Analysis I: Data wrangling, descriptive statistics, t-tests

PNI Summer Internship 2020

Mai Nguyen

Overview

- Introduction to today's data set
- Data wrangling & getting to know your data
- Gaussian distribution & standard normal distribution
- Describing a distribution
- t-tests: compare one or two groups
- (ANOVA: compare more than two groups)

Today's data set

- Shared by Nina Rouhani in Yael Niv's lab at Princeton (recently defended May 2020!)

Journal of Experimental Psychology: Learning, Memory, and Cognition

Dissociable Effects of Surprising Rewards on Learning and Memory

Nina Rouhani, Kenneth A. Norman, and Yael Niv

Online First Publication, March 19, 2018. http://dx.doi.org/10.1037/xlm0000518

CITATION

Rouhani, N., Norman, K. A., & Niv, Y. (2018, March 19). Dissociable Effects of Surprising Rewards on Learning and Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000518

Today’s data set

- How does **prediction error** affect **learning**?
- Prediction error: difference between what you expect and what you get

Expect

Actual

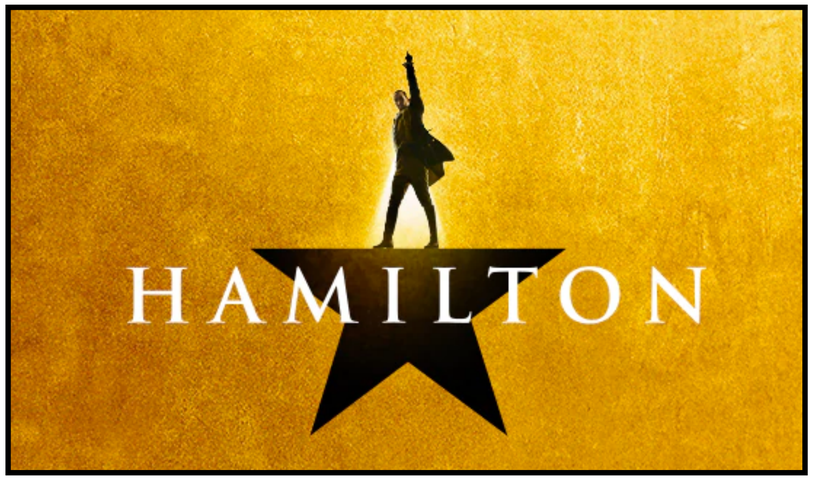
Negative prediction error

Today’s data set

- How does **prediction error** affect **learning**?
- Prediction error: difference between what you expect and what you get

Nothing

Expect

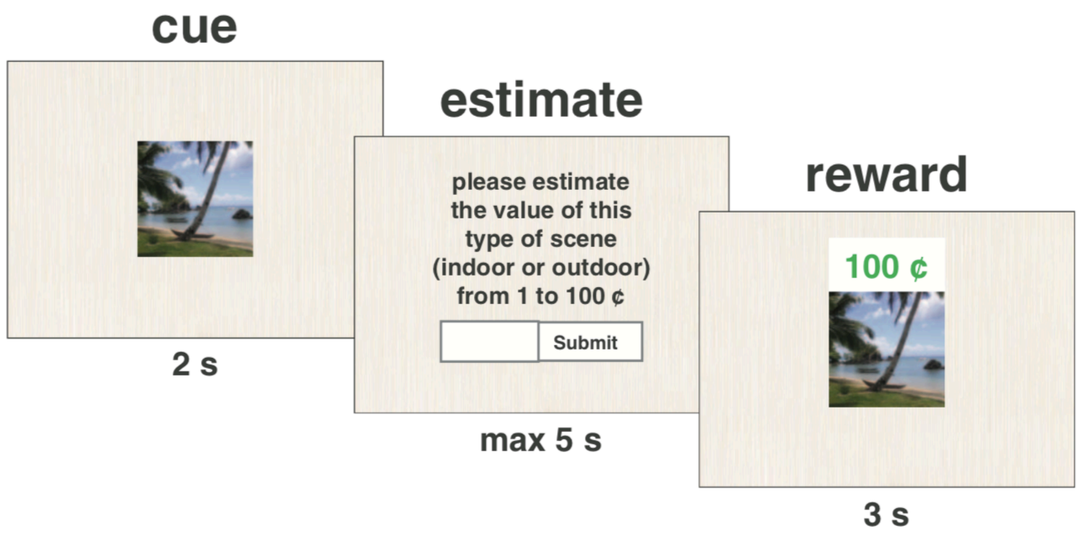


Actual

Positive prediction error

Today's data set

LEARNING room 1



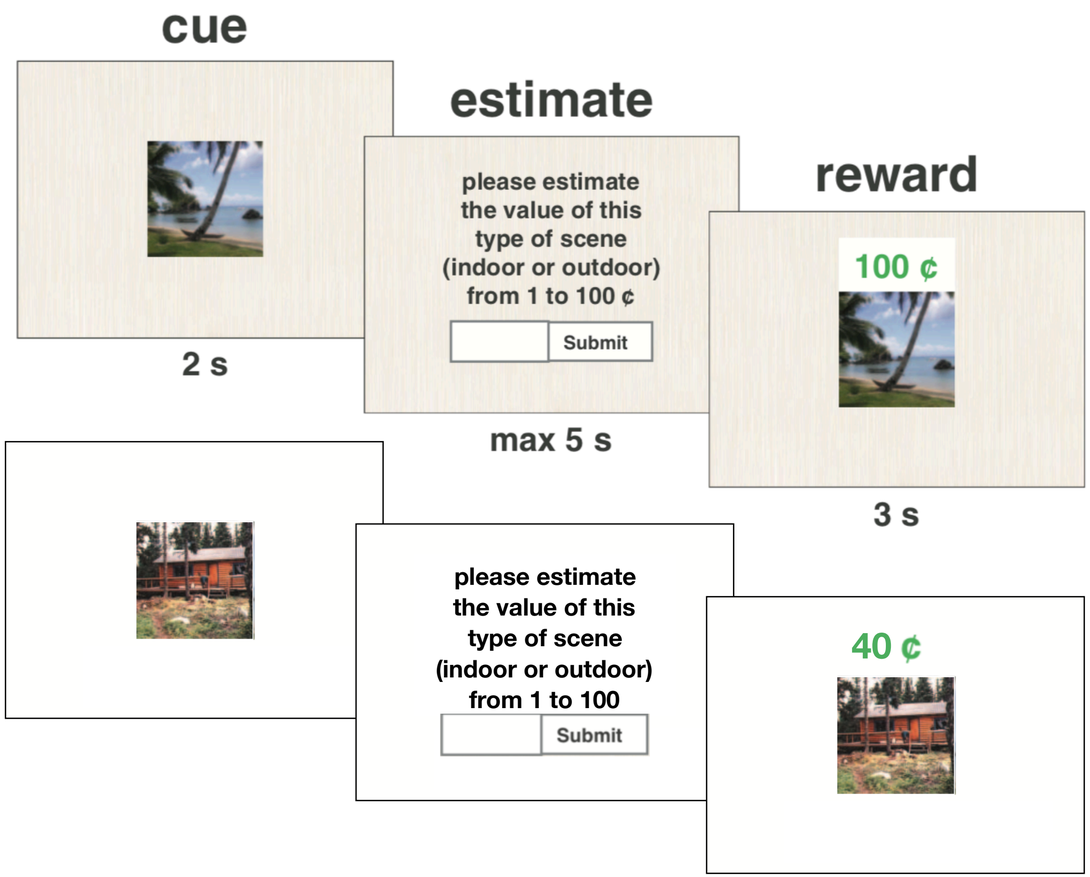
- View images of indoor and outdoor scenes
- Estimate value of scene category, get feedback -> **prediction error**
- Update value estimate -> **learning**

- Each scene varies in value, but the **average value of scene categories** differ:
  - high value: 60 cents
  - low value: 40 cents

Today's data set

LEARNING room 1

LEARNING room 2



- View images of indoor and outdoor scenes
- Estimate value of scene category, get feedback -> **prediction error**
- Update value estimate -> **learning**

- Each scene varies in value, but the **average value of scene categories** differ:
  - high value: 60 cents
  - low value: 40 cents

- Scenes appear in two different rooms which vary in **risk** or **variance**

Today’s data set

- N = 136 participants
- 30 trials per room, 2 rooms (high vs low risk)

Today’s data set

- Questions:
    - Are people learning in either High Risk or Low Risk condition?
    - Does prediction error differ in High Risk versus Low Risk condition? By High vs Low value?
    - Does learning rate differ by High vs Low risk? By High vs Low value?

Data wrangling

- Process of getting data into a useable and useful state
    - Reading into programming env
    - Organizing, structuring
    - Cleaning
    - Enriching/processing
- I’ve been doing this for you for the most part, but we’ll do a little bit of this today (the CSV file you have has already quite a bit of data wrangling in the background)

Data Wrangling

- Steps prior to today:
    - Collate individual subject data files into single file
    - Remove subjects who don't finish the experiment
    - Remove subjects who don’t pass attention checks
    - Insert NaNs for non-responses
    - Calculate learning rate, prediction error and save to file
    - Mai: recode condition values for MATLAB, export to CSV

Data wrangling

1. Can you even open it?

2. Read documentation. What do the different numbers mean? How is the data organized already?

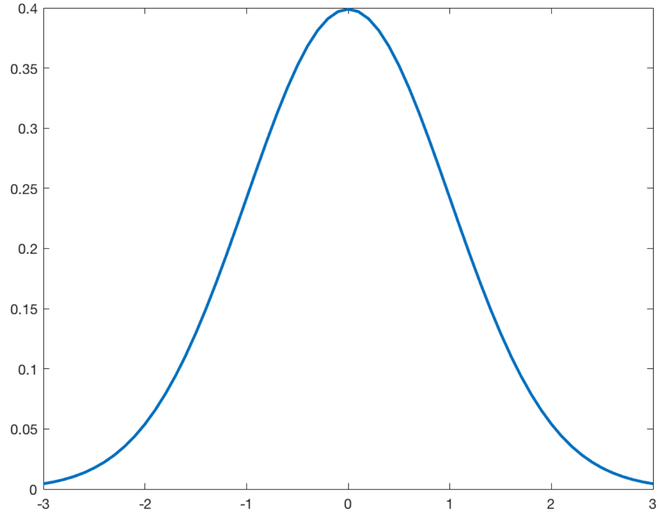
3. Read into MATLAB. Familiarize yourself with the data structure in MATLAB and check against any outside sources (e.g. against what you see in excel)

4. Should the data be reorganized?

5. Get to know your data: histograms, descriptive statistics, etc

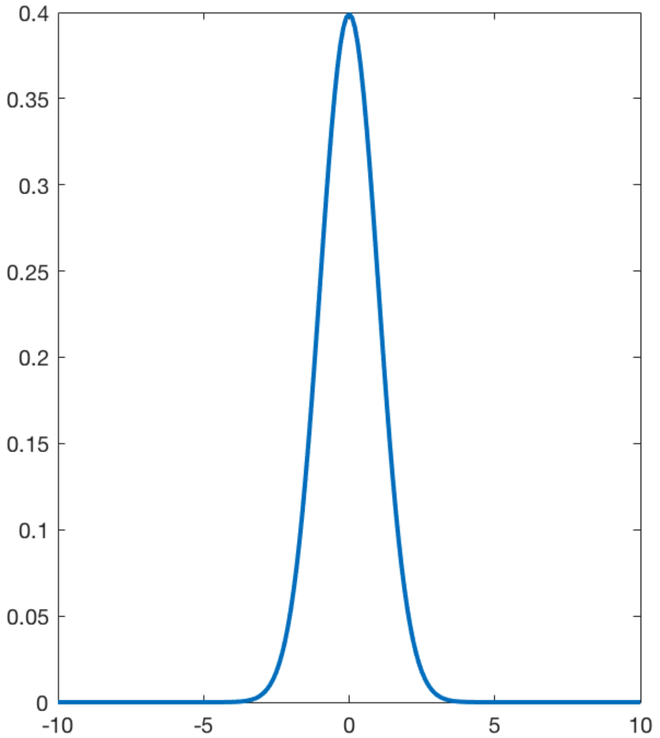
Gaussian distribution

- Also known as: “normal” distribution or “bell curve”
- Defined by mean and standard deviation:
  - mean = (sum of all values) / (# of values) = $\frac{\sum x}{N} = \mu$
  - standard deviation = square root of [(difference from mean)^2 / (# of values)] = $\sqrt{\frac{\sum (x - mean)^2}{N}} = \sigma$

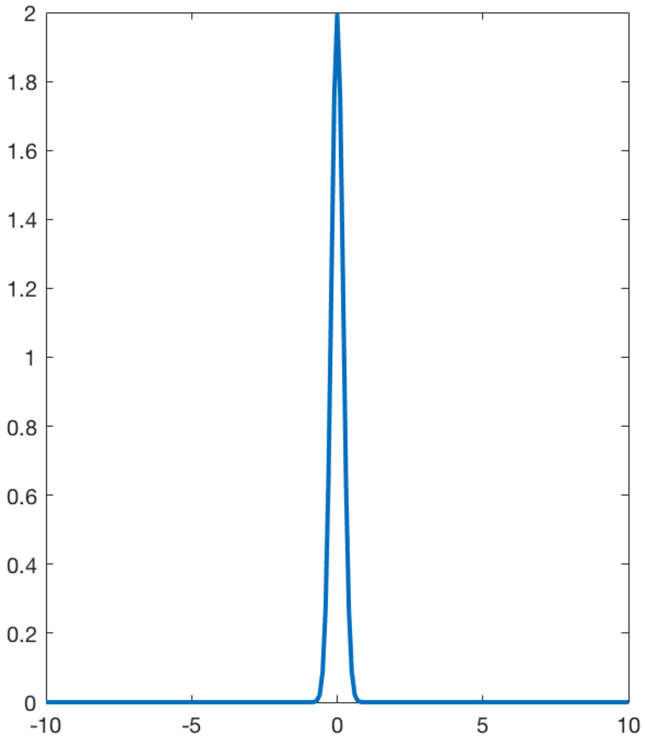


Gaussian distribution

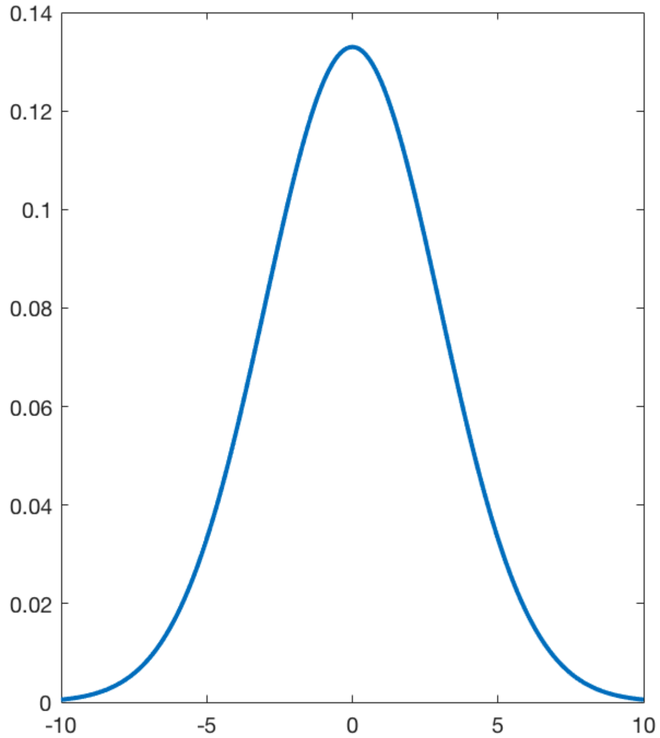
Standard normal



mean = 0
std = 1



mean = 0
std = .25

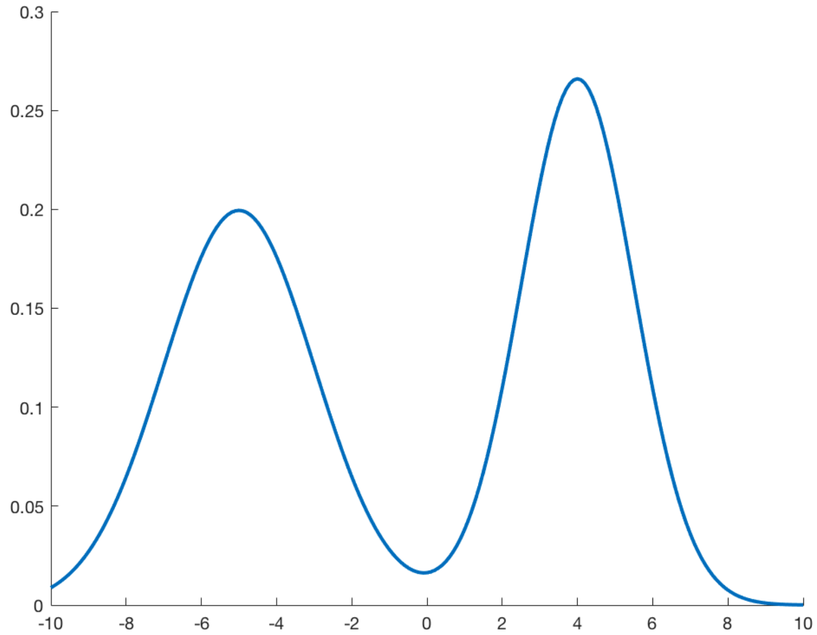


mean = 0
std = 3

Gaussian distributions

- Important because a lot of measurements in psych/neuro are normally distributed (but not all!)

- Often describe data in terms of parameters of normal distribution (this isn’t always appropriate!)



Gaussian distributions

- Important because a lot of measurements in psych/neuro are normally distributed (but not all!)
- Often describe data in terms of parameters of normal distribution (this isn’t always appropriate!)
- Standard parametric stats assume data is normally distributed

Descriptive statistic

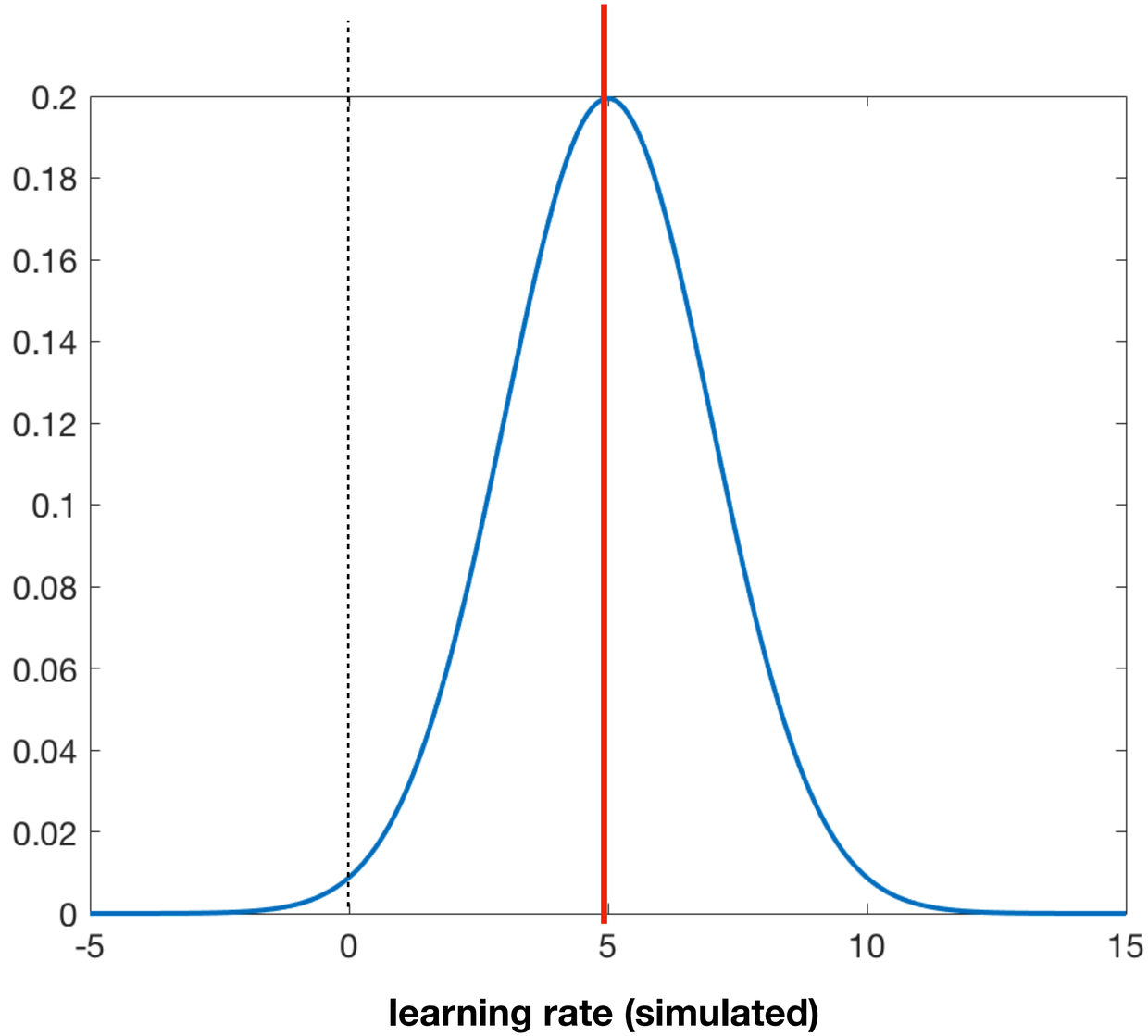
- Mean, median
- Standard deviation
- **Skew, kurtosis**

One sample t-test: is this thing different from 0?

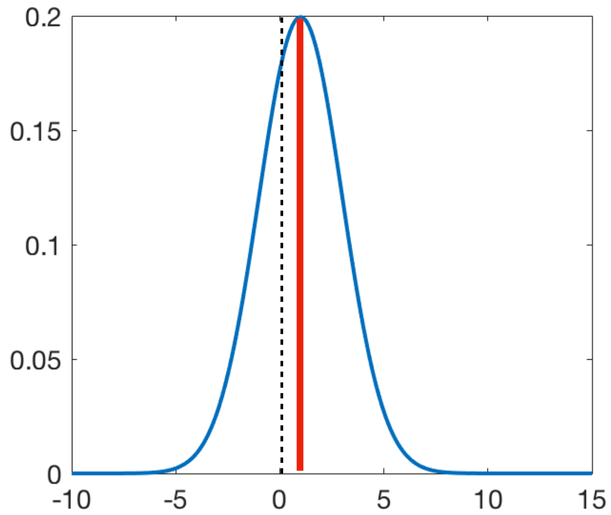
- Are people learning e.g. is the learning rate different from 0?

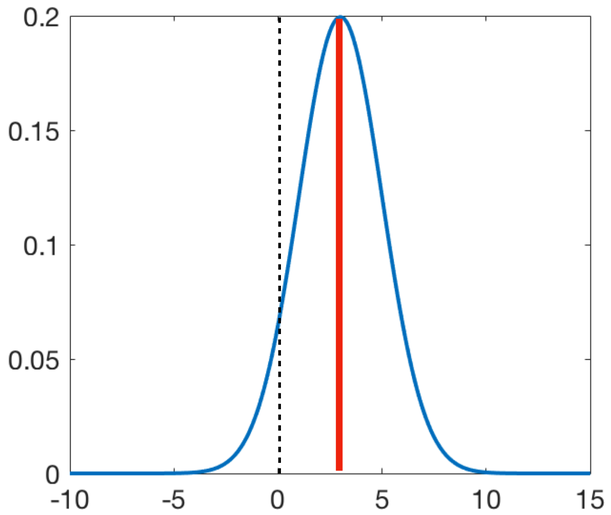
$$t = \frac{\bar{x}}{s/N}$$

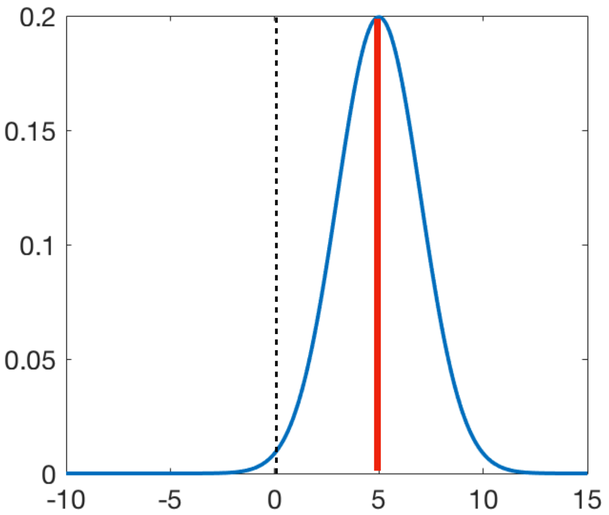
$$t = \frac{\textbf{mean}}{\textbf{measure of variance}}$$



One sample t-test: is this thing different from 0?





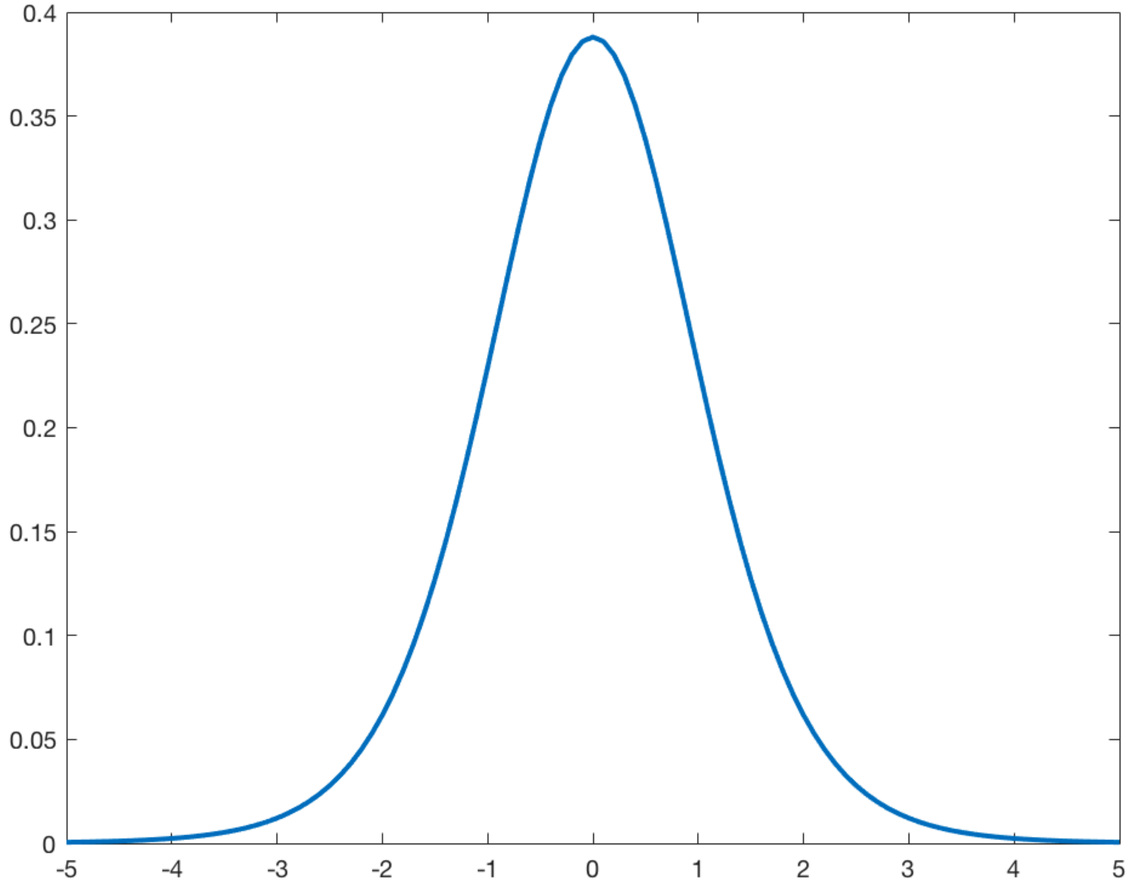


We have a t-stat. Now what?

- Null hypothesis: this thing is not different from 0
- Alternate hypothesis: this thing is different from 0
- **Null hypothesis testing:** if can reject the null hypothesis, then we accept the alternate hypothesis
    - **t-distributions** and **p-values**

We have a t-stat. Now what?

- **t-statistic** is assumed to be drawn from the **t-distribution**
  - Assume null hypothesis is true
  - Repeat exp 10,000 times and calc t-stat each time
  - Resulting distribution of t-stats approximates the t-distribution
- We can't run the experiment 10,000 times, so assume t-stat is from t-distribution



t-distribution with 9 degrees of freedom

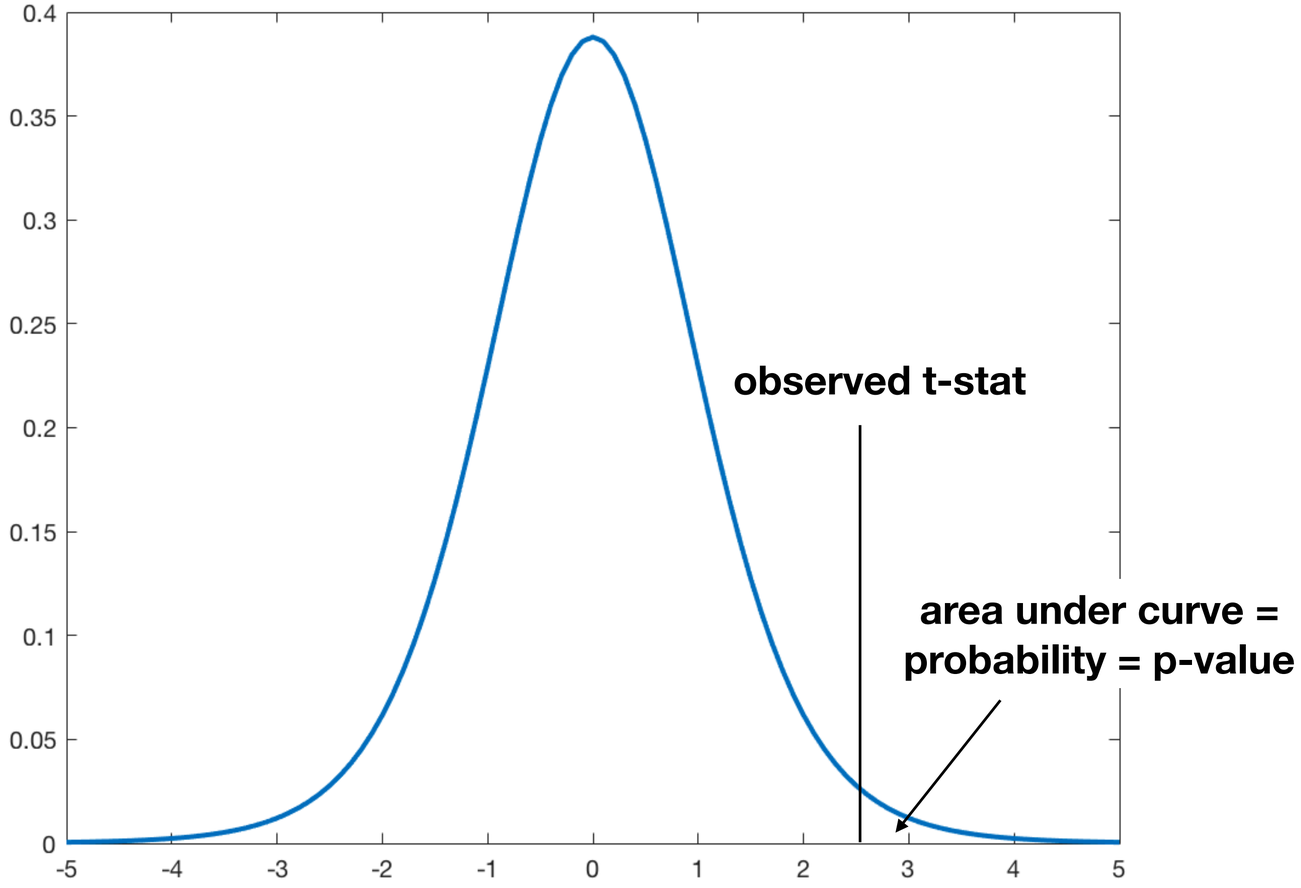
We have a t-stat. Now what?

- **p-value**: under the null hypothesis, how probable is the observed t-statistic?

- By convention:

  - $p < .05$ = significant = publishable = fame and glory

  - $p = .055$ = insignificant = your numbers are bad and you should feel bad



t-distribution with 9 degrees of freedom

One sample t-test: is this thing different from 0?

- `ttest(data)`
- Is the learning rate different from 0 in the low risk and high risk condition?

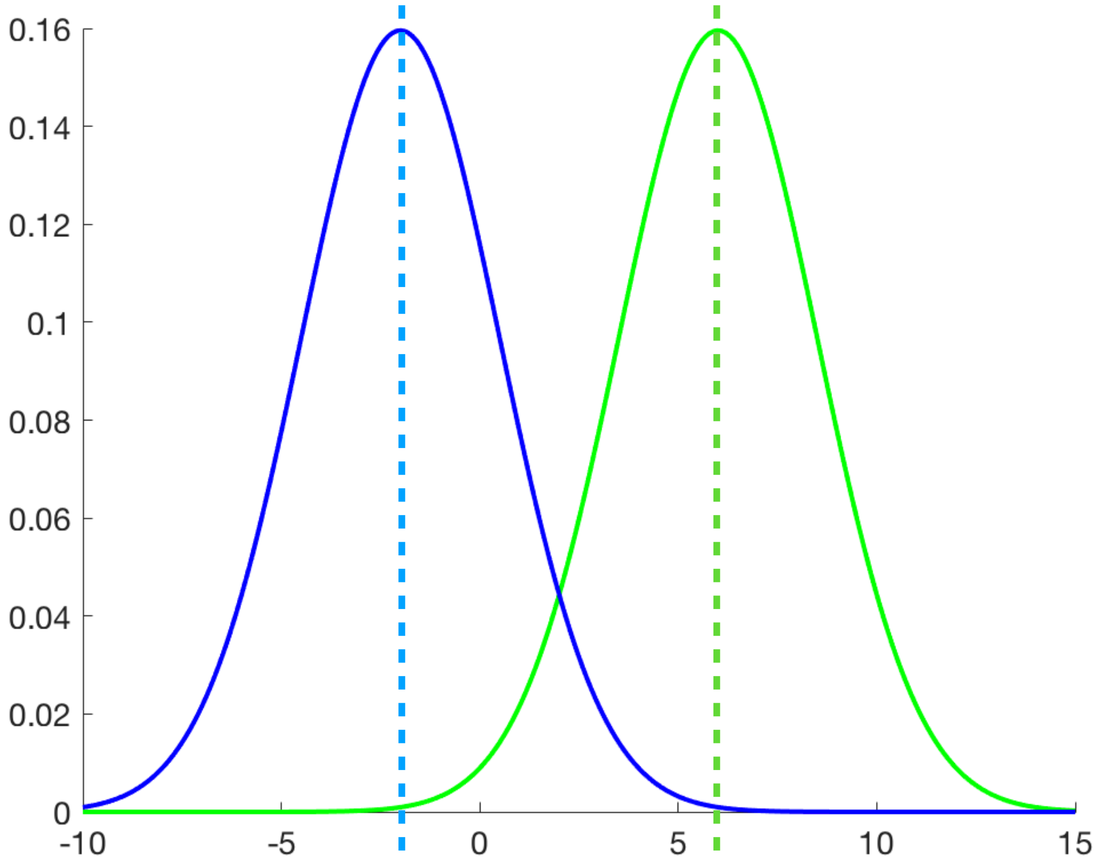
Two-sample t-test: is this thing different from this other thing?

- Two-sample t-tests compare two sets of data
- **Independent two-sample t-tests**: compare two **independent** sets of data
    - e.g. are basketball players taller than jockeys?
    - e.g. do engineering majors have lower GPAs than humanities majors?

Independent two-sample t-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

$$t = \frac{\textbf{difference in means}}{\textbf{measure of variance}}$$



Paired two-sample t-test

- **Paired two-sample t-tests**: compare data collected from the **same** subjects under different conditions or points in time
    - e.g. do students improve their scores before and after training?
    - e.g. do subjects do better in condition1 or in condition2?

- Actually just a one-sample t-test on difference between conds

Two-sample t-test: is this thing different from this other thing?

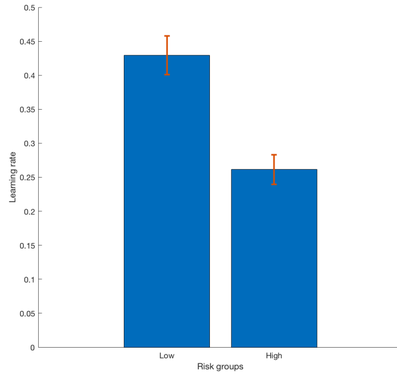
- Independent t-tests: `ttest2(data1, data2)`
- Paired t-tests: `ttest(data)`
- For comparing learning rate data in high-risk versus low-risk condition, should we use independent or paired t-tests?

Write up and visualize your results

- Make a figure - what makes a good figure?
    - Summarizes the relevant data, including error bars if appropriate
    - Clearly labeled axes, titles, and legends (as appropriate)
    - Large enough font and labels to be readable
    - Doesn’t include extraneous information or distracting embellishments
- Exercise: visualize the learning rate results from the last section

Write up and visualize your results

- By convention report statistics: statisticType(degrees of freedom) = statistic value, p = pValue
    - e.g. t(99) = 3.2, p = .032
- Describe your figure using plain English and citing the relevant numbers
    - BAD: Because t(270) = 4.7 and p < .05, we rejected the null hypothesis.
    - GOOD: The learning rate was significantly faster in the low-risk condition than the high-risk condition (t(270) = 4.7, p<.001).



Python aside

- scipy.stats library:
  - `scipy.stats.ttest_1samp()`: one sample t-test
  - `scipy.stats.ttest_ind()`: independent, two-sample t-test
  - `spicy.stats.ttest_rel()`: paired two-sample t-test

Aside: what if you have more than 2 groups?

- Use ANOVA = analysis of variance
- Recommendation: do ANOVAs in R
- If you must:
    - MATLAB: anova1(data), anova2(data, reps)
    - Python: scipy.stats or statsmodel package

One-way ANOVA: are these 3+ things different?

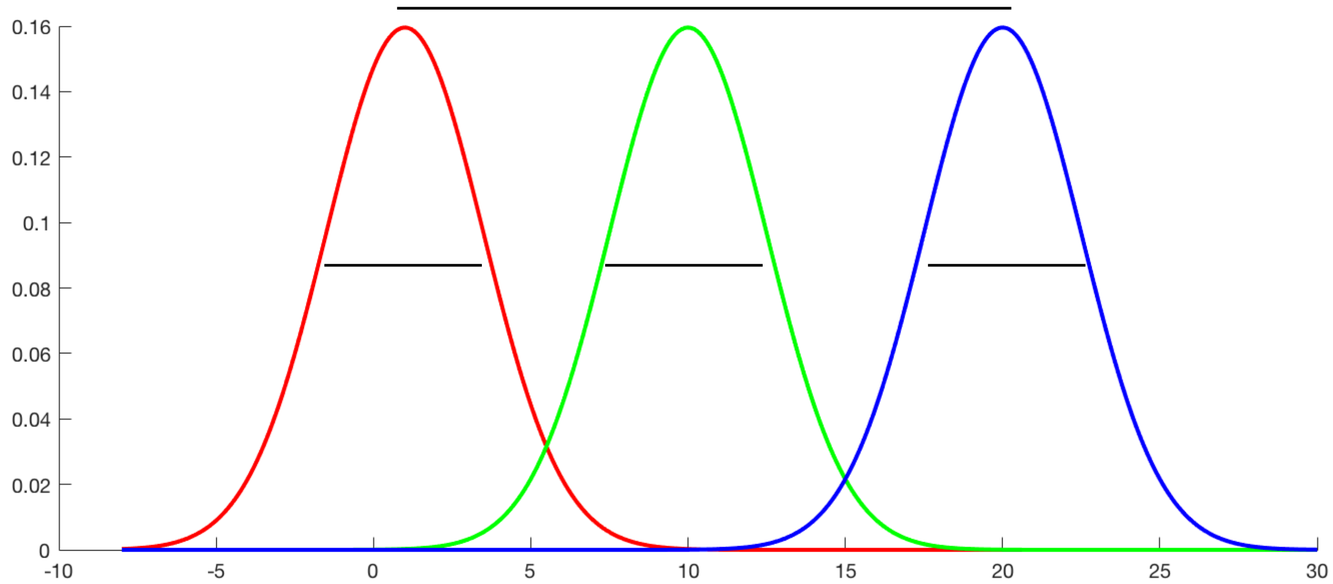
- Often times there are more than two things in the world. How do you compare more than 2 things?
    - e.g. Do students learn better entirely online, entirely in-person, or in a mixture of online and in-person?
    - e.g. How does income 5 years after graduation differ for humanity, science, engineering, or art degrees?
- **One-way ANOVA = analysis of variance**

One-way ANOVA: are these 3+ things different?

$$F = \frac{MS_B}{MS_E}$$

$$MS_B = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N - 1}$$

$$MS_E = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{N - 2}$$



$$\mathbf{F = \frac{\text{variance between groups}}{\text{variance within groups}}}$$

One-way ANOVA: are these 3+ things different?

- `[p, stats] = anova1(data)`

Two-Way ANOVA: are these things different from each other and do they differ in how they differ?

- Multiple independent variables might affect the dependent variable
- In Nina's study:
    - Independent variable 1: risk, high versus low
    - Independent variable 2: reward, 60 cents vs 40 cents

Two-way ANOVA

- **Main effects:** how does the IV by itself affect the DV?
    - Basically a one-way ANOVA
    - Questions: How does risk affect learning rate? How does reward affect learning rate?
- **Interaction effects**: how does the effect of one IV depend on the effect of the other IV?
    - Questions: Does the effect of reward on learning rate depend on the risk level?

Two-way ANOVA