

3. Sampling distribution of a count and proportion

a) Sampling distribution of a count

Suppose we have a population of size N , each individual takes only two values: success or failure, no. of success in population = $N1$

Now we draw an SRS of size n from the above population. Let X = no. of success in the SRS. We call the random variable X a **count** of the sample, and $\hat{p} = X/n$ the **sample proportion**.

To find the distribution of a count, we consider two cases:

Case 1 When n observations in the sample are independent and the probability of a success, p , is the same for each observation, then it satisfies the binomial settings. Then $X \sim B(n, p)$.

Case 2 When n observations are not independent or the probability of a success p is not the same for each observation, then X is not $B(n, p)$, but as a rule of thumb for a count for approximation: If $N > 10n$. $X \sim B(n, p)$ approximately.

b) Sample proportions

The sample proportion does not have a binomial distribution for any of the previous two cases. We translate any question about the proportion into a question about the count. And we have

$$\mu_{\hat{p}} = p \quad \sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

Lecture_11

1

Example: A factory employs several thousand workers, of whom 30% are Hispanic. If the 15 members of the union executive committee were chosen from the workers at random.

- What is the probability that exactly 20% members of the committee are Hispanic?
- What is the probability that 20% or fewer members of the committee are Hispanic?

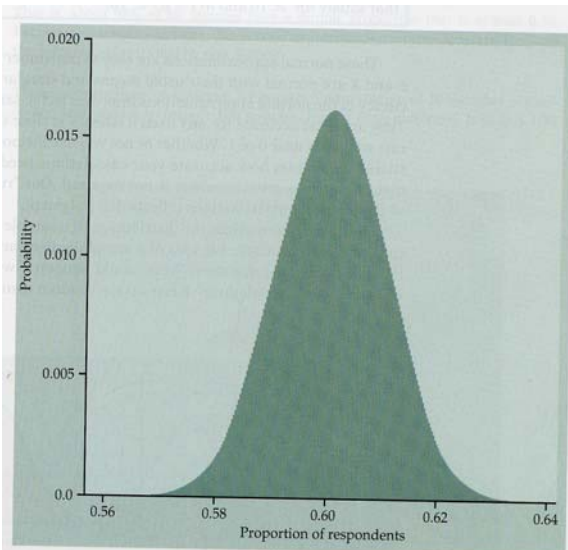


FIGURE 5.2 Probability histogram of the sample proportion \hat{p} based on a binomial count with $n = 2500$ and $p = 0.6$. The distribution is very close to normal.

Lecture_11

2

c) Normal approximation of count and proportion.

Draw an SRS of size n from a large population having population proportion p of successes. Let X be the count of successes in the sample and $\hat{p} = X/n$ the sample proportion of successes. When n is large, the sampling distributions of these statistics are approximately normal:

$$X \sim N(np, \sqrt{np(1-p)})$$
$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

As a rule of thumb, we will use this approximation for values of n and p that satisfy $np \geq 10$ and $n(1-p) \geq 10$

d) The continuity correction for normal approximation.

When we use normal approximation to do binomial distribution calculation, continuity correction enables us to get more precise result.

Example: A selective university would like to have an entering class of 1200 students. Because not all students who are offered admission accept, the college admits more than 1200 students. Past experience shows that about 70% of the students admitted will accept. The college decides to admit 1500 students.

- (a) What are the mean and the standard deviation of the number X of students who accept?
- (b) Find the probability that at least 1000 students accept and the probability that more than 1200 students will accept.

5.2 The sampling distribution of a sample mean

Suppose we select an SRS of size n from the population with a variable X:

$$X_1, X_2, \dots, X_n$$

Then the sample mean $\bar{x} = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = \frac{1}{n} \sum X_i$

is a statistic (random variable).

Assumption: the population is large relative to the sample

$\Rightarrow X_1, X_2, \dots, X_n$ are independent and have the same distribution as the one of X.

1. The mean and standard deviation of \bar{x}

Let μ =mean of population= μ_x

σ =standard deviation of the population= σ_x

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

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

2. The sample distribution of \bar{x}

a) Population distribution is normal

First note that:

- Any linear combination of independent normal random variables is also normally distributed.
- Normal distribution is uniquely determined by mean and standard deviation.

Then we have:

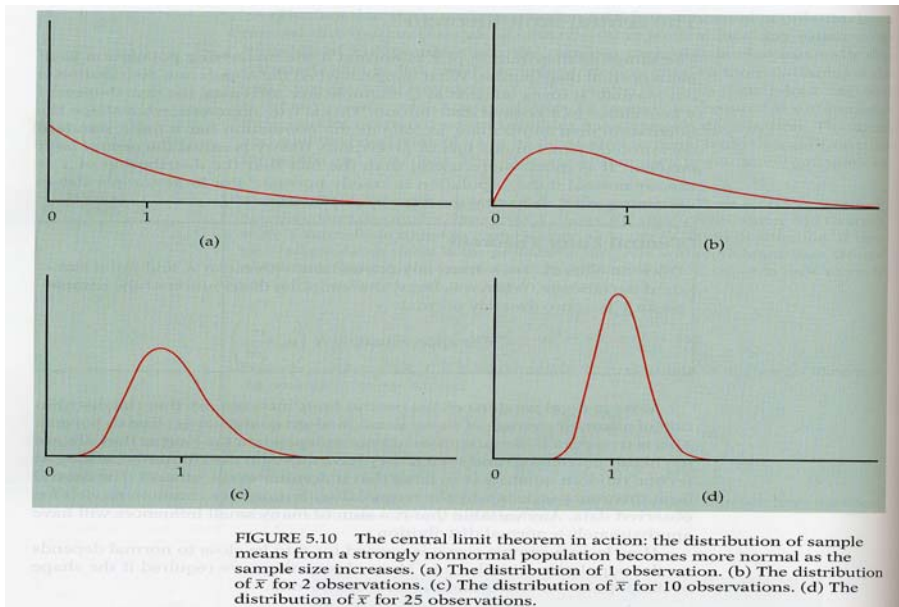
If a population is $N(\mu, \sigma)$, then the sample mean \bar{x} of n independent observations has the $N(\mu, \sigma/\sqrt{n})$ distribution.

b) Population distribution is not normal: The Central Limit Theorem (CLT)

When the size n of an SRS is large, then \bar{x} is approximately $N(\mu, \sigma/\sqrt{n})$ where μ and σ are population mean and standard deviation.

Lecture_11

5



Lecture_11

6

Example: An university posts the grade distributions for its courses online. Suppose that the distribution of grades in a course offered by this university in the spring 2001 semester was

Grade	A	B	C	D	E
Probability	0.18	0.32	0.34	0.09	0.07

- (a) Using the common scale $A=4, B=3, C=2, D=1, F=0$, take X to be the grade of a randomly chosen student. Find the mean μ and standard deviation σ of grades of this course.
- (b) Assume this course is very large. We can take the grades of an SRS of 50 students to be independent of each other. If \bar{X} is the average of these 50 grades, what are the mean and standard deviation of \bar{X} ?
- (c) What is the probability $P(X \geq 3)$ that a randomly chosen student gets a B or better? What is the approximate probability that the grade point average for 50 randomly chosen students is B or better (i.e. $P(\bar{x} \geq 3)$)?