Important Ideas from Last Day

(1) Definition of a Point Estimate

(2) Definition of a Point Estimator

(3) Definition of a Sampling Distribution of an Estimator

(4) Diamond Experiment to Illustrate Concepts

(5) Review of Results from STAT 230 which can be used to determine Sampling Distributions

Important Idea

In a real empirical study you would ONLY have one data set $y_1, y_2, \ldots, y_n$ that you would then used to estimate an attribute of interest (for example, the mean of the Diamond population).

To answer the question “How good is this estimate?” we need to put this experiment in the larger context of repeating the experiment over and over again. This is where the sampling distribution becomes important.

To Do

Read Sections 4.3 - 4.4.

Start End-of-Chapter Problems 1-30.

Today's Lecture

(1) The Sampling Distribution of the Sample Mean for Gaussian Data with Known Variance. (We will do Unknown Variance in Section 4.7.)

(2) Interval Estimation and Interval Estimates

(3) Interval Estimation using Likelihood Intervals

Gaussian Data with Known Standard Deviation

Suppose that for a given set of data measured in centimeters the assumed model $Y_i \backsim G(\mu,0.5)$, $i=1,2,\ldots,n$ is reasonable.

As usual let the observed data be $y_1,y_2,\ldots,y_n$.

Suppose also that we have chosen to use

$$\tilde{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

as the estimator of the unknown mean μ.

Gaussian Data with Known Standard Deviation

In this example we don't need to simulate the sampling distribution of the point estimator of μ since we know

$$\tilde{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim G\left(\mu, \frac{0.5}{\sqrt{n}}\right)$$

Our point estimate of μ based on the observed data is

$$\hat{\mu} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Gaussian Data with Known Standard Deviation

Since the sampling distribution is known we can answer the following question related to how good the estimator is:

Suppose I ask each of you to draw a sample of size $n = 49$ and then use the sample mean to estimate μ.

How often would you expect the estimate to be within 0.1 centimeter of the true value of μ?

Gaussian Data with Known Standard Deviation

If the sample size = n = 49 then we would expect the estimate of μ to be within 0.1 centimeter of the true value of μ, 84 percent of the time.

Exercise:

Show that if the sample size = *n* = 100 then you would expect the estimate of *μ* to be within 0.1 centimeter of the true value of *μ*, 95.5 percent of the time.

Interval Estimation

Questions like:

How good is the estimate?

How often is the estimate within a certain distance of the true value?

What is the uncertainty in our estimate?

all lead naturally to the idea of interval estimation.

Interval Estimation

Suppose that for a certain population we are interested in estimating an attribute using observed data $y_1, y_2, \ldots, y_n$.

Suppose also that the attribute of interest can be represented by the parameter θ in a statistical model.

Suppose we estimate θ using the estimate

$$\hat{\theta} = g(y_1, y_2, \ldots, y_n)$$

Interval Estimation

We would like to quantify the uncertainty in our estimate.

One way we do this is to give an interval of values for θ which are "supported" by the data.

We will see how to use the sampling distribution of

$$\tilde{\theta} = g\left(Y_1, Y_2, \ldots, Y_n\right)$$

in order to quantify this idea of "support".

Interval Estimate

To indicate the uncertainty in an estimate we use an interval estimate of θ of the form

$$[L(y), U(y)]$$

where $L(y)$ and $U(y)$ are both functions of the observed data y.

Note: In most examples the data we consider are of the form $y = (y_1, y_2, \ldots, y_n)$.

Interval Estimate

For example, for Normal data we will see that intervals like

$$\left[\bar{y} - 2 \cdot \frac{s}{\sqrt{n}}, \bar{y} + 2 \cdot \frac{s}{\sqrt{n}}\right]$$

are useful for summarizing the uncertainty in the estimate

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

of the unknown mean *μ*.

Sec. 4.3: Likelihood Intervals

One way of obtaining an interval of values for the unknown parameter which are “reasonable” given the observed data, and which uses an idea that we have already discussed, is to construct a likelihood interval.

Relative Likelihood Function

Recall:

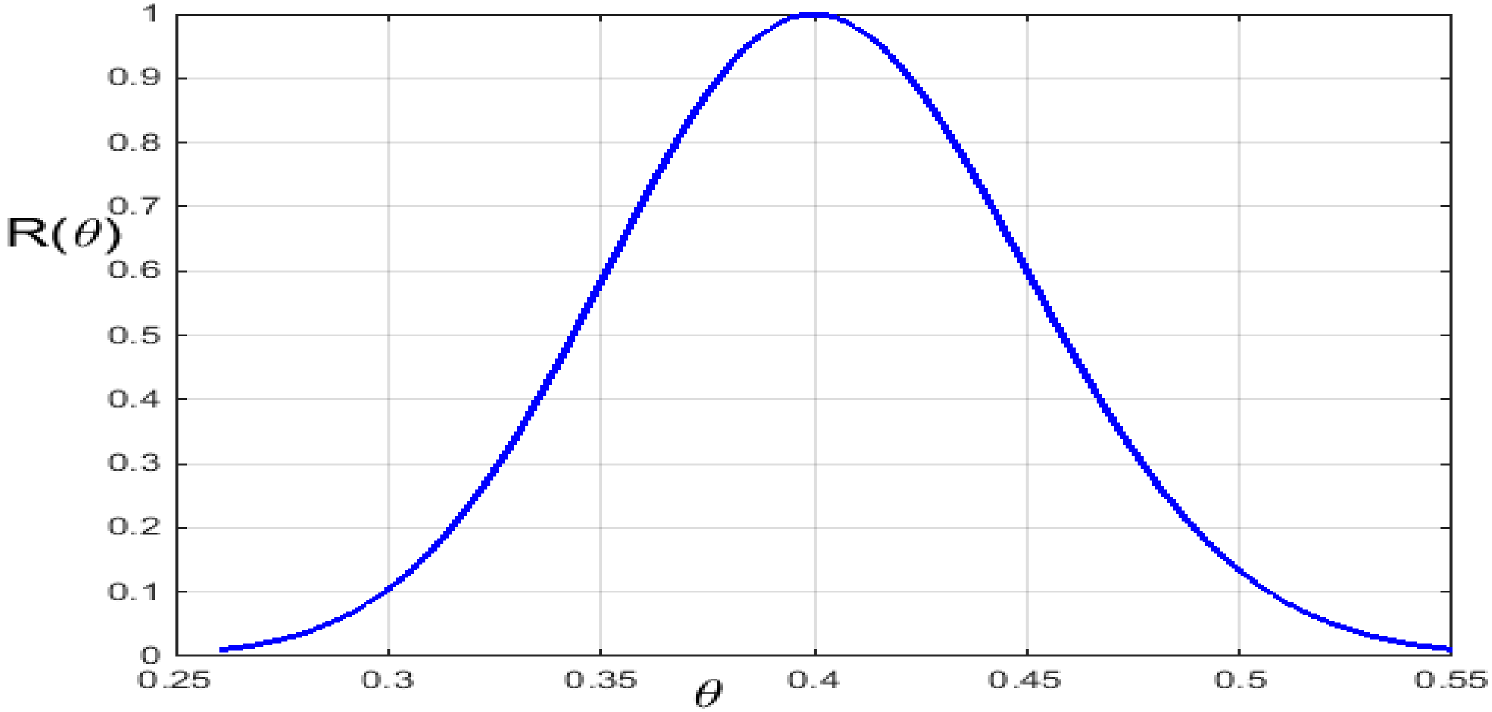
The relative likelihood function is defined as

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}, \quad \theta \in \Omega.$$

$$\text{Note: } 0 \leq R(\theta) \leq 1 \quad \text{for all } \theta \in \Omega$$

$$\text{and } R(\hat{\theta}) = 1.$$

Clicker Question



Given the above $R(\theta)$ which statement is false:

A: $\theta = 0.35$ is a plausible value of θ

B: $\theta = 0.45$ is a plausible value of θ

C: $\theta = 0.53$ is a plausible value of θ

Relative Likelihood Function and Interval Estimates

The relative likelihood function can be used to obtain interval estimates in the following way:

Definition:

A 100*p*% likelihood interval for the parameter θ is the set $\{\theta: R(\theta) \geq p\}$.

Likelihood Intervals - Notes

(1) Remember the values of θ that result in large values of $L(\theta)$ (and hence $R(\theta)$) are the most plausible in the light of the data. (Why?)

(2) The dependence of the interval $\{\theta: R(\theta) \geq p\}$ on the data y is somewhat hidden but remember that $R(\theta) = R(\theta;y)$ is a function of the data y so the endpoints of the interval will depend on the data.

Likelihood Intervals - Notes

(3) $\{\theta: R(\theta) \geq p\}$ is not necessarily an interval unless $R(\theta)$ is unimodal, but this is the case for all models that we will consider in this course.

(4) Choosing $p \in [0.10, 0.15]$ is common.

Interpretation of Likelihood Intervals

Values of θ inside a 10% likelihood interval are plausible values of θ in light of the observed data.

Values of θ inside a 50% likelihood interval are very plausible values of θ in light of the observed data.

Values of θ outside a 10% likelihood interval are implausible values of θ in light of the observed data.

Values of θ outside a 1% likelihood interval are very implausible values of θ in light of the observed data.

Relative Likelihood Function for Binomial - Review

For Binomial data

$$L(\theta) = \binom{n}{y} \theta^{y} (1-\theta)^{n-y} \quad for \;\; 0 < \theta < 1$$

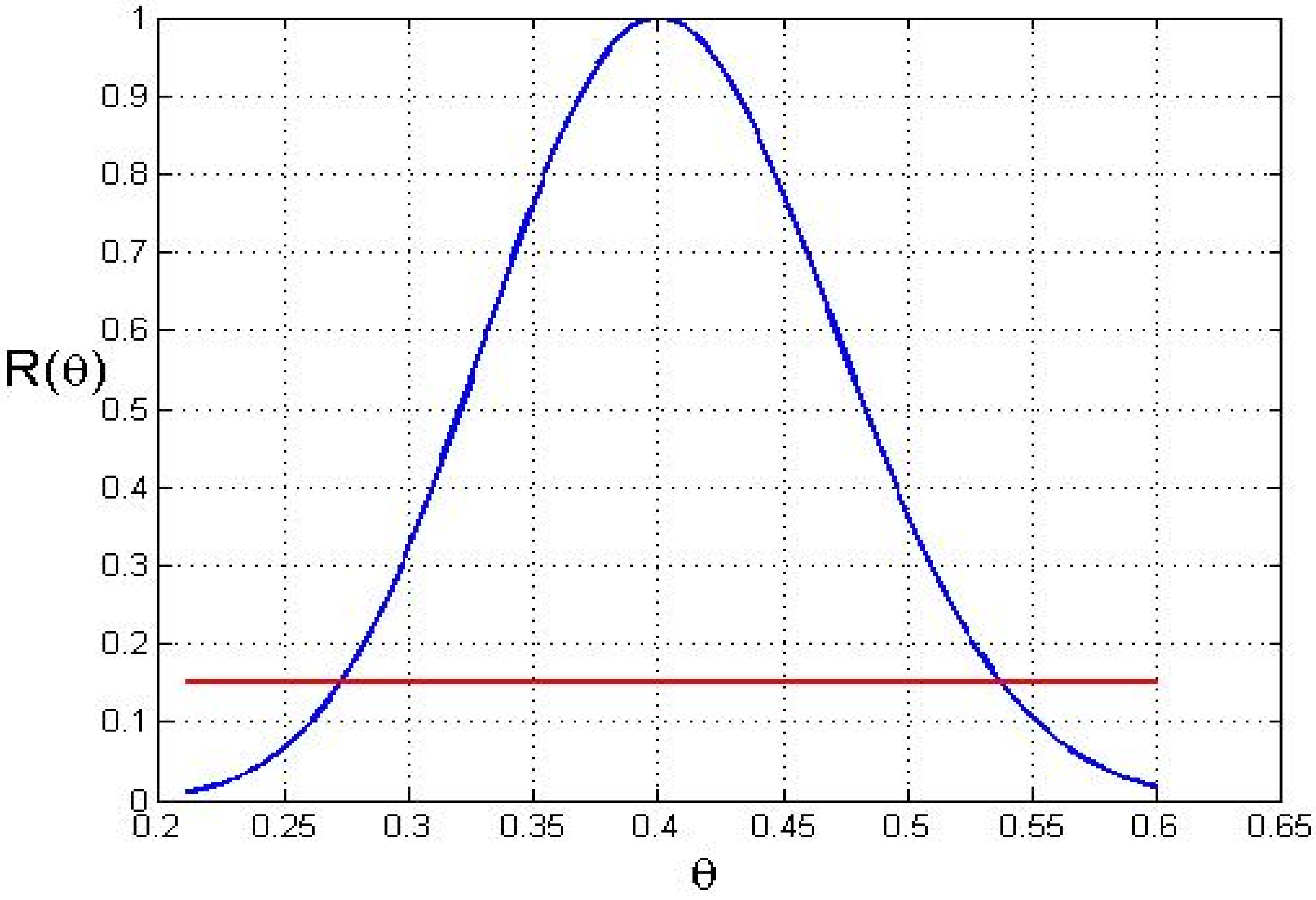
or more simply

$$L(\theta) = \theta^{y} (1-\theta)^{n-y} \quad \text{for} \;\; 0 < \theta < 1$$

The relative likelihood function is

$$R(\theta) = \frac{\theta^{y} (1-\theta)^{n-y}}{\hat{\theta}^{y} (1-\hat{\theta})^{n-y}} \quad \text{for} \;\; 0 < \theta < 1 \;\; \text{where} \;\; \hat{\theta} = \frac{y}{n}$$

Binomial R.L. for *n* = 50 and *y* = 20 with 15% Likelihood Interval



Finding a Likelihood Interval

In general a relative likelihood interval must be determined numerically using software like R.

If we graph the relative likelihood function $R(\theta)$ and then draw in the line $y = p$ we can usually read off the $100p\%$ likelihood interval quite well.

In the Binomial example we can read from the graph that a 15% likelihood interval is equal to [0.27,0.54].

Relative Likelihood Function for Binomial Distribution

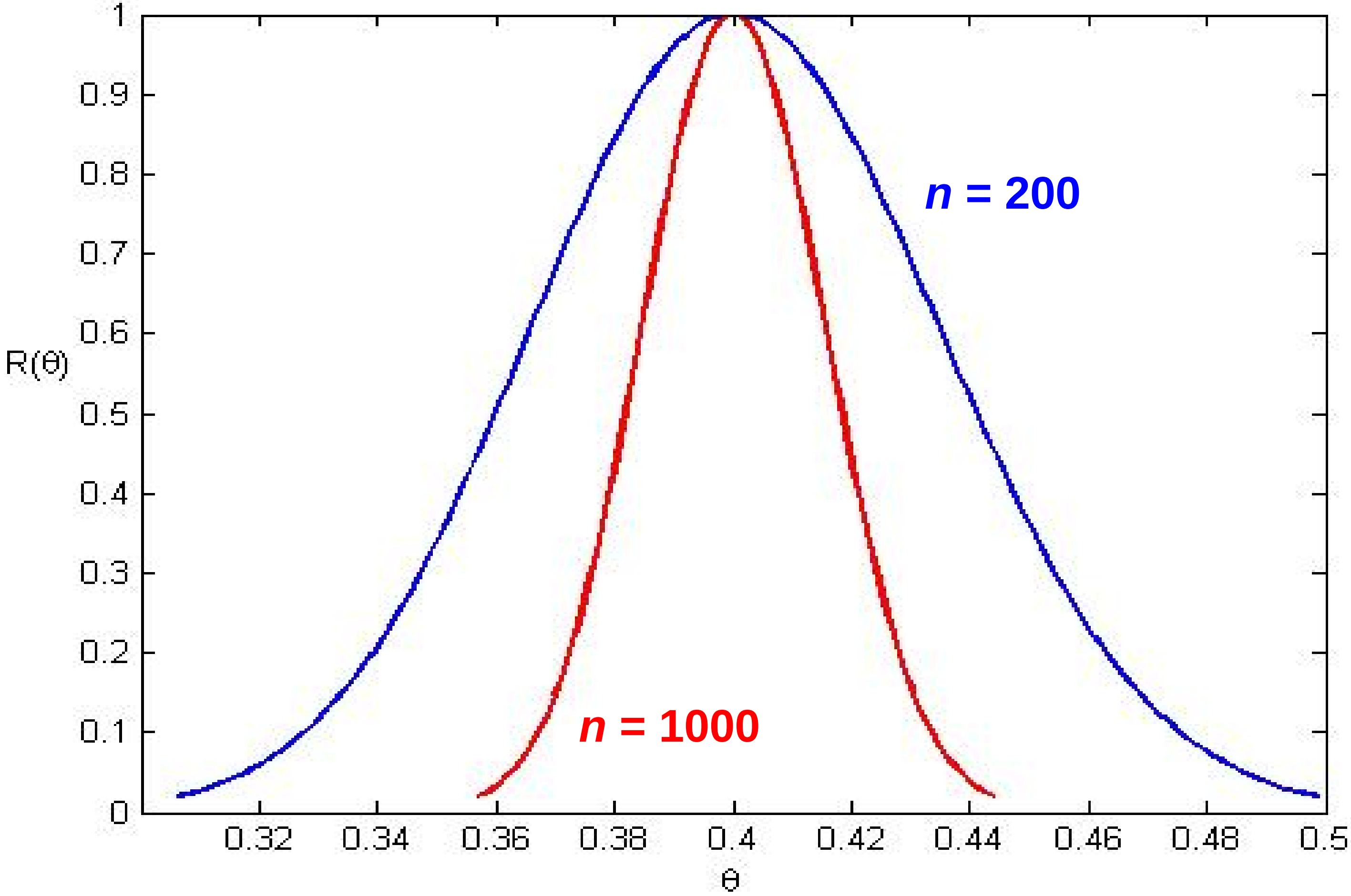
The next slide shows the relative likelihood functions $R(\theta)$ for two sets of data:

Data set 1 : $n = 200$, $y = 80$

Data set 2 : $n = 1000$, $y = 400$.

In each case $\hat{\theta} = 0.40$.

n = 200 versus n = 1000



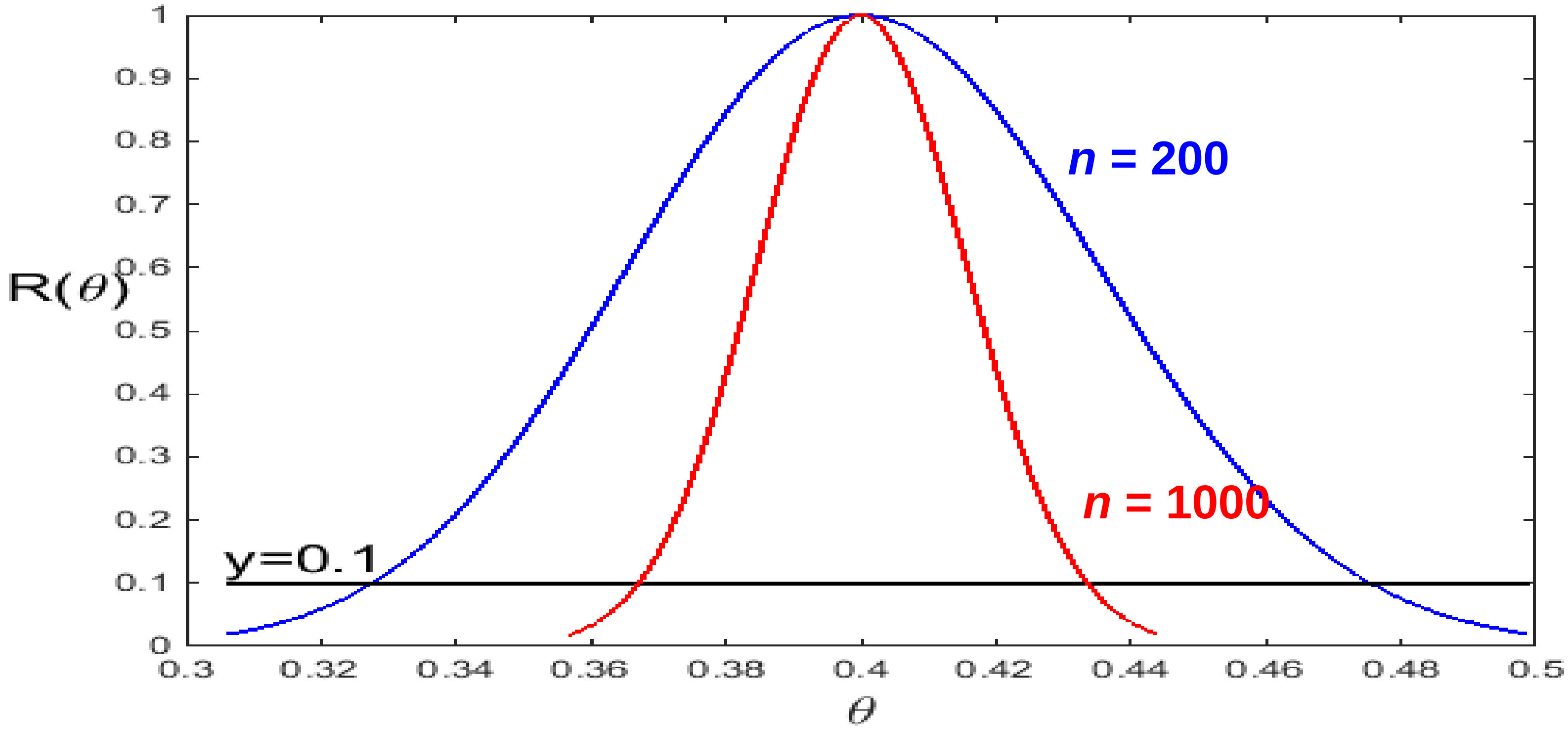
Relative Likelihood Function for Binomial Distribution

The relative likelihood function is more “concentrated” around θ for *n* = 1000.

Why does this make sense?

The likelihood intervals also reflect this.

Relative Likelihood Function for Binomial Distribution



Data set 1 : $R(\theta) \geq 0.1$ for $\theta \in [0.33, 0.47]$

Data set 2 : $R(\theta) \geq 0.1$ for $\theta \in [0.37, 0.43]$

Sample Size and Likelihood Intervals

As the sample size n increases the graph of the relative likelihood function $R(\theta)$ becomes more "concentrated" around θ.

Consequently likelihood intervals become narrower as the sample size increases.

Both these statements reflect the fact that larger data sets contain more information about the unknown parameter θ.