🧭 *Homework Show-Off: "Prove You're Data-Fluent!"*

Nam N Lai - Data 101- October 19 2025

Dataset: Students_Social_Media_Addiction.csv (N = 705)

This report applies four inferential frameworks—z-test & permutation test, confidence interval, chi-square test of independence, and Bayesian reasoning (odds form), all using a **95% confidence level (α = 0.05)**

1) Hypothesis Testing — z-test + Permutation Test (one-tailed):

***Undergrads Are More Addicted—and It's Not a Fluke!* (Δ = +0.250; z = 2.083; p(one-tailed) = 0.0186; perm p(one-tailed) = 0.0186)**

Two group: Academic_Level = 'Undergraduate' and when Academic_Level = 'Graduate' for
testing of Addicted_Score

H_0: The average value of Addicted_Score is the same when Academic_Level = 'Undergraduate' and when Academic_Level = 'Graduate'

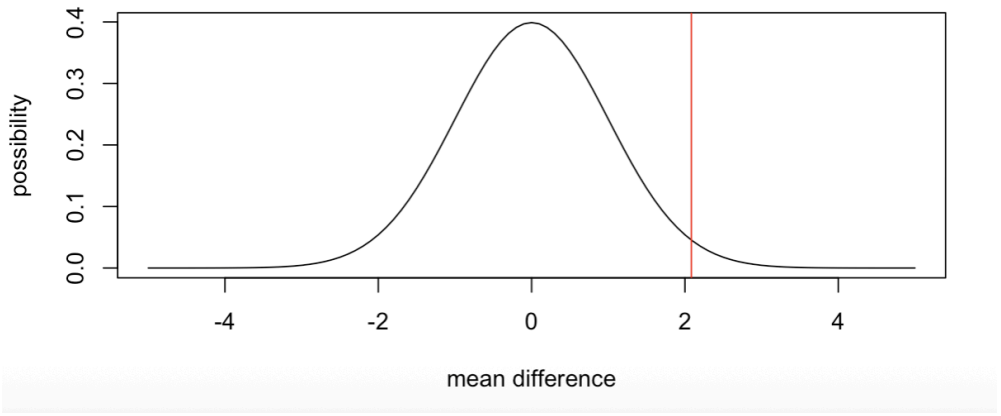
H_1 : The average value of Addicted_Score is greater when Academic_Level = 'Undergraduate' than when Academic_Level = 'Graduate'

Code R:

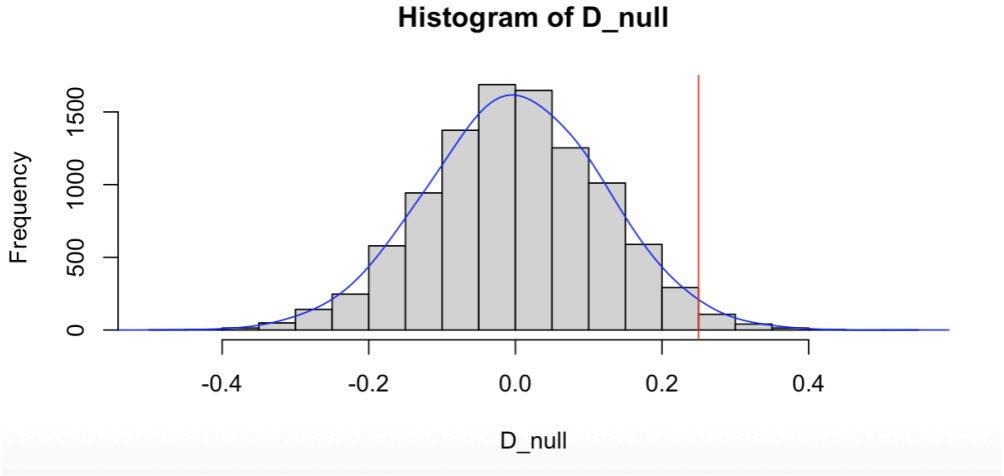
```
m_U <- mean(df$Addicted_Score[df$Academic_Level=="Undergraduate"],
na.rm=TRUE)
m_G <- mean(df$Addicted_Score[df$Academic_Level=="Graduate"],
na.rm=TRUE)
obs_diff <- m_U - m_G
obs_diff

# 3) Z-test (one-tailed): test U - G > 0 ⇒ sub1="Graduate", sub2="Undergraduate"
p_z <- z_test_from_data(df, "Academic_Level", "Addicted_Score",
            "Graduate", "Undergraduate")

# 4) Permutation test (one-tailed, N>=10000)

p_perm <- permutation_test(df, "Academic_Level", "Addicted_Score",
                           10000, "Graduate", "Undergraduate")

cat(p_perm)
```

My result:

- Observed difference of Addicted_Score's mean: **0.2498409** (Undergrad − Graduate)
- z-value: **2.083**; p(one-tailed) = **0.0186**



- Permutation p(one-tailed) = **0.0184 (10,000 permutations)**



-> Observed difference = 0.2498 in the right tail of distribution

Interpretation: There is evidence to show Undergraduate students scored 0.25 points higher on the social media addiction scale compared to graduate students. Both the z-test (p = 0.0186) and permutation test (p = 0.0184) produced p-values well below 0.05, leading us to reject the null hypothesis at the 95% confidence level. The convergence of both methods confirms this is not random noise—the effect is statistically credible. While the "Observed difference" is modest, it reveals a meaningful pattern: younger, less experienced students appear more vulnerable to social media addiction

Headline Justification: There is 0.25-point difference in addiction scores (undergraduate's mean addicted score higher 0.25-point than graduate) with consistent statistical significance across two independent tests (both p ≈ 0.019 < 0.05). The phrase "not a fluke" is justified because the permutation test, which makes no distributional assumptions independently confirms the z-test result, demonstrating the effect is robust and unlikely due to chance

Code implements the test:

- read.csv(...) loads the data to df.
- df$Academic_Level=="Undergraduate" creates a logical mask; the bracket [...] filters rows to compute m_U for that group (same for m_G).
- obs_diff <- m_U - m_G is the sample effect size
- z_test_from_data(...) computes a one-tailed z-test for H_1: μ_Undergrad > μ_Grad, and prints the z and p.
- permutation_test(...) simulates the null by reshuffling labels 10,000 times and estimating the p-value as the fraction of shuffled differences ≥ the observed difference.

2) Confidence Interval — Instagram primary users:

Instagram hits Users Average 4.87 hours/day—and We're Precise: 95% CI [4.719, 5.026] (±0.154 h)

Target quantity: `Avg_Daily_Usage_Hours` among students whose Most_Used_Platform == "Instagram"

Code R:

```
#Confidence Interval for Avg_Daily_Usage_Hours of student most used Instagram
z_score <- 1.96

# 2) Subset: Instagram users
ins <- subset(df, Most_Used_Platform == "Instagram" &
!is.na(Avg_Daily_Usage_Hours))

# 3) Sample size, mean, and standard deviation
n  <- nrow(ins)
m  <- mean(ins$Avg_Daily_Usage_Hours)
s  <- sd(ins$Avg_Daily_Usage_Hours)

# 4) Standard error of the mean (SEM)

sem <- s / (n^(1/2))
# 5) Margin of Error and CI
MOE <- z_score * sem
LCL <- m - MOE
UCL <- m + MOE
round(c(n = n, mean = m, sd = s, SEM = sem, MOE = MOE, LCL = LCL, UCL =
UCL), 3)
```

My result:

- **Sample size:** n = 249
- **Sample mean:** 4.872 hours/day
- **Standard deviation:** 1.238 hours
- **Standard error (SEM):** 0.078 hours
- **Margin of error (MOE):** 0.154 hours
- **95% CI:** [4.719, 5.026] hours/day

Interpretation: If we repeated this study many times with different random samples of 249 Instagram users, approximately 95% of the resulting confidence intervals would capture the true population mean daily usage. Our interval [4.72, 5.026] hours is quite narrow (width ≈ 0.307 hours), suggesting we can estimate the population mean with high precision. Based on our sample of 249 students, we are 95% confident that the average daily usage for ALL Instagram-primary students in the population falls between 4.72 and 5.026 hours, with a margin of error of only ±0.15 hours

Headline Justification: The headline's precision claim is justified by the narrow 95% CI spanning only 0.307 hours (4.72 to 5.026). With a margin of error of just ±0.15 hours (about ±9 minutes), this interval demonstrates high stability of our estimate. The narrow range, which is driven by a robust sample size (n = 249), proves we can confidently place the true average daily usage instagram of students between 4.72 and 5.03 hours—a remarkably precise estimate with minimal uncertainty.

Code implements:

- subset(...) filters students whose Most_Used_Platform == "Instagram" and removes missing values in Avg_Daily_Usage_Hours, creating our sample (n = 249).
- nrow(ins) counts the sample size.
- mean(...) and sd(...) compute the sample mean ($\bar{m}$) and standard deviation (s).
- sem <- s / (n^(1/2)) calculates the standard error of the mean.
- z_score <- 1.96 is the critical value for 95% confidence under the normal distribution; MOE <- 1.96 * sem computes the margin of error.
- LCL and UCL define the 95% confidence interval endpoints [$\bar{m}$ - MOE, $\bar{m}$ + MOE].
- round(c(...), 3) displays all statistics rounded to 3 decimal places.

3) Hypothesis of Independence — Chi-Square Test of Independence

Academic Level and Performance Impact Are Linked—High Schoolers Hit Hardest! ($\chi^2 = 10.79$, df = 2, p = 0.0045)

Two categorical variables:

- Academic_Level: Graduate, High School, Undergraduate
- Affects_Academic_Performance: Yes, No

H_0: Academic_Level and Affects_Academic_Performance are independent

H_1 : Academic_Level and Affects_Academic_Performance are independent are associated

Code R:

```
cont_ac <- table(df$Academic_Level, df$Affects_Academic_Performance)
cont_ac

chisq_ac <- chisq.test(cont_ac)
chisq_ac

# Interpretation:
# H0: Academic_Level and Affects_Academic_Performance are independent
# H1: They are associated
# Decision: if p-value < 0.05 → reject H0 (association exists)

chisq_ac$expected
chisq_ac$observed
```

My result:

- Independence (Null Hypothesis): expected of Academic_Level and Affects_Academic_Performance

	No	Yes
Graduate	116.170213	208.82979
High School	9.651064	17.34894
Undergraduate	126.178723	226.82128

- Observation:

	No	Yes
Graduate	126	199
High School	2	25
Undergraduate	124	229

- χ^2 = **10.793**
- df (degrees of freedom) = **2**
- p = **0.004533**

Interpretation: The chi-square test reveals a statistically significant association between academic level and perceived impact on academic performance (χ^2 = 10.79, df = 2, p = 0.005). We reject the null hypothesis of independence at the 95% confidence level (p ≈ 0.0045 < 0.05). Examining the observed vs. expected counts shows the key driver: **high school students report being affected far more than expected** (observed 25 "Yes" vs. expected 17.35), while they say "No" far less (observed 2 vs. expected 9.65)
-> **Note**: Although the high school sample is modest (n=27), all expected counts exceed the chi-square threshold of 5 (minimum = 9.65), validating the test. The stark deviation (93% vs 64%) is robust enough to drive significance despite the smaller cell, suggesting a genuine developmental pattern rather than statistical artifact

Headline Justification: The headline is justified by the significant chi-square result (χ^2 = 10.79, p = 0.005 < 0.01), confirming academic level and performance impact are not independent. The phrase "High Schoolers Hit Hardest" is supported by the stark deviation in their response pattern: 25 out of 27 high school students (92.6%) report social media affects their academics—far exceeding the expected proportion under independence. This demonstrates a clear, statistically credible link between academic experience level and perceived social media impact

Code implements:

- table(df$Academic_Level, df$Affects_Academic_Performance) builds a contingency table of observed counts for each Academic_Level × Affects_Academic_Performance combination. This is the input to the chi-square test
- chisq.test(cont_ac) performs Pearson's chi-square test of independence. Internally it computes: **$\chi^2 = \Sigma [(O - E)^2 / E]$** where **O** are observed counts and **E** are expected counts under H_0 (independence). Expected counts are calculated: **E_ij = (row_i total × column_j total) / grand total**
- **Degrees of freedom** are **(R - 1)(C - 1)**, where R = number of rows (academic levels) and C = number of columns (impact categories). With 3 academic levels and 2 impact categories: **df = (3 - 1)(2 - 1) = 2**
- chisq_ac$expected prints the expected counts under independence
- chisq_ac$observed returns the observed table for easy comparison

4) Bayesian Reasoning — Posterior via Odds Form

Seeing TikTok? Bet on Academic Impact: 64% Prior Jumps to 94% Posterior (LR = 8.01)

Belief (prior) **:** Affects_Academic_Performance == "Yes"
Observation (evidence) **:** Most_Used_Platform == "TikTok"
Prior Probability: Before knowing which platform, 64.3% of students say social media affects their academics
True Positive: Among students who are affected, 31.8% use TikTok P(TikTok | Affects = "Yes")
False Positive:Among students who are not affected, only 4.0% use TikTok P(TikTok | Affects = "No")
Likelihood Ratio: The ratio of how many times more likely TikTok usage appears among affected students compared to unaffected students (LR = TP / FP)
Posterior Probability: if a student uses TikTok, probability they are affected academically P(Affects | TikTok)

R code:

```
# Step 1: Prior Probability and Odds
n_affects <- sum(df$Affects_Academic_Performance == "Yes", na.rm = TRUE)
n_total   <- sum(!is.na(df$Affects_Academic_Performance))
PriorP <- n_affects / n_total
#PriorP <- mean(df$Affects_Academic_Performance == "Yes", na.rm = TRUE)
PriorOdds <- PriorP / (1 - PriorP)
cat("Prior Probability:", round(PriorP, 3), "\n")
cat("Prior Odds:", round(PriorOdds, 3), "\n")

# Step 2: Observation (Most_Used_Platform == "TikTok")
obs <- df$Most_Used_Platform == "TikTok"

# Step 3: True Positive and False Positive
TP <- sum(df$Most_Used_Platform == "TikTok" &
df$Affects_Academic_Performance == "Yes", na.rm = TRUE) /
  sum(df$Affects_Academic_Performance == "Yes", na.rm = TRUE)
cat("True Positive:", round(TP, 3), "\n")

FP <- sum(df$Most_Used_Platform == "TikTok" &
df$Affects_Academic_Performance == "No", na.rm = TRUE) /
  sum(df$Affects_Academic_Performance == "No", na.rm = TRUE)
cat("False Positive:", round(FP, 3), "\n")

cat("TP:", round(TP, 3), " FP:", round(FP, 3), "\n")

# Step 4: Likelihood Ratio and Posterior
LR <- TP / FP
PostOdds <- LR * PriorOdds
cat(PostOdds)
```

```
PostP <- PostOdds / (1 + PostOdds)
cat("Likelihood Ratio:", round(LR, 2), "\n")
cat("Posterior Probability (Affects | TikTok):", round(PostP * 100, 2), "%\n")
```

My result:

- Prior Probability: 0.643 -> Prior Odds: 1.798
- True Positive: 0.318
- False Positive: 0.04
- Likelihood Ratio: 8.01
- Posterior Probability (Affects | TikTok): 93.51 % -> Posterior Odds: 14.4

Interpretation: Before knowing which platform a student uses, we estimated 64.3% of students experience academic impact from social media (prior). However, observing that a student primarily uses TikTok dramatically updates this belief to 93.5% (posterior). The likelihood ratio of 8.01 reveals that TikTok usage is 8 times more common among affected students than unaffected ones, this is strong Bayesian evidence (LR > 5 indicates substantial support). The Bayesian update transforms our prior odds of 1.8 : 1 into posterior odds of 14.4 : 1, meaning if we encounter a high-used TikTok student, we should be highly confident (>93%) they experience academic consequences

Headline Justification: The headline is justified by the dramatic probability shift from 64.3% (prior) to 93.51 % (posterior), leading to a 45.4% relative increase driven by observing TikTok usage. The likelihood ratio of 8.01 provides strong Bayesian evidence that TikTok is highly diagnostic of academic impact. The phrase "Bet on Impact" reflects the posterior odds of 14.4:1—if you encounter a TikTok user as a student, the data suggests betting 14:1 odds they report being affected. This isn't speculation; it's a rigorous Bayesian update showing TikTok usage as a powerful predictor of academic disruption

Code implements:

Step 1: Prior Probability and Odds

- n_affects <- sum(df$Affects_Academic_Performance == "Yes", na.rm = TRUE) creates a logical vector (TRUE for "Yes", FALSE otherwise) and sum() counts the number of students who report being affected, excluding missing values
- n_total <- sum(!is.na(df$Affects_Academic_Performance)) counts all students with non-missing responses by summing TRUE values from !is.na(...).
- PriorP <- n_affects / n_total computes the **prior probability** P(Affects = "Yes") as the proportion of affected students in the entire dataset (baseline rate before observing platform)

- PriorOdds <- PriorP / (1 - PriorP) converts prior probability to **odds form**: Odds = P/(1-P). Example: 0.643/(1-0.643) = 1.798, meaning "Yes" is 1.798 times more likely than "No" before seeing evidence

Step 2: Define Observation

- obs <- df$Most_Used_Platform == "TikTok" creates a logical vector identifying TikTok users (TRUE = TikTok, FALSE = other platforms). This defines the evidence we're conditioning on, though this variable isn't directly used in subsequent calculations

Step 3: True Positive and False Positive

- TP <- sum(df$Most_Used_Platform == "TikTok" & df$Affects_Academic_Performance == "Yes", na.rm = TRUE) / sum(df$Affects_Academic_Performance == "Yes", na.rm = TRUE) computes **True Positive** = P(TikTok | Affects = "Yes"). The numerator counts students who both use TikTok AND are affected (using & logical AND); the denominator is total affected students.
- FP <- sum(df$Most_Used_Platform == "TikTok" & df$Affects_Academic_Performance == "No", na.rm = TRUE) / sum(df$Affects_Academic_Performance == "No", na.rm = TRUE) computes **False Positive** = P(TikTok | Affects = "No"). The numerator counts students who use TikTok BUT are not affected; denominator is total unaffected students

Step 4: Likelihood Ratio and Posterior Update

- LR <- TP / FP computes the **Likelihood Ratio** = P(TikTok | Yes) / P(TikTok | No), quantifying how much more diagnostic TikTok usage is for affected vs. unaffected students. Example: 0.318 / 0.040 = 8.01, meaning TikTok is 8× more common among affected students. LR > 5 indicates strong evidence.
- PostOdds <- LR * PriorOdds implements **Bayes' Theorem in odds form**: Posterior Odds = Likelihood Ratio × Prior Odds. This updates our belief by multiplying prior odds (1.798:1) by the strength of evidence (LR = 8.01), yielding posterior odds (14.40:1)
- PostP <- PostOdds / (1 + PostOdds) converts **posterior odds back to probability**: P = Odds / (1 + Odds)

What Did I Learn About My Data Story ?

- **The hypothesis tests (z-test and permutation) told me who is more vulnerable:** Undergraduate students score 0.25 points higher on addiction scales than graduate students ($p \approx 0.019$). This suggests that **younger, less academically experienced students lack the self-regulation skills or time management** that graduate students have developed. The convergence of

both parametric and non-parametric tests confirms this isn't a statistical fluke, which undergraduates genuinely struggle more with controlling their social media usage, possibly because they're still developing discipline and haven't yet experienced the high-stakes demands of advanced academic work

- **The confidence interval revealed how much time is being consumed:** Instagram users average 4.87 hours per day, with remarkable precision (95% CI: [4.72, 5.03]). This narrow interval,spanning just 18 minutes can show that **Instagram usage is remarkably consistent across students**, almost like a daily routine or habit. Nearly 5 hours a day represents roughly **1/3 of waking hours** devoted to a single platform, highlighting how deeply social media has become in student life. The stability of this estimate suggests this isn't occasional but habitual behavior
- **The chi-square test revealed high school students drive the entire association** ($\chi^2 = 10.79$, $p = 0.0045$): 92.6% report academic impact versus expected 64%—a 28-point deviation contributing ~87% of χ^2. Undergraduates and graduates show near-independence. **The association exists not because academic level universally predicts impact, but because high schoolers—facing peer pressure, identity struggles, and immature impulse control—represent a uniquely vulnerable developmental window** where social media's grip is tightest
- **Bayesian reasoning identified which platform is the danger zone:** TikTok emerged as a powerful diagnostic marker. if a student uses TikTok, our confidence they're affected jumps from 64% to 94% (LR = 8.01). This dramatic update reveals that **TikTok's algorithm-driven, short-form video format is particularly addictive and academically disruptive**. Unlike passive platforms, TikTok's endless scroll and personalized content create a dopamine loop that's hard to escape. The 8.01 likelihood ratio suggests TikTok isn't just another social media app, it's qualitatively different in its ability to capture attention and interfere with studying

-> Social media addiction among students follows a clear pattern: **younger students (especially high schoolers) are most vulnerable, Instagram consumes nearly 5 hours daily with clock-like regularity, and TikTok stands out as the most academically destructive platform**. From hypothesis testing through Bayesian inference, every statistical lens exposed a different dimension of the same underlying crisis: students are losing control of their time and attention to social media, and those least equipped to fight back, the youngest among them are suffering most