OpenIntro Statistics

CH 04 A: Foundations for Inference

CH 4.1 Variability in Estimates

Young, Underemployed and Optimistic

Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy. An analysis of government economic data suggests that this perception is correct. The recent indicators on the nation's labor market show a decline in the

3

http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession

Young, Underemployed and Optimistic

Coming of Age, Slowly, in a Tough Economy

Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience. And more than one-third (35%) say that, as a result of the poor economy, they have gone back to school. Their personal lives have also been affected: 31% have postponed either getting married or having a baby (22% say they have postponed having a baby and 20% have put off getting married). One-in-four (24%) say they have moved back in with their parents after living on their own.

4

http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession

Margin of error

The general public survey is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States, including an oversample of 346 adults ages 18 to 34. A total of 769 interviews were completed with respondents contacted by landline telephone and 1,279 with those contacted on their cellular phone. Data are weighted to produce a final sample that is representative of the general population of adults in the continental United States. Survey interviews were conducted under the direction of Princeton Survey Research Associates International, in English and Spanish. Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

5

Margin of error

"… 41% believe young adults … are having the toughest time in today's economy"

41% ± 2.9%: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.

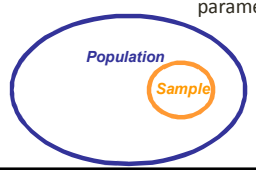
"… Among all 18- to 34- year olds, fully half (49%) say they have taken a job they didn't want just to pay bills … "

49% ± 4.4%: We are 95% confident that 44.6% to 53.4% of 18-34 years olds have taken a job they didn't want just to pay the bills.

6

Reminder: Parameter versus statistic

- Population: the entire group of individuals in which we are interested but can't usually assess directly.
- A **parameter** is a number describing a characteristic of the population. Parameters are usually unknown.
- Sample: the part of the population we actually examine and for which we do have data.
- A **statistic** is a number describing a characteristic of a sample. We often use a statistic to estimate an unknown population parameter.

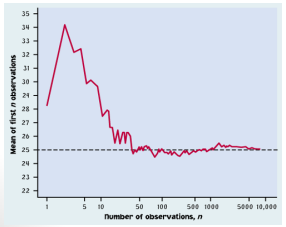


7

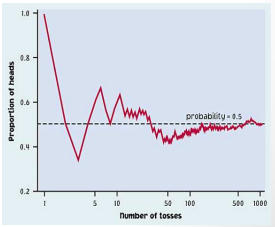
The law of large numbers

Law of large numbers: As the number of randomly-drawn observations (n) in a sample increases,

the mean of the sample ($\bar{x}$) gets closer and closer to the population mean μ (quantitative variable).



the sample proportion ($\hat{p}$) gets closer and closer to the population proportion p (categorical variable).



8

What is a sampling distribution?

The **sampling distribution of a statistic** is the distribution of all possible values taken by the statistic when all possible samples of a fixed size *n* are taken from the population.

It is a theoretical idea—we do not actually build it.

The sampling distribution of a statistic is the **probability distribution** of that statistic.

9

What is a sampling distribution?

When sampling randomly from a given population, the law of large numbers describes what happens when the sample size n is gradually increased.

The sampling distribution describes what happens when we take all possible random samples of a fixed size n.

10

Parameter estimation

We are often interested in ***population parameters***.

Since complete populations are difficult (or impossible) to collect data on, we use ***sample statistics*** as ***point estimates*** for the unknown population parameters of interest.

Sample statistics vary from sample to sample.

Quantifying how sample statistics vary provides a way to estimate the ***margin of error*** associated with our point estimate.

But before we get to quantifying the variability among samples, let's try to understand how and why point estimates vary from sample to sample.

11

Parameter estimation

Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Not the same, but only somewhat different.

12

Sampling Distributions

Sampling Distribution of the Sample Mean

https://istats.shinyapps.io/sampdist_cont/

13

Central Limit Theorem

The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right),$$

where *SE* is represents ***standard error***, which is defined as the standard deviation of the sampling distribution. If σ is unknown, use *s*.

14

Central Limit Theorem

It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population mean.

We won't go through a detailed proof of why $SE = \sigma / \sqrt{n}$, but note that as *n* increases *SE* decreases.

As the sample size increases we would expect samples to yield more consistent sample means, hence the variability among the sample means would be lower.

15

CLT - conditions

Certain conditions must be met for the CLT to apply:

***Independence*:** Sampled observations must be independent.

This is difficult to verify, but is more likely if

- random sampling / assignment is used, and
- if sampling without replacement, n < 10% of the population.

16

CLT - conditions

Certain conditions must be met for the CLT to apply:

***Independence*:**

***Sample size / skew*:** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large.

The more skewed the population distribution, the larger sample size we need for the CLT to apply

For moderately skewed distributions n > 30 is a widely used rule of thumb

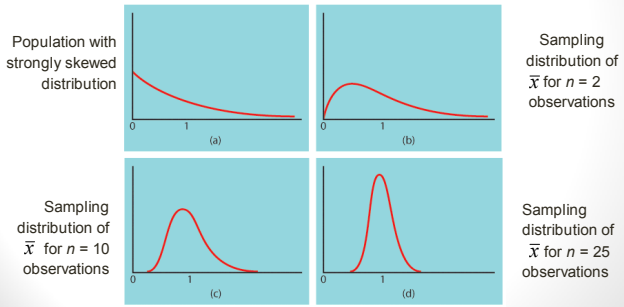
This is also difficult to verify for the population, but we can check it using the sample data, and assume that the sample mirrors the population.

17

CH 4.4 Examining the Central Limit Theorem

The central limit theorem

Central Limit Theorem: When randomly sampling from any population with mean μ and standard deviation σ, **when *n* is large enough,** the sampling distribution of $\bar{x}$ is approximately normal: $N(\mu, \sigma/\sqrt{n})$.



19

How large a sample size?

It depends on the population distribution. More observations are required if the population distribution is far from normal.

A sample size of 25 is generally enough to obtain a normal sampling distribution from a strong skewness or even mild outliers.

A sample size of 40 will typically be good enough to overcome extreme skewness and outliers.

In many cases, n = 25 isn't a huge sample. Thus, even for strange population distributions we can assume a normal sampling distribution of the mean, and work with it to solve problems.

20

IQ scores: population vs. sample

In a large population of adults, the mean IQ is 112 with standard deviation 20. Suppose 200 adults are randomly selected for a market research campaign.

- The distribution of the sample mean IQ is

A) exactly normal, mean 112, standard deviation 20.

B) approximately normal, mean 112, standard deviation 20.

C) approximately normal, mean 112 , standard deviation 1.414.

D) approximately normal, mean 112, standard deviation 0.1.

21

IQ scores: population vs. sample

In a large population of adults, the mean IQ is 112 with standard deviation 20. Suppose 200 adults are randomly selected for a market research campaign.

- The distribution of the sample mean IQ is

A) exactly normal, mean 112, standard deviation 20.

B) approximately normal, mean 112, standard deviation 20.

C) approximately normal, mean 112 , standard deviation 1.414.

D) approximately normal, mean 112, standard deviation 0.1.

Population distribution: $N(\mu = 112; \sigma = 20)$

Sampling distribution for $n = 200$ is $N(\mu = 112; \sigma/\sqrt{n} = 1.414)$

22

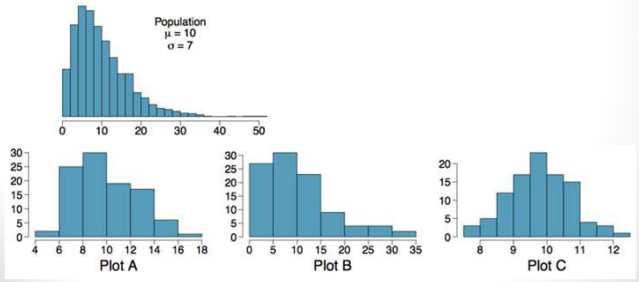
Practice

At top: distribution for a population (μ = 10, σ = 7). Which plot is....

A single random sample of 100 observations from this population. **B**

A distribution of 100 sample means from random samples with size 7. **A**

A distribution of 100 sample means from random samples with size 49. **C**



23

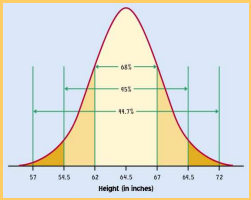
Further properties

The Central Limit Theorem is valid as long as we are sampling many small random events, even if the events have different distributions (as long as no one random event has an overwhelming influence).

Why is this cool?

It explains why so many variables are normally distributed.

Example: Height seems to be determined by a large number of genetic and environmental factors, like nutrition.



So height is very much like our sample mean $\bar{x}$. The "individuals" are genes and environmental factors. Your height is a mean.

Now we have a better idea of why the density curve for height has this shape.

24

Sampling Distributions

Sampling Distribution of the Sample Proportion

https://istats.shinyapps.io/SampDist_Prop/

25

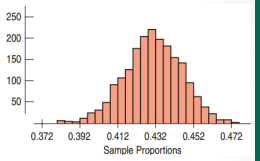
Sampling Distributions

Sampling Distribution for Proportions

Symmetric

Unimodal

Centered at p



The sampling distribution follows the Normal model.

What does the sampling distribution tell us?

The sampling distribution allows us to make statements about where we think the corresponding population parameter is and how precise these statements are likely to be.

26

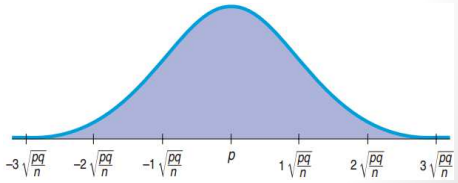
Mean and Standard Deviation

Sampling Distribution for Sample Proportion $\hat{p}$

Mean = p

$$\sigma(\hat{p}) = \frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$$

$$N\left(p, \sqrt{\frac{pq}{n}}\right)$$



27

When Does the Normal Model Work?

Success Failure Condition:
$np \geq 10$, $nq \geq 10$ There must be at least 10 expected successes and failures.

Independent trials: Check for the Randomization Condition.

10% Condition: Sample size less than 10% of the population size

28

Enough Lefty Seats?

13% of all people are left handed.

- A 200-seat auditorium has 15 lefty seats.
- What is the probability that there will not be enough lefty seats for a class of 90 students?

Think→

- **Plan:** 15/90 ≈ 0.167, Want $P(\hat{p} > 0.167)$
- **Model:**
  - ✓ **Independence Assumption:** With respect to lefties, the students are independent.
  - ✓ **10% Condition:** This is out of all people.
  - ✓ **Success/Failure Condition:** $15 \geq 10, 75 \geq 10$

29

Enough Lefty Seats?

Think→

- **Model:** $p = 0.13$, $SD(\hat{p}) = \sqrt{\frac{(0.13)(0.87)}{90}} \approx 0.035$

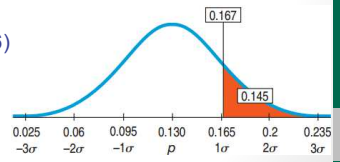
The model is: $N(0.13, 0.035)$

Show→

- Plot $z = \frac{0.167 - 0.13}{0.035} \approx 1.06$
- Mechanics:

$P(\hat{p} > 0.167) = P(z > 1.06)$

≈ 0.1446



30

Enough Lefty Seats?

Tell →

- **Conclusion:** There is about a 14.5% chance that there will not be enough seats for the left handed students in the class.

31

Sampling Distributions

Sampling Distribution of other statistics

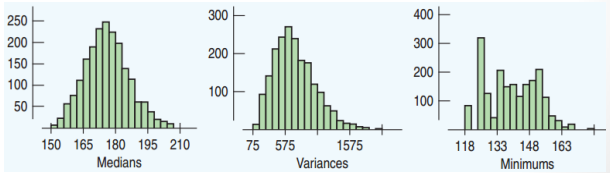
http://onlinestatbook.com/stat_sim/sampling_dist/

32

The Sampling Distribution for Other Statistics

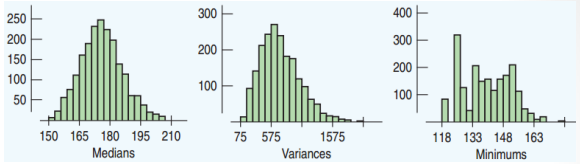
There is a sampling distribution for any statistic, but the Normal model may not fit.

Below are histograms showing results of simulations of sampling distributions.



33

The Sampling Distribution for Other Statistics



- The medians seem to be approximately Normal. !!!!!!
- The variances seem somewhat skewed right.
- The minimums are all over the place.
- In this course, we will focus on the proportions and the means.

34