Extended Answer Section

Answer these questions in the answer book(s) provided.
Ask for extra books if you need them.

1. As part of a study into the effects of the use of oral contraceptive (OC) on blood pressure in women, independent random samples of 10 OC users and 14 non-users were selected from women who were aged 35–39, pre-menopausal and non-pregnant. The data below show their systolic blood pressure in mm Hg:

```
sort(users)
```

```
##  [1] 115.7 122.9 123.0 124.0 125.2 125.9 135.7 139.3 141.3 141.4
```

```
sort(non.users)
```

```
##  [1]  94.7 111.8 117.7 117.7 118.6 119.5 120.1 126.5 127.2 128.1 131.0
## [12] 132.7 132.8 138.7
```

```
fivenum(users)
```

```
## [1] 115.7 123.0 125.6 139.3 141.4
```

```
mean(users)
```

```
## [1] 129.4
```

```
mean(non.users)
```

```
## [1] 122.7
```

```
sp=sqrt((9*var(users)+13*var(non.users))/22)
sp
```

```
## [1] 10.29
```

(a) Obtain a five-number summary of the non-users sample.

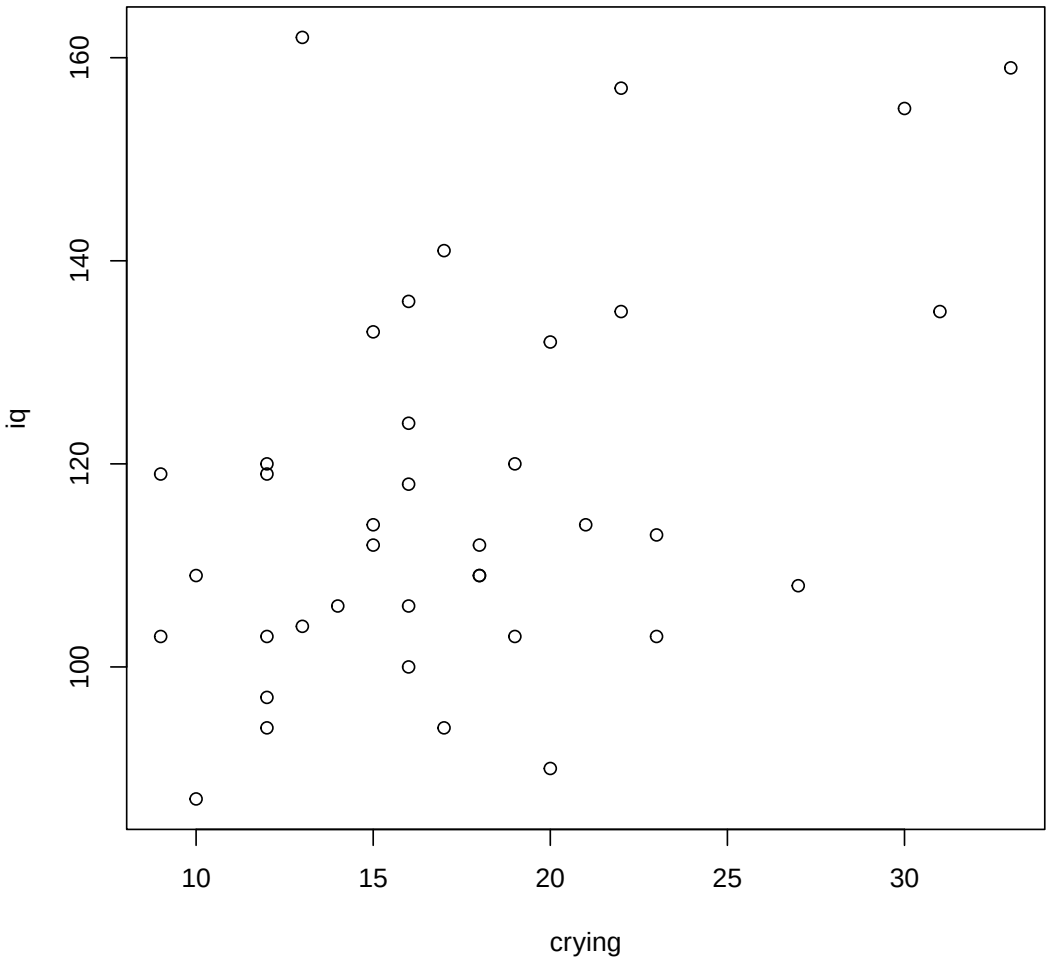
(b) Sketch two boxplots (one for each sample) on the same scale. **Note:** the only outlier is the value 94.7 in the non-users sample.

(c) Based on the boxplots, explain whether or not it seems reasonable to model these as samples from normal populations with the same variance.

(d) Construct a 95% confidence interval for the mean difference between the user and non-user populations, assuming these are normal with a common variance.

2. Babies who cry a lot may be more easily stimulated and this in turn may be an indicator of high IQ. A study involving 38 babies at around 4 days old were provoked into crying and had the intensity measured (number of peaks in the most active 20 seconds). The same babies were then given a standard IQ test at age 3. The data are in R vectors named respectively `crying` and `iq`. A scatterplot appears below.

```
plot(crying,iq)
```



Examine the R output on the following page and answer the questions that follow.

```
x=crying
y=iq
Sxx=sum((x-mean(x))^2)
Sxy=sum((x-mean(x))*(y-mean(y)))
Syy=sum((y-mean(y))^2)
Sxx
```

```
## [1] 1291
```

```
Sxy
```

```
## [1] 1927
```

```
Syy
```

```
## [1] 13901
```

(a) Using the R output above compute the correlation coefficient.

(b) Consider the linear regression model with normal errors where for $i = 1, 2, \ldots, n$, $Y_i \sim N(\alpha+\beta x_i, \sigma^2)$ are independent for known constants $x_1, x_2, \ldots, x_n$ and unknown $\sigma^2 > 0$, α and β.
The observed value of the t-statistic for testing $\beta = 0$ is of the form $t = b/\text{se}(b)$ where $b = S_{xy}/S_{xx}$, $\text{se}(b) = \hat{\sigma}/\sqrt{S_{xx}}$ and

$$\hat{\sigma}^2 = \frac{1}{n-2}\left[S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right].$$

Show that

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where r is the observed value of the correlation coefficient.

(c) Use the preceeding parts to compute upper and lower bounds (the tightest possible using the tables) for the p-value of a test $H_0\colon \beta = 0$ against the alternative $H_1\colon \beta > 0$ based on the crying data above (**hint:** under the model in (b) $(n-2)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-2}$).

3. (a) Suppose X is a random variable taking values 1,2,3,4,5,6 with equal probabilities (i.e. the roll of a fair six-sided die). Verify by direct calculation that

(i) $E(X) = 7/2$;

(ii) $Var(X) = 35/12$.

(b) Suppose now that a fair six-sided die is rolled 3 times independently.

(i) Suppose the faces of each of the dice are coloured, so that 1,3 and 5 are red, 2 and 4 are blue and 6 is green. What is the probability that exactly 1 of each colour is obtained?

(ii) Compute a normal approximation with continuity correction to the probability that the sum of the rolls is at most 6.

(iii) By carefully enumerating all relevant outcomes, determine the exact probability that the sum of the rolls is at most 6.

4. (a) Suppose a genetic theory predicts that two varieties of plant A and B should occur in the proportions 3:1 respectively. If a random sample of 36 such plants reveals 22 of type A and 14 of type B, determine upper and lower bounds (the tightest possible using the χ^2 table) for the p-value of the appropriate χ^2-test of the genetic theory.

(b) Suppose independent counts x and $n - x$ are observed in two categories which are hypothesised to have respective probabilities p and $1 - p$. The $100(1 - \alpha)\%$ Wilson confidence interval for p is defined as all p such that, with $\hat{p} = x/n$,

$$\frac{\sqrt{n}\,|\hat{p} - p|}{\sqrt{p(1-p)}} \leq z_{\alpha/2}$$

where $P(Z \geq z_{\alpha/2}) = \alpha/2$ and $Z \sim N(0, 1)$.
Prove that a value p_0 is in this interval *if and only if* the two-sided χ^2-test of $H_0 \colon p = p_0$ gives a p-value $\geq \alpha$ (**hint**: $P(Z^2 \geq c) = P(|Z| \geq \sqrt{c}) = 2P(Z \geq \sqrt{c})$).

End of Extended Answer Section