Lecture 18

Nancy Pfenning Stats 1000

Chapter 9: Means and Proportions as Random Variables

Recall: in Section 1.3, we stated that the normal curve is idealized: its curve depicts an idealized histogram from a population with infinite possible values, all falling into a precise pattern. We call its mean μ and its standard deviation σ.

Example

> Height (a quantitative variable) of college-aged women in the U.S. is normal with $\mu = 65, \sigma = 2.7$. Of course this is idealized—we haven't measured all of them!

On the other hand, a set of actual observations from a variable x has a mean $\bar{x}$ and a standard deviation s.

Example

> Heights of women in this class have mean $\bar{x} =$_______$0.75in$, standard deviation $s =$_______$0.75in$.

Analogously for categorical variables,

Example

> The proportion of women in the U.S. is $p = .5$.

Example

> The proportion of women in this class is $\hat{p} =$_______$1in$.

In the first and third examples, μ and σ, p are **parameters**, numbers which describe the entire *population*.

In the second and fourth examples, $\bar{x}$ and s, $\hat{p}$ are **statistics**, numbers computed or measured from *sample* data.

Example

> Identify the sample, the population, the statistic, and the parameter of interest in each of the following:
>
> 1. A survey is carried out at a university to estimate the proportion of undergraduates living at home during the current term. **sample**: the undergrads surveyed; **population**: all undergrads at that university; **statistic**: proportion of sampled undergrads living at home; **parameter**: proportion of all undergrads at that university living at home.
> 2. In 1988, investigators chose 400 teachers at random from the National Science Teachers Association list and polled them as to whether or not they believed in the biblical creation. Of 200 respondents, 30% did believe. **sample**: 200 respondents; **population**: National Science Teachers Association members; **statistic**: 30% (proportion of believers in sample); **parameter**: unknown proportion of all NSTA members believing in biblical creation.
> 3. A survey of 1000 households in a certain city found their mean household size to be approximately 3.1 persons. **sample**: 1000 households surveyed; **population**: all households in that city; **statistic**: 3.1 (mean size of sampled households); **parameter**: unknown mean household size in city.
> 4. A balanced coin is flipped 100 times and the percentage of heads is 47%. **sample**: the 100 flips; **population**: all coin flips; **statistic**: 47%; **parameter**: 50% (percentage of all coinflips that would result in heads).

Ultimately, we will measure statistics and use them to draw conclusions about unknown parameters [statistical inference, or reasoning “backward”].

First, we must discover, for a given parameter, how the accompanying statistic tends to behave [reasoning “forward”, which is accomplished through use of the laws of probability]. This forward reasoning process is in a way impractical, because in real life parameters are usually unknown and cannot be “given”. But we need to learn how to reason ”forward” before we can learn to reason “backward”.

In another 100 coinflips, the percentage of heads will easily differ from 47%; another 1000 households wouldn't necessarily have a mean size of 3.1 persons. **Sampling variability** is a fact of life. Although the outcome for any one sample is unknown, it is also a fact that regular and predictable patterns emerge in the long run. We will get a feel for these patterns by examining the **sampling distributions** of means and proportions.

In practice, just one sample is taken from a population of categorical values and the statistic $\hat{p}$, sample proportion, is measured—one time only. In theory, we may consider values of $\hat{p}$ for repeated samples, in order to get an idea of how sample proportion as a *variable* behaves. For samples taken at random, sample proportion $\hat{p}$ is a **random variable**. To get an idea of how such a random variable behaves, we consider its **sampling distribution**: the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Sampling Distribution of Sample Proportion

Example

Suppose the population proportion of blue M&M's in a large bowl is $p = .17$. What kind of values would sample proportion $\hat{p}$ of blue M&M's take for repeated samples of

1. $n = 25$ (a teaspoon)?
2. $n = 75$ (a Tablespoon)?

1. Behavior of sample proportion for samples of size **25** taken from a population whose proportion in the category of interest (blue) is .17:
   (a) **center:** Some sample proportions will be less than .17, some more, but they'll tend to go below .17 just as much as they go above: the **mean** of sample proportion $\hat{p}$ should equal population proportion $p = .17$.
   (b) **spread:** How much the sample proportions vary depends on the size of the sample. If we'd only taken samples of size 5, sample proportion of blues could vary all the way from $\frac{0}{5} = .00$ to $\frac{3}{5} = .60$. For samples of size 25, sample proportions would tend not to vary this much.
   (c) **shape:** The most common sample proportion in the long run will be about .17, with proportions below and above .17 becoming less and less likely; we'd expect a single-peaked symmetric bell-shape with tapering ends. In other words, it should follow the normal curve!
2. Behavior of sample proportion $\hat{p}$ for samples of size **75** taken from a population whose proportion in the category of interest (blue) is $p = .17$:

   This distribution should also be **centered** at .17. There should be less **spread** than for samples of 25. Once again, the **shape** should be normal.

The laws of probability will confirm that what we expect to see in practice should also hold in theory. Under the right circumstances, statistical theory dictates the occurrence of precisely the same phenomena that can be observed in practice. First of all, our theory assumes a binomial model: in order for observations to be approximately independent, the sample size must not be too large relative to population size. Thus, we need a population at least ten times sample size. Next, in order for the Central Limit Theorem to apply, the sample size n must be large enough relative to population shape, which is determined by the value of p. [$p = .5$ means the population is symmetric, and a smaller sample should be adequate; p closer to 0 or 1 means the population is skewed left or right, and a larger sample is needed.] We will require that $np \geq 10$ and $n(1-p) \geq 10$. [Note: this is identical to the requirement for use of a normal approximation to probabilities involving the binomial count X of successes. That's because the sample proportion $\hat{p} = \frac{X}{n}$ has the same shape as X; the scale is simply divided by n.]

If the above conditions hold, then we have the following

Rules for Sample Proportions

If numerous samples or repetitions of the same size are taken,

1. **center:** The mean of the distribution of sample proportion $\hat{p}$ will be the true proportion p from the population. [Thus, $\hat{p}$ is an unbiased estimator for p.]

2. **spread:** Standard deviation of sample proportion is

$$\sqrt{\frac{p(1-p)}{n}}$$

Thus, the spread decreases as sample size increases.

3. **shape:** The frequency curve made from proportions from the various samples will be approximately normal. [Central Limit Theorem]

Applying these rules to our M&M experiment, we can predict that

1. For a teaspoon (sample size 25),
   (a) The histogram of sample proportion values will be centered at population proportion, .17.
   (b) The standard deviation of $\hat{p}$ should be approximately $\sqrt{\frac{.17 * .83}{25}} = .075$
   (c) The histogram should be only roughly normal, because our requirement is not satisfied: $np = 25(.17) = 4.25$ is less than 10.

2. For a Tablespoon (sample size 75),
   (a) The histogram should also be centered at .17.
   (b) The standard deviation should be approximately $\sqrt{\frac{.17 * .83}{75}} = .043$
   (c) The histogram should be closer to normal, because the requirement is satisfied: $np = 75(.17) = 12.75$ and $n(1-p) = 75(.83) = 62.25$ are both greater than 10.

Recall: The Empirical Rule introduced in Chapter 2 stated that for any normal curve with mean μ, standard deviation σ, approximately

1. 68% of values should fall within 1 σ of μ.

2. 95% of values should fall within 2 σ of μ.

3. 99.7% of values should fall within 3 σ of μ.

This enables us to set up **probability intervals** for the sample proportion of blues in a Tablespoon. For samples of size 75, approximately

1. 68% of sample proportions should be within $1 * .043$ of .17, that is, in [.127,.213]

2. 95% of sample proportions should be within $2 * .043$ of .17, that is, in [.084,.256]

3. 99.7% of sample proportions should be within $3 * .043$ of .17, that is, in [.041,.299]

At this point, we should check how well our own sample proportions conformed to the Empirical Rule.

Example

Lacking any further information, one might begin by assuming that the proportion of freshmen taking intro Stats classes is .25. According to survey data, the sample proportion of freshmen among surveyed students is $\frac{35}{445} = .08$. If the population proportion were truly .25, then sample proportion would have mean .25 and standard deviation $\sqrt{\frac{(.25)(1-.25)}{445}} = .02$. The probability of a sample proportion as low as .08, coming from a population with proportion of freshmen equal to .25, would be

$$P(\hat{p} \leq .08) \approx P(Z \leq \frac{.08 - .25}{.02}) = P(Z \leq -8.5) \approx 0$$

I would characterize this as "virtually impossible" and so I now decide not to believe that the overall proportion of freshmen is .25.

Exercise: Assume the proportion of females in all intro Stat classes is $p = .5$. What are the mean and standard deviation of sample proportion, if population proportion were indeed .5? Use the class survey responses to find the sample proportion of females in the class. Then use a normal approximation to find the probability of a sample proportion as high as the one observed, if the population proportion were truly .5. Characterize the results, based on your probability, in words such as "not unusual", "unlikely", "almost impossible", etc. Finally, tell whether you believe p is .5.

Lecture 19

Sampling Distribution of Sample Mean

In practice, just one sample is taken from a population of quantitative values and the statistic $\bar{x}$, sample mean, is measured—one time only. In theory, we may consider values of $\bar{x}$ for repeated samples, in order to get an idea of how sample mean as a *variable* behaves. For samples taken at random, sample mean is a **random variable**, written $\bar{X}$. To get an idea of how such a random variable behaves, we consider its **sampling distribution**: the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Example

The population of possible rolls X for a single die (equally likely values {1,2,3,4,5,6}) has mean $\mu = 3.5$ and standard deviation $\sigma = 1.7$. The sample mean roll $\bar{X}$ of 2 dice takes on various values subject to the laws of chance—it is a random variable. We can summarize its sampling distribution—just as we summarized distributions of data values in Chapter 2—by telling about its **center**, **spread**, and **shape**.

1. Sometimes the mean roll of 2 dice will be less than 3.5, sometimes greater than 3.5. It should be just as likely to get a lower-than-average mean than a higher-than-average mean: the sampling distribution of sample mean roll $\bar{X}$ should be **centered** at 3.5.
2. For the roll of 2 dice, the sample mean roll $\bar{X}$ will have a fair amount of **spread**: sample means all the way from 1 (if two 1's are rolled) to 6 (if two 6's are rolled) are not uncommon.
3. The most likely mean roll is 3.5 (resulting from (1,6), (2,5), (3,4), (4,3), (5,2), or (6,1)). Lower or higher mean rolls are progressively less likely, with 1 (two 1's are rolled) and 6 (two 6's are rolled) being least likely. Thus, the **shape** should be somewhat triangular: highest in the middle at 3.5, descending on either side.

Example

The sample mean roll $\bar{X}$ of 8 dice is also a random variable whose sampling distribution can be summarized by telling its center, spread, and shape.

1. Sometimes the mean roll of 8 dice will be less than 3.5, sometimes greater than 3.5. It should be just as likely to get a lower-than-average mean than a higher-than-average mean: the sampling distribution of sample mean roll $\bar{X}$ should be **centered** at 3.5.
2. For the roll of 8 dice, the distribution of sample mean roll $\bar{X}$ would not be as **spread** as that for 2 dice. All eight 1's or 6's will almost never happen: rolling this many dice at once, there tend to be some low numbers that balance out the high numbers.
3. The most likely mean roll is still 3.5 , with lower or higher mean rolls progressively less likely. But now there is a much better chance of the mean being close to 3.5, and a much worse chance of being as low as 1 or as high as 6: The **shape** of the sampling distribution bulges at the mean 3.5 and tapers away at either end: it appears normal!

Rules For Sample Means

These examples suggest some general results for the sampling distribution of *any* sample mean: Suppose a simple random sample of size n is taken from a population of quantitative values for a random variable X having mean μ and finite standard deviation σ. Then the following hold for the sampling distribution of sample mean $\bar{X}$:

1. The distribution of $\bar{X}$ is centered at μ. Thus, if we are using sample mean $\bar{x}$ (a statistic) to estimate population mean μ (a parameter), we may sometimes under-estimate and sometimes over-estimate, but there will be no systematic tendency either way. Thus, we say $\bar{X}$ is an **unbiased estimator** of μ.

2. The distribution of $\bar{X}$ has more spread for smaller samples, less spread for larger samples. In fact, it can be shown that the **standard deviation** of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$, where σ is population standard deviation. Thus, we can tell precisely how much the spread decreases as sample size increases: increasing from 2 to 8 dice means the spread of $\bar{X}$ decreases from $\frac{1.7}{\sqrt{2}} = 1.2$ to $\frac{1.7}{\sqrt{8}} = .6$.

3. For large sample size n, the sampling distribution of $\bar{X}$ is approximately normal. This is the celebrated **Central Limit Theorem**. A simpler situation is when the population itself is normal. Then sample mean $\bar{X}$ is guaranteed to be normal for *any* sample size n (even $n = 1$!).

We will summarize our results as follows: Take a simple random sample of size n from a population of values of a quantitative variable X and consider sample mean $\bar{X}$. If X is normal with mean μ, standard deviation σ, then $\bar{X}$ is normal with mean μ, standard deviation $\frac{\sigma}{\sqrt{n}}$. Otherwise, $\bar{X}$ is approximately normal with mean μ, standard deviation $\frac{\sigma}{\sqrt{n}}$ for large enough n. (C.L.T.)

How large is large enough? It depends on the shape of the population distribution. More observations are required if the shape of the population distribution is far from normal.

The above rules enable us to set up **probability intervals** for the sample mean roll of 8 dice:

For samples of size 8 coming from a population with mean 3.5, standard deviation 1.7, approximately

1. 68% of sample means should be within $1 * 1.7$ of 3.5, that is, in [2.9, 4.1]

2. 95% of sample means should be within $2 * 1.7$ of 3.5, that is, in [2.3, 4.7]

3. 99.7% of sample means should be within $3 * 1.7$ of 3.5, that is, in [1.7, 5.3]

At this point, we should check how well our own sample means conformed to the Empirical Rule.

Example

Women's heights are normal with mean 64.5, standard deviation 2.5. Pick *one* woman at random. According to the 68-95-99.7 Rule, the probability is

68% that her height X is between 62 and 67 inches

95% that her height X is between 59.5 and 69.5 inches

99.7% that her height X is between 57 and 72 inches.

Now pick a random sample of *25* women. Their sample mean height $\bar{X}$ is normal with mean 64.5, standard deviation $\frac{2.5}{\sqrt{25}} = .5$. The probability is

68% that their sample mean height $\bar{X}$ is between 64 and 65 inches

95% that their sample mean height $\bar{X}$ is between 63.5 and 65.5 inches

99.7% that their sample mean height $\bar{X}$ is between 63 and 66 inches

Thus, the sample of 25 heights has a mean which is, according to the laws of probability, much closer to the true mean than the value for a single height would be. Also, note the tradeoff: *lower* probability of mean height being in a *narrower* interval, *higher* probability of mean height being in a *wider* interval. Such tradeoffs will be encountered later with "confidence intervals".

Example

What is the probability that the height of a randomly chosen woman is less than 63.75 inches?

$$P(X < 63.75) = P(Z < \frac{63.75 - 64.5}{2.5}) = P(Z < -.3) = .3821$$

Example

What is the probability that sample mean height for a random sample of 25 women is less than 63.75 inches?

$$P(\bar{X} < 63.75) = P(Z < \frac{63.75 - 64.5}{\frac{2.5}{\sqrt{25}}}) = P(Z < -1.5) = .0668$$

Thus, it is not unusual for an individual woman to be less than 63.75 inches, but it would be unusual for the mean height of 25 women to be that low.

Example

Household size X in the U.S. has mean $\mu = 2.6$, standard deviation $\sigma = 1.4$.

1. Do you think the population distribution is normal? No: most households will have about 1 or 2 or 3 people, but a few households will be unusually large—the distribution would be right-skewed.
2. Pick a household at random. Find the probability that the household size exceeds 2.7. We can't answer this without knowing the exact distribution; normal tables do not apply.
3. Take a random sample of 10 households. Find the probability that sample mean household size exceeds 2.7. Can't be done: sample size $n = 10$ is too small to expect the Central Limit Theorem to guarantee an approximately normal distribution of $\bar{X}$, so we cannot find probabilities from normal tables.
4. Take a random sample of 100 households. Find the probability that sample mean household size exceeds 2.7. Now $\bar{X}$ is approximately normal with mean 2.6, standard deviation $\frac{1.4}{\sqrt{100}}$) and so

$$P(\bar{X} > 2.7) = P(Z > \frac{2.7 - 2.6}{\frac{1.4}{\sqrt{100}}}) = P(Z > .71) = P(Z < -.71) = .2389$$

Example

We considered the distribution of mean roll of 2 dice and 8 dice. What about the mean roll of 100 dice? It is unlikely to stray far at all from the overall mean of 3.5.

According to the **law of large numbers**, the actually observed mean outcome $\bar{X}$ must approach the mean μ of the population as the number of observations increases.

Example

Presumably, heights (in inches) of young women have a mean of 64.5 and a standard deviation of 2.5. Sample mean height for a random sample of 281 women would have mean 64.5 and standard deviation $\frac{2.5}{sqrt281} = .149$ The observed sample mean for surveyed women's heights is 64.783. The probability of a sample mean this high, coming from a population with mean 64.5, is

$$P(\bar{X} \geq 64.783) = P(Z \geq \frac{64.783 - 64.5}{.149}) = P(Z \geq 1.90) = .0233$$

It's pretty unlikely to get a sample mean as high as 64.783, if population mean were 64.5. We have reason to suspect that the population mean is somewhat higher than 64.5. Some sources report the population mean as being 65; in reality, it could well be somewhere between 64.5 and 65.0.

Exercise: If students each picked a number truly at random from 1 to 20, then their responses would follow a "uniform distribution", with each of the numbers appearing with probability $\frac{1}{20} = .05$. It can be shown that the mean of all the numbers between 1 and 20 is 10.5, and the standard deviation is 5.77. What are the mean and standard deviation of **sample mean** selection for a sample of n students, if their selections are truly random? Use the class survey responses to find the sample mean "random" number selected. Then use a normal approximation to find the probability of a sample mean as high as the one observed, if the population mean were truly 10.5. Characterize the results, based on your probability, in words such as "not unusual", "unlikely", "almost impossible", etc. Finally, tell whether you have statistical evidence of bias in favor of higher numbers.