Lecture #10: Bayes

aka STAT109A, AC209A, CSCIE-109A

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Gumb

Lecture Outline: Bayes

- Inference Review
  - Confidence Intervals
  - Hypothesis Tests
  - Likelihood
- Bayes Formula
- Bayes 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 last week: 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**

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 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 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:

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

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 $z^* = 1.96$)

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.

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?

?

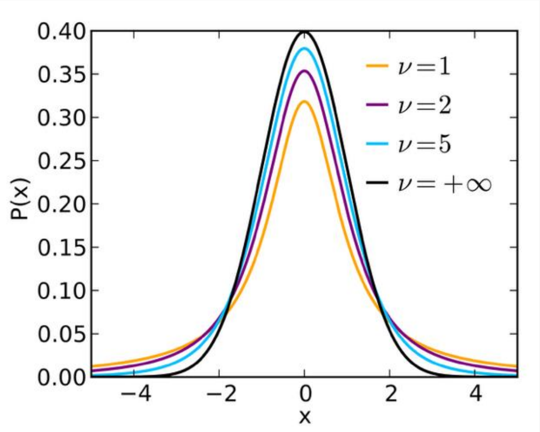
To compare the t-test values of the predictors from our model, $|t - test|$, with the t-tests calculated using permuted data, $|t^R|$, we estimate the probability of observing $|t^R| \geq |t - test|$.

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-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.

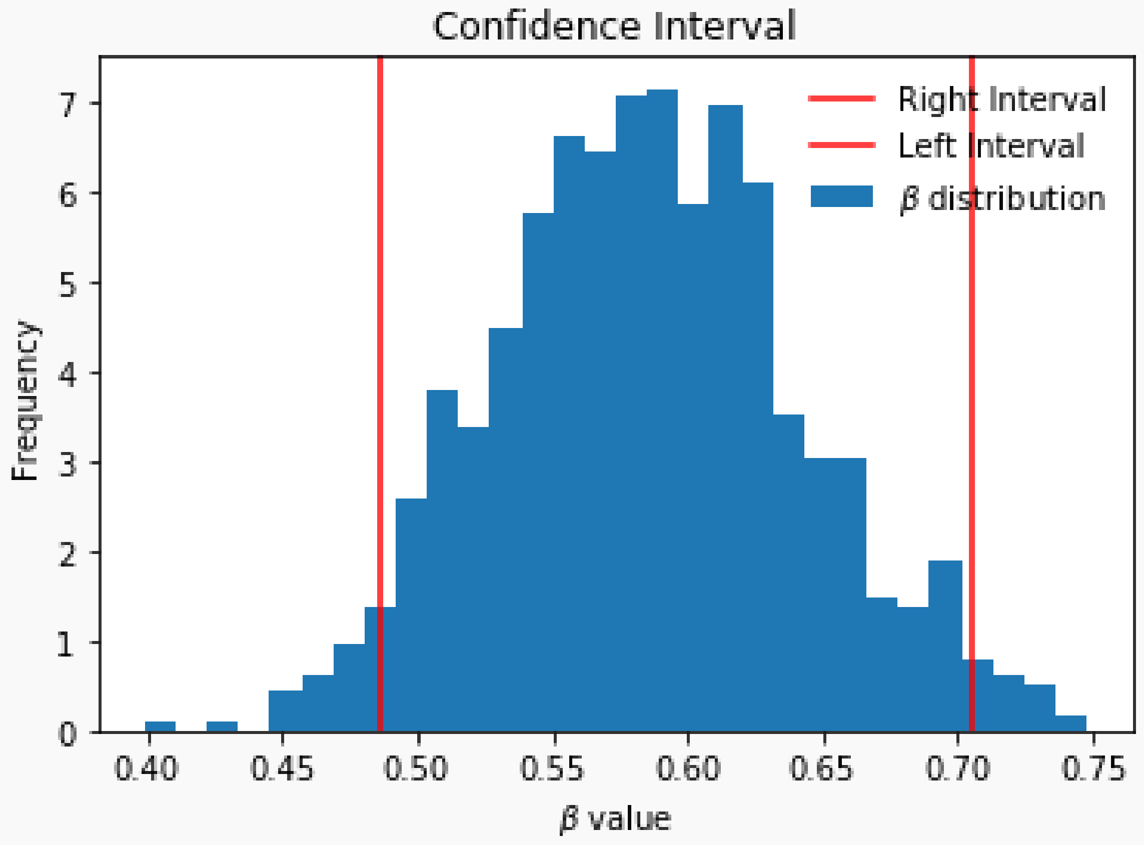
In a permutation test, you resample the data assuming the null hypothesis is true. This can most easily done by shuffling the response variable while keep the columns of the predictors as-is.

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()
```

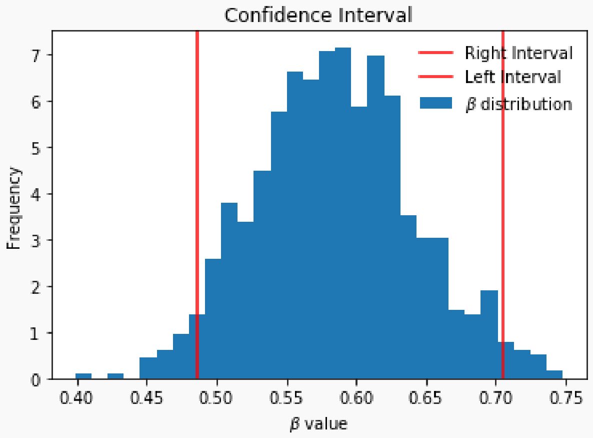
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1949.0670	745.203	-2.615	0.009	-3412.677	-485.457
type[T.multifamily]	-452.2352	77.451	-5.839	0.000	-604.352	-300.119
type[T.singlefamily]	335.7612	54.642	6.145	0.000	228.441	443.081
type[T.townhouse]	-76.4372	56.859	-1.344	0.179	-188.111	35.237
sqft	0.6411	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

CS109A
GAME Time

What happens to this distribution when B (the number of bootstrap samples) increases?



Options (pick all that apply)

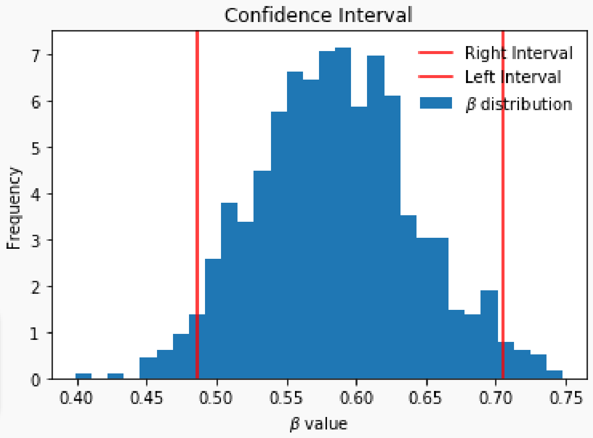
A. The distribution becomes more normal.

B. The variance decreases.

C. The resulting confidence interval becomes narrower.

D. The distribution gets smoother.

What happens to this distribution when B (the number of bootstrap samples) increases?



Options (pick all that apply)

A. The distribution becomes more normal.

B. The variance decreases.

C. The resulting confidence interval becomes narrower.

D. The distribution gets smoother.

Use this output to predict (with 95% uncertainty) the selling price of a 2860 square foot home

Options (pick all that apply)

A. 0.5898 +/- 2(0.023)

B. 0.5898(2860) +/- 2(0.023)

C. 247.4 + 0.5898(2860) +/- 2(0.023)

D. $247.4 + 0.5898(2860) \pm 2\sqrt{0.023^2 + \widehat{\sigma^2}}$

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

Use this output to predict (with 95% uncertainty) the selling price of a 2860 square foot home

Options (pick all that apply)

A. 0.5898 +/- 2(0.023)

B. 0.5898(2860) +/- 2(0.023)

C. 247.4 + 0.5898(2860) +/- 2(0.023)

D. 247.4 + 0.5898(2860) +/- 2 $\sqrt{0.023^2+\widehat{\sigma^2}}$

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

Lecture Outline: Bayes

- Inference Review
    - Confidence Intervals
    - Hypothesis Tests
    - Likelihood
- Bayes Formula
- Bayes 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.

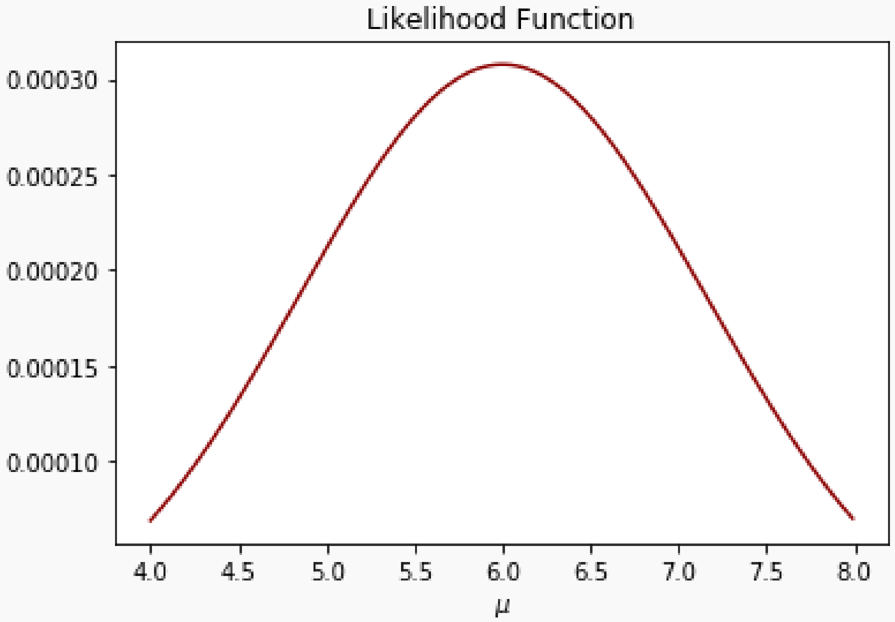
For a set of independent and normally distributed random variables, $X_i \sim N(\mu, \sigma^2)$:

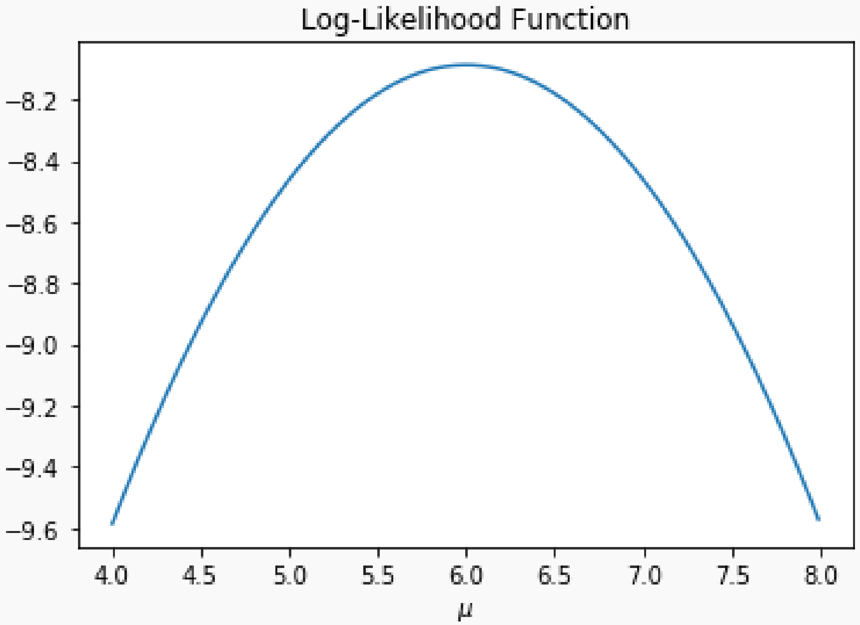
$$\mathcal{L}(\mu, \sigma^2 | x_1, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}$$

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 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

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

Lecture Outline: Bayes

- Inference Review
  - Confidence Intervals
  - Hypothesis Tests
  - Likelihood
- Bayes Formula
- Bayes Inference

Conditional Probability

Let A and B be events describing random experiment/phenomenon. Then the conditional probability of A occurring given B has occurred is defined as:

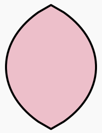
$$P(A|B) = \frac{P(A \text{ } and \text{ } B)}{P(B)}$$

Example: 40% of undergrads in 109A are Stat concentrators and 60% are CS concentrators. 20% of undergrads are both (either joint or double)

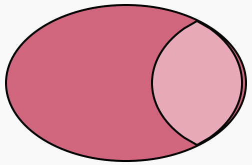
1. Define 2 events to describe this scenario.
2. Are these 2 events disjoint?
3. Are these 2 events independent?
4. Determine P(CS | Stat) and P(Stat | CS).
5. Interpret the two conditional probabilities in the previous part. What do they represent?

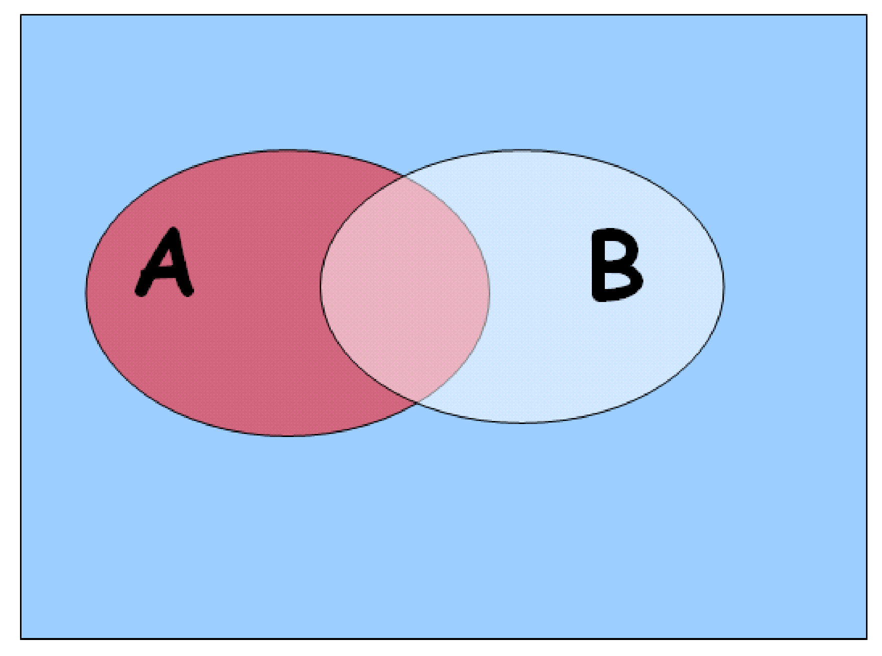
Conditional probability as a Venn Diagram

$$P(B|A) = \frac{P(B \text{ and } A)}{P(A)}$$

$=$ 

―――――――



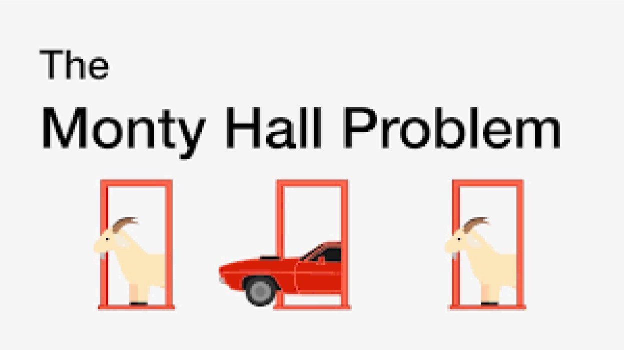


Conditional probability: how much does *B* take up within *A*?

In other words, restricting yourself only to *A*, how much does the intersection with *B* take up?

Very tricky…

- The Monty Hall Problem
  - There are prizes behind 3 doors: two are 'worthless' (like a goat) and one is expensive (like a new car)
  - You are asked to choose one of the 3 doors
  - Then, Monty Hall (from *Let's Make a Deal*) opens one of the other 2 doors and shows you a worthless prize



- **Should you switch doors?**

- **NYTimes take:**

  https://www.nytimes.com/2008/04/08/science/08tier.html

Bayes Rule

- Bayes' rule (formula) provides a way to go from $P(B \mid A)$ to $P(A \mid B)$ (they are in general not equal…)
- If A and B are two events whose probabilities are not 0 or 1:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

- Determine P(Stat | CS) from the fact that P(CS | Stat) = 0.50.

Bayes Rule, another example

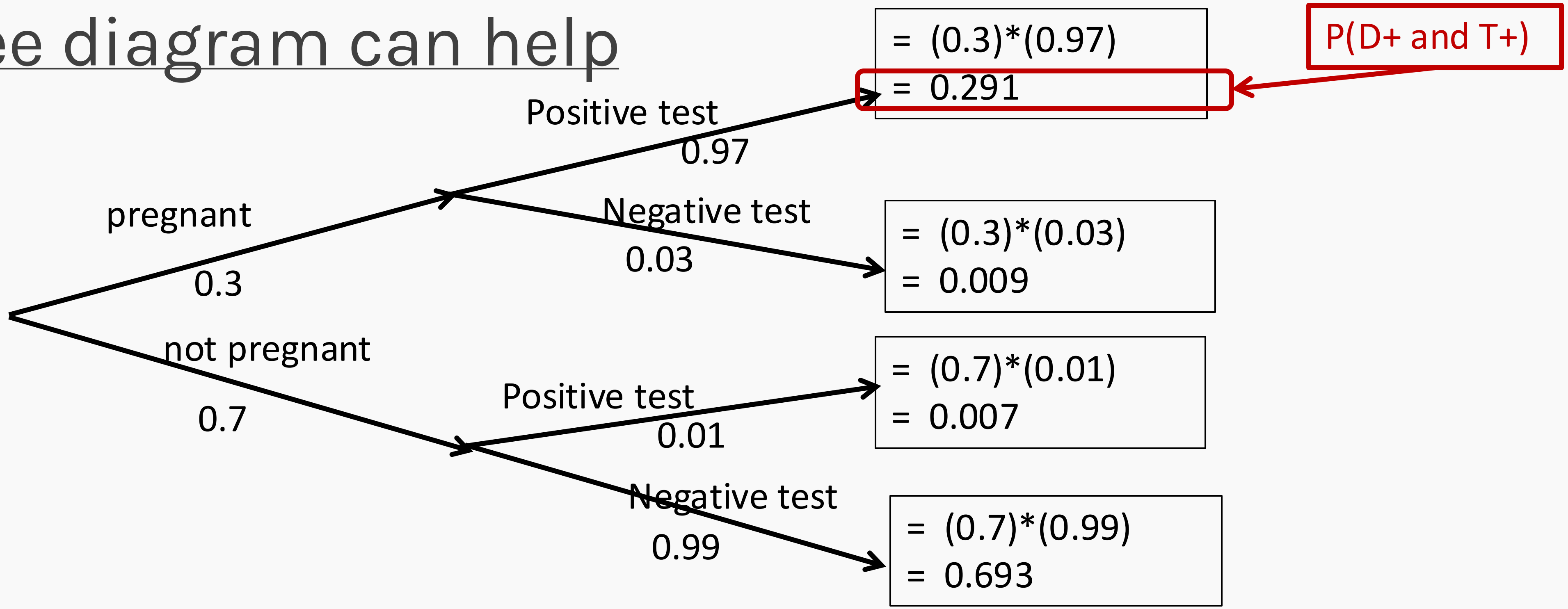
- Pregnancy tests are quite accurate.
- Historical data indicate that:
  - $P(T+ / D+) = 0.97$ (sensitivity)
  - $P(T- / D-) = 0.99$ (specificity)
- Those taking pregnancy tests are truly pregnant less often than one might think: $P(D+) = 0.30$
- We can use Bayes' Rule to determine what we truly care about: you take a pregnancy test and test positive.
- What is the chance you are actually pregnant?

Bayes Rule, worked example

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)}$$

$$= \frac{0.97 \cdot 0.30}{(0.97 \cdot 0.30) + (0.01 \cdot 0.70)} = 0.9765$$

A tree diagram can help



Bayes Rule, priors and posteriors

- Note: this calculation is based off the fact that your chance of being pregnant before taking the test was assumed to be 30%.
  - This is called the **prior probability**.
  - This may not actually be 30%. Maybe you believe you have more like a 50% chance.
- This probability was updated to be 97.65% after testing positive based on the test.
  - This is called the **posterior probability**.
- This change from prior to posterior is essentially *updating* the probability given evidence.
- This can be applied to theory (parameters) and data…

Lecture Outline: Bayes

- Inference Review
  - Confidence Intervals
  - Hypothesis Tests
  - Likelihood
- Bayes Formula
- **Bayesian Inference**

Bayes Rule, for distributions!

- We just saw the simplest form of Bayes' Rule:

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

- What is Bayes' Rule effectively doing?

- How would this be useful for statistical inference?
  *Think: parameters (θ) and data (X).

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Bayes Rule/Inference, for continuous RVs

- This can be rewritten for a set of parameters, θ, treating it as a continuous random variable, in terms of PDFs:

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)}$$

- Let's break this down:

  X : the vector (or matrix) of data: $X_1, X_2, ..., X_n$

  θ : the vector of parameters (or just a scalar).

$f(\mathbf{X}|\theta)$: the likelihood of X_i's

$f(\mathbf{X})$: the marginal pdf of X_i's (just a normalizing constant)

$f(\theta)$: the *prior distribution* of θ

$f(\theta|\mathbf{X})$: the *posterior distribution* of θ

Bayesian Inference: from prior to posterior

- The prior distribution, $f(\theta)$, often is based on a known distribution with it's own set of parameters. These are called *hyperparameters*.
- The marginal PDF of X is the distribution of X ignoring θ. How do you solve for a marginal PDF based on a joint distribution?

$$f(X) = \int_\theta f(x,\theta)d\theta = \int_\theta f(X|\theta)f(\theta)d\theta$$

- By definition, this marginal PDF of X will not involve θ. Thus, it can be though of as a multiplicative normalizing constant with respect to θ.
- So we can write the posterior dist. as proportional to:

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)} \propto f(X|\theta)f(\theta)$$

Bayesian Inference, a very simple example

- You own 3 coins: a fair one (with $p = 0.50$ of landing heads) and two biased coins (one with $p = 0.10$ and the other $p = 0.90$). You reach into your pocket and select one coin at random to flip.
- You flip it 4 times and see 3 heads and one tail.
- Intuitively, which coin(s) do you feel are plausible to have been the one chosen? What if you had to pick just one?
- What is the posterior distribution for p?

$P(p = 0.10 \mid X) = 0.007$, $P(p = 0.50 \mid X) = 0.458$, $P(p = 0.90 \mid X) = 0.535$

- Now which coin do you believe was chosen? Are you certain?
- What would happen if $n = 4$, $k = 2$? What about if $n = 40$ and $k = 30$?
- Note: this parameters space is discrete, which is rarely the case in practice.

Bayesian Perspective

- So how is this Bayesian approach different from the Frequentist approach (which typically only uses the likelihood function)?
- It also relies on a prior distribution. So an analyst has to place some *a priori* probability on the distribution of the parameter.
- This adds some extra uncertainty into the approach. Different analysts can come up at the same problem with different priors, and thus get different results ☹
- But this is really no different than different Frequentists making different assumptions on the data (independence, specific properties of the underlying distribution of the X_i's, etc…)

Bayesian Probability of θ

- The other difference from a Frequentist's approach is now we have distribution(s) of the parameter(s) (both the prior and the posterior distributions).
- So what is this probability distribution really measuring?
- A Frequentist's "definition" of probability: the long run expected **frequency** of an occurrence of a random variable if an experiment is performed an infinite number of times. Can only be applied to random things.
- A Bayesian's "definition" of probability: a measure or description of belief or plausibility…and can be applied to any unknown quantities ☺ Random entities **or** unknown latent variables/parameters.
- Sounds a whole lot like a Frequentist's use of the word *confidence* in a Confidence Interval!

Bayesian's Prior and Posterior

- A Bayesian's prior distribution, $f(\theta)$, captures one's prior belief or experience of the parameter. This belief should be updated based on what? The data!!! $X_1,\ldots,X_n$
- And the posterior distribution, $f(\theta \,/\, X_1,\ldots,X_n)$, can be thought of exactly this way: as a measure of belief on the parameter given the data seen in the sample.
- And how should this belief be updated? Weighted based on the likelihood!
- So more likely values of θ will have more bearing on the posterior, given the data we see.
- So once the data is fixed at what is actually measured, then the posterior will be weighted towards values of θ that agree with those measurement.

Bayes Approaches to Frequentist Ideas

- Bayesian inferences on the parameters, θ, can then be based solely on the posterior distribution. Which makes life simple!
- The posterior is not exactly a sampling distribution though. Why not?
- But the posterior is a measure of uncertainty of the parameter, and can be used to examine the uncertainty of an estimator.
- The posterior can also be used to calculate Bayesian analogues to Frequentist inferential techniques: interval estimates and hypothesis tests!

Which is better: Bayes or Frequentist?

- So which should we use: the Bayesian approach or the Frequentist approach?
- It depends on the setting. And depends on who you are doing the work for.
- Frequentist approaches are classical approaches, and were developed first because they were easy to solve.
- Bayesian approaches usually are more computationally intensive, and only recently (10+ years) have taken off.
- In practice in modern times, both approaches are often used for the same data and both analyses are presented.
- Both often give quite similar results.
- At the very least, we first have to define what an estimator is in the Bayesian paradigm…

Bayesian Normal-Normal Model

- Let $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ and where σ^2 is known (maybe from a previous study). Let's put a prior on $\mu \sim N(\mu_0, \sigma_0^2)$.
- What are the parameter(s) and the hyperparameters?
- Write down the prior:
- Write down the likelihood:
- Write down the normalizing constant (the denominator):

Normal-Normal Model: Posterior Result

- So the posterior distribution is:

$$\mu|X = N\left(\frac{\sigma^2\mu_0 + n\sigma_0^2\bar{X}}{\sigma^2 + n\sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}\right)$$

- So what?
- The posterior distribution for the mean of a normal distribution, given the data, only depends on the sample data in terms of the sample mean. The posterior of μ is normally dist. (if we start with a prior that is normally dist.).
- What is the posterior mean estimator (the mean of this distribution)?
- The posterior mean of μ is a weighted average of the prior mean, μ_0, and n-times the sample mean. So what happens to the effect of the prior on the posterior (and the estimator) as n increases?
    - The variance of the posterior decreases as n increases.