Week 3

Introduction to Likelihood Inference

University of New Orleans
POLI 6003
Rich Frank
September 13, 2012

Altman and Brehm (2003) takeaways

- Details are important.
- Software default settings can have an effect.
  - Always keep track of software and version used.
- Rounding/truncating
- Random number generators
  - Will become important when we discuss out-of-sample predictions of marginal effects.

- Even pros make mistakes.
- Ability to replicate is essential.

Distributions

- We have some uncertainty as to how the data were generated.

- We have to be clear as to what extent we think the sample represents the population as well as our uncertainty about the data generation process.

Linking DVs, distributions, and models

- King (1989) spends a chapter introducing a number of distributions that we will be seeing this semester.
  - Bernoulli
  - Binomial
  - Extended beta-binomial
  - Poisson
  - Negative binomial
  - Normal
  - Log-normal

Probability

- Probability and likelihood differ from one another principally by how they treat the data and model in relation to one another.

- Probability theory presumes some given model (or set of parameters) and seeks to estimate the data (given those parameters).

- As King (1989:9) puts it:

$$Y \sim f(y|\, \theta, \alpha)$$

- And

$$\theta = g(X, \beta)$$

- So our data, Y , has a probability distribution given by parameters θ and α, and θ is a function of some variables X and their parameters, β.

- All this comprise King’s model so the normal probability statement appears is:

$$Pr(y|M) \equiv Pr(\text{data}|\text{model})$$

- Some problems with using probability model:
  - 1. Presumes that data are random and unknown.
  - 2. Assumes that the model is known.

- What would be useful to have is the inverse probability:

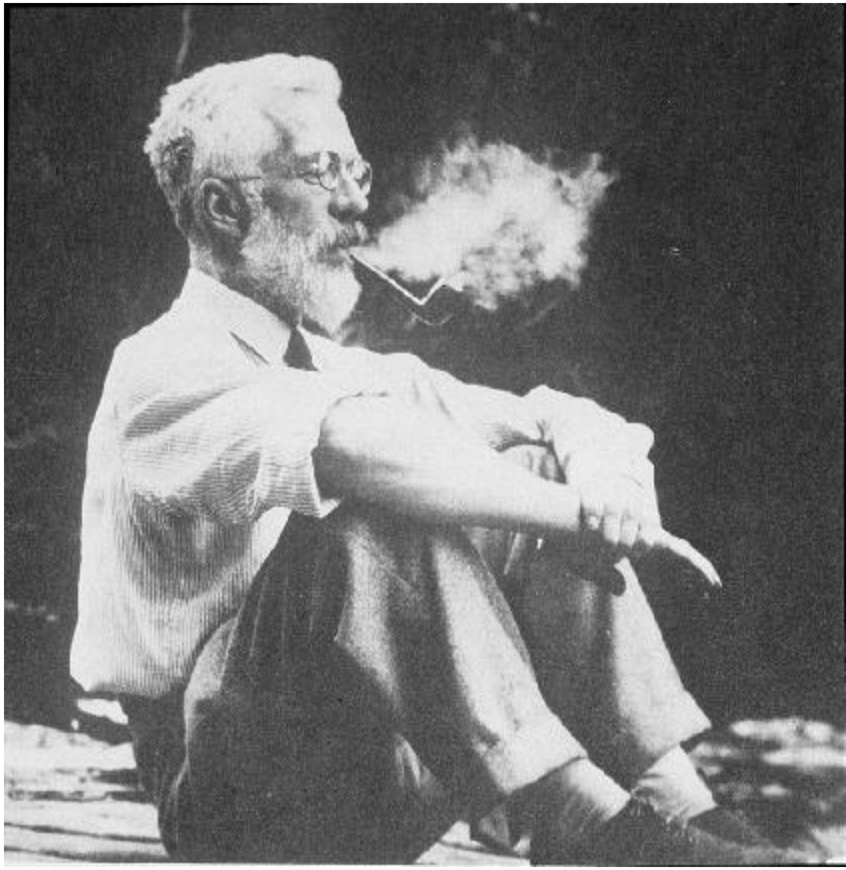
$$Pr(M \mid y)$$

- But that requires knowledge or strong assumptions about the important elements of the unknown model, θ.

- Bayesian methods make some (often weak) assumptions about the model given what we know about the world, but these are beyond the scope of this class.

- To a certain extent complexity theory also makes similar assumptions to generate data.

- As a result likelihood methods of inference have become popular.
- Attributed to the mathematician R.A. Fischer.
- Likelihood methods estimate the model given the data.



- Likelihood depends on the following axiom:

$$L(\tilde{\theta} \mid y, M^*) \equiv L(\tilde{\theta} \mid y)$$
$$= k(y)Pr(y|\tilde{\theta})$$
$$\alpha \;\; Pr(y|\tilde{\theta})$$

  - Where $\tilde{\theta}$ represents the hypothetical value of θ
  - $k(y)$ = the constant of proportionality, which is constant across all hypothetical $\tilde{\theta}$ but represents the way (the functional form) that the data shape $\tilde{\theta}$. This enables us to estimate likelihood as a measure of relative (rather than absolute) uncertainty.

- Thus likelihood is proportional to traditional probability where the constant *k(y)* is an unknown function of the data.

- The uncertainty is relative to other possible functions of y and the hypothetical values of $\tilde{\theta}$.

- Therefore, it measures the relative likelihood of a specific hypothetical $\tilde{\beta}$ producing the data we observe.

Examples

- Let's think about this a bit less theoretically.
- Okay, so the goal of likelihood is to estimate parameters given the data.
  - Consider the data fixed.
  - And consider a distribution of parameters $\tilde{\theta}$ we want to find the $\tilde{\theta}$ that is most likely to have generated the data we see.

Suppose we have a model where

$$Y \sim \mathrm{N}(\mu, \sigma^2)$$

$$\mathrm{E}(\mathrm{Y}) = \mu$$

$$Var(Y) = \sigma^2$$

- So the model presumes the data are normally distributed.
- We want to estimate the parameters μ and σ^2
- We want to find values of the mean and variance that are most likely to have produced the data, *Y*.

- Imagine that we have data on annual measures of presidential approval over an 8 year period.

$$Y = [\ 54\ \ 53\ \ 49\ \ 61\ \ 58\ \ 62\ \ 50\ \ 52]'$$

- We want to know the chances that the data are drawn from a distribution of mean μ and variance σ^2.
- What do you think is the likely mean of the distribution?
- Maximum likelihood is a more formal and systematic way of finding the parameters of the distribution most likely to have generated the data.

- If we assume that the data are normally distributed, then the PDF is given by:

$$Pr(Y = y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right]}$$

- So we can compute the probability of any particular observation in the distribution by solving the equation using that value from the data.

- However we are more interested in the *joint* probability of all of the 8 observations of presidential approval rather than just 1 from a distribution with a particular mean and variance.
- Assuming that the observations are independent of each other (a leap in this case) the joint PDF is equal to the product of the marginal probabilities.

$$Pr(A \text{ and } B) = Pr(\mathrm{A}) \cdot Pr(B)$$

- So the joint probability is given by:

$$Pr(Y) = y_i \forall i) = L(Y|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right]}$$

- This formula assumes the parameters are given while we want to estimate them.
- Fortunately the likelihood of the parameters is proportional to the probability of the data given the parameters.

$$L(\mu, \sigma^2 | \mathrm{Y}) = k(y) \prod_{i=1}^{n} f_{normal}(Y | \mu, \sigma^2)$$

$$\propto \prod_{i=1}^{n} f_{normal}(Y | \mu, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right]}$$

- This is the likelihood function, the constant k(y) ensures that the probability and the likelihood are proportional (∝).
- Now we can use this function to establish which values of the mean and variance are most likely to have generated the data.

- We can do this by hand (at least just this once).
- First, let's pick a value of μ, and for convenience set σ^2 to 1. (King tells us that this is the stylized Normal distribution—it relies on the independence of the mean and variance in the Normal distribution).

- Let $\mu = 53$.

$$L = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(54-53)^2}{2\sigma^2}\right]} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(53-53)^2}{2\sigma^2}\right]}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(49-53)^2}{2\sigma^2}\right]} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(61-53)^2}{2\sigma^2}\right]}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(58-53)^2}{2\sigma^2}\right]} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(62-53)^2}{2\sigma^2}\right]}$$

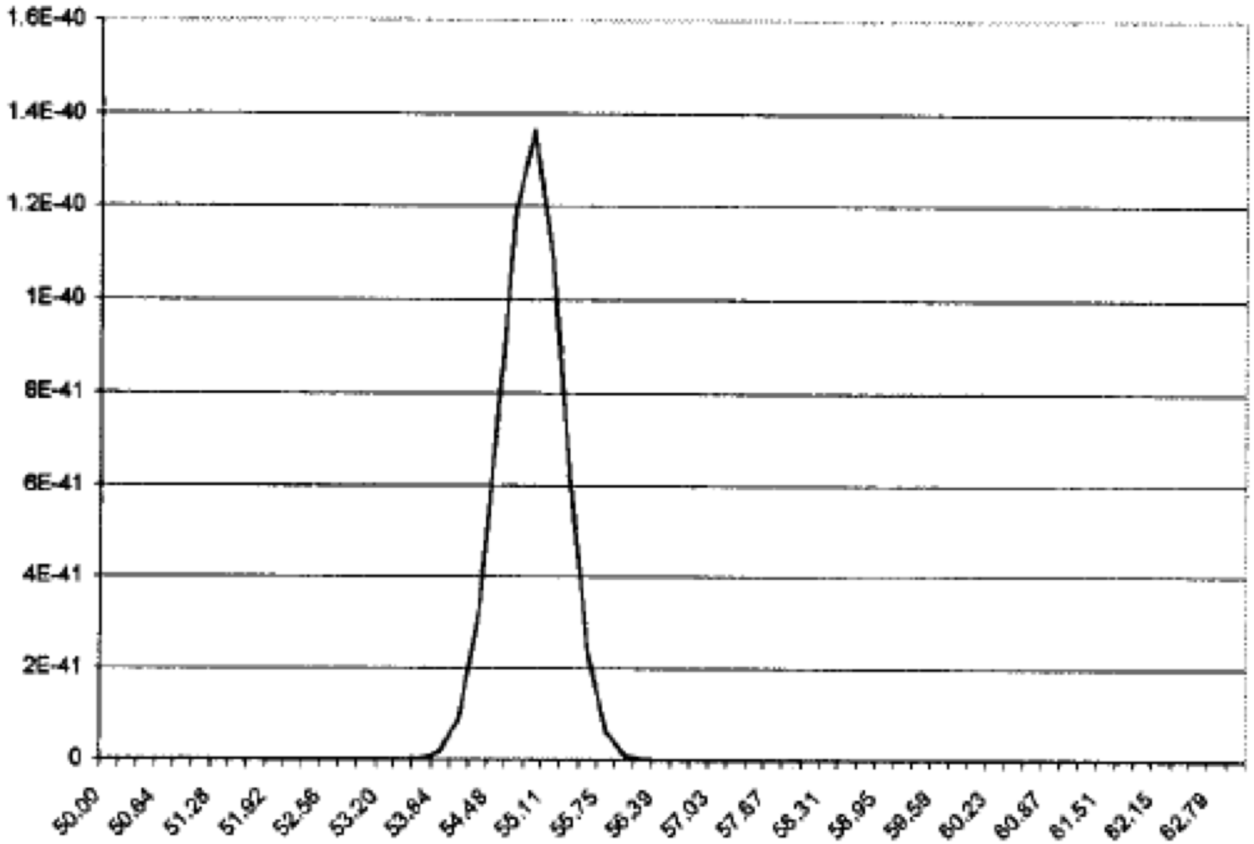
$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(50-53)^2}{2\sigma^2}\right]} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(52-53)^2}{2\sigma^2}\right]} \approx 1.752\text{e-}45$$

- That is a really small number. Thus we have little reason to think that 53 is the mean of the distribution that generated the data.

- So what we want to do is to do this same calculation for a number of different possible values of μ.

- The largest of the likelihoods will be the maximum of the likelihood function.

- The most efficient way to represent this is in a figure.



- Likelihood estimate of mean presidential approval

- From the figure it looks like the maximum is approximately 55. This is the ML estimate of μ.

- Another way of saying it is that the value of μ that maximizes the likelihood function is 55.

- As you can see this is a bit clunky to do, partly because products are harder to deal with than sums.

- Fortunately, we can transform the likelihood function above by any monotonic form (like the natural log).

$$ln\, L(\mu, \sigma^2 |\, \mathrm{Y}) = ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right]}$$

$$= \sum \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right]} \right]$$

$$= -\frac{1}{2}(\ln(2\pi)) - \frac{1}{2}(\ln(\sigma^2)) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - u_i)^2\right]$$

- Using this log-likelihood function we conduct similar calculations to what we did above, plot the estimates, and visually find the maximum.

- This method is similar to what numerical methods do without the graphics.

- Let’s try a different example.
- Suppose we have data on 20 states about whether they have adopted a lottery.

$$Y = [\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ ]'$$

- The probit likelihood function is given by:

$$ln\ L = \sum_{i=1}^{n} y_i \ln\big(\Phi(X\beta)\big) + (1 - y_i)\ln(1 - \Phi(X\beta))$$

- This is doable. For each of the 20 observations (y_i) we multiply y_i by the log of $\Phi(X\beta)$ and then add 1- y_i multiplied by the log of 1- $\Phi(X\beta)$.

- $\Phi(X\beta)$ would be the independent variables multiplied by their coefficients, summed (a z-score) and then evaluated on the normal CDF thus giving us a probability.

- However, we have no independent variables in this example.

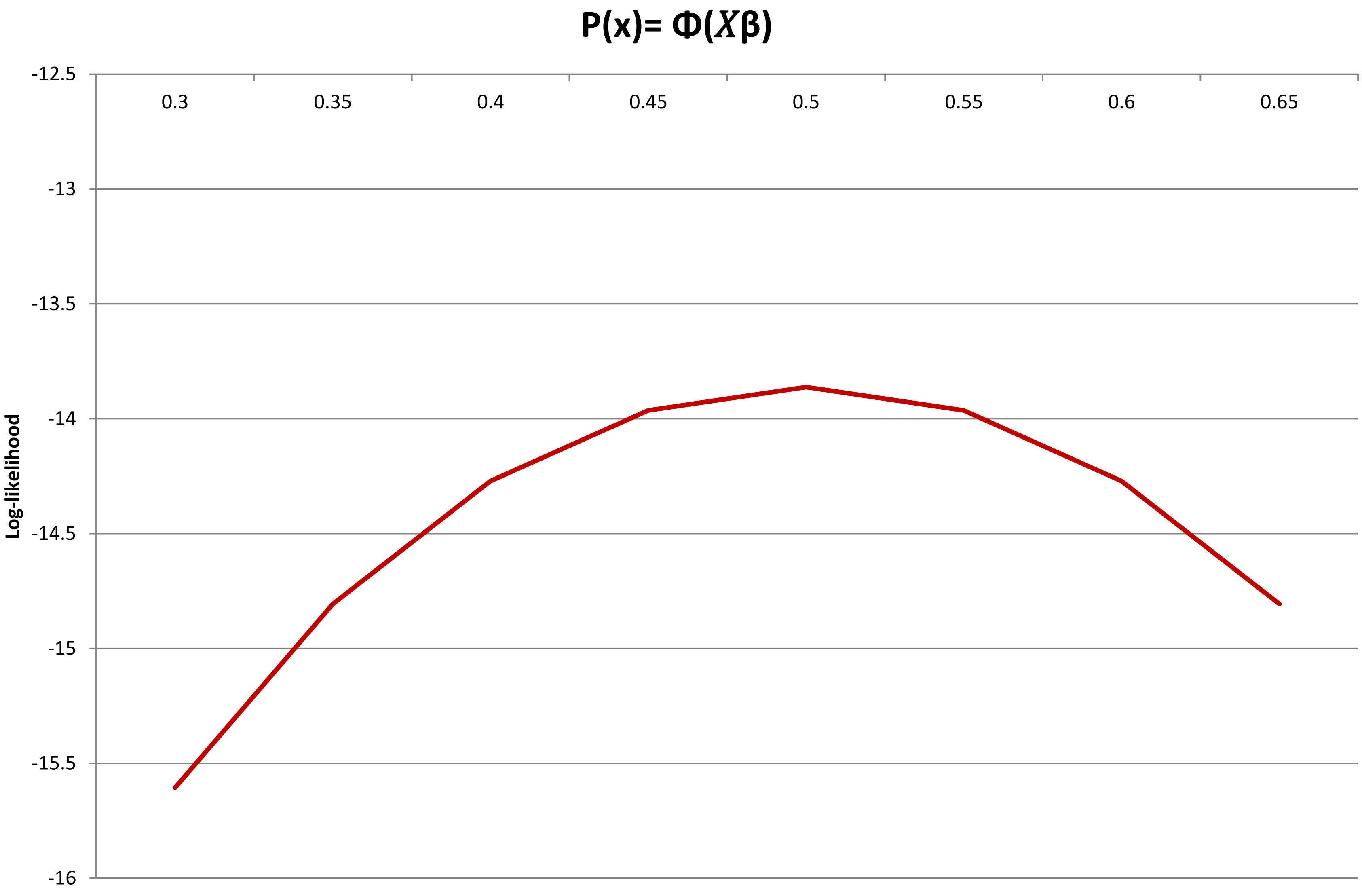
- Instead, like the above example we will test different values of probability as the mean probability responsible for generating the data on lottery adoption.

	p(z)=.3	p(z)=.35	p(z)=.4	p(z)=.45	p(z)=.5	p(z)=.55	p(z)=.6	p(z)=.65
Y	0.3	.35	.4	.45	.5	.55	.6	.65
0	-.356675	-0.430783	-.510826	-.597837	-.693147	-.798508	-.916291	-1.049822
0	-.356675	-0.430783	-.510826	-.597837	-.693147	-.798508	-.916291	-1.049822
0	-.356675	-0.430783	-.510826	-.597837	-.693147	-.798508	-.916291	-1.049822
0	-.356675	-0.430783	-.510826	-.597837	-.693147	-.798508	-.916291	-1.049822
0	-.356675	-0.430783	-.510826	-.597837	-.693147	-.798508	-.916291	-1.049822
0	-.356675	-0.430783	-.510826	-.597837	-.693147	-.798508	-.916291	-1.049822
0	-.356675	-0.430783	-.510826	-.597837	-.693147	-.798508	-.916291	-1.049822
0	-.356675	-0.430783	-.510826	-.597837	-.693147	-.798508	-.916291	-1.049822
0	-.356675	-0.430783	-.510826	-.597837	-.693147	-.798508	-.916291	-1.049822
0	-.356675	-0.430783	-.510826	-.597837	-.693147	-.798508	-.916291	-1.049822
1	-1.203973	-1.049822	-.916291	-.798508	-.693147	-.597837	-.510826	-.430783
1	-1.203973	-1.049822	-.916291	-.798508	-.693147	-.597837	-.510826	-.430783
1	-1.203973	-1.049822	-.916291	-.798508	-.693147	-.597837	-.510826	-.430783
1	-1.203973	-1.049822	-.916291	-.798508	-.693147	-.597837	-.510826	-.430783
1	-1.203973	-1.049822	-.916291	-.798508	-.693147	-.597837	-.510826	-.430783
1	-1.203973	-1.049822	-.916291	-.798508	-.693147	-.597837	-.510826	-.430783
1	-1.203973	-1.049822	-.916291	-.798508	-.693147	-.597837	-.510826	-.430783
1	-1.203973	-1.049822	-.916291	-.798508	-.693147	-.597837	-.510826	-.430783
1	-1.203973	-1.049822	-.916291	-.798508	-.693147	-.597837	-.510826	-.430783
1	-1.203973	-1.049822	-.916291	-.798508	-.693147	-.597837	-.510826	-.430783
SUM	**-15.60648**	**-14.80605**	**-14.27116**	**-13.96345**	**-13.86294**	**-13.96345**	**-14.27116**	**-14.80605**

For example for the first column

$\ln L = 0 * \ln(.3) + (1-0) * \ln(1-.3) +$
$\ln L = 0 * \ln(.3) + (1-0) * \ln(1-.3) +$
$\ln L = 0 * \ln(.3) + (1-0) * \ln(1-.3) +$
$\ln L = 0 * \ln(.3) + (1-0) * \ln(1-.3) +$
$\ln L = 0 * \ln(.3) + (1-0) * \ln(1-.3) +$
$\ln L = 0 * \ln(.3) + (1-0) * \ln(1-.3) +$
$\ln L = 0 * \ln(.3) + (1-0) * \ln(1-.3) +$
$\ln L = 0 * \ln(.3) + (1-0) * \ln(1-.3) +$
$\ln L = 0 * \ln(.3) + (1-0) * \ln(1-.3) +$
$\ln L = 0 * \ln(.3) + (1-0) * \ln(1-.3) +$
$\ln L = 1 * \ln(.3) + (1-1) * \ln(1-.3) +$
$\ln L = 1 * \ln(.3) + (1-1) * \ln(1-.3) +$
$\ln L = 1 * \ln(.3) + (1-1) * \ln(1-.3) +$
$\ln L = 1 * \ln(.3) + (1-1) * \ln(1-.3) +$
$\ln L = 1 * \ln(.3) + (1-1) * \ln(1-.3) +$
$\ln L = 1 * \ln(.3) + (1-1) * \ln(1-.3) +$
$\ln L = 1 * \ln(.3) + (1-1) * \ln(1-.3) +$
$\ln L = 1 * \ln(.3) + (1-1) * \ln(1-.3) +$
$\ln L = 1 * \ln(.3) + (1-1) * \ln(1-.3) +$
$\ln L = 1 * \ln(.3) + (1-1) * \ln(1-.3) = -15.60648$

Likelihood estimates of lottery adoption



- So the maximum is at .5.

- This should not be a surprise because out of the 20 observations there are 10 zeros and 10 ones.

Plug and chug iterative search

- So now we have a rough intuition as to what your software is doing when it is estimating an ML model (e.g. logit).
- The software takes a starting value of β (either zero or an OLS estimate) to estimate the log-likelihood (LL).
- It takes the first derivative of the LL with respect of the parameters to find the gradient.
- The gradient tells us the slope of a line tangent to the curve at the point of the LL estimate.
- If the gradient is positive then the L is increasing in β.
- It then increases the estimate of β and try again.
  - If the slope is negative it decreases the estimate.

- Once the first derivative is approaching zero, it stops and looks at the second derivative.
- If it is negative then it has reached a maximum.
- The flatter the slope the harder it can be to determine that we are at the top.
- The second derivative tells us how quickly the slope is changing, so we know how large of a step to take.

Properties of MLEs

- **Consistency**—They are asymptotically consistent. As sample size increases, the estimates increasingly resemble the actual population parameters. As a result MLEs are good large sample estimators (what is large?)

- **Asymptotic normalcy**—The MLE parameters are distributed according to the standard multivariate normal no mater what distribution assumptions you make in your model. This allows us to describe them using z-scores.

Properties of MLEs

- **Asymptotic efficiency**—basically this means that MLE has the smallest asymptotic variance of any estimators that are also consistent and asymptotically normal.

- **Invariance**—If Θ_{ML} is a vector of ML estimates, and $g(\Theta)$ is a continuous function of Θ, then $g(\Theta_{ML})$ is a consistent estimator of $g(\Theta)$. So if we transform variables, we can retransform the estimates without losing interpretive ability.

- Again, we can only speak about relative likelihood not absolute likelihood of our estimates given a set of data.

- Next week, we will begin to talk about likelihood inference applied to binary dependent variables.