Human-Computer Interaction

Statistics II

Inferential + Regression

Professor Bilge Mutlu

Today's Agenda

- **From estimates to tests:** H_0 and H_1, *p*, two ways to be wrong
- **χ^2**: categorical associations
- **Student's t**: two means
- **ANOVA**: 3+ means (one-/two-way, within/between)
- **Regression**: correlation, linear & logistic
- **Assumptions:** checking them, and what to do when they fail
- In-class: *Pick the Test, Run One* → the Analysis Plan assignment

Two running examples

A • Stories

Stories written **alone**, with **one AI idea**, or with **five** (GPT-4).

Rated for **novelty** and **usefulness**, 1–9.[1]

B • Debates

Debates with a **human** or **GPT-4**, **personalized** or not.

Outcome: shift **toward the opponent**, 1–5.[2]

---

[1] Doshi, A. R. & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), eadn5290.

[2] Salvi, F., Horta Ribeiro, M., Gallotti, R. & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 9, 1645–1653.

From estimates to tests

Oct 30: how big? Today: could it be zero?

A · Stories: stories written with five AI ideas were rated **0.31 points** more novel than stories written alone, **95% CI [0.05, 0.58]**.

The test asks: suppose AI ideas did nothing. How often would chance alone produce a gap this big between two groups this size?

```
two$condition <- factor(two$condition, levels = c("ai_5_ideas", "human_only"))
t.test(novelty ~ condition, data = two)   # five AI ideas minus alone
# t = 2.333, df = 182.31, p-value = 0.02074
# 95 percent confidence interval: 0.048 0.578
```

Same data, same CI. The test adds a **p-value**: how surprising the gap would be if there were no real difference.

From research hypothesis to H_0 and H_1

Oct 9 form: *Y will be higher at level 1 of X than at level 2, under Z.*

A · Stories: *Stories written with five AI ideas will be rated more novel than stories written alone.*

- **Null, H_0:** $\mu_{5\ ideas} = \mu_{alone}$ (no difference in the population)
- **Alternative, H_1:** $\mu_{5\ ideas} \neq \mu_{alone}$ (a difference in **either direction**)

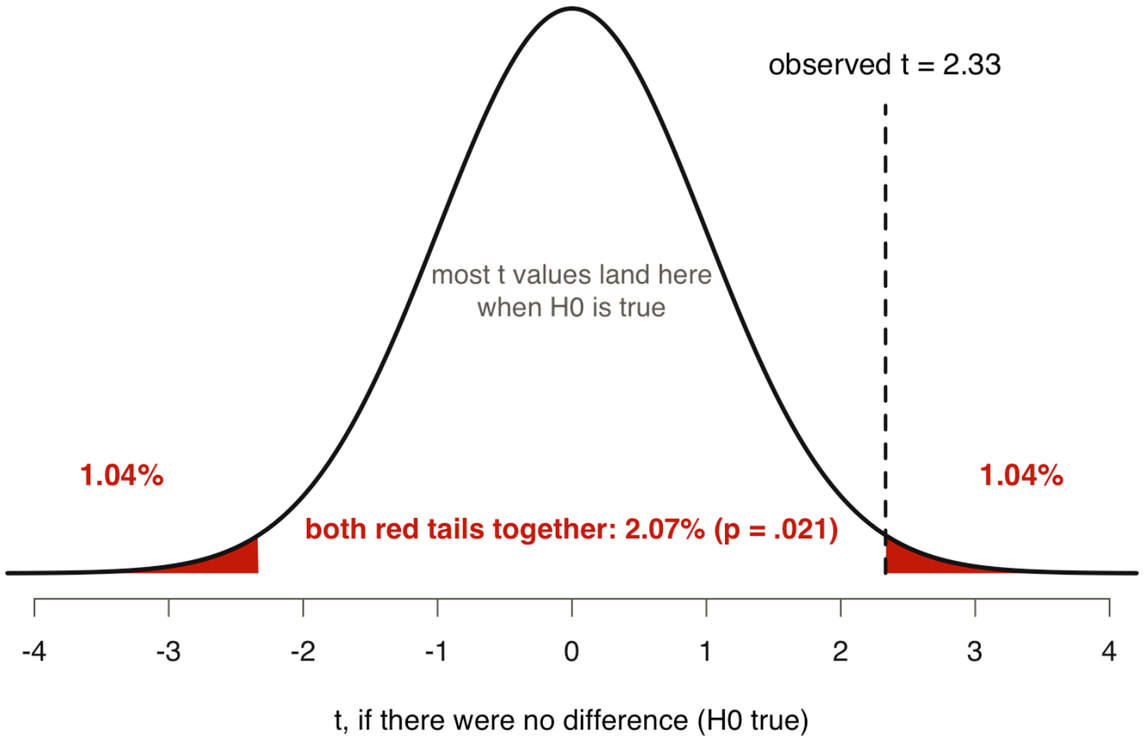
The hypothesis predicts a direction, but the standard test is **two-sided**: a gap either way counts against H_0.

The test is a test of H_0. We look for evidence against it; we never prove

The logic of a test

1. **Assume H_0** is true
2. **Signal / noise:**
$$t = \frac{difference}{standard\ error} = \frac{0.31}{0.134} = 2.33$$
3. **How surprising?** Under H_0, a *t* at least 2.33 away from 0, **in either direction** (two-sided), happens **2.07%** of the time: $p = .021$

p-value: the probability of data at least this extreme, *if H_0 were true*.



Deciding: α, and two ways to be wrong

$p \leq \alpha \rightarrow$ reject H_0 · $p > \alpha \rightarrow$ **fail to reject** H_0 (not "accept")

In HCI, α = .05, set before the data.

	H_0 true *(no real effect)*	**H_0 false *(real effect)***
Reject H_0	**Type I** error: false positive	Correct: effect detected
Fail to reject	Correct	**Type II** error: false negative

How often is each error?

- **α (alpha):** the Type I rate you accept. At **.05**, a 1-in-20 chance of a false positive when there is no effect
- **β (beta):** the Type II rate. The chance of missing an effect that is there
- **Power = 1 – β:** the chance of detecting an effect that is there. Aim for **.80**

You choose α. β depends on the effect size and the sample size: small samples, large β.

p and the CI tell the same story

- **They agree:** the 95% CI excludes 0 exactly when two-sided $p < .05$. Here: [0.05, 0.58], $p = .021$
- **The CI tells you more:** how big the effect is, and how precise
- **Significant does not mean important:** the CI's low end, 0.05, is a trivial gain on a 1–9 scale

Report estimate, CI, then test: *"0.31 points more novel, 95% CI [0.05, 0.58], t(182.3) = 2.33, p = .021."*

Common misreadings of p[3]

- *"p = .02: a 2% chance H_0 is true."* No: it is the chance of a result this extreme **if H_0 were true**, not the chance that H_0 is true
- *"A small p means a big effect."* No: with a large sample, tiny effects get small p's. Judge size by the effect size (*d*, *r*) and its CI
- *"p > .05 means no effect."* No: the study may be too small to detect it. If the CI still includes effects that would matter, the question is open
- *"p = .049 and p = .051 are different findings."* No: nearly the same

---

[3] Wasserstein, R. L. & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133.

[4] Rosnow, R. L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44(10), 1276–1284.

Recap: how do we choose a test?

- Determined by the *data types* of the **input** and **output** variables.
- Data types, from Oct 9: *Nominal, Ordinal, Interval, Ratio*
- Interval/ratio distributions can be *normal* or *non-normal*: the shape you plotted on Oct 30

	Nominal	Categorical (2+)	Ordinal	Quant. Discrete	Quant. Non-Normal	Quant. Normal
Nominal	Chi-squared, Fisher's	Chi-squared	Chi-squared Trend, Mann-Whitney	Mann-Whitney	Mann-Whitney	Student's *t*
Categorical (2+)	Chi-squared	Chi-squared	Kruskal-Wallis	Kruskal-Wallis	Kruskal-Wallis	ANOVA
Ordinal	Chi-squared Trend, Mann-Whitney		Spearman rank	Spearman rank	Spearman rank	Spearman rank, linear regression
Quant. Discrete	Logistic regression			Spearman rank	Spearman rank	Spearman rank, linear regression
Quant. Non-Normal	Logistic regression				Plot data–Pearson, Spearman	Pearson, Spearman & linear regression
Quant. Normal	Logistic regression				Linear regression	Pearson, linear regression

Which methods do we cover?

- χ^2
- Student's t
- ANOVA
- Regression (linear & logistic)

Chi-squared: categorical associations

The χ^2 family

- **Pearson χ^2:** observed vs. expected counts, $\chi^2 = \sum \frac{(O-E)^2}{E}$; use when expected counts ≥ 5.
- **Likelihood ratio (G^2):** MLE-based; converges to χ^2.
- **Fisher's exact:** exact probability under fixed margins; use for small counts.

Compare χ^2 to its distribution at the appropriate **df** → a **p-value**, as on the *t* slide.

Could people tell they were debating GPT-4? (B · Debates)

Summarize data in a **contingency table** cross-tabulating the categorical variables.

```
d <- read.csv("debates.csv")
tab <- table(opponent = d$opponent, thought = d$perceivedOpponent)
chisq.test(tab)

#          thought
# opponent  ai human
#    gpt4  225    75     75% said "AI"
#    human 216   234     48% said "AI"
#
# X-squared = 53.057, df = 1, p-value = 3.24e-13
```

Effect size: **$\varphi = \sqrt{(\chi^2 / n)} = \sqrt{(53.06 / 750)} = .27$**, a correlation for two yes/no variables. People could tell, often, but half of those facing a person also said "AI."

Student's t-test: two means

The *t*-test

Do the means of **two groups** differ?

$$t = \frac{signal}{noise} = \frac{difference}{variability} = \frac{\mu_t - \mu_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$$

Assumes: independence and roughly **normal** scores in each group (checked at the end of today).

Does design change the *t*-test?

Yes, two types:

1. **Unpaired:** the two distributions come from *different* populations.
2. **Paired:** the two distributions come from the *same* population.

Unpaired t: do five AI ideas raise novelty? (A · Stories)

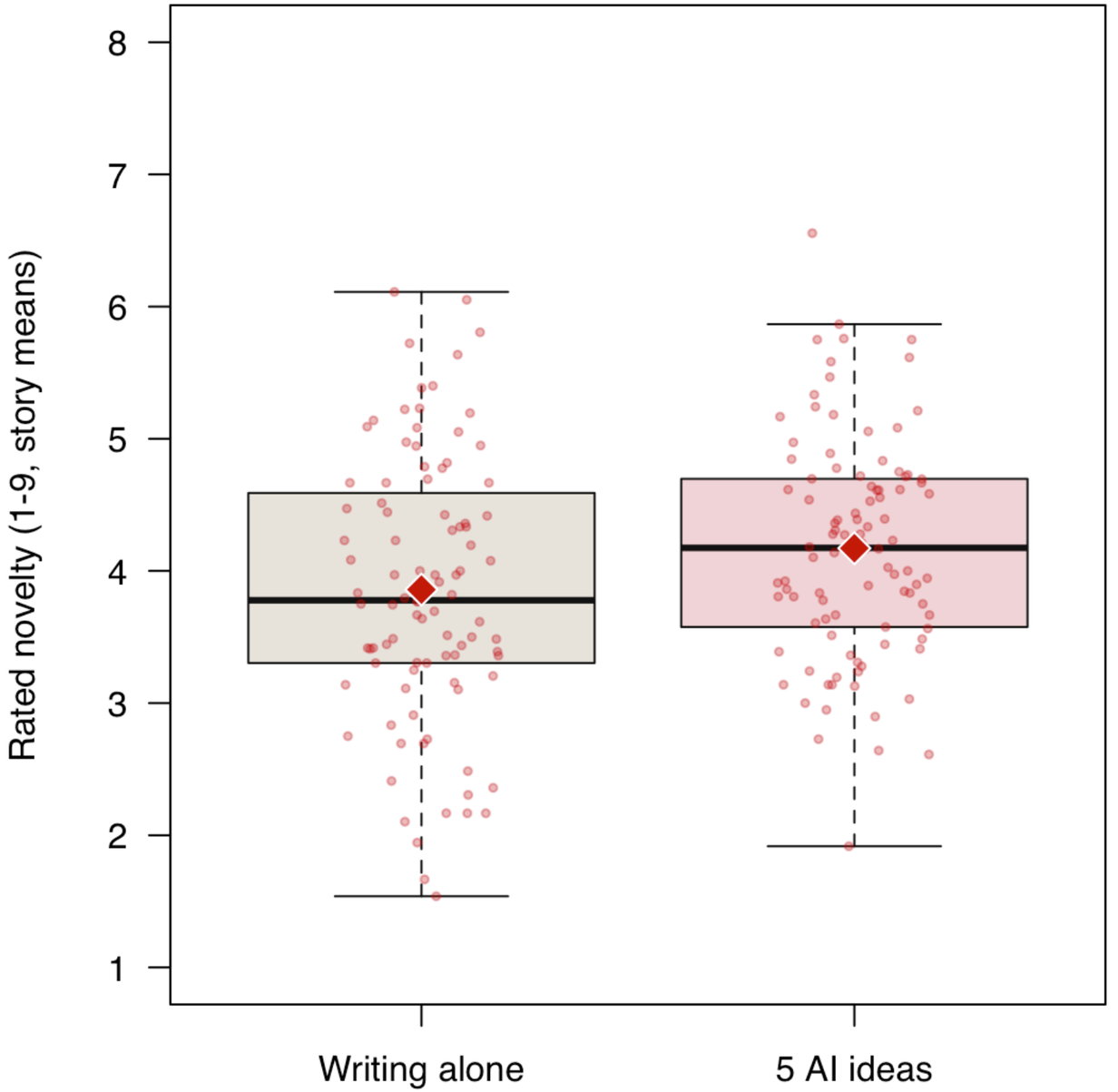
Novelty index (mean of novel, original, rare; story means), **writing alone** vs. **five AI ideas**.

```
two <- subset(s, condition %in% c("human_only", "ai_5_ideas"))
t.test(novelty ~ condition, data = two)

# t = -2.333, df = 182.31, p-value = 0.02074
# 95 percent CI: -0.578 -0.048
# mean human_only 3.86 · ai_5_ideas 4.17
```

Report: **0.31 points higher, 95% CI [0.05, 0.58]**

Effect size: **Cohen's *d* = 0.34, 95% CI [0.05, 0.62]**, small to medium



Paired t: did a personalized GPT-4 move people? (B · Debates)

Same 150 participants, agreement with the opponent's side **before** and **after** the debate.

```
g <- subset(d, opponent == "gpt4" & personalized == "yes")
t.test(g$sideAgreementPostTreatment,
       g$sideAgreementPreTreatment, paired = TRUE)

# t = 1.451, df = 149, p-value = 0.1489
# mean difference 0.14, 95% CI [-0.05, 0.33]
```

Not significant is not "no effect." The CI runs from a small move away to a third of a point toward.

Why not just run multiple t-tests?

$$H_0 : \mu_1 = \mu_2 = \mu_3, \alpha = .05 \cdot 3 \text{ pairwise tests: } (1 - \alpha)^3 \approx .86$$

Chance of at least one false positive: $1 - .86 \approx .14$, not $.05$ ➜ inflated **Type I error**.

Each test spends α once. Three tests, three chances to be wrong.

Correcting for multiple comparisons

- **Bonferroni:** p × number of tests; strict
- **Holm:** same protection, more power; **use this one**
- **Or omnibus first:** ANOVA, then post hoc (next)

A · Stories, novelty	**Uncorrected**	**Bonferroni**	**Holm**
five AI ideas vs. alone	**.021**	.062	.062
one AI idea vs. alone	.163	.488	.325
five vs. one AI idea	.427	1.000	.427

ANOVA: three or more means

ANOVA [5]

Analyzes differences among group means via the variation among vs. within groups. **Assumes**, like the *t*-test: independence, roughly **normal** residuals, similar group variances.

Procedures:

1. One-way (single factor)
2. Two-way (two factors)

Models:

1. Fixed effects (between)
2. Random effects (within)

[5] Wikipedia: Analysis of variance.

The *F*-statistic

$$F = \frac{\sigma_{explained}}{\sigma_{unexplained}} = \frac{SS_{treatment}/(k-1)}{SS_{error}/(n-k)}$$

k: number of groups · n: sample size

```
m <- aov(novelty ~ condition, data = s)   # all three conditions
summary(m)
#              Df Sum Sq Mean Sq F value Pr(>F)
# condition     2   4.87  2.4363   2.597 0.0763
# Residuals   290 272.11  0.9383
```

Test: $F(2, 290) = 2.60$, $p = .076$. The omnibus test does not reach .05.

Effect size: $\eta^2 = SS_{condition}/SS_{total} = 4.87/276.98 = .018$, small

ANOVA tells us whether, not which

Omnibus test significant → make **pairwise comparisons**:

1. **A priori** (planned contrasts)
2. **Post hoc** (e.g., Tukey HSD)

```
m <- aov(usefulness ~ condition, data = s)   # F(2, 290) = 4.39, p = .013
TukeyHSD(m)
#                          diff   lwr   upr  p adj
# ai_5_ideas-ai_1_idea     0.27 -0.08  0.62  0.165
# human_only-ai_1_idea    -0.17 -0.53  0.18  0.488
# human_only-ai_5_ideas   -0.44 -0.80 -0.09  0.010
```

Two-way: opponent × personalization (*B • Debates*)

```
m <- aov(shiftTowardOpponent ~ opponent * personalized, data = d)
#                         Df Sum Sq Mean Sq F value  Pr(>F)
# opponent                 1   12.4  12.377   8.148 0.00443
# personalized             1    1.9   1.921   1.264 0.26118
# opponent:personalized    1    0.4   0.400   0.264 0.60778
# Residuals              746 1133.2   1.519
```

shift toward opponent	not personalized	personalized
GPT-4	−0.02	0.14
human	−0.22	−0.16

A main effect of **opponent**; no reliable effect of **personalization** and no interaction, in this analysis.

Beyond one-way

- **Within-participants:** repeated-measures / random-effects ANOVA, `Error(Participant.ID/Group)`
- **Two-way:** `y ~ A * B`: main effects + interaction
- **Mixed effects:** between + within factors together
- **ANCOVA:** add a continuous covariate (e.g., writers' creative ability in A · Stories); simplify the model if it has no effect

Regression

Testing a correlation

Oct 30: *r* describes how two variables co-vary. Testing it: **H_0: ρ = 0** in the population.

```
cor.test(s$dat_score, s$novelty)
# t = 3.203, df = 289, p-value = 0.001512
# 95 percent confidence interval: 0.072 0.294
# cor 0.185
```

Writers with higher creative ability wrote stories rated more novel: ***r* = .19, 95% CI [.07, .29], *p* = .002**. r^2 = .03: they share 3% of the variance.

Linear regression

Models a **continuous outcome** from one or more predictors:
$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$

```
m <- lm(novelty ~ condition + dat_score, data = s)
summary(m)   # rounded
#                      Estimate   SE     t       p
# conditionai_1_idea       0.23 0.14  1.69    .093
# conditionai_5_ideas      0.34 0.14  2.45    .015
# dat_score                0.03 0.01  3.30    .001
# R-squared: .055
```

β: change in *y* per unit of *x*, others held fixed. Report **β, CI, p**, and R^2

Regression: assumptions & variants

- **Assumptions:** a straight-line relationship; independent observations; **equal spread** of the errors at every value of *x* (homoscedasticity); roughly normal errors
- **Logistic regression:** for a **yes/no outcome**; report the **odds ratio**
  - B · Debates: facing GPT-4, the **odds** of saying "AI" were **3.4 times** those facing a person, **OR = 3.42 [2.46, 4.74]**
- **Mixed-effects:** for **repeated measures**, such as many ratings per

Assumptions: checking them, and what to do when they fail

Checking assumptions: *t*, ANOVA, regression

- **Normality:** of the *residuals* (scores minus their group mean), not the raw outcome. **Plot it:** `qqnorm(residuals(m))`
- **Equal variances:** Welch's *t* is R's default and does not assume them; for ANOVA, compare group SDs, or use `oneway.test()`
- **Independence:** a property of the **design**, not a test: one row per story (A), debate pairs (B)

A · Stories, novelty: group SDs 0.84 to 1.04; residuals close to normal (Shapiro-Wilk $p = .52$). ANOVA fits.

When they fail: rank-based tests

Assumes normal	Rank-based alternative	In R
Unpaired *t*	Mann-Whitney (Wilcoxon rank-sum)	`wilcox.test(y ~ g)`
Paired *t*	Wilcoxon signed-rank	`wilcox.test(x, y, paired = TRUE)`
One-way ANOVA	Kruskal-Wallis	`kruskal.test(y ~ g)`
Pearson *r*	Spearman ρ	`cor.test(x, y, method = "spearman")`

A · Stories: Mann-Whitney $p = .031$ (t: .021); Kruskal-Wallis $p = .10$ (ANOVA: .076). Same conclusions.

An effect size for every test

Test	Effect size	Small · medium · large
χ^2 (2 × 2)	φ (Cramér's *V* if larger)	.1 · .3 · .5
t	Cohen's *d*	.2 · .5 · .8
ANOVA	η^2 (partial η^2 with 2+ factors)	.01 · .06 · .14
Correlation	*r*	.1 · .3 · .5
Linear regression	β with its CI; R^2	in the outcome's units
Logistic regression	odds ratio, exp(β)	no standard anchors

Report an effect size + CI with every p-value. Significance is not magnitude.

In class: *Pick the Test, Run One* (30 min)

In your pod, for **your own study** (or a practice set), starting from item 5 of your Oct 30 write-up:

- **10 min · Map hypotheses to tests:** H_0 and H_1, then **t / χ^2 / ANOVA / regression**
- **10 min · Assumptions & effect size** for each test
- **10 min · Run one** and write the result sentence

Keep it: it starts your Nov 13 write-up.

Grab the worksheet

- One copy per pod: File → Make a copy
- Stuck on a test? The **test picker** asks one question at a time
- Every example from today is in the **R lab**
- Both are linked from the practice data page

da.gd/vEfS1



No quantitative data?

Use a **practice set** as a thought experiment, and keep it through Dec 4. We grade the method, not the fit.

- **A** · GenAI and story writing
- **B** · Debating GPT-4

da.gd/d8BEE



Assignment: Inferential Analysis Plan

Individual write-up, started with your team in class. **Due Fri Nov 13, 9:30 am.** Bring it to the Analysis/Results Crit.

1. **Hypotheses:** H1, H2 … as testable inferential claims
2. **Test per hypothesis:** chosen test + *why* (variable types, design) + assumption checks
3. **Effect sizes & CIs:** what you report with each p-value (from M8)
4. **Regression model** *(if applicable)*: linear or logistic; predictors and outcome

What's Next?

- **Friday Nov 13, CRIT: Analysis/Results Crit (M10)**
  - Bring this **Analysis Plan** + your **preliminary results**, the artifact you crit
- **Project milestone:** data collection **wrapping**; next team status on Canvas Fri Nov 20, 9:30