Basic Statistics

Dr. Sudipta Das

Department of Computer Science,
Ramakrishna Mission Vivekananda Educational & Research Institute

Outline I

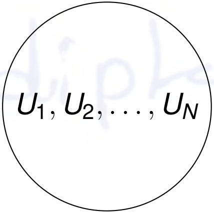
1. Sampling Theory
   - Sampling from Finite Population
   - Sampling from Theoretical Population
     - Sampling from Normal Population
     - Large Sampling

Chapter 7: Sampling Theory

Sampling I

Performing the study of a **population** with the help of a **sample**

- Population: - A collection/aggregate of all the units/individuals possessing a common characteristics.



Population

- Let's consider a population of size N units, say $(U_1, U_2, \ldots, U_N)$.
    - N is called population size
- Let Y be the character under study, having values $Y_1^{(p)}, Y_2^{(p)}, \ldots, Y_N^{(p)}$ for those N units.

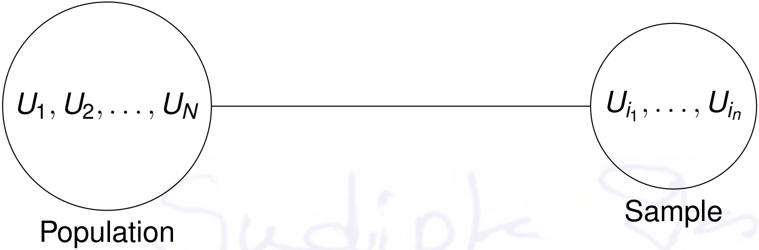
- Three types of population
  1. Real and finite
     - the population of students studying data analytic courses
  2. Infinite or hypothetical
     - the population of all the tosses that can be made with a coin
     - the population of all the stars in our galaxy
  3. Theoretical in nature
     - the binomial population or normal population etc.
- A population can be studied either by complete enumeration (census) or by choosing some units from population (sampling)

Sampling IV

Census or Complete enumeration:

- By census, we mean inspection/enumeration/study of **all the units** belonging to a population.
  - If we undertake a census, we have knowledge of $Y_1^{(p)}, Y_2^{(p)}, \ldots, Y_N^{(p)}$ individually.
- Problem with census
  - Naturally, census possible for a finite population only.
    - For infinite population census is impossible
  - For finite population there may be restriction in time, cost or space.
    - Destructive sampling: to study the length of life (in hours) of the bulbs produced by a company.

Sampling V



Sampling:

- This consists of selection/inspection of **only a few units** of population
- The selected/inspected units as a group is called a sample.

Sampling VI

- Types of sampling
  1. Non-probabilistic (judgmental) sampling
     - Not helpful
  2. Probabilistic sampling
     - It helps in statistical analysis and decision making.
- Based on the study of probability, the present study of *sampling* marks the beginning the learning of statistics beyond the descriptive phase.

Sampling from Finite Population I

- Simple Random Sampling (SRS):
  - Suppose we draw n units from the population of N units, as our sample.
    - n is called sample size
  - The sample will be called simple random sample, if **every possible sample of size n has the same probability of being selected**

Sampling from Finite Population II

- Three types of Simple Random Sampling (SRS)
  1. SRS with replacement (SRSWR)
  2. SRS without replacement (SRSWOR)
  3. SRS without replacement unordered (SRSWOR unordered)

Sampling from Finite Population III

- Simple Random Sampling with replacement (SRSWR)
  - In this case, after the first unit is selected at random from all the N units, it is replaced to the population, and, the second unit is drawn, again at random, from all the N units of the population.
  - This procedure is repeated n times to get an SRS of size n drawn with replacement from the population
- SRSWR for a population of size N,
  - total number of possible sample of size n is

$$N^n$$

  - and selection probability of each sample is $1/N^n$

Sampling from Finite Population IV

- Simple Random Sampling without replacement (SRSWOR)
  - Here after the first unit is selected at random from all the N units, it is not replaced to the population, and, the second unit is drawn, again at random, from the remaining $N-1$ units of the population.
  - This procedure is repeated n times to get an SRS of size n drawn without replacement from the population
- SRSWOR for a population of size N,
  - total number of possible sample of size n is

$$N(N-1)(N-2)\cdots(N-n+1) = {}^{N}P_{n}$$

  - and selection probability of each sample is $\dfrac{1}{{}^{N}P_{n}}$

- Simple Random Sampling without replacement unordered (SRSWOR unordered)
  - If the order of appearance of the units in the sample can be ignored, then we have the unordered SRSWOR procedure
- SRSWOR unordered for a population of size N,
  - total number of possible sample of size n is

$$\frac{N(N-1)(N-2)\cdots(N-n+1)}{1\times 2\times\cdots\times n} = {}^{N}C_{n}$$

  - and selection probability of each sample is $\dfrac{1}{{}^{N}C_{n}}$

Sampling from Finite Population VI

- Example: Population $\{U_1, U_2, U_3, U_4\}$ N=4 with $n = 2$

SRSWR

$$\left.\begin{array}{cccc} U_1,U_1 & U_1,U_2 & U_1,U_3 & U_1,U_4 \\ U_2,U_1 & U_2,U_2 & U_2,U_3 & U_2,U_4 \\ U_3,U_1 & U_3,U_2 & U_3,U_3 & U_3,U_4 \\ U_4,U_1 & U_4,U_2 & U_4,U_3 & U_4,U_4 \end{array}\right\} 16$$

SRSWOR(ordered)

$$\left.\begin{array}{cccc} -- & U_1,U_2 & U_1,U_3 & U_1,U_4 \\ U_2,U_1 & - & U_2,U_3 & U_2,U_4 \\ U_3,U_1 & U_3,U_2 & -- & U_3,U_4 \\ U_4,U_1 & U_4,U_2 & U_4,U_3 & -- \end{array}\right\} 12$$

SRSWOR(unordered)

$$\left.\begin{array}{cccc} -- & U_1,U_2 & U_1,U_3 & U_1,U_4 \\ -- & - & U_2,U_3 & U_2,U_4 \\ -- & -- & -- & U_3,U_4 \\ -- & -- & -- & -- \end{array}\right\} 6$$

- Objective: To know about a parameter of the population from the collected sample

Sampling from Finite Population VII

- Parameter: A parameter is a real-valued measurable function of the population values, say

$$\theta = f(Y_1^{(p)}, Y_2^{(p)}, \ldots, Y_N^{(p)}),$$

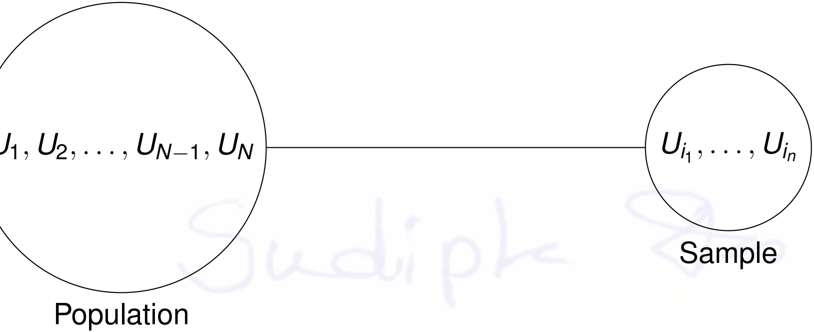
where $(Y_1^{(p)}, Y_2^{(p)}, \ldots, Y_N^{(p)})$ are the population values of a finite population

Sampling from Finite Population VIII

- Population parameters, e.g.,
  1. Population mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i^{(p)}$
  2. Population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i^{(p)} - \mu\right)^2$
  3. Population proportion of some character:

$$p = \frac{X^{(p)}}{N},$$

where $X^{(p)}$ =number of units possessing the character in the population



- Population values $Y(U_1) = Y_1^{(p)}, Y(U_2) = Y_2^{(p)}, \ldots, Y(U_N) = Y_N^{(p)}$
- Sample values $Y(U_{i_1}) = Y_1, Y(U_{i_2}) = Y_2, \ldots, Y(U_{i_n}) = Y_n$, where $\{i_1, \ldots, i_n\}$ are integers from $\{1, 2, \ldots, N\}$
  - Note that the sample values are RANDOM VARIABLE

Sampling from Finite Population X

- Statistic: A statistic is a real valued measurable function of the sample values $(Y_1, Y_2, \ldots, Y_n)$, say

$$T = T(Y_1, Y_2, \ldots, Y_n),$$

For example

  - Sample mean $= \bar{Y} = \dfrac{Y_1 + Y_2 + \ldots + Y_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$
  - Sample variance $= s_Y^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$
  - Sample proportion of some character $= \hat{p} = \dfrac{X}{n}$, where $X =$number of units possessing the character in the sample

- Note:
  - Any statistics, being a function of n random variables $(Y_1, Y_2, \ldots, Y_n)$, is itself a random variable, having a certain probability distribution.

- Sampling Distribution of a statistic:
  - is the probability distribution of the statistic (which is a random variable) for repeated sampling each of size n from the same population.

- Expectation and Standard error of a statistic T with sampling distribution given by a p.m.f $f(t)$
  - Expectation of T
  $$\mu_T = E(T) = \sum_t t \times f(t),$$
  which is the mean of the sampling distribution
  - Standard error of T
  $$\sigma_T = SE(T) = +\sqrt{var(T)},$$
  where $var(T) = E(T - \mu_T)^2 = \sum_t (t - \mu_T)^2 \times f(t)$.

Sampling from Finite Population XIII

Study on Sample Mean

Problem:

- A population has 4 values ($Y_1^{(P)} = 1$, $Y_2^{(P)} = 2$, $Y_3^{(P)} = 4$, $Y_4^{(P)} = 5$) and we draw a sample of size 2 from this population. Derive the sampling distribution of the sample mean. Also find the expectation and the standard error of the sample mean.

Solution steps:

- Population size, $N = 4$ and Sample size, $n = 2$,
- Population mean, $\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i^{(P)} = 3.0$
- Population variance, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i^{(P)}\right)^2 - \mu^2 = 2.50$
- Population standard deviation, $\sigma = \sqrt{2.50}$

Table: Possible sample and the sample mean ($\bar{Y}$) in SRSWR

Sample No	Sample Values	Prob	Sample Mean	Sample No	Sample Values	Prob	Sample Mean
1	(1,1)	1/16	1.0	9	(4,1)	1/16	2.5
2	(1,2)	1/16	1.5	10	(4,2)	1/16	3.0
3	(1,4)	1/16	2.5	11	(4,4)	1/16	4.0
4	(1,5)	1/16	3.0	12	(4,5)	1/16	4.5
5	(2,1)	1/16	1.5	13	(5,1)	1/16	3.0
6	(2,2)	1/16	2.0	14	(5,2)	1/16	3.5
7	(2,4)	1/16	3.0	15	(5,4)	1/16	4.5
8	(2,5)	1/16	3.5	16	(5,5)	1/16	5.0

Table: Sampling distribution of the sample mean ($\bar{Y}$) in SRSWR

$\bar{Y}$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	Total
Prob	1/16	2/16	1/16	2/16	4/16	2/16	1/16	2/16	1/16	1

Sampling from Finite Population XVI

- Mean of the sample mean ($\bar{Y}$), in SRSWR:

$$\text{Expectation of } \bar{Y} = E[\bar{Y}] = 3.0$$

- Variance of the sample mean ($\bar{Y}$), in SRSWR:

$$\text{Variance of } \bar{Y} = E[\bar{Y}^2] - E^2[\bar{Y}] = 1.25$$

- Standard error of the sample mean ($\bar{Y}$), in SRSWR:

$$\text{Square root of Variance of } \bar{Y} = SE[\bar{Y}] = \sqrt{1.25}$$

- Note that
  - $E[\bar{Y}] = \mu = \text{ population mean} = 3.0$
  - $SE[\bar{Y}] = \dfrac{\sigma}{\sqrt{n}} = \dfrac{\text{population s.d}}{\sqrt{\text{sample size}}} = \dfrac{\sqrt{2.5}}{\sqrt{2}}$

Table: Possible sample and the sample mean ($\bar{Y}$) in SRSWOR

Sample No	Sample Values	Prob	Sample Mean
1	(1,2)	1/6	1.5
2	(1,4)	1/6	2.5
3	(1,5)	1/6	3.0
4	(2,4)	1/6	3.0
5	(2,5)	1/6	3.5
6	(4,5)	1/6	4.5

Table: Sampling distribution of the sample mean ($\bar{Y}$) in SRSWOR

$\bar{Y}$	1.5	2.5	3.0	3.5	4.5	Total
Prob	1/6	1/6	2/6	1/6	1/6	1

Sampling from Finite Population XVIII

- Mean of the sample mean ($\bar{Y}$), in SRSWOR:

$$\text{Expectation of } \bar{Y} = E[\bar{Y}] = 3.0$$

- Variance of the sample mean ($\bar{Y}$), in SRSWOR:

$$\text{Variance of } \bar{Y} = E[\bar{Y}^2] - E^2[\bar{Y}] = 5/6$$

- Standard error of the sample mean ($\bar{Y}$), in SRSWR:

$$\text{Square root of Variance of } \bar{Y} = SE[\bar{Y}] = \sqrt{5/6}$$

- Note that
  - $E[\bar{Y}] = \mu = 3.0$
  - $SE[\bar{Y}] = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} = \frac{\sqrt{2.5}}{\sqrt{2}} \times \sqrt{\frac{4-2}{4-1}}$

Theorem

- Consider a finite population of N units having mean μ and variance σ^2. Suppose we draw a simple random sample of size n from the above population. If $\bar{Y}$ denotes the sample mean, then
  - For SRSWR,
    - $E[\bar{Y}] = \mu$
    - $Var(\bar{Y}) = \dfrac{\sigma^2}{n} \Rightarrow SE[\bar{Y}] = \dfrac{\sigma}{\sqrt{n}}$
  - For SRSWOR,
    - $E[\bar{Y}] = \mu$
    - $Var(\bar{Y}) = \dfrac{\sigma^2}{n}\dfrac{N-n}{N-1} \Rightarrow SE[\bar{Y}] = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$

Remarks

- The factor $\dfrac{N-n}{N-1}$ is called the **finite population correction factor.**
- If $n > 1$, then $SE(\bar{Y}|SRSWOR) < SE(\bar{Y}|SRSWR)$
- For a fixed n, as $N \to \infty$, the $SE(\bar{Y}|SRSWOR) \to SE(\bar{Y}|SRSWR)$
  - As sampling from a practically infinite population, it is immaterial whether the units already drawn are returned or not before drawing the next unit.

Sampling from Finite Population XXI

Sketch of proof (SRSWR)

- Support of the i^{th} sample unit (Y_i) is $\{Y_1^{(P)}, Y_2^{(P)}, \ldots, Y_k^{(P)}, \ldots, Y_l^{(P)}, \ldots, Y_N^{(P)}\}$
- $P(Y_i = Y_k^{(P)}) = \frac{1}{N}$
- $P(Y_i = Y_k^{(P)}, Y_j = Y_l^{(P)}) = \frac{1}{N^2}$
- $E(Y_i) = \sum_{k=1}^{N} Y_k^{(P)} P(Y_i = Y_k^{(P)}) = \frac{1}{N} \sum_{k=1}^{N} Y_k^{(P)} = \mu$
- $Var(Y_i) = \sum_{k=1}^{N} (Y_k^{(P)} - \mu)^2 P(Y_i = Y_k^{(P)}) = \frac{1}{N} \sum_{k=1}^{N} (Y_k^{(P)} - \mu)^2 = \sigma^2$
- $Cov(Y_i, Y_j) = 0$
- $E[\bar{Y}] = E\left[\frac{1}{n} \sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\left[\sum_{i=1}^{n} EY_i\right] = \frac{1}{n} \times n\mu = \mu$
- $Var[\bar{Y}] = Var\left[\frac{1}{n} \sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n} Var(Y_i) + 2\sum_{i=1}^{n} \sum_{j=i+1}^{n} Cov(Y_i, Y_j)\right] = \frac{1}{n^2}\left[n\sigma^2\right] = \frac{\sigma^2}{n}$

Sampling from Finite Population XXII

Sketch of proof (SRSWOR)

- Support of the i^{th} sample unit (Y_i) is $\{Y_1^{(P)}, Y_2^{(P)}, \ldots, Y_k^{(P)}, \ldots, Y_l^{(P)}, \ldots, Y_N^{(P)}\}$
- $P(Y_i = Y_k^{(P)}) = \frac{{}^{N-1}P_{n-1}}{{}^{N}P_n} = \frac{1}{N}$
- $P(Y_i = Y_k^{(P)}, Y_j = Y_l^{(P)}) = \frac{{}^{N-2}P_{n-2}}{{}^{N}P_n} = \frac{1}{N(N-1)}$
- $E(Y_i) = \sum_{k=1}^{N} Y_k^{(P)} P(Y_i = Y_k^{(P)}) = \frac{1}{N}\sum_{k=1}^{N} Y_k^{(P)} = \mu$
- $Var(Y_i) = \sum_{k=1}^{N}(Y_k^{(P)} - \mu)^2 P(Y_i = Y_k^{(P)}) = \frac{1}{N}\sum_{k=1}^{N}(Y_k^{(P)} - \mu)^2 = \sigma^2$
- $Cov(Y_i, Y_j) = \sum_{\substack{k=1 \\ k\neq l}}^{N}\sum_{l=1}^{N}(Y_k^{(P)} - \mu)(Y_l^{(P)} - \mu)P(Y_i = Y_k^{(P)}, Y_j = Y_l^{(P)}) = -\frac{\sigma^2}{N-1}$
- $E[\bar{Y}] = E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\left[\sum_{i=1}^{n} EY_i\right] = \frac{1}{n} \times n\mu = \mu$
- $Var[\bar{Y}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n} Var(Y_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} Cov(Y_i, Y_j)\right] = \frac{1}{n^2}\left[n\sigma^2 + 2\,{}^{n}C_2\left(\frac{-\sigma^2}{N-1}\right)\right] = \frac{\sigma^2}{n}\frac{N-n}{N-1}$

Sampling from Finite Population XXIII

Study on Sample Proportion

- Suppose we interested in estimating the population proportion of some character (e.g. proportion of smokers in a city), say p.
- We draw an SRS of size n and let $\hat{p}$ be the corresponding sample proportion
  - Sample Proportion, $\hat{p}$ is a statistic
  - We are interested to know the expectation and standard error of $\hat{p}$

Sampling from Finite Population XXIV

- Define the N population values as

$$Y_i^{(P)} = \begin{cases} 1, & \text{if the } i^{th} \text{unit posses the character} \\ 0, & \text{otherwise} \end{cases}$$

- Thus
  - Population mean $= \mu = \frac{1}{N}\sum_{i=1}^{N} Y_i^{(P)} = \frac{X^{(P)}}{N} = p =$ Population proportion
  - Population variance $= \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}\left(Y_i^{(P)}\right)^2 - p^2 = p(1-p)$
- Therefore, the n sample values are

$$Y_i = \begin{cases} 1, & \text{if the } i^{th} \text{unit posses the character} \\ 0, & \text{otherwise} \end{cases}$$

  - Sample mean $= \bar{y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{X}{n} = \hat{p} =$ Sample proportion

Theorem

- If $\hat{p}$ be the sample proportion for a simple random sample of size n drawn from a population of size N having population proportion p, then
  - For SRSWR,
    - $E[\hat{p}] = p$
    - $Var(\hat{p}) = \frac{p(1-p)}{n} \Rightarrow SE[\bar{Y}] = \sqrt{\frac{p(1-p)}{n}}$
  - For SRSWOR,
    - $E[\hat{p}] = p$
    - $Var(\hat{p}) = \frac{p(1-p)}{n} \times \frac{N-n}{N-1} \Rightarrow SE[\bar{Y}] = \sqrt{\frac{p(1-p)}{n} \times \frac{N-n}{N-1}}$

Sampling from Theoretical Population I

- (I.I.D.) Random Sampling: A number of random variables is selected from a population of **identical** random variables and the random variables are selected **independently** one from another
    - n, (sample size) is the number of selected random variables
- Note that
    - Any function T of observable random variables $X_1, \ldots, X_n$ that does not depend on any unknown parameters is called a statistic.
    - The probability distribution of the sample statistic T is called the sampling distribution of T

Sampling from Theoretical Population II

Theorem:

- Let $X_1, \ldots, X_n$ be a random sample of size n from a population with mean μ and variance σ^2. Then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean having

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

and

$$Var(\bar{X}) = \frac{\sigma^2}{n} \Rightarrow SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

  - Note: The sample means become more and more reliable as an estimate of μ as the sample size is increased,

- **Normal Population:** The random variables (units) in the population are normally distributed.

Normal Random Variable

- A random variable X is said to be normal if its probability density function is given by

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

where $-\infty < \mu < \infty$ and $0 < \sigma < \infty$.

  - Notation: $X \sim N(\mu, \sigma)$
  - μ is called the expectation/mean of X, i.e., $E[X] = \mu$
  - σ^2 is called the variance of X, i.e., $E\left[X - \mu\right]^2 = \sigma^2$
  - Moment Generating Function (mgf): $M_X(t) = E\left(e^{tX}\right) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

Standardized Normal Random Variable

- If X is normal random variable with mean μ and variance σ^2, then the random variable
$$Z = \frac{X - \mu}{\sigma}$$
is said to be standardized normal random variable.
  - We denote this by $Z \sim N(0, 1)$

Sampling from Normal Population IV

- Probability density function of a standardized normal random variable Z is given by

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

  - Mean: $\mu_Z = 0$
  - Variance: $\sigma_Z^2 = 1$
  - MGF: $M_Z(t) = e^{\frac{1}{2}t^2}$

Distribution of Sample Mean ($\bar{X}$)

- *Theorem 1*: Let $\{X_1, \ldots, X_n\}$ be a random sample of size n, drawn from a normal population with mean μ and variance σ^2. Then

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

is Normally distributed with mean μ and variance $\dfrac{\sigma^2}{n}$.

- Thus,

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

Sampling from Normal Population VI

χ^2- distribution

- If Z_is are n independent standardized normal variables, then the random variable
$$K = \sum_{i=1}^{n} Z_i^2$$
is said to have a Chi-square distribution with n degrees of freedom.
  - We denote this by $K \sim \chi_n^2$

Sampling from Normal Population VII

- Probability density function of a (centralized) χ^2 random variable K with degree of freedoms n, is given by

$$f_K(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}.$$

  - Mean: $\mu_K = n$
  - Variance: $\sigma_K^2 = 2n$
  - MGF: $M_K(t) = (1-2t)^{-n/2}$, for $t < \frac{1}{2}$

Sampling from Normal Population VIII

Some observations

- Suppose the random sample $\{X_1, \ldots, X_n\}$ is drawn from a normal population with mean μ and variance σ^2. Equivalently, $X_i \sim N(\mu, \sigma^2)$. Then

$$Z_i = (X_i - \mu)/\sigma, \text{ for } i = 1, \ldots, n$$

  are independent standard normal random variables.
- The square of standard normal random variables

$$Z_i^2 = \left(\frac{X_i - \mu}{\sigma}\right)^2 \text{ for } i = 1, \ldots, n$$

  has a χ^2-distribution with 1 degrees of freedom.
  - MGF: $M_{Z_i^2}(t) = (1 - 2t)^{-1/2}$, for $t < \frac{1}{2}$

Sampling from Normal Population IX

- *Theorem 2*: Suppose the random sample $\{X_1, \ldots, X_n\}$ is drawn from a $N(\mu, \sigma^2)$ distributed population. Then $Z_i = (X_i - \mu)/\sigma, i = 1, \ldots, n$ are independent standard normal random variables. Thus the random variable

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

has a χ^2-distribution with n degrees of freedom.

  - MGF: $M(t) = (1 - 2t)^{-n/2}$, for $t < \frac{1}{2}$

Distribution of Sample Variance (S^2)

- *Theorem 3*: If $\{X_1, \ldots, X_n\}$ is a random sample from a normal population with the mean μ and variance σ^2, then the random variable
$$\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2}$$
has a χ^2-distribution with $n-1$ degrees of freedom, where
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$
The sample mean $\bar{X}$ and sample variance S^2 are independent, also.
- Sketch of proof:

$$\underbrace{\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2}_{n} = \sum_{i=1}^{n}\left[\frac{(X_i-\bar{X})+(\bar{X}-\mu)}{\sigma}\right]^2 = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 + n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 = \underbrace{\frac{(n-1)S^2}{\sigma^2}}_{n-1} + \underbrace{\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2}_{1}$$

t- distribution

- If Y and Z are two independent random variables, such that $Y \sim \chi^2_n$ and $Z \sim N(0, 1)$, then the random variable

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

  is said to have a (Student) t-distribution with n degrees of freedom.
    - We denote this by $T \sim t_n$

Sampling from Normal Population XII

- Probability density function of a T random variable with degree of freedom n, is given by

$$f_T(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

  - Mean: $\mu_T = 0$
  - Variance: $\sigma_T^2 = \begin{cases} \frac{n}{n-2}, & \text{for } n > 2 \\ 1, & \text{for } 1 < n \leq 2 \end{cases}$
  - MGF: $M_T(t)$ is undefined

Distribution of Sample Mean standardized by Sample Variance

- *Theorem 4*: If $\bar{X}$ and S^2 are the mean and the variance of a random sample of size n, drawn from a normal population with the mean μ and variance σ^2, then the statistic (random variable)

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

  has a $t-$distribution with $n - 1$ degrees of freedom

F distribution

- If U and V are two independent chi-square random variables with n_1 and n_2 degrees of freedom, respectively. Then the random variable

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

is said to have an F-distribution with (n_1, n_2) degrees of freedom.

  - We denote this by $F \sim F_{n_1,n_2}$

- Probability density function of a (centralized) F random variable with degrees of freedom n_1 and n_2, is given by

$$f_F(x) = \frac{1}{x\,\mathrm{B}\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \sqrt{\frac{(n_1 x)^{n_1} n_2^{n_2}}{(n_1 x + n_2)^{n_1+n_2}}}.$$

  - Mean: $\dfrac{n_2}{n_2 - 2}$ for $n_2 > 2$
  - Variance: $\dfrac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}$ for $n_2 > 4$
  - MGF: $M_F(t)$ does not exist

Distribution of Ratio of Sample Variances

- *Theorem 5*: Let two independent random samples of size n_1 and n_2 be drawn from two normal populations with variances σ_1^2, and σ_2^2, respectively. If the variances of the random samples are given by S_1^2 and S_2^2, respectively, then the statistic (random variable)

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

has a $F-$distribution with $(n_1 - 1), (n_2 - 1)$ degrees of freedom.

- *Corollary*: Under the equality assumption of two population variances (i.e., $\sigma_1^2 = \sigma_2^2$) the statistic (random variable)

$$F = \frac{S_1^2}{S_2^2}$$

has a $F-$distribution with $(n_1 - 1), (n_2 - 1)$ degrees of freedom

Large Sampling from Any Population I

Distribution of Large Sample Mean $(\bar{X})$

- *Central Limit Theorem (CLT)*: Suppose $\{X_1, \ldots, X_n\}$, a random sample of size n, is drawn from a population (*not necessarily normal*) with mean μ and finite variance σ^2.
  Then the standardized sample mean

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