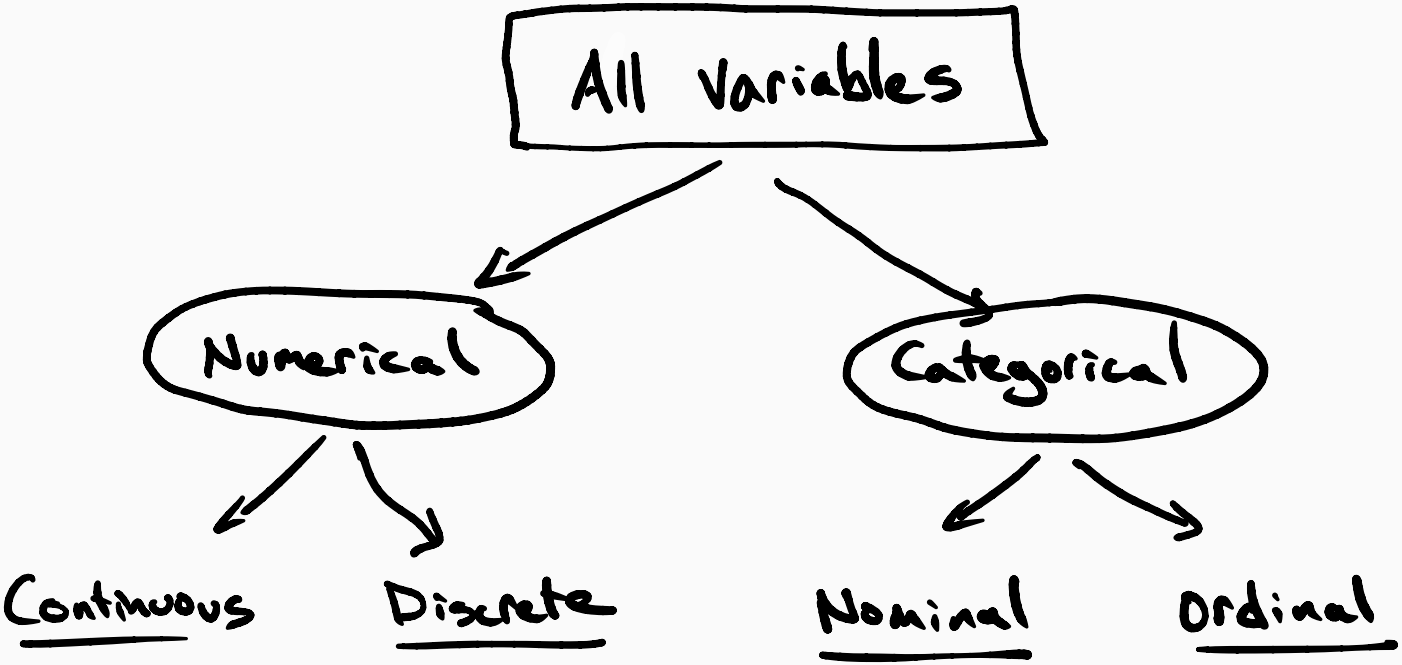
# Taxonomy of Inference

# Recall: Taxonomy of Data



# Taxonomy of Inference

The method of analysis should be driven by:

1. Research Question
2. Data

Let

- $X$: 2-level categorical variable
- $Y$: >2-level categorical variable
- $W$: Numerical (continuous) variable

# Categorical Data

visualize: bar plots

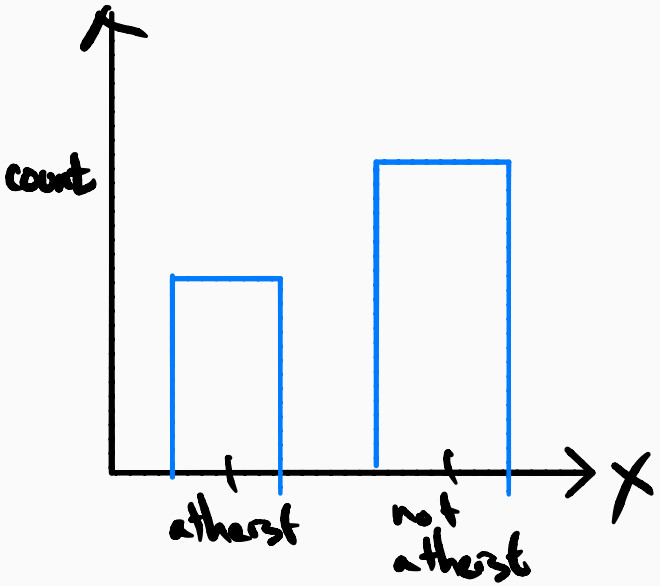
parameters of interest: $p$

| breakfast |
| --- |
| "cold cereal" |
| "cold cereal" |
| "pancakes" |
| ⋮ |

# $X$: One 2-level categorical variable

E.g. Ask a person if they are "atheist" or "not atheist".

Visualize: simple bar plot

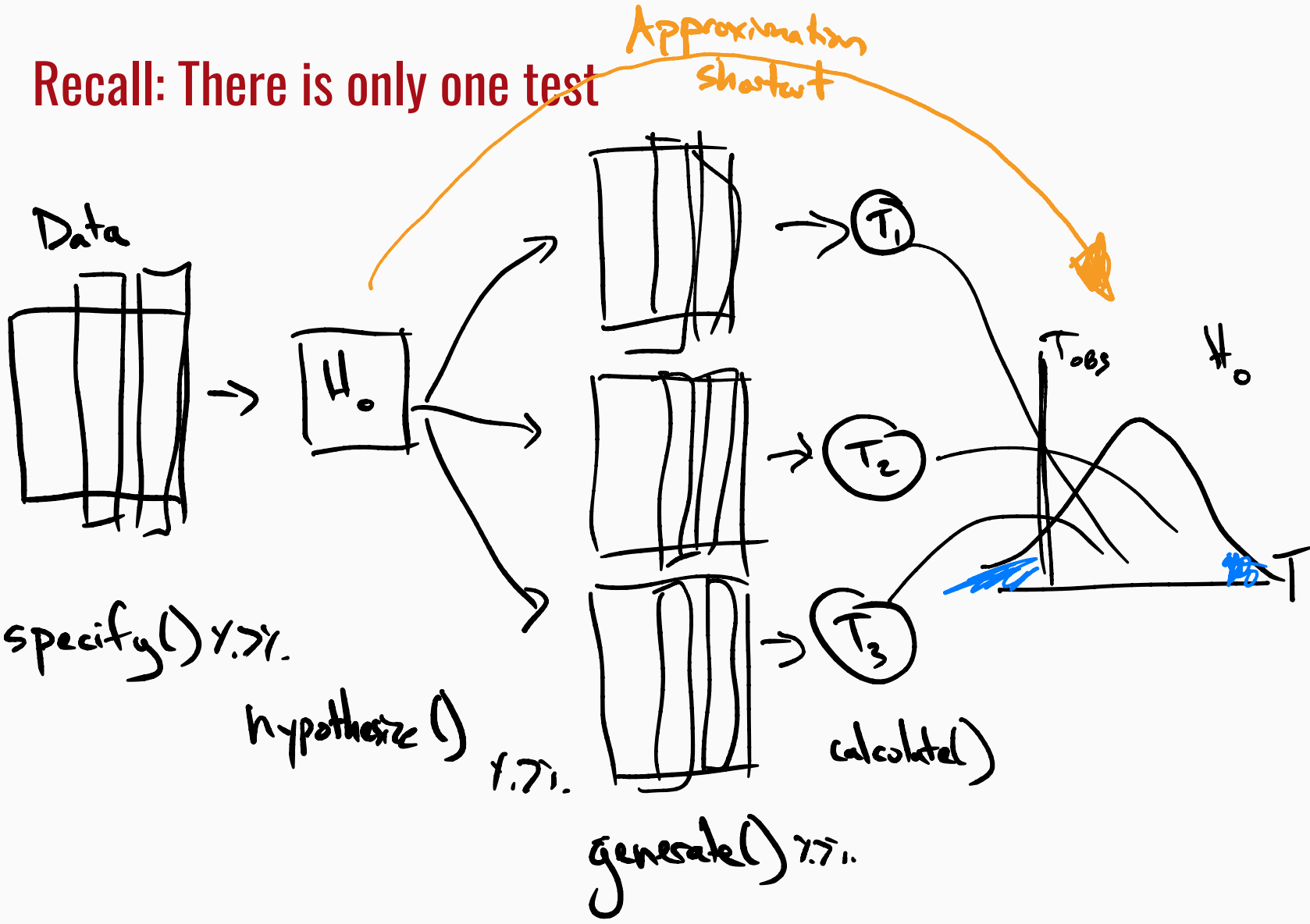



Parameter of interest: $p$: proportion of "successes"

Sampling Dist. of $\hat{p}$:

- Exact: Binomial
- Simulation
- $\frac{\hat{p} - p_0}{SE} \sim \text{Normal}(0,1)$ if $np \geq 10$ and $n(1-p) \geq 10$

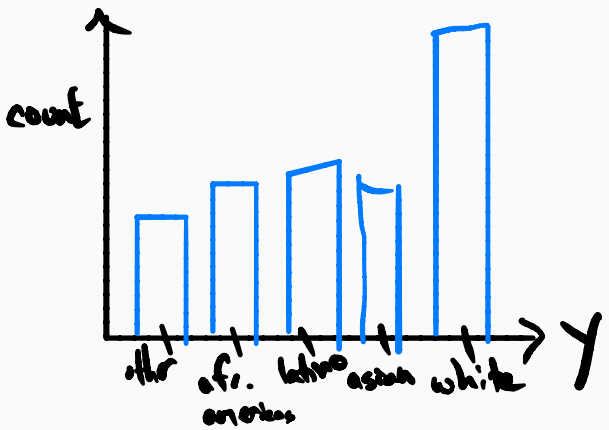
# Recall: There is only one test



# $Y$: One > 2-level categorical variable

E.g. Ethnicity of Reed's First year class.

visualize: barplot




parameters of interest: $p_1, p_2, \ldots p_k$

Inferential Technique:

Goodness of Fit test.

→ calculate $\chi^2$ stat

$H_0: p_1 = p_1^0, p_2 = p_2^0 \ldots, p_k = p_k^0$

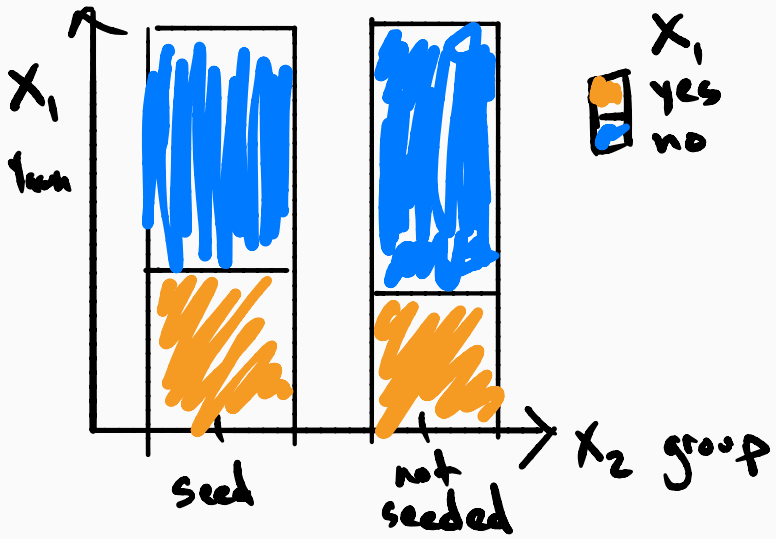
Get null dist of $\chi^2$ using

- Simulation
- $\chi^2$ dist df: $k-1$

# $X_1 \sim X_2$: Two 2-level categorical variables

Eg: Yawning and being in the seeded groups.

Visualize: barplot




parameters of interest: $p_1 - p_2$ ← diff in props of people in 2 group that yawned.

Null dist of $\hat{p}_1 - \hat{p}_2$:

- permute
- $\frac{\hat{p}_1 - \hat{p}_2 - (p_1^* - p_2^*)}{SE} \sim N(0,1)$

if - $np \geq 10$ - ind. obs.
- $n(1-p) \geq 10$

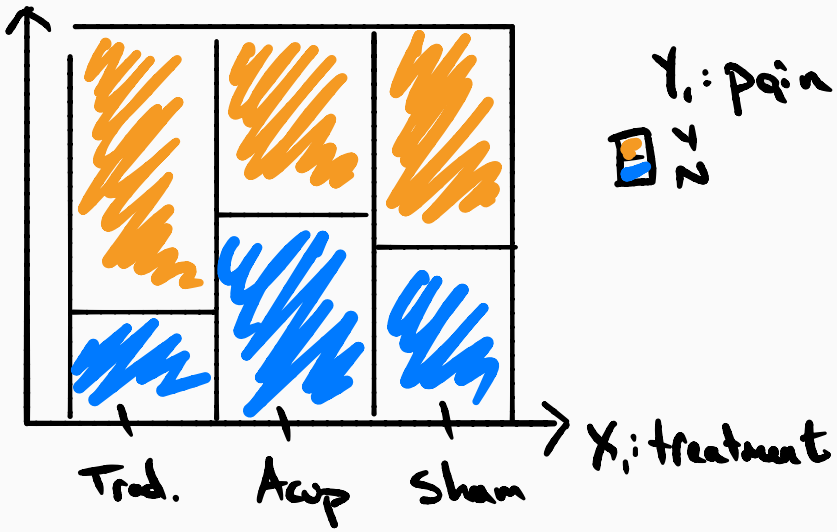
for both groups.

# $Y_1 \sim Y_2$: Two > 2-level categorical variables

$Y_1 \sim X_1$

Eg: Does Acupuncture relieve pain?

Visualized bar plot




Inferential Test of Independence

Question:

- Use $\chi^2$ statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E}$$

np: G.f

p(A,B) = P(A)P(B)

- Find null:
  - permutation
  - $\chi^2$ dist df: (R-1)(C-1)

# Numerical Data

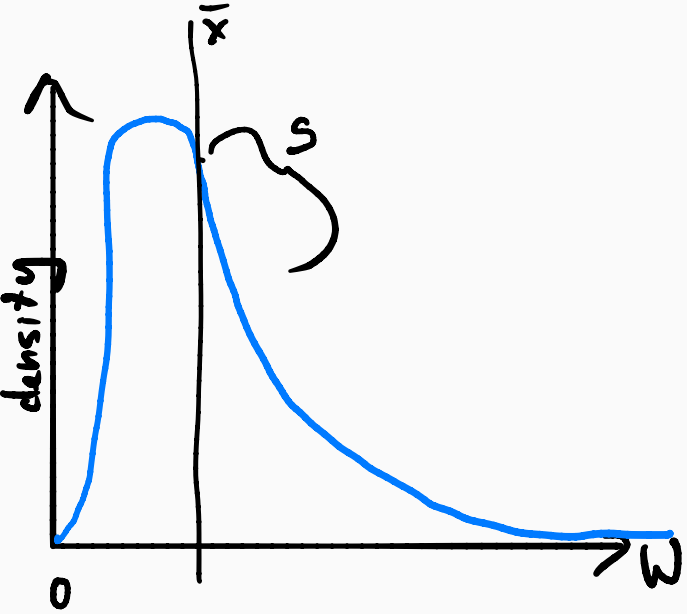
- visualize: - boxplot
  - density
  - histogram
- parameters $\mu$
  $\sigma$
  $q_{.7}$
  max
  .
  .
  .

| weight in kg |
| --- |
| 1.035 |
| .917 |
| .984 |
| . |
| . |
| . |
| . |

# $W$: One numerical variable

E.g. Individual incomes in U.S.

Visualize: density, histogram, boxplot.




parameter of interest: $\mu$

sampling dist of $\bar{x}$: $\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t_{df=n-1}$ if - ind obs.
- W being nearly normal

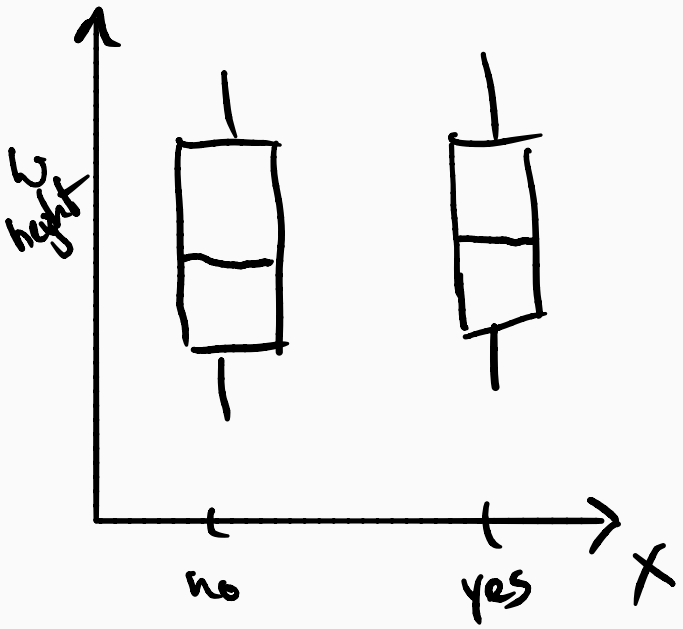
CLT

- as n increases the "nearly normal" assumption can be relaxed.

# $W \sim X$: One numerical, one 2-level categorical variable

E.g. Does average height differ between physically active and not "active" teens?

visualize: side by side boxplot




parameter of interest $\mu_1 - \mu_2$

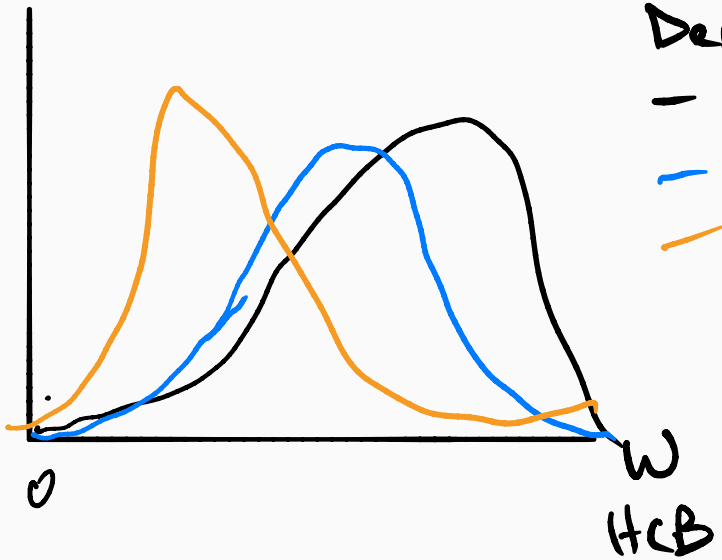
sampling dist $\dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t_{df = \min(n_1 - 1, n_2 - 1)}$

if - ind obs.
- W nearly normal

# $W \sim Y$: One numerical, one > 2-level categorical variable

E.g: HCB and Wolf River Depth.

Visualize: Overlaid density plots.




Depth
- Bottom
- Mid
- Surface

Inferential Question: $H_0: \mu_1 = \mu_2 = \mu_3$

Test Statistic: $F = \frac{MSG}{MSE}$

Null Dist of F

1. Permute
2. F-dist
   - ind obs.
   - normal in each group
   - constant variance