大數據統計分析與預測 第十章作業(Hypothesis test)

Created by 黃彥鈞 Weber YC, Huang (m946108006)

Q3. What is a p-value? What does the p-value mean in words?

In statistics, the p-value is the probability of obtaining the observed results of a test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

Q7. Explain the analogy between type I and type II errors in a test of hypothesis and the false positive and false negative results that occur in diagnostic testing.

In statistical hypothesis testing, a type I error is the rejection of a true null hypothesis (also known as a "false positive" finding or conclusion), while a type II error is the non-rejection of a false null hypothesis (also known as a "false negative" finding or conclusion)

Q9. The distribution of diastolic blood pressures for the population of female diabetics between the ages of 30 and 34 has an unknown mean $\mu_d$and standard deviation $\sigma_d = 9.1\, mm\ Hg$. It may be useful to physicians to know whether the mean of this population is equal to the mean diastolic blood pressure of the general population of females in this group, $74.4\, mm\ Hg$.

(a) What is the null hypothesis of the appropriate test?

$H_0 : \mu = 74.4\, mm\ Hg$

(b) What is the alternative hypothesis?

$H_0 : \mu \neq 74.4\, mm\ Hg$

(c) A sample of ten diabetic women is selected；their mean diastolic blood pressure is $\overline{x}_d = 84\, mm\ Hg$. Using this information, conduct a two-sided test at the $\alpha = 0.05$ level of significance. What is the p-value of the test?

> *R-code :*

```
# Q9. Calculate 2 sided p_value
cal_pval <- function(){
  sample_avg <- as.numeric(readline(prompt="Enter Sample avg: "))
  h0 <- as.numeric(readline(prompt="Enter null hypothesis: "))
  sd <- as.numeric(readline(prompt="Enter standard deviation (popluation): "))
  size <- as.numeric(readline(prompt="Enter sample size: "))
  cat(' Sample avg :', sample_avg, '\n', 'Null hypothesis :', h0, '\n', 'Standard deviation (popluation) :'
      , sd, '\n', 'Sample size :', size, '\n')

  z <- (sample_avg-h0)/(sd/sqrt(size))

  alpha = 0.05
  z.half.alpha = qnorm(1-alpha/2)
  ans <- c(-z.half.alpha, z.half.alpha)
  cat('\n', 'Area', ans, '\n','Z :', z)

  p = 2*pnorm(-abs(z))
  cat('\n','P-value :', p)
}

cal_pval() # Run code
```

> *Output :*

```
Sample avg : 84
Null hypothesis : 74.4
Standard deviation (popluation) : 9.1
Sample size : 10

Area -1.959964 1.959964
Z : 3.336029

 ==== Answer ====
P-value : 0.0008498424
```

(d) What conclusion do you draw from the results of the test?

Since P-value < 0.05, we reject H_0 and conclude the mean diastolic blood pressure for the population of female diabetics between the ages 30-34 is not equal to 74.4 $mm\ Hg$.

(e) Would your conclusion have been different if you had chosen $\alpha = 0.01$ instead of $\alpha = 0.05$?

Since P-value < 0.01, the conclusion will remain as the same.

Q11. Body mass index is calculated by dividing a person's weight by the square of his or her height； it is a measure of the extent to which the individual is overweight. For the population of middle-aged men who later develop diabetes mellitus, the distribution of baseline body mass indices is approximately normal with an unknown mean μ and standard deviation σ . A sample of 58 men selected from this group has mean $\overline{x} = 25\ kg/m^2$ and the standard deviation $s = 2.7\ kg/m^2$.

(a) Construct a 95% confidence interval for population mean μ .

Since the sigma is unknown, so we should conduct the t-distribution

> *R-code :*

```
# Q11.
# T distribution calculate confidence interval
t_dis_ci <- function(){
  x <- as.numeric(readline(prompt="Enter sample mean: "))
  sd <- as.numeric(readline(prompt="Enter sample standard
deviation: "))
  n <- as.numeric(readline(prompt="Enter size: "))
  ci_rate <- as.numeric(readline(prompt="Enter
confidential rate: "))
  cat('Sample mean :', x,'\n')
  cat('Sample standard deviation :', sd,'\n')
  cat('Size :', n,'\n')
  cat('Confidential rate :', ci_rate*100, '%','\n')

  ci <- 1 - ((1 - ci_rate)/2)
  error <- qt(ci,df = n-1)*sd/sqrt(n)
  left <- x - error
  right <- x + error
  cat('\n', '==== Answer ====', '\n')
  cat('Confidence interval :', '(',left, ',', right, ')')
}

t_dis_ci()# Run code
```

> *Output :*

```
Sample mean : 25
Sample standard deviation : 2.7
Size : 58
Confidential rate : 95 %

 ==== Answer ====
Confidence interval : ( 24.29007 , 25.70993 )
```

(b) At the 0.05 level of significance, test whether the mean baseline body mass index for the population of middle-aged men who do develop diabetes is equal to $24 \, kg/m^2$, the mean of the population for men who do not. What is the $p - value$ of the test?

$H_0 \, : \mu = 24 \, kg/m^2$

$H_A \, : \mu \neq 24 \, kg/m^2$

> *R-code :*

```
cal_pval_t <- function(){
  sample_avg <- as.numeric(readline(prompt="Enter Sample avg: "))
  h0 <- as.numeric(readline(prompt="Enter null hypothesis: "))
  sd <- as.numeric(readline(prompt="Enter Sample standard deviation: "))
  size <- as.numeric(readline(prompt="Enter sample size: "))
  cat(' Sample avg :', sample_avg, '\n', 'Null hypothesis :', h0, '\n', 'Sample standard deviation :'
      , sd, '\n', 'Sample size :', size, '\n')

  z <- (sample_avg-h0)/(sd/sqrt(size))

  cat('\n','t :', z, '\n')

  p = 2*pnorm(-abs(z))
  cat('\n', '==== Answer ====', '\n')
  cat('P-value :', p)
}

cal_pval_t() # Run code
```

> *Output :*

```
Sample avg : 25
Null hypothesis : 24
Sample standard deviation : 2.7
Sample size : 58

t : 2.820657

 ==== Answer ====
P-value : 0.004792546
```

(c) What do you conclude?

Since P-value < 0.05, we reject H_0 and conclude the mean baseline body mass index for the population of men who later develop diabetes mellitus is not equal to 24 kg/m^2.

(d) Based on the 95% confidence interval, would you have expected to reject or not to reject the null hypothesis? Why?

Since the value 24 doesn't lie in the 95% confidence interval for μ , we should expect the null hypothesis would be rejected.

Q14. Data from Framinghan study allow us to compare distributions of initial serum cholesterol levels for two populations of males: those who go on to develop coronary heart disease and those who do not. The mean serum cholesterol levels of the population of men who do not develop heart disease is $\mu = 219\ mg/100\ ml$ and the standard deviation is $\sigma = 41\ mg/100\ ml$. Suppose, however, that you do not know the true population mean； instead, you hypothesize that μ is equal to $244\ mg/100\ ml$. This is the mean initial serum cholesterol level of men who eventually develop the disease. Since it is believed that the mean serum cholesterol level for the men who do not develop the heart disease cannot be higher then the mean level who do, a one-sided test conducted at the $\alpha = 0.05$ level of significance is appropriate.

(a) What is the probability of making a Type I error?

Since it is a one-sided(left-tail) test, the null hypothesis and alternative hypothesis are :

$H_0 : \mu \geq 244$

$H_A : \mu < 244$

the Type I error would be reject H_0, that is, false rejecting $H_0 : \mu \geq 244$

(b) If a sample of size 25 is selected from the population of men who do not go on to develop coronary heart disease, what is the probability a making a Type II error?

> *R-code :*

```
# Q14.
cal_t2error_lower <- function(){
  mu <- as.numeric(readline(prompt="Enter population avg:
"))
  h0 <- as.numeric(readline(prompt="Enter null hypothesis:
"))
  sigma <- as.numeric(readline(prompt="Enter population
standard deviation: "))
```

```
  size <- as.numeric(readline(prompt="Enter sample size:
"))
  afa <- as.numeric(readline(prompt = "Enter alpha: "))
  cat(' Pop mean :', mu, '\n', 'Null hypothesis :',
      h0, '\n', 'Pop standard deviation :',
      sigma, '\n', 'Sample size :', size, '\n', 'Alpha :',
afa, '\n')
  se <- sigma/sqrt(size) # standard error
  q <- qnorm(afa, mean=h0, sd=se)
  t2error <-pnorm(q, mean=mu, sd=se, lower.tail=FALSE)

  cat('\n', '==== Answer ====', '\n')
  cat('Type II error rate :', t2error*100, '%', '\n')
  cat('Power :', (1-t2error)*100, '%')
}
cal_t2error_lower()
```

> *Output :*

```
 Pop mean : 219
 Null hypothesis : 244
 Pop standard deviation : 41
 Sample size : 25
 Alpha : 0.05

 ==== Answer ====
Type II error rate(Beta) : 8.017032 %
Power(1-Beta) : 91.98297 %
```

(c) What is the power of the test?

$Power(1-\beta) = 91.98\,\%$

(d) How could you increase the power?

The factor that influence $power(1-\beta)$ are :

- Type I error rate (α)
- Wide of $|\mu_1 - \mu_0|$
- Sample size

In fact, the α is commonly set as 0.05, and the wide of $|\mu_1 - \mu_0|$ is not easy to modify. The quick way to increase power is to select the correct sample size.

$$size(n) = \frac{((Z_\alpha + Z_\beta) \times \sigma)^2}{(|\mu_1 - \mu_0|)^2}$$

For example, if we give the sample mean difference $|\mu_1 - \mu_0|(\Delta) = 2$, $\sigma = 5$, the sample size will be $54.91 \approx 55$

> *R-code :*

```
# determine the perfect sample size
# install.packages('pwr')
library(pwr)
delta = 2
sigma = 5
d = delta/sigma
pwr.t.test(d=d, sig.level=.05, power = .90, type =
'one.sample', alternative = 'greater')
```

> *Output :*

```
One-sample t test power calculation

             n = 54.90553
             d = 0.4
     sig.level = 0.05
         power = 0.9
   alternative = greater
```

(e) You wish to test the null hypothesis

$H_0 : \mu \geq 244\, mg/100\, ml$

against the alternative

$H_A : \mu < 244\, mg/100\, ml$

at the $\alpha = 0.05$ level of significance. If the true population mean is as low as $219\, mg/100\, ml$, you want to risk only a 5% chance of failing to reject H_0. How large a sample would be required?

$\alpha = 0.05$
$\beta = 0.05$
$\Delta = 219 - 244 = -25$
$\sigma = 41$

> *R-code :*

```
library(pwr)
delta = -25
sigma = 41
d = delta/sigma
pwr.t.test(d=d, sig.level=.05, power = .95, type =
'one.sample', alternative = 'less')
```

> *Output :*

```
 One-sample t test power calculation

              n = 30.51692
              d = -0.6097561
      sig.level = 0.05
          power = 0.95
    alternative = less
```

- Answer : 30.51

(f) How would the sample size change if you were willing to risk a 10% chance of failing to reject a false null hypothesis?

$\alpha = 0.05$
$\beta = 0.10$
$\Delta = 219 - 244 = -25$
$\sigma = 41$

> *R-code :*

```
library(pwr)
delta = -25
sigma = 41
d = delta/sigma
pwr.t.test(d=d, sig.level=.05, power = .95, type =
'one.sample', alternative = 'less')
```

> *Output :*

```
One-sample t test power calculation

              n = 24.45157
              d = -0.6097561
      sig.level = 0.05
          power = 0.9
    alternative = less
```

- Answer : 24.45