Group Lima PS 6

Tasheena, Azka & Daniel

15.

Suppose that n measurements are to be taken under a treatment condition and another n measurements are to be taken independently under a control condition. It is thought that the standard deviation of a single observation is about 10 under both conditions. How large should n be so that 95% confidence interval for $\mu_X - \mu_Y$ has a width of 2? Use normal distribution rather than the t distribution since n will turn out to be rather large.

$$2\Big(z(\frac{\alpha}{2})\sigma\sqrt{(\frac{1}{n}+\frac{1}{m})}\Big) = 2$$

$$z(\frac{\alpha}{2})\sigma\sqrt{\frac{2}{n}} = 1$$

Substituting $\alpha = .05$ and $\sigma = 10$:

$$1.96(10)\sqrt{\frac{2}{n}} = 1$$

Solving for n:

$$n = 768.32$$

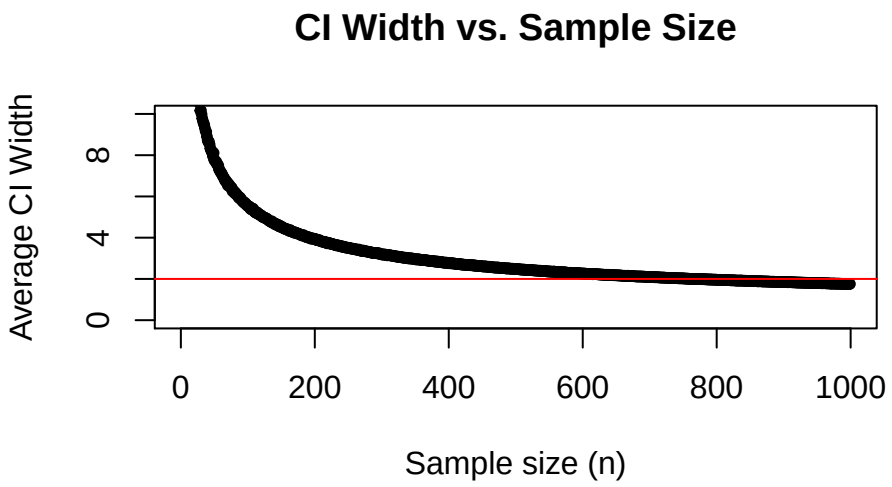
$$\textit{So } n \textit{ should be } 769$$

Empirical: To find the number of sums with width greater than 2, we find the cutoff for when the number of trials exceeds width of 2. To simulate, we generate two samples from a normal distirbution. We then use t-test to assess differernces in means of the two samples and then calculate the corresponding confidence interval. The upper and lower bounds of the confidence interval are then used to compute the width. This process is repeated 100 times and the averaged confidence intervals are calculated. The number of intervals with width greater than 2 provide a cutoff for the n value that gives a width of 2.

```
set.seed(10)
confintsum <- rep(0,1000)
confintsum[1] = 1000 #because we don't use a sample size of 1
numsim = 100
for(j in 1:numsim){
  for(sampsize in 2:1000){
    xsample <- (rnorm(sampsize, mean = 2, sd = 10))
    ysample <- (rnorm(sampsize, mean = 0, sd = 10))
    cf<-t.test(xsample, ysample)$conf.int
    confintsum[sampsize] = confintsum[sampsize] + cf[2] - cf[1]
    }
}
CIavgwidth <- confintsum/numsim
tally(CIavgwidth > 2)
```

```
##
##  TRUE FALSE
##   766   234
```

```r
plot(CIavgwidth[2:1000], ylim = c(0, 10), pch = 20, xlab = "Sample size (n)", ylab = "Average CI Width")
abline(h = 2, col = "red")
```



16

Referring to Problem 15, how large should n be so that the test of $H_0 : \mu_X = \mu_Y$ against the one-sided alternative $H_A : \mu_X > \mu_Y$ has a power of .5 if $\mu_X - \mu_Y = 2$ and $\alpha = .10$?

$$1 - \Phi\Big(z(\frac{\alpha}{2}) - \frac{\triangle}{\sigma}\sqrt{\frac{n}{2}}\Big)$$

$$1 - \Phi\Big(1.28 - \frac{2}{10}\sqrt{\frac{n}{2}}\Big) = .5$$

$$(1.28 - .2\sqrt{\frac{n}{2}}) = 0$$

$$.2\sqrt{\frac{n}{2}} = 1.28$$

$$\sqrt{\frac{n}{2}} = .064$$

$$n = 81.92$$

$$n \approx 82$$

Empirical: Power is defined as the probability of rejecting the null distribution given the null distribution is false. It can also equally be represented as 1 - type II error where type II error is the probability of failing to reject the null when the alternative is true. To find the n that has power of 0.5 if the difference in means is 2 and the alpha level is 0.10, we find the probability of the type II error and corresponding power and then tally the samples with power of 0.5 by finding the cutoff region.

```r
type2sum <- rep(0,100)
set.seed(100)
for(j in 1:1000)
    {
  for(i in 2:100)
      {
```

```
      xsample <- rnorm(i, mean = 2, sd = 10)
      ysample <- rnorm(i, mean = 0, sd = 10)
      if (t.test(xsample, ysample, alternative = "greater")$p.value>0.1)
        {
            type2sum[i] = type2sum[i] + 1
        }
    }
}
powersum = 1- type2sum/1000 #power is 1 - typeII error
powersum[1] = 0
tally(powersum<0.5) #find the cutoff region
```

```
##
##  TRUE FALSE
##    82    18
```

23

Let $X_1, ..., X_n$ be i.i.d. with cdf F, and let $Y_1, ..., Y_m$ be i.i.d. with cdf G. The hypothesis to be tested is that $F = G$. Suppose for simplicity that $m + n$ is even so that in the combined sample of X 's and Y 's, $\frac{m+n}{2}$ observations are less than the median and $\frac{m+n}{2}$ are greater.

a. As a test statistic, consider T, the number of X's less than the median of the combined sample. Show that T follows a hypergeometric distribution under the null hypothesis:

$$P(T = t) = \frac{\binom{(m+n)/2}{t} * \binom{(m+n)/2}{n-t}}{\binom{m+n}{n}}$$

a. We motivate our analytical solution with actual data where one vector contains 3 values, second has 2 and where k is set to 2. We then compute all paired differences and order our data as shown below:

```
m <- c(3,4,6)
n <- c(5,7)
sort(apply(expand.grid(c(3,4,6), -c(5,7)) , 1,sum)) #ordered set of differences
```

```
## [1] -4 -3 -2 -1 -1  1
```

A Hypergeometric distribution is used to find the number of successes (trials) in a sample drawn without replacement where the probability of success is not the same on successive trials. If we consider the possible number of X's and strive to get the number of values less than and greater than the median of the combined sample, this is intuitively a yes/no probability i.e. a sample will either be below or above the combined median. Consequently this situation intuitively mimics a bernoulli case which follows into a hypergeometric distribution, which is also appropriate since we are finding the number of X's (trials) less than the median of the combined sample.

Under null, F=G, half of the values will be less than the median and half will be below the median. To find the number of X's less than the median of the combined sample, we first consider the t that are in the first half and then the n-t remaining values that are in the second half. This gives all possible combinations of values that are less than the median of the combined sample. We then divide this by the total number of ways of choosing the number of X's from the combined sample.

Number of ways there will be 't' X's in the first half of the ordered combined sample = $\binom{(m+n)/2}{t}$, where n = # of X's and m = # of Y's.

Number of ways there will be 'n-t' X's in the second half of the ordered combined sample = $\binom{(m+n)/2}{n-t}$.

Number of ways the X's are positioned in the combined ordered sample = $\binom{(m+n)}{n}$.

Therefore, the probability that there are t X's below the combined median (i.e. in the first half of the ordered combined sample) =

$$P(T = t) = \frac{\binom{(m+n)/2}{t} * \binom{(m+n)/2}{n-t}}{\binom{m+n}{n}}$$

which follows a hypergeometric distribution.

b. Show how to find a confidence interval for the difference between the median of F and the median of G under the shift model, $G(x) = F(x - \triangle)$.

On page 442 in Rice, the confidence interval for the shift model (difference in median) is

$C = [D_{(k)}, D_{(mn-k+1)})$, where $D_{(1)}, D_{(2)}, ..., D_{(mn)}$ denotes the ordered mn differences $Y_j - X_i$. Rice encourages us to consider the case m = 3, n = 2, k = 2. From Rice's confidence interval shown above, the value at $D_{(k)}$ is -3 and value at $D_{(mn-k+1)}$ is $D_{(6-2+1=5)}$ which corresponds to value of -1. So our confidence interval according to this formula would be (-3, -1) at 59% confidence level. Our analytical solution is validated by wilcox.test which also returns a confidence interval of (-3, -1). We also calculated the confidence interval for k = 1, which comes out to be (-4, 1) at 86% confidence level. The 59% confidence level was calculated by finding the transition point from k = 1 or (-3,-1) to k = 2 or (-4, 1) while 86% confidence level was found by finding the transition point from k = 2 or (-4, 1) to error (conf.level not achievable). We observe the inversion of the confidence interval where the confidence interval for k = 1 (-3,-1) does not include 0 and is rejected while confidence interval is not rejected for interval (-4, 1) since 0 is included.

```
wilcox.test(c(3,4,6),c(5,7),conf.int=TRUE, conf.level = 0.59) #59% confidence interval of (-3, -1) at k
```

```
##
##  Wilcoxon rank sum test
##
## data:  c(3, 4, 6) and c(5, 7)
## W = 1, p-value = 0.4
## alternative hypothesis: true location shift is not equal to 0
## 59 percent confidence interval:
##  -3 -1
## sample estimates:
## difference in location
##                   -1.5
```

```
wilcox.test(c(3,4,6),c(5,7),conf.int=TRUE, conf.level = 0.86) #86% confidence interval of (-4, 1) at k
```

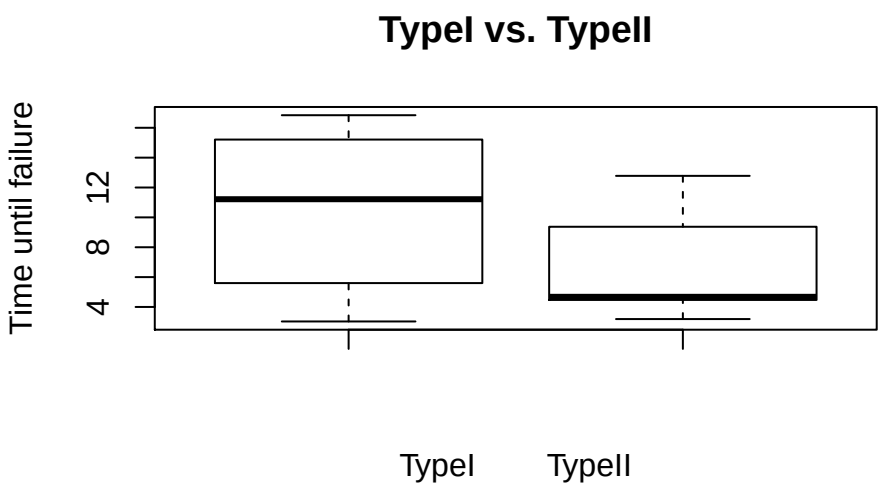
```
##
##  Wilcoxon rank sum test
##
## data:  c(3, 4, 6) and c(5, 7)
## W = 1, p-value = 0.4
## alternative hypothesis: true location shift is not equal to 0
```

```
## 86 percent confidence interval:
##  -4  1
## sample estimates:
## difference in location
##                   -1.5
```

c. Apply the results (a) and (b) to the data of Problem 21.

Sorting the data and depicting differences between the types through boxplots

```
TypeI <- c(3.03, 5.53, 5.60, 9.30, 9.92, 12.51, 12.95, 15.21, 16.04, 16.84)
TypeII <- c(3.19, 4.26, 4.47, 4.53, 4.67, 4.69, 12.78, 6.79, 9.37, 12.75)
boxplot(TypeI, TypeII, main = "TypeI vs. TypeII", ylab= "Time until failure", xlab = "TypeI        Type
```




```
TotalSamp<-c(TypeI, TypeII)
sort(apply(expand.grid(c(TypeI), -c(TypeII)) , 1, sum)) #ordered pair of differences
```

```
##   [1] -9.75 -9.72 -7.25 -7.22 -7.18 -7.15 -6.34 -3.84 -3.77 -3.76 -3.48
##  [12] -3.45 -2.86 -2.83 -1.66 -1.64 -1.50 -1.44 -1.26 -1.23 -1.19 -0.27
##  [23] -0.24 -0.16 -0.07  0.17  0.20  0.55  0.84  0.86  0.91  0.93  1.00
##  [34]  1.06  1.07  1.13  1.27  1.34  2.34  2.41  2.43  2.46  2.51  3.13
##  [45]  3.14  3.26  3.29  3.58  4.06  4.09  4.61  4.63  4.77  4.83  5.04
##  [56]  5.23  5.25  5.39  5.45  5.66  5.72  5.84  6.11  6.16  6.67  6.73
##  [67]  7.47  7.82  7.84  7.98  8.04  8.25  8.26  8.28  8.42  8.42  8.48
##  [78]  8.69  9.25  9.32  9.76 10.05 10.52 10.54 10.68 10.74 10.95 11.35
##  [89] 11.37 11.51 11.57 11.78 12.02 12.15 12.17 12.31 12.37 12.58 12.85
## [100] 13.65
```

Applying results to part a:

$$P(T=t) = \frac{\binom{(10+10)/2}{t} * \binom{(10+10)/2}{10-t}}{\binom{10+10}{10}}$$

$$P(T=t) = \frac{\binom{(20)/2}{t} * \binom{(20)/2}{10-t}}{\binom{20}{10}}$$

Applying results to part b:

```r
wilcox.test(c(TypeI),c(TypeII),conf.int=TRUE, conf.level = 0.95) #testing for differences in means with
```

```
##
##  Wilcoxon rank sum test
##
## data:  c(TypeI) and c(TypeII)
## W = 75, p-value = 0.06301
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -0.16  8.48
## sample estimates:
## difference in location
##                   4.35
```

At alpha 0.05, confidence interval for difference in means is (-0.16, 8.48). Since this interval contains 0, we would fail to reject that there are no mean differences between TypeI and TypeII.