USC Viterbi
School of Engineering

Experimental Design

CSCI 699 Computational Human-Robot Interaction

Instructor: Stefanos Nikolaidis

Experiments

- Scientific experiments are typically informed by **theories**, also called hypotheses

- We typically assume that there is some variable X that causally influences some other variable Y.

- We can do an experiment to generate knowledge about how changing X influences Y in a given direction.

Experiments

- Experiments were originally used in medicine, to test drug effectiveness.

- There would be two groups of patients, one administered treatment A, another treatment B, and they test whether treatment A has a significantly better effect on treating acne compared to treatment B.

- To generate meaningful conclusions, we need to manipulate only the variable and that we are studying.

Experiments in HRI

- Say we want to test whether algorithm A performs better than algorithm B.

- It is important to keep everything else constant: same robot, same computer, same initial and final conditions

Independent Variables

- When testing medical treatments, it could be drug A vs drug B
- When testing HRI algorithms, it could be algorithm A or algorithm B

Dependent Variables

- What we try to measure
- If we want to test the effect of a drug on treating acne, we would like to have some indication on the level of the acne before and after the drug was administered.
- If we want to test whether the robot has learned the task, we can compare the distance between the robot's trajectory to a default trajectory

Population

- Where we are going to collect all data from
- Population of patients with a level of acne
- People that interact with the robot

Hypotheses

- A hypothesis describes a relationship between an independent variable and a dependent variable.
  - Independent variable *x* affects dependent variable *y*.

- A hypothesis may also make claims about the direction of the relationship
  - *x* positively affects *y,* or *x* negatively affects *y*

Hypotheses

- A hypothesis describes a relationship between an independent variable and a dependent variable.
  - Independent variable *x* affects dependent variable *y*.
- A hypothesis may also make claims about the direction of the relationship
  - *x* positively affects *y,* or *x* negatively affects *y*
- How to make hypotheses?
  - Intuition
  - Observation

Pilot Studies

- Studies with a small number of people (sometimes friends / lab-mates)

- Studies help with:

Pilot Studies

- Studies with a small number of people (sometimes friends / lab-mates)

- Studies help with:
  - Identify trends
  - Identify issues with design (ambiguous instructions, debugging)

Experimental Design

- **Between Subjects**
  - Two groups: a group A that takes drug A, a group B that takes drug B, or a group that works with algorithm A, a second group that works with algorithm B.

- **Within Subjects**
  - Same group: first they take drug A, then they take drug B
  - Work with robot A, then with robot B

Experimental Design

- **Within vs Between Subjects**
  - Ordering effect ✗
  - Potentially undesirable effects (fatigue) ✗
  - Smaller number of samples required ✓
  - Avoid individual differences affecting the results ✓

Experimental Design

- **Between Subjects**
  - Two groups: a group A that takes drug A, a group B that takes drug B, or a group that works with algorithm A, a second group that works with algorithm B.
- **Within Subjects**
  - Same group: first they take drug A, then they take drug B
  - Work with robot A, then with robot B
- **Mixed Design**
  - Experiment is within subjects with respect to some variables and between subjects with respect to some other variables

Mixed Design Example

- First group is with a manufacturing robot that executes algorithm A and a humanoid robot that executes algorithm A

- Second group is with a manufacturing robot that executes algorithm B and a humanoid robot that executes algorithm B.

Mixed Design Example

- First group is with a manufacturing robot that executes algorithm A and a humanoid robot that executes algorithm A
- Second group is with a manufacturing robot that executes algorithm B and a humanoid robot that executes algorithm B.
- How can we remove the bias from people seeing algorithm B with the manufacturing robot first?

Mixed Design Example

- First group is with a manufacturing robot that executes algorithm A and a humanoid robot that executes algorithm A

- Second group is with a manufacturing robot that executes algorithm B and a humanoid robot that executes algorithm B.

- How can we remove the bias from people seeing algorithm B with the manufacturing robot first?

  counterbalance the order

Mixed Design Example

- First group is with a manufacturing robot that executes algorithm A and a humanoid robot that executes algorithm A

- Second group is with a manufacturing robot that executes algorithm B and a humanoid robot that executes algorithm B.

- How can we remove the bias from people seeing algorithm B with the manufacturing robot first?

  counterbalance the order

  have a training period

Confounds

- These are factors that affect the outcomes
- Example: “estrogen treatment was significantly correlated with positive health”
- What are possible confounds?

Confounds

- These are factors that affect the outcomes
- Example: “estrogen treatment was significantly correlated with positive health”
- What are possible confounds?
- In human-robot interaction: algorithm parameter tuning

Researcher Bias

- Can be implicit in how researchers provide instruction to participants (encouraging them towards one practice vs the other)

- Bias can be reduced by:
  - Having a script
  - Having someone that does not know the conditions / hypotheses do the experiment

Researcher Bias

- Can be implicit in how researchers provide instructions to participants (encouraging them towards one practice vs the other)

- Bias can be reduced by:
  - Having a script
  - Having someone that does not know the conditions / hypotheses do the experiment

- Data Annotation

Researcher Bias

- Can be implicit in how researchers provide instructiosn to participants (encouraging them towards one practice vs the other)

- Bias can be reduced by:
  - Having a script
  - Having someone that does not know the conditions / hypotheses do the experiment

- Data Annotation
  - Automate data annotation
  - Have independent annotators, test for consistency

Additional Considerations

- Learning: The response of a human subject changes as an experiment proceeds, because they gain skill or knowledge

- Mitigation Strategies:
  - Train subjects
  - Test a lot of subjects and vary the order of the test sequence

Additional Considerations

- Learning: The response of a human subject changes as an experiment proceeds, because they gain skill or knowledge

- Mitigation Strategies:
  - Train subjects
  - Test a lot of subjects and vary the order of the test sequence

- Fatigue

Additional Considerations

- Learning: The response of a human subject changes as an experiment proceeds, because they gain skill or knowledge

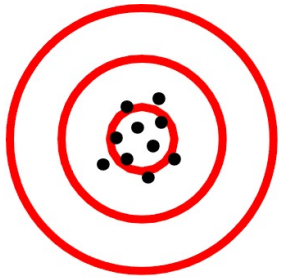
- Mitigation Strategies:
  - Train subjects
  - Test a lot of subjects and vary the order of the test sequence

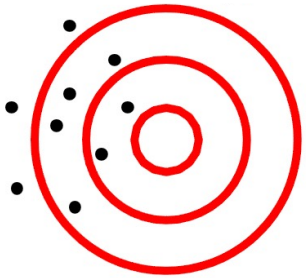
- Fatigue
  - Subject fatigue
  - Experimenter fatigue
  - Experimental materials fatigue

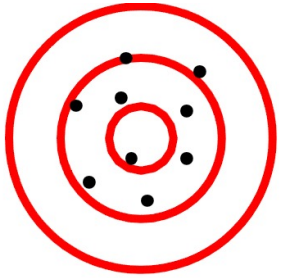
Validity / Reliability

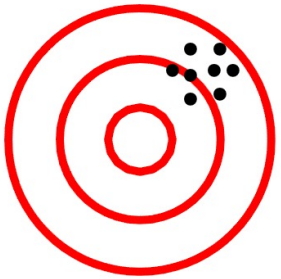
- Validity: Do we measure what we want to measure? (measure the right thing)
- Reliability (error, variance): measure the thing right
- Poor validity example: test a human-robot interaction algorithm that you want to use in a factory at USC undergrads
- Poor reliability example: every time you do an IQ test you get different results

Validity / Reliability









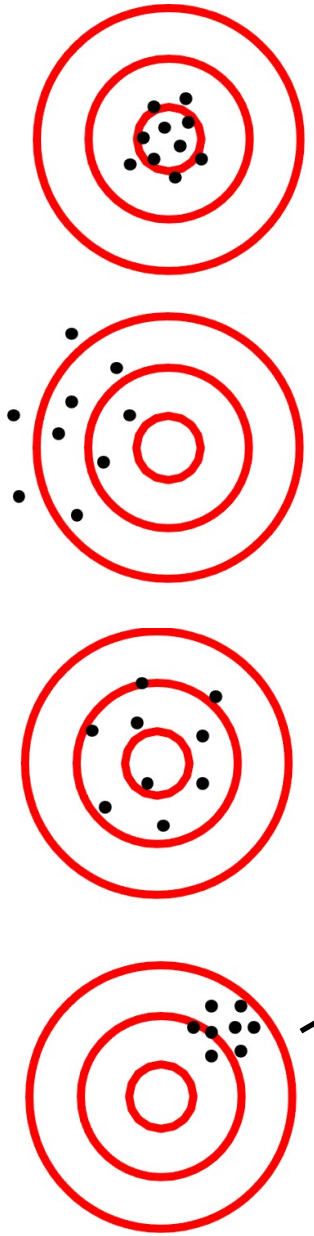
a. Low V, low R

b. Low V, high R

c. Low R, high V

d. High R, high V

Validity / Reliability



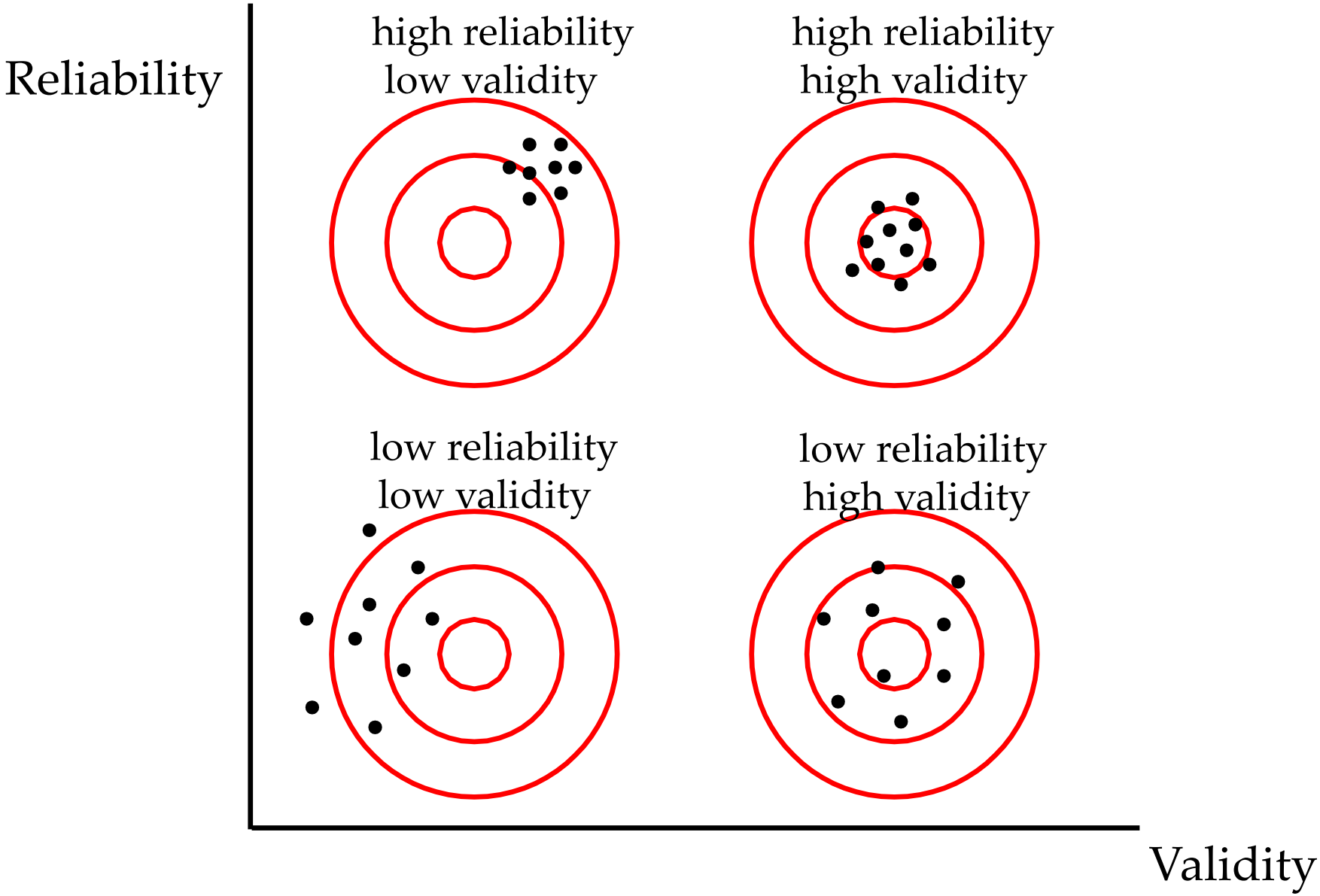
a. Low V, low R

b. Low V, high R

c. High V, low R

d. High R, high V

Validity / Reliability



Example: Likert Scale Measurements

- Having multiple items (questions) improves reliability

Table 2: Muir's questionnaire.

1. To what extent can the robot's behavior be predicted from moment to moment?
2. To what extent can you count on the robot to do its job?
3. What degree of faith do you have that the robot will be able to cope with similar situations in the future?
4. Overall how much do you trust the robot?

User Study Execution

- Informed Consent
- Script
  - We need to provide clear, specific instructions of what we want participants to do
- Study Execution
  - Audio and record as much as possible since unexpected / surprising events often lead to new hypotheses
- Subjective Responses / Objective Measures
  - Likert Scale
  - Open-ended Responses
  - Ask the same questions in different ways (positive / negative)
- Debriefing
  - If there was deception, correct their impression

IRB Approval

- Stanford prison experiment

- Implementation of rules to preclude any harmful treatment of participants

- Potential benefit for science outweighs the possible risk for physical and psychological harm.

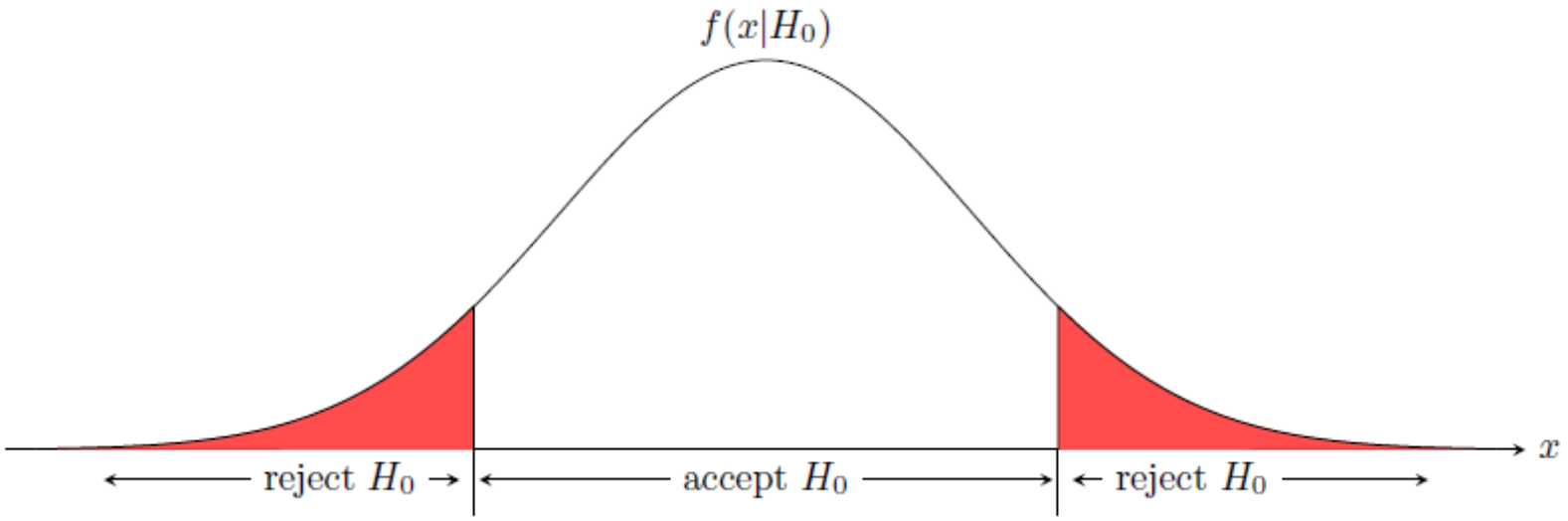
Statistical Analysis

- H_0: the *null hypothesis*. This is the default assumption for the model generating the data

- H_A: the *alternative hypothesis*. If we reject the null hypothesis we accept this alternative as the best explanation for the data

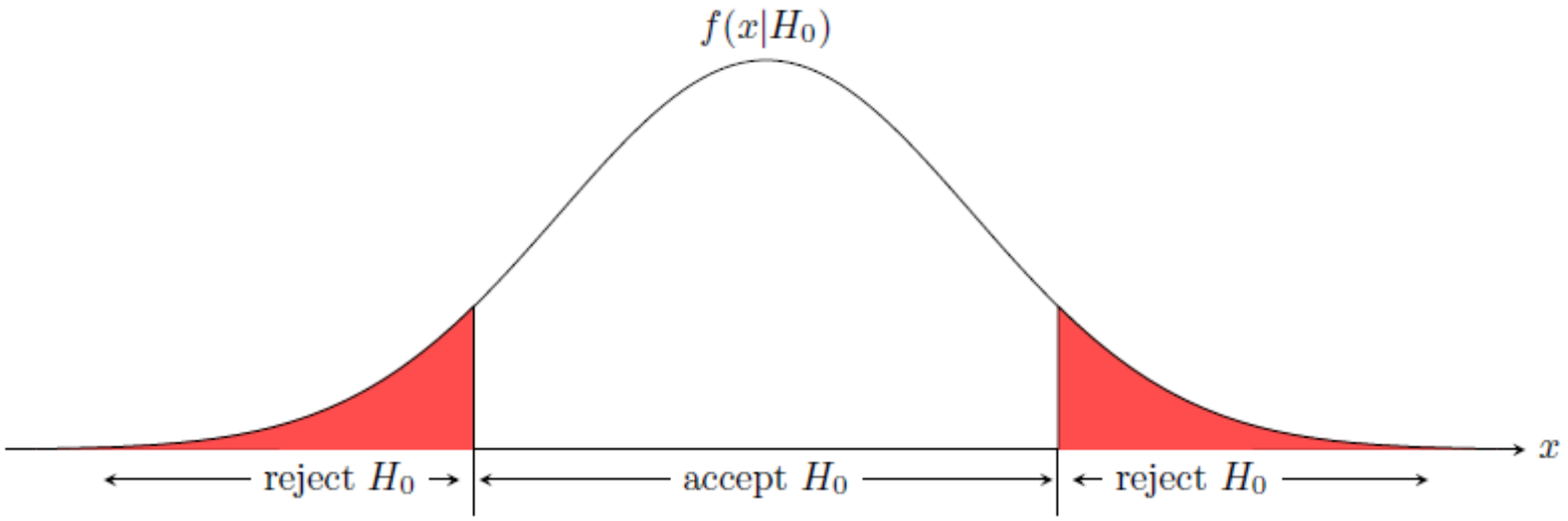
Statistical Analysis

- Rejection region: if x is in the rejection region, we reject H_0 in favor of H_A
- *Significance level*: Probability we incorrectly reject the null hypothesis (typically equal to 0.05)



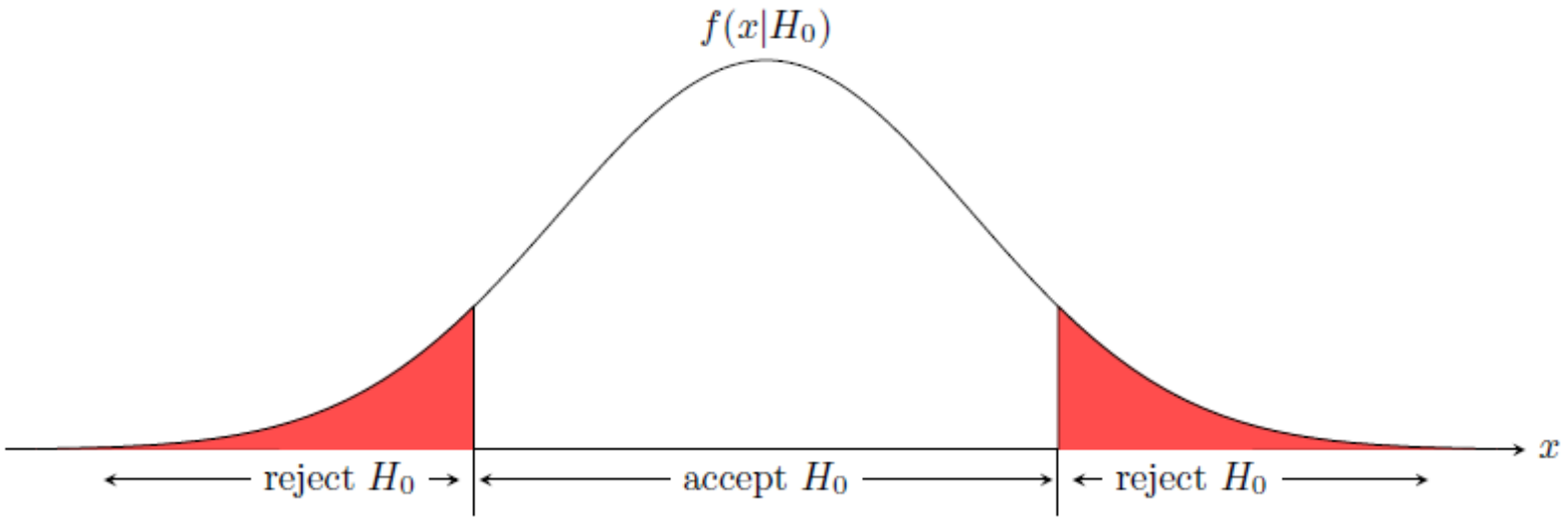
Statistical Analysis

- Rejection region: if x is in the rejection region, we reject H_0 in favor of H_A
- *Significance level*: Probability we incorrectly reject the null hypothesis (typically equal to 0.05)
- p value: Probability, assuming the null hypothesis, of seeing data at least as extreme as the experimental data.



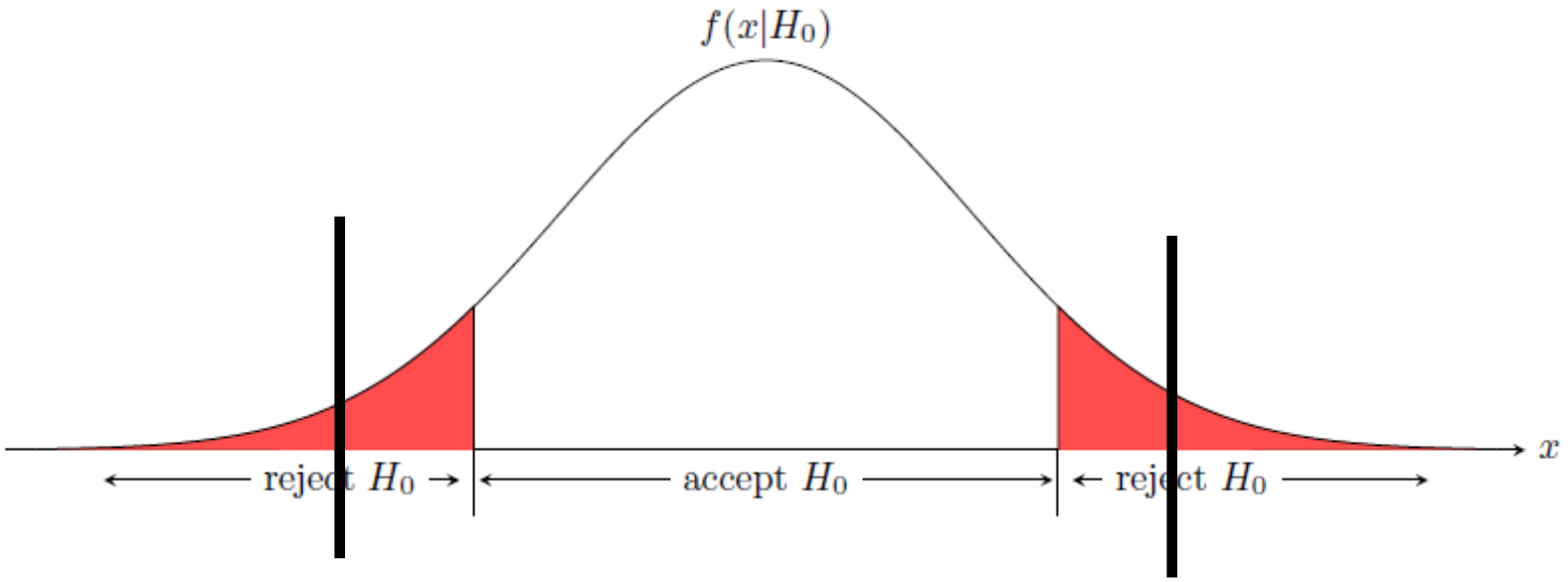
Statistical Analysis

- Rejection region: if x is in the rejection region, we reject H_0 in favor of H_A
- *Significance level*: Probability we incorrectly reject the null hypothesis (typically equal to 0.05)
- p value: Probability, assuming the null hypothesis, of seeing data at least as extreme as the experimental data.



Statistical Analysis

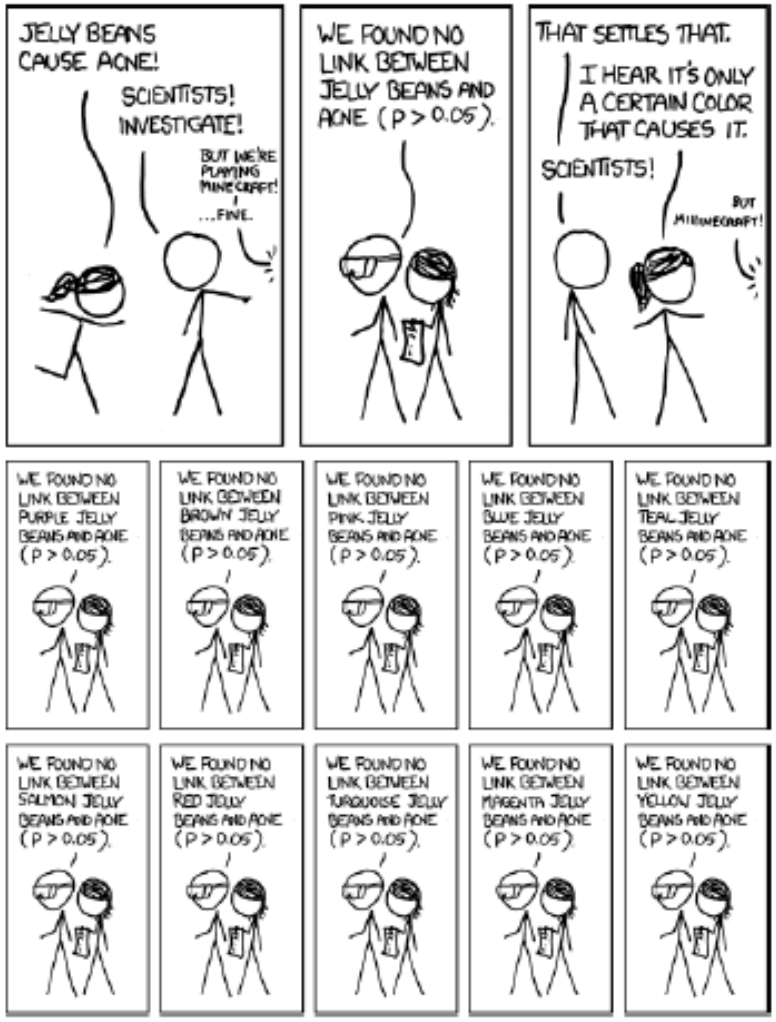
- Rejection region: if x is in the rejection region, we reject H_0 in favor of H_A
- *Significance level*: Probability we incorrectly reject the null hypothesis (typically equal to 0.05)
- p value: Probability, assuming the null hypothesis, of seeing data at least as extreme as the experimental data.

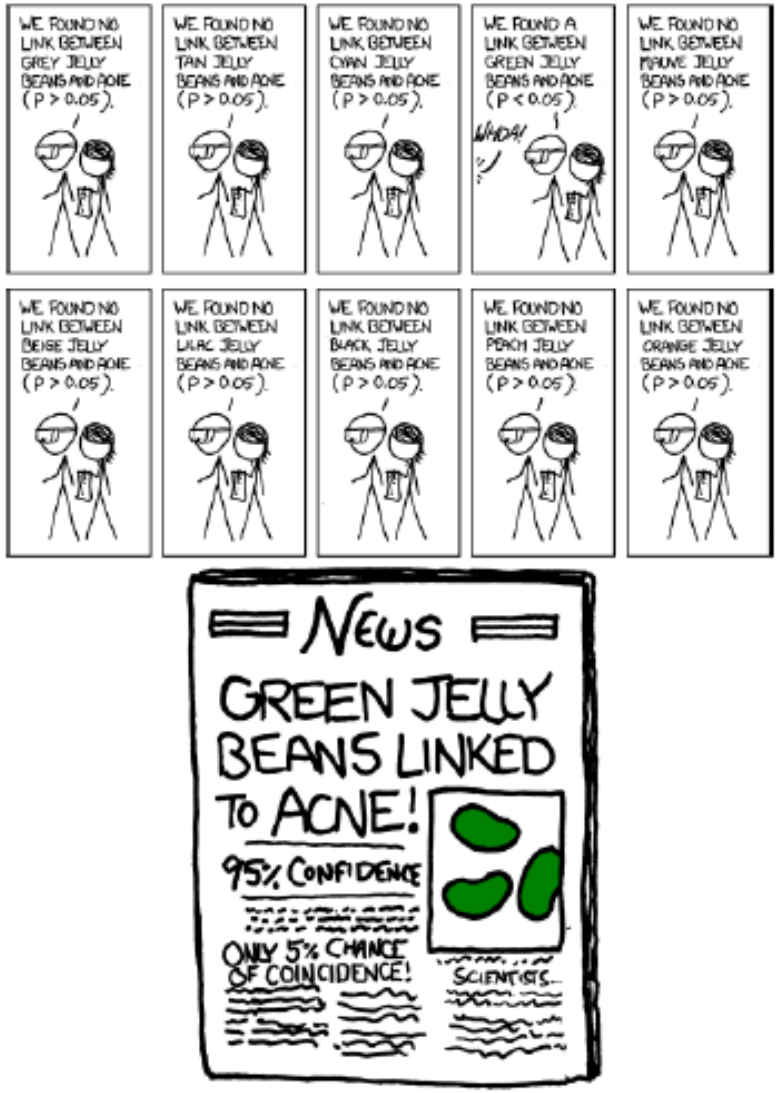


Errors

- Type I error: false detection of an effect that is not present
- Type II error: failure to detect an effect that is present

Type I Error





http://xkcd.com/882/

Type I error

$P(\geq 1 \text{ error in } 20 \text{ comparisons}) =$

Type I error

$$P(\geq 1 \text{ error in } 20 \text{ comparisons}) = 1 - P(\text{no error})$$

Type I error

$$P(\geq 1 \text{ error in } 20 \text{ comparisons}) = 1 - P(\text{no error})$$
$$= 1 - 0.95^{20} = 0.64$$

Type I error

$$P(\geq 1 \text{ error in } 20 \text{ comparisons}) = 1 - P(\text{no error})$$
$$= 1 - 0.95^{20} = 0.64$$

How to reduce the change of Type I error?

Type I error

$$P(\geq 1 \text{ error in } 20 \text{ comparisons}) = 1 - P(\text{no error})$$
$$= 1 - 0.95^{20} = 0.64$$

How to reduce the change of Type I error?

Bonferroni Correction

Ordinal Data

- “The robot is trustworthy”

1. strongly disagree
2. disagree
3. neither agree nor disagree
4. agree
5. strongly agree

Ordinal Data

- “The robot is trustworthy”

1. strongly disagree
2. disagree
3. neither agree nor disagree
4. agree
5. strongly agree

(disagree + agree) / 2 = (neither + neither) / 2

Common Tests in HRI

https://stats.idre.ucla.edu/other/mult-pkg/whatstat

Independent variables	Type of dependent variables	Test
1 (2 levels, between subjects)	Interval and normal	Unpaired two-tailed t-td t-test
	ordinal	Wilcoxon-Mann Whitney test
1 (2 levels, dependent groups)	Interval and normal	Paired two-tailed t-test
	ordinal	Wilcoxon signed rank test
1 (multiple levels, dependent)	Interval and normal	one-way repeated measures ANOVA
Multiple independent variables	Interval and normal	factorial ANOVA