MATH 254 - Statistical Modeling and Applications - Lab 4

STEMI Hospital data - Permutation Tests

Tural Sadigov

9/21/22

Table of contents

0.1 NAME: ______________

0.2 Who are you working with? ______________

We will work on the same STEMI Hospital data from Lab 3. Remember that in 2002-2003, a study was conducted on in-hospital deaths from myocardial infarction with ST elevation (STEMI). In this mini-project, we will investigate the relationship between observed deaths and expected deaths within each category using permutations tests for true correlation coefficient and true slope of the regression model. You can use your codes from Lab 3 as a reference.

In linear regression, we assume that there is a TRUE linear relationship between two variables (Observed deaths and Expected deaths) with some normal error, z, that has mean 0 and some unknown variance, σ^2:

$$ObservedDeaths = \beta_0 + \beta_1(ExpectedDeaths) + z$$

This linear model has three UNKNOWN (true) parameters: β_0, β_1, σ. Here β_0 is the true y-intercept, β_1 is the true slope and σ is the true standard deviation of the error. We will estimate them using the data at hand, and fit a model. Our estimators will be $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}$, and our fitted model will be

$$\widehat{ObservedDeaths} = \hat{\beta}_0 + \hat{\beta}_1(ExpectedDeaths)$$

Load libraries and data.

```
# libraries
library(tidyverse)
library(infer)
# install.packages('broom')
library(broom)


# Import Hospital data from GitHub
hospital = read_csv(...)

# view the  the hospital data
hospital

# data wrangling
# create (mutate) expected_deaths column
# rename Deaths as observed_deaths
# select expected_deaths and observed_deaths only
hospital_df <-
  hospital %>%
  ... %>%
  ... %>%
  ...

# view the data we will be using
hospital_df
```

0.3 Linear Regression - true slope

1. Use `lm()` to obtain the fitted slope, $\hat{\beta}_1$, of the linear regression model? (Just turn on eval, and make sure you understand the code)

```
broom::tidy(lm(observed_deaths ~ expected_deaths,
   data = hospital_df))
```

```
# OR
lm(observed_deaths ~ expected_deaths,
   data = hospital_df) %>%
  broom::tidy()

# One could also use infer package
# hospital_df %>%
#  specify(observed_deaths ~ expected_deaths) %>%
#  fit()
```

Can this be attributed to the chance alone? Conduct a Permutation test with clearly defined steps below:

2. Define the population parameter of interest in the context of the problem. State Null and Alternative Hypothesis, and the significance level.
3. Create Permutation distribution of the fitted slope, $\hat{\beta}_1$. Calculate p-value using the permutation distribution.

```
slope_hat <-
  hospital_df %>%
  specify(observed_deaths ~ expected_deaths) %>%
  calculate(stat = ...)

slope_hat

set.seed(2022)
null_dist <-
  hospital_df %>%
  specify(...) %>%
  hypothesize(null = 'independence') %>%
  generate(reps = ..., type = ...) %>%
  calculate(stat = 'slope')

null_dist %>%
  visualize() +
  shade_p_value(obs_stat = ...,
                direction = "two-sided")

null_dist %>%
  get_p_value(obs_stat = ...,
              direction = "two-sided")
```

4. Make a decision with your choice of significance level in the context of the problem.
5. Could it be the case that you made a wrong decision? Explain.

0.4 True Correlation coefficient

6. What is the sample correlation coefficient between expected deaths and observed deaths? Do you think the correlation is weak, medium or strong?

```
hospital_df %>%
  summarise(r = cor(..., ...))
```

Can this sample correlation coefficient be attributed to chance alone? Conduct a Permutation test with clearly defined steps below:

7. Define the parameter of interest in the context of the problem. State Null and Alternative Hypothesis, and the significance level.
8. Create Permutation distribution of observed sample correlation, $\hat{p}$. Calculate p-value using the permutation distribution.

```
correlation_hat <-
  hospital_df %>%
  specify(observed_deaths ~ expected_deaths) %>%
  calculate(stat = ...)

correlation_hat

set.seed(2023)
null_dist <-
  hospital_df %>%
  specify(...) %>%
  hypothesize(null = ...) %>%
  generate(reps = 1000, type = ...) %>%
  calculate(stat = ...)

null_dist %>%
  visualize() +
  shade_p_value(obs_stat = ...,
                direction = "two-sided")

null_dist %>%
  get_p_value(obs_stat = ...,
              direction = "two-sided")
```

9. Make a decision with your choice of significance level in the context of the problem.
10. Could it be the case that you made a wrong decision? Explain.