Data Mining
W4240 Section 001

Prof. Giovanni Motta

Columbia University, Department of Statistics

September 21, 2015

Outline

Outline

Administrative Notes

- Reminder: syllabus $\rightarrow$ forum $\rightarrow$ TAs $\rightarrow$ professors
- Did I register for this class?
- HW01 due next Wednesday September 23 online before class
- Today: Probability
- Next time: Probability & dimension-reduction

Outline

Bernoulli Distribution

- Coin toss with parameter y
- $X \sim Ber(y)$
- (or perhaps $X|Y = y \sim Ber(y)$)

$$p(X = k) = \begin{cases} 1 & y \\ 0 & 1 - y \end{cases}$$

- $\mathbb{E}(X)$ y
- $\psi_X(t)$ $y \exp(t) + (1 - y)$
- $\mathrm{Var}(X)$ $y(1 - y)$
- y... just a parameter

Outline

Describing Random Variables

- Probability mass (or density): $f_X(x) = \frac{\partial}{\partial x} F_X(x)$
- Cumulative mass (or distribution): $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$
- Expectation: a weighted average, which describes the mean of a random variable

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} t\, f_X(t)dt$$

- Variance: the second moment around the mean, which describes the spread of a random variable

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[X - \mathbb{E}(X)\right]^2 \\ &= \mathbb{E}\left(X^2\right) - \left[\mathbb{E}(X)\right]^2\end{aligned}$$

- Moment Generating Function:

$$\psi(t) = \mathbb{E}\left(\exp\{tX\}\right)$$

Describing Random Variables

... but there are other ways to describe spread of a random variable

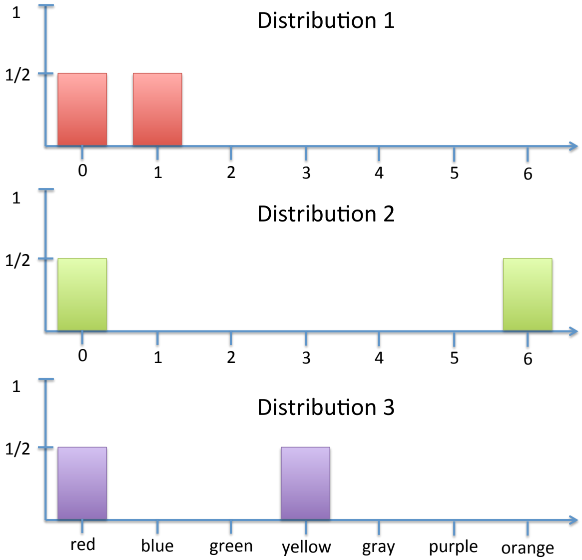
$$\text{Let } p_X(x) := \mathbb{P}(X = x)$$

Entropy: the expectation of the negative log density, which describes the uncertainty or unpredictability of a random variable

$$H(X) = -\sum_x p_X(x) \log p_X(x)$$

Note: $\lim_{p \to 0^+} p \log p = 0$ (exercise!)

Entropy vs. Variance



Entropy of a Bernoulli Random Variable

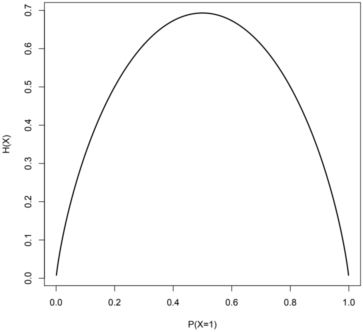
- Flip a coin with probability of heads π

$$X \sim Ber(\pi)$$
$$p(X = k) = \pi^k (1 - \pi)^{1-k}, \qquad k = 0, 1$$

- $X \sim Bernoulli(\pi)$: which value of π maximizes entropy? Which minimizes?

Entropy of a Bernoulli Random Variable

```
pr.X0=seq(0,1,length.out=1000)
pr.X1=1-pr.X0
H=-pr.X0*log(pr.X0)-pr.X1*log(pr.X1)
plot(pr.X1,H,type="l",lwd=2,xlab="P(X=1)",ylab="H(X)")
```



Review of multiple random variables

- Probability is the foundation for dealing with data
- Joint cdf:
$$F_{XY}(a,b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{XY}(x,y)dxdy$$
- Marginalization:
$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$$
- Conditioning:
$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$
- Independence:
$$f_{XY}(x,y) = f_X(x)f_Y(y)$$
- Bayes Rule:
$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_x f_{Y|X}(y|x)f_X(x)dx}$$

Describing Two Random Variables

You have probably already learned about one way to describe how much the distribution of one random variable, X, tells you about the distribution of another, Y:

Covariance: a measure of linear relationship between variables

$$Cov(X, Y) = E\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right]$$

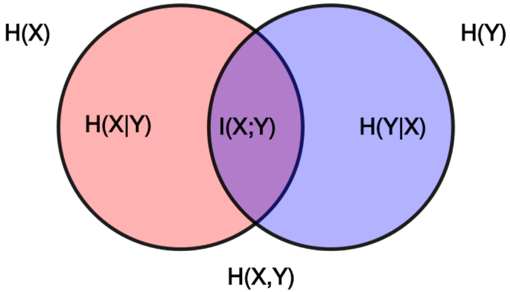
Describing Two Random Variables

... but again, there are other ways to describe dependencies.

Mutual Information: measures the mutual dependence between two random variables

$$
\begin{aligned}
I(X;Y) &= \sum_x \sum_y f(x,y) \log \frac{f(x,y)}{f(x)f(y)} \\
&= H(X) + H(Y) - H(X,Y) \\
&= H(Y) - H(Y \mid X) \left(= H(Y) - \sum_x f(x) H(Y \mid X = x) \right) \\
&= H(X) - H(X \mid Y) \\
&= I(Y;X)
\end{aligned}
$$

Mutual Information



[from Wikipedia]

Why Mutual Information?

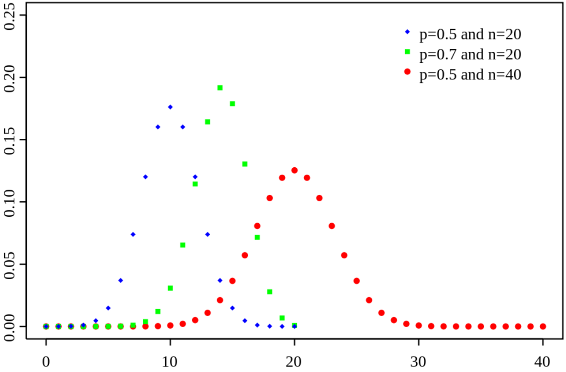
Suppose we have weather data where we would like to predict whether it is a nice day given Temperature (Low, Med, High) and whether it is cloudy (Y/N):

Temp	Cloudy	Nice Day
High	N	Y
Low	Y	N
Med	Y	N
Med	Y	Y
Low	N	N

What is $I(Nice\ Day\,;\,Cloudy)$? What is $I(Nice\ Day\,;\,Temp)$? Which is a better predictor?

Binomial Distribution

$$P(Z = k) = \binom{n}{k} p^k (1-p)^{(n-k)}$$



[wikipedia]

Binomial Distribution

- Number of successes in n trials
- $Z \sim Binom(n, p)$

$$P(Z = k) = \binom{n}{k} p^k (1 - p)^{(n-k)}$$

- $\mathbb{E}(X)$ np
- In R, you can get random numbers from a binomial distribution with `rbinom(n, size, prob)`.
- $\mathrm{Var}(X)$ $np(1 - p)$
- Variance of the sum of independent random variables? prove
- Sometimes mgf, sometimes independence, etc.

Multinomial distribution

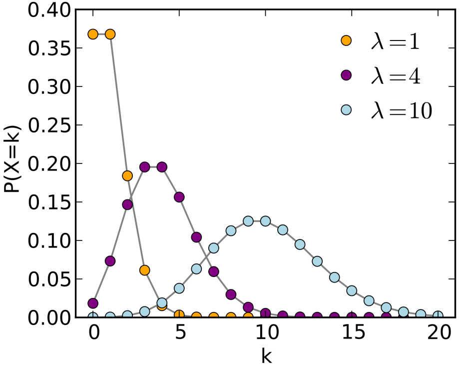
- An extension of the binomial distribution to K categories (instead of 2)—n trials each with K possible outcomes
- It is parameterized by a point on a probability simplex $\pi = (\pi_1, \ldots, \pi_K)$ where $\sum_{k=1}^{K} \pi_k = 1$. Here π_k is the probability of choosing category k.
- (what's a probability simplex?)
- Let $x_k \in \{0, ..., n\}$ be the outcome for category k.

$$(X_1, \ldots, X_K) \sim Multi(n, \pi)$$
$$p(x_1, \ldots, x_K \,|\, \pi) = \frac{n!}{x_1! \ldots x_K!} \pi_1^{x_1} \ldots \pi_K^{x_K}$$

- In `R` you can get random multinomial numbers with `rmultinom(n, size, prob)`.

Poisson Random Variables in Pictures

$$P(W = k) = \frac{\lambda^k}{k!}\exp(-\lambda)$$



[wikipedia]

Poisson Distribution

- Number of events with average λ (arrivals, decay, etc)

$$P(X = k) = \frac{\lambda^k}{k!}\exp(-\lambda)$$

- $\mathbb{E}(X)$ λ
- $\psi_X(t)$ $\exp(\lambda(\exp(t) - 1))$
- $\mathrm{Var}(X)$ λ
- mgf of sums of independent variables X_i? $\prod_i \psi_{X_i}(t)$
- Sum of independent Poisson random variables with rates λ_1 and λ_2?
- (skipping Poisson processes)

Outline

Useful Distributions: Continuous

Continuous random variables are slightly different than discrete:

- probability of any specific atom is 0: $P(X = x) = 0$ (so $p(x) \neq P(x)$)
- have *probability density function* (pdf), $p(x)$, that integrates to 1
$$\int_{-\infty}^{\infty} p(x)dx = 1$$
- to find probability of event A (eg, $A = \{X \in (-5, 2)\}$),
$$\begin{aligned} P(A) &= \int_A p(x)dx \\ &= \int_{-5}^{2} p(x)dx \end{aligned}$$

Uniform Random Variables

- Any point in a given interval $[a, b]$ has equal probability density
- $P(U \in [a, b])$? 1
- Probability density function? $f(u) = \frac{1}{b-a}\mathbb{1}(u \in [a, b])$
- $P(U \in [c, d])$? $\frac{\min(0,\min(d,b)-\max(c,a))}{b-a}$

$$P(U \in A) = \frac{1}{b-a}\int_A \mathbb{1}(u \in [a, b])du$$

- U is called a *Uniform* random variable $U \sim Unif(a, b)$

Exponential Random Variables

- Reminder: $X \sim Exp(\lambda)$ is called an *Exponential* random variable if:
$$f_X(x) = \lambda \exp\{-\lambda x\} \mathbb{1}(x > 0)$$

- $\mathbb{E}(X)$? $\frac{1}{\lambda}$

- $\psi_X(t)$? $\frac{\lambda}{\lambda - t}$

Gamma Distribution

- Reminder: We say $X \sim Gamma(\alpha, \beta)$ if:

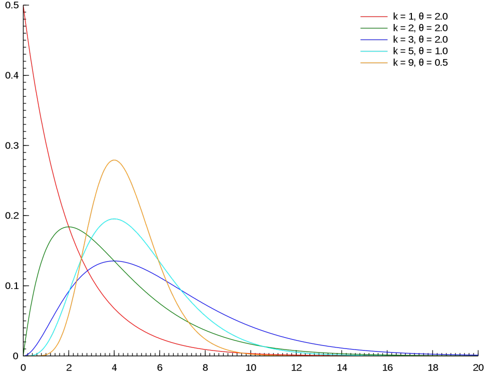
$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\} \mathbb{1}(x > 0)$$

- What is $\Gamma(\alpha)$? (let $\beta = 1$ for simplicity)
  - $\int_0^\infty x^{\alpha-1} \exp\{-x\} dx$? $\Gamma(\alpha)$
  - $\Gamma(\alpha)$ generalizes $(\alpha-1)!$ to non-integer values
  - Useful property: $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$
  - $\Gamma(1, \beta)$ for any β? $Exp(\beta)$

Gamma in Pictures

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\}$$

(note suppression of $\mathbb{1}(x > 0)$)



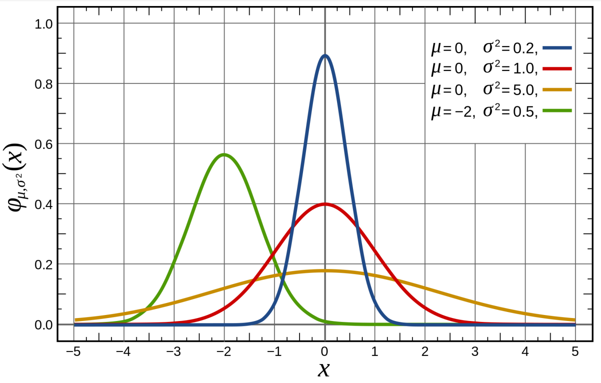
[textbook]

Gamma and Exponential example

- Light bulbs X_i fail according to $Exp(\lambda)$. Call Z the averaged failure times of n independent light bulbs.
  - $P(X_i > \tau)$? $\exp(-\lambda\tau)$
  - $\mathbb{E}(Z)$? $\frac{1}{\lambda}$
  - $\mathrm{Var}(Z)$? $\frac{1}{n\lambda^2}$
  - (really important statistical fact!)
  - Let $Y = nZ$ be the sum. $f_Y(y)$? $Y \sim Gamma(n, \lambda)$
  - $f_Z(z)$? $Z \sim Gamma(n, n\lambda)$
  - $\mathrm{Var}(Z)$? $\frac{n}{(n\lambda)^2}$

Normal pictures

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



[wikipedia]

Normal distribution

- Reminder: $X \sim \mathcal{N}(\mu, \sigma^2)$ if:

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

- $\psi_X(t)$? $\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$
- $\mathbb{E}(X)$? μ
- $\mathrm{Var}(X)$? σ^2
- Distribution of $Y = aX + b$? $\mathcal{N}(a\mu + b, a^2\sigma^2)$
- $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ indep.; Distribution of $\sum_i X_i$? $\mathcal{N}\left(\sum_i \mu_i, \sum_i \sigma_i^2\right)$
- $X_i \sim_{iid} \mathcal{N}(\mu, \sigma^2)$; Distribution of $\frac{1}{n}\sum_i X_i$? $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Normal cdf

- $F_X(x) = \int_{-\infty}^{x} f_X(u)du$ is ugly (not closed-form)

- Define it away. Take $Z \sim \mathcal{N}(0,1)$, and define:

$$\Phi(z) = \int_{-\infty}^{z} f_Z(u)du = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du$$

- $F_X(x)$ in terms of $\Phi(\cdot)$? $\Phi\left(\frac{x-\mu}{\sigma}\right)$

Normal problem example

- The ith mouse in a colony has weight $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Call Z the averaged weights of n independent mice.
  - $\mathbb{E}(Z)$? μ
  - $\mathrm{Var}(Z)$? $\frac{\sigma^2}{n}$
  - Let $W = Z - \mu$; $f_W(w)$? $\mathcal{N}(0, \frac{\sigma^2}{n})$
  - (really important statistical fact!)
  - What is the probability that Z is more than $2\frac{\sigma}{\sqrt{n}}$ from μ?

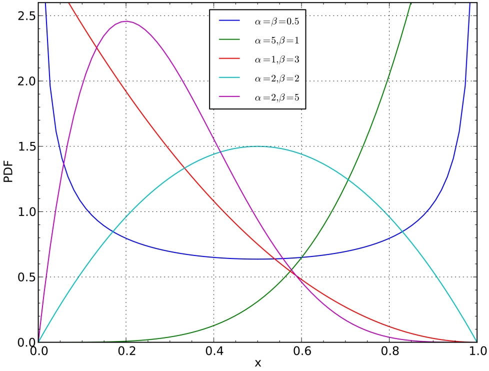
$$
\begin{aligned}
P\left(|W| \geq 2\frac{\sigma}{\sqrt{n}}\right) &= P\left(W < -2\frac{\sigma}{\sqrt{n}}\right) + P\left(W > 2\frac{\sigma}{\sqrt{n}}\right) \\
&= \Phi(-2) + (1 - \Phi(2)) \\
&= 2\Phi(-2) \\
&\approx 0.05
\end{aligned}
$$

  - statistical confidence intervals...

Beta distribution

- We say $X \sim Beta(\alpha, \beta)$ if:

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}(x \in [0,1])$$



[wikipedia]

Another review example

- n devices are produced by a factory, each with an independent probability p of being defective.
- X is # defective devices. What is $f_X(x)$? $f_X(x) = Binom(n, p)$
- How many defective devices will I get on average? $\mathbb{E}(X) = np$
- Now you know that one or more devices has been found defective. How many defective devices do I expect, given this information?

$$\mathbb{E}(X|X \geq 1) = \sum_{k=0}^{n} k \frac{P(X = k, X \geq 1)}{P(X \geq 1)}$$

- Note: $P(X \geq 1) = 1 - P(X = 0) = 1 - (1-p)^n$
- Note: $P(X = k, X \geq 1) = P(X = k)\mathbb{1}(k > 0)$
- which implies:

$$\mathbb{E}(X|X \geq 1) = \frac{np}{1 - (1-p)^n}$$

Outline

Statistical Models

Statistical models describe the relationships between random variables

- Suppose that data come from an unknown distribution
- See data, infer properties of the distribution (state of the world!)

Example properties:

- bias of a coin
- the average age of a student
- the average volatility of a stock
- the average volatility of a stock given that the S&P 500 had a 2% change yesterday

Statistical Models

Desirable model features:

- model is smaller than data (data compression)
- efficient explanation of past (low error, low complexity)
- efficient prediction of future (high generalization, low complexity)

To meet these goals:

- number of possible models is limited through construction
- models are determined by parameters (but not necessarily a finite number)

Statistical Models

Simplest type of model: fit a probability distribution to data

Examples:

- Binary data (a series of 0/1 outcomes)
  - fit a Bernoulli distribution
- Continuous data (a series of values between $-\infty$ and ∞)
  - fit a Gaussian distribution (or Cauchy distribution or gamma, etc)
- Categorical data (K categories)
  - fit a multinomial distribution

Fitting Parametric Distributions

Suppose that we see random variables $X_1, \ldots, X_n$. If we fit a distribution to the data, we are assuming the data are *independent and identically distributed (i.i.d.)* from the distribution:

- (independent) X_i is independent from X_j for all $i \neq j$
- (identically distributed) X_i has the same distribution as X_j for all $i \neq j$

Are i.i.d.:

- n flips of the same coin
- n rolls of the same die

Not i.i.d.:

- a sequence of n words from a text (not independent)
- a sequence of n coin flips where each flip is of a different–and possibly unfair–coin (not identically distributed)

Statistical Models: Parameters

Parameters are values that index a statistical model.

Example: Bernoulli random variables

- suppose that $X \sim Ber(\pi)$: $p(1) = \pi$, $p(0) = 1 - \pi$
- the likelihood of seeing $X = x$ given parameter π is

$$p(x \mid \pi) = \pi^x (1 - \pi)^{1-x}$$

- changing π leads to different distributions

Example: Gaussian random variables

- a Gaussian distribution has two parameters, μ and σ^2
- the likelihood of x given (μ, σ^2) is

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Example: Modeling Coin Flips with a Bernoulli Distribution

Data: flip a coin n times, get observations $x_1, \ldots, x_n$ (0 for tails, 1 for heads)

Model: fit with Bernoulli distribution, try to find best value for π given data

Model assumptions:

- data are binary
- data are i.i.d.

Assumptions are perfectly met. How to choose π?

Outline

Maximum likelihood estimation

- Choose model parameters to maximize the likelihood of the data $X_1, ..., X_n$
  - (recent centenarian)
  - Probably the most important method in statistics.
  - **Probably the most important method in statistics.**
- Why is this a sensible thing to do? (draw pictures)

Likelihood

The *likelihood* of a model is a function of a set of parameters given a set of observed outcomes.

- use likelihood to select parameters given data
- for data x and parameter θ,

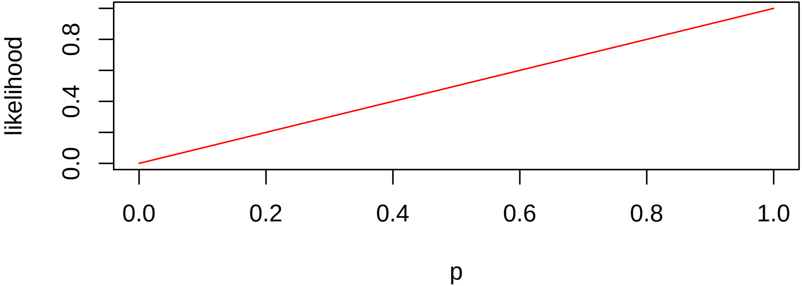
$$\mathcal{L}(\theta \mid x) = p(x \mid \theta)$$

Example: n coin flips, outcomes $x_1, \ldots, x_n$ (1 for heads, 0 for tails)

$$\begin{aligned}
\mathcal{L}(\pi \mid x_1, \ldots, x_n) &= p(x_1, \ldots, x_n \mid \pi) \\
&= \prod_{i=1}^{n} \pi^{x_i} (1-\pi)^{1-x_i} \\
&= \pi^{\sum_{i=1}^{n} x_i} (1-\pi)^{\sum_{i=1}^{n} (1-x_i)}
\end{aligned}$$

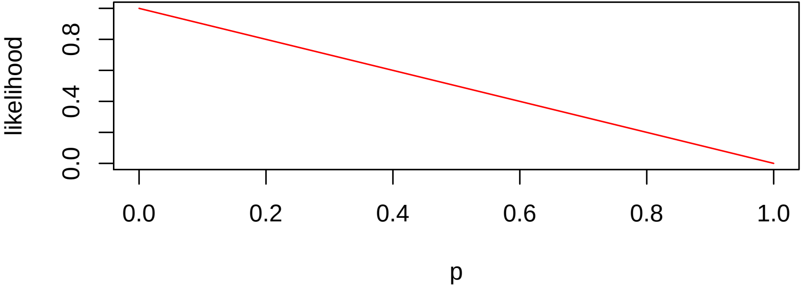
Likelihood

Example: $X_i \sim Ber(\pi)$, observed H; $\mathcal{L}(\pi) = \pi$



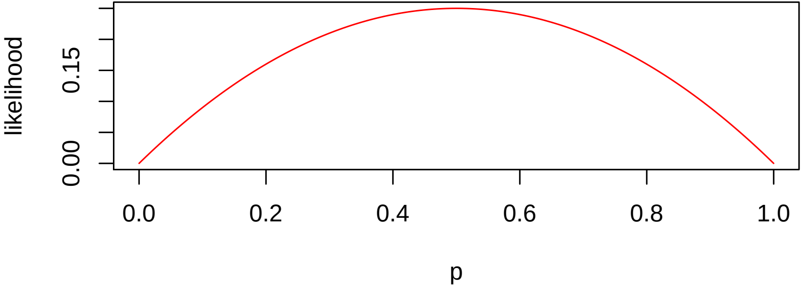
Likelihood

Example: $X_i \sim Ber(\pi)$, observed T; $\mathcal{L}(\pi) = 1 - \pi$



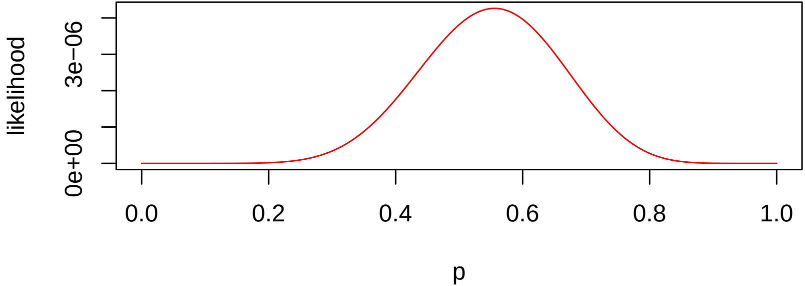
Likelihood

Example: $X_i \sim Ber(\pi)$, observed H, T; $\mathcal{L}(\pi) = \pi(1-\pi)$



Likelihood

Example: $X_i \sim Ber(\pi)$, observed 10 H, 8 T; $\mathcal{L}(\pi) = \pi^{10}(1-\pi)^8$



Maximum Likelihood

Idea: maximize the likelihood function to find the parameter θ! The θ found in this manner is called the *maximum likelihood estimate*, $\hat{\theta}^{MLE}$.

$$\hat{\theta}^{MLE} = \arg\max \mathcal{L}(\theta \,|\, x_1, \ldots, x_n)$$

Properties:

- consistent: if this is the correct data generating distribution, then our estimate will become correct as we get more data
- ...but can get silly estimates with small amounts of data ($\hat{\pi}^{MLE}$ given H observed is 1)

Finding the Maximum Likelihood

Try taking derivative of $\mathcal{L}$ and setting equal to 0.

Example: $X_i \sim Ber(\pi)$

$$\begin{aligned} \frac{d}{d\pi}\mathcal{L}(\pi \mid x_1, \ldots, x_n) &= \frac{d}{d\pi}\, \pi^{\sum_{i=1}^{n} x_i}(1-\pi)^{\sum_{i=1}^{n}(1-x_i)} \\ &= \end{aligned}$$

Finding the Maximum Likelihood

Instead, maximize the *log likelihood*,

$$\ell(\theta \,|\, x_1, \ldots, x_n) = \log\left(\mathcal{L}(\theta \,|\, x_1, \ldots, x_n)\right)$$

- log increasing function $\rightarrow \log(p(\theta))$ has same arg max as $p(\theta)$
- log turns products into sums, exponents into multiples

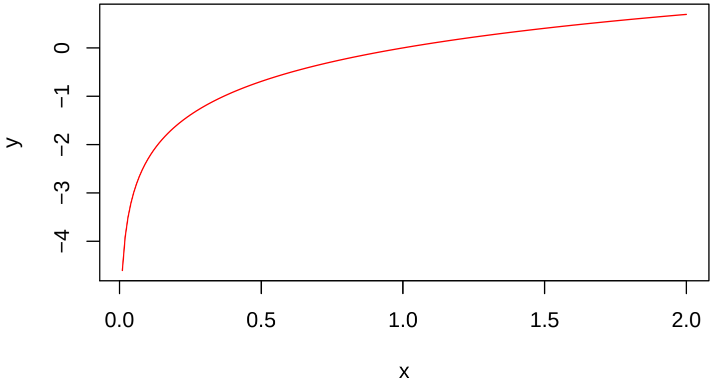
Max vs. Arg Max

- *max* is the maximal value
- *arg max* is the argument that generates maximal value
- Ex:
  - $\max\ (1 - x^2) = 1$
  - $\arg\max\ (1 - x^2) = 0$

Finding the Maximum Likelihood

The log likelihood (x is original, y is log value),

$$\ell(\theta \mid x_1, \ldots, x_n) = \log\left(\mathcal{L}(\theta \mid x_1, \ldots, x_n)\right)$$



Finding the Maximum (Log) Likelihood

Try taking derivative of ℓ and setting equal to 0.

Example: $X_i \sim Ber(\pi)$

$$\begin{aligned}\ell(\pi \mid x_1, \ldots, x_n) &= \log\left(\pi^{\sum_{i=1}^n x_i}(1-\pi)^{\sum_{i=1}^n (1-x_i)}\right)\\ \frac{d}{d\pi}\ell(\pi \mid x_1, \ldots, x_n) &= \\ \hat{\pi} &= \end{aligned}$$

Finding the Maximum (Log) Likelihood

Find the Gaussian MLE:

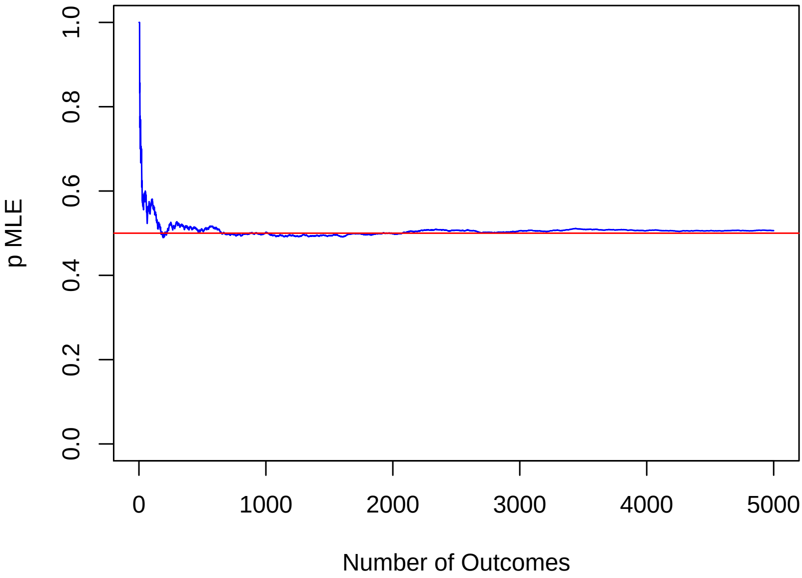
$$
\begin{aligned}
\mathcal{L}(\mu, \sigma^2 \mid x) &= \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\
&= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\
\ell(\mu, \sigma^2 \mid x) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\frac{\partial}{\partial\mu}\ell(\mu, \sigma^2 \mid x) &= \\
\hat{\mu} &= \\
\frac{\partial}{\partial\sigma^2}\ell(\hat{\mu}, \sigma^2 \mid x) &= \\
\hat{\sigma^2} &=
\end{aligned}
$$

Example: Bernoulli MLE

Flip a coin 5,000 times:

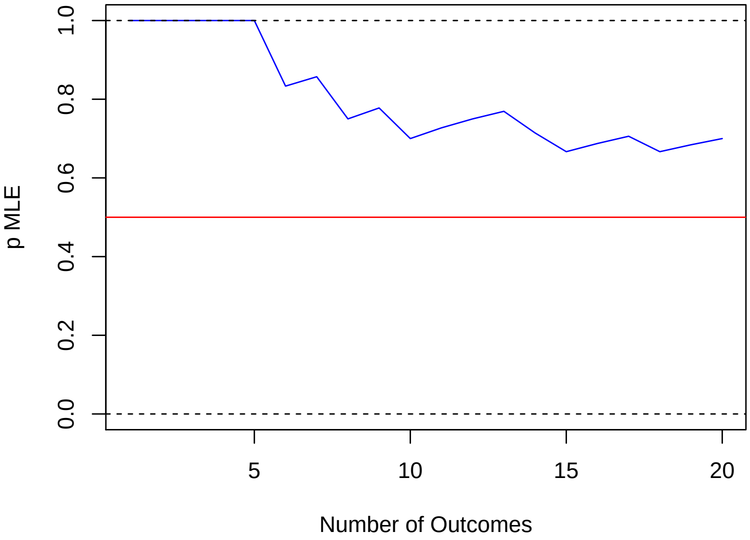
1 1 1 1 1 0 1 0 1 0 1 1 1 0 0 1 1 0 1 1 0 0 0 1 0 0 1 0 1 0 1 0 1 1
0 0 1 1 1 0 1 1 0 1 0 1 0 1 1 1 0 1 0 1 0 1 0 0 0 0 1 0 0 0 0 1 1 1
1 1 1 0 1 0 1 0 1 1 1 1 0 0 0 0 1 0 1 0 1 1 1 1 1 0 1 1 0 1 1 1 0 0
1 1 1 0 0 1 0 0 0 1 0 1 1 0 0 1 0 1 1 0 0 0 0 1 0 1 1 0 1 0 0 1 0 0
0 0 1 0 0 1 1 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 0 1 0 1 1 0 0 0 0 1 1 0
0 0 0 0 0 1 0 1 1 0 0 1 1 0 0 0 0 1 0 0 1 1 1 0 0 0 1 1 0 1 0 1 1 1
1 0 0 1 1 0 1 0 0 1 0 0 1 1 1 0 1 1 0 1 1 1 1 0 1 0 1 0 0 1 1 1 1 1
1 1 0 1 0 1 1 0 1 0 1 1 1 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 1 1 1 1 0
1 1 0 0 1 0 1 0 0 0 1 1 1 1 0 1 1 1 1 1 1 0 1 0 1 1 0 1 0 0 0 0 0 1
1 0 1 1 1 0 0 0 1 1 0 0 1 0 0 0 0 1 1 1 0 1 0 1 1 0 1 1 0 0 1 1 0 1
0 0 0 1 1 0 1 0 0 1 0 1 0 0 0 0 0 0 1 1 1 1 0 1 1 1 0 1 1 0 1 0 0 0
1 1 0 1 1 0 0 0 0 0 1 0 1 1 1 0 0 0 0 0 1 1 1 1 1 0 1 1 1 0 0 1 0 1
0 1 0 0 1 0 0 0 0 1 0 1 1 1 0 0 1 1 0 1 1 1 0 0 1 1 1 0 0 1 0 0 1 1
0 0 1 1 0 0 0 1 0 0 1 1 0 0 1 0 0 0 1 0 1 0 0 0 0 0 1 1 1 1 0 .....

Example: Bernoulli MLE



Example: Bernoulli MLE

First 20 outcomes: 1 1 1 1 1 0 1 0 1 0 1 1 1 0 0 1 1 0 1 1



Overfitting

Overfitting is a modeling problem that occurs when a model describes the noise in the training data rather than the underlying relationship between the data.

Here, the MLE says $\pi = 1$ because we have randomly seen a few 1's in a row.

Overfitting leads to poor predictions for unseen data.

When Should I Use the MLE?

Use when:

- data are i.i.d. from a parametric distribution
- you have a lot of data (100's or more observations)
- if less data, you can put a *prior* on the estimator to get more reasonable estimates (W4640 is Bayesian Statistics)

More examples:

- $X_i \sim_{iid} Exp(\theta)$. What is θ^*?
- $X_i \sim_{iid} Unif(a, b)$. What is a^* and b^*?

Outline

Homework and Next Time

Next Time: We will be moving to *An Introduction to Statistical Learning* by James, Witten, Hastie and Tibshirani.

☞ **Read Sections 10.1 & 10.2 for Wednesday September 23**

☞ **Homework**: turn in homework next Wednesday **before class**