Lecture #9: Probability and the MLE

aka STAT109A, AC209A, CSCIE-109A

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Gumb

Lecture Outline: Probability

- Today's Example
- Review
- Probability and Random Variables
  - Normal Distribution
  - Binomial Distribution
- Likelihood
- Statistical Inference

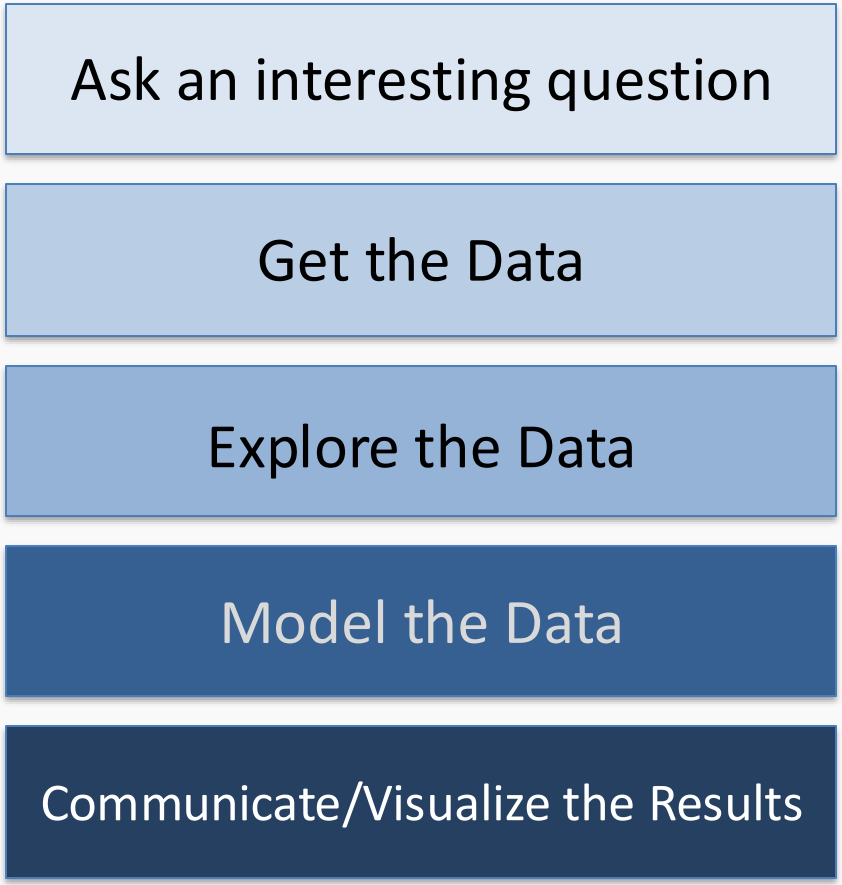
Today's Data Science Example

Recall the data science process:

Today's question: how much are your professors' homes worth?

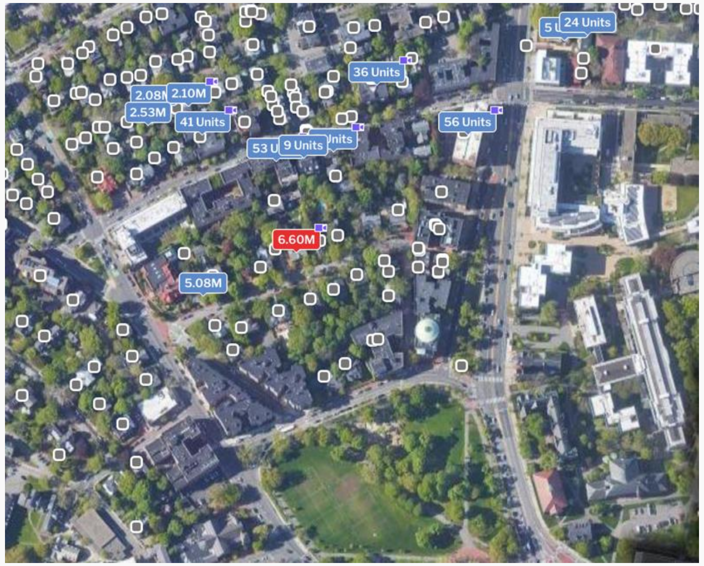
Related question: what variables are associated with selling prices of homes in the Cambridge-Somerville* area?

*Both Pavlos and Kevin live in Somerville



Today's Data Science Example

Today's Data are from Redfin.com:



A place to find historical selling prices of homes in any location (and plotted on a map). Today we're looking at the Cambridge & Somerville area, residential properties, that have sold in the last 12 months.

	Address	Location	Price	Beds	Baths	Sq.Ft.	$/Sq.Ft.	On Redfin	
	20A Lafayette S...	East Arlington	$1,425,000	4	4.5	2,503	$569	—	
	45 Marlboro St	Belmont	$1,590,000	5	4	4,194	$379	—	
	28 Betts Rd	Belmont	$1,250,000	3	3	1,913	$653	—	
	97 Shaw Rd	Shaw Estates	$2,186,000	4	2.5	2,844	$769	—	

- Ask an interesting question
- Get the Data
- Explore the Data
- Model the Data
- Communicate/Visualize the Results

*Note: Kevin did a *little* preprocessing

Today's Data Science Example

- **Response (y):**
  - Selling price (in \$1000s) for $n = 592$ homes around Cambridge and Somerville
- **Predictors (X):**
  - Condo, single family, multi-family, or townhouse
  - Number of bedrooms
  - Number of bathrooms
  - Floor space in square feet
  - Lot size in square feet
  - The year the home was built
  - Distance to Harvard Sq T-stop (in km)

```
print(homes.dtypes)
```

```
date              object
type              object
address           object
city              object
zip                int64
price              int64
beds               int64
baths            float64
neighborhood      object
sqft               int64
lotsize          float64
year             float64
hoa              float64
url               object
mls                int64
latitude         float64
longitude        float64
dist             float64
dtype: object
```

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Today’s Data Science Example

Explore the Data

Notice anything concerning?
What does this indicate? What should we do?

```
homes["zip"]=homes["zip"].astype(str)
homes.describe().round(5)
```

	price	beds	baths	sqft	lotsize	year	hoa	mls	latitude	longitude	dist
count	5.920000e+02	592.00000	592.00000	592.00000	183.00000	591.00000	402.00000	5.920000e+02	592.00000	592.00000	592.00000
mean	1.244197e+06	2.92568	2.05828	1690.04561	3860.93989	1926.58545	352.30100	7.308775e+07	42.38443	-71.11148	2.17682
std	7.810677e+05	1.56242	0.96753	953.69978	1858.56503	46.57422	262.84504	3.623761e+04	0.01220	0.01653	0.88593
min	1.227940e+05	0.00000	1.00000	336.00000	1007.00000	1828.00000	1.00000	7.291816e+07	42.35605	-71.16066	0.44139
25%	8.000000e+05	2.00000	1.00000	1000.00000	2557.50000	1900.00000	200.00000	7.305637e+07	42.37497	-71.12367	1.50989
50%	1.028000e+06	3.00000	2.00000	1407.00000	3580.00000	1910.00000	286.00000	7.309318e+07	42.38459	-71.10973	2.07152
75%	1.450000e+06	3.00000	2.50000	2162.00000	4562.50000	1947.50000	443.50000	7.311370e+07	42.39333	-71.10031	2.82374
max	7.500000e+06	9.00000	7.00000	7530.00000	10454.00000	2024.00000	2984.00000	7.316201e+07	42.41408	-71.07172	4.60435

Today's Data Science Example

Explore the Data

```
homes['lotsize'] = homes['lotsize'].fillna(0)
homes['hoa'] = homes['hoa'].fillna(0)
homes['price'] = homes['price']/1000
```

What did we do? Is this reasonable?

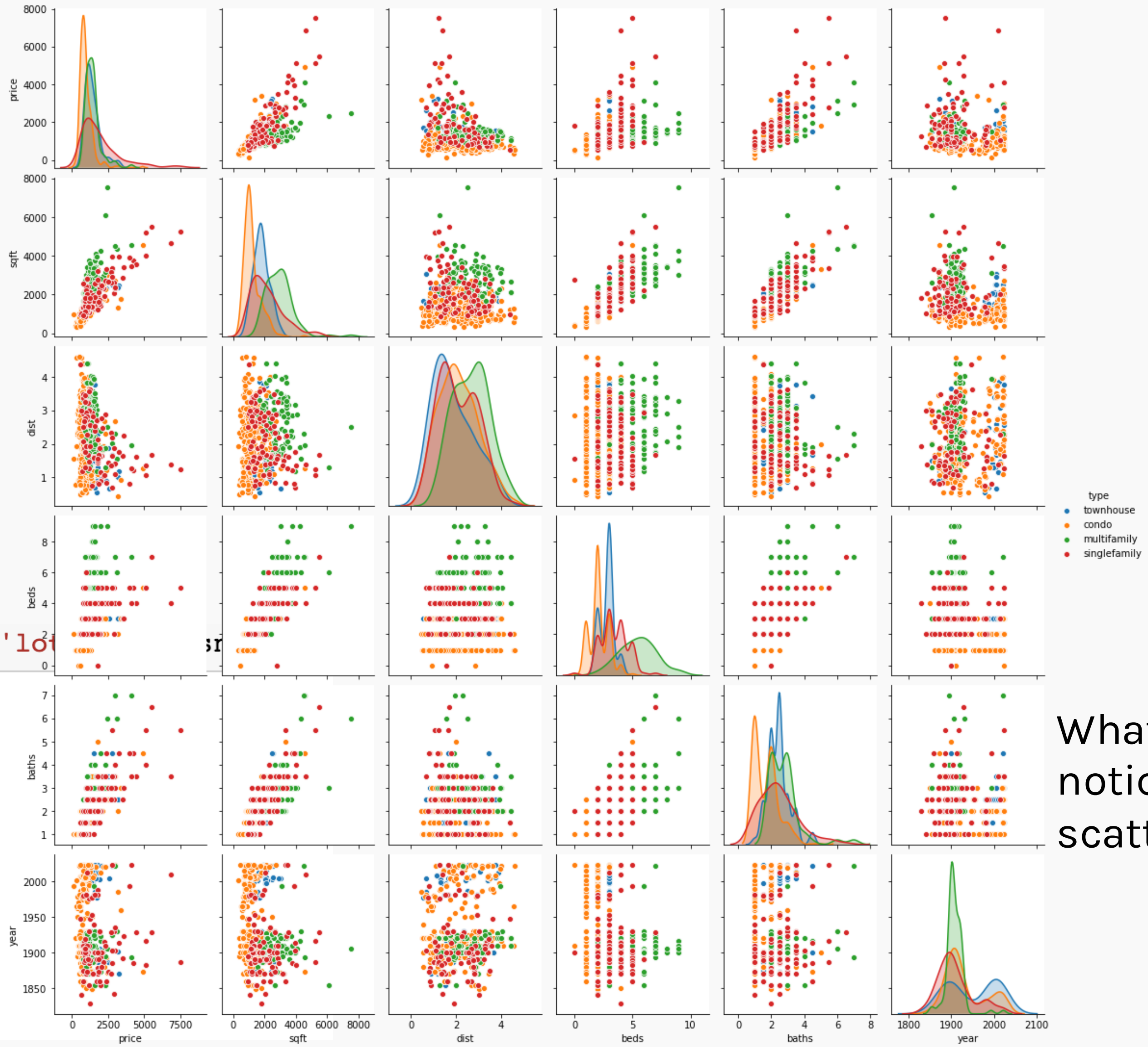
```
pd.crosstab(homes['type'],homes['hoa'].isnull())
```

hoa	False	True
type		
condo	334	7
multifamily	0	92
singlefamily	1	90
townhouse	67	1

```
pd.crosstab(homes['type'],homes['lot
```

lotsize	False	True
type		
condo	0	341
multifamily	92	0
singlefamily	91	0
townhouse	0	68

```
sns.pairplot(homes[['price','sqft', 'type',  'dist' , 'beds' , 'baths' , 'year']], hue="type");
```

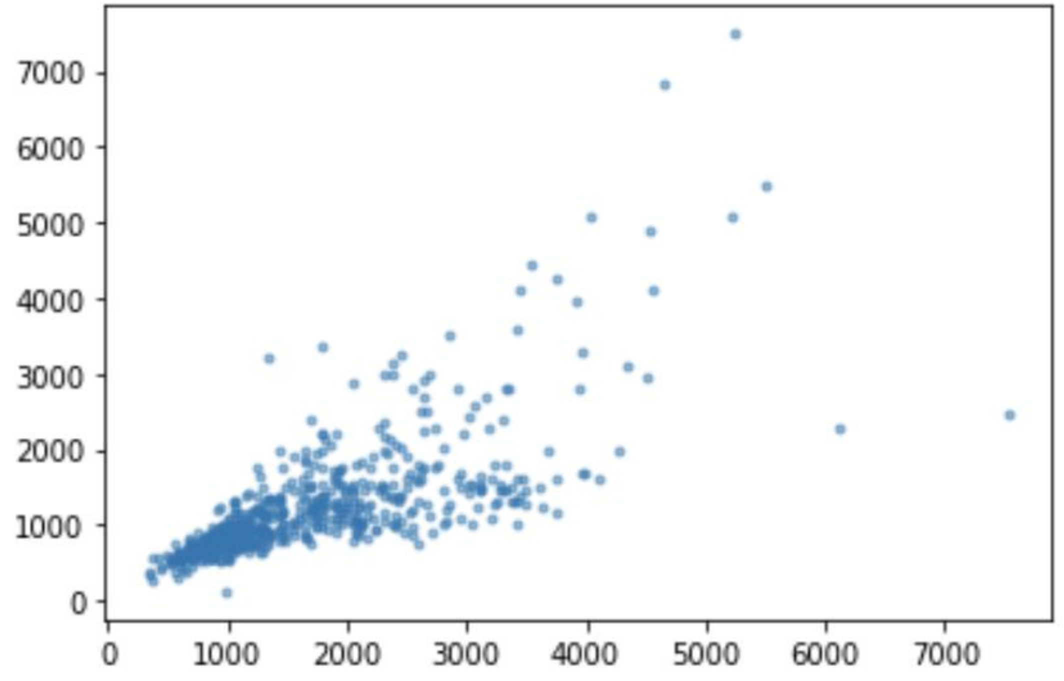


What do you notice in the scatterplots?

Today's Data Science Example

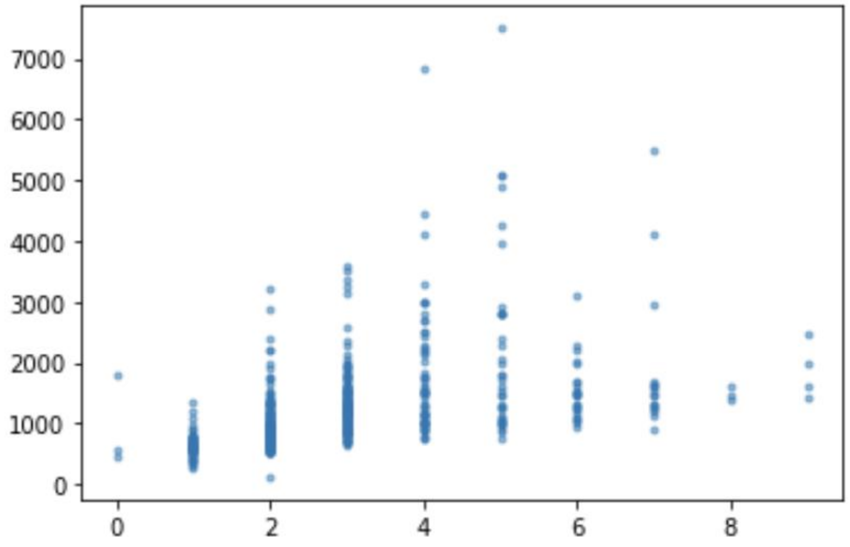
Explore the Data

```
plt.scatter(x = homes['sqft'], y = homes['price'],
            alpha = 0.5, marker = '.');
```

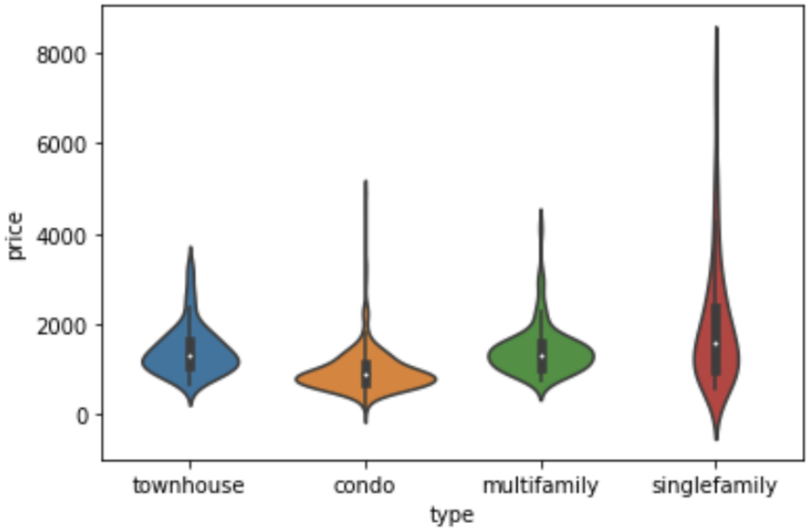


What do you notice? Any causes for concern?

```
plt.scatter(x = homes['beds'], y = homes['price'],
            alpha = 0.5, marker = '.');
```




```
sns.violinplot(x = homes['type'], y = homes['price']);
```



Today's Data Science Example

How should we model the data? Key: **price** (in thousands of $) is the response variable.

Linear Regressions make sense. Why?

Let's ask lots of questions ☺ Like…

- What associations do we care about?
- What predictors/features are likely to be strongly associated with the response?
- What might be nonlinear?
- What interactions are likely to be present?
- What other data considerations should we make?

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Today's Data Science Example

Model the Data

What are the coefficient interpretations?

Notice anything interesting? What does this indicate?

```
type_dummies = pd.get_dummies(homes['type'], drop_first=True)
homes=pd.concat([homes,type_dummies],axis=1)
```

```
X_full = homes[['multifamily', 'singlefamily', 'townhouse',
                'sqft',  'dist', 'beds', 'baths', 'year']]
regress_full = sk.linear_model.LinearRegression().fit(X = X_full, y = homes['price'])

pd.DataFrame({'coef': np.append(regress_full.intercept_,regress_full.coef_)},
             index = np.append("intercept",X_full.columns))
```

	coef
intercept	-1949.067039
multifamily	-452.235208
singlefamily	335.761220
townhouse	-76.437171
sqft	0.641074
dist	-173.542970
beds	-89.934486
baths	198.464604
year	1.229999

```
regress_sqft = sk.linear_model.LinearRegression().fit(X = homes[['sqft']], y = homes['price'])
print("Intercept =",regress_sqft.intercept_.round(2),", Slope =",regress_sqft.coef_[0].round(4))

regress_beds = sk.linear_model.LinearRegression().fit(X = homes[['beds']], y = homes['price'])
print("Intercept =",regress_beds.intercept_.round(2),", Slope =",regress_beds.coef_[0].round(4))
```

```
Intercept = 247.44 , Slope = 0.5898
Intercept = 570.39 , Slope = 230.3084
```

Today's Data Science Example

Model the Data

The library **`statsmodels`** makes our life WAY simpler (my R is showing):

```
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

```
fullmodel_sm = smf.ols(formula = "price ~ sqft + type + dist + beds + baths + year",
                       data = homes).fit()
fullmodel_sm.params
#fullmodel_sm.summary()
```

```
Intercept              -1927.782817
type[T.multifamily]     -455.751842
type[T.singlefamily]     332.773270
type[T.townhouse]        -78.756789
sqft                       0.641839
dist                    -173.274319
beds                     -90.246936
baths                    199.724130
year                       1.218183
dtype: float64
```

Lecture Outline: Probability

- Today's Example
- **Review**
- Probability and Random Variables
  - Normal Distribution
  - Binomial Distribution
- Likelihood
- Statistical Inference

CS109A
GAME Time

When do we use cross-validation?

Options

A. To choose the best k in a k-NN model.

B. To choose the best λ in a Ridge/LASSO model.

C. To choose the best predictors in a linear regression model.

D. To choose between families of models.

When should we standardize predictors?

Options

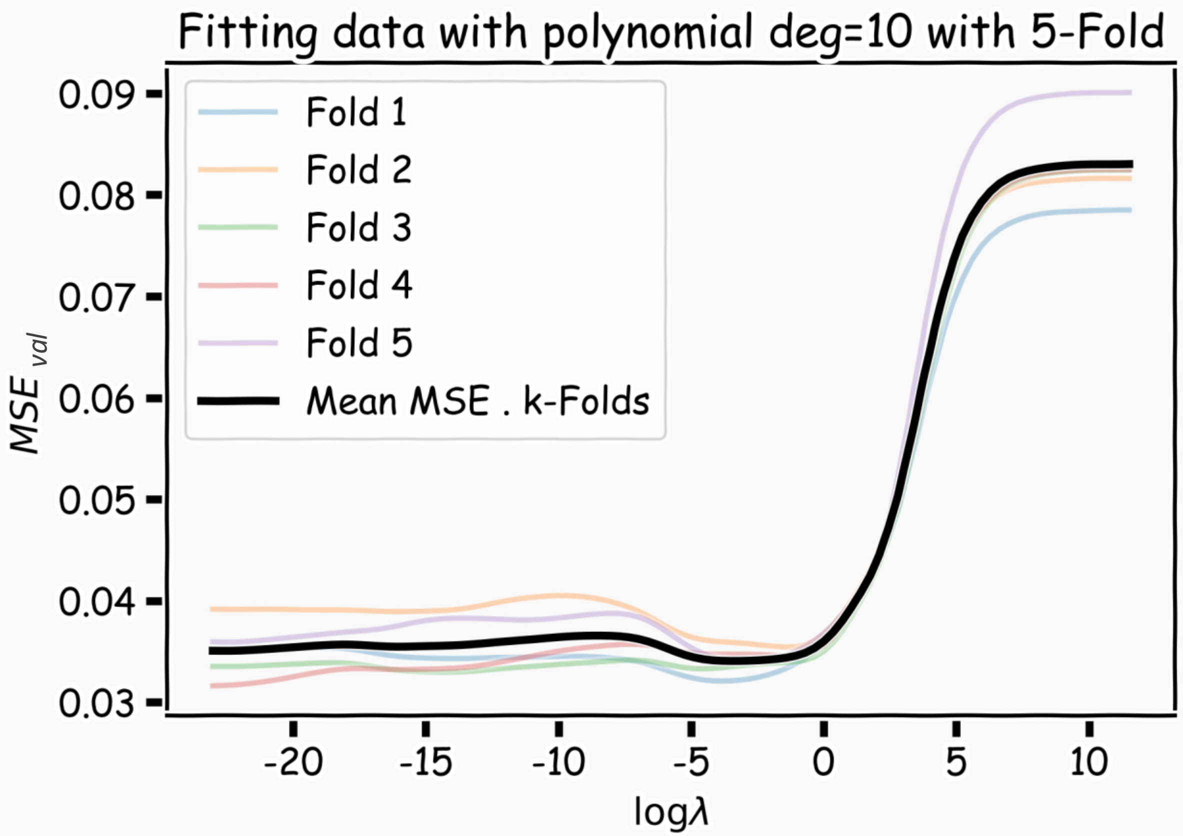
A. Always.

B. Whenever we use k-NN.

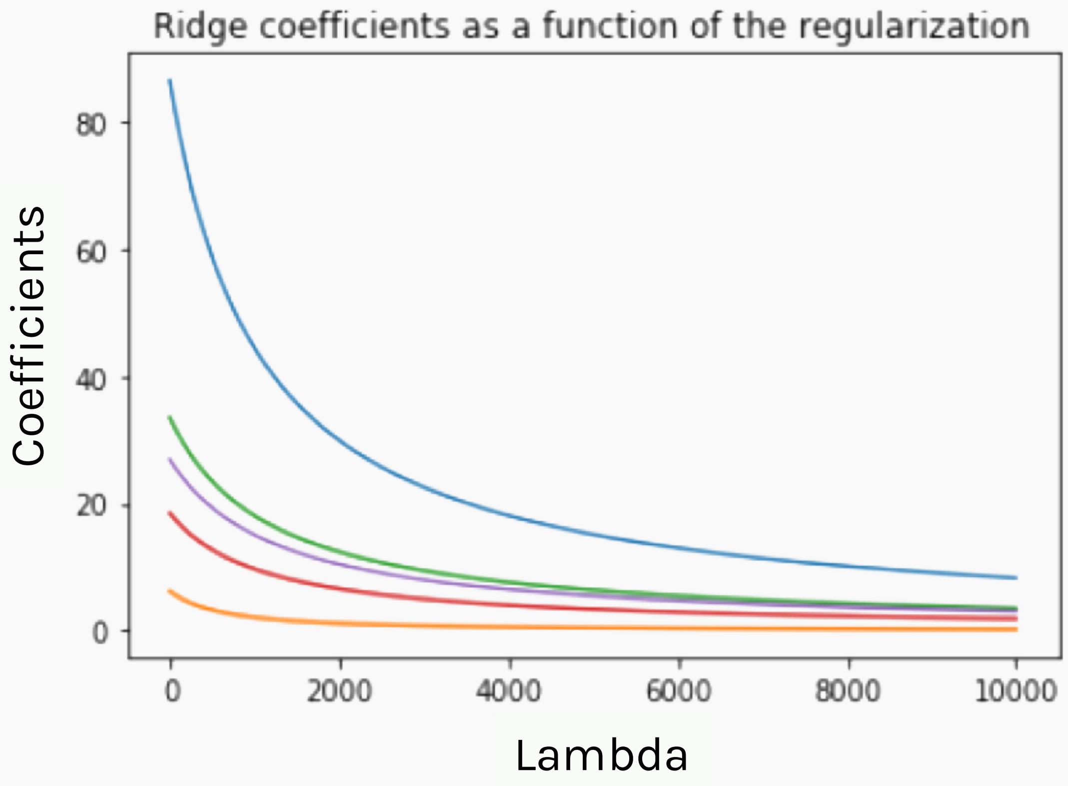
C. Whenever we use a Ridge/LASSO model.

D. Whenever we want to treat the transformed predictors more equally.

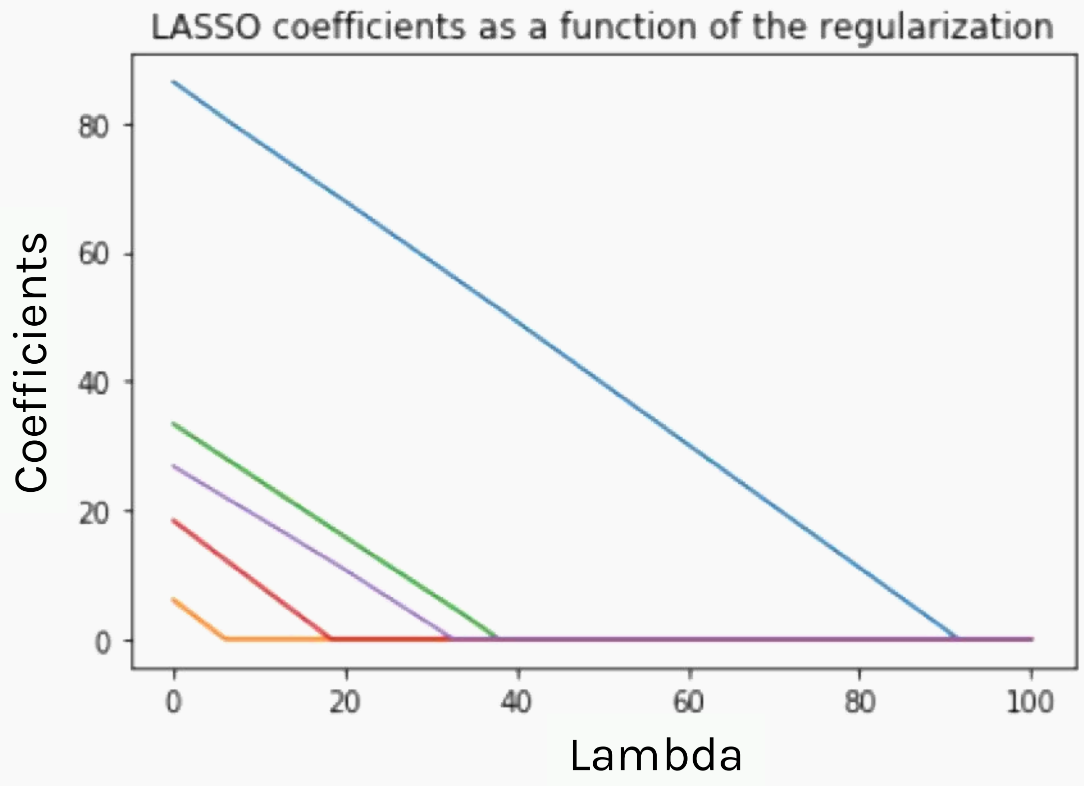
Ridge regularization with **cross-validation** only: step by step



Ridge and LASSO visualized

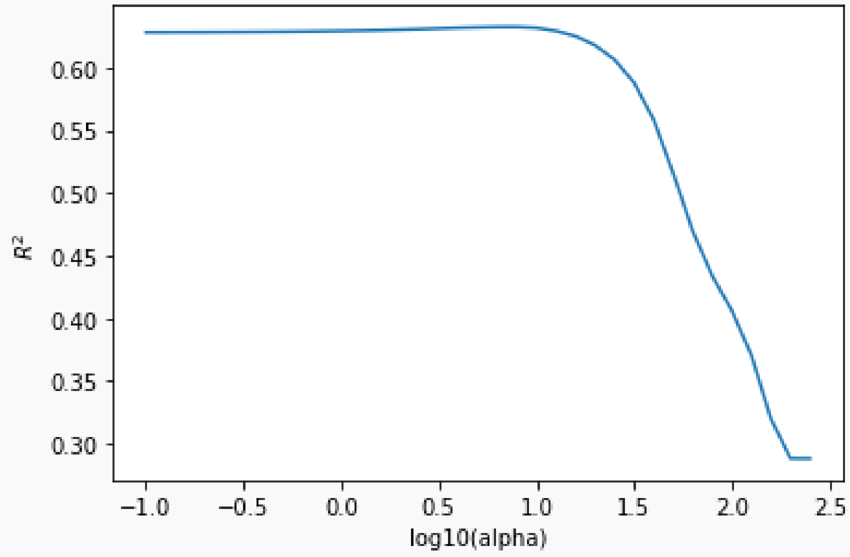


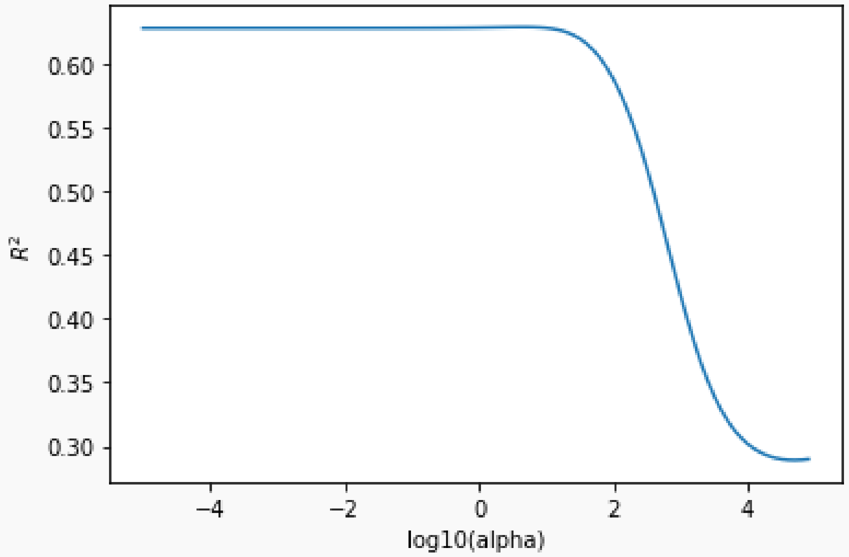
The values of the coefficients decrease as lambda increases, but they are not nullified.

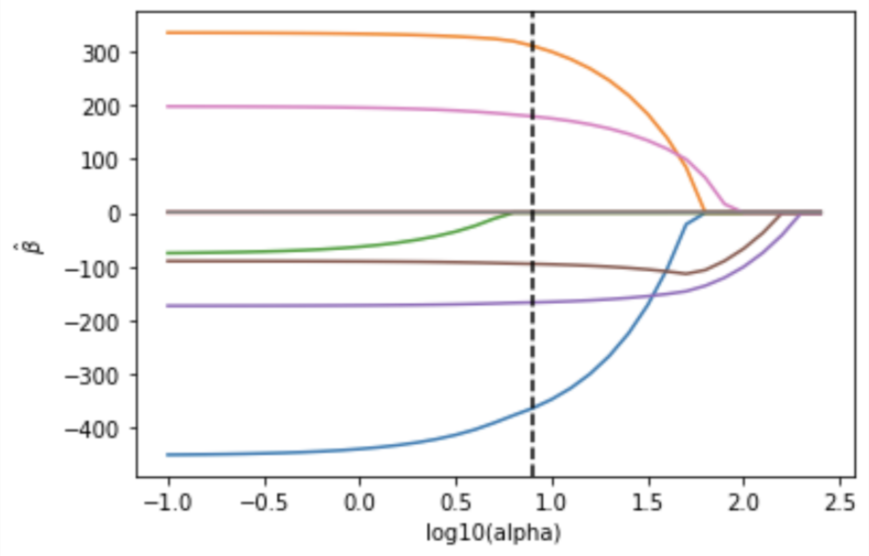


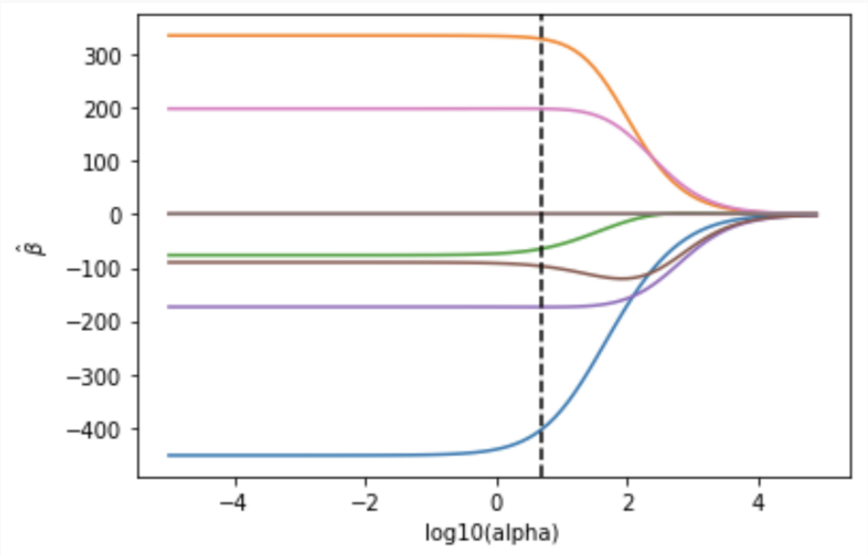
The values of the coefficients decrease as lambda increases and are nullified fast.

Ridge and LASSO: 2 real *trajectory* plots



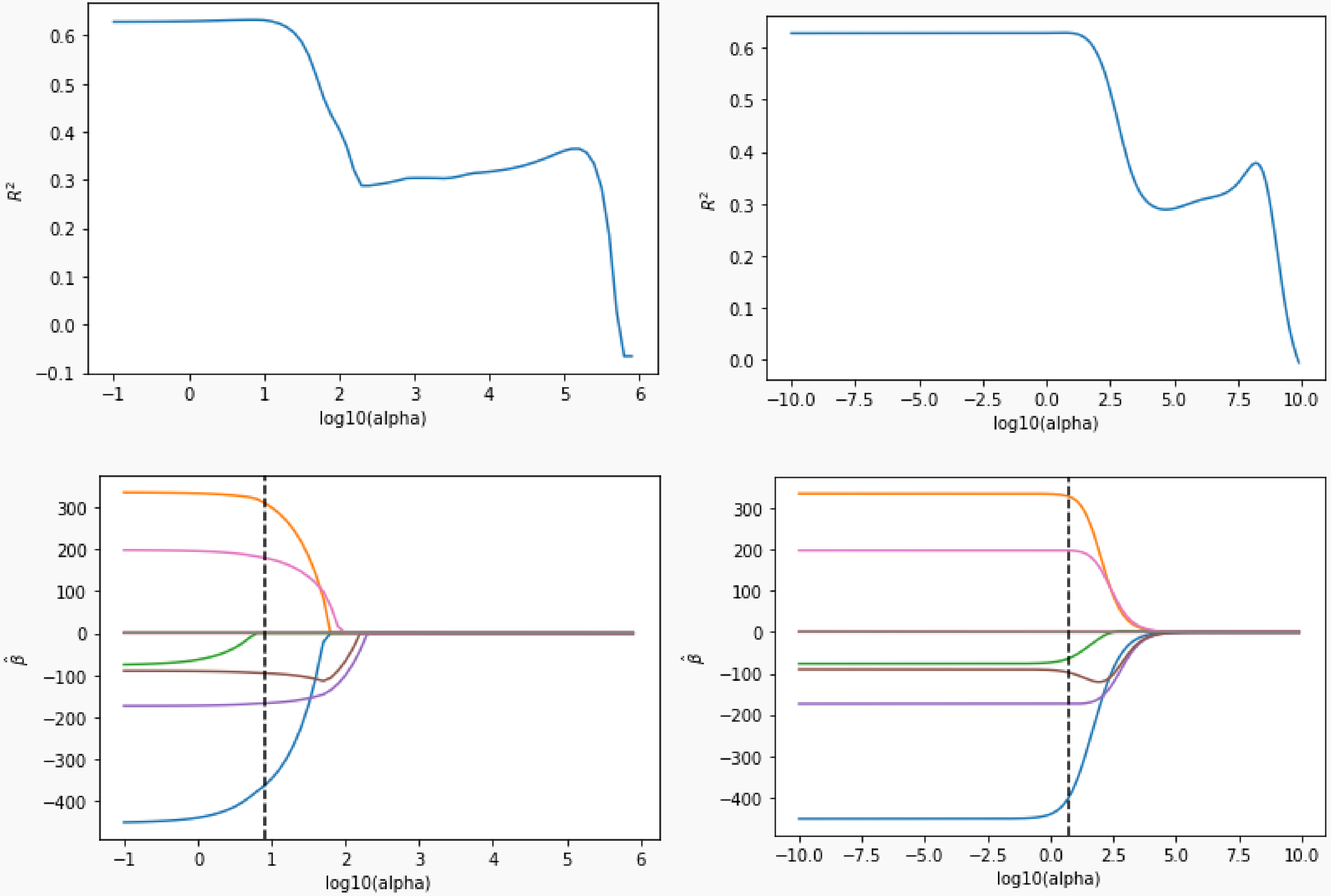






- Which is Ridge vs. Lasso?
- Which is the better predictive model?
- Which variable is least important? 2nd least?
- What is MSE in the OLS model? What about the intercept only model?
- Is there evidence of overfitting? Of multicollinearity?

Ridge and LASSO: 2 real *trajectory* plots



- Which is Ridge vs. Lasso?
- Which is the better predictive model?
- Which variable is least important? 2nd least?
- What is MSE in the OLS model? What about the intercept only model?
- Is there evidence of overfitting? Of multicollinearity?


BELIEVE

Lecture Outline: Probability

- Today's Example
- Review
- **Probability and Random Variables**
    - Normal Distribution
    - Binomial Distribution
- Likelihood
- Statistical Inference

What is probability?

Q: What is probability?

A: A common definition: the long-run, relative frequency* of a random phenomenon/experiment/event.

Q: What values can probabilities take on?

A: Any value between 0 and 1 (including the endpoints).

Q: Why do we care?

A: Because data can be thought of as random realizations of a *data generating process* (whether though sampling or a theoretical construct).

*Note: this ignores the Bayesian definition of probability: a measure of belief.

What is a random variable?

In the context of data, we often describe their potential **numeric** outcomes (before collecting the data) as random variables. That is:

Let's perform a survey of Harvard students and ask the question: do you primarily use a Mac (vs. PC vs. Linux/Ubuntu, etc.)?

Let X_1 be the observed response for the first person we are going to ask. Then X_1 can be thought of as a random variable. ($X_1 = 1$ implies 'Mac', $X_1 = 0$ implies anything else).

*Technically a random variable is a function that takes possible outcomes of random phenomenon (responses of 'Mac', 'PC', etc.) and maps them to numeric values.

What is a probability distribution?

A **probability distribution** is any function (formula, table, or graph) that assigns probabilities (or relative frequencies) to all the possible outcomes of a random variable.

Typically they are written as a formula (called a probability mass function or probability density function, or as its cumulative distribution function).

In our 'Mac' example, we could define the probability distribution as a table:

x	$P(X = x)$
0	$1 - p$
1	p

Which could be summarized as the formula, for $x \in \{0,1\}$:

$$P(X = x) = p^x(1 - p)^{1-x}$$

The goal of our study would be to estimate p.

Discrete vs. Continuous

There are two major types of random variables: **discrete** (can only take on specific values) and **continuous** (can take on any value within a range).

The probability distribution function is defined differently for these two types:

A **probability mass function** (PMF) is a function that gives the probability of getting a specific value for a discrete random variable.

A **probability density function** (PDF) is a function that gives the relative likelihood of a specific value for a continuous random variable (the height of the curve). This is usually written as $f(x)$

*Note: probabilities for a continuous random variable can be represented as areas under the curve, and thus $P(X = x) = 0$ since there is no width.

Joint Distributions

What happens to these probability distributions (PMFs and PDFs) when there are multiple random variables involved (aka, multiple observations in a data set)?

Let $f(x_1, x_2, \dots, x_n)$ be the **joint distribution** of n separate random variables. If they all come from the same generative marginal distribution, $f(x_i)$, and are **independent**, what is the resulting distribution?

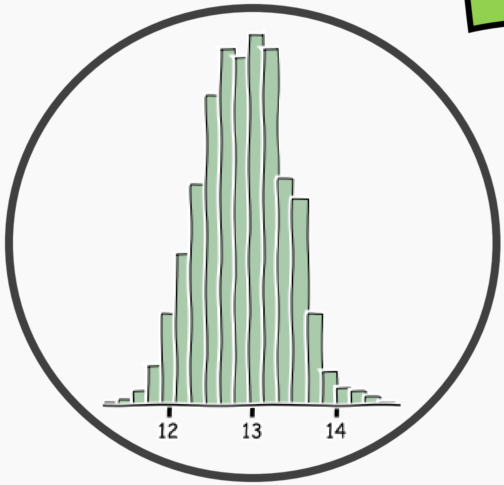
$$f(x_1, x_2, \dots, x_n) = f(x_1) \cdot f(x_2) \cdots f(x_n) = \prod_{i=1}^{n} f(x_i)$$

What does independent data mean, anyway? When does this breakdown?

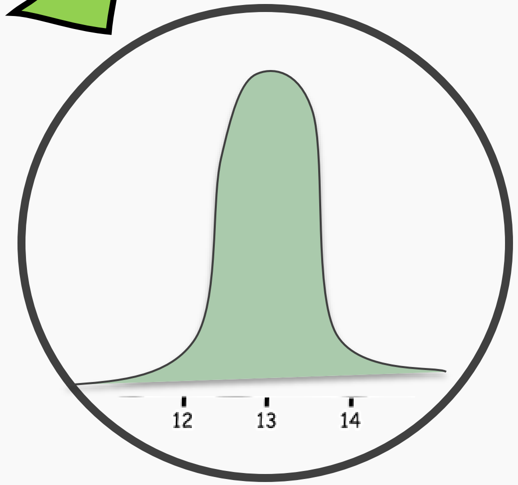
PDF vs PMF

Random Variable

X



Discrete Random Variable



Continuous Random Variable

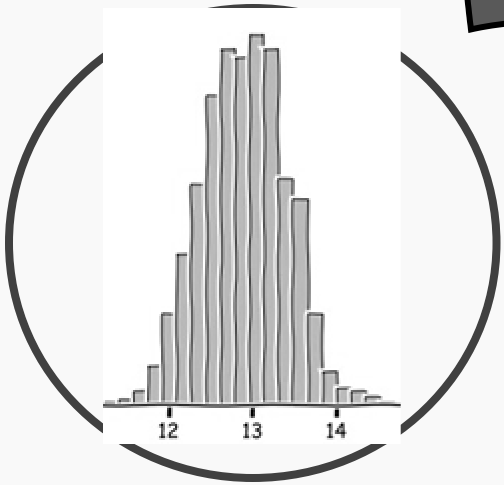
Lecture Outline: Probability

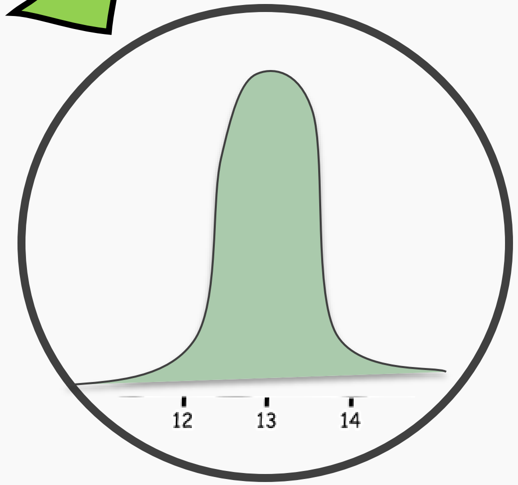
- Today’s Example
- Review
- Probability and Random Variables
    - **Normal Distribution**
    - Binomial Distribution
- Likelihood
- Statistical Inference

PDF vs PMF

Random Variable

X





~~Discrete Random Variable~~

Continuous Random Variable

The Normal Distribution

Let X be a **normally distributed** random variable. Then $X \sim N(\mu, \sigma^2)$, and X has probability density function (PDF):

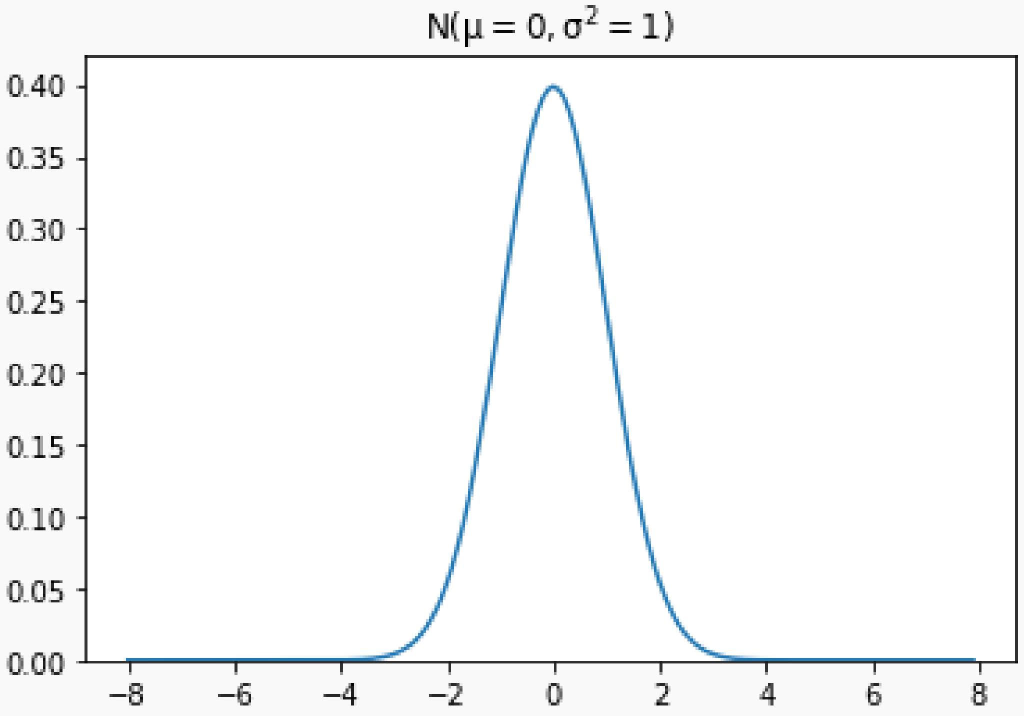
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

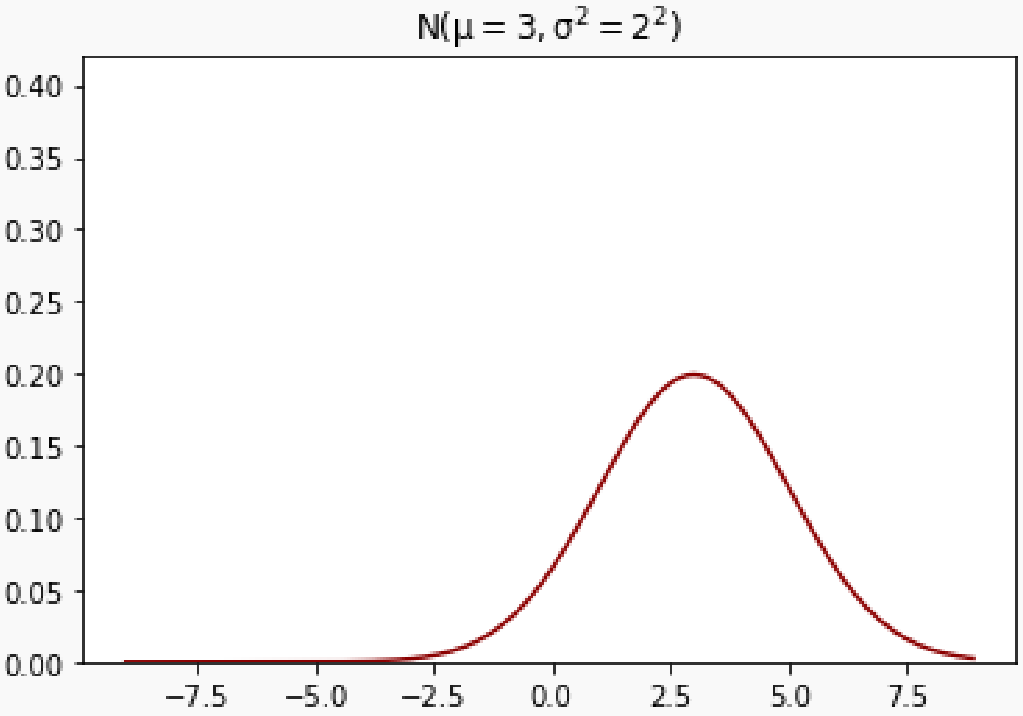
The normal distribution (sometimes called the Gaussian) is often referred to as the bell-shaped curve. But the normal distribution isn't the only one that is bell-shaped: t distributions are also bell-shaped, for example.

The standard normal distribution is a special case: $Z \sim N(0,1)$.

Any normal random variable can be standardized using the formula $Z = \frac{X-\mu}{\sigma}$.

The Normal Distribution Examples





A normal distribution has mean μ and standard deviation σ.

Central Limit Theorem

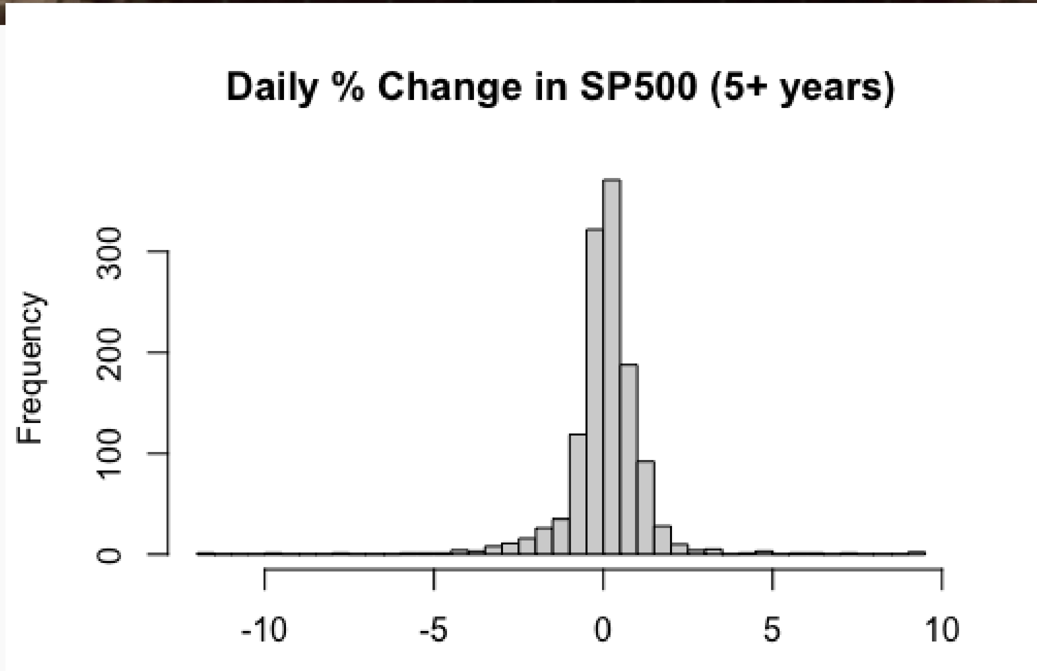
Why is the normal distribution used so often?

The **Central Limit Theorem**: random variables that are averages or sums of many other random variables will be approximately normally distributed.

More specifically: if $X_1, X_2, \ldots, X_n$ are independent random variables (representing individual observations of data) with mean μ and standard deviation σ (not necessarily normal themselves), then the sample mean $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ will have approximate distribution:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

I See Normal Distributions



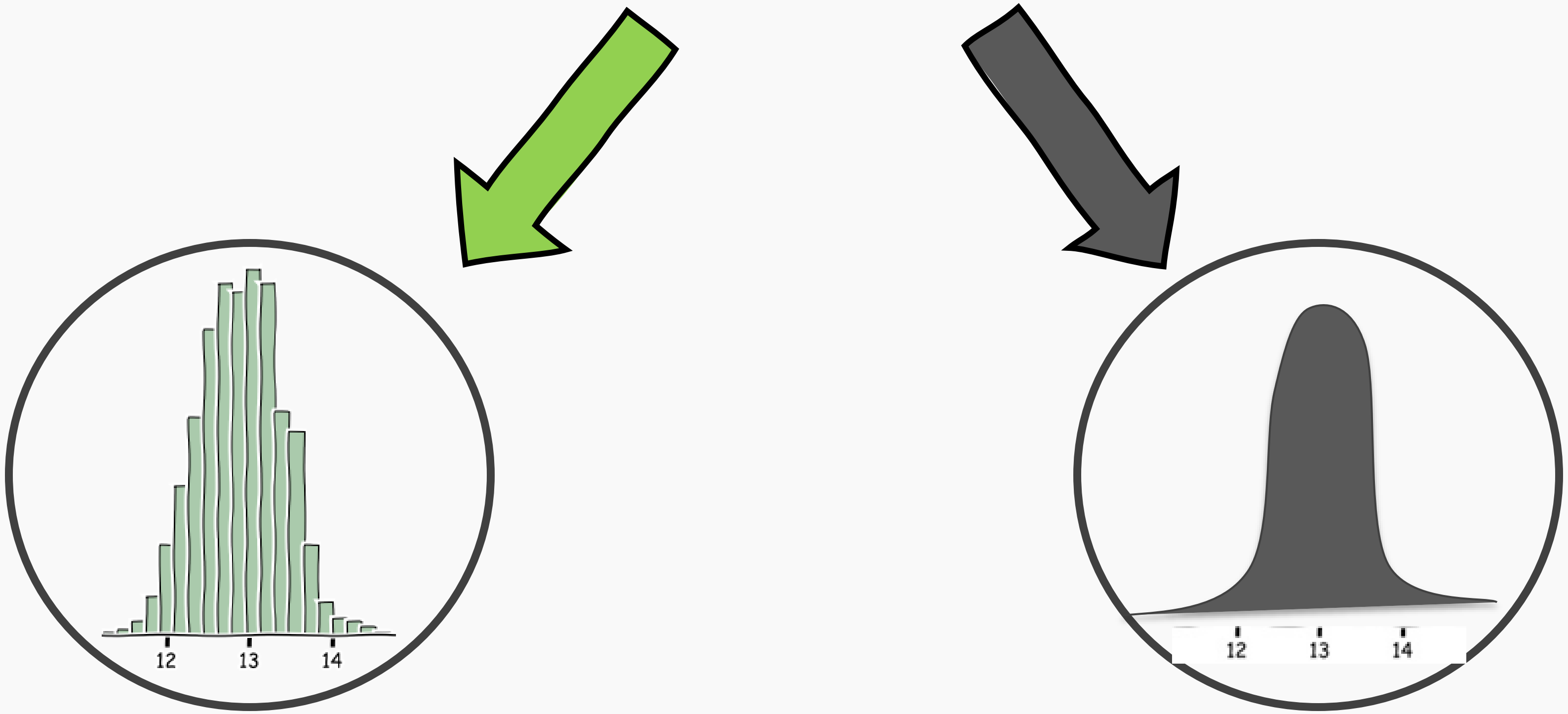
Lecture Outline: Probability

- Today's Example
- Review
- Probability and Random Variables
    - Normal Distribution
    - **Binomial Distribution**
- Likelihood
- Statistical Inference

PDF vs PMF

Random Variable

X



Discrete Random Variable

Continuous Random Variable

The Binomial Distribution

Let X be a random variable that counts the number of successes in a fixed number of independent trials (n) with fixed probability of success (p) in each trial. Then X is said to have a **binomial distribution**. This is often written as: $X \sim Binom(n, p)$, and X has probability mass function (PMF):

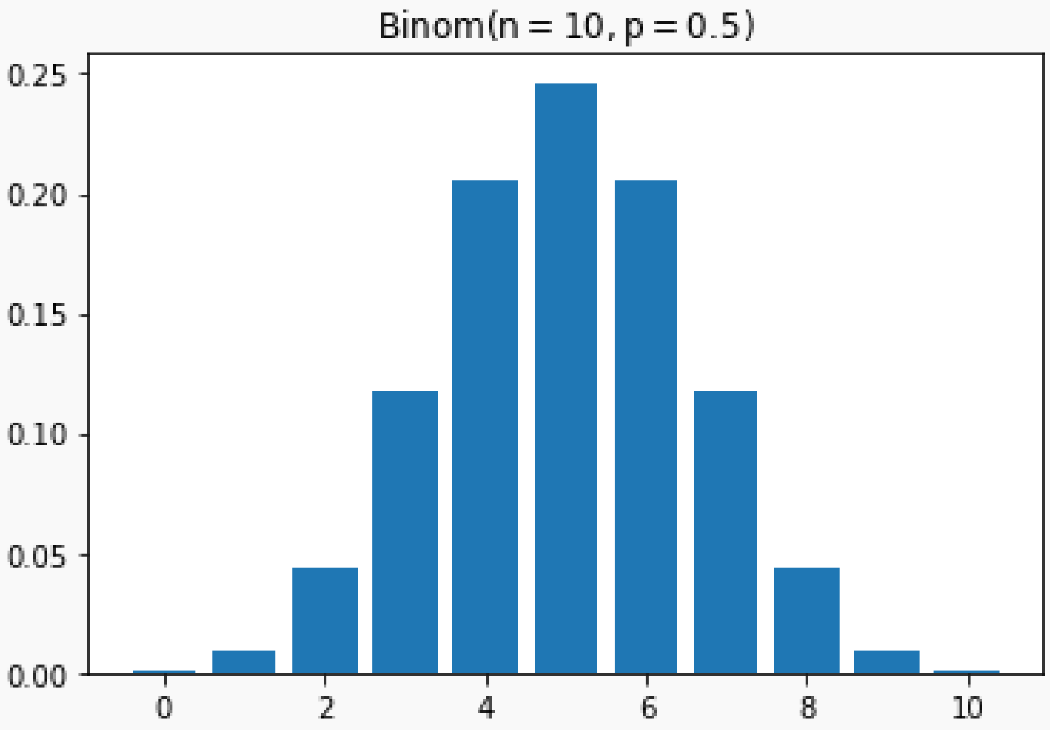
$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

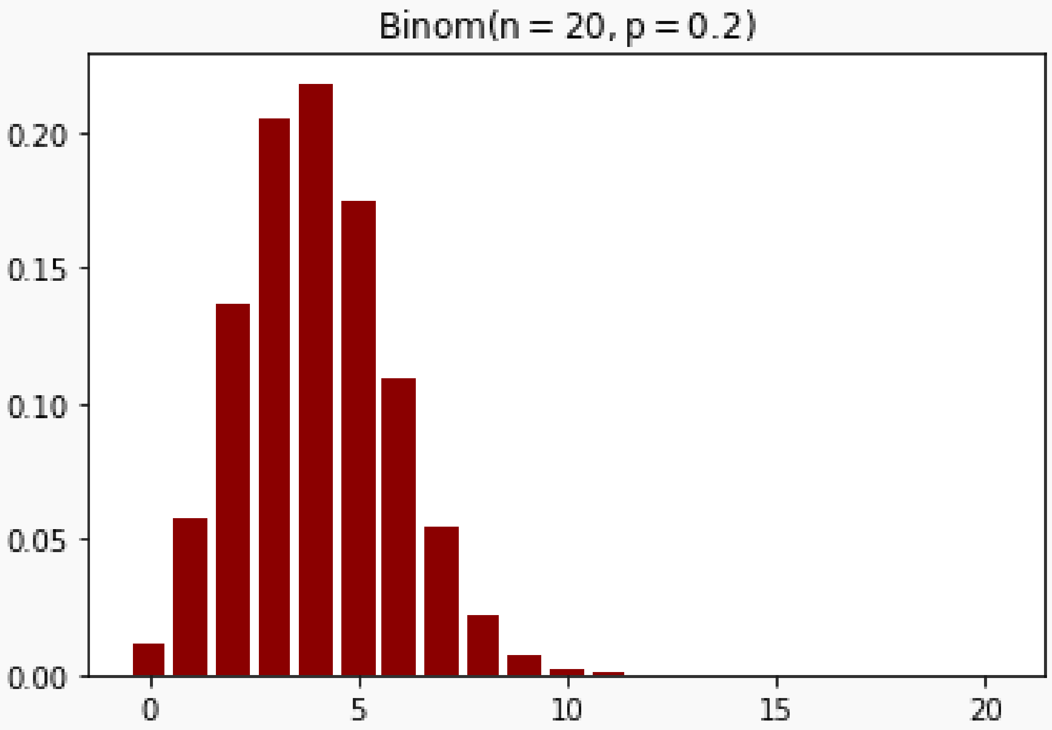
Think counting the number of heads when flipping a biased coin n times.

The binomial distribution is useful to describe polling data (proportion of people who will vote for Biden), survey data (will you take 109B in the Spring), or any data that are binary!

The **Bernoulli distribution** is a special case when $n = 1$. This is the distribution that describes our `Mac` example.

Binomial Distribution Examples





A binomial distribution has mean np and standard deviation $\sqrt{np(1-p)}$.

Lecture Outline: Probability

- Today's Example
- Review
- Probability and Random Variables
  - Normal Distribution
  - Binomial Distribution
- **Likelihood: a roadmap for statistical inference**
- Statistical Inference

The Probability of Data

In a typical probability problem (like in Stat 104 or 110), you would be told something like “20% of Harvard College students are collegiate athletes. What is the probability that there are 50 athletes in a random sample of 200 students from Harvard College?”

$$P(X = 50) = \binom{200}{50}(0.20)^{50}(0.80)^{150} = 0.0149$$

$$P(X \geq 50) = \sum_{50}^{200}\binom{200}{x}(0.20)^{x}(0.80)^{200-x} = 0.0494$$

<u>An alternative question</u>: what is more likely to occur: 50 athletes or 40 athletes in a sample of 200 students? How can we make the determination?

Inference: the inverse of probability

In the last problem, how did we know that the statement "20% of Harvard College students are collegiate athletes" is accurate? Where did this come from?

In most applications, the true population parameter (here, the proportion in all of Harvard College) is unknown. What we get to observe is the data, and we want to make a statement about the unknown parameter. So a more poignant question would be:

"There are 50 athletes in a random sample of 200 students from Harvard College. Is a binomial distribution with $p = 0.2$ or $p = 0.25$ more reasonable?"

This approach of using the data to make a statement about a parameter (in a statistical model) is called **inference**.

The idea of likelihood

The **likelihood** approach to inference is based on exactly what was presented in the last slide: given observed values of data (summarized by specific sample statistics), what values of the model's parameters are likely?

It simply just flips a PDF or PMF on its head: instead of writing this function with the data (X) as the unknown, it uses the same function but uses the parameter(s) as the unknown(s). The **likelihood function**, $\mathcal{L}$, measures how well a model (and its set of parameters) describes the observed data.

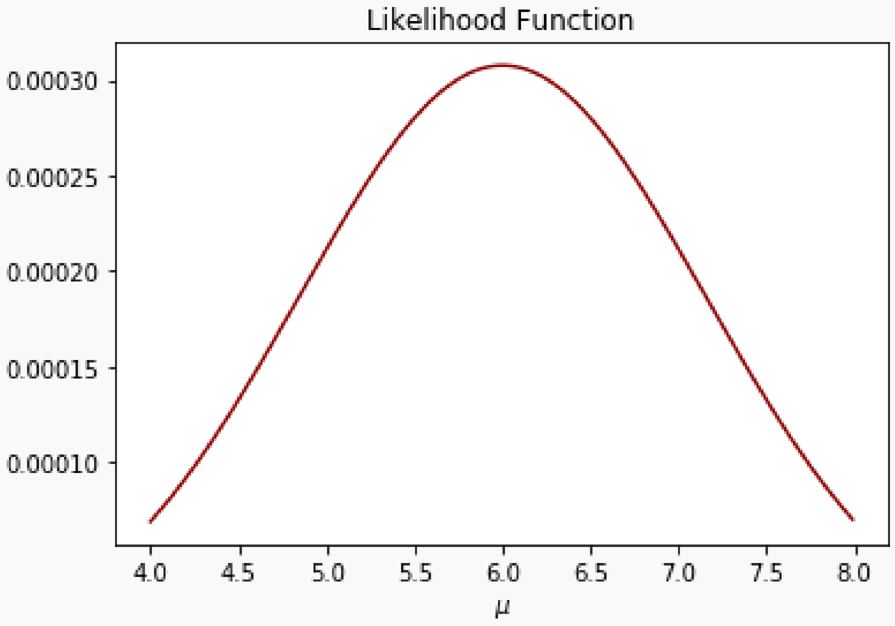
For a set of independent and normally distributed random variables, $X_i \sim N(\mu, \sigma^2)$:

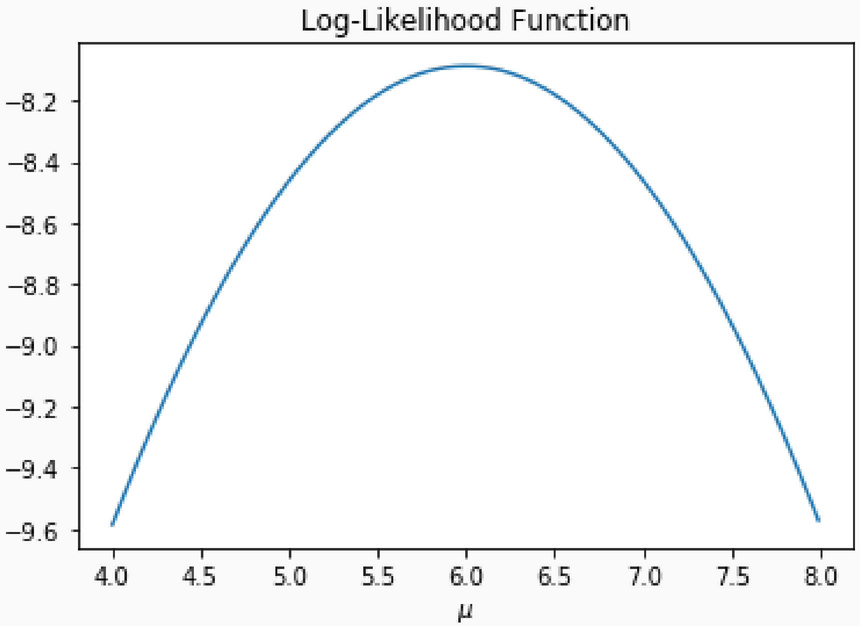
$$\mathcal{L}(\mu, \sigma^2 | x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}$$

Likelihood function example

3 observations are collected [3, 5, 10] that are thought to come from a normal distribution with unknown mean, μ, but is known to have a variance of $\sigma^2 = 2^2$ (yes, this is **very** contrived).

Let's plot the likelihood and log-likelihood functions:





Maximizing the likelihood

In order to choose the best Normal distribution to describe a set of data, we should maximize the likelihood that chooses the best set of parameters given the data.

The **maximum likelihood estimates** for a statistical model are those that maximize the likelihood function given the observed data.

How do we do this mathematically? How could we do this computationally?

With Math: Take [partial] derivatives w.r.t. the unknown parameters (called the score equations), set to zero, and solve!

With Computers: Gradient descent! (of the negative log-likelihood)

The Simple Linear Regression Model

We've defined the linear regression model to predict the *i*-th observation's response, Y_i, from a predictor, X_i, to be:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

For any random variable, ϵ, that has zero mean, then:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

The error term, ϵ_i, represents the distance the observation lies from the line in the vertical distance (direction of *Y*).

The Probabilistic Regression Model

If we assume that $\epsilon_i \sim N(0, \sigma^2)$

This regression model can be rewritten as:

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

The likelihood of a measurement having value Y_i given X_i for a model β_0, β_1:

$$L(\beta_0, \beta_1, \sigma^2 | Y_i, X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{Y_i - (\beta_0 + \beta_1 X_i)}{\sigma}\right)^2}$$

The Probabilistic Regression Model

The likelihood of a measurement having value Y_i given X_i for a model β_0, β_1

$$L(\beta_0, \beta_1, \sigma^2 | Y_i, X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{Y_i - (\beta_0 + \beta_1 X_i)}{\sigma}\right)^2}$$

This formulation allows us to write out the **joint** likelihood function for this probability model.

The joint likelihood function for this probability model becomes:

$$L(\beta_0, \beta_1, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{Y_i - (\beta_0 + \beta_1 X_i)}{\sigma}\right)^2}$$

The Likelihood of Linear Regression

The joint likelihood function for this probability model becomes:

$$L(\beta_0, \beta_1, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{Y_i - (\beta_0 + \beta_1 X_i)}{\sigma}\right)^2}$$

Which leads to the log-likelihood:

$$l(\beta_0, \beta_1, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = \ln(L(\beta_0, \beta_1, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X})) = -\sum_{i=1}^{n} \ln\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2$$

What should we do with this log-likelihood?

?

What does this function look eerily similar to? What does maximizing this function lead to with regards to the best estimates of β_0, β_1?

The Likelihood of Linear Regression

Instead of maximizing the log-likelihood we can minimize the
negative-log-likelihood:

$$-l(\beta_0, \beta_1, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{n} \ln\left(\sqrt{2\pi\sigma^2}\right) + \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2$$

Which is equivalent to minimizing

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2$$

A look ahead: likelihood for binary outcomes

Let $Y \sim Bern(p)$. What is the joint likelihood function? What is the log-likelihood?

Likelihood:

$$L(p|\,\boldsymbol{Y}) = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i}$$

log-likelihood:

$$l(p|\,\boldsymbol{Y}) = \sum_{i=1}^{n} Y_i \ln(p) + \sum_{i=1}^{n} (1-Y_i)\ln(1-p)$$

Why do we care?

What if we let $p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots)}}$? This is logistic regression!

Lecture Outline: Probability

- Today's Example
- Review
- Probability and Random Variables
  - Normal Distribution
  - Binomial Distribution
- Likelihood: a roadmap for statistical inference
- **Statistical Inference**

Inference: connecting estimates to the bigger picture

The estimated model to predict **price** from **sqft** only was:

$$\hat{y}_i = 247.44 + 0.5898x_i$$

Review from earlier today: what is the underlying theoretical model for this simple linear regression (aka, the population model)?

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

What's the difference between the two? What's the connection?

The estimates from the data ($\hat{\beta}_0 = 247.44$ and $\hat{\beta}_1 = 0.5898$) are just one guess (based on a single sample of 592 homes) of what the line would be if all homes in the Cambridge/Somerville were sold.

Beyond Point Estimates

$$\hat{y}_i = 247.44 + 0.5898x_i$$

OK, those point estimates of the parameters are great, but how accurate is $\hat{\beta}_1 = 0.5898$? Is a true $\beta_1 = 0.60$ reasonable? How about 0.70? How about 0?

In order to assess these questions, we need to get a sense of the variability of our estimate(s)…they won't be 100% on target. That way we can build a range of plausible values of the true β_1 around our estimate $\hat{\beta}_1$. This is called a…….

Confidence Interval

There are many ways to build a confidence interval. We will see the 2nd of two options in today's class (the two most common approaches):

1. Using Bootstrap resamples
2. **Using formulas based on probability theory**

Oh, and one more thing…

In simple regression, the estimates are calculated to be:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = r \cdot \frac{s_y}{s_x}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

In multiple regression, the estimates are calculated to be:

$$\hat{\vec{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\boldsymbol{X}^T\vec{y}$$

What other parameter have we ignored? This will be useful going forward.

The estimate of the residual variance is (p is the number of predictors):

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - (p+1)}$$

It is just a *corrected* version of MSE (based on the # of β coef's used in the model).

Confidence intervals for the predictors' estimates: **Standard Errors**

We can empirically estimate the standard deviations $\hat{\sigma}_{\hat{\beta}}$ which are called the **standard errors,** $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ through bootstrapping.

Alternatively:

If we know the variance σ_ϵ^2 of the noise ϵ, we can compute $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ analytically using the formulae below (no need to bootstrap):

$$SE\left(\hat{\beta}_0\right) = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$$

$$SE\left(\hat{\beta}_1\right) = \frac{\sigma_\epsilon}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Where n is the number of observations

$\bar{x}$ is the mean value of the predictor.

Standard Errors

In practice, we do not know the value of σ_ϵ since we do not know the exact distribution of the noise ϵ.

We can empirically estimate σ_ϵ, from the data and our regression line:

$$\hat{\sigma}_\epsilon = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - (p+1)}} = \sqrt{\frac{n \cdot MSE}{n - p - 1}}$$

Standard Errors based on probability theory

$$\widehat{SE}(\hat{\beta}_0) = \hat{\sigma}_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$$

More data: $n \uparrow$ and $\sum_i (x_i - \bar{x})^2 \uparrow \Longrightarrow SE \downarrow$

Wider coverage: Var (x), aka $\sum_i (x_i - \bar{x})^2 \uparrow \Longrightarrow SE \downarrow$

More "precise" data: $\sigma_\epsilon^2 \downarrow \Rightarrow SE \downarrow$

$$\widehat{SE}\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}_\epsilon}{\sqrt{\sum_i (x_i - \bar{x})^2}} = \frac{\hat{\sigma}_\epsilon}{\sqrt{n \cdot s_x^2}}$$

Better model: $(y_i - \hat{f}) \downarrow \Longrightarrow \hat{\sigma}_\epsilon \downarrow \Longrightarrow SE \downarrow$

$$\hat{\sigma}_\varepsilon = \sqrt{\sum \frac{\left(y_i - \hat{f}(x)\right)^2}{n - p - 1}}$$

Question: What happens to the $\widehat{\beta_0}$, $\widehat{\beta_1}$ under these scenarios?

Standard Errors in Multiple Regression

In multiple regression, the standard error formulas are a bit more complicated. Recall the linear algebra version of the estimates:

$$\hat{\vec{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\boldsymbol{X}^T\vec{y}$$

What is $\mathrm{Var}\left(\hat{\vec{\beta}}\right)$? What are its dimensions?

$$\widehat{\mathrm{Var}}\left(\hat{\vec{\beta}}\right) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}_\varepsilon^2$$

The standard errors are the diagonal elements of this resulting covariance matrix.

*Note: it takes a little bit of matrix algebra to derive this result.

Confidence Intervals (formula based)

A 95% confidence interval for the true *slope* (β_j) in a linear regression model can then be calculated based on these formulas:

$$\hat{\beta}_1 \pm t^* \cdot \widehat{SE}\left(\hat{\beta}_1\right)$$

where t^* is the *critical value* (aka, quantile) from a t-distribution with $df = n - (p + 1)$ that puts 2.5% probability in each tail.

Note: $t^* \approx 2$ (if n is very, very large, this becomes $z^* = 1.96$)

Hypothesis Testing

Hypothesis testing is a formal process through which we evaluate the validity of a statistical hypothesis by considering evidence **for** or against the hypothesis gathered by **random sampling** of the data.

1. State the hypotheses, typically a **null hypothesis**, H_0 and an **alternative hypothesis**, H_A, that is the negation of the former.
2. Choose a type of analysis, i.e. how to use sample data to evaluate the null hypothesis. Typically, this involves choosing a single test statistic.
3. **Sample** data and compute the test statistic.
4. Use the value of the test statistic (or the p-value) to either reject or not reject the null hypothesis.
5. Restate the conclusion in context of the problem.

Hypothesis Testing

1. State Hypothesis:

Null hypothesis:

H_0: There is no relation between X_j and Y in the model ($\beta_j = 0$).

The alternative:

H_A: There is some relation between X_j and Y in the model ($\beta_j \neq 0$).

2. Choose test statistic

$$t\text{–}test = \frac{\hat{\beta}_1}{\widehat{SE}\left(\hat{\beta}_1\right)}$$

Hypothesis Testing

3. Sample:

Using probability theory (or permutations) we can estimate $\hat{\beta}_1$, its standard error, and the $t - test$ statistic.

4. **Reject or not reject the hypothesis:**

We compute *p-value* , the probability of observing any value equal to $|t|$ or larger, from random data.

If p-value < p-value-threshold (α) we reject the null.

5. **Restate the conclusion in context of the problem:**

What is the direction of the relationship? What is the magnitude? Is the relationship surprising? Are there any possible confounders?

P-value

Why the absolute value?

?

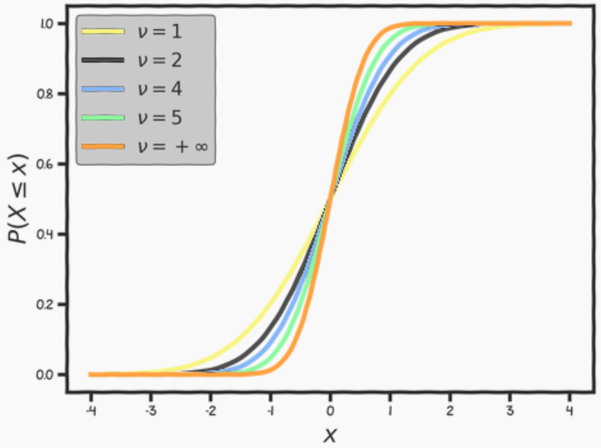
To compare the t-test values of the predictors from our model, $|t - test|$, with the t-tests calculated using permuted data, $|t^R|$, we estimate the probability of observing $|t^R| \geq |t - test|$.

We call this probability the p-value:

$$p - value = P(|t^R| \geq |t - test|)$$

Small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance.
It is common to use $p\text{-value} < 0.05$ as the threshold for significance.

To calculate the p-value we use the cumulative distribution function (CDF) of the student-t. `stats model` a python library has a build-in function `stats.t.cdf()` which can be used to calculate this.



Permutation Tests: a side note

Should you use a bootstrap approach to perform a hypothesis test?

While this is tempting, this is **not advisable**. Why?

It is a technical issue: the bootstrap approach is prone to inflating Type I error: you conclude there is an association when there really is not one.

In order to preserve the state Type I error (presumably at 5%), you should instead perform a permutation test: another resampling method.

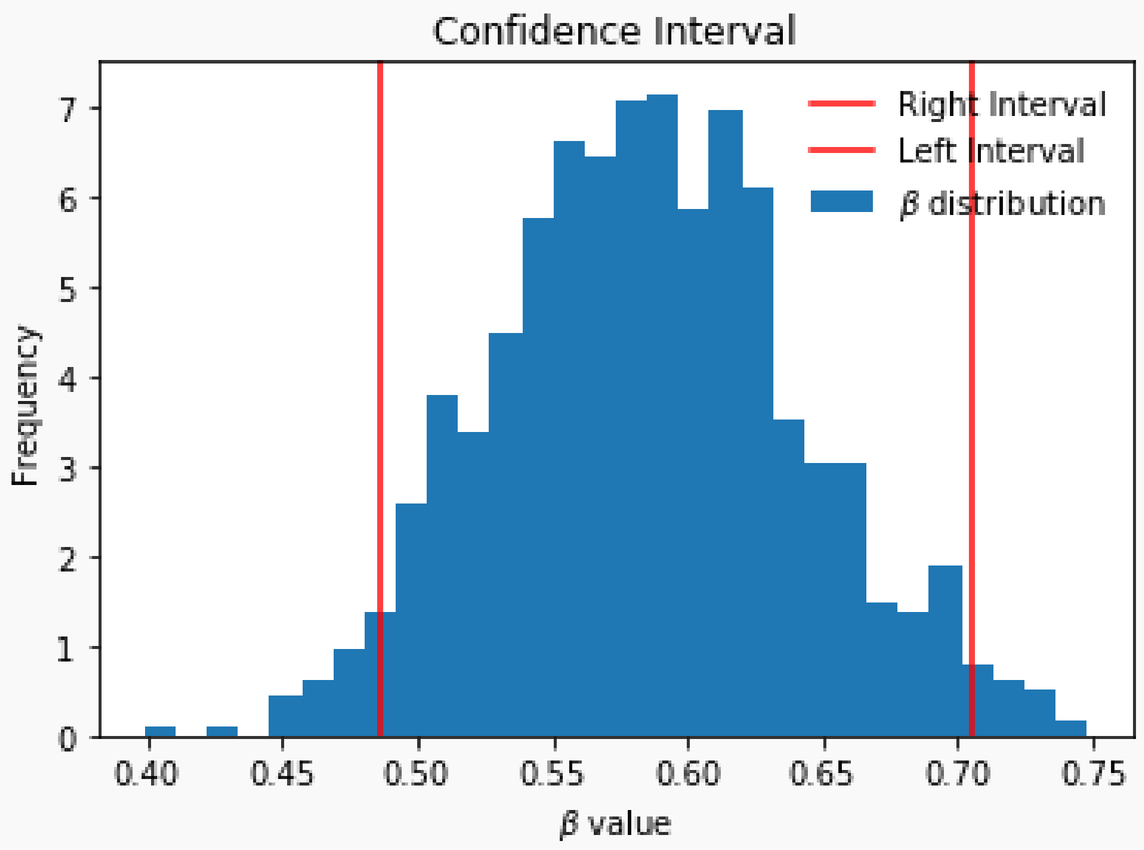
In a permutation test, you resample the data assuming the null hypothesis is true. This can most easily done by shuffling the response variable while keep the columns of the predictors as-is.

Inference via statsmodels vs. bootstrapping

```python
beta1_CI = (np.percentile(beta1_list,2.5),np.percentile(beta1_list,97.5))

print(f'The beta1 confidence interval is {round(beta1_CI[0],3),round(beta1_CI[1],3)}')
```

```
The beta1 confidence interval is (0.487, 0.705)
```




```python
sqftmodel_sm = smf.ols(formula = "price ~ sqft",
                       data = homes).fit()
sqftmodel_sm.summary()
```

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.519
Model:	OLS	**Adj. R-squared:**	0.518
Method:	Least Squares	**F-statistic:**	635.6
Date:	Tue, 03 Oct 2023	**Prob (F-statistic):**	9.97e-96
Time:	22:00:05	**Log-Likelihood:**	-4566.2
No. Observations:	592	**AIC:**	9136.
Df Residuals:	590	**BIC:**	9145.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	247.4382	45.388	5.452	0.000	158.296	336.581
sqft	0.5898	0.023	25.211	0.000	0.544	0.636

Omnibus:	325.423	**Durbin-Watson:**	1.725
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4390.598
Skew:	2.123	**Prob(JB):**	0.00
Kurtosis:	15.648	**Cond. No.**	3.95e+03

Inference via statsmodels

```
fullmodel_sm = smf.ols(formula = "price ~ sqft + dist + beds + baths + year + type",
                       data = homes).fit()
fullmodel_sm.summary()
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1949.0670	745.203	-2.615	0.009	-3412.677	-485.457
type[T.multifamily]	-452.2352	77.451	-5.839	0.000	-604.352	-300.119
type[T.singlefamily]	335.7612	54.642	6.145	0.000	228.441	443.081
type[T.townhouse]	-76.4372	56.859	-1.344	0.179	-188.111	35.237
sqft	0.6411	0.044	14.720	0.000	0.556	0.727
dist	-173.5430	20.099	-8.634	0.000	-213.018	-134.067
beds	-89.9345	23.532	-3.822	0.000	-136.152	-43.717
baths	198.4646	31.332	6.334	0.000	136.928	260.002
year	1.2300	0.388	3.169	0.002	0.468	1.992

Dep. Variable:	price	**R-squared:**	0.733
Model:	OLS	**Adj. R-squared:**	0.729
Method:	Least Squares	**F-statistic:**	200.0
Date:	Tue, 03 Oct 2023	**Prob (F-statistic):**	1.14e-161
Time:	22:00:14	**Log-Likelihood:**	-4391.8
No. Observations:	592	**AIC:**	8802.
Df Residuals:	583	**BIC:**	8841.
Df Model:	8		
Covariance Type:	nonrobust		

Omnibus:	259.016	**Durbin-Watson:**	1.914
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4084.354
Skew:	1.507	**Prob(JB):**	0.00
Kurtosis:	15.510	**Cond. No.**	1.18e+05

Take home message

By taking a probabilistic approach to linear regression and assuming the residuals are normally distributed, we see that **maximizing the likelihood** for this model is equivalent to **minimizing mean squared error** around the line!

So, if we believe our residuals are normally distributed, then minimizing mean square error is a natural choice.

But by choosing this specific probability model, we get much more than simply motivation for our loss function. We get *instructions* on how to perform inferences as well ☺

CIs and Hypothesis tests can be performed via resampling or using formulas!

CS109A
GAME Time

What is the goal of visualization?

Options

A. To explore the distributions of variables.

B. To explore the data to build hypotheses.

C. To communicate results of your models.

D. To trick and manipulate your audience.