Z, Chi-Squared, T and Cauchy Distributions

North Carolina 2004 Births Dataset can be downloaded here: https://schs.dph.ncdhhs.gov/interactive/query/births/bd_2009andearlier.cfm

N, Z, χ^2 and T distributions

```
ncb<-read.csv('./Data/NCBirths2004.csv')
head(ncb)
```

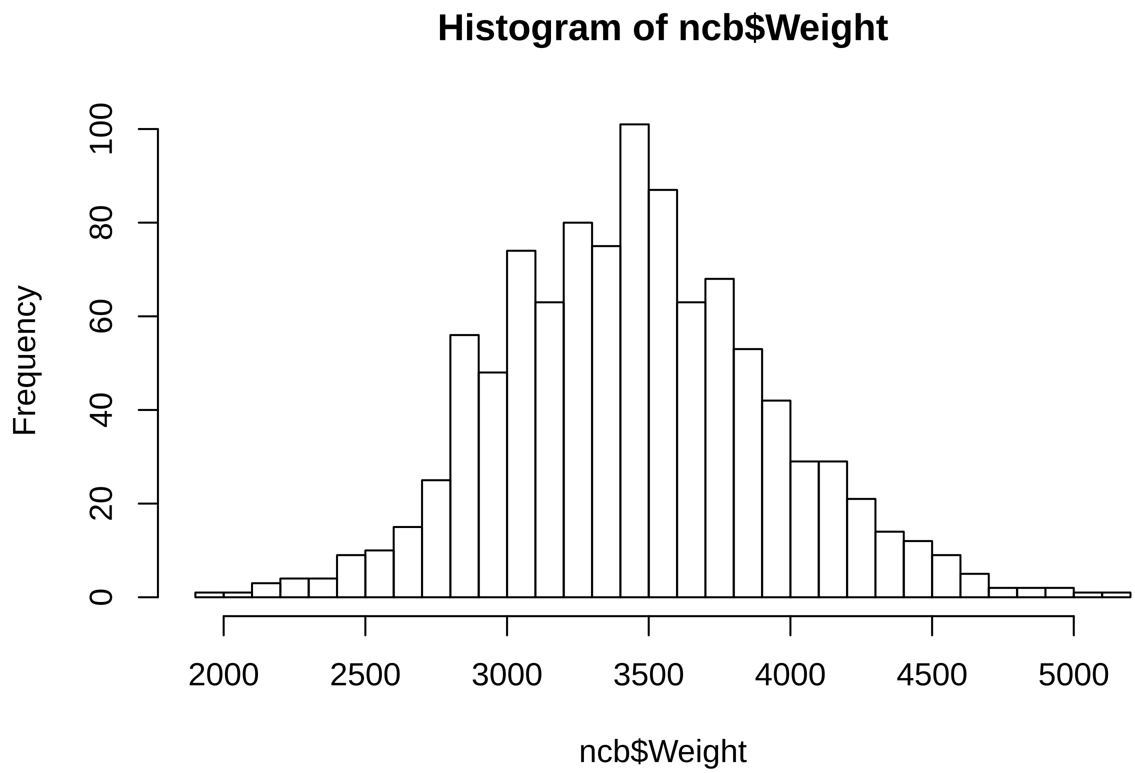
```
##   ID MothersAge Tobacco Alcohol Gender Weight Gestation
## 1  1      30-34      No      No   Male   3827        40
## 2  2      30-34      No      No   Male   3629        38
## 3  3      35-39      No      No Female   3062        37
## 4  4      20-24      No      No Female   3430        39
## 5  5      25-29      No      No   Male   3827        38
## 6  6      35-39      No      No Female   3119        39
```

```
str(ncb)
```

```
## 'data.frame':    1009 obs. of  7 variables:
##  $ ID        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MothersAge: Factor w/ 8 levels "15-19","20-24",..: 4 4 5 2 3 5 2 2 2 3 ...
##  $ Tobacco   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Alcohol   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Gender    : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 1 1 2 2 1 ...
##  $ Weight    : int  3827 3629 3062 3430 3827 3119 3260 3969 3175 3005 ...
##  $ Gestation : int  40 38 37 39 38 39 40 40 39 39 ...
```

I will use birth weight. The distribution is approximately normal with population mean 3448 grams and standard deviation 488 grams.

```
hist(ncb$Weight,breaks='FD')
```




```r
mu<-mean(ncb$Weight)
sigma<-sd(ncb$Weight)
mu;sigma
```

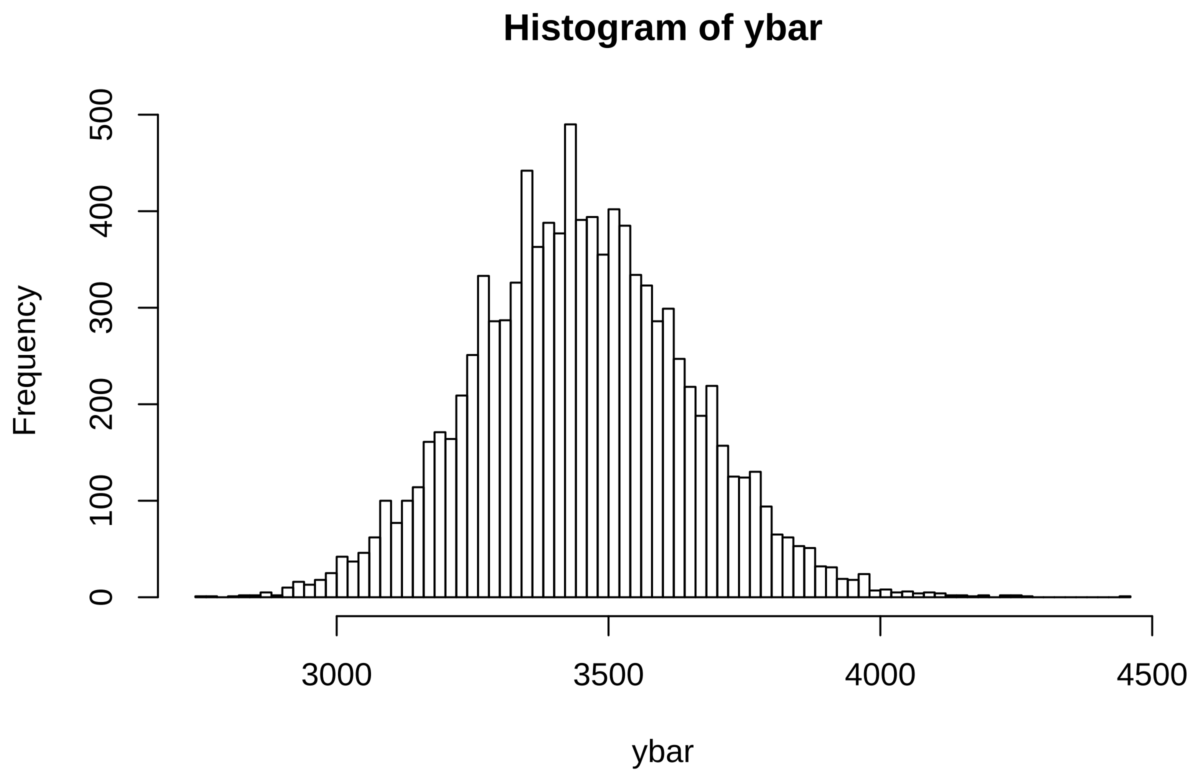
```
## [1] 3448.26
```

```
## [1] 487.736
```

Let's say that I want to know the average weight of North Carolina births in 2004 but I can only see a sample of 6. While I won't know the true mean I may be able to get a good idea of how close I'll come because of the central limit theorem. If I take the average of a sample of 6 observations many times I get an approximately normal distribution. The average of n i.i.d. random variables X_i converges to a Normal distribution as n goes to infinity.

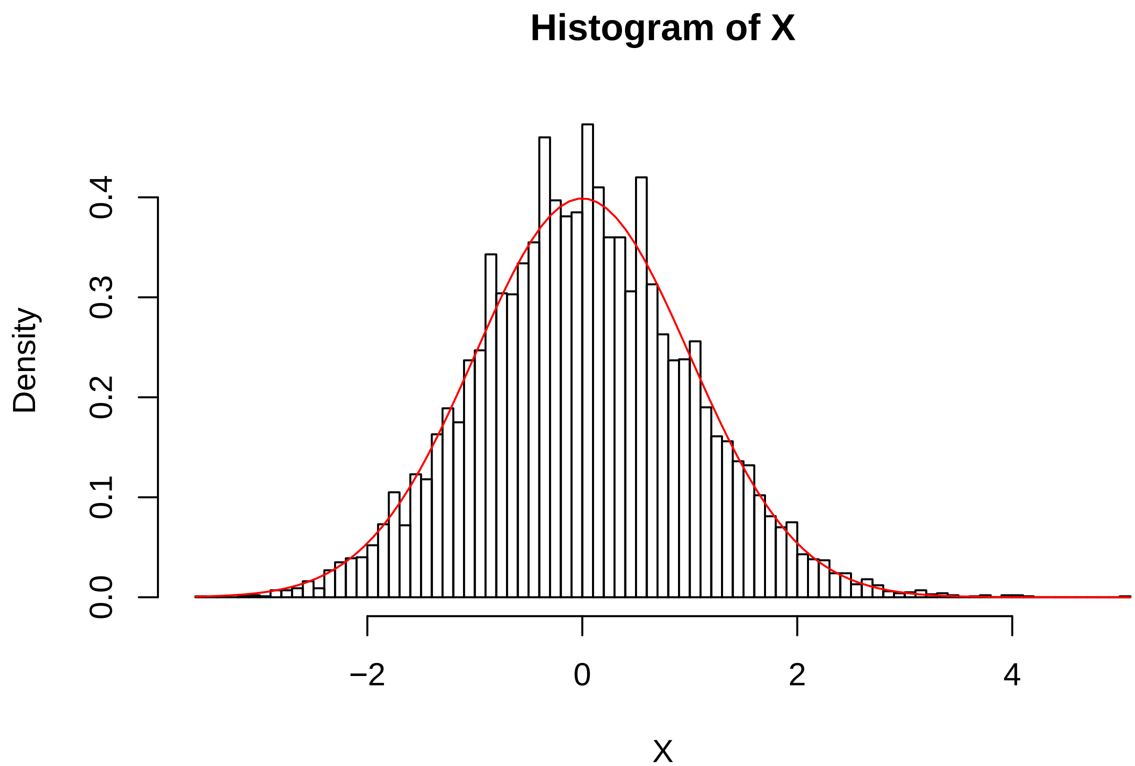
$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} = N(\mu_X, \sigma_X/\sqrt{n})$$

```r
n<-6; N <- 10000; ybar <- numeric(N)
for(i in 1:N){
  y <- sample(ncb$Weight, n)    #sample without replacement
  ybar[i] <- mean(y)
}
hist(ybar,breaks = "FD")
```



And, of course, standardizing each result gets very close to a Standard normal distribution.

```
X<-(ybar-mu)/(sigma/sqrt(n))
hist(X,freq=FALSE,breaks="FD")
curve(dnorm(x),add = TRUE, col = "red")
```



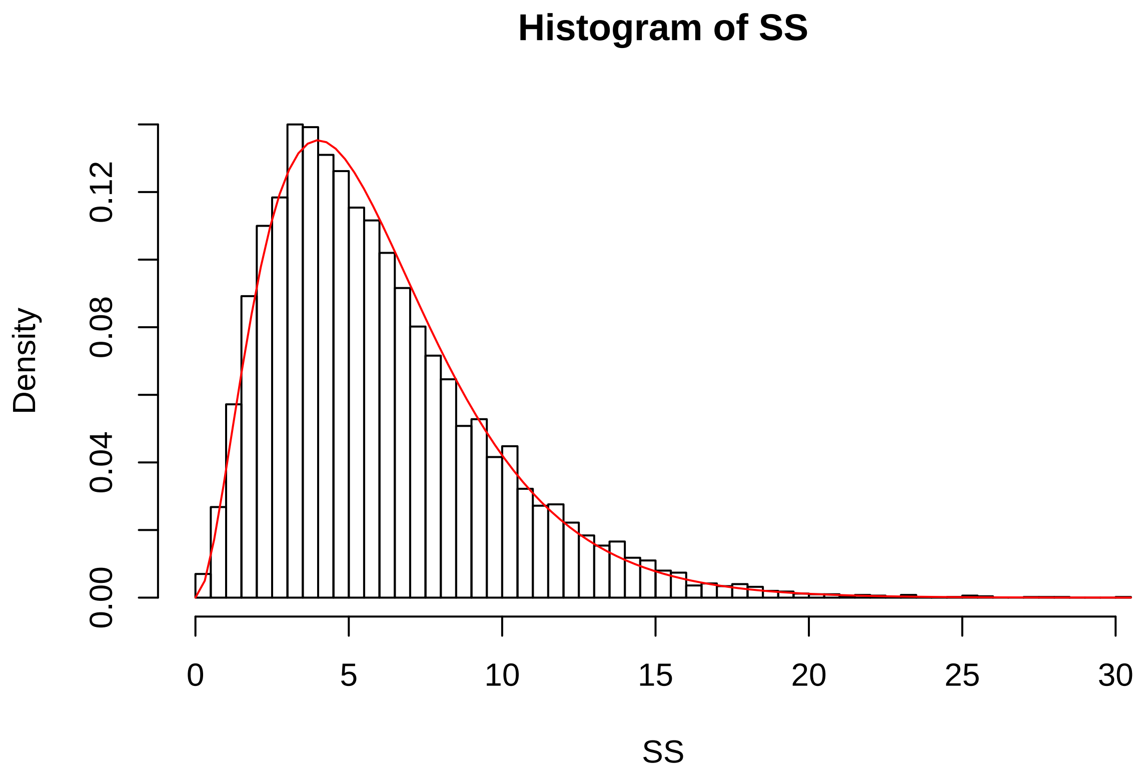

```
mean(X);sd(X)
```

```
## [1] 0.01049366
```

```
## [1] 0.9964116
```

Furthermore, the sum of the squares of the standardized samples have an approximately χ^2 distribution with df=6 for the 6 observations in each sample. $\sum_{i=1}^{n} Z^2 = \chi^2(n)$

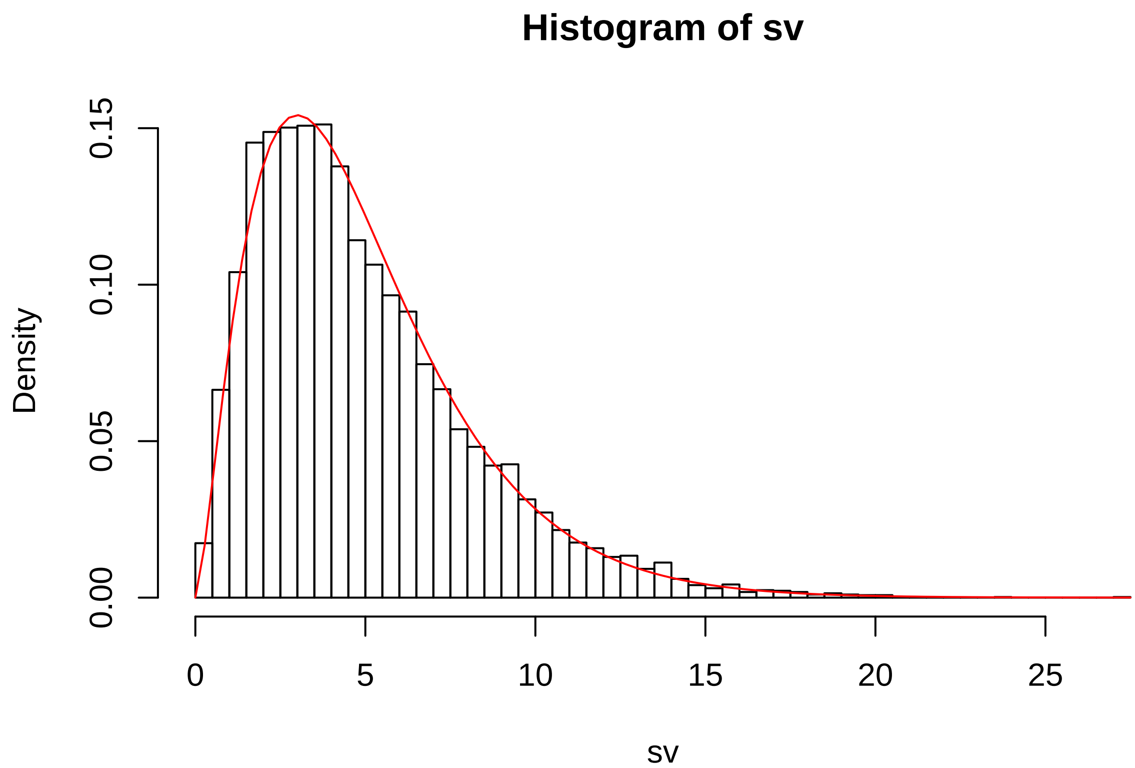
```
SS<-numeric(N)
for(i in 1:N){
  y <- sample(ncb$Weight, n)    #sample without replacement
  SS[i]<-sum(((y-mu)/sigma)^2)
}
hist(SS,freq=FALSE,breaks="FD")
curve(dchisq(x,n),add=TRUE, col="red")
```



However, I still have a problem in that I only see a sample of 6 weights and so even though I know that my observed average falls within a guassian sampling distribution, I do not know the true variance of the population, so my estimate of the variance of the sampling distribuiton is subject to error. I can get around this problem by finding a sampling distribution which doesn't require the population variance.

To get there I'll first model the sampling distribution of sample variances. The χ^2 distribution is the sum of squares of n i.i.d. standard normal random variables, so the distribution of sample variances multipled by $(n-1)$ and standardizd by dividing by σ^2, has an approximiately χ^2 distribution with df=n-1=5.

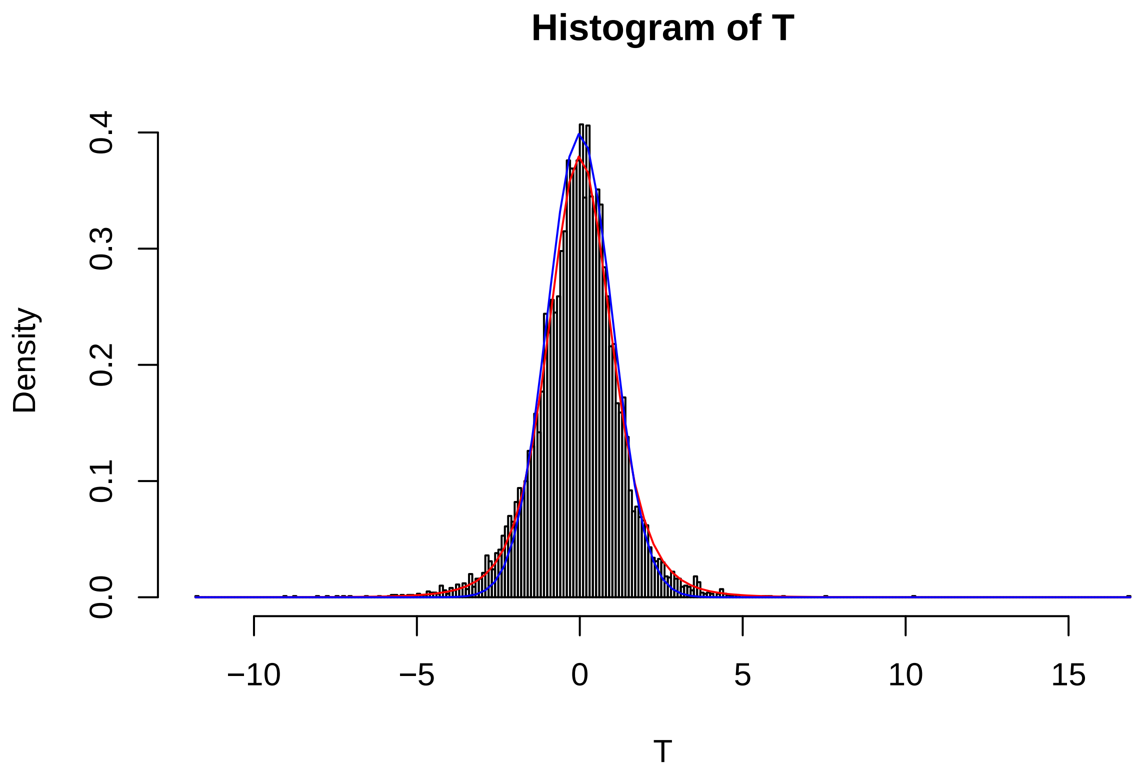
```
sv <- numeric(N)
for (i in 1:N) {
  y <- sample(ncb$Weight, n)
  sv[i] <- (n-1)*var(y)/(sigma^2)
}
hist(sv, freq=FALSE, breaks = "FD")
curve(dchisq(x,n-1), add = TRUE, col = "red")
```



The T distribution is a normal distribution divided by a χ^2 distribution. so I can studentize the sample mean and the resulting random variable has an approximately Student T distribuiton with df=n-1=5. Because we divide the two, population variance is not necessary.

$$\frac{Z}{\sqrt{\chi^2/n}}$$

```
T <- numeric(N)
for (i in 1:N) {
  y <- sample(ncb$Weight, n)
  T[i] <- (mean(y)-mu)/sqrt((var(y)/n))
}
hist(T, freq=FALSE, breaks = "FD")
curve(dt(x,n-1), add = TRUE,col = "red")
#compare to normal:
curve(dnorm(x),add=TRUE, col="blue")
```



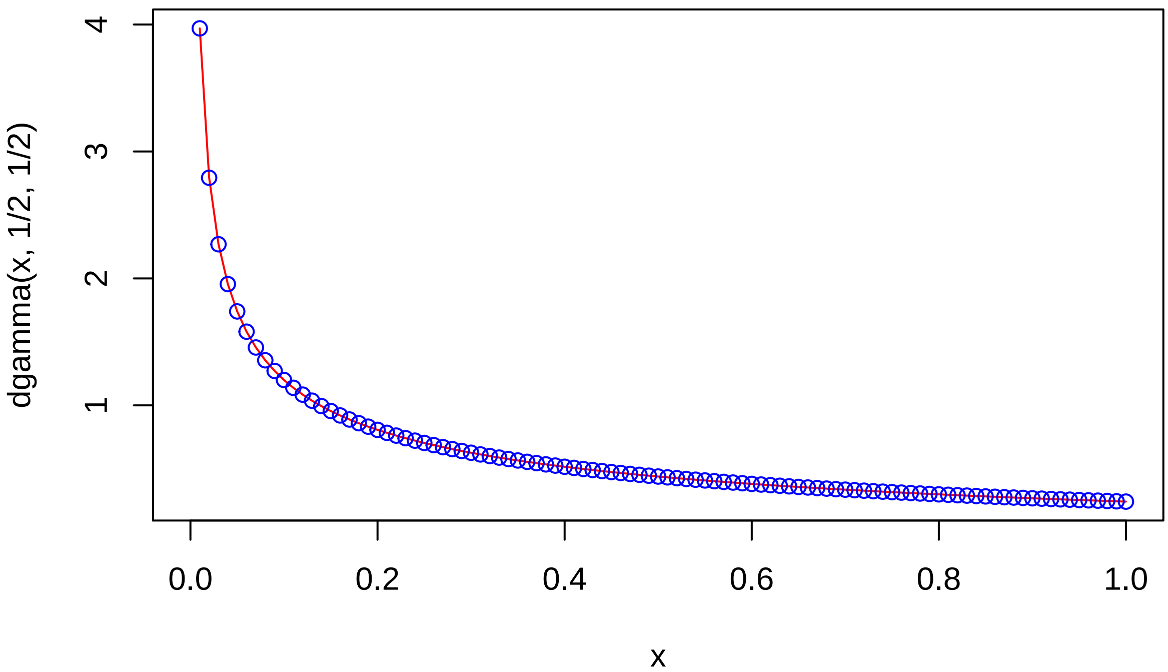

```
#close to t dist. Better fit than St Normal
```

Using the T distribution allows me to make inferences about my sample average compared to the population average, without knowing the population variance. Now I can build a confidence interval around my 6 sample weights. The T distribution has larger tails which makes sense given our uncertainity about the population variance. As n gets larger, it approaches a Normal distribution.

The T Distribution Can Have Undefined Variance

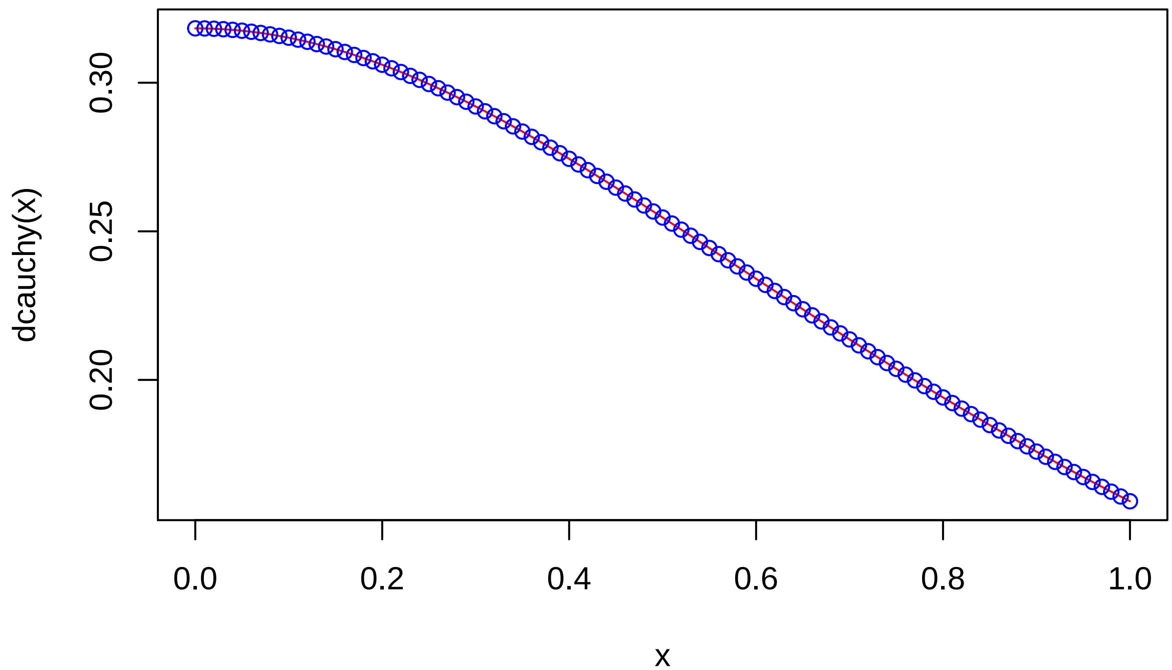
When we only have one observation in our sample, the $\chi^2(1)$ distribution is the same as a Gamma(1/2,1/2).

```
curve(dgamma(x,1/2,1/2), col = "red")
curve(dchisq(x,1), add=TRUE, col = "blue",type="p")
```



A T distribution with df=1 becomes $(Z/\sqrt{Z^2/1}) = Z/Z$ which is a Cauchy distribution. A Cauchy distribution is an example of a distribution with undefined variance, which means that the Central Limit Theorem doesn't hold.

```r
curve(dcauchy(x), col = "red")
curve(dt(x,1), add=TRUE, col = "blue",type="p")
```



So even though the area under the curve is finite,

```
f<-function(x) dt(x,1)
integrate(f,-Inf,Inf)
```

```
## 1 with absolute error < 1.6e-10
```

attempting to calculate the variance leads to "'probably divergent".

```
# These are commented out because knitR fails with integrate() error
# f<-function(x) x^2*dt(x,1)
# integrate(f,-Inf,Inf)
# # At df=1 this is Cauchy distribution which has undefinted variance and no moment generating function
# f<-function(x) (1/pi)*(x^2)*(1/(1+x^2))
# integrate(f,-Inf,Inf)
```

The mean is also undefined. Although integrate() tells me it's 0, this is clearly not true.

```
f<-function(x) x*dt(x,1)
integrate(f,-Inf,Inf)
```

```
## 0 with absolute error < 0
```

```
#using large but unequal limits of integration does not give Expectations clsoe to 0
integrate(f,-100,200) #.221
```

```
## 0.2206237 with absolute error < 1.6e-07
```

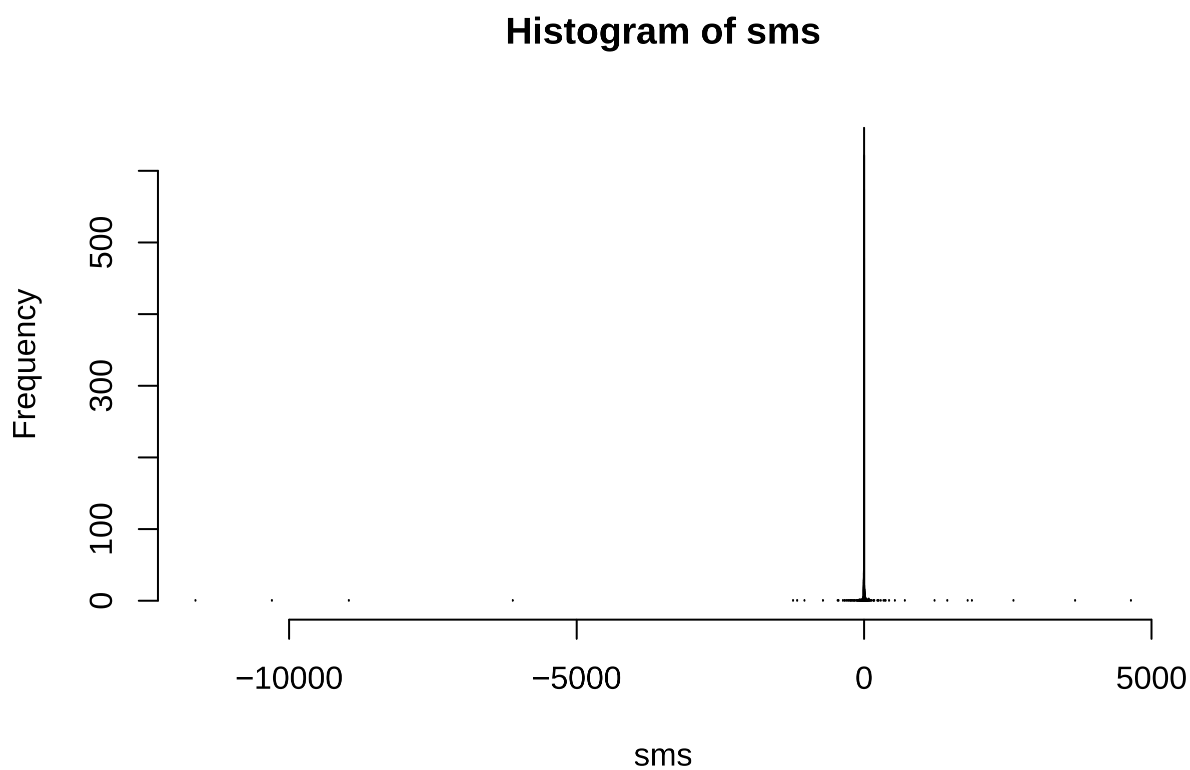
```
integrate(f,-9000,500) #-.92
```

```
## -0.9200333 with absolute error < 3.2e-06
```

```r
integrate(f,-9000,777777) # 1.419
```

```
## 1.419412 with absolute error < 3.1e-06
```

```r
#Histogram of sample means. Although 0 is most common, there are large outliters. The sample means are
N<-10000;n<-1000; sms<-numeric(N)
for (i in 1:N) {
  samps<-rt(n,1)
  sms[i]<-mean(samps)
}
hist(sms,breaks = "FD")
```



A T distribution with 2 degrees of freedom also has divergent variance, although its expecation is defined.

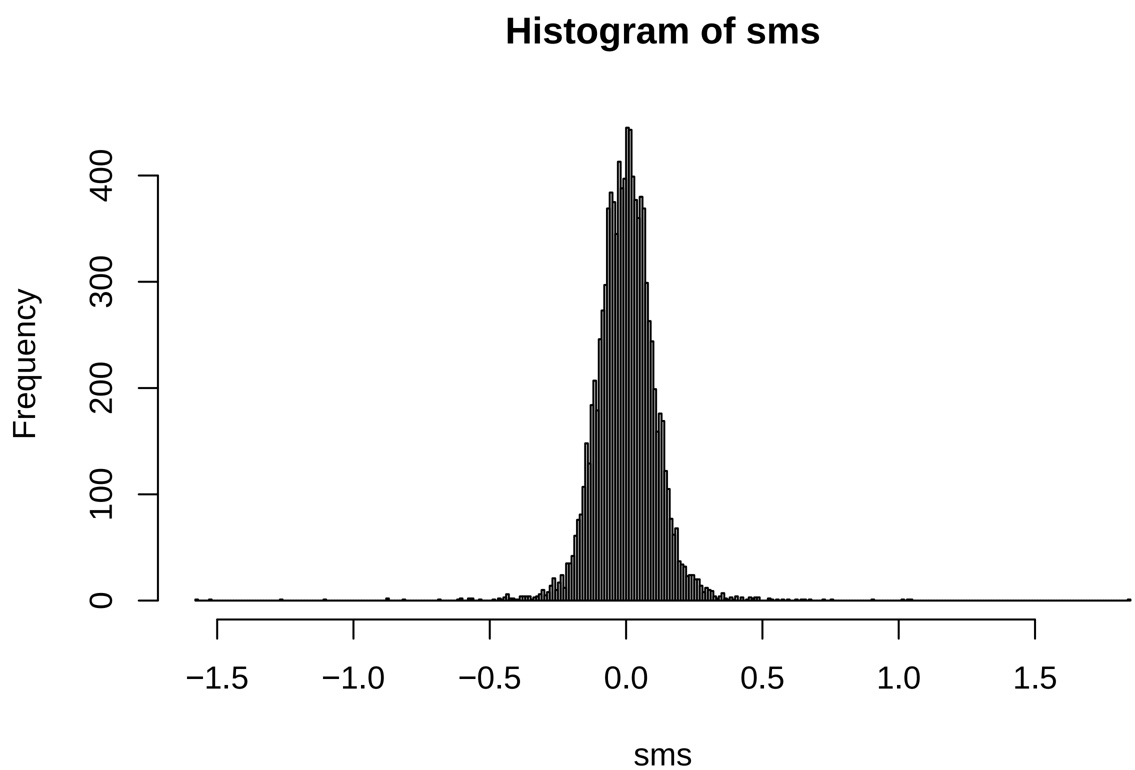
```r
#Commented out because knitR fails with integrate() error

#Variance calculation fails
# f<-function(x) x^2 *dt(x,2)
# integrate(f,-Inf,Inf) #Error: max number of subdivisions reached
# # this integral is divergent
# f<-function(x) x^2 * (1/((2+x^2)^(3/2)))
# integrate(f,-Inf,Inf)

#expectation
f<-function(x) x*dt(x,2)
integrate(f,0,Inf) # .7071
```

```
## 0.7071068 with absolute error < 1.1e-09
```

```r
integrate(f,-Inf,0)# -.7071
```

```
## -0.7071068 with absolute error < 1.1e-09
```

```r
#Histogram of sample means - these sample means are reliably close to 0
N<-10000;n<-1000; sms<-numeric(N)
for (i in 1:N) {
  samps<-rt(n,2)
  sms[i]<-mean(samps)
}
hist(sms,breaks = "FD")
```




```r
max(sms);min(sms)
```

```
## [1] 1.847588
```

```
## [1] -1.573718
```

At df=3 and higher, both expectation and variance are defined.