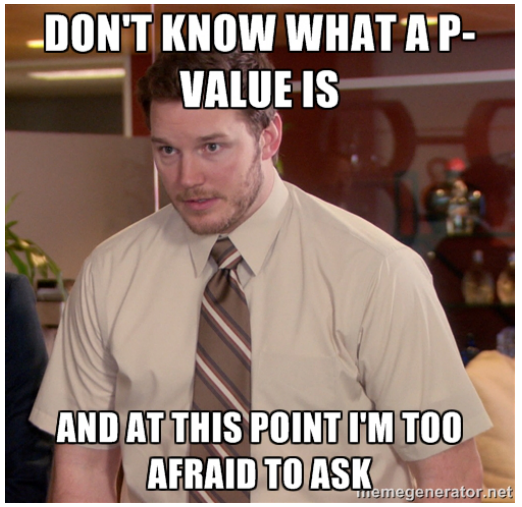
Decision Analysis (DA) Prep

mm/dd/yyyy - Steve Mortimer, XXXXXXXXXXX, XXXXXXXX

The Purpose or Goal

What do we hope to accomplish?

- Give you some statistics basics
- Introduce the core DA concepts, but slower & more background
- Answer any questions/concerns



We are not here to give you
"the answers" to the DA courses!

Source: Wikipedia

Agenda

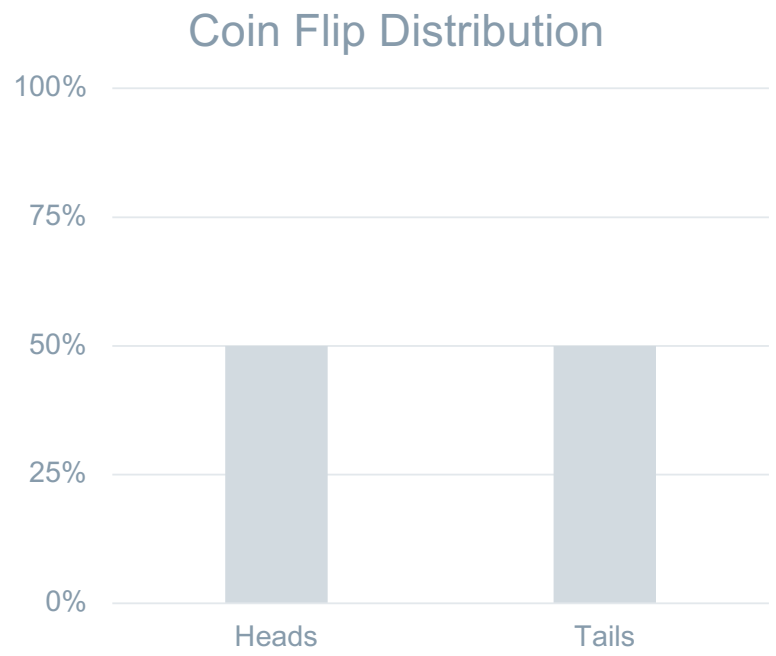
The items we're covering

1. **Distributions**
2. Statistical Tests
3. Regression
4. Decision (Logic) Trees

What is a Distribution?

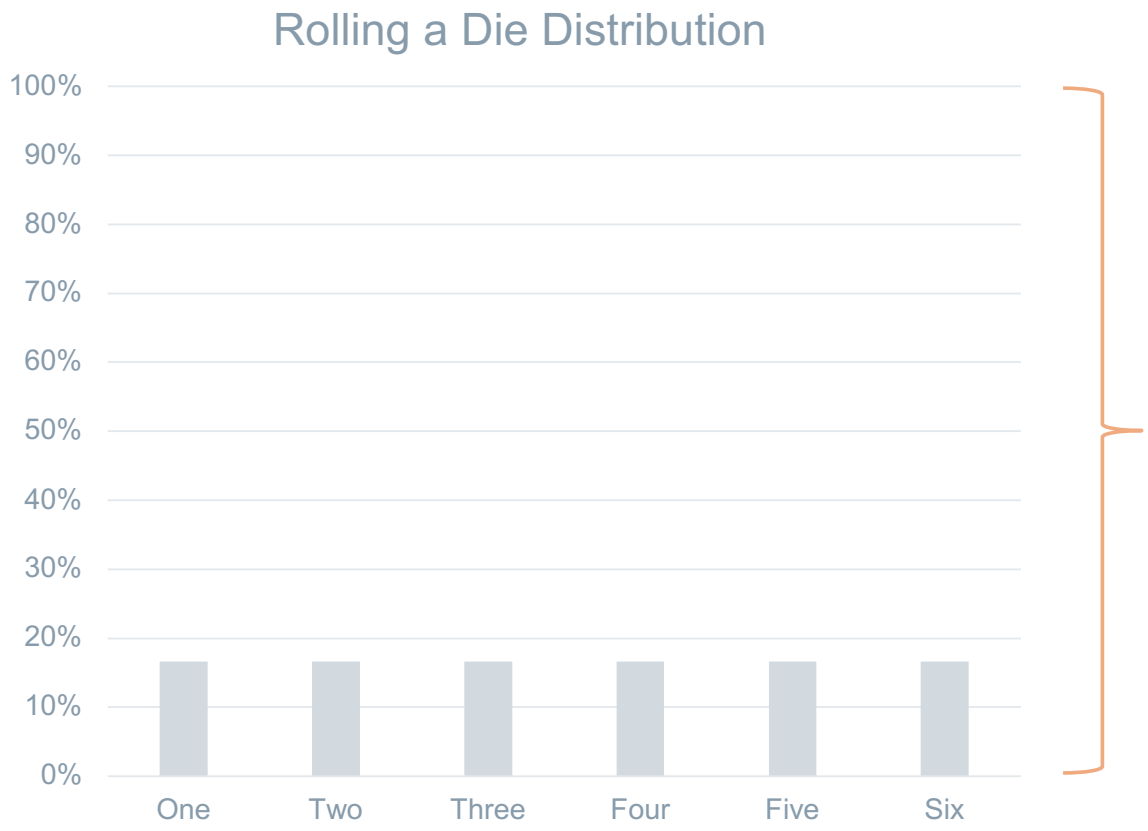
Consider flipping a coin

A distribution shows the number times each possible outcome will occur relative to one another



What is a Distribution?

Consider rolling a die



Why all the way to 100%?

Distributions cover **ALL** possible outcomes so they always sum to 100%

Continuous vs. Discrete Distributions

Classifying different types of distributions

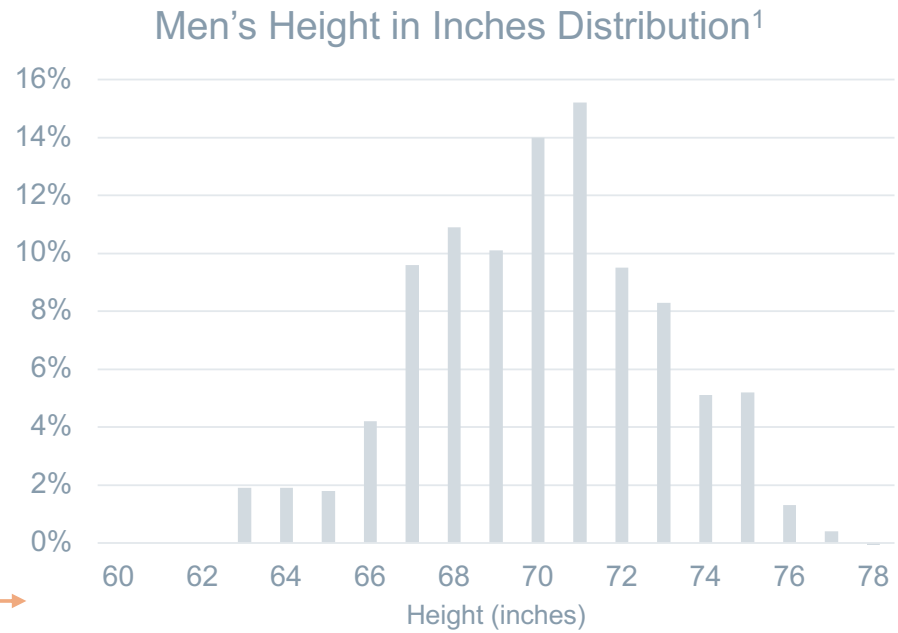
Discrete

A finite (countable) number of possible outcomes

Continuous

An infinite* number of possible outcomes

* Typically something that is a number and has many possible values even if we are able to count every possible value



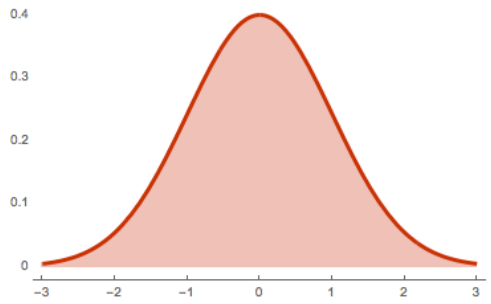
1) https://www2.census.gov/library/publications/2010/compendia/statab/130ed/tables/11s0205.pdf

This is so boring.
What does this have to do with DA?!

The Distributions You Need to Know

Defined by their shape

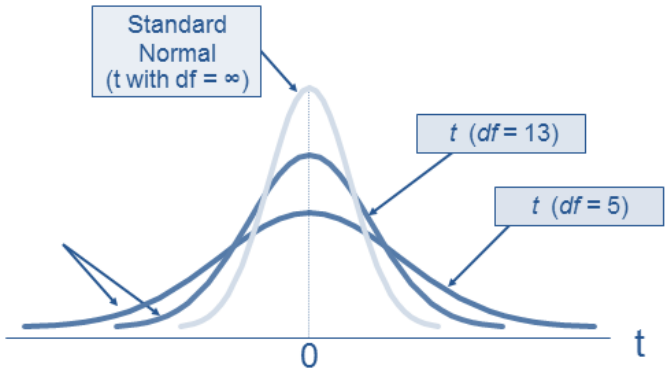
Normal ("Z")



Summary:

- "Bell shaped"
- Symmetric (same on both sides)
- Defined by mean (center) and variance (spread)

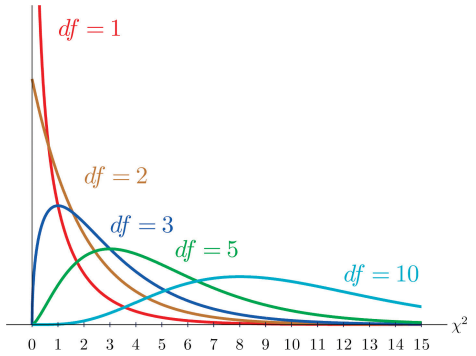
Student's T ("T")



Summary:

- A little wider than Normal dist.
- Also symmetric
- Defined by the degrees of freedom

Chi-Squared (χ^2)



Summary:

- "Long tailed"
- **Not** symmetric
- Also defined by its degrees of freedom

Agenda

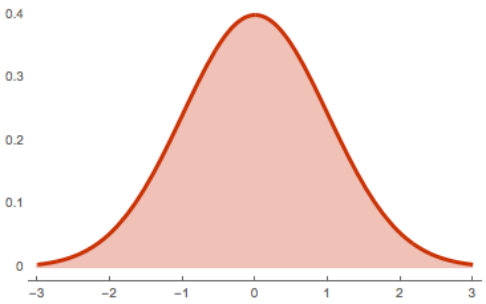
The items we're covering

1. Distributions
2. **Statistical Tests**
3. Regression
4. Decision (Logic) Trees

You Can't Do Statistical Tests Without Distributions!

Matching the distributions to statistical tests

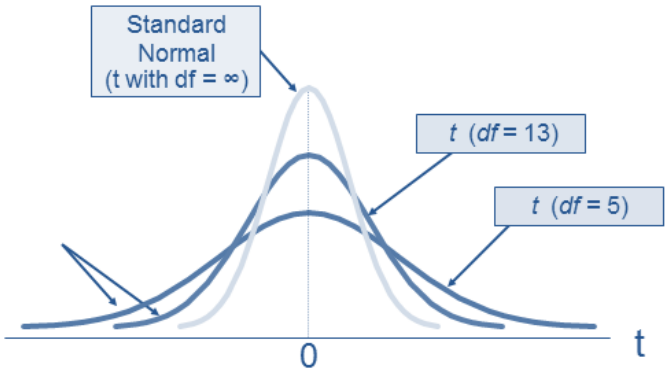
Normal ("Z")



One Sample Test ("Z-test")

Is our one sample different from what we expect?

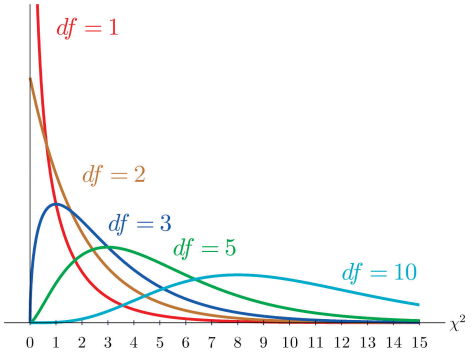
Student's T ("T")



Two Sample Test ("T-test")

Are our two samples different from each other?

Chi-Squared (χ^2)



Test of Independence & Goodness of Fit

Are two variables independent of each other

– OR –

Is the distribution a close fit to another distribution?

One Sample Test with Normal Distribution

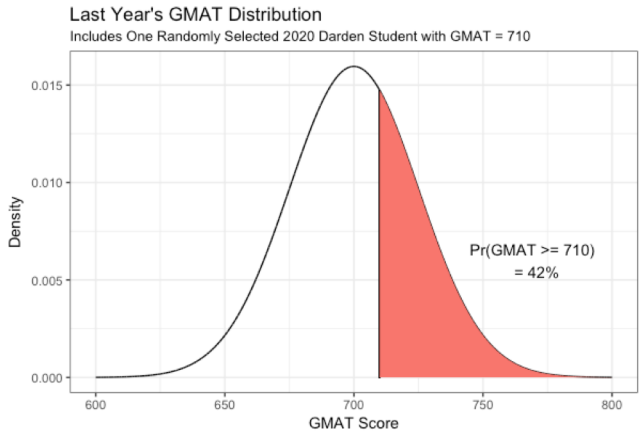
Are Darden students' GMAT scores higher than last year's students' scores?

Last year the average was 700. We are told that GMAT scores have a standard deviation = 25.

1 Student: 710 GMAT

GMAT = 710 is 0.4 standard deviations above the last year's average:

$$\frac{(710 - 700)}{25} = \frac{10}{25} = 0.4$$

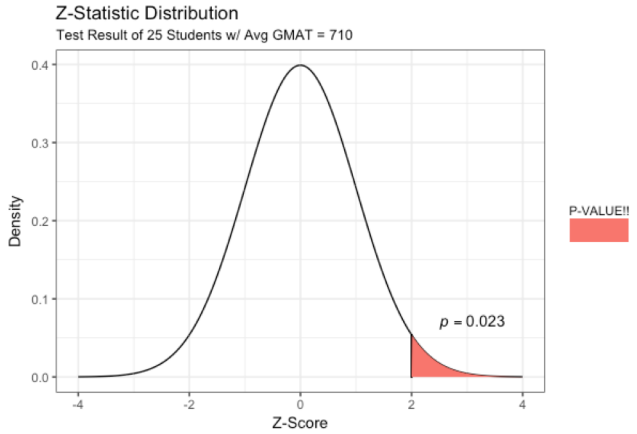


25 Students: 710 GMAT

Having an average GMAT = 710 across 25 students is 2 standard deviations above last year's average:

$$\frac{(710 - 700)}{25/\sqrt{25}} = \frac{10}{5} = 2$$

WHAT HAPPENED HERE?!?

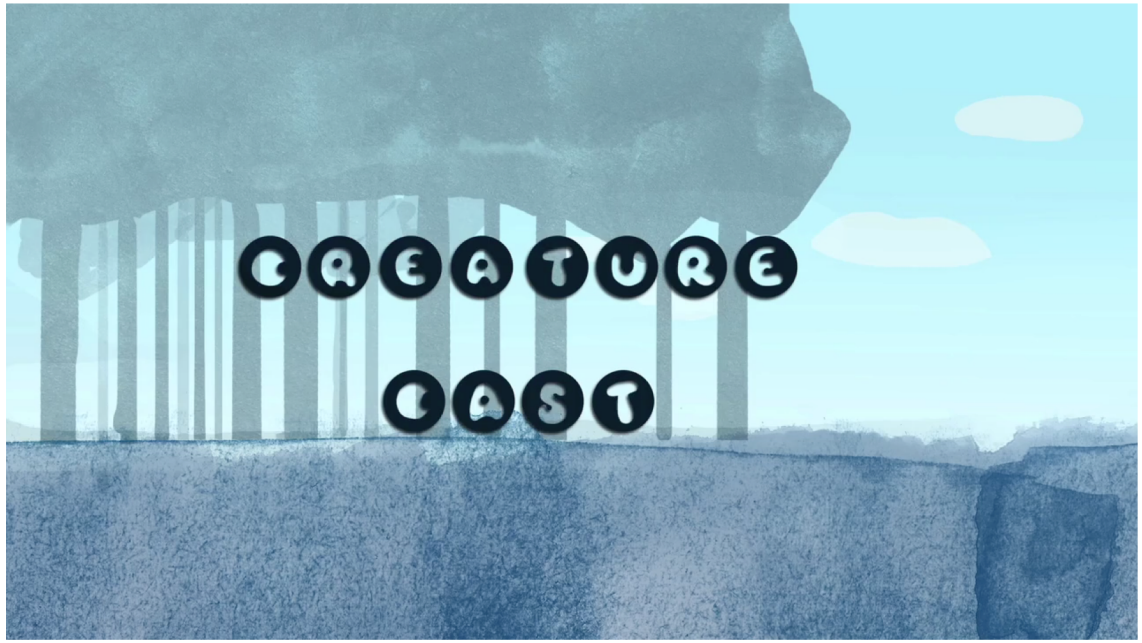


The Central Limit Theorem

More observations in a sample shrinks the std. dev. of the average

Test Statistic ("Z")

$$Z = \frac{(Sample\ Mean\ - \mu)}{\frac{\sigma}{\sqrt{n}}}$$



https://vimeo.com/75089338

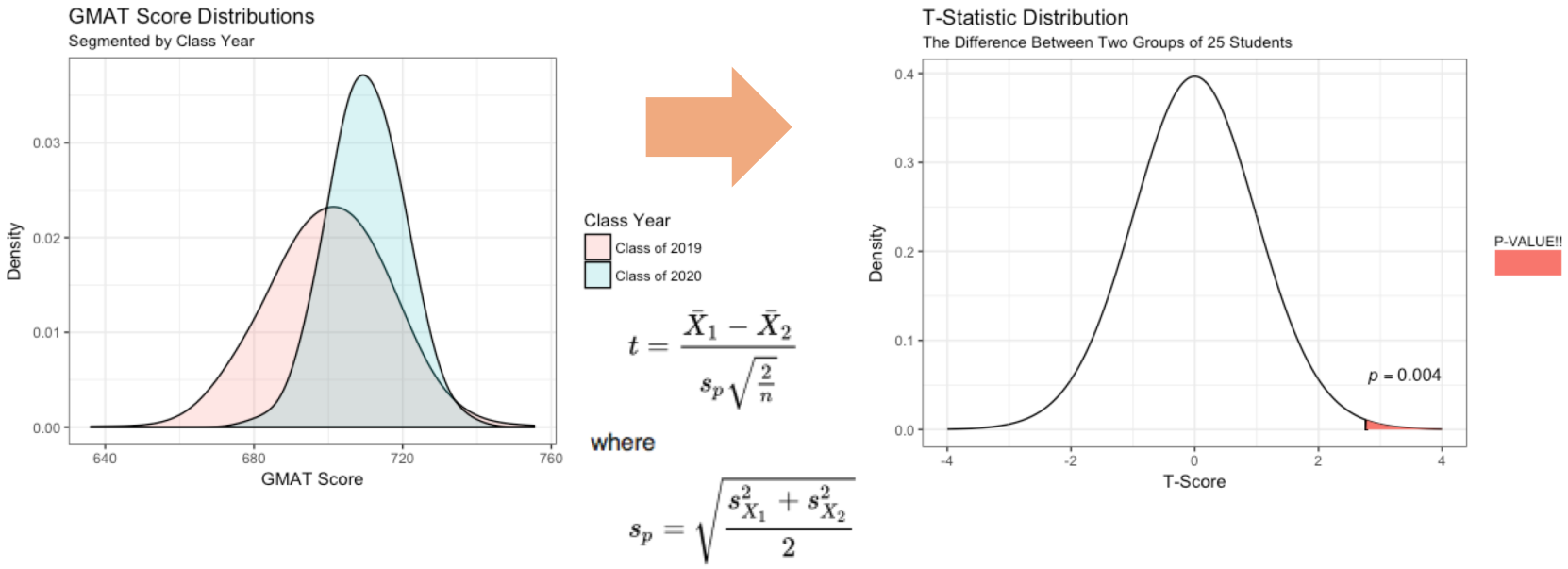
Two Sample Test (“T”)

How to draw a comparison if both groups are samples

Scenario:

We randomly sampled 25 students from Class of 2019 and found – GMAT Avg: 700, Std. Dev: 15

We sampled 25 students from Class of 2020 and found – GMAT Avg: 710, Std. Dev: 10



We use the data to create a statistic with a distribution we know and see where it falls!

Chi-Squared Test ("χ^2")

Comparing Counts of Data

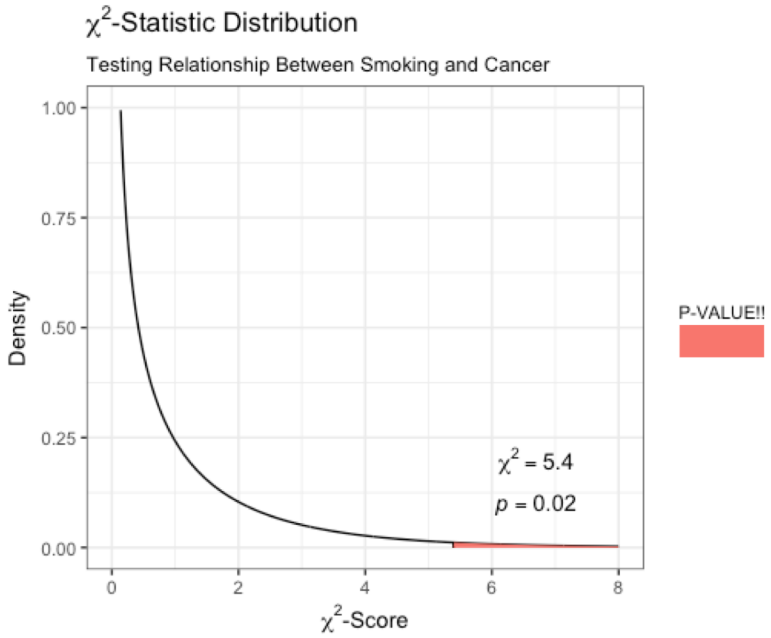
Counts of things are always zero or more (positive) so there's a different distribution for positive data

Observed	No Cancer	Cancer	**Total**
Non-Smoker	37	13	**50**
Smoker	27	23	**50**
Total	**64**	**36**	**100**

Expected	No Cancer	Cancer	**Total**
Non-Smoker	32	18	**50**
Smoker	32	18	**50**
Total	**64**	**36**	**100**

Test Statistic (χ^2)

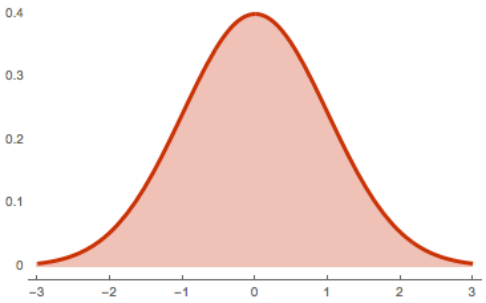
$$X^2 = \sum_{i=1}^{Rows} \sum_{j=1}^{Columns} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$



The Area Under the Distribution

How to get the area using only Excel

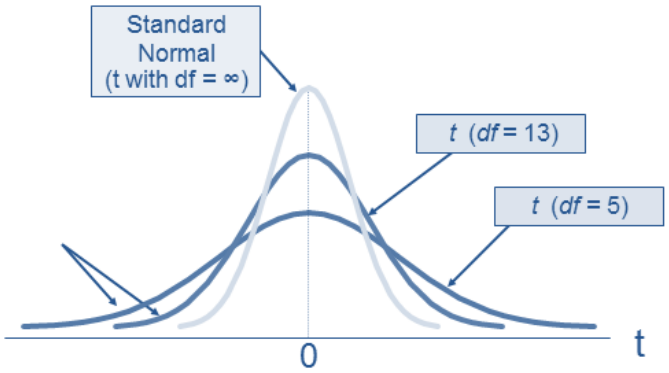
Normal ("Z")



GMAT One Sample Example:

1 – NORMSDIST(2) = 0.023

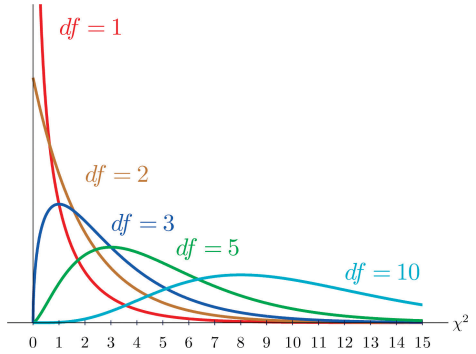
Student's T ("T")



GMAT Two Sample Example:

1 – TDIST(2.7735 ,48) = 0.004

Chi-Squared (χ^2)



Smoking/Cancer Example:

CHIDIST(5.4, 1) = 0.02

Statistical Tests are Easy! Decide on the statistic you want to use, then figure out area under curve

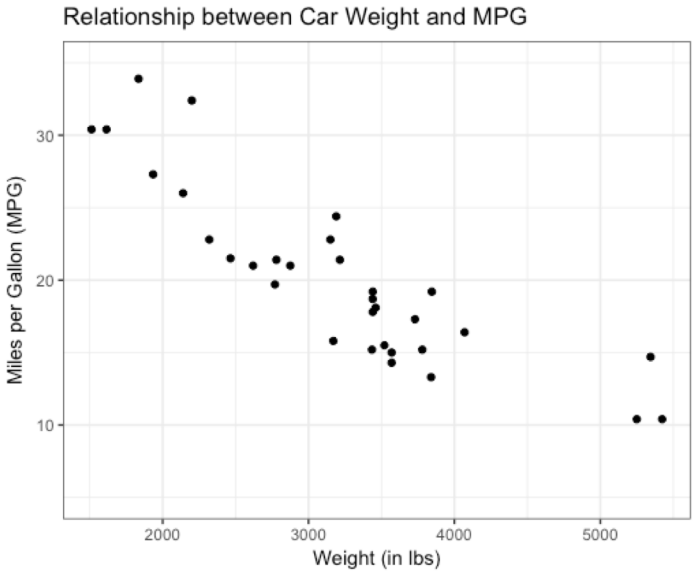
Agenda

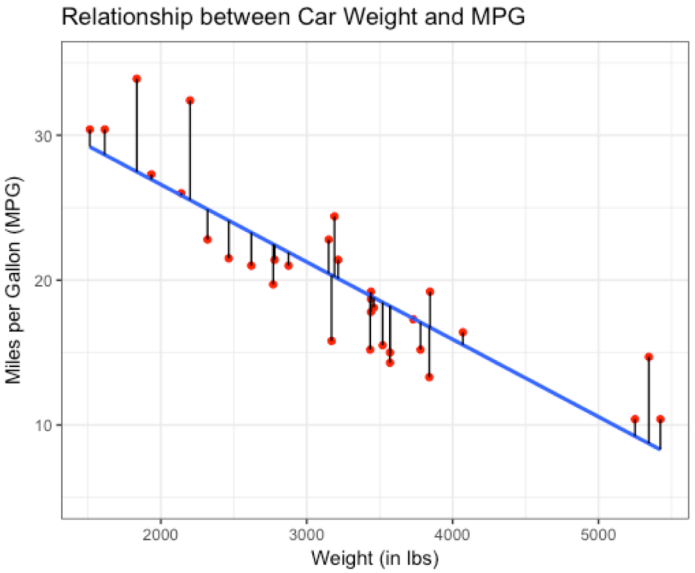
The items we're covering

1. Distributions
2. Statistical Tests
3. **Regression**
4. Decision (Logic) Trees

Linear Regression

Fit a line between points





$$MPG = 37.285 - .005344 * Weight$$

"Intercept" (where line crosses if Weight is zero)

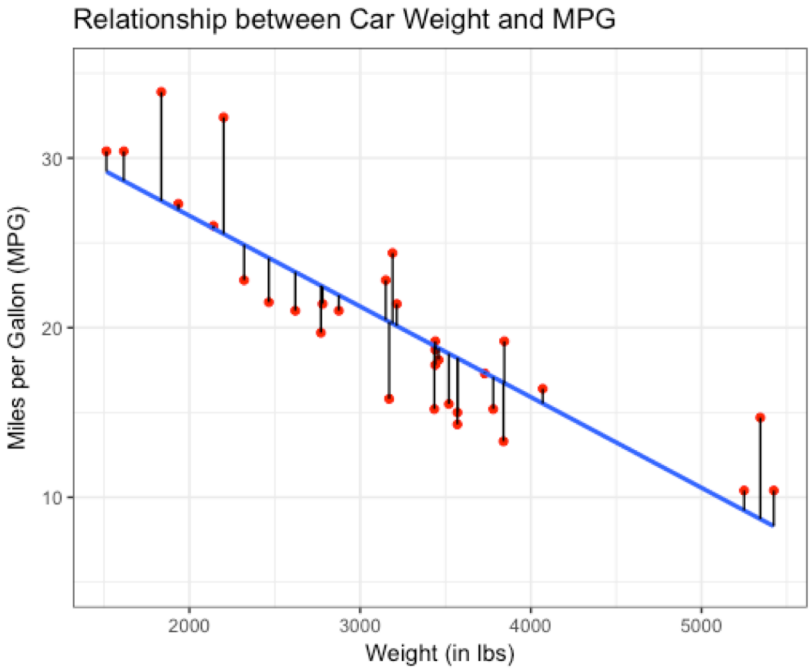
"Coefficient" (slope or change that every pound of weight affects the MPG)

What if a car weights 3,000lbs? What would the MPG be?

$$37.285 - .005344 * 3000 = 21.25mpg$$

Validating Your Regression Model

It involves more than just fitting a line



$$MPG = 37.285 - .005344 * Weight$$

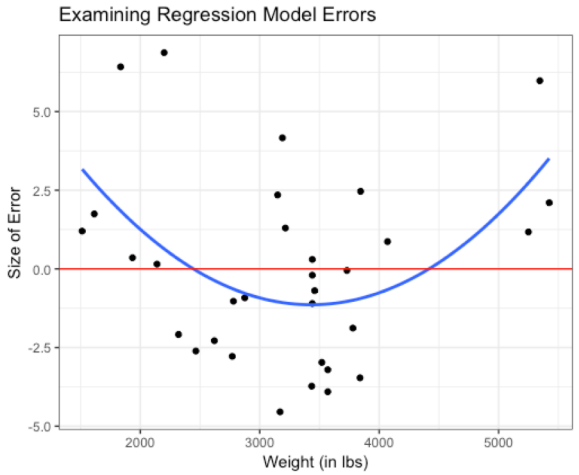
$$R^2 = .7528 \qquad Standard\ Error = 3.046$$

Is *Weight* significantly related to *MPG*? Or is it a fluke?

Coefficients are distributed based on the T-distribution, so we can do a statistical test on them to see if significant!

$$T = -9.56 \rightarrow p = 0.000000000129$$

When my model is wrong (residual), is it always wrong in the same way (systematic error)?



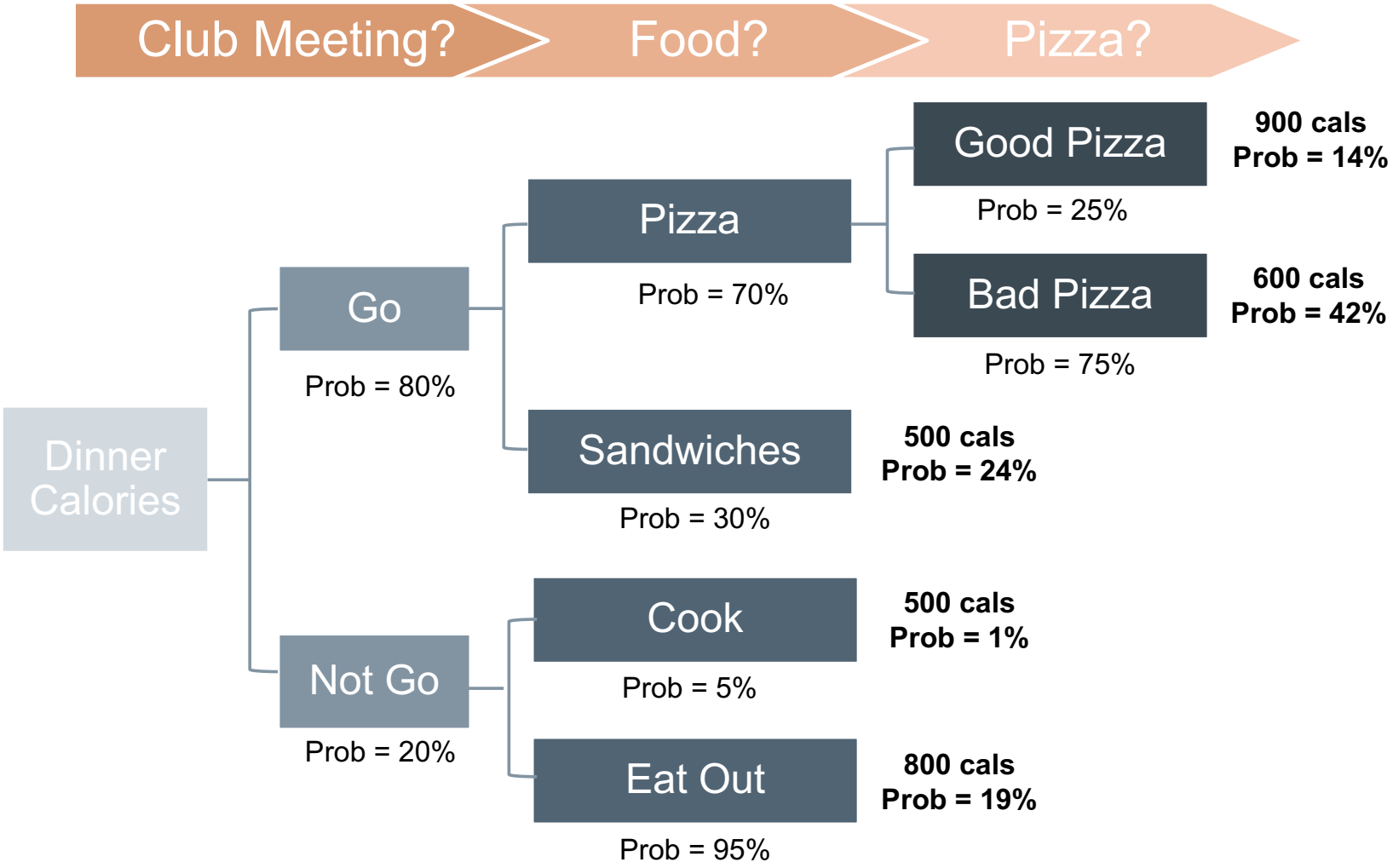
Agenda

The items we're covering

1. Distributions
2. Statistical Tests
3. Regression
4. **Decision (Logic) Trees**

Estimating Your Average Dinner Calories

Use a Decision (Logic) Tree to assess the possibilities



Decision (Logic) Tree Outputs

Shows us the range of outcomes and expected

Questions?

In Person: Ask Now!

DSC Email: DardenDSC@darden.virginia.edu

Steve Mortimer: MortimerS19@darden.virginia.edu


DATA
SCIENCE
CLUB