Post-exam 2 practice questions –solutions
18.05, Spring 2014

1 Confidence intervals

To practice for the exam use the t and z-tables supplied at the end of this file. Be sure to learn to use these tables. Note the t and z-tables give left tail probabilities and the χ^2-table gives right tail critical values.

1. (a) We compute the data mean and variance $\bar{x} = 65$, $s^2 = 35.778$. The number of degrees of freedom is 9. We look up the *critical value* $t_{9,0.025} = 2.262$ in the t-table The 95% confidence interval is

$$\left[\bar{x} - \frac{t_{9,0.025}s}{\sqrt{n}},\, \bar{x} + \frac{t_{9,0.025}s}{\sqrt{n}}\right] = \left[65 - 2.262\sqrt{3.5778},\, 65 + 2.262\sqrt{3.5778}\right] = [60.721,\, 69.279]$$

On the exam you will be expected to be able to use the t-table. We won't ask you to compute by hand the mean and variance of 10 numbers.

95% confidence means that in 95% of experiments the random interval will contain the true θ. It is not the probability that θ is in the given interval. That depends on the prior distribution for θ, which we don't know.

(b) We can look in the z-table or simply remember that $z_{0.025} = 1.96$. The 95% confidence interval is

$$\left[\bar{x} - \frac{z_{0.025}\sigma}{\sqrt{n}},\, \bar{x} + \frac{z_{0.025}\sigma}{\sqrt{n}}\right] = \left[65 - \frac{1.96 \cdot 5}{\sqrt{10}},\, 65 + \frac{1.96 \cdot 5}{\sqrt{10}}\right] = [61.901,\, 68.099]$$

This is a narrower interval than in part (a). There are two reasons for this, first the true variance 25 is smaller than the sample variance 35.8 and second, the normal distribution has narrower tails than the t distribution.

(c) We use the normal-normal update formulas to find the posterior pdf for θ.

$$a = \frac{1}{16}, \quad b = \frac{10}{25}, \quad \mu_{\text{post}} = \frac{a60 + b65}{a + b} = 64.3, \quad \sigma^2_{\text{post}} = \frac{1}{a+b} = 2.16.$$

The posterior pdf is $f(\theta|\text{data}) = \text{N}(64.3, 2.16)$. The posterior 95% probability interval for θ is

$$\left[64.3 - z_{0.025}\sqrt{2.16},\, 64.3 + z_{0.025}\sqrt{2.16}\right] = [61.442,\, 67.206]$$

(d) There's no one correct answer; each method has its own advantages and disadvantages. In this problem they all give similar answers.

2. Suppose we have taken data $x_1, \ldots, x_n$ with mean $\bar{x}$. The 95% confidence interval for the mean is $\overline{x} \pm z_{0.025}\dfrac{\sigma}{\sqrt{n}}$. This has width $2\, z_{0.025}\dfrac{\sigma}{\sqrt{n}}$. Setting the width equal to 1 and substitituting values $z_{0.025} = 1.96$ and $\sigma = 5$ we get

$$2 \cdot 1.96\frac{5}{\sqrt{n}} = 1 \;\Rightarrow\; \sqrt{n} = 19.6.$$

So, $n = (19.6)^2 = \boxed{384.}$.

If we use our rule of thumb that $z_{0.025} = 2$ we have $\sqrt{n}/10 = 2 \Rightarrow n = 400$.

3. We need to use the studentized mean $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$..

We know $t \sim t(n-1) = t(48)$. So we use the $m = 48$ line of the t table and find $t_{0.05} = 1.677$. Thus,

$$P(-1.677 < \frac{\bar{x} - \mu}{s/\sqrt{n}} < 1.677 \mid \mu) = 0.90.$$

Unwinding this, we get the 90% confidence interval for μ is

$$\left[\bar{x} - \frac{s}{\sqrt{n}} \cdot 1.677, \bar{x} + \frac{s}{\sqrt{n}} \cdot 1.677\right] = \left[92 - \frac{0.75}{7} \cdot 1.677, 92 + \frac{0.75}{7} \cdot 1.677\right] = \boxed{[91.82, 92.18].}$$

4. **(a)** The rule-of-thumb is that a 95% confidence interval is $\bar{x} \pm 1/\sqrt{n}$. To be within 1% we need

$$\frac{1}{\sqrt{n}} = 0.01 \Rightarrow n = 10000.$$

Using $z_{0.025} = 1.96$ instead the 95% confidence interval is

$$\bar{x} \pm \frac{z_{0.025}}{2\sqrt{n}}.$$

To be within 1% we need

$$\frac{z_{0.025}}{2\sqrt{n}} = 0.01 \Rightarrow n = 9604.$$

Note, we are still using the standard Bernoulli approximation $\sigma \leq 1/2$.

(b) The 90% confidence interval is $\overline{x} \pm z_{0.05} \cdot \frac{1}{2\sqrt{n}}$. Since $z_{0.05} = 1.64$ and $n = 400$ our confidence interval is

$$\overline{x} \pm 1.64 \cdot \frac{1}{40} = \overline{x} \pm 0.041$$

If this is entirely above 0.5 we have $\overline{x} - 0.041 > 0.5$, so $\overline{x} > 0.541$. Let T be the number out of 400 who prefer A. We have $\overline{x} = \frac{T}{400} > 0.541$, so $\boxed{T > 216}$.

5. A 95% confidence means about $5\% = 1/20$ will be wrong. You'd expect about 2 to be wrong.

With a probability $p = 0.05$ of being wrong, the number wrong follows a Binomial$(40, p)$ distribution. This has expected value 2, and standard deviation $\sqrt{40(0.05)(0.95)} = 1.38$. 10 wrong is (10-2)/1.38 = 5.8 standard deviations from the mean. This would be surprising.

2 χ^2 confidence interval

6. We have $n = 27$ and $s^2 = 5.86^2$. If we fix a hypothesis for σ^2 we know

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

We used R to find the critical values. (Or use the χ^2 table at the end of this file.)

```
c025 = qchisq(0.975,26) = 41.923
c975 = qchisq(0.025,26) = 13.844
```

The 95% confidence interval for σ^2 is

$$\left[\frac{(n-1)\cdot s^2}{c_{0.025}}, \frac{(n-1)\cdot s^2}{c_{0.975}}\right] = \left[\frac{26\cdot 5.86^2}{41.923}, \frac{26\cdot 5.86^2}{13.844}\right] = [21.2968,\ 64.4926]$$

We can take square roots to find the 95% confidence interval for σ

$$[4.6148,\ 8.0307]$$

On the exam we will give you enough of a table to compute the critical values you need for χ^2 distributions.

3 Bootstrapping

7. (a) Step 1. We have the point estimate $p \approx \hat{p} = 0.30303$.
Step 2. Use the computer to generate many (say 10000) size 100 samples. (These are called the bootstrap samples.)
Step 3. For each sample compute $p^* = 1/\bar{x}^*$ and $\delta^* = p^* - \hat{p}$.
Step 4. Sort the δ^* and find the critical values $\delta_{0.95}$ and $\delta_{0.05}$. (Remember $\delta_{0.95}$ is the 5th percentile etc.)
Step 5. The 90% bootstrap confidence interval for p is

$$[\hat{p} - \delta_{0.05},\ \hat{p} - \delta_{0.95}]$$

(b) It's tricky to keep the sides straight here. We work slowly and carefully:
The 5th and 95th percentiles for $\bar{x}^*$ are the 10th and 190th entries

$$2.89, \quad 3.72$$

(Here again there is some ambiguity on which entries to use. We will accept using the 11th or the 191st entries or some interpolation between these entries.)
So the 5th and 95th percentiles for p^* are

$$1/3.72 = 0.26882, \quad 1/2.89 = 0.34602$$

So the 5th and 95th percentiles for $\delta^* = p^* - \hat{p}$ are

$$-0.034213, \quad 0.042990$$

These are also the 0.95 and 0.05 critical values.
So the 90% CI for p is

$$[0.30303 - 0.042990,\ 0.30303 + 0.034213] = [0.26004,\ 0.33724]$$

8. (a) The steps are the same as in the previous problem except the bootstrap samples are generated in different ways.

Step 1. We have the point estimate $q_{0.5} \approx \hat{q}_{0.5} = 3.3$.
Step 2. Use the computer to generate many (say 10000) size 100 resamples of the original data.
Step 3. For each sample compute the median $q^*_{0.5}$ and $\delta^* = q^*_{0.5} - \hat{q}_{0.5}$.
Step 4. Sort the δ^* and find the critical values $\delta_{0.95}$ and $\delta_{0.05}$. (Remember $\delta_{0.95}$ is the 5th percentile etc.)
Step 5. The 90% bootstrap confidence interval for $q_{0.5}$ is

$$[\hat{q}_{0.5} - \delta_{0.05},\ \hat{q}_{0.5} - \delta_{0.95}]$$

(b) This is very similar to the previous problem. We proceed slowly and carefully to get terms on the correct side of the inequalities.

The 5th and 95th percentiles for $q^*_{0.5}$ are

$$2.89, \quad 3.72$$

So the 5th and 95th percentiles for $\delta^* = q^*_{0.5} - \hat{q}_{0.5}$ are

$$[2.89 - 3.3, \quad 3.72 - 3.3] = [-0.41, \quad 0.42]$$

These are also the 0.95 and 0.05 critical values.
So the 90% CI for p is

$$[3.3 - 0.42,\ 3.3 + 0.41] = [2.91,\ 3.71]$$

4 Linear regression/Least squares

9. (a) The density f_{ε_i} for the normal distribution is known.

$$f(y_i \,|\, a, b, x_i, \sigma) = f_{\varepsilon_i}(y_i - ax_i - b) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(y_i - ax_i - b)^2}{2\sigma^2}}.$$

(b) (i) The y values are $8, 2, 1$. The likelihood function is a product of the densities found in part (a)

$$f(y\text{-data} \,|\, a, b, \sigma, x\text{-data}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^3 \mathrm{e}^{-((8-a-b)^2 + (2-3a-b)^2 + (1-5a-b)^2)/2\sigma^2}$$

$$\ln(f(y\text{-data} \,|\, a, b, \sigma, x\text{-data})) = -3\ln(\sigma) - \frac{3}{2}\ln(2\pi) - \frac{(8-a-b)^2 + (2-3a-b)^2 + (1-5a-b)^2}{2\sigma^2}$$

(ii) We just copy our answer in part (i) replacing the explicit values of x_i and y_i by their symbols

$$f(y_1, \ldots, y_n \,|\, a, b, \sigma, x_1, \ldots, x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \mathrm{e}^{-\sum_{j=1}^n (y_j - ax_j - b)^2/2\sigma^2}$$

$$\ln(f(8, 3, 2 \,|\, a, b, \sigma)) = -n\ln(\sigma) - \frac{n}{2}\ln(2\pi) - \sum_{j=1}^n (y_j - ax_j - b)^2/2\sigma^2$$

(c) We set partial derivatives to 0 to try and find the MLE. (Don't forget that σ is a contstant.)

$$\begin{aligned}\frac{\partial}{\partial a}\ln(f(8,3,2\,|\,a,b,\sigma)) &= -\frac{-2(8-a-b)-6(2-3a-b)-10(1-5a-b)}{2\sigma^2}\\ &= \frac{-70a-18b+38}{2\sigma^2}\\ &= 0\\ &\Rightarrow 70a+18b=38\end{aligned}$$

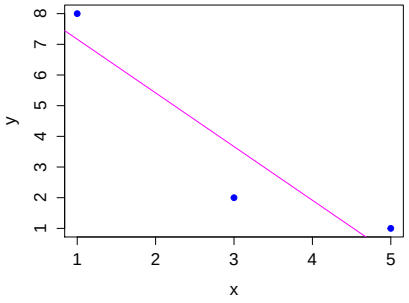
$$\begin{aligned}\frac{\partial}{\partial b}\ln(f(8,3,2\,|\,a,b,\sigma)) &= -\frac{-2(8-a-b)-2(2-3a-b)-2(1-5a-b)}{2\sigma^2}\\ &= \frac{-18a-6b+22}{2\sigma^2}\\ &= 0\\ &\Rightarrow 18a+6b=22\end{aligned}$$

We have two simultaneous equations: $70a+18b=38$, $18a+6b=22$.
These are easy to solve, e.g. first eliminate b and solve for a. We get

$$a=-\frac{7}{4} \qquad b=\frac{107}{12}$$

You can use R to plot the data and the regression line you found in part (c). Here's the R code I used to make the plot

```
x = c(1,3,5)
y = c(8,2,1)
a = -7/4
b = 107/12
plot(x,y,pch=19,col="blue")
abline(a=b,b=a, col="magenta")
```



10. The correlation between x and y is the same as the coefficient b_1 of the best fit line to the standardized data

$$u_i=\frac{x_i-\bar{x}}{\sqrt{s_{xx}}}, \quad v_i=\frac{y_i-\bar{y}}{\sqrt{s_{yy}}}$$

11. The total squared error is

$$S(a)=\sum\left(y_i-\frac{a}{x_i}\right)^2.$$

Taking the derivative and setting it to 0 gives

$$S'(a)=\sum -\frac{2}{x_i}\left(y_i-\frac{a}{x_i}\right)=0$$

This implies

$$a\sum \frac{1}{x_i^2} = \sum \frac{y_i}{x_i} \;\Rightarrow\; \boxed{\hat{a} = \frac{\sum y_i/x_i}{\sum 1/x_i^2}.}$$

12.

(a) We're given $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. Since $ax_i + b$ is a constant, Y_i is simply a shift of ε_i. Thus

$$\begin{aligned} E(Y_i) &= ax_i + b + E(\varepsilon_i) = ax_i + b \\ \mathrm{Var}(Y_i) &= \mathrm{Var}(\varepsilon_i) = \sigma^2. \end{aligned}$$

Since a shifted normal random variable is still normal (you should be able to show this by transforming cdf's) we have

$$Y_i \sim \mathrm{N}(ax_i + b, \sigma^2).$$

(b) The density for a normal distribution is known

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \mathrm{e}^{-\frac{(y_i - ax_i - b)^2}{2\sigma^2}}.$$

(c)

$$\begin{aligned} f(\text{data} \,|\, \sigma, a, b) &= f_{Y_1}(y_1) f_{Y_2}(y_2) \cdots f_{Y_n}(y_n) \\ &= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2 \right). \end{aligned}$$

(d) The log likelihood is

$$\ln(f(\text{data} \,|\, \sigma, a, b)) = -\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

If σ is constant then the only part of the log likelihood that varies is the sum in the last term. So, the maximum likelihood is found by maximizing this sum:

$$-\sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Notice the minus sign out front. This is exactly the same as minimizing

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

This last expression is the expression minimized by least squares. Therefore, under our normality assumptions, the values of a and b are the same for MLE and least squares.

18.05 Introduction to Probability and Statistics
Spring 2014