AMRITA
VISHWA VIDYAPEETHAM

18MAT289

STATISTICAL INFERENCE THEORY

LAB REPORT

SREEHARI P SREEDHAR
CB.SC.I5DAS20032

Table of Contents

Covariance

The covariance of two variables x and y in a data set measures how the two are linearly related. A positive covariance would indicate a positive linear relationship between the variables, and a negative covariance would indicate the opposite.

The sample covariance is defined in terms of the sample means as:

Cov = Sum((xi - xBar)(yi - yBar)) / (n - 1)

Similarly, the population covariance is defined in terms of the population mean µx, µy as:

Cov = Sum((xi - µx)(yi - µy)) / N

Problem

Find the covariance of eruption duration and waiting time in the data set faithful. Observe if there is any linear relationship between the two variables.

Solution

```
duration = faithful$eruptions
waiting = faithful$waiting

cov(duration, waiting)

## [1] 13.97781
```

Correlation

The correlation coefficient of two variables in a data set equals to their covariance divided by the product of their individual standard deviations. It is a normalized measurement of how the two are linearly related.

Formally, the sample correlation coefficient is defined by the following formula, where sx and sy are the sample standard deviations, and sxy is the sample covariance.

r = sxy/(sx * sy)

Similarly, the population correlation coefficient is defined as follows, where σx and σy are the population standard deviations, and σxy is the population covariance.

rho = σxy/(σx * σy)

If the correlation coefficient is close to 1, it would indicate that the variables are positively linearly related and the scatter plot falls almost along a straight line with positive slope. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it would indicate a weak linear relationship between the variables.

Problem

Find the correlation coefficient of eruption duration and waiting time in the data set faithful. Observe if there is any linear relationship between the variables.

Solution

```
cor(duration, waiting)

## [1] 0.9008112
```

Binomial Distribution

The binomial distribution is a discrete probability distribution. It describes the outcome of n independent trials in an experiment. Each trial is assumed to have only two outcomes, either success or failure. If the probability of a successful trial is p, then the probability of having x successful outcomes in an experiment of n independent trials is as follows.

f(x) = nCx * p^x * (1 - p)^(n-x)

Problem

Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

Solution

```
pbinom(4, size = 12, prob = 0.2)

## [1] 0.9274445
```

Poisson Distribution

The Poisson distribution is the probability distribution of independent event occurrences in an interval. If λ is the mean occurrence per interval, then the probability of having x occurrences within a given interval is:

f(x) = λ^x * e^(- λ) / x!

Problem

If there are twelve cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.

Solution

```
ppois(16, lambda = 12, lower = FALSE)

## [1] 0.101291
```

Normal Distribution

The normal distribution is defined by the following probability density function, where μ is the population mean and σ^2 is the variance.

f(x) = e^(-(x - μ)^2 / 2 * σ^2) / σ * sqrt(2 * pi)

Problem

Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?

Solution

```
pnorm(84, mean = 72, sd = 15.2, lower.tail = FALSE)

## [1] 0.2149176
```

Point Estimation of Population Mean and Proportion

```
library(MASS)
```

Problem 1

Find a point estimate of mean university student height with the sample data from survey.

Solution

```
height.survey = survey$Height

mean(height.survey, na.rm = TRUE)

## [1] 172.3809
```

Problem 2

Find a point estimate of the female student proportion from survey.

Solution

```
gender.response = na.omit(survey$Sex)
n = length(gender.response)

k = sum(gender.response == 'Female')

k/n

## [1] 0.5
```

Interval Estimation of Population Mean with Known Variance

For random sample of sufficiently large size, the end points of the interval estimate at (1 − α) confidence level is given as follows:

xBar +- zα/2 * σ / sqrt(n)

Problem

Assume the population standard deviation σ of the student height in survey is 9.48. Find the margin of error and interval estimate at 95% confidence level.

Solution

```
height.response = na.omit(survey$Height)

n = length(height.response)

sigma = 9.48
se = sigma/sqrt(n)

se

## [1] 0.6557453

E = qnorm(0.975) * se

E

## [1] 1.285237

mean(height.response) + c(-E, E)

## [1] 171.0956 173.6661
```

Interval Estimation of Population Mean with Unknown Variance

For random samples of sufficiently large size, and with standard deviation s, the end points of the interval estimate at (1 −α) confidence level is given as follows:

xBar +- t(α/2) * s / sqrt(n)

Problem

Without assuming the population standard deviation of the student height in survey, find the margin of error and interval estimate at 95% confidence level.

Solution

```
n = length(height.response)

s = sd(height.response)
```

```
se = s/sqrt(n)

se

## [1] 0.6811677

E = qt(0.975, df = n - 1)*se

E

## [1] 1.342878

mean(height.response) + c(- E, E)

## [1] 171.0380 173.7237
```

Sample Size of Population Mean

The formula below provides the sample size needed under the requirement of population mean interval estimate at (1 −α) confidence level, margin of error E, and population variance σ^2. Here, zα⁄2 is the 100(1 − α⁄2) percentile of the standard normal distribution.

n = (zα⁄2)^2 * σ^2 / E^2

Problem

Assume the population standard deviation σ of the student height in survey is 9.48. Find the sample size needed to achieve a 1.2 centimeters margin of error at 95% confidence level.

Solution

```
z = qnorm(0.975)
sigma = 9.48

E = 1.2

(z * sigma / E)^2

## [1] 239.7454
```

Interval Estimation of Population Proportion

If the samples size n and population proportion p satisfy the condition that np ≥ 5 and n(1 − p) ≥ 5, than the end points of the interval estimate at (1 − α) confidence level is defined in terms of the sample proportion as follows.

pBar +- zα⁄2 * sqrt(p * (1 - p) / n)

Problem

Compute the margin of error and estimate interval for the female students proportion in survey at 95% confidence level.

Solution

```
gender.response = na.omit(survey$Sex)

n = length(gender.response)
k = sum(gender.response == 'Female')

pbar = k/n

pbar

## [1] 0.5

se = sqrt( pbar * (1 - pbar) / n )

se

## [1] 0.03254723

E = qnorm(0.975) * se

E

## [1] 0.06379139

pbar + c(- E, E)

## [1] 0.4362086 0.5637914
```

Sample Size of Population Proportion

The formula below provides the sample size needed under the requirement of population proportion interval estimate at (1 − α) confidence level, margin of error E, and planned proportion estimate p. Here, zα⁄2 is the 100(1 − α⁄2) percentile of the standard normal distribution.

n = (zα⁄2)^2 * p * (1 - p) / E^2

Problem

Using a 50% planned proportion estimate, find the sample size needed to achieve 5% margin of error for the female student survey at 95% confidence level.

Solution

```
z = qnorm(0.975)
p = 0.5
```

```
E = 0.05

p * (1 - p) * (z / E)^2

## [1] 384.1459
```

Lower Tail Test of Population Mean with Known Variance

h0: m >= m0

z = (xBar - m0) / (σ / sqrt(n))

Reject h0 if z <= -zα

Problem

Suppose the manufacturer claims that the mean lifetime of a light is more than 10000 hours. In a sample of 30 light bulbs, it was found that they only last 9900 hours on average. Assume the population standard deviation is 120 hours. At 0.05 significance level, can we reject the claim by manufacturer?

Solution

```
xbar = 9900
n=30
m0= 10000
sigma = 120
z = (xbar-m0)/(sigma/sqrt(n))

z

## [1] -4.564355

a = 0.05
z.a = qnorm(1 - a)

-z.a

## [1] -1.644854
```

-1.644854 > -4.564355 We **reject** the null hypothesis.

Upper Tail Test of Population Mean with Known Variance

h0: m <= m0

z = (xBar - m0) / (σ / sqrt(n))

Reject h0 if z >= zα

Problem

Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the population standard deviation is 0.25 grams. At 0.05 significance level, we can reject the claim on the food label.

Solution

```
n = 35
xbar = 2.1
m0 = 2
sigma = 0.25
z = (xbar-m0)/(sigma/sqrt(n))

z

## [1] 2.366432

a = 0.05
z.a = qnorm(1 - a)

z.a

## [1] 1.644854
```

2.366432 > 1.644854 We **reject** the null hypothesis.

Two-Tailed Test of Population Mean with Known Variance

h0: m = m0

z = (xBar - m0) / (σ / sqrt(n))

Reject h0 if z >= | zα |

Problem

Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4kg. In a sample of 35 penguins at the same time this year in the same colony, the mean penguin weight is 14.6kg. Assume the population standard deviation is 2.5kg. At 0.05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

Solution

```
xbar = 14.6
m0 = 15.4
sigma = 2.5
n = 35
z = (xbar-m0)/(sigma/sqrt(n))
```

```
z

## [1] -1.893146

a = 0.05
z.a = qnorm(1 - a)

c(-z.a, z.a)

## [1] -1.644854  1.644854
```

-1.644854 <= -1.893146 <= 1.644854 We **do not reject** the null hypothesis.

Lower Tail Test of Population Mean with Unknown Variance

h0: m >= m0

t = (xBar - m0) / (s / sqrt(n))

Reject h0 if t <= -tα

Problem

Suppose the manufacturer claims that the mean lifetime of a light is more than 10000 hours. In a sample of 30 light bulbs, it was found that they only last 9900 hours on average. Assume the sample population standard deviation is 125 hours. At 0.05 significance level, can we reject the claim by manufacturer?

Solution

```
xbar = 9900
n = 30
m0 = 10000
s = 125
t = (xbar - m0)/(s/sqrt(n))

t

## [1] -4.38178

a = .05
t.a = qt(1 - a, df = n-1)

-t.a

## [1] -1.699127
```

-4.38178 < -1.699127 We **reject** the null hypothesis.

Upper Tail Test of Population Mean with Unknown Variance

h0: m >= m0

t = (xBar - m0) / (s / sqrt(n))

Reject h0 if t >= tα

Problem

Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 25 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the sample population standard deviation is 0. grams. At 0.05 significance level, we can reject the claim on the food label.

Solution

```
n = 25
xbar = 2.1
m0 = 2
s = 0.3
t = (xbar - m0)/(s/sqrt(n))

t

## [1] 1.666667

a = .05
t.a = qt(1 - a, df = n -1)

t.a

## [1] 1.710882
```

1.710882 > 1.666667 We **reject** the null hypothesis.

Two-Tailed Test of Population Mean with Unknown Variance

h0: m = m0

t = (xBar - m0) / (s / sqrt(n))

Reject h0 if t >= | tα |

Problem

Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 25 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the sample standard deviation is 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

Solution

```
xbar = 14.6
m0 = 15.4
s = 2.5
n = 25
t = (xbar-m0)/(s/sqrt(n))

t
```

```
## [1] -1.6
```

```
a = .05
t.a.half = qt(1 - a/2, df = n-1)
c(-t.a.half, t.a.half)
```

```
## [1] -2.063899  2.063899
```

-2.063899 <= -1.6 <= 2.063899
We **do not reject** the null hypothesis.

Lower Tail Test of Population Proportion

h0: p >= p0

z = (pbar - p0)/sqrt(p0*(1 - p0)/n)

Reject h0 if z <= -zα

Problem

Suppose 60% of citizens voted in last election. 85 out of 148 people in a telephone survey said that they voted in current election. At 0.5 significance level, can we reject the null hypothesis that the proportion of voters in the population is above 60% this year?

Solution

```
pbar = 85/148
p0 = 0.6
n = 148
z = (pbar - p0)/sqrt(p0*(1 - p0)/n)

z
```

```
## [1] -0.6375983
```

```
a = .05
z.a = qnorm(1 - a)
```

```
-z.a

## [1] -1.644854
```

-0.6375983 !< -1.644854
We **do not reject** the null hypothesis.

Alternative Solution

```
p = pnorm(z)

p

## [1] 0.2618676
```

0.2618676 > .05
We **do not reject** the null hypothesis.

Upper Tail Test of Population Proportion

h0: p >= p0

z = (pbar - p0)/sqrt(p0*(1 - p0)/n)

Reject h0 if z >= zα

Problem

Suppose that 12% of apples harvested in an orchard last year was rotten. 30 out of 214 apples in a harvest sample this year turns out to be rotten. At .05 significance level, can we reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year?

Solution

```
pbar = 30/214
p0 = 0.12
n = 214

z = (pbar - p0)/sqrt(p0*(1 - p0)/n)

z

## [1] 0.908751

a = 0.05
z.a = qnorm(1 - a)

z.a

## [1] 1.644854
```

0.908751 !< 1.644854
We **do not reject** the null hypothesis.

Alternative Solution

```
p = pnorm(z, lower.tail = FALSE)

p

## [1] 0.1817408
```

0.1817408 > 0.05
We **do not reject** the null hypothesis.

Two-Tailed Test of Population Proportion

h0: p >= p0

z = (pbar - p0)/sqrt(p0*(1 - p0)/n)

Reject h0 if z >= | zα |

Problem

Suppose a coin toss turns up 12 heads out of 30 trials. At .05 significance level, can one reject the null hypothesis that the coin toss is fair?

```
pbar = 12/30
p0 = .5
n = 30
z = (pbar - p0)/sqrt(p0*(1 - p0)/n)

z

## [1] -1.095445

a = .05
z.a.half = qnorm(1 - a/2)
c(-z.a.half, z.a.half)

## [1] -1.959964  1.959964
```

-1.959964 <= -1.095445 <= 1.959964
We **do not reject** the null hypothesis.

Alternative Solution

```
p = 2*pnorm(z, lower.tail = FALSE)

p

## [1] 1.726678
```

1.726678 > 0.05
We **do not reject** the null hypothesis.

T-Test

h0: mD = delta

T0 = (DBar - delta) / (SD/sqrt(n))

Problem

From the following data set, test whether two serum uric acid population have the same mean.

sample1 : 1.2, 0.8, 1.1, 0.7, 0.9, 1.1, 1.5, 0.8, 1.6, 0.9 sample2 : 1.7, 1.5, 2.0, 2.1, 1.1, 0.9, 2.2, 1.8, 1.3, 1.5

Solution

```
sample1 = c(1.2, 0.8, 1.1, 0.7, 0.9, 1.1, 1.5, 0.8, 1.6, 0.9)
sample2 = c(1.7, 1.5, 2.0, 2.1, 1.1, 0.9, 2.2, 1.8, 1.3, 1.5)

t.test(sample1, sample2)

##
##  Welch Two Sample t-test
##
## data:  sample1 and sample2
## t = -3.3046, df = 16.145, p-value = 0.004431
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.902565 -0.197435
## sample estimates:
## mean of x mean of y
##      1.06      1.61
```

0.004431 < 0.05
We **reject** the null hypothesis.

F-Test

h0: (σ1)^2 = (σ2)^2

f0 = (s1)^2 / (s2)^2

Problem

A quality control supervisor for an automobile manufacturer is concerned with uniformity in the number of defects in cars coming off the assembly line. If one assembly line has significantly more variability in the number of defects, then changes have to be made.

A -> N = 23 mean = 15 Std. Dev = 20
B -> N = 23 mean = 17 Std. Dev = 16

Solution

```
s1 = 20
s2 = 16

N = 23

df = N - 1

F = s1^2 / s2^2

F

## [1] 1.5625

critical_val = qf(0.05, df, df, lower.tail = FALSE)

critical_val

## [1] 2.04777
```

1.5625 < 2.04777
We **do not reject** the null hypothesis.

Paired T-Test

h0 = mD = 0

t0 = dBar/ (sd / sqrt(n))

Problem 1

A school of athletes has been taken a new instructor and want to test the effectiveness of the new type of training proposed by comparing the average time of 10 runners in the 200 meters and the time in seconds before and after training for each athletes is given below.

Before Training : 13.8,14.4,13.7,16.5,18.1,20.1,13.5,16.2,15.3,12.2

After Training : 13.6,14.5,12.9,16.1,17.7,20.912.9,16.8,16.9,12.0

Solution

```
x = c(13.8,14.4,13.7,16.5,18.1,20.1,13.5,16.2,15.3,12.2)
y = c(13.6,14.5,12.9,16.1,17.7,20.9,12.9,16.8,16.9,12.0)

t.test(x,y, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = -0.21331, df = 9, p-value = 0.8358
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.5802549  0.4802549
## sample estimates:
## mean difference
##           -0.05
```

As p = 0.8358 > 0.05, h0: m1 - m2 = 0 is **not rejected**.

Problem 2

A firm manufacturing rivers wants to limit variations in their length as much as possible. The lengths(in cm) of 10 rivers manufactured by a new process are given as :

2.15,1.99,2.05,2.12,2.17,2.01,1.98,2.03,2.25,1.93

Examine whether the new process can be considered superior to the old, if the old population has standard deviation 0.145 cm.

Solution

```
x = c(2.15,1.99,2.05,2.12,2.17,2.01,1.98,2.03,2.25,1.93)
n = 10
df = n-1
sigma0 = 0.145
v = var(x)

v
```

```
## [1] 0.01010667
```

```
chitest = df*v/(sigma0^2)

chitest
```

```
## [1] 4.326278
```

```
alpha = 0.05

qchisq(alpha,df, lower.tail=TRUE)
```

```
## [1] 3.325113
```

As 4.326278 > 3.325113 , h0 is **not rejected**.

```
alpha = 0.01

qchisq(alpha,df, lower.tail=TRUE)

## [1] 2.087901
```

As 4.326278 > 2.087901 , h0 is **not rejected**.

Chi Squared Test for Goodness of Fit

chiSq = Sum((fi - ei)^2 / ei)

Problem

The following shows the distribution of the digits in the number chosen random from a telephone directory:
Frequency : 3026,3107,2997,2996,3075,2933,3107,2972,2964,2853

Expected : 3000,3000,3000,3000,3000,3000,3000,3000,3000,3000

Solution

```
frequency = c(3026,3107,2997,2996,3075,2933,3107,2972,2964,2853)
expected = c(3000,3000,3000,3000,3000,3000,3000,3000,3000,3000)


chisq.test(frequency, p= expected, rescale.p = TRUE)

##
##  Chi-squared test for given probabilities
##
## data:  frequency
## X-squared = 19.085, df = 9, p-value = 0.02448

p = c(0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1)


chisq.test(frequency,p=p)

##
##  Chi-squared test for given probabilities
##
## data:  frequency
## X-squared = 19.085, df = 9, p-value = 0.02448
```

Chi Squared Test of Independence

chiSq = Sum((fij - eij)^2 / eij)

Problem 1

In order to determine the possible effect of a chemical treatment on the rate of germination of cotton seeds, a pot culture was conducted. The results are given below. Significance at 0.05.

Solution

```
Sample2 = matrix(c(118, 120, 22, 40), nrow = 2, ncol = 2)

Sample2

##      [,1] [,2]
## [1,]  118   22
## [2,]  120   40

chisq.test(Sample2)

## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  Sample2
## X-squared = 3.3808, df = 1, p-value = 0.06596
```

As p = 0.06596 > 0.05, h0: Attributes are independent is **not rejected**.

Problem 2

The severity of a disease and blood group were studied in a research project.

```
DisxBlood = matrix(c(51, 105, 384, 40, 103, 527, 10, 25, 125, 9, 17, 104), nr
ow = 3, ncol = 4)

DisxBlood

##      [,1] [,2] [,3] [,4]
## [1,]   51   40   10    9
## [2,]  105  103   25   17
## [3,]  384  527  125  104
```

Solution

```
chisq.test(DisxBlood)

## 
##  Pearson's Chi-squared test
## 
## data:  DisxBlood
## X-squared = 12.237, df = 6, p-value = 0.05689
```

As p = 0.2003 > 0.05, h0: Attributes are independent is **not rejected**.

```
qchisq(0.95, df = 6)

## [1] 12.59159
```

As chisq = 12.237 < chisq(table) = 12.59159, h0: Attributes are independent is **not rejected**.

Problem 3

A public opinion poll surveyed a simple random sample of 1000 voters.

```
GenxParty = matrix(c(220, 270, 170, 320, 70, 70), nrow = 2, ncol = 3)

GenxParty

##      [,1] [,2] [,3]
## [1,]  220  170   70
## [2,]  270  320   70
```

Solution

```
chisq.test(GenxParty)

## 
##  Pearson's Chi-squared test
## 
## data:  GenxParty
## X-squared = 15.81, df = 2, p-value = 0.0003688
```

As p = 0.0003688 < 0.05, h0: Attributes are independent is **rejected**.

```
qchisq(0.95, df = 2)

## [1] 5.991465
```

As chisq = 15.81 < chisq(table) = 5.991465, h0: Attributes are independent is **rejected**.