Student Crimtab Data for Simulation of T-Distribution

coop711

2015년 3월 25일

Data Manipulation

- Data Frame으로 정리하고, 다시 long format 으로 변환. height를 인치로 변환하면 어떻게 될까?

```
dimnames(crimtab.2)[[2]]<-as.numeric(dimnames(crimtab.2)[[2]])/2.54
crimtab.2
```

```
##       56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77
## 9.4    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 9.5    0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 9.6    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 9.7    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 9.8    0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 9.9    0  0  1  0  1  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 10     1  0  0  1  2  0  2  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
## 10.1   0  0  0  1  3  1  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0
## 10.2   0  0  2  2  2  1  0  2  0  1  0  0  0  0  0  0  0  0  0  0  0  0
## 10.3   0  1  1  3  2  2  3  5  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 10.4   0  0  1  1  2  3  3  4  3  3  0  0  0  0  0  0  0  0  0  0  0  0
## 10.5   0  0  0  1  3  7  6  4  3  1  3  1  0  1  0  0  0  0  0  0  0  0
## 10.6   0  0  0  1  4  5  9 14  6  3  1  0  0  1  0  0  0  0  0  0  0  0
## 10.7   0  0  1  2  4  9 14 16 15  7  3  1  2  0  0  0  0  0  0  0  0  0
## 10.8   0  0  0  2  5  6 14 27 10  7  1  2  1  0  0  0  0  0  0  0  0  0
## 10.9   0  0  0  0  2  6 14 24 27 14 10  4  1  0  0  0  0  0  0  0  0  0
## 11     0  0  0  2  6 12 15 31 37 27 17 10  6  0  0  0  0  0  0  0  0  0
## 11.1   0  0  0  3  3 12 22 26 24 26 24  7  4  1  0  0  0  0  0  0  0  0
## 11.2   0  0  0  3  2  7 21 30 38 29 27 20  4  1  0  0  0  0  0  0  0  1
## 11.3   0  0  0  1  0  5 10 24 26 39 26 24  7  2  0  0  0  0  0  0  0  0
## 11.4   0  0  0  0  3  4  9 29 56 58 26 22 10 11  0  0  0  0  0  0  0  0
## 11.5   0  0  0  0  0  5 11 17 33 57 38 34 25 11  2  0  0  0  0  0  0  0
## 11.6   0  0  0  0  2  1  4 13 37 39 48 38 27 12  2  2  0  1  0  0  0  0
## 11.7   0  0  0  0  0  2  9 17 30 37 48 45 24  9  9  2  0  0  0  0  0  0
## 11.8   0  0  0  0  1  0  2 11 15 35 41 34 29 10  5  1  0  0  0  0  0  0
## 11.9   0  0  0  0  1  1  2 12 10 27 32 35 19 10  9  3  1  0  0  0  0  0
## 12     0  0  0  0  0  0  1  4  8 19 42 39 22 16  8  2  2  0  0  0  0  0
## 12.1   0  0  0  0  0  0  0  2  4 13 22 28 15 27 10  4  1  0  0  0  0  0
## 12.2   0  0  0  0  0  0  1  2  5  6 23 17 16 11  8  1  1  0  0  0  0  0
## 12.3   0  0  0  0  0  0  0  0  4  8 10 13 20 23  6  5  0  0  0  0  0  0
## 12.4   0  0  0  0  0  0  1  1  1  2  7 12  4  7  7  1  0  0  1  0  0  0
## 12.5   0  0  0  0  0  0  0  1  0  1  3 12 11  8  6  8  0  2  0  0  0  0
## 12.6   0  0  0  0  0  0  0  0  0  1  0  3  5  7  8  6  3  1  1  0  0  0
## 12.7   0  0  0  0  0  0  0  0  0  1  1  7  5  5  8  2  2  0  0  0  0  0
## 12.8   0  0  0  0  0  0  0  0  0  0  1  2  3  1  8  5  3  1  1  0  0  0
## 12.9   0  0  0  0  0  0  0  0  0  0  0  1  2  2  0  1  1  0  0  0  0  0
## 13     0  0  0  0  0  0  0  0  0  0  3  0  1  0  1  0  2  1  0  0  0  0
## 13.1   0  0  0  0  0  0  0  0  0  0  0  1  1  0  0  0  0  0  0  0  0  0
## 13.2   0  0  0  0  0  0  0  0  0  0  1  1  0  1  0  3  0  0  0  0  0  0
## 13.3   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  1  0  0  0
## 13.4   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 13.5   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0
```

```
crimtab.3<-crimtab.2
crimtab.3.df<-as.data.frame(crimtab.3, stringsAsFactors = F)
head(crimtab.3.df)
```

```
##   Var1 Var2 Freq
## 1  9.4   56    0
## 2  9.5   56    0
## 3  9.6   56    0
## 4  9.7   56    0
## 5  9.8   56    0
## 6  9.9   56    0
```

```
str(crimtab.3.df)
```

```
## 'data.frame':    924 obs. of  3 variables:
##  $ Var1: chr  "9.4" "9.5" "9.6" "9.7" ...
##  $ Var2: chr  "56" "56" "56" "56" ...
##  $ Freq: int  0 0 0 0 0 0 1 0 0 0 ...
```

```
crimtab.3.df$finger<-as.numeric(crimtab.3.df$Var1)
crimtab.3.df$height<-as.numeric(crimtab.3.df$Var2)
str(crimtab.3.df)
```

```
## 'data.frame':    924 obs. of  5 variables:
##  $ Var1  : chr  "9.4" "9.5" "9.6" "9.7" ...
##  $ Var2  : chr  "56" "56" "56" "56" ...
##  $ Freq  : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ finger: num  9.4 9.5 9.6 9.7 9.8 9.9 10 10.1 10.2 10.3 ...
##  $ height: num  56 56 56 56 56 56 56 56 56 56 ...
```

```
crimtab.3.long<-apply(crimtab.3.df[,4:5], 2, function(x) rep(x, crimtab.3.d
f[,3]))
str(crimtab.3.long)
```

```
##  num [1:3000, 1:2] 10 10.3 9.9 10.2 10.2 10.3 10.4 10.7 10 10.1 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:2] "finger" "height"
```

Student 의 Simulation 재현 ————————————

- 3,000장의 카드를 잘 섞는 것은 sample() 이용.

```
r  crimtab.shuffle<-crimtab.3.long[sample(1:3000),]
```

- 표본의 크기가 4인 750개의 표본을 만드는 작업은 rep() 이용.

```
crimtab.shuffle.sample.id<-as.factor(rep(1:750, each=4))
```

- 각 표본의 평균과 표준편차 계산에는 tapply() 이용.

```
crimtab.shuffle.sample.mean.finger<-tapply(crimtab.shuffle[,1],crimtab.shuffl
e.sample.id,mean)
crimtab.shuffle.sample.sd.finger<-tapply(crimtab.shuffle[,1],crimtab.shuffle.sa
mple.id,sd)
```

- t-통계량 계산. Student는 표준편차 계산에서 분모에 n을 사용하고 히스토그램을 그려 비교하였으나 자유도 3인 t-분포와 비교하기 위하여 $t = \frac{\bar{X}_n - \mu}{SD/\sqrt{n}}$을 계산함.

```
crimtab.shuffle.sample.t<-(crimtab.shuffle.sample.mean.finger-mean(crimtab.3.lo
ng[,1]))/(crimtab.shuffle.sample.sd.finger/sqrt(4))
```

- 계산한 t-통계량 값들의 평균과 표준편차, 히스토그램을 그리고 자유도 3인 t-분포의 밀도함수 및 표준정규곡선과 비교. 우선 모두 같은 값들이 나와서 분모가 0인 경우가 있는지 파악. 있으면 모평균과 비교하여 양수인 경우 +6, 음수인 경우 -6 값 부여(Student가 한 일)

```
t.inf<-is.infinite(crimtab.shuffle.sample.t)
crimtab.shuffle.sample.t[t.inf]
```

```
##  66
## Inf
```

```
crimtab.shuffle.sample.t[t.inf]<-6*sign(crimtab.shuffle.sample.t[t.inf])
```

- 문제되는 값이 없는 것을 확인하고, 평균과 표준편차 계산. 자유도 n인 t-분포의 평균과 표준편차은 각각 0과 $\sqrt{\frac{n}{n-2}}$임을 상기할 것.

```
mean(crimtab.shuffle.sample.t)
```

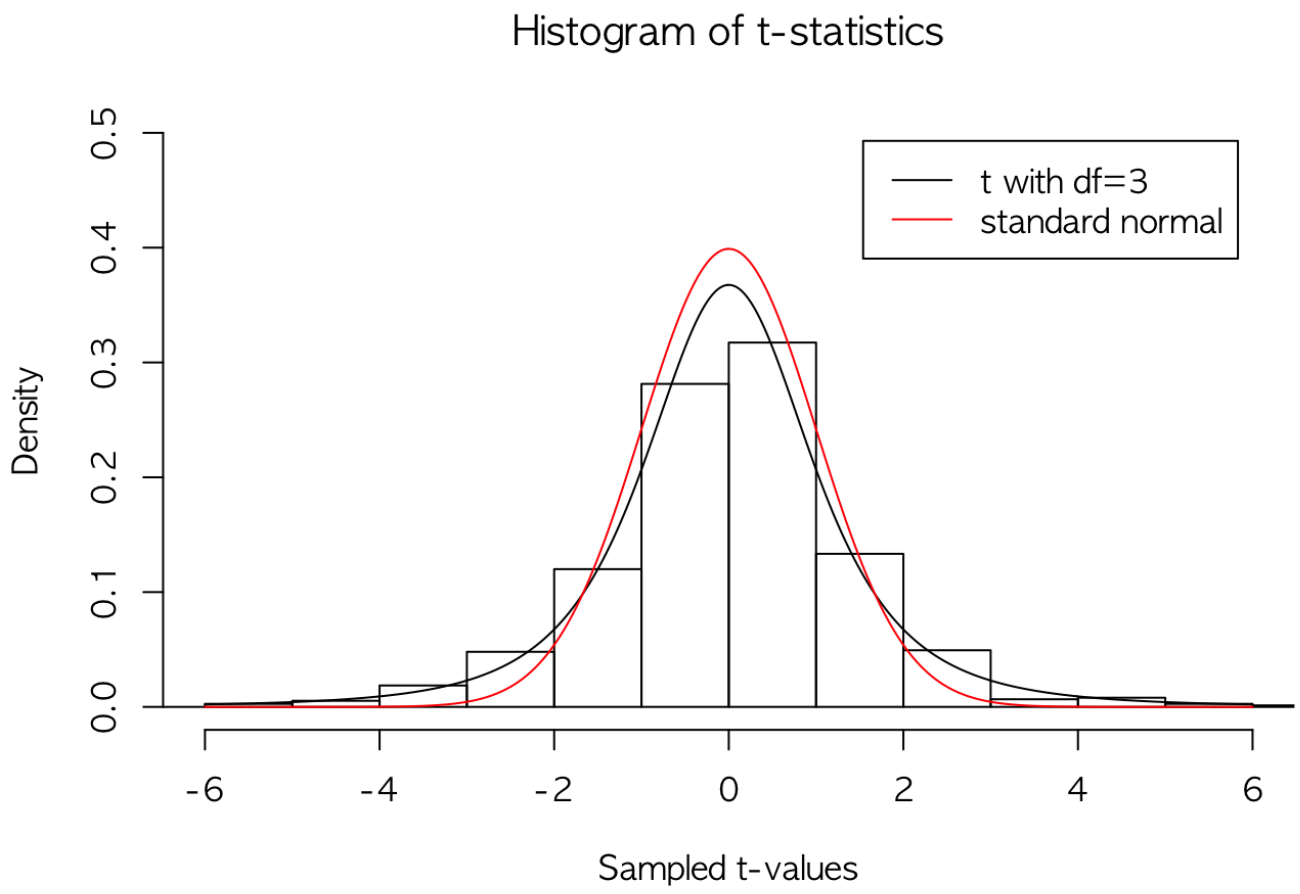
```
## [1] -0.03145874
```

```
sd(crimtab.shuffle.sample.t)
```

```
## [1] 1.55559
```

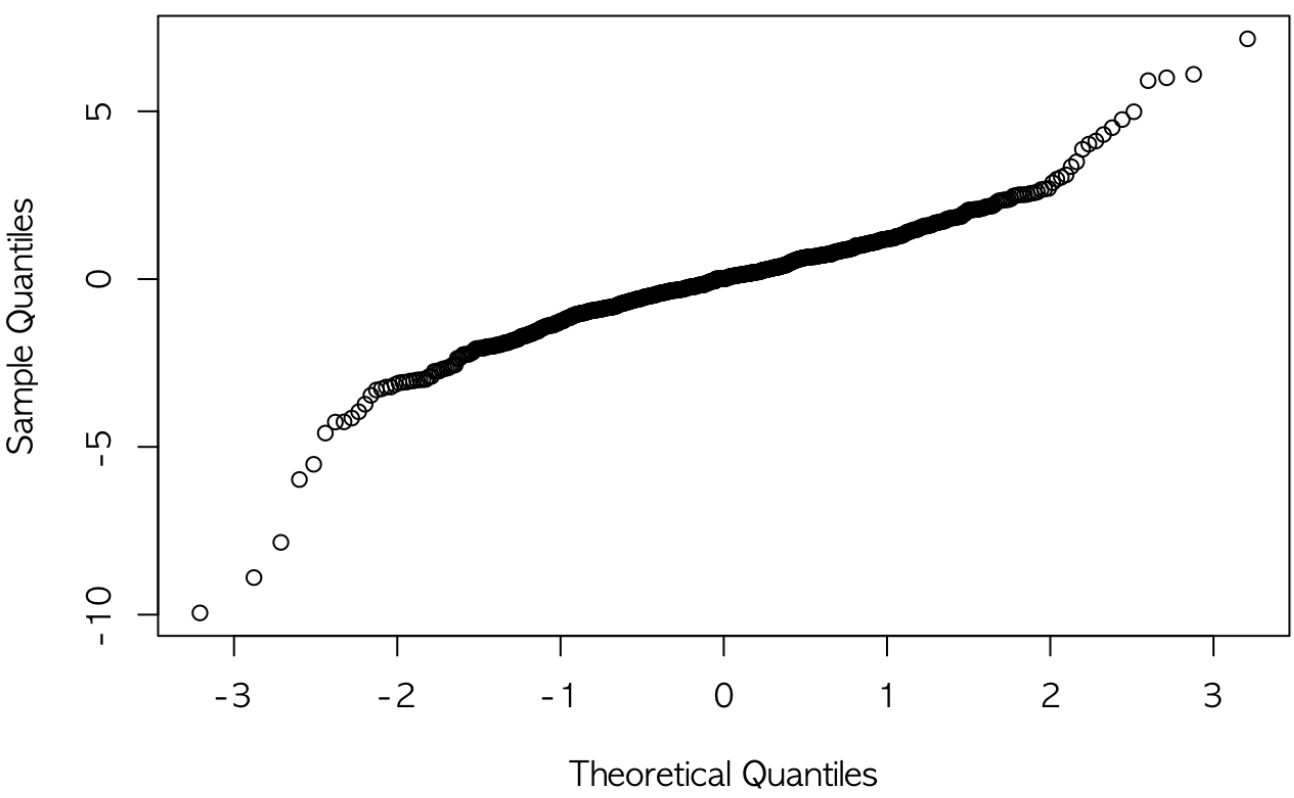
- t-통계량들의 히스토그램을 그리고, 자유도 3인 t의 밀도함수, 표준정규분포 밀도함수와 비교.

```
hist(crimtab.shuffle.sample.t,prob=T,nclass=20,xlim=c(-6,6), ylim=c(0,0.5), mai
n="Histogram of t-statistics", xlab="Sampled t-values")
lines(seq(-6,6,by=0.01), dt(seq(-6,6, by=0.01), df=3))
lines(seq(-6,6,by=0.01), dnorm(seq(-6,6, by=0.01)), col="red")
legend("topright", inset=0.05, lty=1, col=c("black","red"), legend=c("t with d
f=3", "standard normal"))
```



- qqnorm() 을 그려보면 정규분포와 꼬리에서 큰 차이가 난다는 것을 알 수 있음.

```
qqnorm(crimtab.shuffle.sample.t)
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
5
0
-5
-10
-3
-2
-1
0
1
2
3