

SML 201 – Week 6

John D. Storey

Spring 2016

Contents

Probability and Statistics	3
Roles In Data Science	3
Central Dogma of Inference	4
Data Analysis Without Probability	4
Probability	4
Sample Space	4
Sample Spaces We Consider	5
Mathematical Probability	5
Two Events	5
Conditional Probability	5
Independence	6
Bayes Theorem	6
Random Variables	6
Definition	6
Discrete Random Variables	6
Continuous Random Variables	7
Example: Discrete PMF	7
Example: Discrete CDF	7
Probabilities of Events Via Discrete CDF	8
Example: Continuous PDF	9
Example: Continuous CDF	9
Probabilities of Events Via Continuous CDF	10
Example: Continuous RV Event	11
Note on PMFs and PDFs	11

Sample Vs Population Statistics	11
Expected Value	12
Variance	12
Random Variables in R	12
Discrete RVs	13
Uniform (Discrete)	13
Uniform (Discrete) PMF	13
Uniform (Discrete) in R	14
Binomial	14
Binomial PMF	14
Binomial in R	15
Poisson	15
Poisson PMF	16
Poisson in R	16
Continuous RVs	17
Uniform (Continuous)	17
Uniform (Continuous) PDF	17
Uniform (Continuous) in R	18
Exponential	18
Exponential PDF	19
Exponential in R	19
Normal	20
Normal PDF	20
Normal in R	21
Central Limit Theorem	22
Linear Transformation of a RV	22
Sums of Random Variables	22
Means of Random Variables	22
Statement of the CLT	22
Example: Calculations	23
Example: Plot	23

Statistical Inference	24
Data Collection as a Probability	24
Parameters and Statistics	24
Sampling Distribution	25
Example: Fair Coin?	25
Example (cont'd)	25
Example (cont'd)	25
Central Dogma of Inference	27
Extras	27
License	27
Source Code	27
Session Information	27

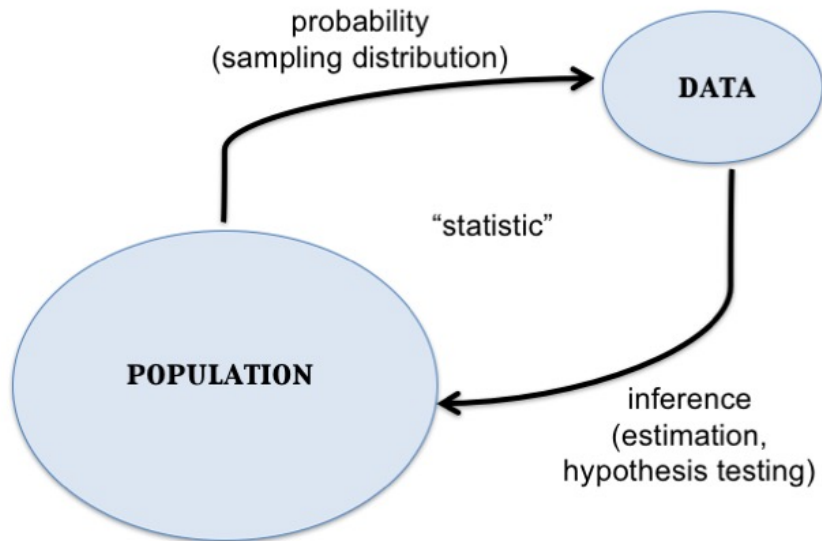
Probability and Statistics

Roles In Data Science

Probability and statistics are required in data science whenever the goals include:

1. Characterizing randomness or “noise” in the data
2. Quantifying uncertainty in models we build or decisions we make from the data
3. Predicting future observations or decisions in the face of uncertainty

Central Dogma of Inference



Data Analysis Without Probability

It is possible to do data analysis without probability and formal statistical inference.

Exploratory data analysis and visualization tend to not involve probability or formal statistical inference.

Important problems machine learning do not involve probability or statistical inference.

Recall Figure 2.1 from *Elements of Data Analytic Style* (shown in Week 1).

Probability

Sample Space

- The **sample space** S is the set of all **outcomes**
- We are interested in calculating probabilities on relevant subsets of this space, called **events**: $A \subseteq S$
- Examples —

- Two coin flips: $S = \{HH, HT, TH, TT\}$
- Netflix movie rating: $S = \{1, 2, 3, 4, 5\}$
- Number of lightning strikes on campus: $S = \{0, 1, 2, 3, \dots\}$
- Height of adult humans in meters: $S = [0, \infty)$

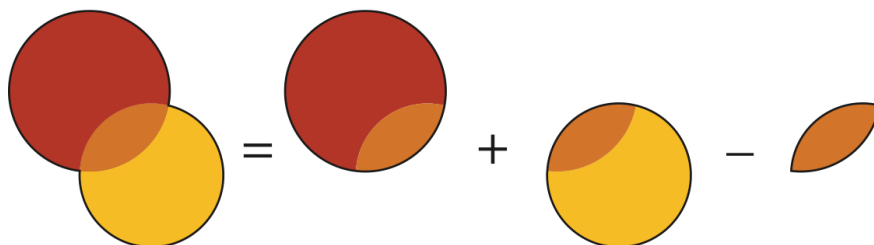
Sample Spaces We Consider

- $S = \{1, 2, \dots, n\}$
- $S = \{0, 1, 2, 3, \dots\}$
- $S = [0, \infty)$
- $S = \mathbb{R} = (-\infty, \infty)$

Mathematical Probability

1. The probability of any even A is such that $0 \leq \Pr(A) \leq 1$
2. If S is the sample space then $\Pr(S) = 1$
3. Let A^c be all outcomes from S that are not in A (called the *complement*); then $\Pr(A) + \Pr(A^c) = 1$
4. If A and $B = \emptyset$, then $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$, where \emptyset is the empty set

Two Events



$$\Pr[A \text{ or } B] = \Pr[A] + \Pr[B] - \Pr[A \text{ and } B]$$

Conditional Probability

An important calculation in probability and statistics is the conditional probability. We can consider the probability of an event A , conditional on the fact that we are restricted to be within event B . This is defined as:

$$\Pr(A|B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)}$$

Independence

Two events A and B by definition independent when:

- $\Pr(A|B) = \Pr(A)$
- $\Pr(B|A) = \Pr(B)$
- $\Pr(A \text{ and } B) = \Pr(A)\Pr(B)$

All three of these are equivalent.

Bayes Theorem

A common approach in statistics is to obtain a conditional probability of two events through the opposite conditional probability and their marginal probability. This is called Bayes Theorem:

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

Random Variables

Definition

We will define a **random variable** (rv) to be a variable that takes values according to a probability distribution.

We define a random variable through its **probability mass function** (pmf) for discrete rv's or its **probability density function** (pdf) for continuous rv's. We can also define it through its **cumulative distribution function** (cdf). The pmf/pdf determines the cdf, and vice versa.

Note: There's a more technical and rigorous definition of a rv, but it does not affect how we will use random variables.

Discrete Random Variables

A discrete rv X takes on a discrete set of values such as $S = \{1, 2, \dots, n\}$ or $S = \{0, 1, 2, 3, \dots\}$.

Its distribution is characterized by its pmf:

$$f(x) = \Pr(X = x) \text{ for } x \in S.$$

Its cdf is:

$$F(y) = \Pr(X \leq y) = \sum_{x \leq y} \Pr(X = x).$$

Continuous Random Variables

A continuous rv X takes on a continuous set of values such as $S = [0, \infty)$ or $S = \mathbb{R} = (-\infty, \infty)$.

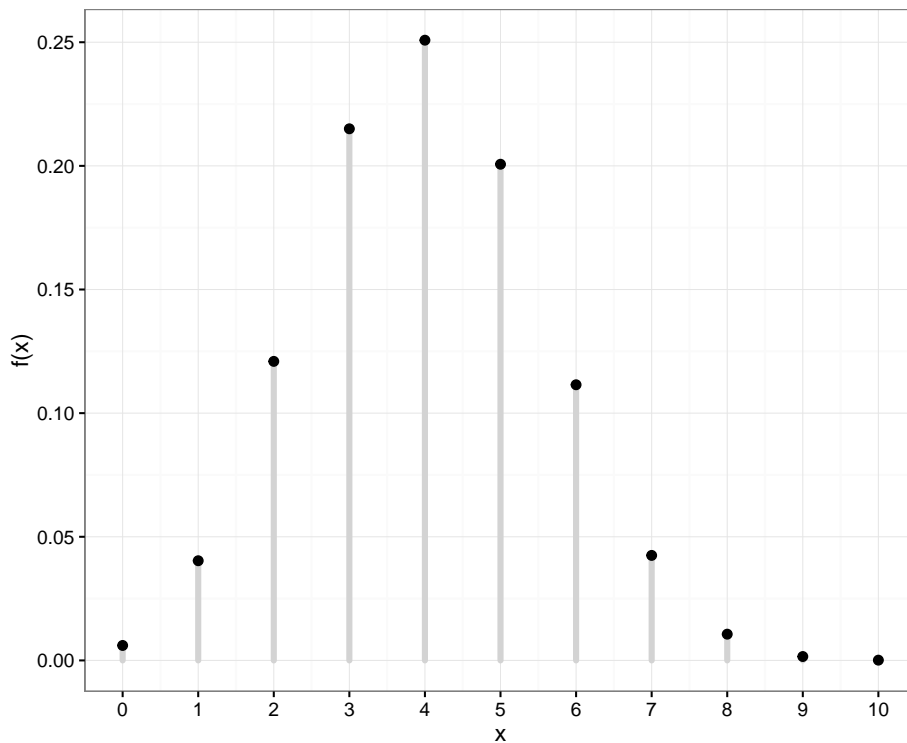
The probability X takes on any specific value is 0. Its pdf $f(x)$ therefore gives an infinitesimal, local, relative probability.

For $S = (-\infty, \infty)$, its cdf is:

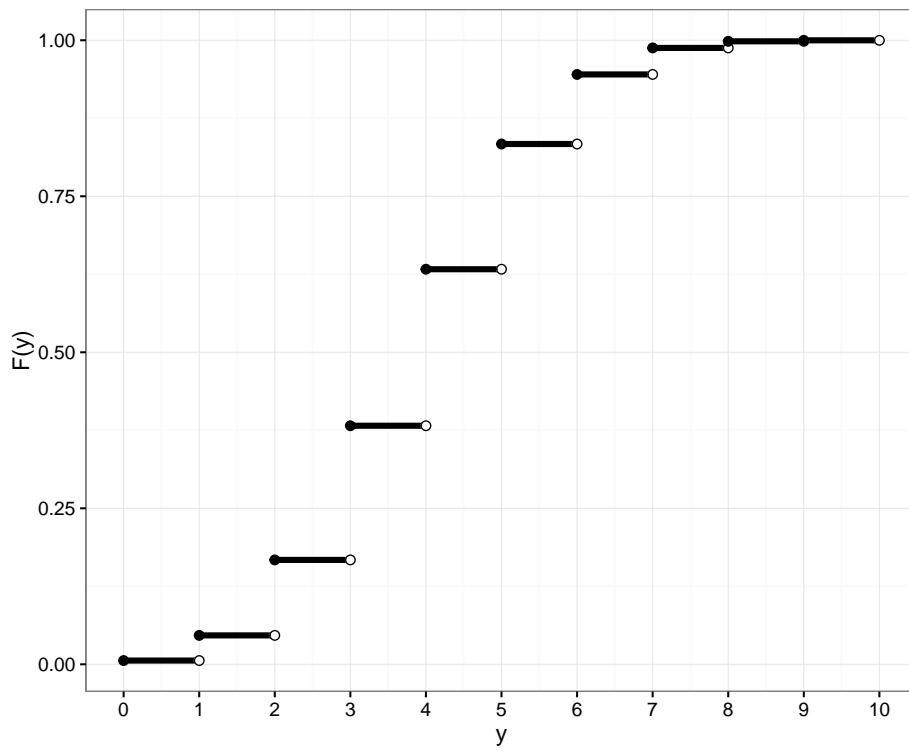
$$F(y) = \Pr(X \leq y) = \int_{-\infty}^y f(x)dx.$$

When $S = [0, \infty)$, the integral starts at 0.

Example: Discrete PMF



Example: Discrete CDF

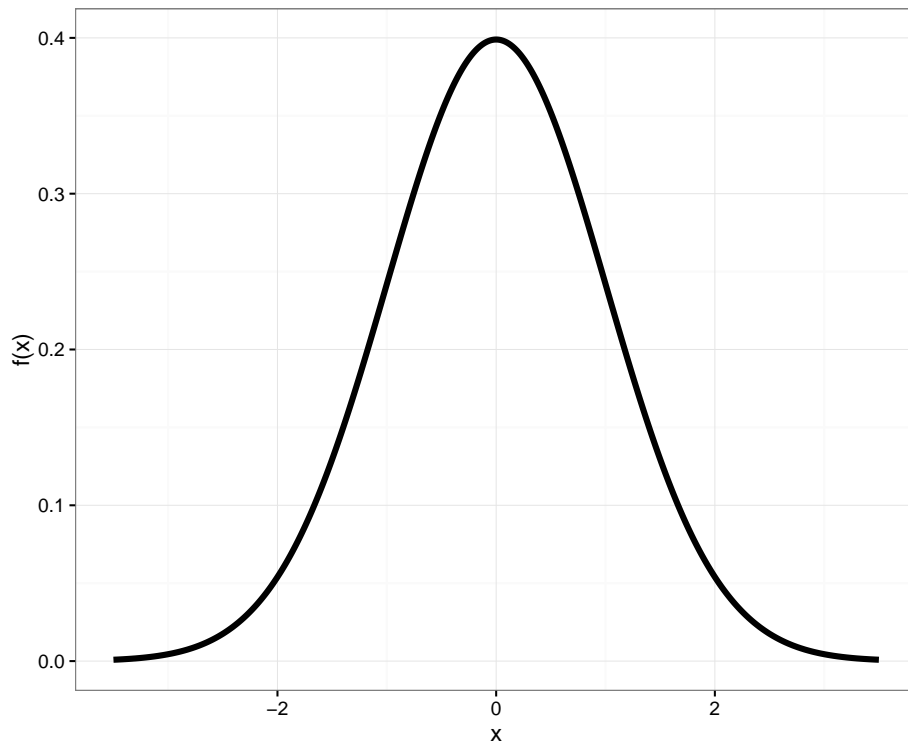


Probabilities of Events Via Discrete CDF

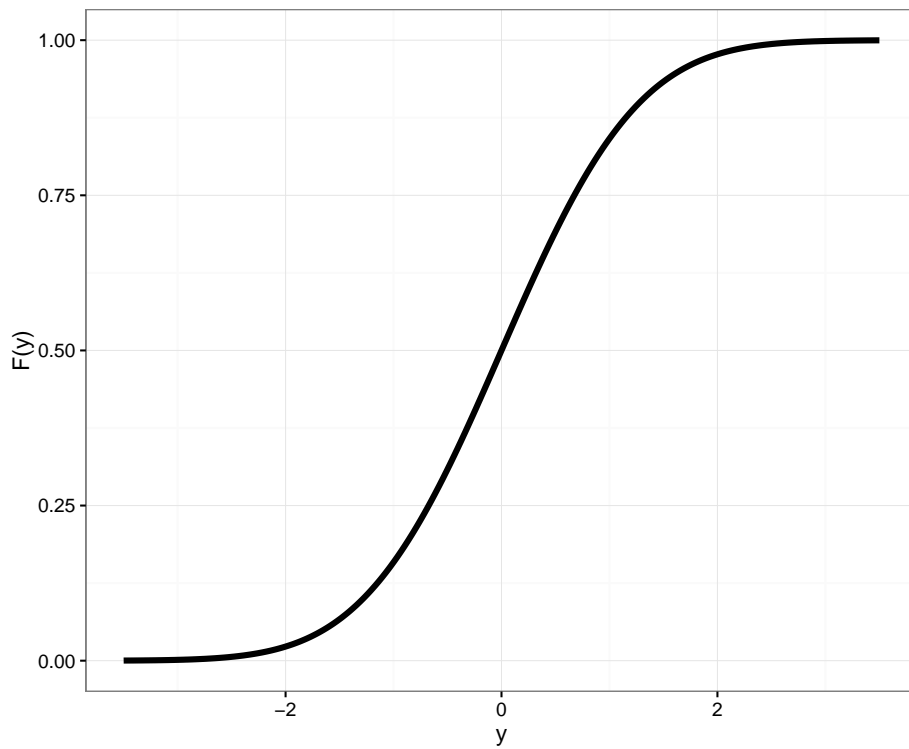
Examples:

Probability	CDF	PMF
$\Pr(X \leq b)$	$F(b)$	$\sum_{x \leq b} f(x)$
$\Pr(X \geq a)$	$1 - F(a - 1)$	$\sum_{x \geq a} f(x)$
$\Pr(X > a)$	$1 - F(a)$	$\sum_{x > a} f(x)$
$\Pr(a \leq X \leq b)$	$F(b) - F(a - 1)$	$\sum_{a \leq x \leq b} f(x)$
$\Pr(a < X \leq b)$	$F(b) - F(a)$	$\sum_{a < x \leq b} f(x)$

Example: Continuous PDF



Example: Continuous CDF

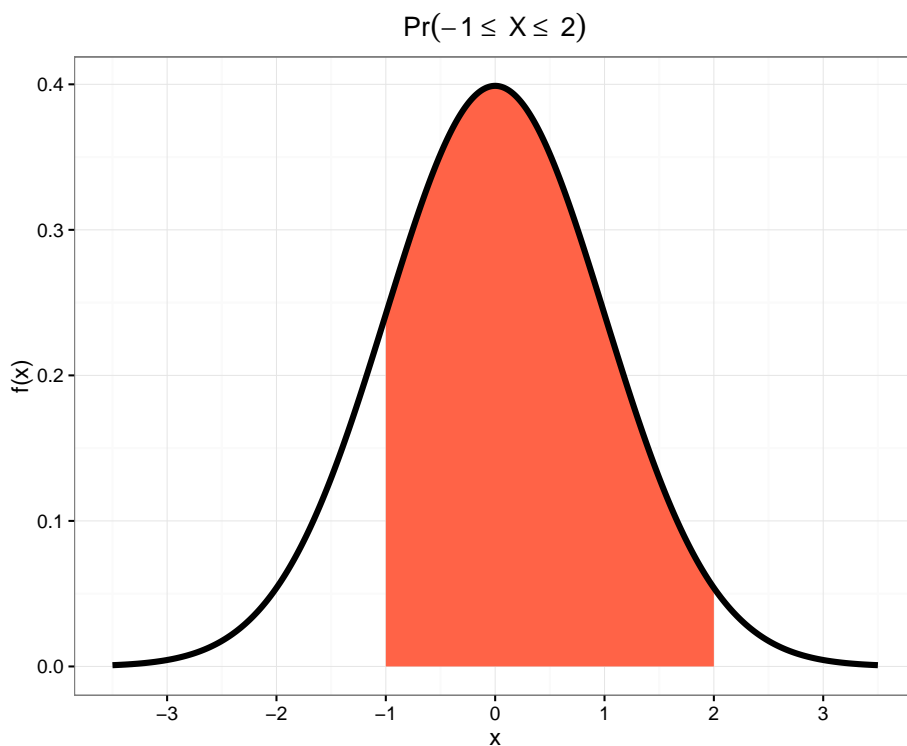


Probabilities of Events Via Continuous CDF

Examples:

Probability	CDF	PDF
$\Pr(X \leq b)$	$F(b)$	$\int_{-\infty}^b f(x)dx$
$\Pr(X \geq a)$	$1 - F(a)$	$\int_a^{\infty} f(x)dx$
$\Pr(X > a)$	$1 - F(a)$	$\int_a^{\infty} f(x)dx$
$\Pr(a \leq X \leq b)$	$F(b) - F(a)$	$\int_a^b f(x)dx$
$\Pr(a < X \leq b)$	$F(b) - F(a)$	$\int_a^b f(x)dx$

Example: Continuous RV Event



Note on PMFs and PDFs

PMFs and PDFs are defined as zero outside of the sample space S . That is:

$$f(x) = 0 \text{ for } x \notin S$$

Also, they sum or integrate to 1:

$$\sum_{x \in S} f(x) = 1$$

$$\int_{x \in S} f(x) dx = 1$$

Sample Vs Population Statistics

We earlier discussed measures of center and spread for a set of data, such as the mean and the variance.

Analogous measures exist for probability distributions.

These are two types are distinguished by calling those on data “sample” measures and those on probability distributions “population” measures.

Expected Value

The **expected value**, also called the “population mean”, is a measure of center for a rv. It is calculated in a fashion analogous to the sample mean:

$$E[X] = \sum_{x \in S} x f(x) \quad (\text{discrete})$$

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx \quad (\text{continuous})$$

Variance

The **variance**, also called the “population variance”, is a measure of spread for a rv. It is calculated in a fashion analogous to the sample variance:

$$\text{Var}(X) = E[(X - E[X])^2]; \quad \text{SD}(X) = \sqrt{\text{Var}(X)}$$

$$\text{Var}(X) = \sum_{x \in S} (x - E[X])^2 f(x) \quad (\text{discrete})$$

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx \quad (\text{continuous})$$

Random Variables in R

The pmf/pdf, cdf, quantile function, and random number generator for many important random variables are built into R. They all follow the form, where `<name>` is replaced with the name used in R for each specific distribution:

- `d<name>`: pmf or pdf
- `p<name>`: cdf
- `q<name>`: quantile function or inverse cdf
- `r<name>`: random number generator

To see a list of random variables, type `?Distributions` in R.

Discrete RVs

Uniform (Discrete)

This simple rv distribution assigns equal probabilities to a finite set of values:

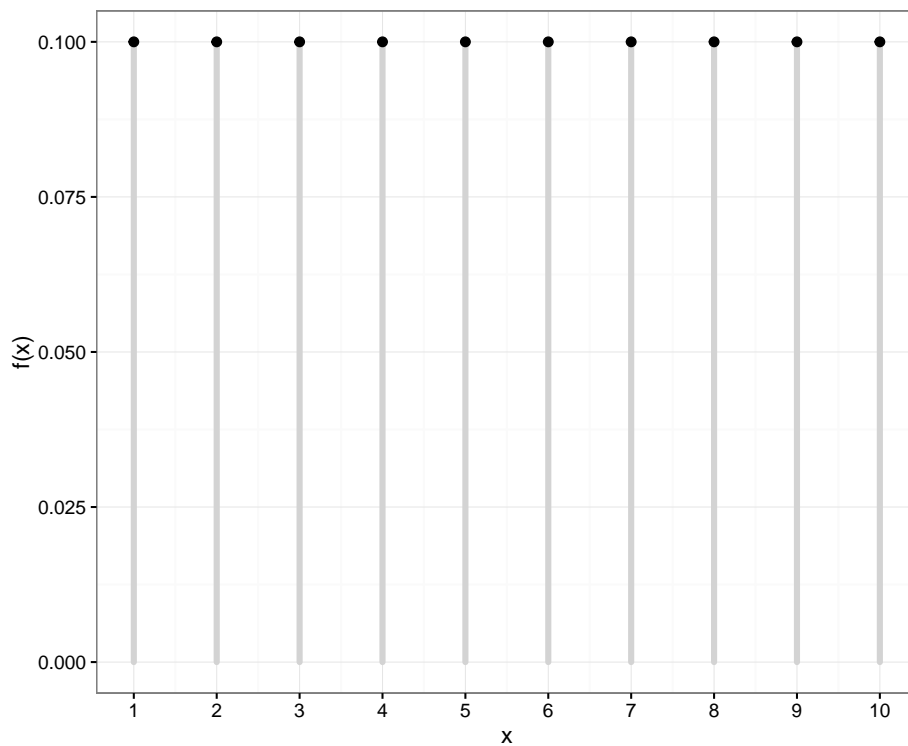
$$X \sim \text{Uniform}\{1, 2, \dots, n\}$$

$$S = \{1, 2, \dots, n\}$$

$$f(x) = 1/n \text{ for } x \in S$$

$$\mathbb{E}[X] = \frac{n+1}{2}, \quad \text{Var}(X) = \frac{n^2-1}{12}$$

Uniform (Discrete) PMF



Uniform (Discrete) in R

There is no family of functions built into R for this distribution since it is so simple. However, it is possible to generate random values via the `sample` function:

```
> n <- 20L
> sample(x=1:n, size=10, replace=TRUE)
[1] 13  4  8  1 13  3  1  8  2  6
>
> x <- sample(x=1:n, size=1e6, replace=TRUE)
> mean(x) - (n+1)/2
[1] -0.005421
> var(x) - (n^2-1)/12
[1] -0.008098145
```

Binomial

$$X \sim \text{Binomial}(n, p)$$

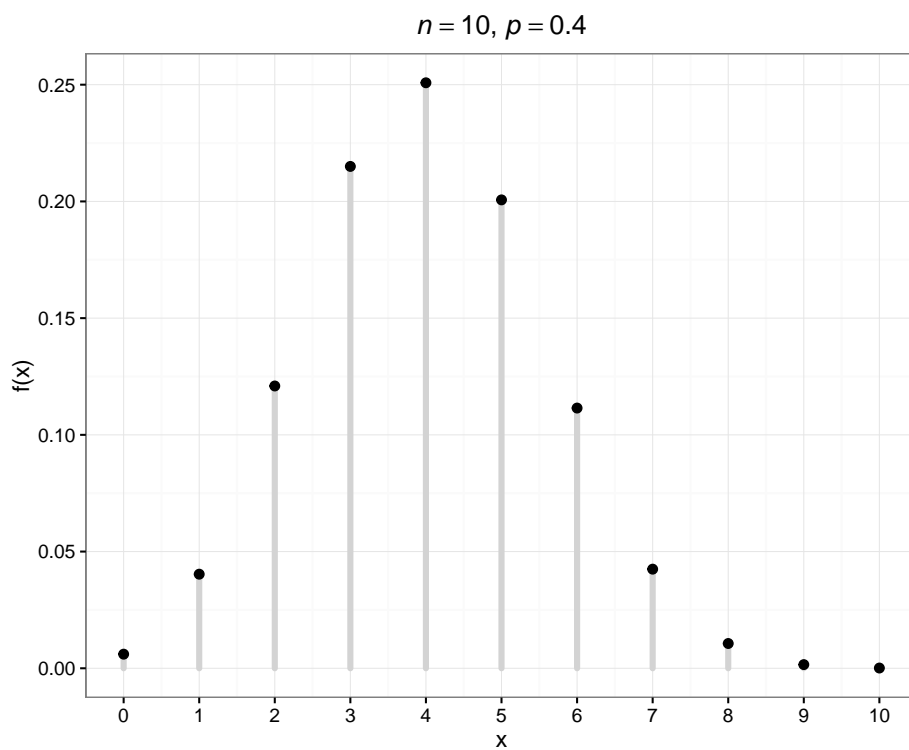
$$S = \{1, 2, \dots, n\}$$

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x \in S$$

$$E[X] = np, \text{ Var}(X) = np(1-p)$$

Note that $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the number of unique ways to choose x items from n without respect to order.

Binomial PMF



Binomial in R

```
> str(dbinom)
function (x, size, prob, log = FALSE)
```

```
> str(pbinom)
function (q, size, prob, lower.tail = TRUE, log.p = FALSE)
```

```
> str(qbinom)
function (p, size, prob, lower.tail = TRUE, log.p = FALSE)
```

```
> str(rbinom)
function (n, size, prob)
```

Poisson

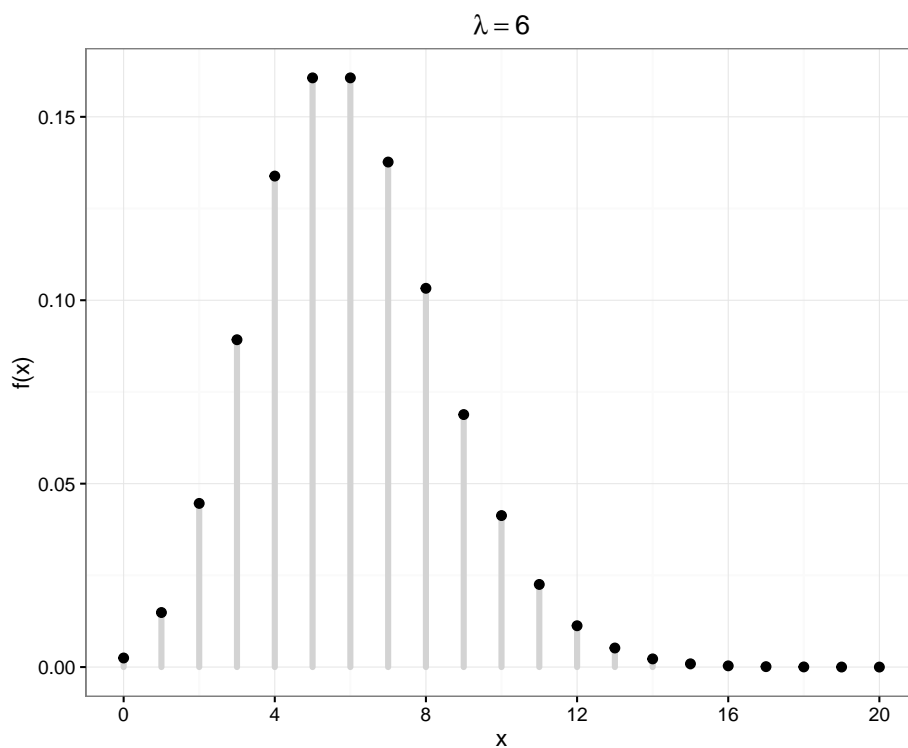
$$X \sim \text{Poisson}(\lambda)$$

$$S = \{0, 1, 2, 3, \dots\}$$

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x \in S$$

$$E[X] = \lambda, \text{ Var}(X) = \lambda$$

Poisson PMF



Poisson in R

```
> str(dpois)
function (x, lambda, log = FALSE)
```



```
> str(ppois)
function (q, lambda, lower.tail = TRUE, log.p = FALSE)
```

```
> str(qpois)
function (p, lambda, lower.tail = TRUE, log.p = FALSE)
```

```
> str(rpois)
function (n, lambda)
```

Continuous RVs

Uniform (Continuous)

$$X \sim \text{Uniform}(0, 1)$$

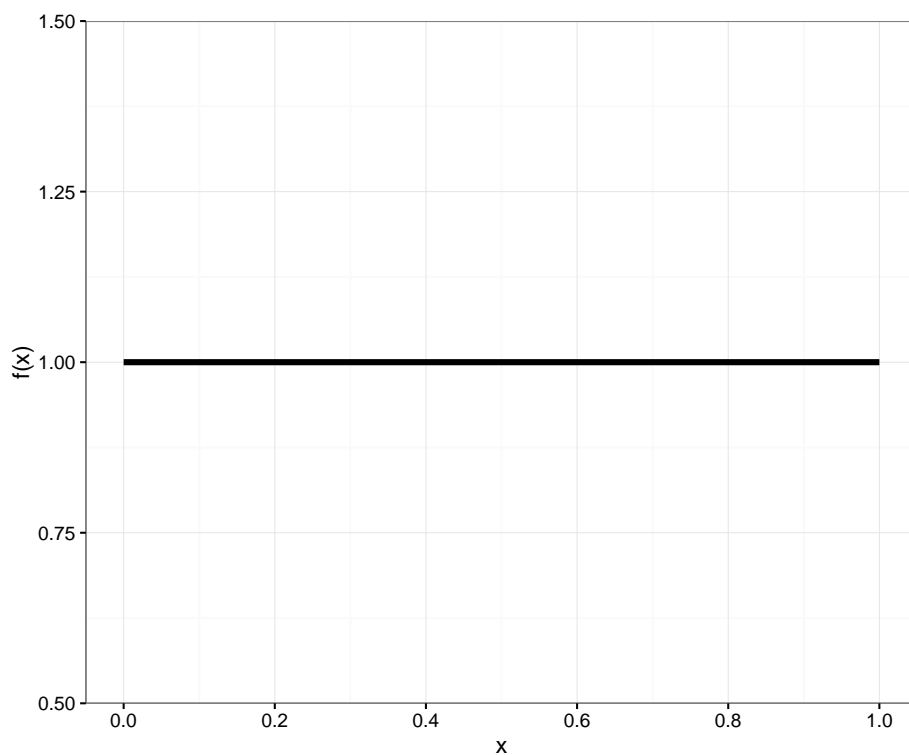
$$S = [0, 1]$$

$$f(x) = 1 \text{ for } x \in S$$

$$F(y) = y \text{ for } y \in S$$

$$E[X] = 1/2, \text{ Var}(X) = 1/12$$

Uniform (Continuous) PDF



Uniform (Continuous) in R

```
> str(dunif)
function (x, min = 0, max = 1, log = FALSE)
```

```
> str(punif)
function (q, min = 0, max = 1, lower.tail = TRUE, log.p = FALSE)
```

```
> str(qunif)
function (p, min = 0, max = 1, lower.tail = TRUE, log.p = FALSE)
```

```
> str(runif)
function (n, min = 0, max = 1)
```

Exponential

$$X \sim \text{Exponential}(\lambda)$$

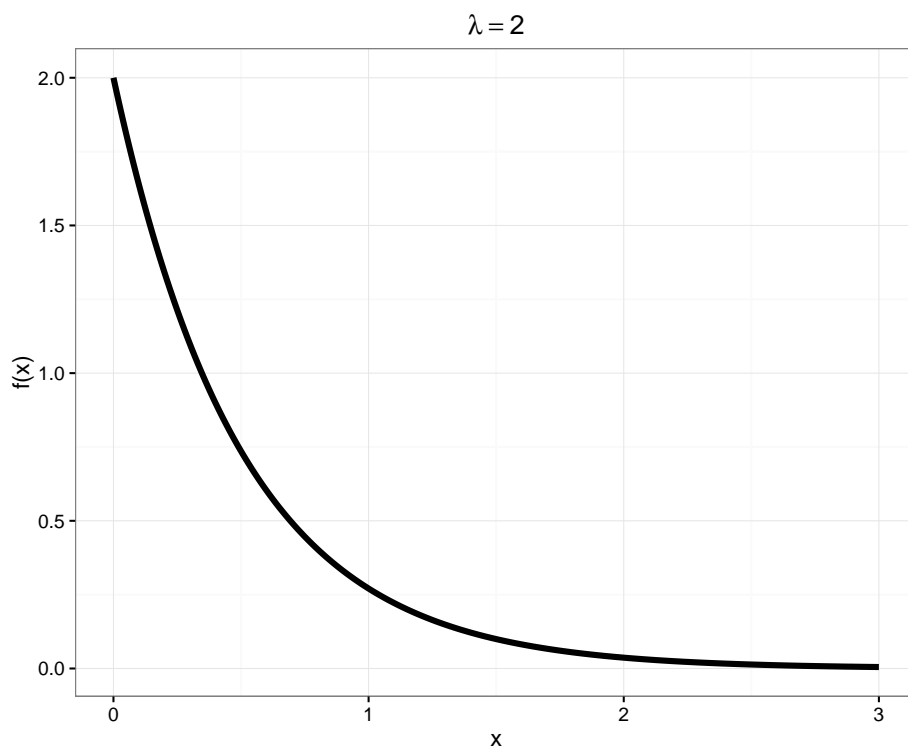
$$S = [0, \infty)$$

$$f(x) = \lambda e^{-\lambda x} \text{ for } x \in S$$

$$F(y) = 1 - e^{-\lambda y} \text{ for } y \in S$$

$$E[X] = \frac{1}{\lambda}, \text{ Var}(X) = \frac{1}{\lambda^2}$$

Exponential PDF



Exponential in R

```
> str(dexp)
function (x, rate = 1, log = FALSE)
```

```
> str(pexp)
function (q, rate = 1, lower.tail = TRUE, log.p = FALSE)
```

```
> str(qexp)
function (p, rate = 1, lower.tail = TRUE, log.p = FALSE)
```

```
> str(rexp)
function (n, rate = 1)
```

Normal

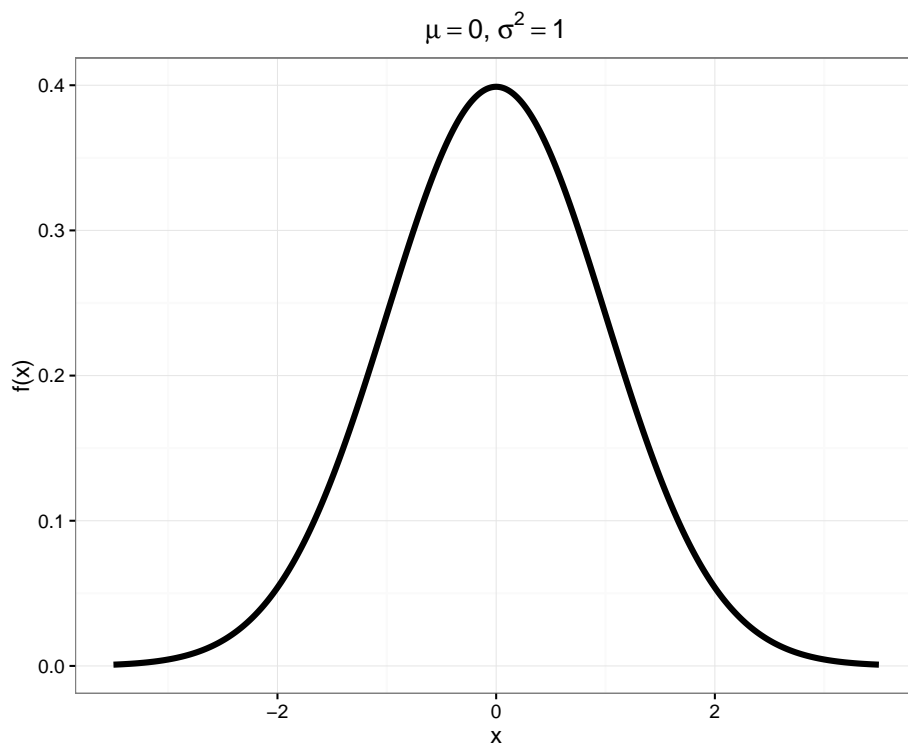
$$X \sim \text{Normal}(\mu, \sigma^2)$$

$$S = (-\infty, \infty)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } x \in S$$

$$E[X] = \mu, \text{ Var}(X) = \sigma^2$$

Normal PDF



Normal in R

```
> str(dnorm) #notice it requires the STANDARD DEVIATION, not the variance
function (x, mean = 0, sd = 1, log = FALSE)
```

```
> str(pnorm)
function (q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
```

```
> str(qnorm)
function (p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
```

```
> str(rnorm)
function (n, mean = 0, sd = 1)
```

Central Limit Theorem

Linear Transformation of a RV

Suppose that X is a random variable and that a and b are constants. Then:

$$\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$$

$$\text{Var}(a + bX) = b^2\text{Var}(X)$$

Sums of Random Variables

If X_1, X_2, \dots, X_n are independent and identically distributed (iid) random variables, then:

$$\mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i]$$

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$$

Means of Random Variables

Suppose X_1, X_2, \dots, X_n are independent and identically distributed (iid) random variables. Let $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ be their sample mean. Then:

$$\mathbb{E}[\bar{X}] = \mathbb{E}[X_i]$$

$$\text{Var}(\bar{X}) = \frac{1}{n} \text{Var}(X_i)$$

Statement of the CLT

Suppose X_1, X_2, \dots, X_n are iid rv's with population mean $\mathbb{E}[X_i] = \mu$ and variance $\text{Var}(X_i) = \sigma^2$.

Then for “large n ”, $\sqrt{n}(\bar{X} - \mu)$ approximately follows the $\text{Normal}(0, \sigma^2)$ distribution.

As $n \rightarrow \infty$, this approximation becomes exact.

Example: Calculations

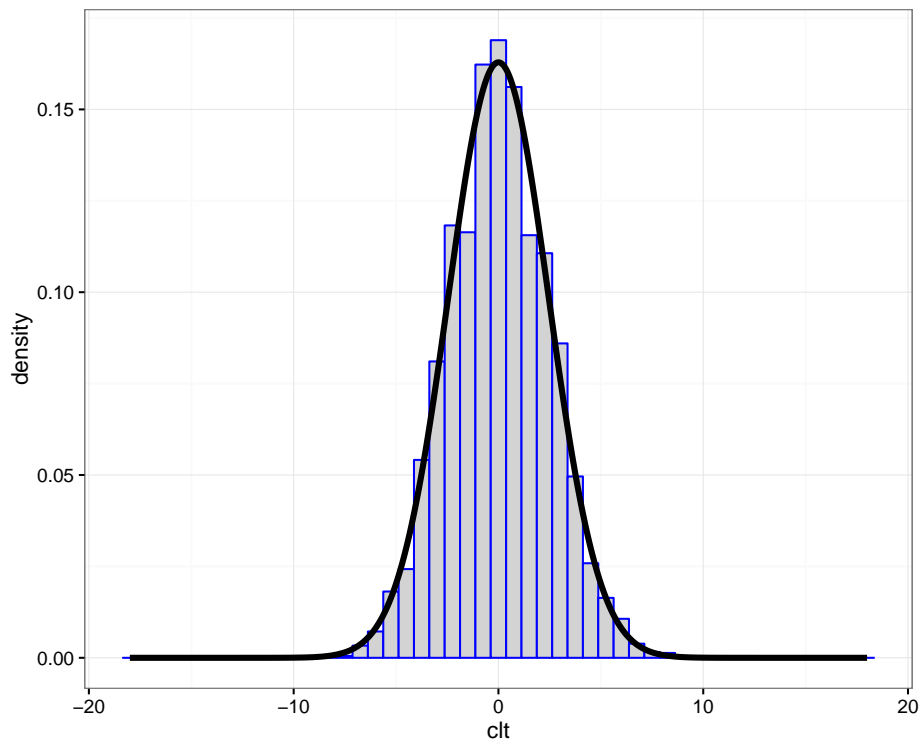
Let X_1, X_2, \dots, X_{40} be iid $\text{Poisson}(\lambda)$ with $\lambda = 6$.

We will form $\sqrt{40}(\bar{X} - 6)$ over 10,000 realizations and compare their distribution to a $\text{Normal}(0, 6)$ distribution.

```
> x <- replicate(n=1e4, expr=rpois(n=40, lambda=6),
+               simplify="matrix")
> x_bar <- apply(x, 2, mean)
> clt <- sqrt(40)*(x_bar - 6)
>
> df <- data.frame(clt=clt, x = seq(-18,18,length.out=1e4),
+                 y = dnorm(seq(-18,18,length.out=1e4),
+                           sd=sqrt(6)))
```

Example: Plot

```
> ggplot(data=df) +
+   geom_histogram(aes(x=clt, y=..density..), color="blue",
+                 fill="lightgray", binwidth=0.75) +
+   geom_line(aes(x=x, y=y), size=1.5)
```



Statistical Inference

Data Collection as a Probability

- Suppose data are collected in such a way that it is randomly observed according to a probability distribution
- If that probability distribution can be parameterized, then it is possible that the parameters describe key characteristics of the population of interest
- **Statistical inference** reverse engineers this process to estimate the unknown values of the parameters and express a measure of uncertainty about these estimates

Parameters and Statistics

- A **parameter** is a number that describes a population
 - A parameter is often a fixed number
 - We usually do not know its value

- A **statistic** is a number calculated from a sample of data
- A statistic is used to estimate a parameter

Sampling Distribution

The **sampling distribution** of a statistic is the probability distribution of the statistic under repeated realizations of the data from the assumed data generating probability distribution.

The sampling distribution is how we connect an observed statistic to the population.

Example: Fair Coin?

Suppose I claim that a specific coin is fair, i.e., that it lands on heads or tails with equal probability.

I flip it 20 times and it lands on heads 16 times.

1. My data is $x = 16$ heads out of $n = 20$ flips.
2. My data generation model is $X \sim \text{Binomial}(20, p)$.
3. I form the statistic $\hat{p} = 16/20$ as an estimate of p .

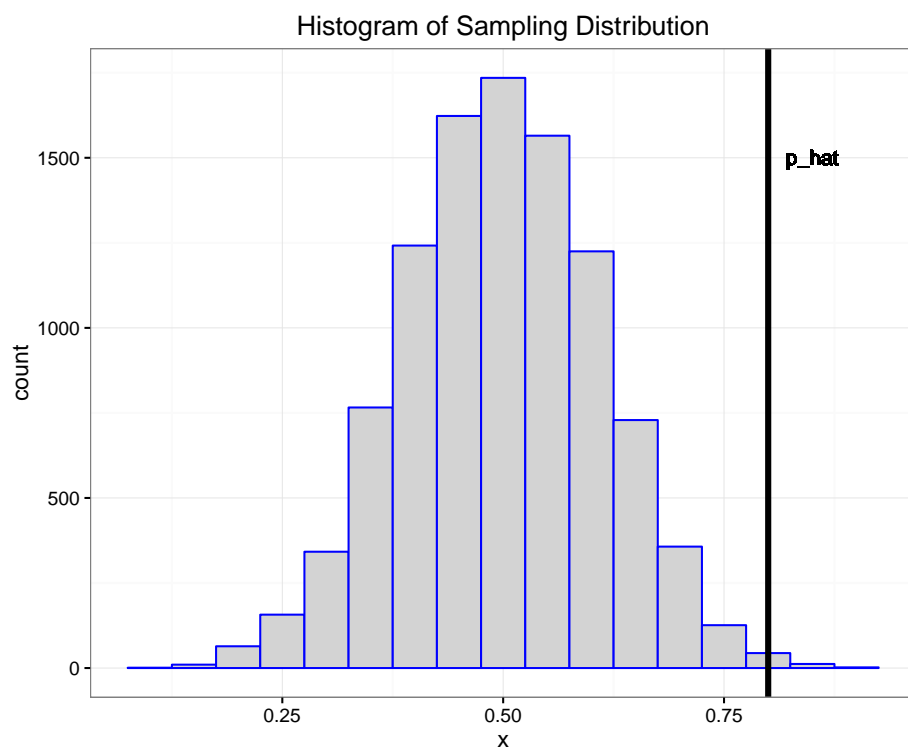
Example (cont'd)

Let's simulate 10,000 times what my estimate would look like if $p = 0.5$ and I repeated the 20 coin flips over and over.

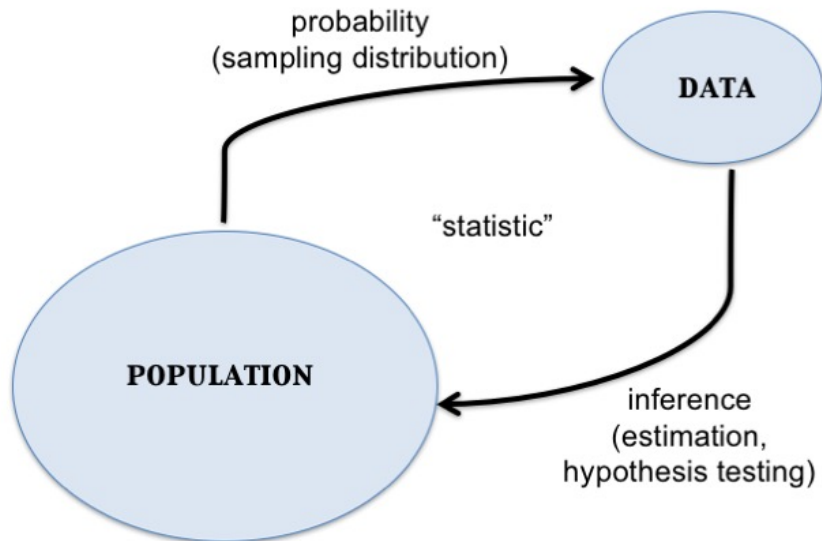
```
> x <- replicate(n=1e4, expr=rbinom(1, size=20, prob=0.5))
> sim_p_hat <- x/20
> my_p_hat <- 16/20
```

What can I do with this information?

Example (cont'd)



Central Dogma of Inference



Extras

License

<https://github.com/SML201/lectures/blob/master/LICENSE.md>

Source Code

<https://github.com/SML201/lectures/tree/master/week6>

Session Information

```
> sessionInfo()  
R version 3.2.3 (2015-12-10)  
Platform: x86_64-apple-darwin13.4.0 (64-bit)  
Running under: OS X 10.11.3 (El Capitan)
```

```
locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats      graphics  grDevices  utils      datasets  methods
[7] base

other attached packages:
[1] ggplot2_2.1.0  knitr_1.12.3   magrittr_1.5
[4] devtools_1.10.0

loaded via a namespace (and not attached):
[1] Rcpp_0.12.3      digest_0.6.9    plyr_1.8.3
[4] grid_3.2.3       gtable_0.2.0    formatR_1.2.1
[7] evaluate_0.8     scales_0.4.0    stringi_1.0-1
[10] rmarkdown_0.9.5  labeling_0.3     tools_3.2.3
[13] stringr_1.0.0    munsell_0.4.3   yaml_2.1.13
[16] colorspace_1.2-6 memoise_1.0.0    htmltools_0.3
```