Review for final exam: in-class solutions
MIT 18.05 Spring 2022

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. 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 $\sigma \le 1/2$.

(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 4. **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.

Problem 5. **(Confidence intervals)**

A statistician chooses 20 randomly selected class days and counts the number of students present in 18.05. They find a standard deviation of 4.06 students If the number of students present is normally distributed, find the 95% confidence interval for the population standard deviation of the number of students in attendance.

Solution: We have $n = 20$ and $s^2 = 4.06^2$. 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.)

```
c025 = qchisq(0.975,19) = 32.852
c975 = qchisq(0.025,19) = 8.907
```

The 95% confidence interval for σ^2 is

$$\left[\frac{(n-1)\cdot s^2}{c_{0.025}}, \frac{(n-1)\cdot s^2}{c_{0.975}}\right] = \left[\frac{19\cdot 4.06^2}{32.852}, \frac{19\cdot 4.06^2}{8.907}\right] = [9.53,\ 35.16]$$

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

$$[3.09,\ 5.93]$$

Problem 6. **Linear regression (least squares)**

(a) *Set up fitting the least squares line through the points* $(1,1)$, $(2,1)$, *and* $(3,3)$.

(a) Solution: The model is $y_i = a + bx_i + \varepsilon_i$, where ε_i is random error. We assume the errors are independent with mean 0 and the same variance for each i (homoscedastic).

The total error squared is

$$E^2 = \sum (y_i - a - bx_i)^2 = (1-a-b)^2 + (1-a-2b)^2 + (3-a-3b)^2$$

The least squares fit is given by the values of a and b which minimize E^2. We solve for them by setting the partial derivatives of E^2 with respect to a and b to 0. In R we found that $a = -1/3$, $b = 1$.

Problem 7. ***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.

Problem 8. ***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 9. (NHST chi-square)
A study of recidivism (repeat offenses) of juvenile offenders used an experimental design with

random assignment of juveniles to experimental intervention (Family Group Counseling) or control group (diversion programs). 70 out of 200 people in the control group re-offended and 30 out of 200 people in the experimental group re-offended.

Use a chi-square significance test to test whether the recidivism rates within 6 months for the two experimental groups are significantly different at a significance level of 0.05.

Solution: We will use a chi-square test for homogeneity. Remember we need to use all the data!. For hypotheses we have:

H_0: the re-offense rate is the same for both groups.
H_A: the rates are different.

Here is the table of counts. The computation of the expected counts is explained below.

	Control group		Experimental group		
	observed	expected	observed	expected	
Re-offend	70	50	30	50	100
Don't re-offend	130	150	170	150	300
	200		200		400

The expected counts are computed as follows. Under H_0 the re-offense rates are the same, say θ. To find the expected counts we find the MLE of θ using the combined data:

$$\hat{\theta} = \frac{\text{total re-offend}}{\text{total subjects}} = \frac{100}{400}.$$

Then, for example, the expected number of re-offenders in the control group is $200 \cdot \hat{\theta} = 50$. The other expected counts are computed in the same way.

The chi-square test statistic is

$$X^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}} = \frac{20^2}{50} + \frac{20^2}{150} + \frac{20^2}{50} + \frac{20^2}{150} \approx 8 + 2.67 + 8 + 2.67 \approx 21.33.$$

Finally, we need the degrees of freedom: $df = 1$ because this is a two-by-two table and $(2-1)\cdot(2-1) = 1$. (Or because we can freely fill in the count in one cell and still be consistent with the marginal counts 200, 200, 100, 300, 400 used to compute the expected counts.)

From the χ^2 table: $p = P(X^2 > 21.33 | df = 1) < 0.01$.

Conclusion: we reject H_0 in favor of H_A. The experimental intervention appears to be effective.

18.05 Introduction to Probability and Statistics
Spring 2022