### Sample Proportions

Start with a population of units and a variable that is categorical with two categories on the basis of some attribute that each unit either succeeds in having, or fails to have. These categories are generically referred to as “Success” and “Failure” (however, the use of these words does not place a value judgment on either category). A proportion \( p \) of the population have the designated attribute and are referred to as Successes. If \( p \) is close to 1, then nearly all units have the attribute; if close 0 nearly none; close to 0.5 is nearly an even split. \( p \) is a parameter.

Some reference sources (including textbooks) use the symbol \( q \) as a shortcut for the Failure proportion: \( q = (1 - p) \). In this document we’ll stick to writing out \((1 - p)\).

**Example 1**

Consider the population of all American adults. It is reported that 40% of all Americans believe in evolution. Let’s take it for granted this is true. The units are American adults; the variable is “whether or not one believes in evolution.” Then \( p = 0.40 \) or 40%, and we are (without assigning a value judgment) specifying that “belief in evolution” constitutes a Success, and lack of belief a Failure. If \( p = 0.40 \) then \( q = 0.60 \).

It could certainly be done the other way around, and then \( p = 0.60 \). (So we always attach \( p \) to what we define as a Success.) If this is the case, then \( q = 0.40 \). Choosing belief in evolution to go with Success is natural, because the initial statement reported 40% as believers in evolution.

Next consider a simple random sample (SRS) of \( n \) individuals from the population. \( \hat{p} \) is the proportion of Successes among units in the sample. To obtain this value we would first count \( X \) the number of sampled units who are Successes, then divide by \( n \): \( \hat{p} = X/n \). (The caret \(^\wedge\) indicates “sample proportion” as well as “estimate of.” The symbol \( \hat{p} \) is often spoken “p-Hat” but it is better called “the sample proportion estimating the population proportion \( p \).”)

\( \hat{p} \) is a statistic: it varies depending on which sample has been chosen. This sampling variability is characterized by a distribution – the distribution of the sample proportion.

**Example 1**

A pollster will obtain a simple random sample of 800 American adults. The results are summarized with \( \hat{p} \), the proportion of the 800 selected adults that believe in evolution. This sample proportion varies from sample to sample. In one sample it is found that 303 of the adults believe in evolution. \( 303/800 = 0.3788 \) or 37.88%. For this sample the proportion of sampled adults that do not believe in evolution is \( \hat{q} = (1 - \hat{p}) = 1 - 0.3788 = 0.6212 = 62.12\% \).
The Big Picture

What we are concerned with is a new unit/variable pairing that is derived from this situation. The units of concern for us are samples of size \( n \), while the variable is the proportion of Successes in the sample. So even though the original variable is categorical, the variable we care about – the sample proportion \( \hat{p} \) – is quantitative.

Using either extensive simulation or the mathematical theory of statistics, we come about the following rules for the distribution of sample proportions.

Distribution of Sample Proportions

All statements below pertain to samples that are randomly selected. We insist on random sampling – or sampling that is equivalent to random – in order to make proper probability statements about samples. You should check this condition before concluding that your formal statistical analysis is accurate.

1. The mean of the distribution \( \hat{p} \) of is \( p \). In mathematical shorthand this is expressed as follows: \( \mu_p = p \). That’s a pretty simple formula. What matters is that you understand what it expresses…

   If we are look at all possible samples of size \( n \), and compute a sample proportion for each, the mean of all these sample proportions is \( p \), the population proportion. (The technical term for this almost trivial result is unbiased. When a statistic used to estimate a parameter is, on average, equal to the parameter, we say the estimating procedure is unbiased. This is a good thing.)

For the most part our analysis of variation among sample proportions applies to situations when the Binomial distribution applies.† Technically this requires sampling with replacement. However, if sampling is done from a finite population, then as long as the population is at least 20 times the sample size‡, the results below apply. Consequently you should check whether this condition is met before making claims about the standard deviation of sample proportions.

2. If sampling is with replacement, or the population size is at least 20 times the sample size‡ (or both) then the standard deviation of the distribution of \( \hat{p} \) is

\[
\sigma_{\hat{p}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}
\]

---

* If you don’t know about that, don’t worry. It’s not required. We’ll avoid using Binomial distributions by setting criteria for which a Normal distribution can be used.
† This condition is met in almost all statistical analyses.
‡ If not, there is another not-too-complex formula that suffices. To avoid confusing matters, we’ll only consider the situations that most commonly occur in practice.
Two expressions are given. The first shows how similar the result is to results for standard deviations of sample means: $\sqrt{p(1-p)}$ replaces $\sigma$. The second tends to be used as it’s a little more computationally efficient.

These formulas for mean and standard deviation are a good thing, because when the distribution of sample proportions is approximately Normal (it often is), then an accompanying mean and standard deviation provide a complete characterization of the variability in sample proportions.

Proportions are special cases of means. Consequently, the Central Limit Theorem applies, and the distribution of all possible sample proportions will – under the right conditions – be an approximately Normal distribution.

- The closer $p$ is to 0.5, the closer the distribution is to Normal.
- The larger $n$ is, the closer the distribution is to Normal.

If $p$ is the Success proportion, then $np$ is simply the mean, or expected, count of Successes in the sample; $n(1-p)$ is the mean, or expected, count of Failures in the sample.

3. A fairly simple rule distinguishes situations for which the Normal distribution applies. It can be stated in two ways:

- If $n$ and $p$ are such that both $np$ and $n(1-p)$ are at least $10^\S$, then the distribution of sample proportions is adequately modeled by the Normal distribution. **

This is equivalent to the following.

- If $n$ and $p$ is such that both the expected number of Successes and Failures in the sample are at least 10, then the distribution of sample proportions is adequately modeled by the Normal distribution.

---

\S This condition is also commonly met. It says: You can’t use a Normal if you’re expecting only a handful or two or units falling into either of the categories. If the condition is not met, then the Binomial distribution must be used directly. We will not worry about this case here – except to say that while the methods are more involved, final results have the same meaning.

** Some references (textbooks) drop the criteria from 10 to 5. This is a little too low. There are adjustments and tweaks to the method that make 5 acceptable, but we will not pursue those here, in the interest of simplicity.
Example 1

Consider the population of all American adults. It is reported that 40% of all Americans believe in evolution. Let’s take it for granted this is true.

A pollster will obtain a simple random sample of 800 American adults.

a) Describe (in words) the applicable parameter \( p \).

\( p \) is the proportion of all American adults who believe in evolution.

b) Give a value for the parameter.

\( p = 0.40 \).

c) Describe (in words) the statistic \( \hat{p} \) that estimates this parameter.

\( \hat{p} \) is the proportion of American adults in a sample (of size 800) who believe in evolution.

d) Identify a value for the mean of the distribution of \( \hat{p} \).

\( \mu_{\hat{p}} = p = 0.40 \)

e) Is the population size at least 20 times larger than the sample size \( n \)?

The population here is immense – over 230 million (230,000,000). You don’t need to know the exact number. 20 times larger than the sample is \( 20 \times 800 = 16,000 \), and the population we’re talking about here is obviously far larger than this.

f) Identify the standard deviation of the distribution of \( \hat{p} \).

\[
\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.40 \times 0.60}{800}} = \sqrt{0.0003} = 0.01732
\]

It’s a good idea to not round any intermediate computations (let the calculator do the entire operation at once) and report at least 3 and preferably 4 significant figures with this result.

g) Determine whether both \( np \) and \( n(1-p) \) are at least 10.

\( np = 800(0.4) = 320 \). This is the expected number of believers (in evolution) in the sample; the expected number of non-believers is then \( 800 - 320 = 480 \): \( n(1-p) = 800(0.6) = 480 \). Since both of these are at least 10, a Normal distribution is appropriate:
h) Describe the distribution of $\hat{p}$. Indicate whether or not it is adequately modeled by a Normal distribution, and if so, state the mean and standard deviation.

The distribution of sample proportions is approximately Normal, with mean 0.40 and standard deviation 0.01732.

Let’s solve and interpret a probability problem: What is the probability that fewer than 37% of the sampled adults believe in evolution? Interpret the result in terms of the units.

We can approximate this with a Normal probability calculation. The mean of the distribution is 0.4, with standard deviation of 0.01732. The Z-score for a proportion of 0.37 is:

$$Z = \frac{0.37 - 0.40}{0.01732} = \frac{-0.03}{0.01732} = -1.732$$

An outcome of 37% for the sample is 1.732 standard deviations below what happens on average over all samples.

The probability is 0.0416. This is an approximate probability. Approximately 4.16% of all samples of size 800 [units] will have less than 37% of adults who believe in evolution. [The variable here is the sample percentage.]

**A good idea!** With any small probability, obtain its reciprocal.

For example: $1/0.0416 = 24.04$. So we can say – in a way that probably is more meaningful to the average ear – the following: On average, 1 in 24 samples of size 800 have less than 37% who favor the President.
Exercises Part A

For each of 1 – 7 below do the following:

a) Describe (in words) the applicable parameter \( p \). “\( p \) is …”

b) Give a value for the parameter: \( p = \underline{\text{________}} \)

c) Describe (in words) the statistic \( \hat{p} \) that estimates this parameter. “\( \hat{p} \) is …”

d) Identify the value for the mean of the distribution of \( \hat{p} \).

e) Is the population size at least 20 times larger than the sample size \( n \)? If not, say so and go no further – omit parts f through h.

f) Obtain the standard deviation for the distribution of \( \hat{p} \) ?

g) Determine whether both \( np \) and \( n(1 - p) \) are at least 10. Use this to judge whether the distribution of \( \hat{p} \), the sample proportion, is adequately modeled by a Normal distribution. If not, say so and go no further – omit part h.

h) State the distribution of the sample proportion \( \hat{p} \).

1. Assume that 52% (= 0.52) of all New York State voters have a favorable view of the President. We randomly sample 1000 voters.

2. A company produces thousands of electrical components. 0.96 of all components work properly, the remainder are defective. A random sample of 100 components is selected.

3. It is reported that 5% of all Americans are homosexual. A random sample of 3000 Americans is taken.

4. It is reported that 5% of all Americans are homosexual. A random sample of 30 Americans is taken.

5. 15% of all SUNY Oswego students call Oswego County their home county. 1500 students are randomly selected. (Assume there are 9,000 students at SUNY Oswego. This number is fairly accurate.)

6. A fair coin is tossed 200 times.

7. 25% of the 10,000+ deer in Sherwood Forest carry a virus. 125 deer are randomly selected and tested for the virus.

Check your answers before proceeding.
Exercises Part B

Now compute some probabilities. You will use the Normal distribution. (Technically all these probabilities are approximate probabilities.) When you have your probability, interpret it in terms of units and variable. (The units will be samples of size \( n \); the variable will be the sample proportion.) Problem numbers for 1 – 7 match those from Part A.

1. What is the probability that fewer than 50% of the sampled voters have a favorable view of the President? That is, find the probability that \( \hat{p} < .50 \). This is written \( P(\hat{p} < .50) \).

2. We cannot use the Normal distribution here because \( n(1 - p) \) is only 4. So #2 is omitted.

3. What’s the probability that over 6% of the sampled people are homosexual? In light of this probability, what would you conclude if 6% or more of the sampled people are homosexual?

4. We cannot use the normal distribution here because \( np \) is only 1.5. However, compare the standard deviation to that of #3. If the sample size is increased 100-fold (from 30 to 3000 here) by how much is the standard deviation increased/decreased?

5. We cannot do any more here because the population is not at least 20 times the sample size. The standard deviation formula we have is not appropriate. So #5 is omitted.

6. What’s the probability of obtaining between 90 and 110 heads? Note: This is equivalent to obtaining between 45% and 55% heads, so you need to compute \( P(0.45 < \hat{p} < 0.55) \).

7. What’s the probability that no more than 14% of the sampled deer carry the virus, \( P(\hat{p} < 0.14) \)? If the study were actually completed and we found that 14% of the sampled deer carried the virus, what would you conclude?

8. 25% of the millions of deer roaming Siberia carry a virus. 125 deer are randomly selected and tested for the virus. What’s the probability that no more than 14% of the sampled deer carry the virus?

9. 25% of the millions of deer roaming Siberia carry a virus. 60 deer are randomly selected and tested for the virus. What’s the probability that no more than 14% of the sampled deer carry the virus?

10. Compare your results to 7 and 8 above. Does the size of the population matter?

11. Compare your results to 8 and 9 above. Does the size of the sample matter?
Solutions Part A

1. a) $p$ is the proportion of all NYS voters who have a favorable view of the President.
   b) $p = 0.52$
   c) $\hat{p}$ is the proportion of the 1000 sampled voters who intend to vote for Lazio.
   d) 0.52
   e) The population is much larger than 20 times the sample size.
   f) The standard deviation is 0.01580.
   g) $np = 520$ and $n(1 - p) = 480$ – both a well above 10. So the distribution of sample proportions is approximately Normal.
   h) Approximately Normal with mean 0.52 and standard deviation 0.01580.

2. a) $p$ is the proportion of all components that work properly.
   b) $p = 0.96$
   c) $\hat{p}$ is the proportion of sampled components that work properly.
   d) 0.96
   e) The population size is certainly at least 20 times the sample size: $20 \times 100 = 2000$ and the population size is given as “thousands” which is at least 2,000.
   f) 0.01960.
   g) $n(1 - p) = 4$. The distribution of sample proportions is not approximately Normal. Stop here.

3. a) $p$ is the proportion of all Americans that are homosexual.
   b) $p = 0.05$
   c) $\hat{p}$ is the proportion of the 3000 sampled people who are homosexual.
   d) 0.05.
   e) Again, the population is way more than 20 times the sample size.
   f) 0.00398.
   g) $np = 150$ and $n(1 - p) = 2850$ – both a well above 10. So the distribution of sample proportions is approximately Normal.
   h) Normal with mean 0.05 and standard deviation 0.00398.

4. a) $p$ is the proportion of all Americans that are homosexual.
   b) $p = 0.05$
   c) $\hat{p}$ is the proportion of the 30 sampled people who are homosexual.
d) 0.05 

e) (Similar to #3.)

f) The standard deviation is 0.03979. Notice that the standard deviation here is 10 times larger than the value from #3. The situations are similar except that in #4 the sample size is $100 = 10^2$ times smaller than for #3.

g) $np = 30(0.05) = 1.5$. This is below 10: The distribution of $\hat{p}$ will not be approximately Normal. (However, the mean of 0.05 and standard deviation 0.00398 do still apply.) Stop here.

5. a) $p$ is the proportion of all SUNY Oswego students who come from Oswego County. 

b) $p = 0.15$

c) $\hat{p}$ is the proportion of the 1500 sampled students who come from Oswego County.

d) 0.15

e) The population size is not 20 times the sample size. Stop here.

6. a) $p$ is the proportion of tosses that result in a head. 

b) $p = 0.50$

c) $\hat{p}$ is the proportion of the 200 tosses that are heads.

d) 0.50

e) The population here is essentially infinite: One could toss a coin forever. So any sample size meets the 20 Times Criterion.

f) The standard deviation is 0.03536.

g) $np = 100$ and $n(1 - p) = 100$. Both are at least 10. So the distribution of $\hat{p}$ is approximately Normal.

h) Normal with mean 0.5 and standard deviation 0.03536.

7. a) $p$ is the proportion of all deer that carry the virus.

b) $p = 0.25$

c) $\hat{p}$ is the proportion of the 125 sampled deer that carry the virus.

d) 0.25

e) The population is way larger than $20 \times 125 = 2500$.

f) The standard deviation is 0.03873.

g) $np = 31.25$ and $n(1 - p) = 93.75$. Both are at least 10. So the distribution of $\hat{p}$ is approximately Normal.

h) Normal with mean 0.25 and standard deviation 0.03873.
Solutions Part B

1. The distribution of sample proportions is approximately Normal with mean 0.52 and standard deviation 0.01580. The Z-score for 0.50 is then

\[
Z = \frac{0.50 - 0.52}{0.0158} = \frac{0.02}{0.0158} = -1.266
\]

This gives a probability of 0.1028 (10.28%). 10.28% of all samples of 1000 New York voters will have less than 50% of the sample favoring the President.

2. Nothing to answer.

3. The distribution of sample proportions is approximately Normal with mean 0.05 and standard deviation 0.00398. The Z-score for 0.06 is (0.06 – 0.05) / 0.00398 = 2.5126. The probability is 0.0060 or 0.6%. This is about 1/1667. It’s much less than 1%. Consider all possible samples of 3000 Americans. In only 0.0006 = 0.06% = 1 in 1667 of all of these samples are there more than 6% homosexuals. Since this is rather unlikely to occur when the true figure is 5%, we’d be suspicious of the true figure if we actually observed a result of 6% for a sample of 3000 people. We have evidence then that the true figure is larger than 5%.

4. If the sample size is increased by 100 times, the standard deviation is decreased 10 times.

5. Nothing to answer.

6. 0.8427. Look at all possible samples/sequences of 200 coin tosses: Approximately 84.27% of the samples have between 45% and 55% heads occurring. (About 84.27% of the time you toss a coin 200 times, you’ll get between 45% and 55% heads. This percent refers to all conceivable times anyone might ever toss a fair coin 200 times.)

7. 0.002254 = 0.2254%. That’s about 1/444. So: For a mere 1 out of 444 of all possible samples of 125 deer will we find less than 14% of the sample carrying the virus.

If this study was actually conducted and we learned that 14% of the sampled deer carried the virus we would be very suspicious of the claimed figure of 25% – a result of 14% strongly suggests that the virus rate has decreased. (We might also be suspicious that the sample is not truly random. Random selection of things like deer is very difficult to accomplish in the field. Each deer should be equally likely to be included in the sample when selection is implemented. Given that some deer are a lot harder to find and capture than others, this could be very hard to accomplish.)

8. 0.002254 (the same as in #7).

9. 0.02455 – about 1/41 (different from #7).

10. The size of the population does not matter.
This is generally true except when the 20 Times / 5% Rule comes into play. When this happens (small population) the standard deviation formula we have cannot be used. A variation of if exists, which does depend upon the population size.

A general observation then is that as long as a population is “large” (where 20 times the sample size is the formal criterion), it doesn’t really matter how large! This also implies that the accuracy obtained in sampling does not depend on the population size, nor on the percent of the population that is sampled (again – unless that percent is over 5%). All that really matters is the sample size.

11. *The size of the sample does matter.* Smaller samples are less likely to give a “representative” result.