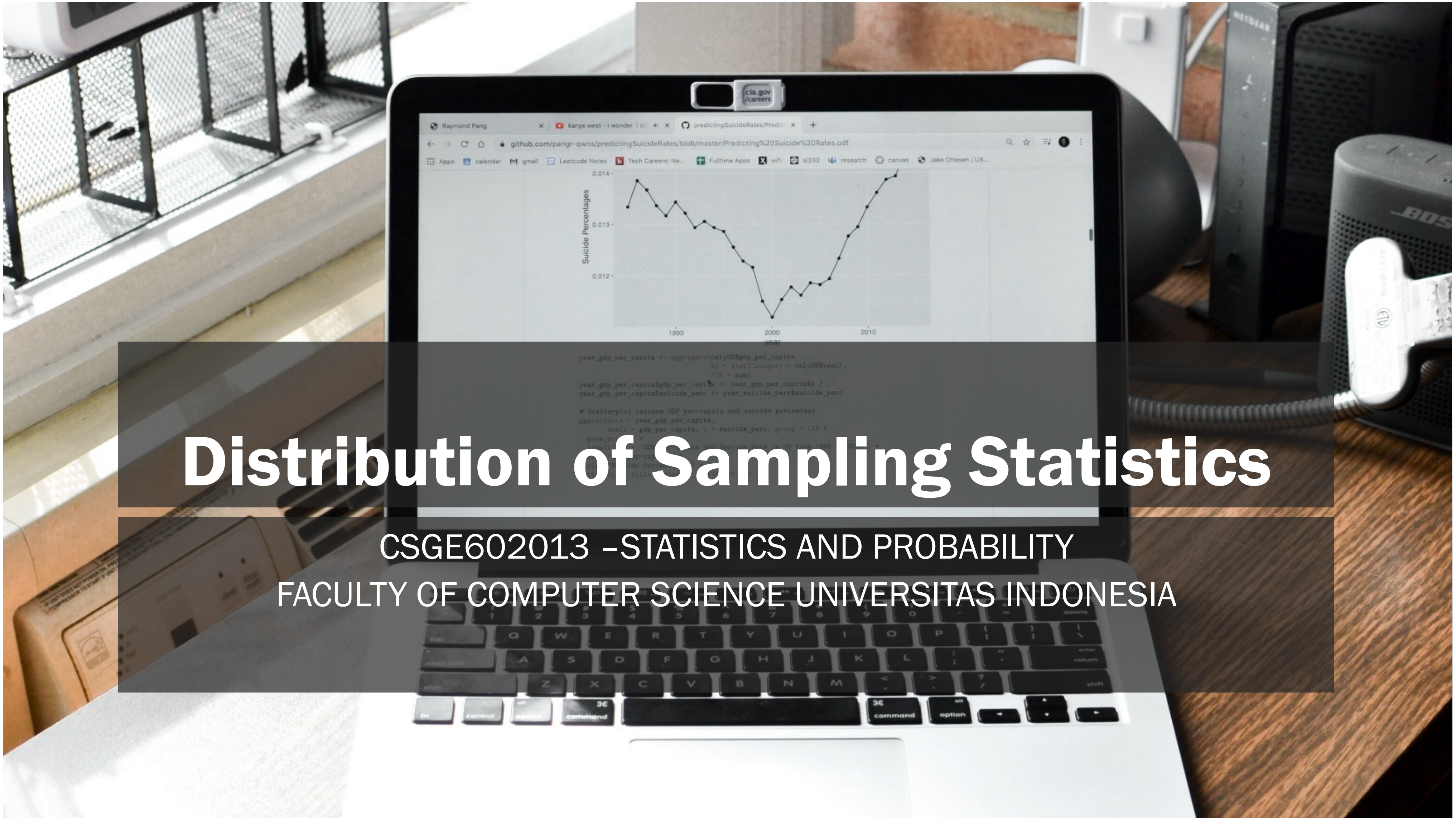


Distribution of Sampling Statistics

CSGE602013 –STATISTICS AND PROBABILITY

FACULTY OF COMPUTER SCIENCE UNIVERSITAS INDONESIA

References

- Introduction to Probability and Statistics for Engineers & Scientists, 4th ed., Sheldon M. Ross, Elsevier, 2009.
- A First Course in Probability, 8th Edition. Sheldon M. Ross
- Applied Statistics for the Behavioral Sciences, 5th Edition, Hinkle., Wiersma., Jurs., Houghton Mifflin Company, New York, 2003.
- Probability and Statistics for Engineers & Scientists, 4th Edition. Anthony J. Hayter, Thomson Higher Education

Outline

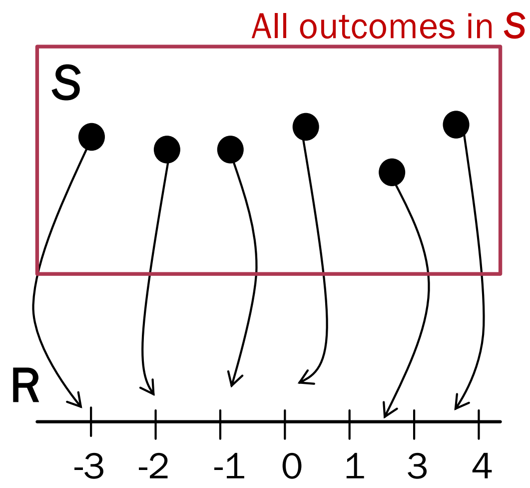
- Preliminaries
- Central Limit Theorem
- Distribution of Sample Mean
- Distribution of Proportion

PRELIMINARIES

Recall: Random Variable

- Random Variable X, is a function that assigns a numerical value $X(s)$ to each possible outcome in an experiment.

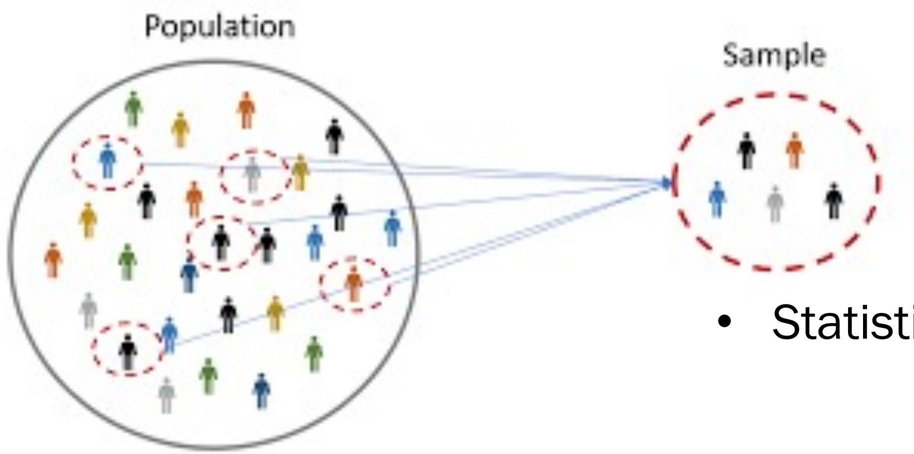
$$X : S \rightarrow R \qquad (or\ X(s) \in R, \forall s \in S)$$



- Each occurrence of the RV follows a certain **probability distribution**

Recall: Parameters & Statistics

- A parameter is to a population as a statistic is to a sample

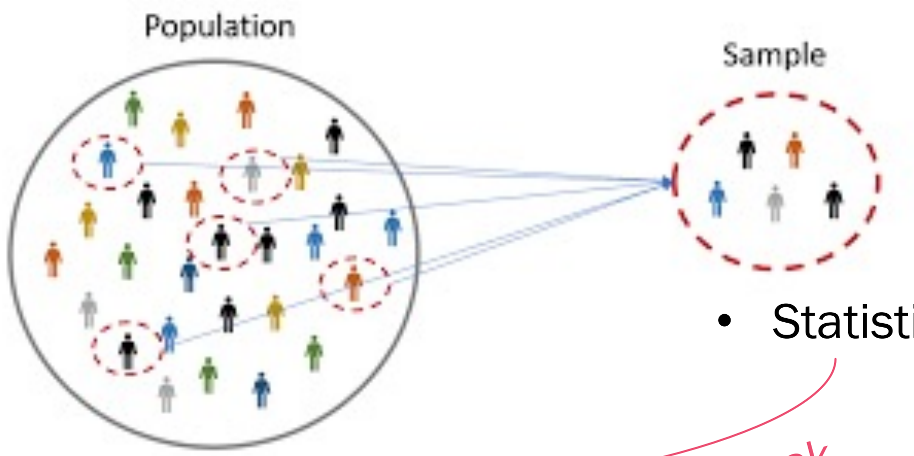


- Statistics

- Parameter

Inference from Samples

- Our goal is to make inferences about a (population) distribution F using the samples taken from F.



- Statistics

- Parameter

This is our task

- We need some assumptions

Assumptions

There is an underlying **probability distribution** of the population's parameters.

↓

Each measurable value of every member of the population can be viewed as **independent RVs with that distribution**

↓

Thus the randomly chosen **sample data are also independent RVs** following that distribution

Samples as Independent RVs

- If $\boldsymbol{X_1, X_2, \ldots, Xn}$ are **independent** random variables having a common distribution **F**, then we say that they constitute a ***sample*** (or ***random sample***) of size **n** from the distribution **F.**

n data samples

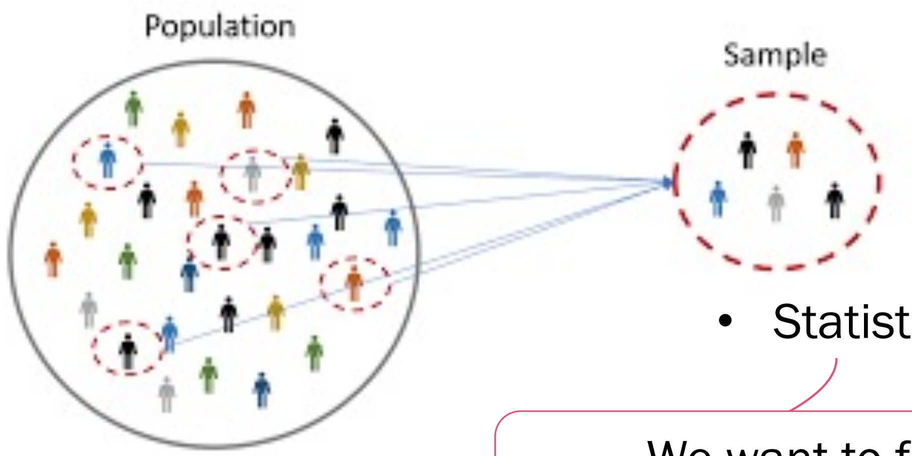
- X_1
- X_2
- ...
- X_n

Each are an independent RV with distribution F

- Inference Problems
  - **Parametric inference problem:** *F* is specified up to a set of unknown parameters
  - **Nonparametric inference problem:** nothing is assumed about *F*

Parametric Inference

- Our goal is to make inferences about a (population) distribution F using the samples taken from F.



- Statistics
- Parameter

We want to find the **probability distributions** of sample statistics

Inference from Statistics

- With the following data samples:

n data samples
- X_1
- X_2
- ...
- X_n

Each are an independent RV with dist F

- What can I measure?
  - Sample Mean
  - Sample Variance

Central Limit Theorem

CENTRAL LIMIT THEOREM

Central Limit Theorem

n data samples $\left.\begin{array}{l} X_1 \\ X_2 \\ \ldots \\ X_n \end{array}\right\}$ Each are an independent RV with dist F

Expectation μ, Variance σ^2

$$X_i \sim F(\mu, \sigma^2)$$

- Then, for **a large n**, the distribution of $X_1 + X_2 + \cdots + X_n$ is **approximately normal** with

$$X_1 + X_2 + \ldots + X_n \sim N(\underbrace{n\mu}_{\text{mean}}, \underbrace{n\sigma^2}_{\text{variance}})$$

Central Limit Theorem (2)

- For this **large n,** we attempt to convert it into a standard normal random variable

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \sim N(0,1)$$

- The cumulative distribution function(CDF) is as follows:

$$P\left(\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} < x\right) = P(Z < x)$$
$$= \Phi(x)$$

Contoh

- Sebuah perusahaan asuransi mempunyai 25,000 pemegang polis asuransi kendaraan. Bila klaim tahunan seorang pemegang polis adalah sebuah variabel acak dengan *mean* 320 dan standar deviasi 540, aproksimasikan probabilitas bahwa total klaim tahunan melebihi 8.3 juta!

Contoh

- Sebuah perusahaan asuransi mempunyai 25,000 pemegang polis asuransi kendaraan. Bila klaim tahunan seorang pemegang polis adalah sebuah variabel acak dengan *mean* 320 dan standar deviasi 540, aproksimasikan probabilitas bahwa total klaim tahunan melebihi 8.3 juta!
- Asumsikan X adalah VA yang merupakan total yearly claim. Number the policy holders, and let X_i denote the yearly claim of policy holder i.
- With n = 25000, we have from the central limit theorem that $X = \sum_{i=1}^{n} X_i$ will have approximately a normal distribution with $X_1 + X_2 + \ldots + X_n \sim N(n\mu, n\sigma^2)$

$$\mu = 320 \times 25000 = 8 \times 10^6$$
$$\sigma = 540\sqrt{25000} = 8.5381 \times 10^4$$

$$X = \sum_{i=1}^{n} X_i \sim N(8 \times 10^6, (8.5381 \times 10^4)^2)$$

Contoh

- Sebuah perusahaan asuransi mempunyai 25,000 pemegang polis asuransi kendaraan. Bila klaim tahunan seorang pemegang polis adalah sebuah variabel acak dengan *mean* 320 dan standar deviasi 540, aproksimasikan probabilitas bahwa total klaim tahunan melebihi 8.3 juta!

$$X = \sum_{i=1}^{n} X_i \sim N(8\times10^6, (8.5381\times10^4)^2)$$

$$P\left(X > 8.3\times10^6\right) = P\left(\frac{X - 8\times10^6}{8.5381\times10^4} > \frac{8.3\times10^6 - 8 \quad \times10^6}{8.5381\times10^4}\right)$$

$$\approx P\left(Z > 3.51\right)$$

$$\approx 1 - P(Z \le 3.51)$$

$$\approx 1 - \Phi(3.51)$$

$$\approx 0.00023$$

DISTRIBUTION OF SAMPLE MEAN

Recall: Sample Mean

- Let X_1, X_2, ..., X_n be a sample of values from a population having *expectation* **μ** and *variance* **σ**^{2}.
- The sample mean is:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$\bar{X}$ is also a **random variable**

Sample Mean (2)

$$E[\overline{X}] = E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right]$$
$$= \frac{1}{n}\left(E[X_1] + \ldots + E[X_n]\right)$$
$$= \mu$$

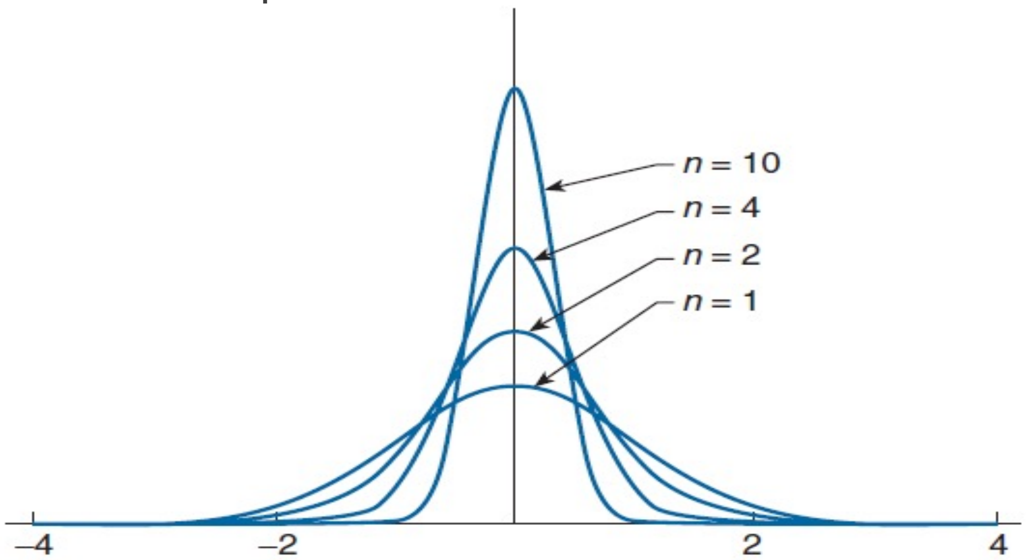
$$Var(\overline{X}) = Var\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)$$
$$= \frac{1}{n^2}\left(Var(X_1) + \ldots + Var(X_n)\right)$$
$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

By independence

- Where μ and σ^2 are the population mean and variance, respectively

The Sample Mean Approximating the Normal Distribution

- $\bar{X}$ is also centered about the population mean μ, but its spread becomes more and more reduced as the sample size increases.



Densities of sample means from a standard normal population.

Approximate Distribution of The Sample Mean

- Let $\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{Xn}$ be a sample of values from a population having expectation $\boldsymbol{\mu}$ and *variance* $\boldsymbol{\sigma^2}$.
- We know from the central limit theorem that $\bar{X}$ is approximately normal when sample size n is large, which is:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Approximate Distribution of The Sample Mean (2)

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Where $E[\overline{X}] = \mu$

$$Var(\overline{X}) = \frac{\sigma^2}{n} \quad SD(\overline{X}) = \sqrt{Var(\overline{X})} = \sigma / \sqrt{n}$$

- The standard normal distribution

$$\frac{\overline{X} - E[\overline{X}]}{SD(\overline{X})} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Contoh

- Berat badan dari populasi pekerja mempunyai mean 167 (*pounds*) dan standar deviasi 27.
  a. Bila sampel sebanyak 36 pekerja dipilih, aproksimasikan probabilitas bahwa sample mean dari berat badan mereka di antara 163 dan 171.
  b. Aproksimasikan lagi seperti di (a) jika ukuran sampel 144 pekerja.

Contoh

- Berat badan dari populasi pekerja mempunyai mean 167 (*pounds*) dan standar deviasi 27.
  a. Bila sampel sebanyak 36 pekerja dipilih, aproksimasikan probabilitas bahwa sample mean dari berat badan mereka di antara 163 dan 171.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \begin{aligned} &\mu = 167 \\ &\sigma / \sqrt{n} = 27 / \sqrt{36} = 4.5 \end{aligned}$$

$$\begin{aligned} P\left(163 < \bar{X} < 171\right) &= P\left(\frac{163-167}{4.5} < \frac{\bar{X}-167}{4.5} < \frac{171-167}{4.5}\right) \\ &\approx P\left(-0.8889 < Z < 0.8889\right) \\ &\approx 2P(Z < 0.8889) - 1 \\ &\approx 0.6259 \end{aligned}$$

Contoh

- Berat badan dari populasi pekerja mempunyai mean 167 (*pounds*) dan standar deviasi 27.
  b. Aproksimasikan lagi seperti di (a) jika ukuran sampel 144 pekerja.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \begin{aligned} \mu &= 167 \\ \sigma / \sqrt{n} &= 27 / \sqrt{144} = 2.25 \end{aligned}$$

$$\begin{aligned} P\left(163 < \bar{X} < 171\right) &= P\left(\frac{163-167}{2.25} < \frac{\bar{X}-167}{2.25} < \frac{171-167}{2.25}\right) \\ &\approx 2P\left(Z < 1.7778\right) - 1 \\ &\approx 0.9246 \end{aligned}$$

How Large a Sample is Needed ?

- **A general rule of thumb** is that one can be confident of the normal approximation whenever the sample size **n** is at least 30.

∴ no matter how **non normal** the underlying population distribution is, the sample mean of a sample of size at least 30 will be **approximately normal.**

- In most cases, the normal approximation is valid for much smaller sample sizes.

How Large a Sample is Needed ?

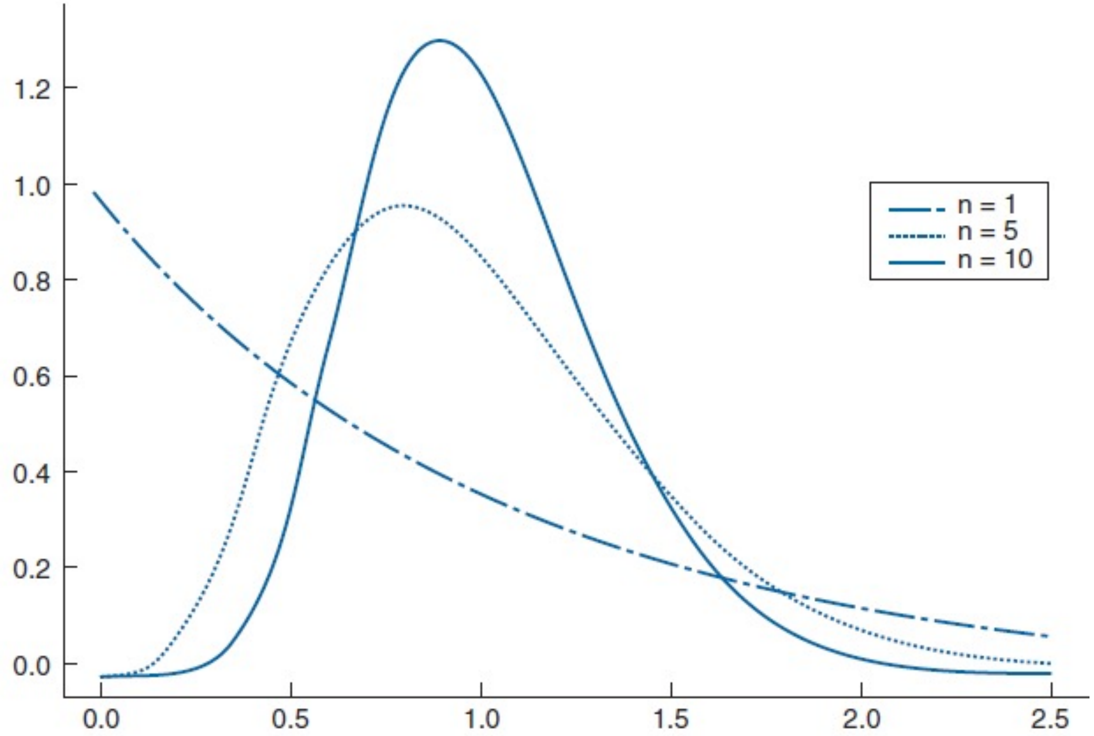


FIGURE 6.4 *Densities of the average of n exponential random variables having mean 1.*

DISTRIBUTION OF PROPORTION

Sampling from A Finite Population

- Consider a population of N elements, and suppose that p is the proportion of the population that has a certain characteristic of interest; that is
  - $\boldsymbol{Np}$ elements have this characteristic
  - $\boldsymbol{N(1-p)}$ do not
- A sample of size n from this population is said to be a random sample if each of the $\binom{N}{n}$ possibly selected population subsets of size n is equally likely to be selected as the sample

Sampling from A Finite Population (2)

- Suppose a random sample of size n has been chosen from population of size N.
- For $i = 1, 2, \dots, n$, let

$$X_i = \begin{cases} 1 & \text{If the } \mathbf{i^{th}} \text{ member of the sample has the characteristic} \\ 0 & \text{Otherwise} \end{cases}$$

*) Each X_i is thus a Bernoulli RV.

- When the population size N is large with respect to the sample size n, then $\boldsymbol{X_1, X_2, \dots, Xn}$ are approximately **independent.**

Sampling from A Finite Population (3)

- If we sum up $n\ \boldsymbol{X_i}$

$$X = \sum_{i=1}^{n} X_i$$

*) Recall each X_i is a Bernoulli RV.

- Each X_i is either 1 (success) or 0. Thus X is the total number of success in n trials.
- Since X_i is independent, then X would be a **Binomial RV **)** with parameters n and p.

$$X \sim B(n, p)$$

$$E[X] = np$$

$$Var(X) = np(1-p)$$

Sampling from A Finite Population (4)

- The sample mean also shows the proportion **p** of the members that posses the characteristics

$$\bar{X} = \frac{X}{n} = \sum_{i=1}^{n} X_i / n$$

- If the underlying population is large in relation to the sample size, we can then infer

$$E[\overline{X}] = E[X / n] = p$$

$$Var(\overline{X}) = \frac{1}{n^2} Var(X) = \frac{p(1-p)}{n}$$

$$SD(\overline{X}) = \sqrt{\frac{p(1-p)}{n}}$$

Exercise

- Suppose that 45 percent of the population favors a certain candidate in an upcoming election. If a random sample of size 200 is chosen, find
  - (a) the expected value and standard deviation of the number of members of the sample that favor the candidate;
  - (b) the probability that more than half the members of the sample favor the candidate.

Exercise (cont.)

- Suppose that 45 percent of the population favor a certain candidate in an upcoming election. If a random sample of size 200 is chosen, find
  - (a) the expected value and standard deviation of the number of members of the sample that favor the candidate;
  - (b) the probability that more than half the members of the sample favor the candidate.

Q & A