CHAPTER 15: Sampling Distributions

Basic Practice of Statistics

7th Edition

Lecture PowerPoint Slides

In Chapter 15, we cover …

- 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

Sampling Terminology (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 fixed unknown number that describes the population. Example: population mean

- 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. Example: Sample mean

- Remember ***p*** and ***s***: ***p***arameters come from ***p***opulations and ***s***tatistics come from ***s***amples.

- We write μ (the Greek letter mu) for the mean of the population and σ (the Greek letter sigma) for the standard deviation of the population. We write $\bar{x}$ ("x-bar") for the mean of the sample and s for the standard deviation of the sample.

Parameter vs Statistic

A properly chosen sample of 1600 people across the United States was asked if they regularly watch a certain television program, and 24% said *yes*. The *parameter* of interest here is the true proportion of all people in the U.S. who watch the program, while the *statistic* is the value 24% obtained from the sample of 1600 people.

Parameter vs. Statistic

- The mean of a population is denoted by μ – this is a parameter.
- The mean of a sample is denoted by $\bar{x}$ – this is a statistic. $\bar{x}$ is used to estimate μ.
- The true proportion of a population with a certain trait is denoted by p – this is a parameter.
- The proportion of a sample with a certain trait is denoted by $\hat{p}$ ("*p-hat*") – this is a statistic. $\hat{p}$ is used to estimate p.

Parameters and Statistics (1 of 2)

A **parameter** is a number that describes the __________.

a) population

b) sample

c) statistic

d) None of the answer options is correct.

Parameters and Statistics (2 of 2)

A ____________ is a number that can be computed from the sample data to estimate an unknown population parameter.

a) population

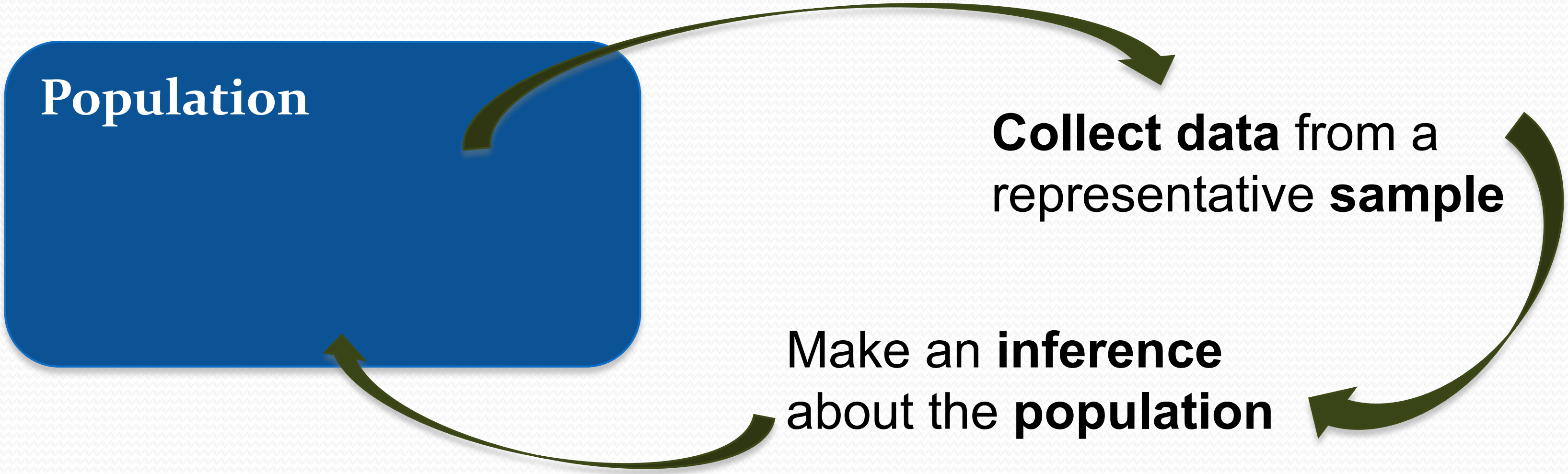
b) parameter

c) statistic

d) None of the answer options is correct.

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.



Collect data from a representative **sample**

Make an **inference** about the **population**

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. ($\bar{x}$ gets closer to 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.

The Law of Large Numbers (Gambling)

- The “house” in a gambling operation is not gambling at all
  - the games are defined so that the gambler has a negative expected gain per play (the true mean gain is negative)
  - each play is independent of previous plays, so the *law of large numbers* guarantees that the average winnings of a large number of customers will be close the the (negative) true average

Law of Large Numbers (1 of 2)

As the sample size gets larger, the sample mean gets closer to the _________.

a) population mean

b) population variance

c) population standard deviation

d) population median

Law of Large Numbers (2 of 2)

- _______________ assures us that, if we measure enough subjects, the statistic $\bar{x}$ will eventually get very close to the unknown parameter μ.

a) Benford's law

b) The law of large numbers

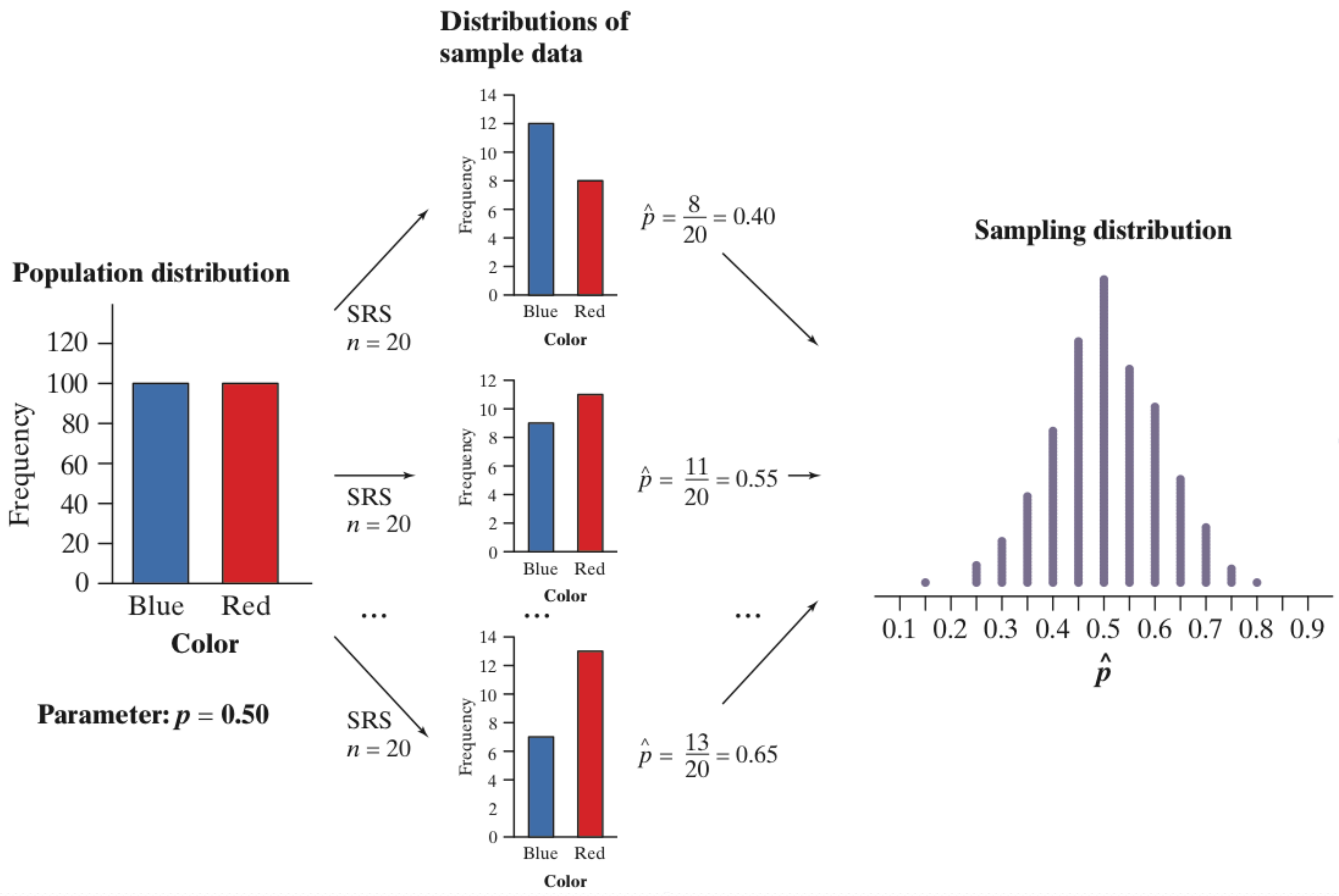
c) Conditional probability

d) The continuous probability model

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.
- If we use software to imitate chance behavior to carry out tasks such as exploring sampling distributions, this 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.

Population distributions vs. sampling distributions



Population distribution

Parameter: $p = 0.50$

SRS $n = 20$

SRS $n = 20$

SRS $n = 20$

Distributions of sample data

$\hat{p} = \frac{8}{20} = 0.40$

$\hat{p} = \frac{11}{20} = 0.55$

$\hat{p} = \frac{13}{20} = 0.65$

Sampling distribution

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

- When we choose many SRSs from a population, the sampling distribution of the sample mean is centered at the population mean μ and is less spread out than the population distribution.

MEAN AND STANDARD DEVIATION OF A SAMPLE MEAN

- 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 $\boldsymbol{\mu}$** and **standard deviation $\boldsymbol{\sigma/\sqrt{n}}$**

- 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.

Case Study

Does This Wine Smell Bad?

Dimethyl sulfide (DMS) is sometimes present in wine, causing “off-odors”. Winemakers want to know the odor threshold – the lowest concentration of DMS that the human nose can detect. Different people have different thresholds, and of interest is the mean threshold in the population of all adults.

Case Study

Does This Wine Smell Bad?

Suppose the mean threshold of all adults is $\mu=25$ micrograms of DMS per liter of wine, with a standard deviation of $\sigma=7$ micrograms per liter and the threshold values follow a bell-shaped (normal) curve.

Where should 95% of all individual threshold values fall?

- mean plus or minus two standard deviations

$$25 - 2(7) = 11$$

$$25 + 2(7) = 39$$

- 95% should fall between 11 & 39

- What about the ***mean*** (average) of a sample of *10* adults? What values would be expected?

Sampling Distribution

- What about the ***mean*** (average) of a sample of *10* adults? What values would be expected?
- Answer this by thinking: "What would happen if we took many samples of *10* subjects from this population?"
  - take a large number of samples of 10 subjects from the population
  - calculate the sample mean (x-bar) for each sample
  - make a histogram of the values of x-bar
  - examine the graphical display for shape, center, spread

Case Study

Does This Wine Smell Bad?

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

Many (1000) samples of *n*=10 adults from the population were taken and the resulting histogram of the 1000 x-bar values is on the next slide.

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

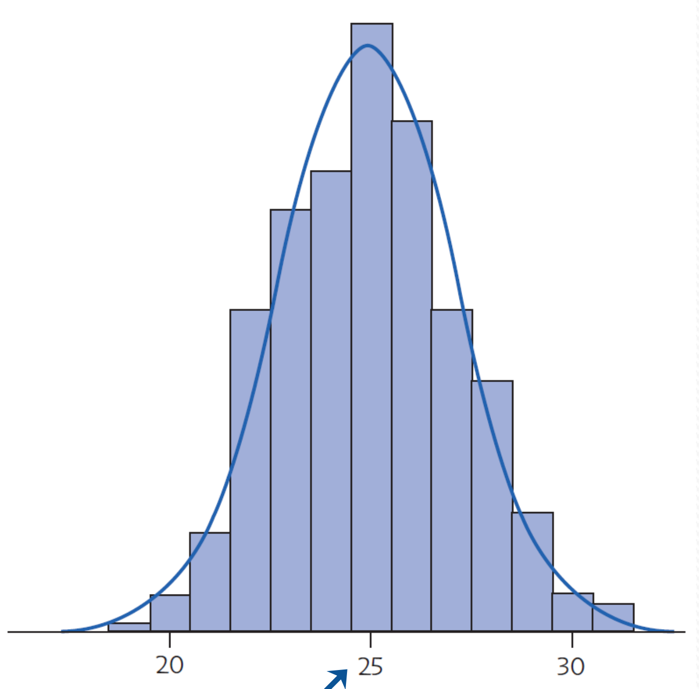
Does This Wine Smell Bad?

$$\xrightarrow{\text{SRS size } 10} \bar{x} = 26.42$$

$$\xrightarrow{\text{SRS size } 10} \bar{x} = 24.28$$

$$\xrightarrow{\text{SRS size } 10} \bar{x} = 25.22$$

$$\vdots$$



Population,
mean $\mu = 25$

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

What can we say about the shape, center, and spread of this distribution?

- Shape: It looks Normal !
- Center: The mean of the 1000 $\overline{x}$'s is 24.95. That is, the distribution is centered very close to the population mean $\mu = 25$.
- Spread: The standard deviation of the 1000 $\overline{x}$'s is 2.217, notably smaller than the standard deviation $\sigma = 7$ of the population of individual subjects.

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

- 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 are less variable than the results of small samples.

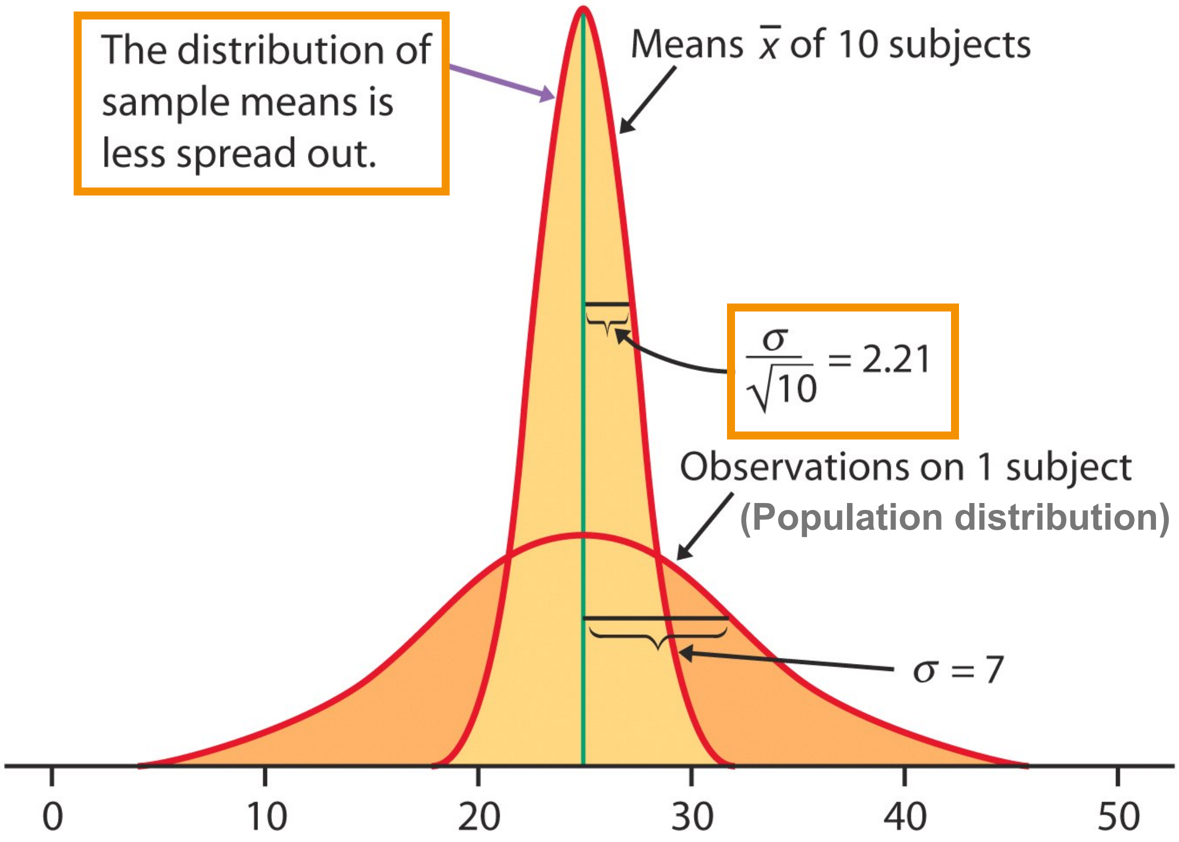
- Not only is the standard deviation of the distribution of $\bar{x}$ smaller than the standard deviation of individual observations, it gets smaller as we take larger samples. The results of large samples are less variable than the results of small samples.

  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.

Case Study

Does This Wine Smell Bad?

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


The distribution of sample means is less spread out.

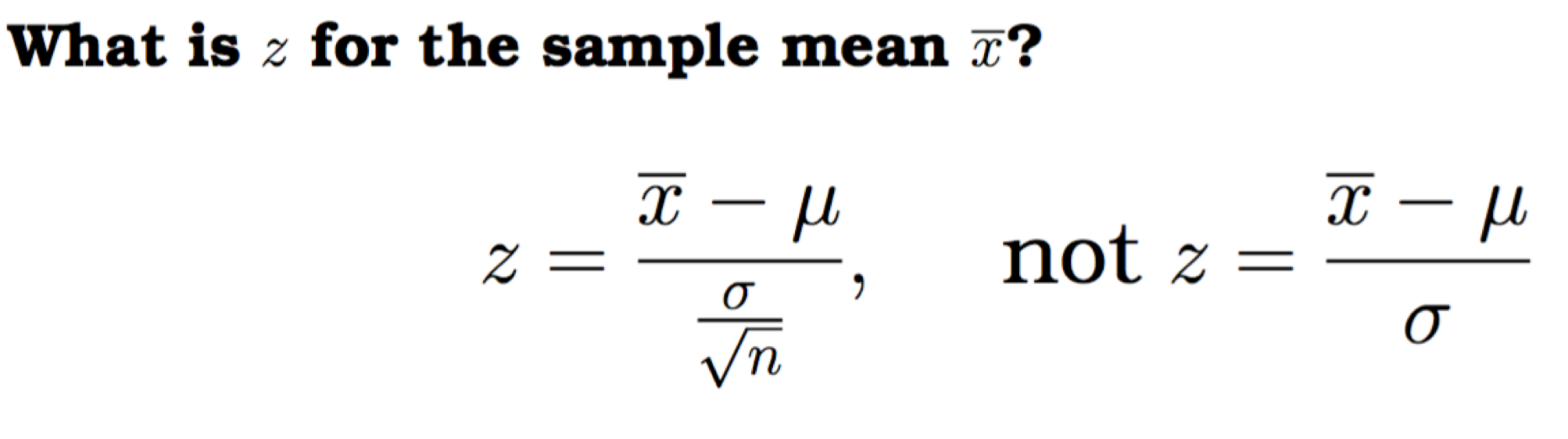
The sampling distribution of $\bar{x}$ (part III)

- 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.



What is z for the sample mean $\overline{x}$?

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}, \quad \text{not } z = \frac{\overline{x} - \mu}{\sigma}$$

Sampling Distribution (1 of 3)

The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from different populations.

a) true

b) false

Sampling Distributions (2 of 3)

____________ is "correct on the average" in many samples. How close the estimator falls to the parameter in most samples is determined by the __________ of the sampling distribution.

a) An unbiased estimator; spread/variability

b) The law of large numbers; mean

c) An unbiased estimator; mean

d) The law of large numbers; median

Sampling Distributions (3 of 3)

- For any population mean μ, the standard deviation of the distribution of $\overline{x}$ gets smaller as we take larger samples. Thus, the results of large samples are less variable than the results of small samples.

a) true

b) false

Standard Deviation of Sample Mean

In 2012, high school seniors reported drinking an average of 3.4 alcoholic drinks with a variance of 4.1, a substantial drop since the late 1990s. A simple random sample of 100 high school seniors is to be taken. What is the standard deviation of $\bar{x}$, the sample mean number of drinks per student?

a) SQR [3.4/100]
b) SQR [4.1]
c) 4.1/3.4
d) SQR [4.1/100]

The central limit theorem

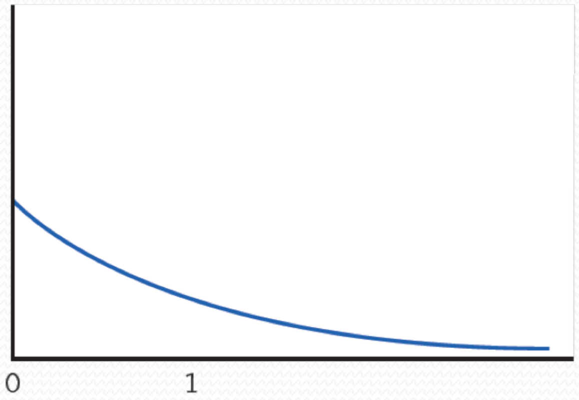
- Most population distributions are not Normal. What is the shape of the sampling distribution of sample means when the population distribution isn't Normal?
- A remarkable fact is that as the sample size increases, the distribution of sample means changes its shape: it looks less like that of the population 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, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \text{ is approximately } N\left(\mu,\ {}^{\sigma}/_{\sqrt{n}}\right)$$

- 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.

Central limit theorem: example (part I)



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?

The central limit theorem states that the sampling distribution of the mean time spent working on the 70 units has:

$$\mu_{\bar{X}} = \mu = 1,$$

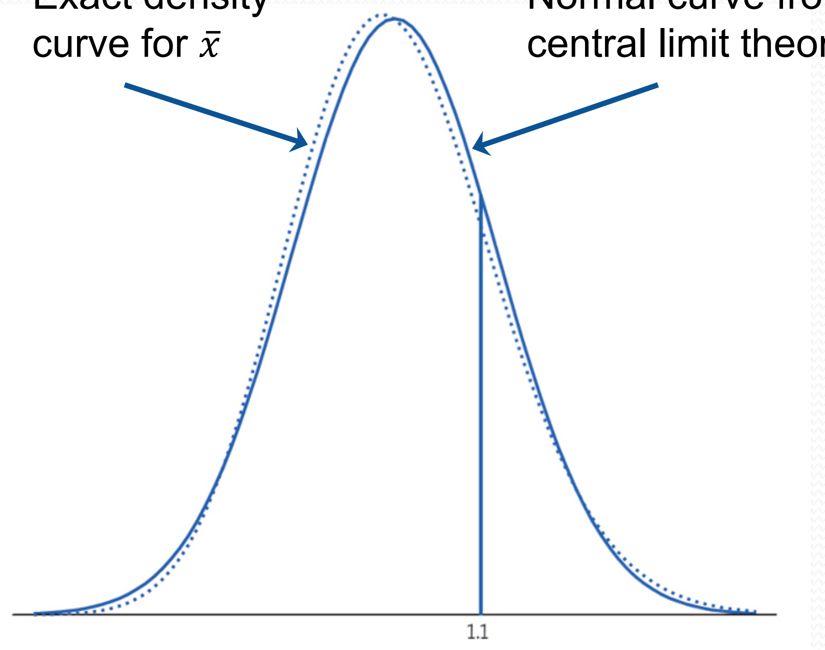
$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1}{\sqrt{70}} = 0.12,$$

and a Normal distribution shape.

Central limit theorem: example (part II)

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

The sampling distribution of the mean time spent working is approximately $N(1,\ 0.12)$ since $n = 70 \geq 30$.



Exact density curve for $\bar{x}$

Normal curve from the central limit theorem

$$z = \frac{1.1 - 1}{0.12} = 0.83$$

$$P(\bar{x} > 1.1) = P(Z > 0.83)$$

$$= 1 - 0.7967 = 0.2033$$

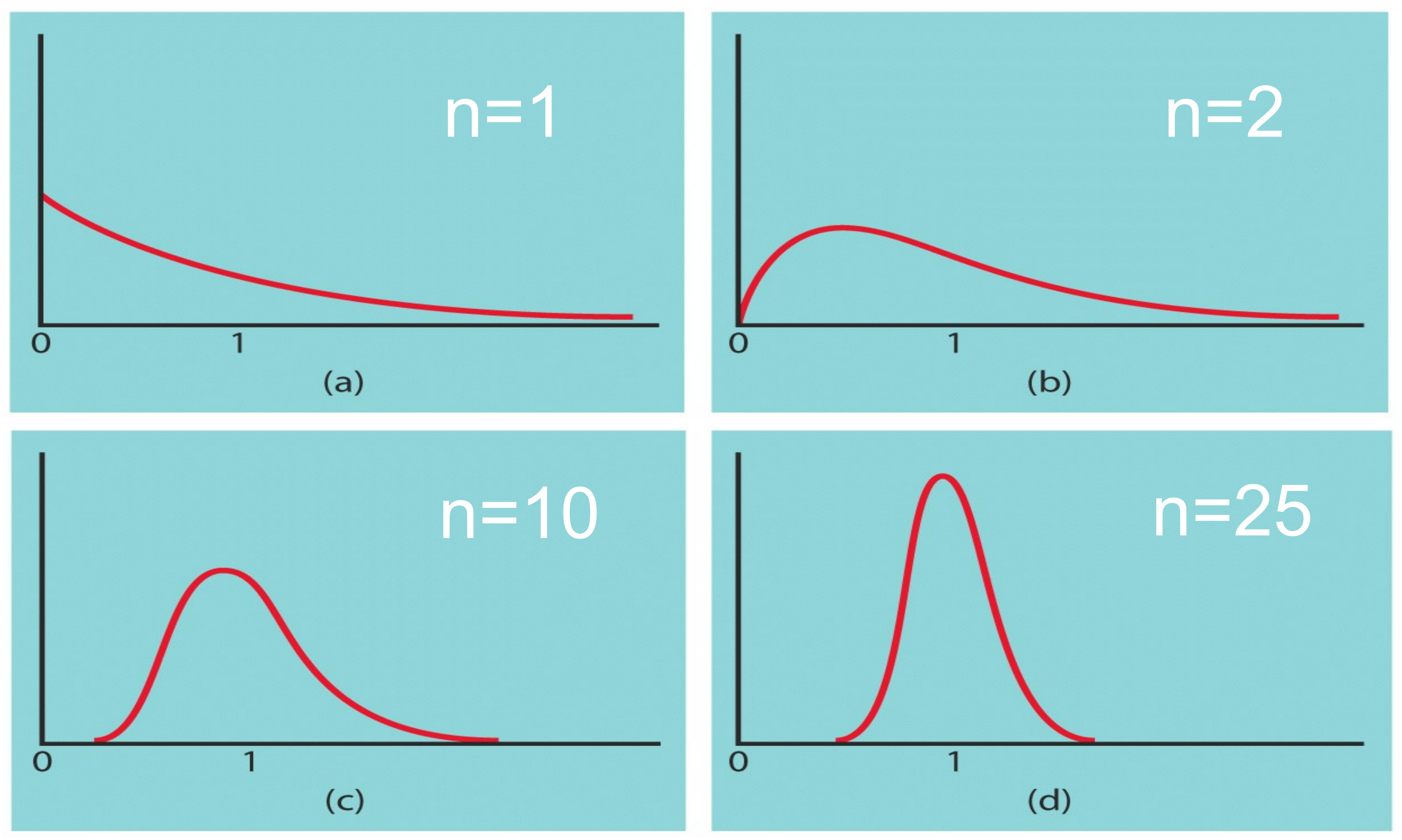
If you budget 1.1 hours per unit, there is a 20% chance the technicians will not complete the work within the budgeted time.

Central Limit Theorem: Sample Size

- How large must *n* be for the CLT to hold?
  - depends on how far the population distribution is from Normal
    - the further from Normal, the larger the sample size needed
    - a sample size of **25 or 30** is typically large enough for any population distribution encountered in practice
    - recall: if the population is Normal, any sample size will work ($n \geq 1$)

Central Limit Theorem:

Sample Size and Distribution of $\overline{x}$



Central Limit Theorem (1 of 2)

- As the sample size increases, the distribution of $\overline{x}$ (sample mean) changes shape: It looks less like that of the population and more like a Normal distribution. When the sample is large enough, the distribution of $\overline{x}$ is very close to Normal. This is true no matter what the shape of the population distribution, as long as the population has a finite standard deviation σ. This famous fact of probability theory is called:

a) the law of large numbers.
b) the Normal theorem.
c) the central limit theorem.
d) the distributive theorem.

Central Limit Theorem (2 of 2)

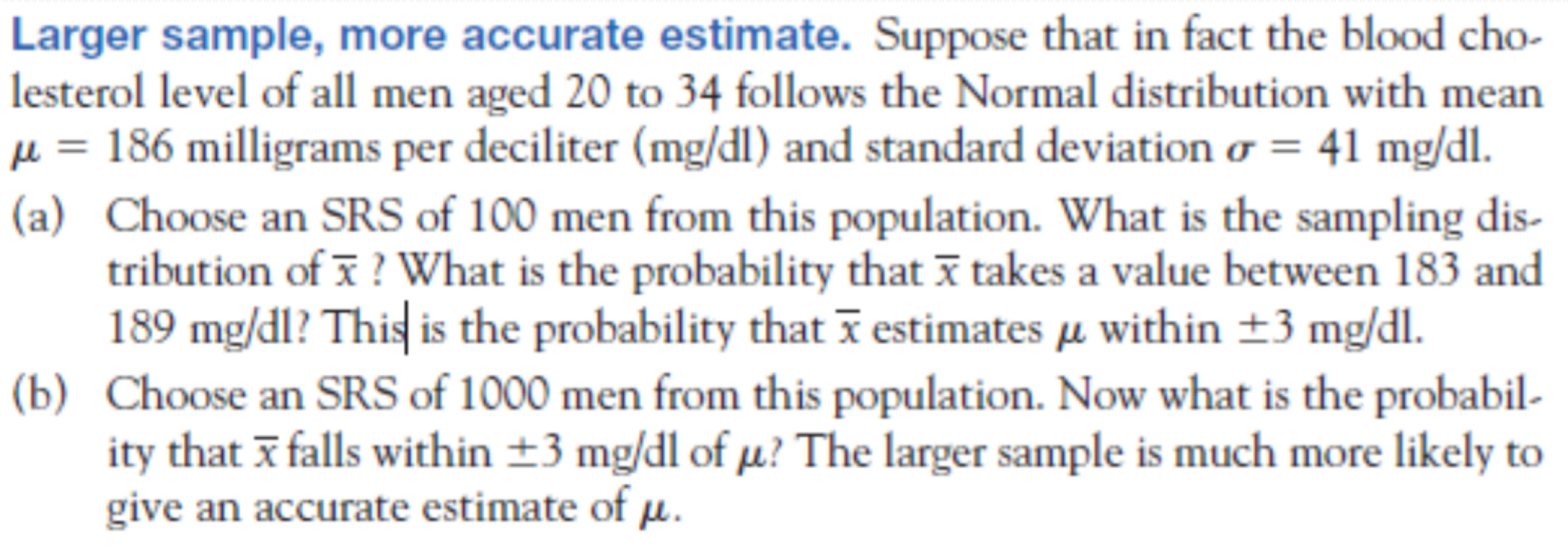
A sample of size 64 is taken from a distribution with mean 100 and standard deviation 24. The sample mean will have a distribution that is approximately ________________ with standard deviation ______________.

a) Normal; 24/64

b) binomial; 24/64

c) Normal; 24/8

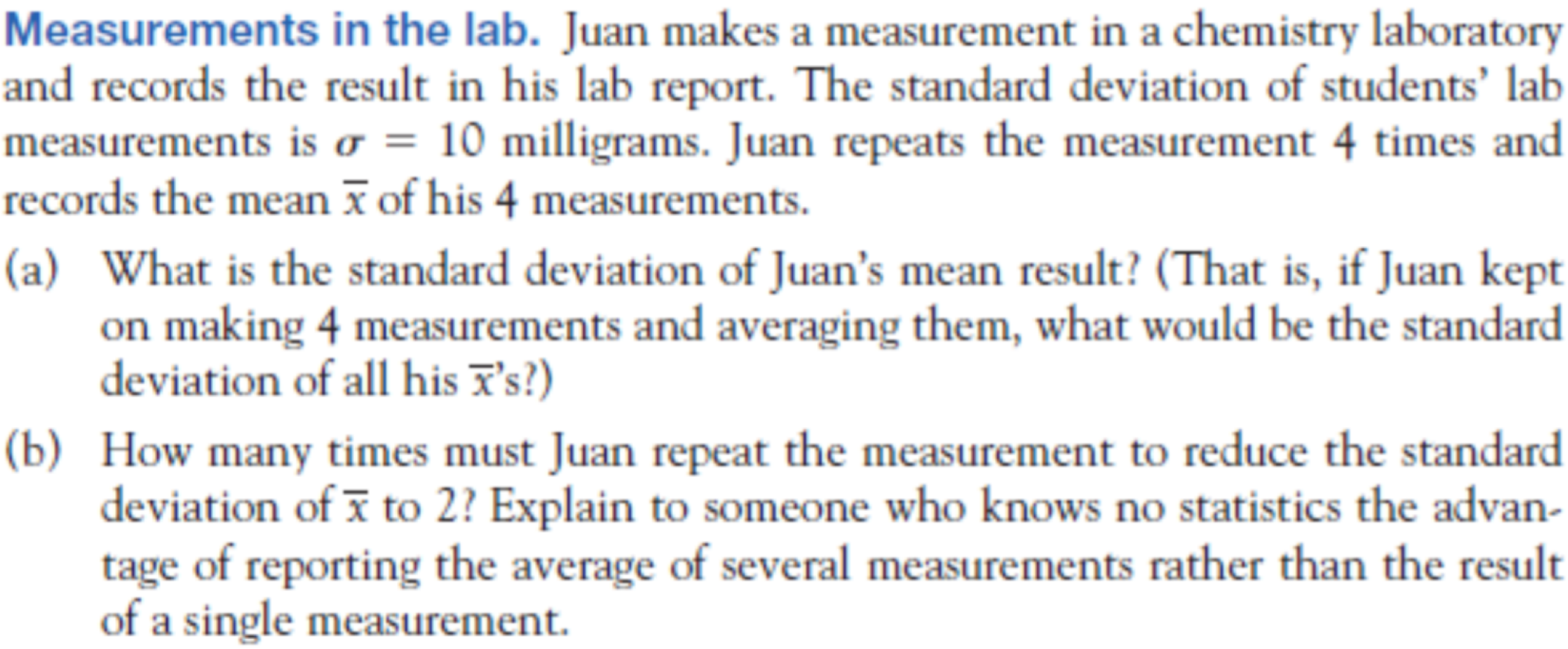
d) binomial; 24/8



Larger sample, more accurate estimate. Suppose that in fact the blood cholesterol level of all men aged 20 to 34 follows the Normal distribution with mean $\mu = 186$ milligrams per deciliter (mg/dl) and standard deviation $\sigma = 41$ mg/dl.

(a) Choose an SRS of 100 men from this population. What is the sampling distribution of $\overline{x}$? What is the probability that $\overline{x}$ takes a value between 183 and 189 mg/dl? This is the probability that $\overline{x}$ estimates μ within ± 3 mg/dl.

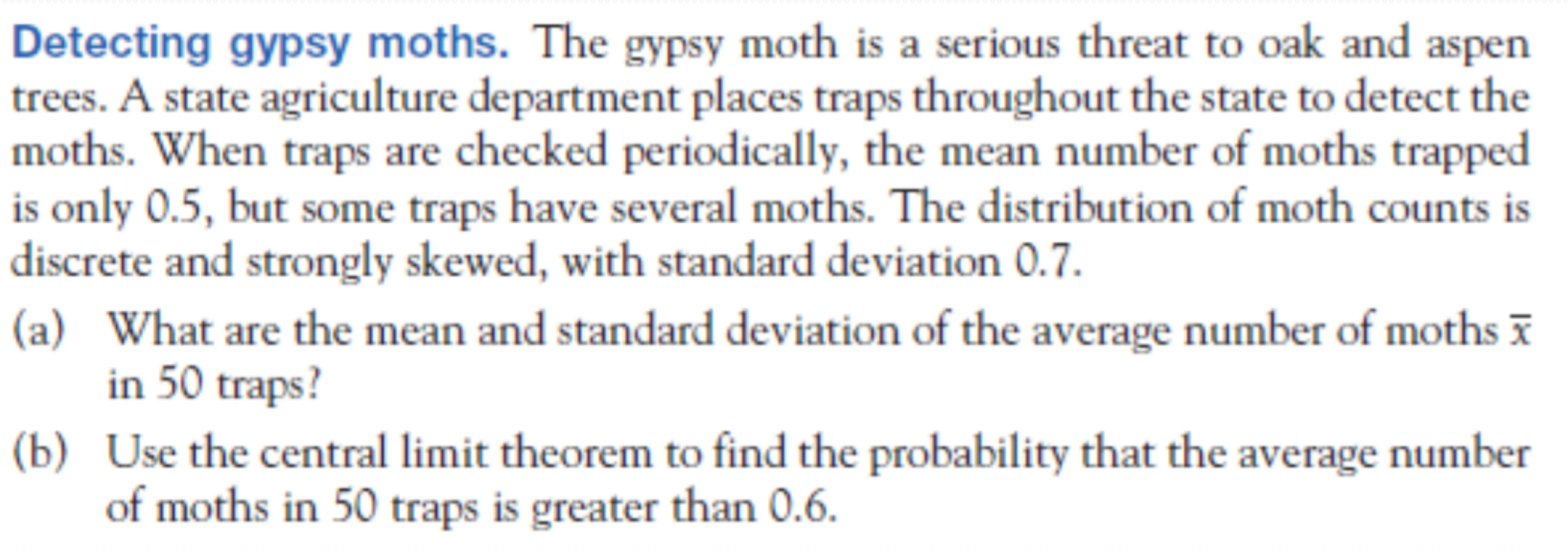
(b) Choose an SRS of 1000 men from this population. Now what is the probability that $\overline{x}$ falls within ± 3 mg/dl of μ? The larger sample is much more likely to give an accurate estimate of μ.



Measurements in the lab. Juan makes a measurement in a chemistry laboratory and records the result in his lab report. The standard deviation of students' lab measurements is $\sigma = 10$ milligrams. Juan repeats the measurement 4 times and records the mean $\bar{x}$ of his 4 measurements.

(a) What is the standard deviation of Juan's mean result? (That is, if Juan kept on making 4 measurements and averaging them, what would be the standard deviation of all his $\bar{x}$'s?)

(b) How many times must Juan repeat the measurement to reduce the standard deviation of $\bar{x}$ to 2? Explain to someone who knows no statistics the advantage of reporting the average of several measurements rather than the result of a single measurement.



Detecting gypsy moths. The gypsy moth is a serious threat to oak and aspen trees. A state agriculture department places traps throughout the state to detect the moths. When traps are checked periodically, the mean number of moths trapped is only 0.5, but some traps have several moths. The distribution of moth counts is discrete and strongly skewed, with standard deviation 0.7.

(a) What are the mean and standard deviation of the average number of moths $\overline{x}$ in 50 traps?

(b) Use the central limit theorem to find the probability that the average number of moths in 50 traps is greater than 0.6.