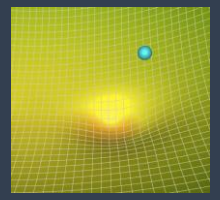
What can I say about ...	 FITTING	 HOLDOUT	 BAYESIANISM	 FREQUENTISM
... which model to choose?	≈	👍	👍	
... parameter values?	≈	≈	👍	Confidence intervals: canvas estimates across the multiverse
... predicted outcomes?			👍	
... whether my model is adequate?	❌	❌	❌	

Exercise 9.2.3 (Confidence interval for comparing groups).

We are given data $x = [x_1, \dots, x_m]$ which we believe is $N(\mu, \sigma^2)$ and further data $y = [y_1, \dots, y_n]$ which we believe is $N(\mu + \delta, \sigma^2)$. Find a 95% confidence interval for $\hat{\delta}$.

The MLEs are

$\hat{\mu} = \bar{x}$

$\hat{\delta} = \bar{y} - \bar{x}$

$\hat{\sigma} = \cdots$

```
x = [4.3, 5.1, 6.1, 6.8, 7.4, 8.8, 9.9]
y = [8.3, 8.5, 8.9]
m,n = len(x), len(y)
```

① Decide on a readout function

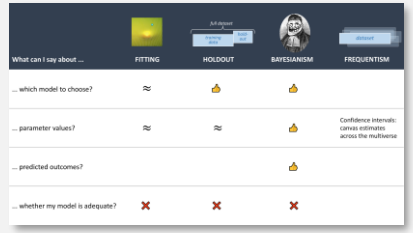
```
def t(x,y): return np.mean(y) - np.mean(x)
```

② Simulate synthetic versions of the dataset

```
# fit a parametric model
μ̂,δ̂ = np.mean(x), np.mean(y) - np.mean(x)
σ̂ = np.sqrt((np.sum((x-μ̂)**2 + np.sum((y-μ̂-δ̂)**2))/(m+n))
def rxy_star():
    return (np.random.normal(loc=μ̂, scale=σ̂, size=m),
            np.random.normal(loc=μ̂ + δ̂, scale=σ̂, size=n))
```

```
# fit a non-parametric model
def rxy_star():
    return (np.random.choice(x, size=m),
            np.random.choice(y, size=n))
```

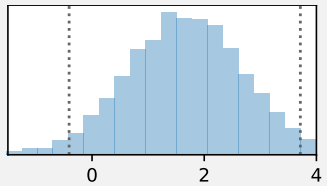
This is exactly what the top row of the table answers!



How should I choose between these two models?

③ Sample the readout statistic and report its spread

```
t_ = [t(*rxy_star()) for _ in range(10000)]
lo,hi = np.quantile(t_, [.025, .975])
plt.hist(t_)
```



THE TOTALITARIAN RULE

When you fit a model by maximizing $\Pr(\text{data}; \text{params})$ use ALL the data and ALL the parameters.

When you resample, resample ALL the data.

Once we've tried all the models we can think of
and chosen the best, what next —
do we go with it and publish?
or do we keep brainstorming for new models?



FREQUENTISM

1. Eyeball our model's fit, to see if we can spot any specific improvements
   e.g. compute the residuals $\varepsilon_i = y_i - \mathrm{pred}(x_i)$ and look for patterns

2. Ask: is it plausible that our model might have generated the dataset?



Confidence intervals:
canvas estimates
across the multiverse



... whether my model is adequate?





Hypothesis
Testing

RESEARCH QUESTION

Can you taste the difference between milk-first versus tea-first?

EXPERIMENT

Make 4 cups of each style, shuffle them, and ask taster to label them



RESEARCH QUESTION

Can you taste the difference between milk-first versus tea-first?

EXPERIMENT

Make 4 cups of each style, shuffle them, and ask the taster to label them

DATASET

8 pairs of (truth,label)

HYPOTHESIS / MODEL

There's no difference. Thus the dataset arose by chance, from purely random choice



Fisher's hypothesis testing

Let x be the dataset.

State a null hypothesis H_0, i.e. a probability model for the dataset

1. Choose a test statistic
$$t : \text{dataset} \mapsto \mathbb{R}$$

2. Define a random synthetic dataset X^*, what we might see if H_0 were true.

3. Look at the histogram of $t(X^*)$.
Let p be the probability of seeing a value as extreme or more so than the observed $t(x)$.

A low p-value is a sign that H_0 should be rejected.

cup id	truth	taster
1	milk	milk
2	tea	tea
⋮		

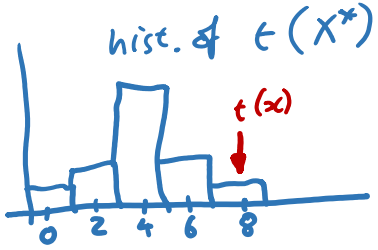
x = taster's assignment of labels

H_0: the taster can't tell the difference and just assigned labels randomly

t(x) = #correct

def X*(): return random permutation of {t,t,t,t,m,m,m,m}

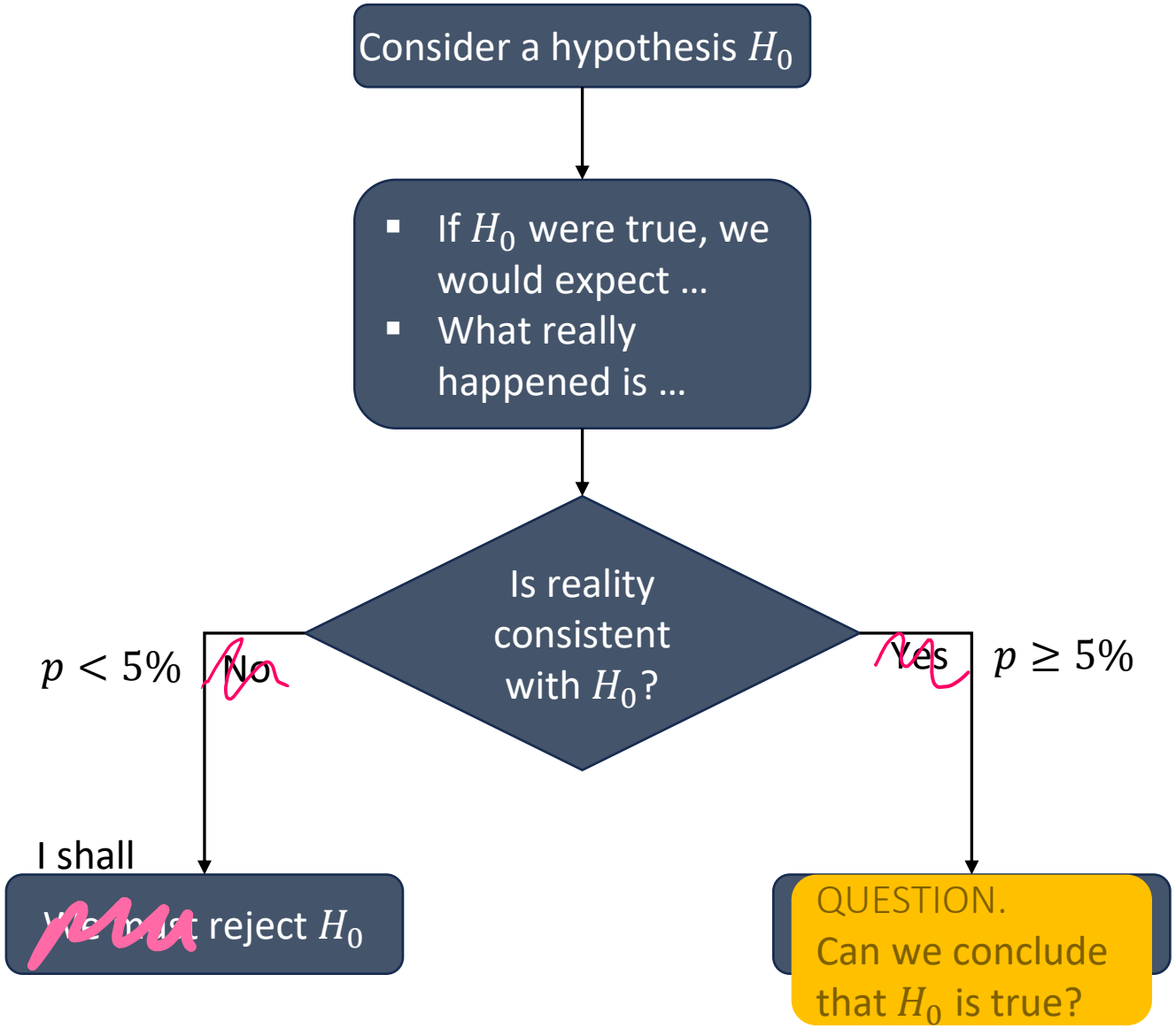
What would be the distribution of the test statistic if H_0 were true?



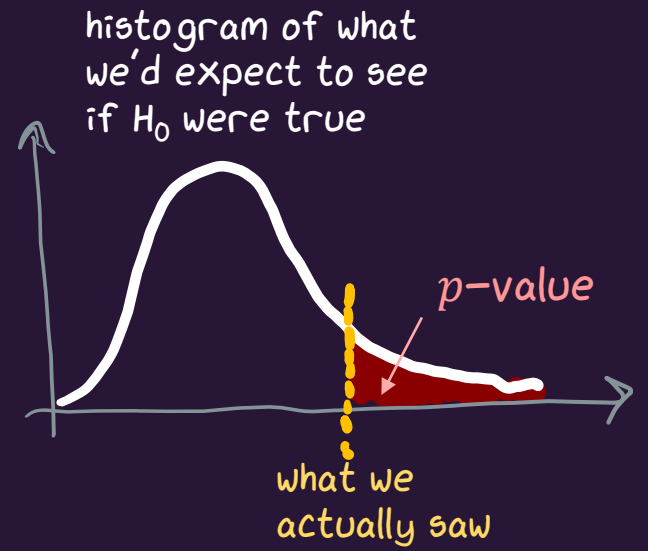
$p = \mathbb{P}(t(X^*) \geq t(x)) = 1.4\%$

p<5% so we shall reject H_0.
"The evidence against H_0 is statistically significant (p=1.4%)"

Hypothesis testing asks whether a proposed probability model H_0 is consistent with the dataset.



- Because of noise, this isn't a simple yes/no decision.
- It's about *degree of consistency,* which we measure by the p-value. Instead of yes/no, we can go by $p < 5\%$ / $p \geq 5\%$.
- The p-value is the probability, according to the model H_0, of seeing something at least as extreme as what we actually saw.



Example 9.3.2, 9.5.2.

I have a dataset with readings from two groups, $x = [x_1, \dots, x_m]$ and $y = [y_1, \dots, y_n]$. Test whether the two groups are significantly different, using the test statistic $\bar{y} - \bar{x}$.

Parametrically:

H_0: X, Y are both sampled from $N(\mu, \sigma^2)$

(estimate these params from all of the data (Totalitarian rule).

Non-parametrically:

H_0: X, Y are both samples drawn from the same distribution.

estimate this dist. by the emp. dist. of all of the data.

① Define the test statistic

```
def t(x,y): return np.mean(y) - np.mean(x)
```

② Simulate synthetic versions of the dataset, assuming H_0 is true

```
xy = np.concatenate([x,y])
μ̂ = np.mean(xy)
σ̂ = np.sqrt(np.mean((xy - μ̂)**2))
def rxy_star():
    return (np.random.normal(loc=μ̂, scale=σ̂, size=len(x)),
            np.random.normal(loc=μ̂, scale=σ̂, size=len(y)))
```

```
xy = np.concatenate([x,y])
def rxy_star():
    return (np.random.choice(xy, size=len(x)),
            np.random.choice(xy, size=len(y)))
```

③ Sample the test statistic under H_0 and compute the p-value

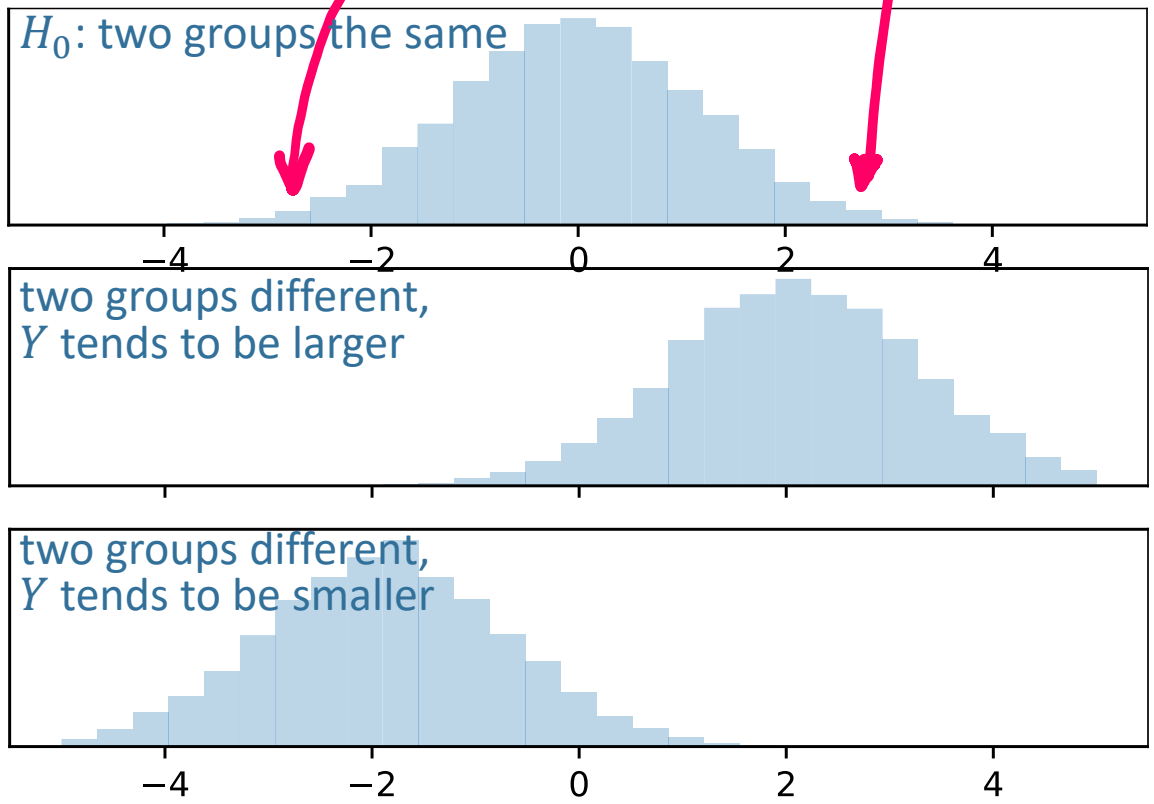
```
t_ = np.array([t(*rxy_star()) for _ in range(10000)])
p =  2 * min(np.mean(t_ >= t(x,y)), np.mean(t_ <= t(x,y)))
```

Decide what "more extreme" means. This decides whether we do a one-tailed or two-tailed test.

What does “more extreme” mean?

What counts as ‘more extreme’?

- Plot the histogram for $t(X^*)$, assuming H_0 is true
- Also plot the histogram for some scenarios where H_0 is false
- Do the alternatives push $t(X^*)$ bigger, or smaller, or either?
  This determines what ‘more extreme’ means — either one-tailed or two-tailed.



Thus, if the observed t lies at either extreme, it's evidence against H_0: "the two groups are the same"

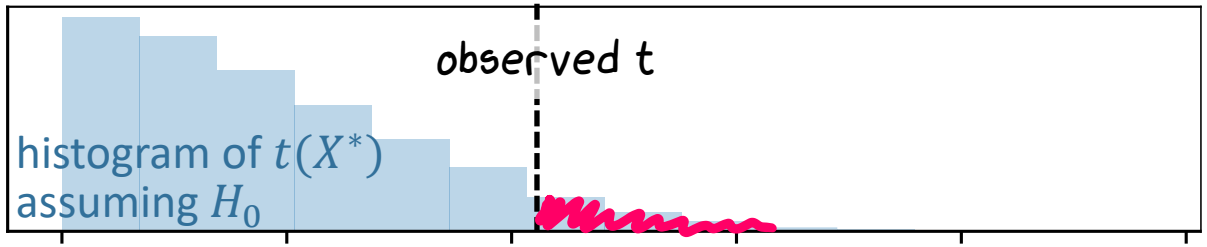
Large +ve values of t are evidence against H_0

Large −ve values of t are evidence against H_0

The p-value is

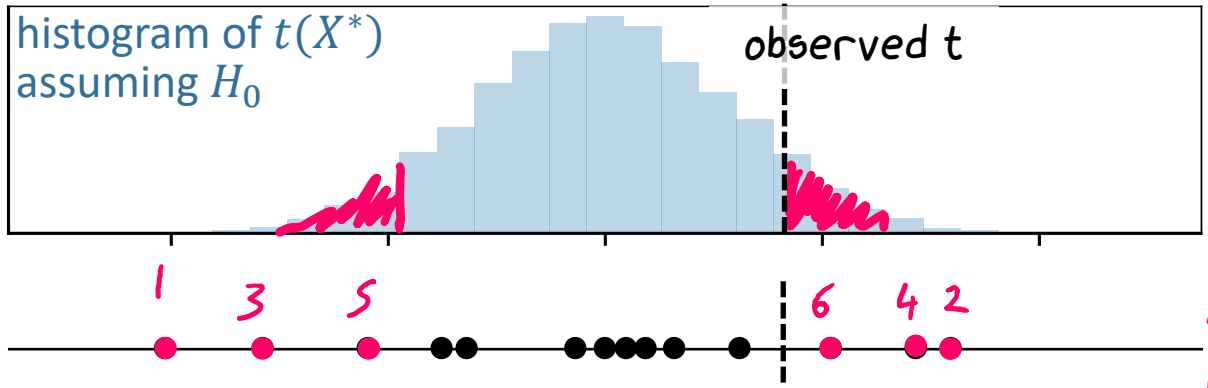
$$p := \mathbb{P}\left(\begin{array}{c} t(X^*) \text{ at least} \\ \text{as extreme as } t(x) \end{array} \;\middle|\; H_0 \text{ is true}\right) \approx \begin{array}{c} \text{fraction of synthetic datasets } X^* \text{ with} \\ t(X^*) \text{ at least as extreme as } t(x) \end{array}$$

If we've decided on a one-tailed test ("reject H_0 if the observed t is large") ...




```
p = np.mean(t_ >= t(x))
```

If we've decided on a two-tailed test ("reject H_0 if the observed t is large +ve or -ve") ...




```
p = 2 * min(np.mean(t_ >= t(x)),
            np.mean(t_ <= t(x)))
```

"6 of my samples of t(X*) are more extreme than t(x)."

Where do test statistics come from?

Example 9.3.2.
I have a dataset with readings from two groups, $x = [x_1, \dots, x_m]$ and $y = [y_1, \dots, y_n]$. Test whether the two groups are significantly different, using the test statistic $\bar{y} - \bar{x}$.

In some problems, it's natural to express H_0 as a *constraint* on the parameters of a more general model H_1.

e.g. H_1: $X \sim N(\mu, \sigma^2)$ $Y \sim N(\nu, \sigma^2)$, $H_0: \mu = \nu$

e.g. H_1: $X \sim N(\mu, \sigma^2)$ $Y \sim N(\mu + \delta, \sigma^2)$, $H_0: \delta = 0$.

- **For the test statistic:**
  estimate the parameters under H_1, $\hat{\mu} = \bar{x}$, $\hat{\nu} = \bar{y}$. OR $\hat{\delta} = \bar{y} - \bar{x}$
  and let t measure how close they are to meeting the constraint.

e.g. $t = |\bar{y} - \bar{x}|$. If H_0 is true, this will be small ≈ 0. So $t \gg 0$ is evidence against H_0. So do a 1-tailed test.

e.g. $t = \bar{y} - \bar{x}$. If H_0 is true, this will be ≈ 0. So $t \gg 0$ or $t \ll 0$ is evidence against H_0. So do a 2-tailed test.

- **For resampling:**
  Fit H_0 by maximizing the likelihood of the data under the constraint.
  Then, sample from this fitted model.

ie $\max_{\mu,\nu,\sigma : \mu = \nu} \log Pr(\text{data}; \mu, \nu, \sigma)$ or equivalently $\max_{\lambda,\sigma} \log Pr(\text{data}; \lambda, \lambda, \sigma)$

Exercise 9.3.3 (Equality of group means).

We are given three groups of observations from three different systems

$$x = [7.2, 7.3, 7.8, 8.2, 8.8, 9.5]$$
$$y = [8.3, 8.5, 9.2]$$
$$z = [7.4, 8.5, 9.0]$$

Do all three groups have the same mean?

① State a general model H_1 then express our H_0 as a constraint on its parameters

H_1: $X \sim N(a, \sigma^2)$ $Y \sim N(b, \sigma^2)$ $Z \sim N(c, \sigma^2)$

H_0: $a = b = c =: \mu$

② Estimate the parameters under H_1, then let t measure how close they are to meeting the constraint

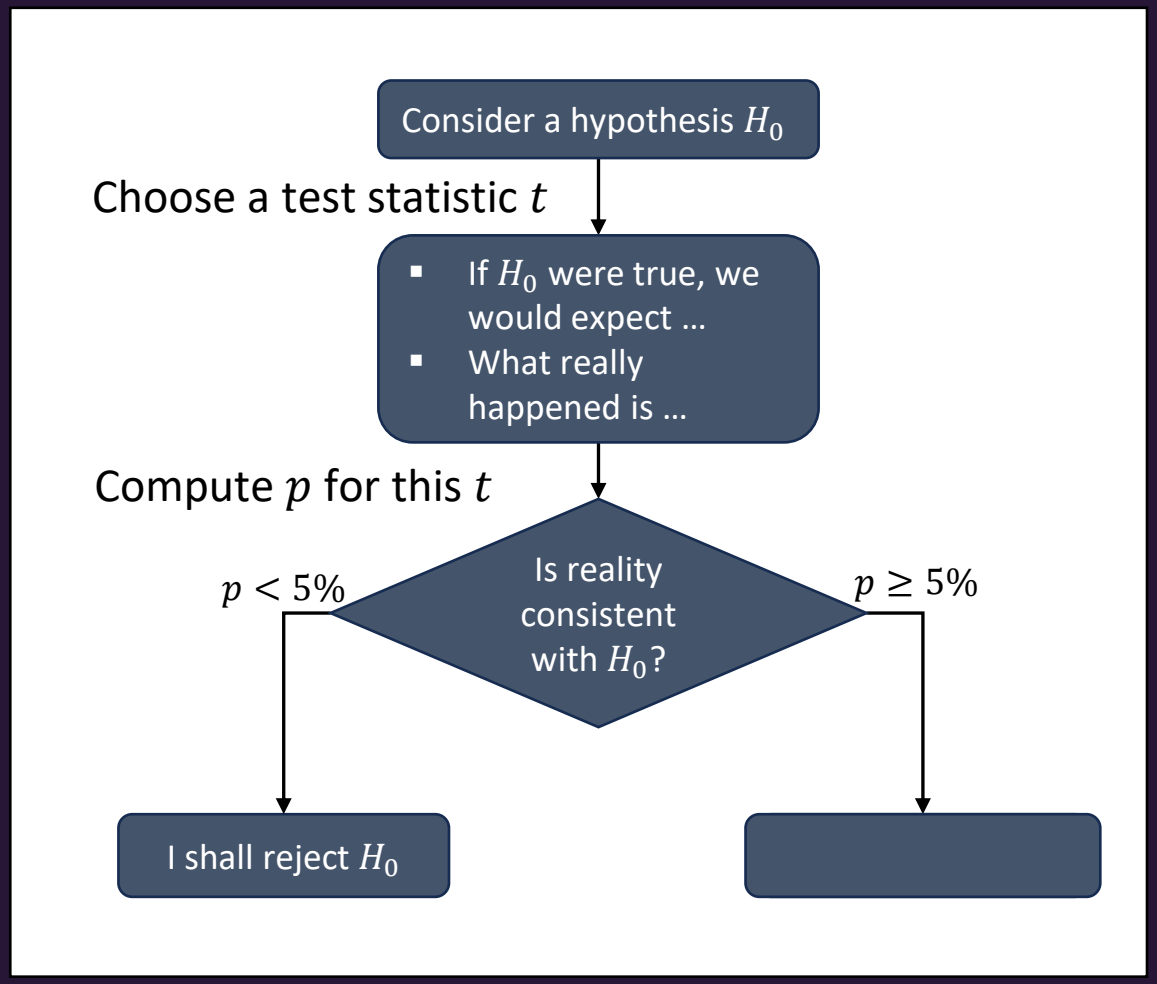
$\hat{a}$ $\hat{b}$ $\hat{c}$.

e.g. $t = (\hat{a} - \hat{b})^2 + (\hat{b} - \hat{c})^2 + (\hat{c} - \hat{a})^2$

e.g. $t = (\hat{a} - \hat{\mu})^2 + (\hat{b} - \hat{\mu})^2 + (\hat{c} - \hat{\mu})^2$ where $\hat{\mu}$ = MLE for my H_0 model, all data comes from $N(\mu, \sigma^2)$

③ For resampling, fit the model under the H_0 constraint

How do I formulate my H_0 for the question I want to investigate?

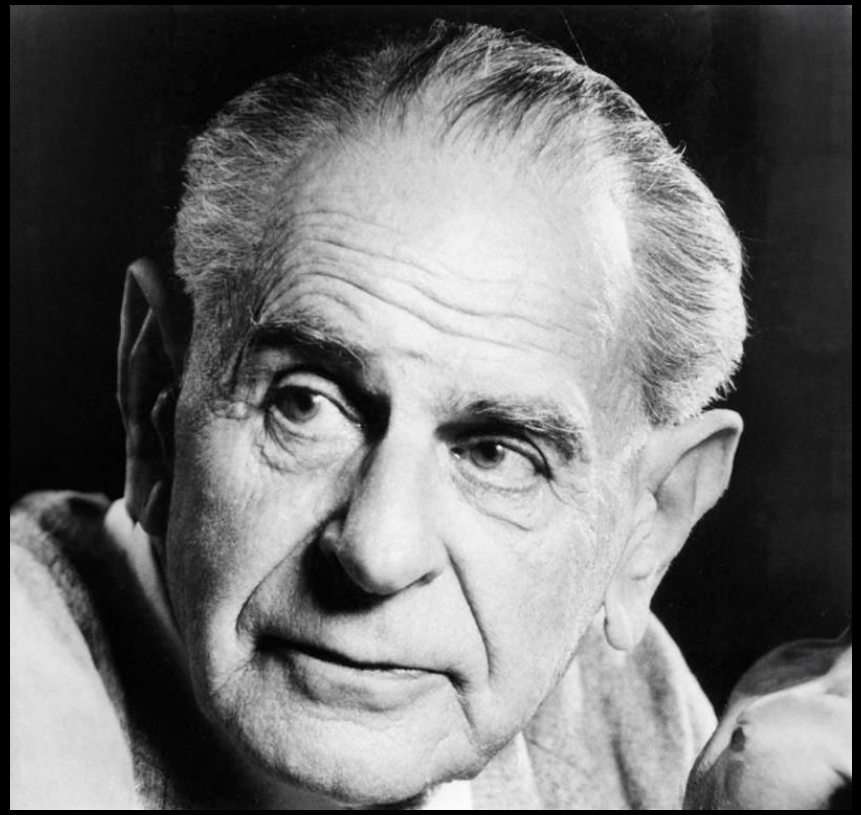


"Data science is quantitative rhetoric"



We only get a definite publishable conclusion if we reject H_0.

Anything we want to argue, we have to phrase it as "reject H_0" for a suitable H_0.



"Every genuine scientific theory must be falsifiable. It is easy to obtain evidence in support of virtually any theory; the evidence only counts if it is the positive result of a genuinely risky prediction."

Karl Popper (1902–1994)

EXERCISE.

Here are marks for IA Algorithms questions.
I think there might be a gender bias.
What H_0 do I choose?

gender	mark
F	19
F	14
M	18
O	11
M	16
⋮	⋮



H_0: I think everyone gets pretty much the same marks, regardless of gender

Null hypothesis:

Let's introduce a richer model, $H_1: \text{Mark} \sim \mu_{\text{gender}} + N(0, \sigma^2)$
and express H_0 as a constraint: $\mu_F = \mu_M = \mu_O$

H_1: I think gender affects marks



Test statistic: $t = (\hat{\mu}_F - \hat{\mu})^2 + (\mu_M - \hat{\mu})^2 + (\hat{\mu}_O - \hat{\mu})^2$ where $\hat{\mu}_F, \hat{\mu}_M, \hat{\mu}_O$ are MLE under H_1 and $\hat{\mu}$ is MLE under H_0

Resampling: Fit H_0 which says $\text{Mark} \sim N(\mu, \sigma^2)$ then sample from this fitted $N(\hat{\mu}, \hat{\sigma}^2)$

Here are marks for IA Algorithms questions.
I think there might be a gender bias.



Under the general H_1 model

$$\text{Mark} \sim \mu_{\text{gender}} + N(0, \sigma^2)$$

I propose $H_0: \mu_M = \mu_F = \mu_O$

gender	mark
F	19
F	14
M	18
O	11
M	16
⋮	⋮

QUESTION. Suppose we reject H_0. Does this mean we believe that the means μ_g are not all equal?

- Our H_0 claims several things at the same time:
  (means are equal) & (variances are equal) & (all are Gaussian) & (all are independent)
- If we reject H_0, we reject it in its entirety.
  We might be rejecting it because of evidence against any one or more of its subclaims.

NEXT LECTURE

THE GREAT GENERALIZATION

SMACKDOWN

BAYESIANIST v FREQUENTIST

