MATH11400 Statistics 1 2010–11

Homepage http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/

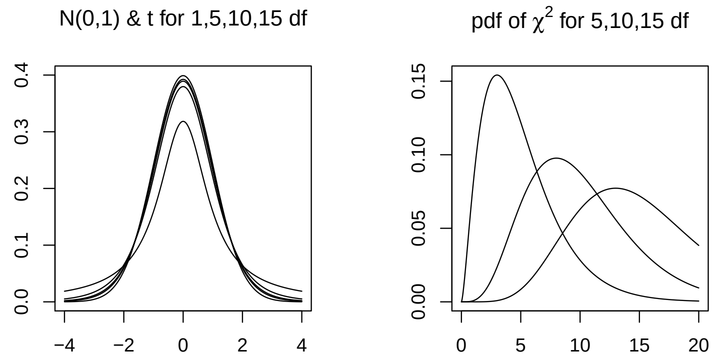
Solution Sheet 8

1. The plots are shown below. I plotted the probability density function for the $N(0,1)$ distribution over the range $(-4,4)$, and added the probability density function for the t_1 distribution, with the commands:

```
> range<-seq(-4,4,0.01)
> plot(range,dnorm(range), type="l", ylim=c(0,0.4))
> lines(range,dt(range,1))
```

Extra plots of the pdf from t_5, t_{10} and t_{15} can be added with additional commands of the form `lines(range,dt(range,***))`, replacing `***` by `5,10` and `15` as required. The χ^2 plots were done similarly, by setting an appropriate range and using the command `dchisq` with appropriate degrees of freedom.

To tell which pdf is which, note that the peak of the t_r distribution increases upwards towards that of the $N(0,1)$ with increasing r while the peak of the χ^2_r distribution moves right with increasing r.



2. (a) From your notes, $\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2 \sim \chi^2_{n-1} = \text{Gamma}((n-1)/2, 1/2)$, with mean $((n-1)/2)/(1/2) = (n-1)$ and variance $((n-1)/2)/(1/2)^2 = 2(n-1)$.

$$\begin{aligned}
\text{Now } S^2 &= \textstyle\sum_{i=1}^n (X_i - \overline{X})^2/(n-1) = (\sigma^2/(n-1)) \sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2 \\
\text{so } \mathrm{E}(S^2) &= (\sigma^2/(n-1))\,\mathrm{E}[\textstyle\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2] = (\sigma^2/(n-1))(n-1) = \sigma^2 \\
\text{and } \mathrm{Var}(S^2) &= (\sigma^2/(n-1))^2\,\mathrm{Var}[\textstyle\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2] = (\sigma^2/(n-1))^2 2(n-1) \\
&= 2\sigma^4/(n-1).
\end{aligned}$$

Similarly

$$\begin{aligned}
\hat{\sigma}^2_{mle} &= \textstyle\sum_{i=1}^n (X_i - \overline{X})^2/n = (\sigma^2/n) \sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2 \\
\text{so } \mathrm{E}(\hat{\sigma}^2_{mle}) &= (\sigma^2/n)\,\mathrm{E}[\textstyle\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2] = (\sigma^2/n)(n-1) = \sigma^2(n-1)/n \\
\text{and } \mathrm{Var}(\hat{\sigma}^2_{mle}) &= (\sigma^2/n))^2\,\mathrm{Var}[\textstyle\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2] = (\sigma^2/n)^2 2(n-1) \\
&= 2\sigma^4(n-1)/n^2.
\end{aligned}$$

(b) The bias of an estimator $\hat{\sigma}^2$ for σ^2 is defined as $\mathrm{E}(\hat{\sigma}^2 - \sigma^2)$. Thus the bias of S^2 is $\mathrm{E}(S^2 - \sigma^2) = \mathrm{E}(S^2) - \sigma^2 = \sigma^2 - \sigma^2 = 0$, so S^2 is unbiased as an estimator of σ^2. Similarly, the bias of $\hat{\sigma}^2_{mle}$ is $\mathrm{E}(\hat{\sigma}^2_{mle} - \sigma^2) = \mathrm{E}(\hat{\sigma}^2_{mle}) - \sigma^2 = \sigma^2(n-1)/n - \sigma^2 = -\sigma^2/n$, so $\hat{\sigma}^2_{mle}$ is a biased estimator of σ^2 with bias $-\sigma^2/n$.

The mean square error of an estimator $\hat{\sigma}^2$ for σ^2 is defined as $\mathrm{E}[(\hat{\sigma}^2 - \sigma^2)^2] = \mathrm{Var}(\hat{\sigma}^2) + [\mathrm{E}(\hat{\sigma}^2) - \sigma^2]^2 = \mathrm{Var}(\hat{\sigma}^2) + [\text{bias}(\hat{\sigma}^2)]^2$. Thus the mean square error of S^2 is $\mathrm{Var}(S^2) + [\text{bias}(S^2)]^2 = 2\sigma^4/(n-1) + 0 = 2\sigma^4/(n-1)$, and the mean square error of $\hat{\sigma}^2_{mle}$ is $\mathrm{Var}(\hat{\sigma}^2_{mle}) + [\text{bias}(\hat{\sigma}^2_{mle})]^2 = 2\sigma^4(n-1)/n^2 + [-\sigma^2/n]^2 = \sigma^4(2n-1)/n^2$.

Note that $(2n-1)/n^2 < 2/(n-1)$, so overall the maximum likelihood estimator of σ^2 has smaller mean square error than the sample variance, even though it is biased.

3. (a) X_i/θ and Y_i/θ are i.i.d. $\mathrm{N}(0,1)$, so $T_i/\theta_i^2 \sim \chi^2_2$ by 6.7 and 6.8. By 6.13, χ^2_2 is the same as $\mathrm{Gamma}(1, 1/2)$, which is in turn the same as $\mathrm{Exponential}(1/2)$. Therefore, $T_i \sim \mathrm{Exponential}(1/(2\theta^2))$.

Thus $t_1, t_2, \ldots, t_n$ are an observed random sample from this distribution, which has probability density function $f_T(t;\theta) = (1/2\theta^2)\exp(-t/2\theta^2)$ for $t > 0$. Then $(\partial/\partial\theta)\log f_T(t;\theta) = -2/\theta + t/\theta^3$. Thus the likelihood equation is

$$0 = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_T(t_i;\theta) = -2n/\theta + \frac{1}{\theta^3}\sum_{i=1}^{n} t_i$$

and the solution is $\widehat{\theta}_{\text{mle}} = \sqrt{\sum_{i=1}^{n} t_i/2n}$.

(b) Working directly in terms of x_i and y_i, both of these are drawn from $\mathrm{N}(0,\theta^2)$, with density function $f_X(x;\theta) = (1/\sqrt{2\pi\theta^2})\exp(-x^2/(2\theta^2))$. Now $(\partial/\partial\theta)\log f_X(x;\theta) = -1/\theta + x^2/\theta^3$. So now the likelihood equation is

$$0 = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\{\log f_X(x_i;\theta) + \log f_X(y_i;\theta)\} = -2n/\theta + \frac{1}{\theta^3}\sum_{i=1}^{n}(x_i^2 + y_i^2).$$

Since $t_i = x_i^2 + y_i^2$, this is the same equation, so has the same solution.

(c) From sheet 2, question 2, the method of moments estimator based on the x sample alone is $\widehat{\theta}_{\text{mom}} = \sqrt{\sum_i x_i^2/n}$. Working the question through again, with both x and y data simultaneously, gives $\widehat{\theta}_{\text{mom}} = \sqrt{\sum_{i=1}^{n} t_i/2n}$ again. So all three versions of the estimation problem have the same solution.

4. We would usually assume that the data are:

- the observed values of a simple random sample of size $n = 9$
- from the Normal distribution $N(\mu, \sigma^2)$ with unknown values of both μ and σ^2.

Note that, although the stem and leaf plot of the data in **R** shown below looks roughly symmetric and bell-shaped, it is not clear how the new cars were chosen or how comparable the driving conditions were, so the observations may not necessarily be a *simple random sample* from the population of fuel consumption figures for *all* cars of this type.

```
> stem(fuel)
  The decimal point is at the |
   9 | 9
  10 | 00
  10 | 579
  11 | 2
  11 | 8
  12 | 1
```

Summary values of the data are:

$$
\begin{array}{llllll}
n = 9 & \sum_{j=1}^n x_j & = 97.1 \quad \text{giving} & \bar{x} & = \sum_{j=1}^n x_j/n & = 10.78889 \\
 & \sum_{j=1}^n x_j^2 & = 1052.759 & s^2 & = \sum_{j=1}^n (x_j - \bar{x})^2/(n-1) & = 0.6447611
\end{array}
$$

From your notes, for a simple random sample of size n from the $N(\mu, \sigma^2)$ distribution, a $100(1-\alpha)\%$ confidence interval (c_L, c_U) for the population mean μ is given by

$$c_L = \bar{X} - t_{n-1;\alpha/2} S/\sqrt{n} \qquad \text{and} \qquad c_U = \bar{X} + t_{n-1;\alpha/2} S/\sqrt{n}.$$

Now $n - 1 = 8$, $\alpha = 0.1$ (since we want a 90% confidence interval), and from **R** or statistical tables $t_{8;0.05} = 1.860$. Combining this with the data gives

$$
\begin{array}{lllllll}
c_L & = & 10.7889 - 1.860 \times \sqrt{0.6448}/\sqrt{9} & = & 10.2912 & \simeq & 10.291 \\
c_U & = & 10.7889 + 1.860 \times \sqrt{0.6448}/\sqrt{9} & = & 11.2866 & \simeq & 11.287
\end{array}
$$

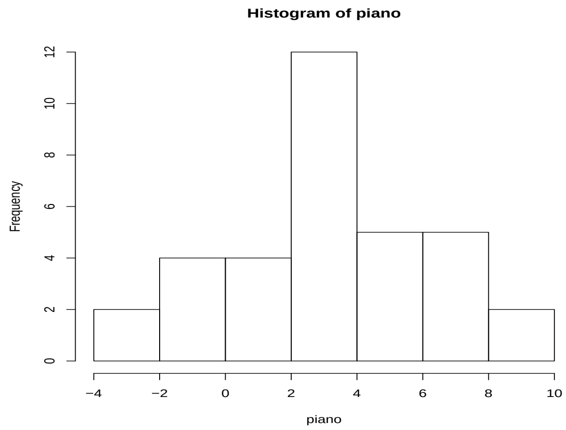
and under our assumptions the required 90% confidence interval for μ is $(10.291, 11.287)$

5. Assume that the data are:

   - the observed values of a simple random sample of size $n = 34$
   - from a distribution with unknown population mean μ and population variance σ^2.

   The sample size $n = 34$ is reasonably large and the sample histogram below is roughly symmetric and bell-shaped, so the central limit theorem enables us to assume $\sqrt{n}(\bar{X} - \mu)/S \simeq t_{n-1}$. However, no information is given about how the children were chosen, or what population they were chosen from, and in practice we might want to explore these aspects further. Summary values of the data are:

$$
\begin{array}{llllll}
n = 34 & \sum_{j=1}^n x_j & = 124 \quad \text{giving} & \bar{x} & = \sum_{j=1}^n x_j/n & = 3.647059 \\
 & \sum_{j=1}^n x_j^2 & = 750 & s^2 & = \sum_{j=1}^n (x_j - \bar{x})^2/(n-1) & = 9.023173
\end{array}
$$



If we assume $\sqrt{n}(\bar{X} - \mu)/S \simeq t_{n-1}$, then using the same argument as that given in your notes for a simple random sample of size n from the $N(\mu, \sigma^2)$ distribution, a $100(1-\alpha)\%$ confidence interval (c_L, c_U) for the population mean μ is given by

$$c_L = \bar{X} - t_{n-1;\alpha/2} S/\sqrt{n} \qquad \text{and} \qquad c_U = \bar{X} + t_{n-1;\alpha/2} S/\sqrt{n}.$$

Now $n - 1 = 33$, $\alpha = 0.05$ (since we want a 95% confidence interval), and from **R** or statistical tables $t_{33;0.025} = 2.0345$ (for the tables you have to interpolate between $t_{32;0.025} = 2.037$ and $t_{34;0.025} = 2.032$). Combining this with the data gives

$$\begin{aligned} c_L &= 3.6470 - 2.0345 \times \sqrt{9.0232}/\sqrt{34} &= 2.5990 &\simeq 2.6 \\ c_U &= 3.6470 + 2.0345 \times \sqrt{9.0232}/\sqrt{34} &= 4.6951 &\simeq 4.7 \end{aligned}$$

and the required 95% confidence interval for μ is $(2.6, 4.7)$