Monte Carlo methods

Monte Carlo Simulation

Computers can be used to generate pseudo-random numbers. For most practicaly purposes these pseudo-random numbers can be used to immitate real random variables. This permits us to examine properties of random variables using a computer instead of theoretical or analytical derivations. One very useful aspect of this concept is that we can create *simulated* data to test out ideas or competing methods without actually having to perform laboratory experiments.

Simulations can also be used to check theoretical or analytical results. Also, many of the theoretical results we use in statistics are based on asymptotics: they hold when the sample size goes to infinity. In practice we never have an infinite number of samples so we may want to know how well the theory works with our actual sample size. Sometimes we can answer this question analytically, but not always. Simulations are extremely useful in these cases.

The technique we used to motivate random variables and null distribution was a type of monte carlo simulation. We had access to population data and generated samples at randome. Here we introduce a new dataset and focus specifically on Monte Carlos simulations. The dataset contains baby weights and several covariants, one of which is whether the mother smokes. Babies of smoking mothers tend to weigh slightly less. Say we want to know if a sample size of 10 is enough to use the central limit theorem to approximate the distribution of the t-statistic as normal with mean 0 and standard deviation 1. We will use Monte Carlo simulations to determine if10 large enough to use this approximantion? Let's use a monte carlo simulation to corroborate.

Below is the code we used to obtain random sample and then the difference:

```
# library(downloader)
# url<-"https://raw.githubusercontent.com/genomicsclass/dagdata/master/inst/extdata/babies.txt"
# filename <- tempfile()
# download(url,destfile="babies.txt")
dat <- read.table("babies.txt",header=TRUE)
smokers <- sample(dat$bwt[dat$smoke==1],10)
nonsmokers <- sample(dat$bwt[dat$smoke==0],10)
mean(smokers)-mean(nonsmokers)
```

```
## [1] -2.3
```

But as we have learned this is a random variabe and different random samples give us a different answer

```
for(i in 1:10) {
  smokers <- sample(dat$bwt[dat$smoke==1],10)
  nonsmokers <- sample(dat$bwt[dat$smoke==0],10)
  cat("observed difference = ",mean(smokers)-mean(nonsmokers),"ounces\n")
}
```

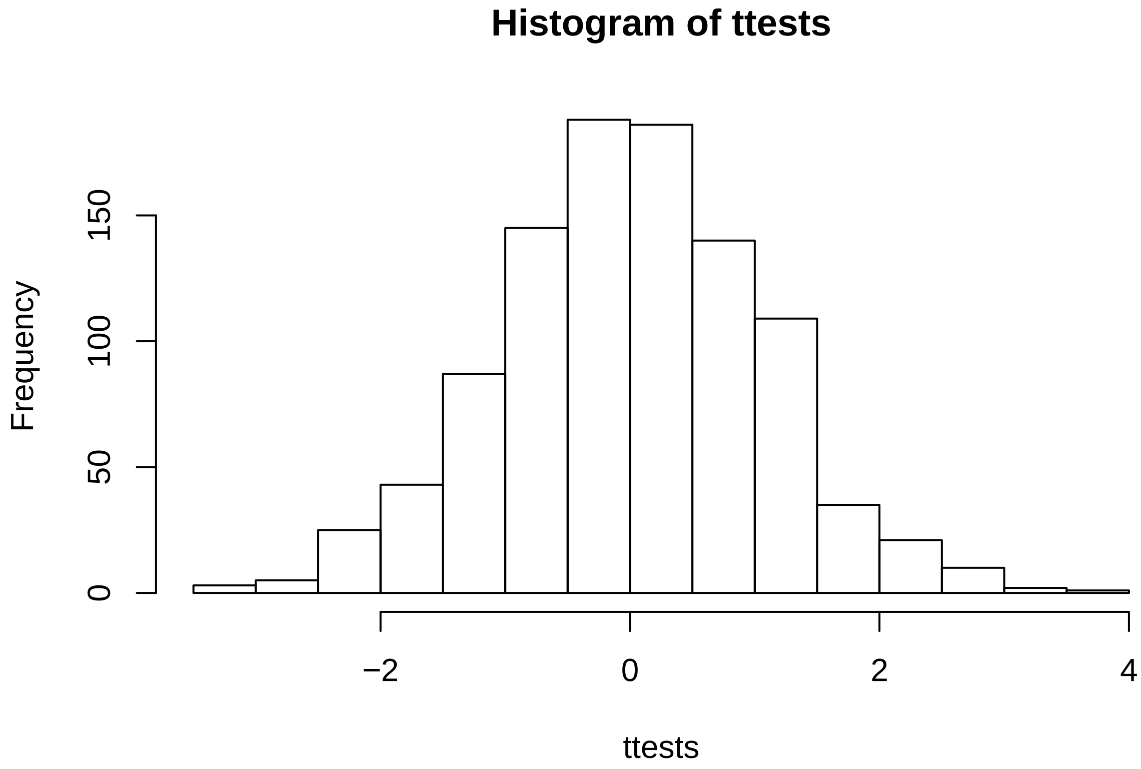
```
## observed difference =  -9.3 ounces
## observed difference =  -13.1 ounces
## observed difference =  -20.1 ounces
## observed difference =  -11 ounces
## observed difference =  -12.3 ounces
## observed difference =  -11.3 ounces
## observed difference =  -4.1 ounces
## observed difference =  -0.4 ounces
## observed difference =  -15.2 ounces
## observed difference =  -3.7 ounces
```

As noted earlier, in practice we can afford to measure so many samples, but on a computer it is as easy as writing a loop. Let's take 1,000 random samples under the null and re-computing the t-statistic:

```
ttestgenerator <- function(n) {
  # note that here we have a false "smokers" group where we actually
  # sample from the nonsmokers. this is because we are modeling the *null*
  smokers = sample(dat$bwt[dat$smoke==0], n)
  nonsmokers = sample(dat$bwt[dat$smoke==0], n)
  return((mean(smokers)-mean(nonsmokers))/sqrt(var(smokers)/n + var(nonsmokers)/n))
  }
ttests <- replicate(1000, ttestgenerator(10))
```

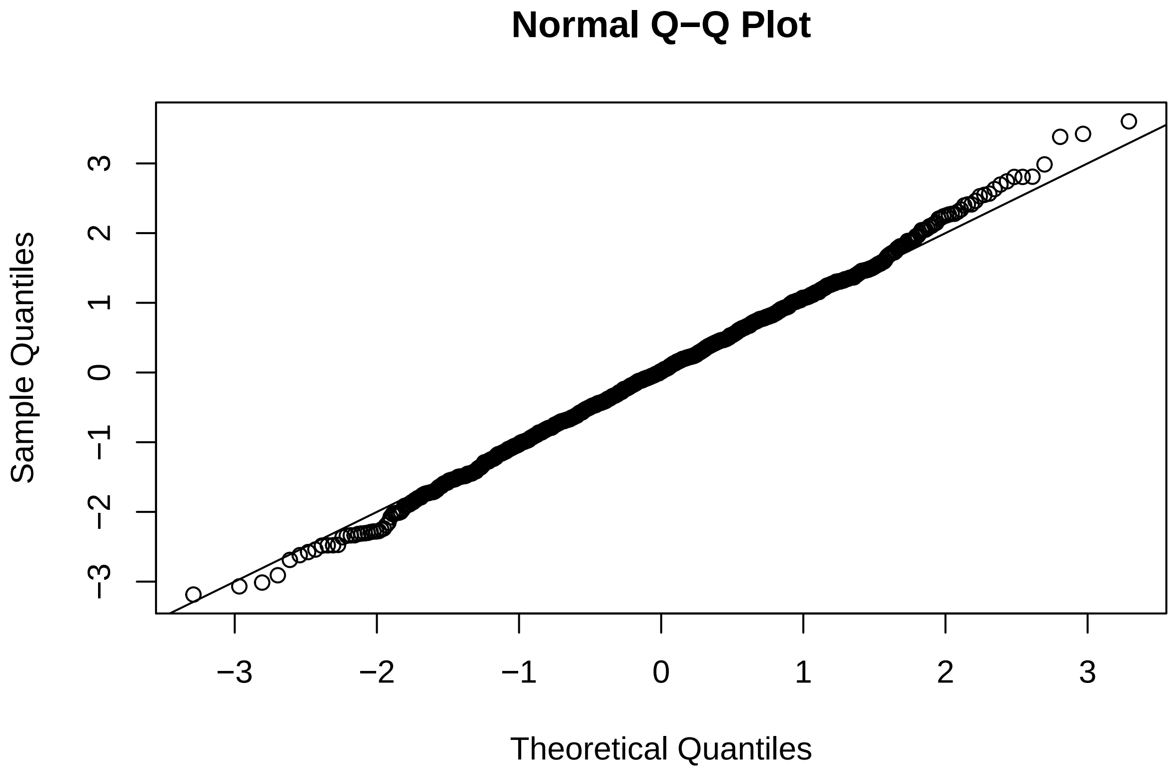
With 1,000 simulated ocurrences of this random variable we can now get a gimplse of it's distribution

```
hist(ttests)
```



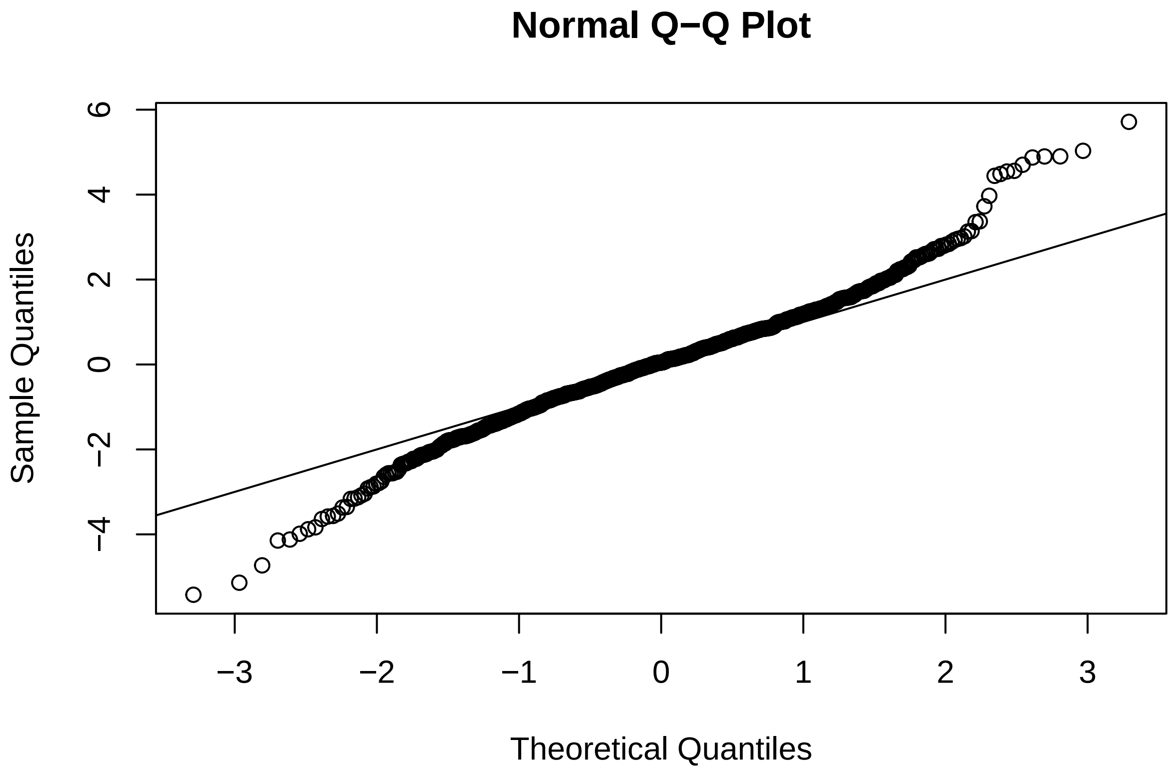
Now let's check on the theory used previously. Under the null hypothesis the difference in means is 0. To recreate this with our simulation we will sample non-smokers twice: there can't be a difference in population average if we sample from the same population. So is the distribution of this t-statistic well approximated by the normal distribution?

```
qqnorm(ttests)
abline(0,1)
```



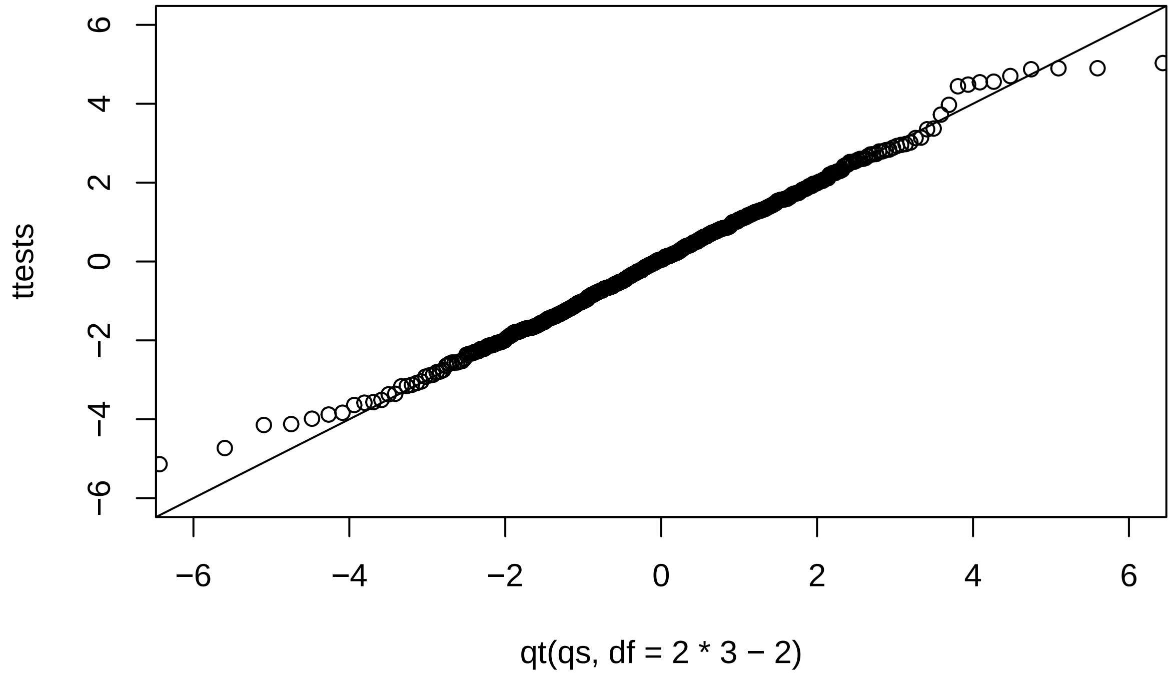
This looks like a very good approximation. So for this particular population a sample size of 10 was large enough to use the CLT approximation. How about 3?

```
ttests <- replicate(1000, ttestgenerator(3))
qqnorm(ttests)
abline(0,1)
```



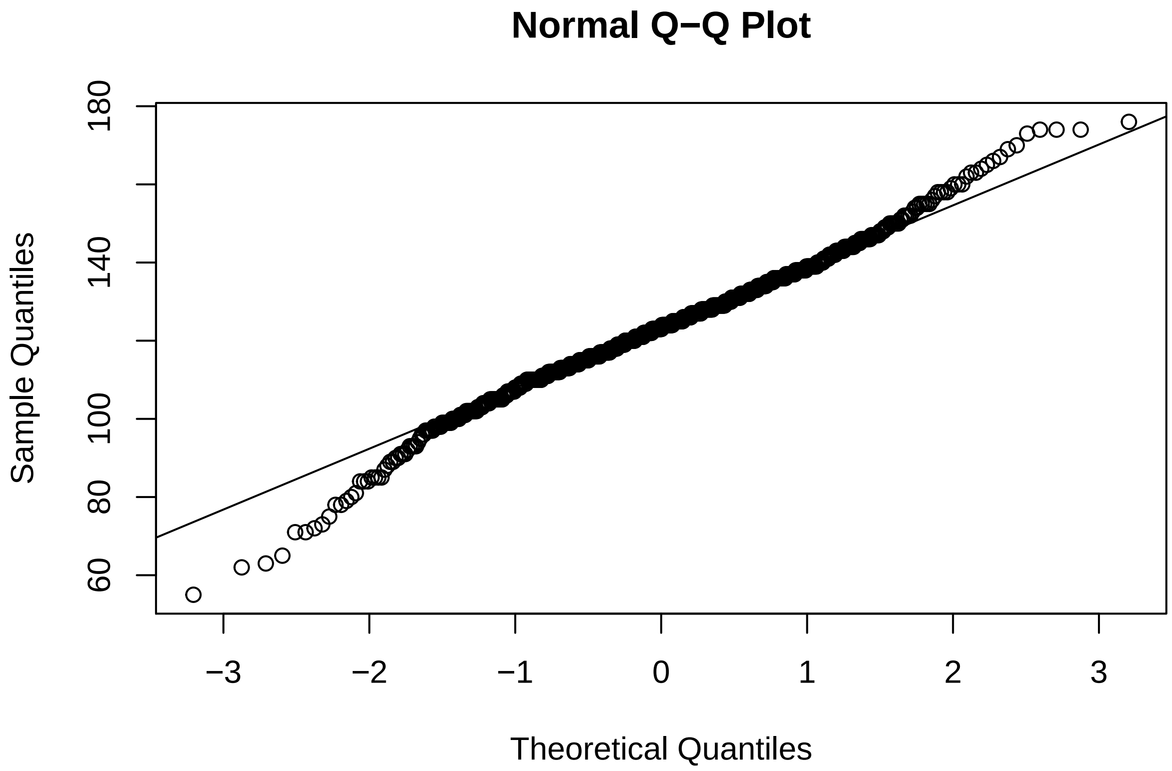
Now we see that the large quantiles (refered to by statisticians as the *tails*) are large than expected. In the previous module we explained that when the sample size is not large enough and the *population values* follow a normal distribution then the t-distribution is a better approximation. Our simulation results seem to confirm this:

```
qs <- (seq(0,999)+0.5)/1000
qqplot(qt(qs,df=2*3-2),ttests,xlim=c(-6,6),ylim=c(-6,6))
abline(0,1)
```



The t-distribution is a much better approximation in this case but it is still not perfect. This is due to the fact that the original data is not that well approximated by the normal distribution.

```
qqnorm(dat$bwt[dat$smoke==0])
qqline(dat$bwt[dat$smoke==0])
```



Parametric simulations for the observations

In the previous section we sampled from the entire population. In many cases we don't have access to data from the entire population. In these cases we can simulate the populaton data as well, using what is called a "parametric simulation". This means that we take parameters from the real data (here the mean and the standard deviation), and plug these into a model (here the normal distribution). This is acually the most common form of Monte Carlo simulation.

For the case of wieghts we could use:

```
nonsmokerweights <- rnorm(5000,
                          mean=mean(dat$bwt[dat$smoke==0]),
                          sd=sd(dat$bwt[dat$smoke==0]))
```

and repeat the entire excercise.

Optional homework:

1. How different are the N(0,1) and t-distribution when degrees of freedom are 18? How about 4?
2. For the case with 10 samples, what is the distribution of the sample median weight for smokers? How does the mean, median of the distribution compare to the population median?
3. Repeat the code above but simulated income for American and Canadians. Is the median income the same? is the mean income the same?