0521 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 $\overline{x} \pm s$, $\overline{x} \pm 2s$, $\overline{x} \pm 3s$ to 68%, 95%, 99.7%, resp.
3. Methods using technology or more advanced math tools (not covered in this course).

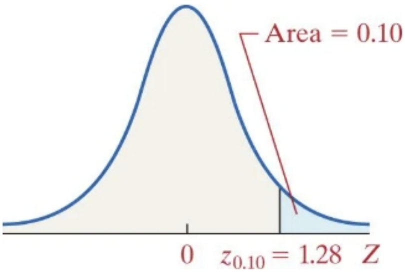
In some specific situation, we do have conventional standards to tell if an approximation is good or bad.

z_α

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

Definition of z_α

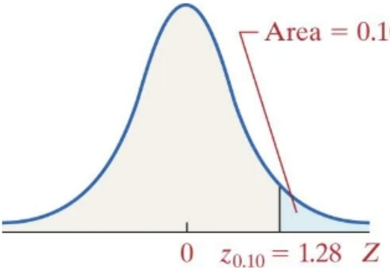
Given a number $0 \leq \alpha \leq 1$, z_α is defined by $P(z \geq z_\alpha = \alpha)$.



Example

Definition of z_α

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



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

Property of z_α

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 choosen 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.

Sampling distribution

Important concepts: Population Parameter vs. Sample Statistic.

	Population Parameter	**Sample Statistic**
Mean	μ	$\overline{x}$
Median	η	M
Variance	σ^2	s^2
Standard Deviation	σ	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 $\overline{x}$ of a random sample of size n?

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

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

Theorem 1.

Fix a population of size N with population mean μ and population standard deviation σ. Fix a sample size n such that $n < 0.05N$. Consider the random variable $\overline{x}$ of sample mean of random samples of size n. Then the random variable $\overline{x}$ has mean

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

and a standard deviation

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

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

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

Theorem.

If a random variable x is (approximately) normally distributed, then the probability distribution of $\overline{x}$ would also be approximately normal. This result is independent of n except for the assumption that $n < 0.05N$.

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

Central Limit Theorem

As n increases, the distribution of $\overline{x}$ becomes more and more normal. When $n \geq 30$, we claim that the bell curve is a good approximation of the distribution of $\overline{x}$, i.e. the distribution of $\overline{x}$ 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 $\overline{x}$?

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 $\overline{x}$?

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

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	μ	$\overline{x}$
Median	η	M
Variance	σ^2	s^2
Standard Deviation	σ	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 $p = \dfrac{a}{N}$ 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 $p = \dfrac{123694}{652442} = 0.19$.

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 $\hat{p} = \dfrac{b}{n}$ 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 $\mathbf{Age} \geq \mathbf{60}$. Suppose in my sample there are 5 people who are over 60 years old. Then $\hat{p} = \dfrac{5}{20} = 0.25$.

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.

Assume $n \leq 0.05N$, the following results hold

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. $\mu_{\hat{p}} = p$ regardless of what n we choose.
3. $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$.

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 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 $\hat{p} = \dfrac{b}{n}$ 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 3.1×10^8. 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 $P(\hat{p} \leq 12\%)$.