

DSCI 552 Lab 2 Solutions

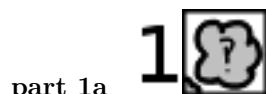
General instructions



- This assignment is to be completed in R, submitting both a .Rmd markdown file you create in RStudio (you can add your answers directly to this one) along with a rendered .md file.
- By submit, we mean upload these file to your repository for this assignment on Github.
- In the main README.md of you lab/assignment repository, create a table that lists and links to all the important files for this assignment.
- When answering the questions, in addition to R code, you must write a couple sentences to explain what you are doing, and what result you obtained.
- Make it easy for others to run your code
 - In exactly one, very early R chunk, load any necessary packages, so your dependencies are obvious.
 - In exactly one, very early R chunk, import anything coming from an external file. This will make it easy for someone to see which data files are required, edit to reflect their locals paths if necessary, etc. There are situations where you might not keep data in the repo itself.
 - Pretend you are someone else. Clone a fresh copy of your own repo from GitHub, fire up a new RStudio session and try to knit your R markdown file. Does it “just work”? It should!

```
# Load external files
titanic <- read.csv("titanic.csv")
```

Exercise 1 - Basic components of a hypothesis test



part 1a

Given the research question below and the sample dataset, write testable null and alternative hypotheses:

Research question: Does strength training (*i.e.*, lifting weights) during training for a half marathon increase speed on race day?

Sample fake data:


note: assume $N = 50$ for each strength training group

runner	strength training (yes/no)	race day speed (km/h)
Lindsay Lohan	yes	11.94
Kevin Bacon	no	9.43
Justin Bieber	no	9.33
Simon Cowell	no	9.64
Kim Kardashian West	yes	9.65
Pauly Shore	yes	9.67
...

The null hypothesis is: the average race day speed for people have strength training is the same


as those who don't have strength training.

The alternative hypothesis is: The average race day speed for people have strength training is higher than those who don't have strength training.

part 1b 


Identify a reasonable test statistic to use for your hypothesis test and justify why you chose it.

One possible choice is the difference in the sample average race day speed between people have strength training and people don't have strength training. It is chosen because it estimates the difference in the population average race day speed between the two groups, i.e., assessing the hypothesis of interest directly. To construct a meaningful test statistic, one might have to adjust this difference in sample averages by its estimated standard error.

part 1c 

Describe the expected distribution of the test statistic if the null hypothesis is true.



With sample size large enough, both the sample averages from the two groups follow normal distribution so their difference also follows normal distribution. Under null, the mean of the normal distribution of difference in sample averages is 0 but the variance is unknown. The adjusted difference, which is the test statistic suggested in part 1c, follows t distribution with d.f. depending on the sample size.

part 1d 

Describe how you would convert the test statistic to a p-value.

The p-value is the probability that a t distributed random variable is larger than the observed test statistic.

Exercise 2a - Hypothesis tests on continuous response variables (t-test)

1  2 

You are interested in whether or not there is a difference between survivorship and mean fare paid for men in passenger classes 1 & 2 (i.e., pool the data from `pclass 1 & 2`, or `pclass != 3`) from the [Titanic dataset](#). To answer this question do the following:

- define the response and explanatory variables
- state the null and alternative hypotheses
- make a visualization of the data using `ggplot2::geom_boxplot()`
- test your null hypothesis using a t-test (use the `t.test()` function in R)

- explain and interpret your results in 2-3 sentences (report both the p-value from the test, as well as the confidence intervals)

Solutions

```
##1.
names(titanic)

## [1] "pclass" "survived" "name" "sex" "age"
## [6] "sibsp" "parch" "ticket" "fare" "cabin"
## [11] "embarked" "boat" "body" "home.dest"

kable(some(titanic)) # see list of a few passengers
```

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin
245	1	0	Ross, Mr. John Hugo	male	36.0	0	0	13049	40.1250	A/50...
403	2	1	Duran y More, Miss. Florentina	female	30.0	1	0	SC/PARIS 2148	13.8583	...
456	2	0	Howard, Mr. Benjamin	male	63.0	1	0	24065	26.0000	...
518	2	0	Nesson, Mr. Israel	male	26.0	0	0	244368	13.0000	E...
582	2	0	Watson, Mr. Ennis Hastings	male	NA	0	0	239856	0.0000	...
644	3	1	Asplund, Miss. Lillian Gertrud	female	5.0	4	2	347077	31.3875	...
768	3	0	Demetri, Mr. Marinko	male	NA	0	0	349238	7.8958	...
1130	3	0	Pettersen, Mr. Johan Emil	male	25.0	1	0	347076	7.7750	...
1172	3	0	Sage, Master. William Henry	male	14.5	8	2	CA. 2343	69.5500	...
1261	3	1	Turja, Miss. Anna Sofia	female	18.0	0	0	4138	9.8417	...

```
#The response variable is the fare and the explanatory variable is the survival state

fare_surv_class12 <- titanic %>% filter(pclass != 3 & sex == "male") %>%
  select(fare, survived, pclass)
attach(fare_surv_class12)

some(fare_surv_class12) # see list of a few male passengers from 1st and 2nd classes
```

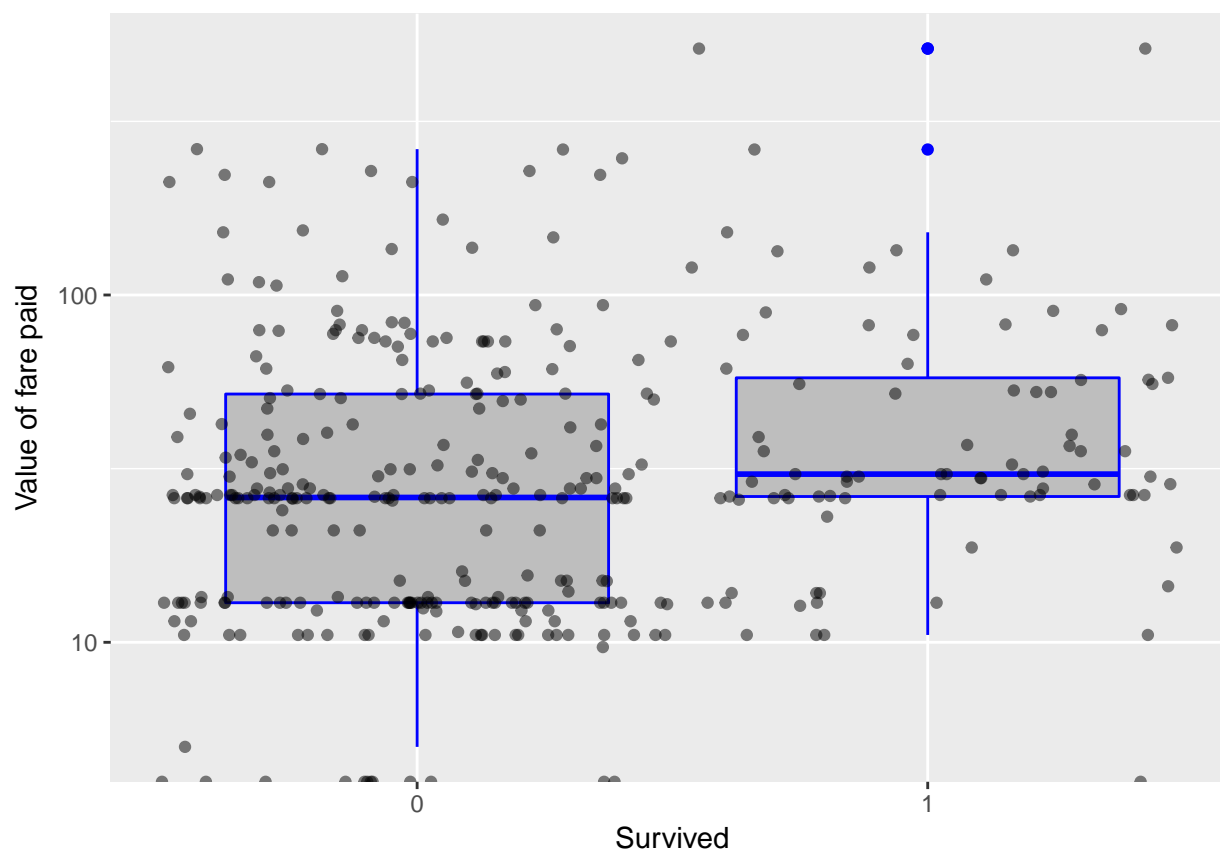
```
##      fare survived pclass
## 24  31.0000      0      1
## 29  47.1000      0      1
## 52  56.9292      1      1
## 91 211.5000      0      1
## 122  66.6000      0      1
## 238  10.5000      0      2
## 248  13.0000      0      2
## 253  65.0000      0      2
## 254  73.5000      0      2
## 267  31.5000      0      2
```

```
## 2
```

Null hypothesis: There is no difference in mean fare price between men that survived, and men who perished.

Alternative hypothesis: The mean fare prices between men who survived, and men who perished are different.

```
##3.
ggplot(fare_surv_class12, aes(factor(survived), fare)) +
  geom_boxplot(fill="grey", colour="blue") +
  geom_jitter(width = 0.5, alpha = 0.5) + scale_y_log10() +
  xlab("Survived") + ylab("Value of fare paid")
```



```
##4.

#t= (\bar{x}_1 - \bar{x}_2) / \sqrt{(s^2/n_1 + s^2/n_2)}
#df = n1 + n2 - 2
t.test(fare_surv_class12$fare ~ fare_surv_class12$survived,
  paired = FALSE,
  var.equal = TRUE,
  alternative="two.sided")
```

```
##
## Two Sample t-test
##
## data: fare_surv_class12$fare by fare_surv_class12$survived
## t = -2.161, df = 348, p-value = 0.03138
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -30.296847 -1.425311
```

```
## sample estimates:
## mean in group 0 mean in group 1
##      41.57061      57.43169

#or

#t= (\bar{x}_1 - \bar{x}_2)/\sqrt{(s_1^2/n_1 + s_2^2/n_2)}
#df = (s_1^2/n_1 + s_2^2/n_2)/(s_1^2/n_1/(n_1-1) + s_2^2/n_2/(n_2-1))
t.test(fare_surv_class12$fare ~ fare_surv_class12$survived,
       paired = FALSE,
       var.equal = FALSE,
       alternative="two.sided")

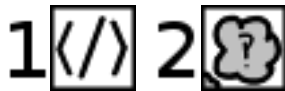
##
## Welch Two Sample t-test
##
## data: fare_surv_class12$fare by fare_surv_class12$survived
## t = -1.7051, df = 106.46, p-value = 0.09109
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -34.302228  2.580071
## sample estimates:
## mean in group 0 mean in group 1
##      41.57061      57.43169

##5.
```

Assuming equal variance, the p value is 0.03138 which is less than the alpha level of 0.05. The 95% confidence interval of the difference in means doesn't contain zero: (-30.29, -1.43). We thus reject the null hypothesis at the 0.05 significance level. The difference in means is negative within the 95% confidence interval, which means the mean fare price of the perished men is generally lower than the mean fare price of the men who survived.

Assuming unequal variance, the p value is 0.09 and there is less evidence to reject the null hypothesis. The 95% confidence interval of the difference in means contains zero: (-34.30, 2.58). We thus fail to reject the null hypothesis at the 0.05 significance level.

Exercise 2b - Hypothesis tests on continuous response variables (permutation)



Perform a permutation test to answer the question from Exercise 2a. For an example of how to do a permutation test see [this tutorial](#). In completing this exercise, ensure that you:

- carry out a permutation test from scratch to test your null hypothesis, this should include:
- generate a histogram of the test statistic (difference in means) from the permutation of the explanatory variable
- label your observed difference in means on the histogram with a vertical line
- calculate a p-value and interpret it in regards to whether or not you should reject the null hypothesis
 - note - calculate the p-value by counting how many permuted mean-differences are larger than the one we observed in our actual data, and then dividing this by the number of items in our permutation distribution

- explain and interpret your results in 2-3 sentences
- repeat the permutation analysis but use the `coin` package in R. Do you get the same results? Discuss.

Solutions

```
## 1
set.seed(1234)

some(fare_surv_class12) # see list of a few male passengers from 1st and 2nd classes

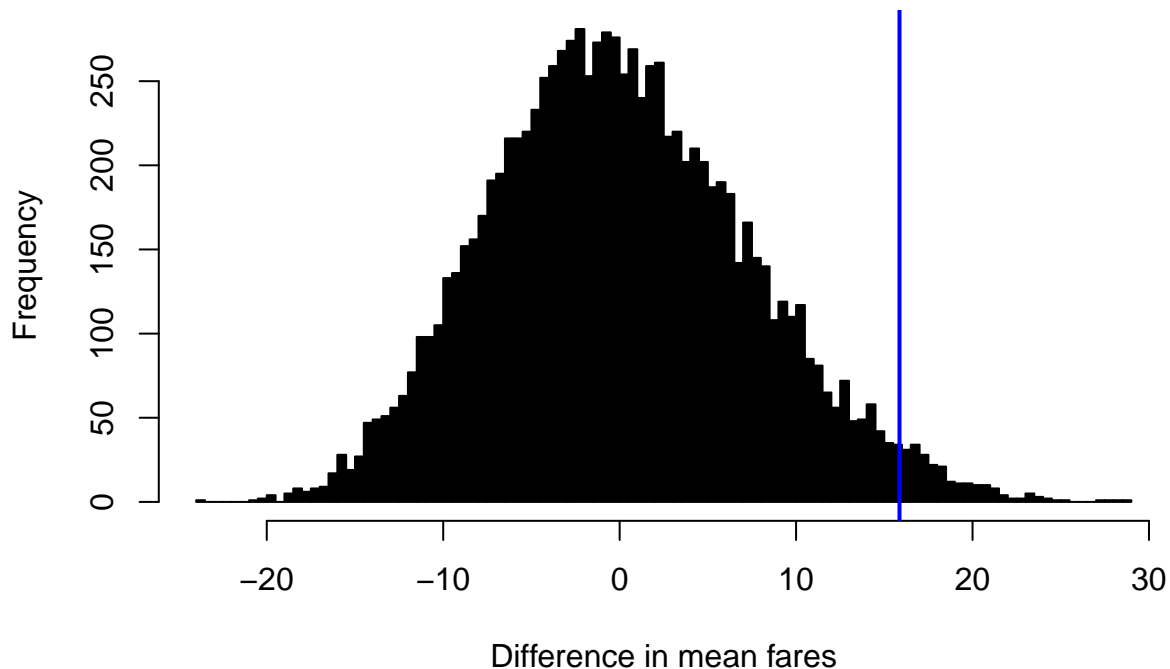
##      fare survived pclass
## 4      0.0000      0      1
## 40     25.5875      0      1
## 80     93.5000      0      1
## 176    61.3792      1      1
## 213     26.2500      0      2
## 217     73.5000      0      2
## 218     13.0000      0      2
## 221     31.5000      0      2
## 228     13.0000      0      2
## 298     10.5000      0      2

nrow(fare_surv_class12) #number of passengers

## [1] 350

dist <- replicate(10000, diff(by(fare, sample(survived, size=length(survived), replace=FALSE), mean)))

## 2,3
hist(dist, col = "black", breaks = 100, main = "", xlab = "Difference in mean fares")
abline(v = diff(by(fare, survived, mean)), col = "blue", lwd = 2)
```



```
## 4
#The p-value obtained via permutation by hand is:

sum(abs(dist) >= abs(diff(by(fare, survived, mean))))/10000 # two-tailed test

## [1] 0.0297

# repeat the permutation but use `coin` package
#"The p-value obtained via permutation using the coin package is:
independence_test(fare ~ survived)

##
## Asymptotic General Independence Test
##
## data: fare by survived
## Z = 2.1497, p-value = 0.03158
## alternative hypothesis: two.sided
```

Manually or using the `coin` package, the difference in mean fare paid by men who survived and men who perished is statistically significant at the 0.05 significance level.

Exercise 2c - Hypothesis tests on continuous response variables (reflection)



In ~ half a page, discuss: - the similarities & differences of the t-test and permutation test approaches, and the results obtained in Exercises 2a and 2b. - when you might choose one of these approaches over the other.

Solutions

Similarities:

- requires a test statistic
- returns a p-value

Differences:

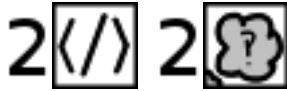
- permutation takes numerous resamples (without replacement) from the observed data to generate a distribution to be used in the test, while the t-test assumes that the test statistic is normally distributed.
- the t-test accounts for degrees of freedom and allows us to generate a confidence interval, while the permutation test does not generate a confidence interval.

Which approach?

- when the number of observations is large (roughly > 20) I would choose the t-test over permutation because it also calculates the confidence interval of the test statistic.
- when my number of observations is small (roughly < 20) I would either bootstrap or permute instead of using the t-test because these methods do not make the assumption that the test statistic is normally distributed (which they might not be with such a small number of observations).

Suggested reading: These are some material I found online that would help to understand when to choose which test. [1](#), [2](#), [3](#).

Exercise 3 - Permutation tests for categorical variables



Use permutation test to determine if in the third passenger class (`pclass == 3 & age < 18`) significantly less male children survived the sinking of the titanic compared to female children? For an example of how to do a permutation test see [this tutorial](#).

- define the response and explanatory variables
- state the null and alternative hypotheses
- make a visualization of the data using `ggplot2::geom_bar` (plot proportions, not counts - *hint: try using `geom_bar(position = "fill")`*)
- make a 2x2 contingency table for the counts, as well as one for the proportions
- carryout a permutation test from scratch to test your null hypothesis, this should include:
- generate a histogram of the test statistic (difference in probabilities) from the permutation of the explanatory variable
- label your observed difference in probabilities on the histogram with a vertical line
- calculate a p-value and interpret it in regards to whether or not you should reject the null hypothesis
 - note - calculate the p-value by counting how many permuted proportion-differences are larger than the one we observed in our actual data, and then dividing this by the number of items in our permutation distribution
- explain and interpret your results in 2-3 sentences
- repeat the permutation analysis but use the `coin` package in R. Do you get the same results? Discuss.

Solutions

1

The response variables is survivorship and the explanatory variables is gender

```
Q3dat <-titanic %>% filter(pclass==3 & age<18) %>% select(pclass,survived,sex,age)
head(Q3dat)
```

```
##   pclass survived    sex    age
## 1      3         0  male 13.0000
## 2      3         0  male 16.0000
## 3      3         1 female 16.0000
## 4      3         1  male  0.8333
## 5      3         0  male  4.0000
## 6      3         0 female  6.0000
```

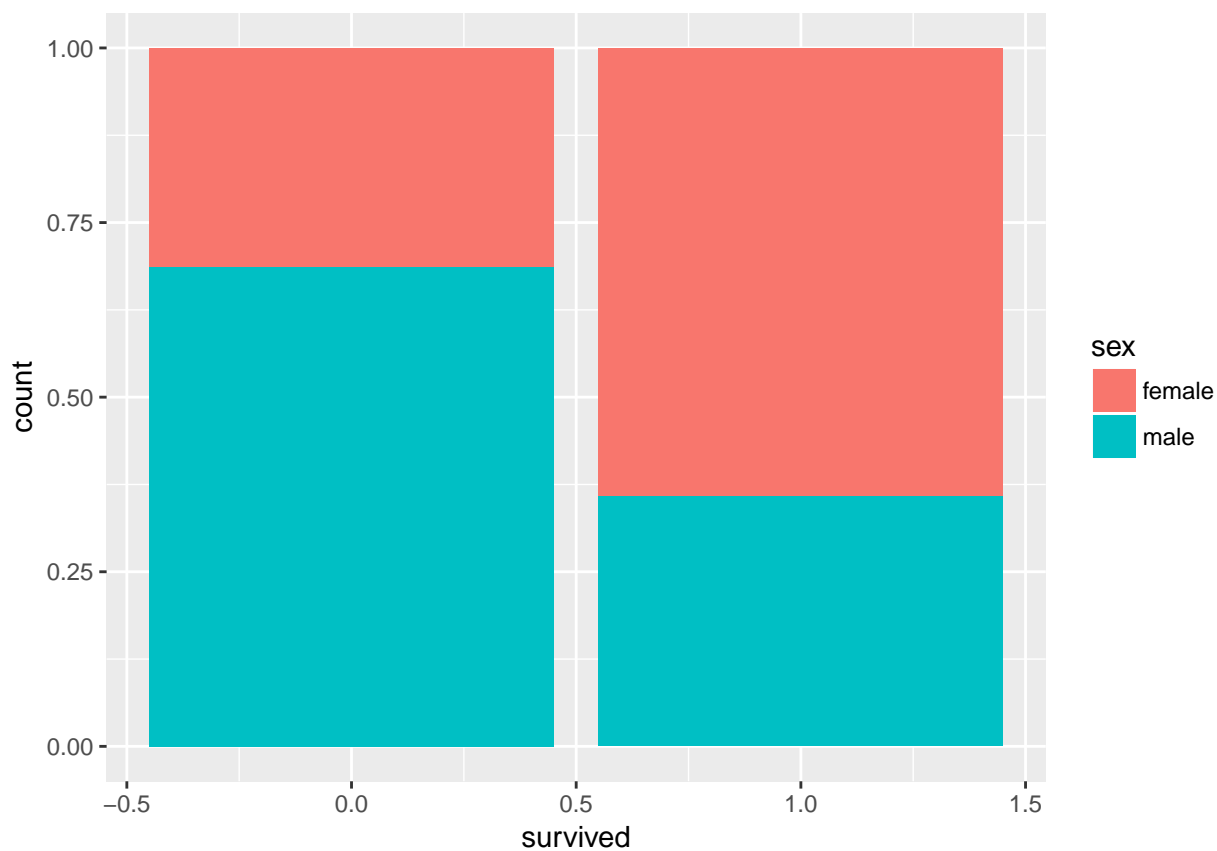
2

Null hypothesis: From the sinking of the Titanic, the proportion of survived female children is the same as the proportion of survived male children in the third passenger class.

Alternative hypothesis: There is a difference in the proportion of male versus female children in the third passenger class who survived the sinking of the Titanic.

3 Visualization


```
ggplot(Q3dat, aes(survived, fill=sex)) + geom_bar(position = "fill")
```



4 Contingency table of proportions

```
table(Q3dat$sex,Q3dat$survived) / nrow(Q3dat)
```

```
##
##           0           1
##  female 0.1981132 0.2358491
##  male   0.4339623 0.1320755
```

5 Permutation test

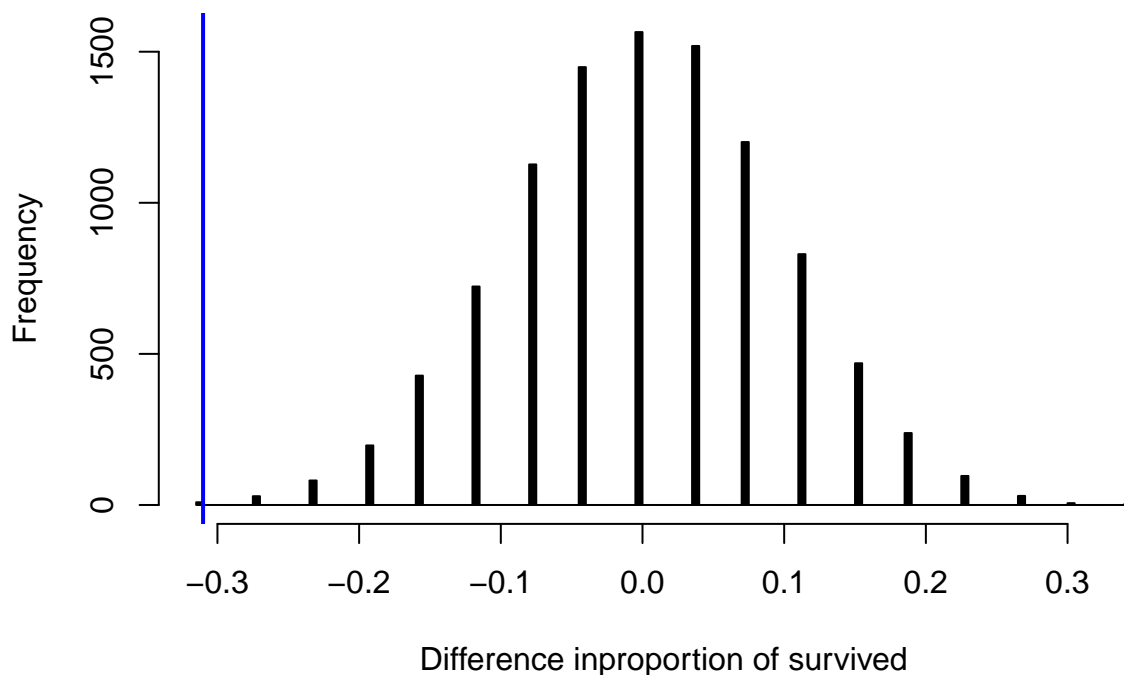
Observed difference in proportion of survived between female and male children

```
obs_diff <- diff(by(Q3dat$survived,Q3dat$sex,mean))
obs_diff
```

```
## [1] -0.3101449
```

Histogram of difference in proportion of survived from permutation

```
## generate permutation diff
set.seed(1234)
dist <- replicate(10000, diff(by(Q3dat$survived, sample(Q3dat$sex, size=length(Q3dat$sex), replace=FALSE),
## construct histogram
hist(dist, col = "black", breaks = 100, main = "", xlab = "Difference inproportion of survived")
abline(v = obs_diff, col = "blue", lwd = 2)
```



p-value computation

```
sum(abs(dist)>abs(obs_diff))/length(dist)
```

```
## [1] 3e-04
```

The p-value is 0.0003, and the null hypothesis is safely rejected for any reasonable nominal level, say 0.05.

4 Interpretation

It is highly unlikely that the population proportions of survived are the same for male and female children.

5 Coin package

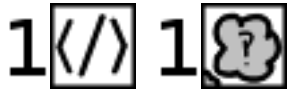
```
independence_test(Q3dat$survived~Q3dat$sex)
```

```
##
## Asymptotic General Independence Test
```

```
##
## data: Q3dat$survived by Q3dat$sex (female, male)
## Z = 3.2662, p-value = 0.00109
## alternative hypothesis: two.sided
```

The p-value reported by the coin package and the permutation test conducted is slightly different although both indicate strong tendency to reject the null hypothesis. The difference in p-value is potentially due to the difference in technical implementation or methodology used in coin.

Exercise 4 - Fisher's Exact test



- Perform an exact test (specifically, Fisher's exact test by using the `fisher.test()` function in R) to answer the question from Exercise 3.
- Explain and interpret your results in 2-3 sentences (report both the p-value from the test, as well as the confidence intervals).

Solutions

```
## 1

children <- titanic %>% filter(pclass == 3 & age < 18) %>% select(survived, sex, age)
some(children) # list of a few passengers in 3rd class who are minors
```

```
##      survived    sex age
## 16           0 female 17
## 38           0  male 15
## 45           0  male  1
## 57           1 female  4
## 61           1  male  6
## 68           0  male  2
## 84           0  male  8
## 88           1 female  1
## 90           0  male  4
## 105          1 female 15
```

```
count_table <- table(children$sex, children$survived)
colnames(count_table) <- c("Perished", "Survived")
count_table
```

```
##
##      Perished Survived
##  female      21      25
##  male       46      14
```

```
fisher.test(count_table)
```

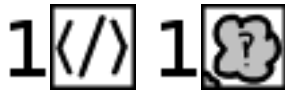
```
##
## Fisher's Exact Test for Count Data
##
## data:  count_table
## p-value = 0.001251
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1017841 0.6350648
## sample estimates:
## odds ratio
##  0.2592803

## 2
```

The null hypothesis is that the odds of death of either female or male child is the same. In other words, we expect gender to be independent of death rate.

The p value reported by Fisher's Exact Test is 0.001251, which is less than 0.05. So it is safe to reject the null hypothesis that the odds ratio is 1 at 0.05 significance level. The alternate hypothesis suggests that the odds of death of a female child is about 75% less than that of a male child.

Exercise 5 - Chi-square test



- Perform an Pearson's Chi-square test using the `chisq.test(..., correct = FALSE)` function in R to answer the question from Exercise 3.
- Explain and interpret your results in 2-3 sentences (you need only report the p-value for this exercise).

Solution

1 Pearson's Chi-square test

```
## chisq.test requires counting data as input
table(Q3dat$sex,Q3dat$survived)

##
##           0  1
## female 21 25
## male   46 14

chisq.test(table(Q3dat$sex,Q3dat$survived),correct = FALSE)

##
## Pearson's Chi-squared test
##
## data:  table(Q3dat$sex, Q3dat$survived)
## X-squared = 10.77, df = 1, p-value = 0.001032
```

2 Interpretation

The small p-value leads to a rejection of the null at 5% significant level. This implies that the population proportions of survived is likely to be different for female and male children.

Exercise 6 - Reflection



In ~ half a page, discuss: - the similarities and differences in approach and the results obtained between the 3 different tests you used to assess the same null hypothesis in Exercises 3, 4 & 5. - when you might choose one of these approaches over the other.

Solutions

Permutation test

Take several samples from the total possible permutations of the two categorical variables, and see what proportion of those samples gives a result that is higher than the observed value. This proportion is the p-value and the alpha level can be set at 5%. The p-value is to be interpreted as the probability of observing our result or an even higher value if the null hypothesis is true.

Fisher's exact test

Enumerate all permutations of the outcomes, then see what proportion of them is higher than the one observed. This proportion is the p-value, which is to be interpreted as the probability of observing our result or a higher value if the null hypothesis is true.

Chi-square test

Calculate the expected frequency for each categorical variable under the null hypothesis. For each cell of the contingency table, find the squared difference between the expectation and the outcome, and divide it by the expectation. Add up all these values to find your test statistic, and see where it falls on the Chi-square distribution with the appropriate number of degrees of freedom. The p-value is the probability of seeing a result at least as high as our observation under the null hypothesis.

Which approach?

- use the chi-squared test when N is large and the proportions are large
- use the fisher test when N is small and/or proportions are small
- use permutation if N is either small or large and the response variable is a real number

Suggested reading: [When to choose which statistic test](#), [Exact tests](#)

Exercise 7 - type I & II errors

Write ~ 2-4 sentences for each question listed below:



Given the following null and alternative hypotheses, write what a corresponding Type I error, Type II error, and what Power would be.


Ho: There is no difference in the median height between males & females

Ha: There is a difference in the median height between males & females

Type I error: Ho is rejected when it is true, i.e., the test result indicates that there is strong evidence that the median heights between males and females are not the same when their median heights are the same


Type II error: H_0 is not rejected when the H_0 is false (i.e., H_a is true), i.e., the test result indicates that there is no strong evidence that the median heights between males and females are not the same when their median heights are not the same

Power: Power is equal to $1 - \beta$ (probability of a Type II error). The power in this case is the probability for the test to correctly reject the null hypothesis

7(b) 


You choose an alpha level of .01 and then analyze your data. (a) What is the probability that you will make a Type I error given that the null hypothesis is true?

1%

7(c) 


Are more likely to make a Type I error when using a small sample than when using a large sample? Explain why or why not.

Usually type I error is controlled at the significance level roughly over any sample size. In case the null distribution of the test statistic is approximated by its asymptotic distribution, the type I error could be better controlled when sample size is larger.

7(d) 


Is a researcher risks making a Type I error any time the null hypothesis is rejected? Explain why or why not.

Yes, a researcher always risks making Type I error when null is rejected. The statement is false only if the test can reject H_0 correctly without uncertainty, i.e., the truth is known.

7(e) 

Give an example a Type II error, and of when and why one might occur.

Type II error occurs when H_0 is falsely rejected. For example, a patient gets negative result in a HIV test, but he/she indeed is infected with HIV.

7(f) 

Define effect size and explain its relationship with power.

The effect size is the minimum deviation from the null hypothesis that you hope to detect. The larger the effect size, the larger the probability to detect the difference, i.e., the larger power.

Exercise 8 - predicting sample size (optional)



Resource on power analysis in R: <http://www.statmethods.net/stats/power.html>

Using the built in `chickwts` dataset, we are interested in determining the sample size needed to observe a significant difference between mean `weight` from the `meatmeal` and `casein` diets. Remember from lecture 5 that when we did a t-test to compare the mean weight between these two diets, and we found that the diets were not statistically different ($p > 0.05$) but the sample size was small...

To accomplish this, perform a power analysis using: - `pwr::pwr.t.test` function in R (assume you will be collecting equal sample sizes for each group in your next experiment) with the arguments: - `type = "two.sample"` - `sig.level = 0.05` - For effect size (`d`), use the observed effect size from the data - the effect size, `d` is defined as difference in means of weight for the two diet groups divided by the pooled standard deviation (square root of the average of the two standard deviations). - use `ggplot2` to plot power versus sample size for power = 0.2, 0.4, 0.6, 0.8 & 1.0 - explain and interpret your results in 2-3 sentences

Solutions

```
some(chickwts) # see table
```

```
##      weight      feed
## 3      136 horsebean
## 5      217 horsebean
## 9      143 horsebean
## 10     140 horsebean
## 13     181  linseed
## 23     243  soybean
## 37     423 sunflower
## 42     226 sunflower
## 68     216  casein
## 70     283  casein
```

```
weight_meat <- chickwts %>% filter(feed == "meatmeal")
weight_casein <- chickwts %>% filter(feed == "casein")
```

```
# pooled variance
```

```
pooled_var <- ((length(weight_casein)-1)*var(weight_casein$weight) +
               (length(weight_meat)-1)*var(weight_meat$weight)) /
               (length(weight_meat) + length(weight_casein) - 2)
```

```
pooled_var
```

```
## [1] 4181.905
```

```
# effect size
```

```
effect_size <- abs(mean(weight_casein$weight) - mean(weight_meat$weight)) / sqrt(pooled_var)
```

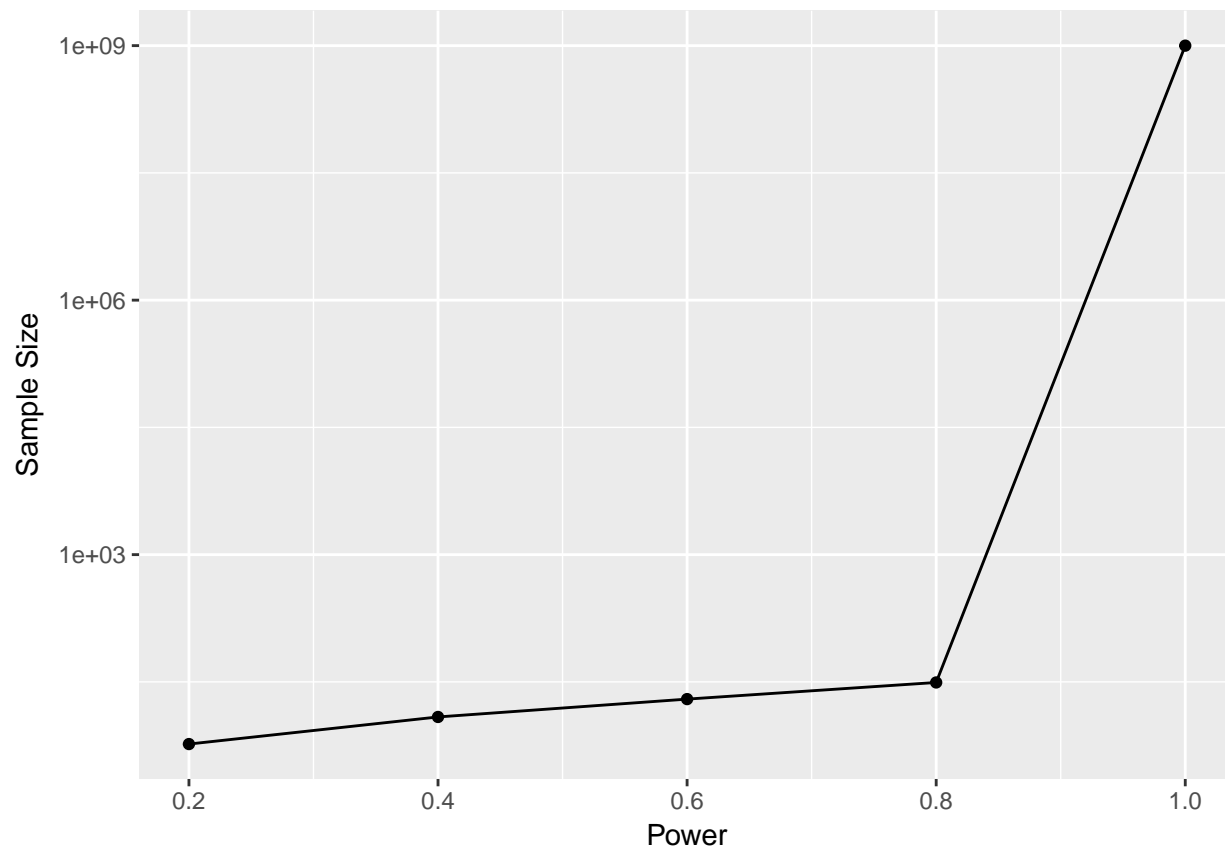
```
effect_size
```

```
## [1] 0.7217556
```

```
# Power analysis
pow <- seq(0.2, 1.0, by = 0.2)
sample_size <- c()
for (i in pow) {
  x <- pwr.t.test(d=effect_size, sig.level = 0.05, power = i, type="two.sample")
  sample_size <- c(sample_size, x$n)
}
print(sample_size)
```

```
## [1] 5.847457e+00 1.220121e+01 1.980978e+01 3.112412e+01 1.000000e+09
```

```
# Plot
size_vs_power <- data.frame(pow, sample_size)
ggplot(size_vs_power, aes(x = pow, y = sample_size)) +
  geom_point() +
  geom_line() +
  scale_y_log10() + xlab("Power") + ylab("Sample Size")
```



The plot tells us that when the effect size, significance level, and type of a test is determined, the increase in power will lead to a larger sample size. Larger power means less chance of making Type II error. So if we want to eliminate Type II error, we will need to increase the sample size.

Exercise 9 - predicting effect size (optional)



- from the hypothesis test in Exercise 8, do a power analysis (using `pwr::pwr.t.test` function in R) where we set `sig.level = 0.05` and the `n = 100` (for each group)
- use `ggplot2` to plot power versus effect size for 5 different, reasonable effect sizes (you calculated the effect size from the actual data in the previous exercise, so try some values above and below that).
- explain and interpret your results in 2-3 sentences

Solution

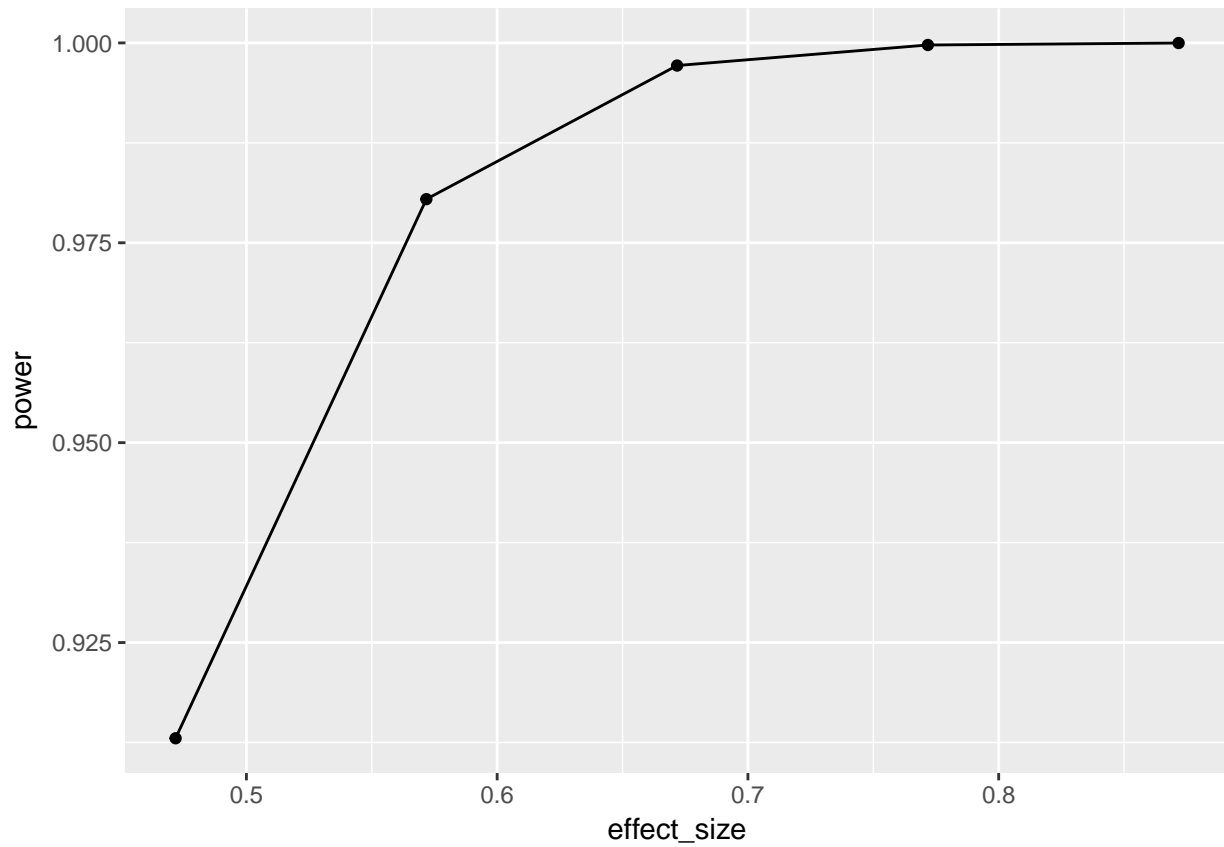
1 Power analysis

```
effect_size
```

```
## [1] 0.7217556
```

```
d_list <- seq(effect_size - 0.25, effect_size + 0.15, by = 0.1)
pow_list <- c()
for (i in d_list) {
  x <- pwr.t.test(d = i, n = 100, sig.level = 0.05, type = "two.sample")
  pow_list <- c(pow_list, x$power)
}

power_vs_effect <- data.frame(power = pow_list, effect_size = d_list)
ggplot(power_vs_effect, aes(x = effect_size, y = power)) +
  geom_point() +
  geom_line()
```



2 It can be seen that the power increases as the effect size increases given a fixed sample size. Thus, if the two groups means are farther away from each other, then it is easier to be detected by the t test.