Chapter 11 – Sampling Distribution

Stat 115

Definition 11.1 : Random Sample (finite population)

Suppose we select *n* distinct elements from a population consisting of *N* elements, using a particular probability sampling method.

Let X_1=measure taken from the 1st element in the sample

X_2=measure taken from the 2nd element in the sample

…

X_n=measure taken from the nth element in the sample

Then, (X_1, X_2,…,X_n) is called a **random sample of size n from a finite population**.

Definition 11.2 : Random Sample (infinite population)

Let X_1=measure taken from the 1st element in the sample
X_2=measure taken from the 2nd element in the sample
…
X_n=measure taken from the nth element in the sample

Then $(X_1, X_2,\ldots,X_n)$ is called a **random sample of size n from an infinite population** if the values of $X_1, X_2,\ldots,X_n$ are n independent observations generated from the same cumulative distribution function (cdf), F(**.**). This common cdf or its corresponding probability mass/density function, f(**.**), is called the **parent population** or the **distribution of the population.**

Definition 11.3: Statistic

Suppose $(X_1, X_2,\ldots,X_n)$ is a random sample. A **statistic** is a random variable that is a function of $X_1, X_2,\ldots,X_n$.

Example 11.1: Statistic

Suppose (X_1, X_2,…,X_n) is a random sample.

a) $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ is a random variable that is a function of X_1, X_2,…,X_n. Thus, $\bar{X}$ is a statistic.

b) $S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$ is a random variable that is a function of X_1, X_2,…,X_n. Thus, S^2 is a statistic.

Definition 1.4: Sampling Distribution of a Statistic

The **sampling distribution of a statistic** is its probability distribution.

Example 1: Sampling Distribution of the Mean

Consider 4 observations making up the population values of a random variable X having the probability distribution

$f(x) = ¼,\ x = 0,1,2,3$

Note that $\mu = E(X) = 3/2$ and $\sigma^2 = Var(X) = 5/4$

Con't of Example

Suppose we list all possible samples of size 2, with replacement, and for each sample compute for the value of the sample mean, $\overline{X}$:

No.	Sample	$\overline{X}$
1	0,0	0.0
2	0,1	0.5
3	0,2	1.0
4	0,3	1.5
5	1,0	0.5
6	1,1	1.0
7	1,2	1.5
8	1,3	2.0

No.	Sample	$\overline{X}$
9	2,0	1.0
10	2,1	1.5
11	2,2	2.0
12	2,3	2.5
13	3,0	1.5
14	3,1	2.0
15	3,2	2.5
16	3,3	3.0

$\overline{x}$

=

)

Con’t of Example

Find the sampling distribution of the mean.

$$\overline{X}$$

$$P(\overline{X} = \overline{x})$$

Read Examples 11.2, 11.3, and 11.4

Definition 11.5: Standard error

The standard deviation of a statistic is called its **standard error**.

Standard error is a measure of reliability of the statistic.

Notes on Standard Error

- A small standard error indicates that the computed values of our statistic in the different samples generated are close to one another, so that even if we know that the value of a statistic varies from one sample to another, a small standard error gives us an assurance that at least the variation among their values is not too large.

- A small standard error means that the realized values of the statistic under repeated sampling are, on the average, very close to the average value of the statistic.

Theorem 11.1

If $(X_1, X_2,\ldots,X_n)$ is a random sample from a **finite population** of size N and whose values are generated using simple random sampling without replacement then $E(\bar{X})=\mu$ and

$Var(\bar{X})= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$, where $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$ and

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}.$$

Theorem 11.2

If (X_1, X_2,…,X_n) is a random sample from an **infinite population** then E($\overline{X}$)=μ and

Var($\overline{X}$)= $\frac{\sigma^2}{n}$, where $\mu = E(X_i)$ and $\sigma^2 = E(X_i - \mu)^2$.

Using Example 1

Find the mean of $\overline{X}$ and standard error of $\overline{X}$

Read Examples 11.6, 11.7

Some Notes on Theorems 11.1 and 11.2

- Both theorems state that E($\overline{X}$) = μ. This means that in both sampling schemes, the values of $\overline{X}$ in the different samples generated are all centered about μ, the parameter we are estimating.
- The standard error of $\overline{X}$ is smaller for populations where σ^2 is small.

Some Notes on Theorems 11.1 and 11.2

- Theorems 11.1 and 11.2 provide us with mathematical evidence to our claim in Chapter 3 that our estimates under simple random sampling (with or without replacement) are more reliable when the elements in the population are homogeneous with respect to the characteristic under study.
- Increasing the sample size, n, in either sampling scheme will decrease the standard error of $\overline{X}$

Some Notes on Theorems 11.1 and 11.2

- This tells us that if we use either SRSWR or SRSWOR and we want the value of $\overline{X}$ from one sample to the other to be very near μ, then all we need to do is increase the sample size, n.

Homework

- DO Exercises for Section 11.1, p. 390-391, nos. 1 and 2

11.2 Central Limit Theorem

Theorem 11.3 Central Limit Theorem

If $\bar{X}$ is the mean of a random sample of size n from a **large or infinite population** with mean μ and variance σ^2, then the sampling distribution of $\bar{X}$ is approximately normally distributed with mean, $E(\bar{X})=\mu$, and variance, $Var(\bar{X})=\sigma^2/n$, when n is sufficiently large.

Notes: CLT

- CLT states that when the sample size is sufficiently large, we can use the normal distribution to approximate the sampling distribution of $\overline{X}$.
- What is remarkable about the CLT is that it does not state any requirement about the distribution of the population, aside from having mean μ and variance σ^{2}.

Notes: CLT

- The normal approximation will hold for population distributions that are either discrete or continuous.
- The normal approximation will hold for population distributions that are either symmetric or skewed.
- We can use the approximation even for random samples from finite populations so long as N is very large.

Notes: CLT

- When do we consider the sample size, n, to be sufficiently large? In most situations, the normal approximation will be good if $n \geq 30$.
- If the distribution of the population is not very different from the normal distribution, then the approximation will be good even if $n < 30$.
- In fact, if the population is normally distributed, then $\overline{X}$ will be normally distributed even for a sample of size 1.

Example: CLT

An electrical firm manufactures electric light bulbs that have a length of life which is normally distributed with mean and standard deviation equal to 500 and 50 hours, respectively. Find the probability that a random sample of 15 bulbs will have an average life of less than 475 hours.

T-distribution

t-distribution

- Very similar to the standard normal distribution
- The graph of the pdf of the t-distribution is also a bell-shaped curve that is symmetric about 0.
- Its tails will also approach the x-axis without ever touching it

t-distribution

Difference with Standard normal distribution:

- t- distribution has a larger variance than a standard normal distribution
- However, as the degrees of freedom increases, the variance of the t-distribution goes to 1, so that when the degrees of freedom is at least 30, the t-distribution is almost the same as the standard normal distribution.

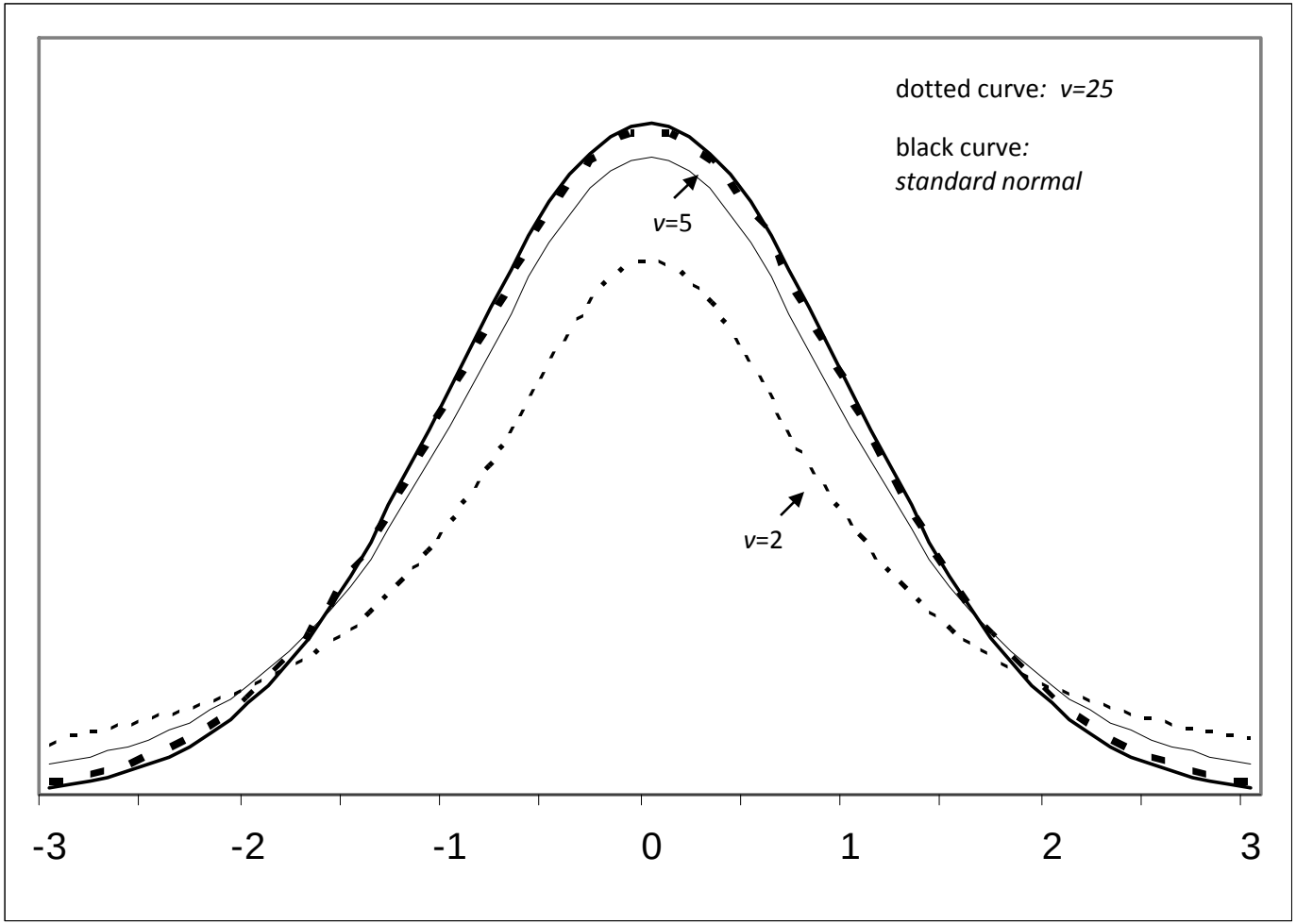

FIGURE 11.1. GRAPH OF THE T-DISTRIBUTION WITH VARYING DEGREES OF FREEDOM (*V*) AND THE STANDARD NORMAL DISTRIBUTION

t-distribution

If $\bar{X}$ and s^2 are the mean and variance, respectively, of a random sample of size n taken from a population which is normally distributed with mean μ and variance σ^2, then

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

is a random variable having the t-distribution with v = n – 1 degrees of freedom

- Notation: $T \sim t_{\alpha, v=n-1}$

t- distribution

Area under the curve

- Just like any continuous probability distribution, the probability that a random sample produces a t-value falling between any two specified values is equal to the area under the curve of the t-distribution between any two ordinates corresponding to the specified values.

t-distribution

Notation:

- t_{α} is the t-value leaving an area of α in the right-tail of the t-distribution. That is, $T \sim t_{\alpha,v}$ then t_{α} is such that $P(T > t_{\alpha,v}) = \alpha = P(T < -t_{\alpha,v})$

- Since the t-distribution is symmetric about zero, $t_{1-\alpha,v} = -t_{\alpha,v}$

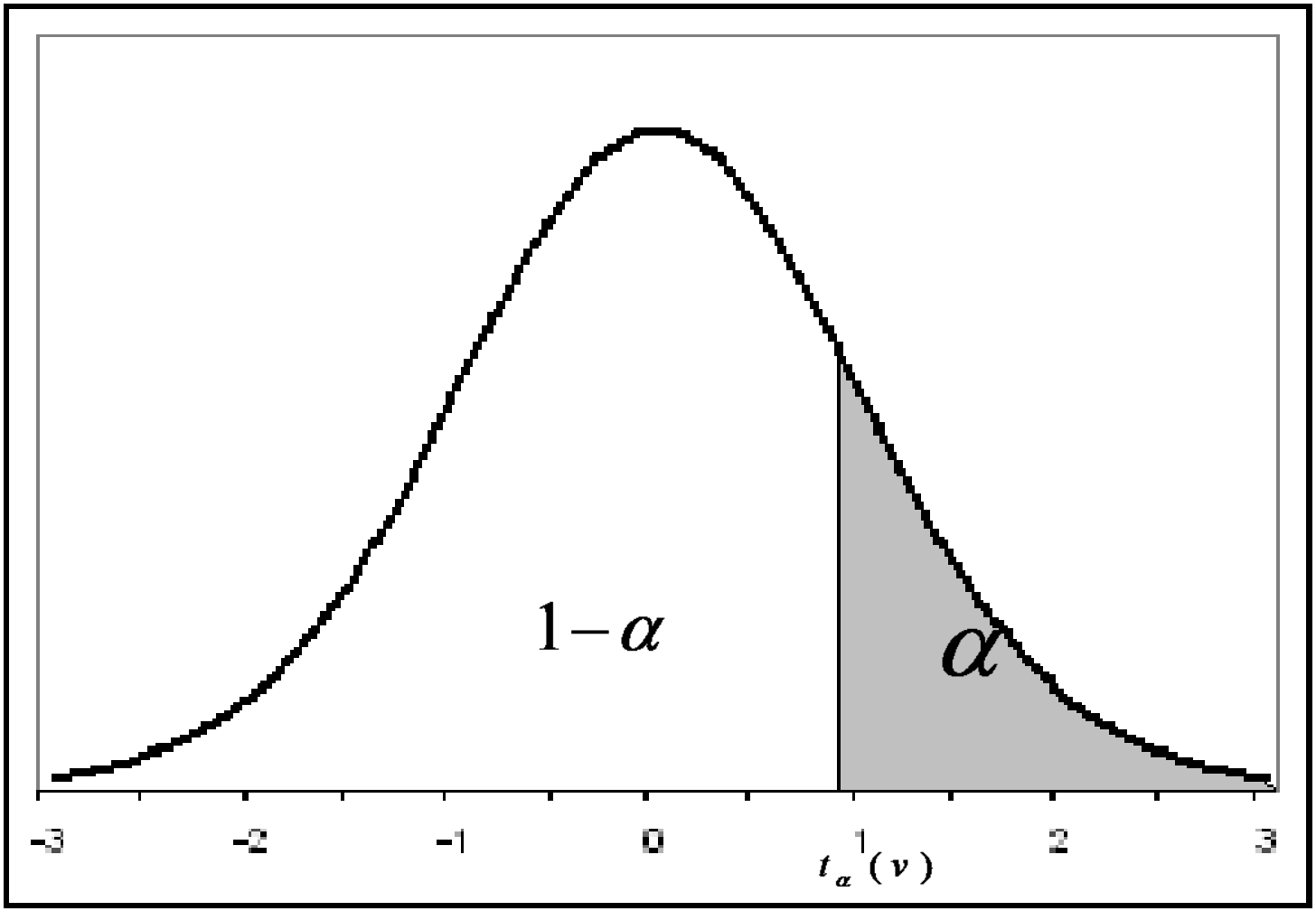

FIGURE 11.2. GRAPHICAL REPRESENTATION OF THE $100(1-\alpha)^{TH}$ PERCENTILE, $T_\alpha(V)$.

Examples: t- distribution

1. Find the following values on the t-table:
   a) $t_{0.025}$ when v = 14
   b) $t_{0.99}$ when v = 10

2. Find k such that P(k < T < 2.807) = 0.945 when T ~ $t_{\alpha,23}$

3. A manufacturing firm claims that batteries used in their electronic games will last an average of 30 hours. To maintain this average, 16 batteries are tested each month. If the computed t-value falls between $-t_{0.025}$ and $t_{0.025}$, the firm is satisfied with its claim. What conclusion should the firm draw from a sample that has mean $\overline{X}$ = 27.5 hours and standard deviation s = 5 hours? Assume the distribution of battery lives to be approximately normal.

Read Example 11.10

Chi-Square Distribution

- Just like the t-distribution, the chi-square distribution has a single parameter called the DEGREES OF FREEDOM.
- If X is a random variable that follows a chi-square distribution with v degrees of freedom, then we write $X \sim \chi^2_{(v)}$ where χ is the small Greek letter chi (read as ki).

Chi-square Distribution

- The pdf of the chi-square distribution is positive for positive real numbers only; elsewhere, its value is 0.
- Its mean is equal to its degrees of freedom.
- Its variance is twice its degrees of freedom.
- Thus, as the degrees of freedom increases, both the mean and variance will also increase.

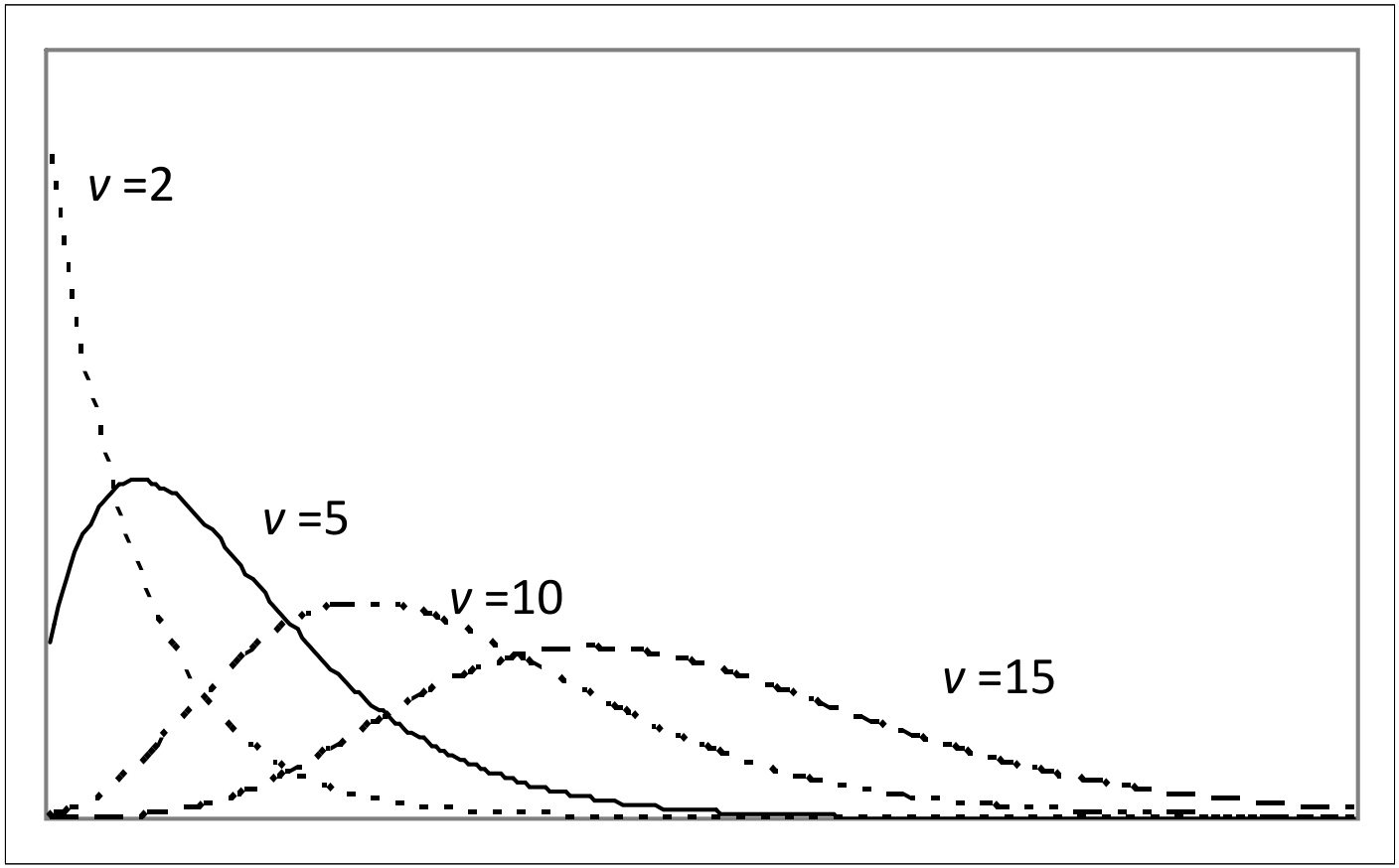

FIGURE 11.3. GRAPH OF THE CHI-SQUARE DISTRIBUTION WITH VARYING DEGREES OF FREEDOM (*V*).

Chi-square distribution

- Skewed to the right
- Its skewness is more pronounced for smaller degrees of freedom.
- As the degrees of freedom increases, its distribution becomes more symmetric.
- If $X \sim \chi^2_{(v)}$, then $\chi^2_{(v)}$ satisfies the condition that the $P(X \leq \chi^2_{(v)}) = 1 - \alpha$, or equivalently, $P(X > \chi^2_{(v)}) = \alpha$.

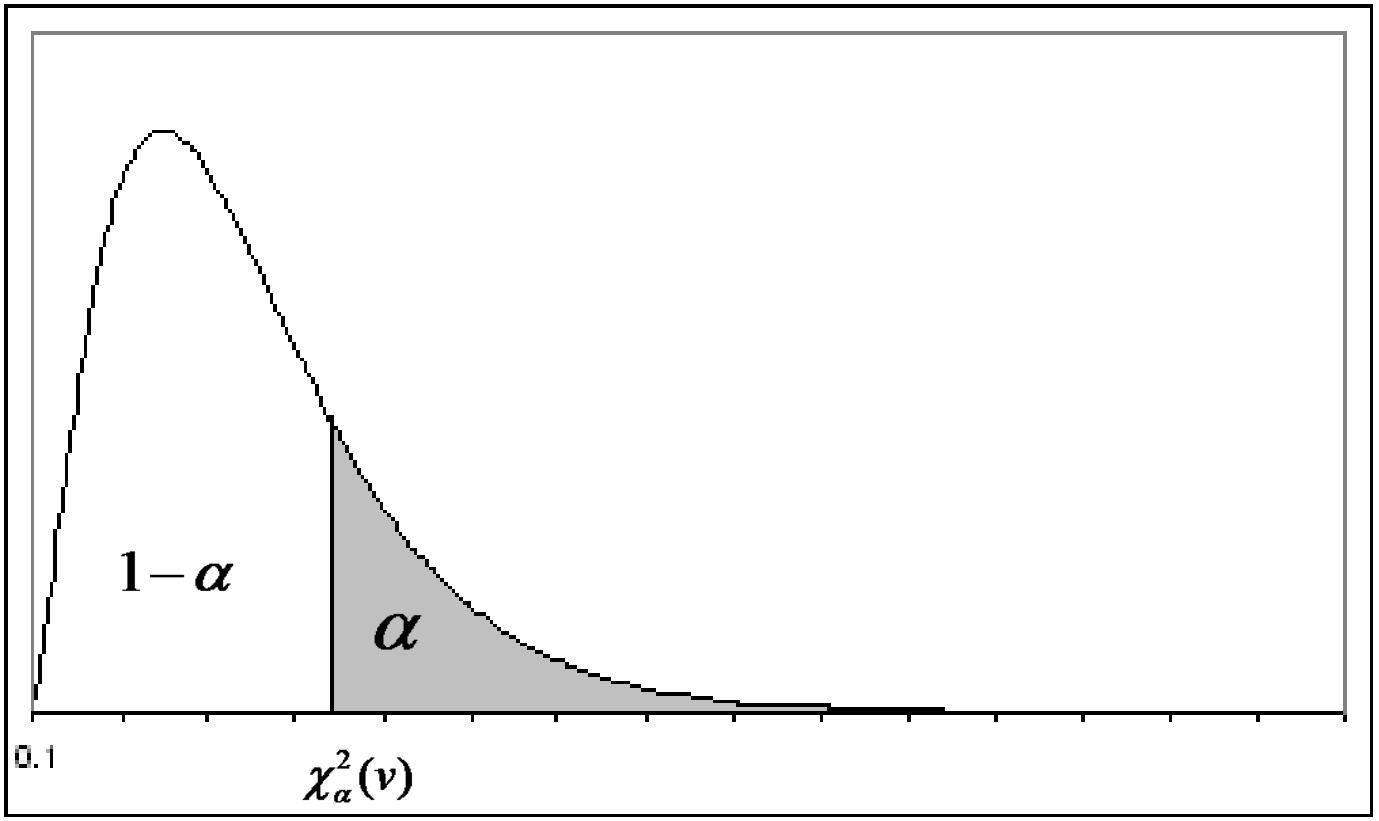

FIGURE 11.4. GRAPHICAL REPRESENTATION OF THE $100(1-\alpha)^{TH}$ PERCENTILE, $\chi^2_\alpha(V)$.

Example 11.11

Suppose 11.11: Suppose $X \sim \chi^2_{\alpha(V=10)}$. WE CAN USE TABLE B.3, APPENDIX B, TO DETERMINE THE FOLLOWING:

A) $P(X > 18.307)$

B) $P(X < 20.483)$

C) $\chi^2_{0.01(V=10)}$

F-Distribution

- The F-distribution has 2 parameters: (i) the numerator degrees of freedom, v_1; and, (ii) the denominator degrees of freedom, v_2.

- This distribution is directly related to the chi-square distribution. If there are two independent random variables, X and Y, such that $X \sim \chi^2(v_1)$ and $Y \sim \chi^2(v_2)$ then the random variable, , will follow an F-distribution with v_1 and v_2 degrees of freedom.
-
- This is why we call v_1 and v_2 the numerator and denominator degrees of freedom, respectively. These were the original degrees of freedom of the χ^2 random variables in the F-ratio.

F-Distribution

- Just like the chi-square distribution, the pdf of the F-distribution is positive for positive real numbers only; elsewhere, its value is 0.

- The graph of the pdf is also skewed to the right. In general, distributions with higher degrees of freedom are less skewed.

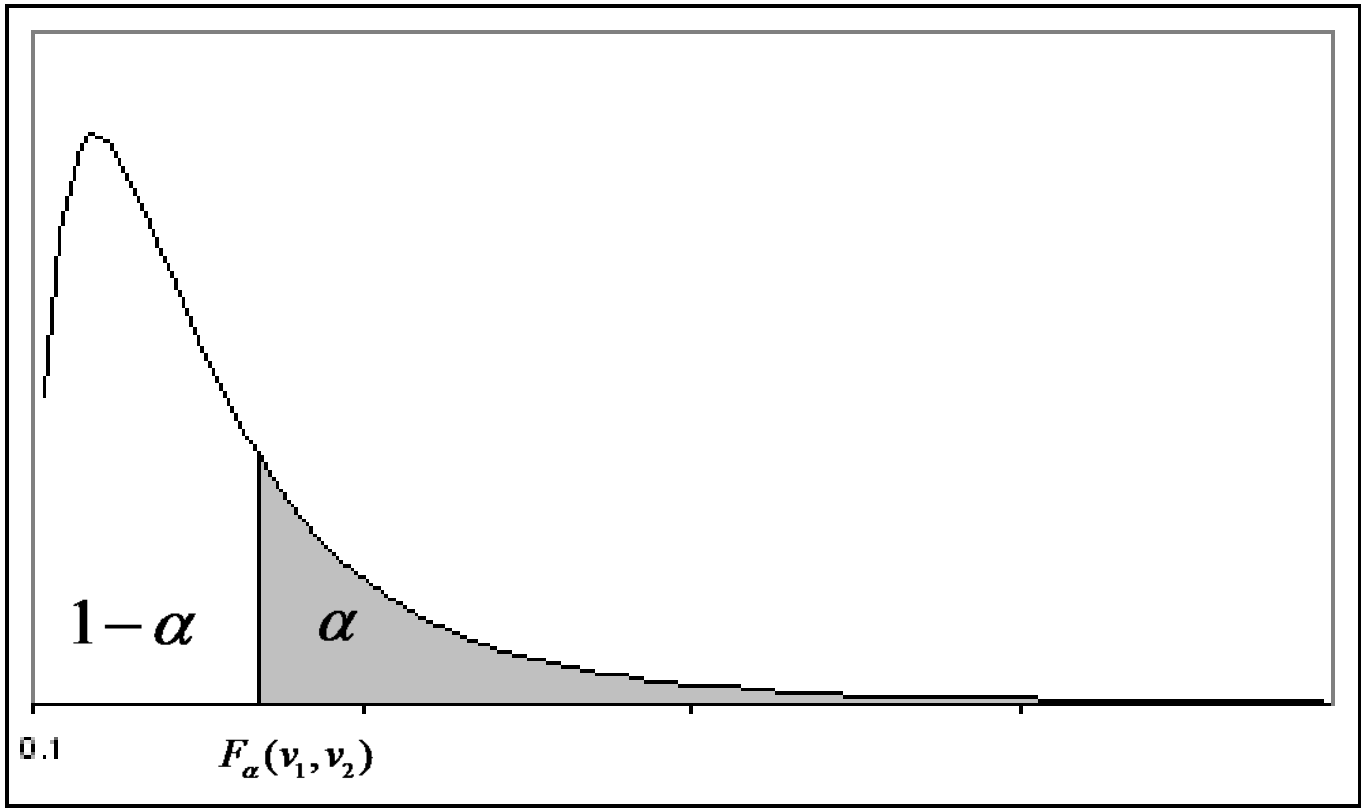

FIGURE 11.5. GRAPHICAL REPRESENTATION OF THE $100(1-\alpha)^{TH}$ PERCENTILE, F_α

F-Distribution

- Table B.4, Appendix B, presents the $100(1-\alpha)^{th}$ percentile, of an F-distribution with integral-valued degrees of freedom ranging from 1 to 30 for selected values of α.

- We will denote the $100(1-\alpha)^{th}$ percentile of an F-distribution with ν_1 and ν_2 degrees of freedom by $F_\alpha(\nu_1,\nu_2)$. If $X \sim F(\nu_1,\nu_2)$ then $F_\alpha(\nu_1,\nu_2)$ satisfies the condition that the $P(X \leq F_\alpha(\nu_1,\nu_2))=1-\alpha$, or equivalently, $P(X>F_\alpha(\nu_1,\nu_2))=\alpha$. Figure 11.5 shows the graphical representation of $F_\alpha(\nu_1,\nu_2)$.

11.3 Sampling from the Normal Distribution

- Suppose $(X_1, X_2, \ldots, X_n)$ is a random sample satisfying the condition that $X_i \sim \text{Normal}(\mu, \sigma^2)$ for i=1,2,…,n. Define the statistics, $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ as the sample mean and $S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$ as the sample variance, where n is the sample size. The sampling distribution of $\overline{X}$ is also normal with mean μ and variance σ^2.

TABLE 11.1. SAMPLING DISTRIBUTIONS OF STATISTICS BASED ON A RANDOM SAMPLE FROM A NORMAL DISTRIBUTION

STATISTIC	*SAMPLING DISTRIBUTION*	**PARAMETER/S**
$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$	STANDARD NORMAL DISTRIBUTION	MEAN = 0 VARIANCE = 1
$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$	T-DISTRIBUTION	DEGREES OF FREEDOM: V = N – 1
$X^2 = \frac{(n-1)S^2}{\sigma^2}$	CHI-SQUARE DISTRIBUTION	DEGREES OF FREEDOM: V = N – 1

Sampling Distributions of Statistics Based on Two Independent Random Samples from Normal Distributions

- THERE ARE INSTANCES WHEN WE WISH TO COMPARE THE BEHAVIOR OF TWO DIFFERENT POPULATIONS.
- SUPPOSE $(X_1,X_2,\ldots,X_{n1})$ is a random sample of size n_1 satisfying the condition that $X_i \sim \text{Normal}(\mu_X, \sigma_X^2)$ for $i=1,2,\ldots,n_1$.
- Suppose we take another random sample of size n_2, $(Y_1,Y_2,\ldots,Y_{n2})$, whose selection is independent of the selection of the sample from the first population, and this time satisfying the condition that $Y_i \sim \text{Normal}(\mu_Y, \sigma_Y^2)$ for $i=1,2,\ldots,n_2$.

Sampling Distributions of Statistics Based on Two Independent Random Samples from Normal Distributions

- Define the statistics, $\bar{X} = \frac{\sum_{i=1}^{n_1} X_i}{n_1}$ as the sample mean, $S_X^2 = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2}{n_1 - 1}$ as the sample variance, where n_1 is the sample size of the sample taken from the first normal population.
- Likewise, define $\bar{Y} = \frac{\sum_{i=1}^{n_2} Y_i}{n_2}$ as the sample mean, as the $S_Y^2 = \frac{\sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_2 - 1}$ sample variance, where n_2 is the sample size of the sample taken from the second normal population.

TABLE 11.2. SAMPLING DISTRIBUTION OF STATISTICS BASED ON TWO INDEPENDENT RANDOM SAMPLES FROM NORMAL DISTRIBUTIONS

STATISTIC	*SAMPLING DISTRIBUTION*	**PARAMETER/S**
$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}$	STANDARD NORMAL DISTRIBUTION	MEAN = 0 VARIANCE = 1
$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $S_p = \sqrt{\frac{(n_1 - 1)S_x^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}}$	T-DISTRIBUTION IF $\sigma_X^2 = \sigma_Y^2$	DEGREES OF FREEDOM: $V = N_1 + N_2 - 2$
$F = \frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2}$	F-DISTRIBUTION	DEGREES OF FREEDOM: $V_1 = N_1 - 1$ $V_2 = N_2 - 1$