به نام خدا

آمار و احتمال مهندسی
دانشگاه صنعتی شریف - دانشکده مهندسی کامپیوتر

پاییز 1401

تمرین عملی ششم

Amirreza Azari

99101087

سوال اول

فرض کنید شما در کاشان در یک کارخانه گلاب گیری مشغول به کار هستید. دوست شما ویلیام گاست، مایل ها آنورتر در دوبلین مشغول کاری مشابه شماست منتها هنوز مقاله معروف خود یعنی student-t را منتشر نکرده است و شما اطلاعی از توزیع t ندارید.

فرض کنید مقدار گلابی که از یک گل محمدی استخراج میشود از توزیع نرمال با میانگین ۱۵۰ و انحراف معیار ۱۰ پیروی میکند. و گل های باغ شما تعداد زیادی گل محمدی دارد که هیچگاه تمام نمیشود.

شما که در آمار و احتمال دستی بر آتش دارید، با قضیه حد مرکزی آشنا هستید و می‌دانید که در صورتی که n عدد بزرگی باشد و $\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$ داریم:

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Z$$

که در آن X_iها مستقل و از توزیع یکسان با میانگین μ و واریانس σ^2 هستند.

شما کنجکاو میشوید که اگر نمونه ای که از جامعه داریم محدود باشد (به هر حال همه که مانند شما چنین باغ بزرگی ندارند!) و واریانس جامعه را نداشته باشیم و به جای آن از واریانس نمونه استفاده کنیم. چه اتفاقی خواهد افتاد؟
به عبارتی اگر x_iها نمونه کوچک ما باشد و داشته باشیم:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n-1}$$

در آن صورت

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

از چه توزیعی تبعیت خواهد کرد؟
در نتیجه خودتان دست به کار میشوید تا با استفاده از شرایطی که برایتان مهیاست جواب سوالتان را پیدا کنید.

ابتدا در سلول زیر تابعی بنویسید که با ورودی گرفتن نمونه‌ها و میانگین جامعه مقدار t را محاسبه کند:

In [2]:
```
find_t <- function(df, v){
  return((mean(df) - v) / (( sqrt(sum((df - mean(df))^2) / (length(df) - 1)) ) / sqrt(length(df))) )
}
```

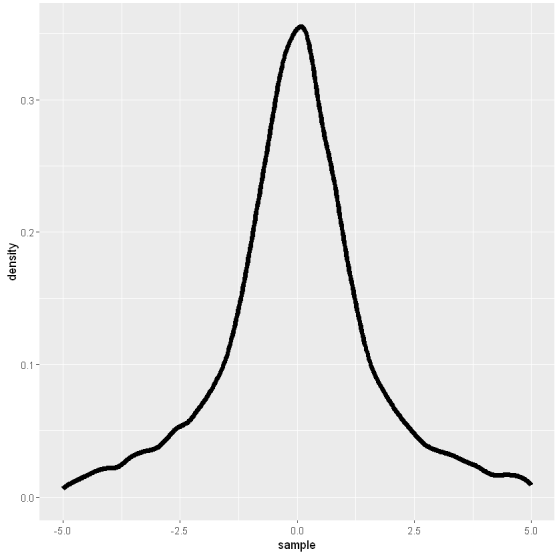
حال شما برای اینکه توزیع t را بدست بیاورید باید به دفعات متعددی از گل‌هایتان نمونه گیری انجام داده و مقادیر t آن ها را بدست بیاورید.
در سلول زیر تابعی بنویسید که که با گرفتن تعداد نمونه‌گیری‌ها و اندازه نمونه از گل های باغتان نمونه گیری انجام دهد و با استفاده از تابعی که در قسمت قبل نوشتید مقادیر t متناظر آنها را برگرداند.

In [3]:
```
funct <- function(n, amount) {
  t <- 0
  values <- rnorm(n , mean = 150 , sd = 10)
  sample <- sapply(1:n, function(i) sample(values, amount))

  for(i in 1:n){
    df <- c(sample[1:amount, i])
    t[i] <- find_t(df , mean(values))
  }

  return(t)
}
```

با استفاده از تابعی که نوشتید به تعداد ۱۰۰۰۰ بار و با اندازه ۲ نمونه گیری کنید و نتیجه را در قالب یک نمودار نمایش دهید.
برای اینکه بازه نمایش را محدود کنید میتوانید از xlim (https://ggplot2.tidyverse.org/reference/lims.html) استفاده کنید.

In [5]:
```
library(ggplot2)
sample <- funct(10000, 2)
plt <- ggplot() +
    geom_density(mapping = aes(sample), size = 2) + xlim(-5,5)
plt
```

```
Warning message:
"Removed 1270 rows containing non-finite values (`stat_density()`)."
```

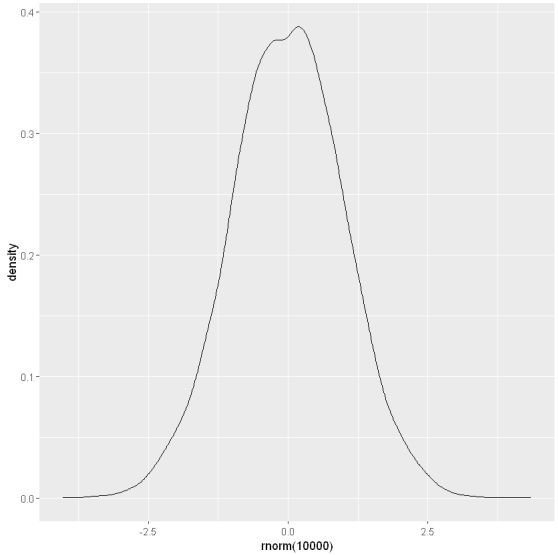


بعد از کشیدن نمودار و مشاهده آن به شباهتش با توزیع نورمال استاندارد پی می‌برید و خوشحال میشوید که CLT برای این شرایط نیز صادق است و به یک تعمیم برای قضیه حد مرکزی دست یافته‌اید!
اما وقتی بیشتر دقت میکنید حس می‌کنید یک جای این نمودار میلنگد

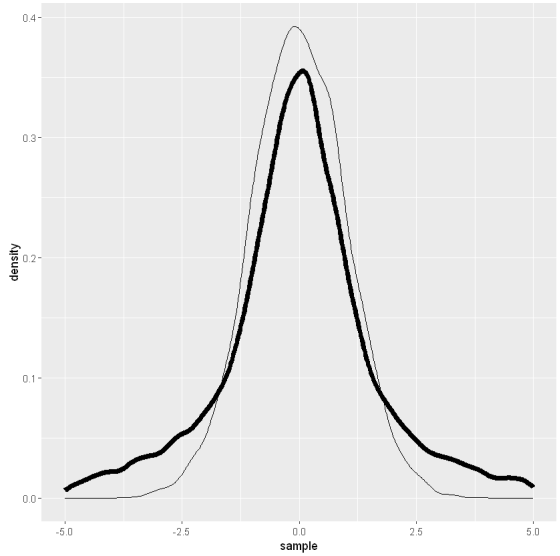
در نتیجه نمودار قبل را در کنار نمودار توزیع نرمال استاندارد رسم میکنیم:

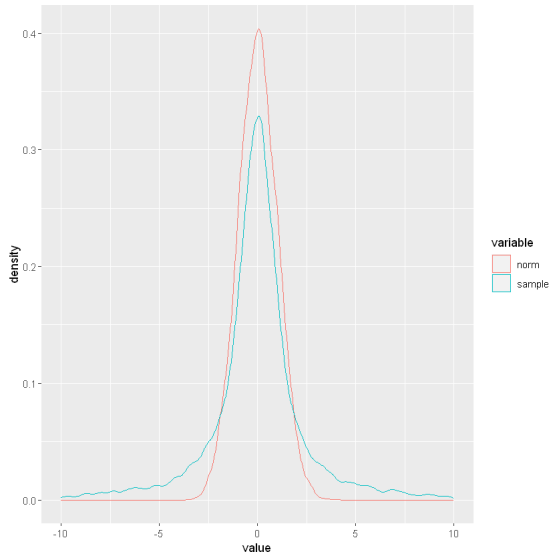
In [7]:
```
plt2 <- ggplot() +
    geom_density(mapping = aes(rnorm(10000)))
plt2
plt + geom_density(mapping = aes(rnorm(10000)))
library(reshape2)
norm <- rnorm(10000)
df <- melt(data.frame(norm, sample))
ggplot(df, aes(value, color = variable)) +
    geom_density() +
    xlim(-10, 10)
```

```
Warning message:
"Removed 1270 rows containing non-finite values (`stat_density()`)."
```




```
No id variables; using all as measure variables

Warning message:
"Removed 625 rows containing non-finite values (`stat_density()`)."
```





دو نمودار را با هم مقایسه کنید. چه تفاوتی میان آنها وجود دارد؟

نمودار ما پایین تر از نمودار نرمال می باشد و واریانس نمونه با واریانس نرمال تفاوت زیادی دارد

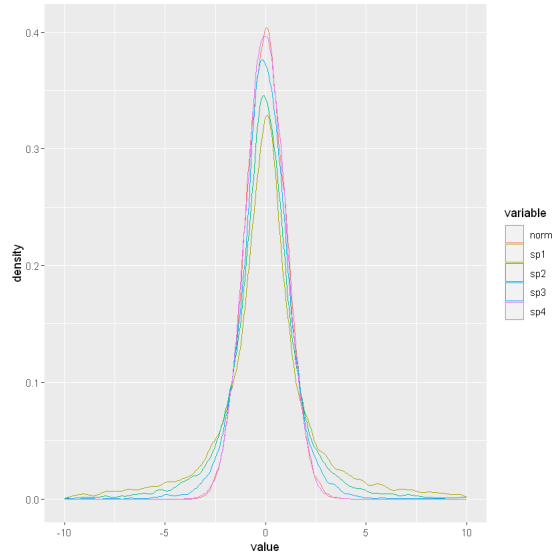
متاسفانه این دونمودار با وجود شباهت زیادی که به هم دارند از توزیع های متفاوتی می‌آیند!
شما نسبت به توزیعی که بدست آوردید کنجکاوتر شده و سعی میکنید آن را به ازای اندازه نمونه‌های مختلف بررسی کنید.

به ازای اندازه نمونه های $n = 2, n = 3, n = 6, n = 100$ مانند قبل نمونه گیری های ۱۰۰۰۰ تایی انجام داده و آن ها را داخل یک dataframe ذخیره کنید. همچنین یک ستون نیز برای توزیع نرمال استاندارد به دیتافریم اضافه کنید.

سپس در یک نمودار، توزیع های مربوط به هر یک از n های مختلف و توزیع استاندارد نرمال را رسم کرده و به هر یک از نمودار ها، یک رنگ جدا اختصاص دهید به طوری که از همدیگر قابل تمیز باشند.
راهنمایی: در رابطه با تابع melt از پکیج reshape2 تحقیق کنید.

In [10]:
```
sp1 <- funct(10000, 2)
sp2 <- funct(10000, 3)
sp3 <- funct(10000, 6)
sp4 <- funct(10000, 100)

df <- data.frame(norm, sp1, sp2, sp3, sp4)
ggplot(melt(df), aes(value, color = variable)) +
    geom_density() +
    xlim(-10, 10)
```

```
No id variables; using all as measure variables

Warning message:
"Removed 681 rows containing non-finite values (`stat_density()`)."
```



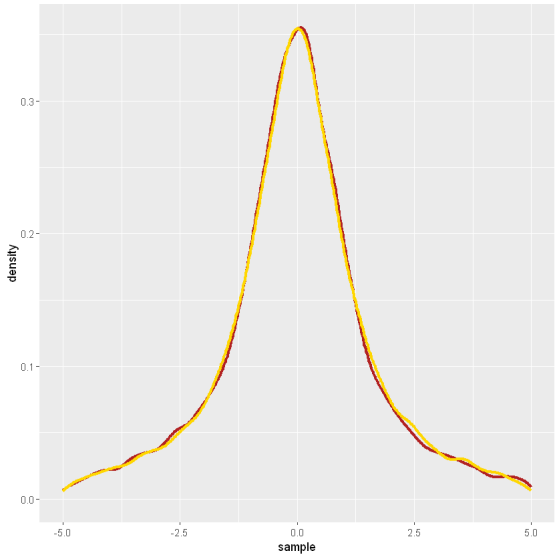
از نموداری که رسم کردید و مقایسه توزیع ها چه نتیجه‌ای میگیرید؟

هر چقدر تعداد نمونه ها بیشتر بشود، به توزیع نرمال نزدیک تر می شویم و نمودار بیشتر شبیه به نمودار نرمال عمل می کند.

روزی دوست شما، ویلیام گاست به کاشان می‌آید. شما با او در رابطه با توزیع جدیدی که کشف کردید صحبت میکنید. او می‌گوید که اتفاقا درباره این توزیع مقاله ای چاپ کرده و توابع مربوط به این توزیع را به کتابخانه پایه R اضافه کرده است. او همچنین میگوید که این توزیع شامل پارامتری به نام درجه آزادی یا ν می‌باشد که در اینجا برابر با $n-1$ است.
با استفاده از دیتافریمی که ساختید توزیع مربوط به $n=2$ و توزیع t متناظر با آن در R را در یک نمودار رسم کنید و آن دو را با هم مقایسه کنید. آیا اکنون بنظر شما این دو توزیع یکسانند؟

In [11]:
```
new_t <- rt(10000,1)

ggplot() + geom_density(aes(sample), col = "firebrick", size=1.4) +
    geom_density(aes(new_t),col = "gold" , size=1.4) +
    xlim(-5,5)
```

```
Warning message:
"Removed 1270 rows containing non-finite values (`stat_density()`)."
Warning message:
"Removed 1271 rows containing non-finite values (`stat_density()`)."
```



In [ ]:
```
# این 2 توزیع تقریبا یکسان هستند.
```

فرض کنید کسی به شما ۶ عدد گل محمدی میدهد و ادعا میکند که از باغ شما چیده است. شما به او شک می‌کنید و فرآیند گلاب‌گیری را روی آنها انجام داده و مقادیر گلاب بدست آماده از آنها برابر ۱۳۰, ۱۲۰, ۱۳۶, ۱۴۵, ۱۵۰, ۱۴۰ میلی‌لیتر شده است. آیا بنظر شما او راست می‌گوید؟

از
t-test
برای پاسخگویی استفاده می کنیم.

فرض صفر و فرض جایگزین را در این مورد بیان کنید

H0:
گل ها از باغ ما است
H1:
گلها از باغ ما نیست

In [10]:
```
# آزمون دو طرفه است.
```

مقدار p-value را یکبار با استفاده از دیتافریمی که دارید و یکبار با استفاده از توابع کتابخانه‌ای R بدست آورید و آن دو را با هم مقایسه کنید.

In [12]:
```
third_t <- find_t(c(130, 120, 136, 145, 150, 140), mean(rnorm(1000000 , mean = 150 , sd = 10)))
p_value <- 2 * pt(third_t, 5)
p_value

cons <- 0
for(i in 1:10000){
  if(df$sp3[i] <= third_t || df$sp3[i] >= (-1) * third_t)
    cons <- cons + 1
}

p_value <- cons / length(df$sp3)
p_value
```

0.0304032836427777

0.0302

In [26]:
```
# اعداد بسیار نزدیک به هم می باشند و با مقایسه آنها با 0.05، فرض صفر رد می شود.
```

سوال دوم

یک کمپانی سازنده توپ گلف تصمیم دارد که مقاومت توپ های خود را افزایش دهد. برای این کار لایه‌ای بر توپ های خود می‌افزاید. تست مقاومت برای توپ های جدید رضایت‌بخش است و توپ های جدید مقاومت بیشتری نسبت به توپ های قبلی دارند ولی قبل از تولید انبوه یکی از محققان این شرکت ادعا میکند که توپ های جدید نسبت به توپ های قبلی مسافت کوتاه تری را می‌پیمایند. برای این کار او در 40 بار در شرایط مختلف توپ ها را با دستگاهی پرتاب کرد تا تنها عامل تفاوت بین دو پرتاب نوع توپ ها باشند. این محقق برای اثبات ادعایش با استفاده از تست جایگشتی از شما کمک خواسته برای این کار ابتدا مجموعه دادگان شامل عملکرد توپ های فعلی و جدید را از روی فایلGolf.csv بخوانید:

In [13]:
```
golf <- read.csv("Golf.csv")
head(golf, 10)
```

A data.frame: 10 × 2

	Current	New
	<int>	<int>
1	264	277
2	261	269
3	267	263
4	272	266
5	258	262
6	283	251
7	258	262
8	266	289
9	259	286
10	270	264

ستون اول دادگان شامل عملکرد توپ های فعلی و ستون دوم آن شامل عملکرد متناظر توپ های جدید می‌باشد. تفاوت میانگین دو عملکرد را برای دو نوع توپ محاسبه کنید.

In [14]:
```
main_mean <- mean(golf$Current) - mean(golf$New)
main_mean
```

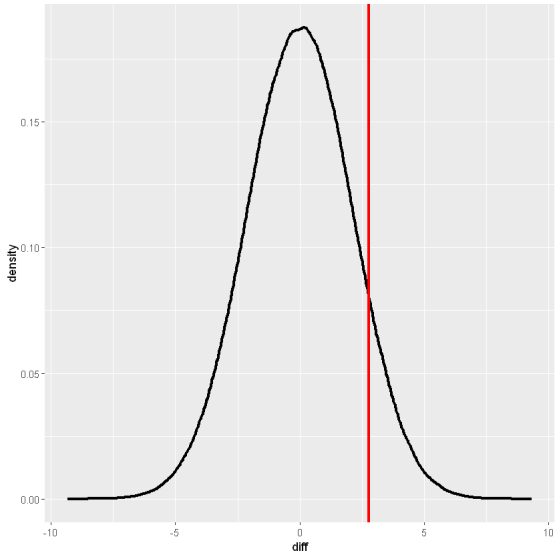
2.77499999999998

تفاوت میانگین را برای جایشگت های مختلف داده محاسبه کنید و در نهایت با استفاده از هیستوگرام مکان داده مشاهده شده را معین کنید.

In [15]:
```
cal_mean <- function(x, y) (mean(x) - mean(y))
sample_vector <- c(golf$Current, golf$New)

diff <- replicate(500000, {i <- sample(1:80, 40); cal_mean(sample_vector[i], sample_vector[-i])})
summary(diff)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
-9.325000 -1.425000  0.025000  0.005692  1.425000  9.325000 
```

In [16]:
```
library(ggplot2)
ggplot(data.frame(dist = diff), aes(diff)) +
        geom_density(size = 1.25) +
        geom_vline(xintercept=main_mean, col="red", size = 1.2)
```



p-value را محاسبه کنید، آیا فرض صفر رد می شود؟

In [17]:
```
(sum(diff > main_mean) + sum(diff < (-1*main_mean))) / length(diff)
```

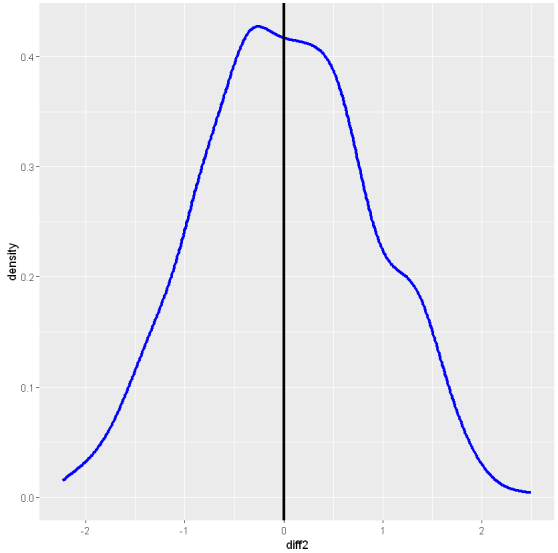
0.183416

In [42]:
```
# خیر زیرا از 0.05 بزرگتر می باشد.
```

پس از انجام آزمایش او به قدرت تست جایگشتی در مسئله شرکت توپ گلف سازی شک کرد و تصمیم گرفت با انجام آزمایش هایی کارایی آن برای مسئله فعلی را مورد جدال قرار دهد! برای این کار او قصد دارد 1000 بار آزمون جایگشتی را برای دو دسته داده که هر دو از توزیع نرمال با میانگین صفر ولی

انحراف معیار های متفاوت 1 و 2 امده‌اند اجرا کند و p-value های به دست امده را بررسی کند. ابتدا به مدت 1000 مرتبه تست جایگشتی را روی دو دسته با اندازه 10 و 20 اجرا کنید.

In [18]:
```
sample_1 <- rnorm(10, 0, 1)
sample_2 <- rnorm(20, 0, 2)
all_sample <- c(sample_1, sample_2)
diff2 <- replicate(1000, {i <- sample(1:30, 20); cal_mean(all_sample[i], all_sample[-i])})
ggplot(data.frame(dist = diff2), aes(diff2)) +
        geom_density(size = 1.25, col="blue") +
        geom_vline(xintercept=0, size = 1.3)
```



In [27]:
```
ps <- matrix(0, nrow = 10000, ncol = 30)
difference <- abs(mean(sample_1)-mean(sample_2))
all = c(sample_1,sample_2)
# calculating p_values:
p_values <- 0

for(j in 1:1000){
    for(i in 1:1000){
      ps[i,] <- sample(all, size = 30, replace = FALSE)
    }
    d <-0
    for(i in 1:1000){
      d[i] <- abs(mean(ps[i,1:10]) - mean(ps[i,11:30]))
    }
    greater = d[d >= difference]
    pValue = length(greater) / 1000
    p_values[j] = pValue
}
```

بازه صفر تا یک را از نقاط 0.05 و 0.1 و 0.5 و 0.9 و 0.95 به بازه های کوچکتر تقسیم کنید و تعداد p-value هایی که در هر کدام از بازه ها می‌افتند را محاسبه کنید.

In [28]:
```
length(p_values[p_values<0.05])
length(p_values[p_values>=0.05 & p_values<0.1])
length(p_values[p_values>=0.1 & p_values<0.5])
length(p_values[p_values>=0.5 & p_values<0.9])
length(p_values[p_values>=0.9 & p_values<0.95])
length(p_values[p_values>0.95])
# ---------------------------------------------------
# Maybe we can use another way:
sum(diff2 <= 0.05)
sum(diff2 > 0.05 & diff2 <= 0.1)
sum(diff2 > 0.1 & diff2 <= 0.5)
sum(diff2 > 0.5 & diff2 <= 0.9)
sum(diff2 > 0.9 & diff2 <= 0.95)
sum(diff2 > 0.95)
```

```
0
```

```
0
```

```
1000
```

```
0
```

```
0
```

```
0
```

```
508
```

```
2
```

```
53
```

```
30
```

```
3
```

```
404
```

اگر انحراف معیار دو دسته را عوض کنیم نتایج چه تغییری خواهند کرد؟
با توجه به نتایج به دست آمده توضیح دهید که آگر از آزمون جایگشت برای مقایسه میانگین ها استفاده کنیم با چه مشکلاتی مواجه خواهیم شد؟

In [24]:
```r
# با انحراف معیار جدید امتحان می کنیم
sample_1 <- rnorm(10, 0, 1.5)
sample_2 <- rnorm(20, 0, 7.5)
all_sample <- c(sample_1, sample_2)
diff2 <- replicate(1000, {i <- sample(1:30, 20); cal_mean(all_sample[i], all_sample[-i])})
ggplot(data.frame(dist = diff2), aes(diff2)) +
        geom_density(size = 1.25, col="blue") +
        geom_vline(xintercept=0, size = 1.3)

ps <- matrix(0, nrow = 10000, ncol = 30)
difference <- abs(mean(sample_1)-mean(sample_2))
all = c(sample_1,sample_2)
# calculating p_values:
p_values <- 0

for(j in 1:1000){
    for(i in 1:1000){
      ps[i,] <- sample(all, size = 30, replace = FALSE)
    }
    d <-0
    for(i in 1:1000){
      d[i] <- abs(mean(ps[i,1:10]) - mean(ps[i,11:30]))
    }
    greater = d[d >= difference]
    pValue = length(greater) / 1000
    p_values[j] = pValue
}

length(p_values[p_values<0.05])
length(p_values[p_values>=0.05 & p_values<0.1])
length(p_values[p_values>=0.1 & p_values<0.5])
length(p_values[p_values>=0.5 & p_values<0.9])
length(p_values[p_values>=0.9 & p_values<0.95])
length(p_values[p_values>0.95])

# با انحراف معیار جدید امتحان می کنیم
sample_1 <- rnorm(10, 0, 2)
sample_2 <- rnorm(20, 0, 11)
all_sample <- c(sample_1, sample_2)
diff2 <- replicate(1000, {i <- sample(1:30, 20); cal_mean(all_sample[i], all_sample[-i])})
ggplot(data.frame(dist = diff2), aes(diff2)) +
        geom_density(size = 1.25, col="blue") +
        geom_vline(xintercept=0, size = 1.3)

ps <- matrix(0, nrow = 10000, ncol = 30)
difference <- abs(mean(sample_1)-mean(sample_2))
all = c(sample_1,sample_2)
# calculating p_values:
p_values <- 0

for(j in 1:1000){
    for(i in 1:1000){
      ps[i,] <- sample(all, size = 30, replace = FALSE)
    }
    d <-0
    for(i in 1:1000){
      d[i] <- abs(mean(ps[i,1:10]) - mean(ps[i,11:30]))
    }
    greater = d[d >= difference]
    pValue = length(greater) / 1000
    p_values[j] = pValue
}

length(p_values[p_values<0.05])
```

```
length(p_values[p_values>=0.05 & p_values<0.1])
length(p_values[p_values>=0.1 & p_values<0.5])
length(p_values[p_values>=0.5 & p_values<0.9])
length(p_values[p_values>=0.9 & p_values<0.95])
length(p_values[p_values>0.95])
```

0

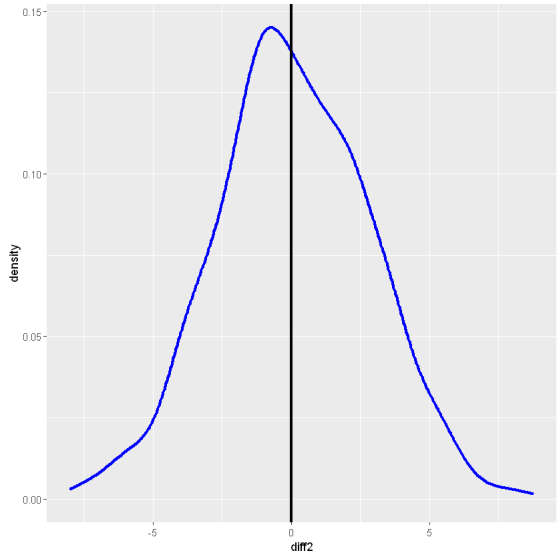
0

0

0

0

1000



0

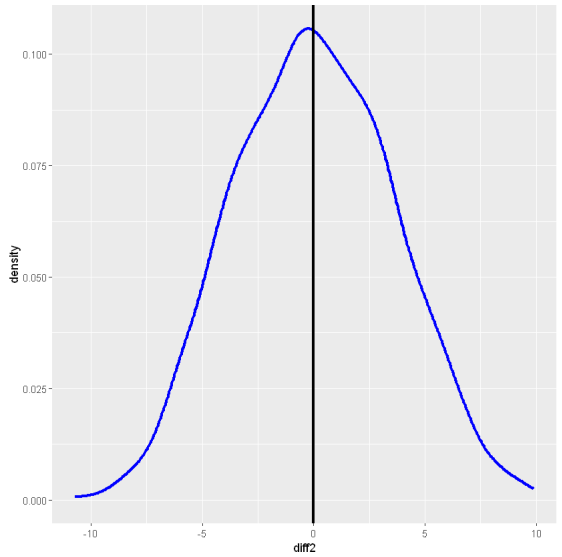
0

1000

0

0

0



چون دو توزیع نرمال می باشند، به نتیجه غلط خواهیم رسید