Lesson 7

Sampling Distribution of the Mean and the Central Limit Theorem

Review: Probability Distributions

- Any characteristic that can be measured or categorized is called a *variable*.
- If the variable can assume a number of different values such that any particular outcome is determined by chance it is called a *random variable.*
- Every random variable has a corresponding *probability distribution*.
- The probability distribution applies the theory of probability to describe the behavior of the random variable.

Review: Probability Distributions

- So far we've covered the following probability distribution
  - Probability tables to describe the distributions of Nominal variables
  - Probability density curves for continuous variables – particularly the Normal Distribution
  - Probability distributions for discrete variables including
    - Binomial Distribution
    - Poisson Distribution

Sampling Distributions

- Review: a statistic is a numerical value used as a summary measure for a sample

- Statistics are random variables that have different values from sample to sample
- Since statistics are random variables, they have probability distributions
- Probability distributions for *statistics* are called sampling distributions.

Sampling Distribution of a Statistic

- Definition of sampling distribution: The probability distribution of a statistic that results from all possible samples of a given size is the sampling distribution of the statistic.

- If you know the sampling distribution of the statistic you can generalize results from samples to the population using
  - Confidence intervals for estimating parameters
  - Hypothesis tests
  - Other statistical inference methods

Preview of Sampling Distributions

The sampling distributions we will cover in this course are

- Normal distribution for
  - Sample mean if population standard deviation is known
  - Sample Proportion if $n*\pi > 5$ and $n*(1-\pi) > 5$ (Lesson 9)
- t-distribution for
  - Sample mean if standard deviation is estimated from the sample
- Chi-square distribution for
  - Chi-square statistic used to test for independence between two categorical variables (Lesson12)
- F-distribution for
  - Ratio of two variances (Lesson 13)

Lesson 7 Outline

- Sampling Distribution of the Sample Mean
- Central Limit Theorem
- SE of the mean
- Calculating probabilities from the sampling distribution of the mean
- Introduction to t-distribution

Sampling vs. Population Distribution

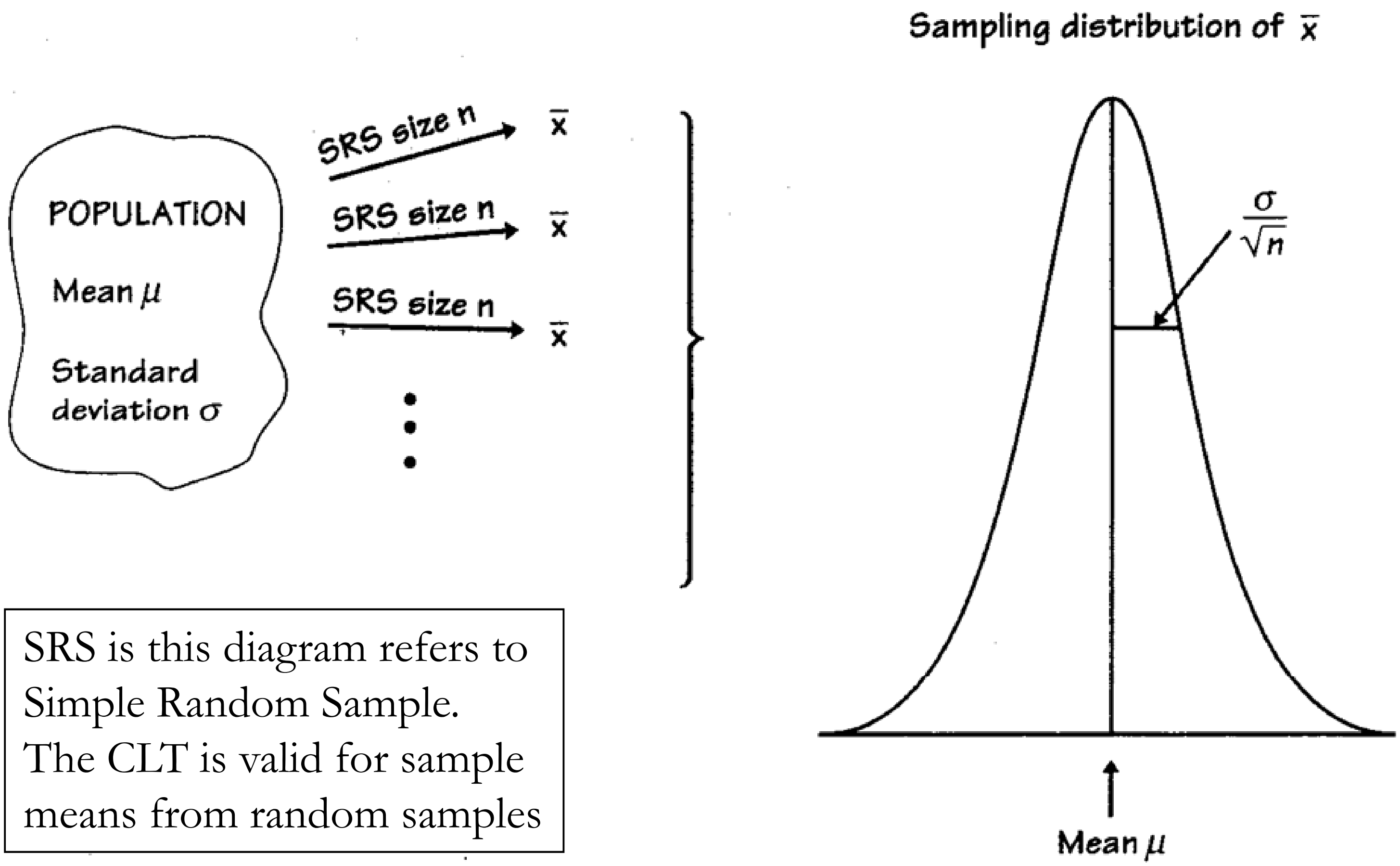
- The distribution of the individual observations: the population distribution

- The distribution of the sample means derived from samples drawn from the population: the sampling distribution

Central Limit Theorem

- The Central Limit Theorem (CLT) is based on the sampling distribution of sample means from all possible samples of size n drawn from the population.
- However, to apply the CLT it isn't necessary to generate all possible samples of size n and calculate the mean for each sample to determine the sampling distribution.
- If the population mean and standard deviation are known, the sampling distribution of the sample mean is also known:

$$\bar{x} \sim N\left(\mu, \ \frac{\sigma}{\sqrt{n}} \right)$$



SRS is this diagram refers to Simple Random Sample. The CLT is valid for sample means from random samples

FIGURE The sampling distribution of a sample mean $\bar{x}$ has mean μ and standard deviation $\sigma/\sqrt{n}$. The distribution is normal if the population distribution is normal; it is approximately normal for large samples in any case.

Central Limit Theorem Illustration

- View the following website for an interactive illustration of the Central Limit Theorem

  http://davidmlane.com/hyperstat/A14043.html

Standard Error of the Mean

- The Standard Error of the Mean (SEM) measures the variability in the sampling distribution of sample means

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

- As sample size increases, the SEM decreases because the sample size (n) is in the denominator of the calculation

Standard Deviation vs. Standard Error

- *Standard deviation* measures the variability in the population and is based on measurements of individual observations
- *Standard error* is the standard deviation of the statistic and measures the variability of the statistic from repeated samples
  - The sampling distribution of ANY statistic has a standard error. The SEM is the SE for the sampling distribution of the mean

Birth weight example

- Birth weights over a long period of time at a certain hospital show a normal distribution with mean μ of 112 oz and a standard deviation σ of 20.6 oz.

- Calculate the probability that the next infant born weighs between 107 and 117 oz.
  - This is calculating the probability for an individual – use the population distribution given above

- Calculate the probability that the mean birth weight for the next 25 infants is between 107 and 117 oz.
  - This is calculating the probability for a sample mean – use the CLT and sampling distribution of the sample mean for samples of size n= 25

Distribution of Birth weights Single Observation

- Birth weights over a long period of time at a certain hospital show a normal distribution with mean μ of 112 oz and a standard deviation σ of 20.6 oz.
- Calculate the probability that the next infant born weighs between 107 and 117 oz.

Distribution of birth weights

- Since birth weights are normally distributed, the NORMDIST function in Excel can be used to find the probability that the next infant born is between 107 – 117 oz

=NORMDIST(117,112,20.6,1)=NORMDIST(107,112,20.6, 1)

= 0.192

- The probability that the next infant born weighs between 107 – 117 oz. = 0.192
- Since this is population data we can also state that 19.2% of infants born at this hospital weigh between 107 – 117 oz.

Distribution of Birth weights Mean Birth Weight

- Birth weights over a long period of time at a certain hospital show a normal distribution with mean μ of 112 oz and a standard deviation σ of 20.6 oz.
- Calculate the probability that the mean birth weight for the next 25 infants is between 107 and 117 oz.

Sampling Distribution of Mean Birth weights

- Birth weights over a long period of time at a certain hospital show a mean μ of 112 oz and a standard deviation σ of 20.6 oz.
- Calculate the probability that the <u>mean</u> birth weight of the next 25 infants born will fall between 107 and 117 oz.
- First determine the mean and SEM for this sampling distribution. From the CLT we know:

$$\bar{x} \sim N\left(112,\ \frac{20.6}{\sqrt{25}}\right) \text{ with standard error for } \bar{\text{x}} = \frac{20.6}{\sqrt{25}} = 4.12$$

Birth weight example

- Since the sampling distribution of sample means has a normal distribution, the NORMDIST function in Excel can be used to find the probability that the mean birthweight of the next 25 infants is between 107 – 117 oz

=NORMDIST(117,112,4.12,1)=NORMDIST(107,112,4.12, 1)

= 0.774

- The probability that the mean birth weight of the next 25 infants is between 107 and 117 oz = 0.774

Comparing the Results of the Two Probability Calculations

- P (next infant born weighs 107-117 oz) = 0.192
- P (mean weight of next 25 infants is 107-117) = 0.774

- The probability that the mean birth weight for 25 infants is between 107 – 117 oz. is greater than the probability that an individual birth weight is between 107 – 117 oz because the sampling distribution of the mean is less dispersed than the population distribution.

Wing length example

- The distribution of wing lengths of butterflies in Baja, CA has a mean value (μ) of 4 cm and variance (σ^2) of 25 cm^2

- *We don't know if butterfly wing length is normally distributed or not.*

- What is the probability that a sample mean wing length calculated from 64 butterflies will fall between 3.5 cm and 4.5 cm?

Sampling distribution of mean wing lengths

- We know from the CLT that, regardless of the distribution of the population data, the sample means from a sample of size 64 are <u>normally</u> distributed with
  - Mean = population mean = 4

$$\mathrm{SE} = \frac{5}{\sqrt{64}} = 0.625$$

- We want to find the probability that the sample mean wing length is between 3.5 and 4.5 cm

$$3.5 \le \bar{x} \le 4.5$$

Sampling distribution of mean wing length

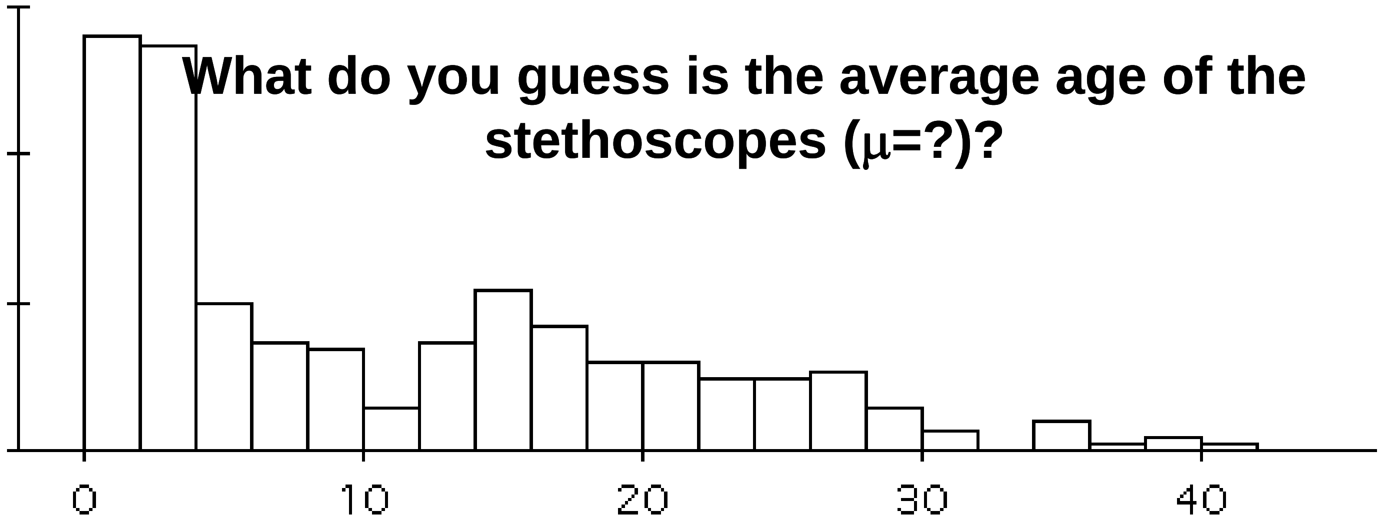
- Normal distribution
- Mean = 4
- SEM = 0.625

- Use the NORMDIST function to find the area between 3.5 and 4.5

  NORMDIST(4.5, 4, 0.625, 1) – NORMDIST(3.5, 4, 0.625, 1) = 0.576

- The probability that the mean wing length for a sample of 64 butterfly wings is between 3.5 and 4.5 cm = 0.576.

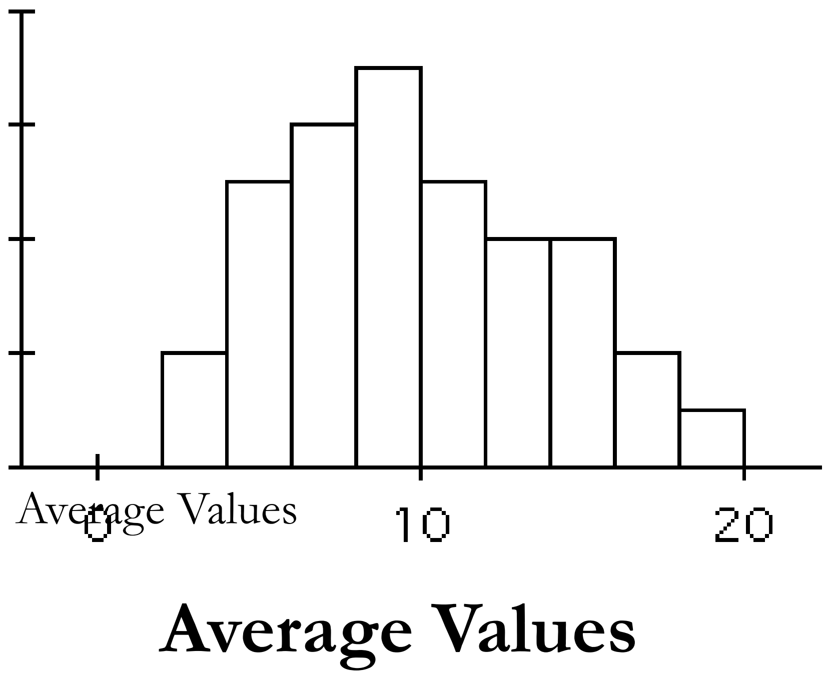
Age of ALL Stethoscopes in a Hospital System



What do you guess is the average age of the stethoscopes (μ=?)?

1. 5 years
2. 10 years
3. 15 years
4. 20 years
5. 25 years

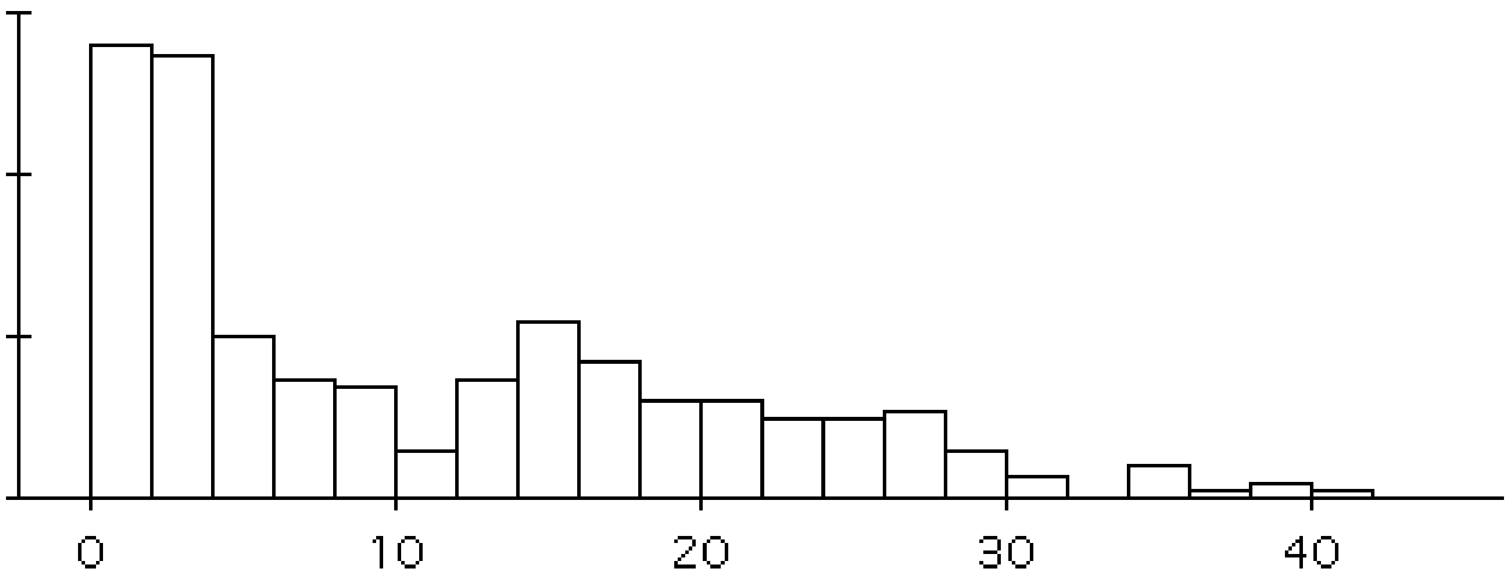
What do you guess is the average age of the stethoscopes?



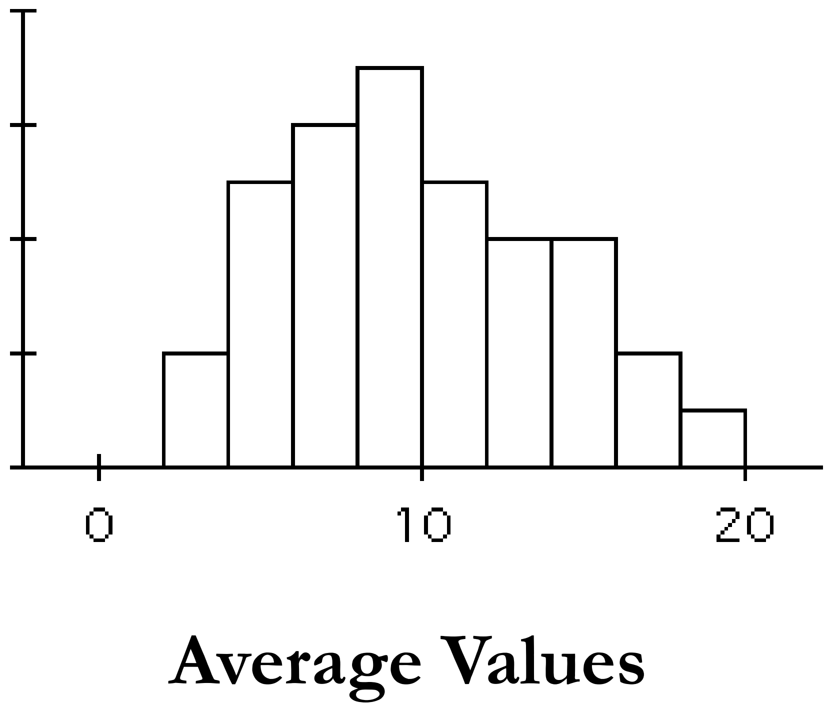
1. 5 years
2. 10 years
3. 15 years
4. 20 years
5. 25 years

Age of ALL Stethoscopes in a Hospital System

Suppose standard deviation of the age of the stethoscopes is $\sigma = 18$.

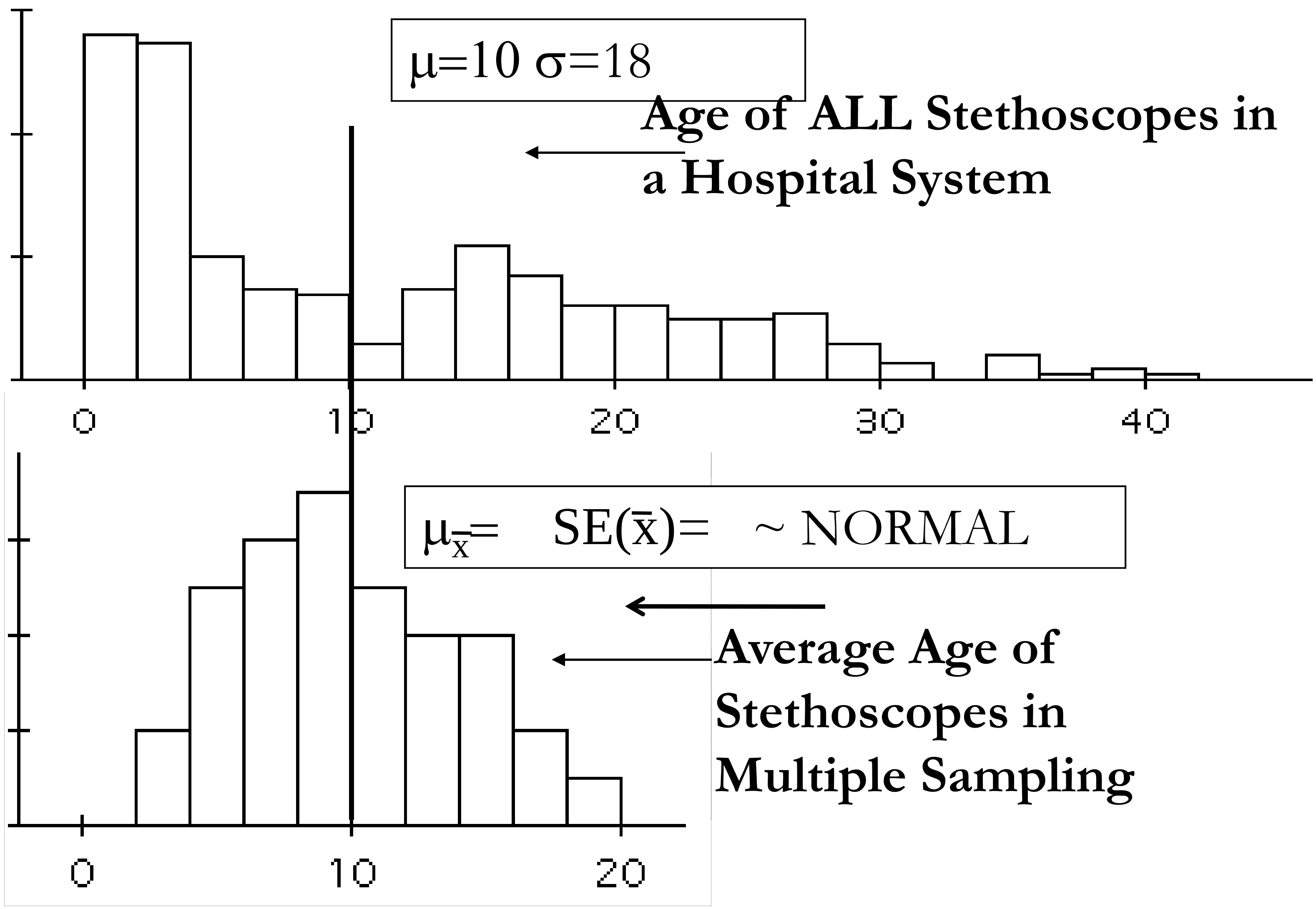


Recall: you took many samples of size n=36 and plotted these averages. What is the standard deviation of these possible averages (called the standard error)?



1. 2
2. 3
3. 5
4. 15

The Central Limit Theorem!!



Sampling distribution when σ is unknown

- The sampling distribution of the sample mean is normally distributed (n>30) when the population standard deviation (σ) is known.

- Often σ is unknown and is estimated by the sample standard deviation (*S*).

- When σ is unknown the SE is calculated as

$$\mathrm{SE} = \frac{S}{\sqrt{n}}$$

The CLT when σ is unknown

- The sampling distribution of the sample mean is distributed as a Student's T distribution with n-1 degrees of freedom when the population standard deviation is unknown and one of these conditions is met:
  - N> 40 and the population is unimodal
  - N>15 and the population is approximately normal.
  - N any size, and the population is normal.

$$\bar{x} \sim t_{n-1}\left(\mu,\ \frac{s}{\sqrt{n}}\right)$$

History of Student's t-distribution

- William Gosset 1876-1937

- Gosset worked as a chemist in the Guinness brewery in Dublin and did important work on statistics. He discovered the form of the t-distribution and invented the t-test to handle small samples for quality control in brewing. He published his findings (in July 1908!) under the name "Student".

- This is why the t-test is sometimes referred to as "Student's t-test."

t-distribution

- The t-distribution is similar to the standard normal distribution
  - It is unimodal and symmetric about the mean
  - The mean of the t-distribution = 0
  - The total area under the t-distribution curve = 1
- There are some differences between the t-distribution and the standard normal
  - The t-distribution has larger tail areas (the tail areas are the areas at either end of the curve)
  - The t-distribution is Indexed by degrees of freedom (df) which are equal to the sample size - 1.

t-distribution and standard normal distribution

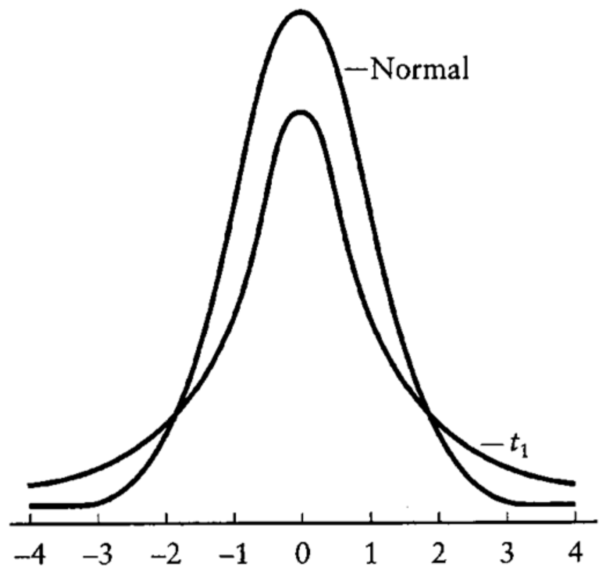


FIGURE 9.2
The standard normal distribution and Student's *t* distribution with 1 degree of freedom

As n increases, the t-distribution is closer to the Standard Normal distribution
Values of the t-distribution are called t-coefficients

Areas under the t-distribution curve

- The Excel function TDIST will give the tail area under the t-curve for a specified t-coefficient.

  TDIST(t, deg_freedom, 1 or 2)
- Specify the value of the t-coefficient (t), the degrees of freedom (n-1) and
  - ‘1’ to obtain the area in one tail or
  - ‘2’ to obtain the area in both tails.
- The TDIST function only accepts positive values of the t-coefficient. By symmetry the area in the positive ‘tail’ is equal to the area in the negative ‘tail’

TDIST function examples

- P(T>2.5) for a t-distribution with 9 degrees of freedom: TDIST(2.5, 9, 1) = 0.017

- P(T<-2.5) for a t-distribution with 9 degrees of freedom: TDIST(2.5, 9, 1) = 0.017

- P(T<-2.5) + P(T>2.5) for a t-distribution with 9 degrees of freedom: TDIST(2.5, 9, 2) = 0.034

NOTE: TDIST function operates differently from the NORMSDIST function

Compare t distribution areas to Standard Normal Curve Areas

- The area > z-score of 2.5 on the standard normal curve:

  1 – NORMSDIST(2.5) = 0.0062

- The area < z-score of -2.5 on the standard normal curve:

  NORMSDIST(-2.5) = 0.0062

- The two tail areas: area beyond $\pm$ 2.5

  2*NORMSDIST(-2.5) = 0.0124

Notice that the tail areas are greater for the t-distribution with 9 df than for the standard normal distribution.

t-coefficient notation

t-coefficients are notated with the degrees of freedom as a subscript

- t-coefficient for a sample size of 10: t_9
- t-coefficient for a sample size of 18: t_{17}

Sampling Distribution of the Sample Mean: Overview

- By the Central Limit Theorem, means of observations with known standard deviation are normally distributed for large enough sample sizes regardless of the distribution of the observations.
- When the population standard deviation is unknown, the sampling distribution of the means is a t-distribution with n-1 df.
- The SEM measures the variability of the distribution of means.

Readings and Assignments

- Reading:
  - Chapter 4 pgs 80 – 90
  - Chapter 5 pgs. 95-101
- Spend some time thinking about the relationships between the population distribution and the sampling distribution of sample means
  - Draw pictures of the distributions
  - Work through Lesson 7 Practice Exercises
  - Excel Module 7 – examples of sampling distribution of the mean when σ is known.