Math C067 Sampling Distributions

Sample Mean and Sample Proportion

Richard Beigel

Some time between April 16, 2007 and April 16, 2007

Examples of Sampling

- A pollster may try to estimate the proportion of voters in favor of a particular presidential candidate by polling a random sample of voters.
- A statistician may want to estimate the average starting income of new college graduates by finding the mean of a random sample of new college graduates.

Numbers

- Population size: N
- Sample size: n (usually much smaller than N)

Sampling with Replacement Choose 1 person, Choose a 2nd person (who could be the same as the first), Choose a 3rd person (who could be the same as the first and/or second), ..., Choose an nth person

Sampling without Replacement Choose n different people

Rules of thumb

- Our formulas will be **correct for sampling with replacement**.
- Our formulas will be **approximately correct for sampling without replacement** provided that the population size N is large.

Sample Mean Suppose that we are trying to estimate a random variable that is based on a large population, e.g., we want to estimate average income.

- Random Variable that we wish to estimate: X
- Expected value (mean) of X: μ
- Standard deviation of X: σ

We define a random variable $\overline{X}$ called the **Sample Mean** via the following random process:

- Evaluate X at n random points, i.e., choose n random people and find out how much each of them earns
- Find the average of those n values

Formulas

- The expected value of $\overline{X}$ is the same as the the expected value of X, i.e.,

$$\mu_{\overline{X}} = \mu$$

(correct even if we sample without replacement).

- The variance of $\overline{X}$ is smaller than the variance of X by a factor of n (the sample size), i.e.,

$$\sigma^2_{\overline{X}} = \frac{1}{n}\sigma^2$$

- The standard deviation of $\overline{X}$ is smaller than the standard deviation of X by a factor of $\sqrt{n}$ (the sample size), i.e.,

$$\sigma_{\overline{X}} = \frac{1}{\sqrt{n}}\sigma$$

Example: Test scores. One million students take a standardized test. Their average score is 500. The standard deviation of their scores is 100. A statistician chooses 400 students at random with replacement and computes the mean $\overline{X}$ of their test scores. What is the expected value of $\overline{X}$? What is the standard deviation of $\overline{X}$?

- $\mu_{\overline{X}} = \mu = 500$
- $\sigma_{\overline{X}} = \frac{1}{\sqrt{n}}\sigma = \frac{1}{\sqrt{400}} \cdot 100 = \frac{1}{20} \cdot 100 = 5$

Note: these formulas do not depend in any way on the population size N.

Example: Test scores. One million students take a standardized test. Their average score is 500. The standard deviation of their scores is 100. A statistician chooses a group of 400 different students at random and computes the mean $\overline{X}$ of their test scores. What is the expected value of $\overline{X}$? What is the standard deviation of $\overline{X}$?

- $\mu_{\overline{X}} = \mu = 500$. This formula works with or without replacement.
- Because the population size is large we can use the same formula we used with replacement to **estimate** the standard deviation.

$$\sigma_{\overline{X}} \approx \frac{1}{\sqrt{n}}\sigma = \frac{1}{\sqrt{400}} \cdot 100 = \frac{1}{20} \cdot 100 = 5$$

Example: Average income 10 million people work in Metropolis. Their average income is \$100,000. The standard deviation of their incomes is 2500. A statistician chooses 100 different Metropolis' workers at random and computes their average income I. Was the sample chosen with replacement or without replacement? What is the expected value of I? What is the standard deviation of I? What is the probability that I is between \$99,500 and \$100,500.

- I is just a sample mean, so we can call it $\overline{X}$. (But the problem statement can call it anything. It's up to you to recognize that I is the sample mean.)

$$\mu_I = \mu_{\overline{X}} = \mu = 100000$$

-

$$\sigma_I = \sigma_{\overline{X}} \approx \frac{1}{\sqrt{n}}\sigma = \frac{1}{\sqrt{100}} \cdot 2500 = \frac{1}{10} \cdot 2500 = 250$$

- **The Sample Mean is approximately normally distributed** so we can use the methods of chapter 6 to estimate $\Pr[99500 \leq \overline{X} \leq 100500]$. We are looking at 2 standard deviations on each side of the mean so

$$\Pr[99500 \leq \overline{X} \leq 100500] \approx 0.4772 + 0.4772 = 0.9544$$

Note: You can approximate a Sample Mean by a Normal Distribution provided that $n \geq 30$.

Sample Proportion

- Suppose that a fraction (proportion) p of a population favors candidate C for president.
- We take a random sample consisting of n people
- Let X be a random variable that denotes the number of people in the sample who favor candidate C
- Then the distribution of X is $B(n, p)$.
- Let $\hat{P}$ be a random variable that denotes the fraction (proportion) of people in the sample that favor candidate C
- Then $\hat{P} = \frac{1}{n}X$
- $\hat{P}$ is called the Sample Proportion
- $E(X) = np$, $E(\hat{P}) = p$
- $\text{Var}(X) = np(1-p)$, $\text{Var}(\hat{P}) = \frac{p(1-p)}{n}$
- $\sigma_X = \sqrt{np(1-p)}$, $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$

Example: An Election Suppose that 55 percent of the voters favor Candidate C for president. If a random sample of 1000 voters is chosen, estimate the probability that between 52 and 58 percent of the voters in the sample will favor Candidate C.

To answer this question we can work with $B(n, p)$ and estimate $\Pr[520 \leq X \leq 580]$ or we can work directly with the Sample Proportion and ask whether $\Pr[0.52 \leq \hat{P} \leq 0.58]$. Both random variables X and $\hat{P}$ are approximately normally distributed.

Solve this on your own and then compare your answer to the answers on the next page.

Solved Example: An Election Suppose that 55 percent of the voters favor Candidate C for president. If a random sample of 1000 voters is chosen, estimate the probability that between 52 and 58 percent of the voters in the sample will favor Candidate C.

To answer this question we can work with $B(n, p)$ and estimate $\Pr[520 \leq X \leq 580]$ or we can work directly with the Sample Proportion and ask whether $\Pr[0.52 \leq \hat{P} \leq 0.58]$. Both random variables X and $\hat{P}$ are approximately normally distributed.

Solve this on your own and then compare your answer to the answers on the next page.

Solution using Binomial Distribution

- $E[X] = np = 1000 \cdot 0.55 = 550.$
- $\text{Var}[X] = npq = np(1-p) = 1000 \cdot 0.55 \cdot 0.45 = 247.5$
- $\sigma_X = \sqrt{\text{Var}[X]} = \sqrt{247.5} \approx 15.73$

$$\begin{aligned} \Pr[520 \leq X \leq 580] &= \Pr[519.5 \leq X \leq 580.5] \\ &= \Pr[519.5 \leq X \leq 550] + \Pr[550 < X \leq 580.5] \end{aligned}$$

The first interval consists of $\frac{550-519.5}{15.73} \approx 1.94$ standard deviations. The second interval consists of $\frac{580.5-550}{15.73} \approx 1.94$ standard deviations. Therefore

$$\Pr[520 \leq X \leq 580] \approx 0.4738 + 0.4738 = 0.9476$$

Did you get the correct answer? If not, go back and try again.

Solution using Sample Proportion

- $E[\hat{P}] = p = 0.55$
- $\text{Var}[\hat{P}] = \frac{pq}{n} = \frac{p(1-p)}{n} = \frac{0.55 \cdot 0.45}{1000} = 0.0002475$
- $\sigma_{\hat{P}} = \sqrt{\text{Var}[\hat{P}]} = \sqrt{0.0002475} \approx 0.01573$

$$\Pr[0.52 \leq \hat{P} \leq 0.58] = \Pr[0.52 \leq \hat{P} \leq 0.55] + \Pr[0.55 < \hat{P} \leq 0.58]$$

The first interval consists of $\frac{0.55-0.52}{0.01573} \approx 1.91$ standard deviations. The second interval consists of $\frac{0.58-0.55}{0.01573} \approx 1.91$ standard deviations. Therefore

$$\Pr[0.52 \leq \hat{P} \leq 058] \approx 0.4719 + 0.4719 = 0.9438$$

(The answer we got by using the Binomial Distribution was more accurate because we expanded the interval by 0.5 in both directions.)

Did you get the correct answer? If not, go back and try again.

Solved Example: 1936 Presidential Election Suppose that 61% of the voters favor Candidate R for President. You poll a random sample consisting of 3000 voters. What is the probability that a majority of your sample favors Candidate R?

Try this yourself before reading on.

Solved Example: 1936 Presidential Election Suppose that 61% of the voters favor Candidate R for President. You poll a random sample consisting of 3000 voters. What is the probability that a majority of your sample favors Candidate R?

Solution using Binomial Distribution Let X denote the number of voters in your sample who favor Candidate R. The distribution of X is $B(3000, 0.61)$. We wish to estimate $\Pr[X > 1500]$, which is the same as $\Pr[X \geq 1501]$.

- $E[X] = np = 3000 \cdot 0.61 = 1830.$
- $\text{Var}[X] = npq = np(1-p) = 3000 \cdot 0.61 \cdot 0.39 = 713.7.$
- $\sigma_X = \sqrt{\text{Var}[X]} = \sqrt{713.7} \approx 26.72$

$$\begin{aligned} \Pr[X \geq 1501] &= \Pr[X \geq 1500.5] \\ &= \Pr[1500.5 \leq X] \\ &= \Pr[1500.5 \leq X \leq 1830] + \Pr[1830 < X] \end{aligned}$$

The first interval consists of approximately $\frac{1830-1500.5}{26.72} \approx 12.33$ standard deviations. The second interval consists of half of the distribution. Therefore,

$$\Pr[X \geq 1501] \approx 0.5 + 0.5 = 1.0$$

Solution using Sample Proportion

- $E[\hat{P}] = p = 0.61$
- $\text{Var}[\hat{P}] = \frac{pq}{n} = \frac{p(1-p)}{n} = \frac{0.61 \cdot 0.39}{3000} = 0.0000793$
- $\sigma_{\hat{P}} = \sqrt{\text{Var}[\hat{P}]} = \sqrt{0.0000793} \approx 0.008905$

$$\begin{aligned} \Pr[\hat{P} > 0.5] &= \Pr[0.5 < \hat{P}] \\ &= \Pr[0.5 < \hat{P} \leq 0.61] + \Pr[0.61 < \hat{P}] \end{aligned}$$

The first interval consists of approximately $\frac{0.61-0.5}{0.008905} \approx 12.35$ standard deviations. The second interval consists of half of the distribution. Therefore,

$$\Pr[\hat{P} > 0.5] \approx 0.5 + 0.5 = 1.0$$

Gallup's Gambit In a 1935, George Gallup came up with a clever marketing strategy for his opinion polls. Gallup bet that he would correctly predict the results of the 1936 Presidential Election. He promised to refund all subscription fees for the entire year if his prediction was incorrect.

Gallup did predict the Election results correctly based on a sample of approximately 3000 voters. Reader's Digest **incorrectly** predicted the result based on a sample of 10,000,000 voters.

How could Gallup predict correctly based on such a small sample? Because the actual election wasn't really close, even his small sample provided a comfort level of 12 standard deviations.

How could Reader's Digest predict incorrectly based on such a large sample? Their sample was not representative, but consisted mainly of people who had a car or a telephone.

Lesson: Randomness is much more important than sample size. (Accuracy in sampling is also important, since some people lie to pollsters. There is a science to designing questionnaires.)

Some recent elections have been decided by less than a 1% margin. It is interesting to look at how Gallup's poll might have turned out, had the 1936 Election not been a landslide.

Example: What if the Election had been close? Suppose that 50.5% of the voters favor Candidate R for President. You poll a random sample consisting of 3000 voters. What is the probability that a majority of your sample favors Candidate R?

Try this one yourself before reading the solution.

Example: What if the Election had been close? Suppose that 50.5% of the voters favor Candidate R for President. You poll a random sample consisting of 3000 voters. What is the probability that a majority of your sample favors Candidate R?

Solution using Binomial Distribution Let X denote the number of voters in your sample who favor Candidate R. The distribution of X is $B(3000, 0.505)$. We wish to estimate $\Pr[X > 1500]$, which is the same as $\Pr[X \geq 1501]$.

- $E[X] = np = 3000 \cdot 0.505 = 1515$.
- $\text{Var}[X] = npq = np(1-p) = 3000 \cdot 0.505 \cdot 0.495 = 749.925$.
- $\sigma_X = \sqrt{\text{Var}[X]} = \sqrt{749.925} \approx 27.38$

$$\begin{aligned} \Pr[X \geq 1501] &= \Pr[X \geq 1500.5] \\ &= \Pr[1500.5 \leq X] \\ &= \Pr[1500.5 \leq X \leq 1515] + \Pr[1515 < X] \end{aligned}$$

The first interval consists of approximately $\frac{1515-1500.5}{27.38} \approx 0.53$ standard deviations. The second interval consists of half of the distribution. Therefore,

$$\Pr[X \geq 1501] \approx 0.2019 + 0.5 = 0.7019$$

Thus Gallup's chance of predicting the Election correctly would have been about 70%, assuming he used a representative sample. If his sample was skewed (like *Reader's Digest*'s), then his chance could have been much smaller.

Solution using Sample Proportion

- $E[\hat{P}] = p = 0.505$
- $\text{Var}[\hat{P}] = \frac{pq}{n} = \frac{p(1-p)}{n} = \frac{0.505 \cdot 0.495}{3000} = 0.000083325$
- $\sigma_{\hat{P}} = \sqrt{\text{Var}[\hat{P}]} = \sqrt{0.000083325} \approx 0.009128$

$$\begin{aligned} \Pr[\hat{P} > 0.5] &= \Pr[0.5 < \hat{P}] \\ &= \Pr[0.5 < \hat{P} \leq 0.505] + \Pr[0.505 < \hat{P}] \end{aligned}$$

The first interval consists of approximately $\frac{0.505-0.5}{0.009128} \approx 0.55$ standard deviations. The second interval consists of half of the distribution. Therefore,

$$\Pr[\hat{P} > 0.5] \approx 0.2088 + 0.5 = 0.7088$$

Thus Gallup's chance of predicting the Election correctly would have been about 71%, assuming he used a representative sample. If his sample was skewed (like *Reader's Digest*'s), then his chance could have been much smaller.

Example: How good was Gallup's Sample? Only 54% of Gallup's sample favored Roosevelt.

Suppose that 61% of the voters favor Candidate R for President. You poll a random sample consisting of 3000 voters. What is the probability that 54% or less your sample favors Candidate R?

Try this yourself before reading on.

Solved Example: How good was Gallup's Sample? Suppose that 61% of the voters favor Candidate R for President. You poll a random sample consisting of 3000 voters. What is the probability that 54% or less your sample favors Candidate R?

Solution using Binomial Distribution Let X denote the number of voters in your sample who favor Candidate R. The distribution of X is $B(3000, 0.61)$. 54% of 3000 is 1620. We wish to estimate $\Pr[X \leq 1620]$.

- $E[X] = np = 3000 \cdot 0.61 = 1830.$
- $\text{Var}[X] = npq = np(1-p) = 3000 \cdot 0.61 \cdot 0.39 = 713.7.$
- $\sigma_X = \sqrt{\text{Var}[X]} = \sqrt{713.7} \approx 26.72$

$$\begin{aligned} \Pr[X \leq 1620] &= \Pr[X \leq 1620.5] \\ &= \Pr[X \leq 1830] - \Pr[1620.5 < X \leq 1830] \end{aligned}$$

The first interval consists of half of the distribution. The second interval consists of $\frac{1830-1620.5}{26.72} \approx 7.84$ standard deviations. Therefore,

$$\Pr[X \leq 1620] \approx 0.5 - 0.5 = 0.0$$

This suggests that Gallup's sample was not close to uniform, i.e., not representative of the population. Perhaps Gallup just got lucky. They say that Fortune favors the bold.

Solution using Sample Proportion

- $E[\hat{P}] = p = 0.61$
- $\text{Var}[\hat{P}] = \frac{pq}{n} = \frac{p(1-p)}{n} = \frac{0.61 \cdot 0.39}{3000} = 0.0000793$
- $\sigma_{\hat{P}} = \sqrt{\text{Var}[\hat{P}]} = \sqrt{0.0000793} \approx 0.008905$

$$\Pr[\hat{P} \leq 0.54] = \Pr[\hat{P} \leq 0.61] - \Pr[0.54 < \hat{P} \leq 0.61]$$

The first interval consists of approximately $\frac{0.61-0.54}{0.008905} \approx 7.86$ standard deviations. The second interval consists of half of the distribution. Therefore,

$$\Pr[\hat{P} \leq 0.54] \approx 0.5 - 0.5 = 0.0$$

This suggests that Gallup's sample was not close to uniform, i.e., not representative of the population. Perhaps Gallup just got lucky. They say that Fortune favors the bold.

Gallup vs. the Tout If the 1936 Election had been a close one (decided by only a 1% margin), there is substantial probability (30% or more) that Gallup would have predicted it wrong. Was Gallup taking a big risk? Perhaps Gallup would have taken a larger sample if it looked like the election was going to be close. Or, perhaps Gallup had nothing to lose.

Suppose that you are gambler. A **tout** is a person who sells predictions, usually written on something called a "tout sheet". For example, the tout might predict whether the Phillies will beat the Cardinals tonight. The tout charges $1 for his prediction but he always offers a money-back guarantee. Key facts about tout sheets:

1. Tout sheets are worthless. Why? In reality, the tout will tell one customer that Philadelphia will win and tell another customer that Philadelphia will lose. So half of the predictions are guaranteed to be correct. (Not like Gallup Polls)

2. The tout never loses. When he predicts correctly, he earns $1. When he predicts incorrectly he earns $0. (Like Gallup, although Gallup also extended his offer to previous customers.)

3. Like most something-for-nothing schemes, selling tout sheets is illegal. (Not like Gallup.)

4. Touts thrive on repeat business. People who win are likely to return to the tout for another prediction, which is usually more expensive than the first. (Like Gallup, although I don't think he jacked up his rates.)

5. A tout's customers are known in gambling circles as "suckers." (Like Gallup's customers, if they were fooled by the gambit.)