Lab 02A: Week 5 (Solutions)

Contents

1 Quick quiz
- 1.1 Bias
- 1.2 Hypothesis test

2 Group work (Power)

3 Questions
- 3.1 Blood alcohol readings
- 3.2 Life satisfaction
- 3.3 Power, effect size and sample size

4 For practice after the computer lab
- 4.1 Git + GitHub
- 4.2 Number of televisions per household
- 4.3 Drawing conclusions from rejection regions

The **specific aims** of this lab are:

- to practice statistical thinking with summary statistics
- to learn some practical skills related to data cleaning
- to produce a data analytic report in R Markdown
- to review hypothesis testing framework in the context of testing for means

The unit **learning outcomes** addressed are:

- LO1 Formulate domain/context specific questions and identify appropriate statistical analysis.
- LO2 Extract and combine data from multiple data resources.
- LO3 Construct, interpret and compare numerical and graphical summaries of different data types including large and/or complex data sets.
- LO5 Identify, justify and implement appropriate parametric or non-parametric statistical tests.
- LO8 Create a reproducible report to communicate outcomes using a programming language.

1 Quick quiz

1.1 Bias

1.1 Bias

After watching the biggest loser on TV, Tommy became self conscious about his size and booked in to the doctor for a check up. Tommy went to the pub that week and realised most of his friends at the pub were bigger than him. This survey made Tommy conclude that he was smaller than average. Because of this conclusion Tommy cancelled his doctor appointment. Tommy's health is now at risk because he was not aware of what kind of bias?

a. Selection bias

b. Measurement bias

c. Response bias

d. Confounding

a. Selection bias

1.2 Hypothesis test

Suppose that Sandy wishes to test the hypotheses $H_0\colon\ \mu = 60$ against $H_1\colon\ \mu > 60$ for a normal population with an unknown standard deviation. A random sample of size 30 from this population gave her a mean of 63 and a standard deviation of 9. The observed value, $t_0 = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$, of the test statistic under the null hypothesis is:

a. 2.03

b. 0.20

c. 1.83

d. 0.18

e. 2.83

c. $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{63 - 60}{9/\sqrt{30}} = 1.8257$

Hence, Sandy should perform a

a. one-sided t test and calculate the p-value using $P(t_{29} < t_0)$

b. two-sided t test and calculate the p-value using $2P(t_{29} > t_0)$

c. two-sided t test and calculate the p-value using $P(t_{29} \neq t_0)$

d. one-sided t test and calculate the p-value using $P(t_{29} > t_0)$

e. one-sided t test and calculate the p-value using $P(|\bar{X} - 60| > t_0)$

d. one-sided t test and calculate the p-value using $P(t_{29} > 1.83)$

2 Group work (Power)

In your break out rooms, consider the following questions.

1. What are the two types of errors you can make when performing a hypothesis test?
2. What is the **power** of a test?
3. When doing a simple test for a mean, discuss how varying the level of significance α, the alternative population mean μ_1, the sample size n and the population standard deviation σ could affect $\text{Power}(\mu_1)$. This web app might help you think about this visually.
4. Suppose you have n observations, $X_1, X_2, \ldots, X_n$, are independent and identically distributed $N(\mu, 1)$, and we wish to test $H_0\colon\ \mu = 0$ vs $H_1\colon\ \mu > 0$. Use the `pwr.t.test()` function from the **pwr** package to find the smallest sample size, n, we could choose to be at least 90% sure of finding evidence at the 5% level when $\mu = 1$?

1. The power of a (binary) hypothesis test is the probability that the test correctly rejects the null hypothesis when a specific alternative hypothesis is true: $\text{power} = P(\text{reject } H_0 \mid H_1 \text{ is true})$.
2. As α increases, the critical value *decreases* and the power increases. As μ_1 increases the power increases. As n increases, the power increases. As σ increases, the power decreases. When using the web app, note that Cohen's d is the difference between the means divided by the standard deviation.
3. $n = 11$ is the smallest sample size required to achieve 90% power with a 5% significance level and Cohen's d of 1.

```
pwr::pwr.t.test(d = 1, sig.level = 0.05, power = 0.9, type = "one.sample",
    alternative = "greater")
```

```

     One-sample t test power calculation 

              n = 10.08107
              d = 1
      sig.level = 0.05
          power = 0.9
    alternative = greater
```

3 Questions

3.1 Blood alcohol readings

Source: Larsen and Marx (2012Question 6.2.7).

The following are 30 blood alcohol determinations made by Analyzer GTE-10, a three-year-old unit that may be in need of re-calibration. All 30 measurements were made using a test sample on which a properly adjusted machine would give a reading of 12.6%.

```
bac = c(12.3, 12.7, 12.6, 13.1, 13.2, 12.8, 13.1, 12.9, 13.1, 12.4, 13.6,
    12.7, 12.6, 13.1, 12.4, 12.6, 13.3, 12.6, 12.4, 13.1, 12.9, 12.6, 12.7,
    12.5, 12.4, 12.4, 12.6, 12.7, 12.4, 12.9)
n = length(bac)
xbar = mean(bac)
s = sd(bac)
c(n, xbar, s)
```

```
[1] 30.0000000 12.7566667  0.3244978
```

a. Let μ denote the true average reading that the Analyzer GTE-10 would give on the test sample. Write out the hypothesis for the Analyzer GTE-10 being faulty.

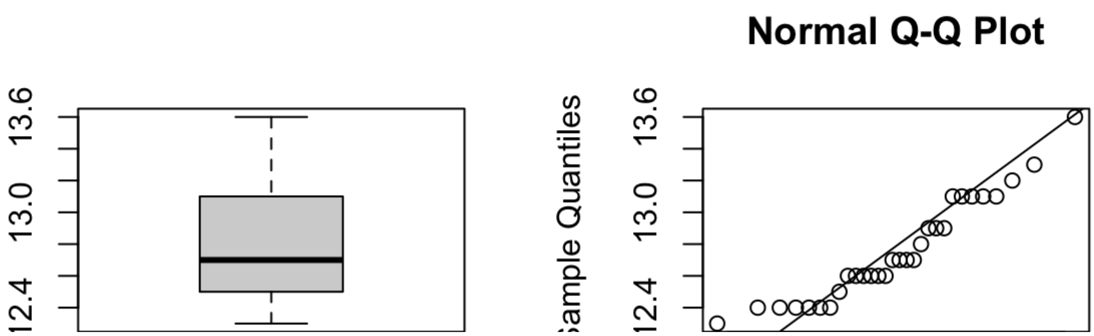
Hypothesis: $H_0\colon\ \mu = 12.6$ vs $H_1\colon\ \mu \neq 12.6$

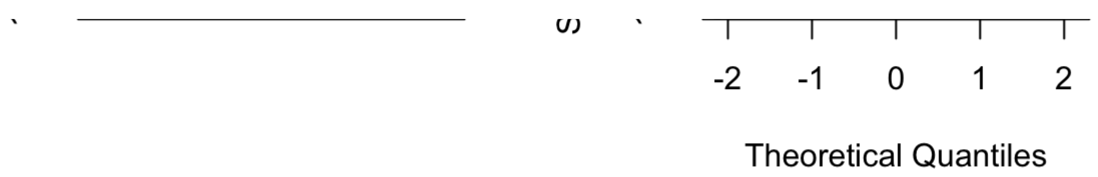
b. What are the assumptions of this test? Are they satisfied?

Assumptions: X_i are iid rv and follow $N(\mu, \sigma^2)$.

We can check for the normality assumption using a boxplot (looking for symmetry) or a normal quantile-quantile plot (looking for the points being close to the straight line).

```
par(mfrow = c(1, 2))
boxplot(bac)
qqnorm(bac)
qqline(bac)
```





There's some indication of some skewness in the boxplot and QQ plot, but it's not too bad, so we can say the normality assumption is approximately satisfied.

c. Assuming that the readings are normally distributed, the test statistic will follow a t distribution. How many degrees of freedom will the test have?

Test statistic: $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$. Under H_0, $T \sim t_{n-1}$. So the degrees of freedom are $30 - 1 = 29$.

d. Calculate the observed test statistic.

Observed test statistic: $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{12.75 - 12.6}{0.324/\sqrt{30}} = 2.644$

```
t0 = (mean(bac) - 12.6)/(sd(bac)/sqrt(length(bac)))
t0
```

```
[1] 2.644389
```

e. The critical value at the 5% level of significance is 2.045. At the level of significance $\alpha = 0.05$, what is your conclusion?

In an exam you might be given this sort of output:

```
qt(c(0.9, 0.95, 0.975), 29)
```

```
[1] 1.311434 1.699127 2.045230
```

```
qt(c(0.9, 0.95, 0.975), 30)
```

```
[1] 1.310415 1.697261 2.042272
```

Decision: Reject H_0 in favour of H_1 as the observed test statistic is larger than the critical value.

f. Perform the test in R using the `t.test()` function.

```
t.test(bac, mu = 12.6, alternative = "two.sided")
```

```
	One Sample t-test

data:  bac
t = 2.6444, df = 29, p-value = 0.01307
alternative hypothesis: true mean is not equal to 12.6
95 percent confidence interval:
 12.63550 12.87784
sample estimates:
mean of x
 12.75667
```

3.2 Life satisfaction

A research study was conducted to examine the differences between older and younger adults on perceived life satisfaction. Thirty older adults (between 40 and 50) and thirty younger adults (between 20 and 30) were give a life satisfaction test (known to have high reliability and validity). Scores on the measure range from 0 to 60 with high scores indicative of high life satisfaction with low scores indicative of low life satisfaction. Carefully study the R output below before answering the questions.

```r
Young = c(24, 26, 40, 29, 29, 41, 32, 19, 23, 25, 37, 31, 31, 29, 24,
          42, 32, 13, 33, 25, 20, 26, 20, 23, 23, 15, 34, 29, 20, 38)
Old = c(27, 26, 45, 34, 34, 45, 36, 20, 22, 24, 35, 31, 26, 41, 31,
        37, 31, 12, 38, 26, 22, 27, 21, 31, 23, 24, 27, 33, 22, 40)
# old fashioned method
c(length(Young), length(Old))
```

```
[1] 30 30
```

```r
c(mean(Old), mean(Young), mean(Old-Young))
```

```
[1] 29.700000 27.766667  1.933333
```

```r
c(sd(Old), sd(Young), sd(Old-Young))
```

```
[1] 7.835155 7.435624 4.176563
```

```r
# tidyverse method
library("tidyverse")
dat = tibble(
  satisfaction = c(Young, Old),
  age = c(rep("Young", length(Young)),
          rep("Old", length(Old)))
)
dat %>%
  group_by(age) %>%
  summarise(
```

```
  summarise(
    n = n(),
    mean = mean(satisfaction),
    sd = sd(satisfaction)
  )
```

```
# A tibble: 2 × 4
  age         n  mean    sd
  <chr> <int> <dbl> <dbl>
1 Old      30  29.7  7.84
2 Young    30  27.8  7.44
```

```
qt(c(0.025,0.05,0.1),58)
```

```
[1] -2.001717 -1.671553 -1.296319
```

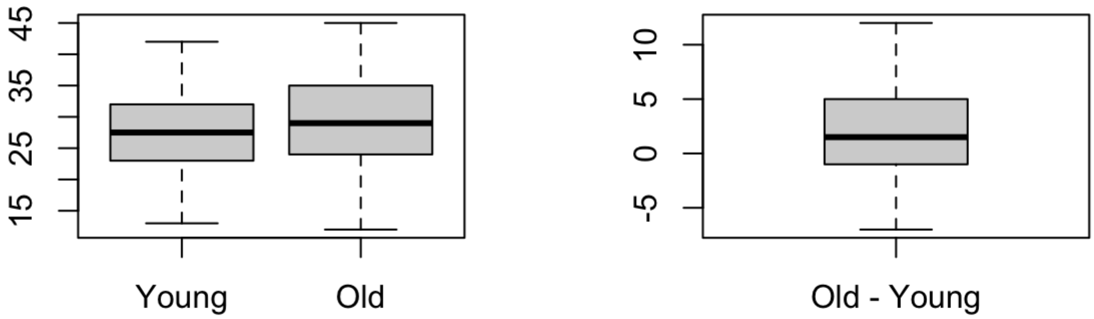
```
qnorm(c(0.025,0.05,0.1))
```

```
[1] -1.959964 -1.644854 -1.281552
```

```
qt(c(0.025,0.05,0.1),29)
```

```
[1] -2.045230 -1.699127 -1.311434
```

```
par(mfrow = c(1,2))
boxplot(Young,Old,names = c('Young','Old'))
boxplot(Old - Young)
axis(1, at = 1, labels = 'Old - Young')
```



a. Which hypothesis test should be used?

Two sample t-test.

b. State the null and alternative hypotheses.

Let μ_{O} and μ_{Y} be the population mean perceived life satisfaction levels for old and young people respectively.

$H_0\colon\ \mu_{\text{O}} - \mu_{\text{Y}} = 0$ vs $H_1\colon\ \mu_{\text{O}} - \mu_{\text{Y}} \neq 0$

Or $H_0\colon\ \mu_{\text{O}} = \mu_{\text{Y}}$ vs $H_1\colon\ \mu_{\text{O}} \neq \mu_{\text{Y}}$.

c. What assumptions have been made in order for the test to be valid? Are they reasonable? Give reasons.

Assumptions: - $X_1, \ldots, X_n$ are iid $N(\mu_X, \sigma^2)$ - $Y_1, \ldots, Y_n$ are iid $N(\mu_Y, \sigma^2)$ - X_i and Y_i are independent of each other

- Both boxplots for young and old look symmetric with no outliers - justifying the assumption of normality (or could mention that n is large)
- Young seems to have slightly lower spread than old, but not concerning for assumption of equal variance.
- Independence assumption doesn't appear to be violated based on the way the data was collected in the question proposed.

d. Calculate the appropriate test statistic.

$$n_{\text{O}} = n_{\text{Y}} = 30$$

$$s_p = \sqrt{\frac{(n_{\text{Y}}-1)s_C^2+(n_{\text{O}}-1)s_M^2}{n_{\text{Y}}+n_{\text{O}}-2}} = \sqrt{\frac{29\times 7.435624^2+29\times 7.835155^2}{58}} = 7.638002$$

$$t_0 = \frac{29.7-27.76667}{7.638002\sqrt{\frac{1}{30}+\frac{1}{30}}} = 0.98033$$

d. What is the conclusion and why?

```
qt(c(0.9, 0.95, 0.975), 58)
```

```
[1] 1.296319 1.671553 2.001717
```

The observed test statistic is less than 2.00 so we do not reject H_0 at the 5% level of significance.

f. Perform the test in R using the `t.test()` function.

```r
t.test(Old, Young, var.equal = TRUE)
```

```

	Two Sample t-test

data:  Old and Young
t = 0.98033, df = 58, p-value = 0.331
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.014301  5.880968
sample estimates:
mean of x mean of y 
 29.70000  27.76667 
```

```r
t.test(satisfaction ~ age, data = dat, var.equal = TRUE)
```

```

	Two Sample t-test

data:  satisfaction by age
t = 0.98033, df = 58, p-value = 0.331
alternative hypothesis: true difference in means between group Old and group Young is not eq
95 percent confidence interval:
 -2.014301  5.880968
sample estimates:
  mean in group Old mean in group Young 
           29.70000            27.76667 
```

After talking to the person who is in charge of the study, you find out that the old and young adults are NOT independent groups. Instead, thirty young adults were given the life satisfaction test when they enrolled in the study and then given the test again ten years later.

```r
paired_dat = tibble(Young, Old) %>%
  mutate(
    Difference = Old - Young
  )
glimpse(paired_dat)
```

```
Rows: 30
Columns: 3
$ Young      <dbl> 24, 26, 40, 29, 29, 41, 32, 19, 23, 25, 37, 31, 3…
$ Old        <dbl> 27, 26, 45, 34, 34, 45, 36, 20, 22, 24, 35, 31, 2…
$ Difference <dbl> 3, 0, 5, 5, 5, 4, 4, 1, -1, -1, -2, 0, -5, 12, 7,…
```

```r
paired_dat %>%
  summarise(
    n = n(),
    mean_old = mean(Old),
    mean_young = mean(Young),
    mean_diff = mean(Difference),
```

```
    sd_old = sd(Old),
    sd_young = sd(Young),
    sd_diff = sd(Difference)
  )
```

```
# A tibble: 1 × 7
      n mean_old mean_young mean_diff sd_old sd_young sd_diff
  <int>    <dbl>      <dbl>     <dbl>  <dbl>    <dbl>   <dbl>
1    30     29.7       27.8      1.93   7.84     7.44    4.18
```

g. Explain why the previous test is now inappropriate.

The observations are paired and thus no longer independent.

h. Which hypothesis test should now be used?

We should used a paired t-test.

i. Repeat steps b-e.

Let μ_d be the population mean difference in perceived life satisfaction levels between when someone is old and young.

H_0: $\mu_\mathrm{d} = 0$ vs H_1: $\mu_\mathrm{d} \neq 0$.

$t_0 = \frac{1.93-0}{4.17/\sqrt{30}} = 2.5354$

$P(t_{29} < -2.04) = 0.025$, therefore $P(|t_{29}| > 2.5354) < 0.05$

There is enough evidence to reject the null hypothesis.

j. Perform the test in R using the `t.test()` function.

```
t.test(paired_dat$Difference)
```

```
	One Sample t-test

data:  paired_dat$Difference
t = 2.5354, df = 29, p-value = 0.01688
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.3737791 3.4928876
sample estimates:
mean of x
 1.933333
```

3.3 Power, effect size and sample size

These questions have been adapted from chapter 11 of Nordmann and McAleer (2021).

3.3.1 Calculating effect size given a test statistic

The effect size (Cohen's d) is $|\mu_1 - \mu_0|/\sigma$. According to Cohen (1988), the effect size is small (.2 to .5), medium (.5 to .8) or large (> .8). In the context of a one sample t-test we can estimate Cohen's d from $t_0 = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$, the observed test statistic, as $d = t_0/\sqrt{n}$.

You run a one-sample t-test and discover a significant effect, R reports the test statistic to be 3.24 with 25 degrees of freedom and $p < .05$. Using the above formulas, calculate d and determine whether the effect size is small, medium or large.

```
n = 25 + 1
d = 3.24/sqrt(n)
d
```

```
[1] 0.6354163
```

The effect size, 0.64, is categorised as a medium effect size.

3.3.2 Using `pwr.t.test()`

Using `pwr.t.test()`, find how many people per group (n) you would need to detect an effect size of $d = .4$ with 80% power at the 5% significance level in a two sample t-test with a two sided alternative hypothesis. How does this change as the effect size changes (e.g. from 0.1 to 0.8)? Plot the sample size required to achieve a power of 80% against effect size.

```
library(pwr)
pwr.t.test(d = 0.4, power = 0.8, sig.level = 0.05, alternative = "two.sided",
    type = "two.sample")
```

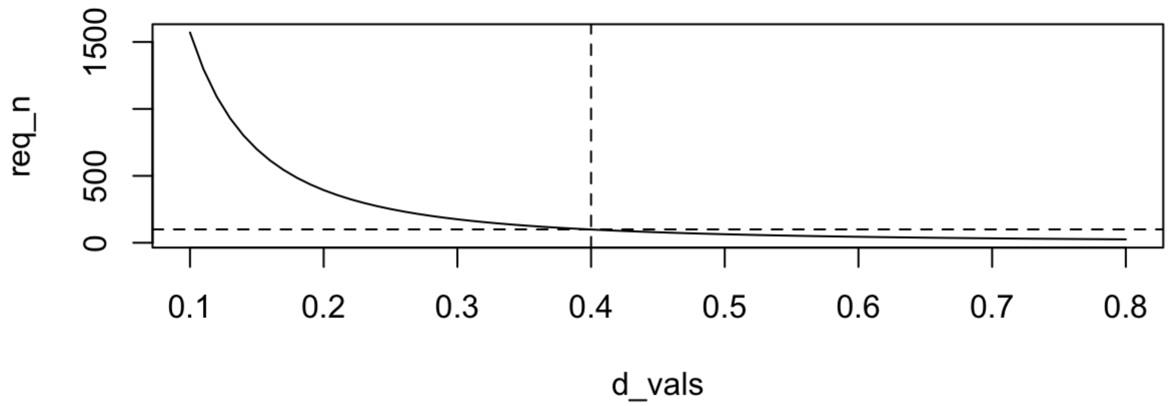
```
     Two-sample t test power calculation

              n = 99.08032
              d = 0.4
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

```
NOTE: n is number in *each* group
```

We would need 100 people in each group to achieve a power of at least 80%.

```
d_vals = seq(0.1, 0.8, 0.01)
req_n = NULL
for (i in seq_along(d_vals)) {
    req_n[i] = pwr.t.test(d = d_vals[i], power = 0.8, sig.level = 0.05,
        alternative = "two.sided", type = "two.sample")$n
}
plot(d_vals, req_n, type = "l")
abline(v = 0.4, lty = 2)
abline(h = 100, lty = 2)
```



As the effect size decreases, the sample size required to achieve 80% power increases dramatically.

4 For practice after the computer lab

Here are some additional questions for you to attempt in your own time.

- Look at questions 6.2.1 to 6.2.11; 6.4.1 to 6.4.9; and 9.2.1 to 9.2.20 from Larsen and Marx (2012).

4.1 Git + GitHub

We will do this in the live lecture in Week 5, but try yourself first.

1. Create a new repository on the University of Sydney GitHub enterprise page. It's best to give it the same name as your RStudio project, in the links below we refer to it as `repo_name`. You should probably make it private, otherwise anyone with a unikey can access it.
2. Link it with a RStudio project containing your Lab 2A solutions, and commit your changes. Details

here.

3. Go to the GitHub repo Settings page [https://github.sydney.edu.au/YOUR_UNIKEY/REPOSITORY_NAME/settings] and activate GitHub pages for the master branch of your repository. Wait a minute or two and you should be able to access your compiled HTML file at: `https://pages.github.sydney.edu.au/unikey/repo_name/lab2a.html`

4. Try making a change in your Rmd file, recompile your HTML locally, commit and push your changes to the GitHub repo. You should see your changes reflected at that webpage automatically after a minute or so.

4.2 Number of televisions per household

A random sample of 250 households in a large city revealed that the mean number of television per household was 2.76 with a sample standard deviation of 1.8. Can we conclude at the 5% significance level that the true number of televisions per household is at least 2.5? Assuming that the sample standard deviation is a good estimate of the population standard deviation, compute the power of this when the true mean is 3. What is the probability of a type II error?

The hypotheses are $H_0\colon\ \mu = 2.5$ vs $H_1\colon\ \mu > 2.5$. We have the true standard deviation $\sigma = 1.8$, $n = 250$ and $\bar{x} = 2.76$.

The test statistic is

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.76 - 2.5}{1.8/\sqrt{250}} = 2.284$$

The p-value is

$$P(t_{249} > 2.284) = 1 - P(t_{249} \leq 2.284) = 1 - 0.9884 = 0.0116$$

```
pt(2.284, df = 249)  # P(t_{249}\leq 2.284)
```

```
[1] 0.988393
```

```
pt(2.284, df = 249, lower.tail = FALSE)  # P(t_{249} > 2.284)
```

```
[1] 0.01160698
```

Noting that the critical value $P(t_{249} < c) = 0.95 \implies c = 1.65$.

```
qt(0.95, 249)
```

```
[1] 1.650996
```

The rejection region for sample mean $\bar{x}$ is:

$\frac{x-\mu_0}{s/\sqrt{n}} \geq c \Rightarrow \bar{x} \geq \mu_0 + c\frac{\sigma}{\sqrt{n}} = 2.5 + 1.65 \times \frac{1.8}{\sqrt{250}} = 2.688.$

The power under the alternative, $\mu = 3$, is

$$\text{Power}(3) = P(\text{reject } H_0 \mid H_0 \text{ is false and } \mu = 3) = P(\bar{X} > 2.688 \mid \bar{X} \sim N(3, 1.8^2/250))$$

```r
# because n is so large, we can get do an approximate power
# calculation using the normal distribution (instead of a non-central
# t distribution)
pnorm(2.688, mean = 3, sd = 1.8/sqrt(250), lower.tail = FALSE)
```

```
[1] 0.996934
```

```r
# using the pwr package
pwr::pwr.t.test(n = 250, d = (3 - 2.5)/1.8, sig.level = 0.05, type = "one.sample",
    alternative = "greater")
```

```

     One-sample t test power calculation 

              n = 250
              d = 0.2777778
      sig.level = 0.05
          power = 0.9968831
    alternative = greater
```

Hence the type II error is

$$\begin{aligned}\beta(3) &= P(\text{Do not reject } H_0 \mid H_0 \text{ is false and } \mu = 3)\\ &= 1 - \text{Power}(3)\\ &= 1 - 0.9969\\ &= 0.0031\end{aligned}$$

4.3 Drawing conclusions from rejection regions

Suppose that Candy wishes to test the hypotheses $H_0\colon\ \mu = 50$ against $H_1\colon\ \mu < 50$ for a normal population with an unknown standard deviation. A random sample of size 25 from this population gave her a mean of 48 and a standard deviation of 4. Compute the rejection region on the sample mean at the $\alpha = 0.05$ level of significance and draw your conclusion. Explain your decision to Candy.

The appropriate quantile is c such that $P(t_{24} < c) = 0.05 \implies c = -1.711$. The rejection region is

$$\bar{x} < \mu_0 + c\frac{s}{\sqrt{n}} = 50 - 1.711 \cdot \frac{4}{\sqrt{25}} = 48.6312$$

since the sample mean is 48 which is less than 48.6312, we reject H_0.

R code (comparable to what may be given in an exam context).

```r
qnorm(c(0.01, 0.025, 0.05, 0.1))
```

```
[1] -2.326348 -1.959964 -1.644854 -1.281552
```

```r
pnorm(c(-3.21, -2.89, -2.75, -2.28, -0.39, -0.064, -0.71))
```

```
[1] 0.0006636749 0.0019262091 0.0029797632 0.0113038442 0.3482682735
[6] 0.4744851134 0.2388520681
```

```r
qt(c(0.01, 0.025, 0.05, 0.1), 24)
```

```
[1] -2.492159 -2.063899 -1.710882 -1.317836
```

```r
qt(c(0.01, 0.025, 0.05, 0.1), 25)
```

```
[1] -2.485107 -2.059539 -1.708141 -1.316345
```

References

Champely, Stephane. 2020. *Pwr: Basic Functions for Power Analysis*. https://CRAN.R-project.org/package=pwr.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge academic.

Larsen, Richard J., and Morris L. Marx. 2012. *An Introduction to Mathematical Statistics and Its Applications*. 5th ed. Boston, MA: Prentice Hall.

Nordmann, Emily, and Phil McAleer. 2021. *Fundamentals of Quantitative Analysis*. https://psyteachr.github.io/quant-fun-v2/power-and-error.html.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.