Set_3.Rmd

Alexander Williams

2024-09-20

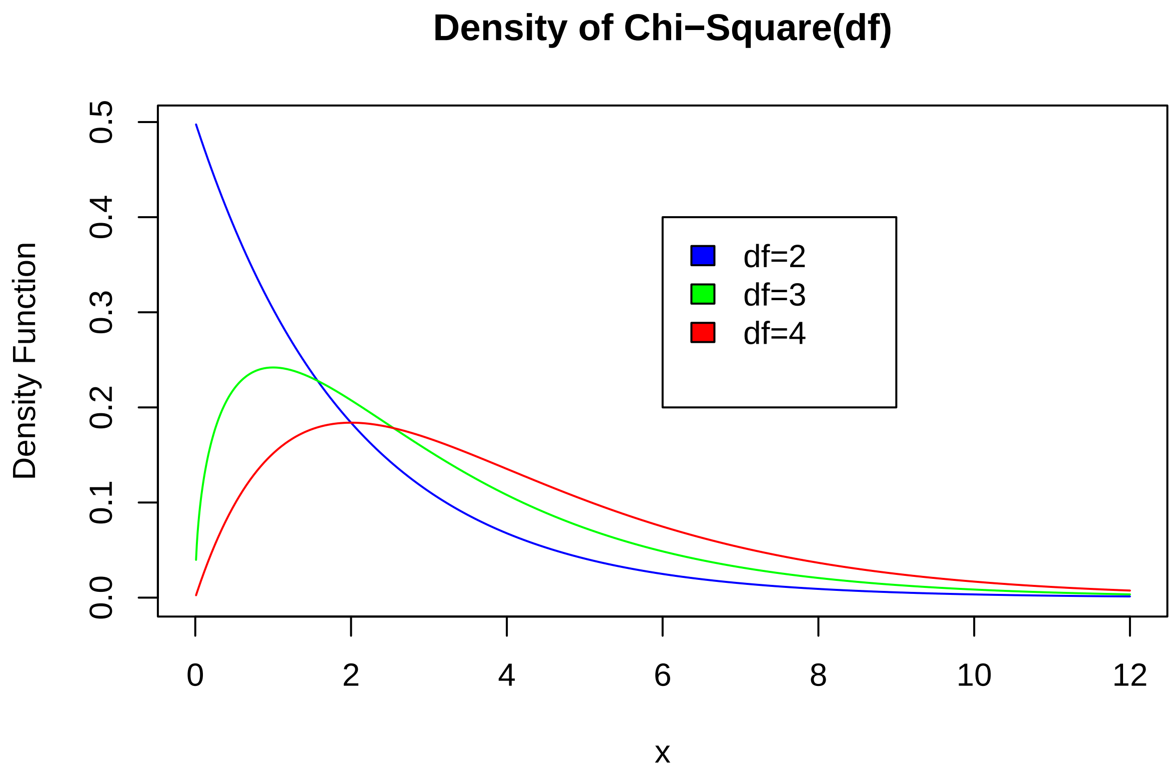
Review of Probability Distributions

Chi-square Distribution

(pronounced kai)

if $Z \sim N(0,1)$ we say the random variable defined by $X = Z^2$ is $\chi^2_{(1)}$.

```r
## make several plots of the Chi-Square density function
x<-seq(0.01,12,0.01)
y.df2<-dchisq(x,df=2) # df = degrees of freedom
y.df3<-dchisq(x,df=3)
y.df4<-dchisq(x,df=4)
plot(c(0,12),
     c(0,max(y.df2,y.df3,y.df4)),
     type='n', #don't plot the points, plot line
     ylab='Density Function',
     xlab='x')
title('Density of Chi-Square(df)') # name
lines(x,y.df2,col='blue')
lines(x,y.df3,col='green')
lines(x,y.df4,col='red')
legend(x=c(6,9), # where you want the legend, what you want in it
       y=c(0.4,.2),
       legend=c('df=2','df=3','df=4'),
       fill=c('blue','green','red'))
```



if X is chi-squared with 4 degrees of freedom, $X \sim \chi^2_{(4)}$ compute P(X>=4). If Y is chi-squared with 3 degrees of freedom, $Y \sim \chi^2_{(3)}$ compute P(Y>=4)

```
# pchisq computes P(X<=q), so 1 - pchisq gets >= probability
x.prob<- 1 - pchisq(q=4, df=4)
x.prob
```

```
## [1] 0.4060058
```

```
y.prob<- 1 - pchisq(q=4, df=3)
y.prob
```

```
## [1] 0.2614641
```

if X and Y are independent, how do we compute $P(X + Y \geq 4)$? since X has 4 degrees, Y has 3 degrees, we add them together to get 7 degrees

```
1 - pchisq(q=4, df=7)
```

```
## [1] 0.7797774
```

if $X \sim \chi^2_{(4)}$, computer median and 0.7 quantile of the distribution

the median is the 0.5 quantile, so $P(X \leq q_{0.5})$. Quantile means "Find the value on the curve such that the area to the left of this point $= q_p$

```
# qchisq gives the quantile, given percentage p
q5 <- qchisq(p=0.5, df=4)
q5
```

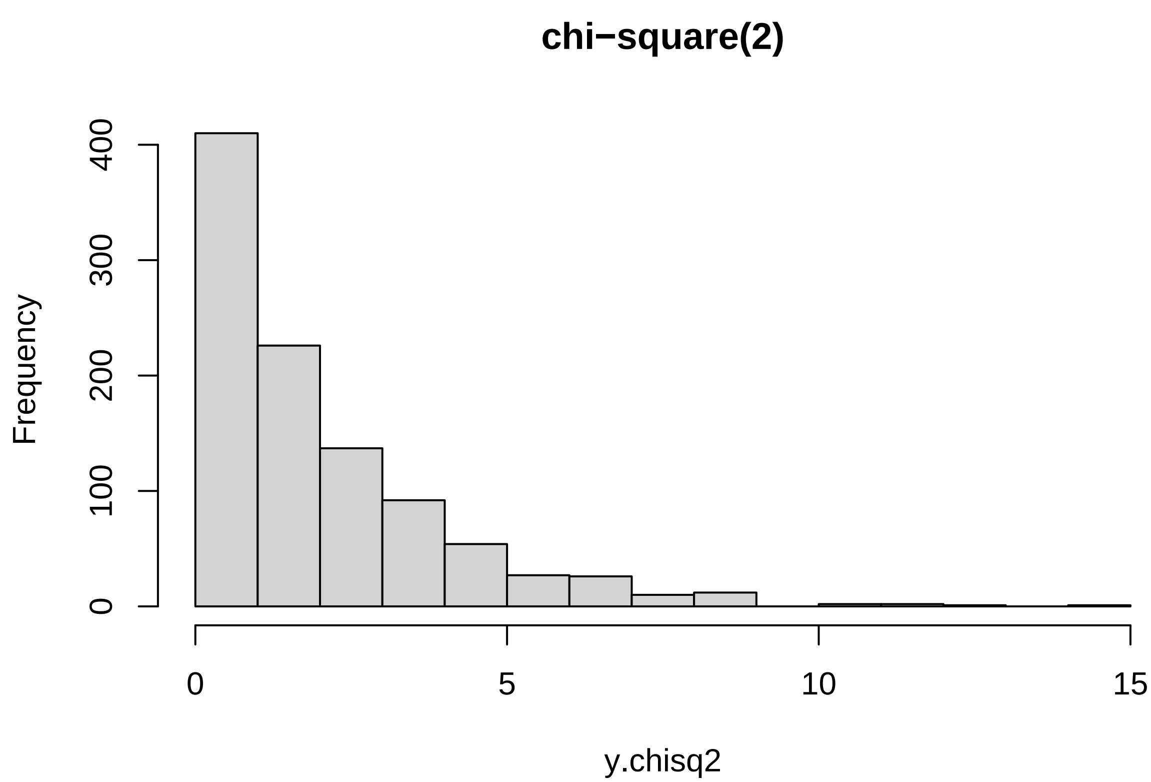
```
## [1] 3.356694
```

```
# do the same thing for the 0.7 quantile
q7 <- qchisq(p=0.7, df=4)
q7
```

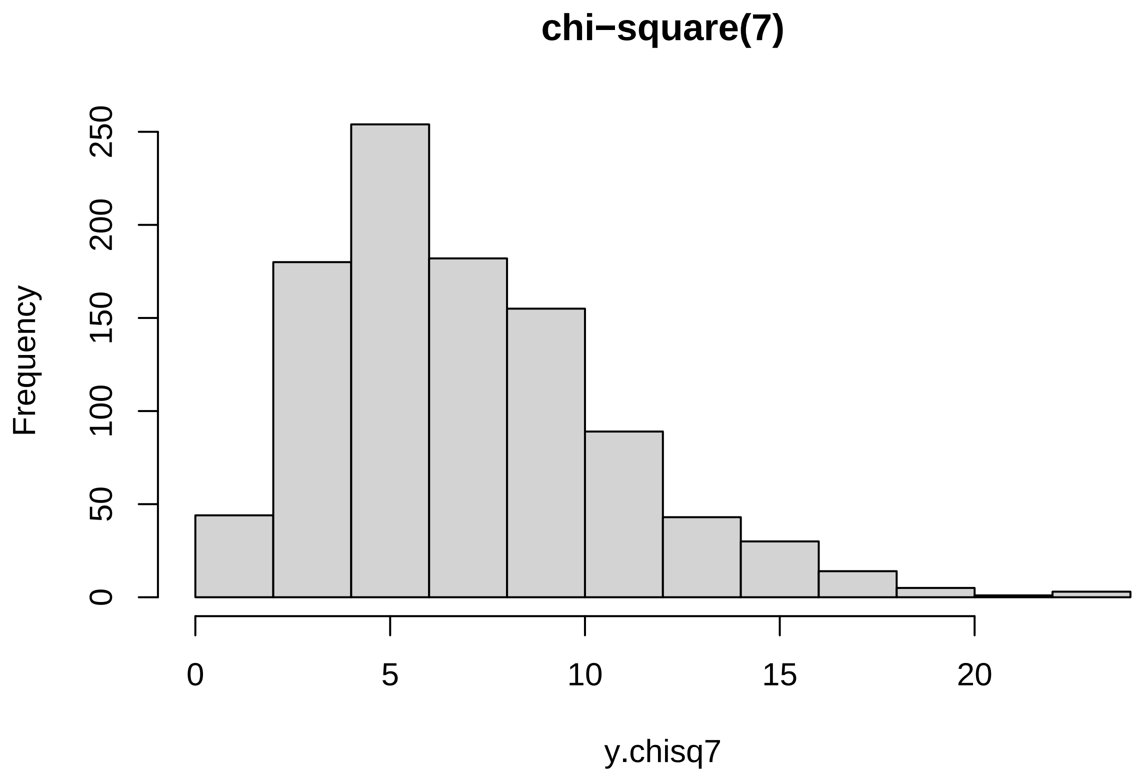
```
## [1] 4.878433
```

We can simulate this distribution onin R

```
y.chisq2<-rchisq(n=1000,df=2)
y.chisq7<-rchisq(n=1000,df=7)
hist(y.chisq2,main="chi-square(2)")
```




```
hist(y.chisq7,main="chi-square(7)")
```



t-distribution

used in small data sets to approximate normal distributions if $Z \sim N(0,1)$ and $W \sim \chi^2_{(n)}$, Z and W are assumed independent, then a t_n distribution is defined as:

$$X = \frac{Z}{\sqrt{W/n}}$$

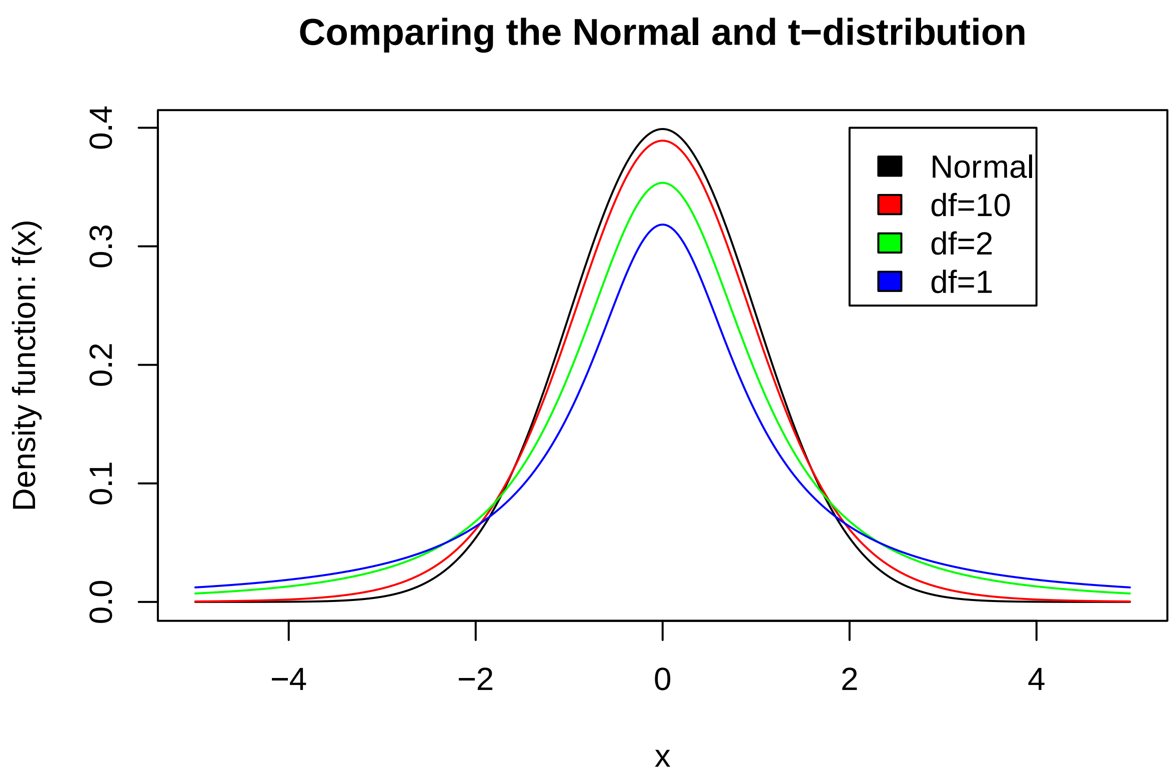
We can plot t-distribution with different degrees of freedom

```
# create a plot comparing the densities of the normal, t and chi-square distributions
x<-seq(-5,5,.01)
y.norm<-dnorm(x,mean=0,sd=1)
y.t1<-dt(x,df=1)
y.t2<-dt(x,df=2)
y.t10<-dt(x,df=10)

# set up the plot area
plot(c(min(x),max(x)),
    c(min(y.norm,y.t1,y.t2,y.t10),
    max(y.norm,y.t1,y.t2,y.t10)),
    type="n",
    xlab="x",
    ylab="Density function: f(x)")

title("Comparing the Normal and t-distribution")
```

```r
lines(x,y.norm)
lines(x,y.t10,col="red")
lines(x,y.t2,col="green")
lines(x,y.t1,col="blue")
legend(x=c(2,4),
       y=c(.4,.25),
       legend=c('Normal','df=10','df=2','df=1'),
       fill=c('black','red','green','blue'))
```



you can see that it approaches normal as degrees of freedom increase the case where the degrees of freedom=1 is special as the tail decays so slowly, even the mean doesn't exist. It's useful for finding counter examples

if X is t-dist with 3 degrees of freedom, compute P(X>=4) if Y is t-dist with 10 degrees of freedom, compute P(Y>=4)

```r
x.prob<- 1 - pt(q=4, df=3)
x.prob
```

```
## [1] 0.01400423
```

```r
y.prob<- 1 - pt(q=4, df=10)
y.prob
```

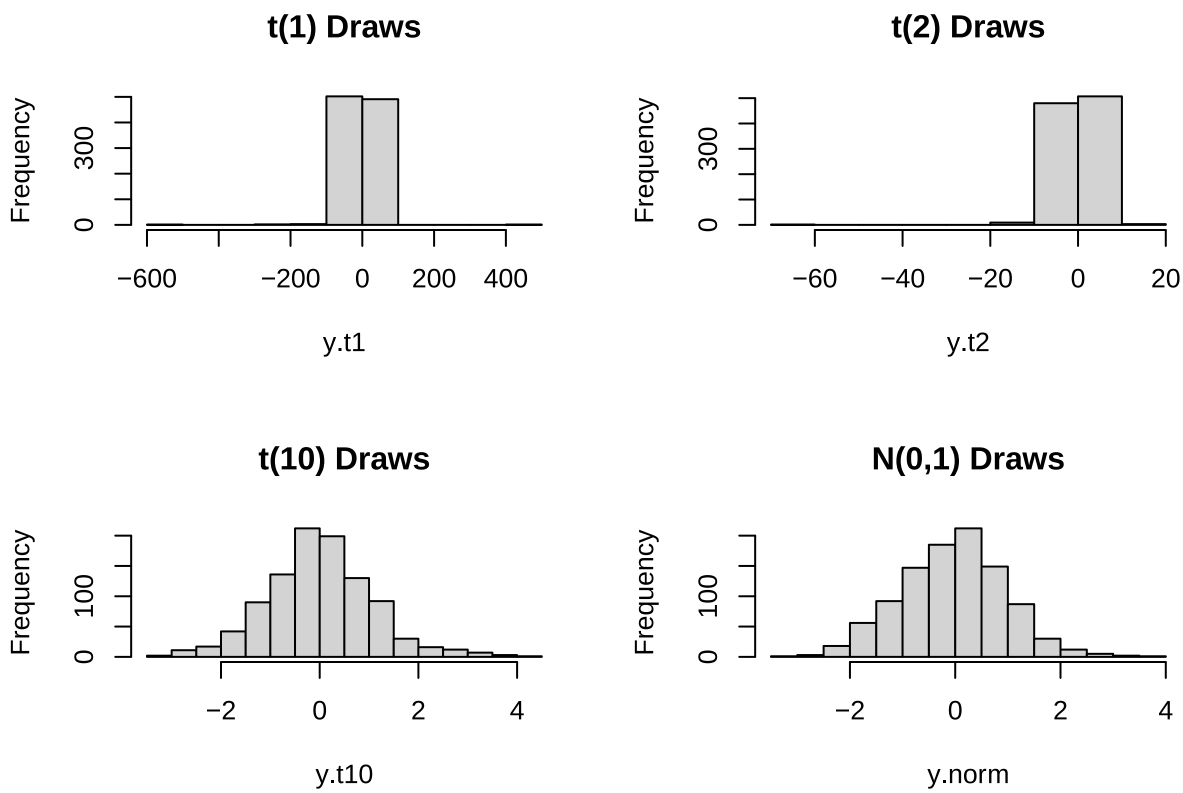
```
## [1] 0.001259166
```

lets simulate the t_n distribution

```r
y.t1<-rt(n=1000,df=1)
y.t2<-rt(n=1000,df=2)
```

```r
y.t10<-rt(n=1000,df=10)
y.norm<-rnorm(n=1000,mean=0,sd=1)


par(mfrow=c(2,2))
hist(y.t1,main='t(1) Draws')
hist(y.t2,main='t(2) Draws')
hist(y.t10,main='t(10) Draws')
hist(y.norm,main='N(0,1) Draws')
```



Poisson distribution

given a number of discreet events over a period of time/space, what are the odds of X events in some time?

useful for counts, things like number of infections in a tree

$X \sim Poisson(\lambda)$, where λ is number of events that occurred in an amount of time

this distribution has $E[X] = Var[X] = \lambda$, and uses the `pois` term in R. `ppois(), qpois(), rpois()...`

Binomial Distribution

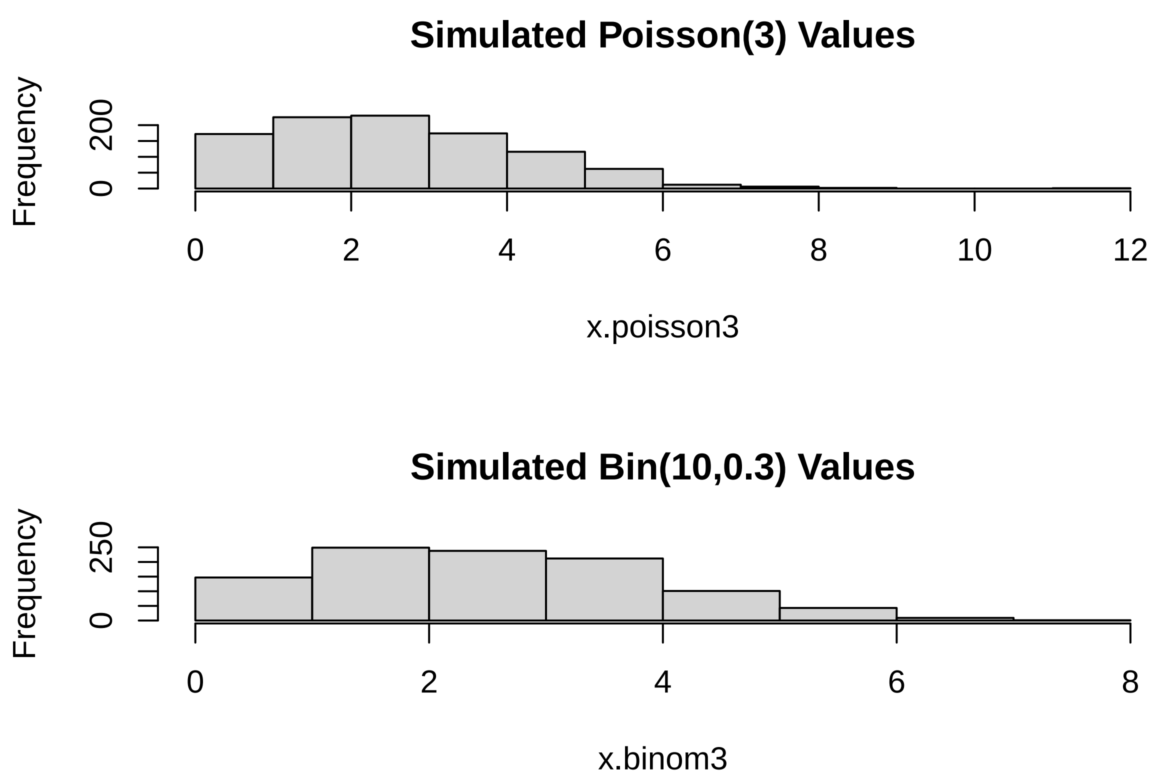
good old discrete distribution, given n events with probability p of happening.

$X \sim Bin(n, p)$

$E[X] = np$, and $Var[X] = np(1 - p)$.

in R, uses binom, pbinom(), qbinom(), rbinom()

```
# simulate 1000 values from a Poisson distribution with E[X] = 3
x.poisson3<- rpois(n=1000, lambda=3)
# simulate 1000 values from a binomial distribution with E[X] =
x.binom3<-rbinom(n=1000,size=10,prob=0.3)
par(mfrow=c(2,1))
hist(x.poisson3, main='Simulated Poisson(3) Values')
hist(x.binom3, main='Simulated Bin(10,0.3) Values')
```



you can see that poisson and binomial distributions can be quite similar, but it's important to be able to determine which distribution a dataset might be.

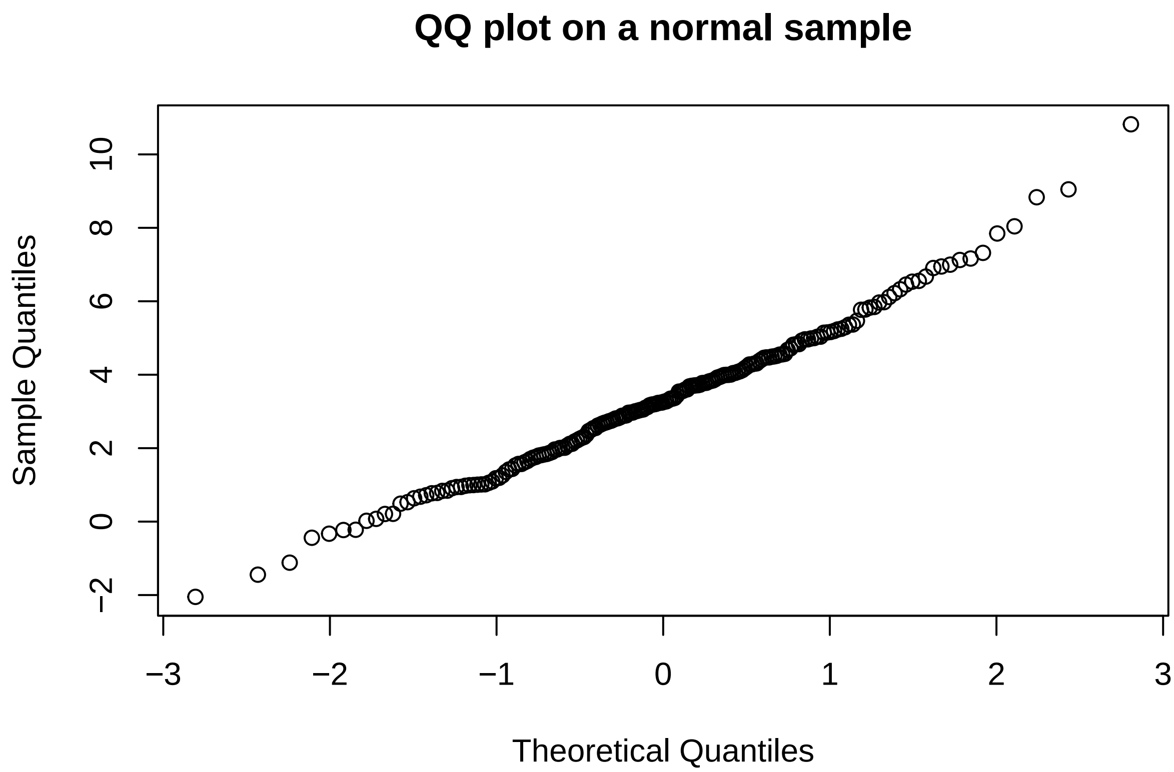
Quantile-Quantile plots

graphical method to determine if data comes from a particular distribution. Questions like "do you think this is a normal distribution?"

1. sort data $y_1, ..., y_n$ in ascending data, leading to so-called order-statistics $y_{(1)}, ..., y_{(n)}$
2. consider theoretical distribution of interest and consider a hypothetical sample $X_1, ..., X_n$, and it's order statistics $X_{(1)}, ..., X_{(n)}$
3. compare the sampled order statistics against the expected order statistics $E[X_{(1)}], ..., E[X(n)]$

A good fit results in a linear plot

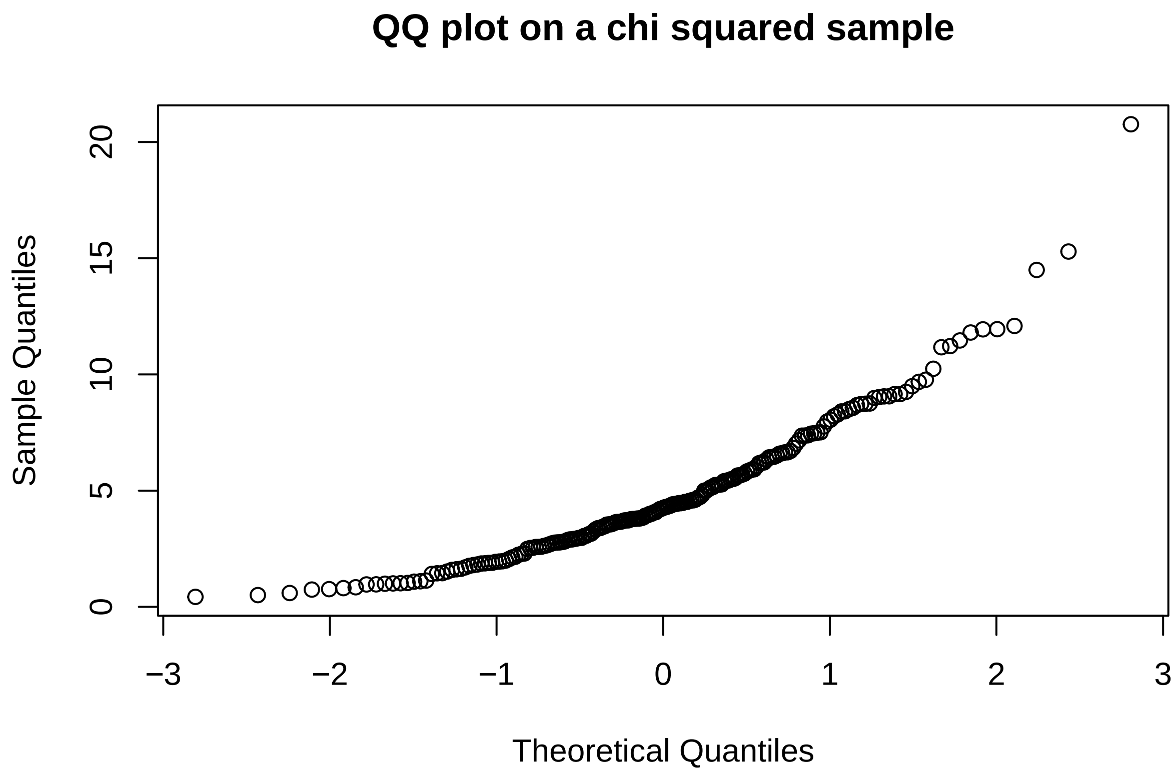
```
x.norm <- rnorm(n=200,mean=3,sd=2)
qqnorm(x.norm,main="QQ plot on a normal sample") #creates quantile plot against normal distribution
```



The line is linear, but you'll notice it doesn't have a slope of 1 and doesn't go through the origin. This is because the mean and standard deviation are not the same, but the sample is normal.

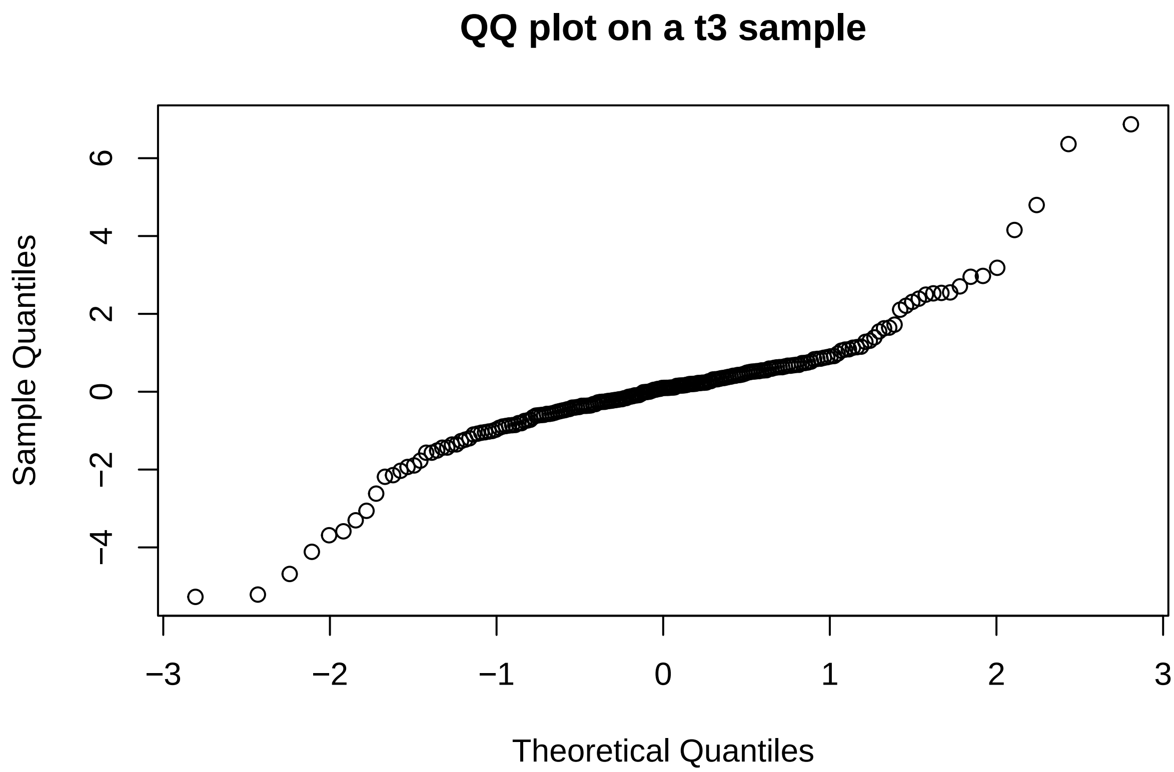
What if we try against non-normal distributions?

```
x.chi <- rchisq(n=200, df=5)
qqnorm(x.chi,main="QQ plot on a chi squared sample")
```



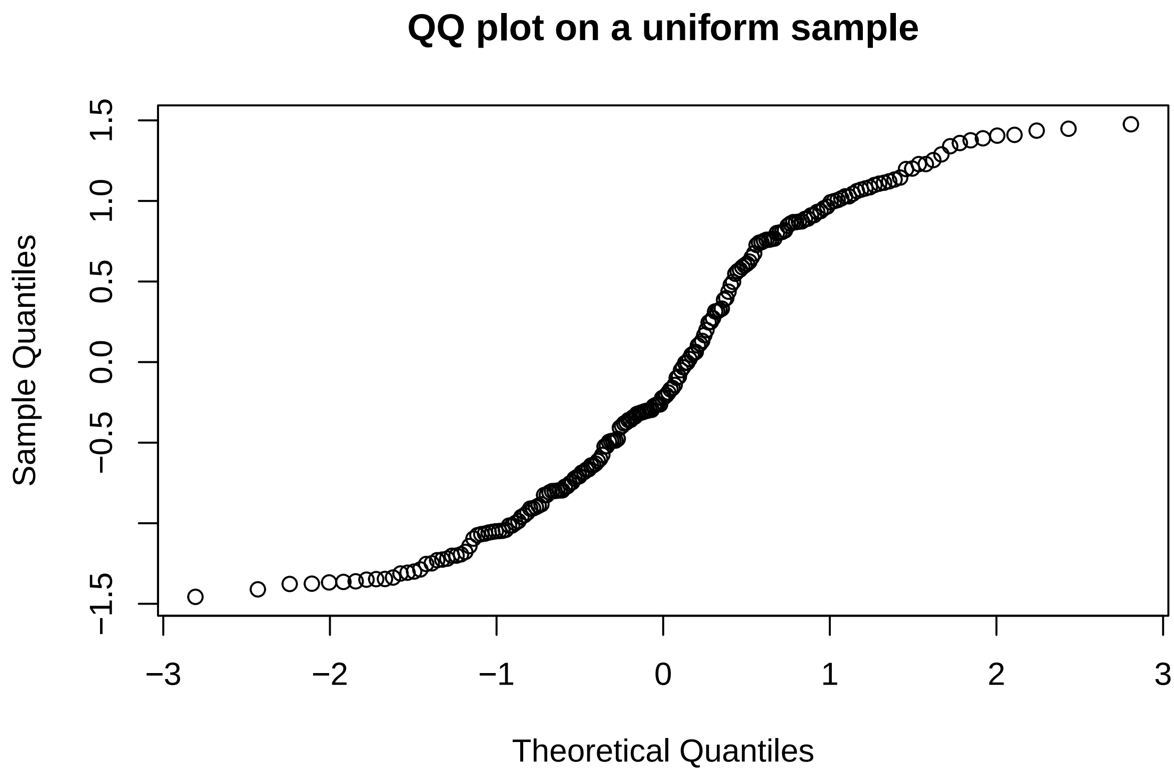
This graph is bow-shaped, showing that the plot is skewed. The right tail has more weight than the left tail. A normal distribution is symmetrical, so this is not normal.

```r
x.t <-rt(n=200, df=3)
qqnorm(x.t,main="QQ plot on a t3 sample")
```



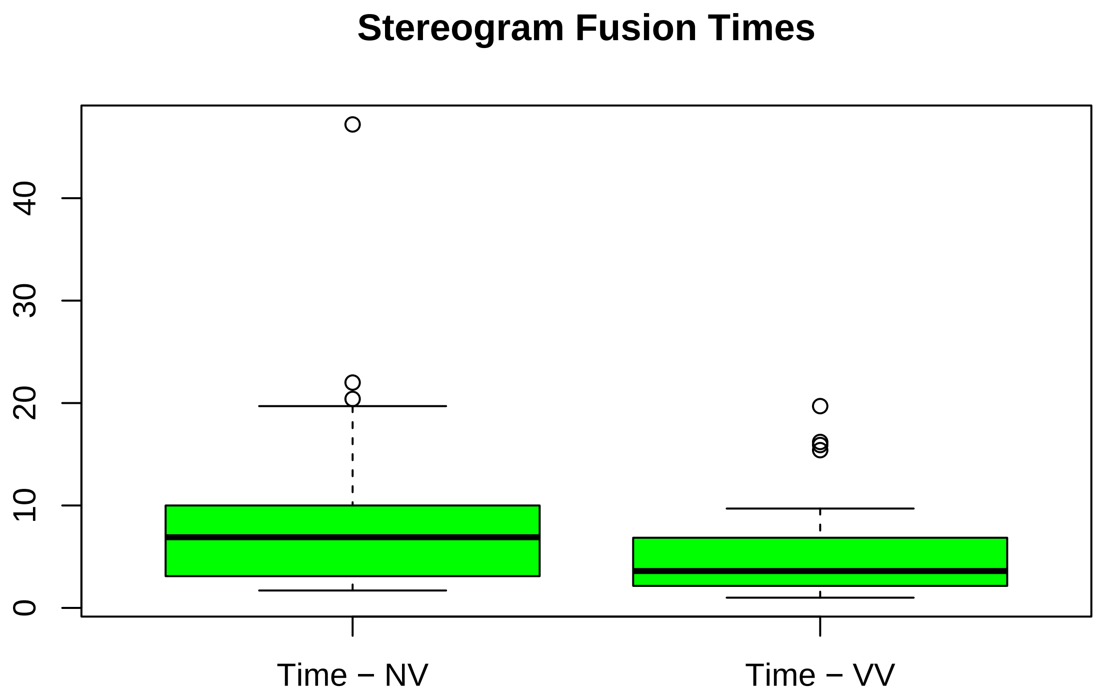
This graph deviates in the tails, with more weight in them then expected. This makes sense for a T-distribution, as it decays slower than normal distributions.

```
x.uniform <-runif(n=200, min=-1.5, max=1.5)
qqnorm(x.uniform,main="QQ plot on a uniform sample")
```

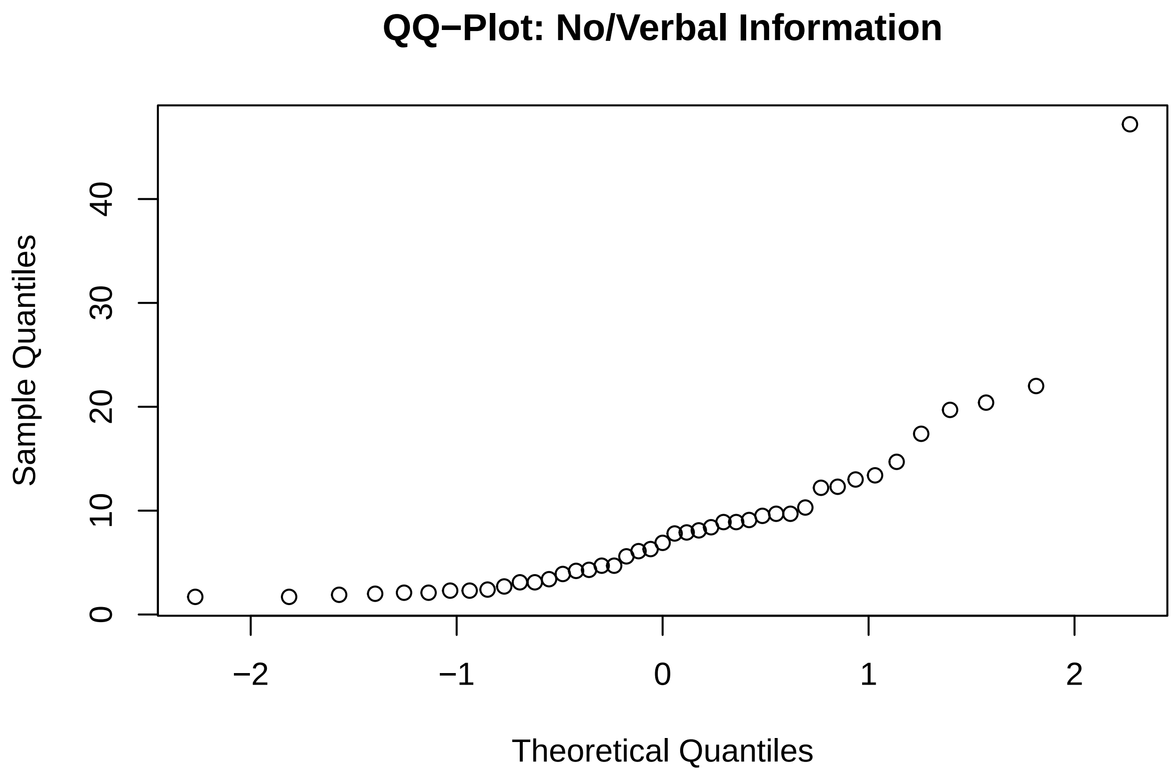


this is the opposite of a t-distribution plot. The left tail bends up and the right tail bends down, meaning both are lighter than in a normal distribution. _____ # Example: Stereogram dataset

```
stereograms<-read.table(file="~/Documents/STAT 359/data/stereograms.txt",
                         sep="",
                         header=TRUE)
time.NV<-stereograms$fusion_time[stereograms$group=='NV']
time.VV<-stereograms$fusion_time[stereograms$group=='VV']
boxplot(time.NV,time.VV,col='green',names=c('Time - NV','Time - VV'))
title('Stereogram Fusion Times')
```




```
qqnorm(time.NV,main='QQ-Plot: No/Verbal Information')
```


QQ−Plot: No/Verbal Information
Sample Quantiles
Theoretical Quantiles
0
10
20
30
40
−2
−1
0
1
2