MSSC 6010/Comp. Probability

Instructor: Mehdi Maadooliat

Chapter 6

Department of Mathematics, Statistics and Computer Science

Special thanks to Prof. Ana Militino for providing the original slides of the book.

Chapter 6

Sampling and Sampling Distributions

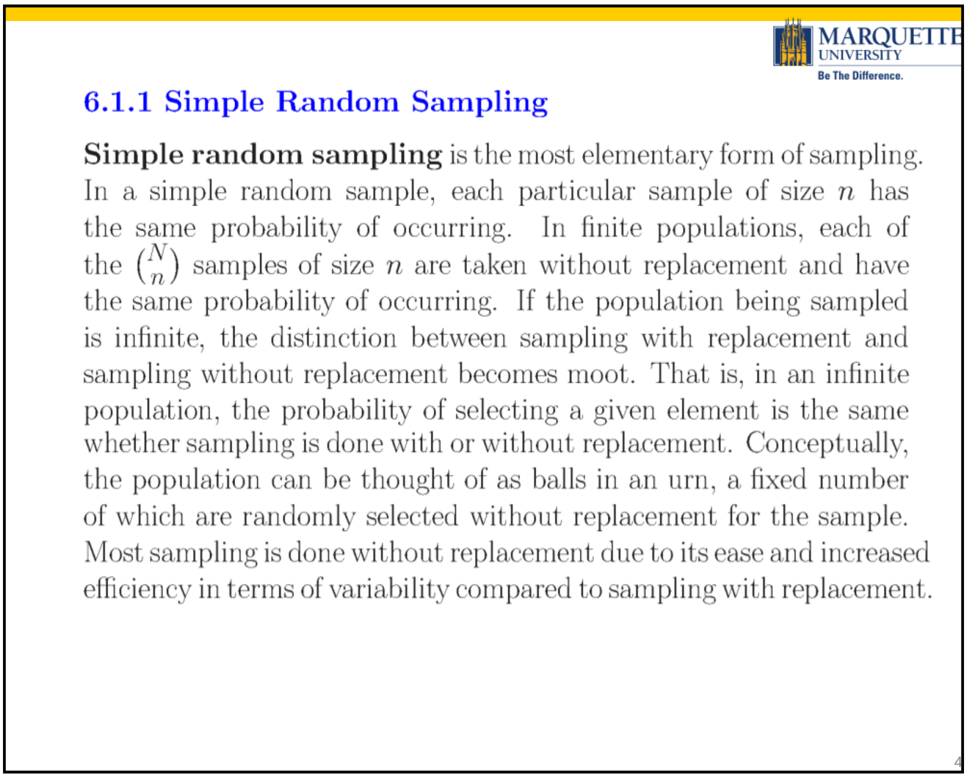
6.1 Sampling

The objective of statistical analysis is to gain knowledge about certain properties in a population that are of interest to the researcher. When the population is small, the best way to study the population of interest is to study all of the elements in the population one by one. This process of collecting information on the entire population of interest is called a **census**. However, it is usually quite challenging to collect information on an entire population of interest. Not only do monetary and time constraints prevent a census from being taken easily, but also the challenges of finding all the members of a population can make gathering an accurate census all but impossible. Under certain conditions, random selection of certain elements actually returns more reliable information than can be obtained by using a census. Standard methods used to learn about the characteristics of a population of interest include simulation, designed experiments, and sampling.

Simulation studies typically generate numbers according to a researcher specified model. For a simulation study to be successful, the chosen simulation model must closely follow the real life process the researcher is attempting to simulate. For example, the effects of natural disasters, such as earthquakes, on buildings and highways are often modeled with simulation.

Sampling is the most frequently used form of collecting information about a population of interest. Many forms of sampling exist, such as random sampling, simple random sampling, systematic sampling, and cluster sampling. It will be assumed that the population from which one is sampling has size N and that the sample is of size $n < N$.

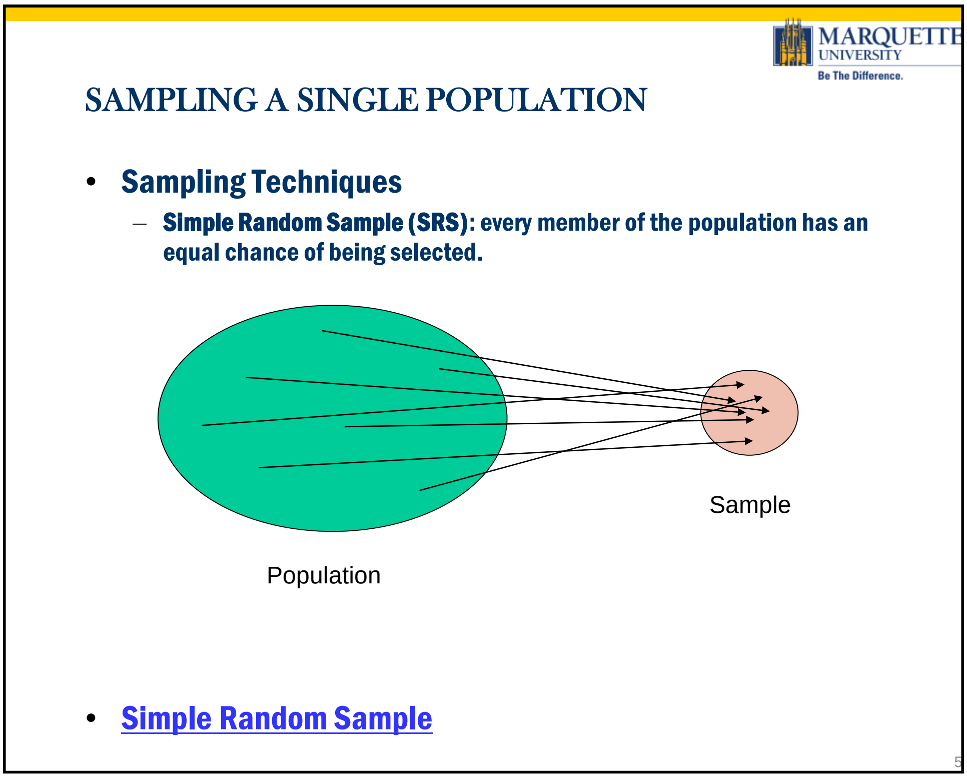
Random sampling is the process of selecting n elements from a population where each of the n elements has the same probability of being selected, namely $\frac{1}{N}$. More precisely, the random variables $X_1, X_2, \ldots, X_n$ form a random sample of size n from a population with a **pdf** $f(x)$ if $X_1, X_2, \ldots, X_n$ are mutually independent random variables such that the marginal **pdf** of each X_i is $f(x)$. The statement "$X_1, X_2, \ldots, X_n$ are independent and identically distributed, (i.i.d.), random variables with **pdf** $f(x)$" is often used to denote a random sample. The objective of random sampling is to obtain a representative sample of the population that can be used to make generalizations about the population.



6.1.1 Simple Random Sampling

Simple random sampling is the most elementary form of sampling. In a simple random sample, each particular sample of size n has the same probability of occurring. In finite populations, each of the $\binom{N}{n}$ samples of size n are taken without replacement and have the same probability of occurring. If the population being sampled is infinite, the distinction between sampling with replacement and sampling without replacement becomes moot. That is, in an infinite population, the probability of selecting a given element is the same whether sampling is done with or without replacement. Conceptually, the population can be thought of as balls in an urn, a fixed number of which are randomly selected without replacement for the sample. Most sampling is done without replacement due to its ease and increased efficiency in terms of variability compared to sampling with replacement.

4



SAMPLING A SINGLE POPULATION

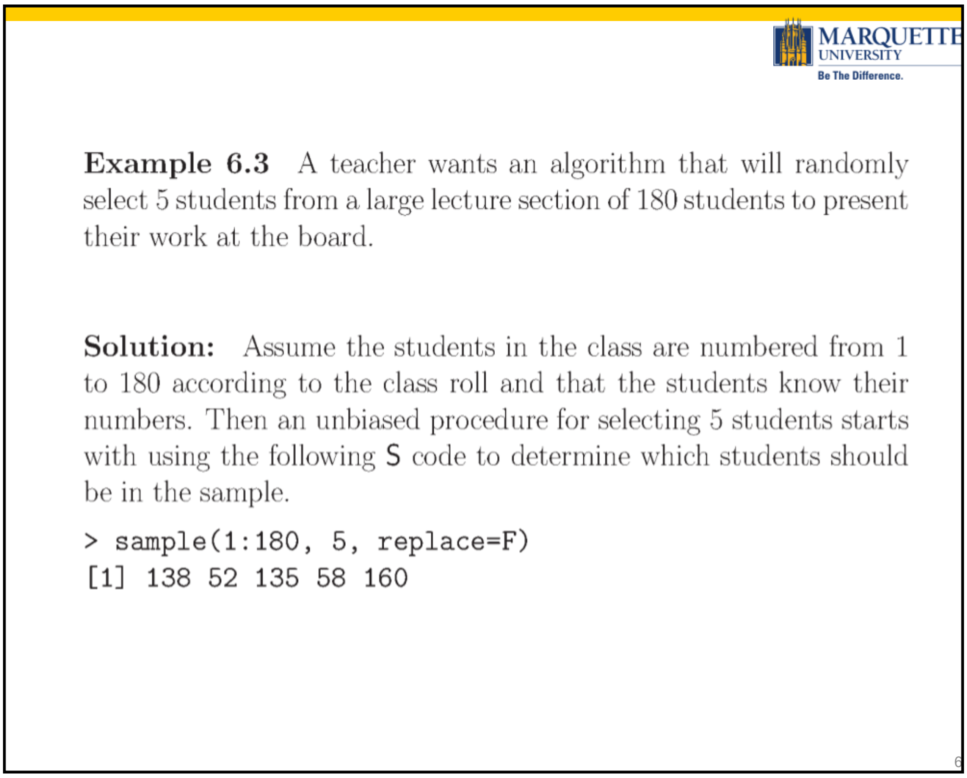
- **Sampling Techniques**
  - **Simple Random Sample (SRS): every member of the population has an equal chance of being selected.**

Sample

Population

- **Simple Random Sample**

5

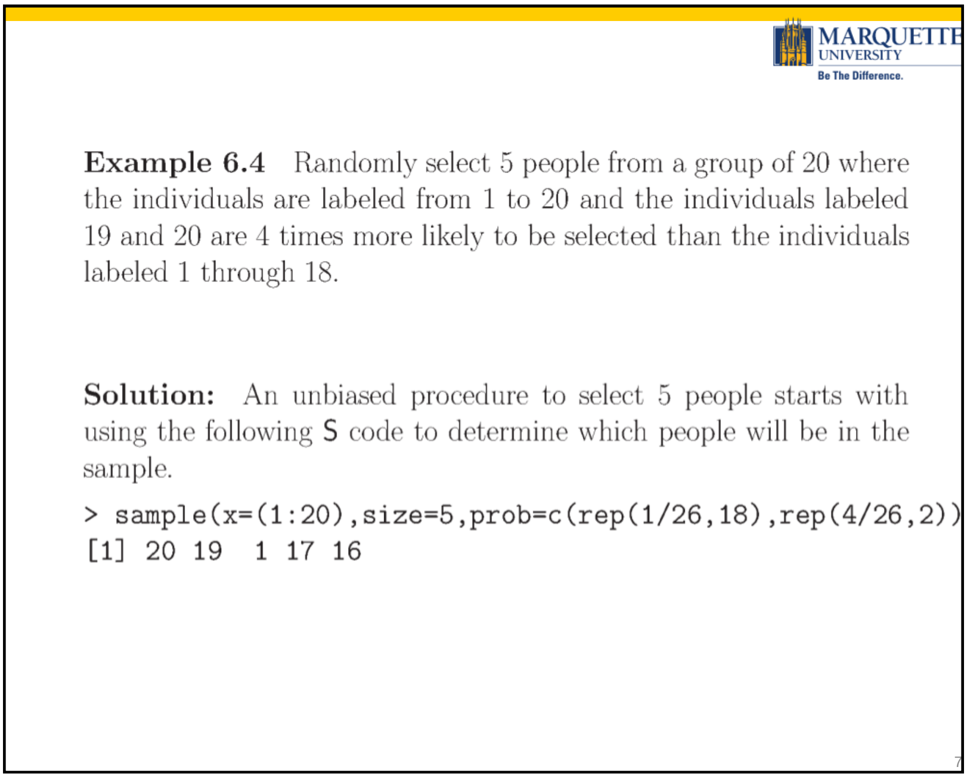


Example 6.3 A teacher wants an algorithm that will randomly select 5 students from a large lecture section of 180 students to present their work at the board.

Solution: Assume the students in the class are numbered from 1 to 180 according to the class roll and that the students know their numbers. Then an unbiased procedure for selecting 5 students starts with using the following S code to determine which students should be in the sample.

```
> sample(1:180, 5, replace=F)
[1] 138 52 135 58 160
```

6

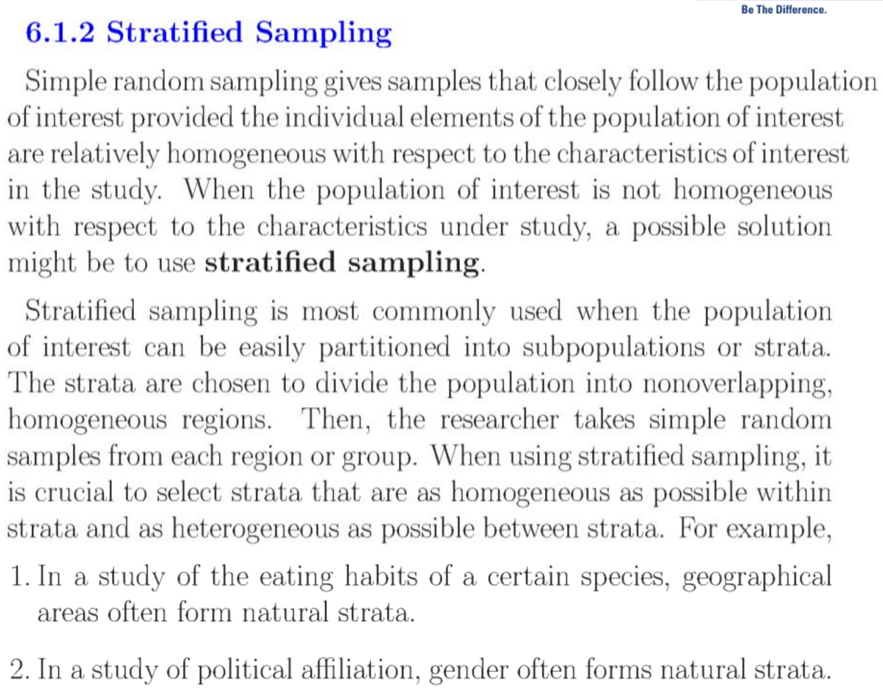


Example 6.4 Randomly select 5 people from a group of 20 where the individuals are labeled from 1 to 20 and the individuals labeled 19 and 20 are 4 times more likely to be selected than the individuals labeled 1 through 18.

Solution: An unbiased procedure to select 5 people starts with using the following S code to determine which people will be in the sample.

```
> sample(x=(1:20),size=5,prob=c(rep(1/26,18),rep(4/26,2))
[1] 20 19  1 17 16
```

7



6.1.2 Stratified Sampling

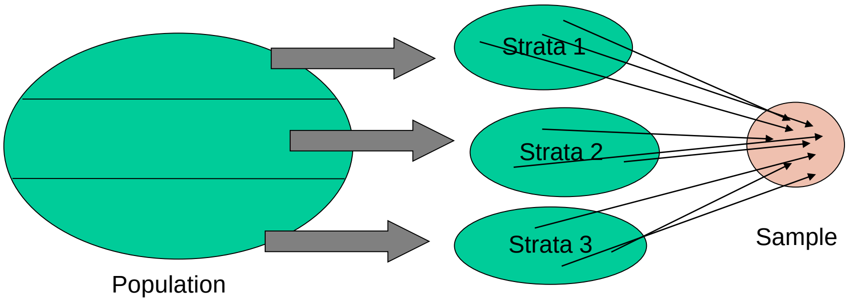
Simple random sampling gives samples that closely follow the population of interest provided the individual elements of the population of interest are relatively homogeneous with respect to the characteristics of interest in the study. When the population of interest is not homogeneous with respect to the characteristics under study, a possible solution might be to use **stratified sampling**.

Stratified sampling is most commonly used when the population of interest can be easily partitioned into subpopulations or strata. The strata are chosen to divide the population into nonoverlapping, homogeneous regions. Then, the researcher takes simple random samples from each region or group. When using stratified sampling, it is crucial to select strata that are as homogeneous as possible within strata and as heterogeneous as possible between strata. For example,

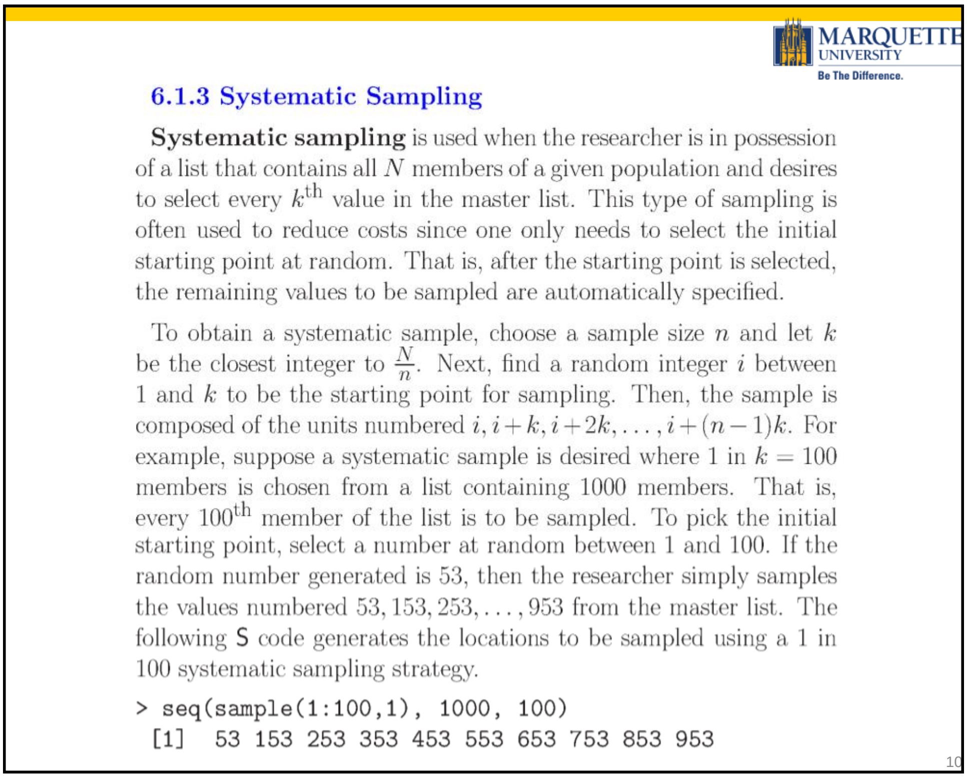
1. In a study of the eating habits of a certain species, geographical areas often form natural strata.
2. In a study of political affiliation, gender often forms natural strata.

SAMPLING A SINGLE POPULATION

- **Sampling Techniques**
  - **Stratified Random Sample**: Divide the sample into several strata. Then take a SRS from each stratum.



- Advantage: Each stratum is guaranteed to be randomly sampled
- Example: Obtain a list of all SSN for individuals in the U.S. who are over 65. Divide up the SSNs into region of the country (time zones). Then randomly sample 30 from each time zone.

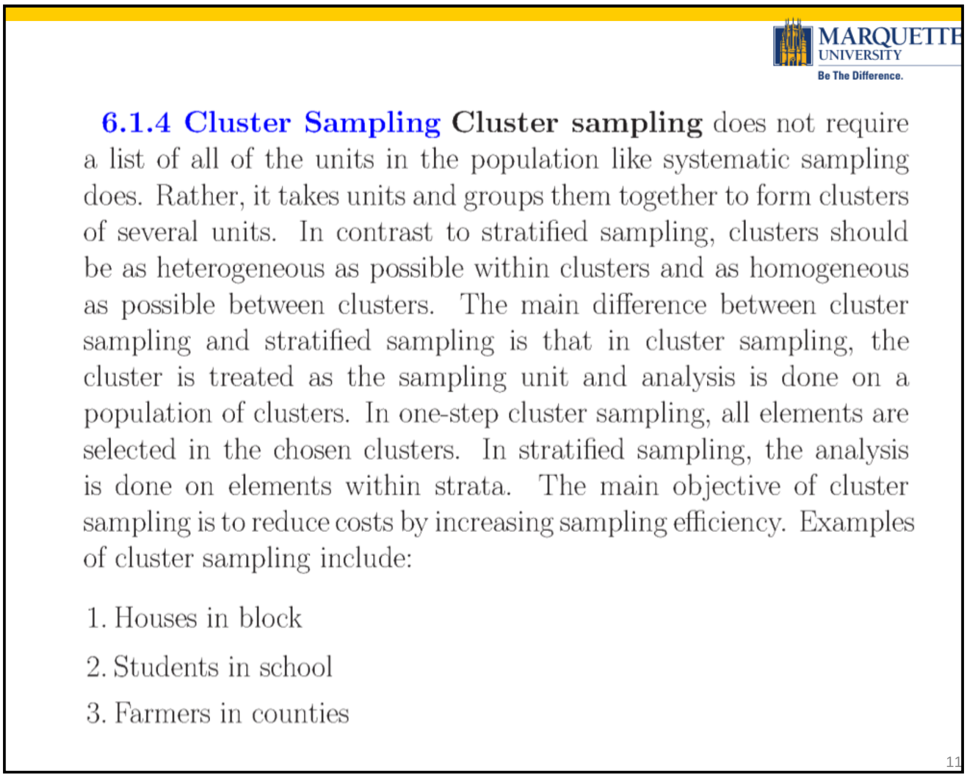


6.1.3 Systematic Sampling

Systematic sampling is used when the researcher is in possession of a list that contains all N members of a given population and desires to select every k^{th} value in the master list. This type of sampling is often used to reduce costs since one only needs to select the initial starting point at random. That is, after the starting point is selected, the remaining values to be sampled are automatically specified.

To obtain a systematic sample, choose a sample size n and let k be the closest integer to $\frac{N}{n}$. Next, find a random integer i between 1 and k to be the starting point for sampling. Then, the sample is composed of the units numbered $i, i+k, i+2k, \ldots, i+(n-1)k$. For example, suppose a systematic sample is desired where 1 in $k = 100$ members is chosen from a list containing 1000 members. That is, every 100^{th} member of the list is to be sampled. To pick the initial starting point, select a number at random between 1 and 100. If the random number generated is 53, then the researcher simply samples the values numbered $53, 153, 253, \ldots, 953$ from the master list. The following S code generates the locations to be sampled using a 1 in 100 systematic sampling strategy.

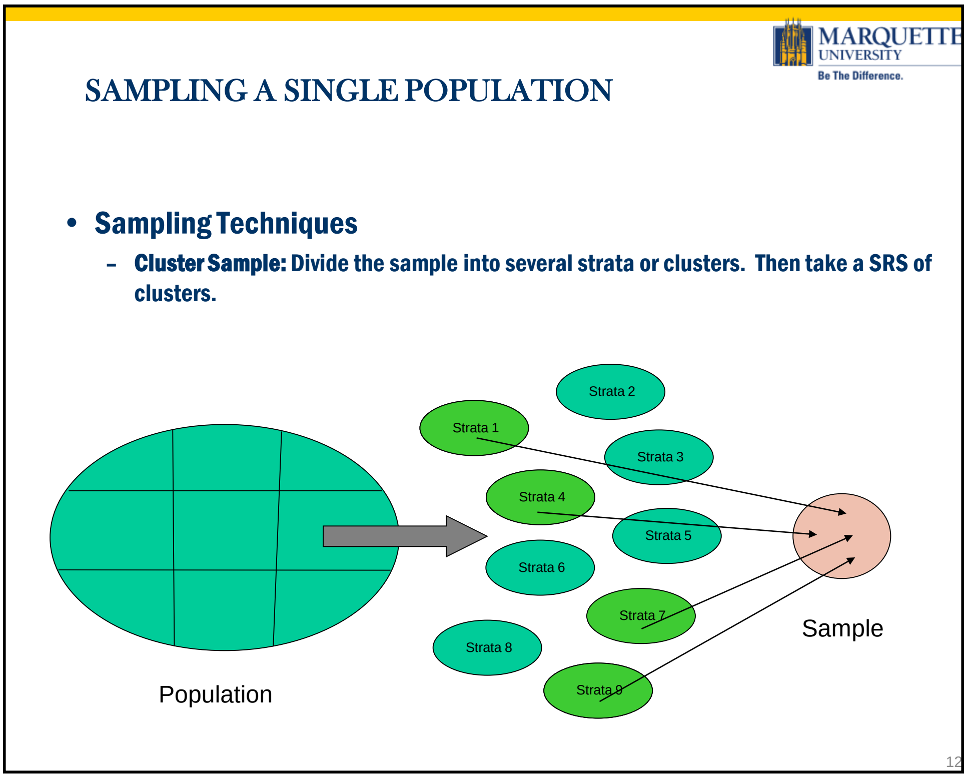
```
> seq(sample(1:100,1), 1000, 100)
 [1]  53 153 253 353 453 553 653 753 853 953
```



6.1.4 Cluster Sampling

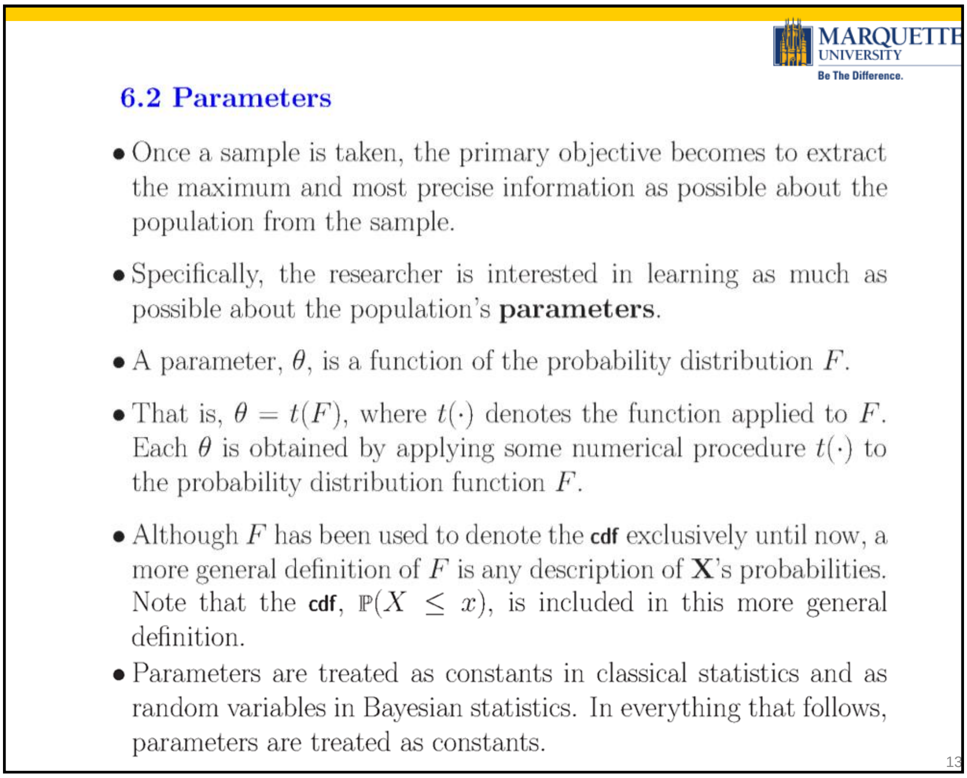
Cluster sampling does not require a list of all of the units in the population like systematic sampling does. Rather, it takes units and groups them together to form clusters of several units. In contrast to stratified sampling, clusters should be as heterogeneous as possible within clusters and as homogeneous as possible between clusters. The main difference between cluster sampling and stratified sampling is that in cluster sampling, the cluster is treated as the sampling unit and analysis is done on a population of clusters. In one-step cluster sampling, all elements are selected in the chosen clusters. In stratified sampling, the analysis is done on elements within strata. The main objective of cluster sampling is to reduce costs by increasing sampling efficiency. Examples of cluster sampling include:

1. Houses in block
2. Students in school
3. Farmers in counties



SAMPLING A SINGLE POPULATION

- **Sampling Techniques**
  - **Cluster Sample: Divide the sample into several strata or clusters. Then take a SRS of clusters.**

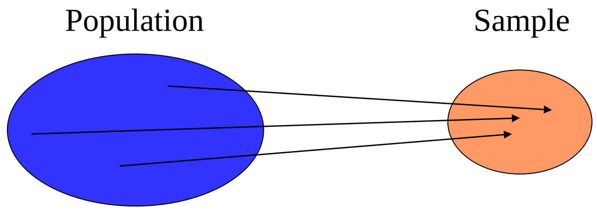


6.2 Parameters

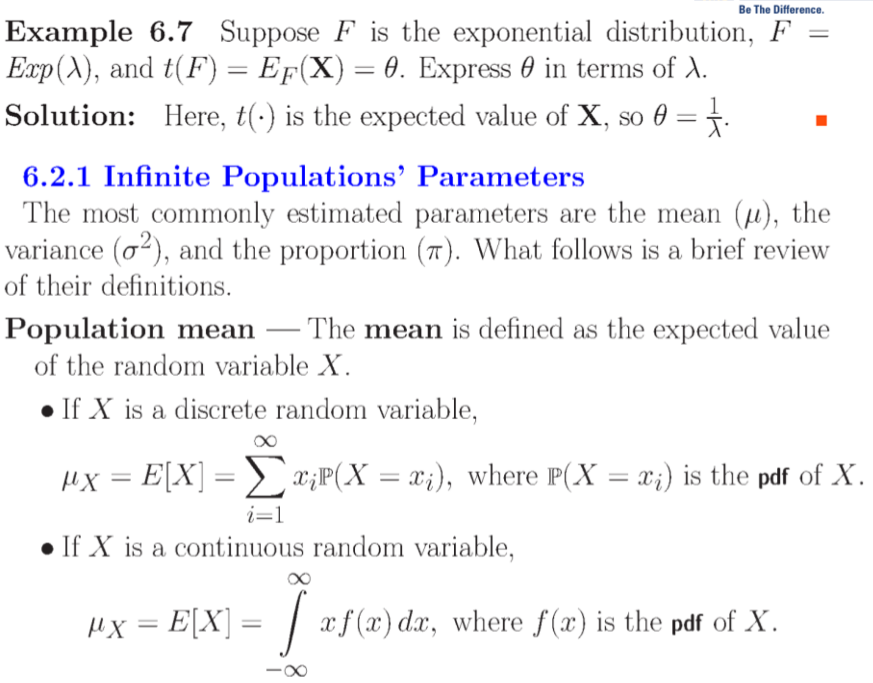
- Once a sample is taken, the primary objective becomes to extract the maximum and most precise information as possible about the population from the sample.
- Specifically, the researcher is interested in learning as much as possible about the population's **parameters**.
- A parameter, θ, is a function of the probability distribution F.
- That is, $\theta = t(F)$, where $t(\cdot)$ denotes the function applied to F. Each θ is obtained by applying some numerical procedure $t(\cdot)$ to the probability distribution function F.
- Although F has been used to denote the **cdf** exclusively until now, a more general definition of F is any description of $\mathbf{X}$'s probabilities. Note that the **cdf**, $\mathbb{P}(X \leq x)$, is included in this more general definition.
- Parameters are treated as constants in classical statistics and as random variables in Bayesian statistics. In everything that follows, parameters are treated as constants.

INFERENCE OVERVIEW

- We use sample statistics to make inference about population parameters



	Population	Sample
Mean:	μ	$\overline{X}$
Standard Deviation:	σ	s
Proportion:	π	$\hat{\pi}$



Example 6.7 Suppose F is the exponential distribution, $F = Exp(\lambda)$, and $t(F) = E_F(\mathbf{X}) = \theta$. Express θ in terms of λ.

Solution: Here, $t(\cdot)$ is the expected value of $\mathbf{X}$, so $\theta = \frac{1}{\lambda}$. ■

6.2.1 Infinite Populations' Parameters

The most commonly estimated parameters are the mean (μ), the variance (σ^2), and the proportion (π). What follows is a brief review of their definitions.

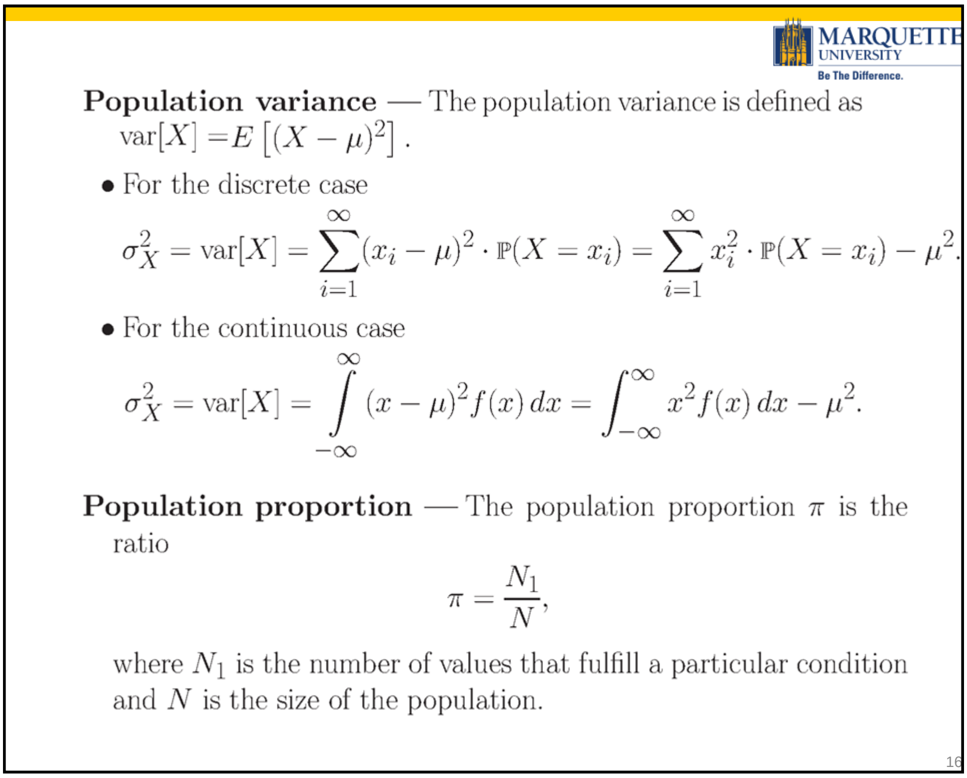
Population mean — The **mean** is defined as the expected value of the random variable X.

- If X is a discrete random variable,

$$\mu_X = E[X] = \sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i), \text{ where } \mathbb{P}(X = x_i) \text{ is the } \textbf{pdf} \text{ of } X.$$

- If X is a continuous random variable,

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f(x)\, dx, \text{ where } f(x) \text{ is the } \textbf{pdf} \text{ of } X.$$



Population variance — The population variance is defined as $\mathrm{var}[X] = E\left[(X-\mu)^2\right]$.

- For the discrete case

$$\sigma_X^2 = \mathrm{var}[X] = \sum_{i=1}^{\infty}(x_i-\mu)^2 \cdot \mathbb{P}(X=x_i) = \sum_{i=1}^{\infty} x_i^2 \cdot \mathbb{P}(X=x_i) - \mu^2.$$

- For the continuous case

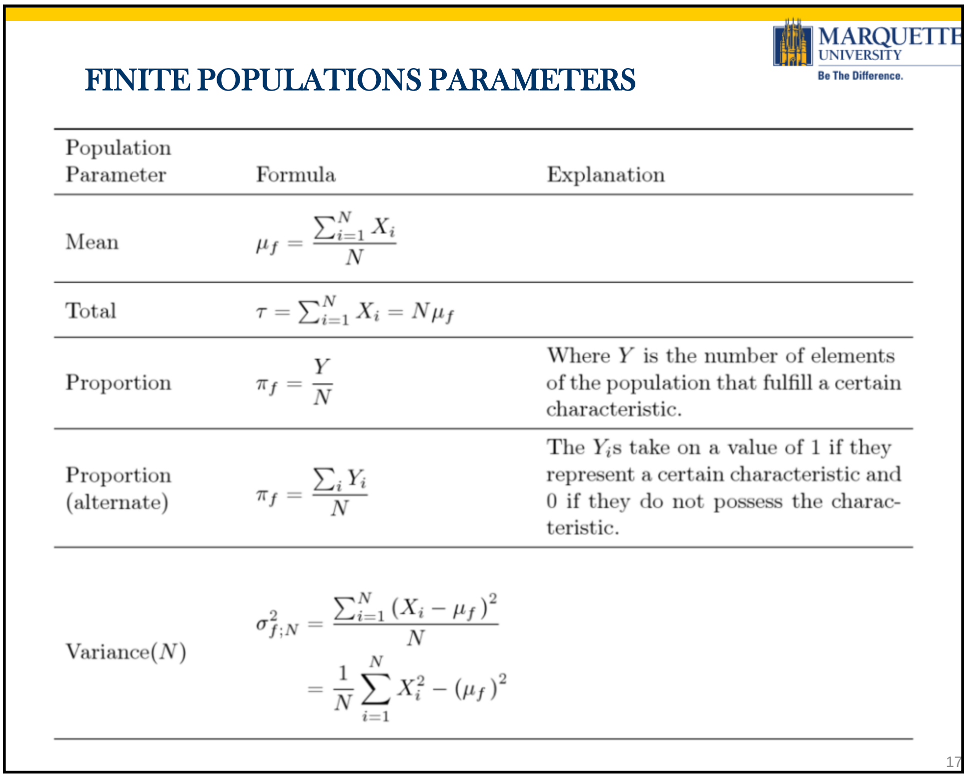
$$\sigma_X^2 = \mathrm{var}[X] = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2.$$

Population proportion — The population proportion π is the ratio

$$\pi = \frac{N_1}{N},$$

where N_1 is the number of values that fulfill a particular condition and N is the size of the population.

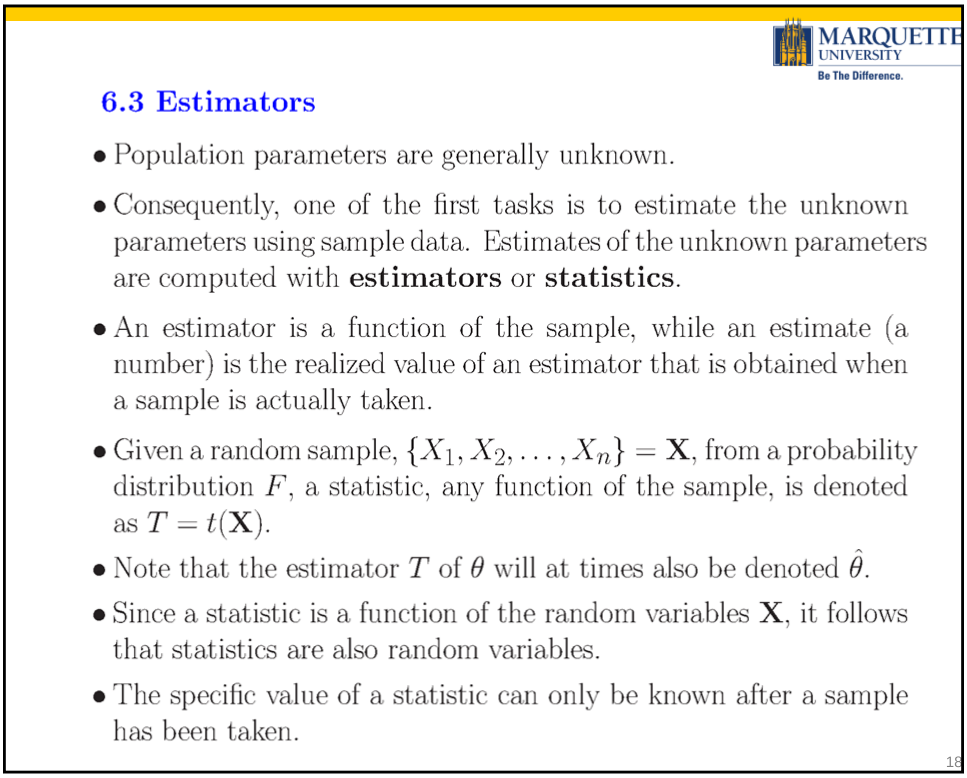
16



FINITE POPULATIONS PARAMETERS

Population Parameter	Formula	Explanation
Mean	$\mu_f = \frac{\sum_{i=1}^{N} X_i}{N}$	
Total	$\tau = \sum_{i=1}^{N} X_i = N\mu_f$	
Proportion	$\pi_f = \frac{Y}{N}$	Where Y is the number of elements of the population that fulfill a certain characteristic.
Proportion (alternate)	$\pi_f = \frac{\sum_i Y_i}{N}$	The Y_is take on a value of 1 if they represent a certain characteristic and 0 if they do not possess the characteristic.
Variance(N)	$\sigma_{f;N}^2 = \frac{\sum_{i=1}^{N}(X_i-\mu_f)^2}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - (\mu_f)^2$	

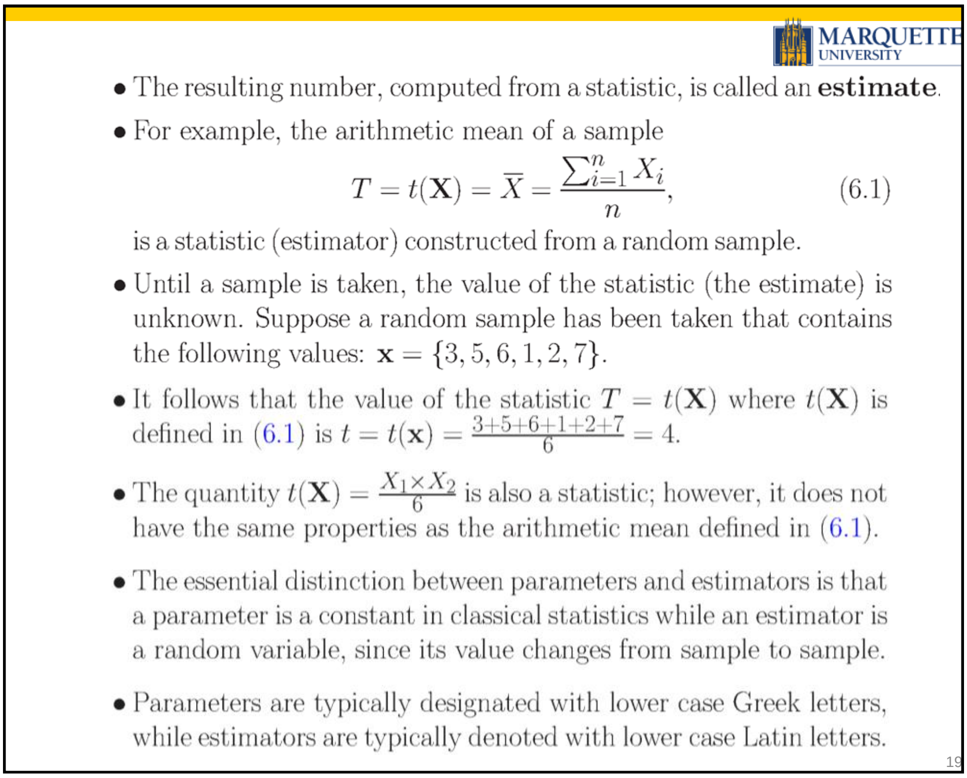
17



6.3 Estimators

- Population parameters are generally unknown.
- Consequently, one of the first tasks is to estimate the unknown parameters using sample data. Estimates of the unknown parameters are computed with **estimators** or **statistics**.
- An estimator is a function of the sample, while an estimate (a number) is the realized value of an estimator that is obtained when a sample is actually taken.
- Given a random sample, $\{X_1, X_2, \ldots, X_n\} = \mathbf{X}$, from a probability distribution F, a statistic, any function of the sample, is denoted as $T = t(\mathbf{X})$.
- Note that the estimator T of θ will at times also be denoted $\hat{\theta}$.
- Since a statistic is a function of the random variables $\mathbf{X}$, it follows that statistics are also random variables.
- The specific value of a statistic can only be known after a sample has been taken.

18



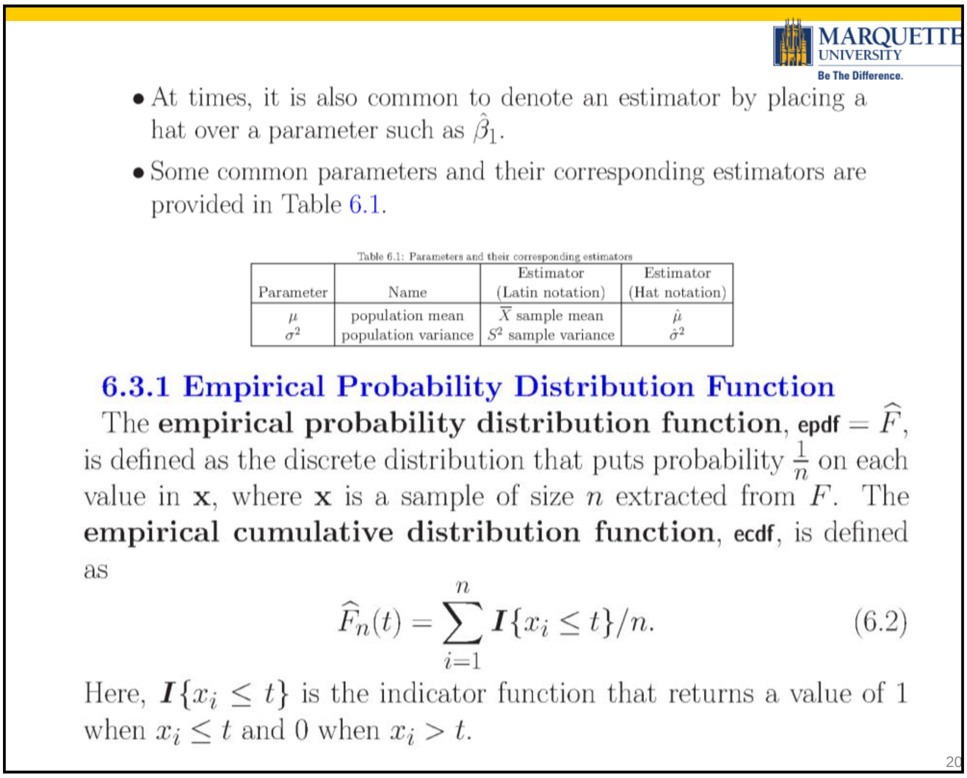
- The resulting number, computed from a statistic, is called an **estimate**.
- For example, the arithmetic mean of a sample

$$T = t(\mathbf{X}) = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \tag{6.1}$$

is a statistic (estimator) constructed from a random sample.

- Until a sample is taken, the value of the statistic (the estimate) is unknown. Suppose a random sample has been taken that contains the following values: $\mathbf{x} = \{3, 5, 6, 1, 2, 7\}$.
- It follows that the value of the statistic $T = t(\mathbf{X})$ where $t(\mathbf{X})$ is defined in (6.1) is $t = t(\mathbf{x}) = \frac{3+5+6+1+2+7}{6} = 4$.
- The quantity $t(\mathbf{X}) = \frac{X_1 \times X_2}{6}$ is also a statistic; however, it does not have the same properties as the arithmetic mean defined in (6.1).
- The essential distinction between parameters and estimators is that a parameter is a constant in classical statistics while an estimator is a random variable, since its value changes from sample to sample.
- Parameters are typically designated with lower case Greek letters, while estimators are typically denoted with lower case Latin letters.

19



- At times, it is also common to denote an estimator by placing a hat over a parameter such as $\hat{\beta}_1$.
- Some common parameters and their corresponding estimators are provided in Table 6.1.

Table 6.1: Parameters and their corresponding estimators

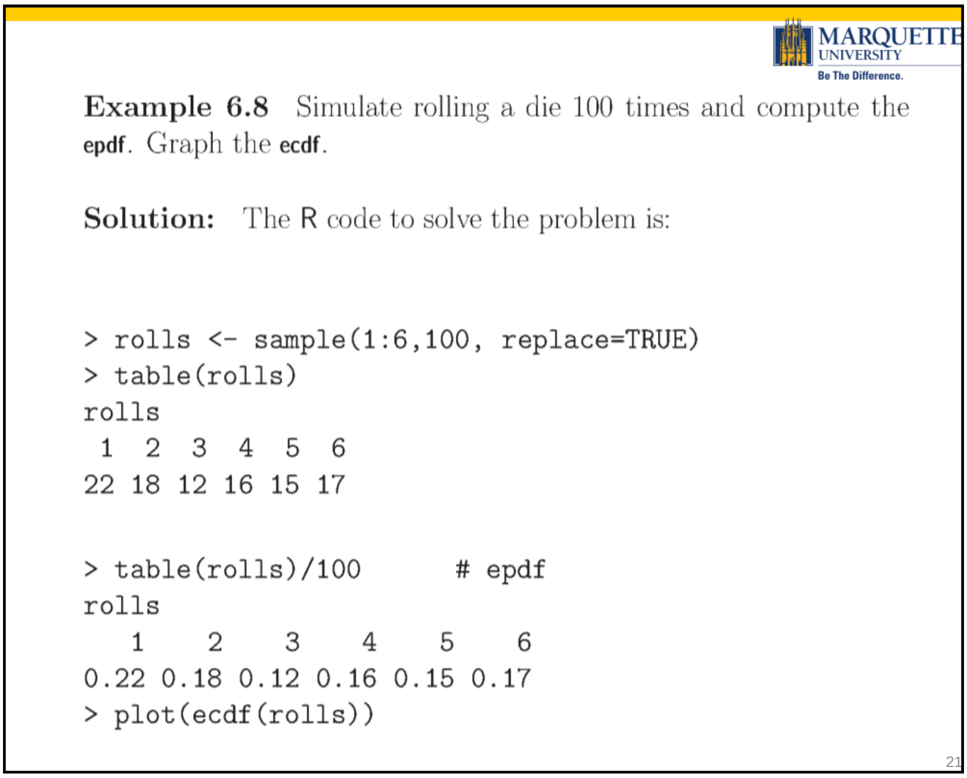
Parameter	Name	Estimator (Latin notation)	Estimator (Hat notation)
μ	population mean	$\overline{X}$ sample mean	$\hat{\mu}$
σ^2	population variance	S^2 sample variance	$\hat{\sigma}^2$

6.3.1 Empirical Probability Distribution Function

The **empirical probability distribution function**, **epdf** $= \widehat{F}$, is defined as the discrete distribution that puts probability $\frac{1}{n}$ on each value in $\mathbf{x}$, where $\mathbf{x}$ is a sample of size n extracted from F. The **empirical cumulative distribution function**, **ecdf**, is defined as

$$\widehat{F}_n(t) = \sum_{i=1}^{n} \boldsymbol{I}\{x_i \leq t\}/n. \tag{6.2}$$

Here, $\boldsymbol{I}\{x_i \leq t\}$ is the indicator function that returns a value of 1 when $x_i \leq t$ and 0 when $x_i > t$.

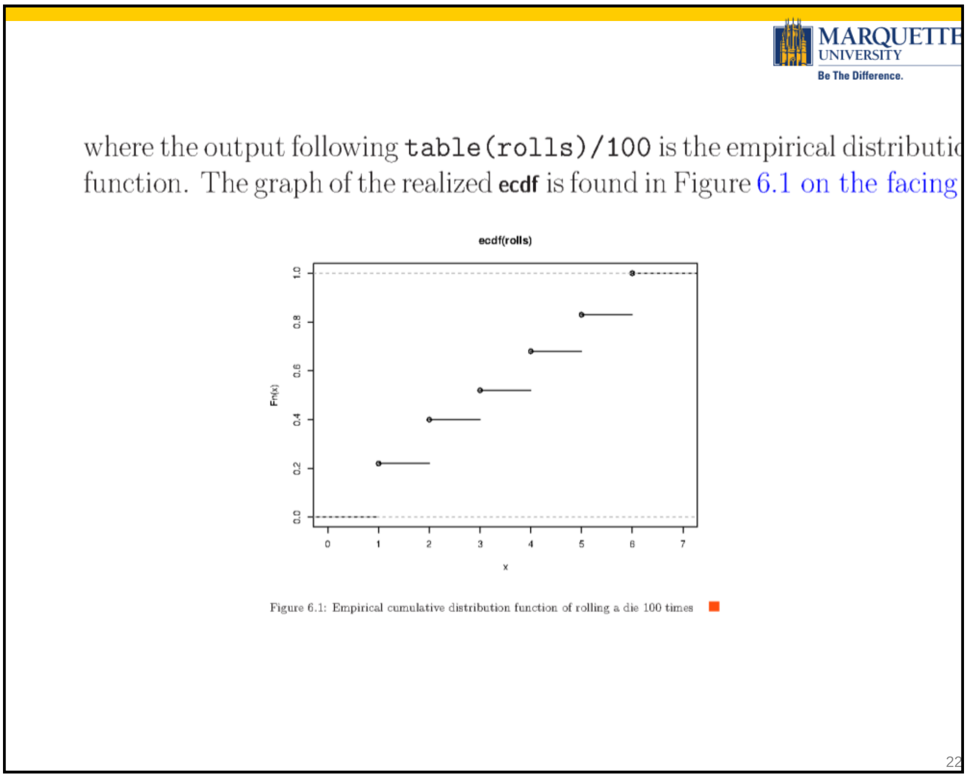


Example 6.8 Simulate rolling a die 100 times and compute the **epdf**. Graph the **ecdf**.

Solution: The R code to solve the problem is:

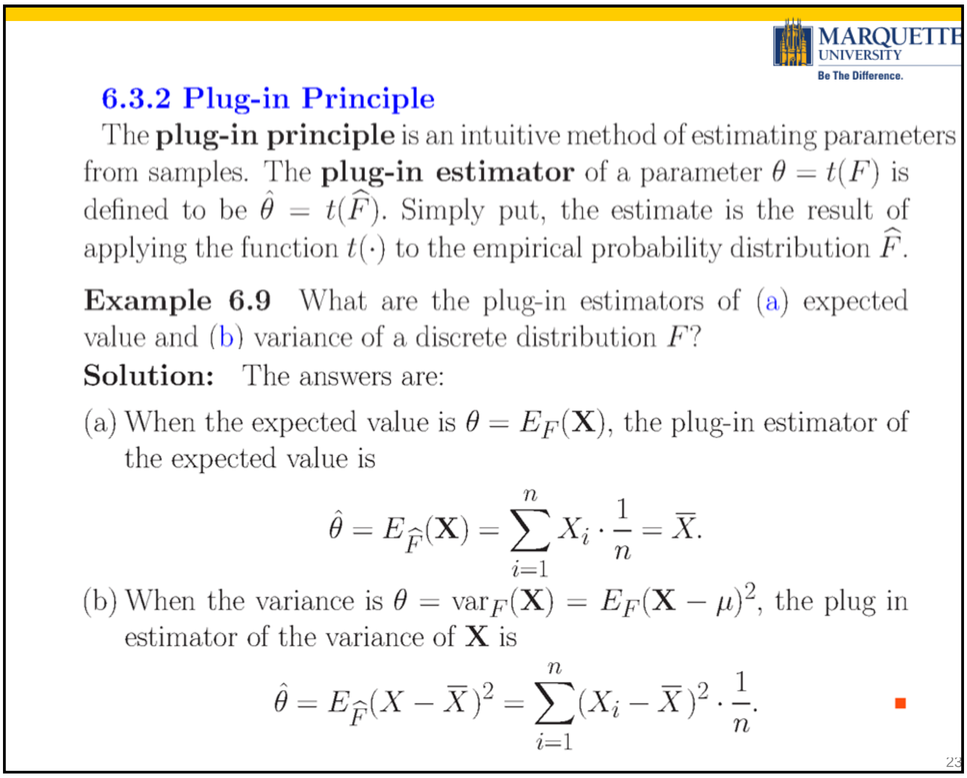
```
> rolls <- sample(1:6,100, replace=TRUE)
> table(rolls)
rolls
 1  2  3  4  5  6
22 18 12 16 15 17

> table(rolls)/100        # epdf
rolls
   1    2    3    4    5    6
0.22 0.18 0.12 0.16 0.15 0.17
> plot(ecdf(rolls))
```



where the output following `table(rolls)/100` is the empirical distributic function. The graph of the realized **ecdf** is found in Figure 6.1 on the facing

Figure 6.1: Empirical cumulative distribution function of rolling a die 100 times ■



MARQUETTE UNIVERSITY
Be The Difference.

6.3.2 Plug-in Principle

The **plug-in principle** is an intuitive method of estimating parameters from samples. The **plug-in estimator** of a parameter $\theta = t(F)$ is defined to be $\hat{\theta} = t(\widehat{F})$. Simply put, the estimate is the result of applying the function $t(\cdot)$ to the empirical probability distribution $\widehat{F}$.

Example 6.9 What are the plug-in estimators of (a) expected value and (b) variance of a discrete distribution F?

Solution: The answers are:

(a) When the expected value is $\theta = E_F(\mathbf{X})$, the plug-in estimator of the expected value is

$$\hat{\theta} = E_{\widehat{F}}(\mathbf{X}) = \sum_{i=1}^{n} X_i \cdot \frac{1}{n} = \overline{X}.$$

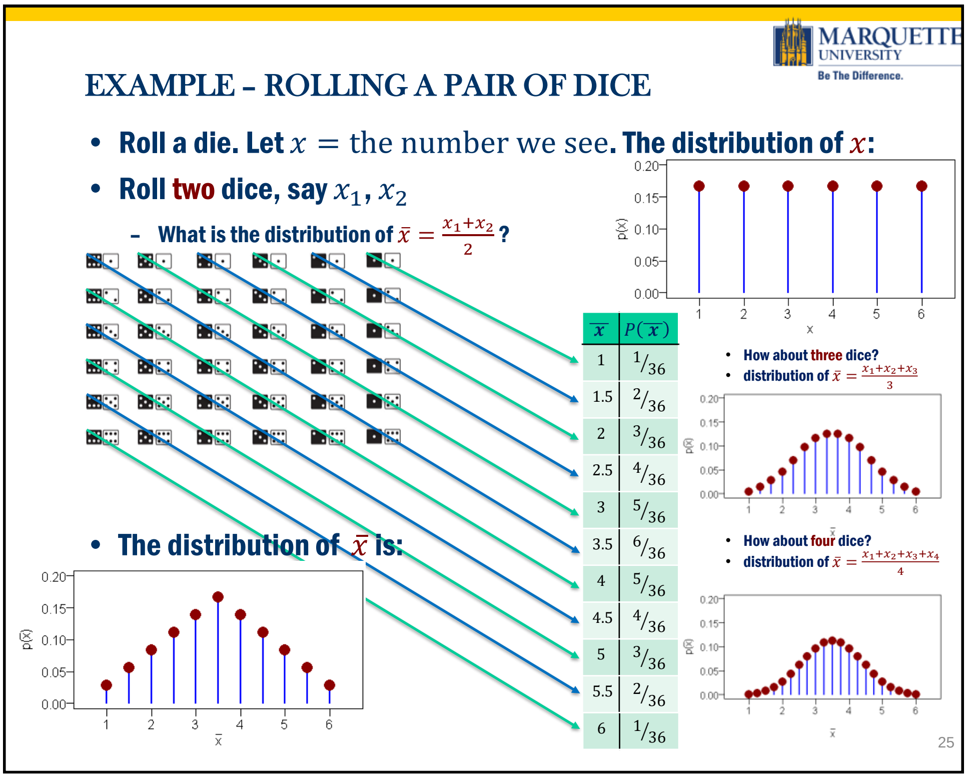
(b) When the variance is $\theta = \mathrm{var}_F(\mathbf{X}) = E_F(\mathbf{X} - \mu)^2$, the plug in estimator of the variance of $\mathbf{X}$ is

$$\hat{\theta} = E_{\widehat{F}}(X - \overline{X})^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 \cdot \frac{1}{n}.$$ ■

MARQUETTE UNIVERSITY
Be The Difference.

6.4 Sampling Distribution of $\overline{X}$

- Suppose 10 college students are randomly selected from the population of college students in the state of Colorado and compute the mean age of the sampled students.
- If this process were repeated three times, it is unlikely any of the computed sample means would be identical.
- Likewise, it is not likely that any of the three computed sample means would be exactly equal to the population mean.
- However, these sample means are typically used to estimate the unknown population mean.
- So, how can the accuracy of the sampled value be assessed?



MARQUETTE UNIVERSITY
Be The Difference.

EXAMPLE – ROLLING A PAIR OF DICE

- Roll a die. Let $x =$ the number we see. The distribution of x:
- Roll two dice, say x_1, x_2
  - What is the distribution of $\bar{x} = \frac{x_1+x_2}{2}$?

$\overline{x}$	$P(\overline{x})$
1	$^1/_{36}$
1.5	$^2/_{36}$
2	$^3/_{36}$
2.5	$^4/_{36}$
3	$^5/_{36}$
3.5	$^6/_{36}$
4	$^5/_{36}$
4.5	$^4/_{36}$
5	$^3/_{36}$
5.5	$^2/_{36}$
6	$^1/_{36}$

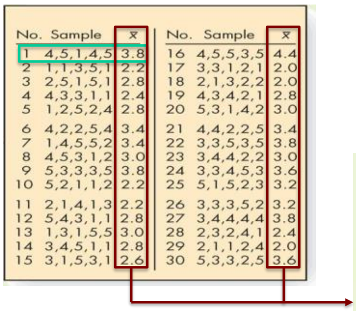
- How about three dice?
- distribution of $\bar{x} = \frac{x_1+x_2+x_3}{3}$

- How about four dice?
- distribution of $\bar{x} = \frac{x_1+x_2+x_3+x_4}{4}$

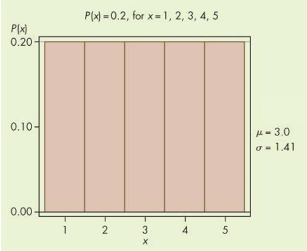
- The distribution of $\bar{x}$ is:

EXAMPLE 2 – CREATING A SAMPLING DISTRIBUTION OF SAMPLE MEANS

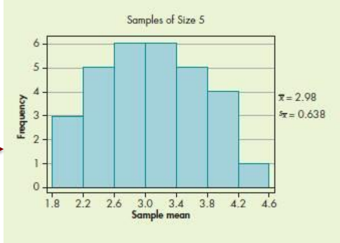
- **Let's consider a population that consists of five equally likely integers:** $1, 2, 3, 4$, **and** 5.
- **Randomly choose 30 samples of size 5 from this population.**



No.	Sample	$\bar{x}$	No.	Sample	$\bar{x}$
1	4,5,1,4,5	3.8	16	4,5,5,3,5	4.4
2	1,1,3,5,1	2.2	17	3,3,1,2,1	2.0
3	2,5,1,5,1	2.8	18	2,1,3,2,2	2.0
4	4,3,3,1,1	2.4	19	4,3,4,2,1	2.8
5	1,2,5,2,4	2.8	20	5,3,1,4,2	3.0
6	4,2,2,5,4	3.4	21	4,4,2,2,5	3.4
7	1,4,5,5,2	3.4	22	3,3,5,3,5	3.8
8	4,5,3,1,2	3.0	23	3,4,4,2,2	3.0
9	5,3,3,3,5	3.8	24	3,3,4,5,3	3.6
10	5,2,1,1,2	2.2	25	5,1,5,2,3	3.2
11	2,1,4,1,3	2.2	26	3,3,3,5,2	3.2
12	5,4,3,1,1	2.8	27	3,4,4,4,4	3.8
13	1,3,1,5,5	3.0	28	2,3,2,4,1	2.4
14	3,4,5,1,1	2.8	29	2,1,1,2,4	2.0
15	3,1,5,3,1	2.6	30	5,3,3,2,5	3.6

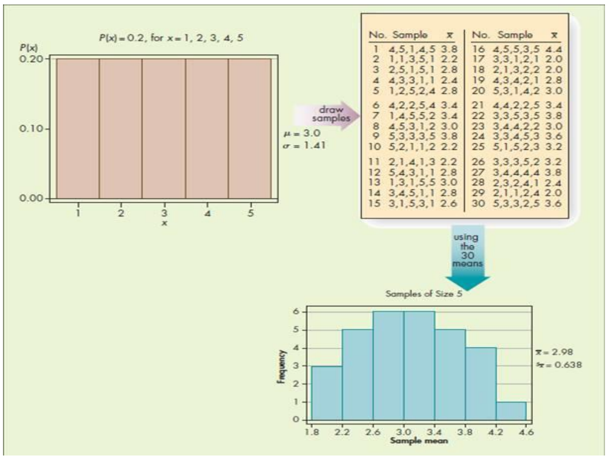


The Population: Theoretical Probability Distribution



Frequency Distribution of Sample Means

EXAMPLE 2 – CREATING A SAMPLING DISTRIBUTION OF SAMPLE MEANS



No.	Sample	$\bar{x}$	No.	Sample	$\bar{x}$
1	4,5,1,4,5	3.8	16	4,5,5,3,5	4.4
2	1,1,3,5,1	2.2	17	3,3,1,2,1	2.0
3	2,5,1,5,1	2.8	18	2,1,3,2,2	2.0
4	4,3,3,1,1	2.4	19	4,3,4,2,1	2.8
5	1,2,5,2,4	2.8	20	5,3,1,4,2	3.0
6	4,2,2,5,4	3.4	21	4,4,2,2,5	3.4
7	1,4,5,5,2	3.4	22	3,3,5,3,5	3.8
8	4,5,3,1,2	3.0	23	3,4,4,2,2	3.0
9	5,3,3,3,5	3.8	24	3,3,4,5,3	3.6
10	5,2,1,1,2	2.2	25	5,1,5,2,3	3.2
11	2,1,4,1,3	2.2	26	3,3,3,5,2	3.2
12	5,4,3,1,1	2.8	27	3,4,4,4,4	3.8
13	1,3,1,5,5	3.0	28	2,3,2,4,1	2.4
14	3,4,5,1,1	2.8	29	2,1,1,2,4	2.0
15	3,1,5,3,1	2.6	30	5,3,3,2,5	3.6

- **Looks like, under certain assumptions, the Sampling distribution of the sample mean approaches the normal distribution. (Sampling distribution applet)**

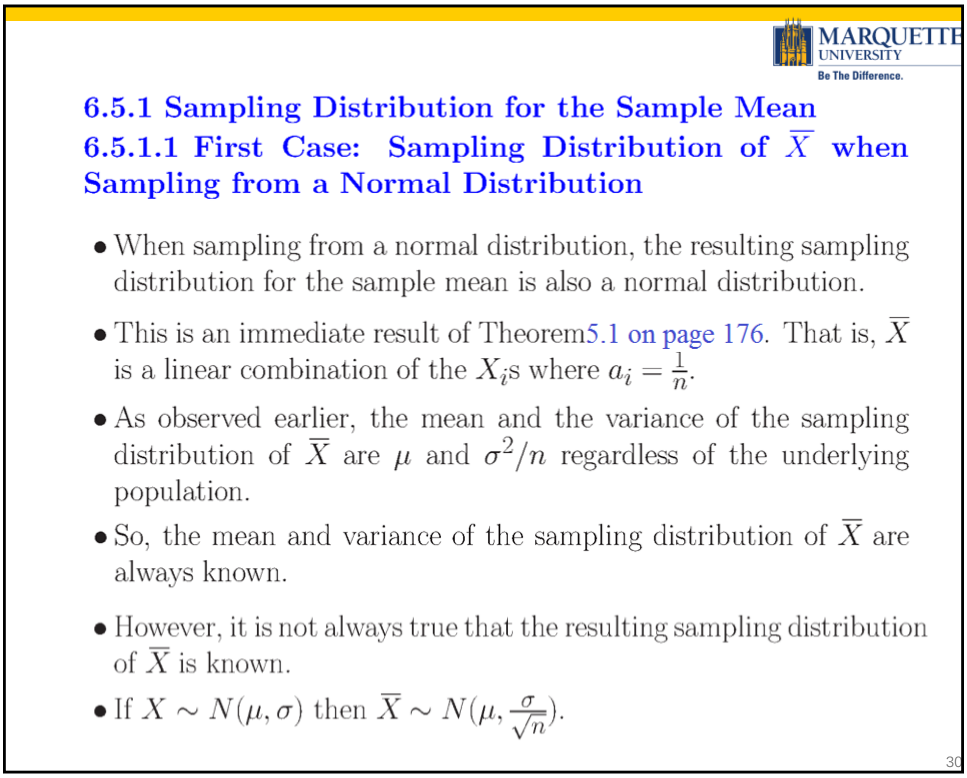
- To assess the accuracy of a value (estimate) returned from a statistic, the probability distribution of the statistic of interest is used to place probabilistic bounds on the sampling error.
- The probability distribution associated with all of the possible values a statistic can assume is called the **sampling distribution** of the statistic.
- This section presents the sampling distribution of the sample mean. Before discussing the sampling distribution of $\overline{X}$, the mean and variance of $\overline{X}$ for any random variable X are highlighted.
- If X is a random variable with mean μ and variance σ^2 and if a random sample $X_1, \ldots, X_n$ is taken, the expected value and variance of $\overline{X}$ are written

$$E\left[\overline{X}\right] = \mu_{\overline{X}} = \mu, \tag{6.3}$$

$$\operatorname{var}\left[\overline{X}\right] = \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}. \tag{6.4}$$

6.5 Sampling Distribution for a Statistic from an Infinite Population

- Consider a population from which k random samples, each of size n, are taken. In general, if given k samples, k different values for the sample mean will result.
- If k is very large, theoretically infinite, the values of the means from each of the samples, denoted $\overline{X}_i$ for each sample i, will be random variables with a resulting distribution referred to as the sampling distribution of the sample mean.
- The sampling distribution of a statistic, $t(X)$, is the resulting probability distribution for $t(X)$ calculated by taking an infinite number of random samples of size n.
- The resulting sampling distribution will typically not coincide with the distribution of the parent population.

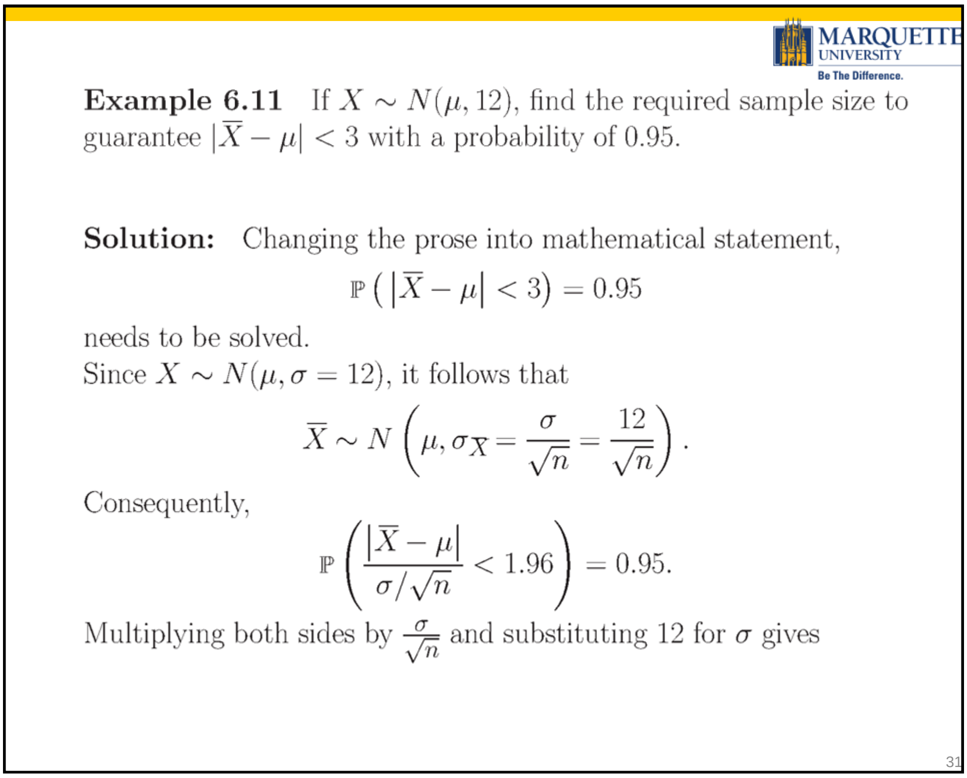


6.5.1 Sampling Distribution for the Sample Mean

6.5.1.1 First Case: Sampling Distribution of $\overline{X}$ when Sampling from a Normal Distribution

- When sampling from a normal distribution, the resulting sampling distribution for the sample mean is also a normal distribution.
- This is an immediate result of Theorem5.1 on page 176. That is, $\overline{X}$ is a linear combination of the X_is where $a_i = \frac{1}{n}$.
- As observed earlier, the mean and the variance of the sampling distribution of $\overline{X}$ are μ and σ^2/n regardless of the underlying population.
- So, the mean and variance of the sampling distribution of $\overline{X}$ are always known.
- However, it is not always true that the resulting sampling distribution of $\overline{X}$ is known.
- If $X \sim N(\mu, \sigma)$ then $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

30



Example 6.11 If $X \sim N(\mu, 12)$, find the required sample size to guarantee $|\overline{X} - \mu| < 3$ with a probability of 0.95.

Solution: Changing the prose into mathematical statement,

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| < 3\right) = 0.95$$

needs to be solved.
Since $X \sim N(\mu, \sigma = 12)$, it follows that

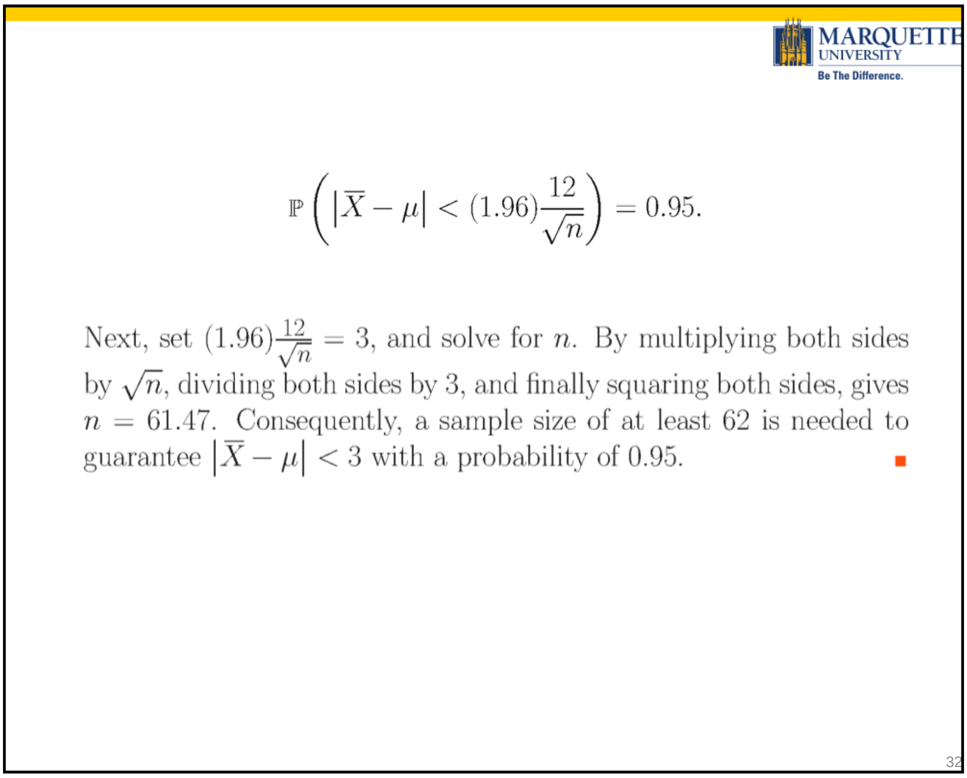
$$\overline{X} \sim N\left(\mu, \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{n}}\right).$$

Consequently,

$$\mathbb{P}\left(\frac{|\overline{X} - \mu|}{\sigma/\sqrt{n}} < 1.96\right) = 0.95.$$

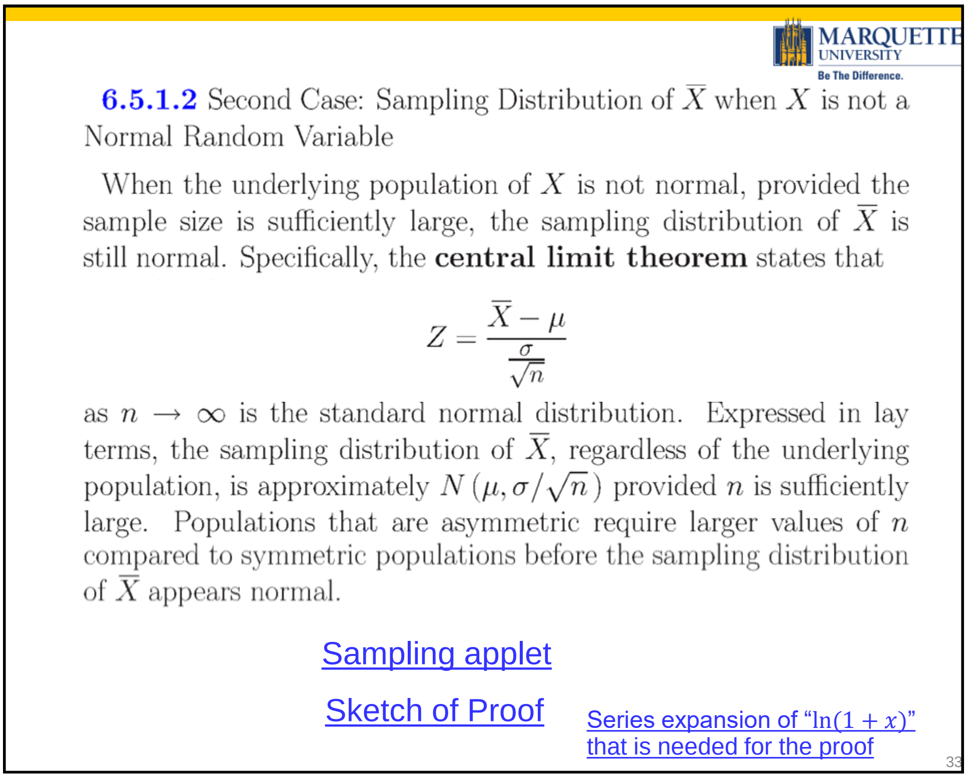
Multiplying both sides by $\frac{\sigma}{\sqrt{n}}$ and substituting 12 for σ gives

31



$$\mathbb{P}\left(\left|\overline{X}-\mu\right| < (1.96)\frac{12}{\sqrt{n}}\right) = 0.95.$$

Next, set $(1.96)\frac{12}{\sqrt{n}} = 3$, and solve for n. By multiplying both sides by $\sqrt{n}$, dividing both sides by 3, and finally squaring both sides, gives $n = 61.47$. Consequently, a sample size of at least 62 is needed to guarantee $\left|\overline{X}-\mu\right| < 3$ with a probability of 0.95. ■



6.5.1.2 Second Case: Sampling Distribution of $\overline{X}$ when X is not a Normal Random Variable

When the underlying population of X is not normal, provided the sample size is sufficiently large, the sampling distribution of $\overline{X}$ is still normal. Specifically, the **central limit theorem** states that

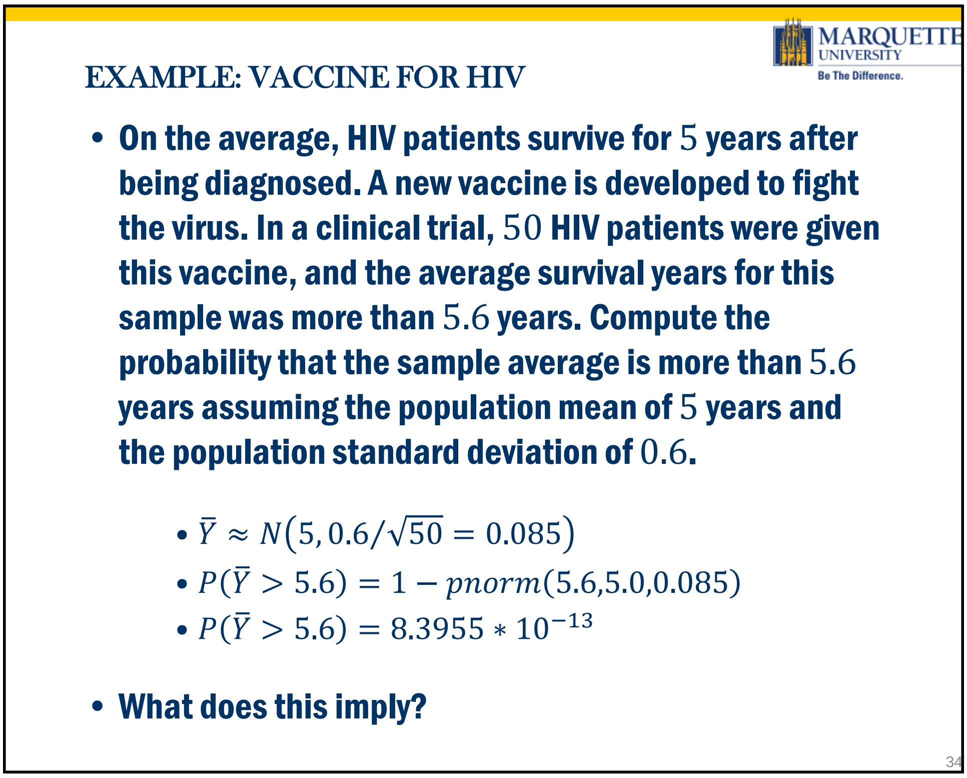
$$Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

as $n \to \infty$ is the standard normal distribution. Expressed in lay terms, the sampling distribution of $\overline{X}$, regardless of the underlying population, is approximately $N\left(\mu, \sigma/\sqrt{n}\right)$ provided n is sufficiently large. Populations that are asymmetric require larger values of n compared to symmetric populations before the sampling distribution of $\overline{X}$ appears normal.

Sampling applet

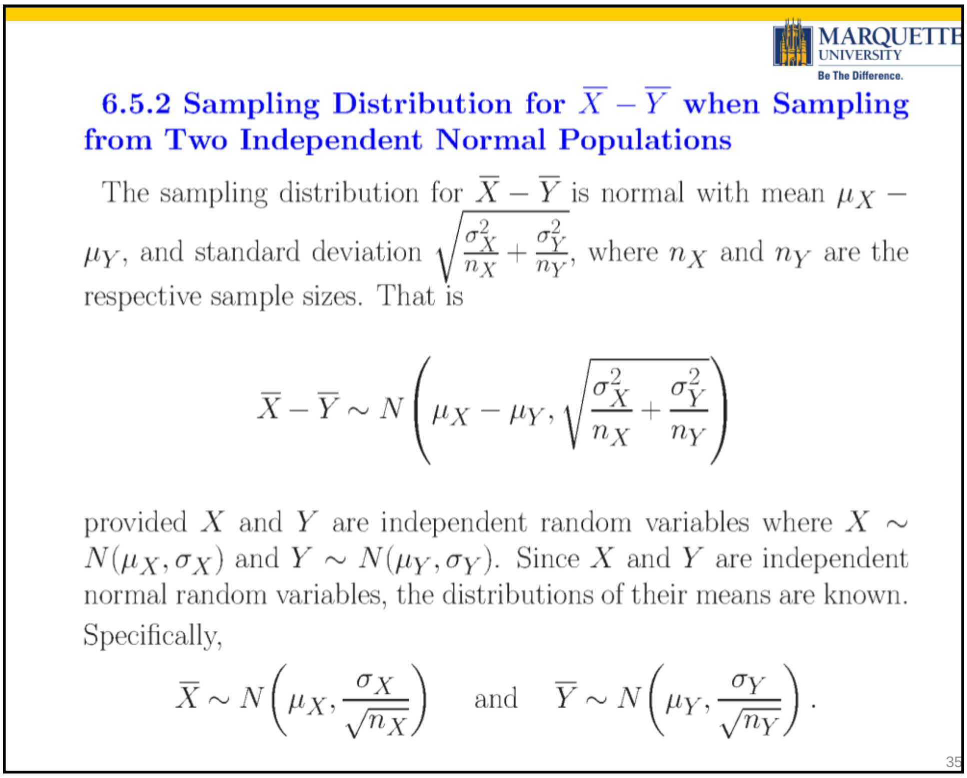
Sketch of Proof

Series expansion of "$\ln(1+x)$" that is needed for the proof



EXAMPLE: VACCINE FOR HIV

- **On the average, HIV patients survive for 5 years after being diagnosed. A new vaccine is developed to fight the virus. In a clinical trial, 50 HIV patients were given this vaccine, and the average survival years for this sample was more than 5.6 years. Compute the probability that the sample average is more than 5.6 years assuming the population mean of 5 years and the population standard deviation of 0.6.**
  - $\bar{Y} \approx N\left(5, 0.6/\sqrt{50} = 0.085\right)$
  - $P(\bar{Y} > 5.6) = 1 - pnorm(5.6,5.0,0.085)$
  - $P(\bar{Y} > 5.6) = 8.3955 * 10^{-13}$
- **What does this imply?**



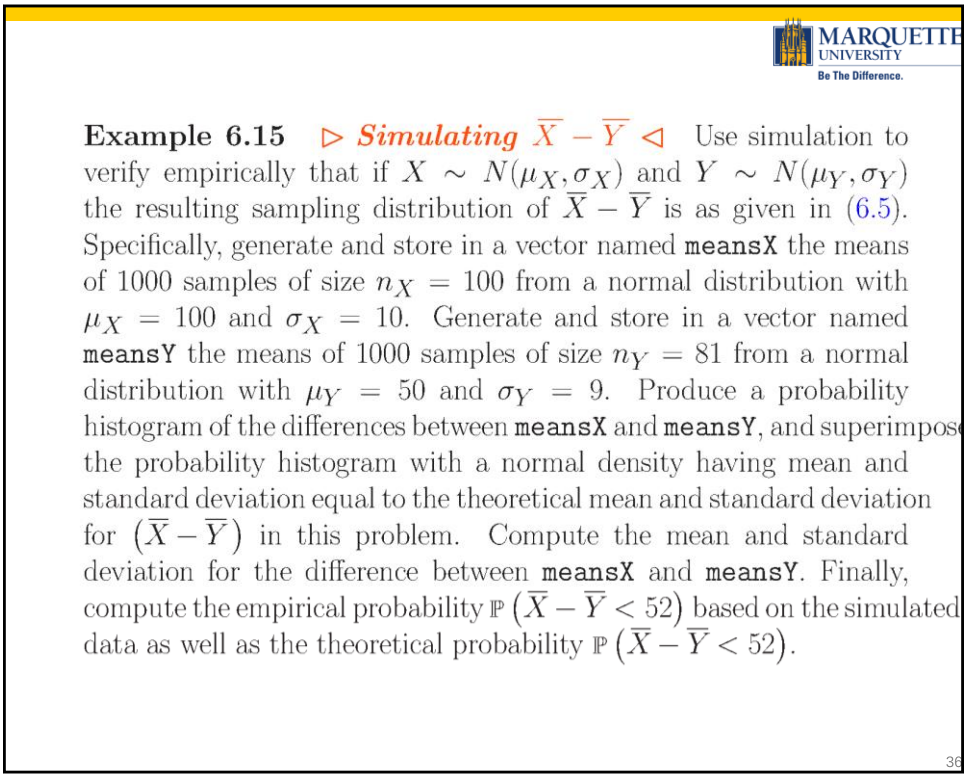
6.5.2 Sampling Distribution for $\bar{X} - \bar{Y}$ when Sampling from Two Independent Normal Populations

The sampling distribution for $\bar{X} - \bar{Y}$ is normal with mean $\mu_X - \mu_Y$, and standard deviation $\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$, where n_X and n_Y are the respective sample sizes. That is

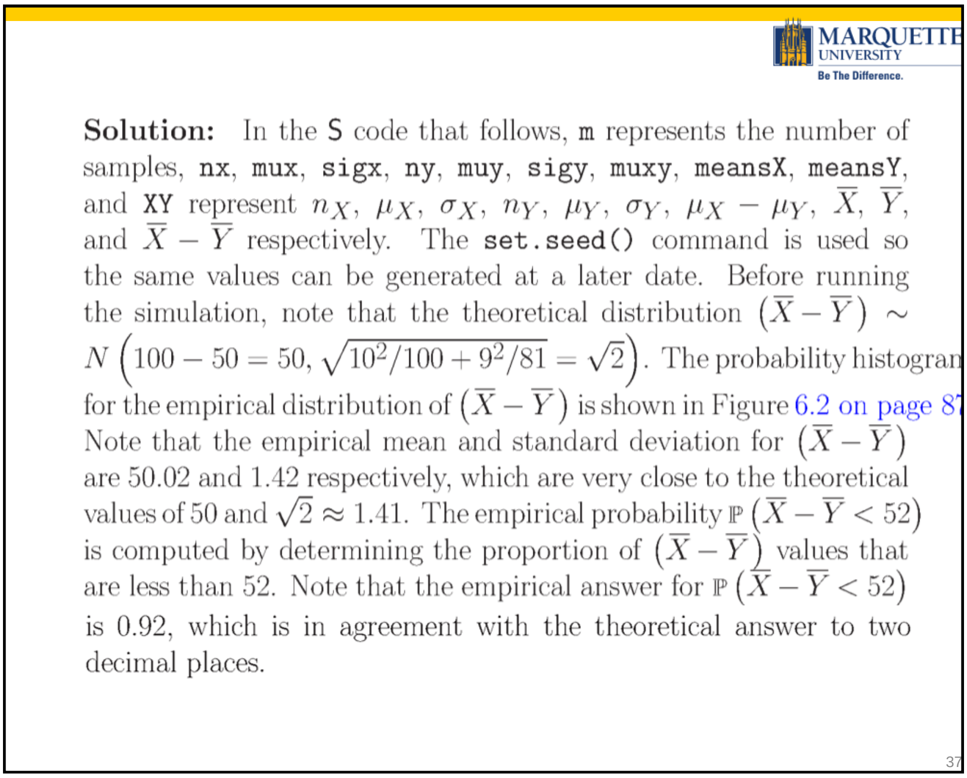
$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right)$$

provided X and Y are independent random variables where $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$. Since X and Y are independent normal random variables, the distributions of their means are known. Specifically,

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n_X}}\right) \quad \text{and} \quad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y}{\sqrt{n_Y}}\right).$$



Example 6.15 ▷ ***Simulating*** $\overline{X} - \overline{Y}$ ◁ Use simulation to verify empirically that if $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$ the resulting sampling distribution of $\overline{X} - \overline{Y}$ is as given in (6.5). Specifically, generate and store in a vector named **`meansX`** the means of 1000 samples of size $n_X = 100$ from a normal distribution with $\mu_X = 100$ and $\sigma_X = 10$. Generate and store in a vector named **`meansY`** the means of 1000 samples of size $n_Y = 81$ from a normal distribution with $\mu_Y = 50$ and $\sigma_Y = 9$. Produce a probability histogram of the differences between **`meansX`** and **`meansY`**, and superimpos the probability histogram with a normal density having mean and standard deviation equal to the theoretical mean and standard deviation for $\left(\overline{X} - \overline{Y}\right)$ in this problem. Compute the mean and standard deviation for the difference between **`meansX`** and **`meansY`**. Finally, compute the empirical probability $\mathbb{P}\left(\overline{X} - \overline{Y} < 52\right)$ based on the simulated data as well as the theoretical probability $\mathbb{P}\left(\overline{X} - \overline{Y} < 52\right)$.



Solution: In the S code that follows, **`m`** represents the number of samples, **`nx`**, **`mux`**, **`sigx`**, **`ny`**, **`muy`**, **`sigy`**, **`muxy`**, **`meansX`**, **`meansY`**, and **`XY`** represent n_X, μ_X, σ_X, n_Y, μ_Y, σ_Y, $\mu_X - \mu_Y$, $\overline{X}$, $\overline{Y}$, and $\overline{X} - \overline{Y}$ respectively. The **`set.seed()`** command is used so the same values can be generated at a later date. Before running the simulation, note that the theoretical distribution $\left(\overline{X} - \overline{Y}\right) \sim N\left(100 - 50 = 50, \sqrt{10^2/100 + 9^2/81} = \sqrt{2}\right)$. The probability histogran for the empirical distribution of $\left(\overline{X} - \overline{Y}\right)$ is shown in Figure 6.2 on page 8 Note that the empirical mean and standard deviation for $\left(\overline{X} - \overline{Y}\right)$ are 50.02 and 1.42 respectively, which are very close to the theoretical values of 50 and $\sqrt{2} \approx 1.41$. The empirical probability $\mathbb{P}\left(\overline{X} - \overline{Y} < 52\right)$ is computed by determining the proportion of $\left(\overline{X} - \overline{Y}\right)$ values that are less than 52. Note that the empirical answer for $\mathbb{P}\left(\overline{X} - \overline{Y} < 52\right)$ is 0.92, which is in agreement with the theoretical answer to two decimal places.

R CODE:

```
set.seed(17)
m <- 1000
nx <- 100; ny <- 81
mux <- 100; sigx <- 10
muy <- 50; sigy <- 9
muxy <- mux - muy
sigxy <- sqrt((sigx^2/nx) + (sigy^2/ny))
meansX <- array(0, m) # Array of m zeros
meansY <- array(0, m) # Array of m zeros
for(i in 1:m) {meansX[i] <- mean(rnorm(nx, mux, sigx))}
for(i in 1:m) {meansY[i] <- mean(rnorm(ny, muy, sigy))}
XY <- meansX - meansY
ll <- muxy - 3.4 * sigxy
ul <- muxy + 3.4 * sigxy
hist(XY, prob = T, xlab = "xbar-ybar", nclass = "scott",
col="cyan", xlim = c(ll, ul), ylim = c(0, 0.3))
lines(seq(ll, ul, 0.05),dnorm(seq(ll,ul,0.05),muxy,sigxy),
col=1, lwd=2)
```

R CODE (CON'T) :

```
print(round(c(mean(XY), sqrt(var(XY))), 2))
[1] 50.02 1.42
sum(XY < 52)/1000
[1] 0.92
round(pnorm(52, 50, sqrt(2)), 2)
[1] 0.92
```

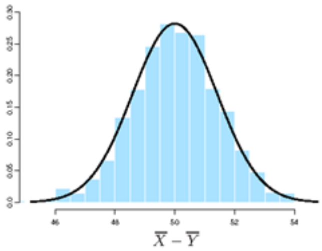
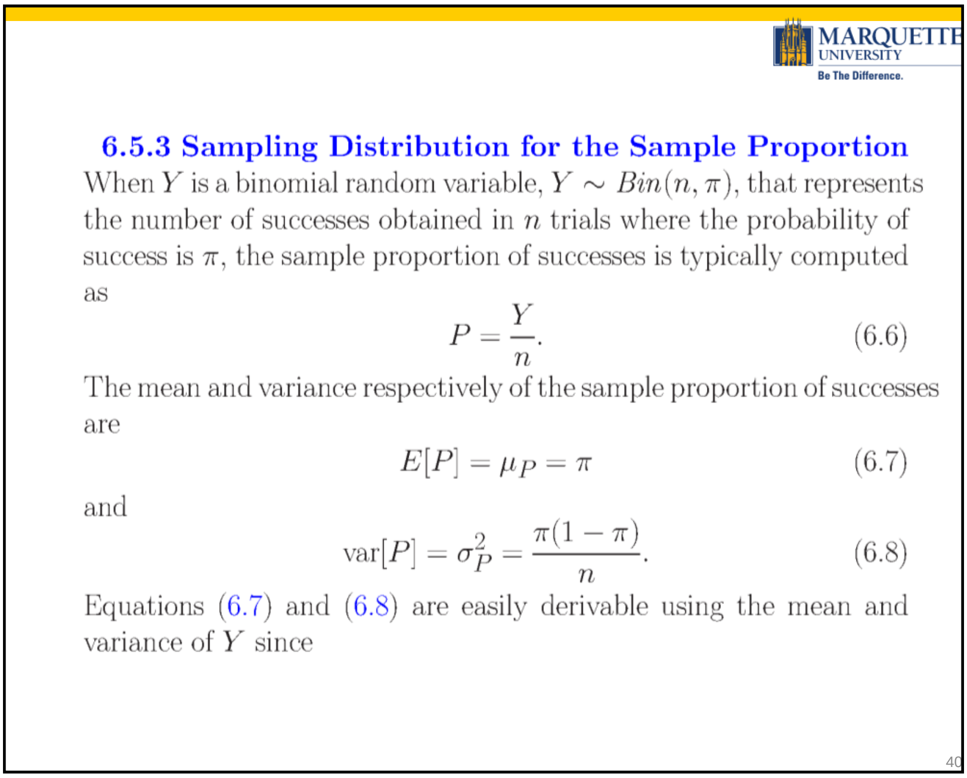


Figure 6.2: Probability histogram for simulated distribution of $(\overline{X} - \overline{Y})$ with superimposed normal density with $\mu = 50$, and $\sigma = \sqrt{2}$.



6.5.3 Sampling Distribution for the Sample Proportion

When Y is a binomial random variable, $Y \sim Bin(n, \pi)$, that represents the number of successes obtained in n trials where the probability of success is π, the sample proportion of successes is typically computed as

$$P = \frac{Y}{n}. \tag{6.6}$$

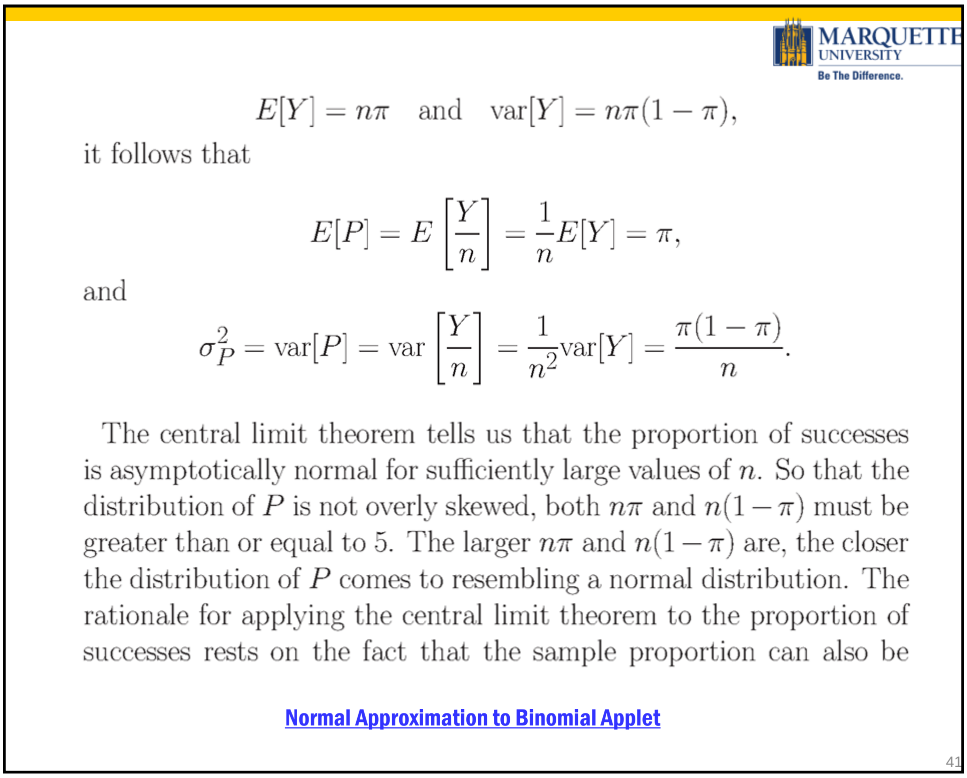
The mean and variance respectively of the sample proportion of successes are

$$E[P] = \mu_P = \pi \tag{6.7}$$

and

$$\text{var}[P] = \sigma_P^2 = \frac{\pi(1-\pi)}{n}. \tag{6.8}$$

Equations (6.7) and (6.8) are easily derivable using the mean and variance of Y since



$$E[Y] = n\pi \quad \text{and} \quad \text{var}[Y] = n\pi(1-\pi),$$

it follows that

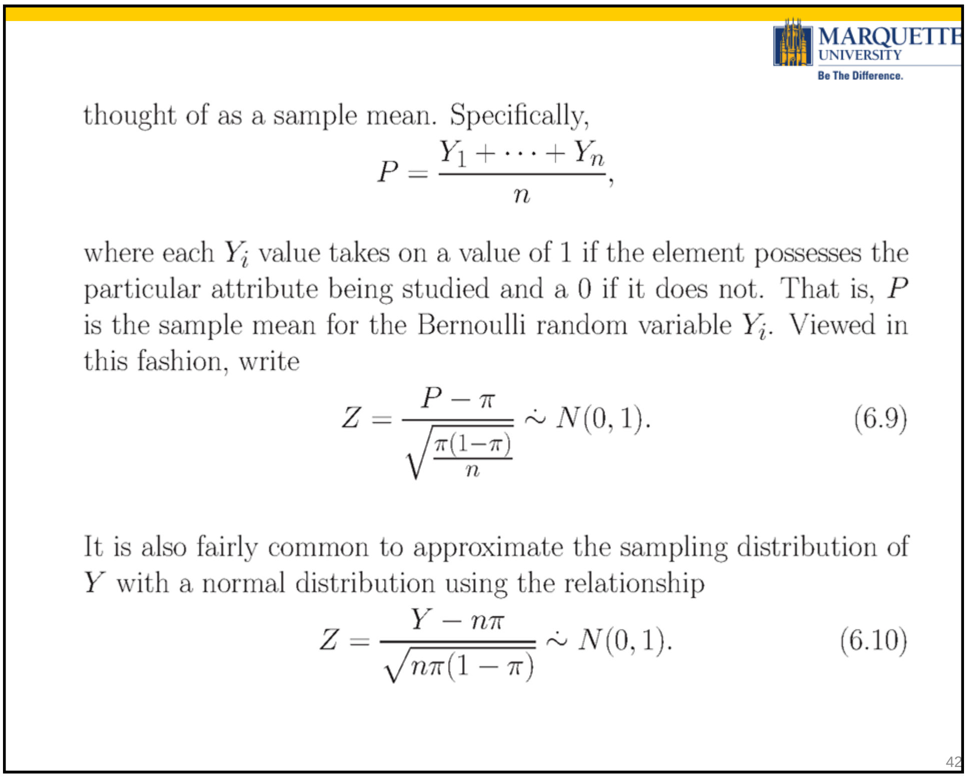
$$E[P] = E\left[\frac{Y}{n}\right] = \frac{1}{n}E[Y] = \pi,$$

and

$$\sigma_P^2 = \text{var}[P] = \text{var}\left[\frac{Y}{n}\right] = \frac{1}{n^2}\text{var}[Y] = \frac{\pi(1-\pi)}{n}.$$

The central limit theorem tells us that the proportion of successes is asymptotically normal for sufficiently large values of n. So that the distribution of P is not overly skewed, both $n\pi$ and $n(1-\pi)$ must be greater than or equal to 5. The larger $n\pi$ and $n(1-\pi)$ are, the closer the distribution of P comes to resembling a normal distribution. The rationale for applying the central limit theorem to the proportion of successes rests on the fact that the sample proportion can also be

Normal Approximation to Binomial Applet



thought of as a sample mean. Specifically,

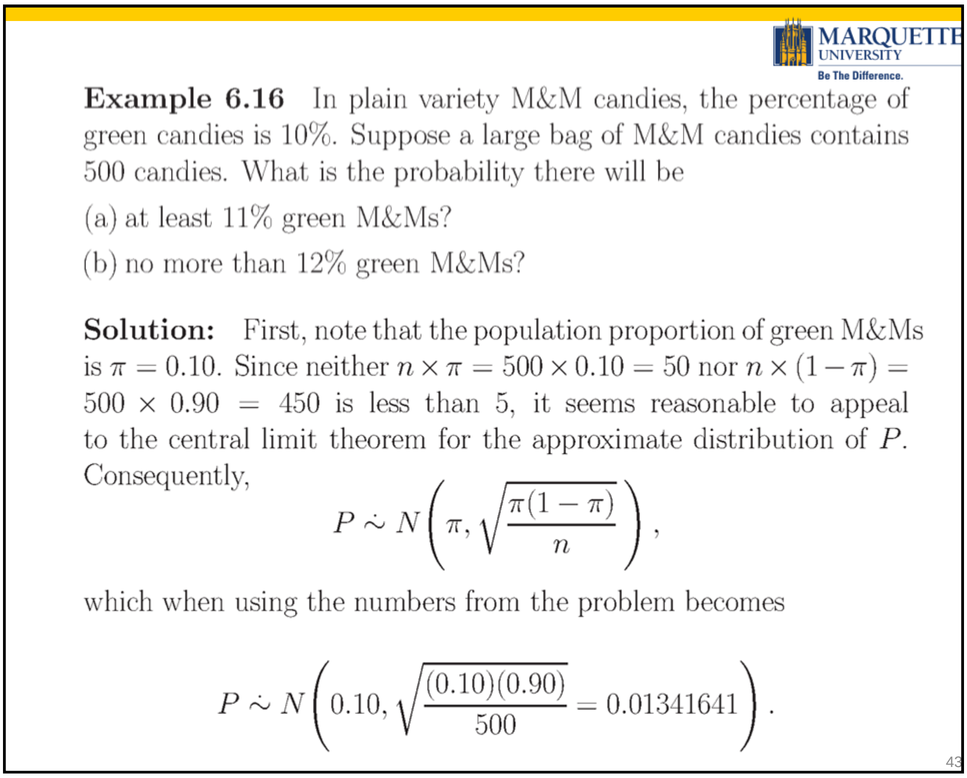
$$P = \frac{Y_1 + \cdots + Y_n}{n},$$

where each Y_i value takes on a value of 1 if the element possesses the particular attribute being studied and a 0 if it does not. That is, P is the sample mean for the Bernoulli random variable Y_i. Viewed in this fashion, write

$$Z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \mathrel{\dot{\sim}} N(0,1). \tag{6.9}$$

It is also fairly common to approximate the sampling distribution of Y with a normal distribution using the relationship

$$Z = \frac{Y - n\pi}{\sqrt{n\pi(1-\pi)}} \mathrel{\dot{\sim}} N(0,1). \tag{6.10}$$



Example 6.16 In plain variety M&M candies, the percentage of green candies is 10%. Suppose a large bag of M&M candies contains 500 candies. What is the probability there will be

(a) at least 11% green M&Ms?

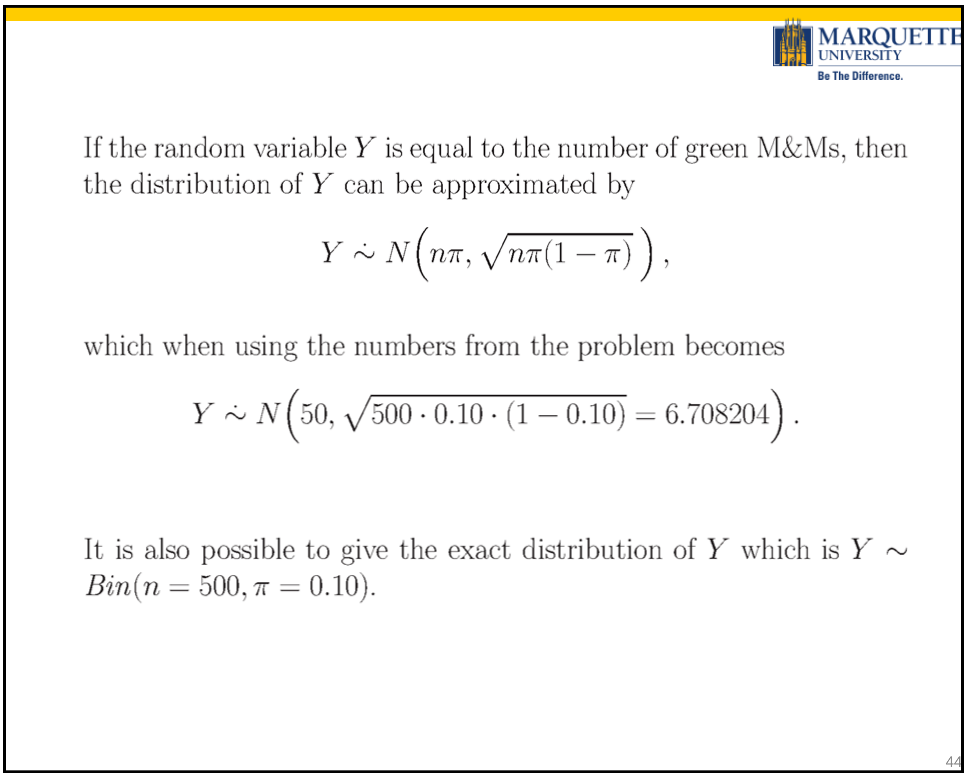
(b) no more than 12% green M&Ms?

Solution: First, note that the population proportion of green M&Ms is $\pi = 0.10$. Since neither $n \times \pi = 500 \times 0.10 = 50$ nor $n \times (1 - \pi) = 500 \times 0.90 = 450$ is less than 5, it seems reasonable to appeal to the central limit theorem for the approximate distribution of P. Consequently,

$$P \mathrel{\dot{\sim}} N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right),$$

which when using the numbers from the problem becomes

$$P \mathrel{\dot{\sim}} N\left(0.10, \sqrt{\frac{(0.10)(0.90)}{500}} = 0.01341641\right).$$



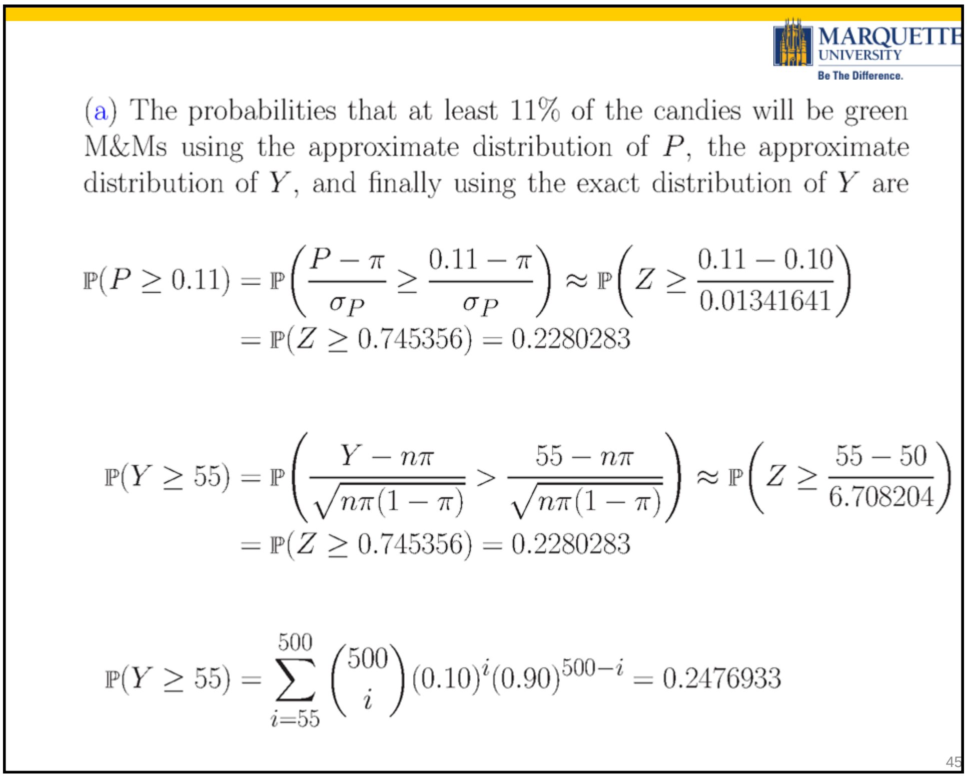
If the random variable Y is equal to the number of green M&Ms, then the distribution of Y can be approximated by

$$Y \dot{\sim} N\left(n\pi, \sqrt{n\pi(1-\pi)}\right),$$

which when using the numbers from the problem becomes

$$Y \dot{\sim} N\left(50, \sqrt{500 \cdot 0.10 \cdot (1-0.10)} = 6.708204\right).$$

It is also possible to give the exact distribution of Y which is $Y \sim Bin(n = 500, \pi = 0.10)$.



(a) The probabilities that at least 11% of the candies will be green M&Ms using the approximate distribution of P, the approximate distribution of Y, and finally using the exact distribution of Y are

$$\begin{aligned}\mathbb{P}(P \geq 0.11) &= \mathbb{P}\left(\frac{P-\pi}{\sigma_P} \geq \frac{0.11-\pi}{\sigma_P}\right) \approx \mathbb{P}\left(Z \geq \frac{0.11-0.10}{0.01341641}\right)\\ &= \mathbb{P}(Z \geq 0.745356) = 0.2280283\end{aligned}$$

$$\begin{aligned}\mathbb{P}(Y \geq 55) &= \mathbb{P}\left(\frac{Y-n\pi}{\sqrt{n\pi(1-\pi)}} > \frac{55-n\pi}{\sqrt{n\pi(1-\pi)}}\right) \approx \mathbb{P}\left(Z \geq \frac{55-50}{6.708204}\right)\\ &= \mathbb{P}(Z \geq 0.745356) = 0.2280283\end{aligned}$$

$$\mathbb{P}(Y \geq 55) = \sum_{i=55}^{500} \binom{500}{i} (0.10)^i (0.90)^{500-i} = 0.2476933$$

(b) The probability that no more than 12% of the candies will be green M&Ms is

$$\mathbb{P}(P \le 0.12) = \mathbb{P}\left(\frac{P-\pi}{\sigma_P} \le \frac{0.12-\pi}{\sigma_P}\right) \approx \mathbb{P}\left(Z \le \frac{0.12-0.10}{0.01341641}\right)$$
$$= \mathbb{P}(Z \le 1.490712) = 0.9319814$$

$$\mathbb{P}(Y \le 60) = \mathbb{P}\left(\frac{Y-n\pi}{\sqrt{n\pi(1-\pi)}} > \frac{60-n\pi}{\sqrt{n\pi(1-\pi)}}\right) \approx \mathbb{P}\left(Z \le \frac{60-50}{6.708204}\right)$$
$$= \mathbb{P}(Z \le 1.490712) = 0.9319814$$

$$\mathbb{P}(Y \le 60) = \sum_{i=0}^{60} \binom{500}{i} (0.10)^i (0.90)^{500-i} = 0.9381745.$$

R CODE:

The following S commands compute the answers for (a) and (b).

```
➢ 1 - pnorm(0.11,0.10,sqrt(0.1*0.9/500))
[1] 0.2280283
➢ 1 - pnorm(55,500*0.1,sqrt(500*0.1*0.9))
[1] 0.2280283
➢ 1 - pbinom(54,500,0.10)
[1] 0.2476933

➢ pnorm(0.12,0.10,sqrt(0.1*0.9/500))
[1] 0.9319814
➢ pnorm(60,500*.10,sqrt(500*0.1*0.9))
[1] 0.9319814
➢ pbinom(60,500,0.1)
[1] 0.9381745
```

The astute observer will notice that the approximations are not equal to the exact answers. This is due to the fact that a continuous distribution has been used to approximate a discrete distribution.

The accuracy of the answers can be improved by applying what is called a **continuity correction**. Using the continuity correction, (6.9) and (6.10) become

$$Z = \frac{P \pm \frac{0.5}{n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \dot{\sim} N(0,1) \qquad (6.11)$$

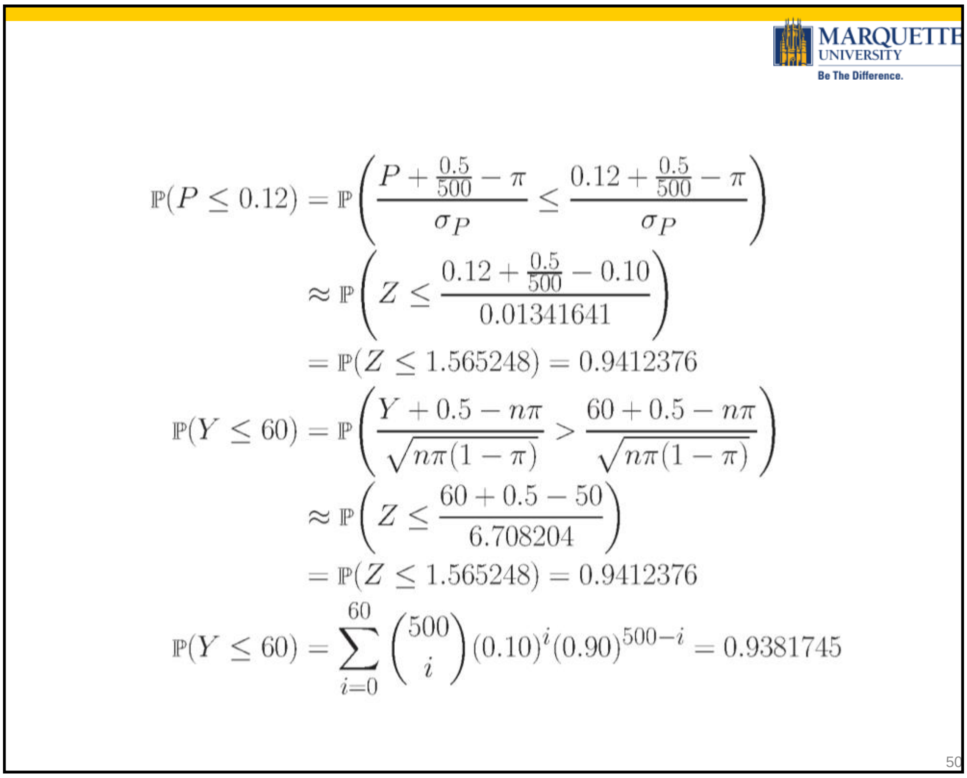
and

$$Z = \frac{Y \pm 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}} \dot{\sim} N(0,1). \qquad (6.12)$$

Normal Approximation to Binomial Applet

When solving less than or equal type inequalities, add the continuity correction; and when solving greater than or equal type inequalities, subtract the continuity correction. Notice how much closer the approximat are to the exact answers when the appropriate continuity corrections are applied.

$$\begin{aligned}
\mathbb{P}(P \geq 0.11) &= \mathbb{P}\left(\frac{P - \frac{0.5}{500} - \pi}{\sigma_P} \geq \frac{0.11 - \frac{0.5}{500} - \pi}{\sigma_P}\right) \\
&\approx \mathbb{P}\left(Z \geq \frac{0.11 - \frac{0.5}{500} - 0.10}{0.01341641}\right) \\
&= \mathbb{P}(Z \geq 0.6708204) = 0.2511675 \\
\mathbb{P}(Y \geq 55) &= \mathbb{P}\left(\frac{Y - 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}} > \frac{55 - 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) \\
&\approx \mathbb{P}\left(Z \geq \frac{55 - 0.5 - 50}{6.708204}\right) \\
&= \mathbb{P}(Z \geq 0.6708204) = 0.2511675 \\
\mathbb{P}(Y \geq 55) &= \sum_{i=55}^{500} \binom{500}{i} (0.10)^i (0.90)^{500-i} = 0.2476933
\end{aligned}$$



$$\mathbb{P}(P \le 0.12) = \mathbb{P}\left(\frac{P + \frac{0.5}{500} - \pi}{\sigma_P} \le \frac{0.12 + \frac{0.5}{500} - \pi}{\sigma_P}\right)$$

$$\approx \mathbb{P}\left(Z \le \frac{0.12 + \frac{0.5}{500} - 0.10}{0.01341641}\right)$$

$$= \mathbb{P}(Z \le 1.565248) = 0.9412376$$

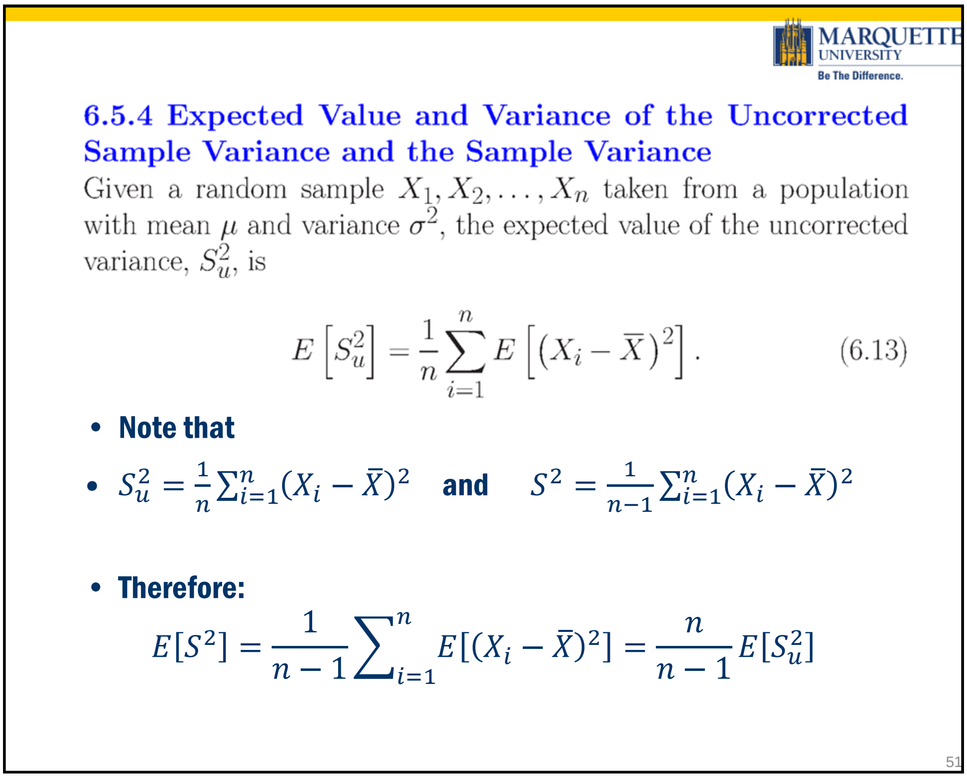
$$\mathbb{P}(Y \le 60) = \mathbb{P}\left(\frac{Y + 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}} > \frac{60 + 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

$$\approx \mathbb{P}\left(Z \le \frac{60 + 0.5 - 50}{6.708204}\right)$$

$$= \mathbb{P}(Z \le 1.565248) = 0.9412376$$

$$\mathbb{P}(Y \le 60) = \sum_{i=0}^{60} \binom{500}{i} (0.10)^i (0.90)^{500-i} = 0.9381745$$

50



6.5.4 Expected Value and Variance of the Uncorrected Sample Variance and the Sample Variance

Given a random sample $X_1, X_2, \ldots, X_n$ taken from a population with mean μ and variance σ^2, the expected value of the uncorrected variance, S_u^2, is

$$E\left[S_u^2\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[\left(X_i - \overline{X}\right)^2\right]. \qquad (6.13)$$

- **Note that**
- $S_u^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ **and** $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- **Therefore:**

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

51

Expanding the right hand side of (6.13) gives

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

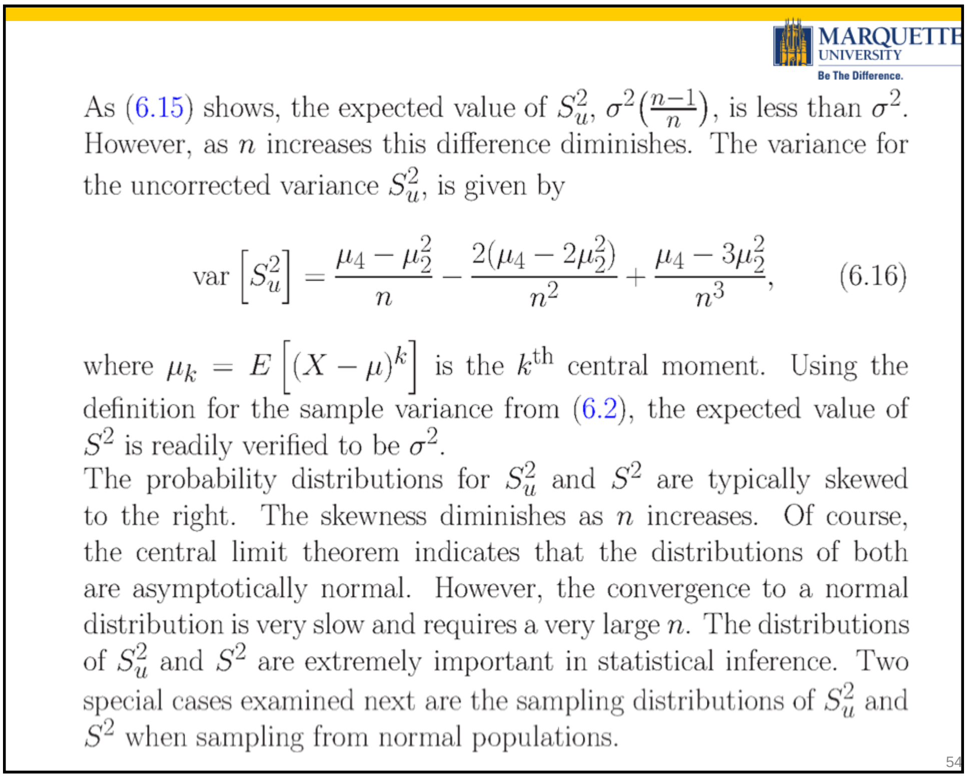
- **Remember that:** $S_u^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$

Substituting $\sum_{i=1}^{n}(X_i-\mu)^2-n\left(\mu-\overline{X}\right)^2$ for $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$ in (6.13) gives

$$E\left[S_u^2\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}(X_i-\mu)^2-n\left(\mu-\overline{X}\right)^2\right]$$

$$E\left[S_u^2\right]=\frac{1}{n}\left(n\sigma^2-n\frac{\sigma^2}{n}\right) \qquad (6.15)$$

$$\begin{aligned}E\left[S_u^2\right]&=\sigma^2-\frac{\sigma^2}{n}\\ &=\sigma^2\left(\frac{n-1}{n}\right).\end{aligned}$$

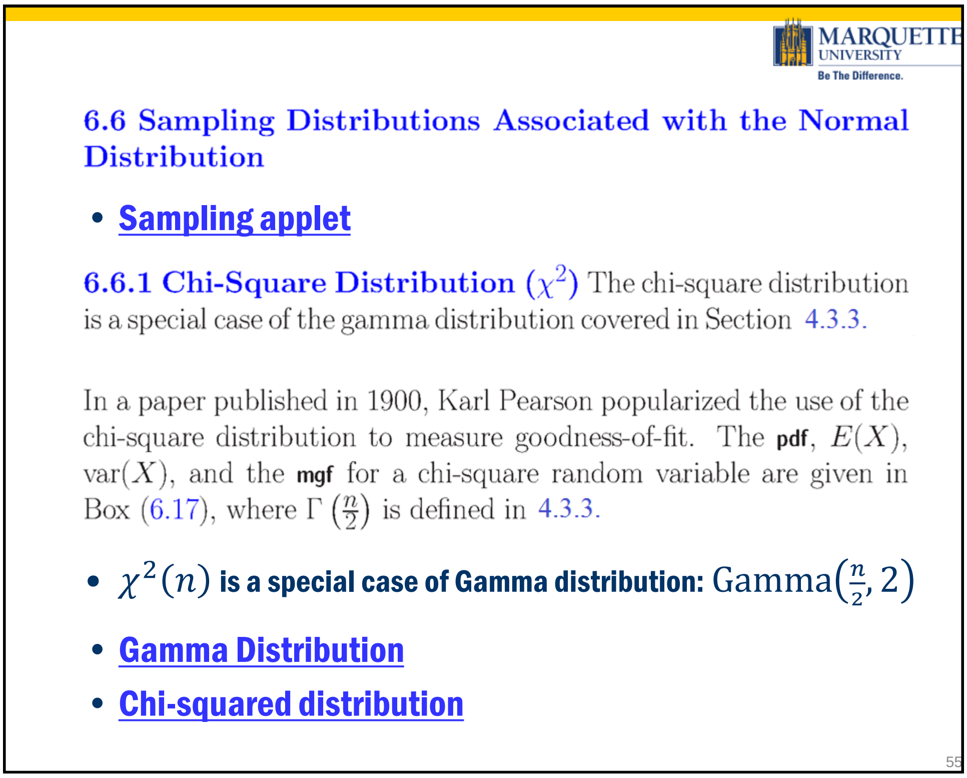


As (6.15) shows, the expected value of S_u^2, $\sigma^2\left(\frac{n-1}{n}\right)$, is less than σ^2. However, as n increases this difference diminishes. The variance for the uncorrected variance S_u^2, is given by

$$\mathrm{var}\left[S_u^2\right] = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}, \qquad (6.16)$$

where $\mu_k = E\left[(X-\mu)^k\right]$ is the k^{th} central moment. Using the definition for the sample variance from (6.2), the expected value of S^2 is readily verified to be σ^2.

The probability distributions for S_u^2 and S^2 are typically skewed to the right. The skewness diminishes as n increases. Of course, the central limit theorem indicates that the distributions of both are asymptotically normal. However, the convergence to a normal distribution is very slow and requires a very large n. The distributions of S_u^2 and S^2 are extremely important in statistical inference. Two special cases examined next are the sampling distributions of S_u^2 and S^2 when sampling from normal populations.



6.6 Sampling Distributions Associated with the Normal Distribution

- **Sampling applet**

6.6.1 Chi-Square Distribution (χ^2) The chi-square distribution is a special case of the gamma distribution covered in Section 4.3.3.

In a paper published in 1900, Karl Pearson popularized the use of the chi-square distribution to measure goodness-of-fit. The **pdf**, $E(X)$, $\mathrm{var}(X)$, and the **mgf** for a chi-square random variable are given in Box (6.17), where $\Gamma\left(\frac{n}{2}\right)$ is defined in 4.3.3.

- $\chi^2(n)$ **is a special case of Gamma distribution:** $\mathrm{Gamma}\left(\frac{n}{2}, 2\right)$
- **Gamma Distribution**
- **Chi-squared distribution**

$$\text{Chi-Square Distribution} \quad X \sim \chi^2_n$$

$$f(x) = \begin{cases} \dfrac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}} \cdot x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \qquad (6.17)$$

$$E[X] = n$$

$$\text{var}[X] = 2n$$

$$M_X(t) = (1-2t)^{-\frac{n}{2}} \text{ for } t < \frac{1}{2}$$

The chi-square distribution is strictly dependent on the parameter, n, called the **degrees of freedom**. In general, the chi-square distribution is unimodal and skewed to the right. Three different chi-square distributions are represented in Figure 6.3 on page 112.

$$\Gamma\left(\tfrac{n}{2}\right) = \int_0^\infty x^{n/2-1} e^{-x} dx$$

56

The notation used with the chi-square distribution to indicate α of the distribution is in the left tail when the distribution has n degrees of freedom is $\chi^2_{\alpha;n}$. For example, $\chi^2_{0.95;10}$ denotes the value such that 95% of the area is to the left of said value in a χ^2_{10} distribution. To find the value corresponding to $\chi^2_{0.95;10}$, use the S command `qchisq(p,df)` where `p` is the area to the left (probability) and `df` is the degrees of freedom. The command

```
> qchisq(.95,10)
[1] 18.30704
```

which says that $\mathbb{P}(\chi^2_{10} < 18.31) = 0.95$

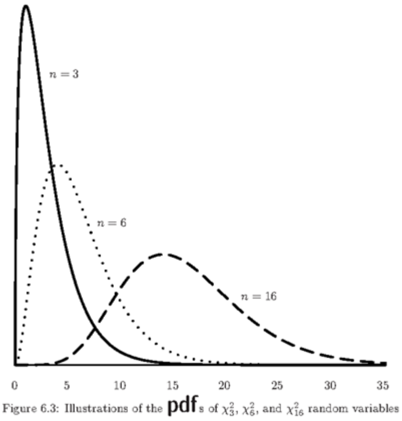
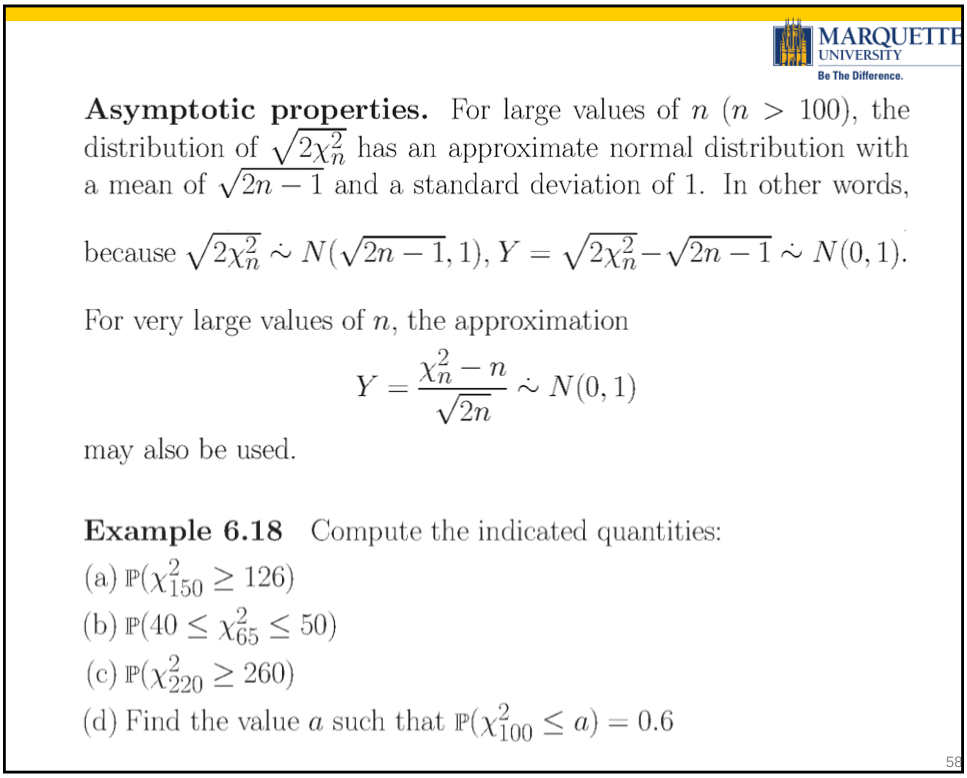


Figure 6.3: Illustrations of the pdfs of χ^2_3, χ^2_6, and χ^2_{16} random variables

57



Asymptotic properties. For large values of n ($n > 100$), the distribution of $\sqrt{2\chi_n^2}$ has an approximate normal distribution with a mean of $\sqrt{2n-1}$ and a standard deviation of 1. In other words,

because $\sqrt{2\chi_n^2} \dot\sim N(\sqrt{2n-1}, 1)$, $Y = \sqrt{2\chi_n^2} - \sqrt{2n-1} \dot\sim N(0,1)$.

For very large values of n, the approximation

$$Y = \frac{\chi_n^2 - n}{\sqrt{2n}} \dot\sim N(0,1)$$

may also be used.

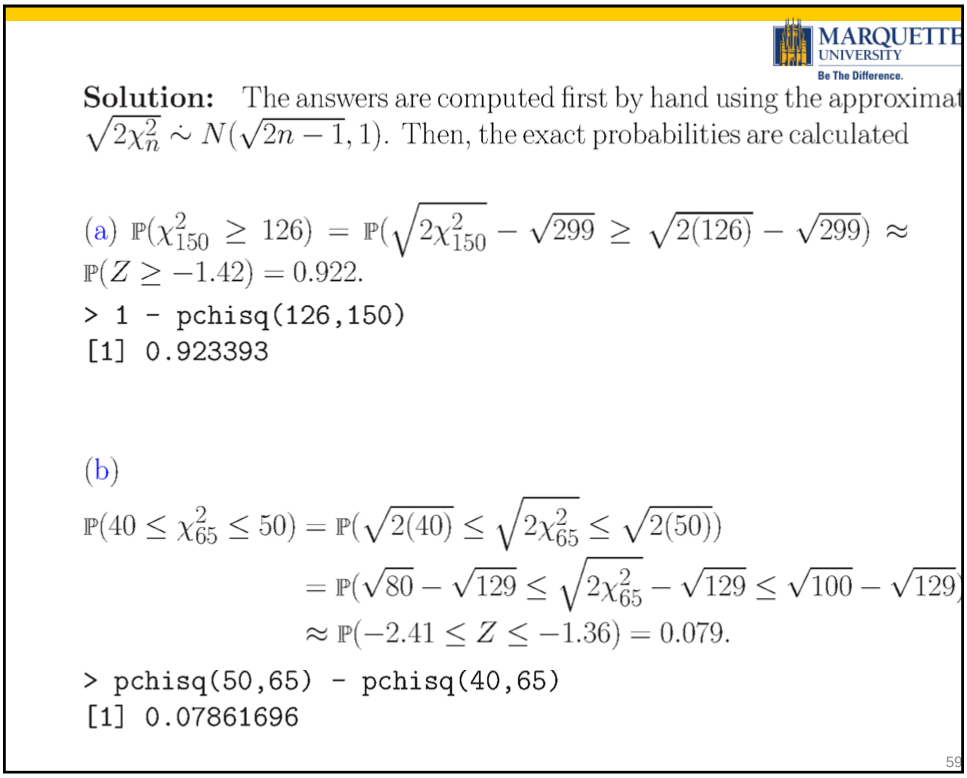
Example 6.18 Compute the indicated quantities:

(a) $\mathbb{P}(\chi_{150}^2 \geq 126)$

(b) $\mathbb{P}(40 \leq \chi_{65}^2 \leq 50)$

(c) $\mathbb{P}(\chi_{220}^2 \geq 260)$

(d) Find the value a such that $\mathbb{P}(\chi_{100}^2 \leq a) = 0.6$



Solution: The answers are computed first by hand using the approximat $\sqrt{2\chi_n^2} \dot\sim N(\sqrt{2n-1}, 1)$. Then, the exact probabilities are calculated

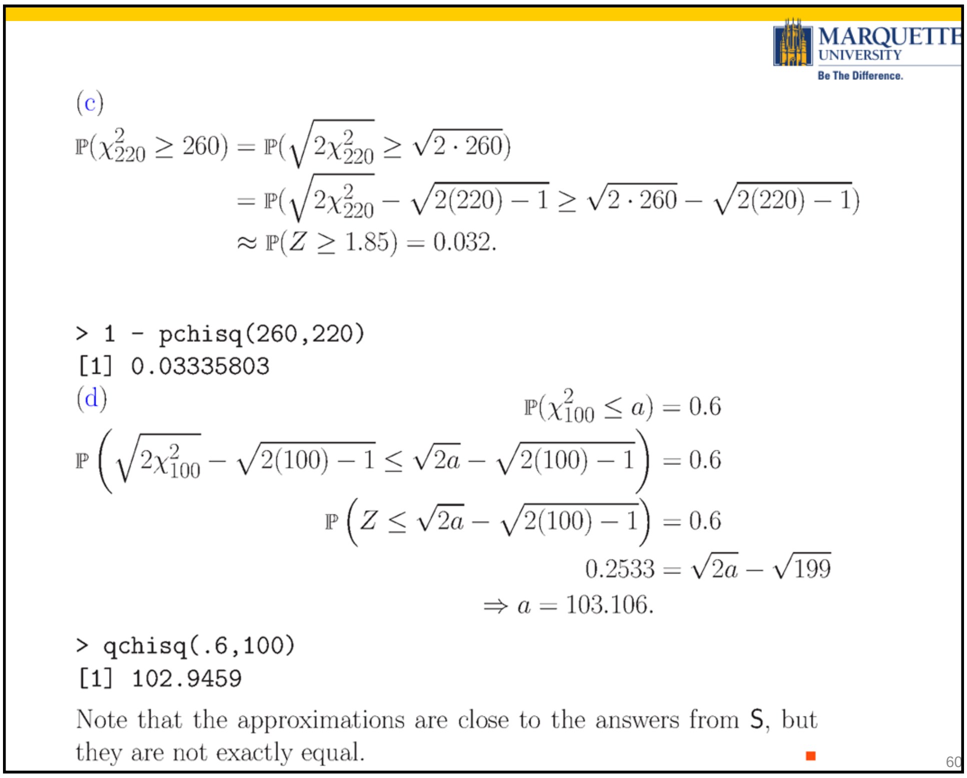
(a) $\mathbb{P}(\chi_{150}^2 \geq 126) = \mathbb{P}(\sqrt{2\chi_{150}^2} - \sqrt{299} \geq \sqrt{2(126)} - \sqrt{299}) \approx \mathbb{P}(Z \geq -1.42) = 0.922.$

```
> 1 - pchisq(126,150)
[1] 0.923393
```

(b)

$$\begin{aligned}\mathbb{P}(40 \leq \chi_{65}^2 \leq 50) &= \mathbb{P}(\sqrt{2(40)} \leq \sqrt{2\chi_{65}^2} \leq \sqrt{2(50)}) \\ &= \mathbb{P}(\sqrt{80} - \sqrt{129} \leq \sqrt{2\chi_{65}^2} - \sqrt{129} \leq \sqrt{100} - \sqrt{129}) \\ &\approx \mathbb{P}(-2.41 \leq Z \leq -1.36) = 0.079.\end{aligned}$$

```
> pchisq(50,65) - pchisq(40,65)
[1] 0.07861696
```



(c)

$$\begin{aligned}
\mathbb{P}(\chi^2_{220} \ge 260) &= \mathbb{P}(\sqrt{2\chi^2_{220}} \ge \sqrt{2\cdot 260}) \\
&= \mathbb{P}(\sqrt{2\chi^2_{220}} - \sqrt{2(220)-1} \ge \sqrt{2\cdot 260} - \sqrt{2(220)-1}) \\
&\approx \mathbb{P}(Z \ge 1.85) = 0.032.
\end{aligned}$$

```
> 1 - pchisq(260,220)
[1] 0.03335803
```

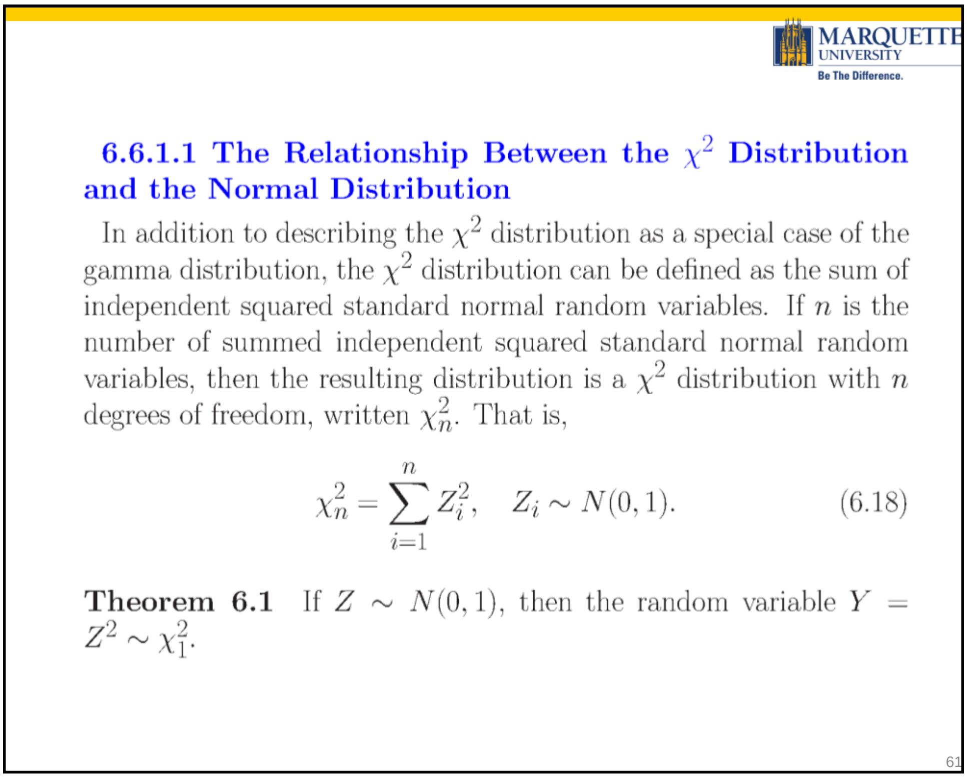
(d)

$$\begin{aligned}
\mathbb{P}(\chi^2_{100} \le a) &= 0.6 \\
\mathbb{P}\left(\sqrt{2\chi^2_{100}} - \sqrt{2(100)-1} \le \sqrt{2a} - \sqrt{2(100)-1}\right) &= 0.6 \\
\mathbb{P}\left(Z \le \sqrt{2a} - \sqrt{2(100)-1}\right) &= 0.6 \\
0.2533 &= \sqrt{2a} - \sqrt{199} \\
\Rightarrow a &= 103.106.
\end{aligned}$$

```
> qchisq(.6,100)
[1] 102.9459
```

Note that the approximations are close to the answers from S, but they are not exactly equal. ■

60



6.6.1.1 The Relationship Between the χ^2 Distribution and the Normal Distribution

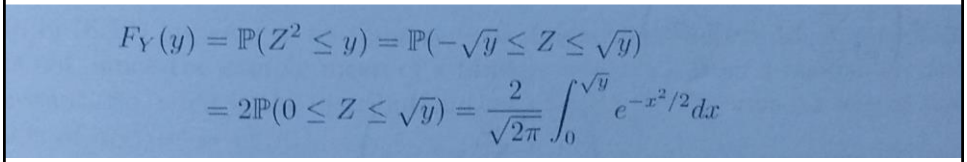
In addition to describing the χ^2 distribution as a special case of the gamma distribution, the χ^2 distribution can be defined as the sum of independent squared standard normal random variables. If n is the number of summed independent squared standard normal random variables, then the resulting distribution is a χ^2 distribution with n degrees of freedom, written χ^2_n. That is,

$$\chi^2_n = \sum_{i=1}^{n} Z_i^2, \quad Z_i \sim N(0,1). \qquad (6.18)$$

Theorem 6.1 If $Z \sim N(0,1)$, then the random variable $Y = Z^2 \sim \chi^2_1$.

61

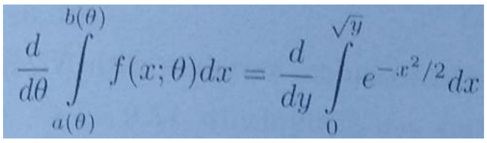
Theorem 6.1 If $Z \sim N(0,1)$, then the random variable $Y = Z^2 \sim \chi_1^2$.



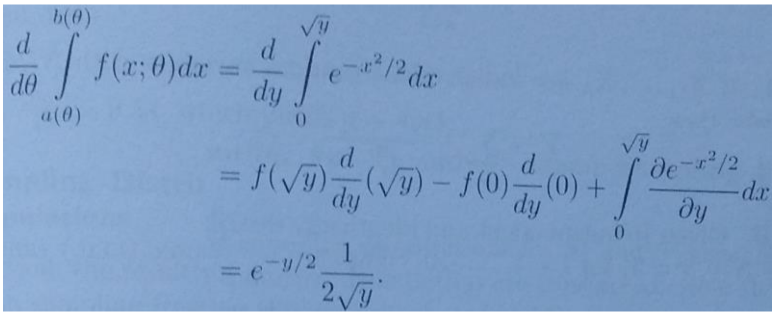
$$F_Y(y) = \mathbb{P}(Z^2 \le y) = \mathbb{P}(-\sqrt{y} \le Z \le \sqrt{y})$$
$$= 2\mathbb{P}(0 \le Z \le \sqrt{y}) = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-x^2/2} dx$$

- **Recall Leibniz integral rule:**

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\int_{a(\theta)}^{b(\theta)} f(x,\theta)\,\mathrm{d}x\right) = \int_{a(\theta)}^{b(\theta)} \partial_\theta f(x,\theta)\,\mathrm{d}x + f\big(b(\theta),\theta\big)\cdot b'(\theta) - f\big(a(\theta),\theta\big)\cdot a'(\theta)$$

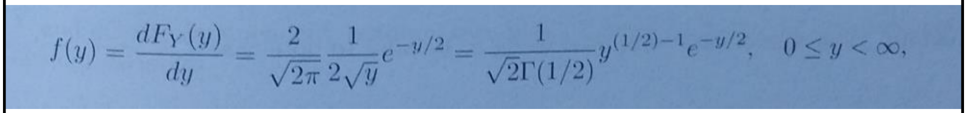


$$\frac{d}{d\theta}\int\limits_{a(\theta)}^{b(\theta)} f(x;\theta)dx = \frac{d}{dy}\int\limits_0^{\sqrt{y}} e^{-x^2/2}dx$$



$$\frac{d}{d\theta}\int\limits_{a(\theta)}^{b(\theta)} f(x;\theta)dx = \frac{d}{dy}\int\limits_0^{\sqrt{y}} e^{-x^2/2}dx$$
$$= f(\sqrt{y})\frac{d}{dy}(\sqrt{y}) - f(0)\frac{d}{dy}(0) + \int\limits_0^{\sqrt{y}} \frac{\partial e^{-x^2/2}}{\partial y}dx$$
$$= e^{-y/2}\frac{1}{2\sqrt{y}}.$$

- **Taking the derivative of $F_Y(y)$ yields**



$$f(y) = \frac{dF_Y(y)}{dy} = \frac{2}{\sqrt{2\pi}}\frac{1}{2\sqrt{y}}e^{-y/2} = \frac{1}{\sqrt{2}\Gamma(1/2)}y^{(1/2)-1}e^{-y/2}, \quad 0 \le y < \infty,$$

which is the pdf for χ_1^2

Change of Variable

Given a random variable x, with probability distribution function $f_X(x|\theta)$, we often would like to know the probability distribution of a random variable y, that is a function $y(x)$ of x, $y=y(x)$.

Change of Variable

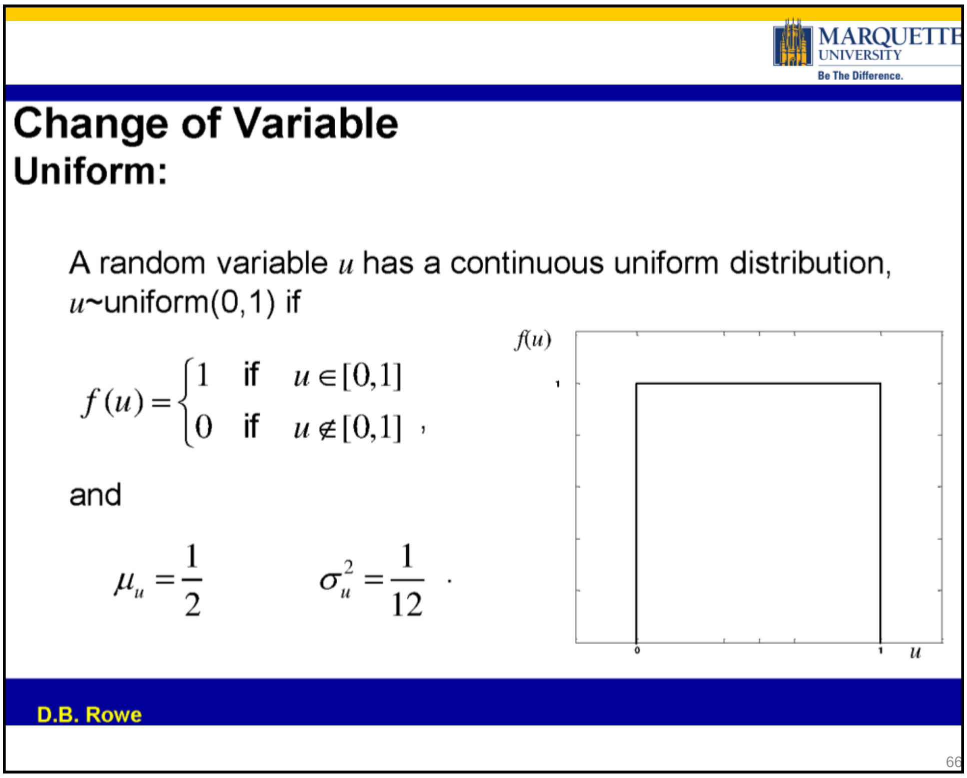
Let $y=y(x)$ be a one-to-one transformation with inverse transformation $x=x(y)$.

Then, if $f_X(x|\theta)$ is the PDF of x, the PDF of y can be found as

$$f_Y(y|\theta)=f_X(x(y)|\theta)\times|J(x\rightarrow y)|$$

where $J(x\rightarrow y)=\dfrac{dx(y)}{dy}$.

Suppress PDF subscripts.



Change of Variable

Uniform:

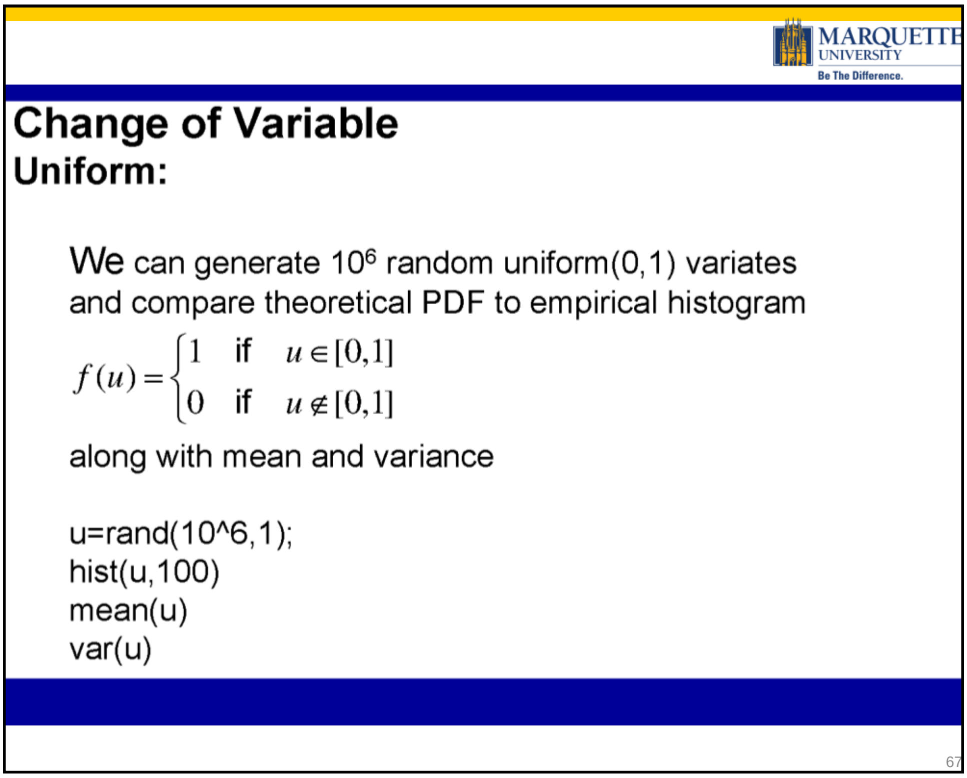
A random variable u has a continuous uniform distribution, u~uniform(0,1) if

$$f(u)=\begin{cases}1 & \text{if} \quad u\in[0,1] \\ 0 & \text{if} \quad u\notin[0,1]\end{cases},$$

and

$$\mu_u=\frac{1}{2} \qquad \sigma_u^2=\frac{1}{12}\ .$$

D.B. Rowe



Change of Variable

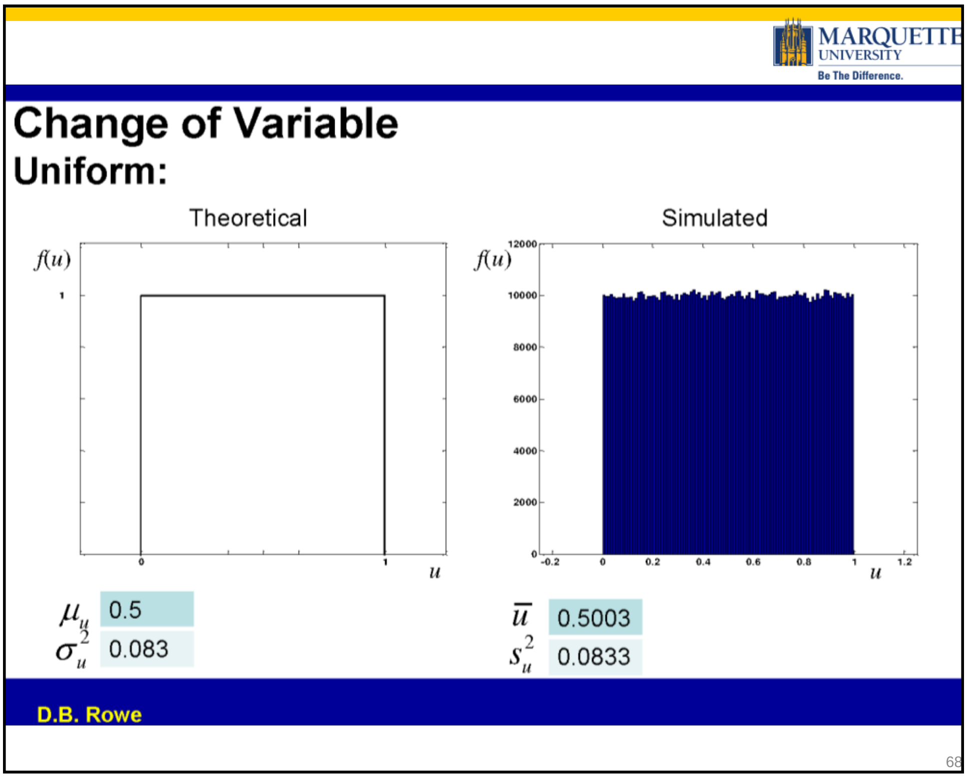
Uniform:

We can generate 10^6 random uniform(0,1) variates and compare theoretical PDF to empirical histogram

$$f(u)=\begin{cases}1 & \text{if} \quad u\in[0,1] \\ 0 & \text{if} \quad u\notin[0,1]\end{cases}$$

along with mean and variance

```
u=rand(10^6,1);
hist(u,100)
mean(u)
var(u)
```



Change of Variable

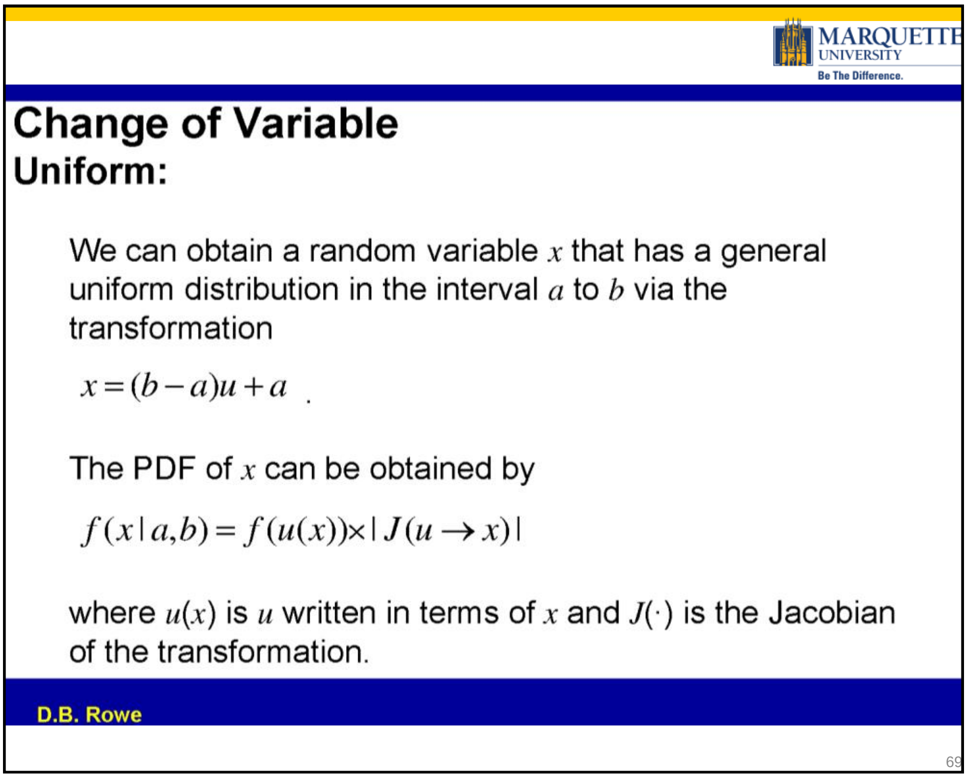
Uniform:

μ_u	0.5
σ_u^2	0.083

$\bar{u}$	0.5003
s_u^2	0.0833

D.B. Rowe

68



Change of Variable

Uniform:

We can obtain a random variable x that has a general uniform distribution in the interval a to b via the transformation

$$x = (b-a)u + a \quad .$$

The PDF of x can be obtained by

$$f(x \mid a,b) = f(u(x)) \times \mid J(u \to x) \mid$$

where $u(x)$ is u written in terms of x and $J(\cdot)$ is the Jacobian of the transformation.

D.B. Rowe

69

Change of Variable

Uniform:

The original variable u in terms of the new variable is

$$u(x) = \frac{x-a}{b-a}$$

and the Jacobian of the transformation is

$$J(u \to x) = \frac{du(x)}{dx} = \frac{1}{b-a} \quad .$$

This yields

$$f(x \mid a,b) = f(u(x)) \times | J(u \to x) | = 1 \times \left| \frac{1}{b-a} \right| .$$

Change of Variable

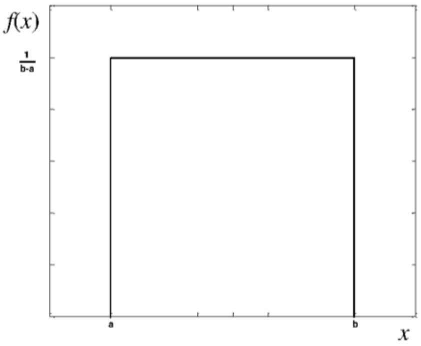
Uniform:

A random variable x has a continuous uniform distribution, x~uniform(a,b) if

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if} \quad x \in [a,b] \\ 0 & \text{if} \quad x \notin [a,b] \end{cases} ,$$

where, $a,b \in \mathbb{R}$, $a < b$.

Note that u=0 mapped to x=a and u=1 mapped to x=b.



Change of Variable
Uniform:

We can generate 10^6 random uniform(a,b) variates and compare theoretical PDF to empirical histogram

$$f(x)=\begin{cases}\dfrac{1}{b-a} & \text{if} \quad x\in[a,b]\\ 0 & \text{if} \quad x\notin[a,b]\end{cases}$$

along with mean & variance by transforming random variates

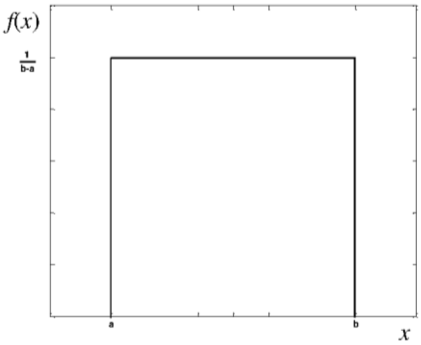
```
a=1;,b=2;
x=a+(b-a)*u;
hist(x,100)
mean(x), var(x)
```

10^6 uniform variates
u=rand(10^6,1);

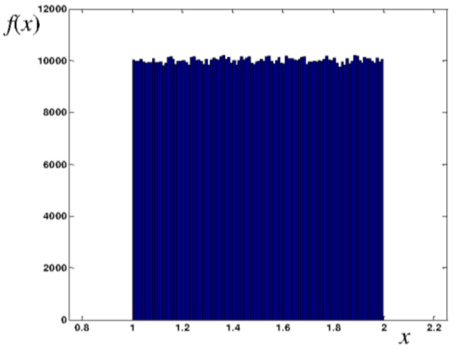
Change of Variable
Uniform:

$$a=1 \qquad \mu_x=\frac{b+a}{2} \qquad \sigma_x^2=\frac{(b-a)^2}{12}$$
$$b=2$$

Theoretical



Simulated



μ_x	1.5
σ_x^2	0.083

$\bar{x}$	1.5003
s_x^2	0.0833

Change of Variable

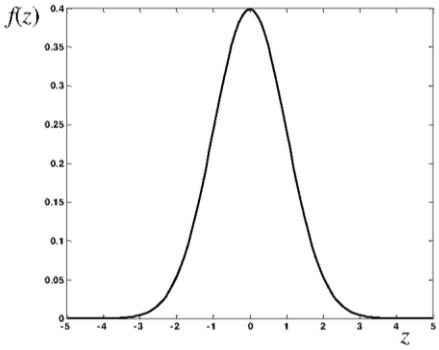
Normal: The same process can be applied.

A random variable z has a standard normal distribution, z~normal(0,1) if

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2} \quad ,$$

where $z\in\mathbb{R}$ and

$$\mu_z=0 \qquad \sigma_z^2=1 \ .$$



Change of Variable

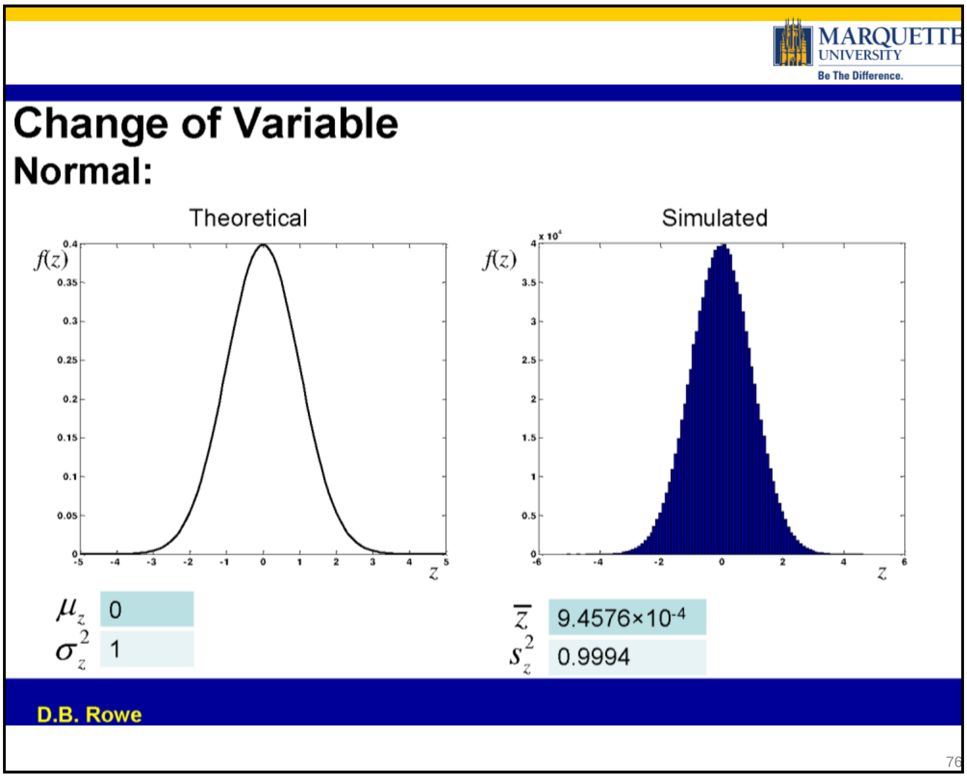
Normal:

We can generate 10^6 random normal(0,1) variates and compare theoretical PDF to empirical histogram

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$$

along with mean and variance

```
z=randn(10^6,1);
hist(z,(-5:.1:5))
mean(z), var(z)
xlim([-5 5])
```



MARQUETTE UNIVERSITY Be The Difference.

Change of Variable

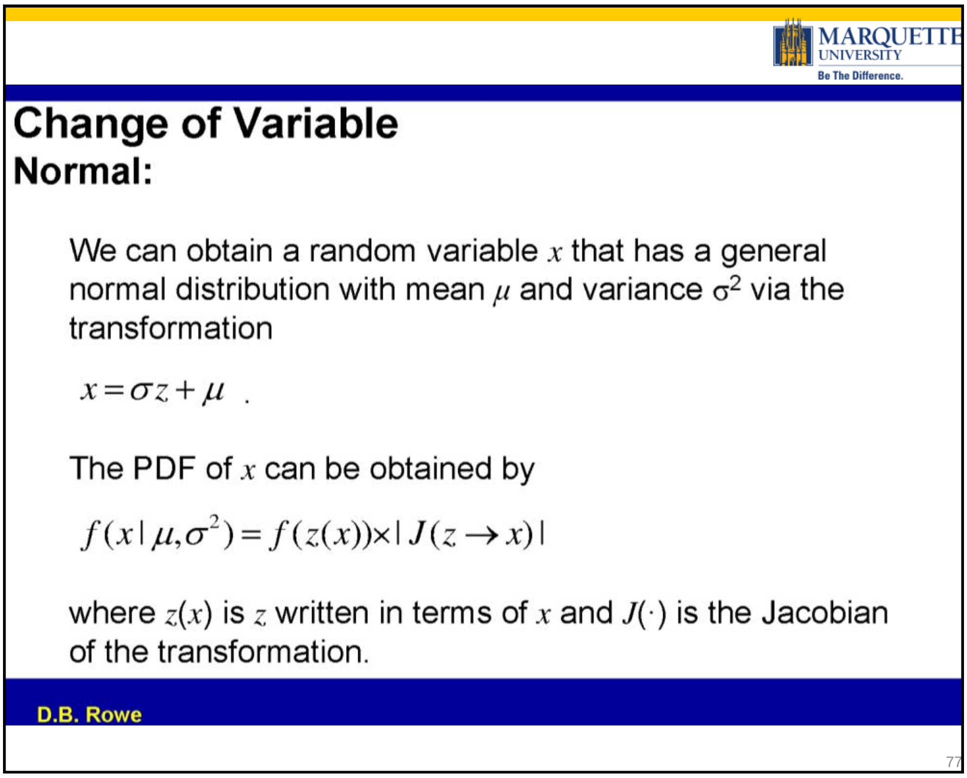
Normal:

μ_z	0
σ_z^2	1

$\bar{z}$	9.4576×10^{-4}
s_z^2	0.9994

D.B. Rowe

76



MARQUETTE UNIVERSITY Be The Difference.

Change of Variable

Normal:

We can obtain a random variable x that has a general normal distribution with mean μ and variance σ^2 via the transformation

$$x = \sigma z + \mu \quad .$$

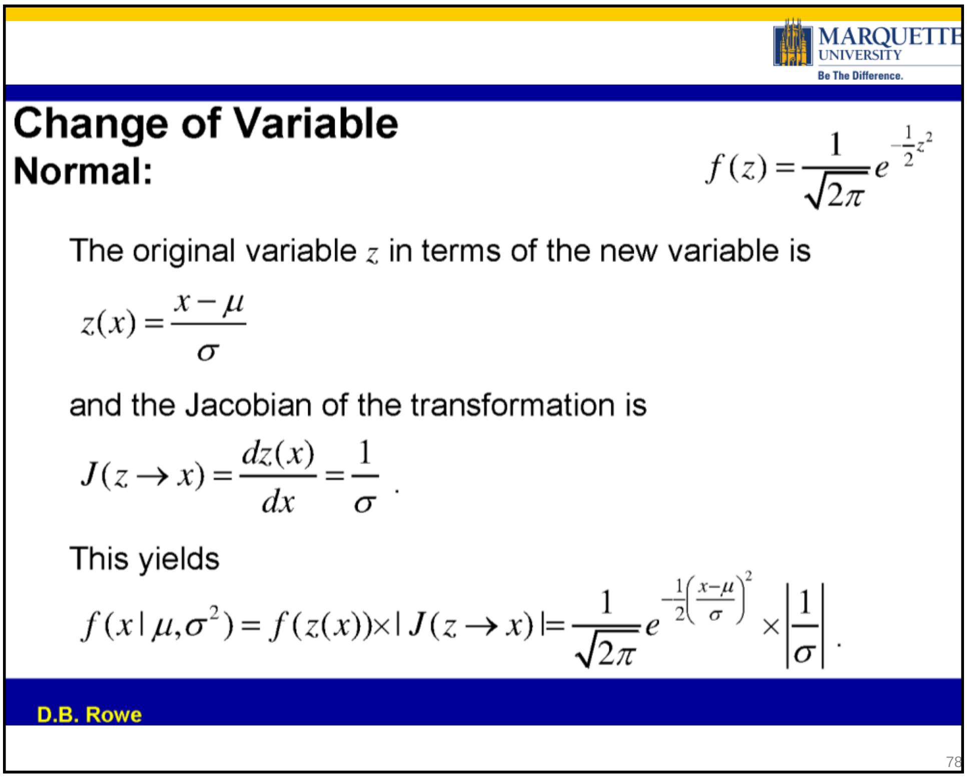
The PDF of x can be obtained by

$$f(x \mid \mu, \sigma^2) = f(z(x)) \times | J(z \rightarrow x) |$$

where $z(x)$ is z written in terms of x and $J(\cdot)$ is the Jacobian of the transformation.

D.B. Rowe

77



MARQUETTE UNIVERSITY
Be The Difference.

Change of Variable
Normal:

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$$

The original variable z in terms of the new variable is

$$z(x)=\frac{x-\mu}{\sigma}$$

and the Jacobian of the transformation is

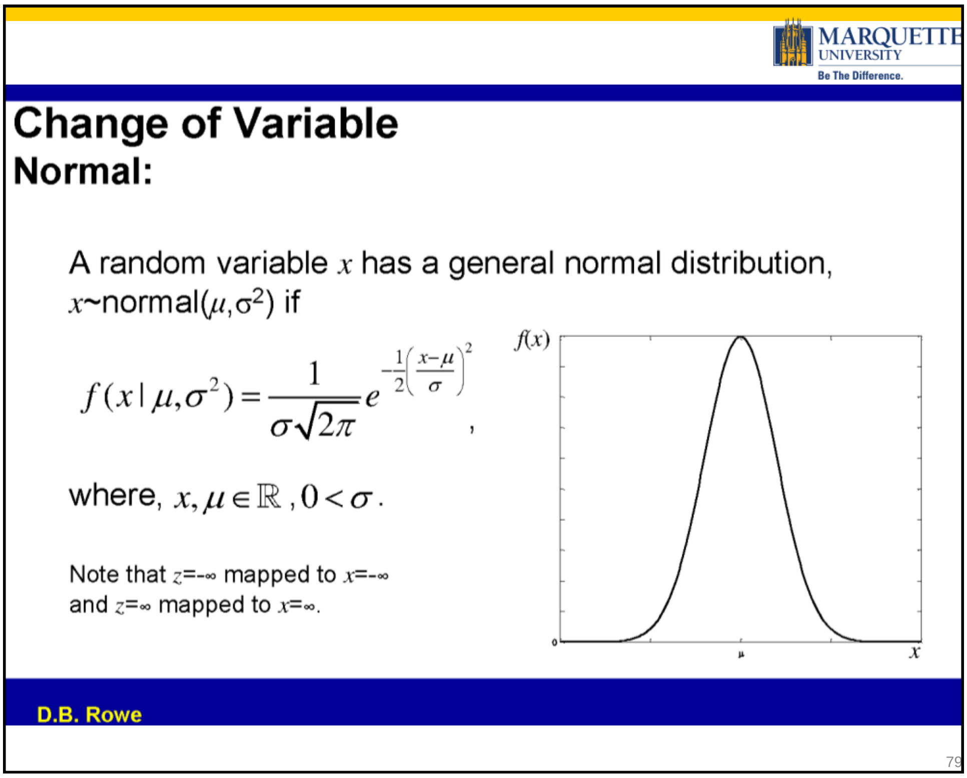
$$J(z\rightarrow x)=\frac{dz(x)}{dx}=\frac{1}{\sigma}\ .$$

This yields

$$f(x\,|\,\mu,\sigma^2)=f(z(x))\times|\,J(z\rightarrow x)|=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\times\left|\frac{1}{\sigma}\right|\ .$$

D.B. Rowe

78



MARQUETTE UNIVERSITY
Be The Difference.

Change of Variable
Normal:

A random variable x has a general normal distribution, x~normal(μ,σ^2) if

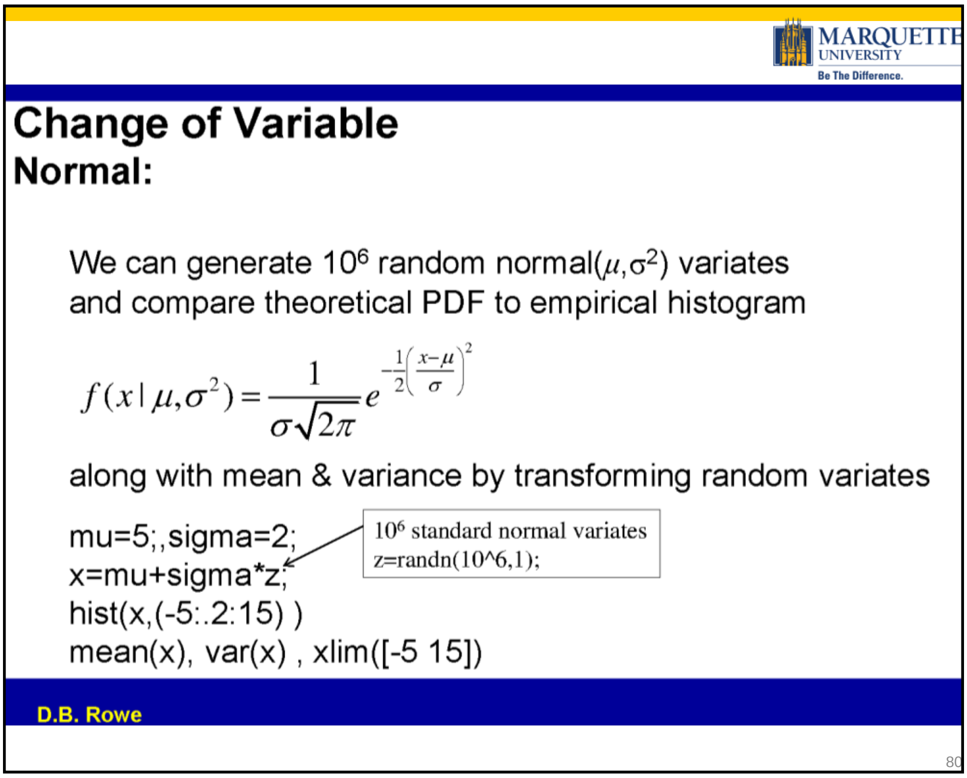
$$f(x\,|\,\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\ ,$$

where, $x,\mu\in\mathbb{R}\ ,0<\sigma$.

Note that z=-∞ mapped to x=-∞
and z=∞ mapped to x=∞.

D.B. Rowe

79



MARQUETTE UNIVERSITY Be The Difference.

Change of Variable

Normal:

We can generate 10^6 random normal(μ,σ^2) variates and compare theoretical PDF to empirical histogram

$$f(x\mid\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

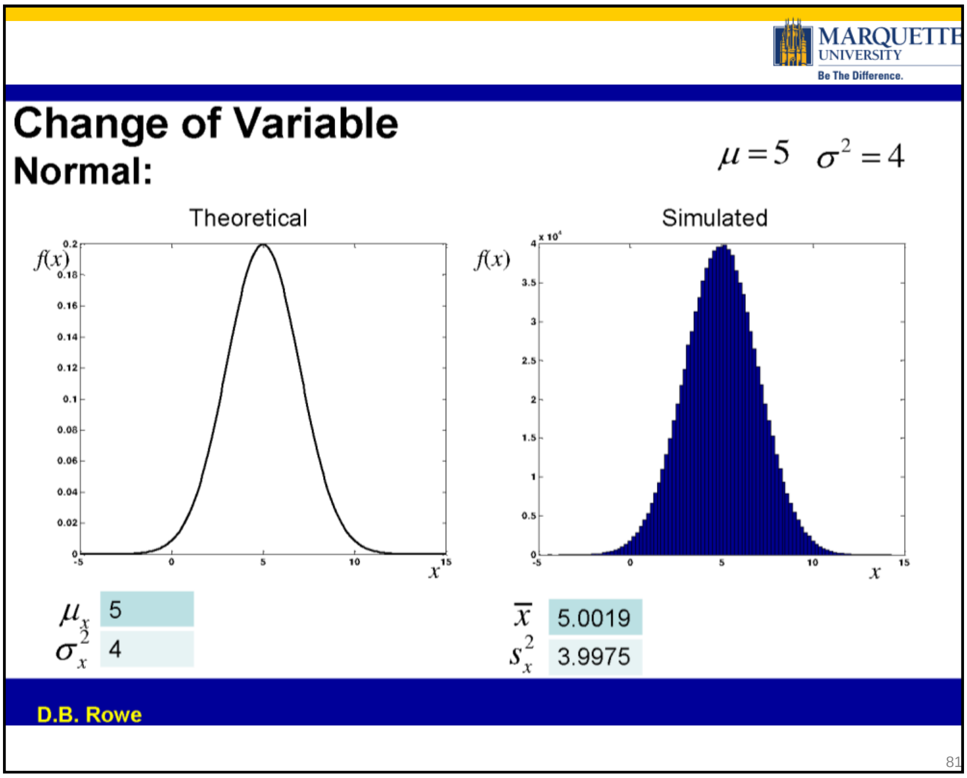
along with mean & variance by transforming random variates

```
mu=5;,sigma=2;
x=mu+sigma*z;
hist(x,(-5:.2:15) )
mean(x), var(x) , xlim([-5 15])
```

10^6 standard normal variates
z=randn(10^6,1);

D.B. Rowe

80



MARQUETTE UNIVERSITY Be The Difference.

Change of Variable

Normal:

$\mu=5 \quad \sigma^2=4$

Theoretical

Simulated

μ_x 5
σ_x^2 4

$\bar{x}$ 5.0019
s_x^2 3.9975

D.B. Rowe

81

Change of Variable

This process can be used to find the distribution of more than linear functions $y=y(x)$ of random variables.

For example, let x~normal(μ,σ^2).

Assume we want to know the distribution of $y=\left(\frac{x-\mu}{\sigma}\right)^2$.

We can determine $f(y)$ through the transformation of variable procedure.

$$f_Y(y\,|\,\theta)=f_X(x(y)\,|\,\theta)\times|\,J(x\rightarrow y)\,|$$

Homework problem.

CHANGE OF VARIABLE
NOT ONE-TO-ONE

Theorem 6.1 If $Z \sim N(0,1)$, then the random variable $Y = Z^2 \sim \chi^2_1$.

Let $y=y(z)$ be a not one-to-one transformation,
(i.e. $y=z^2$, then $z_1(y)=+\sqrt{y}$ and $z_2(y)=-\sqrt{y}$.)

We can still perform the change of variable by breaking up the transformation into pieces that are1-to-1.

$$f_Y(y\,|\,\theta)=\sum_j f_Z(z_j(y)\,|\,\theta)\times\left|\frac{dz_j(y)}{dy}\right|$$

i.e. $f_Y(y\,|\,\theta)=f_Z(\sqrt{y}\,|\,\theta)\left|\frac{1}{2\sqrt{y}}\right|+f_Z(-\sqrt{y}\,|\,\theta)\left|\frac{-1}{2\sqrt{y}}\right|$

$$f_Y(y)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(\sqrt{y})^2}{2}}\times\frac{1}{2\sqrt{y}}+\frac{1}{\sqrt{2\pi}}e^{-\frac{(-\sqrt{y})^2}{2}}\times\frac{1}{2\sqrt{y}}$$

$$=\frac{1}{\sqrt{2}}\times\frac{1}{\sqrt{\pi}}\times\frac{1}{\sqrt{y}}\times e^{-\frac{y}{2}} \quad = \quad \frac{1}{\sqrt{2}\Gamma(1/2)}y^{(1/2)-1}e^{-\frac{y}{2}}, \qquad y>0. \blacksquare$$

Theorem 6.1 If $Z \sim N(0,1)$, then the random variable $Y = Z^2 \sim \chi^2_1$.

Corollary 6.1 If $X \sim N(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$, and $Z^2 \sim \chi^2_1$.

Theorem 6.2 If $X_1, \ldots, X_r$ are independent random variables with chi-square distributions $\chi^2_{n_1}, \ldots, \chi^2_{n_r}$ respectively, then

$$Y = \sum_{i=1}^{r} X_i \sim \chi^2_s, \quad \text{where} \quad s = \sum_{i=1}^{r} n_i.$$

- **Proof:**

$$M_Y(t) = E[e^{tY}] = \prod_{i=1}^{r} E[e^{tX_i}] = \prod_{i=1}^{r} M_{X_i}(t)$$

$$= \prod_{i=1}^{r} (1-2t)^{-\frac{n_i}{2}} = (1-2t)^{-\frac{1}{2}\sum_{i=1}^{r} n_i}$$

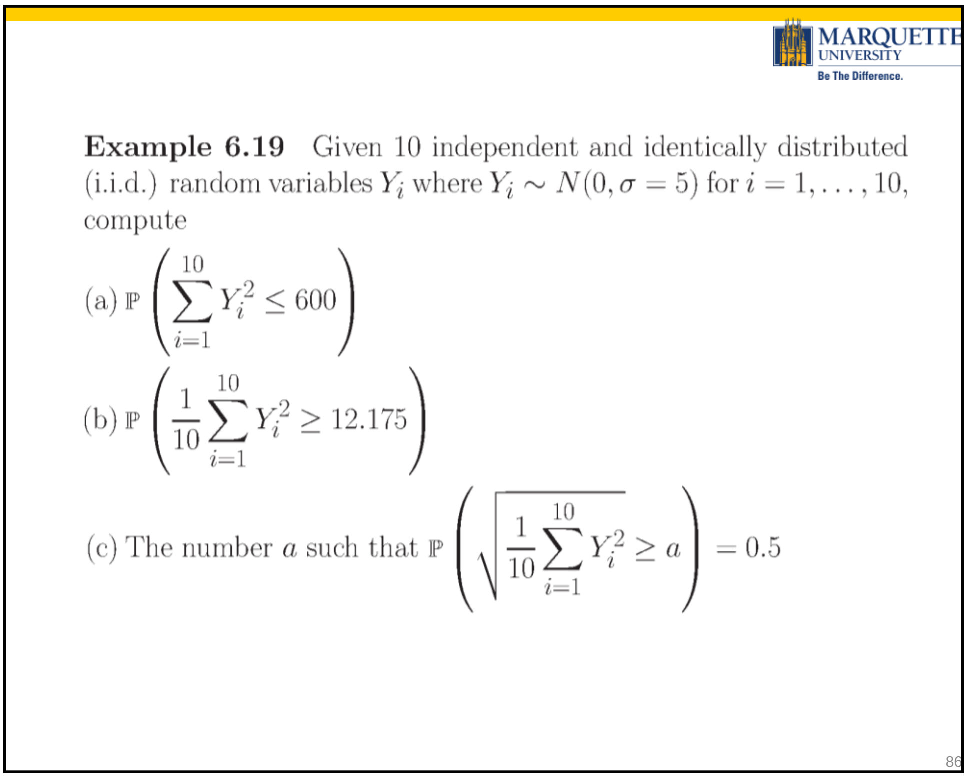
One of the properties of χ^2 distributions is that of reproducibility. In other words, the sum of independent χ^2 random variables is also a χ^2 distribution with degrees of freedom equal to the sum of the degrees of freedom of each of the independent χ^2 random variables.

Corollary 6.2 If $X_1, \ldots, X_n$ are independent random variables following a $N(0,1)$ distribution, then

$$Y = \sum_{i=1}^{n} X_i^2 \sim \chi^2_n.$$

Corollary 6.3 If $X_1, \ldots, X_n$ are independent random variables with $N(\mu_i, \sigma_i)$ distributions respectively, then

$$Y = \sum_{i=1}^{n} \frac{(X_i - \mu_i)^2}{\sigma_i^2} \sim \chi^2_n.$$

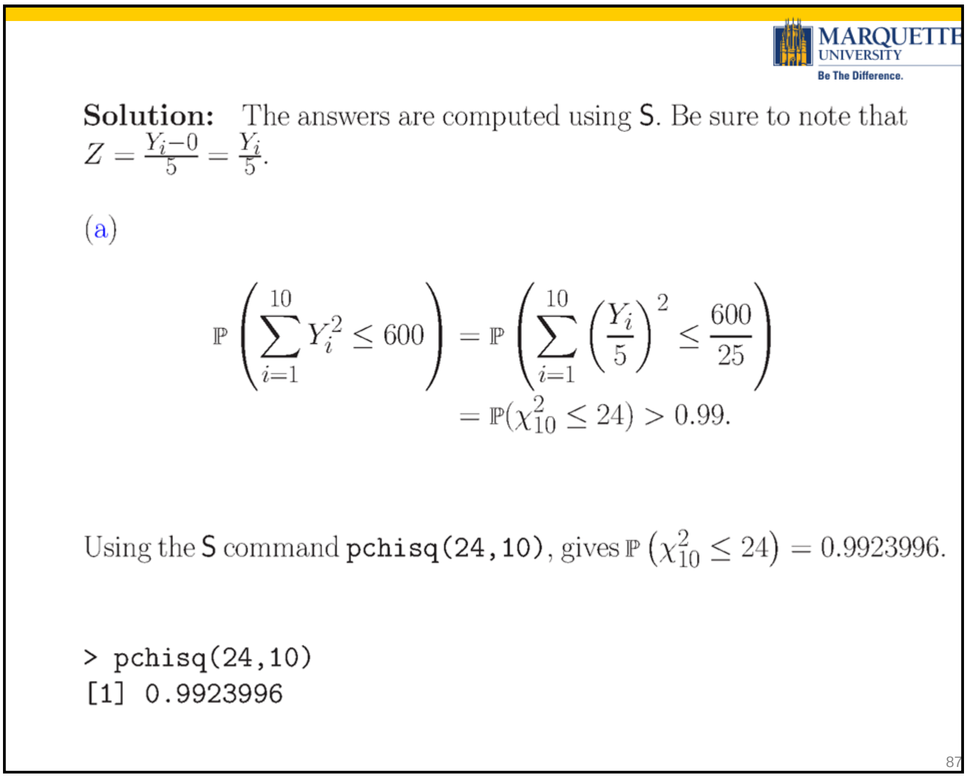


Example 6.19 Given 10 independent and identically distributed (i.i.d.) random variables Y_i where $Y_i \sim N(0, \sigma = 5)$ for $i = 1, \ldots, 10$, compute

(a) $\mathbb{P}\left(\sum_{i=1}^{10} Y_i^2 \leq 600\right)$

(b) $\mathbb{P}\left(\frac{1}{10}\sum_{i=1}^{10} Y_i^2 \geq 12.175\right)$

(c) The number a such that $\mathbb{P}\left(\sqrt{\frac{1}{10}\sum_{i=1}^{10} Y_i^2} \geq a\right) = 0.5$



Solution: The answers are computed using S. Be sure to note that $Z = \frac{Y_i - 0}{5} = \frac{Y_i}{5}$.

(a)

$$\mathbb{P}\left(\sum_{i=1}^{10} Y_i^2 \leq 600\right) = \mathbb{P}\left(\sum_{i=1}^{10}\left(\frac{Y_i}{5}\right)^2 \leq \frac{600}{25}\right)$$
$$= \mathbb{P}(\chi^2_{10} \leq 24) > 0.99.$$

Using the S command `pchisq(24,10)`, gives $\mathbb{P}\left(\chi^2_{10} \leq 24\right) = 0.9923996$.

```
> pchisq(24,10)
[1] 0.9923996
```

(b)

$$\mathbb{P}\left(\frac{1}{10}\sum_{i=1}^{10} Y_i^2 \geq 12.175\right) = \mathbb{P}\left(\sum_{i=1}^{10}\left(\frac{Y_i}{5}\right)^2 \geq \frac{12.175(10)}{25}\right)$$

$$= \mathbb{P}(\chi^2_{10} \geq 4.87) = 0.90.$$

```
> 1 - pchisq(4.87,10)
[1] 0.8996911
```

(c)

$$\mathbb{P}\left(\frac{1}{10}\sum_{i=1}^{10} Y_i^2 \geq a^2\right) = \mathbb{P}\left(\sum_{i=1}^{10}\left(\frac{Y_i}{5}\right)^2 \geq \frac{10a^2}{25}\right)$$

$$= \mathbb{P}\left(\chi^2_{10} \geq \frac{10a^2}{25}\right) = 0.5$$

Using the S command `qchisq()`, the value $\chi^2_{10,0.50} = 9.34$ is calculated.

```
> qchisq(0.50;10)
[1] 9.341818
```

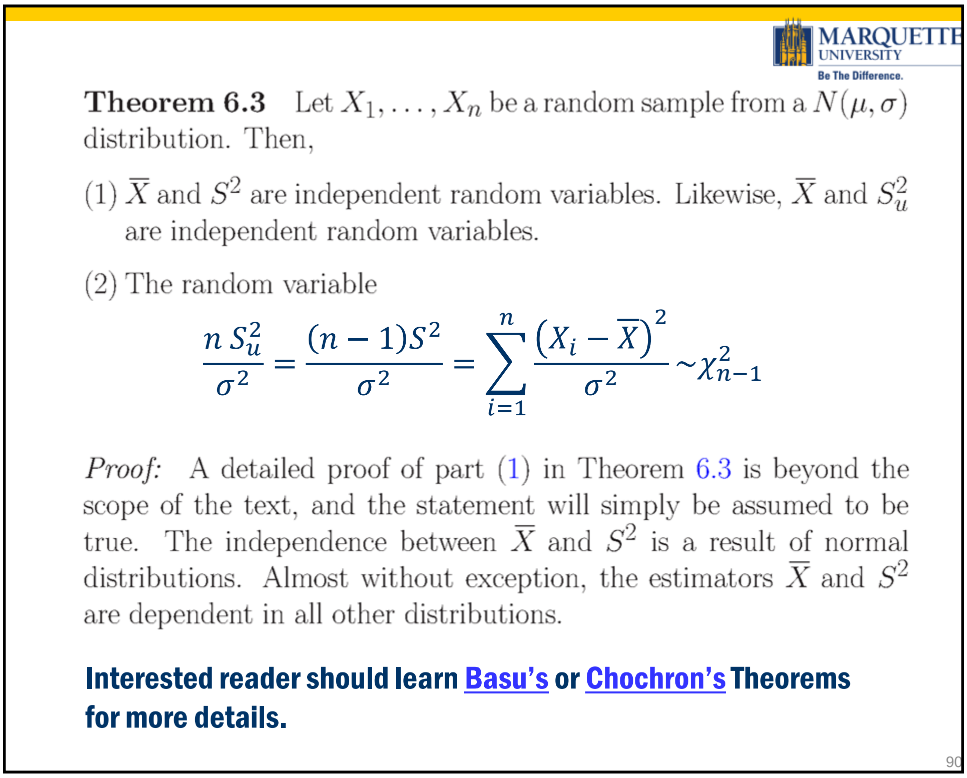
Consequently, $\frac{10a^2}{25} = 9.34$, which yields $a = 4.83$. ■

6.6.1.2 Sampling Distribution for S_u^2 and S^2 when Sampling from Normal Populations

In this section, the resulting sampling distributions for S_u^2 and S^2 given in Table 6.2 on page 44 when sampling from a normal distribution are considered. Note that $\sum_{i=1}^{n}(X_i - \overline{X})^2 = nS_u^2 = (n-1)S^2$ and that dividing this by σ^2 yields

$$\sum_{i=1}^{n}\frac{\left(X_i - \overline{X}\right)^2}{\sigma^2} = \frac{nS_u^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \tag{6.19}$$

The first term in (6.19) appears to be some type of standardized normal random variable. However, it is not since the sample mean of a random variable is itself a random variable and not a constant. So, what is the distribution then of nS_u^2/σ^2? Theorem 6.3 on the next page tells us that the distribution of nS_u^2/σ^2 is χ^2_{n-1}.



Theorem 6.3 Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma)$ distribution. Then,

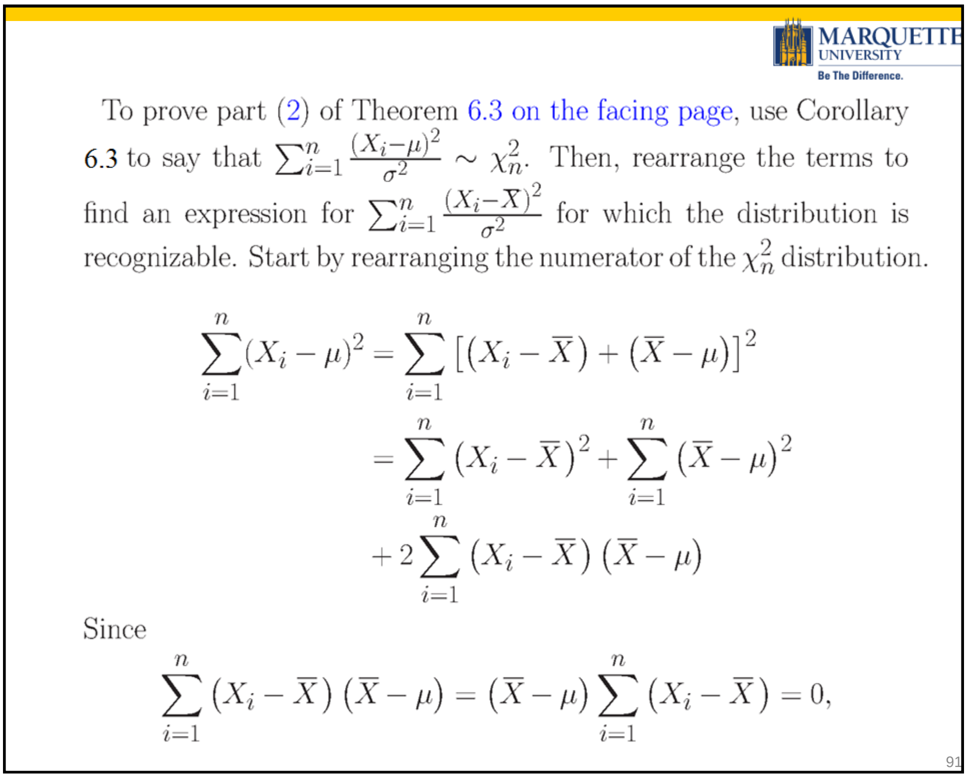
(1) $\overline{X}$ and S^2 are independent random variables. Likewise, $\overline{X}$ and S_u^2 are independent random variables.

(2) The random variable

$$\frac{n\,S_u^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2_{n-1}$$

Proof: A detailed proof of part (1) in Theorem 6.3 is beyond the scope of the text, and the statement will simply be assumed to be true. The independence between $\overline{X}$ and S^2 is a result of normal distributions. Almost without exception, the estimators $\overline{X}$ and S^2 are dependent in all other distributions.

Interested reader should learn Basu's or Chochron's Theorems for more details.

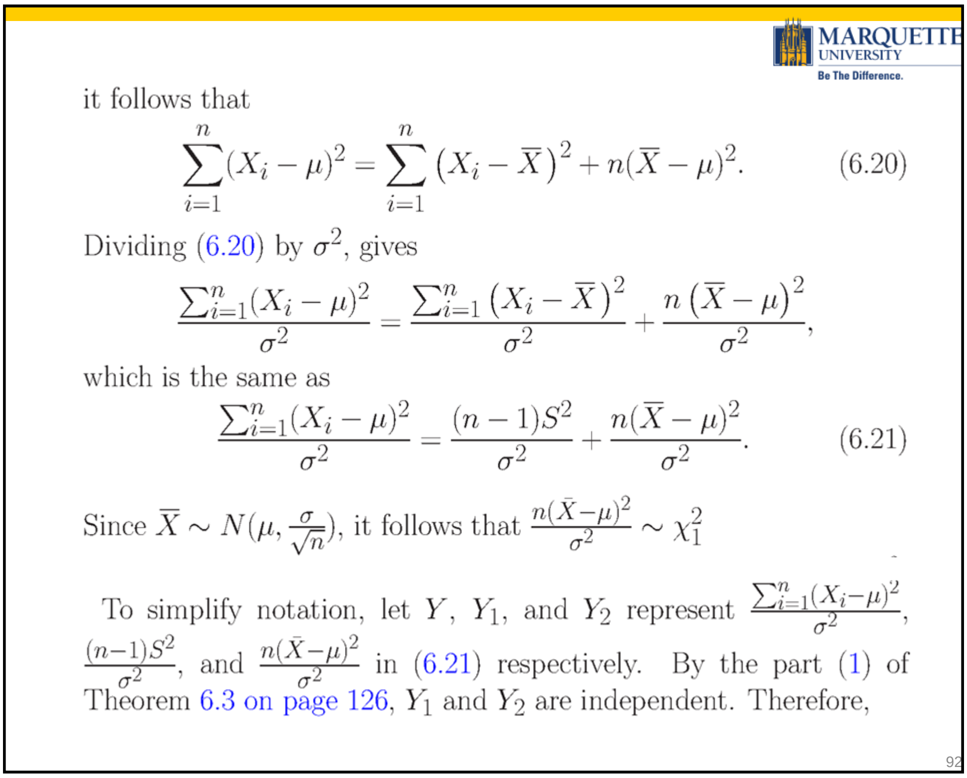


To prove part (2) of Theorem 6.3 on the facing page, use Corollary 6.3 to say that $\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2_n$. Then, rearrange the terms to find an expression for $\sum_{i=1}^{n} \frac{(X_i-\overline{X})^2}{\sigma^2}$ for which the distribution is recognizable. Start by rearranging the numerator of the χ^2_n distribution.

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

Since

$$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(\overline{X}-\mu\right) = \left(\overline{X}-\mu\right)\sum_{i=1}^{n}\left(X_i-\overline{X}\right) = 0,$$



it follows that

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+n(\overline{X}-\mu)^2. \qquad (6.20)$$

Dividing (6.20) by σ^2, gives

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sigma^2}+\frac{n\left(\overline{X}-\mu\right)^2}{\sigma^2},$$

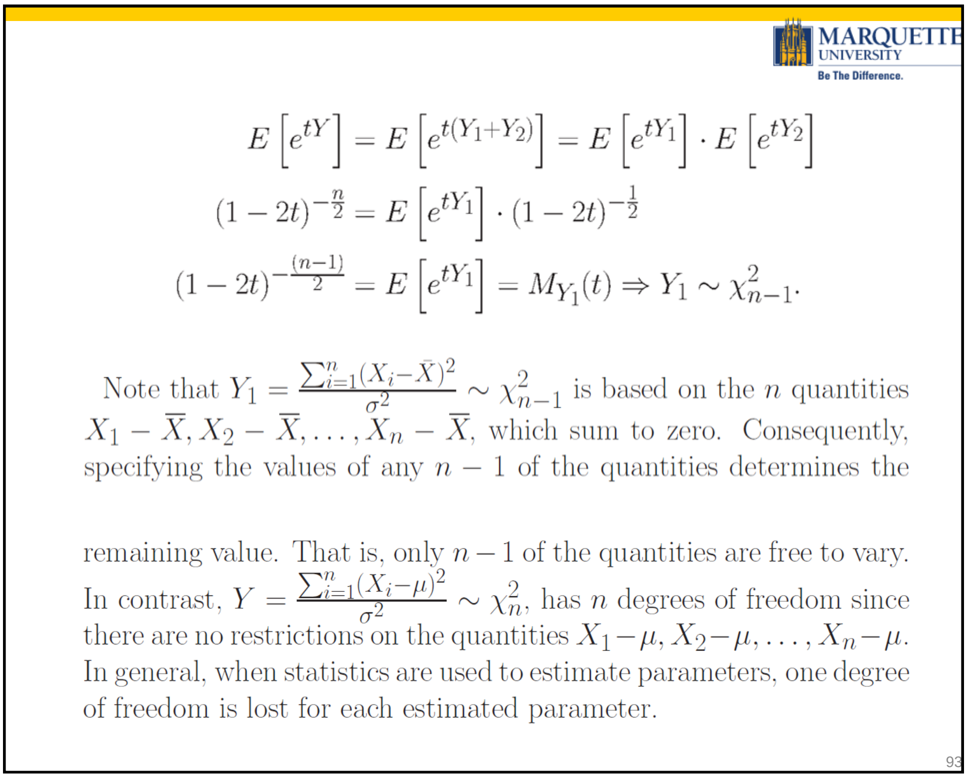
which is the same as

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}=\frac{(n-1)S^2}{\sigma^2}+\frac{n(\overline{X}-\mu)^2}{\sigma^2}. \qquad (6.21)$$

Since $\overline{X}\sim N(\mu,\frac{\sigma}{\sqrt{n}})$, it follows that $\frac{n(\bar{X}-\mu)^2}{\sigma^2}\sim\chi_1^2$

To simplify notation, let Y, Y_1, and Y_2 represent $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}$, $\frac{(n-1)S^2}{\sigma^2}$, and $\frac{n(\bar{X}-\mu)^2}{\sigma^2}$ in (6.21) respectively. By the part (1) of Theorem 6.3 on page 126, Y_1 and Y_2 are independent. Therefore,

92



$$E\left[e^{tY}\right]=E\left[e^{t(Y_1+Y_2)}\right]=E\left[e^{tY_1}\right]\cdot E\left[e^{tY_2}\right]$$

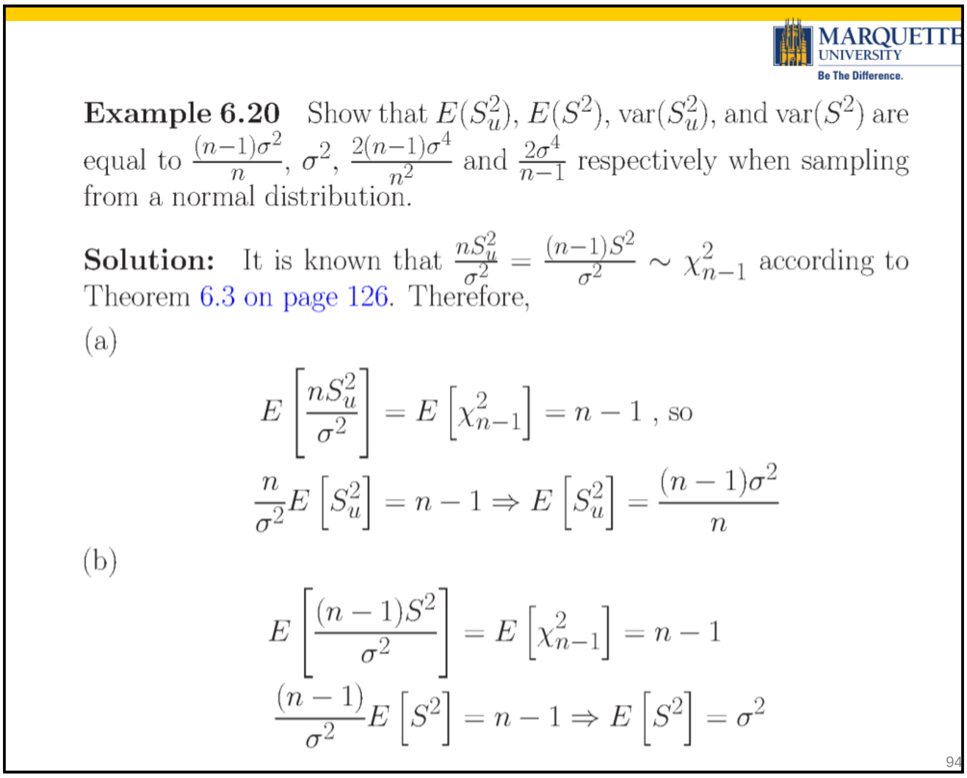
$$(1-2t)^{-\frac{n}{2}}=E\left[e^{tY_1}\right]\cdot(1-2t)^{-\frac{1}{2}}$$

$$(1-2t)^{-\frac{(n-1)}{2}}=E\left[e^{tY_1}\right]=M_{Y_1}(t)\Rightarrow Y_1\sim\chi^2_{n-1}.$$

Note that $Y_1=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}\sim\chi^2_{n-1}$ is based on the n quantities $X_1-\overline{X},X_2-\overline{X},\ldots,X_n-\overline{X}$, which sum to zero. Consequently, specifying the values of any $n-1$ of the quantities determines the

remaining value. That is, only $n-1$ of the quantities are free to vary. In contrast, $Y=\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}\sim\chi^2_n$, has n degrees of freedom since there are no restrictions on the quantities $X_1-\mu,X_2-\mu,\ldots,X_n-\mu$. In general, when statistics are used to estimate parameters, one degree of freedom is lost for each estimated parameter.

93



Example 6.20 Show that $E(S_u^2)$, $E(S^2)$, $\text{var}(S_u^2)$, and $\text{var}(S^2)$ are equal to $\frac{(n-1)\sigma^2}{n}$, σ^2, $\frac{2(n-1)\sigma^4}{n^2}$ and $\frac{2\sigma^4}{n-1}$ respectively when sampling from a normal distribution.

Solution: It is known that $\frac{nS_u^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ according to Theorem 6.3 on page 126. Therefore,

(a)

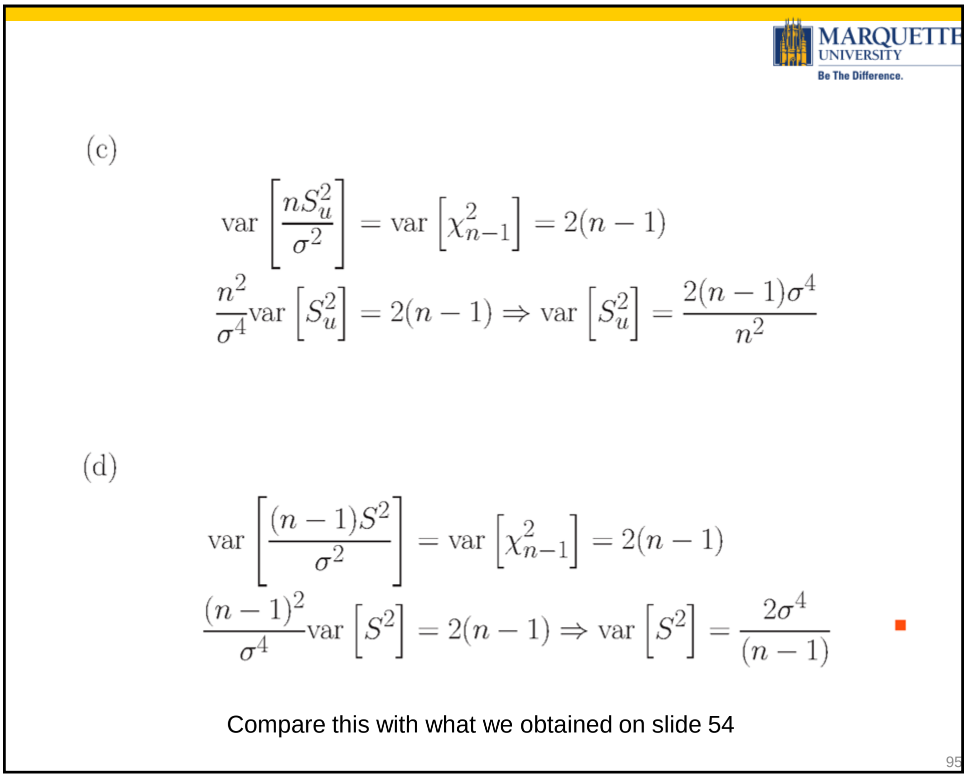
$$E\left[\frac{nS_u^2}{\sigma^2}\right] = E\left[\chi^2_{n-1}\right] = n-1 \text{ , so}$$

$$\frac{n}{\sigma^2}E\left[S_u^2\right] = n-1 \Rightarrow E\left[S_u^2\right] = \frac{(n-1)\sigma^2}{n}$$

(b)

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = E\left[\chi^2_{n-1}\right] = n-1$$

$$\frac{(n-1)}{\sigma^2}E\left[S^2\right] = n-1 \Rightarrow E\left[S^2\right] = \sigma^2$$



(c)

$$\text{var}\left[\frac{nS_u^2}{\sigma^2}\right] = \text{var}\left[\chi^2_{n-1}\right] = 2(n-1)$$

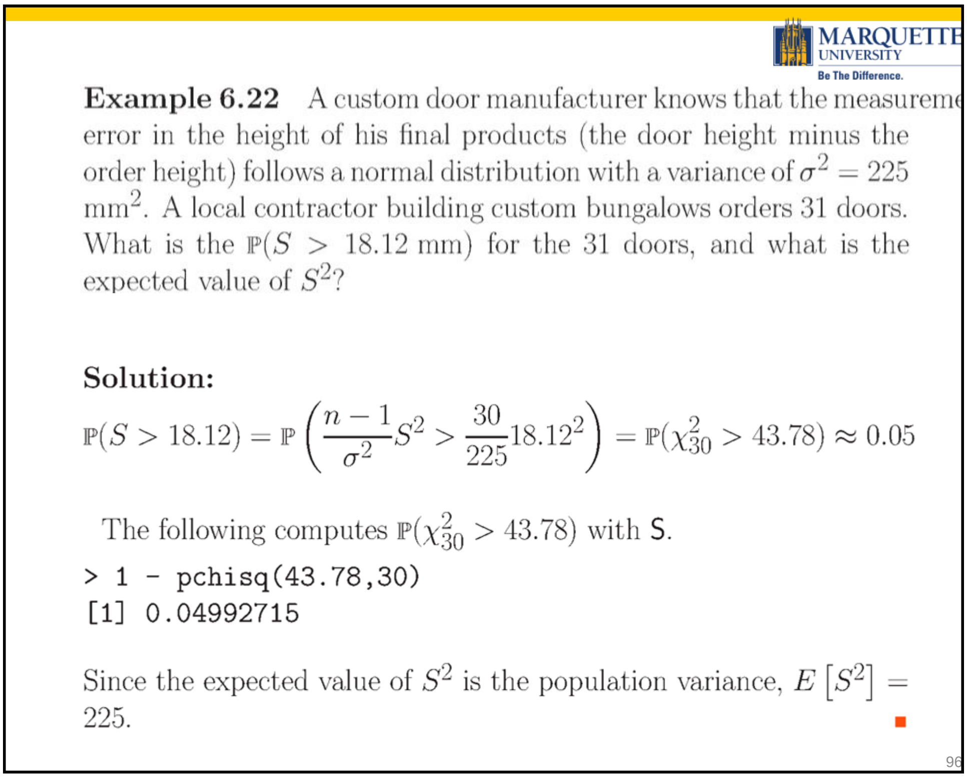
$$\frac{n^2}{\sigma^4}\text{var}\left[S_u^2\right] = 2(n-1) \Rightarrow \text{var}\left[S_u^2\right] = \frac{2(n-1)\sigma^4}{n^2}$$

(d)

$$\text{var}\left[\frac{(n-1)S^2}{\sigma^2}\right] = \text{var}\left[\chi^2_{n-1}\right] = 2(n-1)$$

$$\frac{(n-1)^2}{\sigma^4}\text{var}\left[S^2\right] = 2(n-1) \Rightarrow \text{var}\left[S^2\right] = \frac{2\sigma^4}{(n-1)} \quad ■$$

Compare this with what we obtained on slide 54



Example 6.22 A custom door manufacturer knows that the measureme error in the height of his final products (the door height minus the order height) follows a normal distribution with a variance of $\sigma^2 = 225$ mm^2. A local contractor building custom bungalows orders 31 doors. What is the $\mathbb{P}(S > 18.12 \text{ mm})$ for the 31 doors, and what is the expected value of S^2?

Solution:

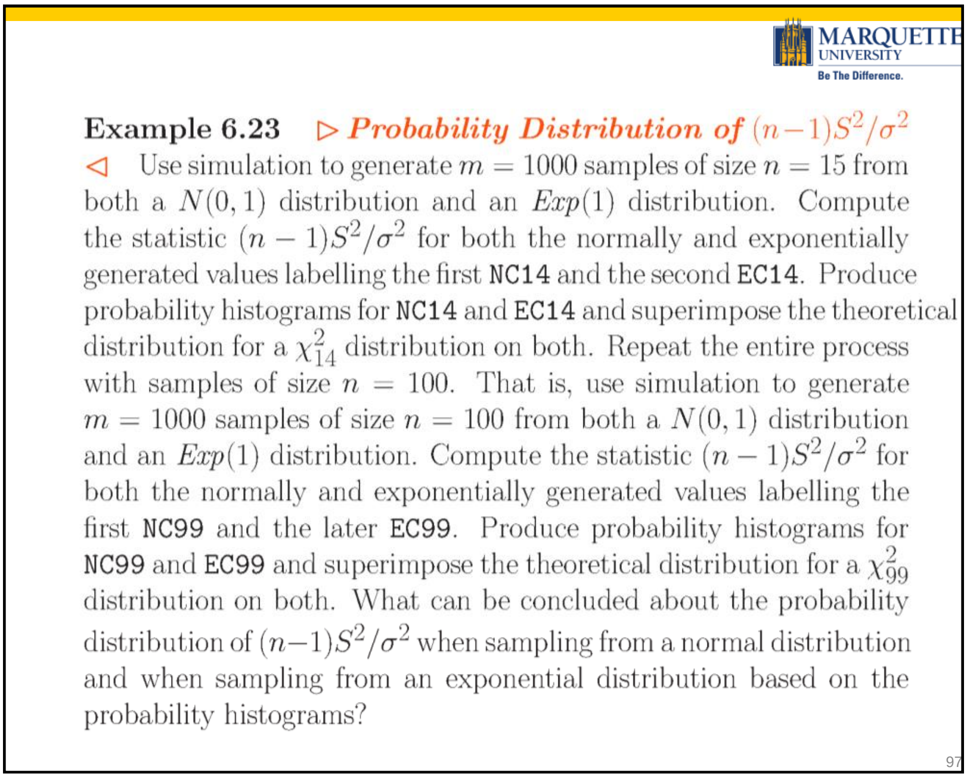
$$\mathbb{P}(S > 18.12) = \mathbb{P}\left(\frac{n-1}{\sigma^2}S^2 > \frac{30}{225}18.12^2\right) = \mathbb{P}(\chi^2_{30} > 43.78) \approx 0.05$$

The following computes $\mathbb{P}(\chi^2_{30} > 43.78)$ with S.

```
> 1 - pchisq(43.78,30)
[1] 0.04992715
```

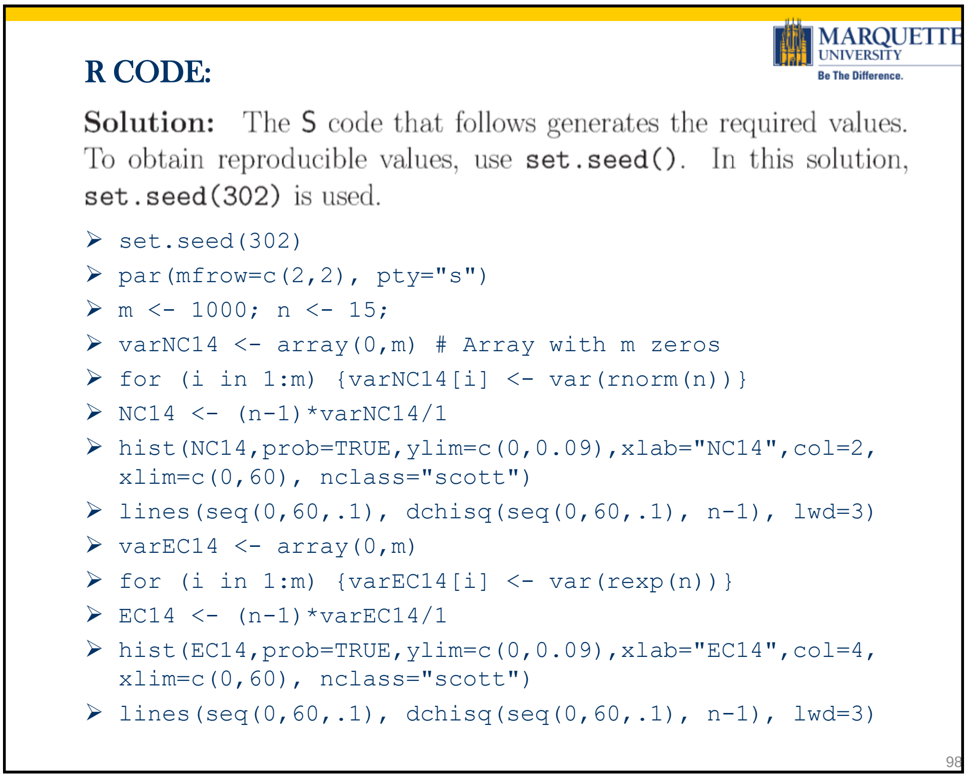
Since the expected value of S^2 is the population variance, $E\left[S^2\right] = 225$. ■

96



Example 6.23 ▷ ***Probability Distribution of*** $(n-1)S^2/\sigma^2$ ◁ Use simulation to generate $m = 1000$ samples of size $n = 15$ from both a $N(0,1)$ distribution and an $Exp(1)$ distribution. Compute the statistic $(n-1)S^2/\sigma^2$ for both the normally and exponentially generated values labelling the first `NC14` and the second `EC14`. Produce probability histograms for `NC14` and `EC14` and superimpose the theoretical distribution for a χ^2_{14} distribution on both. Repeat the entire process with samples of size $n = 100$. That is, use simulation to generate $m = 1000$ samples of size $n = 100$ from both a $N(0,1)$ distribution and an $Exp(1)$ distribution. Compute the statistic $(n-1)S^2/\sigma^2$ for both the normally and exponentially generated values labelling the first `NC99` and the later `EC99`. Produce probability histograms for `NC99` and `EC99` and superimpose the theoretical distribution for a χ^2_{99} distribution on both. What can be concluded about the probability distribution of $(n-1)S^2/\sigma^2$ when sampling from a normal distribution and when sampling from an exponential distribution based on the probability histograms?

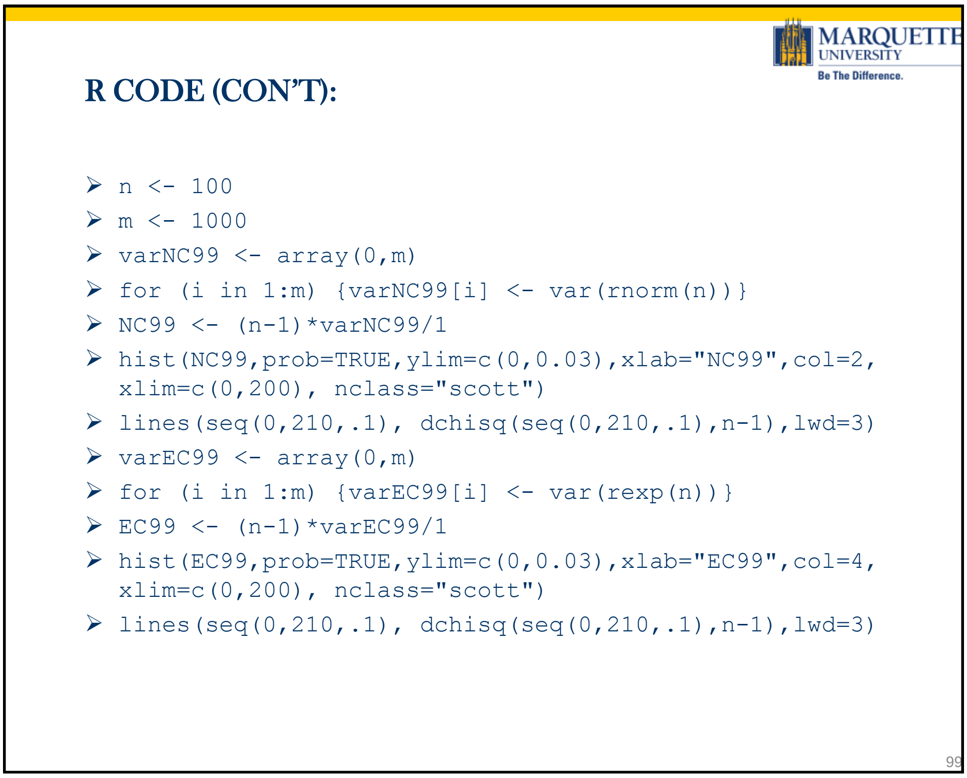
97



R CODE:

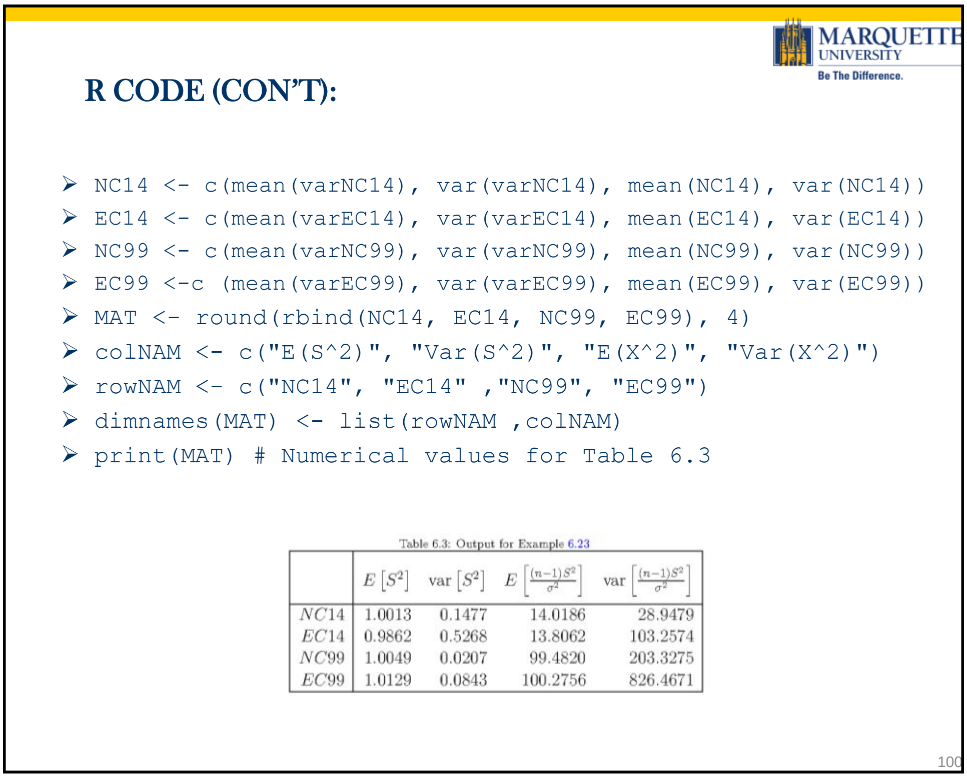
Solution: The S code that follows generates the required values. To obtain reproducible values, use `set.seed()`. In this solution, `set.seed(302)` is used.

```
set.seed(302)
par(mfrow=c(2,2), pty="s")
m <- 1000; n <- 15;
varNC14 <- array(0,m) # Array with m zeros
for (i in 1:m) {varNC14[i] <- var(rnorm(n))}
NC14 <- (n-1)*varNC14/1
hist(NC14,prob=TRUE,ylim=c(0,0.09),xlab="NC14",col=2,
xlim=c(0,60), nclass="scott")
lines(seq(0,60,.1), dchisq(seq(0,60,.1), n-1), lwd=3)
varEC14 <- array(0,m)
for (i in 1:m) {varEC14[i] <- var(rexp(n))}
EC14 <- (n-1)*varEC14/1
hist(EC14,prob=TRUE,ylim=c(0,0.09),xlab="EC14",col=4,
xlim=c(0,60), nclass="scott")
lines(seq(0,60,.1), dchisq(seq(0,60,.1), n-1), lwd=3)
```



R CODE (CON'T):

```
n <- 100
m <- 1000
varNC99 <- array(0,m)
for (i in 1:m) {varNC99[i] <- var(rnorm(n))}
NC99 <- (n-1)*varNC99/1
hist(NC99,prob=TRUE,ylim=c(0,0.03),xlab="NC99",col=2,
xlim=c(0,200), nclass="scott")
lines(seq(0,210,.1), dchisq(seq(0,210,.1),n-1),lwd=3)
varEC99 <- array(0,m)
for (i in 1:m) {varEC99[i] <- var(rexp(n))}
EC99 <- (n-1)*varEC99/1
hist(EC99,prob=TRUE,ylim=c(0,0.03),xlab="EC99",col=4,
xlim=c(0,200), nclass="scott")
lines(seq(0,210,.1), dchisq(seq(0,210,.1),n-1),lwd=3)
```

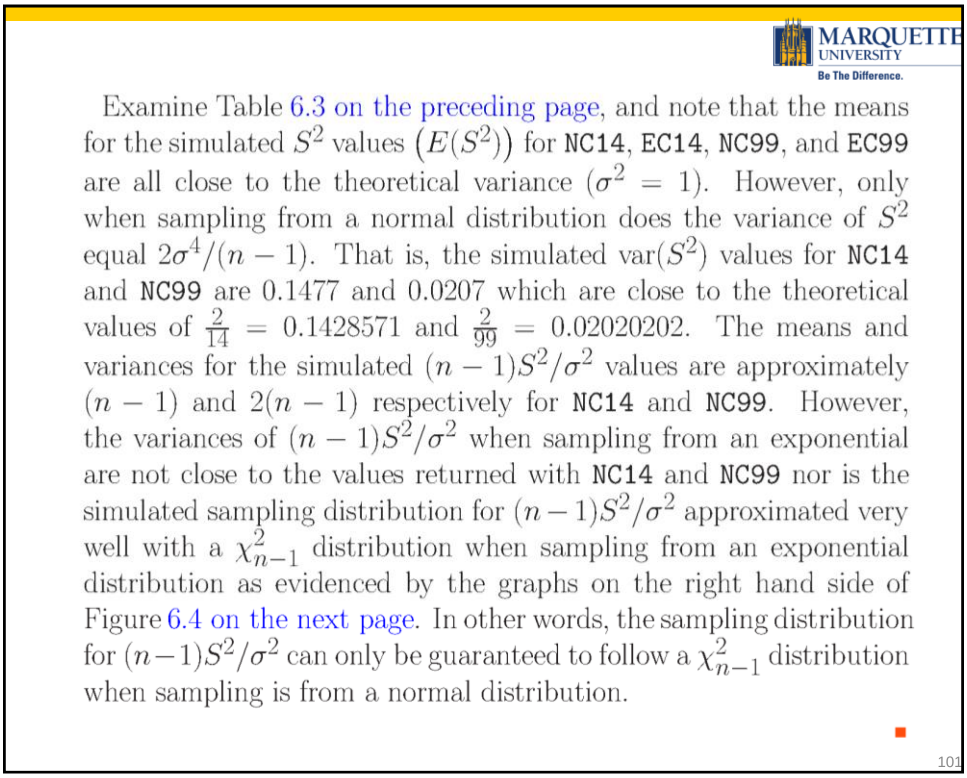


R CODE (CON'T):

```
➢ NC14 <- c(mean(varNC14), var(varNC14), mean(NC14), var(NC14))
➢ EC14 <- c(mean(varEC14), var(varEC14), mean(EC14), var(EC14))
➢ NC99 <- c(mean(varNC99), var(varNC99), mean(NC99), var(NC99))
➢ EC99 <-c (mean(varEC99), var(varEC99), mean(EC99), var(EC99))
➢ MAT <- round(rbind(NC14, EC14, NC99, EC99), 4)
➢ colNAM <- c("E(S^2)", "Var(S^2)", "E(X^2)", "Var(X^2)")
➢ rowNAM <- c("NC14", "EC14" ,"NC99", "EC99")
➢ dimnames(MAT) <- list(rowNAM ,colNAM)
➢ print(MAT) # Numerical values for Table 6.3
```

Table 6.3: Output for Example 6.23

	$E\left[S^2\right]$	$\text{var}\left[S^2\right]$	$E\left[\frac{(n-1)S^2}{\sigma^2}\right]$	$\text{var}\left[\frac{(n-1)S^2}{\sigma^2}\right]$
$NC14$	1.0013	0.1477	14.0186	28.9479
$EC14$	0.9862	0.5268	13.8062	103.2574
$NC99$	1.0049	0.0207	99.4820	203.3275
$EC99$	1.0129	0.0843	100.2756	826.4671



Examine Table 6.3 on the preceding page, and note that the means for the simulated S^2 values $\left(E(S^2)\right)$ for NC14, EC14, NC99, and EC99 are all close to the theoretical variance ($\sigma^2 = 1$). However, only when sampling from a normal distribution does the variance of S^2 equal $2\sigma^4/(n-1)$. That is, the simulated var(S^2) values for NC14 and NC99 are 0.1477 and 0.0207 which are close to the theoretical values of $\frac{2}{14} = 0.1428571$ and $\frac{2}{99} = 0.02020202$. The means and variances for the simulated $(n-1)S^2/\sigma^2$ values are approximately $(n-1)$ and $2(n-1)$ respectively for NC14 and NC99. However, the variances of $(n-1)S^2/\sigma^2$ when sampling from an exponential are not close to the values returned with NC14 and NC99 nor is the simulated sampling distribution for $(n-1)S^2/\sigma^2$ approximated very well with a χ^2_{n-1} distribution when sampling from an exponential distribution as evidenced by the graphs on the right hand side of Figure 6.4 on the next page. In other words, the sampling distribution for $(n-1)S^2/\sigma^2$ can only be guaranteed to follow a χ^2_{n-1} distribution when sampling is from a normal distribution. ■

GRAPH OUTPUT:

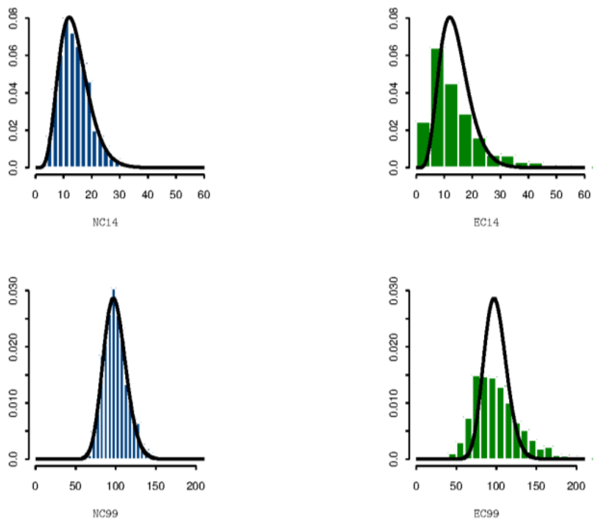


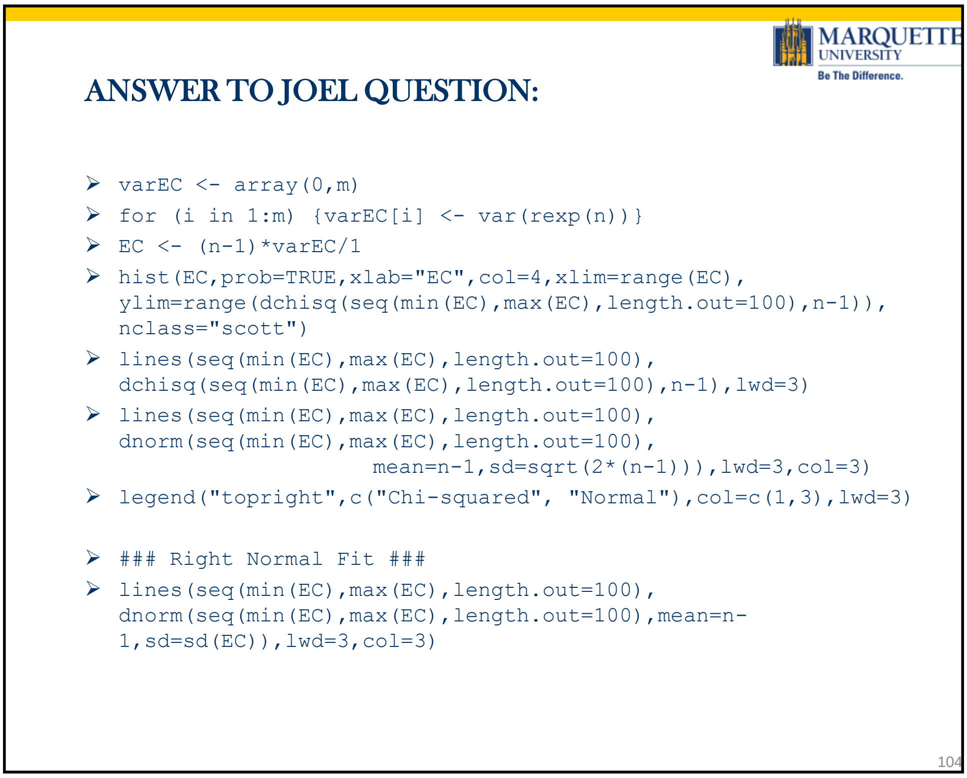
Figure 6.4: Probability histograms for simulated distributions of $\frac{(n-1)S^2}{\sigma^2}$ when sampling from normal and exponential distributions. NC14 designates the

JOEL QUESTION:

- **I tried the simulations with $n = 500$ and $n = 1000$. The $\frac{(n-1)S^2}{\sigma^2}$ from the exponential doesn't appear to be approaching χ^2 even for $n = 500$ or $n = 1000$. What do you think?**

```
➢ n <- 500
➢ m <- 1000
➢ par(mfrow=c(2,1))

➢ varNC <- array(0,m)
➢ for (i in 1:m) {varNC[i] <- var(rnorm(n))}
➢ NC <- (n-1)*varNC/1
➢ hist(NC,prob=TRUE,xlab="NC",col=2,xlim=range(NC),
  nclass="scott")
➢ lines(seq(min(NC),max(NC),length.out=100),
  dchisq(seq(min(NC),max(NC),length.out=100),n-1),lwd=3)
➢ lines(seq(min(NC),max(NC),length.out=100),
  dnorm(seq(min(NC),max(NC),length.out=100),
                      mean=n-1,sd=sqrt(2*(n-1))),lwd=3,col=3)
➢ legend("topright",c("Chi-squared", "Normal"),col=c(1,3),lwd=3)
```

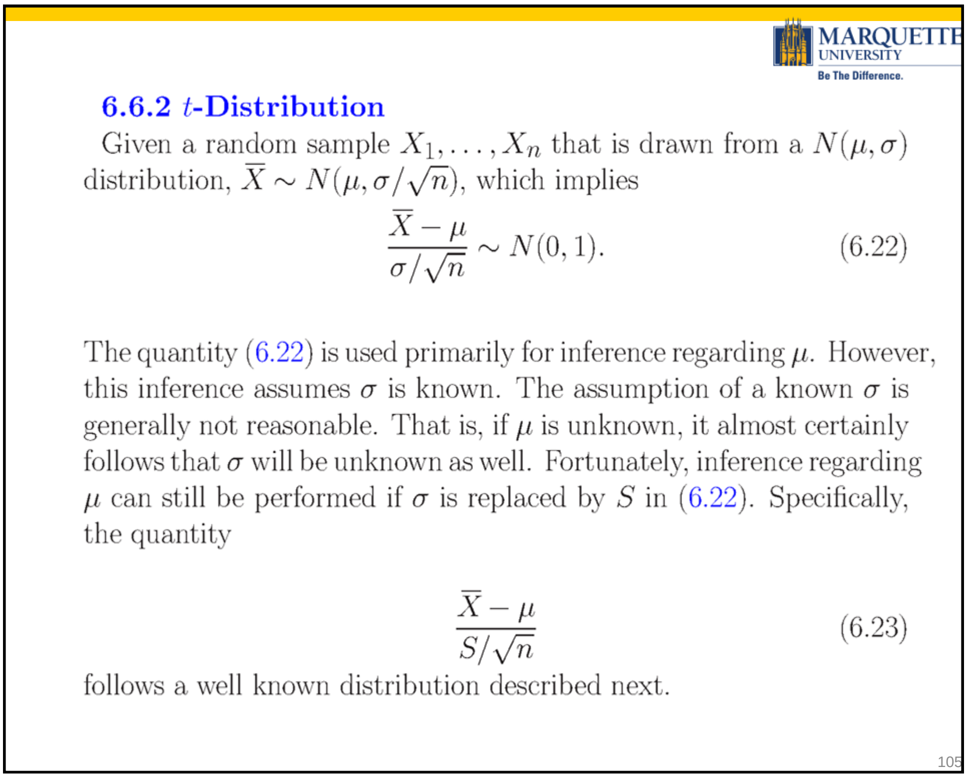


ANSWER TO JOEL QUESTION:

```
➢ varEC <- array(0,m)
➢ for (i in 1:m) {varEC[i] <- var(rexp(n))}
➢ EC <- (n-1)*varEC/1
➢ hist(EC,prob=TRUE,xlab="EC",col=4,xlim=range(EC),
  ylim=range(dchisq(seq(min(EC),max(EC),length.out=100),n-1)),
  nclass="scott")
➢ lines(seq(min(EC),max(EC),length.out=100),
  dchisq(seq(min(EC),max(EC),length.out=100),n-1),lwd=3)
➢ lines(seq(min(EC),max(EC),length.out=100),
  dnorm(seq(min(EC),max(EC),length.out=100),
                    mean=n-1,sd=sqrt(2*(n-1))),lwd=3,col=3)
➢ legend("topright",c("Chi-squared", "Normal"),col=c(1,3),lwd=3)

➢ ### Right Normal Fit ###
➢ lines(seq(min(EC),max(EC),length.out=100),
  dnorm(seq(min(EC),max(EC),length.out=100),mean=n-
  1,sd=sd(EC)),lwd=3,col=3)
```

104



6.6.2 t-Distribution

Given a random sample $X_1, \ldots, X_n$ that is drawn from a $N(\mu, \sigma)$ distribution, $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$, which implies

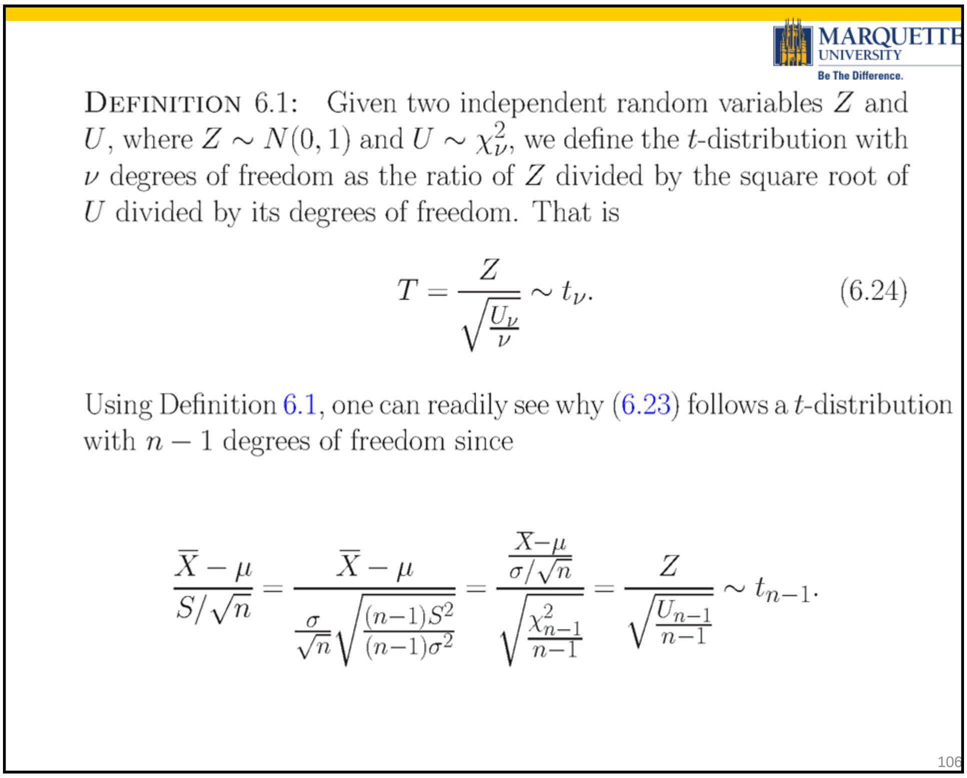
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1). \tag{6.22}$$

The quantity (6.22) is used primarily for inference regarding μ. However, this inference assumes σ is known. The assumption of a known σ is generally not reasonable. That is, if μ is unknown, it almost certainly follows that σ will be unknown as well. Fortunately, inference regarding μ can still be performed if σ is replaced by S in (6.22). Specifically, the quantity

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{6.23}$$

follows a well known distribution described next.

105

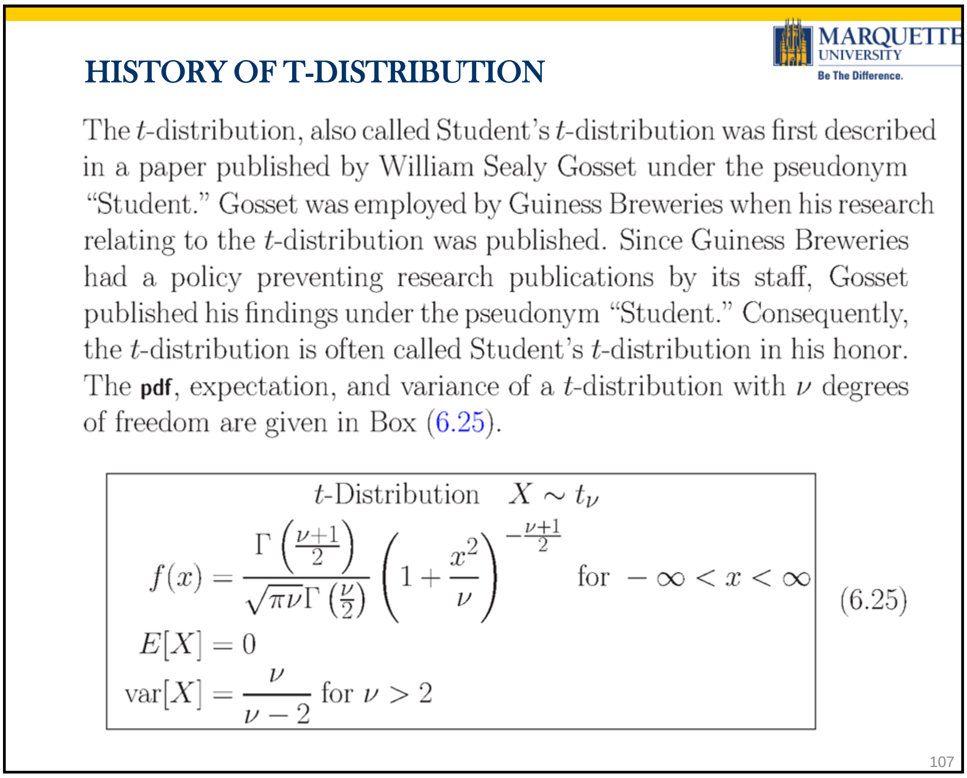


Definition 6.1: Given two independent random variables Z and U, where $Z \sim N(0,1)$ and $U \sim \chi^2_\nu$, we define the t-distribution with ν degrees of freedom as the ratio of Z divided by the square root of U divided by its degrees of freedom. That is

$$T = \frac{Z}{\sqrt{\frac{U_\nu}{\nu}}} \sim t_\nu. \tag{6.24}$$

Using Definition 6.1, one can readily see why (6.23) follows a t-distribution with $n-1$ degrees of freedom since

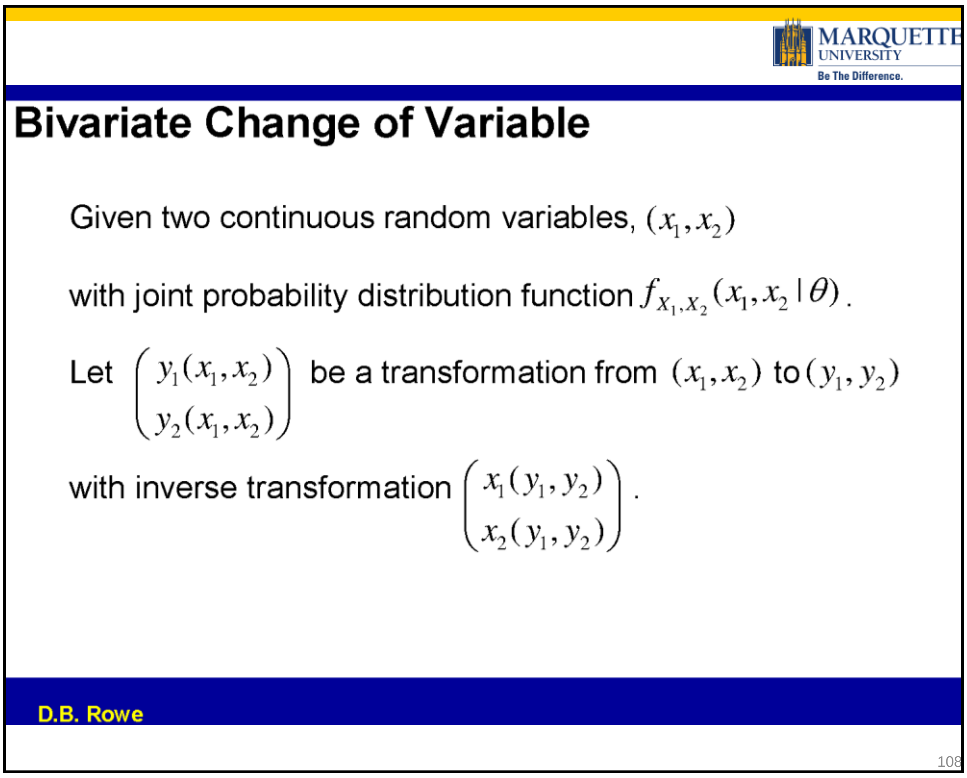
$$\frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = \frac{Z}{\sqrt{\frac{U_{n-1}}{n-1}}} \sim t_{n-1}.$$



HISTORY OF T-DISTRIBUTION

The t-distribution, also called Student's t-distribution was first described in a paper published by William Sealy Gosset under the pseudonym "Student." Gosset was employed by Guiness Breweries when his research relating to the t-distribution was published. Since Guiness Breweries had a policy preventing research publications by its staff, Gosset published his findings under the pseudonym "Student." Consequently, the t-distribution is often called Student's t-distribution in his honor. The **pdf**, expectation, and variance of a t-distribution with ν degrees of freedom are given in Box (6.25).

$$\boxed{\begin{array}{l} \qquad\qquad t\text{-Distribution} \quad X \sim t_\nu \\ f(x) = \dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\dfrac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for } -\infty < x < \infty \\ E[X] = 0 \\ \text{var}[X] = \dfrac{\nu}{\nu-2} \text{ for } \nu > 2 \end{array}} \tag{6.25}$$



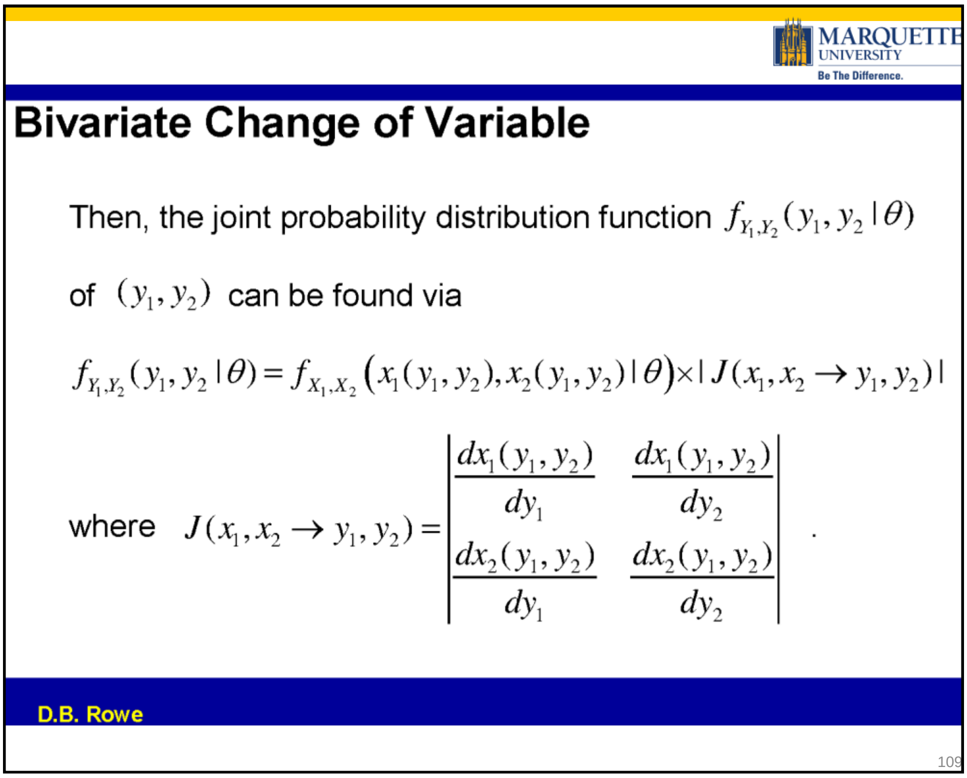
Bivariate Change of Variable

Given two continuous random variables, (x_1,x_2)

with joint probability distribution function $f_{X_1,X_2}(x_1,x_2\,|\,\theta)$.

Let $\begin{pmatrix} y_1(x_1,x_2) \\ y_2(x_1,x_2) \end{pmatrix}$ be a transformation from (x_1,x_2) to (y_1,y_2)

with inverse transformation $\begin{pmatrix} x_1(y_1,y_2) \\ x_2(y_1,y_2) \end{pmatrix}$.



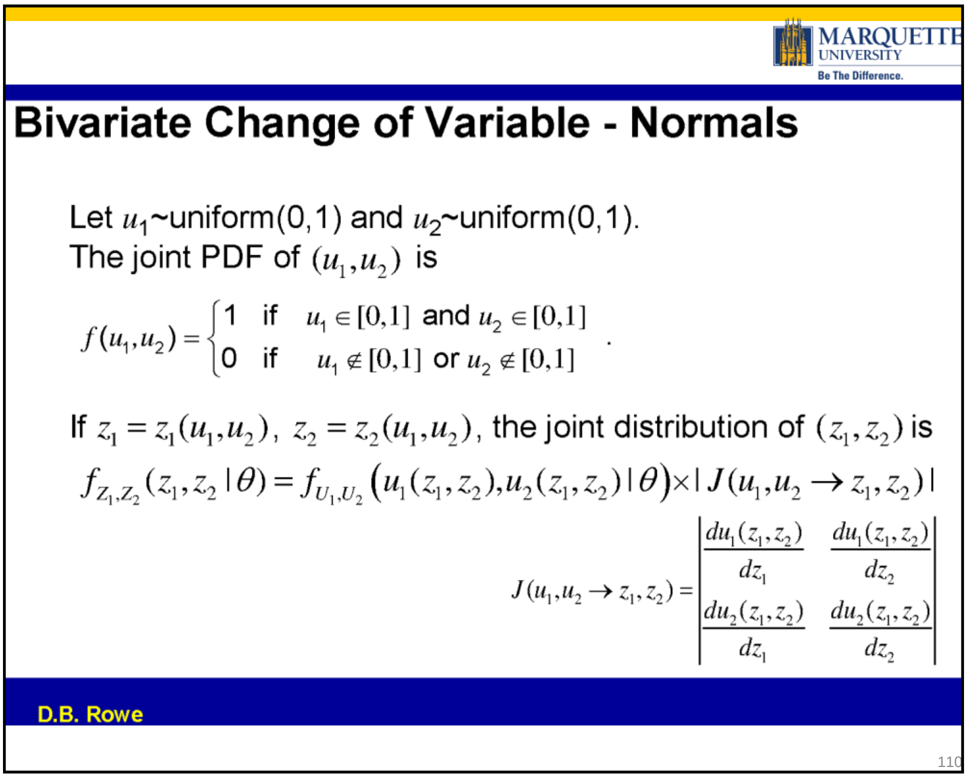
Bivariate Change of Variable

Then, the joint probability distribution function $f_{Y_1,Y_2}(y_1,y_2\,|\,\theta)$

of (y_1,y_2) can be found via

$$f_{Y_1,Y_2}(y_1,y_2\,|\,\theta)=f_{X_1,X_2}\big(x_1(y_1,y_2),x_2(y_1,y_2)\,|\,\theta\big)\times|\,J(x_1,x_2\rightarrow y_1,y_2)\,|$$

where $J(x_1,x_2\rightarrow y_1,y_2)=\begin{vmatrix} \dfrac{dx_1(y_1,y_2)}{dy_1} & \dfrac{dx_1(y_1,y_2)}{dy_2} \\ \dfrac{dx_2(y_1,y_2)}{dy_1} & \dfrac{dx_2(y_1,y_2)}{dy_2} \end{vmatrix}$.



Bivariate Change of Variable - Normals

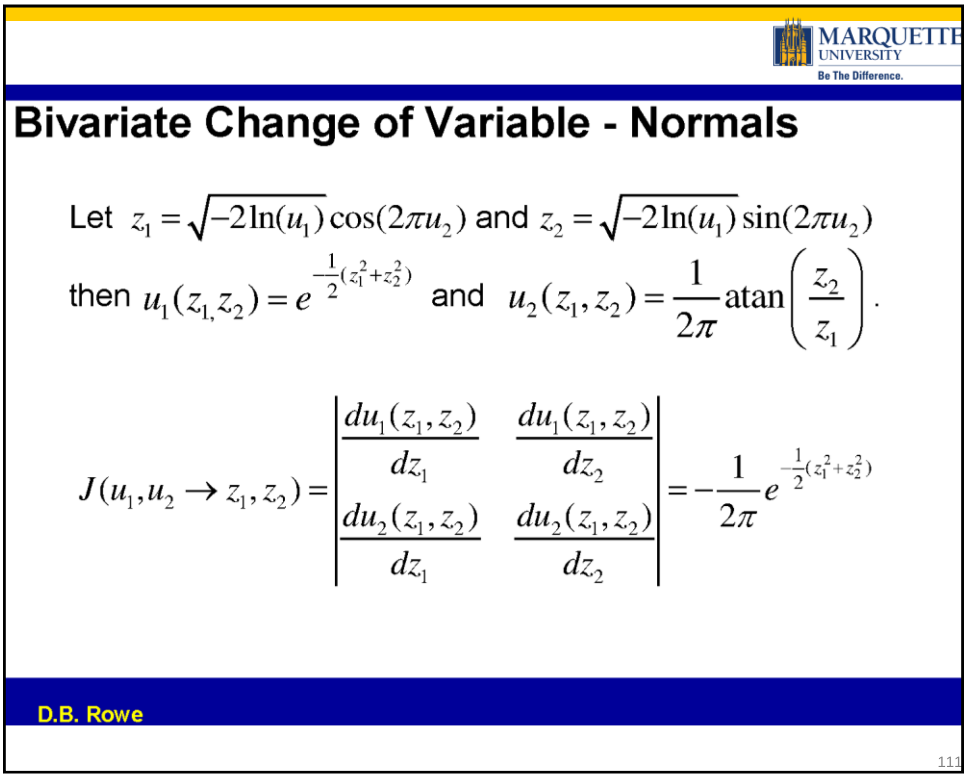
Let u_1~uniform(0,1) and u_2~uniform(0,1).
The joint PDF of (u_1,u_2) is

$$f(u_1,u_2)=\begin{cases}1 & \text{if} \quad u_1\in[0,1] \text{ and } u_2\in[0,1]\\ 0 & \text{if} \quad u_1\notin[0,1] \text{ or } u_2\notin[0,1]\end{cases}.$$

If $z_1=z_1(u_1,u_2)$, $z_2=z_2(u_1,u_2)$, the joint distribution of (z_1,z_2) is

$$f_{Z_1,Z_2}(z_1,z_2\,|\,\theta)=f_{U_1,U_2}\left(u_1(z_1,z_2),u_2(z_1,z_2)\,|\,\theta\right)\times|\,J(u_1,u_2\to z_1,z_2)\,|$$

$$J(u_1,u_2\to z_1,z_2)=\begin{vmatrix}\dfrac{du_1(z_1,z_2)}{dz_1} & \dfrac{du_1(z_1,z_2)}{dz_2}\\ \dfrac{du_2(z_1,z_2)}{dz_1} & \dfrac{du_2(z_1,z_2)}{dz_2}\end{vmatrix}$$



Bivariate Change of Variable - Normals

Let $z_1=\sqrt{-2\ln(u_1)}\cos(2\pi u_2)$ and $z_2=\sqrt{-2\ln(u_1)}\sin(2\pi u_2)$

then $u_1(z_{1,}z_2)=e^{-\frac{1}{2}(z_1^2+z_2^2)}$ and $u_2(z_1,z_2)=\dfrac{1}{2\pi}\text{atan}\left(\dfrac{z_2}{z_1}\right)$.

$$J(u_1,u_2\to z_1,z_2)=\begin{vmatrix}\dfrac{du_1(z_1,z_2)}{dz_1} & \dfrac{du_1(z_1,z_2)}{dz_2}\\ \dfrac{du_2(z_1,z_2)}{dz_1} & \dfrac{du_2(z_1,z_2)}{dz_2}\end{vmatrix}=-\frac{1}{2\pi}e^{-\frac{1}{2}(z_1^2+z_2^2)}$$

Bivariate Change of Variable - Normals

Therefore,

$$f_{Z_1,Z_2}(z_1,z_2\,|\,\theta)=f_{U_1,U_2}\big(u_1(z_1,z_2),u_2(z_1,z_2)\,|\,\theta\big)\times|\,J(u_1,u_2\to z_1,z_2)\,|$$

which upon insertion yields

$$\begin{aligned} f_{Z_1,Z_2}(z_1,z_2\,|\,\theta) &= \frac{1}{2\pi}e^{-\frac{1}{2}(z_1^2+z_2^2)} \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z_1^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z_2^2}\quad. \end{aligned}$$

Joint PDF factors thus independent

This means $z_1 \sim N(0,1)$, $z_2 \sim N(0,1)$, z_1 and $z_2$are independent.

Bivariate Change of Variable - Normals

Generate 10^6 independent uniform(0,1)'s.

The first half of the 10^6 standard uniform random variates were used as u_1's and the second half used as u_2's.

Take each (u_1,u_2) pair to produce a (z_1,z_2) pair.

$$z_1=\sqrt{-2\ln(u_1)}\cos(2\pi u_2) \qquad z_2=\sqrt{-2\ln(u_1)}\sin(2\pi u_2)$$

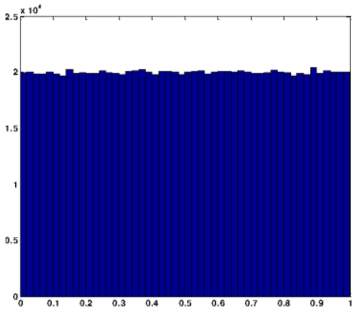
(z_1,z_2) are independent normally distributed.

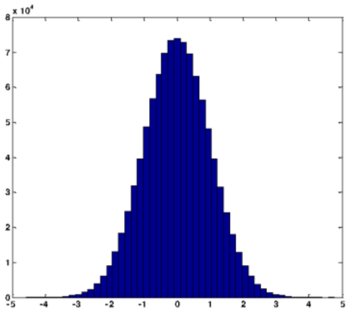
Bivariate Change of Variable - Normals

```
n=10^6;
u1=rand(n/2,1);
u2=rand(n/2,1);
figure(1)
hist([u1;u2],50)
```

```
z1=sqrt(-2*log(u1)).*cos(2*pi*u2);
z2=sqrt(-2*log(u1)).*sin(2*pi*u2);
figure(2)
hist([z1;z2],50)
```

```
[mean(u1),var(u1)]
[mean(u2),var(u2)]
[mean(z1),var(z1)]
[mean(z2),var(z2)]
[corr(u1,u2),corr(z1,z2)]
```





```
 0.5000   0.0832
 0.5006   0.0833
 0.0000   1.0011
-0.0016   0.9970
 0.0025   0.0013
```

↑

Uncorrelated and since normal are independent

Bivariate Change of Variable - Student-t

We showed that if x_i~normal(μ,σ^2) for i=1,…,n, then

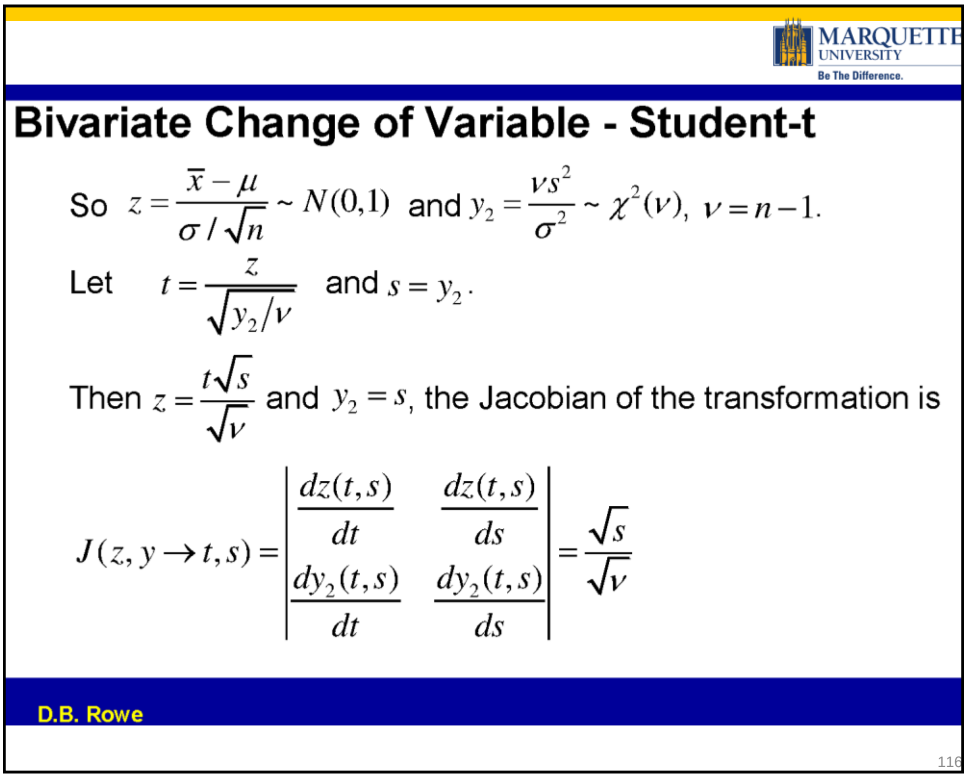
the distribution of $\bar{x} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ and $z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

and that the distribution of $y_2 = \sum_{i=1}^{n}\left(\dfrac{x_i-\bar{x}}{\sigma}\right)^2 \sim \chi^2(n-1)$.

Note that $y_2 = \dfrac{(n-1)s^2}{\sigma^2}$.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

It turns out that z and $\dfrac{(n-1)s^2}{\sigma^2}$ are statistically independent!



Bivariate Change of Variable - Student-t

So $z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $y_2 = \dfrac{\nu s^2}{\sigma^2} \sim \chi^2(\nu),\ \nu = n-1.$

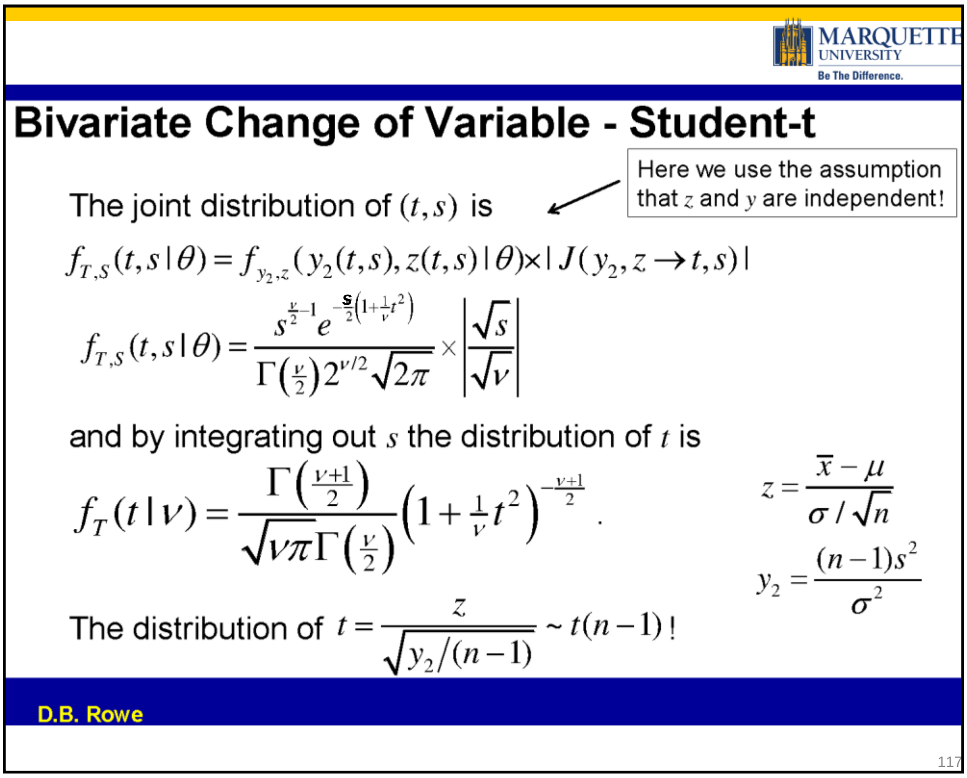
Let $t = \dfrac{z}{\sqrt{y_2/\nu}}$ and $s = y_2$.

Then $z = \dfrac{t\sqrt{s}}{\sqrt{\nu}}$ and $y_2 = s$, the Jacobian of the transformation is

$$J(z,y \to t,s) = \begin{vmatrix} \dfrac{dz(t,s)}{dt} & \dfrac{dz(t,s)}{ds} \\ \dfrac{dy_2(t,s)}{dt} & \dfrac{dy_2(t,s)}{ds} \end{vmatrix} = \frac{\sqrt{s}}{\sqrt{\nu}}$$

D.B. Rowe

116



Bivariate Change of Variable - Student-t

The joint distribution of (t,s) is

Here we use the assumption that z and y are independent!

$$f_{T,S}(t,s \mid \theta) = f_{y_2,z}(y_2(t,s), z(t,s) \mid \theta) \times | J(y_2, z \to t,s) |$$

$$f_{T,S}(t,s \mid \theta) = \frac{s^{\frac{\nu}{2}-1} e^{-\frac{s}{2}\left(1+\frac{1}{\nu}t^2\right)}}{\Gamma\left(\frac{\nu}{2}\right) 2^{\nu/2}\sqrt{2\pi}} \times \left| \frac{\sqrt{s}}{\sqrt{\nu}} \right|$$

and by integrating out s the distribution of t is

$$f_T(t \mid \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1+\tfrac{1}{\nu}t^2\right)^{-\frac{\nu+1}{2}}.$$

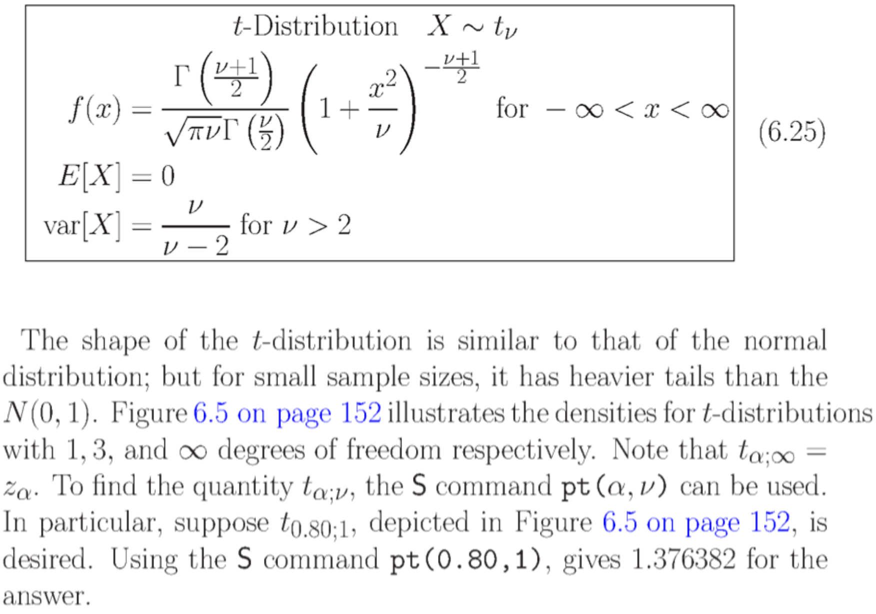
$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

$$y_2 = \frac{(n-1)s^2}{\sigma^2}$$

The distribution of $t = \dfrac{z}{\sqrt{y_2/(n-1)}} \sim t(n-1)$!

D.B. Rowe

117



$$t\text{-Distribution} \quad X \sim t_\nu$$

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for } -\infty < x < \infty \qquad (6.25)$$

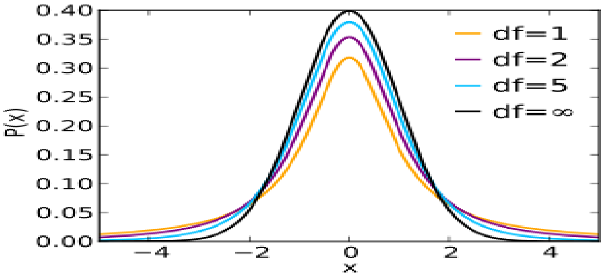
$$E[X] = 0$$

$$\text{var}[X] = \frac{\nu}{\nu-2} \text{ for } \nu > 2$$

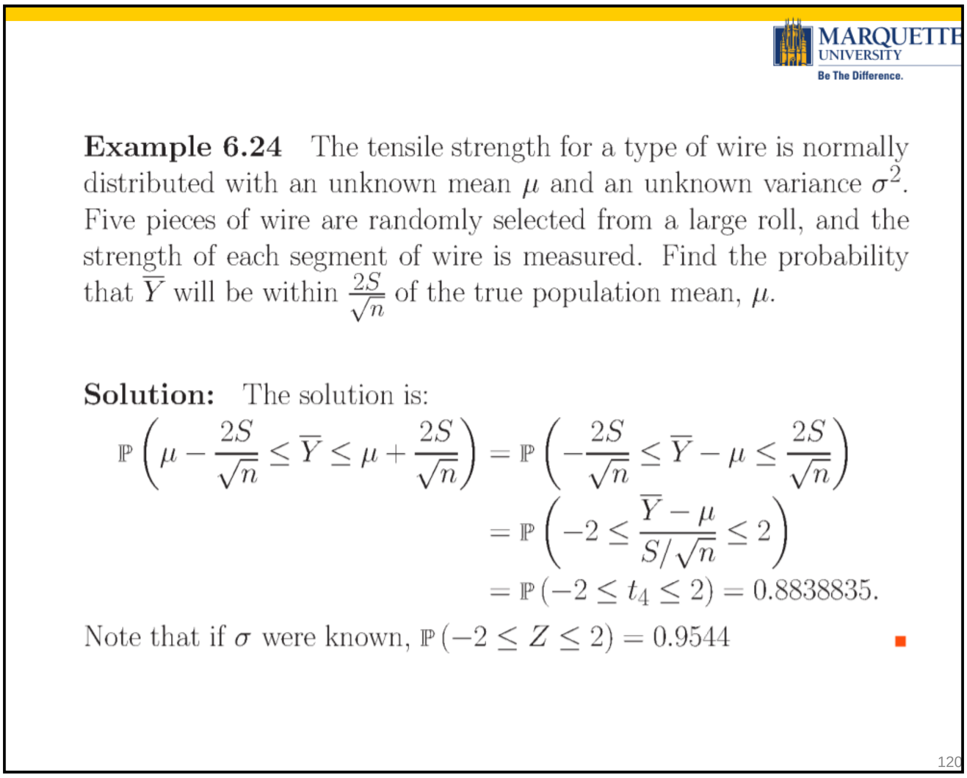
The shape of the t-distribution is similar to that of the normal distribution; but for small sample sizes, it has heavier tails than the $N(0,1)$. Figure 6.5 on page 152 illustrates the densities for t-distributions with $1, 3$, and ∞ degrees of freedom respectively. Note that $t_{\alpha;\infty} = z_\alpha$. To find the quantity $t_{\alpha;\nu}$, the S command `pt(`α, ν`)` can be used. In particular, suppose $t_{0.80;1}$, depicted in Figure 6.5 on page 152, is desired. Using the S command `pt(0.80,1)`, gives 1.376382 for the answer.

REMARKS ON T-DISTRIBUTION

1. **t-dist. has similar shape as $N(0,1)$, but is flatter than $N(0,1)$.**
2. **The t-distribution is symmetric around 0 as $N(0,1)$ is.**
3. **It has a range from $-\infty$ to ∞ as the range of $N(0,1)$.**
4. **Unlike $N(0,1)$, the t-distribution depends on the degrees of freedom df.**
5. **As the df increases, t-distribution approaches to $N(0,1)$.**



- **Applet: t-distribution vs $N(0,1)$**

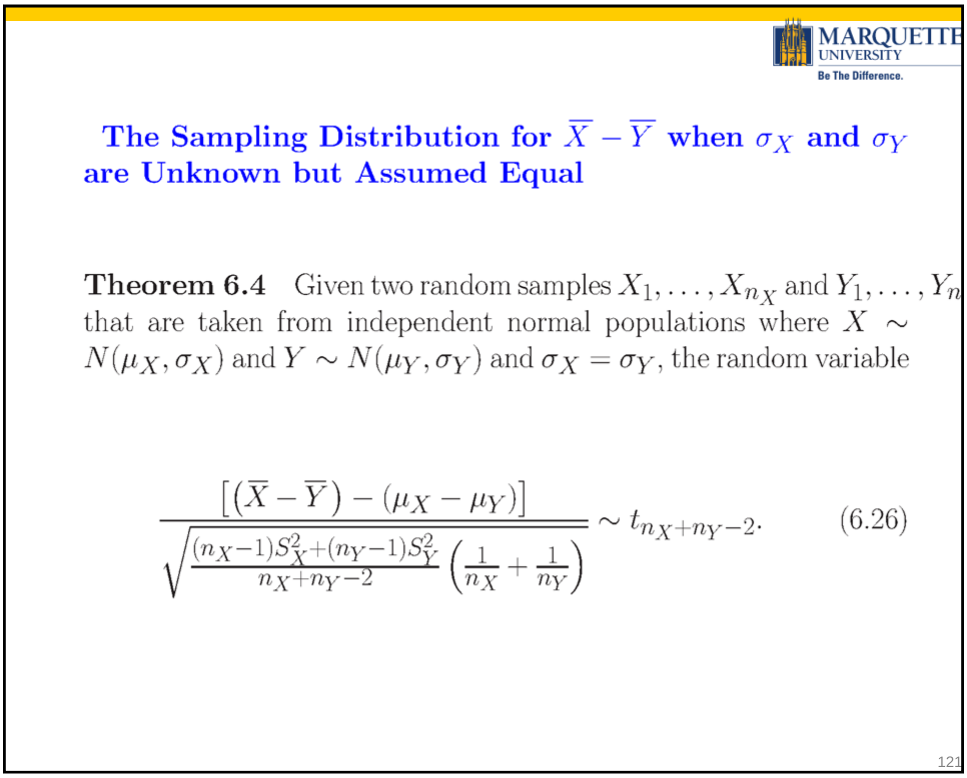


Example 6.24 The tensile strength for a type of wire is normally distributed with an unknown mean μ and an unknown variance σ^2. Five pieces of wire are randomly selected from a large roll, and the strength of each segment of wire is measured. Find the probability that $\overline{Y}$ will be within $\frac{2S}{\sqrt{n}}$ of the true population mean, μ.

Solution: The solution is:

$$\begin{aligned}\mathbb{P}\left(\mu - \frac{2S}{\sqrt{n}} \le \overline{Y} \le \mu + \frac{2S}{\sqrt{n}}\right) &= \mathbb{P}\left(-\frac{2S}{\sqrt{n}} \le \overline{Y} - \mu \le \frac{2S}{\sqrt{n}}\right)\\ &= \mathbb{P}\left(-2 \le \frac{\overline{Y} - \mu}{S/\sqrt{n}} \le 2\right)\\ &= \mathbb{P}\left(-2 \le t_4 \le 2\right) = 0.8838835.\end{aligned}$$

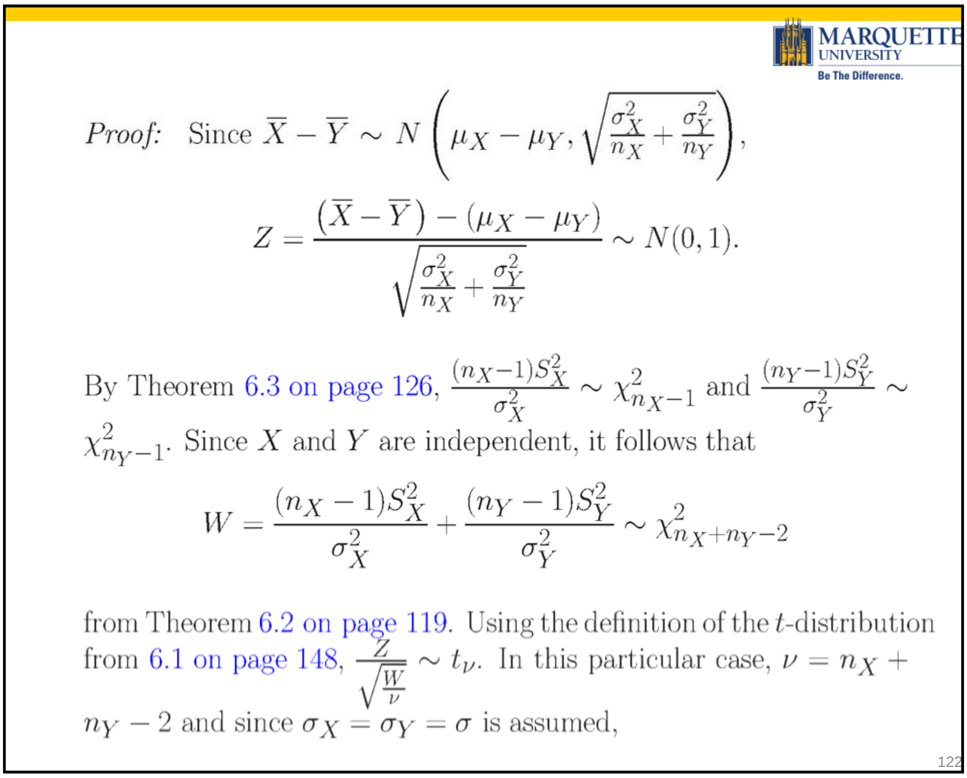
Note that if σ were known, $\mathbb{P}\left(-2 \le Z \le 2\right) = 0.9544$ ■



The Sampling Distribution for $\overline{X} - \overline{Y}$ when σ_X and σ_Y are Unknown but Assumed Equal

Theorem 6.4 Given two random samples $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_n$ that are taken from independent normal populations where $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$ and $\sigma_X = \sigma_Y$, the random variable

$$\frac{\left[\left(\overline{X} - \overline{Y}\right) - (\mu_X - \mu_Y)\right]}{\sqrt{\frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim t_{n_X+n_Y-2}. \qquad (6.26)$$



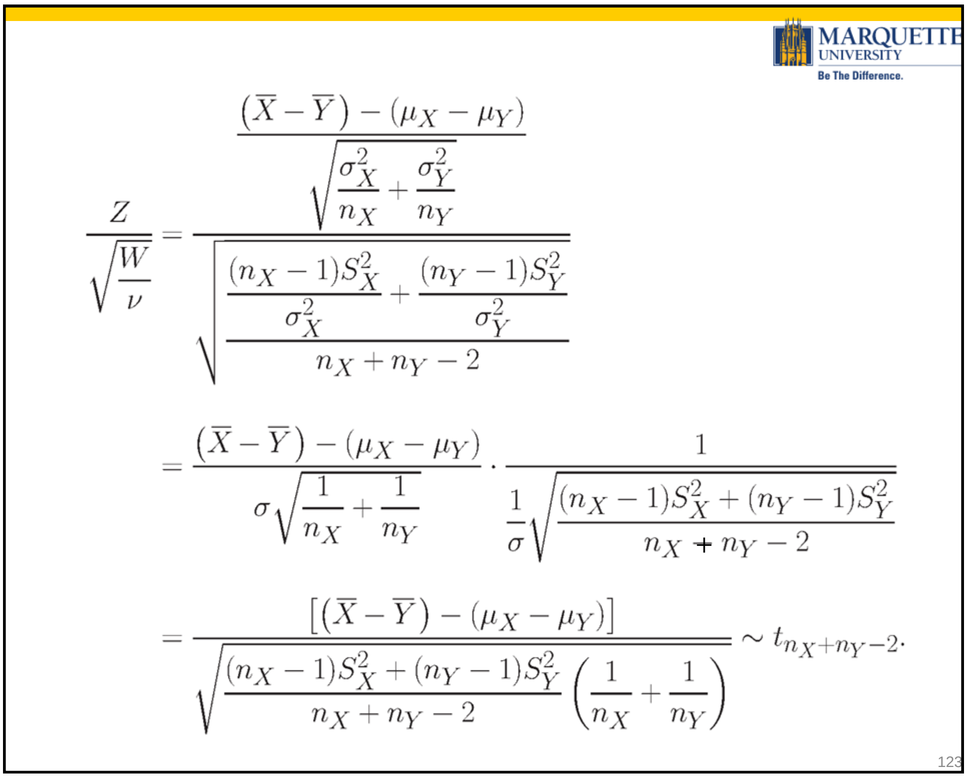
Proof: Since $\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right)$,

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0,1).$$

By Theorem 6.3 on page 126, $\frac{(n_X-1)S_X^2}{\sigma_X^2} \sim \chi^2_{n_X-1}$ and $\frac{(n_Y-1)S_Y^2}{\sigma_Y^2} \sim \chi^2_{n_Y-1}$. Since X and Y are independent, it follows that

$$W = \frac{(n_X-1)S_X^2}{\sigma_X^2} + \frac{(n_Y-1)S_Y^2}{\sigma_Y^2} \sim \chi^2_{n_X+n_Y-2}$$

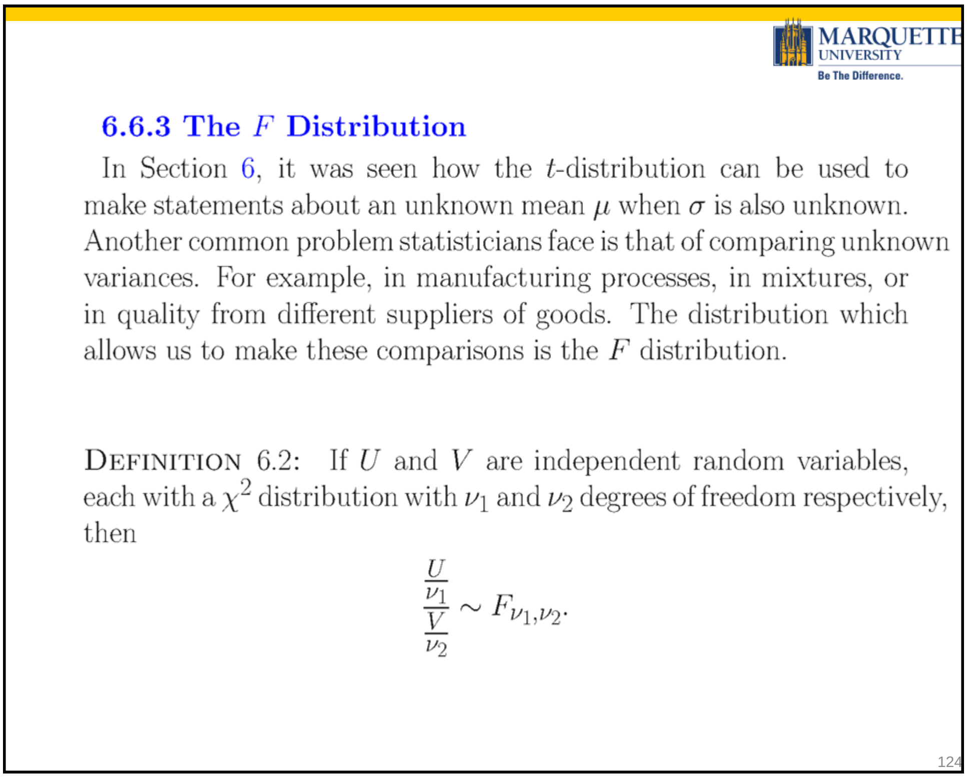
from Theorem 6.2 on page 119. Using the definition of the t-distribution from 6.1 on page 148, $\frac{Z}{\sqrt{\frac{W}{\nu}}} \sim t_\nu$. In this particular case, $\nu = n_X + n_Y - 2$ and since $\sigma_X = \sigma_Y = \sigma$ is assumed,



$$\frac{Z}{\sqrt{\frac{W}{\nu}}} = \frac{\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}}{\sqrt{\frac{\frac{(n_X-1)S_X^2}{\sigma_X^2} + \frac{(n_Y-1)S_Y^2}{\sigma_Y^2}}{n_X+n_Y-2}}}$$

$$= \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \cdot \frac{1}{\frac{1}{\sigma}\sqrt{\frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2}}}$$

$$= \frac{[(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)]}{\sqrt{\frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim t_{n_X+n_Y-2}.$$

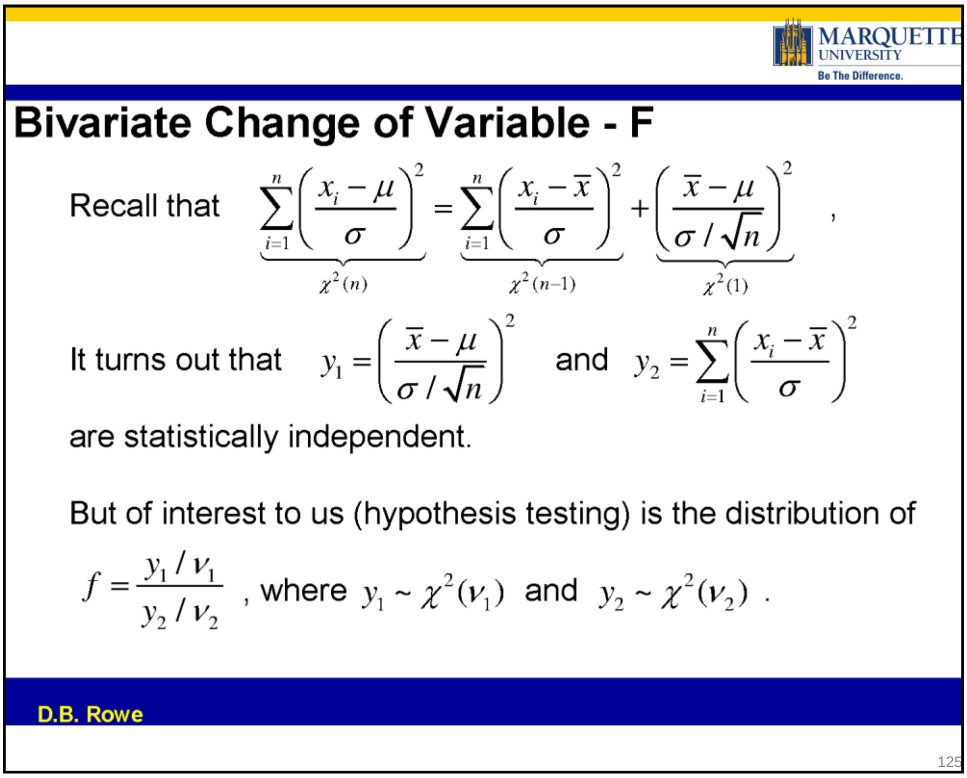


6.6.3 The F Distribution

In Section 6, it was seen how the t-distribution can be used to make statements about an unknown mean μ when σ is also unknown. Another common problem statisticians face is that of comparing unknown variances. For example, in manufacturing processes, in mixtures, or in quality from different suppliers of goods. The distribution which allows us to make these comparisons is the F distribution.

DEFINITION 6.2: If U and V are independent random variables, each with a χ^2 distribution with ν_1 and ν_2 degrees of freedom respectively, then

$$\frac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}} \sim F_{\nu_1,\nu_2}.$$



Bivariate Change of Variable - F

Recall that $\underbrace{\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}_{\chi^2(n)} = \underbrace{\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2}_{\chi^2(n-1)} + \underbrace{\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2}_{\chi^2(1)}$,

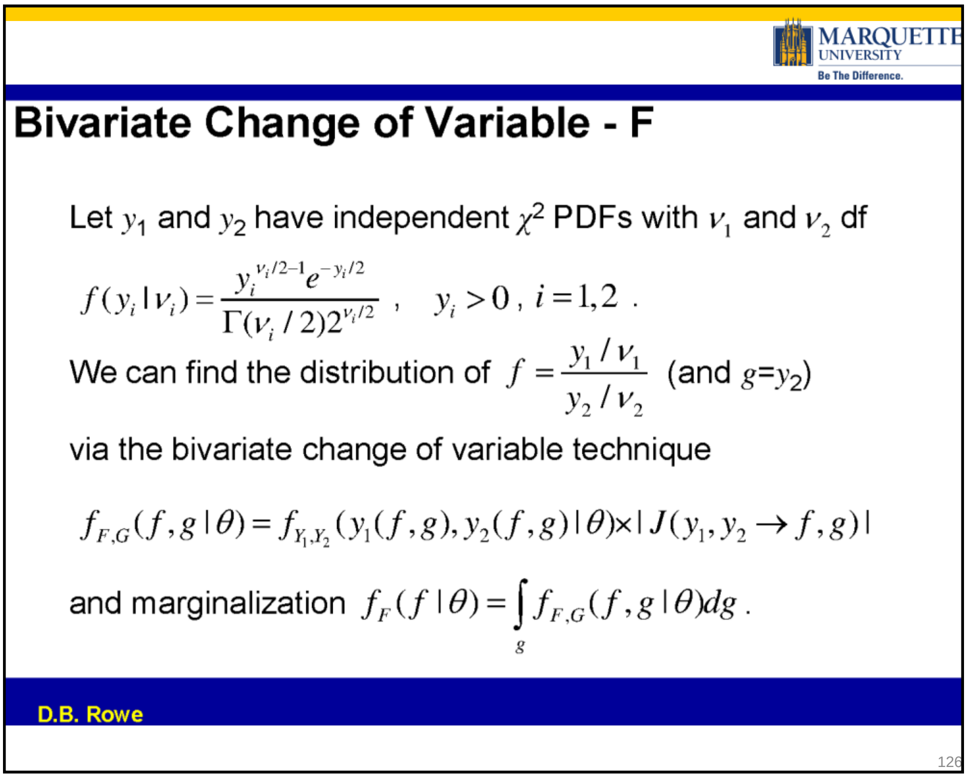
It turns out that $y_1 = \left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2$ and $y_2 = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$

are statistically independent.

But of interest to us (hypothesis testing) is the distribution of

$f = \frac{y_1/\nu_1}{y_2/\nu_2}$, where $y_1 \sim \chi^2(\nu_1)$ and $y_2 \sim \chi^2(\nu_2)$.

D.B. Rowe



Bivariate Change of Variable - F

Let y_1 and y_2 have independent χ^2 PDFs with ν_1 and ν_2 df

$$f(y_i \mid \nu_i) = \frac{y_i^{\nu_i/2-1} e^{-y_i/2}}{\Gamma(\nu_i/2) 2^{\nu_i/2}} \ , \quad y_i > 0 \ , \ i = 1,2 \ .$$

We can find the distribution of $f = \dfrac{y_1/\nu_1}{y_2/\nu_2}$ (and g=y_2)

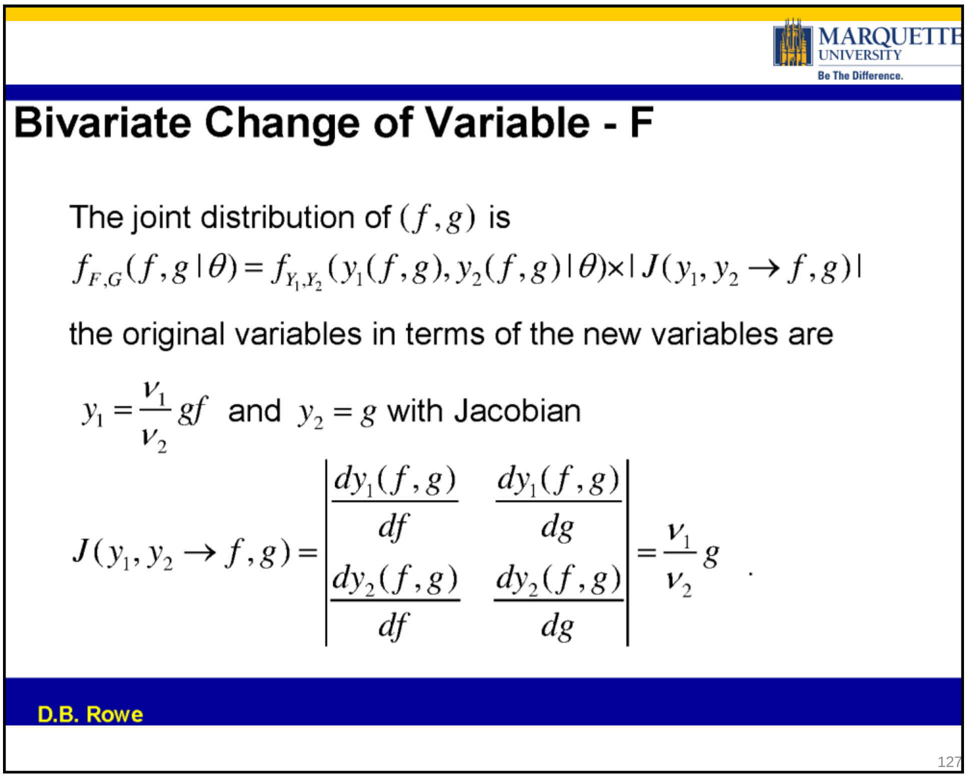
via the bivariate change of variable technique

$$f_{F,G}(f,g \mid \theta) = f_{Y_1,Y_2}(y_1(f,g), y_2(f,g) \mid \theta) \times | J(y_1,y_2 \to f,g) |$$

and marginalization $f_F(f \mid \theta) = \displaystyle\int_g f_{F,G}(f,g \mid \theta) dg$.

D.B. Rowe

126



Bivariate Change of Variable - F

The joint distribution of (f,g) is

$$f_{F,G}(f,g \mid \theta) = f_{Y_1,Y_2}(y_1(f,g), y_2(f,g) \mid \theta) \times | J(y_1,y_2 \to f,g) |$$

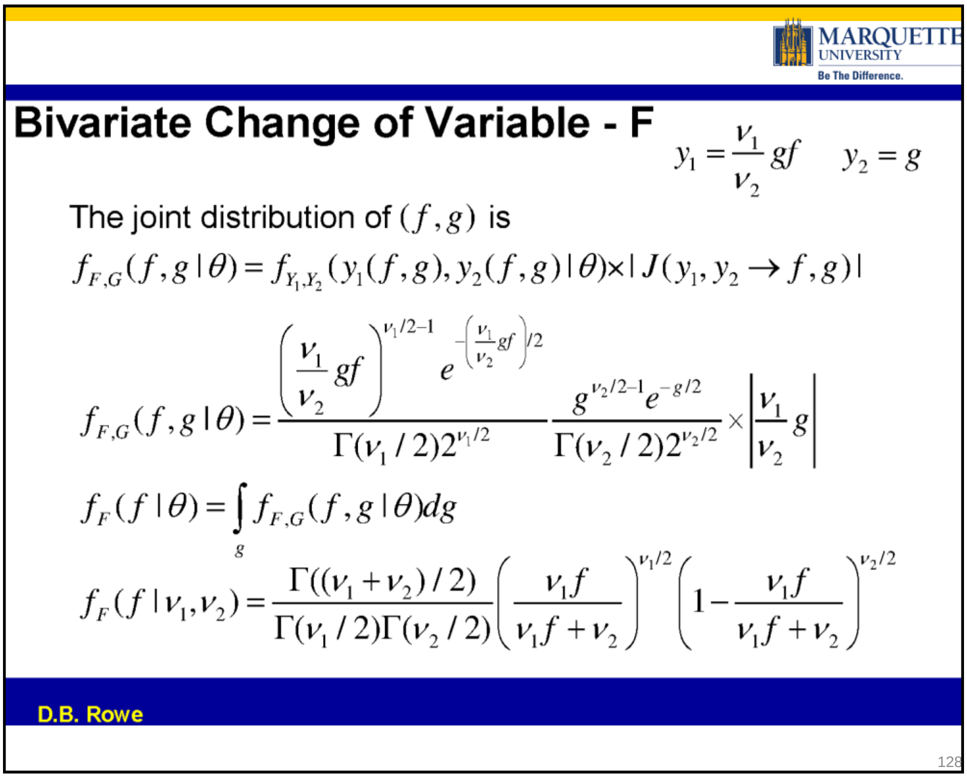
the original variables in terms of the new variables are

$y_1 = \dfrac{\nu_1}{\nu_2} gf$ and $y_2 = g$ with Jacobian

$$J(y_1,y_2 \to f,g) = \begin{vmatrix} \dfrac{dy_1(f,g)}{df} & \dfrac{dy_1(f,g)}{dg} \\ \dfrac{dy_2(f,g)}{df} & \dfrac{dy_2(f,g)}{dg} \end{vmatrix} = \frac{\nu_1}{\nu_2} g \quad .$$

D.B. Rowe

127



Bivariate Change of Variable - F

$$y_1 = \frac{\nu_1}{\nu_2} gf \qquad y_2 = g$$

The joint distribution of (f,g) is

$$f_{F,G}(f,g \mid \theta) = f_{Y_1,Y_2}(y_1(f,g), y_2(f,g) \mid \theta) \times | J(y_1,y_2 \to f,g) |$$

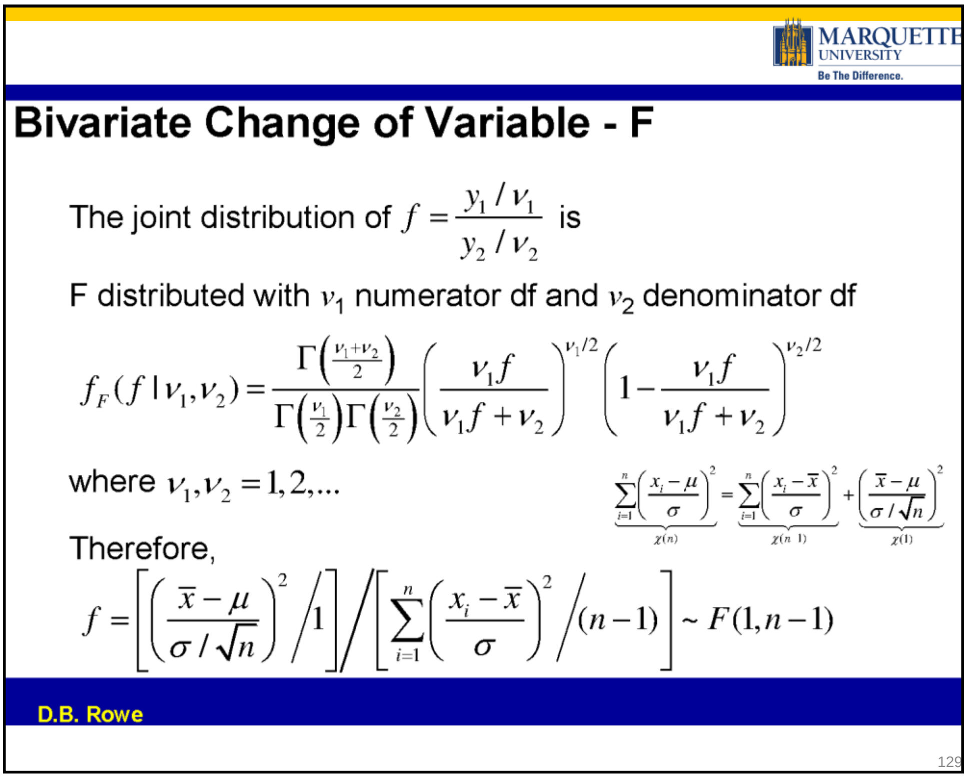
$$f_{F,G}(f,g \mid \theta) = \frac{\left(\frac{\nu_1}{\nu_2} gf\right)^{\nu_1/2-1} e^{-\left(\frac{\nu_1}{\nu_2} gf\right)/2}}{\Gamma(\nu_1/2) 2^{\nu_1/2}} \frac{g^{\nu_2/2-1} e^{-g/2}}{\Gamma(\nu_2/2) 2^{\nu_2/2}} \times \left| \frac{\nu_1}{\nu_2} g \right|$$

$$f_F(f \mid \theta) = \int_g f_{F,G}(f,g \mid \theta) dg$$

$$f_F(f \mid \nu_1,\nu_2) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1 f}{\nu_1 f + \nu_2}\right)^{\nu_1/2} \left(1 - \frac{\nu_1 f}{\nu_1 f + \nu_2}\right)^{\nu_2/2}$$

D.B. Rowe

128



Bivariate Change of Variable - F

The joint distribution of $f = \dfrac{y_1/\nu_1}{y_2/\nu_2}$ is

F distributed with ν_1 numerator df and ν_2 denominator df

$$f_F(f \mid \nu_1,\nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1 f}{\nu_1 f + \nu_2}\right)^{\nu_1/2} \left(1 - \frac{\nu_1 f}{\nu_1 f + \nu_2}\right)^{\nu_2/2}$$

where $\nu_1, \nu_2 = 1, 2, \ldots$

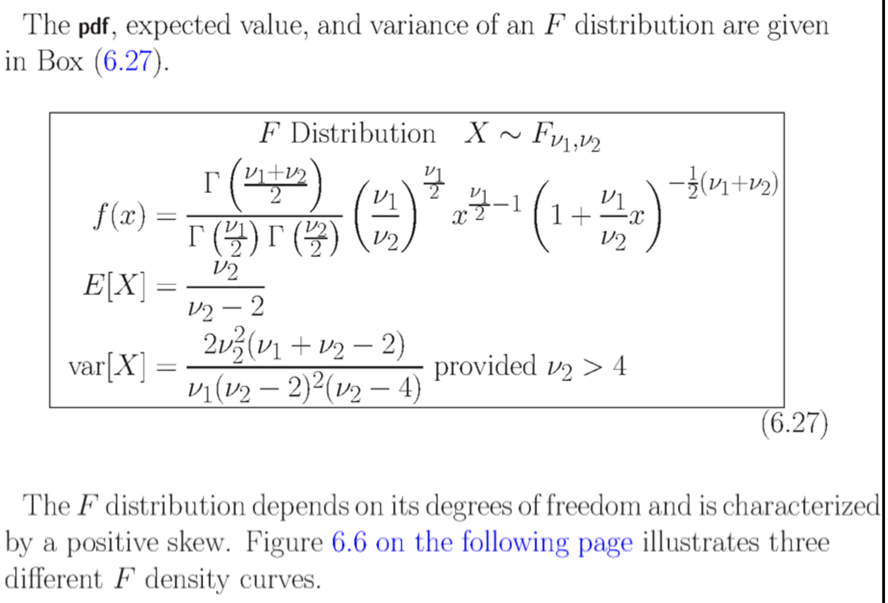
$$\underbrace{\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}_{\chi(n)} = \underbrace{\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2}_{\chi(n\ 1)} + \underbrace{\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2}_{\chi(1)}$$

Therefore,

$$f = \left[\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2 \Big/ 1\right] \Big/ \left[\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 \Big/ (n-1)\right] \sim F(1,n-1)$$

D.B. Rowe

129



The **pdf**, expected value, and variance of an F distribution are given in Box (6.27).

F Distribution $\quad X \sim F_{\nu_1,\nu_2}$

$$f(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}\left(1+\frac{\nu_1}{\nu_2}x\right)^{-\frac{1}{2}(\nu_1+\nu_2)}$$

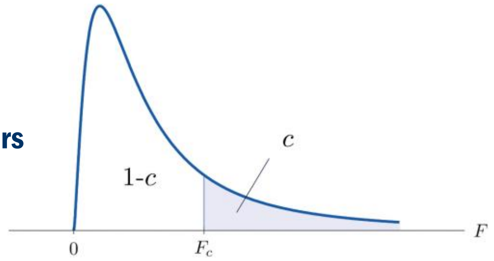
$$E[X] = \frac{\nu_2}{\nu_2-2}$$

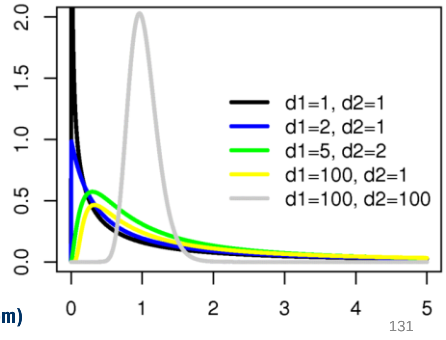
$$\text{var}[X] = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \text{ provided } \nu_2 > 4 \tag{6.27}$$

The F distribution depends on its degrees of freedom and is characterized by a positive skew. Figure 6.6 on the following page illustrates three different F density curves.

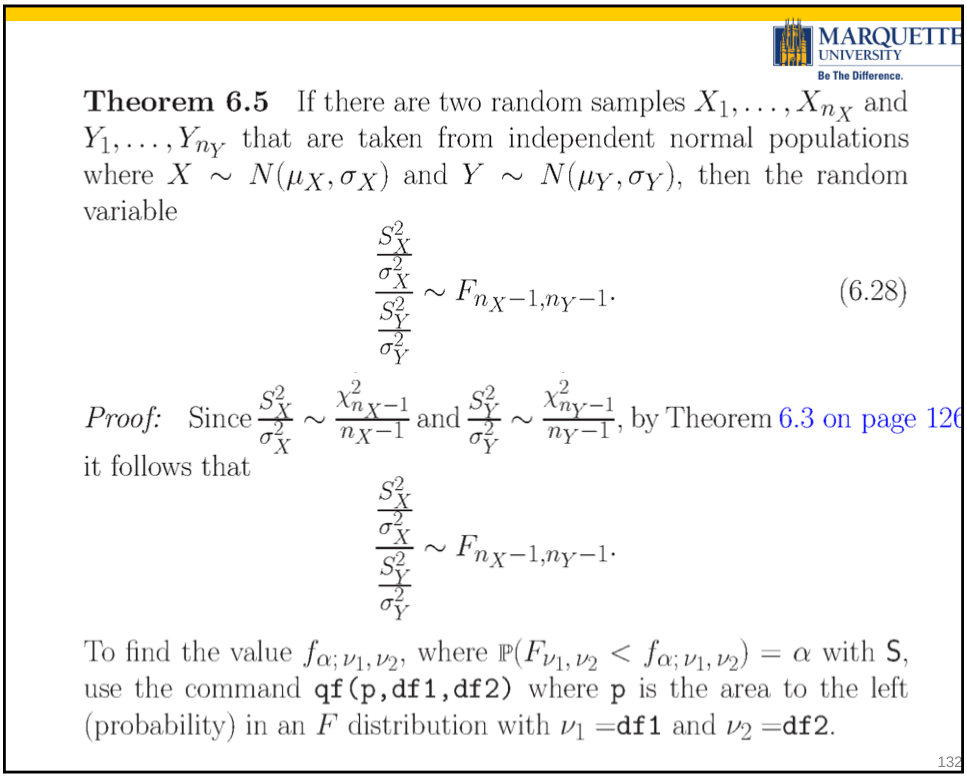
F DISTRIBUTION

- **Right skewed distribution**
- **Defined over positive numbers**
- **Parameters: df$_1$=ν_1, df$_2$=ν_2**
- **How to write:**
  - $F(\nu_1, \nu_2)$
- $F(\nu_1,\nu_2) = \frac{\frac{\chi^2(\nu_1)}{\nu_1}}{\frac{\chi^2(\nu_2)}{\nu_2}}$
- **F Calculator**
- **Ti-84: Fcdf(lower, upper, dfNumer, dfDenom)**





Source: Wikipedia



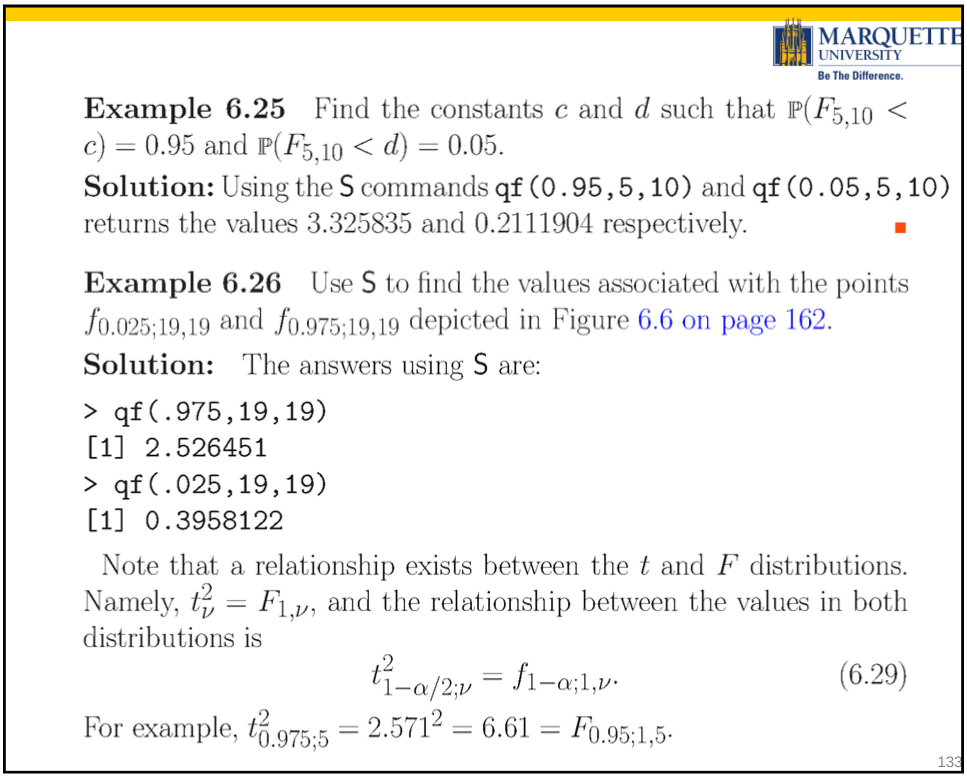
Theorem 6.5 If there are two random samples $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ that are taken from independent normal populations where $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$, then the random variable

$$\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \sim F_{n_X-1,n_Y-1}. \tag{6.28}$$

Proof: Since $\frac{S_X^2}{\sigma_X^2} \sim \frac{\chi^2_{n_X-1}}{n_X-1}$ and $\frac{S_Y^2}{\sigma_Y^2} \sim \frac{\chi^2_{n_Y-1}}{n_Y-1}$, by Theorem 6.3 on page 126 it follows that

$$\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \sim F_{n_X-1,n_Y-1}.$$

To find the value $f_{\alpha;\,\nu_1,\nu_2}$, where $\mathbb{P}(F_{\nu_1,\,\nu_2} < f_{\alpha;\,\nu_1,\,\nu_2}) = \alpha$ with S, use the command `qf(p,df1,df2)` where `p` is the area to the left (probability) in an F distribution with $\nu_1 =$`df1` and $\nu_2 =$`df2`.



Example 6.25 Find the constants c and d such that $\mathbb{P}(F_{5,10} < c) = 0.95$ and $\mathbb{P}(F_{5,10} < d) = 0.05$.

Solution: Using the S commands `qf(0.95,5,10)` and `qf(0.05,5,10)` returns the values 3.325835 and 0.2111904 respectively. ■

Example 6.26 Use S to find the values associated with the points $f_{0.025;19,19}$ and $f_{0.975;19,19}$ depicted in Figure 6.6 on page 162.

Solution: The answers using S are:

```
> qf(.975,19,19)
[1] 2.526451
> qf(.025,19,19)
[1] 0.3958122
```

Note that a relationship exists between the t and F distributions. Namely, $t_\nu^2 = F_{1,\nu}$, and the relationship between the values in both distributions is

$$t^2_{1-\alpha/2;\nu} = f_{1-\alpha;1,\nu}. \tag{6.29}$$

For example, $t^2_{0.975;5} = 2.571^2 = 6.61 = F_{0.95;1,5}$.

Bivariate Change of Variable - F/Student-t

We just showed that

$$f=\frac{y_1/\nu_1}{y_2/\nu_2}\sim F(1,n-1) \text{ where } y_1=\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2 \text{ and } y_2=\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$$

Recall that we showed that

$$t=\frac{z}{\sqrt{y_2/(n-1)}}\sim t(n-1) \text{ where } z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \text{ and } y_2=\frac{(n-1)s^2}{\sigma^2}\,?$$

What this means is, when ν_1 =1, f=t^2 !

$$t^2=f=\left[\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2\Big/1\right]\Big/\left[\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2\Big/(n-1)\right]$$

Bivariate Change of Variable - normal, χ^2, t, F

Recap: u_1 and u_2~uniform(0,1) and independent

$$z_1=\sqrt{-2\ln(u_1)}\cos(2\pi u_2) \qquad z_2=\sqrt{-2\ln(u_1)}\sin(2\pi u_2)$$

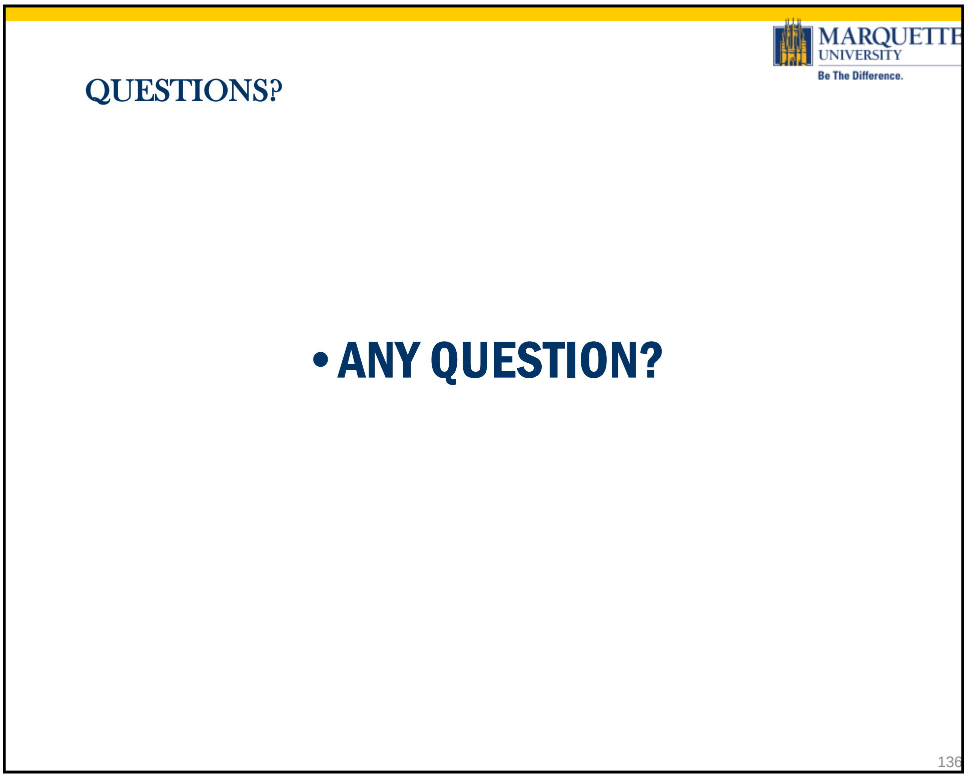
$z_1\sim N(0,1)$, $z_2\sim N(0,1)$, z_1 and z_2 are independent

$$x_i=\sigma z_i+\mu\sim N(\mu,\sigma^2)\,,\quad \bar{x}\sim N\left(\mu,\frac{\sigma^2}{n}\right)\,,\quad z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

$$y_1=\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2\sim\chi^2(1)\,,\quad y_2=\frac{(n-1)s^2}{\sigma^2}\sim\chi^2(n-1)$$

y_1 and y_2 are independent

$$t=\frac{z}{\sqrt{y_2/(n-1)}}\sim t(n-1)\,,\qquad f=\frac{y_1/1}{y_2/(n-1)}\sim F(1,n-1)\,.$$



MARQUETTE UNIVERSITY

Be The Difference.

QUESTIONS?

- ANY QUESTION?

136