Siteng Zhang, Jina Park, Marco Ma

# Monte Carlo Simulation to Identify Independence and $t$-distribution

### Introduction

The student's theorem have stated that following the definition of a t-distributed RV, the independence of $\bar{X}$ and $S^2$ with r degrees of freedom iff will lead to $T = \sqrt{n}\frac{\bar{X}-\mu}{S} \sim t_{n-1}$. Thus, it aroused our interest to investigate possible dependents of $Z \sim N(0,1)$ and $U \sim \chi_r{}^2$ such that $T = \frac{z}{\sqrt{U/r}} \sim t_r$. Different techniques such as Monte Carlo simulations and transformations methods are being applied while solving the questions. Results were being obtained throughout the exploration, which $Z$ and $U$ are not independent as $X_1$ and $S^2$ are not independent.

### Student's Theorem

According to Student's Theorem, if $X_1,\ldots,X_n \sim N(\mu,\frac{\sigma^2}{n})$, the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu,\frac{\sigma^2}{n})$ and sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$ are independent. The independence of $\bar{X}$ and $S^2$ leads to $T = \sqrt{n}\frac{\bar{X}-\mu}{S} \sim t_{n-1}$. This follows the definition of a $t$-distributed RV.

### The definition of a $t$-distributed RV

Let $Z \sim N(0,1)$ and $U \sim \chi_r{}^2$ such that $Z$ and $U$ are independent. A RV $T$ is said to have a $t$-distribution with $r$ degrees of freedom iif $T = \frac{Z}{\sqrt{U/r}} \sim t_r$.

### Independence and $t$-distribution

Suppose $X_1,\ldots,X_n \sim N(0,1)$. Let $Z = X_1$ and $U = (n-1)S^2$, so that $Z \sim N(0,1)$ and $U \sim \chi_{n-1}{}^2$. It follows that $Z$ and $U$ are not independent if $X_1$and $S^2$ are not independent according to the theorem and definition explained above.

Suppose we consider the case $n = 2$, $X_1,X_2 \sim N(0,1)$, so that $Z = X_1$ and $U = S^2 = \frac{1}{2}(X_1 - X_2)^2$.

We first obtain the joint distribution of $X$. Since $X_1$ and $X_2$ are standard normal distributions, we have the following equation.

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x_1^2}{2}} * \frac{1}{\sqrt{2\pi}}e^{-\frac{x_2^2}{2}} = \frac{1}{2\pi}e^{-\frac{(x_1^2+x_2^2)}{2}}, (x_1,x_2) \in R^2$$

In addition, suppose $Y = (Z,U)^T, X = (X_1,X_2)^T$.

We use the transformation method by equating and arranging the vectors $Y$ and $X$ with respect to $x_1$ and $x_2$ as follows;

$$\begin{pmatrix} z = x_1 \\ u = \frac{1}{2}(x_1 - x_2)^2 \end{pmatrix} \qquad \begin{pmatrix} x_1 = z \\ x_2 = x_1 \pm \sqrt{2u} \end{pmatrix}$$

$$\Leftrightarrow$$

Thus, we have $x_1 = z$ and $x_2 = x_1 \pm \sqrt{2u}$.

Since we have two different values for $x_2$, there are two ways to define range as following;

$$R_1 = \{(x_1, x_2) | x_1 < x_2, x_1, x_2 \in R\}$$
$$R_2 = \{(x_1, x_2) | x_1 \geq x_2, x_1, x_2 \in R\}$$

For $R_1 = \{(x_1, x_2) | x_1 < x_2, x_1, x_2 \in R\}$, we have $x_1 = z$ and $x_2 = x_1 + \sqrt{2u}$ where $z \in R, u > 0$, hence, we obtain the following for $|J|$.

$$|J| = \begin{bmatrix} \dfrac{\partial x_1}{\partial z} & \dfrac{\partial x_1}{\partial u} \\ \dfrac{\partial x_2}{\partial x} & \dfrac{\partial x_2}{\partial u} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & \frac{1}{\sqrt{2u}} \end{bmatrix} = \frac{1}{\sqrt{2u}}$$

and the following joint distribution.

$$\begin{aligned} g_{Z,U}(z,u) &= \frac{1}{2\pi}\frac{1}{\sqrt{2u}} e^{\frac{z^2+(x_1+\sqrt{2u})^2}{2}} I_{(-\infty,+\infty)}(z) I_{(0,+\infty)}(u) \\ &= \frac{1}{2\pi\sqrt{2u}} e^{\frac{z^2+(z+\sqrt{2u})^2}{2}} I_{(-\infty,+\infty)}(z) I_{(0,+\infty)}(u) \\ &= \frac{1}{2\pi\sqrt{2u}} e^{-\frac{z^2+z^2+2z\sqrt{2u}+2u}{2}} I_{(-\infty,+\infty)}(z) I_{(0,+\infty)}(u) \\ &= \frac{1}{2\pi\sqrt{2u}} e^{-(z^2+z\sqrt{2u}+u)} I_{(-\infty,+\infty)}(z) I_{(0,+\infty)}(u) \end{aligned}$$

Similarly, for $R_2 = \{(x_1, x_2) | x_1 \geq x_2, x_1, x_2 \in R\}$, we obtain $|J|$ and joint distribution as follows.

$$|J| = \begin{bmatrix} \dfrac{\partial x_1}{\partial z} & \dfrac{\partial x_1}{\partial u} \\ \dfrac{\partial x_2}{\partial x} & \dfrac{\partial x_2}{\partial u} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & -\frac{1}{\sqrt{2u}} \end{bmatrix} = |-\frac{1}{\sqrt{2u}}| = \frac{1}{\sqrt{2u}}$$

$$\begin{aligned} g_{Z,U}(z,u) &= \frac{1}{2\pi}\frac{1}{\sqrt{2u}} e^{\frac{z^2+(x_1-\sqrt{2u})^2}{2}} I_{(-\infty,+\infty)}(z) I_{(0,+\infty)}(u) \\ &= \frac{1}{2\pi\sqrt{2u}} e^{\frac{z^2+(z-\sqrt{2u})^2}{2}} I_{(-\infty,+\infty)}(z) I_{(0,+\infty)}(u) \\ &= \frac{1}{2\pi\sqrt{2u}} e^{-\frac{z^2+z^2-2z\sqrt{2u}+2u}{2}} I_{(-\infty,+\infty)}(z) I_{(0,+\infty)}(u) \\ &= \frac{1}{2\pi\sqrt{2u}} e^{-(z^2-z\sqrt{2u}+u)} I_{(-\infty,+\infty)}(z) I_{(0,+\infty)}(u) \end{aligned}$$

Therefore, we obtain the the joint distribution of $x_1$ and $x_2$ as follows;

$$\begin{aligned} g_{Z,U}(z,u) &= (\frac{1}{2\pi\sqrt{2u}} e^{-(z^2+z\sqrt{2u}+u)} + \frac{1}{2\pi\sqrt{2u}} e^{-(z^2-z\sqrt{2u}+u)}) I_{(-\infty,+\infty)}(z) I_{(0,+\infty)}(u) \\ &= \frac{1}{2\pi\sqrt{2u}} (e^{-(z^2-z\sqrt{2u}+u)} + e^{-(z^2+z\sqrt{2u}+u)}) I_{(-\infty,+\infty)}(z) I_{(0,+\infty)}(u) \end{aligned}$$

Suppose we have independent $Z$ and $U$ such as

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, z \in R$$

$$f_U(u) = \frac{1}{2\pi\sqrt{2u}} \int_{-\infty}^{+\infty} (e^{-(z^2 - z\sqrt{2u} + u)} + e^{-(z^2 + z\sqrt{2u} + u)}) dz I_{(0,+\infty)}(u) = \frac{1}{2\pi\sqrt{u}} e^{-(u + \frac{u^2}{2})} I_{(0,+\infty)}(u)$$

Then, the joint distribution would be

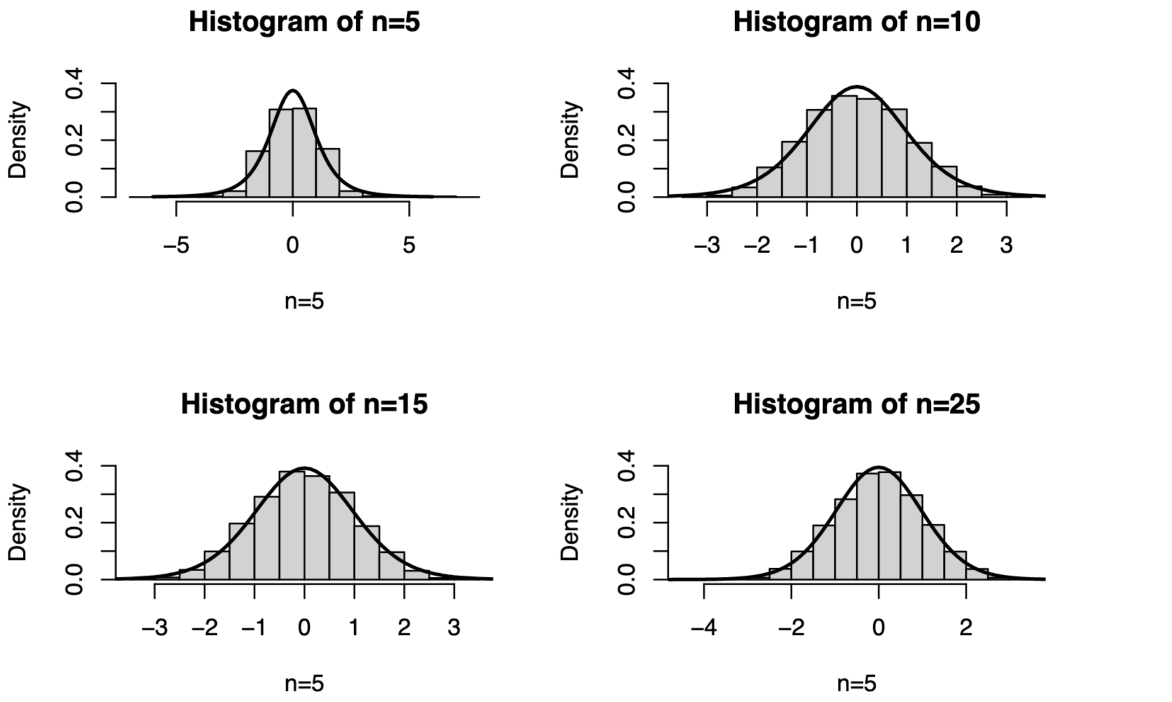
$$f_Z(z) f_U(u) = \frac{1}{2\pi\sqrt{u}} e^{-(u + \frac{u^2}{2} + \frac{z^2}{2})} I_{(-\infty,+\infty)}(z) I_{(0,+\infty)}(u)$$

which does not equal to $g_{Z,U}(z,u)$.

Therefore, $Z$ and $U$ are not independent as $X_1$ and $S^2$ are not independent.

## Monte Carlo Simulations

Now we are conducting Monte Carlo simulation to compute the $t$-RV $T = \frac{z}{\sqrt{U/n-1}}$, where $Z = X_1$ and $U = (n-1)S^2$. 10,000 random samples were generated from $X_1, \cdots, X_n \sim N(0,1)$ for fixed $n$ = 5, 10, 15, and 25, each of the random samples is computing the distribution of $T = \frac{z}{\sqrt{U/n-1}}$. The following histograms are demonstrating the distribution of $T = \frac{z}{\sqrt{U/n-1}}$ and the density curves of the $\boldsymbol{t_{n-1}}$distribution of each $n$.




**Comparing quantiles of $\boldsymbol{T = \frac{z}{\sqrt{U/n-1}}}$ and $\boldsymbol{t_{n-1}}$ distributions**

```
##              n=5         t4       n=10         t9       n=15        t14
## 7.5%  -1.5308138 -1.7781922 -1.4744623 -1.5737358 -1.4687825 -1.5230951
## 80%    0.9812642  0.9409646  0.9005082  0.8834039  0.8446103  0.8680548
## 85%    1.1763139  1.1895669  1.0936191  1.0997162  1.0423004  1.0762802
## 90%    1.4033529  1.5332063  1.3553463  1.3830287  1.2920688  1.3450304
## 95%    1.7341207  2.1318468  1.7096401  1.8331129  1.6216817  1.7613101
## 99.5%  2.7750624  4.6040949  2.4718489  3.2498355  2.3987240  2.9768427
##             n=25        t24
## 7.5%  -1.4805887 -1.4871358
## 80%    0.8840362  0.8568555
## 85%    1.0692162  1.0593189
## 90%    1.3147886  1.3178359
## 95%    1.6651119  1.7108821
## 99.5%  2.4099883  2.7969395
```

t4, t9, t14, and t24 stand for true quantiles of the $t_{n-1}$ distributions, respectively.

### Conclusion

From the table above, we observe that $T$ does not follow $t_{n-1}$ distribution because there is a significant difference between estimated value and true value at p=0.995. Since there is a significant difference after 10000 simulations, the independence of $Z$ and $U$ is both necessary and sufficient for $T$ to have a $t_{n-1}$ distribution. When $n$ increases, the difference of quantiles is not significant. There is a situation in which it would be reasonable to approximate the distribution of $T$ with the $t_{n-1}$ distribution, such as $n$ is greater than 25.

APPENDIX

The following is R-code used to conduct Monte Carlo simulation. What' s written after # is the explanation of the codes.

```
#Produce the same random values
set.seed(2022)

#Set the function
Tvalue = function(X,n){
  Z <- X[[1]]
  U <- (n-1)*var(X)
  t <- Z/sqrt(U/(n-1))
  return(t)
}

#Do the simulation
R=10000
T5 <- rep(0,R)
T10 <- rep(0,R)
T15 <- rep(0,R)
T25 <- rep(0,R)
for (i in 1:R){
  X5 <- rnorm(5,0,1)
  X10 <- rnorm(10,0,1)
  X15 <- rnorm(15,0,1)
  X25 <- rnorm(25,0,1)
  T5[i] <- Tvalue(X5,5)
  T10[i] <- Tvalue(X10,10)
  T15[i] <- Tvalue(X15,15)
  T25[i] <- Tvalue(X25,25)
}

#Plot the histogram
x=seq(-5,5,0.01)
par(mfrow=c(2,2))
hist(T5,freq=FALSE,main="Histogram of n=5",xlab="n=5",ylim = c(0, 0.4))
curve(dt(x, 4), -6, 6, add=T,lwd = 2, ylim = c(0, 0.4))
hist(T10,freq=FALSE,main="Histogram of n=10",xlab="n=5",ylim = c(0, 0.4))
curve(dt(x, 9), -6, 6, add=T,lwd = 2, ylim = c(0, 0.4))
hist(T15,freq=FALSE,main="Histogram of n=15",xlab="n=5",ylim = c(0, 0.4))
curve(dt(x, 14), -6, 6, add=T,lwd = 2, ylim = c(0, 0.4))
hist(T25,freq=FALSE,main="Histogram of n=25",xlab="n=5",ylim = c(0, 0.4))
curve(dt(x, 24), -6, 6, add=T,lwd = 2, ylim = c(0, 0.4))


#Set the function
qvalue=function(x){
```

```
  q <- c(quantile(x,0.075),quantile(x,0.8),quantile(x,0.85),quantile(x,0.9),quanti
le(x,0.95),quantile(x,0.995))
  return(q)
}
q_t = function(n){
  q <- c(qt(0.075,n-1),qt(0.8,n-1),qt(0.85,n-1),qt(0.9,n-1),qt(0.95,n-1),qt(0.995,
n-1))
  return(q)
}

#Create the quantile table
q5 <- qvalue(T5)
q10 <- qvalue(T10)
q15 <- qvalue(T15)
q25 <- qvalue(T25)
q_t5 <- q_t(5)
q_t10 <- q_t(10)
q_t15 <- q_t(15)
q_t25 <- q_t(25)
q <-c('0.075','0.8','0.85','0.9','0.95','0.995')
q_table<-data.frame(q5,q_t5,q10,q_t10,q15,q_t15,q25,q_t25)
colnames(q_table)<-list('n=5','t4','n=10','t9','n=15','t14','n=25','t24')
print(q_table)
```