Sampling Distributions

TOPIC 12

- Sample Proportion
- Sample Mean
- Central Limit Theorem
- Difference Between Two Independent Sample Proportions
- Difference Between Two Independent Sample Means
- The t-distribution
- The Chi-Square Distribution
- The Standard Error

The *population* is the complete set of items of interest. A *sample* is a part of a population used to represent the population. The population mean μ and population standard deviation σ are examples of *population parameters.* The sample mean $\bar{x}$ and the sample standard deviation s are examples of *statistics.* Statistics are used to make inferences about population parameters. While a population parameter is a fixed quantity, statistics vary depending on the particular sample chosen. The probability distribution showing how a statistic varies is called a *sampling distribution.* The sampling distribution is *unbiased* if its mean is equal to the associated population parameter.

TIP

A sampling distribution is not the same thing as the distribution of a sample.

We want to know some important truth about a population, but in practical terms this truth is unknowable. What's the average adult human body weight? What proportion of people have high cholesterol? What we can do is carefully collect data from as large and representative a group of individuals as possible and then use this information to estimate the value of the population parameter. How close are we to the truth? We know that different samples would give different estimates, and so sampling error is unavoidable. What is wonderful, and what we will learn in this Topic, is that we can quantify this sampling error! We can make statements like "the average weight must almost surely be within 5 pounds of 176 pounds" or "the proportion of people with high cholesterol must almost surely be within ±3% of 37%."

SAMPLING DISTRIBUTION OF A SAMPLE PROPORTION

Whereas the mean is basically a quantitative measurement, the proportion represents essentially a qualitative approach. The interest is simply in the presence or absence of some attribute. We count the number of yes responses and form a proportion. For example, what proportion of drivers wear seat belts? What proportion of SCUD missiles can be intercepted? What proportion of new stereo sets have a certain defect?

This separation of the population into "haves" and "have-nots" suggests that we can make use of our earlier work on binomial distributions. We also keep in mind that, when n (trials, or in this case sample size) is large enough, the binomial can be approximated by the normal.

In this topic we are interested in estimating a population proportion p by considering a single sample proportion $\hat{p}$. This sample proportion is just one of a whole universe of sample proportions, and to judge its significance we must know how sample proportions vary. Consider the set of proportions from all possible samples of a specified size n. It seems reasonable that these proportions will cluster around the population proportion (the sample proportion is an unbiased statistic of the population proportion) and that the larger the chosen sample size, the tighter the clustering.

How do we calculate the mean and standard deviation of the set of population proportions? Suppose the sample size is n and the actual population proportion is p. From our work on binomial distributions, we remember that the mean and standard deviation for the number of successes in a given sample are pn and $\sqrt{np(1-p)}$, respectively, and for large n the complete distribution begins to look "normal."

Here, however, we are interested in the proportion rather than in the number of successes. From Topic Two we remember that when we multiply or divide every element by a constant, we multiply or divide both the mean and the standard deviation by the same constant. In this case, to change number of successes to proportion of successes, we divide by n:

$$\mu_{\hat{p}} = \frac{pn}{n} = p \quad \text{and} \quad \sigma_{\hat{p}} = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}$$

Furthermore, if each element in an approximately normal distribution is divided by the same constant, it is reasonable that the result will still be an approximately normal distribution.

Thus the principle forming the basis of the following discussion is

> Start with a population with a given proportion p. Take all samples of size n. Compute the proportion in each of these samples. Then
>
> 1. the set of all sample proportions is approximately normally distributed (often stated: the distribution of sample proportions is approximately normal).
> 2. the mean $\mu_{\hat{p}}$ of the set of sample proportions equals p, the population proportion.
> 3. the standard deviation $\sigma_{\hat{p}}$ of the set of sample proportions is approximately equal to $\sqrt{\frac{p(1-p)}{n}}$.
>
> Alternatively, we say that the sampling distribution of $\hat{p}$ is approximately normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.

Since we are using the normal approximation to the binomial, both np and $n(1-p)$ should be at least 10. Furthermore, in making calculations and drawing conclusions from a specific sample, it is important that the sample be a *simple random sample.*

Finally, because sampling is usually done without replacement, the sample cannot be too large; the sample size n should be no larger than 10% of the population. (We're actually worried about *independence*, but randomly selecting a relatively small sample allows us to assume independence. Of course, it's always better to have larger samples—it's just that if the sample is large relative to the population, then the proper inference techniques are different from those taught in introductory statistics classes.)

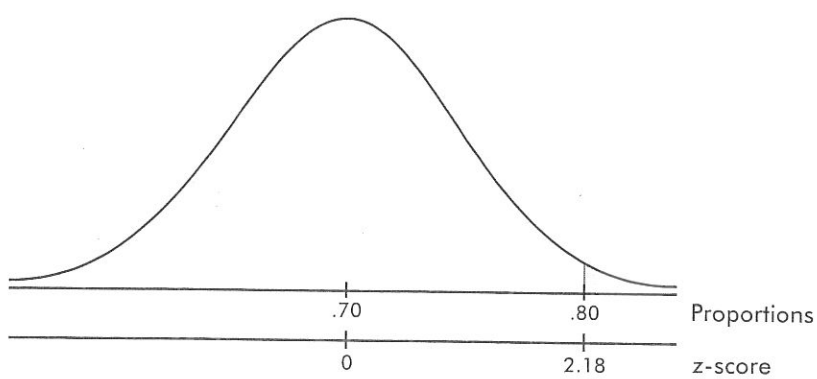
EXAMPLE 12.1

Suppose that 70% of all dialysis patients will survive for at least 5 years. In a simple random sample (SRS) of 100 new dialysis patients, what is the probability that the proportion surviving for at least 5 years will exceed 80%?

Answer: Both $np = (100)(.7) = 70 > 10$ and $n(1-p) = (100)(.3) = 30 > 10$, and our sample is clearly less than 10% of all dialysis patients. So the set of sample proportions is approximately normally distributed with mean .70 and standard deviation

$$\sigma_p = \sqrt{\frac{(.7)(.3)}{100}} = .0458$$

With a z-score of $\frac{.80-.70}{.0458} = 2.18$, the probability that the sample proportion will exceed 80% is $1 - .9854 = .0146$. [normalcdf(.8, 1, .7, .0458) = .0145.]



EXAMPLE 12.2

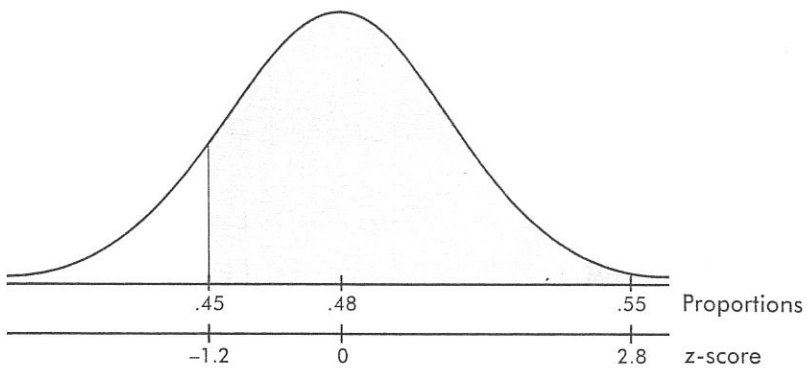
It is estimated that 48% of all motorists use their seat belts. If a police officer observes 400 cars go by in an hour, what is the probability that the proportion of drivers wearing seat belts is between 45% and 55%?

(continued)

Answer: Both $np = (400)(.48) = 192 > 10$ and $n(1 - p) = (400)(.52) = 208 > 10$, and our sample is clearly less than 10% of all motorists. So the set of sample proportions is approximately normally distributed with mean .48 and standard deviation

$$\sigma_p = \sqrt{\frac{(.48)(.52)}{400}} = .0250$$

The z-scores of .45 and .55 are $\frac{.45-.48}{.0250} = -1.2$ and $\frac{.55-.48}{.0250} = 2.8$, respectively. From Table A, the area between .45 and .55 is .9974 – .1151 = .8823. Thus there is a .8823 probability that between 45% and 55% of the drivers are wearing seat belts. [normalcdf(.45, .55, .48, .0250) = .8824.]



SAMPLING DISTRIBUTION OF A SAMPLE MEAN

Suppose we are interested in estimating the mean μ of a population. For our estimate we could simply randomly pick a single element of the population, but then we would have little confidence in our answer. Suppose instead that we pick 100 elements and calculate their average. It is intuitively clear that the resulting sample mean has a greater chance of being closer to the mean of the whole population than the value for any individual member of the population does.

When we pick a sample and measure its mean $\bar{x}$, we are finding exactly one sample mean out of a whole universe of sample means. To judge the significance of a single sample mean, we must know how sample means vary. Consider the set of means from all possible samples of a specified size. It is both apparent and reasonable that the sample means are clustered around the mean of the whole population; furthermore, these sample means have a tighter clustering than the elements of the original population. In fact, we might guess that the larger the chosen sample size, the tighter the clustering.

How do we calculate the standard deviation $\sigma_{\bar{x}}$ of the set of sample means? Suppose the variance of the population is σ^2 and we are interested in samples of size n. Sample means are obtained by first summing together n elements and then dividing by n. A set of sums has a variance equal to the sum of the variances associated with the original sets. In our case, $\sigma^2_{\text{sums}} = \sigma^2 + \cdots + \sigma^2 = n\sigma^2$. When each element of a set is divided by some constant, the new variance is the old one divided by the square of the constant. Since the sample means are obtained by dividing the sums by n, the variance of the sample means is obtained by dividing the variance of the

sums by n^2. Thus if $\sigma_{\bar{x}}$ symbolizes the standard deviation of the sample means, we find that

$$\sigma_{\bar{x}}^2 = \frac{\sigma_{\text{sums}}^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

In terms of standard deviations, we have $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

We have shown the following:

> Start with a population with a given mean μ and standard deviation σ. Compute the mean of all samples of size n. Then the mean of the set of sample means will equal μ, the mean of the population, and the standard deviation $\sigma_{\bar{x}}$ of the set of sample means will be approximately equal to $\frac{\sigma}{\sqrt{n}}$, that is, the standard deviation of the whole population divided by the square root of the sample size.

Note that the variance of the set of sample means varies directly as the variance of the original population and inversely as the size of the samples, while the standard deviation of the set of sample means varies directly as the standard deviation of the original population and inversely as the square root of the size of the samples.

EXAMPLE 12.3

Suppose that tomatoes weigh an average of 10 ounces with a standard deviation of 3 ounces and a store sells boxes containing 12 tomatoes each. If customers determine the average weight of the tomatoes in each box they buy, what will be the mean and standard deviation of these averages?

Answer: We have samples of size 12. The mean of these sample means will equal the population mean, 10 ounces. The standard deviation of these sample means will equal $\frac{3}{\sqrt{12}} = 0.866$ ounces.

Note that while giving the mean and standard deviation of the set of sample means, we did not describe the shape of the distribution. If we are also given that the original population is normal, then we can conclude that the set of sample means has a normal distribution.

EXAMPLE 12.4

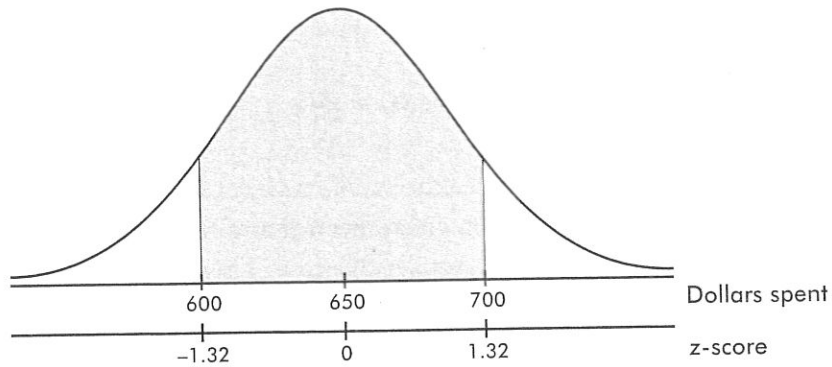
Suppose that the distribution for total amounts spent by students vacationing for a week in Florida is normally distributed with a mean of \$650 and a standard deviation of \$120. What is the probability that an SRS of 10 students will spend an average of between \$600 and \$700?

Answer: The mean and standard deviation of the set of all sample means of size 10 are:

$$\mu_{\bar{x}} = 650 \text{ and } \sigma_{\bar{x}} = \frac{120}{\sqrt{10}} = 37.95$$

(continued)

The z-scores of 600 and 700 are $\frac{600-650}{37.95} = -1.32$ and $\frac{700-650}{37.95} = 1.32$, respectively. Using Table A, we find the desired probability is .9066 – .0934 = .8132. [normalcdf(600, 700, 650, 37.95) = .8123.]



CENTRAL LIMIT THEOREM

We assumed above that the original population had a normal distribution. Unfortunately, few populations are normal, let alone exactly normal. However, it can be shown mathematically that no matter how the original population is distributed, if n is large enough, then the set of sample means is approximately normally distributed. For example, there is no reason to suppose that the amounts of money that different people spend in grocery stores are normally distributed. However, if each day we survey 30 people leaving a store and determine the average grocery bill, these daily averages will have a nearly normal distribution.

The following principle forms the basis of much of what we discuss in this topic and in those following. It is a simplified statement of the *central limit theorem* of statistics.

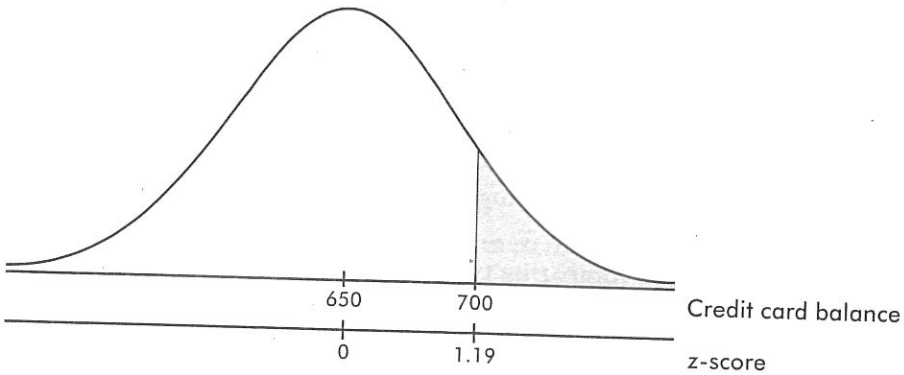
> Start with a population with a given mean μ, a standard deviation σ, and any shape distribution whatsoever. Pick n sufficiently large (at least 30) and take all samples of size n. Compute the mean of each of these samples. Then
>
> 1. the set of all sample means is approximately normally distributed (often stated: the distribution of sample means is approximately normal).
> 2. the mean of the set of sample means equals μ, the mean of the population.
> 3. the standard deviation $\sigma_{\bar{x}}$ of the set of sample means is approximately equal to $\frac{\sigma}{\sqrt{n}}$, that is, equal to the standard deviation of the whole population divided by the square root of the sample size.
>
> Alternatively, we say that the sampling distribution of $\bar{x}$ is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

While we mention $n \geq 30$ as a rough rule of thumb, $n \geq 40$ is often used, and n should be chosen even larger if more accuracy is required or if the original population is far from normal. As with proportions, we have the assumptions of a simple random sample and of sample size n no larger than 10% of the population.

EXAMPLE 12.5

Suppose that the average outstanding credit card balance for young couples is \$650 with a standard deviation of \$420. In an SRS of 100 couples, what is the probability that the mean outstanding credit card balance exceeds \$700?

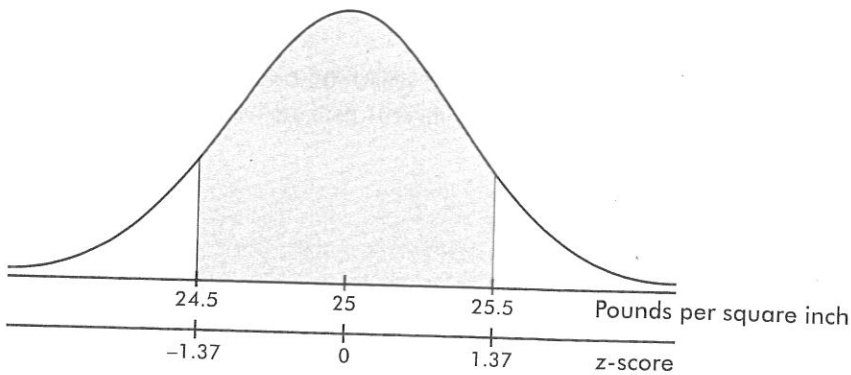
Answer: The sample size is over 30, we have an SRS, and our sample is less than 10% of all couples with outstanding balances, and so by the central limit theorem the set of sample means is approximately normally distributed with mean 650 and standard deviation $\frac{420}{\sqrt{100}} = 42$. With a z-score of $\frac{700-650}{42} = 1.19$, the probability that the sample mean exceeds 700 is 1 –.8830 = .1170. [normalcdf(700, 10000, 650, 42) = .1169.]



EXAMPLE 12.6

The strength of paper coming from a manufacturing plant is known to be 25 pounds per square inch with a standard deviation of 2.3. In a simple random sample of 40 pieces of paper, what is the probability that the mean strength is between 24.5 and 25.5 pounds per square inch?

Answer: We have a large ($n = 40$) SRS that is still smaller than 10% of all papers coming from the plant. $\mu_{\bar{x}} = 25$ and $\sigma_{\bar{x}} = \frac{2.3}{\sqrt{40}} = 0.364$. The z-scores of 24.5 and 25.5 are $\frac{24.5-25}{0.364} = -1.37$ and $\frac{25.5-25}{0.364} = 1.37$, respectively. The probability that the mean strength in the sample is between 24.5 and 25.5 pounds per square inch is .9147 – .0853 = .8294. [normalcdf(24.5, 25.5, 25, .364) = .8304.]



Don't be confused by the several different distributions being discussed! First, there's the distribution of the original population, which may be uniform, bell-shaped, strongly skewed – anything at all. Second, there's the distribution of the data in the sample, and the larger the sample size, the more this will look like the population distribution. Third, there's the distribution of the means of many samples of a given size, and the amazing fact is that this *sampling distribution* can be described by a normal model, regardless of the shape of the original population.

SAMPLING DISTRIBUTION OF A DIFFERENCE BETWEEN TWO INDEPENDENT SAMPLE PROPORTIONS

Numerous important and interesting applications of statistics involve the comparison of two population proportions. For example, is the proportion of satisfied purchasers of American automobiles greater than that of buyers of Japanese cars? How does the percentage of surgeons recommending a new cancer treatment compare with the corresponding percentage of oncologists? What can be said about the difference between the proportion of single parents on welfare and the proportion of two-parent families on welfare?

Our procedure involves comparing two sample proportions. When is a difference between two such sample proportions significant? Note that we are dealing with one difference from the set of all possible differences obtained by subtracting sample proportions of one population from sample proportions of a second population. To judge the significance of one particular difference, we must first determine how the differences vary among themselves. Remember that the variance of a set of differences is equal to the sum of the variances of the individual sets; that is,

$$\sigma_d^2 = \sigma_1^2 + \sigma_2^2$$

Now if

$$\sigma_1 = \sqrt{\frac{p_1(1-p_1)}{n_1}} \quad \text{and} \quad \sigma_2 = \sqrt{\frac{p_2(1-p_2)}{n_2}}$$

then

$$\sigma_d^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \quad \text{and} \quad \sigma_d = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$