Lab 04: Hypothesis Testing SOLUTIONS

PSTAT 100: Spring 2024 (Instructor: Ethan P. Marzban)

MEMBER 1 (NetID 1) MEMBER 2 (NetID 2)
MEMBER 3 (NetID 3)

May 6, 2024

Required Packages

```
library(ottr)        # for checking test cases (i.e. autograding)
library(pander)      # for nicer-looking formatting of dataframe outputs
library(tidyverse)   # for graphs, data wrangling, etc.
library(gridExtra)   # for multipanel graphs
```

Logistical Details

Logistical Details

- This lab is due by **11:59pm on Wednesday, May 8, 2024**.
- Collaboration is allowed, and encouraged!
  - If you work in groups, list ALL of your group members' names and NetIDs (not Perm Numbers) in the appropriate spaces in the YAML header above.
  - Please delete any "MEMBER X" lines in the YAML header that are not needed.
  - No more than 3 people in a group, please.
- Ensure your Lab properly renders to a `.pdf`; non-`.pdf` submissions will not be graded and will receive a score of 0.
- Ensure all test cases pass (test cases that have passed will display a message stating `"All tests passed!"`)

Lab Overview and Objectives

In this lab, we will discuss:

- One-sample tests for the mean
- Two-sample t-tests
- ANOVA (Analysis of Variance)

Recap: Hypothesis Testing

Recall that in **hypothesis testing**, we use data to assess claims made about a given population parameter. For example, given a population mean μ, our **null** hypothesis may take the form

$$H_0: \ \mu = \mu_0$$

for some specified value of $\mu_0 \in \mathbb{R}$, and our **alternative** hypothesis would take one of the following four forms:

- $H_A: \ \mu < \mu_0$ (**lower-tailed**)
- $H_A: \ \mu > \mu_0$ (**upper-tailed**)
- $H_A: \ \mu \neq \mu_0$ (**two-sided**)
- $H_A: \ \mu = \mu_A, \ \mu_A \neq \mu_0$ (**simple-vs-simple**)

We saw in lecture that a two-sided test for the mean when the population standard deviation σ is unknown takes the form

$$\text{Reject } H_0 \text{ when } \left| \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \right| > F^{-1}_{t_{n-1}}\left(1 - \frac{\alpha}{2}\right)$$

where $|(\overline{X}_n - \mu_0)/(S_n/\sqrt{n})|$ is called our **test statistic**, $F^{-1}_{t_{n-1}}(\cdot)$ denotes the quantile (i.e. inverse-cdf) function of the t_{n-1} distribution, and α is our **significance level** (set before the start of our analyses).

! Question 1

What value is at the 87.5^{th} percentile of the t_{42} (i.e. $t-$distribution with 48 degrees of freedom)? Assign your answer to a variable called `t_quant_1`. **Hint:** `qt()`.

Solution:

```
## replace this line with your code

t_quant_1 <- qt(0.875, 48)
```

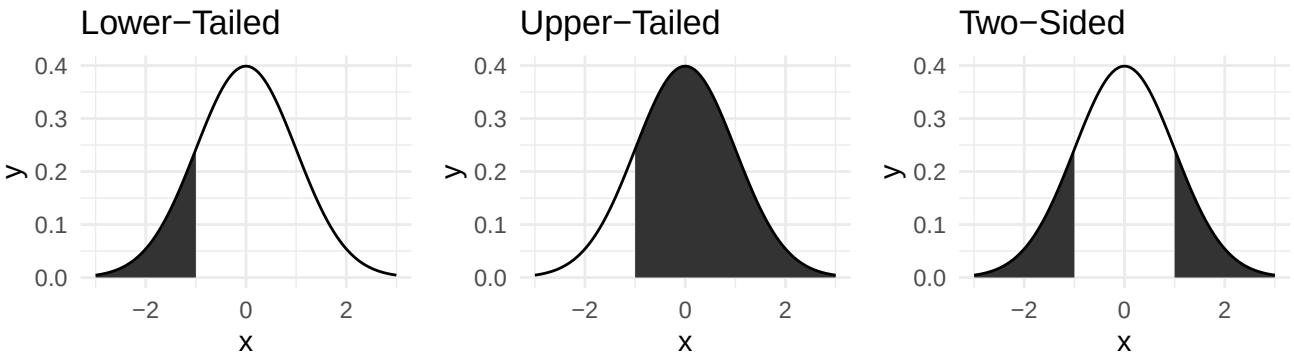
Answer Check:

```
# DO NOT EDIT THIS LINE
invisible({check("tests/q1.R")})
```

```
All tests passed!
```

Additionally, recall that hypothesis tests can be reformulated to be in terms of ***p*-values** (as opposed to **critical values**, as in the formulation above). We define a p-value to be the probability of, under

the null, observing something as or more extreme (in the direction of the alternative) as what was observed. Rather than trying to memorize formulas for p-values, I recommend drawing a picture. For instance, if we observe a test statistic value of -1, here are the diagrams corresponding to the lower-tailed, upper-tailed, and two-sided p-values:



! Question 2

Consider a sample of size $n = 87$ taken from a population with unknown mean μ and unknown standard deviation σ. The sample mean of these 87 values is 3.9 and the sample standard deviation is 1.46. Suppose we wish to test the hypotheses:

$$\begin{bmatrix} H_0: & \mu = 3.5 \\ H_A: & \mu < 3.5 \end{bmatrix}$$

Compute the p-value of the observed value of the test statistic; assign this value to a variable called `p_val_1`. **Hint:** `pt()`.

Solution:

```
## replace this line with your code

p_val_1 <- pt((3.9 - 3.5) / (1.46 / sqrt(87)), 86)
```

Answer Check:

```
# DO NOT EDIT THIS LINE
invisible({check("tests/q2.R")})
```

```
All tests passed!
```

Testing Across Two Groups

In certain cases, it may be desired to test whether or not two populations have the same mean. For example, we might ask ourselves: is the true average (mean) commute time of all Los Angelites the same as the true average (mean) commute time of New Yorkers?

More mathematically, consider two populations $\mathcal{P}_1$ (with mean μ_1 and standard deviation σ_1) and $\mathcal{P}_2$ (with mean μ_2 and standard deviation σ_2). The null hypothesis we wish to test can be formulated as

$$H_0 : \ \mu_1 = \mu_2$$

and some possible alternatives are:

- $H_A : \ \mu_1 < \mu_2$ (**lower-tailed**)
- $H_A : \ \mu_1 > \mu_2$ (**upper-tailed**)
- $H_A : \ \mu_1 \neq \mu_2$ (**two-sided**)

It's customary to reparametrize the null and alternative hypotheses to be in terms of parameter *differences*:

$$H_0 : \ \mu_1 - \mu_2 = 0$$

and

- $H_A : \ \mu_1 - \mu_2 < 0$ (**lower-tailed**)
- $H_A : \ \mu_1 - \mu_2 > 0$ (**upper-tailed**)
- $H_A : \ \mu_1 - \mu_2 < 0$ (**two-sided**)

The reason we do so is, if we view $\delta := \mu_1 - \mu_2$ as its own parameter, our test can be rephrased as a test solely on δ - that is, we can effectively treat the problem as a one-sample problem (which we are now very familiar with).

Now, consider a sample $X \sim \mathcal{P}_1$ of size n_1 and $Y \sim \mathcal{P}_2$ of size n_2 (note that we are allowing our two samples to be of different sizes!). An unbiased estimator for δ is $\Delta := \overline{X}_{n_1} - \overline{Y}_{n_2}$, and hence it makes sense to formulate a test statistic to be in terms of this difference. Note that, if we assume independence both within our samples and across our samples,

$$\begin{aligned} \text{Var}(\Delta) &:= \text{Var}\left(\overline{X}_{n_1} - \overline{Y}_{n_2}\right) \\ &= \text{Var}\left(\overline{X}_{n_1}\right) + \text{Var}\left(\overline{Y}_{n_2}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

Hence, a natural test statistic (assuming *unknown* population standard deviations) is

$$\text{TS} := \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_1}}} \tag{1}$$

where S_X^2 and S_Y^2 denote the sample variances of our samples X and Y, repsectively. It turns out that the exact distribution of TS is unknown, but very well-approximated by a $t-$distribution with degrees of freedom given by the **Satterthwaite Approximation**:

$$\text{df} = \text{round}\left\{ \frac{\left[\left(\frac{s_X^2}{n_1}\right) + \left(\frac{s_Y^2}{n_2}\right)\right]^2}{\frac{\left(\frac{s_X^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_Y^2}{n_2}\right)^2}{n_2 - 1}} \right\}$$

where s_X^2 and s_Y^2 denote the observed instances of S_X^2 and S_Y^2, respectively, and round$(\cdot)$ denotes the rounding function [e.g. round$(4.2) = 4$]

! Question 3

Suppose we have a sample X of size $n_1 = 50$ from Population 1 and a sample Y of size $n_2 = 60$ from Population 2. If $s_X^2 = 4.3$ and $s_Y^2 = 3.2$, compute the degrees of freedom of the t-distribution that the test statistic defined in Equation (1) follows. Store your answer in a variable called `df1`.

Solution:

```
## replace this line with your code

sx2 <- 4.3
sy2 <- 3.2
n1 <- 50
n2 <- 60

numerator <- ((sx2 / n1) + (sy2 / n2))^2
denom <- ((sx2 / n1)^2 / (n1 - 1)) + ((sy2 / n2)^2 / (n2 - 1))

df1 <- round(numerator / denom)
```

Answer Check:

```
# DO NOT EDIT THIS LINE
invisible({check("tests/q3.R")})
```

```
All tests passed!
```

! Question 4

Refer to the situation in question 3 above, and assume that the mean of X is 5.5 and the mean of Y is 5.2. Compute a two-sided p-value (i.e. a p-value assuming a two-sided alternative) of the observed test statistic. Store your answer in a variable called `pval2`.

Solution:

```
## replace this line with your code

test_stat <- (5.5 - 5.2) / sqrt((sx2 / n1) + (sy2 / n2))

pval2 <- 2*pt(abs(test_stat), df = df1, lower.tail = F)
```

Answer Check:

```
# DO NOT EDIT THIS LINE
invisible({check("tests/q4.R")})
```

```
All tests passed!
```

Now, there exists a function in `R` called `t.test()` which is most often used to conduct **two-sample *t*-tests** (which is precisely the test we just developed above).

! Question 5

Consider the following vectors `x` and `y`, and interpret them as two samples from two different populations

```
x <- c(1, 2, 3, 4, 5)
y <- c(2, 2, 3, 3, 4, 5)
```

First conduct a two-sided *t*-test **by hand** (i.e. using only basic `R` functions and **NOT** using `t.test()`). Then, reconduct your test using `t.test()`, and compare results.

Solution:

```
## replace this line with your code
xbar <- mean(x)
ybar <- mean(y)

sx2 <- var(x)
sy2 <- var(y)
n1 <- length(x)
n2 <- length(y)

ts <- (xbar - ybar) / sqrt((sx2 / n1) + (sy2 / n2))

numerator <- ((sx2 / n1) + (sy2 / n2))^2
denom <- ((sx2 / n1)^2 / (n1 - 1)) + ((sy2 / n2)^2 / (n2 - 1))
```

The observed value of the test statistic is -0.1953662, and the degrees of freedom are 7.2679146. Compare this with the output of `t.test()`:

```
t.test(x, y, alternative = "two.sided")
```

```
    Welch Two Sample t-test

data:  x and y
t = -0.19537, df = 7.2679, p-value = 0.8505
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 -2.168948  1.835615
sample estimates:
mean of x mean of y
 3.000000  3.166667
```

Answer Check:

There is no autograder for this question.

ANOVA

A natural question that arises is: how can we compare means across *several* (i.e. more than 3 groups). For example, suppose we want to determine whether the average (mean) air pollution levels are the same across three different cities.

More concretely, consider k populations $\mathcal{P}_1, \cdots, \mathcal{P}_3$ with means $\mu_1, \cdots, \mu_k$ and variances $\sigma_1^2, \cdots, \sigma_k^2$. Additionally, consider testing

$$\begin{bmatrix} H_0: & \mu_1 = \mu_2 = \cdots = \mu_k \\ H_A: & \text{At least one of the means are different} \end{bmatrix}$$

Such a set of hypotheses can be tested using what is known as an **Analysis of Variance** (or **ANOVA**, for short). Here's the main idea of how ANOVA works. Suppose we have samples (potentially of different sizes) from each population. Even if all populations have the same means, we wouldn't be surprised in our samples had slightly different observed sample means. This is because there will be some baseline variability due to chance. What ANOVA seeks to do is compare the variances within and across samples and see whether or not the overall variability exceeds what we would expect due to chance alone (which would lead credence *away* from the null, that the populations all have the same mean).

In the interest of time, I'll bypass the theoretical derivations of ANOVA and jump straight to how we can perform an ANOVA in `R`. There are actually a couple of functions which can be used for conducting an ANOVA - we'll use the function `aov()` [and we'll return to ANOVA in a couple of weeks].

One thing I want to make very clear is the fact that the alternative hypothesis in ANOVA is **NOT** $H_A: \mu_1 \neq \mu_2 \neq \cdots \neq \mu_k$. Rather, the alternative is simply that *at least one* of the group means differs from the rest.

As a concrete example, consider the following (fictional) situation. Suppose we want to determine whether or not the average scores of students on a particular exam differ significantly based on class

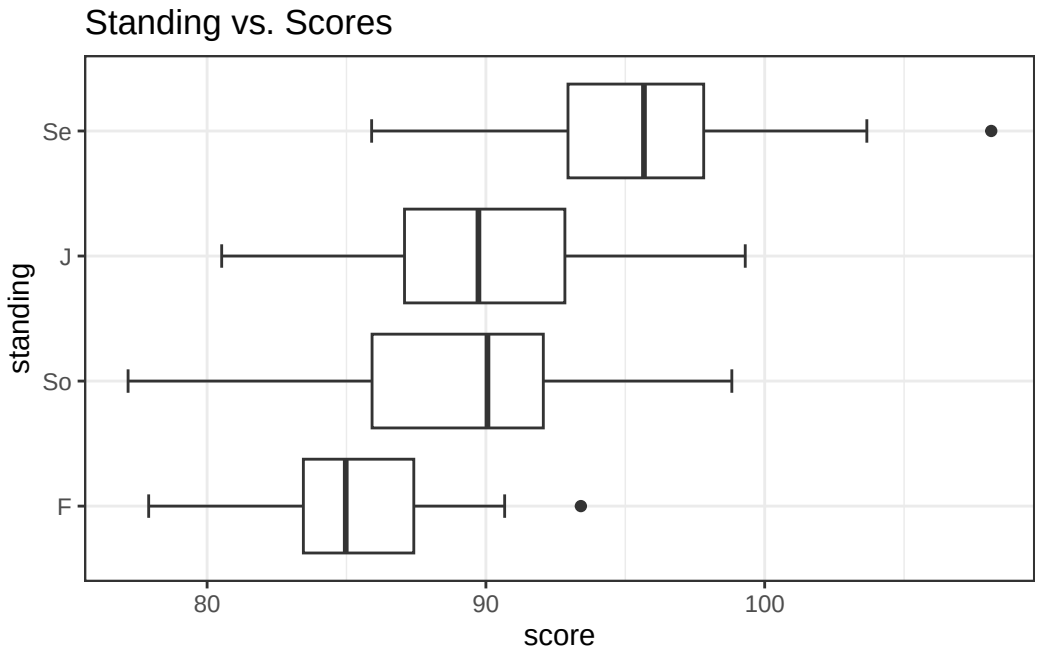
standing (i.e. freshmen, sophomore, junior, senior). Additionally, suppose we collect the following (fictional) data:

```
freshmen <- rnorm(50, 85, 3)
sophomores <- rnorm(60, 90, 5)
juniors <- rnorm(70, 90, 4)
seniors <- rnorm(40, 95, 5)

scores <- data.frame(
  score = c(freshmen, sophomores, juniors, seniors),
  standing = factor(
    c(rep("F", length(freshmen)),
                rep("So", length(sophomores)),
                rep("J", length(juniors)),
                rep("Se", length(seniors))
                ),
    ordered = T,
    levels = c("F", "So", "J", "Se")
  )
)
```

As a first pass, we can generate a side-by-side boxplot:

```
scores %>%
  ggplot(aes(y = standing,
             x = score)) +
  geom_boxplot(staplewidth = 0.25) +
  theme_bw() +
  ggtitle("Standing vs. Scores")
```



Based on the boxplot alone, it looks like there are some clear differences in the scores across the different class standings. To formally test this using an ANOVA, we use:

```
aov(score ~ standing, data = scores) %>% summary()
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
standing      3   2402   800.8   46.71 <2e-16 ***
Residuals   216   3703    17.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We'll discuss the different components of the output in a few weeks (after we discuss regression). For now, focus on the `Pr(>F)` column - this is (essentially) a p-value of the hypotheses posited at the start of this section.

! Question 6

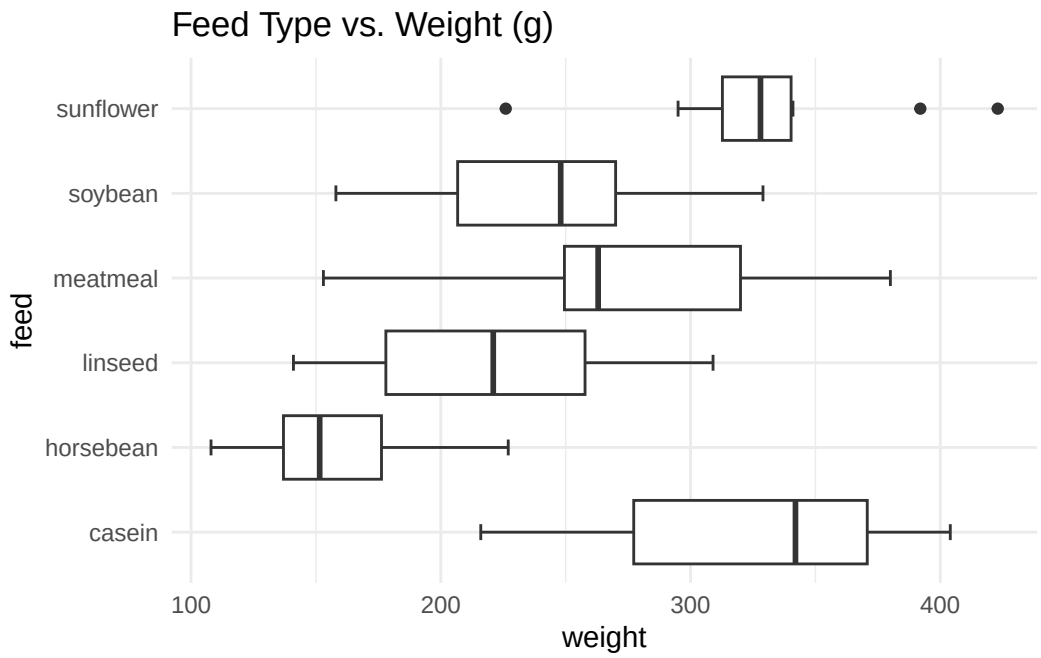
One of the datasets built into `R` is called `chickwts`, and contains the weights of various chickens placed on one of six different feed supplements.

a) Generate a side-by-side boxplot of the weight (in grams) vs. feed type. Based on the graph, does there appear to be a difference in average (mean) weight across the different feed types?

b) Conduct an ANOVA to test whether there is a statistically significant difference in average chick weights across the different feed types.

Solution:

```
## replace this line with your code;
## feel free to add more code chunks as you see fit.
chickwts %>%
  ggplot(aes(x = weight, y = feed)) +
  geom_boxplot(staplewidth = 0.25) +
  theme_minimal() +
  ggtitle("Feed Type vs. Weight (g)")
```



Based on this boxplot, it appears as though chick weights do differ across the different feed types. To formally test this with an ANOVA, we use

```
aov(weight ~ feed, chickwts) %>% summary()
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
feed         5 231129   46226   15.37 5.94e-10 ***
Residuals   65 195556    3009
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sure enough, the p-value is `5.94e-10` (i.e. 5.94×10^{-10}), which is much smaller than a 0.05 level of significance meaning:

> At a 5% level of significance, there is evidence to suggest that there exists a difference in average (mean) chick weight across the different feed types.

Answer Check:

There is no autograder for this question.

One key point that we will return to later is that we need to assume our samples are drawn from a normal distribution - otherwise, the results of an ANOVA may be skewed. Again, we'll discuss how to check this assumption (as well as how to bypass it) later in the course.

Submission Details

1) Check that all of your tables, plots, and code outputs are rendering correctly in your final `.pdf`.

2) Check that you passed all of the test cases (on questions that have autograders). You'll know that you passed all tests for a particular problem when you get the message "All tests passed!".

3) Submit **ONLY** your `.pdf` to Gradescope. Make sure to match pages to your questions - we'll be lenient on the first few labs, but after a while failure to match pages will result in point penalties.