Name _____

Spring 2021 **STAT 587-2** **Exam II**
(45 points)

Instructions:

- You are allowed to use any resource except aid from another individual.
- Aid from another individual, will automatically earn you a 0.

Answer:

```
library("tidyverse")

## -- Attaching packages --------------------------------------- tidyverse 1.3.0
--
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts()
--
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

1. For the following questions, determine the most appropriate model among the following:

   - binomial: $Y \sim Bin(n, \theta)$
   - normal: $Y \sim N(\mu, \sigma^2)$
   - paired normal: $Y_1 - Y_2 \sim N(\mu, \sigma^2)$
   - two binomial: $Y_g \sim Bin(n_g, \theta_g)$ for $g = 1, 2$
   - two normal: $Y_{g,i} \sim N(\mu_g, \sigma_g^2)$ for $g = 1, 2$

   (a) The square footage of houses in Ames.
   Answer: normal

   (b) The number of apple trees in an orchard that started budding this past weekend.
   Answer: binomial

   (c) Comparing the number of planes requiring a repair within one year for planes that were repaired in Atlanta versus those repaired in Minneapolis.
   Answer: two binomial

   (d) Comparing the run times of a set of algorithms where each algorithm is run twice: once with subroutine A and once with subroutine B.
   Answer: paired normal

   (e) Comparing drying times for concrete made with two different ratios of small to large aggregates.
   Answer: two normal

2. Determine what is known about the two-sided p-value from the two-sided confidence interval. For each of these questions consider the data model $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$, the p-value (p) obtained from testing $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$, and the two-sided $100(1-a)\%$ confidence interval (CI) obtained from the formula:

$$\overline{y} \pm t_{1-a/2,n-1} s/\sqrt{n}.$$

Determine what can be said about the p-value.

(a) The 95% CI contains 0.
Answer: $p > 0.05$

(b) The 90% CI does not contain 0.
Answer: $p < 0.1$

(c) The 80% CI has 0 as one of its endpoints.
Answer: $p = 0.2$

(d) The 70% CI does not contain 0.
Answer: $p < 0.3$

(e) The 85% CI contains 0.
Answer: $p > 0.15$

(f) The upper endpoint of a 60% CI is 0.
Answer: $p = 0.4$

3. A simulation study is performed to understand the win probability of a game of soltaire, e.g. `https://solitaired.com/`. In the study, the player won 15 out of 70 games.

   Answer:

```
y = 15
n = 70
theta_hat = y/n
```

   (a) Determine the MLE for the win probability.
   Answer: $\hat{\theta} = 15/70 = 0.2142857$

   (b) Construct an equal-tailed 90% confidence interval based on the CLT.
   Answer:

```
a = 0.1
```

      i. Determine the critical value.
      Answer:

```
crit_value = qnorm(1-a/2)
```

      $z_{1-a/2} = 1.6448536$

      ii. Determine the standard error.
      Answer:

```
se = sqrt(theta_hat*(1-theta_hat)/n)
```

      $\sqrt{\hat{\theta}(1-\hat{\theta})/n} = 0.0490433$

      iii. Determine the CI endpoints.
      Answer:

```
ci = theta_hat + c(-1,1) * crit_value * se
```

      Thus 90% CI is $(0.1336166, 0.2949548)$.

   (c) Does the CLT provide a reasonable CI? Why or why not?
   Answer: Yes, n is relatively large and θ does not appear to be too close to 0 or 1.

4. The Ames Water Treatment plant routinely tests the incoming water for total hardness. The following is output from the analysis for February 2021 where hardness is measured in parts per million (ppm).

Answer:

```
set.seed(20210329)
y = rnorm(40, mean = 416, sd = sqrt(1000))
t = t.test(y)
```

```
t.test(y)


##
##  One Sample t-test
##
## data:  y
## t = 87.897, df = 39, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  417.2500 436.9058
## sample estimates:
## mean of x
##  427.0779
```

Answer: For the following questions, we will use the significant digits shown on the output.

```
n = 39+1
sample_mean = 427.0779
sample_sd = sample_mean / (87.897 / sqrt(n)) # from formula for t-statistic, see bel
```

Answer the following questions based on this output.

(a) Determine the number of samples measured.
Answer: 40

(b) Determine the sample average.
Answer: 427.0779

(c) Determine the sample standard deviation.

Answer: Recall that, by default,

$$t = (\overline{y} - 0)/(s/\sqrt{n}) \qquad \text{and thus} \qquad s = \overline{y}/(t/\sqrt{n}).$$

```
sample_sd
```

```
## [1] 30.73003
```

You could also determine this using the CI and p-value (although the pvalue won't be very accurate).

(d) Using our default prior, construct a 80% credible interval for the true mean hardness.

Answer: The formula is

$$\overline{y} \pm t_{(n-1,1-a/2)} s/\sqrt{n}.$$

```
a = 0.2
sample_mean + c(-1,1) * qt(1-a/2, df = n-1) * sample_sd / sqrt(n)
```

```
## [1] 420.7437 433.4121
```

(e) Using our default prior, determine the posterior probability that the true mean hardness is less than 430 ppm.

Answer:

```
t = (430 - sample_mean) / (sample_sd / sqrt(n))
```

Recall that $\mu|y \sim t_{n-1}(\overline{y}, s^2/\sqrt{n})$. Calculate

$$\begin{aligned} P(\mu < 430|y) &= P\left(\frac{\mu-\overline{y}}{s/\sqrt{n}} < \frac{430-427.0779}{30.7300341/\sqrt{40}} \middle| y\right) \\ &= P(T_{n-1} < 0.6013981|y) \\ &= 0.7244726 \end{aligned}$$

5. The file `nursing_homes.csv` contains data from a random sample of nursing homes in Iowa and Minnesota and whether or not each nursing home had an outbreak of COVID-19. Answer the following questions based on the data in this file.

   Answer: This code creates the data set

```
set.seed(20210329)
n = c(96,55)
d = rbind(data.frame(state = "Iowa",
                     home = 1:n[1],
                     outbreak = sample(c("Yes","No"),
                                       size = n[1],
                                       replace = TRUE,
                                       prob = c(.7,.3))),
          data.frame(state = "Minnesota",
                     home = 1:n[2],
                     outbreak = sample(c("Yes","No"),
                                       size = n[2],
                                       replace = TRUE,
                                       prob = c(.6,.4))))

write.csv(d, file = "nursing_homes.csv", row.names = FALSE)
```

   This code reads the data

```
d = read.csv("nursing_homes.csv")
s = d %>%
  group_by(state) %>%
  summarize(n = n(),
            y = sum(outbreak == "Yes"),
            p = y/n)
s
```

```
## # A tibble: 2 x 4
##   state         n     y     p
## * <chr>     <int> <int> <dbl>
## 1 Iowa         96    56 0.583
## 2 Minnesota    55    35 0.636
```

   (a) How many observations of nursing homes in Iowa are there?

   Answer:

```
s %>% filter(state == "Iowa") %>% select(n)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    96
```

(b) In our data, how many nursing homes in Iowa had an outbreak?

Answer:

```
s %>% filter(state == "Iowa") %>% select(y)
```

```
## # A tibble: 1 x 1
##       y
##   <int>
## 1    56
```

(c) In our data, what is the proportion of nursing homes in Iowa that had an outbreak?

Answer:

```
s %>% filter(state == "Iowa") %>% select(p)
```

```
## # A tibble: 1 x 1
##       p
##   <dbl>
## 1 0.583
```

(d) Using `prop.test`, find the p-value for a test of equality of the proportion of outbreaks in Iowa compared to Minnesota.

Answer:

```
prop.test(s$y, s$n)$p.value
```

```
## [1] 0.6397671
```

(e) Based on this p-value and a significance level of 0.05, what conclusion would you make?

Answer: Fail to reject the null hypothesis since $p > 0.05$.

(f) Using `prop.test`, find the equal-tail 90% confidence interval for the difference in proportion of outbreaks in Iowa compared to Minnesota (Iowa - Minnesota).

Answer:

```
prop.test(s$y, s$n, conf.level = 0.9)$conf.int
```

```
## [1] -0.20235982  0.09629922
## attr(,"conf.level")
## [1] 0.9
```