CHAPTER 4: DISTRIBUTIONS OF SAMPLING STATISTICS

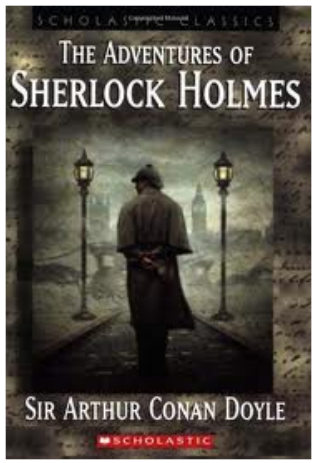
STATISTICS (FERM)
Lecturer: Nguyen Minh Quan, PhD
quannm@hcmiu.edu.vn

CONTENTS

Statistical Inference

"We have got to the deductions and the inferences", said inspector Lestrade, winking at me.
"I find it's hard enough to tackle facts, Holmes, without flying away after theories and fancies"

Inspector Lestrade to Sherlock Holmes.
The Boscome Valley Mystery



Introduction

- Statistics is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of conclusions.
- By suitably sampling from this collection, and then analyzing the sampled items, one hopes to be able to draw some conclusions about the collection as a whole.
- It is necessary to make some assumptions about the relationship between the sample and population.
- In this chapter, we will be concerned with the probability distributions of two important statistics: the sample mean and the sample variance. **[Ref. Chapter 6 in the textbook by S. Ross]**

The sample mean

Definition

If $X_1, \cdots, X_n$ are independent random variables having a common distribution F, then we say that they constitute a sample (sometimes called a random sample) from the distribution F.

Definition

Let $X_1, \cdots, X_n$ be a sample from a distribution F. The sample mean is defined by

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

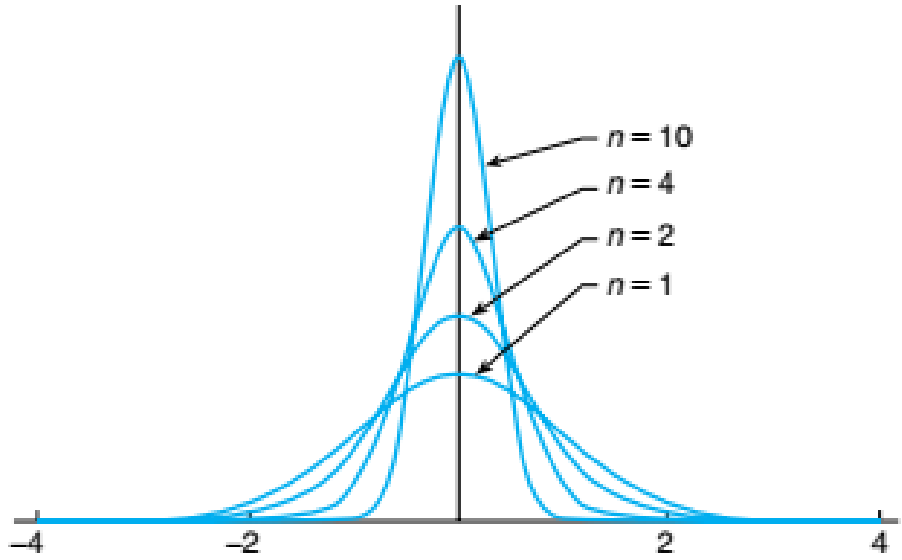
For population, the quantities μ and σ^2 are called the population mean and the population variance, respectively.

Note: $\bar{X}$ is a RV.

The sample mean

Theorem

$$E\left[\bar{X}\right] = \mu, \quad Var\left(\bar{X}\right) = \frac{\sigma^2}{n}$$



Densities of sample means from a standard normal population.

$\bar{X}$ is also centered about the population mean μ, but its spread becomes more and more reduced as the sample size increases.

The Central Limit Theorem

Theorem

Let $X_1, X_2, ..., X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for n large, the distribution of $X_1 + \cdots + X_n$ is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

It follows from the central limit theorem that $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$ is approximately a standard normal random variable; thus, for n large,

$$P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} < x\right) \simeq P(Z < x)$$

where Z is a standard normal random variable.

The Central Limit Theorem

Example

An insurance company has $25,000$ automobile policy holders. If the yearly claim of a policy holder is a random variable with mean 320 and standard deviation 540, approximate the probability that the total yearly claim exceeds 8.3 million.

Solution

Let X denote the total yearly claim. Number the policy holders, and let X_i denote the yearly claim of policy holder i. With $n = 25,000$, we have from the central limit theorem that $X = \sum_{i=1}^{n} X_i$ will have approximately a normal distribution with mean $320 \times 25,000 = 8 \times 10^6$ and standard deviation $540\sqrt{25000} = 8.5381 \times 10^4$.

The Central Limit Theorem

Example (Cont.)

$$P\left(X > 8.3 \times 10^6\right) = P\left(\frac{X - 8 \times 10^6}{8.5381 \times 10^4} > \frac{8.3 \times 10^6 - 8 \times 10^6}{8.5381 \times 10^4}\right)$$

$$\approx P\left(Z > 3.51\right) = 0.0002 \ (Z \text{ is standard normal})$$

Applications of the CLT in binomial RV

Since a binomial random variable X having parameters (n, p) represents the number of successes in n independent trials when each trial is a success with probability p, we can express it as $X = X_1 + ... + X_n$.

$$X_i = \begin{cases} 1 \text{ if the ith trial is a success} \\ 0 \text{ otherwise} \end{cases}$$

We recall that $E(X_i) = p$ and $Var(X_i) = p(1-p)$.

Corollary

For n large $\frac{X-np}{\sqrt{np(1-p)}}$ will approximately be a standard normal random variable.

The Central Limit Theorem

Example

The ideal size of a first-year class at a particular college is 150 students. The college, knowing from past experience that, on the average, only 30 percent of those accepted for admission will actually attend, uses a policy of approving the applications of 450 students. Compute the probability that more than 150 first-year students attend this college.

Solution

Let X denote the number of students that attend; then assuming that each accepted applicant will independently attend, it follows that X is a binomial random variable with parameters $n = 450$ and $p = 0.3$.

$$P(X > 150.5) = P\left(\frac{X - (450)(0.3)}{\sqrt{450(0.3)(0.7)}} \geqslant \frac{150.5 - (450)(0.3)}{\sqrt{450(0.3)(0.7)}}\right)$$

$$\approx P(Z > 1.59) = 0.06 \ (Z \text{ is standard normal})$$

Approximate Distribution of the Sample Mean

- Let $X_1, ..., X_n$ be a sample from a population having mean μ and variance σ^2. The central limit theorem can be used to approximate the distribution of the sample mean

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

- It follows from the central limit theorem that X will be approximately normal when the sample size n is large.
- Since the sample mean has expected value μ and standard deviation $\sigma/\sqrt{n}$, it then follows that

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

Approximate Distribution of the Sample Mean

Example

The weights of a population of workers have mean 167 and standard deviation 27.
(a) If a sample of 36 workers is chosen, approximate the probability that the sample mean of their weights lies between 163 and 170.
(b) Repeat part (a) when the sample is of size 144.

Solution

Let Z be a standard normal random variable.
(a) It follows from the central limit theorem that X is approximately normal with mean 167 and standard deviation $27/\sqrt{36} = 4.5$. Therefore,

$$\frac{\bar{X} - 167}{4.5} \sim N(0, 1)$$

$$P\left(163 < \bar{X} < 170\right) = P\left(\frac{163 - 167}{4.5} < \frac{\bar{X} - 167}{4.5} < \frac{170 - 167}{4.5}\right)$$

Approximate Distribution of the Sample Mean

Example

Solution (Cont.)

$$\approx P\left(-0.89 < Z < 0.67\right) \approx 0.5619$$

(b) $\frac{\bar{X}-163}{27/\sqrt{144}} \sim N\left(0,1\right) \Rightarrow \frac{\bar{X}-163}{2.25} \sim N\left(0,1\right)$

$$P\left(163 < \bar{X} < 170\right) = P\left(\frac{163-167}{2.25} < \frac{\bar{X}-167}{2.25} < \frac{170-167}{2.25}\right)$$

$$\approx P\left(-1.78 < Z < 1.33\right) \approx 0.8698$$

Thus increasing the sample size from 36 to 144 increases the probability from 0.5619 to 0.8698.

Approximate Distribution of the Sample Mean

Exercise

The mean annual cost of automobile insurance is \$939 (CNBC, 2006). Assume that the standard deviation is $\sigma = \$245$.
(a) What is the probability that a simple random sample of automobile insurance policies will have a sample mean within \$25 of the population mean for each of the following sample sizes: 30, 50, 100, and 400?
(b) What is the advantage of a larger sample size when attempting to estimate the population mean?

Approximate Distribution of the Sample Mean

Exercise

An astronomer wants to measure the distance from her observatory to a distant star. However, due to atmospheric disturbances, any measurement will not yield the exact distance d. As a result, the astronomer has decided to make a series of measurements and then use their average value as an estimate of the actual distance. If the astronomer believes that the values of the successive measurements are independent random variables with a mean of d light years and a standard deviation of 2 light years, how many measurements need she make to be at least 95 percent certain that her estimate is accurate to within ± 0.5 light years?

Hint

$$\frac{\bar{X} - d}{2/\sqrt{n}} \sim N(0, 1)$$

$$P\left(-0.5 < \bar{X} - d < 0.5\right) \geqslant 0.95 \Rightarrow \Phi\left(\frac{\sqrt{n}}{4}\right) \geqslant 0.975 \Rightarrow n \geqslant 62$$

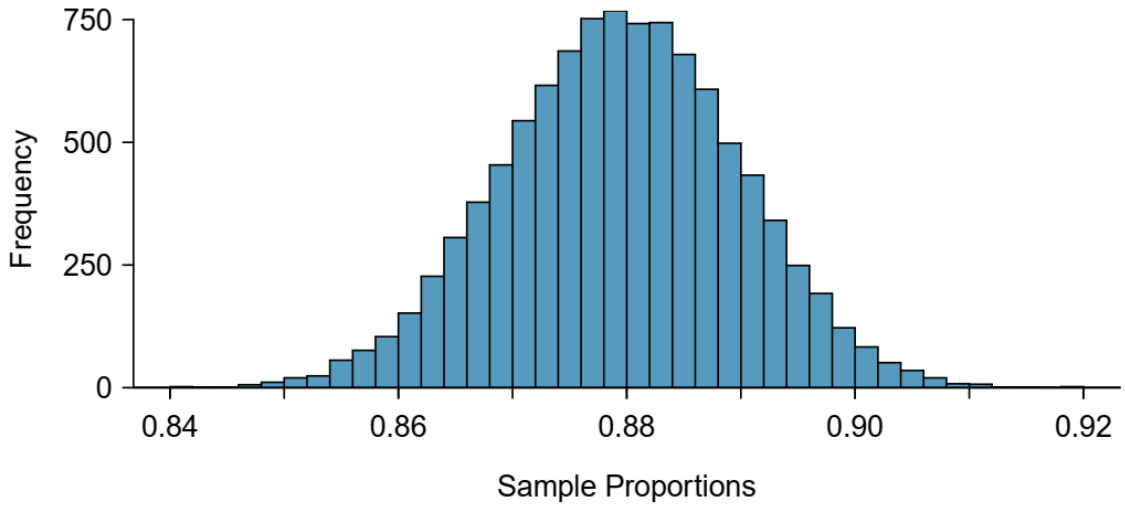
How Large a Sample Is Needed?

- In general, the answer depends on the population distribution of the sample data.
- If the underlying population distribution is normal, then the sample mean $\bar{X}$ will also be normal regardless of the sample size.
- A general rule of thumb is that one can be confident of the normal approximation whenever the sample size n is at least 30.

The Central Limit Theorem for the sample proportion $\hat{p}$

- Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$. Take a poll of 1000 American adults on this topic, the estimate $\hat{p}$ (pronounced $p-$hat, so called a point estimate of p) would not be perfect, but how close might we expect the sample proportion in the poll would be to 88%?

- The figure below is a histogram of 10,000 sample proportions ($p = 0.88, n = 1000$):



The Central Limit Theorem for the sample proportion $\hat{p}$

- Central Limit Theorem: For n large, the sample proportion $\hat{p}$ follows a normal distribution with the following mean and standard error (standard deviation):

$$\mu_{\hat{p}} = p, \quad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$, which is called the success-failure condition.
- In practice we do not know p. We can use the **approximate condition** for the success-failure condition: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.

The Central Limit Theorem for the sample proportion $\hat{p}$

Example

The population proportion is $p = 0.3$. What is the probability that a sample proportion will be within 0.04 of the population proportion for each of the following sample sizes?
(a) $n = 200$
(b) $n = 500$
(c) $n = 1000$
(d) What is the advantage of a larger sample size?

Hint: Use the CLT to compute $P(|\hat{p} - p| < 0.04)$ or $P(-0.04 < \hat{p} - p < 0.04)$ for $n = 200, 500, 1000$ and observe the results.

The Central Limit Theorem for the sample proportion $\hat{p}$

Exercise

The president of Doerman Distributors, Inc., believes that 30% of the firm's orders come from first-time customers. A random sample of 100 orders will be used to estimate the proportion of first-time customers.
(a) Assume that the president is correct and $p = 0.30$. What is the sampling distribution of $\hat{p}$ for this study?
(b) What is the probability that the sample proportion $\hat{p}$ will be between 0.2 and 0.4?

The sample variance S^2

Definition

The statistic S^2, defined by

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

is called the sample variance. $S = \sqrt{S^2}$ is called the sample standard deviation.

Theorem

The expected value of the sample variance S^2 is equal to the population variance σ^2:

$$E(S^2) = \sigma^2$$

Key proof: Use identity $\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n(\bar{X})^2$.

The sample variance S^2

We would like to compute their distributions.

Recall the distribution of the sample mean

$E(\bar{X}) = \mu$ and $Var(\bar{X}) = \sigma^2/n$.

That is, $\bar{X}$, the average of the sample, is normal with a mean equal to the population mean but with a variance reduced by a factor of 1/n. It follows from this that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is a standard normal random variable.

How about the sample variance S^2?

Joint Distribution of $\bar{X}$ and S^2

Identity (relation between S^2 and $\bar{X}$)

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}+\left[\frac{\sqrt{n}\,(\bar{X}-\mu)}{\sigma}\right]^2$$

This implies the df of S^2 is $n-1$:

Theorem

If $X_1,\cdots,X_n$ is a sample from a normal population having mean μ and variance σ^2, then $\bar{X}$ and S^2 are independent random variables, with $\bar{X}$ being normal with mean μ and variance σ^2/n and $(n-1)S^2/\sigma^2$ being chi-square with $n-1$ degrees of freedom.

Joint Distribution of $\bar{X}$ and S^2

Example

The time it takes a central processing unit to process a certain type of job is normally distributed with mean 20 seconds and standard deviation 3 seconds. If a sample of 15 such jobs is observed, what is the probability that the sample variance will exceed 12?

Solution

$$P\left(S^2 > 12\right) = P\left(\frac{14S^2}{9} > \frac{14}{9}12\right)$$

$$= P\left(\chi^2_{14} > 18.67\right) = 0.1779$$

Joint Distribution of $\bar{X}$ and S^2

Corollary

Let $X_1, \cdots, X_n$ be a sample from a normal population with mean μ. If $\bar{X}$ denotes the sample mean and S the sample standard deviation, then

$$\sqrt{n}\frac{(\bar{X}-\mu)}{S} \sim t_{n-1}.$$

Key proof: Recall that a t-random variable with n degrees of freedom is defined as the distribution of

$$\frac{Z}{\sqrt{\chi_n^2/n}}$$

Apply this fact for $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ and $\chi^2_{n-1} = (n-1)S^2/\sigma^2$.

Sampling from a Finite Population

Consider a population of N elements, and suppose that p is the proportion of the population that has a certain characteristic of interest.

Definition

A sample of size n from this population is said to be a random sample if it is chosen in such a manner that each of the $\binom{N}{n}$ population subsets of size n is equally likely to be the sample.

Sampling from a Finite Population

Let

$$X_i = \begin{cases} 1, & \text{if the ith member of the sample has the characteristic} \\ 0, & \text{otherwise} \end{cases}$$

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

X is equal to the number of members of the sample that possess the characteristic.

Fact

$$P(X_i = 1) = p, \text{ and } P(X_i = 0) = 1 - p.$$

Sampling from a Finite Population

Fact: The random variables $X_1, X_2, \cdots, X_n$ are not independent.

Observations

$$P(X_2 = 1|X_1 = 1) = \frac{Np - 1}{N - 1}, \text{ and } P(X_2 = 1|X_1 = 0) = \frac{Np}{N - 1}.$$

Remark

When the population size N is large with respect to the sample size n, then $X_1, X_2, \cdots, X_n$ are approximately independent and we will take the distribution of X to be binomial with

$$E(X) = np \text{ and } SD(X) = \sqrt{np(1 - p)}$$

Sampling from a Finite Population

Example

Suppose that 45 percent of the population favors a certain candidate in an upcoming election. If a random sample of size 200 is chosen, find
(a) the expected value and standard deviation of the number of members of the sample that favor the candidate;
(b) the probability that more than half the members of the sample favor the candidate.

Solution

(a) $E(X) = 200(0.45) = 90$ and $SD(X) = \sqrt{200(0.45)(0.55)} = 7.0356$
(b) Using the normal approximation to the binomial and continuity correction:

$$P(X \geqslant 101) = P(X \geqslant 100.5)$$

$$= P\left(\frac{X - 90}{7.0356} \geqslant \frac{100.5 - 90}{7.0356}\right) \simeq P(Z \geqslant 1.4924) \simeq 0.0678$$

Sampling from a Finite Population

Example

According to the U.S. Department of Agriculture's World Livestock Situation, the country with the greatest per capita consumption of pork is Denmark. In 1994, the amount of pork consumed by a person residing in Denmark had a mean value of 147 pounds with a standard deviation of 62 pounds. If a random sample of 25 Danes is chosen, approximate the probability that the average amount of pork consumed by the members of this group in 1994 exceeded 150 pounds.

Hint: $P(\bar{X} > 150) \simeq 0.404$, where $\bar{X}$ is the sample mean of the 25 sample values.

–END OF CHAPTER 4. THANK YOU!–