Lecture 1

Hisam Sabouni, PhD

January 2020

Overview

ECON 382 is the first econometrics course at the graduate level. The course focuses on the basic econometric theory with computer applications. We will cover all of the traditional econometrics technique starting from the simple and the multiple regression models to the simultaneous equation system estimation. In addition, we will introduce more advanced models which utilize longitudinal samples. In all these models, the emphasis will be on statistical inference and we will also pay attention to the application of the models in industrial organization, labor, health, and elsewhere in economics and social sciences. It is hoped that throughout the course, students will gain a thorough knowledge and understanding of econometrics theory and also develop useful skills in applying the methods to the empirical work. The focus will be on empirical work rather than on theoretical topics.

Why do we care?

Econometrics is the primary tool for economists to test their theories and understanding of the world using data. Econometrics is essentially just good old fashioned statistics tweaked to be suited towards problems economists faces such as answering policy questions and forecasting. One of the primary concern economists face is the issue of *causality*. Causality goes beyond the idea of simple association between two or more variables and establishes clear channels of relationships.

Economists face difficulty establishing causality because we primarily rely on observational data. In the natural sciences causal relationships are formed by controlled experiments where there are clear control and treatment groups. The groups outcomes are compared to estimate the effect the treatment has *ceteris paribus*. One should note however that experiments are actually conducted in economics in labs and in the field. These experimental settings try to follow the natural sciences through randomized control trials. However, given the nature of economic problems there are certain ethical and moral challenges faced with conducting large scale economic experiments (As there are for medical tests).

Our objective in this class is to introduce you to the core concepts in econometrics so that you can take your constructed economic models and conduct empirical analysis (use data to test your ideas). Before we can jump into the exciting world of econometrics. we will first review the core ideas in statistics that we will need to make sure everyone has a strong foundation in as well as review the tools we will be using throughout the class.

Tools

We will rely on the R programming language to build our models and analyze economic data. R can be downloaded at the following URL: https://cran.r-project.org/. In addition to R, we will be using the R studio integrated development environment (IDE), which allows us to create smart documents and program in a better environment that R provides. R studio can be downloaded for free at the following URL: https://www.rstudio.com/products/rstudio/download/. One of the great features of R is that we can easily load in data from pretty much anywhere really quickly. Furthermore, R is an open-source programming language where developers from all over the world contribute to the create of packages (https://cran.r-project.org/web/packages/available_packages_by_name.html). A package we will quite often rely upon is quantmod, which will allow us to pull data from FRED directly into R.

Packages need to be loaded in each R session in order to be used (i.e. every time you open R the package will remain loaded throughout your session unless you explicit *detach* the package.)

```r
#install.packages('quantmod')
library(quantmod)
```

In addition to using R we will also be using Python to learn how to collect data from the internet. To install Python on your machine please download and install Anaconda https://www.anaconda.com/download/.

Lastly you will need a latex installed on your machine to convert your RMarkdown files into pdf documents as well as to more easily write down all of your equations.

For Mac: http://www.tug.org/mactex/2016/downloading.html (Links to an external site.)

For Windows: https://miktex.org/download

Citation Notice

All of these notes are **heavily** borrowed from our textbook by Jefferey Wooldridge!

Fundamentals of Probability

Experiment: "An experiment is any procedure that can, at least in theory, be infinitely repeated and has a well-defined set of outcomes."

Random variable: "A variable that takes on numerical values and has an outcome that is determined by an experiment."

Bernoulli random variable: A random variable that can only take on the values of zero and one. $X \sim Bernoulli(\theta)$ which is read as "X has a Bernoulli distribution" with probability of success equal to θ".

In the case of a Bernoulli random variable we can denote the likelihood of the random variable taking on a 1 as θ and taking on a 0 as $1 - \theta$, given that the probabilities of all outcomes must sum to one.

$$P(X = 1) = \theta$$

$$P(X = 0) = 1 - \theta$$

In general, the probability density function (pdf) of a discrete random variable X summarizes the information concerning the possible outcomes of X and the corresponding probabilities.

Continuous random variables: Takes on any real value with zero probability. Here the idea is that a continuous random variable X can take on so many possible values that we cannot count them or match them up with the positive integers, so logical consistency dictates that X can take on each value with probability zero. We can define a probability density function for continuous random variables which will provide the information about the likely outcomes of the random variable. In general the cumulative distribution function (cdf) nicely summarizes the distribution of continuous random variables by measuring:

$$F(x) = P(X \leq x)$$

Where here the cdf tells us the likelihood of X, a random variable, being drawn that is less than a real number x.

Joint Distributions and Independence: If we have two discrete random variables X and Y, then (X, Y) have a joint distribution which is fully described by the joint probability density function of (X, Y):

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

The random variables X and Y are said to be independent iff (if and only iff):

$$f_{X,Y}(x, y) = f_x(x)f_y(y)$$

Hot hand fallacy?: A basketball player shooting two free throws. Let X be the Bernoulli random variable equal to one if she or he makes the first free throw, and zero otherwise. Let Y be a Bernoulli random variable if he or she makes the second free throw. How should we think about $P(X = 1, Y = 1)$? If a player gets 'hot', higher likelihood making shots if made past shots, then $P(X = 1, Y = 1) \neq P(X = x)P(Y = y)$. If the likelihood making the second shot is not conditional on the first shot and each shot is in fact independent of past shots then $P(X = 1, Y = 1) = P(X = x)P(Y = y)$.

This same logical can be extended to the studying joint distribution of 3-shots, 4-shots, etc. In the case of independence (no hot-hand) what we have is a sequence of independent Bernoulli trials (flipping a coin again and again). The probability density function of a series of independent Bernoulli trials is given by the binomial distribution:

$$f(x) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x},\ x = 0, 1, 2, \ldots, n$$

Here n is the number of trials (free-throws), x is the number of 'successful' trials (making a free-throw), and θ is the likelihood of 'success' (likelihood of making a free-throw). The first term $\frac{n!}{x!(n-x)!}$ gives us the total number of possible combinations of the events where the order does not matter. This tells us how many possible combinations there are to observe x successful outcomes given n trials.

Example: If a hotel has 100 available rooms, the hotel might be interested in knowing how many rooms they can book every night. Should the hotel take reservations for all 100 rooms? What if some of the people that made reservations cancel their reservation? What if the hotel takes more than 100 reservations, they might be then interested in $P(X > 100)$, where X is the number of clients that arrive for their reservation. If we assume that each room reservation is independent of other reservations then we can deploy our binomial distribution to study the likelihood of us overbooking and irritating customers! For example, suppose we decide to take on 120 reservations and the likelihood of a customer showing up is 70%. What is the probability that more than 100 customers will show up?

Lets write some code and to solve the problem:

```
#Using the Binomial Distribution
#?pbinom
binomial_distribution <- function(n,x,theta){



}
```

Covariance

Let $\mu_x = E(X)$ and $\mu_y = E(Y)$ and consider the random variable $(X - \mu_x)(Y - \mu_y)$. If X and Y are both above their mean then $(X - \mu_x)(Y - \mu_y) > 0$. If X and Y are both below their mean then $(X - \mu_x)(Y - \mu_y) > 0$. If $X > \mu_x$ and $Y < \mu_y$ or $X < \mu_x$ and $Y > \mu_y$ then $(X - \mu_x)(Y - \mu_y) < 0$.

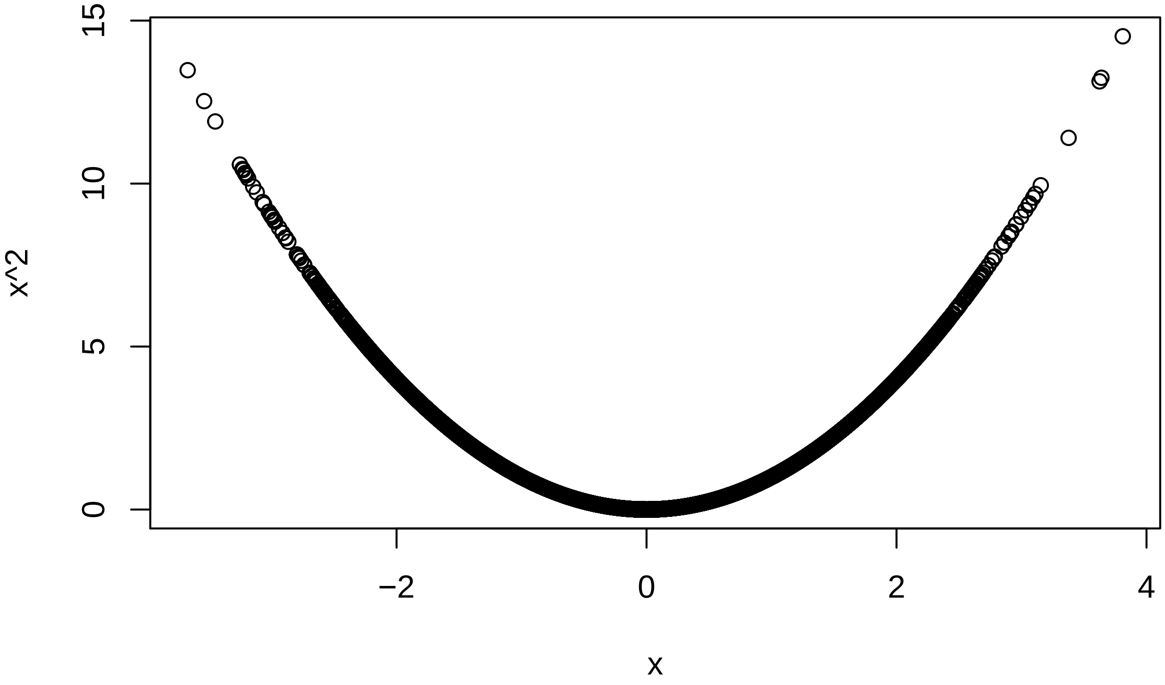
In general, the covariance $\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)]$ tells us about the co-movements of X and Y. When $\sigma_{xy} > 0$, then, on average, when X is above its mean Y is also above its mean. When $\sigma_{xy} < 0$, then, on average, when X is above its mean Y is below its mean (or vice versa). Here it is important to note that covariance measures the amount of **linear dependence** between two random variables.

Lets expand this: $E[(X - \mu_x)(Y - \mu_y)]$

If X and Y are independent, then $Cov(X, Y) = 0$.

However $Cov(X, Y) = 0$ does not imply that X and Y are independent.

```
set.seed(1)
#10,000 draws from a standard normal distribution
x <- rnorm(10000)
x_2 <- x^2
plot(x,x_2,xlab='x',ylab='x^2',main='')
```



```
sum((x - mean(x))*(x_2 - mean(x_2)))/(length(x) - 1)
```

```
## [1] -0.03775601
```

```
cov(x,x_2)
```

```
## [1] -0.03775601
```

Between any two random variables the absolute value of the covariances is bounded by the product of their standard deviations (Cauchy-Schwartz inequality):

$$|Cov(X,Y)| \leq \sigma_x \sigma_y$$

We can use this inequality (or equality of $X = aY$), to normalize our measure of covariance into what is known as the correlation coefficient:

$$Corr(X,Y) = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}$$

Maintaining the sign that emerges from the estimated covaraince (for $\sigma_x > 0$ and $\sigma_y > 0$) we have that:

$$-1 \leq Corr(X,Y) \leq 1$$

```
sum((x - mean(x))*(x_2 - mean(x_2)))/(length(x) - 1)/(sd(x)*sd(x_2))
```

```
## [1] -0.02596296
```

```
cor(x,x_2)
```

```
## [1] -0.02596296
```

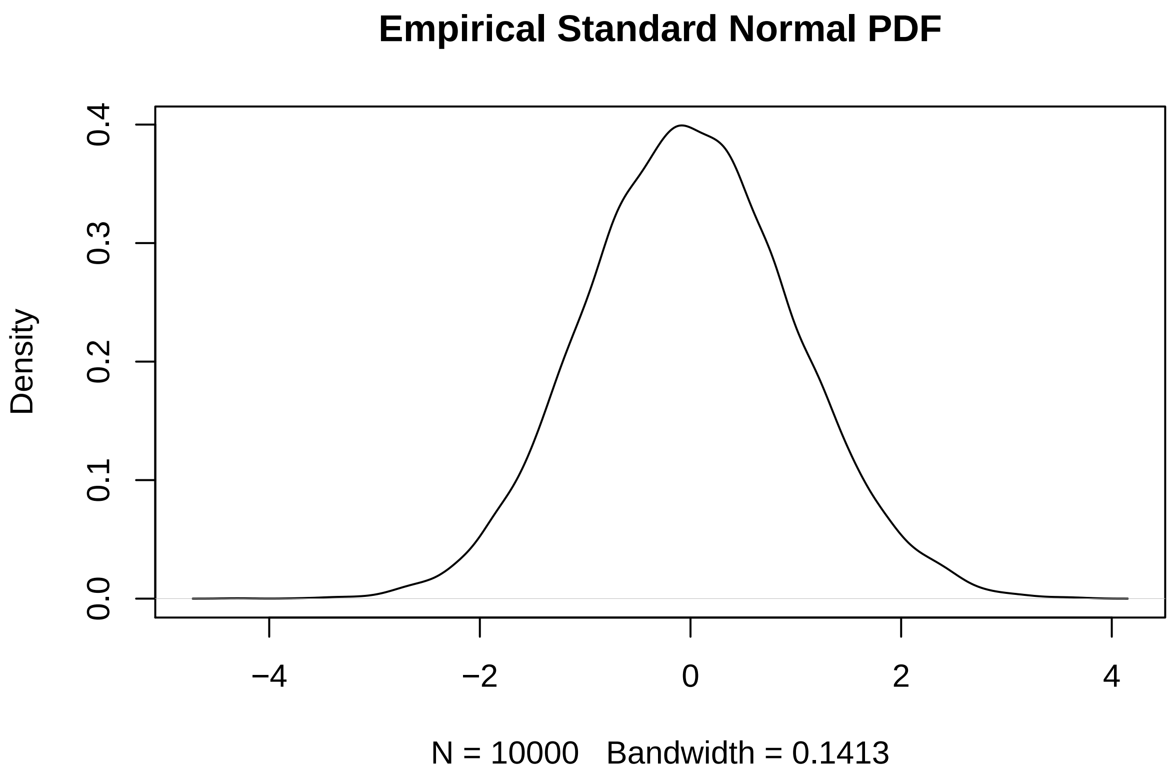
The Normal Distribution

The normal distribution and those derived from it are the most widely used distributions in statistics and econometrics. This is primarily due to the ease of working with the normal distribution along with the fact that many random variables in nature tend to follow a normal distribution (height, weight, innate ability, etc.). The probability density function of a random variable X that follows a normal distribution can be written as:

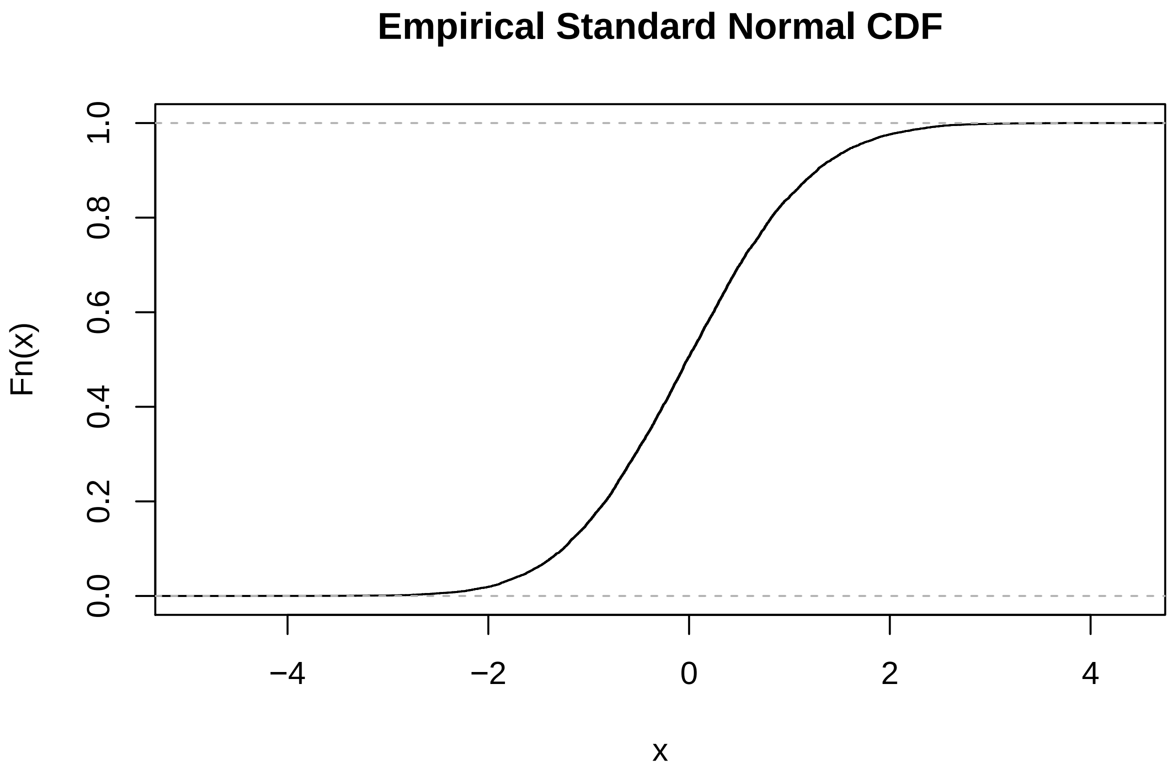
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp[-(x-\mu)^2/2\sigma^2], \text{ where } -\infty < x < \infty$$

Here $\mu = E(X)$ and $\sigma^2 = Var(X)$ (written more generally as $X \sim Normal(\mu, \sigma^2)$). There is no closed form solution for the cumulative distribution of the normal distribution there are however empirical estimates:

```
normal_data <- rnorm(10000)
plot(density(normal_data),typ='l',main='Empirical Standard Normal PDF')
```



```r
empirical_cdf <- ecdf(normal_data)
plot(empirical_cdf,main='Empirical Standard Normal CDF')
```



Lets download some data from the National Longitudinal Study and take a look at the distributions of a number of variables. We will use data that comes inside of an R package called modelr

```
#Uncomment the line below and run it to install the package. After install re-comment
#install.packages('modelr')
library(modelr)
#After loading in the package we can acces some of the data that comes with it. We wil
head(heights)
```

```
## # A tibble: 6 x 8
##   income height weight   age marital  sex    education  afqt
##    <int>  <dbl>  <int> <int> <fct>    <fct>      <int> <dbl>
## 1  19000     60    155    53 married  female        13  6.84
## 2  35000     70    156    51 married  female        10 49.4
## 3 105000     65    195    52 married  male          16 99.4
## 4  40000     63    197    54 married  female        14 44.0
## 5  75000     66    190    49 married  male          14 59.7
## 6 102000     68    200    49 divorced female        18 98.8
```

```
nrow(heights) #7006 rows of data
```

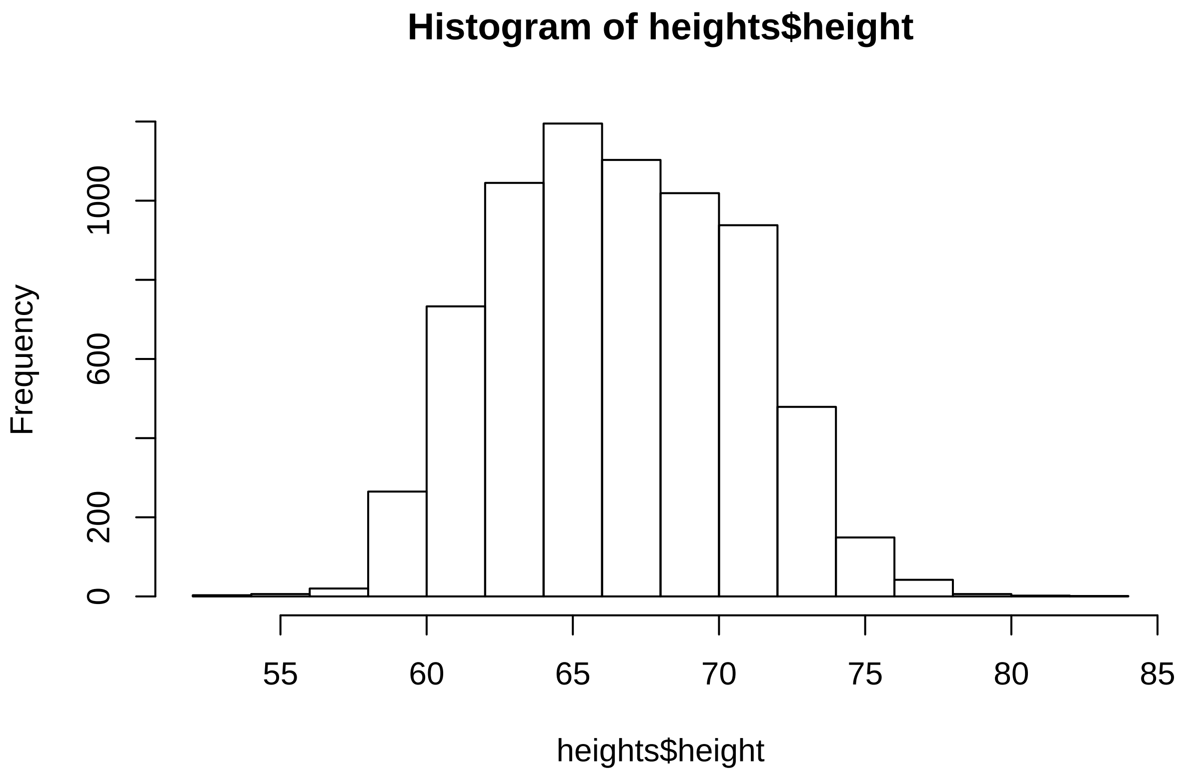
```
## [1] 7006
```

```
colnames(heights)
```

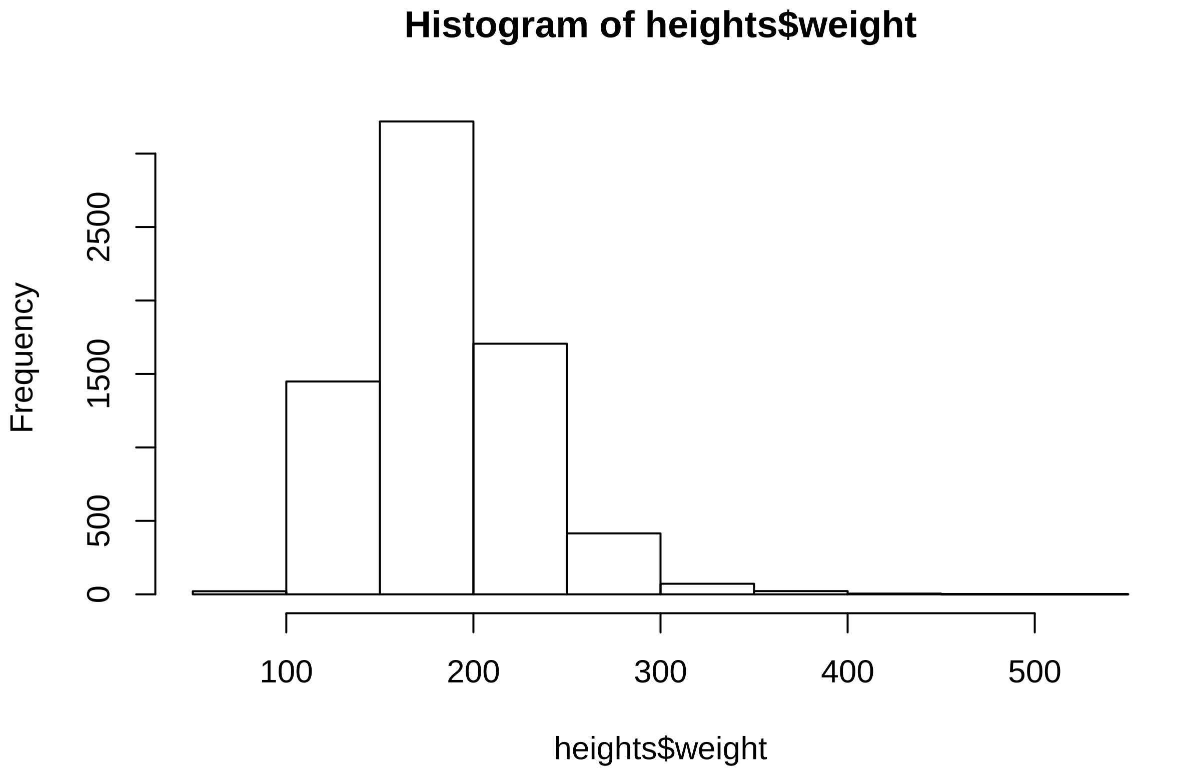
```
## [1] "income"    "height"    "weight"    "age"       "marital"   "sex"
## [7] "education" "afqt"
```

Here we have data on income, height, weight, age, marital status, sex, education, and a percentile score on the Armed Forces Qualification Test (a measure of innate ability like IQ). Lets take a look at the distributions for height and weight by creating histograms:

```
hist(heights$height)
```



```
hist(heights$weight)
```

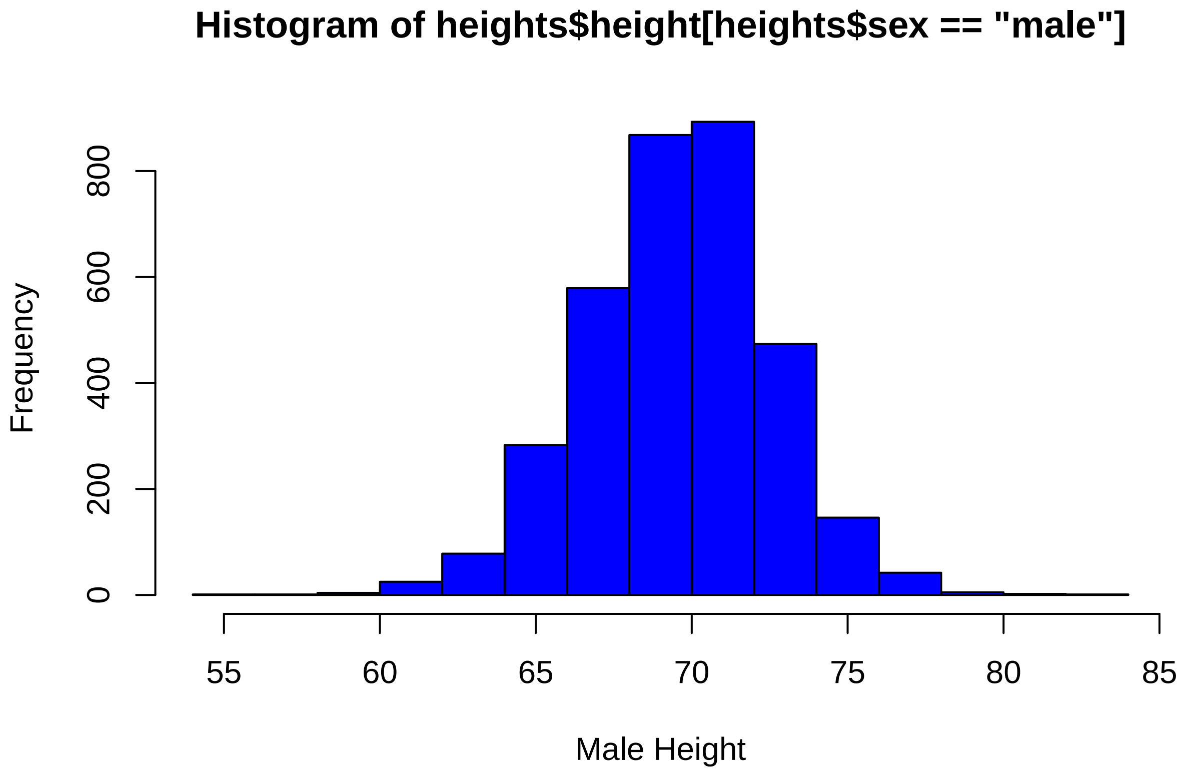


We can also breakdown the categories further by other variables. For instance we can compare survey respondents heights based on their sex:

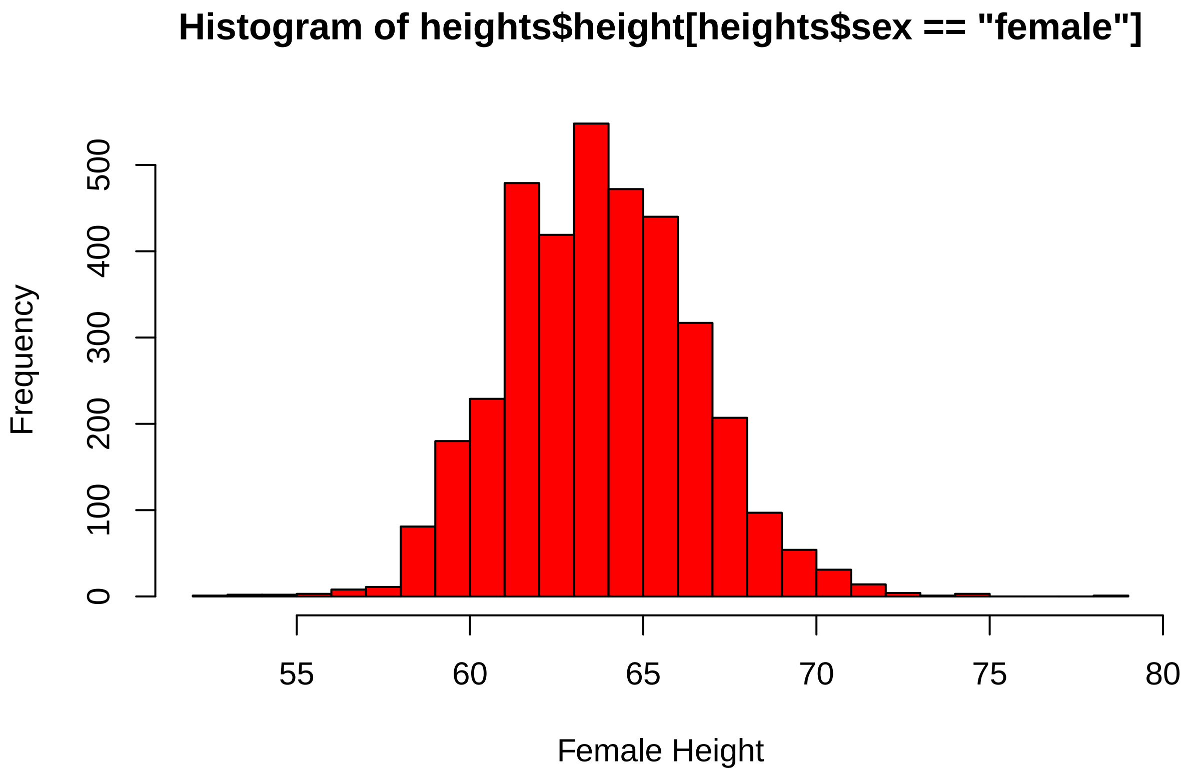
```
#How many reported male and females in sample?
table(heights$sex)
```

```
##
##   male female
##   3402   3604
```

```
hist(heights$height[heights$sex == 'male'],col='blue',xlab='Male Height',breaks = 20)
```



```
hist(heights$height[heights$sex == 'female'],breaks = 20,col='red',xlab='Female Height')
```



```
#What is the average male and female height?


#What is the standard deviation of male and female height?


#How can we compare our distributions to a normal distribution?
```