STA305/1004 - 2017 Homework #2 - Solutions

1. In a noninferiority trial, the new treatment is tested to investigate if it is at least similar to an existing therapy in terms of efficacy. In other words, a noninferiority trial is interested in examining whether a new treatment is not worse than the standard treatment by a prespecified noninferiority margin $\delta < 0$. The threshold δ specifies the lower bound beyond which the experimental drug is considered unacceptably inferior to the standard drug. The hypothesis test of interest is

$$H_0 : \theta \leq \delta \text{ versus } H_1 : \theta > \delta,$$

where, $\theta = \mu_1 - \mu_2$, the difference in means between the experimental treatment (treatment 1) and the standard of care (treatment 2). Rejecting H_0 means that the experimental treatment is claimed to be noninferior to (not worse than) the standard treatment. For example if $\theta = 0$ and $\delta = -0.5$ then the alternative hypothesis states that the experimental treatment is noninferior to the standard treatment if the mean difference of the outcome in the experimental versus the standard arm is at least -0.5.

Assume that the outcomes of a noninferiority trial are $Y_{1j} \sim N\left(\mu_1, \sigma^2\right), j = 1, \ldots, n_1$ in the group receiving the experimental treatment with, and the outcomes in the group receiving the standard treatment are $Y_{2j} \sim N\left(\mu_2, \sigma^2\right), j = 1, \ldots, n_2$. In the following questions assume that σ^2 is known.

(a) Specify the values of the test statistic for which the null hypothesis is rejected (i.e., the rejection region of the test) at level α.

H_0 is rejected at level α if

$$\frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} - \delta}{\sigma\sqrt{(1/n_1 + 1/n_2)}} > z_\alpha,$$

where $\bar{Y}_{i\cdot} = \frac{1}{n_1}\sum_{k=1}^{n_1} Y_{ik}, i = 1, 2$, and z_α is the $100(1-\alpha)$ percentile of the $N(0,1)$.

(b) Show that the power of this test is given by

$$\Phi\left(\left(\frac{\theta - \delta}{\sigma\sqrt{1/n_1 + 1/n_2}}\right) - z_\alpha\right),$$

for $\theta > \delta$, where $\Phi(\cdot)$ is the CDF, and z_α is the $100(1-\alpha)^{th}$ percentile of the $N(0,1)$.

Assume that H_1 is true, and β is the probability of a type II error. Then $\theta > \delta$, and the probability that the test rejects is:

$$\begin{aligned}
1 - \beta &= P\left(\frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} - \delta}{\tilde{\sigma}} > z_\alpha\right) \\
&= P\left(\frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\tilde{\sigma}} > z_\alpha + \frac{\delta}{\tilde{\sigma}}\right) \\
&= P\left(\frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\tilde{\sigma}} - \frac{\theta}{\tilde{\sigma}} > z_\alpha + \frac{\delta}{\tilde{\sigma}} - \frac{\theta}{\tilde{\sigma}}\right) \\
&= P\left(Z > z_\alpha + \frac{\delta}{\tilde{\sigma}} - \frac{\theta}{\tilde{\sigma}}\right) \\
&= P\left(Z < -z_\alpha - \frac{\delta}{\tilde{\sigma}} + \frac{\theta}{\tilde{\sigma}}\right) \\
&= \Phi\left(\frac{\theta - \delta}{\tilde{\sigma}} - z_\alpha\right),
\end{aligned}$$

where, $\tilde{\sigma} = \sigma\sqrt{(1/n_1 + 1/n_2)}$.

(c) If the numbers of patients in the two groups are different with $n_1 = rn_2$ then show that

$$n_2 = \frac{(1+1/r)\sigma^2(z_\alpha + z_\beta)^2}{(\theta - \delta)^2},$$

where β is the probability of a type II error, and $\theta > \delta$. (HINT: Use the power function from the previous part)

$$\begin{aligned}
1 - \beta &= \Phi\left(\frac{\theta - \delta}{\tilde{\sigma}} - z_\alpha\right) \\
&\Rightarrow \Phi^{-1}\left(1 - \beta\right) = \frac{\theta - \delta}{\tilde{\sigma}} - z_\alpha \\
&\Rightarrow z_\beta + z_\alpha = \frac{\theta - \delta}{\tilde{\sigma}} \\
&\Rightarrow z_\beta + z_\alpha = \frac{\theta - \delta}{\sigma\sqrt{(1/rn_2 + 1/n_2)}} \\
&\Rightarrow \frac{1}{n_2}\left(\frac{1}{r} + 1\right) = \left(\frac{\theta - \delta}{\sigma\left(z_\beta + z_\alpha\right)}\right)^2 \\
&\Rightarrow n_2 = \frac{(1+1/r)\sigma^2(z_\alpha + z_\beta)^2}{(\theta - \delta)^2}
\end{aligned}$$

(d) A pharmaceutical company is interested in establishing non-inferiority of a test drug compared to the standard treatment. The primary efficacy parameter is the percent change (from the beginning of the study) in LDL - low density lipidproteins. The company considers a difference of 0.05 to be clinically important, and assumes that the true difference in mean LDL between treatment groups is 0.0, and the standard deviation is 0.10. An equal number of subjects will be recruited into both treatment groups. How many subjects should be recruited into each treatment group so that the study has 90% power at the 5% significance level? Use R to calculate the answer using the formula in the previous part of the question. Hand in your R code

In this case $r = 1, \alpha = 0.05, \theta = 0.01, \delta = -0.05, \beta = 0.1, \sigma^2 = (0.1)^2$

```
((1+1)*(0.10)^2*(qnorm(1-.05)+qnorm(1-.1))^2)/(0-(-.05))^2
```

```
## [1] 68.51078
```

2. An engineer would like to design a study to compare the average lifetime of a certain electronic component under two different conditions. Components will be randomized to the two different conditions A and B. The distribution of the component's lifetime is known to follow an exponential distribution with rate λ. The density function of this distribution is

$$f(x) = \lambda \exp\left(-\lambda x\right), x \geq 0.$$

   The distribution depends on a single parameter $\lambda > 0$. The mean and standard deviation of the exponential distribution is λ^{-1}. This parameterization is in units corresponding to the reciprocal of time.

The engineer plans to compare the mean lifetimes between the two groups. She has hypothesized that the expected lifetime of the components under condition A is 10.00 months, and 6.67 months under condition B.

(a) If 100 batteries are tested in each group then what is the power of the two sample t-test to detect a difference in means at the 5% significance level? (HINT: A random sample from an exponential distribution can be generated in R using the function `rexp(n = ,rate = )`.) Hand in your R code.

In order to assume that the data follow an exponential distribution in the two-sample t-test we can simulate power. In group A $\lambda = 1/10$ and in group B $\lambda = 1/6.67$.

```r
set.seed(154)
N <- 1000
n <- 100
res <- numeric(N)
res <- replicate(N,t.test(rexp(n = n,rate = 1/10),
                          rexp(n = n,rate = 1/6.67),var.equal = F)$p.value)
sum(res<=0.05)/N
```

```
## [1] 0.803
```

The power of the two-sample t-test is approximately 0.803.

(b) The engineer contacts a statistician to check if her experiment is properly designed. The engineer tells the statistician her plan. The statistician quickly points out that one of the major assumptions of the t-test is that the data are assumed to be independent samples from normal distributions. She suggests using a nonparametric method called the Mann-Whitney Test (also called the Wilcoxon Rank sum test). This test doesn't assume that the data follow any particular distribution form. This test is implemented in R via the function `wilcox.test()`. For example, the P-value of the Mann-Whitney test for comparing the means of two random samples of size 50 from exponential distributions with $\lambda_1 = 0.1$, and $\lambda_2 = 0.2$ can be calculated in R.

```r
x <- rexp(n = 50,rate = 0.1)
y <- rexp(n = 50,rate = 0.2)
wilcox.test(x,y)$p.value # the P-value from the test
```

```
## [1] 0.0008586496
```

Calculate the power for the engineer's study if she uses the Mann-Whitney test to compare means. Hand in your R code.

We can use the same code as before except use the Mann-Whitney test instead of the t-test.

```r
set.seed(154)
N <- 1000
n <- 100
res <- numeric(N)
res <- replicate(N,wilcox.test(rexp(n = n,rate = 1/10),
                          rexp(n = n,rate = 1/6.67))$p.value)
sum(res<=0.05)/N
```

```
## [1] 0.694
```

The power of the Mann-Whitney test is approximately 0.694.

(c) Which statistical test should the engineer use to compare mean lifetimes under the two conditions? If she uses the test you recommend then will she be guaranteed that her experiment will detect a statistically significant difference between the two treatment groups? Explain your reasoning.

The engineer should use the t-test since the test is more powerful. What has been shown with these two power calculations is that even though the normality assumption of the t-test assumptions is not satisfied the t-test is robust against violation of the normality assumption. The Wilcoxon test does not assume that the data are normal, but the test clearly has lower power. So, if she uses the Wilcoxon test instead of a t-test then she will be less likely to detect a difference in lifetimes between the two conditions. Therefore, it makes sense to recommend the t-test.

She is not guaranteed to find a difference, but if a difference exists then she will have a higher probability of detecting it with the t-test. In particular, if the conditions don't have an effect on lifetime then she will only have a 5% chance of stating that there is a difference, but if a difference exists then she will have an 80% chance of detecting it with the t-test.

3. Consider the following hypothetical example. Suppose that eight patients were assigned to one of two medical operations by a physician: $Y(0)$ is the number of years lived after standard surgery; and $Y(1)$ is the number of years lived after new surgery. The hypothetical data given below shows all potential outcomes under the two different treatments.

$Y(0)$	$Y(1)$
13	14
6	0
4	1
5	2
6	3
6	1
8	10
8	9

The doctor assigns a patient to the surgical treatment (new or standard) such that the patient will maximize years their lived after surgery, and if there is no difference he flips a coin.

(a) Is the assignment mechanism that the doctor used ignorable? Explain your answer.

The assignment mechanism depends on the potential outcomes. Therefore, the assignment mechanism is nonignorable.

(b) What is the true average causal effect of treatment? Which treatment is better on average? Explain your answer.

The true average causal effect can be calculated by computing the mean difference.

```
Y0 <- c(13,6,4,5,6,6,8,8)
Y1 <- c(14,0,1,2,3,1,10,9)
mean(Y1)-mean(Y0)
```

```
## [1] -2
```

The true average causal effect is $\bar{Y}_1 - \bar{Y}_0$ =-2. This means that $\bar{Y}_1 = \bar{Y}_0 - 2$. In other words, patients live two years less on the new treatment compared to the old treatment. Therfore, the standard treatment is better.

(c) If we only observed the outcomes under the treatment that the doctor assigned then what is the observed treatment effect? Is it different compared to the true average causal treatment effect?

If we only observe the outcomes under the treatment that the doctor assigned then observed treatment effect is mean difference in the observed data.

```
Y0.obs <- c(6,4,5,6,6)
Y1.obs <- c(14,10,9)
mean(Y1.obs)-mean(Y0.obs)
```

```
## [1] 5.6
```

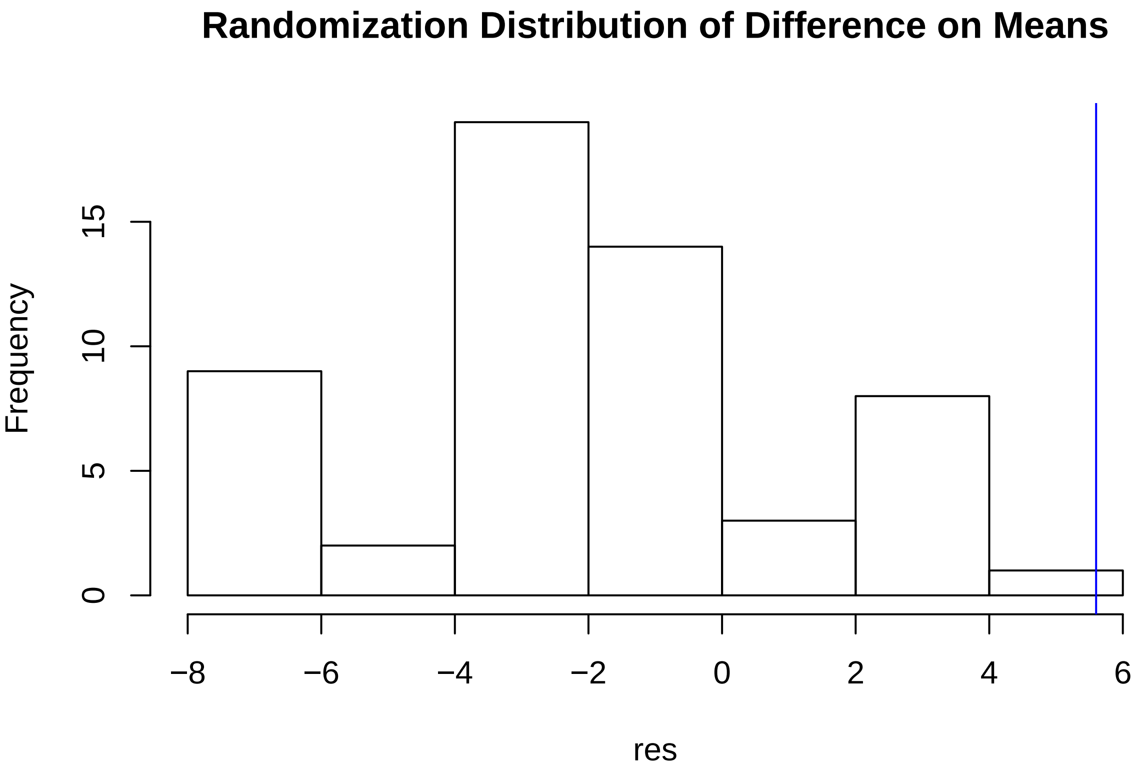
The observed mean difference is 5.6 $\neq$ -2. The observed mean difference is what we would observe if the study were analyzed as if patients were randomized to the two treatments.

(d) The data that you used to calculate the true treatment effect in (b) can be used to calculate the mean difference in all possible treatment assignments (the randomization distribution). If the treatment assignment would have been chosen at random then would the inference based on the difference of means be correct? Compare the latter to how the doctor assigned treatment. Is it misleading to use the difference in means to compare the treatments using the doctor's treatment assignment? Explain your reasoning.

In this question we know the true average difference between the treatments is -2, the observed difference is +5.6. The observed difference would lead to a conclusion that the new treatment adds 5.6 years of life compared to the old treatment. So, using the oberved difference to infer the true difference will be misleading.

The doctor assigned five subjects to the standard treatment and three subjects to the new treatment. This can be done in $\binom{8}{3} = \binom{8}{5} = 56$ ways. To calculate the randomization distribution calculate all 56 possible differences in means. This is done using the R code below.

```
Y0 <- c(13,6,4,5,6,6,8,8)
Y1 <- c(14,0,1,2,3,1,10,9)
mean.obs <- mean(Y1.obs)-mean(Y0.obs)
N <- choose(8,3)
index <-combn(1:8,3)
res <- numeric(N)
for (i in 1:N)
{
  res[i] <- mean(Y1[index[,i]])-mean(Y0[-index[,i]])
}
hist(res,main="Randomization Distribution of Difference on Means")
abline(v=mean.obs,col="blue")
```



```
tbar <- mean(res)
tbar
```

```
## [1] -2
```

The mean of the randomization distribution is -2 the same as the true mean difference. This implies that the inference would have be correct, on average, had our treatment assignment been randomly chosen from one of the $\binom{8}{3}$ treatment assignments. Drawing a treatment assignment at random does not depend on the potential outcomes thus the treatment assignment would be ignorable.

The doctor assigned treatment based on potential outcomes. We observed the most extreme assignment that even had the wrong sign. Therefore the difference in observed means is entirely misleading using the doctor's treatment assignment. The difference of the observed estimates is a good estimate of the average causal effect only if the assignment is chosen at random.