General Linear Models

Statistical Inference

– O.Örsan Özener

Statistical inference

- **Statistical inference** aims at learning characteristics of the population from a sample; the population characteristics are *parameters* and sample characteristics are *statistics*.

Statistical model

- A **statistical model** is a representation of a complex phenomena that generated the data.
  - It has mathematical formulations that describe relationships between random variables and parameters.
  - It makes assumptions about the random variables, and sometimes parameters.
  - A general form: data = model + residuals
  - Model should explain most of the variation in the data
  - Residuals are a representation of a lack-of-fit, that is of the portion of the data unexplained by the model.

Real Life Data

- **Empirical problem:** Class size and educational output
  - Policy question: What is the effect on test scores (or some other outcome measure) of reducing class size by one student per class? by 8 students/class?
  - We must use data to find out (is there any way to answer this *without* data?)

The California Test Score Data Set

All K-6 and K-8 California school districts (*n* = 420)

Variables:

- 5th grade test scores (Stanford-9 achievement test, combined math and reading), district average
- Student-teacher ratio (STR) = no. of students in the district divided by no. full-time equivalent teachers

Initial look at the data:

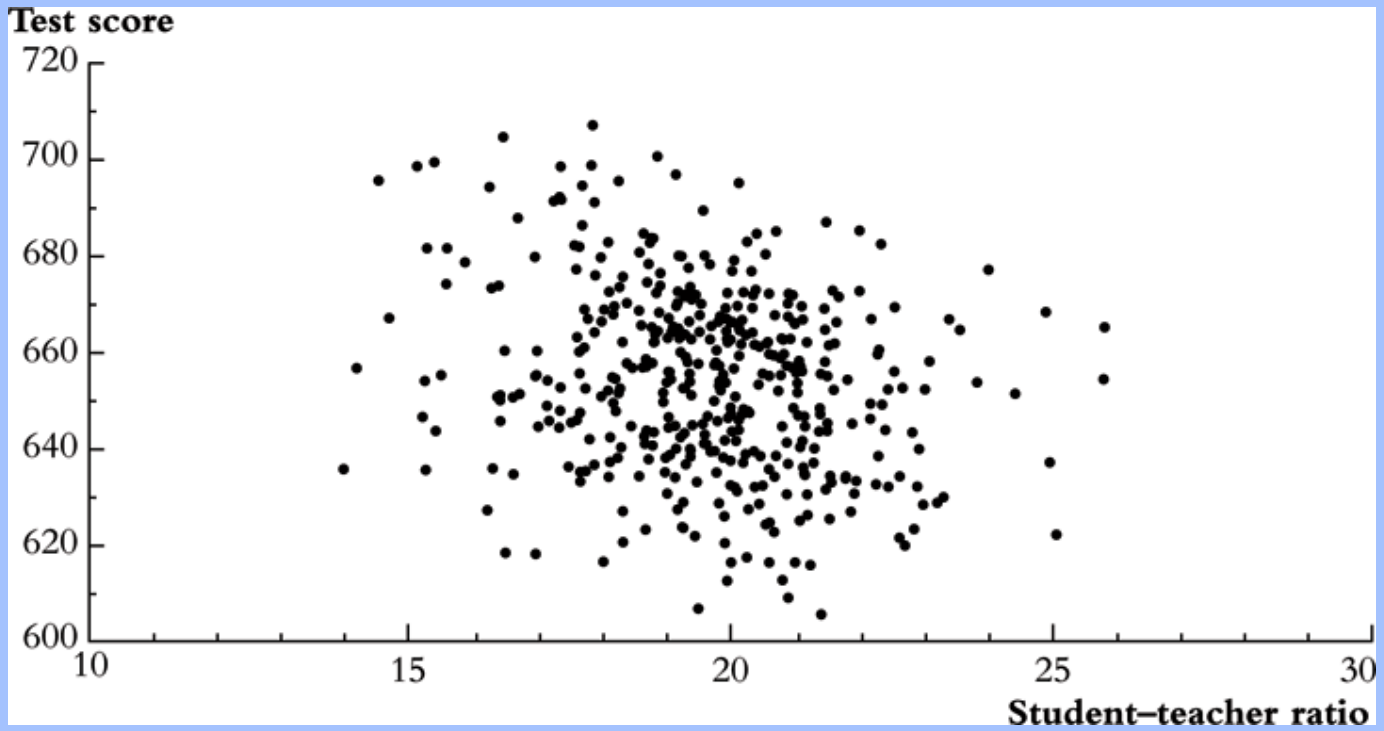
TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1999

			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	654.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

This table doesn't tell us anything about the relationship between test scores and the *STR*.

Do districts with smaller classes have higher test scores?

Scatterplot of test score v. student-teacher ratio



What does this figure show?

We need to get some numerical evidence on whether districts with low STRs have higher test scores – but how?

1. Compare average test scores in districts with low STRs to those with high STRs ("***estimation***")
2. Test the "null" hypothesis that the mean test scores in the two types of districts are the same, against the "alternative" hypothesis that they differ ("***hypothesis testing***")
3. Estimate an interval for the difference in the mean test scores, high v. low STR districts ("***confidence interval***")

Initial data analysis: Compare districts with "small" (STR < 20) and "large" (STR ≥ 20) class sizes:

Class Size	Average score ($\overline{Y}$)	Standard deviation (sB_{YB})	*n*
Small	657.4	19.4	238
Large	650.0	17.9	182

1. ***Estimation*** of Δ = difference between group means
2. ***Test the hypothesis*** that Δ = 0
3. Construct a ***confidence interval*** for Δ

This framework allows rigorous statistical inferences about moments of population distributions using a sample of data from that population …

1. **Estimation**
2. Testing
3. Confidence Intervals

Estimation

$\overline{Y}$ is the natural estimator of the mean. But:

a) What are the properties of $\overline{Y}$?

b) Why should we use $\overline{Y}$ rather than some other estimator?

- Y_{1} (the first observation)
- maybe unequal weights – not simple average
- median(Y_{1},…, Y_{n})

The starting point is the sampling distribution of $\overline{Y}$ …

The sampling distribution of $\overline{Y}$

$\overline{Y}$ is a random variable, and its properties are determined by the ***sampling distribution*** of $\overline{Y}$

- The individuals in the sample are drawn at random.
- Thus the values of ($Y\text{B}_{1\text{B}}$, …, $Y\text{B}_{n\text{B}}$) are random
- Thus functions of ($Y\text{B}_{1\text{B}}$, …, $Y\text{B}_{n\text{B}}$), such as $\overline{Y}$, are random: had a different sample been drawn, they would have taken on a different value
- The distribution of $\overline{Y}$ over different possible samples of size *n* is called the ***sampling distribution*** of $\overline{Y}$.
- The mean and variance of $\overline{Y}$ are the mean and variance of its sampling distribution, $E(\overline{Y})$ and $\text{var}(\overline{Y})$.

Things we want to know about the sampling distribution:

- What is the mean of $\bar{Y}$?
  - If *E*($\bar{Y}$) = true μ (mean),then $\bar{Y}$ is an ***unbiased*** estimator of μ
- What is the variance of $\bar{Y}$?
  - How does var($\bar{Y}$) depend on *n* (famous 1/*n* formula)
- Does $\bar{Y}$ become close to μ when *n* is large?
  - Law of large numbers: $\bar{Y}$ is a ***consistent*** estimator of μ
- $\bar{Y} - \mu$ appears bell shaped for *n* large…is this generally true?
  - In fact, $\bar{Y} - \mu$ is approximately normally distributed for *n* large (Central Limit Theorem)

Mean and variance of sampling distribution of $\overline{Y}$, ctd.

$$E(\overline{Y}) = \mu_Y$$

$$\text{var}(\overline{Y}) = \frac{s_Y^2}{n}$$

Implications:

1. $\overline{Y}$ is an *unbiased* estimator of μ_Y (that is, $E(\overline{Y}) = \mu_Y$)
2. var($\overline{Y}$) is inversely proportional to *n*
    1. the spread of the sampling distribution is proportional to $1/\sqrt{n}$
    2. Thus the sampling uncertainty associated with $\overline{Y}$ is proportional to $1/\sqrt{n}$ (larger samples, less uncertainty, but square-root law)

The sampling distribution of $\bar{Y}$ when *n* is large

For small sample sizes, the distribution of $\bar{Y}$ is complicated, but if *n* is large, the sampling distribution is simple!

1. As *n* increases, the distribution of $\bar{Y}$ becomes more tightly centered around μ_Y (the *Law of Large Numbers*)
2. Moreover, the distribution of

$$\frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}}$$

becomes normal (the *Central Limit Theorem*)

Why Use $\bar{Y}$ To Estimate μ_Y?

- $\bar{Y}$ is unbiased: $E(\bar{Y}_p) = \mu_Y$
- $\bar{Y}$ is consistent: $\bar{Y} \rightarrow \mu_Y$
- $\bar{Y}$ is the "least squares" estimator of μ_Y; $\bar{Y}$ solves,

$$\min_m \sum_{i=1}^{n} (Y_i - m)^2$$

so, $\bar{Y}$ minimizes the sum of squared "residuals"

optional derivation

$$\frac{d}{dm}\sum_{i=1}^{n}(Y_i - m)^2 = \sum_{i=1}^{n}\frac{d}{dm}(Y_i - m)^2 = 2\sum_{i=1}^{n}(Y_i - m)$$

Set derivative to zero and denote optimal value of m by $\hat{m}$:

$$\sum_{i=1}^{n} Y = \sum_{i=1}^{n}\hat{m} = n\hat{m} \text{ or } \hat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

Why Use $\overline{Y}$ To Estimate μ_Y, ctd.

- $\overline{Y}$ has a smaller variance than all other *linear unbiased* estimators: consider the estimator, $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, where $\{a_i\}$ are such that $\hat{\mu}_Y$ is unbiased; then var($\overline{Y}$) ≤ var($\hat{\mu}_Y$)

- $\overline{Y}$ isn't the only estimator of μ_Y – can you think of a time you might want to use the median instead?

1. Estimation

$$\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = \frac{1}{n_{\text{small}}} \sum_{i=1}^{n_{\text{small}}} Y_i - \frac{1}{n_{\text{large}}} \sum_{i=1}^{n_{\text{large}}} Y_i$$

= 657.4 – 650.0

= 7.4

Is this a large difference in a real-world sense?

- Standard deviation across districts = 19.1
- Difference between 60th and 75th percentiles of test score distribution is 667.6 – 659.4 = 8.2
- This is a big enough difference to be important for school reform discussions, for parents, or for a school committee?

Hypothesis Testing

The ***hypothesis testing*** problem (for the mean): make a provisional decision based on the evidence at hand whether a null hypothesis is true, or instead that some alternative hypothesis is true. That is, test

- H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) > \mu_{Y,0}$ (1-sided, >)
- H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) < \mu_{Y,0}$ (1-sided, <)
- H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) \neq \mu_{Y,0}$ (2-sided)

Some terminology for testing statistical hypotheses:

p-value = probability of drawing a statistic (e.g. $\overline{Y}$) at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true.

The ***significance level*** of a test is a pre-specified probability of incorrectly rejecting the null, when the null is true.

Calculating the p-value based on $\overline{Y}$:

p-value = $\Pr_{H_0}[|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{act} - \mu_{Y,0}|]$

Where $\overline{Y}^{act}$ is the value of $\overline{Y}$ actually observed (nonrandom)

Calculating the p-value, ctd.

- To compute the *p*-value, you need the to know the sampling distribution of $\bar{Y}$, which is complicated if *n* is small.
- If *n* is large, you can use the normal approximation (CLT):

p-value = $\Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$,

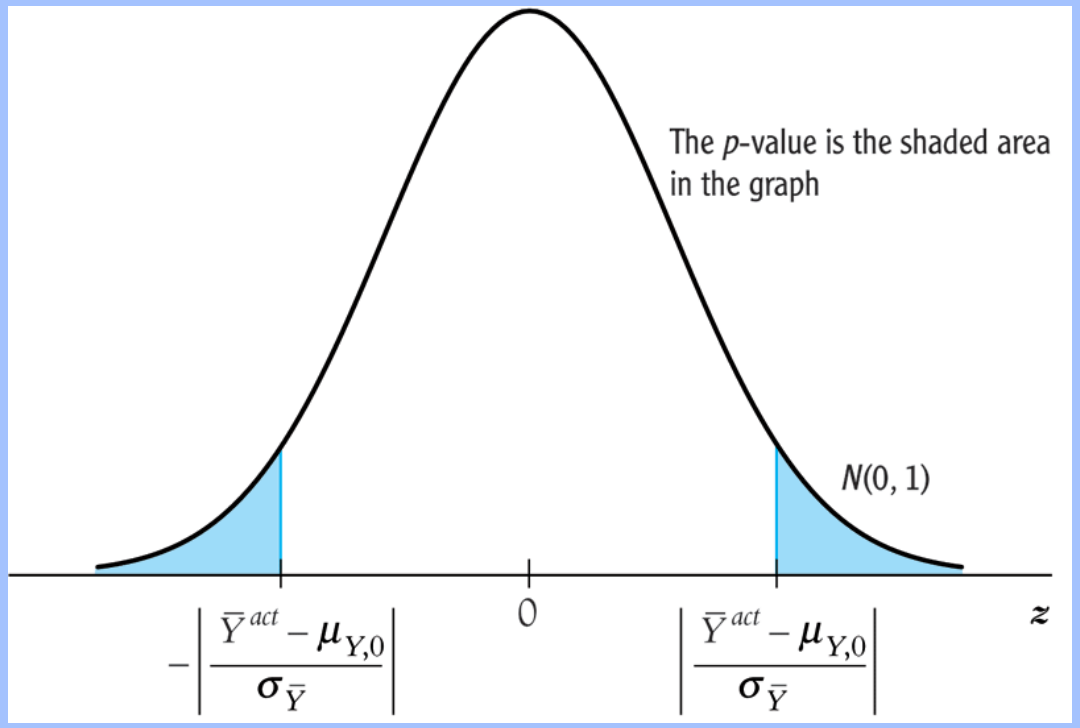
$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right|\right]$$

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{s_{\bar{Y}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_{\bar{Y}}}\right|\right]$$

$\cong$ probability under left+right $N(0,1)$ tails

where $s_{\bar{Y}}$ = std. dev. of the distribution of $\bar{Y} = \sigma_Y/\sqrt{n}$.

Calculating the p-value with σ_Y known:



- For large *n*, *p*-value = the probability that a *N*(0,1) random variable falls outside $|(\bar{Y}^{act} - \mu_{Y,0})/ s_{\bar{Y}}|$
- In practice, $s_{\bar{Y}}$ is unknown – it must be estimated

Estimator of the variance of Y:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 =$$ "sample variance of Y"

Fact:

If $(Y_1,\ldots,Y_n)$ are i.i.d. and $E(Y^4) < \infty$, then

$$s_Y^2 \xrightarrow{p} \sigma_Y^2$$

Why does the law of large numbers apply?

- Because s_Y^2 is a sample average;
- Technical note: we assume $E(Y^4) < \infty$ because here the average is not of Y_i, but of its square;

Computing the p-value with s_Y^2 estimated:

p-value = $\Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$,

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right|\right]$$

$$\cong \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right|\right] \quad \text{(large } n\text{)}$$

so

$$p\text{-value} = \Pr_{H_0}[|t| > |t^{act}|] \quad \textbf{(} s_Y^2 \textbf{estimated)}$$

$\cong$ probability under normal tails outside $|t^{act}|$

where $t = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}$ (the usual t-statistic)

What is the link between the *p*-value and the significance level?

- The significance level is prespecified. For example, if the prespecified significance level is 5%,
  - you reject the null hypothesis if $|t| \geq 1.96$.
  - Equivalently, you reject if $p \leq 0.05$.
  - The *p*-value is sometimes called the ***marginal significance level***.
  - Often, it is better to communicate the *p*-value than simply whether a test rejects or not – the *p*-value contains more information than the "yes/no" statement about whether the test rejects.

At this point, you might be wondering,...
What happened to the *t*-table and the degrees of freedom?

Digression: the Student *t* distribution

If Y_i, $i = 1,\ldots, n$ is i.i.d. $N(\mu_Y, s_Y^2)$, then the *t*-statistic has the Student *t*-distribution with $n - 1$ degrees of freedom.

The critical values of the Student *t*-distribution is tabulated in the back of all statistics books. Remember the recipe?

1. Compute the *t*-statistic
2. Compute the degrees of freedom, which is $n - 1$
3. Look up the 5% critical value
4. If the *t*-statistic exceeds (in absolute value) this critical value, reject the null hypothesis.

2. Hypothesis testing

Difference-in-means test: compute the *t*-statistic,

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$

(remember this?)

- where $SE(\bar{Y}_s - \bar{Y}_l)$ is the "standard error" of $\bar{Y}_s - \bar{Y}_l$, the subscripts *s* and *l* refer to "small" and "large" STR districts, and $s_s^2 = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(Y_i - \bar{Y}_s)^2$ (etc.)

Compute the difference-of-means *t*-statistic:

Size	$\bar{Y}$	$s_{\bar{Y}_B}$	*n*
small	657.4	19.4	238
large	650.0	17.9	182

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657.4 - 650.0}{\sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}}} = \frac{7.4}{1.83} = 4.05$$

|t| > 1.96, so reject (at the 5% significance level) the null hypothesis that the two means are the same.

Confidence Intervals

- A 95% ***confidence interval*** for μ_Y is an interval that contains the true value of μ_Y in 95% of repeated samples.
- *Digression*: What is random here? The values of $Y_1,...,Y_n$ and thus any functions of them – including the confidence interval. The confidence interval will differ from one sample to the next. The population parameter, μ_Y, is not random; we just don't know it.

Confidence intervals, ctd.

A 95% confidence interval can always be constructed as the set of values of μ_Y not rejected by a hypothesis test with a 5% significance level.

$$\{\mu_Y: \left|\frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}}\right| \leq 1.96\} = \{\mu_Y: -1.96 \leq \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \leq 1.96\}$$

$$= \{\mu_Y: -1.96 \leq -\mu_Y \leq 1.96 \frac{s_Y}{\sqrt{n}}\}$$

$$= \{\mu_Y \in (\bar{Y} - 1.96 \frac{s_Y}{\sqrt{n}}, \bar{Y} + 1.96 \frac{s_Y}{\sqrt{n}})\}$$

This confidence interval relies on the large-n results that $\bar{Y}$ *is approximately normally distributed and* $s_Y^2 \xrightarrow{p} s_Y^2$.

3. Confidence interval

A 95% confidence interval for the difference between the means is,

$$(\bar{Y}_s - \bar{Y}_l) \pm 1.96 \times SE(\bar{Y}_s - \bar{Y}_l)$$

$$= 7.4 \pm 1.96 \times 1.83 = (3.8, 11.0)$$

Two equivalent statements:

1. The 95% confidence interval for Δ doesn't include 0;
2. The hypothesis that Δ = 0 is rejected at the 5% level.

Summary:

- Theory of estimation (sampling distribution of)
- Theory of hypothesis testing (large-*n* distribution of *t*-statistic and computation of the *p*-value)
- Theory of confidence intervals (constructed by inverting the test statistic)

Promotion Strategies

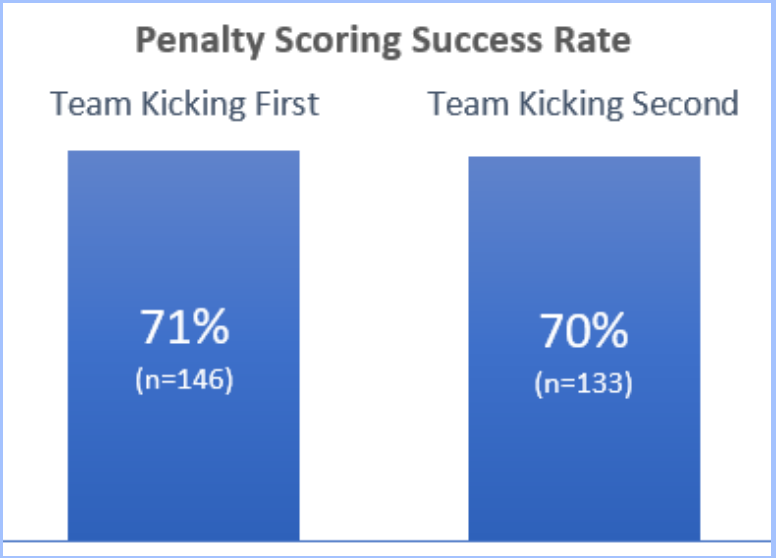
Penalty Kick

Loss Aversion: The psychological effect where people will work harder to avoid a perceived loss than for the equivalent gain.

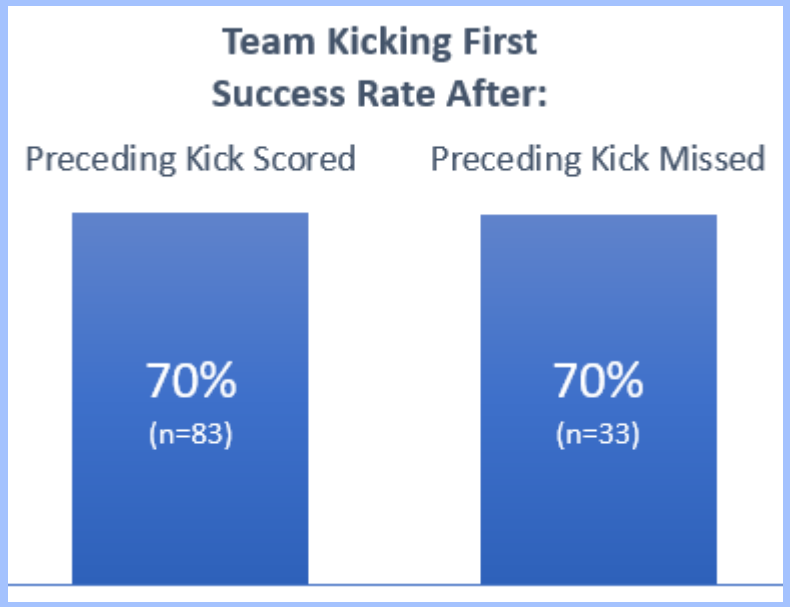
Observation: For the player kicking missing mean a negative result for that round. For the second kicker, if the first kicker scores then “par” would be the same for him, but if the first kicker missed scoring is the “birdie” as that would be a gain for the round.

Hypothesis: The team kicking first will be more likely to win given this effect as the second team might not try as hard based on the outcomes.

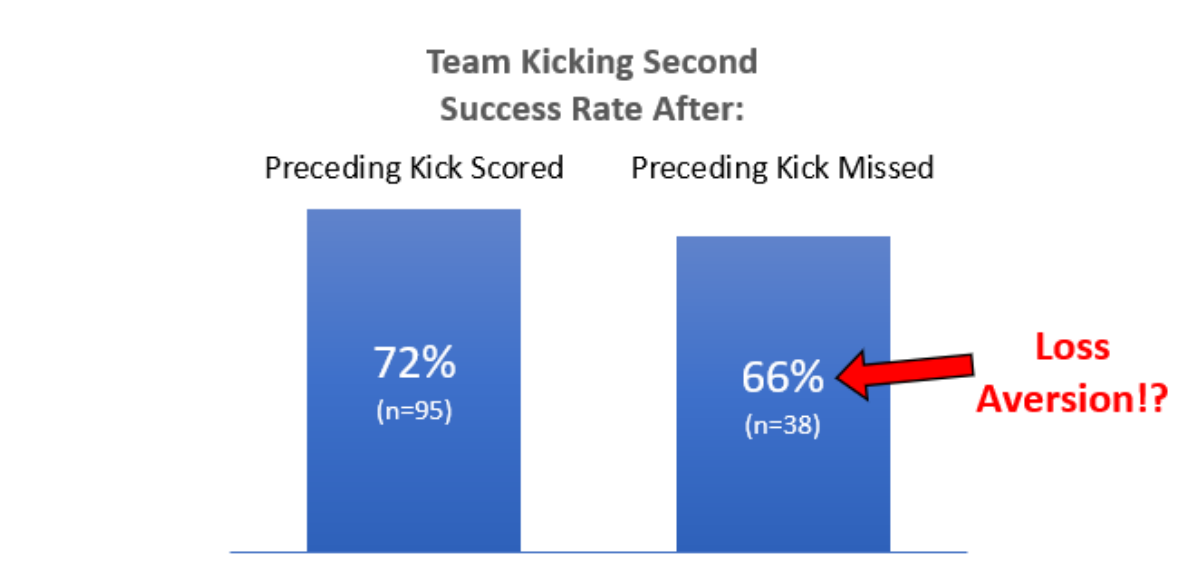
Penalty Kick - Data



Penalty Kick - Data



Penalty Kick - Data



The important question: is it a significant difference?

Penalty Kick – Conclusion

Null Hypothesis: The null hypothesis is that the first kick of a round doesn't affect the outcome of the second, the alternate hypothesis is that it does

P-value: P-value = 0.5107

Conclusion: Fail to reject the null hypothesis so it could be random noise!

Parameter Estimation

- How can I estimate model parameters from data?
- What should I worry about when choosing between estimators?
- Is there some optimal way of estimating parameters from data?
- How should I make statements about certainty regarding estimates and hypotheses?

Parameter Estimation

- We can formulate most questions in statistics in terms of making statements about underlying parameters
- We want to devise a framework for estimating those parameters and making statements about our certainty
- Several different approaches to making such statements
  - Moment estimators
  - Likelihood
  - Bayesian estimation

Moment Estimation

- One way of estimating parameter values – moment methods
- In such techniques parameter values are found that match sample moments (mean, variance, etc.) to those expected
- E.g. for random variables X_1, X_2, etc. sampled from a $N(\mu,\sigma^2)$ distribution

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \longrightarrow \hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{n}{n-1}s^2$$

$$E(\bar{X}) = \mu$$

$$E(s^2) = \frac{n-1}{n}\sigma^2$$

Example: fitting a gamma distribution

- The gamma distribution is parameterized by a shape parameter, α, and a scale parameter, β

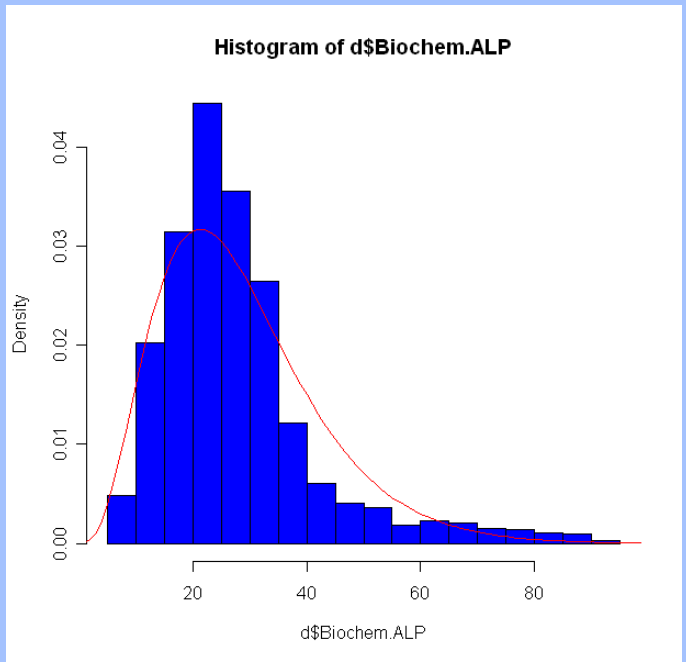
$$f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

- The mean of the distribution is α/β and the variance is α/β^2

- We can fit a gamma distribution by looking at the first two sample moments

$$\hat{\beta} = \frac{\overline{X}}{\frac{1}{n-1}\sum_i (X_i - \overline{X})^2}$$

$$\hat{\alpha} = \overline{X}\hat{\beta}$$



Alkaline phosphatase measurements in 2019 mice

$\alpha = 4.03$

$\beta = 0.14$

Bias

- Although the moment method looks sensible, it can lead to biased estimators

- Bias is measured by the difference between the expected estimate and the truth

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- However, bias is not the only thing to worry about
  - For example, the value of the first observation is an unbiased estimator of the mean for a Normal distribution. However it is a rubbish estimator

- We also need to worry about the variance of an estimator

The bias-variance trade off

- Some estimators may be biased
- Some estimators may have large variance
- Which is better?
- A simple way of combining both metrics is to consider the mean-squared error of an estimator $\hat{\theta}$

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \left[E(\hat{\theta} - \theta)\right]^2$$

Example

- Consider two ways of estimating the variance of a Normal distribution from the sample variance

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

$$\hat{\sigma}_A^2 = s^2$$

$$\hat{\sigma}_B^2 = \frac{n}{n-1}s^2$$

- The second estimator is unbiased, but the first estimator has lower MSE
- Actually, there is a third estimator, which is even more biased than the first, but which has even lower MSE

$$\hat{\sigma}_C^2 = \frac{n}{n+1}s^2$$

Least squares estimation

- A commonly-used approach to fitting models to data is called **least squares estimation**

- This attempts to minimize the sum of the squares of residuals
  - A residual is the difference between an observed and a fitted value

- An important point to remember is that minimizing LS is not the only thing to worry about when fitting model
  - Over-fitting

Is there an optimal way to estimate parameters?

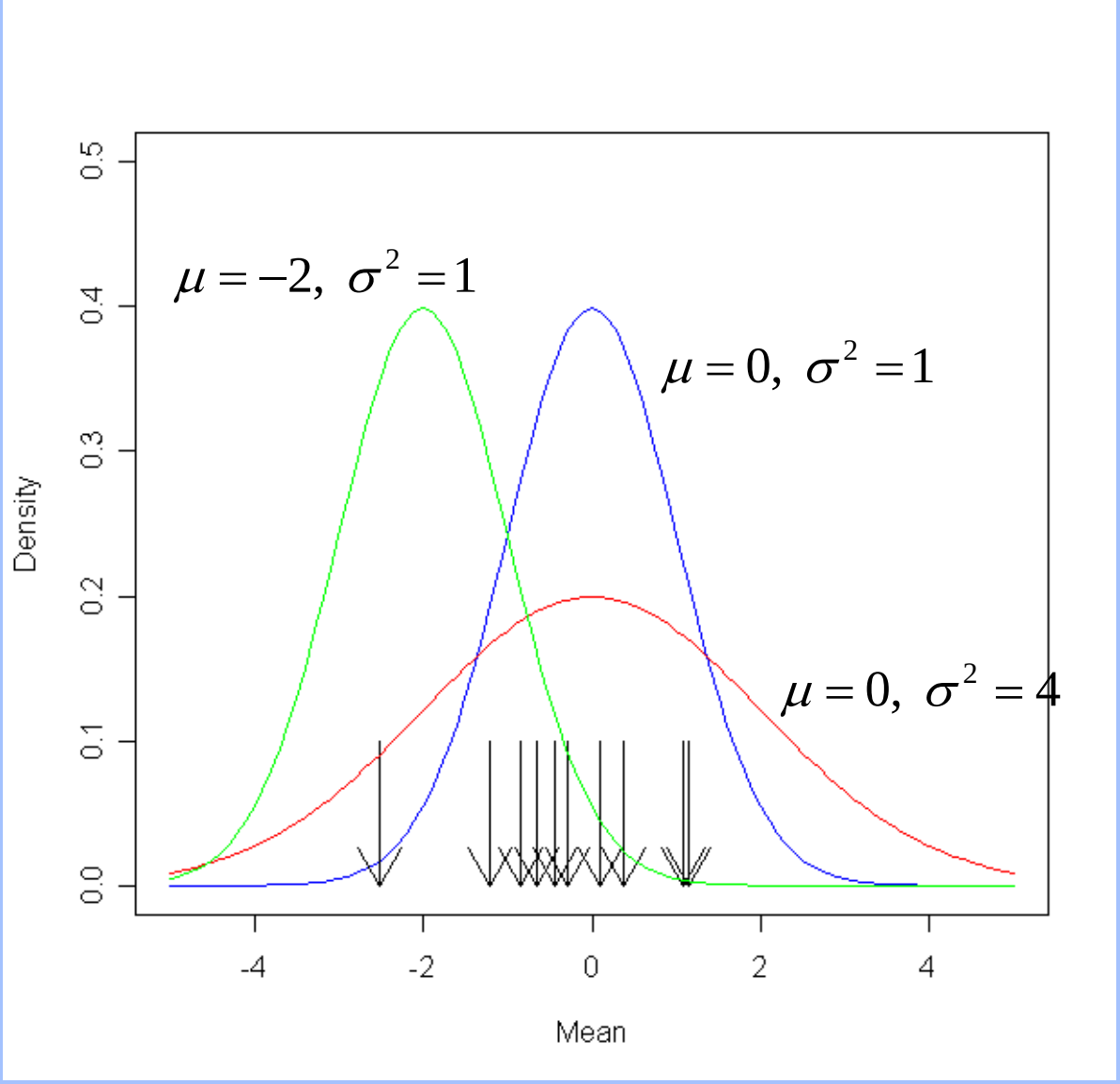
- For any model the maximum information about model parameters is obtained by considering the **likelihood** function

- The **likelihood function** is proportional to the probability of observing the data given a specified parameter value

- One natural choice for point estimation of parameters is the **maximum likelihood estimate**, the parameter values that maximize the probability of observing the data

- The maximum likelihood estimate (mle) has some useful properties (though is not always optimal in every sense)

An intuitive view on likelihood



An example

- Suppose we have data generated from a Poisson distribution. We want to estimate the parameter of the distribution

- The probability of observing a particular random variable is

$$P(X;\mu) = \frac{e^{-\mu}\mu^{X}}{X!}$$

- If we have observed a series of iid Poisson RVs we obtain the joint likelihood by multiplying the individual probabilities together

$$P(X_1, X_2, \ldots, X_n;\mu) = \frac{e^{-\mu}\mu^{X_1}}{X_1!} \times \frac{e^{-\mu}\mu^{X_2}}{X_2!} \times \ldots \times \frac{e^{-\mu}\mu^{X_n}}{X_n!}$$

$$L(\mu;\mathbf{X}) = \prod_i e^{-\mu}\mu^{X_i}$$

$$L(\mu;\mathbf{X}) = e^{-n\mu}\mu^{n\bar{X}}$$

Comments

- Note in the likelihood function the factorials have disappeared. This is because they provide a constant that does not influence the relative likelihood of different values of the parameter

- It is usual to work with the **log likelihood** rather than the likelihood. Note that maximizing the log likelihood is equivalent to maximizing the likelihood

- We can find the **mle** of the parameter analytically

$$L(\mu; \mathbf{X}) = e^{-n\mu} \mu^{n\bar{X}}$$

$$\ell(\mu; \mathbf{X}) = -n\mu + n\bar{X} \log \mu$$

Take the natural log of the likelihood function

$$\frac{d\ell}{d\mu} = -n + \frac{n\bar{X}}{\mu}$$

Find where the derivative of the log likelihood is zero

$$\hat{\mu} = \bar{X}$$

Note that here the mle is the same as the moment estimator

Properties of the maximum likelihood estimate

- The maximum likelihood estimate can be found either analytically or by numerical maximization

- The mle is **consistent** in that it converges to the truth as the sample size gets infinitely large

- The mle is **asymptotically efficient** in that it achieves the minimum possible variance (the Cramér-Rao Lower Bound) as $n \rightarrow \infty$

- However, the mle is often **biased** for finite sample sizes
  - For example, the mle for the variance parameter in a normal distribution is the sample variance

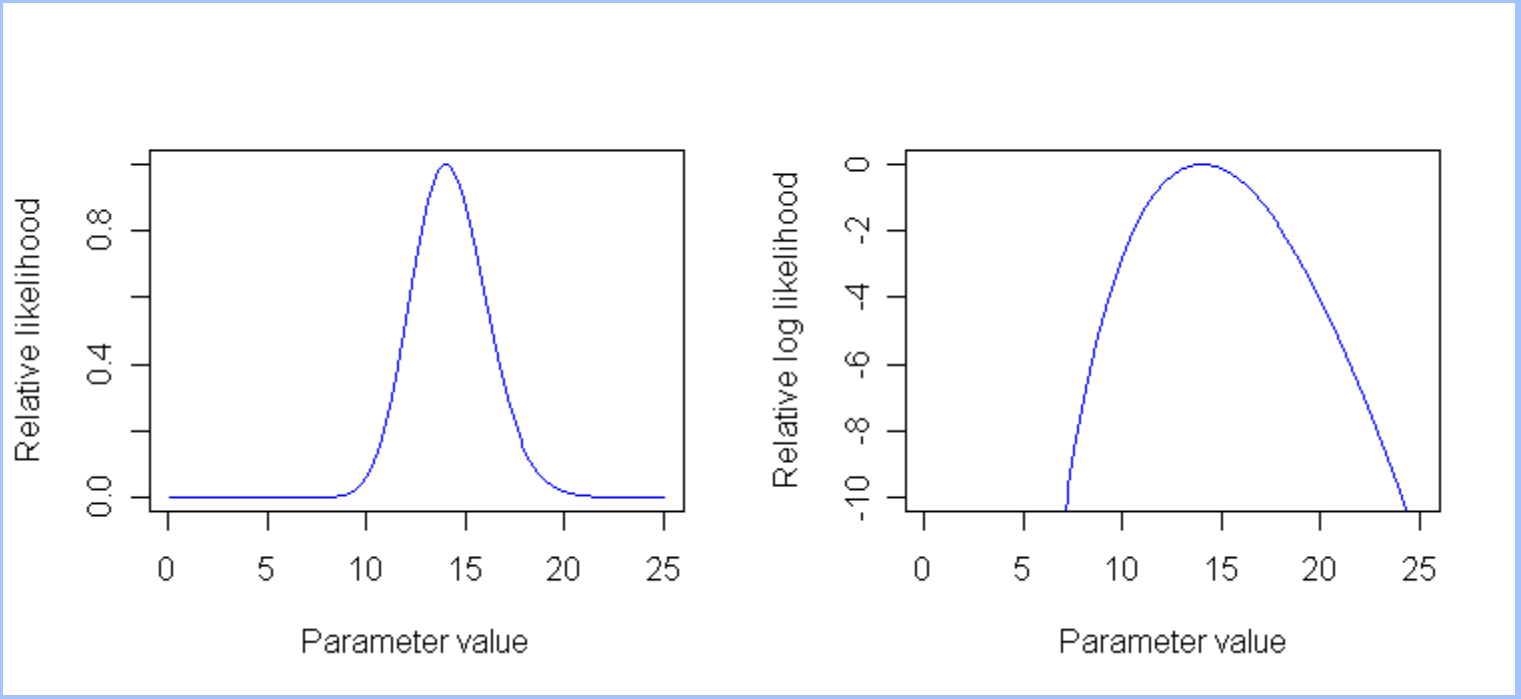
Comparing parameter estimates

- Obtaining a point estimate of a parameter is just one problem in statistical inference
- We might also like to ask how good different parameter values are
- One way of comparing parameters is through **relative likelihood**
- For example, suppose we observe counts of 12, 22, 14 and 8 from a Poisson process
- The maximum likelihood estimate is 14. The relative likelihood is given by

$$\frac{L(\mu;\mathbf{X})}{L(\hat{\mu};\mathbf{X})} = e^{-n(\mu-\hat{\mu})}\left(\frac{\mu}{\hat{\mu}}\right)^{n\bar{X}}$$

Using relative likelihood

- The relative likelihood and log likelihood surfaces are shown below



Example 1

- A sample of ten new bike helmets manufactured by a certain company is obtained. Upon testing, it is found that the first, third, and tenth helmets are flawed, whereas the others are not.

- Let $p = P$(flawed helmet), i.e., p is the proportion of all such helmets that are flawed.

- Define (Bernoulli) random variables $X_1, X_2, \ldots, X_{10}$ by

$$X_1 = \begin{cases} 1 \text{ if 1st helmet is flawed} \\ 0 \text{ if 1st helmet isn't flawed} \end{cases} \quad \cdots \quad X_{10} = \begin{cases} 1 \text{ if 10th helmet is flawed} \\ 0 \text{ if 10th helmet isn't flawed} \end{cases}$$

Example 1

- Then for the obtained sample, $X_1 = X_3 = X_{10} = 1$ and the other seven X_i's are all zero.

- The probability mass function of any particular X_i is $p^{x_i}(1-p)^{1-x_i}$, which becomes p if $x_i = 1$ and $1 - p$ when $x_i = 0$.

- Now suppose that the conditions of various helmets are independent of one another.

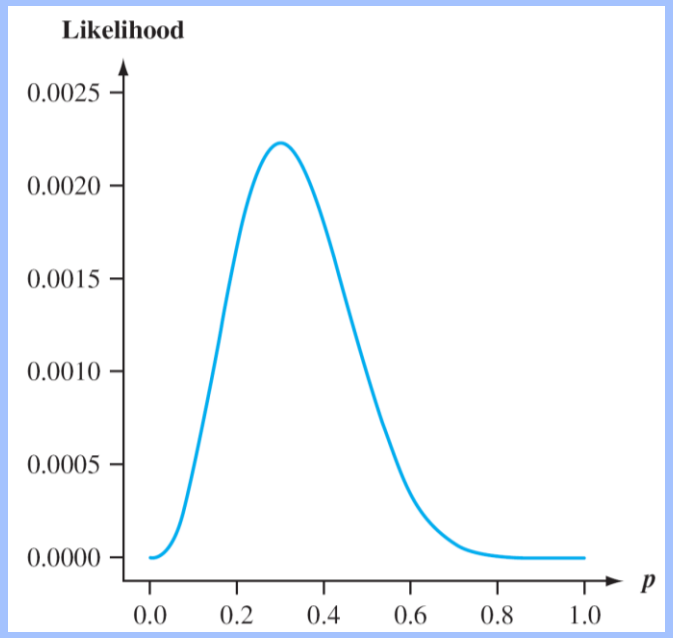
- This implies that the X_i's are independent, so their joint probability mass function is the product of the individual pmf's.

Example 1

- Thus the joint pmf evaluated at the observed X_i's is
- $$f(x_1, \ldots, x_{10}; p) = p(1-p)p \cdots p = p^3(1-p)^7$$
- Suppose that $p = .25$. Then the probability of observing the sample that we actually obtained is $(.25)^3(.75)^7 = .002086$.
- If instead $p = .50$, then this probability is $(.50)^3(.50)^7 = .000977$. For what value of p is the obtained sample most likely to have occurred? That is, for what value of p is the joint pmf as large as it can be? What value of p maximizes the joint pmf?

Example 1

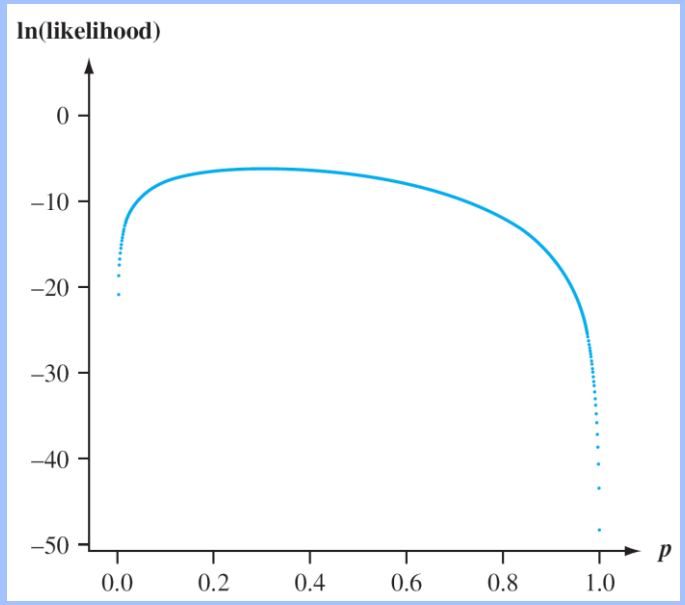
- Figure below shows a graph of the *likelihood* as a function of *p*. It appears that the graph reaches its peak above *p* = .3 = the proportion of flawed helmets in the sample.



Graph of the likelihood (joint pmf)

Example 1

- Figure below shows a graph of the natural logarithm of likelihood function; since *ln*[g(*u*)] is a strictly increasing function of *g*(*u*), finding *u* to maximize the function *g*(*u*) is the same as finding *u* to maximize *ln*[g(*u*)].



Graph of the natural logarithm of the likelihood

Example 1

We can verify our visual impression by using calculus to find the value of p that maximizes.

Working with the natural log of the joint pmf is often easier than working with the joint pmf itself, since the joint pmf is typically a product so its logarithm will be a sum.

Here

$$\ln[\, f\,(x_1, \ldots, x_{10}; p)] = \ln[p^3(1-p)^7] = 3\ln(p) + 7\ln(1-p)$$

Example 1

Thus

$$\frac{d}{dp}\{\ln[f(x_1, \ldots, x_{10}; p)]\} = \frac{d}{dp}\{3\ln(p) + 7\ln(1 - p)\}$$

$$= \frac{3}{p} + \frac{7}{1 - p}(-1)$$

$$= \frac{3}{p} - \frac{7}{1 - p}$$

[the (-1) comes from the chain rule in calculus].

Example 1

Equating this derivative to 0 and solving for p gives $3(1 - p) = 7p$, from which $3 = 10p$ and so $p = 3/10 = .30$ as conjectured.

That is, our point estimate is $\hat{p} = .30$. It is called the *maximum likelihood estimate* because it is the parameter value that maximizes the likelihood (joint pmf) of the observed sample.

In general, the second derivative should be examined to make sure a maximum has been obtained, but here this is obvious from the figure before.

Example 1

Suppose that rather than being told the condition of every helmet, we had only been informed that three of the ten were flawed.

Then we would have the observed value of a binomial random variable X = the number of flawed helmets.

The pmf of X is $\binom{10}{x} p^x (1 - p)^{10-x}$. For x = 3, this becomes $\binom{10}{3} p^3 (1 - p)^7$. The binomial coefficient $\binom{10}{3}$ is irrelevant to the maximization, so again $\hat{p}$ = .30.

Maximum Likelihood Estimation

The likelihood function tells us how likely the observed sample is as a function of the possible parameter values.

Maximizing the likelihood gives the parameter values for which the observed sample is most likely to have been generated—that is, the parameter values that “agree most closely” with the observed data.

Example 2

- Suppose $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ. Because of independence, the likelihood function is a product of the individual pdf's:

$$f(x_1, \ldots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdot \ldots \cdot (\lambda e^{-\lambda x_n})$$

$$= \lambda^n e^{-\lambda \Sigma x_i}$$

- The natural logarithm of the likelihood function is

$$\ln[\, f\,(x_1, \ldots, x_n\,;\, \lambda)] = n \ln(\lambda) - \lambda \Sigma x_i$$

Example 2

- Equating $(d/d\lambda)$[ln(likelihood)] to zero results in $n/\lambda - \Sigma x_i = 0$, or $\lambda = n/\Sigma x_i = 1/\overline{x}.$

- Thus the mle is $\hat{\lambda} = 1/\overline{X};$ it is identical to the method of moments estimator

- It is not an unbiased estimator, since

$$E(1/\overline{X}) \neq 1/E(\overline{X})].$$

Example 3

- Let $X_1, \ldots, X_n$ be a random sample from a normal distribution. The likelihood function is

$$f(x_1, \ldots, x_n; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/(2\sigma^2)}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\Sigma(x_i-\mu)^2/(2\sigma^2)}$$

- so

$$\ln[f(x_1, \ldots, x_n; \mu, \sigma^2)] = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \Sigma(x_i - \mu)^2$$

Example 3

- To find the maximizing values of μ and σ^2, we must take the partial derivatives of ln(f) with respect to μ and σ^2, equate them to zero, and solve the resulting two equations.

- Omitting the details, the resulting mle's are

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{\sum (X_i - \overline{X})^2}{n}$$

- The mle of σ^2 is not the unbiased estimator, so two different principles of estimation (unbiasedness and maximum likelihood) yield two different estimators.

Newton-Raphson Algorithm

- Let x_0 be a good estimate of r and let $r = x_0 + h$. Since the true root is r, and $h = r - x_0$, the number h measures how far the estimate x_0 is from the truth.
- Since h is 'small,' we can use the linear (tangent line) approximation to conclude that

$$0 = f(r) = f(x_0 + h) \approx f(x_0) + hf'(x_0),$$

- Therefore, unless $f'(x_0)$ is close to 0,

$$h \approx - f(x_0)/f'(x_0) \ .$$

Newton-Raphson Algorithm

- It follows that

$$r = x_0 + h \approx x_0 - f(x_0)/f'(x_0)$$

- Our new improved estimate x_1 of r is therefore given by

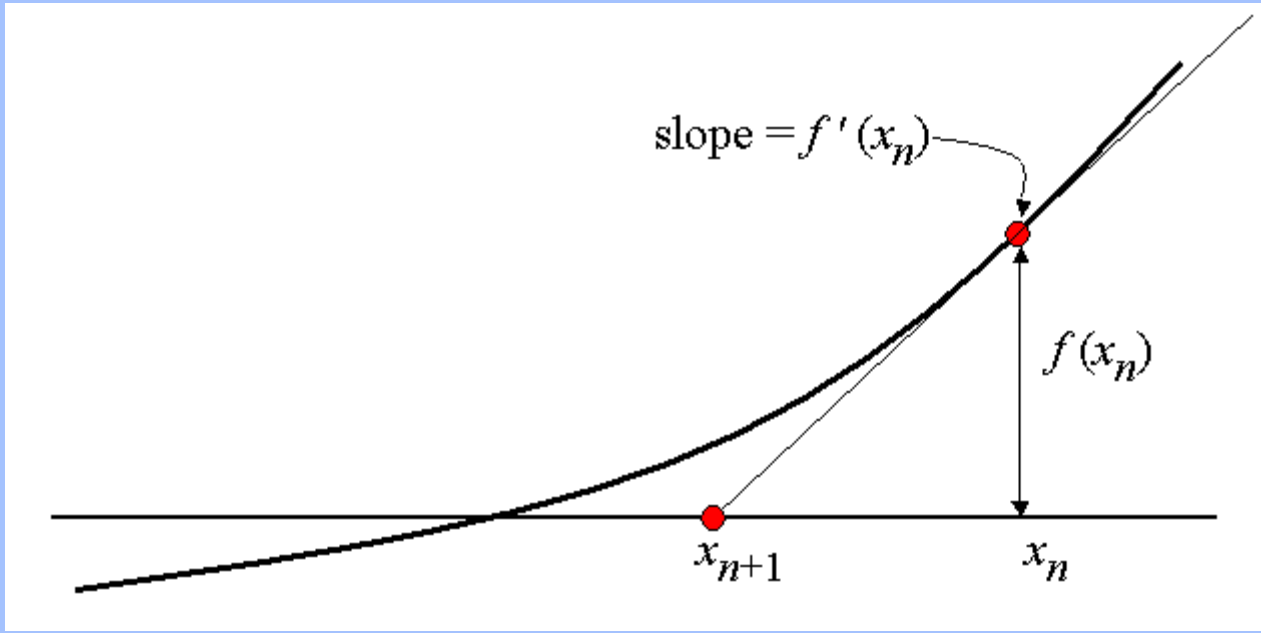
$$x_1 = x_0 - f(x_0)/f'(x_0)$$

- Continue in this way. If x_n is the current estimate, then the next estimate x_{n+1} is given by

$$x_{n+1} = x_n - f(x_n)/f'(x_n)$$

Newton-Raphson Algorithm

- Geometrically,



Calculating the MLE of a Binomial Sampling Model

- Our log-likelihood function is:

$$l(\pi|y) = yln(\pi) + (\mathrm{n} - y)ln(1-\pi)$$

where n is the sample size, y is the number of successes, and π is the probability of a success. The first derivative of the log-likelihood function is

$$l'(\pi|y) = \frac{y}{\pi} + (\mathrm{n} - y)\frac{-1}{1-\pi}$$

and the second derivative of the log-likelihood function is

$$l''(\pi|y) = -\frac{y}{\pi^2} - \frac{n-y}{(1-\pi)^2}$$

Calculating the MLE of a Binomial Sampling Model

- Analytically, we know that the MLE is $\hat{\pi} = \frac{y}{n}$.
- For the sake of example, suppose n = 5 and y = 2. Analytically, we know that the MLE = 0.4. Let's see how the Newton Raphson algorithm works in this situation.
- We begin by setting a tolerance level to 0.01 Next we make an initial guess as to the MLE. Suppose π_0 = 0.55, then $l'(\pi_0|y) \approx -3.03$ which is larger in absolute value than our tolerance of 0.01. Thus we set

$$\pi_1 \leftarrow \pi_0 - \frac{l'(\pi_0|y)}{l''(\pi_0|y)} \approx 0.40857$$

Calculating the MLE of a Binomial Sampling Model

- Now we calculate $l'(\pi_1|y) \approx -0.1774$ which is still larger in absolute value than our tolerance of 0.01. Thus we set

$$\pi_2 \leftarrow \pi_1 - \frac{l'(\pi_1|y)}{l''(\pi_1|y)} \approx 0.39994$$

$l'(\pi_2|y) \approx 0.0012$ which is smaller in absolute value than our tolerance of 0.01 so we can stop.

And of course $0.39994 \approx 0.4$

Let's go back to the original policy question:

What is the effect on test scores of reducing STR by one student/class?

Have we answered this question?

FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.