Swarthmore College Honors Exam
Statistics
Spring 2023

Instructions to students:

- Recall you can use your laptop during this exam for the sole purpose of using R-studio to analyze the data associated with this exam.
- You will receive an email at 9 am from Usha Jenemann ((ujenema1@swarthmore.edu) that contains the data you will need for the exam. Please open that and move it to your laptop for the duration of the exam.
- At the end of the exam:
  - either create pdfs or take snapshots of any tables or other produced-work that you want to submit to the examiner;
  - email that to Usha Jenemann (ujenema1@swarthmore.edu) so that it can be included with your written work.
  - Hand in your completed green book and any scratch paper to the proctor.

Paired t-test: Let $(x_1, y_1), \ldots, (x_n, y_n)$ be pairs of values. Under certain assumptions the means, μ_X and μ_Y, can be compared using the test statistic value

$$t = \frac{\overline{d}}{s_D/\sqrt{n}}$$

where $\overline{d}$ is the average of $x_1 - y_1, \ldots, x_n - y_n$, and s_D is the sample standard deviation for these differences. Use `t.test(x,y,paired=T)` or `t.test(x-y)` with `x=c(`$x_1, \ldots, x_n$`)` and `y=c(`$y_1, \ldots, y_n$`)` to see that the degrees of freedom is $n-1$. A 95% confidence interval for $\mu_X - \mu_Y$ is

$$(\overline{d} - t_* \, s_D/\sqrt{n}, \ \overline{d} + t_* \, s_D/\sqrt{n}).$$

Compute t_* using `qt(.975,`$n-1$`)`.

(Pooled) two-sample t-test: Let $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_n)$ be values from distributions with means μ_X and μ_Y and the same variance σ^2. Under certain assumptions the means can be compared using the test statistic value

$$t = \frac{\overline{x} - \overline{y}}{s_P\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

where ${s_p}^2 = \frac{(m-1){s_X}^2 + (n-1){s_Y}^2}{m+n-2}$ is the pooled estimate of σ^2. Use `t.test(x,y,var.equal=T)` with `x=c(`$x_1, \ldots, x_m$`)` and `y=c(`$y_1, \ldots, y_n$`)` to see that the degrees of freedom is $m+n-2$. A 95% confidence interval for $\mu_X - \mu_Y$ is

$$\left(\overline{x} - \overline{y} - t_* \, s_P\sqrt{\frac{1}{m} + \frac{1}{n}}, \ \overline{x} - \overline{y} + t_* \, s_P\sqrt{\frac{1}{m} + \frac{1}{n}}\right).$$

Compute t_* using `qt(.975,`$m+n-2$`)`.

Welch two-sample t-test: Let $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_n)$ be values from distributions with means μ_X and μ_Y. Under certain assumptions the means can be compared using the test statistic value

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{{s_X}^2}{m} + \frac{{s_Y}^2}{n}}}.$$

Use `t.test(x,y)` with `x=c(`$x_1, \ldots, x_m$`)` and `y=c(`$y_1, \ldots, y_n$`)` to see that the approximate degrees of freedom is given by

$$\nu = \frac{\left({s_X}^2/m + {s_Y}^2/n\right)^2}{\frac{({s_X}^2/m)^2}{m-1} + \frac{({s_Y}^2/n)^2}{n-1}}.$$

An approximate 95% confidence interval for $\mu_X - \mu_Y$ is

$$\left(\overline{x} - \overline{y} - t_* \sqrt{\frac{{s_X}^2}{m} + \frac{{s_Y}^2}{n}}, \ \overline{x} - \overline{y} + t_* \sqrt{\frac{{s_X}^2}{m} + \frac{{s_Y}^2}{n}}\right).$$

Compute t_* using `qt(.975,`ν`)`.

Suppose `filename.txt` and `filename.csv` are files in the `Downloads` folder of `username`.

`setwd("~/Downloads")` sets your working directory on a Mac
`setwd("C:/Users/username/Downloads")` sets your working directory on a PC
`data=read.table("filename.txt", header=T)` reads a plain text (ASCII) file into R
`data=read.csv("filename.csv")` reads a comma separated values file into R

Suppose $x_1, \ldots, x_n$ is a list of n numbers, and suppose d is a number:

`x=c(`$x_1, \ldots, x_n$`)` has the property that `x[i]` is the number x_i, for $1 \le i \le n$
`length(x)` is n, the length of the list $x_1, \ldots, x_n$
`y=c(`$y_1, \ldots, y_n$`)` has the property that `y[i]` is the number y_i, for $1 \le i \le n$
`x[-m]` is the list of numbers $x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_n$
`c(x,y)` is the list of numbers $x_1, \ldots, x_n, y_1, \ldots, y_n$
`cbind(x,y)` is the list of pairs $(x_1, y_1), \ldots, (x_n, y_n)$
`quantile(x)`, for n odd, lists the min, Q1, median, Q3, and max
`quantile(x,type=2)`, for n even, lists the min, Q1, median, Q3, and max
`1:n` or `seq(1,n)` is the list of numbers $1, 2, \ldots, n$, but `seq(1,n)` has other options
`d*x` is the list of numbers $dx_1, \ldots, dx_n$
`x/d` is the list of numbers $\frac{x_1}{d}, \ldots, \frac{x_n}{d}$
`x+d` is the list of numbers $x_1 + d, \ldots, x_n + d$
`log(x)` is the list of numbers $\ln(x_1), \ldots, \ln(x_n)$
`exp(x)` is the list of numbers $e^{x_1}, \ldots, e^{x_n}$
`sum(x)` is the number $x_1 + \cdots + x_n$
`sample(x,m)` is a list $x_{j_1}, \ldots x_{j_m}$ where $j_1, \ldots, j_m$ are chosen (randomly) from $1, \ldots, n$.
`order(x)` is the list $j_1, \ldots, j_n$ where $x_{j_1}, \ldots, x_{j_n}$ are $x_1, \ldots, x_n$ from smallest to largest
`hist(x)` displays a histogram for the numbers $x_1, \ldots, x_n$
`boxplot(x)` displays a boxplot for the numbers $x_1, \ldots, x_n$
`barplot(x)` displays side-by-side bars of heights $x_1, \ldots, x_n$
`barplot(cbind(x))` displays stacked bars whose heights are $x_1, \ldots, x_n$ from bottom to top
`qqnorm(x)` displays a normal quantile plot for the numbers $x_1, \ldots, x_n$
`mean(x)` is the average $\overline{x}$ (the sample mean) of the numbers $x_1, \ldots, x_n$
`var(x)` is the sample variance $\frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1}$
`sd(x)` is the sample standard deviation, which is the square root of the sample variance

Suppose `y=c(`$y_1, \ldots, y_n$`)`, where $y_1, \ldots, y_n$ is a list of numbers:

`x*y` is the list of numbers $x_1 y_1, \ldots, x_n y_n$
`x/y` is the list of numbers $\frac{x_1}{y_1}, \ldots, \frac{x_n}{y_n}$, provided $y_1, \ldots, y_n$ are all nonzero
`x+y` is the list of numbers $x_1 + y_1, \ldots, x_n + y_n$
`boxplot(x,y)` displays side-by-side boxplots for $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$
`c(x,y)` is the list of numbers $x_1, \ldots, x_n, y_1, \ldots, y_n$
`cbind(x,y)` is the table with $x_1, \ldots, x_n$ in the first column and $y_1, \ldots, y_n$ in the second
`cbind(x,y)[i,]` is the list of numbers, x_i and y_i, in the i^{th} row of the table
`cor(x,y)` is the sample correlation $r = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}}$

Ordinary least squares fits a line $\hat{f}(x) = b_0 + b_1 x$ to data points $(x_1, y_1), \ldots, (x_n, y_n)$

`plot(x,y)` displays the data points a coordinate plane
`lm(y~x)` is the list of the intercept b_0 and slope b_1 of the fitted line
`summary(lm(y~x))` gives information about the linear model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for $1 \leq i \leq n$
`summary(lm(y~x))$coefficients[1,1]` is the estimate b_0 of the (true) intercept β_0
`summary(lm(y~x))$coefficients[2,1]` is the estimate b_1 of the (true) slope β_1
`summary(lm(y~x))$r.squared` is the square r^2 of the sample correlation r
`fitted(lm(y~x))` is the list of numbers $b_0 + b_1 x_1, \ldots, b_0 + b_1 x_n$
`resid(lm(y~x))` is the list of numbers $y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n$, where $\hat{y}_i = b_0 + b_1 x_i$ for $1 \leq i \leq n$

Logical values `T` or `TRUE` and `F` or `FALSE` are interpreted as 1 or 0 when added or multiplied

`T==T` is `T`, `T==F` is `F`, `F==T` is `F`, and `F==F` is `T`
`T&T` is `T`, `T&F` is `F`, `F&T` is `F`, and `F&F` is `T`, so interpret "`&`" as "and"
`T|T` is `T`, `T|F` is `T`, `F|T` is `T`, and `F|F` is `F`, so interpret "`|`" as "or"

Suppose $k_1, \ldots, k_n$ and $\ell_1, \ldots, \ell_n$ are lists of logical values

`k=c(`$k_1, \ldots, k_n$`)` has the property that `k[i]` is the logical value k_i, for $1 \leq i \leq n$
ℓ`=c(`$\ell_1, \ldots, \ell_n$`)` has the property that ℓ`[i]` is the logical value ℓ_i, for $1 \leq i \leq n$
`k[-m]` is the list of logical values $k_1, \ldots, k_{m-1}, k_{m+1}, \ldots, k_n$
`k&`ℓ is the list of logical values k_1`&`ℓ_1, … k_n`&`ℓ_n.
`k==`ℓ is the list of logical values k_1`==`ℓ_1, … , k_n`==`ℓ_n

Suppose `x` and `y` are lists of numbers or words, and d is a number or word

`rep(d,n)` is the list $d, \ldots, d$ of length n
`x==d` is the list of logical values $k_1, \ldots, k_n$ where k_i is `T` if $x_i = d$, and k_i is `F` if $x_i \neq d$
`x>d` is the list of logical values $k_1, \ldots, k_n$ where k_i is `T` if $x_i > d$, and k_i is `F` if $x_i \not> d$
`x>=d` is the list of logical values $k_1, \ldots, k_n$ where k_i is `T` if $x_i \geq d$, and k_i is `F` if $x_i \not\geq d$
`x<d` is the list of logical values $k_1, \ldots, k_n$ where k_i is `T` if $x_i < d$, and k_i is `F` if $x_i \not< d$
`x<=d` is the list of logical values $k_1, \ldots, k_n$ where k_i is `T` if $x_i \leq d$, and k_i is `F` if $x_i \not\leq d$
`x==y` is the list of logical values $k_1, \ldots, k_n$ where k_i is `T` if $x_i = y_i$, and k_i is `F` if $x_i \not\leq y_i$
`x>y` is the list of logical values $k_1, \ldots, k_n$ where k_i is `T` if $x_i > y_i$, and k_i is `F` if $x_i \not> y_i$
`x>=y` is the list of logical values $k_1, \ldots, k_n$ where k_i is `T` if $x_i \geq y_i$, and k_i is `F` if $x_i \not\geq y_i$
`x<y` is the list of logical values $k_1, \ldots, k_n$ where k_i is `T` if $x_i < y_i$, and k_i is `F` if $x_i \not< y_i$
`x<=y` is the list of logical values $k_1, \ldots, k_n$ where k_i is `T` if $x_i \leq y_i$, and k_i is `F` if $x_i \not\leq y_i$

Suppose $x_1, \ldots, x_n$ are values of a quantitative random variable; and suppose $v_1, \ldots, v_n$ and $w_1, \ldots, w_n$ are levels of two categorical random variables. For $1 \leq i \leq n$, suppose x_i is the value of the quantitative variable for one of n randomly selected individuals, and v_i and w_i are the levels of the categorical variables for that same individual. If $v_1, \ldots, v_n$ are character strings, R will know that they are values of a categorical random variable. Otherwise use `as.factor(c(`v_1`,...,`v_n`))` instead of `c(`v_1`,...,`v_n`)` in the commands below.

`table(c(`$v_1, \ldots, v_n$`))` is a list of the number of times each level among $v_1, \ldots, v_n$ occurs
`pie(table(c(`$v_1, \ldots, v_n$`)))` has a slice for each level, sized by the number of times it occurs

`boxplot(c(`$x_1, \ldots, x_n$`)~c(`v_1`,...,`v_n`))` displays boxplots for the levels among $v_1, \ldots, v_n$; the boxplot for a level is the boxplot for values x_i such that (x_i, v_i) has v_i equal to that level. From left to right, boxplots are displayed in alphabetical (or numerical) order of the levels `table(c(`$v_1,\dots,v_n$`),c(`$w_1,\dots,w_n$`))` is a table each of whose entries is the number of pairs (v_i, w_i), for $1 \le i \le n$, where v_i is the level for that row and w_i is the level for that column.

Suppose `m` is a table with r rows and c columns, where the entries in each column are either all numbers, all character strings, or all logical values:

`m[i,j]` is the entry in row i and column j, for $1 \le i \le r$ and $1 \le j \le c$
`m[i,]` or `m[i,1:c]` is the list of entries in row i
`m[,j]` or `m[1:r,j]` is the list of entries in column j
`t(m)` is the table with c rows and r columns whose j^{th} row is the j^{th} column of `m`

Suppose the entries of the table `m` are numbers and `z=c(`$z_1,\dots,z_r$`)` is a list of numbers:

`sum(m)` is the sum of all of the entries in the table
`rowSums(m)` is the list $x_1,\dots,x_r$ where x_i is the sum of the entries in row i of the table `m`
`m/z` has entry `m[i,j]/`z_i in its i^{th} row and j^{th} column, where $1 \le i \le r$ and $1 \le j \le c$
`rowMeans(m)` is the list $x_1,\dots,x_r$ where x_i is the average of the entries in row i of `m`
`barplot(m)` displays c side-by-side stacked barplots; the heights, from bottom to top, of the bars in the j^{th} stacked barplot are `m[1,j]`, …, `m[r,j]`.

Suppose Z is a standard normal random variable and x is a number.

`dnorm(z)` is the density function f_Z evaluated at z; that is $\frac{1}{\sqrt{2\pi}}e^{-z^2/2}$
`pnorm(z)` is $P(Z \le z) = \int_{-\infty}^{z} f_Z(t)dt$, the area to the left of the vertical line through z
`qnorm(p)`, for $0 < p < 1$ is the value of z such that $P(Z \le z) = p$
`rnorm(n)` displays data values $x_1,\dots,x_n$ for n iid random standard normal random variables

Commands can be modified using options: `help(command)` displays documentation. So `help(pnorm)` gives options for `pnorm` to find $P(X \le x)$ for any normally distributed random variable X. Documentation for exponential distributions is at `help(pexp)`, for lognormal distributions is at `help(plnorm)`, and for binomial distributions is at `help(pbinom)`.

Miscellaneous commands and options

`dir()` lists the files in the working directory you've set
`names(data)` lists words in the header of "filename" after a read command defining `data`
`attach(data)` executes all commands `x=c(`$x_1,\dots,x_n$`)` where `x` heads $x_1,\dots,x_n$ in `data`.

`x[-c(`$x_{j_1},\dots,x_{j_m}$`)]` is the list $x_1,\dots,x_n$ with $x_{j_1},\dots,x_{j_m}$ omitted
`matrix(`$c(x_1,\dots,x_{mn})$`,ncol=n)` has entries $x_{m(j-1)+1},\dots,x_{mj}$ in column j, for $1 \le j \le n$.
`write(c(`$x_1,\dots,x_{mn}$`),"new.txt",ncolumns=m,sep="\t")` creates "new.txt" with entries $x_{n(i-1)+1},\dots,x_{ni}$ in row i, for $1 \le i \le m$.

`par(mfrow=c(m,n))` tells R to put the next mn plots in an array with m rows and n columns
`plot(x,y,type="n")` displays what `plot(x,y)` would have, but without the points
`points(x,y)` issued after a `plot` command adds points to the plot
`abline(`b_0`,`b_1`)` after a command like `plot` adds a line with intercept b_0 and slope b_1
`abline(v=?)` after a command like `plot` adds a vertical line through ? on the horizontal

`q()` quits R; don't save your working directory

Lynne Butler (lbutler@haverford.edu)

1. In January 2021, the US Centers for Disease Control and Prevention published the following percentiles for the height (in centimeters) of 5,092 males aged 20 and over from a nationally representative sample.

Percentile								
5th	10th	15th	25th	50th	75th	85th	90th	95th
162.8	165.8	167.6	170.1	175.4	180.2	182.9	184.7	187.4

(a) These data are consistent with the claim that heights of adult men in the US are normally distributed. Why? Your explanation should refer to a plot you provide.

(b) Using only sample quartiles of the distribution of men's heights, estimate the mean of the distribution. What property of the normal distribution justifies your calculation?

(c) Using only the first and third sample quartiles of the distribution of men's heights, estimate the variance of the distribution. Justify your calculation.

2. Weather stations at Philadelphia International Airport (PHL) and Boston Logan International Airport (BOS) record daily high and low temperatures, MAX (°F) and MIN (°F). *Diurnal temperature range* is DTR = MAX − MIN. Records for a sample of 90 summer days and 90 winter days are given in `PHLwin.csv`, `PHLsum.csv`, `BOSwin.csv` and `BOSsum.csv`. The airports are 280 miles apart; both are on the east coast of the United States.

Of interest in this problem are four population parameters:

μ_{Pw} = mean DTR at PHL in winter (°F)
μ_{Ps} = mean DTR at PHL in summer (°F)
μ_{Bw} = mean DTR at BOS in winter (°F)
μ_{Bs} = mean DTR at BOS in summer (°F)

(a) Find a 95% confidence interval for $\mu_{Ps} - \mu_{Pw}$. Do these data provide strong evidence that mean DTR in summer exceeds mean DTR in winter at PHL? If so, by how many °F? What command in R did you use to compute the confidence interval? Provide plot(s) that must be checked to see that the assumptions of the test are reasonable, and say what you concluded by looking at the plot(s).

(b) At significance level .05, test the null $H_0 : \mu_{Ps} - \mu_{Pw} = \mu_{Bs} - \mu_{Bw}$ against the two-sided alternative. Provide a plot that justifies your test procedure. Provide an explanation of your conclusion that can be understood by a non-statistician who is interested in comparing the difference between mean DTR in summer and winter at PHL with the difference between mean DTR in summer and winter at BOS. What command in R did you use to conduct the test? Hint: Rephrase the null.

3. Let b and m be real numbers and let $\sigma > 0$. Suppose X is normally distributed and the conditional distribution of Y given $X = x$ is $N(b + mx, \sigma^2)$ for every real number x. Consider the random variables $E(Y|X)$ and $\text{Var}(Y|X)$.

(a) What is an expression for $E(Y|X)$ in terms of X, $E(X)$, $\text{Var}(X)$, b, m and σ? Why?

(b) What is an expression for $E(E(Y|X))$ in terms of $E(X)$, $\text{Var}(X)$, b, m and σ? Why?

(c) What is an expression for $\text{Var}(E(Y|X))$ in terms of $E(X)$, $\text{Var}(X)$, b, m and σ? Why?

(d) What is an expression for $\text{Var}(Y|X)$ in terms of X, $E(X)$, $\text{Var}(X)$, b, m and σ? Why?

(e) Explain why your answers to (c) and (d) are consistent with the *Mathematica* calculation below. Hint: $\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))$.

```
Integrate[1 / Sqrt[2 Pi VarX] E^(-1 / (2 VarX) (x - EX)^2)
  1 / Sqrt[2 Pi sigma^2] E^(-1 / (2 sigma^2) (y - b - m x)^2),
 {x, -Infinity, Infinity},
 Assumptions → {sigma > 0, VarX > 0, b ∈ Reals, m ∈ Reals}]
```

$$\frac{e^{-\frac{(\mathbf{b+EX\,m-y})^2}{2\,(\mathbf{sigma}^2+\mathbf{m}^2\,\mathbf{VarX})}}}{\sqrt{2\pi}\,\sqrt{\mathbf{sigma}^2+\mathbf{m}^2\,\mathbf{VarX}}}$$

(f) What is an expression for $\text{Cov}(X, Y)$ in terms of $E(X)$, $\text{Var}(X)$, b, m and σ? Why? Hint: $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$.

Primer on Vaccine Efficacy: Vaccine efficacy is estimated from a trial in which half the subjects receive the vaccine, while the other half receive a placebo. For example, the efficacy of the **Moderna** vaccine, mRNA-1273, could be estimated as $1 - \frac{11}{185}$ because the trial had 11 symptomtic Covid-19 cases out of 15,210 vaccinated subjects and 185 cases out of 15,210 non-vaccinated subjects. Of the total number of cases, $11 + 185 = 196$, a fraction $\hat{p} = 11/196$ were in the vaccinated group. Vaccine efficacy is defined as $\theta = 1 - \frac{p}{1-p}$, so could be estimated as

$$\hat{\theta} = 1 - \frac{\hat{p}}{1-\hat{p}} = 1 - \frac{11/196}{1 - 11/196} = 1 - \frac{11/196}{185/196} = 1 - \frac{11}{185} \approx 94.1\%$$

The computation reported on 4 February 2021 in the *New England Journal of Medicine* adjusted for the fact that not all subjects received two shots and were surveilled for the same length of time.

4. A credible interval of (.903, .976) was reported on 31 December 2020 in *The New England Journal of Medicine* for the efficacy of the monovalent **Pfizer-BioNTech** COVID-19 vaccine.

Efficacy End Point	BNT162b2		Placebo		Vaccine Efficacy, % (95% Credible Interval)‡	Posterior Probability (Vaccine Efficacy >30%)§
	No. of Cases	Surveillance Time (n)†	No. of Cases	Surveillance Time (n)†		
	(N=18,198)		(N=18,325)			
Covid-19 occurrence at least 7 days after the second dose in participants without evidence of infection	8	2.214 (17,411)	162	2.222 (17,511)	95.0 (90.3–97.6)	>0.9999

* The total population without baseline infection was 36,523; total population including those with and those without prior evidence of infection was 40,137.

† The surveillance time is the total time in 1000 person-years for the given end point across all participants within each group at risk for the end point. The time period for Covid-19 case accrual is from 7 days after the second dose to the end of the surveillance period.

‡ The credible interval for vaccine efficacy was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

§ Posterior probability was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

(a) Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p) and $X = \sum_{i=1}^n X_i$. If the prior for the parameter p is Beta$(\alpha, 1)$, where $\alpha > 0$, what is the posterior for p given $X = x$, where $x \in \{0, 1, \ldots n\}$?

(b) What justifies the use, mentioned in the table's footnote, of a "beta-binomial model"? Hint: What is a "success" and what is a "failure"? (See the primer on the previous page and take the perspective of the virus that would like to circumvent vaccine-induced immunity.)

(c) What is wrong with the R code below, which uses the counts 8 and 162 but finds the 95% CI (.904, .976), not (.903, .976), for vaccine efficacy θ? Hint: The total surveillance time is $2.214 + 2.222 = 4.436$ thousand person years.

```
> 1-qbeta(.025,.700102+8,1+162)/(1-qbeta(.025,.700102+8,1+162))
[1] 0.9762552
> 1-qbeta(.975,.700102+8,1+162)/(1-qbeta(.975,.700102+8,1+162))
[1] 0.9035199
```

(d) What is the mean of a Beta$(\alpha, 1)$ distribution, where $\alpha > 0$? Hint: The Beta(α, β) distribution, for $\alpha > 0$ and $\beta > 0$, has pdf $f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$. (Recall that $\Gamma(k+1) = k\Gamma(k)$ if $k > 0$.)

(e) Which distribution, Beta$(.700102, 1)$ or Beta$(.7, 1)$, has mean equal to the Bernoulli parameter value that gives vaccine efficacy 30%? (FDA guidance for industry in June 2020 specified estimated vaccine efficacy be at least 50%, with a CI lower limit greater than 30%.)

(f) Explain one of the two calculations below of the credible interval provided in the table.

```
> 1-qbeta(.025,.700102+8*2.218/2.214,1+162*2.218/2.222)/(1-qbeta(.025,.700102+8*2.218/2.214,1+162*2.218/2.222))
[1] 0.976155
> 1-qbeta(.975,.700102+8*2.218/2.214,1+162*2.218/2.222)/(1-qbeta(.975,.700102+8*2.218/2.214,1+162*2.218/2.222))
[1] 0.9032201
```

```
> 1-qbeta(.025,.7+8*2.218/2.214,1+162*2.218/2.222)/(1-qbeta(.025,.7+8*2.218/2.214,1+162*2.218/2.222))
[1] 0.9761554
> 1-qbeta(.975,.7+8*2.218/2.214,1+162*2.218/2.222)/(1-qbeta(.975,.7+8*2.218/2.214,1+162*2.218/2.222))
[1] 0.9032209
```