Golas, Ekal

UNIVERSITY OF TEXAS AT DALLAS | 800 W CaMPBELL RD RICHARDSON TX 75080

MINI PROJECT 4

Statistical methods for data science



Table of Contents

[Answers 2](#_Toc435108584)

[Exercise 1 2](#_Toc435108585)

[Exercise 2 3](#_Toc435108586)

[Exercise 3 4](#_Toc435108587)

[R Code 5](#_Toc435108588)

[Exercise 1 5](#_Toc435108589)

[Exercise 2 5](#_Toc435108590)

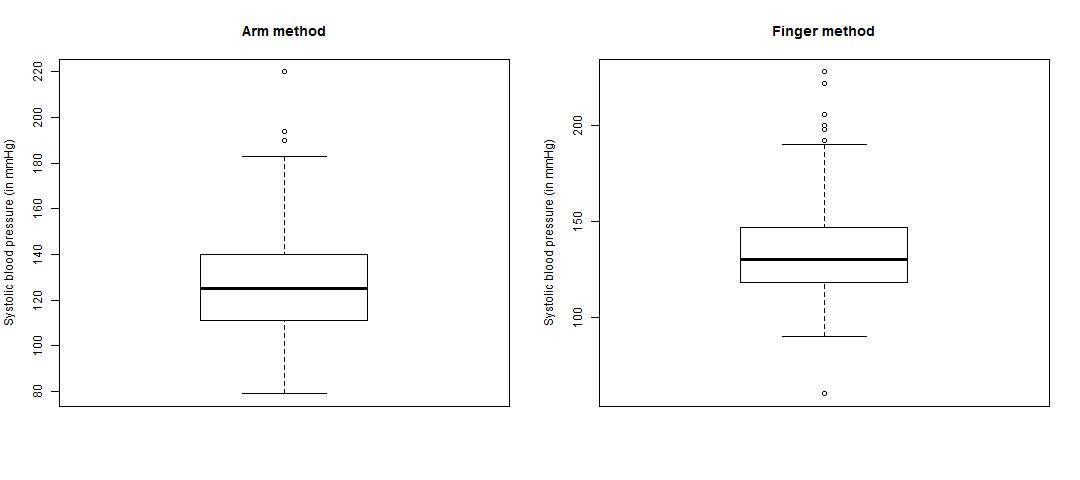
[Exercise 3 6](#_Toc435108591)

# Answers

## Exercise 1

The answers to the questions stated in (Choudhary, 2015) are as follows:-

1. **Perform an exploratory analysis of the data by examining the distributions of the measurements from the two methods using boxplots. Comment on what you see. Do the two distributions seem similar? Justify your answer**

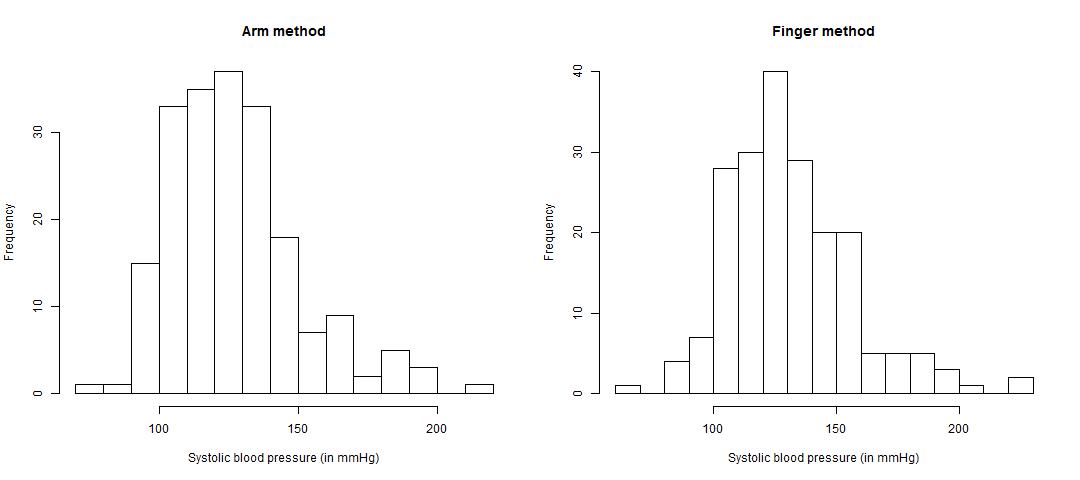
The results are displayed as follows:-

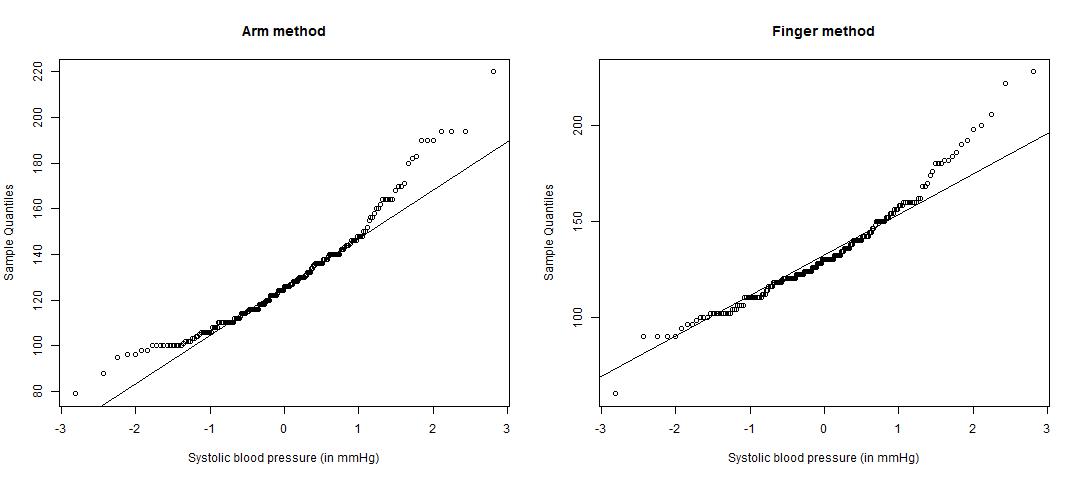
From the above plots, the distributions do not seem similar because:-

* Finger method has lot more outliers than the arm method
* Range of finger method is more compact than arm method
* Finger method is right-skewed while arm method looks symmetric

1. **Use histograms and QQ plots to examine the shapes of the two distributions. Comment on what you see. Does the assumption of normality seem reasonable? Justify your answer.**

The results are displayed as follows:-





From the above plots, the assumption of normality does not seem reasonable because for both distributions, the high end and the low end does not adhere to the normal distribution line.

1. **Construct an appropriate 95% confidence interval for the difference in the means of the two methods. Interpret your results. Can we conclude that the two methods have identical means? Justify your answer. What assumptions, if any, did you make to construct the interval? Do the assumptions seem to hold?**

The 95% CI is [-9.1109747, 0.5209747]. Since 0 falls in this interval, it is likely that the two means are identical. To derive this conclusion, the following assumptions were made:-

* Data comes from a normal distribution – As stated above, that is only partially correct
* Variance is not known and is assumed to be not equal
* Hence, CI is found by using t-distribution where the degrees of freedom are given by Satterthwaite approximation

1. **Perform an appropriate 5% level test to see if there is any difference in the means of the two methods. Be sure to clearly set up the null and alternative hypotheses. State your conclusion. What assumptions, if any, did you make to construct the interval? Do they seem to hold?**

With the 5% level test, the CI is found to be the same as above. The hypothesis was setup as:-

* Null : Difference in means of the two methods is 0
* Alternate : Difference in means of the two methods is not 0

With the level test, there is not enough evidence to reject the null hypothesis. So it can be concluded that the difference in means of two methods is 0. Following assumptions were made to derive this-:

* Variance was not assumed to be equal for this test
* Data was modelled by t-distribution

1. **Do the results from (c) and (d) seem consistent? Justify your answer.**

Yes the results from C and D seem consistent as the assumptions made for both the tests are the same and their results match.

## Exercise 2

The answers to the questions are stated as follows:-

1. **Set up the null and alternative hypotheses.**

Null: Mean of the population is 10, µ = 10

Alternate: Mean of the population is greater than 10, µ > 10

1. **Which test would you use? What is the test statistic? What is the null distribution of the test statistic?**

Here t-test would be used where the test statistic is T = (mean – sample.mean) / (sample.sd / sqrt(sample.size)). The null distribution of the test statistic follows a t distribution.

1. **Compute the observed value of the test statistic**

The observed value for test statistic is 1.974186

1. **Compute the p-value of the test using the usual way.**

P-value computed is 0.03153941

1. **Estimate the p-value of the test using Monte Carlo simulation. How do your answers in (d) and (e) compare?**

P-value computed using Monte Carlo simulation is 0.02494. It is generally lesser than the p-value obtained the usual way.

1. **State your conclusion at 5% level of significance.**

At 5% level of significance, as p-value < 0.05, we reject the null hypothesis. Hence, mean of the population is greater than 10.

## Exercise 3

The answers to the questions are stated as follows:-

1. **Construct an appropriate 95% confidence interval for the difference in mean credit limits of all credit cards issued in January 2011 and in May 2011. Interpret your results. Be sure to justify your choice of the interval.**

The 95% confidence interval for the difference in mean is computed as [-150.8982, -49.1018]. So it can be concluded that mean credit limit for credit cards issues in May is greater than mean credit limit of credit cards issued in January 2011 by an interval of 49 to 150.

1. **Perform an appropriate 5% level test to see if the mean credit limit of all credit cards issued in May 2011 is greater than the same in January 2011. Be sure to specify the hypotheses you are testing, and justify the choice of your test. State your conclusion.**

The hypothesis in this case is:-

* Null: Mean credit limit of all credit cards issued in May is as same as January
* Alternate: Mean credit limit of all credit cards issued in May is greater than the same in January 2011

Doing a t-test for the above, the P-value computed is 6.178294e-05 which is practically close to 0. Hence we reject the null hypothesis. So, it is concluded that mean credit limit for credit cards issues in May is greater than mean credit limit of credit cards issued in January 2011

# R Code

## Exercise 1

# Read the data from file

data = read.table("bp.txt", header = TRUE)

# Construct boxplots for finger and arm method

jpeg("Exercise 1 - boxplots.jpg", width = 1080, height = 480)

par(mfrow = c(1, 2))

boxplot(data$armsys, main = "Arm method", range = 1.5, ylab = "Systolic blood pressure (in mmHg)")

boxplot(data$fingsys, main = "Finger method", range = 1.5, ylab = "Systolic blood pressure (in mmHg)")

dev.off()

# Plot histograms for both methods

jpeg("Exercise 1 - histograms.jpg", width = 1080, height = 480)

par(mfrow = c(1, 2))

hist(data$armsys, main = "Arm method", xlab = "Systolic blood pressure (in mmHg)", breaks = 20)

hist(data$fingsys, main = "Finger method", xlab = "Systolic blood pressure (in mmHg)", breaks = 20)

dev.off()

# Construct QQ plots for both methods

jpeg("Exercise 1 - QQPlots.jpg", width = 1080, height = 480)

par(mfrow = c(1, 2))

qqnorm(data$armsys, main = "Arm method", xlab = "Systolic blood pressure (in mmHg)")

qqline(data$armsys)

qqnorm(data$fingsys, main = "Finger method", xlab = "Systolic blood pressure (in mmHg)")

qqline(data$fingsys)

dev.off()

# Compute the CI numerically using satterthwaite approx

S.x = var(data$armsys)

S.y = var(data$fingsys)

n.x = length(data$armsys)

n.y = length(data$fingsys)

v = (S.x / n.x + S.y / n.y) ^ 2 / (S.x^2 / (n.x^2 \* (n.x - 1)) + S.y ^ 2 / (n.y ^ 2 \* (n.y - 1)))

CI = mean(data$armsys) - mean(data$fingsys) + c(-1, 1) \* qt(1 - 0.05 / 2, v) \* sqrt(S.x / n.x + S.y / n.y)

CI

# Do a t test to get the confidence interval for the difference in means

t.test(data$armsys, data$fingsys, conf.level = 0.95, var.equal = FALSE)

## Exercise 2

# Define the data

sample.mean = 9.02

sample.sd = 2.22

sample.size = 20

H0.mean = 10

# Compute test statistic

test = (H0.mean - sample.mean) / (sample.sd / sqrt(sample.size))

test

# Get the p-value using t distribution

p.value = (1 - pt(abs(test), sample.size - 1))

p.value

# Get p-value using monte carlo simulation

p.value.sim = function(){

count = 0

for (i in 1:1000){

x = rnorm(sample.size, mean = sample.mean, sd = sample.sd)

if (mean(x) > H0.mean) {

count = count + 1

}

}

return (count / 1000)

}

# Simulate 100 times

sim.pvalue = replicate(100, p.value.sim())

# State at 5% significance

sum(sim.pvalue) / 100

## Exercise 3

# Define the data

n.x = 400

x.mean = 2635

x.sd = 365

n.y = 500

y.mean = 2887

y.sd = 412

# Get degrees of freedom

df.satterth.approx = function(n.x, n.y, s.x, s.y) {

num = ((s.x ^ 2 / n.x) + (s.y ^ 2 / n.y)) ^ 2

denom = (s.x ^ 4 / ((n.x ^ 2 \* (n.x - 1)))) + (s.y ^ 4/(n.y ^ 2 \* (n.y - 1)))

return (num / denom)

}

df.est = df.satterth.approx(n.x, n.y, x.sd, y.sd)

# Construct 95% CI for difference in mean

CI = n.x - n.y + c(-1, 1) \* qt(1 - 0.05 / 2, df = df.est) \* sqrt(x.sd ^ 2 / n.x + y.sd ^ 2 / n.y)

CI

# Perform a 5% level test

tstat = (n.x - n.y) / sqrt((x.sd ^ 2 / n.x) + (y.sd ^ 2 / n.y))

pval = (1 - pt(abs(tstat), df.est))

pval