STA305 Homework 2

Rui Qiu c3qiurui

2016-02-19

Question 1

(a)

The probability requested in the problem can be interpreted as the power of the second test, that is, the probability of receiving a significant result in the second experiment against the alternative hypothesis defined by the result of the first experiment. In the special case of a test of a mean with known variance, one would compute the power of the test against the hypothesis that the population mean equals the mean of the first sample.

Since the sample size of the second experiment is half that of the first, the computed probability of obtaining $z > 1.645$ is shown as following:

```
pow.z.test <- function(alpha,mu1,mu0,sigma,n) {
  arg1 <- qnorm(1-alpha)-(mu1-mu0)/(sigma/sqrt(n))
  1-pnorm(arg1)
}
```

This is the code to calculate the power of one-tailed z-test, and we plug in the following values `alpha=0.05, mu1=2.1*1/sqrt(20), mu0=0, sigma=1, n=10` because:

$$\frac{\overline{X}-0}{\sigma/\sqrt{20}} = 2.1,$$
$$\overline{X} = 2.1 \cdot \frac{\sigma}{\sqrt{20}},$$

```
> pow.z.test(0.05,2.1/sqrt(20),0,1,10)
[1] 0.4364683
> pow.z.test(0.05,2.1*0.7/sqrt(20),0,0.7,10)
[1] 0.4364683
```

Note that a change of sigma does not affect the result in fact.
Besides by simulation, we can also calculate such power rigourously:

$$
\begin{aligned}
1-\beta &= P(\text{reject } H_0 \mid H_1 \text{ is true}) \\
&= P((\overline{X}-0) \geq \frac{1.645\sigma}{\sqrt{10}} \mid \mu = \frac{2.1\sigma}{\sqrt{20}}) \\
&= P(\frac{\overline{X}-2.1\sigma/\sqrt{20}}{\sigma/\sqrt{10}} \geq \frac{1.645\sigma/\sqrt{10}-2.1\sigma/\sqrt{20}}{\sigma/\sqrt{10}}) \\
&= P(z_2 \geq 0.16) \\
&= 0.436441
\end{aligned}
$$

Therefore the probability that the results of experiment #2 will be significant is at 5% level by a one-tailed z-test is around 0.44.

(b)

Let $\theta = \mu = \frac{z\sigma}{\sqrt{n_1}} = 2.1\sigma/\sqrt{20}$.
Then the power at this this θ is given by:

$$
\begin{aligned}
1-\beta &= P(z_2 \geq z_\alpha - \frac{z\sigma/\sqrt{n_1}}{\sigma\sqrt{\frac{1}{n_2}}}) \\
\beta &\approx \Phi(z_\alpha - \frac{z}{\sqrt{\frac{n_1}{n_2}}}) \\
z &= (z_\beta + z_\alpha)\sqrt{\frac{n_1}{n_2}} \\
z^2 &= (z_\beta + z_\alpha)^2 \frac{n_1}{n_2} \\
n_2 &= \left[\frac{(z_\beta + z_\alpha)\sqrt{n_1}}{z}\right]^2
\end{aligned}
$$

(c)

```r
sizefunc <- function(beta) {
  zbeta <- qnorm(1-beta)
  res <- ((1.645+zbeta)*sqrt(20)/2.1)^2
  return(res)
}
```

We are required to calculate the sample size for $1-\beta$ equal to $0.7, 0.8, 0.9$ respectively, i.e., $\beta = 0.3, 0.2, 0.1$.

```
> sizefunc(0.3)
[1] 21.34376
> sizefunc(0.2)
[1] 28.04211
> sizefunc(0.1)
[1] 38.8422
```

Therefore, in order to increase the probability of significant results, we should conduct the experiment #2 with larger sample size. Specifically, we need at least $22, 29, 39$ samples for power $0.7, 0.8$ and 0.9.

(d)

Since his original plan only has a power around 0.44, in other words, if the study was repeated 100 times, only 44 of the 100 studies will reject H_0. So the study is not good enough, it's highly possible that the research would not get a significant result as he did in experiment #1.

Question 2

(a)

According to the problem, the sample size of Acetazolamide arm is 300, and the sample size of placebo arm is 150.

Note that we want to determine whether acetazolamide reduces mechanical ventilation duration, so we are conducting a one-tailed two sample t-test.

So $\alpha = 0.05, n_1 = 150, \mu_1 = 27, n_2 = 300, \mu_2 = 24, \sigma = 10.$

```
twosampttestpow <- function(alpha,n1,n2,mu1,mu2,sigma) {
  delta <- mu1-mu2
  t.crit <- qt(1-alpha,n1+n2-2)
  t.gamma <- delta/(sigma*sqrt(1/n1+1/n2))
  t.power <- 1-pt(t.crit,n1+n2-2,ncp=t.gamma)
  return(t.power)
}
```

Plug in and calculate:

```
> twosampttestpow(0.05,150,300,27,24,10)
[1] 0.9115901
```

Therefore, the power of the two-sample t-test using the 5% significance level is 0.91.

(b)

Can calculate the power using a simulation of 50 times.

```
> set.seed(2301)
> pvals <- replicate(100, t.test(27+10*rt(n=150,df=5),24+10*rt(n=300,df=5), alternative = "
greater", var.equal=T)$p.value)
> sum(pvals<=0.05)/100
[1] 0.76
```

So the power (with duration in each arm following a location-scale t_5 distribution) is about 0.76.

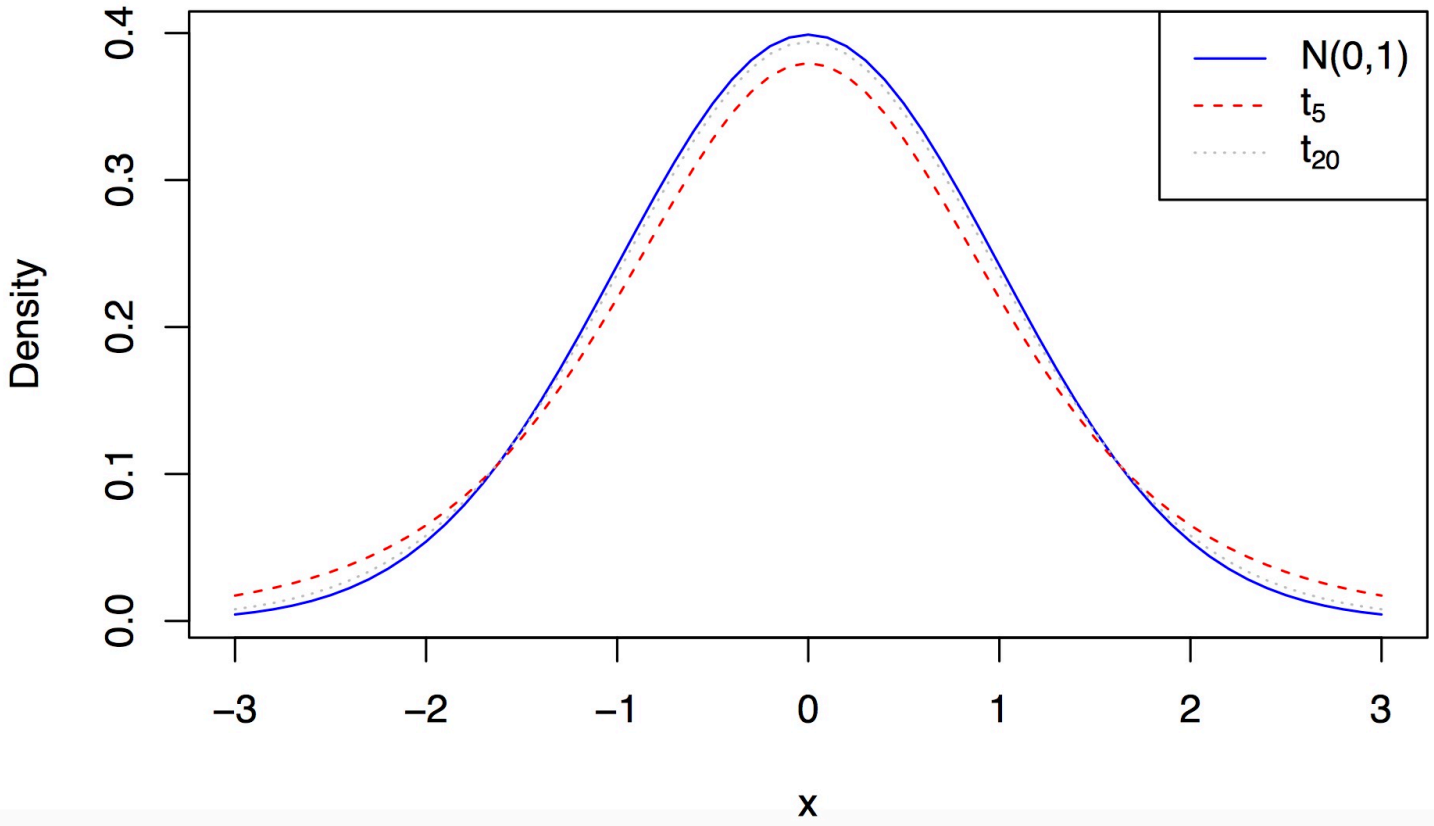
(c)

Similarly:

```
> set.seed(2302)
> pvals <- replicate(100, t.test(27+10*rt(n=150,df=20),24+10*rt(n=300,df=20),alternative =
"greater", var.equal=T)$p.value)
> sum(pvals<=0.05)/100
[1] 0.88
```

So this time, as the degree of freedom increases, the power is closer to the power of normal distribution. In fact, the power is 0.88.

(d)



As we can observe from the plot, the smaller degree of freedom `df` is, the larger tail(s) our target distribution has, so that the p-value is less likely to be smaller than our significance level 0.05, therefore, the power is smaller.

Thus, the case whose arms follow t_{20} distribution has larger power than its counterpart with t_5 distribution in arms. ($0.88 > 0.76$.)

And if we compare t-distribution with normal distribution, it is not hard to find that both t-distributions (in this case) have larger tail(s) than normal distribution, so we have different powers.

Question 3

(a)

Student	$Y(0)$	$Y(1)$	T_i
1	**95**	79	0
2	76	**77**	1
3	**61**	54	0
4	**53**	47	0
5	**71**	69	0
6	81	**82**	1
7	65	**70**	1

The data we actually would observe under such assignment mechanism is $Y(T_i)$, which only contains 7 different values and

- the observed average grade of all traditional course takers is $\overline{Y(0)} = (95 + 61 + 53 + 71)/4 = 70.$
- the observed average grade of all inverted course takers is $\overline{Y(1)} = (77 + 82 + 70)/3 = 76.33.$

(b)

Yes, it is **non-ignorable** because the assignment mechanism is not independent of potential outcomes.

(c)

No, we **cannot draw** the correct causal conclusion about the effectiveness of two methods based on the observed data. Specifically, by comparing the two sample means, we have $\overline{Y(1)} > \overline{Y(0)}$ indicating the inverted class has higher students grades than traditional class.
However, if we calculate the two sample means of the professor's perfect predicted data,
$72 = \overline{Y(0)} > \overline{Y(1)} = 68.$
This is because the observed data is misleading, as we pretend we are having unconfounded treatment but in fact we did not.

(d)

```
> y0 <- c(76,53,65)
> y1 <- c(79,54,69,82)
> grades <- c(y0,y1)
> N <- choose(7,3) # 3 students in traditional, 4 students in inverted
> res <- numeric(N)
> library(combinat)
> index <- combn(1:7,3)
> for (i in 1:N) {
+     res[i] <- mean(grades[index[,i]])-mean(grades[-index[,i]])
+ }
> obs <- mean(y0)-mean(y1)
> sum(res<=obs)/N
[1] 0.2285714
```

Note that the professor assigned 3 students to traditional class and 4 students to inverted class, then there are in total 35 (7 choose 3) possible treatment assignments.

The assumption is $Y_i(0) = Y_i(1) = Y^{\text{obs}}$ for all i, i.e., there is no difference between the grades from two different methods on the same unit.

Consequently, the null hypothesis H_0 in the randomization test is that two sample means $\mu_0 = \mu_1$.

The assumption is **valid** because in reality, at most one of the potential outcomes can be realized and thus observed, so we can 'assume' potential outcomes are the same.

The p-value calculated is $0.23 > 0.05$, so we fail to reject null hypothesis. Thus this randomization **does not lead to the correct causal inference** because we cannot see the difference through such test.

Question 4

(a)

Proof:
Our goal is to prove that the distribution of the Z and observable variable $\boldsymbol{x}$ are conditional independent on propensity score $e = e(\boldsymbol{x}) = P(Z = 1 \mid \boldsymbol{x})$.

Recall the definition of propensity score, which is:

$$e(\boldsymbol{x}) = P(Z = 1 \mid \boldsymbol{x}) = E[Z \mid \boldsymbol{x}].$$

As Z is a binary variable, the distribution of Z is determined by its mean. Therefore, once conditioned on propensity score $e(\boldsymbol{x})$, it is irrelevant whether the mean of Z is computed further conditioning on $\boldsymbol{x}$ or not. It is equivalent to show:

$$E[Z \mid \boldsymbol{x}, e] = E[Z \mid e] \cdots \cdots \cdots \cdots \star$$

Let's consider the LHS of $\star$ first:

$$E[Z \mid \boldsymbol{x}, e] = E[Z \mid \boldsymbol{x}] = e$$

The first equivalence is true because e is a function of $\boldsymbol{x}$, so if conditioning on $\boldsymbol{x}$, then information about e is redundant.

The second equivalence is true because of the definition of propensity score 1.

Now we consider the RHS of the starred equation. We want to show $E[Z \mid e] = e$ obviously. Here we need to use $E(X) = E(E(X \mid Y))$, where X can be defined as $Z|e$, Y can be defined as $\boldsymbol{x}|e$.

Plug in, we get:

$$
\begin{aligned}
E[Z \mid e] &= E\{E[Z \mid \boldsymbol{x}, e] \mid e\} \\
&= E[e \mid e] \\
&= e
\end{aligned}
$$

So we proved $E[Z \mid \boldsymbol{x}, e] = E[Z \mid e]$, then we are done.

Why is this $\star$ equivalent to $P(\boldsymbol{x}, Z \mid e) = P(\boldsymbol{x} \mid e)P(Z \mid e)$?

Because if $\star$ is satisfied, Z has nothing to do with $\boldsymbol{x}$ when conditioning on e.

(b)

Propensity score matching balances the observed covariates.
But randomization balances the observed, unobserved covariates, and even the potential responses.