Human-Computer Interaction

Statistics I: Descriptive Statistics + Effect Sizes & CIs

Professor Bilge Mutlu

Today's Agenda

- Statistical data analysis
- Descriptive statistics
- Effect sizes & confidence intervals *(estimation before testing)*
- Factor analysis: from items to scales
- Hands-on

Why do we need statistics?

Descriptive

Inspect the *sample*: quality, shape, spread

Describes your data

Today, plus estimation

Inferential

Generalize to the *population*: relationships

Claims beyond your data

Methods 9, Nov 6

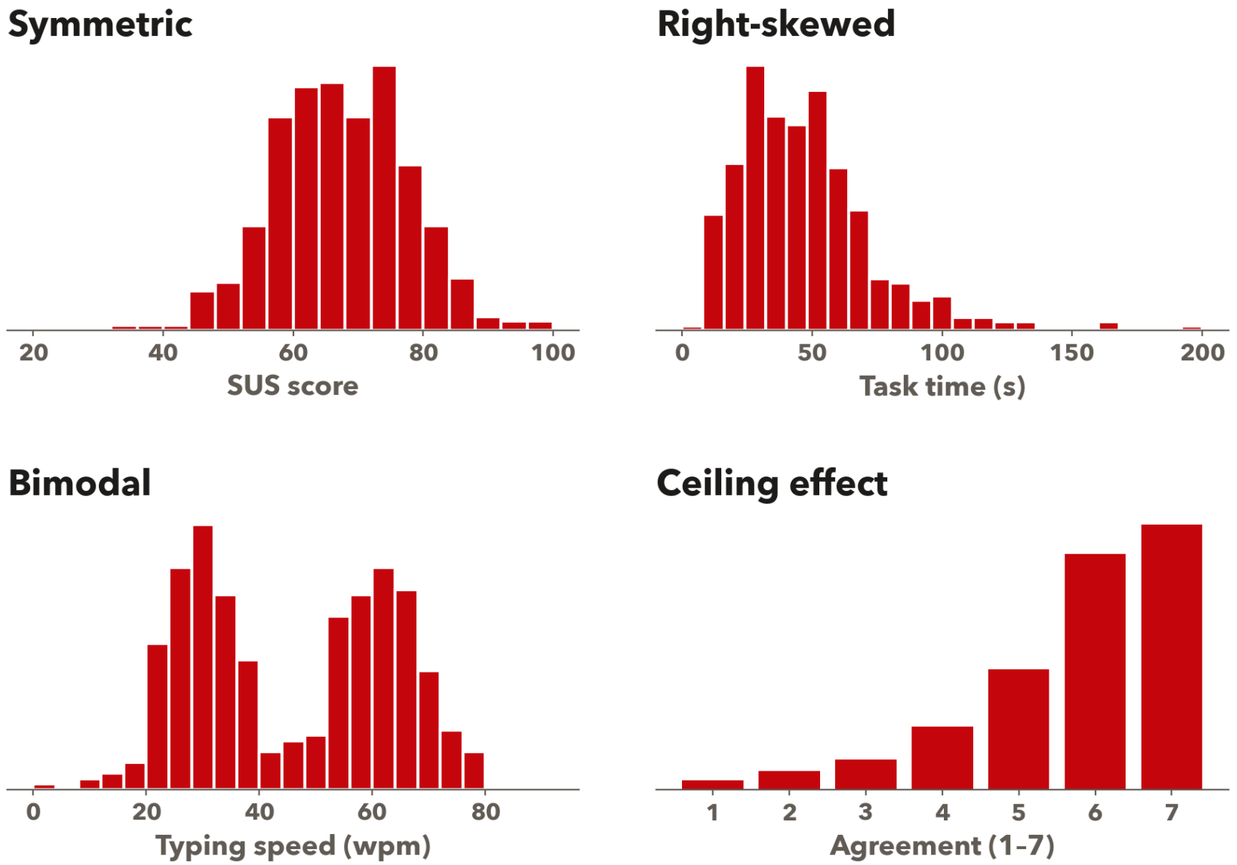
Descriptive Statistics

What do we look at in univariate analysis?

1. **Distribution:** what does the shape look like? *(discrete/ordinal/ continuous)*
2. **Central tendency:** where is the majority of the data? *(continuous)*
3. **Dispersion:** how far does it deviate from the center? *(continuous)*

Distribution: what shape is it?

- **Plot it first:** histogram if continuous, bar chart of counts if discrete or ordinal
- **Symmetric:** mean and SD summarize it well
- **Skewed:** a long tail pulls the mean; report the median + IQR
- **Bimodal:** two groups mixed together; one mean describes no one
- **Ceiling (or floor):** answers pile up at the end of the scale, leaving little room to show differences

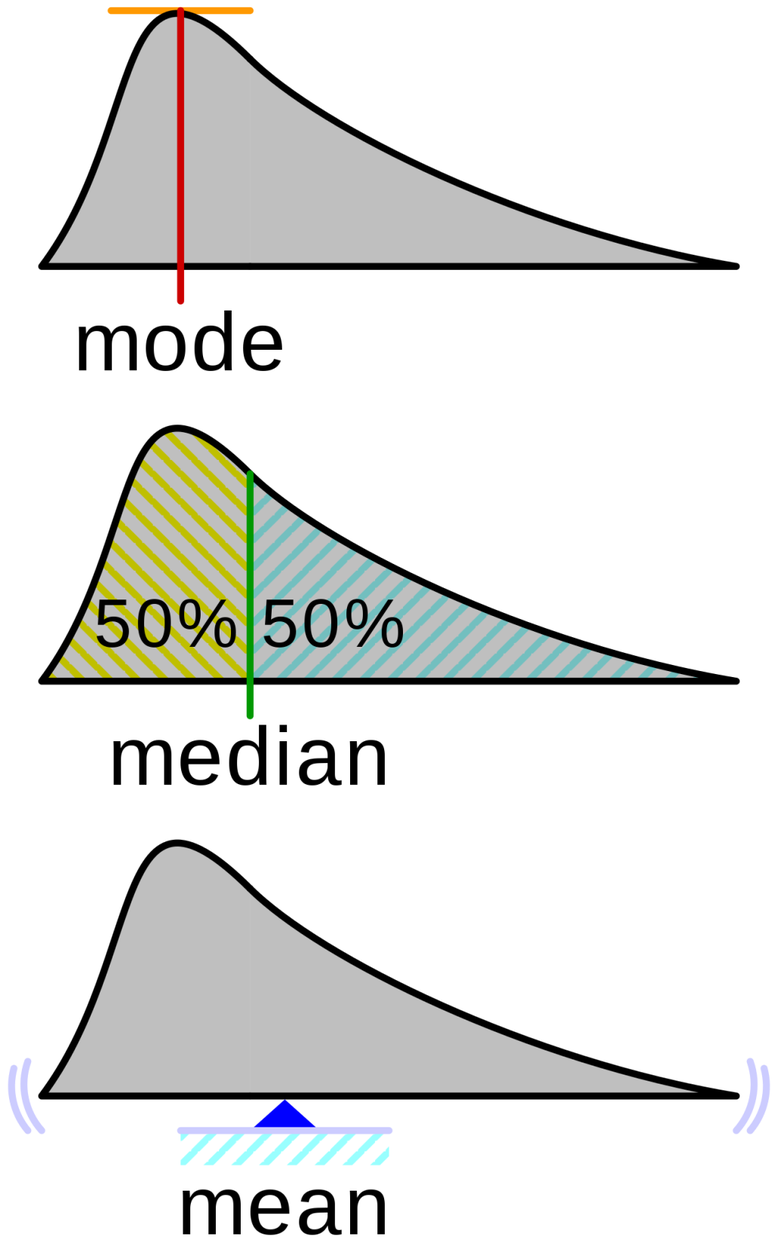


Central tendency[1]

Mean: arithmetic average, $\sum \frac{x}{n}$

Median: middle value when ordered

Mode: most frequent value



[1] Cmglee, own work, CC BY-SA 3.0.

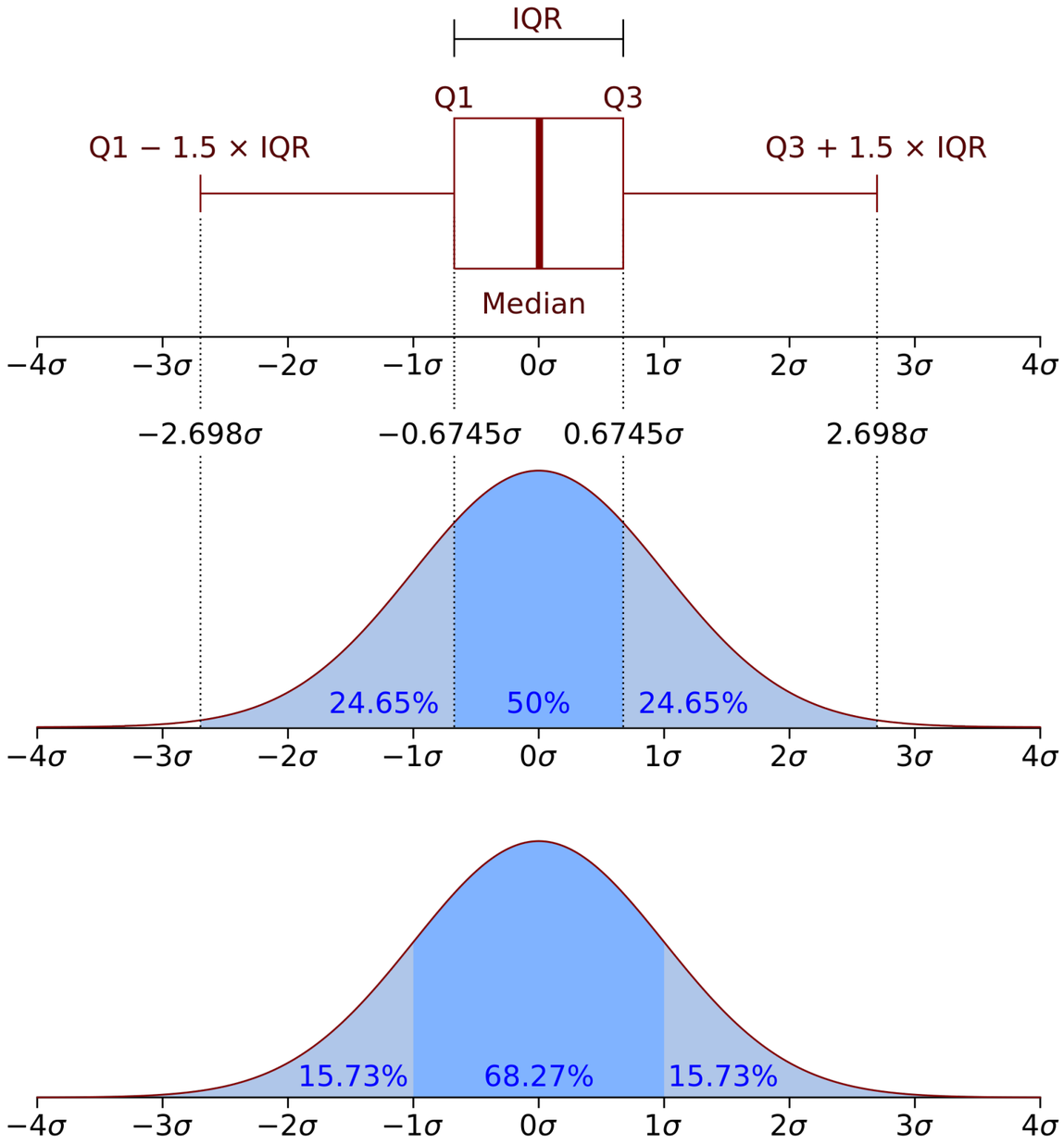
Dispersion[2]

Range: largest minus smallest

Quartiles / IQR: the distribution split into four

Variance: mean squared deviation from the mean

Standard deviation: the spread, in the data's own units



[2] Jhguch, en.wikipedia, CC BY-SA 2.5.

Descriptive statistics in R

- `describe(var)`: all descriptive statistics *(psych package)*
- `hist(var)`: histogram
- `plot(density(var))`: density plot
- `boxplot(var)`: box plot
- `plot(var1, var2)`: scatterplot

Correlation as description

Correlation & covariance measure how two variables *co-vary*; correlation is the normalized form.

- *Pearson's* r (continuous) · *Spearman's* ρ (one/both discrete or non-normal)
- r ranges −1 to 1; r^2 = the share of variance shared *(e.g.,* $r = .70 \Rightarrow 49\%$*)*
- Correlation can serve **descriptive** *or* inferential purposes

From "is there an effect?" to "how big?"

The p-value asks: *would we see this if nothing were going on?* A yes/no.

The more useful question is **estimation**:

- **How large** is the effect?
- **How precisely** do we know it?

Report the size of the thing, with your uncertainty about it, before any p-value.

Effect sizes[3]

An **effect size** puts the magnitude of a difference or relationship on an interpretable scale.

Cohen's d: standardized mean difference

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$$

Rough anchors: .2 small · .5 medium · .8 large

Correlation r: strength of a linear relationship (also an effect size)

r: −1 → 0 → 1

Report the size, not just the sign.

[3] Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, 25(1), 7–29. doi:10.1177/0956797613504966

Confidence intervals

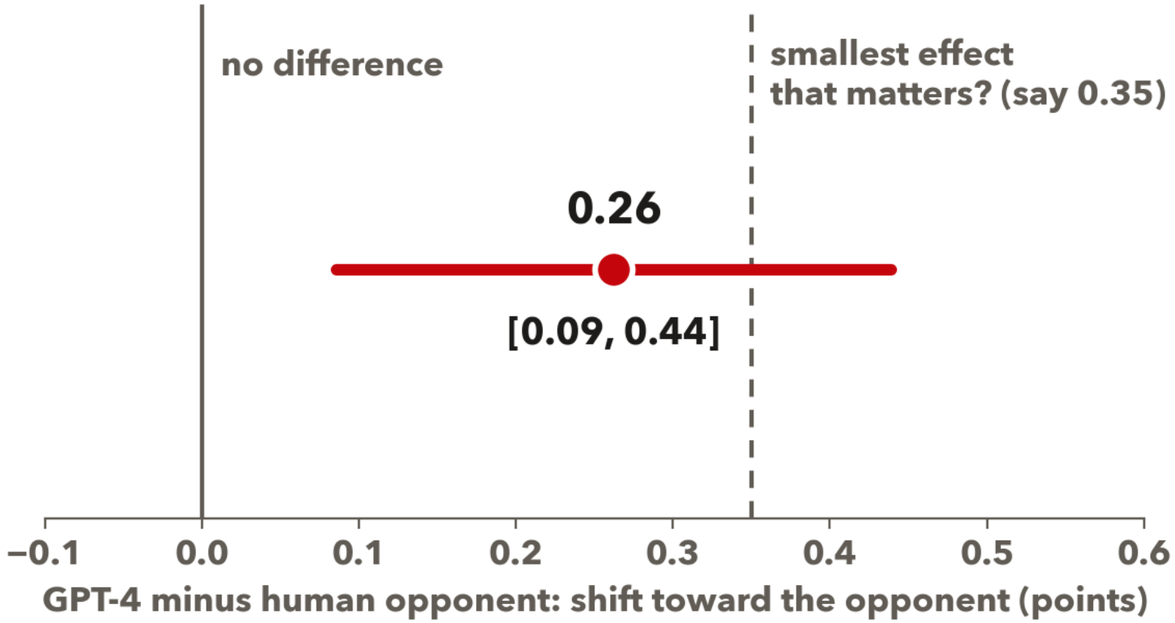
A **CI** is the range of effect sizes compatible with your data: your *precision made visible*.

- **Estimation sentence:** *"The difference is ≈ X, 95% CI [lo, hi]."*
- A **wide** CI → a small/noisy sample; the data can't yet pin the effect down
- A CI that spans zero → the direction itself is still uncertain

Write the estimation sentence before you reach for a p-value.

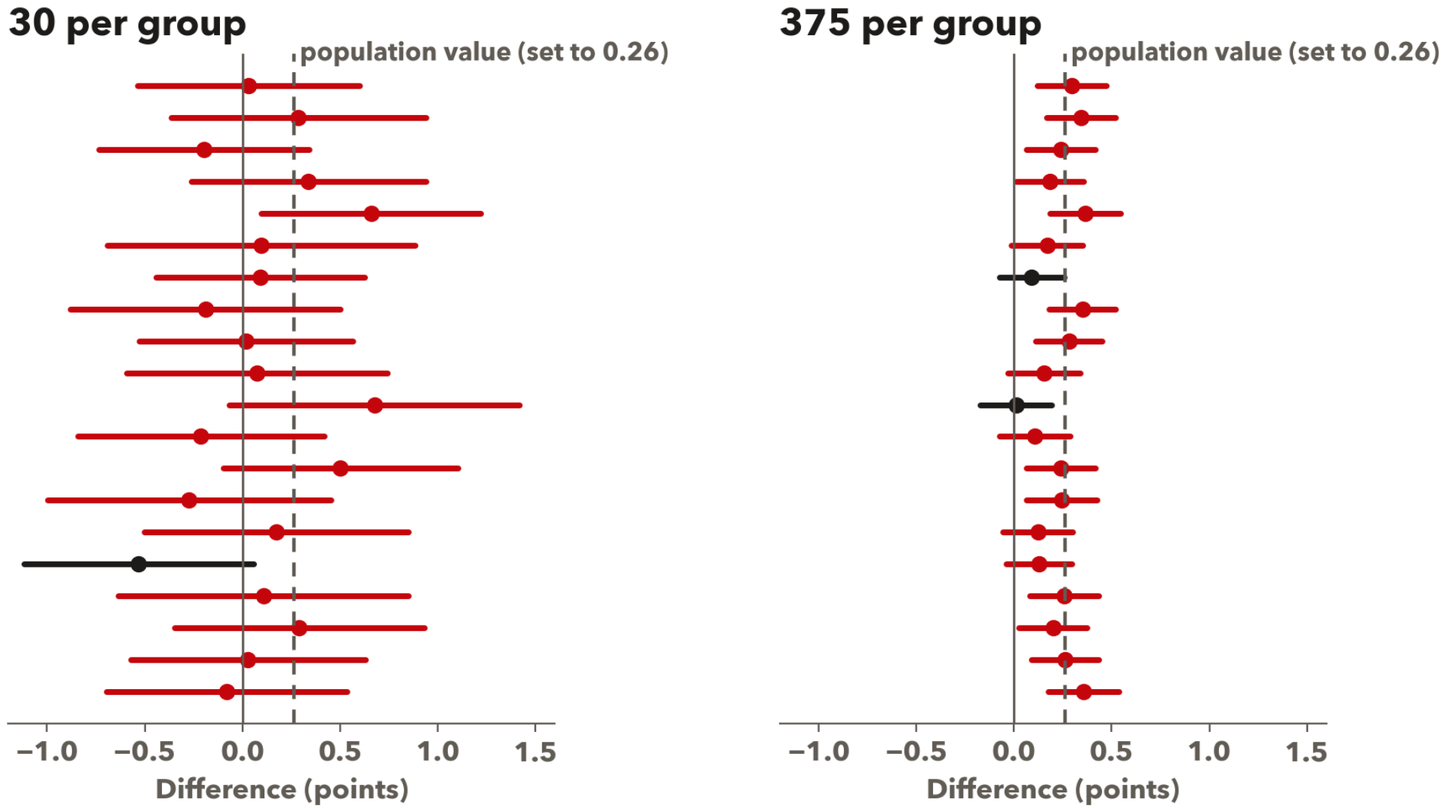
Reading a CI[4]

1. **Where is it?** The estimate: 0.26 points
2. **How wide?** ±0.17: narrow on a −4 to +4 scale, wide next to 0.26
3. **What does it rule out?** Zero, so the direction is clear. Not 0.35, so whether it matters is still open (0.35 decided based on SESOI, e.g., by decision boundary, anchors, scale-based, distribution-based)
4. **Say it:** "0.26 points more, 95% CI [0.09, 0.44]"



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

The dance of the CIs[3]



[3] Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, 25(1), 7–29. doi:10.1177/0956797613504966

Common misreadings

- *"95% of the data fall inside it."* No: it's about the mean, not individual scores
- *"Overlapping CIs mean no difference."* Not always: half an arm of overlap is still $p \approx .05$[5]
- *"95% chance the population value is in this one."* The 95% is the method's long-run rate
- *"A replication will fall inside it 95% of the time."* About 83%, on average[6]

---

[5] Cumming, G. & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, 60(2), 170–180.

[6] Cumming, G. & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, 11(3), 217–227.

From items to scales: factor analysis

Twelve ratings, or two scales?

Practice set A: each story rated on **12 items**, 1–9[7]

novel · original · rare · appropriate · feasible · publishable · enjoyed · well written · boring · funny · twist · future

Oct 16: **factor analysis** (do items group as intended?), then **alpha** for each group (do its items agree?)

Authors' indices: **novelty** = novel, original, rare; **usefulness** = appropriate, feasible, publishable. Does the data agree?

---

[7] 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.

EFA vs. CFA

Exploratory (EFA): what groups exist in my items? No structure is assumed; the data suggest one.

Confirmatory (CFA): does a structure I specify in advance fit? Items are assigned to factors up front, and the fit is tested.

EFA finds a structure; CFA tests one, ideally on new data.

The steps

1. **Prepare:** reverse-key negative items; choose the unit of analysis
2. **How many factors?** Parallel analysis, scree plot
3. **Extract and rotate:** make the loadings readable
4. **Read the loadings:** name the factors, find problem items
5. **Build scales:** average the items; check alpha

Iterate: drop or move an item, rerun.

Which rows? Ratings vs. stories

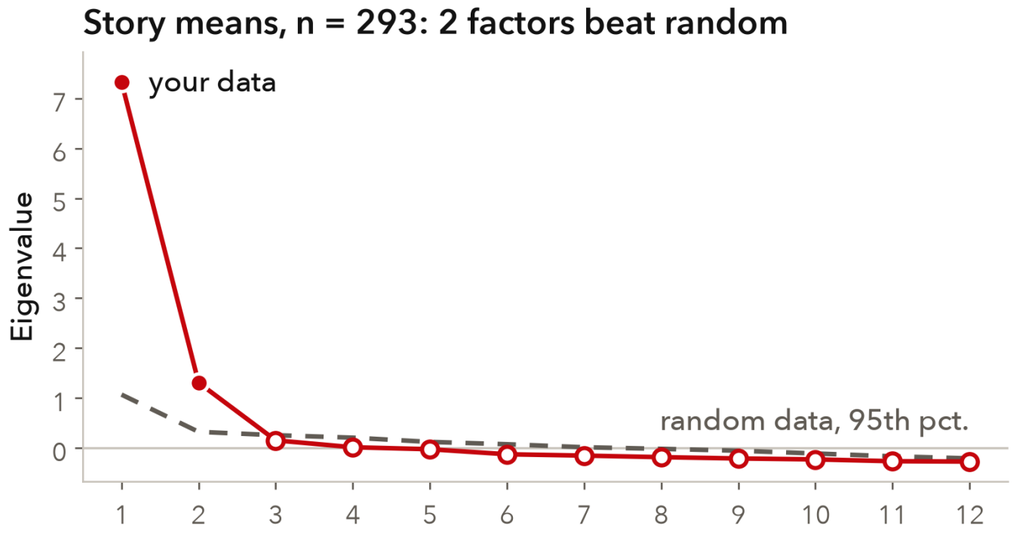
3,519 ratings from 600 evaluators on 293 stories

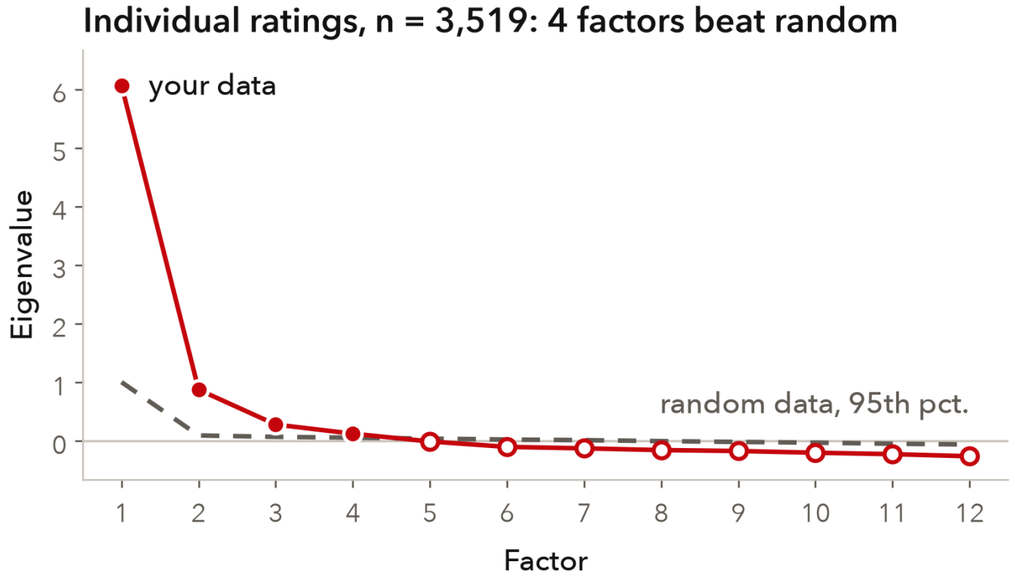
- Each evaluator rated about 6 stories; each story has about 12 ratings
- So rows are not independent, and evaluator habits, like rating everything high, add shared variance to every item
- The indices describe **stories**

Our choice: average each item per story and analyze the **293 story means**

How many factors?

- **Eigenvalue:** how much variance a factor accounts for
- **Parallel analysis:**[8] keep the factors whose eigenvalues beat those of random data of the same size
- Story means: **2**. Ratings: 4, but factors 3 and 4 are tiny
- "Eigenvalue above 1" alone is among the least accurate rules[9]





[8] Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185.

[9] Costello, A. B. & Osborne, J. W. (2005). Best practices in exploratory factor analysis. *Practical Assessment, Research & Evaluation*, 10(7).

Rotation

Extraction finds the factors; **rotation** turns them so each item loads high on one factor and near zero on the others.

- **Varimax** (orthogonal): factors forced to be uncorrelated
- **Oblimin** (oblique): factors allowed to correlate

Here the two factors correlate **.61**, so oblique fits better.[9]

[9] Costello, A. B. & Osborne, J. W. (2005). Best practices in exploratory factor analysis. *Practical Assessment, Research & Evaluation*, 10(7).

Reading the loadings

Loading: how strongly an item goes with a factor, read like a correlation. Colored: .32 or higher.[10]

- **Novelty:** novel, original, rare, twist, future
- **Quality:** appropriate, feasible, publishable, enjoyed, well written, boring
- **funny:** low communality, .28; drop it
- **enjoyed:** cross-loads under varimax, .36, not under oblimin

	Varimax		Oblimin		
	Novelty	Quality	Novelty	Quality	Shared
novel	**.78**	**.55**	**.79**	.23	.91
original	**.84**	**.45**	**.90**	.09	.91
rare	**.87**	**.38**	**.96**	.00	.91
twist	**.65**	**.34**	**.70**	.05	.54
future	**.57**	**.37**	**.60**	.13	.47
funny	**.51**	−.14	**.67**	**−.42**	**.28**
appropriate	.15	**.88**	−.07	**.93**	.79
feasible	.23	**.92**	.01	**.94**	.90
publishable	.25	**.89**	.06	**.89**	.85
enjoyed	**.36**	**.85**	.19	**.79**	.85
well written	.20	**.88**	−.01	**.91**	.81
boring, reversed	**.40**	**.75**	.27	**.66**	.72

Story means, n = 293. Colored: |loading| ≥ .32. Shared = communality, red if below .40.

[10] Tabachnick & Fidell's minimum, cited in Costello & Osborne (2005).

Build the scales, check alpha

Scale score = the mean of its items. Then alpha for each scale.

Scale	Items	Alpha
Novelty, ours	5	.92
Quality, ours	6	.96
Novelty, authors' items	3	.98
Usefulness, authors' items	3	.95

Story means, n = 293. Dropping **twist** raises our novelty alpha to .94; adding **funny** lowers it to .89.

Compare with the authors' indices

Alpha	Story means	Single ratings	Paper
Novelty: novel, original, rare	.98	.94	.92
Usefulness: appropriate, feasible, publishable	.95	.88	.89

- **Same stories score high either way:** per-story scores on our scales and the authors' indices correlate .96 (novelty) and .97 (quality vs. usefulness)
- So the extra items add little; the EFA backs the authors' grouping
- Report the unit your alpha is computed on

Factor analysis in R

```
library(psych)    # fa.parallel, fa, alpha; also install GPArotation
r <- read.csv("ai-creativity-ratings.csv")
items <- names(r)[4:15]                              # the 12 rating items
r$boring <- 10 - r$boring                            # reverse-key
s <- aggregate(r[items], list(story = r$story_id), mean)

fa.parallel(s[items], fm = "minres", fa = "fa")  # how many factors
fit <- fa(s[items], nfactors = 2, rotate = "varimax", fm = "minres")
print(fit$loadings, cutoff = .32)
alpha(s[c("novel", "original", "rare")])          # reliability
```

Full, commented script: practice-data page, *ai-creativity-factor-analysis.R*

In-class activity: *describe before you test* (30 min)

In your pod, use your team's data (or practice set A or B):

1. **(12 min) Descriptives.** Central tendency + spread for key measures; one distribution plot; note shape and outliers.
2. **(18 min) Effect size + CI.** For one comparison, compute an **effect size** (*d* / *r*) **and its CI**; write the **estimation sentence** before any p-value. What does the CI width say about your sample?

Keep it: it starts your Nov 6 write-up.

Grab the worksheet

- One copy per pod: File → Make a copy
- Nothing to submit today
- It starts your *Descriptive Summary + Effect Sizes*

da.gd/T5kZEE



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: *descriptives + effect sizes*

Due Fri Nov 6, 9:30 am · ≈1–2 pp, individual, on Canvas

1. **Measures & data:** variables, types, your *n*
2. **Descriptives:** center + spread, one distribution plot
3. **Effect sizes with CIs**, as estimation sentences
4. **Reading the estimate:** what the CI width supports
5. **Next step:** the inferential test you'll run, and why

Graded on completeness, soundness, feasibility (rubric on Canvas).

What's Next

- **Monday: XR & Spatial Computing** (Nov 2)
  - **Read:** Slater (2009) · Zhang et al., *VRCopilot* (UIST 2024)
  - **Seminar post** on Canvas by 9:30
- **Friday, Methods 9: Statistics II** (Nov 6)
  - **Read:** the textbook **Ch. 4**, inferential tests + regression
  - Bring your **descriptive summary + effect sizes** write-up
  - **Team project status** on Canvas by 9:30