Post Exam 2 Practice Questions –solutions, 18.05, Spring 2022

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.

Problem 1. **Basketball**

Suppose that against a certain opponent the number of points the MIT basketball team scores is normally distributed with unknown mean θ and unknown variance, σ^2.

Suppose that over the course of the last 10 games between the two teams MIT scored the following points:

59, 62, 59, 74, 70, 61, 62, 66, 62, 75

(a) *Compute a 95% t–confidence interval for θ. Does 95% confidence mean that the probability θ is in the interval you just found is 95%?*

Solution: 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) *Now suppose that you learn that $\sigma^2 = 25$. Compute a 95% z–confidence interval for θ. How does this compare to the interval in (a)?*

Solution: 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) *Let X be the number of points scored in a game. Suppose that your friend is a confirmed Bayesian with a priori belief $\theta \sim N(60, 16)$ and that $X \sim N(\theta, 25)$. He computes a 95% probability interval for θ, given the data in part (a). How does this interval compare to the intervals in (a) and (b)?*

Solution: 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) *Which of the three intervals constructed above do you prefer? Why?*

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

Problem 2. Confidence interval 2

The volume in a set of wine bottles is known to follow a $N(\mu, 25)$ distribution. You take a sample of the bottles and measure their volumes. How many bottles do you have to sample to have a 95% confidence interval for μ with width 1?

Solution: Suppose we have taken data $x_1, \dots, x_n$ 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 $\sqrt{n}/10 = 2 \ \Rightarrow\ n = 400$.

Problem 3. *Suppose data $x_1, \dots, x_n$ are i.i.d. and drawn from $N(\mu, \sigma^2)$, where μ and σ are unknown.*

Suppose a data set is taken and we have $n = 49$, sample mean $\bar{x} = 92$ and sample standard deviation $s = 0.75$.

Find a 90% confidence interval for μ.

Solution: 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

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

Problem 4. Polling confidence intervals

You do a poll to see what fraction p of the population supports candidate A over candidate B.

(a) *How many people do you need to poll to know p to within 1% with 95% confidence.*

Solution: 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) *Let p be the fraction of the population who prefer candidate A. If you poll 400 people, how many have to prefer candidate A so that the 90% confidence interval is entirely above $p = 0.5$.*

Solution: 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 $\boxed{T > 216}$.

Problem 5. **Confidence intervals 3**

Suppose you made 40 confidence intervals with confidence level 95%. About how many of them would you expect to be "wrong'? That is, how many would not actually contain the parameter being estimated? Should you be surprised if 10 of them are wrong?

Solution: 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 Chi-square confidence interval

Problem 6. **Hotel**

A statistician chooses 27 randomly selected dates, and when examining the occupancy records of a particular motel for those dates, finds a standard deviation of 5.86 rooms rented. If the number of rooms rented is normally distributed, find the 95% confidence interval for the population standard deviation of the number of rooms rented.

Solution: 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

Problem 7. ***Parametric bootstrap***
Suppose we have a sample of size 100 drawn from a geom(p) distribution with unknown p. The MLE estimate for p is given by by $\hat{p} = 1/\bar{x}$. Assume for our data $\bar{x} = 3.30$, so $\hat{p} = 1/\bar{x} = 0.30303$.

(a) *Outline the steps needed to generate a parametric basic bootstrap 90% confidence interval.*

Solution: 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) *Suppose the following sorted list consists of 200 bootstrap means computed from a sample of size 100 drawn from a geometric(0.30303) distribution. Use the list to construct a 90% basic CI for p.*

```
2.68 2.77 2.79 2.81 2.82 2.84 2.84 2.85 2.88 2.89
2.91 2.91 2.91 2.92 2.94 2.94 2.95 2.97 2.97 2.99
3.00 3.00 3.01 3.01 3.01 3.03 3.04 3.04 3.04 3.04
3.04 3.05 3.06 3.06 3.07 3.07 3.07 3.08 3.08 3.08
3.08 3.09 3.09 3.10 3.11 3.11 3.12 3.13 3.13 3.13
3.13 3.15 3.15 3.15 3.16 3.16 3.16 3.16 3.17 3.17
3.17 3.18 3.20 3.20 3.20 3.21 3.21 3.22 3.23 3.23
3.23 3.23 3.23 3.24 3.24 3.24 3.24 3.25 3.25 3.25
3.25 3.25 3.25 3.26 3.26 3.26 3.26 3.27 3.27 3.27
```

```
3.28 3.29 3.29 3.30 3.30 3.30 3.30 3.30 3.30 3.31
3.31 3.32 3.32 3.34 3.34 3.34 3.34 3.35 3.35 3.35
3.35 3.35 3.36 3.36 3.37 3.37 3.37 3.37 3.37 3.37
3.38 3.38 3.39 3.39 3.40 3.40 3.40 3.40 3.41 3.42
3.42 3.42 3.43 3.43 3.43 3.43 3.44 3.44 3.44 3.44
3.44 3.45 3.45 3.45 3.45 3.45 3.45 3.45 3.46 3.46
3.46 3.46 3.47 3.47 3.49 3.49 3.49 3.49 3.49 3.50
3.50 3.50 3.52 3.52 3.52 3.52 3.53 3.54 3.54 3.54
3.55 3.56 3.57 3.58 3.59 3.59 3.60 3.61 3.61 3.61
3.62 3.63 3.65 3.65 3.67 3.67 3.68 3.70 3.72 3.72
3.73 3.73 3.74 3.76 3.78 3.79 3.80 3.86 3.89 3.91
```

Solution: The basic interval requires an algebraic pivot, so 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% basic CI for p is

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

Problem 8. ***Empirical bootstrap***

Suppose we had 100 data points $x_1, \ldots x_{100}$ with sample median $\hat{q_{0.5}} = 3.3$.

(a) *Outline the steps needed to generate an empirical percentile bootstrap 90% confidence interval for the median $q_{0.5}$.*

Solution: For the percentile bootstrap, we don't have to pivot, so the algebra is a little shorter.

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 bootstrap sample compute and save the bootstrap median $q^*_{0.5}$.

Step 4. Find the quantiles $c_{0.05}$ and $c_{0.95}$. (Remember $c_{0.05}$ is the 5th percentile in the list of bootstrap medians, etc.)

Step 5. The 90% percentile bootstrap confidence interval for $q_{0.5}$ is

$$[c_{0.05},\ c_{0.95}]$$

(b) *Suppose now that the sorted list in the previous problem consists of 200 empirical bootstrap medians computed from resamples of size 100 drawn from the original data. Use the list to construct a 90% precentile CI for $q_{0.5}$.*

Solution: The list covers steps 1-3 in part (a). Since it is sorted, step 4 is straightforward. The 5th and 95th percentiles for $q^*_{0.5}$ are

$$2.89, \quad 3.72$$

(Here we just took the 10th and 190th values. We could have interpolated between the 9th and 10th, and 190th and 191st entries, but this would not change our answer to two decimal places.)

The above interval is our empirical percentile bootstrap confidence interval for the median.

4 Linear regression/Least squares

Problem 9. *In this problem we will use maximum likelihood estimates to develop Gauss' method of least squares for fitting lines to data.*

Suppose you have bivariate data, that is, a sequence of pairs $(x_1, y_1), \dots, (x_n, y_n)$. A common model is that there is a linear relationship between x and y, so in principle the data should lie exactly along a line. However, since data has random noise this will not be the case. What we can do is look for the line that best fits the data. To do this we will use a model called simple linear regression.

For bivariate data the simple linear regression model assumes that the x_i are not random and that, for some values of the parameters a and b, we have

$$y_i = ax_i + b + \text{ random noise}$$

To be more precise, we will assume that the value y_i is drawn from a random variable of the form

$$Y_i \sim ax_i + b + \varepsilon_i$$

where ε_i is a normal random variable with mean 0 and variance σ^2. We assume all of the random variables ε_i are independent and that σ is the same for all i.

Notes. *1. The goal in simple linear regression is to find the line, i.e. the values of a and b that best fit the data.*

2. We think of ε_i as the random measurement error.

3. Remember that (x_i, y_i) are not variables. They are data values.

(a) *The distribution of Y_i depends on a, b, σ and x_i. So we write its pdf as*

$$f(y_i \,|\, a, b, x_i, \sigma).$$

Give the formula for the likelihood function corresponding to one random value y_i. (Hint: $y_i - ax_i - b \sim N(0, \sigma^2)$.)

Solution: We know that $y_i - ax_i - b = \varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. Therefore,

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

(b) *For general data $(x_1, y_1), \ldots, (x_n, y_n)$ give the likelihood and log likelihood functions (again as functions of a, b, and σ).*

Solution: The likelihood function is just the product of the likelihoods found in part (a). We'll let $L(a, b, \sigma)$ be the likelihood of the data and $l(a, b, \sigma)$ be the log likelihood.

$$L(a, b, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \mathrm{e}^{-\sum_{i=1}^n (y_i - ax_i - b)^2 / 2\sigma^2}$$

$$l(a, b, \sigma) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \sum_{i=1}^n \frac{(y_i - ax_i - b)^2}{2\sigma^2}.$$

(c) *Find the maximum likelihood estimate for a and b by taking partial derivatives of the log likelihood function and setting them equal to 0.*

Solution: The partial derivatives of $l(a, b, \sigma)$ are

$$\frac{\partial l}{\partial a} = \sum_{i=1}^n 2x_i(y_i - ax_i - b)$$

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

Here is the algebra use in setting these to 0 and solving for a and b:

$$\sum_{i=1}^n 2x_i(y_i - ax_i - b) = 0 \;\Rightarrow\; a\sum x_i^2 + b\sum x_i = \sum x_i y_i$$

$$\sum_{i=1}^n 2(y_i - ax_i - b) = 0 \;\Rightarrow\; a\sum x_i + b\,n = \sum y_i$$

It will be convenient to use the following notations:

$$S_x = \sum_{i=1}^n x_i \qquad S_y = \sum_{i=1}^n y_i \qquad \bar{x} = \frac{S_x}{n} \qquad \bar{y} = \frac{S_y}{n}$$

$$S_{xx} = \sum_{i=1}^n x_i^2 \qquad S_{xy} = \sum_{i=1}^n x_i y_i$$

With this notation our equations become

$$aS_{xx} + bS_x = S_{xy}, \quad aS_x + bn = S_y.$$

Solving the second equation for b gives

$$b = \frac{S_y - aS_x}{n} = \boxed{\bar{y} - a\bar{x}}$$

Putting this into the first equation and solving for a gives

$$a = \frac{S_{xy} - \bar{y}S_x}{S_{xx} - S_x\bar{x}} = \boxed{\frac{S_{xy}/n - \bar{x}\bar{y}}{S_{xx}/n - \bar{x}^2}}.$$

(The last expression came by dividing both numerator and denominator of the middle expression by n.)

Note: We could also find the MLE for σ, but, since we don't need it for fitting the line, we don't bother.

(d) *Suppose we have data $(1, 8)$, $(3, 2)$, $(5, 1)$. Use your answer in part (b) to find the value of a and b which gives the MLE for the best fitting line to the data.*

We have $n = 3$ points. First we compute:

$$S_x = \sum x_i = 9, \qquad \bar{x} = 3, \qquad S_y = \sum y_i = 11, \qquad \bar{y} = 11/3.$$
$$S_{xx} = \sum x_i^2 = 35, \qquad S_{xy} = \sum x_i y_i = 19$$

So,

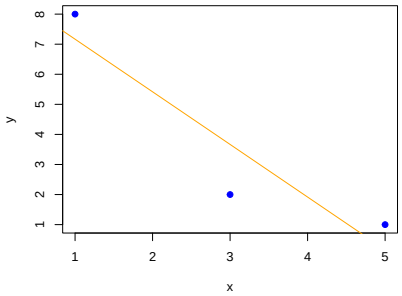
$$a = \frac{S_{xy}/n - \bar{x}\bar{y}}{S_{xx}/n - \bar{x}^2} = \frac{19/3 - 3 \cdot 11/3}{35/3 - 9} = \boxed{-\frac{7}{4}}$$
$$b = \bar{y} - a\bar{x} = \frac{11}{3} + \frac{21}{4} = \boxed{\frac{107}{12}}$$

(e) *Use R to plot the data and the regression line you found in problem (1c). The commands* `plot(x, y, pch=19)` *and* `abline()` *will come in handy.*

Print the plot and turn it in.

Solution: 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="orange")
```



Problem 10. *What is the relationship between correlation and least squares fit line?*

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

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

Problem 11. *You have bivariate data (x_i, y_i). You have reason to suspect the data is related by $y_i = a/x_i + U_i$ where U_i is a random variable with mean 0 and variance σ^2 (the same for all i).*

Find the least squares estimate of a.

Solution: 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

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

Problem 12. ***Least Squares and MLE***

In this problem we will see that the least squares fit of a line is just the MLE assuming the error terms are normally distributed.

For bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$, the simple linear regression model says that y_i is a random value generated by a random variable

$$Y_i = ax_i + b + \varepsilon_i$$

where a, b, x_i are fixed (not random) values, and ε_i is a random variable with mean 0 and variance σ^2.

(a) *Suppose that each $\varepsilon_i \sim N(0, \sigma^2)$. Show that $Y_i \sim N(ax_i + b, \sigma^2)$.*

(b) *Give the formula for the pdf $f_{Y_i}(y_i)$ of Y_i.*

(c) *Write down the likelihood of the data as a function of a, b, and σ.*

(a) Solution: 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

$$E[Y_i] = ax_i + b + E[\varepsilon_i] = ax_i + b$$
$$\mathrm{Var}(Y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2.$$

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) Solution: 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) Solution:

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

(d) Solution: The log likelihood is

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

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:

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

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

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

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 2022