Chapter 1

Sampling Distributions

The purpose of sampling and experimentation is to make inference about the population, and sampling distribution plays a pivotal role in connecting them. This chapter is also the bridge to connect all the knowledge we learned in previous chapter with the following of chapters of **statistical inference**.

1.1 What is sampling distribution?

Statistics and Parameters

The purpose of sampling is to make inferences about the population. Before we move on we have to clarify two concepts:**statistics** and **parameters**.

A **statistic** is a random variable used to describe the characteristic of a sample.

A **parameter** is a number that describes some characteristic of the population.

Take a look at the following two examples.

(1) Each month, the Current Population Survey (CPS) interviews a random sample of individuals in about 60,000 U.S. households. The CPS uses the proportion of unemployment in this sample to estimate the the national unemployment rate.

(2) Selected a random sample of 100 students from a high school. Use the average BMI(Body Mass Index) of this sample to estimate the average BMI of all the students in this high school.

In the above two examples, the unemployment rate of the sample and the average BMI of the sample are **statistics**. And the reason why statistics are random variables is that for different samples, they may take different values.

In the above two examples, "the national unemployment rate" and "the average BMI of all the students in this high school" are **parameters**. Parameters of a population are numbers, not random variables, because they never change, though we may don't know the real values.

Some of the conventional notations for statistics are $\widehat{p}, \overline{x}, s_X^2$, they are sample proportion, sample mean and sample standard deviation. The corresponding notations of parameters are p, μ, σ^2.

In AP exam, don't use wrong notations, otherwise points will be deducted. We will learn more conventional notations.

Sampling distributions

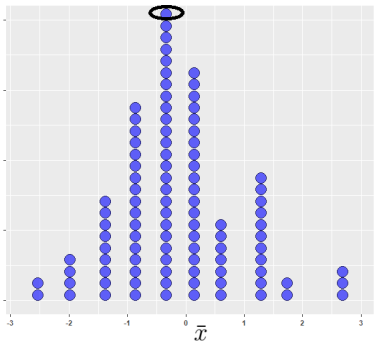
Since a statistic is a random variable, and its value changes for different samples, its distribution may be described if we draw many many samples.

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.

Table 1.1 gives a simple random sample with sample size 25, sampled from a population with distribution $N(0, 5)$. The mean of this sample $\bar{x} = -0.418$. Figure 1.1 shows the sampling distribution of $\bar{x}$ of 100 such samples. The circled point is a possible point for the sample given in table 1.1.

1.033	0.767	0.915	1.824	8.305	-9.884	1.677	-1.214	-0.837
-2.392	-0.441	-4.682	0.634	1.910	-5.532	5.244	-10.623	-0.321
0.960	8.066	-8.139	-9.221	-1.955	8.784	4.667		

Table 1.1: One SRS from population with population distribution N(0, 5)



Is the distribution of the sample data in table 1.1 the same as the sampling distribution in figure 1.1?

Figure 1.1: The sampling distribution of $\bar{x}$ of 100 samples with size 25

Distinguish **the sampling distribution**, **the distribution of the sample data** and **the population distribution**.

Draw simple random samples with sample size 20 from a population with 50 black balls and 150 white balls. The statistics of the proportion of black balls in the samples is denoted $\hat{p}$. Figure 1.2 gives the sampling distribution $\hat{p}$ of 100 such samples.

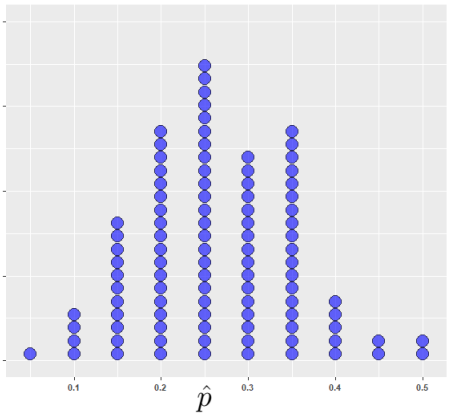


Figure 1.2: Sampling distribution of $\hat{p}$

Unbiased estimator

A statistic is a random variable used to estimate the corresponding parameter and sometimes this statistic is called an **estimator** of the parameter. In the following, we will use *estimator* and *statistic* interchangeablely.

Each value of this estimator is an **estimate**. If the mean(expected value) of the this estimator equals the parameter to be estimated, it is an **unbiased estimator**.

$\widehat{p}, \overline{x}, s_X^2$ are unbiased estimators of p, μ, σ^2, which means

$$\mu_{\widehat{p}} = p, \qquad \mu_{\overline{x}} = \mu, \qquad \mu_{s_X^2} = \sigma^2.$$

Take a look at figure 1.1 and figure 1.2, the center of the sampling distribution is approximately around the parameter to be estimated.

Variability of a statistic

When a statistic is used to estimate a parameter, we expect it is an unbiased estimator, that is the center of the sampling distribution equal the parameter and the estimation is *accurate*. We also expect the sampling distribution of the statistic is small, in this way the estimation is more *precise*

If the sample is more representative of the population, the estimates are more close to the parameter, the variability of the estimator is smaller. Previously, we learned *stratified random sample* is more representative than SRS, and the sampling distribution of those samples has a smaller variability. The other way to make a sample more representative is to increase sample size. The larger the sample size the smaller the variability of the sampling distribution of the statistics. Latter we will learn how to quantify the relationship between the sample size and the variability of the distribution of the statistics. But, we need to know one fact now:

The **variability of a statistic** is described by the spread of its sampling distribution. Larger samples give smaller spreads. The spread of the sampling distribution does not depend much on the size of the population, as long as the population is at least 10 times larger than the sample.

In figure 1.3 take the bull's–eys as the parameter to be estimated.

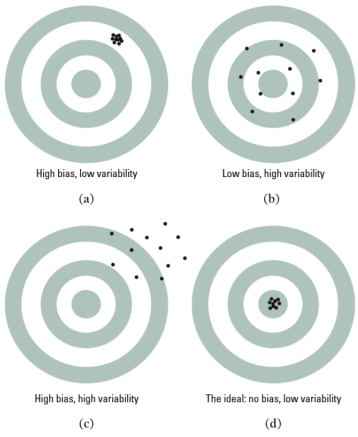
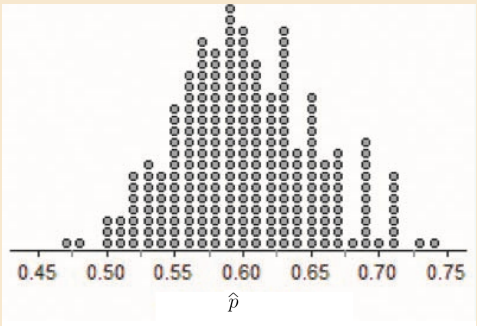


Figure 1.3: Illustration of the variability and bias

Doing Homework!

A school newspaper article claims that 60% of the students at a large high school did all their assigned homework last week. Some skeptical AP® Statistics students want to investigate whether this claim is true, so they choose an SRS of 100 students from the school to interview. What values of the sample proportion $\hat{p}$ would be consistent with the claim that the population proportion of students who completed all their homework is $p = 0.6$? To find out, we used Fathom software to simulate choosing 250 SRSs of size $n = 100$ students from a population in which $p = 0.60$. The figure below is a dotplot of the sample proportion $\hat{p}$ of students who did their homework.



(a) There is one dot on the graph at 0.73. Explain what this value represents.

(b) Describe the distribution. Are there any obvious outliers?

(c) Would it be surprising to get a sample proportion of 0.45 or lower in an SRS of size 100 when $p = 0.6$? Justify your answer.

(d) Suppose that 45 of the 100 students in the actual sample say that they did all their homework last week. What would you conclude about the newspaper articles claim? Explain.

1.2 Sampling distribution of $\hat{p}$

Suppose there are N balls, the proportion of black balls is p, and the others are white balls. Pick an SRS of size n(without replacement), and the proportion of black balls in this sample is $\hat{p} = \frac{X}{n}$, where X is the number of black balls in the sample. Let's find out the sampling distribution of the statistic $\hat{p}$.

(1) **Use classical probability model to calculate $\mathbf{P}(\hat{p} = r)$.**

Can you make a simple deduction about the results on the right?

The number of all possible outcomes with $\hat{p} = r$ is given by

$$ {}_{nr}\mathbf{C}_{Np} \times {}_{n(-r)}\mathbf{C}_{N}(1-p). $$

The number of all possible outcomes is given by: ${}_{n}\mathbf{C}_{N}$

$$ \mathbf{P}(\hat{p} = r) = \frac{{}_{nr}\mathbf{C}_{Np} \times {}_{n(-r)}\mathbf{C}_{N}(1-p)}{{}_{n}\mathbf{C}_{N}} $$

(2) **Use binomial distribution to calculate$\mathbf{P}(\hat{p} = r)$.**

Suppose $n \leq 10\%\, N$, the **10 % condition** is met. The outcome for the individuals in the sample are approximately independent with each other. X is approximately $\mathbf{B}(n, n\hat{p})$

$$ \mathbf{P}(\hat{p}) = {}_{nr}\mathbf{C}_{n}\, p^{nr}\, (1-p)^{n-nr} $$

(3) **Use Normal distribution**

In addition to the **10% condition**, the **large counts condition** is met($np \leq 10, \quad n(1-p) \leq 10$). Then X is approximately normal.

$$ X \sim \mathbf{N}(\mu_X, \sigma_X), \qquad \mu_X = np, \quad \sigma_X = \sqrt{np(1-p)} $$

$\hat{p} = \frac{X}{n}$ is normally distributed according to the knowledge of *transforming random variables*, $\hat{p} \sim N(\mu_{\hat{p}}, \sigma_{\hat{p}})$. The $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$ are given bellow.

What conditions must be met in order to use the formulas for $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?

$$ \mu_{\hat{p}} = \frac{1}{n}\mu_X = p $$

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

From the above equations we can come up with two conclusions

Explain why we can come up with those two conclusions?

I. $\hat{p}$ is an unbiased estimator of p.

II. If the sample size increases by timing factor C, the standard deviation of $\hat{p}$ decreases by timing factor $\frac{1}{\sqrt{C}}$.

Do you go to church?

The Gallup Poll asked a random sample of 1785 adults whether they attended church during the past week. Let $\hat{p}$ be the proportion of people in the sample who attended church. A newspaper report claims that 40% of all U.S. adults went to church last week. Suppose this claim is true.

(a) What is the mean of the sampling distribution of $\hat{p}$. Why?

(b) Find the standard deviation of the sampling distribution of $\hat{p}$. Check to see if the 10% condition is met.

(c) Is the sampling distribution of $\hat{p}$ approximately normal? Check to see if the large counts condition is met?

(d) Of the respondents, 44% said they did attend church last week. Find the probability of obtaining a sample of 1785 adults in which 44% or more say they attended church last week if the newspaper report's claim is true. Dose this poll give convincing evidence against the claim? Explain.

1.3 The sampling distribution of $\bar{x}$

Choose an SRS of size n from a population of size N, and variable X_i is the measurement on the ith individual in the sample.

(1) **Suppose the population is normally distributed** $N(\mu, \sigma)$.

$$X_i \sim N(\mu, \sigma), \quad \mu_{X_i} = \mu, \quad \sigma_{X_i} = \sigma, \qquad i = 1, 2, \cdots, n.$$

$$\text{sample mean:} \quad \overline{x} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mu_{\overline{x}} = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Why the **10% condition** is required for the formula of $\sigma_{\overline{x}}$ here?

Is the **10% condition** required for the formula of $\mu_{\overline{x}}$?

- **If 10% condition ($\mathbf{n \leq 10\%\, N}$) is met.**

$$\mathbf{Var}(\overline{x}) = (\frac{1}{n})^2\ \mathbf{Var}(\sum_{i=1}^{n} X_i) = (\frac{1}{n})^2\ \sum_{i=1}^{n} \mathbf{Var}(X_i) = \frac{\sigma^2}{n}$$

$$\sigma_{\overline{x}} = \sqrt{\mathbf{Var}(\overline{x})} = \frac{\sigma}{\sqrt{n}}$$

If the **10% condition** ($\mathbf{n \leq 10\%\, N}$) is met and the population is normally distributed, then, according to the knowledge of "combination of independent normal random variables" in *chapter 5*, $\overline{x}$ is normally distributed.

$$\overline{x} \sim N(\mu_{\overline{x}}, \sigma_{\overline{x}})$$

The $\mu_{\overline{x}}$ and $\sigma_{\overline{x}}$ are given by the above formulas.

(2) **Don't know whether the population distribution is normal or not**.

The sampling distribution of a statistic is closely related to the sampling distribution. However in the case of sample distribution, there is a very important and elegant theorem, called **central limit theorem(CLT)**, which guarantees the sampling distribution of sample mean $\overline{x}$ is approximately normal as the sample size n increases. Let's take a look at some simulations.

We will sample from a population with probability density function

$$f(x) = e^{-x} \qquad x \in [0, +\infty).$$

The density curve of this distribution is strongly right skewed as shown by figure 1.4.

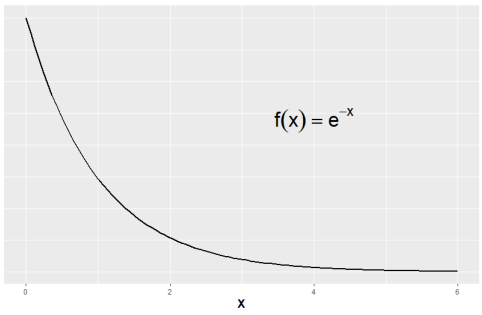


Figure 1.4: A strongly right–skewed distribution

Figure 1.5 shows the sampling distribution of $\overline{x}$ of 500 samples of different sample sizes: $n = 5, 10, 20, 40, 80, 160$. We can tell that the distribution become more and more symmetric. If you pay attention to the range of the distribution, which is a measurement of the variability of the sampling distribution of $\overline{x}$, you will find the range decreases as the sample size increase. This verifies what we learned before about the relationship between sample size and variability.

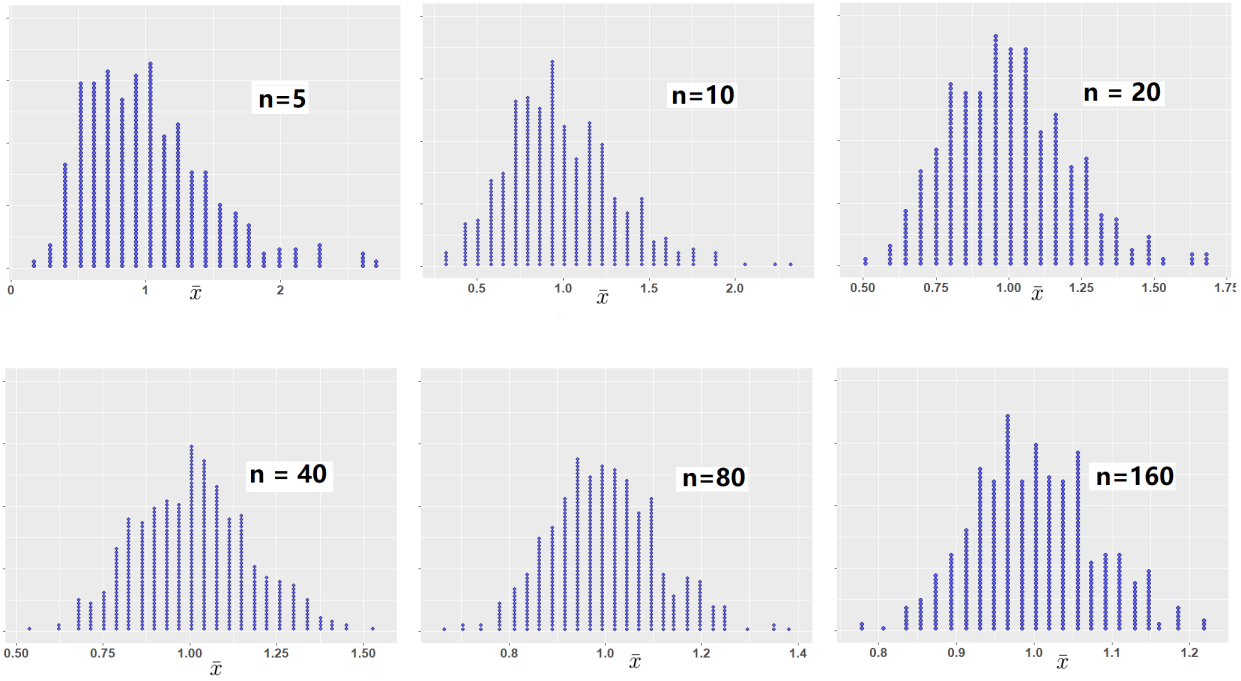


Figure 1.5: Demonstration of **Central Limit Theorem(CLT)**

Central Limit Theorem(CLT)

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 $\overline{x}$ is approximately Normal, $N(\mu, \frac{\sigma}{\sqrt{n}})$.

As a rule of thumb, if $\mathbf{n \geq 30}$, n is large enough invoke **CTL**.

(3) The t-distribution

Previously we learned that if the **10% condition**($n \leq 10\%\, N$) is met and either **(1) the population is normally distributed**, or (2) **the sample is a large sample** (**n** $\geq$ **30**), then the sampling distribution of $\overline{x}$ is approximately normal.

$$\overline{x} \sim N(\mu_{\overline{x}}, \sigma_{\overline{x}}), \qquad \mu_{\overline{x}} = \mu, \quad \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}},$$

where μ and σ are population mean and population standard deviation.

What if the population standard deviation σ is unknown? In this case, the sampling distribution can be described by **t-distribution**.

What is the distribution of

$$z = \frac{\overline{x} - \mu}{s_X/\sqrt{n}}$$

If $\overline{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, then $t = \frac{\overline{x} - \mu}{s_X/\sqrt{n}}$ follows a **t-distribution** with **df** $= n - 1$, where **df** is the **degree of freedom**.

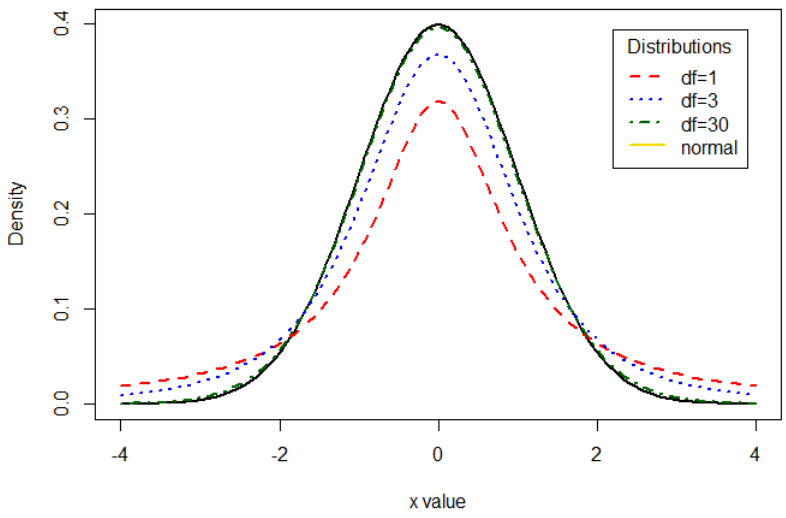


Figure 1.6: t–distributions and Normal distribution

Figure 1.6, gives t-distributions with different degree of freedom and standard normal distribution. we can have the following statements by reading this figure.

- The t-distribution is symmetric with center 0
- The t-distribution has a larger variance than $N(0, 1)$
- The larger the degree of freedom(**df**), the smaller the variance of the t-distribution.
- As **df** increases, the a-distribution approximates to standard normal distribution $N(0, 1)$.

The Principal's Suspicion

A report claims that the height of high school kids follows a Normal distribution with mean $\mu = 168\,\text{cm}$ and standard deviation $\sigma = 9\,\text{cm}$. A high school principal suspect that the average height of the students in this school is higher. In order to test the principal's suspicion, an SRS of 9 students were drawn.

Assume the report's claim is true. Solving the following problems.

(a) Find the probability that a randomly selected young woman is taller than 175 cm. Show your work.

(b) Find the probability that the mean height of the SRS exceeds 175 cm. Show your work.

(c) The heights of the 9 students in the SRS are given below, in centemeters.

$$172 \quad 178 \quad 168 \quad 183 \quad 180 \quad 165 \quad 175 \quad 177 \quad 170$$

Find the statistic t of this sample.

(d) If the population standard deviation($\sigma = 9\,\text{cm}$) is unknown. Find the probability that a randomly chosen SRS of size 9 with t larger than the value in the proceeding problem.

(e) What can you say about the principal's suspicion.

Are We Richer?

The yearly income of Chinese people is strongly right-skewed. According to the report from the government, the average yearly income of Chinese people is $74,318$ ¥ and the standard deviation of the income of the nation is $30,000$ ¥ in 2017.

The leader of a province thought people in this province might be richer. In order to test this though, an SRS of size 100 is drawn, with sample mean $77,000$ ¥ and the standard deviation $40,000$¥ .

Answer the following questions while assuming the report from the government is true for this province as well.

(a) Find the probability a randomly chosen person with a yearly income 77,000 ¥ or higher?

(b) Find the probability a randomly chosen SRS with sample mean 77,000 ¥ or higher?

(c) Suppose the population standard deviation is unknown and and the standard deviation of the sample is $40,000$¥ ., find the t value of this sample.

(d) Find the probability that a randomly chosen SRS of size 100 has t value larger than the value in previous problem.

(e) Do you think people are richer in this province according to the result in (d)?