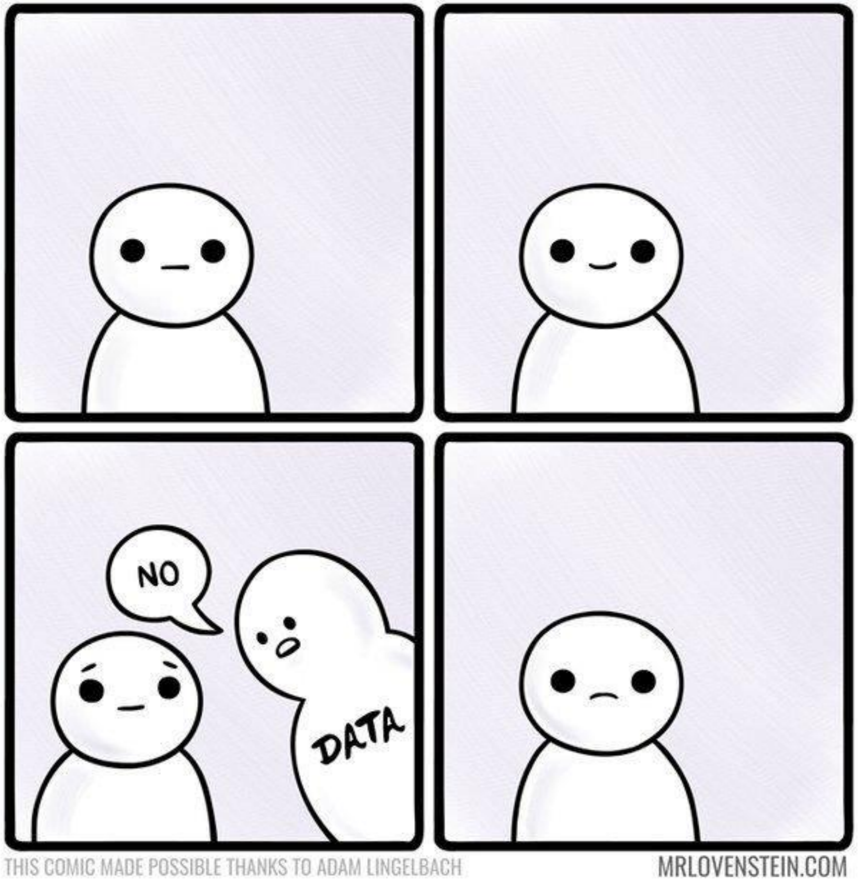
Lecture 5 – Probability & Hypothesis Testing



February 5th, 2026

Spring 2026

Today’s Learning Outcomes

1. Be able to explain the problems with the hypothesis testing framework

2. Be able to explain what is a p-value means and it’s relationship to α

3. Know (roughly) the difference between frequentist and Bayesian approaches

Hypothesis Testing (basic)

- In stereotyped hypothesis testing we typically set up two hypotheses

- H_0 = null hypothesis
  - Typically, a model of no or random pattern of occurrences between variables

- H_1 = alternative hypothesis
  - Our proposed explanation for an observed pattern

Hypotheses

- We do not actually ever test the alternative hypothesis!

- Instead most tests are set up to reject the null hypothesis, but that does not mean the alternative hypothesis is automatically correct
  - They can both be wrong!

- Ex. We have to sets of numbers of equal length (x_1 and x_2)
  - H_0: mean(x_1) = mean(x_2)
  - H_1: mean(x_1) ≠ mean(x_2)

Rejection of the null does not tell me anything about how different or in what direction the difference in means is

Testing Hypotheses

- Statistical tests often rely on the use of p-values to determine “objectively” whether some difference or correlation is meaningful or not

- p-value varies from 0 to 1, and represents the % chance that your observed result could be due to random chance based on your model of the data

- The acceptable level of risk of accepting a difference as significant when it might actually not be is the significance level, symbolized as α

p-values

- We set α <u>prior</u> to performing a test that gives us back a p-value
  - A p-value below α is a rejection of the null hypothesis
  - A p-value greater than α fails to reject the null hypothesis

- The most common p-value below which a test is considered significant is 0.05 = 1 in 20 chance
  - Literally chosen arbitrarily for space-saving, copyright avoidance, and possible personal dislikes

The Origin of p-values

“We were surprised to learn, in the course of writing this paper, that the $p < 0.05$ cutoff was established as a competitive response to a disagreement over book royalties between two foundational statisticians. In the early 1920s, Kendall Pearson, whose income depended on the sale of extensive statistical tables, was unwilling to allow Ronald A. Fisher to use them in his new book. To work around this barrier, Fisher created a method of inference based on only two values: p-values of 0.05 and 0.01”

> The value for which P=0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.

- R.A. Fisher in “Statistical tables for biological, agricultural and medical research”

T-test Example

- The t-test is used to determine whether two sets of values were drawn from the same distribution
  - Simplest version assumes datasets are independent, are approximately normal, and have similar variances
  - In Python the function is `ttest_ind`

```
import numpy as np
from scipy.stats import ttest_ind
# generate two samples like rnorm(20, 0, 1)
x1 = np.random.normal(0, 1, 20)
x2 = np.random.normal(0, 1, 20)
# two-sample t-test
t_stat, p_value = ttest_ind(x1, x2)
ttest_ind(x1, x2)
```

20 values each from the exact same normal distribution

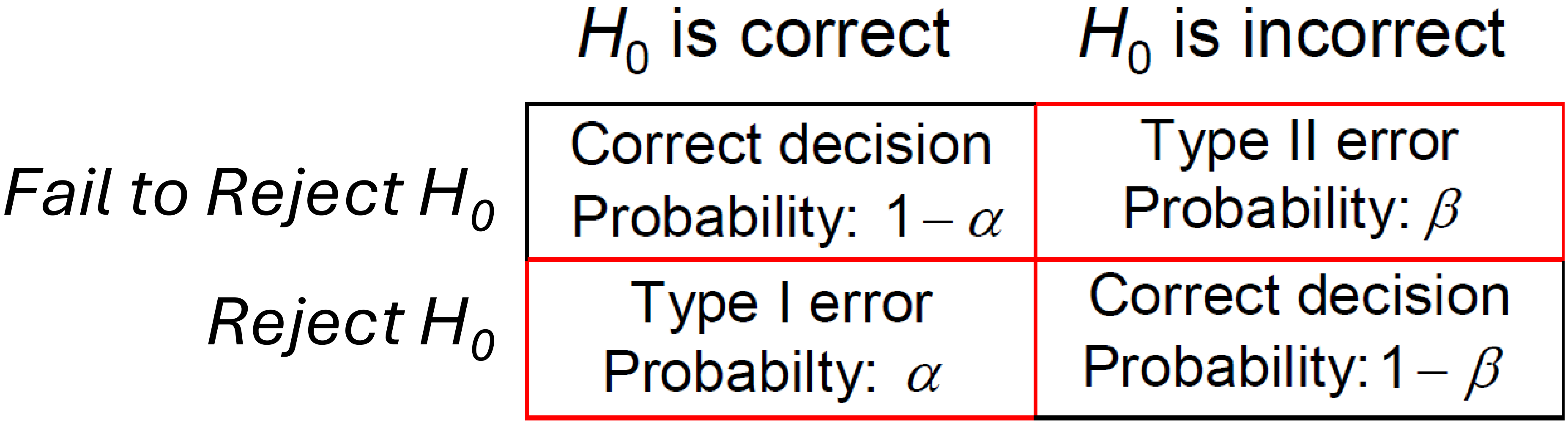
p-value > 0.05, fail to reject

Output:

```
TtestResult(statistic=np.float64(0.09120247539013474),
pvalue=np.float64(0.9278107889418081), df=np.float64(38.0))
```

Essentially test whether the normal distributions of each set overlap enough that they could be from one distribution

p-values, α, and β



	H_0 is correct	H_0 is incorrect
Fail to Reject H_0	Correct decision Probability: $1-\alpha$	Type II error Probability: β
Reject H_0	Type I error Probabilty: α	Correct decision Probability: $1-\beta$

- In the previous slide we correctly rejected the null hypothesis but with real data we are never actually sure which of the 4 boxes above we are in for any given test
  - β is the probability of failing to reject the null when it actually is incorrect (1 – β = power)

p-values, α, and β

	H_0 is correct	H_0 is incorrect
Fail to Reject H_0	Correct decision Probability: $1-\alpha$	Type II error Probability: β
Reject H_0	Type I error Probabilty: α	Correct decision Probability: $1-\beta$

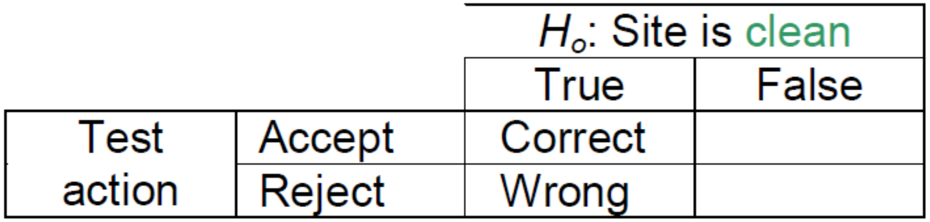
```
import numpy as np
from scipy.stats import ttest_ind
# generate two samples like rnorm(20, 0, 1)
x1 = np.random.normal(0, 1, 20)
x2 = np.random.normal(0, 1, 20)
# two-sample t-test
t_stat, p_value = ttest_ind(x1, x2)
ttest_ind(x1, x2)

Output:
TtestResult(statistic=np.float64(-2.2383316840237413),
pvalue=np.float64(0.031136755070637314), df=np.float64(38.0))
```

20 values each from the exact same normal distribution

p-value < 0.05, reject
Type 1 error

What Should Your Null Hypothesis Be?

- When we run tests the null is often set up to fail but this can matter a lot (we don’t check if the alternative is actually supported)
-

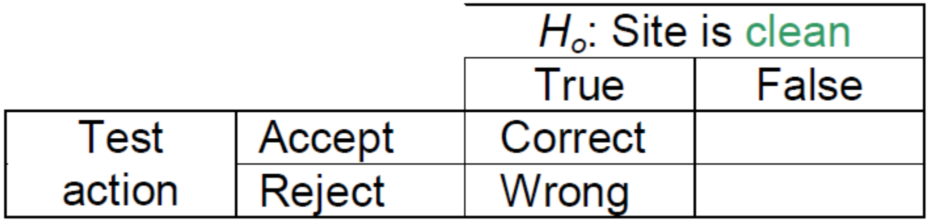


		H_o: Site is clean	
		True	False
Test action	Accept	Correct	
	Reject	Wrong	

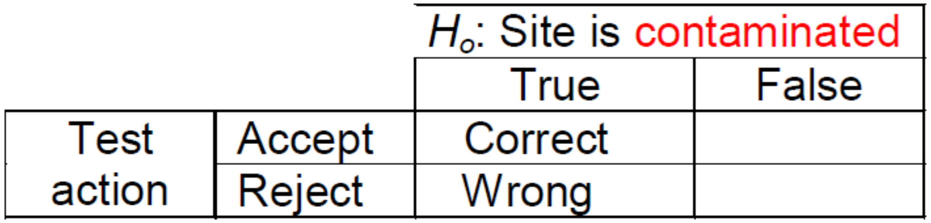
A. Wrong rejection means the site is declared contaminated when it is actually clean, which would lead to unnecessary cleaning.

What Should Your Null Hypothesis Be?

- When we run tests the null is often set up to fail but this can matter a lot (we don't check if the alternative is actually supported)
-



		H_o: Site is clean	
		True	False
Test action	Accept	Correct	
	Reject	Wrong	



		H_o: Site is contaminated	
		True	False
Test action	Accept	Correct	
	Reject	Wrong	

A. Wrong rejection means the site is declared contaminated when it is actually clean, which would lead to unnecessary cleaning.

B. Now, the wrong decision declares a contaminated site clean. No action, prolongs a health hazard.

What Should Your Null Hypothesis Be?

- When we run tests the null is often set up to fail but this can matter a lot (we don’t check if the alternative is actually supported)

- Also, not always clear what a null hypothesis should be
  - Association between geographic range and extinction risk
    - Even if actually independent we expect them to have a positive correlation due to both correlating with sampling

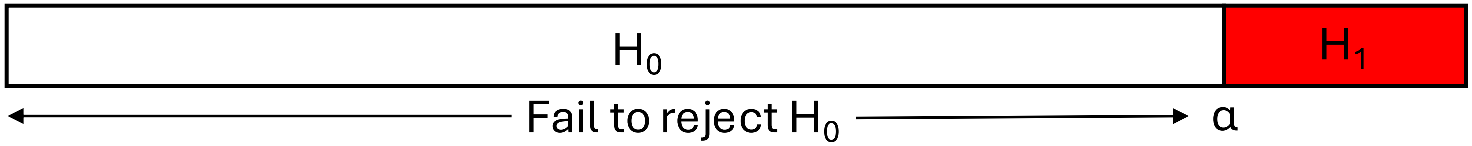
Testing Expectations

- Let’s check whether the p-value represents what it is supposed to

- Example using the P_Rand() function in the Jupyter Notebook for lecture 5
    - Function which generates random samples from the same distribution and calculates a p-value many times from a correlation many times
    - Expectation is that with α = 0.05 we should reject the null hypothesis incorrectly (type 1 error) 5% of the time

Problems with Traditional Hypothesis Testing

- Inherently a binary system with either fail or reject



- Does not say anything about the strength of association and often not even the direction
  - Smaller p-value not indicative of strength, just probability

- Can do multiple pairwise tests but still binary and hard to compare alternative models

Significance Versus Importance

- As sample size increases our ability to confidently reject a null hypothesis also increases if the null hypothesis is indeed false

- Can also detect smaller differences with large sample size, but run the risk of worrying about real but not meaningful differences
  - Not all differences actually matter

Hypothetical Deer Fossil Example

From here, population of deer femur bone lengths

```
# small n
n = 25
deer1 = np.random.normal(loc=295.1, scale=1, size=n)
deer2 = np.random.normal(loc=295.2, scale=1, size=n)
print('n=25 t-test:', ttest_ind(deer1, deer2))
```

not significant
incorrectly fail to reject

```
n=25 t-test: TtestResult(statistic=np.float64(-
1.3825527924034435), pvalue=np.float64(0.17320112522574393),
df=np.float64(48.0))
```

Hypothetical Deer Fossil Example

From here, population of deer femur bone lengths

Even though actually different, it’s not a difference that functionally matters. Any two populations are going to be different

```
# large n
n = 1000
deer1 = np.random.normal(loc=295.1, scale=1, size=n)
deer2 = np.random.normal(loc=295.2, scale=1, size=n)
print('n=25 t-test:', ttest_ind(deer1, deer2))
```

significant
correctly reject

```
n=1000 t-test: TtestResult(statistic=np.float64(-
2.891312779051431), pvalue=np.float64(0.0038778119319773 91),
df=np.float64(1998.0))
```

Geology Scenario

- We are trying to detect a difference in silica percent from two eruptions because we think it is relevant to the observed difference in their eruption style
  - In reality they ARE different so if we take a large enough sample size we would be able to distinguish that difference, but the actual difference might be so slight (let's say 0.1%) that it does not really make sense as the explanation for the different eruption styles

- NEED TO CHECK YOUR ABSOLUTE EXPLANATORY POWER AND YOUR UNDERSTANDING OF THE SYSTEM AS A WHOLE!

Frequentist & Bayesian

- Tests using p-values are inherently a frequentist approach
    - A single estimated value (p) based on rigid assumptions of how data behave (such as normality)

- Bayesian approaches instead rely on prior belief and updates of that belief as more information is accumulated
    - Updated as new data are incorporated, still rely on a model but in reality a whole family of models with a range of possible estimated values and their associated likelihoods

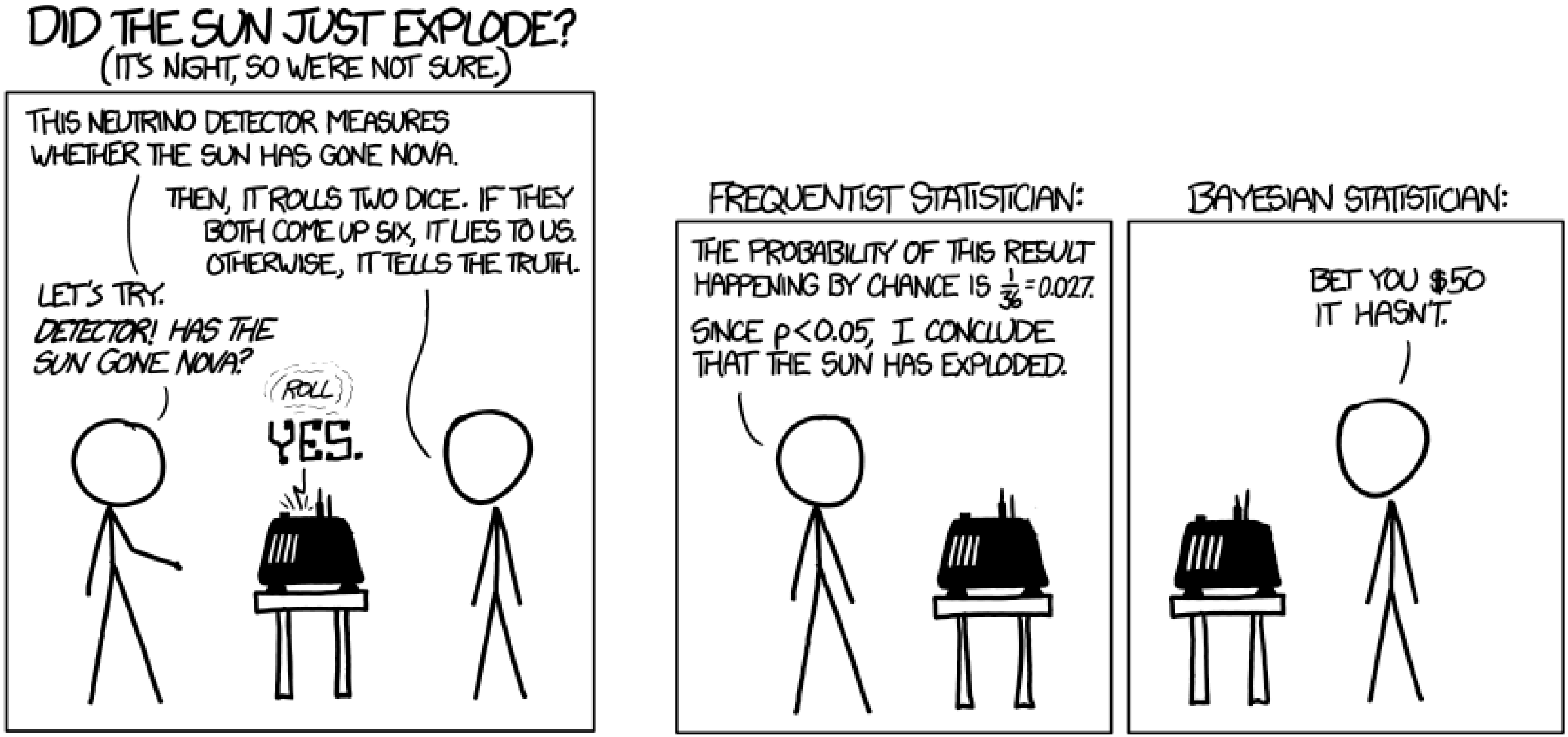
Frequentist Vs. Bayesian Example

- To determine whether they should bring an umbrella today a frequentist approach would be to look up the number of days in a month with rain let’s say the average is 14/31 days and concludes that it is most likely not raining

- A Bayesian approach to the same problem might be to look out the window and observe the number of people carrying umbrellas. You see that ~2/3 of people seem to be carrying them. Based on your previous experience you conclude that it is likely to rain today and bring an umbrella

Limits of Frequentist Approaches

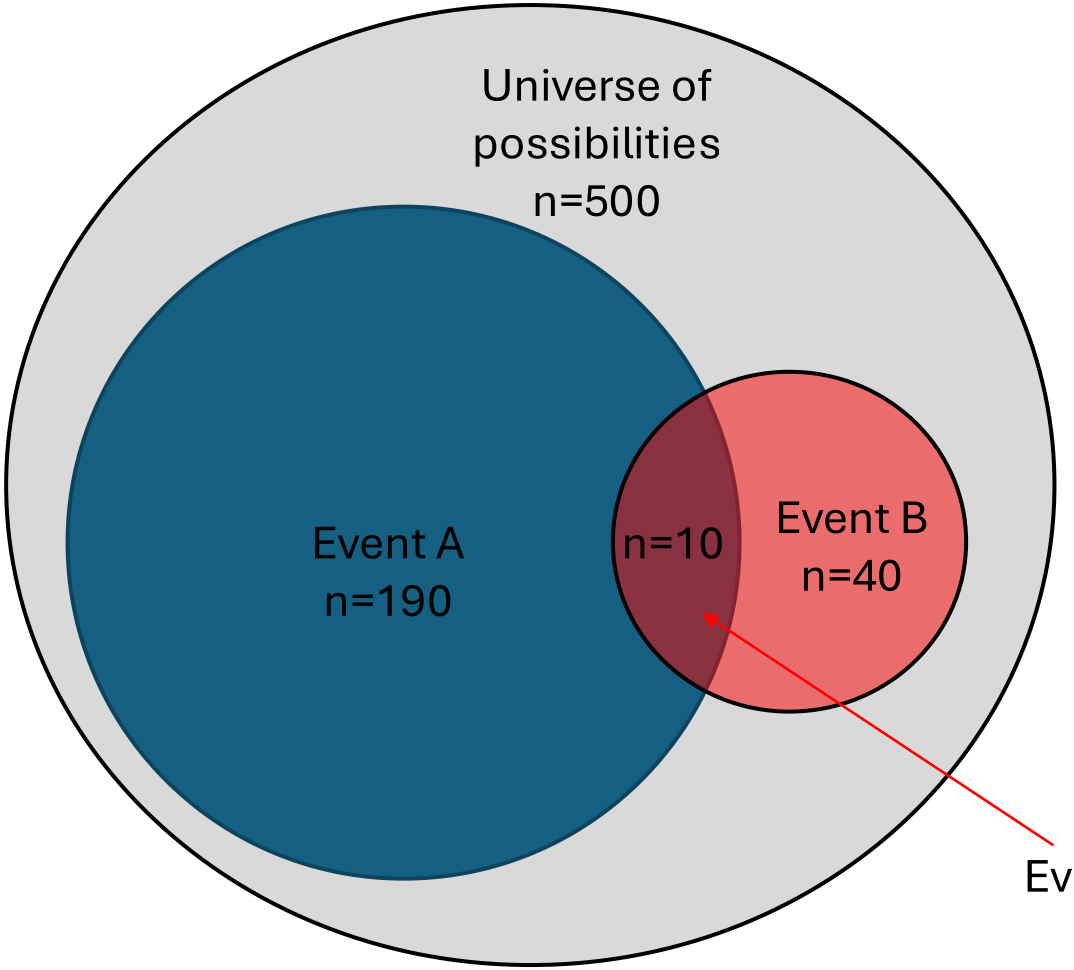
- At extremes frequentist approaches can lead to very wrong conclusions when applied blindly


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}=0.027$. SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Frequentist

I have two events of interest
Event A = presence of an oil deposit
Event B = presence of a salt dome

How likely is it that a new well has both a salt dome and an oil deposit?



Event A and B

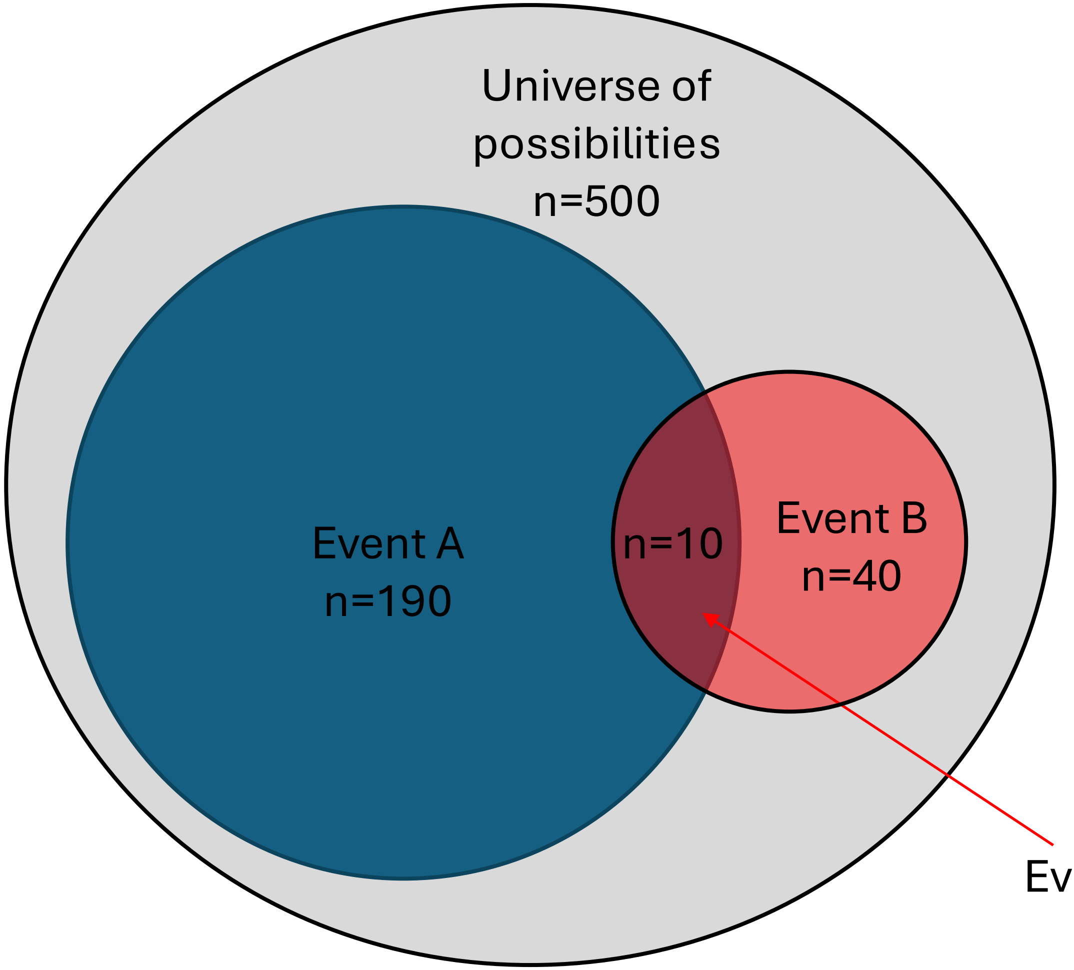
10 wells have both a salt dome and oil

10/500 = 0.02 = 2% chance of oil deposit
based on this approach

Fails to take account of prior knowledge about how common event A is independent of event B

Bayesian

A newly drilled well has a salt dome.
What is the probability it also has oil deposits?
= P(AB)



Event A and B = P(AB)

$P(A|B) * P(B) = P(B|A)*P(A)$

↓

$P(A|B) = [P(B|A)*P(A)] / P(B)$

$P(A|B) = [10/200 * 200/500] / [50/500]$

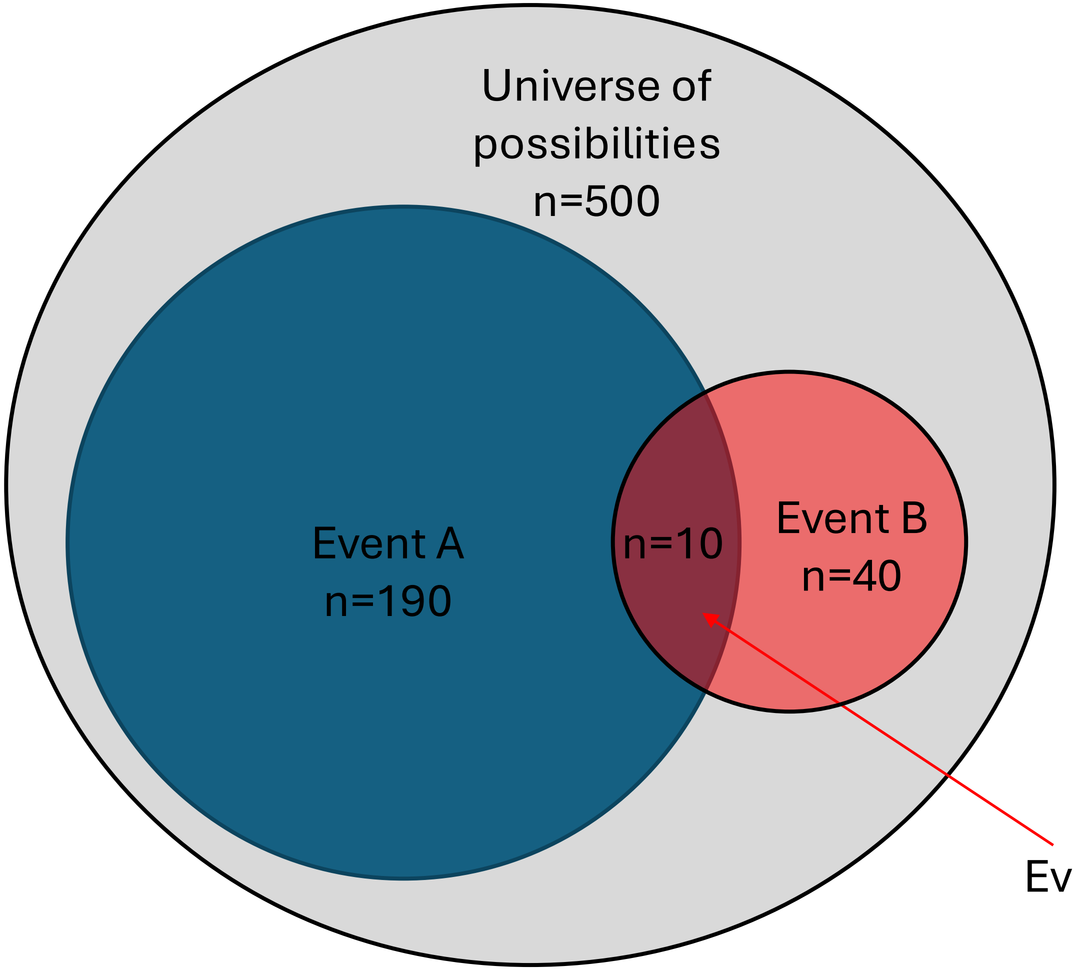
$P(A|B) = [0.05 * 0.4] / [0.1]$

$P(A|B) = 0.02 / 0.1 = 0.2$

= 20% chance of oil given presence of a salt dome

Bayesian

A newly drilled well has oil deposits.
What is the probability it also has a salt dome?
= P(AB)



Event A and B = P(AB)

$$P(B|A) * P(A) = P(A|B)*P(B)$$

↓

$$P(B|A) = [P(A|B)*P(B)] / P(A)$$

$$P(B|A) = [10/50 * 50/500] / [200/500]$$

$$P(A|B) = [0.2* 0.1] / [0.4]$$

$$P(A|B) = 0.02 / 0.1 = 0.05$$

= 5% chance of a salt dome given presence of oil deposits

Bayes Theorem

- Bayes theorem P(A|B) = [P(B|A)*P(A)] / P(B)
  - Probability of A given event B = (probability of event B given A * probability of A) / probability of B

- P(A) is the prior, starting points which may be altered by the data if there is support for it

- Bayesian approaches are all couched as probabilities so I can directly compare the likelihood of one model to another (no obvious analog for p-values)

*Another example explanation

*Venn diagram visualization/explanation

Given some measurements about a volcano today, what is the probability it will produce an eruption larger than size X in the next year?

Frequentist Model

1. Collect lots of data

- Many volcanoes
- Observations of many variables (gas emissions, ground deformation, seismicity, etc.)
- For each volcano: did it erupt above size X within 1 year after the observations? (Yes/No)

2. Fit a standard statistical model
   - Example: logistic regression
   - This produces a single predicted probability between 0 and 1.

3. Act on that probability
   - You might set a threshold like

- If probability > 0.3 → issue alert
- Else → no alert

Given some measurements about a volcano today, what is the probability it will produce an eruption larger than size X in the next year?

Bayesian Model

1. Same data collection
2. Bayesian thinking - parameters are random
   - The model coefficients (e.g., effect of gas flux or seismicity) are treated as random variables.
   - You use both ***Prior beliefs*** (what you know before data), and ***Likelihood*** (what the data says) to generate posterior distributions (updated beliefs).
   - Output = a range of possible parameter values,
     - You *will NOT* get: Coefficient for seismicity = 0.7
     - You *WILL* get: Coefficient for seismicity ~ Normal(0.7, ± 0.15)
     - Meaning: You are unsure about the true value, and the model reflects that uncertainty.

- 3. Prediction is also a distribution
  - Using the posterior parameter distributions and the current volcano data, the model gives
  - Probability of eruption next year: X ~ 0.18–0.41 (95% credible interval)
    - Instead of a single number, you get a range of plausible probabilities.

Possible outcomes

Frequentist

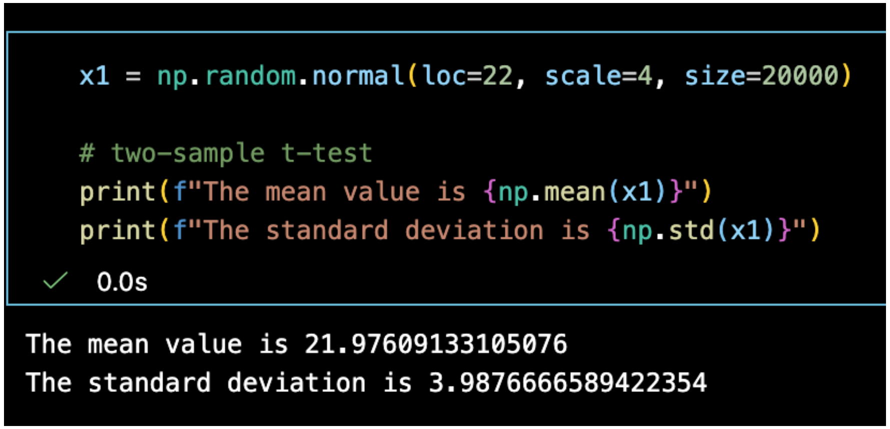
- Threshold breached, but not eruption → everybody is mad, model rejected (rightly/wrongly), back to the drawing board on model form
- Threshold breached, eruption occurs → everybody is thankful, model continues to be used (rightly, wrongly), value recalculated from additional observation but because models was successful unlikely to change much

Bayesian

- Predicts eruption, but not eruption → everybody is mad but good chance model outputs had some part below threshold, model updated with new information and parameters recalibrated, model may be substantially revised
- Predicts eruption, eruption occurs → everybody is thankful, model updated with new information and parameters recalibrated, model unlikely to change much since it performed well

Frequentist Vs Bayesian Summary

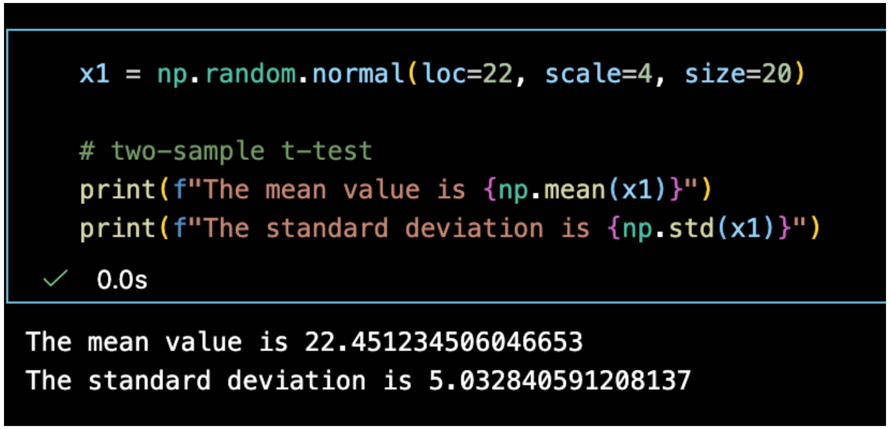
- Given a set of values from a normal distribution a frequentist would estimate the average and standard deviation of the underlying population by calculating the average and standard deviation of those numbers
  - Single point estimate for each parameter needed to define a normal distribution
  - At large sample sizes this works pretty well, but breaks down as you lower sample size below ~30



```
x1 = np.random.normal(loc=22, scale=4, size=20000)

# two-sample t-test
print(f"The mean value is {np.mean(x1)}")
print(f"The standard deviation is {np.std(x1)}")
```

✓ 0.0s

```
The mean value is 21.97609133105076
The standard deviation is 3.9876666589422354
```



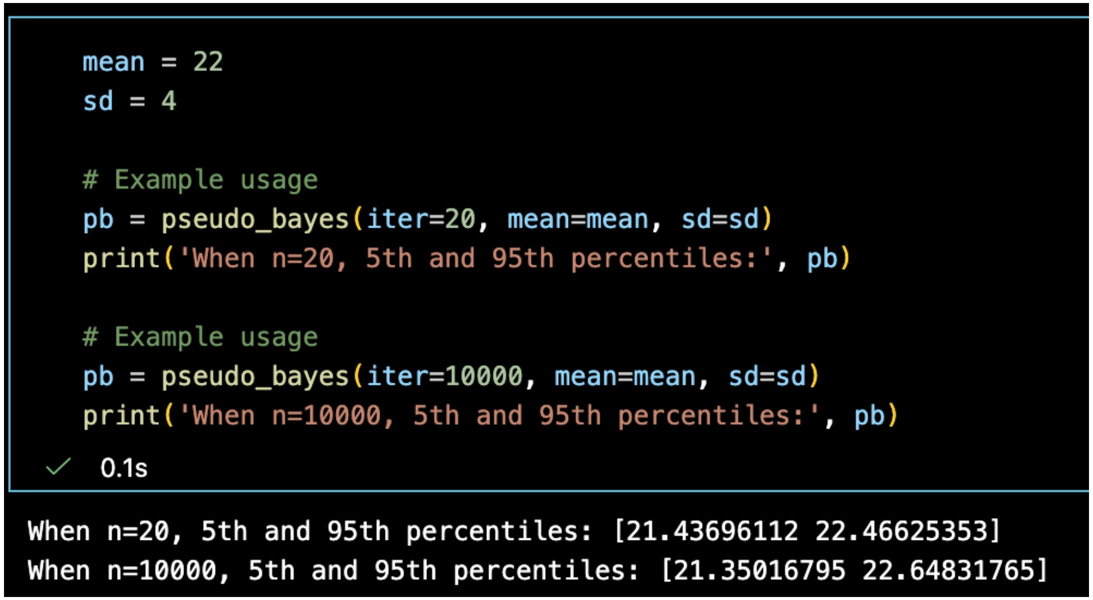
```
x1 = np.random.normal(loc=22, scale=4, size=20)

# two-sample t-test
print(f"The mean value is {np.mean(x1)}")
print(f"The standard deviation is {np.std(x1)}")
```

✓ 0.0s

```
The mean value is 22.451234506046653
The standard deviation is 5.032840591208137
```

Frequentist Vs Bayesian Summary

- Given a set of values from a normal distribution a Bayesian would have a prior estimate for the mean and standard deviation and then calculate a range of possible values and their likelihoods
  - Distribution of possible values of the mean and standard deviation t may have come from and their relative densities



```
mean = 22
sd = 4

# Example usage
pb = pseudo_bayes(iter=20, mean=mean, sd=sd)
print('When n=20, 5th and 95th percentiles:', pb)

# Example usage
pb = pseudo_bayes(iter=10000, mean=mean, sd=sd)
print('When n=10000, 5th and 95th percentiles:', pb)
```

✓ 0.1s

```
When n=20, 5th and 95th percentiles: [21.43696112 22.46625353]
When n=10000, 5th and 95th percentiles: [21.35016795 22.64831765]
```

Range of likely population mean

Today’s Learning Outcomes

1. Be able to explain the problems with the hypothesis testing framework

2. Be able to explain what is a p-value means and it’s relationship to α

3. Know (roughly) the difference between frequentist and Bayesian approaches

Eruption of size X prediction in the next year

Frequentist model

- Observe lots of data from multiple variables for volcanoes along with information about whether they did, or did not, erupt at size X within a year after said observation (# of yes versus # of no divided by total set). Given the current observation model (ex. logistic regression) will spit out a prediction ranging from 0 to 1. Might have a threshold for actions based on the value which is returned.

Bayesian model

- Same as above in terms of data collection but use the frequentist observations of number of eruptions > size X to estimate the parameters (coefficients) that correspond to each of the variables of interest which were observed. Assumes those variables are drawn from a random sample of some kind and will give back a range of values for each parameter in your model. Use that range of variables and the current observed data to predict the range of probability an eruption will occur in the next year above size X