18.05 Problem Set 11, Spring 2022 Solutions

Problem 1. (20: 10,5,5 pts.) **(Height)**
Suppose μ is the average height of a college male. You measure the heights (in inches) of twenty college men, getting data $x_1, \ldots, x_{20}$, with sample mean $\overline{x} = 69.55$ in. and sample variance $s^2 = 14.26$ in^2. Suppose that the x_i are drawn from a normal distribution with unknown mean μ and unknown variance σ^2.

(a) *Using $\overline{x}$ and s^2, construct a 90% t–confidence interval for μ.*

Solution: We have $n = 20$ and $\alpha = 0.1$ so

$$t_{\alpha/2} = \texttt{qt(0.05,19) = 1.7291}.$$

Thus the 90% t-confidence interval is given by

$$\left[\overline{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}},\ \overline{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right] \approx [68.09,\ 71.01]$$

(b) *Now suppose you are told that the height of a college male is normally distributed with standard deviation 3.77 in. Using the same data as in part (a), construct a 90% z–confidence interval for μ.*

Solution: We have

$$z_{\alpha/2} = \texttt{qnorm(0.05) = 1.6448}.$$

So the 90% z-confidence interval is given by

$$\left[\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] \approx [68.16,\ 70.94]$$

(c) *In (b), how many people in total would you need to measure to bring the* **width** *of the 90% z–confidence interval down to 1 inch?*

Solution: We need n such that $2 \cdot z_{0.05} \cdot \sigma/\sqrt{n} = 1$. So

$$n = (2 \cdot z_{0.05} \cdot \sigma)^2 = (2 \cdot 1.6448 \cdot 3.77)^2 \approx 153.8.$$

Since you need a whole number of people the answer is $\boxed{n = 154}$.

Problem 2. (10 pts.) **Confidence intervals from standardized statistics**
The Beta distribution arises in a surprising way: draw a sample of size n from a uniform(0,1) distribution and let w_2 be the second smallest value. Then it turns out that

$$w_2 \sim Beta(2, n-1).$$

Now suppose you draw a sample of size n from a uniform$(0, a)$ distribution, where a is unknown. If we let y_2 be the second smallest data value then the standardized order statistic

$$y_2/a \sim Beta(2, n-1).$$

Use y_2 and `qbeta` in R, to define a 95% confidence interval for a. (Because n and y_2 are not given this will be a general formula not numbers.)

Finally, supposing $n = 9$ and $y_2 = 1.5$, give the 95% confidence interval for a.

Solution: This is a problem about using a standardized statistic and 'pivoting' to compute confidence intervals. We are told that $y_2/a \sim \text{Beta}(2, n-1)$. If $c_{0.025}$ amd $c_{0.975}$ are the critical values for $\text{Beta}(2, n-1)$ then this means that

$$P\left(c_{0.975} < \frac{y_2}{a} < c_{0.025} \,|\, a\right) = 0.95.$$

Here, in R notation, $c_{0.975} =$ `qbeta(0.025, 2, n-1)` and $c_{0.025} =$ `qbeta(0.975, 2, n-1)`. Doing some algebra to isolate a in the middle, this becomes

$$P\left(\frac{y_2}{c_{0.025}} < a < \frac{y_2}{c_{0.975}} \,|\, a\right) = 0.95.$$

If $n = 9$ we have

$$c_{0.025} = \texttt{qbeta(0.975, 2, 8)} = 0.482, \qquad c_{0.975} = \texttt{qbeta(0.025, 2, 8)} = 0.0281$$

So, in this case, the 95% confidence interval is

$$\boxed{\left[\frac{y_2}{c_{0.025}}, \frac{1.5}{c_{0.975}}\right] \approx \left[\frac{1.5}{0.482}, \frac{1.5}{0.0281}\right] \approx [3.1, 53.3]}.$$

Problem 3. (35: 15,10,10 pts.) **Various variances**
Consider a sample of size n drawn from a Bernoulli(θ) distribution. (That is, a draw from a binomial(n, θ) distribution.) In constructing a confidence interval the conservative estimate is that the variance of the underlying Bernoulli distribution is $\sigma^2 = 1/4$ –this is conservative because for any θ we know that $\sigma^2 \le 1/4$.

(a) *In this problem we want to compare how well normal distribution using the conservative estimate matches the one using the true variance*

(i) Let $\theta = 0.5$ and $n = 250$. Make a plot that includes

- *the pmf $p(x|\theta)$ of the binomial(n, θ) distribution.*
- *the pdf of the normal distribution with the same mean and variance as the binomial$((n, \theta))$ distribution*
- *the normal distribution with the same mean (as the binomial distribution) and conservative variance to your plot.*

Note: *The conservative variance for a Bernoulli(θ) distribution is 1/4. So the conservative variance for a binomial(n, θ) is $n/4$.*

Note. *It is not reasonable to compare the probabilities in a pmf to the densities in a pdf. In order to make them comparable, but also to make the graphs readable, you should plot the pmf as points, but think of them as the top of a density histogram with bin width 1 and breaks on the half integers.*

(ii) Make a similar plot for $\theta = 0.3$ and $n = 250$.

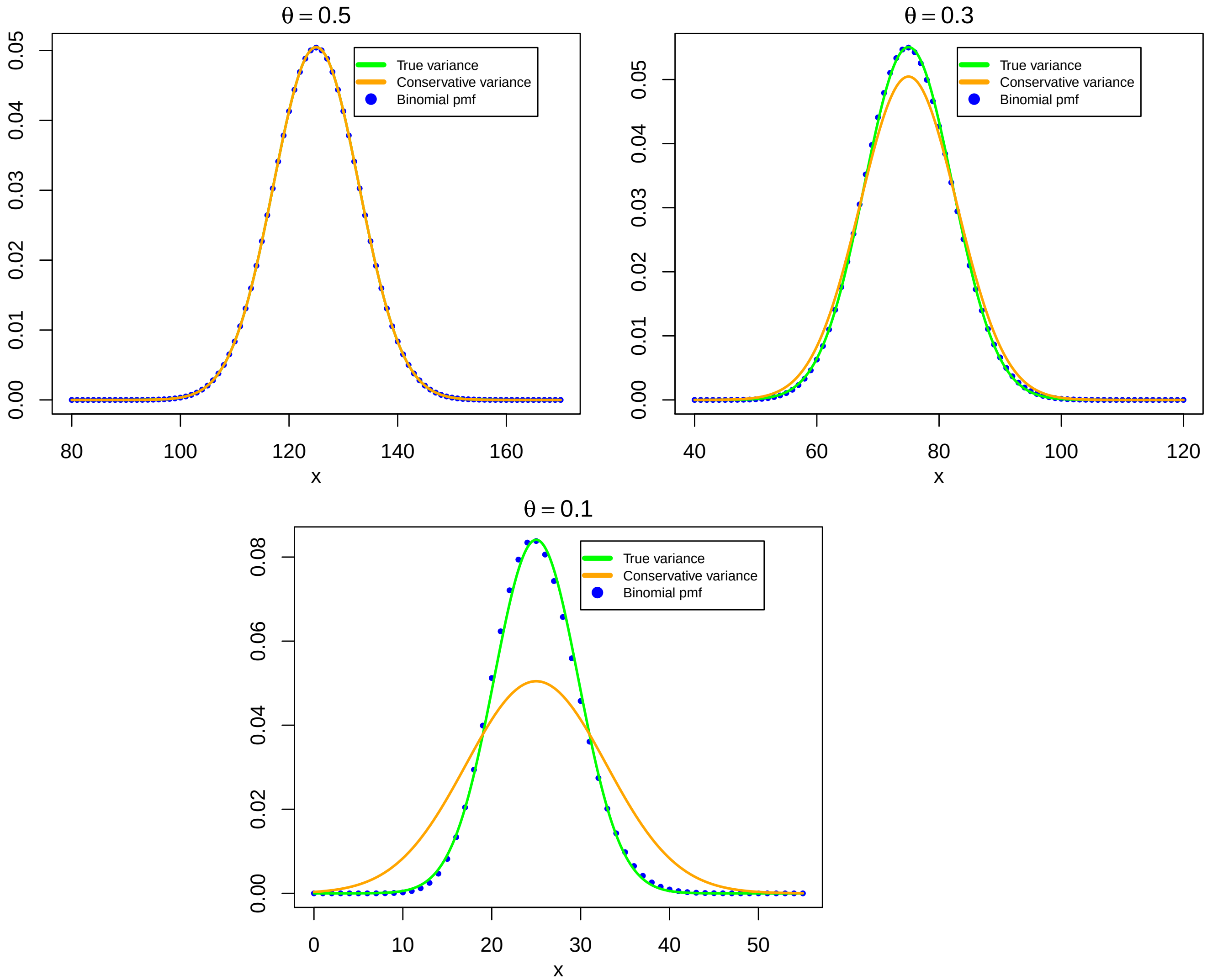
(iii) Make a similar plot for $\theta = 0.1$ and $n = 250$.

In each case, how close are each of the normal distributions to the binomial distribution? How do the two normal distributions differ? Based on your plots, for what range of θ do you think the conservative normal distribution is a reasonable approximation for binomial(n, θ) with large n?

Solution: We have $x \sim$ binomial(n, θ), so $E[X] = n\theta$ and $\text{Var}(X) = n\theta(1-\theta)$. The conservative variance is just $\frac{n}{4}$. So the distributions being plotted are

$$\text{binomial}(250,\ \theta), \quad \text{N}(250\theta,\ 250\theta(1-\theta)), \quad \text{N}(250\theta,\ 250/4).$$

Note, the whole range is from 0 to 250, but we only plotted the parts where the graphs were not essentially 0.



We notice that for each θ the blue dots lie very close to the green (true variance) curve. So the $\text{N}(n\theta, n\theta(1-\theta))$ distribution is quite close to the binomial(n, θ) distribution for each of the values of θ considered. In fact, this is true for all θ by the Central Limit Theorem. For $\theta = 0.5$ the conservative variance is the exact variance. For $\theta = 0.3$ the conservative variance works well: it has smaller peak and slightly fatter tail. For $\theta = 0.1$ the conservative approximation breaks down and is not very good.

In summary we can say two things about the conservative variance:

1. It gives good results for θ near 0.5 and breaks down for extreme values of θ.

2. Since the conservative variance overestimates the variance (the conservative graphs are shorter and wider) it gives us a confidence interval that is larger than is strictly necessary. That is a nominal 95% conservative interval actually has a greater than 95% confidence level.

(b) *Suppose θ is the probability of success, and that the result of an experiment was 140 successes out of 250 trials. Find 80% and 95% confidence intervals for θ using the conservative variance. (For the 95% interval use the rule-of-thumb that $z_{0.025} = 2$.)*

Solution: Using the conservative variance, we know that $\bar{x}$ is approximately $\mathrm{N}(\theta, 1/4n)$. For an 80% confidence interval, we have $\alpha = 0.2$ so

$$z_{\alpha/2} = \text{ qnorm(0.9,0,1)} = 1.2815.$$

So the 80% confidence interval for θ is given by

$$\left[\bar{x} - \frac{z_{0.1}}{2\sqrt{n}}, \bar{x} + \frac{z_{0.1}}{2\sqrt{n}}\right] = [0.5195, 0.6005]$$

For the 95% confidence interval, we use the rule-of-thumb that $z_{0.025} \approx 2$. So the confidence interval is

$$\left[\bar{x} - \frac{1}{\sqrt{n}}, \bar{x} + \frac{1}{\sqrt{n}}\right] = [0.497, 0.623]$$

It's okay to have used the exact value of $z_{0.025}$. This gives a confidence interval:

$$\left[\bar{x} - \frac{1.96}{2\sqrt{n}}, \bar{x} + \frac{1.96}{2\sqrt{n}}\right] = [0.498, 0.622]$$

(c) *Using the same data as in part (b), find an 80%* **posterior** *probability interval for θ using a flat prior, i.e. Beta$(1, 1)$. Center your interval between the 0.1 and 0.9 quantiles. Compare this with the 80% confidence interval in part (b).*

Hint: Use `qbeta(p, a, b)` to do the computation.

Solution: With prior Beta$(1, 1)$, if we observe x successes in 250 trials, then the posterior is Beta$(1 + x, 1 + 250 - x)$. In our case $x = 140$. So, using R we get the 80% posterior probability interval:

$$\text{prob\_interval } = [\texttt{qbeta}(0.1, 141, 111), \texttt{qbeta}(0.9, 141, 111)] \approx [0.52, 0.60]$$

This is quite close to the 80% confidence interval. Though the two intervals have **very different** technical meanings, we see that they are consistent (and numerically close). Both give a type of estimate of θ.