

Hypothesis testing

Rakibul Islam Prince

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
library(pacman)
p_load(dplyr, GGally, ggplot2, ggthemes, ggvis, httr, lubridate, plotly, rio, rmarkdown, shiny, stringr)
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

```
stroke_data<-import("../data/healthcare-dataset-stroke-data.csv")
stroke_data<-na.omit(stroke_data)
clean_stroke_data<-stroke_data[!apply(stroke_data=="N/A",1,any),]
clean_g_stroke_data<-clean_stroke_data %>%
  filter(gender!="Other")
clean_g_stroke_data$bmi<-as.numeric(clean_g_stroke_data$bmi)
str(clean_g_stroke_data)
```

```
## 'data.frame':    4908 obs. of  12 variables:
##  $ id           : int  9046 31112 60182 1665 56669 53882 10434 60491 12109 12095 ...
##  $ gender       : chr   "Male" "Male" "Female" "Female" ...
##  $ age          : num   67 80 49 79 81 74 69 78 81 61 ...
##  $ hypertension : int    0 0 0 1 0 1 0 0 1 0 ...
##  $ heart_disease : int    1 1 0 0 0 1 0 0 0 1 ...
##  $ ever_married  : chr    "Yes" "Yes" "Yes" "Yes" ...
##  $ work_type     : chr    "Private" "Private" "Private" "Self-employed" ...
##  $ Residence_type : chr    "Urban" "Rural" "Urban" "Rural" ...
##  $ avg_glucose_level: num   229 106 171 174 186 ...
##  $ bmi           : num   36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
##  $ smoking_status : chr    "formerly smoked" "never smoked" "smokes" "never smoked" ...
##  $ stroke        : int    1 1 1 1 1 1 1 1 1 1 ...
```

```
clean_g_stroke_data$hypertension<-as.factor(clean_g_stroke_data$hypertension)
clean_g_stroke_data$heart_disease<-as.factor(clean_g_stroke_data$heart_disease)
clean_g_stroke_data$stroke<-as.factor(clean_g_stroke_data$stroke)
data<-clean_g_stroke_data
sample_n(data,5)
```

```
##      id gender age hypertension heart_disease ever_married work_type
## 1 72435 Female  37              0              0          Yes  Private
```

```
## 2 39714 Male 12 0 0 No children
## 3 64864 Male 63 1 0 Yes Private
## 4 27380 Female 36 0 0 Yes Private
## 5 30480 Male 48 0 0 Yes Private
## Residence_type avg_glucose_level bmi smoking_status stroke
## 1 Urban 217.11 29.1 never smoked 0
## 2 Urban 64.08 18.2 Unknown 0
## 3 Rural 60.17 23.5 smokes 0
## 4 Rural 74.14 31.2 formerly smoked 0
## 5 Urban 85.54 32.2 smokes 0
```

Let's check whether the mean of average glucose level differs between men and women. null hypothesis $H_0: \mu_1 - \mu_2 = 0$ alternative hypothesis $H_A: \mu_1 - \mu_2 \neq 0$ (two-sided test) $\alpha = 0.05$

```
avg_glu_m <- data$avg_glucose_level[data$gender == "Male"]
print(paste("male avg glucose level:", mean(avg_glu_m)))
```

```
## [1] "male avg glucose level: 108.131720537046"
```

```
avg_glu_f <- data$avg_glucose_level[data$gender == "Female"]
print(paste("female avg glucose level:", mean(avg_glu_f)))
```

```
## [1] "female avg glucose level: 103.329913703832"
```

```
glu_m_mean <- mean(avg_glu_m)
glu_m_len <- length(avg_glu_m)
glu_m_var <- var(avg_glu_m)
glu_f_mean <- mean(avg_glu_f)
glu_f_len <- length(avg_glu_f)
glu_f_var <- var(avg_glu_m)
```

```
mu0 = 0
mean_diff <- glu_m_mean - glu_f_mean
se <- sqrt(glu_m_var / glu_m_len + glu_f_var / glu_f_len)
```

```
t <- (mean_diff - mu0) / se
print(paste("calculate Z-score:", t))
```

```
## [1] "calculate Z-score: 3.55782983108098"
```

```
print(paste("For 95% CI Z-score:", round(qnorm(0.975), 3)))
```

```
## [1] "For 95% CI Z-score: 1.96"
```

Here, we are seeing that our calculated Z-score is way outside of 95% CI. So, we reject the null hypothesis.

```
# Confidence Interval
round(mean_diff + c(-1, 1) * qnorm(0.975) * se, 3)
```

```
## [1] 2.157 7.447
```

here 95% interval doesn't contain 0. so we reject the null hypothesis.

```
#using R library
t.test(data$avg_glucose_level ~ data$gender, var.equal = FALSE)
```

```
##
## Welch Two Sample t-test
##
## data: data$avg_glucose_level by data$gender
```

```
## t = -3.6739, df = 4089.7, p-value = 0.0002419
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
## -7.364237 -2.239377
## sample estimates:
## mean in group Female    mean in group Male
##          103.3299          108.1317
```

We believe that average BMI for the population is larger than 30 (regardless of the gender). Let's check our beliefs:

null hypothesis $H_0: \mu_0 = 30$ alternative hypothesis $H_A: \mu_0 > 30$ (one-sided test) $\alpha = 0.05$

```
bmi = data %>% select(bmi) %>% summarise(mean=mean(bmi), sd=sd(bmi))
bmi
```

```
##      mean      sd
## 1 28.89456 7.85432
```

```
length(data$bmi)
```

```
## [1] 4908
```

```
mu_o <- 30
alpha <- 0.05
z <- (bmi$mean - mu_o) / (bmi$sd / sqrt(length(data$bmi)))
```

```
print(paste("calculate Z-score:", z))
```

```
## [1] "calculate Z-score: -9.86004516573277"
```

```
print(paste("For 95% CI Z-score:", round(qnorm(0.95), 3)))
```

```
## [1] "For 95% CI Z-score: 1.645"
```

calculate z-score is less than .95 quantile.so, we accept the Null hypothesis.

```
#using R library
t.test(data$bmi, mu=mu_o, alternative="greater")
```

```
##
## One Sample t-test
##
## data: data$bmi
## t = -9.86, df = 4907, p-value = 1
## alternative hypothesis: true mean is greater than 30
## 95 percent confidence interval:
##  28.71012      Inf
## sample estimates:
## mean of x
## 28.89456
```

If the claim is: H_0 : ratio of strokes among men and women is the same: $p_m - p_f = 0$ H_A : ratio of strokes among men and women is different: $p_m - p_f \neq 0$ $\alpha = 0.05$

```
# Number of males and females with stroke
stroke_m <- nrow(data %>% filter(gender == "Male", stroke == 1))
stroke_f <- nrow(data %>% filter(gender == "Female", stroke == 1))

# Total number of males and females
```

```

total_m <- nrow(data %>% filter(gender == "Male"))
total_f <- nrow(data %>% filter(gender == "Female"))

# Proportion test
result <- prop.test(x = c(stroke_m, stroke_f), n = c(total_m, total_f), alternative="two.sided")

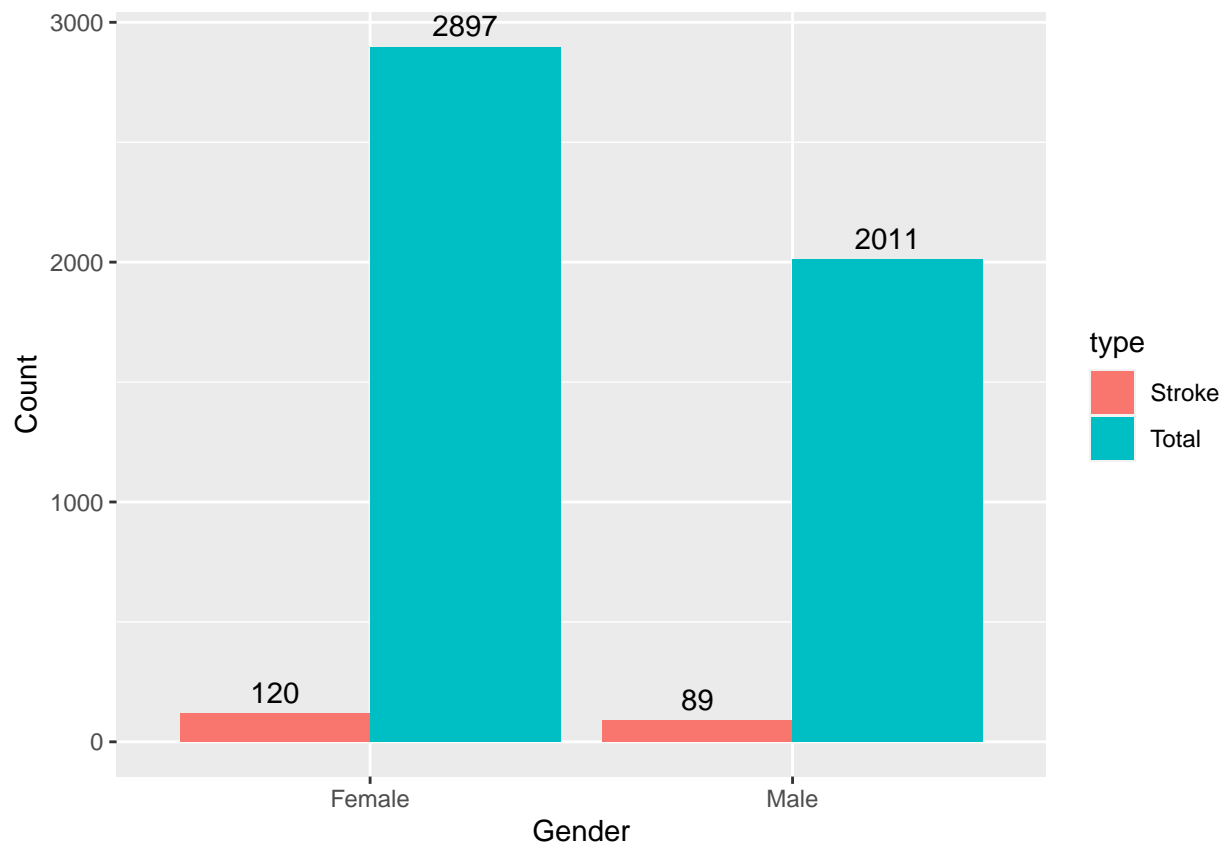
# Print result
print(result)

##
## 2-sample test for equality of proportions with continuity correction
##
## data: c(stroke_m, stroke_f) out of c(total_m, total_f)
## X-squared = 0.16955, df = 1, p-value = 0.6805
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.009138838 0.014807694
## sample estimates:
##      prop 1      prop 2
## 0.04425659 0.04142216

# Create a data frame
bar_data <- data.frame(
  gender = rep(c("Male", "Female"), 2),
  count = c(stroke_m, stroke_f, total_m, total_f),
  type = rep(c("Stroke", "Total"), each = 2)
)

# Create the bar plot
ggplot(bar_data, aes(x = gender, y = count, fill = type)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = count), vjust = -0.5, position = position_dodge(0.9)) +
  labs(y = "Count", x = "Gender")

```



```
print(paste0("Calculated p-value is: ", round(result$p.value, 3)))
```

```
## [1] "Calculated p-value is: 0.681"
```

```
print(paste0("Significance level alpha is: 0.05"))
```

```
## [1] "Significance level alpha is: 0.05"
```

As, $p\text{-value} > \alpha$: accept H_0 That means, proportion of both Males and females having stroke is quite similar.