Review of Statistical Theory
Part 2

Dragos Ailoae
dailoae@gradcenter.cuny.edu

Advanced Economics and Business Statistics
ECON-4400w - Spring 2022

Brooklyn College
Feb 9, 2022

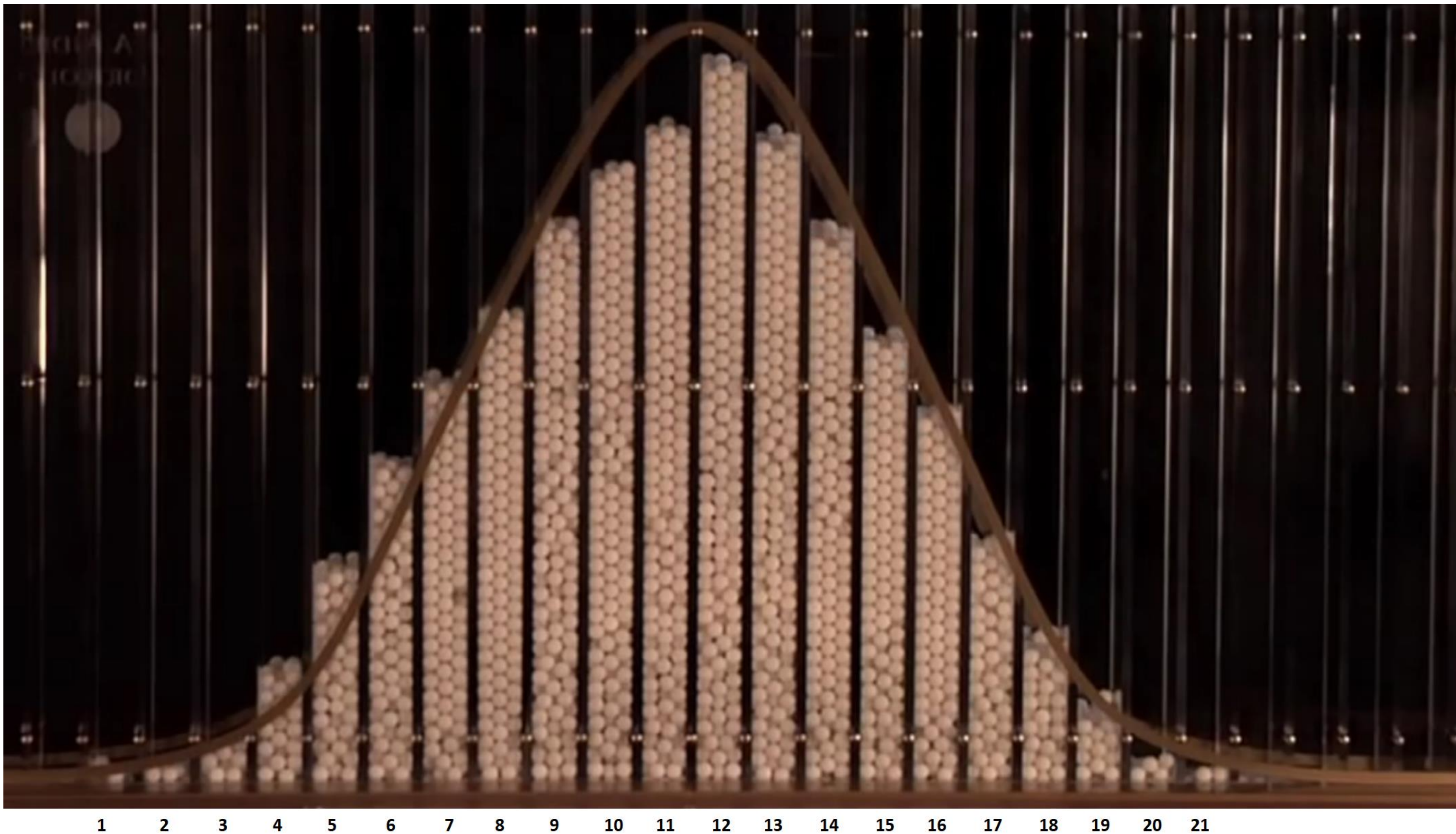
Review of Statistical Theory

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

<u>The probability framework for statistical inference</u>

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

Sampling



https://imgur.com/gallery/hfBzEWM

(d) Distribution of a sample of data drawn randomly from a population: $Y_1, \ldots, Y_n$

Population

- The group or collection of all possible entities of interest (school districts). We will think of populations as infinitely large

We will assume 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 $(Y_1, Y_2, \ldots, Y_n)$, where Y_i = value of Y for the i^{th} individual (district, entity) sampled

Distribution of $Y_1,\ldots, Y_n$ under 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 and 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, under simple random sampling, Y_1 and Y_2 are independently and identically distributed (***i.i.d.***).
  - More generally, under simple random sampling, $\{Y_i\}$, $i = 1,\ldots, n$, are i.i.d.

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 mean. But:

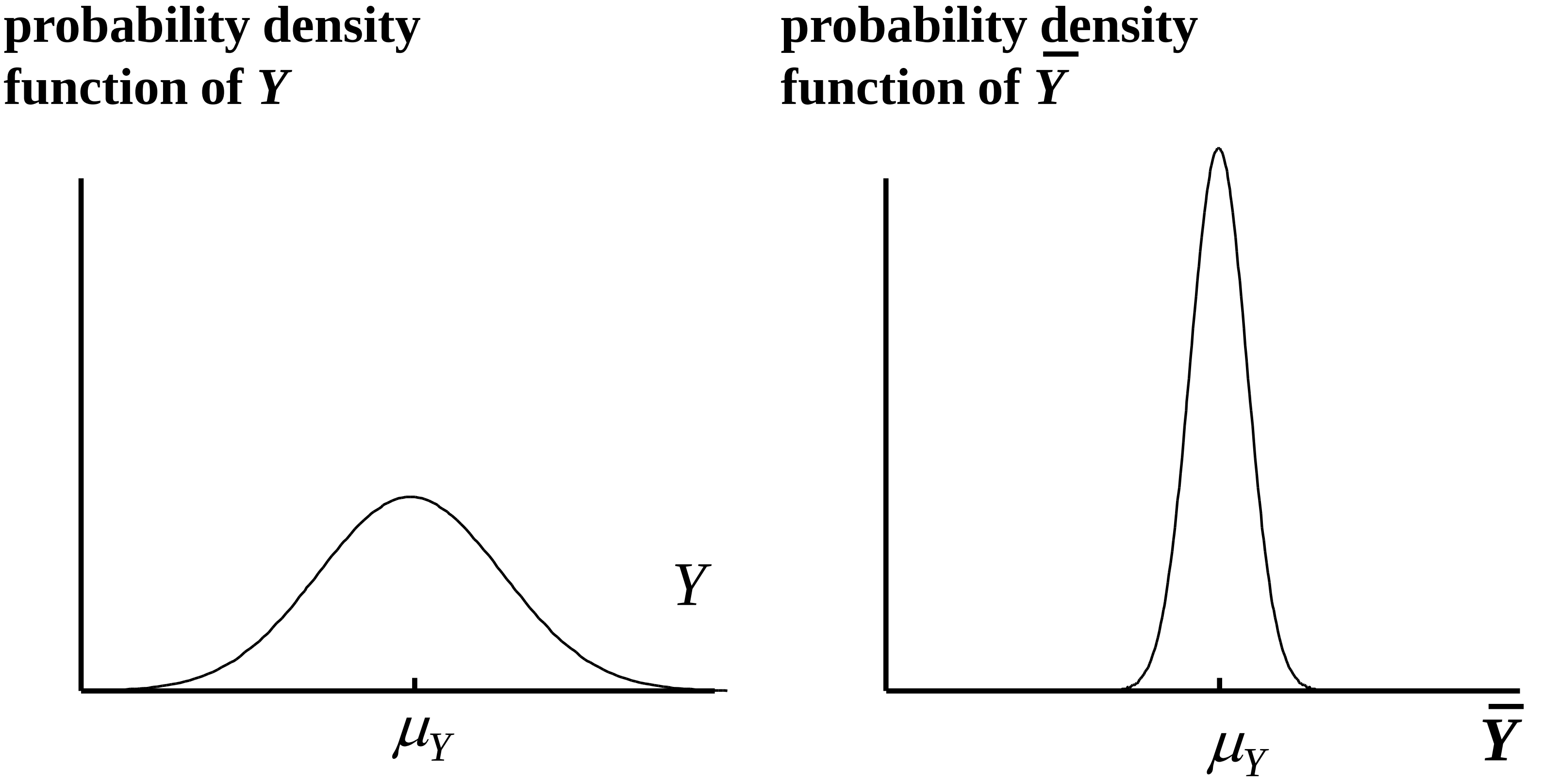
a) What are the properties of $\bar{Y}$?

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

- Y_1 (the first observation)
- maybe unequal weights – not simple average
- median($Y_1,\ldots, Y_n$)

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

Sampling and Estimators



In this illustration, Y and $\bar{Y}$ are both centered around μ_Y, but the dispersion differs.

(a) The sampling distribution of $\bar{Y}$

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

- The individuals in the sample are drawn at random.
- Thus the values of (Y_1, …, Y_n) are random
- Thus functions of (Y_1, …, 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.

The mean and variance of the sampling distribution of $\bar{Y}$ (1 of 3)

- General case – that is, for Y_i i.i.d. **from any distribution**

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

- Variance:

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

The mean and variance of the sampling distribution of $\bar{Y}$ (2 of 3)

so $$\text{var}(\bar{Y}) = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)\right]^2$$

$$= E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)\right] \times \left[\frac{1}{n}\sum_{j=1}^{n}(Y_j - \mu_Y)\right]\right\}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[(Y_i - \mu_Y)(Y_j - \mu_Y)\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{cov}(Y_i, Y_j)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2$$

$$= \frac{\sigma_Y^2}{n}$$

The mean and variance of the sampling distribution of $\bar{Y}$ (3 of 3)

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

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

Implications:

1. $\bar{Y}$ is an *unbiased* estimator of μ_Y (that is, $E(\bar{Y}) = \mu_Y$)
2. $\text{var}(\bar{Y})$ 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 $\bar{Y}$ is proportional to $1/\sqrt{n}$ (larger samples, less uncertainty, but square-root law)

Things we want to know about the sampling distribution:

- What is the mean of $\bar{Y}$?
  - If $E(\bar{Y}) = \mu$, 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)

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 $\bar{Y} - \mu_Y$ becomes normal (the *Central Limit Theorem*)

The *Law of Large Numbers*:

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.

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

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

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

("$\bar{Y} \xrightarrow{p} \mu_Y$" means "$\bar{Y}$ converges in probability to μ_Y").

(*the math*: as $n \to \infty$, $\text{var}(\bar{Y}) = \dfrac{\sigma_Y^2}{n} \to 0$, which implies that $\Pr[|\bar{Y} - \mu_Y| < \varepsilon] \to 1$.)

The *Central Limit Theorem* (CLT) (1 of 3)

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_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 larger is *n*, the better is the approximation.**

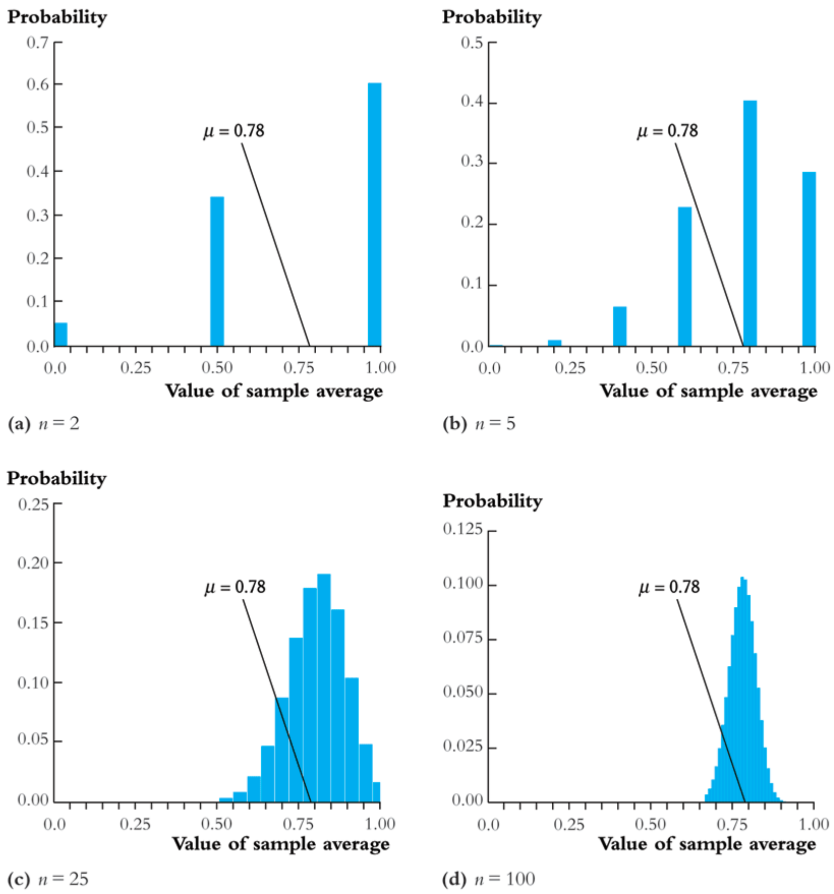
Summary: The Sampling Distribution of $\bar{Y}$

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

- The exact (finite sample) sampling distribution of $\bar{Y}$ 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 (the population distribution)
- When n is large, the sampling distribution simplifies:
  - $\bar{Y} \xrightarrow{p} \mu Y$ (Law of large numbers)
  - $\dfrac{\bar{Y} - E(\bar{Y})}{\sqrt{\mathrm{var}(\bar{Y})}}$ is approximately $N(0,1)$ (CLT)

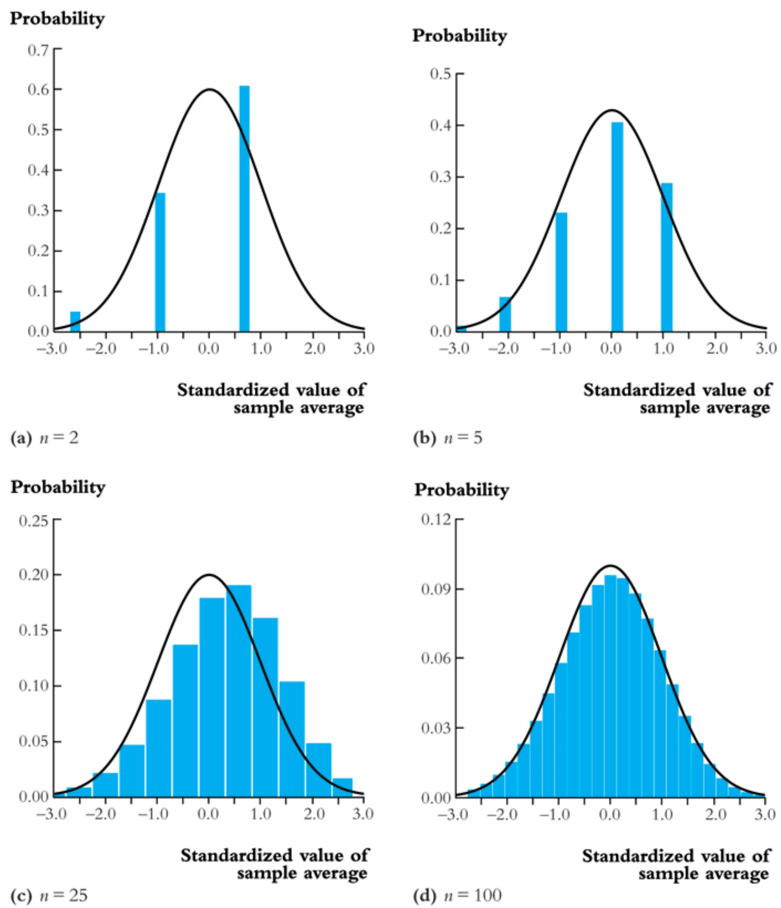
The *Central Limit Theorem* (CLT) (2 of 3)

Sampling distribution of $\bar{Y}$ when Y is Bernoulli, p = 0.78:



The *Central Limit Theorem* (CLT) (3 of 3)

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



(b) Why Use $\bar{Y}$ To Estimate μ_Y?

- $\bar{Y}$ is unbiased: $E(\bar{Y}) = \mu_Y$

- $\bar{Y}$ is consistent: $\bar{Y} \xrightarrow{p} \mu_Y$

- $\bar{Y}$ has a smaller variance than all other *linear unbiased* estimators: consider the estimator, $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, where $\{a_i\}$ are such that $\hat{\mu}_Y$ is unbiased; then $\operatorname{var}(\bar{Y}) \leq \operatorname{var}(\hat{\mu}_Y)$ (proof: SW, Ch. 17)

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