18.05 Problem Set 9, Spring 2018

(due outside 4-174 Monday, May 7 at 9:30 AM)

Problem 1. (12 pts.) **Z**
Suppose we have n data points drawn from a $\mathrm{N}(\mu, 5^2)$ distribution, where the value of μ is unknown.

(a) Suppose we have 16 data points and that the sample mean is $\overline{x} = 20$. Construct a precise 95% confidence interval for the mean μ.

(b) For the data in part (a) also construct 50% and 99% confidence intervals.

(c) Explain why the confidence interval for the mean μ when $n = 100$ is always narrower than the confidence interval when $n = 10$.

(d) What is the smallest value of n so that the 95% confidence interval for the mean will have width less than 1.0? (Still with $\sigma^2 = 5^2$.)

Problem 2. (18 pts.) **Mayflies**
Adult mayflies live anywhere from 30 minutes to 1 day, depending on the species. Data for one species was collected by tracking 10 mayflies. The recorded lifespans in hours were

17.68, 13.69, 11.22, 11.05, 13.86, 14.47, 14.50, 13.47, 10.04, 13.10

(a) Compute a 95% confidence interval for the mean lifetime of this species of mayfly.

(b) What assumptions did you make in part (a)?

(c) Consider the null hypothesis that the mean lifetime of this species is 14 hours versus the two-sided alternative that the mean is not 14 hours. Use your answer in part (a) say whether or not you would reject the null hypothesis at significance level 0.05.

(d) Compute a 95% confidence interval for the standard deviation of distribution of the lifetime of a mayfly.

(e) Based on the sample variance of this data estimate the number of data points you would need to make the width of the 95% confidence interval for the mean less than or equal to 1 hour.

(f) Is the value of n in part (e) guaranteed to be sufficient? Explain your reasoning.

Problem 3. (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

$$w_2 \sim \mathrm{beta}(2, n-1).$$

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

$$y_2/a \sim \text{beta}(2, n-1).$$

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.)

Problem 4. (12 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 $\sigma^2 \leq 1/4$.

(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, θ)) 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 $n/4$.

(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 think the conservative normal distribution is a reasonable approximation for binomial(n, θ) with large n?

(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 $z_{0.025} = 2$.)

(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.

Problem 5. (15 pts.) **Frequentists don't give probabilities for hypotheses**
When given a 95% confidence interval many people often mistake this to mean that there is a 95% probability 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 θ from the choices 0, 0.2, 0.4, 0.6, 0.8, 1.0.
- Draw a sample of size n from a $\mathrm{N}(\theta, 1)$ distribution.
- Compute the 95% confidence interval for θ.
- Check if θ is in the confidence interval.
- Record the result.

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

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

θ	0	0.2	0.4	0.6	0.8	1.0
$p(\theta)$	0.95	0.01	0.01	0.01	0.01	0.01

Suppose I pick a random θ from this prior distribution and draw a sample of size 100 from a $\mathrm{N}(\theta, 1)$ distribution. I do not tell you the value of θ, but I do tell you that the sample mean is $\overline{x} = 0.32$.

We both know the prior distribution for the choices, the sample size ($n = 100$) and the variance used in drawing the sample ($\sigma^2 = 1$). So we both know that the 95% confidence interval for θ is

$$\overline{x} \pm 1.96/10 = [0.124,\ 0.516].$$

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

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

Are either of these probabilities close to 95%?