Name ____________________

Spring 2018	**STAT 401C**	**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. Please answer the following questions using your own words. If I find that you have copied your answers from the internet, you will receive a 0.

   (a) What is the population? (5 points)
   Answer: All the units being studied.

   (b) What is a sample? (5 points)
   Answer: The subset of the population that you actually have data on.

   (c) For the model $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$ name two statistics. (5 points)
   Answer: sample mean ($\overline{y}$) and sample standard deviation (s) although there are many others that you could have answered, e.g. quantiles

   (d) Describe what it means for an estimator to be unbiased. (5 points)
   Answer: An estimator $\hat{\theta}$ is unbiased for a population parameter θ if $E[\hat{\theta}] = \theta$ where the expectation is taken with respect to the data being random.

2. Iowa State University has a goal that 95% of PhD Engineering students are employed full time within 6 months of graduation. Of the 69 individuals who graduated in the spring of 2016 with a PhD in Engineering, 67 of those individual were employed full time within 6 months of graduation.

   (a) State an appropriate model for these data. (5 points)

   Answer: Let Y be the number of employed students. Then a reasonable model is $Y \sim Bin(n, \theta)$ where $n = 69$ and $y = 67$.

   (b) Compute an exact pvalue for the hypothesis test with null hypothesis that the 6-month employment probability is less than or equal to 95% versus the alternative that the 6-month employment probability is greater than 95%. (5 points)

   Answer:

```
y <- 67
n <- 69
binom.test(y, n, p=.95, alternative="greater")$p.value

## [1] 0.3231703
```

   (c) Using the default prior, calculate the posterior probability that the 6-month employment probability is greater than 95%. (5 points)

   Answer:

```
a <- b <- 1
1-pbeta(.95, a+y, b+n-y)

## [1] 0.6862642
```

   (d) Using the default prior, calculate an upper one-sided 99% credible interval. (5 points)

   Answer:

```
qbeta(c(.01,1), a+y,b+n-y)

## [1] 0.8852796 1.0000000
```

3. A research group at Iowa State believes they have developed a material just as hard as diamond. They tested 9 randomly chosen samples of this material. The sample average Vickers hardness was 8 Mg/mm^2 and the sample standard deviation was 4 Mg/mm^2. Assume the samples are independent, normally distributed with a common mean and variance. Also assume the default prior.

   (a) Calculate a 95% credible interval for the true mean hardness. (5 points)

   Answer:

```
n = 9
ybar = 8
s = 4
ybar + qt(c(.025,.975),n-1)*s/sqrt(n)

## [1]  4.925328 11.074672
```

   (b) Calculate a 95% credible interval for the true standard deviation of hardness. (5 points)

   Answer:

```
1/sqrt(qgamma(c(.975,.025), (n-1)/2, s^2*(n-1)/2))

## [1] 2.701828 7.663084
```

   (c) Diamond is known to have a Vickers hardness of 10 Mg/mm^2. Calculate the posterior probability the mean hardness of this material is greater than that of diamond. (5 points)

   Answer:

```
1-pt((10-ybar)/(s/sqrt(n)), n-1)

## [1] 0.08600165
```

   (d) The predictive distribution for a new observation $\tilde{Y}$ is $\tilde{Y} \sim t_{n-1}(\overline{y}, s^2(1+1/n))$. Calculate the probability a new observation is over 10 Mg/mm^2. (5 points)

   Answer:

$$
\begin{aligned}
P(\tilde{Y} > 10) &= P\left(\frac{\tilde{Y}-\overline{y}}{s\sqrt{1+1/n}} > \frac{10-8}{4\sqrt{1+1/9}}\right) \\
&= P(T_{n-1} > 0.4743416) \\
&= 1 - P(T_{n-1} < 0.4743416)
\end{aligned}
$$

```
1-pt((10-ybar)/(s*sqrt(1+1/n)), df = n-1)

## [1] 0.3239668
```

4. In an attempt to slow down traffic, police installed a stationary radar gun with a sign that displays the speed of cars approaching it. Prior to placing the sign, police randomly selected 100 cars passing that position and recorded 65 cars in excess of the posted speed limit. After placing the sign, police randomly selected 100 cars passing the same position and recorded 55 cars in excess of the posted speed limit.

   (a) Calculate an approximate pvalue for the hypothesis test with alternative hypothesis that the true population proportion decreased after the sign was posted. (5 points)

   Answer:

```
y <- c(65,55)
n <- c(100,100)
prop.test(y,n, alternative="greater")$p.value                # or

## [1] 0.09696543

prop.test(y,n, alternative="greater", correct=FALSE)$p.value # or

## [1] 0.07445734

1 - pnorm((0.65-0.55)/sqrt(0.65*(1-0.65)+0.55*(1-0.55))*sqrt(100))

## [1] 0.07339654
```

   (b) Calculate an approximate 90% confidence interval for the decrease in true population proportion after the sign was posted. (5)

   Answer:

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

## [1] -0.02336371  0.22336371
## attr(,"conf.level")
## [1] 0.9

prop.test(y,n, conf.level=0.9, correct=FALSE)$conf.int

## [1] -0.01336371  0.21336371
## attr(,"conf.level")
## [1] 0.9

(0.65-0.55) + c(-1,1) * qnorm(0.95) * sqrt(0.65*(1-0.65)+0.55*(1-0.55))/sqrt(100)

## [1] -0.01336371  0.21336371
```

   (c) Calculate an approximate posterior probability that the true population proportion decreased after the sign was posted. (5 points)

   Answer: Assume $Y_i \overset{ind}{\sim} Bin(n_i, \theta_i)$ and $\theta_i \overset{ind}{\sim} Be(a, b)$ with $a = b = 1$.

```
a <- b <- 1
nr <- 1e5
diff <- rbeta(nr, a+y[1], b+n[1]-y[1]) - rbeta(nr, a+y[2], b+n[2]-y[2])
mean(diff>0)
```

```
## [1] 0.92277
```

(d) Calculate an approximate 90% credible interval for the decrease in true population proportion after the sign was posted. (5 points)

Answer:

```
quantile(diff, probs = c(.05,.95))

##          5%         95%
## -0.01459887  0.20914547
```

5. The file `carbon_dating.csv` contains carbon dating ages (in years before present time) on replicate measurements wood. For the following questions, let $Y_i \stackrel{ind}{\sim} N(\mu, \sigma^2)$ where Y_i is the carbon dated age of the i replicate piece of wood.

   (a) Calculate the maximum likelihood estimator for μ. (5 points)
   Answer:

```
d <- read.csv("carbon_dating.csv")
mean(d$carbon_age)

## [1] 4504.4
```

   (b) Calculate the maximum likelihood estimator for σ^2. (5 points)
   Answer:

```
n <- length(d$carbon_age)
var(d$carbon_age)*(n-1)/n

## [1] 823.04
```

   (c) Calculate a 99% confidence interval for μ. (5 points)
   Answer:

```
d <- read.csv("carbon_dating.csv")
t.test(d$carbon_age, conf.level = 0.99)$conf.int

## [1] 4473.322 4535.478
## attr(,"conf.level")
## [1] 0.99
```

   (d) Calculate a pvalue for the hypothesis test $H_0 : \mu \leq 4500$ vs $H_A : \mu > 4500$. (5 points)
   Answer:

```
t.test(d$carbon_age, mu=4500, alternative="greater")$p.value

## [1] 0.3281749
```