CHAPTER 7
Sampling Distributions

Objectives

- Parameters and statistics
- Statistical estimation and the law of large numbers
- Sampling distributions
- The sampling distribution of $\bar{x}$
- The central limit theorem
- Sampling distributions and statistical significance

Parameters and statistics

- As we begin to use sample data to draw conclusions about a wider population, we must be clear about whether a number describes a sample or a population.

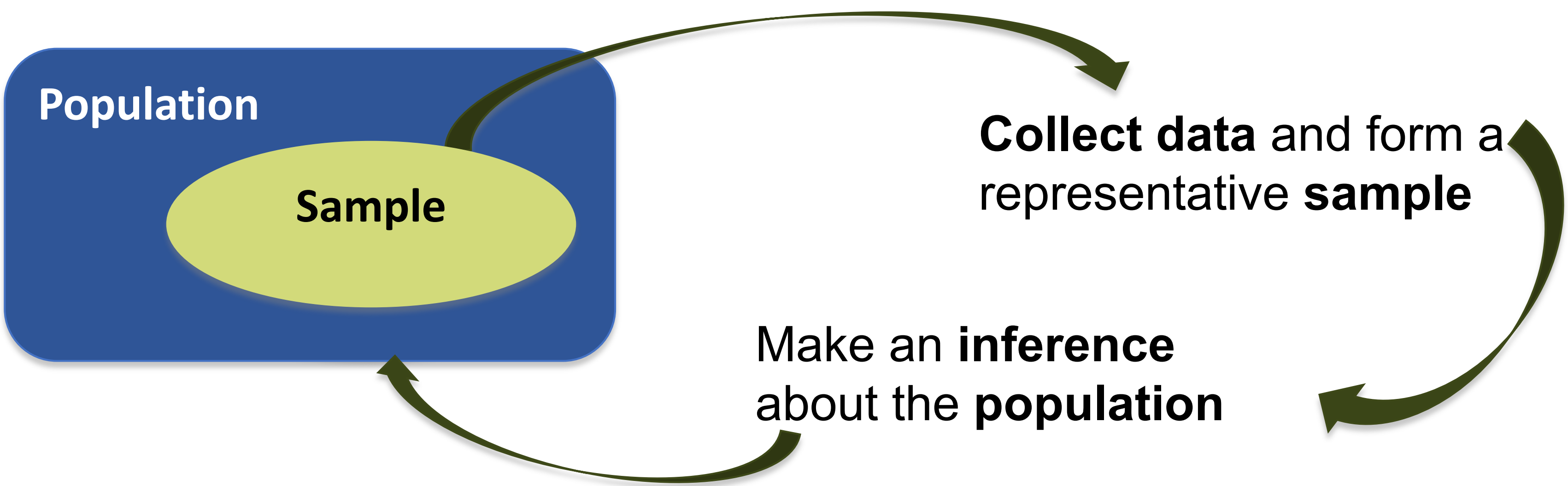
- A **parameter** is a number that describes the population. In practice, the value of a parameter is not known because we can rarely examine the entire population.

- A **statistic** is a number that can be computed from the sample data without making use of any unknown parameters. In practice, we often use a statistic to estimate an unknown parameter.

- Remember ***p*** and ***s***: ***p***arameters come from ***p***opulations and ***s***tatistics come from ***s***amples.
- μ (mu) for the mean of the population
- σ (sigma) for the standard deviation of the population.
- $\bar{x}$ ("x-bar") for the mean of the sample and s for the standard deviation of the sample.

Statistical estimation

- The process of **statistical inference** involves using information from a sample to draw conclusions about a wider population.
- Different random samples yield different statistics. We need to be able to describe the **sampling distribution** of possible statistic values in order to perform statistical inference.
- We can think of a statistic as a **random variable** because it takes numerical values that describe the outcomes of the random sampling process. Therefore, we can examine its probability distribution using concepts we learned in earlier chapters.



The law of large numbers

- If $\bar{x}$ is rarely exactly right and varies from sample to sample, why is it nonetheless a reasonable estimate of the population mean μ?
- Here is one answer: If we keep taking larger and larger samples, the statistic $\bar{x}$ is guaranteed to get closer and closer to the parameter μ.

LAW OF LARGE NUMBERS

- Draw observations at random from any population with finite mean μ. As the number of observations drawn increases, the mean $\bar{x}$ of the observed values tends to get closer and closer to the mean μ of the population.

Sampling distributions

- The law of large numbers assures us that if we measure enough subjects, the statistic $\bar{x}$ will eventually get very close to the unknown parameter μ.
- If we took every one of the possible samples of a certain size, calculated the sample mean for each, and graphed all of those values, we'd have a sampling distribution.
- Using software to imitate chance behavior to carry out tasks such as exploring sampling distributions is called simulation.
- The population distribution of a variable is the distribution of values of the variable among all individuals in the population.

- The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.
- Be careful: The population distribution describes the individuals that make up the population. A sampling distribution describes how a statistic varies in many samples from the population.

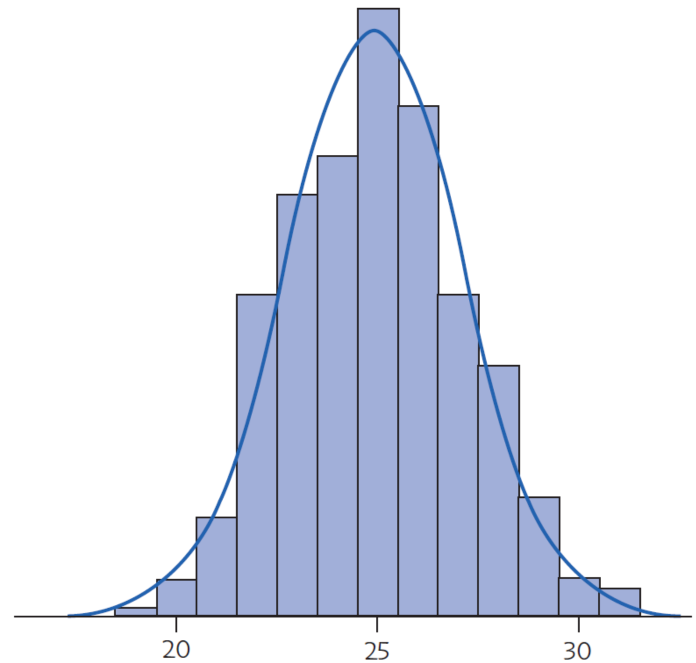
The sampling distribution of $\bar{x}$ (illustrated)

Mean threshold of all adults to smell sulfate in wine is μ=25 with a standard deviation of σ=7, and the threshold values follow a bell-shaped (normal) curve.

Randomly select 10 adults, the mean $\bar{x}$ is 26.42. What's the distribution of the sample mean $\bar{X}$?

Population,
mean $\mu = 25$

$$\left.\begin{array}{l} \xrightarrow{\text{SRS size 10}} \bar{x} = 26.42 \\ \xrightarrow{\text{SRS size 10}} \bar{x} = 24.28 \\ \xrightarrow{\text{SRS size 10}} \bar{x} = 25.22 \\ \qquad\qquad\vdots \end{array}\right\}$$



Sampling distribution of $\bar{x}$ of 10 adults,
mean $\mu_{\bar{x}} = 25$

Recall:

Random variable: Variable that has a single numerical value determined by chance for each outcome of an experiment.

Probability distribution: A graph, table, or formula that gives the probability for each value of the random variable.

New:

Sampling distribution of the sample mean:

Probability distribution of the sample mean is obtained when we repeatedly draw samples of the same size n from the same population

The Mean and Standard Deviation of a Sample mean $\overline{x}$

- Suppose that $\bar{x}$ is the mean of an SRS of size n drawn from a large population with mean μ and standard deviation σ. Then the sampling distribution of $\bar{x}$ has **mean μ** and **standard deviation $\sigma/\sqrt{n}$**

- Because the mean of the statistic $\bar{x}$ is always equal to the mean μ of the population (that is, the sampling distribution of $\bar{x}$ is centered at μ), we say the statistic $\bar{x}$ is an **unbiased estimator** of the parameter μ.

  Note: on any particular sample, $\bar{x}$ may fall above or below μ.

The sampling distribution of $\bar{x}$

- Because the standard deviation of the sampling distribution of $\bar{x}$ is $\sigma/\sqrt{n}$, the averages are less variable than individual observations, and averages of large sample are less variable than the averages of small samples.
- Not only is the standard deviation of the distribution of $\bar{x}$ smaller than the standard deviation of individual observations, it gets smaller as we take larger samples. The results of large samples are less variable than the results of small samples.

Note: While the standard deviation of the distribution of $\bar{x}$ gets smaller, it does so at the rate of $\sqrt{n}$, not n. To cut the sampling distribution's standard deviation in half, for instance, you must take a sample four times as large, not just twice as large.

The shape of the sampling distribution of $\bar{x}$

- We have described the center and variability of the sampling distribution of a sample mean $\bar{x}$, but not its shape. The shape of the sampling distribution depends on the shape of the population distribution.

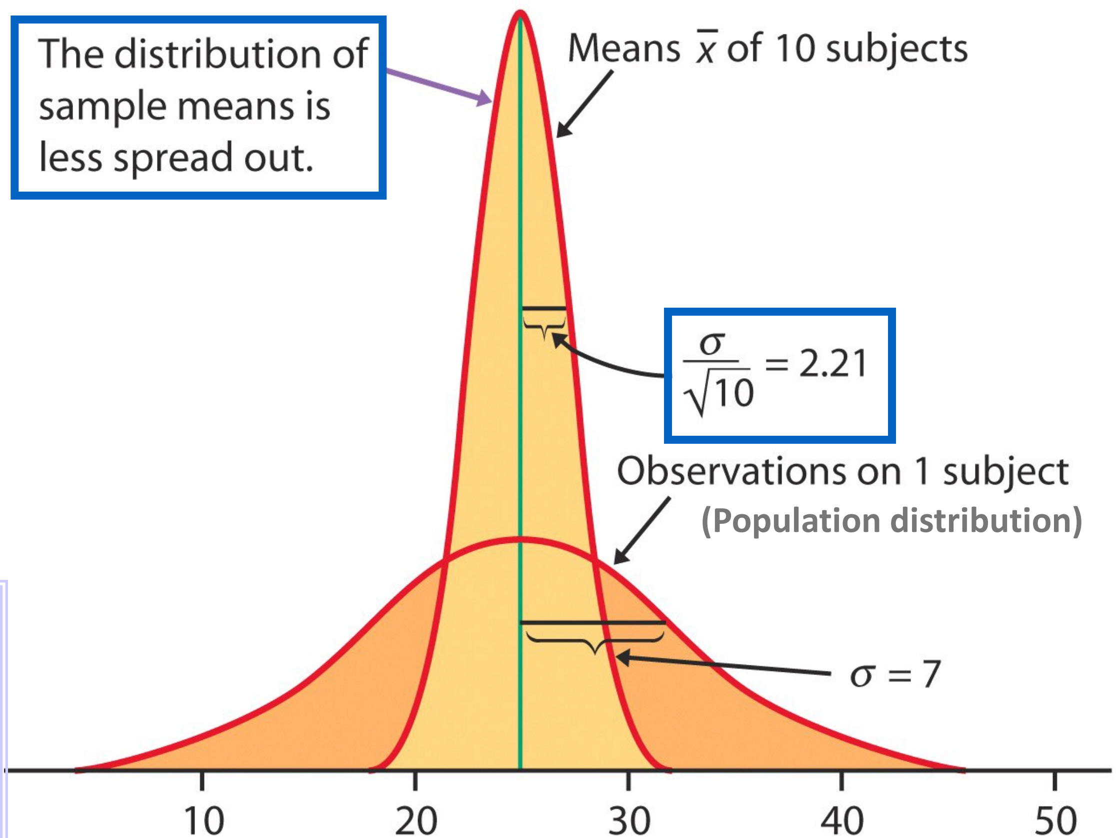
- In one important case there is a simple relationship between the two distributions: if the population distribution is Normal, then so is the sampling distribution of the sample mean.

SAMPLING DISTRIBUTION OF A SAMPLE MEAN

- If individual observations have the $N(\mu,\ \sigma)$ distribution, then the sample mean $\bar{x}$ of an SRS of size n has the $N(\mu,\ {}^{\sigma}/_{\sqrt{n}})$ distribution.

Example: When population distribution is Normal: Does This Wine Smell Bad?

Mean threshold of all adults is μ=25 with a standard deviation of σ=7, and the threshold values follow a bell-shaped (normal) curve.



If the population is $N(\mu, \sigma)$ then the sample means distribution is $N(\mu, \sigma/\sqrt{n})$.

When population distribution are not normal: The central limit theorem

- Most population distributions are not Normal. What is the shape of the sampling distribution of sample means when the population distribution isn't Normal?
- A remarkable fact is that as the sample size increases, the distribution of sample means changes its shape: it looks less like that of the population distribution and more like a Normal distribution!

- Draw an SRS of size n from any population with mean μ and finite standard deviation σ. The **central limit theorem** says that when *n* is large ($n \geq 30$), the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

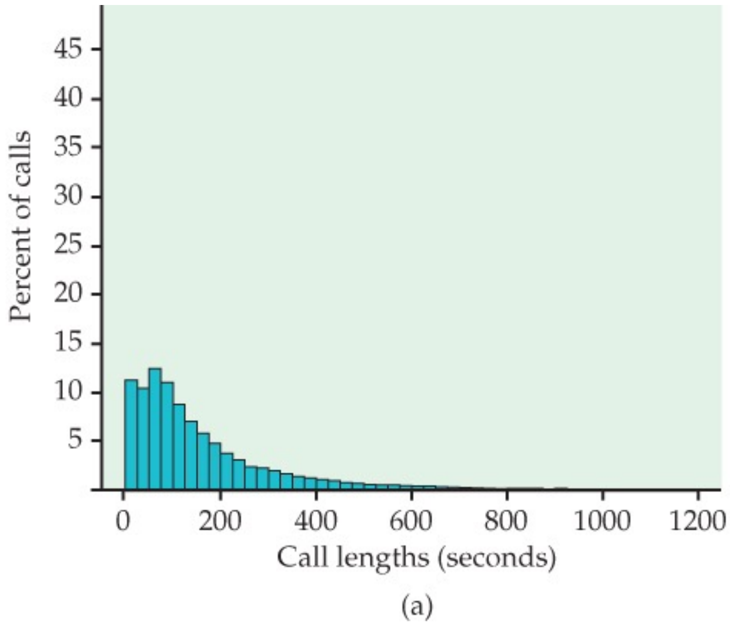
$$\bar{x} \text{ is approximately } N\left(\mu,\ {}^{\sigma}/_{\sqrt{n}}\right) \text{or } z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,\ 1)$$

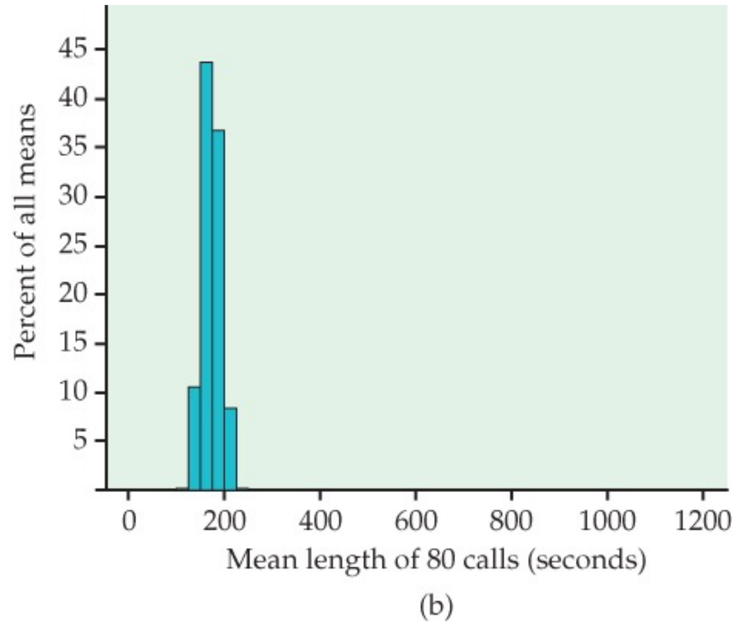
- The central limit theorem allows us to use Normal probability calculations to answer questions about sample means from many observations, even when the population distribution is not Normal.

When population distribution is not Normal Distribution

Population distribution

Sampling distribution

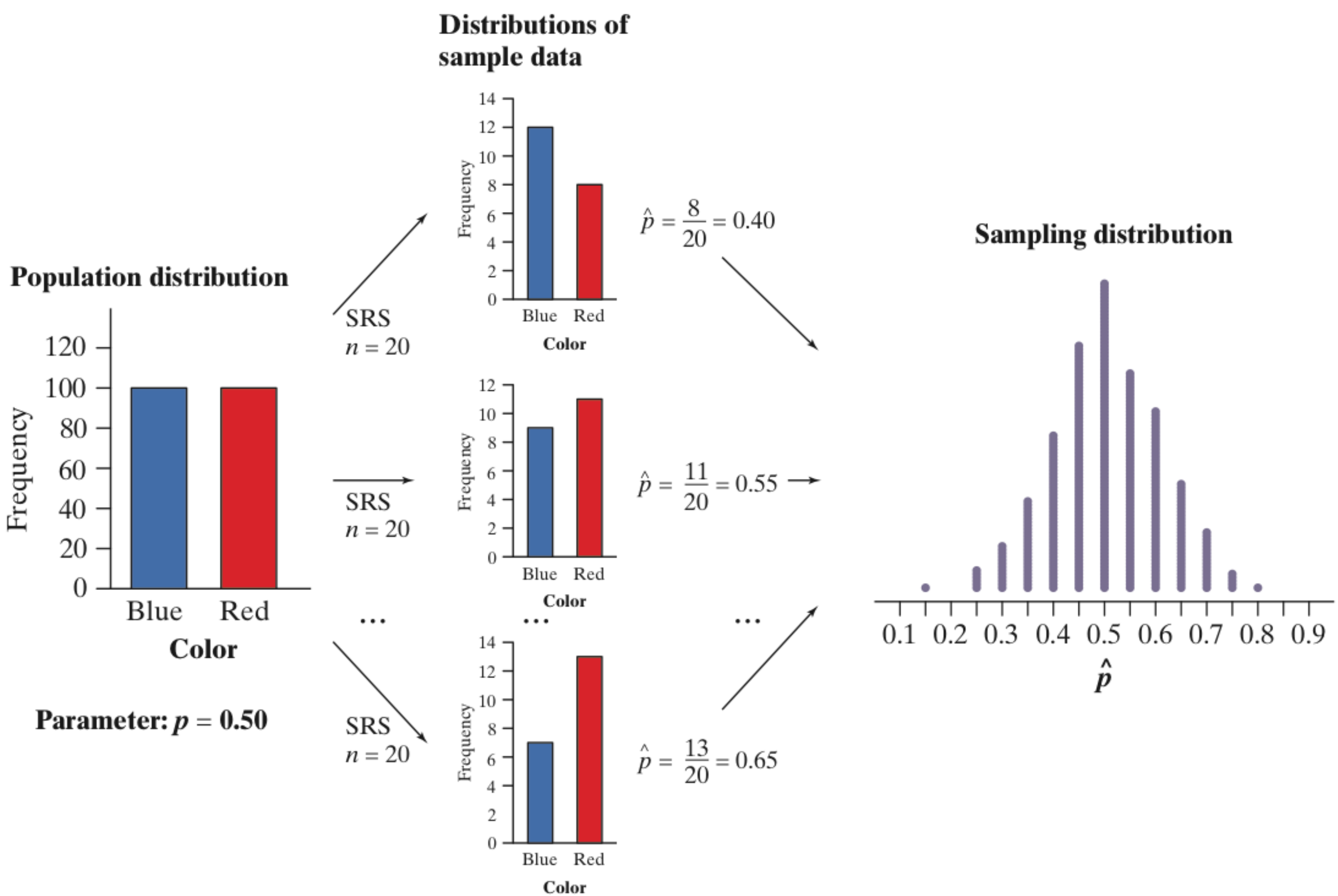




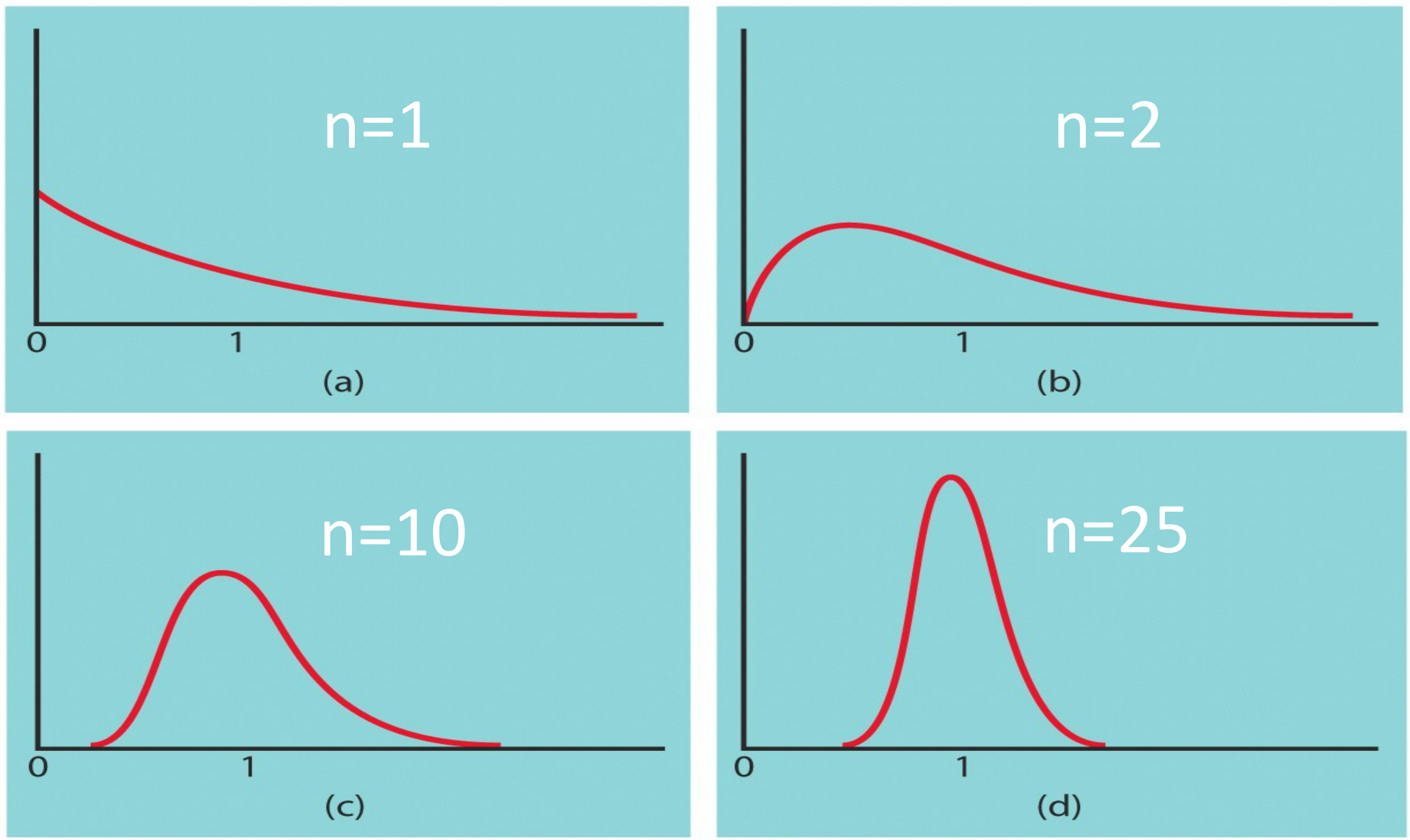
(a). The distribution of lengths of all customer service calls received by a bank in a month.

(a). The distribution of the sample mean $\bar{x}$ of size 80 with 500 random samples from this population.

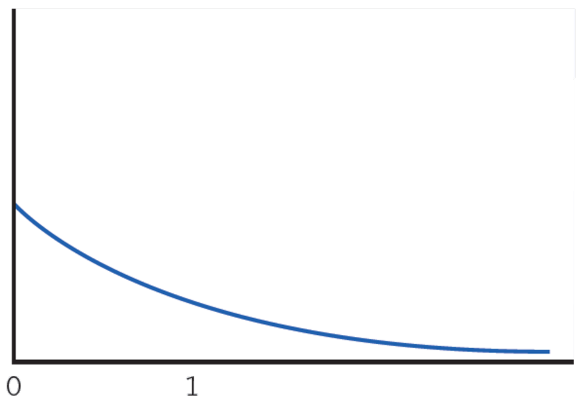
Central limit theorem: When population distribution is not normal



Central Limit Theorem:
Sample Size and Distribution of $\bar{x}$



Central limit theorem: example



Based on service records from the past year, the time (in hours) that a technician requires to complete preventative maintenance on an air conditioner follows the distribution that is strongly right-skewed and whose most likely outcomes are close to 0. The mean time is $\mu = 1$ hour and the standard deviation is $\sigma = 1$.

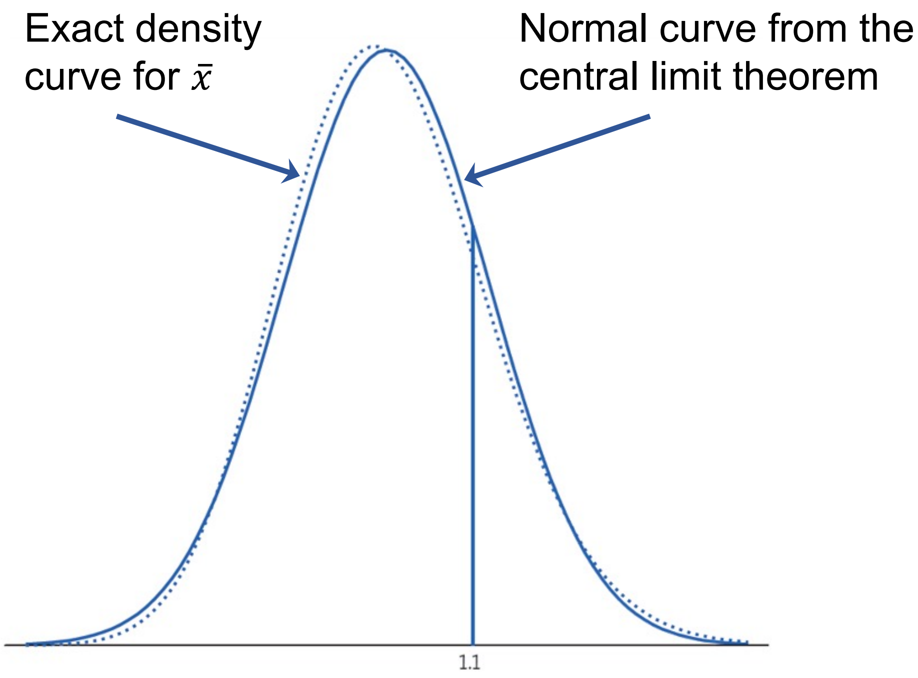
Your company will service an SRS of 70 air conditioners. You have budgeted 1.1 hours per unit. Will this be enough?

Based on the central limit theorem, what's sampling distribution of the mean time spent working on the 70 units?

Central limit theorem: example

Your company will service an SRS of 70 air conditioners. You have budgeted 1.1 hours per unit. Will this be enough? Calculate P($\bar{x}$>1.1).

since $n = 70 \geq 30$,the sampling distribution of the mean time spent working is approximately $N(\ ?,\ \ ?\ )$



If you budget 1.1 hours per unit, there is a ? % chance the technicians will not complete the work within the budgeted time.

Sampling distributions and statistical significance

- The sampling distribution of a sample statistic is determined by the particular sample statistic we are interested in, the distribution of the population of individual values from which the sample statistic is computed, and the method by which samples are selected from the population.

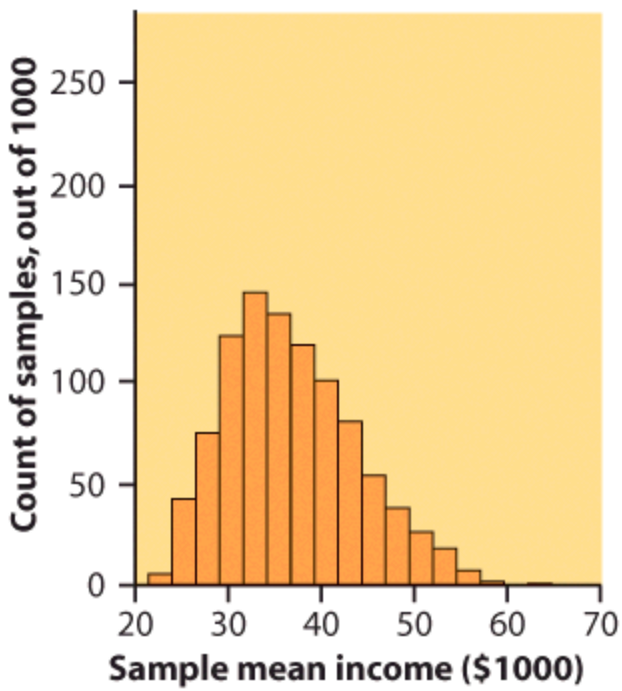
- The sampling distribution allows us to determine the probability of observing any particular value of the sample statistic in another such sample from the population.

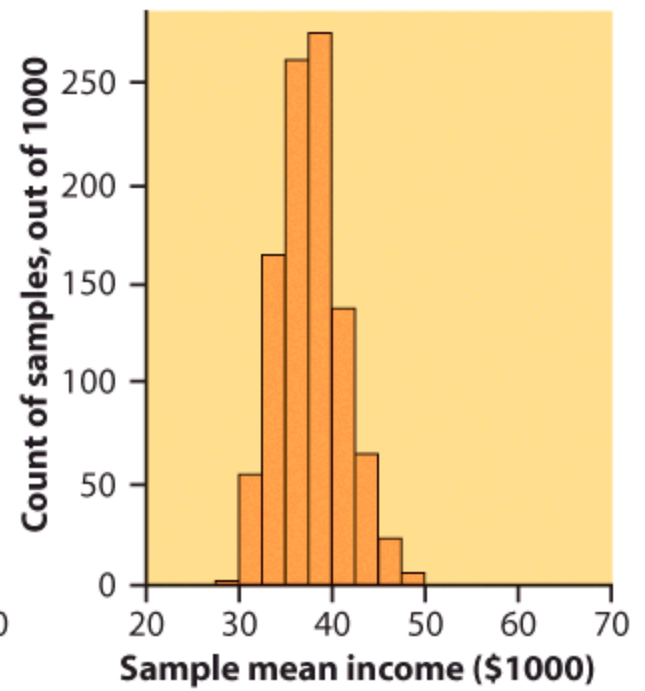
Income distribution

Let’s consider the very large database of individual incomes from the Bureau of Labor Statistics as our population. It is strongly right skewed.

- We take 1000 SRSs of 100 incomes, calculate the sample mean for each, and make a histogram of these 1000 means.
- We also take 1000 SRSs of 25 incomes, calculate the sample mean for each, and make a histogram of these 1000 means.

Which histogram corresponds to samples of size 100? 25?





IQ scores: population vs. sample

In a large population of adults, the mean IQ is 112 with standard deviation 20.
Suppose 200 adults are randomly selected for a market research campaign.

- The distribution of the sample mean IQ is:

  A) Exactly normal, mean 112, standard deviation 20

  B) Approximately normal, mean 112, standard deviation 20

  C) Approximately normal, mean 112 , standard deviation 1.414

  D) Approximately normal, mean 112, standard deviation 0.1

Application of the central limit theorem

Hypokalemia is diagnosed when blood potassium levels are below 3.5mEq/dl. Let's assume that we know a patient whose measured potassium levels vary daily according to a normal distribution $N(\mu = 3.8, \sigma = 0.2)$.

If only one measurement is made, what is the probability that this patient will be diagnosed with Hypokalemia?

Instead, if measurements are taken on 4 separate days, what is the probability of a diagnosis with Hypokalemia?