Welcome to STA302/STA1001

Mark Ebden, 7-13 September 2017, Week 1

With grateful acknowledgment to Alison Gibbs and Becky Lin

Arabian Nights #20-24: The Tale of the Three Apples

Week 1

- About this course
- Quick review of Statistics, through an R lens
- Introduction to Linear Regression

Course websites

- U of T portal: portal.utoronto.ca
- *Piazza* discussion forum: piazza.com/utoronto.ca/fall2017/sta3021001
- Recognized Study Groups: studygroups.artsci.utoronto.ca

Recognized Study Groups

Benefits of joining or starting a study group may include:

- Guaranteed regular study time
- A better understanding of material by working with your peers
- Meeting people in your program
- Gaining transferable leadership skills
- Access to Faculty of Arts & Science resources and support
- Receiving Co-Curricular credit

Online you can sign up to lead a study group, or find out which exist already

Climb "aboaRd"...

Let's look at R and RStudio now.

In your own time: because courses such as CSC108 are based in Python, you may like to save on your desktop a bilingual dictionary of Python-to-R: `http://mathesaurus.sourceforge.net/matlab-python-xref.pdf`

Python	R
`random.random((10,))`	`runif(10)`
`random.uniform((10,))`	
`random.uniform(2,7,(10,))`	`runif(10, min=2, max=7)`
`random.uniform(0,1,(6,6))`	`matrix(runif(36),6)`
`random.standard_normal((10,))`	`rnorm(10)`

What is Linear Regression?

"As with most statistical analyses, the goal of regression is to **summarize** observed data as simply, usefully and elegantly as possible." (Weisberg 2014)

In the case of simple linear regression, our summarizing model is:

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$$
$$\text{var}(Y|X = x) = \sigma^2$$

and we make some assumptions about the errors (the difference between actual values of y and what was expected).

We are modelling the *statistical relationship* between two variables.

Section 1 versus 2

This point in the slides is where we reached in the Thursday-morning class.

The Montreal Protocol, enacted 1 January 1989

Was the Montreal Protocol (MP) effective in reducing the atmospheric concentration of CFCs?

- There is a public database giving the concentration of CFCs from month to month: `ftp://aftp.cmdl.noaa.gov/data/hats/cfcs/cfc11/insituGCs/RITS/monthly/mlo_F11_MM.dat`
- Estimate two average measurements:
  - μ_1 for CFCs *before* the MP (1987-1989)
  - μ_2 for CFCs *after* the MP (1994-2000)
- Two-sample t-test: $H_0 : \mu_1 = \mu_2$, versus $H_a : \mu_1 \neq \mu_2$
- Result: $p \ll 0.05$, giving ample evidence that mean CFC concentration is not the same pre- and post-MP

The world will always need statisticians!

- The mean concentration is *higher* post-MP than pre-MP
- Why is the MP considered such a success? e.g. see www.ozoneheroes.org:

- Our fatal error was not to look at the data
- Linear regression would have been much more suitable here

Quick Review of Statistics

Example dataset:

```
set.seed(100)
# Generate 5 observations from a N(60,10^2):
dat <- round(rnorm(5,mean = 60,sd = 10),1)
dat
```

```
## [1] 55.0 61.3 59.2 68.9 61.2
```

Distributions

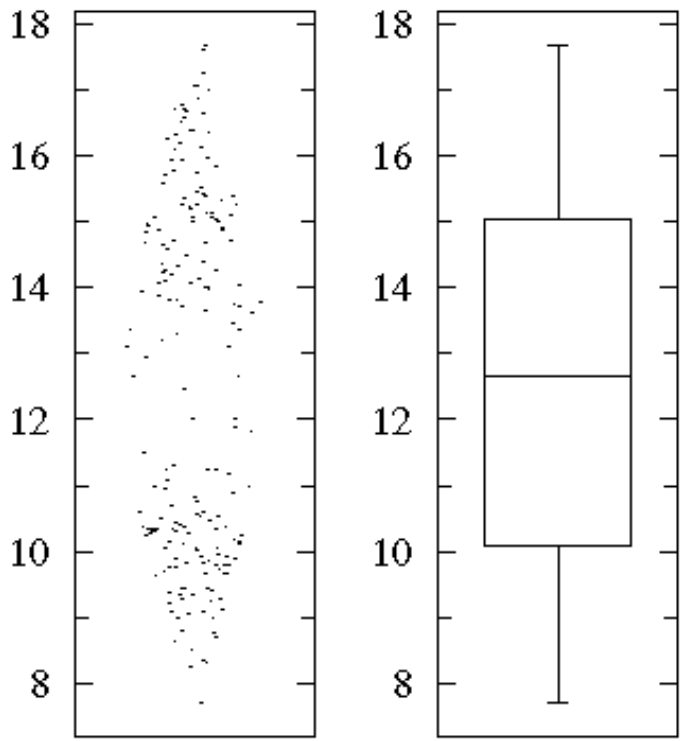
Distributions can be displayed numerically or graphically. An example of a numerical display in R is:

```
summary(dat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.00   59.20   61.20   61.12   61.30   68.90
```

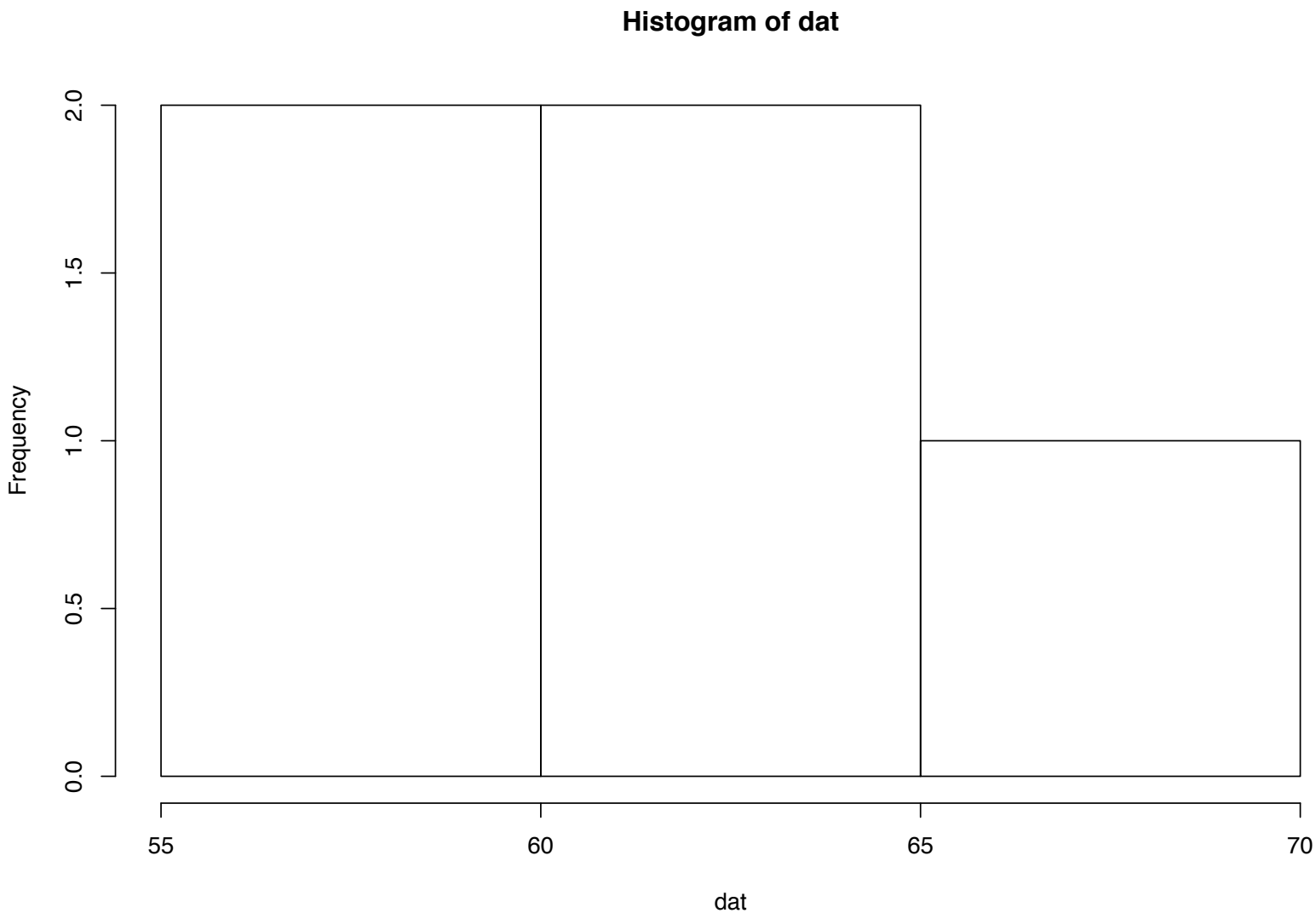
Distributions

Examples of graphical summaries of data sets in R are from the `boxplot` and `beeswarm` commands, or `hist` will give you the *histogram* (next slide).



Distributions

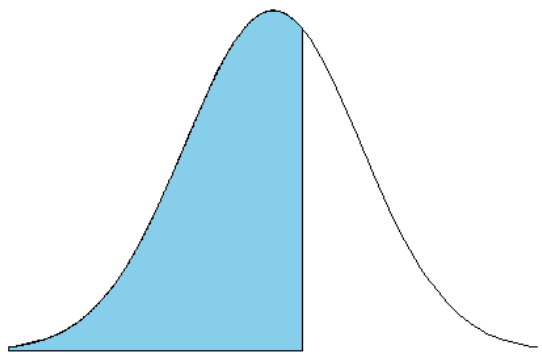
```
hist(dat)
```



The CDF for Continuous Distributions

- A continuous random variable X is fully characterized by its *density function* $f(x)$
- $f(x) \geq 0$, f is piecewise continuous, and $\int_{-\infty}^{\infty} f(x)dx = 1$
- The *cumulative distribution function* (CDF) of X is defined as:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$



Continuous Distributions

- If f is continuous at x then $F'(x) = f(x)$ (fundamental theorem of calculus)
- The CDF can be used to calculate the probability that X falls in the interval (a, b). This is the area under the density curve which can also be expressed in terms of the CDF:

$$P(a < X < b) = \int_a^b f(x)dx = F(b) - F(a)$$

Going back to R...

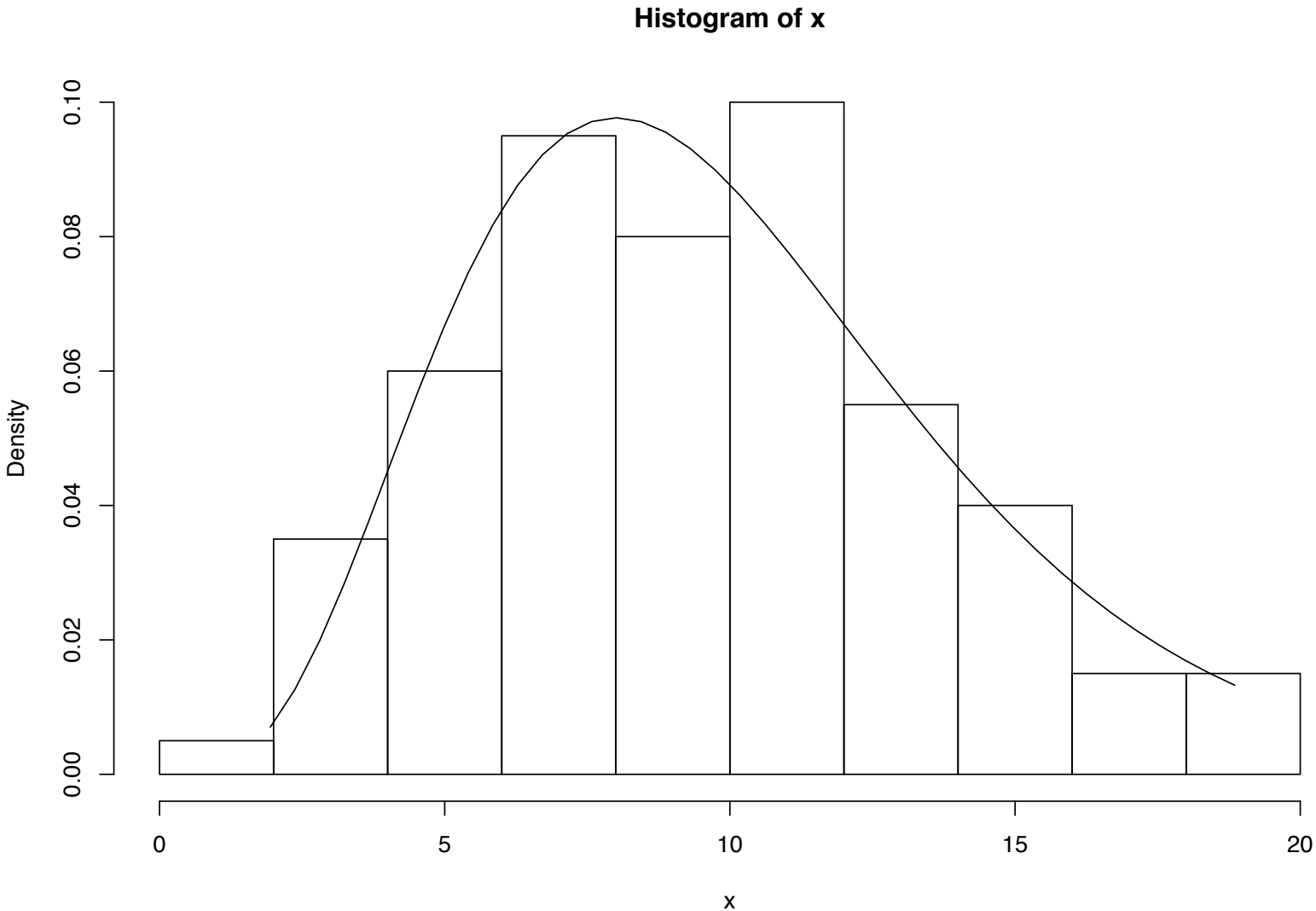
- In R, a list of all the common distributions can be obtained by the command `help("distributions")`
- For example, the normal density and CDF are given by `dnorm()` and `pnorm()`
- In such functions:
  - `d` = density function
  - `p` = CDF
  - `q` = quantile function
  - `r` = random sample

An example: χ^2

- The function `rchisq()` gives observations from a χ^2 distribution
- `rchisq(100,10)` will sample 100 times from a χ^2_{10} distribution
- On the next slide, the true density function of the χ^2_{10} is superimposed over the histogram of the sample. Here's the full code:

```
x<- rchisq(100,10) # draw a sample of 100 from chi-squared 10
h <- hist(x,freq=FALSE) # create the histogram
# superimpose chi-squared density over histogram
xfit <- seq(min(x),max(x),length=40)
yfit <- dchisq(xfit,10) # chi-square density
lines(xfit,yfit)
```

An example: χ^2

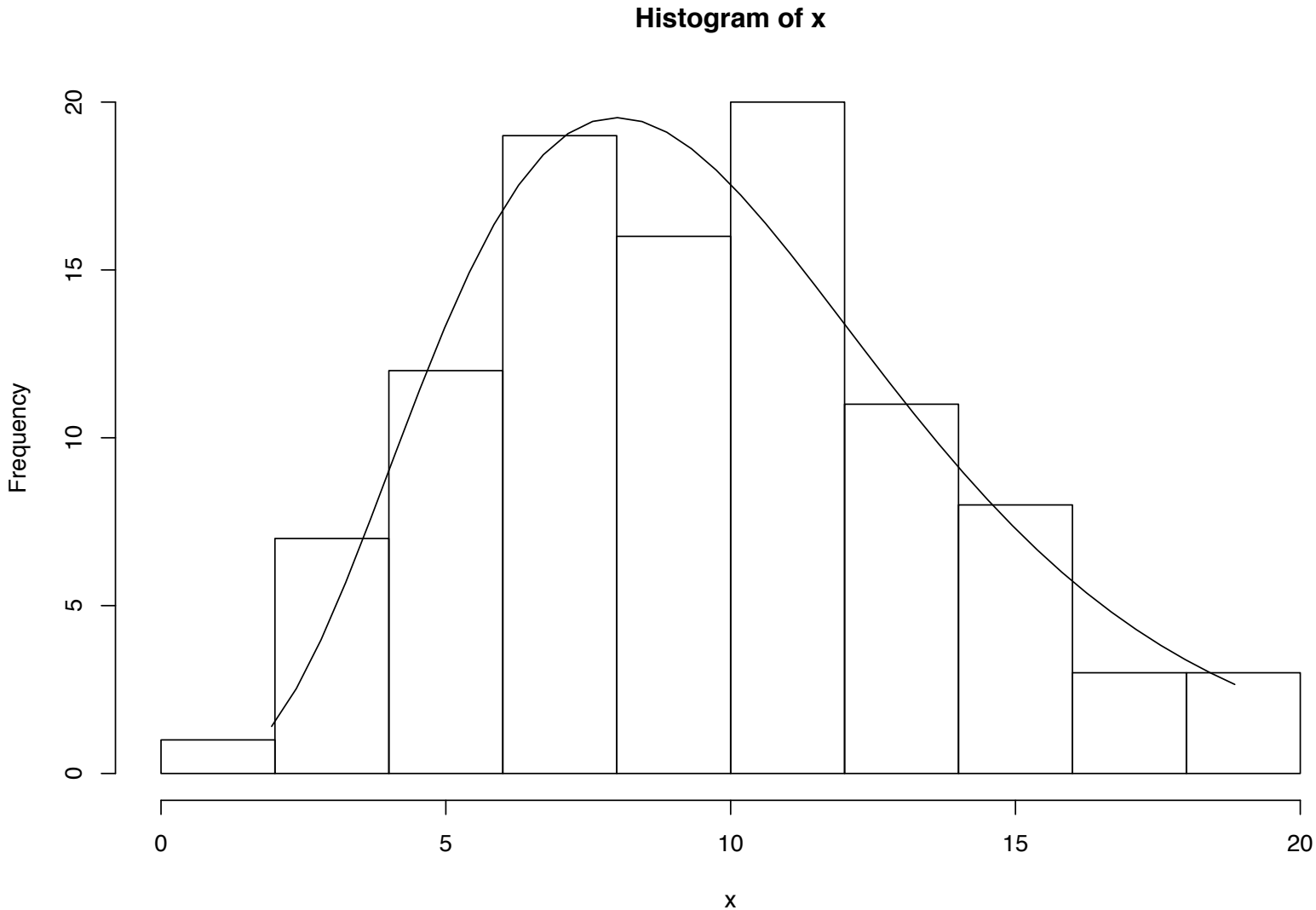


Alternative plot

Instead of plotting the density function, you may prefer to plot the frequency counts.

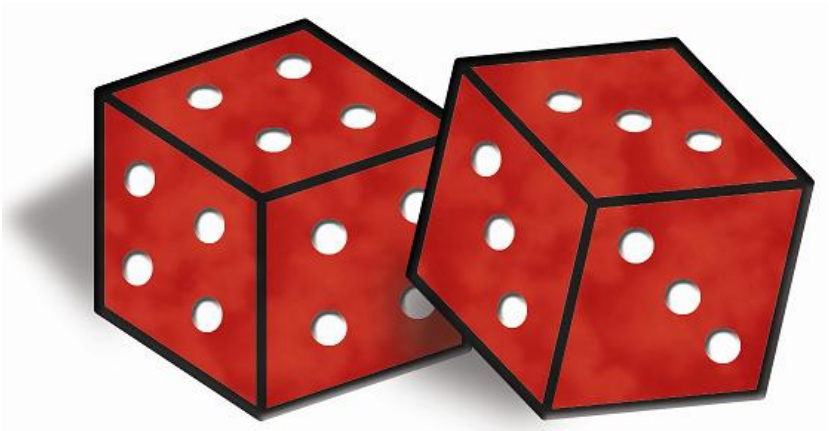
```
h <- hist(x) # create the histogram
# superimpose chi-squared density over histogram
xfit <- seq(min(x),max(x),length=40)
yfit <- dchisq(xfit,10) # chi-square density
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit,yfit)
```

Alternative plot



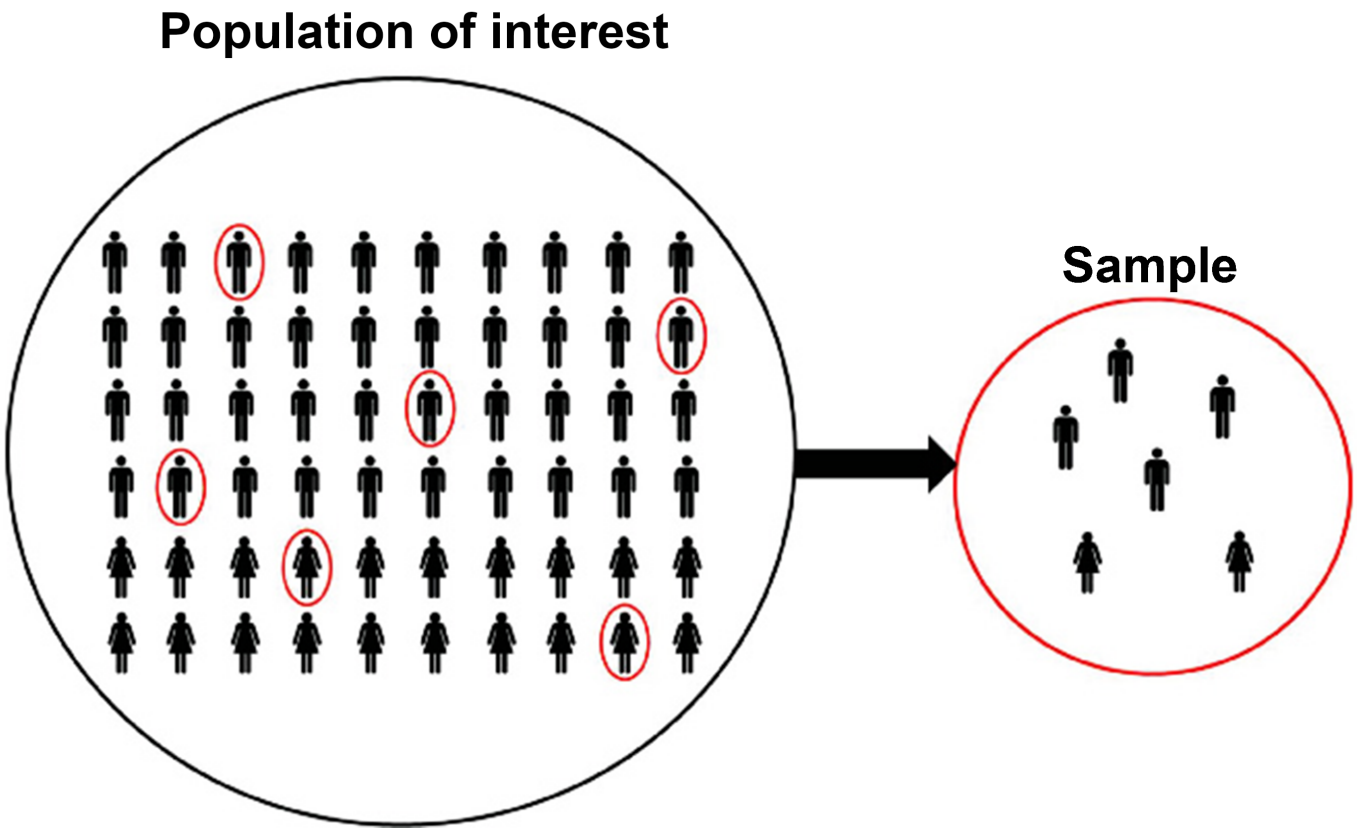
Reminder about Randomness

- In the previous slides we were simulating drawing from a distribution. What does that mean?
- A *random drawing* is one in which each member of the population has an equal chance of being selected.
- The hypothesis of random sampling may not apply to real data
- For example, cold days are usually followed by cold days, so daily temperature is not directly representable by random drawings



Reminder about the data we're discussing

The total aggregate of observations that might occur as a result of repeatedly performing a particular operation is called a *population* of observations. The observations that actually occur are some kind of *sample* from the population.



Parameters and Statistics

What is the difference between a parameter and a statistic?

- A *parameter* is a population quantity and a *statistic* is a quantity based on a sample drawn from the population

Example: The population of all adult (18+ years old) males in Toronto, Canada.

- Suppose that there are N adult males and the quantity of interest, y, is age
- A sample of size n is drawn from this population
- The population mean is $\mu = \sum_{i=1}^{N} y_i/N$ which is a parameter
- The sample mean is $\bar{y} = \sum_{i=1}^{n} y_i/n$ which is a statistic
- We can use x instead of y to describe this

Back to Distributions: The Gaussian

- The density function of the normal distribution with mean μ and standard deviation σ is:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

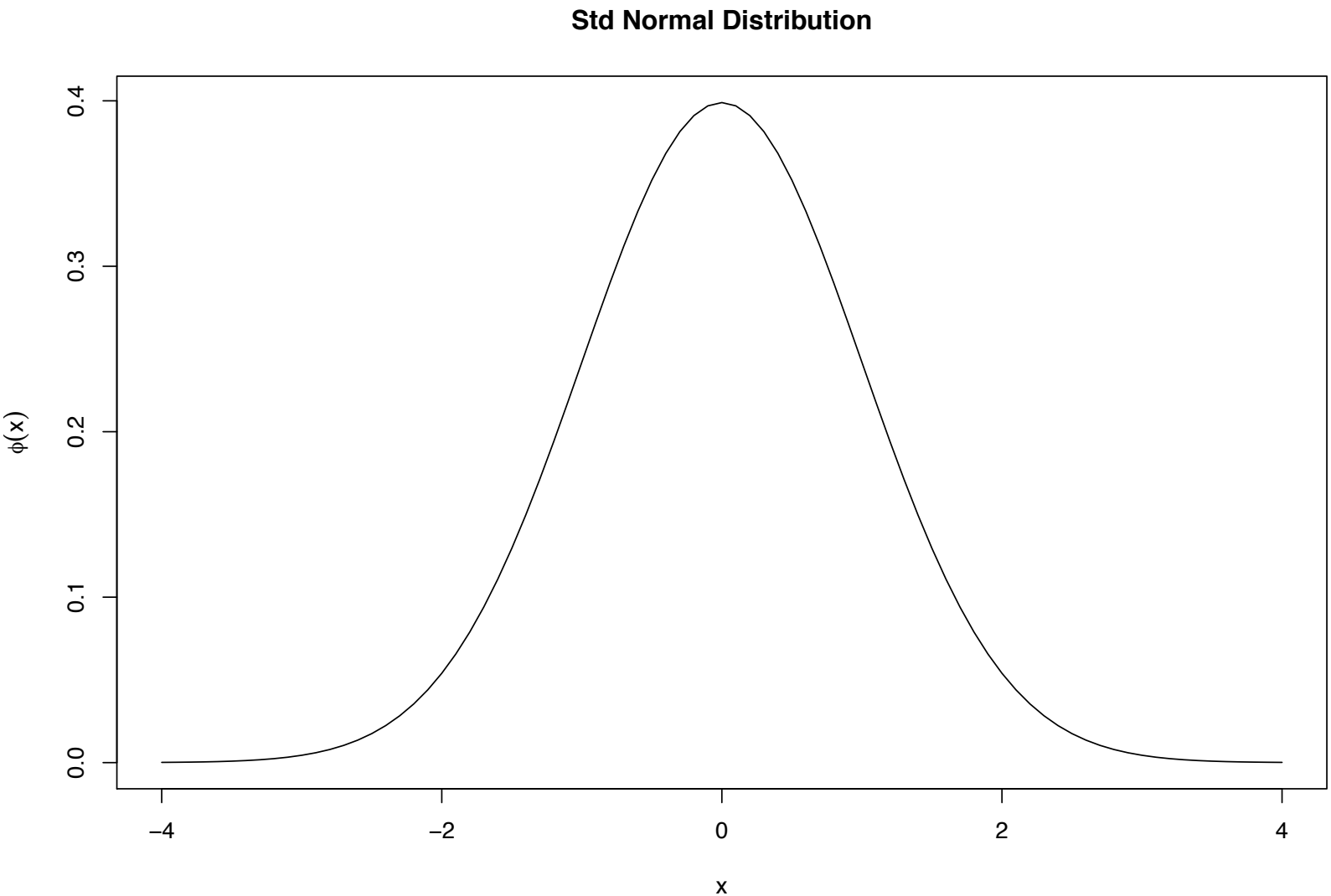
- The cumulative distribution function (CDF) of a $\mathcal{N}(0,1)$ distribution is:

$$\Phi(x) = P(X < x) = \int_{-\infty}^{x} \phi(x)dx$$

This isn't available in closed form, so use R.

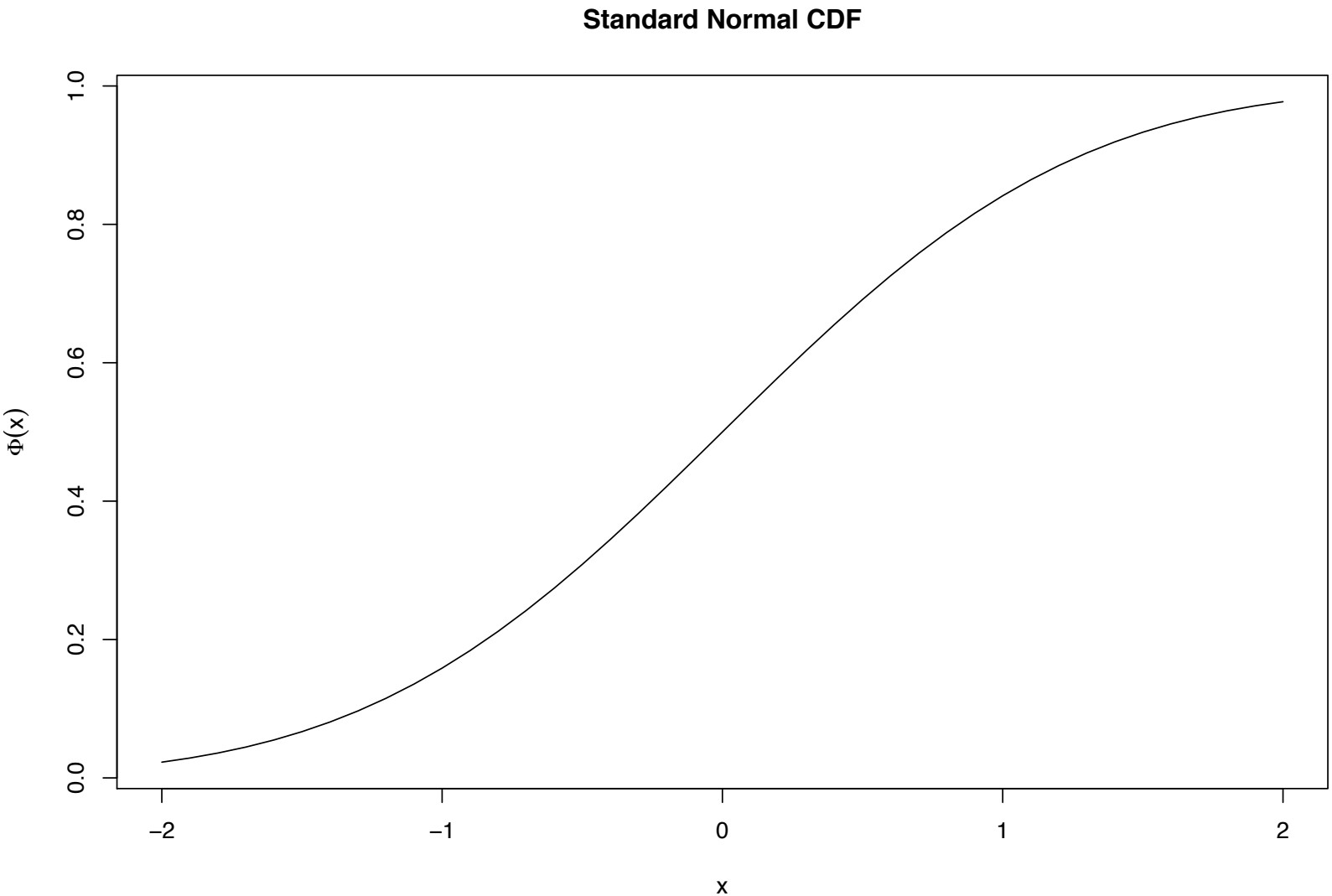
The Normal Distribution

```
x <- seq(-4,4,by=0.1)
plot(x,dnorm(x),type="l",main = "Std Normal Distribution",
     ylab=expression(paste(phi(x))))
```



The Normal Distribution

```
plot(x <- seq(-2,2,by=0.1),pnorm(x),type="l",
     xlab="x",ylab=expression(paste(Phi(x))),
     main = "Standard Normal CDF")
```



The Normal Distribution

- A random variable X that follows a normal distribution with mean μ and variance σ^2 will be denoted by

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right).$$

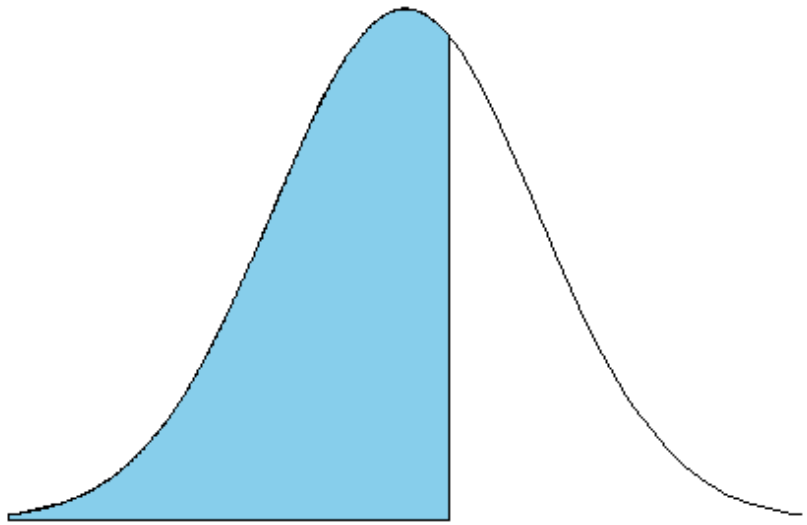
- If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ and $Z \sim \mathcal{N}(0,1)$, then

$$Z = \frac{X - \mu}{\sigma}$$

  This gives what is known as a Z-value or standardized variable
- Exercise: Double check that $E(Z) = 0$ and $\text{var}(Z) = 1$.

The Normal Distribution: Question for You

Let $X \sim \mathcal{N}(5, 3)$. How might we use R to find $P(X < 6)$?



1. `dnorm(6, mean = 5, sd = sqrt(3)) # Or just dnorm(6,5,sqrt(3))`
2. `pnorm(6, mean = 5, sd = sqrt(3))`
3. `qnorm(6, mean = 5, sd = sqrt(3))`
4. `rnorm(6, mean = 5, sd = sqrt(3))`

Next question: How would we use R to find $P(4 < X < 6)$? (See answer slide.)

Central Limit Theorem

- This theorem helps explain the massive importance of the normal distribution.
- The central limit theorem states that if $X_1, X_2, \ldots$ is an independent sequence of identically distributed random variables with mean $\mu = E(X_i)$ and variance $\sigma^2 = \text{var}(X_i)$ then

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) = \Phi(x),$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$ and $\Phi(x)$ is the standard normal CDF. This means that the distribution of $\bar{X}$ is approximately $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Central Limit Theorem

Example: A fair coin is flipped 50 times. What is the distribution of the average number of heads?

- Label the outcomes $X_1, \ldots, X_{50}$, where $X_i = 1$ if the toss is a head and $X_i = 0$ if the toss is a tail
- Since the coin is fair, $P(X_i = 1) = 0.5, i = 1, \ldots, 50$. The average number of heads is $\sum_{i=1}^{50} X_i/50$.
- Because $E(X_i) = 0.5$ and

$$\text{Var}(X_i) = p(1-p) = 0.5(1-0.5) = 0.25,$$

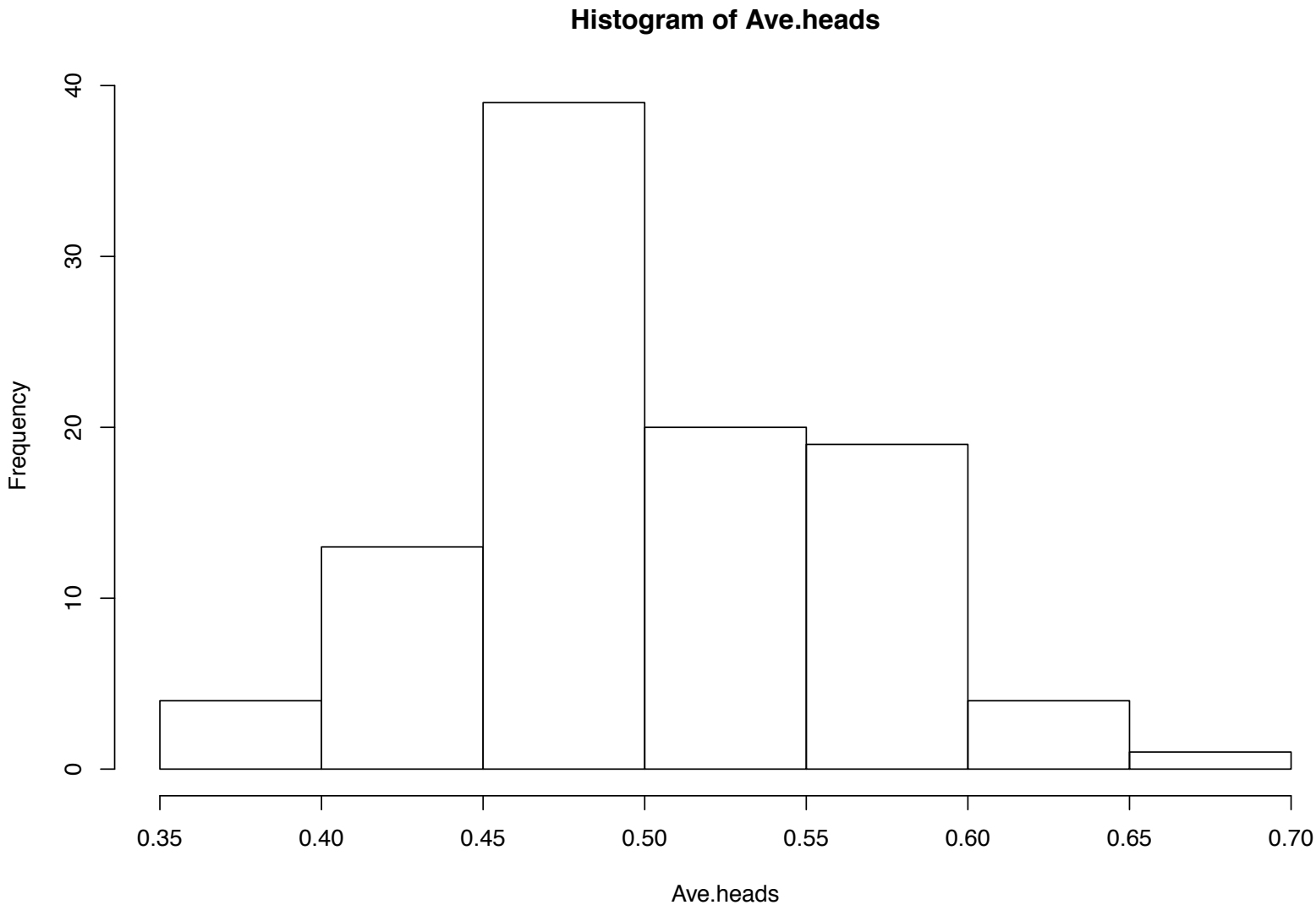
it follows that

$$\sum_{i=1}^{50} X_i/50 \overset{\text{approx}}{\sim} \mathcal{N}(0.5, 0.25/50).$$

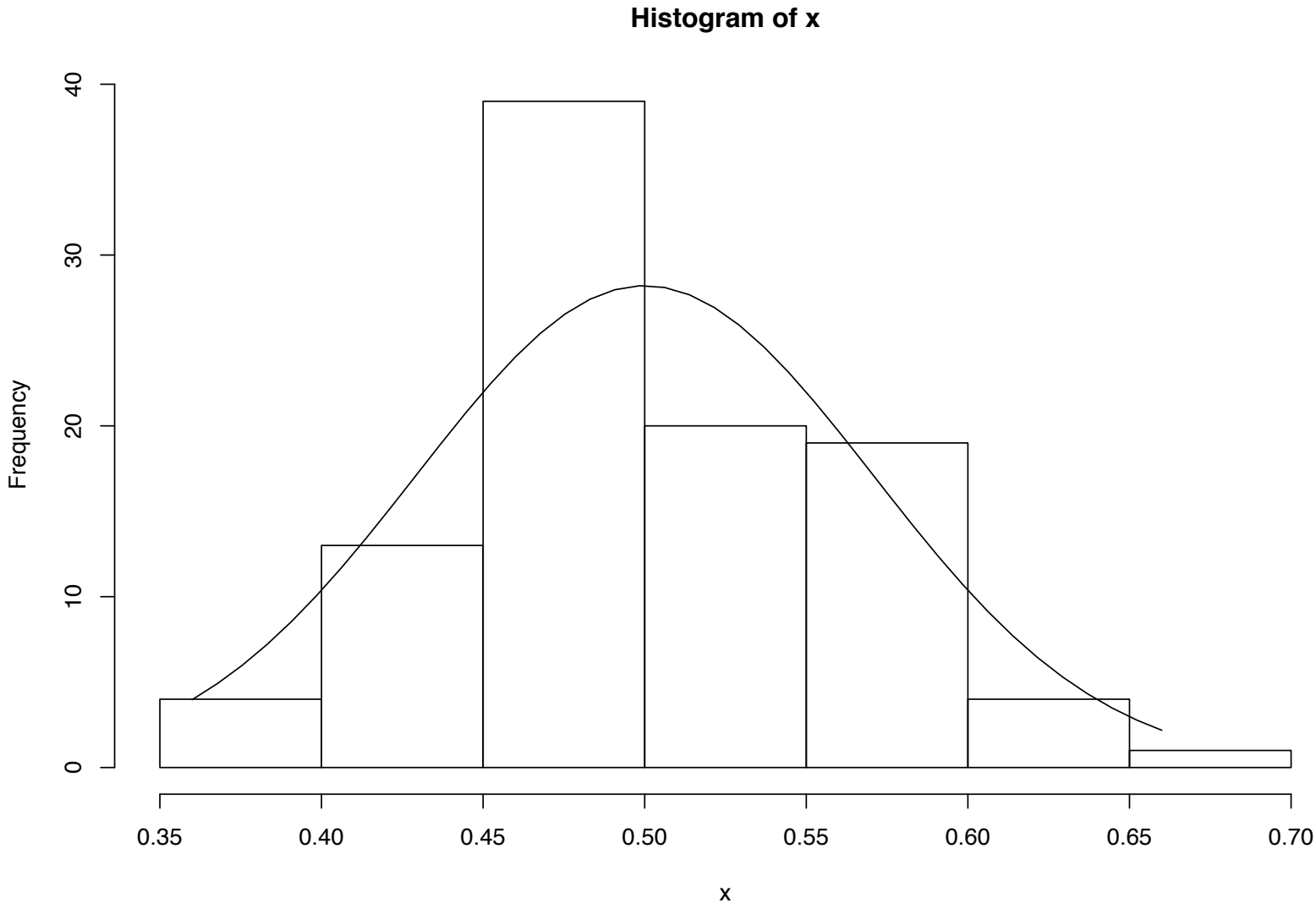
- **How to do this in R:** Typing `rbinom(n,50,0.5)` will draw n times from a $p = 0.5$ binomial distribution with 50 trials

Central Limit Theorem

```
set.seed(100)
Total.heads <- rbinom(100,50,0.5); Ave.heads <- Total.heads/50;
hist(Ave.heads)
```



Central Limit Theorem



How do the normal and χ^2 distributions relate?

- Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variables that have a $\mathcal{N}(0, 1)$ distribution.
- The distribution of

$$\sum_{i=1}^{n} X_i^2$$

  has a χ^2 distribution on n degrees of freedom or χ^2_n.
- The mean of a χ^2_n distribution is n and its variance is $2n$.

Example Application of the χ^2 Distribution

- Let $X_1, X_2, ..., X_n$ be independent with a $\mathcal{N}(\mu, \sigma^2)$ distribution
- The distribution of the sample variance $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ turns out to be:

$$S^2 \sim [\sigma^2/(n-1)]\ \chi^2_{n-1}$$

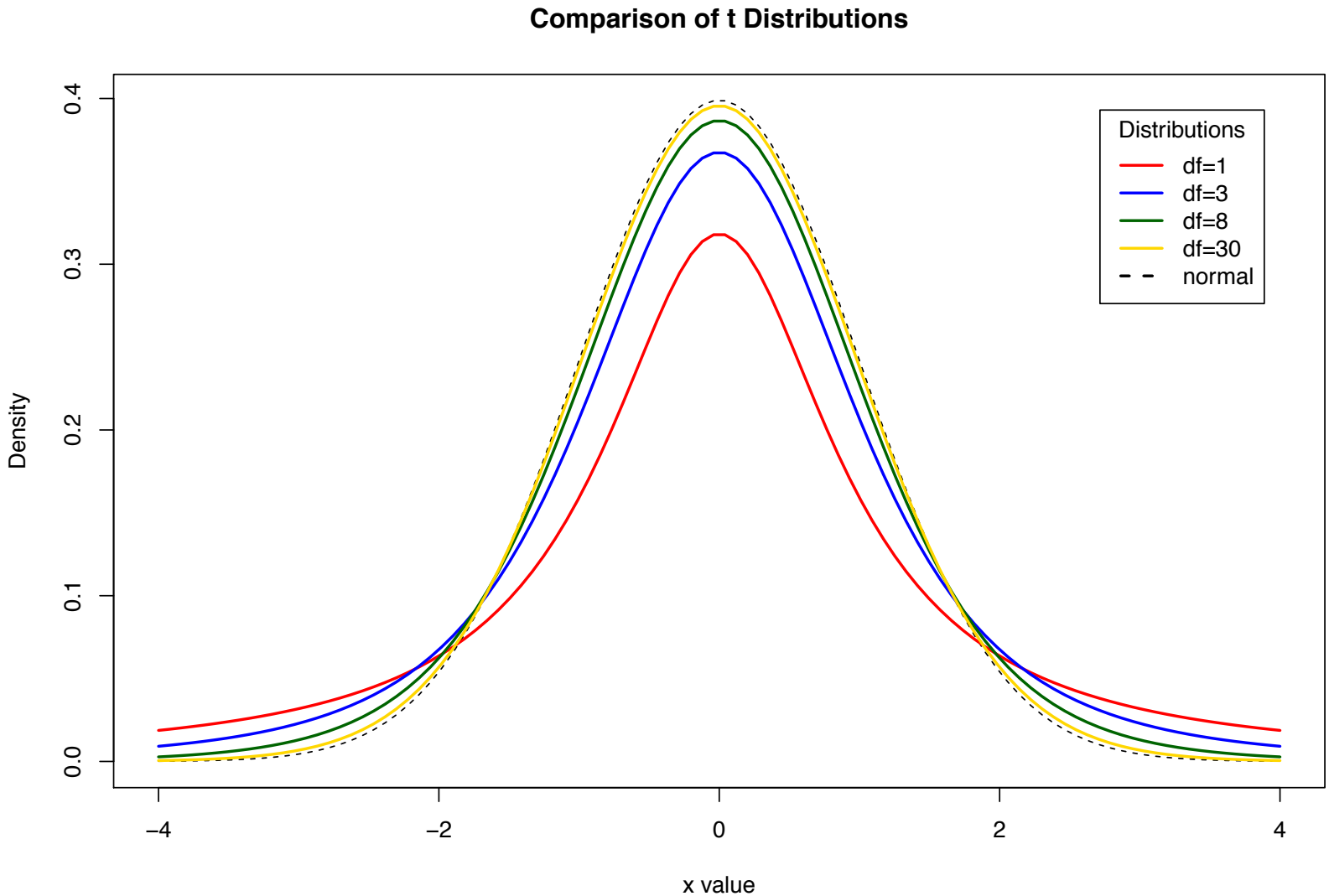
The t Distribution

- If $X \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_n$ then $\frac{X}{\sqrt{W/n}}$ has a t distribution on n degrees of freedom or $\frac{X}{\sqrt{W/n}} \sim t_n$
- Let $X_1, X_2, \ldots$ be an independent sequence of identically distributed random variables that have a $\mathcal{N}(0,1)$ distribution
- Example application: If $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$, then t_{n-1} is the distribution of

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

The t Distribution

The tails are heavier than those for a normal distribution



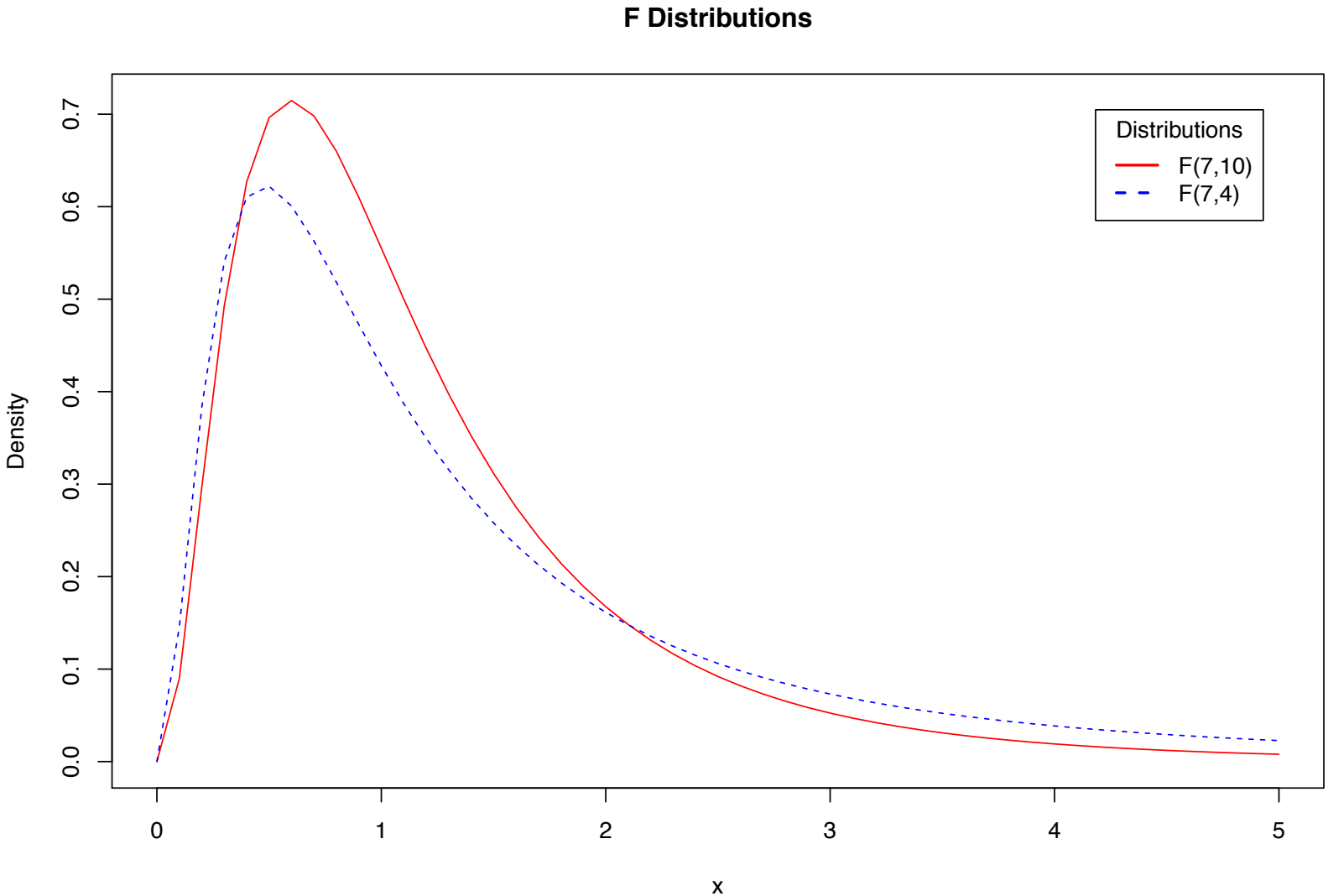
The F Distribution

- Let $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ be independent. Consider

$$W = \frac{X/m}{Y/n}$$

- We say $W \sim F_{m,n}$ where $F_{m,n}$ denotes the F **distribution** on m, n degrees of freedom
- The F distribution is right skewed (see graph on next page)
- For $n > 2, E(W) = n/(n-2)$
- It also follows that the square of a t_n random variable follows an $F_{1,n}$

The F Distribution



Distributions at a glance

Distribution	Description	Example use
χ^2_n	$\sum_n \mathcal{N}^2$	χ^2 test
t_n	$\frac{\mathcal{N}}{\sqrt{\chi^2_n/n}}$	cf. Gaussian means, e.g. t test
$F_{m,n}$	$\frac{\chi^2_m/m}{\chi^2_n/n}$	cf. Gaussian variances, ANOVA/F test

- About this course
- Quick review of Statistics, through an R lens
- **Introduction to Linear Regression**

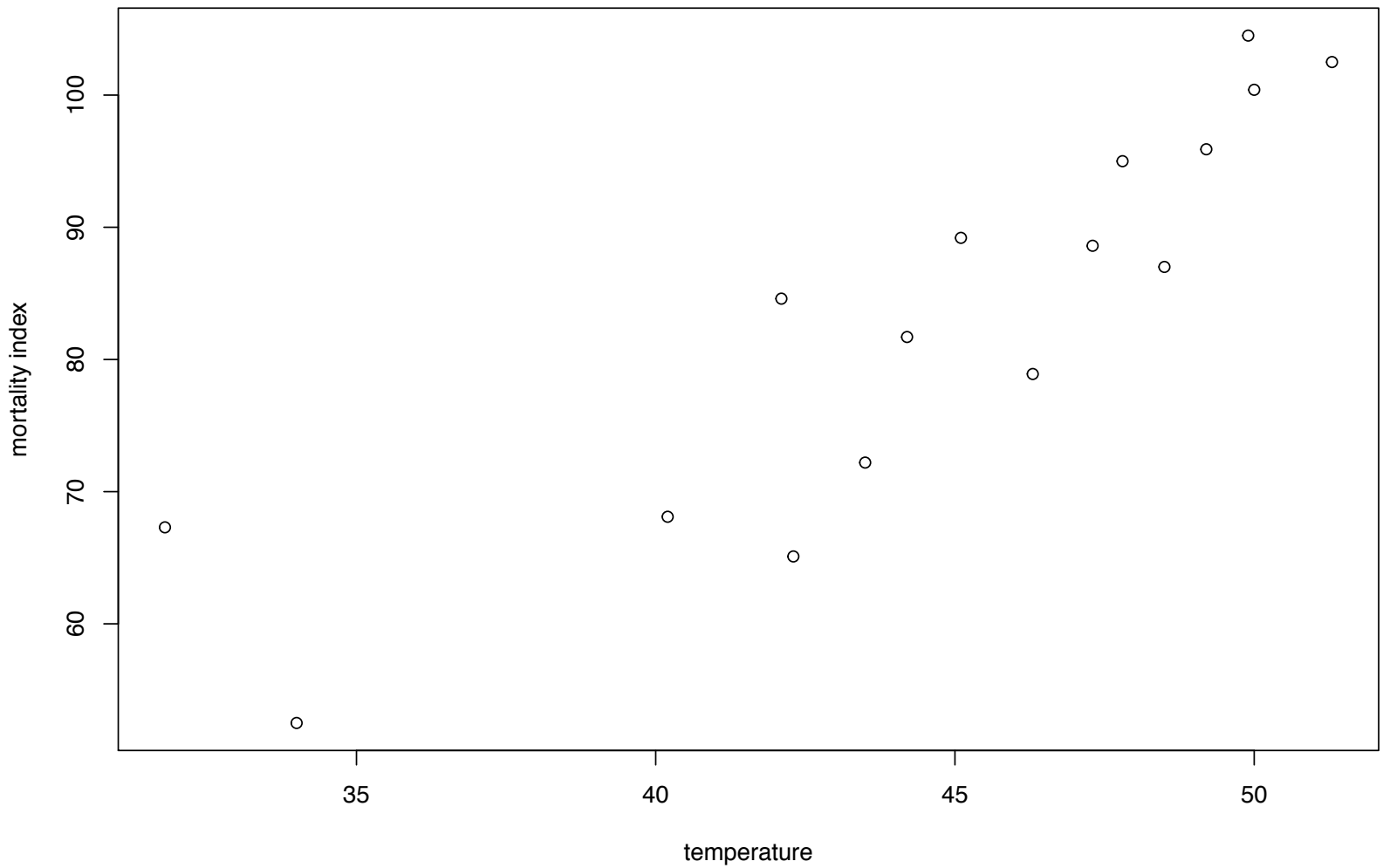
Lea (1965) discussed the relationship between mean annual temperature (^oF) and mortality index (100 = England/Wales) for a type of breast cancer in women taken from regions in northern Europe. (Example from Wu and Hamada, 2009.)

The data are shown below.

```
M <- c(102.5, 104.5, 100.4, 95.9, 87.0, 95.0, 88.6, 89.2,
       78.9, 84.6, 81.7, 72.2, 65.1, 68.1, 67.3, 52.5)
T <- c(51.3, 49.9, 50.0,49.2, 48.5, 47.8, 47.3, 45.1,
       46.3, 42.1, 44.2, 43.5, 42.3, 40.2, 31.8, 34.0)
```

Introduction to Linear Regression

```
plot(T,M,xlab="temperature",ylab="mortality index")
```



Introduction to Linear Regression

A linear regression model of mortality versus temperature is obtained by estimating the intercept and slope in the equation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $i \in \{1, ..., n\}$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

The values of β_0 (y intercept) and β_1 (slope) that minimize the sum of squares

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

are called the least-squares estimators.

We will show in this course that the least squares estimators are given by:

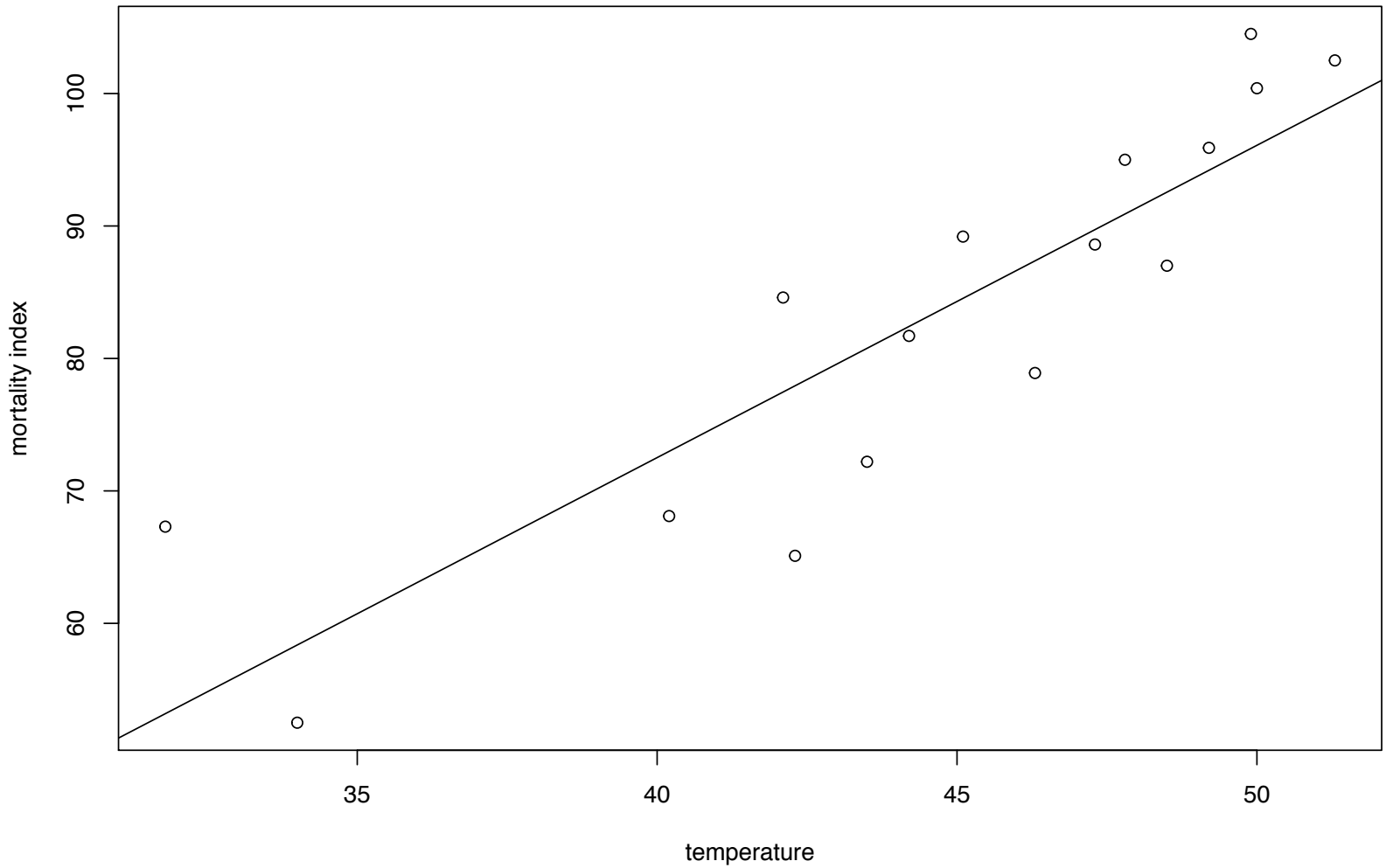
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = r\frac{S_y}{S_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where:

- S_{xx} is the variance of x, and S_{xy} is the covariance of x and y
- r is the correlation between y and x
- S_x and S_y are the sample standard deviations of x and y respectively

Adding a regression line to the plot



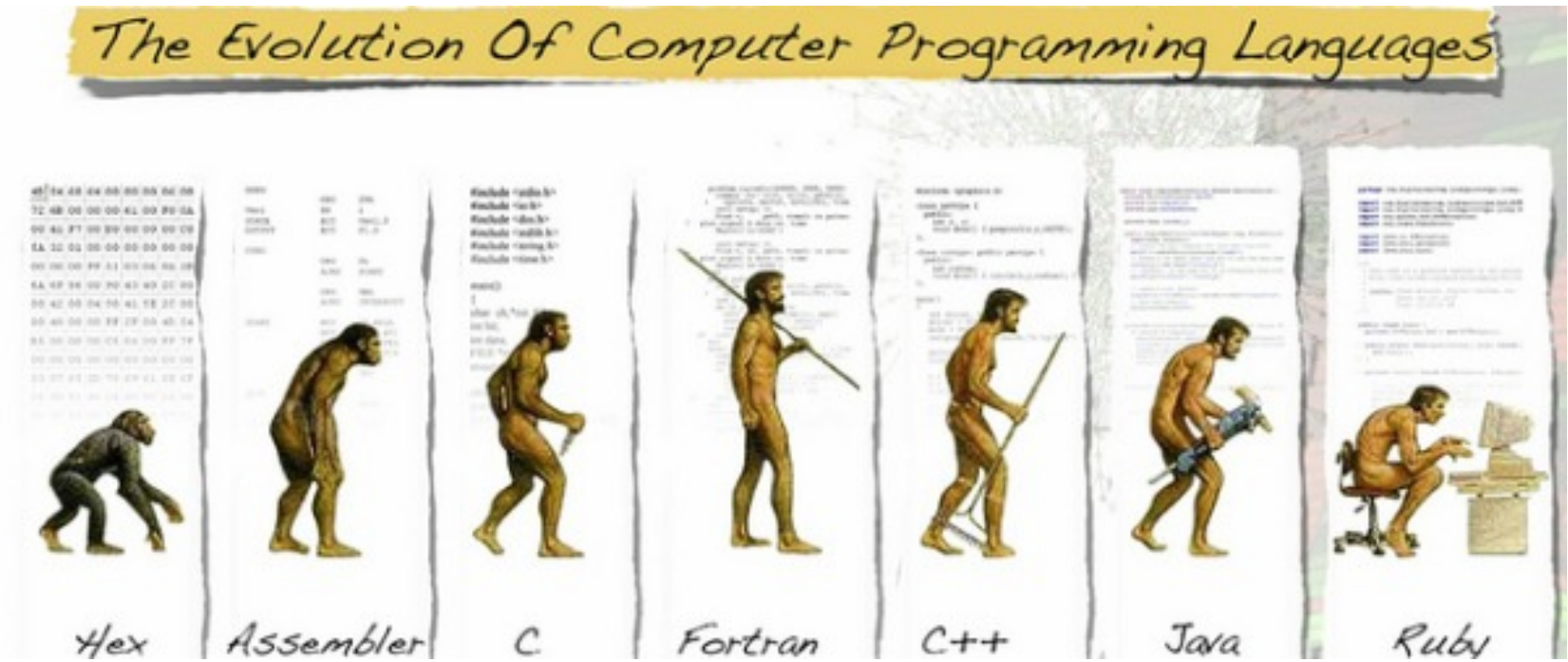
Ok we have a line, but...

As per the preface of our textbook,

> "it makes sense to base inferences or conclusions only on valid models."

Epilogue: A secret benefit to using R is *quick reports*

- And *quick presentations* (shhh...)
- Thanks to RMarkdown
- File extension: .rmd instead of .r



Next steps

- Your questions and comments
- Homework #1 (not for credit) should be available on the course website by Monday
- Over the weekend:
  - Brush up on R if you need to!
    - Check out the resources in the syllabus
    - Can you compile the .R and .Rmd files appearing on Portal soon?
  - Regression etc:
    - Chapter 1 of our textbook is optional but might inspire you