Central Limit Theorem

General Idea: Regardless of the population distribution model, as the sample size increases, the **sample mean** tends to be normally distributed around the population mean, and its standard deviation shrinks as n increases.

Certain conditions must be met to use the CLT.

- The samples must be independent
- The sample size must be “big enough”

CLT Conditions

Independent Samples Test

- **“Randomization”:** Each sample should represent a random sample from the population, or at least follow the population distribution.
- **“10% Rule”:** The sample size must not be bigger than 10% of the entire population.

Large Enough Sample Size

- Sample size n should be large enough so that $np \geq 10$ and $nq \geq 10$

Example: Is CLT appropriate?

It is believed that nearsightedness affects about 8% of all children. 194 incoming children have their eyesight tested. Can the CLT be used in this situation?

- Randomization: We have to assume there isn't some factor in the region that makes it more likely these kids have vision problems.
- 10% Rule: The population is “all children” - this is in the millions. 194 is less than 10% of the population.
- np=194*.08=15.52, nq=194*.92=176.48

We have to make one assumption when using the CLT in this situation.

Central Limit Theorem (Sample Mean)

- X_1, X_2, ..., X_n are *n* random variables that are independent and identically distributed with mean *μ* and standard deviation *σ*.
- $\overline{X} = (X_1+X_2+...+X_n)/n$ is the sample mean
- We can show $E(\overline{X})=\mu$ and $SD(\overline{X})=\sigma/\sqrt{n}$
- CLT states: $$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \rightarrow N(0,1)$$

  as n→∞

Implication of CLT

- We have: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \rightarrow N(0,1)$
- Which means $\bar{X} \rightarrow N(\mu, \sigma^2/n)$
- So the sample mean can be approximated with a normal random variable with mean *μ* and standard deviation *σ√n.*

Proportions of a Sample

Let's say we have a population with probability *p* of a certain characteristic (and *q*=1-*p*). We have a random sample of *n* from the population. What is the **mean** and **standard deviation** of the proportion of our sample that has the characteristic?

- We can use the CLT if n is large enough
- If X is the number of times the characteristic is found in our sample, **p**=X/n, our sample proportion, has mean **p** and standard deviation **$\sqrt{(pq/n)}$**

Central Limit Theorem (Sample Sum)

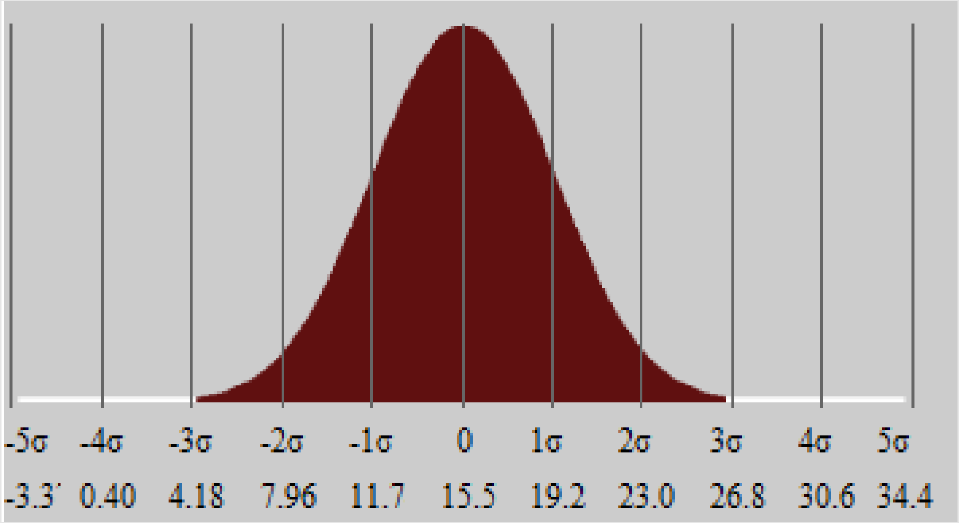
- X_1, X_2, ..., X_n are *n* random variables that are independent and identically distributed with mean *μ* and standard deviation *σ*.
- $S_n = X_1+X_2+...+X_n$ is the sample sum
- We can show $E(S_n)=n\mu$ and $SD(S_n)=\sigma\sqrt{n}$
- CLT states: $$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to N(0,1)$$

  as n→∞

Applications of CLT

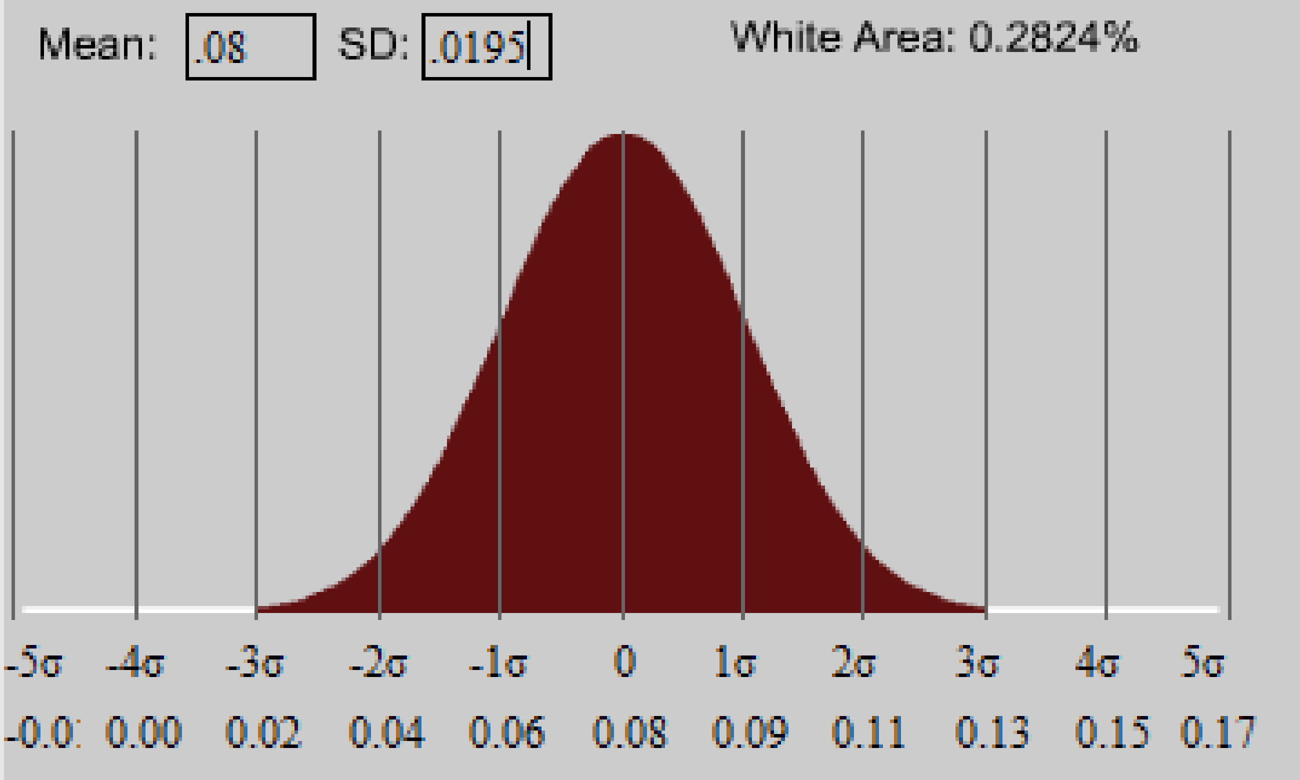
It is believed that nearsightedness affects about 8% of all children. 194 incoming children have their eyesight tested. Can the CLT be used in this situation?

- $\overline{X}$ should be approximately normally distributed have a mean of .08*194=15.52 and SD of √(.08*.92*194)=3.7787



CLT for Proportions

- How is the proportion of nearsighted children distributed?
- Divide by n=194: mean is .08, SD is .0195


Mean: .08 SD: .0195 White Area: 0.2824%

Applying the 68-95-99.7 Rule

- You can be 68% sure the sample mean is within 1 standard deviation (of the population mean)
- You are 95% sure the sample mean is within 2 standard deviations
- You are 99.7% sure the sample mean is within 3 standard deviations.
- **To cover virtually all possibilities, we can go 3 standard deviations from the sample mean.**

Example: Nearsightedness (cont)

With 192 incoming children, what is a reasonable range of nearsighted children the school can expect?

- Because 3 standard deviations covers 99.7% of the data, we use this for 'reasonable'.
- 3 standard deviations is 11.3361.
- 15.52-11.361=4.1839, 15.52+11.361=26.881
- The school should expect between 4 and 27 nearsighted children.

Example: College Retention Rates

Nationally 74% of college freshman continue as sophomores. A particular school has 486 out of all 600 freshmen stay. Is this unusual?

- We consider “unusual” anything above or below 3 standard deviations from the mean.
- We expect the retention to be $.74*600=444$, with a standard deviation of $\sqrt{npq}=\sqrt{(600*.74*.26)}=10.744$
- $\mu+3\sigma=444+3*10.744=476.232$, so 486 is “unusual”

Example: Pregnancy Length

Human pregnancies follow a normal distribution with mean of 268 days and s.d. 11 days. We study the mean pregnancy length of 70 women (call this random variable $\overline{X}$). What is this statistic's Expected Value and S.D.?

- It is reasonable to use the CLT (conditions are met)
- $\overline{X}$ is large enough to approximate with a normal distribution, $E(\overline{X})=268$, $SD(\overline{X})=11/\sqrt{70}=1.314$

Example: Dice Game

Awards for a dice game are as follows:

• Roll and Odd number:	$0
• Roll a 2 or a 4:	$2
• Roll a 6:	$26

- $E(X)=\$5$, $SD(X)=\sqrt{[E(X^2)-5^2]}=9.43$

(We already know how to find these)

Example: Dice Game (cont.)

- If you play the dice game 30 times, what is the expected value and standard deviation of your winnings? What is the probability you win at least $200?
- Let $S_{30}=X_1+X_2+...+X_{30}$
- $E(S_{30})=30*\$5=\$150$
- $SD(S_{30})=9.43*\sqrt{30}=51.65$
- $P(S_{30}\geq 200)$=normalcdf(200, 1000, 150, 51.65) =.1665

Example: AP Scores

- A Teacher has a class of 68 students taking the AP Physics test. Assuming they are typical of the population, the result of whose scores are given, what is the probability the average score will be at least 3?

Score	1	2	3	4	5
Probability	21.5%	18.8%	24.7%	21.7%	13.3%

Example: AP Scores (cont.)

Score	1	2	3	4	5
Probability	21.5%	18.8%	24.7%	21.7%	13.3%

- If X_1 is the score of a student, $E(X_1)=2.865$, $SD(X_1)=1.334$
- If $\overline{X}$ is the average of all 68 students, CLT states $E(\overline{X})=2.865$ and $SD(\overline{X})=1.334/\sqrt{68}=.162$
- $P(\overline{X}\geq 3)=\text{normalcdf}(3, 5, 2.865,.162)=.202$

Example: Restaurant Tips

A waitress earns varying tips with a mean of $10.90 and sd of $5.60. Assume on the weekend she gets 60 tips.

- What is the probability she earns $750 or more?
- Oh the best 5% of weekends, she earns at least how much $$? (i.e. what does she make less than on 95% of weekends)?

Example: Restaurant Tips (cont)

- We can use CLT to answer the problems.
- Let S_{60} be the sum of 60 tips following this model. $E(S_{60})=60*10.90 = \$654$, $SD(S_{60})=5.60*\sqrt{60}=\$43.3774$
- $P(S_{60} \geq 750)=$ normalcdf(750, 1000, 654, 43.3774)=.0134
- Upper 5th percentile is invNorm(.95, 654, 43.3774)=$725.35

Sums of Independent Normal r.v.s

- If you have two random variables, each following a normal distribution:
- $X \sim N(\mu_X, \sigma^2_X)$ and $Y \sim N(\mu_Y, \sigma^2_Y)$
- Then Let W=X+Y
- $W \sim N(\mu_X + \mu_Y, \sigma^2_X + \sigma^2_Y)$
- So the sum of two indpt. Normal random variables is Normal, its mean is the sum of their means and its variance is the sum of their variances.
- **Note: its sd is NOT the sum of their sds!!**

Example: Comparing IQ Scores

IQ Scores are normally distributed. Students at University A have IQ scores mean 130 and sd 6. At University B the mean is 120 and sd is 9. If you compare a random student from each university, what is the probability the IQ score from the University A student will be higher by 5 points or more?

- IQ Scores from University A ~ μ_A=130 σ_A=6
- IQ Scores from University B ~ μ_B=120 σ_B=9

Example: IQ Scores (cont)

- Let A and B be the samples from these populations. Let's define X=A-B, the difference of their scores.
- $E(X)=E(A-B)=130-120=10$
- $SD(X)=\sqrt{Var(X)}=\sqrt{(Var(A)+Var(B))}$ $=\sqrt{(6^2+9^2)}=10.8167$
- So X ~ Normal with $\mu_X=10$ $\sigma_X=10.8167$
- $P(X\geq 5)=\text{normalcdf}(5, 100, 10, 10.8167)$ $=.678$

Example: IQ Scores (cont)

Recall B~Normal with μ_B=120 σ_B=9. What is the probability the average of 3 University B students' IQs is at least 125?

- If B_1, B_2 and B_3 are the IQ scores of 3 univ. B students, Let $\overline{B}=(B_1+B_2+B_3)/3$
- $E(\overline{B})=120$,
- $SD(\overline{B})=SD[\ (B_1+B_2+B_3)/3\ ]$
  $=SD(B_1+B_2+B_3)/3=\sqrt{3}*9/3=5.1962$
- $\overline{B}$~Normal $\mu_{\overline{B}}=120$ $\sigma_{\overline{B}}=5.1962$ so
  $P(\overline{B}\geq 125)$=normalcdf(125,200,120,5.1962)=.1680

Example: IQ Scores (cont)

Say we want to compare the average of 3 random Univ. A students' IQ scores with the average of 3 random students from Univ. B. What is the probability the Univ. A students' avg score is at least 5 points higher?

- First define $\overline{A}$ similar to $\overline{B}$ on the previous page, and find $\overline{A}$~Normal with $\mu_{\overline{A}}=130$ $\sigma_{\overline{A}}=3.4641$.
- Define a new $\overline{X}=\overline{A}-\overline{B}$. We know $\overline{X}$ will be normal, so we just need its mean and standard deviation.

Example: IQ Scores (cont)

- $E(\overline{X})=E(\overline{A}-\overline{B})=E(\overline{A})-E(\overline{B})=130-120=10$
- $SD(\overline{X})=SD(\overline{A}-\overline{B})=\sqrt{}Var(\overline{A}-\overline{B}) =\sqrt{}[Var(\overline{A})+Var(\overline{B})] =\sqrt{}(3.4641^2+5.1962^2) =6.245$
- So the probability the A students' avg IQ is at least 5 points higher than the B students is the same as the probability $\overline{X}\geq 5$.
- $Pr(\overline{X}\geq 5)$=normalcdf(5, 100, 10, 6.245)