Contents

[آزمون 1: آزمون برای مقایسه میانگین یک جامعه با عدد معلوم (واریانس جامعه باید معلوم باشد) 5](#_Toc30953002)

[آزمون 2: آزمون برای میانگین‏های دو جامعه (واریانس‏های دو جامعه باید معلوم و مساوی باشند) 11](#_Toc30953003)

[آزمون 3: آزمون برای میانگین‏های دو جامعه (واریانس دو جامعه باید معلوم و نامساوی باشند) 18](#_Toc30953004)

[آزمون 4: آزمون برای یک نسبت (توزیع دو جمله‏ای) 25](#_Toc30953005)

[آزمون 5: آزمون برای تساوی دو نسبت (توزیع دو جمله‏ای) 32](#_Toc30953006)

[آزمون 6: آزمون برای مقایسه دو مقدار شمارشی (توزیع پواسون) 37](#_Toc30953007)

[آزمون 7: آزمون برای میانگین یک جامعه (واریانس جامعه نامعلوم است) 41](#_Toc30953008)

[آزمون 8: آزمون برای میانگین‏های دو جامعه (واریانس‏های دو جامعه نامعلوم ولی مساوی می‏باشند) 47](#_Toc30953009)

[آزمون 9: آزمون برای میانگین‏های دو جامعه (واریانس‏های دو جامعه نامعلوم و نامساوی می‏باشند) 53](#_Toc30953010)

[آزمون 10: آزمون برای میانگین‏های دو جامعه (روش مقایسه‏های جفت شده) 59](#_Toc30953011)

[آزمون 11: آزمون برای ضریب رگرسیونی 64](#_Toc30953012)

[آزمون 12: آزمون یک ضریب همبستگی 71](#_Toc30953013)

[آزمون 13: آزمون یک ضریب همبستگی 77](#_Toc30953014)

[آزمون 14: آزمون برای دو ضریب همبستگی 84](#_Toc30953015)

[آزمون 15: آزمون برای واریانس یک جامعه 94](#_Toc30953016)

[آزمون 16: آزمون برای واریانس‏های دو جامعه (آزمون نسبت واریانس‏ها) 99](#_Toc30953017)

[آزمون 17: آزمون برای واریانس‏های دو جامعه (با مشاهدات همبسته) 104](#_Toc30953018)

[آزمون 18: آزمون هتلینگ برای دو سری از میانگین‏های جامعه 109](#_Toc30953019)

[آزمون 19: آزمون تشخیصی برای منشا یک نمونه تایی 115](#_Toc30953020)

[آزمون 20: آزمون دیکسون برای نقاط دورافتاده 119](#_Toc30953021)

[آزمون 21: آزمون برای نسبت‏های همبسته 126](#_Toc30953022)

[آزمون 22: آزمون (خی‏دو) برای یک مقدار فرض شده به عنوان واریانس جامعه 131](#_Toc30953023)

[آزمون 23: آزمون برای دو مقدار شمارشی (توزیع پواسون) 136](#_Toc30953024)

[آزمون 24: آزمون برای بررسی نرمال بودن یک جامعه 141](#_Toc30953025)

[آزمون 25: آزمون کوکران برای نقاط دورافتاده واریانس 145](#_Toc30953026)

[آزمون 26: آزمون وارون ویلکاکسون (آزمون ) 150](#_Toc30953027)

[آزمون 27: آزمون مجموع رتبه‏های ویلکاکسون – من – ویتنی برای دو جامعه 155](#_Toc30953028)

[آزمون 28: آزمون همبستگی رتبه‏ای اسپیرمن (برای مشاهدات جفت شده) 159](#_Toc30953029)

[آزمون 29: آزمون همبستگی رتبه‏ای کندال (برای مشاهدات جفت شده) 165](#_Toc30953030)

[آزمون 30: آزمون برای بررسی خطی بودن رگرسیون 171](#_Toc30953031)

# آزمون 1: آزمون [[1]](#footnote-2) برای مقایسه میانگین یک جامعه با عدد معلوم (واریانس جامعه باید معلوم باشد)

**هدف**

بررسی معنی‏دار بودن تفاوت بین میانگین مفروض یک جامعه و یک عدد معلوم (مثلاً میانگین یک نمونه ).

* فرضیه صفر ():
* فرضیه مقابل (): یا یا

نکته: این آزمون به لحاظ معلوم فرض شدن واریانس جامعه، معمولاً در موارد کمی در عمل اتفاق می‏افتد، ولی به لحاظ نظری در صورت وجود شرایط از اهمیت برخوردار است.

**پیش فرض‏ها**

1. واریانس جامعه[[2]](#footnote-3) باید معلوم باشد. (در صورتی که واریانس جامعه نامعلوم بود، می‏توان از آزمون برای میانگین یک جامعه استفاده کرد (آزمون 7).)
2. در صورتی که جامعه به صورت نرمال توزیع شده باشد، آزمون دقیق است. ولی اگر توزیع جامعه نرمال نباشد، آزمون باز هم تقریب خوبی به دست می‏دهد.

**روش**

از جامعه‏ای با میانگین مفروض و واریانس معلوم ، یک نمونه تصادفی با حجم گرفته شده و میانگین نمونه محاسبه می‏گردد. آماره آزمون تحت فرضیه صفر () برابر است با:

که می‏توان آماره آزمون را با استفاده از آزمون یک دنباله[[3]](#footnote-4) یا دو دنباله[[4]](#footnote-5)، با ناحیه بحرانی[[5]](#footnote-6) در سطح ، با توزیع نرمال استاندارد[[6]](#footnote-7) مقایسه کرد.

**مثال**

عملکرد گندم در استان خراسان رضوی تا سال آبی 97-1396، دارای توزیع نرمال با میانگین 95/2843 کیلوگرم در هکتار و انحراف استاندارد 01/587 کیلوگرم در هکتار بوده است. در پایان سال آبی 98-1397 به صورت تصادفی 50 مزرعه انتخاب و با میانگین‏گیری، عملکرد متوسط گندم 08/3462 کیلوگرم در هکتار به دست آمد. در سطح معنی‏داری ، آیا فرضیه صفر (میانگین عملکرد گندم استان خراسان رضوی در سال آبی 98-1397 تفاوتی با میانگین عملکرد گندم تا سال آبی 1396-97 ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های نمونه‏ تصادفی انتخاب شده از عملکرد گندم در سال آبی 98-1397، نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "../r/func.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "ggplot2", "DescTools")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if**(**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data <- **read.csv**(file = "../data/Data\_Test\_01.csv", header = TRUE)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |
| --- | --- | --- | --- |
| ## |  | Sample | Yield |
| ## | 1 | 1 | 3755 |
| ## | 2 | 2 | 4843 |
| ## | … | … | … |
| ## | 49 | 49 | 2471 |
| ## | 50 | 50 | 2802 |

**گام 2: بررسی محدودیت‏ها**

1- مقدار واریانس جامعه معلوم می‏باشد (012/587).

2- فرض شده است که عملکرد گندم در استان خراسان رضوی تا سال آبی 97-1396، دارای توزیع نرمال می‏باشد.

برای توصیف تغییرات نمونه‏ها، از نمودار چندک-چندک نرمال[[7]](#footnote-8) برای ارزیابی نرمال بودن توزیع نمونه‏ها، نمودار هیستوگرام[[8]](#footnote-9) برای بررسی تغییرات نمونه‏ها حول میانگین و نمودار جعبه‏ای[[9]](#footnote-10) برای بررسی تغییرات توزیع نمونه‏ها حول میانه استفاده شده است. برای این منظور به ترتیب از توابع **،**  و موجود در بسته استفاده شده است.

در نمودار جعبه‏ای، از جعبه‏ای برای نمایش فاصله بین چارک اول و سوم استفاده می‏شود و خطی در داخل جعبه، میانه (چارک دوم) را مشخص می‏کند. خارج از جعبه حداقل و حداکثر مقدار داده را نیز مشخص می‌نمایند. گهگاه نمونه‏های خارج از محدوده نیز به صورت نقاطی نشان داده می‏شوند.

*# visualizing sample data:*  
*# normality plot*  
nPlot = ggpubr**::ggqqplot**(data = data**$**Yield, ylab = "Yield (kg/ha)")  
  
*# histogram plot*  
hPlot = ggpubr**::gghistogram**(data = data**$**Yield, binwidth = 1000, add = "mean",   
 fill = "gray", xlab = "Yield (kg/ha)",  
 add.params = **list**(size = 1.2, linetype = 2))  
*# box plot*  
bPlot = ggpubr**::ggboxplot**(data = data**$**Yield,  
 add = **c**("mean"), fill = "gray", width = 0.5, xlab = "Year",  
 ylab = "Yield (kg/ha)", orientation = "horizontal") **+**  
 ggplot2**::geom\_hline**(yintercept = 2843.95, linetype = "dashed",  
 color = "red", size = 1.5) **+**  
 ggplot2**::annotate**(geom = "text", x = 1.3, y = 2700, label = "P. Mean",  
 color = "red", angle = 90)  
  
*# arrange on one page*  
ggpubr**::ggarrange**(ggpubr**::ggarrange**(hPlot, nPlot, nrow = 1, ncol = 2),  
 bPlot, nrow = 2, ncol = 1, heights = **c**(4, 3))



**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین میانگین مفروض یک جامعه و یک عدد معلوم، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و -مقدار آن به دست آید.

# method 1  
# setting initial parameter values: population  
mu.p = 2843.95  
sd.p = 587.01  
  
# setting initial parameter values: sample  
n.x = nrow(x = data)  
mean.x = mean(x = data$Yield)  
  
# test statistic  
Z = (mean.x - mu.p) / (sd.p / sqrt(n.x))  
  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
if (alternative != "two.sided") {  
 lower.critical = qnorm(p = (1 - alpha), lower.tail = FALSE)  
 upper.critical = qnorm(p = (1 - alpha), lower.tail = TRUE)  
} else {  
 lower.critical = qnorm(p = (1 - (alpha / 2)), lower.tail = FALSE)  
 upper.critical = qnorm(p = (1 - (alpha / 2)), lower.tail = TRUE)  
}  
  
p\_value = pnorm(q = Z, lower.tail = TRUE)  
p\_value = switch(EXPR = alternative,  
 two.sided = 2 \* min(p\_value, 1 - p\_value),  
 less = p\_value,  
 greater = 1 - p\_value)  
  
**cat**("Test Statistic: ", Z, "\n")  
**cat**("Critical Values: ", **c**(**-**lower.critical, upper.critical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: 7.4460  
## Critical Values: -1.96 1.96  
## P-value: 9.6217e-14

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع نرمال استاندارد با استفاده از تابع در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "norm", statistic = Z, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله ، مقدار بحرانی در برابر با 96/1 است (جدول شماره 1). مقدار محاسبه شده آماره آزمون تحت فرضیه صفر ()، بیشتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنی است که تفاوت معنی‏داری بین عملکرد گندم در سال آبی 98-1397 با میانگین عملکرد گندم تا سال آبی 97-1396 در سطح خطای 5 درصد وجود دارد.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین میانگین مفروض یک جامعه و یک عدد معلوم، از تابع موجود در بسته استفاده شده است.

# method 2  
# setting initial parameter values: population  
mu.p = 2843.95  
sd.p = 587.01  
  
# setting initial parameter values: sample  
x = data$Yield  
  
# use "ZTest" function from "DescTools" package  
DescTools::ZTest(x = x, mu = mu.p, sd\_pop = sd.p,   
 alternative = "two.sided", conf.level = 0.95)

## One Sample z-test  
## data: x  
## z = 7.446, Std. Dev. Population = 587.01, p-value = 9.622e-14  
## alternative hypothesis: true mean is not equal to 2843.95  
## 95 percent confidence interval:  
## 3299.373 3624.787  
## sample estimates:  
## mean of x: 3462.08

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 2: آزمون [[10]](#footnote-11) برای مقایسه میانگین‏های دو جامعه (واریانس‏های دو جامعه باید معلوم و مساوی باشند)

**هدف**

بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های معلوم و مساوی.

* فرضیه صفر ():
* فرضیه مقابل (): یا یا

نکته: این آزمون به لحاظ معلوم فرض شدن واریانس دو جامعه، معمولاً در موارد کمی در عمل اتفاق می‏افتد، ولی به لحاظ نظری در صورت وجود شرایط از اهمیت برخوردار است.

**پیش فرض‏ها**

1. واریانس هر دو جامعه باید معلوم و با هم مساوی باشند. (در صورتی که واریانس دو جامعه نامعلوم بود، می‏توان از آزمون برای میانگین‏های دو جامعه استفاده کرد (آزمون 8).)
2. در صورتی که هر دو جامعه به صورت نرمال توزیع شده باشند، آزمون دقیق است. ولی اگر توزیع دو جامعه نرمال نباشد، آزمون باز هم تقریب خوبی به دست می‏دهد.

**روش**

از دو جامعه‏ با میانگین‏های و ، نمونه‏های تصادفی مستقل با حجم‏های و گرفته شده، و میانگین نمونه‏ها ( و ) محاسبه می‏گردد. آماره آزمون تحت فرضیه صفر () برابر است با:

که می‏توان آماره آزمون را با استفاده از آزمون یک دنباله یا دو دنباله، با ناحیه بحرانی در سطح ، با توزیع نرمال استاندارد مقایسه کرد.

**مثال**

کارخانه‏‏ای دارای دو خط تولید نوار آبیاری پلاکدار (تیپ) می‏باشد. خط تولید شماره یک از پلاک‏های خارجی، و خط تولید شماره دو از پلاک‏های ایرانی با مشخصات مشابه، برای تولید نوار تیپ پلاکدار استفاده می‏کنند. برای بررسی عملکرد این دو نوار آبیاری، به صورت تصادفی 40 مزرعه گندم که از تیپ نوع و 50 مزرعه گندم که از تیپ نوع استفاده کرده‏اند انتخاب با میانگین‏گیری، عملکرد متوسط گندم برای هر کدام به ترتیب 35/3319 و 82/3030 کیلوگرم در هکتار به دست آمد. همچنین انحراف استاندارد هر دو جامعه با هم برابر و 074/616 کیلوگرم در هکتار است. در سطح معنی‏داری ، آیا فرضیه صفر (تفاوتی بین میانگین عملکرد گندم در مزارعی که از تیپ نوع استفاده کرده‏اند نسبت به مزارعی که از تیپ نوع استفاده کرده‏اند وجود ندارد) رد خواهد شد (با فرض اینکه همه شرایط یکسان باشد)؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های نمونه تصادفی انتخاب شده از مزارعی که از تیپ و تیپ استفاده کرده‏اند، نمایش داده شده است.

# remove all variable  
rm(list = ls())  
  
# load required functions  
source(file = "func.R")  
  
# load required packages  
wants = c("dplyr", "psych", "pacman", "ggpubr", "DescTools", "miscor")  
has = wants %in% rownames(x = installed.packages())  
if (any(!has)) install.packages(wants[!has])  
pacman::p\_load(char = wants)  
  
# load required data  
data = read.csv(file = "data/Data\_Test\_02.csv", header = TRUE)  
  
# show data  
psych::headTail(x = data, top = 2, bottom = 2)

|  |  |  |  |
| --- | --- | --- | --- |
| ## |  | Tape | Yield |
| ## | 1 | A | 3462 |
| ## | 2 | A | 3473 |
| ## | … | … | … |
| ## | 89 | B | 3952 |
| ## | 90 | B | 3190 |

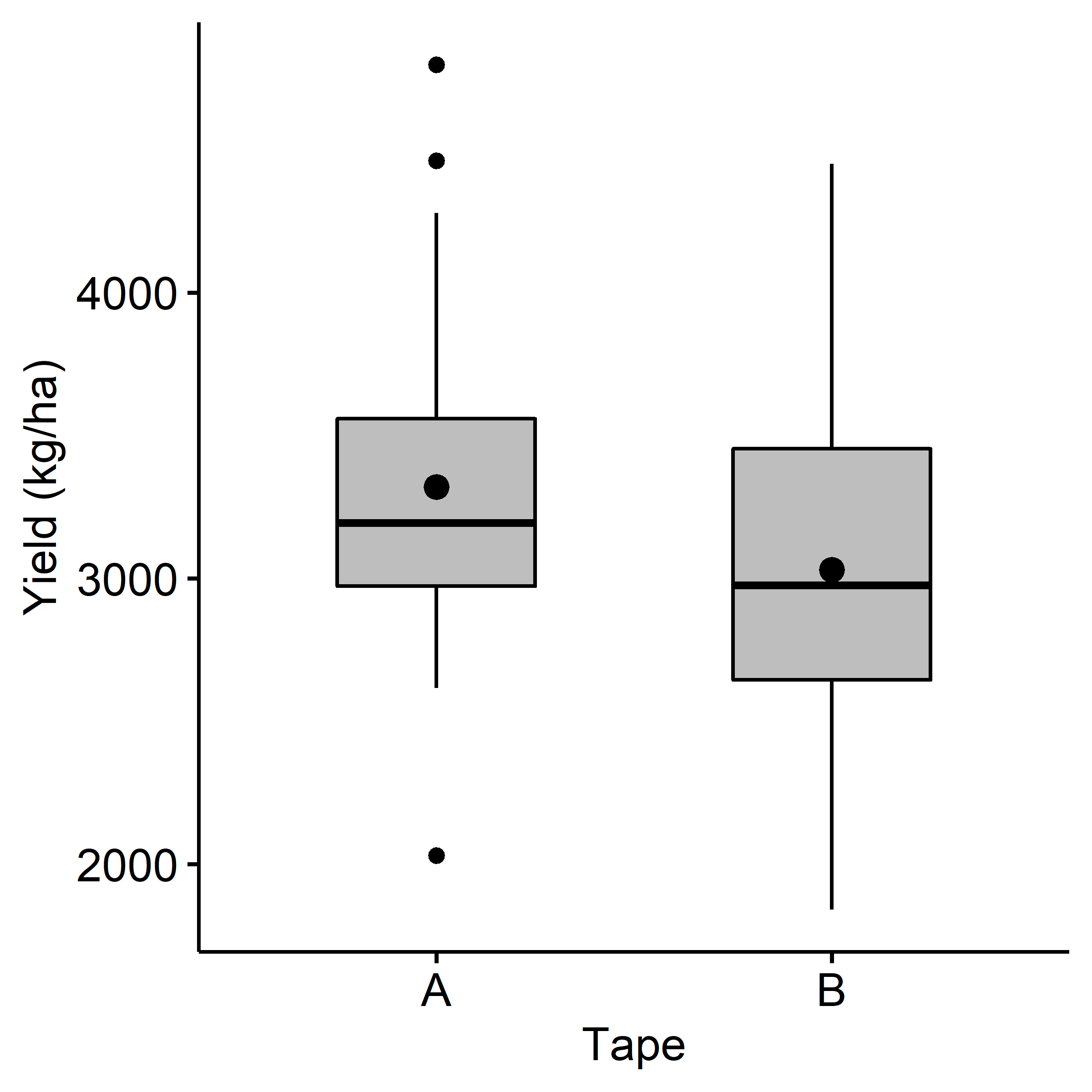
**گام 2: بررسی محدودیت‏ها**

1- فرض شده است که واریانس هر دو جامعه معلوم و با هم برابر است (0742/616).

2- فرض شده است که عملکرد گندم هر دو جامعه به صورت نرمال توزیع شده است.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها با استفاده از تابع موجود در بسته رسم شده است.

# visualizing samples with box plots  
ggpubr::ggboxplot(data = data, x = "Tape", y = "Yield",   
 add = c("mean"), fill = "gray", width = 0.5,   
 xlab = "Tape", ylab = "Yield (kg/ha)")



**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های معلوم و مساوی، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و -مقدار آن به دست آید.

# method 1  
# setting initial parameter values: population  
sd.p = 616.074  
  
# setting initial parameter values: sample  
sample.A = data %>% filter(Tape == "A") %>% select(Yield) %>% unlist()  
n.A = length(x = sample.A)  
mean.A = mean(x = sample.A)  
  
sample.B = data %>% filter(Tape == "B") %>% select(Yield) %>% unlist()  
n.B = length(x = sample.B)  
mean.B = mean(x = sample.B)  
  
# test statistic  
Z = (mean.A - mean.B) / (sd.p \* sqrt((1/n.A) + (1/n.B)))  
  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
if (alternative != "two.sided") {  
 lower.critical = qnorm(p = (1 - alpha), lower.tail = FALSE)  
 upper.critical = qnorm(p = (1 - alpha), lower.tail = TRUE)  
} else {  
 lower.critical = qnorm(p = (1 - (alpha / 2)), lower.tail = FALSE)  
 upper.critical = qnorm(p = (1 - (alpha / 2)), lower.tail = TRUE)  
}  
  
p\_value = pnorm(q = Z, lower.tail = TRUE)  
p\_value = switch(EXPR = alternative,  
 two.sided = 2 \* min(p\_value, 1 - p\_value),  
 less = p\_value,  
 greater = 1 - p\_value)  
cat("Test Statistic:", Z, "\n")  
cat("Critical Values:", c(lower.critical, upper.critical), "\n")  
cat("P-value:", p\_value, "\n")

## Test Statistic: 2.2078  
## Critical Values: -1.96 1.96  
## P-value: 0.0272

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع نرمال استاندارد با استفاده از تابع در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "norm", statistic = Z, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله ، مقدار بحرانی در برابر با 96/1 است (جدول شماره 1). مقدار محاسبه شده آماره آزمون تحت فرضیه صفر بیشتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که فرضیه یکسان بودن عملکرد گندم در مزارعی که از تیپ نوع و مزارعی که از تیپ نوع استفاده کرده‏اند، رد می‏شود.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های معلوم و مساوی، از تابع موجود در بسته استفاده شده است.

# method 2  
# setting initial parameter values: population  
sd.p = 616.074  
  
# setting initial parameter values: sample  
sample.A = data %>% filter(Tape == "A") %>% select(Yield) %>% unlist()  
sample.B = data %>% filter(Tape == "B") %>% select(Yield) %>% unlist()  
  
# use "ZTest" function from "DescTools" package  
ZTest(x = sample.A, y = sample.B, alternative = "two.sided",  
 mu = 0, sd\_pop = sd.p, conf.level = 0.95)

## Two Sample z-test  
## data: sample.A and sample.B  
## z = 2.2078, Std. Dev. Population = 616.07, p-value = 0.0273  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 32.3843 544.6757  
## sample estimates:  
## mean of x mean of y  
## 3319.35 3030.82

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشد. این آزمون را می‏توان به صورت یک دنباله، با ناحیه بحرانی در سطح ، با توزیع نرمال استاندارد نیز مقایسه کرد.

# method 2  
# setting initial parameter values: population  
sd.p = 616.074  
  
# setting initial parameter values: sample  
sample.A = data %>% filter(Tape == "A") %>% select(Yield) %>% unlist()  
sample.B = data %>% filter(Tape == "B") %>% select(Yield) %>% unlist()  
  
# use "ZTest" function from "DescTools" package  
ZTest(x = sample.A, y = sample.B, alternative = "greater",  
 mu = 0, sd\_pop = sd.p, conf.level = 0.95)

## Two Sample z-test  
## data: sample.A and sample.B  
## z = 2.2078, Std. Dev. Population = 616.07, p-value = 0.0136  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 73.5658 Inf  
## sample estimates:  
## mean of x mean of y  
## 3319.35 3030.82

در آزمون یک دنباله ، مقدار بحرانی در برابر با 64/1 است (جدول شماره 1). مقدار محاسبه شده آماره آزمون تحت فرضیه صفر ، بیشتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که عملکرد گندم در مزارعی که از تیپ نوع استفاده کرده‏اند بیشتر از مزارعی است که از تیپ نوع استفاده کرده‏اند و این تفاوت معنی‏دار است.

# آزمون 3: آزمون [[11]](#footnote-12) برای مقایسه میانگین‏های دو جامعه (واریانس دو جامعه باید معلوم و نامساوی باشند)

**هدف**

بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های معلوم و نامساوی.

* فرضیه صفر ():
* فرضیه مقابل (): یا یا

**پیش فرض‏ها**

1. واریانس هر دو جامعه باید معلوم باشد. (در صورتی که واریانس دو جامعه نامعلوم بود، می‏توان از آزمون برای میانگین‏های دو جامعه استفاده کرد (آزمون 9).)
2. در صورتی که هر دو جامعه به صورت نرمال توزیع شده باشند، آزمون دقیق است. ولی اگر توزیع دو جامعه نرمال نباشد، آزمون باز هم تقریب خوبی به دست می‏دهد.

**روش**

از دو جامعه‏ با میانگین‏های و و واریانس‏های و ، نمونه‏های تصادفی مستقل با حجم‏های و گرفته شده، و میانگین نمونه‏ها ( و ) محاسبه می‏گردد. آماره آزمون تحت فرضیه صفر برابر است با:

که می‏توان آماره آزمون را با استفاده از آزمون یک دنباله یا دو دنباله، با ناحیه بحرانی در سطح ، با توزیع نرمال استاندارد مقایسه کرد.

**مثال**

برای بررسی عملکرد دو رقم زودرس و پر بازده گندم (سیروان و پیشگام) در استان خراسان رضوی، به صورت تصادفی 55 مزرعه گندم که با رقم سیروان و 50 مزرعه گندم که با رقم پیشگام کشت شده بودند انتخاب و با میانگین‏گیری، عملکرد متوسط گندم برای هر کدام به ترتیب 98/3252 و 20/2979 کیلوگرم در هکتار به دست آمد. همچنین انحراف استاندارد دو جامعه ارقام سیروان و پیشگام، به ترتیب 50/694 و 36/864 کیلوگرم در هکتار است. در سطح معنی‏داری ، آیا فرضیه صفر (تفاوتی بین میانگین عملکرد گندم برای دو رقم سیروان و پیشگام در استان خراسان رضوی وجود ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های نمونه تصادفی انتخاب شده، از مزارعی که از دو رقم سیروان و پیشگام استفاده کرده‏اند، نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "func.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "BSDA")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_03.csv", header = TRUE)  
  
# show data  
**headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |
| --- | --- | --- | --- |
| ## |  | Cultivar | Yield |
| ## | 1 | Sirvan | 3420 |
| ## | 2 | Sirvan | 4382 |
| ## | … | … | … |
| ## | 104 | Pishgam | 3723 |
| ## | 105 | Pishgam | 3710 |

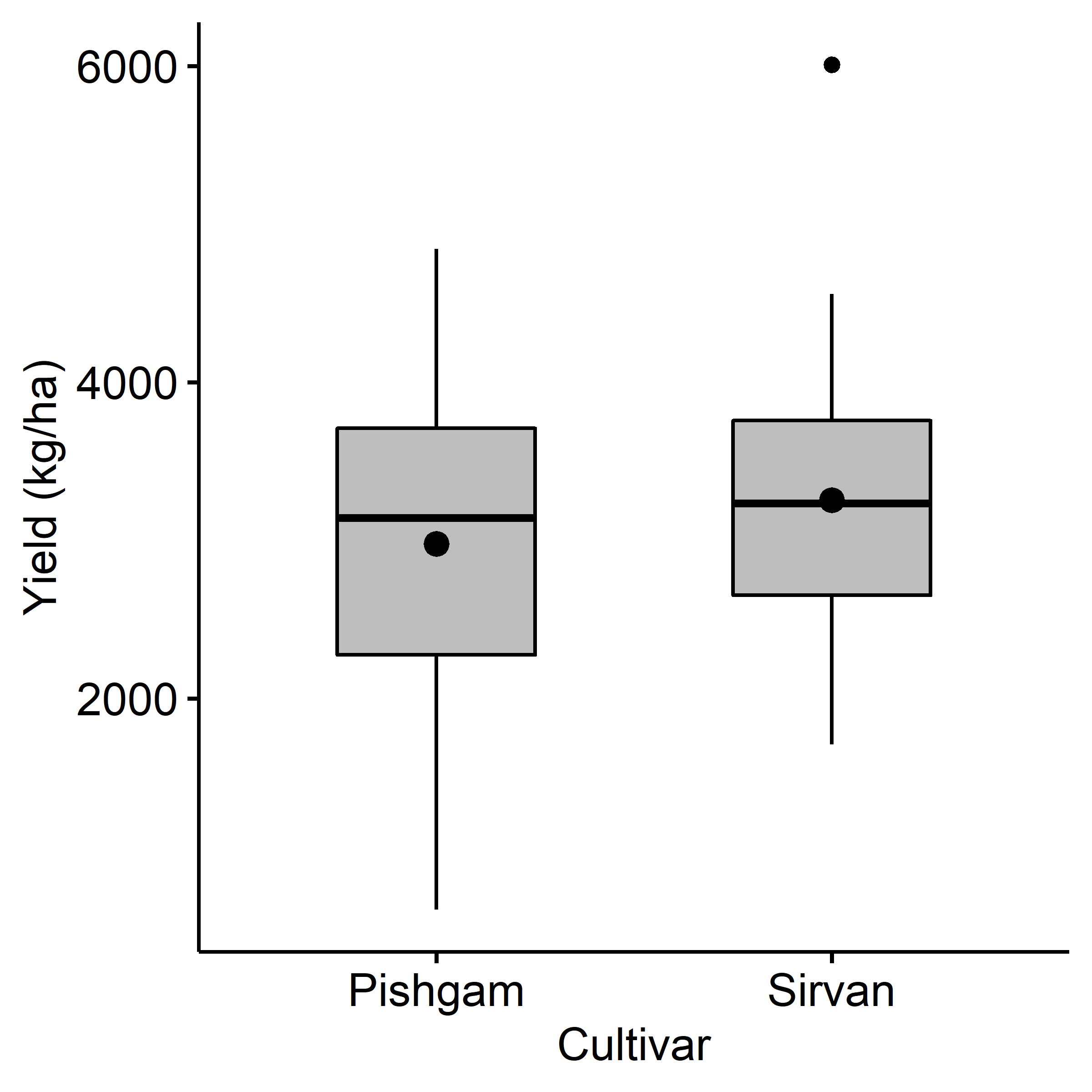
**گام 2: بررسی محدودیت‏ها**

1- فرض شده است که واریانس هر دو جامعه معلوم است (رقم سیروان: 502/694، رقم پیشگام: 362/864).

2- فرض شده است که عملکرد گندم هر دو جامعه به صورت نرمال توزیع شده است.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها با استفاده از تابع موجود در بسته رسم شده است.

# visualizing samples with box plots  
ggpubr::ggboxplot(data = data, x = "Cultivar", y = "Yield",   
 add = c("mean"), fill = "gray", width = 0.5,   
 xlab = "Cultivar", ylab = "Yield (kg/ha)")



**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های معلوم و نامساوی، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و -مقدار آن به دست آید.

# method 1  
# setting initial parameter values: population  
sd.p.A = 694.50 # Sirvan  
sd.p.B = 864.36 # Pishgam  
  
# setting initial parameter values: sample  
sample.A = data %>% filter(Cultivar == "Sirvan") %>% select(Yield) %>% unlist()  
n.A = length(x = sample.A)  
mean.A = mean(x = sample.A)  
  
sample.B = data %>% filter(Cultivar == "Pishgam") %>% select(Yield) %>% unlist()  
n.B = length(x = sample.B)  
mean.B = mean(x = sample.B)  
  
# test statistic  
Z = (mean.A - mean.B) / (sqrt(((sd.p.A ^ 2)/n.A) + ((sd.p.B ^ 2)/n.B)))  
  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
if (alternative != "two.sided") {  
 lower.critical = qnorm(p = (1 - alpha), lower.tail = FALSE)  
 upper.critical = qnorm(p = (1 - alpha), lower.tail = TRUE)  
} else {  
 lower.critical = qnorm(p = (1 - (alpha / 2)), lower.tail = FALSE)  
 upper.critical = qnorm(p = (1 - (alpha / 2)), lower.tail = TRUE)  
}  
  
p\_value = pnorm(q = Z, lower.tail = TRUE)  
p\_value = switch(EXPR = alternative,  
 two.sided = 2 \* min(p\_value, 1 - p\_value),  
 less = p\_value,  
 greater = 1 - p\_value)  
cat("Test Statistic:", Z, "\n")  
cat("Critical Values:", c(lower.critical, upper.critical), "\n")  
cat("P-value:", p\_value, "\n")

## Test Statistic: 1.7770  
## Critical Values: -1.96 1.96  
## P-value: 0.0754

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع نرمال استاندارد با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "norm", statistic = Z, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله ، مقدار بحرانی در برابر با 96/1 است (جدول شماره 1). مقدار محاسبه شده آماره آزمون تحت فرضیه صفر کمتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که عملکرد گندم برای دو رقم سیروان و پیشگام در استان خراسان رضوی تفاوت معنی‏داری با یکدیگر ندارند.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های معلوم و نامساوی، از تابع موجود در بسته استفاده شده است.

# method 2  
# setting initial parameter values: population  
sd.p.A = 694.50 # Sirvan  
sd.p.B = 864.36 # Pishgam  
  
# setting initial parameter values: sample  
sample.A = data **%>%** **filter**(Cultivar **==** "Sirvan") **%>%** **select**(Yield) **%>%** **unlist**()  
sample.B = data **%>%** **filter**(Cultivar **==** "Pishgam") **%>%** **select**(Yield) **%>%** **unlist**()  
  
# use "z.test" function from "BSDA" package  
**z.test**(x = sample.A, y = sample.B, alternative = "two.sided", mu = 0,   
 sigma.x = sd.p.A, sigma.y = sd.p.B, conf.level = 0.95)

## Two-sample z-Test  
## data: sample.A and sample.B  
## z = 1.778, p-value = 0.0754  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -28.0268 575.5905  
## sample estimates:  
## mean of x mean of y  
## 3252.98 2979.20

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشد. این آزمون را می‏توان به صورت یک دنباله، با ناحیه بحرانی در سطح ، با توزیع نرمال استاندارد نیز مقایسه کرد.

# method 2  
# setting initial parameter values: population  
sd.p.A = 694.50 # Sirvan  
sd.p.B = 864.36 # Pishgam  
  
# setting initial parameter values: sample  
sample.A = data **%>%** **filter**(Cultivar **==** "Sirvan") **%>%** **select**(Yield) **%>%** **unlist**()  
sample.B = data **%>%** **filter**(Cultivar **==** "Pishgam") **%>%** **select**(Yield) **%>%** **unlist**()  
  
# use "z.test" function from "BSDA" package  
**z.test**(x = sample.A, y = sample.B, alternative = "greater", mu = 0,   
 sigma.x = sd.p.A, sigma.y = sd.p.B, conf.level = 0.95)

## Two-sample z-Test  
## data: sample.A and sample.B  
## z = 1.778, p-value = 0.0377  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 20.4960 NA  
## sample estimates:  
## mean of x mean of y  
## 3252.98 2979.20

در آزمون یک دنباله ، مقدار بحرانی در برابر با 64/1 است (جدول شماره 1). مقدار محاسبه شده آماره آزمون تحت فرضیه صفر ، بیشتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که عملکرد گندم برای رقم سیروان بیشتر از عملکرد گندم برای رقم پیشگام در استان خراسان رضوی می‏باشد و این تفاوت معنی‏دار است.

# آزمون 4: آزمون برای مقایسه یک نسبت[[12]](#footnote-13) در جامعه با عدد معلوم (توزیع دو جمله‏ای[[13]](#footnote-14))

**هدف**

بررسی معنی‏دار بودن تفاوت بین یک نسبت مفروض و یک نسبت مشاهده شده .

* فرضیه صفر ():
* فرضیه مقابل (): یا یا

**پیش فرض‏ها**

1. این آزمون تقریبی بوده و فرض می‏کند که تعداد مشاهدات نمونه به اندازه کافی بزرگ است (یعنی ). این فرض توجیه آن است که بتوان از تقریب نرمال بجای دو جمله‏ای استفاده کرد.

**روش**

یک نمونه تصادفی عضوی از جامعه‏ای که در آن فرض بر این است که نسبت متعلق به یک گروه خاص است، انتخاب می‏گردد. سپس نسبت عناصر متعلق به آن گروه خاص در نمونه تصادفی، محاسبه می‏گردد. آماره آزمون تحت فرضیه صفر برابر است با:

که می‏توان آماره آزمون را با استفاده از آزمون یک دنباله یا دو دنباله، با ناحیه بحرانی در سطح ، با توزیع نرمال استاندارد مقایسه کرد.

**مثال**

شرکت پاکان بذر اصفهان مدعی است که درصد قوه نامیه بذر اسطوخودوس این شرکت 90 درصد می‏باشد. به منظور بررسی این ادعا، با استفاده از روش کیسه‏ای قوه نامیه بذرها مورد ارزیابی قرار می‏گیرد. در این روش بر روی ظرف مخصوص کاغذ فیلتر که لبه‏های آن از دو طرف ظرف بیرون آمده قرار داده می‏شود. سپس به صورت کاملا ً تصادفی 100 دانه بذر اسطوخودوس انتخاب و بر روی این کاغذ مرطوب پخش کرده و سپس روی آن‏ها با لبه‏های به بیرون کشیده شده پوشیده می‏شود. سپس درب ظرف با دریچه مخصوص بسته می‏شود. پس از اتمام مدت تعیین شده برای جوانه‏زنی، تعداد بذرهای جوانه زده 82 عدد شمارش می‏شود. در سطح معنی‏داری ، آیا فرضیه صفر (تفاوت معنی‏داری بین درصد قوه نامیه بذر اسطوخودوس نمونه‏ تصادفی (82 درصد) نسبت به درصد قوه نامیه بذر اسطوخودوس اعلام شده توسط شرکت پاکان بذر اصفهان (90 درصد) وجود ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏های مورد نیاز فراخوانی می‏شود.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "func.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)

**گام 2: بررسی محدودیت‏ها**

تعداد مشاهدات نمونه به اندازه کافی بزرگ است (100 عدد).

**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین یک نسبت مفروض و یک نسبت مشاهده شده، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و -مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
P0 = 0.9  
  
x = 82  
n = 100  
P = x **/** n  
  
# test statistic  
Z = (P **-** P0) **/** **sqrt**((P0 **\*** (1 **-** P0)) **/** n)  
  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
**if** (alternative **!=** "two.sided") {  
 lower.critical = **qnorm**(p = (1 **-** alpha), lower.tail = FALSE)  
 upper.critical = **qnorm**(p = (1 **-** alpha), lower.tail = TRUE)  
} **else** {  
 lower.critical = **qnorm**(p = (1 **-** (alpha **/** 2)), lower.tail = FALSE)  
 upper.critical = **qnorm**(p = (1 **-** (alpha **/** 2)), lower.tail = TRUE)  
}  
  
p\_value = **pnorm**(q = Z, lower.tail = TRUE)  
p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value)  
  
**cat**("Test Statistic:", Z, "\n")  
**cat**("Critical Values:", **c**(lower.critical, upper.critical), "\n")  
**cat**("P-value:", p\_value, "\n")

## Test Statistic: -2.6667  
## Critical Values: -1.96 1.96  
## P-value: 0.0077

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع نرمال استاندارد با استفاده از تابع در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "norm", statistic = Z, alpha = 0.05,  
 alternative = "two.sided")



در آزمون دو دنباله ، مقدار بحرانی در برابر با 96/1 است (جدول شماره 1). قدر مطلق مقدار محاسبه شده آماره آزمون تحت فرضیه صفر بزرگتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که ادعای شرکت پاکان بذر اصفهان مبنی بر اینکه درصد قوه نامیه بذر اسطوخودوس این شرکت 90 درصد است، رد خواهد شد.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین یک نسبت مفروض و یک نسبت مشاهده شده، از تابع موجود در بسته استفاده شده است.

# method 2  
# setting initial parameter values  
P0 = 0.9  
  
x = 82  
n = 100  
  
# use "prop.test" function from "stats" package  
**prop.test**(x = x, n = n, p = P0, alternative = "two.sided",   
 conf.level = 0.95, correct = FALSE)

## 1-sample proportions test without continuity correction  
## data: x out of n, null probability P0  
## X-squared = 7.1111, df = 1, p-value = 0.007661  
## alternative hypothesis: true p is not equal to 0.9  
## 95 percent confidence interval:  
## 0.7333 0.8830  
## sample estimates:  
## p: 0.82

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشد. ذکر این نکته ضروری است که آماره آزمون در این روش از جذر **X-squared** به دست می‏آید.

# آزمون 5: آزمون برای تساوی دو نسبت[[14]](#footnote-15) (توزیع دو جمله‏ای[[15]](#footnote-16))

**هدف**

بررسی فرضیه تساوی دو نسبت و از عناصر دو جامعه، بر اساس نمونه‏های انتخاب شده از هر دو جامعه.

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. این آزمون تقریبی بوده و فرض می‏کند که تعداد مشاهدات نمونه به اندازه کافی بزرگ است (یعنی ). این فرض توجیه آن است که بتوان از تقریب نرمال بجای دو جمله‏ای استفاده کرد.

**روش**

با فرض اینکه و نسبت‏های یک صفت مشخص در دو جامعه باشند، نمونه‏های تصادفی به حجم و از جوامع انتخاب و نسبت‏های متناظر و آن‏ها محاسبه می‏گردد. آماره آزمون برابر است با:

با در نظر گرفتن فرضیه صفر ، آماره تقریبا به صورت نرمال استاندارد توزیع شده است. در نتیجه می‏توان آماره آزمون را با استفاده از آزمون یک دنباله یا دو دنباله، با ناحیه بحرانی در سطح ، با توزیع نرمال استاندارد مقایسه کرد.

**مثال**

نرخ تعداد روزهای بارانی (روزهایی با بارندگی بیشتر از 1/0 میلیمتر) در سال 2018 در ایستگاه خرم‏آباد، 230/0 و در ایستگاه مشهد، 132/0 به دست آمده است. در سطح معنی‏داری ، آیا فرضیه صفر (تفاوت معنی‏داری بین نرخ تعداد روزهای بارانی در سال 2018 در دو ایستگاه خرم‏آباد و مشهد وجود ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی می‏شوند. سپس وقوع (1) و عدم وقوع (0) بارش برای هر روز ایستگاه‏های خرم آباد (A) و مشهد (B) محاسبه میگردد. در ادامه داده‏های سال 2018 به عنوان نمونه تعیین می‏گردند. در انتها نیز قسمتی از داده‏های نمونه نمایش داده می‏شود.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if**(**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_05.csv", header = TRUE)  
  
# Prepare Data:  
data = data **%>%** **mutate**(A\_occ = **ifelse**(test = A\_prec **>=** 0.1, yes = 1, no = 0),  
 B\_occ = **ifelse**(test = B\_prec **>=** 0.1, yes = 1, no = 0))  
  
data\_s = data **%>%** **filter**(Year **>=** 2018)  
  
# Show Data:  
psych**::headTail**(x = data\_s, top = 2, bottom = 2)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## |  | Year | Month | Day | A\_prec | B\_prec | A\_occ | B\_occ |
| ## | 1 | 2018 | 1 | 1 | 0 | 0 | 0 | 0 |
| ## | 2 | 2018 | 1 | 2 | 4.9 | 0 | 1 | 0 |
| ## | … | … | … | … | … | … | … | … |
| ## | 364 | 2018 | 12 | 30 | 0 | 0.3 | 0 | 1 |
| ## | 365 | 2018 | 12 | 31 | 0 | 0 | 0 | 0 |

**گام 2: بررسی محدودیت‏ها**

تعداد مشاهدات نمونه به اندازه کافی بزرگ است: برای به دست آوردن تعداد مشاهدات نمونه از تابع **nrow** استفاده شده است.

# assumption checking  
# 1. the number of observations in the sample is sufficiently large (i.e. n>=30).  
**cat**("Number of Observations:", **nrow** (x = data\_s), "\n")

## Number of Observations: 365

**گام 3: محاسبه آماره آزمون**

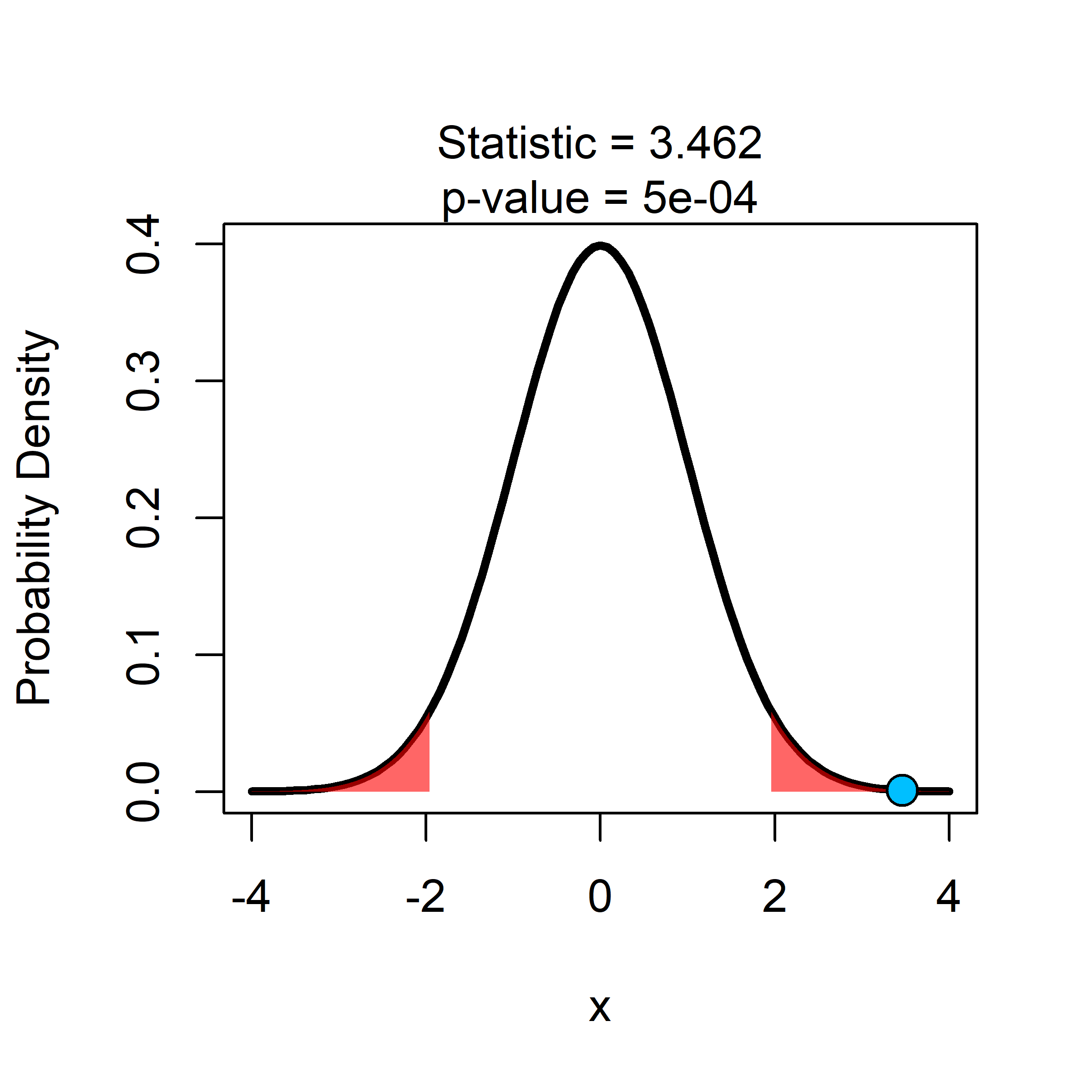
محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی فرضیه تساوی دو نسبت از عناصر دو جامعه، بر اساس نمونه‏های انتخاب شده از هر دو جامعه، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values: khorramabad station  
p\_A = **mean**(x = data\_s**$**A\_occ, na.rm = TRUE)  
x\_A = **sum**(data\_s**$**A\_occ, na.rm = TRUE)  
n\_A = **length**(x = data\_s**$**A\_occ)  
  
# setting initial parameter values: mashhad station  
p\_B = **mean**(x = data\_s**$**B\_occ, na.rm = TRUE)  
x\_B = **sum**(data\_s**$**B\_occ, na.rm = TRUE)  
n\_B = **length**(x = data\_s**$**B\_occ)  
  
# test statistic  
P = ((p\_A **\*** n\_A) **+** (p\_B **\*** n\_B)) **/** (n\_A **+** n\_B)  
Z = (p\_A **-** p\_B) **/** **sqrt**(P **\*** (1 **-** P) **\*** ((1 **/** n\_A) **+** (1 **/** n\_B)))  
  
# compute the critical values at 0.05 significance level (two tailed)  
alpha = 0.05  
two.sided = TRUE  
Zcritical = **qnorm**(p = (1 **-** alpha**/**2))  
  
# p-value (two tailed)  
p\_value = **pnorm**(q = Z, lower.tail = FALSE)  
**if** (two.sided) { p\_value = 2 **\*** p\_value  
 **if** (p\_value **>** 1) { p\_value = 2 **-** p\_value } }  
**cat**("Test Statistic: ", Z, "\n")  
**cat**("Critical Values: ", **c**(**-**Zcritical, Zcritical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: 3.4620  
## Critical Values: -1.96 1.96  
## P-value: 0.0005

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع نرمال استاندارد با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "norm", statistic = Z, alpha = 0.05,  
 alternative = "two.sided")



در آزمون دو دنباله ، مقدار بحرانی در برابر با 96/1 است (جدول شماره 1). مقدار محاسبه شده برای بیشتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که تفاوت معنی‏داری بین نرخ تعداد روزهای بارانی در سال 2018 در دو ایستگاه خرم‏آباد و مشهد وجود دارد.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین یک نسبت مفروض و یک نسبت مشاهده شده، از تابع **prop.test** موجود در بسته **stats** استفاده شده است.

# method 2  
# setting initial parameter values: khorramabad station  
p\_A = **mean**(x = data\_s**$**A\_occ, na.rm = TRUE)  
x\_A = **sum**(data\_s**$**A\_occ, na.rm = TRUE)  
n\_A = **length**(x = data\_s**$**A\_occ)  
# setting initial parameter values: mashhad station  
p\_B = **mean**(x = data\_s**$**B\_occ, na.rm = TRUE)  
x\_B = **sum**(data\_s**$**B\_occ, na.rm = TRUE)  
n\_B = **length**(x = data\_s**$**B\_occ)  
  
# use "prop.test" function from "stats" package  
**prop.test**(x = **c**(x\_A, x\_B), n = **c**(n\_A, n\_B), alternative = "two.sided",   
 conf.level = 0.95, correct = FALSE)

## 2- sample test for equality of proportions without continuity correction  
## data: c(x\_A, x\_B) out of c(n\_A, n\_B)  
## X-squared = 11.985, df = 1, p-value = 0.0005  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.0433 0.1540  
## sample estimates:  
## prop 1 prop 2  
## 0.2301 0.1315

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند. ذکر این نکته ضروری است که آماره آزمون در این روش از جذر **X-squared** به دست می‏آید.

# آزمون 6: آزمون برای مقایسه دو مقدار شمارشی[[16]](#footnote-17) (توزیع پواسون[[17]](#footnote-18))

**هدف**

بررسی معنی‏دار بودن تفاوت بین دو مقدار شمارشی.

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. این آزمون تقریبی بوده و فرض می‏کند که تعداد مقادیر شمارشی به اندازه کافی بزرگ است (یعنی ). این فرض توجیه آن است که بتوان از تقریب نرمال بجای پواسون استفاده کرد.

**روش**

با فرض اینکه و دو مقدار شمارشی مشاهده شده در مدت زمان و باشند، دو فراوانی متوسط آن‏ها به ترتیب برابر است با و . برای آزمون فرضیه برابری فراوانی‏های متوسط، از آماره آزمون زیر استفاده می‏شود:

می‏توان آماره آزمون را با استفاده از آزمون یک دنباله یا دو دنباله، با ناحیه بحرانی در سطح ، با توزیع نرمال استاندارد مقایسه کرد.

**مثال**

تعداد روزهای تابستانی (دمای حداکثر روزانه بیشتر از 25 درجه سانتی‏گراد باشد) در سال، یکی از شاخص‏های معتبر حدی دما می‏باشد. این شاخص برای دو ایستگاه هواشناسی سینوپتیک کرمان و زاهدان در بازه زمانی سال‏های 1979 تا 2006 محاسبه شده است. به صورت کاملا تصادفی برای ایستگاه کرمان 15 سال و برای ایستگاه زاهدان 12 سال از این شاخص انتخاب می‏گردد. مجموع تعداد روزهای تابستانی برای ایستگاه کرمان و مشهد به ترتیب 2795 و 2518 روز و فراوانی‏ متوسط رخداد این شاخص در بازه انتخاب شده به ترتیب 33/186 و 83/209 روز در سال به دست می‏آید. در سطح معنی‏داری ، آیا فرضیه صفر (تفاوت معنی‏داری بین فراوانی‏ متوسط رخداد شاخص تعداد روزهای تابستانی در دو ایستگاه کرمان و زاهدان وجود ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی می‏شوند و در انتها نیز قسمتی از داده‏ها نمایش داده می‏شود.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if**(**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_06.csv", header = TRUE)  
  
# Prepare Data:  
**set.seed**(33)  
A = data **%>%** **select**(Kerman) **%>%** **sample\_n**(size = 15) **%>%** **unlist**()  
B = data **%>%** **select**(Zahedan) **%>%** **sample\_n**(size = 12) **%>%** **unlist**()  
  
# Show Data:  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Year | Kerman | Zahedan |
| ## | 1 | 1979 | 175 | 203 |
| ## | 2 | 1980 | 184 | 214 |
| ## | … | … | … | … |
| ## | 27 | 2005 | 196 | 223 |
| ## | 28 | 2006 | 200 | 229 |

**گام 2: بررسی محدودیت‏ها**

تعداد مقادیر شمارشی به اندازه کافی بزرگ است: برای به دست آوردن تعداد روزهای تابستانی از تابع **sum** استفاده شده است.

# assumption checking  
# 1. the counts are large enough (i.e. n>=30)  
**cat**("Number of Counts (Kerman):", **sum**(A), "\n")  
**cat**("Number of Counts (Zahedan):", **sum**(B), "\n")

## Number of Counts (Kerman): 2795  
## Number of Counts (Zahedan): 2518

**گام 3: محاسبه آماره آزمون**

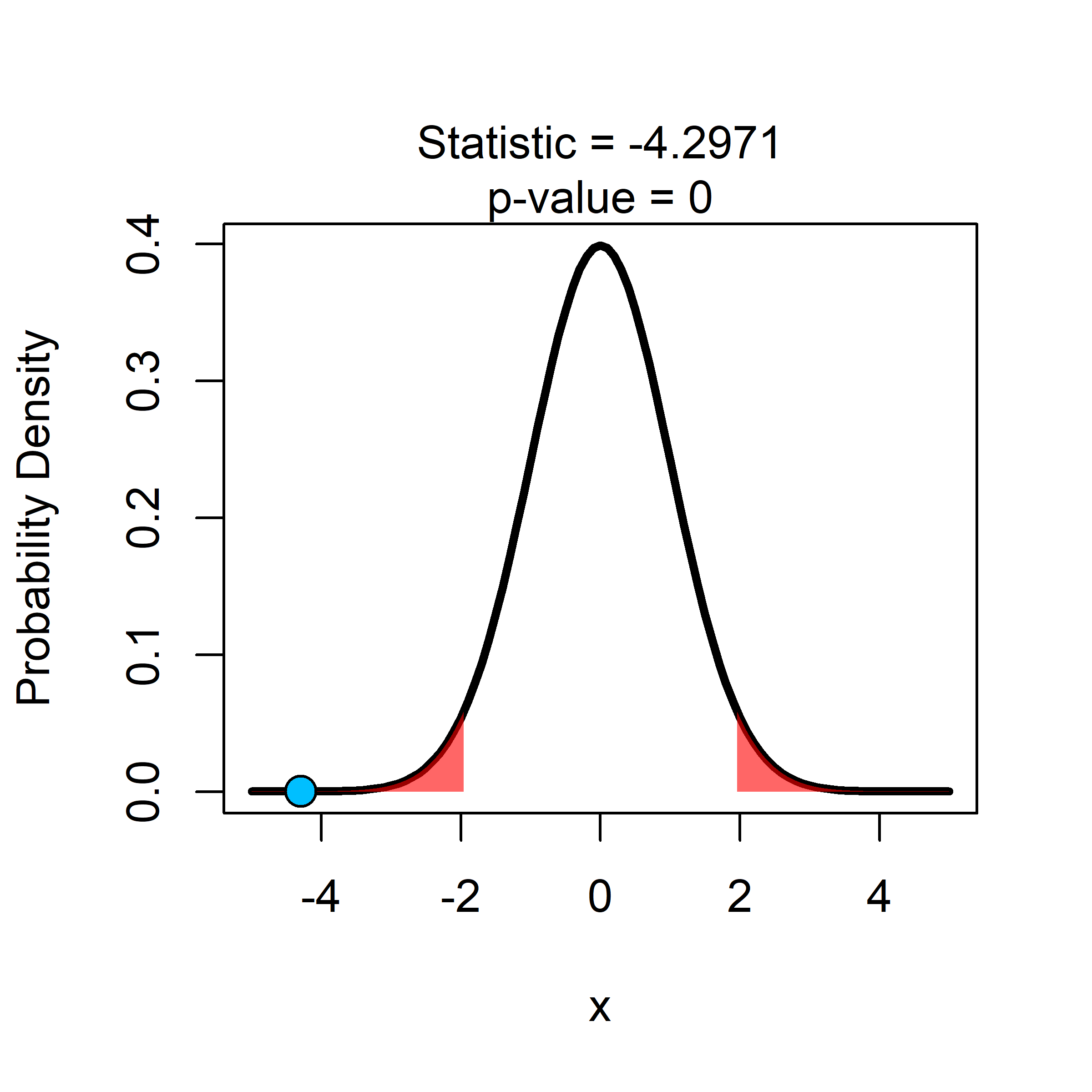
برای بررسی معنی‏دار بودن تفاوت بین دو مقدار شمارشی، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# setting initial parameter values: kerman station   
n\_A = **sum**(A)  
t\_A = **length**(x = A)  
R\_A = n\_A **/** t\_A  
  
# setting initial parameter values: zahedan station  
n\_B = **sum**(B)  
t\_B = **length**(x = B)  
R\_B = n\_B **/** t\_B  
  
# test statistic  
Z = (R\_A **-** R\_B) **/** **sqrt**((R\_A **/** t\_A) **+** (R\_B **/** t\_B))  
  
# compute the critical values at 0.05 significance level (two tailed)  
alpha = 0.05  
two.sided = TRUE  
Zcritical = **qnorm**(p = (1 **-** alpha**/**2))  
  
# p-value (two tailed)  
p\_value = **pnorm**(q = Z, lower.tail = FALSE)  
  
**if** (two.sided) { p\_value = 2 **\*** p\_value  
 **if** (p\_value **>** 1) { p\_value = 2 **-** p\_value } }  
**cat**("Test Statistic: ", Z, "\n")  
**cat**("Critical Values: ", **c**(**-**Zcritical, Zcritical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: -4.2971  
## Critical Values: -1.96 1.96  
## P-value: 1.7308e-05

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع نرمال استاندارد با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "norm", statistic = Z, alpha = 0.05,  
 alternative = "two.sided")



در آزمون دو دنباله ، مقدار بحرانی در برابر با 96/1 است (جدول شماره 1). مقدار محاسبه شده برای بیشتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که تفاوت معنی‏داری بین فراوانی‏ متوسط رخداد شاخص تعداد روزهای تابستانی در دو ایستگاه کرمان و زاهدان وجود دارد.

# آزمون 7: آزمون [[18]](#footnote-19) برای مقایسه میانگین یک جامعه با عدد معلوم (واریانس جامعه نامعلوم است)

**هدف**

بررسی معنی‏دار بودن تفاوت بین میانگین مفروض یک جامعه و یک عدد معلوم (مثلاً میانگین یک نمونه ).

* فرضیه صفر ():
* فرضیه مقابل (): یا یا

**پیش فرض‏ها**

1. اگر واریانس جامعه[[19]](#footnote-20) معلوم باشد، «آزمون برای میانگین یک جامعه (آزمون 1)» کارآمدتر خواهد بود.
2. در صورتی که جامعه به صورت نرمال توزیع شده باشد، آزمون دقیق است. ولی اگر توزیع جامعه نرمال نباشد، آزمون باز هم تقریب خوبی به دست می‏دهد.

**روش**

از جامعه‏ای با میانگین مفروض و واریانس نامعلوم، یک نمونه تصادفی با حجم گرفته شده است و میانگین نمونه محاسبه و سپس انحراف معیار نمونه با استفاده از فرمول زیر محاسبه می‏گردد:

آماره آزمون تحت فرضیه صفر () برابر است با:

که می‏توان آماره آزمون را با توزیع ی استیودنت[[20]](#footnote-21) با درجه‏های آزادی[[21]](#footnote-22) مقایسه کرد. آزمون می‏تواند به هر یک از دو صورت یک دنباله[[22]](#footnote-23) یا دو دنباله[[23]](#footnote-24) انجام گیرد.

**مثال**

میانگین عملکرد چغندرقند استان خراسان رضوی تا سال آبی 97-1396، 58700 کیلوگرم در هکتار بوده است. در پایان سال آبی 98-1397 به صورت تصادفی 50 مزرعه انتخاب و با میانگین‏گیری، عملکرد متوسط نمونه چغندرقند 04/62739 کیلوگرم در هکتار و انحراف معیار نمونه چغندرقند نیز 33/14179 کیلوگرم در هکتار به دست می‏آید. در سطح معنی‏داری ، آیا فرضیه صفر (میانگین عملکرد چغندرقند استان خراسان رضوی در سال آبی 98-1397 تفاوتی با میانگین عملکرد چغندرقند تا سال آبی 97-1396 ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده، و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "DescTools")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if**(**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data <- **read.csv**(file = "data/Data\_Test\_07.csv", header = TRUE)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Year | Sample | Yield |
| ## | 1 | 1397-98 | 1 | 53930 |
| ## | 2 | 1397-98 | 2 | 86128 |
| ## | … | … | … | … |
| ## | 29 | 1397-98 | 29 | 61925 |
| ## | 30 | 1397-98 | 30 | 39046 |

**گام 2: بررسی محدودیت‏ها**

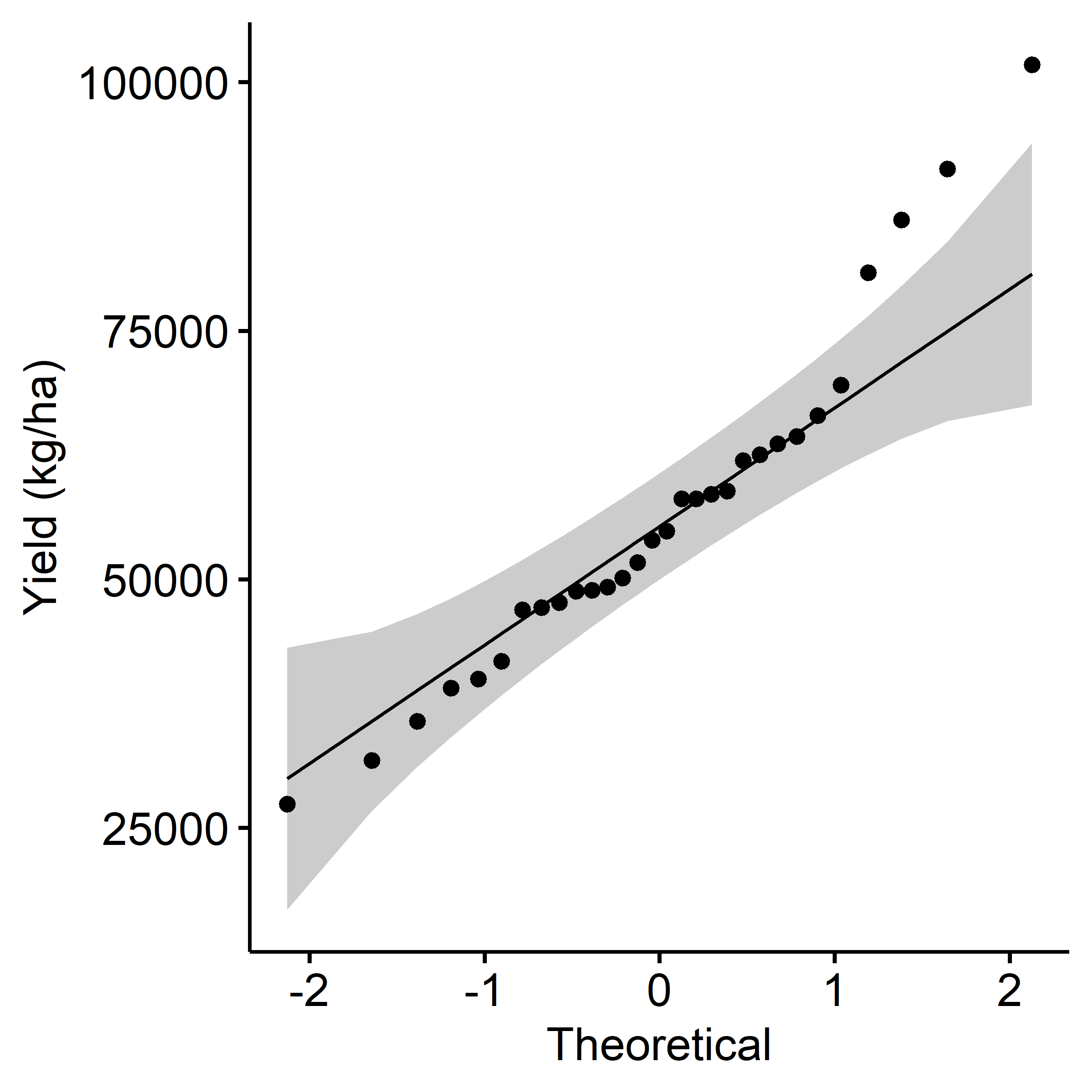
در صورتی که جامعه به صورت نرمال توزیع شده باشد، آزمون دقیق است: با توجه به اینکه مشخصات و داده‏های جامعه در اختیار نمی‏باشد، به جای آن به بررسی اینکه آیا نمونه از توزیع نرمال[[24]](#footnote-25) پیروی می‏کنند یا خیر، پرداخته شده است. برای انجام این آزمون از تابع **shapiro.test** استفاده شده است.

# use Shapiro-Wilk normality test for sample data  
**shapiro.test**(x = data**$**Yield)

## Shapiro-Wilk normality test  
## data: data$ Yield  
## W = 0.9468, p-value = 0.1386

با توجه به اینکه **p-مقدار[[25]](#footnote-26)** آزمون بیشتر از 05/0 می‏باشد، فرضیه صفر رد نشده و می‏توان نتیجه گرفته که داده‏های نمونه دارای توزیع نرمال می‏باشند. از طرفی می‏توان با رسم نمودار چندک چندک[[26]](#footnote-27) نیز نرمال بودن توزیع نمونه را مورد ارزیابی قرار داد. برای این منظور، از تابع **ggqqplot** موجود در بسته **ggpubr** استفاده شده است.

# normality plot for sample data  
ggpubr**::ggqqplot**(data = data**$**Yield, ylab = "Yield (kg/ha)")



همانطور که دیده می‏شود، فرض نرمال بودن توزیع نمونه‏ها مورد تایید قرار می‏گیرد. نمودارهای حاصل، تقریبا خط راستی هستند که از مرکز مختصات گذر کرده‏اند.

**گام 3: محاسبه آماره آزمون**

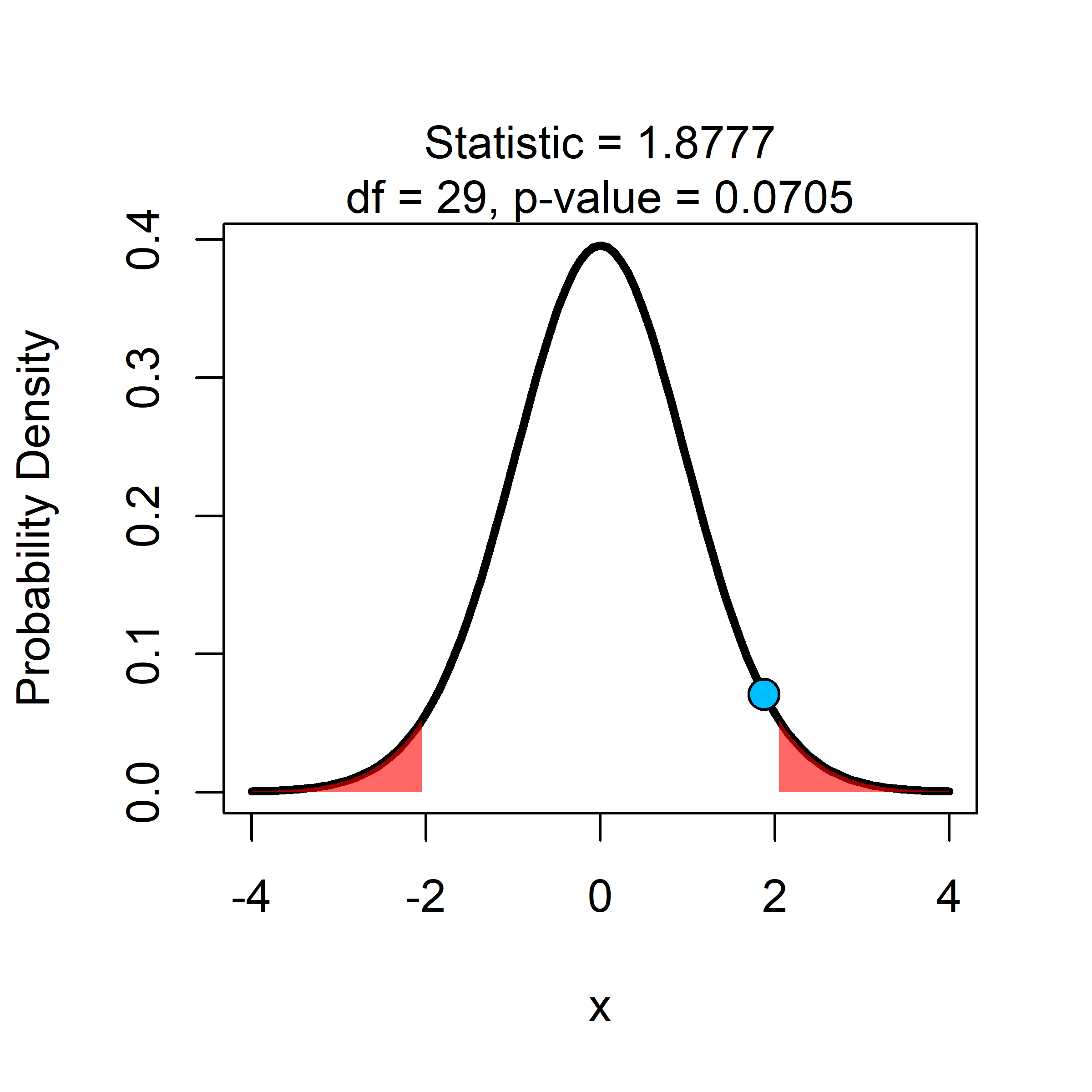
محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، بررسی معنی‏دار بودن تفاوت بین میانگین مفروض یک جامعه و میانگین یک نمونه، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values: population  
p = **c**("1392-93" = 43415, "1393-94" = 50260, "1394-95" = 54472,  
 "1395-96" = 54348, "1396-97" = 51248)  
mp = **mean**(x = p)  
  
# setting initial parameter values: sample  
x = data **%>%** **select**(Yield) **%>%** **unlist**()  
nx = **length**(x = x)  
mx = **mean**(x = x)  
sx = **sqrt**(**sum**((x **-** mx) **^** 2) **/** (nx **-** 1)) # sd(x = x)  
  
# test statistic  
t = (mx **-** mp) **/** (sx **/** **sqrt**(nx))  
  
# compute the critical values at 0.05 significance level (two tailed)  
alpha = 0.05  
two.sided = TRUE  
Tcritical = **qt**(p = (1 **-** alpha**/**2), df = nx **-** 1)  
  
# p-value (two tailed)  
p\_value = **pt**(q = t, df = nx **-** 1, lower.tail = FALSE)  
  
**if** (two.sided) {  
 p\_value = 2 **\*** p\_value  
 **if** (p\_value **>** 1) { p\_value = 2 **-** p\_value }  
}  
**cat**("Test Statistic: ", t, "\n")  
**cat**("Degrees of Freedom:", nx **-** 1, "\n")  
**cat**("Critical Values: ", **c**(**-**Tcritical, Tcritical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: 1.878  
## Degrees of Freedom: 29  
## Critical Values: -2.0452 2.0452  
## P-value: 0.0705

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع استیودنت با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "t", statistic = t, df = df, alpha = 0.05,  
 alternative = "two.sided")



در آزمون دو دنباله ، مقدار بحرانی در برابر با 0452/2 است (جدول شماره 2). مقدار محاسبه شده برای کمتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که تفاوت معنی‏داری بین میانگین عملکرد چغندرقند استان خراسان رضوی در سال آبی 98-1397 با میانگین عملکرد چغندرقند در بین سال‏های آبی 93-1392 تا 97-1396 وجود ندارد.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های معلوم و مساوی، از تابع **TTestA** موجود در بسته **DescTools** استفاده شده است.

# method 2  
# setting initial parameter values: population  
p = **c**("1392-93" = 43415, "1393-94" = 50260, "1394-95" = 54472,  
 "1395-96" = 54348, "1396-97" = 51248)  
mp = **mean**(x = p)  
  
# setting initial parameter values: sample  
x = data **%>%** **select**(Yield) **%>%** **unlist**()  
nx = **length**(x = x)  
mx = **mean**(x = x)  
sx = **sqrt**(**sum**((x **-** mx) **^** 2) **/** (nx **-** 1))  
  
# use "TTestA" function from "DescTools" package  
DescTools**::TTestA**(mx = mx, sx = sx, nx = nx, mu = mp,  
 alternative = "two.sided", conf.level = 0.95)

## One Sample t-test  
## data: x  
## t = 1.8777, df = 29, p-value = 0.0705  
## alternative hypothesis: true mean is not equal to 50748.6  
## 95 percent confidence interval:  
## 50228.8 62920.0  
## sample estimates:  
## mean of x: 56574.4

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند. این آزمون را می‏توان به صورت یک دنباله، با ناحیه بحرانی در سطح ، با توزیع نرمال استاندارد نیز مقایسه کرد.

# use "TTestA" function from "DescTools" package  
DescTools**::TTestA**(mx = mx, sx = sx, nx = nx, mu = mp,  
 alternative = "greater", conf.level = 0.95)

## One Sample t-test  
## data: x  
## t = 1.8777, df = 29, p-value = 0.0353  
## alternative hypothesis: true mean is greater than 50748.6  
## 95 percent confidence interval:  
## 51302.63 Inf  
## sample estimates:  
## mean of x: 56574.4

در آزمون یک دنباله ، مقدار بحرانی در برابر با 699/1 است (جدول شماره 2). مقدار محاسبه شده برای بیشتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که میانگین عملکرد چغندرقند استان خراسان رضوی در سال آبی 98-1397 بیشتر از میانگین عملکرد چغندرقند در بین سال‏های آبی 93-1392 تا 97-1396 می‏باشد و این تفاوت معنی‏دار است.

# آزمون 8: آزمون [[27]](#footnote-28) برای میانگین‏های دو جامعه (واریانس‏های دو جامعه نامعلوم ولی مساوی می‏باشند)

**هدف**

بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های نامعلوم ولی مساوی.

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. اگر واریانس‏های دو جامعه[[28]](#footnote-29) معلوم باشد، «آزمون برای میانگین‏های دو جامعه (آزمون 2)» کارآمدتر خواهد بود.
2. در صورتی که هر دو جامعه به صورت نرمال توزیع شده باشند، آزمون دقیق است. ولی اگر توزیع دو جامعه نرمال نباشند، آزمون باز هم تقریب خوبی به دست می‏دهد.

**روش**

از دو جامعه‏ با میانگین‏های و ، نمونه‏های تصادفی مستقل با حجم‏های و گرفته شده، و میانگین نمونه‏ها ( و ) محاسبه می‏گردد. سپس مجموع توان‏های دوم[[29]](#footnote-30) نمونه‏ها با استفاده از فرمول‏های زیر محاسبه می‏گردد:

بهترین تخمین از واریانس جامعه توسط فرمول زیر به دست می‏آید:

آماره آزمون برابر است با:

که می‏توان آماره آزمون را با توزیع ی استیودنت[[30]](#footnote-31) با درجه‏های آزادی[[31]](#footnote-32) مقایسه کرد. آزمون می‏تواند به هر یک از دو صورت یک دنباله[[32]](#footnote-33) یا دو دنباله[[33]](#footnote-34) انجام گیرد.

**مثال**

کارخانه‏‏ ایرانی تولید کننده آبپاش‏های آبیاری بارانی می‏خواهد بررسی کند که، آیا ضریب یکنواختی آبپاش تولید شده توسط این شرکت (ADF) با ضریب یکنواختی آبپاش نمونه خارجی (Nelson) در شرایط مشخص (متوسط سرعت باد 3 کیلومتر بر ساعت، فشار کارکرد 4 بار و فواصل آبپاش‏ها 2525 متر) یکسان است یا نه؟. برای بررسی عملکرد این دو نوع آبپاش، به صورت تصادفی 18 آبپاش تولید شده توسط این شرکت (ADF) و 15 آبپاش نمونه خارجی (Nelson) انتخاب و ضریب یکنواختی این آبپاش‏ها با استفاده از ضریب یکنواختی کریستیانسن (CU) محاسبه شد. با میانگین‏گیری، ضریب یکنواختی آبپاش‏های ADF و Nelson به ترتیب 38/81 و 57/84 درصد و مجموع توان‏های دوم نمونه‏ها به ترتیب 42/317 و 87/96 می‏باشد. با قبول فرضیه یکسان بودن واریانس دو جامعه، در سطح معنی‏داری ، آیا فرضیه صفر (تفاوتی بین میانگین ضریب یکنواختی آبپاش‏های نوع ADF و Nelson وجود ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if**(**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_08.csv", header = TRUE)  
  
# prepare data  
ADF = data **%>%** **filter**(Sprinkler **==** "ADF") **%>%** **select**(CU) **%>%** **unlist**()  
Nelson = data **%>%** **filter**(Sprinkler **==** "Nelson") **%>%** **select**(CU) **%>%** **unlist**()  
  
# Show Data:  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |
| --- | --- | --- | --- |
| ## |  | Sprinkler | CU |
| ## | 1 | Nelson | 82.0 |
| ## | 2 | Nelson | 83.8 |
| ## | … | … | … |
| ## | 26 | ADF | 81.1 |
| ## | 27 | ADF | 85.8 |

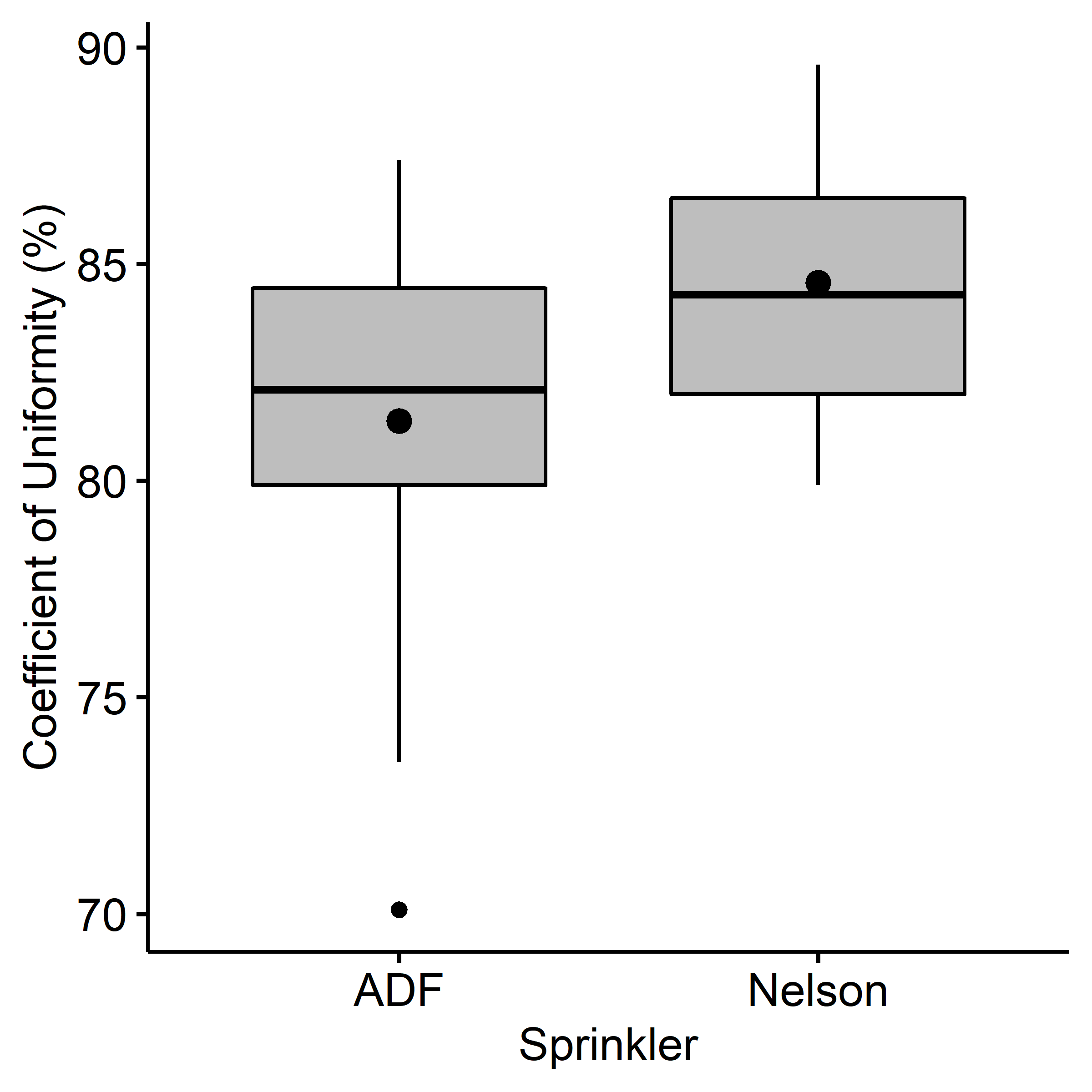
**گام 2: بررسی محدودیت‏ها**

1- فرض می‏شود واریانس دو جامعه با هم برابر هستند.

2- فرض می‏شود هر دو جامعه به صورت نرمال توزیع شده باشند.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها رسم شده است. در این نمودار از جعبه‏ای برای نمایش فاصله بین چارک اول و سوم استفاده می‏شود و خطی در داخل جعبه، میانه (چارک دوم) را مشخص می‏کند. خارج از جعبه حداقل و حداکثر مقدار داده را نیز مشخص می‌نمایند. گهگاه نمونه‏های خارج از محدوده نیز به صورت نقاطی نشان داده می‏شوند. برای این منظور، از تابع **ggboxplot** موجود در بسته **ggpubr** استفاده شده است.

# visualizing samples with box plots  
ggpubr**::****ggboxplot**(data = data, x = "Sprinkler", y = "CU",   
 fill = "gray", add = "mean", ylab = "Coefficient of Uniformity (%)")



**گام 3: محاسبه آماره آزمون**

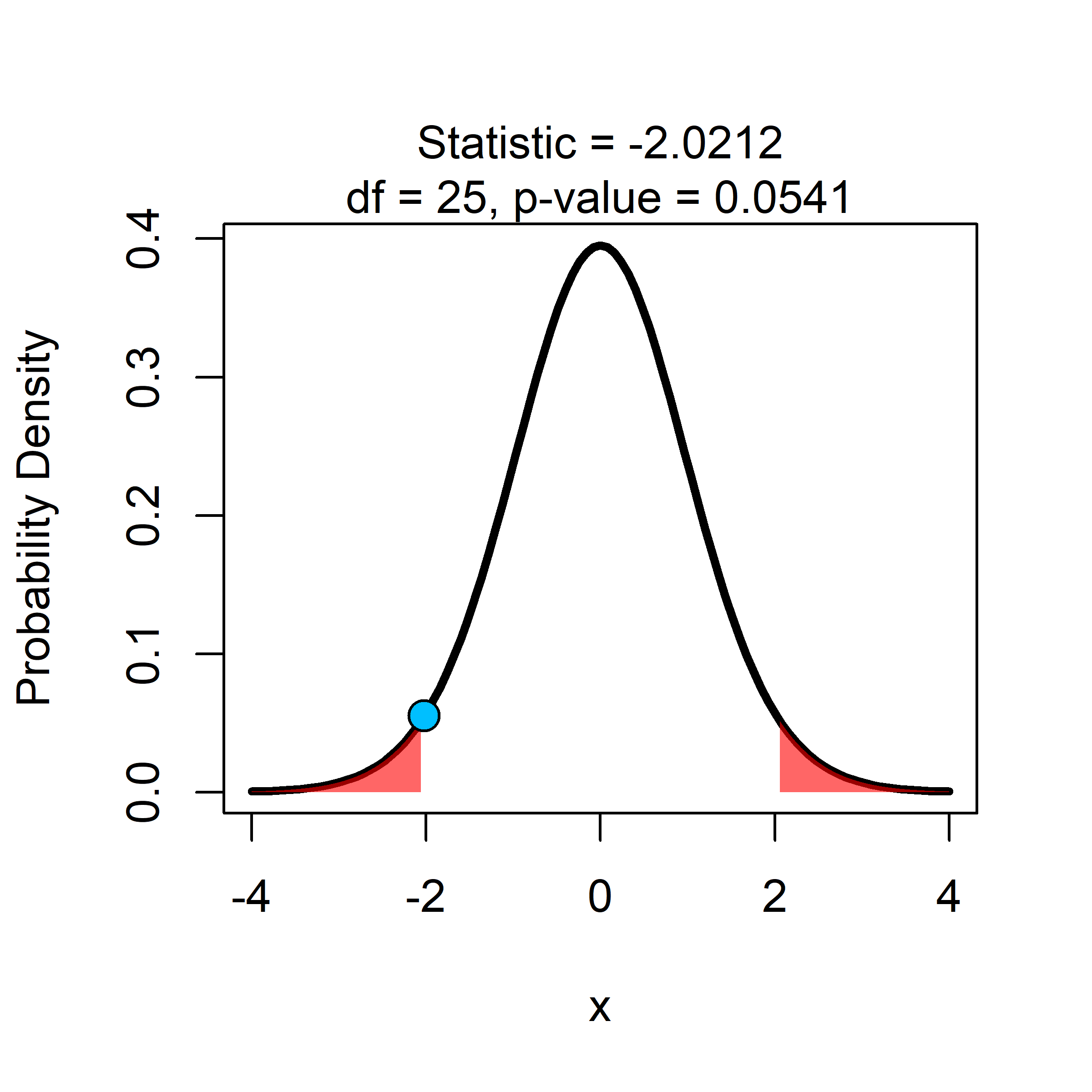
محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های نامعلوم ولی مساوی، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
n\_ADF = **length**(x = ADF)  
m\_ADF = **mean**(x = ADF)  
n\_Nelson = **length**(x = Nelson)  
m\_Nelson = **mean**(x = Nelson)  
  
# sums of squares  
s\_ADF = **sum**((ADF **-** m\_ADF) **^** 2)  
s\_Nelson = **sum**((Nelson **-** m\_Nelson) **^** 2)  
# population variance  
s = (s\_ADF **+** s\_Nelson) **/** (n\_ADF **+** n\_Nelson **-** 2)  
  
# test statistic  
t = (m\_ADF **-** m\_Nelson) **/** **sqrt**(s **\*** ((1 **/** n\_ADF) **+** (1 **/** n\_Nelson)))  
  
# compute the critical values at 0.05 significance level (two tailed)  
alpha = 0.05  
two.sided = TRUE  
Tcritical = **qt**(p = (1 **-** alpha**/**2), df = (n\_ADF **+** n\_Nelson **-** 2))  
# p-value (two tailed)  
p\_value = **pt**(q = t, df = (n\_ADF **+** n\_Nelson **-** 2), lower.tail = FALSE)  
  
**if** (two.sided) {  
 p\_value = 2 **\*** p\_value  
 **if** (p\_value **>** 1) { p\_value = 2 **-** p\_value }  
}  
**cat**("Test Statistic: ", t, "\n")  
**cat**("Degrees of Freedom:", (n\_ADF **+** n\_Nelson **-** 2), "\n")  
**cat**("Critical Values: ", **c**(**-**Tcritical, Tcritical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: -2.0212  
## Degrees of Freedom: 25  
## Critical Values: -2.0595 2.0595  
## P-value: 0.0541

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع استیودنت با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "t", statistic = t, df = df, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله ، مقدار بحرانی در برابر با 0595/2 است (جدول شماره 2). مقدار محاسبه شده برای کمتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که تفاوت معنی‏داری بین میانگین ضریب یکنواختی آبپاش‏های نوع ADF و Nelson وجود ندارد.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های نامعلوم ولی مساوی، از تابع **t.test** موجود در بسته **stats** استفاده شده است.

# method 2  
# use "t.test" function from "stats" package  
**t.test**(x = ADF, y = Nelson, var.equal = TRUE,  
 alternative = "two.sided", conf.level = 0.95)

## Two Sample t-test  
## data: ADF and Nelson  
## t = -2.0212, df = 25, p-value = 0.0541  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -6.4338 0.0605  
## sample estimates:  
## mean of x mean of y  
## 81.3800 84.5667

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 9: آزمون [[34]](#footnote-35) برای میانگین‏های دو جامعه (واریانس‏های دو جامعه نامعلوم و نامساوی می‏باشند)

**هدف**

بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های نامعلوم و نامساوی.

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. اگر واریانس‏های دو جامعه[[35]](#footnote-36) معلوم باشد، «آزمون برای میانگین‏های دو جامعه (آزمون 3)» کارآمدتر خواهد بود.
2. در صورتی که هر دو جامعه به صورت نرمال توزیع شده باشند یا حجم نمونه‏ها به اندازه کافی بزرگ باشند آزمون دقیق است.
3. آزمون باید فقط برای فرض (آزمون دو دنباله) مورد استفاده قرار بگیرد.

**روش**

از دو جامعه‏ با میانگین‏های و ، نمونه‏های تصادفی مستقل با حجم‏های و گرفته شده، و میانگین نمونه‏ها ( و ) محاسبه می‏گردد. سپس واریانس نمونه‏ها با استفاده از فرمول‏های زیر محاسبه می‏گردد:

آماره آزمون برابر است با:

که می‏توان آماره آزمون را با توزیع ی استیودنت[[36]](#footnote-37) با درجه‏های آزادی به دست آمده از فرمول زیر مقایسه کرد.

**مثال**

کارخانه‏‏ ایرانی تولید کننده آبپاش‏های آبیاری بارانی می‏خواهد بررسی کند که، آیا ضریب یکنواختی آبپاش تولید شده توسط این شرکت (ADF) با ضریب یکنواختی آبپاش نمونه خارجی (Nelson) در شرایط مشخص (متوسط سرعت باد 3 کیلومتر بر ساعت، فشار کارکرد 4 بار و فواصل آبپاش‏ها 2525 متر) یکسان است یا نه؟. برای بررسی عملکرد این دو نوع آبپاش، به صورت تصادفی 18 آبپاش تولید شده توسط این شرکت (ADF) و 15 آبپاش نمونه خارجی (Nelson) انتخاب و ضریب یکنواختی این آبپاش‏ها با استفاده از ضریب یکنواختی کریستیانسن (CU) محاسبه شد. با میانگین‏گیری، ضریب یکنواختی آبپاش‏های ADF و Nelson به ترتیب 38/81 و 57/84 درصد و مجموع توان‏های دوم نمونه‏ها به ترتیب 42/317 و 87/96 می‏باشد. با قبول فرض یکسان نبودن واریانس دو جامعه، در سطح معنی‏داری ، آیا فرضیه صفر (تفاوتی بین میانگین ضریب یکنواختی آبپاش‏های نوع ADF و Nelson وجود ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if**(**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_09.csv", header = TRUE)  
  
# show data:  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |
| --- | --- | --- | --- |
| ## |  | Sprinkler | CU |
| ## | 1 | Nelson | 82.0 |
| ## | 2 | Nelson | 83.8 |
| ## | … | … | … |
| ## | 26 | ADF | 81.1 |
| ## | 27 | ADF | 85.8 |

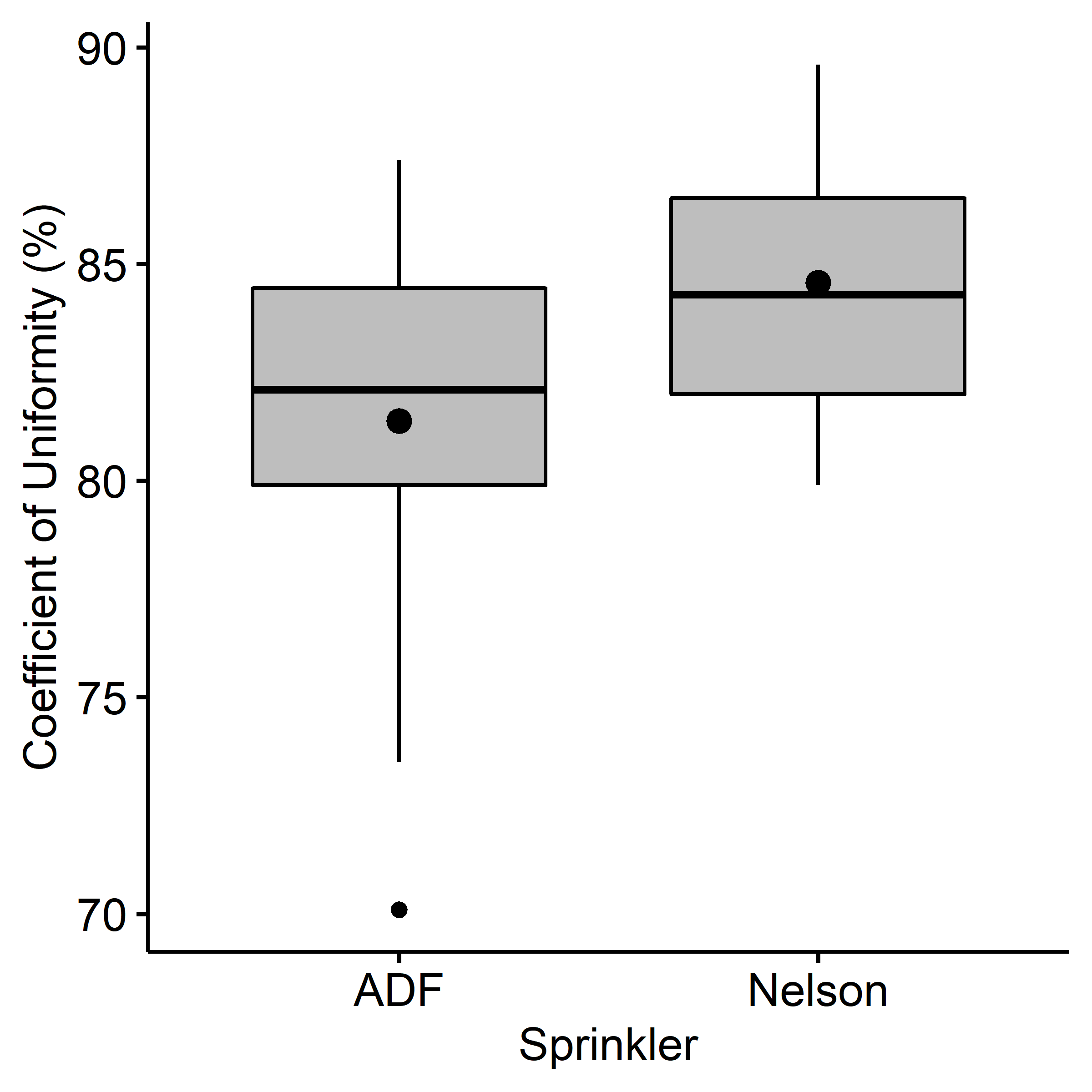
**گام 2: بررسی محدودیت‏ها**

1- آزمون فقط به صورت دو دنباله مورد استفاده قرار بگیرد.

2- فرض می‏شود هر دو جامعه به صورت نرمال توزیع شده باشند.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها رسم شده است. برای این منظور، از تابع **ggboxplot** موجود در بسته **ggpubr** استفاده شده است.

# visualizing samples with box plots  
ggpubr**::ggboxplot**(data = data, x = "Sprinkler", y = "CU",   
 fill = "gray", add = "mean", ylab = "Coefficient of Uniformity (%)")



**گام 3: محاسبه آماره آزمون**

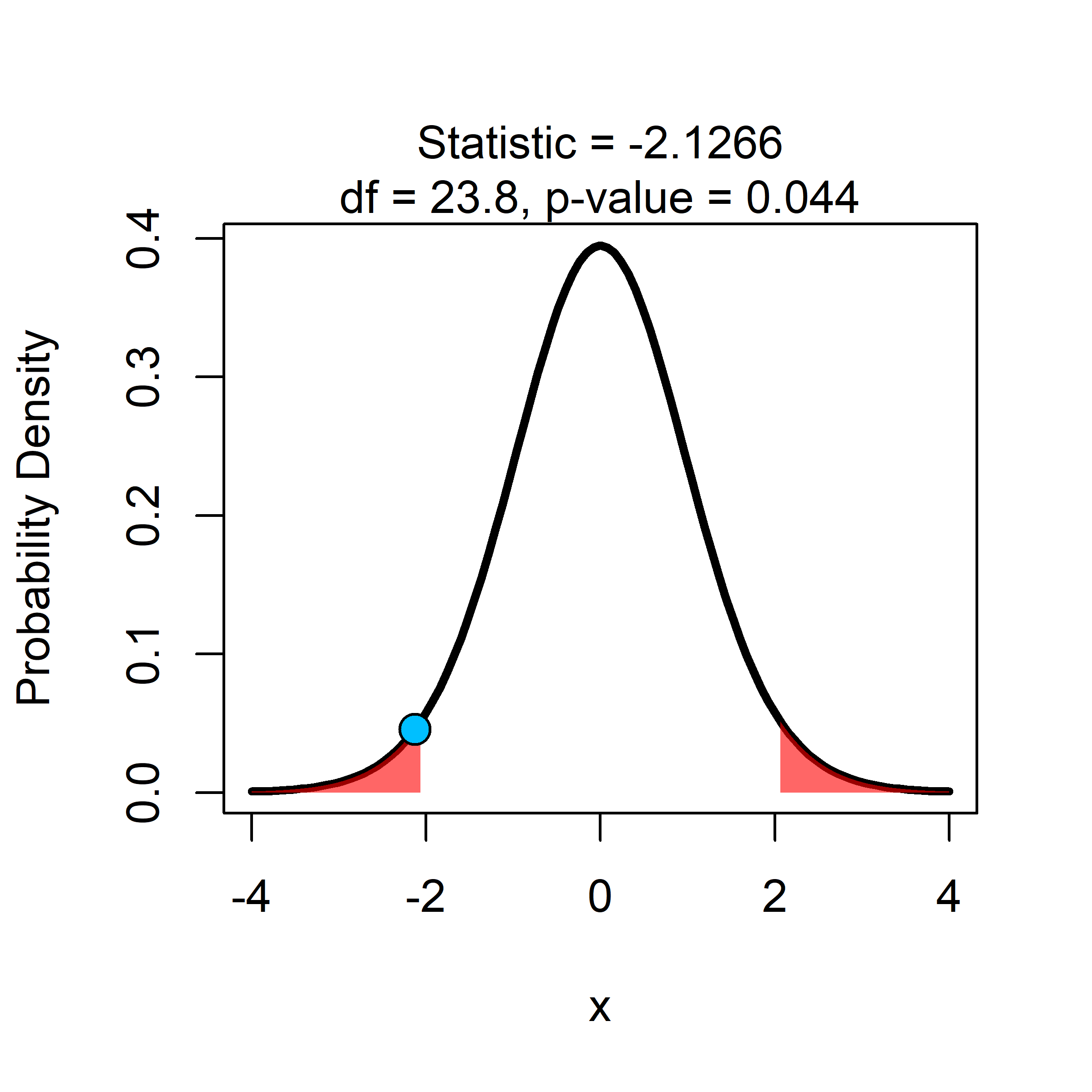
محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های نامعلوم و نامساوی، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
A = data **%>%** **filter**(Sprinkler **==** "ADF") **%>%** **select**(CU) **%>%** **unlist**()  
nA = **length**(x = A) # sample size (ADF)  
mA = **mean**(x = A) # sample mean (ADF)  
sA = (**sum**((A **-** mA) **^** 2)) **/** (nA **-** 1) # sample variance (ADF)  
  
B = data **%>%** **filter**(Sprinkler **==** "Nelson") **%>%** **select**(CU) **%>%** **unlist**()  
nB = **length**(x = B) # sample size (Nelson)  
mB = **mean**(x = B) # sample mean (Nelson)  
sB = (**sum**((B **-** mB) **^** 2)) **/** (nB **-** 1) # sample variance (Nelson)  
  
# test statistic  
t = (mA **-** mB) **/** **sqrt**((sA **/** nA) **+** (sB **/** nB))  
  
# degrees of freedom  
df\_numerator = ((sA **/** nA) **+** (sB **/** nB)) **^** 2  
df\_denominator = ((sA **^** 2) **/** ((nA **^** 2) **\*** (nA **-** 1))) **+** ((sB **^** 2) **/** ((nB **^** 2) **\*** (nB **-** 1)))  
df = df\_numerator **/** df\_denominator  
  
# compute the critical values at 0.05 significance level (two tailed)  
alpha = 0.05  
two.sided = TRUE  
Tcritical = **qt**(p = (1 **-** alpha**/**2), df = df)  
  
# p-value (two tailed)  
p\_value = **pt**(q = t, df = df, lower.tail = FALSE)  
  
**if** (two.sided) {  
 p\_value = 2 **\*** p\_value  
 **if** (p\_value **>** 1) {p\_value = 2 **-** p\_value }  
}  
  
  
**cat**("Test Statistic: ", t, "\n")  
**cat**("Degrees of Freedom:", df, "\n")  
**cat**("Critical Values: ", **c**(**-**Tcritical, Tcritical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: -2.1266  
## Degrees of Freedom: 23.7646  
## Critical Values: -2.0650 2.0650  
## P-value: 0.0440

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع استیودنت با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "t", statistic = t, df = df, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله ، مقدار بحرانی در برابر با 065/2 است (جدول شماره 2). مقدار محاسبه شده برای کمتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که تفاوت معنی‏داری بین میانگین ضریب یکنواختی آبپاش‏های نوع ADF و Nelson وجود ندارد.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه با واریانس‏های نامعلوم و نامساوی، از تابع **t.test** موجود در بسته **stats** استفاده شده است.

# method 2  
x = data **%>%** **filter**(Sprinkler **==** "ADF") **%>%** **select**(CU) **%>%** **unlist**()  
y = data **%>%** **filter**(Sprinkler **==** "Nelson") **%>%** **select**(CU) **%>%** **unlist**()  
# use "t.test" function from "stats" package  
**t.test**(x = x, y = y, var.equal = FALSE,  
 alternative = "two.sided", conf.level = 0.95)

## Welch Two Sample t-test  
## data: x and y  
## t = -2.1266, df = 23.765, p-value = 0.0440  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -6.2810 -0.0924  
## sample estimates:  
## mean of x mean of y  
## 81.3800 84.5667

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 10: آزمون [[37]](#footnote-38) برای میانگین‏های دو جامعه (روش مقایسه‏های جفت شده[[38]](#footnote-39))

**هدف**

بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه و . هیچگونه فرضی برای واریانس‏های دو جامعه وجود ندارد.

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. مشاهدات دو نمونه باید به صورت جفتی (زوجی) جمع‏آوری شده باشند. صرفنظر از تفاوت‏های دو جامعه، مشاهدات در هر زوج باید تحت شرایط یکسان یا تقریبا یکسان جمع‏آوری شده باشند.
2. در صورتی که هر دو جامعه به صورت نرمال توزیع شده باشند، آزمون دقیق است. ولی اگر توزیع دو جامعه نرمال هم نباشند، آزمون باز هم تقریب خوبی به دست می‏دهد.

**روش**

در ابتدا تفاوت بین هر جفت از مشاهدات محاسبه، و سپس واریانس تفاضل‏ نمونه‏ها، با فرض جفت مشاهده، توسط فرمول زیر محاسبه می‏گردد:

درصورتی که میانگین نمونه‏ها و باشد، آماره آزمون برابر است با:

که می‏توان آماره آزمون را با توزیع ی استیودنت[[39]](#footnote-40) با درجه‏های آزادی[[40]](#footnote-41) مقایسه کرد. آزمون می‏تواند به هر یک از دو صورت یک دنباله[[41]](#footnote-42) یا دو دنباله[[42]](#footnote-43) انجام گیرد.

**مثال**

هدف از انجام آزمایشی بررسی تاثیر کود حیوانی بر روی خاک‏های شور در مناطق خشک می‏باشد. برای اجرای این پژوهش دو محدوده که از لحاظ خصوصیات فیزیکی و شیمیایی یکسانی بودند، انتخاب گردید. دو تیمار آزمایشی شامل شاهد (عدم مصرف کود) و خاک اصلاحی (اضافه کردن 15 درصد کود حیوانی) و 10 تکرار در نظر گرفته شده است. قبل و بعد از مخلوط کردن کود حیوانی، نمونه‏های خاک از عمق 5 سانتیمتری برداشته و بعضی از ویژگی‏های فیزیکی و شیمیایی خاک از جمله pH آن مورد تجزیه و تحلیل قرار گرفت. در سطح معنی‏داری ، آیا فرضیه صفر (تفاوتی بین pH خاک قبل و بعد از اضافه کردن کود حیوانی وجود ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "reshape2")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if**(**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_10.csv", header = TRUE)  
# show data:  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |
| --- | --- | --- | --- |
| ## |  | Control | Treatment |
| ## | 1 | 7.41 | 7.56 |
| ## | 2 | 7.81 | 7.27 |
| ## | … | … | … |
| ## | 9 | 7.79 | 8.42 |
| ## | 10 | 7.19 | 7.31 |

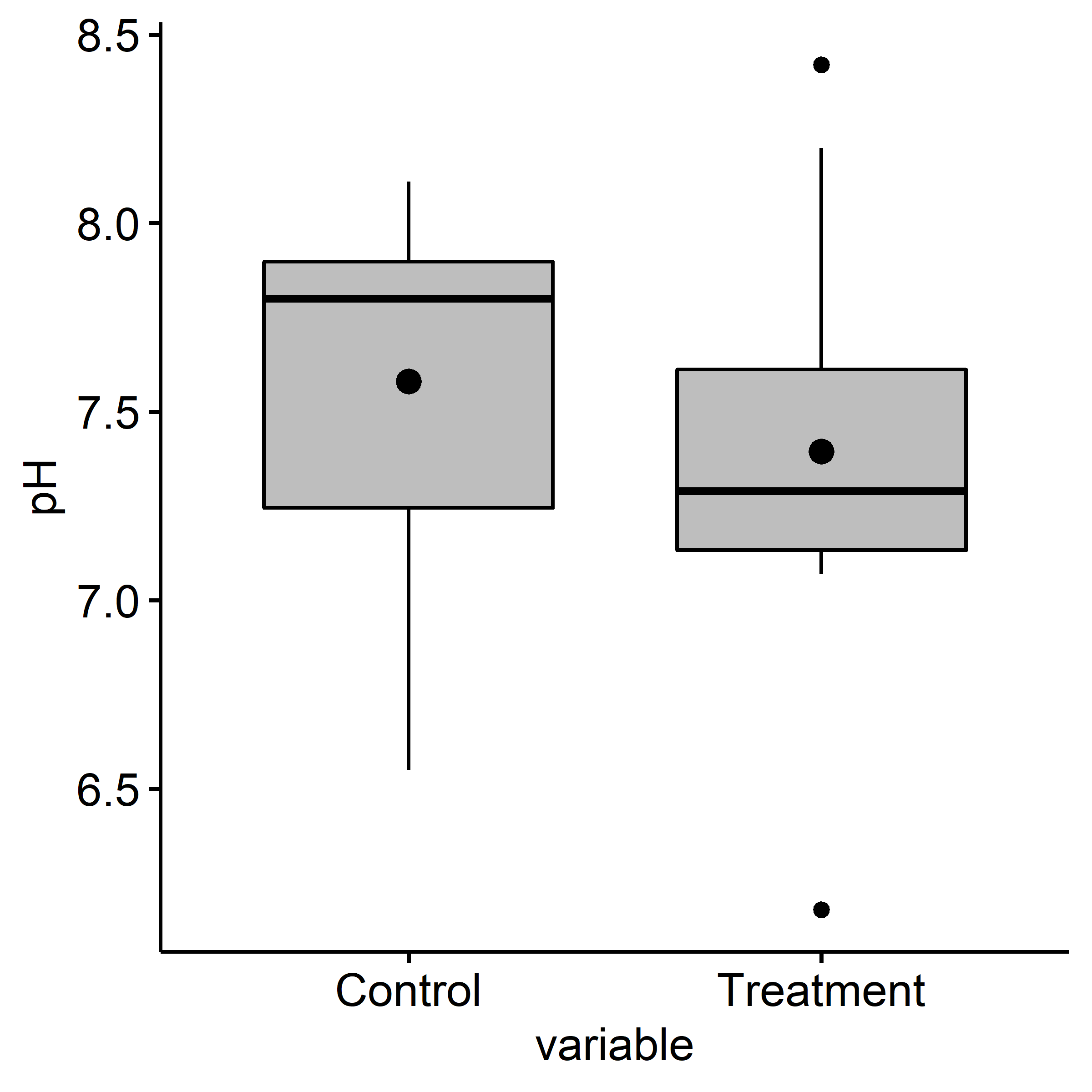
**گام 2: بررسی محدودیت‏ها**

1- مشاهدات دو نمونه به صورت جفتی جمع‏آوری شده‏اند.

2- فرض می‏شود هر دو جامعه به صورت نرمال توزیع شده باشند.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها رسم شده است. برای این منظور، از تابع **ggboxplot** موجود در بسته **ggpubr** استفاده شده است.

# visualizing samples with box plots  
dataPlot = data **%>%** reshape2**::melt**(value.name = "pH")  
ggpubr**::ggboxplot**(data = dataPlot, x = "variable", y = "pH",   
 fill = "gray", add = "mean", ylab = "pH")



**گام 3: محاسبه آماره آزمون**

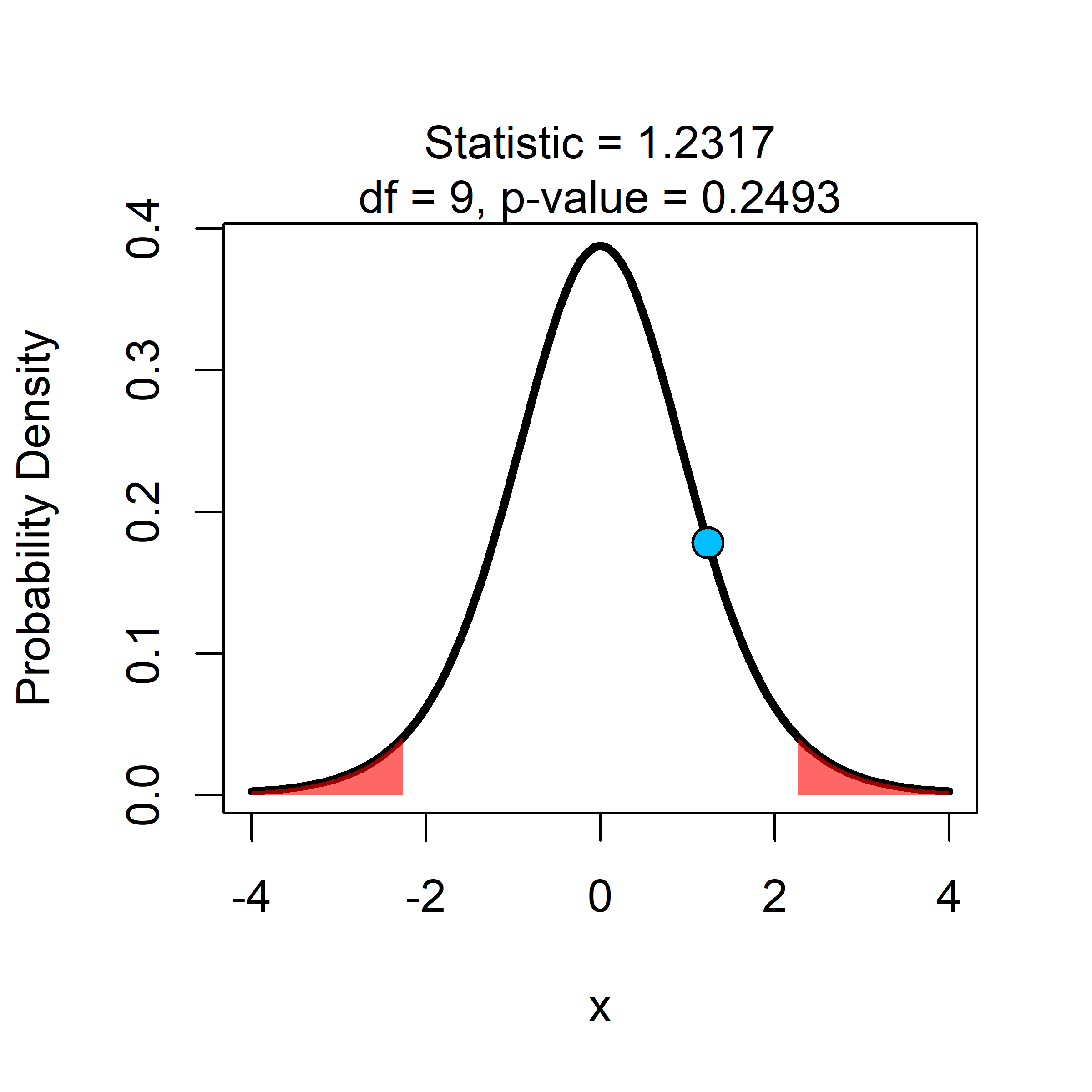
محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
x = data**$**Control  
mx = **mean**(x = x)  
y = data**$**Treatment  
my = **mean**(x = y)  
n = **nrow**(x = data)  
  
# differences d are formed for each pair of observations  
d = x **-** y  
md = **mean**(x = d)  
  
# variance of the difference  
s = **sum**((d **-** md) **^** 2 **/** (n **-** 1))  
  
# test statistic  
t = (mx **-** my) **/** **sqrt**(s **/** n)  
  
# degrees of freedom  
df = n **-** 1  
  
# compute the critical values at 0.05 significance level (two tailed)  
alpha = 0.05  
two.sided = TRUE  
Tcritical = **qt**(p = (1 **-** alpha**/**2), df = df)  
  
# p-value (two tailed)  
p\_value = **pt**(q = t, df = df, lower.tail = FALSE)  
  
**if** (two.sided) {  
 p\_value = 2 **\*** p\_value  
 **if** (p\_value **>** 1) {p\_value = 2 **-** p\_value }  
}  
**cat**("Test Statistic: ", t, "\n")  
**cat**("Degrees of Freedom:", df, "\n")  
**cat**("Critical Values: ", **c**(**-**Tcritical, Tcritical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: 1.2317  
## Degrees of Freedom: 9  
## Critical Values: -2.2622 2.2622  
## P-value: 0.2493

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع استیودنت با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "t", statistic = t, df = df, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله ، مقدار بحرانی در برابر با 2622/2 است (جدول شماره 2). مقدار محاسبه شده برای کمتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که تفاوت معنی‏داری بین میانگین pH خاک قبل و بعد از اضافه کردن کود حیوانی وجود ندارد.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین میانگین‏های دو جامعه، از تابع **t.test** موجود در بسته **stats** استفاده شده است.

# method 2  
# use "t.test" function from "stats" package  
x = data**$**Control  
y = data**$**Treatment  
**t.test**(x = x, y = y, paired = TRUE,  
 alternative = "two.sided", conf.level = 0.95)

## Paired t-test  
## data: x and y  
## t = 1.2317, df = 9, p-value = 0.2493  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.1556 0.5276  
## sample estimates:  
## mean of the differences:  
## 0.186

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 11: آزمون [[43]](#footnote-44) برای ضریب رگرسیونی[[44]](#footnote-45)

**هدف**

بررسی معنی‏دار بودن ضریب رگرسیونی روی .

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

پیش فرض‏های معمول برای رگرسیون باید برقرار باشد، مانند:

1. متغیر به ازای هر مقدار باید دارای توزیع نرمال باشد.
2. به ازای هر مقدار مشخص از ، واریانس مقادیر باید ثابت باشد.

**روش**

به منظور تخمین یک رابطه رگرسیون خطی به شکل ، نمونه‏ای جفتی (زوج) از نقاط مورد نیاز است. ضریب رگرسیونی نامیده می‏شود، و برای آزمون فرضیه صفر در ابتدا برآورد نمونه‏ای[[45]](#footnote-46) زیر محاسبه می‏گردد:

واریانس ها و ها پیرامون خط رگرسیون[[46]](#footnote-47) به شرح زیر محاسبه می‏شوند:

که و به ترتیب میانگین‏های ها و ها هستند. آماره آزمون برابر است با:

که می‏توان آماره آزمون را با توزیع ی استیودنت[[47]](#footnote-48) با درجه‏های آزادی[[48]](#footnote-49) مقایسه کرد. به دلیل آنکه مقدار می‏تواند مثبت یا منفی باشد، فرضیه صفر آزمون باید به صورت دو دنباله[[49]](#footnote-50) در نظر گرفته شود. هرچند، فرضیه صفر آزمون می‏تواند به صورت یک دنباله در نظر گرفته شود اگر فرضیه مقابل سویی[[50]](#footnote-51) باشد.

**مثال**

در پژوهشی هدف ارائه مدل ساده رگرسیون خطی برای برآورد غلظت اورانیوم، با استفاده از نمونه‏های اندازه‏گیری شده کیفیت آب زیرزمینی دشت مشهد می‏باشد. بدین منظور 53 نمونه کیفیت آب زیرزمینی از دشت مشهد تهیه و مقدار غلظت اورانیوم آن‏ها بر حسب میکروگرم بر لیتر و کل مواد جامد محلول آن‏ها بر حسب میلی‏گرم بر لیتر اندازه‏گیری شد. بر اساس نمونه‏های موجود، ضریب رگرسیونی ، 5818/0، واریانس ها، 234/4690 و واریانس ها، 005/3597 به دست آمد. با توجه به اطلاعات فوق و در سطح معنی‏داری ، آیا فرضیه صفر (ضریب رگرسیون رابطه خطی بین بارندگی سالانه ایستگاه مشهد و سبزوار برابر با صفر بوده و رگرسیون معنی‏داری وجود ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "ggExtra", "jtools")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if**(**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_11.csv", header = TRUE)  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Year | Mashhad | Sabzevar |
| ## | 1 | 1959 | 234.5 | 145.4 |
| ## | 2 | 1960 | 174.5 | 113.3 |
| ## | … | … | … | … |
| ## | 59 | 2017 | 238.1 | 134.1 |
| ## | 60 | 2018 | 237.4 | 139.7 |

**گام 2: بررسی محدودیت‏ها**

متغیر به ازای هر مقدار دارای توزیع نرمال می‏باشد: در این مرحله برای بررسی اینکه آیا داده‏های بارندگی سالانه دو ایستگاه مشهد و سبزوار از توزیع نرمال[[51]](#footnote-52) پیروی می‏کنند یا خیر، از **«آزمون شاپیرو ویلک[[52]](#footnote-53)»** استفاده شده است. برای انجام این آزمون از تابع **shapiro.test** بسته **stats** استفاده شده است.

# assumption checking:  
# 1. the variable y follows a normal distribution for each value of x.  
# use shapiro-wilk normality test  
**shapiro.test**(x = data**$**Mashhad)

## Shapiro-Wilk normality test  
## data: data$Mashhad  
## W = 0.9723, p-value = 0.1897

# use Shapiro-Wilk normality test  
**shapiro.test**(x = data**$**Sabzevar)

## Shapiro-Wilk normality test  
## data: data$Sabzevar  
## W = 0.9716, p-value = 0.1753

با توجه به اینکه **p-مقدار[[53]](#footnote-54)** هر دو آزمون بیشتر از 05/0 می‏باشد، فرضیه صفر رد نشده و می‏توان نتیجه گرفت که داده‏های بارندگی سالانه دو ایستگاه مشهد و سبزوار از توزیع نرمال پیروی می‏کنند. از طرفی می‏توان با رسم نمودار چندک چندک[[54]](#footnote-55) نیز نرمال بودن داده‏ها را مورد ارزیابی قرار داد. برای این منظور، از تابع **ggqqplot** موجود در بسته **ggpubr** استفاده شده است.

# normality plot  
ggpubr**::ggqqplot**(data = data**$**Mashhad, ylab = "Mashhad Prec. (mm)")  
ggpubr**::ggqqplot**(data = data**$**Sabzevar, ylab = "Sabzevar Prec. (mm)")

|  |  |
| --- | --- |
|  |  |

همانطور که دیده می‏شود، فرض نرمال بودن توزیع دو جامعه مورد تایید قرار می‏گیرد. از طرفی به منظور تجسم رابطه بین این دو ایستگاه، نمودار نقطه‏ای آن‏ها با استفاده از تابع **ggscatter** از بسته **ggpubr** رسم شده است.

# visualize data using scatter plots  
data **%>%** ggpubr**::ggscatter**(x = "Mashhad", y = "Sabzevar",  
 add = "reg.line", conf.int = TRUE,  
 xlab = "Mashhad Prec. (mm)",  
 ylab = "Sabzevar Prec. (mm)") **%>%**   
 ggExtra**::ggMarginal**(type = "boxplot", fill = "gray90")



**گام 3: محاسبه آماره آزمون**

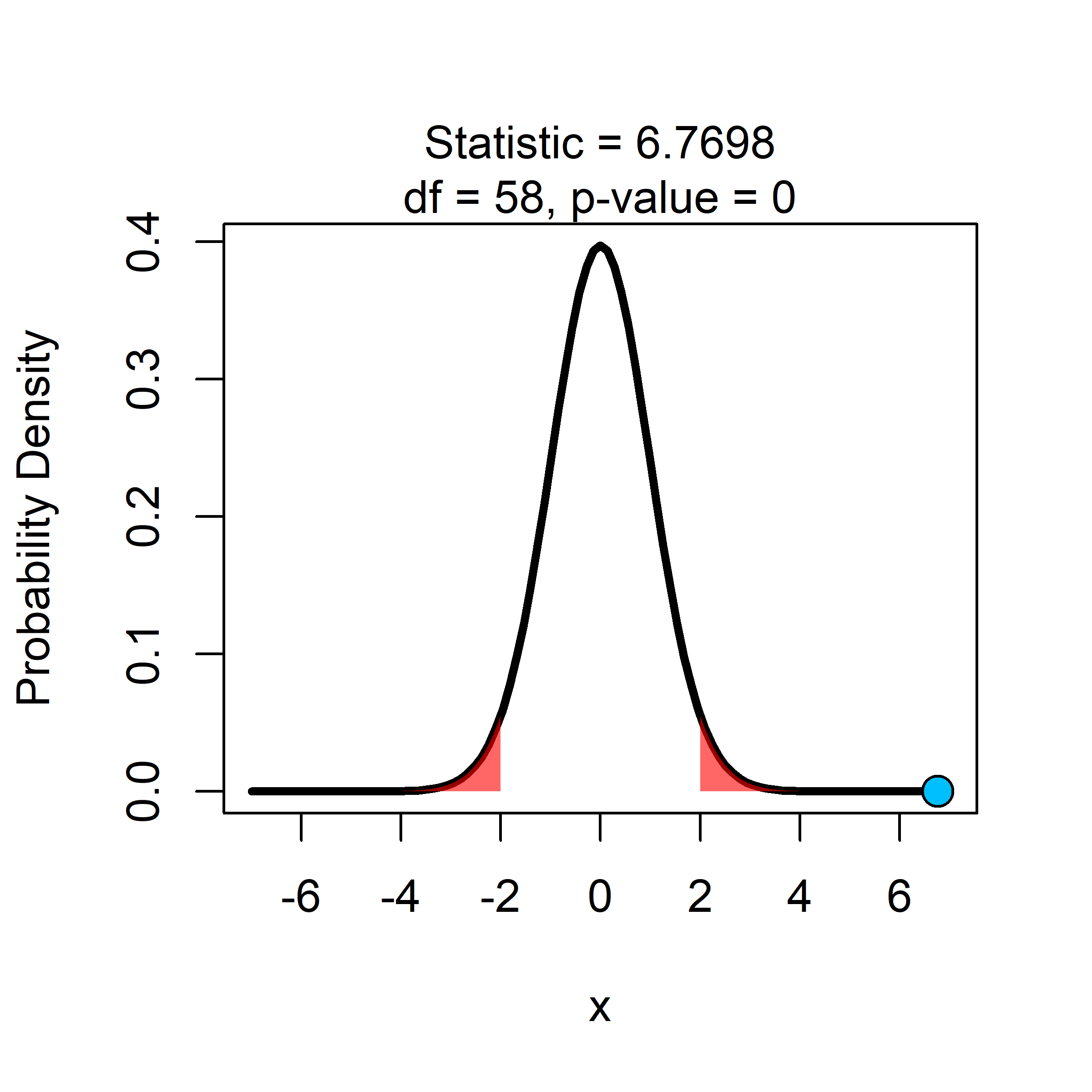
محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن ضریب رگرسیونی روی ، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
x = data**$**Mashhad  
x.m = **mean**(x = x)  
y = data**$**Sabzevar  
y.m = **mean**(x = y)  
n = **nrow**(x = data)  
  
# regression coefficient  
b.numerator = **sum**(x **\*** y) **-** ((**sum**(x) **\*** **sum**(y)) **/** n)  
b.denominator = **sum**(x **^** 2) **-** ((**sum**(x) **^** 2)**/** n)  
b = b.numerator **/** b.denominator  
  
# variance of x  
s.x = **sum**((x **-** x.m) **^** 2) **/** (n **-** 1)  
  
# variance of y  
s.y = **sum**((y **-** y.m) **^** 2) **/** (n **-** 1)  
  
# variance of yx  
s.yx = ((n **-** 1) **/** (n **-** 2)) **\*** (s.y **-** (s.x **\*** (b **^** 2)))  
# test statistic  
t = **sqrt**(((b **^** 2) **\*** s.x **\*** (n **-** 1)) **/** (s.yx))  
  
# degrees of freedom  
df = n **-** 2  
  
# compute the critical values at 0.05 significance level (two tailed)  
alpha = 0.05  
two.sided = TRUE  
Tcritical = **qt**(p = (1 **-** alpha**/**2), df = df)  
  
# p-value (two tailed)  
p\_value = **pt**(q = t, df = df, lower.tail = FALSE)  
  
**if** (two.sided) {  
 p\_value = 2 **\*** p\_value  
 **if** (p\_value **>** 1) { p\_value = 2 **-** p\_value }  
}  
**cat**("Test Statistic: ", t, "\n")  
**cat**("Degrees of Freedom:", df, "\n")  
**cat**("Critical Values: ", **c**(**-**Tcritical, Tcritical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: 6.7698  
## Degrees of Freedom: 58  
## Critical Values: -2.0017 2.0017  
## P-value: 7.1315e-09

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع استیودنت با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "t", statistic = t, df = df, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله ، مقدار بحرانی در برابر با 2 است (جدول شماره 2). مقدار محاسبه شده برای بیشتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که ضریب رگرسیون رابطه خطی بین بارندگی سالانه ایستگاه مشهد و سبزوار مخالف با صفر بوده و رگرسیون معنی‏داری بین این دو ایستگاه وجود دارد. باید به این نکته توجه داشت که این آزمون در رابطه با چگونگی کیفیت پیش‏بینی توسط ارائه نمی‏دهد، و تنها معنی‏دار بودن رگرسیون را بررسی می‏کند.

در روش دوم، بررسی معنی‏دار بودن ضریب رگرسیونی روی ، از تابع **lm** موجود در بسته **stats** استفاده شده است.

# method 2  
# use "lm" function from "stats" package  
mod = **lm**(formula = Sabzevar **~** Mashhad, data = data)  
jtools**::summ**(model = mod)

## MODEL INFO:  
## Observations: 60  
## Dependent Variable: Sabzevar  
## Type: OLS linear regression   
## Standard errors: OLS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## |  | Est. | S.E | t val. | p |
| ## | (Intercept) | 39.57 | 22.06 | 1.79 | 0.08 |
| ## | Mashhad | 0.58 | 0.09 | 6.77 | 0.00 |

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 12: آزمون [[55]](#footnote-56) یک ضریب همبستگی[[56]](#footnote-57)

**هدف**

بررسی تفاوت معنی‏داری آماری ضریب همبستگی یک نمونه با عدد صفر.

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. فرض می‏شود که مقادیر و از یک توزیع نرمال دو متغیره[[57]](#footnote-58) باشند و رابطه بین آن‏ها نیز به صورت خطی می‏باشد. در صورتی که فرضیه صفر (): مورد نظر باشد، به آزمون 13 (آزمون یک ضریب همبستگی) مراجعه شود.

**روش**

با در نظر گرفتن نمونه‏ای جفتی (زوج) از نقاط ، ضریب همبستگی با استفاده از فرمول زیر محاسبه می‏گردد:

برای آزمون فرضیه صفر ، آماره آزمون برابر است با:

که می‏توان آماره آزمون را با توزیع ی استیودنت[[58]](#footnote-59) با درجه‏های آزادی[[59]](#footnote-60) مقایسه کرد. این تست معمولا به صورت دو دنباله[[60]](#footnote-61) در نظر گرفته می‏شود اما می‏توان آن را در شرایط خاص به صورت یک دنباله نیز بکار برد.

**مثال**

در پژوهشی، هدف بررسی همبستگی بین داده‏های بارندگی سالانه ایستگاه‏های سینوپتیک بندرانزلی و رشت می‏باشد. بدین منظور داده‏های 60 سال بارندگی سالانه (1959 تا 2018) این دو ایستگاه انتخاب شد. بر اساس داده‏های موجود، ضریب همبستگی ، 7076/0 به دست آمد. با توجه به اطلاعات فوق و در سطح معنی‏داری ، آیا فرضیه صفر (ضریب همبستگی برابر با صفر است و رابطه‏ای بین بارندگی سالانه دو ایستگاه بندرانزلی و رشت وجود ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
rm(list = ls())  
  
# load required functions  
source(file = "muFunc.R")  
  
# load required packages  
wants = c("pacman", "dplyr", "psych", "ggpubr", "ggExtra", "PerformanceAnalytics")  
has = wants %in% rownames(x = installed.packages())  
if(any(!has)) install.packages(wants[!has])  
pacman::p\_load(char = wants)  
  
# load required data  
data = read.csv(file = "data/Data\_Test\_12.csv", header = TRUE)  
# show data  
psych::headTail(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Year | Anzali | Rasht |
| ## | 1 | 1959 | 2408.2 | 1459.9 |
| ## | 2 | 1960 | 1971.7 | 1148.4 |
| ## | … | … | … | … |
| ## | 59 | 2017 | 1609 | 931.5 |
| ## | 60 | 2018 | 1584.5 | 948.4 |

**گام 2: بررسی محدودیت‏ها**

فرض می‏شود که مقادیر و از یک توزیع نرمال دو متغیره باشند: در این مرحله برای بررسی اینکه آیا داده‏های بارندگی سالانه دو ایستگاه بندرانزلی و رشت از توزیع نرمال پیروی می‏کنند یا خیر، از **«آزمون شاپیرو ویلک[[61]](#footnote-62)»** استفاده شده است. برای انجام این آزمون از تابع **shapiro.test** بسته **stats** استفاده شده است.

# assumption checking  
# 1. it is assumed that the x and y values originate from a bivariate normal distribution.  
# use shapiro-wilk normality test  
**shapiro.test**(x = data**$** Anzali)

## Shapiro-Wilk normality test  
## data: data$ Anzali  
## W = 0.9612, p-value = 0.0539

# use Shapiro-Wilk normality test  
**shapiro.test**(x = data**$** Rasht)

## Shapiro-Wilk normality test  
## data: data$ Rasht  
## W = 0.9704, p-value = 0.1523

با توجه به اینکه **p-مقدار** هر دو آزمون بیشتر از 05/0 می‏باشد، فرضیه صفر رد نشده و می‏توان نتیجه گرفت که داده‏های بارندگی سالانه دو ایستگاه بندرانزلی و رشت از توزیع نرمال پیروی می‏کنند. از طرفی می‏توان با رسم نمودار چندک چندک نیز نرمال بودن داده‏ها را مورد ارزیابی قرار داد. برای این منظور، از تابع **ggqqplot** موجود در بسته **ggpubr** استفاده شده است.

# normality plot  
ggpubr**::ggqqplot**(data = data**$**Anzali, ylab = "Anzali Prec. (mm)")  
ggpubr**::ggqqplot**(data = data**$**Rasht, ylab = "Rasht Prec. (mm)")

|  |  |
| --- | --- |
|  |  |

همانطور که دیده می‏شود، فرض نرمال بودن توزیع دو جامعه مورد تایید قرار می‏گیرد.

فرض می‏شود رابطه بین و به صورت خطی می‏باشد. به منظور تجسم رابطه بین این دو ایستگاه، نمودار نقطه‏ای آن‏ها با استفاده از تابع **chart.Correlation** از بسته **PerformanceAnalytics** و تابع **ggscatter** از بسته **ggpubr** رسم شده است.

# 2. that the relationship is linear  
PerformanceAnalytics**::chart.Correlation**(R = data[, **c**("Anzali", "Rasht")])  
  
# visualize data using scatter plots  
data **%>%** ggpubr**::ggscatter**(x = "Anzali", y = "Rasht",  
 add = "reg.line", conf.int = TRUE,  
 xlab = "Anzali Prec. (mm)",  
 ylab = "Rasht Prec. (mm)") **%>%**  
 ggExtra**::ggMarginal**(type = "boxplot", fill = "gray90")

|  |  |
| --- | --- |
|  |  |

همانطور که مشخص است رابطه بین بارندگی سالانه دو ایستگاه بندرانزلی و رشت به صورت خطی می‏باشد.

**گام 3: محاسبه آماره آزمون**

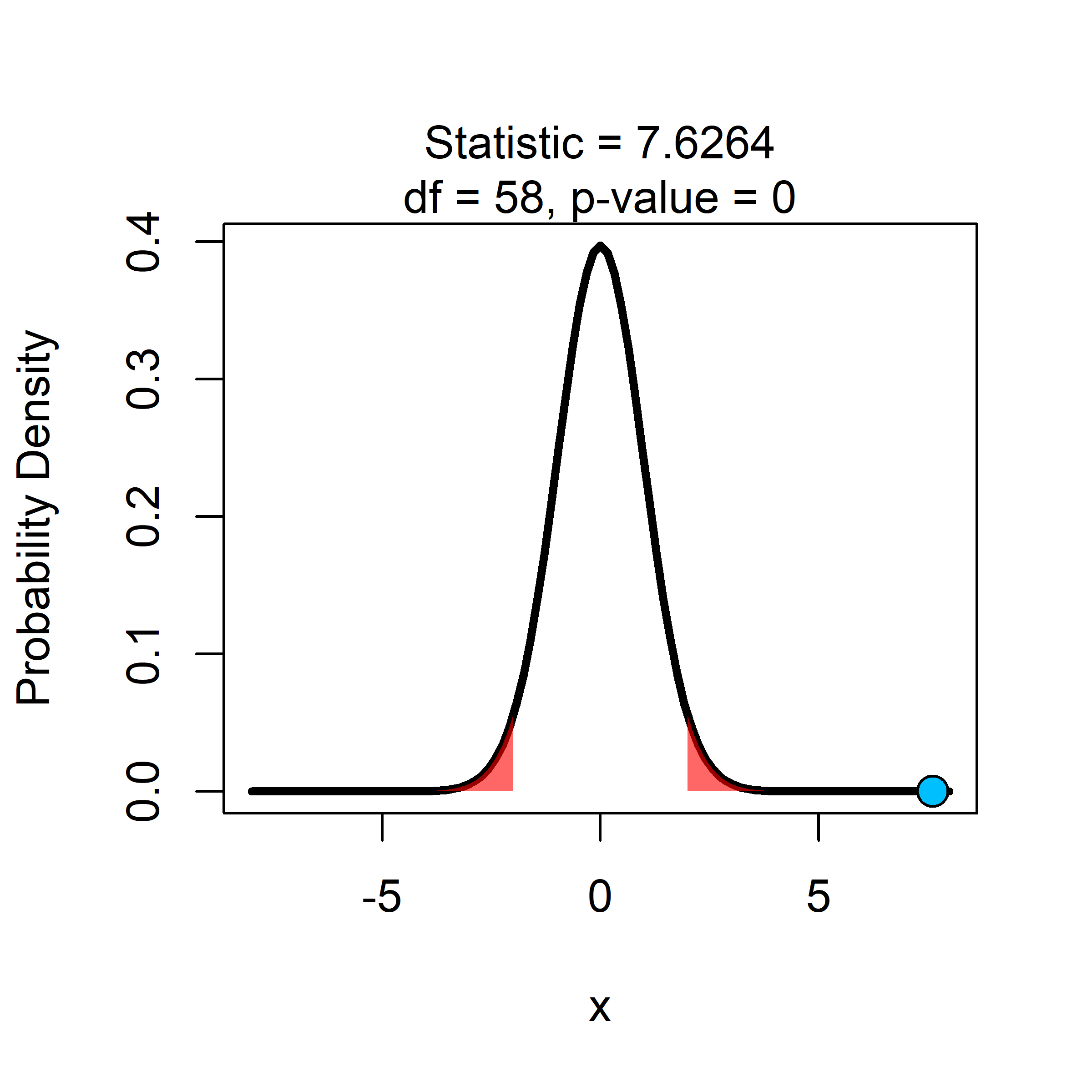
محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی تفاوت معنی‏داری آماری ضریب همبستگی یک نمونه با عدد صفر، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
x = data**$**Anzali  
xm = **mean**(x = x)  
y = data**$**Rasht  
ym = **mean**(x = y)  
n = **length**(x)  
  
# correlation coefficient (r)  
r.numerator = **sum**((x **-** xm) **\*** (y **-** ym))  
r.denominator = **sqrt**(**sum**((x **-** xm) **^** 2) **\*** **sum**((y **-** ym) **^** 2))  
r = r.numerator **/** r.denominator  
  
# test statistic  
t = (r **/** (**sqrt**(1 **-** (r **^** 2)))) **\*** **sqrt**(n **-** 2)  
  
# degrees of freedom  
df = n **-** 2  
# compute the critical values at 0.05 significance level (two tailed)  
alpha = 0.05  
two.sided = TRUE  
Tcritical = **qt**(p = (1 **-** alpha**/**2), df = df)  
# p-value (two tailed)  
p\_value = **pt**(q = t, df = df, lower.tail = FALSE)  
**if** (two.sided) {  
 p\_value = 2 **\*** p\_value  
 **if** (p\_value **>** 1) { p\_value = 2 **-** p\_value }  
}  
**cat**("Test Statistic:", t, "\n")  
**cat**("Degrees of Freedom:", df, "\n")  
**cat**("Critical Values:", **c**(**-**Tcritical, Tcritical), "\n")  
**cat**("P-value:", p\_value, "\n")

## Test Statistic: 7.6264  
## Degrees of Freedom: 58  
## Critical Values: -2.0017 2.0017  
## P-value: 2.5956e-10

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع استیودنت با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "t", statistic = t, df = df, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله ، مقدار بحرانی در برابر با 2 است (جدول شماره 2). مقدار محاسبه شده برای بیشتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که ضریب همبستگی رابطه بین بارندگی سالانه ایستگاه بندرانزلی و رشت مخالف با صفر بوده و یک ارتباط خطی بین این دو ایستگاه برقرار است.

در روش دوم، برای بررسی تفاوت معنی‏داری آماری ضریب همبستگی یک نمونه با عدد صفر، از تابع **cor.test** موجود در بسته **stats** استفاده شده است.

# method 2  
# use "cor.test" function from "stats" package  
x = data**$**Anzali  
y = data**$**Rasht  
**cor.test**(x = x, y = y, alternative = "two.sided", conf.level = 0.95)

## Pearson's product-moment correlation  
## data: x and y  
## t = 7.6264, df = 58, p-value = 2.596e-10  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5530 0.81507  
## sample estimates: cor: 0.7076

# آزمون 13: آزمون [[62]](#footnote-63) یک ضریب همبستگی[[63]](#footnote-64)

**هدف**

بررسی معنی‏دار بودن تفاوت بین ضریب همبستگی و یک مقدار مشخص از ضریب همبستگی جامعه[[64]](#footnote-65) ().

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. مقادیر و از توزیع‏های نرمال بر گرفته شده باشند.
2. واریانس مقادیر مستقل از مقادیر باشد.
3. رابطه بین و خطی باشد.

در صورتی که شرایط فوق محقق نشده باشد، باید از آزمون همبستگی رتبه‏ای کندال[[65]](#footnote-66) استفاده کرد.

**روش**

با ضریب همبستگی نمونه‏ای[[66]](#footnote-67) ، و استفاده از تبدیل فیشر[[67]](#footnote-68) می‏توان نوشت:

توزیع با میانگین و انحراف استاندارد تقریباً نرمال می‏باشد، که:

آماره آزمون با استفاده از معادله زیر محاسبه می‏گردد:

**مثال**

در منابع، ضریب همبستگی بین خاک و تولید یک نوع علوفه خشک بر حسب تن در هکتار، 75/0 گزارش شده است. به منظور بررسی این موضوع، داده‏های خاک و تولید یک نوع علوفه خشک 20 مزرعه جمع‏آوری و ضریب همبستگی داده‏های این 20 مزرعه، 81/0 بدست آمد. در سطح معنی‏داری ، آیا فرضیه صفر (ضریب همبستگی خاک و تولید یک نوع علوفه خشک تفاوتی با عدد 75/0 ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "ggExtra",   
 "PerformanceAnalytics", "miscor")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_13.csv", header = TRUE)  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Year | Mashhad | Torbat |
| ## | 1 | 1959 | 234.5 | 258.7 |
| ## | 2 | 1960 | 174.5 | 263.7 |
| ## | … | … | … | … |
| ## | 59 | 2017 | 238.1 | 180.9 |
| ## | 60 | 2018 | 237.4 | 197.3 |

**گام 2: بررسی محدودیت‏ها**

فرض می‏شود مقادیر و از توزیع‏های نرمال بر گرفته شده باشند: در این مرحله برای بررسی اینکه آیا داده‏های بارندگی سالانه دو ایستگاه مشهد و تربت‏حیدریه از توزیع نرمال پیروی می‏کنند یا خیر، از **«آزمون شاپیرو ویلک[[68]](#footnote-69)»** استفاده شده است. برای انجام این آزمون از تابع **shapiro.test** بسته **stats** استفاده شده است.

# assumption checking  
# 1. the x and y values originate from normal distributions.  
# use shapiro-wilk normality test  
**shapiro.test**(x = data**$**Mashhad)

## Shapiro-Wilk normality test  
## data: data$Mashhad  
## W = 0.9723, p-value = 0.1897

# use Shapiro-Wilk normality test  
**shapiro.test**(x = data**$**Torbat)

## Shapiro-Wilk normality test  
## data: data$Torbat  
## W = 0.9664, p-value = 0.0974

با توجه به اینکه **p-مقدار** هر دو آزمون بیشتر از 05/0 می‏باشد، فرضیه صفر رد نشده و می‏توان نتیجه گرفت که داده‏های بارندگی سالانه دو ایستگاه مشهد و سبزوار از توزیع نرمال پیروی می‏کنند. از طرفی می‏توان با رسم نمودار چندک چندک نیز نرمال بودن داده‏ها را مورد ارزیابی قرار داد. برای این منظور، از تابع **ggqqplot** موجود در بسته **ggpubr** استفاده شده است.

# normality plot  
ggpubr**::ggqqplot**(data = data**$**Mashhad, ylab = "Mashhad Prec. (mm)")  
ggpubr**::ggqqplot**(data = data**$**Torbat, ylab = "Torbat Prec. (mm)")

|  |  |
| --- | --- |
|  |  |

همانطور که دیده می‏شود، فرض نرمال بودن توزیع دو جامعه مورد تایید قرار می‏گیرد.

فرض می‏شود که واریانس مقادیر مستقل از مقادیر است.

فرض می‏شود رابطه بین و به صورت خطی می‏باشد. به منظور تجسم رابطه بین این دو ایستگاه، نمودار پراکندگی[[69]](#footnote-70) آن‏ها با استفاده از تابع **chart.Correlation** از بسته **PerformanceAnalytics** و تابع **ggscatter** از بسته **ggpubr** رسم شده است. نمودار پراکندگی یک از ساده‏ترین روش‏ها برای بررسی همبستگی و ارتباط بین دو متغیر است. این نمودار، نوع و جهت رابطه را به طور بصری ارائه می‏دهد و می‏توان با مشاهده نمودار از نوع رابطه بین دو متغیر و جهت (خطی یا غیر خطی و مثبت یا منفی) و شدت رابطه آگاهی تقریبی یافت.

# 2. that the relationship is linear  
PerformanceAnalytics**::chart.Correlation**(R = data[, **c**("Mashhad", "Torbat")])  
  
# visualize data using scatter plots  
data **%>%** ggpubr**::ggscatter**(x = "Mashhad", y = "Torbat",  
 add = "reg.line", conf.int = TRUE,  
 xlab = "Mashhad Prec. (mm)",  
 ylab = "Torbat Prec. (mm)") **%>%**  
 ggExtra**::ggMarginal**(type = "boxplot", fill = "gray90")

|  |  |
| --- | --- |
|  |  |

همانطور که مشخص است رابطه بین بارندگی سالانه دو ایستگاه مشهد و تربت‏حیدریه به صورت خطی می‏باشد.

**گام 3: محاسبه آماره آزمون**

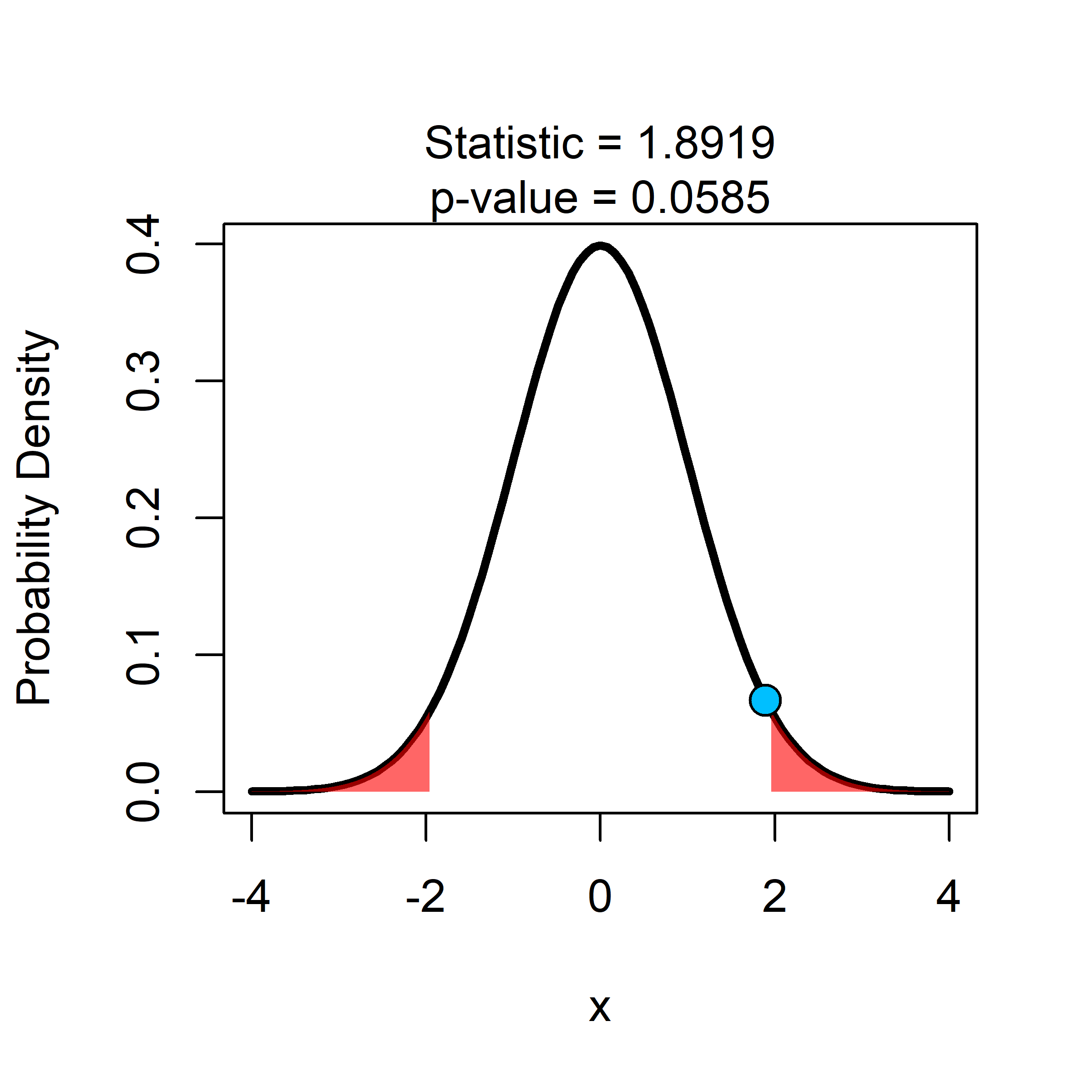
محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین ضریب همبستگی و یک مقدار مشخص از ضریب همبستگی جامعه، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
data = data **%>%** **filter**(Year **>=** 2004)  
x = data**$**Mashhad  
x.m = **mean**(x = x)  
y = data**$**Torbat  
y.m = **mean**(x = y)  
n = **length**(x)  
  
# specified value of correlation coefficient  
rho0 = 0.69  
  
# correlation coefficient (r)  
r.numerator = **sum**((x **-** x.m) **\*** (y **-** y.m))  
r.denominator = **sqrt**(**sum**((x **-** x.m) **^** 2) **\*** **sum**((y **-** y.m) **^** 2))  
r = r.numerator **/** r.denominator  
  
  
# fisher Z-transformation  
Z1 = 0.5 **\*** **log**((1 **+** r) **/** (1 **-** r))  
  
# the distribution of Z1 is approximately normal with mean  
Z1.m = 0.5 **\*** **log**((1 **+** rho0) **/** (1 **-** rho0))  
  
# and standard deviation:  
Z1.sd = 1 **/** **sqrt**(n **-** 3)  
  
# Test Statistic  
Z = (Z1 **-** Z1.m) **/** Z1.sd  
  
# compute the critical values at 0.05 significance level (two tailed)  
alpha = 0.05  
two.sided = TRUE  
Zcritical = **qnorm**(p = (1 **-** alpha**/**2))  
  
# p-value (two tailed)  
p\_value = **pnorm**(q = Z, lower.tail = FALSE)  
  
**if** (two.sided) {  
 p\_value = 2 **\*** p\_value  
 **if** (p\_value **>** 1) { p\_value = 2 **-** p\_value }  
}  
**cat**("Test Statistic:", t, "\n")  
**cat**("Critical Values:", **c**(**-**Tcritical, Tcritical), "\n")  
**cat**("P-value:", p\_value, "\n")

## Test Statistic: 1.8919  
## Critical Values: -1.96 1.96  
## P-value: 0.0585

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع نرمال استاندارد با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "norm", statistic = Z, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله ، مقدار بحرانی در برابر با 96/1 است (جدول شماره 1). مقدار محاسبه شده برای کمتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. بدان معنا است که می‏توان یک ضریب همبستگی واحد، بین داده‏های بارش مجموع سالانه دو ایستگاه سینوپتیک مشهد و تربت حیدریه بین سال‏های 1959 تا 2018 ارائه کرد.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین ضریب همبستگی و یک مقدار مشخص از ضریب همبستگی جامعه، از تابع **test.cor** موجود در بسته **miscor** استفاده شده است.

# method 2  
# setting initial parameter values  
r = **cor**(x = data**$**Mashhad, y = data**$**Torbat)  
n = **nrow**(x = data)  
rho0 = 0.69  
  
# use "test.cor" function from "miscor" package  
miscor**::test.cor**(r = r, n = n, rho0 = rho0, conf.level = 0.90, reduced = TRUE)

## Statistical test for the product-moment correlation coefficient  
## H0: rho == 0.69 versus H1: rho != 0.69  
## z = 1.892, p-value = 0.0585  
## Sample estimate r: 0.884  
## Two-sided 90% CI: [0.726, 0.953]

# آزمون 14: آزمون [[70]](#footnote-71) برای دو ضریب همبستگی[[71]](#footnote-72)

**هدف**

بررسی معنی‏دار بودن تفاوت بین ضرایب همبستگی برای یک جفت از متغیرهای انتخاب شده از دو نمونه متفاوت و تفاوت بین دو مقدار مشخص و .

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. مقادیر و از توزیع‏های نرمال بر گرفته شده باشند.
2. واریانس مقادیر مستقل از مقادیر باشد.
3. رابطه بین و خطی باشد.

**روش**

ضریب همبستگی نمونه‏ای[[72]](#footnote-73) ، همانند آنچه در آزمون 13 بیان شد، محاسبه شده و برای نمونه اول:

توزیع با میانگین و انحراف استاندارد تقریباً نرمال می‏باشد، که:

و همچنین برای نمونه دوم:

توزیع با میانگین و انحراف استاندارد تقریباً نرمال می‏باشد، که:

آماره آزمون با استفاده از معادله زیر محاسبه می‏گردد:

که در آن:

می‏باشد. دارای توزیع نرمال با میانگین صفر و واریانس 1 می‏باشد.

**مثال**

در پژوهشی روی 30 نمونه تصادفی از دو رقم لوبیا (بوفی و کندی)، ضریب همبستگی بین درصد سوختگی برگ‏ها و مقدار سدیم آب آبیاری بررسی شده است. ضریب همبستگی برای رقم بوفی و کندی به ترتیب 46/0 و 74/0 بدست آمد. در سطح معنی‏داری ، آیا فرضیه صفر (ضریب همبستگی بین درصد سوختگی برگ‏ها و مقدار سدیم آب آبیاری دو رقم لوبیا با یکدیگر برابر است) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "ggExtra",  
 "PerformanceAnalytics", "miscor", "cocor")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_14.csv", header = TRUE)  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## |  | Year | T.mashhad | SSH.mashhad | T.sabzevar | SSH.sabzevar |
| ## | 1 | 1979 | 14.42 | 8.27 | 18.47 | 8.06 |
| ## | 2 | 1980 | 13.87 | 7.64 | 18.68 | 7.69 |
| ## | … | … | … | … | … | … |
| ## | 39 | 2017 | 16.28 | 8.61 | 18.8 | 9.08 |
| ## | 40 | 2018 | 16.51 | 7.91 | 19.13 | 8.39 |

**گام 2: بررسی محدودیت‏ها**

فرض می‏شود مقادیر و از توزیع‏های نرمال بر گرفته شده باشند: در این مرحله برای بررسی اینکه آیا داده‏های متوسط دمای سالانه و متوسط ساعات آفتابی سالانه دو ایستگاه مشهد و سبزوار از توزیع نرمال پیروی می‏کنند یا خیر، از **«آزمون شاپیرو ویلک[[73]](#footnote-74)»** استفاده شده است. برای انجام این آزمون از تابع **shapiro.test** بسته **stats** استفاده شده است.

ایستگاه مشهد

# assumption checking: mashhad station  
# 1. the x and y values originate from normal distributions.  
# use shapiro-wilk normality test  
**shapiro.test**(x = data**$**T.mashhad)

## Shapiro-Wilk normality test  
## data: data$T.mashhad  
## W = 0.9687, p-value = 0.3271

# use Shapiro-Wilk normality test  
**shapiro.test**(x = data**$**SSH.mashhad)

## Shapiro-Wilk normality test  
## data: data$SSH.mashhad  
## W = 0.9729, p-value = 0.4416

ایستگاه سبزوار

# use shapiro-wilk normality test  
**shapiro.test**(x = data**$**T.sabzevar)

## Shapiro-Wilk normality test  
## data: data$T.sabzevar  
## W = 0.9713, p-value = 0.3949

# use Shapiro-Wilk normality test  
**shapiro.test**(x = data**$**SSH.sabzevar)

## Shapiro-Wilk normality test  
## data: data$SSH.sabzevar  
## W = 0.9684, p-value = 0.3196

با توجه به اینکه **p-مقدار** هر چهار آزمون بیشتر از 05/0 می‏باشد، فرضیه صفر رد نشده و می‏توان نتیجه گرفت که داده‏های متوسط دمای سالانه و متوسط ساعات آفتابی سالانه دو ایستگاه مشهد و سبزوار از توزیع نرمال پیروی می‏کنند. از طرفی می‏توان با رسم نمودار چندک چندک نیز نرمال بودن داده‏ها را مورد ارزیابی قرار داد. برای این منظور، از تابع **ggqqplot** موجود در بسته **ggpubr** استفاده شده است.

ایستگاه مشهد

# normality plot  
**ggqqplot**(data = data**$**T.mashhad, ylab = "Mashhad Temp. (C)")  
**ggqqplot**(data = data**$**SSH.mashhad, ylab = "Mashhad Sunshine (h)")

|  |  |
| --- | --- |
|  |  |

ایستگاه سبزوار

# normality plot  
**ggqqplot**(data = data**$**T.sabzevar, ylab = "Sabzevar Temp. (C)")  
**ggqqplot**(data = data**$**SSH.sabzevar, ylab = "Sabzevar Sunshine (h)")

|  |  |
| --- | --- |
|  |  |

همانطور که دیده می‏شود، فرض نرمال بودن توزیع دو جامعه مورد تایید قرار می‏گیرد.

فرض می‏شود که واریانس مقادیر مستقل از مقادیر است.

فرض می‏شود رابطه بین و به صورت خطی می‏باشد. به منظور تجسم رابطه بین داده‏های متوسط دمای سالانه و متوسط ساعات آفتابی سالانه در هر دو ایستگاه، نمودار پراکندگی[[74]](#footnote-75) آن‏ها با استفاده از تابع **chart.Correlation** از بسته **PerformanceAnalytics** و تابع **ggscatter** از بسته **ggpubr** رسم شده است.

ایستگاه مشهد

# 2. the relationships are linear.  
PerformanceAnalytics**::chart.Correlation**(R = data[, **c**("T.mashhad", "SSH.mashhad")])  
  
# visualize data using scatter plots  
data **%>%** ggpubr**::ggscatter**(x = "T.mashhad", y = "SSH.mashhad",  
 add = "reg.line", conf.int = TRUE,  
 xlab = "Mashhad Temp. (C)",  
 ylab = "Mashhad Sunshine (h)") **%>%**  
 ggExtra**::ggMarginal**(type = "boxplot", fill = "gray90")

|  |  |
| --- | --- |
|  |  |

ایستگاه سبزوار

# 2. the relationships are linear.  
PerformanceAnalytics**::chart.Correlation**(R = data[, **c**("T.sabzevar", "SSH.sabzevar")])  
  
# visualize data using scatter plots  
data **%>%** ggpubr**::ggscatter**(x = "T.sabzevar", y = "SSH.sabzevar",  
 add = "reg.line", conf.int = TRUE,  
 xlab = "Sabzevar Temp. (C)",  
 ylab = "Sabzevar Sunshine (h)") **%>%**  
 ggExtra**::ggMarginal**(type = "boxplot", fill = "gray90")

|  |  |
| --- | --- |
|  |  |

همانطور که مشخص است رابطه بین داده‏های متوسط دمای سالانه و متوسط ساعات آفتابی سالانه در هر دو ایستگاه به صورت خطی می‏باشد.

**گام 3: محاسبه آماره آزمون**

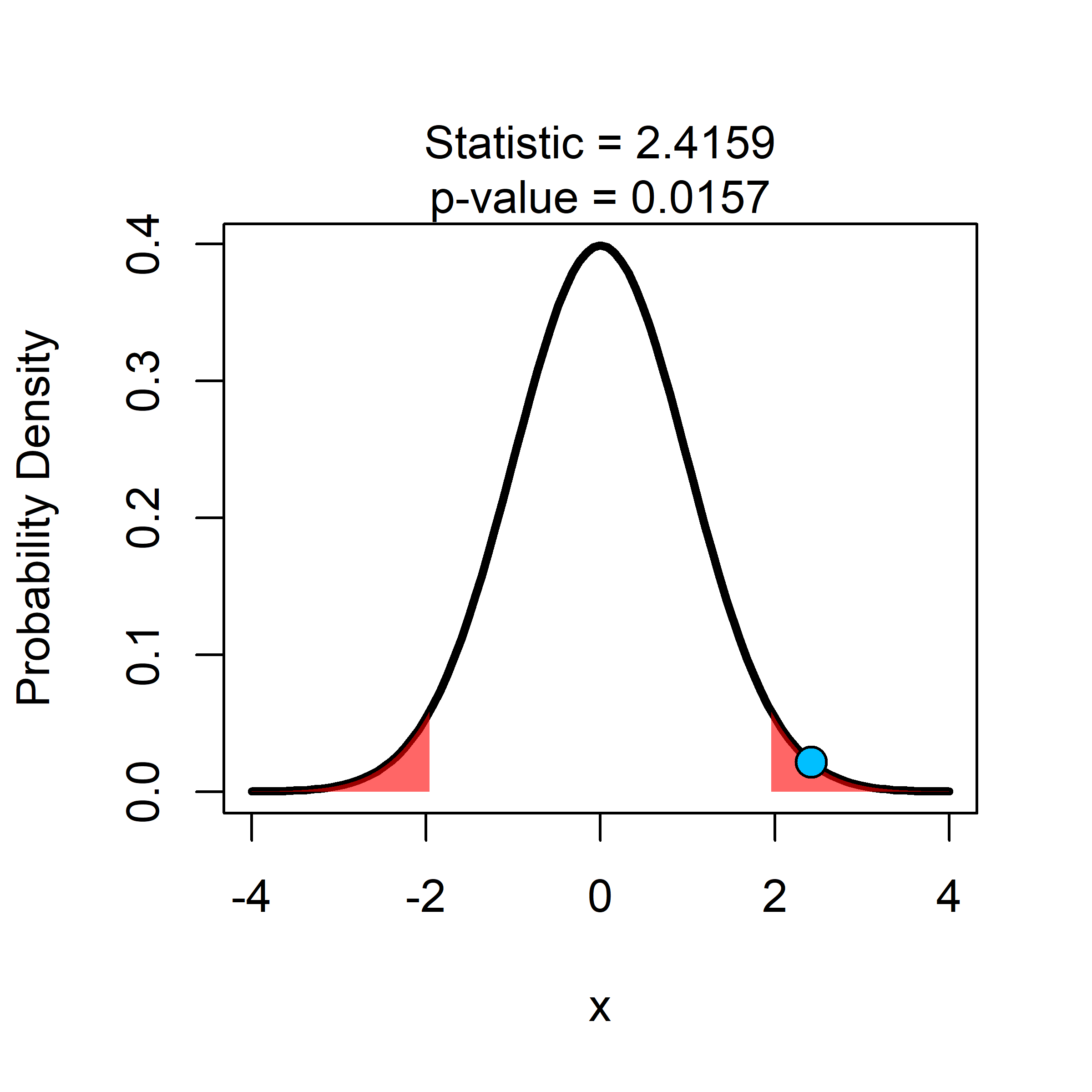
محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین ضرایب همبستگی برای یک جفت از متغیرهای انتخاب شده از دو نمونه متفاوت، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# mashhad station  
# setting initial parameter values  
x1 = data**$**T.mashhad  
x1.m = **mean**(x = x1)  
y1 = data**$**SSH.mashhad  
y1.m = **mean**(x = y1)  
n1 = **length**(x1)  
  
# correlation coefficient (r)  
r1.numerator = **sum**((x1 **-** x1.m) **\*** (y1 **-** y1.m))  
r1.denominator = **sqrt**(**sum**((x1 **-** x1.m) **^** 2) **\*** **sum**((y1 **-** y1.m) **^** 2))  
r1 = r1.numerator **/** r1.denominator  
  
# fisher Z-transformation  
Z1.mashhad = 0.5 **\*** **log**((1 **+** r1) **/** (1 **-** r1))  
  
# standard deviation  
Z1.sd.mashhad = 1 **/** **sqrt**(n1 **-** 3)  
  
# sabzevar station  
# setting initial parameter values  
x2 = data**$**T.sabzevar  
x2.m = **mean**(x = x2)  
y2 = data**$**SSH.sabzevar  
y2.m = **mean**(x = y2)  
n2 = **length**(x2)  
  
# correlation coefficient (r)  
r2.numerator = **sum**((x2 **-** x2.m) **\*** (y2 **-** y2.m))  
r2.denominator = **sqrt**(**sum**((x2 **-** x2.m) **^** 2) **\*** **sum**((y2 **-** y2.m) **^** 2))  
r2 = r2.numerator **/** r2.denominator  
  
# fisher Z-transformation  
Z1.sabzevar = 0.5 **\*** **log**((1 **+** r2) **/** (1 **-** r2))  
# standard deviation  
Z1.sd.sabzevar = 1 **/** **sqrt**(n2 **-** 3)  
# test Statistic  
sigma = **sqrt**((Z1.sd.mashhad **^** 2) **+** (Z1.sd.sabzevar **^** 2))  
Z = (Z1.mashhad **-** Z1.sabzevar) **/** sigma  
# compute the critical values at 0.05 significance level (two tailed)  
alpha = 0.05  
two.sided = TRUE  
Zcritical = **qnorm**(p = (1 **-** alpha**/**2))  
# p-value (two tailed)  
p\_value = **pnorm**(q = Z, lower.tail = FALSE)  
**if** (two.sided) {  
 p\_value = 2 **\*** p\_value  
 **if** (p\_value **>** 1) { p\_value = 2 **-** p\_value }  
}  
**cat**("Test Statistic:", t, "\n")  
**cat**("Critical Values:", **c**(**-**Tcritical, Tcritical), "\n")  
**cat**("P-value:", p\_value, "\n")

## Test Statistic: 2.4159  
## Critical Values: -1.96 1.96  
## P-value: 0.0157

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع نرمال استاندارد با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "norm", statistic = Z, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله ، مقدار بحرانی در برابر با 96/1 است (جدول شماره 1). مقدار محاسبه شده برای بیشتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که ضرایب همبستگی به دست آمده در دو ایستگاه از نظر آماری با یکدیگر تفاوت دارند.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین ضریب همبستگی و یک مقدار مشخص از ضریب همبستگی جامعه، از تابع **test.cor** موجود در بسته **miscor** استفاده شده است.

# method 2  
r1 = **cor.test**(x = data**$**T.mashhad, y = data**$**SSH.mashhad,  
 alternative = "two.sided", conf.level = 0.95)  
  
r2 = **cor.test**(x = data**$**T.sabzevar, y = data**$**SSH.sabzevar,  
 alternative = "two.sided", conf.level = 0.95)  
  
# use "cocor.indep.groups" function from "cocor" package  
**cocor.indep.groups**(r1.jk = r1**$**estimate, n1 = n1,  
 r2.hm = r2**$**estimate, n2 = n2,  
 alternative = "two.sided")

## Results of a comparison of two correlations based on independent groups  
## Comparison between r1.jk = 0.6997 and r2.hm = 0.2959  
## Difference: r1.jk - r2.hm = 0.4038  
## Group sizes: n1 = 40, n2 = 40  
## Null hypothesis: r1.jk is equal to r2.hm  
## Alternative hypothesis: r1.jk is not equal to r2.hm (two-sided)  
## Alpha: 0.05  
## fisher1925: Fisher's z (1925)  
## z = 2.4159, p-value = 0.0157  
## Null hypothesis rejected

# آزمون 15: آزمون [[75]](#footnote-76) برای واریانس یک جامعه

**هدف**

بررسی معنی‏دار بودن تفاوت بین واریانس یک نمونه و واریانس مفروض یک جامعه .

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. فرض می‏شود که نمونه از یک جامعه با توزیع نرمال استخراج شده باشد.

**روش**

نمونه‏ای تصادفی با حجم و مقادیر ، و ... و در نظر گرفته شده، و روابط زیر محاسبه می‏گردد:

فرضیه صفر آزمون که واریانس جامعه با مقدار برابر است با آماره آزمون زیر مورد بررسی قرار می‏گردد:

که می‏توان آماره آزمون را با توزیع خی‏دو[[76]](#footnote-77) با درجه‏های آزادی[[77]](#footnote-78) مقایسه کرد. آزمون می‏تواند به هر یک از دو صورت یک دنباله[[78]](#footnote-79) یا دو دنباله[[79]](#footnote-80) انجام گیرد.

**مثال**

شرکتی دانش بنیان بذر گوجه فرنگی گلخانه‏ای تولید کرده و مدعی است که در شرایط مشخص (دمای مشخص گلخانه، تهویه مناسب و ...)، متوسط وزن هر میوه آن 190 گرم با انحراف معیار 15± گرم می‏باشد. برای بررسی این موضوع از گلخانه‏ای که تمام شرایط مشخص شده را رعایت کرده است در موقع برداشت محصول، به صورت کاملا تصادفی 100 عدد گوجه فرنگی انتخاب شد. میانگین و واریانس نمونه‏های انتخابی به ترتیب برابر است با 125/191 گرم و 549/295 می‏باشد. با قبول فرض نرمال بودن توزیع جامعه، در سطح معنی‏داری ، آیا فرضیه صفر (ادعای شرکت دانش بنیان در مورد انحراف معیار 15± گرم وزن میوه‏ها) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggplot2", "ggExtra", "TeachingDemos")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_15.csv", header = TRUE)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

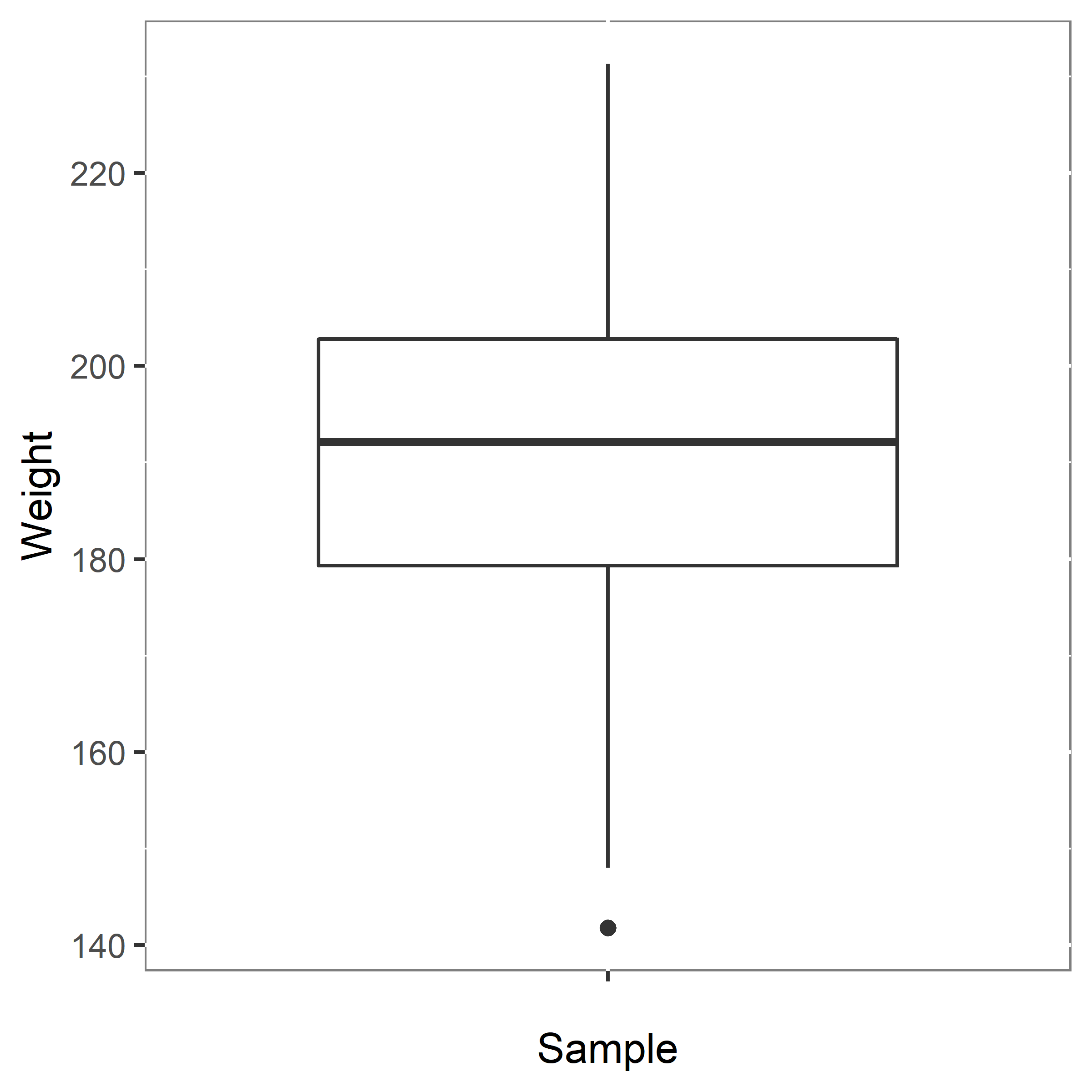
|  |  |  |  |
| --- | --- | --- | --- |
| ## |  | Sample | Weight |
| ## | 1 | 1 | 188.8 |
| ## | 2 | 2 | 159.2 |
| ## | … | … | … |
| ## | 99 | 99 | 148 |
| ## | 100 | 100 | 181.3 |

**گام 2: بررسی محدودیت‏ها**

1- فرض می‏شود که نمونه از یک جامعه با توزیع نرمال استخراج شده باشد.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها رسم شده است. در این نمودار از جعبه‏ای برای نمایش فاصله بین چارک اول و سوم استفاده می‏شود و خطی در داخل جعبه، میانه (چارک دوم) را مشخص می‏کند. خارج از جعبه حداقل و حداکثر مقدار داده را نیز مشخص می‌نمایند. گهگاه نمونه‏های خارج از محدوده نیز به صورت نقاطی نشان داده می‏شوند. برای این منظور، از تابع **ggboxplot** موجود در بسته **ggpubr** استفاده شده است.

# visualizing samples with box plots  
data **%>%** **ggplot**(**aes**(x = "", y = Weight)) **+**  
 **geom\_boxplot**() **+** **xlab**(label = "Sample") **+** **ylab**("Weight") **+**  
 **theme**(panel.background = **element\_rect**(fill = "white", colour = "grey50"))



**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین واریانس یک نمونه و واریانس مفروض یک جامعه، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
x = data**$**Weight  
n = **length**(x = x)  
x.m = **mean**(x = x)  
x.v = (**sum**((x **-** x.m) **^** 2)) **/** (n **-** 1)  
  
# population variance  
sigmasq = 225  
  
# test statistic  
chi2 = ((n **-** 1) **\*** x.v) **/** (sigmasq)  
  
# degrees of freedom  
df = n **-** 1  
  
  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
**if** (alternative **!=** "two.sided") {  
 lower.critical = **qchisq**(p = (1 **-** alpha), df = df, lower.tail = FALSE)  
 upper.critical = **qchisq**(p = (1 **-** alpha), df = df, lower.tail = TRUE)  
} **else** {  
 lower.critical = **qchisq**(p = (1 **-** (alpha **/** 2)), df = df, lower.tail = FALSE)  
 upper.critical = **qchisq**(p = (1 **-** (alpha **/** 2)), df = df, lower.tail = TRUE)  
}  
  
p\_value = **pchisq**(q = chi2, df = df, lower.tail = TRUE)  
p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value)  
 **cat**("Test Statistic: ", t, "\n")  
**cat**("Degrees of Freedom:", (n\_ADF **+** n\_Nelson **-** 2), "\n")  
**cat**("Critical Values: ", **c**(**-**Tcritical, Tcritical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: 130.0414  
## Degrees of Freedom: 99  
## Critical Values: 73.361 128.422  
## P-value: 0.0397

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "chisq", statistic = chi2, df = df, alpha = 0.05,  
 alternative = "two.sided", from = 50, to = 150)



همانطور که مشخص است، در آزمون دو دنباله ، مقادیر بحرانی در برابر با 36/73 و 42/128 است (جدول شماره 5). مقدار محاسبه شده برای بیشتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که ادعای شرکت دانش بنیان رد می‏شود و تغییرپذیری وزن میوه‏ها بیشتر از 15 گرم می‏باشد.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین واریانس یک نمونه و واریانس مفروض یک جامعه ، از تابع **t.test** موجود در بسته **stats** استفاده شده است.

# method 2  
# setting initial parameter values  
x = data**$**Weight  
sigmasq = 225  
  
# use "sigma.test" function from "TeachingDemos" package  
**sigma.test**(x = x, sigmasq = sigmasq, alternative = "two.sided", conf.level = 0.95)

## One sample Chi-squared test for variance  
## data: x  
## X-squared = 130.04, df = 99, p-value = 0.0397  
## alternative hypothesis: true variance is not equal to 225  
## 95 percent confidence interval:  
## 227.8372 398.8396  
## sample estimates:  
## var of x: 295.5486

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 16: آزمون [[80]](#footnote-81) برای واریانس‏های دو جامعه (آزمون نسبت واریانس‏ها[[81]](#footnote-82))

**هدف**

بررسی معنی‏دار بودن تفاوت بین واریانس‏های دو جامعه.

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. دو جامعه باید از توزیع نرمال پیروی کنند. (لزومی به برابری میانگین دو جامعه نمی‏باشد.)

**روش**

از جامعه اول، نمونه‏های تصادفی با حجم و با مقادیر ، ، ...، و از جامعه دوم، نمونه‏های تصادفی با حجم و با مقادیر ، ، ...، انتخاب و مقادیر زیر محاسبه می‏گردند:

در نتیجه آماره آزمون برابر است با:

با در نظر گرفتن فرضیه صفر، واریانس دو جامعه با هم برابر است، می‏توان آماره آزمون را با توزیع [[82]](#footnote-83) و با درجه‏های آزادی[[83]](#footnote-84) مقایسه کرد. آزمون می‏تواند به هر یک از دو صورت یک دنباله[[84]](#footnote-85) یا دو دنباله[[85]](#footnote-86) انجام گیرد.

**مثال**

مجموعه داده گل زنبق[[86]](#footnote-87)، یا مجموعه داده زنبق فیشر یک مجموعه داده چند متغیره است که توسط رانلد فیشر در سال ۱۹۳۶ معرفی شد. این مجموعه داده 4 ویژگی از 3 نوع گونه به نام‏های setosa،versicolor و virginica را شامل می‏شود. 4 ویژگی استفاده شده عبارت‏اند از طول و عرض گلبرگ[[87]](#footnote-88) و طول و عرض کاسبرگ[[88]](#footnote-89). از هر نوع گل50 نمونه وجود دارد بنابراین داده‏های این مجموعه را می‏توان شامل 150 مشاهده دانست که هر کدام از مشاهدات شامل 4 ویژگی می‏باشند. محققی در نظر دارد برابری واریانس‏های عرض کاسبرگ دو گونه setosa و virginica را مورد آزمون قرار دهد. واریانس دو گونه setosa و virginica به ترتیب برابر با 14/0 و 10/0 سانتیمتر می‏باشد. در سطح معنی‏داری ، آیا فرضیه صفر (تفاوتی بین واریانس‏های عرض کاسبرگ این دو گونه از گل زنبق وجود ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
# load required functions  
**source**(file = "muFunc.R")  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_16.csv", header = TRUE)  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

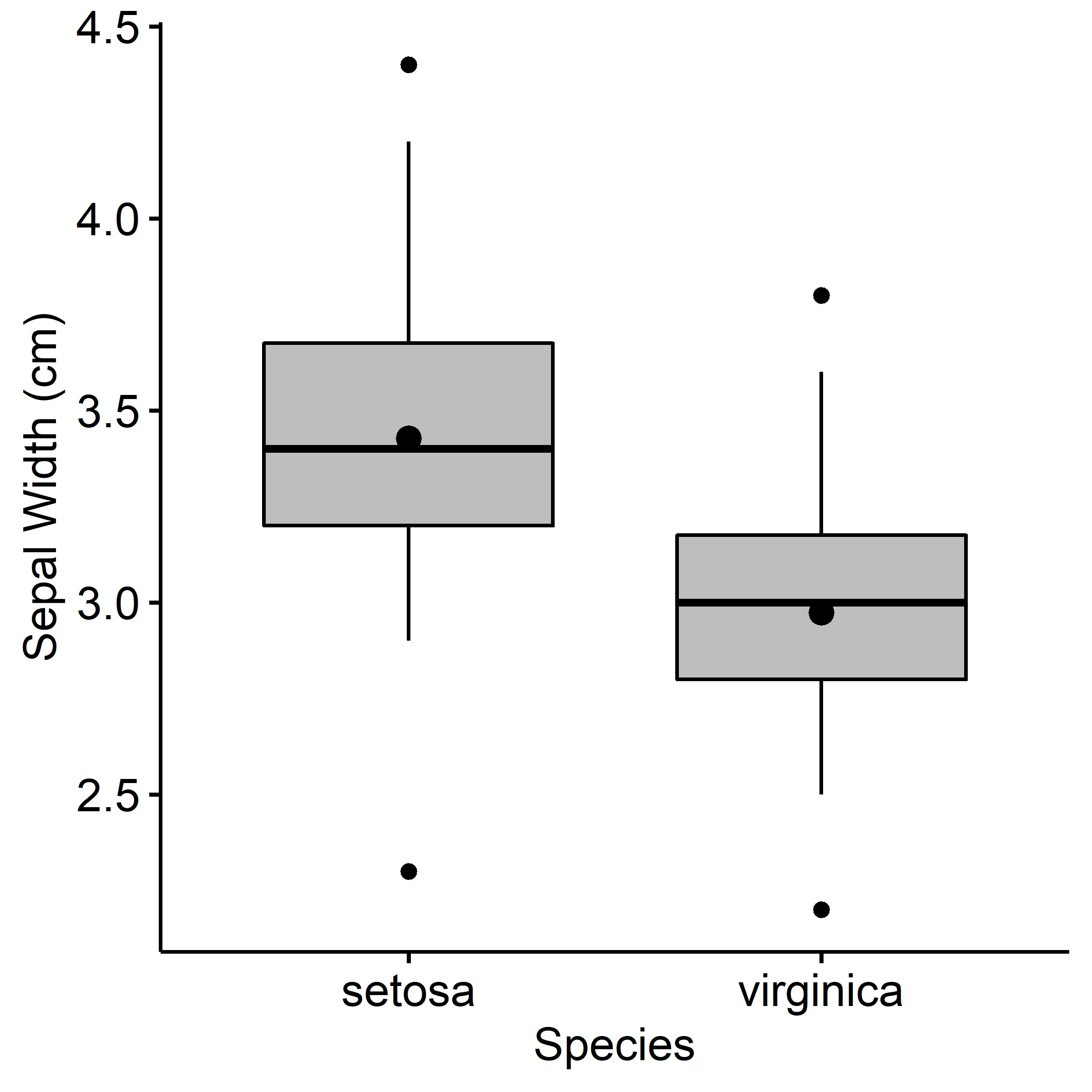
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## |  | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
| ## | 1 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| ## | 2 | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| ## | … | … | … | … | … | … |
| ## | 149 | 6.2 | 3.4 | 5.4 | 2.3 | virginica |
| ## | 150 | 5.9 | 3.0 | 5.1 | 1.8 | virginica |

**گام 2: بررسی محدودیت‏ها**

1- فرض می‏شود هر دو جامعه به صورت نرمال توزیع شده باشند.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها رسم شده است. برای این منظور، از تابع **ggboxplot** موجود در بسته **ggpubr** استفاده شده است.

dataPlot = data **%>%** **filter**(Species **!=** "versicolor")  
  
# visualizing samples with box plots  
ggpubr**::ggboxplot**(data = dataPlot, x = "Species", y = "Sepal.Width",   
 fill = "gray", add = "mean", ylab = "Sepal Width (cm)")



**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین واریانس‏های دو جامعه، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
x = data **%>%** **filter**(Species **==** "setosa") **%>%** **select**(Sepal.Width) **%>%** **unlist**()  
x.n = **length**(x = x)  
x.m = **mean**(x = x)  
x.v = **var**(x = x)  
  
y = data **%>%** **filter**(Species **==** "virginica") **%>%** **select**(Sepal.Width) **%>%** **unlist**()  
y.n = **length**(x = y)  
y.m = **mean**(x = y)  
y.v = **var**(x = y)  
  
# test statistic  
FF = x.v **/** y.v  
  
# degrees of freedom  
df1 = x.n **-** 1  
df2 = y.n **-** 1  
  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
**if** (alternative **!=** "two.sided") {  
 lower.critical = **qf**(p = (1 **-** alpha), df1 = df1, df2 = df2, lower.tail = FALSE)  
 upper.critical = **qf**(p = (1 **-** alpha), df1 = df1, df2 = df2, lower.tail = TRUE)  
} **else** {  
 lower.critical = **qf**(p = (1 **-** (alpha **/** 2)), df1 = df1, df2 = df2, lower.tail = FALSE)  
 upper.critical = **qf**(p = (1 **-** (alpha **/** 2)), df1 = df1, df2 = df2, lower.tail = TRUE)  
}  
p\_value = **pf**(q = FF, df1 = df1, df2 = df2, lower.tail = TRUE)  
p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value)  
**cat**("Test Statistic: ", FF, "\n")  
**cat**("Degrees of Freedom:", c(df1, df2), "\n")  
**cat**("Critical Values: ", **c**(lower.critical, upper.critical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: 1.3816  
## Degrees of Freedom: 49 49  
## Critical Values: 0.5675 1.7622  
## P-value: 0.2614

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "f", statistic = FF, df1 = df1, df2 = df2,  
 alpha = 0.05, alternative = "two.sided", from = 0, to = 3)



همانطور که مشخص است، در آزمون دو دنباله ، مقادیر بحرانی در برابر با (76/1 – 57/0) است (جدول شماره 3). مقدار محاسبه شده برای بین مقادیر بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که تفاوت معنی‏داری بین واریانس‏های عرض کاسبرگ در این دو گونه از گل زنبق وجود ندارد.

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین واریانس‏های دو جامعه ، از تابع **var.test** موجود در بسته **stats** استفاده شده است.

# method 2  
# setting initial parameter values  
x = data **%>%** **filter**(Species **==** "setosa") **%>%** **select**(Sepal.Width) **%>%** **unlist**()  
y = data **%>%** **filter**(Species **==** "virginica") **%>%** **select**(Sepal.Width) **%>%** **unlist**()  
  
# use "var.test" function from "stats" package  
**var.test**(x = x, y = y, alternative = "two.sided", conf.level = 0.95)

## F test to compare two variances   
## data: x and y  
## F = 1.3816, num df = 49, denom df = 49, p-value = 0.2614  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.7840 2.4346  
## sample estimates:  
## ratio of variances:  
## 1.381578

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 17: آزمون [[89]](#footnote-90) برای واریانس‏های دو جامعه (با مشاهدات همبسته[[90]](#footnote-91))

**هدف**

بررسی معنی‏دار بودن تفاوت بین واریانس‏های دو جامعه، هنگامیکه بین جفت مشاهدات همبستگی وجود داشته باشد.

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. فرض بر این است که مشاهدات به صورت زوجی جمع‏آوری شده‏اند و بین آن‏ها همبستگی وجود دارد. دو جامعه نیز باید از توزیع نرمال پیروی کنند.

**روش**

نمونه‏ای تصادفی با حجم و به صورت زوجی با مقادیر ، ، ...، انتخاب می‏گردد. نسبت واریانس مشابه آزمون 16 محاسبه می‏گردد.

همچنین ضریب همبستگی نیز با استفاده از معادله زیر محاسبه می‏گردد:

در نتیجه آماره آزمون برابر است با:

می‏توان آماره آزمون را با درجه‏های آزادی[[91]](#footnote-92) و مقادیر بحرانی به دست آمده از جدول 6 مقایسه کرد. در اینجا فرضیه صفر برابر است با ، هنگامیکه ضریب همبستگی جامعه مخالف صفر باشد. مقدار باید بزرگتر از 1 باشد.

**مثال**

ترویج و آموزش کشاورزی نوعی آموزش غیر رسمی به روستاییان و علاقه‌مندان به کشاورزی برای آشنائی و استفاده از فناوری‏های نو جهت افزایش تولید و بهره‏وری و درآمد و ارتقای سطح زندگی تولیدکنندگان کشاورزی و رسیدن به اهداف توسعه کشاورزی و روستائی است. جهاد کشاورزی استان خراسان رضوی در نظر دارد تاثیر یک برنامه ترویجی را بر میزان درآمد کشاورزان یک منطقه مورد ارزیابی قرار دهد. از اینرو 45 کشاورز منطقه انتخاب و میزان درآمد آن‏ها پرسیده می‏شود. بعد از دو سال از اجرای برنامه ترویجی نیز از همان 45 کشاورز درآمد آن‏ها پرسیده می‏شود. در سطح معنی‏داری ، آیا فرضیه صفر (تفاوتی بین واریانس درآمد کشاورزان قبل و بعد از اجرای برنامه ترویجی وجود ندارد) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
# load required functions  
**source**(file = "muFunc.R")  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "EnvStats")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_17.csv", header = TRUE)  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Sample | Before | After |
| ## | 1 | 1 | 79 | 101 |
| ## | 2 | 2 | 62 | 88 |
| ## | … | … | … | … |
| ## | 44 | 44 | 8 | 4 |
| ## | 45 | 45 | 38 | 24 |

**گام 2: بررسی محدودیت‏ها**

1- مشاهدات به صورت زوجی جمع‏آوری شده‏اند.

2- فرض می‏شود هر دو جامعه به صورت نرمال توزیع شده باشند.

3- بین زوج مشاهدات همبستگی وجود دارد.

به منظور تجسم رابطه بین زوج مشاهدات، نمودار نقطه‏ای آن‏ها با استفاده از تابع **chart.Correlation** از بسته **PerformanceAnalytics** و تابع **ggscatter** از بسته **ggpubr** رسم شده است.

# 1. correlation exists between the paired observations  
PerformanceAnalytics**::chart.Correlation**(R = data[, **c**("Before", "After")])  
  
data **%>%** ggpubr**::ggscatter**(x = "Before", y = "After",  
 add = "reg.line", conf.int = TRUE,  
 xlab = "Income - Before (Million Rial)",  
 ylab = "Income - After (Million Rial)") **%>%**  
 ggExtra**::ggMarginal**(type = "boxplot", fill = "gray90")

|  |  |
| --- | --- |
|  |  |

**گام 3: محاسبه آماره آزمون**

برای بررسی معنی‏دار بودن تفاوت بین واریانس‏های دو جامعه، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# setting initial parameter values  
x = data**$**Before  
x.m = **mean**(x = x)  
x.v = **var**(x = x)  
y = data**$**After  
y.m = **mean**(x = y)  
y.v = **var**(x = y)  
n = **nrow**(data)  
  
# correlation coefficient (r)  
r.numerator = **sum**((x **-** x.m) **\*** (y **-** y.m))  
r.denominator = **sqrt**(**sum**((x **-** x.m) **^** 2) **\*** **sum**((y **-** y.m) **^** 2))  
r = r.numerator **/** r.denominator  
  
# variance ratio (F)  
**if** (x.v **>=** y.v) {  
 FF = x.v **/** y.v  
} **else** {  
 FF = y.v **/** x.v  
}  
# test statistic  
gamma.F = (FF **-** 1) **/** **sqrt**(((F **+** 1) **^** 2) **-** (4 **\*** F **\*** (r **^** 2)))  
# degrees of freedom  
df = n **-** 2  
# compute the critical values  
alpha = 0.05  
alternative = "two.sided"  
  
**if** (alternative **!=** "two.sided") {  
 t.critical = **qt**(p = (1 **-** alpha), df = df, lower.tail = FALSE)  
 r.critical = **sqrt**((t.critical**^**2) **/** ((t.critical**^**2) **+** df))  
} **else** {  
 t.critical = **qt**(p = (1 **-** (alpha **/** 2)), df = df, lower.tail = FALSE)  
 r.critical = **sqrt**((t.critical**^**2) **/** ((t.critical**^**2) **+** df))  
}  
**cat**("Test Statistic:", gamma.F, "\n")  
**cat**("Degrees of Freedom:", df, "\n")  
**cat**("Critical Values:", **c**(**-**r.critical, r.critical), "\n")

## Test Statistic: 0.8675  
## Degrees of Freedom: 43  
## Critical Values: -0.294 0.294

همانطور که مشخص است، در آزمون دو دنباله ، مقادیر بحرانی در برابر با (294/0 – 294/-0) است (جدول شماره 6). مقدار محاسبه شده برای بزرگتر از مقادیر بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که تفاوت معنی‏داری بین واریانس درآمد کشاورزان قبل و بعد از اجرای برنامه ترویجی وجود دارد.

# آزمون 18: آزمون هتلینگ[[92]](#footnote-93) برای دو سری از میانگین‏های جامعه

**هدف**

به منظور مقایسه نتایج دو آزمایش، که هر یک از آن‏ها دارای نتایج مختلفی هستند. به عبارت دیگر، آیا الگوی میانگین به دست آمده از آزمایش اول با الگوی میانگین به دست آمده از آزمایش دوم توافق دارد؟

* فرضیه صفر ():
* فرضیه صفر ():

**پیش فرض‏ها**

1. فرض می‏شود همه متغیرها مستقل از یکدیگرند و از توزیع نرمال چند متغیره پیروی می‏کنند (متغیرها معمولا همبسته[[93]](#footnote-94) هستند).

**روش**

نتایج دو آزمایش با و مشخص شده‏اند. به منظور ساده سازی، تعداد متغیرها به سه مورد ، و محدود می‏شود. تعداد مشاهدات هر دو آزمایش با و نشان داده می‏شود. برای پیدا کردن پارامترهای ، و لازم است دستگاه سه معادله سه مجهول زیر حل گردد:

که در آن:

برای سایر جملات نیز تعاریف مشابهی وجود دارد.

هتلینگ به صورت زیر تعریف می‏شود:

و در نهایت آماره آزمون برابر است با:

که می‏توان آماره آزمون را با توزیع [[94]](#footnote-95) و با درجه‏های آزادی[[95]](#footnote-96) مقایسه کرد. در اینجا ، تعداد متغیرها می‏باشد.

**مثال**

در این مثال هم از مجموعه داده گل زنبق[[96]](#footnote-97) استفاده می‏شود. این مجموعه داده 4 ویژگی از 3 نوع گونه به نام‏های setosa،versicolor و virginica را شامل می‏شود. 4 ویژگی استفاده شده عبارت‏اند از طول و عرض گلبرگ[[97]](#footnote-98) و طول و عرض کاسبرگ[[98]](#footnote-99). از هر نوع گل50 نمونه وجود دارد بنابراین داده‏های این مجموعه را می‏توان شامل 150 مشاهده دانست که هر کدام از مشاهدات شامل 4 ویژگی (متغیر) می‏باشند. محققی در نظر دارد بداند که آیا دو گونه setosa و versicolor از نظر میانگین چهار متغیر یکسان است یا خیر؟ (با فرض مساوی بودن ماتریس کوواریانس آن‏ها).

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "MVTests")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_18.csv", header = TRUE)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

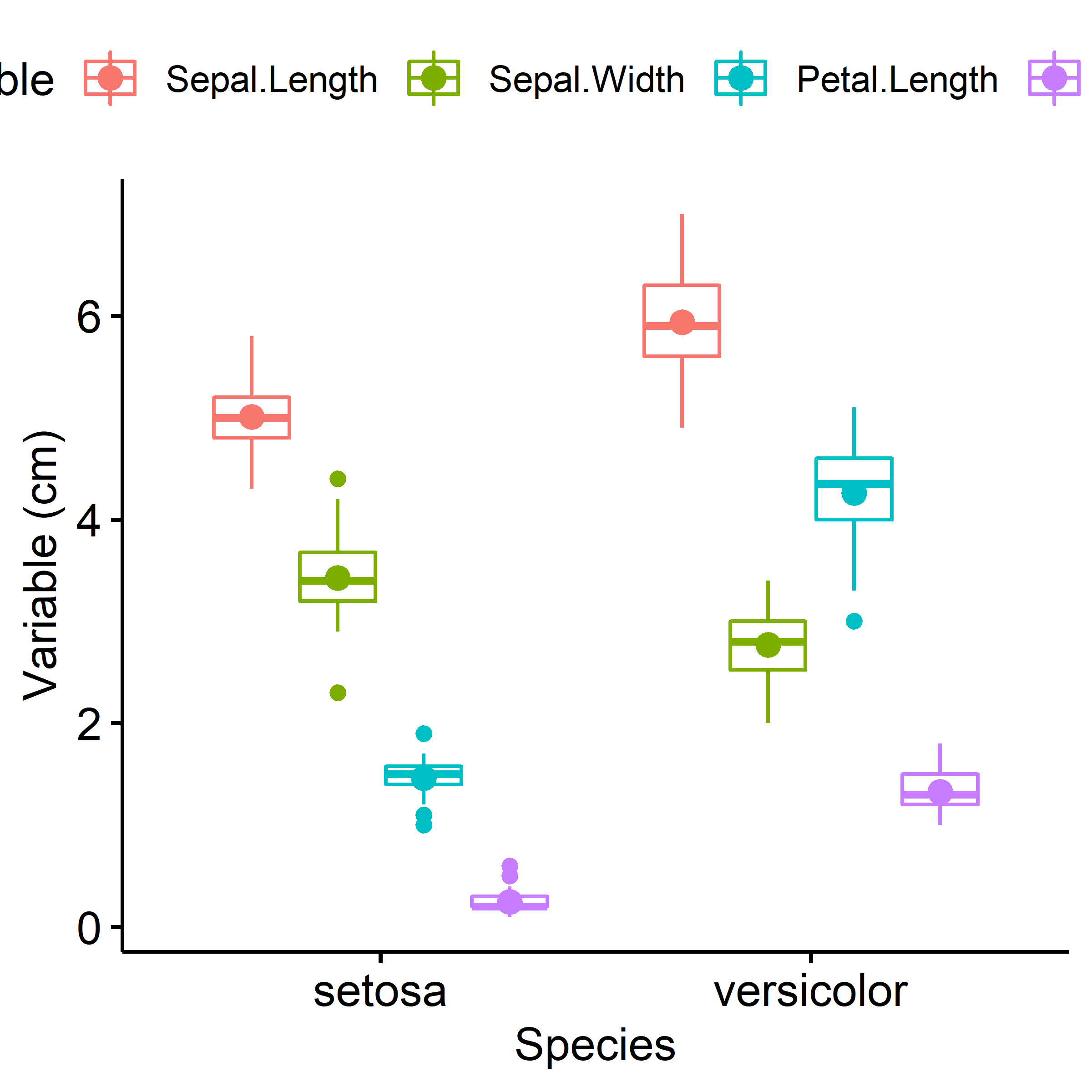
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## |  | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
| ## | 1 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| ## | 2 | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| ## | … | … | … | … | … | … |
| ## | 149 | 6.2 | 3.4 | 5.4 | 2.3 | virginica |
| ## | 150 | 5.9 | 3.0 | 5.1 | 1.8 | virginica |

**گام 2: بررسی محدودیت‏ها**

1- فرض می‏شود هر دو جامعه به صورت نرمال توزیع شده باشند و از هم مستقل هستند.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها رسم شده است. برای این منظور، از تابع **ggboxplot** موجود در بسته **ggpubr** استفاده شده است.

dataPlot = data **%>%** **filter**(Species **!=** "virginica") **%>%**   
 reshape2**::melt**(id.vars = "Species")  
  
# visualizing samples with box plots  
ggpubr**::ggboxplot**(data = dataPlot, x = "Species", y = "value", color = "variable",   
 add = "mean", ylab = "Variable (cm)")



**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی توافق بین الگوی میانگین به دست آمده از آزمایش اول با الگوی میانگین به دست آمده از آزمایش دوم، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
XY = **function**(X, Y) {  
 X = **unlist**(x = X)  
 Y = **unlist**(x = Y)  
 result = **sum**((X **-** **mean**(X)) **\*** (Y **-** **mean**(Y)))  
 **return**(result)  
}  
data = data **%>%** **filter**(Species **!=** "virginica")  
  
A = data **%>%** **filter**(Species **==** "setosa") **%>%** **select**(**-**Species)  
A.m = **colMeans**(x = A)  
A.n = **nrow**(x = A)  
B = data **%>%** **filter**(Species **==** "versicolor") **%>%** **select**(**-**Species)  
B.m = **colMeans**(x = B)  
B.n = **nrow**(x = B)  
n = A.n **+** B.n  
p = **ncol**(x = A)  
  
# solve the equations to find the statistics a, b, c and d (AA \* x = BB)  
data\_AA = **c**()  
**for** (i **in** 1**:**p) {  
 **for** (j **in** 1**:**p) {  
 data\_AA = **c**(data\_AA, (**XY**(A[i], A[j]) **+** **XY**(B[i], B[j])))  
 }  
}  
AA = **array**(data = data\_AA, dim = **c**(p, p))  
BB = **c**()  
**for** (i **in** 1**:**p) {  
 BB = **c**(BB, ((n **-** 2) **\*** (A.m[i] **-** B.m[i])))  
}  
coef = **solve**(a = AA, b = BB)  
  
# Hotelling’s T2  
T2.a = (A.n **\*** B.n) **/** (A.n **+** B.n)  
T2.b = **c**()  
**for** (i **in** 1**:**p) {  
 T2.b = **c**(T2.b, (coef[i] **\*** (A.m[i] **-** B.m[i])))  
}  
T2.b = **sum**(T2.b)  
T2 = T2.a **\*** T2.b  
# test statistic   
FF = ((A.n **+** B.n **-** p **-** 1) **/** (p **\*** (A.n **+** B.n **-** 2))) **\*** T2  
  
# degrees of freedom  
df1 = p  
df2 = A.n **+** B.n **-** p **-** 1  
  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
**if** (alternative **!=** "two.sided") {  
 lower.critical = **qf**(p = (1 **-** alpha), df1 = df1, df2 = df2, lower.tail = FALSE)  
 upper.critical = **qf**(p = (1 **-** alpha), df1 = df1, df2 = df2, lower.tail = TRUE)  
} **else** {  
 lower.critical = **qf**(p = (1 **-** (alpha **/** 2)), df1 = df1, df2 = df2, lower.tail = FALSE)  
 upper.critical = **qf**(p = (1 **-** (alpha **/** 2)), df1 = df1, df2 = df2, lower.tail = TRUE)  
}  
p\_value = **pf**(q = FF, df1 = df1, df2 = df2, lower.tail = TRUE)  
p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value)  
  
**cat**("Hotelling T Sqaure Statistic:", T2, "\n")  
**cat**("Test Statistic:", FF, "\n")  
**cat**("Degrees of Freedom:", **c**(df1, df2), "\n")  
**cat**("Critical Values:", **c**(lower.critical, upper.critical),"\n")  
**cat**("P-value:", p\_value, "\n")

## Hotelling T Sqaure Statistic: 2580.839  
## Test Statistic: 625.4583  
## Degrees of Freedom: 4 95  
## Critical Values: 0.1202 2.9237  
## P-value: 0.0

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "f", statistic = FF, df1 = df1, df2 = df2,  
 alpha = 0.05, alternative = "two.sided", from = 0, to = 4)



همانطور که مشخص است، در آزمون دو دنباله ، مقادیر بحرانی در برابر با (92/2 – 12/0) است (جدول شماره 3). مقدار محاسبه شده برای بیشتر از مقادیر بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که توافقی بین الگوی میانگین به دست آمده از آزمایش اول با الگوی میانگین به دست آمده از آزمایش دوم وجود ندارد.

در روش دوم، برای بررسی توافق بین الگوی میانگین به دست آمده از آزمایش اول با الگوی میانگین به دست آمده از آزمایش دوم، از تابع **MVTests** موجود در بسته **TwoSamplesHT2** استفاده شده است.

# method 2  
# setting initial parameter values  
data = data **%>%** **filter**(Species **!=** "virginica") **%>%** **select**(**-**Species)  
group = **c**(**rep**(x = 1, 50), **rep**(x = 2, 50))  
  
# use "TwoSamplesHT2" function from "MVTests" package  
result = MVTests**::TwoSamplesHT2**(data = data, group = group,   
 alpha = 0.05, Homogenity = TRUE)  
  
**summary**(object = result)

## Two Independent Samples Hotelling T Square Test   
## Hotelling T Sqaure Statistic = 2580.839   
## F value = 625.458, df1 = 4, df2 = 95, p-value: <2e-16

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 19: آزمون تشخیصی[[99]](#footnote-100) برای منشا یک نمونه تایی

**هدف**

به منظور تعیین منشا یک سری از مقادیر متغیر تصادفی، هنگامیکه این سری قطعا از یکی از دو جامعه انتخاب شده باشند.

**پیش فرض‏ها**

این تست یک قانون تصمیم‏گیری فراهم می‏کند که ارتباط نزدیکی با آزمون هتلینگ (آزمون 18) دارد. از اینرو محدودیت‏های آزمون 18 در اینجا نیز برقرار است.

**روش**

با توجه به نمادهای استفاده شده در آزمون هتلینگ، نمونه‏ای از هر دو جامعه گرفته شده و معادله‏های زیر برای هر دو جامعه محاسبه می‏گردند:

برای تعیین منشا سری نمونه انتخابی (که با نماد نشان داده خواهد شد)، معادله زیر محاسبه می‏گردد:

حال اگر بود، می‏توان گفت که سری نمونه انتخابی متعلق به جامعه می‏باشد، و اگر می‏توان گفت که سری نمونه انتخابی متعلق به جامعه می‏باشد.

**مثال**

مثال آزمون 18 را در نظر بگیرید، حال اگر 15 نمونه از یک گونه گل زنبق انتخاب شده باشند، این نمونه‏های انتخابی به کدام یک از گونه‏های گل زنبق (setosa و versicolor) تعلق دارند؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "MVTests")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_18.csv", header = TRUE)  
data\_s = **read.csv**(file = "data/Data\_Test\_19.csv", header = TRUE)  
# show data  
psych**::headTail**(x = data\_s, top = 2, bottom = 2)

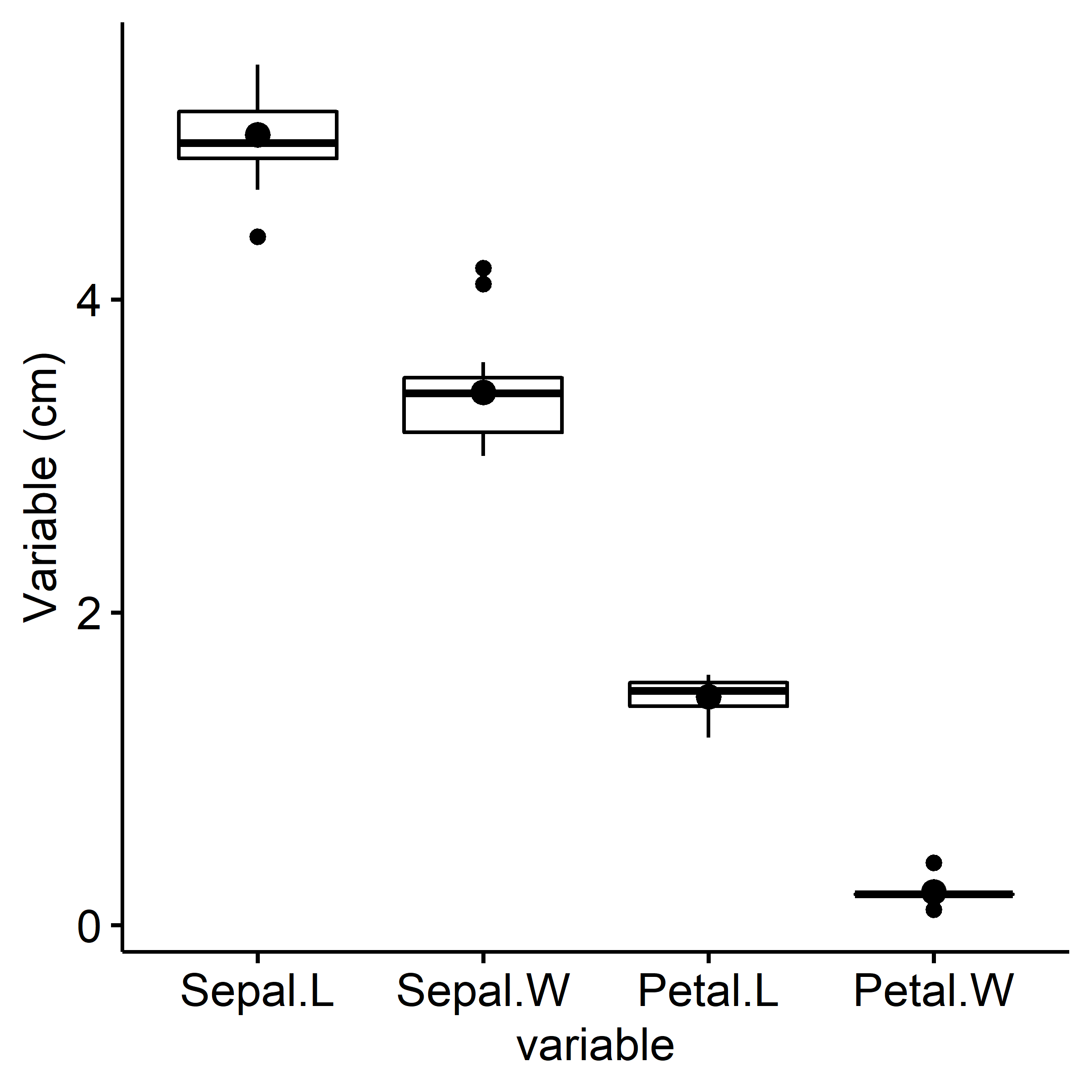
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## |  | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
| ## | 1 | 5 | 3 | 1.6 | 0.2 |
| ## | 2 | 5 | 3.4 | 1.6 | 0.4 |
| ## | … | … | … | … | … |
| ## | 14 | 4.4 | 3 | 1.3 | 0.2 |
| ## | 15 | 5.1 | 3.4 | 1.5 | 0.2 |

**گام 2: بررسی محدودیت‏ها**

1- فرض می‏شود هر دو جامعه به صورت نرمال توزیع شده باشند و از هم مستقل هستند.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها رسم شده است. برای این منظور، از تابع **ggboxplot** موجود در بسته **ggpubr** استفاده شده است.

# visualizing samples with box plots  
dataPlot = data\_s **%>%** reshape2**::melt**()  
ggpubr**::ggboxplot**(data = dataPlot, x = "variable", y = "value", add = "mean",  
 ylab = "Variable (cm)")



**گام 3: محاسبه آماره آزمون**

برای تعیین منشا یک سری از مقادیر متغیر تصادفی، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون آن به دست آید.

# method 1  
# setting initial parameter values  
XY = **function**(X, Y) {  
 X = **unlist**(x = X)  
 Y = **unlist**(x = Y)  
 result = **sum**((X **-** **mean**(X)) **\*** (Y **-** **mean**(Y)))  
 **return**(result)  
}  
data = data **%>%** **filter**(Species **!=** "virginica")  
  
A = data **%>%** **filter**(Species **==** "setosa") **%>%** **select**(**-**Species)  
A.m = **colMeans**(x = A)  
A.n = **nrow**(x = A)  
  
B = data **%>%** **filter**(Species **==** "versicolor") **%>%** **select**(**-**Species)  
B.m = **colMeans**(x = B)  
B.n = **nrow**(x = B)  
  
n = A.n **+** B.n  
p = **ncol**(x = A)  
  
# solve the equations to find the statistics a, b, c and d (AA \* x = BB)  
data\_AA = **c**()  
**for** (i **in** 1**:**p) {  
 **for** (j **in** 1**:**p) {  
 data\_AA = **c**(data\_AA, (**XY**(A[i], A[j]) **+** **XY**(B[i], B[j])))  
 }  
}  
AA = **array**(data = data\_AA, dim = **c**(p, p))  
  
BB = **c**()  
**for** (i **in** 1**:**p) {  
 BB = **c**(BB, ((n **-** 2) **\*** (A.m[i] **-** B.m[i])))  
}

coef = **solve**(a = AA, b = BB)  
  
# discriminant test  
DA = **sum**(coef **\*** A.m)  
DB = **sum**(coef **\*** B.m)  
  
S = data\_s  
S.m = **colMeans**(x = S)  
DS = **sum**(coef **\*** S.m)  
  
**if** (**abs**(DA **-** DS) **<** **abs**(DB **-** DS)) {  
 **cat**("S Series Belongs to Population A")  
} **else** {  
 **cat**("S Series Belongs to Population B")  
}

## S Series Belongs to Population A

# آزمون 20: آزمون دیکسون[[100]](#footnote-101) برای نقاط دورافتاده[[101]](#footnote-102)

**هدف**

بررسی معنی‏دار بودن تفاوت بین یک مقدار حدی[[102]](#footnote-103) (بیش از حد) مشکوک و سایر مقادیر، در یک نمونه.

* فرضیه صفر (): مقدار حدی مشکوک نمونه، متعلق به جامعه مورد بررسی می‏باشد.
* فرضیه مقابل (): مقدار حدی مشکوک نمونه، یک مشاهده دورافتاده می‏باشد.

**پیش فرض‏ها**

1- حجم نمونه باید بزرگتر از 3 باشد.

2- جامعه‏ای که نمونه از آن انتخاب شده است باید دارای توزیع نرمال باشد.

**روش**

نمونه‏ای با حجم انتخاب می‏شود. در صورتی که این نمونه دارای مقدار حدی باشد، از نزدیکترین مشاهده مجاور آن داده‏ها به صورت صعودی یا نزولی مرتب می‏گردد. در این حالت مقدار حدی، بزرگترین یا کوچکترین مقدار خواهد بود. سری مرتب شده با نمادهای ، ، ...، نشان داده می‏شود. در این صورت آماره آزمون برابر است با:

مقادیر بحرانی برای از جدول 8 به دست می‏آید. فرضیه صفر، مقدار حدی متعلق به نمونه می‏باشد، هنگامی رد می‏شود که مقدار محاسبه شده بیشتر از مقدار بحرانی باشد.

**مثال**

محققی از مجموعه داده‏های بارش سالانه ایستگاه سینوپتیک اهواز نمونه‏ای کاملا تصادفی با حجم 24 سال انتخاب کرده است. این محقق در نظر دارد بررسی کند که آیا کمترین و بیشترین مقدار بارش سالانه در این نمونه، مشاهده دورافتاده می‏باشند یا خیر؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "outliers")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_21.csv", header = TRUE)  
  
**set.seed**(1)  
sample = data **%>%** **sample\_n**(size = 24)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Year | Station | Prec. |
| ## | 1 | 1959 | Ahvaz | 120.1 |
| ## | 2 | 1960 | Ahvaz | 117.9 |
| ## | … | … | … | … |
| ## | 59 | 2017 | Ahvaz | 77.5 |
| ## | 60 | 2018 | Ahvaz | 307.8 |

**گام 2: بررسی محدودیت‏ها**

1- حجم نمونه باید بزرگتر از 3 ‏باشد.

# assumption checking  
# 1. the sample size should be greater than 3.  
**cat**("Sample Size Greater Than 3:", (**nrow**(sample) **>** 3), "\n")

## Sample Size Greater Than 3: TRUE

2- جامعه‏ای که نمونه از آن انتخاب شده است باید دارای توزیع نرمال باشد: در این مرحله برای بررسی اینکه آیا داده‏های بارندگی سالانه ایستگاه اهواز از توزیع نرمال پیروی می‏کنند یا خیر، از «آزمون شاپیرو ویلک» استفاده شده است. برای انجام این آزمون از تابع **shapiro.test** بسته **stats** استفاده شده است.

# 2. the population which is being sampled is assumed normal.  
# use Shapiro-Wilk normality test  
**shapiro.test**(x = data**$**Prec)

## Shapiro-Wilk normality test  
## data: data$Prec  
## W = 0.9696, p-value = 0.1389

با توجه به اینکه p-مقدار آزمون بیشتر از 05/0 می‏باشد، فرضیه صفر رد نشده و می‏توان نتیجه گرفت که داده‏های بارندگی سالانه ایستگاه اهواز از توزیع نرمال پیروی می‏کند. از طرفی می‏توان با رسم نمودار چندک چندک نیز نرمال بودن داده‏ها را مورد ارزیابی قرار داد. برای این منظور، از تابع **ggqqplot** موجود در بسته **ggpubr** استفاده شده است.

# normality plot  
ggpubr**::ggqqplot**(data = data**$**Prec, ylab = "Precipitation (mm)")  
  
# visualize data using density plot  
ggpubr**::ggdensity**(data = data, x = "Prec", fill = "gray",  
 xlab = "Precipitation (mm)", add = **c**("mean"))

|  |  |
| --- | --- |
|  |  |

همانطور که دیده می‏شود، فرض نرمال بودن توزیع جامعه مورد تایید قرار می‏گیرد.

**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین یک مقدار حدی مشکوک و سایر مقادیر در یک نمونه، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
x = **sort**(x = sample**$**Prec)  
n = **length**(x = x)  
# select type  
**if** (n **<=** 7) {  
 type = 10  
} **else** **if** (n **>** 7 **&** n **<=** 10) {  
 type = 11  
} **else** **if** (n **>** 10 **&** n **<=** 13) {  
 type = 21  
} **else** {  
 type = 22  
}  
# test statistic  
# check not the value with lowest difference from the mean  
**if** (type **==** 10) {  
 Qmin = (x[2] **-** x[1]) **/** (x[n] **-** x[1])  
} **else** **if** (type **==** 11) {  
 Qmin = (x[2] **-** x[1]) **/** (x[n **-** 1] **-** x[1])  
} **else** **if** (type **==** 21) {  
 Qmin = (x[3] **-** x[1]) **/** (x[n **-** 1] **-** x[1])  
} **else** {  
 Qmin = (x[3] **-** x[1]) **/** (x[n **-** 2] **-** x[1])  
}  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
**if** (alternative **!=** "two.sided") {  
 dixon.critical = outliers**::qdixon**(p = alpha, n = n, type = type)  
} **else** {  
 dixon.critical = outliers**::qdixon**(p = (alpha **/** 2), n = n, type = type)  
}  
p\_value = outliers**::pdixon**(q = Qmin, n = n, type = type)  
p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value)  
  
**cat**("Lowest Value:", x[1], "\n")  
**cat**("Test Statistic: ", Qmin, "\n")  
**cat**("Critical Values: ", dixon.critical, "\n")  
**cat**("P-value: ", p\_value, "\n")

## Lowest Value: 72.9  
## Test Statistic: 0.1864  
## Critical Values: 0.452  
## P-value: 0.8998

همانطور که مشخص است، مقدار بحرانی در برابر با 452/0 است (جدول شماره 8). مقدار محاسبه شده برای آماره آزمون، 186/0 کوچکتر از مقادیر بحرانی بوده (p-مقدار بزرگتر از 05/0 می‏باشد) و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که کمترین مقدار نمونه، متعلق به جامعه مورد بررسی بوده و مشاهده دورافتاده نمی‏باشد.

# test statistic  
# check not the value with largest difference from the mean  
**if** (type **==** 10) {  
 Qmax = (x[n] **-** x[n **-** 1]) **/** (x[n] **-** x[1])  
} **else** **if** (type **==** 11) {  
 Qmax = (x[n] **-** x[n **-** 1]) **/** (x[n] **-** x[2])  
} **else** **if** (type **==** 21) {  
 Qmax = (x[n] **-** x[n **-** 2]) **/** (x[n] **-** x[2])  
} **else** {  
 Qmax = (x[n] **-** x[n **-** 2]) **/** (x[n] **-** x[3])  
}  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
**if** (alternative **!=** "two.sided") {  
 dixon.critical = outliers**::qdixon**(p = alpha, n = n, type = type)  
} **else** {  
 dixon.critical = outliers**::qdixon**(p = (alpha **/** 2), n = n, type = type)  
}  
p\_value = outliers**::pdixon**(q = Qmax, n = n, type = type)  
p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value)  
  
**cat**("Highest Value:", x[n], "\n")  
**cat**("Test Statistic: ", Qmax, "\n")  
**cat**("Critical Values: ", dixon.critical, "\n")  
**cat**("P-value: ", p\_value, "\n")

## Highest Value: 498.8  
## Test Statistic: 0.4560  
## Critical Values: 0.452  
## P-value: 0.0464

همانطور که مشخص است، مقدار بحرانی در برابر با 452/0 است (جدول شماره 8). مقدار محاسبه شده برای آماره آزمون، 456/0 بیشتر از مقادیر بحرانی بوده (p-مقدار کمتر از 05/0 می‏باشد) و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که بیشترین مقدار نمونه، متعلق به جامعه مورد بررسی نبوده و مشاهده دورافتاده می‏باشد.

در ادامه برای نمایش مشاهدات دورافتاده در نمونه‏، نمودار جعبه‏ای آن‏ها رسم شده است. در این نمودار از جعبه‏ای برای نمایش فاصله بین چارک اول و سوم استفاده می‏شود و خطی در داخل جعبه، میانه (چارک دوم) را مشخص می‏کند. خارج از جعبه حداقل و حداکثر مقدار داده را نیز مشخص می‌نمایند. مشاهدات دورافتاده نیز به صورت نقاطی خارج از محدوده نشان داده می‏شوند. برای این منظور، از تابع **ggboxplot** موجود در بسته **ggpubr** استفاده شده است.

# visualizing samples with box plots  
ggpubr**::ggboxplot**(data = sample, x = "Station", y = "Prec",   
 fill = "gray", add = "mean", xlab = "", ylab = "Precipitation (mm)")



در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین یک مقدار حدی مشکوک و سایر مقادیر در یک نمونه، از تابع **dixon.test** موجود در بسته **outliers** استفاده شده است.

# method 2  
# use "dixon.test" function from "outliers" package  
# check not the value with lowest difference from the mean  
outliers**::dixon.test**(x = sample**$**Prec, opposite = TRUE, two.sided = TRUE)

## Dixon test for outliers  
## data: sample$Prec  
## Q = 0.1864, p-value = 0.8998  
## alternative hypothesis: lowest value 72.9 is an outlier

# check not the value with largest difference from the mean  
outliers**::dixon.test**(x = sample**$**Prec, opposite = FALSE, two.sided = TRUE)

## Dixon test for outliers  
## data: sample$Prec  
## Q = 0.45603, p-value = 0.04642  
## alternative hypothesis: highest value 498.8 is an outlier

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 21: آزمون برای نسبت‏های همبسته[[103]](#footnote-104)

**هدف**

بررسی تفاوت معنی‏داری آماری بین دو نسبت همبسته در نظرسنجی‏ها. همچنین می‏توان برای کاربردهای عمومی هم مورد استفاده قرار گیرد.

* فرضیه صفر (): دو نسبت همبسته تفاوت معنی‏داری با هم ندارند.
* فرضیه مقابل (): دو نسبت همبسته تفاوت معنی‏داری با هم دارند.

**پیش فرض‏ها**

1 – از افراد مشابه در دو زمان، بر اساس بلی – خیر مورد سوال واقع شوند.

2- حجم نمونه‏ باید به اندازه کافی بزرگ باشد.

**روش**

N نفر به سوالی، قبل و بعد از یک برنامه خاص، به صورت بلی – خیر پاسخ می‏دهند. با توجه به پاسخ‏ها می‏توان جدول دو طرفه زیر را تشکیل داد:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | نظرسنجی 1 | | |  | |
|  | خیر | بلی |  | |  |
|  | b | a | بلی | | نظرسنجی 2 |
|  | d | c | خیر | |
| N |  |  |  | |  |

به منظور تصمیم‏گیری برای این که آیا این برنامه خاص تغییر معنی‏داری در نسبت پاسخ‏های بلی داشته است یا خیر، آماره آزمون زیر محاسبه می‏گردد:

که در آن

می‏باشد.

**مثال**

از 171 نفر از کشاورزان حوضه زاینده‏رود در دو استان اصفهان و چهارمحال و بختیاری که در چندین سال گذشته از آب رودخانه استفاده کرده‌اند یا در حال حاضر استفاده می‌کنند، در مورد محدود شدن استفاده از منابع آب، قبل و بعد از یک برنامه آموزشی، به صورت درست یا نادرست، مورد سوال قرار گرفته‏اند. آیا این برنامه آموزشی در تغییر پاسخ کشاورزان، قبل و بعد از برنامه آموزشی تاثیرگذار بوده است یا خیر؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
# load required functions  
**source**(file = "muFunc.R")  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_23.csv", header = TRUE)  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | User\_ID | Before | After |
| ## | 1 | 1 | NO | YES |
| ## | 2 | 2 | NO | NO |
| ## | … | … | … | … |
| ## | 170 | 170 | YES | NO |
| ## | 171 | 171 | NO | NO |

**گام 2: بررسی محدودیت‏ها**

حجم نمونه‏ اندازه کافی بزرگ می‏باشد.

**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی تفاوت معنی‏داری آماری بین دو نسبت همبسته، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# two-way table  
result = **table**(data**$**Before, data**$**After)  
# number of people  
N = **sum**(result)  
# test statistic  
a = result[2,2]  
b = result[1,2]  
c = result[2,1]  
d = result[1,1]  
sigma = **sqrt**(((b **+** c) **-** (((b **-** c) **^** 2) **/** N)) **/** (N **\*** (N **-** 1)))  
Z = (b **-** c) **/** (N **\*** sigma)  
  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
**if** (alternative **!=** "two.sided") {  
 lower.critical = **qnorm**(p = (1 **-** alpha), lower.tail = FALSE)  
 upper.critical = **qnorm**(p = (1 **-** alpha), lower.tail = TRUE)  
} **else** {  
 lower.critical = **qnorm**(p = (1 **-** (alpha **/** 2)), lower.tail = FALSE)  
 upper.critical = **qnorm**(p = (1 **-** (alpha **/** 2)), lower.tail = TRUE)  
}  
  
p\_value = **pnorm**(q = Z, lower.tail = TRUE)  
p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value)  
  
**cat**("Two-way Table:", "\n")  
result  
**cat**("Test Statistic:", Z, "\n")  
**cat**("Critical Values:", **c**(lower.critical, upper.critical),"\n")  
**cat**("P-value:", p\_value, "\n")

## Two-way Table:  
## NO YES  
## NO 21 30  
## YES 39 81  
## Test Statistic: -1.0840  
## Critical Values: -1.96 1.96  
## P-value: 0.2784

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع نرمال استاندارد با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "norm", statistic = Z, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، مقادیر بحرانی در برابر با (96/1 – 96/1-) است (جدول شماره 1). مقدار محاسبه شده برای بین مقادیر بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که دو نسبت همبسته تفاوت معنی‏داری با هم ندارند و این برنامه آموزشی در تغییر پاسخ کشاورزان، تاثیرگذار نبوده است.

در روش دوم، برای بررسی تفاوت معنی‏داری آماری بین دو نسبت همبسته، از تابع **mcnemar.test** موجود در بسته **stats** استفاده شده است.

# method 2  
# use "mcnemar.test" function from "stats" package  
**mcnemar.test**(x = result, correct = F)

## McNemar's Chi-squared test  
## data: result  
## McNemar's chi-squared = 1.1739, df = 1, p-value = 0.2786

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 22: آزمون (خی‏دو) برای یک مقدار فرض شده به عنوان واریانس جامعه

**هدف**

بررسی تفاوت معنی‏داری آماری بین واریانس جامعه و یک مقدار فرض شده .

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. جامعه‏ای که نمونه از آن انتخاب شده است باید دارای توزیع نرمال باشد.

**روش**

واریانس نمونه با استفاده از معادله زیر محاسبه می‏گردد:

در این صورت آماره آزمون برابر است با:

که می‏توان آماره آزمون را با توزیع [[104]](#footnote-105) با درجه‏های آزادی[[105]](#footnote-106) مقایسه کرد.

**مثال**

اندازه‏گیری رطوبت خاک به روش بازتاب زمانی امواج (TDR) یک روش غیرمستقیم، بیخطر و سریع است که بر اساس ثابت دی‏الکتريك خاک عمل می‏نماید. این روش برای اکثر خاک‏های دارای شرایط معمولی و متعارف نتایج قابل قبولی می‏دهد، اما نتایج این روش برای خاک‏های با شرایط نامتعارف ازجمله شوری و ماده آلی زیاد با خطا همراه می‏باشد. شرکت دانش بنیانی مدعی شده است که حسگر تدفینی برای روش TDR طراحی کرده است دارای دقت بسیار بالایی در برآورد رطوبت بافت‏های مختلف خاک در شرایط نامتعارف می‏باشد. بر اساس گفته این شرکت واریانس خطاهای این دستگاه 10 درصد می‏باشد. به صورت تصادفی این حسگر در 30 نقطه مزرعه‏ای قرار گرفته و رطوبت خاک به روش TDR و وزنی (به عنوان رطوبت واقعی) به طور همزمان در این نقاط اندازه‏گيری شد. این نمونه‏ها دارای میانگین خطای 43/0 درصد و واریانس خطای 8/18 درصد می‏باشند. با توجه به اطلاعات فوق و در سطح معنی‏داری ، آیا فرضیه صفر (رطوبت خاک‏های اندازه‏گیری شده با دستگاه TDR دارای واریانس خطای 10 درصد می‏باشند) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "EnvStats")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_24.csv", header = TRUE)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Sample | TDR | Gravimetric |
| ## | 1 | 1 | 11 | 18 |
| ## | 2 | 2 | 14 | 18 |
| ## | … | … | … | … |
| ## | 29 | 29 | 13 | 10 |
| ## | 30 | 30 | 11 | 9 |

**گام 2: بررسی محدودیت‏ها**

فرض می‏شود جامعه‏ای که نمونه از آن انتخاب شده است دارای توزیع نرمال می‏باشد.

به منظور تجسم رابطه بین این دو مقدار اندازه‏گیری، نمودار نقطه‏ای آن‏ها با استفاده از تابع **ggscatter** از بسته **ggpubr** رسم شده است.

# visualize data using scatter plots  
data **%>%** ggpubr**::ggscatter**(x = "Gravimetric", y = "TDR",  
 add = "reg.line", conf.int = TRUE,  
 xlab = "Gravimetric Soil Moisture (%)",  
 ylab = "TDR Soil Moisture (%)") **%>%**  
 ggExtra**::ggMarginal**(type = "boxplot", fill = "gray90")



**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی تفاوت معنی‏داری آماری بین واریانس جامعه و یک مقدار فرض شده، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
x = data**$**Gravimetric **-** data**$**TDR  
x.m = **mean**(x = x)  
x.n = **length**(x = x)  
p.v = 10  
  
# sample variance  
x.v = (**sum**((x **-** x.m) **^** 2)) **/** (x.n **-** 1)  
  
# test statistic  
chi2 = (x.v **/** p.v) **\*** (x.n **-** 1)  
  
# degrees of freedom  
df = x.n **-** 1  
  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
**if** (alternative **!=** "two.sided") {  
 lower.critical = **qchisq**(p = (1 **-** alpha), df = df, lower.tail = FALSE)  
 upper.critical = **qchisq**(p = (1 **-** alpha), df = df, lower.tail = TRUE)  
} **else** {  
 lower.critical = **qchisq**(p = (1 **-** (alpha **/** 2)), df = df, lower.tail = FALSE)  
 upper.critical = **qchisq**(p = (1 **-** (alpha **/** 2)), df = df, lower.tail = TRUE)  
}  
  
p\_value = **pchisq**(q = chi2, df = df, lower.tail = TRUE)  
p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value)  
  
**cat**("Test Statistic:", chi2, "\n")  
**cat**("Degrees of Freedom:", df, "\n")  
**cat**("Critical Values:", **c**(lower.critical, upper.critical), "\n")  
**cat**("P-value:", p\_value, "\n")

## Test Statistic: 54.53667  
## Degrees of Freedom: 29  
## Critical Values: 16.0471 45.7223  
## P-value: 0.0056

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "chisq", statistic = chi2, df = df, alpha = 0.05,  
 alternative = "two.sided", from = 10, to = 60)



همانطور که مشخص است، مقادیر بحرانی در برابر با (72/45 – 05/16) است (جدول شماره 5). مقدار محاسبه شده برای بیشتر از مقادیر بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که رطوبت خاک‏های اندازه‏گیری شده با دستگاه TDR دارای واریانس خطای بیشتری از 10 درصد می‏باشند و ادعای شرکت رد می‏گردد.

در روش دوم، برای بررسی تفاوت معنی‏داری آماری بین واریانس جامعه و یک مقدار فرض شده، از تابع **varTest** موجود در بسته **EnvStats** استفاده شده است.

# method 2  
# use "varTest" function from "EnvStats" package   
p.v = 10  
EnvStats**::varTest**(x = x, sigma.squared = p.v, alternative = "two.sided", conf.level = 0.95)

## Chi-Squared Test on Variance  
## data: x  
## Chi-Squared = 54.537, df = 29, p-value = 0.0056  
## alternative hypothesis: true variance is not equal to 10  
## 95 percent confidence interval:  
## 11.9278 33.9854  
## sample estimates:  
## variance: 18.8058

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 23: آزمون برای دو مقدار شمارشی (توزیع پواسون[[106]](#footnote-107))

**هدف**

بررسی تفاوت معنی‏داری آماری بین دو مقدار شمارش شده (بر اساس یک توزیع پواسون).

* فرضیه صفر ():
* فرضیه مقابل ():

**پیش فرض‏ها**

1. مقادیر شمارشی دارای توزیع پواسون بوده و دو نمونه در شرایط مشابه‏ای به دست آمده‏اند.

**روش**

فرض می‏شود که و نماد میانگین دو جامعه و و تعداد مقادیر شمارش شده باشند. به منظور آزمون فرضیه صفر ، آماره آزمون برابر است با:

که می‏توان آماره آزمون را با توزیع و با درجه‏های آزادی مقایسه کرد. هر گاه مقادیر شمارشی در بازه‏های زمانی متفاوت و به دست آمده باشند، لازم است نسبت‏های مقادیر شمارشی و با هم مقایسه شوند. از این رو آماره آزمون در حالت برابر است با:

که می‏توان آماره آزمون را با توزیع و با درجه‏های آزادی مقایسه کرد.

**مثال**

هدف از پژوهشی بررسی این است که آیا تعداد روزهای کاملا ابری (مقدار ساعات آفتابی آن روز برابر با صفر باشد) ایستگاه‏های هواشناسی رشت و رامسر در سال 1397 با هم برابر است یا خیر؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_25.csv", header = TRUE)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Day | Rasht | Ramsar |
| ## | 1 | 1 | 0 | 0 |
| ## | 2 | 2 | 7.6 | 4.3 |
| ## | … | … | … | … |
| ## | 364 | 364 | 2 | 0.4 |
| ## | 365 | 365 | 0 | 0.3 |

**گام 2: بررسی محدودیت‏ها**

مقادیر شمارشی دارای توزیع پواسون بوده و دو نمونه در شرایط مشابه‏ای به دست آمده‏اند.

به منظور تجسم رابطه بین این دو مقدار شمارشی، نمودار موزاییکی[[107]](#footnote-108) آن‏ها با استفاده از تابع **mosaicplot** از بسته **graphics** رسم شده است. عدد 1 نشان دهنده روز کاملا ابری و عدد صفر نشان دهنده روزی که کاملا ابری نبوده است.

# visualize data using scatter plots  
dataPlot = **data.frame**(Rasht = **ifelse**(test = data**$**Rasht **==** 0, yes = 1, no = 0),  
 Ramsar = **ifelse**(test = data**$**Ramsar **==** 0, yes = 1, no = 0))  
  
**mosaicplot**(x = **table**(dataPlot), main = "")



**گام 3: محاسبه آماره آزمون**

برای بررسی تفاوت معنی‏داری آماری بین دو مقدار شمارش شده، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# setting initial parameter values  
x = data**$**Rasht  
x.t = **length**(x = x)  
x.n = **sum**(x **==** 0)  
y = data**$**Ramsar  
y.t = **length**(x = y)  
y.n = **sum**(y **==** 0)  
  
**if** (x.t **==** y.t) {  
 # test statistic  
 FF = x.n **/** (y.n **+** 1)  
   
 # degrees of freedom  
 df1 = 2 **\*** (y.n **+** 1)  
 df2 = 2 **\*** x.n  
} **else** {  
 # test statistic  
 FF = ((1**/**x.t) **\*** (x.n **+** 0.5)) **/** ((1**/**y.t) **\*** (y.n **+** 0.5))  
   
 # degrees of freedom  
 df1 = (2 **\*** x.n) **+** 1  
 df2 = (2 **\*** y.n) **+** 1  
}  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
**if** (alternative **!=** "two.sided") {  
 lower.critical = **qf**(p = (1 **-** alpha), df1 = df1, df2 = df2, lower.tail = FALSE)  
 upper.critical = **qf**(p = (1 **-** alpha), df1 = df1, df2 = df2, lower.tail = TRUE)  
} **else** {  
 lower.critical = **qf**(p = (1 **-** (alpha **/** 2)), df1 = df1, df2 = df2, lower.tail = FALSE)  
 upper.critical = **qf**(p = (1 **-** (alpha **/** 2)), df1 = df1, df2 = df2, lower.tail = TRUE)  
}  
  
p\_value = **pf**(q = FF, df1 = df1, df2 = df2, lower.tail = TRUE)  
p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value)  
  
**cat**("Test Statistic:", FF, "\n")  
**cat**("Degrees of Freedom:", **c**(df1, df2), "\n")  
**cat**("Critical Values:", **c**(lower.critical, upper.critical), "\n")  
**cat**("P-value:", p\_value, "\n")

## Test Statistic: 1.1096  
## Degrees of Freedom: 146 162  
## Critical Values: 0.7270 1.3718  
## P-value: 0.5179

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "f", statistic = FF, df1 = df1, df2 = df2,  
 alpha = 0.05, alternative = "two.sided", from = 0.5, to = 2)



همانطور که مشخص است، مقادیر بحرانی در برابر با (37/1 – 73/0) است (جدول شماره 3). مقدار محاسبه شده برای بین دو مقادیر بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که تعداد روزهای کاملا ابری ایستگاه‏های هواشناسی رشت و رامسر از نظر آماری تفاوت معنی‏داری با یکدیگر ندارند.

# آزمون 24: آزمون برای بررسی نرمال بودن یک جامعه

**هدف**

بررسی معنی‏دار بودن تفاوت بین یک توزیع فراوانی بر اساس یک نمونه گرفته شده و توزیع فراوانی نرمال[[108]](#footnote-109).

* فرضیه صفر (): نمونه از یک جامعه با توزیع فراوانی نرمال گرفته شده است.
* فرضیه مقابل (): نمونه از یک جامعه با توزیع فراوانی نرمال گرفته نشده است.

**پیش فرض‏ها**

1. این آزمون موقعی کاربردی می‏باشد که نمونه از جامعه‏ای با توزیع پیوسته[[109]](#footnote-110) گرفته شده باشد.

**روش**

این آزمون بسیار ساده‏تر از آزمون گشتاور فیشر (آزمون 20) می‏باشد. در ابتدا انحراف استاندارد نمونه و دامنه نمونه محاسبه می‏گردد.

آماره آزمون (دامنه استیودنت شده) برابر است با:

مقادیر بحرانی برای از جدول 14 به دست می‏آید. فرضیه صفر، نمونه از یک جامعه با توزیع فراوانی نرمال گرفته شده است، هنگامی رد می‏شود که مقدار محاسبه شده خارج از محدوده مقادیر بحرانی به دست آمده از جدول 14 باشد.

**مثال**

محققی در نظر دارد بررسی کند که آیا مجموعه 113 سال داده‏های مجموع بارش سالانه ایستگاه پاریس فرانسه، از توزیع نرمال پیروی میکند یا خیر؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_33.csv", header = TRUE)  
cv\_table = **read.csv**(file = "data/Table\_14.csv", header = TRUE)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Year | Station | Prec. |
| ## | 1 | 1900 | Paris | 497.3 |
| ## | 2 | 1901 | Paris | 545.3 |
| ## | … | … | … | … |
| ## | 112 | 2016 | Paris | 657.4 |
| ## | 113 | 2017 | Paris | 740.8 |

**گام 2: بررسی محدودیت‏ها**

نمونه از جامعه‏ای با توزیع پیوسته گرفته شده باشد.

می‏توان با رسم نمودار چندک چندک نیز نرمال بودن داده‏ها را مورد ارزیابی قرار داد. برای این منظور، از تابع **ggqqplot** و **ggdensity** موجود در بسته **ggpubr** استفاده شده است.

# normality plot  
ggpubr**::ggqqplot**(data = data**$**Prec, ylab = "Precipitation (mm)")  
# visualize data using density plot  
ggpubr**::ggdensity**(data = data, x = "Prec", fill = "gray",  
 xlab = "Precipitation (mm)", add = **c**("mean"))

|  |  |
| --- | --- |
|  |  |

**گام 3: محاسبه آماره آزمون**

برای بررسی معنی‏دار بودن تفاوت بین یک توزیع فراوانی بر اساس یک نمونه گرفته شده و توزیع فراوانی نرمال، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# setting initial parameter values  
x = data**$**Prec  
x.n = **length**(x = x)  
x.m = **mean**(x)  
  
# sample standard deviation (s)  
x.s = **sqrt**(x = ((**sum**((x **-** x.m) **^** 2) **/** (x.n **-** 1)))) # x.s = sd(x = x)  
  
# sample range (w)  
x.sorted = **sort**(x = x, decreasing = FALSE)  
x.w = x.sorted[x.n] **-** x.sorted[1]  
  
# test statistic  
q = x.w **/** x.s  
  
# compute the critical values and p-value  
alpha = "0.050"  
  
  
cv = **switch**(EXPR = alpha,  
 "0.000" = cv\_table **%>%** **filter**(n **>=** x.n) **%>%** **select**(a000, b000) **%>%** **slice**(1),  
 "0.005" = cv\_table **%>%** **filter**(n **>=** x.n) **%>%** **select**(a005, b005) **%>%** **slice**(1),  
 "0.010" = cv\_table **%>%** **filter**(n **>=** x.n) **%>%** **select**(a010, b010) **%>%** **slice**(1),  
 "0.025" = cv\_table **%>%** **filter**(n **>=** x.n) **%>%** **select**(a025, b025) **%>%** **slice**(1),  
 "0.050" = cv\_table **%>%** **filter**(n **>=** x.n) **%>%** **select**(a050, b050) **%>%** **slice**(1),  
 "0.100" = cv\_table **%>%** **filter**(n **>=** x.n) **%>%** **select**(a100, b100) **%>%** **slice**(1))  
  
**cat**("Test Statistic:", q, "\n")  
**cat**("Critical Values:", **c**(cv[[1]], cv[[2]]), "\n")  
**if** (q **<** cv[[1]] **&** q **>** cv[[2]]) {  
 **cat**("Reject H0: Sample Distribution Cannot Be Considered as a Normal Distribution", "\n")  
} **else** {  
 **cat**("Accept H0: Sample Distribution Can Be Considered as a Normal Distribution", "\n")  
}

## Test Statistic: 5.6602  
## Critical Values: 4.59 6.18  
## Accept H0: Sample Distribution Can Be Considered as a Normal Distribution

همانطور که مشخص است، مقادیر بحرانی در برابر با (18/6 – 59/4) است (جدول شماره 14). مقدار محاسبه شده برای آماره آزمون، 66/5، بین مقادیر بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که مجموعه 113 سال داده‏های مجموع بارش سالانه ایستگاه پاریس فرانسه، از توزیع نرمال تبعیت میکند.

# آزمون 25: آزمون کوکران[[110]](#footnote-111) برای نقاط دورافتاده واریانس[[111]](#footnote-112)

**هدف**

بررسی معنی‏دار بودن تفاوت بین یک واریانس نسبتا بزرگ و واریانس دیگر.

* فرضیه صفر (): همه مقادیر واریانس‏ها با هم برابر هستند.
* فرضیه صفر (): حداقل یک مقدار واریانس به طور معنی‏دار، بزرگتر از بقیه مقادیر واریانس‏ها می‏باشد.

**پیش فرض‏ها**

1. فرض می‏شود که نمونه از جوامعی با توزیع نرمال گرفته شده باشند.
2. حجم نمونه باید با هم برابر باشد.

**روش**

آماره آزمون برابر است با:

که در اینجا نشان دهنده واریانس امین نمونه می‏باشد که با استفاده از فرمول زیر محاسبه می‏گردد:

مقادیر بحرانی از جدول 15 به دست می‏آید. اگر مقدار آماره آزمون محاسبه شده بزرگتر از مقادیر بحرانی باشد، فرضیه صفر، مقدار واریانس بزرگ تفاوت معنی‏داری با بقیه مقادیر واریانس‏ها ندارد، رد خواهد شد.

**مثال**

به طور کاملا تصادفی 25 مزرعه گندم از 5 شهرستان مشهد، سبزوار، تربت‏حیدریه، نیشابور و قوچان انتخاب و عملکرد آن‏ها به دست آمده است. با فرض اینکه جامعه عملکرد گندم هر یک از این استان‏ها به صورت نرمال توزیع شده باشد، آیا واریانس‏های آن‏ها با هم برابر است؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "outliers")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_34.csv", header = TRUE)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## |  | S1 | S2 | S3 | S4 | S5 |
| ## | 1 | 4521 | 5223 | 4177 | 3073 | 4167 |
| ## | 2 | 2980 | 4569 | 2846 | 1752 | 2593 |
| ## | … | … | … | … | … | … |
| ## | 24 | 3447 | 7905 | 3768 | 4684 | 2949 |
| ## | 25 | 4276 | 5902 | 3635 | 6039 | 3336 |

**گام 2: بررسی محدودیت‏ها**

فرض می‏شود که K نمونه از جوامعی با توزیع نرمال گرفته شده است و حجم آن‏ها نیز با هم برابر است.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها رسم شده است. برای این منظور، از تابع **ggboxplot** موجود در بسته **ggpubr** استفاده شده است.

dataPlot = data **%>%** reshape2**::melt**()  
  
# visualizing samples with box plots  
ggpubr**::ggboxplot**(data = dataPlot, x = "variable", y = "value",  
 add = "mean", xlab = "Station", ylab = "Yield (kg/ha)")



**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی معنی‏دار بودن تفاوت بین یک واریانس نسبتا بزرگ و واریانس دیگر، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
n = n1 = n2 = n3 = n4 = n5 = **nrow**(x = data)  
k = **ncol**(x = data)  
  
d1 = data **%>%** **select**(S1) **%>%** **unlist**()  
v1 = **var**(x = d1)  
  
d2 = data **%>%** **select**(S2) **%>%** **unlist**()  
v2 = **var**(x = d2)  
  
d3 = data **%>%** **select**(S3) **%>%** **unlist**()  
v3 = **var**(x = d3)  
  
d4 = data **%>%** **select**(S4) **%>%** **unlist**()  
v4 = **var**(x = d4)  
  
d5 = data **%>%** **select**(S5) **%>%** **unlist**()  
v5 = **var**(x = d5)  
  
v = **c**(v1, v2, v3, v4, v5)  
  
# test statistic  
C = **max**(v) **/** **sum**(v)  
  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "greater"  
  
**if** (alternative **!=** "two.sided") {  
 cochran.critical = outliers**::qcochran**(p = (1 **-** alpha), n = n, k = k)  
} **else** {  
 cochran.critical = outliers**::qcochran**(p = (1 **-** (alpha **/** 2)), n = n, k = k)  
}  
  
p\_value = outliers**::pcochran**(q = C, n = n, k = k)  
p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value)  
  
**cat**("Test Statistic: ", C, "\n")  
**cat**("Degrees of Freedom: ", c(k, n – 1), "\n")  
**cat**("Critical Values: ", cochran.critical, "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: 0.4607  
## Degrees of Freedom: 5 24  
## Critical Values: 0.3323  
## P-value: 5.0001e-05

همانطور که مشخص است، مقدار بحرانی در برابر با 33/0 است (جدول شماره 15). مقدار محاسبه شده برای بیشتر از مقادیر بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که واریانس عملکرد گندم در نمونه شهرستان نیشابور با بقیه شهرستان‏ها تفاوت دارد (واریانس‏ها همگن نیستند).

در روش دوم، برای بررسی معنی‏دار بودن تفاوت بین یک واریانس نسبتا بزرگ و واریانس دیگر، از تابع **cochran.test** موجود در بسته **outliers** استفاده شده است.

# method 2  
# use "cochran.test" function from "outliers" package  
outliers**::cochran.test**(object = v,  
 data = **rep**(x = n, k),  
 inlying = FALSE)

## Cochran test for outlying variance  
## data: v  
## C = 0.4607, df = 25, k = 5, p-value = 5.001e-05  
## alternative hypothesis: Group 4 has outlying variance  
## sample estimates:  
## 1 2 3 4 5   
## 897858.2 2180636.2 314530.0 3372150 4 554793.9

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 26: آزمون وارون ویلکاکسون[[112]](#footnote-113) (آزمون )

**هدف**

برای آزمون کردن اینکه آیا دو نمونه تصادفی انتخاب شده می‏توانند از دو جامعه با توزیع فراوانی یکسان باشند.

* فرضیه صفر (): توزیع فراوانی دو جامعه یکسان است.
* فرضیه مقابل (): توزیع فراوانی دو جامعه یکسان نیست.

**پیش فرض‏ها**

1. فرض می‏شود که توزیع فراوانی دو جامعه به صورت پیوسته می‏باشد و دو نمونه، تصادفی و مستقل از یکدیگر می‏باشند.

**روش**

از جامعه اول، نمونه‏های تصادفی با حجم و از جامعه دوم، نمونه‏های تصادفی با حجم انتخاب می‏شود. دو نمونه با هم ترکیب شده و به ترتیب صعودی مرتب می‏گردند. آنگاه تعدادی پرش‏ (یا وارون‏)

از یک سری به سری دیگر خواهیم داشت. آماره آزمون برابر است کوچکترین مقدار از بین تعداد وارون‏ها و تعداد غیر وارون‏ها.

با در نظر گرفتن فرضیه صفر، توزیع فراوانی دو جامعه یکسان است، اگر مقدار آماره آزمون از مقدار حدی به دست آمده از جدول 20 بیشتر باشد، فرضیه صفر رد می‏گردد.

**مثال**

متوسط سالانه دمای خاک دو ایستگاه مشهد (40 سال) و کرمان (30 سال) محاسبه شده است. در سطح معنی‏داری ، آیا فرضیه صفر (توزیع فراوانی متوسط سالانه دمای خاک ایستگاه‏های مشهد و کرمان یکسان است) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_49.csv", header = TRUE)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Year | Mashhad | Kerman |
| ## | 1 | 1979 | 5.50 | NA |
| ## | 2 | 1980 | 5.35 | NA |
| ## | … | … | … | … |
| ## | 39 | 2017 | 5.44 | 6.83 |
| ## | 40 | 2018 | 6.05 | 6.98 |

**گام 2: بررسی محدودیت‏ها**

1- توزیع فراوانی دو جامعه به صورت پیوسته می‏باشد.

2- دو نمونه، مستقل از یکدیگر می‏باشند.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها رسم شده است. برای این منظور، از تابع **ggboxplot** موجود در بسته **ggpubr** استفاده شده است.

data = data **%>%**  
 reshape2**::melt**(id = **c**("Year"), variable.name = "Station", value.name = "Soil.Temp") **%>%**  
 **na.omit**()  
  
# Visualize Data Using Box Plots:  
ggpubr**::ggboxplot**(data = data, x = "Station", y = "Soil.Temp",  
 ylab = "Soil Temp. (C)", xlab = FALSE,  
 add = **c**("mean"))



**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای آزمون کردن اینکه آیا دو نمونه تصادفی انتخاب شده می‏توانند از دو جامعه با توزیع فراوانی یکسان باشند، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
x = data **%>%** **filter**(Station **==** "Mashhad") **%>%** **select**(Soil.Temp) **%>%** **unlist**()  
x.n = **length**(x = x)  
y = data **%>%** **filter**(Station **==** "Kerman") **%>%** **select**(Soil.Temp) **%>%** **unlist**()  
y.n = **length**(x = y)  
  
# test Statistic  
r = **rank**(x = **c**(x, y))  
W = **sum**(r[**seq\_along**(along.with = x)]) **-** (x.n **\*** (x.n **+** 1) **/** 2)  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
**if** (alternative **!=** "two.sided") {  
 lower.critical = **qwilcox**(p = (1 **-** alpha), m = x.n, n = y.n, lower.tail = FALSE)  
 upper.critical = **qwilcox**(p = (1 **-** alpha), m = x.n, n = y.n, lower.tail = TRUE)  
} **else** {  
 lower.critical = **qwilcox**(p = (1 **-** (alpha **/** 2)), m = x.n, n = y.n, lower.tail = FALSE)  
 upper.critical = **qwilcox**(p = (1 **-** (alpha **/** 2)), m = x.n, n = y.n, lower.tail = TRUE)  
}  
p\_value = **switch**(EXPR = alternative,  
 two.sided = {  
 **if** (W **>** ((x.n **\*** y.n) **/** 2)) {  
 p = **pwilcox**(q = (W **-** 1), m = x.n, n = y.n, lower.tail = FALSE)  
 } **else** {  
 **pwilcox**(q = W, m = x.n, n = y.n)  
 }  
 **min**(2 **\*** p, 1)  
 },  
 less = **pwilcox**(q = W, m = x.n, n = y.n),  
 greater = **pwilcox**(q = (W **-** 1), m = x.n, n = y.n, lower.tail = FALSE))  
  
**cat**("Test Statistic: ", W, "\n")  
**cat**("Critical Values: ", **c**(lower.critical, upper.critical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Test Statistic: 748  
## Critical Values: 435 765  
## P-value: 0.08

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع ویلکاکسون با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "wilcox", statistic = W, m = x.n, n = y.n,  
 alpha = 0.05, alternative = "two.sided", from = 300, to = 900)



همانطور که مشخص است، در آزمون دو دنباله، مقادیر بحرانی در برابر با (765 - 435) است (جدول شماره 20). مقدار محاسبه شده برای بین مقادیر بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که توزیع فراوانی متوسط سالانه دمای خاک ایستگاه‏های مشهد و کرمان یکسان است.

در روش دوم، برای آزمون کردن اینکه آیا دو نمونه تصادفی انتخاب شده می‏توانند از دو جامعه با توزیع فراوانی یکسان باشند، از تابع **wilcox.test** موجود در بسته **stats** استفاده شده است.

# method 2  
# use "wilcox.test" function from "stats" package  
**wilcox.test**(x = x, y = y, alternative = "two.sided",   
 conf.level = 0.95, exact = FALSE)

## Wilcoxon rank sum test with continuity correction  
## data: x and y  
## W = 748, p-value = 0.08  
## alternative hypothesis: true location shift is not equal to 0

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 27: آزمون مجموع رتبه‏های ویلکاکسون – من – ویتنی[[113]](#footnote-114) برای دو جامعه

**هدف**

برای آزمون کردن اینکه آیا دو نمونه تصادفی انتخاب شده می‏توانند از دو جامعه با میانگین مساوی باشند.

* فرضیه صفر (): میانگین دو جامعه یکسان است.
* فرضیه مقابل (): میانگین دو جامعه یکسان نیست.

**پیش فرض‏ها**

1. فرض می‏شود که توزیع فراوانی دو جامعه به صورت پیوسته با شکل و پراکنش یکسان می‏باشند.

**روش**

از جامعه اول، نمونه‏های تصادفی با حجم و از جامعه دوم، نمونه‏های تصادفی با حجم انتخاب می‏شود. دو نمونه با هم ترکیب شده و به ترتیب صعودی مرتب و رتبه‏بندی می‏گردند. در صورتی که مقادیر مشابه وجود داشته باشد، میانگین شماره رتبه‏ها در نظر گرفته می‏شود. مجموع رتبه‏های دو نمونه محاسبه می‏گردد. نمونه‏ای که حجم کوچکتری دارد، مجموع رتبه‏های آن به عنوان کمیت اول در نظر گرفته می‏شود. از طرفی نشان دهنده مجموع حجم دو نمونه، و نشان دهنده کوچکترین حجم نمونه می‏باشد. کمیت دوم برابر است با:

مقادیر و با مقادیر بحرانی جدول 21 مقایسه می‏شوند. اگر مقدار بحرانی جدول 21 از هر یک از دو کمیت و بزرگتر باشد، فرضیه صفر، میانگین دو جامعه یکسان است، رد خواهد شد.

نکته: اگر حجم دو نمونه با هم برابر باشد، کمیت اول برابر است با کوچکترین مجموع رتبه‏های دو نمونه.

**مثال**

کارخانه‏‏ای دارای دو خط تولید نوار آبیاری پلاکدار (تیپ) می‏باشد. خط تولید شماره یک از پلاک‏های خارجی، و خط تولید شماره دو از پلاک‏های ایرانی با مشخصات مشابه، برای تولید نوار تیپ پلاکدار استفاده می‏کنند. به صورت تصادفی 15 مزرعه گندم که از تیپ نوع و 10 مزرعه گندم که از تیپ نوع استفاده کرده‏اند انتخاب می‏گردد. در سطح معنی‏داری ، آیا فرضیه صفر (تفاوتی بین میانگین عملکرد گندم در مزارعی که از تیپ نوع استفاده کرده‏اند نسبت به مزارعی که از تیپ نوع استفاده کرده‏اند وجود ندارد) رد خواهد شد (با فرض اینکه همه شرایط یکسان باشد)؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_52.csv", header = TRUE)  
  
data = data **%>%**  
 reshape2**::melt**(variable.name = "Tape", value.name = "Yield") **%>% na.omit**()  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |
| --- | --- | --- | --- |
| ## |  | Tape | Yield |
| ## | 1 | A | 4289 |
| ## | 2 | A | 4398 |
| ## | … | … | … |
| ## | 24 | B | 3840 |
| ## | 25 | B | 3673 |

**گام 2: بررسی محدودیت‏ها**

1- توزیع فراوانی دو جامعه به صورت پیوسته با شکل و پراکنش یکسان می‏باشد.

برای توصیف تغییرات نمونه‏ها، نمودار جعبه‏ای آن‏ها رسم شده است. برای این منظور، از تابع **ggboxplot** موجود در بسته **ggpubr** استفاده شده است.

# visualize data using box plots  
ggpubr**::ggboxplot**(data = data, x = "Tape", y = "Yield",  
 ylab = "Yield (kg/ha)", xlab = FALSE,  
 add = **c**("mean"))



**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای آزمون کردن اینکه آیا دو نمونه تصادفی انتخاب شده می‏توانند از دو جامعه با میانگین مساوی باشند، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# setting initial parameter values  
rankedData = data %>%   
 dplyr::mutate(Yield = round(x = Yield, digits = 2)) %>%   
 dplyr::mutate(Rank = rank(x = Yield))  
  
nA = sum(rankedData$Tape == "A")  
RA = sum(rankedData %>% filter(Tape == "A") %>% select(Rank))  
  
nB = sum(rankedData$Tape == "B")  
RB = sum(rankedData %>% filter(Tape == "B") %>% select(Rank))  
  
# size of the combined samples   
N = nA + nB  
# rank sum R of the smaller sample and n denote the size of the smaller sample  
if (nA < nB) {  
 n = nA  
 R = RA  
} else if (nA > nB) {  
 n = nB  
 R = RB  
} else {  
 if (RA < RB) {  
 n = nA  
 R = RA  
 } else {  
 n = nB  
 R = RB  
 }  
}  
  
# second quantity  
Rprime = (n \* (N + 1)) - R  
  
cat("n1:", max(nA, nB), "\n")  
cat("n2:", min(nA, nB), "\n")  
cat("R:", R, "\n")  
cat("R':", Rprime, "\n")

## n1: 15  
## n2: 10  
## R: 120  
## R': 140

با استفاده از جدول 21، مقدار بحرانی آزمون در آزمون دو طرفه و به ازای و ، برابر با 94 است. مقدار محاسبه شده آماره آزمون بزرگتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد نخواهد شد. این بدان معنا است که تفاوتی بین میانگین عملکرد گندم در مزارعی که از تیپ نوع استفاده کرده‏اند نسبت به مزارعی که از تیپ نوع استفاده کرده‏اند وجود ندارد.

# آزمون 28: آزمون همبستگی رتبه‏ای اسپیرمن[[114]](#footnote-115) (برای مشاهدات جفت شده)

**هدف**

بررسی معنی‏دار بودن همبستگی بین دو سری از مشاهدات جفت شده.

* فرضیه صفر (): عدم همبستگی بین دو سری از مشاهدات جفت شده.
* فرضیه مقابل (): وجود همبستگی بین دو سری از مشاهدات جفت شده.

**پیش فرض‏ها**

1. فرض می‏شود که توزیع دو جامعه به صورت پیوسته بوده و مشاهدات و به صورت جفتی انتخاب شده‏اند.

**روش**

مشاهدات به صورت صعودی مرتب شده و مقادیر به آن‏ها نسبت داده می‏شود. برای مشاهدات نیز همین اعمال تکرار می‏شود. برای هر جفت از مشاهدات، تفاوت رتبه‏ها، ، تعیین می‏شود. در اینصورت کمیت برابر است با:

برای نمونه‏های بزرگ آماره آزمون برابر است با:

با در نظر گرفتن فرضیه صفر، عدم همبستگی بین دو سری از مشاهدات جفت شده، می‏توان آماره آزمون را با توزیع نرمال استاندارد مقایسه کرد. آزمون می‏تواند به هر یک از دو صورت یک دنباله[[115]](#footnote-116) یا دو دنباله[[116]](#footnote-117) انجام گیرد. برای نمونه‏های کوچک، آماره آزمون برابر است با:

که با در نظر گرفتن فرضیه صفر، عدم همبستگی بین دو سری از مشاهدات جفت شده، می‏توان آماره آزمون را با مقدار بحرانی به دست آمده از جدول 26 مقایسه کرد.

در هر دو حالت، اگر آماره آزمون بیشتر از مقدار بحرانی باشد، فرضیه صفر، عدم همبستگی بین دو سری از مشاهدات جفت شده، رد می‏شود.

**مثال**

هدف از پژوهشی بررسی وجود همبستگی بین مجموع بارندگی سالانه دو ایستگاه آبادان و اهواز می‏باشد. در سطح معنی‏داری ، آیا فرضیه صفر (عدم همبستگی بین مجموع بارندگی سالانه دو ایستگاه آبادان و اهواز) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "pspearman", "PerformanceAnalytics")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_58.csv", header = TRUE)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Year | Abadab | Ahvaz |
| ## | 1 | 1959 | 102.3 | 120.1 |
| ## | 2 | 1960 | 87 | 117.9 |
| ## | … | … | … | … |
| ## | 59 | 2017 | 73.7 | 77.5 |
| ## | 60 | 2018 | 182.8 | 307.8 |

**گام 2: بررسی محدودیت‏ها**

1- فرض می‏شود که توزیع دو جامعه به صورت پیوسته بوده و مشاهدات و به صورت جفتی انتخاب شده‏اند.

به منظور تجسم رابطه بین این دو ایستگاه، نمودار نقطه‏ای آن‏ها با استفاده از تابع **chart.Correlation** از بسته **PerformanceAnalytics** و تابع **ggscatter** از بسته **ggpubr** رسم شده است.

# visualize data using scatter plots  
PerformanceAnalytics**::chart.Correlation**(R = data[, **c**("Abadan", "Ahvaz")],  
 method = "spearman")  
  
data **%>%** ggpubr**::ggscatter**(x = "Abadan", y = "Ahvaz",  
 add = "reg.line", conf.int = TRUE,  
 xlab = "Abadan Prec. (mm)", ylab = "Ahvaz Prec. (mm)")

|  |  |
| --- | --- |
|  |  |

**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، بررسی معنی‏دار بودن همبستگی بین دو سری از مشاهدات جفت شده، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
x = data**$**Abadan  
y = data**$**Ahvaz  
n = **length**(x = x)  
  
  
# rank the x and y variables  
rank\_x = **rank**(x = x)  
rank\_y = **rank**(x = y)  
  
# the observed value of R statistics  
R = **sum**((rank\_x **-** rank\_y) **^** 2)  
  
**if** (n **>** 10) {  
 # spearman's ρ statistic  
 rho = 1 **-** ((6 **\*** R) **/** (n **\*** ((n **^** 2) **-** 1)))  
   
 test statistic  
 Z = ((6 **\*** R) **-** (n **\*** ((n **^** 2) **-** 1))) **/** (n **\*** (n **+** 1) **\*** **sqrt**(n **-** 1))  
   
 # compute the critical values and p-value  
 alpha = 0.05  
 alternative = "two.sided"  
   
 **if** (alternative **!=** "two.sided") {  
 lower.critical = **qnorm**(p = (1 **-** alpha), lower.tail = FALSE)  
 upper.critical = **qnorm**(p = (1 **-** alpha), lower.tail = TRUE)  
 } **else** {  
 lower.critical = **qnorm**(p = (1 **-** (alpha **/** 2)), lower.tail = FALSE)  
 upper.critical = **qnorm**(p = (1 **-** (alpha **/** 2)), lower.tail = TRUE)  
 }  
   
 p\_value = **pnorm**(q = Z, lower.tail = TRUE)  
 p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value)  
   
 **cat**("Observed Value of R Statistics:", R, "\n")  
 **cat**("Spearman's ρ Statistic:", rho, "\n")  
 **cat**("Test Statistic:", Z, "\n")  
 **cat**("Critical Values:", **c**(lower.critical, upper.critical), "\n")  
 **cat**("P-value:", p\_value, "\n")  
   
 plotDistStat(distribution = "norm", statistic = Z, alpha = 0.05,  
 alternative = "two.sided")  
} **else** {  
 # spearman's ρ statistic  
 rho = 1 **-** ((6 **\*** S) **/** (n **\*** ((n **^** 2) **-** 1)))  
   
 # degrees of freedom  
 df = n **-** 2  
  
 # compute the critical values  
 alpha = 0.05  
   
 lower.critical = **qt**(p = (1 **-** alpha), df = df, lower.tail = FALSE)  
 lower.critical.r = psych**::t2r**(t = lower.critical, df = df)  
   
 upper.critical = **qt**(p = (1 **-** alpha), df = df, lower.tail = TRUE)  
 upper.critical.r = psych**::t2r**(t = upper.critical, df = df)  
   
 **cat**("Observed Value of R Statistics:", R, "\n")  
 **cat**("Spearman's ρ Statistic (Test Statistic):", rho, "\n")  
 **cat**("Critical Values:", **c**(lower.critical.r, upper.critical.r), "\n")  
}

## Observed Value of R Statistics: 9556   
## Spearman's ρ Statistic: 0.7345   
## Test Statistic: -5.6417   
## Critical Values: -1.96 1.96   
## P-value: 1.6842e-08

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع نرمال استاندارد با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.



همانطور که مشخص است، در آزمون دو دنباله، مقدار بحرانی در برابر با 96/1 است (جدول شماره 1). مقدار محاسبه شده برای آماره آزمون بزرگتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که همبستگی بین مجموع بارندگی سالانه دو ایستگاه آبادان و اهواز وجود دارد.

در روش دوم، بررسی معنی‏دار بودن همبستگی بین دو سری از مشاهدات جفت شده، از تابع **cor.test** موجود در بسته **stats** استفاده شده است.

# method 2  
# use "cor.test" function from "stats" package   
**cor.test**(x = data**$**Abadan, y = data**$**Ahvaz,  
 alternative = "two.sided", method = "spearman")

## Spearman's rank correlation rho  
## data: data$Abadan and data$Ahvaz  
## S = 9556, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho: 0.7345

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 29: آزمون همبستگی رتبه‏ای کندال[[117]](#footnote-118) (برای مشاهدات جفت شده)

**هدف**

بررسی معنی‏دار بودن همبستگی بین دو سری از مشاهدات جفت شده.

* فرضیه صفر (): عدم همبستگی بین دو سری از مشاهدات جفت شده.
* فرضیه مقابل (): وجود همبستگی بین دو سری از مشاهدات جفت شده.

**پیش فرض‏ها**

1. فرض می‏شود که توزیع دو جامعه به صورت پیوسته بوده و مشاهدات و به صورت جفتی انتخاب شده‏اند.

**روش**

مشاهدات به صورت صعودی مرتب شده و مقادیر به آن‏ها نسبت داده می‏شود. برای مشاهدات نیز همین اعمال تکرار می‏شود. همه جفت‏های ممکن رتبه‏ها (که تعداد آن‏ها می‏باشد) مورد بررسی قرار می‏گیرد. هر جفت با جفت به طور متوالی و منظم مورد مقایسه قرار می‏گیرد. هنگامیکه و دارای علامت یکسان باشند، امتیاز در نظر گرفته می‏شود. اگر علامت‏ها باهم یکسان نباشند، امتیاز در نظر گرفته می‏شود. اگر حاصل صفر باشد، امتیازی در نظر گرفته نمی‏شود. این امتیازها با هم جمع شده و این مجموع با نمایش داده می‏شود.

برای نمونه‏های بزرگ ، آماره آزمون برابر است با:

با در نظر گرفتن فرضیه صفر، عدم همبستگی بین دو سری از مشاهدات جفت شده، می‏توان آماره آزمون را با توزیع نرمال استاندارد مقایسه کرد. آزمون می‏تواند به هر یک از دو صورت یک دنباله[[118]](#footnote-119) یا دو دنباله[[119]](#footnote-120) انجام گیرد. برای نمونه‏های کوچک با در نظر گرفتن به عنوان آماره آزمون، می‏توان آن را با مقدار بحرانی به دست آمده از جدول 27 مقایسه کرد.

در هر دو حالت، اگر آماره آزمون بیشتر از مقدار بحرانی باشد، فرضیه صفر، عدم همبستگی بین دو سری از مشاهدات جفت شده، رد می‏شود.

**مثال**

هدف از پژوهشی بررسی وجود همبستگی بین میانگین دمای بیشینه سالانه دو ایستگاه آبادان و اهواز می‏باشد. در سطح معنی‏داری ، آیا فرضیه صفر (عدم همبستگی بین میانگین دمای بیشینه سالانه دو ایستگاه آبادان و اهواز) رد خواهد شد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "pspearman", "PerformanceAnalytics")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_59.csv", header = TRUE)  
  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Year | Abadab | Ahvaz |
| ## | 1 | 1959 | 32.44 | 31.71 |
| ## | 2 | 1960 | 33.41 | 33.56 |
| ## | … | … | … | … |
| ## | 59 | 2017 | 35.52 | 35.13 |
| ## | 60 | 2018 | 34.92 | 34.74 |

**گام 2: بررسی محدودیت‏ها**

1- فرض می‏شود که توزیع دو جامعه به صورت پیوسته بوده و مشاهدات و به صورت جفتی انتخاب شده‏اند.

به منظور تجسم رابطه بین این دو ایستگاه، نمودار نقطه‏ای آن‏ها با استفاده از تابع **chart.Correlation** از بسته **PerformanceAnalytics** و تابع **ggscatter** از بسته **ggpubr** رسم شده است.

# visualize data using scatter plots  
PerformanceAnalytics**::chart.Correlation**(R = data[, **c**("Abadan", "Ahvaz")],  
 method = "kendall")  
  
data **%>%** ggpubr**::ggscatter**(x = "Abadan", y = "Ahvaz",  
 add = "reg.line", conf.int = TRUE,  
 xlab = "Abadan Max Temp. (C)",  
 ylab = " Ahvaz Max Temp. (C)")

|  |  |
| --- | --- |
|  |  |

**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، بررسی معنی‏دار بودن همبستگی بین دو سری از مشاهدات جفت شده، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# setting initial parameter values  
x = data**$**Abadan  
y = data**$**Ahvaz  
n = **length**(x = x)  
mat = **matrix**(data = 0, nrow = n **-** 1, ncol = n **-** 1)  
**colnames**(x = mat) = 1**:**(n **-** 1)  
**rownames**(x = mat) = 2**:**n  
# calculate for each yi, count the number of yj > yi (Nc: Concordant Pairs ) and   
# the number of yj < yi (Nd: Discordant Pairs)  
**for** (i **in** 1**:**(n **-** 1)) {  
 **for** (j **in** (i **+** 1)**:**n) {  
 val = (y[j] **-** y[i]) **\*** (x[j] **-** x[i])  
 **if** (val **>** 0) {  
 mat[j **-** 1, i] = 1  
 } **else** {  
 mat[j **-** 1, i] = -1  
 }  
 }  
}  
  
# total number of concordant pairs (Nc) - Total number of discordant pairs (Nd)  
K = **sum**(mat)  
# kendall correlation distance  
tau = K **/** (n **\*** (n **-** 1) **/** 2)  
# a z-score for rho which approximately follows a standard normal distribution  
# under the null hypothesis.  
Z = K **/** (**sqrt**((n **\*** (n **-** 1) **\*** (2 **\*** n **+** 5)) **/** 18))  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "two.sided"  
  
**if** (alternative **!=** "two.sided") {  
 lower.critical = **qnorm**(p = (1 **-** alpha), lower.tail = FALSE)  
 upper.critical = **qnorm**(p = (1 **-** alpha), lower.tail = TRUE)  
} **else** {  
 lower.critical = **qnorm**(p = (1 **-** (alpha **/** 2)), lower.tail = FALSE)  
 upper.critical = **qnorm**(p = (1 **-** (alpha **/** 2)), lower.tail = TRUE)  
}  
  
p\_value = **pnorm**(q = Z, lower.tail = TRUE)  
p\_value = **switch**(EXPR = alternative,  
 two.sided = 2 **\*** **min**(p\_value, 1 **-** p\_value),  
 less = p\_value,  
 greater = 1 **-** p\_value  
  
**cat**("Nc - Nd: ", K, "\n")  
**cat**("tau: ", tau, "\n")  
**cat**("z-score: ", Z, "\n")  
**cat**("Critical Values: ", **c**(lower.critical, upper.critical), "\n")  
**cat**("P-value: ", p\_value, "\n")

## Nc - Nd: 1256  
## tau: 0.71  
## Test Statistic: 8.01  
## Critical Values: -1.96 1.96   
## P-value: 1.11e-15

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع نرمال استاندارد با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "norm", statistic = Z, alpha = 0.05,  
 alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله، مقدار بحرانی در برابر با 96/1 است (جدول شماره 1). مقدار محاسبه شده برای آماره آزمون بزرگتر از مقدار بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که همبستگی بین میانگین دمای بیشینه سالانه دو ایستگاه آبادان و اهواز وجود دارد.

در روش دوم، بررسی معنی‏دار بودن همبستگی بین دو سری از مشاهدات جفت شده، از تابع **cor.test** موجود در بسته **stats** استفاده شده است.

# method 2  
# use "cor.test" function from "stats" package   
**cor.test**(x = data**$**Abadan, y = data**$**Ahvaz,  
 alternative = "two.sided", method = "kendall")

## Kendall's rank correlation tau  
## data: data$Abadan and data$Ahvaz  
## z = 8.01, p-value = 7.049e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho: 0.71

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

# آزمون 30: آزمون برای بررسی خطی بودن رگرسیون

**هدف**

بررسی خطی بودن رگرسیون بین متغیر و متغیر .

* فرضیه صفر ():
* فرضیه صفر ():

**پیش فرض‏ها**

به ازای مقادیر ، ها دارای توزیع نرمال و مستقل می‏باشند. جزء خطاها نیز دارای توزیع نرمال و مستقل با میانگین صفر می‏باشند.

**روش**

در این آزمون خطی بودن یا نبودن یک رابطه رگرسیونی بررسی می‏شود. در ابتدا فرض زیر مورد بررسی قرار می‏گیرد:

با توجه به فرضیه صفر

با درجه آزادی ، مجموع مربعات با توجه به رگرسیون برابر است با:

با درجه آزادی یک. و در نهایت آماره آزمون، نسبت میانگین مربعات، برابر است با:

که می‏توان آماره آزمون را با توزیع [[120]](#footnote-121) و با درجه‏های آزادی[[121]](#footnote-122) مقایسه کرد.

**مثال**

در پژوهشی هدف بررسی خطی بودن رابطه بین مجموع بارندگی سالانه ایستگاه‏های مشهد و سبزوار می‏باشد. در آزمون 81 وجود رابطه بین مجموع بارندگی سالانه دو ایستگاه بررسی و مشخص گردید، آیا این رابطه خطی می‏باشد؟

**کدهای زبان برنامه‏نویسی R**

**گام 1: آماده‏سازی داده‏ها**

در ابتدا بسته‏‏ها و داده‏های مورد نیاز فراخوانی شده و سپس قسمتی از داده‏های ورودی نمایش داده شده است.

# remove all variable  
**rm**(list = **ls**())  
  
# load required functions  
**source**(file = "muFunc.R")  
  
# load required packages  
wants = **c**("pacman", "dplyr", "psych", "ggpubr", "ggExtra")  
has = wants **%in%** **rownames**(x = **installed.packages**())  
**if** (**any**(**!**has)) **install.packages**(wants[**!**has])  
pacman**::p\_load**(char = wants)  
  
# load required data  
data = **read.csv**(file = "data/Data\_Test\_82.csv", header = TRUE)  
# show data  
psych**::headTail**(x = data, top = 2, bottom = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## |  | Year | Mashhad | Sabzevar |
| ## | 1 | 1959 | 234.5 | 145.4 |
| ## | 2 | 1960 | 174.5 | 112.2 |
| ## | … | … | … | … |
| ## | 59 | 2017 | 238.1 | 134.1 |
| ## | 60 | 2018 | 237.4 | 139.7 |

**گام 2: بررسی محدودیت‏ها**

1- فرض می‏شود به ازای مقادیر X، Yها دارای توزیع نرمال و مستقل می‏باشند.

به منظور تجسم رابطه بین این دو ایستگاه، نمودار نقطه‏ای آن‏ها با استفاده از تابع **ggscatter** از بسته **ggpubr** رسم شده است..

# visualize data using scatter plots  
data %>% ggpubr::ggscatter(x = "Mashhad", y = "Sabzevar",   
 add = "reg.line", conf.int = TRUE,  
 xlab = "Mashhad Prec. (mm)",  
 ylab = "Sabzevar Prec. (mm)") %>%  
 ggExtra::ggMarginal(type = "boxplot", fill = "gray90")



**گام 3: محاسبه آماره آزمون**

محاسبه آماره آزمون با دو روش انجام گرفته است. در روش اول، برای بررسی خطی بودن رگرسیون بین متغیر و متغیر ، محاسبات از ابتدا و به صورت مرحله به مرحله انجام گرفته است تا مقدار آماره آزمون و p-مقدار آن به دست آید.

# method 1  
# x variable  
x = data$Mashhad  
x.m = mean(x = x)  
  
# y variable  
y = data$Sabzevar  
y.m = mean(x = y)  
  
# length data  
n = nrow(x = data)  
  
# regression coefficient (b)  
b = (sum(x \* y) - ((sum(x) \* sum(y)) / n)) / (sum(x ^ 2) - (sum(x) ^ 2 / n))  
  
# total sum of squares (SST)  
SST = sum((y - y.m) ^ 2)  
  
# regression sum of squares (SSR)  
SSR <- sum((b \* (x - x.m)) ^ 2)  
  
# residual sum of squares (SSE)  
SSE <- sum((y - y.m - (b \* (x - x.m))) ^ 2)  
  
# mean squares  
SSE.m <- SSE / (n - 2)  
SSR.m <- SSR / 1  
  
# ratio of mean squares (F)  
FF = SSR.m / SSE.m  
  
# degrees of freedom  
df1 = 1  
df2 = n - 2  
# compute the critical values and p-value  
alpha = 0.05  
alternative = "greater"  
if (alternative != "two.sided") {  
 lower.critical = qf(p = (1 - alpha), df1 = df1, df2 = df2, lower.tail = FALSE)  
 upper.critical = qf(p = (1 - alpha), df1 = df1, df2 = df2, lower.tail = TRUE)  
} else {  
 lower.critical = qf(p = (1 - (alpha / 2)), df1 = df1, df2 = df2, lower.tail = FALSE)  
 upper.critical = qf(p = (1 - (alpha / 2)), df1 = df1, df2 = df2, lower.tail = TRUE)  
}  
p\_value = pf(q = FF, df1 = df1, df2 = df2, lower.tail = TRUE)  
p\_value = switch(EXPR = alternative,  
 two.sided = 2 \* min(p\_value, 1 - p\_value),  
 less = p\_value,  
 greater = 1 - p\_value)  
  
cat("Test Statistic:", FF, "\n")  
cat("Degrees of Freedom:", c(df1, df2), "\n")  
cat("Critical Values:", c(lower.critical, upper.critical), "\n")  
cat("P-value:", p\_value, "\n")

## Test Statistic: 45.8306  
## Degrees of Freedom: 1 58  
## Critical Values: 0.0040 4.0069  
## P-value: 7.13e-09

برای درک بهتر نتایج به دست آمده، مقدار آماره آزمون و توزیع با استفاده از تابع **plotDistStat** در یک نمودار رسم شده‏اند.

plotDistStat(distribution = "f", statistic = FF, df1 = df1, df2 = df2,  
 alpha = 0.05, alternative = "two.sided")



همانطور که مشخص است، در آزمون دو دنباله، مقادیر بحرانی در برابر با (00/4 – 00/0) است (جدول شماره 3). مقدار محاسبه شده برای بیشتر از مقادیر بحرانی بوده و در نتیجه فرضیه صفر رد خواهد شد. این بدان معنا است که رابطه خطی بین مجموع بارندگی سالانه ایستگاه مشهد و سبزوار برقرار است.

در روش دوم، برای بررسی خطی بودن رگرسیون بین متغیر و متغیر ، از تابع **lm** موجود در بسته **stats** استفاده شده است.

# method 2  
mod <- **lm**(formula = Sabzevar **~** Mashhad, data = data)  
  
**summary**(mod) ## Two Independent Samples Hotelling T Square Test   
## Hotelling T Sqaure Statistic = 2580.839   
## F value = 625.458, df1 = 4, df2 = 95, p-value: <2e-16

## Call:  
## lm(formula = Sabzevar ~ Mashhad, data = data)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -79.419 -30.501 -1.627 23.623 136.452   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 39.57 22.06 1.79 0.0781 .   
## Mashhad 0.58 0.09 6.77 7.13e-09 \*\*\*  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## Residual standard error: 45.21 on 58 degrees of freedom  
## Multiple R-squared: 0.4414, Adjusted R-squared: 0.4318   
## F-statistic: 45.83 on 1 and 58 DF, p-value: 7.131e-09

# assumption checking  
# 1. The error terms are normally and independently distributed with mean zero.  
shapiro.test(x = mod$residuals)

## Shapiro-Wilk normality test  
## data: mod$residuals  
## W = 0.96278, p-value = 0.06438

mean(mod$residuals)

## [1] 7.465093e-16

نتایج به دست آمده در روش دوم با نتایج روش اول برابر می‏باشند.

**R مقدماتی**

|  |
| --- |
| **استفاده از Help** |

**دسترسی به فایل‏های کمکی**

سه روش کلی برای استفاده از Help برنامه R وجود دارد:

* **نیاز به کمک در مورد عملکرد تابعی خاص:** به طور مثال نیاز است در مورد ساختار نوشتن تابع mean و آرگمان‏های ورودی آن اطلاعاتی به دست آید.

?mean



* **جستجو برای یک کلمه یا عبارتی خاص:** به طور مثال نیاز است در مورد چگونگی استفاده از میانگین وزنی در R اطلاعاتی به دست آید.

help.search(pattern = "weighted mean")



* **نیاز به راهنمایی در مورد یک بسته (Package):** به طور مثال نیاز است در مورد بسته dplyr اطلاعاتی کسب شود.

**help**(package = "dplyr")



**گرفتن اطلاعات بیشتر در مورد یک شیء‏ (Object)**

* **دریافت خلاصه‏ای از ساختار یک شیء:** به طور مثال نیاز است خلاصه‏ای از دیتاست **iris** ارائه شود.

data("iris")  
str(object = iris)

## 'data.frame': 150 obs. of 5 variables:  
## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...  
## $ Sepal.Width: num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  
## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...  
## $ Petal.Width: num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...  
## $ Species: Factor w/ 3 levels "setosa", "versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

* **پیدا کردن کلاس یک شیء:** به طور مثال نیاز است کلاس دیتاست **iris** مشخص شود.

class(x = iris)

## "data.frame"

|  |
| --- |
| **استفاده از بسته‏ها** |

بیش از ۵,۰۰۰ بسته آماده برای زبان **R** وجود دارد که می‏توان آن‏ها را دانلود و از آن‏ها استفاده کرد.

* **دانلود و نصب یک بسته از CRAN[[122]](#footnote-123):** در مثال زیر نحوه نصب بسته **dplyr** نشان داده شده است.

install.packages (pkgs = "dplyr")

* **فراخوانی یک بسته و استفاده از تمام توابع آن:** در مثال زیر بسته **dplyr** فراخوانی می‏شود.

library(package = dplyr)

* **استفاده از یک تابع خاص درون یک بسته، بدون فراخوانی بقیه توابع آن بسته:** در مثال زیر تابع **select** از بسته **dplyr** فراخوانی می‏شود.

dplyr::select(.data = iris, Species)

* **بارگزاری دیتاست‏های از پیش ساخته شده به درون محیط کاری R:** در مثال زیر دیتاست **iris** از بسته **datasets** به درون محیط کاری **R** فراخوانی می‏شود.

data(iris)

|  |
| --- |
| **دایرکتوری جاری[[123]](#footnote-124)** |

دایرکتوری جاری، فولدر پیش فرضی است که خواندن و نوشتن داده‏ها از داخل آن انجام می‏شود.

* **نمایش دایرکتوری جاری:** با اجرای دستور زیر، محل دایرکتوری جاری برنامه R نشان داده می‏شود.

**getwd()**

## "C:/Users/Pooya/OneDrive/Academic/Books/StatisticalTests"

* تغییر دایرکتوری جاری: با اجرای دستور زیر محل دایرکتوری جاری به آدرس “C:/Users/Pooya” تغییر پیدا می‏کند.

setwd(dir = "C:/Users/Pooya")

|  |
| --- |
| **بردارها[[124]](#footnote-125)** |

یک بردار آرایه‏ای یک بعدی از شیء‏ها با نوع یکسان است. یک بردار فقط دارای یک سطر یا ستون است. طول یک بردار برابر تعداد داده‏های موجود در آن می‏باشد. هر داده دارای یک ایندکس مخصوص است. با استفاده از ایندکس می‏توان به آن داده خاص در بردار ارجاع نمود. برای تعریف یک بردار می‏توان از تابع c() استفاده نمود.

**ایجاد یک بردار**

* **ایجاد یک بردار با استفاده از تابع c():** با اجرای دستور زیر یک بردار با مقادیر 1، 2، 10 و 20 در محیط کاری برنامه R ایجاد می‏شود.

**c**(1, 2, 10, 20)

## 1 2 10 20

* **ایجاد یک بردار با استفاده از تابع c() که هر عنصر دارای نام باشد:** در مثال زیر هدف ایجاد یک بردار با اطلاعات زیر می‏باشد.

Age = 30, ID = 12, Year = 2020

c("Age" = 30, "ID" = 12, "Year" = 2020)

## Age ID Year  
## 30 12 2020

* **ایجاد یک بردار از توالی اعداد با گام 1:** با استفاده از عملگر “:” و اجرای دستور زیر یک بردار با مقادیر 2 تا 8 با گام 1 ایجاد می‏شود.

2**:**8

## 2 3 4 5 6 7 8

* **ایجاد یک بردار از توالی اعداد با گام‏های مختلف با استفاده از تابع seq:** در مثال زیر هدف ایجاد برداری بین اعداد 2 تا 4 با گام 5/0 با استفاده از تابع seq می‏باشد.

seq(from = 2, to = 4, by = 0.5)

## 2.0 2.5 3.0 3.5 4.0

* **تکرار یک بردار پشت سر هم، به تعداد مورد نیاز با استفاده از تابع rep:** در مثال زیر بردار 1:2، سه بار تکرار می‏شود و برداری جدید با شش عضو ایجاد می‏شود.

rep(x = 1:2, 3)

## 1 2 1 2 1 2

* **تکرار عناصر یک بردار پشت سر هم، به تعداد مورد نیاز با استفاده از تابع rep:** در مثال زیر هر عنصر بردار 1:2، سه بار تکرار می‏شود و برداری جدید با شش عضو ایجاد می‏شود.

**rep**(x = 1**:**2, each = 3)

## 1 1 1 2 2 2

**توابع بردارها**

* **مرتب کردن یک بردار با استفاده از تابع sort:** با استفاده از تابع sort می‏توان عناصر بردار c(1, 5, 2, 0) را از کوچک به بزرگ مرتب کرد. برای اینکه عناصر از بزرگ به کوچک مرتب شوند باید مقدار آرگمان decreasing تابع sort، مساوی با TRUE قرار گیرد.

sort(x = c(1, 5, 2, 0))

## 0 1 2 5

sort(x = c(1, 5, 2, 0), decreasing = TRUE)

## 5 2 1 0

* **معکوس کردن یک بردار با استفاده از تابع rev:** در مثال زیر عناصر بردار c(2, 0, 5) با استفاده از تابع rev معکوس، و برداری جدید ایجاد می‏شود.

rev(x = c(1, 5, 2, 0))

## 0 2 5 1

* **استخراج عناصر منحصر به فرد یک بردار با استفاده از تابع unique:** در مثال زیر تابع unique مقادیر تکراری بردار c(1, 2, 1, 0, 5, 5, 7, 0) را حذف و یک بردار با مقادیر منحصر به فرد بر می‏گرداند.

unique(x = c(1, 2, 1, 0, 5, 5, 7, 0))

## 1 2 0 5 7

* **شمارش تعداد عناصر منحصر به فرد در یک بردار با استفاده از تابع table:** در مثال زیر تابع table تعداد هر عنصر منحصر به فرد در بردار c(1, 2, 1, 0, 5, 5, 7, 0) را به صورت یک جدول بر می‏گرداند.

table(x = c(1, 2, 1, 0, 5, 5, 7, 0))

## x  
## 0 1 2 5 7   
## 2 2 1 2 1

**انتخاب عناصر یک بردار**

می‏توان عناصر موجود در یک بردار را با استفاده از [ ] و شماره ایندکس عناصر انتخاب کرد.

* **انتخاب یک عنصر خاص:** در مثال زیر هدف انتخاب عنصر چهارم از بردار x می‏باشد.

x = c(1, 5, 0, 7, 1, 1, 2)  
x[4]

## 7

* **انتخاب همه به غیر از یک عنصر خاص:** در مثال زیر هدف انتخاب همه عناصر به غیر از عنصر چهارم از بردار x می‏باشد.

x = c(1, 5, 0, 7, 1, 1, 2)  
x[-4]

## 1 5 0 1 1 2

* **انتخاب چندین عنصر پشت سر هم:** در مثال زیر هدف انتخاب عناصر دوم تا چهارم از بردار x می‏باشد.

x = c(1, 5, 0, 7, 1, 1, 2)  
x[2:4]

## 5 0 7

* **انتخاب همه به غیر از چند عنصر خاص پشت سر هم:** در مثال زیر هدف انتخاب همه عناصر به غیر از عناصر دوم تا چهارم از بردار x می‏باشد.

x = c(1, 5, 0, 7, 1, 1, 2)  
x[-(2:4)]

## 1 1 1 2

* **انتخاب چندین عنصر همزمان:** در مثال زیر هدف انتخاب عناصر اول و پنجم از بردار x می‏باشد.

x = c(1, 5, 0, 7, 1, 1, 2)  
x[c (1, 5)]

## 1 1

می‏توان عناصر موجود در یک بردار را با استفاده از مقادیرشان نیز انتخاب کرد.

* **انتخاب عناصری از یک بردار که با یک مقدار خاص برابر هستند:** در مثال زیر هدف انتخاب عناصری از بردار x است که مقادیرشان برابر با 1 می‏باشد.

x = c(1, 5, 0, 7, 1, 1, 2)  
x[x == 1]

## 1 1 1

* **انتخاب عناصری از یک بردار با اعمال شرط روی مقادیر آن:** در مثال زیر هدف انتخاب عناصری از بردار x است که مقادیرشان بزرگتر از 1 می‏باشد.

x = c(1, 5, 0, 7, 1, 1, 2)  
x[x > 1]

## 5 7 2

* **انتخاب عناصری از یک بردار که با بردار دوم مشترک هستند:** در مثال زیر هدف انتخاب عناصری از بردار x است که مقادیرشان با مقادیر بردار c(1, 7) برابر می‏باشد.

x = c(1, 5, 0, 7, 1, 1, 2)  
x[x %in% c(1, 7)]

## 1 7 1 1

می‏توان عناصر موجود در یک بردار را با استفاده اسم آن‏ها نیز انتخاب کرد.

* **انتخاب یک عنصر از یک بردار با استفاده از اسم آن عنصر:** در مثال زیر هدف انتخاب عنصری از بردار x است که اسم آن برابر Year می‏باشد.

x = c("Age" = 30, "ID" = 12, "Year" = 2020)  
x["Year"]

## Year  
## 2020

* **انتخاب چند عنصر از یک بردار با استفاده از اسم آن عناصر:** در مثال زیر هدف انتخاب عناصری از بردار x است که اسم آن‏ها برابر با Age و Year می‏باشد.

x = c("Age" = 30, "ID" = 12, "Year" = 2020)  
x[c("Age", "Year")]

## Age Year  
## 30 2020

|  |
| --- |
| **برنامه‏نویسی** |

برتری زبان R نسبت به نرم‏افزارهای آماری مشابه خود این است که ویژگی‏های یک زبان برنامه‏نویسی در آن مستتر است و مانند سایر زبان‏های برنامه‏نویسی دارای ساختارهای کنترلی می‏باشد.

|  |
| --- |
| **دستورات شرطی** |

ساختارهای تصمیم‏گیری مستلزم این است که برنامه‏نویس یک یا چند شرط را مشخص کند که باید توسط برنامه برآورد یا تست شود، اگر این شرط درست و برقرار باشد، دستور یا دستوراتی باید اجرا شوند، و به صورت دلخواه می‏توان دستوراتی را نوشت که در صورت برقرار نبودن آن شروط اجرا شوند. در حالت کلی ساختار دستورات شرطی در زبان برنامه‏نویسی R به صورت زیر می‏باشد:

|  |
| --- |
| if (condition) {  Do Somethings  } else {  Do Somethings  } |

علاوه بر این اگر دستورات قسمت if و else به صورت تک دستوری باشند، می‏توان از تابع ifelse نیز استفاده نمود. ساختار تابع ifelse به صورت زیر می‏باشد:

|  |
| --- |
| ifelse(test = condition, yes = Do Something, no = Do Something) |

* **ساختارهای شرطی if:** در مثال زیر مقدار متغیر score برابر 15 است. در ساختار دستورات شرطی اگر مقدار score بزرگتر مساوی 10 باشد کلمه pass چاپ خواهد شد، در غیر اینصورت کلمه fail چاپ خواهد شد.

score = 15  
  
if (score >= 10) {  
 print(x = "Pass")  
} else {  
 print(x = "Fail")  
}

## "Pass"

score = 15  
  
ifelse(test = score >= 10, yes = print(x = "Pass"), no = print(x = "Fail"))

## "Pass"

|  |
| --- |
| **حلقه‏ها** |

حلقه‏های for و while در R برای تکرار شدن بخشی از کد استفاده می‏شود. تعداد تکرار با ایجاد شرط برای تابع while و یا تعیین مقدار مشخص یا مقدار متغیر برای تابع for، مشخص می‏شود.

**حلقه for**

حلقه for هنگامی مورد استفاده قرار می‏گیرد که تعداد تکرار مشخص باشد. در حالت کلی ساختار حلقه‏های for در زبان برنامه‏نویسی R به صورت زیر می‏باشد:

|  |
| --- |
| for (variable in sequence) {  Do Somethings  } |

که در آن variable، یک متغییر شمارنده، و sequence می‏تواند یک بردار باشد. دستورات درون بدنه حلقه به تعداد درایه‏های بردار sequence تکرار و اجرا می‏شود.

* **حلقه‏های for:** با اجرای دستورات زیر، دستور print چهار بار و به ازای هر i اجرا خواهد شد.

for (i in 1:4) {  
 print(i)  
}

## 1  
## 2  
## 3  
## 4

**حلقه while**

حلقه while تا زمانی ادامه می‏یابد که شرط برقرار باشد. در حالت کلی ساختار حلقه‏های while در زبان برنامه‏نویسی R به صورت زیر می‏باشد:

|  |
| --- |
| while (condition) {  Do Somethings  } |

که در آن condition، شرط ورود به حلقه است، و تا زمانی که این شرط برقرار باشد حلقه تکرار می‏شود.

* **حلقه‏های while:** با اجرای دستورات زیر، تا موقعی که مقدار i کوچکتر از 5 باشد، دستور print اجرا خواهد شد.

i = 1  
  
while (i < 5) {  
 print(i)  
 i = i + 1  
}

## 1  
## 2  
## 3  
## 4

|  |
| --- |
| **توابع** |

تابع یک بخشی از برنامه است که وظیفه خاصی را انجام می‏دهد. در حالت کلی ساختار توابع در زبان برنامه‏نویسی R به صورت زیر می‏باشد:

|  |
| --- |
| function\_name = function (variables) {  Do Somethings  return(new\_variable)  } |

که در آن function\_name، نامی است که برای تابع انتخاب می‏شود، variables آرگمان‏های ورودی تابع و new\_variable متغیری است که از تابع به عنوان نتیجه، خارج می‏شود.

* **توابع:** با اجرای دستورات زیر، تابعی با عنوان square، که آرگمان x را به عنوان ورودی گرفته و توان دوم آن را به عنوان خروجی تحویل می‏دهد، در محیط کاری برنامه R ایجاد می‏شود. با فراخوانی نام تابع و دادن ورودی به آن، خروجی مورد نظر نمایش داده می‏شود.

square <- **function**(x) {  
 z = x **\*** x  
 **return**(z)  
}  
  
**square**(x = 5)

## 25

|  |
| --- |
| **خواندن و نوشتن داده‏ها** |

**خواندن و نوشتن داده‏هایی که با یک کارکتر خاص از همدیگر جدا شده‏اند**

فرض کنید داده‏های میانگین دما و مجموع بارش سالانه برای سه سال متوالی در فایلی به نام data.txt ذخیره شده باشد. مقادیر هر سطر با فاصله یک تب از هم جدا شده‏اند.



* **تابع read.table:** با اجرای دستورات زیر، محتوای فایل data.txt خوانده شده و در متغیر df ذخیره می‏گردد.

df = read.table(file = "C:/Users/Pooya/Desktop/data.txt",  
 header = TRUE,  
 sep = "\t")  
  
print(df)

## Year Temp Prec  
## 1 2000 15.5 254.2  
## 2 2001 15.8 247.9  
## 3 2002 15.7 254.1

که آرگومان file برای مشخص کردن آدرس فایل data.txt، و آرگومان sep برای مشخص کردن اینکه داده‏ها با چه کارکتری از هم دیگر جدا شده‏اند، استفاده شده است.

* **تابع write.table:** با اجرای دستورات زیر، محتوای متغیر df در فایل newData.txt ذخیره می‏گردد. در فایل جدید داده‏ها با فاصله تب از همدیگر جدا می‏شوند.

write.table(x = df,  
 file = "C:/Users/Pooya/Desktop/newData.txt",  
 sep = "\t")

**خواندن و نوشتن داده‏هایی که با کاما از همدیگر جدا شده‏اند**

این یک حالت خاص از توابع read.table و write.table می‏باشد. فرض کنید داده‏های میانگین دما و مجموع بارش سالانه برای سه سال متوالی در فایلی به نام data.csv ذخیره شده باشد. مقادیر هر سطر با یک کاما از همدیگر جدا شده‏اند.



* **تابع read.csv:** با اجرای دستورات زیر، محتوای فایل data.csv خوانده شده و در متغیر df ذخیره می‏گردد.

df = read.csv(file = "C:/Users/Pooya/Desktop/data.csv",  
 header = TRUE)  
  
print(df)

## Year Temp Prec  
## 1 2000 15.5 254.2  
## 2 2001 15.8 247.9  
## 3 2002 15.7 254.1

* **تابع write.csv:** با اجرای دستورات زیر، محتوای متغیر df در فایل newData.csv ذخیره می‏گردد. در فایل جدید داده‏ها با یک کاما از همدیگر جدا می‏شوند.

write.csv (x = df,  
 file = "C:/Users/Pooya/Desktop/newData.csv")

|  |
| --- |
| **عملگرهای منطقی** |

عملگرهای منطقی برای شرطی کردن اجرای دستورات استفاده می‏شوند. خروجی این عملگرها در صورت برقرار بودن شرط TRUE و در صورتی که شرط برقرار نباشد FALSE می‏باشد.

|  |  |
| --- | --- |
| **عملگر** | **شرح** |
| > | بزرگتر |
| < | کوچکتر |
| >= | بزرگتر یا مساوی |
| <= | کوچکتر یا مساوی |
| == | مساوی |
| != | نامساوی |
| ! | نقیض |
| & | "و" منطقی |
| | | "یا" منطقی |

* **عملگرهای منطقی:** با اجرای دستورات زیر، با توجه به نوع عملگر، خروجی TRUE یا FALSE نمایش داده می‏شود.

5 > 4  
5 < 4  
5 >= 4  
5 <= 4  
5 == 4  
5 != 4  
!4  
(5 **<** 4) **&** (4 **<** 10)  
(5 **<** 4) **|** (4 **<** 10)

## TRUE  
## FALSE  
## TRUE  
## FALSE  
## FALSE  
## TRUE  
## FALSE  
## FALSE  
## TRUE

|  |
| --- |
| **نوع داده‏ها در R** |

در حالت کلی می‏توان داده‏ها را در برنامه R به پنج زیر مجموعه زیر تقسیم‏بندی کرد:

|  |  |
| --- | --- |
| نوع داده | مثال |
| منطقی (logical) | x = TRUE print(class(x = x))  ## "logical" |
| عددی (numeric) | x = 12.5 print(class(x = x))  ## "numeric" |
| صحیح (integer) | x = 2L print(class(x = x))  ## "integer" |
| مختلط (complex) | x = 3 + 2i print(class(x = x))  ## "complex" |
| متن (character) | x = "Hello" print(class(x = x))  ## "character" |

|  |
| --- |
| **عملگر تخصیص** |

برای ذخیره یک ساختار داده باید به آن یک نام اختصاص داد. بدین منظور از عملگرهای تخصیص “=” و “<-” استفاده می‏شود.

* **عملگرتخصیص:** با اجرای دستورات زیر، مقدار 5 در متغیر x و مقدار TRUE در متغیر y ذخیره می‏گردد.

x = 5  
y <- TRUE  
  
print(x)  
print(x)

## 5  
## TRUE

|  |
| --- |
| **مدیریت متغیرها** |

* **مشاهده همه متغیرهای محیط کاری R با استفاده از تابع ls:** با اجرای دستورات زیر، متغیرهای موجود در محیط کاری نمایش داده می‏شوند.

x = 2  
y = TRUE  
z = "Hello"  
  
ls()

## "x" "y" "z"

* **حذف یک متغیر از محیط کاری R با استفاده از تابع rm:** با اجرای دستورات زیر، متغیر x حذف خواهد شد.

x = 2  
y = TRUE  
z = "Hello"  
  
ls()  
**rm**(x)  
**ls**()

## "x" "y" "z"  
## "y" "z"

* **حذف همه متغیرها از محیط کاری R با استفاده از تابع rm:** با اجرای دستورات زیر، همه متغیرها حذف خواهند شد.

x = 2  
y = TRUE  
z = "Hello"  
  
**rm**(list = **ls**())  
**ls**()

## character(0)

|  |
| --- |
| **توابع ریاضی در R** |

* **لگاریتم طبیعی با استفاده از تابع log:** با اجرای دستورات زیر، لگاریتم طبیعی همه عناصر موجود در بردار x محاسبه می‏شود.Inf، به معنی بینهایت و NaN، به معنی وجود نداشتن مقداری است.

x = **c**(1, 25, 0, 4, 1, 100, -1, 44, 100, -1)  
**log**(x = x)

## 0.0 3.2 -Inf 1.4 0.0 4.6 NaN 3.8 4.6 NaN

* **لگاریتم بر پایه 10 با استفاده از تابع log10:** با اجرای دستورات زیر، لگاریتم بر پایه 10 همه عناصر موجود در بردار x محاسبه می‏شود.Inf، به معنی بینهایت و NaN، به معنی وجود نداشتن مقداری است.

x = **c**(1, 25, 0, 4, 1, 100, -1, 44, 100, -1)  
**log10**(x = x)

## 0.0 1.4 -Inf 0.6 0.0 2.0 NaN 1.6 2.0 NaN

* **تابع نمایی () با استفاده از تابع exp:** با اجرای دستورات زیر، تابع نمایی عناصر موجود در بردار x محاسبه می‏شود.

x = **c**(1, 25, 0, 4, 1, 100, -1, 44, 100, -1)  
**exp**(x = x)

## 2.7e+00 7.2e+10 1.0e+00 5.5e+01 2.7e+00 2.7e+43 3.7e-01 1.3e+19 2.7e+43 3.7e-01

* **پیدا کردن مقدار بیشینه عناصر یک بردار با استفاده از تابع max:** با اجرای دستورات زیر، بزرگترین عنصر موجود در بردار x نمایش داده می‏شود.

x = **c**(1, 25, 0, 4, 1, 100, -1, 44, 100, -1)  
**max**(x)

## 100

* **پیدا کردن مقدار کمینه عناصر یک بردار با استفاده از تابع min:** با اجرای دستورات زیر، کوچکترین عنصر موجود در بردار x نمایش داده می‏شود.

x = **c**(1, 25, 0, 4, 1, 100, -1, 44, 100, -1)  
**min**(x)

## -1

* **پیدا کردن مجموع همه عناصر یک بردار با استفاده از تابع sum:** با اجرای دستورات زیر، مجموع همه عناصر موجود در بردار x محاسبه و نمایش داده می‏شود.

x = **c**(1, 25, 0, 4, 1, 100, -1, 44, 100, -1)  
**sum**(x)

## 273

* **پیدا کردن میانگین همه عناصر یک بردار با استفاده از تابع mean:** با اجرای دستورات زیر، میانگین همه عناصر موجود در بردار x محاسبه و نمایش داده می‏شود.

x = **c**(1, 25, 0, 4, 1, 100, -1, 44, 100, -1)  
**mean**(x)

## 27

* **پیدا کردن میانه عناصر یک بردار با استفاده از تابع median:** با اجرای دستورات زیر، میانه عناصر موجود در بردار x محاسبه و نمایش داده می‏شود.

x = **c**(1, 25, 0, 4, 1, 100, -1, 44, 100, -1)  
**median**(x = x)

## 2.5

* **پیدا کردن واریانس همه عناصر یک بردار با استفاده از تابع var:** با اجرای دستورات زیر، واریانس همه عناصر موجود در بردار x محاسبه و نمایش داده می‏شود.

x = **c**(1, 25, 0, 4, 1, 100, -1, 44, 100, -1)  
**var**(x = x)

## 1681

* **پیدا کردن انحراف معیار همه عناصر یک بردار با استفاده از تابع sd:** با اجرای دستورات زیر، انحراف معیار همه عناصر موجود در بردار x محاسبه و نمایش داده می‏شود.

x = **c**(1, 25, 0, 4, 1, 100, -1, 44, 100, -1)  
**sd**(x = x)

## 41

* **محاسبه ضریب همبستگی بین دو بردار با استفاده از تابع cor:** با اجرای دستورات زیر، ضریب همبستگی بین دو بردار x و y محاسبه و نمایش داده می‏شود.

x = **c**(2, 8, 4, 5, 1)  
y = **c**(1, 7, 4, 5, 3)  
**cor**(x = x, y = y)

## 0.9

* **پیدا کردن رتبه عناصر یک بردار با استفاده از تابع rank:** با اجرای دستورات زیر، رتبه هر یک از عناصر موجود در بردار x محاسبه و نمایش داده می‏شود.

x = **c**(1, 25, 0, 4, 1, 100, -1, 44, 100, -1)  
**rank**(x = x)

## 4.5 7.0 3.0 6.0 4.5 9.5 1.5 8.0 9.5 1.5

* **گرد کردن اعداد با استفاده از تابع round:** با اجرای دستورات زیر، همه عناصر بردار x تا دو رقم اعشار گرد می‏شوند.

x = **c**(1.254, 100.24, 5.8975)  
**round**(x = x, digits = 2)

## 1.25 100.24 5.90

|  |
| --- |
| **ماتریس‏ها** |

ماتریس‏ها یکی از شی‏ء‏های زبان برنامه‏نویسی R می‏باشند که در آن، عناصر در یک طرح مستطیلی دو بعدی مرتب می‏شوند. عناصر ماتریس‏ها می‏توانند شامل نوع داده‏های مختلف مانند منطقی، عددی، صحیح، مختلط و متنی باشند. نکته مهم در اینجا این است که همه عناصر موجود در یک ماتریس باید از یک نوع داده باشند.

* **ایجاد یک ماتریس با استفاده از تابع matrix:** با اجرای دستورات زیر، ماتریسی با عناصر بردار c(1, 2, 3, 4, 5, 6) تشکیل می‏شود. این ماتریس دارای 2 سطر و 3 ستون بوده و در صورتی که مقدار آرگمان byrow برابر با TRUE باشد مقادیر بردار، به صورت سطر به سطر ماتریس را تشکیل می‏دهد. در صورتی که مقدار آرگمان byrow برابر با FALSE باشد ماتریس به صورت ستون به ستون با استفاده از مقادیر بردار مورد نظر تشکیل خواهد شد.

x = matrix(data = c(1, 2, 3, 4, 5, 6),  
 nrow = 2,  
 ncol = 3,  
 byrow = TRUE)  
  
print(x)

## [,1] [,2] [,3]  
## [1,] 1 2 3  
## [2,] 4 5 6

x = matrix(data = c(1, 2, 3, 4, 5, 6),  
 nrow = 2,  
 ncol = 3,  
 byrow = FALSE)  
  
print(x)

## [,1] [,2] [,3]  
## [1,] 1 3 5  
## [2,] 2 4 6

* **انتخاب یک سطر از ماتریس:** با اجرای دستورات زیر، تمام عناصر موجود در سطر دوم ماتریس x انتخاب و نمایش داده می‏شود.

x = matrix(data = 1:12,  
 nrow = 4,  
 ncol = 3,  
 byrow = TRUE)  
  
x[2, ]

## 4 5 6

* **انتخاب چند سطر از ماتریس:** با اجرای دستورات زیر، تمام عناصر موجود در سطرهای دوم و چهارم ماتریس x انتخاب و نمایش داده می‏شود.

x = matrix(data = 1:12,  
 nrow = 4,  
 ncol = 3,  
 byrow = TRUE)  
  
x[**c**(2, 4), ]

## [,1] [,2] [,3]  
## [1,] 4 5 6  
## [2,] 10 11 12

* **انتخاب یک ستون از ماتریس:** با اجرای دستورات زیر، تمام عناصر موجود در ستون دوم ماتریس x انتخاب و نمایش داده می‏شود.

x = matrix(data = 1:12,  
 nrow = 4,  
 ncol = 3,  
 byrow = TRUE)  
  
x[ , 2]

## 2 5 8 11

* **انتخاب چند ستون از ماتریس:** با اجرای دستورات زیر، تمام عناصر موجود در ستون‏های دوم و سوم ماتریس x انتخاب و نمایش داده می‏شود.

x = matrix(data = 1:12,  
 nrow = 4,  
 ncol = 3,  
 byrow = TRUE)  
  
x[ , 2:3]

## [,1] [,2]  
## [1,] 2 3  
## [2,] 5 6  
## [3,] 8 9  
## [4,] 11 12

* **انتخاب یک عنصر خاص از ماتریس:** با اجرای دستورات زیر، عنصر موجود در سطر سوم و ستون دو ماتریس x انتخاب و نمایش داده می‏شود.

x = matrix(data = 1:12,  
 nrow = 4,  
 ncol = 3,  
 byrow = TRUE)  
  
x[3, 2]

## 8

|  |
| --- |
| **لیست‏ها** |

در زبان برنامه‏نویسی R، لیست‏ها مجموعه‏ای منظم از عناصر می‏باشند که ضرورتی ندارد که مولفه‏های آن از یک نوع داده و ابعاد آن با هم مساوی باشد. به عنوان مثال یک لیست می‏تواند شامل بردارهای منطقی، عددی، صحیح، اعداد مختلط و متنی باشد.

* **ایجاد یک لیست با استفاده از تابع list:** با اجرای دستورات زیر، یک لیست با عنوان l تشکیل می‏شود. این لیست شامل دو مقدار است. x، که برداری از نوع عددی و مقادیر 1 تا 5 و y، که یک بردار متنی شامل کارکترهای a و b.

l = list(x = 1:5,  
 y = c("a", "b"))  
  
print(l)

## $x  
## 1 2 3 4 5  
## $y  
## "a" "b"

* **انتخاب یک عنصر خاص از یک لیست با شماره ایندکس آن (خروجی یک لیست باشد):** با اجرای دستورات زیر، عنصر‏ دوم (y) انتخاب و نمایش داده می‏شود. خروجی در این حالت یک لیست می‏باشد.

l = list(x = 1:5,  
 y = c("a", "b"))  
  
l[2]

## $y  
## "a" "b"

* **انتخاب یک عنصر خاص از یک لیست با شماره ایندکس آن (خروجی بستگی به ماهیت آن عنصر دارد):** با اجرای دستورات زیر، عنصر‏ دوم (y) انتخاب و نمایش داده می‏شود. خروجی در این حالت یک بردار از کارکترها می‏باشد.

l = list(x = 1:5,  
 y = c("a", "b"))  
  
l[[2]]

## "a" "b"

* **انتخاب یک عنصر خاص از یک لیست با استفاده از نام آن (خروجی یک لیست باشد):** با اجرای دستورات زیر، عنصر‏ اول (x) انتخاب و نمایش داده می‏شود. خروجی در این حالت یک لیست می‏باشد.

l = list(x = 1:5,  
 y = c("a", "b"))  
  
l["x"]

## $x  
## 1 2 3 4 5

* **انتخاب یک عنصر خاص از یک لیست با استفاده از نام آن (خروجی بستگی به ماهیت آن عنصر دارد):** با اجرای دستورات زیر، عنصر‏ اول (x) انتخاب و نمایش داده می‏شود. خروجی در این حالت یک بردار از نوع عددی می‏باشد.

l = list(x = 1:5,  
 y = c("a", "b"))  
  
l$x

## 1 2 3 4 5

|  |
| --- |
| **دیتا فریم‏ها** |

دیتا فریم‏ها از نظر ساختاری زیر مجموعه لیست‏ها حساب می‏شوند ولی شباهت بسیار زیادی به ماتریس‏ها دارند. یک دیتا فریم را می‏توان ماتریسی در نظر گرفت که ستون‏های آن بردارهایی با ساختارهای مختلف داده است. از سوی دیگر می‏توان گفت که دیتا فریم لیستی از بردارهای هم اندازه و با ساختارهای مختلف است.

* **ایجاد یک دیتا فریم با استفاده از تابع data.frame:** با اجرای دستورات زیر، یک دیتا فریم با عنوان df تشکیل می‏شود. این دیتا فریم شامل دو عنصر یا ستون با سه عضو یا سطر می‏باشد. x، که برداری از نوع عددی و مقادیر 1 تا 3 و y، که یک بردار متنی شامل کارکترهای a، b و c می‏باشد.

df = data.frame(x = 1:3,  
 y = c("a", "b", "c"))  
  
print(df)

## x y  
## 1 1 a  
## 2 2 b  
## 3 3 c

* **انتخاب یک ستون خاص از یک دیتا فریم** **با شماره ایندکس آن (خروجی بستگی به ماهیت آن ستون دارد):** با اجرای دستورات زیر، ستون‏ دوم دیتا فریم (y) انتخاب و نمایش داده می‏شود. خروجی در این حالت یک بردار متنی شامل کارکترهای a، b و c می‏باشد.

df = data.frame(x = 1:3,  
 y = c("a", "b", "c"))  
  
df[, 2]

## "a" "b" "c"

* **انتخاب یک ستون خاص از یک دیتا فریم با استفاده از نام آن ستون (خروجی بستگی به ماهیت آن ستون دارد):** با اجرای دستورات زیر، ستون‏ دوم دیتا فریم (y) انتخاب و نمایش داده می‏شود. خروجی در این حالت یک بردار متنی شامل کارکترهای a، b و c می‏باشد.

df = data.frame(x = 1:3,  
 y = c("a", "b", "c"))  
  
df$y

## "a" "b" "c"

* **انتخاب یک ستون خاص از یک دیتا فریم با استفاده از نام آن ستون (خروجی یک دیتا فرم است):** با اجرای دستورات زیر، ستون‏ دوم دیتا فریم (y) انتخاب و نمایش داده می‏شود. خروجی در این حالت یک دیتا فریم با سه سطر و یک ستون می‏باشد.

df = data.frame(x = 1:3,  
 y = c("a", "b", "c"))  
  
df["y"]

## y  
## 1 a  
## 2 b  
## 3 c

* **انتخاب یک سطر خاص از یک دیتا فریم با استفاده شماره ایندکس آن سطر (خروجی یک دیتا فرم است):** با اجرای دستورات زیر، سطر‏ دوم دیتا فریم انتخاب و نمایش داده می‏شود. خروجی در این حالت یک دیتا فریم با یک سطر و دو ستون می‏باشد.

df = data.frame(x = 1:3,  
 y = c("a", "b", "c"))  
  
df[2, ]

## x y  
## 1 2 b

* **انتخاب یک عنصر خاص از یک دیتا فریم با استفاده شماره ایندکس سطر و ستون آن (خروجی بستگی به ماهیت آن عنصر دارد):** با اجرای دستورات زیر، سطر‏ دوم از ستون دوم دیتا فریم انتخاب و نمایش داده می‏شود. خروجی در این حالت یک کارکتر می‏باشد.

df = data.frame(x = 1:3,  
 y = c("a", "b", "c"))  
  
df[2, 2]

## b

* **مشاهده ابتدا و انتهای یک دیتا فریم با استفاده از توابع head و tail:** با اجرای دستورات زیر، سه سطر اول و دو سطر آخر دیتا فرم df نمایش داده می‏شود.

df = data.frame(x = 1:5,  
 y = LETTERS[1:5],  
 z = c(TRUE, TRUE, FALSE, TRUE, FALSE))  
  
head(x = df, n = 3)

## x y z  
## 1 1 A TRUE  
## 2 2 B TRUE  
## 3 3 C FALSE

df = data.frame(x = 1:5,  
 y = LETTERS[1:5],  
 z = c(TRUE, TRUE, FALSE, TRUE, FALSE))  
  
tail(x = df, n = 2)

## x y z  
## 4 4 D TRUE  
## 5 5 E FALSE

* **مشاهده تعداد سطرها، تعداد ستون‏ها و ابعاد یک دیتا فریم به ترتیب با استفاده از توابع nrow، ncol و dim:** با اجرای دستورات زیر، تعداد سطرها، تعداد ستون‏ها و ابعاد دیتا فریم df نمایش داده می‏شود.

df = data.frame(x = 1:5,  
 y = LETTERS[1:5],  
 z = c(TRUE, TRUE, FALSE, TRUE, FALSE))  
  
nrow(x = df)  
**ncol**(x = df)  
**dim**(x = df)

## 5  
## 3  
## 5 3

* **اضافه کردن ستون به یک دیتا فریم با استفاده از تابع cbind:** با اجرای دستورات زیر، بردار w به عنوان ستون جدید به دیتا فریم df اضافه می‏گردد. تعداد درایه‏های بردار w باید با تعداد سطرهای دیتا فریم df برابر باشد.

df = data.frame(x = 1:5,  
 y = LETTERS[1:5],  
 z = c(TRUE, TRUE, FALSE, TRUE, FALSE))  
  
print(x = df)

## x y z  
## 1 1 A TRUE  
## 2 2 B TRUE  
## 3 3 C FALSE  
## 4 4 D TRUE  
## 5 5 E FALSE

w = **c**(10L, 20L, 30L, 40L, 50L)  
  
df = **cbind**(df, w)  
  
**print**(df)

## x y z w  
## 1 1 A TRUE 10  
## 2 2 B TRUE 20  
## 3 3 C FALSE 30  
## 4 4 D TRUE 40  
## 5 5 E FALSE 50

* **اضافه کردن سطر به یک دیتا فریم با استفاده از تابع rbind:** با اجرای دستورات زیر، دیتا فریم w به عنوان سطر جدید به انتهای دیتا فریم df اضافه می‏گردد. تعداد ستون‏های دیتا فریم w و نوع داده هر ستون باید با تعداد ستون‏ها و نوع داده‏های دیتا فریم df برابر باشد.

df = data.frame(x = 1:5,  
 y = LETTERS[1:5],  
 z = c(TRUE, TRUE, FALSE, TRUE, FALSE))  
  
w = data.frame(x = 6, y = "F", z = FALSE)  
  
df = rbind(df, w)  
  
print(df)

## x y z  
## 1 1 A TRUE  
## 2 2 B TRUE  
## 3 3 C FALSE  
## 4 4 D TRUE  
## 5 5 E FALSE  
## 6 6 F FALSE

|  |
| --- |
| **توزیع‏های احتمالی** |

در زبان برنامه نویسی R بیشتر توابع احتمالی وجود دارد. هر تابع دارای چهار شکل متفاوت است:

* دستور dfunc(x, …): عرض تابع را در نقطه x نشان می‏دهد (Density Function).
* دستور pfunc(x, …): مقدار احتمال تجمعی را تا نقطه x نشان می‏دهد (Cumulative Dist.).
* دستور qfunc(p, …): مقدار چندک تابع را به ازای 0 < p < 1 نشان می‏دهد (Quantile).
* دستور rfunc(x, …): نمونه تصادفی از تابع را شبیه‏سازی می‏کند (Random Variates).

در جدول زیر، funcهای مختلف موجود در زبان برنامه‏نویسی R نشان داده شده است:

|  |  |  |
| --- | --- | --- |
| **آرگمان‏های ورودی** | **func** | **توزیع** |
| shape1, shape2 | beta | Beta |
| size, prob | binom | Binomial |
| df, ncp | chisq | Chi Square |
| df1, df2, ncp | f | F |
| shapem, rate, scale | gamma | Gamma |
| location, scale | logis | Logistic |
| meanlog, sdlog | lnorm | Log Normal |
| mean, sd | norm | Normal |
| lambda | pois | Poisson |
| df, ncp | t | t Student |
| shape, scale | Weibull | Weibull |
| min, max | unif | Uniform |

|  |
| --- |
| **نمودارها** |

در زبان برنامه‏نویسی R می‏توان نمودارها متفاوت و مختلفی را رسم نمود.

* **رسم نمودار بردار x با استفاده از تابع plot:** با اجرای دستورات زیر، نمودار پراکندگی بردار x نسبت به شماره ایندکس آن‏ها رسم می‏گردد.

x = rnorm(n = 6, mean = 10, sd = 2)  
  
plot(x = x)



* **رسم نمودار پراکندگی بردار x در مقابل بردار y با استفاده از تابع plot:** با اجرای دستورات زیر، نمودار پراکندگی بردار x در مقابل بردار y رسم می‏گردد.

x = rnorm(n = 6, mean = 10, sd = 2)  
y = rnorm(n = 6, mean = 10, sd = 2)  
  
plot(x = x, y = y)



* **رسم منحنی یک تابع در محدوده تعیین شده با استفاده از تابع curve:** با اجرای دستورات زیر، منحنی توزیع آماری نرمال در فاصله 3- تا 3 رسم می‏گردد..

curve(expr = dnorm(x = x, mean = 0, sd = 1), from = -3, to = 3)



* **رسم هیستوگرام یک بردار با استفاده از تابع hist:** با اجرای دستورات زیر، هیستوگرام بردار x رسم خواهد شد.

x = rnorm(n = 100, mean = 0, sd = 1)  
hist(x = x)



* **رسم نمودار میله‏ای یک بردار با استفاده از تابع barplot:** با اجرای دستورات زیر، نمودار میله‏ای بردار x رسم خواهد شد.

x = rnorm(n = 10, mean = 0, sd = 1)  
barplot(height = x)



* **رسم نمودار جعبه‏ای یک بردار با استفاده از تابع barplot:** با اجرای دستورات زیر، نمودار جعبه‏ای بردار x رسم خواهد شد.

x = rnorm(n = 100, mean = 0, sd = 1)  
boxplot(x = x)



* **رسم نمودار برای بررسی مقایسه توزیع‏ها با استفاده از توابع qqnorm و qqline:** با اجرای دستورات زیر، می‏توان نرمال بودن توزیع داده‏های بردار x را با تابع qqnorm و رسم خط نرمال روی نمودار را با تابع qqline بررسی کرد.

x = rnorm(n = 100, mean = 0, sd = 1)  
qqnorm(y = x)  
qqline(y = x)



* **رسم نمودار برای مقایسه توزیع‏ دو متغیر با استفاده از تابع qqplot:** با اجرای دستورات زیر، می‏توان توزیع مقادیر دو بردار x و y را با هم مقایسه و هم توزیع بودن آن‏ها را بررسی کرد.

x = rnorm(n = 100, mean = 0, sd = 1)  
y = rf(n = 100, df1 = 1, df2 = 20)  
  
qqplot(x = x, y = y)



1. Z-test [↑](#footnote-ref-2)
2. Population Variance [↑](#footnote-ref-3)
3. One-tailed Test [↑](#footnote-ref-4)
4. Two-tailed Test [↑](#footnote-ref-5)
5. Critical Region [↑](#footnote-ref-6)
6. Standard Normal Distribution [↑](#footnote-ref-7)
7. Q-Q Normal Plot [↑](#footnote-ref-8)
8. Histogram Plot [↑](#footnote-ref-9)
9. Box Plot [↑](#footnote-ref-10)
10. Z-test [↑](#footnote-ref-11)
11. Z-test [↑](#footnote-ref-12)
12. Proportion [↑](#footnote-ref-13)
13. Binomial Distribution [↑](#footnote-ref-14)
14. Proportion [↑](#footnote-ref-15)
15. Binomial Distribution [↑](#footnote-ref-16)
16. Counts [↑](#footnote-ref-17)
17. Poisson Distribution [↑](#footnote-ref-18)
18. t-test [↑](#footnote-ref-19)
19. Population Variance [↑](#footnote-ref-20)
20. Student's t-distribution [↑](#footnote-ref-21)
21. Degrees of Freedom [↑](#footnote-ref-22)
22. One-tailed Test [↑](#footnote-ref-23)
23. Two-tailed Test [↑](#footnote-ref-24)
24. Normal Distribution [↑](#footnote-ref-25)
25. P-value [↑](#footnote-ref-26)
26. Q-Q Plot [↑](#footnote-ref-27)
27. t-test [↑](#footnote-ref-28)
28. Population Variance [↑](#footnote-ref-29)
29. Sum of Squares [↑](#footnote-ref-30)
30. Student's t-distribution [↑](#footnote-ref-31)
31. Degrees of Freedom [↑](#footnote-ref-32)
32. One-tailed Test [↑](#footnote-ref-33)
33. Two-tailed Test [↑](#footnote-ref-34)
34. t-test [↑](#footnote-ref-35)
35. Population Variance [↑](#footnote-ref-36)
36. Student's t-distribution [↑](#footnote-ref-37)
37. t-test [↑](#footnote-ref-38)
38. Paired Comparisons [↑](#footnote-ref-39)
39. Student's t-distribution [↑](#footnote-ref-40)
40. Degrees of Freedom [↑](#footnote-ref-41)
41. One-tailed Test [↑](#footnote-ref-42)
42. Two-tailed Test [↑](#footnote-ref-43)
43. t-test [↑](#footnote-ref-44)
44. Regression Coefficient [↑](#footnote-ref-45)
45. Sample Estimate [↑](#footnote-ref-46)
46. Regression Line [↑](#footnote-ref-47)
47. Student's t-distribution [↑](#footnote-ref-48)
48. Degrees of Freedom [↑](#footnote-ref-49)
49. Two-tailed Test [↑](#footnote-ref-50)
50. Directional [↑](#footnote-ref-51)
51. Normal Distribution [↑](#footnote-ref-52)
52. Shapiro-Wilk Test [↑](#footnote-ref-53)
53. P-value [↑](#footnote-ref-54)
54. Q-Q Plot [↑](#footnote-ref-55)
55. t-test [↑](#footnote-ref-56)
56. Correlation Coefficient [↑](#footnote-ref-57)
57. Bivariate Normal Distribution [↑](#footnote-ref-58)
58. Student's t-distribution [↑](#footnote-ref-59)
59. Degrees of Freedom [↑](#footnote-ref-60)
60. Two-tailed Test [↑](#footnote-ref-61)
61. Shapiro-Wilk Test [↑](#footnote-ref-62)
62. Z-test [↑](#footnote-ref-63)
63. Correlation Coefficient [↑](#footnote-ref-64)
64. Population Correlation Coefficient [↑](#footnote-ref-65)
65. Kendall Rank Correlation Test [↑](#footnote-ref-66)
66. Sample Correlation Coefficient [↑](#footnote-ref-67)
67. Fisher's Z-transformation [↑](#footnote-ref-68)
68. Shapiro-Wilk Test [↑](#footnote-ref-69)
69. Scatter Diagram [↑](#footnote-ref-70)
70. Z-test [↑](#footnote-ref-71)
71. Correlation Coefficient [↑](#footnote-ref-72)
72. Sample Correlation Coefficient [↑](#footnote-ref-73)
73. Shapiro-Wilk Test [↑](#footnote-ref-74)
74. Scatter Diagram [↑](#footnote-ref-75)
75. Chi-square Test [↑](#footnote-ref-76)
76. Chi-square Distribution [↑](#footnote-ref-77)
77. Degrees of Freedom [↑](#footnote-ref-78)
78. One-tailed Test [↑](#footnote-ref-79)
79. Two-tailed Test [↑](#footnote-ref-80)
80. F-test [↑](#footnote-ref-81)
81. Variance Ratio Test [↑](#footnote-ref-82)
82. Fisher Distribution [↑](#footnote-ref-83)
83. Degrees of Freedom [↑](#footnote-ref-84)
84. One-tailed Test [↑](#footnote-ref-85)
85. Two-tailed Test [↑](#footnote-ref-86)
86. Iris Flower Data Set [↑](#footnote-ref-87)
87. Petal [↑](#footnote-ref-88)
88. Sepal [↑](#footnote-ref-89)
89. F-test [↑](#footnote-ref-90)
90. Correlated Observations [↑](#footnote-ref-91)
91. Degrees of Freedom [↑](#footnote-ref-92)
92. Hotelling’s T2-test [↑](#footnote-ref-93)
93. Correlated [↑](#footnote-ref-94)
94. Fisher Distribution [↑](#footnote-ref-95)
95. Degrees of Freedom [↑](#footnote-ref-96)
96. Iris Flower Data Set [↑](#footnote-ref-97)
97. Petal [↑](#footnote-ref-98)
98. Sepal [↑](#footnote-ref-99)
99. Discriminant Test [↑](#footnote-ref-100)
100. Dixon’s Test [↑](#footnote-ref-101)
101. Outliers [↑](#footnote-ref-102)
102. Extreme Value [↑](#footnote-ref-103)
103. Correlated Proportions [↑](#footnote-ref-104)
104. Chi-square Distribution [↑](#footnote-ref-105)
105. Degrees of Freedom [↑](#footnote-ref-106)
106. Poisson Distribution [↑](#footnote-ref-107)
107. Mosaic Plots [↑](#footnote-ref-108)
108. Normal Frequency Distribution [↑](#footnote-ref-109)
109. Continuous Distribution [↑](#footnote-ref-110)
110. Cochran's Test [↑](#footnote-ref-111)
111. Variance Outliers [↑](#footnote-ref-112)
112. Wilcoxon Inversion Test [↑](#footnote-ref-113)
113. Wilcoxon–Mann–Whitne Rank Sum Test [↑](#footnote-ref-114)
114. Spearman Rank Correlation Test [↑](#footnote-ref-115)
115. One-tailed Test [↑](#footnote-ref-116)
116. Two-tailed Test [↑](#footnote-ref-117)
117. Kendall Rank Correlation Test [↑](#footnote-ref-118)
118. One-tailed Test [↑](#footnote-ref-119)
119. Two-tailed Test [↑](#footnote-ref-120)
120. Fisher Distribution [↑](#footnote-ref-121)
121. Degrees of Freedom [↑](#footnote-ref-122)
122. Comprehensive R Archive Network [↑](#footnote-ref-123)
123. Working Directory [↑](#footnote-ref-124)
124. Vectors [↑](#footnote-ref-125)