

Introduction to Econometrics

The statistical analysis of economic (and related) data

Brief Overview of the Course

Economics suggests interesting relations, often with policy implications, but virtually never suggests quantitative magnitudes of causal effects.

- What is the price elasticity of cigarettes?
- What is the effect of reducing class size on student achievement?
- What is the effect on earnings of a year of education?
- What is the effect on output growth of a 1 percentage point increase in interest rates by the Fed?

The focus of this course is the use of statistical and econometric methods to quantify causal effects

Ideally, we would like an experiment:

Cigarette prices; class size; returns to education; Fed

But almost always we must use observational (nonexperimental) data. Observational data poses major challenges: consider estimation of returns to education

- confounding effects (omitted factors)
- simultaneous causality
- “correlation does not imply causation”

In this course you will:

- Learn methods for estimating causal effects using observational data;
- Learn some tools that can be used for other purposes, for example forecasting using time series data;
- Focus on applications – theory is used only as needed to understand the “why”s of the methods;
- Learn to produce (you do the analysis) and consume (evaluate the work of others) econometric applications; and
- Practice “producing” in your problem sets.

Review of Probability and Statistics

(SW Chapters 2,3)

Empirical problem: Class size and educational output

- Policy question: What is the effect of reducing class size by one student per class? by 8 students/class?
- What is the right output measure (“dependent variable”)?
 - parent satisfaction
 - student personal development
 - future adult welfare and/or earnings
 - performance on standardized tests

What do data say about the class size/test score relation?

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:

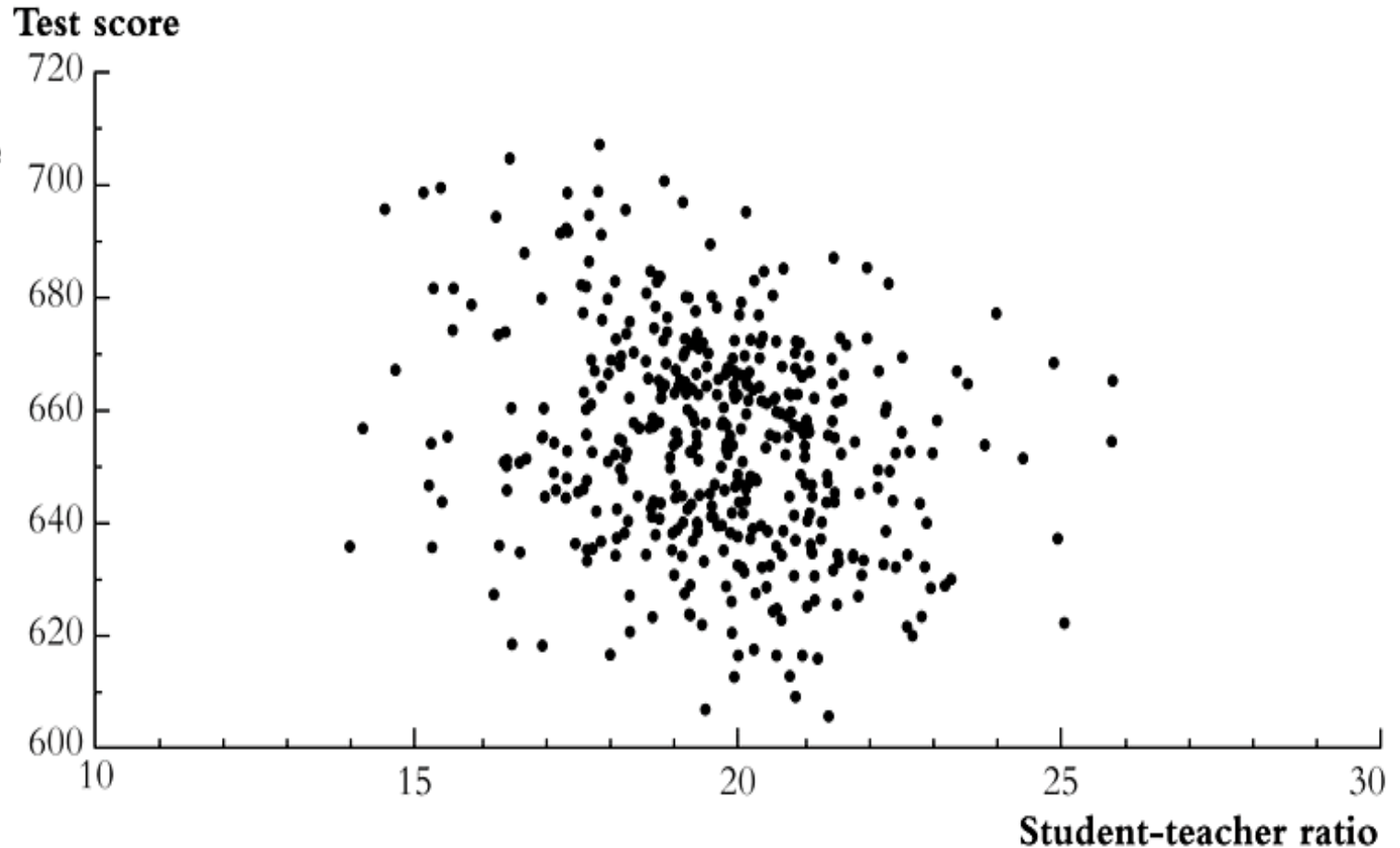
TABLE 4.1 Summary of the Distribution of Student-Teacher Ratios and Fifth-Grade Test Scores for 420 K-8 Districts in California in 1998

	Average	Standard Deviation	Percentile						
			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

Do districts with smaller classes have higher test scores?

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 .



How can we get some numerical evidence on whether districts with low STRs have higher test scores?

1. Compare average test scores in districts with low STRs to those with high STRs (“*estimation*”)
2. Test the 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 (\bar{Y})	Standard deviation (s_Y)	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 $\Delta = 0$
3. *Construct a confidence interval* for Δ

1. Estimation

$$\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = 657.4 - 650.0 = 7.4$$

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

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

2. Hypothesis testing

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

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

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_Y	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.

3. Confidence interval

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

$$\begin{aligned}(\bar{Y}_s - \bar{Y}_l) \pm 1.96 \times SE(\bar{Y}_s - \bar{Y}_l) \\ = 7.4 \pm 1.96 \times 1.83 = (3.8, 11.0)\end{aligned}$$

Two equivalent statements:

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

This should all be familiar. But:

1. What is the underlying framework that justifies all this?
2. Estimation: Why estimate Δ by $\bar{Y}_s - \bar{Y}_l$?
3. Testing: What is the standard error of $\bar{Y}_s - \bar{Y}_l$, really?
Why reject $\Delta = 0$ if $|t| > 1.96$?
4. Confidence intervals (interval estimation): What is a confidence interval, really?

- 1. The probability framework for statistical inference**
2. Estimation
3. Testing
4. Confidence Intervals

Population

- The group or collection of entities of interest
- Here, “all possible” school districts
- “All possible” means “all possible” circumstances that lead to specific values of STR, test scores
- We will think of populations as infinitely large; the task is to make inferences from a sample from a large population

Random variable Y

- Numerical summary of a random outcome
- Here, the numerical value of district average test scores (or district STR), once we choose a year/district to sample.

Population distribution of Y

- The probabilities of different values of Y that occur in the population, for ex. $\Pr[Y = 650]$ (when Y is discrete)
- or: The probabilities of sets of these values, for ex. $\Pr[Y > 650]$ (when Y is continuous).

“Moments” of the population distribution

mean = expected value

$$= E(Y)$$

$$= \mu_Y$$

= long-run average value of Y over repeated realizations of Y

$$\text{variance} = E(Y - \mu_Y)^2$$

$$= \sigma_Y^2$$

= measure of the squared spread of the distribution

$$\text{standard deviation} = \sqrt{\text{variance}} = \sigma_Y$$

Conditional distributions

- The distribution of Y , given value(s) of some other random variable, X
- Ex: the distribution of test scores, given that $STR < 20$

Moments of conditional distributions

- conditional mean = mean of conditional distribution
 $= E(Y|X = x)$ (*important notation*)
- conditional variance = variance of conditional distribution
- *Example:* $E(\text{Test scores}|STR < 20)$, the mean of test scores for districts with small class sizes

The difference in means is the difference between the means of two conditional distributions:

$$\Delta = E(\text{Test scores} | STR < 20) - E(\text{Test scores} | STR \geq 20)$$

Other examples of conditional means:

- Wages of all female workers ($Y = \text{wages}$, $X = \text{gender}$)
- One-year mortality rate of those given an experimental treatment ($Y = \text{live/die}$; $X = \text{treated/not treated}$)

The conditional mean is a new term for the familiar idea of the group mean

Inference about means, conditional means, and differences in conditional means

We would like to know Δ (test score gap; gender wage gap; effect of experimental treatment), but we don't know it.

Therefore we must collect and use data that permits making statistical inferences about Δ .

- Experimental data
- Observational data

Simple random sampling

- Choose an individual (district, entity) at random from the population

Randomness and data

- Prior to sample selection, the value of Y is random because the individual selected is random
- Once the individual is selected and the value of Y is observed, then Y is just a number – not random
- The data set is (Y_1, Y_2, \dots, Y_n) , where $Y_i =$ value of Y for the i^{th} individual (district, entity) sampled

Implications of simple random sampling

Because individuals #1 and #2 are selected at random, the value of Y_1 has no information content for Y_2 . Thus:

- Y_1, Y_2 are *independently distributed*
- Y_1 and Y_2 come from the same distribution, that is, Y_1, Y_2 are *identically distributed*
- That is, a consequence of simple random sampling is that Y_1 and Y_2 are independently and identically distributed (*i.i.d.*).
- More generally, under simple random sampling, $\{Y_i\}$, $i = 1, \dots, n$, are i.i.d

1. The probability framework for statistical inference
- 2. Estimation**
3. Testing
4. Confidence Intervals

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

- What are the properties of this estimator?
- Why should we use \bar{Y} rather than some other estimator?
 - Y_1 (the first observation)
 - maybe unequal weights – not simple average
 - $\text{median}(Y_1, \dots, Y_n)$

To answer these questions we need to characterize the *sampling distribution* of \bar{Y}

- The individuals in the sample are drawn at random.
- Thus the values of (Y_1, \dots, Y_n) are random
- Thus functions of (Y_1, \dots, Y_n) , such as \bar{Y} , are random: had a different sample been drawn, they would have taken on a different value
- The distribution of \bar{Y} over different possible samples of size n is called the *sampling distribution* of \bar{Y} .
- The mean and variance of \bar{Y} are the mean and variance of its sampling distribution, $E(\bar{Y})$ and $\text{var}(\bar{Y})$.
- To compute $\text{var}(\bar{Y})$, we need the *covariance*

The *covariance* between r.v.'s X and Z is,

$$\text{cov}(X,Z) = E[(X - \mu_X)(Z - \mu_Z)] = \sigma_{XZ}$$

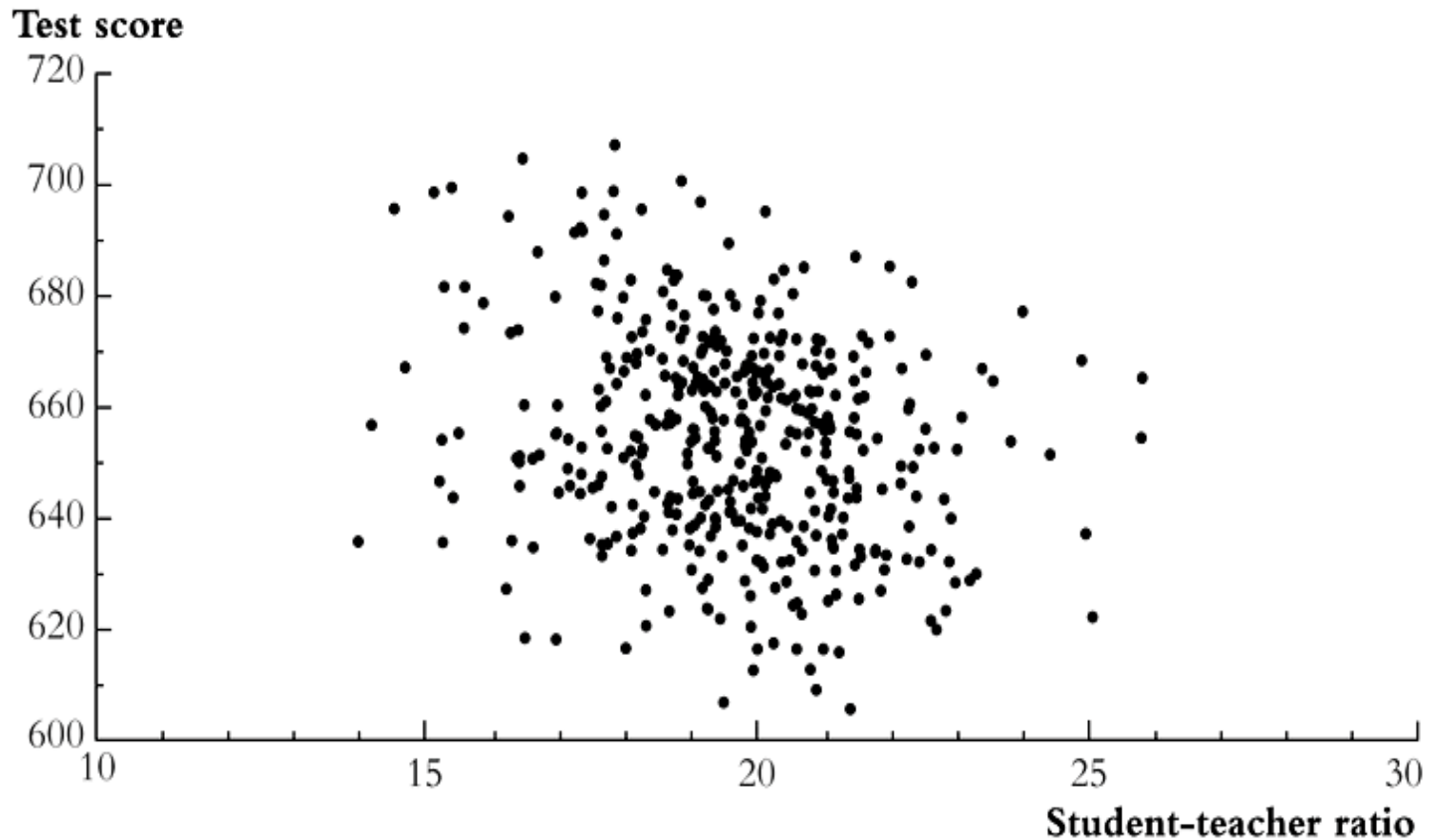
- The covariance is a measure of the linear association between X and Z ; its units are units of $X \times$ units of Z
- $\text{cov}(X,Z) > (<) 0$: X and Z positive (negative) relation between X and Z
- If X and Z are independently distributed, then $\text{cov}(X,Z) = 0$ (but not vice versa!!)
- The covariance of a r.v. with itself is its variance:

$$\text{cov}(X,X) = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = \sigma_X^2$$

The covariance between *Test Score* and *STR* is negative:

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 .



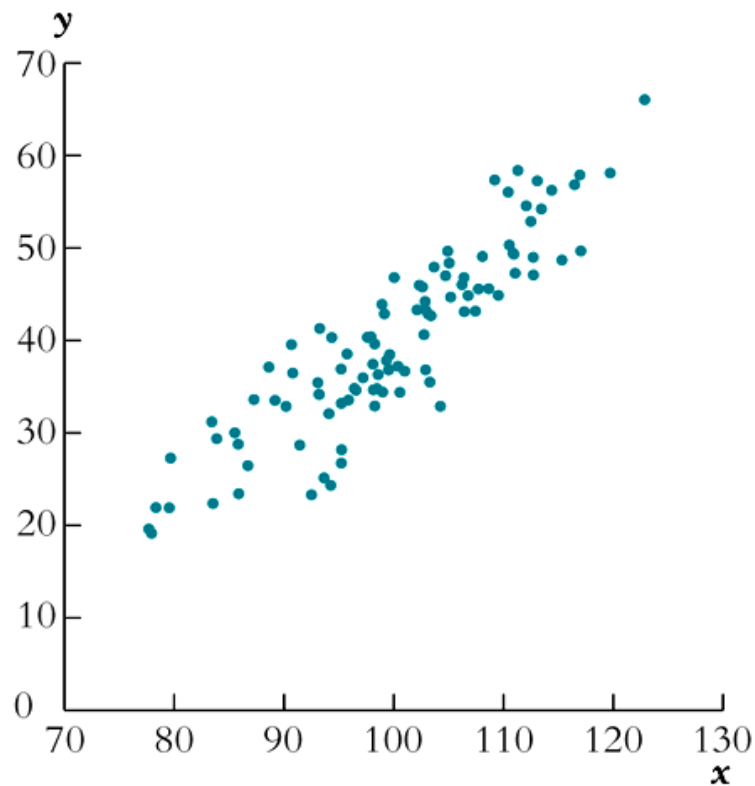
The *correlation coefficient* is defined in terms of the covariance:

$$\text{corr}(X,Z) = \frac{\text{cov}(X,Z)}{\sqrt{\text{var}(X)\text{var}(Z)}} = \frac{\sigma_{XZ}}{\sigma_X\sigma_Z} = r_{XZ}$$

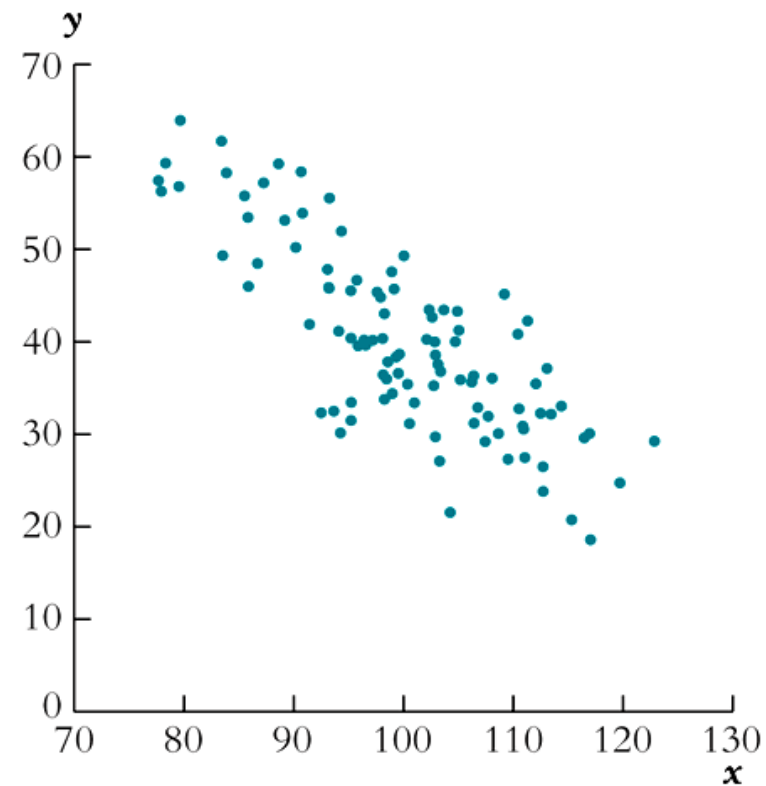
- $-1 \leq \text{corr}(X,Z) \leq 1$
- $\text{corr}(X,Z) = 1$ means perfect positive linear association
- $\text{corr}(X,Z) = -1$ means perfect negative linear association
- $\text{corr}(X,Z) = 0$ means no linear association
- If $E(X|Z) = \text{const}$, then $\text{corr}(X,Z) = 0$ (not necessarily vice versa however)

The correlation coeff. measures linear association

FIGURE 3.3 Scatterplots for Four Hypothetical Data Sets



(a) Correlation = +0.9

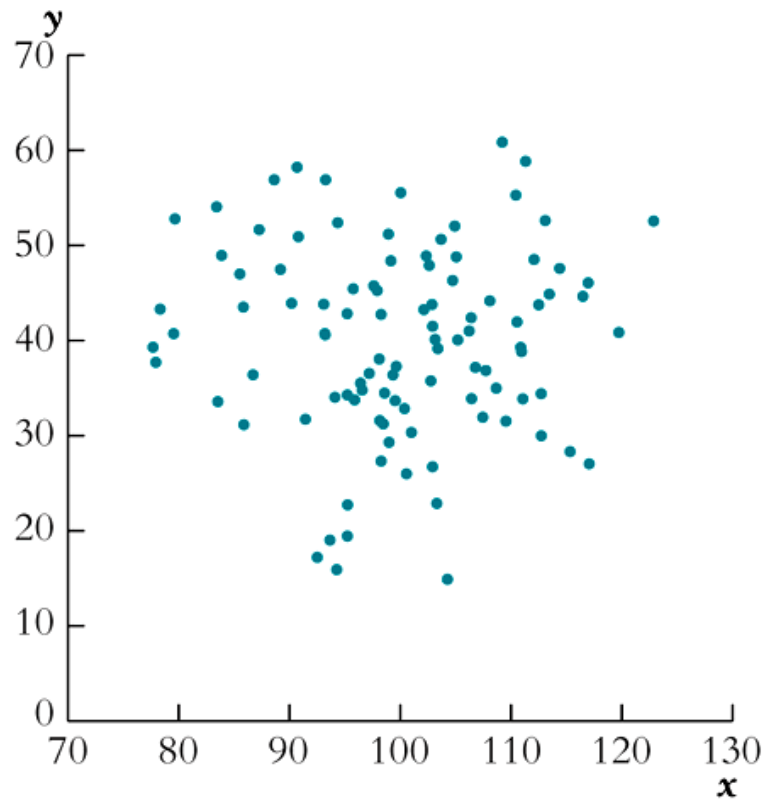


(b) Correlation = -0.8

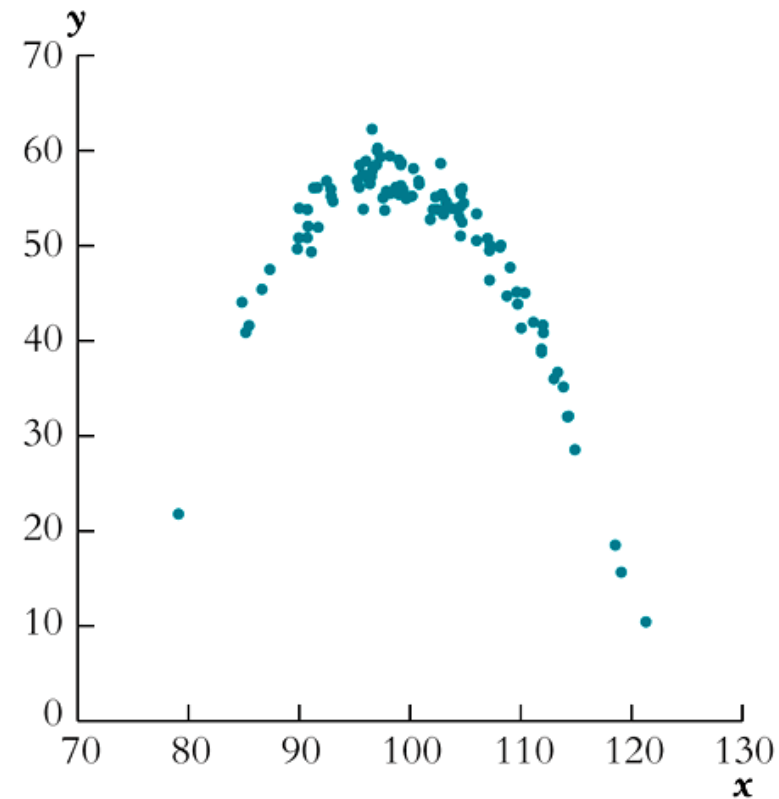
The scatterplots in Figures 3.3a and 3.3b show strong linear relationships between X and Y . In Figure 3.3c, X is independent of Y and the two variables are uncorrelated. In Figure 3.3d, the two variables also are uncorrelated even though they are related nonlinearly.

The correlation coeff. measures linear association

FIGURE 3.3 Scatterplots for Four Hypothetical Data Sets



(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)

The scatterplots in Figures 3.3a and 3.3b show strong linear relationships between X and Y . In Figure 3.3c, X is independent of Y and the two variables are uncorrelated. In Figure 3.3d, the two variables also are uncorrelated even though they are related nonlinearly.

The mean and variance of the sampling distribution of \bar{Y}

$$\text{mean: } E(\bar{Y}) = E\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n} \sum_{i=1}^n E(Y_i) = \frac{1}{n} \sum_{i=1}^n \mu_Y = \mu_Y$$

$$\begin{aligned} \text{Variance: } \text{var}(\bar{Y}) &= E[\bar{Y} - E(\bar{Y})]^2 \\ &= E[\bar{Y} - \mu_Y]^2 \\ &= E\left[\frac{1}{n} \sum_{i=1}^n (Y_i - \mu_Y)\right]^2 \\ &= E\left[\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (Y_i - \mu_Y)(Y_j - \mu_Y)\right] \end{aligned}$$

so

$$\begin{aligned}\text{var}(\bar{Y}) &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n E(Y_i - \mu_Y)(Y_j - \mu_Y) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{cov}(Y_i, Y_j) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{var}(Y_i) + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \text{cov}(Y_i, Y_j) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sigma_Y^2 + 0 \\ &= \frac{\sigma_Y^2}{n}\end{aligned}$$

Summary: $E(\bar{Y}) = \mu_Y$ and $\text{var}(\bar{Y}) = \frac{\sigma_Y^2}{n}$.

Implications:

- \bar{Y} is an *unbiased* estimator of μ_Y (that is, $E(\bar{Y}) = \mu_Y$)
- $\text{var}(\bar{Y})$ is inversely proportional to n
- spread of sampling distribution is proportional to $1/\sqrt{n}$
- in this sense, the sampling uncertainty arising from using \bar{Y} to make inferences about μ_Y is proportional to $1/\sqrt{n}$

What about the entire sampling distribution of \bar{Y} , not just the mean and variance?

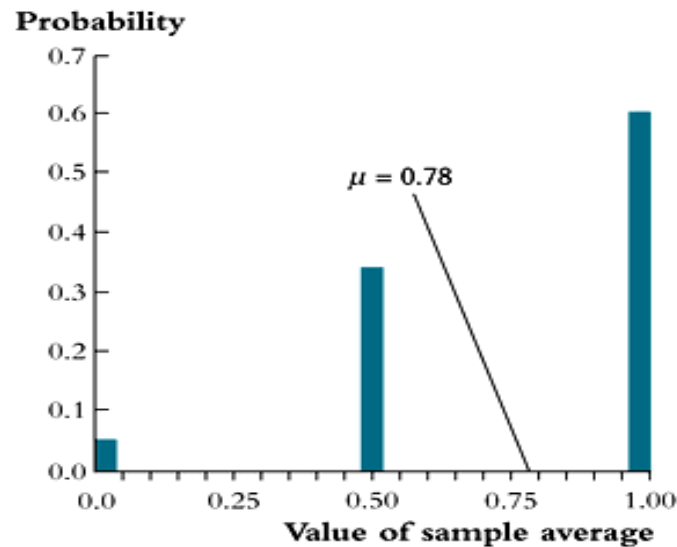
In general, the exact sampling distribution of \bar{Y} is very complicated and depends on the population distribution of Y .

Example: Suppose Y takes on 0 or 1 (a *Bernoulli* random variable) with the probability distribution,

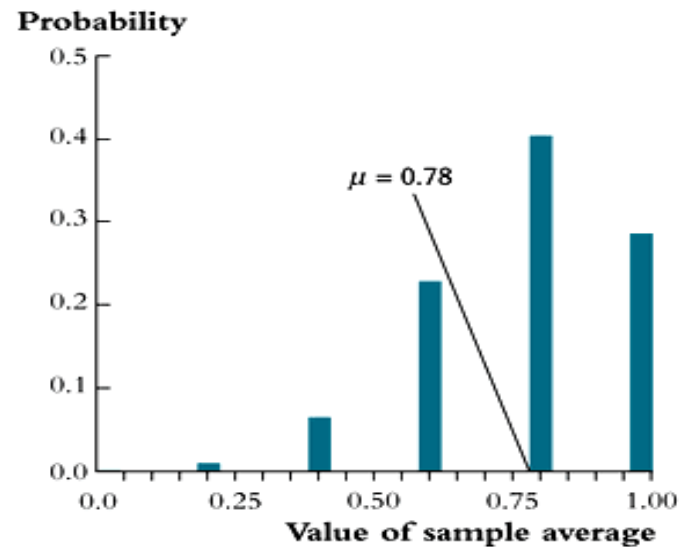
$$\Pr[Y = 0] = .22, \Pr(Y = 1) = .78$$

Then $E(Y) = .78$ and $\sigma_Y^2 = .78 \times (1 - .78) = 0.1716$

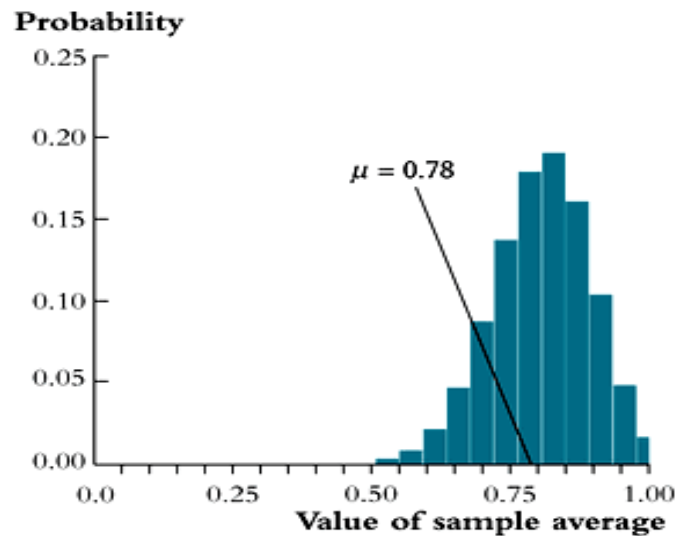
FIGURE 2.6 Sampling Distribution of the Sample Average of n Bernoulli Random Variables



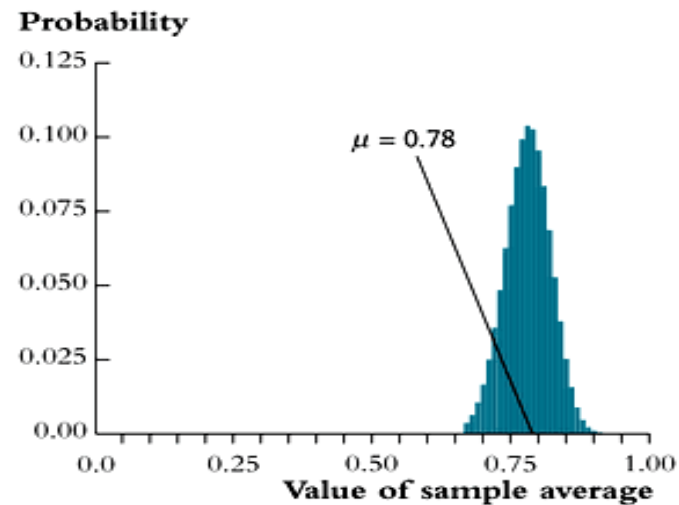
(a) $n = 2$



(b) $n = 5$



(c) $n = 25$



(d) $n = 100$

The distributions are the sampling distributions of \bar{Y} , the sample average of n independent Bernoulli random variables with $p = \Pr(Y_i = 1) = 0.78$ (the probability of a fast commute is 78%). The variance of the sampling distribution of \bar{Y} decreases as n gets larger, so the sampling distribution becomes more tightly concentrated around its mean $\mu = 0.78$ as the sample size n increases.

For small sample sizes, the distribution of \bar{Y} is complicated.

BUT: when n is large, it is not!

(1) As n increases, the distribution of \bar{Y} becomes more tightly centered around μ_Y : the sampling uncertainty decreases as n increases (recall that $\text{var}(\bar{Y}) = \sigma_Y^2/n$)

An estimator is *consistent* if the probability that its falls within an interval of the true population value tends to one as the sample size increases.

The *Law of Large Numbers*:

If (Y_1, \dots, Y_n) are i.i.d. and $\sigma_Y^2 < \infty$, then \bar{Y} is a consistent estimator of μ_Y , that is,

$$\Pr[|\bar{Y} - \mu_Y| < \varepsilon] \rightarrow 1 \text{ as } n \rightarrow \infty$$

which can be written, $\bar{Y} \xrightarrow{p} \mu_Y$ (“ \bar{Y} converges in probability to μ_Y ”)

(Proof: as $n \rightarrow \infty$, $\text{var}(\bar{Y}) = \frac{\sigma_Y^2}{n} \rightarrow 0$, which implies that

$$\Pr[|\bar{Y} - \mu_Y| < \varepsilon] \rightarrow 1.)$$

(2) Central limit theorem (CLT): If (Y_1, \dots, Y_n) are i.i.d. and $0 < \sigma_Y^2 < \infty$, then when n is large the distribution of \bar{Y} is well approximated by a normal distribution:

- \bar{Y} is approximately distributed $N(\mu_Y, \frac{\sigma_Y^2}{n})$ (“normal distribution with mean μ_Y and variance σ_Y^2/n ”)

- $\sqrt{n}(\bar{Y} - \mu_Y)/\sigma_Y$ is approximately distributed $N(0,1)$ (standard normal)

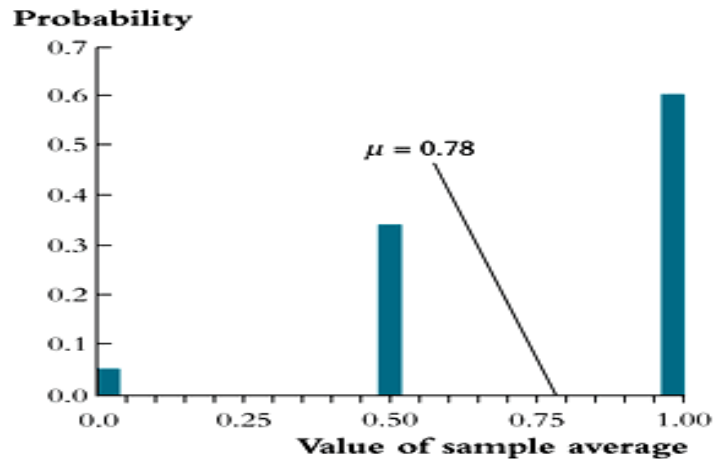
- That is, “standardized” $\bar{Y} = \frac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}} = \frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{n}}$ is

approximately distributed as $N(0,1)$

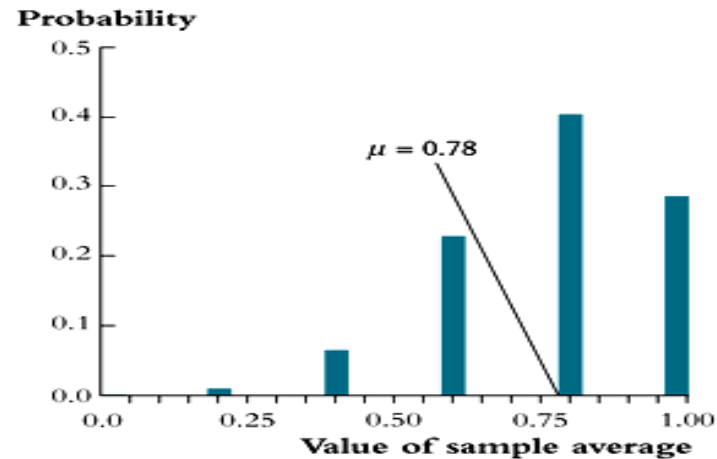
- The approximation gets better as n increases

Example: Y has Bernoulli distribution, $p = 0.78$:

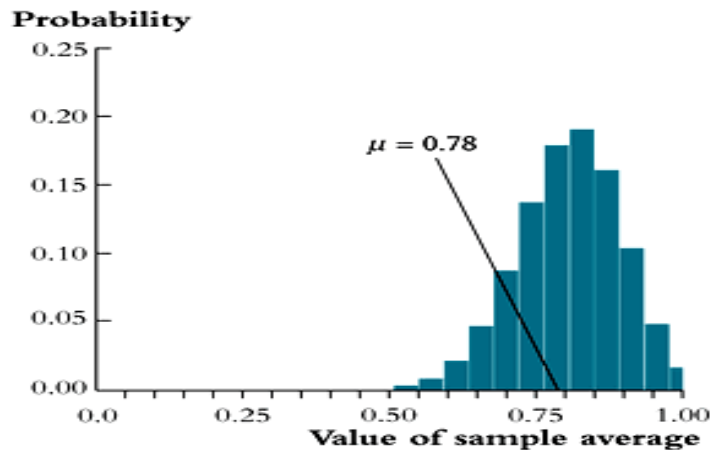
FIGURE 2.6 Sampling Distribution of the Sample Average of n Bernoulli Random Variables



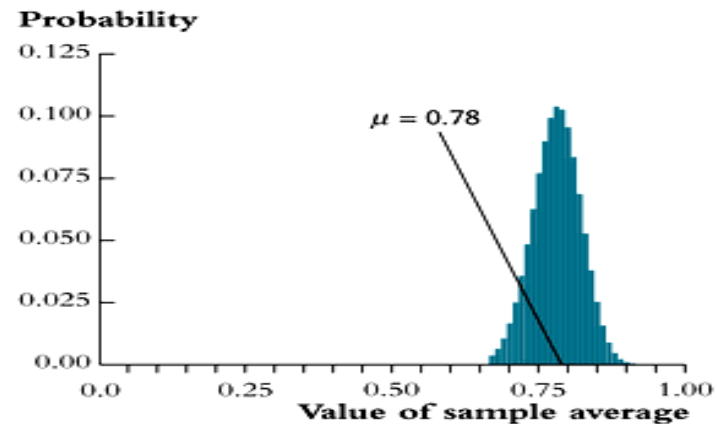
(a) $n = 2$



(b) $n = 5$



(c) $n = 25$

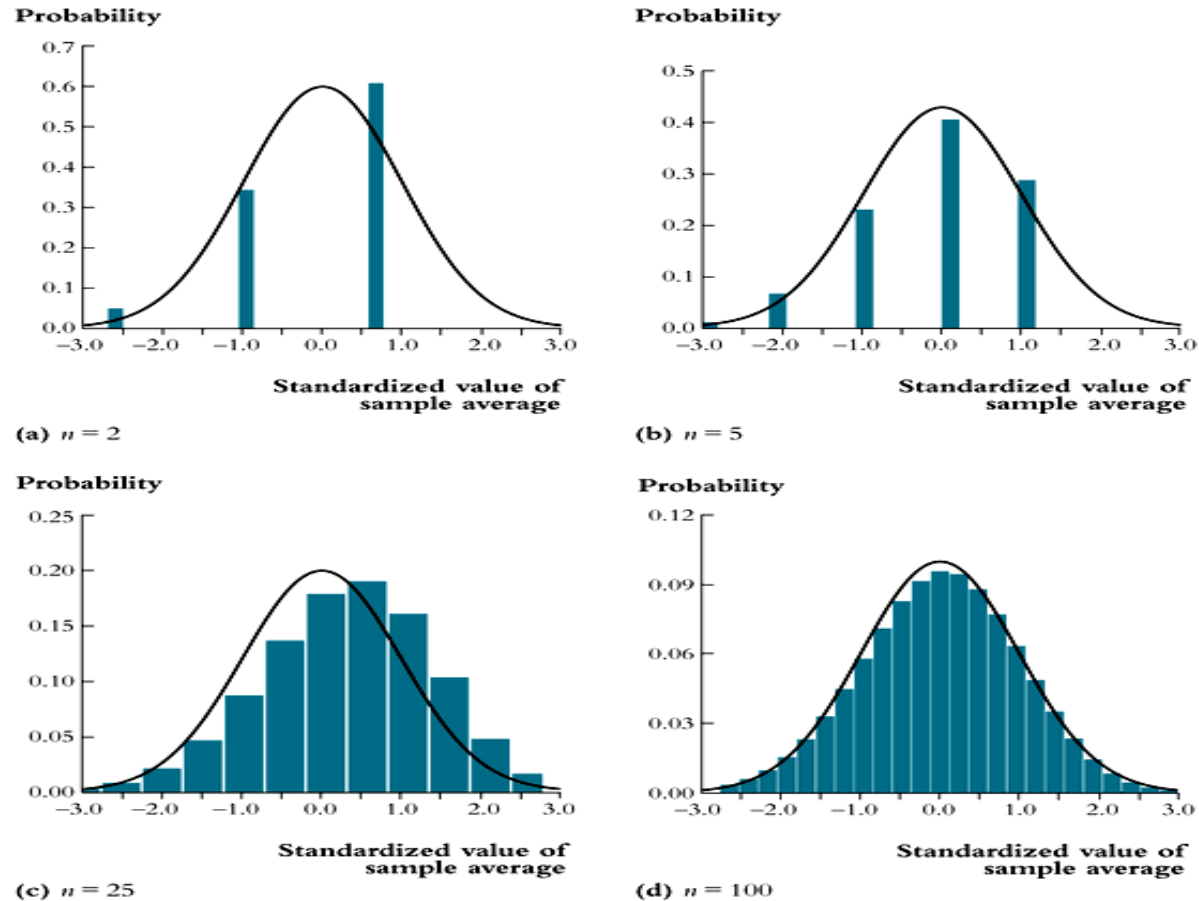


(d) $n = 100$

The distributions are the sampling distributions of \bar{Y} , the sample average of n independent Bernoulli random variables with $p = \Pr(Y_i = 1) = 0.78$ (the probability of a fast commute is 78%). The variance of the sampling distribution of \bar{Y} decreases as n gets larger, so the sampling distribution becomes more tightly concentrated around its mean $\mu = 0.78$ as the sample size n increases.

Same example: distribution of $\frac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}}$:

FIGURE 2.7 Distribution of the Standardized Sample Average of n Bernoulli Random Variables with $p = .78$



The sampling distribution of \bar{Y} in Figure 2.6 is plotted here after standardizing \bar{Y} . This centers the distributions in Figure 2.6 and magnifies the scale on the horizontal axis by a factor of \sqrt{n} . When the sample size is large, the sampling distributions are increasingly well approximated by the normal distribution (the solid line), as predicted by the central limit theorem.

Summary: for (Y_1, \dots, Y_n) i.i.d. with $0 < \sigma_Y^2 < \infty$,

- The exact (finite sample) sampling distribution of \bar{Y} has mean μ_Y (“ \bar{Y} is an unbiased estimator of μ_Y ”) and variance σ_Y^2/n
- Other than its mean and variance, the exact distribution of \bar{Y} is complicated and depends on the distribution of Y
- $\bar{Y} \xrightarrow{p} \mu_Y$ (Law of large numbers)
- $\frac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}}$ is approximately distributed $N(0,1)$ (CLT)

So, why use \bar{Y} to estimate μ_Y ?

- unbiasedness: $E(\bar{Y}) = \mu_Y$
- consistency: $\bar{Y} \xrightarrow{p} \mu_Y$
- \bar{Y} is the “least squares” estimator of μ_Y ; \bar{Y} solves,

$$\min_m \sum_{i=1}^n (Y_i - m)^2 \quad (\text{calculus; or see App. 3.2})$$

- \bar{Y} has a smaller variance than all other *linear unbiased* estimators: consider the estimator,

$$\hat{\mu}_Y = \frac{1}{n} \sum_{i=1}^n a_i Y_i, \text{ where } \{a_i\} \text{ are such that } \hat{\mu}_Y \text{ is unbiased;}$$

then $\text{var}(\bar{Y}) \leq \text{var}(\hat{\mu}_Y)$.

1. The probability framework for statistical inference
2. Estimation
- 3. Hypothesis Testing**
4. Confidence intervals

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

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

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

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

p-value = probability of drawing a statistic (e.g. \bar{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 \bar{Y} :

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

where \bar{Y}^{act} is the value of \bar{Y} actually observed (nonrandom)

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

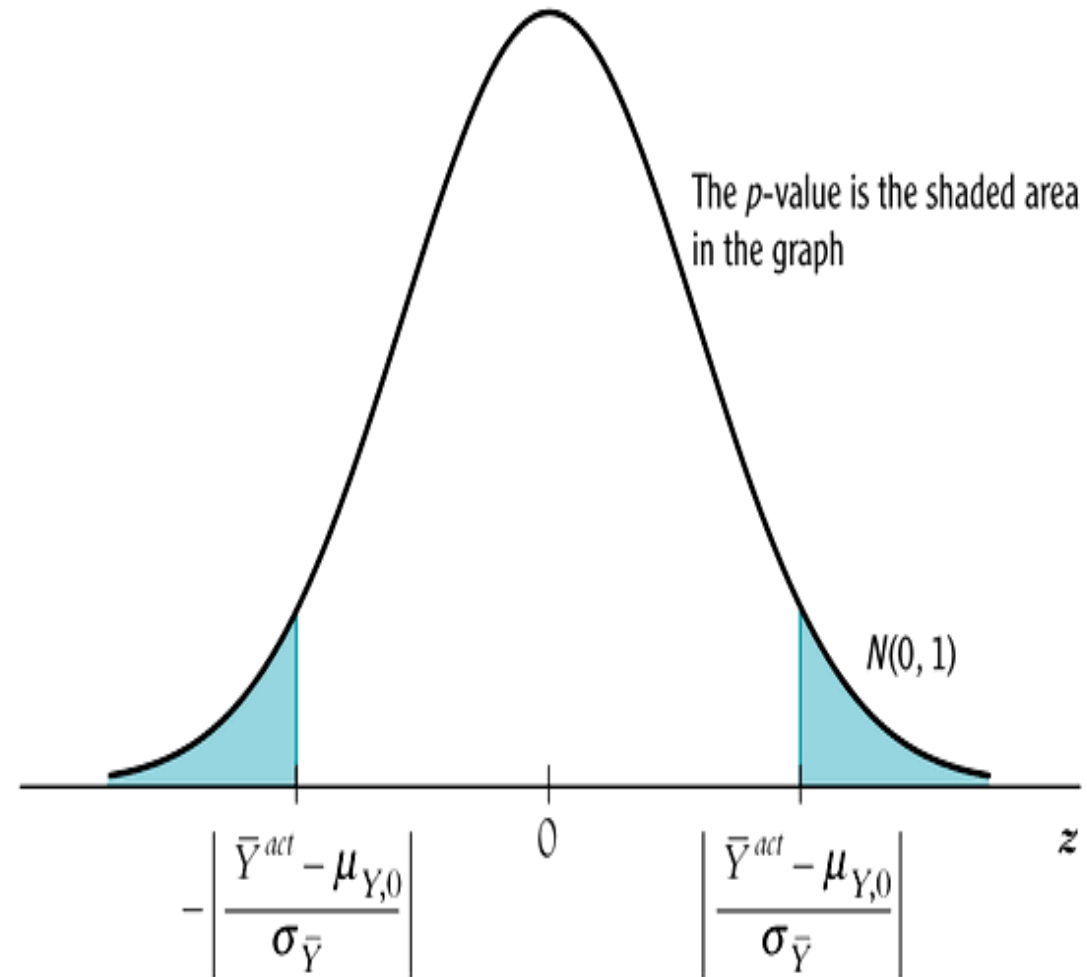
To compute the p -value, you need the distribution of \bar{Y} .
If n is large, we can use the large- n normal approximation:

$$\begin{aligned} p\text{-value} &= \Pr_{H_0} [|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|], \\ &= \Pr_{H_0} \left[\left| \frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}} \right| > \left| \frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}} \right| \right] \\ &\approx \text{probability under left+right } N(0,1) \text{ tails} \end{aligned}$$

Let $\sigma_{\bar{Y}}$ denote the std. dev. of the distribution of \bar{Y} :

FIGURE 3.1 Calculating a p -value

The p -value is the probability of drawing a value of \bar{Y} that differs from $\mu_{Y,0}$ by at least as much as \bar{Y}^{act} . In large samples, \bar{Y} is distributed $N(\mu_{Y,0}, \sigma_{\bar{Y}}^2)$ under the null hypothesis, so $(\bar{Y} - \mu_{Y,0})/\sigma_{\bar{Y}}$ is distributed $N(0, 1)$. Thus the p -value is the shaded standard normal tail probability outside $\pm |(\bar{Y}^{act} - \mu_{Y,0})/\sigma_{\bar{Y}}|$.



In practice, $\sigma_{\bar{Y}}$ is unknown – it too must be estimated

Estimator of the variance of Y:

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

Fact:

If (Y_1, \dots, 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; see Appendix 3.3
- Technical note: we assume $E(Y^4) < \infty$ because here the average is not of Y_i , but of its square; see App. 3.3

Computing the p -value with σ_Y^2 estimated:

$$\begin{aligned} p\text{-value} &= \Pr_{H_0} [|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|], \\ &= \Pr_{H_0} \left[\left| \frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}} \right| > \left| \frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}} \right| \right] \\ &\approx \Pr_{H_0} \left[\left| \frac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}} \right| > \left| \frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y / \sqrt{n}} \right| \right] \quad (\text{large } n) \\ &= \Pr_{H_0} [|t| > |t^{act}|] \\ &\approx \text{probability under normal tails} \end{aligned}$$

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

The p -value and the significance level

With a prespecified significance level (e.g. 5%):

- reject if $|t| > 1.96$
- equivalently: reject if $p < 0.05$.
- The p -value is sometimes called the *marginal significance level*.

The Student t -distribution

If Y is distributed $N(\mu_Y, \sigma_Y^2)$, then the t -statistic has the Student t -distribution (tabulated in back of all stats books)

Some comments:

- For $n > 30$, the t -distribution and $N(0,1)$ are very close
- The assumption that Y is distributed $N(\mu_Y, \sigma_Y^2)$ is rarely plausible in practice (income? number of children?)
- The t -distribution is an historical artifact from days when sample sizes were very small
- In this class, we won't use the t distribution – we rely solely on the large- n approximation given by the CLT

1. The probability framework for statistical inference
2. Estimation
3. Testing
- 4. 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.

(What is random here? the confidence interval – it will differ from one sample to the next; the population parameter, μ_Y , is not random, we just don't know it.)

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.

$$\begin{aligned}\{\mu_Y: \left| \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \right| < 1.96\} &= \{\mu_Y: -1.96 < \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} < 1.96\} \\ &= \{\mu_Y: -1.96 \frac{s_Y}{\sqrt{n}} < \bar{Y} - \mu_Y < 1.96 \frac{s_Y}{\sqrt{n}}\} \\ &= \{\mu_Y: (\bar{Y} - 1.96 \frac{s_Y}{\sqrt{n}}, \bar{Y} + 1.96 \frac{s_Y}{\sqrt{n}})\}\end{aligned}$$

This confidence interval relies on the large- n results that

\bar{Y} is approximately normally distributed and $s_Y^2 \xrightarrow{P} \sigma_Y^2$

Summary:

From the assumptions of:

- (1) simple random sampling of a population, that is, $\{Y_i, i = 1, \dots, n\}$ are i.i.d.
- (2) $0 < E(Y^4) < \infty$

we developed, for large samples (large n):

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

Are assumptions (1) & (2) plausible in practice? Yes

Original policy question:

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

Have we answered this question?

- We examined Δ = the difference in means, small v. large classes
- But Δ doesn't really answer the policy question.
- Rather, the object of policy interest is $\frac{\Delta \text{Test score}}{\Delta STR}$
- But this is the slope of a line relating test score and *STR*
- So somehow we need to estimate this slope...