Big Data BUS 41201

Week 1: Large-scale Inference

Veronika Ročková

University of Chicago Booth School of Business

http://faculty.chicagobooth.edu/veronika.rockova/

Getting oriented

- ✓ Introduction: goals, material, syllabus
- ✓ Data Science, Machine Learning and Big Data
- ✓ Computing
  - ⇝ R: examples, resources, and how we'll learn.
  - ⇝ Big Data, distribution, storage, connectivity.
- ✓ Visualization, Statistics, and Dimension Reduction
- ✓ Testing and Discovery: False Discovery Rates

This is a class about Inference at Scale

We're here to make sure you have the tools for making good decisions based on large and complicated data.

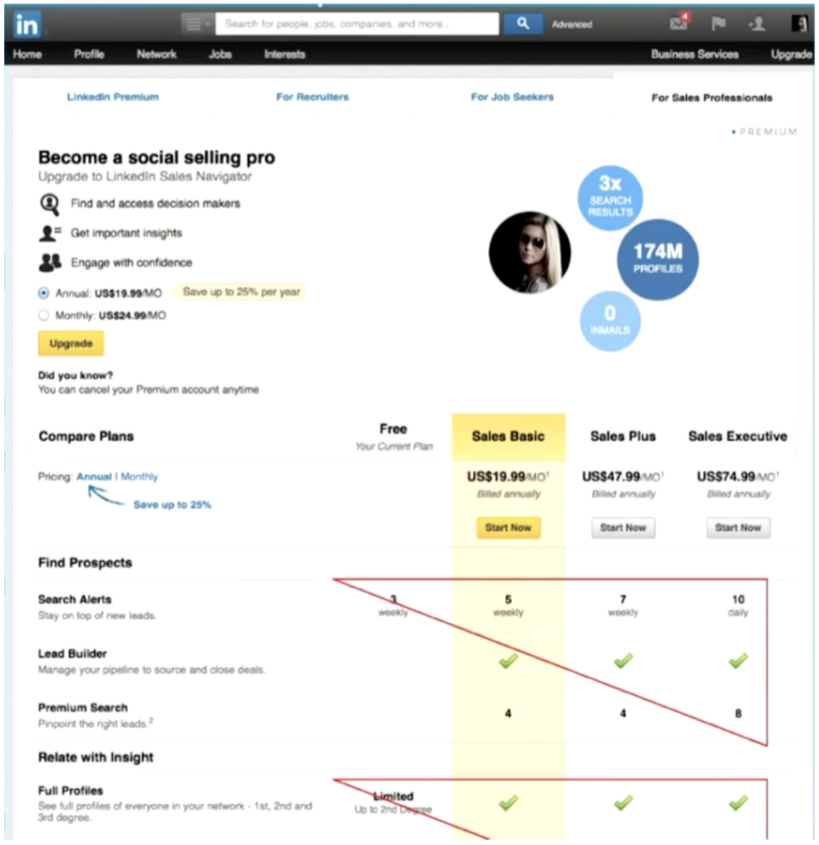
A mix of practice and principles

- Solid understanding of essential statistical principles.
- Concrete analysis ability and best-practice guidelines.

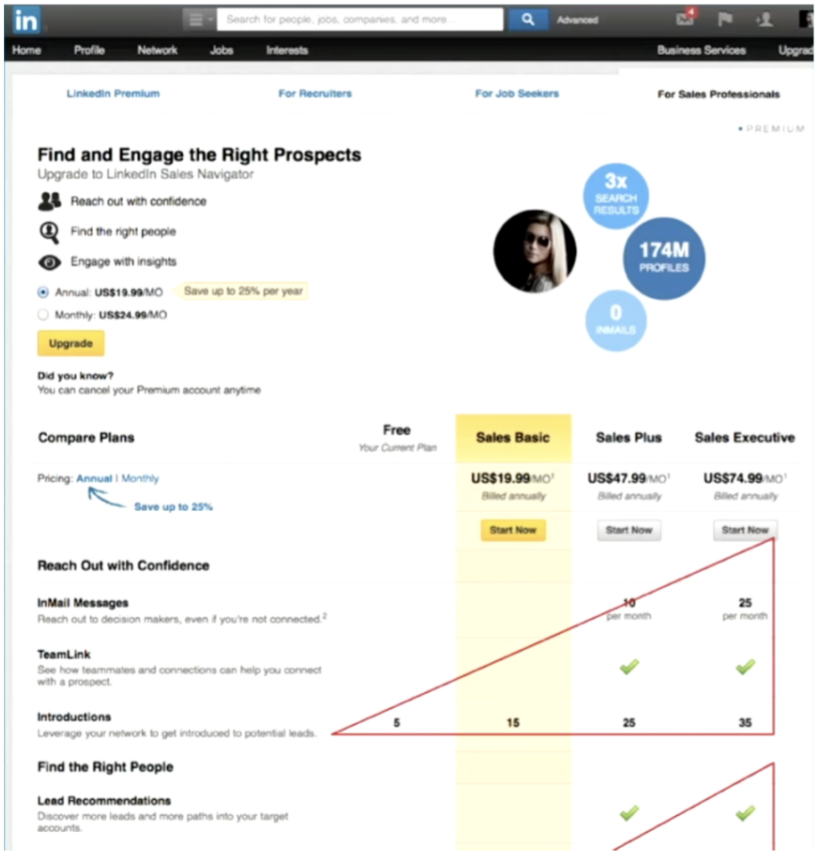
We'll learn what to trust, how to use it, *and how to learn more*.

A hands-on subject: the idea that MBAs can just pass data analysis off to number crunchers is an out-of-date cartoon.

...to whet your appetite



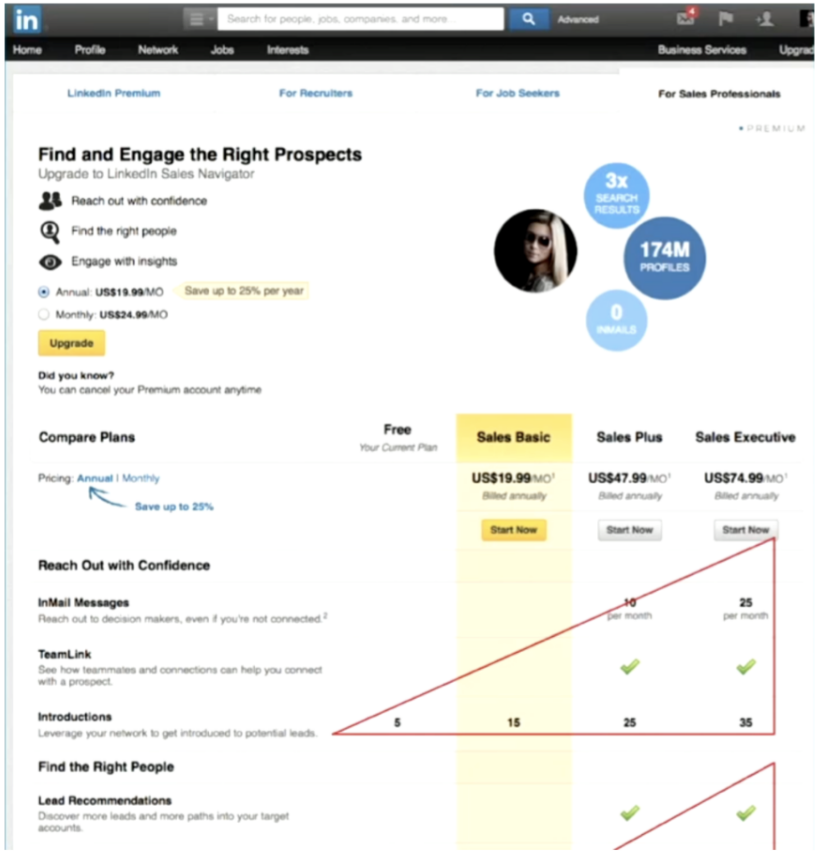
A



B

[Credit: Iavor Bojinov]

...to whet your appetite



B

10% increase in spending.

[Credit: Iavor Bojinov]

What is in a name?

Big Data, Econometrics, Statistics, and Machine Learning

There are many labels for what we do...

Econometrics $\rightarrow$ Statistics
$\rightarrow$ Data Mining / Big Data / Data Science
$\rightarrow$ Machine Learning (ML) and Artificial Intelligence (AI)

Along this spectrum, you move from heavy focus on what things you are measuring (what real phenomena they correspond to) to a more pragmatic 'useful is true' pattern discovery approach.

The similarities are much bigger than any distinctions.

The 'BD' name comes from computer scientists working to do aggregation on data that is *too big to fit on a single machine*.

"*Statistics* means the practice or science of collecting and analyzing numerical data in large quantities"*.

"*Data Science* means the practice of liberating and creating meaning from data using scientific methods"*.

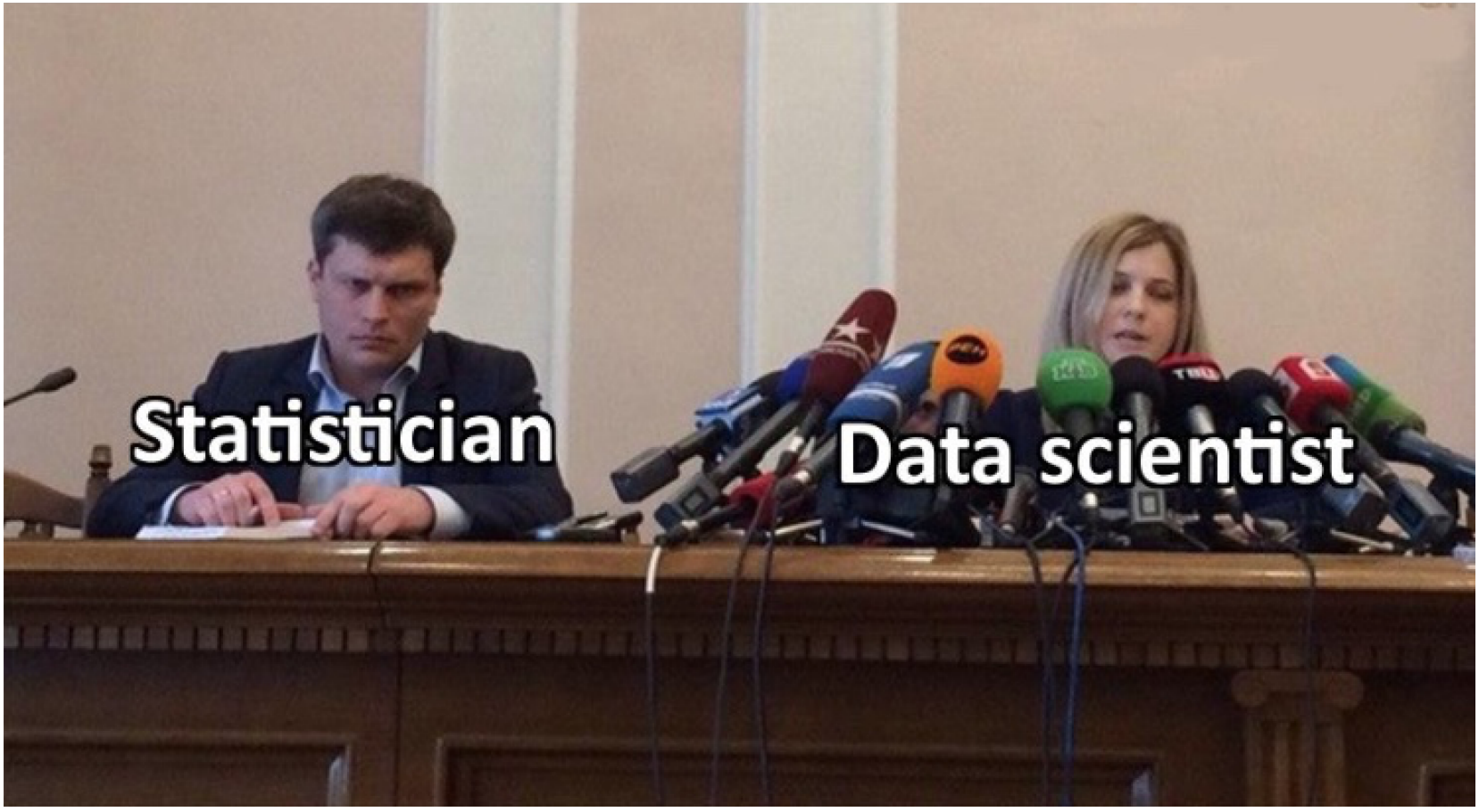
It is the umbrella term for inference in a world that is messier that in your old statistic textbook.

Big Data is focused on actionable knowledge extraction from very large datasets (integral in business and industrial applications).

- Infer patterns from complex high dimensional data.
- Simplicity and scalability of algorithms is essential.
- We keep an eye on both *useful* and *true*.
- The end product is a *decision*.

*D. Donoho "50 Years of Data Science"

Today's Data Science Movement



"There will be a shortage of talent necessary for organization to take advantage of big data. By 2018, the United States alone could face a shortage of 140 000-190 000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."

Big data: The next frontier for innovation, competition, and productivity (McKinsey Report 2011)

"Statistical thinking not only helps make scientific discoveries, but it quantifies the reliability, reproducibility and general uncertainty associated with these discoveries. Because one can easily be fooled by complicated biases and patterns arising by chance, and because statistics has matured around making discoveries from data, statistical thinking will be integral to Big Data challenges."

Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society (ASA White Paper 2014)

Today's Data Science Movement

Google's Chief Economist Hal Varian on Statistics and Data

POSTED TO QUOTES, STATISTICS | NATHAN YAU

> *The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.*
>
> *I think statisticians are part of it, but it's just a part. You also want to be able to visualize the data, communicate the data, and utilize it effectively. But I do think those skills – of being able to access, understand, and communicate the insights you get from data analysis – are going to be extremely important. Managers need to be able to access and understand the data themselves.*

A big aspect of Big Data is 'pattern discovery' or 'data mining'
Good DM is about inferring useful signal at massive scale.

Our goal is to summarize really high dimensional data in such a way that you can relate it to *structural models* of interest.

$\Rightarrow$ Variable Selection and Dimension Reduction

We also want to *predict!* If things don't change too much...

$\Rightarrow$ Probabilistic Prediction and Classification Rules

We need to constantly beware of false discovery.

What does it mean to be 'big'?

Big in both the number of observations (size 'n')
and in the number of variables (dimension 'p').

In these settings, you cannot:
Look at each individual variable and make a decision (*t-tests*).
Choose amongst a small set of candidate models (*F-test*).
Plot every variable to look for interactions or transformations.

Some BD tools are straight out of previous statistics classes (linear regression) and some are totally new (trees, PCA).
All require a different approach when n and p get really big.

Course Schedule

subject to change...

[1] Data: Computing, plotting, and principles. [False?] discovery

[2] Regression: A grand overview, linear and logistic

[3] Model Selection: Penalties, information criteria, cross-validation

[4] Treatment Effects: HD controls, AB testing, bootstrap

[5] Classification: Multinomials, KNN, sensitivity/specificity

Midterm!

[6] Networks: Co-occurence, directed graphs, Page Rank, community detection

[7] Clustering: Mixture models, k-means, and association rules

[8] Factors: Latent variables, factor models, PCA, and PLS

[9] Trees: CART and random forests, ensembles

The Big Data Team

Kenichiro McAlinn
Chicago Booth
kenmcalinn@gmail.com

Grant Gannaway
U of C Economics Dept
grant.gannaway@gmail.com

Yuexi Wang
Chicago Booth
yxwang99@uchicago.edu

We'll be working with real data analysis examples

- Mining client information: Who buys your stuff, what do they pay, what do they think of your new product?
- Online behavior tracking: Who is on what websites, what do they buy, how do/can we affect behavior?
- Collaborative filtering: predict preferences from people who do what you do; space-time recommender engines.
- Text mining: Connect blogs/emails/news to sentiment, beliefs, or intent. Parsing unstructured data, e.g. EMR.
- Big covariates: mining data to predict asset prices; using unstructured data as controls in observational studies.

Many are applicable to marketing, but we're far more general.

All of our analysis will be conducted in R

This is the real deal: industrial strength software for data analysis. It's free, cross platform, and hugely capable.

Academics (stats, marketing/finance, genetics, engineering), companies (EBay, Google, Microsoft, Boeing, Citadel, IBM), and governments (Rand, DOE National Labs, Navy) use R.

Since R is free, you'll always be able to use it.

Great source *Introductory Statistics with R* by Peter Dalgaard

A huge strength of R is that it is open-source.
This is also why it is sometimes a bit unpolished.

R has a core, to which you can add contributed packages. These add-ons are as varied as the people who write them. Some are specific, others general, some are great, some not so great.

R is not without flaws, but neither are the other options.
e.g. python, but the community of stats developers is smaller and you need to be a more careful programmer.

Some students prefer to wrap R in an IDE; e.g. R-studio.

The barrier of entry for R is its command line interface:

You type commands to get what you want.

The learning curve can be steep, but is very worthwhile.

- You have code of saved commands, so you know what you've done and can easily repeat similar operations.
- Interacting with computers in this way is a Big Data skill.

All code for lectures and homework will be available online.
The best way to learn software is through imitation.

There are a ton of resources: see the website and **syllabus**.

Computing: R

To start, click on  or  or just type 'R' in a terminal.

At its most basic, R's just a fancy calculator. (e.g., `*`,`/`,`+`,`-`).

Everything is based on assigning names
(<- works pretty much the same as =).

`A <- 2` → 2.

`B <- c(1,2,3)` → 1 2 3 (same as `B=1:3`).

`C = A + B[3]` → 5.

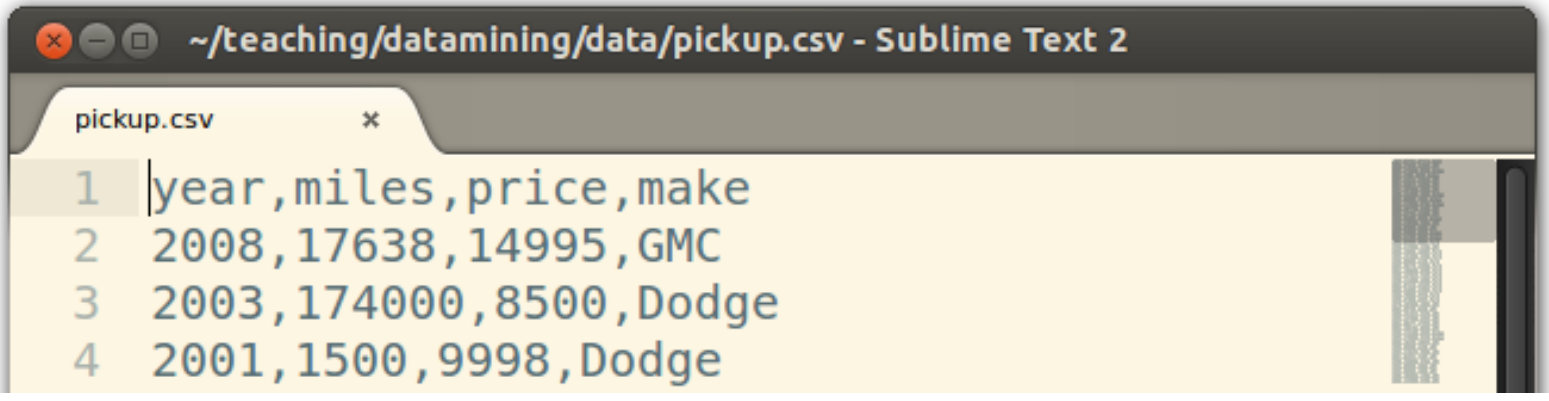
The `c( )` function and `:` operator build vectors.
`length(B)` will tell you how many elements it holds (3).

To inspect a variable, just type the name (do this often!).

R can read almost any data format

First, set the working directory to where you store data.
It's good practice to create folders just for this.
Use `cmd-D` (Mac), `File` → `ChangeDir` (Windows)
or just type `setwd('/path/to/my/working/dir')`.
Set a default with *preferences* or shortcut *properties*.

In this directory, you'll store flat files: text tables of data.
We'll use mostly `.csv` files: comma separated values.
Any file in Excel can be saved as '.csv', and vice versa.



To load data, type `trucks <- read.csv("pickup.csv")`.
The data are then in working memory (RAM).

`trucks` is a `dataframe`: a matrix with names.
You can use index names and numbers to access the data.

```
trucks[1,] # the first observation
trucks[1:10,] # the first 10 observations
trucks[,1] # the first variable (year)
trucks$year # same thing
trucks[,'year'] # same thing again
trucks[trucks$miles>200000,] # some real clunkers
```

And call R's functions on the data.

```
nrow(trucks) # sample size
summary(trucks) # summary of each variable
```

Basic Elements in R: numeric, factor, logical, character

The values in our dataframes all have a class.

- numeric values are just numbers (1, 2, 0.56, 10.2).
- factor variables are categories with levels ('lowfat','reg')
- logical values are either TRUE or FALSE.
- character strings are just words or messages ('hi mom').

We have plenty of tools to investigate and manipulate these:

`as.numeric(X)`, `factor(X)`, `class(X)`, `levels(X)`

R has functions that look like `f(arg1, arg2, ...)`.

e.g., create new variables: `lprice = log(trucks$price)`.

And add them to your data: `truck$lprice = lprice`.

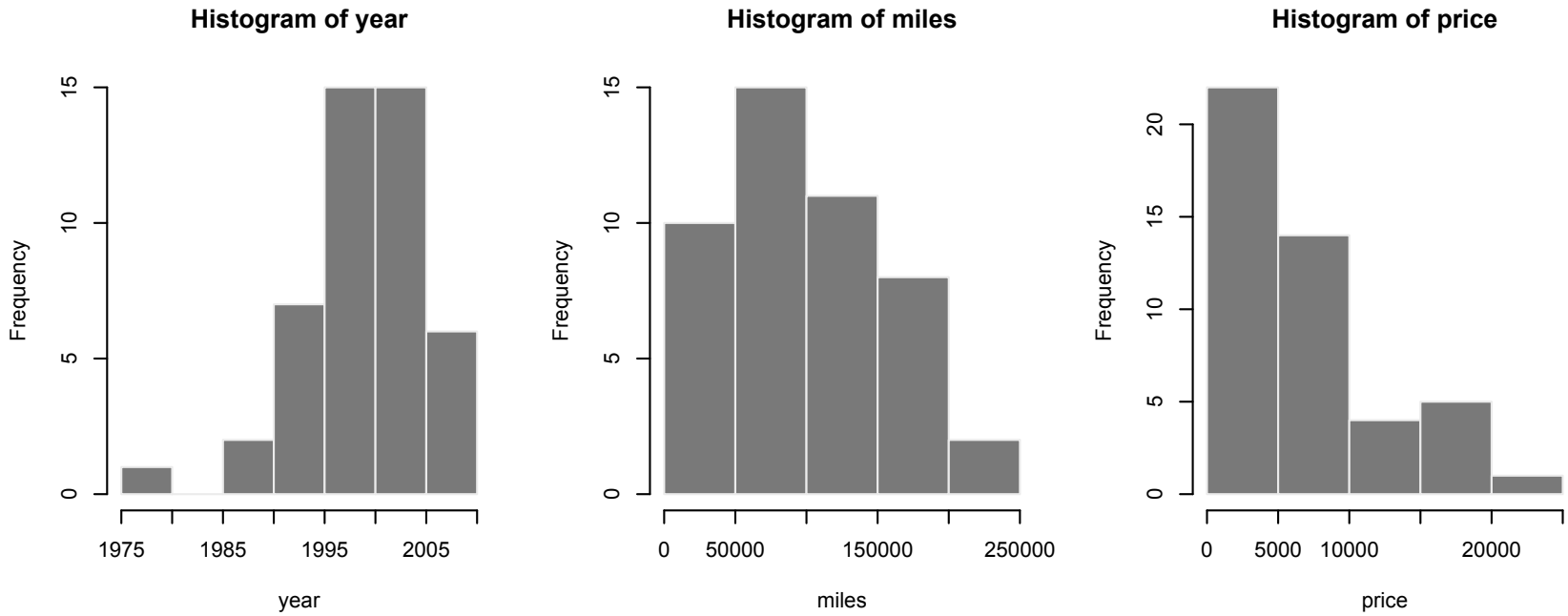
To find out how a function works, type `?f` or `help(f)`.

Plotting is super intuitive

Use `plot(mydata$X, lY)` or `plot(lY ~ mydata$X)`

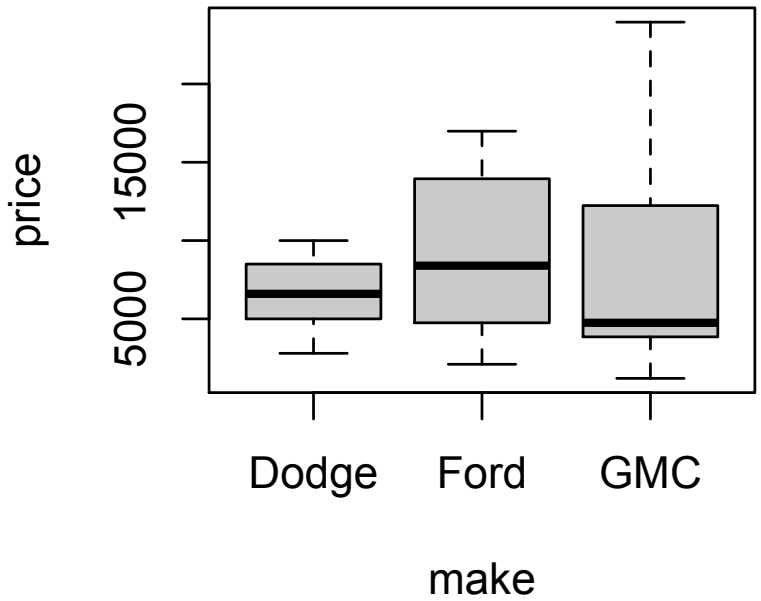
Let's look at some basic plots...

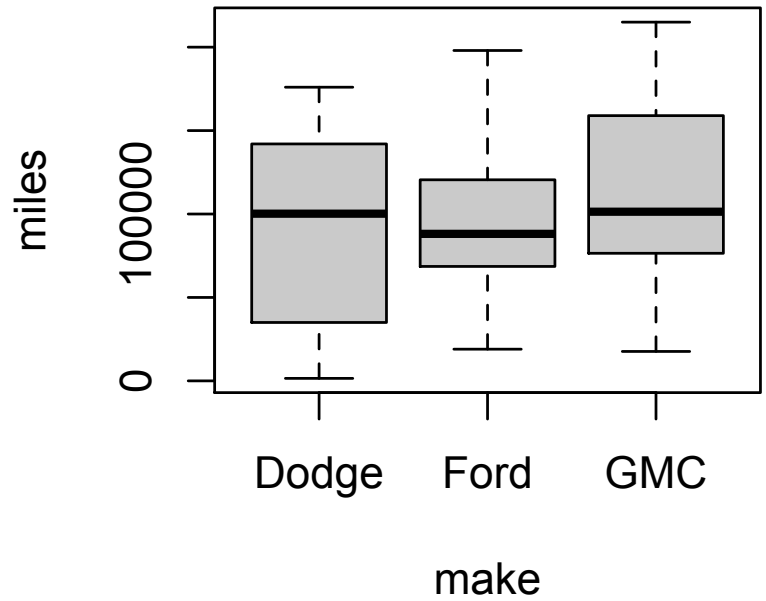
The simple histogram for continuous variables



Data is binned and plotted bar height is the count in each bin.

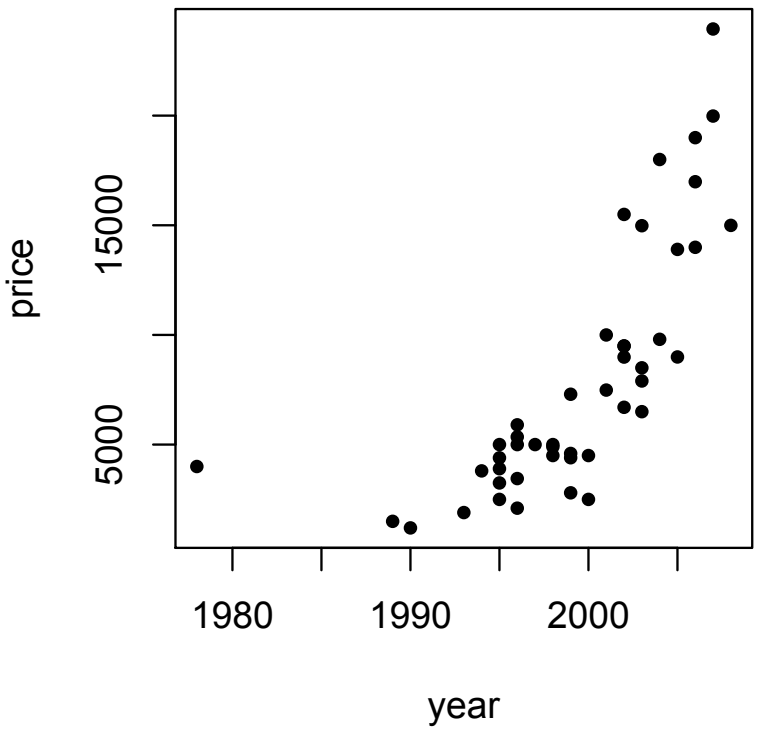
Boxplots: summarizing conditional distributions

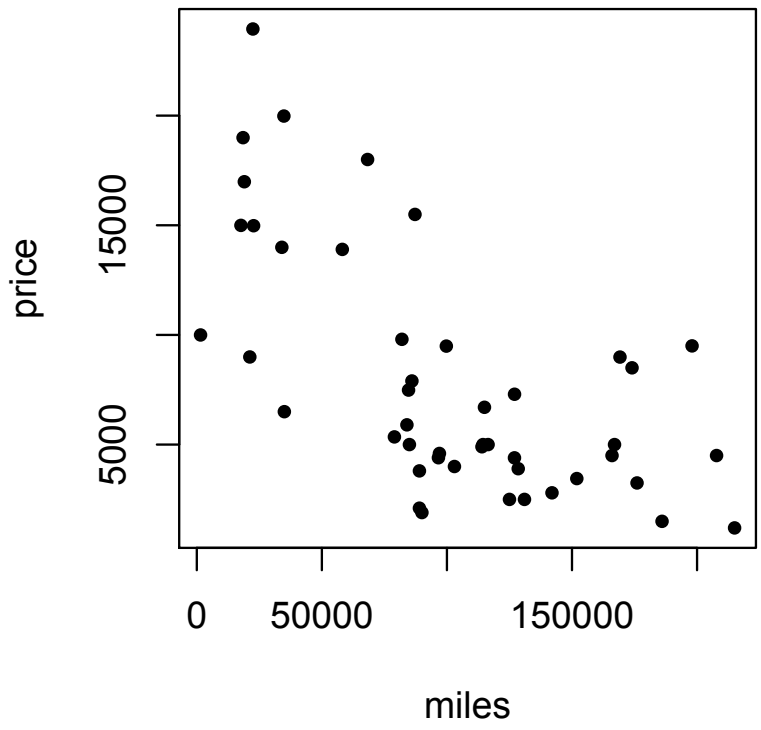




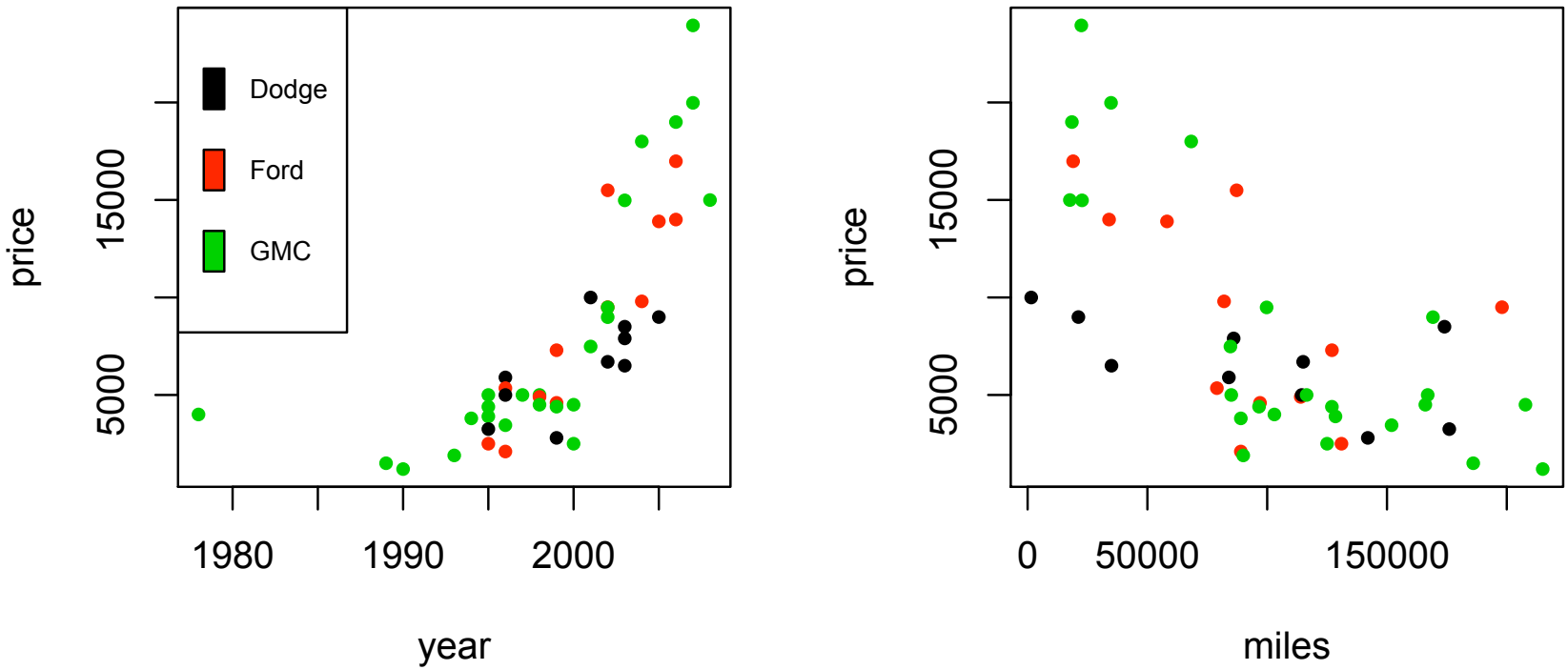
The box is the Interquartile Range (IQR; i.e., 25^{th} to 75^{th} %), with the median in bold. The whiskers extend to the most extreme point which is no more than 1.5 times the IQR width from the box.

Use scatterplots to compare variables.





And color them to see another dimension.



The scatterplot is more powerful than you think...

Scatterplots are a fundamental unit of statistics.

If you're able to find and compare meaningful low-dimensional summary statistics, then you are winning the DM game.

- Humans are good at comparing a few variables.
- If we can put it in a picture, we can build intuition.
- Prediction is easier in low dimensions.

The key to good graphics is to reduce high-dimensional data to a few very informative summary statistics, then plot them.
We'll focus on info visualization throughout this course.

A note on data visualization

Data visualization is an essential part of *exploratory data analysis (EDA)* (histograms, scatter-plots and time series plots)

Some guidelines

- *reducing dimension* of your data to a few rich variables for comparison. Can be just picking two features to scatterplot, or can involve more complicated projections.
- check for *outliers and pattern anomalies*, prior to fitting models
- *effective communication* of results – with shapes, space, and color – for a given set of variable observations.

Regression is king

```
fit <- glm(log(price)~ year+make, data=trucks)
summary(fit)  # glm = generalized linear model
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -195.22731   25.18120  -7.753 1.24e-09 ***
year           0.10196    0.01259   8.099 4.07e-10 ***
makeFord       0.13987    0.19786   0.707    0.484
makeGMC        0.16202    0.17586   0.921    0.362
```

This is the model (familiarize yourself with the notation!)

$$\mathbb{E}[\log(\texttt{price})] = \beta_0 + \texttt{year}\beta_{\texttt{year}} + \mathbb{1}_{[\texttt{ford}]}\beta_{\texttt{ford}} + \mathbb{1}_{[\texttt{gmc}]}\beta_{\texttt{gmc}}.$$

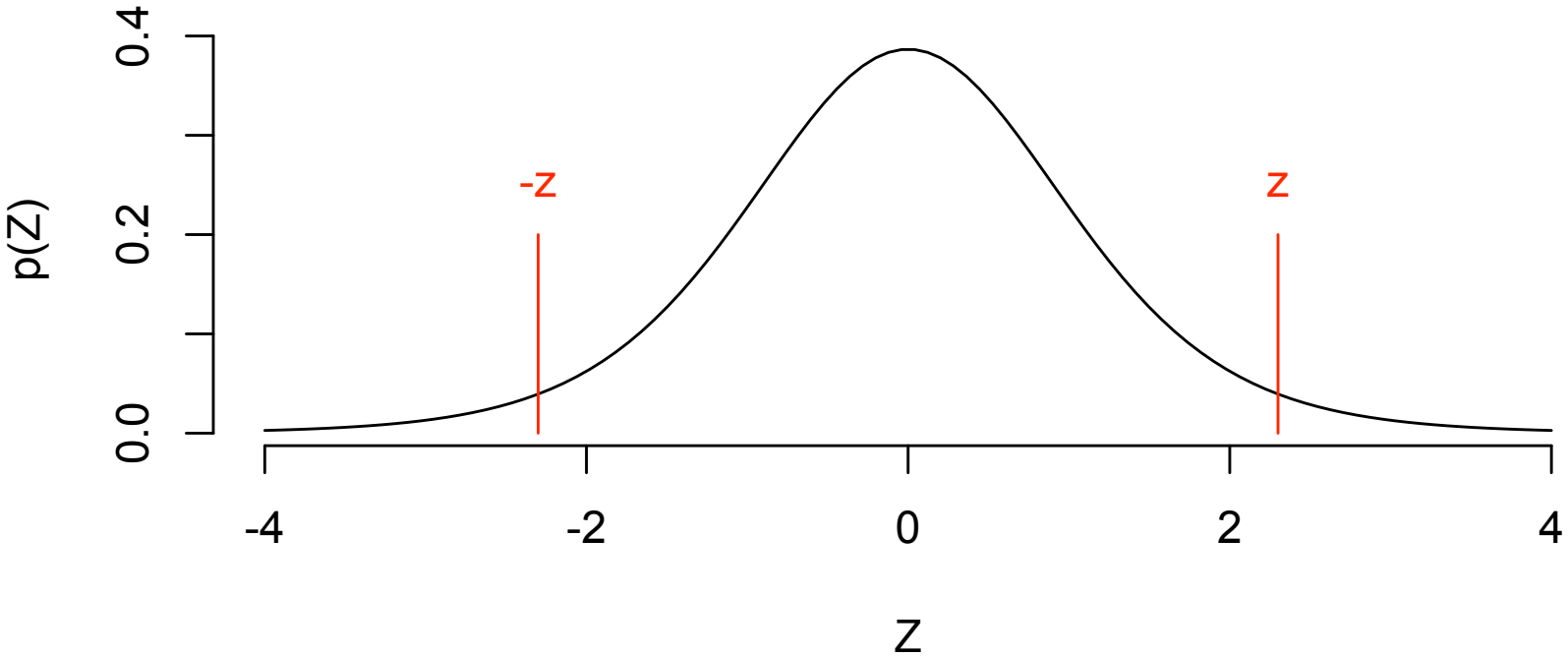
See other models and syntax in `pickups.R`.

We'll see more next week: review the basics before then.

Review: Hypothesis Testing

What is `Pr(>|t|)`? Why the `***`?

For a test of $\beta = 0$ vs $\beta \neq 0$, the test stat is $z_\beta = \hat{\beta}/s_{\hat{\beta}}$:
how many standard deviations is our estimate away from zero?
The p-value is then $\mathrm{P}(|Z| > |z_{\hat{\beta}}|)$, with $Z \sim \mathrm{N}(0, 1)$.



A single test

A p-value is probability of observing a test statistic that is more extreme than what we would observe *if the null hypothesis H_0 were true*.

$$\mathrm{P}(Z \text{ as extreme as observed or larger} \mid H_0)$$

Testing procedure: Choose a cut-off 'α' for your p-value 'p', and conclude significance (e.g., variable association) for $p < \alpha$.

This is justified as giving only α probability of a false-positive.

$$\mathrm{P}(p < \alpha \mid H_0) = \alpha$$

For example, in regression, we reject the hypothesis $\{\beta \neq 0\}$ only if its p-value is less than the accepted risk of a false discovery for each coefficient.

p-values

There is a more-or-less agreed upon scale for interpreting p-values

α	0.10	0.05	0.025	0.01	0.001
Strength of evidence	borderline	moderate	substantial	strong	overwhelming

The smaller the p-value the more decisively we reject. p-value should not to be confused with probability of H_0 being true!

Big data challenges for p-values

☹ With big enough sample sizes (i.e. n is large), it is easy to reject even when the estimated effect $\hat{\beta}$ is small.

☹ With many parameters (i.e. p is large), performing many tests without properly accounting for *multiplicity* can culminate in many false discoveries.

Large-scale testing

We wish to test p simultaneous null hypotheses

$$H_{01}, H_{02}, \ldots, H_{0p}$$

with a common procedure.

Out of the p hypotheses, N_0 are true Nulls and $N_1 = p - N_0$ are true non-Nulls.

		Decision	
Actual	Not-reject	Reject	
Noise	True Non-discovery (TN)	False Discoveries (FD)	N_0
Signal	False Non-discovery (FN)	True Discoveries (TD)	N_1
	p-R	R	

The problem is to choose a procedure that balances the two types of errors.

The problem of multiplicity

α is for a single test. If you repeat many tests, about $\alpha \times 100\%$ of the null tests should erroneously pop up as significant ☹.

Suppose that 5 of 100 regression coefficients are actually influential, and that you happen to find all of them significant.

Test the remaining 95 of them at level $\alpha = 0.05$:

Since you reject H_0 for about 5% of the useless 95 variables, $4.75/(5 + 4.75) \approx 50\%$ of significant tests are false discoveries!

This is called the False Discovery Proportion (FDP).

FDP depends on α and on the number of true non-Null hypotheses. It can be really big with a small true non-Null rate.

False Discovery Rate

Big data is about making *many* tough decisions.

Instead of focusing on single tests, we'll consider

$$\text{FD Proportion} = \frac{\#\text{ false positives}}{\#\text{ tests called significant}} = \frac{FD}{R}$$

FDP is a property of our fitted model. We can't know it.

But we can control its expectation:

False Discovery Rate, FDR $= \mathbb{E}$[FDP].

It is the multivariate (aggregate) "analogue" of α.

If all tests are tested at the α level, we have $\alpha = \mathbb{E}[FD/N_0]$, whereas FDR$=\mathbb{E}[FD/R]$.

False Discovery Rate control

Suppose we want to be sure that FDR $\leq q$ (say, 0.1).

The Benjamini + Hochberg (BH) algorithm:

- Rank your p p-values, smallest to largest,

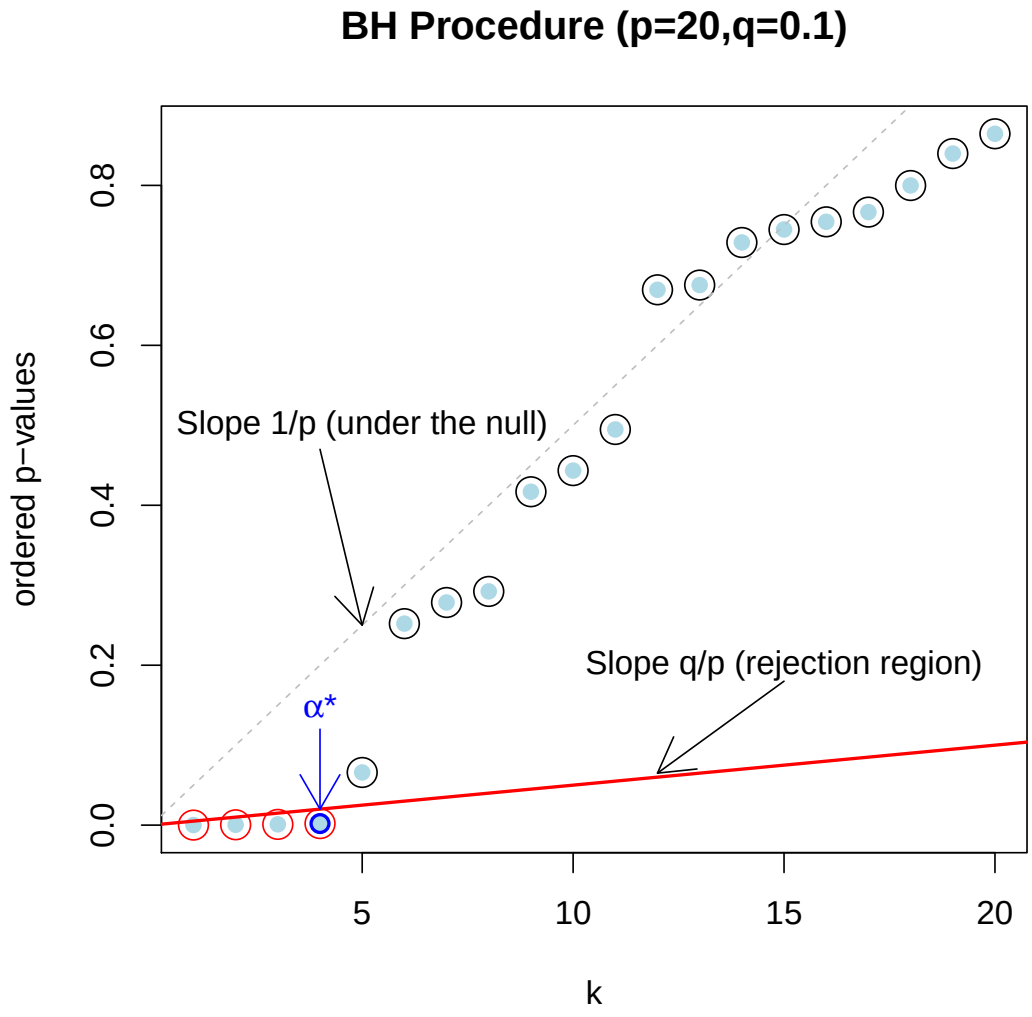
$$p_{(1)} \leq \cdots \leq p_{(p)}.$$

- Set the p-value cut-off as $\alpha^\star = \max \left\{ p_{(k)} : p_{(k)} \leq q\frac{k}{p} \right\}$.

If your rejection region is p-values $\leq \alpha^\star$, then FDR $\leq q$.

Caution: assumes (rough) independence between tests.

♦ Understanding FDR control



Declare significance for all points below the line.

FDR roundup

We introduced the problem of *multiplicity* by saying that given α (p-value cutoffs) can lead to big FDR: $\alpha \to q(\alpha)$

B+H reverse the relationship – it's a recipe for the $\alpha(q)$ that will give you whatever FDR q you want: $q \to \alpha^\star(q)$

FDR is *the* way to summarize risk when you have many tests.

B+H offers a way of combining multiple tests with principled bound on the overall error and, at the same time, power to detect.

You'll never think about testing the same again! ☺

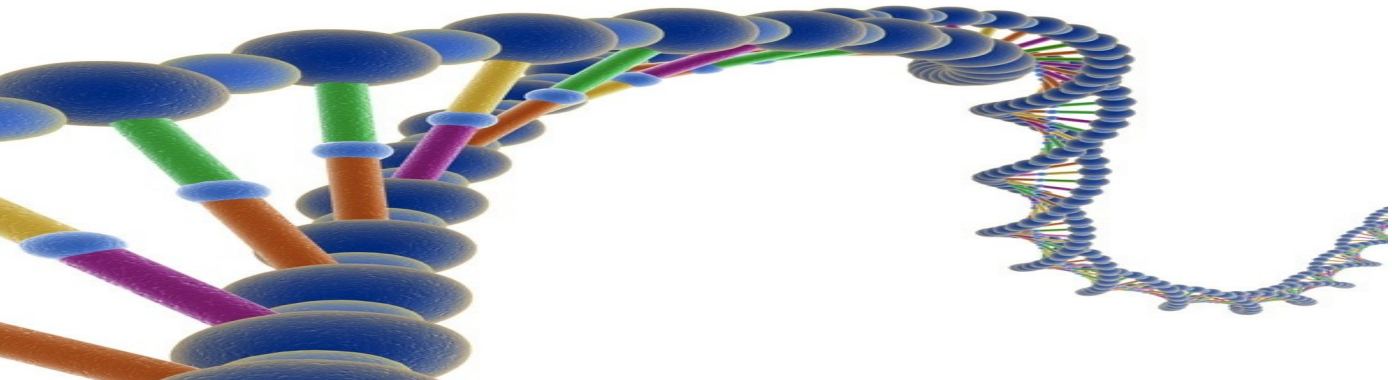
Example: multiple testing in GWAS

GWAS: genome-wide association studies
Scan large DNA sequences for association with disease.

Single-nucleotide polymorphisms (SNPs) are paired DNA locations that vary across chromosomes. The allele that occurs most often is major (A), and the other is minor (a).

Question: Which variants are associated with increased risk?

Then investigate why + how.



Cholesterol

Willer et al, Nat Gen 2013 describe meta-analysis of GWAS for Cholesterol levels. We'll focus on the 'bad' LDL Cholesterol.

At each of 2.5 million SNPs, they fit the linear regression

$$\mathbb{E}[LDL] = \alpha + \beta AF$$

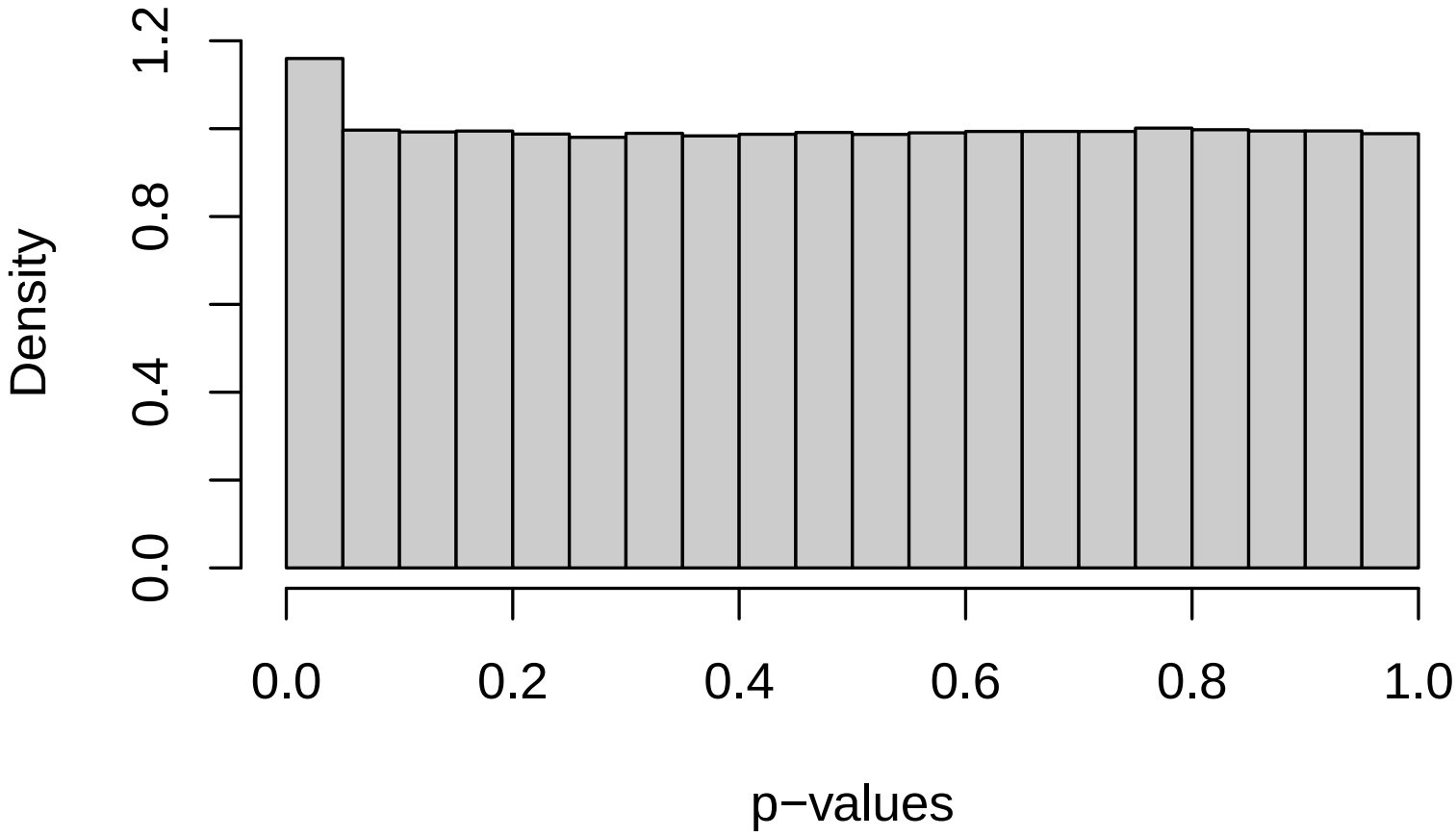
Where AF is allele frequency for the 'trait increasing allele'.

2.5 mil SNP locations

$$\Rightarrow \text{2.5 mil tests of } \beta \neq 0$$

$$\Rightarrow \text{2.5 mil p-values!}$$

Cholesterol GWAS P-values: Which are significant?

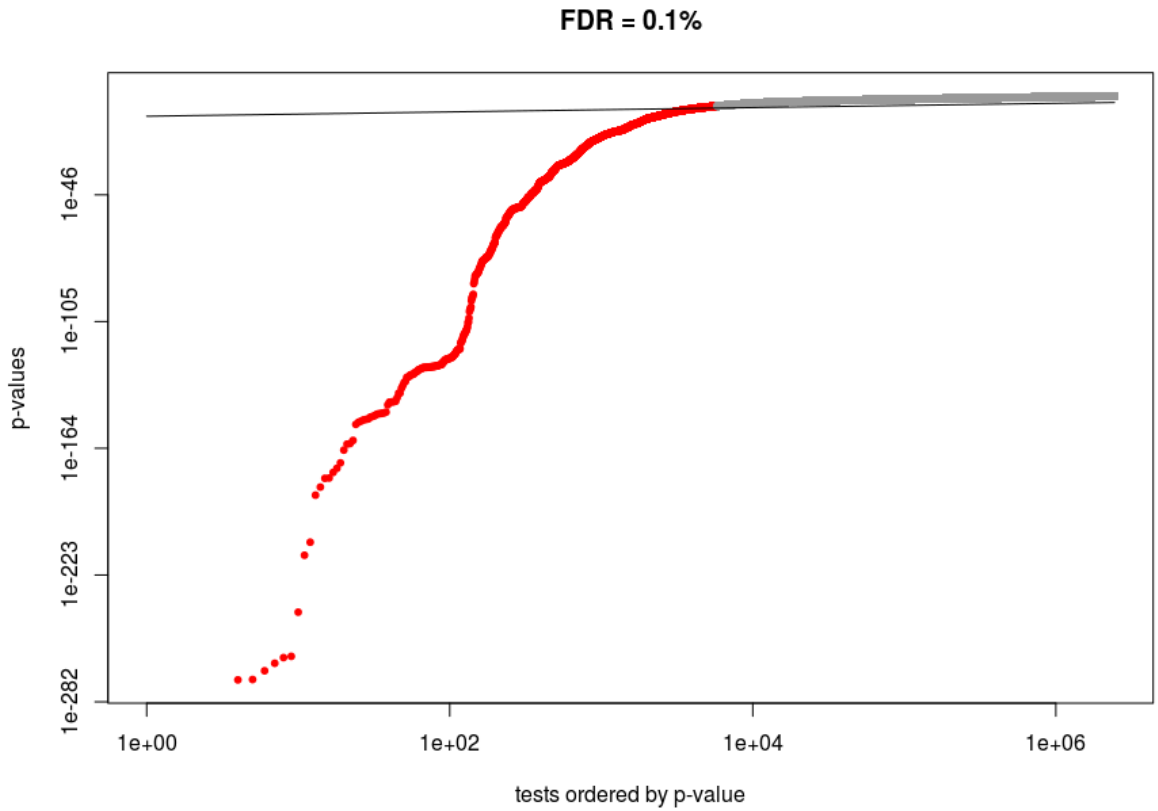


The tiny spike down by zero is our only hope for discovery.

Recall: p-values from the null distribution are uniform.

Controlling the False Discovery Rate

The slope of the FDR-cut line is q/[# of variables].



Lots of action!

4000 significant tests at FDR of 1e−3

(so only 4-5 are false discoveries).

All there is to know about FDR

p-values from the null distribution are uniform, and N of them ranked and plotted should lie along a line with slope $1/N$.

FDP is the number of false discoveries divided by the number of test results called significant. You don't know the FDP.

You can control $FDR = \mathbb{E}[FDP]$ to be $\leq q$ amongst p tests

- rank and plot p-values against rank$/p$
- draw a line with slope q/p
- find the max point where p-values cross this line, and use that point to set your rejection region.

R roundup

Quitting R: usually save the script, not the workspace.
The workspace is an .rda image, while the .R script is just text.

Stay up-to-date: Make sure you have latest versions.
Stay organized: keep track of your work.
Keep things simple, and don't panic.

Consider using shared drives to collaborate. Or, even better, create a group github repo and use version control.

R is a fantastic tool and there is tons you can do, but don't worry about learning everything right away.

Plenty of resources out there. Also don't hesitate to use me and your friends (and google!) to avoid frustration.

Amazon Reviews

The dataset consists of 13 319 reviews for selected products on Amazon from Jan-Oct 2012.

Reviews include *product information*, *ratings*, *and a plain text review*.

We will look for words associated with good/bad ratings.

Week 1 Homework

See `HW1_start.R` for code to get you started.

Assignment in `HW1_assignment.pdf`

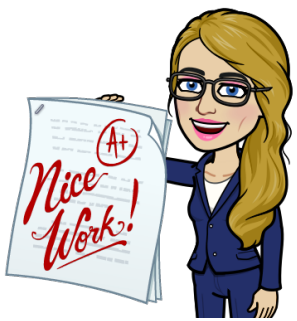
You are encouraged to use R markdown

R markdown sample code in `HW1_assignment.rmd`

You are encouraged to work in groups of 5.

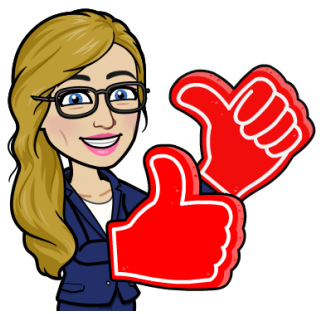
Tell me what you find (and how).

Homeworks are out of 5:



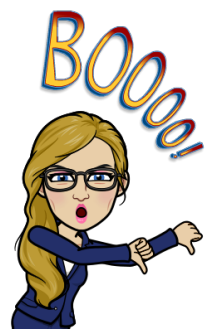
Amazing

Great



Good

OK



Ouch