confident that the given interval $[42.8, 47.8]$ contains the true value of μ since we are told it is a 95% confidence interval. In other words, we hope we were one of the "lucky" 95% who constructed an interval containing the true value of μ. **Warning:** $P\left(\mu \in [42.8, 47.8]\right) = 0.95$ is an **incorrect** statement.

(b) An approximate 95% confidence interval for the proportion of Canadians whose mobile phone is a smartphone is

$$\begin{aligned} \hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} &= 0.45 \pm 1.96\sqrt{\frac{0.45(0.55)}{1000}} = 0.45 \pm 0.03083498 \\ &= [0.419, 0.481] \end{aligned}$$

(c) We need n such that

$$n \geq \left(\frac{1.96}{0.02}\right)^2 (0.5)^2 = 2401$$

A sample size of 2401 or larger should be used.

4.14 (a) If Y is the number who support this information then $Y \sim \text{Binomial}(n, \theta)$. An approximate 95% confidence interval is given by

$$\begin{aligned} & 0.7 \pm 1.96\sqrt{\frac{0.7(0.3)}{200}} \\ = \; & 0.7 \pm 0.06351 \\ = \; & [0.6365,\ 0.7635] \end{aligned}$$

(b) The Binomial model assumes that the 200 people represent 200 independent trials. If 100 of the people interviewed were 50 married couples then the two people in a couple are probably not independent with respect to their views.

4.15 Let $Y =$ number of women who tested positive. Assume that model $Y \sim \text{Binomial}(n, \theta)$. Since

$$P\left(-2.5758 \leq Z \leq 2.5758\right) = 2P\left(Z \leq 2.5758\right) - 1 = 2\left(0.995\right) - 1 = 0.99$$

an approximate 99% confidence interval is given by

$$\begin{aligned} & \hat{\theta} \pm 2.58\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \\ = \; & \frac{64}{29000} \pm 2.5758\sqrt{\frac{\frac{64}{29000}(\frac{28936}{29000})}{29000}} \\ = \; & 0.002206897 \pm 0.000709781 \\ = \; & [0.0015, 0.0029] \end{aligned}$$

The Binomial model assumes that the $29,000$ women represented $29,000$ independent trials and that the probability that a randomly chosen women is HIV positive is equal to θ. The women may not represent independent trials and the probability that a randomly chosen women is HIV positive may be higher among certain high risk women such as women who are intravenous drug users.

4.16 (a) Since

$$\begin{aligned} 0.95 \; & \approx \; P\left(-1.96 \leq \frac{\bar{Y} - \theta}{\sqrt{\bar{Y}/n}} \leq 1.96\right) \\ & = \; P\left(\bar{Y} - 1.96\sqrt{\bar{Y}/n} \leq \theta \leq \bar{Y} + 1.96\sqrt{\bar{Y}/n}\right) \end{aligned}$$

therefore the interval $\left[\bar{y} - 1.96\sqrt{\bar{y}/n},\ \bar{y} + 1.96\sqrt{\bar{y}/n}\right]$ is an approximate 95% confidence interval for θ.

(b) An approximate 95% confidence interval for θ in the data in Chapter 2, Problem 10 is

$$\left[\frac{1669}{696} - 1.96\sqrt{\left(\frac{1669}{696}\right)/696},\ \frac{1669}{696} + 1.96\sqrt{\left(\frac{1669}{696}\right)/696}\right]$$
$$= [2.282942, 2.513035]$$

(c) A 15% likelihood interval for θ obtained using the R commands

```
n<-696
t<-1669/696
uniroot(function(x) (exp(n*t*log(x/t)+n*(t-x))-0.15),
lower=2.2,upper=2.4)$root
uniroot(function(x) (exp(n*t*log(x/t)+n*(t-x))-0.15),
lower=2.4,upper=2.6)$root
```

is

$$[2.285465, 2.514146]$$

The intervals are very similar since the sample size $n = 696$ is very large.

4.17 For Company A the approximate 95% confidence interval is $[17.5, 22.5]$ and for Company B the approximate 95% confidence interval is $[9.73, 13.60]$. These intervals are similar but not identical to the intervals in (c) since $n = 12$ is small. The intervals would be more similar for a larger value of n.

4.18 (a) Since

$$E\left(Y^k\right) = \int_0^\infty y^k \frac{y}{\theta^2} e^{-y/\theta} dy = \int_0^\infty \frac{y^{k+1}}{\theta^2} e^{-y/\theta} dy \quad \text{let } x = y/\theta$$
$$= \frac{1}{\theta^2}\int_0^\infty (x\theta)^{k+1} e^{-x}\theta dx = \theta^k \int_0^\infty x^{k+1} e^{-x} dx = \theta^k \Gamma(k+2)$$

therefore

$$E(Y) = \theta\Gamma(3) = 2\theta,\ \ E\left(Y^2\right) = \theta^2\Gamma(4) = 6\theta^2$$
$$Var(Y) = E\left(Y^2\right) - [E(Y)]^2 = 6\theta^2 - (2\theta)^2 = 2\theta^2$$

as required.

(b) The likelihood function is

$$L(\theta) = \prod_{i=1}^n \frac{y_i}{\theta^2} e^{-y_i/\theta} = \left(\prod_{i=1}^n y_i\right)\theta^{-2n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^n y_i\right) \quad \text{for } \theta > 0$$

or more simply

$$L(\theta) = \theta^{-2n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^n y_i\right) \quad \text{for } \theta > 0$$

The log likelihood function is

$$l(\theta) = -2n\log\theta - \frac{1}{\theta}\sum_{i=1}^{n} y_i \quad \text{for } \theta > 0$$

and

$$l'(\theta) = -\frac{2n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} y_i = \frac{1}{\theta^2}\left(\sum_{i=1}^{n} y_i - 2n\theta\right) \quad \text{for } \theta > 0$$

Now $l'(\theta) = 0$ if

$$\theta = \frac{1}{2n}\sum_{i=1}^{n} y_i = \frac{1}{2}\bar{y}$$

(Note a First Derivative Test could be used to confirm that $l(\theta)$ has an absolute maximum at $\theta = \bar{y}/2$.) The maximum likelihood estimate of θ is $\hat{\theta} = \frac{\bar{y}}{2}$.

(c)

$$E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} 2\theta = \frac{1}{n}(2n\theta) = 2\theta$$

and

$$Var\left(\bar{Y}\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} 2\theta^2 = \frac{1}{n^2}\left(2n\theta^2\right) = \frac{2\theta^2}{n}$$

(d) Since $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables then by the Central Limit Theorem

$$\frac{\bar{Y} - 2\theta}{\theta\sqrt{2/n}} \text{ has approximately a } G(0,1) \text{ distribution}$$

If $Z \sim N(0,1)$

$$P(-1.96 \le Z \le 1.96) = 0.95$$

Therefore

$$P\left(-1.96 \le \frac{\bar{Y} - 2\theta}{\theta\sqrt{2/n}} \le 1.96\right) \approx 0.95$$

(e) Since

$$\begin{aligned} 0.95 &\approx P\left(-1.96 \le \frac{\bar{Y} - 2\theta}{\theta\sqrt{2/n}} \le 1.96\right) \\ &= P\left(\bar{Y} - 1.96\theta\sqrt{2/n} \le 2\theta \le \bar{Y} + 1.96\theta\sqrt{2/n}\right) \\ &= P\left(\bar{Y}/2 - 0.98\theta\sqrt{2/n} \le \theta \le \bar{Y}/2 + 0.98\theta\sqrt{2/n}\right) \end{aligned}$$

an approximate 95% confidence interval for θ is

$$\left[\hat{\theta} - 0.98\hat{\theta}\sqrt{2/n},\ \hat{\theta} + 0.98\hat{\theta}\sqrt{2/n}\right]$$

where $\hat{\theta} = \bar{y}/2$.

(f) For these data the maximum likelihood estimate of θ is

$$\hat{\theta} = \frac{\bar{y}}{2} = \frac{88.92/18}{2} = 2.47$$

and the approximate 95% confidence interval for θ is

$$2.47 \pm 0.98\,(2.47)\sqrt{\frac{2}{18}} = [1.66,\ 3.28]$$

4.19 (a)

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{2}\theta^3 t_i^2 \exp\left(-\theta t_i\right) = \left(\frac{1}{2^n}\prod_{i=1}^{n} t_i^2\right)\theta^{3n}\exp\left(-\theta\sum_{i=1}^{n} t_i\right)$$

or more simply

$$L(\theta) = \theta^{3n}\exp\left(-\theta\sum_{i=1}^{n} t_i\right) \quad \text{for } \theta > 0$$

The log likelihood function is

$$l(\theta) = 3n\log\theta - \theta\sum_{i=1}^{n} t_i \qquad \frac{dl}{d\theta} = \frac{3n}{\theta} - \sum_{i=1}^{n} t_i$$

Solving $l(\theta) = 0$, we obtain the maximum likelihood estimate

$$\hat{\theta} = \frac{3n}{\sum_{i=1}^{n} t_i}$$

The relative likelihood function is

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^{3n}\exp\left(-\theta\sum_{i=1}^{n} t_i\right)}{\hat{\theta}^{3n}\exp\left(-\hat{\theta}\sum_{i=1}^{n} t_i\right)} \quad \text{for } \theta > 0$$

But

$$\hat{\theta} = \frac{3n}{\sum_{i=1}^{n} t_i}$$

so

$$\sum_{i=1}^{n} t_i = \frac{3n}{\hat{\theta}}$$

Therefore

$$\begin{aligned} R(\theta) &= \frac{\theta^{3n}\exp\left(-\theta\frac{3n}{\hat{\theta}}\right)}{\hat{\theta}^{3n}\exp\left(-3n\right)} \\ &= \left(\frac{\theta}{\hat{\theta}}\right)^{3n}\exp\left[3n\left(1-\frac{\theta}{\hat{\theta}}\right)\right] \quad \text{for } \theta > 0 \end{aligned}$$

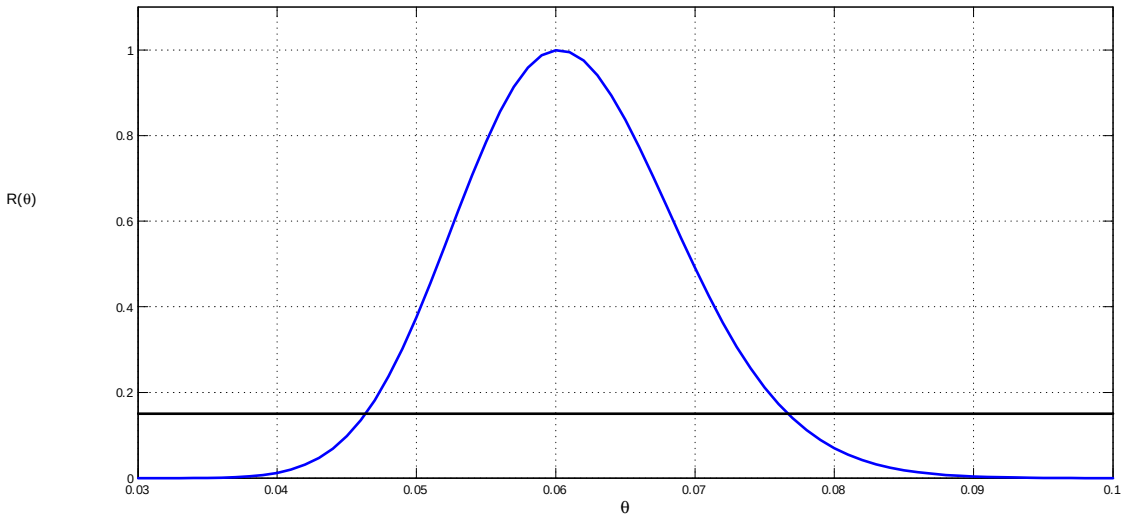


Figure 4.7: Relative Likelihood for Light Bulb Data

(b) Since $n = 20$ and $\sum_{i=1}^{20} t_i = 996$, therefore $\hat{\theta} = 3\,(20)\,/996 = 0.06024$. Reading from the graph in Figure 4.7 or by solving $R(\theta) = 0.15$ using the `uniroot` function in R, we obtain the 15% likelihood interval $[0.0463,\ 0.0768]$ which is an approximate 95% confidence interval for θ.

(c)

$$\begin{aligned} E\,(T) &= \frac{1}{2}\int_0^\infty \theta^3 t^3 e^{-\theta t} dt = \frac{1}{2}\int_0^\infty (\theta t)^3 e^{-(\theta t)} dt \\ &= \frac{1}{2\theta}\int_0^\infty x^3 e^{-x} dx \quad \text{(by letting } x = \theta t\text{)} \\ &= \frac{1}{2\theta}\Gamma(4) = \frac{1}{2\theta} 3! = \frac{3}{\theta} \end{aligned}$$

and a 95% approximate confidence interval for $E\,(T) = 3/\theta$ is

$$\left[\frac{3}{0.0768}, \frac{3}{0.0463}\right] = [39.1,\ 64.8]$$

(d)

$$\begin{aligned} p(\theta) &= P(T \le 50) = \frac{\theta^3}{2}\int_0^{50} t^2 e^{-\theta t} dt \\ &= \frac{\theta^3}{2}\left[\frac{-2500}{\theta} e^{-50\theta} - \frac{100}{\theta^2} e^{-50\theta} + \frac{2}{\theta^2}\left(-\frac{1}{\theta} e^{-50\theta} + \frac{1}{\theta}\right)\right] \\ &= 1 - \left(1250\theta^2 + 50\theta + 1\right) e^{-50\theta} \end{aligned}$$

Since

$$p\left(0.0463\right) = 1 - \left[1250\left(0.0463\right)^2 + 50\left(0.0463\right) + 1\right] e^{-50(0.0463)} = 0.408$$

and

$$p\left(0.0768\right) = 1 - \left[1250\left(0.0768\right)^2 + 50\left(0.0768\right) + 1\right] e^{-50(0.0768)} = 0.738$$

the confidence intervals for $p(\theta)$ using the model is $[0.408,\ 0.738]$.
The confidence interval for p using the Binomial model is

$$\begin{aligned} \hat{p} \pm 1.96\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} &= \frac{11}{20} \pm 1.96\sqrt{\frac{\left(11/20\right)\left(9/20\right)}{20}} \\ &= 0.55 \pm 0.218 \\ &= [0.332,\ 0.768] \end{aligned}$$

The Binomial model involves fewer model assumptions but gives a less precise (wider) interval.

4.20 (a)

(i) If $X \sim \chi^2\left(10\right)$ then $P(X \leq 2.6) \approx P\left(X \leq 2.558\right) = 0.01$ and $P(X > 16) \approx 1 - P\left(X \leq 15.987\right) = 1 - 0.9 = 0.1$.

(ii) If $X \sim \chi^2\left(4\right)$ then $P(X > 15) \approx 1 - P\left(X \leq 14.86\right) = 1 - 0.995 = 0.005$.

(iii) If $X \sim \chi^2\left(40\right)$ then $P\left(X \leq 24.4\right) \approx P\left(X \leq 24.433\right) = 0.025$ and $P\left(X \leq 55.8\right) \approx P\left(X \leq 55.758\right) = 0.95$.
If $Y \sim N\left(40, 80\right)$ then

$$\begin{aligned} P\left(Y \leq 24.4\right) &= P\left(Z \leq \frac{24.4 - 40}{\sqrt{80}}\right) \quad \text{where } Z \sim N\left(0,1\right) \\ &= P\left(Z \leq -1.74\right) \\ &= 1 - P\left(Z \leq 1.74\right) \\ &= 1 - 0.95907 \\ &= 0.04093 \approx 0.041 \end{aligned}$$

and

$$\begin{aligned} P\left(Y \leq 55.8\right) &= P\left(Z \leq \frac{55.8 - 40}{\sqrt{80}}\right) \quad \text{where } Z \sim N\left(0,1\right) \\ &= P\left(Z \leq 1.77\right) \\ &= 0.96164 \approx 0.96 \end{aligned}$$

If $X \sim \chi^2\left(40\right)$ then the graph of the probability density function of X will be fairly symmetric about the mean $E\left(X\right) = 40$ and very similar to the graph of the probability density function of a $N\left(40, 80\right)$ random variable. We note that $P\left(X \leq 55.8\right) = 0.95$ is close to $P\left(Y \leq 55.8\right) = 0.96$ while $P\left(X \leq 24.4\right) = 0.025$ and $P\left(Y \leq 24.4\right) = 0.041$ are not as close.

(iv) If $X \sim \chi^2(25)$ then solving $P(X \leq a) = 0.025$ and $P(X > b) = 0.025$ gives $a = 13.120$ and $b = 40.646$.

(v) If $X \sim \chi^2(12)$ then solving $P(X \leq a) = 0.05$ and $P(X > b) = 0.05$ gives $a = 5.226$ and $b = 21.026$.

(b)

(i) $P(X \leq 2.6) = \texttt{pchisq}(2.6, 10) = 0.01621621$,
$P(X > 16) = 1 - P(X \leq 16) = 1 - \texttt{pchisq}(16, 10) = 0.0996324$.

(ii) $P(X > 15) = 1 - P(X \leq 15) = 1 - \texttt{pchisq}(15, 4) = 0.004701217$.

(iii) $P(X \leq 24.4) = \texttt{pchisq}(24.4, 40) = 0.02469984$ and
$P(X \leq 55.8) = \texttt{pchisq}(55.8, 40) = 0.950383$.

(iv) $a = \texttt{qchisq}(0.025, 25) = 13.11972$ and $b = \texttt{qchisq}(0.975, 25) = 40.64647$

(v) $a = \texttt{qchisq}(0.05, 12) = 5.226029$ and $b = \texttt{qchisq}(0.95, 12) = 21.02607$

(c)

(i) If $X \sim \chi^2(1)$ then

$$\begin{aligned} P(X \leq 2) &= P\left(|Z| \leq \sqrt{2}\right) \quad \text{where } Z \sim N(0,1) \\ &= 2P(Z \leq 1.41) - 1 \\ &= 2(0.92073) - 1 \\ &= 0.84146 \end{aligned}$$

and

$$\begin{aligned} P(X > 1.4) &= 1 - P\left(|Z| \leq \sqrt{1.4}\right) \quad \text{where } Z \sim N(0,1) \\ &= 2\left[1 - P(Z \leq 1.18)\right] \\ &= 2(1 - 0.88100) \\ &= 0.23800 \end{aligned}$$

(ii) If $X \sim \chi^2(2) = \text{Exponential}(2)$ then

$$\begin{aligned} P(X &\leq 2) = 1 - e^{-2/2} \\ &= 1 - e^{-1} \approx 0.632 \end{aligned}$$

and

$$\begin{aligned} P(X &> 3) = e^{-3/2} \\ &= e^{-1.5} \approx 0.223 \end{aligned}$$

(d) If $X \sim G(3,2)$ then $\left(\frac{X-3}{2}\right)^2 \sim \chi^2(1)$. Since $Y_i \sim \textit{Exponential}(2)$, $i = 1, 2, \ldots, 5$ independently and $\textit{Exponential}(2)$ is the same distribution as $\chi^2(2)$,therefore $W = \sum_{i=1}^{5} Y_i + \left(\frac{X-3}{2}\right)^2 \sim \chi^2(10+1)$ or $\chi^2(11)$.

(e) If $X_i \sim \chi^2(i)$, $i = 1, 2, \ldots, 10$ independently then $\sum_{i=1}^{10} X_i \sim \chi^2\left(\sum_{i=1}^{10} i\right)$ or $\chi^2(55)$.

4.21 (a)

$$\begin{aligned}
\int_0^\infty \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} y^{\frac{k}{2}-1} e^{-\frac{y}{2}} dy &= \frac{1}{\Gamma\left(\frac{k}{2}\right)} \int_0^\infty \left(\frac{y}{2}\right)^{\frac{k}{2}-1} e^{-\frac{y}{2}} \frac{dy}{2} \quad \text{let } x = \frac{y}{2} \\
&= \frac{1}{\Gamma\left(\frac{k}{2}\right)} \int_0^\infty x^{\frac{k}{2}-1} e^{-x} dx \\
&= \frac{1}{\Gamma\left(\frac{k}{2}\right)} \Gamma\left(\frac{k}{2}\right) \quad \text{since } \int_0^\infty x^{\alpha-1} e^{-x} dx = \Gamma(\alpha) \\
&= 1
\end{aligned}$$

(b) See Figure 4.8. As k increases the probability density function becomes more symmetric about the line $y = k$.

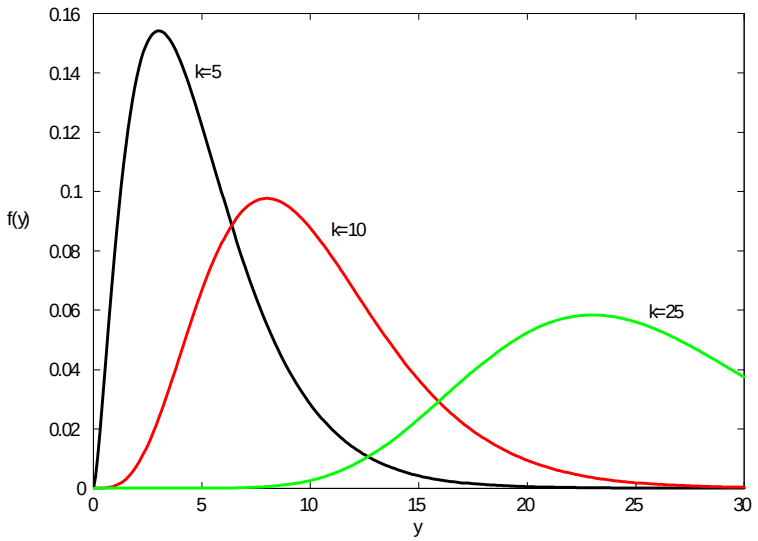


Figure 4.8: Chi-squared probability density functions for $k = 5, 10, 25$

(c)

$$\begin{aligned}
M(t) &= E\left(e^{Yt}\right) = \int_0^\infty \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} y^{\frac{k}{2}-1} e^{-\frac{y}{2}} e^{yt} dy \\
&= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \int_0^\infty y^{\frac{k}{2}-1} e^{-(\frac{1}{2}-t)y} dy \quad \text{converges for } t < \frac{1}{2} \\
&= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})(\frac{1}{2}-t)^{\frac{k}{2}}} \int_0^\infty x^{\frac{k}{2}-1} e^{-x} dx \quad \text{by letting } x = \left(\frac{1}{2} - t\right) y \\
&= \left[2^{\frac{k}{2}}(\frac{1}{2}-t)^{\frac{k}{2}}\right]^{-1} = (1-2t)^{-\frac{k}{2}} \quad \text{for } t < \frac{1}{2}
\end{aligned}$$

Therefore

$$\begin{aligned} M^{'}(0) &= E(Y) = -\frac{k}{2}(1-2t)^{-\frac{k}{2}-1}(-2)|_{t=0} = k \\ M^{''}(0) &= E\left(Y^2\right) = -\frac{k}{2}\left[-\left(\frac{k}{2}+1\right)\right](1-2t)^{-\frac{k}{2}-2}(-2\times -2)|_{t=0} = k^2+2k \\ Var(Y) &= k^2+2k-k^2 = 2k \end{aligned}$$

(d) $W_i \sim \chi^2(k_i)$ has moment generating function $M_i(t) = (1-2t)^{-k_i/2}$. Therefore $S = \sum_{i=1}^{n} W_i$ has moment generating function

$$M_s(t) = \prod_{i=1}^{n} M_i(t) = (1-2t)^{-\sum_{i=1}^{n} k_i/2}$$

which is the moment generating function of a χ^2 distribution with degrees of freedom equal to $\sum_{i=1}^{n} k_i$. Therefore $S \sim \chi^2\left(\sum_{i=1}^{n} k_i\right)$ as required.

4.22 (a) The graph is given in Figure 4.9. As k increases the graphs become more and more like the graph of the $N(0,1)$ probability density function and for $k = 25$ there is little difference between the $t(25)$ probability density function and the $N(0,1)$ probability density function.

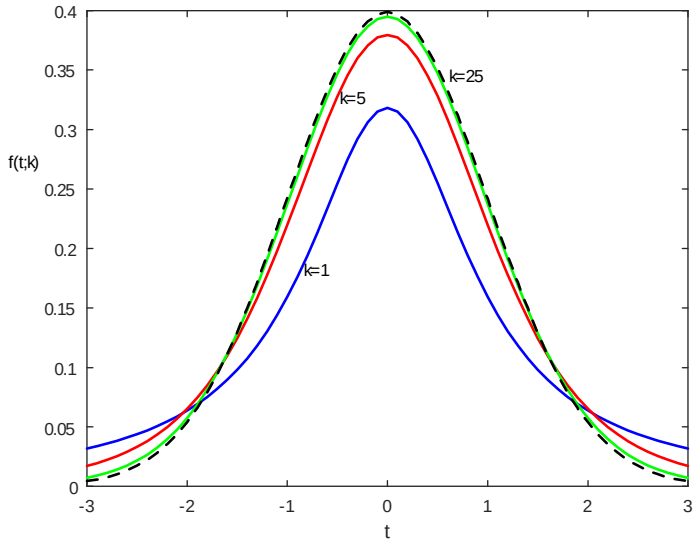


Figure 4.9: Graphs of the $t(k)$ pd.f. for $k = 1, 5, 25$ and the $N(0,1)$ p.d.f. (dashed line)

(b)

$$\begin{aligned}\frac{d}{dt}f(t;k) &= \frac{d}{dt}c_k\left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}} = c_k\left(-\frac{k+1}{2}\right)\left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}-1}\frac{2t}{k} \\ &= t\cdot c_k\left(-\frac{k+1}{k}\right)\left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}-1} = 0 \quad \text{if} \quad t=0\end{aligned}$$

Since $\frac{d}{dt}f(t;k) > 0$ if $t < 0$ and $\frac{d}{dt}f(t;k) < 0$ if $t > 0$ then by the First Derivative Test $f(t;k)$ has a global maximum at $t = 0$.

(c)

$$\begin{aligned}E(T) &= E\left(\frac{Z}{\sqrt{\frac{U}{k}}}\right) = E(Z)\,E\left(\frac{1}{\sqrt{\frac{U}{k}}}\right) \\ &\qquad \text{since } Z \text{ and } U \text{ are independent random variables} \\ &= 0 \quad \text{since } E(Z) = 0\end{aligned}$$

(d)

(i) If $T \sim t(10)$ then $P(T \le 0.88) = 0.8$,

$$P(T \le -0.88) = P(T > 0.88) \approx 1 - P(T \le 0.8791) \approx 1 - 0.8 = 0.2$$

and

$$\begin{aligned}P(|T| &\le 0.88) = P(-0.88 \le T \le 0.88) = P(T \le 0.88) - P(T \le -0.88) \\ &= P(T \le 0.88) - [1 - P(T \le 0.88)] = 2P(T \le 0.88) - 1 \\ &\approx 2P(T \le 0.8791) - 1 = 2(0.8) - 1 = 0.6\end{aligned}$$

(ii) If $T \sim t(17)$ then

$$\begin{aligned}P(|T| &\ge 2.90) = 2P(T \ge 2.90) \ \text{ by symmetry} \\ &= 2[1 - P(T \le 2.90)] \approx 2[1 - P(T \le 2.8982)] = 2(1 - 0.995) = 0.01\end{aligned}$$

(iii) If $T \sim t(30)$ then

$$\begin{aligned}P(T &\le -2.04) = P(T \ge 2.04) = 1 - P(T \le 2.04) \\ &\approx 1 - P(T \le 2.0423) = 1 - 0.975 = 0.025\end{aligned}$$

and if $Z \sim N(0,1)$ then

$$\begin{aligned}P(Z &\le -2.04) = 1 - P(Z \le 2.04) \\ &= 1 - 0.97932 = 0.02068\end{aligned}$$

and these values are close.
If $T \sim t(30)$ then $P(T \le 0.26) \approx P(T \le 0.2556) = 0.6$ which is close to $P(Z \le 0.26) = 0.60257$ if $Z \sim N(0,1)$.

(iv) If $T \sim t(18)$ then $P(T \leq 2.1009) = 0.975$ so $P(T \geq 2.1009) = 0.025$ and by symmetry $P(T \leq -2.1009) = 0.025$. Therefore $a = -2.1009$ and $b = 2.1009$.

(v) If $T \sim t(13)$ then $P(T \leq 1.7709) = 0.95$ so $P(T \geq 1.7709) = 0.05$ and by symmetry $P(T \leq -1.7709) = 0.05$. Therefore $a = -1.7709$ and $b = 1.7709$.

(e)

(i) $P(T \leq -0.88) = \texttt{pt}(-0.88, 10) = 0.1997567$ and $P(|T| \leq 0.88) = 2P(T \leq 0.88) - 1 = 2 \times \texttt{pt}(0.88, 10) - 1 = 0.6004867$

(ii) $P(|T| \geq 2.90) = 2P(T \geq 2.90) = 2\left[1 - P(T \leq 2.90)\right] = 2\left[1 - \texttt{pt}\left(2.90, 17\right)\right] = 0.009962573$

(iii) $P(T \leq -2.04) = \texttt{pt}(-2.04, 30) = 0.02511979$ and $P(T \leq 0.26) = \texttt{pt}(0.26, 30) = 0.60168$

(iv) $a = \texttt{qt}(0.025, 18) = -2.100922$ and $b = \texttt{qt}(0.975, 18) = 2.100922$

(v) $a = \texttt{qt}(0.05, 13) = -1.770933$ and $b = \texttt{qt}(0.95, 13) = 1.770933$

4.23

$$\begin{aligned}
\lim_{k\to\infty} f(t;k) &= \lim_{k\to\infty} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} \\
&= \lim_{k\to\infty} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k}{2}} \left(1 + \frac{t^2}{k}\right)^{-\frac{1}{2}} \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) \quad \text{for } t \in \Re
\end{aligned}$$

since

$$\begin{aligned}
\lim_{k\to\infty} c_k &= \frac{1}{\sqrt{2\pi}} \\
\lim_{k\to\infty} \left(1 + \frac{t^2}{k}\right)^{-\frac{1}{2}} &= 1 \\
\lim_{k\to\infty} \left(1 + \frac{t^2}{k}\right)^{-\frac{k}{2}} &= \exp\left(-\frac{1}{2}t^2\right) \quad \text{since} \quad \lim_{y\to\infty} \left(1 + \frac{a}{n}\right)^{bn} = e^{ab}
\end{aligned}$$

4.24 (a) Since

$$W_i = Y_i - \bar{Y} = Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i = \left(1 - \frac{1}{n}\right) Y_i - \frac{1}{n}\sum_{j\neq i} Y_j \quad \text{for } i = 1, 2, \ldots, n$$

therefore W_i is a linear combination of $Y_1, Y_2, \ldots, Y_n$ and therefore a linear combination of independent Normal random variables.

(b)

$$E(W_i) = E\left(Y_i - \bar{Y}\right) = E(Y_i) - E\left(\bar{Y}\right) = \mu - \mu = 0 \quad \text{for } i = 1, 2, \ldots, n$$

Now $Cov(Y_i, Y_j) = 0$ if $i \neq j$ (since the $Y_i's$ are independent random variables) and $Cov(Y_i, Y_j) = \sigma^2$ if $i = j$ (since $Cov(Y_i, Y_i) = Var(Y_i) = \sigma^2$). This implies

$$\begin{aligned} Cov\left(Y_i, \bar{Y}\right) &= Cov\left(Y_i, \frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n} Cov\left(Y_i, \sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n} Cov(Y_i, Y_i) = \frac{1}{n} Var(Y_i) = \frac{\sigma^2}{n} \end{aligned}$$

Therefore

$$\begin{aligned} Var(W_i) &= Var\left(Y_i - \bar{Y}\right) = Var(Y_i) + Var\left(\bar{Y}\right) - 2Cov\left(Y_i, \bar{Y}\right) \\ &= \sigma^2 + \frac{\sigma^2}{n} - 2\left(\frac{\sigma^2}{n}\right) = \sigma^2\left(1 - \frac{1}{n}\right) \end{aligned}$$

(c)

$$\begin{aligned} Cov(W_i, W_j) &= Cov\left(Y_i - \bar{Y}, Y_j - \bar{Y}\right) \quad i \neq j \\ &= Cov(Y_i, Y_j) - Cov\left(Y_i, \bar{Y}\right) - Cov\left(\bar{Y}, Y_j\right) + Cov\left(\bar{Y}, \bar{Y}\right) \\ &= 0 - \frac{\sigma^2}{n} - \frac{\sigma^2}{n} + Var\left(\bar{Y}\right) \\ &= -\frac{2\sigma^2}{n} + \frac{\sigma^2}{n} = -\frac{\sigma^2}{n} \end{aligned}$$

4.25 Since $P(-a \leq Z \leq a) = p$ where $Z \sim G(0, 1)$ and

$$\frac{\bar{Y} - \theta}{\bar{Y}/\sqrt{n}} \text{ has approximately a } G(0, 1) \text{ distribution.}$$

then

$$\begin{aligned} p &\approx P\left(-a \leq \frac{\bar{Y} - \theta}{\bar{Y}/\sqrt{n}} \leq a\right) \\ &= P\left(\bar{Y} - a\bar{Y}/\sqrt{n} \leq \theta \leq \bar{Y} + a\bar{Y}/\sqrt{n}\right) \end{aligned}$$

and therefore $\bar{y} \pm a\bar{y}/\sqrt{n}$ is an approximate $100p\%$ confidence interval for θ.

4.26 (a) For these data $\bar{y} = \frac{1}{30}(11400) = 380$. An approximate 90% confidence interval for θ is $380 \pm (1.645)\, 380/\sqrt{30} = [265.9, 494.1]$.

(b) From Example 2.3.2

$$L(\theta) = \theta^{-n} e^{-n\bar{y}/\theta} \quad \text{for } \theta > 0 \quad \text{and} \quad \hat{\theta} = \bar{y}.$$

Therefore

$$\begin{aligned} R(\theta) &= \frac{L(\theta)}{L(\hat{\theta})} = \frac{L(\theta)}{L(\bar{y})} = \frac{\theta^{-n}e^{-n\bar{y}/\theta}}{(\bar{y})^{-n}e^{-n}} = \left(\frac{\bar{y}}{\theta}\right)^n e^{n(1-\bar{y}/\theta)} \\ &= \left[\frac{\bar{y}}{\theta}e^{(1-\bar{y}/\theta)}\right]^n \quad \text{for } \theta > 0 \end{aligned}$$

For the given data $n = 30$ and $\hat{\theta} = \frac{1}{30}(11400) = 380$ and

$$R(\theta) = \left[\frac{380}{\theta}e^{(1-380/\theta)}\right]^{30} \quad \text{for } \theta > 0$$

From the inverse Normal table

$$\begin{aligned} 0.90 &= P(|Z| \leq 1.6449) \quad \text{where } Z \sim N(0,1) \\ &= P\left(W \leq (1.6449)^2\right) \quad \text{where } W \sim \chi^2(1) \\ &= P(W \leq 2.7057) \end{aligned}$$

Since (see Section 4.6) $\{\theta : \Lambda(\theta) \leq 2.7057\} = \left\{\theta : 2l(\hat{\theta}) - 2l(\theta) \leq 2.7057\right\}$ is an approximate 90% confidence interval. Therefore

$$\begin{aligned} &\left\{\theta : 2l(\hat{\theta}) - 2l(\theta) \leq 2.7057\right\} \\ &= \left\{\theta : R(\theta) \geq e^{-2.7057/2}\right\} \\ &= \{\theta : R(\theta) \geq 0.2585\} \end{aligned}$$

which implies that a 26% likelihood interval is an approximate 90% confidence interval.
Using the uniroot function in R and

$$R(\theta) = \left[\frac{380}{\theta}e^{(1-380/\theta)}\right]^{30} \quad \text{for } \theta > 0$$

we obtain the interval as [285.5, 521.3]. Alternatively the likelihood interval can be determined approximately from a graph of the relative likelihood function. See Figure 4.10. The intervals [265.9, 494.1] and [285.5, 521.3] are not very close. The reason for this is that the relative likelihood function is not symmetric about $\hat{\theta}$.

(c) Since $P(X \leq m) = 1 - e^{-m/\theta} = 0.5$, therefore $m = -\theta\log(0.5) = \theta\log 2$ and the confidence interval for m is $[285.5\log 2,\ 521.3\log 2] = [197.9,\ 361.3]$ by using the confidence interval for θ obtained in (b).

4.27 (a) Let $F(y) = P(Y \leq y)$ be the cumulative distribution function of Y. For $w > 0$,

$$G(w) = P(W \leq w) = P\left(\frac{2Y}{\theta} \leq w\right) = P\left(Y \leq \frac{\theta w}{2}\right) = F\left(\frac{\theta w}{2}\right)$$

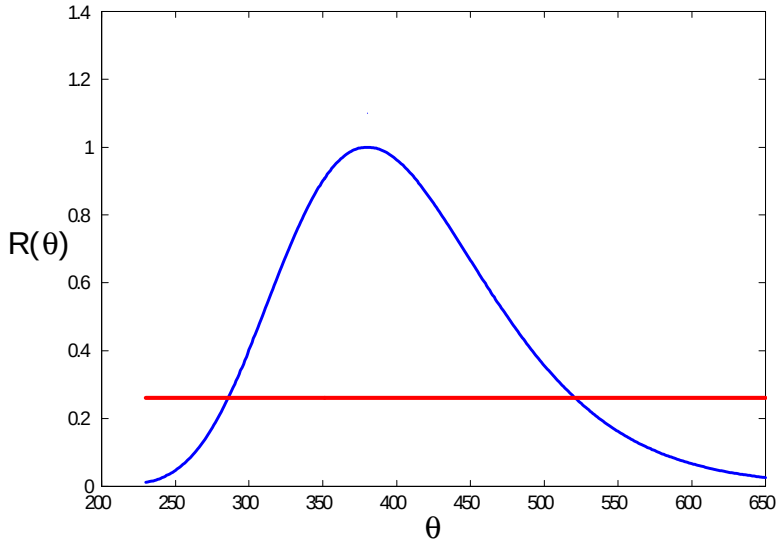


Figure 4.10: Relative likelihood function for survival times for AIDS patients

The probability density function of W is

$$\begin{aligned} g(w) &= \frac{d}{dw}G(w) = \frac{d}{dw}F\left(\frac{\theta w}{2}\right) \\ &= f\left(\frac{\theta w}{2}\right)\frac{d}{dw}\left(\frac{\theta w}{2}\right) = \frac{1}{\theta}e^{-(\frac{\theta w}{2})/\theta}\left(\frac{\theta}{2}\right) \\ &= \frac{1}{2}e^{-\frac{w}{2}} \quad \text{for } w > 0 \end{aligned}$$

which is the probability density function of a $\chi^2(2)$ random variable.

(b) Let $W_i = 2Y_i/\theta$, $i = 1, 2, \ldots, n$ independently. Then by part (a) $W_i \sim \chi^2(2)$, $i = 1, 2, \ldots, n$ independently. Since the sum of n independent $\chi^2(2)$ random variables has a χ^2 distribution with degrees of freedom equal to $\sum_{i=1}^{n} 2 = 2n$, therefore

$$U = \sum_{i=1}^{n} W_i = \sum_{i=1}^{n} \frac{2Y_i}{\theta} \sim \chi^2(2n)$$

as required.

(c) Using the Chi-squared table find a and b such that $P(U \leq a) = \frac{1-p}{2} = P(U \geq b)$

where $U \sim \chi^2(2n)$. Since

$$\begin{aligned} p &= P\left(a \leq \sum_{i=1}^{n} \frac{2Y_i}{\theta} \leq b\right) = P\left(\frac{1}{b} \leq \frac{\theta}{2\sum_{i=1}^{n} Y_i} \leq \frac{1}{a}\right) \\ &= P\left(\frac{2\sum_{i=1}^{n} Y_i}{b} \leq \theta \leq \frac{2\sum_{i=1}^{n} Y_i}{a}\right) \end{aligned}$$

then a $100p\%$ confidence interval for θ is given by

$$\left[\frac{2\sum_{i=1}^{n} y_i}{b}, \frac{2\sum_{i=1}^{n} y_i}{a}\right]$$

(d) Since

$$P(U \leq 43.19) = \frac{1-0.9}{2} = 0.05 = P(U \geq 79.08) \quad \text{where } W \sim \chi^2(60)$$

a 90% confidence interval for θ

$$\left[\frac{2(11400)}{79.082}, \frac{2(11400)}{43.188}\right] = [288.3,\ 527.9]$$

which is very close to the approximate 90% likelihood-based confidence interval $[285.5,\ 521.3]$ but not close to the approximate confidence interval $[265.9, 494.1]$ based on the asymptotic Normal pivotal.

4.28 Since $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, n$ independently, therefore

$$\frac{Y_i - \mu}{\sigma} \sim G(0,1) \quad \text{for } i = 1, 2, \ldots, n \text{ independently}$$

and

$$\left(\frac{Y_i - \mu}{\sigma}\right)^2 \sim \chi^2(1) \quad \text{for } i = 1, 2, \ldots, n \text{ independently}$$

since this distribution of the square of a $G(0,1)$ random variable is $\chi^2(1)$. Since the sum of n independent $\chi^2(1)$ random variables has a χ^2 distribution with degrees of freedom equal to $\sum_{i=1}^{n} 1 = n$, therefore

$$U = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$$

Since U is a function of the data and the unknown parameter σ (μ is known) whose distribution is completely known, therefore U is a pivotal quantity.

To construct a $100p\%$ confidence interval for σ^2 find values a and b such that

$$P(W \leq a) = \frac{1-p}{2} = P(W \geq b) \quad \text{where } W \sim \chi^2(n)$$

Since

$$\begin{aligned} p &= P\left(a \leq \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 \leq b\right) \\ &= P\left(\frac{\sum_{i=1}^{n} (Y_i - \mu)^2}{b} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n} (Y_i - \mu)^2}{a}\right) \\ &= P\left(\sqrt{\frac{\sum_{i=1}^{n} (Y_i - \mu)^2}{b}} \leq \sigma \leq \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \mu)^2}{a}}\right) \end{aligned}$$

a $100p\%$ confidence interval for σ^2 is given by

$$\left[\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{b}, \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{a}\right]$$

and a $100p\%$ confidence interval for σ is given by

$$\left[\sqrt{\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{b}}, \sqrt{\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{a}}\right]$$

4.29 (a) Since the points in the qqplot in Figure 4.11 lie reasonably along a straight line the Gaussian model seems reasonable for these data.

(b) A suitable study population for this study would be common octopi in the Ria de Vigo. The parameter μ represents the mean weight in grams of common octopi in the Ria de Vigo. The parameter σ represents the standard deviation of the weights in grams of common octopi in the Ria de Vigo.

(c) Since $P(T \leq 2.1009) = (1 + 0.95)/2 = 0.975$,

$$\hat{\mu} = \bar{y} = \frac{20340}{19} = 1070.526 \quad \text{and} \quad s = \left[\frac{1}{18}(884095)\right]^{1/2} = 221.62$$

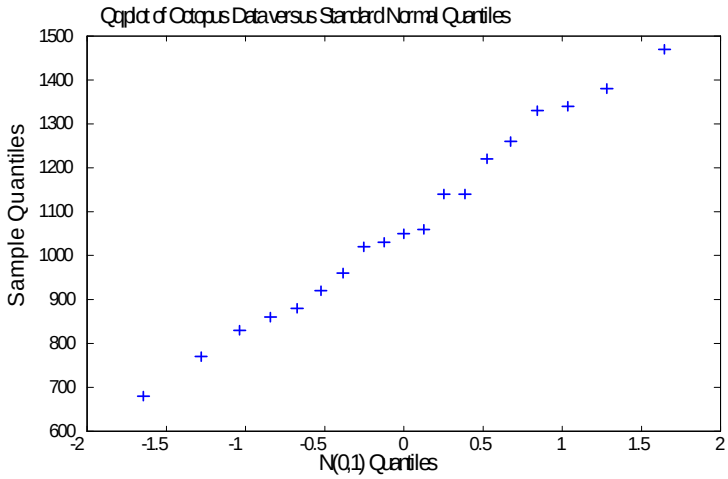


Figure 4.11: Qqplot for octopus data

therefore a 95% confidence interval for μ is

$$\begin{aligned} & 1070.526 \pm 2.1009\,(221.62)\,/\sqrt{19} = 1070.526 \pm 106.817 \\ = \;& [963.709, 1177.343] \end{aligned}$$

Since the value $\mu = 1100$ grams is well within this interval then the researchers could conclude that based on these data the octopi in the Ria de Vigo are reasonably healthy based on their mean weight.

(d) Since $P\,(W \leq 9.391) = 0.05 = P\,(W \geq 28.869)$ where $W \sim \chi^2\,(18)$ a 90% confidence interval for σ for the given data is

$$\left[\left(\frac{884095}{28.869}\right)^{1/2}, \left(\frac{884095}{9.391}\right)^{1/2}\right] = \left[(306.24)^{1/2}, (941.42)^{1/2}\right] = [175.00, 306.83]$$

4.30 (a) Qqplots of the weights for females and males separately are shown in Figures 4.12 and 4.13. In both cases the points lie reasonably along a straight line so it is reasonable to assume a Normal model for each data set.

(b) Using R we obtain $P\,(T \leq 1.976013) = 0.975$ where $T \sim t\,(149)$.
A 95% confidence interval for the mean weight of females is

$$\begin{aligned} & \left[\bar{y}_f - 1.976013 s_f/\sqrt{150}, \bar{y}_f + 1.976013 s_f/\sqrt{150}\right] \\ = \;& \left[70.4432 - (1.976013)\,(12.5092)\,/\sqrt{150}, 70.4432 + (1.976013)\,(12.5092)\,/\sqrt{150}\right] \\ = \;& [68.425, 72.461] \end{aligned}$$

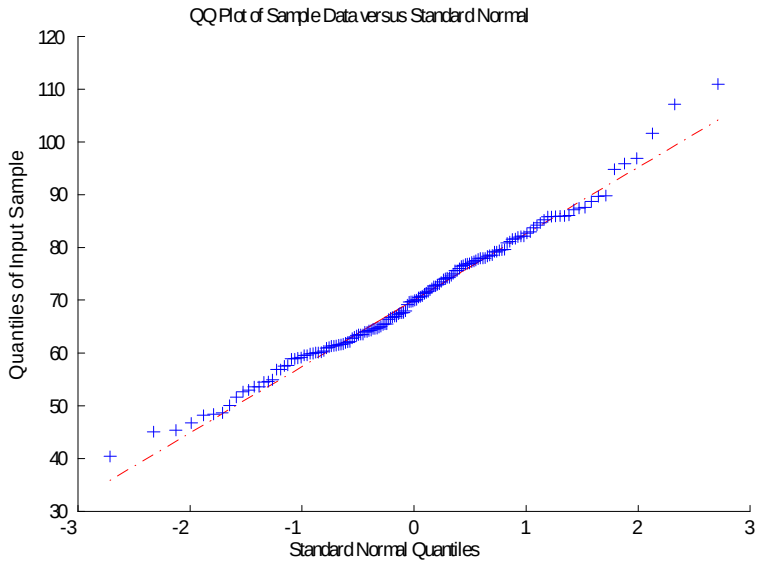


Figure 4.12: Qqplot of female weights

A 95% confidence interval for the mean weight of males is

$$\begin{aligned} & \left[\bar{y}_m - 1.976013 s_m/\sqrt{150}, \bar{y}_m + 1.976013 s_m/\sqrt{150}\right] \\ = & \left[82.5919 - (1.976013)(12.8536)/\sqrt{150}, 82.5919 + (1.976013)(12.8536)/\sqrt{150}\right] \\ = & [80.518, 84.666] \end{aligned}$$

We note that the interval for females and the interval for males have no values in common. The mean weight for males is higher than the mean weight for females.

(c) To obtain confidence intervals for the standard deviations we note that the pivotal quantity $(n-1)S^2/\sigma^2 = 149S^2/\sigma^2$ has a $\chi^2(149)$ distribution. Using R we have $P(W \leq 117.098) = 0.025 = P(W \geq 184.687)$ where $W \sim \chi^2(149)$. A 95% confidence interval given by

$$\left[\sqrt{\frac{149s^2}{184.687}}, \sqrt{\frac{149s^2}{117.098}}\right]$$

For the females we obtain

$$\begin{aligned} & \left[\sqrt{\frac{149\,(156.4806)}{184.687}}, \sqrt{\frac{149\,(156.4806)}{117.098}}\right] \\ = & \left[\sqrt{126.2439}, \sqrt{199.1119}\right] = [11.236, 14.111] \end{aligned}$$

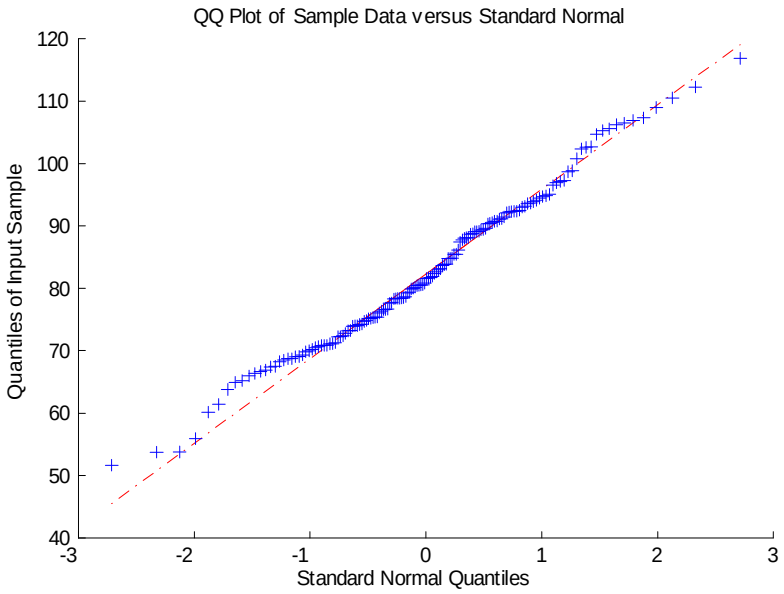


Figure 4.13: Qqplot of male weights

For the males we obtain

$$\left[\sqrt{\frac{149\,(165.2162)}{184.687}}, \sqrt{\frac{149\,(165.2162)}{117.098}}\right]$$
$$= \left[\sqrt{133.2915}, \sqrt{210.2274}\right] = [11.545, 14.499]$$

These intervals are quite similar.

4.31 (a) A suitable study population consists of the detergent packages produced by this particular detergent packaging machine. The parameter μ corresponds to the mean weight of the detergent packages produced by this detergent packaging machine. The parameter σ is the standard deviation of the weights of the detergent packages produced by this detergent packaging machine.

(b) For these data

$$\begin{aligned} \bar{y} &= \frac{4803}{16} = 300.1875 \\ s^2 &= \frac{1}{15}\left[1442369 - 16\,(300.1875)^2\right] = 37.89583 \\ s &= 6.155959 \end{aligned}$$

Since $P\,(T \leq 2.1314) = (1 + 0.95)/2 = 0.975$ where $T \sim t\,(15)$, a 95% confidence interval for μ is

$$\begin{aligned} 300.1875 \pm (2.1314)\,(6.155959)\,/\sqrt{16} &= 300.1875 \pm 3.2803 \\ &= [296.91,\ 303.47] \end{aligned}$$

Since $P(W \leq 6.262) = (1 - 0.95)/2 = 0.025$ and $P(W \leq 27.488) = (1 + 0.95)/2 = 0.975$, a 95% confidence interval for σ

$$\left[\sqrt{\frac{(15)(37.89583)}{27.488}},\ \sqrt{\frac{(15)(37.89583)}{6.262}}\right] = [4.55,\ 9.53]$$

(c) Since $P(T \leq 2.1314) = (1 + 0.95)/2 = 0.975$ where $T \sim t(15)$, a 95% prediction interval for Y is

$$\begin{aligned} & 300.1875 \pm (2.1314)(6.155959)\sqrt{1 + \frac{1}{16}} \\ = \ & 300.1875 \pm 13.5249 \\ = \ & [286.7,\ 313.7] \end{aligned}$$

4.32 (a) A suitable study population consists of the radon detectors sold at the Home Depot in Waterloo. The parameter μ corresponds to the mean level of radon in picocuries per liter detected by the detectors in the study population when placed in a chamber for three days and exposed to 105 picocuries per liter. The parameter σ is the standard deviation of these levels made by the radon detectors in the study population.

(b) For the radon data

$$n = 12, \quad \bar{y} = 104.1333 \quad \text{and} \quad s = \left[\frac{1}{11}\sum_{i=1}^{12}(y_i - \bar{y})^2\right]^{1/2} = 9.3974$$

From the t table, $P(T \leq 2.20) = (1 + 0.95)/2 = 0.975$ where $T \sim t(11)$. A 95% confidence interval for μ is

$$\begin{aligned} & \left[104.1333 - 2.20(9.3974)/\sqrt{12},\ 104.1333 + 2.20(9.3974)/\sqrt{12}\right] \\ & = [104.1333 - 5.9682,\ 104.1333 + 5.9682] \\ & = [98.1652,\ 110.1015] \end{aligned}$$

which does contain the value $\mu = 105$.

(c) From the Chi-squared table, $P(W \leq 3.816) = 0.025$ and $P(W \leq 21.920) = (1 + 0.95)/2 = 0.975$ where $W \sim \chi^2(11)$. A 95% confidence interval for σ is

$$\left[\sqrt{\frac{971.43}{21.920}}, \sqrt{\frac{971.43}{3.816}}\right] = [6.6571, 15.9551]$$

(d) Since the value $\mu = 105$ is near the center of the 95% confidence interval for μ, the data support the conclusion that the detector is accurate, that is, that the detector is not giving biased readings. The confidence interval for σ, however, indicates that the precision of the detectors might be of concern. The 95%

confidence interval for σ suggests that the standard deviation could be as large as 16 parts per billion. As a statistician you would need to rely on the expertise of the researchers for a decision about whether the size of the σ is scientifically significant and whether the precision of the detectors is too low. You would also point out to the researchers that this evidence is based on a fairly small sample of only 12 detectors.

(e) A 95% prediction interval for Y, the reading for the new radon detector exposed to 105 picocuries per liter of radon over 3 days, is

$$\left[104.1333 - 2.20\,(9.3974)\left(1+\frac{1}{12}\right)^{1/2},\ 104.1333 + 2.20\,(9.3974)\left(1+\frac{1}{12}\right)^{1/2}\right]$$
$$= [104.1333 - 21.5185,\ 104.1333 + 21.5185]$$
$$= [82.6148,\ 125.6519]$$

(f) For a 95% confidence interval for μ of width $2d$ we chose $n \approx (1.96\sigma/d)^2$. For this problem $d = 3$. Since σ is unknown we estimate it using $\sigma \approx s = 9.4$. Therefore $n \approx \left(\frac{1.96\sigma}{3}\right)^2 \approx \left(\frac{1.96}{3}\right)^2 (9.4)^2 = 37.7$. Since 12 observations have already been taken, the researchers should use at least $38 - 12 = 26$ more detectors. We note that this calculation depends on an estimate of σ from a small sample ($n = 12$) and the value 1.96 is from the Normal table rather than the t table in which the values are greater than 2. Therefore the researchers should be advised to use more than 26 additional detectors. Note that the upper limit of the 95% confidence interval for σ is 16 and $\left(\frac{2}{3}\right)^2 (16)^2 = 113.8$ which gives a rough upper limit on the number of detectors to use.

4.33 (a) The combined likelihood function for μ is

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2\sigma_1^2}(x_i-\mu)^2\right] \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2\sigma_2^2}(y_i-\mu)^2\right] \\ &= (2\pi)^{-(n+m)/2}\,\sigma_1^{-m}\sigma_2^{-n} \exp\left\{-\frac{1}{2\sigma_1^2}\left[\sum_{i=1}^{m}(x_i-\bar{x})^2 + m(\bar{x}-\mu)^2\right]\right\} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma_2^2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2\right]\right\} \end{aligned}$$

or more simply ignoring constants

$$L(\mu) = \exp\left[\frac{-m}{2\sigma_1^2}(\bar{x}-\mu)^2 - \frac{n}{2\sigma_2^2}(\bar{y}-\mu)^2\right] \quad \text{for } \mu \in \Re$$

since σ_1^2 and σ_2^2 are known. The log likelihood function is

$$l(\mu) = -\frac{m}{2\sigma_1^2}(\bar{x}-\mu)^2 - \frac{n}{2\sigma_2^2}(\bar{y}-\mu)^2$$

Solving

$$l'(\mu) = \frac{m}{\sigma_1^2}(\bar{x} - \mu) + \frac{n}{\sigma_2^2}(\bar{y} - \mu) = \frac{\left(m\sigma_2^2\bar{x} + n\sigma_1^2\bar{y}\right) - \left(m\sigma_2^2 + n\sigma_1^2\right)\mu}{\sigma_1^2\sigma_2^2} = 0$$

gives the maximum likelihood estimate for μ as

$$\begin{aligned}\hat{\mu} &= \frac{m\sigma_2^2\bar{x} + n\sigma_1^2\bar{y}}{m\sigma_2^2 + n\sigma_1^2} = \frac{\left(m/\sigma_1^2\right)\bar{x} + \left(n/\sigma_2^2\right)\bar{y}}{m/\sigma_1^2 + n/\sigma_2^2} \\ &= \frac{w_1\bar{x} + w_2\bar{y}}{w_1 + w_2}\end{aligned}$$

where $w_1 = m/\sigma_1^2$ and $w_2 = n/\sigma_2^2$. We first note that both $\bar{x}$ and $\bar{y}$ are both estimates of μ and it makes sense to take a weighted average of the two estimates to get a better estimate of μ. If the sample sizes n and m are not equal it makes sense to weight the estimate that is a function of more observations. It also makes sense that the mean of the observations that come from a distribution with smaller variance is a better estimate of μ and should be given more weight. By examining the weights w_1 and w_2 we can see that the estimate $\hat{\mu}$ does satisfies both of these requirements.

(b) Since the observations in $\bar{x}$ are observations from a distribution with larger variability then we don't want to take just an average of $\bar{x}$ and $\bar{y}$. We would choose an estimate that weights $\bar{y}$ more that $\bar{x}$ since $\bar{y}$ is a better estimate.

(c)

$$\begin{aligned}Var(\tilde{\mu}) &= Var\left(\frac{\overline{X} + 4\overline{Y}}{5}\right) = \frac{1}{25}\left[Var(\overline{X}) + 16Var(\bar{Y})\right] \\ &= \frac{1}{25}\left[\frac{1}{10} + 16\left(\frac{0.25}{10}\right)\right] = 0.02\end{aligned}$$

and $\sqrt{Var(\tilde{\mu})} = 0.1414$.

$$Var\left(\frac{\overline{X} + \overline{Y}}{2}\right) = \frac{1}{4}\left[Var(\overline{X}) + Var(\overline{Y})\right] = \frac{1}{4}\left(\frac{1}{10} + \frac{0.25}{10}\right) = 0.03125$$

and $\sqrt{Var\left(\frac{\overline{X}+\overline{Y}}{2}\right)} = 0.1768$. We can clearly see now that $\tilde{\mu}$ has a smaller standard deviation than the estimator $\left(\overline{X} + \overline{Y}\right)/2$.

SOLUTIONS TO CHAPTER 5 PROBLEMS

5.1 (a) The model $Y \sim \text{Binomial}(n, \theta)$ is appropriate in the case in which the experiment consists of a sequence of n independent trials with two outcomes on each trial (Success and Failure) and $P(\text{Success}) = \theta$ is the same on each trial. In this experiment the trials are the guesses. Since the deck is reschuffled each time it seems reasonable to assume the guesses are independent. It also seems reasonable to assume that the women's ability to guess the number remains the same on each trial. To test the hypothesis that the women is guessing at random the appropriate null hypothesis would be $H_0 : \theta = \frac{1}{5} = 0.2$.

(b) For $n = 20$ and $H_0 : \theta = 0.2$, we have $Y \sim \text{Binomial}(20, 0.2)$ and $E(Y) = 20(0.2) = 4$. We use the test statistic or discrepancy measure $D = |Y - E(Y)| = |Y - 4|$. The observed value of D is $d = |8 - 4| = 4$. Then

$$\begin{aligned} p - value &= P(D \geq 4; H_0) = P(|Y - 4| \geq 4; H_0) \\ &= P(Y = 0) + P(Y \geq 8) \\ &= \binom{20}{0}(0.2)^0(0.8)^{20} + \sum_{y=8}^{20}\binom{20}{y}(0.2)^y(0.8)^{20-y} \\ &= 1 - \sum_{y=1}^{7}\binom{20}{y}(0.2)^y(0.8)^{20-y} = 0.04367 \quad \text{calculated using R} \end{aligned}$$

Since $0.01 < p - value = 0.04367 < 0.05$ there is evidence based on the data against $H_0 : \theta = 0.2$. These data suggest that the woman might have some special guessing ability.

(c) For $n = 100$ and $H_0 : \theta = 0.2$, we have $Y \sim \text{Binomial}(100, 0.2)$, $E(Y) = 100(0.2) = 20$ and $Var(Y) = 100(0.2)(0.8) = 16$. We use the test statistic or discrepancy measure $D = |Y - E(Y)| = |Y - 20|$. The observed

value of D is $d = |32 - 20| = 12$. Then

$$\begin{aligned} p-value &= P\left(D \geq 12; H_0\right) = P\left(|Y - 20| \geq 12\right) \\ &= P(Y \leq 8) + P(Y \geq 32) \\ &= \sum_{y=0}^{8} \binom{100}{y} (0.2)^y (0.8)^{100-y} + \sum_{y=32}^{100} \binom{100}{y} (0.2)^y (0.8)^{100-y} \\ &= 1 - \sum_{y=9}^{31} \binom{100}{y} (0.2)^y (0.8)^{100-y} = 0.004 \quad \text{calculated using R} \end{aligned}$$

or

$$\begin{aligned} p-value &= P\left(D \geq 12; H_0\right) = P\left(|Y - 20| \geq 12\right) \\ &\approx P\left(|Z| \geq \frac{12}{\sqrt{16}}\right) \quad \text{where } Z \sim N(0,1) \\ &= 2\left[1 - P\left(Z \leq 3\right)\right] = 2\left(1 - 0.99865\right) = 0.0027 \end{aligned}$$

Since $0.001 < p - value < 0.01$ there is strong evidence based on the data against $H_0 : \theta = 0.2$. These data suggest that the woman has some special guessing ability. Note that we would not conclude that it has been proven that she does have special guessing ability!

5.2 Assuming $H_0 : \theta = 10$ is true $Y \sim$ Poisson(10). For the discrepancy measure $D = \max(0, Y - 10)$

$$\begin{aligned} p-value &= P(D \geq 15; H_0) = P\left(\max(0, Y - 10) \geq 15; H_0\right) \\ &= P(Y \geq 25) \quad \text{if } Y \sim \text{Poisson}\left(10\right) \\ &= 1 - \sum_{y=0}^{24} \frac{10^y e^{-10}}{y!} = 0.000047 \quad \text{calculated using R} \end{aligned}$$

or

$$\begin{aligned} p-value &= P(D \geq 15; H_0) = P(Y \geq 25) \\ &\approx P\left(Z \geq \frac{25 - 10}{\sqrt{10}}\right) \quad \text{where } Z \sim N(0,1) \\ &= P\left(Z \geq 4.74\right) \approx 0 \end{aligned}$$

Since $p-value \approx 0$ there is very strong evidence based on the data against $H_0 : \theta = 10$.

5.3 (a) A qqplot of the data is given in Figure 5.1. Since the points in the qqplot lie reasonably along a straight line it seems reasonable to assume a Normal model for these data.

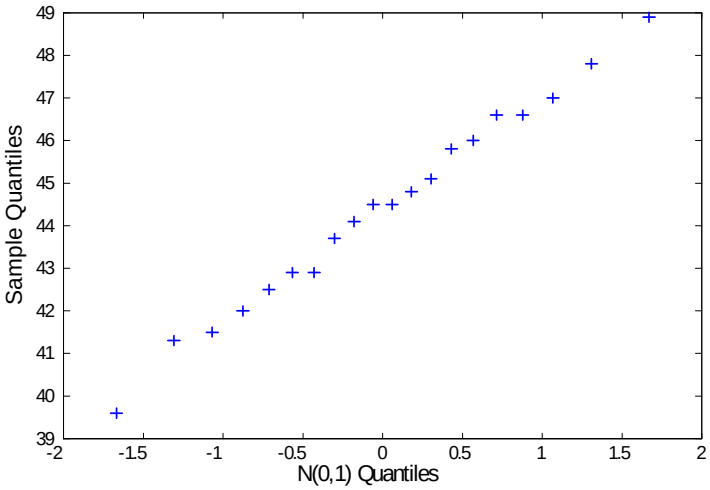


Figure 5.1: Qqplot for Dioxin data

(b) A study population is a bit difficult to define in this problem. One possible choice is to define the study population to be all measurements that could be taken on a given day by this instrument on a standard solution of 45 parts per billion dioxin. The parameter μ corresponds to the mean measurement made by this instrument on the standard solution. The parameter σ corresponds to the standard deviation of the measurements made by this instrument on the standard solution.

(c) For these data

$$\bar{y} = \frac{888.1}{20} = 44.405 \quad \text{and} \quad s = \left[\frac{39545.03 - 20\left(44.405\right)^2}{19}\right]^{1/2} = 2.3946$$

To test $H_0 : \mu = 45$ we use the test statistic

$$D = \frac{\left|\bar{Y} - 45\right|}{S/\sqrt{20}} \quad \text{where } T = \frac{\bar{Y} - 45}{S/\sqrt{20}} \sim t\left(19\right)$$

The observed value of D is

$$d = \frac{\left|44.405 - 45\right|}{2.3946/\sqrt{20}} = 1.11$$

and

$$\begin{aligned} p - value &= P\left(D \geq d; H_0\right) \\ &= P\left(|T| \geq 1.11\right) \quad \text{where } T \sim t\left(19\right) \\ &= 2\left[1 - P\left(T \leq 1.11\right)\right] \\ &= 0.2803 \quad \text{calculated using R} \end{aligned}$$

Alternatively using the t table we have $P(T \leq 0.8610) = 0.8$ and $P(T \leq 1.3277) = 0.9$ so

$$2(1-0.9) \leq p-value \leq 2(1-0.8)$$
$$\text{or } \ 0.2 \leq p-value \leq 0.4$$

In either case since the $p-value$ is larger than 0.1 and we would conclude that, based on the observed data, there is no evidence against the hypothesis $H_0 : \mu = 45$. (Note: This does not imply the hypothesis is true!).

A $100p\%$ confidence interval for μ based on the pivotal quantity

$$T = \frac{\bar{Y} - 45}{S/\sqrt{20}} \sim t(19)$$

is given by

$$\left[\bar{y} - as/\sqrt{20}, \bar{y} + as/\sqrt{20}\right]$$

where $P(T \leq a) = (1+p)/2$. From the t table we have $P(T \leq 2.093) = (1+0.95)/2 = 0.975$. Therefore the 95% confidence interval for μ is

$$\left[\bar{y} - 2.093s/\sqrt{20}, \bar{y} + 2.093s/\sqrt{20}\right] = [43.28, 45.53]$$

Based on these data it would appear that the new instrument is working as it should be since there was no evidence against $H_0 : \mu = 45$. We might notice that the value $\mu = 45$ is not in the center of the 95% confidence interval but closer to the upper endpoint suggesting that the instrument might be under reading the true value of 45. It would be wise to continue testing the instrument on a regular basis on a known sample to ensure that the instrument is continuing to work well.

(d) To test $H_0 : \sigma^2 = \sigma_0^2$ we use the test statistic

$$U = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

For $n = 20$ and $H_0 : \sigma^2 = 4$, or equivalently $H_0 : \sigma = 2$, we have

$$U = \frac{19S^2}{4} \sim \chi^2(19)$$

The observed value of U is

$$u = \frac{19(5.7342)}{4} = 27.23745$$

and

$$p-value = 2P(U \geq 27.23745) \quad \text{where } U \sim \chi^2(19)$$
$$= 0.1985 \quad \text{calculated using R}$$

Alternatively using the Chi-squared table we have $P(U \geq 27.204) = 1-0.9 = 0.1$ so $p-value \approx 2(0.1) = 0.2$. In either case, since the $p-value$ is larger than 0.1, we would conclude that there is no evidence against the hypothesis $H_0 : \sigma^2 = 4$ based on the observed data.

A $100p\%$ confidence interval for σ based on the pivotal quantity

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

is given by

$$\left[\sqrt{\frac{(n-1)s^2}{b}}, \sqrt{\frac{(n-1)s^2}{a}}\right]$$

where $P(U \leq a) = (1-p)/2 = P(U \geq b)$. For $n = 20$ and $p = 0.95$ we have $P(U \leq 8.907) = 0.025 = P(U \geq 32.852)$ and the confidence interval for σ is

$$\left[\sqrt{\frac{19(5.7342)}{32.852}}, \sqrt{\frac{19(5.7342)}{8.907}}\right] = [1.82, 3.50]$$

Based on these data there is no evidence to contradict the manufacturer's claim that the variability in measurements is less than two parts per billion. Note however that the confidence for σ does contain values of σ larger than 2 so again it would be wise to continue testing the instrument on a regular basis on a known sample to ensure that the instrument is continuing to work well.

(e) For the new data the observed value of the discrepancy measure

$$D = \frac{\left|\bar{Y} - \mu_0\right|}{S/\sqrt{n}}$$

is

$$d = \frac{|44.1 - 45|}{2.1/\sqrt{25}} = 2.1429$$

and

$$\begin{aligned} p - value &= P(D \geq d; H_0) \\ &= P(|T| \geq 2.1429) \quad \text{where } T \sim t(24) \\ &= 2[1 - P(T \leq 2.1429)] \\ &= 0.04247 \quad \text{calculated using R} \end{aligned}$$

Alternatively using the t table we have $P(T \leq 2.0639) = 0.975$ and $P(T \leq 2.4922) = 0.99$ so

$$\begin{gathered} 2(1-0.99) \leq p - value \leq 2(1-0.975) \\ \text{or} \quad 0.02 \leq p - value \leq 0.05 \end{gathered}$$

In either case since $0.02 \leq p - value \leq 0.05$ and there is evidence against the hypothesis $H_0 : \mu = 45$ based on the data.

Since $P(T \leq 2.0639) = 0.975$ a 95% confidence interval for μ is

$$\left[44.1 - 2.0639\,(2.1)/\sqrt{25}, \bar{y} + 2.0639\,(2.1)/\sqrt{25}\right] = [43.23, 44.97]$$

which only contains values of μ less than 45. Based on these data it would appear that the new instrument is giving measurements on average which are below the true value of 45 parts per billion and therefore the new instrument needs to be adjusted.

For these new data a statistically significant result of under measuring has been determined. The question of whether this result is of practical significance can only be answered by the people who use these results to make a decision. With many labs results decisions are made based on whether the observed measurement is within a certain interval which is considered to "safe"or not. Dioxins are poisonous to humans. Unfortunately dioxins are present in the food we eat. The 95% confidence interval suggests that the new instrument is giving results which are under reporting by $1 - 2$ parts per billion on average. What we need now is an expert on dioxin who can tell us how much a difference $1 - 2$ parts per billion makes in the context of how these results are used in the hospital lab.

(f) Here is the R code plus the output:

```
y<-c(44.1,46,46.6,41.3,44.8,47.8,44.5,45.1,42.9,44.5,
+ 42.5,41.5,39.6,42,45.8,48.9,46.6,42.9,47,43.7)
> t.test(y,mu=45,conf.level=0.95) # test hypothesis mean=45

One Sample t-test

data: y
t = -1.1112, df = 19, p-value = 0.2803
alternative hypothesis: true mean is not equal to 45
95 percent confidence interval:
43.28429 45.52571
sample estimates:
mean of x
44.405

> # and gives 1 95% confidence interval
> df<-length(y)-1 # degrees of freedom
> s2<-var(y) # sample variance
> p<-0.95 # p=0.95 for 95% confidence interval
> a<-qchisq((1-p)/2,df) # lower value from Chi-squared dist'n
```

```
> b<-qchisq((1+p)/2,df) # upper value from Chi-squared dist'n
> c(s2*df/b,s2*df/a) # confidence interval for sigma squared
[1] 3.31634 12.23256
> c(sqrt(s2*df/b),sqrt(s2*df/a)) # confidence interval for sigma
[1] 1.821082 3.497508
> sigma0sq<-2^2 # test hypotheis sigma=2 or sigmasq=4
> chitest<-s2*df/sigma0sq
> q<-pchisq(chitest,df)
> min(2*q,2*(1-q))
[1] 0.1984887
```

5.4 To test $H_0 : \mu = 45$ when $\sigma^2 = 4$ is known we use the discrepancy measure

$$D = \frac{\left|\bar{Y} - 45\right|}{2/\sqrt{20}} \quad \text{where } Z = \frac{\bar{Y} - 45}{2/\sqrt{20}} \sim N(0,1)$$

The observed value of D is

$$d = \frac{|44.405 - 45|}{2/\sqrt{20}} = 1.33$$

and

$$\begin{aligned} p - value &= P(D \geq d; H_0) \\ &= P(|Z| \geq 1.33) \quad \text{where } Z \sim N(0,1) \\ &= 2\left[1 - P(Z \leq 1.33)\right] = 2(1 - 0.90824) \\ &= 0.18352 \end{aligned}$$

Since $p - value > 0.1$ there is no evidence to contradict the manufacturer's claim that $H_0 : \mu = 45$ based on the data.

A 95% confidence interval for μ is given by

$$\begin{aligned} &\left[44.405 - 1.96(2)/\sqrt{20}, 44.405 + 1.96(2)/\sqrt{20}\right] \\ = \quad &[43.52, 45.29] \end{aligned}$$

5.5 To test the hypothesis $H_0 : \mu = 105$ we use the discrepancy measure or test statistic

$$D = \frac{\left|\bar{Y} - 105\right|}{S/\sqrt{12}}$$

where

$$S = \left[\frac{1}{11}\sum_{i=1}^{12}\left(Y_i - \bar{Y}\right)^2\right]^{1/2}$$

and

$$T = \frac{\bar{Y} - 105}{S/\sqrt{12}} \sim t(11)$$

assuming the hypothesis $H_0 : \mu = 105$ is true.
For these data $\bar{y} = 104.13$, $s^2 = 88.3115$ and $s = 9.3974$. The observed value of the discrepancy measure D is

$$d = \frac{|\bar{y} - 105|}{s/\sqrt{12}} = \frac{|104.13 - 105|}{9.3974/\sqrt{12}} = 0.3194$$

and

$$\begin{aligned} p - value &= P(D \geq d; H_0) \\ &= P(|T| \geq 0.3194) \quad \text{where } T \sim t(11) \\ &= 2[1 - P(T \leq 0.3194)] = 2(0.3777) \\ &= 0.7554 \quad \text{calculated using R} \end{aligned}$$

Alternatively using the t table in the Course Notes we have $P(T \leq 0.2596) = 0.6$ and $P(T \leq 0.5399) = 0.7$ so

$$\begin{gathered} 2(1 - 0.7) \leq p - value \leq 2(1 - 0.6) \\ \text{or} \quad 0.6 \leq p - value \leq 0.8 \end{gathered}$$

In either case since the $p - value$ is much larger than 0.1 and we would conclude that, based on the observed data, there is no evidence against the hypothesis $H_0 : \mu = 105$. (Note: This does not imply the hypothesis is true!)

5.6 To test $H_0 : \sigma^2 = \sigma_0^2$ when μ is known we use the test statistic

$$U = \frac{\sum_{i=1}^{12} (Y_i - \mu)^2}{\sigma_0^2} \sim \chi^2(n)$$

For $n = 12$, $\mu = 105$ and $H_0 : \sigma^2 = 100$, we have

$$U = \frac{\sum_{i=1}^{12} (Y_i - 105)^2}{100} \sim \chi^2(12)$$

Since

$$\begin{aligned} \sum_{i=1}^{12} (y_i - 105)^2 &= \sum_{i=1}^{12} y_i^2 - 2(105) \sum_{i=1}^{12} y_i + 12(105)^2 \\ &= 131096.44 - 210(1249.6) + 12(105)^2 = 980.44 \end{aligned}$$

The observed value of U is

$$u = \frac{980.44}{100} = 9.8044$$

and

$$\begin{aligned} p-value &= 2P\left(U \leq 9.8044\right) \quad \text{where } U \sim \chi^2\left(12\right) \\ &= 0.73 \quad \text{calculated using R} \end{aligned}$$

Alternatively using the Chi-squared table in the Course Notes we have $P\left(U \leq 9.034\right) = 0.3$ so $p-value > 2\left(0.3\right) = 0.6$. In either case since the $p-value$ is larger than 0.1 and we would conclude that, based on the observed data, there is no evidence against the hypothesis $H_0 : \sigma^2 = 100$.

From the Chi-squared table $P\left(U \leq 4.404\right) = 0.025$ and $P\left(U \leq 23.337\right) = 0.975$ where $U \sim \chi^2\left(12\right)$. Since

$$\begin{aligned} 0.95 &= P\left(4.404 \leq U \leq 23.337\right) = P\left[4.404 \leq \frac{\sum_{i=1}^{12}\left(Y_i - 105\right)^2}{\sigma^2} \leq 23.337\right] \\ &= P\left[\frac{\sum_{i=1}^{12}\left(Y_i - 105\right)^2}{23.337} \leq \sigma^2 \leq \frac{\sum_{i=1}^{12}\left(Y_i - 105\right)^2}{4.404}\right] \\ &= P\left[\sqrt{\frac{\sum_{i=1}^{12}\left(Y_i - 105\right)^2}{23.337}} \leq \sigma \leq \sqrt{\frac{\sum_{i=1}^{12}\left(Y_i - 105\right)^2}{4.404}}\right] \end{aligned}$$

a 95% confidence interval for σ is given by

$$\left[\sqrt{\frac{\sum_{i=1}^{12}\left(y_i - 105\right)^2}{23.337}}, \sqrt{\frac{\sum_{i=1}^{12}\left(y_i - 105\right)^2}{4.404}}\right] = \left[\sqrt{\frac{980.44}{23.337}}, \sqrt{\frac{980.44}{4.404}}\right] = [6.482, 14.921]$$

5.7 (a) The respondents to the survey are students who heard about the online referendum and then decided to vote. These students may not be representative of all students at the University of Waterloo. For example, it is possible that the students who took the time to vote are also the students who most want a fall study break. Students who don't care about a fall study break probably did not bother to vote. This is an example of sample error. Any online survey such as this online referendum has the disadvantage that the sample of people who choose to vote are not necessarily a representative sample of the study population of interest. The advantage of online surveys is that they are inexpensive and easy to

conduct. To obtain a representative sample you would need to select a random sample of all students at the University of Waterloo. Unfortunately taking such a sample would be much more time consuming and costly then conducting an online referendum.

(b) A suitable target population would be the $30,990$ eligible voters. This would also be the study population. Note that all undergraduates were able to vote but it is not clear how the list of undergraduates is determined.

(c) The attribute of interest is the proportion of the $30,990$ eligible voters (the study population) who would respond yes to the question. The parameter θ in the Binomial model corresponds to this attribute. A Binomial model assumes independent trials (students) which might not be a valid assumption. For example, if groups of students, say within a specific faculty, all got together and voted, their responses may not be independent events.

(d) The maximum likelihood estimate of θ based on the observed data is

$$\hat{\theta} = \frac{4440}{6000} = 0.74$$

Since this estimate is not based on a random sample it is not possible to say how accurate this estimate is.

(e) An approximate 95% confidence interval for θ is given by

$$0.74 \pm 1.96\sqrt{\frac{0.74\,(0.26)}{6000}} = 0.74 \pm 0.01 = [0.73, 0.75]$$

(f) Since $\theta = 0.7$ is not a value contained in the approximate 95% confidence interval $[0.73, 0.75]$ for θ, therefore the approximate $p - value$ for testing $H_0 : \theta = 0.7$ is less than 0.05. (Note that since $\theta = 0.7$ is far outside the interval, the $p - value$ would be much smaller than 0.05.)

5.8 (a) If $H_0 : \theta = 3$ is true then since Y_i has a Poisson distribution with mean 3, $i = 1, 2, \ldots, 25$ independently, then $\sum_{i=1}^{25} Y_i$ has a Poisson distribution with mean $3 \times 25 = 75$. The discrepancy measure

$$D = \left| \sum_{i=1}^{25} Y_i - 75 \right| = \left| \sum_{i=1}^{25} Y_i - E\left(\sum_{i=1}^{25} Y_i \right) \right|$$

is reasonable since it is measuring the agreement between the data and $H_0 : \theta = 3$ by using the distance between the observed value of $\sum_{i=1}^{25} Y_i$ and its expected value $E\left(\sum_{i=1}^{25} Y_i \right) = 75$.

For the given data, $\sum_{i=1}^{25} y_i = 51$. The observed value of the discrepancy measure is

$$d = \left| \sum_{i=1}^{25} y_i - 75 \right| = |51 - 75| = 24$$

and

$$\begin{aligned} p - value &= P\left(D \geq d; H_0\right) \\ &= P\left(\left| \sum_{i=1}^{25} Y_i - 75 \right| \geq 24; H_0\right) \\ &= \sum_{x=0}^{51} \frac{75^x e^{-75}}{x!} + \sum_{x=99}^{\infty} \frac{75^x e^{-75}}{x!} \\ &= 1 - \sum_{x=52}^{98} \frac{75^x e^{-75}}{x!} \\ &= 0.006716 \quad \text{calculated using R} \end{aligned}$$

Since $0.001 < p - value < 0.01$ we would conclude that there is strong evidence against the hypothesis $H_0 : \theta = 3$ based on the data.

(b) If Y_i has a Poisson distribution with mean θ and variance θ, $i = 1, 2, \ldots, n$ independently then by the Central Limit Theorem

$$\frac{\bar{Y} - E\left(\bar{Y}\right)}{\sqrt{Var\left(\bar{Y}\right)}} = \frac{\bar{Y} - \theta}{\sqrt{\theta/n}}$$

has approximately a $N\left(0, 1\right)$ distribution.

(c) If $H_0 : \theta = 3$ is true then $E\left(\bar{Y}\right) = 3$. The discrepancy measure $D = \left|\bar{Y} - 3\right|$ is reasonable for testing $H_0 : \theta = 3$ since it is measuring the agreement between the data and $H_0 : \theta = 3$ by using the distance between the observed value of $\bar{Y}$ and its expected value $E\left(\bar{Y}\right) = 3$.

The observed value of the discrepancy measure is

$$d = |\bar{y} - 3| = \left| \frac{51}{25} - 3 \right| = |2.04 - 3| = 0.96$$

and also

$$\frac{|\bar{y} - 3|}{\sqrt{3/25}} = \frac{0.96}{\sqrt{3/25}} = 2.77$$

Therefore

$$\begin{aligned} p - value &= P\left(D \geq d; H_0\right) \\ &= P\left(\left|\bar{Y} - 3\right| \geq 0.96; H_0\right) \\ &\approx P\left(|Z| \geq \frac{0.96}{\sqrt{3/25}}\right) \quad \text{where } Z \sim N\left(0, 1\right) \\ &= 2\left[1 - P\left(Z \leq 2.77\right)\right] = 0.005584 \end{aligned}$$

The approximate $p-value$ of 0.005584 is close to the $p-value$ calculated in (a) which is the exact $p-value$. Since we are only interested in whether the $p-value$ is bigger than 0.1 or between 0.1 and 0.05 etc. we are not as worried about how good the approximation is. In this example the conclusion about H_0 is the same for the approximate $p-value$ as it is for the exact $p-value$.

5.9 The observed value of the likelihood ratio test statistic for testing $H_0 : \theta = 3$ is

$$\begin{aligned}\lambda(3) &= -2\log R(3) = -2\log\left[\left(\frac{3}{2.04}\right)^{51} e^{25(2.04-3)}\right] \\ &= -2\log(0.01315) = 8.6624\end{aligned}$$

and

$$\begin{aligned}p-value &= P(\Lambda(3) \geq 8.6624; H_0) \\ &\approx P(W \geq 8.6624) \quad \text{where } W \sim \chi^2(1) \\ &= P\left(|Z| \geq \sqrt{8.6624}\right) \quad \text{where } Z \sim N(0,1) \\ &= 2[1 - P(Z \leq 2.94)] = 0.00328\end{aligned}$$

The $p-value$ is close to the $p-values$ calculated in $8(a)$ and $8(c)$.

5.10 Since

$$R(\theta) = \left[\frac{3.6}{\theta} e^{(1-3.6/\theta)}\right]^{20} \quad \text{for } \theta > 0$$

then

$$R(5) = \left[\frac{3.6}{5} e^{(1-3.6/5)}\right]^{20} = 0.3791$$

The observed value of the likelihood ratio test statistic for testing $H_0 : \theta = 5$ is

$$\lambda(5) = -2\log R(5) = -2\log(0.3791) = 1.9402$$

Therefore

$$\begin{aligned}p-value &\approx P(W \geq 1.9402) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{1.9402}\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2[1 - P(Z \leq 1.39)] = 2(1 - 0.91774) = 0.16452\end{aligned}$$

and since $p-value > 0.1$ there is no evidence against $H_0 : \theta = 5$ based on the data. The approximate 95% confidence interval for θ is $[2.40, 5.76]$ which contains the value $\theta = 5$. This also implies that the $p-value > 0.05$ and so the approximate confidence interval is consistent with the test of hypothesis.

5.11 Since

$$r(\theta) = 15\log[2.3(\theta+1)] - 34.5(\theta+1) + 15 \quad \text{for } \theta > -1$$

then

$$r(-0.1) = 15\log[2.3(-0.1+1)] - 34.5(-0.1+1) + 15 = -5.1368$$

The observed value of the likelihood ratio test statistic for testing $H_0 : \theta = -0.1$ is

$$\lambda(-0.1) = -2r(-0.1) = -2(-5.1368) = 10.2735$$

Therefore

$$\begin{aligned} p-value &\approx P(W \geq 10.2735) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{10.2735}\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2[1 - P(Z \leq 3.21)] = 2(1 - 0.99934) = 0.00132 \end{aligned}$$

and since $0.001 < p-value < 0.01$ there is strong evidence against $H_0 : \theta = -0.1$ based on the data. The approximate 95% confidence interval for θ is $[-0.75, -0.31]$ which does not contain the value $\theta = -0.1$. This also implies that the $p-value < 0.05$ and so the approximate confidence interval is consistent with the test of hypothesis.

5.12 Since

$$R(\theta) = \frac{\theta^{16}(1-\theta)^{66}}{(8/41)^{16}(33/41)^{66}} \quad \text{for } 0 < \theta \leq \frac{1}{2}$$

then

$$R(0.18) = \frac{(0.18)^{16}(1-0.18)^{66}}{(8/41)^{16}(33/41)^{66}} = 0.9397$$

The observed value of the likelihood ratio test statistic for testing $H_0 : \theta = 0.18$ is

$$\lambda(0.18) = -2\log R(0.18) = -2\log(0.9397) = 0.1244$$

Therefore

$$\begin{aligned} p-value &\approx P(W \geq 0.1244) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{0.1244}\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2[1 - P(Z \leq 0.35)] = 2(1 - 0.63683) = 0.72634 \end{aligned}$$

and since $p-value > 0.1$ there is no evidence against $H_0 : \theta = 0.18$ based on the data. The approximate 95% confidence interval for θ is $[0.12, 0.29]$ which contains the value $\theta = 0.18$. This also implies that the $p-value > 0.05$ and so the approximate confidence interval is consistent with the test of hypothesis.

5.13 (a) The maximum likelihood estimate of θ is $\hat{\theta} = 18698.6/20 = 934.93$. The agreement between the plot of the empirical cumulative distribution function and the cumulative distribution function of an Exponential(934.93) random variable given in Figure 5.2 indicates that the Exponential is reasonable.

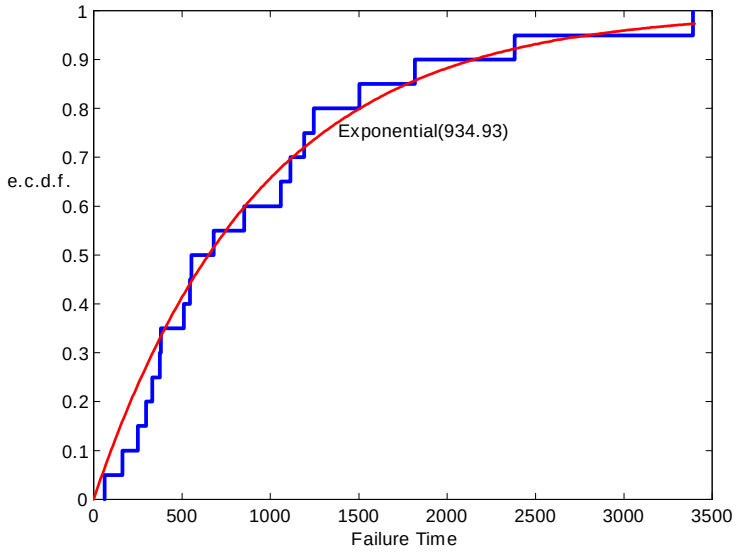


Figure 5.2: Empirical cdf and Exponential(934.93) cdf for failure times

(b) The observed value of the likelihood ratio statistic for testing $H_0 : \theta = \theta_0$ for Exponential data is (see Example 5.3.2) is

$$\lambda(\theta_0) = -2\log R(\theta_0) = -2\log\left[\left(\frac{\hat{\theta}}{\theta_0}\right)^n e^{n\left(1-\hat{\theta}/\theta_0\right)}\right]$$

where $\hat{\theta} = \bar{y}$. For $n = 20$, $\hat{\theta} = \bar{y} = 934.93$ and $\theta_0 = 1000$ we have

$$\lambda(1000) = -2\log\left[\left(\frac{934.93}{1000}\right)^{20} e^{20(1-934.93/1000)}\right] = 2\log(0.95669) = 0.0885$$

with

$$\begin{aligned} p-value &\approx P(W \geq 0.0885) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{0.0885}\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2\left[1 - P(Z \leq 0.30)\right] = 2(1 - 0.61791) = 0.76418 \end{aligned}$$

Since $p - value > 0.1$ there is no evidence against the hypothesis $H_0 : \theta = 1000$ based on the observed data.

5.14 A test statistic that could be used will be to test the mean of the generated sample. The mean should be closed to 0.5 if the random number generator is working well.

5.15 (a) For each given region the assumptions of independence, individuality and homogeneity would need to hold for the number of events per person per year.

(b) Assume the observations $y_1, y_2, \ldots, y_K$ from the different regions are independent. Since $Y_j \sim \text{Poisson}(P_j\theta_j t)$ then the likelihood function for $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ is

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{K} \frac{(P_j\theta_j t)^{y_j}\, e^{-P_j\theta_j t}}{y_j!}$$

or more simply

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{K} \theta_j^{y_j} e^{-P_j\theta_j t}$$

and the log likelihood function is

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{K} [y_j \log\theta_j - P_j\theta_j t]$$

Since

$$\frac{\partial l}{\partial \theta_j} = \frac{y_j}{\theta_j} - P_j t = \frac{y_j - (P_j t)\,\theta_j}{\theta_j} = 0$$

for $\theta_j = y_j/(P_j t)$, the maximum likelihood estimate of θ_j is $\hat{\theta}_j = y_j/(P_j t)$, $j = 1, 2, \ldots, K$. So

$$\begin{aligned} l(\hat{\boldsymbol{\theta}}) &= \sum_{j=1}^{K} \left[y_j \log\hat{\theta}_j - P_j\hat{\theta}_j t\right] = \sum_{j=1}^{K} \left[y_j \log\left(\frac{y_j}{P_j t}\right) - y_j\right] \\ &= \sum_{j=1}^{K} y_j \left[\log\left(\frac{y_j}{P_j t}\right) - 1\right] \end{aligned}$$

The likelihood function assuming $H_0 : \theta_1 = \theta_2 = \cdots = \theta_K$ is given by

$$L(\theta) = \prod_{j=1}^{K} \theta^{y_j} e^{-P_j\theta t}$$

with log likelihood function

$$l(\theta) = \left(\sum_{j=1}^{K} y_j\right) \log\theta - \theta t \sum_{j=1}^{K} P_j$$

Since

$$l'(\theta) = \frac{1}{\theta}\sum_{j=1}^{K} y_j - \sum_{j=1}^{K} P_j t = \frac{1}{\theta}\left[\sum_{j=1}^{K} y_j - \theta t \sum_{j=1}^{K} P_j\right] = 0$$

if $\theta = \sum_{j=1}^{K} y_j / \sum_{j=1}^{K} P_j t$, the maximum likelihood estimate of θ assuming

$H_0 : \theta_1 = \theta_2 = \cdots = \theta_K$ is $\hat{\theta}_0 = \sum_{j=1}^{K} y_j / t \sum_{j=1}^{K} P_j$. So

$$\begin{aligned} l(\hat{\theta}_0) &= \left(\sum_{j=1}^{K} y_j\right) \log \hat{\theta}_0 - \hat{\theta}_0 \sum_{j=1}^{K} P_j t \\ &= \left(\sum_{j=1}^{K} y_j\right) \log \left(\frac{\sum_{j=1}^{K} y_j}{t \sum_{j=1}^{K} P_j}\right) - \left(\frac{\sum_{j=1}^{K} y_j}{t \sum_{j=1}^{K} P_j}\right) \sum_{j=1}^{K} P_j t \\ &= \left(\sum_{j=1}^{K} y_j\right) \left[\log \left(\sum_{j=1}^{K} y_j / t \sum_{j=1}^{K} P_j\right) - 1\right] \end{aligned}$$

The likelihood ratio test statistic for testing $H_0 : \theta_1 = \theta_2 = \cdots = \theta_K$ is

$$\begin{aligned} \Lambda &= 2l(\tilde{\boldsymbol{\theta}}) - 2l(\tilde{\theta}_0) \\ &= 2\sum_{j=1}^{K} Y_j \left[\log\left(\frac{Y_j}{P_j t}\right) - 1\right] - 2\left(\sum_{j=1}^{K} Y_j\right) \left[\log \left(\sum_{j=1}^{K} Y_j / t \sum_{j=1}^{K} P_j\right) - 1\right] \end{aligned}$$

The observed value of Λ is

$$\begin{aligned} \lambda &= 2l(\hat{\boldsymbol{\theta}}) - 2l(\hat{\theta}_0) \\ &= 2\sum_{j=1}^{K} y_j \left[\log\left(\frac{y_j}{P_j t}\right) - 1\right] - 2\left(\sum_{j=1}^{K} y_j\right) \left[\log \left(\sum_{j=1}^{K} y_j / t \sum_{j=1}^{K} P_j\right) - 1\right] \end{aligned}$$

The $p - value$ is

$$P\left(\Lambda \geq \lambda; H_0\right) \approx P\left(W \geq \lambda\right) \quad \text{where } W \sim \chi^2\left(K-1\right)$$

(c) For the given data

$$\hat{\boldsymbol{\theta}} = \left(\frac{27}{5\left(2025\right)}, \frac{18}{5\left(1116\right)}, \frac{41}{5\left(3210\right)}, \frac{29}{5\left(1687\right)}, \frac{31}{5\left(2840\right)}\right) \text{ and } \hat{\theta}_0 = \frac{146}{5\left(10878\right)}$$

$\lambda = 3.73$ and $p - value \approx P(W \geq 3.73) = 0.44$ where $W \sim \chi^2\left(4\right)$. Since $p - value > 0.1$ there is no evidence based on the data against $H_0 : \theta_1 = \theta_2 = \cdots = \theta_5$, that is, that the rates are equal.

5.16 Let μ_G = mean of the Poisson model for Gretzky and μ_C = mean of the Poisson model for Crosby. From Example 5.4.3

$$\begin{aligned} l(\boldsymbol{\theta}) &= l(\mu_G, \mu_C) \\ &= n\bar{x}\log\left(\mu_G\right) - n\mu_G + m\bar{y}\log\left(\mu_C\right) - m\mu_C \end{aligned}$$

For the Gretzky data $n = 696$ and $\bar{x} = \frac{1669}{696} = 2.398$. For the Crosby data $m = 783$ and $\bar{y} = \frac{1027}{783} = 1.3116$. Now $\hat{\boldsymbol{\theta}} = (\bar{x}, \bar{y}) = (2.398, 1.3116)$, $\hat{\mu} = \frac{n\bar{x}+m\bar{y}}{n+m} = 1.8229$ and $\hat{\boldsymbol{\theta}}_0 = (\hat{\mu}, \hat{\mu}) = (1.8229, 1.8229)$. The observed value of the likelihood ratio statistic is

$$\begin{aligned}\lambda &= 2\left[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_0)\right] = 2\left[l(2.398, 1.3116) - l(1.8229, 1.8229)\right] \\ &= 2\left(-957.6534 + 1077.313\right) = 239.320\end{aligned}$$

and $p - value \approx P(W \geq 239.320) \approx 0$ where $W \sim \chi^2(1)$. Since $p - value \approx 0$ there is very strong evidence based on the data against the hypothesis of equal means.

5.17 (a) $\tilde{\mu} = \overline{Y}$, $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})$, $\hat{\mu}_0 = \mu_0$, $\tilde{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_0)$, and $\Lambda(\mu_0) = n\log\left(\frac{\tilde{\sigma}_0^2}{\tilde{\sigma}^2}\right)$.

$$\frac{\tilde{\sigma}_0^2}{\tilde{\sigma}^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_0)}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y}) + n(\overline{Y} - \mu_0)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})}$$

so that

$$\Lambda(\mu_0) = n\log\left[1 + \frac{n(\overline{Y} - \mu_0)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})}\right] = n\log\left(1 + \frac{T^2}{n-1}\right)$$

SOLUTIONS TO CHAPTER 6 PROBLEMS

The following identities, proved in the Chapter 1 problems, will be used in problems 1 and 2.

$$\begin{aligned}
0 &= \sum_{i=1}^{n}(y_i-\bar{y}) = \sum_{i=1}^{n}(x_i-\bar{x}) \\
S_{yy} &= \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i(y_i-\bar{y}) = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \\
S_{xx} &= \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i(x_i-\bar{x}) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\
S_{xy} &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} x_i(y_i-\bar{y}) = \sum_{i=1}^{n}(x_i-\bar{x})y_i
\end{aligned}$$

6.1 If $a_i = \frac{(x_i-\bar{x})}{S_{xx}}$ then

$$\begin{aligned}
\sum_{i=1}^{n} a_i &= \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_{xx}} = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}) = \frac{1}{S_{xx}}(0) = 0 \\
\sum_{i=1}^{n} a_i x_i &= \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_{xx}}x_i = \frac{1}{S_{xx}}\sum_{i=1}^{n} x_i(x_i-\bar{x}) = \frac{S_{xx}}{S_{xx}} = 1 \\
\sum_{i=1}^{n} a_i^2 &= \sum_{i=1}^{n}\left[\frac{(x_i-\bar{x})}{S_{xx}}\right]^2 = \left(\frac{1}{S_{xx}}\right)^2\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{S_{xx}}{(S_{xx})^2} = \frac{1}{S_{xx}}
\end{aligned}$$

If $b_i = \frac{1}{n} + (x-\bar{x})\frac{(x_i-\bar{x})}{S_{xx}}$ then

$$\begin{aligned}
\sum_{i=1}^{n} b_i &= \sum_{i=1}^{n}\left[\frac{1}{n} + (x-\bar{x})\frac{(x_i-\bar{x})}{S_{xx}}\right] = \frac{1}{n}\sum_{i=1}^{n} 1 + \frac{(x-\bar{x})}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}) = 1 + 0 = 1 \\
\sum_{i=1}^{n} b_i x_i &= \sum_{i=1}^{n}\left[\frac{1}{n} + (x-\bar{x})\frac{(x_i-\bar{x})}{S_{xx}}\right]x_i = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{(x-\bar{x})}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})x_i \\
&= \bar{x} + \frac{(x-\bar{x})}{S_{xx}}S_{xx} = \bar{x} + (x-\bar{x}) = x
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{n} b_i^2 &= \sum_{i=1}^{n}\left[\frac{1}{n}+(x-\bar{x})\frac{(x_i-\bar{x})}{S_{xx}}\right]^2 \\
&= \left(\frac{1}{n}\right)^2\sum_{i=1}^{n}1+2\left(\frac{1}{n}\right)\frac{(x-\bar{x})}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})+\left(\frac{x-\bar{x}}{S_{xx}}\right)^2\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \left(\frac{1}{n}\right)^2(n)+0+\left(\frac{x-\bar{x}}{S_{xx}}\right)^2 S_{xx} \\
&= \frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}
\end{aligned}$$

6.2

$$\begin{aligned}
0 &= \frac{\partial l}{\partial \alpha}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)=\frac{1}{\sigma^2}\left[\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}\alpha-\beta\sum_{i=1}^{n}x_i\right] \\
&= \frac{1}{\sigma^2}\left[n\left(\frac{\sum_{i=1}^{n}y_i}{n}\right)-n\alpha-n\beta\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)\right]=\frac{n}{\sigma^2}(\bar{y}-\alpha-\beta\bar{x}) \\
\text{so } \alpha &= \bar{y}-\beta\bar{x}
\end{aligned}$$

$$\begin{aligned}
0 &= \frac{\partial l}{\partial \beta}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)\,x_i=\frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_iy_i-\alpha\sum_{i=1}^{n}x_i-\beta\sum_{i=1}^{n}x_i^2\right) \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}x_iy_i-(\bar{y}-\beta\bar{x})\sum_{i=1}^{n}x_i-\beta\sum_{i=1}^{n}x_i^2\right] \quad \text{since } \alpha=\bar{y}-\beta\bar{x} \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}x_i(y_i-\bar{y})+\beta\bar{x}\sum_{i=1}^{n}x_i-\beta\sum_{i=1}^{n}x_i^2\right]=\frac{1}{\sigma^2}\left[\sum_{i=1}^{n}x_i(y_i-\bar{y})-\beta\sum_{i=1}^{n}x_i(x_i-\bar{x})\right] \\
\text{so } \beta &= \frac{\sum_{i=1}^{n}x_i(y_i-\bar{y})}{\sum_{i=1}^{n}x_i(x_i-\bar{x})}=\frac{S_{xy}}{S_{xx}}
\end{aligned}$$

Therefore the maximum likelihood estimates are

$$\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}, \quad \hat{\beta}=\frac{S_{xy}}{S_{xx}}$$

$$\begin{aligned}
\frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2=0 \\
\text{so } \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2
\end{aligned}$$

so the maximum likelihood estimate of σ^2 is

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{\alpha}-\hat{\beta}x_i\right)^2$$

Since

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2 &= \sum_{i=1}^{n}(y_i-\bar{y}+\hat{\beta}\bar{x}-\hat{\beta}x_i)^2 \\ &= \sum_{i=1}^{n}(y_i-\bar{y})^2-2\tilde{\beta}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})+\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= S_{yy}-2\hat{\beta}S_{xy}+\hat{\beta}\left(\frac{S_{xy}}{S_{xx}}\right)S_{xx}=S_{yy}-\hat{\beta}S_{xy}\end{aligned}$$

therefore

$$\hat{\sigma}^2=\frac{1}{n}\left(S_{yy}-\hat{\beta}S_{xy}\right)$$

6.3 (a) The maximum likelihood estimates of α and β are

$$\begin{aligned}\hat{\beta} &= \frac{S_{xy}}{S_{xx}}=\frac{2325.20}{2802.00}=0.8298 \\ \hat{\alpha} &= \bar{y}-\hat{\beta}\bar{x}=133.56-\left(\frac{2325.20}{2802.00}\right)(43.20)=97.7111\end{aligned}$$

and an unbiased estimate of σ^2 is

$$s_e^2=\frac{1}{n-2}(S_{yy}-\hat{\beta}S_{xy})=\frac{1}{23}\left[3284.16-\left(\frac{2325.20}{2802.00}\right)(2325.20)\right]=58.89677$$

(b) The scatterplot with fitted line and the residual plots shown in Figure 6.1.The

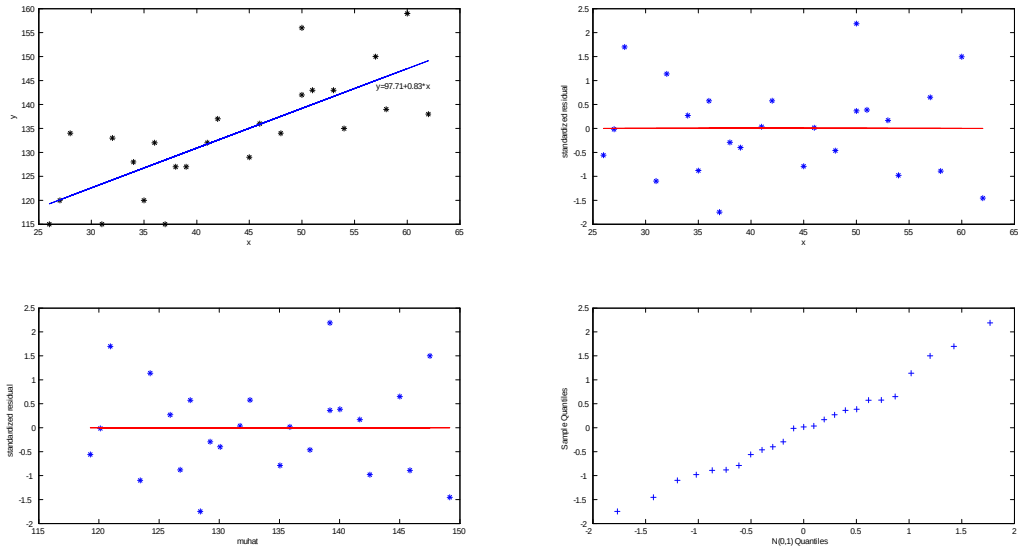


Figure 6.1: Scatterplot and residual plots for nurses data

scatterplot is checking whether the response variate can be modeled by a random

variable whose mean is a linear function of the explanatory variate and whose standard deviation is constant over the range of values of the explanatory variate. If these assumptions hold then we would expect to see the observed points lying reasonably along the fitted line and the variability about the fitted line being reasonably constant over the range of values of the explanatory variate. For these data the points do lie rougly along the fitted line and the variability about the curve is reasonably constant. Note that there is quite a bit of variability about the fitted line.

The plot of the standardized residuals versus the explanatory variate and the plot of the standardized residuals versus the fitted values are both checking whether the response variate can be modeled by a Gaussian random variable whose mean is a linear function of the explanatory variate and whose standard deviation is constant over the range of values of the explanatory variate. If these assumptions hold then we would expect to see the standardized residuals lying in roughly a horizontal band between -3 and $+3$ and cut roughly in half by the line $r^* = 0$. For these data the points do lie in such a band about the line $r^* = 0$.

The qqplot of the standardized residuals is checking whether the response variate can be modeled by a Gaussian random variable whose mean is a linear function of the explanatory variate and whose standard deviation is constant over the range of values of the explanatory variate. In other words the qqplot is checking whether the standardized residuals reasonably follow a $G(0,1)$ distribution. If these assumptions hold then we would expect to see the points scattered about the straight line with more variability about the line at each end. For these data the points do lie reasonably along the line with more variability in the tails.

Based on these observations we would conclude that the simple linear regression model fits these data well. We note however that the sample size of 25 is not very large.

(c) Since $P(T \leq 2.0687) = 0.975$ where $T \sim t\,(23)$ and

$$s_e = \left[\frac{1}{n-2}(S_{yy} - \hat{\beta} S_{xy})\right]^{1/2} = 7.674423$$

therefore a 95% confidence interval for β is

$$\hat{\beta} \pm 2.0687\,(7.674423)\,/\sqrt{2802.0} = 0.8298 \pm 0.2999 = [0.53, 1.13]$$

(d) Since $P(T \leq 1.7139) = 0.95$ where $T \sim t\,(23)$, a 90% confidence interval for the mean systolic blood pressure of nurses aged $x = 35$ is

$$\begin{aligned} & \hat{\alpha} + \hat{\beta}(35) \pm 1.7139\,(7.6744)\left[\frac{1}{25} + \frac{(35 - 43.20)^2}{2802.00}\right]^{1/2} \\ = \;& 126.7553 \pm 3.3274 = [123.43, 130.08] \end{aligned}$$

(e) Since $P(T \leq 2.8073) = 0.995$ where $T \sim t\,(23)$, a 99% prediction interval for the systolic blood pressure of a nurse aged $x = 50$ is

$$\begin{aligned} & \hat{\alpha} + \hat{\beta}(50) \pm 2.8073\,(7.6744)\left[1 + \frac{1}{25} + \frac{(50 - 43.20)^2}{2802.00}\right]^{1/2} \\ = \quad & 139.2029 \pm 22.14463 = [117.06, 161.35] \end{aligned}$$

6.4

(a) This is an observational study since the person conducting the study is not in control of the STAT 230 final grades (the explanatory variate).

(b) A possible Problem for this study is to study the relationship between STAT 231 final grades and STAT 230 final grades. In particular the researcher might want to know, since STAT 230 is a prerequisite for STAT 231, whether students with higher (lower) STAT 230 final grades also have higher (lower) STAT 231 final grades. This is a descriptive Problem.

(c) A unit in this study is a student. A suitable target population for this study would be all students enrolled in STAT 231 in the winter term 2013 and in all subsequent terms.

(d) There are two variates which are STAT 230 final grade and STAT 231 final grade. These variates are both discrete variates since only integer grades are assigned.

(e) It makes sense to define $x =$ STAT 230 final grade as the explanatory variate and $y =$ STAT 231 final grade as the response variate since STAT 230 is a prerequisite for STAT 231 and the researcher is interested in how a student's STAT 230 final grades affects their STAT 231 final grade.

(f) A suitable study population for this study would be all students enrolled in STAT 231 in the winter term 2013. A possible source of study error might be that the students enrolled in winter 2013 were systematically different from students taking STAT 231 in subsequent terms with respect to the attributes of interest.

(g) The sampling protocol was to select 30 students at random from all students enroled in STAT 231 in winter 2013.

(h) It important for the students to be chosen at random from the group of students taking STAT 231 to ensure that the students in the sample are representative of all the students enroled in STAT 231 in winter 2013. If the first 30 students in an alphabetized list of all students were chosen then this could be a source of sample error since it is possible that the students with last names beginning with letters at the beginning of the alphabet are systematically different (higher or lower STAT marks on average) from the other students.

(i) The variates are measured by taking the students grades during the term and calculating a final grade according to the marking scheme. A possible source of measurement error might be that certain grades were recorded or calculated incorrectly.

(j) The sample correlation is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.8182$$

(k) See the top left plot in Figure 6.2 for a scatterplot of the data. The points lie reasonably about a straight line although there is quite a bit of variability about the line.

(l) Fitted line: $y = -4.0667 + 0.9944x$

Least squares estimate of α: $\hat{\alpha} = -4.0667$

Maximum likelihood estimate of β: $\hat{\beta} = 0.9944$

Unbiased estimate of σ: $s_e = 9.4630$

(m) The scatterplot with fitted line and the residual plots shown in Figure 6.2 show no unusual patterns. The size of the dataset is quite small. The simple linear regression model fits the data well.

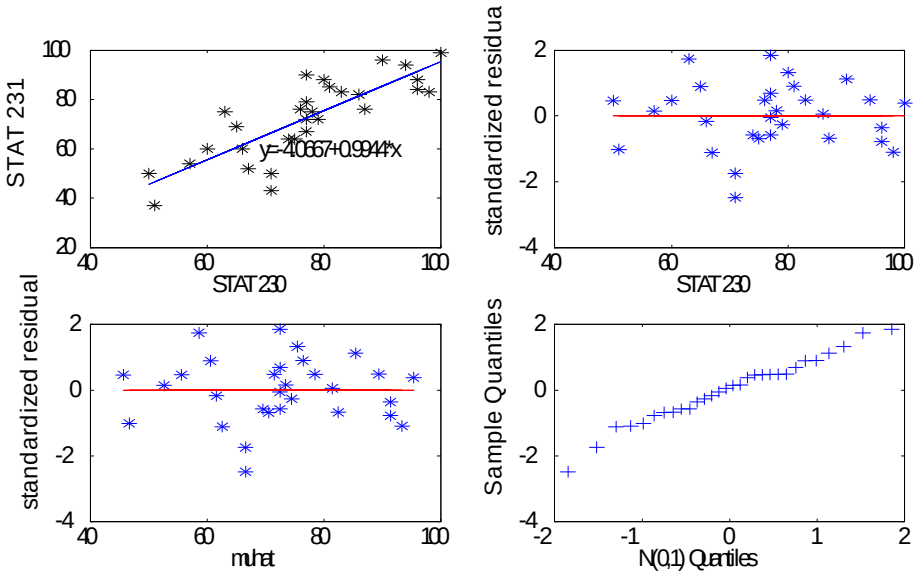


Figure 6.2: Scatterplot and residual plots for STAT 230/231 final grades

(n) The parameter $\mu(x) = \alpha + \beta x$ corresponds to the mean STAT 231 final grade for students with a STAT 230 final grade of x in the study population. The parameter σ represents the variability in the response variate Y in the study

population which is assumed to be the same for each value of the explanatory variate x in the study population.

(o) The parameter β corresponds to the change in the mean STAT 231 final grade in the study population for a one mark increase in STAT 230 final grade.

To test the hypothesis $H_0 : \beta = 0$ we use the test statistic

$$D = \frac{\left|\tilde{\beta} - 0\right|}{S_e/\sqrt{S_{xx}}}$$

For these data the observed value is

$$d = \frac{\left|\hat{\beta} - 0\right|}{s_e/\sqrt{S_{xx}}} = 7.5304$$

and

$$\begin{aligned} p - value &= 2\left[1 - P\left(T \leq 7.5304\right)\right] \quad \text{where } T \sim t\left(28\right) \\ &\approx 0 \end{aligned}$$

Since $p - value \approx 0$ there is very strong evidence based on the data against the hypothesis $H_0 : \beta = 0$. Since the simple linear regression model is reasonable for these data, the data suggest that there is a linear relationship between STAT 231 final grades and STAT 230 final grades. This relationship is also observed in the scatterplot.

(p) To test the hypothesis that there is no relationship $H_0 : \beta = 1$ we use the test statistic

$$D = \frac{\left|\tilde{\beta} - 1\right|}{S_e/\sqrt{S_{xx}}}$$

For these data the observed value is

$$d = \frac{\left|\hat{\beta} - 1\right|}{s_e/\sqrt{S_{xx}}} = 0.0428$$

and

$$\begin{aligned} p - value &= 2\left[1 - P\left(T \leq 0.0428\right)\right] \quad \text{where } T \sim t\left(28\right) \\ &= 0.9662 \quad \text{calculated using R} \end{aligned}$$

Since $p - value \gg 0.1$ therefore there is no evidence based on the data against the hypothesis $\beta = 1$.

The hypothesis $H_0 : \beta = 1$ means that we are hypothesizing that, in the study population, for every one mark increase in STAT 230 final grade there is a one mark increase in the mean STAT 231 final grade.

(q) Since $P(T \leq 2.0484) = (1 + 0.95)/2 = 0.975$ where $T \sim t(28)$ a 95% confidence interval for β is

$$\hat{\beta} \pm 2.0484\,(9.4630)\,/\sqrt{5135.8667} = 0.9944 \pm 0.2705 = [0.7239, 1.2648]$$

Since this interval contains the value $\beta = 1$ but does not contain the value $\beta = 0$, the confidence interval is consistent with the $p - values$ determined in (o) and (p).

Interpretation of the confidence interval: Suppose we were able to repeat the experiment (select 30 students at random from the study population and record their STAT 230 and STAT 231 final grades) a large number of times and each time we construct a 95% confidence interval for β for the observed data. Then, approximately 95% of the constructed intervals would contain the true, but unknown value of β. We say that we are 95% confident that our interval $[0.7239, 1.2648]$ contains the true value of β.

(r) A 95% confidence interval for the mean STAT 231 final grade for students with a STAT 230 final grade of $x = 75$ is

$$\begin{aligned} & -4.0667 + (0.9944)\,(75) \pm 2.0484\,(9.4630)\left[\frac{1}{30} + \frac{(75 - 76.7333)^2}{5135.8667}\right]^{1/2} \\ = \;& 70.50979 \pm 3.56994 = [66.9, 74.1] \end{aligned}$$

Note: Be sure to use all decimal places during your calculations otherwise rounding will give different final results.

A 95% confidence interval for the mean STAT 231 final grade for students with a STAT 230 final grade of $x = 50$ is

$$\begin{aligned} & -4.0667 + (0.9944)\,(50) \pm 2.0484\,(9.4630)\left[\frac{1}{30} + \frac{(50 - 76.7333)^2}{5135.8667}\right]^{1/2} \\ = \;& 45.65095 \pm 8.05046 = [37.6, 53.7] \end{aligned}$$

The 95% confidence interval for $x = 50$ if wider. This is because there are fewer observations near $x = 50$ which is close to the smallest observed value of x while $x = 75$ is very close to the mean value of x. There are many more observations close to the mean of the explanatory variate. Therefore there is less uncertainty in the estimates of the mean response near the center of the observed data.

(s) A 95% prediction interval for STAT 231 final grade for a student with a STAT 230 final grade of $x = 75$ is

$$\begin{aligned} & -4.0667 + (0.9944)\,(75) \pm 2.0484\,(9.4630)\left[1 + \frac{1}{30} + \frac{(75 - 76.7333)^2}{5135.8667}\right]^{1/2} \\ = \;& 70.51 \pm 19.71 = [50.8, 90.2] \end{aligned}$$

This interval is much wider than the interval in (r). The interval is wider because it is an interval for a future observation (random variable) whereas the interval in (r) is an interval for an unknown constant. The interval is particularly wide because the estimate of σ which is $s_e = 9.4630$ marks is quite large. Recall that the parameter σ corresponds to the variability about the line $y = \alpha + \beta x$ which we have already observed is large for these data.

One way to obtain a better prediction would be to collect data on more explanatory variates which could possibly explain the variability in STAT 231 final grades better than just using one explanatory variate.

(t) A 90% confidence interval for σ is

$$\left[(9.462966)\sqrt{\frac{30-2}{41.33714}},\ (9.462966)\sqrt{\frac{30-2}{16.92788}}\right] = [7.788182, 12.17041]$$

In the context of this study the parameter σ represents the variability in STAT 231 final grades for a given STAT 230 final grade. The estimate of σ which is $s_e = 9.462966$ indicates that there is large variability in STAT 231 final grades for a given STAT 230 final grade.

6.5 (a) The maximum likelihood estimate of α and β are

$$\begin{aligned}\hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{22769.645}{6283.422} = 3.6238\\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 187.975 - \left(\frac{22769.645}{6283.422}\right)(43.03) = 32.0444\end{aligned}$$

The fitted line is $y = 32.04 + 3.62x$.

(b) The scatterplot with fitted line and the residual plots shown in Figure 6.3 show no unusual patterns. There is one residual value which is larger than 3 for $x = 50.3$. The simple linear regression model fits these data.

(c) Since

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \text{sample correlation } = r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

therefore

$$r = \hat{\beta}\left(\frac{S_{xx}}{S_{yy}}\right)^{1/2} \quad \text{or} \quad \hat{\beta} = r\left(\frac{S_{yy}}{S_{xx}}\right)^{1/2}$$

(d) Since $P(T \le 2.1009) = 0.975$ where $T \sim t(18)$ and

$$s_e = \left[\frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy})\right]^{1/2} = 100.6524$$

therefore a 95% confidence interval for β is

$$\hat{\beta} \pm 2.1009\,(100.6524)\,/\sqrt{6283.422} = 3.6238 \pm 2.6677 = [0.9561, 6.2915]$$

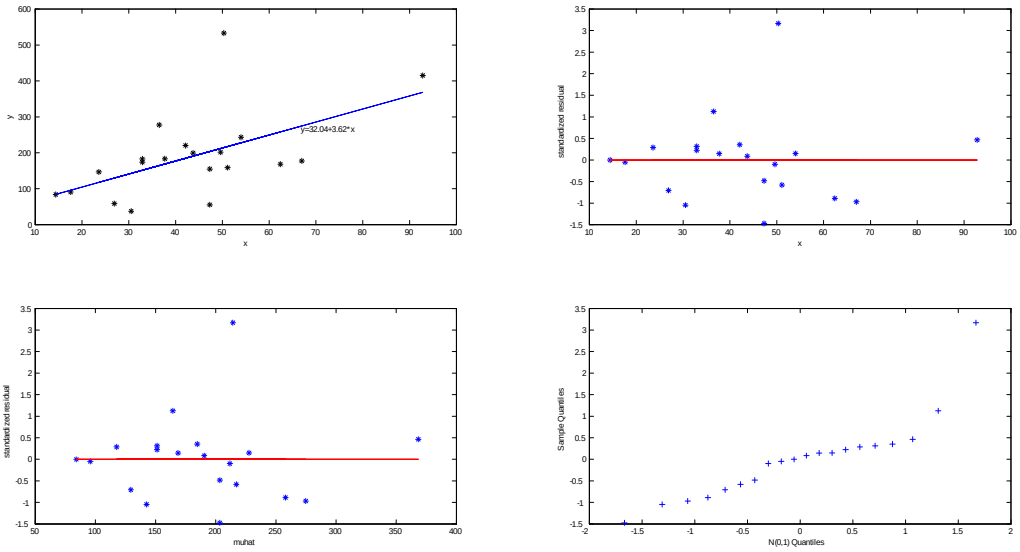


Figure 6.3: Scatterplot and residual plots for actor data

The study population is all the actors listed at boxofficemojo.com/people/. The parameter β represents the mean change in the amount grossed by a movie for a unit change in the value of an actor. However, since the 20 data points were obtained by taking the first 20 actors in the list, the sample is not a random sample. If actors with last names starting with letters at the beginning of the alphabet are more or less successful then other actors then the estimate of β might be biased.

(e) To test $H_0 : \beta = 0$ we use the test statistic

$$D = \frac{\left|\tilde{\beta} - 0\right|}{S_e/\sqrt{S_{xx}}}$$

For these data the observed value is

$$d = \frac{\left|\hat{\beta} - 0\right|}{s_e/\sqrt{S_{xx}}} = 2.85$$

and

$$\begin{aligned} p - value &= 2\left[1 - P\left(T \leq 2.85\right)\right] \quad \text{where } T \sim t\left(18\right) \\ &= 0.0106 \quad \text{calculated using R} \end{aligned}$$

Since $0.01 < p - value < 0.05$ there is evidence against $H_0 : \beta = 0$ based on the data. Note that this is consistent with the fact that the 95% confidence

interval for β does not contain the value $\beta = 0$. Based on this $p - value$ and the scatterplot, we would conclude that the data suggest a linear relationship between the amount grossed by a movie and the value of an actor.

(f) Since $P(T \leq 2.1009) = 0.975$ where $T \sim t\,(18)$, a 95% confidence interval for the mean amount grossed by movies for actors whose value is $x = 50$ is

$$\begin{aligned} & 32.0444 + (3.6238)\,(50) \pm 2.1009\,(100.6524)\left[\frac{1}{20} + \frac{(50 - 43.03)^2}{6283.422}\right]^{1/2} \\ = \;& 213.2326 \pm 50.8090 = [162.4236, 264.0417] \end{aligned}$$

A 95% confidence interval for the mean amount grossed by movies for actors whose value is $x = 100$ is

$$\begin{aligned} & 32.0444 + (3.6238)\,(100) \pm 2.1009\,(100.6524)\left[\frac{1}{20} + \frac{(100 - 43.03)^2}{6283.422}\right]^{1/2} \\ = \;& 394.4209 \pm 159.1644 = [235.2565, 553.5853] \end{aligned}$$

The largest observed x value is $x = 92.8$. By constructing a confidence interval for the mean amount grossed by movies for actors whose value is $x = 100$, we are assuming that the linear relationship hold beyond the observed data.

6.6 For these data

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{-3316.6771}{22.9453} = -144.5469$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 548.9700 - (-144.5469)\,(0.9543) = 686.9159$$
$$s_e^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy}) = \frac{1}{28}\left[489624.723 - (-144.5469)\,(-3316.6771)\right] = 364.6199$$
$$s_e = 19.0950$$

(a) The fitted line is $y = 686.9159 - 144.5469$

(b) See Figure 6.4.

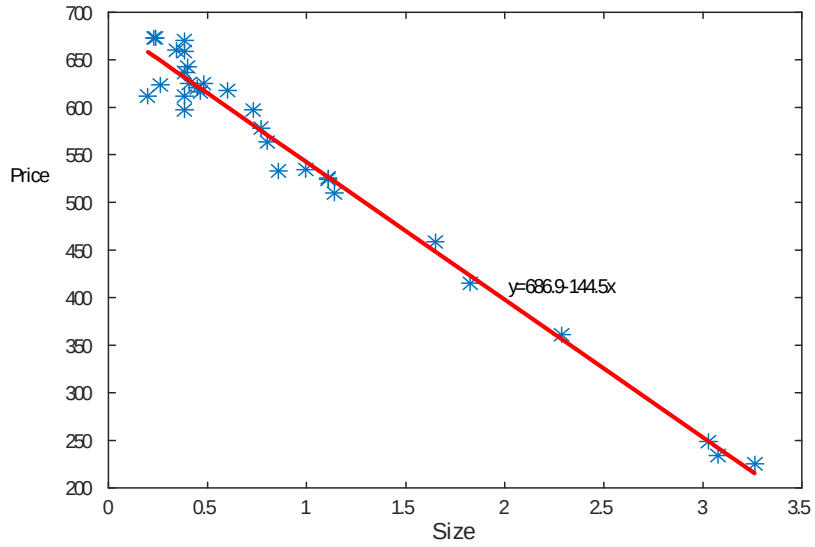


Figure 6.4: **Scatterplot and fitted line for building price versus size**

(c) Since $\hat{\beta}$ is negative this implies that the larger sized buildings tend to sell for less per square meter. The estimate $\hat{\beta} = -144.55$ indicates a drop in average price of \$144.55 per square meter for each increase of one unit in x; remember x's units are $m^2(10^5)$.

(d) Since $P(T \leq 2.0484) = 0.975$ for $T \sim t\,(28)$, a 95% confidence interval for $\mu(4.47)$ is

$$\begin{aligned} & \hat{\mu}(4.47) \pm 2.0484 s_e \sqrt{\frac{1}{30} + \frac{(4.47 - \bar{x})^2}{S_{xx}}} \\ = \; & \$40.79 \pm \$29.58 \\ = \; & [\$11.21,\ \$70.37] \end{aligned}$$

(e) A 95% prediction interval for Y when $x = 4.47$ is

$$\begin{aligned} & \tilde{\mu}(4.47) \pm 2.0484 s_e \sqrt{1 + \frac{1}{30} + \frac{(4.47 - \bar{x})^2}{22.945}} \\ = \; & \$40.79 \pm \$49.04 \\ = \; & [-\$8.25,\ \$89.83] \end{aligned}$$

The lower limit is negative, which is nonsensical. This happened because we were using a Gaussian model (Gaussian random variables Y can be positive or negative) in a setting where the price Y must be positive. Nonetheless, the

Gaussian model fits the data reasonably well. We might just truncate the prediction interval and take it to be [0, \$89.83].

(f) It is better to use the predication interval since we are only interested in the assessed value for one building. Note however that the value $x = 4.47$ is well outside the interval of observed x values which was $[0.20, 3.26]$) in the data set of 30 buildings. Thus any conclusions we reach are based on an assumption that the linear model $E(Y|x) = \alpha + \beta x$ applies beyond $x = 3.26$ at least as far as $x = 4.47$. This may or may not be true, but we have no way to check it with the data we have.

6.7 (a) Recall this was a regression of the form $E(Y_i) = \alpha + \beta x_{1i}$ where $x_{1i} = x_i^2$, and x_i = bolt diameter. Now $n = 30$, $\hat{\alpha} = 1.6668$, $\hat{\beta} = 2.8378$, $s_e = 0.05154$, $S_{xx} = 0.2244$, $\bar{x}_1 = 0.11$. A point estimate of the mean breaking strength at $x_1 = (0.35)^2 = 0.1225$ is

$$\hat{\mu}(0.1225) = \hat{\alpha} + \hat{\beta}(0.1225) = 1.667 + 2.838(0.1225) = 2.01447$$

A confidence interval for $\mu(0.1225)$ is

$$\hat{\mu}(0.1225) \pm as_e\sqrt{\frac{1}{n} + \frac{(0.1225 - \bar{x}_1)^2}{S_{xx}}}$$

From the t table, $P(T \leq 2.0484) = 0.975$ where $T \sim t(28)$. The 95% confidence interval is

$$\begin{aligned} & 2.01447 \pm 2.0484(0.05154)\sqrt{\frac{1}{30} + \frac{(0.1225 - 0.11)^2}{0.2244}} \\ = \quad & 2.01447 \pm 0.01932 = [1.9952, 2.0338] \end{aligned}$$

(b) A 95% prediction interval for the strength at $x_1 = (0.35)^2 = 0.1225$ is

$$\begin{aligned} & \hat{\mu}(0.1225) \pm as_e\sqrt{1 + \frac{1}{n} + \frac{(0.1225 - \bar{x}_1)^2}{S_{xx}}} \\ = \quad & 2.01447 \pm 2.0484(0.05154)\sqrt{1 + \frac{1}{30} + \frac{(0.1225 - \bar{x}_1)^2}{0.2244}} \\ = \quad & 2.01447 \pm 0.10732 = [1.9072, 2.1218] \end{aligned}$$

This interval is wider since it is an interval estimate for a single observation (a random variable) at $x_1 = 0.35$ rather than an interval estimate for a mean (a constant).

(c) Since Y represents the mean strength of the bolt of diameter $x = 0.35$, then based on the assumed model $Y \sim G(\alpha + \beta(0.1225), \sigma)$. Since α, β and σ are

unknown we estimate them using $\hat{\alpha} = 1.6668$, $\hat{\beta} = 2.8378$, and $s_e = 0.05154$ and use $Y \sim G\,(2.01447, 0.05154)$. Since $V \sim G\,(1.60, 0.10)$ independently of $Y \sim G\,(2.01447, 0.05154)$ then $V - Y \sim G\left(1.60 - 2.01447, \sqrt{(0.1)^2 + (0.05154)^2}\right)$ or $V - Y \sim G\,(-0.41447, 0.1125)$. Therefore an estimate of $P\,(V > Y)$ is

$$\begin{aligned} \hat{P}\,(V > Y) &= \hat{P}\,(V - Y > 0) = P\left(Z > \frac{0 - (-0.41447)}{0.1125}\right) \quad \text{where } Z \sim G\,(0,1) \\ &= 1 - P\,(Z \leq 3.68) \approx 0 \end{aligned}$$

6.8 (a)

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{2818.556835}{2818.946855} = 0.9999$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 23.5505 - 23.7065 \times 0.9999 = -0.1527$$

The scatterplot with fitted line and the residual plots shown in Figure 6.5 show no unusual patterns. The model fits the data well.

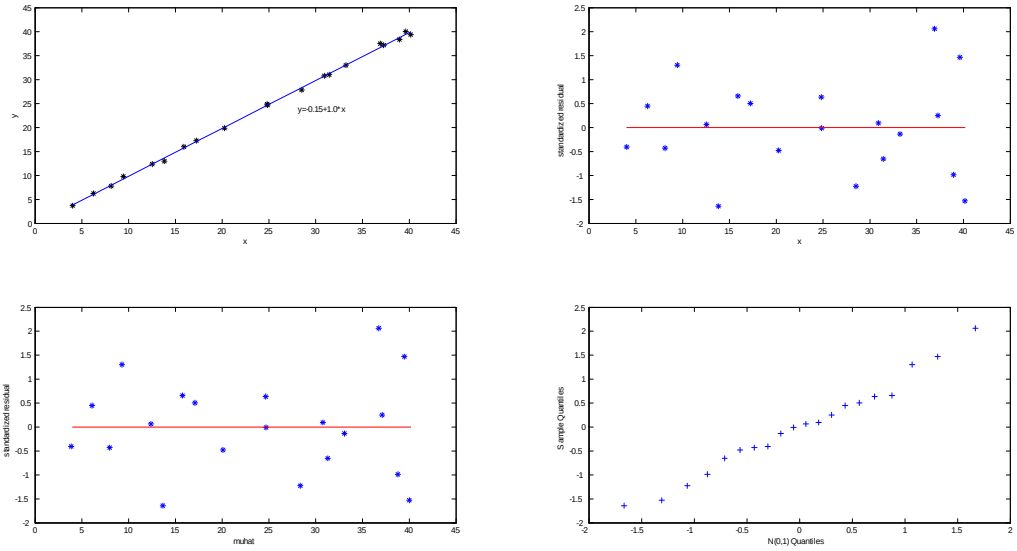


Figure 6.5: Scatterplot and residual plots for cheap versus expensive procedures

(b) Since $P(T \leq 2.1009) = 0.975$ where $T \sim t\,(18)$ and

$$\begin{aligned} s_e &= \left(\frac{S_{yy} - \hat{\beta} S_{xy}}{n-2}\right)^{1/2} \\ &= \left[\frac{2820.862295 - (0.9998616)\,(2818.556835)}{18}\right]^{1/2} = 0.3870 \end{aligned}$$

a 95% confidence interval for β is

$$0.9999 \pm 2.1009\,(0.3870)\,/\sqrt{2818.946855} = [0.9845,\ 1.0152]\,.$$

Since the value $\beta = 1$ is inside the 95% confidence interval for β we know the $p - value$ for testing $H_0 : \beta = 1$ is greater than 0.05. Alternatively

$$p - value = 2\left[1 - P\left(T \leq \frac{\left|\hat{\beta} - 1\right|}{s_e/\sqrt{S_{xx}}}\right)\right] = 2\left[1 - P\left(T \leq 0.019\right)\right] = 0.99$$

Since $p - value > 0.1$ there is no evidence against $H_0 : \beta = 1$ based on the data. A 95% confidence interval for α is

$$-0.1527 \pm 2.1009(0.3870)\sqrt{\frac{1}{20} + \frac{(0 - 23.7065)^2}{2818.946855}} = [-0.5587,\ 0.2533]$$

Since $\alpha = 0$ is inside the 95% confidence interval for α we know the $p - value$ for testing $H_0 : \alpha = 0$ is greater than 0.05. Alternatively

$$p{-}value = 2\left[1 - P\left(T \leq \frac{|\hat{\alpha} - 0|}{s_e\sqrt{\frac{1}{n} + \frac{(0-\bar{x})^2}{S_{xx}}}}\right)\right] = 2\left[1 - P\left(T \leq 0.7903\right)\right] = 0.4396$$

Since $p - value > 0.1$ there is no evidence against $H_0 : \alpha = 0$ based on the data. The question of interest is how well the cheaper way of determining concentrations compares with the more expensive way. To put this question in terms of the model we first note that the assumed model is

$$Y_i \sim G(\alpha + \beta x_i, \sigma) \quad \text{for } i = 1, 2, \ldots, n \text{ independently}$$

If the cheaper way worked perfectly then the measurements using the cheaper way would be identical to the more expensive way plus some variability. That is, the model would be

$$Y_i \sim G(x_i, \sigma) \quad \text{for } i = 1, 2, \ldots, n \text{ independently}$$

This means we are interested in whether the model with $\beta = 1$ and $\alpha = 0$ fits the data well. This is the reason why we test the hypotheses $H_0 : \beta = 1$ and $H_0 : \alpha = 0$.

(c) The scatterplot plus the fitted line indicates good agreement between the cheaper way of determining concentrations and the more expensive way. The points lie quite close to the fitted line. The data suggest that the cheaper way of determining concentrations is quite accurate since the cheaper way does not appear to consistently give values which are systematically above (or below) the concentration determined by the more expensive way.

(d) Since the fitted model is

$$y = -0.1527 + 0.9999x$$

the point estimate of the y-intercept is $\hat{\alpha} = -0.1527$ which is slightly negative which suggests the cheaper way is giving values lower than the true concentration as determined by the more expensive way. However, the confidence interval for α was $[-0.5587, 0.2533]$ which certainly includes the value $\alpha = 0$ as well as values of α above and below zero. The data do not suggest the cheaper way is giving lower values. If the confidence interval only contained negative values then this would be strong evidence that the cheaper way is giving lower values.

6.9 (a) The likelihood function is for β is

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta x_i)^2\right] \quad \text{for } \beta \in \Re$$

or more simply

$$L(\beta) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2\right] \quad \text{for } \beta \in \Re$$

The log likelihood function is

$$l(\beta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2 \quad \text{for } \beta \in \Re$$

Maximizing $l(\beta)$ is equivalent to minimizing $g(\beta) = \sum_{i=1}^{n}(y_i - \beta x_i)^2$ which is the criterion for determining the least squares estimate of β.
Solving

$$l'(\beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)x_i = 0$$

we obtain both the maximum likelihood estimate and the least squares estimate of β given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

(b) Note that

$$\tilde{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} = \sum_{i=1}^{n}\left(\frac{x_i}{\sum_{i=1}^{n} x_i^2}\right) Y_i = \sum_{i=1}^{n} a_i Y_i \quad \text{where } a_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$$

so $\tilde{\beta}$ is a linear combination of independent Normal random variables and therefore has a Normal distribution. Since

$$E(\tilde{\beta}) = \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{i=1}^{n} x_i^2} \right) E(Y_i) = \frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i (\beta x_i) = \frac{\beta}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i^2 = \beta$$

and

$$Var(\tilde{\beta}) = \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{i=1}^{n} x_i^2} \right)^2 Var(Y_i) = \frac{1}{\left[\sum_{i=1}^{n} x_i^2\right]^2} \sum_{i=1}^{n} x_i^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

therefore

$$\tilde{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \right)$$

(c)

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \hat{\beta} x_i)^2 &= \sum_{i=1}^{n} \left(y_i^2 - 2 x_i y_i \hat{\beta} + x_i^2 \hat{\beta}^2 \right) \\
&= \sum_{i=1}^{n} y_i^2 - 2 \underbrace{\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right]}_{\hat{\beta}} \sum_{i=1}^{n} x_i y_i + \underbrace{\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right]^2}_{\hat{\beta}^2} \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} y_i^2 - 2 \frac{\left[\sum_{i=1}^{n} x_i y_i\right]^2}{\sum_{i=1}^{n} x_i^2} + \frac{\left[\sum_{i=1}^{n} x_i y_i\right]^2}{\sum_{i=1}^{n} x_i^2} \\
&= \sum_{i=1}^{n} y_i^2 - \frac{\left[\sum_{i=1}^{n} x_i y_i\right]^2}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

as required.

(d) Find a in the t table such that $P(-a \leq T \leq a) = 0.95$ where $T \sim t(n-1)$. Then since

$$\begin{aligned}
0.95 &= P\left(-a \leq \frac{\tilde{\beta} - \beta}{S_e / \sqrt{\sum_{i=1}^{n} x_i^2}} \leq a \right) \\
&= P\left(\tilde{\beta} - a S_e / \sqrt{\sum_{i=1}^{n} x_i^2} \leq \beta \leq \tilde{\beta} + a S_e / \sqrt{\sum_{i=1}^{n} x_i^2} \right)
\end{aligned}$$

a 95% confidence interval for β is given by

$$\left[\hat{\beta} - as_e/\sqrt{\sum_{i=1}^{n} x_i^2}\ ,\ \hat{\beta} + as_e/\sqrt{\sum_{i=1}^{n} x_i^2}\right]$$

(e) Define the discrepancy measure

$$D = \frac{|\tilde{\beta} - \beta_0|}{S_e/\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Under the null hypothesis $H_0 : \beta = \beta_0$, the $p - value$ is given by

$$P\left(|T| > \frac{|\hat{\beta} - \beta_0|}{s_e/\sqrt{\sum_{i=1}^{n} x_i^2}}\right) = 2\left[1 - P\left(T \leq \frac{|\hat{\beta} - \beta_0|}{s_e/\sqrt{\sum_{i=1}^{n} x_i^2}}\right)\right]$$

where $T \sim t(n-1)$.

6.10 (a)

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{13984.5554}{14058.9097} = 0.9947$$

and the fitted model is $y = 0.9947x$.

(b) The scatterplot with fitted line, the residual plots, and the qqplot of the residuals are given in Figure 6.6. If the model is correct we should see the points in the scatterplot lying about the fitted line with no unusual pattern, the residual plots should look like a band of points about the line $r = 0$, and the qqplot should be a set of points which lie reasonably about a straight line with more variability at each end. These behaviours are what were observe in the graphs below and so the model appears to fit the data well.

(c) Since $P(T \leq 2.0930) = 0.975$ where $T \sim t(19)$ and

$$\begin{aligned} s_e &= \left[\frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2}\right)\right]^{1/2} \\ &= \left[\frac{1}{19}\left(13913.3833 - \frac{13984.5554^2}{14058.9097}\right)\right]^{1/2} = 0.3831 \end{aligned}$$

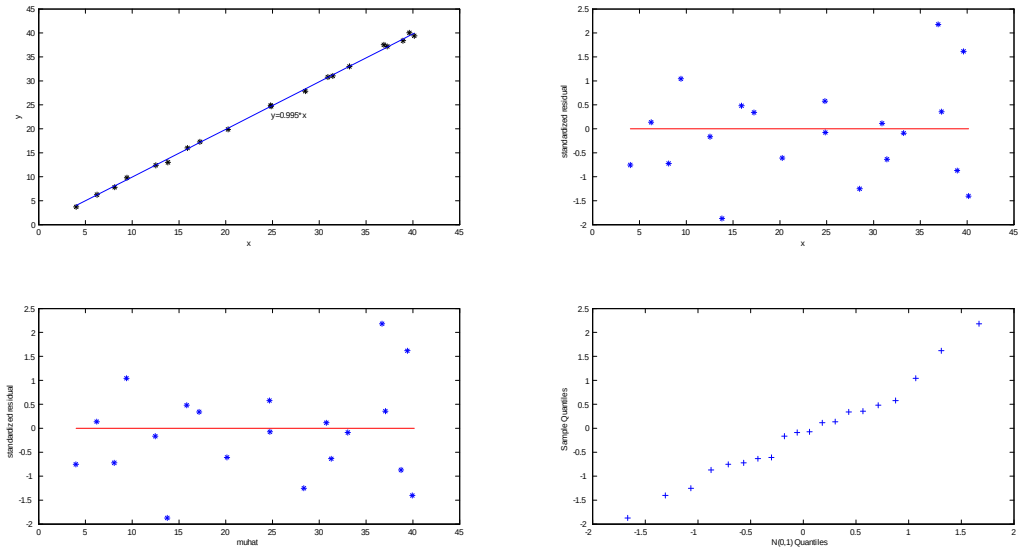


Figure 6.6: Scatterplot and residual plots for model through the origin

a 95% confidence interval for β is given by

$$\hat{\beta} \pm as_e/\sqrt{\sum_{i=1}^{n} x_i^2} = 0.9947 \pm 2.0930(0.3831)/\sqrt{14058.9097} = [0.9879,\ 1.0015]$$

For testing $H_0 : \beta = 1$ we have

$$p - value = 2\left[1 - P\left(T \leq \frac{|0.9947 - 1|}{0.3831/\sqrt{14058.9097}}\right)\right]$$

$$\begin{aligned} &= 2\left[1 - P\left(T \leq 1.640361\right)\right] \quad \text{where } T \sim t\left(19\right) \\ &= 0.1174 \quad \text{calculated using R} \end{aligned}$$

Since $p - value > 0.1$, there is no evidence against $H_0 : \beta = 1$ based on the data.

(d) Based on this analysis we would conclude that the simpler model $Y \sim G(\beta x_i, \sigma)$ is an adequate model for these data as compared to the model $Y_i \sim G(\alpha + \beta x_i, \sigma)$.

6.11 (a) The maximum likelihood estimates of α and β are

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{6175}{2155.2} = 2.8652 \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 30.3869$$

and $y = 30.3869 + 2.8652x$ is the equation of the fitted line.

(b) The scatterplot with fitted line and the residual plots are shown in Figure 6.7. There are a few large negative residuals but overall the model seems reasonable.

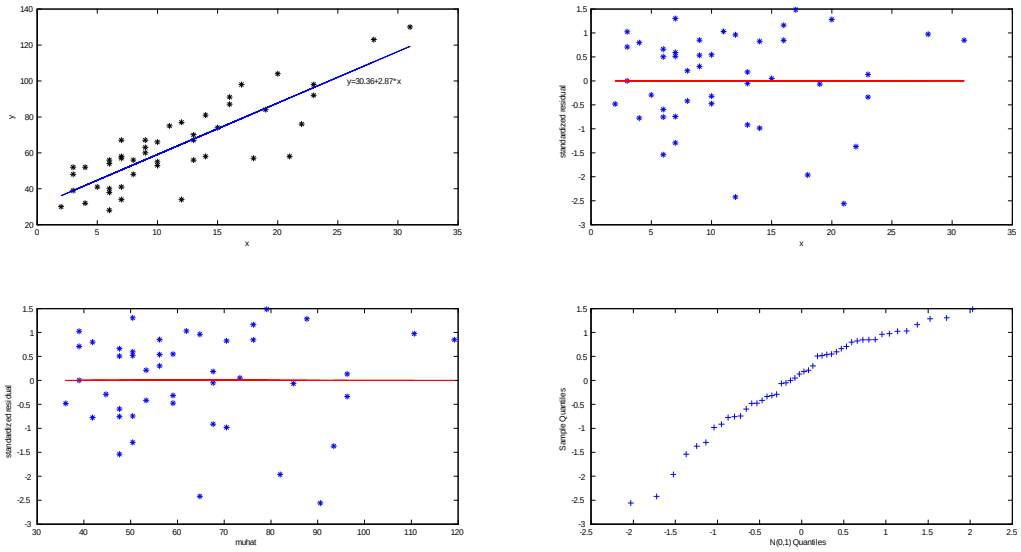


Figure 6.7: Scatterplot and residual plots for death rate due to cirrhosis of the liver versus wine consumption

(c) An estimate of σ is

$$s_e = \left[\frac{S_{yy} - \hat{\beta} S_{xy}}{n-2}\right]^{1/2} = \left[\frac{24801.1521 - (2.8652)(6175.1522)}{44}\right]^{1/2} = 12.7096$$

Since

$$p - value = 2\left[1 - P\left(T \leq \frac{\left|\hat{\beta} - 0\right|}{s_e/\sqrt{S_{xx}}}\right)\right] = 2\left[1 - P\left(T \leq 10.47\right)\right] \approx 0$$

there is very strong evidence based on the data against $H_0 : \beta = 0$. The data suggest that there is a linear relationship between wine consumption per capita and the death rate from cirrhosis of the liver.

(d) Since $P(T \leq 2.0154) = 0.975$ where $T \sim t\,(44)$ a 95% confidence interval for β is

$$2.8652 \pm 2.0154\,(12.7096)\,/\sqrt{2155.1522} = [2.3135, 3.4171]$$

6.12 (a) The command summary(RegModel)$coefficients gives the output:

```
Estimate    Std.  Error    t value     Pr(>|t|)
(Intercept) 1.16113913  0.005428554   213.8947    1.283890e-123
x    -0.06206624 0.003353336   -18.5088    1.954991e-32
```

The fitted line is $y = 1.16113913 - 0.06206624x$ where $y =$ BodyDensity and $x =$ Skinfold.

(b) For the hypothesis $H_0 : \beta = 0$ the value of the test statistic is -18.5088 and the $p - value$ is 1.954991×10^{-32} or approximately 0. Since $p - value \approx 0$ there is very strong evidence based on the data against the hypothesis $H_0 : \beta = 0$. Since the plots in (d) suggest the simple linear regression model is reasonable, therefore the data suggest there is a linear relationship between body density and skinfold measurement.

(c) An estimate of σ is $s_e = 0.007877322$.

(d) The plots are given in Figure 6.8. The scatterplot and residual plots indicate that the model fits the data well.

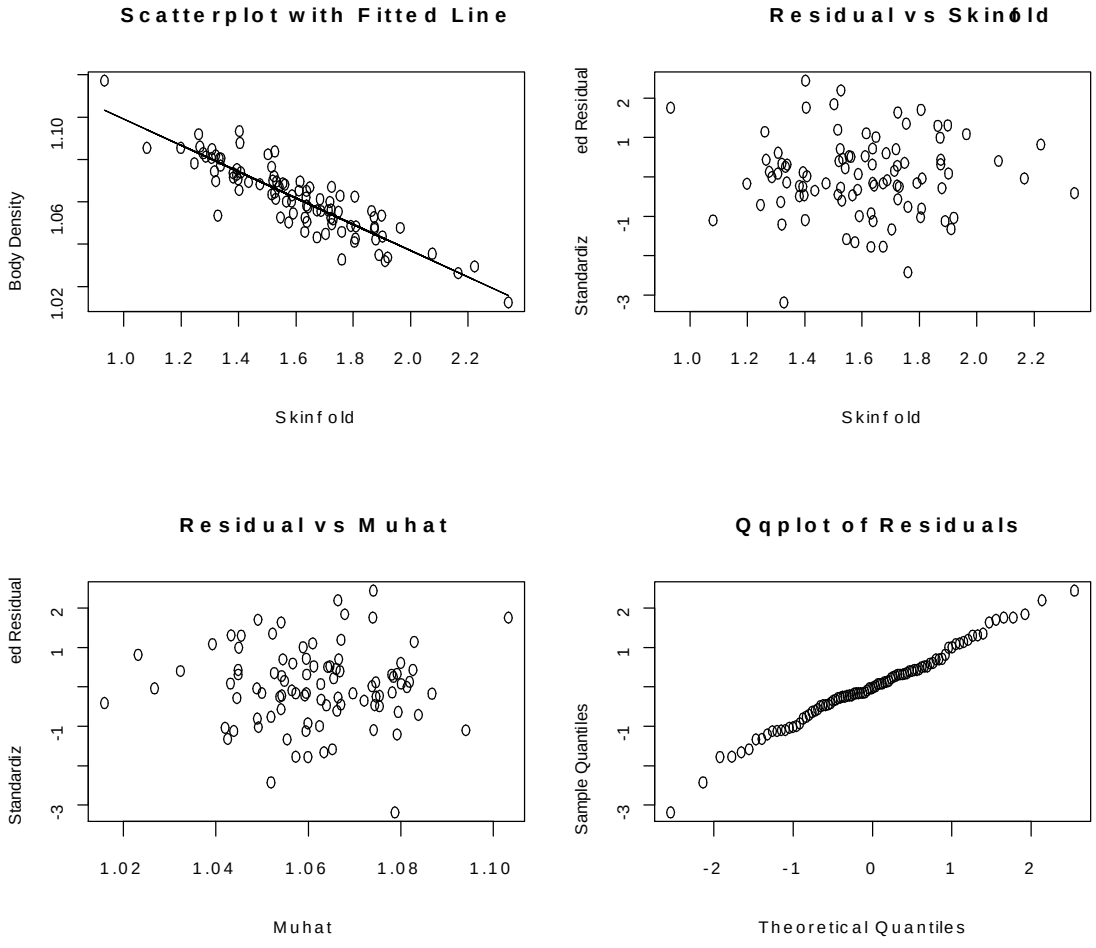


Figure 6.8: Fitted Line and Residual Plots for Skinfold Data

(e) From the R output

```
> # 95% Confidence interval for slope
> confint(RegModel,level=0.95)
2.5 %      97.5 %
(Intercept) 1.15035436   1.17192390
x     -0.06872823  -0.05540425
```

the 95% confidence interval for β is $[-0.06872823, -0.05540425]$.

(f) From the R output

```
> # 90% confidence interval for mean response at x=2
> predict(RegModel,data.frame("x"=2),interval="confidence",level=0.90)
fit      lwr     upr
1 1.037007 1.034394 1.03962
```

the 90% confidence interval for the mean body density for a skinfold measurement of 2 is $[1.034394, 1.03962]$

(g) From the R output

```
> # 99% prediction interval for response at x=1.8
> predict(RegModel,data.frame("x"=1.8),interval="prediction",level=0.99)
fit    lwr      upr
1 1.04942 1.028503 1.070336
```

a 99% prediction interval for the body density of a male with skinfold measurement of $x = 1.8$ is $[1.028503, 1.070336]$.

(h) From the R output

```
> a<-qchisq(0.025,df)
> b<-qchisq(0.975,df)
> int<-c(se*sqrt(df/b),se*sqrt(df/a))
> cat("95% confidence interval for sigma:  ",int)
95% confidence interval for sigma:  0.006875574 0.009223456
```

the 95% confidence interval for σ is $[0.006875574,\ 0.009223456]$

(i) Skinfold measurements seem to provide a reasonable approximation to body density measurements. However we notice that the range of body density measurements is $[1.0126, 1.1171]$ with a width of 0.1045 and that the 99% prediction interval for the body density of a male with skinfold measurement of $x = 1.8$ has width 0.041833 which is approximately one third the width of the range of measurements. There is a fair bit of uncertainty in approximating the body density using the skinfold measurement. The decision to use the approximation or not would depend on issues such as what the body density measurement is to be used for and how accurate it needs to be and how much more cost and effort is required to measure body density directly. Note that an accurate body density measurement is usually done by weighing a person under water.

6.13 (a)

$$\bar{x} = 191.7871 \qquad \bar{y} = 20.0276$$
$$S_{xx} = 2291.3148 \quad S_{yy} = 447.8497 \quad S_{xy} = 1008.8246$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1008.8246}{2291.3148} = 0.44028$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 20.0276 - \left(\frac{1008.8246}{2291.3148}\right)(191.7871) = -64.4128$$

The fitted line is $y = -64.4128 + 0.44028x$. The scatterplot and residual plot are given in the top two panels of Figure 6.9. Both graphs show a distinctive pattern. In the scatterplot as x increases the points lie above the line, then below then above. Correspondingly in the residual plot as x increases the residuals are positive then negative then positive. In the residual plot the points do not lie in a horizontal about the line $\hat{r}_i = 0$ which suggests that the linear model is not adequate.

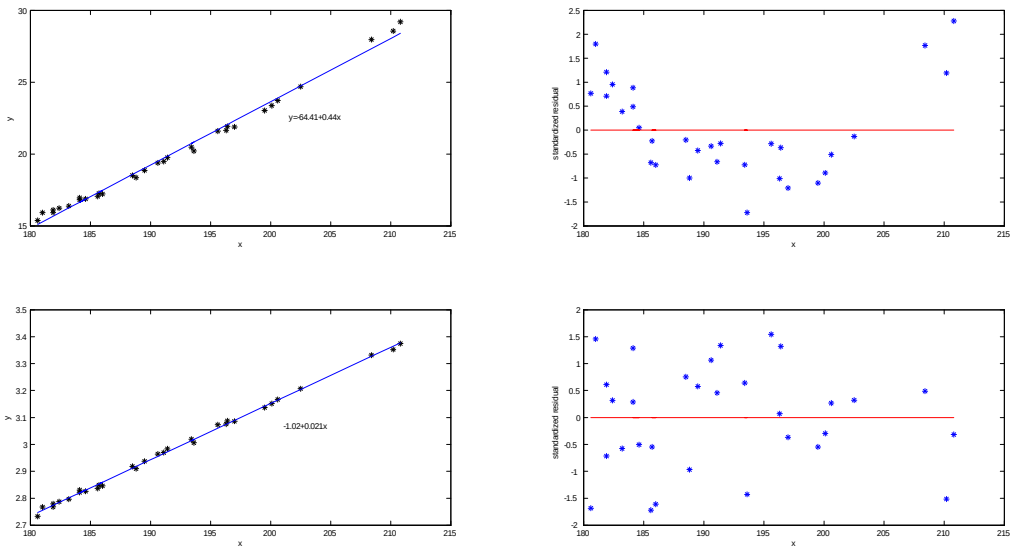


Figure 6.9: Fitted lines and residual plots for atmospheric pressure data

(b)

$$\bar{x} = 191.7871 \qquad \bar{y} = 2.9804$$
$$S_{xx} = 2291.3148 \quad S_{yy} = 1.00001 \quad S_{xy} = 47.81920$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{47.81920}{2291.3148} = 0.02087$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 2.9804 - \left(\frac{47.81920}{2291.3148}\right)(191.7871) = -1.02214$$

The fitted line is $z = -1.02214 + 0.02087x$. The scatterplot and residual plots are given in the bottom two panels of Figure 6.9. In both of these plots we do not observe any unusual patterns. There is no evidence to contradict the linear model for log(pressure) versus temperature.

(c) Based on the scatterplots and residual plots in Figure 6.9 it is clear that the model log(pressure) versus temperature is a much better fit than the model pressure versus temperature. Although the model log(pressure) versus temperature is a good fit to the data this does not "prove" that the theory's model is correct - only that there is no evidence to disprove it.

(d) Since $P(T \leq 2.0452) = 0.975$ where $T \sim t(29)$, and

$$\begin{aligned} s_e &= \left(\frac{S_{yy} - \hat{\beta} S_{xy}}{n-2} \right)^{1/2} \\ &= \left[\frac{1.00001 - (0.02087)(47.81920)}{29} \right]^{1/2} = 0.00838894 \end{aligned}$$

a 95% confidence interval for the mean log atmospheric pressure at a temperature of $x = 195$ is

$$\begin{aligned} & -1.02214 + (0.02087)(195) \pm 2.0452 (0.008389) \left[\frac{1}{31} + \frac{(195 - 191.7871)^2}{2291.3148} \right]^{1/2} \\ = \; & 3.04747 \pm 0.00329 = [3.04418, 3.05076] \end{aligned}$$

which implies a 95% confidence interval for the mean atmospheric pressure at a temperature of $x = 195$ is

$$[\exp(3.04418), \exp(3.05076)] = [20.9927, 21.1313]$$

6.14 (a) We assume that the study population is the set of all Grade 3 students who are being taught the same curriculum. (For example in Ontario all Grade 3 students must be taught the same Grade 3 curriculum set out by the Ontario Government.) The parameter μ_1 represents the mean score on the DRP test if all Grade 3 students in the study population took part in the new directed readings activities for an 8-week period.
The parameter μ_2 represents the mean score on the DRP test for all Grade 3 students in the study population without the directed readings activities.
The parameter σ represents the standard deviation of the DRP scores for all Grade 3 students in the study population which is assumed to be the same whether the students take part in the new directed readings activities or not.

(b) The qqplot of the responses for the treatment group and the qqplot of the responses for the control group are given in Figures 6.10 and 6.11. Looking at these

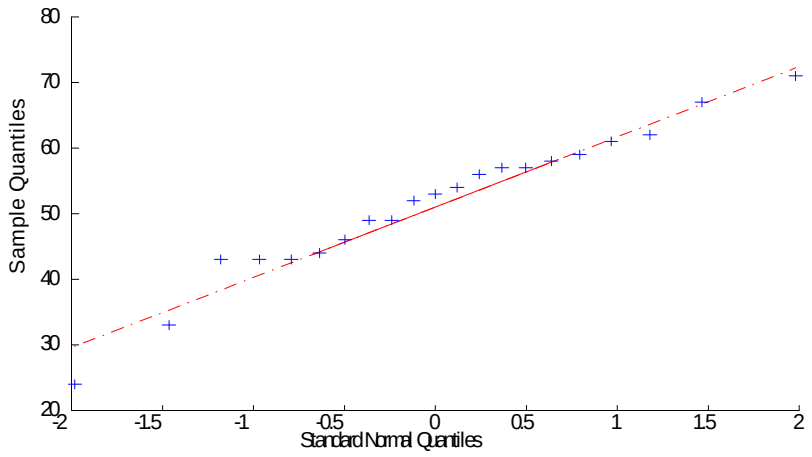


Figure 6.10: Normal Qqplot of the Responses for the Treatment Group

plots we see that the points lie reasonably along a straight line in both plots and so we would conclude that the normality assumptions seem reasonable.

(c) For the given data

$$s_p = \left[\frac{1}{21+23-2}\left(2423.2381+6469.7391\right)\right]^{1/2} = 14.5512$$

Also $P\left(T \leq 2.018\right) = 0.975$ where $T \sim t\left(42\right)$. A 95% confidence interval for the difference in the means, $\mu_1 - \mu_2$ is

$$\begin{aligned} & 51.4762 - 41.5217 \pm \left(2.018\right)\left(14.5512\right)\sqrt{\frac{1}{21}+\frac{1}{23}} \\ = \;& 9.9545 \pm 8.8628 = \left[1.0916,\ 18.8173\right] \end{aligned}$$

(d) To test the hypothesis of no difference between the means, that is, to test the hypothesis $H_0 : \mu_1 = \mu_2$ we use the discrepancy measure

$$D = \frac{\left|\bar{Y}_1 - \bar{Y}_2 - 0\right|}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - 0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t\left(n_1+n_2-2\right)$$

assuming $H_0 : \mu_1 = \mu_2$ is true. The observed value of D for these data is

$$d = \frac{\left|\bar{y}_1 - \bar{y}_2 - 0\right|}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{\left|51.4762 - 41.5217 - 0\right|}{14.5512\sqrt{\frac{1}{21}+\frac{1}{23}}} = 2.2666$$

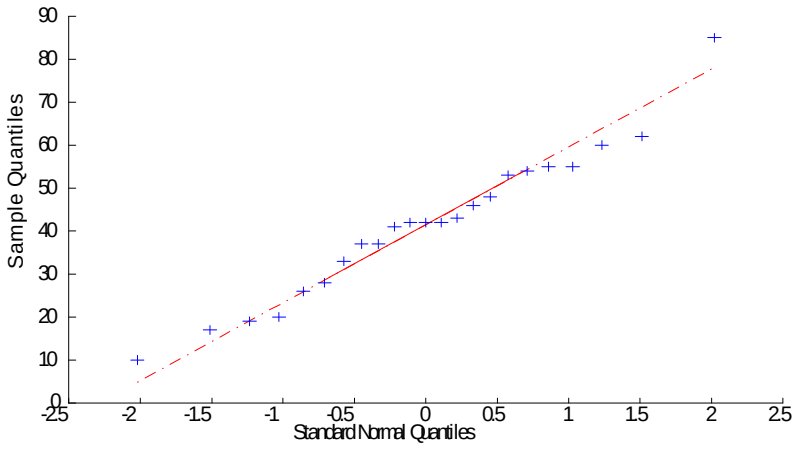


Figure 6.11: Normal Qqplot for the Responses in the Control Group

and

$$\begin{aligned} p-value &= 2\left[1-P\left(T \leq 2.2666\right)\right] \quad \text{where } T \sim t\left(42\right) \\ &= 0.02863 \quad \text{calculated using R} \end{aligned}$$

Since $0.01 < p-value < 0.05$, there is evidence against $H_0 : \mu_1 = \mu_2$ based on the data.

Although the data suggest there is a difference between the treatment group and the control group we **cannot conclude that the difference is due to the the new directed readings activities**. The difference could simply be due to the differences in the two Grade 3 classes. Since randomization was **not** used to determine which student received the treatment and which student was in the control group, the difference in the DRP scores could have existed before the treatment was applied.

(e) Here is the output from running t.test in R

> # t test for hypothesis of no difference in means
> # and 95% confidence interval for mean difference mu
> # note that R uses mu = mu_control - mu_treament
> t.test(DRP~Group,data=treatmentvscontroldata,var.equal=TRUE,conf.level=0.95)

Two Sample t-test

data: DRP by Group
t = -2.2666, df = 42, p-value = 0.02863
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:

-18.817650 -1.091253
sample estimates:
mean in group Control mean in group Treatment
41.52174 51.47619

6.15 (a) The pooled estimate of variance is

$$s_p = \sqrt{\frac{209.02961 + 116.7974}{18}} = 4.25$$

From the t table, $P(T < 1.734) = 0.95$ where $T \sim t\,(18)$. The 90% confidence interval is

$$10.693 - 6.750 \pm 1.734\,(4.25)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = [0.647, 7.239]$$

(b) We test the hypothesis $H_0 : \mu_1 = \mu_2$ or equivalently $H_0 : \mu_1 - \mu_2 = 0$ using the discrepancy measure

$$D = \frac{|\bar{Y}_1 - \bar{Y}_2|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The observed value of this statistic is

$$d = \frac{|10.693 - 6.750|}{4.25\sqrt{\frac{1}{10} + \frac{1}{10}}} = 2.074$$

with

$$\begin{aligned} p - value &= 2\,[1 - P(T \leq 2.074)] \quad \text{where } T \sim t(18) \\ &= 0.0527 \quad \text{calculated using R} \end{aligned}$$

Since $0.05 < p - value < 0.1$, there is weak evidence against $H_0 : \mu_1 - \mu_2 = 0$ based on the data.

(c) We repeat the above using as data $z_{ij} = \log(y_{ij})$. The sample means are 2.248, 1.7950 and the sample variances are 0.320, 0.240 respectively. The pooled estimate of variance is $s_p = \sqrt{\frac{0.320+0.240}{2}} = 0.529$. The observed value of the discrepancy measure is

$$d = \frac{|2.248 - 1.795 - 0|}{0.529\sqrt{\frac{1}{10} + \frac{1}{10}}} = 1.9148$$

with

$$\begin{aligned} p - value &= 2\,[1 - P(T \leq 1.9148)] \quad \text{where } T \sim t(18) \\ &= 0.0715 \quad \text{calculated using R} \end{aligned}$$

Since $0.05 < p - value < 0.1$, there is evidence against $H_0 : \mu_1 - \mu_2 = 0$ based on the data.

(d) One could check the Normality assumption with qqplots for each of the variates y_{ij} and $z_{ij} = \log(y_{ij})$ although with such a small sample size these will be difficult to interpret.

6.16 (a) The pooled estimate of the common standard deviation σ is

$$s_p = \sqrt{\frac{3050 + 2937}{58}} = 10.1599$$

Using R, $P(T \leq 2.0017) = 0.975$ where $T \sim t\,(58)$. The 95% confidence interval for $\mu_1 - \mu_2$ is

$$120 - 114 \pm 2.0017\,(10.1599)\sqrt{\frac{1}{30} + \frac{1}{30}} = 6 \pm 5.2511 = [0.7489, 11.2511]$$

(b) Since

$$d = \frac{|120 - 114 - 0|}{10.1599\sqrt{\frac{1}{30} + \frac{1}{30}}} = 2.2872$$

with

$$\begin{aligned} p - value &= 2\,[1 - P(T \leq 2.2872)] \quad \text{where } T \sim t(58) \\ &= 0.0259 \quad \text{calculated using R} \end{aligned}$$

Since $0.01 < p - value < 0.05$, there is evidence against the hypothesis of no difference based on the data. This is consistent with the fact that the 95% confidence interval for $\mu_1 - \mu_2$ did not contain the value $\mu_1 - \mu_2 = 0$.

6.17 Let μ_1 be the mean log failure time for welded girders and μ_2 be the mean score for log failure time for repaired welded girders. The pooled estimate of the common standard deviation σ is

$$s_p = \sqrt{\frac{13\,(0.0914) + 9\,(0.0422)}{22}} = 0.26697$$

From the t table, $P(T < 2.0739) = 0.975$ where $T \sim t\,(22)$. The 95% confidence interval for $\mu_1 - \mu_2$ is

$$14.564 - 14.291 \pm 2.0739\,(0.26697)\sqrt{\frac{1}{14} + \frac{1}{10}} = 0.273 \pm 0.22924 = [0.04376, 0.50224]$$

Since

$$d = \frac{|14.564 - 14.291 - 0|}{0.26697\sqrt{\frac{1}{14} + \frac{1}{10}}} = 2.4698$$

with

$$\begin{aligned} p - value &= 2\,[1 - P(T \leq 2.4698)] \quad \text{where } T \sim t(22) \\ &= 0.02175 \quad \text{calculated using R} \end{aligned}$$

Since $0.01 < p-value < 0.05$, there is evidence against the hypothesis of no difference based on the data. This is consistent with the fact that the 95% confidence interval for $\mu_1 - \mu_2$ did not contain the value $\mu_1 - \mu_2 = 0$.

6.18 (a) For the female coyotes we have

$$\bar{y}_f = 89.24, \quad s_f^2 = 42.87887, \quad n_f = 40$$

For the male coyotes we have

$$\bar{y}_m = 92.06, \quad s_m^2 = 44.83586, \quad n_m = 43$$

Since $n_f = 40$ and $n_m = 43$ are reasonably large we have that $\bar{Y}_f$ has approximately a $N(89.24, 42.87887/40)$ distribution and $\bar{Y}_m$ has approximately a $N(92.06, 44.83586/43)$ distribution. An approximate 95% confidence interval for $\mu_f - \mu_m$ is given by

$$89.24 - 92.06 \pm 1.96\sqrt{\frac{42.87887}{40} + \frac{44.83586}{43}} = [-5.67, 0.03]$$

The value $\mu_f - \mu_m = 0$ is just inside the right hand endpoint and the $p-value$ for testing $H_0 : \mu_f - \mu_m = 0$ would be close to 0.05 so there is weak evidence based on the data of a difference between mean length for male and female coyotes. Since the interval contains mostly negative values the data suggest the mean length for males is slightly larger than for females.

(b) Using $Y_1 \sim N(89.24, 42.87887)$, $Y_2 \sim N(92.06, 44.83586)$ and $Y_1 - Y_2 \sim N(89.24 - 92.06, 42.87887 + 44.83586)$ or $Y_1 - Y_2 \sim N(-2.82, 87.71473)$ we estimate $P(Y_1 > Y_2) = P(Y_1 - Y_2 > 0)$ as

$$P\left(Z > \frac{0 - (-2.82)}{\sqrt{87.71473}}\right) = 1 - P(Z \leq 0.30) = 1 - 0.61791 = 0.38209$$

(c) Since $P(T \leq 2.0227) = 0.975$ where $T \sim t(39)$ a 95% confidence interval the mean length of female coyotes is

$$89.24 \pm 2.0227\sqrt{42.87887/40} = 89.24 \pm 2.0942 = [87.1457, 91.3342]$$

Since $P(T \leq 2.0181) = 0.975$ where $T \sim t(42)$ a 95% confidence interval the mean length of male coyotes is

$$92.06 \pm 2.0181\sqrt{44.83586/43} = 92.06 \pm 2.06073 = [89.9993, 94.1207]$$

6.19 The pooled estimate of the common standard deviation is

$$s_p = \sqrt{\frac{0.608 + 0.35569}{22}} = 0.2093$$

Since $P(T \leq 2.0739) = 0.975$ where $T \sim t\,(22)$, a 95% confidence interval for $\mu_1 - \mu_2$ is

$$1.370 - 1.599 \pm 2.0739\,(0.2093)\sqrt{\frac{1}{12} + \frac{1}{12}} = [-0.4064, -0.0520]$$

This interval does not contain $\mu_1 - \mu_2 = 0$ and only contains negative values. The data suggest that $\mu_1 < \mu_2$, that is, the mean reaction time for the "Alcohol" group is less than the mean reaction time for the "Non-Alcohol" group. We are not told the units of these reaction times so it is unclear whether this difference is of practical significance.

6.20 (a) We assume that the observed differences are a random sample from a $G\,(\mu, \sigma)$ distribution. An estimate of σ is

$$s = \sqrt{\frac{17.135}{7}} = 1.5646$$

Since $P(T \leq 2.3646) = 0.975$ where $T \sim t\,(7)$, a 95% confidence interval for μ is

$$1.075 \pm 2.3646\,(1.5646)\,/\sqrt{8} = 1.075 \pm 1.3080 = [-0.2330, 2.3830]$$

(b) If the natural pairing is ignored an estimate of the common standard deviation is

$$s_p = \sqrt{\frac{535.16875 + 644.83875}{14}} = 9.18075$$

Since $P(T \leq 2.1448) = 0.975$ where $T \sim t\,(14)$, a 95% confidence interval for $\mu_1 - \mu_2$ is

$$23.6125 - 22.5375 \pm 2.1448\,(9.18075)\sqrt{\frac{1}{8} + \frac{1}{8}} = [-8.7704, 10.9204]$$

We notice that although both intervals in (a) and (b) are centered at the value 1.075, the interval in (b) is very much wider.

(c) A matched pairs study allows for a more precise comparison since differences between the 8 pairs have been eliminated. That is by analyzing the differences we do not need to worry that there may have been large differences in the 8 cars which were used in the study with respect to other explanatory variates which might affect gas mileage (the response variate) such as size of engine, make of car, etc.

6.21 (a) We assume that the study population is the set of all factories of similar size. The parameter μ represents the mean difference in the number of staff hours per month lost due to accidents before and after the introduction of an industrial safety program in the study population.

(b) For these data

$$s = \left[\frac{1}{7}\left(1148.79875\right)\right]^{1/2} = 12.8107$$

From the t table $P\left(T \leq 2.3646\right) = 0.975$ where $T \sim t\left(7\right)$. A 95% confidence interval for μ is

$$-15.3375 \pm 2.364624\left(12.8107\right)/\sqrt{8} = -15.3375 \pm 10.71002 = [-26.04752, -4.627484]$$

(c) The observed discrepancy measure is

$$d = \frac{|\bar{y} - 0|}{s/\sqrt{n}} = \frac{|-15.3375 - 0|}{12.8107/\sqrt{8}} = 3.386309$$

and

$$\begin{aligned} p - value &= 2\left[1 - P\left(T \leq 3.386309\right)\right] \quad \text{where } T \sim t\left(7\right) \\ &= 0.01166 \quad \text{calculated using R} \end{aligned}$$

Since the $0.01 < p - value < 0.05$ there is evidence against the hypothesis $H_0 : \mu = 0$ based on the data.
Since this experimental study was conducted as a matched pairs study, an analysis of the differences, $y_i = y_{1ii} - y_{2i}$, allows for a more precise comparison since differences between the 8 pairs have been eliminated. That is by analyzing the differences we do not need to worry that there may have been large differences in the safety records between factories due to other variates such as differences in the management at the different factories, differences in the type of work being conducted at the factories etc. Note however that a drawback to the study was that we were not told how the 8 factories were selected. To do the analysis above we have assumed that the 8 factories are a random sample from the study population of all similar size factories but we do not know if this is the case.

6.23 (a) Since two algorithms are each run on the same 20 sets of numbers we analyse the differences $y_i = y_{Ai} - y_{Bi}$, $i = 1, 2, \ldots, 20$. Since $P(T < 2.8609) = 0.995$ where $T \sim t\left(19\right)$, we obtain the confidence interval

$$0.409 \pm 2.8609\left(0.487322\right)/\sqrt{20} = [0.097, 0.721]$$

These values are all positive indicating strong evidence based on the data against $H_0 : \mu_A - \mu_B = 0$ $(p - value < 0.01)$, that is, the data suggest that algorithm B is faster.

(b) To check the Normality assumption we plot a qqplot of the differences. See Figure 6.12. The data lie reasonably along a straight line and therefore a Normal model is reasonable.

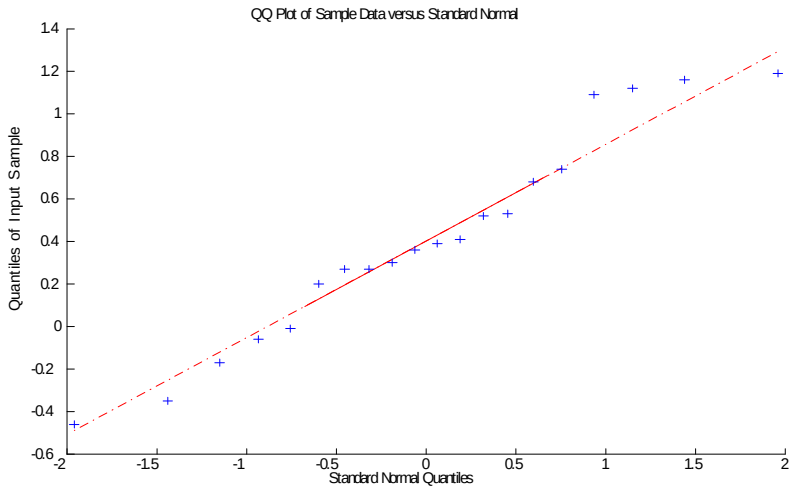


Figure 6.12: Qqplot for sorting algorithm data.

(c) We can estimate the probability by using the fact that $Y_A - Y_B \sim G(\mu, \sigma)$. We estimate the parameters using $\hat{\mu} = 0.40$ and $s = 0.487322$. Since

$$\begin{aligned} P(Y_A > Y_B) &= P(Y_A - Y_B > 0) = P\left(Z > \frac{0 - 0.409}{0.487322}\right) \\ &= P\left(Z > -0.84\right) = P\left(Z < 0.84\right) = 0.80 \quad \text{where } Z \sim N\left(0, 1\right) \end{aligned}$$

an estimate of the probability that algorithm B sorts a randomly selected list faster than A is 0.80.

(d) An estimate of p is $\hat{p} = 15/20 = 0.75$ and an approximate 95% is given by

$$\begin{aligned} \hat{p} \pm 1.96\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} &= 0.75 \pm 1.96\sqrt{\frac{0.75\left(0.25\right)}{20}} \\ &\text{or} \quad [0.56, 0.94] \end{aligned}$$

(e)

$$s_p = \sqrt{\frac{1.4697 + 0.9945}{2}} = 1.1100$$

Using R we have $P(T < 2.7116) = (1 + 0.99)/2 = 0.995$ where $T \sim t(38)$. The interval, assuming common variance, is

$$\bar{y}_1 - \bar{y}_2 \pm a s_p \sqrt{\frac{1}{20} + \frac{1}{20}} = 0.409 \pm 2.7116(1.1100)\sqrt{\frac{1}{20} + \frac{1}{20}}$$

or

$$[-0.543, 1.361]$$

This second interval $[-0.543, 1.361]$ is much wider than the first interval $[0.097, 0.721]$ biased on the paired experiment and unlike the first interval, it contains the value

zero. Unlike the paired design, independent samples of the same size (20 different problems run with each algorithm) is too small to demonstrate the superiority of algorithm B. The independent samples is a less efficient way to analyse the difference. This is why in computer simulations, it is essential to be able to run different simulations using the same random number seed.

(f) Here is the R output for doing the t tests and confidence intervals for the paired analysis and the unpaired analysis:
> t.test(Time~Algorithm,data=sortingdata,paired=TRUE,conf.level=0.99)

Paired t-test

data: Time by Algorithm
t = 3.7534, df = 19, p-value = 0.001346
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
0.09724793 0.72075207
sample estimates:
mean of the differences
0.409

t.test(Time~Algorithm,data=sortingdata,paired=F,var.equal=T,conf.level=0.99)

Two Sample t-test

data: Time by Algorithm
t = 1.1652, df = 38, p-value = 0.2512
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
-0.5427918 1.3607918
sample estimates:
mean in group A mean in group B
4.7375 4.3285

SOLUTIONS TO CHAPTER 7 PROBLEMS

7.1 Here is the R output

```
> y<-c(556,678,739,653,725,714,566,797) # observed frequencies
> e<-sum(y)/8 # expected frequencies
> lambda<-2*sum(y*log(y/e)) # observed value of LR statistic
> lambda
[1] 74.10284
> df<-7 # degrees for freedom for this example equal 7
> 1-pchisq(lambda,df) # p-value for LR test
[1] 2.181588e-13
> d<-sum((y-e)^2/e) # observed value of Pearson goodness of fit statistic
> d
[1] 72.86367
> 1-pchisq(d,df) # p-value for Pearson goodness of fit test
[1] 3.890221e-13
```

In both cases there is very strong evidence against the hypothesis that the distribution of colours is uniform.

7.2 From Table 2.3 in the Course Notes we have the observed frequencies and the expected frequencies calculated using the Poisson model with the mean θ estimated by the sample mean $\hat{\theta} = 3.8715$. In order to use the χ^2 approximation the last four classes have been combined so that the expected frequency in all classes is at least 5. The observed value of the likelihood ratio statistic is

$$2\sum_{j=1}^{12} f_j \log\left(\frac{f_j}{e_j}\right) = 2\left[57\log\left(\frac{57}{54.31}\right) + 203\log\left(\frac{203}{210.28}\right) + \cdots + 6\log\left(\frac{6}{5.80}\right)\right] = 14.01$$

(remember log = ln). Note that this value has been calculated using the expected frequencies calculated in R and not the rounded frequencies displayed in the table.

There are 12 rows in the table so $k = 12$. There is only one unknown parameter θ to be estimated under the hypothesized Poisson(θ) model so $p = 1$. The degrees of freedom for the Chi-squared approximation equal $k - 1 - p = 12 - 1 - 1 = 10$.

Number of α-particles detected: j	Observed Frequency: f_j	Expected Frequency: e_j
0	57	54.31
1	203	210.28
2	383	407.06
3	525	525.31
4	532	508.44
5	408	393.69
6	273	254.03
7	139	140.50
8	45	67.99
9	27	29.25
10	10	11.32
≥ 11	6	5.80
Total	2608	2607.98

Since

$$\begin{aligned} p-value &= P(\Lambda \geq 14.01; H_0) \approx P(W \geq 14.01) \quad \text{where } W \sim \chi^2(10) \\ &= 0.1725 \quad \text{calculated using R} \\ &> 0.1 \end{aligned}$$

there is no evidence against the Poisson model based on the observed data.
The observed value of the goodness of fit statistic is

$$\sum_{j=1}^{12} \frac{(f_j - e_j)^2}{e_j} = \frac{(57-54.3)^2}{54.31} + \frac{(203-210.3)^2}{210.28} + \cdots + \frac{(6-5.80)^2}{5.80} = 12.96$$

and

$$\begin{aligned} p-value &= P(\Lambda \geq 12.96; H_0) \\ &\approx P(W \geq 12.96) \quad \text{where } W \sim \chi^2(10) \\ &= 0.2259 > 0.1 \quad \text{calculated using R} \end{aligned}$$

so again there is no evidence against the Poisson model based on the observed data.
Here is R code to do this analysis using the likelihood ratio statistic:

```
th<-10097/2608  # estimate of theta = mean interruption time
# observed frequencies for collapsed table
f<-c(57,203,383,525,532,408,273,139,45,27,10,6)
# expected frequencies based on Poisson model
e<-2608*c(dpois(0:10,th),1-ppois(10,th))
lambda<-2*sum(f*log(f/e))    # observed value of LR statistic
pvalue<-1-pchisq(lambda,3)   # p-value for LR test
c(lambda,pvalue)
```

7.3 The table below contains the observed and expected frequencies for the data for Wayne Gretzky from Chapter 2, Problem 10. The categories 7 and ≥ 8 have been combined to obtain an expected frequency of at least five.

Number of Points in a Game: y	Observed Number of Games with y points: f_y	Expected Number of Games with y points: e_y
0	69	63.27
1	155	151.71
2	171	181.90
3	143	145.40
4	79	87.17
5	57	41.81
6	14	16.71
≥ 7	8	8.04
Total	696	696

The observed value of the likelihood ratio statistic is $\lambda = 7.491$ and the approximate $p-value$ is $P(W > 7.491) = 0.2778$ where $W \sim \chi^2(6)$, so there is no evidence against the Poisson model based on the observed data.

7.4 The table below contains the observed and expected frequencies for the data for Sidney Crosby from Chapter 2, Problem 11. The categories 5 and ≥ 6 have been combined to obtain an expected frequency of at least five.

Number of Points in a Game: y	Observed Number of Games with y points: f_y	Expected Number of Games with y points: e_y
0	219	210.93
1	259	276.66
2	185	181.43
3	90	79.32
4	24	26.01
≥ 5	6	8.65
Total	783	783

The observed value of the likelihood ratio statistic is $\lambda = 3.971$ and the approximate $p-value$ is $P(W > 3.971) = 0.4099$ where $W \sim \chi^2(4)$, so based on the data there is no evidence against the Poisson model.

7.5 (a) The total number of defectives among the $250 \times 12 = 3000$ items inspected is

$$80 \times 1 + 31 \times 2 + 19 \times 3 + 11 \times 4 + 5 \times 5 + 1 \times 6 = 274$$

and the maximum likelihood estimate of θ = the proportion of defectives is

$$\hat{\theta} = \frac{274}{3000} = 0.09133$$

We want to test the hypothesis that the number of defectives in a box is Binomial$(12, \theta)$. Under this hypothesis and using $\hat{\theta} = 0.091333$ we obtain the expected numbers in each category

Number of defective	0	1	2	3	4	5	≥ 6	Total
e_i	79.21	95.54	52.82	17.70	4	0.64	0.08	250

where

$$e_i = 250\binom{12}{i}\hat{\theta}^i(1-\hat{\theta})^{12-i} \quad \text{for } i = 0, 1, \ldots, 5$$

and the last category is obtained by subtraction. Since the expected numbers in the last three categories are all less than 5 we pool these categories to improve the Chi-squared approximation and obtain

Number of defective	0	1	2	3	≥ 4	Total
f_i (e_i)	103(79.21)	80(95.54)	31(52.82)	19(17.7)	17(4.72)	250

The observed value of the likelihood ratio statistic is

$$2\left[103\log\left(\frac{103}{79.21}\right) + 80\log\left(\frac{80}{95.54}\right) + \cdots + 17\log\left(\frac{17}{4.72}\right)\right] = 38.8552$$

(Remember log = ln.) Under the null hypothesis we had to estimate the parameter θ. The degrees of freedom are $4 - 1 = 3$. The approximate $p - value$ is $P(W > 38.8552) \approx 0$ where $W \sim \chi^2(3)$, so based on the data there is very strong evidence that the Binomial model does not fit.

(b) The likely reason that the Binomial model does not fit well is that defects usually occur in batches which would result in more cartons with no defects than one would expect.

7.6 (a) Here is the R code and output for this problem:

```
> y<-c(70,75,63,59,81,92,75,100,63,58) # observed frequencies
> e<-sum(y)/10   # expected frequencies
> df<-9    # degrees of freedom = 10-1 = 9
> # Likelihood Ratio Goodness of Fit Test
> lambda<-2*sum(y*log(y/e))
> pvalue<-1-pchisq(lambda,df)
> c(lambda,pvalue)
[1] 23.604947153 0.004971575
> # Pearson goodness of fit statistic
> d<-sum((y-e)^2/e)
> pvalue<-1-pchisq(d,df)
> c(d,pvalue)
[1] 24.298913043 0.003852929
```

Since the $p - value < 0.01$ there is strong evidence based on the data against the hypothesis that the machine is operating in a truly "random" fashion.

(b) Let Λ_i be the likelihood ratio statistic for testing $H_0 : \theta_j = 0.1$, $j = 1, 2, \ldots, 9$ for position i, $i = 1, 2, \ldots, 6$. To test the hypothesis $H_0 : \theta_j = 0.1$, $j = 1, 2, \ldots, 9$ for all positions we could use the test statistic $D = \max_{1 \le i \le 6} \Lambda_i$ (Why does this make sense?). For these data we have from (a) that the observed value of D is $d = 23.60494715$. Therefore

$$\begin{aligned}
p - value &= P\left(D \ge 23.605\right) \\
&= P\left(\max_{1 \le i \le 6} \Lambda_i \ge 23.605\right) \\
&= 1 - P\left(\max_{1 \le i \le 6} \Lambda_i \le 23.605\right) \\
&= 1 - P\left(\Lambda_1 \le 23.605, \Lambda_2 \le 23.605, \ldots, \Lambda_6 \le 23.605\right) \\
&= 1 - P\left(\Lambda_1 \le 23.605\right) P\left(\Lambda_2 \le 23.605\right) \cdots P\left(\Lambda_6 \le 23.605\right) \\
&\quad \text{since } \Lambda_1, \Lambda_2, \ldots, \Lambda_6 \text{ are independent random variables} \\
&\approx 1 - \left(1 - 0.004971575\right)^6 \\
&= 0.02946115
\end{aligned}$$

Since $p - value \approx 0.029 < 0.05$ there is evidence based on the data against the hypothesis that all six machines are operating in a truly random fashion.

7.7 (a) This process can be thought of as an experiment in which we observe y_i = the number of non-zero digits (Failures) until the first zero (Success) for $i = 1, 2, \ldots, 50$ and $P\left(\text{Success}\right) = 0.1$. Therefore the Geometric(0.1) distribution is an appropriate model for these data.

(b) The data in a frequency table are:

# between 2 zeros	0	1	2	3	4	5	6	7	8	10	12
# of occurrences	6	4	9	3	5	2	2	3	2	2	1
# between 2 zeros	13	14	15	16	18	19	20	21	22	26	
# of occurrences	1	1	1	1	1	1	1	1	2	1	

The expected frequencies are

$$e_j = 50\,(0.1)\,(1-0.1)^j\,, \quad j = 0, 1, \ldots.$$

To obtain expected frequencies of at least five we join adjacent categories to obtain:

Observation between two 0's	0	1	$2-3$	$4-5$	$6-7$	$8-10$	≥ 11	Total
Observed Frequency.: f_j	6	4	12	7	5	4	12	50
Expected Frequency.: e_i	5	4.5	7.695	6.233	5.049	5.833	15.691	50

The observed value of the likelihood ratio statistic is $\lambda = 3.984$. The degrees of freedom for the Chi-squared approximation are $7-1=6$. Since

$$\begin{aligned} p-value &\approx P\,(W \geq 3.984) \quad \text{where} \;\; W \sim \chi^2\,(6) \\ &\approx 0.68 > 0.1 \end{aligned}$$

there is no evidence based on the data against the hypothesis that the Geometric(0.1) distribution is a good model for these data.

7.8 (a) For $n=2$, the likelihood function is

$$L_2\,(\theta_2) = \left[\binom{2}{0}(1-\theta_2)^2\right]^{23}\left[\binom{2}{1}\theta_2\,(1-\theta_2)\right]^{44}\left[\binom{2}{2}\theta_2^2\right]^{13} \quad \text{for } 0<\theta_2<1$$

or more simply

$$L_2\,(\theta_2) = (1-\theta_2)^{2(23)}\,\theta_2^{44}\,(1-\theta_2)^{44}\,\theta_2^{2(13)} = \theta_2^{70}\,(1-\theta_2)^{90} \quad \text{for } 0<\theta_2<1$$

which is maximized for

$$\hat{\theta}_2 = \frac{70}{160} = 0.4375$$

For $n=3$

$$\begin{aligned} L_3\,(\theta_3) &= (1-\theta_3)^{3(10)}\,\theta_3^{25}\,(1-\theta_3)^{2(25)}\,\theta_3^{2(48)}\,(1-\theta_3)^{1(48)}\,\theta_3^{3(13)} \\ &= \theta_3^{160}\,(1-\theta_3)^{128} \quad \text{for } 0<\theta_3<1 \end{aligned}$$

which is maximized for

$$\hat{\theta}_3 = \frac{160}{288} = 0.5556$$

For $n = 4$

$$\begin{aligned} L_4(\theta_4) &= (1-\theta_4)^{4(5)} \theta_4^{30} (1-\theta_4)^{3(30)} \theta_4^{2(34)} (1-\theta_3)^{2(34)} \\ &\quad \times \theta_4^{3(22)} (1-\theta_4)^{1(22)} \theta_4^{4(5)} \\ &= \theta_4^{184} (1-\theta_4)^{200} \quad \text{for } 0 < \theta_4 < 1 \end{aligned}$$

which is maximized for

$$\hat{\theta}_4 = \frac{184}{384} = 0.4792$$

The expected frequencies assuming the Binomial model, are calculated using

$$e_{nj} = y_{n+} \binom{n}{j} \hat{\theta}_n^j \left(1 - \hat{\theta}_n\right)^{n-j} \quad \text{for } j = 0, 1, \ldots, n; \quad n = 2, 3, 4$$

and are given below:

		Number of females $= j$					Total number of litters
	e_{nj}	0	1	2	3	4	y_{n+}
Litter	2	25.3125	39.375	15.3125			80
Size $= n$	3	8.4280	31.6049	39.5062	16.4609		96
	4	7.0643	25.9964	35.8751	22.0034	5.0608	96

For $n = 2$ the observed value of the likelihood ratio statistic is

$$2\left[23 \log\left(\frac{23}{25.3125}\right) + 44 \log\left(\frac{44}{39.375}\right) + 13 \log\left(\frac{13}{15.3125}\right)\right] = 1.11$$

The degrees of freedom are $3 - 1 - 1 = 1$ since θ_2 was estimated. Since

$$\begin{aligned} p - value &\approx P(W \geq 1.11) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{1.11}\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2\left[1 - P\left(Z \leq 1.05\right)\right] \\ &= 0.29220 > 0.1 \end{aligned}$$

there is no evidence based on the data against the Binomial model. Similarly for $n = 3$, we obtain $\lambda = 4.22$ and $P(W \geq 4.22) = 0.12$ where $W \sim \chi^2(2)$ and there is no evidence based on the data against the Binomial model. For $n = 4$, $\lambda = 1.36$ and $P(W \geq 1.36) = 0.71$ where $W \sim \chi^2(3)$ and there is also no evidence based on the data against the Binomial model.

(b) The likelihood function for $\theta_1, \theta_2, \theta_3, \theta_4$ is

$$\begin{aligned} L(\theta_1, \theta_2, \theta_3, \theta_4) &= \theta_1^{12}(1-\theta_1)^8 \theta_2^{70}(1-\theta_2)^{90} \theta_3^{160}(1-\theta_3)^{128} \theta_4^{184}(1-\theta_4)^{200} \\ \text{for } 0 &< \theta_n < 1; \quad n = 1, 2, 3, 4 \end{aligned}$$

Under the hypothesis $H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta$ the likelihood function is

$$\begin{aligned} L(\theta) &= \theta^{12}(1-\theta)^8 \theta^{70}(1-\theta)^{90} \theta^{160}(1-\theta)^{128} \theta^{184}(1-\theta)^{200} \\ &= \theta^{12+70+160+184}(1-\theta)^{8+90+128+200} \\ &= \theta^{426}(1-\theta)^{426} \quad \text{for } 0 < \theta < 1 \end{aligned}$$

which is maximized for $\hat{\theta} = \frac{426}{852} = 0.5$. The expected frequencies, assuming H_0 are calculated using

$$e_{nj} = y_{n+}\binom{n}{j}(0.5)^n \quad \text{for } j = 0, 1, \ldots, n; \quad n = 2, 3, 4$$

and are given below:

		Number of females $= j$					Total number
	e_{nj}	0	1	2	3	4	of litters $= y_{n+}$
Litter	1	10	10				20
Size $= n$	2	20	40	20			80
	3	12	36	36	12		96
	4	6	24	36	24	6	96

The observed value of the likelihood ratio statistic is

$$2\left[8\log\left(\frac{8}{10}\right) + 12\log\left(\frac{12}{10}\right) + \cdots + 22\log\left(\frac{22}{24}\right) + 5\log\left(\frac{5}{6}\right)\right] = 14.27$$

The degrees of freedom for the Chi-squared approximation equal $[(2-1)+(3-1)+(4-1)+(5-1)]-1 = 9$. Since

$$\begin{aligned} p-value &\approx P(W \geq 14.27) \quad \text{where } W \sim \chi^2(9) \\ &= 0.113 \quad \text{calculated using R} \\ &> 0.1 \end{aligned}$$

there is no evidence based on the data against the hypothesis $\theta_1 = \theta_2 = \theta_3 = \theta_4$.

7.9 The observed frequencies are:

y_{ij}	Tall wife	Medium wife	Short wife	Total
Tall husband	18	28	19	65
Medium husband	20	51	28	99
Short husband	12	25	9	46
Total	50	104	56	210

The expected frequencies are:

e_{ij}	Tall wife	Medium wife	Short wife	Total
Tall husband	$\frac{65\times50}{210} = 15.476$	$\frac{65\times104}{210} = 32.191$	17.333	65
Medium husband	$\frac{99\times50}{210} = 23.571$	$\frac{99\times50}{210} = 49.029$	26.400	99
Short husband	10.952	22.781	12.267	46
Total	50	104	56	210

The observed value of the likelihood ratio statistic is

$$\begin{aligned}\lambda &= 2[18\log\left(\frac{18}{15.476}\right) + 28\log\left(\frac{28}{32.191}\right) + 19\log\left(\frac{19}{17.333}\right) \\ &\quad +20\log\left(\frac{20}{23.571}\right) + 51\log\left(\frac{51}{49.029}\right) + 28\log\left(\frac{28}{26.400}\right) \\ &\quad +12\log\left(\frac{12}{10.952}\right) + 25\log\left(\frac{25}{22.781}\right) + 9\log\left(\frac{9}{12.267}\right)] \\ &= 3.1272\end{aligned}$$

The degrees of freedom for the Chi-squared approximation are $(3-1)(3-1) = 4$. Since

$$\begin{aligned}p - value &\approx P(W \geq 3.1272) \quad \text{where } W \sim \chi^2(4) \\ &= 0.5368 \quad \text{calculated using R} \\ &> 0.1\end{aligned}$$

there is no evidence based on the data against the hypothesis that the heights of husbands and wives are independent.

7.10 (a) The expected frequencies are:

$y_{ij}\,(e_{ij})$	Both	Mother	Father	Neither	Total
Above Average	$\frac{30\times50}{100} = 15$	$\frac{16\times50}{100} = 8$	$\frac{18\times50}{100} = 9$	18	50
Below Average	15	8	9	18	50
Total	30	16	18	36	100

The observed value of the likelihood ratio statistic is $\lambda = 10.8$. The degrees of freedom for the Chi-squared approximation are $(4-1)(2-1) = 3$ and

$$\begin{aligned}p - value &\approx P(W \geq 10.8) \quad \text{where } W \sim \chi^2(3) \\ &= 0.013 \quad \text{calculated using R}\end{aligned}$$

Since $0.01 < p - value < 0.05$, there is evidence based on the data against the hypothesis that birth weight is independent of parental smoking habits.

(b) The expected frequencies depending on whether the mother is a smoker or non-smoker are:

Mother smokes

e_{ij}	Father smokes	Father non-smoker	Total
Above average	$\frac{30\times15}{46} = 9.78$	5.22	15
Below average	20.22	10.78	31
Total	30	16	46

Mother non-smoker

$y_{ij}\,(e_{ij})$	Father smokes	Father non-smoker	Total
Above average	$\frac{18\times35}{54} = 11.67$	23.33	35
Below average	6.33	12.67	19
Total	18	36	54

For the Mother smokes table, the observed value of the likelihood ratio statistic is $\lambda = 0.2644$. The degrees of freedom for the Chi-squared approximation are $(2-1)(2-1) = 1$ and

$$\begin{aligned} p-value &\approx P(W \geq 0.2644) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{0.2644}\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2\left[1 - P(Z \leq 0.51)\right] = 0.60710 \end{aligned}$$

For the Mother non-smoker table, the observed value of the likelihood ratio statistic is $\lambda = 0.04078$. The degrees of freedom for the Chi-squared approximation equal $(2-1)(2-1) = 1$ and

$$\begin{aligned} p-value &\approx P(W \geq 0.04078) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{0.04078}\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2\left[1 - P(Z \leq 0.20)\right] = 0.83997 \end{aligned}$$

Since $p - value > 0.1$ in both cases, there is no evidence based on the data against the hypothesis that, given the smoking habits of the mother, birth weight is independent of the smoking habits of the father.

7.11 The expected frequencies are:

	Normal	Enlarged	Much enlarged	Total
Carrier present	$\frac{516\times72}{1398} = 26.57$	$\frac{589\times72}{1398} = 30.33$	15.09	72
Carrier absent	489.43	558.67	277.91	1326
Total	516	589	293	1398

The observed value of the likelihood ratio statistic is 7.3209 with

$$\begin{aligned} p-value &\approx P(W \geq 7.3209) = 0.026 \quad \text{where } W \sim \chi^2(2) = \text{Exponential}(2) \\ &= e^{-7.3209/2} = 0.02572 \end{aligned}$$

Since $0.01 < p-value < 0.05$, there is evidence based on the data against the hypothesis that the two classifications are independent.

7.12 The expected frequencies are given in brackets

	Employed	Unemployed	Total
No certificate, diploma or degree	66 [70.984]	10 [5.016]	76
High school diploma or equivalent	185 [187.734]	16 [13.266]	201
Postsecondary certificate, diploma or degree	683 [675.282]	40 [47.718]	723
Total	934	66	1000

The observed value of the likelihood ratio statistic is $\lambda = 6.1673$. The degrees of freedom for the Chi-squared approximation are $(3-1)(2-1) = 2$ and

$$p-value \approx P(W \geq 6.1673) = 0.0458 \quad \text{where } W \sim \chi^2(2)$$

Since $p-value < 0.05$, there is evidence based on the data to contradict the hypothesis that employment status is independent of educational level.

7.13 (a) The expected frequencies are:

e_{ij}	3 boys	2 boys	2 girls	3 girls	Total
Mother under 30	$\frac{29\times11}{64}$ $= 4.9844$	$\frac{29\times18}{64}$ $= 8.1563$	$\frac{29\times22}{64}$ $= 9.96883$	5.8906	29
Mother over 30	6.0156	9.8438	12.0313	7.1094	35
Total	11	18	22	13	64

The observed value of the likelihood ratio statistic is $\lambda = 0.5587$. The degrees of freedom for the Chi-squared approximation are $(4-1)(2-1) = 3$. Since

$$\begin{aligned} p-value &\approx P(W \geq 0.5587) \quad \text{where } W \sim \chi^2(3) \\ &= 0.9058 \quad \text{calculated using R} \\ &> 0.1 \end{aligned}$$

there is no evidence based on the data to contradict the hypothesis of no association between the sex distribution and age of the mother.

(b) The expected frequencies are:

y = no. of boys	3	2	1	0	Total
Observed Frequency	11	18	22	13	64
Expected Frequency	$64\,(0.5)^3$ $= 8$	$64\binom{3}{2}\,(0.5)^3$ $= 24$	$64\binom{3}{1}\,(0.5)^3$ $= 24$	8	64

The observed value of the likelihood ratio statistic is $\lambda = 5.4441$. The degrees of freedom for the Chi-squared approximation are $4 - 1 = 3$. Since

$$\begin{aligned} p - value &\approx P(W \geq 5.4441) \quad \text{where } W \sim \chi^2(3) \\ &= 0.1420 \quad \text{calculated using R} \\ &> 0.1 \end{aligned}$$

there is no evidence based on the data against the Binomial$(3, 0.5)$ model.

7.14 (a) The expected frequencies are:

e_{ij}	Rust-Proofed	Not Rust Proofed	Total
Rust present	$\frac{42\times 50}{100} = 21$	21	42
Rust absent	29	29	58
Total	50	50	100

The observed value of the likelihood ratio statistic is likelihood ratio statistic is

$$\lambda = 2\left[14\log\left(\frac{14}{21}\right) + 28\log\left(\frac{28}{21}\right) + 36\log\left(\frac{36}{29}\right) + 22\log\left(\frac{22}{29}\right)\right] = 8.1701$$

with

$$\begin{aligned} p - value &\approx P\left(W \geq 8.1701\right) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{8.1701}\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2\left[1 - P\left(Z \leq 2.86\right)\right] \\ &= 0.0042587 \end{aligned}$$

Since $0.001 < p - value < 0.01$ there is strong evidence against the hypothesis that the probability of rust occurring is the same for rust-proofed and non-rust-proofed cars based on the observed data.

7.15 The data in a two way table are:

y_{ij} $[e_{ij}]$	Cold	No Cold	Total
Vitamin C	20 [25]	80 [75]	100
Placebo	30 [25]	70 [75]	100
Total	50	150	200

If the probability of catching the cold is the same for each group, then an estimate of this probability is $\frac{50}{200} = 0.25$. The expected frequencies and observed frequencies are shown in the table. The original model consists of two independent Binomial models each with their own unknown parameter. Under the null hypothesis that the probability of catching a cold is the same for both groups the model is two independent Binomial models with only one unknown parameter. Therefore the degrees of freedom for the Chi-squared approximation are $2-1=1$. The observed value of the likelihood ratio statistic is

$$2\left[20\log\left(\frac{20}{25}\right)+80\log\left(\frac{80}{75}\right)+30\log\left(\frac{30}{25}\right)+70\log\left(\frac{70}{75}\right)\right]=2.6807$$

Since

$$\begin{aligned} p-value &\approx P(W\geq 2.68) \quad \text{where } W\sim\chi^2(1) \\ &= 2\left[1-P\left(Z\leq\sqrt{2.68}\right)\right] \quad \text{where } Z\sim N(0,1) \\ &= 2\left[1-P\left(Z\leq 1.64\right)\right] \\ &= 0.10157>0.1 \end{aligned}$$

there is no evidence based on the data against the hypothesis that the probability of catching a cold during the study period was the same for each group.

7.16 The data in a two way table are:

y_{ij} $[e_{ij}]$	Correct	Not Correct	Total
A	1328 [1333]	72 [67]	1400
B	1338 [1333]	62 [67]	1400
Total	2666	134	2800

If the probability of an error is the same for each algorithm, then an estimate of this probability is $\frac{134}{2800} = 0.0479$. The expected frequencies and observed frequencies are shown in the table. The original model consists of two independent Binomial models each with their own unknown parameter. Under the null hypothesis that the probability of an error is the same for both algorithms the model is two independent Binomial models with only one unknown parameter. Therefore the degrees of freedom for the Chi-squared approximation are $2-1=1$. The observed value of the likelihood ratio statistic is

$$2\left[1328\log\left(\frac{1328}{1333}\right)+72\log\left(\frac{72}{67}\right)+1338\log\left(\frac{1338}{1333}\right)+62\log\left(\frac{62}{67}\right)\right]=0.7845$$

Since

$$\begin{aligned} p-value &\approx P(W\geq 0.7845) \quad \text{where } W\sim\chi^2(1) \\ &= 2\left[1-P\left(Z\leq\sqrt{0.7845}\right)\right] \quad \text{where } Z\sim N(0,1) \\ &= 2\left[1-P\left(Z\leq 0.89\right)\right] \\ &= 0.37346>0.1 \end{aligned}$$

there is no evidence based on the data against the hypothesis that the probability of error is the same for each algorithm.

7.17 The expected frequencies are:

e_{ij}	Mon	Tue	Wed	Thu	Total
Lose	$\frac{12\times35}{56}$ $= 7.5$	$\frac{13\times35}{56}$ $= 8.125$	$\frac{16\times35}{56}$ $= 10$	9.375	35
Win	4.5	4.875	6	5.625	21
Total	12	13	16	15	56

Note that two of the frequencies are just under 5, however they are close enough to 5 that we will not collapse the table. The observed value of the likelihood ratio statistic is $\lambda = 0.8727337$. The degrees of freedom for the Chi-squared approximation are $(4-1)(2-1) = 3$. Since

$$\begin{aligned} p-value &\approx P(W \geq 0.8727337) \quad \text{where } W \sim \chi^2(3) \\ &= 0.8320023 \quad \text{calculated using R} \\ &> 0.1 \end{aligned}$$

there is no evidence based on the data to contradict the hypothesis that the probability of winning is the same across the four weekdays Monday to Thursday.

7.18 (a) Under the hypothesis $H_0 : \theta_{12} = \theta_{21} = \theta$ the likelihood function is

$$L(\theta_{11}, \theta) = \theta_{11}^{y_{11}} \theta^{y_{12}+y_{21}} (1 - 2\theta - \theta_{11})^{y_{22}}$$

and the log likelihood function is

$$l(\theta_{11}, \theta) = y_{11} \log \theta_{11} + (y_{12} + y_{21}) \log \theta + y_{22} \log (1 - 2\theta - \theta_{11})$$

Solving

$$\begin{aligned} \frac{\partial l}{\partial \theta_{11}} &= \frac{y_{11}}{\theta_{11}} - \frac{y_{22}}{1-2\theta-\theta_{11}} = 0 \\ \frac{\partial l}{\partial \theta} &= \frac{y_{12}+y_{21}}{\theta} - \frac{2y_{22}}{1-2\theta-\theta_{11}} = 0 \end{aligned}$$

gives

$$\hat{\theta}_{11} = \frac{y_{11}}{n}, \quad \hat{\theta}_{12} = \hat{\theta}_{21} = \frac{y_{12}+y_{21}}{2n}, \quad \hat{\theta}_{22} = \frac{y_{22}}{n}$$

(b) For the given data

$$\hat{\theta}_{11} = \frac{1325}{1400} = 0.9464, \quad \hat{\theta}_{12} = \hat{\theta}_{21} = \frac{16}{2800} = 0.0057, \quad \hat{\theta}_{22} = \frac{59}{1400} = 0.0421$$

The expected frequencies under H_0 are given in brackets

		B		
		Correct	Incorrect	
A	Correct	1325 [1325]	3 [8]	
	Incorrect	13 [8]	59 [59]	
				1400

The observed value of the likelihood ratio statistic is

$$2\left[0+3\log\left(\frac{3}{8}\right)+13\log\left(\frac{13}{8}\right)+0\right]=6.7382$$

Note that the number of correct y_{11} and the number of incorrect y_{22} for both algorithms do not affect the value of the likelihood ratio statistic. In the unconstrained model there were three parameters $(\theta_{11},\theta_{12},\theta_{21})$ and under H_0 there were two parameters (θ_{11},θ) so the degrees of freedom of the Chi-squared approximation are $3-2=1$.

$$\begin{aligned} p-value &\approx P(W\geq 6.7382) \quad \text{where } W\sim\chi^2(1) \\ &= 2\left[1-P\left(Z\leq\sqrt{6.7382}\right)\right] \quad \text{where } Z\sim N(0,1) \\ &= 2\left[1-P\left(Z\leq 2.60\right)\right] \\ &= 0.00932 \end{aligned}$$

Since $p-value<0.01$, there is strong evidence based on the data against the hypothesis that the probability of error is the same for each algorithm which is a completely different conclusion compared with the experiment that was not paired.

SOLUTIONS TO CHAPTER 8 PROBLEMS

8.1 (a) The observed and expected frequencies (in square brackets) assuming independence are given in the table where $e_{11} = (3301 \times 28358)/50267 = 1862.25$, $e_{12} = (3301 \times 15328)/50267 = 1006.57$ and all other expected frequencies can be determined by subtraction.

No. of cigarettes	0	$1-20$	>20	Total
Weight ≤ 2.5	1322 [1862.25]	1186 [1006.57]	793 [432.17]	3301
Weight > 2.5	27036 [26495.75]	14142 [14321.42]	5788 [6148.83]	46966
Total	28358	15328	6581	50267

The observed value of the likelihood ratio statistic is 480.644. Since

$$\begin{aligned} p-value &\approx P(W \geq 480.644) \quad \text{where } W \sim \chi^2(2) = Exponential(2) \\ &\approx 0 \end{aligned}$$

there is very strong evidence based on the data against the hypothesis that birth weight is independent of the mother's smoking habits. The data suggest that lower birth weights are associated with mothers who smoke more.

(b) Since this is an observational study, evidence of an association does not imply a causal relationship. In particular the researchers cannot conclude that if the mothers stopped smoking then birth weights would increase.

The researchers would need to conduct an experimental study in which they controlled how much the mothers smoked in order to conclude that the evidence of a relationship between mother's smoking habits and birth weights implies a causal relationship. Of course a study in which the researchers "controlled" the smoking habits of the mothers would be very difficult to conduct.

(c) An association between the smoking habits of fathers and birth weights is to be expected since there is probably an association between the smoking habits of the

fathers and the smoking habits of the mothers. That is, the association between the smoking habits of fathers and birth weights is a result of the association between the smoking habits of the fathers and the smoking habits of the mothers together with the association between the smoking habits of the mothers and birth weights.

8.2 (a) The observed and expected frequencies (in square brackets) assuming independence are given in the table.

	Mark ≤ 80	Mark > 80	Total
Standard Lecture	60 $\left[\frac{75\times106}{150} = 53\right]$	15 [22]	75
CAI	46 [53]	29 [22]	75
Total	106	44	150

The observed value of the likelihood ratio statistic is 6.3874 and

$$\begin{aligned} p-value &\approx P\left(W \geq 6.3874\right) \quad \text{where } W \sim \chi^2\left(1\right) \\ &= 2\left[1 - P\left(Z \leq \sqrt{6.3874}\right)\right] \quad \text{where } Z \sim N\left(0,1\right) \\ &= 2\left[1 - P\left(Z \leq 2.53\right)\right] = 0.0114 \end{aligned}$$

Since $0.01 < p - value < 0.05$, there is evidence based on the data against the hypothesis of independence, that is, against the hypothesis that marks are independent of whether the student received the standard lecture or some CAI.

(b) In order to conclude that CAI increases the chances of achieving a mark over 80%, randomization of the students to either a standard lecture or to CAI would need to have been done.

8.3 (a) The observed and expected frequencies (in square brackets) assuming independence are given in the table.

	Admitted	Not Admitted	Total
Male Applicants	3738 $\left[\frac{8442\times5232}{12763} = 3460.67\right]$	4704 [4981.33]	8442
Female Applicants	1494 [1771.33]	2827 [2549.67]	4321
Total	5232	7531	12763

The observed value of the likelihood ratio statistic is $\lambda = 112.398$. Since

$$\begin{aligned} p-value &\approx P\left(W \geq 112.398\right) \quad \text{where } W \sim \chi^2\left(1\right) \\ &= 2\left[1 - P\left(Z \leq \sqrt{112.398}\right)\right] \quad \text{where } Z \sim N\left(0,1\right) \\ &= 2\left[1 - P\left(Z \leq 10.60\right)\right] \approx 0 \end{aligned}$$

there is very strong evidence based on the data against the hypothesis of independence, that is, against the hypothesis that whether the student is admitted or not is independent of their sex.

(b) Only Program A shows any evidence of non-independence, and that is in the direction of a lower admission rate for males.

(c) This is an example of Simpson's Paradox. The association is observed in the collapsed table since in the table broken down by program we observe that over 50% of the men applied to programs A and B which had higher admission rates while over 50% of the women applied to programs C - F which had much lower admission rates.

8.4 (a) Since $n_1 = n_2 = 100$ we will use the test statistic

$$D = \frac{\left|\bar{Y}_1 - \bar{Y}_2\right|}{\sqrt{\frac{S_1^1}{100} + \frac{S_2^1}{100}}}$$

The observed value of the test statistic

$$d = \frac{|11.7 - 12.0|}{\sqrt{\frac{(2.1)^2}{100} + \frac{(2.4)^2}{100}}} = 0.9407$$

Since

$$\begin{aligned} p - value &\approx P\left(|Z| \geq 0.94\right) \quad \text{where } Z \sim N\left(0,1\right) \\ &= 2\left[1 - P\left(Z \leq 0.94\right)\right] \\ &= 2\left[1 - P\left(Z \leq 0.94\right)\right] \\ &= 0.34722 > 0.1 \end{aligned}$$

there is no evidence based on the data against the hypothesis of no difference between the mean amount of rust for rust-proofed cars as compared to non-rust-proofed cars.

(b) Since the cars were not randomly assigned to rust-proofing or not, a variate that the manufacturer is not aware of which is not rust-proofing could have had an effect on the results. For example, maybe the cars that were rust-proofed were owned by drivers who lived in areas where salt is used frequently in winter and therefore they had decided to use rust-proofing to reduce the effects of the salt. The drivers who did not chose rust-proofing might live in areas where driving conditions do not affect the rusting of cars. It would have been better to use randomization to decide which of the cars received rust-proofing and which did not. In this way the variates that affect the rusting of cars that the manufacturer is not aware of are balanced in the two groups.

8.5 This study is an observational study based only on data from the United States. A causal relationship cannot be concluded only on the basis of these data. To establish a causal relationship a strong association would need to be observed in numerous studies in many countries. Other possible sources of confounding variates would need to be examined in these studies to determine if they could explain the association. A pathway by which drinking wine causes cirrhosis of the liver would need to be established.

8.6 This is an experimental study since Hooker observed the boiling point of water at many different elevation levels. (We don't know how he chose these levels.) We are assuming that his method for boiling water and for measuring water temperature and atmospheric pressure were controlled as much as possible at the different elevations to avoid other variates affecting the relationship. Recall that Hooker was interested in using the boiling point of water as the explanatory variate and atmospheric pressure as the response variate since measuring the boiling point would give travelers a quick way to estimate elevation, using the known relationship between elevation and barometric pressure, and the model relating pressure to boiling point. The causal relationship actually works in the reverse direction, that is, it is atmospheric pressure which is causing the change in the boiling point of water. This conclusion however requires an argument based on physics. Pressure on the surface of water tends to keep the water molecules contained. As pressure increases, water molecules need additional heat to gain the speed necessary for escape. Lowering the pressure lowers the boiling point because the molecules need less speed to escape.

8.7 It is important that the subject does not whether they are receiving the treatment since if they do know they might think the treatment is working just because they know that they are receiving a treatment (the placebo effect). It is important that the physician not know whether the subject is receiving the treatment or not since knowing might affect their decision about whether the treatment is working or not.

SAMPLE TESTS

Sample Midterm Test 1

1. Multiple choice questions.

(a) Which one of the following can be appropriately modeled using a Binomial model?

A: time between arrivals of buses at a bus stop

B: number of tosses needed until we get 10 Heads in total when tossing a coin

C: closing price in dollars of a stock

D: number of calls received by a call center in one hour

E: number of people in a sample drawn at random from a large population that have a certain disease

(b) Which one of the following statements is **FALSE**?

A: Pie charts and bar charts are suitable for representing categorical data.

B: In a relative frequency histogram the height of each rectangle is equal to k times the number of observations in the interval for some positive constant k.

C: The sample variance of a data set cannot be determined from a boxplot.

D: A run chart is a good way to summarize data collected over time.

(c) Which one of the following statements is **FALSE**?

A: $L(\theta)$ and $l(\theta) = \log L(\theta)$ are maximized for the same value of θ.

B: $L(\theta)$ and $l(\theta) = \log L(\theta)$ have the same concavity near their maximum value.

C: $L(\theta)$ and $l(\theta) = \log L(\theta)$ have the same shape.

D: $l(\theta) = \log L(\theta)$ is a one-to-one function of $L(\theta)$.

(d) Which one of the following statements is **FALSE**?

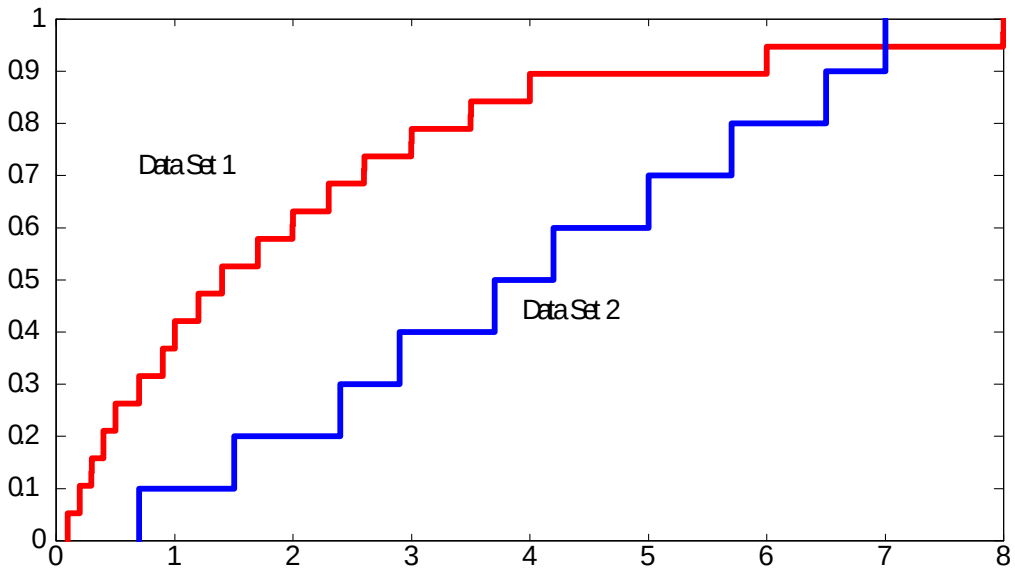
A: If y successes are observed in n Bernoulli trial with $P(\text{Success}) = \theta$ then the maximum likelihood estimate of θ is $\hat{\theta} = y/n$.

B: For an observed random sample $y_1, y_2, \ldots, y_n$ from a Poisson(θ) distribution the maximum likelihood estimate of θ is $\hat{\theta} = \bar{y}$.

C: For an observed random sample $y_1, y_2, \ldots, y_n$ from a Exponential(θ) distribution the maximum likelihood estimate of θ is $\hat{\theta} = \bar{y}$.

D: For an observed random sample $y_1, y_2, \ldots, y_n$ from a $G(\mu, \sigma)$ distribution the maximum likelihood estimate of $\theta = (\mu, \sigma^2)$ is $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = \left(\bar{y},\ \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)$.

(e) In the graph below the empirical cumulative distribution function is graphed for two different data sets. The observations in each data set are unique.



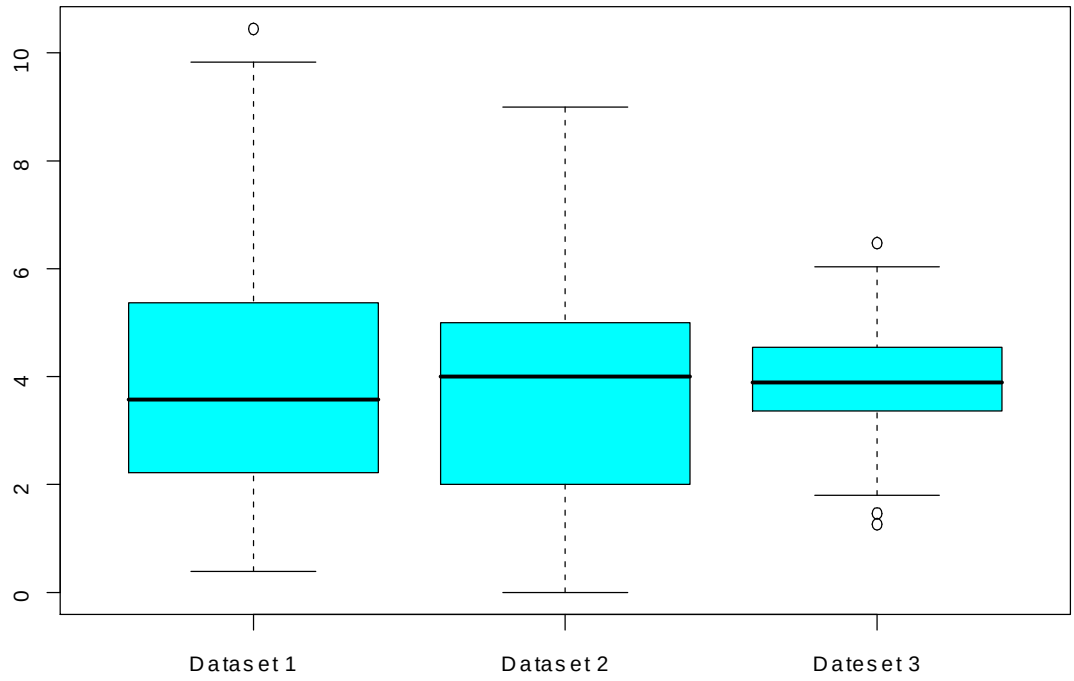
Which one of the following statements is **TRUE**?

A: There are more observations in Data Set 2 than in Data Set 1.
B: All the values in both data sets are positive.
C: For Data Set 2, $\hat{F}(5) = 0.6$.
D: The skewness of Data Set 1 is negative.

(f) Suppose a and b are positive constants. The function $G(\theta) = \theta^a (1-\theta)^b$, $\theta > 0$ is maximized for:

A: $\frac{a}{a+b}$
B: $\frac{a}{b}$
C: $\frac{b}{a}$
D: b
E: None of the above.

(g) In the figure below are boxplots for 3 different datasets. Assume all 3 datasets are unimodal.



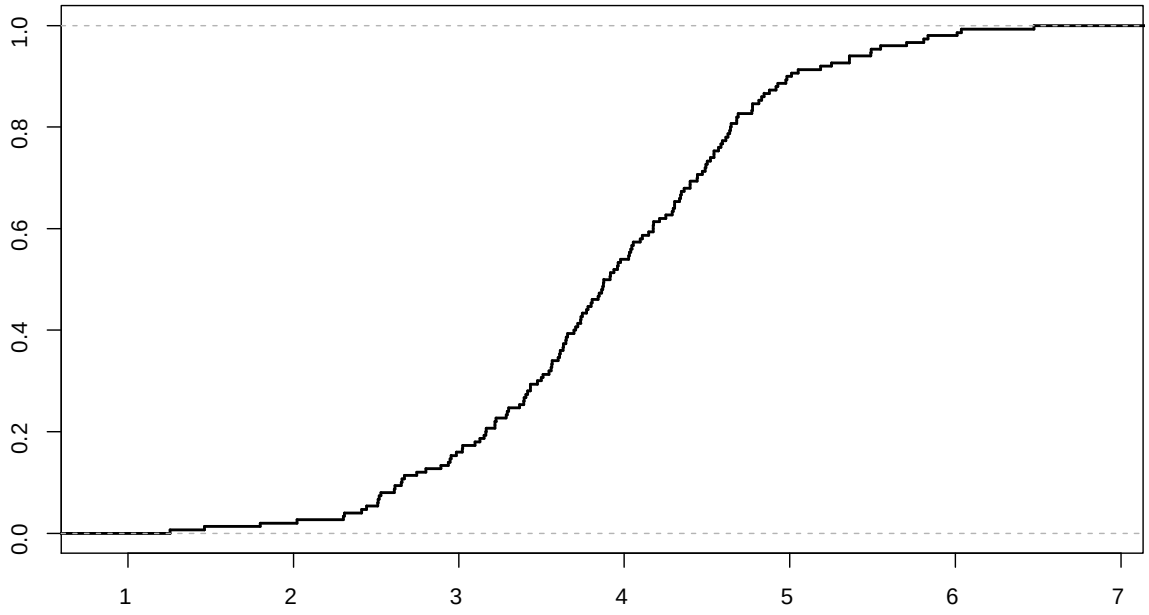
Which one of the following statements is **FALSE**?

A: Dataset 3 has the smallest variability.
B: Dataset 1 is negatively skewed.
C: Dataset 3 is the most bell-shaped.
D: Dataset 2 has the largest sample median
E: Dataset 1 has the largest range.

(h) The correlation between two variates x and y can be computed in R using which of the following commands:

A: cov(x, y)
B: cor(x, y)
C: cov(x, y)(var(x)*var(y))
D: cor(x, y)/(var(x)*var(y))
E: None of the above.

(i) Which of the following commands in R would produce the following plot for the variate y?



A: ecdf(y)
B: hist(y)
C: boxplot(y)
D: plot(ecdf(y))
E: None of the above.

(j) Consider the following R console output:

fivenum(y)
[1] 0.0013 0.2805 0.6747 1.3374 8.6917

The IQR of the variate y is given by:

A: 0.6747
B: 1.0569
C: 8.6904
D: None of the above.

2. In modelling the number of transactions of a certain type received by a central computer for a company with many on-line terminals the Poisson distribution can be used. If the transactions arrive at random at the rate of θ per minute then the probability of y transactions in a time interval of length t minutes is

$$P(Y = y; \theta) = f(y; \theta) = \frac{(\theta t)^y}{y!} e^{-\theta t} \quad \text{for } y = 0, 1, \ldots \text{ and } \theta \geq 0 \tag{11.1}$$

(a) Suppose $y_1, y_2, \ldots, y_n$ were the number of transaction recorded in n independent $t = 1$ minute intervals. Find the the maximum likelihood estimate of θ based on the model (11.1) and these data. Clearly show all your steps.

(b) Suppose that for $n = 200$ independent $t = 1$ minute intervals the observed frequencies were those given in the table below. Assuming $\theta = \hat{\theta} = 2.1$, complete the following table of expected frequencies. Comment on how well the model fits the data.

	0	1	2	3	4	≥ 5	Total
Observed Frequency	28	45	56	40	21	10	200
Expected Frequency						12.425	200

(c) What is the maximum likelihood estimate of the probability that during a 2 minute interval there are no transactions?

3. Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the $G(0, \sigma)$ distribution with probability density function

$$f(y; \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/(2\sigma^2)} \text{ for } y \in \Re \text{ and } \sigma > 0$$

(a) Find the likelihood function $L(\sigma)$ and the maximum likelihood estimate $\hat{\sigma}$ based on the observed data $y_1, y_2, \ldots, y_n$. Clearly show all your steps.

(b) Show that the relative likelihood function $R(\sigma)$ is given by

$$R(\sigma) = \left(\frac{\hat{\sigma}}{\sigma}\right)^n \exp\left\{\frac{n}{2}\left[1 - \frac{(\hat{\sigma})^2}{\sigma^2}\right]\right\} \quad \text{for } \sigma > 0$$

Note: $\exp(x) = e^x$

(c) Suppose $\hat{\sigma} = 1.2$ for a given data set. If $Y \sim G(0, \sigma)$ then determine the maximum likelihood estimate of $P(Y > 0.3; \sigma)$.

4. The data below are the final grades of 90 students in a second year statistics course:

99	96	95	94	94	93	93	92	92	92
91	91	91	90	90	90	89	89	88	88
88	87	87	86	86	86	86	85	85	85
85	85	84	84	84	83	83	82	82	82
82	81	81	80	80	79	79	79	78	78
77	77	77	76	76	75	75	75	74	74
73	73	72	71	71	70	70	70	69	69
68	68	68	67	66	66	65	64	64	63
61	60	59	57	54	54	53	48	47	42

For these data

$$\sum_{i=1}^{90} y_i = 6987, \quad \sum_{i=1}^{90} y_i^2 = 555863, \quad \text{and sample kurtosis} = 3.0211$$

A relative frequency histogram and qqplot for these data are given below:

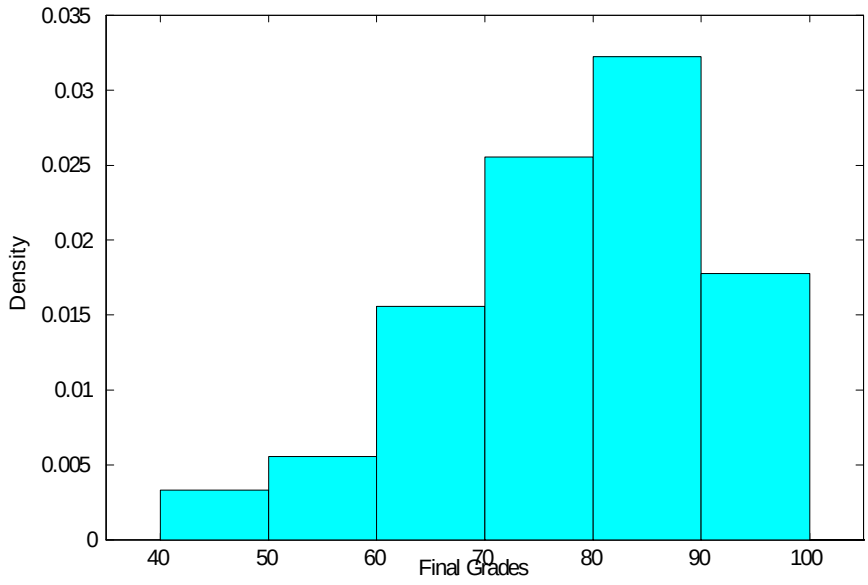


Figure 11.2: Relative Frequency Histogram of Final Grades

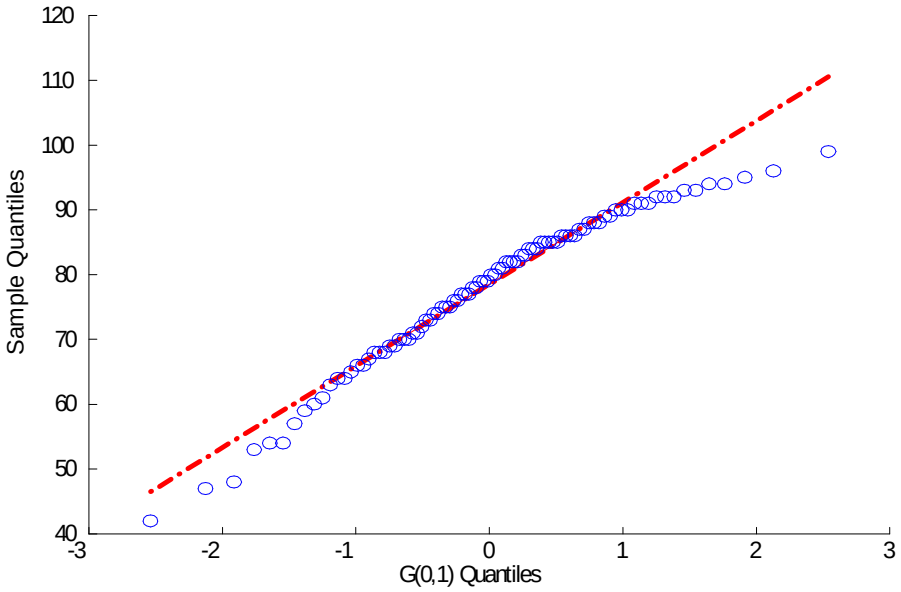


Figure 11.3: Qqplot of Final Grades

Answer questions $(a) - (e)$ based on the given information.

(a) The five-number summary for these data is:

_____, _____, _____, _____, _____

(b) For these data:

$$\text{sample mean } = \bar{y} = \_\_\_\_\_$$

and

$$\text{sample standard deviation} = s = \_\_\_\_\_$$

(c) For these data the sample skewness would be (Circle the letter corresponding to your choice.):

A: negative
B: approximately zero
C: positive
D: not enough information to tell

(d) For these data determine the proportion of observations in the interval $[\bar{y} - s, \bar{y} + s]$. Compare this with $P(Y \in [\mu - \sigma, \mu + \sigma])$ where $Y \sim G(\mu, \sigma)$.

(e) Find the interquartile range (IQR) for these data. Show that $IQR = 1.349\sigma$ for data from a Gaussian distribution.

(f) Using both the numerical and graphical summaries for these data, assess whether it is reasonable to assume a Gaussian model for these data. **You must support your conclusion with reasons written in complete sentences.**

5. Answer the questions below based on the following article (condensed) which appeared in the Globe & Mail newspaper on January 22, 2014:

Early engagement key to getting girls into science careers, Canadian study says

Girls are almost three times more likely to consider careers in science, math and engineering if they participate in science fairs and summer camps – particularly in the early grades – according to a new Canadian report. The study by researchers at Mount Saint Vincent University in Halifax also suggests that good grades and teacher influence matters less than exposure to these outside-the-classroom activities.

The findings come at a time when governments are reaching out to young women in an effort to persuade them to consider the so-called STEM fields of learning – science, technology, engineering and mathematics – and organizations have stepped up their mentoring efforts. Learning experts say it is crucial to reach girls before their enthusiasm wanes and they drop science and math courses, which are optional in high school. "I think this is a wake-up call. We need to increase the engagement level, and we need to encourage it from a young age," said the study's lead investigator, Tamara Franz-Odendaal, an associate professor at the university.

Prof. Franz-Odendaal and her team surveyed about 600 students in Grades 7 through 9 last year from the provinces New Brunswick, Nova Scotia and Prince Edward Island. The data were collected using an online survey that students completed during school hours. They found girls who engaged in activities, such as science fairs, competitions and engineering summer camps, were 2.7 times more likely to consider a STEM career. For boys, the influence was statistically insignificant.

(a) What type of study is this and why?

(b) Define the Problem for this study.

(c) Is the type of Problem descriptive, causative, or predictive? Explain why.

(d) What are the two most important variates in this study and what is their type?

(e) Define a suitable target population for this study.

(f) Define a suitable study population for this study.

(g) Give a possible source of study error for this study in relation to your answers to to (e) and (f).

(h) What information is given about the sampling protocol for this study?

(i) Give a possible source of sample error for this study based on the information you have stated in (h).

(j) What type of numerical summary is the number 2.7 mentioned in the article?

Sample Midterm Test 2

1. Between 10:00 am September 26 and 10:00 pm September 28 2016 the Federation of Students at the University of Waterloo conducted a referendum. The question was:

Which one of the following options do you support?

(1) Keep the mandatory, refundable $4.75 per academic term fee for WPIRG (Waterloo Public Interest Research Group).

(2) Remove the mandatory, refundable $4.75 per academic term fee for WPIRG (Waterloo Public Interest Research Group).

All eligible undergraduates were informed by email to cast their ballot online. Of the $31,380$ eligible voters, 8788 voted, and 7156 chose option 2.

(a) The Federation of Students used an empirical study to determine whether or not students supported the removal of the WPIRG fee. The Plan step of the empirical study involved using an online referendum. Using complete sentences give at least one advantage and at least one disadvantage of using the online referendum in this context.

(b) Assume the model $Y \sim Binomial\,(8788, \theta)$ where Y = number of students who chose option (2): "Remove the mandatory, refundable $4.75 per academic term fee for WPIRG." What does the parameter θ represent in this study? Using complete sentences indicate how valid you think the Binomial model is and why?

(c) The maximum likelihood estimate of θ based on the observed data is

_____. (You do not need to derive this estimate.)

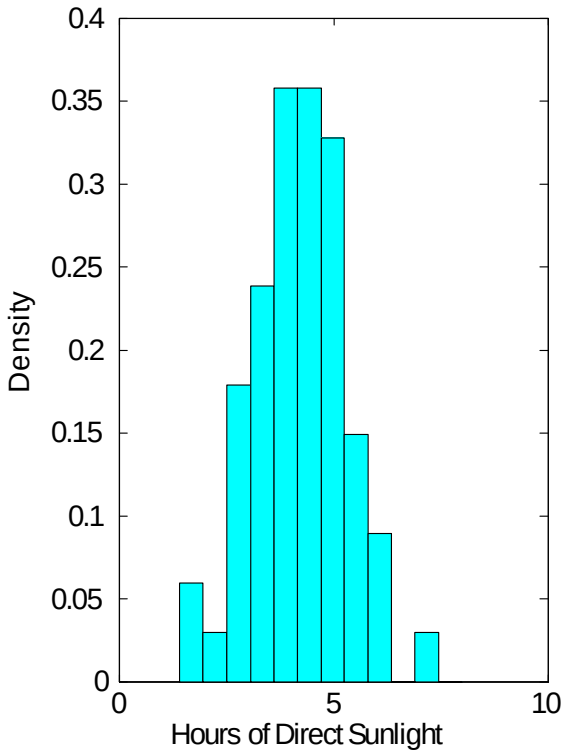
(d) The $p - value$ for testing the hypothesis $H_0 : \theta = 0.8$ is approximately

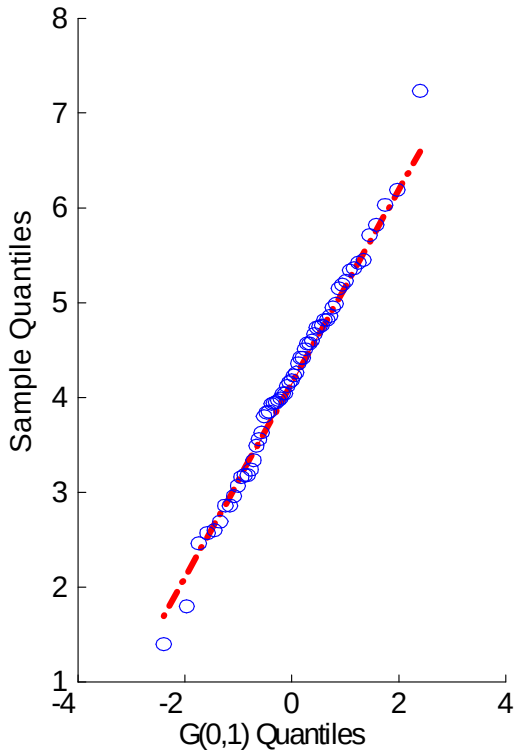
_____ . **Show all your steps.**

(e) State your conclusion regarding the hypothesis $H_0 : \theta = 0.8$ in a sentence.

(f) **By reference to your answer in** (d), indicate whether the value $\theta = 0.8$ is inside an approximate 95% confidence interval or not. Justify your answer but do not construct the interval.

2. To decide whether to install solar panels on the roof of her house a homeowner records the number of hours of full sunlight on her roof for 61 consecutive days in June and July. A relative frequency histogram and a qqplot for these data are given below:Let y_i = number





of hours of full sunlight on the i'th day, $i = 1, 2, \ldots, 61$. For these data

$$\sum_{i=1}^{61} y_i = 255.28 \quad \text{and} \quad \sum_{i=1}^{61} (y_i - \bar{y})^2 = 71.5607$$

To analyze these data the model $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, 61$ independently is assumed.

(a) Using complete sentences indicate how reasonable the Gaussian model is for these data.

(b) In a complete sentence explain clearly what the parameter μ represents.

(c) For these data the maximum likelihood estimate of μ is _____
and the maximum likelihood estimate of σ is _____.
(You do not need to derive these estimates.)

(d) A 99% confidence interval for μ based on these data is (**show all your steps**):

_____

(e) The Solar Energy Association recommends that the average number of hours of full sunlight in a day should be at least 4 to generate enough energy to make solar panels

worthwhile. Using complete sentences indicate what the homeowner should conclude about whether or not it is worthwhile placing solar panels on the roof of her house. Note any limitations of her study.

(f) The $p-value$ for testing $H_0 : \sigma = 1$ is between

_____ and _____. **Show all your steps.**

3. Suppose $Y \sim Exponential\,(\theta)$ with probability density function

$$f(y;\theta) = \frac{1}{\theta} e^{-y/\theta} \quad \text{for } y > 0 \text{ and } \theta > 0$$

(a) Use Change of Variable to show that $W = \frac{2Y}{\theta}$ has probability density function given by

$$g\,(w) = \frac{1}{2} e^{-w/2} \quad \text{for} \quad w > 0$$

which is the probability density function of a $\chi^2(2)$ random variable.

(b) Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the $Exponential\,(\theta)$ distribution. Use your result from (a) and theorem(s) that you have learned in class to show that

$$U = \sum_{i=1}^{n} \frac{2Y_i}{\theta} \sim \chi^2\,(2n)$$

(c) Explain clearly how the pivotal quantity U given in (b) can be used to obtain a two-sided, equal tailed, $100p\%$ confidence interval for θ.

(d) Suppose $W \sim \chi^2\,(20)$ and let a and b be such that $P\,(W \leq a) = 0.025 = P\,(W \geq b)$.

Then $a =$ _____ and $b =$ _____. (Use all the decimal places from the table.)

(e) Suppose $y_1, y_2, \ldots, y_{10}$ is an observed random sample from the $Exponential\,(\theta)$ distribution with

$$\sum_{i=1}^{10} y_i = 62.4$$

(i) Using your results from (c) and (d), a 95% confidence interval for θ based on the pivotal quantity U is:

_____. Show your work.

(ii) An approximate 95% confidence interval for θ based on the asymptotic Normal pivotal quantity is:

_____. Show your work. Compare this interval with the interval you obtained in (i).

4. **Circle the letter corresponding to your choice.**

(a) Suppose $Y \sim Binomial\,(n, \theta)$. An experiment is to be conducted in which data y are to be collected to estimate θ. To ensure that the width of the approximate 90% confidence interval for θ is no wider that $2\,(0.03)$, the sample size n should be at least:

A: 1068
B: 2401
C: 752
D: 267
E: 188

(b) For a Binomial experiment the approximate 95% confidence interval for θ based on the asymptotic Normal pivotal quantity was 0.75 ± 0.05. Which statement is **TRUE**?

A: $P\,(\theta \in [0.7, 0.8]) = 0.95$.
B: The interval $[0.7, 0.8]$ is also a 15% likelihood interval.
C: We are 95% confident that $\theta = \hat{\theta}$.
D: If the Binomial experiment was repeated 100 times independently and an approximate 95% confidence interval was constructed each time then approximately 95 of these intervals would contain the true value of θ.
E: None of the above.

(c) Which statement is **FALSE**?

A: A 15% likelihood interval is an approximate 95% confidence interval.
B: A 10% likelihood interval is an approximate 90% confidence interval.
C: Likelihood intervals must usually be found numerically or from a graph of the relative likelihood function.
D: Likelihood intervals are as good or better than approximate confidence intervals based on asymptotic Normal pivotal quantities.
E: The likelihood ratio statistic is an asymptotic pivotal quantity.

(d) Suppose $Y_i \sim G\,(\mu, \sigma)\,,\ i = 1, 2, \ldots, n$ independently. Let

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \quad \text{where} \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

The distribution of T is:

A: $G(0, 1)$
B: $G(0, \sigma)$
C: $t\,(n)$
D: $t\,(n-1)$
E: $\chi^2\,(n-1)$

(e) Suppose we have n independent observations from a $G(\mu, \sigma)$ distribution. Which statement is **TRUE**?

A: $\tilde{\mu}$ is a point estimate of μ.

B: If σ is known then $\bar{y} \pm 1.645\sigma/\sqrt{n}$ is a 95% confidence interval for μ.

C: If σ is unknown then $\bar{y} \pm as/\sqrt{n}$ is a 95% confidence interval for μ if $P(T \leq a) = 0.975$ and $T \sim t(n-1)$.

D: If σ is unknown then $\bar{y} \pm as/\sqrt{n}$ is a 95% confidence interval for μ if $P(T \leq a) = 0.95$ and $T \sim t(n-1)$.

E: S is the maximum likelihood estimator of σ.

(f) Data are collected in an experiment to test the null hypothesis H_0 using the test statistic D. The $p - value$ for testing H_0 is equal to

A: the probability that the null hypothesis H_0 is true.

B: the probability that the alternative hypothesis H_A is true.

C: the probability of obtaining a value of D as unusual or more unusual than the observed value of D if H_0 is true.

D: the probability of obtaining a value of D as unusual or more unusual than the observed value of D if the alternative hypothesis H_A is true.

E: None of the above.

(g) Which statement is **FALSE**?

A: For Binomial data the likelihood ratio statistic is a continuous random variable.

B: The distribution of the likelihood ratio statistic based on a random sample $Y_1, Y_2, \ldots, Y_n$ is approximately $\chi^2(1)$ for large n.

C: For Exponential data, the likelihood ratio statistic is a continuous random variable.

D: For Binomial(n, θ) data, an approximate 95% confidence interval for θ based on the asymptotic Normal pivotal quantity can contain values outside the interval $[0, 1]$.

E: For Exponential(θ) data, an approximate 95% confidence interval for θ based on a 15% likelihood interval only contains values of θ greater than zero.

For questions (h) and (i), suppose that a data set is assumed to be a random sample from a $G(\mu, \sigma)$ distribution where μ and σ are unknown. Suppose also that the data set is stored in the variable y and the following code has been run in R:

```
ybar<-mean(y)
n<-length(y)
s2<-var(y)
s<-sqrt(s2)
```

(h) Which of the following R commands gives a 95% confidence interval for the mean μ?

A: c(ybar-qnorm(0.975,0,1)*s/sqrt(n),ybar+qnorm(0.975,0,1)*s/sqrt(n))
B: c(ybar-qnorm(0.95,0,1)*s/sqrt(n),ybar+qnorm(0.95,0,1)*s/sqrt(n))
C: c(ybar+qt(0.05,n-1)*s/sqrt(n),ybar+qt(0.95,n-1)*s/sqrt(n))
D: c(ybar-qt(0.975,n)*s/sqrt(n),ybar+qt(0.975,n)*s/sqrt(n))
E: c(ybar-qt(0.975,n-1)*s/sqrt(n),ybar+qt(0.975,n-1)*s/sqrt(n))

(i) Which of the following R commands gives a 95% confidence interval for the standard deviation σ?

A: c(sqrt((n-1)*s2/qchisq(0.025,n-1)),sqrt((n-1)*s2/qchisq(0.975,n-1)))
B: c(sqrt((n-1)*s2/qchisq(0.975,n-1)),sqrt((n-1)*s2/qchisq(0.025,n-1)))
C: c(sqrt((n-1)*s2/qchisq(0.05,n-1)),sqrt((n-1)*s2/qchisq(0.95,n-1)))
D: c(sqrt((n-1)*s2/qchisq(0.025,n)),sqrt((n-1)*s2/qchisq(0.975,n)))
E: c(sqrt((n-1)*s2/qchisq(0.975,n)),sqrt((n-1)*s2/qchisq(0.025,n)))

(j) Suppose that a data set is assumed to be a random sample from an *Exponential* (θ) distribution. Suppose also that the data set is stored in the variable y and the following code has been run in R:

```
thetahat<-mean(y)
n<-length(y)
```

Which of the following R commands does **NOT** give the observed value of the likelihood ratio statistic evaluated at theta for these data?

A: 2*(log((theta/thetahat)^n)+n*(thetahat/theta-1))
B: 2*log((theta/thetahat)^n*exp(n*(1-theta/thetahat)))
C: -2*n*(log(thetahat/theta)+1-thetahat/theta)
D: -2*log((thetahat/theta)^n*exp(n*(1-thetahat/theta)))
E: -2*n*log((thetahat/theta)*exp(1-thetahat/theta))

Sample Final Exam

1. Let Y = the number of children in the family of a randomly chosen child. A proposed probability function for Y is

$$f(y;\theta) = P(Y = y;\theta) = y\theta^{y-1}(1-\theta)^2 \quad \text{for } y = 1, 2, \ldots; \quad 0 \leq \theta < 1 \tag{1}$$

Suppose 50 children are chosen at random and the observed data are $y_1, y_2, \ldots, y_{50}$ with $\sum_{i=1}^{50} y_i = 100$.

(a) Assume that the model (1) holds. Find the maximum likelihood estimate of θ based on these data. Clearly show all your steps.

(b) Assuming the model (1) holds and $\hat{\theta} = \frac{1}{3}$, complete the following table of expected frequencies.

Number of Children	1	2	3	≥ 4	Total
Observed Frequency	21	15	8	6	50
Expected Frequency				5.5556	50

(c) Using your results from (b) and the likelihood ratio test statistic, test the hypothesis that model (1) is a suitable model for these data. Be sure to give the approximate $p-value$ and give your conclusion in a sentence.

2. A random sample of 50 students from STAT 231 were asked the length of time (in hours) they spent in completing the first R assignment. The data were:

0.02	0.02	0.03	0.04	0.05	0.05	0.06	0.09	0.09	0.10
0.10	0.10	0.13	0.13	0.14	0.16	0.16	0.17	0.17	0.18
0.19	0.20	0.21	0.22	0.22	0.27	0.27	0.28	0.31	0.37
0.38	0.39	0.43	0.47	0.47	0.47	0.47	0.49	0.52	0.61
0.62	0.63	0.64	0.74	0.77	0.90	0.99	1.37	2.00	2.02

Let Y_i = length of time spent completing the first R assignment by student i, $i = 1, 2, \ldots, 50$. A proposed model for the data is that Y_i has probability density function

$$f(y;\theta) = \frac{\theta}{(1+y)^{\theta+1}} \quad \text{for } y \geq 0 \text{ and } \theta > 0 \tag{2}$$

(*a*) The data were stored in the variable y in R. Answer (*i*)-(*iii*) using the following output from R:

> sum(y)
[1] 19.9
> sum(log(1+y))
[1] 14.960
> fivenum(y)
[1] 0.020 0.130 0.245 0.490 2.020

(*i*) sample median = _____

(*ii*) IQR = _____

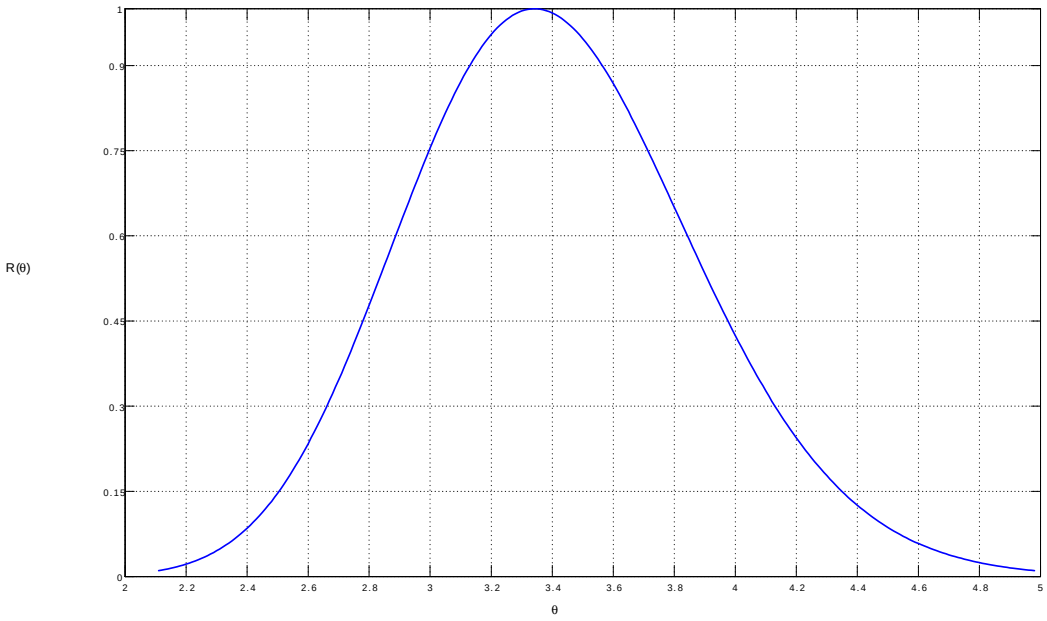
(*iii*) sample skewness is positive / negative (circle your choice)

(*b*) Assume that the 50 observations represent a random sample from the distribution with probability density function given by (2). Derive the maximum likelihood estimate of θ based on these data. Show all your steps clearly.

(*c*) Briefly describe two graphical summaries that could be used to check the fit of the model (2) to the data.

(*d*) A graph of $R(\theta)$ for the given data and model is given below. An approximate 95% confidence interval for θ is [_____, _____]

(Use 2 decimal places.)



(*e*) Test $H_0 : \theta = 4.3$ using the likelihood ratio test statistic. You may use the fact that $R(4.3) = 0.1774$. Be sure to give the approximate $p - value$ and a conclusion.

3. Researchers in the Kinesiology Department at a very large university with highly ranked sports teams were interested in comparing two exercise programs for treating sprained ankles. Ninety-two athletes who attended the university sports injury clinic for sprained ankles were randomly assigned to two different programs. In Program 1, 41 athletes were asked to complete a series of stretching exercises followed by icing. In Program 2, 51 athletes were asked to complete a series of stretching exercises followed by acupuncture. Let y_{ij} = the number of days until return to sports activity for athlete j in Program i, $j = 1, 2, \ldots, n_i$ and $i = 1, 2$. To analyze these data assume the model for data from Program 1 is,

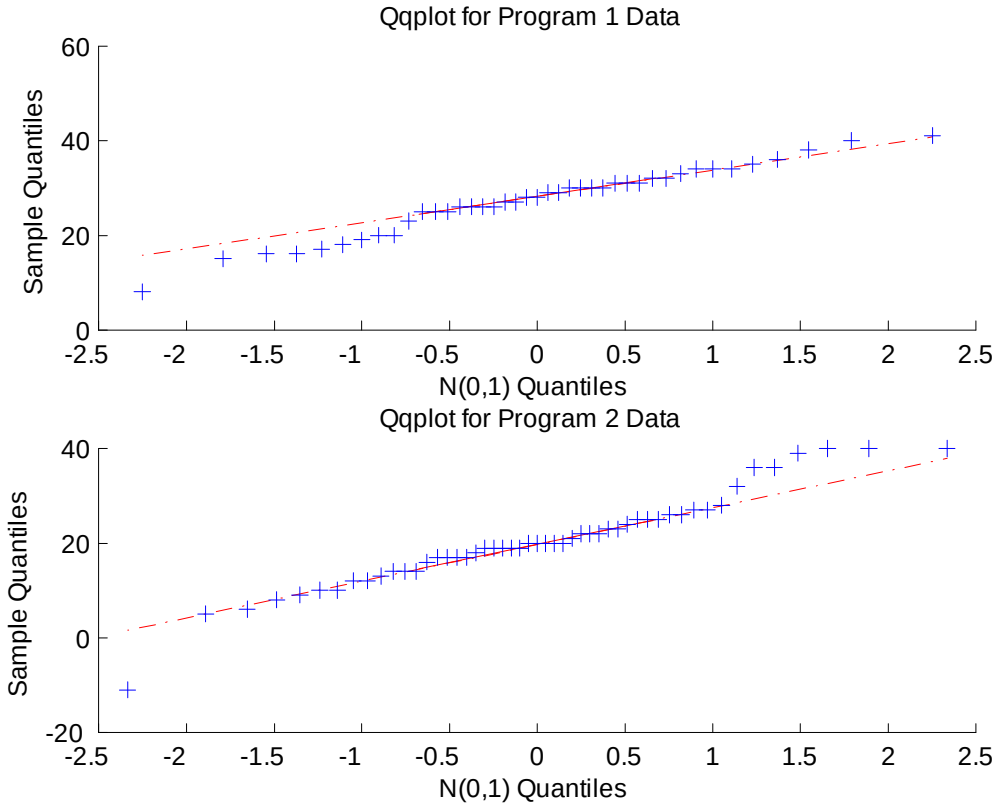
$$Y_{1j} \sim G(\mu_1, \sigma), \quad j = 1, 2, \ldots, 41 \text{ independently}$$

and independently, the model for data from Program 2 is,

$$Y_{2j} \sim G(\mu_2, \sigma), \quad j = 1, 2, \ldots, 51 \text{ independently}$$

where μ_1, μ_2 and σ are unknown parameters.

The observed data gave the following summaries:



Program 1	$\bar{y}_1 = 27.3415$	$s_1^2 = 51.8805$
Program 2	$\bar{y}_2 = 20.4314$	$s_2^2 = 95.1302$

(a) Based on the qqplots what can you conclude about the validity of the Gaussian model assumptions? Your answer should be in complete sentences.

(b) Determine a 99% confidence interval for $\mu_1 - \mu_2$.

(c) On the basis of the confidence interval you constructed in (b), what can you say about the $p - value$ associated with a test of the hypothesis $H_0 : \mu_1 = \mu_2$? (You do not need to do this test.) What is your conclusion regarding the hypothesis $H_0 : \mu_1 = \mu_2$?

(d) With reference to a suitable study population, what conclusions can the researchers draw from this study? Indicate any limitations to these conclusions.

(e) The same study was conducted at a different large university but with 100 athletes in Program 1 and 100 athletes in Program 2. The $p-value$ for testing $H_0 : \mu_1 = \mu_2$ was equal to 0.028 and a 95% confidence for $\mu_1 - \mu_2$ was $[1.1, 3.4]$. Explain in complete sentences the difference between a result which is statistically significant and a result which is of practical significance in the context of this larger study.

4. The term "white coat hypertension" is a name give to the phenomenon that occurs when a person's blood pressure is higher when it is taken in a medical setting than when it is taken at home.

To study this effect a doctor measured the systolic blood pressures in mm Hg of 26 patients at home and in the doctor's office. Systolic blood pressure measurements were taken at home and in the doctor's office both using the same home blood pressure monitor. Thirteen patients were randomized to take the first blood pressure reading at home while the remaining 13 took the first blood pressure reading at the doctor's office. The reading for the doctor's office (y_{1i}) and at home (y_{2i}) as well as the difference ($y_i = y_{1i} - y_{21}$) for each of the 26 patients are given below.

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13
y_{1i}	138	158	96	143	135	110	124	135	184	172	110	176	141
y_{21}	134	154	87	137	129	100	121	128	176	168	104	174	139
$y_i = y_{1i} - y_{21}$	4	4	9	6	6	10	3	7	8	4	6	2	2

Subject	14	15	16	17	18	19	20	21	22	23	24	25	26
y_{1i}	129	141	127	128	152	151	151	140	112	141	151	137	146
y_{21}	124	134	114	125	146	146	152	136	112	133	152	132	143
$y_i = y_{1i} - y_{21}$	5	7	13	3	6	5	-1	4	0	8	2	5	3

To analyze these data the model, $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, 26$ independently, is assumed where μ and σ are unknown parameters.

The measurements taken at the doctor's office were stored in the variable y1 and the measurements taken at home were stored in the variable y2 in R.

The following code was run in R:

```
t.test(y1,y2,mu=0,paired=TRUE,conf.level=0.95)
s<-sd(y1-y2)
cat("s = ", s)
```

The output obtained was:

```
Paired t-test
data:  y1 and y2
t = 8.3098, df = 25, p-value = 1.167e-08
alternative hypothesis:  true difference in means is not equal to 0
95 percent confidence interval:
3.789707 6.287217
sample estimates:
mean of the differences
5.038462
> s<-sd(y1-y2)
>
> cat("s = ", s)
s = 3.091676
```

(*a*) Explain clearly in a full sentence what the hypothesis $H_0 : \mu = 0$ means in the context of this study.

(*b*) Based on the R output on the previous page give a point estimate of μ and a 95% confidence interval for μ. Use all the decimal places in the output.

(*c*) Based on the R output on the previous page what is the $p - value$ for testing the hypothesis $H_0 : \mu = 0$. Clearly state your conclusion regarding the hypothesis $H_0 : \mu = 0$ in a sentence.

(*d*) Construct a 95% confidence interval for σ using the information from the R output.

(*e*) This experiment is a matched pairs experiment. Explain in full sentences why this type of design is better than a design in which 52 subjects are randomly divided into two groups of 26 with one group having their blood pressure taken in the doctor's office while the other group has their blood pressure taken at home.

(*f*) If a difference in mean systolic blood pressure was found to be statistically significant can you conclude that the difference is due to where the blood pressure is taken (doctor's office versus home)? Explain.

5. An instructor of a second year course in statistics used clickers in her lectures in the winter 2016 term. She is interested in the relationship between clicker grades and final grades in the course. Her data consist of (x_i, y_i), $i = 1, 2, \ldots, 82$ where x_i = clicker grade out of 5 and y_i = final grade out of 100. To analyze these data the simple linear regression model, $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, 82$ independently, is assumed where α, β and σ are unknown parameters and the $x_i's$ are assumed to be known constants.

Suppose the data set $x_1, x_2, \ldots, x_{82}$ are stored in the vector x and the data set $y_1, y_2, \ldots, y_{82}$ are stored in the vector y in R. The following code was run in R:

```
RegModel<-lm(y~x)
summary(RegModel)$coefficients
```

The output obtained was:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	47.185	3.343	14.115	< 2e-16
x	5.191	0.817	6.353	1.2e-08

(*a*) Answer the following questions based on this information. Use all the decimals given in the output.

The least squares estimate of β is_____.

The maximum likelihood estimate of α is_____.

The equation of the fitted least squares line is_____.

An estimate of the mean increase in final grade for a unit increase in clicker grade is _____.

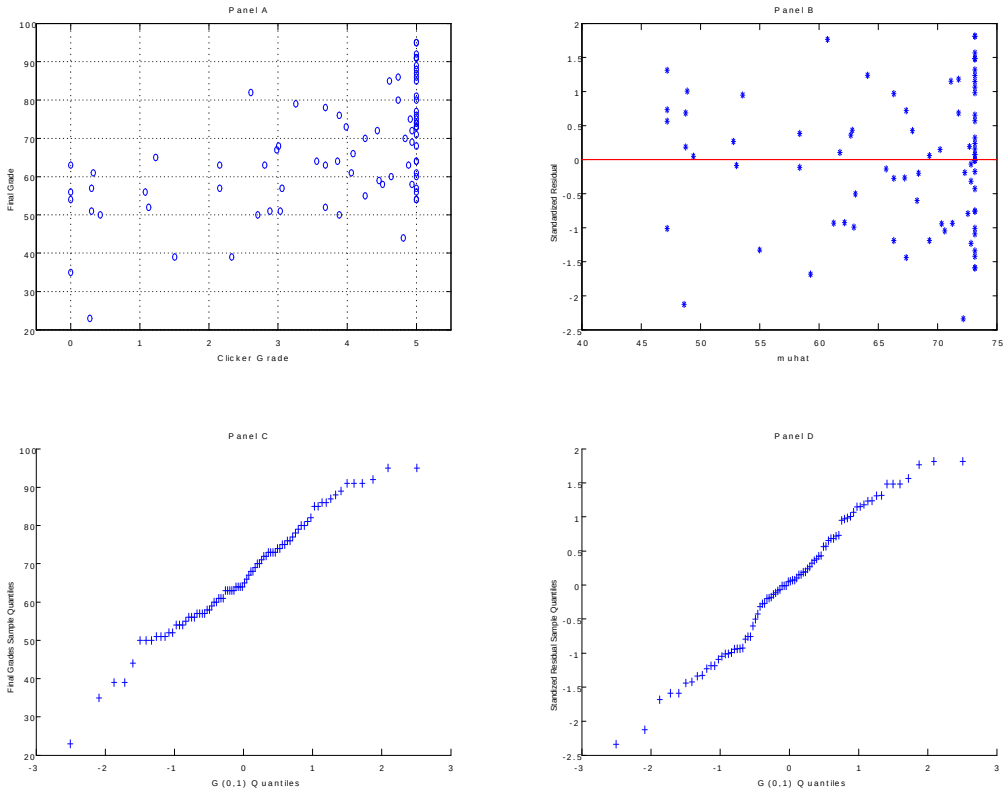
An estimate of the standard deviation of the maximum likelihood estimator $\tilde{\beta}$ is _____.

The value of the test statistic for testing $H_0 : \beta = 0$ is equal to _____.

The $p - value$ for testing $H_0 : \beta = 0$ is equal to _____.

State your conclusion with justification regarding the hypothesis $H_0 : \beta = 0$ in a sentence.

(*b*) Draw the fitted line on the scatterplot (Panel A below).



(*c*) Which of the plots in panels A-D of Figure 1 are relevant for drawing conclusions about the validity of the assumed regression model for these data? What conclusions can be drawn from these plots about the validity of the assumed model? Indicate clearly what you should see if the assumptions are valid. Your answer should be written in complete sentences.

(*d*) If evidence of a relationship is found can you conclude that a higher clicker grade results is the cause of a higher final grade?

(e) The following additional code was run:

```
xbar<-mean(x)
Sxx<-(length(x)-1)*var(x)
se<-summary(RegModel)$sigma
cat("xbar = ", xbar,", Sxx = ", Sxx, ", se = ", se)
```

The output obtained was:

```
xbar = 3.75405 , Sxx = 217.3591 , se = 12.0447
```

Based on this output and the previous R output given, determine a 95% prediction interval for the final grade for a student who has a clicker mark of 4.2. Show your work.

6. In November 2016 the Ipsos Market Research Company conducted a telephone survey of 1000 adults aged 18 and over in Canada. Participants were asked "Do you agree/disagree that finding holiday gifts that people will like is difficult?" Whether the adult was a Baby Boomer (born 1946-1964), Gen X'er (born 1965-77) or a Millennial (born after 1977) was also recorded. The results were:

Difficult / Generation	Millennial	Gen'X	Baby Boomers	Total
Agree	182	270	268	720
Disagree	88	80	112	280
Total	270	350	380	1000

(a) Is this an experimental or observational study? Explain.

(b) State the two variates of interest for this study and give their type.

(c) Use the likelihood ratio test statistic to test the hypothesis of no relationship (i.e. test for independence) between the two variates.

(d) For this study suggest, with reasons, a suitable target population and study population.

(e) Give a possible source of study error in relationship to your answer in (c).

(f) Let θ be the proportion of the study population who agree that finding holiday gifts that people will like is difficult. Give a point estimate of θ and an approximate 95% confidence interval for θ based on the observed data.

SOLUTIONS TO SAMPLE TESTS

Sample Midterm Test 1 Solutions

1. (a) E (b) B (c) C (d) D (e) B (f) A (g) B (h) B (i) D (j) B

2. (a) Since $t = 1$, the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} = \left(\prod_{i=1}^{n} \frac{1}{y_i!} \right) \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} \quad \text{for} \quad \theta \geq 0$$

or more simply (ignoring constants with respect to θ)

$$L(\theta) = \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} = \theta^{n\bar{y}} e^{-n\theta} \quad \text{for } \theta > 0 \text{ since } n\bar{y} = \sum_{i=1}^{n} y_i$$

The log likelihood function is

$$l(\theta) = n\bar{y} \log \theta - n\theta \quad \text{for} \quad \theta > 0$$

Solving

$$\frac{d}{d\theta} l(\theta) = \frac{n\bar{y}}{\theta} - n = \frac{n\bar{y} - n\theta}{\theta} = \frac{n(\bar{y} - \theta)}{\theta} = 0$$

gives the maximum likelihood estimate $\hat{\theta} = \bar{y}$.

(b)

	0	1	2	3	4	≥ 5	Total
Observed Frequency	28	45	56	40	21	10	200
Expected Frequency	24.491	51.432	54.003	37.802	19.846	12.425	200

$$e_j = 200 \cdot \frac{(2.1)^j}{j!} e^{-2.1} \quad \text{for } j = 0, 1, 2, 3, 4, 5$$

The agreement between the observed and expected frequencies seems quite good. The model appears to fit the data well.

(c) By the Invariance Property of maximum likelihood estimates, the maximum likelihood estimate of the probability that during a $t = 2$ minute interval there are no transactions is

$$\frac{\left(2\hat{\theta}\right)^0}{0!}e^{-2\hat{\theta}} = e^{-2(2.1)} = e^{-4.2} = 0.015$$

3. (a) The likelihood function is

$$\begin{aligned} L(\sigma) &= \prod_{i=1}^{n} f(y_i;\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-y_i^2/\left(2\sigma^2\right)} \quad \text{for } \sigma > 0 \\ &= (2\pi)^{-n/2} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right) \end{aligned}$$

or more simply

$$L(\sigma) = \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right) \quad \text{for } \sigma > 0$$

The log likelihood is

$$l(\sigma) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 \quad \text{for } \sigma > 0$$

Solving

$$\frac{d}{d\sigma}l(\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n} y_i^2 = \frac{1}{\sigma^3}\left(-n\sigma^2 + \sum_{i=1}^{n} y_i^2\right) = 0$$

gives the maximum likelihood estimate

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2}$$

(b)

$$\begin{aligned} R(\sigma) &= \frac{L(\sigma)}{L(\hat{\sigma})} = \frac{\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right)}{\hat{\sigma}^{-n}\exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n} y_i^2\right)} \quad \text{for } \sigma > 0 \\ &= \left(\frac{\hat{\sigma}}{\sigma}\right)^n \frac{\exp\left(-\frac{1}{2\sigma^2}n\hat{\sigma}^2\right)}{\exp\left(-\frac{n}{2}\right)} \quad \text{since } n\hat{\sigma}^2 = \sum_{i=1}^{n} y_i^2 \\ &= \left(\frac{\hat{\sigma}}{\sigma}\right)^n \exp\left\{\frac{n}{2}\left[1 - \frac{(\hat{\sigma})^2}{\sigma^2}\right]\right\} \quad \text{for } \sigma > 0 \end{aligned}$$

as required.

(c) By the Invariance Property of maximum likelihood estimates, the maximum likelihood estimate of $P(Y > 0.3; \sigma)$ is

$$\begin{aligned} P(Y > 0.3; \hat{\sigma}) &= P\left(Z > \frac{0.3 - 0}{1.2}\right) \quad \text{where } Z \sim G(0,1) \\ &= 1 - P(Z < 0.25) = 1 - 0.5987 = 0.4013 \approx 0.401 \end{aligned}$$

4. (a) The five-number summary for these data is:

$$\underline{\quad 42 \quad}, \ \underline{\quad 69.5 \quad}, \ \underline{\quad 79.5 \quad}, \ \underline{\quad 87 \quad}, \ \underline{\quad 99 \quad}$$

$$\begin{aligned} q(0.25) &= \frac{1}{2}\left(y_{(22)} + y_{(23)}\right) = \frac{1}{2}(69 + 70) = 69.5 \\ q(0.5) &= \frac{1}{2}\left(y_{(45)} + y_{(46)}\right) = \frac{1}{2}(79 + 80) = 79.9 \\ q(0.75) &= \frac{1}{2}\left(y_{(68)} + y_{(69)}\right) = \frac{1}{2}(87 + 87) = 87 \end{aligned}$$

(b) For these data:

$$\text{sample mean } = \bar{y} = \frac{6987}{90} = 77.633$$

and

$$\text{sample standard deviation} = s = \sqrt{\frac{1}{89}\left[555863 - \frac{(6987)^2}{90}\right]} = 12.288$$

(c) A

(d) The proportion of observations in the interval $[\bar{y} - s, \bar{y} + s] = [65.345,\ 89.921]$ is $60/90 = 0.667$.

If $Y \sim G(\mu, \sigma)$ then

$$\begin{aligned} P(Y \in [\mu - \sigma, \mu + \sigma]) &= P(|Y - \mu| \leq \sigma) = P\left(\frac{|Y - \mu|}{\sigma} \leq 1\right) \\ &= P(|Z| \leq 1) = 2P(Z \leq 1) - 1 \quad \text{where } Z \sim N(0,1) \\ &= 2(0.84134) - 1 = 0.68268 \\ &\approx 0.683 \end{aligned}$$

The proportion of observations in the interval (0.667) is slightly smaller than what would be expected for Gaussian data (0.683).

(e) For these data the

$$IQR = 87 - 69.5 = 17.5$$

To show that $IQR = 1.349\sigma$ for Gaussian data we need to solve

$$0.5 = P(|Y - \mu| \leq c\sigma) = P\left(\frac{|Y - \mu|}{\sigma} \leq c\right) \quad \text{for } c \text{ if } Y \sim G(\mu, \sigma)$$

From the $N(0,1)$ table $P(|Z| \leq c) = 2P(Z \leq c) - 1 = 0.5$ holds if $c = 0.6745$. Therefore

$$IQR = 2(0.6745)\sigma = 1.349\sigma$$

for Gaussian data.

(e) For Gaussian data we expect the relative frequency histogram to be approximately symmetric. The relative frequency histogram for these data is negatively skewed with a left tail.

For Gaussian data we expect the sample mean and sample median to be approximately equal. For these data the sample median = 79.5 > sample mean = 77.633.

For Gaussian data we expect the sample kurtosis to be close to 3. The sample kurtosis for these data equals 3.0211 which is quite close to 3.

For Gaussian data we expect the points to be scattered about a straight line with more variability about the line at both ends. The shape of this qqplot is very U-shaped.

The proportion of observations in the interval $[\bar{y} - s, \bar{y} + s]$ (0.667) is slightly smaller than we would expect for Gaussian data (0.683).

For Gaussian data we expect the IQR to be close in value to $1.349s = 1.349(12.288) = 16.58$. For these data $IQR = 17.5$ which is larger than expected.

Based on these observations, the Gaussian model is not the best model for these data.

The relative frequency histogram suggests that a model which is negatively skewed would be more appropriate for these data.

5. (a) This study would best be described as a sample survey since the population of interest (students in Grades 7 to 9 in New Brunswick, Nova Scotia, and Prince Edward Island) is finite. As well the purpose of the study was just to learn about this population and the researchers did not attempt to change or control any of the variates for the sampled units. It should be noted that, since the survey was voluntary, the sample would not be a representative sample.

(b) The Problem is to examine the relationship between participation in activities such as science fairs, competitions and engineering camps and the likelihood of considering careers in science, math and engineering among students in the early grades.

(c) This is a descriptive type Problem since the researchers wanted to know whether participating in activities such as science fairs, competitions and engineering camps is associated with whether girls would be more likely to consider a STEM career. The researchers were only observing variates to determine attributes of the study population.
Important Note: This is NOT a causative problem because the researchers were not in control of assigning any of the variates in this study.

(d) One important variate is whether or not the student participated in activities such as science fairs, competitions and engineering camps. This is a categorical variate.

The other important variate was whether or not the student would consider a STEM career. This is a also a categorical variate.

(e) A suitable target population for this study is the set of all students in Grades 7 to 9 in the provinces of New Brunswick, Nova Scotia and Prince Edward Island.

OR

A suitable target population for this study is the set of all students in Grades 7 to 9 in Canada.

(f) A suitable study population for this study is the set of all students in Grades 7 to 9 in the schools chosen by the researchers in the provinces of New Brunswick, Nova Scotia and Prince Edward Island. The schools are not specified in the article but it would have been impossible for the researchers to go to every school in these 3 provinces.

OR

A suitable study population for this study is the set of all students in Grades 7 to 9 in the provinces of New Brunswick, Nova Scotia and Prince Edward Island.

(g) If the target population is the set all students in Grades 7 to 9 in the provinces of New Brunswick, Nova Scotia and Prince Edward Island and the study population is the set of all students in Grades 7 to 9 in the schools chosen by the researchers in these provinces then a possible source of study error is that the students in the schools chosen by the researchers might be systematically different from the students in all schools. For example, it might be that the researchers only included schools in large cities and not schools in rural areas. Students in rural schools may have less access to science fairs, competitions and engineering camps.

OR

If the target population is the set all students in Grades 7 to 9 in Canada and the study population is the set of all students in Grades 7 to 9 in the provinces of New Brunswick, Nova Scotia and Prince Edward Island then a possible source of study error is that the students in the provinces of New Brunswick, Nova Scotia and Prince Edward Island might be systematically different from the students in Canada. For example, it might be that the schools in the provinces of New Brunswick, Nova Scotia and Prince Edward Island which have a much smaller population have less government funding for activities such as science fairs, competitions and engineering camps.

(h) The article indicates that the data were collected using an online survey and that the students completed the survey during school hours. No information is given about whether students were required to complete the survey or not.

(i) A possible source of sample error is that the survey was a voluntary survey. It could be that students who completed the survey are students who are generally more engaged in all activities and therefore might also be more likely to engage in other activities such as science fairs, competitions and engineering camps as compared to the students in the study population who did not volunteer to complete the survey.

(j) The numerical summary 2.7 is a relative risk. It is the relative risk among girls of considering a STEM career in the group who participate in activities such as science fairs, competitions and engineering camps as compared to those who do not participate.

Sample Midterm Test 2 Solutions

1. (a) The respondents to the survey are students who heard about the online referendum and then decided to vote in the referendum. These students may not be representative of all students at the University of Waterloo. For example, it is possible that the students who took the time to vote are also the students who most want the to remove the WPIRG fee. Students who don't care about the WPIRG fee probably did not bother to vote. This is an example of sample error. Any online survey such as this **online referendum has the disadvantage that the sample of people who choose to vote are not necessarily a representative sample of the study population of interest**. The **advantage of online surveys is that they are inexpensive and easy to conduct.** To obtain a representative sample you would need to select a random sample of all students at the University of Waterloo. Unfortunately taking such a sample would be much more time consuming and costly than conducting an online referendum.

(b) The parameter θ represents the proportion of the $31,380$ eligible undergraduate voters (the study population) who support option (2).

Important Note: The parameters in the model are always related to attributes of interest in the **study population** not in the sample.

There are two possible outcomes (Yes or No) on each trial (student) which is consistent with a Binomial model. A Binomial model also assumes independent trials. This assumption may not be a valid. For example, if groups of students, say within a specific faculty, got together and decided how to vote, their responses would not be independent events.

Since a student may not vote more than once, the sample of 8788 students is actually drawn without replacement from the finite population of $31,380$ students. If the sample was drawn at random (it was not) then we could justify the Binomial model using the Binomial approximation to the Hypergeometric.

(c) The maximum likelihood estimate of θ based on the observed data is 0.814.

$$\hat{\theta} = \frac{7156}{8788} = 0.814292$$

(d) To test $H_0 : \theta = 0.8$ we use the test statistic

$$D = |Y - n\theta_0| = |Y - (8788)(0.8)| = |Y - 7030.4|$$

with observed value

$$d = |7156 - 7030.4| = 125.6$$

and

$$\begin{aligned} p-value &= P\left(D \geq d; H_0\right) \\ &= P\left(|Y-7030.4| \geq 125.6\right) \quad \text{if } Y \sim Binomial\left(8788, 0.8\right) \\ &\approx P\left(|Z| \geq \frac{125.6}{\sqrt{8788\left(0.8\right)\left(1-0.8\right)}}\right) \quad \text{where } Z \sim G\left(0,1\right) \\ &= 2\left[1-P\left(Z \leq 3.35\right)\right] \\ &= 2\left(1-0.9996\right) \\ &= 0.0008 \end{aligned}$$

The $p-value$ for testing the hypothesis $H_0 : \theta = 0.8$ is approximately 0.001.

(e) Since the approximate $p-value$ for testing $H_0 : \theta = 0.8$ is less than 0.001 we would conclude that, based on the data, there is very strong evidence against the hypothesis $H_0 : \theta = 0.8$.

(f) Since the approximate $p-value$ for testing $H_0 : \theta = 0.8$ is less than 0.001 which is less than 0.05 then we know that the value $\theta = 0.8$ is not inside an approximate 95% confidence interval.

2. (a) Since the relative frequency histogram looks reasonably bell-shaped and the points in the qqplot lie reasonably along a straight line, the Gaussian model seems reasonable for these data.

(b) The parameter μ represents the mean number of hours of full sunlight in a day on the homeowner's roof over a year which is the study population.

Important Note: The parameters in the model are always related to attributes of interest in the **study population** not in the sample.

(c) For these data the maximum likelihood estimate of μ is

$$\frac{255.28}{61} = 4.18492 = 4.185$$

and the maximum likelihood estimate of σ is

$$\hat{\sigma} = \left[\frac{1}{61}\left(71.5607\right)\right]^{1/2} = 1.0831095 = 1.083$$

(d) Note that

$$s = \left[\frac{1}{60}\left(71.5607\right)\right]^{1/2} = 1.092098$$

Since $P(T \leq 2.6603) = 0.995$ where $T \sim t(60)$ a 99% confidence interval for μ is given by

$$\begin{aligned} & \bar{y} \pm 2.6603 s/\sqrt{61} \\ = & \frac{255.28}{61} \pm 2.6603\,(1.092098)/\sqrt{61} \\ = & 4.18492 \pm 0.371987 \\ = & [3.812931, 4.556905] \\ = & [3.813,\ 4.557] \end{aligned}$$

(e) The point estimate of μ is $\hat{\mu} = \bar{y} = 4.185$ which is greater than 4. However the 99% confidence interval for μ is $[3.813, 4.557]$ which contains values less than 4. Therefore based on the data there are values of μ which are less than 4 which are reasonable in light of the observed data. Since values of μ which are less than 4 are reasonable in light of the observed data and since the Solar Energy Association recommends that the average number of hours of full sunlight in a day should be at least 4, the landowner should conclude that there is not enough evidence to suggest placing solar panels on her roof.

Note also that she only took observations in two particular months (June and July) which may be the months with most sunlight. It would be a better idea to take measurements over the different months of the year in order to make an informed decision about whether solar panels are worthwhile.

(f) We use the test statistic

$$U = \frac{(n-1)\,S^2}{\sigma_0^2} = \frac{60S^2}{(1)^2} = 60S^2 \sim \chi^2\,(60) \quad \text{if} \quad H_0 : \sigma = 1 \quad \text{is true}$$

The observed value is $u = 60s^2 = 71.5607$.

$$p - value = 2P\,(U \geq 71.5607) \quad \text{where } U \sim \chi^2\,(60)$$

From the Chi-squared table

$$P\,(U \geq 59.335) = 1 - 0.5 = 0.5 \quad \text{and} \quad P\,(U \geq 74.397) = 1 - 0.9 = 0.1$$

Therefore $0.2 < p - value < 1$.

3. (a) For $w \geq 0$,

$$\begin{aligned} G\,(w) &= P\,(W \leq w) = P\left(\frac{2Y}{\theta} \leq w\right) = P\left(Y \leq \frac{\theta w}{2}\right) \\ &= F\left(\frac{\theta w}{2}\right) \quad \text{where } F\,(y) = P\,(Y \leq y) \text{ is the c.d.f. of } Y \end{aligned}$$

Therefore

$$\begin{aligned} g\,(w) &= G'\,(w) = f\left(\frac{\theta w}{2}\right)\frac{d}{dw}\left(\frac{\theta w}{2}\right) = \frac{1}{\theta}e^{-\left(\frac{\theta w}{2}\right)/\theta}\left(\frac{\theta}{2}\right) \\ &= \frac{1}{2}e^{-w/2} \quad \text{for } w \geq 0 \end{aligned}$$

as required.

(b) Since the sum of independent Chi-squared random variables has a Chi-squared distribution
with degrees of freedom equal to the sum of the degrees of freedom of the Chi-squared random variables in the sum, and since $\frac{2Y_i}{\theta} \sim \chi^2(2)$, $i = 1, 2, \ldots, n$ therefore

$$U = \sum_{i=1}^{n} \frac{2Y_i}{\theta} \sim \chi^2\left(\sum_{i=1}^{n} 2\right) \quad \text{or} \quad \chi^2(2n) \quad \text{as required}$$

(c) Using the Chi-squared table find a and b such that $P(U \leq a) = \frac{1-p}{2}$ and $P(U \leq b) = \frac{1+p}{2}$ where $U \sim \chi^2(2n)$.

Since

$$\begin{aligned} p &= P\left(a \leq \sum_{i=1}^{n} \frac{2Y_i}{\theta} \leq b\right) = P\left(\frac{1}{b} \leq \frac{\theta}{2\sum_{i=1}^{n} Y_i} \leq \frac{1}{a}\right) \\ &= P\left(\frac{2\sum_{i=1}^{n} Y_i}{b} \leq \theta \leq \frac{2\sum_{i=1}^{n} Y_i}{a}\right) \end{aligned}$$

then a $100p\%$ confidence interval for θ is given by

$$\left[\frac{2\sum_{i=1}^{n} y_i}{b}, \frac{2\sum_{i=1}^{n} y_i}{a}\right]$$

(d) If $W \sim \chi^2(20)$ then $P(W \leq 9.591) = 0.025 = P(W \geq 34.170)$ so $a = 9.591$ and $b = 34.170$.

(e)

(i) A 95% confidence interval for θ based on the pivotal quantity U is $[3.652, 13.012]$.

$$\begin{aligned} \left[\frac{2\sum_{i=1}^{n} y_i}{b}, \frac{2\sum_{i=1}^{n} y_i}{a}\right] &= \left[\frac{2(62.4)}{34.170}, \frac{2(62.4)}{9.591}\right] \\ &= [3.65232,\ 13.01220] \end{aligned}$$

(ii) An approximate 95% confidence interval for θ based on the asymptotic Normal pivotal quantity is $[2.372, 10.108]$.

$$\begin{aligned} \bar{y} \pm 1.96\frac{\bar{y}}{\sqrt{n}} &= \frac{62.4}{10} \pm 1.96\frac{(62.4/10)}{\sqrt{10}} \\ &= 6.24 \pm 3.867592 \\ &= [2.372408,\ 10.107592] \end{aligned}$$

The intervals are quite different which is what you would expect since the result in (i) is exact while the result in (ii) is based on an approximation which is poor since $n = 10$ is small.

4. (a) C (b) D (c) B (d) D (e) C (f) C (g) A (h) E (i) B (j) B

Sample Final Exam Solutions

1. (a)

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{50} y_i \theta^{y_i-1}(1-\theta)^2 = \left(\prod_{i=1}^{50} y_i\right) \theta^{\sum_{i=1}^{50} y_i - 50} (1-\theta)^{100} \\ &= \left(\prod_{i=1}^{50} y_i\right) \theta^{100-50} (1-\theta)^{100} = \left(\prod_{i=1}^{50} y_i\right) \theta^{50} (1-\theta)^{100} \quad \text{for } 0 \leq \theta < 1 \end{aligned}$$

or more simply

$$L(\theta) = \theta^{50}(1-\theta)^{100} \quad \text{for } 0 \leq \theta < 1$$

The log likelihood function is

$$l(\theta) = 50 \log \theta + 100 \log(1-\theta) \quad \text{for } 0 < \theta < 1$$

Since

$$\frac{d}{d\theta} l(\theta) = \frac{50}{\theta} - \frac{100}{1-\theta} = \frac{50 - 50\theta - 100\theta}{\theta(1-\theta)} = \frac{50 - 150\theta}{\theta(1-\theta)} = 0$$

if

$$\theta = \frac{50}{150} = \frac{1}{3} = 0.333$$

therefore the maximum likelihood estimate for θ is

$$\hat{\theta} = \frac{50}{150} = \frac{1}{3} = 0.333$$

(b)

Number of Children	1	2	3	≥ 4	Total
Observed Frequency	21	15	8	6	50
Expected Frequency	22.222	14.815	7.407	5.5556	50

$$\begin{aligned} e_1 &= 50\left(\frac{2}{3}\right)^2 = 22.222222 \\ e_2 &= 50(2)\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right) = 14.8148148 \\ e_3 &= 50(3)\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^2 = 7.4074074 \\ \text{or } e_3 &= 50 - (22.222222 + 14.8148148 + 5.5556) = 7.4073632 \end{aligned}$$

(c) The observed value of the likelihood ratio statistic is

$$\begin{aligned}\lambda &= 2\left[21\log\left(\frac{21}{22.2222}\right)+15\log\left(\frac{15}{14.8148148}\right)+8\log\left(\frac{8}{7.4074074}\right)+6\log\left(\frac{6}{5.5556}\right)\right]\\ &= 0.1516 = 0.152\end{aligned}$$

(Remember log = ln.) The degrees of freedom are $4-1-1=2$.

$$\begin{aligned}p-value &\approx P\left(W\geq 0.1516\right) \quad \text{where } W\sim\chi^2\left(2\right)=Exponential\left(2\right)\\ &= e^{-0.1516/2}\\ &= 0.927\end{aligned}$$

Therefore based on these data there is no evidence against the hypothesis that model (1) is a suitable model for these data.

2. (a)

(i) sample median = 0.245

(ii) IQR = 0.36

(iii) sample skewness is positive / negative

(b)

$$L\left(\theta\right)=\prod_{i=1}^{50}\frac{\theta}{\left(1+y_i\right)^{\theta+1}}=\theta^{50}\left[\prod_{i=1}^{50}\left(1+y_i\right)\right]^{-\theta-1} \quad \text{for } \theta>0$$

or more simply

$$L\left(\theta\right)=\theta^{50}\left[\prod_{i=1}^{50}\left(1+y_i\right)\right]^{-\theta} \quad \text{for } \theta>0$$

The log likelihood function is

$$\begin{aligned}l\left(\theta\right) &= 50\log\theta-\theta\log\left[\prod_{i=1}^{50}\left(1+y_i\right)\right]\\ &= 50\log\theta-\theta\sum_{i=1}^{50}\log\left(1+y_i\right) \quad \text{for } \theta>0\end{aligned}$$

Since

$$\frac{d}{d\theta}l\left(\theta\right)=\frac{50}{\theta}-\sum_{i=1}^{50}\log\left(1+y_i\right)=\frac{1}{\theta}\left(50-\theta\sum_{i=1}^{50}\log\left(1+y_i\right)\right)=0$$

if

$$\theta=\frac{50}{\sum_{i=1}^{50}\log\left(1+y_i\right)}=\frac{50}{\log\left[\prod_{i=1}^{50}\left(1+y_i\right)\right]}=\frac{50}{14.960}=3.342$$

Therefore the maximum likelihood estimate of θ based on these data is

$$\hat{\theta}=3.342$$

(c) Superimpose a graph of the probability density function of the assumed model on a relative frequency histogram for the observed data and see how well they agree.

Superimpose a graph of the cumulative distribution function of the assumed model on the graph of the empirical cumulative distribution function for the observed data and see how well they agree.

(d) An approximate 95% confidence interval for θ is [2.50 , 4.35]

(e) The observed value of the likelihood ratio statistic is

$$-2\log R(4.3) = -2\log(0.1774) = 3.458696 = 3.459$$

and

$$\begin{aligned} p-value &\approx P(W \geq 3.459) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{3.459}\right)\right] \quad \text{where } Z \sim G(0,1) \\ &= 2[1 - P(Z \leq 1.860)] \\ &= 2(1 - 0.96856) \\ &= 0.0629 \end{aligned}$$

Since $p-value \approx 0.0629$ there is some (weak) evidence against $H_0 : \theta = 4.3$ based on these data.

3. (a) For both Programs 1 and 2 the points lie reasonably along a straight line with more variability at both ends of the line which is what we expect for Gaussian data. These qqplots suggest the Gaussian assumption is reasonable for both groups.

(b)

$$s_p^2 = \frac{40(51.8805) + 50(95.1302)}{90} = 75.9081, \qquad s_p = 8.7125$$

$$P(T \leq 2.6316) = \frac{1 + 0.99}{2} = 0.995 \quad \text{where } T \sim t(90)$$

A 99% confidence interval for $\mu_1 - \mu_2$ is

$$\begin{aligned} & 27.3415 - 20.4314 \pm (2.6316)(8.7125)\sqrt{\frac{1}{41} + \frac{1}{51}} \\ = \; & 6.9101 \pm 4.8093 \\ = \; & [2.1009, 11.7193] \\ = \; & [2.101, 11.719] \end{aligned}$$

(c) The 99% confidence interval for $\mu_1 - \mu_2$ does not contain the value zero.

Therefore the $p-value$ for testing $H_0 : \mu_1 = \mu_2$ will be less than 0.01.

Since the $p-value < 0.01$, therefore there is strong evidence against the hypothesis $H_0 : \mu_1 = \mu_2$ based on the observed data.

(d) A reasonable study population consists of the population of athletes at the large university who attended the university sports injury clinic for hamstring injuries during the time of the study.

Since the 99% confidence interval for $\mu_1 - \mu_2$ does not contain the value zero, the data suggest that there is a difference in the mean number of days until return to sports activity for the two different exercise programs. These conclusions only apply to the study population. It would not be correct to make a statement about the mean difference in days until return to sports activity between the two exercise programs at another university.

A drawback of the study is that the study was only done at one university and only involved one sports injury clinic. The results of this study suggest conducting similar studies at other universities to see if Program 2 does in fact reduce the mean number of days to return to sports activity at other universities.

(e) In this larger study the difference in mean number of days until return to sports activity was found to be statistically significant. However a difference of 1 to 3 days until return to sports activity might not be large enough to be of practical significance for athletes at a university unless of course they are varsity athletes who are trying to return from injury to play in an important season game.

4. (a) The hypothesis $H_0 : \mu = 0$ means that there is no difference in mean blood pressure measurements between the doctor's office and at home.

(b) A point estimate of μ is 5.038462 and a 95% confidence interval for μ is $[3.789707, 6.287217]$.

(c) The $p - value$ for testing the hypothesis $H_0 : \mu = 0$ is $1.167e - 08$ or approximately zero. Since $p - value \approx 0$ there is very strong evidence against the hypothesis $H_0 : \mu = 0$ based on these data.

(d) From the Chi-squared table

$$P(W \leq 13.12) = 0.025 \quad \text{and} \quad P(W \leq 40.646) = 0.975$$

where $W \sim \chi^2(25)$.

A 95% confidence interval for σ is

$$\begin{aligned} & \left[\frac{s\sqrt{25}}{\sqrt{40.646}}, \frac{s\sqrt{25}}{\sqrt{13.12}} \right] \\ = & \left[\frac{(3.091676)(5)}{\sqrt{40.646}}, \frac{(3.091676)(5)}{\sqrt{13.12}} \right] \\ = & [2.42468, 4.26773] \\ = & [2.425, 4.268] \end{aligned}$$

(e) For both designs $\bar{Y}_1 - \bar{Y}_2$ is a point estimator of the mean difference $\mu_1 - \mu_2$.

For two independent samples the variance of this estimator is

$$Var\left(\bar{Y}_1 - \bar{Y}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

and for the paired experiment the variance of the estimator is

$$Var\left(\bar{Y}_1 - \bar{Y}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2Cov\left(Y_{1i}, Y_{2i}\right)$$

A sample of dependent pairs (Y_{1i}, Y_{2i}) is better than two independent random samples for estimating $\mu_1 - \mu_2$ since the difference $\mu_1 - \mu_2$, can be estimated more accurately (smaller variance and shorter confidence intervals) if $Cov\left(Y_{1i}, Y_{2i}\right) > 0$.

In this example, an analysis of the differences allows for a more precise comparison since differences between the 26 subjects have been eliminated, that is, by analyzing the differences we do not need to worry that there may have been large differences in blood pressure between the 26 subjects due to other variates such as sex, age, smoker/non-smoker, etc.

(f) Since this is a experimental study in which the researcher controlled where the blood pressure measurement was taken and randomized the order of the place (doctor's office or at home), we are able to conclude that a statistically significant difference in mean blood pressure is due to where the measurement was taken.

5. (a) [8]

The least squares estimate of β is 5.191.

The maximum likelihood estimate of α is 47.185.

The equation of the fitted least squares line is $y = 47.185 + 5.191x$.

An estimate of the mean increase in final grade for a unit increase in clicker grade is 5.191.

An estimate of the standard deviation of the maximum likelihood estimator $\tilde{\beta}$ is 0.817.

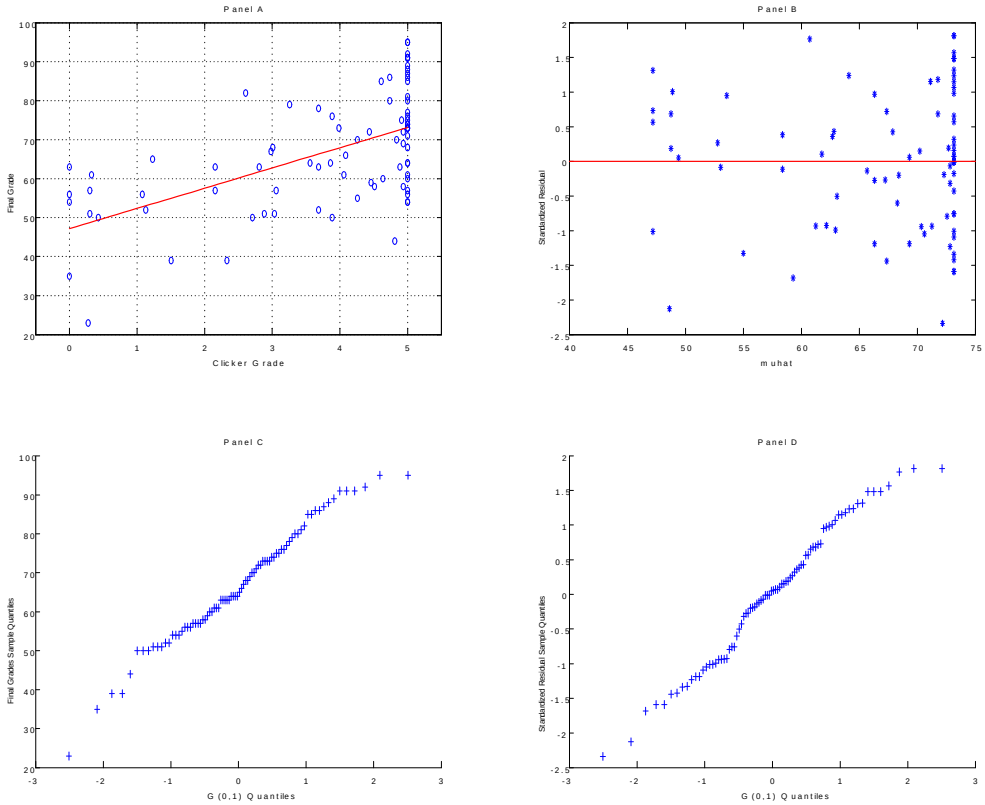
The value of the test statistic for testing $H_0 : \beta = 0$ is equal to 6.353.

The $p - value$ for testing $H_0 : \beta = 0$ is equal to $1.2e-08$ or $1.202e-08$.

State your conclusion with justification regarding the hypothesis $H_0 : \beta = 0$ in a sentence.

Since $p - value \approx 0$ there is very strong evidence against the hypothesis $H_0 : \beta = 0$ based on these data.

(b)



(c) The scatterplot in Panel A with the fitted line, the standardized residual plot in Panel B and the qqplot of the standardized residuals in Panel D are all relevant for drawing conclusions about the validity of the assumed regression model.

Panel C is **not relevant** for drawing conclusions about the validity of the assumed regression model.

(1) In Panel A we are looking to see if the observed points lie reasonably along the fitted lie which they do. Note that there is quite a bit of variability about the line.

(2) In Panel B we are looking to see if the points lie in roughly a horizontal band about the line $\hat{r}^* = 0$ which they do. No systematic patterns are observed.

(3) In Panel D we are looking to see if the points lie reasonably along a straight line which they do.

Based on these plots the model assumptions seem reasonable.

(d) This is an observational study and not an experimental study. The instructor is not in control of the clicker marks (the explanatory variate) and therefore a causal relationship cannot be concluded. In particular the instructor cannot conclude that a higher clicker grade results in a higher final grade. An alternative explanation is that a student's intelligence or their study habits cause both the higher clicker grade and the higher final grade.

(e) Based on this output and the previous R output given, determine a 95% prediction interval for the final grade for a student who has a clicker mark of 4.2. Show your work.

From the t table $P(T \leq 1.9901) = \frac{1+0.95}{2} = 0.975$ where $T \sim t(80)$.

The 95% prediction interval is

$$\begin{aligned} & 47.185 + 5.191\,(4.2) \pm 1.9901\,(12.0447)\sqrt{1 + \frac{1}{82} + \frac{(4.2 - 3.75405)^2}{217.3591}} \\ & 68.9872 \pm 24.1268 \\ = \; & [44.860,\ 93.114] \end{aligned}$$

6.

Difficult / Generation	Millennial	Gen'X	Baby Boomers	Total
	182	270	268	
Agree	[194.4]	[252]	[273.6]	720
Disagree	88	80	112	280
	[75.6]	[98]	[106.4]	
Total	270	350	380	1000

(a) This is an observational study since no variates were manipulated by the researchers.

(b) One variate is age category (Baby Boomer, Gen X'er or Millennial) which is a categorical variate.

The other variate is whether the person agrees or disagrees with the statement that finding holiday gifts that people will like is difficult. This is also a categorical variate.

(c) The expected frequencies under the hypothesis of independence are given in the table in square brackets.

$$e_{11} = \frac{720 \times 270}{1000} = 194.4, \quad e_{12} = \frac{720 \times 350}{1000} = 252$$

and the remaining frequencies can be determined by subtraction.

The observed value of the likelihood ratio statistic is

$$\begin{aligned}\lambda &= 2[182\log\left(\frac{182}{194.4}\right)+270\log\left(\frac{270}{252}\right)+268\log\left(\frac{268}{273.6}\right)\\ &\quad +88\log\left(\frac{88}{75.6}\right)+80\log\left(\frac{80}{98}\right)+112\log\left(\frac{112}{106.4}\right)]\\ &= 7.930\end{aligned}$$

$$\begin{aligned}p-value &\approx P\left(W\geq 7.930\right) \quad \text{where } W\sim\chi^2\left(2\right)\\ &= e^{-7.930/2}=0.019\end{aligned}$$

Since $p-value\approx 0.019$ there is evidence against the hypothesis of no relationship based on the data.

(d) A suitable target population is Canadian adults 18 and over.

The study population is Canadian adults 18 and over with a telephone since only a telephone survey was conducted.

(e) Study error arises when the attributes in the study population differ from the attributes in the target population. In this study a possible source of study error is that only adults with telephones were in the study population. For example suppose only adults with land line telephones were contacted. People with land lines may be systematically different with respect to how difficult they think finding holiday gifts is.

(f) A point estimate of θ is

$$\hat{\theta}=\frac{720}{1000}=0.72$$

An approximate 95% confidence interval is

$$\begin{aligned}&0.72\pm 1.96\sqrt{\frac{\left(0.72\right)\left(0.28\right)}{1000}}\\ =\ &0.72\pm 0.027829\\ =\ &[0.692,0.748]\end{aligned}$$

DISTRIBUTIONS AND STATISTICAL TABLES

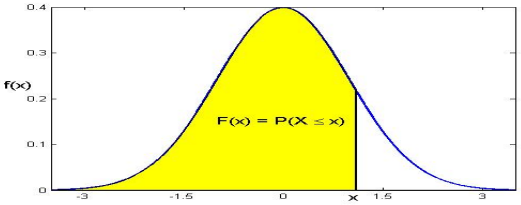
Summary of Discrete Distributions

Notation and Parameters	**Probability Function** $f(y)$	**Mean** $E(Y)$	**Variance** $Var(Y)$	**Moment Generating Function** $M(t)$
Discrete Uniform(a,b) $b \geq a$ a,b integers	$\frac{1}{b-a+1}$ $y = a, a+1, \ldots, b$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\frac{1}{b-a+1}\sum_{x=a}^{b} e^{tx}$ $t \in \Re$
Hypergeometric(N,r,n) $N = 1,2,\ldots$ $n = 0,1,\ldots,N$ $r = 0,1,\ldots,N$	$\frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$ $y = \max(0, n-N+r), \ldots, \min(r,n)$	$\frac{nr}{N}$	$\frac{nr}{N}(1-\frac{r}{N})\frac{N-n}{N-1}$	Not tractable
Binomial(n,p) $0 \leq p \leq 1,\ q = 1-p$ $n = 1,2,\ldots$	$\binom{n}{y}p^y q^{n-y}$ $y = 0,1,\ldots,n$	np	npq	$(pe^t + q)^n$ $t \in \Re$
Bernoulli(p) $0 \leq p \leq 1,\ q = 1-p$	$p^y q^{1-y}$ $y = 0,1$	p	pq	$pe^t + q$ $t \in \Re$
Negative Binomial(k,p) $0 < p \leq 1,\ q = 1-p$ $k = 1,2,\ldots$	$\binom{y+k-1}{y}p^k q^y$ $= \binom{-k}{y}p^k(-q)^y$ $y = 0,1,\ldots$	$\frac{kq}{p}$	$\frac{kq}{p^2}$	$\left(\frac{p}{1-qe^t}\right)^k$ $t < -\ln q$
Geometric(p) $0 < p \leq 1,\ q = 1-p$	pq^y $y = 0,1,\ldots$	$\frac{q}{p}$	$\frac{q}{p^2}$	$\frac{p}{1-qe^t}$ $t < -\ln q$
Poisson(μ) $\mu \geq 0$	$\frac{e^{-\mu}\mu^y}{y!}$ $y = 0,1,\ldots$	μ	μ	$e^{\mu(e^t-1)}$ $t \in \Re$
Multinomial$(n;p_1,p_2,\ldots,p_k)$ $0 \leq p_i \leq 1$ $i = 1,2,\ldots,k$ and $\sum_{i=1}^{k} p_i = 1$	$f(y_1,y_2,\ldots,y_k) = \frac{n!}{y_1!y_2!\cdots y_k!}p_1^{y_1}p_2^{y_2}\cdots p_k^{y_k}$ $y_i = 0,1,\ldots,n$ $i = 1,2,\ldots,k$ and $\sum_{i=1}^{k} y_i = n$	$E(Y_i) = np_i$ $i = 1,2,\ldots,k$	$Var(Y_i)$ $= np_i(1-p_i)$ $i = 1,2,\ldots,k$	$M(t_1,t_2,\ldots,t_k)$ $=(p_1e^{t_1}+p_2e^{t_2}+\cdots$ $+p_{k-1}e^{t_{k-1}}+p_k)^n$ $t_i \in \Re$ $i = 1,2,\ldots,k-1$

Summary of Continuous Distributions

Notation and Parameters	**Probability Density Function** $f(y)$	**Mean** $E(Y)$	**Variance** $Var(Y)$	**Moment Generating Function** $M(t)$
Uniform(a,b) $b > a$	$\frac{1}{b-a}$ $a \le y \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$ $t \neq 0$; 1 $t = 0$
Exponential(θ) $\theta > 0$	$\frac{1}{\theta}e^{-y/\theta}$ $y \ge 0$	θ	θ^2	$\frac{1}{1-\theta t}$ $t < \frac{1}{\theta}$
$\mathrm{N}(\mu,\sigma^2) = G(\mu,\sigma)$ $\mu \in \Re,\ \sigma^2 > 0$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(y-\mu)^2/(2\sigma^2)}$ $y \in \Re$	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$ $t \in \Re$
$\chi^2(k)$ $k = 1,2,\ldots$	$\frac{y^{(k/2)-1}e^{-y/2}}{2^{k/2}\Gamma(k/2)}$ $y > 0$ $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$	k	$2k$	$(1-2t)^{-k/2}$ $t < \frac{1}{2}$
$t(k)$ $k = 1,2,\ldots$	$\frac{c_k}{(1+\frac{y^2}{k})^{(k+1)/2}}$ $y \in \Re$ $c_k = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}$	0 if $k = 2,3,\ldots$ DNE if $k = 1$	$\frac{k}{k-2}$ if $k = 3,4,\ldots$ DNE if $k = 1,2$	DNE

N(0,1) Cumulative Distribution Function



This table gives values of $F(x) = P(X \le x)$ for $X \sim N(0,1)$ and $x \ge 0$

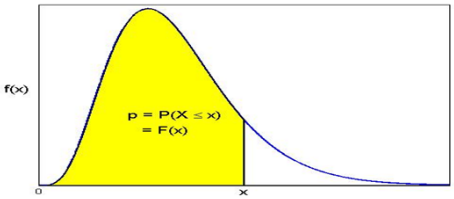
x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983

N(0,1) Quantiles:

This table gives values of $F^{-1}(p)$ for $p \ge 0.5$

p	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.075	0.08	0.09	0.095
0.5	0.0000	0.0251	0.0502	0.0753	0.1004	0.1257	0.1510	0.1764	0.1891	0.2019	0.2275	0.2404
0.6	0.2533	0.2793	0.3055	0.3319	0.3585	0.3853	0.4125	0.4399	0.4538	0.4677	0.4959	0.5101
0.7	0.5244	0.5534	0.5828	0.6128	0.6433	0.6745	0.7063	0.7388	0.7554	0.7722	0.8064	0.8239
0.8	0.8416	0.8779	0.9154	0.9542	0.9945	1.0364	1.0803	1.1264	1.1503	1.1750	1.2265	1.2536
0.9	1.2816	1.3408	1.4051	1.4758	1.5548	1.6449	1.7507	1.8808	1.9600	2.0537	2.3263	2.5758

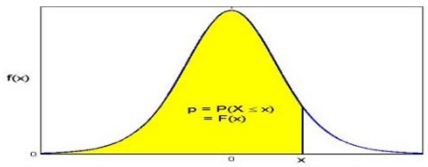
Chi-Squared Quantiles



This table gives values of x for p = P(X ≤ x) = F(x)

df\p	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.000	0.000	0.001	0.004	0.016	2.706	3.842	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.146	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.054	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.391	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.430	104.210
80	51.172	53.540	57.153	60.391	64.278	96.578	101.880	106.630	112.330	116.320
90	59.196	61.754	65.647	69.126	73.291	107.570	113.150	118.140	124.120	128.300
100	67.328	70.065	74.222	77.929	82.358	118.500	124.340	129.560	135.810	140.170

Student t Quantiles



This table gives values of x for $p = P(X \le x) = F(x)$, for $p \ge 0.6$

df \ p	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	0.3249	0.7265	1.3764	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088	636.6192
2	0.2887	0.6172	1.0607	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271	31.5991
3	0.2767	0.5844	0.9785	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145	12.9240
4	0.2707	0.5686	0.9410	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	0.2672	0.5594	0.9195	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
6	0.2648	0.5534	0.9057	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076	5.9588
7	0.2632	0.5491	0.8960	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853	5.4079
8	0.2619	0.5459	0.8889	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008	5.0413
9	0.2610	0.5435	0.8834	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968	4.7809
10	0.2602	0.5415	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869
11	0.2596	0.5399	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247	4.4370
12	0.2590	0.5386	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296	4.3178
13	0.2586	0.5375	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520	4.2208
14	0.2582	0.5366	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874	4.1405
15	0.2579	0.5357	0.8662	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328	4.0728
16	0.2576	0.5350	0.8647	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862	4.0150
17	0.2573	0.5344	0.8633	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458	3.9651
18	0.2571	0.5338	0.8620	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105	3.9216
19	0.2569	0.5333	0.8610	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794	3.8834
20	0.2567	0.5329	0.8600	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518	3.8495
21	0.2566	0.5325	0.8591	1.3232	1.7207	2.0796	2.5176	2.8314	3.5272	3.8193
22	0.2564	0.5321	0.8583	1.3212	1.7171	2.0739	2.5083	2.8188	3.5050	3.7921
23	0.2563	0.5317	0.8575	1.3195	1.7139	2.0687	2.4999	2.8073	3.4850	3.7676
24	0.2562	0.5314	0.8569	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668	3.7454
25	0.2561	0.5312	0.8562	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502	3.7251
26	0.2560	0.5309	0.8557	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350	3.7066
27	0.2559	0.5306	0.8551	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210	3.6896
28	0.2558	0.5304	0.8546	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082	3.6739
29	0.2557	0.5302	0.8542	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962	3.6594
30	0.2556	0.5300	0.8538	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460
40	0.2550	0.5286	0.8507	1.3031	1.6839	2.0211	2.4233	2.7045	3.3069	3.5510
50	0.2547	0.5278	0.8489	1.2987	1.6759	2.0086	2.4033	2.6778	3.2614	3.4960
60	0.2545	0.5272	0.8477	1.2958	1.6706	2.0003	2.3901	2.6603	3.2317	3.4602
70	0.2543	0.5268	0.8468	1.2938	1.6669	1.9944	2.3808	2.6479	3.2108	3.4350
80	0.2542	0.5265	0.8461	1.2922	1.6641	1.9901	2.3739	2.6387	3.1953	3.4163
90	0.2541	0.5263	0.8456	1.2910	1.6620	1.9867	2.3685	2.6316	3.1833	3.4019
100	0.2540	0.5261	0.8452	1.2901	1.6602	1.9840	2.3642	2.6259	3.1737	3.3905
>100	0.2535	0.5247	0.8423	1.2832	1.6479	1.9647	2.3338	2.5857	3.1066	3.3101