Name ______________________________

Spring 2019 **STAT 587C** **Exam II**

(100 points)

Instructions:

- Full credit will be given only if you show your work.
- The questions are not necessarily ordered from easiest to hardest.
- You are allowed to use any resource except aid from another individual.
- Aid from another individual, will automatically earn you a 0.

1. SpaceX has recovered the first stage booster of its Falcon 9 rocket in 35 of 42 attempts. Assume the number of successes have a binomial distribution with an unknown true probability of success. For the following questions, assume the default prior for the true probability of success. (4 points each)

   (a) What is the posterior distribution for the true probability of success?
   Answer:

   ```
   a = b = 1
   y = 35
   n = 42
   ```

   The posterior is $\theta|y \sim Be(36, 8)$.

   (b) What is the posterior expectation for the true probability of success?
   Answer: The posterior expectation is

   $$E[\theta|y] = \frac{a+y}{a+b+n} = \frac{36}{44} = 0.8181818.$$

   (c) Provide an equal-tail 90% credible interval for the true probability of success.
   Answer:

   ```
   qbeta(c(.05,.95), a+y, b+n-y)

   ## [1] 0.7158740 0.9039095
   ```

   (d) Calculate the probability the true probability of success is greater than 0.9.
   Answer:

   ```
   1-pbeta(0.9, a+y, b+n-y)

   ## [1] 0.06066935
   ```

   (e) Are these data a random sample from a population? Explain why or why not.
   Answer: No, these data are not a random sample from a population. From one perspective, they are the entire population.

2. During the spring, the city of Ames flushes all of its fire hydrants in order to remove sediment from the water line leading to the fire hydrant. To understand how often the fire hydrants should be flushed, the city selects a random sample of the fire hydrants in Ames and measures the amount of dry sediment removed when performing the flush. In 16 hydrants, the city found a sample average of 2 kilograms (kg) of dry sediment and a sample standard deviation of 1.5 kg. For the following questions, assume the dry sediment measurements are independent and normally distributed with mean μ and variance σ^2 and assume the default prior for the mean and variance. (4 points each)

Answer:

```
n = 16
ybar = 2
s = 1.5
```

(a) State the marginal posterior for the population variance.

Answer: $\sigma^2|y \sim IG\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \overset{d}{=} IG\,(7.5, 16.875)$

(b) State the marginal posterior for the population mean.

Answer: $\mu|y \sim t_{n-1}(\overline{y}, s^2/n) \overset{d}{=} t_{15}(2, 0.140625)$

(c) Provide an 80% credible interval for the population mean.

Answer:

```
a <- 0.2
ybar + c(-1,1)*qt(1-a/2, df = n-1)*s/sqrt(n)

## [1] 1.497273 2.502727
```

(d) Let $\tilde{Y}$ be the amount of dry sediment in an unmeasured fire hydrant water line. The predictive distribution for $\tilde{Y}$ is $t_{n-1}(\overline{y}, s^2[1+1/n])$. Determine the probability this amount will be less than 3 kg.

Answer:

$$P\left(\tilde{Y} < 3|y\right) = P\left(\frac{\tilde{Y}-\overline{y}}{s\sqrt{1+1/n}} < \frac{3-\overline{y}}{s\sqrt{1+1/n}}\middle| y\right) = P\left(T_{n-1} < \frac{3-\overline{y}}{s\sqrt{1+1/n}}\right)$$

```
pt((3-ybar)/(s*sqrt(1+1/n)), df = n-1)

## [1] 0.7362202
```

(e) Explain why the normal distribution may not be a very good model for these data.

Answer: With the observed mean and standard deviation there is a relatively high probability of having negative amounts of dry sediment, e.g.

$$P\left(\tilde{Y} < 0\right) = 0.1076977.$$

3. Iowa State University researchers have developed a method of testing soil health using tea bags buried in an agricultural field. After 30 days, the tea bag is recovered, dried, and weighed. Lower weight indicates healthier soil due to tea leaves decomposing. The file `tea.csv` contains measurements of tea bag weights in grams (g) from agricultural fields that have prairie strips, a treatment that is designed to increase soil health. For the following questions, assume the tea bag weights are independent and normally distributed with mean μ and variance σ^2. (4 points each)

   Answer:

   ```
   tea = read.csv("tea.csv")
   n = nrow(tea)
   ```

   (a) Calculate the maximum likelihood estimator for the population mean of tea bag weights.

   Answer:

   ```
   mean(tea$weight)

   ## [1] 1.153333
   ```

   (b) Calculate the maximum likelihood estimator for the population variance in tea bag weights.

   Answer:

   ```
   var(tea$weight)*(n-1)/n

   ## [1] 0.05979722
   ```

   (c) Construct a 99% confidence interval for the mean tea bag weight.

   Answer:

   ```
   t.test(tea$weight, conf.level = 0.99)$conf.int

   ## [1] 1.010190 1.296477
   ## attr(,"conf.level")
   ## [1] 0.99
   ```

   (d) From previous experience, the researchers know the mean weight of tea bags in standard agricultural fields is 1.2 g. Calculate a p-value for the null hypothesis $H_0 : \mu \geq 1.2$.

   Answer:

   ```
   t.test(tea$weight, mu = 1.2, alternative = "less")$p.value

   ## [1] 0.1847837
   ```

   (e) If researchers believe fields with prairie strips will have healthier soil than standard fields, is the null hypothesis in the previous question the appropriate null hypothesis? Explain why or why not.

   Answer: Yes, it is appropriate since standard fields have mean tea bag weight of 1.2 g and healthy soil has lower weight. Thus, we would expect the fields with prairie strips should have mean weight less than 1.2 g which would be the alternative hypothesis.

4. Let $Y_i \overset{ind}{\sim} Bin(n_i, \theta_i)$ for $i = 1, 2$. For the following questions, assume the data are random. (5 points each)

(a) Calculate

$$E\left[\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right].$$

Answer:

$$\begin{aligned} E\left[\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right] &= \frac{E[Y_1]}{n_1} - \frac{E[Y_2]}{n_2} \\ &= \frac{n_1\theta_1}{n_1} - \frac{n_2\theta_2}{n_2} \\ &= \theta_1 - \theta_2 \end{aligned}$$

(b) Calculate

$$Var\left[\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right].$$

Answer:

$$\begin{aligned} Var\left[\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right] &= \frac{Var[Y_1]}{n_1^2} + \frac{Var[Y_2]}{n_2^2} \\ &= \frac{n_1\theta_1(1-\theta_1)}{n_1^2} + \frac{n_2\theta_2(1-\theta_2)}{n_2^2} \\ &= \frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2} \end{aligned}$$

(c) Calculate a standard error for

$$\frac{Y_1}{n_1} - \frac{Y_2}{n_2}.$$

Answer: Let $\hat{\theta}_1 = y_1/n_1$ and $\hat{\theta}_2 = y_2/n_2$, then

$$\begin{aligned} SE\left[\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right] &= \sqrt{Var\left[\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right]} \\ &= \sqrt{\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}} \\ &= \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}} \end{aligned}$$

(d) Provide a formula for computing an approximate 100(1-a)% confidence interval for $\theta_1 - \theta_2$?

Answer: Using the CLT, an approximate 100(1-a)% confidence interval is

$$\hat{\theta}_1 - \hat{\theta}_2 \pm z_{a/2} SE\left[\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right]$$

where these quantities are defined in the previous answer.

5. The US FDA is currrently overseeing a clinical trial for the drug *selonsertib* which is aimed at patients who have late-stage fatty liver disease. Patients at Marshall University Hospital in Huntington, West Virginia who have late-stage fatty liver disease can enroll in the clinical trial and will be randomly assigned either selonsertib or placebo (a sugar pill). Patients who enroll will have their *aspartate transaminase* levels measured before starting the pill regiment and again one year later. Doctors will compare how much these levels change for the selonsertib group compared with the placebo group.

   (a) Describe the population being studied. (3 points)

   Answer: There are multiple possible answers here, but likely the population of interest is all US patients with late-stage fatty liver disease.

   (b) Describe the sample. (3 points)

   Answer: Individuals at Marshall University Hospital in Huntington, West Virginia who have late-stage fatty liver disease and enroll in the clinical trial.

   (c) Is this a random sample? Explain why or why not. (2 points)

   Answer: No, clearly not random as individuals need to opt in to the trial and the individuals are from one hospital.

   (d) Describe a reasonable model for these data to address the scientific question of interest. (12 points)

   Answer: Let D_{ig} be the difference in aspartate transaminase (end of the year minus the beginning of the year) for individual i in treatment group g where $g = 1, 2$ for control/placebo and treatment group respectively and $i = 1, \ldots, n_g$. Then $D_{ig} \overset{ind}{\sim} N(\mu_g, \sigma_g^2)$ although you could assume $\sigma_1 = \sigma_2$.