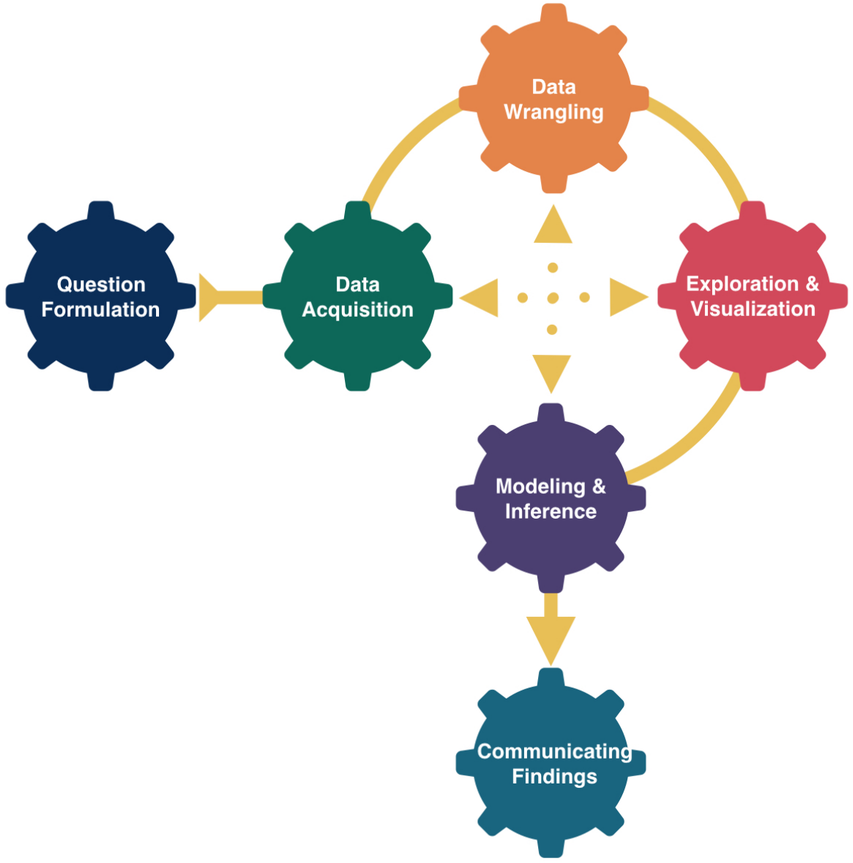


Model Guidance

Kelly McConville

Stat 100

Week 8 | Fall 2023

Announcements

- Oct 30th: Hex or Treat Day in Stat 100
  - Wear a Halloween costume and get either a hex sticker or candy!!

Goals for Today

- Finish up: Regression with polynomial explanatory variables
- Modeling guidance
- Sampling variability
- Sampling distributions

Linear Regression

Model Form:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Linear regression is a flexible class of models that allow for:

- Both quantitative and categorical **explanatory** variables.
- **Multiple** explanatory variables.
- **Curved** relationships between the response variable and the explanatory variable.
- BUT the **response variable is quantitative**.

Example: Movies

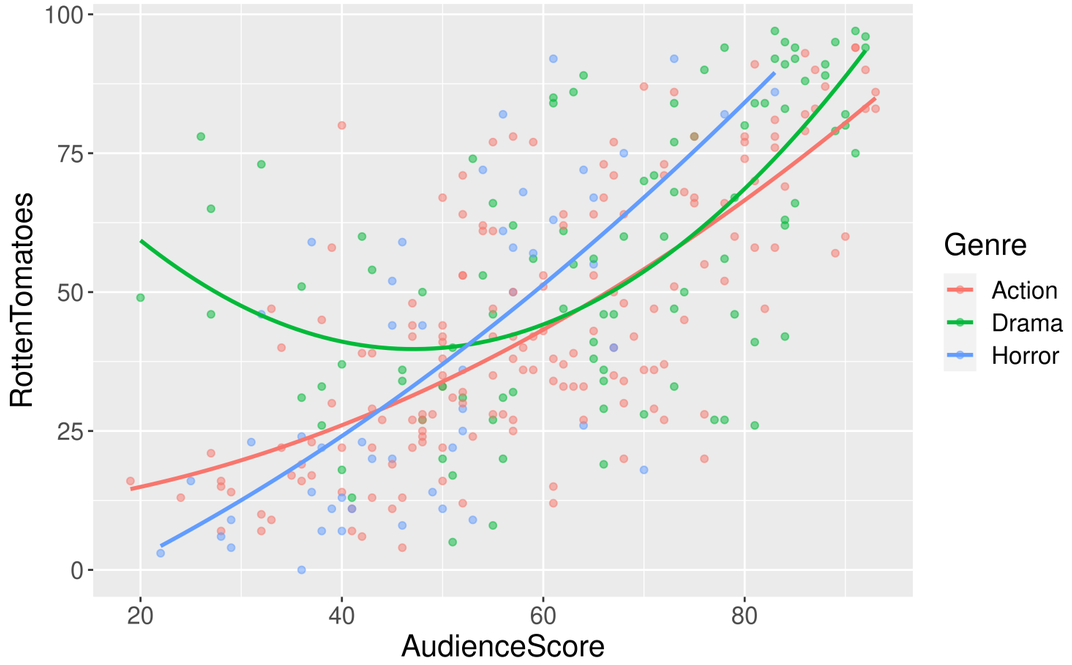
Let's model a movie's critic rating using the audience rating and the movie's genre.

```
1 library(tidyverse)
2 movies <- read_csv("https://www.lock5stat.com/datasets2e/HollywoodMovies.csv")
3
4 # Restrict our attention to dramas, horrors, and actions
5 movies2 <- movies %>%
6   filter(Genre %in% c("Drama", "Horror", "Action")) %>%
7   drop_na(Genre, AudienceScore, RottenTomatoes)
8 glimpse(movies2)
```

```
Rows: 313
Columns: 16
$ Movie            <chr> "Spider-Man 3", "Transformers", "Pirates of the Carib…
$ LeadStudio       <chr> "Sony", "Paramount", "Disney", "Warner Bros", "Warner…
$ RottenTomatoes   <dbl> 61, 57, 45, 60, 20, 79, 35, 28, 41, 71, 95, 42, 18, 2…
$ AudienceScore    <dbl> 54, 89, 74, 90, 68, 86, 55, 56, 81, 52, 84, 55, 70, 6…
$ Story            <chr> "Metamorphosis", "Monster Force", "Rescue", "Sacrific…
$ Genre            <chr> "Action", "Action", "Action", "Action", "Action", "Ac…
$ TheatersOpenWeek <dbl> 4252, 4011, 4362, 3103, 3778, 3408, 3959, 3619, 2911,…
$ OpeningWeekend   <dbl> 151.1, 70.5, 114.7, 70.9, 49.1, 33.4, 58.0, 45.3, 19.…
$ BOAvgOpenWeekend <dbl> 35540, 17577, 26302, 22844, 12996, 9791, 14663, 12541…
$ DomesticGross    <dbl> 336.53, 319.25, 309.42, 210.61, 140.13, 134.53, 131.9…
$ ForeignGross     <dbl> 554.34, 390.46, 654.00, 245.45, 117.90, 249.00, 157.1…
$ WorldGross       <dbl> 890.87, 709.71, 963.42, 456.07, 258.02, 383.53, 289.0…
$ Budget           <dbl> 258.0, 150.0, 300.0, 65.0, 140.0, 110.0, 130.0, 110.0…
```

Coming Back to Our Exploratory Data Analysis

```
ggplot(data = movies2,
       mapping = aes(x = AudienceScore,
                     y = RottenTomatoes,
                     color = Genre)) +
  geom_point(alpha = 0.5) +
  stat_smooth(method = lm, se = FALSE,
        formula = y ~ poly(x, degree = 2))
```



Fitting the Polynomial Model

```
mod2 <- lm(RottenTomatoes ~ poly(AudienceScore, degree = 2, raw = TRUE)*Genre, data = movies2)
library(moderndive)
get_regression_table(mod2, print = TRUE)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	9.922	14.851	0.668	0.505	-19.301	39.145
poly(AudienceScore, degree = 2, raw = TRUE)1	0.098	0.515	0.191	0.849	-0.916	1.113
poly(AudienceScore, degree = 2, raw = TRUE)2	0.008	0.004	1.788	0.075	-0.001	0.016
Genre: Drama	88.923	24.538	3.624	0.000	40.638	137.208
Genre: Horror	-23.767	31.054	-0.765	0.445	-84.876	37.342
poly(AudienceScore, degree = 2, raw = TRUE)1:GenreDrama	-2.608	0.840	-3.107	0.002	-4.260	-0.956
poly(AudienceScore, degree = 2, raw = TRUE)2:GenreDrama	0.019	0.007	2.785	0.006	0.006	0.032

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
poly(AudienceScore, degree = 2, raw = TRUE)1:GenreHorror	0.574	1.223	0.469	0.639	-1.833	2.981
poly(AudienceScore, degree = 2, raw = TRUE)2:GenreHorror	-0.001	0.012	-0.061	0.951	-0.024	0.022

Linear Regression & Curved Relationships

Form of the Model:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

But why is it called **linear** regression if the model also handles for curved relationship??

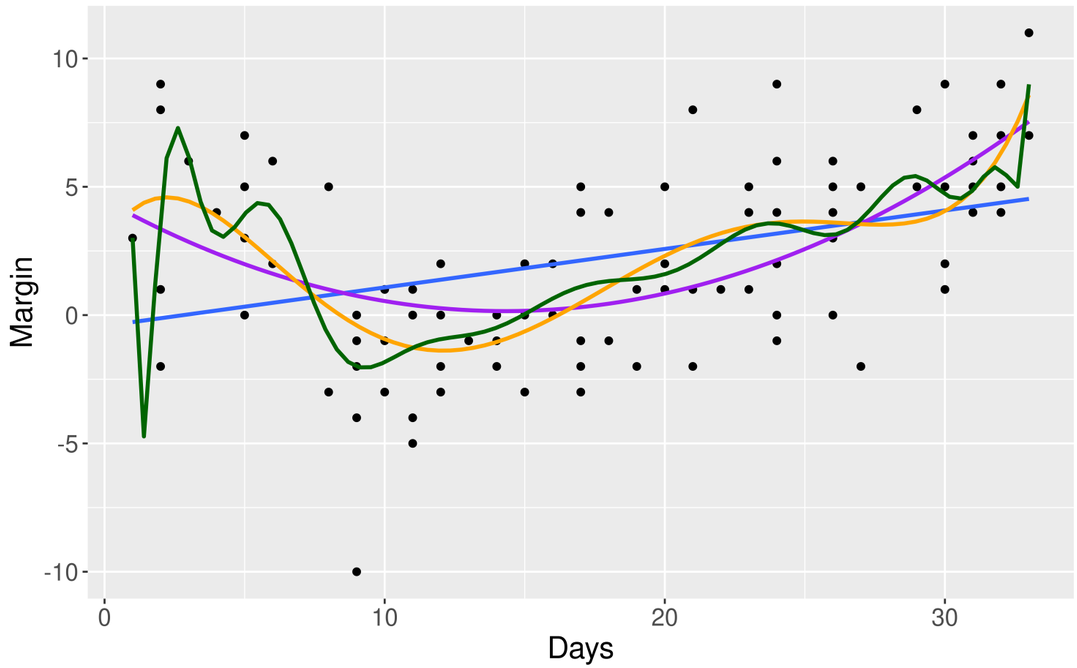
Model Building Guidance

What degree of polynomial should I include in my model?

Guiding Principle: Capture the general trend, not the noise.

$$y = f(x) + \epsilon$$
$$y = \text{TREND} + \text{NOISE}$$

Returning the 2008 Election Example:



Model Building Guidance

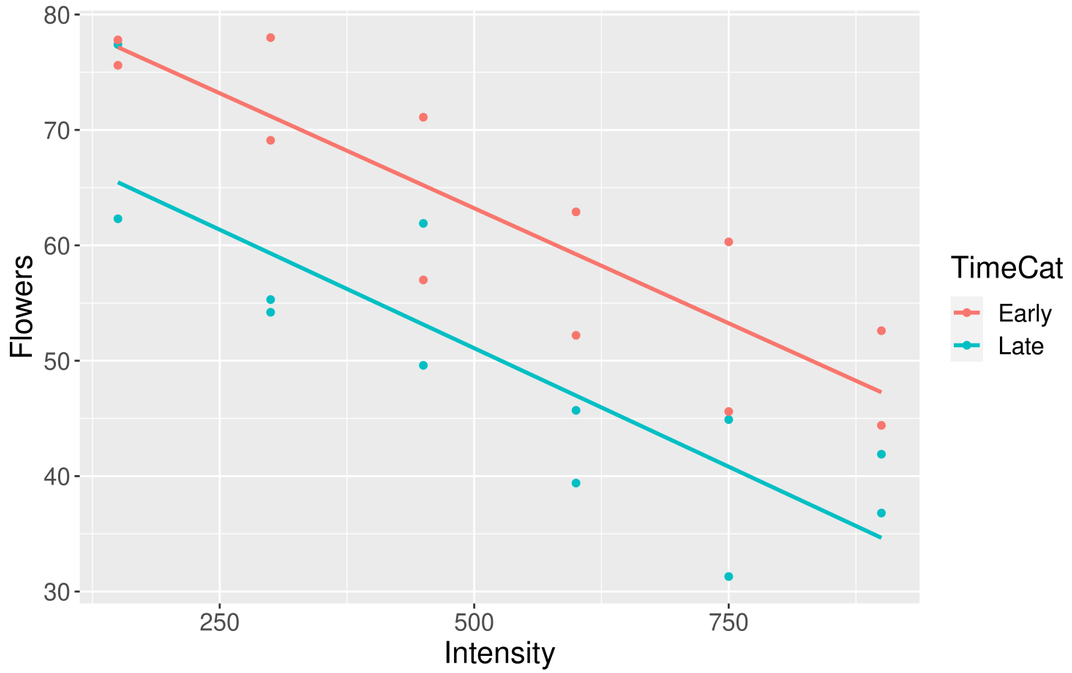
Shouldn't we always include the interaction term?

Guiding Principle: Occam's Razor for Modeling

> "All other things being equal, simpler models are to be preferred over complex ones." – ModernDive

Guiding Principle: Consider your modeling goals.

- The equal slopes model allows us to control for the intensity of the light and then see the impact of being in the early or late timing groups on the number of flowers.



- Later in the course will learn statistical procedures for determining whether or not a particular term should be included in the model.

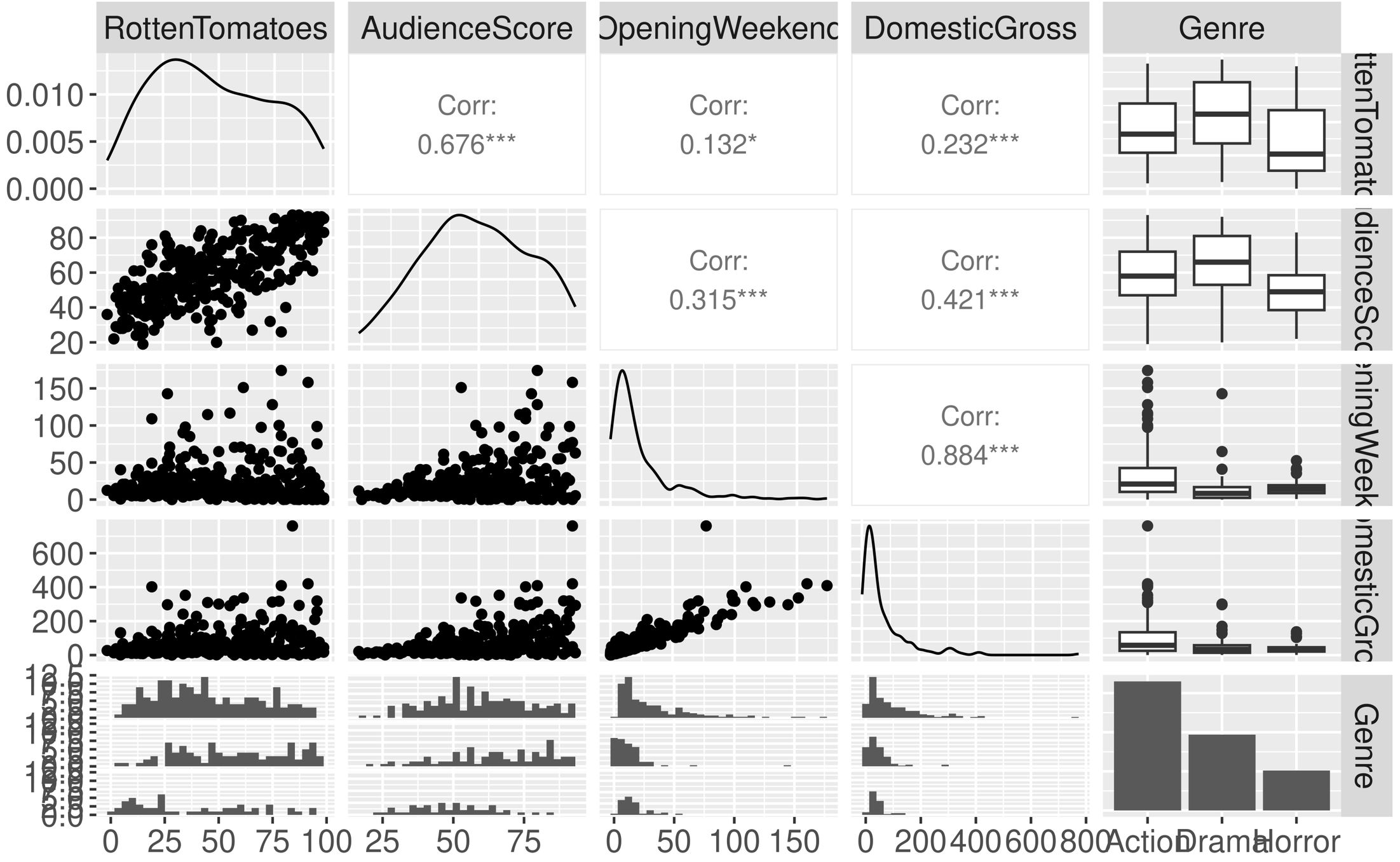
What if I want to include more than 2 explanatory variables??

Model Building Guidance

We often have several potential explanatory variables. How do we determine which to include in the model and in what form?

Guiding Principle: Include explanatory variables that attempt to explain **different** aspects of the variation in the response variable.

```
library(GGally)
movies2 %>%
  select(RottenTomatoes, AudienceScore, OpeningWeekend,
         DomesticGross, Genre) %>%
  ggpairs()
```


RottenTomatoes
AudienceScore
OpeningWeekend
DomesticGross
Genre
Corr:
0.676***
Corr:
0.132*
Corr:
0.232***
Corr:
0.315***
Corr:
0.421***
Corr:
0.884***
0.010
0.005
0.000
80
60
40
20
150
100
50
0
600
400
200
0
0 25 50 75 100
25 50 75
0 50 100 150
0 200 400 600 800
Action
Drama
Horror

Model Building Guidance

We often have several potential explanatory variables. How do we determine which to include in the model and in what form?

Guiding Principle: Include explanatory variables that attempt to explain **different** aspects of the variation in the response variable.

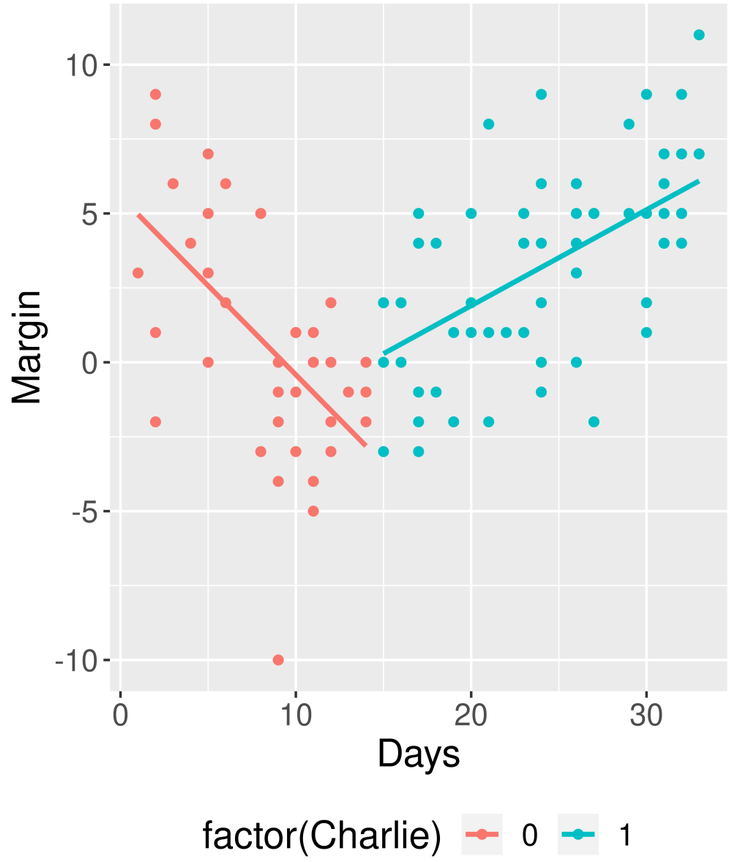
```
mod_movies <- lm(RottenTomatoes ~ AudienceScore + DomesticGross + Genre, data = movies2)
get_regression_table(mod_movies, print = TRUE)
```

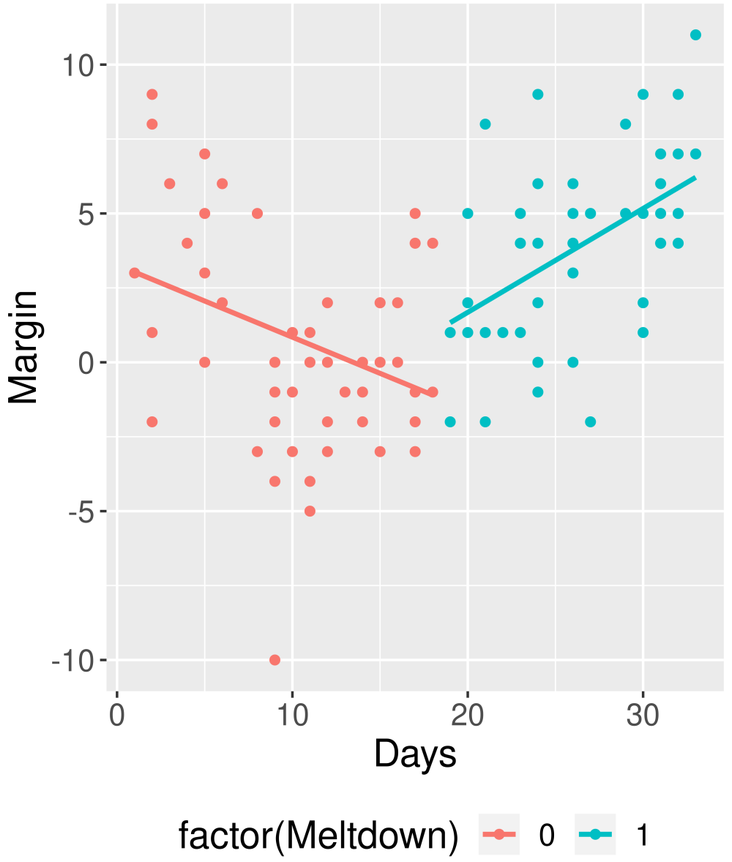
term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-12.472	4.083	-3.055	0.002	-20.506	-4.439
AudienceScore	0.975	0.072	13.590	0.000	0.834	1.117
DomesticGross	-0.006	0.015	-0.431	0.667	-0.035	0.023
Genre: Drama	6.117	2.644	2.314	0.021	0.916	11.319
Genre: Horror	2.058	3.141	0.655	0.513	-4.121	8.238

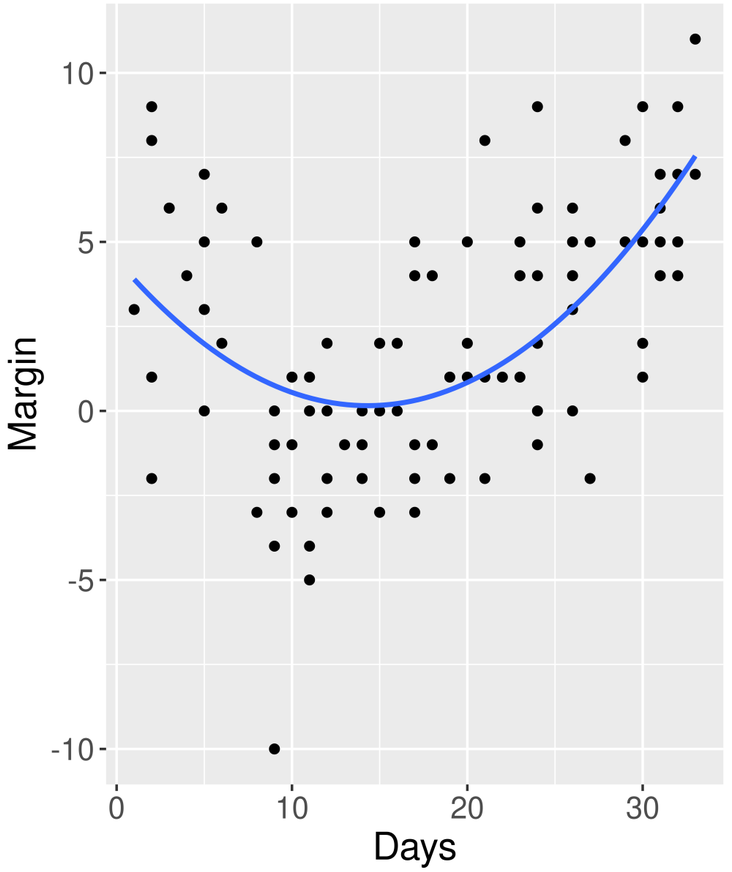
Model Building Guidance

We often have several potential explanatory variables. How do we determine which to include in the model and in what form?

Guiding Principle: Use your modeling motivation to determine how much you weigh **interpretability** versus **prediction accuracy** when choosing the model.


Margin
Days
10
5
0
-5
-10
0
10
20
30
factor(Charlie)
0
1


Margin
Days
10
5
0
-5
-10
0
10
20
30
factor(Meltdown)
0
1

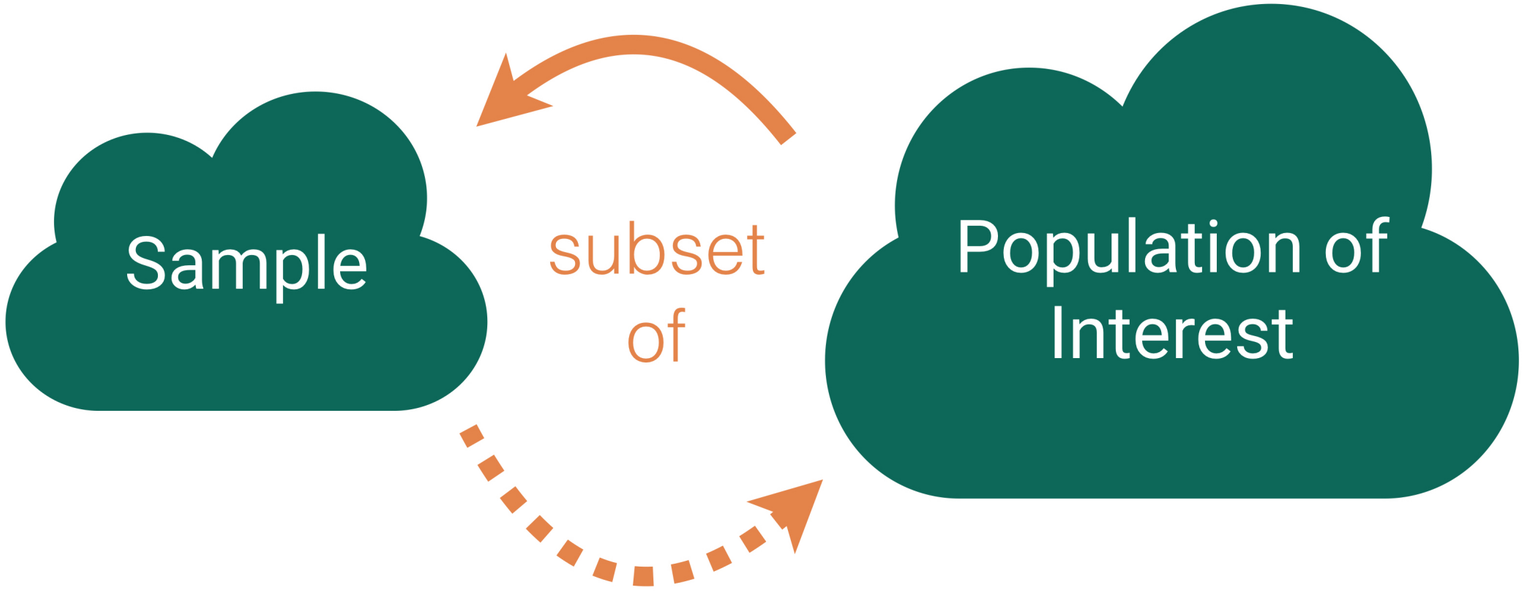

Margin
Days
10
5
0
-5
-10
0
10
20
30

Model Building

- We will come back to methods for model selection.
- Key ideas:
  - Determining the **response** variable and the potential **explanatory** variable(s)
  - Writing out the **model form** and understanding the terms
  - **Building** and **visualizing** linear regression models in **R**
  - **Comparing** different potential models

Shift Gears: Statistical Inference

The ❤️ of statistical inference is quantifying uncertainty




```r
library(tidyverse)
ce <- read_csv("data/fmli.csv")
summarize(ce, meanFINCBTAX = mean(FINCBTAX))
```

```
# A tibble: 1 × 1
  meanFINCBTAX
         <dbl>
1       62480.
```

The ❤️ of statistical inference is quantifying uncertainty

```
1 library(tidyverse)
2 ce <- read_csv("data/fmli.csv")
3 summarize(ce, meanFINCBTAX = mean(FINCBTAX))
```

```
# A tibble: 1 × 1
  meanFINCBTAX
         <dbl>
1       62480.
```

Now we need to distinguish between the **population** and the **sample**

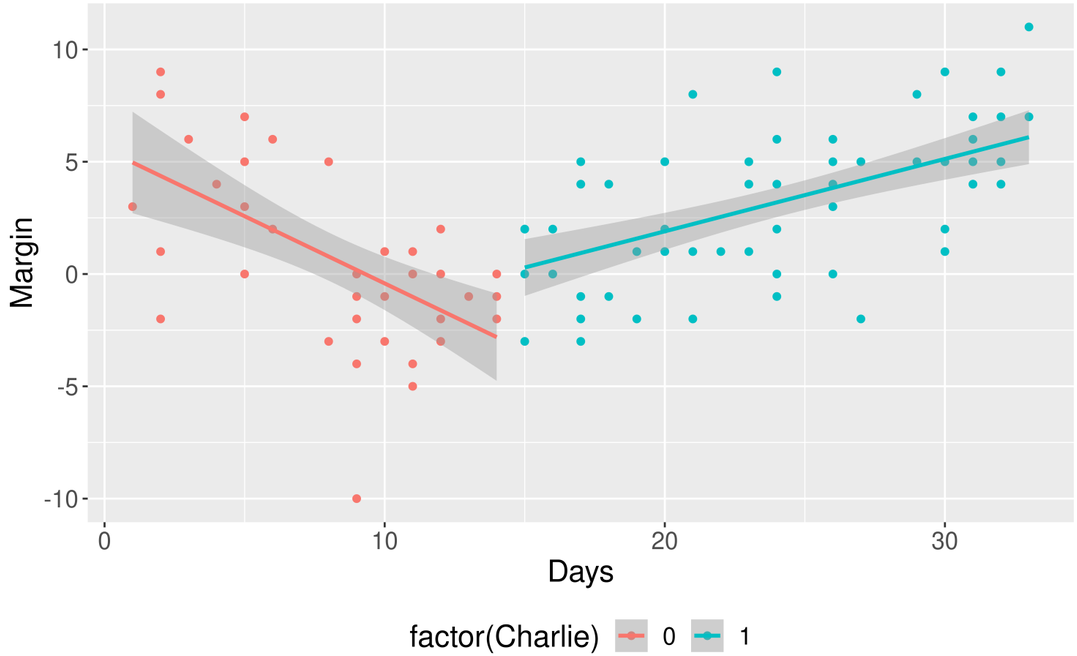
- **Parameters**:
  - Based on the **population**
  - Unknown then if don't have data on the whole population
  - EX: β_o and β_1
  - EX: μ = population mean
- **Statistics**:
  - Based on the **sample** data
  - Known
  - Usually estimate a population parameter
  - EX: $\hat{\beta}_o$ and $\hat{\beta}_1$
  - EX: $\bar{x}$ = sample mean

Quantifying Our Uncertainty

R has been giving us uncertainty estimates:

```r
library(Stat2Data)
data("Pollster08")

ggplot(Pollster08, aes(x = Days,
                       y = Margin,
                       color = factor(Charlie))) +
  geom_point() +
  stat_smooth(method = "lm", se = TRUE) +
  theme(legend.position = "bottom")
```



Quantifying Our Uncertainty

R has been giving us uncertainty estimates:

```r
library(Stat2Data)
data("Pollster08")
modPoll <- lm(Margin ~ Days*factor(Charlie), data = Pollster08)
library(moderndive)
get_regression_table(modPoll)
```

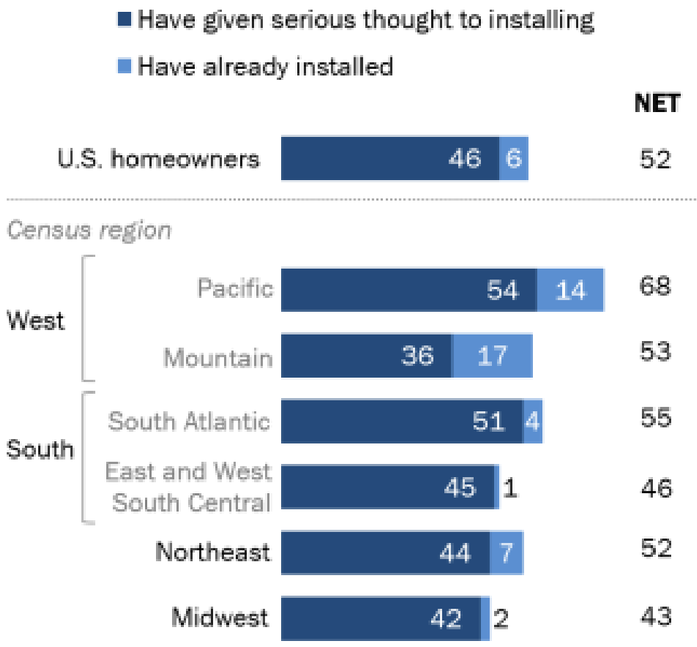
```
# A tibble: 4 × 7
  term                  estimate std_error statistic p_value lower_ci upper_ci
  <chr>                    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept                5.57      1.09       5.11       0    3.40     7.73
2 Days                    -0.598     0.121     -4.96       0   -0.838   -0.359
3 factor(Charlie): 1     -10.1       1.92      -5.25       0  -13.9     -6.29
4 Days:factor(Charlie)1    0.921     0.136      6.75       0    0.65     1.19
```

Quantifying Our Uncertainty

The news and journal articles are also giving us uncertainty estimates:

More than four-in-ten U.S. homeowners are considering residential solar panels

% of U.S. homeowners who say they ____ solar panels at home



Note: Based on homeowners. Respondents who gave other responses or did not give an answer are not shown.
Source: Survey conducted Oct. 1-13, 2019.

PEW RESEARCH CENTER

Note: The findings are based on a survey conducted Oct. 1-13, 2019, among 3,627 U.S. adults on Pew Research Center's American Trends Panel. The margin of sampling error for the full sample is plus or minus 2.1 percentage points. The margin of error for the 2,564 U.S. homeowners is plus or minus 2.5 percentage points. See full topline results.

Quantifying Our Uncertainty

The news and journal articles are also giving us uncertainty estimates:

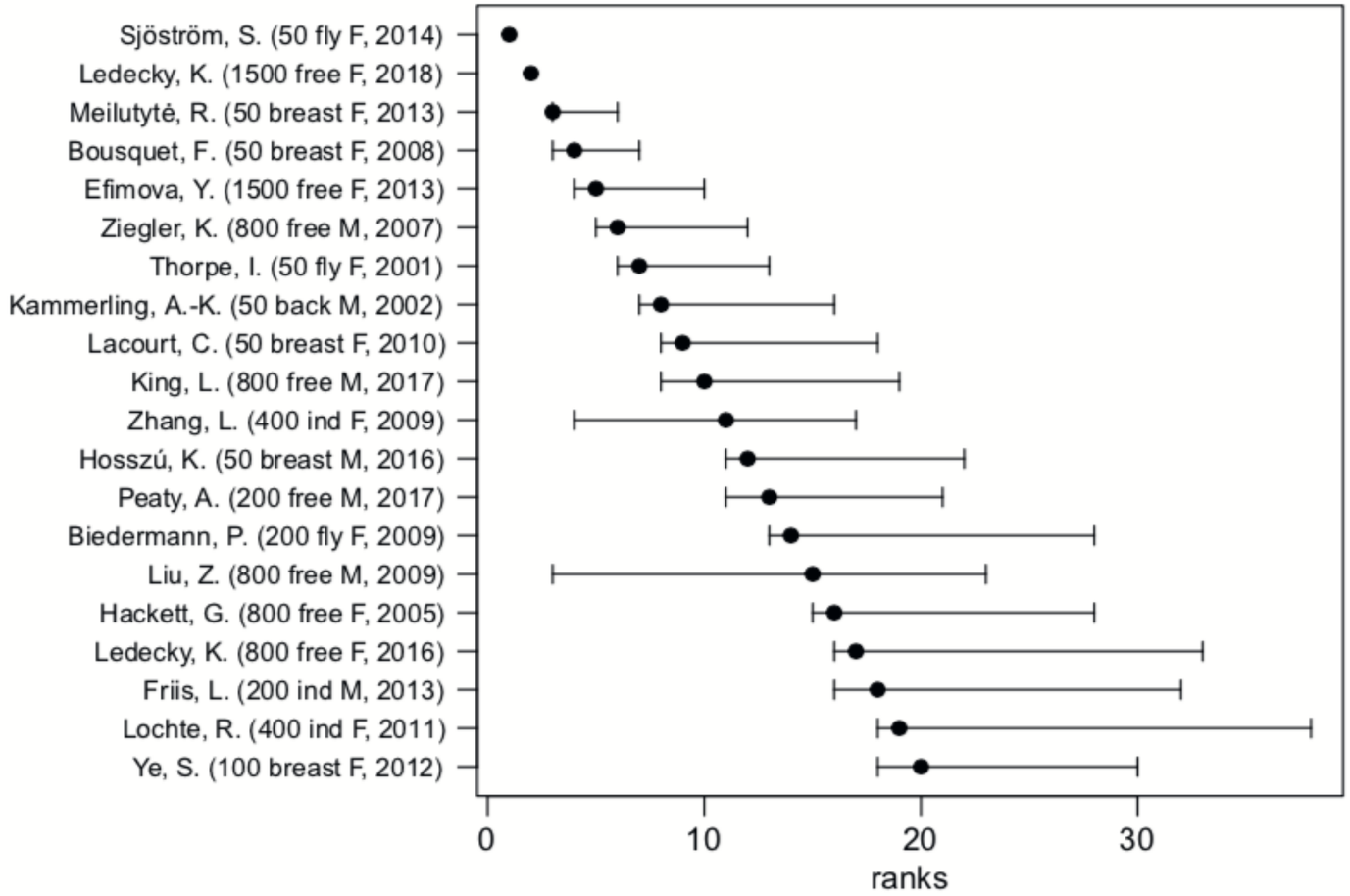


Figure 2: Ranking of the top 20 personal best swims (and their swimmers), 2001–2018, with 95% confidence intervals. Better-ranked swimmers are higher on the *y*-axis.

Statistical Inference

Goal: Draw conclusions about the population based on the sample.

Main Flavors

- Estimating numerical quantities (parameters).
- Testing conjectures.

Estimation

Goal: Estimate a (population) parameter.

Best guess?

- The corresponding (sample) statistic

Example: Are GIFs just another way for people to share videos of their pets?

via GIPHY

Want to estimate the proportion of GIFs that feature animals.

Estimation

Key Question: How accurate is the statistic as an estimate of the parameter?

Helpful Sub-Question: If we take many samples, how much would the statistic vary from sample to sample?

Need two new concepts:

- The **sampling variability** of a statistic
- The **sampling distribution** of a statistic

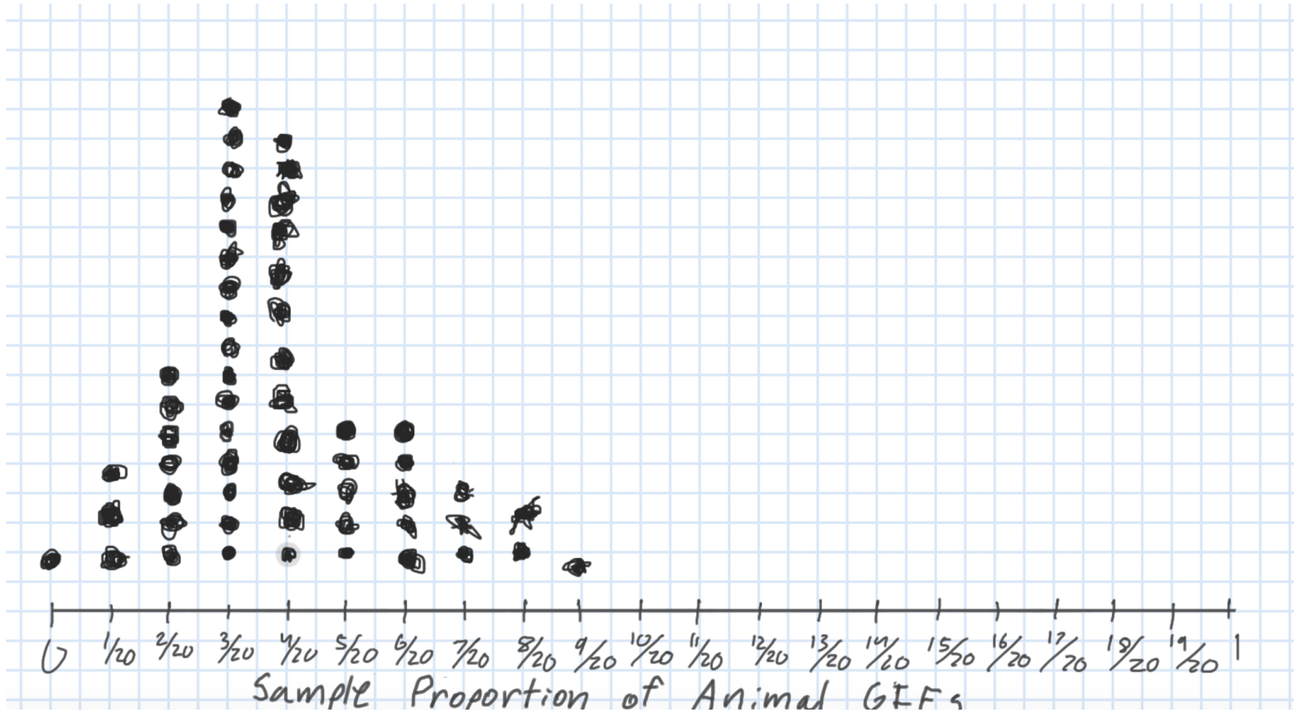
Let's learn about these ideas through an activity! Go to bit.ly/stat100gif.

Sampling Distribution of a Statistic

Steps to Construct an (Approximate) Sampling Distribution:

1. Decide on a sample size, n.
2. Randomly select a sample of size n from the population.
3. Compute the sample statistic.
4. Put the sample back in.
5. Repeat Steps 2 - 4 many (1000+) times.

Sampling Distribution of a Statistic



- Center? Shape?
- Spread?
  - Standard error = standard deviation of the statistic

- What happens to the center/spread/shape as we increase the sample size?
- What happens to the center/spread/shape if the true parameter changes?