Unit 3: Foundations for inference
Lecture 1: Variability in estimates and CLT

Statistics 101

Mine Çetinkaya-Rundel

September 24, 2013

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

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.

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.

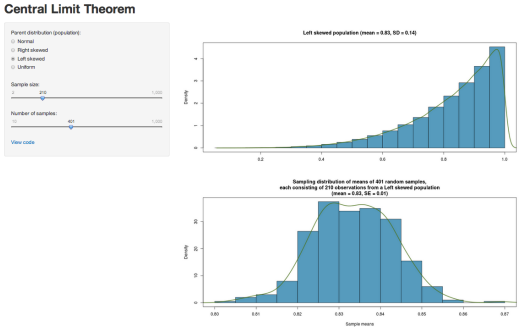
- 41% ± 2.9%: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.
- 49% ± 4.4%: We are 95% confident that 44.6% to 53.4% of 18-34 years olds have taken a job they didn't want just to pay the bills.

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.

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?

http://rundel.dyndns.org:3838/CLT/



Try for yourself:

- Pick normal distribution for the population. How does the shape of the sampling distribution change as n changes?
- Pick a non-normal distribution for the population. How does the shape of the sampling distribution change as n changes?
- Does the shape of the sampling distribution change when the number of samples changes?

Central limit theorem

Central limit theorem

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

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

If σ is unknown, use s.

- So it wasn't a coincidence that the sampling distributions we saw earlier were symmetric.
- We won't go into the proving why $SE = \frac{\sigma}{\sqrt{n}}$, 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.

CLT - conditions

Certain conditions must be met for the CLT to apply:

1. *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.
2. *Sample size/skew:* Either the population distribution is normal or
   $n > 30$ and the population distribution is not extremely skewed (the more skewed the distribution, the higher n necessary for the CLT to apply).

   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.

CLT - sample size/skew condition

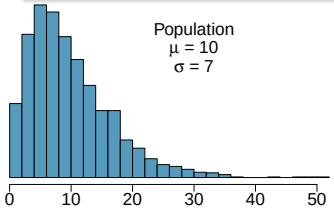
Clicker question

Which of the below visualizations is not appropriate for checking the shape of the distribution of the sample, and hence the population?

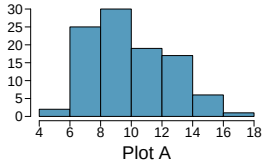
(a) histogram
(b) boxplot
(c) normal probability plot
(d) barplot

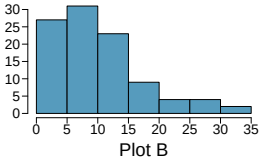
Clicker question

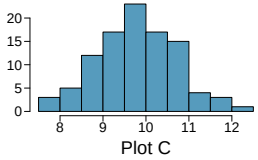
Four plots: Determine which plot (A, B, or C) is which.
(1) At top: distribution for a population ($\mu = 10, \sigma = 7$),
(2) a single random sample of 100 observations from this population,
(3) a distribution of 100 sample means from random samples with size 7, and
(4) a distribution of 100 sample means from random samples with size 49.



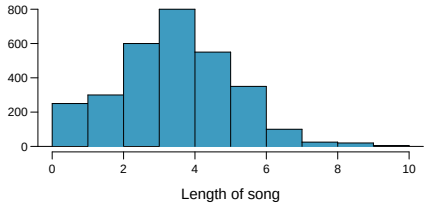
(a) A - (3); B - (2); C - (4)
(b) A - (2); B - (3); C - (4)
(c) A - (3); B - (4); C - (2)
(d) A - (4); B - (2); C - (3)







Suppose an iPod has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes. Calculate the probability that a randomly selected song lasts more than 5 minutes.



The distribution of the lengths of these songs is slightly right skewed to the right, therefore it is not reasonable to use a normal model to estimate this probability.
However, we can approximate the probability using the histogram.

$$P(X > 5) = \frac{350 + 100 + 25 + 20 + 5}{3000} = \frac{500}{3000} = 0.17$$

You are about to take a trip to visit your parents and the drive is 6 hours. You make a random playlist of 100 songs. What is the probability that your playlist lasts the entire drive?

We are looking for the probability that the total length of 100 songs is more than 6 hours = 360 minutes.

$$P(X_1 + X_2 + \cdots + X_{100} > 360\ min) = ?$$

Clicker question

Which of the following is equivalent to the desired probability above?

(a) P(each song on the playlist lasts at least 3.6 minutes)
(b) P(average song length is at least 3.6 minutes)

In order to calculate $P(\bar{X} > 3.6\ min)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = 3.45, SE = \frac{1.63}{\sqrt{100}} = 0.163\right)$$