University of Glasgow

User Interaction COMPSCI2031

Dr Florian Weidner Florian.Weidner@glasgow.ac.uk
Dr Ilyena Hirskyj-Douglas ilyena.hirskyj-douglas@glasgow.ac.uk

Recap: What we did last week

- Ethnography and Interviews
- Interview Task
- Reading: Ethnography in different contexts

User Interaction Topics

- ✅ HCI History and Introduction
- ✅ Usability and Heuristics
- ✅ Heuristic Evaluation and Human Cognition
- ✅ Human Perception and Capabilities
- ✅ Experimental Design & Variables Research
- ✅ Personas and Scenarios
- ✅ Surveys in HCI
- ✅ Ethnography
9. Statistical Methods
10. Theories in HCI
11. Models of Interaction
12. Large Scale and Mobile HCI
13. User-Centered Design
14. Ethics in User Testing
15. Revision & Example Exams

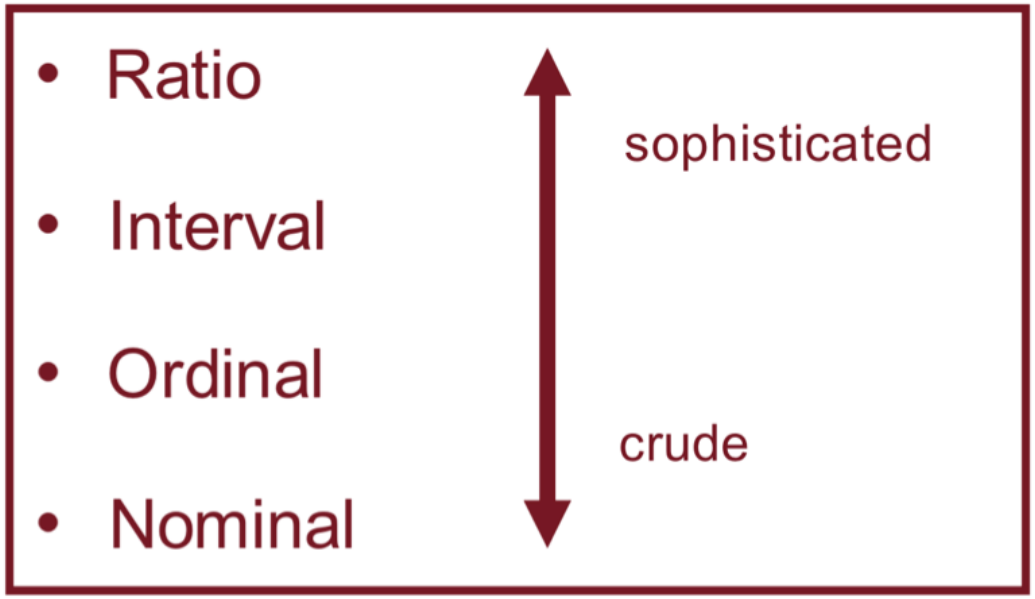
Statistics for User Studies

Lecture 9

Analysing Data from User Studies

- Providing “descriptive statistics” of quantitative data is the bare minimum in quantitative contexts (not always needed!)
  - Average, distribution, standard deviation
- Helps us to
  - make claims
  - infer causal relationships
  - hypothesis test
- Goal: Have you shown that your product is “better”?

Measurement Scales



- How to determine?
- From the types of computations possible with each measurement
- Nature of data

Measurement Scales

- Nominal / categorical
  - Labels or names
  - These could be numbers; but can’t do computations with them.
  - E.g. Happy, sad labels. ID numbers
- Ordinal
  - Can put the values in a ranking, but not equally spaced
  - e.g. ordered list of favourite films
  - Can do > or < comparisons, but not valid to calculate means

Measurement Scales

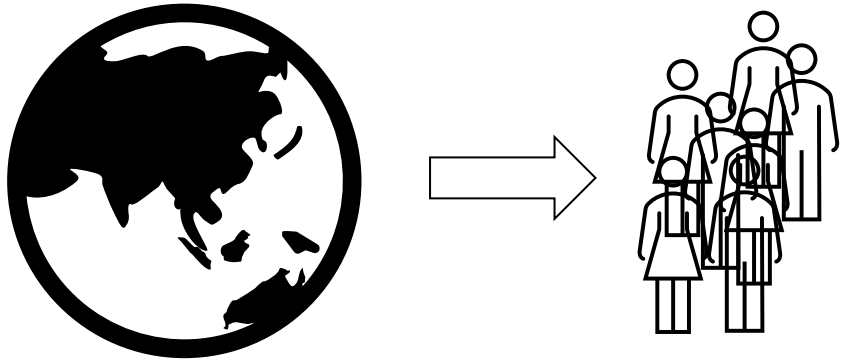
- Interval
  - Equal distances between adjacent values, but no absolute zero
- Can compute mean
  - E.g., Rating scale
  - Sometimes treated as Ordinal or Interval:
    - Important to know which if you want to compute means.
    - Treating as Interval OK if options are equally spaced and centred at neutral value.

Measurement Scales

- Ratio
  - Do have absolute zero
    - e.g. time, distance, counts of events
  - Support many calculations
  - add, divide, mean, standard deviation

Population and Sample

- Limited access to Population (all)
- We use of Samples (subset)
  → Proxy for Population
- Task Measurement
- Estimation
  → The average result from the sample is used to estimate the average result for the entire user population.



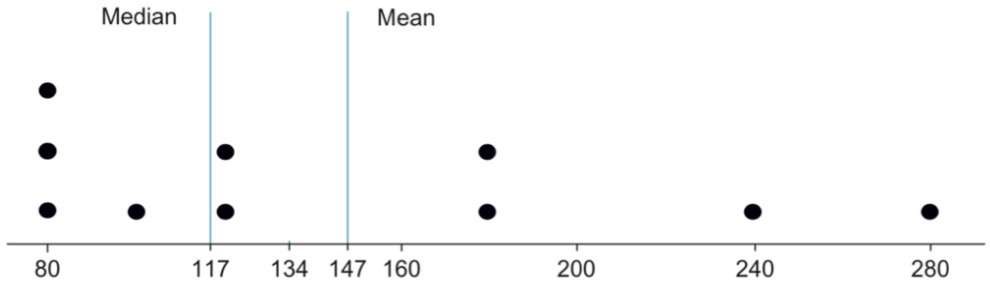
Descriptive Statistics

Descriptive Statistics

- Measures of Central Tendency: Mean, Median, and Mode
  - The mean is simple to calculate, but also provide little (or potentially misleading) information
    - Typically only useful if normally distributed data
  - The median may differ significantly from the mean, and this can insight into the “shape” of the data
- Standard Deviation describes the spread of the data
  - Estimate of average difference of values from the mean
- Plotting distributions tells you much more than simple values

Descriptive Statistics: Central Tendency

- $Mean = \frac{sum\ of\ all}{number}$
  → mean of 2, 4, 6 is $\frac{2+4+6}{3} = 4$
- *Median* = middle number of sorted values of 1, 3, 5 is 3
- *Mode* = data point which occurs most; of 2, 2, 3, 4, 4, 4 is 4



Descriptive Statistics: Mean and Mode

- Mean (X)
  = (5+5+3+3)/4
  = 4
- Mean(Y)
  = (20+22+40+42)/4
  = 31
- Modes?
  → No single mode as no repeating numbers!

X (user rating)	Y (Users age)
5	20
5	22
3	40
3	42

Standard Deviation

You only need to know what it means, not how to calculate it ☺

- Measure of how spread-out numbers are in a dataset.
- It tells you how much the numbers vary from the average (mean).

1. Find mean
2. Find difference of each value from mean
3. Square differences
4. Add up all squared differences
5. Average
6. Take square root.

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - Mean)^2}$$

Standard Deviation

- Back to our example!
- Mean (X) = 4
- Mean(Y) = 31
- SD(X) = 1
- SD(Y) = 10

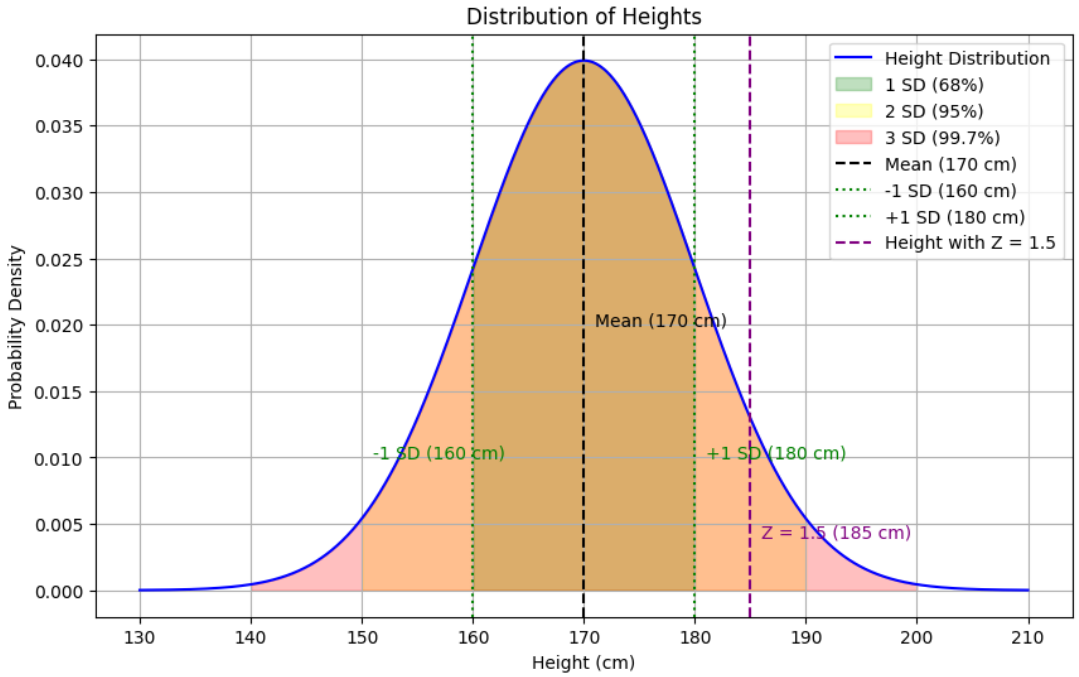
→ What does this tell us?

→ Is the SD large or small?

X (user rating)	Y (Users age)
5	20
5	22
3	40
3	42

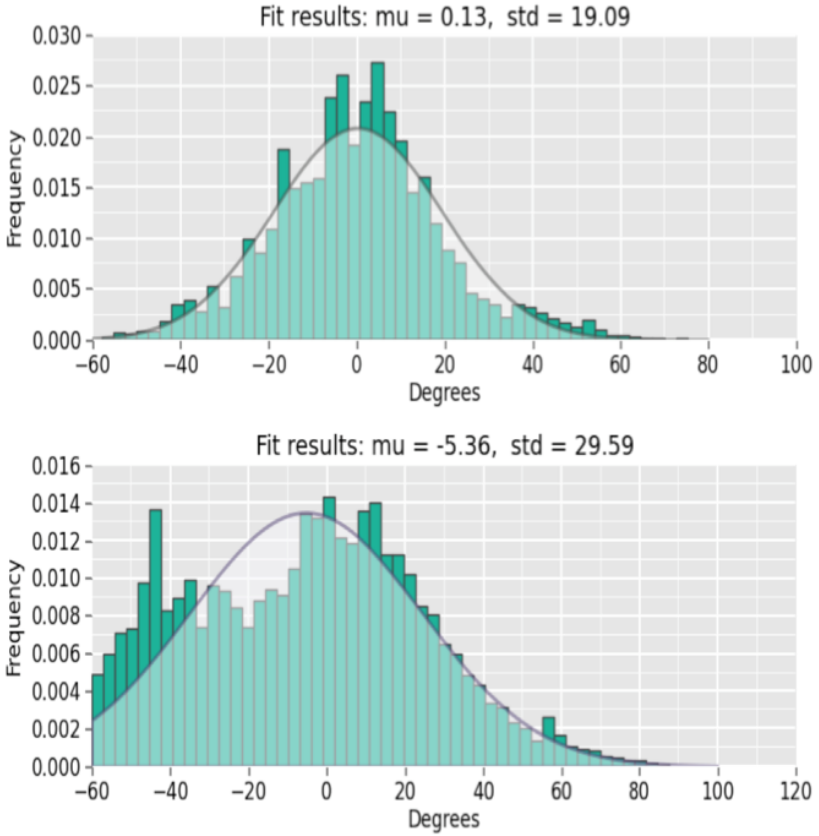
Standard Deviation and Normal Distribution

- Standard deviation?
  - Spread of data
  - Consistency
- Normal distribution?
  - Bell shaped, symmetrical curve
  - Common in nature!
- Z-Score?
  - how many sample values fall within n std devs of mean
  - 0 → value is at mean
  - >0 → value is above mean
  - $Z = \frac{Value - Mean}{StDev} = \frac{185 - 170}{10}$

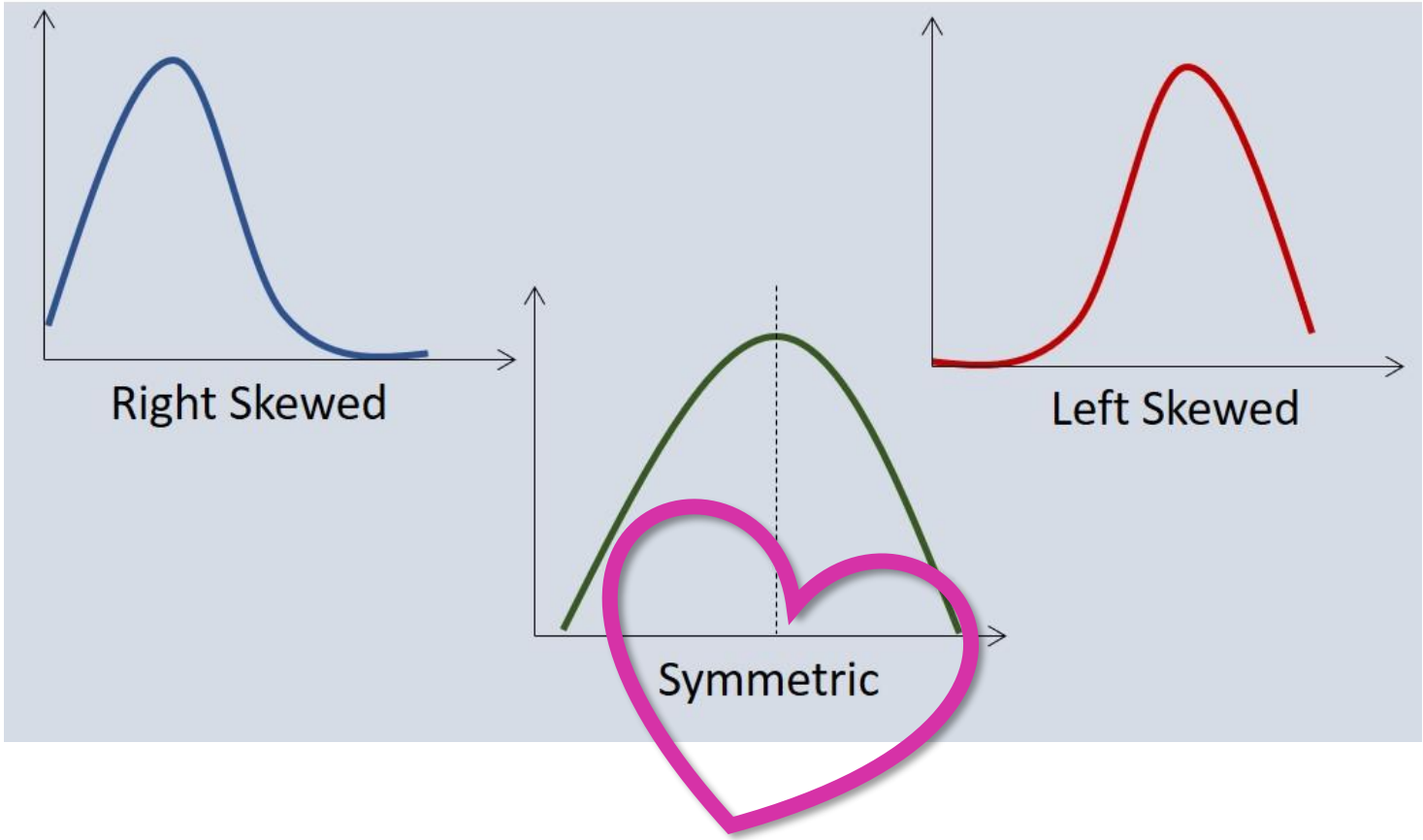


Standard Deviation for Evaluation Data

- Problem! → With human participants, the data is typically not normal distributed 🥹
- Error Rate?
- Outliers?
- Multimodal?

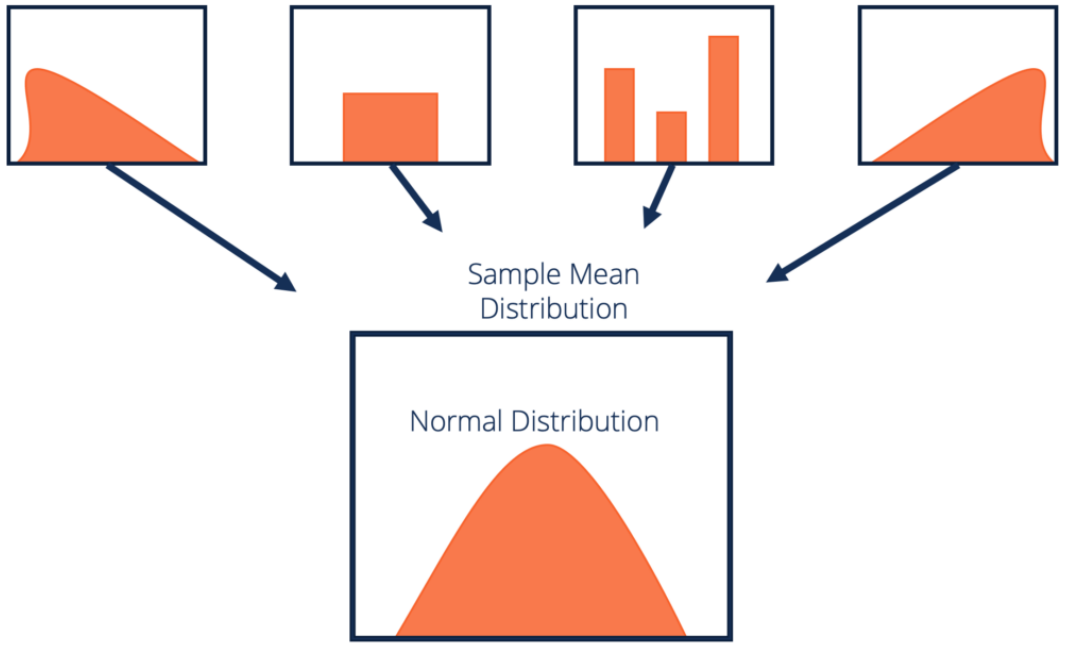


Data Skew



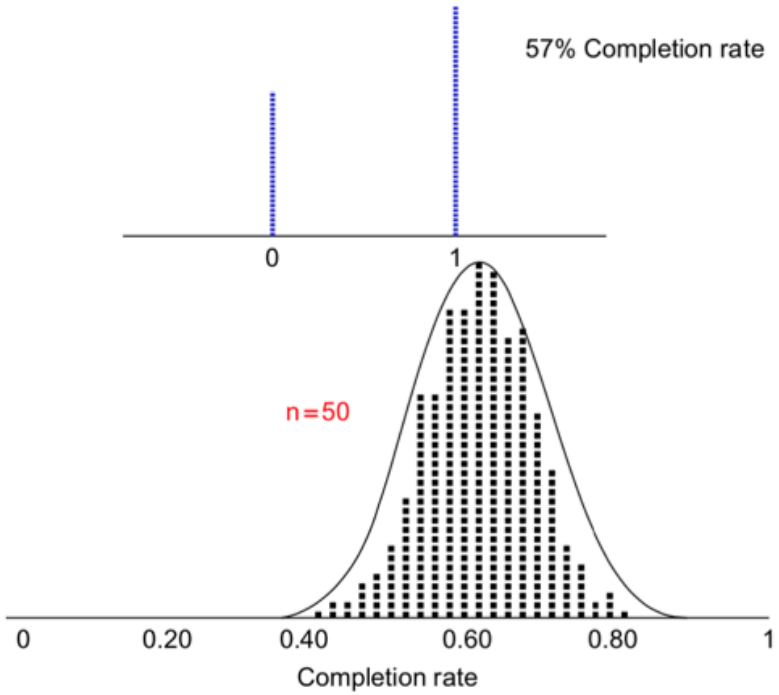
Central Limit Theorem

If you take many random samples from any population and calculate the mean of each sample, the distribution will be normal (bell-shaped, approximately), regardless of the original population's distribution.



Central Limit Theorem

- Technically: As the sample size approaches infinity…
- For us: > 30 (even smaller for interval data)
- Even applies to binary data! Example:
  - Completion (y/n)
  - Completion rate

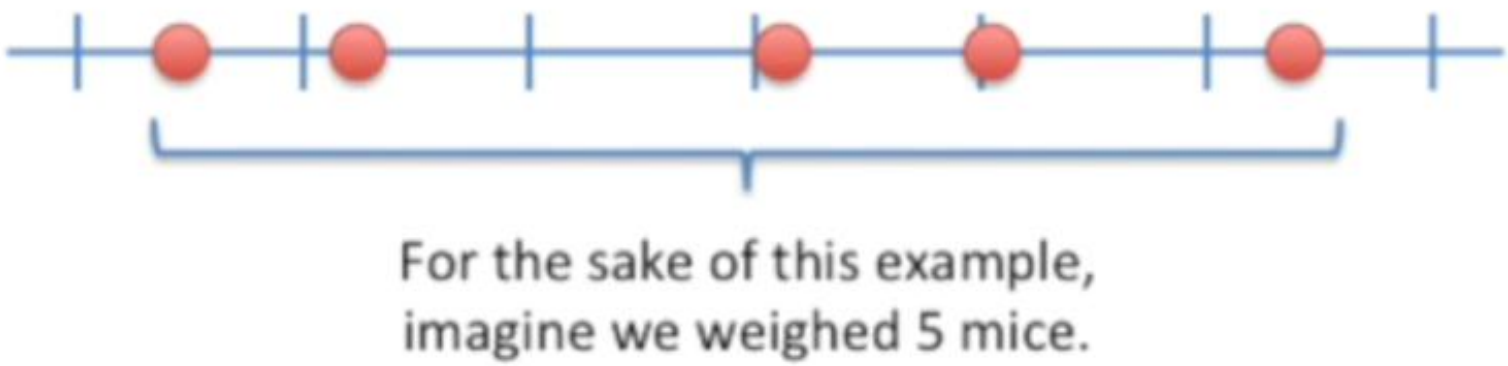


Central Limit Theorem: Implications

- Many statistical hypothesis tests (e.g. t-test) assume normal distribution of data
  - If data non-normally distributed (e.g. skewed), will these tests be invalid?
  - If sample size is large enough, CLT tells us that the distribution of sample means approximate a normal distribution
  - And so, we can use these hypothesis tests! 😊

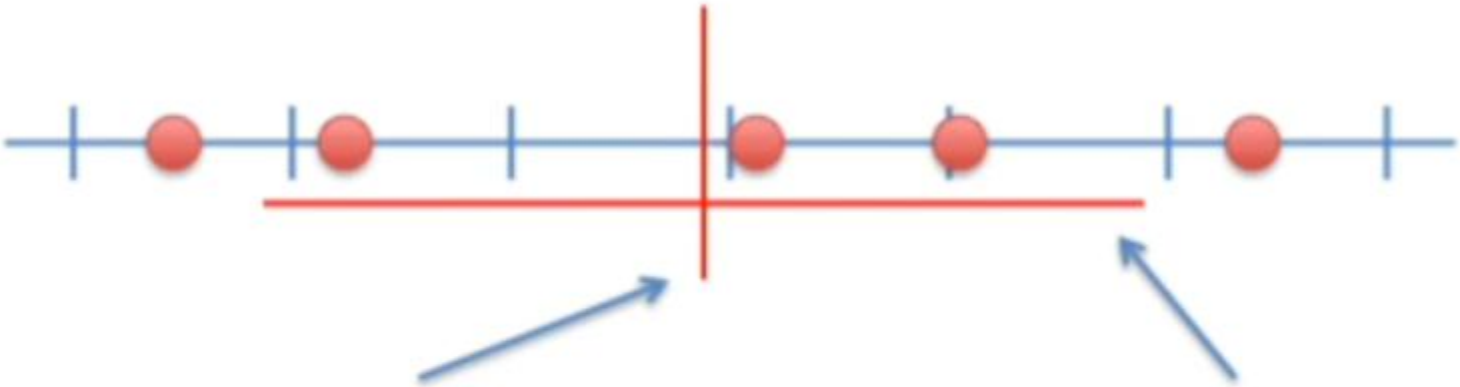
Standard Error

= standard deviation of means (plural!)



For reference: https://www.youtube.com/watch?v=A82brFpdr9g

Standard Error

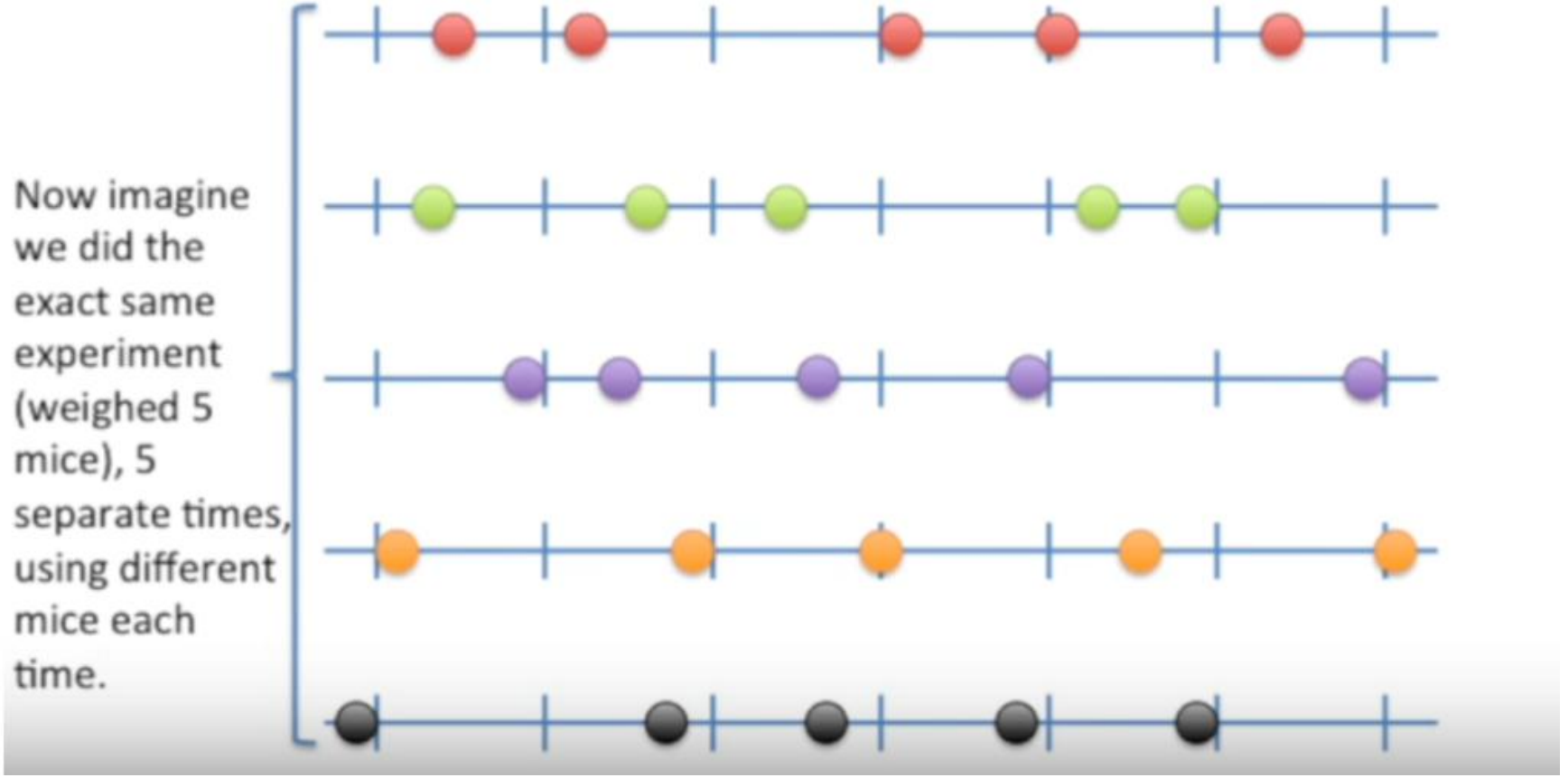


This is the average (or mean) of the values we measured.

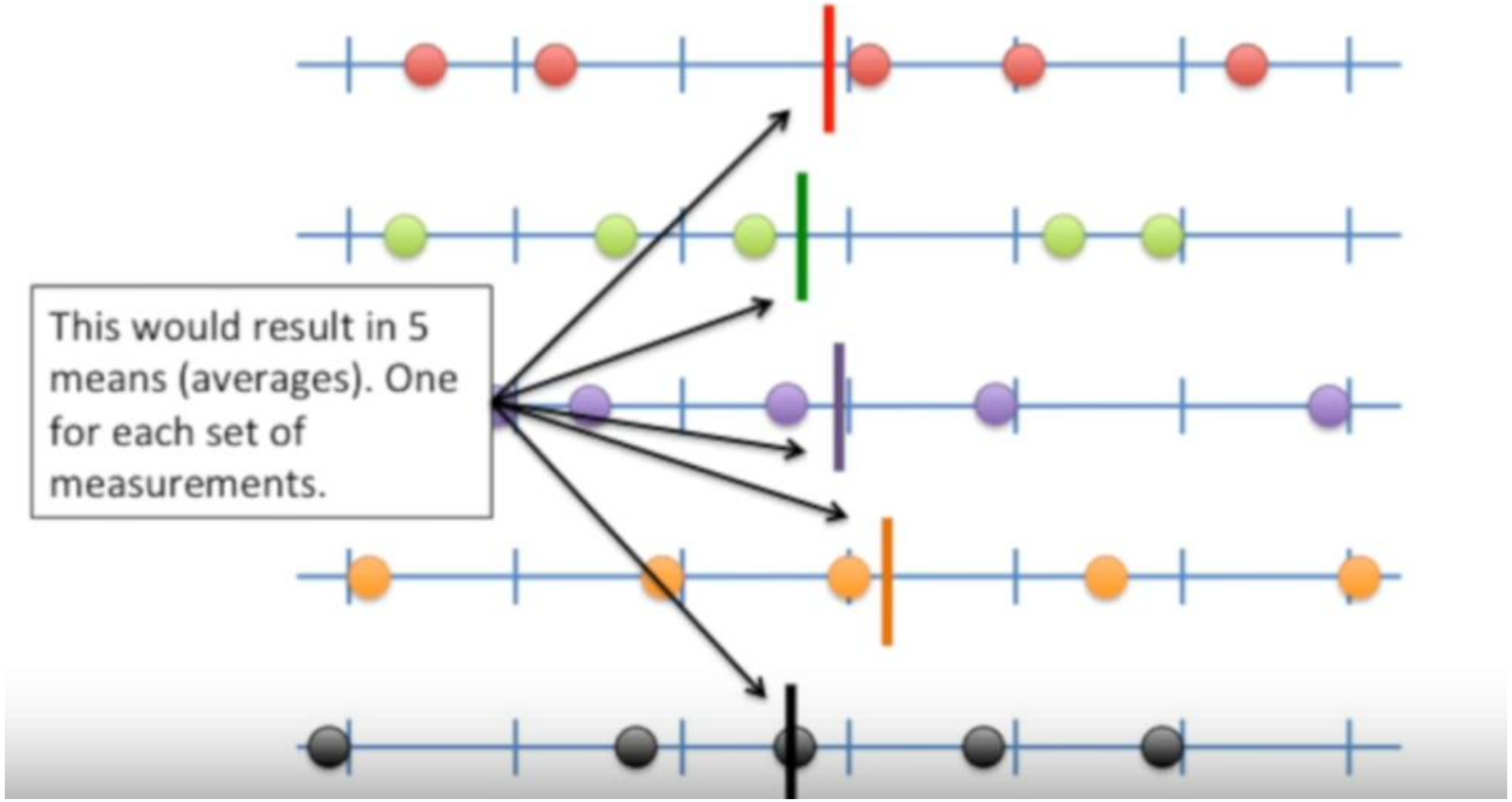
This is the **standard deviation** on both sides of the mean.

It quantifies of how much the data are spread out.

Standard Error



Standard Error

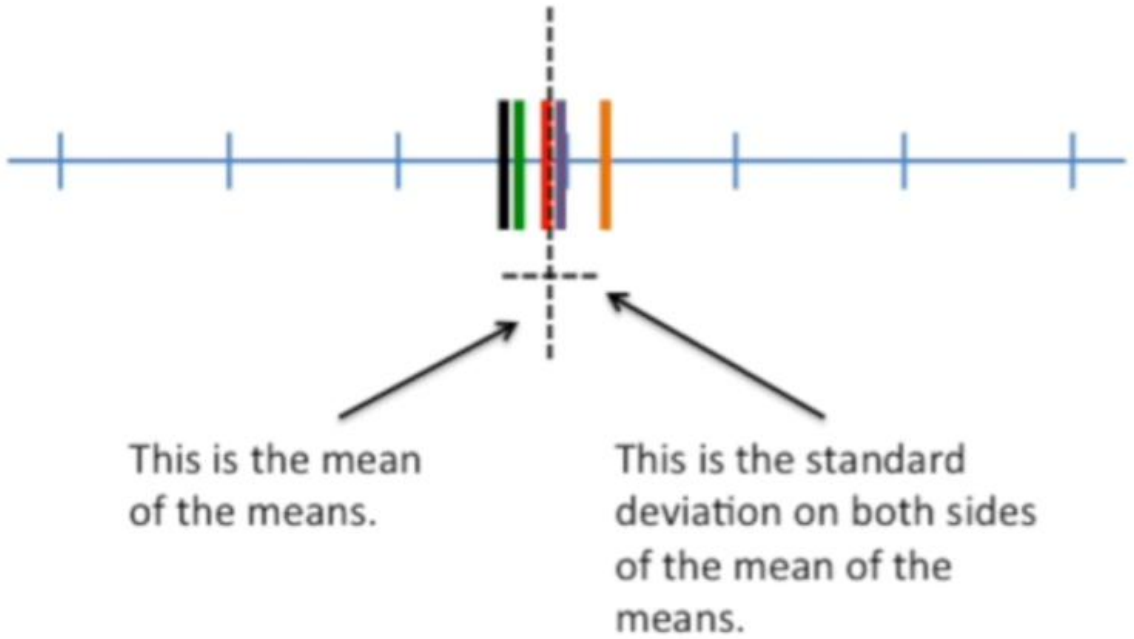


Standard Error

Here's what it would look like if we plotted all 5 means on the same number line.

Standard Error

The standard deviation of the means is called The Standard Error.



Standard Error

- But we don’t want to
  - run multiple experiments
  - take multiple samples!
- How we model standard error?

→ Estimate as: $SE = \frac{SD}{\sqrt{N}}$

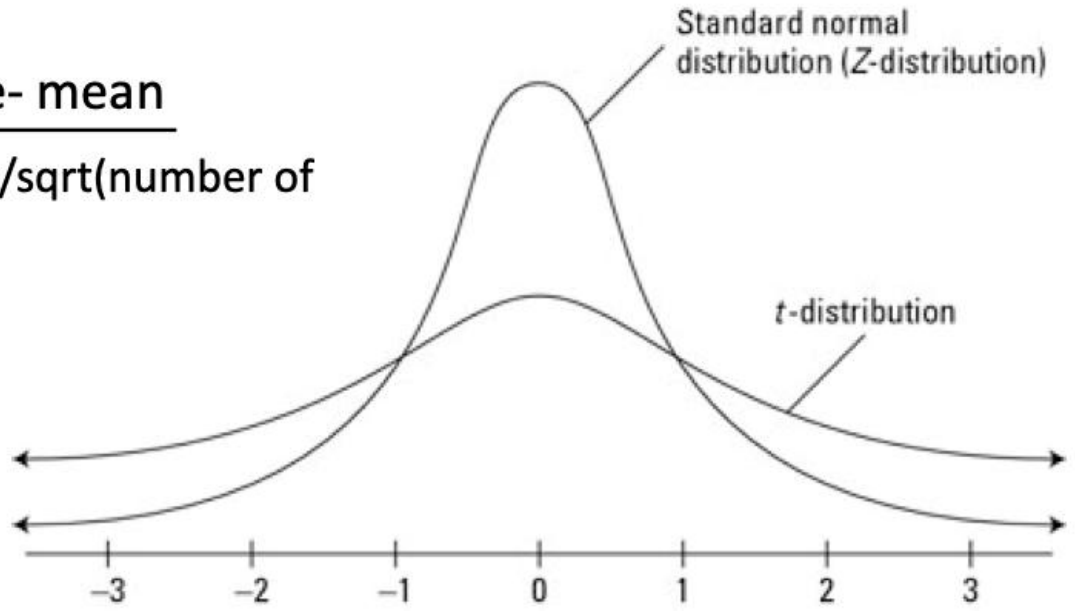
Why bother? Sample vs. Population!

- Standard deviation is about **variability** within a single dataset → sample
- Standard error is about **precision** of the sample mean as an estimate → population

t-distribution

- We cannot always know about distributions, means, SD and so on of population, only our sample
- Student's t-distribution, t-scores rather than z-scores

$$\text{t-score} = \frac{\text{value- mean}}{\text{std dev/sqrt(number of points)}}$$



For reference: https://www.jmp.com/en/statistics-knowledge-portal/t-test/t-distribution

Hypothesis testing

Statistically significant differences.

Hypothesis Testing

- Hypothesis testing to prove something measurable about a system.
- Start with key question, e.g. is A better than B?
- Consider how you will measure “better”,
  - Does using A result in faster completion times than B?
  - Does using A lead to less errors than B?
- Phrase this as a **falsifiable** statement,
  - e.g. There is no difference between A and B. (null hypothesis or H_0)
  - **Goal: Rejecting the null hypothesis → there’s a difference**

Hypothesis Testing

- We look for sufficient evidence (instead of definitive proof)
  - Science changes, often safer to say evidence rather than proof
- Evidence to reject H_0
- We use statistical test for those: there are lots! 😥
- What are stats tests testing and telling you?
  - How likely is it that two samples are from the same distribution
  - How confident are we that they're different? → confidence interval
  - By how much are they different? → p-value 🦄

For reference: https://github.com/valentin-schwind/statistics-decision-tree

Hypothesis Testing

- Consider an example comparing a mouse to a trackpad
  Null hypothesis H_0: There is no difference between user performance in using these two input devices for an object selection task.
- Collect data about user performance for both devices.
- Perform hypothesis testing.
- See if we can reject null hypothesis H_0 or not
  - Reject H0 → there is probably a difference
  - Fail to reject H0 → there's probably no difference (but we don't know)
- How to do this test?

Hypothesis Testing: T-Tests

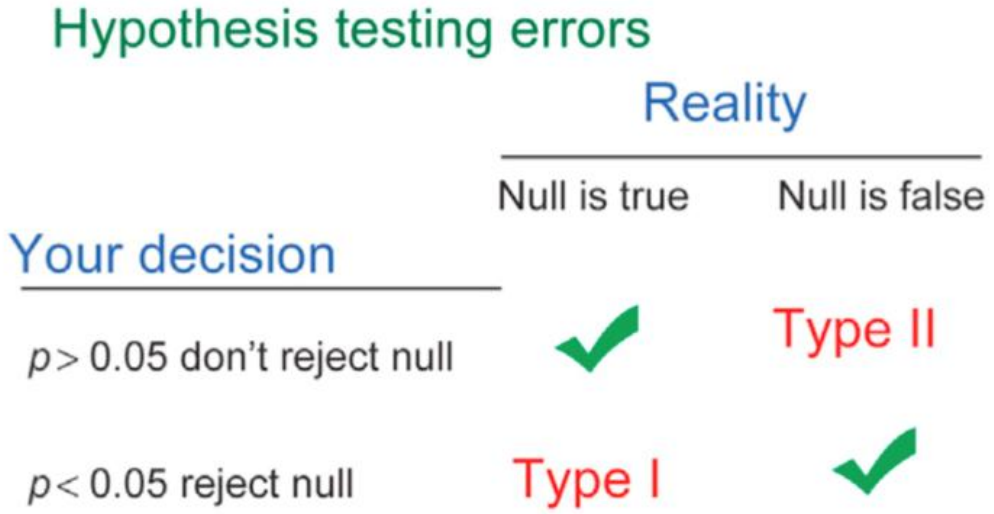
- There are many different tests. One of the most common is t-test
- Assumptions it makes:
  - Data follows a normal distribution
  - Data drawn from interval/ratio data
- What it does:
  - Compares sample means
  - Computes our 🪄 **p-value** 🪄
  - Computes confidence interval (CI)

reliability of the estimated mean

- Value between 0 and 1 (percentage)
- Lots of confusion about this! Careful!
- "How likely it is to get this results if the H_0 is true." or "How likely is it that difference is by chance?" or "Low *p* means the null hypothesis unlikely to be true."
- In HCI: if < **0.05**, less than 🪄 **5%** 🪄 chance that difference is by chance
  → we reject H_0

Hypothesis Testing: Errors

- Remember: H_0 = no difference
- p: Likelihood that difference is by chance.
- Type 1: False Positive
  Acting on something that is not real.
- Type 2: False Negative
  Missing something that is real.



Errors in Statistical Testing

- Type 1 and 2 errors can never be avoided entirely
  → can reduce their likelihood by increasing sample size.
- Much of statistical theory is around avoided these two types of errors
- Statistical tests by nature are probabilistic
  → we cannot know for certain whether conclusions are correct
- Significance level can help with Type 1* (False Positives) with
  → 5% chance of incorrectly rejecting the true null hypothesis
- Adjusting significance level: reducing one increases the other

*Type 1: Incorrectly rejecting H_0. Assuming there is a difference when there is none.
Type 2: Incorrectly failing to reject H_0. Assuming there is no difference when there is one

Correlations

Relationships between two variables

Correlations: Relationships between two variables

You only need to know what it means, not how to calculate it ☺

- Correlation coefficient (Pearson's r) → measures statistical relationship between two variables: X and Y
- E.g., from surveys you can get
  - X = rating of user satisfaction
  - Y = user age
- Measures relationship strength and direction (positive or negative).
- Practically, we will compute sample correlation coefficient

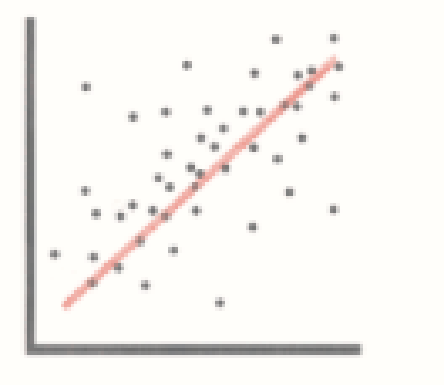
- $S_{X,Y} = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \overline{y})$
  → Sum of product of differences from mean
- $S_{X,X} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$
  → Sum of squares for X
- $S_{Y,Y} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$
  → Sum of squares for Y
- → $r(X,Y) = \frac{S_{X,Y}}{\sqrt{S_{XX}\,S_{YY}}}$

$\bar{x} = mean\ of\ all\ x'$
$x_i = i'th\ x\ value\ (individual\ x)$
$N = sample\ size$

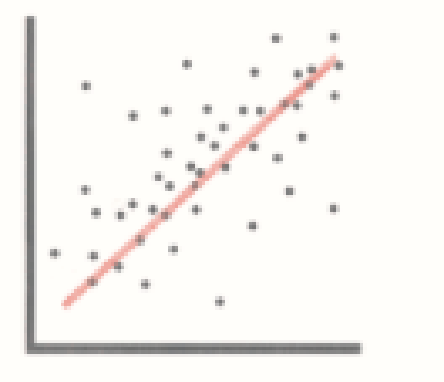
Comparing Correlation Plots

- Correlation coefficient (r) is between -1 to 1.
- 1 means perfect positive correlation, -1 means perfect negative correlation, and 0 means no correlation.

Example Comparing Images: R

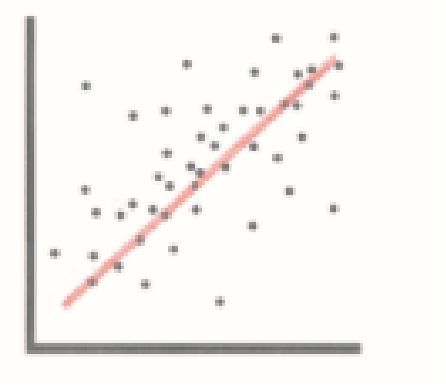


Example Comparing Images: R

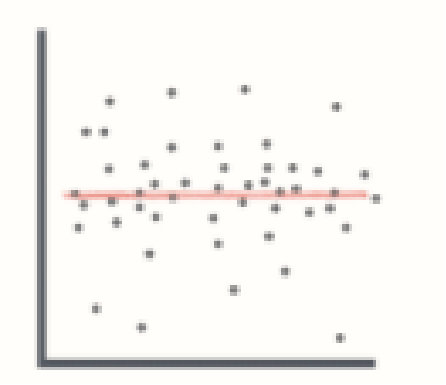


- Positive

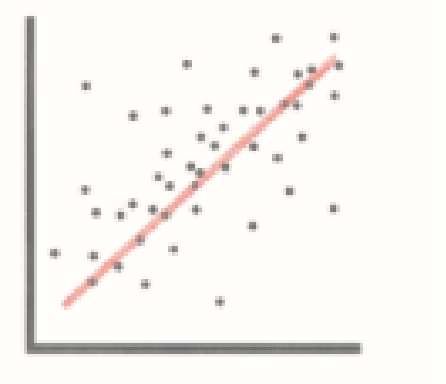
Example Comparing Images: R



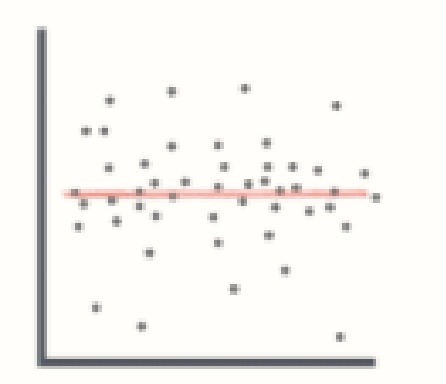
- Positive



Example Comparing Images: R



- Positive



- No Correlation

Questions?
Comments?
Concerns?

User Interaction Topics

- ✅ HCI History and Introduction
- ✅ Usability and Heuristics
- ✅ Heuristic Evaluation and Human Cognition
- ✅ Human Perception and Capabilities
- ✅ Experimental Design & Variables Research
- ✅ Personas and Scenarios
- ✅ Surveys in HCI
- ✅ Ethnography
- ✅ Statical Methods
- Theories in HCI
- Models of Interaction
- Large Scale and Mobile HCI
- User-Centered Design
- Ethics in User Testing
- Revision & Example Exams

Statistics Task

Feel free to use online calculators!

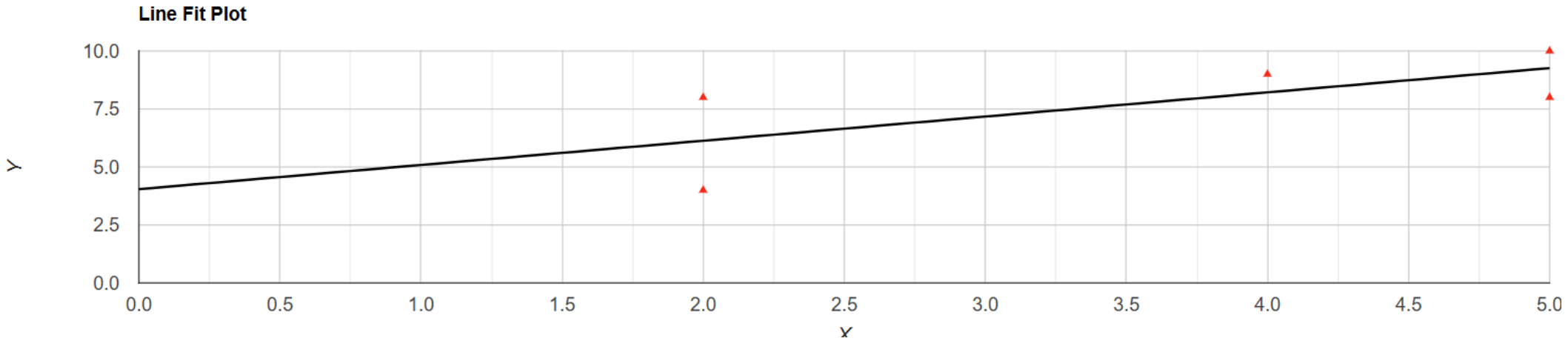
- Calculate:
  - The average user satisfaction
  - The variability or spread UX
  - Typical user satisfaction
  - Most frequent UX rating
- For 1.-4., assign mode, mean, median, and standard deviation to each of the above.
- Calculate Pearson's r and describe what it means
- Post results in Teams chats ☺

- Fictional Dataset:

Participant ID	User Satisfaction Rating	User Experience UX
1	5	10
2	5	8
3	4	9
4	2	8
5	2	4

Statistics Task: Class Discussion

- https://www.statskingdom.com/correlation-calculator.html
- https://www.socscistatistics.com/tests/pearson/default2.aspx



Reading

- Reading: Sauro & Lewis, Quantifying User Experience: Appendix A Crash Course in Fundamental Statistical Concepts File (whole chapter)
- https://www.sciencedirect.com/science/article/pii/B9780128023082000126