Biostatistics Homework 7

By 𝕃umi (张鹿鸣 12112618)

1. Vaccine

1.1)

TFTFTFFFT

I'm wondering, the ARV/ARU ratio should also be the odds ratio of vaccination among the people who get the disease right?

I think it's right.

1.2)

```
data.frame(
  `mRNA-1273` = c(11, 15181 - 11, 15181),
  Placebo = c(185, 15170 - 185, 15170),
  Total = c(11 + 185, 15181 + 15170 - 11 - 185, 15181 + 15170),
  row.names = c("COVID 19", "No Symptoms", "Total")
)
```

```
##             mRNA.1273 Placebo Total
## COVID 19           11     185   196
## No Symptoms     15170   14985 30155
## Total           15181   15170 30351
```

1.3)

From 1.1, we know that

$$\text{vaccine efficacy} = (1 - \frac{ARV}{ARU}) \times 100\%$$

Thus the vaccine efficacy for our mRNA vaccine is:

$$\text{vaccine efficacy}_{\text{mRNA}} = (1 - \frac{11/15181}{185/15170}) \times 100\% \approx 94.06$$

1.4)

H_0: Vaccination and infection status are independent.

H_1: Vaccination and infection status are not independent.

For the expected values in the table, I will just list an equation here and ignore the calculations. Suppose e_{ij} is the expected value on the ith row and jth column:

$$e_{ij} = \frac{\sum_{k=1}^{i} a_{kj}}{\sum_{m=1}^{i}\sum_{n=1}^{j} a_{mn}} \times \sum_{k=1}^{j} a_{ik}$$

R helped me to calculate the table, which is the output of the following code:

```r
vaccine <- data.frame(
  `mRNA-1273` = c(11, 15181 - 11),
  Placebo = c(185, 15170 - 185),
  row.names = c("COVID 19", "No Symptoms")
)
chisq_vaccine <- chisq.test(vaccine)
chisq_vaccine$expected
```

```
##               mRNA.1273     Placebo
## COVID 19       98.03552    97.96448
## No Symptoms 15082.96448 15072.03552
```

```r
chisq_vaccine
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  vaccine
## X-squared = 153.82, df = 1, p-value < 2.2e-16
```

As we can see, the χ^2 score is extremely high (153.8177), and thus the adjusted p-value is extremely low, 2.538393×10^{-35} to be more preciese.

Here we performed a continuity correction, if we didn't perform the correction, the p-value and statistics should be:

```r
chisq.test(vaccine, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  vaccine
## X-squared = 155.6, df = 1, p-value < 2.2e-16
```

The uncorrected p-value is 1.035122×10^{-35}, which is also very small.

1.5)

First we should ensure that our samples are randomly selected and that they are independent from each other.

Then, we should make sure that the number in all the cells are larger than ten, so that we can approximate the binomial distribution to a normal distribution.

2. Using Effect Size Too

2.1)

A wrong interpretation:

From the paper, we can know that if we would like to use the Cohen's d to calculate the effect size, then:

$$d = \frac{M_1 - M_0}{s}$$

Where M_1 and M_0 are the mean of the two groups. s is the standard deviation of either groups.

The table we are using:

```r
blood <- data.frame(
  A = c(1188, 670),
  `Non-A` = c(2506, 1105),
  row.names = c("Healthy", "COVID-19")
)
blood
```

```
##              A Non.A
## Healthy   1188  2506
## COVID-19   670  1105
```

```r
mean_A <- mean(blood[,1])
mean_nonA <- mean(blood[,2])
sd_A <- sd(blood[,1])
eff <- (mean_A - mean_nonA) / sd_A
eff
```

```
## [1] -2.392969
```

The correct interpretation:

I don't think it's that right. Since the mean number in A and nonA don't have a real meaning. I googled it again and Google told me that it is best to use the odds ratio for a 2 by 2 contingency tables. I think I misunderstood the article. Well, maybe we should calculate it again.

$$OR = \frac{ad}{bc} = \frac{2506 \times 670}{1188 \times 1105} \approx 1.279$$

In this case, the diseased are considered the event of interest.

2.2)

D

2.3)

The effect size in the first problem can be calculated by:

$$OR = \frac{\text{COVID}_{\text{vaccine}} \times \text{no COVID}_{\text{no vaccine}}}{\text{COVID}_{\text{no vaccine}} \times \text{no COVID}_{\text{vaccine}}} = \frac{11 \times 14985}{185 \times 15170} = 0.0587$$

Which is quite small, indicating that taking the vaccine might have a huge impact getting COVID. The RNA vaccine really lowers the odds of getting COVID.

2.4)

The effect size tell us in a normalized scale how much the value in your experiment group differs from the control group. We should care about it because p-values only tell us how significant something is going to happen, but we cannot directly see the difference quantitatively. It might be very significant due to a very large sample size, while the actual difference between the two groups are small.

3. The White Stock

3.1)

i.

The equation for Pearson's Correlation Coefficient is:

$$r = \frac{\sigma(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

```
stork <- data.frame(
  Storks = c(100,300,1,5000,9,140,3300,2500,4,5000,5,30000,1500,5000,8000,150,25000),
  Birth_rate = c(83,87,118,117,59,774,901,106,188,124,551,610,120,367,439,82,1576)
)
r <- cov(stork$Storks, stork$Birth_rate) / sqrt(var(stork$Storks) * var(stork$Birth_rate))
r
```

```
## [1] 0.6202653
```

Here we directly used the R function to calculate the covariance, which is

```
cov(stork$Storks, stork$Birth_rate)
```

```
## [1] 2246592
```

We can also calculate the covariance using the formula:

$$\sigma(X,Y) = E(XY) - E(X)E(Y)$$

Which gives us:

```
mean(stork$Storks * stork$Birth_rate) - mean(stork$Storks) * mean(stork$Birth_rate)
```

```
## [1] 2114440
```

Which is different because R function cov() uses a different formula:

$$\sigma(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

So we should do some adjustments:

```
n <- length(stork$Storks)
(mean(stork$Storks * stork$Birth_rate) - mean(stork$Storks) * mean(stork$Birth_rate)) * n / (n - 1)
```

```
## [1] 2246592
```

Now it's the same.

The Pearson correlation coefficient is 0.6202.

ii)

Based on the correlation coefficient, there is a positive linear relationship between the Storks(pairs) and Birth rate (10^3/yr) .

iii)

We can calculate a t-score based on the correlation coefficient.

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} \sim T(n-2)$$

Where n is the number of samples. So our t-score is,

$$t = 0.6202 \times \sqrt{\frac{17-2}{1-0.6202^2}} \approx 3.062$$

And the corresponding p-value:

```
pt(3.062, 15, lower.tail = F) * 2
```

```
## [1] 0.007907499
```

iv)

For a simple linear regression using OLS,

$$a = \frac{\sigma(X,Y)}{\text{Var}(X)}, b = \bar{y} - a\bar{x}$$

So based on our previous answers, if we use our own calculation of covariance:

```
a_adjust <- 2114440 / var(stork$Storks)
b_adjust <- mean(stork$Birth_rate) - a_adjust * mean(stork$Storks)
c(a_adjust, b_adjust)
```

```
## [1]   0.0270999 233.5979093
```

If we use R's calculation of the covariance:

```
a <- 2246592 / var(stork$Storks)
b <- mean(stork$Birth_rate) - a * mean(stork$Storks)
c(a, b)
```

```
## [1]   0.02879364 225.02869336
```

We can also use R's function to calculate the slope and intersection:

```
model <- lm(stork$Birth_rate ~ stork$Storks)
model
```

```
##
## Call:
## lm(formula = stork$Birth_rate ~ stork$Storks)
##
## Coefficients:
##  (Intercept)  stork$Storks
##    225.02869       0.02879
```

v)

There is a difference between the ways you calculate the covariance. Because we are calculating the covariance of a sample, we need to follow the formula $\sigma(X,Y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$, instead of $E(XY) - E(X)E(Y)$. Since the population mean is unknown, and one degree of freedom is used to estimate $\bar{x}$ and $\bar{y}$.

3.2)

So, if we want to see the correlation between area and birth rate:

```
area <- data.frame(
  Area = c(28750,83860,30520,111000,43100,544000,357000,132000,
           41900,93000,301280,312680,92390,237500,504750,41290,779450),
  Birth_rate = c(83,87,118,117,59,774,901,106,188,124,551,610,120,367,439,82,1576)
)
r_area <- cov(area$Area, area$Birth_rate) / sqrt(var(area$Area) * var(area$Birth_rate))
r_area
```

```
## [1] 0.9225445
```

As we can see, the pearson correlation coefficient of area and birth rate is 0.9225, which is pretty high. Indicating there is a positive linear relationship between the area of the country and the birth rate.

We can also calculate the t-score:

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} = 0.9225 \times \sqrt{\frac{17-2}{1-0.9225^2}} \approx 9.2591$$

```
pt(9.2591, 15, lower.tail = F) * 2
```

```
## [1] 1.362273e-07
```

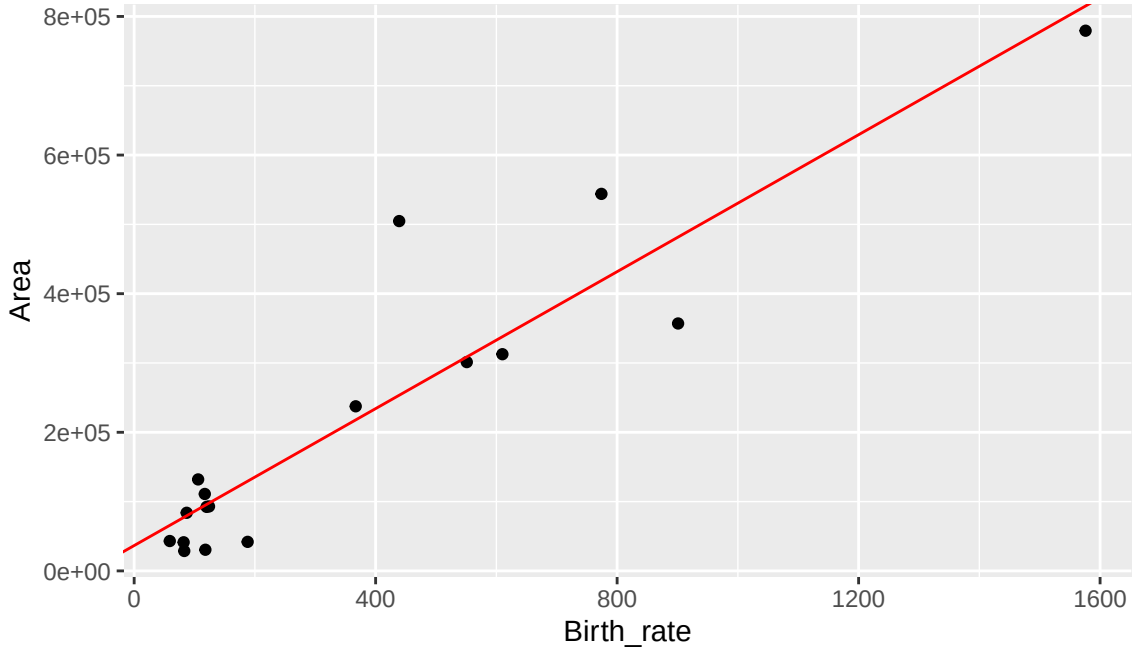
The p-value is 1.362×10^{-7}, which is extremely small.

```
lm(area$Area ~ area$Birth_rate)
```

```
##
## Call:
## lm(formula = area$Area ~ area$Birth_rate)
##
## Coefficients:
##     (Intercept)  area$Birth_rate
##           36553              494
```

The coefficients for the simple linear regression are slope = 494, intercept = 36553. As we can see from the plot below, the function fits quite well.

```
ggplot(data = area, mapping = aes(Birth_rate, Area)) +
  geom_point() +
  geom_abline(slope = 494, intercept = 36553, colour = "red")
```



3.3)

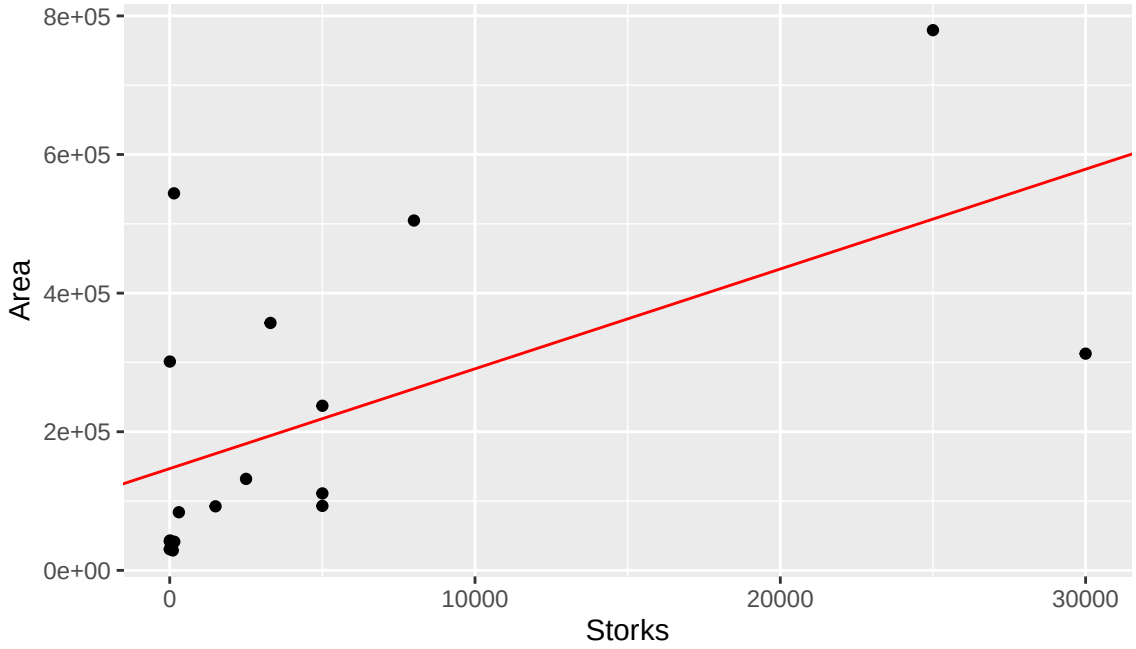
We might need to perform another correlation check between the area and the amount of Storks.

```
cor(area$Area, stork$Storks)
```

```
## [1] 0.5793423
```

We can see there is a slightly positive correlation between the land area of the country and the number of storks observed in that country.

```
area_stork <- lm(area$Area ~ stork$Storks)
ggplot(mapping = aes(stork$Storks,area$Area)) +
  geom_point() + xlab('Storks') + ylab('Area') +
  geom_abline(slope = coef(area_stork)[2], intercept = coef(area_stork)[1], colour = 'red')
```



With the very strong correlation between the birth rate and country size, the not-so-strong correlation of stork numbers and country size might be amplified, resulting in a strong correlation between the stork population and birth rate.

3.4)

As described in 3.3), the confounding variable in this study might be the area of the country. Which is reasonable. The larger the country, the more birds it might "contain", and it have a larger birth rate since they might have a larger population.

Thus, the area and both stork population and birth rate have positive correlations. Leading to a misjudgment. I don't know if China have these lot storks, so if we include some Asian countries, like China or India, this correlation might not be that obvious.