Data Analytics

Yong Zheng

Illinois Institute of Technology

Chicago, IL, 60616, USA

School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

Schedule

- Quick Reviews
- Assignment #1
- Hypothesis Testing by Using R

Schedule

- Quick Reviews
  - Use Sample to Estimate Population
    - Input: Sample data and confidence level
    - Output: confidence interval

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{y} \pm t_{\alpha/2}s_{\bar{y}} = \bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

  - Hypothesis Testing
    - Elements, steps and methods to make decisions
    - One-sample hypothesis testing
      - Large vs small sample size
    - Two-sample hypothesis testing
      - Paired or independent samples
      - Large vs small sample size

Schedule

- Quick Reviews
- Assignment #1
- Hypothesis Testing by Using R

Schedule

- Quick Reviews
- Assignment #1
- Hypothesis Testing by Using R
  - We introduce R practice by using the data in Hypothesis Testing_Using R.zip
  - You will do your own practice by using our data in Case Study 1 – Student Grades

Schedule

- Quick Reviews
- Hypothesis Testing by Using R
  - We introduce R practice by using the data in Hypothesis Testing_Using R.zip
  - You will do your own practice by using our data in Case Study 1 – Student Grades

Data

- Hypothesis Testing_Using R.zip
  - Unzip it
  - d1.txt, d2.txt → grades in ITMD 525 and 527
  - d1 = grades on two classes, independent sample
  - d2 = grades on two classes, paired sample

Statistical Inference

- There are two ways for us to estimate or infer the population parameter, such as population mean:
  1) By estimating its value
     For example: estimate the age of people in USA
  2) By testing hypothesis about its value
     For example:
     Method-1 is better than method 2.
     Students in 527(04) are better than 527(01).
     The average of working hours/day is no more than 8

Statistical Inference by Estimation

- Produce a confidence interval
  1) If n >= 30, normal distribution, z value

     $$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

  2) Otherwise, t distribution, t value

     $$\bar{y} \pm t_{\alpha/2}s_{\bar{y}} = \bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

Statistical Inference by Estimation

- Produce a confidence interval in R

  Method 1: Calculate them in R

  Method 2: Run z-test or t-test

Statistical Inference by Estimation

- Produce a confidence interval: large sample size

  If n >= 30, normal distribution, z value

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

```
> x = mean(d525)
> s = sd(d525)
> n = length(d525)
> n
[1] 40
> err = qnorm(0.975)*s/sqrt(n)
> left = x - err
> right = x + err
> left
[1] 40.06214
> right
[1] 58.23786
```

95% confidence level

Statistical Inference by Estimation

- Produce a confidence interval: small sample size

  If n < 30, t distribution, t value

$$\bar{y} \pm t_{\alpha/2} s_{\bar{y}} = \bar{y} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$, α = 1 – confidence level

```
> d525_small = data[1:20, 1]
> x = mean(d525_small)
> s = sd(d525_small)
> n = length(d525_small)
> n
[1] 20
> df = n-1
> err = qt(0.975, df)*s/sqrt(n)
> left = x - err
> right = x + err
> left
[1] 40.91888
> right
[1] 68.48112
```

95% confidence level

Statistical Inference

- There are two ways for us to estimate or infer the population parameter, such as population mean:
  1) By estimating its value
     For example: estimate the age of people in USA
  2) By testing hypothesis about its value
     For example:
     Method-1 is better than method 2.
     Students in 527(04) are better than 527(01).
     The average of working hours/day is no more than 8

Hypothesis Testing

- One sample
  - H0: average grade in ITMD525 is 60
  - Ha: average grade in ITMD525 is not 60
  - Level of significance: 0.05
- If sample size is large enough, n>=30

```
z.test(x, y = NULL, alternative = "two.sided", mu = 0, sigma.x = NULL,
sigma.y = NULL, conf.level = 0.95)
```

You need to install the package “BSDA”

Hypothesis Testing

- One sample
  - H0: average grade in ITMD525 is 60
  - Ha: average grade in ITMD525 is not 60
  - Level of significance: 0.05
- If sample size is large enough, n>=30

```
> z.test(d525,NULL,alternative="two.sided",mu=60,sigma.x=sd(d525),conf.level=0.95)

        One-sample z-Test

data:  d525
z = -2.34, p-value = 0.01928
alternative hypothesis: true mean is not equal to 60
95 percent confidence interval:
 40.06214 58.23786
sample estimates:
mean of x
    49.15
```

Hypothesis Testing

- One sample
  - H0: average grade in ITMD525 is 60
  - Ha: average grade in ITMD525 is not 60
  - Level of significance: 0.05
- If sample size is small, n<30

```
t.test(x, y = NULL,
    alternative = c("two.sided", "less", "greater"),
    mu = 0, paired = FALSE, var.equal = FALSE,
    conf.level = 0.95, ...)
```

Hypothesis Testing

- One sample
  - H0: average grade in ITMD525 is 60
  - Ha: average grade in ITMD525 is not 60
  - Level of significance: 0.05
- If sample size is small, n<30

```
> t.test(d525_small,NULL,alternative="two.sided",mu=60,paired=F,conf.level=0.95)

        One Sample t-test

data:  d525_small
t = -0.80494, df = 19, p-value = 0.4308
alternative hypothesis: true mean is not equal to 60
95 percent confidence interval:
 40.91888 68.48112
sample estimates:
mean of x
     54.7
```

Hypothesis Testing

- Two Samples: Independent, e.g., d1.txt
  - H0: average grade in ITMD525 and ITMD527 is same
  - Ha: They are different
  - Level of significance: 0.05
- If sample size is large, n>=30

```
z.test(x, y = NULL, alternative = "two.sided", mu = 0, sigma.x = NULL,
sigma.y = NULL, conf.level = 0.95)
```

You need to install the package “BSDA”

Hypothesis Testing

- Two Samples: Independent, e.g., d1.txt
  - H0: average grade in ITMD525 and ITMD527 is same
  - Ha: They are different
  - Level of significance: 0.05
- If sample size is large, n>=30

```
> z.test(d525,d527,alternative="two.sided",mu=0,sigma.x=sd(d525),sigma.y=sd(d527),conf.level=0.95)

	Two-sample z-Test

data:  d525 and d527
z = -0.92982, p-value = 0.3525
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.181196   6.481196
sample estimates:
mean of x mean of y 
    49.15     55.00 
```

Hypothesis Testing

- Two Samples: Independent, e.g., d1.txt
  - H0: average grade in ITMD525 and ITMD527 is same
  - Ha: They are different
  - Level of significance: 0.05
- If sample size is small, n<30

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

Note: make sure paired=F, var.equal=T

Hypothesis Testing

- Two Samples: Independent, e.g., d1.txt
  - H0: average grade in ITMD525 and ITMD527 is same
  - Ha: They are different
  - Level of significance: 0.05
- If sample size is small, n<30

```
> t.test(d525,d527,alternative="two.sided",mu=0,paired=F,var.equal=T,conf.level=0.95)

        Two Sample t-test

data:  d525 and d527
t = -0.9272, df = 88, p-value = 0.3564
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.38839   6.68839
sample estimates:
mean of x mean of y
    49.15     55.00
```

Hypothesis Testing

- Two Samples: Paired, e.g., d2.txt
  - H0: Students in ITMD525 & ITMD527 perform the same
  - Ha: Their performance are different
  - Level of significance: 0.05
- If sample size is large, n>=30

```
> diff=m527-m525
> z.test(diff,NULL,alternative="two.sided",mu=0,sigma.x=sd(diff),sigma.y=NULL,conf.level=0.95)

        One-sample z-Test

data:  diff
z = 0.88942, p-value = 0.3738
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -7.462589 19.862589
sample estimates:
mean of x
      6.2
```

Hypothesis Testing

- Two Samples: Paired, e.g., d2.txt
  - H0: Students in ITMD525 & ITMD527 perform the same
  - Ha: Their performance are different
  - Level of significance: 0.05
- If sample size is small, n<30

```
t.test(x, y = NULL,
      alternative = c("two.sided", "less", "greater"),
      mu = 0, paired = FALSE, var.equal = FALSE,
      conf.level = 0.95, ...)
```

Note: make sure paired=T, var.equal=T; x and y have same sample size

Hypothesis Testing

- Two Samples: Paired, e.g., d2.txt
  - H0: Students in ITMD525 & ITMD527 perform the same
  - Ha: Their performance are different
  - Level of significance: 0.05
- If sample size is small, n<30

```
> m525_small=m525[1:20]
> m527_small=m527[1:20]
> t.test(m525_small,m527_small,alternative="two.sided",mu=0,paired=T,var.equal=T,conf.level=0.95)

        Paired t-test

data:  m525_small and m527_small
t = -0.79922, df = 19, p-value = 0.434
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26.59844  11.89844
sample estimates:
mean of the differences
                  -7.35
```

Schedule

- Quick Reviews
- Hypothesis Testing by Using R
  - We introduce R practice by using the data in Hypothesis Testing_Using R.zip
  - You will do your own practice by using our data in Case Study 1 – Student Grades

Practice By Yourself

- Steps
  - Understand your data
  - Find some attributes you are interested in, and propose some hypothesis
  - Follow the steps of hypothesis testing
    - Write down the H0 and H1
    - Make a decision about one-tailed vs two-tailed
    - Define confidence level
    - Make conclusions by using correct tests according to sample size