

# 0522 Slides

Cesai Li

May 2025

Key questions to ask:

- 1 If we know a data set is (approximately) a normal distribution, how can we extract information we need from the normality? (Be careful with **Population** vs. **Sample** vs. **Model** )
- 2 Given a data set, how can we tell if it can be approximated by a normal distribution?

**Example.** The GRE is a test required for admission to many US graduate schools. Suppose students' scores on some GRE test can be approximated by a normal distribution with mean 150 and standard deviation 10. What proportion of the students scored between 155 and 160? **Answer: Approximately 14.98%**

# Assessing normality

Given a variable  $x$ , how do we know if  $x$  is approximately a normal variable? (I.e. if the probability density function  $f(x)$  can be approximated by a normal distribution.)

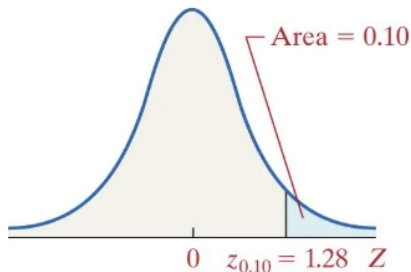
- 1 Method 1. Plot  $f(x)$  and compare its graph to that of a normal distribution.
- 2 Method 2. Compare  $\bar{x} \pm s$ ,  $\bar{x} \pm 2s$ ,  $\bar{x} \pm 3s$  to 68%, 95%, 99.7%, resp.
- 3 Methods using technology or more advanced math tools (not covered in this course).

**In some specific situation, we do have conventional standards to tell if an approximation is good or bad.**

Be careful: In the textbook the term 'z-score' has two different meanings. Here we do not call  $z_\alpha$  a 'z-score'.

### Definition of $z_\alpha$

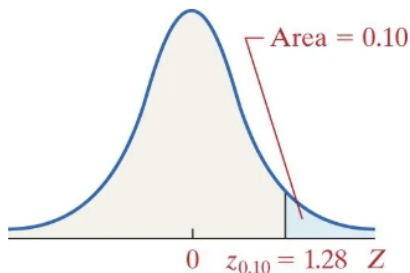
Given a number  $0 \leq \alpha \leq 1$ ,  $z_\alpha$  is defined by  $P(Z \geq z_\alpha) = \alpha$ .



# Example

## Definition of $z_\alpha$

Given a number  $0 \leq \alpha \leq 1$ ,  $z_\alpha$  is defined by  $P(z \geq z_\alpha) = \alpha$ .



Use the Standard Normal Distribution Table to find out  $Z_{0.3}$ .

# Property of $z_\alpha$

Look at the diagram, the symmetry implies the following result.

## Lemma

For any  $0 \leq \alpha \leq 1$  we have  $-z_\alpha = z_{1-\alpha}$ .

**Random sample** is a sample chosen from a population at random.  
**Simple random sample** is a random sample such that each individual in the population has an equal chance of being chosen.

In this course, the two terms are used interchangeably.

# Sampling distribution

**Important concepts:** Population Parameter vs. Sample Statistic.

	Population Parameter	Sample Statistic
Mean	$\mu$	$\bar{X}$
Median	$\eta$	$M$
Variance	$\sigma^2$	$s^2$
Standard Deviation	$\sigma$	$s$

**Definition.** (Sampling distribution of a Sample Statistic.)

Fix a population and a sample size  $n$ . A sampling distribution of a sample statistic is the probability distribution for values of this statistics computed from any sample of size  $n$



# Sampling distribution of the sample mean

Given a population of size  $N$ . What do we know about the sample mean  $\bar{x}$  of a random sample of size  $n$ ?

- 1 What are the possible values of  $\bar{x}$ ?
- 2 What if we choose multiple samples of size  $n$  and compare their  $\bar{x}$ ?

**Example.** Consider a population  $\{1, 2, 0\}$  and a fixed sample size 2. Then the possible samples are  $\{1, 2\}$ ,  $\{2, 0\}$ , and  $\{1, 0\}$ . They have sample mean 1.5, 1, 0.5, resp.

Fix a population and a sample size  $n$ . The sample mean  $\bar{x}$  of a random sample of size  $n$  can be viewed as a **random variable**. Then we may consider its probability distribution, mean  $\mu_{\bar{x}}$ , and standard deviation  $\sigma_{\bar{x}}$ .

Fix a population of size  $N$  with population mean  $\mu$  and population standard deviation  $\sigma$ . Fix a sample size  $n$ . Consider the random variable  $\bar{x}$  of sample mean of random samples of size  $n$ .

## Theorem

the random variable  $\bar{x}$  has mean

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

regardless of  $n$  or  $N$ . If  $n < 0.05N$ , the variable  $\bar{x}$  has a standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

$\sigma_{\bar{x}}$  is also called the **standard error of the mean**. Note that  $\bar{x}$  as a random variable depends on  $n$ , while its mean  $\mu_{\bar{x}}$  is independent of  $n$ , its standard deviation  $\sigma_{\bar{x}}$  does depend on  $n$ .

# Describe the probability distribution of $\bar{x}$ as a variable

## Theorem.

If a random variable  $x$  is (approximately) normally distributed, then the probability distribution of  $\bar{x}$  would also be approximately normal. This result is independent of  $n$ . **This result does not require that  $n < 0.05N$ .**

In the case that the distribution of  $x$  is not normal, we have the following theorem.

## Central Limit Theorem

**If  $n < 0.05N$ ,** as  $n$  increases, the distribution of  $\bar{x}$  becomes more and more normal. When  $n \geq 30$ , we claim that the bell curve is a good approximation of the distribution of  $\bar{x}$ , i.e. the distribution of  $\bar{x}$  is approximately normal.

**RMK. Theorem 1** holds regardless of the distribution of  $x$  (normal or not).

Note: In this specific situation, we have a yes-or-no answer for whether the bell curve is a good approximation, i.e. we have a clear standard to evaluate whether an approximation is good or bad.

### Example.

Suppose a population of size 100 can be approximated by the standard normal distribution. If we randomly pick a sample of size 4 from this population, what is the possibility that the mean of this sample is more than  $\frac{1}{2}$ ?

# Example

Suppose a population of size 100 is a numerical data set with population mean  $\mu = 80$  and population standard deviation  $\sigma = 7$ . Fix a sample size  $n = 49$ , what do we know about the distribution of  $\bar{x}$ ?

- 1 **We can not say the distribution of  $\bar{x}$  is approximately normal (in other words, normality is not guaranteed) because to apply the Central Limit Theorem we will (by the conventional standard we use in this course) need  $n < 0.05N$ , which is not satisfied in this case.**
- 2 We can say  $\mu_{\bar{x}} = 80$ .
- 3 We can not say  $\sigma_{\bar{x}} = 7/\sqrt{49} = 1$  because to apply the standard deviation formula we need  $n < 0.05N$ , which is not satisfied in this case.

In this course, we assume  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$  and CLT normality only hold if  $n < 0.05N$ . Since  $n = 49$  is 49% of  $N = 100$ , these simplifications do not apply.

# Example

Suppose a population of size 100 is a numerical data set with population mean  $\mu = 80$  and population standard deviation  $\sigma = 7$ . Fix a sample size  $n = 4$ , what do we know about the distribution of  $\bar{x}$ ?

- 1 Is the bell curve a good approximation?
- 2 Do we know  $\mu_{\bar{x}}$ ?
- 3 Do we know  $\sigma_{\bar{x}}$ ?

# Example

Suppose a population of size 1000 is a numerical data set with population mean  $\mu = 80$  and population standard deviation  $\sigma = 7$ . If we randomly pick a sample of size 49 from this population, what is the possibility that the mean of this sample is between 79 and 81?



# Sampling distribution

**Important concepts:** Population Parameter vs. Sample Statistic.

	Population Parameter	Sample Statistic
Mean	$\mu$	$\bar{x}$
Median	$\eta$	$M$
Variance	$\sigma^2$	$s^2$
Standard Deviation	$\sigma$	$s$
Proportion	Population Proportion $p$	Sample Proportion $\hat{p}$

**RMK.** Unlike other parameter&statistics pairs, for  $p$  and  $\hat{p}$  to make sense, we do not require the variable  $x$  to be a random variable.

# Example of proportion

## Definition. (Population Proportion.)

Fix a population of size  $N$ . Let  $C$  be an abstract characteristic that an individual in this population can have. Then the population proportion of  $C$  in this population is  $p = \frac{a}{N}$  where  $a$  is the number of individuals in this population that have this characteristic  $C$ .

**Example.** Let my population be all residents in Boston. Let my characteristic be **Age**  $\geq$  **60**. Then  $p = \frac{123694}{652442} = 0.19$ .

# Example of proportion

## Definition. (Sample Proportion.)

Fix a population. Let  $C$  be an abstract characteristic that an individual in this population can have. Let's obtain a sample of size  $n$  from this population. Then the sample proportion of  $C$  in this population is  $\hat{p} = \frac{b}{n}$  where  $b$  is the number of individuals in this sample that have this characteristic  $C$ .

**Example.** Let's obtain a sample of size 20 from the population of all residents in Boston. Let my characteristic be **Age  $\geq 60$** . Suppose in my sample there are 5 people who are over 60 years old. Then

$$\hat{p} = \frac{5}{20} = 0.25.$$

# Idea of sample proportion distribution

We may obtain 1000 simple random samples of size 20 and analysis this data set to estimate the distribution of  $\hat{p}$ . More generally, suppose we have a large number ( $m$ ) of simple random samples of size  $n$  and we want to use this data set to estimate the distribution of  $\hat{p}$ . **Note that as long as  $m$  is big enough,  $m$  does not have a big effect on the (shape of) distribution.**

[diagram]

Terminology: distribution  $\leftrightarrow$  probability distribution function/probability density function. Fix some  $n$ ,  $\hat{p}$  can be viewed as a variable. Be careful that  $\hat{p}$  depends on  $n$ !

Fix a population of size  $N$  and a characteristic. Let  $p$  be the population portion of this characteristic in this population.

Let  $\mu_{\hat{p}}$  denote the mean of the sample proportion, and let  $\sigma_{\hat{p}}$  denote the standard deviation of the sample proportion. Note that  $\sigma_{\hat{p}}$  depends on our choice of  $n$ , but this dependency is not reflected in its notation.

**Theorem.** Assume that any sampling is random.

$\mu_{\hat{p}} = p$ . This holds regardless of  $n$  or  $N$ .

**Theorem.** Assume  $n \leq 0.05N$  and that any sampling is random

- 1 As  $n$  increases, the shape of the distribution of the sample proportion becomes approximately normal. When  $np(1 - p) \geq 10$ , we say the bell curve is a good approximation of the distribution of  $\hat{p}$ .

- 2 
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}}.$$

## Example 2.1

The assumption  $n \leq 0.05N$  is important. (See textbook p.439.)

**Example.** Suppose a population is of size  $N = 10^{12}$ . Suppose a characteristic of this population has  $p = 76\%$ . Describe the distribution of  $\hat{p}$  when  $n = 60$ .

**Answer.** The population size  $N$  is very large, so the condition  $n = 60 \leq N$  is satisfied. Moreover,

$$np(1 - p) = 60 * 0.76 * (1 - 0.76) = 10.944 \geq 10.$$

Since these two conditions are both satisfied, we may claim that the distribution of  $\hat{p}$  is approximately normal, with

$$\mu_{\hat{p}} = p = 0.76$$

and

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}} = 0.055.$$

# Why do we want to describe the distribution of $\hat{p}$

Suppose we obtain 1000 simple random samples of size 20 and analysis this data set to estimate the distribution of  $\hat{p}$ .

If our only purpose is to estimate  $p$ , then instead of obtaining **1000 simple random samples of size 20**, we may simply obtain a big simple random sample of a large enough size  $n$ . Then calculate  $\hat{p} = \frac{b}{n}$  for this big sample, where  $b$  is the number of individuals in this sample with that characteristic. We may then claim that  $p \cong \hat{p}$ .

Being able to describe the distribution of  $\hat{p}$  is more powerful than this. A description of the distribution of  $\hat{p}$  contains more information than its mean  $\mu_{\hat{p}} = p$ , so it does more than estimating the population proportion  $p$ .

## Example 2.2

### Question

Suppose a population is of size  $3.1 \times 10^8$ . Suppose a characteristic of this population has  $p = 15\%$ . In a simple random sample of size 120, what is the probability that less than 12% of this sample has this characteristic?

### In other words (Ex4 p.439)

According to NHS, 15% of all Americans have hearing trouble. In a random sample of 120 Americans, what is the probability at most 12% have hearing trouble?

**Answer.** We are looking for  $P(\hat{p} \leq 12\%)$ , where  $\hat{p}$  is sample proportion with sample size 120, viewed as a variable. Can we say the distribution of this variable  $\hat{p}$  is approximately normal? If so, we can use change of variable and the Standard Normal Distribution Table to calculate  $P(\hat{p} \leq 12\%)$ .