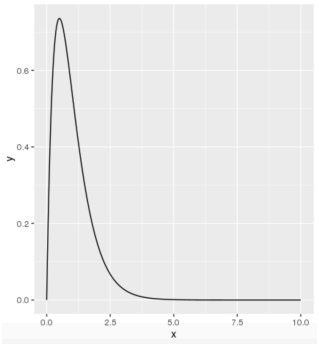
ST 516 - Homework 7 Paul ReFalo 11/11/17

1. (4 points) In lab you investigated the validity of a test to violations of assumptions. In this question you will examine the validity of t-based confidence intervals in different situation: small sample sizes and non-Normal distributions. We will use the case of a Gamma(2, 2) distribution, which has a mean of 1, to investigate.

   . (a) Make a plot in ggplot2 of a Gamma(shape = 2, rate = 2) density curve.




```
> x <- seq(from = 0, to = 10, by = 0.01)
> y <- dgamma(x, shape = 2, rate = 2, log = FALSE)

> qplot(x, y, geom = "line") # Plot Exponential(1)
```

   . (b) Write a function that (i) draws a sample of size n from a Gamma(2, 2) distribution, (ii) performs a t-test with t.test(), (iii) extracts the 95% confidence interval, and (iv) returns TRUE if the interval contains the true mean, and FALSE if it does not. (Hint: it's easiest to write code that works for one specific example, that is, pick a sample size, and write steps (i), (ii), (iii) and (iv), then turn that code into a function).

```
> gammaInterval <- function(n = 25) {
+   g <- rgamma(n, shape = 2, rate = 2)
+   Gresults <- t.test(g, conf.level = 0.95)
+
+   if (Gresults$conf.int[1] <= 1 & 1 <= Gresults$conf.int[2]) {
+     return(TRUE)
+   } else {
+     return(FALSE)
+   }
+ }
```

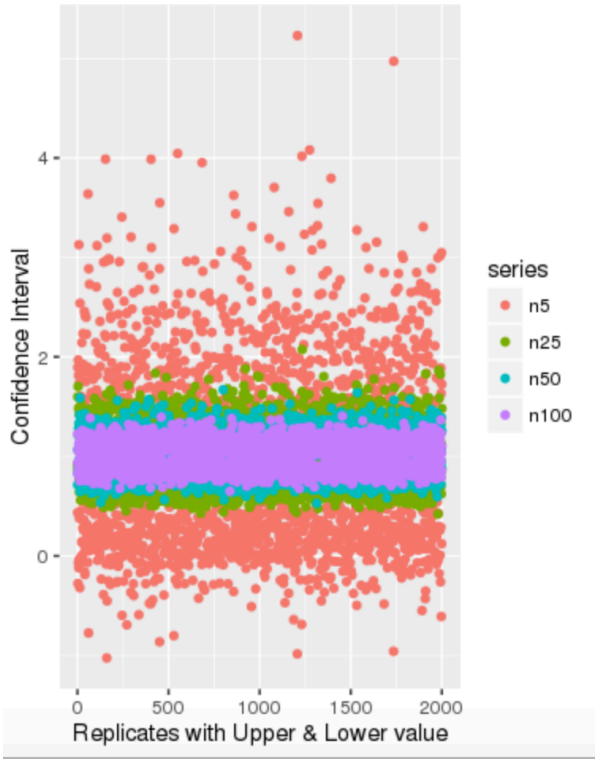
. (c) Use your function, along with mean() and replicate(), to find the proportion of t-based 95% confidence intervals in 100,000 samples of size 5, that contain the true Gamma(2,2) population mean.

```
> simGamma5 <- mean(replicate(100000, gammaInterval(5)))
> simGamma5
[1] 0.91479
```

. (d) Now repeat for samples of size 25, 50, and 100.

```
> simGamma25 <- mean(replicate(100000, gammaInterval(25)))
> simGamma25
[1] 0.93519
> simGamma50 <- mean(replicate(100000, gammaInterval(50)))
> simGamma50
[1] 0.94165
> simGamma100 <- mean(replicate(100000, gammaInterval(100)))
> simGamma100
[1] 0.94683
```

. (e) In two sentences interpret what is happening to the confidence intervals as the sample size increases, and why.



The confidence interval is decreasing as sample size gets larger, that is, confidence interval and sample size are inversely proportional. This is because of the robustness of t-based procedures to normality and skew given sufficiently large sample sizes.

Visual aid at the left. I just wanted to see what the confidence interval points all graphed together might look like. See R script for plot construction.

2. (2 points) The Great Britain Office of Population Census and Surveys collected data on a random sample of 170 married, opposite sex, couples in Britain, recording the age (in years) and heights (in cm) of the husbands and wives. They conduct a two-sample t-test to compare the mean height of husbands to the mean height of wives.

- Assumption: Independence is most likely violated in this study. While the pairs were selected at random the two people in the pair lack independence since people tend to marry others of similar height to themselves, thereby creating a cluster effect.

- Robustness against violation: Getting more data on dependent data sets won't help. The study likely isn't robust against cluster effects.

- Improvement: Conduct the test again with unpaired, randomly selected men and women and report both results (paired and unpaired). If the results of the two studies are the same, this demonstrates robustness to the violation. If they differ this is evidence against robustness of the proximity effect violation.

3. (2 points) In a study on the differences in diet between high and low income households, 50 low income and 50 high income households are randomly selected. Every adult in each household records their calorific intake for one week, and this is summarised to a daily average for each person. In total there are 110 adults in the high income households and 96 adults in the low income houses. The mean average daily caloric intake is compared between adults living in low and high income households using a two sample t-test.

- Assumption: Independence is most likely violated in this study. Since households were randomly selected but individuals in each household are part of the data set this creates a proximity effect.

- Robustness against violation: Getting more data on dependent data sets won't help. The study likely isn't robust against cluster effects.

- Improvement: Conduct the test again with randomly selected adults in both high and low incomes. Analyze the results and report both results. If the results of the two studies are the same, this demonstrates robustness to the violation. If they differ this is evidence against robustness of the proximity effect violation.

4. (2 points) In an effort to quantify gender inequality in income, the State of Oregon collects a random sample of 2000 residents with comparable qualifications and years of experience (in practice this is really hard to do, but for the purpose of this problem assume it was done well). They compare the mean income of females to the mean income of males using a two-sample t-test.

- Assumption: Data normality and scale is likely violated here. Try a log scale and reanalyze the data.
- Robustness against violation: The data needs to be reanalyzed on a different scale so robustness or lack thereof will be reflexed in the two analysis.
- Improvement: Re-plot the data on a log scale on the x-axis and reanalyze. Analyze the results and report both results. If the results of the two studies are the same, this demonstrates robustness to the violation. If they differ this is evidence against robustness of the scaling violation.