Sampling Distribution Models

Sampling Distribution of a Proportion

On October 4, 2004, less than a month before the presidential election, a Gallup Poll asked 1016 national adults, aged 18 or older, whom they supported for President; 49% said they'd chosen John Kerry. A Rasmussen Poll taken just a few weeks later found 45.9% of 1000 likely voters supporting Kerry. Was one poll "wrong?"

We want to imagine the results from all the random samples of size 1000 that we did not take. What would the histogram of all the samples look like? Where do you expect the center of the histogram to be? We can simulate. We want to simulate a bunch of those random samples of 1000 that we did not really draw. We choose the same probability (*p* = 0.49). The following histogram of the proportions a saying they would vote for Kerry for 2000 independent samples of 1000 voters when the true proportion is *p* = 0.49. What type of shape is the graph? The center is at *p* and what is the standard deviation? We saw with binomial distribution the standard deviation is $\sqrt{npq}$. Now we want the standard deviation of the proportion of successes, $\hat{p}$. The sample proportion $\hat{p}$ is the number of successes divided by the number of trials, n, so the standard deviation is

. $\sigma(\hat{p}) = \dfrac{\sqrt{np(1-p)}}{n} = \sqrt{\dfrac{pq}{n}}$ (where $q = (1-p)$)

So for this problem the proportion is p = 0.49 with a standard deviation of

$$\sigma(\hat{p}) = \sqrt{\frac{pq}{n}}\sqrt{\frac{(0.49)(0.51)}{1000}} = 0.0158 \text{ or } 1.58\%.$$

Since it is normal, we know that 95% of normally distributed values are within 2 standard deviations of the mean, so we should not be surprised if 95% of various polls gave results that were near 49% but varied above or below that by no more than 3.16% (1.58 x 2). The proportions supporting Kerry found in the two polls above 49% and 45.9% are both consistent with a true proportion of 49%. This is what we mean by **sampling error**. It is not really an *error* at all, but just *variability* you'd expect to see from one sample to another. A better term would be **sampling variability**.

In other words, if we draw repeated random samples of the same size, *n*, from the same population and measure the proportion, $\hat{p}$, we get for each sample, then the collection of these proportions will pile up around the underlying population proportion, *p,* in such a way that a histogram of the sample proportions can be modeled by a normal model.

Most models are useful only when specific assumptions are true. Here there are two:

1. The sampled value must be independent of each other.
2. The sample size, *n*, must be large enough.

The following are the conditions before using the normal model:

1. Randomization Condition: The sample should be a SRS. Or at least the sampling method was not biased and the sample should represent the population.
2. 10% Condition: The sample size, *n*, must be no larger than 10% of the population. For the polls, the population is so large, so the 1000 that were sampled is a small fraction of the population.
3. Success/Failure Condition: The sample size has to be big enough so that both *np* and *nq* are at least 10. We need to expect at least 10 successes and at least 10 failures to have

enough data for sound conclusions. For the polls, a “success” might be voting for Kerry. With p = 0.49, we expect 1000 x 0.49 = 490 successes and 1000 x 0.51 = 510 failures.

Think of proportions from random samples as random quantities and then say something this specific about their distribution is a fundamental insight. No longer is a proportion something we just compute for a set of data. We see it as a random quantity that has a distribution. We call that distribution the **sampling distribution model** for the proportion. This allows us to determine the amount of variation we should expect in samples. Suppose we spin a coin 100 times in order to decide whether it is fair or not. If we get 52 heads, we’re probably not surprised. Are we surprised to get 90 heads? What about 64 heads? In this case we need a sampling distribution model. The sampling model quantifies the variability, telling us how surprising a sample proportion is.
The Sampling Distribution Model for Proportions: Provided that the sample values are independent and the sample size is large enough, the sampling distribution of $\hat{p}$ is modeled by a normal model with mean $\mu(\hat{p}) = p$ and standard deviation $\sigma(\hat{p}) = \sqrt{\frac{pq}{n}}$.

.

Example: Suppose that about 13% of the population is left-handed. A 200-seat school auditorium has been built with 15 “lefty seats,” seats that have a built-in desk on the left rather than the right arm of the chair. In a class of 90 students, what is the probability that there will not be enough seats for the left-handed students?
Answer: Since 15 out of 90 is16.7%, we need to find the probability of finding more than 16.7% left-handed students out of a sample of 90 if the proportion of lefties is 13%.
Check the conditions:
1. Randomization: 90 students in the class can be thought of as a random sample of students.
2. 10% Condition: 90 is surely less than 10% of the population of all students.
3. Success/Failure Condition: $np = 90(0.13) = 11.7 \geq 10$ & $nq = 90(0.87) = 78.3 \geq 10$.

The population proportion is p = 0.13. Since the conditions are met, we will model the sampling distribution of p with a normal model with mean 0.13 and standard deviation of

$$\sigma(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.13)(0.87)}{90}} = 0.035 .$$

Now find the z-score and look on the table of normal probabilities.

$z = \frac{\hat{p} - p}{\sigma(\hat{p})} = \frac{0.167 - 0.13}{0.035} 1.06$, so $P(\hat{p}) > 0.167 = P(z > 1.06) = 0.1446$. There is about a 14.5% chance that there will not be enough seats for the left-handed students in the class.

Sampling Distribution of a Mean

The winter 1995 issue of *Chance* magazine gave data on the length of the overtime period for all 251 National Hockey League (NHL) playoff games between 1970 and 1993 that went into overtime. In the hockey playoff, the overtime period ends as soon as one of the teams scores a goal. The figure displays a histogram of the data. The graph shows that although most overtime periods lasted less than 20 minutes, a few games had long overtime periods. If we think of the 251 values as a population, the histogram shows the distribution of values in that population. We found the mean of these values to be μ = 9.841, so that is the balance point for the population histogram. The median value for the population is 8.000. For each of the sample sizes n = 5, 10, 20, and 30, we selected 500 random samples of size *n*. Then the histograms were constructed for each of the four sample sizes. The histograms are shown below. As with the samples from a normal population, the averages of the 500 means for the four different sample sizes are all close to the population mean μ = 9.841. Comparison of the four $\bar{x}$ histograms also show as *n* increases, the histogram's spread about the center decreases. So, $\bar{x}$ is less variable for a large sample size than it is for a small sample size. The fifth histogram below is a histogram based on narrower class intervals for the $\bar{x}$ values from samples of size 30. This figure shows that for *n* = 30, the histogram has a shape much like a normal curve. This is predicted by the ***Central Limit Theorem***.

The sampling distribution of *any* mean becomes more nearly normal as the sample size grows; *this is true regardless of the shape of the population distribution*.

Central Limit Theorem: The mean of a random sample has a sampling distribution whose shape can be approximated by a normal model. The larger the sample, the better the approximation will be. Do not mistakenly think the CLT says that the data are normally distributed as long as the sample is large enough. As samples get larger, we expect the distribution of the data to look more like the population from which it is drawn. You can collect a sample of CEO salaries for the next 1000 years, but the histogram will be skewed to the right. The CLT does not talk about the distribution of the data from the sample. It talks about the sample means and sample proportions of many different random samples drawn from the same population. We never draw all those samples, so the CLT is talking about an imaginary distribution—the sampling distribution model.

The Sampling Distribution Model for a Mean:

When a random sample is drawn from any population with mean and standard deviation, its sample mean has a sampling distribution with the same mean but whose stand deviation is. No matter what population the random sample comes from, the shape of the sampling distribution is approximately normal as long as the sample size is large enough. The larger the sample used, the more closely the normal approximates the sampling distribution for the mean (Central Limit Theorem n $\geq$ 30).

Summary:

1. $\mu_{\bar{x}} = \mu$

2. $\sigma(\bar{x}) = \dfrac{\sigma}{\sqrt{n}}$. This rule is exact if the population is infinite and is approximately correct when the population is finite if no more than 10% of the population is included in the sample.

3. When the population distribution is normal, the sampling distribution of $\bar{x}$ is also normal for any sample size *n*.

4. When is sufficiently large, the sampling distribution of $\bar{x}$ is well approximated by a normal curve, even when the population distribution is itself not normal.

If n is large or if the population distribution is normal, then the standardized variable $z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ has (at least approximately) a standard normal distribution.

Example: Suppose that mean adult weight is 175 pounds with a standard deviation of 25 pounds. An elevator in a building has a weight limit of 10 persons or 2000 pounds. What is the probability that the 10 people who get on the elevator overload its weight limit?

Check the conditions!!
Randomization: We will assume that the 10 people getting on the elevator are a random sample from the population.
Independence: It is reasonable to think that the weights of 10 randomly sampled people will be independent.
10% Condition: 10 people are surely less than 10%of the population of possible elevator riders.
Large enough sample: The distribution of population weights is roughly symmetric, so the sample of 10 seems large enough.

Since the conditions are satisfied, the Central Limit Theorem says that the sampling distribution of $\bar{x}$ has a normal model with mean 175 and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{25}{\sqrt{10}} = 7.91$. Now find the standardized variable, z.
$z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{200-175}{7.91} = 3.16$, so $P(\bar{x}) > 200 = P(z > 3.16) = 0.0008$. The chance that a random collection of 10 adults will exceed the elevator's weight limit is only 0.0008. So, if they are a random sample, it is quite unlikely that 10 people will exceed the total weight limit allowed on the elevator.

Both of the sampling distributions we looked at are normal. We know for proportions, $\sigma(\hat{p}) or\, \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$, and for means, $\sigma(\bar{x}) or\, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. These are great if we know, or pretend that we know, $p\ or\ \sigma$, and sometimes we'll do that. Often we know only the observed proportion, $\hat{p}$, or the sample standard deviation, *s*. We use what we know and we estimate. That may not seem like a big deal, but it gets a special name. Whenever we estimate the standard deviation of a sampling distribution, we call it a **standard error**. (Not a great name because it is not standard and nobody made an error. But it is shorter than saying, "the estimated standard deviation of the sampling distribution of the sample statistic.")

For a sample proportion, $\hat{p}$, the standard error is $SE(\hat{p}) or\, SE_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$.

For the sample mean, $\bar{x}$, the standard error is $SE(\bar{x}) or\, SE_{\bar{x}} = \frac{s}{\sqrt{n}}$.