QCB 408 / 508 – Notes on Week 1

Siena Dumas Ang

2020-03-01

Summary

Week 1: We began by covering general course logistics. We took a look at an introduction to statistics and genomics, including understanding what a genome is, the central dogma of molecular biology, and data types we will use in this course. Using the central dogma of statistical inference, we learned aspects of working with data and defined concepts in probability. Moving to random variables, we looked at continuous and discrete probability distributions and looked at an example with the Hardy-Weinberg equilibrium.

- Course Logistics
- Introduction to statistics and genomics
- Probability
- Random Variables
- Probability Distributions
- Hardy-Weinberg equilibrium

Course Logistics

Contacts

John Storey jstorey@princeton.edu Office Hours: Mondays, 12:20-1:20

Andrew Bass ajbass@princeton.edu Office Hours: Tuesdays, 1:00-3:00

Important Sites:

- Blackboard
- Piazza
- https://jdstorey.org/fsg/

Scribed Notes

Sign up for one week of the semester.

Homework and Final

Blackboard, used for assignments. 7 free days to spend across HW/Scribed Notes. (Final not included.)

Take-home final and homeworks should be submitted as .Rmd and PDF.

Additional Resources

- YARP - Yet Another R Primer
- R Programming for Data Science

Introduction to Statistics and Genomics

Genomics Introduction

DNA

DNA, deoxyribonuclic acid, is composed of nucleotides

- adenine (A)
- thymine (T)
- cytosine (C)
- guanine (G)

A and T always pair together.

C and G always pair together.

https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics

What is a Genome?

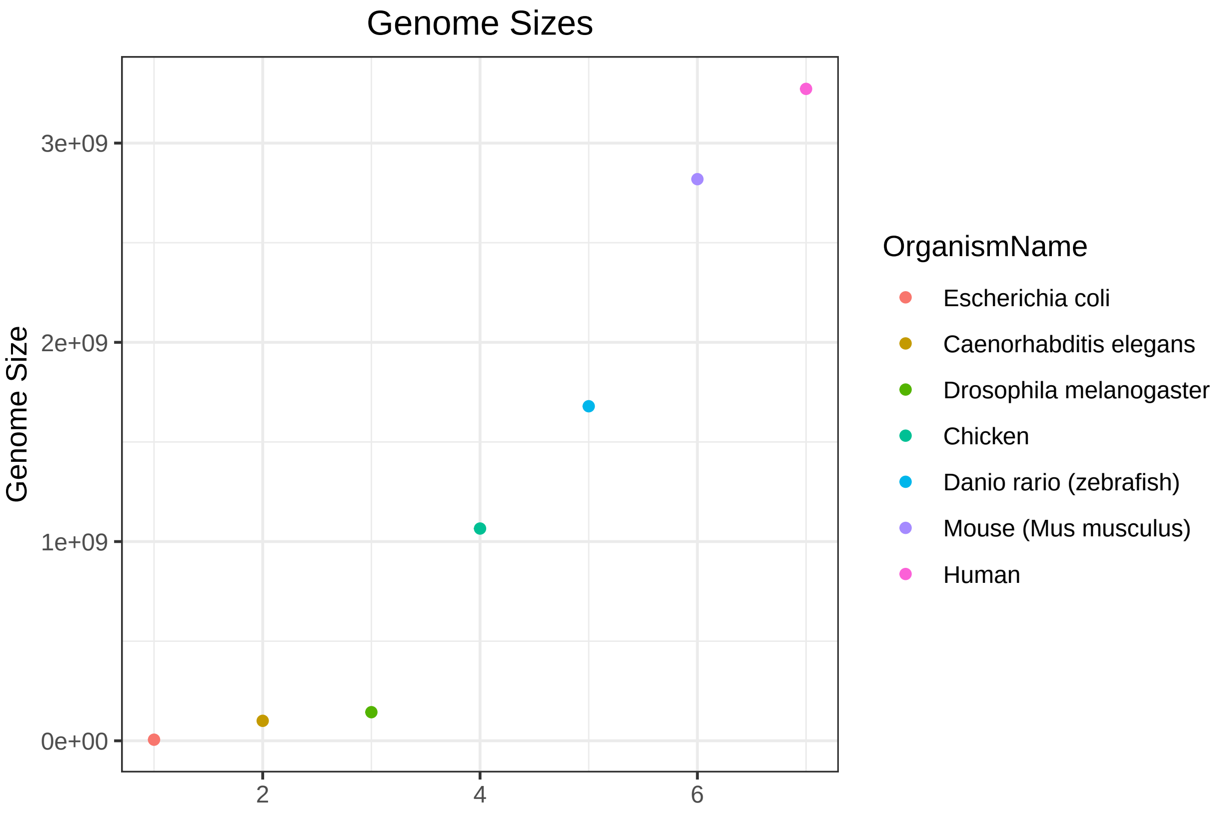
A genome sequence is a base sequence of a single strand (template or reverse complement) of DNA.

"OMICS"

- Genomics is the study of a genome as a whole (both DNA and spatial).
- Transcriptomics is the study of RNA.
- Proteomics is the study of proteins, such as studying protein-protein interactions.

While the human genome is >1B bases, organisms have genomes of varying sizes and complexity.

```
> ecoli <- 5594600
> human <- 3272090000
> drosophila <- 143726000
> chicken <- 1065370000
> zebrafish <- 1679200000
> mouse <- 2818970000
> celegans <- 100286000
>
> genomeinfo <- data.frame("genome" = c(ecoli, human, drosophila, chicken, zebrafish, mouse, celegans),
> genomeinfo <- genomeinfo[order(genomeinfo$genome),]
>
> genomeinfo$rank <- 1:nrow(genomeinfo)
>
> genomeinfo$OrganismName <- factor(genomeinfo$OrganismName, levels=genomeinfo$OrganismName)
>
> ggplot(genomeinfo, aes(x=rank, y=genome, color=OrganismName)) + geom_point() +
+   labs(title="Genome Sizes", x=' ', y='Genome Size')
```



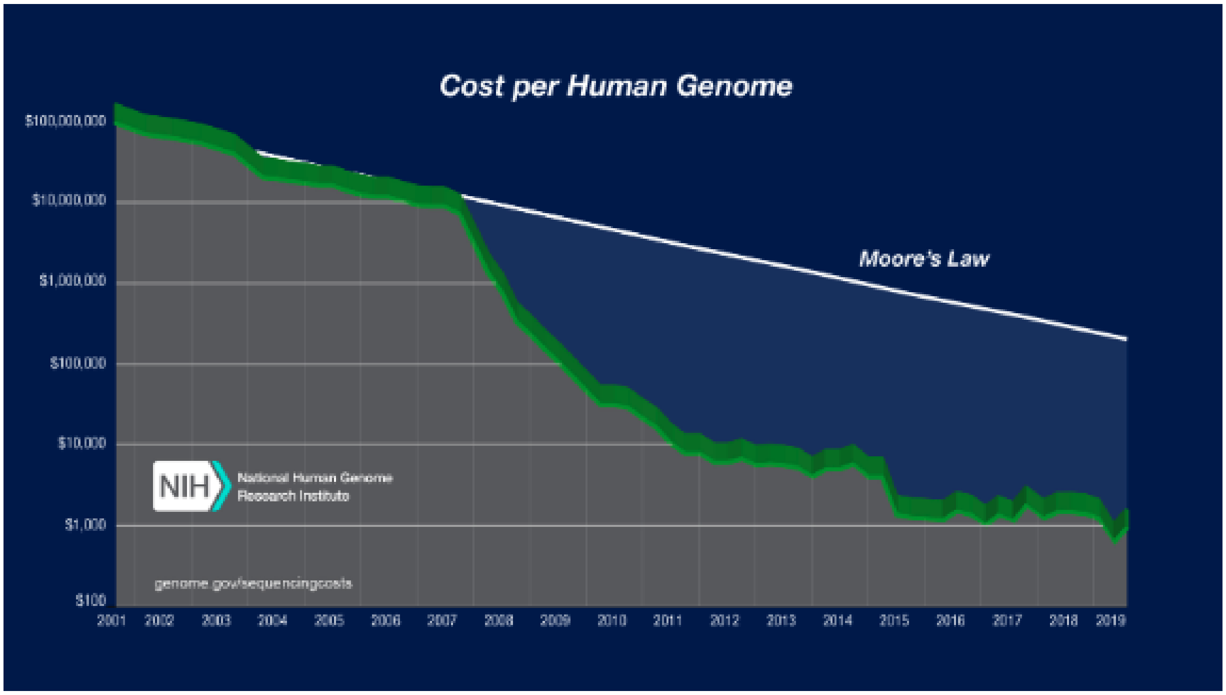
Genome Size Source

DNA Sequencing

The process of reading DNA (and RNA) is called sequencing.

With the rise of next-generation sequencing (NGS), we have been able to sequence genomes in an easy, high-throughput, cost effective way–generating all kinds of exciting data to be analyzed.

Sequencing instruments such as those produced by Illumina (the most widely used technology for DNA sequencing) often use a method called "sequencing by synthesis", appending flourescent bases and reading the flourescent signal to call a base. This has become a cost-effective and highly accurate way to sequence DNA.



Cost per genome data

https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Sequencing usually produces ~150bp (base pair) fragments of DNA (reads), however, new technologies such as PacBio and Oxford Nanopore generate longer reads in exchange for accuracy.

Becase the cost to sequence human genomes has become so reasonable, we are able to investigate widely available sequencing data such as genotype and RNA-seq data in this class.

Motivation

Central Dogma of molecular biology

$$DNA \rightarrow transcription \rightarrow RNA \rightarrow translation \rightarrow Protein$$

From NCBI, "the coded genetic information hard-wired into DNA is transcribed into individual transportable cassettes, composed of messenger RNA (mRNA); each mRNA cassette contains the program for synthesis of a particular protein (or small number of proteins)."

Chromatin: accessibility of DNA

SNP Data

Recall that:

- A & T pair together
- C & G pair together

A single nucleotide polymorphism (SNP) is a single location difference between two genomes. The coding for a SNP is the chromosome and base pair number.

After all samples have been collected and coded, we end up with a matrix of all individuals x SNPs.

Example:

Individual	Count
CC	0
CT	1
TT	2

Here, we count the number of T alleles present. However, the allele that gets counted is arbitrary and can change. For example, one could choose to count the oldest allele or the most common allele.

SNPs can be common (for example, 5%).

SNPs can be rare (where most people only have one allele).

GWAS

A Genome-Wide Association Study (GWAS) associates genetic variation with a disease.

By considering patients with the disease in question and comparing their DNA to individuals without the disease, we allocate SNPs as disease-specific SNPs or non-disease SNPs.

Statistics is involved in many aspects of the pipeline:

- Base calling A/C/T/G on genotyping platform
- Designing the GWAS experiment, for example figuring out how to get enough people
- Doing the analysis, for example which SNPs are of biomedical interest?

Examples of diseases one could consider: - Hyptertension - Height - Alzheimers - Diabetes - Hair Color

Allele frequency in human populations

HGDP project: Human Genome Diversity Project, a large study which aims to understand human genetic diversity.

Median differentiated SNP in HGDP data: in the middle of how much things vary throughout the world.

Due to genetic drift!

- With infinite populations, can go to steady state.
- With finite populations, individuals tend to mate with those near them, etc.

Understanding the distribution of genetic variation gives insight into:

- The migration history of humans
- Global health: understanding health impacts on individuals throughout the world. We should know how genetic variation is distributed.
- Environmental exposures: for example, when the environment predisposes you to a disease and genetic drift lines up with environmental exposure. How do we untangle random (chance) vs. environmental risk for a disease?

In this course, we will look at both finite & randomly mating populations as well as finite population size with drift, considering how to model at genome-wide levels while also taking allele frequencies in the population into account.

RNA-seq

During transcription, mRNA is transcribed off DNA, and different levels of transcription occur.

How to measure transcription:

- Single cell or 1M cells (you can vary this)
- Pull down mRNA (extract and isolate)

- Process (convert to single-stranded probe)
- Fragment
- Sequence
- Figure out which fragment came from which gene
- Quantify (get the number mapped to each gene)

Read depth: Average coverage

Common data types include gene expression data and genome-wide genotype data.

Statistics Introduction

Data Analysis

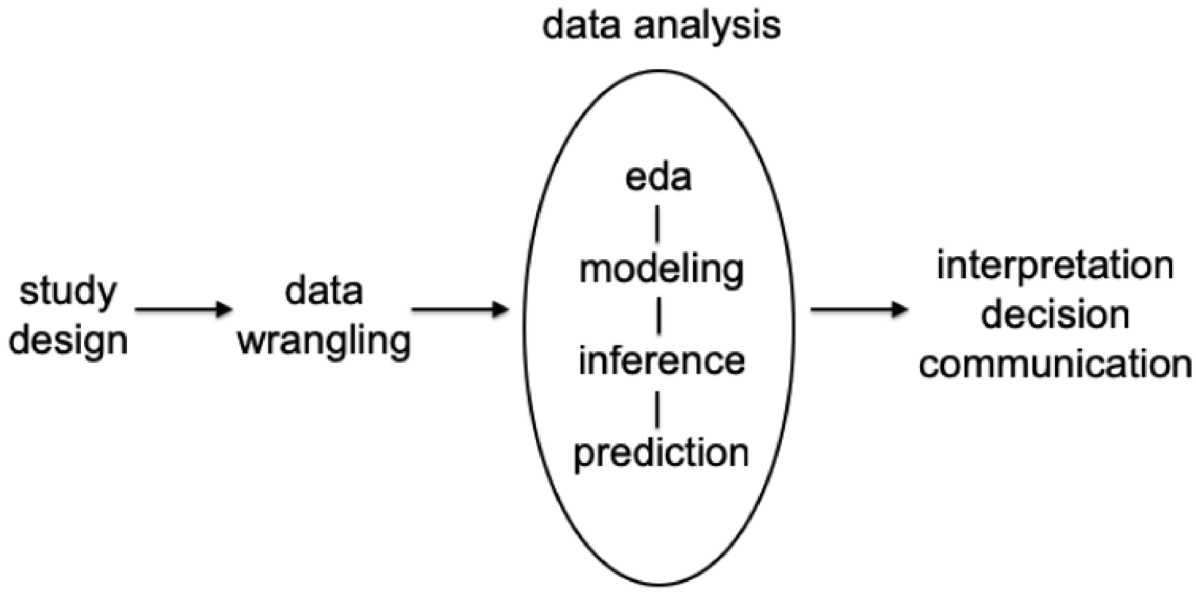


Figure 2.1: Central Dogma of Statistics

Study design

- Design for carrying out the study & collecting data.
- “Study design is an area that is almost solely studied by statisticians and it is one of the core strengths of the field of statistics.” - John Storey, FAS

Data Wrangling

- Get data into a form that can be analyzed in a straightforward format

Data Analysis

- Exploratory data analysis (eda): getting a general idea of the data’s key features (often through summaries and visualizations). Also focused on communication: taking raw datasets and turning it in to pieces of information that can be communicated. An example. This field was pioneered by John Tukey.
- Modeling: Creating a model for the data, for example, modeling height or genotype.
- Inference: Inferring things about the entire population. (Estimation and hypothesis testing.)

- Prediction: Build a model such that if you only partially sample, you can predict what is missing. For example, predicting what someone will buy on Amazon with their click patterns, or predicting what color hair a child will have based on the color of their parents' hair.

Statistics: study design, exploratory data analysis, modeling, and inference.

Machine Learning: Prediction

Data Science: Data wrangling and exploratory data analysis.

Both science and engineering share the fundamental goal of learning from data.

- Science (what we are doing in this course) - understanding the world.
- Engineering - Manipulating the world.

Central dogma of statistical inference

From John Storey, FAS:

> Probabilistic modeling and/or statistical inference are required when 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

We often have inaccessible populations (a data point you can't measure), so we collect a sample.

For example:

- Can't ask everyone to fill out a political survey, so we randomly sample.
- Can't measure the total number of cereal boxes purchased in America, so we randomly sample region by region.
- Can't flip a quarter forever

Data (sample/subset of population) is collected through a mechanism represented by a probability distribution.

For example:

- For a political survey, each person has an equal probability of being selected.
- For cereal box purchases, each store within a region has an equal probability of being selected to track purchases.
- Flip a coin a finite number of times

The probability model (sampling distribution) only works if it accurately represents how the data was generated.

When we just have data and no probability model, we can use EDA to describe the data.

The central dogma of statistical inference relies on having collected data that follows a reasonable probability distribution.

Probability

Definitions

$$Probability\ space: \Omega, \mathcal{F}, \Pr$$

$$\Omega = set\ of\ random\ outcomes, sample\ space$$

$$\Pr = Probability\ measure$$

$$A = any\ subset\ of\ \Omega$$

$$For\ events\ A \subseteq \Omega, can\ calculate\ \Pr(A)$$

For example:

$$\Omega = Heads/Tails$$

$$A = Heads$$

What about continuous sample space? For example, height or rational numbers?

$$\mathcal{F} = \sigma - algebra, all\ events\ A\ where\ \Pr(A)\ is\ meaningful$$

$\sigma - algebra$ in measure theory: $\sigma - algebra$ on a set generated by subsets with well-defined probabilities will also have well defined probabilities.

$\rightarrow$ we won't use it, but powerful mathematics

Examples:

$$Examples\ of\ \Omega$$

- 2 Coin Flips
$$\Omega = \{TT, HH, HT, TH\}$$
- All diploid genotypes at a particular locus
$$\Omega = \{CC, CT, TT\}$$
- Haploid genotypes
$$\Omega = \{C, T\}$$
- Stock returns
$$\Omega = \mathbb{R}$$
- Height
$$\Omega = [0, \infty]$$

Mathematical Probability

1. The probability of any event (outcome) A is such that
$$0 \leq \ \Pr(A) \leq 1$$
2. The probability of the space is 1.
$$\Pr(\Omega) = 1$$
3. Let A^c be the complement of A (everything not in A in the sample space), then
$$\Pr(A) + \Pr(A^c) = 1$$
4. For any n events such that
$$A_i \cup A_j \neq \varnothing, \forall i \neq j$$
$$\Pr(\bigcup_{i=1}^{n} A_j) = \sum_{j=1}^{n} \Pr(A_j)$$

Probability of $A \bigcup B$:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

Online resource for basic statistics: Joe Blitzstein.

Conditional Probability

“In probability, conditioning is everything” - JS

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

Taking what is in A&B, and recalibrating the restricted space by what is in B (Pr(B)).

Independence

Events A and B are independent if (all equivalent):

- $Pr(A|B) = Pr(A)$
- $Pr(B|A) = Pr(B)$
- $Pr(A \cap B) = Pr(A)Pr(B)$

$Pr(A|B) = Pr(A)$ is really saying that you have no extra information about A given that you are in B.

Bayes Theorem

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

Proof: From conditional probability

$$\Pr(A \cap B) = \Pr(B|A)\Pr(A) = \Pr(A|B)\Pr(B)$$

Law of Total Probability

Given events $A_1, A_2, ..., A_n$ such that $A_i \cap A_j = \varnothing, \forall i \neq j$ and $\bigcup A_i = \Omega$. (Events $A_1, ..., A_n$ partition Ω).

For any event B,

$$\Pr(B) = \sum_{j=1}^{n} \Pr(B|A_i)\Pr(A_i)$$

Proof:

$$A_i \cap B, i = 1, ..., n$$

$$\bigcup_{i=1}^{n} A_i \cap B = B\ and\ disjoint$$

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B \cap A_i) = \Pr(B|A_i)\Pr(A_i)$$

(If independent, this theorem doesn’t help you because $\Pr(B|A_i) = \Pr(B)$)

Random Variables

A random variable (rv) $\mathcal{X}$ is a function

$$\mathcal{X} : \Omega \to \mathbb{R}$$

In other words, take any outcome $w \in \Omega$, the function $\mathcal{X}(w)$ produces a real value. The “range” of $\mathcal{X}$ is:

$$\mathcal{R} = \{\mathcal{X}(w) : w \in \Omega\}, where\ \mathcal{R} \subseteq \mathbb{R}$$

Example:

$$\Omega = set\ of\ (SNP)\ genotypes = \{CC, CT, TT\}$$

$\mathcal{X}(CC) = 0,\ \Pr(X = 0) = \Pr(\{CC\})$

$\mathcal{X}(CT) = 1,\ \Pr(X = 1) = \Pr(\{CT\})$

$\mathcal{X}(TT) = 2,\ \Pr(X = 2) = \Pr(\{TT\})$

Probability Distribution of random variables

Cumulative Distribution Function (cdf):

$$F(y) = \Pr(\mathcal{X} \leq y)$$

Example: $F(1) = \Pr(\mathcal{X} \leq 1) = \Pr(\{CC, CT\}$

Discrete rv's have a discrete range $\mathcal{R}$.

Example:

$$\mathcal{R} = \{0, 1, 2, 3, ..., 10\}$$

$$\mathcal{R} = \{0, 1, 2, 3, ...\}$$

$$\mathcal{R} = \{4, 8, 23, 92\}$$

Continuous rv's have a continuous range $\mathcal{R}$.

Example:

$$\mathcal{R} = [0, 1]$$

$$\mathcal{R} = \mathbb{R}$$

Properties of CDFs (FAS):

- They are right continuous with left limits
- $lim_{x \to \infty} F(x) = 1$
- $lim_{x \to -\infty} F(x) = 0$
- The right derivative of $F(x)$ equals $f(x)$

Probability mass or density functions

Probability Mass Function (pmf): Discrete rv's.

$$f(x) = \Pr(\mathcal{X} = x), \forall x \in \mathcal{R}$$

$$f(x) = F(x) - F(b)\ as\ b \uparrow x$$

For example, F(1) - F(.9)

Probability Density Function (pdf): Continuous rv's.

$$f(x) = \frac{d}{dx} F(x)$$

Discrete

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

$$F(y) = \sum_{x \in \mathcal{R}} f(x) = \Pr(\mathcal{X} \leq y)$$

Continuous

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

In real life, we don't measure things continuously, only to a certain number of significant digits–but we can model it continuously.

Note that $\Pr(\mathcal{X} = x) = 0$.

Median of a distribution (aka rv)

A value y such that $F(y) = 0.5$

Expected value of "population mean" (sample mean)

Discrete

$$E(\mathcal{X}) = \sum_{x \in \mathcal{R}} xf(x)$$

ie, each value x can take multiplied by the probability of that value occurring.

Continuous

$$E(\mathcal{X}) = \int xf(x)dx$$

Measure Theory

$$E(\mathcal{X}) = \int xdF(x)dx$$

Population Variance

How much a population varies from the mean.

$$Var(\mathcal{X}) = E[(\mathcal{X} - E[\mathcal{X}])^2]$$

Discrete

$$Var(\mathcal{X}) = \sum (x - E[\mathcal{X}])^2 f(x)$$

Continuous

$$Var(\mathcal{X}) = \int (x - E[\mathcal{X}])^2 f(x)dx$$

Standard Deviation

$$SD(\mathcal{X}) = \sqrt{Var(\mathcal{X})}$$

Covariance

How do rv's $\mathcal{X}$ and $\mathcal{Y}$ coordinately vary around their expected means?

$$Cov(\mathcal{X}, \mathcal{Y}) = E[(\mathcal{X} - E(\mathcal{X}))(\mathcal{Y} - E(\mathcal{Y}))]$$
$$Var(\mathcal{X}) = Cov(\mathcal{X}, \mathcal{X})$$

Correlation: take covariance and make it a unitless measure

$$Cor(\mathcal{X}, \mathcal{Y}) = \frac{Cov(\mathcal{X}, \mathcal{Y})}{SD(\mathcal{X})SD(\mathcal{Y})}$$

Covariance can be any real number.

$$-1 \leq Cor(\mathcal{X}, \mathcal{Y}) \leq 1$$

Probability Distributions: Discrete rv's

Summary:

Distribution	$\mathcal{R}$	$E(\mathcal{X})$	$Var(\mathcal{X})$
Uniform	$\{1,2,3,...,n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Bernoulli	$\{0,1\}$	p	$p(1-p)$
Binomial	$\{0,1,...,n\}$	np	$np(1-p)$
Poisson	$\{0,1,2,...\}$	λ	λ

Uniform

$$X \sim Uniform(\{1,2,3,...,n\})$$

$$\mathcal{R} = \{1,2,3,...,n\}$$

$$E[\mathcal{X}] = \frac{n+1}{2}$$

$$Var(\mathcal{X}) = \frac{n^2-1}{12}$$

$$Proof : E[\mathcal{X}] = \sum_{x \in \mathcal{R}} xf(x) = \sum_{i=1}^{n} i\frac{1}{n} = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$$

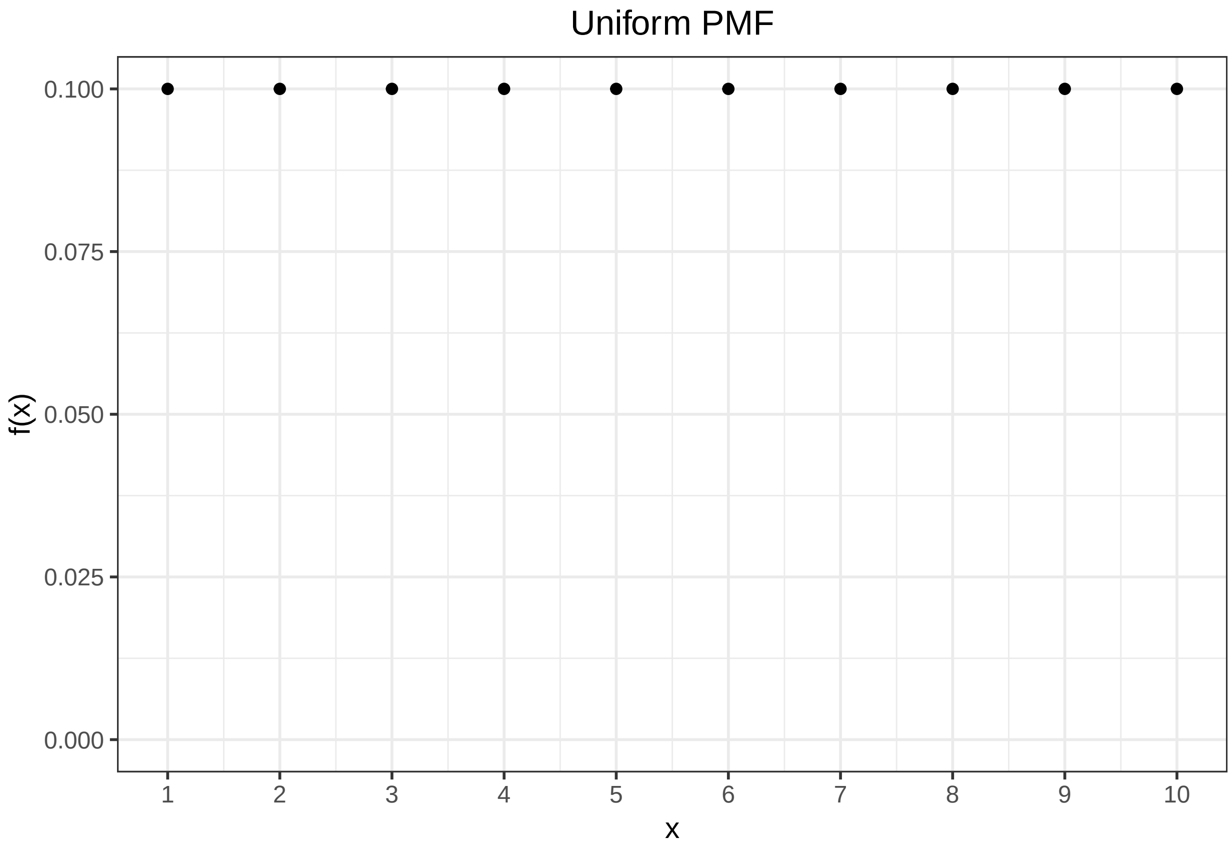
$$Variance\ Proof$$

In R, you can generate random values uniformly using, `sample`

```
> n <- 20L
> X <- sample(x=1:n,size=1e6, replace=TRUE)
> mean(X) - (n+1)/2
[1] -0.007326
> var(X) - (n^2-1)/12
[1] 0.0243156
```

Uniform PMF

```
> library(ggplot2)
> pmf <- data.frame(("n"=1:10), "f"=rep(.1, 10))
>
> ggplot(pmf, aes(x=n, y=f)) + geom_point() + scale_x_continuous(breaks=1:10) +
+ ylim(0,.1) + labs(title="Uniform PMF", x='x', y='f(x)')
```



$\mathcal{R} = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$

Bernoulli

$$X \sim Bernoulli(p)$$
$$\mathcal{R} = \{0, 1\}$$

$$f(x; p) = (1 - p)^{1-x} p^x$$
$$f(0; p) = 1 - p, f(1; p) = p$$

ie, probability of success is p with only two outcomes.

$$E[\mathcal{X}] = p = 0 * f(0) + 1 * f(1)$$
$$Var(\mathcal{X}) = p(1 - p)$$

Binomial

$$X \sim Binomial(n, p)$$

$$\mathcal{R} = \{0, 1, ..., n\}$$

Sum of n independent Bernoulli(p).

For example, the total number of heads across n flips.

$$f(x;p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

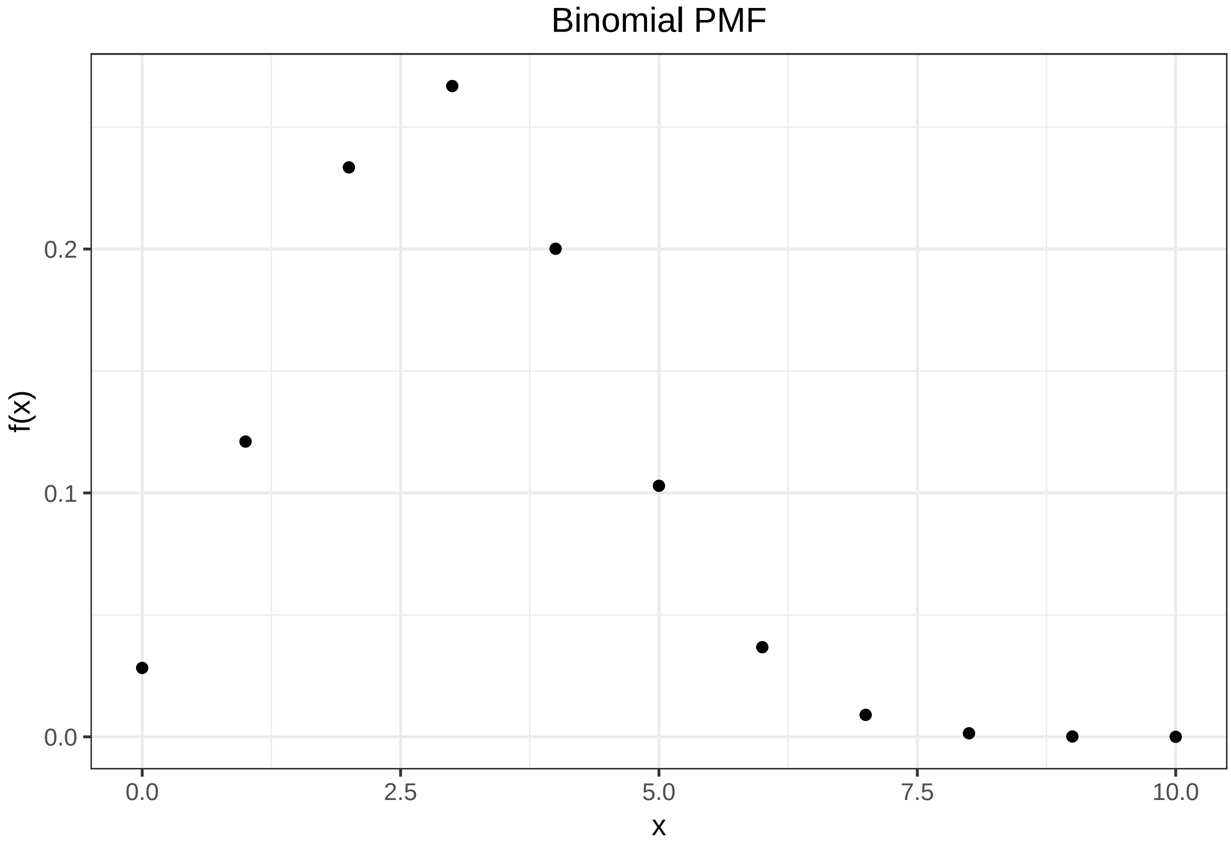
The number of ways to choose x from n without order.

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

In R,

```
dbinom
pbinom
qbinom
rbinom
```

```
> pmf <- data.frame(("n"=0:10), "f"=dbinom(0:10, prob=.3, size=10))
> ggplot(pmf, aes(x=n, y=f)) + geom_point() +
+ labs(title="Binomial PMF", x='x', y='f(x)')
```



$$n = 10, p = .3$$

Hardy-Weinberg Equilibrium

Equilibrium achieved with infinite population size, random mating, nonoverlapping discrete generations, no selection, etc. (absence of other evolutionary influences)

Source

To do Hardy-Weinberg, need genotype, rv for mother, and rv for father.

Example

Under Hardy-Weinberg equilibrium

$$\mathcal{X} = \#T\ alleles$$

$$\mathcal{X} \sim Binomial(2, p)$$
$$where\ p\ is\ the\ allele\ frequency\ of\ T$$
$$\Pr(\mathcal{X} = 0) = \Pr(\{CC\}) = (1-p)^2$$
$$\Pr(\mathcal{X} = 1) = \Pr(\{CT\}) = 2p(1-p)$$

$$\Pr(\mathcal{X} = 2) = \Pr(\{TT\}) = p^2$$

Binomial is the starting point for modeling most populations-things get interesting when you deviate from this.

For example, with genotype frequencies, we typically see $\Pr(\mathcal{X} = 1) = 2p(p-1)$, but often when non-random mating occurs this is less than expected!

Poisson

Poisson is the starting point for modeling RNAseq.

$$X \sim Poisson(\lambda)$$

λ is the rate of an event for a particular space, time unit.

$$\mathcal{R} = \{0, 1, 2, ...\}$$
$$f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Examples: - Lightning strikes per sq. kilometer per year - Defects coming off an assembly line

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

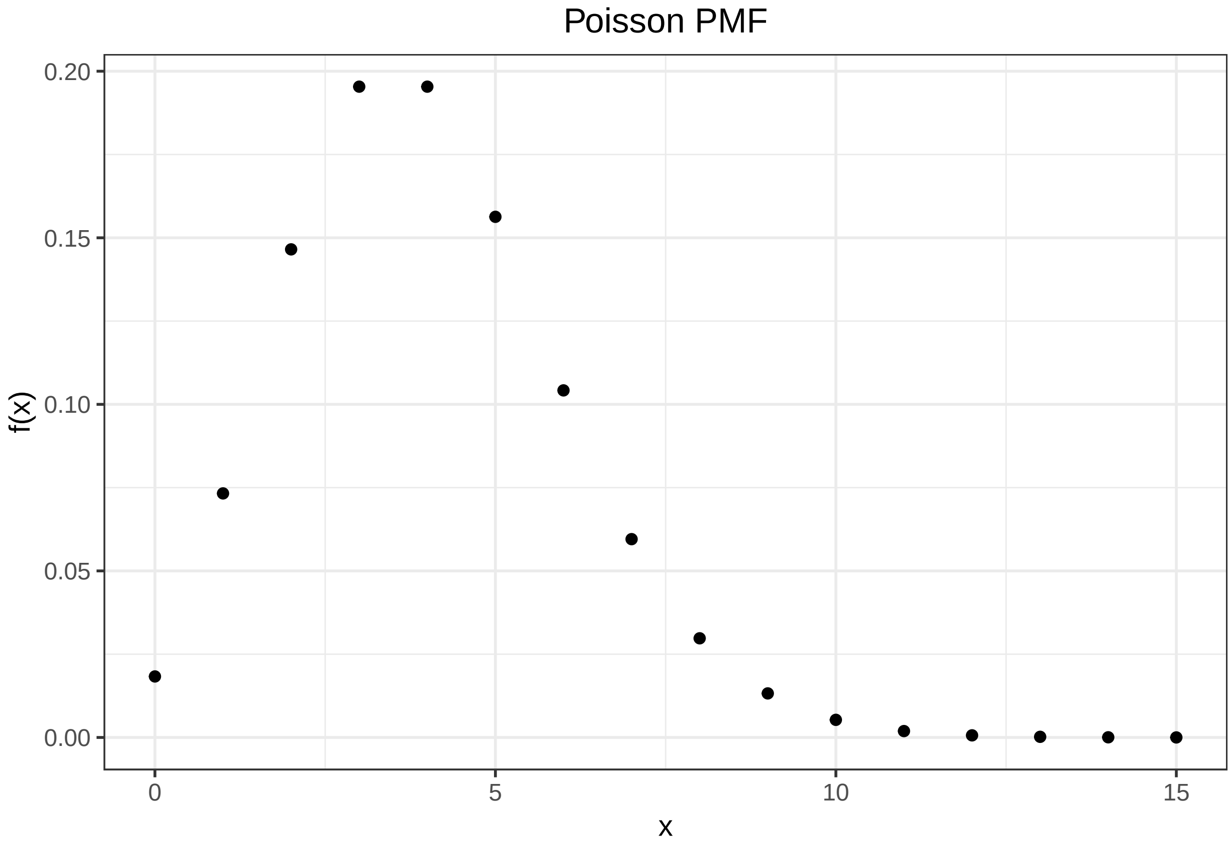
Variance = mean.

In R,

```
dpois (pmf)
ppois (cdf)
qpois (quantile)
rpois (randomdraws)

?Distributions
```

```
> pmf <- data.frame(("n"=0:15), "f"=dpois(0:15, lambda=4))
> ggplot(pmf, aes(x=n, y=f)) + geom_point() +
+ labs(title="Poisson PMF", x='x', y='f(x)')
```



$$\lambda = 4$$

Probability Distributions: Continuous rv's

Summary:

Distribution	$\mathcal{R}$	$E(\mathcal{X})$	$Var(\mathcal{X})$
Uniform	$[0,1]$	$\frac{1}{2}$	$\frac{1}{12}$
Beta	$(0,1)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Uniform

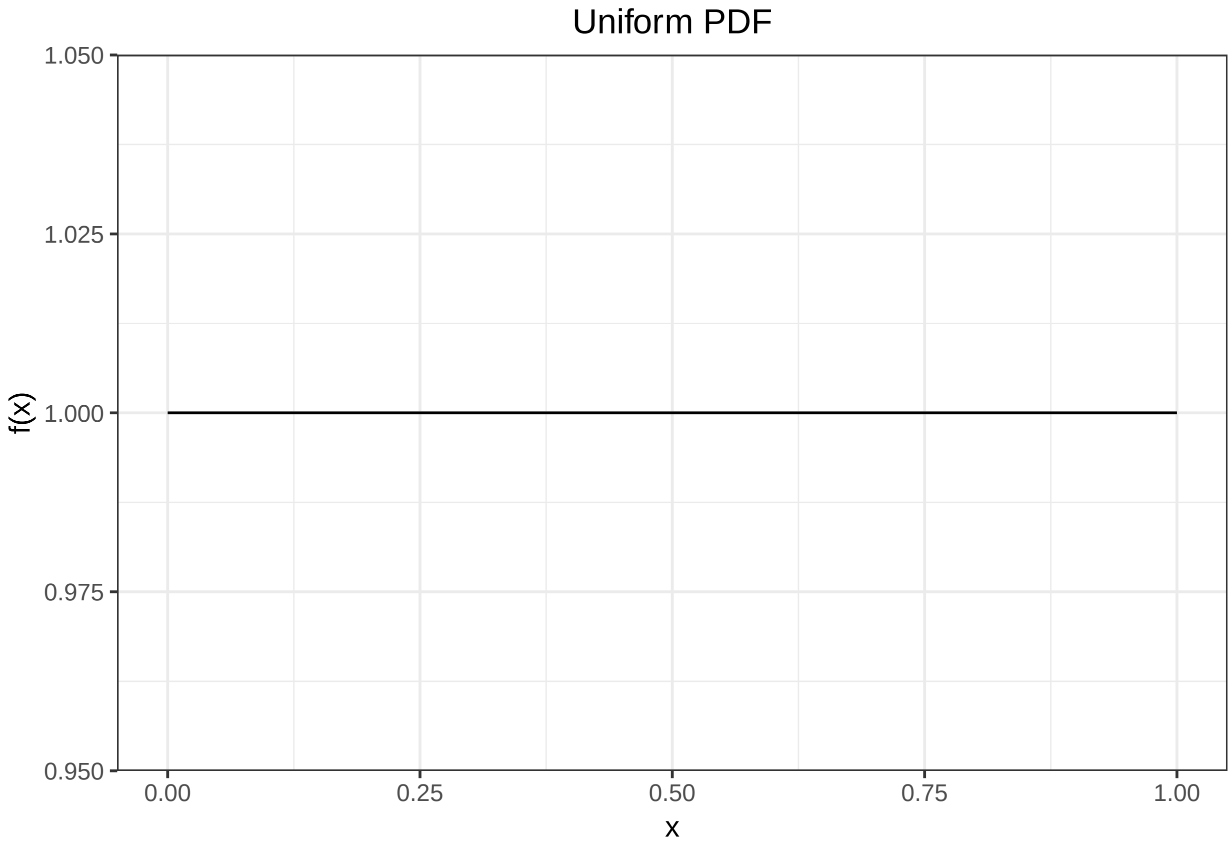
$$\mathcal{X} \sim Uniform(0,1)$$

$$\mathcal{R} = [0,1]$$
$$f(x) = 1,\ \ x \in [0,1]$$
$$F(y) = y,\ \ y \in [0,1]$$

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

$$\mathcal{X} \sim Uniform(0, \theta)$$
$$f(x; \theta) = \frac{1}{\theta}$$
$$F(x; \theta) = \frac{y}{\theta}$$
$$\mathcal{R} = [0, \theta]$$

```
> library(ggplot2)
> xn <- seq(0, 1, .2)
> pmf <- data.frame(("n"=xn), "f"=dunif(xn))
>
> ggplot(pmf, aes(x=n, y=f)) + geom_line() +  labs(title="Uniform PDF", x='x', y='f(x)')
```



Example: Distribution of p-values is Uniform(0,1) when the null hypothesis is true.

Beta

$$\mathcal{X} \sim Beta(\alpha, \beta), \alpha, \beta > 0$$
$$\mathcal{R} = (0, 1)$$

This is useful for probabilities! For example, to generate p for a Bernoulli distribution, use Beta distribution!

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ x \in (0,1)$$

Scale parameters are there so $\int_0^1 f(x;\alpha,\beta)dx = 1$.

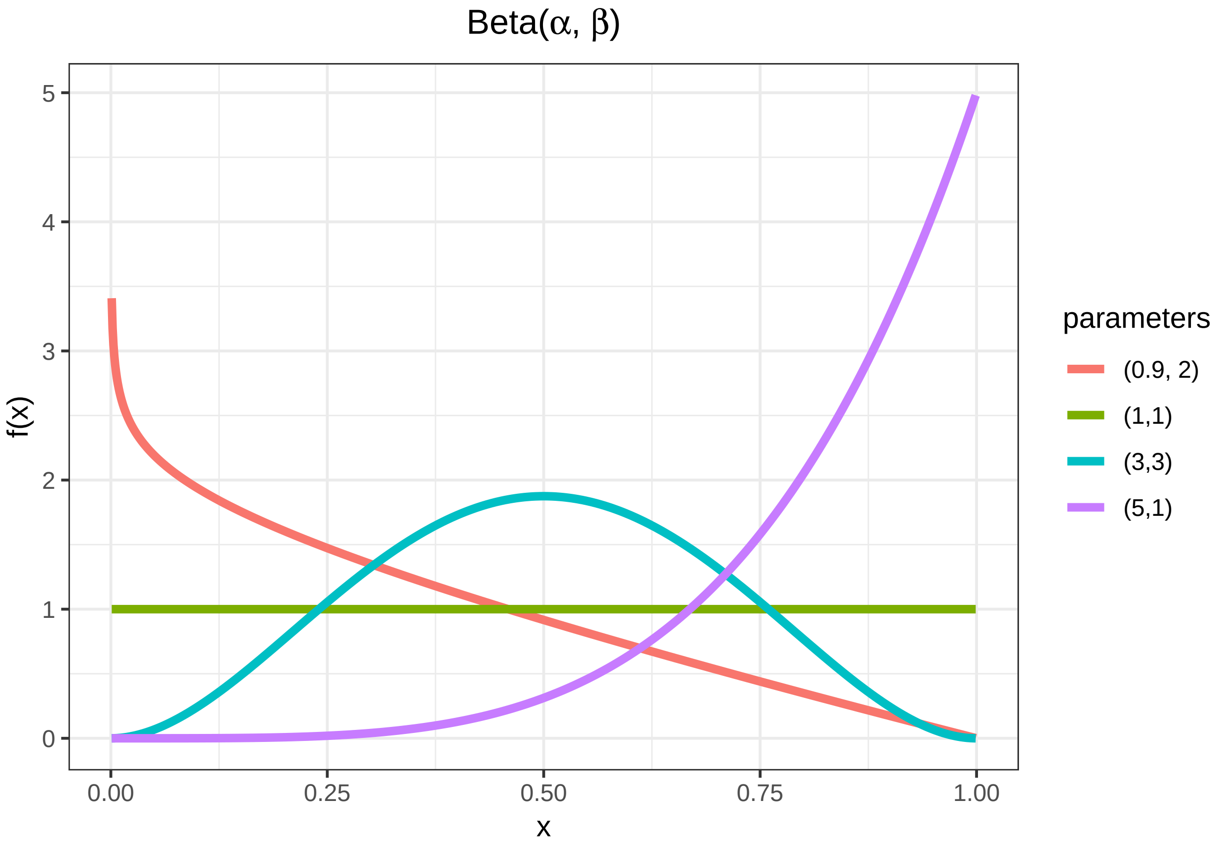
$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$$

In R, `help(gamma)`.

```
dbeta
pbeta
qbeta
rbeta
```

$$E[\mathcal{X}] = \frac{\alpha}{\alpha+\beta}$$

$$Var(\mathcal{X}) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$



The Beta PDF comes in a variety of shapes. (PDF examples from FAS.)

Session Information

```
> sessionInfo()
R version 3.6.0 (2019-04-26)
```

```
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS  10.15.3

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

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   base

other attached packages:
 [1] gapminder_0.3.0 forcats_0.4.0   stringr_1.4.0   dplyr_0.8.1
 [5] purrr_0.3.2     readr_1.3.1     tidyr_0.8.3     tibble_2.1.1
 [9] ggplot2_3.1.1   tidyverse_1.2.1 knitr_1.22

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.1       cellranger_1.1.0 pillar_1.4.0     compiler_3.6.0
 [5] plyr_1.8.4       tools_3.6.0      digest_0.6.18    lubridate_1.7.4
 [9] jsonlite_1.6     evaluate_0.13    nlme_3.1-140     gtable_0.3.0
[13] lattice_0.20-38  pkgconfig_2.0.2  rlang_0.3.4      cli_1.1.0
[17] rstudioapi_0.10  yaml_2.2.0       haven_2.1.0      xfun_0.7
[21] withr_2.1.2      xml2_1.2.0       httr_1.4.0       hms_0.4.2
[25] generics_0.0.2   grid_3.6.0       tidyselect_0.2.5 glue_1.3.1
[29] R6_2.4.0         readxl_1.3.1     rmarkdown_1.12   modelr_0.1.4
[33] magrittr_1.5     backports_1.1.4  scales_1.0.0     htmltools_0.3.6
[37] rvest_0.3.3      assertthat_0.2.1 colorspace_1.4-1 labeling_0.3
[41] stringi_1.4.3    lazyeval_0.2.2   munsell_0.5.0    broom_0.5.2
[45] crayon_1.3.4
```