Applied Statistics and Econometrics
Lecture 2

Giuseppe Ragusa

Luiss University

gragusa@luiss.it
http://gragusa.org/

February 14, 2017

Luiss University

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:

(You should already know how to interpret this table)

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

			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	665.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

Figure 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?

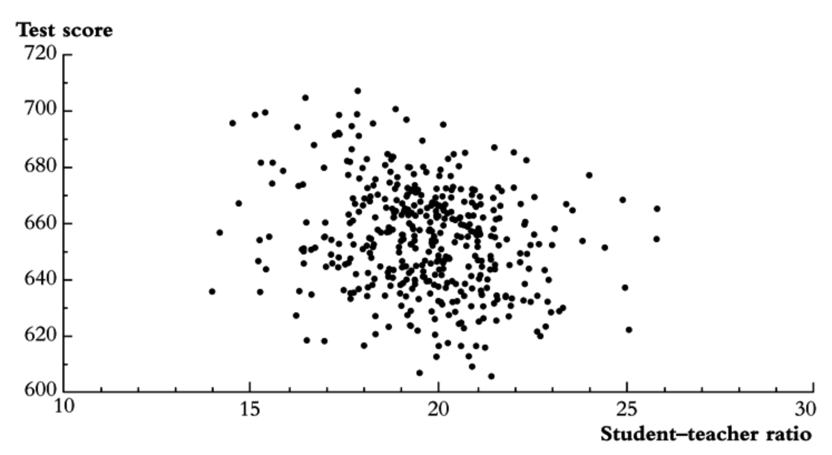


Figure 2: Scatterplot str and testscore

Approach

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

"Estimation"

Compare average test scores in districts with low STRs to those with high STRs

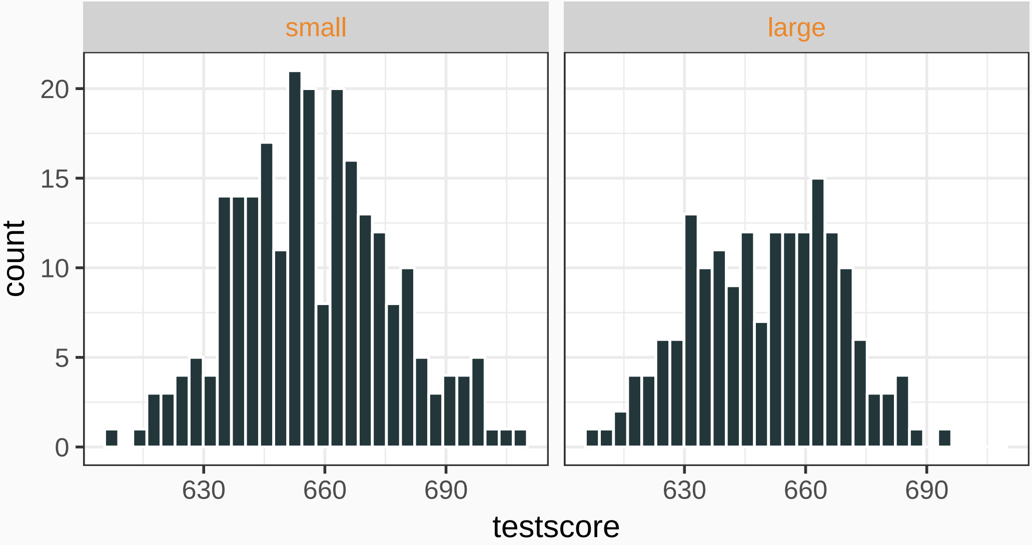
"Hypothesis testing"

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

"Confidence interval"

Estimate an interval for the difference in the mean test scores, high v. low STR districts

Initial data analysis



Initial data analysis

		testscore	
str	n	mean	sd
small	239	657.25	19.39
large	181	650.08	17.85
all	420	654.16	19.05

Steps

1. Estimation of $\Delta = \bar{Y}_{small} - \bar{Y}_{large}$ (difference between group means)
2. Test the hypothesis that $\Delta = 0$
3. Construct a confidence interval for Δ

Estimation

$$
\begin{aligned}
\bar{Y}_{small} - \bar{Y}_{large} =& \frac{1}{n_{small}} \sum_{i \in \text{small}} Y_i - \frac{1}{n_{large}} \sum_{i \in \text{large}} Y_i \\
=& 657.25 - 650.08 \\
=& 7.17
\end{aligned}
$$

Estimation

$$
\begin{aligned}
\bar{Y}_{small} - \bar{Y}_{large} =& \frac{1}{n_{small}} \sum_{i \in \text{small}} Y_i - \frac{1}{n_{large}} \sum_{i \in \text{large}} Y_i \\
=& 657.25 - 650.08 \\
=& 7.17
\end{aligned}
$$

- **Question:** Is this a large difference in a real-world sense?

Estimation

$$\begin{aligned}\bar{Y}_{small} - \bar{Y}_{large} =& \frac{1}{n_{small}} \sum_{i \in \text{small}} Y_i - \frac{1}{n_{large}} \sum_{i \in \text{large}} Y_i \\ =& 657.25 - 650.08 \\ =& 7.17\end{aligned}$$

- **Question:** Is this a large difference in a real-world sense?
    - Standard deviation across districts = 19.05

Estimation

$$
\begin{aligned}
\bar{Y}_{small} - \bar{Y}_{large} =& \frac{1}{n_{small}} \sum_{i \in \text{small}} Y_i - \frac{1}{n_{large}} \sum_{i \in \text{large}} Y_i \\
=& 657.25 - 650.08 \\
=& 7.17
\end{aligned}
$$

- **Question:** Is this a large difference in a real-world sense?
    - Standard deviation across districts = 19.05
    - Difference between 60th and 75th percentiles of test score distribution is $667.6 - 659.4 = 8.2$

Estimation

$$\begin{aligned}\bar{Y}_{small} - \bar{Y}_{large} =& \frac{1}{n_{small}} \sum_{i \in \text{small}} Y_i - \frac{1}{n_{large}} \sum_{i \in \text{large}} Y_i \\ =& 657.25 - 650.08 \\ =& 7.17\end{aligned}$$

- **Question:** Is this a large difference in a real-world sense?
    - Standard deviation across districts $= 19.05$
    - 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

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

$$t = \frac{\bar{Y}_{small} - \bar{Y}_{large}}{\sqrt{\frac{s^2_{small}}{n_{small}} + \frac{s^2_{large}}{n_{large}}}} = \frac{\bar{Y}_{small} - \bar{Y}_{large}}{SE(Y_{small} - Y_{large})}$$

where $SE(Y_{small} - Y_{large})$ is the standard error of $\bar{Y}_{small} - \bar{Y}_{large}$ and

$$s_{small} = \frac{1}{n_{small} - 1} \sum_{i \in small} (Y_i - \bar{Y})^2, \quad s_{large} = \frac{1}{n_{large} - 1} \sum_{i \in large} (Y_i - \bar{Y})^2.$$

Compute the difference-of-means t-statistic:

		testscore	
str	n	mean	sd
large	239	657.25	19.39
small	181	650.08	17.85
All	420	654.16	19.05

$$t = \frac{\bar{Y}_{small} - \bar{Y}_{large}}{SE(Y_{small} - Y_{large})} = \frac{657.25 - 650.08}{\sqrt{\frac{19.39^2}{239} + \frac{17.85^2}{182}}} = \frac{7.17}{1.82} = 3.93$$

t-test

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

Confidence interval

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

$$(\bar{Y}_{small} - \bar{Y}_{large}) \pm 1.96 \times SE(\bar{Y}_{small} - \bar{Y}_{large}) = 7.17 \pm 1.96 \times 1.82 = (3.6, 10.7)$$

- Two equivalent statements:
  1. The 95% confidence interval for $\bar{Y}_{small} - \bar{Y}_{large}$ doesn't include 0;
  2. The null hypothesis that $\bar{Y}_{small} - \bar{Y}_{large} = 0$ vs. a dual sided alternative is rejected at the 5% significance level.

What comes next...

- The mechanics of estimation, hypothesis testing, and confidence intervals should be familiar
- These concepts extend directly to regression and its variants
- Before turning to regression, however, we will review some of the underlying theory of estimation, hypothesis testing, and confidence intervals:
    - why do these procedures work, and why use these rather than others?
    - So we will review the intellectual foundations of statistics and econometrics

Review of Statistical Theory

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

The probability framework for statistical inference

- Population, sample
- Random variable, and distribution
- Moments of a distribution (mean, variance, standard deviation, covariance, correlation)
- Conditional distributions and conditional means
- Distribution of a sample of data drawn randomly from a population: $Y_1, \ldots, Y_n$

Population and sample

Population

- The group or collection of all possible entities of interest (school districts)
- We will think of populations as infinitely large (∞ is an approximation to "very big")

Sample

A sample is a **subset** selected from the population

Population and sample

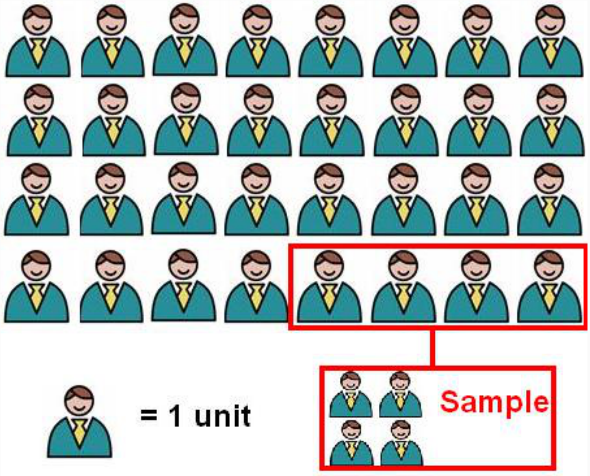


Figure 3: Population and sample

Random variables and probability distributions

Random variable X

- Numerical summary of a random outcome (district average test score, district str)

Random variables and probability distributions

Random variable X

- Numerical summary of a random outcome (district average test score, district str)

Probability distribution of X

- The probabilities of different values of Y that occur in the population, for ex. $\Pr[X = 650]$ (when X is <u>discrete</u>)
- or: The probabilities of sets of these values, for ex. $\Pr[640 \leq Y \leq 660]$ (when X is <u>continuous</u>)
    - in this case the probability is expressed through probability density fincion (p.d.f.)

Probability distribution

If X is continuous, the probability of X is expressed as

$$\Pr[a \leq X \leq b] = \int_a^b f(x)dx,$$

where $f(x)$ is the p.d.f. of X.

Notation

If the random variable X has a normal distribution, we say write

$$X \sim N(\mu, \sigma^2).$$

Probability distribution

A very important distribution is the normal (or Gaussian) distribution. The normal distribution has a bell-shaped p.d.f. which is formally given by:

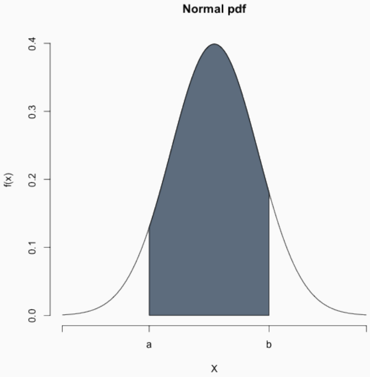
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where μ and σ are parameters that we will see have an important interpretation.

Probability distribution

The probability is the area under the bell shaped p.d.f.

$$\Pr[a \leq X \leq b] = \int_a^b f(x)dx$$



Probability distribution

An other important distribution is the chi-squared distribution:

$$f(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, & x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

where

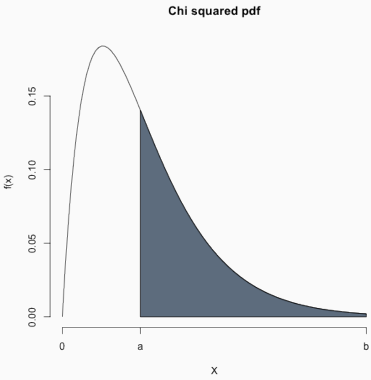
- $\Gamma(\cdot)$ is a complicated function(called Gamma function)
- ν is a parameter (this parameter indicated the "degrees of freedom" of the χ^2 distribution—we often say that $X \sim \chi^2_d$

Notation

If the random variable X has a chi-squared distribution with ν degrees of freedom, we write

$$X \sim \chi^2_\nu.$$

Probability distribution



The probability is the area under the p.d.f.

$$\Pr[a \leq X \leq b] = \int_a^b f(x)dx$$

Probability distribution

An other important distribution is the t-student distribution:

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where

- $\Gamma(\cdot)$ is a complicated function (called Gamma function)
- ν is a parameter (this parameter denotes the "degrees of freedom"—we often say that $X \sim t(\nu)$

Notation

If the random variable X has a t-student distribution with ν degrees of fredom, we write

$$X \sim t(\nu).$$

Probability distribution

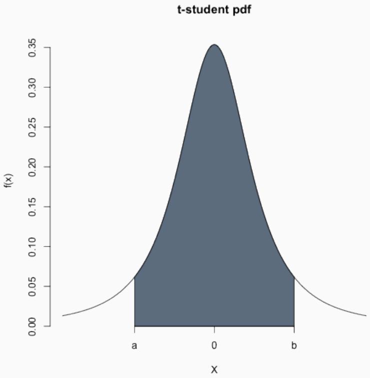


Figure 5:

The probability is the area under the p.d.f.

$\int^{b}$

Moments of a population distribution

mean (long-run average value of Y over repeated realizations)

$$E(X) := \int x f(x) dx$$

The shorthand for the expected value of a r.v. X is μ_X.

Moments of a population distribution

mean (long-run average value of Y over repeated realizations)

$$E(X) := \int x f(x) dx$$

The shorthand for the expected value of a r.v. X is μ_X.

variance (measure of the squared spread of the distribution)

$$E(X - \mu_X)^2 := \int (x - \mu_x)^2 f(x) dx$$

The shorthand for the variance of a r.v. X is σ_X^2.

Moments, cont'd

skewness (measure of asymmetry of a distribution)

$$\frac{E[(Y-\mu_Y)^3]}{\sigma_Y^3}$$

- skewness $= 0$: distribution is symmetric
- skewness $>$ ($<$) 0: distribution has long right (left) tail

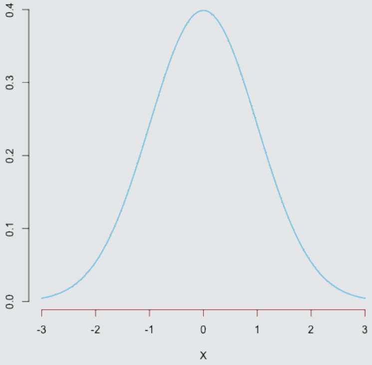
kurtosis (measure of mass in tails)

$$\frac{E[(Y-\mu_Y)^4]}{\sigma_Y^4}$$

- kurtosis $= 3$: normal distribution
- skewness ¿ 3: heavy tails ("leptokurtotic")

Moments, cont'd

$X \sim N(0,1)$

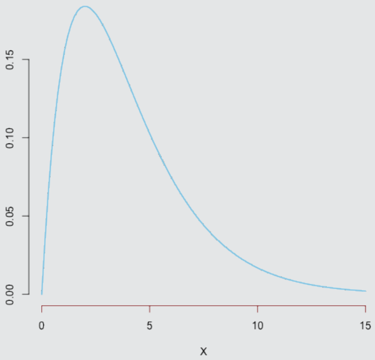


Moments

$$\begin{aligned}
\mu_X &= E(X) = 0 \\
\sigma_X^2 &= \mathrm{Var}(X) = 1 \\
\sigma_X &= \sqrt{\mathrm{Var}(X)} = 1 \\
\mathrm{skew}(X) &= \frac{E(X-\mu_X)^3}{\sigma_X^3} = 0 \\
\mathrm{kurt}(X) &= \frac{E(X-\mu_X)^4}{\sigma_X^4} = 3
\end{aligned}$$

Moments, cont'd

$X \sim \chi^2_\nu$

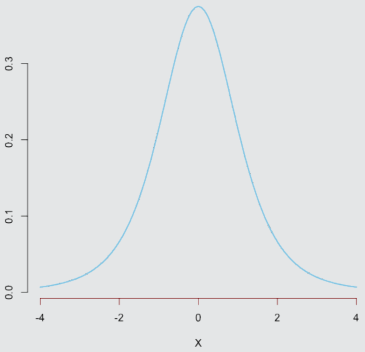


Moments

$$\begin{aligned}
\mu_X &= E(X) = \nu \\
\sigma_X^2 &= \mathrm{Var}(X) = 2\nu \\
\sigma_X &= \sqrt{\mathrm{Var}(X)} = \sqrt{2\nu} \\
\mathrm{skew}(X) &= \frac{E(X-\mu_X)^3}{\sigma_X^3} = \sqrt{8/\nu} \\
\mathrm{kurt}(X) &= \frac{E(X-\mu_X)^4}{\sigma_X^4} = 12/\nu
\end{aligned}$$

Moments, cont'd

$X \sim t(\nu)$



Moments

$$
\begin{aligned}
\mu_X =& E(X) = 0, \text{ if } \nu > 1 \\
\sigma_X^2 =& \mathrm{Var}(X) = \nu/(\nu - 2) \\
\sigma_X =& \sqrt{\mathrm{Var}(X)} = \sqrt{\nu/(\nu - 2)} \\
\mathrm{skew}(X) =& \frac{E(X - \mu_X)^3}{\sigma_X^3} = 0 \\
\mathrm{kurt}(X) =& \frac{E(X - \mu_X)^4}{\sigma_X^4} = 6/(\nu - 4)
\end{aligned}
$$

Random variables: joint distributions and covariance

- Random variables X and Z have a joint distribution
- The covariance between X and Z is

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

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

$$cov(X,X) = E[(X-\mu_X)(X-\mu_X)] = E[(X-\mu_X)^2]$$

Conditional distributions and conditional means

Conditional distributions

The distribution of Y, given value(s) of some other random variable, X

Conditional expectations and conditional moments

- conditional mean = mean of conditional distribution

$$E(Y|X = x) = \int yf(y|X = x)dy$$

- conditional variance = variance of conditional distribution

$$Var(Y|X = x) = \int yf(y|X = x)dy$$

Example (Example:)

$E(Testscores|STR < 20)$ = the mean of test scores among districts with small class sizes

Difference in (conditional) mean

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

$$\Delta = E[testscore|str < 20] - E[testscore|str \geq 20]$$

Other examples of conditional means:

- Wages of all female workers (Y = wages, X = gender)
- Mortality rate of those given an experimental treatment (Y live/die; X = treated/not treated)

Important fact: mean independence

Take two random variables, say U and X. Then is

$$E[U|X = x] = \text{constant}, \quad \text{for all } x$$

then

$$cov(U, X) = 0, \quad E[U] = \text{constant}.$$

We say in this case that U is conditional mean independent from X.

Notice that, $cov(X, U) = 0$ does not imply $E[U|X] = \text{constant}$.

Distribution of a sample drawn randomly from a population

Let Y denote a variable of interest, for instance

$$Y = \{\text{net wage of italian full time employees}\}$$

.

Think of $(Y_1, Y_2, \ldots, Y_n)$ as the collection of wages of n workers drawn from the population

- **Prior to sample selection**, the wages $(Y_1, \ldots, Y_n)$ are random variables because the workers are randomly selected
- **Once the worker is selected** and the value of Y is observed, then $(Y_1, \ldots, Y_n)$ are just an array of numbers - not random

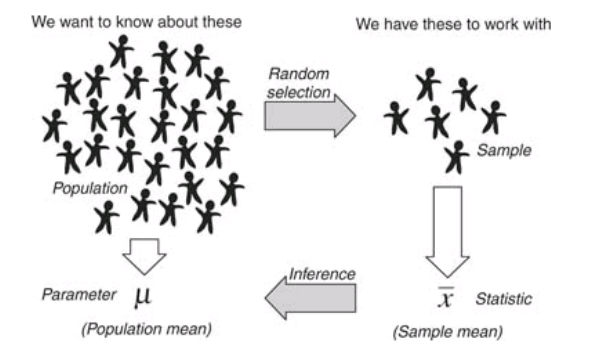
Simple random sampling (or iid)

We will assume simple random sampling, that is, entities (district, entity) are drawn at random from the population.

In this case we will say that $(Y_1, \ldots, Y_n)$ is a family of *independent, identically distributed* (i.i.d.) random variables.

- Y_j and Y_k are independent, that is, the value Y_j has no information content for Y_k (independently)
- The probability distribution of each r.v. is the same (identically)

Sampling distribution



Framework

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

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

Estimation

$\bar{Y}$ is the natural estimator of the expected value of Y, μ_Y. But:

1. What are the properties of $\bar{Y}$?
2. Why should we use $\bar{Y}$ rather than some other estimator?
    - y_1 (the first observation)
    - maybe unequal weights - not simple average
    - $median(Y_1, ..., Y_n)$

The sampling distribution of $\bar{Y}$

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

- The sample is i.i.d.

The sampling distribution of $\bar{Y}$

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

- The sample is i.i.d.
- Thus the values of $(Y_1, \ldots, Y_n)$ are random

The sampling distribution of $\bar{Y}$

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

- The sample is i.i.d.
- Thus the values of $(Y_1, \ldots, Y_n)$ are random
- Thus functions of $(Y_1, \ldots, Y_n)$, such as $\bar{Y}$, are random:
  - had a different sample been drawn, they would have taken on a different value

The sampling distribution of $\bar{Y}$

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

- The sample is i.i.d.
- Thus the values of $(Y_1, \ldots, Y_n)$ are random
- Thus functions of $(Y_1, \ldots, Y_n)$, 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 sampling distribution of $\bar{Y}$

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

- The sample is i.i.d.
- Thus the values of $(Y_1, \ldots, Y_n)$ are random
- Thus functions of $(Y_1, \ldots, Y_n)$, 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 $var(\bar{Y})$.

The sampling distribution of $\bar{Y}$

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

- The sample is i.i.d.
- Thus the values of $(Y_1, \ldots, Y_n)$ are random
- Thus functions of $(Y_1, \ldots, Y_n)$, 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 $var(\bar{Y})$.
- The concept of the sampling distribution underpins all of econometrics.

Example: Bernoulli distribution

Suppose Y takes on 0 or 1 (a Bernoulli random variable) with

$$Y = \begin{cases} 0 & p = .22 \\ 1 & p = .78 \end{cases}$$

Then

$$E(Y) = p \times 1 + (1-p) \times 0 = p = .78$$

and

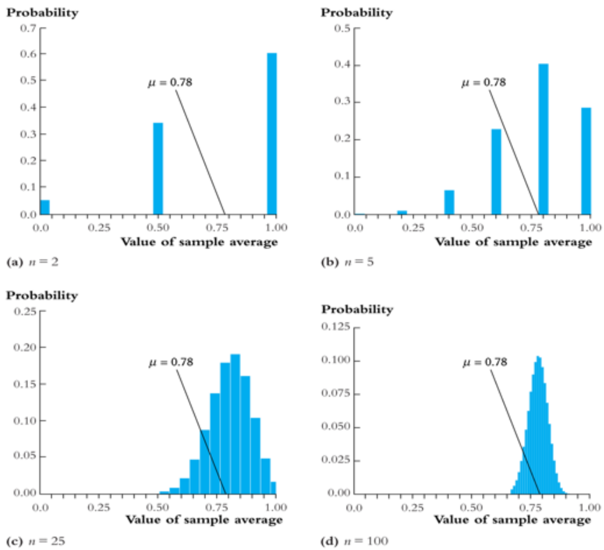
$$\sigma_Y^2 = E[Y - E(Y)]^2 = p(1-p) = .78 \times (1 - .78) = 0.1716$$

The sampling distribution of $\bar{Y}$ depends on n.

Consider n = 2. The sampling distribution of $\bar{Y}$ is,

- $\Pr(\bar{Y} = 0) = .22^2 = .0484$
- $\Pr(\bar{Y} = 1/2) = 2 \times .22 \times .78 = .3432$
- $\Pr(\bar{Y} = 1) = .78^2 = .6084$

The sampling distribution of $\bar{Y}$ when Y is Bernoulli ($p = .78$):



Things we want to know about the sampling distribution:

- What is the mean of $\bar{Y}$?
    - If $E(\bar{Y}) = \mu = .78$, then $\bar{Y}$ is an unbiased estimator of μ

Things we want to know about the sampling distribution:

- What is the mean of $\bar{Y}$?
  - If $E(\bar{Y}) = \mu = .78$, then $\bar{Y}$ is an unbiased estimator of μ
- What is the variance of $\bar{Y}$?
  - How does $var(\bar{Y})$ depend on n?
  - Does $\bar{Y}$ become close to μ when n is large?
    - Law of large numbers: $\bar{Y}$ is a consistent estimator of μ?

Things we want to know about the sampling distribution:

- What is the mean of $\bar{Y}$?
  - If $E(\bar{Y}) = \mu = .78$, then $\bar{Y}$ is an unbiased estimator of μ
- What is the variance of $\bar{Y}$?
  - How does $var(\bar{Y})$ depend on n?
  - 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 $\bar{Y}$

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

and

$$var(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

Implications:

1. $\bar{Y}$ is an unbiased estimator of μ_Y, (that is, $E(\bar{Y}) = \mu_Y$)
2. $var(\bar{Y})$ is inversely proportional to n
    - the spread of the sampling distribution is proportional to $1/n$
    - Thus the sampling uncertainty associated with is proportional to 1/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 becomes more tightly centered around μ_Y (the Law of Large Numbers)
2. Moreover, the distribution of $\bar{Y} - \mu_Y$ becomes normal (the Central Limit Theorem)

The Law of Large Numbers (LLN)

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.

Theorem (LLN)

If $(Y1, \ldots, Yn)$ are i.i.d. and $\sigma_Y^2 < \infty$, then $\bar{Y}$ is a consistent estimator of μ_Y, that is,

$$\Pr[|\bar{Y} - \mu_Y| < \epsilon] \to 1 \textit{ as } n \to \infty$$

which can be written, $\bar{Y} \xrightarrow{p} \mu_Y$

The Central Limit Theorem (CLT):

If $(Y_1, \ldots, Y_n)$ 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^2/n)

The Central Limit Theorem (CLT):

If $(Y_1, \ldots, Y_n)$ 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^2/n)
- $\sqrt{n}(\bar{Y} - \mu_Y)/\sigma_Y$ is approximately distributed $N(0, 1)$ (standard normal)

The Central Limit Theorem (CLT):

If $(Y_1, \ldots, Y_n)$ 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^2/n)
- $\sqrt{n}(\bar{Y} - \mu_Y)/\sigma_Y$ is approximately distributed $N(0, 1)$ (standard normal)
- $\sqrt{n}(\bar{Y} - \mu_Y)/s_Y$ is approximately distributed $N(0, 1)$ (standard normal)

The Central Limit Theorem (CLT):

If $(Y_1, \ldots, Y_n)$ 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^2/n)
- $\sqrt{n}(\bar{Y} - \mu_Y)/\sigma_Y$ is approximately distributed $N(0, 1)$ (standard normal)
- $\sqrt{n}(\bar{Y} - \mu_Y)/s_Y$ is approximately distributed $N(0, 1)$ (standard normal)
- The larger is n, the better are these approximations.

Summary: The Sampling Distribution of $\bar{Y}$

For $Y_1, \ldots, Y_n$ i.i.d. with $0 < \sigma_Y^2 < \infty$

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

Why use $\bar{Y}$ to estimate μ_Y?

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

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

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

Set derivative to zero and denote optimal value of m by

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

Setting the derivative to zero $m = \frac{1}{n} \sum_{i=1} Y_i = \bar{Y}$.

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

- $\bar{Y}$ has a smaller variance than all other linear unbiased estimators:

Example

consider the estimator, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, where $\{a_i\}$ are such that $\bar{\mu}$ is unbiased;

- then $\mathrm{var}(\hat{\mu}) \geq \mathrm{var}(\bar{Y})$

Estimator of the variance of Y

A good estimator of σ_Y^2 is the sample variance of Y

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

Facts:

If $(Y_1, \ldots, Y_n)$ are i.i.d. and $E(Y_4) < \infty$, then $s_Y^2 \xrightarrow{p} \sigma_Y^2$ and, also,

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

Why does the law of large numbers apply?

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

Actually:

population quantity	alternative notation	sample quantity
$E(Y)$	μ_Y	$\bar{Y}$
$\mathrm{Var}(Y)$	σ_Y^2	s_Y^2
$\sqrt{\mathrm{Var}(Y)}$	σ_Y	s_Y
$\mathrm{cov}(Y, X)$		s_{YX}
$\mathrm{corr}(Y, X)$		ρ_{XY}

Sample ¡-¿ Quantities

All these sample quantities are all "good" estimators of the population quantities, in the sense that they are all consistent.

Where are we?

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