Five

Sampling and Sampling Distributions

1 Population and Sample

The aim of *Statistical Inference* is to say something about the population based on a sample.

Definition 1 (Population & 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 categorical value.

So each member of a population can be regarded as a value of a random variable.

Note that a population can be finite or infinite.

FINITE POPULATION

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

For example, it can be

- *the monthly income of Singaporeans;*
- *all the books in the Central Library; or*
- *the CAP scores of students in NUS.*

Infinite Population

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

For example, it can be

- *the results of all possible rolls of a pair of dice;*
- *the depths at all conceivable positions of a lake; or*
- *the PSI level in the air at various parts of Singapore.*

REMARK

Some finite populations are so large that in theory we assume them to be infinite, since it may be impractical/uneconomical to observe all its values.

2 Random Sampling

We often know that the population belongs to (or can be modeled using) a known (family of) distribution(s).

However, the values of parameters (for example, p, μ or σ) that specify the distribution(s) are unknown.

For example:

- A pollster is sure that the responses to his “agree/disagree” question will follow a binomial distribution, but p, the proportion of those who “agree” in the population, is unknown.
- An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean μ and the standard deviation σ of the yields are unknown.

Thus we rely on a sample to learn about these parameters and study the properties of the population.

- The sample should be representative of the population. We have different types of sampling schemes attempting to do that. For the probability methods, it is possible to fully describe the quantitative properties of the sample.

- We will focus on the simple random sample. It is often known simply as a random sample.

Definition 2 (Simple Random Sample)

A set of n members taken from a given population is called a ***sample*** *of size n.*

A ***simple random sample (SRS)*** *of n members is a sample that is chosen such that every subset of n observations of the population has the same probability of being selected.*

Remark

With simple random sampling, everyone has the same chance of inclusion in the sample, so it is fair.

It tends to yield a sample that resembles the population. This reduces the chance that the sample is seriously biased in some way, leading to inaccurate inferences about the population.

EXAMPLE 5.1 (DRUG EXPERIMENT)

Suppose that a researcher in a medical center plans to compare two drugs for some adverse condition. She has four patients with this condition, and she wants to randomly select two to use each drug. Denote the four patients by P_1, P_2, P_3, and P_4.

In selecting $n = 2$ subjects to use the first drug, the six possible samples are

$$(P_1, P_2), (P_1, P_3), (P_1, P_4), (P_2, P_3), (P_2, P_4), (P_3, P_4).$$

REMARK

More generally, let N denote the population size. The population has $\binom{N}{n}$ possible samples of size n.

For large values of N and n, one can use software easily to select the sample from a list of the population members using a random number generator.

Sampling from an Infinite Population

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.

Definition 3 (Simple Random Sample: Infinite Population)
Let X be a random variable with certain probability distribution $f_X(x)$.

Let $X_1, X_2, \ldots, X_n$ be n independent random variables each having the same distribution as X. Then $(X_1, X_2, \ldots, X_n)$ is called a ***random sample of size n*** *from a population with distribution $f_X(x)$.*

The joint probability function of $(X_1, X_2, \ldots X_n)$ is given by

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

where $f_X(x)$ is the probability function of the population.

3 Sampling Distribution of Sample Mean

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.

Definition 4 (Statistic)
*Suppose a random sample of n observations $(X_1, \ldots, X_n)$ has been taken. A function of $(X_1, \ldots, X_n)$ is called a **statistic**.*

Example 5.2 (Sample Mean)

The **sample mean**, defined as

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

is a statistic.

If the values in a random sample are observed and they are $(x_1, \ldots, x_n)$, then the realization of the statistic $\overline{X}$ is given by

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

Example 5.3 (Sample Variance)

The **sample variance**, defined as

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

is a statistic.

Similarly, if the values in a random sample are observed and they are $(x_1, \ldots, x_n)$, then the realization of the statistic S^2 is given by

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

STATISTICS ARE RANDOM VARIABLES

- *Note that X_1 is a random variable and so are $X_2, \ldots, X_n$.*

- *Thus $\overline{X}$ and S^2 are random variables as well.*

- *As many random samples are possible from the same population, we expect the statistic to vary somewhat from sample to sample.*

- *Hence a statistic is a random variable. It is meaningful to consider the probability distribution of a statistic.*

Definition 5 (Sampling Distribution)
*The probability distribution of a statistic is called a **sampling distribution**.*

Two Results

We next present two key results about the sampling distribution of the sample mean.

- Theorem 6 provides formulas for the center and the spread of the sampling distribution.
- Theorem 9 describes the shape of the sampling distribution, showing that it is often approximately normal.

Theorem 6 (Mean and Variance of $\overline{X}$)
For random samples of size n taken from an infinite population with mean μ_X and variance σ_X^2, the sampling distribution of the sample mean $\overline{X}$ has mean μ_X and variance $\dfrac{\sigma_X^2}{n}$. That is,

$$\mu_{\overline{X}} = E\left(\overline{X}\right) = \mu_X \quad \textit{and} \quad \sigma_{\overline{X}}^2 = \operatorname{var}\left(\overline{X}\right) = \frac{\sigma_X^2}{n}.$$

VALIDITY OF $\overline{X}$ AS AN ESTIMATOR FOR μ_X

- *The expectation of* $\overline{X}$ *is equal to the population mean* μ_X.
- *In "the long run",* $\overline{X}$ *does not introduce any systematic bias as an estimator of* μ_X. *So* $\overline{X}$ *can serve as a valid estimator of* μ_X.
- *For an infinite population, when* n *gets larger and larger,* σ_X^2/n, *the variance of* $\overline{X}$, *becomes smaller and smaller, that is, the accuracy of* $\overline{X}$ *as an estimator of* μ_X *keeps improving.*

Definition 7 (Standard Error)

The spread of a sampling distribution is described by its standard deviation, which is called the ***standard error****.*

The standard deviation of the sampling distribution of $\overline{X}$ *is called the standard error of* $\overline{X}$*. We denote it by* $\sigma_{\overline{X}}$.

Remark

The standard error of $\overline{X}$ describes how much $\bar{x}$ tends to vary from sample to sample of size n.

The symbol $\sigma_{\overline{X}}$ (instead of σ) and the terminology standard error (instead of standard deviation) distinguishes this measure from the standard deviation σ of the population.

Because σ_X^2/n decreases as n increases, $\overline{X}$ tends to be closer to μ_X as n increases. The result that $\overline{X}$ converges to μ_X as n grows indefinitely is called the **Law of Large Numbers**.

> **Theorem 8 (Law of Large Numbers (LLN))**
> *If $X_1, \ldots, X_n$ are independent random variables with the same mean μ and variance σ^2, then for any $\varepsilon \in \mathbb{R}$,*
>
> $$P(|\overline{X} - \mu| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

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 $\overline{X}$ is close to μ_X, as n gets larger.

4 Central Limit Theorem

The result that the sampling distribution of $\overline{X}$ is approximately normal is called the **Central Limit Theorem**.

Theorem 9 (Central Limit Theorem (CLT))

If $\overline{X}$ is the mean of a random sample of size n taken from a population having mean μ and finite variance σ^2, then, as $n \to \infty$,

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

or equivalently

$$\overline{X} \to N\left(\mu, \frac{\sigma^2}{n}\right).$$

WHAT IS THE BIG DEAL?

The Central Limit Theorem states that, under rather general conditions, for large n, sums and means of random samples drawn from a population follows the normal distribution closely.

Note that if the random sample comes from a normal population, $\overline{X}$ is normally distributed regardless of the value of n.

Rule of Thumb

The Central Limit Theorem says that, if you take the mean of a large number of independent samples, then the distribution of that mean will be approximately normal.

- *If the population you are sampling from is symmetric with no outliers, a good approximation to normality appears after as few as 15-20 samples.*
- *If the population is moderately skewed, such as exponential or χ^2, then it can take between 30-50 samples before getting a good approximation.*
- *Data with extreme skewness, such as some financial data where most entries are 0, a few are small, and even fewer are extremely large, may not be appropriate for the Central Limit Theorem even with 1000 samples.*

Example 5.4 (Bowling League)

In a bowling league season, bowlers bowl 50 games and the average score is ranked at the end of the season. Historically, John averages 175 a game with a standard deviation of 30. What is the probability that John will average more than 180 this season?

Solution:
We do not know the distribution of X, but we know that $\mu = 175$, $\sigma = 30$ and $n = 50$. Let $\overline{X}$ be the sample mean.

By CLT, we can approximate $\overline{X}$ by $N(\mu, \sigma^2/n)$. The question asks for the probability

$$\begin{aligned} P(\overline{X} > 180) &= P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > \frac{180 - \mu}{\sigma/\sqrt{n}}\right) \\ &\approx P(Z > 1.18) = 0.119. \end{aligned}$$

5 Other Sampling Distributions

We next describe the χ^2, t, and F distributions, which are examples of distributions that are derived from random samples from a normal distribution.

The emphasis is on understanding the relationships between the random variables and how they can be used to describe distributions related to the sample statistics $\overline{X}$ and S^2.

Your goal should be to get comfortable with the idea that sample statistics have known distributions.

Definition 10 (The χ^2 Distribution)

*Let Z be a standard normal random variable. A random variable with the same distribution as Z^2 is called a **χ^2 random variable with one degree of freedom**.*

*Let $Z_1, \ldots, Z_n$ be n independent and identically distributed standard normal random variables. A random variable with the same distribution as $Z_1^2 + \cdots + Z_n^2$ is called a **χ^2 random variable with n degrees of freedom**.*

REMARK

We denote a χ^2 random variable with n degrees of freedom as $\chi^2(n)$.

Properties of χ^2 Distributions

1. *If* $Y \sim \chi^2(n)$, *then* $E(Y) = n$ *and* $\operatorname{var}(Y) = 2n$.

2. *For large* n, $\chi^2(n)$ *is approximately* $N(n, 2n)$.

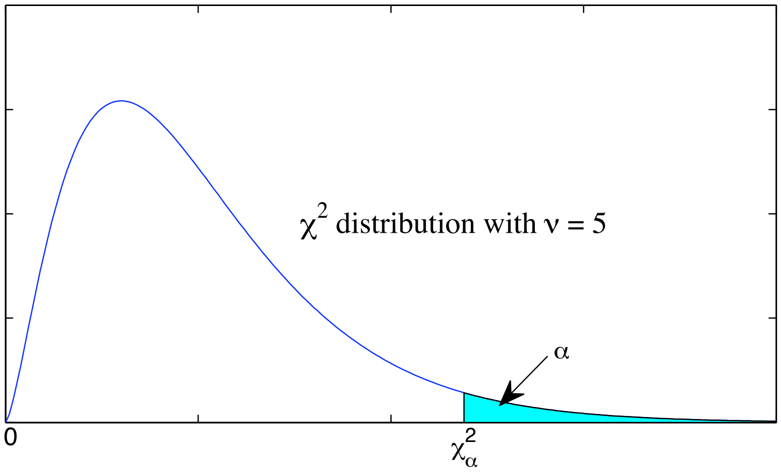
3. *If* Y_1 *and* Y_2 *are independent* χ^2 *random variables with* m *and* n *degrees of freedom respectively, then* $Y_1 + Y_2$ *is a* χ^2 *random variable with* $m+n$ *degrees of freedom.*

4. *The* χ^2 *distribution is a family of curves, each determined by the degrees of freedom* n. *All the density functions have a long right tail.*

Definition 11

Define $\chi^2(n;\alpha)$ *such that for* $Y \sim \chi^2(n)$,

$$P(Y > \chi^2(n;\alpha)) = \alpha.$$



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

Recall that for $X_1, \ldots, X_n$ independent and identically distributed with $E(X) = \mu$ and $\mathrm{var}(X) = \sigma^2$, the sample variance is defined as

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

Though it can be shown that $E(S^2) = \sigma^2$, the sampling distribution of the random variable S^2 has little practical application in statistics.

We shall instead 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 12
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

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

has a χ^2 distribution with $n-1$ degrees of freedom.

Definition 13 (The t-Distribution)
Suppose $Z \sim N(0,1)$ and $U \sim \chi^2(n)$. If Z and U are independent, then

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

*follows the **t-distribution with n degrees of freedom**.*

Properties of the t-Distribution

- *The t-distribution with n degrees of freedom, also called the Student's t-distribution, is denoted by $t(n)$.*

- *The t-distribution approaches $N(0,1)$ as the parameter $n \to \infty$. When $n \geq 30$, we can replace it by $N(0,1)$.*

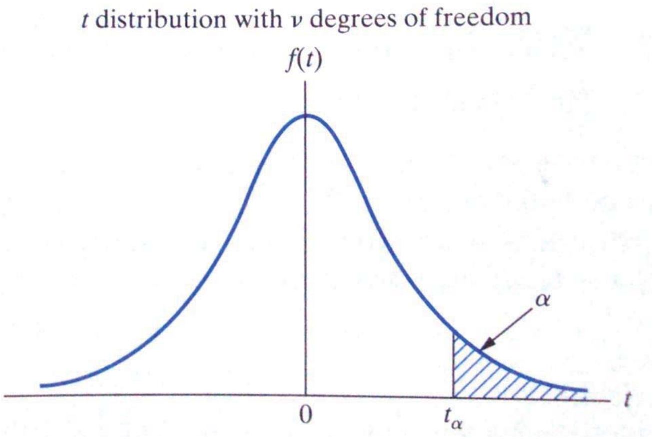
- *If $T \sim t(n)$, then $E(T) = 0$ and $\operatorname{var}(T) = n/(n-2)$ for $n > 2$.*

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

Definition 14

Define $t_{n;\alpha}$ such that for $T \sim t(n)$,

$$P(T > t_{n;\alpha}) = \alpha.$$



The Importance of the t-Distribution

The t-distribution will play an important role in the later chapters, where it appears as the result of random sampling.

The following theorem establishes the connection between a random sample $X_1, \ldots, X_n$ and the t-distribution.

Theorem 15

If $X_1, \ldots, X_n$ are independent and identically distributed normal random variables with mean μ and variance σ^2, then

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

follows a t-distribution with $n-1$ degrees of freedom.

Example 5.5 (Midterm Score)

The lecturer of a class announced that the mean score of the midterm is 16 out of 30. A student doubts it, so he randomly chose 5 classmates and asked them for their scores: 20, 19, 24, 22, 25.

Should the student believe that the mean score is 16? Assume the scores are approximately normally distributed.

Solution:
The student has $n = 5$ sampled data

$$x_1 = 20, x_2 = 19, x_3 = 24, x_4 = 22, x_5 = 25.$$

If $\mu = 16$,

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - 16}{S/\sqrt{5}}$$

should follow a t-distribution with $5 - 1 = 4$ degrees of freedom.

With the observed data $\bar{x} = 22$ and $s = 2.55$ so

$$t = \frac{22 - 16}{2.55/\sqrt{5}} = 5.26.$$

Using software, $P(t(4) > 5.26) = 0.003$. This says that there is only a 0.003 chance that T is 5.26 (or larger), provided the lecturer is telling the truth that $\mu = 16$.

So should the student believe him based on his findings?

Drink Beer and do Statistics!

The t-distributions were discovered by William S. Gosset in 1908. Gosset was a statistician employed by the Guinness brewing company which had stipulated that he not publish under his own name. He therefore wrote under the pen name "Student".

For a biography of Gosset, browse to
`http://www-history.mcs.st-andrews.ac.uk/Biographies/Gosset.html`

Definition 16 (The F-Distribution)

Suppose $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$ are independent. Then the distribution of the random variable

$$F = \frac{U/m}{V/n}$$

is called a ***F-distribution with (m,n) degrees of freedom***.

PROPERTIES OF THE F-DISTRIBUTION

- *The F-distribution with (m,n) degrees of freedom is denoted by $F(m,n)$.*

- *If $X \sim F(m,n)$, then*

$$E(X) = \frac{n}{n-2}, \quad \textit{for } n > 2$$

and

$$\text{var}(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \quad \textit{for } n > 4.$$

- *If* $F \sim F(n,m)$*, then* $1/F \sim F(m.n)$*. This follows immediately from the definition of the* F*-distribution.*

- *Values of the* F*-distribution can be found in the statistical tables or software. The values of interests are* $F(m,n;\alpha)$ *such that*

$$P(F > F(n,m;\alpha)) = \alpha,$$

  where $F \sim F(m,n)$*.*

- *It can be shown that*

$$F(m,n;1-\alpha) = 1/F(n,m;\alpha).$$

EXAMPLE 5.6

For example,

$$F(4,5;0.05) = 5.19$$

means that $P(F > 5.19) = 0.05$, where $F \sim F(4,5)$.