Basic math

MAXIMUM LIKELIHOOD

Factorial

Factorial

$$\begin{aligned} x! &= x\,(x-1)\,(x-2)\cdots 1 \text{ for integers } x \geq 2 \\ &= \prod_{i=1}^{x} i \\ 3! &= 3 \times 2 \times 1 = 6 \end{aligned}$$

The factorial of 1 is 1, and, perhaps counterintuitively the factorial of 0 is also defined as 1.

$$0! = 1! = 1$$

Combinatorial Function

The combinatorial "choose" function is

1. The number of ways to choose x unique items from a larger set of n unique items, without regard to order
2. Also, the number of ways to order a set of binary outcomes

Combinatorial Function

Combinatorial function n choose x

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!} \text{ for non-negative integers } n \text{ and } x \text{ with } n \geq x$$

EX:

How many ways are there to draw three letters from the letters in GRAPH (without worrying about their order)?

"My fruit salad is a combination of apples, grapes and bananas" We don't care what order the fruits are in, they could also be "bananas, grapes and apples" or "grapes, apples and bananas", its the same fruit salad.

- Combination: Picking a team of 3 people from a group of 10. $C(10, 3) = 10!/(7! * 3!) = 10 * 9 * 8/(3 * 2 * 1) = 120.$

CDF and pdf

Probability density function (pdf) probability that random variable X takes on specific value equal to x (for discrete distribution)

Cumulative Distribution Function (CDF) probability that random variable X takes any of a range of values $\leq x$

Normal Distribution CDF and pdf

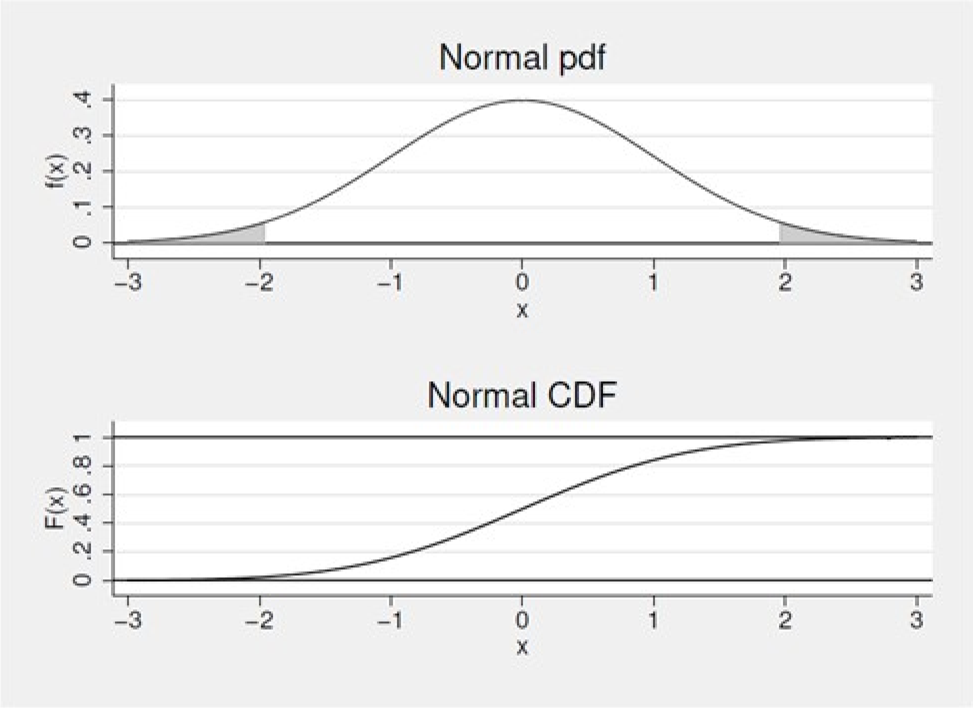


Figure 1.2: The pdf and CDF for the standard normal distribution.

pdf for Discrete Distribution

The *probability density function* (pdf) for a discrete random variable X is the probability that X equals x, denoted $f(x)$.

$$f(x) = \Pr(X = x)$$

The discrete pdf has the following properties:

1. $0 \leq f(x) \leq 1 \ \forall x$
2. $\sum_{i=1}^{k} f(x_i) = 1$ for x that takes on k values

CDF for Discrete Distribution

The *Cumulative Distribution Function* (CDF) for a discrete random variable X is the probability that X is less than or equal to x, denoted $F(x)$.

$$F(x) = \Pr(X \leq x)$$

The discrete CDF is equal to the sum of the corresponding discrete pdfs.

$$F(x) = \sum_{i=1}^{x} f(x_i)$$

The discrete CDF has the following properties:

1. $F(x)$ is monotonically increasing in x
2. $F(-\infty) = F(\text{less than lowest value of } x) = 0$
3. $F(\infty) = F(\text{highest value of } x) = 1$

Five Coin Spin

Spin a coin 5 times (n = 5)

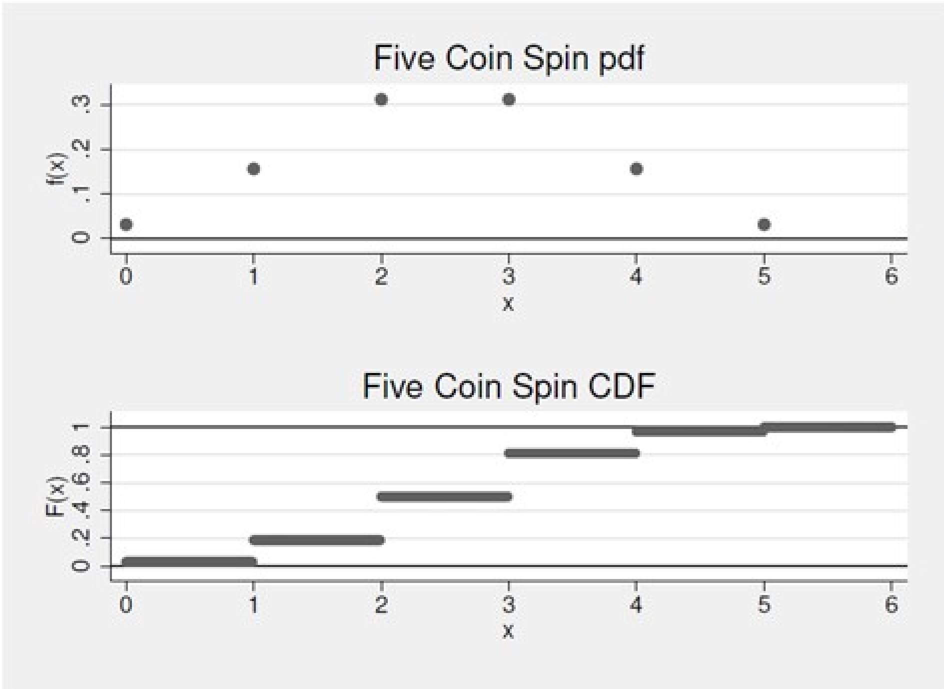
Number of heads X is a random variable

Possible outcomes for X are 0, 1, 2, 3, 4, 5

pdf(2) = Pr(X = 2) = .3125 → $\binom{5}{2}(.5)^2(.5)^3$

CDF(2) = Pr($X \leq 2$) = .5

Five Coin Spin pdf and CDF



Discrete variables, so just dots

Figure 1.1: The pdf and CDF for the number of heads after five coin spins.

CDF for Continuous Distribution

The *Cumulative Distribution Function* (CDF) for a continuous random variable X is the probability that X is less than or equal to x, denoted $F(x)$.

$$F(x) = \Pr\left(X \leq x\right) = \int\limits_{-\infty}^{x} f\left(u\right) du$$

The continuous CDF has the following properties:

1. $F(x)$ is monotonically increasing in x
2. $F(-\infty) = F(\text{less than lowest value of } x) = 0$
3. $F(\infty) = F(\text{highest value of } x) = 1$

pdf for Continuous Distribution (1)

The *probability density function* (pdf) for a continuous random variable X is the derivative of the CDF with respect to x, denoted $f(x)$.

$$f(x) = \frac{dF(x)}{dx}$$

Normal Distribution CDF and pdf

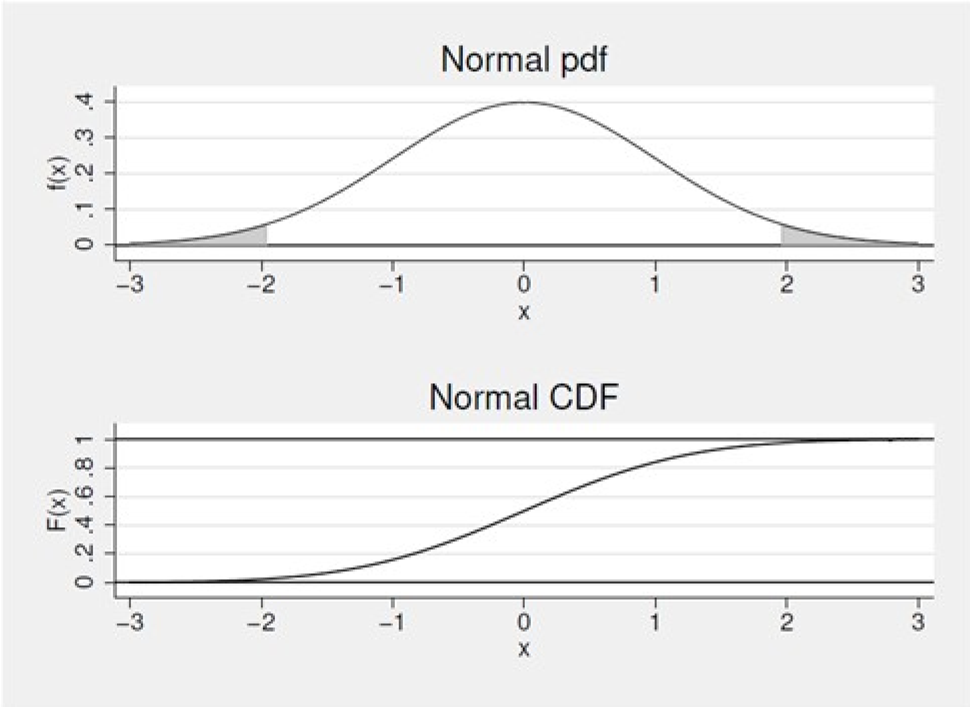


Figure 1.2: The pdf and CDF for the standard normal distribution.

4 Discrete Distributions

Binomial: predict number of heads after spinning a coin n times

Geometric: predict number of times n to get heads once

Poisson: approximation to binomial when p is small

Negative Binomial: predict number of trials needed to get outcome r times

Binomial

In more general terms, suppose a random event has two outcomes, A and B. After n trials, how many outcomes will be A? If X is a random variable with a binomial distribution, then the pdf is

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

where

$$\begin{aligned} X &= \text{number of times that get outcome A in } n \text{ trials (number of heads)} \\ n &= \text{number of independent trials (number of coin spins)} \\ p &= \Pr(\text{outcome A}) \\ x &= \text{specific number } (e.g., 0, 1, 2, 3)\text{; } x \text{ is not a random variable.} \end{aligned}$$

Binomial Example

The mean and variance of the random variable X are

$$
\begin{aligned}
E[X] &= np \\
Var[X] &= np(1-p)
\end{aligned}
$$

<u>Example</u> *For the example of three coin spins, let $n = 3$ and $p = .5$. The number of heads X can take on values of 0, 1, 2, and 3.*

$$
f(0) = \frac{3!}{0!(3-0)!}(.5)^0(1-.5)^{3-0} = 1 \times 1 \times (.5)^3 = \frac{1}{8}
$$

Geometric

(predict number of times *n* to get heads once)

Construct the pdf by calculating probabilities of number of trials:

1 trial with probability p

2 trials with probability $(1-p)\,p$

n trials with probability $(1-p)^{n-1}\,p$

If N is a random variable with a geometric distribution, then the pdf, mean, and variance are

$$\begin{aligned} f(n) &= \Pr(N=n) = p\,(1-p)^{n-1} \\ E\,[N] &= \frac{1}{p} \\ Var\,[N] &= \frac{(1-p)}{p^2} \end{aligned}$$

The mean and variance both increase as p decreases. Unlike the other distributions, $f\,(0) \equiv 0$. There is always at least one outcome A.

Geometric Example

Example *What is the probability that first roll of a 5 will happen on the 4th roll?*

$$f(4) = \left(\frac{5}{6}\right)^3 \left(\frac{1}{6}\right) = .09645$$

Poisson

(approximation to binomial when *p* is small)

If X is a random variable with a Poisson distribution, then the pdf is

$$\begin{aligned} f(x) &= \lim_{\substack{n\to\infty \\ np=\mu}} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{e^{-np}(np)^x}{x!} \end{aligned}$$

Replace np with μ.

$$\begin{aligned} f(x) &= \frac{e^{-\mu}\mu^x}{x!} \\ \mu &= np \end{aligned}$$

Approximation to binomial when p is very small. Why? Because q≈ 1, so np≈ npq

The mean and variance both equal μ.

$$E[X] = \mu$$

$$Var[X] = \mu$$

Poisson model is rarely used, use negative binomial

Negative Binomial

How many trials until get outcome A the rth time? This is a natural generalization of the geometric distribution. It generalizes the geometric to r outcomes instead of one ($X = r$).

If N is a random variable and has a negative binomial distribution, then the pdf, mean, and variance are

$$
\begin{aligned}
f(n) &= \Pr(N = n) = \binom{n + r - 1}{n} p^r (1-p)^n \\
\mathrm{E}[N] &= \frac{r(1-p)}{p} \\
\mathrm{Var}[N] &= \frac{r(1-p)}{p^2}
\end{aligned}
$$

Negative Binomial Example

Example *A sumo wrestler at the level of Ōzeki is told that he must win 3 tournaments to be promoted to Yokozuna, the highest level (this is a gross simplification of what actually happens). What is the probability that a sumo wrestler is promoted to Yokozuna on the 5th try? Assume that $p = .2$.*

$$f(2) = \binom{2+3-1}{2} (.2)^3 (.8)^2 = 0.03072$$

Maximum Likelihood Estimation (MLE)

Least Squares Optimization.

Maximum Likelihood Estimation

Both are optimization procedures that involve searching for different model parameters.

Least squares optimization is an approach to estimating the parameters of a model by seeking a set of parameters that results in the smallest squared error between the predictions of the model (*yhat*) and the actual outputs (*y*), averaged over all examples in the dataset, so-called mean squared error.

Fat Albert

How would you test hypothesis that euro coins are not fair (i.e. $\Pr(H) \neq 0.5$)?

If spin coin 100 times, get *h* heads, then what is best guess at $\Pr(H)$?

Fundamental Idea of MLE

MLE= Best guess

Fundamental idea of MLE: Out of all possible parameter values, choose the value that is most likely to produce the observed data.

However, $\hat{p} = h/100$ is not the only possible answer because p is measured with uncertainty. But of all possible true values of p, $h/100$ is the most likely to produce the observed result.

Coin Spins: HT

<u>**Example**</u> *Suppose that after two spins of a coin, the outcomes are H and T. Consider the likelihood of this result given different values of p. For example, if $p = .1$, then the probability of HT is $.09 = .1 \times .9$. However, of all possible values of p, the value that is most likely to produce the observed data (HT) is $p = .5$ as shown in the following table.*

p	$(1-p)$	$p(1-p)$
.1	.9	.09
.2	.8	.16
.3	.7	.21
.4	.6	.24
.49	.51	.2499
.499	.501	.249999
.5	**.5**	**.25**

One head and one tail

100 Coin Spins: 45H, 55 T

In a more realistic example, consider spinning the coin 100 times. The binomial distribution gives the probability of observing a certain number of heads.

Example *Suppose that one hundred spins of a coin results in 45 heads. Consider the likelihood of this result given different values of p. Of all possible values of p, the value that is most likely to produce the observed data (HT) is $p = .45$. The table below and Figure 2.1 show both the probability of getting 45 heads in 100 spins, and the logarithm of that probability.*

Likelihood function

p	$\boxed{\binom{100}{45} p^{45}(1-p)^{55}}$	$\ln\left(\binom{100}{45} p^{45}(1-p)^{55}\right)$
.3	.00054873	−7.5079016
.4	.04781118	−3.0404958
.44	.07838320	−2.5461456
.45	**.07998750**	**−2.5258849**
.46	.07839174	−2.5460367
.5	.04847430	−3.0267216
.6	.00082912	−7.0951469

Log is a monotonic transformation in that the biggest value also has the biggest log

Graph of Likelihood Functions

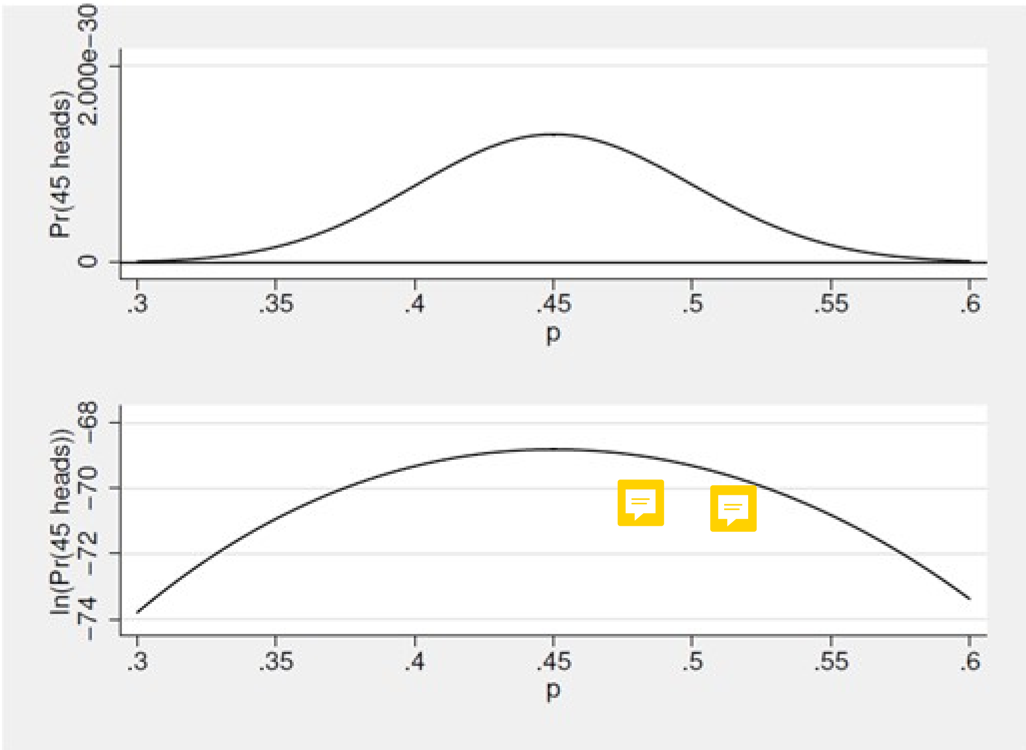


Figure 2.1: The likelihood and log likelihood functions of 45 heads after 100 spins, as a function of p.

Recipe for Solved Problem

1. Write the likelihood function L in terms of p
2. Take the logarithm of L, ($\ell = \ln(L)$)

   (why take log? Because in most cases, it makes step 3 much easier)
3. Take the derivative of ℓ with respect to p
4. Set derivative = 0, solve for p
5. Check second-order conditions

Binomial Example for Solved Problem

1

$$L = p^h (1-p)^t \tag{2.1}$$

2.

$$\ell = \ln(L) = h\ln(p) + t\ln(1-p)$$

3

$$\frac{d\ell}{dp} = \frac{h}{p} - \frac{t}{1-p}$$

4

$$\frac{d\ell}{dp} = 0 \quad \Rightarrow \quad h(1-p) = tp \quad \Rightarrow \quad \hat{p} = \frac{h}{h+t}$$

5.

$$\frac{d^2\ell}{dp^2} = -\frac{h}{p^2} - \frac{t}{(1-p)^2} < 0$$

Choose Function

What happened to the choose function?

Step 1 dropped the choose function

What would happen if you included the choose function? Think about Step 3.

→ if we don't take ln, then we have to add the choose part. Take the derivative and you will get the same thing.

Binomial Owls (1)

Example *A biologist observes how often an owl is successful in catching mice for its chicks. What is the probability p that an owl catches a mouse on any given attempt?*

Data	Number of mice caught given number of tries
Model	Binomial
Observations	$X = 15$ and $n = 60$
Unknown parameter	p

Binomial Owls (2)

$$
\begin{aligned}
L &= p^{15}\left(1-p\right)^{45} \\
\ell &= 15\ln\left(p\right)+45\ln\left(1-p\right) \\
\frac{d\ell}{dp} &= \frac{15}{p}+\frac{-45}{1-p}=0 \\
\hat{p} &= \frac{1}{4} \\
\frac{d^2\ell}{dp^2} &= -\frac{15}{p^2}-\frac{45}{\left(1-p\right)^2}<0
\end{aligned}
$$

Geometric Lottery (1)

Example *Five people purchase instant lottery tickets. Each person buys tickets until they purchase a winning instant lottery ticket. For example, the first person buys 6 losing tickets, then wins on the seventh. What is the probability p of purchasing a winning instant lottery ticket?*

Data	Number of purchases until get winning ticket
Model	Geometric (could also use binomial)
Observations	$N = 7, 12, 11, 22, 3$ and $x = 1$
Unknown parameter	p

Geometric Lottery (2)

$$
\begin{aligned}
L &= \prod_{i=1}^{5} (1-p)^{N_i - 1}\, p \\
&= \left((1-p)^6\, p\right)\left((1-p)^{11}\, p\right)\left((1-p)^{10}\, p\right)\left((1-p)^{21}\, p\right)\left((1-p)^2\, p\right) \\
&= (1-p)^{50}\, p^5 \\
\ell &= 50 \ln(1-p) + 5 \ln(p) \\
\frac{d\ell}{dp} &= \frac{-50}{1-p} + \frac{5}{p} = 0 \\
\hat{p} &= \frac{1}{11} \longrightarrow =5/55 \\
\frac{d^2\ell}{dp^2} &= -\frac{50}{(1-p)^2} - \frac{5}{p^2} < 0
\end{aligned}
$$

Poisson Carl Sagan (1)

Example *Carl Sagan used the Hubble telescope to search for life in other solar systems (n = 100 billion), and he finds 14 with life. He then used the Hubba-Hubble telescope to search for life in solar systems in five other galaxies of similar size and finds 2, 21, 0, 6, and 5 with life. What is the expected number of planets with life in a solar system, $\mu = np$?*

Data	Number of solar systems with life in each galaxy
Model	Poisson (could also use binomial)
Observations	$X = 14, 2, 21, 0, 6, 5$
Unknown parameter	$\mu = np$

Poisson Carl Sagan (2)

$$\begin{aligned}
L &= \prod_{i=1}^{6} \frac{e^{-\mu}\mu^{X_i}}{X_i!} = \frac{e^{-6\mu}\mu^{\sum X_i}}{\prod_{i=1}^{6} X_i!} \\
\ell &= -6\mu + \left(\sum_{i=1}^{6} X_i\right)\ln\mu - \sum_{i=1}^{6}\ln(X_i!) \\
\frac{d\ell}{d\mu} &= -6 + \frac{\boxed{\sum_{i=1}^{6} X_i}}{\mu} = 0 \quad \longrightarrow \quad =48 \\
\hat{\mu} &= \frac{\sum_{i=1}^{6} X_i}{6} = 8 \\
\hat{p} &= \frac{\hat{\mu}}{n} = \frac{8}{100 \text{ billion}} \\
\frac{d^2\ell}{d\mu^2} &= -\frac{\sum_{i=1}^{6} X_i}{\mu^2} < 0
\end{aligned}$$

Negative Binomial Salesman (1)

Example *A door-to-door salesman on commission must work until he makes four sales in a day. What is the probability p that the salesman makes a sale each visit?*

Data	Number of extra visits (no sale) each day for a week
Model	Negative binomial
Observations	$N = 14, 17, 13, 21, 15$
Unknown parameter	p

Negative Binomial Salesman (2)

$$
\begin{aligned}
L &= \prod_{i=1}^{5} p^4 (1-p)^{N_i} \\
&= \left(p^4 (1-p)^{14}\right) \left(p^4 (1-p)^{17}\right) \left(p^4 (1-p)^{13}\right) \left(p^4 (1-p)^{21}\right) \left(p^4 (1-p)^{15}\right) \\
&= p^{20} (1-p)^{80} \\
\ell &= 20 \ln(p) + 80 \ln(1-p) \\
\frac{d\ell}{dp} &= \frac{20}{p} - \frac{80}{1-p} = 0 \\
\hat{p} &= \frac{1}{5} \longrightarrow =(4*5)/(20+80)=20/100 \\
\frac{d^2\ell}{dp^2} &= -\frac{20}{p^2} - \frac{80}{(1-p)^2} < 0
\end{aligned}
$$

MLE: how does the computer do it? It makes educated guesses!

Starting values: 0 is a good place to start

Method of guessing: usually take derivatives, is it positive? If yes, then it means the slope is increasing → increase the value of the parameter.

Stopping rule: stop this iterative process when the change in parameters is small.

Model may not work (converge) when there are multiple peaks.

MLE Properties

Consistency: meaning if you have enough data, you will be close to the truth

Asymptotic normality: meaning you can use Normal theory for hypothesis testing (e.g. z-test)

Asymptotic efficiency: meaning smallest variance

Rules of thumb regarding sample size

- N > 500 = fine; N < 100 can be worrisome
  - Results aren't necessarily wrong if N<100;
  - But it is a possibility; and hard to know when problems crop up
- Plus ~10 cases per independent variable
- Eliason (1993) suggests minimum N~60 for up to 5 independent variables.