BACS HW - Week 5

108070023

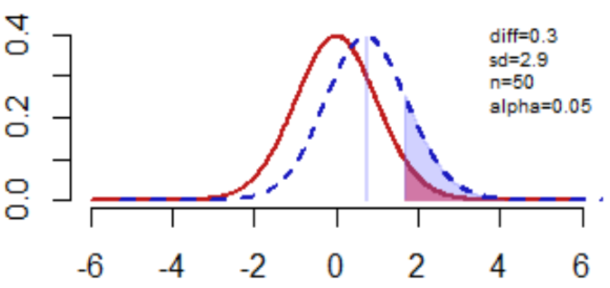
Question 1) Your colleague, a data analyst in your organization, is working on a hypothesis test where he has sampled product usage information from customers who are using a new smartwatch. He wishes to test whether the mean (x_i) usage time is higher than the usage time of the company's previous smartwatch released two years ago (μ_o):

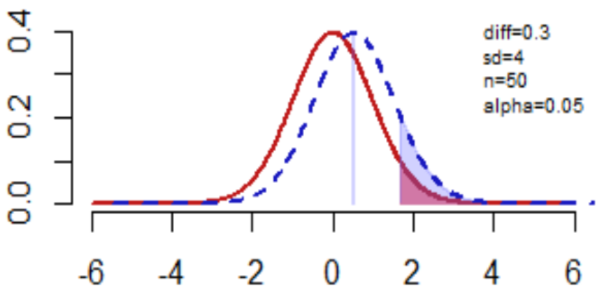
H_{null}: The mean usage time of the new smartwatch is the same or less than for the previous smartwatch.

H_{alt}: The mean usage time is greater than that of our previous smartwatch.

After collecting data from just n=50 customers, he informs you that he has found diff=0.3 and sd=2.9. Your colleague believes that we cannot reject the null hypothesis at alpha of 5%.

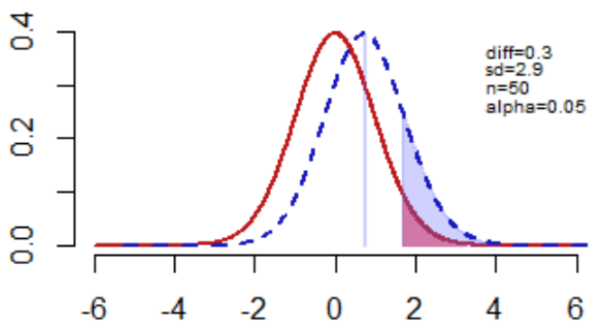
a. You discover that your colleague wanted to target the general population of Taiwanese users of the product. However, he only collected data from a pool of young consumers, and missed many older customers who you suspect might use the product *much less* every day.

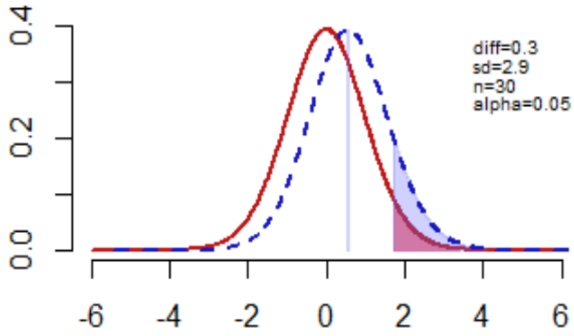




(1) Variance creates random error.
(2) Adding old customers that we suspect to use the product much less means that **the variation (standard deviation) will rises**. Therefore, we compare the original data with the higher variation one (set sd=4)
(3) The t-value will fall (t= diff/(sd/sqrt(n))). From the graph, we can see that the t-value declines, and causes the alternative distribution to move left. Since power is the area of alternative distribution from the cut-off value (95%), **the power will decrease.**
(4) From (3), we know that the power of this scenario declines. Beta is the probability of Type II error, if power(1-beta) declines, meaning that beta increases, so **Type II error will become more likely.**

b. You find that 20 of the respondents are reporting data from the wrong wearable device, so they should be removed from the data. These 20 people are just like the others in every other respect.





(1) It creates neither of the two errors. (The twenty users to be removed is just like the others in every other respect, so variance doesn't change)

(2) Reducing numbers of sample, then n declines.

(3) The t-value will decrease(t=diff/(sd/sqrt(n)). From the graph, we can see that the t-value declines, and causes the alternative distribution to move left. Since power is the area of alternative distribution from the cut-off value(95%), **the power will decrease.**

(4) From (3),we know that the power of this scenario declines. Beta is the probability of Type II error, if power(1-beta) declines, meaning that beta increases, so **Type II error will becomes more likely.**

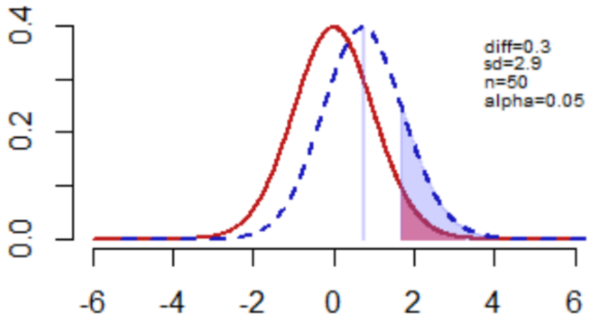
c. A very annoying professor visiting your company has criticized your colleague's "95% confidence" criteria, and has suggested relaxing it to just 90%.

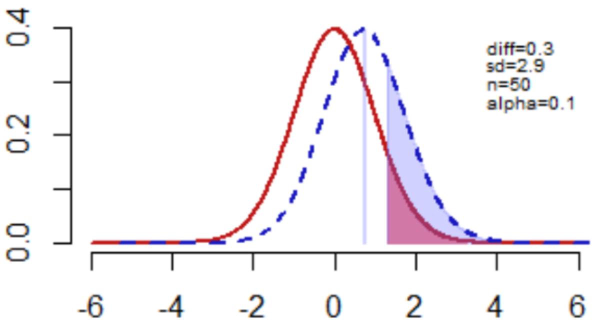
(1) It creates neither of the two errors.

(2) Adjusting the confidence criteria means that **the alpha will rises from 5% to 10%.** Therefore,we compare the original data with the higher alpha one.

(3) The cut-off value falls. Since power is the area of alternative distribution from the cut-off value(90%),if the cur-off value decreases, then **the power will increase.** We can also observe it in the graph.

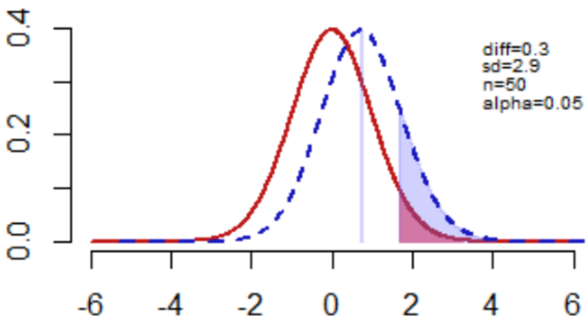
(4) From (2), we can see that alpha increases, which makes **type I error more likely**.

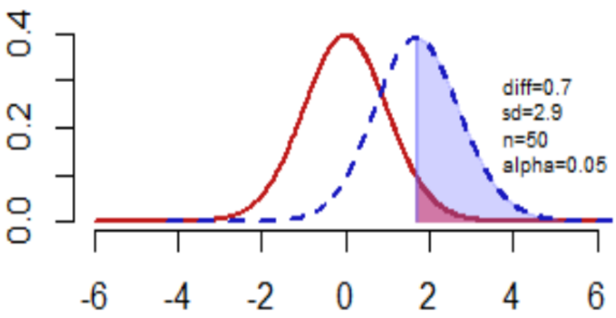




d. Your colleague has measured usage times on five weekdays and taken a daily average. But you feel this will underreport usage for younger people who are very active on weekends, whereas it over-reports usage of older users.

(1) It creates systematic error.(bias increases)
(2) Measuring five days will increase the bias, since it will underreport younger people who are very active on weekends.
(3) The t-value increases(t= diff/(sd/sqrt(n))). Since power is the area of alternative distribution from the cut-off value(90%),if the alternative distribution moves right, then **the power will increase.** We can also observe it in the graph.
(4) Since bias increases, **type I error will be more likely**.





Question 2) Let's return to the *strictly fictional scenario* (but with real data) from last week's Verizon dataset. Imagine this time that Verizon claims that they *take no more than 7.6 minutes on average* (single-tail test) to repair phone services for its customers. The file verizon.csv has a recent sample of repair times collected by the New York Public Utilities Commission, who seeks to verify this claim at 99% confidence.

a. Recreate the traditional hypothesis test of last week using high-level built-in functions of R:

   i. Use the `t.test()` function to conduct a one-sample, one-tailed t-test: report 99% confidence interval of the mean, t-value, and p-value

```
> data<-read.csv("verizon.csv")$Time
> hyp_mean<-7.6
> t.test(data,mu=hyp_mean,alternative="less",conf.level=0.99)

	One Sample t-test

data:  data
t = 2.5608, df = 1686, p-value = 0.9947
alternative hypothesis: true mean is less than 7.6
99 percent confidence interval:
     -Inf 9.360414
sample estimates:
mean of x
 8.522009
```

   ii. Use the `power.t.test()` function to tell us the power of the test

```
> #2-a-ii
> power.t.test(n=length(data),delta=mean(data)-
hyp_mean,sd=sd(data),sig.level = 0.01,alternative ="one.sided")

     Two-sample t test power calculation

              n = 1687
          delta = 0.9220095
             sd = 14.78848
      sig.level = 0.01
          power = 0.3028078
    alternative = one.sided
```

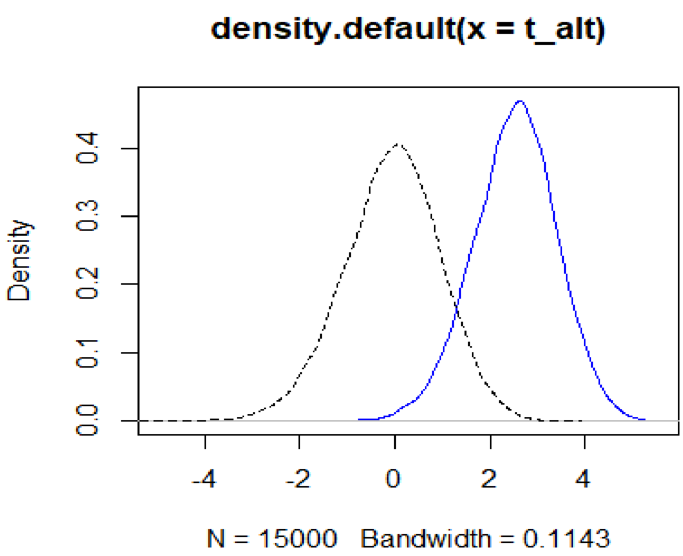
b. Let's use *bootstrapped hypothesis testing* to re-examine this problem:

i. Retrieve the original t-value from traditional methods (above)

```
> #2-b-i
> t_val<-(mean(data)-hyp_mean)/(sd(data)/sqrt(length(data)))
> t_val
[1] 2.560762
```

ii. Bootstrap the null and alternative t-distributions

```
> #2-b-ii
> bootstrap_null_alt<-function(sample0,hyp_mean){
+   resample<-sample(sample0,length(sample0),replace = T)
+   resample_se<-sd(resample)/sqrt(length(resample))
+   alt<-(mean(resample)-hyp_mean)/resample_se
+   null<-(mean(resample)-mean(sample0))/resample_se
+   c(alt,null)
+ }
> bootstrap_data<-
replicate(15000,bootstrap_null_alt(data,hyp_mean))
> t_alt<-bootstrap_data[1,]
> t_null<-bootstrap_data[2,]
> plot(density(t_alt),col="blue",xlim=range(t_null,t_alt))
> lines(density(t_null),lty="dashed")
```



iii. Find the 99% cutoff value for critical null values of t (from the bootstrapped null); What should our test conclude when comparing the original t-value to the 99% cutoff value?

```
> #2-a-iii
> cutoff_99<-quantile(t_null,probs=0.99) > cutoff_99
     99%
2.165597
```

The original t-value(2.560762) is bigger than the 99% cutoff value(2.165597), which means the t-value isn't under the confidence level we set. Therefore, **we can reject Verizon's claim**(takes no more than 7.6 minutes on average to repair phone services for its customers.).

iv. Compute the p-value and power of our bootstrapped test

```
> #2-b-iv
> #compute null probability distribution
> null_prob<-ecdf(t_null)#the cumulative prob dist for t_null
data
> p_val<-1-null_prob(t_val)
> p_val
[1] 0.002466667
> #compute alternative probability distribution and
> alt_prob<-ecdf(t_alt)
> power<-1-alt_prob(cutoff_99)
> power
[1] 0.6703333
```

The p-value is under 0.05 and the power is bigger than 50%, both of them indicate that **we can reject the null hypothesis**.