

Math 183

Statistical Methods

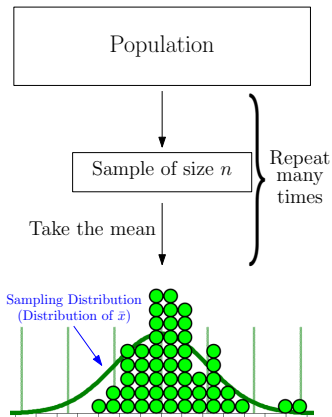
Eddie Aamari
S.E.W. Assistant Professor

`eaamari@ucsd.edu`
`math.ucsd.edu/~eaamari/`
AP&M 5880A

Today: Chapter 4 (end)

- Central limit theorem and consequences
- Conditions to have a normal sampling distribution of a mean statistic
- Introduction to inference for other parameters than the mean

Central Limit Theorem

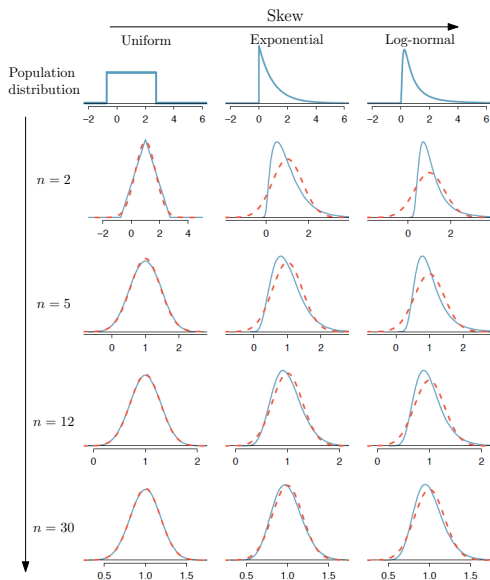


The Central Limit Theorem:

The distribution of \bar{x} is approximately normal. **This is true regardless of the population distribution, provided that n is large.**

The more skewed the population is, the larger n must be to ensure the sampling distribution is approximately normal.

Central Limit Theorem and Skew



As the sample size grows, the sampling distribution tends to look more and more normal. The greater the skew in the population, the higher n must be to get a normal sampling distribution.

General rule:

- If the population has no skew or moderate skew, $n \geq 30$ will generate roughly normal sampling distribution
- For extreme skew, $n \geq 60$ or $n \geq 100$ is needed.

Central Limit Theorem: Conditions

When doing statistical inference (C.I.'s or hypothesis testing), you rely on the SE of the sampling distribution and the assumption of normality for this curve. To use this, you must have:

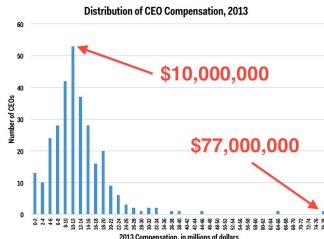
- Independence of data: Knowing one piece of data should not help you predict other data. Since this is hard to check, this is usually replaced with the:
 - Randomization condition (data were chosen randomly)
 - 10% condition ($n < 10\%$ of population size)
- Normal sampling distribution. You can either:
 - Check near-normality in the histogram of your sample.
 - If it fails, non-normal samples are OK as n gets larger (see previous slide).

Checking Conitions: Example

John wants to create a 95% C.I. for the mean salary of US CEO's. He randomly chooses 8 CEO's and finds the mean of their salaries. He builds his C.I. and feels satisfied. What is good/wrong with this approach?

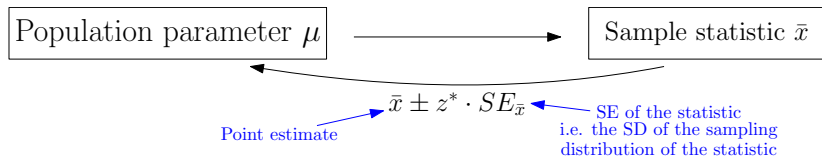
His CEO's salaries are independent because he meets the randomization condition and the 10% condition. (# CEO's > 80 in the US)

The sample size $n = 8$ is problematic if CEO salaries are somewhat skewed or have large outliers.



The sampling distribution of \bar{x} is probably not normal enough to do inference.

Stop Being Mean!



Other parameters we might estimate:

- Proportions p
(of UCSD students coming from California, say)
- Correlations ρ
(how correlated are Twitter followers and FB friend counts, say)
- Standard deviations σ
(spread in the number of failure of a device in a year, say)

Confidence Intervals for Other Estimators/Statistics

In general, we can calculate

$$\text{C.I.} = \text{point estimate} \pm z^* \cdot SE_{stat}.$$

- z^* : should behave as we learned if the sampling distribution is normal
- SE_{stat} : we have to learn it to do inference for other statistics.

Example: You draw a sample of 40 UCSD students to learn what percentage use the app Snapchat (...!). 22 students say they do.

If $SE_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$, find a 95% C.I. for the true parameter value.

Here, our point estimate (best guess) is $\hat{p} = \frac{22}{40} = 0.55$. Thus

$$SE_{\hat{p}} = \sqrt{\frac{0.55 \times 0.45}{40}} \simeq 0.079.$$

Confidence Intervals for Other Estimators/Statistics

For 95% confidence, we recall $z^* = 1.96$.

Our C.I. is $0.55 \pm 1.96 \times 0.079 = (0.396, 0.704)$.

We are 95% confident that the true percentage p of Snapchat users among UCSD students is between 39.6% and 70.4%.

Remark: Every parameter has a specific letter (μ, p), and there is a specific corresponding letter for the statistic (\bar{x}, \hat{p}).

Pay attention to these to know if you are thinking about the population or sample.

Beers

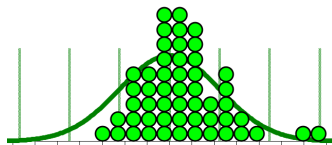
Do men and women of the same weight (say, 180 lbs) have different blood alcohol contents (BAC's) when they drink the same number of beers (say, 4)?

You randomly pick 10 UCSD men of weight 180 and give each 4 beers. You calculate the BAC of each and average the results to get $\bar{x}_M = 0.13$. You do the same for women and get $\bar{x}_F = 0.15$.

Population parameter: We really care about $\mu_F - \mu_M$, the difference in average BAC's of 180 lb men/women who drink 4 beers.

Sample statistic: (point estimate) $\bar{x}_F - \bar{x}_M$, the difference in BAC's in our sample. Here we get $0.15 - 0.13 = 0.02$ as our estimate.

Beers




Our sample difference of 0.02 is only one difference of means in a sampling distribution that has all the possible differences in means we would get from all the samples that could be taken.

Here, the sampling distribution for the idea $\bar{x}_F - \bar{x}_M$ is normal under modest assumptions. We later learn its SE, which we assume to be 0.01 in this problem

Hence, our 95% C.I. is $0.02 \pm 1.96 \cdot 0.01 = (0.0004, 0.0396)$

We are 95% confidence that $\mu_F - \mu_M$ is in this interval. Said differently, we are 95% confident that 180 lb women will have a BAC that is between 0.0004 and 0.0396 higher than 180 lb men after 4 beers.

Beers

<div></div> <div>BLOOD ALCOHOL CONTENT (BAC) Table for Male (M) / Female (F)</div>										
Number of Drinks		Body Weight in Pounds								Driving Condition
		100	120	140	160	180	200	220	240	
0	M	.00	.00	.00	.00	.00	.00	.00	.00	Only Safe Driving Limit
	F	.00	.00	.00	.00	.00	.00	.00	.00	
1	M	.06	.05	.04	.04	.03	.03	.03	.02	Driving Skills Impaired
	F	.07	.06	.05	.04	.04	.03	.03	.03	
2	M	.12	.10	.09	.07	.07	.06	.05	.05	
	F	.13	.11	.09	.08	.07	.07	.06	.06	
3	M	.18	.15	.13	.11	.10	.09	.08	.07	Legally Intoxicated
	F	.20	.17	.14	.12	.11	.10	.09	.08	
4	M	.24	.20	.17	.15	.13	.12	.11	.10	
	F	.26	.22	.19	.17	.15	.13	.12	.11	
5	M	.30	.25	.21	.19	.17	.15	.14	.12	
	F	.33	.28	.24	.21	.18	.17	.15	.14	
<div>Subtract .01% for each 40 minutes of drinking.</div> <div>1 drink = 1.5 oz. 80 proof liquor, 12 oz. 5% beer, or 5 oz. 12% wine.</div> <div>Fewer than 5 persons out of 100 will exceed these values.</div>										

Notice “Fewer than 5 persons out of 100 will exceed these values”.

Am I Normal?

Does every sampling distribution turn out to be approximately normal? **No!**

The Central Limit Theorem tells us only that the sampling distribution for \bar{x} is normal (assuming n is large enough and the population is not that skewed).

What about the sampling distribution for

Statistic	\hat{p}	$\bar{x}_1 - \bar{x}_2$	s_x^2
Normal for $n \gg 1$?	Yes	Yes	No

Situations where the normal does not approximate the sampling distribution:

- Small sample sizes for sampling distributions that usually do have a normal approximation
- Statistics whose sampling distribution truly are not normal (e.g. s_x^2)

Hypothesis Testing for Other Statistics

1) Define the parameter you will study

Let μ be the average stress level of students having dog therapy.

A different type of parameter would be needed here: p, σ, ρ, \dots

2) State H_0 and H_A :

H_0 : $\mu = 81$, and H_A : $\mu < 81$.

The parameter from 1) appears here, but the structure is similar

3) Find the model for the sampling distribution

Assuming H_0 , $\bar{X}_n \simeq N(\mu, \sigma/\sqrt{n}) = \dots$

The sampling distribution for a mean is often normal. Other statistics may have a non-normal sampling distribution

4) Draw a picture showing the sampling distribution. Shade the area for the p -value.

Once we know the shape from 3), we can draw the picture

5) Find the area using technology or table and conclude

Different tables or R function needed here, but same method.

Your Turn!

The weight of quarters in the US is normally distributed with a mean of 5.67g and SD of 0.07g.

If we set $X = N(5.67, 0.07)$, what does X represent?

1. The probability of getting a quarter of a particular weight
2. The weight of a random quarter we find on the street
3. The average weight of a quarter
4. The sampling distribution of quarter average weights

Answer: 2.

Your Turn!

Recall that $X = N(5.67, 0.07)$. What does $P(X \leq 5.67)$ equal?

1. The probability that a random quarter weighs less than 5.67g
2. The percentage of quarters with weights less than 5.67g
3. 0.5
4. The probability that the quarter in my pocket weighs less than 5.67g

Answer: 1.,2.,3.

Your Turn!

You pick 5 random quarters and average their weights. Which random variable Y represents what you have done? Recall that $X = N(5.67, 0.07)$.

1. $Y = \frac{5X}{5}$

2. $Y = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$

3. $Y = N(5.67, 5 \times 0.07)$

4. $Y = N(5.67, 0.07/5)$

Answer: 2. (The correct version of 3. and 4. is $Y = N(5.67, 0.07/\sqrt{5})$)

Your Turn!

Which is smaller? $P(X < 5)$ or $P(Y < 5)$?

1. $P(X < 5)$
2. $P(Y < 5)$
3. They are equal
4. Cannot be determined

Answer: 2., since $E(X) = E(Y) = 5.67$ but

$$SD(X) = 0.07 > 0.07/\sqrt{5} = SD(Y).$$

Said differently: it is hard to get an average of stuff be far from the mean.

Your Turn!

Which of these histograms would have the smallest spread?

1. The weights of 100 randoms quarters
2. The weights of 10,000 random quarters
3. The average weights of sets of 10 quarters
4. The average weights of sets of 100 quarters

Answer: 4.

1. and 2. correspond to population distribution.
3. and 4. are sampling distributions, and the larger n , the smaller the standard error is.

Your Turn!

Suppose, instead, that the weights of quarters is very skewed to the right. Which of the following ideas will have a histogram that is approximately normal?

1. The weights of 100 randoms quarters
2. The weights of 10,000 random quarters
3. The average weights of sets of 10 quarters
4. The average weights of sets of 100 quarters

Answer: 4.

If the sampling distribution is skewed, the central limit theorem applies for $n \geq 60$ or $n \geq 100$.