Analysis of Economics Data
Chapter 3 The Sample Mean

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 3: The Sample Mean

- Now consider **statistical inference**
  - extrapolating from sample to population
  - here from sample mean $\bar{x}$ to population mean μ.
- Basic idea is that the sample values $x_1, ..., x_n$ (lower case)
  - are realizations of random variables $X_1, ..., X_n$ (upper case)
- So the **sample mean** $\bar{x} = (x_1 + \cdots + x_n)/n$
  - **is a realization of the random variable** $\bar{X} = (X_1 + \cdots + X_n)/n$
- This chapter: distribution of $\bar{X}$ from underlying distribution of X.
- Next chapter: The two main tools of statistical inference
  - Confidence intervals for the population mean μ
  - Hypothesis tests on μ.

Outline

1. Random Variables
2. Sample Generated by an Experiment
3. Random Samples
4. Properties of the Sample Mean
5. Sampling from a Finite Population
6. Estimation of the Population Mean
7. Nonrepresentative Samples
8. Computer Generation of a Random Sample

- Datasets: COINTOSSMEANS, CENSUSAGEMEANS

3.1 Random Variables

- A **random variable** is a variable whose value is determined by the outcome of an experiment.
- An **experiment** is an operation whose outcome cannot be predicted with certainty.
- Example: the experiment is tossing a coin and the random variable takes value 1 if heads and 0 if tails.
- Example: the experiment is randomly selecting a person from the population and the associated random variable takes value equal to their annual earnings.
- **Standard notation**
  - X (or Y or Z) denotes a **random variable**
  - x (or y or z) denotes the **values taken** by X (or Y or Z).

Example: Coin toss

- Simplest case is a random variable that takes one of only two possible values.
- Consider toss of fair coin with $X = 1$ if heads and $X = 0$ if tails. Then

$$X = \begin{cases} 0 & \text{with probability } 0.5 \\ 1 & \text{with probability } 0.5. \end{cases}$$

Mean of a Random Variable

- **Mean** of X, denoted μ or μ_X
  - is the **probability-weighted average** of all possible values of X in the population.
- μ is also denoted $\mathrm{E}[X]$
  - the **expected value** of the random variable X
  - the long-run average value expected if we draw a value of X at random, draw a second value of X at random, and so on, and then obtain the average of these values.

$$\begin{aligned} \mu \equiv \mathrm{E}[X] &= x_1 \times \Pr[X = x_1] + x_2 \times \Pr[X = x_2] + \cdots \\ &= \textstyle\sum_x x \cdot \Pr[X = x]. \end{aligned}$$

- Note that
  - $\sum_x$ means the sum over all possible values x can take
  - and the possible values of x are denoted x_1, x_2, x_3,...

Example of Mean

- Fair coin toss: X takes values 0 or 1 with equal probabilities

$$\begin{aligned}\mu &= \textstyle\sum_x x \times \Pr[X = x] \\ &= \Pr[X = 0] \times 0 + \Pr[X = 1] \times 1 \\ &= 0.5 \times 0 + 0.5 \times 1 \\ &= 0.5.\end{aligned}$$

- Unfair coin: $X = 1$ with probability 0.6 and $X = 0$ with probability 0.4
  - $\mu = 0 \times 0.4 + 1 \times 0.6 = 0.6.$

Variance and Standard Deviation

- **Variance** σ^2
  - measures the variability in X around μ
  - equals the expected value of $(X-\mu)^2$, the squared deviation of X from the mean μ
  - probability-weighted average of x_1^*, x_2^*, ...

$$\begin{aligned}\sigma^2 &\equiv \mathrm{E}[(X-\mu)^2] \\ &= (x_1-\mu)^2 \times \Pr[X=x_1] + (x_2-\mu)^2 \times \Pr[X=x_2] + \cdots \\ &= \sum_x (x-\mu)^2 \times \Pr[X=x].\end{aligned}$$

- **Population standard deviation** σ is square root of the variance
  - measured in the same units as X.

Example of Variance and Standard Deviation

- Fair coin toss: X takes values 0 or 1 with equal probabilities so $\mu = 0.5$.
- Variance

$$\begin{aligned} \sigma^2 &= \sum_x (x-\mu)^2 \times \Pr[X=x] \\ &= \Pr(0-0.5)^2 \times [X=0] + (1-0.5)^2 \times \Pr[X=1] \\ &= 0.25 \times 0.5 + 0.25 \times 0.5 \\ &= 0.25. \end{aligned}$$

- Standard deviation

$$\sigma = \sqrt{0.25} \simeq 0.5.$$

3.2 Random Samples

- A sample of size n takes values denoted $x_1, ..., x_n$.
- These values are realizations or outcomes of the random variables $X_1, X_2, ..., X_n$.
- Example: four consecutive coin tosses with results tails, heads, heads and heads
  - random variable X_1 has realized value $x_1 = 0$
  - random variable X_2 takes value $x_2 = 1$
  - random variable X_3 takes value $x_3 = 1$
  - random variable X_4 takes value $x_4 = 1$.

Sample Mean is a Random Variable

- **Sample** of size n has observed values $x_1, x_2, ..., x_n$.
  - These are realizations of the random variables $X_1, X_2, ..., X_n$.
- **Sample mean** is the average

$$\bar{x} = (x_1 + x_2 + \cdots + x_n)/n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- This is a realization of the **random variable**

$$\bar{X} = (X_1 + X_2 + \cdots + X_n)/n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Aside: Sample Variance and Standard Deviation

- Similarly any other sample statistic (such as the median) is a realization of a random variable
- In addition to the sample mean we focus on the sample variance and sample standard deviation.
- **Sample variance** is average of squared deviations of x around $\bar{x}$
  - not around μ since μ is unknown

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

- The sample variance is a realization of the **random variable**

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

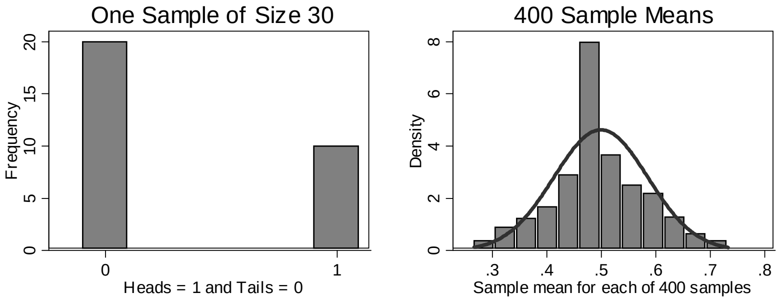
- Taking the square root gives the **sample standard deviation** s which is a realization of the random variable S.

3.3 Sample Generated from an Experiment: Coin Tosses

- We consider a simple experiment that generates many samples
  - hence many sample means $\bar{x}$
  - then summarize the resulting distribution of the many $\bar{x}$.
- **Population:** Outcomes from experiment of tossing a coin
  - $X = 1$ if heads and $X = 0$ if tails
  - Population mean $\mu = \mathrm{E}[X] = 0.5$ and standard deviation $\sigma = 0.5$.
- **Sample:** $n = 30$
  - random sample of size 30 from 30 coin tosses
  - there are 10 heads and 20 tails, so $\bar{x} = 10/30 = 0.333$
  - histogram of this single sample is given in left panel of next slide.

Example: Coin Tosses (continued)

- Left panel: x's from 1 sample of size 30 with 20 heads and 10 tails
- Right panel: $\bar{x}'s$ for 400 samples of size 30



Example: Coin Tosses (continued)

- Randomly draw 400 different samples, each of size 30
  - then $\bar{x}_1 = .333$, $\bar{x}_2 = .500$, $\bar{x}_3 = 533$,....
- Histogram (plus kernel density estimate) for the 400 means from the 400 samples of size 30 is given in right panel of previous slide.
  - roughly centered on the population mean
    - the average of the 400 means is 0.499, close to $\mu = 0.5$.
  - much less variability in these 400 means than in the original population
    - the standard deviation of the 400 means is 0.086
    - much less than the population standard deviation of $\sigma = 0.5$
  - the density estimate is roughly that of the normal.

3.4 Properties of the Sample Mean

- The properties of $\bar{X}$ depend on the properties of $X_1, X_2, ..., X_n$
  - such as the means and variances of $X_1, X_2, ..., X_n$
  - and whether their values depend in part on other values.
- In this chapter we consider the simplest and standard set of assumptions in introductory statistics
  - $X_1, X_2, ..., X_n$ have common mean μ and common variance σ^2
  - $X_1, X_2, ..., X_n$ are statistically independent
    - ★ statistical independence means that the value taken by X_2, for example, is not influenced by the value taken by $X_1, X_3, ..., X_n$.
- In later chapters we relax these assumptions
  - e.g. regression allows for different means for different observations.

Population Assumptions

- **Population**
  - $=$ set of all observations (or experimental outcomes).
- **Sample**
  - $=$ **subset** selected from the population.
- Properties of $\bar{x}$ depend on the random variable $\bar{X}$
  - hence on assumptions about process generating $X_1, X_2, ..., X_n$.
- We assume a **simple random sample** where
  - **A.** X_i has **common mean** μ : $\mathrm{E}[X_i] = \mu$ for all i.
  - **B.** X_i has **common variance** σ^2 : $\mathrm{Var}[X_i] = \sigma^2$ for all i.
  - **C.** X_i is **statistically independent** of $X_j, i \neq j$.
- Shorthand notation: $X_i \sim (\mu, \sigma^2)$
  - means X_i are distributed with mean μ and variance σ^2.

Mean and Variance of the Sample Mean

- Consider $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ for $X_i \sim (\mu, \sigma^2)$.
- The **(population) mean of the sample mean** is

$$\mu_{\bar{X}} \equiv \mathrm{E}[\bar{X}] = \mu.$$

- The **(population) variance of the sample mean** is

$$\sigma^2_{\bar{X}} \equiv \mathrm{E}[(\bar{X} - \mu_{\bar{X}})^2] = \frac{\sigma^2}{n},$$

- The **(population) standard deviation** is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.
- Sample mean is less variable than the underlying data
  - since $\sigma^2_{\bar{X}} < \sigma^2$.
- Sample mean is close to μ as $n \to \infty$
  - since $\mathrm{E}[\bar{X}] = \mu$ and variance $\sigma^2_{\bar{X}} = \sigma^2/n \to 0$ as $n \to \infty$.

Aside: Proof for Mean of the Sample Mean

- Recall
$$\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$$
- Proof uses
  - $\mathrm{E}[aX] = a\mathrm{E}[X]$
  - $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$

  and assumption A (common mean of X_i).
- Then
$$\begin{aligned} \mathrm{E}[\bar{X}] &= \mathrm{E}[\tfrac{1}{n}(X_1 + X_2 + \cdots + X_n)] \\ &= \tfrac{1}{n}\mathrm{E}[X_1 + X_2 + \cdots + X_n] \\ &= \tfrac{1}{n}\{\mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n]\} \\ &= \tfrac{1}{n}\{\mu + \mu + \cdots + \mu\} \\ &= \mu. \end{aligned}$$

Aside: Variance of the Population Mean

- Proof in Appendix B.2 uses that
  - Var$[aX] = a^2$E$[X]$ in general
  - Var$[X+Y]$ =Var$[X]$+Var$[Y]$ for independent variables
  - and assumptions A-C.
- Then

$$\begin{aligned}
\text{Var}[\bar{X}] &= \text{Var}\left[\tfrac{1}{n}(X_1+X_2+\ldots+X_n)\right] \\
&= (\tfrac{1}{n})^2\text{Var}\left[X_1+X_2+\ldots+X_n\right] \\
&= \left(\tfrac{1}{n}\right)^2\left\{\text{Var}\left[X_1\right]+\cdots+\text{Var}\left[X_n\right]\right\} \\
&= \left(\tfrac{1}{n}\right)^2\sigma^2+\cdots+\left(\tfrac{1}{n}\right)^2\sigma^2 \\
&= \left(\tfrac{1}{n}\right)^2\left\{\sigma^2+\cdots+\sigma^2\right\} \\
&= \left(\tfrac{1}{n}\right)^2\times n\sigma^2 \\
&= \tfrac{1}{n}\sigma^2.
\end{aligned}$$

Normal Distribution and the Central Limit Theorem

- We have shown to date that $\bar{X} \sim (\mu, \sigma^2/n)$
- In general, subtracting the mean and dividing by the standard deviation yields a random variable with mean 0 and variance 1.
- So here the standardized variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim (0, 1).$$

- The central limit theorem (a remarkable result) proves normality as the sample size gets large

$$Z \sim N(0, 1) \text{ as } n \rightarrow \infty.$$

- The central limit theorem holds under assumptions A-C
  - and also under some weaker conditions.

Normal Distribution (continued)

- Now convert back to the original $\bar{X}$.
- We have
$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1) \text{ as } n \to \infty.$$
- Then $\bar{X}$ is approximately normally distributed in large samples
$$\bar{X} \sim N(\mu, \sigma^2 / n) \text{ approximately for large } n.$$
- We will use this result to do statistical inference on μ.
- However, the variance σ^2 / n is unknown as σ^2 is unknown
  - we will have to get an estimate
  - replace σ^2 by its estimate s^2
  - where s is the sample standard deviation of X.

Standard Error of the Sample Mean

- **Estimated variance** of $\bar{X}$ is

$$s_{\bar{X}}^2 = \frac{s^2}{n} = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}{n},$$

- **Estimated standard deviation** of $\bar{X}$

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}}{\sqrt{n}}.$$

- $s_{\bar{X}}$ is called the **standard error of the sample mean** $\bar{X}$.
- The term "standard error" means estimated standard deviation
  - various estimators each have a distinct standard error
  - a reported "standard error" in computer output need not be $s_{\bar{X}}$.
- Use the notation

$$se(\bar{X}) = s/\sqrt{n}.$$

Summary for the Sample Mean

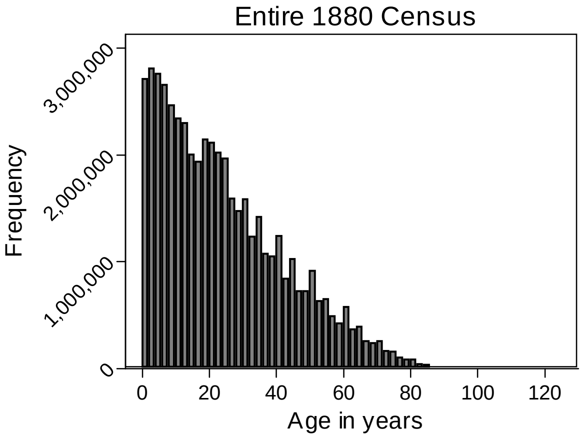
1. Sample values $x_1, ..., x_n$ are observed values of the random variables $X_1, ..., X_n$.
2. Individual X_i have common mean μ and variance σ^2 and are independent.
3. Average $\bar{X}$ of n draws of X_i has mean μ and variance σ^2/n.
4. Standardized statistic $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim (0, 1)$ has mean 0 and variance 1.
5. Z is standard normal as size $n \longrightarrow \infty$ by the central limit theorem.
6. For large n a good approximation is that $\bar{X} \sim N(\mu, \sigma^2/n)$
7. The standard error of $\bar{X}$ equals $s/\sqrt{n}$, where "standard error" is general terminology for "estimated standard deviation".

3.5 Sampling from a Population: 1880 Census

- Now consider an example of sampling from a population.
- **Population:** $N = 50{,}169{,}452$
  - all people recorded as living in the U.S. in 1880
  - the average age is 24.13 years, so $\mu = \mathbf{24.13}$
  - the standard deviation of age is 18.61, so $\sigma = \mathbf{18.61}$
  - histogram is given in the next slide.

Example: 1880 Census (continued)

- Population
  - Probabilities decline with age (clearly not the normal)
  - Peaks due to rounding at five and ten years

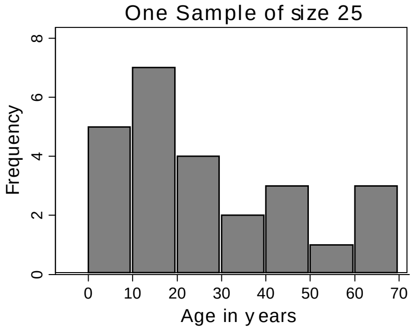


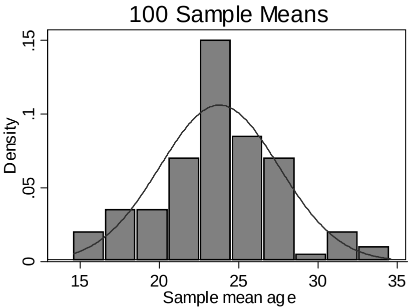
Example: 1880 Census (continued)

- **Single sample:** $n = 25$
  - random sample of size 25 from the entire U.S. population
  - the average age is 27.84, so $\bar{\mathbf{x}}$= **27.84**
  - the standard deviation of age is 20.71, so $\mathbf{s} =$ **20.71**
  - these are **similar to, but not exactly equal** to, μ and σ
  - histogram of $x's$ in a single sample is given in left panel of next slide.
- **Many samples of size 25**
  - randomly draw 100 different samples, each of size 25
  - then $\bar{x}_1 = 27.84$, $\bar{x}_2 = 19.40$, $\bar{x}_3 = 23.28$ years, ...
  - average of the 100 sample means is 23.78, close to $\mu = 24.13$.
  - standard deviation of the 100 means is 3.76, close to $\sigma/\sqrt{n} = 18.61/\sqrt{25} = 3.72$.
  - histogram of $\bar{x}'s$ across 100 samples is given in right panel of next slide.

Example: 1880 Census (continued)

- 100 different means from 100 different samples, each of size 25
  - histogram (left) and kernel density estimate (right)
  - looks like normal with mean μ and standard deviation much less than σ





3.6 Estimation of the Sample Mean

- Desire a good **point estimate** of population mean μ
  - why use $\bar{x}$ rather than some other estimate?
- A desirable estimator of μ has distribution
  - centered on μ
  - with as little variability around μ as possible.

Parameter, Estimator and Estimate

- A **parameter** is a constant that determines in part the distribution of X.
- An **estimator** is a method for estimating a parameter.
- An **estimate** is the particular value of the estimator obtained from the sample.
- For estimation of the mean of X using the sample mean
  - the parameter is μ
  - the estimator is the random variable $\bar{X}$
  - the estimate is the sample value $\overline{x}$.

Unbiased Estimators

- An **unbiased estimator** of a population parameter
  - has expected value that equals the population parameter.
- The sample mean is unbiased for μ
  - since $\mathrm{E}[\bar{X}] = \mu$.

Minimum Variance Estimators

- Other estimators may also be unbiased and consistent for μ
  - e.g. sample median in the case where X is symmetrically distributed
  - discriminate between such estimators using their variance.
- A **best estimator** or **efficient estimator**
  - has **minimum variance** among the class of consistent estimators (or of unbiased estimators).
- Under assumptions A-C the sample mean has variance σ^2/n
  - for X that is normal, Bernoulli, binomial or Poisson no other unbiased estimator has lower variance
  - for X with other distributions the sample mean is often close to having the lowest variance
  - generally the sample mean is used to estimate μ.

Consistent Estimators

- Consistency is a more advanced concept that considers behavior as the sample size goes to infinity.
- A **consistent estimator** of a population parameter
  - is one that is almost certainly arbitrarily close to the population parameter as the sample size gets very large.
- A sufficient condition for consistency is
  - any bias disappears as the sample size gets very large
  - the variance goes to zero as the sample size gets very large
- The sample mean is consistent for μ under assumptions A-C
  - it is unbiased
  - the variance $\sigma^2_{\bar{X}} = \sigma^2 / n \rightarrow 0$ as $n \rightarrow \infty$.

3.7 Samples other than Simple Random Samples

- Recall simple random sample means data are independent and from the same distribution.
- Representative Samples
  - Still from same distribution but no longer statistically independent.
  - Then can adapt methods using an alternative formula for $se(\bar{x})$.
- Nonrepresentative samples
  - Now different observations may have different μ
  - e.g. Survey readers of Golf Digest not representative of population.
  - Big problem.
- Weighted mean can still be used if population weights are known
  - π_i = probability that i^{th} observation is included in the sample.
  - sample weights $w_i = 1/\pi_i$
  - **weighted mean** $\bar{x}_w = \left[\sum_{i=1}^n w_i x_i\right] / \left[\sum_{i=1}^n w_i\right]$.

3.8 Computer Generation of a Random Variable

- A **(pseudo) uniform random number generator**
  - creates values between 0 and 1
  - any value between 0 and 1 is equally likely
  - successive values appear to be independent of each other.
- To simulate 30 coin tosses
  - draw 30 uniform random numbers
  - result is heads if the uniform random number exceeds 0.5
- For Census example
  - if uniform random number is between 0 and $1/N$, where $N =$ 50,169,452, we choose the first person, etcetera
- The sequence depends on the starting value called the **seed**
  - always set the seed (e.g. equal to 10101).

Example Stata Code to give 400 sample means

- The following advanced Stata code obtains the 400 sample means in the coin toss example of Chapter 3.2
  - the program generates one sample of size 30 of x equal 1 or 0
  - the simulate command does this 400 times
  - this gives 400 observations on variable `xbar`.

```
program onesample, rclass
  drop _all
  set obs 30
  generate u = runiform()
  generate x = u > 0.5
  summarize x
  return scalar xbar = r(mean)
end
simulate xbar=r(xbar), seed(10101) reps(400): onesample
summarize
```

Some in-class Exercises

1. Suppose $X = 100$ with probability 0.8 and $X = 600$ with probability 0.2. Find the mean, variance and standard deviation of X.
2. Consider random samples of size 25 from the random variable X that has mean 100 and variance 400. Give the mean, variance and standard deviation of the mean $\overline{X}$.