

MATH 203: PRINCIPLES OF STATISTICS 1
FALL 2024

WHERE DOES THE T-DISTRIBUTION COME FROM ?

If $X_1, \dots, X_n$ are a random sample from the $Normal(\mu, \sigma^2)$ distribution, we know that the sample mean random variable $\bar{X}$ also has a Normal distribution

$$\bar{X} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$$

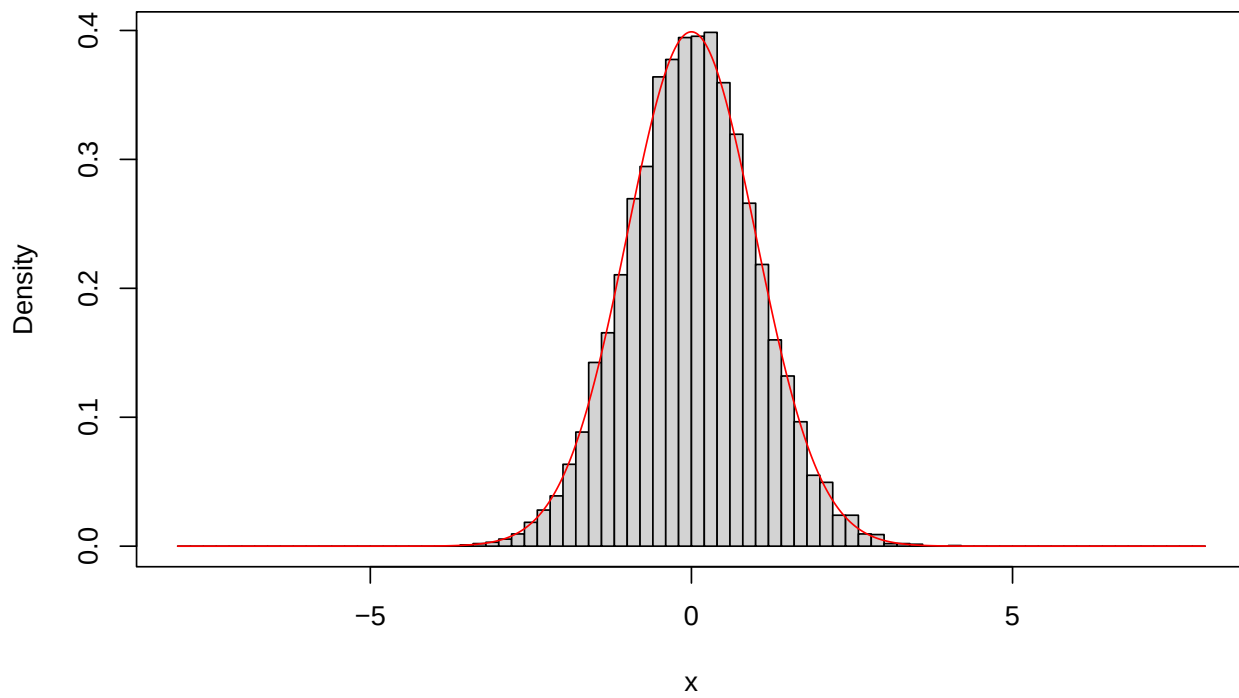
and that we perform the z-transform on $\bar{X}$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

then we can conclude that $Z \sim Normal(0, 1)$. We can demonstrate this result in simulation: we consider $N = 10000$ repeated experiments where we collect random samples of size $n = 8$ from the $Normal(10, 2^2)$ distribution, and form the collection of z statistics assuming the known values of μ and σ^2 . We inspect the histogram of the z statistics, and note that it does follow the $Normal(0, 1)$ distribution. The QQplot confirms Normality.

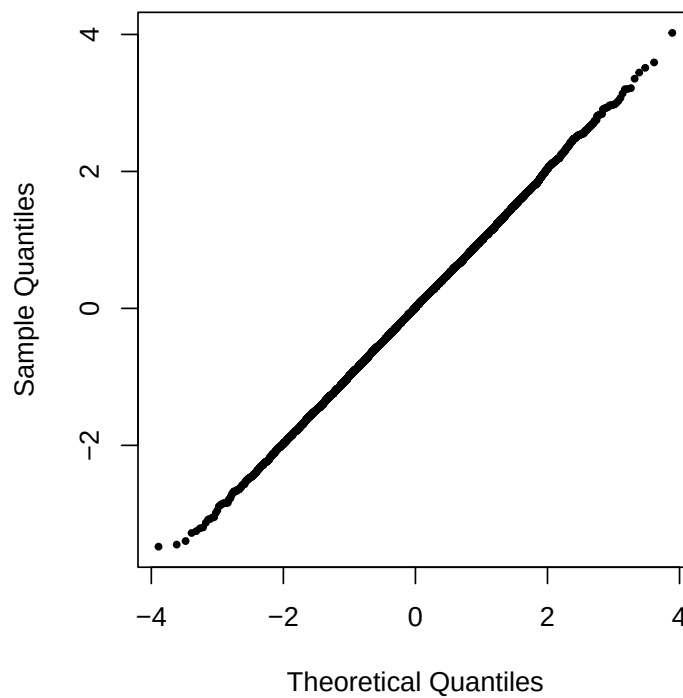
```
set.seed(34)
N<-10000
n<-8
mu<-10
sig<-2
X<-matrix(rnorm(N*n,mu,sig),nrow=N,ncol=n) #Generates the simulated data
xbar<-apply(X,1,mean) #Computes the replicated sample means
zval<-(xbar-mu)/(sig/sqrt(n)) #Computes the replicated z values
xv<-seq(-8,8,by=0.001)
yv1<-dnorm(xv)
par(mar=c(4,4,4,2))
hist(zval,br=seq(-8,8,by=0.2),freq=FALSE,
     main='z computed from 10000 replicates',xlab='x');box()
lines(xv,yv1,col='red')
```

z computed from 10000 replicates



```
par(pty='s')  
qqnorm(zval,pch=19,cex=0.5)
```

Normal Q-Q Plot

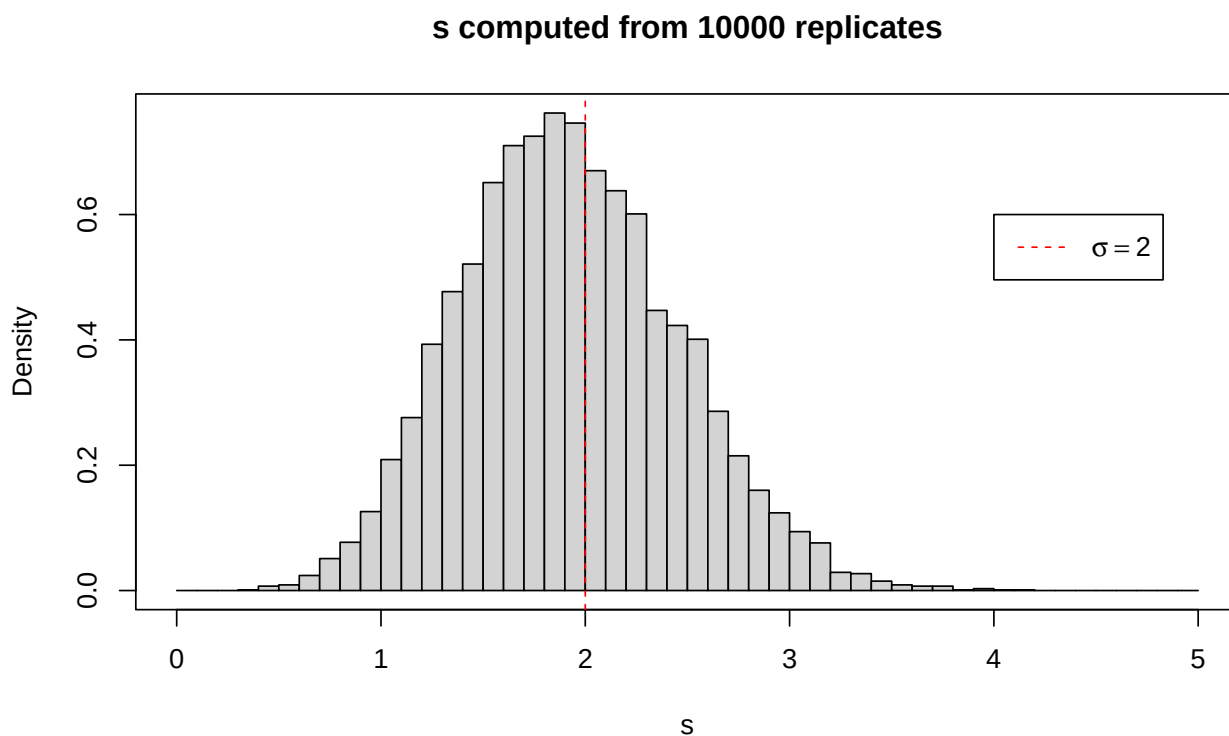


Now suppose that the simple change is made to use an estimate of σ based on the sample. We estimate this population parameter by the sample standard deviation s , where

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

Clearly s will not equal σ exactly, but we would hope that it was close to σ at least when the sample is large. Here $n = 8$, and if we compare the collection of s estimates from the $N = 10000$ replicates with $\sigma = 2$, we see that there is some variation.

```
sval<-apply(X,1,sd)      #Computes the replicated values of s
par(pty='m')
hist(sval,br=seq(0,5,by=0.1),freq=FALSE,
     main='s computed from 10000 replicates',xlab='s');box()
abline(v=sig,lty=2,col='red')
legend(4,0.6,c(expression(sigma==2)),col=c('red'),lty=2)
```



Now suppose instead of computing Z , we compute the observed version of the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

where S is the random variable version of s . We can compute mathematically that this change makes the distribution of T different from the Normal; in fact, we can show that

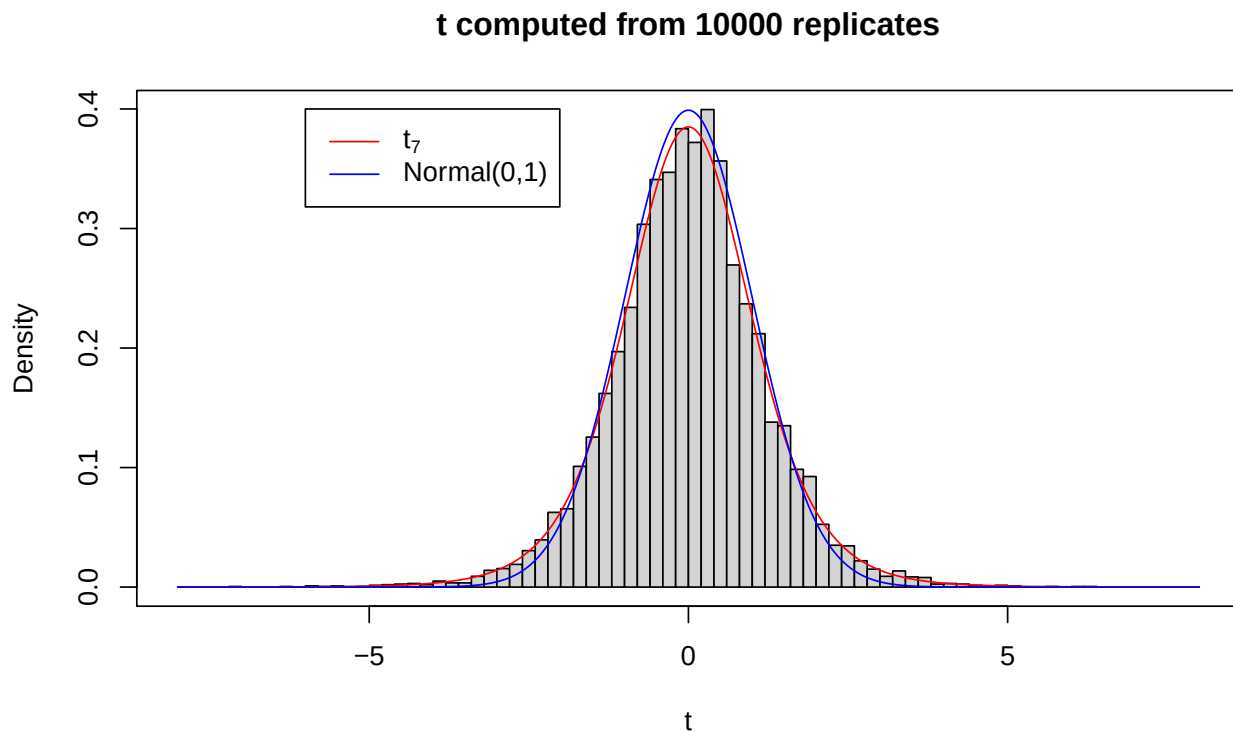
$$T \sim t_{n-1}$$

that is, the distribution is the t -distribution with $n - 1$ degrees of freedom.

```

tval<-(xbar-mu)/(sval/sqrt(n))    #Computes the replicated values of t
yv2<-dt(xv,n-1)
hist(tval,br=seq(-8,8,by=0.2),freq=FALSE,xlab='t',
     main='t computed from 10000 replicates');box()
lines(xv,yv2,col='red'); lines(xv,yv1,col='blue')
legend(-6,0.4,c(expression(t[7]), 'Normal(0,1)'),col=c('red', 'blue'),lty=1)

```



We see that the t -distribution here (specifically, the t_7 distribution), has heavier tails than the Normal. This implies that T has a slightly higher variance than Z , and we can verify that in the sampled values over the 10000 replicates.

```

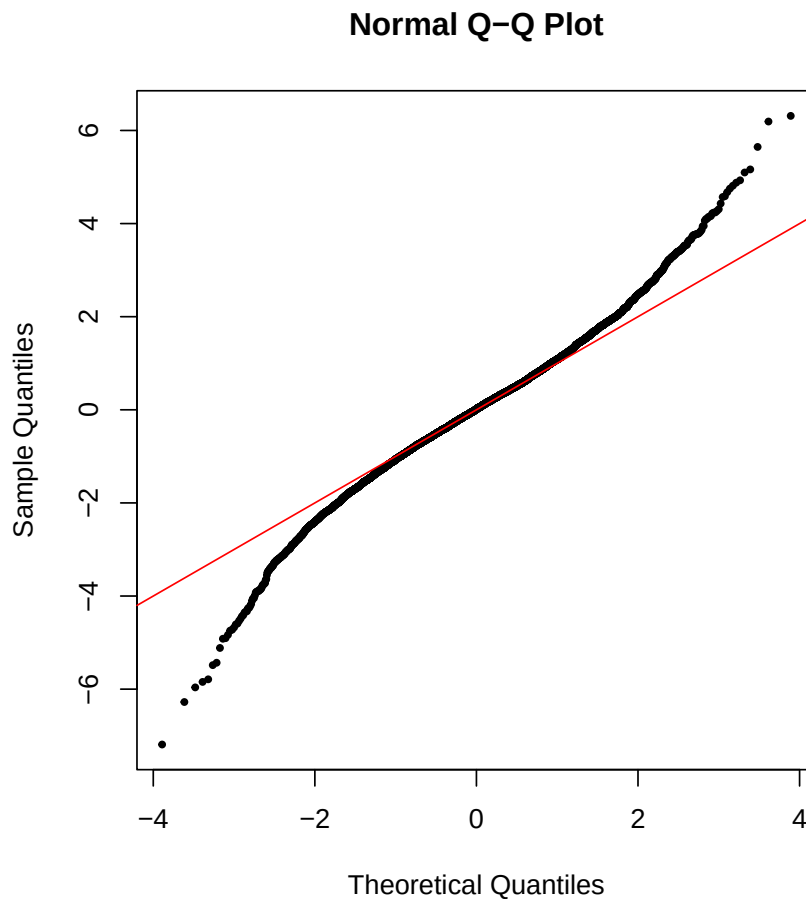
var(zval)    #Sample variance of Z
## [1] 0.9902716

var(tval)    #Sample variance of T
## [1] 1.401097

```

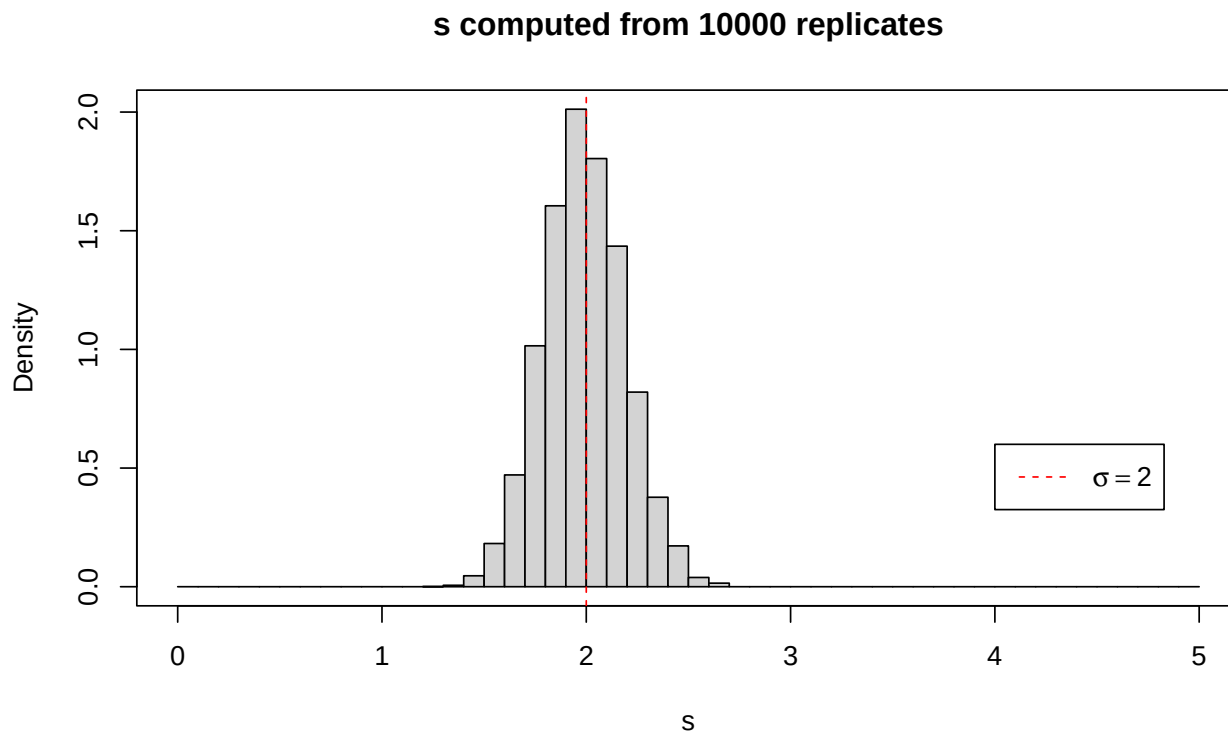
We can also use the QQplot to verify the non-Normal behaviour

```
par(pty='s')
qqnorm(tval,pch=19,cex=0.5)
abline(0,1,col='red')
```



We can repeat the same exercise, but with $n = 50$. Here we expect the distribution of the test statistic to be the t_{49} distribution, but we also know that this distribution is very close to the standard Normal.

```
set.seed(34)
N<-10000
n<-50
mu<-10
sig<-2
X<-matrix(rnorm(N*n,mu,sig),nrow=N,ncol=n) #Generates the simulated data
xbar<-apply(X,1,mean) #Computes the replicated sample means
zval<-(xbar-mu)/(sig/sqrt(n)) #Computes the replicated z values
sval<-apply(X,1,sd) #Computes the replicated values of s
par(pty='m')
hist(sval,br=seq(0,5,by=0.1),freq=FALSE,
     main='s computed from 10000 replicates',xlab='s');box()
abline(v=sig,lty=2,col='red')
legend(4,0.6,c(expression(sigma==2)),col=c('red'),lty=2)
```

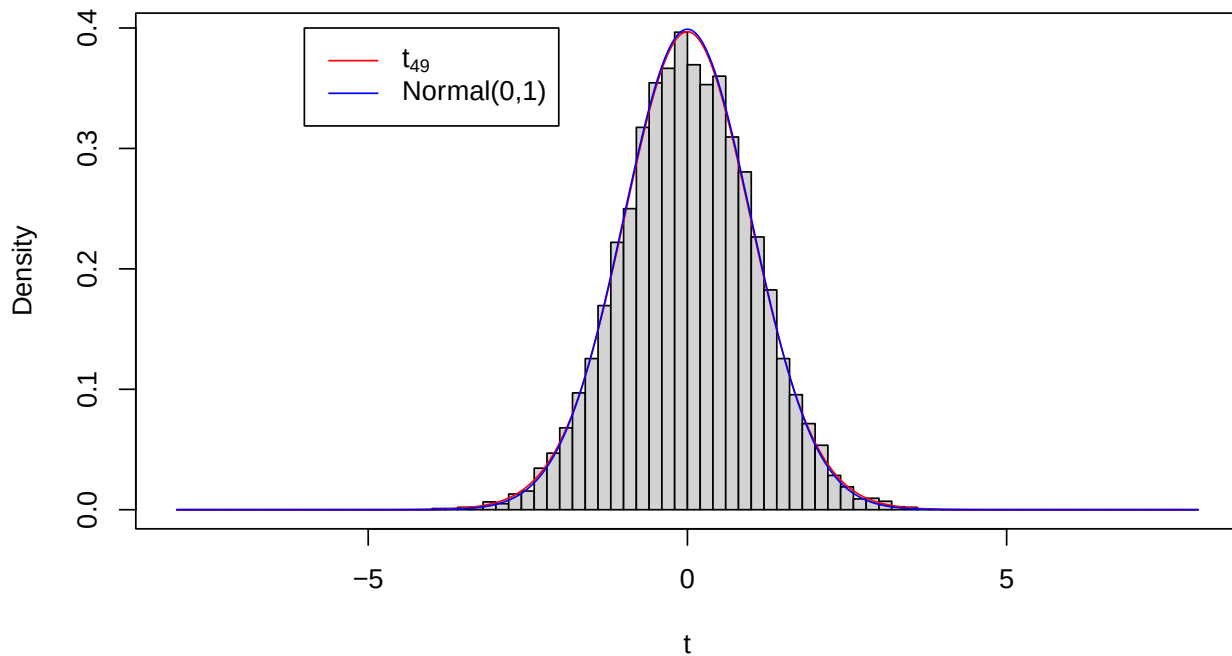


The distribution of the sample standard deviation is much less variable than when $n = 8$; if we summarize the samples of s , we observe that the mean value is 1.988, which is close to the true (data generating) value of σ .

We can now recompute the test statistic and inspect its distribution.

```
tval<-(xbar-mu)/(sval/sqrt(n))    #Computes the replicated values of t
yv2<-dt(xv,n-1)
hist(tval,br=seq(-8,8,by=0.2),freq=FALSE,xlab='t',
     main='t computed from 10000 replicates');box()
lines(xv,yv2,col='red'); lines(xv,yv1,col='blue')
legend(-6,0.4,c(expression(t[49]),'Normal(0,1)'),col=c('red','blue'),lty=1)
```

t computed from 10000 replicates



Here the distribution of T as represented by the sampled values is very close to the standard Normal. However, it is still slightly different: we can inspect the upper a points for the two distributions:

```
aval<-c(0.1,0.05,0.025,0.01,0.005,0.001)
z.a<-qnorm(1-aval)
t.a<-qt(1-aval,n-1)
a.pts<-rbind(z.a,t.a)
colnames(a.pts)<-aval
rownames(a.pts)<-c('Normal','t[49]')
a.pts
```

	0.1	0.05	0.025	0.01	0.005	0.001
Normal	1.281552	1.644854	1.959964	2.326348	2.575829	3.090232
t[49]	1.299069	1.676551	2.009575	2.404892	2.679952	3.265079

```
var(zval) #Sample variance of Z

## [1] 1.011316

var(tval) #Sample variance of T

## [1] 1.054232

par(pty='s')
qqnorm(tval,pch=19,cex=0.5)
abline(0,1,col='red')
```

Normal Q-Q Plot

