Chapter 5

Sampling and Sampling Distributions

Overview

- Population and sample
- Random sampling
- Sampling distribution of sample mean
- Central Limit Theorem and its applications
- Sampling distribution of difference of two sample means
- Chi-square distribution
- Sampling distribution of $(n-1)S^2/\sigma^2$
- t-distribution
- F-distribution

5.1 Population and Sample

- The totality of all possible outcomes or observations of a survey or experiment is called a **population**.
- A **sample** is any subset of a population.
- Every outcome or observation can be recorded as a numerical or a categorical value. Thus each member of a population is a value of a random variable.
- There are two kinds of populations, namely, finite and infinite populations.

Finite Population

- A **finite population** consists of a finite number of elements.

- For instance, it can be
  1. all the citizens of Singapore, or
  2. all the books in the Science Library.

Infinite Population

An **infinite population** is one that consists of an infinitely (countable and uncountable) large number of elements. For instance:

1. The results of all possible rolls of a pair of dice.
2. Random digit numbers taken with replacement from a sample space of 10 digits, namely {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}.
3. The depths at all conceivable positions of a lake.

Remark

- Some finite populations are so large that in theory we assume them to be infinite, such as the population of lives of a certain type of storage battery being manufactured from a factory.

5.2 Random Sampling

5.2.1 Simple Random Sample

- A set of n members taken from a given population is called a **sample** of size n.
- A **simple random sample** of n members is a sample that is chosen in such a way that every subset of n observations of the population has the same probability of being selected.

Example

- We want to choose a simple random sample of size 5 from a group of 20 mice to be used in studying the growth rate of tumors in a cancer research experiment.
- We tag the mice with numbers from 1 to 20
- The following functions in Excel help us to select a random sample
  - "**=rand()**"
  - "**=rank(***cell_number***,** *range*, *order***)**"
    - order: 0 for descending and 1 for ascending
  - "**=index(***range*, *row_number***)**"

Example (Continued)

- Enter numbers 1 to 20 into a range of cells, let say, A1:A20
- Generate 20 random numbers using the function "**=rand()**" and store these numbers into a range of cells, let say, B1:B20.
- Copy and paste (values only) these random numbers into a range of cells, let say, C1:C20
- We select 5 numbers randomly from 1 to 20 by typing "**=index**(*$A$1:$A$20*,**rank(***C1,$C$1:$C$20,0***))**" in a range of cells, let say, D1:D5.

Example (Continued)

	A	B	C	D	E
1	Alex	0.625	0.167	19	Sam
2	Bill	0.229	0.462	11	Kenneth
3	Charles	0.495	0.504	9	Isaac
4	David	0.234	0.564	8	Hardy
5	Edward	0.254	0.81	5	Edward
6	Frank	0.439	0.276	16	
7	Gabriel	0.459	0.989	1	
8	Hardy	0.082	0.637	7	
9	Isaac	0.915	0.468	10	
10	John	0.307	0.353	13	
11	Kenneth	0.303	0.851	3	
12	Lucas	0.044	0.845	4	
13	Michael	0.22	0.193	18	
14	Nathan	0.484	0.334	14	
15	Oliver	0.108	0.417	12	
16	Peter	0.236	0.864	2	
17	Quinn	0.231	0.199	17	
18	Robert	0.35	0.324	15	
19	Sam	0.138	0.086	20	
20	Tom	0.197	0.785	6	

5.2.2 Sampling from a finite population

1. Sampling without replacement

- Given a population, say, {A, B, C, D}, we have the following ${}_4C_2 = 6$ possible samples of size 2.
- They are

$$\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}.$$

- Here the order of the letters is disregarded. It is the case of combinations.

Sampling from a finite population (Continued)

1. Sampling without replacement (Continued)

- In general, there are ${}_N C_n$ samples of size n that can be drawn from a finite population of size N **without replacement**.
- Each sample has equal chance of being selected. Hence each sample has a probability of $\frac{1}{{}_N C_n}$, of being selected.

Sampling from a finite population (Continued)

2. Sampling with replacement

- Using the same population, {A, B, C, D}, there are $4^2 = 16$ samples of size 2 as follows:

  (A, A), (A, B), (A, C), (A, D), (B, A), (B, B), (B, C), (B, D),
  (C, A), (C, B), (C, C), (C, D), (D, A), (D, B), (D, C), (D, D).

Sampling from a finite population (Continued)

2. Sampling with replacement

- Here the order of the letters is taken into consideration. Hence, (A, B) and (B, A) are considered as two different samples. (Why?)
- In general, there are N^n samples of size n that can be drawn from a finite population of size N **with replacement**. Therefore, each sample has the same probability, $1/N^n$, of being selected.

5.2.3 Sampling from an Infinite Population (with or without replacement)

- When lists are available and items are readily numbered, it is easy to draw random samples from finite populations.
- Unfortunately, it is often impossible to proceed in the way we have just described for an infinite population.
- The concept of a random sample from an infinite population is more difficult to explain.
- We use examples to show the characteristics of such a sample.

Example 1

- Suppose we consider the results of 15 tosses of a coin as a sample from the hypothetically infinite population which consists of the results of all possible tosses of the coin.

- If the probability of getting heads is the same for each toss and 15 tosses are independent, we say that the sample is random.

Example 2

- We would also be sampling from an infinite population if we sample with replacement from a finite population, and our sample would be random if

  (1) in each draw all elements of the population have the **same** probability of being selected, and

  (2) successive draws are **independent**.

Example 3

- Consider the population of the sums of all possible rolls of a pair of dice.
- Here the population is considered to be infinite.
- That is, we choose a random sample of size n from a random variable X having probability function given by

x	2	3	4	5	6	7
$f_X(x) = \Pr(X = x)$	1/36	2/36	3/36	4/36	5/36	6/36

x	8	9	10	11	12
$f_X(x) = \Pr(X = x)$	5/36	4/36	3/36	2/36	1/36

Example 3 (Continued)

- To obtain a random sample of size of 100, we simply roll the pair of dice 100 times independently under the same conditions.
- If X_i represents the result on the i-th roll, we then obtain $X_1, X_2, \cdots, X_{100}$.

Example 3 (Continued)

- All the $X_1, X_2, \cdots, X_{100}$ are the random variables which have the **same** distribution as the population random variable X, and evidently $X_1, X_2, \cdots, X_{100}$ are **independent** such that

$$\begin{aligned}
&\Pr(X_1 = x_1, X_2 = x_2, \cdots, \text{and } X_{100} = x_{100}) \\
&= \Pr(X_1 = x_1)\Pr(X_2 = x_2)\cdots\Pr(X_{100} = x_{100}) \\
&= \prod_{i=1}^{100} \Pr(X_i = x_i)
\end{aligned}$$

Definition 5.1

- Let X be a random variable with certain probability distribution, $f_X(x)$.
- Let $X_1, X_2, \cdots, X_n$ be n independent random variables each having the same distribution as X,
- then $(X_1, X_2, \cdots, X_n)$ is called a random sample of size n from a population with distribution $f_X(x)$.
- The joint p.f. (or p.d.f.) of $(X_1, X_2, \cdots, X_n)$ is given by

$$f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n),$$

where $f_X(x)$ is the p.f. (or p.d.f.) of the population.

5.3 Sampling distribution of sample mean

5.3.1 Introduction

- Our main purpose in selecting random samples is to elicit information about the unknown population parameters.
- For instance, we wish to know the proportion of people in Singapore who prefer a certain brand of coffee.
- A large random sample is then selected from the population and the proportion of this sample favouring the brand of coffee in question is calculated.
- This value is now used to make some inference concerning the true proportion in the population.

5.3.2 Statistic and Sampling Distribution

- A function of a random sample $(X_1, X_2, \cdots, X_n)$ is called a **statistic**. (e.g. $\bar{X}$ is a statistic as $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$)
- Hence a statistic is a random variable. It is meaningful to consider the probability distribution of a statistic.
- The probability distribution of a statistic is called a **sampling distribution**.

Sample Mean

- If $X_1, X_2, \cdots, X_n$ represent a random sample of size n, then the sample mean is defined by the statistic

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

- If the values in a random sample are observed and they are $x_1, x_2, \cdots, x_n$, then the realization of the statistic $\bar{X}$ is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example 1

- To study the sampling distribution of the sample mean, we consider a discrete uniform population consisting of the values

$$\{3, 5, 7, 9, 11\}$$

with the population size $N = 5$.

- Hence

$$f_X(x) = \frac{1}{5}, \quad \text{for } x = 3, 5, 7, 9, 11$$

Example 1 (Continued)

- The population mean

$$\mu_X = E(X) = \sum_x x\, f_X(x)$$

$$= \frac{1}{5}(3 + 5 + 7 + 9 + 11) = 7.$$

Example 1 (Continued)

- The population variance

$$\sigma_X^2 = V(X) = \sum_x (x - \mu_X)^2 f_X(x)$$

$$= \frac{1}{5}[(3-7)^2 + (5-7)^2 + \cdots + (11-7)^2] = 8.$$

Example 1 (Continued)

- Suppose we list all possible samples of size 2 with replacement, and then for each sample we compute $\bar{X}$.
- There are $5^2 = 25$ possible distinct samples and their means are as follows:

Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$	Sample	$\bar{X}$
(3, 3)	3	(5, 3)	4	(7, 3)	5	(9, 3)	6	(11, 3)	7
(3, 5)	4	(5, 5)	5	(7, 5)	6	(9, 5)	7	(11, 5)	8
(3, 7)	5	(5, 7)	6	(7, 7)	7	(9, 7)	8	(11, 7)	9
(3, 9)	6	(5, 9)	7	(7, 9)	8	(9, 9)	9	(11, 9)	10
(3,11)	7	(5,11)	8	(7,11)	9	(9,11)	10	(11,11)	11

Example 1 (Continued)

- The sampling distribution of $\bar{X}$ is now found to be:

$\bar{x}$	3	4	5	6	7
$\Pr(\bar{X} = \bar{x})$	1/25	2/25	3/25	4/25	5/25

$\bar{x}$	8	9	10	11
$\Pr(\bar{X} = \bar{x})$	4/25	3/25	2/25	1/25

Example 1 (Continued)

$$\mu_{\bar{X}} = E(\bar{X}) = \sum_x \bar{x}\, f_{\bar{X}}(\bar{x}) = 3\left(\frac{1}{25}\right) + 4\left(\frac{2}{25}\right) + \cdots + 11\left(\frac{1}{25}\right) = 7.$$

$$E(\bar{X}^2) = 3^2\left(\frac{1}{25}\right) + 4^2\left(\frac{2}{25}\right) + \cdots + 11^2\left(\frac{1}{25}\right) = 53.$$

- Hence

$$\sigma_{\bar{X}}^2 = V(\bar{X}) = E(\bar{X}^2) - \left(E(\bar{X})\right)^2 = 53 - 7^2 = 4.$$

- Therefore we have $\mu_{\bar{X}} = \mu_X$ and $\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{2}$ where 2 is the sample size.

Sampling Distribution of Sample Mean

Theorem 5.1

- For random samples of size n taken from an **infinite population** or from a **finite population with replacement** having population mean μ and population standard deviation σ,
- the **sampling distribution of the sample mean** $\bar{X}$ has its mean and variance given by

$$\mu_{\bar{X}} = \mu_X \quad \text{and} \quad \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}.$$

Sampling Distribution of Sample Mean (Continued)

Theorem 5.1 (Continued)

That is,

$$E(\bar{X}) = E(X) \quad \text{and} \quad V(\bar{X}) = \frac{V(X)}{n}.$$

Law of Large Number (LLN)

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a population having any distribution with mean μ and **finite** population variance σ^2. Then for any $\epsilon \in \mathbb{R}$,

$$P(|\bar{X} - \mu| > \epsilon) \longrightarrow 0 \text{ as } n \longrightarrow \infty.$$

Law of Large Number (LLN) (Continued)

REMARK:

This says that as the sample size increases, the probability that the sample mean differs from the population mean goes to zero.

Another way of looking at this is that it is increasingly likely that $\bar{X}$ is close to μ, as n gets larger.

Law of Large Number (LLN) (Continued)

Example

Let $X_1, \cdots, X_n$ be a random sample from $U(0,1)$.

Then $E(X) = 1/2$ and $V(X) = 1/12$

Take $n = 3(10^6)$ and $\epsilon = 0.001$.

Let $\bar{X}_n$ be the sample mean from a sample of size $3(10)^6$.

Hence, $E(\bar{X}) = \frac{1}{2}$ and $V(\bar{X}) = \frac{1/12}{3(10^6)}$

Law of Large Number (LLN) (Continued)

Example (Continued)

Consider $\Pr(|\bar{X}_n - 0.5| > 0.001)$.

Find k such that $k\sqrt{V(\bar{X})} = 0.001$ or $k = 0.001\left(6(10^3)\right) = 6$

$$\Pr(|\bar{X}_n - 0.5| > 0.001) = \Pr\left(|\bar{X}_n - 0.5| > 6\sqrt{V(\bar{X})}\right)$$

By Chebyshev's Inequality,

$$\Pr\left(|\bar{X}_n - 0.5| > 6\sqrt{V(\bar{X})}\right) \leq \frac{1}{6^2} = 0.02778$$

5.4 Central Limit Theorem and its applications

Central Limit Theorem

- Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a population having any distribution with mean μ and finite population variance σ^2.
- The sampling distribution of the sample mean $\bar{X}$ is **approximately normal** with mean μ and variance σ^2/n if ***n* is sufficiently large**.
- Hence

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ follows approximately } N(0,1)$$

Central Limit Theorem (Continued)

Sampling distribution properties of $\bar{X}$:

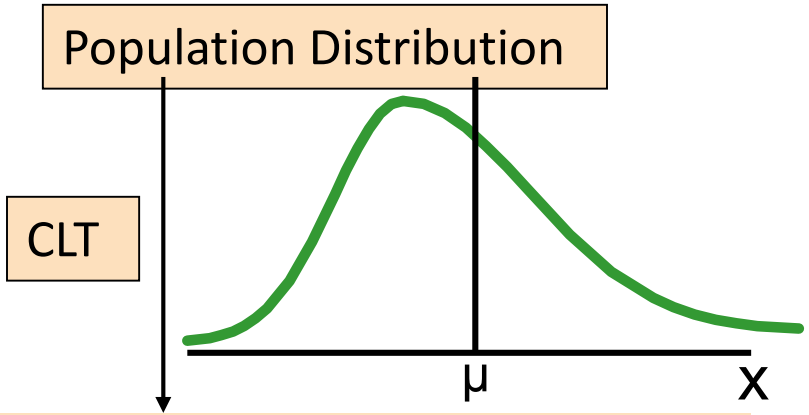
Central Tendency

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

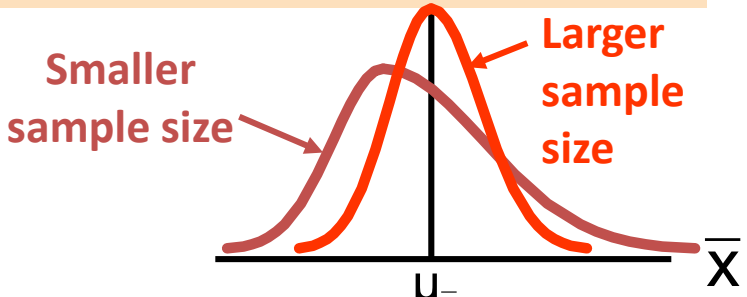
Variation

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

(Sampling with replacement)



Sampling Distribution of $\bar{X}$
(becomes normal as n increases)



Theorem 5.2

- If $X_i, i = 1, 2, \cdots, n$ are $N(\mu, \sigma^2)$, then $\bar{X}$ is $N(\mu, \frac{\sigma^2}{n})$ regardless of the sample size n.

- Similarly, if $X_i, i = 1, 2, \cdots, n$ are approximately $N(\mu, \sigma^2)$, then $\bar{X}$ is approximately $N(\mu, \frac{\sigma^2}{n})$ regardless of the sample size n.

Useful website:

http://onlinestatbook.com/stat_sim/index.html

Example 1

- An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed,

- with mean equal to 800 hours and a standard deviation of 40 hours.

- Find the probability that a random sample of 16 light bulbs will have an average life of less than 775 hours.

Solution to Example 1

- Let X be the lifetime of a light bulb.
- It is given that X approx $\sim N(800, 40^2)$.
- Let $\bar{X}$ be the sample mean of size 16.
- Then $\bar{X}$ approx $\sim N(800, 40^2/16)$.
- Therefore

$$\Pr(\bar{X} < 775) = \Pr\left(\frac{\bar{X} - 800}{40/\sqrt{16}} < \frac{775 - 800}{40/\sqrt{16}}\right) \approx \Pr(Z < -2.5)$$

$$= \Pr(Z > 2.5) = 0.00621.$$

where $Z \sim N(0,1)$

Example 2

- Let $\bar{X}$ denote the mean of a random sample of size 75 from the distribution which has the p.d.f.

$$f_X(x) = 1, \qquad \text{for } 0 < x < 1.$$

- Find $\Pr(0.45 < \bar{X} < 0.55)$.

Solution to Example 2

- It is known that $E(X) = 1/2$, and $V(X) = 1/12$.
- By Central Limit Theorem,

$$\bar{X} \text{ approx} \sim N\left(\frac{1}{2}, \frac{1/12}{75}\right)$$

- Hence

$$\Pr(0.45 < \bar{X} < 0.55) \approx \Pr\left(\frac{0.45 - 0.5}{1/30} < Z < \frac{0.55 - 0.5}{1/30}\right)$$

$$= \Pr(-1.5 < Z < 1.5) = 1 - 2\Pr(Z > 1.5) = 0.8664.$$

Example 3

- A random sample of size 50 is drawn from a Poisson distribution with parameter $\lambda = 0.03$.

- What is the probability that the sum of the sample will be at least 3?

Solution to Example 3

- Let X has a Poisson distribution with parameter $\lambda = 0.03$.
- Then $\mu = \sigma^2 = 0.03$. (Why?)
- Let $S_{50} = \sum_{i=1}^{50} X_i$
- The desired probability is

$$\Pr(S_{50} \geq 3) = \Pr(S_{50} > 2.5) = \Pr\left(\bar{X} > \frac{2.5}{50}\right)$$

$$= \Pr\left(\frac{\bar{X} - 0.03}{\sqrt{0.03/50}} > \frac{0.05 - 0.03}{\sqrt{0.03/50}}\right) \approx \Pr(Z > 0.82) = 0.2061.$$

Note: Continuity correction is used.

Example 4

- The nicotine content in a single cigarette of a particular brand is a random variable with mean $\mu = 0.8$ mg and standard deviation 0.1 mg.
- If an individual smokes five packs (20 cigarettes per pack) of these cigarettes per week,
- what is the probability that the total amount of nicotine consumed in a week is at least 82 mg?

Solution to Example 4

Solution

- 5 packs consist of 100 cigarettes.
- Let $X_i, i = 1, \cdots, 100$ denote the nicotine contents of the 100 cigarettes.
- Then X_i's form a random sample from a distribution with mean $\mu = 0.8$ and $\sigma = 0.1$.

Solution to Example 4 (Continued)

- Applying the **Central Limit Theorem**, we have

$$\bar{X} \text{ approx} \sim N\left(0.8, \frac{0.1^2}{100}\right)$$

- Hence

$$\Pr\left(\sum_{i=1}^{100} X_i \geq 82\right) = \Pr(\bar{X} \geq 0.82) \approx \Pr\left(Z \geq \frac{0.82 - 0.8}{0.01}\right)$$

$$= \Pr(Z \geq 2) = 0.0228.$$

Example 5

- When a batch of a certain chemical product is prepared, the amount of a particular impurity in the batch is a random variable with mean value 4.0 g and standard deviation 1.5 g.

- If 50 batches are independently prepared, what is the approximate probability that the sample average amount of impurity is between 3.5 g and 3.8 g?

Solution to Example 5

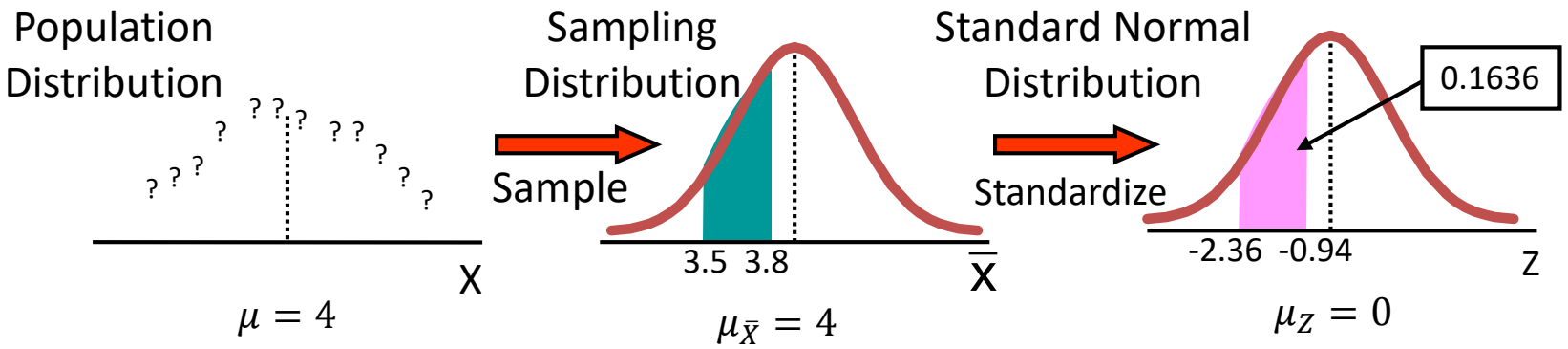
- Since n is large, we apply the **Central Limit Theorem** and we have

$$\bar{X} \text{ approx} \sim N\left(4, \frac{1.5^2}{50}\right)$$

- Hence

$$\begin{aligned}\Pr(3.5 \leq \bar{X} \leq 3.8) &\approx \Pr\left(\frac{3.5-4.0}{1.5/\sqrt{50}} \leq Z \leq \frac{3.8-4.0}{1.5/\sqrt{50}}\right) \\ &= \Pr(-2.357 \leq Z \leq -0.943) \\ &= \Pr(Z \leq -0.943) - \Pr(Z \leq -2.357) \\ &= 0.1636.\end{aligned}$$

Solution to Example 5 (Continued)



5.5 Sampling distribution of the difference of two sample means

Theorem 5.3

- If independent samples of sizes n_1 (≥ 30) and n_2 (≥ 30) are drawn from two populations,
- with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively,
- then the sampling distribution of the differences of the sample means, $\bar{X}_1$ and $\bar{X}_2$, is approximately normally distributed with mean and standard deviation given by

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Proof of Theorem 5.3

$$E(\bar{X}_1) = \mu_1, \qquad E(\bar{X}_2) = \mu_2,$$
$$V(\bar{X}_1) = \frac{\sigma_1^2}{n_1}, \qquad V(\bar{X}_2) = \frac{\sigma_2^2}{n_2}.$$

Hence

$$\mu_{\bar{X}_1 - \bar{X}_2} = E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

Proof of Theorem 5.3 (Continued)

and

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2),$$

since $\bar{X}_1$ and $\bar{X}_2$ are independent.

Therefore

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Proof of Theorem 5.3 (Continued)

- Since $\bar{X}_1$ and $\bar{X}_2$ are approximately normally distributed,

- therefore $\bar{X}_1 - \bar{X}_2$ is also approximately normally distributed.

Remarks

1. Note that if both n_1 and n_2 are greater than or equal to 30, the normal approximation for the distribution of $\bar{X}_1 - \bar{X}_2$ is very good regardless of the shapes of the two population distributions.
2. $$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ approx} \sim N(0,1).$$

Example 1

- The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 year,
- while those of manufacturer B have a mean lifetime of 6.0 years and a standard deviation of 0.8 year.
- What is the probability that a random sample of 36 tubes from manufacturer A will have a mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer B?

Solution to Example 1

- We are given the following information:

  $\mu_A = 6.5, \mu_B = 6.0, \sigma_A = 0.9, \sigma_B = 0.8, n_A = 36$ and $n_B = 49$.
- Using the above theorem, $\bar{X}_A - \bar{X}_B$ will have an approximate normal distribution
- with mean $= 6.5 - 6.0 = 0.5$ and
- standard deviation $= \sqrt{(0.81/36) + (0.64/49)} = 0.189$.
- The desired probability is

$$\Pr(\bar{X}_A - \bar{X}_B \geq 1) \approx \Pr\left(Z \geq \frac{1 - 0.5}{0.189}\right) = 0.00402.$$

5.6 Chi-square distribution

Definition 5.3

- If Y is a random variable with probability density function

$$f_Y(y) = \frac{1}{2^{n/2}\Gamma(n/2)}\, y^{n/2-1}\, e^{-y/2}, \qquad \text{for } y > 0,$$

  and 0 otherwise,

- then Y is defined to have a **chi-square distribution with n degrees of freedom**, denoted by $\chi^2(n)$,

  where n is a positive integer, and $\Gamma(\cdot)$ is the gamma function.

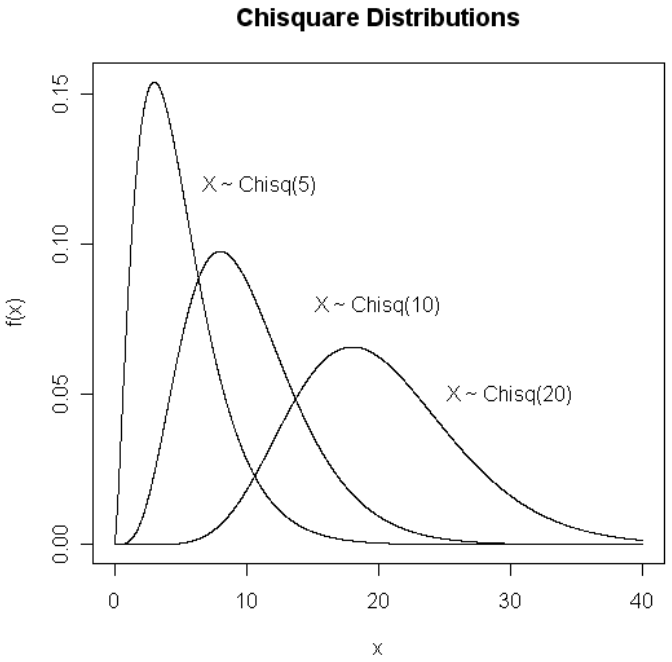
Chi-square distribution (Continued)

- The gamma function, $\Gamma(\cdot)$, is defined by

$$\Gamma(n) = \int_0^\infty x^{n-1}\, e^{-x}\, dx = (n-1)!$$

for $n = 1, 2, 3, \cdots$

- The p.d.f. of χ^2-distribution with $n = 5, 10$ and 25 are shown in the graph on the right side.



Some properties of Chi-square distributions

1. If $Y \sim \chi^2(n)$, then $E(Y) = n$ and $V(Y) = 2n$.
2. For large n, $\chi^2(n)$ approx $\sim N(n, 2n)$.
3. If $Y_1, Y_2, \cdots, Y_k$ are independent chi-square random variables with $n_1, n_2, \cdots, n_k$ degrees of freedom respectively, then

   $Y_1 + Y_2 + \cdots + Y_k$ has a chi-square distribution with $n_1 + n_2 + \cdots + n_k$ degrees of freedom. That is,

$$\sum_{i=1}^{k} Y_i \sim \chi^2\left(\sum_{i=1}^{k} n_i\right)$$

Theorem 5.5

1. If $X \sim N(0,1)$, then $X^2 \sim \chi^2(1)$.
2. Let $X \sim N(\mu, \sigma^2)$, then $[(X-\mu)/\sigma]^2 \sim \chi^2(1)$.
3. Let $X_1, X_2, \cdots, X_n$ be a random sample from a normal population with mean μ, and variance σ^2. Define

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

Then $Y \sim \chi^2(n)$

Use of the χ^2-distribution Table

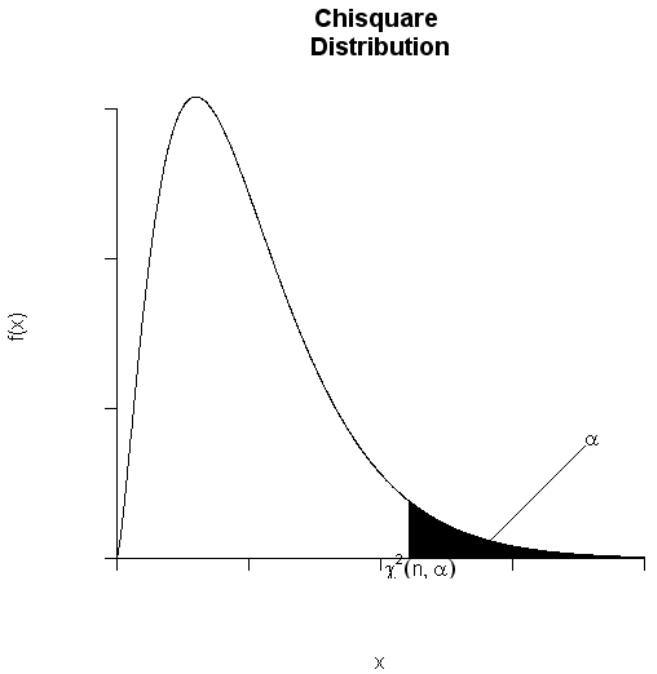
- Let c be a constant satisfying

$$\Pr(Y \geq c) = \int_c^{\infty} f_Y(y)\, dy = \alpha.$$

where $Y \sim \chi^2(n)$.

- We use the notation $\chi^2(n;\alpha)$ to denote this constant c. That is,

$$\Pr(Y \geq \chi^2(n;\alpha)) = \int_{\chi^2(n;\alpha)}^{\infty} f_Y(y)\, dy = \alpha.$$

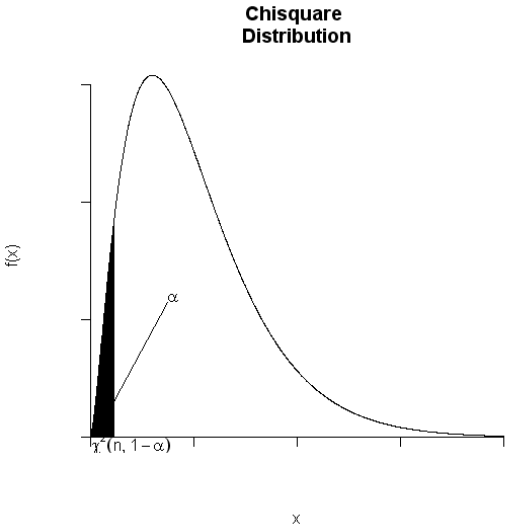


CYM

Use of the χ^2-distribution Table (Continued)

- Similarly, $\chi^2(n; 1-\alpha)$ is the constant satisfying

$$\Pr\left(Y \leq \chi^2(n; 1-\alpha)\right) = \int_0^{\chi^2(n;1-\alpha)} f_Y(y)\, dy = \alpha.$$



Use of the χ^2-distribution Table (Continued)

For example,

- $\chi^2(10; 0.9)$ means $\Pr\big(Y \geq \chi^2(10; 0.9)\big) = 0.9$ or $\Pr\big(Y \leq \chi^2(10; 0.9)\big) = 0.1$.
- From the statistical table on χ^2-distribution, we have $\chi^2(10; 0.9) = 4.865$.
- $\chi^2(10; 0.05)$ means $\Pr\big(Y \geq \chi^2(10; 0.05)\big) = 0.05$.
- From the statistical table on χ^2-distribution, we have $\chi^2(10; 0.05) = 18.307$.

Adobe Acrobat Document

Excel Functions for χ^2-distribution

- The values on previous slide can be obtained using by built-in functions in Microsoft Excel:
- "=CHISQ.INV(α; n)" gives c such that $\Pr(W \leq c) = \alpha$, where $W \sim \chi^2(n)$
- "=CHISQ.DIST(c; n; true)" gives α where $\Pr(W \leq c) = \alpha$.
- For example,

  We have "=CHISQ.INV(0.95; 10)" gives 18.30703 and

  "=CHISQ.DIST(18.30703; 10; true)" gives 0.95.

5.7 The sampling distribution of $(n-1)S^2/\sigma^2$

- Let $X_1, X_2, \cdots, X_n$ be a random sample from a population.
- Then the statistic

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

  is called the **sample variance**.
- The sampling distribution of the random variable S^2 has little practical application in statistics.
- Instead, we shall consider the sampling distribution of the random variable $\frac{(n-1)S^2}{\sigma^2}$ when $X_i \sim N(\mu, \sigma^2)$ for all i.

Theorem 5.5

- If S^2 is the variance of a random sample of size n taken from a **normal** population having the variance σ^2,
- then the random variable

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

- has a chi-square distribution with $n-1$ degrees of freedom. That is,

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

5.8 The *t*-distribution

Definition 5.4

- Suppose $Z \sim N(0,1)$ and $U \sim \chi^2(n)$.
- If Z and U are independent, and let

$$T = \frac{Z}{\sqrt{U/n}}$$

- then the random variable T follows the t-distribution with n degrees of freedom. That is,

$$\frac{Z}{\sqrt{U/n}} \sim t(n)$$

The p.d.f of a *t*-distribution

- If T follows a t-distribution with n degrees of freedom,
- then its p.d.f. is given by

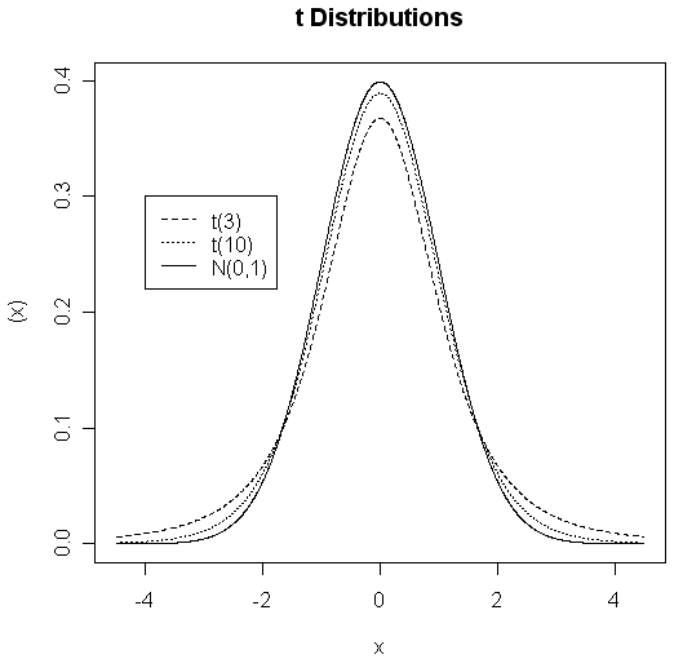
$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}, \qquad -\infty < t < \infty.$$

- The gamma function, $\Gamma(\cdot)$, is defined by

$$\Gamma(n) = \int_0^\infty x^{n-1}\, e^{-x}\, dx = (n-1)! \quad \text{for } n = 1, 2, 3, \cdots$$

Properties of a *t*-distribution

1. The graph of the *t*-distribution is symmetric about the vertical axis and resembles the graph of the standard normal distribution.



Properties of a *t*-distribution (Continued)

2. It can be shown that the p.d.f. of t-distribution with n d.f. is approaching to the p.d.f. of standard normal distribution when $n \to \infty$.

   That is

$$\lim_{n\to\infty} f_T(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2},$$

   as $n \to \infty$.

Properties of a *t*-distribution (Continued)

3. The values of

$$\Pr(T \geq t) = \int_t^{\infty} f_T(x)dx\,,$$

for selected values of n and t are given in a statistical table.

Adobe Acrobat Document

Properties of a *t*-distribution (Continued)

3. (Continued)

   For example,

   - $\Pr(T \geq t_{10;0.05}) = 0.05$ gives $t_{10;0.05} = 1.812$.
   - $\Pr(T \geq t_{10;0.01}) = 0.01$ gives $t_{10;0.01} = 2.602$.

4. If $T \sim t(n)$, then

   $E(T) = 0$ and $V(T) = n/(n-2)$ for $n > 2$.

Adobe Acrobat Document

Excel Functions for *t*-distribution

The values of $\Pr(T \leq t)$ can be obtained by the functions from Microsoft Excel:

"=T.DIST(t; n;TRUE)" gives c.d.f. $\Pr(T \leq t)$. FALSE gives p.d.f $f_T(t)$

"=T.INV(p; n)" gives p-th quantile of a $t(n)$ distribution

For examples

"=T.DIST(2; 10; TRUE)" gives 0.963306

"=T.INV(0.95; 10)" gives 1.812461

Remark

- If the random sample was selected from a normal population, then

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

and

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

- It can be shown that $\bar{X}$ and S^2 are independent, and so are Z and U.

Remark (Continued)

- Therefore,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}}$$

$$= \frac{Z}{\sqrt{U/(n-1)}} \sim t_{n-1},$$

- That is, T has a t-distribution with $n - 1$ d.f.

Example 1

- A manufacturer of light bulbs claims that his light bulbs will burn on the average 500 hours.
- To maintain this average, he tests 25 bulbs each month.
- If the computed t value (i.e. $\frac{\bar{X}-\mu}{s/\sqrt{25}}$) falls between $-t_{24;0.05}$ and $t_{24;0.05}$, he is satisfied with his claim.
- What conclusion should be drawn from a sample that has a mean $\bar{X}$ = 518 hours and a standard deviation s = 40 hours?
- Assume that the distribution of burning times in hours is approximately normal.

Solution to Example 1

- From the table, $t_{24;0.05} = 1.711$. [Excel:"=t.inv(0.95;24)"]
- Therefore, the manufacturer is satisfied with his claim if a sample of 25 bulbs yields a t-value between -1.711 and 1.711.
- If $\mu = 500$, then

  $t = (518 - \mu)/(40/5) = (518 - 500)/8 = 2.25 > 1.711$.
- If $\mu > 500$, then the value of t computed from the sample would be more reasonable.
- Hence the manufacturer is likely to conclude that his bulbs are a better product than he thought.

The *F*-distribution

Definition 5.5

- Let U and V be independent random variables having $\chi^2(n_1)$ and $\chi^2(n_2)$, respectively,
- then the distribution of the random variable,

$$F = \frac{U/n_1}{V/n_2},$$

is called a F distribution with (n_1, n_2) degrees of freedom.

The *F*-distribution (Continued)

Definition 5.5

- The p.d.f. F is given by

$$f_F(x) = \frac{n_1^{n_1/2}\, n_2^{n_2/2}\, \Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \frac{x^{n_1/2-1}}{(n_1 x + n_2)^{(n_1+n_2)/2}},$$

  for $x > 0$ and 0 otherwise.

- It can be shown that $E(X) = n_2/(n_2-2)$, with $n_2 > 2$ and

$$V(X) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}, \text{ with } n_2 > 4$$

Example 1

- Suppose that random samples of sizes n_1 and n_2 are selected from two normal populations with variances σ_1^2 and σ_2^2 respectively.
- From Section 5.7, we know that

$$U = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1)$$

and

$$V = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$$

are independent random variables.

Example 1 (continued)

- Therefore we have

$$F = \frac{U/(n_1 - 1)}{V/(n_2 - 1)} = \frac{\frac{(n_1 - 1)S_1^2/\sigma_1^2}{(n_1 - 1)}}{\frac{(n_2 - 1)S_2^2/\sigma_2^2}{(n_2 - 1)}}$$

$$= \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Theorem 5.7

- If $F \sim F(n, m)$, then $1/F \sim F(m, n)$.
- This theorem follows immediately from the definition of F-distribution.
- Values of the F-distribution can be found in the statistical tables.
- The table gives the values of $F(n_1, n_2; \alpha)$ such that $\Pr\big(F > F(n_1, n_2; \alpha)\big) = \alpha$.

Adobe Acrobat Document

Theorem 5.7 (Continued)

For example

- $F(5,4;0.05) = 6.26$ means $\Pr(F > 6.26) = 0.05$, where $F \sim F(5,4)$.
- $F(4,5;0.025) = 7.39$ means $\Pr(F > 7.39) = 0.025$, where $F \sim F(4,5)$.

<u>Excel commands for F-distribution</u>

"=F.DIST(f; n_1; n_2; true)" gives $\Pr(F \leq f)$

e.g. "=F.DIST(6.26;5;4; true)" gives 0.95

"=F.INV(p; n_1;n_2)" gives c satisfies $\Pr(F \leq c) = p$

e.g. "=F.INV(0.95;4;5)" gives 5.192

Adobe Acrobat Document

Theorem 5.8

$$F(n_1, n_2; 1-\alpha) = 1/\,F(n_2, n_1; \alpha).$$

For example,

- $F(10, 5; 0.95) = 1/F(5, 10; 0.05) = 1/3.33 = 0.30$
  which means $\Pr(F > 0.30) = 0.95$,
  where $F \sim F(10, 5)$.

Example 2

- Let S_1^2 and S_2^2 be the sample variances of independent random samples of sizes $n_1 = 25$ and $n_2 = 31$ taken from normal populations with variances $\sigma_1^2 = 10$ and $\sigma_2^2 = 15$ respectively. Find $\Pr(S_1^2/S_2^2 > 1.26)$.

Solution

$$\Pr\left(\frac{S_1^2}{S_2^2} > 1.26\right) = \Pr\left(\frac{S_1^2/10}{S_2^2/15} > 1.26\left(\frac{15}{10}\right)\right)$$

$$= \Pr(F > 1.89) = 0.05,$$

where $F \sim F(24, 30)$