Advanced Statistics for Social Sciences 1

Statistical Thinking and Data Analysis with R

Instructor: Rachit Dubey

Email: rdubey@ucla.edu

Summary of last week

Normal Distribution

- Normal Distribution and sufficient statistics
- Standard normal distribution

Sampling

- Samples, populations and sampling
- Law of large numbers
- Sampling distributions
- Central Limit Theorem

Lecture 6: Hypothesis Testing

Today's Learning Goals

- Conduct and interpret a hypothesis test (t-test)
- Compute and interpret effect sizes
- Understand confidence intervals

The problem we are solving

We collected a sample. Now what?

Three interconnected questions:

1. "Is this result statistically significant?" → HYPOTHESIS TESTING
2. "How big is the effect?" → EFFECT SIZES
3. "What range captures the parameter?" → CONFIDENCE INTERVALS

What can statistics do for us?

DESCRIBE

- Summarize data
- Visualize patterns
- Estimate parameters

DECIDE

- Test hypotheses
- Compare groups
- Evaluate claims

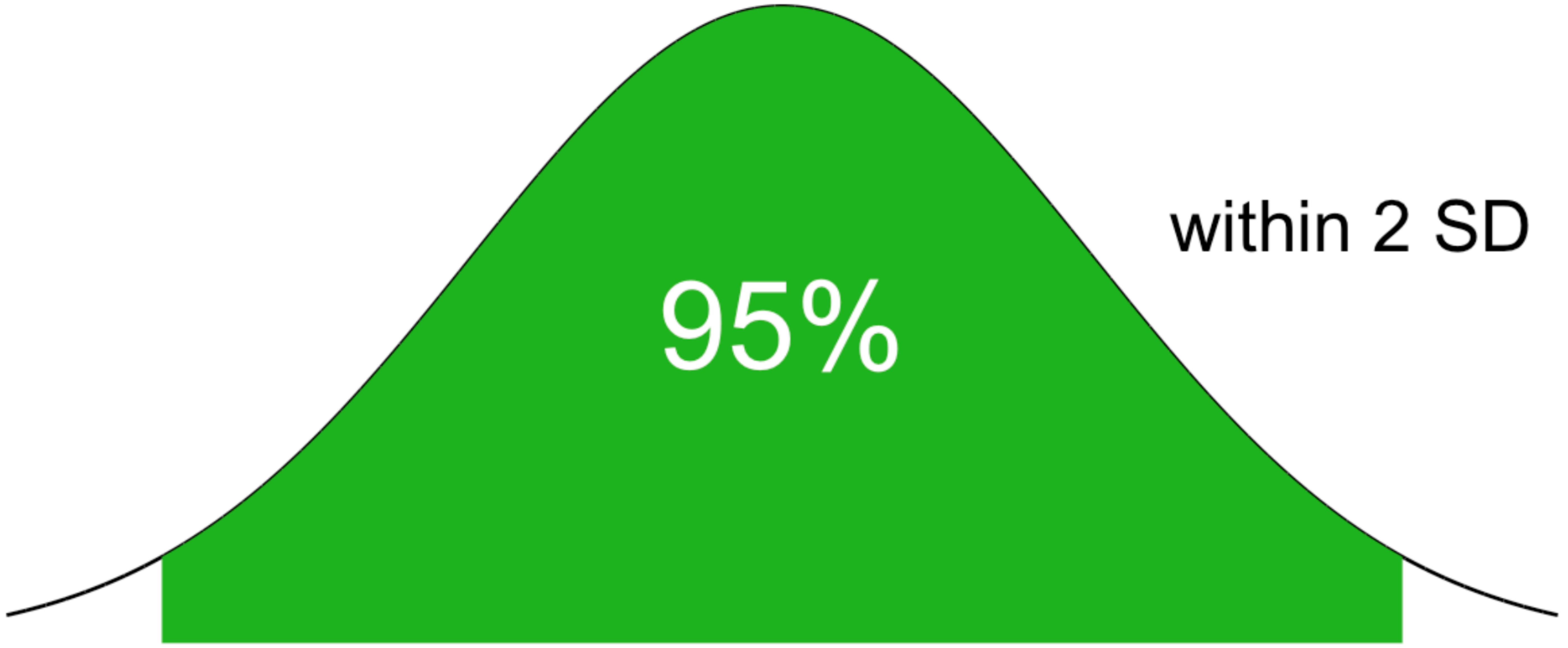
PREDICT

- Build models
- Forecast outcomes
- Classify new data

Key concepts to recap

The Normal Distribution and the ±2 Rule

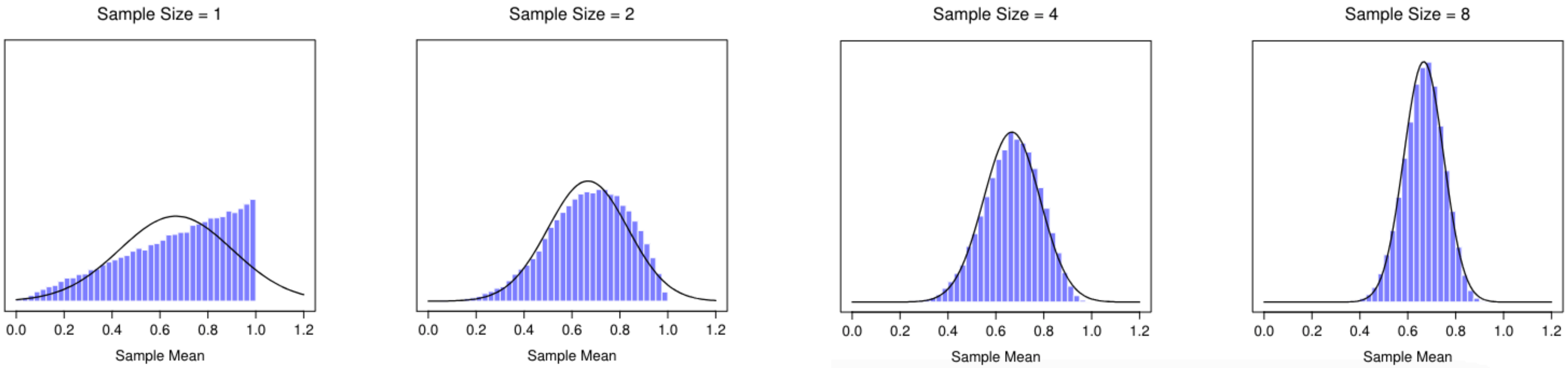
- About 95% of data fall within 2 standard errors of the mean
- This is important for both hypothesis testing AND confidence intervals



Key concepts to recap

Central Limit Theorem

- As long as your sample size isn't tiny, the sampling distribution of the mean will be approximately ***normal*** no matter what your population distribution looks like!



Part 1: Hypothesis Testing

Setting expectations

The theory of hypothesis testing is a mess, and even now there are arguments in statistics about how it 'should' work..

1. NHST has been criticized for 50+ years
2. Despite flaws, it's THE dominant approach in published science
    - Essential to understand for reading literature
    - Alternative approaches exist (Bayesian, effect-size based; more on this later)
3. Our goal: understand it well + use it responsibly
    - Avoid common pitfalls
    - Report complete information
    - Think about practical significance

Key idea

Assume null is true. Then ask: How unlikely are our data under that assumption?

If very unlikely → reject the null

Research hypothesis vs. Statistical hypothesis

RESEARCH HYPOTHESIS

- "Body cameras reduce police use of force"
- "Meditation reduces anxiety"
- "iOS users differ in personality"
- Testable, falsifiable
- About psychology/science

STATISTICAL HYPOTHESIS

- $\mu_{camera} < \mu_{no\_camera}$
- $\mu_{meditation} < \mu_{control}$
- $\mu_{iOS} \neq \mu_{Android}$
- About parameters
- Precise and mathematical

IMPORTANT: Research hypothesis must map ***clearly*** to statistical hypothesis!

Research hypothesis vs. Statistical hypothesis

Question: Is physical activity related to BMI?

Research hypothesis: "BMI is greater for people who don't engage in physical activity"

Key requirements:

✓ Specific and testable

✓ Formulated BEFORE seeing data

✓ Based on theory or prior research

✓ Falsifiable (could be wrong)

Examples of BAD hypotheses:

❌ "People are different" (too vague)

❌ "My data will be interesting" (not scientific)

What p-value is NOT

❌ **"p = probability that null is true"**

THIS IS THE MOST COMMON ERROR

$p \neq P(H_0|\text{data})$

Probability null is true given the data

$p = P(\text{data}|H_0)$

Probability of observing this data IF null were true

❌ **"p = probability I am making a mistake"**

This specific test isn't 5% wrong

In long run, 5% of true nulls rejected

✓ **p = rareness of data under null**

✓ **Smaller p = more evidence against null**

❌ **"Non-significant p means no effect exists"**

Could be Type II error

Maybe just lack of power

❌ **"If I run this study again, I'll get same p"**

Different sample = different p

p varies across replications

Another common misconception

Significant result means the result is important

EXAMPLE:

Study with N=10,000 people

Finding: Diet A produces 1 oz more weight loss than Diet B

$p < 0.05$; Statistically significant

BUT: 1 oz = weight of a few potato chips.. Practically? Who cares.

WHY IT MATTERS:

Big samples detect tiny effects

Statistical significance ≠ Practical significance

RIGHT WAY TO THINK:

Report p-value + effect size + confidence interval (more on this very soon)

Let data speak for itself, don't hide uncertainty

Is *p* rare enough to reject H_0

Convention: Set α (significance level) beforehand

α = 0.05 (most common)

α = 0.01 (more stringent; I recommend this!!)

α = 0.001 (very stringent)

Decision rule:

If $p < \alpha$: REJECT H_0 (result is "significant")

If $p \geq \alpha$: FAIL TO REJECT H_0 (result is "non-significant")

Example:

$p = 0.000072$

$\alpha = 0.01$

$0.000072 < 0.01 \rightarrow$ REJECT H_0

Conclusion: "We have sufficient evidence that BMI differs between active and inactive individuals"

Please remember

0.05 is ARBITRARY..what matters is being transparent!

Putting it together

Step 1: Hypothesis: "Does physical activity relate to BMI?"

Step 2: H_0: BMI_active = BMI_inactive; H_1: BMI_active ≠ BMI_inactive (two-sided)

Step 3: Data collection

Step 4: t = 3.86 (compute t-statistic)

Step 5: p = 0.000145 (two-tailed)

Step 6: Decision: $p < 0.01$ → REJECT H_0

How to report a hypothesis test

"We compared BMI between active vs. inactive using an independent samples t-test. Active individuals ($M = 27$, $SD = 5.2$) showed lower BMI than inactive individuals ($M = 30$, $SD = 9.0$).

This difference was statistically significant, $t(241) = 3.86$, $p < 0.001$."

Components to include:

✓ Test name (t-test, etc.)

✓ Sample sizes

✓ Means and SDs for each group

✓ Direction of difference

✓ Test statistic value

✓ Degrees of freedom

✓ p-value, effect size (more on this soon) and confidence intervals (more on this soon)

✓ What to conclude

A common trap — ALWAYS avoid!!

Significant vs. Not Significant ≠ Evidence of a Real Difference

- Group 1: 33/50 demonstrate X ability compared to control ($p = .03$) ✓ Significant
- Group 2: 29/50 demonstrate X ability compared to control ($p = .32$) ✗ Not Significant

<u>**We can’t say Group 1 differs from Group 2 using this test!!**</u>
We can only say Group 1 differs from control and Group 2 doesn’t differ from control.

A common trap — ALWAYS avoid!!

Significant vs. Not Significant ≠ Evidence of a Real Difference

- Group 1: 33/50 demonstrate X ability compared to control ($p = .03$) ✓ Significant
- Group 2: 29/50 demonstrate X ability compared to control ($p = .32$) ✗ Not Significant

<u>We can't say Group 1 differs from Group 2 using this test!!</u>
We can only say Group 1 differs from control and Group 2 doesn't differ from control.

Key Principle: To claim two groups differ, you must explicitly test that hypothesis. You cannot infer differences by comparing each group separately to chance and then observing different significance levels.

<u>Direct Comparison: Group 1 vs. Group 2 ($p = .54$) — No difference!</u>

Why this (could) have happened: Both groups performed at borderline levels. By chance alone, Group 1 landed just barely above the $p = .05$ threshold while Group 2 didn't. This is random variation, not evidence of true group differences.

Takeaway: Always be wary of this mistake—run a **<u>direct comparison test</u>** if you want to argue groups are different.