18.05 Problem Set 10, Spring 2022 Solutions

Problem 1. (15: 10,5 pts.) **Z**
Suppose we have n data points drawn from a $N(\theta, 9^2)$ distribution, where the value of the mean θ is unknown.

(a) *Suppose that $n = 30$ and that the sample mean is $\overline{x} = 15$. Give the 95% confidence interval for the mean θ. Also give the 70% confidence interval.*

In this problem you should use R to give **precise** *confidence intervals. That is, you should find the critical values to several decimal points of accuracy.*

Solution: We have $\sigma = 9$, $n = 30$, $\overline{x} = 15$. The z-confidence interval for the mean is $\overline{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$.

So, using R, ($z_{\alpha/2}$ = `pnorm(`$1-\alpha/2$`))` the 95% ($\alpha = 0.05$) confidence interval is

$$15 \pm 3.220549 = [11.77945,\ 18.22055]$$

Likewise, the 70% ($\alpha = 0.3$) confidence interval is

$$15 \pm 1.703034 = [13.29697,\ 16.70303]$$

(b) *Explain why the 95% confidence interval for the mean θ is always narrower when $n = 90$ when $n = 30$.*

Solution: The formula for the confidence interval has width $2z_{\alpha/2}\sigma\sqrt{n}$. Since σ and $z_{\alpha/2}$ are fixed, this gets smaller as n increases.

Problem 2. (10 pts.) **(Soda)**
Consider a machine that is known to fill soda cans with amounts that follow a normal distribution with (unknown) mean μ and standard deviation $\sigma = 3$ mL. We measure the volume of soda in a sample of bottles and obtain the following data (in mL):

352, 351, 361, 353, 352, 358, 360, 358, 359

For this problem, you will probably want to use R to make the calculations.

(a) *Construct a precise 95% confidence interval for the mean μ.*

Solution: The sample mean is $\bar{x} = 356$. Since $z_{0.025} \approx 1.96$, $\sigma = 3$ and $n = 9$, the 95% confidence interval is

$$\bar{x} \pm z_{0.025} \cdot \frac{\sigma}{\sqrt{n}} \approx 356 \pm 1.96 = [354.04,\ 357.96]$$

(We can find $z_{0.025}$ in our memory or using R: `z = qnorm(1-0.025)`.)

(b) *Now construct a 98% confidence interval for the mean μ.*

Solution: We have $z_{0.01} =$ `qnorm(0.99)` ≈ 2.33. So the 98% confidence interval is

$$\bar{x} \pm z_{0.01} \cdot \frac{\sigma}{\sqrt{n}} \approx 356 \pm 2.33 = [353.67, 358.33].$$

(c) *Suppose now that σ is not known. Redo parts (a) and (b), and compare your answers to those above.*

Solution: The sample standard deviation is

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \texttt{sd(c(352 351 361 353 352 358 360 358 359))} \approx 3.94.$$

Since $n = 9$ the number of degrees of freedom for the t-statistic is 8.

Redo (a): $t_{8,0.025} =$ `qt(0.975, 8)` ≈ 2.31. So the 95% t-confidence interval is

$$\bar{x} \pm t_{8,0.025} \cdot \frac{s}{\sqrt{n}} \approx 356 \pm 3.03 = [352.97, 359.03].$$

Redo (b): $t_{8,0.01} =$ `qt(0.99, 8)` ≈ 2.896. So the 98% confidence interval is

$$\bar{x} \pm t_{8,0.01} \cdot \frac{s}{\sqrt{n}} \approx 356 \pm 3.89 = [352.20, 359.80].$$

These intervals are larger than the corresponding intervals from parts (a) and (b). The new uncertainly regarding the value of σ means we (generally) have larger intervals for the same level of confidence. This is reflected in the fact that the t distribution has fatter tails than the normal distribution.

Problem 3. (10 pts.) **(Cholesterol)**
In a study on cholesterol levels a sample of 12 men and women was chosen. The plasma cholesterol levels (mmol/L) of the subjects were as follows:

6.0, 6.4, 7.0, 5.8, 6.0, 5.8, 5.9, 6.7, 6.1, 6.5, 6.3, 5.8.

(a) *Estimate the variance of the plasma cholesterol levels with a 95% confidence interval.*

Solution: This is similar to problem 3c. We assume the data is normally distributed with unknown mean μ and variance σ^2.

We have the number of data points $n = 12$. We use the following R code:

```
data = c(6.0, 6.4, 7.0, 5.8, 6.0, 5.8, 5.9, 6.7, 6.1, 6.5, 6.3, 5.8)
s2 = var(data) # sample variance = s^2 = sum((x_i - xbar)^2)/(n-1)
n = length(data)
conf = 0.95
alpha = 1-conf
df = n - 1 # degrees of freedom
c_alpha_2 = qchisq(1-alpha/2, df) # critical value
c_1_minus_alpha_2 = qchisq(alpha/2, df) # critical value

CI = c((n-1)*s2/c_alpha_2, (n-1)*s2/c_1_minus_alpha_2)
cat('s2 =', s2, '\n')
cat('c_alpha_2 =', c_alpha_2, ', c_1_minus_alpha_2 =', c_1_minus_alpha_2, '\n')
cat(conf, 'CI = [', CI[1], ',', CI[2], ']','\n')
```

The code outputs:
`s2 = 0.1535606`

```
c_alpha_2 = 21.92005 , c_1_minus_alpha_2 = 3.815748
0.95 CI = [ 0.07706035 , 0.4426829 ]
```

So the 95% confidence interval is

$$\left[\frac{(n-1)\cdot s^2}{c_{0.025}}, \frac{(n-1)\cdot s^2}{c_{0.975}}\right] \approx [0.077, 0.443].$$

$s^2 = 0.154$ is our point estimate for σ^2 and the confidence interval is our range estimate with 95% confidence.

(b) *What assumptions did you make about the sample in order to make your estimate?*

Solution: We have assumed that the plasma cholesterol levels are independent and normally distributed. This might not be a good assumption because, e.g. cholesterol for men and women might follow different distributions. We'd have to do further exploration to understand this.

Problem 4. (10 pts.) **(Candy)**
A candy manufacturer produces packs of candy targeted to weigh 52 grams. A quality control manager was concerned that the variation in the actual weights was larger than acceptable. That is, They were concerned that some packs weighed significantly less than 52-grams and some weighed significantly more than 52 grams. In order to estimate σ, the standard deviation of the weights of the nominal 52-gram packs, they took a random sample of $n = 10$ packs off of the factory line. The random sample yielded a sample variance of 4.2 grams^2.

(a) *Use the random sample to derive a 95% confidence interval for σ.* Note: the interval is for σ not σ^2.

Solution: We have $n = 10$ and $s^2 = 4.2$. Assuming that the weights are normally distributed with variance σ^2, we know that $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_9$. We have

$$\begin{aligned} c_{0.025} &= \texttt{qchisq(0.975, 9)} &= 19.023 \\ c_{0.975} &= \texttt{qchisq(0.025, 9)} &= 2.7004 \end{aligned}$$

The 95% confidence interval for σ is given by

$$\left[\sqrt{\frac{s^2(n-1)}{c_{0.975}}}, \sqrt{\frac{s^2(n-1)}{c_{0.025}}}\right] \approx [1.4096, 3.7414]$$

(b) *What assumptions did you make about the sample in order to make your estimate.*

Solution: In order to use a χ^2 confidence interval we assumed that the weights of the packs of candy are independent and normally distributed with variance σ^2.

Problem 5. (33: 5,10,18 pts.) **Probabilities for hypotheses**
When given a 95% confidence interval many people often mistake this to mean that there is a 95% probablility that the true value of the parameter is in the confidence interval. This violates the maxim that frequentists don't give probabilities for hypotheses, but the reasoning is subtle. This problem is aimed at understanding the subtlety.

(a) **What 95% does mean.** *Imagine running the following experiment.*

- *Pick a value μ.*
- *Draw a sample of size n from a $N(\mu, 5^2)$ distribution.*
- *Compute the 95% z-confidence interval for μ.*
- *Check if μ is in the confidence interval.*
- *Record the result.*

If you repeat this experiment many times, what percentage of the time will the chosen μ be in the confidence interval? (Hint, this is not a question about the probability of hypothesized values of μ. It's about the probability of random intervals.)

Solution: Approximately 95%. This is exactly the definition of confidence: the probability that a randomly generated interval will contain the value of θ that was used to generate it.

(b) What 95% doesn't mean. *Now let's modify this a bit. Suppose we have the following choices and prior probabilities for μ*

μ	0	1	2	3
$p(\mu)$	0.94	0.02	0.02	0.02

Suppose I pick a random μ from this prior distribution and draw a sample of size 64 from a $N(\mu, 5)$ distribution. I do not tell you the value of μ, but I do tell you that the sample mean is $\overline{x} = 1.6$.

We both know the sample size ($n = 64$) and the variance used in drawing the sample ($\sigma^2 = 5^2$). So we both know that the 95% confidence interval for μ is

$$\overline{x} \pm z_{0.025} \cdot \frac{\sigma}{\sqrt{n}} \approx 1.6 \pm 1.96 \cdot \frac{5}{8} \approx [0.375,\ 2.825].$$

(i) Compute the prior probability that μ is in the above confidence interval.

(ii) Update the probabilities for μ based on the data and compute the posterior probability that μ is in the above confidence interval.

Are either of these probabilities close to 95%? Why is this not surprising?

Solution: (i) The given interval is $[0.375, 2.825]$. From the possible choices of μ, i.e. 0, 1, 2, 3, this interval contains $\mu = 1$ and $\mu = 2$. Using the prior probability table we get

$$P(\mu \text{ is in } [0.375,\ 2.825]\ ) = P(\mu = 1 \text{ or } \mu = 2) = \boxed{0.04}.$$

(ii) We do the usual Bayesian update:

The prior is given above. The sample mean has distribution $\overline{x} \sim \mathrm{N}(\mu, \sigma^2/n) = \mathrm{N}(\mu, (5/8)^2)$. So, the likelihood is

$$f(\overline{x}|\mu) = \frac{1}{\sqrt{2\pi} \cdot (5/8)} \mathrm{e}^{-(\overline{x}-\mu)^2/(50/64)}.$$

We used `dnorm(1.6, μ, 5/8)` to compute the likelihood column in the Bayesian update table below. The posterior was computed as usual:

```
posterior = (prior × likelihood)/sum(prior × likelihood)
```

Here is the Bayesian update table. (The table is rounded to 3 decimal points of precision. Each of the values was computed in R to higher precision.)

Hypothesis μ	prior $p(\mu)$	likelihood $f(1.6 \mid \mu)$	Bayes numerator prior $\times$ likelihood	posterior $p(\mu \mid \overline{x} = 1.6)$
$\mu = 0$	0.94	0.024	0.023	0.537
$\mu = 1$	0.02	0.403	0.008	0.191
$\mu = 2$	0.02	0.520	0.010	0.247
$\mu = 3$	0.02	0.052	0.001	0.025
	1.0		0.042	1.0

Now using the posterior probability column we get the posterior probability μ is in the 95% confidence interval [0.375, 2.825] is

$$P(\mu = 1 \text{ or } \mu = 2 \mid \overline{x} = 1.6) \approx \boxed{0.438}.$$

Neither prior or posterior probability is close to 0.95. This is not surprising since frequentist confidence intervals makes no use of the prior. So, they can't tell us the probabilities that hypothetical values of μ are in any given interval. The fact that both confidence intervals and Bayesian updating use the likelihood is reflected in the fact that the posterior probability just computed is much greater than the prior probability.

(c) *A medical study shows that patients receiving a new cancer treatment survived an average of 15 months longer than patients receiving the conventional treatment. The study also found that the 95% confidence interval for this average was 10 to 20 months.*

Say whether each of the following statements is true or false. False means that the statement does not follow logically from the confidence-interval result.

(i) There is a 95% probability that the true increase in survival time lies between 10 and 20 months.

(ii) The probability that the new treatment increases survival time is at least 95%.

(iii) The confidence interval provides an estimate of the true value of the average increase in survival time.

(iv) The null hypothesis that there is no change in survival time due to the new treatment is likely to be incorrect.

(v) If we ran the same study 100 times, then in approximately 95 of the studies the 95% confidence intervals would contain the true value of the change in survival time.

(vi) We would reject the null hypothesis of no improvement in survival time at the 5% significance level.

Solution: (i) This statement is a posterior probability of a hypothesis, which is not provided by a confidence interval. So, it does not follow logically and therefore should be marked as false.

(ii) Again, this is a probability of a hypothesis, which confidence intervals don't give us.

(iii) This statement is true. This is the basic point of confidence intervals.

(iv) This statement is again a posterior probability of a hypothesis, so it does not follow logically and therefore should be marked as false.

(v) This statement is true. It recognizes that the confidence interval is random and the confidence level refers to probabilities about this randomness.

(Of course, this assumes that our model of the distributions involved is accurate.)

(vi) This statement is true. Let $\overline{x}$ be the experimental average increase in survival time and let μ be the true average. The 95% confidence interval consists of all values μ_0 for which this experimental average would not cause us to reject the null hypothesis H_0: $\mu = \mu_0$. So, since 0 is outside the interval, we would reject the null hypothesis $\mu = 0$.