Sample Proportions and Sampling Distributions

The objective of some statistical applications is to reach a conclusion about a population proportion, p. For example, we may try to estimate an approval rating through a survey, or test a claim about the proportion of defective light bulbs in a shipment based on a random sample. Since p is unknown to us, we must base our conclusion on a sample proportion, $\hat{p}$. However, as we have noted, we know that the value of $\hat{p}$ will vary from sample to sample. The amount of variability will depend on the size of our sample.

Sampling Distribution of a Sample Proportion:

Choose a SRS of size n from a large population with p having some characteristic of interest. Let $\hat{p}$ be the proportion of a sample having that characteristic.

- The mean of the sampling distribution of $\hat{p}$ is p ($\mu_{\hat{p}} = p$)
- The standard deviation is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Because the mean of the sampling distribution of $\hat{p}$ is always equal to the parameter p, the sample proportion $\hat{p}$ is an unbiased estimator of p. The standard deviation of $\hat{p}$ gets smaller as the sample size n increases because n appears in the denominator of the formula for the standard deviation. That is, $\hat{p}$ is less variable in larger samples. What is more, the formula shows just how quickly the standard deviation decreases as n increases. The sample size n is under the square root sign, so to cut the standard deviation in half, we must take a sample four times as large, not just twice as large.

The formula for the standard deviation of $\hat{p}$ doesn't apply when the sample is a large part of the population. You can't use this recipe if you choose an SRS of 50 of the 100 people in a class, for example. In practice, we usually take a sample only when the population is large. Otherwise, we could examine the entire population. Here is a practical guide.

Rule of Thumb 1: (When to use standard deviation formula for $\hat{p}$)

$\sigma_{\hat{p}}$ can be used only if population is at least 10 times the sample size

$N \geq 10n$

Using the Normal Distribution as a approximation for $\hat{p}$

In certain cases we may use the normal distribution and its properties to find the probabilities for $\hat{p}$. The following rule of thumb tells us when this is allowed.

> Rule of Thumb 2:
>
> if $np \geq 10$ and $n(1-p) \geq 10$, then the distribution of $\hat{p}$ is approximately Normal.

Example 7: Based on Census data, we know 11% of US adults are Black. Therefore, $p = 0.11$. We would expect a sample to contain roughly 11% Black representation. Suppose a sample of 1500 adults contains 138 Black individuals. Should we suspect 'undercoverage' in the sampling method?

a. Find $\hat{p}$. $\frac{138}{1500} = .092$

b. Is this lower than what would be expected by chance? That is, we know it is possible that a sample could contain 9.2% Black representation...but is it likely that would happen due to natural variation in a random sampling method?

Check Assumptions: $N \geq 10n$

$N \geq 10(1500)$

$N \geq 15000$

yes more than 15000 US adults

Can we use the normal approximation?

$np \geq 10$ $1500(.11) = 165$ $n(1-p) \geq 10$ $1500(1-.11) = 1335$

Reasonable to approx assume $\hat{p}$ is Normal

Calculate Probability $P(\hat{p} \leq .092)$

$\mu_{\hat{p}} = .11$ $\sigma_{\hat{p}} = \sqrt{\frac{.11(.89)}{1500}} = .0081$

$P(\hat{p} \leq .092) = .0131$

Interpret:

Not very likely to choose a sample with $\hat{p} = .092$ or less

Suspect some bias.

When we solve problems like example 8, you are expected to write all steps out just like we have in class! ALL THE TIME!

Example 8: The development of viral hepatitis subsequent to a blood transfusion can cause serious complications for a patient. The article, "Hepatitis in Patients with Acute Nonlymphatic Leukemia" (Amer. J. of Med.(1983): 413-421) reported that in spite of careful screening for those having a hepatitis antigen, viral hepatitis occurs in 7% of blood recipients. Suppose a new treatment is believed to reduce the incidence of viral hepatitis. The treatment is given to 200 blood recipients and only 6 contract hepatitis. Does it appear that the treatment is effective? That is, is it very likely that we would observe only 6/200 contract hepatitis when 7% of the population is known to do so? Population: all people who have had blood tran.

Let $\hat{p}$ = proportion of sample with hepatitis

find $P(\hat{p} \leq .03) = .0131$

$np \geq 10$ $\quad 200(.07) = 14$

$n(1-p) \geq 10$ $\quad 200(.93) = 186$

So $\hat{p}$ is approx Normal

$\mu_{\hat{p}} = .07$

$\sigma_{\hat{p}} = \sqrt{\frac{.07(.93)}{200}} = .0180$

.03 .07

Since a very unlikely event occured, good evidence treatment is working.

is $N \geq 10n$?

$N \geq 10(200)$

$N \geq 2000$

Reasonable to say more than 2000 blood recipients

Example 9: The article "Should Pregnant Women Move? Linking Risks for Birth Defects with Proximity to Toxic Waste Sites" (Chance (1992): 40-45) reported that in a large study carried out in the state of New York, approximately 30% of the population lived within 1 mile of a hazardous waste site. If an SRS of 400 pregnant women is selected, how likely is it that the sample proportion will be within 5% of the true population proportion? Would this probability be larger or smaller if we selected an SRS of size 500? (You don't need to do a calculation to figure this out...use common sense!)

Population: Residents of New York

$P(.25 \leq \hat{p} \leq .35) = .9710$

$N \geq 10n$

$N \geq 400(6)$

$N \geq 4000$

Reasonable to say NY has more than 4000 residents

$np \geq 10$

$400(.3) = 120$

$n(1-p) \geq 10$

$400(.7) = 280$

So $\hat{p}$ is approx Normal

.25 .3 .35

$\mu_{\hat{p}} = .3$

$\sigma_{\hat{p}} = \sqrt{\frac{.3(.7)}{400}} = .0229$

If $n = 500$, then less variability so higher probability.

Example 10: The article "Thrillers" (Newsweek, Apr. 22, 1985) states, "Surveys tell us that more than half of America's college graduates are avid readers of mystery novels." Assume the true proportion is exactly 0.5. What is the probability that an SRS of 225 college graduates would give a sample proportion greater than 0.6?

Population: all college graduates

$P(\hat{p} \geq .6) = .0013$

$N \geq 10n$

$N \geq 10(225)$

$N \geq 2250$

Reasonable to say more than 2250 college Grads in America

$np \geq 10$

$225(.5) = 112.5$

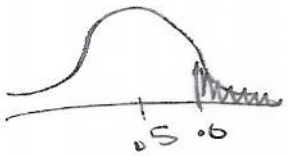
$n(1-p) \geq 10$

$225(.5) \geq 112.5$

So $\hat{p}$ is approx Normal

$\mu_{\hat{p}} = .5$

$\sigma_{\hat{p}} = \sqrt{\frac{.5(.5)}{225}} = .0333$



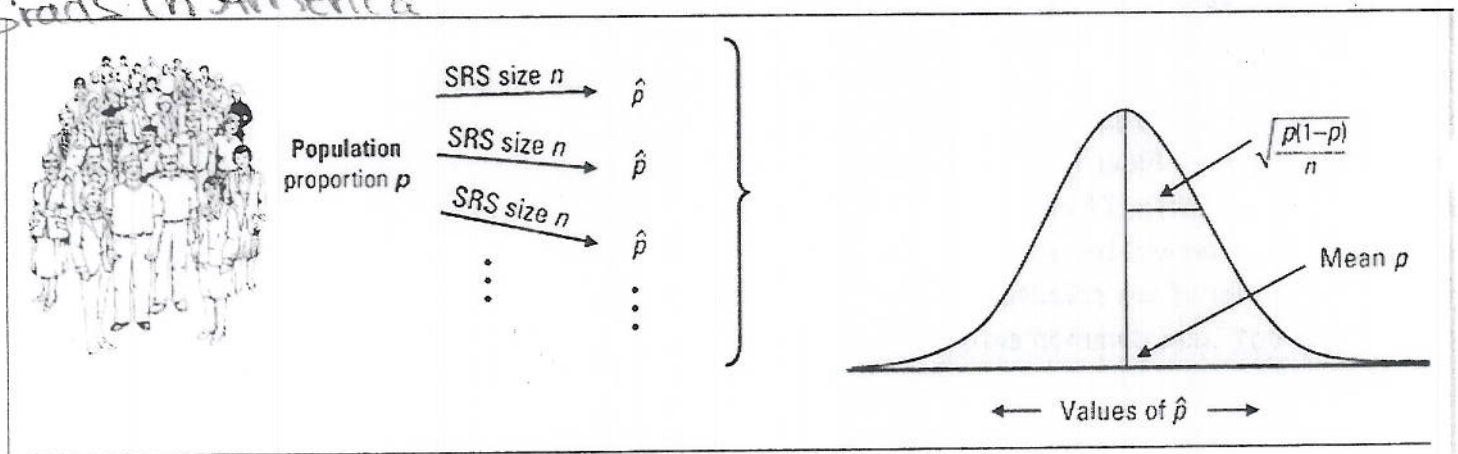


Figure 9.14 *Select a large SRS from a population in which the proportion* p *are successes. The sampling distribution of the proportion* !phat! *of successes in the sample is approximately Normal. The mean is* p *and the standard deviation is* $\sqrt{p(1-p)/n}$.

Sample Means and sampling distributions

When the objective of a statistical application is to reach a conclusion about a population mean, μ, we must consider a sample mean, $\bar{x}$. However, as we have noted, we know that the value of $\bar{x}$ will vary from sample to sample. The amount of variability will depend on the size of our sample.

Sampling Distribution of Sample Means:

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

Then $\mu_{\bar{x}} = \mu$ $\quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Special Notes: The behavior of $\bar{x}$ in repeated samples is much like that of the sample proportion $\hat{p}$:

- The sample mean $\bar{x}$ is an unbiased estimator of the population mean μ.
- The values of $\bar{x}$ are less spread out for larger samples. Their standard deviation decreases at the rate $\sqrt{n}$, so you must take a sample four times as large to cut the standard deviation of $\bar{x}$ in half.
- You should use the recipe $\sigma / \sqrt{n}$ for the standard deviation of $\bar{x}$ only when the population is at least 10 times as large as the sample. This is almost always the case in practice.

Notice that *these facts about the mean and standard deviation of $\bar{x}$ are true no matter what the population distribution looks like.*

Example 11: **Lightning strikes** The number of lightning strikes on a square kilometer of open ground in a year has mean 6 and standard deviation 2.4. (These values are typical of much of the United States.) The National Lightning Detection Network uses automatic sensors to watch for lightning in a sample of 10 square kilometers.

a. What are the mean and standard deviation of $\bar{x}$, the mean number of strikes per square kilometer?

$\mu_{\bar{x}} = 6$ $\quad \sigma_{\bar{x}} = \frac{2.4}{\sqrt{10}} = .7589$

b. Can you calculate the probability that $\bar{x} < 5$? If so, do it. If not, explain why not.

NO

We don't know the shape of the distribution

Sampling Distribution of a Sample Mean from a Normal Population:

Draw a SRS of size n from a Normal population with μ and σ. Then the distribution of $\bar{x}$ is Normal with

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

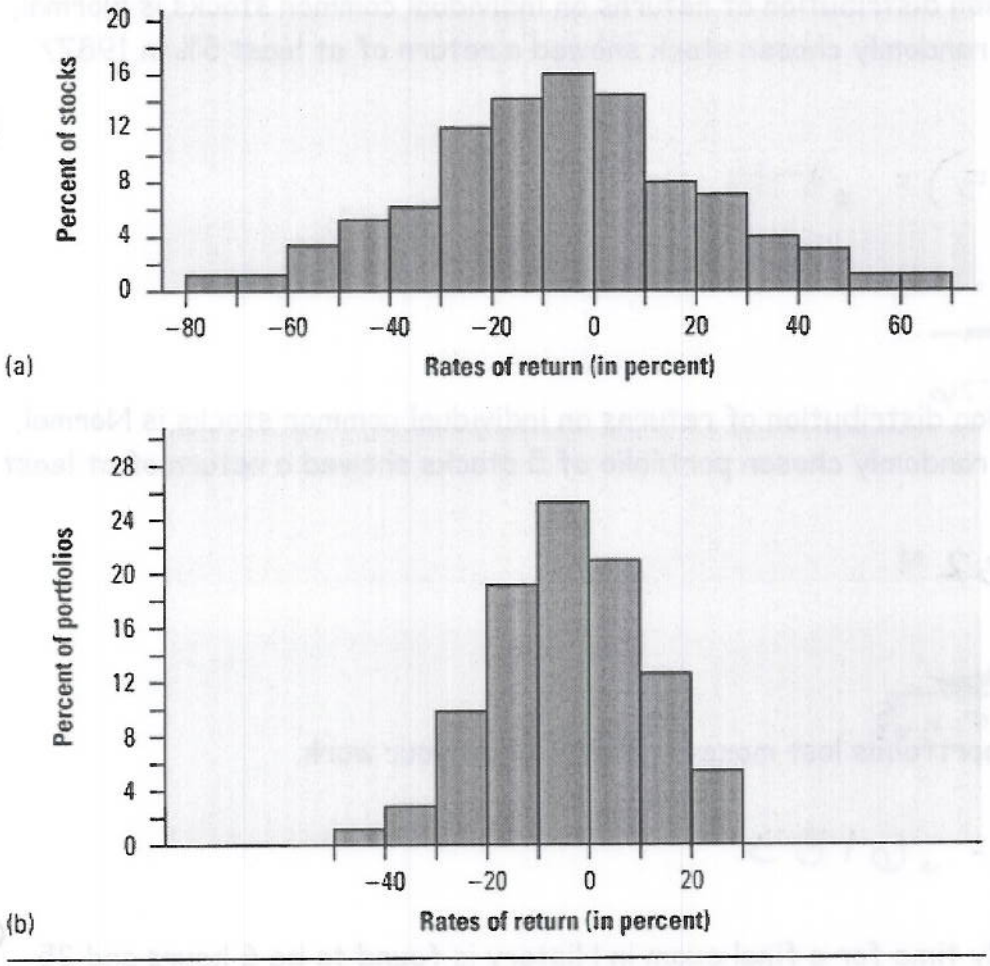


Figure 9.15 ***(a) The distribution of returns for New York Stock Exchange common stocks in 1987. (b) The distributions of returns for portfolios of five stocks in 1987.***

Example 12; **Bull market or bear market?** Investors remember 1987 as the year stocks lost 20% of their value in a single day. For 1987 as a whole, the mean return of all common stocks on the New York Stock Exchange was $\mu = -3.5\%$. (That is, these stocks lost an average of 3.5% of their value in 1987.) The standard deviation of the returns was about $\sigma = 26\%$. Figure 9.15(a) shows the distribution of returns. Figure 9.15(b) is the sampling distribution of the mean returns $\bar{x}$ for all possible samples of 5 stocks.

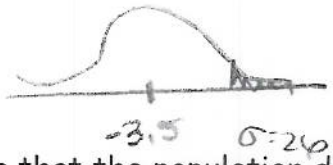
(a) What are the mean and the standard deviation of the distribution in Figure 9.15(b)?

$\mu_{\bar{x}} = -3.5$

$\sigma_{\bar{x}} = \frac{26}{\sqrt{5}} = 11.63$

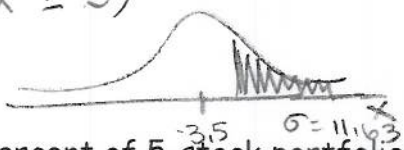
(b) Assuming that the population distribution of returns on individual common stocks is Normal, what is the probability that a randomly chosen stock showed a return of at least 5% in 1987? Show your work.

$P(X \geq 5) = .3719$



(c) Assuming that the population distribution of returns on individual common stocks is Normal, what is the probability that a randomly chosen portfolio of 5 stocks showed a return of at least 5% in 1987? Show your work.

$P(\bar{x} \geq 5) = .2324$



(d) What percent of 5-stock portfolios lost money in 1987? Show your work.

$P(\bar{x} \leq 0) = .6183$

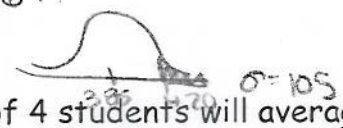
Example 14: The average study time for a final exam in History is found to be 6 hours and 25 minutes with a standard deviation of 1 hour and 45 minutes. Assume the distribution is normal.

$\mu = 385$ min $\quad \sigma = 105$ min

a. What is the probability that a student chosen at random spends more than 7 hours in studying?

36.94%

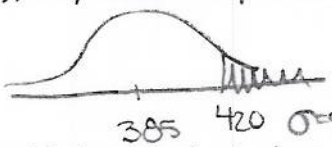
$P(X \geq 420) = .3694$



b. What is the probability that an SRS of 4 students will average more than 7 hours in studying? Compared to a), why does the probability go down?

$\mu_{\bar{x}} = 385$

$\sigma_{\bar{x}} = 105/\sqrt{4} = 52.5$



$P(\bar{x} \geq 420) = .2525$

Less variability due larger sample.

c. What is the probability that a student chosen at random spends less than 4 hours in studying?

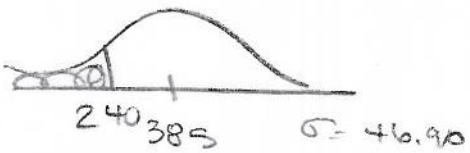
$P(X \leq 240) = .0835$

d. What is the probability that an SRS of 5 students will average less than 4 hours in studying?

$P(\bar{x} \leq 240) = .0001$

$\mu_{\bar{x}} = 385$

$\sigma_{\bar{x}} = 105/\sqrt{5} = 46.96$



Example 14: **Measurements in the lab** Juan makes a measurement in a chemistry laboratory and records the result in his lab report. The standard deviation of students' lab measurements is σ = 10 milligrams. Juan repeats the measurement 3 times and records the mean $\bar{x}$ of his 3 measurements.

(a) What is the standard deviation of Juan's mean result? (That is, if Juan kept on making 3 measurements and averaging them, what would be the standard deviation of all his $\bar{x}$'s?)

$\sigma_{\bar{x}} = \frac{10}{\sqrt{3}} = 5.7735$

(b) How many times must Juan repeat the measurement to reduce the standard deviation of $\bar{x}$ to 3 milligrams? Explain to someone who knows no statistics the advantage of reporting the average of several measurements rather than the result of a single measurement.

want $\sigma_{\bar{x}} = 3$

$3 = \frac{10}{\sqrt{n}}$

$3\sqrt{n} = 10$

$\sqrt{n} = 10/3$ $\quad n = 100/9 \approx 11.11$

12 measurements

Using an average allows for less variability in $\bar{x}$ so its likely to more accurately reflect true μ.

Central Limit Theorem

When dealing with sample means from sufficiently large samples, a useful fact about the sampling distribution arises...regardless of the shape of the population distribution.

Activity: A Penny for Your Thoughts

Your teacher has a jar containing several hundred pennies. Each penny has a 'birth date' indicated by the date on its face. Your task is to take repeated samples of pennies to determine the average birth date of the pennies in the jar. You will take 5 samples of varying sizes and record the average birth date on the dot plots on the board. Sketch the dot plots below and note any patterns you see emerging.