STATISTICS 1 – SOLUTIONS – JUNE 2008

A1. (i)
```
5 | 69
6 | 048
7 | 2
8 | 49
9 | 6
```

(ii) The ordered values are: $56, 59, 60, 64, 68, 72, 84, 89, 96$
The values used to construct the boxplot are the smallest obsn $o_1 = 56$, the lower hinge = median of {data values $\leq$ median} $= o_3 = 60$, the median $o_5 = 68$, the upper hinge = median of {data values $\geq$ median} $= o_7 = 84$, and the largest obsn $o_9 = 96$.

A2. (i) Let X denote the number of rounds required by a randomly chosen guest. $P(X = 0) = 0.2$; $P(X = 1) = 0.3$, $P(X = 2) = 0.5$ so $\mu_X = E(X) = 1.3$, $E(X^2) = 2.3$, $\sigma_X^2 = V(X) = 0.61$
There are $n = 100$ guests, so from the central limit theorem $S = \sum_1^n X_i \simeq N(n\mu_X, n\sigma_X^2) = N(130, 61)$.

(ii) The number of rounds will be enough if $S \leq 150$. Since S takes integer value but the approximating distribution is continuous, a continuity correction will improve the accuracy of the approximation. Let $T \sim N(130, 61)$, then $P(S \leq 150) \simeq P(T \leq 150 + 1/2) = P((T - 130)/\sqrt{(}61) \leq (150.5 - 130)/\sqrt{61} = P(Z \leq 20.5/\sqrt{61})$ [where $Z \sim N(0, 1)$] $= P(Z \leq 2.624756) = 0.9956644 \simeq 0.996$

A3. (i) Assume U and V are independent random variables with $U \sim N(0, 1)$ and $V \sim \chi_r^2$. Then $U/\sqrt{V/r}$ has the t distribution with r degrees of freedom, and we write $U/\sqrt{V/r} \sim t_r$.

(ii) V and W have moment generating functions $\mathcal{M}_V(t) = (1 - 2t)^{-r/2}$ and $\mathcal{M}_W(t) = (1 - 2t)^{-s/2}$. Since V and W are independent $\mathcal{M}_{V+W}(t) = \mathcal{M}_V(t)\mathcal{M}_W(t) = (1 - 2t)^{-r/2}(1 - 2t)^{-s/2} = (1 - 2t)^{-(r+s)/2}$ so $V + W \sim \chi_{r+s}^2$.

A4. (i) The end points of a $100(1-\alpha)\%$ confidence interval for μ are $c_L = \bar{x} - z_{\alpha/2}\sigma_0/\sqrt{n}$ and $c_U = \bar{x} + z_{\alpha/2}\sigma_0/\sqrt{n}$, where $z_{\alpha/2}$ is such that $P(Z > z_\alpha) = \alpha$ when $Z \sim N(0, 1)$.
Here $n = 25$, $\alpha/2 = 0.05$, $z_{0.05} = 1.645$, $\bar{x} = 1.92$, $\sigma_0 = 4$ so

$$c_L = 1.92 - 1.645 \times 4/\sqrt{25} = 1.92 - 1.316 = 0.604$$

$$c_U = 1.92 + 1.645 \times 4/\sqrt{25} = 1.92 + 1.316 = 3.236$$

(ii) (a) It would increase as $1/\sqrt{n}$ would increase – intuitively less observations = less accuracy = wider interval. (b) It would decrease as less confidence = α increase = $z_{\alpha/2}$ decreases as α – intuitively requiring less confidence = smaller interval.

A5. (i) Type I error occurs if we reject H_0 when it is in fact true; Type II error occurs if we accept H_0 when it is in fact false.

(ii) The significance level of the test is P(Type I error); the power is $1-$ P(Type II error).

B1. (15 marks)

(a) [SEEN IN NOTES]

For observations from an Exponential (θ) distribution,

$$\begin{aligned} f(x;\theta) &= \theta e^{-\theta x} \\ \log f(x;\theta) &= \log(\theta) - \theta x \\ \frac{\partial}{\partial\theta}\log f(x;\theta) &= \frac{1}{\theta} - x \\ \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i;\theta) &= \left(\frac{1}{\theta} - x_1\right) + \cdots + \left(\frac{1}{\theta} - x_n\right) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i \end{aligned}$$

so the maximum likelihood estimate $\hat{\theta}$ satisfies the likelihood equation

$$\frac{n}{\hat{\theta}} - \sum_{i=1}^{n} x_i = 0$$

i.e. $$\hat{\theta} = n/\textstyle\sum_1^n x_i$$

(b) [SEEN SIMILAR IN NOTES AND ON PROBLEM SHEETS]

- For two unknown parameters we use the equations involving the two smallest moments of the distribution. Here $\mathrm{E}(X;\alpha,\beta) = \alpha$, $\mathrm{E}(X^2;\alpha,\beta) = \alpha^2 + 2/\beta^2$. Let $m_1 = (x_1 + \cdots + x_n)/n$ and $m_2 = (x_1^2 + \cdots + x_n^2)/n$. Then the method of moments estimates satisfy

$$\mathrm{E}(X;\hat{\alpha},\hat{\beta}) = m_1 \quad \text{and} \quad \mathrm{E}(X^2;\hat{\alpha},\hat{\beta}) = m_2$$

i.e. $$\hat{\alpha} = m_1 \quad \text{and} \quad \hat{\alpha}^2 + 2/\hat{\beta}^2 = m_2$$

The first equation in gives $\hat{\alpha} = m_1$. Substituting this into the second equation gives $2/\hat{\beta}^2 = m_2 - m_1^2$b so $\hat{\beta} = 2/\sqrt{(}m_2 - m_1^2)$.

(c) [MET THESE IDEAS IN THE NOTES]

In both plots the median is above the true value of θ. Moreover both plots are heavily skewed towards the upper end of the range with a large number of upper outliers, so in both cases the mean is likely to be significantly larger than the median and thus even further away from the true value than the median. Thus both estimators have a significan positive bias, with the mom possibly more biased than the mle (at least the median appears further away).

In terms of spread, the IQR is smaller for the mle than the mom, as is the distance between the whiskers, and the outliers don't seem so far from the median. Overall, the mle is appears better in terms of bias and spread – and hence better in terms of mean square error – than the mom.

B2. (15 marks)

(a) [SEEN SIMILAR IN NOTES AND ON PROBLEM SHEETS]

Model assumptions: Let X_i denote the age of the ith unsuccessful entrant and let Y_j denote the age of the jth successful entrant. We assume $X_1, \ldots, X_{12}$ is a simple random sample from the $N(\mu_X, \sigma_X^2)$ distribution, and $Y_1, \ldots, Y_{10}$ is a simple random sample from the $N(\mu_Y, \sigma_Y^2)$ distribution, and that both samples are independent of each other, and that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

Hypotheses: H_0: $\mu_X = \mu_Y$ versus H_A: $\mu_X > \mu_Y$. H_0 corresponds to the mean age of unsuccessful candidates being the same as the mean age of successful candidates; H_1 corresponds to it being greater.

Test Statistic: Here $T = (\bar{X} - \bar{Y})/\sqrt{\dfrac{\hat{\sigma}_X^2}{n} + \dfrac{\hat{\sigma}_Y^2}{m}} = (\bar{X} - \bar{Y})/\hat{\sigma}\sqrt{1/n + 1/m}$ where T has the t distribution with $n + m - 2 = 20$ degrees of freedom when H_0 is true. For the given data,
$n = 12, m = 10, \bar{x} = 49.75, \bar{y} = 45.10, (\bar{x} - \bar{y}) = 4.65$;
$\sum(x_i - \bar{x})^2 = \sum x_i^2 - (\sum x_i)^2/n = 614.25, \sum(y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/m = 422.9$ so
$\hat{\sigma}^2 = [\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2]/(n + m - 2) = 51.8575 = (7.201215)^2$;
$\sqrt{1/n + 1/m} = 0.4281744$;
and the observed test statistic is $t_{obs} = 1.50809$.

Critical region: The critical region of values for which an α-level test would reject H_0 is of the form $C = \{T \geq c^*\}$, where c^* is defined by the condition: $\alpha = P(\text{Reject } H_0 | H_0 \text{ true}) = P(T \geq c^* \mid H_0 \text{ true}) = P(t_{20} \geq c^*)$. Thus, for $\alpha = 0.05$, $c^* = t_{20;0.05} = 1.724718$ giving $C = \{T \geq 1.724718\}$.

Conclusions: The observed test statistic value $t_{obs} = 1.508$ is not in the critical region of the 0.05-level test. Thus there is no strong evidence that would lead us to reject H_0 in favour of H_A, and we conclude that there is no strong evidence that the mean age of unsuccessful entrants is higher than that of successful entrants.

(b) [SEEN SIMILAR IN NOTES AND ON PROBLEM SHEETS]

***p*-value**: For H_A: $\mu_X > \mu_Y$, the values of T which are at least as extreme as t_{obs} are the set $\{T \geq t_{obs}\}$. Thus the p-value = $P(T \geq t_{obs} | H_0 \text{ true}) = P(t_{20} \geq 1.508) = 1 - P(t_{20} \leq 1.508)$. From tables $P(t_{20} \leq 1.5) = 0.9254$ and $P(t_{20} \leq 1.6) = 0.9374$. Thus the p-value $\simeq 1 - 0.9254 = 0.0746$. A more accurate linear interpolation gives $P(t_{20} \leq 1.508) = 0.9254 + 0.08 \times (0.012) = 0.92636$, giving a p-value of $0.07364 \simeq 0.0736$. (R gives 0.07358). This is greater than 0.05, consistent with not rejecting H_0 at the 5% level.

(c) [MET THIS TEST IN THE NOTES AND PROBLEM SHEETS]

If we do not assume $\sigma_X = \sigma_Y$, then an alternative would be to use the test due to Welch with test statistic $T = (\bar{X} - \bar{Y})/\sqrt{\dfrac{\hat{\sigma}_X^2}{n} + \dfrac{\hat{\sigma}_Y^2}{m}}$ where $\hat{\sigma}_X^2 = \sum(X_i - \bar{X})^2/(n-1)$ and $\hat{\sigma}_Y^2 + \sum(Y_j - \bar{Y})^2/(m-1)$. Under H_0, T again has the t distribution but the degrees of freedom ν are much more complicated to compute, since

$$\nu = \left[(S_X^2/n + S_Y^2/m)^2 \Big/ \left(\frac{(S_X^2/n)^2}{n-1} + \frac{(S_Y^2/m)^2}{m-1} \right) \right].$$

B3. (15 marks)

(a) [SEEN SIMILAR IN NOTES AND ON PROBLEM SHEETS]

Put $ss_{yy} = \sum y_i^2 - (\sum y_i)^2/12 = 40.70917$, $ss_{xx} = \sum x_i^2 - (\sum x_i)^2/12 = 485.7292$, $ss_{xy} = \sum y_i x_i - \sum x_i \sum y_i/12 = 122.0704$.
Then the least squares estimates are

$$\hat{\beta} = \frac{\sum y_i x_i - n\bar{y}\bar{x}}{\sum x_i^2 - n\bar{x}^2} = \frac{ss_{xy}}{ss_{xx}} = \frac{122.0704}{485.7292} = 0.2513215$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 5.458333 - 0.2513215 \times 18.80833 = 0.7313954$$

$$\hat{\sigma}^2 = \frac{ss_{yy} - ss_{xy}^2/ss_{xx}}{n-2} = \frac{40.70917 - (122.0704)^2/485.7292}{10} = 1.002931.$$

For $x = 22$ the predictor of Y is $\hat{y} = \hat{\alpha} + \hat{\beta}x = 0.7313954 + 0.2513215 \times 22 = 6.260428$

(b) [MET THESE IDEAS IN THE NOTES]

Under the model, the residuals should be independent observations from $\mathrm{N}(0, \sigma^2)$.

- Substantially more +ve than -ve residual values indicate the assumption $\mathrm{E}(e_i) = 0$ may be violated. Here 6 +ve and 6 -ve so no concern.
- Systematic variation in the size of the residuals indicates the assumption of constant variance may be violated. Some quite large values for larger x but also some smaller ones - so no real concern.
- Systematic variation in the sign of the residuals indicates the linear model assumption my be violated. No sign of systematic variation.
- Significant numbers of outliers in the residuals indicates the Normal distribution assumption may be violated. A couple of quite large residuals, but probably not large enough to cause concerns.

Ovewrall, no reason to believe the model is not an adequate fit to the data.

(c) [SEEN SIMILAR IN NOTES AND ON PROBLEM SHEETS]

$(\hat{\beta} - \beta)/\sqrt{\hat{\sigma}^2/ss_{xx}} \sim t_{n-2}$, so the endpoints of a $100(1-\alpha)\%$ confidence interval for β are given by

$$c_L = \hat{\beta} - t_{n-2,\alpha/2}\sqrt{\hat{\sigma}^2/ss_{xx}} \quad \text{and} \quad c_U = \hat{\beta} + t_{n-2,\alpha/2}\sqrt{\hat{\sigma}^2/ss_{xx}}$$

Here, $n = 12, \alpha = 0.10, t_{n-2,\alpha/2} = t_{10,0.05} = 1.8125, \hat{\beta} = 0.2513215, ss_{xx} = 485.7292, \hat{\sigma}^2 = 1.002931$ giving $c_L = 0.1689615\,; c_U = 0.3336815$.