Quiz 6

Nick Strayer, Alex Sundermann, Linjun Fu, Michael Greer

October 10, 2015

Question 1

Let R21 be the number of rolls to get a "twenty-one-zee". How big of a sample from R21 (Nsamples) do you need to take to estimate E[R21] accurate to 2 decimal places? For this quiz, let's define "accurate to 2 decimal places" as having the 99.7% confidence interval for the estimator have a half-width < 0.0025. Suppose you are certain SD[R21] < 7 and are willing to assume your sample will be large enough that you can rely on the central limit theorem to create your confidence interval for E[R21]. Show your work. You don't need to run any simulations.

$$
\begin{aligned}
2.9677 \cdot \frac{7}{\sqrt{n}} &= 0.0025 \\
\frac{2.9677 \cdot 7}{.0025} &= \sqrt{n} \\
69,048,787 &= n
\end{aligned}
$$

Question 2

Carry forward all the information from question 1. How big of a sample from R21 (Nsamples) do you need to take to estimate SD[R21] accurate to 2 decimal places?

```
#Let's grab the width
intervalWidth <- function(n){
  lower <- ( 49 * qchisq(.0015, df= n-1) )/(n-1)
  upper <- ( 49 * qchisq(.9985, df= n-1) )/(n-1)
  return(abs( sqrt(lower) - sqrt(upper) ) )
}

#Brute force find.
n <- 34500000 #This was chosen by a previous simulation.
while(intervalWidth(n) > .005) n <- n + 1
```

We would need an n of 34525277 to get the same precision.

Question 3

Suppose the sample had a sample mean = 95 and sample sd = 10. Calculate a 95% confidence interval utilizing student's t distribution to allow an unknown true sd. You can do this by hand and/or use any software you like. Stata's cii command is especially helpful. Using two approaches is a nice way to double check your answer. Comment on what you would conclude about the drug in this case. Does it appear to work in reducing nicotine dependence per the BANDS?

$$\bar{x} \pm t_{n-1}\frac{S_n}{\sqrt{n}} = 95 \pm 2.093 * \frac{10}{\sqrt{20}} = 95 \pm 4.68$$

The 95% confidence interval for the sample mean of individuals taking the drug is (90.32, 99.68) which does not include the population mean 100. Therefore, this sample shows evidence supporting that the mean BANDS among individuals receiving the test drug is different than the control group. More specifically, since the all values that fall within the 95% CI exclude 100, we can say that this sample gives evidences that the BANDS mean of individuals on test drug is lower than the population of interest.

```
cii 20 95 10
```

```
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |         20          95    2.236068        90.31986    99.68014
```

Question 4:

> Suppose that for people on the drug, BANDS ~ N(mean=95, sd=12). What is the probability a sample of 20 would yield a 95% confidence interval that excludes 100?

```
Q4 <- function(means=95, sample_size=20, sd=12, excludes=100){
  count <- 0
  iter  <- 10000
  for(i in 1:iter){
      temp <- rnorm(sample_size, mean=means, sd=sd)
      low_bound <- mean(temp) - 2.093*sd(temp)/sqrt(sample_size)
      up_bound  <- mean(temp) + 2.093*sd(temp)/sqrt(sample_size)

      if(low_bound >= excludes | up_bound <= excludes) count = count + 1

      }
  return (count/iter)
}
res4 <- Q4()
```

The proportion of intervals excluding 100 = **0.4278.**

We see that the probability a sample of 20 would yield a 95% confidence interval that excludes 100 is less than 50%. That said, considering the difference between 95 and 100 is only 5, the sample size of 20 is not large enough to exclude 100.

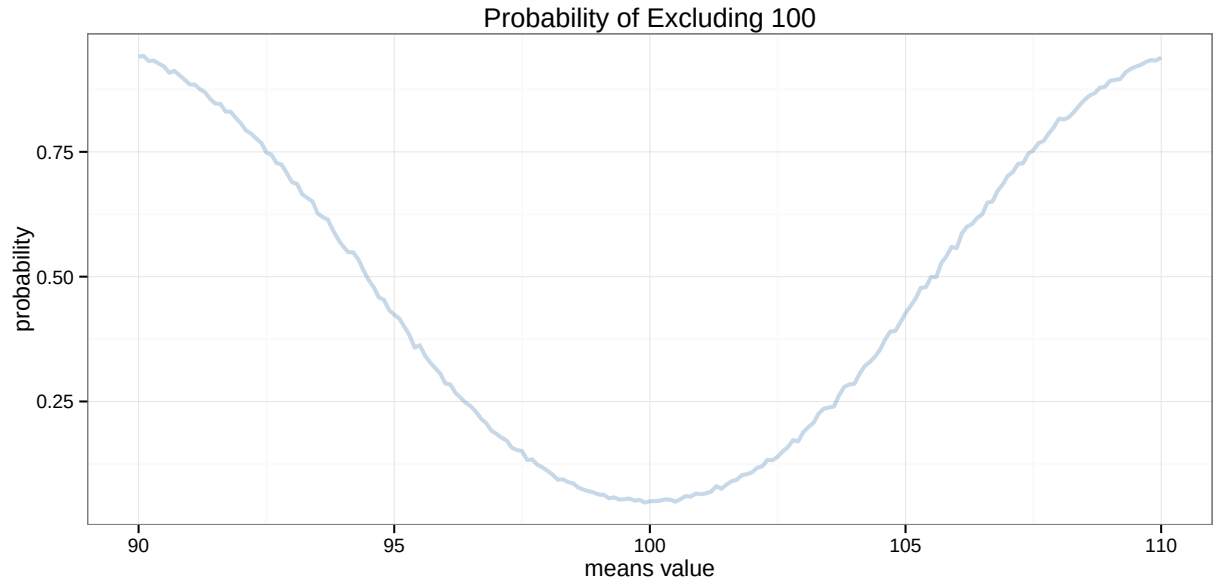
Question 5:

> Extend question 04 to consider a range of means from 90 to 110. Create a plot of the probability of a 95% CI excluding 100 by the true mean.

```r
delta <- 0.1
means <- seq(90, 110, delta)
p     <- numeric(length(means))

for (i in 1:length(means)) p[i] = Q4(means=means[i])

ggplot(data.frame(mean = means, prob = p), aes(x = mean, y = prob)) +
  geom_line(colour = "steelblue", size = 1, alpha = 0.3) + theme_bw() +
  labs(title = "Probability of Excluding 100", x = "means value", y = "probability")
```



The plot is symmetric about the point means = 100. This makes sense as the further away the true mean is from 100, the higher probability the 95% CI has to exclude 100.

Question 6:

> Now suppose a new study with N=21 subjects took pre-drug and post-drug BANDS measurements and recorded whether subjects scores decreased or not. Assuming the chance of getting the exact same score was trivial and that the drug had no effect, the probability of decreasing would be 50%. Suppose 15 of 21 subjects had lower post-drug BANDS. Calculate 95% confidence intervals using the exact, Wilson, and asymptotically Normal approaches. Comment on the study's conclusions for each method.

```r
asy_conf <- binconf(15, 21, alpha=0.05, method="asymptotic",
                    include.x=FALSE, include.n=FALSE, return.df=FALSE)
wil_conf <- binconf(15, 21, alpha=0.05, method="wilson",
                    include.x=FALSE, include.n=FALSE, return.df=FALSE)
ext_conf <- binconf(15, 21, alpha=0.05, method="exact",
                    include.x=FALSE, include.n=FALSE, return.df=FALSE)
```

For this example, we are assuming that we know that the drug has no effect and therefore the chance of having a lower BANDS score after treatment is equal to the chance of having a higher score, 50%. This sample of 21 subjects shows that 71.43% of individuals who take the drug have lower BANDS scores than before treatment and therefore provides evidence suggesting the drug is effectively curbing nicotine dependence. A way to evaluate the strength of the evidence a point estimate provides is to evaluate the associated confidence intervals. The asymptotic normal CI (0.5210709, 0.9075005) excludes 0.50, the expected value if the drug had no effect, and therefore provides stronger evidence that the drug is effective at reducing an individual's BANDS score. The Wilson interval (0.5004362, 0.8618614) barely excludes 0.50 or the interval may be reported as ending on 0.50 depending on the number of decimal places recorded. Therefore, this confidence interval indicates that the evidence of drug efficacy provided by this sample is weak. Finally, the exact CI (0.4782489, 0.8871906) includes 0.50 and therefore the evidence for the drug's effectiveness is weakest according to the CI using this method. In conclusion, when using confidence intervals to assess the strength of the evidence in a point estimate, it is important to consider that the width of the confidence intervals vary depending on method used. When the difference in width between interval methods is the difference between excluding or including the null value, the interpretation of the strength of evidence a point estimate of a study provides may change depending on method of interval calculation used.

Question 7:

> Suppose the true probability of a lower post-drug score was 80%. What is the probability of the 95% CI excluding 50% for each of the CI methods?

```
Q7 <- function(p_lower, methods){
    count <- 0
    iter <- 10000
    for(i in 1:iter){
        temp <- rbinom(21, 1, p_lower) # 1=decrease, 0=non-decrease
        n1 <- length(temp[temp==1]) # find the number of decreasing subjects
        n0 <- length(temp[temp==0]) # find the number of non-decreasing subjects
        conf <- binconf(n1, 21, alpha=0.05, method=methods)
        if(conf[2] > 0.5 | conf[3] < 0.5){
            count = count + 1
        }
    }
    return (count/iter)
}
Q7(0.8, "wilson")
```

```
## [1] 0.8906
```

```
Q7(0.8, "exact")
```

```
## [1] 0.7607
```

```
Q7(0.8, "asymptotic")
```

```
## [1] 0.8937
```

We can see that wilson and asymptotic intervals have the similar probability of excluding 50% while the exact interval has the lowest probability. Recalling what we found in Quiz 7, exact interval always over estimates. That said, it tends to have wider interval centering which makes it harder to exclude 50%. This also agrees with our result in Q6.

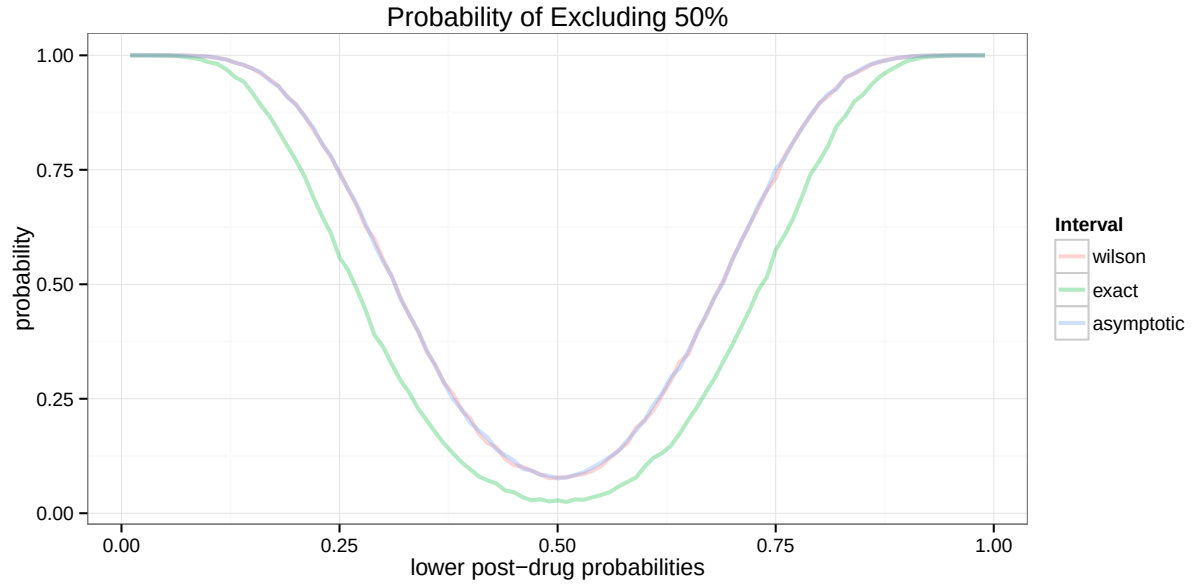
Question 8:

> Extend question 07 to consider a range of lower post-drug probabilities from 1% to 99%. Create a plot of the probability of a 95% CI excluding 50% for each of the CI methods?

```
delta <- 0.01
p <- seq(0.01, 0.99, delta) # create a vector ranging from 1% to 99% with step=1%
excl_p_wil <- rep(0, length(p)) # create a vector of probability of excluding 50%
excl_p_ext <- rep(0, length(p))
excl_p_asy <- rep(0, length(p))
for (i in 1:length(p)){
    excl_p_wil[i] = Q7(p[i], "wilson")
    excl_p_ext[i] = Q7(p[i], "exact")
    excl_p_asy[i] = Q7(p[i], "asymptotic")
}

plot_8 = data.frame(p = p, wilson = excl_p_wil, exact = excl_p_ext, asymptotic = excl_p_asy)
plot_8 = melt(plot_8, id = c("p"))
names(plot_8) = c("p", "Interval", "excl_p")

ggplot(plot_8, aes(x = p, y = excl_p, group = Interval, color = Interval)) +
  geom_line(size = 1, alpha = 0.3) + theme_bw() +
  labs(title = "Probability of Excluding 50%", x = "lower post-drug probabilities", y = "probability")
```



Again, for a range of lower post-drug probabilities from 1% to 99%, we see that wilson and asymptotic intervals have the similar probability of excluding 50% while the exact interval still has the lowest probability. The most interesting point is as follows. In Q5 we found that: for means value 100, the probability of a 95% CI excluding 100 by the true mean is nearly zero. Here in Q8 we find that: for probability of a lower post-drug score 0.5, the probability of a 95% CI excluding 50% is nontrivial! As we can see from the plot, it's still nearly zero for exact interval, but obviously deviates from zero for wilson and asymptotic intervals.