Instructions: Submit one .pdf file that includes the answers to both sets of questions. Also submit your .R script that you used to answer the questions in the R part.

Please feel free to discuss questions on the discussion board.

R Questions

1. (3 points) This question explores the difference between using the Normal distribution and t-distribution as reference distributions for a one sample comparison.

   Begin by setting the seed to 1908 with `set.seed(1908)`, and using `rexp()` to draw a sample of size 10 from an Exponential distribution with rate parameter 1. Name the sample `x2`. Remember you can run `?rexp()` for help on the arguments.

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

      i. First you need to create a vector of x-axis values. The function `seq()` creates a sequence from the `from` argument to the `to` argument, in steps specified by the `by` argument. The appropriate range will depend on the distribution, but from 0 to 10 should be enough for this example:
      ```
      x <- seq(from = 0, to = 10, by = 0.01) # x-axis values
      x
      ```

      ii. Next we find the values of the p.d.f of Exponential(1) distribution at those x-axis values. The function `dexp(x, rate = 1)` gives the value of the Exponential(1) distribution for the values stored as the vector `x`.
      ```
      y <- dexp(x, rate = 1) # distribution function curve
      ```

      iii. Finally you can use `qplot(x, y, geom = "line")` to create a plot. Remember to load the `ggplot2` package.
      ```
      library(ggplot2)
      qplot(x, y, geom = "line") # Plot Exponential(1)
      ```
      This is the population distribution. You might like to try swapping out `geom = "line"` with `geom = "point"`, to see how the `x` and `y` values we created were joined to create the line, or with `geom = "area"` to see a filled version instead.

   (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.

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

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

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

2. (3 points) This question explores the difference between using the Normal approximation and the exact Binomal test for comparing two proportions.

   Draw a single *number of successes* from a Binomial with $n = 20$ with success probability 0.25, and store the result as `x`, using the `rbinom()` function.

(a) Conduct two tests of $H_0 : p = 0.25$ vs. $H_A : p \neq 0.25$, 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.

(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()`.

(c) OK, so far you've figured out how to test whether the true value is inside the confidence interval for draw from the Binomial, using both the exact and approximate approaches. Our goal is to repeat this many times, to assess how close to 95% these CIs are.

First, write a function `approx()` that takes `n` as an argument, draws a single binomal random variable with size `n` and true probability 0.25, conducts an approximate test and returns `TRUE` or `FALSE` depending on whether the 95% CI contains 0.25. Write another function `exact()` that does the same for the exact `binom.test()`.

(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.

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

Conceptual Questions

Answer **two** of the following **three** conceptual questions.

3. (2 points) Ruchdeschel et al. [1] claim that the sex ratio of Ridley's sea turtle (a very rare and endangered sea turtle) moved from a male biased ratio to a female biased ratio.

   They recorded the sex of stranded seas turtles on Cumberland Island. From 1983 to 1989 there were 16 males and 10 females. From 1990 to 2001 there were 19 males and 56 females.

   Is there evidence that the sex ratio of Ridley's sea turtles was male biased in the period 1983-1989, and female biased in the period 1990-2001?

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

   (Note that the more relevant question might be, did the sex ratio change between these two periods? You'll need the tools from the next module to answer that.)

4. (2 points) (From Ex 6. Chapter 4 Statistical Methods. Freund, R.; Mohr, D; Wilson,W. (2010))

   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 casusing 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?

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 peiod 2004-2005 at 20 rural sites on the U.S. West Coast. (Source: National Atmospheric Deposition Program).

```
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)
```

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.

References

[1] Carol Ruckdeschel, C Robert Shoop, and Robert D Kenney. On the sex ratio of juvenile lepidochelys kempii in georgia. *Chelonian Conservation and Biology*, 4(4):858–861, 2005.