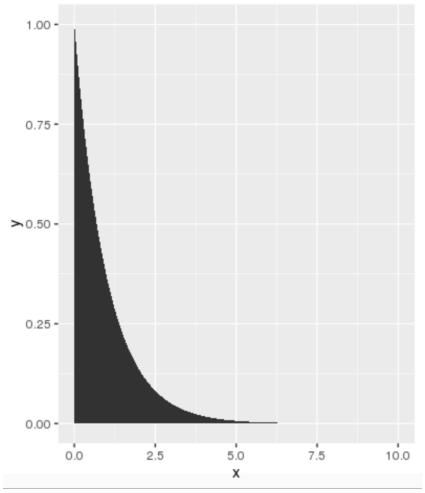
ST 516 - Homework 5
student Paul ReFalo 10/28/17

1(a). It is helpful to be able to picture the Exponential distribution, so follow the steps below to plot the distribution function curve.

```
> x <- seq(from = 0, to = 10, by = 0.01)
> y <- dexp(x, rate = 1)
> qplot(x, y, geom = "area")
> set.seed(1908)
> x2 <- rexp(10, rate = 1)
> x2 # show those numbers
 [1] 0.489901 2.849331 0.946198 1.721328 1.481969
1.659120 0.360343 0.209130 2.668938 2.581284
```



(b) Run a t-test on your size 10 sample, for a null hypothesis that $\mu = 2$, against a two sided alternative, using the t.test() function. Write a summary that includes an interpretation of the p-value and 95% confidence interval.

```
> t.test(x2, mu = 2, conf.level = 0.95) # two-sided by default

        One Sample t-test

data:  x2
t = -1.616, df = 9, p-value = 0.14
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
 0.792493 2.201015
sample estimates:
mean of x
  1.49675
```

The p-value is the probability of observing a result at least as extreme as the data observed, if the null hypothesis is true. A p-value of 0.14 tells us that when the true population mean is 2 (null hypothesis is true), the probability of observing a sample mean at least as far from 2 as 1.49675 is 0.14. We are 95% confident that the population mean is between 0.79 and 2.20.

(c) Calculate a z-statistic (continue to use the sample SD, not population SD), and a p-value based on the normal distribution.

```
> Z <- (mean(x2) - 2)/(sd(x2)/sqrt(length(x2)))
> Z
[1] -1.61648
>
> sd_x2 <- sd(x2)
> sd_x2
[1] 0.984489
>
> # Calculate p-value (two-sided test)
> P <- 2 * pnorm(abs(Z), mean = 0, sd = sd_x2, lower.tail = FALSE)
> P
[1] 0.100601
```

(d) You should find the test statistic is the same for both tests:

- Why is the p-value different?

The Central Limit Theorem at sample size 10 does not apply so sample distribution is not normal. In small samples, the Normal Distribution is not a good estimate of the distribution of the statistic of interest. For this, a t-test would be better.

- Which is more appropriate in real life, where the population standard deviation is usually unknown?

In real life, t-test would be better especially for smaller sample sizes.

2(a). Conduct two tests of $H_0: p = 0.25$ vs. $H_A: p/= 0.25$, using (i) prop.test() without a continuity correction; and (ii) binom.test(). Store the results of (i) as A (for approximate), and (ii) as E (for exact). Note: a 95% confidence interval for is the default for both.

```
> x_q2 <- rbinom(1, 20, p = 0.25)
> x_q2
[1] 4
>
> A <- prop.test(x_q2, n = 20, p = 0.25, conf.level = 0.95, correct = FALSE)
> A
```

Continued on next page

```
      1-sample proportions test without continuity correction

data:  x_q2 out of 20, null probability 0.25
X-squared = 0.2667, df = 1, p-value = 0.606
alternative hypothesis: true p is not equal to 0.25
95 percent confidence interval:
 0.0806577 0.4160174
sample estimates:
  p
0.2

> E <- binom.test(x_q2, n = 20, p = 0.25, conf.level = 0.95)
> E

      Exact binomial test

data:  x_q2 and 20
number of successes = 4, number of trials = 20, p-value = 0.798
alternative hypothesis: true probability of success is not equal to 0.25
95 percent confidence interval:
 0.057334 0.436614
sample estimates:
probability of success
                   0.2
```

(b) A is now a list with nine elements; use str(A) to see for yourself! The lower and upper bound of the confidence interval returned by prop.test() can be retrieved with A$conf.int[1] and A$conf.int[2]. Write a logical statement that returns TRUE if 0.25 is in the 95% confidence interval. Do the same for binom.test().

```
> A$conf.int[1] < 0.25 & 0.25 < A$conf.int[2]
[1] TRUE


> E$conf.int[1] < 0.25 & 0.25 < E$conf.int[2]
[1] TRUE
```

Continued on next page

(c) Write approx() function and exact() function that does the same for the exact binom.test().

```
> approx <- function(n) {
+   x_q2 <- rbinom(1, n, p = 0.25)
+   A <- prop.test(x_q2, n, p = 0.25, conf.level = 0.95, correct = FALSE)
+   #print(A)
+   if (A$conf.int[1] <= 0.25 & 0.25 <= A$conf.int[2]) {
+     return(TRUE)
+   } else {
+     return(FALSE)
+   }
+
+ }
>
> exact <- function(n) {
+   x_q2 <- rbinom(1, n, p = 0.25)
+   E <- binom.test(x_q2, n, p = 0.25, conf.level = 0.95)
+   #print(E)
+   if (E$conf.int[1] <= 0.25 & 0.25 <= E$conf.int[2]) {
+     return(TRUE)
+   } else {
+     return(FALSE)
+   }
+
+ }
>
> approxResult <- approx(20)
> exactResult <- exact(20)
>
> approxResult
[1] TRUE
> exactResult
[1] TRUE
>
> approxReplicates <- replicate(10000, approx(20))
> exactReplicates <- replicate(10000, exact(20))
```

Continued on next page

(d) Now use replicate() to repeat the process 10,000 times for each test, for n = 20. What proportion of intervals covered the true parameter value for each test? How do the two methods compare? If you get any warnings from the use of prop.test() you can ignore them. With 10,000 simulations these estimated probabilities should be within ±0.01 of the truth.

```
> mean(approxReplicates)
[1] 0.9327
> mean(exactReplicates)
[1] 0.9664
```

(e) Now repeat the previous part, for n = 100. How do they compare now?

```
> approxReplicates <- replicate(10000, approx(100))
> exactReplicates <- replicate(10000, exact(100))
>
> mean(approxReplicates)
[1] 0.9508
> mean(exactReplicates)
[1] 0.9614
```

The p-value At low sample sizes, the binomial distribution is more accurate but at higher sample sizes, the prop.test() wins out.

4. (2 points)
Average systolic blood pressure of a normal male is supposed to be about 129. Measurements of systolic blood pressure on a sample of 12 adult males from a community whose dietary habits are suspected of causing high blood pressure are (in R ready format):

bp <- c(115, 134, 131, 143, 130, 154, 119, 137, 155, 130, 110, 138)

Do the data justify the suspicions regarding the blood pressure of this community?

Continued on next page

The hypotheses are

H_0: $\mu = 129$

H_1: $\mu \geq 129$

At $\alpha = 0.01$ the critical value from Table A.2 for 11 degrees of freedom is $|t| > 2.718$ for an upper-tail test.

```
> bp <- c(115, 134, 131, 143, 130, 154, 119, 137, 155, 130, 110, 138)
> var <- ( sum( (bp - mean(bp) )^2 ) ) / length(bp - 1)
> bpMean <- mean(bp)
> t <- (bpMean - 129) / sqrt(var/length(bp))
> t
[1] 1.0381
```

Since the t obtained, 1.0381 is less than the critical value of 2.718 we fail to reject the null hypothesis. The data do not support the suspicion that their dietary habits are resulting in increased blood pressure.

5. (2 points) (Adpated From Ex 22. Chapter 4 Statistical Methods. Freund, R.; Mohr, D; Wilson,W. (2010))
The following data gives the average pH in rain/sleet/snow for the two-year period 2004-2005 at 20 rural sites on the U.S. West Coast. (Source: National Atmospheric Deposition Program).

Is there evidence the median pH is not 5.4?

Conduct an appropriate test, construct a confidence interval and write a summary with your conclusions in the context of the study.

The hypotheses are

H_0: $M = 5.4$

H_1: $M \neq 5.4$

The Z calculated below for a pH of 5.4 in the rain data set gives ~2.68 which is greater than either $z_{\alpha/2} = 1.96$ for 95% which has a confidence interval of 5.305 - 5.36 or $z_{\alpha/2} = 2.576$ for 99% which has a confidence interval of 5.265 - 5.395. We reject the null hypothesis in favor of the alternative hypothesis at both of these intervals and conclude that there is evidence that the median pH is not 5.4.

Continued on next page

```
> rain <- c(5.335, 5.345, 5.380, 5.520, 5.360, 6.285, 5.510, 5.340,
+           5.395, 5.305, 5.190, 5.455, 5.350, 5.125, 5.340, 5.305,
+           5.315, 5.330, 5.115, 5.265)
> rain <- sort(rain)

> obsBelow <- sum(rain < 5.4)
> obsRange <- c(1:20)
> Z <- ((obsBelow/length(rain)) - 0.5) / sqrt(0.5*0.5/length(rain))
> Z
[1] 2.68328
> Z <- ((obsRange/length(rain)) - 0.5) / sqrt(0.5*0.5/length(rain))
> Z
 [1] -4.024922 -3.577709 -3.130495 -2.683282 -2.236068 -1.788854 -1.341641 -0.894427
-0.447214  0.000000  0.447214  0.894427
[13]  1.341641  1.788854  2.236068  2.683282  3.130495  3.577709  4.024922  4.472136
>
> # Zα/2 = 1.96 for 95% CI
>
> CIlow95 <- (20/2) - ((1.96 * sqrt(20)) / 2)
> CIlow95 # 6th smallest
[1] 5.61731
>
> CIhi95 <- (20/2) + ((1.96 * sqrt(20)) / 2)
> CIhi95 # 14th smallest
[1] 14.3827
> rain
 [1] 5.115 5.125 5.190 5.265 5.305 5.305 5.315 5.330 5.335 5.340 5.340 5.345 5.350
5.360 5.380 5.395 5.455 5.510 5.520 6.285
> rain[[6]]
[1] 5.305
> rain[[14]]
[1] 5.36
>
> # Zα/2 = 2.57 for 99% CI
>
> CIlow99 <- (20/2) - ((2.57 * sqrt(20)) / 2)
> CIlow99 # 4th smallest
[1] 4.25331
>
> CIhi99 <- (20/2) + ((2.57 * sqrt(20)) / 2)
> CIhi99 # 16th smallest
[1] 15.7467
>
> rain[[4]]
[1] 5.265
> rain[[16]]
[1] 5.395
```

End of Homework 5