# MATH11400 Statistics 1 2010–11

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

## Solution Sheet 9

1. 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 variance $\sigma^2$ is given by

$$c_L = \frac{\sum_{j=1}^{n}(X_j - \bar{X})^2}{\chi^2_{n-1\,;\,\alpha/2}} \quad \text{and} \quad c_U = \frac{\sum_{j=1}^{n}(X_j - \bar{X})^2}{\chi^2_{n-1\,;\,1-\alpha/2}}.$$

Now $n-1 = 8$, $\sum_1^9 (x_i - \bar{x})^2 = 5.1581$, $\alpha = 0.1$ (since we want a $90\%$ confidence interval), and from **R** (or the annex sheet) $\chi^2_{8\,;\,0.95}$ = `qchisq(0.05,8)= 2.733` and $\chi^2_{8\,;\,0.05}$ = `qchisq(0.95,8)= 15.507`.

Combining this with the data gives $c_L = 5.1581/15.507 = 0.333$, $c_U = 5.1581/2.733 = 1.887$ so under our assumptions the required $90\%$ confidence interval for $\sigma^2$ is $(0.333, 1.887)$.

2. We have looked at these data before, so we can assume that the data are:

- the observed values of a simple random sample of size $n = 25$
- from the Exponential($\theta$) distribution with unknown value of $\theta$.

(a) Summary values of the full data set are: $n = 25$ $\sum_{j=1}^{n} x_j = 95.3$
From your notes, for a simple random sample of size $n$ from the Exponential($\theta$) distribution, a $100(1-\alpha)\%$ confidence interval $(c_L, c_U)$ for $\theta$ is given by

$$c_L = \frac{\chi^2_{2n\,;\,1-\alpha/2}}{2\sum_{i=1}^{n} x_i} \qquad \text{and} \qquad c_U = \frac{\chi^2_{2n\,;\,\alpha/2}}{2\sum_{i=1}^{n} x_i}.$$

Now $2n = 50$, $\alpha = 0.05$ (since we want a $95\%$ confidence interval), and from **R** (or the annex sheet) $\chi^2_{50;0.975}$ =`qchisq(0.025,50)=32.36` and $\chi^2_{50;0.025}$ = `qchisq(0.975,50)=71.42`. Combining this with the data gives

$$\begin{aligned} c_L &= 32.36/(2 \times 95.3) &= 0.1698 &\simeq 0.17 \\ c_U &= 71.42/(2 \times 95.3) &= 0.3747 &\simeq 0.37 \end{aligned}$$

so the required $95\%$ confidence interval for $\theta$ based on the full sample is $(0.17, 0.37)$ and the length of the interval is $0.2$.

(b) Substituting in the $\chi^2$ values from (a), the length of the $95\%$ confidence interval based on a random sample of size 25 is $[(71.42 - 32.36)/2]/\sum_{i=1}^{25} X_i = 19.53/\sum_{i=1}^{25} X_i$. This length will of course vary from sample to sample with the observed values of the $X_i$. However, from the result given, its expected value is $\mathrm{E}(1/\sum_{i=1}^{25} X_i) = \theta/24$. Thus the average length of the interval is $\theta(19.53/24) = (0.814)\theta$.

3. Assume that the interview response data are:

- the observed values of a simple random sample of size $n = 1000$

- from a Bernoulli($\theta$) distribution with unknown values of $\theta$.

Here the sample size $n = 1000$ is very large so the central limit theorem enables us to assume that $\sqrt{n}(\bar{X} - \theta)/\sqrt{\theta(1-\theta)})$ has approximately the $N(0, 1)$ distribution and that the effect of replacing the Bernoulli variance $\theta(1 - \theta)$ by the estimate $\hat{\theta}(1 - \hat{\theta})$ will be negligible, where $\hat{\theta} = \bar{X} = 370/1000 = 0.37$.

Thus, from your notes, a $100(1 - \alpha)\%$ confidence interval $(c_L, c_U)$ for $\theta$ is given by

$$c_L = \bar{X} - z_{\alpha/2}\sqrt{\hat{\theta}(1-\hat{\theta})/n} \qquad \text{and} \qquad c_U = \bar{X} + z_{\alpha/2}\sqrt{\hat{\theta}(1-\hat{\theta})/n}.$$

Now $n - 1 = 8$, $\alpha = 0.01$ (since we want a $99\%$ confidence interval), and from **R** (or the annex sheet, recalling `qnorm` and `pnorm` are inverses of each other) $z_{0.005} =$ `qnorm(0.995)=2.5758`. Combining this with the data gives

$$\begin{aligned} c_L &= 0.37 - 2.5758 \times \sqrt{0.37 \times 0.63/1000} = 0.3307 \simeq 0.331 \\ c_U &= 0.37 + 2.5758 \times \sqrt{0.37 \times 0.63/1000} = 0.4093 \simeq 0.409 \end{aligned}$$

and under our assumptions the required $95\%$ confidence interval for $\theta$ is $(0.331, 0.409)$

4. Again 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 variance $\sigma^2$ is given by

$$c_L = \sum_{j=1}^{n}(X_j - \bar{X})^2/\chi^2_{n-1\,;\,\alpha/2} \quad \text{and} \quad c_U = \sum_{j=1}^{n}(X_j - \bar{X})^2/\chi^2_{n-1\,;\,1-\alpha/2}.$$

Again, $n - 1 = 33$, $\sum_1^{33}(x_i - \bar{x})^2 = 297.7647$, $\alpha = 0.05$, and from **R** (or the annex sheet) we get $\chi^2_{33\,;\,0.975}$ =`qchisq(0.025,33)= 19.05` and $\chi^2_{33\,;\,0.025} =$ `qchisq(0.975,33)= 50.73`.

Combining this with the data gives

$$c_L = 297.7647/50.73 = 5.870 \qquad c_U = 297.7647/19.05 = 15.631$$

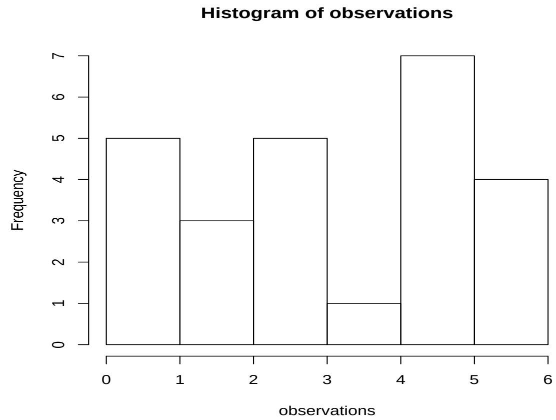
and under our assumptions the required $95\%$ confidence interval for $\sigma^2$ is $(5.870, 15.631)$

5. The sample histogram is shown below. It doesn't look that uniform, but is not that unreasonable for the given sample size.

The relevant summary statistics here are:
$n = 25$ $\quad \sum_{j=1}^{n} x_j = 74.64$ $\quad \bar{x} = 2.9856$ $\quad x_{(25)} = \max\{x_1, \ldots, x_{25}\} = 5.99$.

(a) You are given that $P(X_{(n)}/\theta < v) = v^n$, where here $n = 25$.
Hence $P(X_{(25)}/\theta < v_1) = 0.025$ gives $v_1 = (0.025)^{1/25} = (0.025)^{0.04} = 0.8628$, and
$P(X_{(25)}/\theta > v_2) = 1 - P(X_{(25)}/\theta < v_2) = 0.025$ gives $v_2 = (1 - 0.025)^{0.04} = 0.99990$.
Thus $0.95 = P(0.8628 \le X_{(25)}/\theta \le 0.99990) = P(X_{(25)}/0.99990 \le \theta \le X_{(25)}/0.8628)$ so the interval with end points $(X_{(25)}/0.99990, X_{(25)}/0.8628)$ forms a $95\%$ confidence interval for $\theta$.

For the given data, $x_{(25)} = 5.99$, so a $95\%$ confidence interval computed in this way from the largest observation would have end points $(6.00, 6.94)$ and length $0.94$.




(b) The data are a simple random sample of size $n = 25$ from the U$(0, \theta)$ distribution with mean $\theta/2$ and variance $\theta^2/12$. The sample size is reasonably large and the underlying distribution is symmetric, so, using the CLT, $\bar{X}$ has approximately the $N(\theta/2, \theta^2/12n)$ distribution, i.e. $(2\bar{X} - \theta)/(\theta/\sqrt{3n}) \sim N(0, 1)$. Moreover, we can assume that the effect of replacing $\theta$ by the estimate $\hat{\theta}$ in the variance will not be significant, where $\hat{\theta}_{mom} = 2\bar{X} = 5.9712$. Thus, a $100(1 - \alpha)\%$ confidence interval $(c_L, c_U)$ for $\theta$ is given by

$$c_L = 2\bar{X} - z_{\alpha/2}\hat{\theta}/\sqrt{3n} = 2\bar{X}(1 - (z_{\alpha/2}/\sqrt{3n})); c_U = 2\bar{X} + z_{\alpha/2}\hat{\theta}/\sqrt{3n} = 2\bar{X}(1 + (z_{\alpha/2}/\sqrt{3n})).$$

Now $n = 25$, $\alpha = 0.05$ (since we want a $95\%$ confidence interval), and from **R** (or the annex sheet) $z_{0.025}$ =`qnorm(0.975)= 1.96`, giving $c_L = 4.6198 \simeq 4.62$ and $c_L = 7.3226 \simeq 7.32$. Thus an approximate $95\%$ confidence interval for $\theta$ is $(4.62, 7.32)$, with length $2.70$.

Note that the interval found using $\hat{\theta}_{mle}$ has much shorter length than that found using the $\hat{\theta}_{mom}$ (in fact, the first interval is completely contained within the second). Note also that the lower end point $c_L = 4.62$ of the confidence interval based on $\hat{\theta}_{mom}$ is inconsistent with the fact that we already know $\theta$ MUST be $\geq x_{(25)} = 5.99$; as we saw earlier $\hat{\theta}_{mom}$ is a much less efficient estimate than $\hat{\theta}_{mle}$.

6. **Model assumptions**: (a) The weights of the 25 packets are a simple random sample from the population of weights for all packets produced that day. (b) The population distribution is $N(\mu, 4^2)$, where $\mu$ is unknown.

   **Hypotheses**: $H_0$: $\mu = 200$ versus $H_1$: $\mu \neq 200$.
   The null hypothesis $H_0$ corresponds to *no difference* between the actual mean of the population of weights for that day and the advertised weight of $200g$. The alternative hypothesis $H_1$ corresponds to there being a difference (which could be either positive or negative).

   **Test Statistic**: Since $\bar{X}$ is the natural estimator of $\mu$, we base our test statistic on $\bar{X} - \mu_0 = \bar{X} - 200$. Since the population standard deviation $\sigma_0 = 4$ is known and $n = 25$, we can take as our test statistic $T(X_1, \ldots, X_n) = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0 = 5(\bar{X} - 200)/4$, where $\bar{X} \sim N(\mu, \sigma_0^2/n) = N(\mu, 16/25)$.
   Thus, when $H_0$ is true (i.e. when $\mu = \mu_0 = 200$) we have $T = 5(\bar{X} - 200)/4 \sim N(0, 1)$.

   The data give $\bar{x} = 202.275$ so the observed test statistic is $t_{obs} = 2.84375$.

   ***p*-value**: Since the alternative of interest is $H_1$: $\mu \neq 200$, the values of $T$ which are less consistent with $H_0$ than $t_{obs}$ are the set of values $\{|T| > |t_{obs}|\}$ so

$$\begin{aligned} p\text{-value} &= P(|T| > |t_{obs}||H_0 \text{ true}) = P(|Z| > 2.844) \text{ where } Z \sim N(0,1) \\ &= 2(1 - \Phi(2.844)) = \texttt{2(1-pnorm(2.844))} = 2(1 - 0.9978) = 0.00446. \end{aligned}$$

**Critical region**: Since the alternative of interest is $H_1$: $\mu \neq 200$, the values of $T$ which are less consistent with $H_0$ than a value $t$ are the set of values $\{|T| > |t|\}$. Thus the critical region of values for which the test would reject $H_0$ is of the form $C = \{|T| > c^*\}$. A test has significance level $\alpha$ if P(Reject $H_0|H_0$ true) $= \alpha$. Thus, for a $0.01$-level test, $c^*$ is defined by the condition

$$\begin{aligned} 0.01 &= \alpha = P(\text{Reject } H_0|H_0 \text{ true}) = P(|T| > c^* \mid H_0 \text{ true}) \\ &= P(|Z| > c^*) \ [\text{where } Z \sim N(0,1)] = 2(1 - \Phi(c^*)), \\ \text{so } c^* &= \Phi^{-1}(1 - 0.005) = z_{0.005} = \texttt{qnorm(0.995)} = 2.576 \end{aligned}$$

and the resulting critical region is $C = \{|T| \geq 2.576\}$ .

**Conclusions**: The $p$-value is very small, so there is strong evidence that the data are not consistent with $H_0$ being true. The observed test statistic value $t_{obs} = 2.84375$ falls well within the critical region of the $0.01$-level test, so we would reject $H_0$ in favour of $H_1$, and conclude that the mean of the population of packet weights is not equal to $200g$, at least for that day's production.

Note that a test procedure with significance level $\alpha$ will reject the null hypothesis if the observed $p$-value is less than or equal to $\alpha$. For these data the $p$-value is $0.00446$, so an $\alpha$-level test would reject $H_0$ if and only if $\alpha \geq 0.00446$.

7. **Model assumptions**: (a) The values $X_1, \ldots, X_n$ are a simple random sample of size $n$ from a given population. (b) The population distribution is $N(\mu, 5^2)$, where $\mu$ is unknown.

   **Hypotheses**: $H_0$: $\mu = 100$ versus $H_1$: $\mu > 100$.

   **Test Statistic**: Since $\bar{X}$ is the natural estimator of $\mu$, we base our test statistic on $\bar{X} - \mu_0 = \bar{X} - 100$. Since the population standard deviation $\sigma_0 = 5$ is known we can take as our test statistic $T(X_1, \ldots, X_n) = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0 = \sqrt{n}(\bar{X} - 100)/5$, where $\bar{X} \sim N(\mu, \sigma_0^2/n) = N(\mu, 25/\sqrt{n})$.
   Thus, when $H_0$ is true (i.e. when $\mu = \mu_0 = 100$) we have $T = \sqrt{n}(\bar{X} - 100)/5 \sim N(0,1)$.

   **Sample size**: We are given that the test procedure rejects $H_0$ if and only if $\bar{X} > 102$, so the test procedure rejects $H_0$ if and only if $T > \sqrt{n}(102 - 100)/5 = 2\sqrt{n}/5$.

   For a test procedure with significance level $\alpha$ we require

$$\begin{aligned} \alpha &= P(\text{Reject } H_0|H_0 \text{ true}) = P(T > 2\sqrt{n}/5 \mid H_0 \text{ true}) \\ &= P(Z > 2\sqrt{n}/5) \ [\text{where } Z \sim N(0,1)] = 1 - \Phi(2\sqrt{n}/5). \end{aligned}$$

$$\begin{aligned} \text{Thus } \alpha < 0.05 \quad &\Rightarrow \quad 1 - \Phi(2\sqrt{n}/5) < 0.05 \\ &\Rightarrow \quad \Phi(2\sqrt{n}/5) > 0.95 \\ &\Rightarrow \quad 2\sqrt{n}/5 > \Phi^{-1}(0.95) = z_{0.05} = \texttt{qnorm(0.95)} = 1.645 \\ &\Rightarrow \quad n > 16.9. \end{aligned}$$

   Since the sample size must be an integer, the smallest such $n$ satisfying this inequality is $n = 17$.