MSIT 431 Probability and Statistical Methods

Chapter 5 Sampling Distributions

Dongning Guo

Fall 2017

Chapter 5 Sampling Distributions

5.1 Toward Statistical Inference

5.2 The Sampling Distribution of a Sample Mean

5.3 Sampling Distributions for Counts and Proportions

5.1 Toward statistical inference

- Parameters and statistics
- Sampling variability
- Sampling distribution
- Bias and variability
- Sampling from large populations

Parameters and Statistics

A **parameter** is a number that describes some characteristic of the population. In statistical practice, the value of a parameter is not known because we cannot examine the entire population.

A **statistic** is a number that describes some characteristic of a sample. The value of a statistic can be computed directly from the sample data. We often use a statistic to estimate an unknown parameter.

We write μ (the Greek letter mu) for the population mean and σ for the population standard deviation. We write $\bar{x}$ (x-bar) for the sample mean and s for the sample standard deviation.

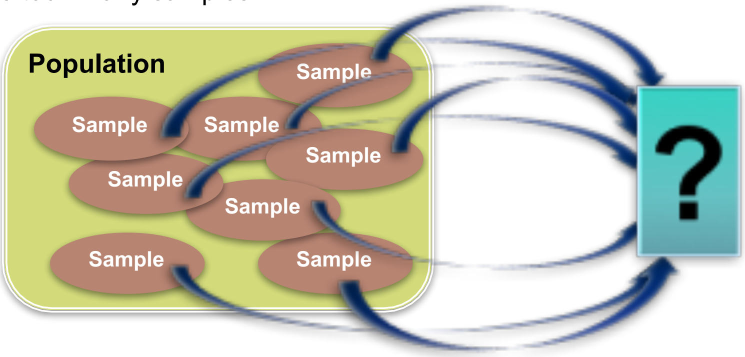
Statistical Estimation

- We estimate a parameter (we don't know its value) using statistics from the sample we have.
- *Different random samples yield different statistics.* So a statistic is in general not equal to the parameter.
- How does the statistic differ from the parameter?
- The **sampling distribution** of a statistic consists of all possible values of the statistic and the relative frequency with which each value occurs.
- We use the statistic and the sampling distribution to carry out **statistical inference** about a wider population.

Sampling Variability

Sampling variability is a term used for the fact that the value of a statistic varies in repeated random sampling.

To make sense of sampling variability, we ask, "What would happen if we took many samples?"



Sampling Distributions

- If we measure enough subjects, the statistic will be very close to the unknown parameter that it is estimating.
- If we took every one of the possible samples of a certain size, calculated the sample mean for each, and made a histogram of all of those values, we'd have a **sampling distribution.**

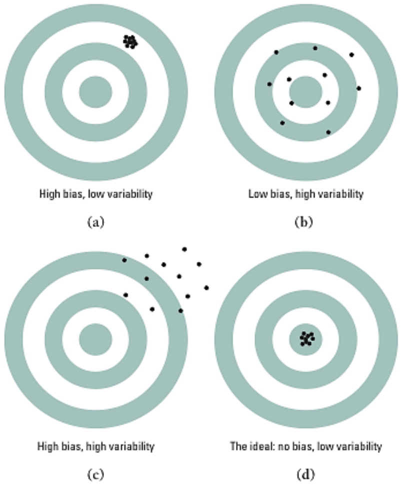
The **population distribution** of a variable is the distribution of values of the variable among all individuals in the population.

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

In practice, it's difficult to take all possible samples of size *n* to obtain the actual sampling distribution of a statistic. Instead, we can use **simulation** to imitate the process of taking many, many samples.

Bias and Variability

We can think of the true value of the population parameter as the bull's-eye on a target and of the sample statistic as an arrow fired at the target. Bias and variability describe what happens when we take many shots at the target.



High bias, low variability (a)

Low bias, high variability (b)

High bias, high variability (c)

The ideal: no bias, low variability (d)

Bias concerns the center of the sampling distribution. A statistic used to estimate a parameter is **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size *n*. Statistics from larger probability samples have smaller spreads.

Managing Bias and Variability

A good sampling scheme must have both small bias and small variability.

To reduce bias, use random sampling.

To reduce variability of a statistic from an SRS, use a larger sample.

The variability of a statistic from a random sample does not depend on the size of the population if it is many times larger than the sample size.

Why Randomize?

1. To eliminate bias in selecting samples from the list of available individuals.

2. The laws of probability allow trustworthy inference about the population.
   - Results from random samples come with a **margin of error** that sets bounds on the size of the likely error.
   - Larger random samples give better information about the population than smaller samples.

5.2 The Sampling Distribution of a Sample Mean

- Population distribution vs. sampling distribution
- The mean and standard deviation of the sample mean
- Sampling distribution of a sample mean
- Central limit theorem

Parameters and Statistics

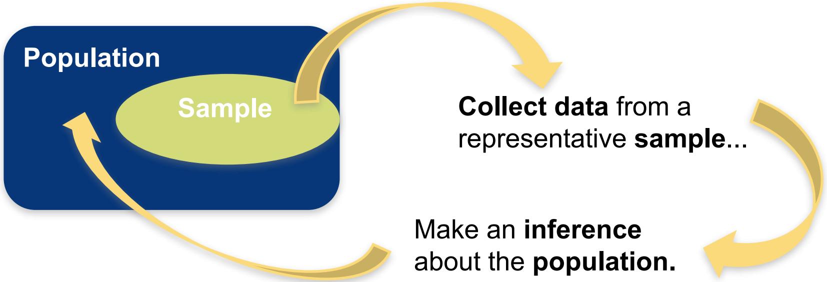
- A **parameter** is a number that describes some characteristic of the population. In statistical practice, the value of a parameter is not known because we cannot examine the entire population.
- A **statistic** is computed directly from the sample data. We often use a statistic to estimate an unknown parameter.
- We write *μ* (the Greek letter mu) for the population mean and *σ* for the population standard deviation. We write (x-bar) for the sample mean and *s* for the sample standard deviation.

Statistical inference

We draw conclusions about a population based on a sample.

Different samples yield different statistics. This is called **sampling variability.** We need to be able to describe the **sampling distribution** of the possible values of a statistic.

We think of a statistic as a random variable. In the parlance of Chapter 4, a statistic has a probability distribution, which is precisely what we mean by sampling distribution.



Collect data from a representative **sample**...

Make an **inference** about the **population.**

Sampling Distributions

The law of large numbers assures us that if we measure enough subjects, the statistic x-bar will eventually get very close to the unknown parameter *μ*.

If we took every one of the possible samples of a certain size, calculated the sample mean for each, and graphed all of those values, we'd have a **sampling distribution.**

> The **population distribution** of a variable is the distribution of values of the variable among all individuals in the population.
>
> The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Mean and Standard Deviation of a Sample Mean

There is no tendency for a sample mean to fall systematically above or below μ, even if the distribution of the raw data is skewed. Thus, the mean of the sampling distribution of a sample mean is an **unbiased estimate** of the population mean μ.

The standard deviation of the sampling distribution of a sample mean measures how much the sample mean varies from sample to sample. It is smaller than the standard deviation of the population by a factor of √*n*.

➔ **Averages are less variable than individual observations.**

The sampling distribution of a sample mean

When we choose many SRSs from a population, the sampling distribution of the sample mean is centered at the population mean *µ* and is less spread out than the population distribution.

The Sampling Distribution of Sample Means

Suppose that $\bar{x}$ is the mean of an SRS of size n drawn from a large population with mean μ and standard deviation σ. Then:

The **mean** of the sampling distribution of $\bar{x}$ is $\mu_{\bar{x}} = \mu$

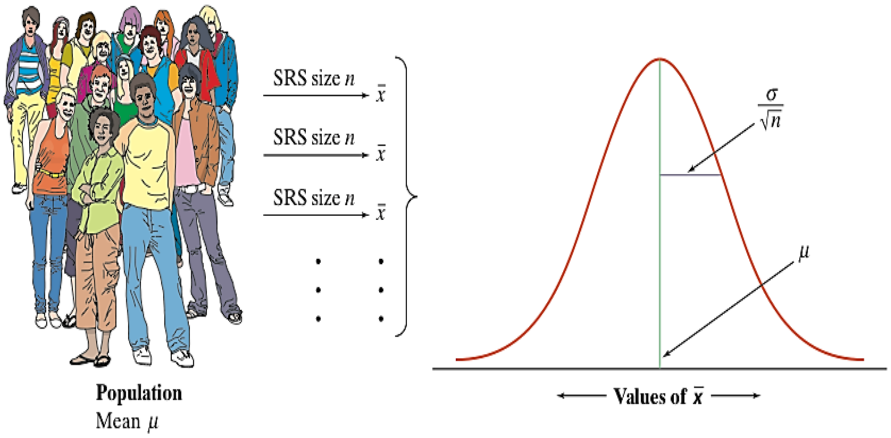
The **standard deviation** of the sampling distribution of $\bar{x}$ is

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

Note : These facts about the mean and standard deviation of $\bar{x}$ are true *no matter what shape the population distribution has.*

If individual observations have the *N*(*µ*,*σ*) distribution, then the sample mean of an SRS of size *n* has the *N*(*µ*, σ/√*n*) distribution, regardless of the sample size *n*.

The sampling distribution of a sample mean



The Central Limit Theorem

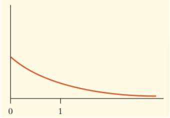
Most population distributions are not Normal. What is the shape of the sampling distribution of sample means when the population distribution isn't Normal?

It is a remarkable fact that, *as the sample size increases, the distribution of sample means begins to look more and more like a Normal distribution!*

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. The **central limit theorem (CLT)** says that when n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Example



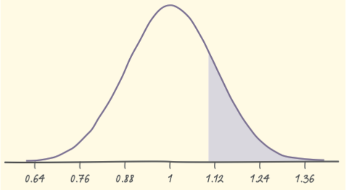
Based on service records from the past year, the time (in hours) that a technician requires to complete preventive maintenance on an air conditioner follows a distribution that is strongly right-skewed, and whose most likely outcomes are close to 0. The mean time is $\mu = 1$ hour and the standard deviation is $\sigma = 1$.

Your company will service an SRS of 70 air conditioners. You have budgeted 1.1 hours per unit. Will this be enough?

The mean and standard deviation of the sampling distribution of the average time spent working on the 70 units are:

$$\mu_{\bar{x}} = \mu = 1 \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{70}} = 0.12$$

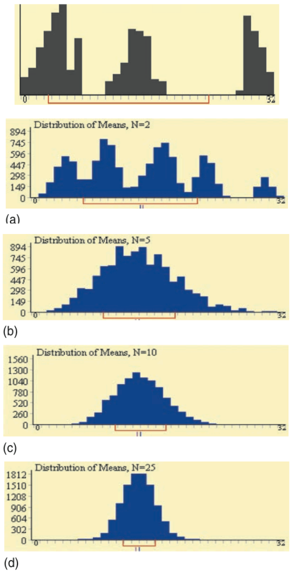
The central limit theorem says that the sampling distribution of the mean time spent working is approximately *N*(1, 0.12) because *n* = 70 ≥ 30.



$$z = \frac{1.1 - 1}{0.12} = 0.83 \qquad \begin{aligned} &P(\bar{x} > 1.1) = P(Z > 0.83) \\ &= 1 - 0.7967 = 0.2033 \end{aligned}$$

If you budget 1.1 hours per unit, there is a 20% chance the technicians will not complete the work within the budgeted time.

A Few More Facts



(a)

(b)

(c)

(d)

Any linear combination of independent Normal random variables is also Normally distributed.

More generally, the central limit theorem notes that the distribution of a sum or average of many small random quantities is close to Normal.

Finally, the central limit theorem also applies to discrete random variables.

5.3 Sampling Distributions for Counts and Proportions

- Binomial distributions for sample counts
- Binomial distributions in statistical sampling
- Finding binomial probabilities
- Binomial mean and standard deviation
- Sample proportions
- Normal approximation for counts and proportions
- Binomial formula

The Binomial Setting

A **binomial setting** arises when we perform several independent trials of the same chance process and record the number of times that a particular outcome occurs. The four conditions (called **BINS**) for a binomial setting are:

- **B**inary? The possible outcomes of each trial can be classified as "success" or "failure."
- **I**ndependent? Trials must be **independent;** that is, knowing the result of one trial must not have any effect on the result of any other trial.
- **N**umber? The number of trials *n* of the chance process must be fixed in advance.
- **S**uccess? On each trial, the probability *p* of success must be the same.

Binomial Distribution

- Consider tossing a coin *n* times. Knowing the outcome of one toss does not change the probability of an outcome on any other toss.
- The number of heads in *n* tosses is a binomial random variable *X*. The probability distribution of *X* is called a **binomial distribution.**

Binomial Distribution
The count *X* of successes in a binomial setting has the **binomial distribution** with parameters *n* and *p*, where *n* is the number of trials of the chance process and *p* is the probability of a success on any one trial. The possible values of *X* are the whole numbers from 0 to *n*.

Form of the binomial distribution

In a binomial setting with *n* trials and success probability *p*, the probability of exactly *x* successes is

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

In this formula, $x!$ means $x(x-1)(x-2)$...2(1). For example, 5!=5(4)(3)(2)(1)=120.

Binomial Formula

Here, we justify the expression given previously for the binomial distribution. First of all, we can find the chance that a binomial random variable takes any value by *adding probabilities for the different ways of getting exactly that many successes in n observations*.

The number of ways of arranging *k* successes among *n* observations is given by the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

for *k* = 0, 1, 2, …, *n.*

25

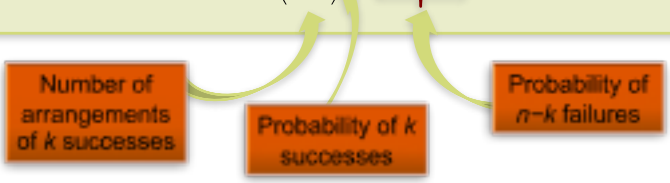
Binomial Probability

The binomial coefficient counts the number of different ways in which *k* successes can be arranged among *n* trials. The binomial probability *P*(*X* = *k*) is this count multiplied by the probability of any one specific arrangement of the *k* successes.

Binomial Probability

If *X* has the binomial distribution with *n* trials and probability *p* of success on each trial, the possible values of *X* are 0, 1, 2, …, *n*. If *k* is any one of these values,

$$P(X = k) = \binom{n}{k} p^k \underbrace{(1-p)^{n-k}}$$



Number of arrangements of *k* successes

Probability of *k* successes

Probability of *n*−*k* failures

26

Example

Each child of a particular pair of parents has probability 0.25 of having blood type O. Suppose the parents have five children.

(a) Find the probability that exactly three of the children have type O blood.

Let *X* = the number of children with type O blood. We know *X* has a binomial distribution with *n* = 5 and *p* = 0.25.

$$P(X=3)=\binom{5}{3}(0.25)^3(0.75)^2=10(0.25)^3(0.75)^2=0.08789$$

(b) What is the probability that more than three of their children have type O blood?

$$\begin{aligned}P(X>3)&=P(X=4)+P(X=5)\\&=\binom{5}{4}(0.25)^4(0.75)^1+\binom{5}{5}(0.25)^5(0.75)^0\\&=5(0.25)^4(0.75)^1+1(0.25)^5(0.75)^0\\&=0.01465+0.00098=0.01563\end{aligned}$$

Binomial Distributions in Statistical Sampling

The binomial distributions are important in statistics when we want to make inferences about the proportion *p* of successes in a population.

Suppose 10% of CDs have defective copy-protection schemes that can harm computers. A music distributor inspects an SRS of 10 CDs from a shipment of 10,000. Let *X* = number of defective CDs in the SRS of size 10.

What is *P*(*X* = 0)? Note: This is not quite a binomial setting. Why?

The actual probability is $P(\text{no defectives})=\frac{9000}{10000}\cdot\frac{8999}{9999}\cdot\frac{8998}{9998}\cdot\ldots\cdot\frac{8991}{9991}=0.3485$

Sampling Distribution of a Count

Choose an SRS of size *n* from a population with proportion *p* of successes. When the population is much larger than the sample, the count *X* of successes in the sample has approximately the binomial distribution with parameters *n* and *p*.

Using the binomial distribution, $P(X=0)=\binom{10}{0}(0.10)^0(0.90)^{10}=0.3487$

Binomial Mean and Standard Deviation

If a count *X* has the binomial distribution based on *n* observations with probability *p* of success, what is its mean *μ*? In general, the mean of a binomial distribution is *μ* = *np.* Here are the facts:

Mean and Standard Deviation of a Binomial Random Variable

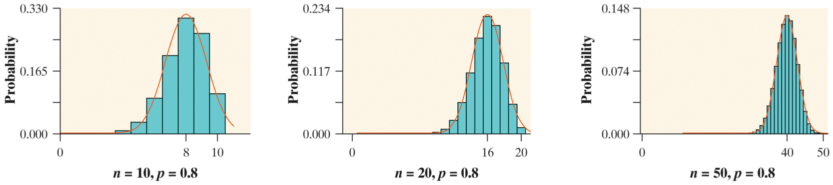
If a count *X* has the binomial distribution with number of trials *n* and probability of success *p*, the **mean** and **standard deviation** of *X* are:

$$\mu_X = np$$
$$\sigma_X = \sqrt{np(1-p)}$$

Note: These formulas work ONLY for binomial distributions. They can't be used for other distributions!

Normal Approximation for Binomial Distributions

As *n* gets larger, something interesting happens to the shape of a binomial distribution.



n = 10, *p* = 0.8 *n* = 20, *p* = 0.8 *n* = 50, *p* = 0.8

Normal Approximation for Binomial Distributions

Suppose that *X* has the binomial distribution with *n* trials and success probability *p*. When *n* is large, the distribution of *X* is approximately Normal with mean and standard deviation

$$\mu_X = np \qquad \sigma_X = \sqrt{np(1-p)}$$

As a rule of thumb, we will use the Normal approximation when *n* is so large that $np \geq 10$ and $n(1-p) \geq 10$.

Sampling Distribution of a Sample Proportion

There is an important connection between the sample proportion $\hat{p}$ and the number of "successes" X in the sample.

$$\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$

Sampling Distribution of a Sample Proportion

Choose an SRS of size n from a population of size N with proportion p of successes. Let $\hat{p}$ be the sample proportion of successes. Then:

The **mean** of the sampling distribution is p.

The **standard deviation** of the sampling distribution is

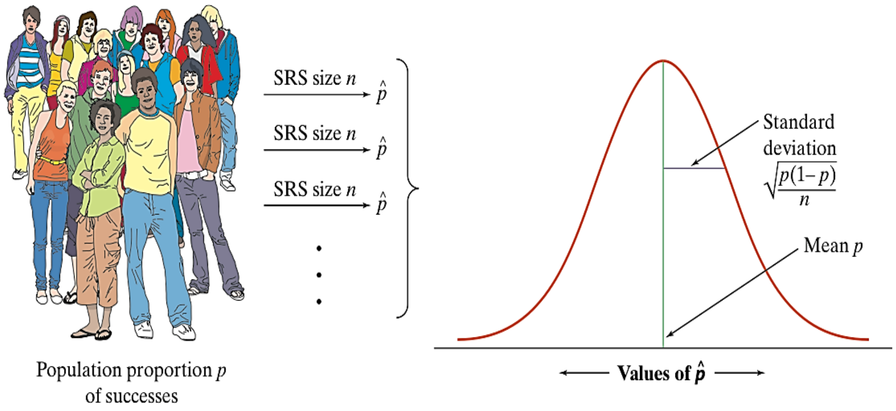
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

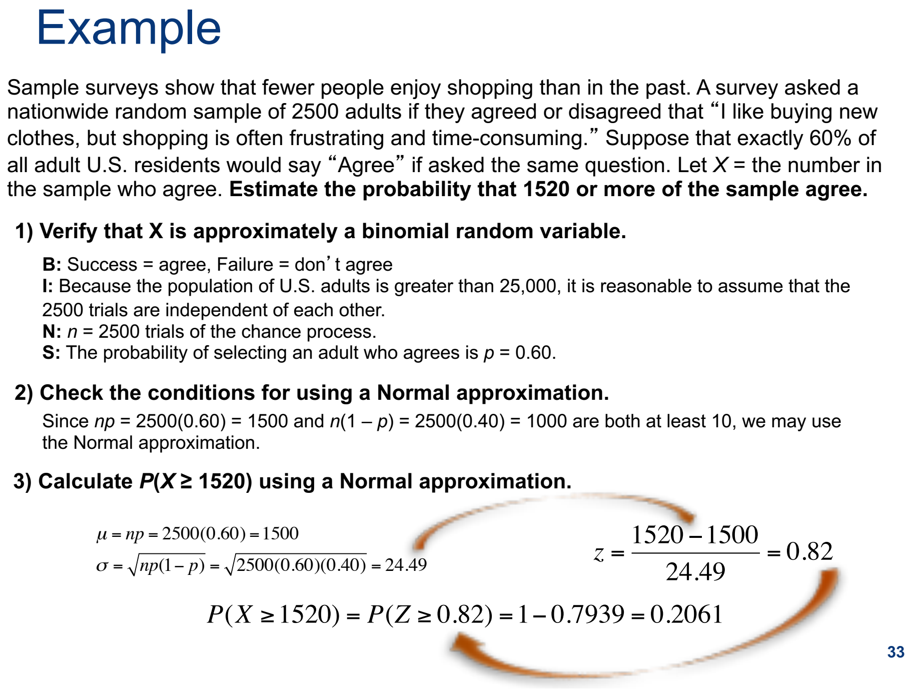
For large n, $\hat{p}$ has approximately the $N(p, \sqrt{p(1-p)/n})$ distribution.

As *n* increases, the sampling distribution becomes **approximately Normal**.

Sampling Distribution of a Sample Proportion

$$\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$





Example

Sample surveys show that fewer people enjoy shopping than in the past. A survey asked a nationwide random sample of 2500 adults if they agreed or disagreed that “I like buying new clothes, but shopping is often frustrating and time-consuming.” Suppose that exactly 60% of all adult U.S. residents would say “Agree” if asked the same question. Let *X* = the number in the sample who agree. **Estimate the probability that 1520 or more of the sample agree.**

1) Verify that X is approximately a binomial random variable.

B: Success = agree, Failure = don’t agree
I: Because the population of U.S. adults is greater than 25,000, it is reasonable to assume that the 2500 trials are independent of each other.
N: *n* = 2500 trials of the chance process.
S: The probability of selecting an adult who agrees is *p* = 0.60.

2) Check the conditions for using a Normal approximation.

Since *np* = 2500(0.60) = 1500 and *n*(1 – *p*) = 2500(0.40) = 1000 are both at least 10, we may use the Normal approximation.

3) Calculate *P*(*X* ≥ 1520) using a Normal approximation.

$$\mu = np = 2500(0.60) = 1500$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{2500(0.60)(0.40)} = 24.49$$

$$z = \frac{1520 - 1500}{24.49} = 0.82$$

$$P(X \geq 1520) = P(Z \geq 0.82) = 1 - 0.7939 = 0.2061$$

33