Plaksha University: Technology Leaders Program

Dr. Nandini Kannan

email : nandini.kannan@plaksha.edu.in

Chapter 3

Sampling Distributions

3.1 Introduction

Statistics as the science that deals with

(a) the collection, organization, and summary of information about a particular topic of interest (*Descriptive Statistics*)

(b) drawing inferences about the population of interest using information obtained from a sample (*Inferential Statistics*)

Definition: A **Parameter** is a numerical measure associated with a population.

Definition: A **Statistic** is a numerical measure associated with the sample.

Example Paper Boat would like to determine the sugar content of Nagpur oranges. How would you help Paper Boat answer this question?

A consumer group wants to determine the fuel efficiency of the new Honda SUV. How would you proceed.

In both cases, identify the parameter and the statistics of interest.

Therefore, Statistics are **Random Variables**.

We would like to know how the statistic changes over different samples.

Definition: The probability distribution of a statistic is called a **sampling distribution**.

Example: For the fuel efficiency example, the sampling distribution of the mean could be attempted as follows:

- Draw a sample of 100 from the population of all vehicles manufactured in a given time period. Compute the sample mean.
- Repeat the process of drawing samples of size 100 several times.
- Each time a sample is selected, the value of the statistic (mean) is calculated.
- Draw the relative frequency histogram of these computed statistics.

If the process is repeated a large number of times, the histogram will provide an approximation of the sampling distribution.

Let $X_1, \ldots, X_n$ be n mutually independent observations made of a particular quantitative phenomenon-for example, blood pressure. We assume that each observation X_i has the same probability distribution. We say that the n observations $X_1, \ldots, X_n$ are **independent and identically distributed (i.i.d.)**.

If X is described by a pmf, then

$$p(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n) = p(x_1) \ldots p(x_n).$$

If X is described by a pdf, then

$$f(x_1, \ldots, x_n) = f(x_1) \ldots f(x_n).$$

$X_1, \ldots, X_n$ are said to be a **random sample of size n from the distribution of X**.

3.2 Sampling Distribution of the Mean

Suppose we are interested in estimating the mean μ_X of a random variable X based on a random sample $X_1, \ldots, X_n$. We can estimate μ_X by the sample average

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

Using properties of expectation, we have

$$\mu_{\bar{X}} = E(\bar{X}) = \mu_X$$

and

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = \frac{\sigma^2_X}{n}.$$

Example: Consider a random sample of $n = 12$ from $U(-1/2, 1/2)$. Let $T = \sum_{i=1}^{n} X_i$. Figure 2.1 shows a histogram of 1000 such sums with a superimposed normal pdf.

Figure 2.2 shows a histogram of the distribution of $X_{(n)}$, the largest order statistic.

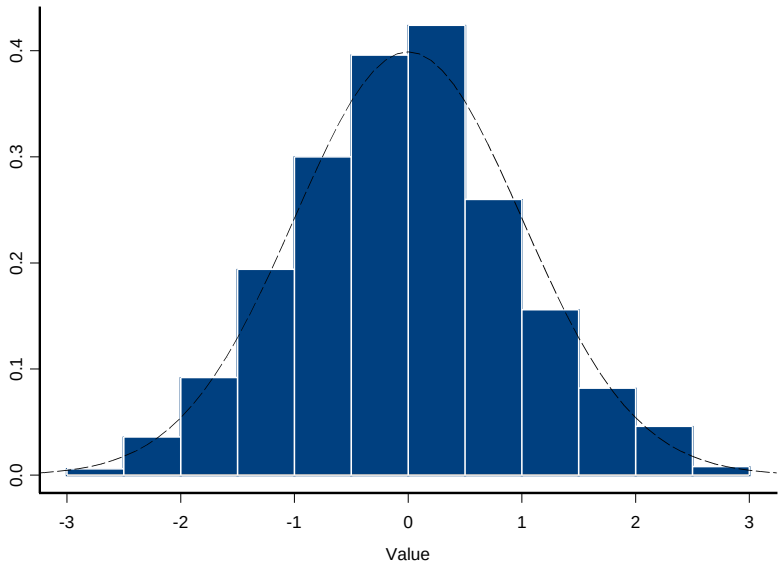

Figure 3.1: Probability Histogram: Sum of 12 Uniform Random Variables

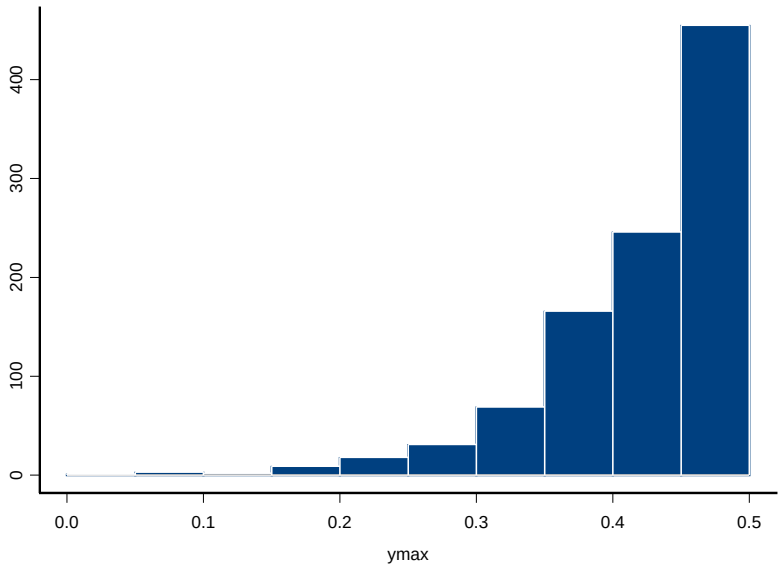

Figure 3.2: Probability Histogram: Max of 12 Uniform Random Variables

3.3 Central Limit Theorem

Example: Suppose X is a discrete random variable with probability distribution given by

x	0	1
p(x)	$\frac{1}{3}$	$\frac{2}{3}$

i.e. the population consists of 0's and 1's, 1/3-rd of the population consists of 0's, 2/3-rd of the population consists of 1's.

We can compute the mean and variance. We have

$$E(X) = \mu = \frac{2}{3} \qquad Var(X) = \sigma^2 = \frac{2}{9}$$

Let X_1, X_2 be a random sample of size 2 drawn from this population. Both X_1 and X_2 can take values 0 and 1. There are four **possible** samples of size 2:

$$(0,0), \qquad (0,1), \qquad (1,0), \qquad (1,1).$$

We compute the average $\overline{X}_2$ and total T_2 for all possible samples. The table is shown below.

X_1	X_2	$\overline{X}_2$	T_2
0	0	0	0
0	1	0.5	1
1	0	0.5	1
1	1	1	2

Thus the average and total are both random variables, and we can compute their probability distributions.

$\overline{X}_2$ can assume values 0, 0.5, and 1.

$$\begin{aligned} P(\overline{X}_2 = 0) &= P(X_1 = 0 \text{ and } X_2 = 0) \\ &= P(X_1 = 0)\ P(X_2 = 0) \qquad \text{(by independence)} \\ &= \left(\frac{1}{3}\right)\left(\frac{1}{3}\right) \\ &= \frac{1}{9} \end{aligned}$$

$$P(\overline{X}_2 = 0.5) = P(X_1 = 0 \text{ and } X_2 = 1) + P(X_1 = 1 \text{ and } X_2 = 0)$$

$$
\begin{aligned}
&= P(X_1 = 0)\ P(X_2 = 1) + P(X_1 = 1)\ P(X_2 = 0) \\
&= \left(\frac{1}{3}\right)\left(\frac{2}{3}\right) + \left(\frac{2}{3}\right)\left(\frac{2}{3}\right) \\
&= \frac{4}{9}
\end{aligned}
$$

$$
\begin{aligned}
P(\overline{X}_2 = 1) &= P(X_1 = 1 \text{ and } X_2 = 1) \\
&= P(X_1 = 1)\ P(X_2 = 1) \\
&= \left(\frac{2}{3}\right)\left(\frac{2}{3}\right) \\
&= \frac{4}{9}
\end{aligned}
$$

Similarly the total T_2 is a random variable taking values 0, 1, and 2. We can compute the probabilities in exactly the same way. The probability distributions can be summarized as follows:

$\overline{X}_2$	0.0	0.5	1.0
prob	$\frac{1}{9}$	$\frac{4}{9}$	$\frac{4}{9}$

We compute the mean and variance to be

$$E(\overline{X}_2) = \frac{2}{3} = \mu$$

$$Var(\overline{X}_2) \;=\; \frac{1}{9} = \frac{\sigma^2}{2}$$

The probability distribution of T_2 is given below.

T_2	0.0	1.0	2.0
prob	$\frac{1}{9}$	$\frac{4}{9}$	$\frac{4}{9}$

We compute the mean and variance to be

$$\begin{aligned} E(T_2) &= \frac{4}{3} = 2\,\mu \\ Var(\overline{X}_2) &= \frac{4}{9} = 2\,\sigma^2 \end{aligned}$$

We can repeat this process by drawing a sample of size 3 from the population. We can compute the probability distributions for $\overline{X}_3$ and T_3. The probability distributions are given below.

$\overline{X}_3$	0.0	$\frac{1}{3}$	$\frac{2}{3}$	1.0
prob	$\frac{1}{27}$	$\frac{6}{27}$	$\frac{12}{27}$	$\frac{8}{27}$

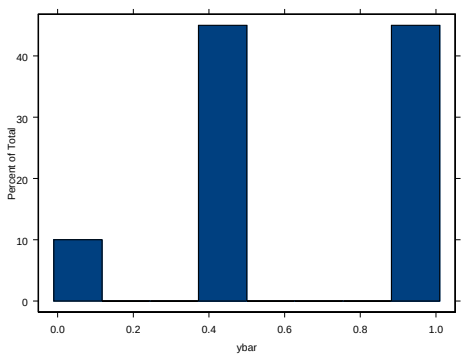

Figure 3.3: Probability Histogram for Average: n=2

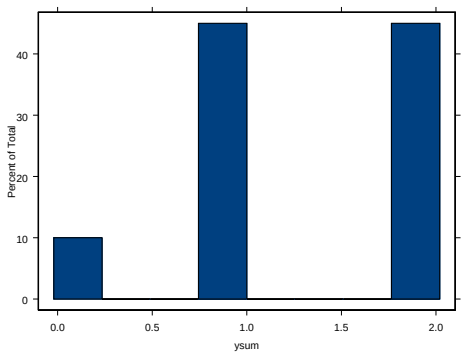

Figure 3.4: Probability Histogram for Sum: n=2

We compute the mean and variance to be

$$\begin{aligned} E(\overline{X}_3) &= \frac{2}{3} = \mu \\ Var(\overline{X}_3) &= \frac{2}{9} = \frac{\sigma^2}{3} \end{aligned}$$

The probability distribution of T_3 is given below.

T_3	0.0	1.0	2.0	3.0
prob	$\frac{1}{27}$	$\frac{6}{27}$	$\frac{12}{27}$	$\frac{8}{27}$

We compute the mean and variance to be

$$\begin{aligned} E(T_3) &= 2 = 3\ \mu \\ Var(\overline{X}_3) &= \frac{2}{3} = 3\ \sigma^2 \end{aligned}$$

Even with a sample of size 3, the histograms are beginning to look symmetric and more like a normal distribution. If we continue this process, we will observe that the probability histograms start resembling the Gaussian Distribution.

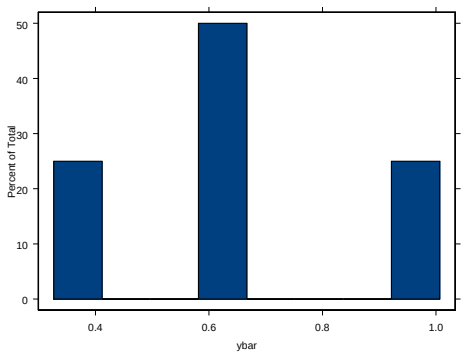

Figure 3.5: Probability Histogram for Average: n=3

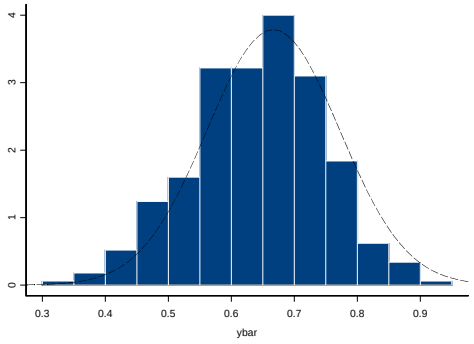

Figure 3.6: Probability Histogram for Average: n=20

Theorem 3.3.1. *Central Limit Theorem: Let $X_1, \ldots, X_n$ be a sequence of iid random variables drawn from a population with finite mean μ, and variance σ^2. Then for large n, the sampling distribution of the sample mean is approximately normal with mean*

$$E(\overline{X}) = \mu$$

and variance

$$Var(\overline{X}) = \frac{\sigma^2}{n}.$$

A similar statement can be written for the total.

Remark. If the population is known to be normal, then the distribution of the sample mean is exactly normal for any sample size n.

Remark. By large n, we usually mean a sample of at least 25 measurements.

Example: A manufacturer of automobile batteries claims that the distribution of the lifetimes of its best battery has an average of 54 months, and a standard deviation of 6 months. Suppose a consumer group decides to check the claim by purchasing a sample of 50 of these batteries and testing them.

(a) Describe the sampling distribution of the average lifetime of a sample of 50 batteries.

(b) What is the probability that the sample has an average life of 52 months or fewer?

Solution Since the sample size is greater than 25, the sampling distribution of the average based on a sample of size 50 is **approximately normally** distributed with mean

$$E(\overline{X}_{50}) = 54 \text{ months}$$

and variance

$$Var(\overline{X}_{50}) = \frac{36}{50}.$$

(b) We want to find

$$P(\overline{X}_{50} \leq 52)$$

We know the distribution is approximately normal, so we can standardize the above using the mean and variance computed in (a).

$$\begin{aligned} P(\overline{X}_{50} \leq 52) &= P((\overline{X}_{50} - 54) \leq (52 - 54)) \\ &= P\left(\frac{(\overline{X}_{50} - 54)}{0.85} \leq \frac{(52 - 54)}{0.85}\right) \\ &= P(Z \leq -2.35) \\ &= 0.0094 \end{aligned}$$

The probability the consumer group will observe a sample average of 52 or less is 0.0094 if the manufacturer's claim is true. If the 50 tested batteries do result in an average of 52 or fewer months, the consumer group will have strong evidence that the manufacturer's claim is untrue. Such an event is very unlikely to happen if the claim is true.

3.4 Normal Approximation to the Binomial

Let $Y \sim Bin(n, p)$. Y is the number of successes that are observed in the n trials and p represents the probability of success in any trial. $Y = X_1 + \ldots + X_n$, where $X_i's$ are iid Bernoulli random variables.

An estimate of p, denoted by $\hat{p}$, is the proportion of successes that are observed in the n trials, i.e.,

$$\hat{p} = \frac{Y}{n}.$$

In order to use the normal approximation to the binomial, we require

$$n\ p \geq 5; n(1-p) \geq 5.$$

Theorem 3.4.1. *The sampling distribution of $\hat{p}$ is approximately normal with mean*

$$p$$

and variance

$$\frac{p\ (1-p)}{n}.$$

Example: An airline has determined that the no-show rate for reservations is 10 %. Suppose the next flight has 100 parties with advance reservations.

a. Find the probability that the number of no-shows is between 20 and 25.

b. Approximate the probability in (a). Justify the approximation.

3.5 Distributions Derived From the Normal

Linear combinations of normally distributed random variables are normal.

Theorem 3.5.1. *Let $X_1, \ldots, X_n$ be independent $N(\mu_i, \sigma_i^2)$ random variables. Let $Y = \sum_{i=1}^n a_i X_i$ be a linear combination of the $X_i's$ with $a_1, \ldots, a_n$ constants. Then $Y \sim N(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2)$.*

3.5.1 The χ^2 distribution

Theorem 3.5.2. *If $Z \sim N(0,1)$, then $Z^2 \sim \chi_1^2$.*

Proof: The pdf of Z is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad -\infty < z < \infty.$$

Let $Y = Z^2$. This defines a transformation between the value of Z and Y that is not one-to-one.

The inverse solutions of $y = z^2$ are $z = \pm\sqrt{y}$. Let $z_1 = -\sqrt{y}$ and $z_2 = \sqrt{y}$.

The Jacobian of the transformation is given by

$$J_1 = \frac{d}{dy}(-\sqrt{y}) = \frac{-1}{2\sqrt{y}}; \qquad J_2 = \frac{d}{dy}(\sqrt{y}) = \frac{1}{2\sqrt{y}}.$$

The pdf of Y is given by

$$g(y) = \frac{1}{\sqrt{2\pi}} e^{-y/2} \left| \frac{-1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{-y/2} \left| \frac{1}{2\sqrt{y}} \right| = \frac{1}{\sqrt{2\pi}}\, y^{1/2-1} e^{-y/2}$$

for $y > 0$.

Since $\int_0^\infty g(y)dy = 1$, we have

$$1 = \frac{1}{\sqrt{2\pi}} \int_0^\infty y^{1/2-1} e^{-y/2} dy = \frac{\Gamma(1/2)}{\sqrt{\pi}},$$

since the function within the integral sign resembles a Gamma random variable with $\alpha = 1/2$ and $\beta = 2$.

Recall the pdf of a $Gamma(\alpha, \beta)$ rv is

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)}\, x^{\alpha-1} e^{-x/\beta}.$$

Therefore $\Gamma(1/2) = \sqrt{\pi}$ and the pdf of Y is given by

$$\begin{cases} \frac{1}{\sqrt{2}\Gamma(1/2)} y^{1/2-1} e^{-y/2}, & y > 0; \\ 0, & \text{otherwise.} \end{cases}$$

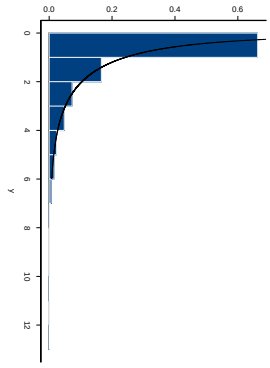

Figure 3.7: Sampling Distribution of Z^2 with superimposed Chi-squared distribution

This is the pdf of a chi-squared random variable with 1 degree of freedom. ■

We also showed that if $X \sim N(\mu, \sigma^2)$, then $(X-\mu)/\sigma \sim N(0,1)$. Therefore $[(X-\mu)/\sigma]^2 \sim \chi^2_1$.

Theorem 3.5.3. *If $U_1, \ldots, U_n$ are iid chi-squared random variables with 1 degree of freedom, then $V = U_1 + \ldots + U_n \sim \chi^2_n$.*

This is the reproductive property of the chi-squared distribution.

Sampling Distribution of S^2

Let $X_1, \ldots, X_n$ be a random sample drawn from a **normal** pop-

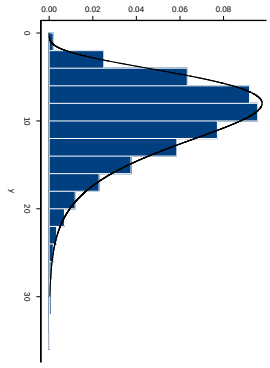

Figure 3.8: Sampling Distribution of Sum of 10 chi-squared Random variables

ulation with mean μ and variance σ^2. The sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is a random variable. We have

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2 \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(\bar{X}-\mu)^2 + 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X}) \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2.
\end{aligned}$$

Dividing each term by σ^2 and substituting $(n-1)S^2$ for $\sum_{i=1}^{n}(X_i -$

$\bar{X})^2$, we have

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X}-\mu)^2}{\sigma^2/n}$$

We know that

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

is a chi-squared random variable with n degrees of freedom. We also know that

$$\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)$$

which implies

$$\frac{(\bar{X}-\mu)^2}{\sigma^2/n}$$

is a chi-squared random variable with 1 degree of freedom. We then have the following result.

Theorem 3.5.4. *Consider a random sample of size n from a normal population with mean μ and variance σ^2. Then*

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

is a chi-squared random variable with $n-1$ degrees of freedom

3.5.2 Student's $t-$ distribution

The Central Limit Theorem allows us to determine the sampling distribution of $\bar{X}$ when σ is known. In many practical applications, the population variance is unknown and must be estimated from the data. This introduces additional variability and produces a distribution that deviates significantly from a standard normal.

Theorem 3.5.5. *Let $Z \sim N(0,1)$ and $V \sim \chi^2(\nu)$. If Z and V are independent, then*

$$T = \frac{Z}{\sqrt{V/\nu}}$$

has a $t-$ distribution with ν degrees of freedom.

Corollary 3.5.6. *Let $X_1, \ldots, X_n$ be a random sample from a* ***normal*** *population with mean μ and variance σ^2. Let*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum\nolimits_{i=1}^{n}(X_i - \bar{X})^2.$$

Then the random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a $t-$ distribution with $\nu = n-1$ degrees of freedom.

- The $t-$distribution is symmetric about 0.
- It is bell-shaped.
- The $t-$distribution is more variable than the Z (standard normal).
- The distribution is characterized by a single parameter ν called the degrees of freedom (df).
- As the df increases, the $t-$distribution gets closer and closer to the normal curve.
- Assumes underlying population is normal: however, if the underlying population is *not* normal but is *"nearly"* bell shaped the distribution of T will be approximately a t.
- Tables of the percentage points for the $t-$ distribution are available for different degrees of freedom.

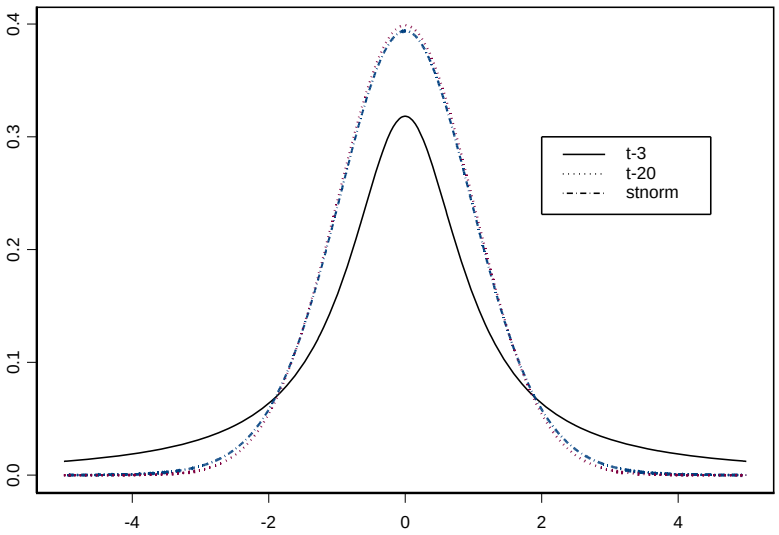

Figure 3.9: The t densities on 1 and 20 df, and the standard normal

3.5.3 The $F-$ distribution

Theorem 3.5.7. *Let $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$ be independent random variables. Then*

$$F = \frac{U/\nu_1}{V/\nu_2}$$

has the $F-$distribution with ν_1 and ν_2 degrees of freedom.

The $F-$distribution is **not** symmetric.

Theorem 3.5.8. *Let $f_\alpha(\nu_1, \nu_2)$ be the value from the $F-$table that cuts off an area of α in the upper tail for ν_1 and ν_2 degrees of freedom. Then*

$$f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_2, \nu_1)}.$$

Theorem 3.5.9. *Let S_1^2 and S_2^2 be the variances corresponding to two independent random samples of size n_1 and n_2 from normal populations with variances σ_1^2 and σ_2^2, respectively. Then*

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an $F-$ *distribution with* $\nu_1 = n_1 - 1$ *and* $\nu_2 = n_2 - 1$ *degrees of freedom.*