BM HW5

JingYao Geng

2020-11-20

Problem 1

Given the non-normal distributions, now you are asked to use an alternative, non-parametric test to assess and comment on the difference in Ig-M levels between the two groups (please ignore unanswered and missing values).

```
# import:
anti = read_csv("./data/Antibodies.csv") %>%
  janitor::clean_names() %>%
  filter(smell != "Unanswered/Others") %>%
  drop_na()

# check the normal approximation assumption:
  n = anti %>%  group_by(smell) %>%  count() # normal: 81, altered: 178. yes!

# tidy:
anti_tidy = anti %>%
  mutate(rank = rank(antibody_ig_m)) %>%
  pivot_wider(
      names_from = "smell",
      values_from = "antibody_ig_m"
    ) %>%
  arrange(rank) %>%
  select(rank, Normal, Altered) # duplicated values observed, rank ties.

# find the ties:
ties = anti_tidy %>% count(rank) %>% filter(n > 1) %>% pull(n)

# the null expectation of T1
n2 = n$n[1] # 178 Altered
n1 = n$n[2] # 81 Normal

# the test statistics T for ties:
t1 = anti_tidy %>% drop_na(Normal) %>% summarise(t1 = sum(rank)) # 9157
t = (abs(t1 - n1 * (n1 + n2 + 1) / 2) - 1/2) #1372.5
T = t/sqrt(n1*n2/12 * (n1 + n2 + 1 - sum(ties*(ties^2 - 1))/((n1 + n2)*(n1 + n2 + 1)))) # 2.455714
t_crit = qnorm(0.975) #1.96

# wilcoxon Rank-Sum test:
test = wilcox.test(anti_tidy %>% pull(Normal), anti_tidy %>% pull(Altered), mu = 0)
# we need to add the n1(n1+1)/2 term, for the same value of t1 = 9157
```

```
test$statistic = test$statistic + n1*(n1 + 1)/2; test
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  anti_tidy %>% pull(Normal) and anti_tidy %>% pull(Altered)
## W = 9157, p-value = 0.01406
## alternative hypothesis: true location shift is not equal to 0
```

After ignoring the unanswered and missing values from the antibodies dataset, we have 2 smell groups: "Normal" and "Altered" in terms of the Ig-M levels. There are 178 observations in the "Altered" group and 81 observations in the "Normal" group. We will use **Wilcoxon Rank-Sum test**: the non parametric equivalent of the Two Sample Independent t-test.

Hypotheses to be tested are:

H_0 : The medians of IgM level of the two groups (Normal and Altered) are equal.

H_1 : The medians of IgM level of the two groups are not equal.

- Normal-Approximation is satisfied: n_{normal} *and* $n_{altered} \geq 10$

Test Statistics: With ties, the test statistic is: $T = \frac{|T_1 - \frac{n_1(n_1+n_2+1)}{2}| - \frac{1}{2}}{\sqrt{(n_1 n_2/12)[(n_1+n_2+1) - \sum_{i=1}^{g} t_i(t_i^2-1)/(n_1+n_2)(n_1+n_2-1)]}}$

Decision Rule:

Reject H_0 : $T > z_{1-\alpha/2}$, with p_value $= 2 \times [1 - \Phi(T)]$

Fail to reject H_0, *otherwise.*

Conclusion

Based on the p-value 0.01406 from the Wilcoxon Rank-Sum test, we reject the null, and conclude that the medians of IgM level are not euqal for the Normal group and Altered group at significance level of 0.05. Moreover, we find the test statistic T is equal to 2.455714, which is greater than the t_crit value of 1.96. This indicates that we reject the null hypotheses as well.

Problem 3

```
gpa = read_csv("./data/GPA.csv") %>%
  janitor::clean_names()
gpa_r = lm(gpa~act, data = gpa)
summary(gpa_r)
##
## Call:
## lm(formula = gpa ~ act, data = gpa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
```

```
## act          0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
anova(gpa_r)
## Analysis of Variance Table
##
## Response: gpa
##            Df Sum Sq Mean Sq F value   Pr(>F)
## act         1  3.588  3.5878  9.2402 0.002917 **
## Residuals 118 45.818  0.3883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
tidy(gpa_r) %>% knitr::kable()
```

Generate a scatter plot and test whether a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). Use a level of significance of 0.05.

term	estimate	std.error	statistic	p.value
(Intercept)	2.1140493	0.3208948	6.587982	0.0000000
act	0.0388271	0.0127730	3.039777	0.0029166

```
qt(0.975, 118) #1.980272
## [1] 1.980272
```

Hypotheses:

$H0 : \beta_1 = \beta_{10}$

$H1 : \beta_1 \neq \beta_{10}$ $(where \;\; \beta_{10} = 0)$

Test statistics

$t_{stat} = \frac{\hat{\beta_1} - \beta_{10}}{se(\hat{\beta_1})} = \frac{0.03883 - 0}{0.01277} \cong 3.0407$

$t_{n-2,1-\alpha/2} = t_{118,0.975} \cong 1.980272$

Decision Rule

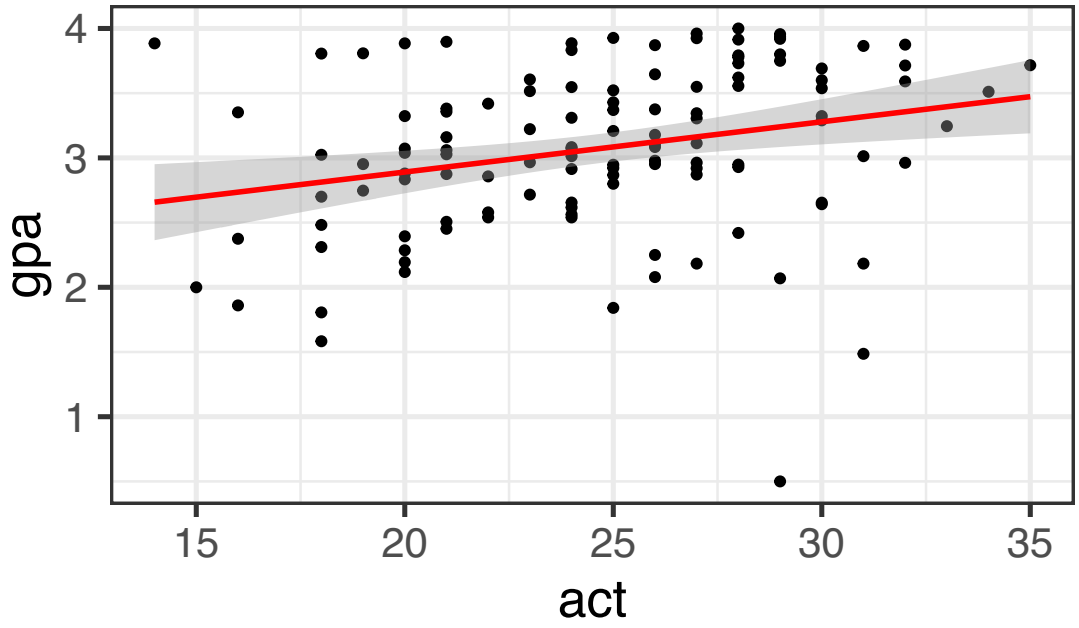
Reject H0 if $|t| > t_{n-2,1-\alpha/2}$ Fail to reject H0, otherwise.

Conclusion

Since $t_{stats} = 3.0407 > t_{118,0.975} = 1.980272$, we reject the null hypothese and conclude that $\beta_1 \neq \beta_{10}$, at 0.05 significant level, there is a significant linear association between student's ACT score and GPA at the end of the freshman year.

**Scatter plot with regression and 95% confidence band:

```
gpa %>%
  ggplot(aes(act, gpa)) +
  geom_point() +
  theme_bw(base_size = 20) +
  geom_smooth(method = 'lm', se = TRUE, color = 'red')
```



Estimated regression line equation The estimated slope is 0.0388271, estimated intercept is 2.1140493. $\hat{GPA} = 2.1140493 \ + \ 0.0388271 \times ACT$

95% confidence interval for 1

A 95% confidence interval for the true slope is:

$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} * se(\hat{\beta}_1)$

where $se(\hat{\beta}_1) = \sqrt{MSE / \sum_{i=1}^{n}(X_i - \bar{X})^2}$

```
confint(gpa_r,level = 0.95)
##                  2.5 %     97.5 %
## (Intercept) 1.47859015 2.74950842
## act         0.01353307 0.06412118
```

Thus, the 95% confidence interval for the true slope is **(0.01353307, 0.06412118)**. It does not include zero in the interval.

95% confidence interval when ACT is 28

```
predict(gpa_r, data.frame(act = 28), interval = "confidence", level = 0.95)
##        fit      lwr      upr
## 1 3.201209 3.061384 3.341033
```

The 95% interval estimate of the mean freshman GPA for students whose ACT test score is 28 is between 3.061384 and 3.341033. We are 95% confident that the true mean freshman GPA for students with a ACT score of 28 lies in interval (3.061384, 3.341033).

95% prediction interval when ACT is 28

```
predict(gpa_r, data.frame(act = 28), interval = "prediction", level = 0.95)
##        fit      lwr      upr
## 1 3.201209 1.959355 4.443063
```

The 95% prediction interval for Anne when she obtained a ACT score of 28 is between 1.959355 and 4.443063.

The prediction interval is wider than the confidence interval because dor prediction interval, we have another term to account for: the error term. We have a larger standard error in the prediction interval, which causes the prediction interval wider than the confidence interval.

Problem 2

(a) $Y_i = \beta_0 + \beta_1 X + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$

Then, we have $Y_i \sim N(\beta_0 + \beta_1 X, \sigma^2)$

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$

Log-likelihood: $\ell(\beta_0, \beta_1, \sigma^2)$

$$= \log\left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}\right]$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n} \frac{(Y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

(1) $\frac{\partial}{\partial \beta_0} \log L(\beta_0, \beta_1, \sigma^2) = 0$

$$\frac{\partial}{\partial \beta_0} \log L(\beta_0, \beta_1, \sigma^2) = \frac{\partial}{\partial \beta_0}\left[-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n} \frac{(Y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right]$$

$$= \frac{\partial}{\partial \beta_0}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \overset{\text{set}}{=} 0$$

$$\therefore \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 n\bar{x}$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 n\bar{x}}{n}$$

$$= \bar{y} - \hat{\beta}_1 \bar{x}$$

(2) $\frac{\partial}{\partial \beta_1} \log L(\beta_0, \beta_1, \sigma^2) = 0$

$$\frac{\partial}{\partial \beta_1} \log L(\beta_0, \beta_1, \sigma^2) = \frac{\partial}{\partial \beta_1}\left[-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n} \frac{(Y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right]$$

$$= \frac{\partial}{\partial \beta_1}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\cdot 2(-x_i)$$

$$= -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) \overset{\text{set}}{=} 0$$

$$\sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} \beta_0 X_i - \sum_{i=1}^{n} \beta_1 X_i^2 = 0$$

$$\sum_{i=1}^{n} Y_i X_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

$$\sum_{i=1}^{n} Y_i X_i = (\bar{Y} - \beta_1 \bar{X}) \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

$$\sum_{i=1}^{n} Y_i X_i = \bar{Y} \sum_{i=1}^{n} X_i - \beta_1 \bar{X} \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

$$\sum_{i=1}^{n} Y_i X_i = \bar{Y} \sum_{i=1}^{n} X_i - n \beta_1 \bar{X}^2 + \beta_1 \sum_{i=1}^{n} X_i^2$$

$$\sum_{i=1}^{n} Y_i X_i = n \bar{X} \bar{Y} + \beta_1 \left(\sum_{i=1}^{n} X_i^2 - n \bar{X}^2 \right)$$

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^{n} X_i^2 - n \bar{X}^2} = \operatorname{corr}[Y_i, X_i] \cdot \frac{sd(Y_i)}{sd(X_i)}$$

(b) $e_i = y_i - \hat{y}_i$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

$$= \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$$

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$= \sum_{i=1}^{n} y_i - n \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$= n \bar{y} - n \hat{\beta}_0 - n \hat{\beta}_1 \bar{x}$$

$$\therefore \hat{\beta}_0 = \frac{n \bar{y} - n \hat{\beta}_1 \bar{x}}{n}$$

$$= \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\therefore \sum_{i=1}^{n} e_i = n \bar{y} - n \hat{\beta}_1 \bar{x} - n [\bar{y} - \hat{\beta}_1 \bar{x}]$$

$$= n \bar{y} - n \hat{\beta}_1 \bar{x} - n \bar{y} + n \hat{\beta}_1 \bar{x}$$

$$= 0.$$