TUT206 OCT 11

Correction!

Question 7: What does it mean if the 95% CI includes 0?

A. Reject null hypothesis
B. Significant effect
C. Cannot reject null hypothesis
D. Parameter is 0

Answer:
C. Cannot reject null hypothesis

Only in the context of $H0 : mu = 0$!!!!!!!!

How to review?

-> see Topic: Midterm + HW05 (utoronto.ca)

Cheat sheet? -> see STA 130 (64 unread) | Piazza QA

Notebook LM! -> podcast

Chatgpt

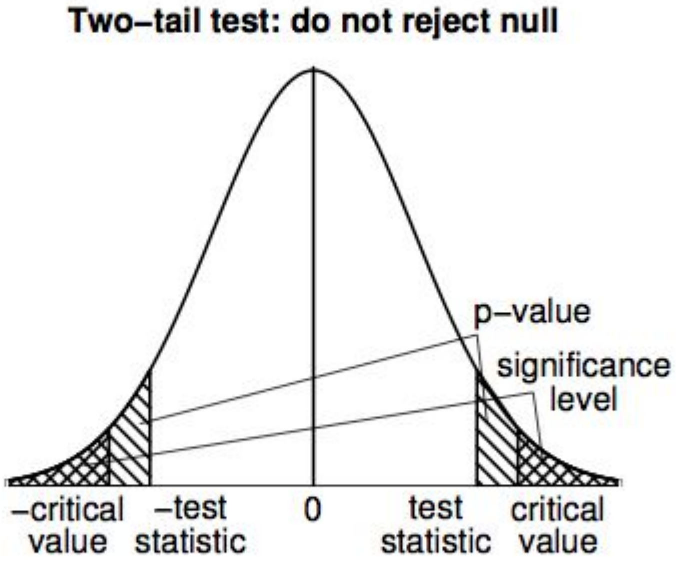
Past midterm

Office hour

Peer review!

Test statistic? what?

a test statistic is a statistic of the sample you took, like the mean or median of a sample



Demo1

$$p\text{-value} = \frac{\text{Number of simulated statistics ''as or more extreme''}}{\text{Total number of simulations}}$$

Simulate Data Under the Null Hypothesis

Simulated_statistics: results of the **simulated statistics** generated under the **null hypothesis** (H0)

Observed Statistic: **observed proportion** of patients whose health score improved

Demo1

A p-value is the the probability that a statistic is as or more extreme than the observed statistic if the null hypothesis is true

If the p-value is less than or equal to the significance level, the null hypothesis is rejected. If the p-value is greater than the significance level, the null hypothesis is not rejected.

- **The p-value can be interpreted as the strength of evidence against the null hypothesis**.
- For example, a p-value of 0.05 means that there is a 5% chance of observing a statistic as or more extreme than the observed statistic if the null hypothesis is true. This would be considered moderate evidence against the null hypothesis.

Demo1

- The p-value will be based on the proportion of simulated sample means that fall within these shaded areas.!!

Demo1: Synthetic (Simulated) Sampling Distribution under the Null Hypothesis (H_0) vs. Bootstrapped Sampling Distribution

1. Synthetic (Simulated) Sampling Distribution under H_0:

- **Goal**: This distribution is created under the assumption that the null hypothesis (**H_0**) is true.
- **Assumption**: It assumes no effect or no difference (in this case, that the health score change is random). The distribution of the test statistic is generated as if H_0 is correct.
- **How it's generated**:
  - We simulate new datasets by randomly assigning outcomes (such as health score improvement or not) based on the assumption that the vaccine has no effect.
  - For each simulated dataset, we compute the test statistic (e.g., the proportion of improvements).
  - We repeat this process many times (e.g., 10,000 simulations), creating a distribution of test statistics.
- **Purpose**:
  - To compare the **observed statistic** from the actual data to this distribution.
  - The p-value is calculated as the proportion of simulated statistics that are as extreme or more extreme than the observed statistic. If the p-value is small, we have evidence to reject H_0.

Demo1: Synthetic (Simulated) Sampling Distribution under the Null Hypothesis (H_0) vs. Bootstrapped Sampling Distribution

2. Bootstrapped Sampling Distribution of a Test Statistic:

- **Goal**: This distribution is generated without making assumptions about H_0. Instead, it estimates the sampling variability of the statistic based on the observed data.
- **Assumption**: It **does not** assume H_0 is true or false. The bootstrapped distribution is based on the assumption that the observed data represents the population.
- **How it's generated**:
  - We take repeated samples **with replacement** from the actual observed dataset.
  - For each bootstrapped sample, we compute the test statistic (e.g., mean health score change).
  - This process is repeated many times (e.g., 1,000 or more bootstrapped samples), creating a distribution of test statistics.
- **Purpose**:
  - To estimate the **sampling distribution** of the statistic directly from the data, which helps us understand the variability in our sample mean (or other statistic).
  - A **bootstrapped confidence interval** can be calculated directly from the percentiles of this distribution (e.g., the 2.5th and 97.5th percentiles for a 95% CI).

Demo1: Synthetic (Simulated) Sampling Distribution under the Null Hypothesis (H_0) vs. Bootstrapped Sampling Distribution

Key Differences:

Aspect	Synthetic Sampling under H_0	Bootstrapped Sampling
Assumption	Null hypothesis (H_0) is true (no effect).	No assumption about H_0; relies on observed data.
Purpose	To compare observed statistic to a distribution generated under H_0.	To estimate the sampling variability of the statistic from data.
How It's Generated	Simulate new datasets assuming H_0 is true.	Resample with replacement from the observed dataset.
Comparison	Used to calculate a p-value by checking the extremeness of the observed statistic under H_0.	Used to calculate confidence intervals or uncertainty around the statistic.
Use in Hypothesis Testing	Directly tests H_0, checks how likely the observed statistic is under H_0.	Helps check if the null hypothesis value (e.g., 0) is within the bootstrapped CI.

Demo2

P-value: The probability of observing a test statistic as extreme or more extreme than the one observed, assuming the **null hypothesis (H_0)** is true

Significance Level (alpha): This is the pre-set threshold (commonly alpha = 0.05) for determining whether or not to reject the null hypothesis. It represents the **risk** we are willing to take of making a **Type I error** (wrongly rejecting H_0 when it's actually true).

Demo2

Type I and II Errors

Decision	Null Hypothesis is True	Null Hypothesis is False
Reject Null	**Type I Error (α chance this results from an i.i.d. sample)**	Correct Decision
Fail to Reject	Correct Decision	**Type II Error (β chance this results from an i.i.d. sample)**

Demo2

How are confidence intervals related to hypothesis testing?

A: If the null hypothesis value lies **outside** the confidence interval, you **reject H_0**. Otherwise, you fail to reject it.

What is the correct way to interpret a 95% confidence interval?

A: You can say "We are 95% confident the interval contains the true parameter." You **cannot** say "there's a 95% chance the parameter is in the interval."

Demo2

Is the p-value the probability that H_0 is true?

A: **No.** The p-value tells you the probability of observing your data, or more extreme data, **assuming H_0 is true**, not the probability that H_0 itself is true.

Q: Why is a confidence interval better than a hypothesis test?

A: A confidence interval provides more information than a hypothesis test because it gives a range of plausible values for the population parameter, not just a binary decision (reject or fail to reject H_0). It also helps to estimate the size of the effect.

Demo2

Q: Which is better, a wider or a narrower confidence interval?

A: A narrower confidence interval is generally better because it indicates a more precise estimate of the population parameter. However, it should still capture the true value with the desired confidence level. A wider interval suggests more uncertainty in the estimate.

Demo2

Common Mistakes (Hell and Wrath of Scott)

- **Misunderstanding of P-values**:
    - **P-value ≠ Probability that H_0 is true**: A p-value doesn't tell you the probability that the null hypothesis is true, only the probability of observing your data or something more extreme **assuming H_0 is true**.
    - **P-value ≠ Risk of Type I Error**: The p-value is not the same as your α level. The p-value is the **calculated** probability based on the data, while α is the **pre-set threshold**.
    - **Confidence Interval Misinterpretation**: Saying "there's a 95% probability that the parameter is in this interval" is **incorrect**. The parameter is fixed. Instead, the correct interpretation is: "We are 95% confident that the interval contains the true parameter."

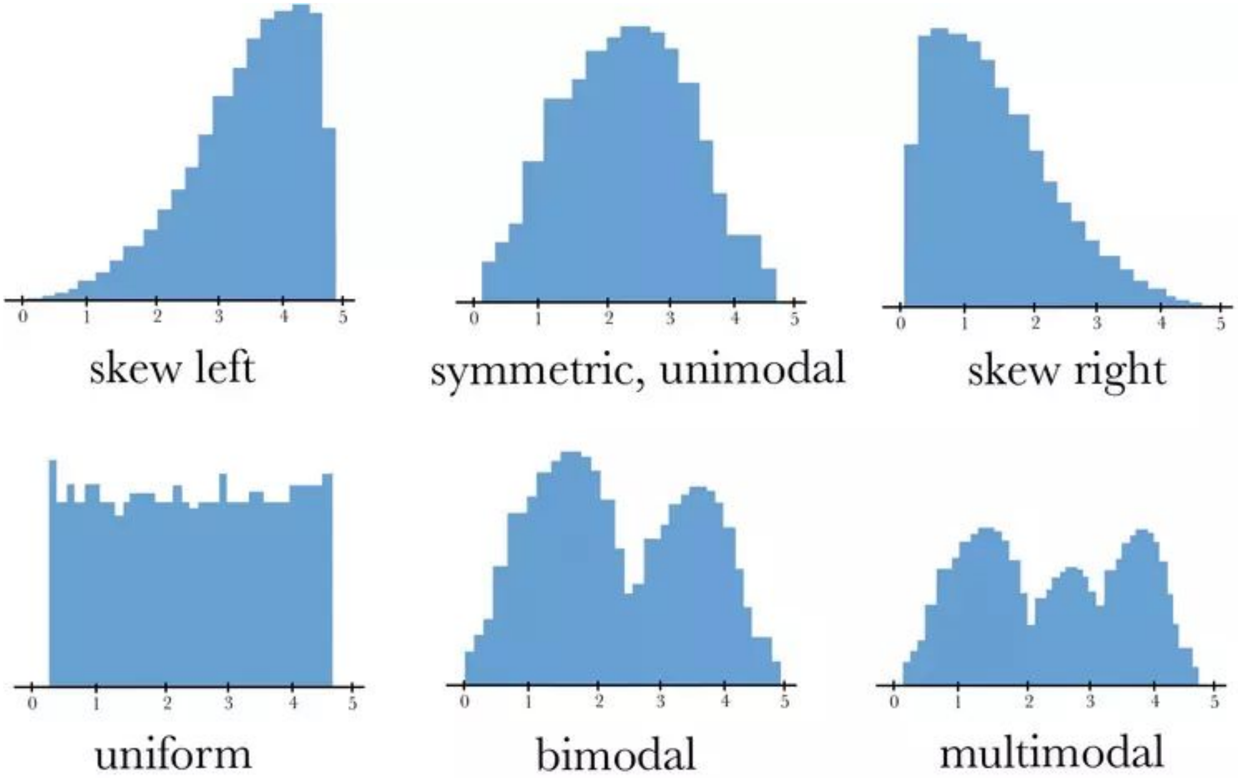
Demo2

p-value	Evidence
$p > 0.1$	No evidence against the null hypothesis
$0.1 \geq p > 0.05$	Weak evidence against the null hypothesis
$0.05 \geq p > 0.01$	Moderate evidence against the null hypothesis
$0.01 \geq p > 0.001$	Strong evidence against the null hypothesis
$0.001 \geq p$	Very strong evidence against the null hypothesis

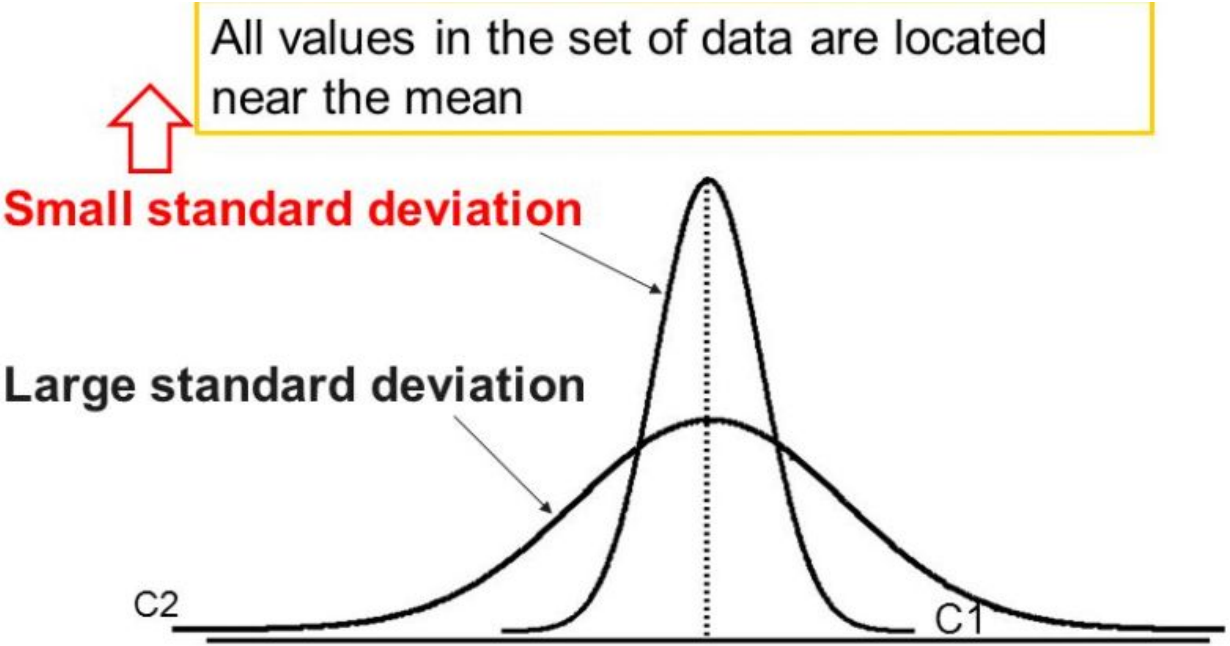
Review

Recap

Recap

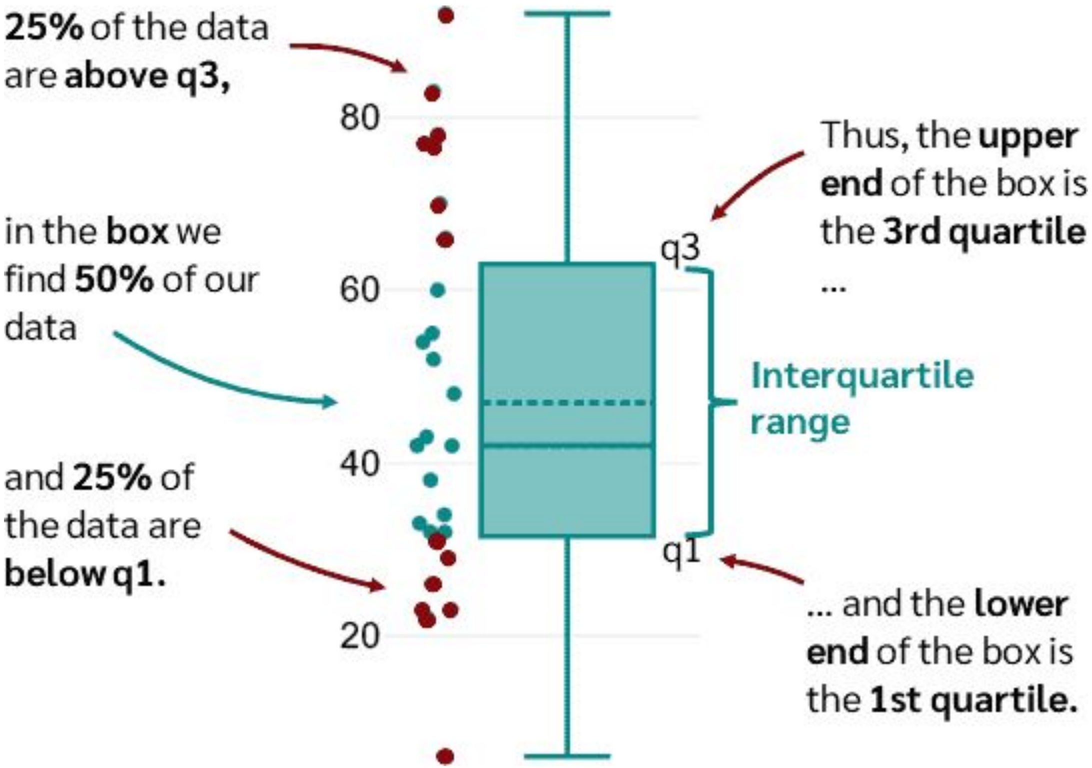

0 1 2 3 4 5
skew left
0 1 2 3 4 5
symmetric, unimodal
0 1 2 3 4 5
skew right
0 1 2 3 4 5
uniform
0 1 2 3 4 5
bimodal
0 1 2 3 4 5
multimodal

Recap



All values in the set of data are located near the mean
Small standard deviation
Large standard deviation
C2
C1

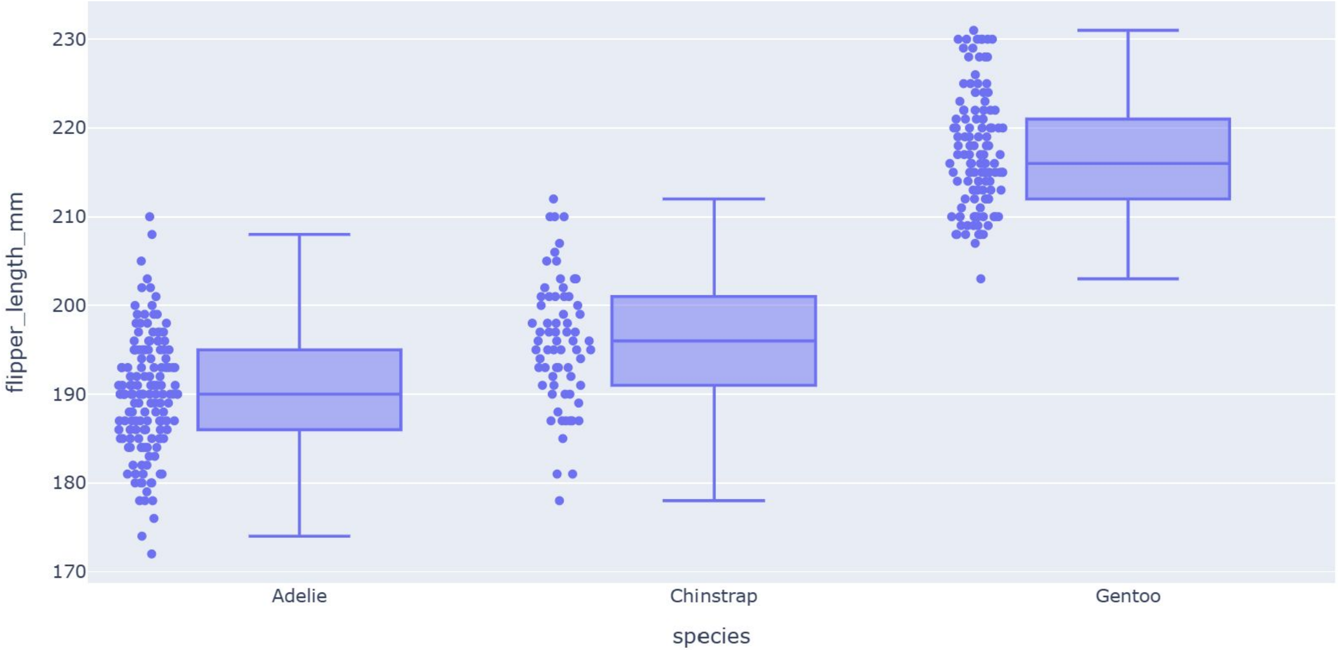
Recap



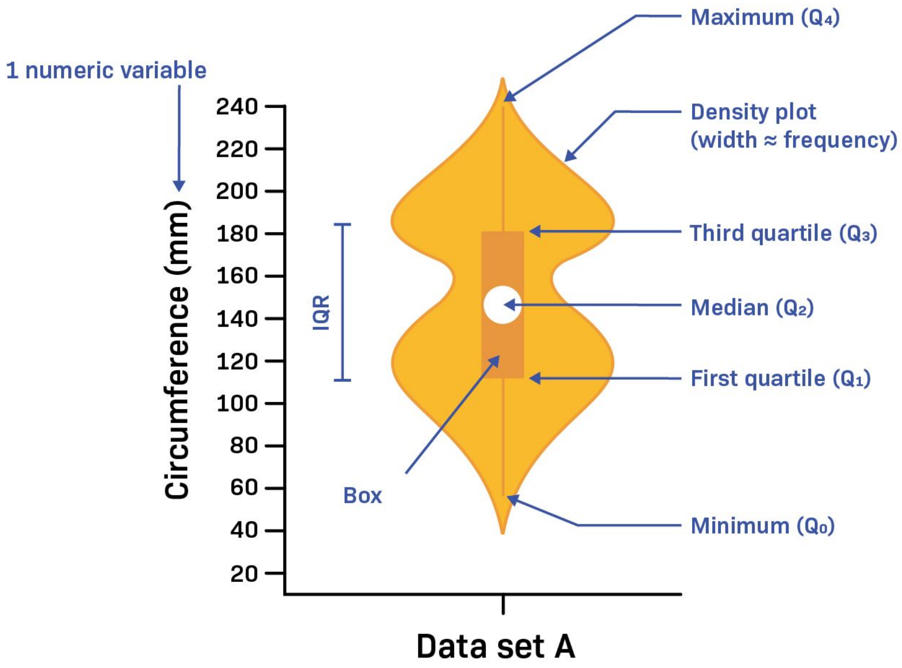
25% of the data are **above q3,**
in the **box** we find **50%** of our data
and **25%** of the data are **below q1.**
80
60
40
20
q3
q1
Thus, the **upper end** of the box is the **3rd quartile** ...
Interquartile range
... and the **lower end** of the box is the **1st quartile.**

Recap

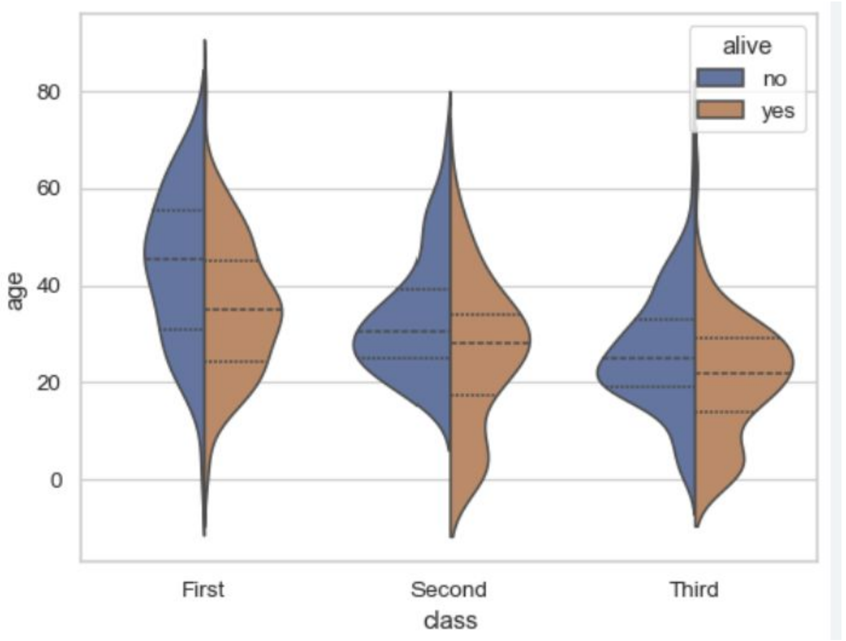
```
import plotly.express as px
fig.show()  # USE `fig.show(renderer="png")
```



Recap

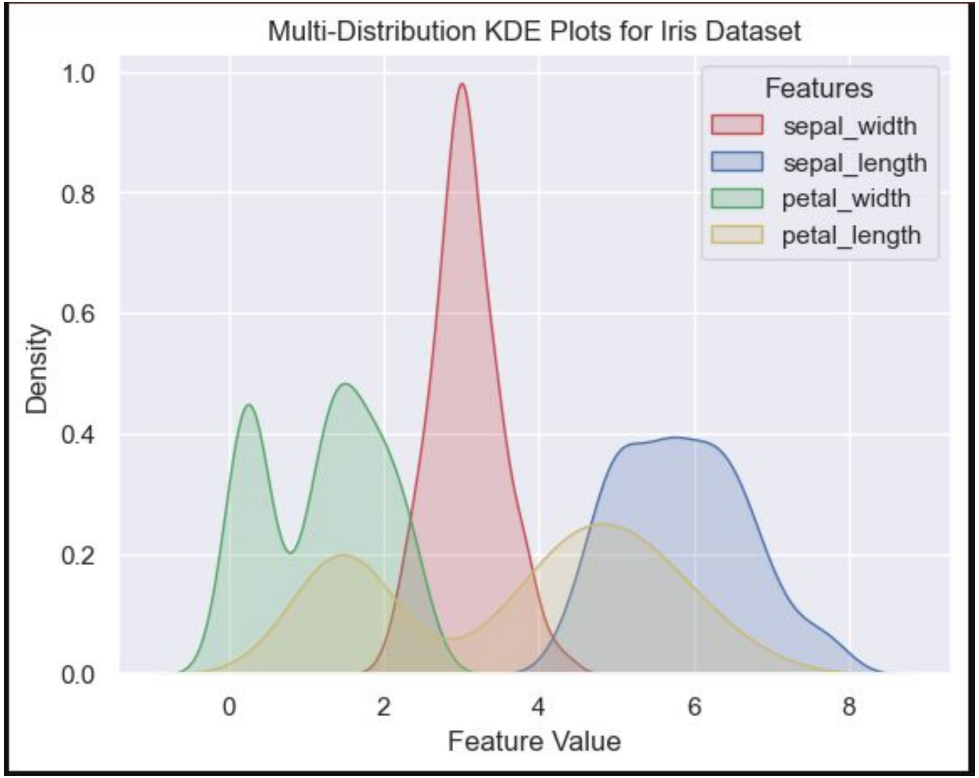


1 numeric variable
Maximum (Q4)
Density plot
(width ≈ frequency)
Third quartile (Q3)
Median (Q2)
First quartile (Q1)
Box
Minimum (Q0)
IQR
Circumference (mm)
240
220
200
180
160
140
120
100
80
60
40
20
Data set A



alive
no
yes
age
80
60
40
20
0
First
Second
Third
class

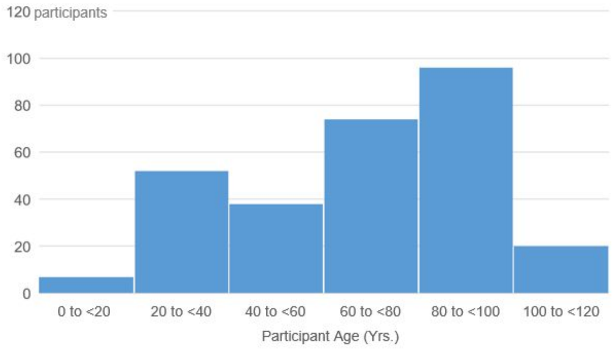
Recap



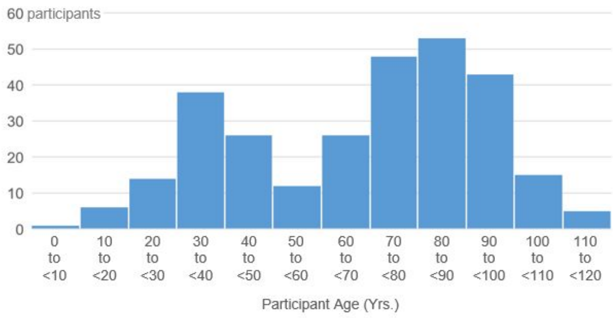
Multi-Distribution KDE Plots for Iris Dataset
Features
sepal_width
sepal_length
petal_width
petal_length
1.0
0.8
0.6
0.4
0.2
0.0
Density
0
2
4
6
8
Feature Value

Recap

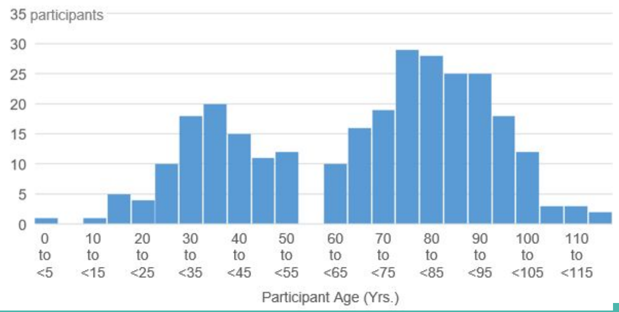
6 bins:



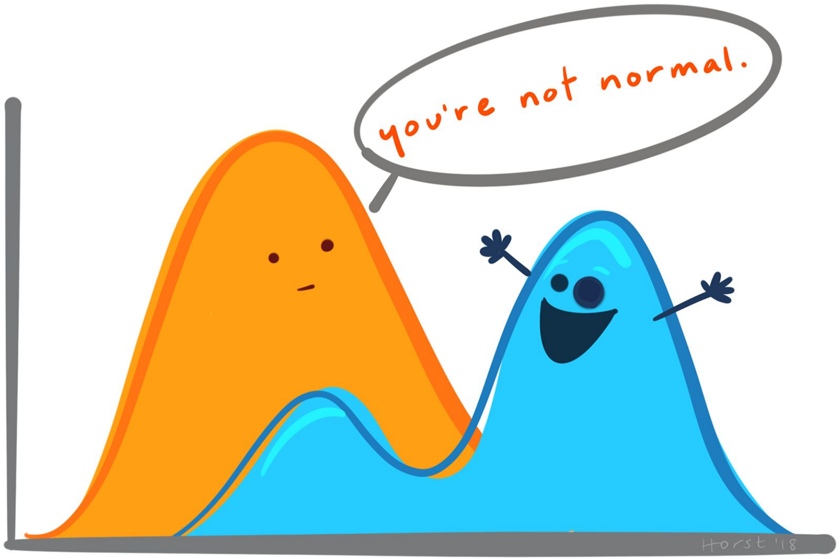
12 bins:



24 bins:



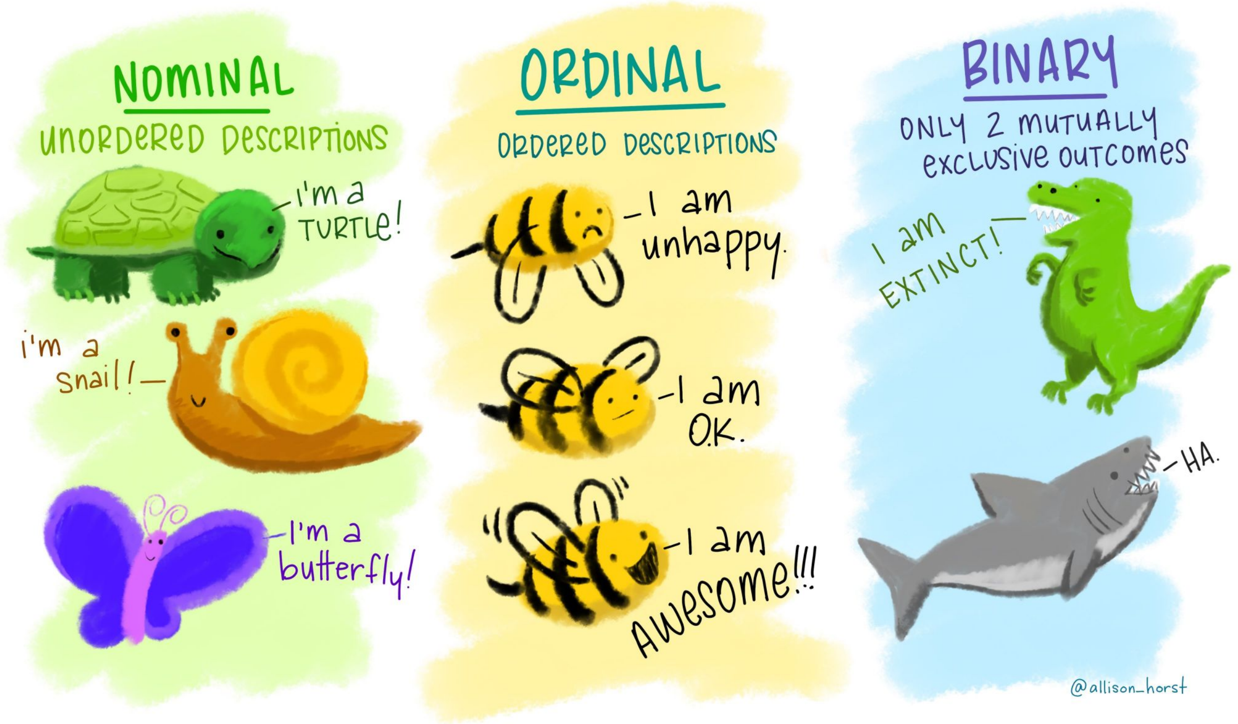
Recap



Recap


CONTINUOUS
measured data, can have ∞ values within possible range.
I am 3.1" tall
I weigh 34.16 grams
DISCRETE
observations can only exist at limited values, often counts.
I have 8 legs and 4 spots!
@allison_horst

Recap



NOMINAL
unordered descriptions
I'm a TURTLE!
i'm a snail!
I'm a butterfly!
ORDINAL
ordered descriptions
I am unhappy.
I am OK.
I am Awesome!!!
BINARY
only 2 mutually exclusive outcomes
I am EXTINCT!
HA.
@allison_horst

Recap: bootstrapping

sample = **np.random.choice**(original_data, size=len(original_data), **replace=True)**

Bootstrapping is a resampling technique used to estimate the sampling distribution of a statistic by taking repeated samples with replacement from the original sample.

Recap: sample mean vs mean

Aspect	Population Mean (Mean)	Sample Mean
Denotation	μ	$\bar{x}$
What it Represents	The average of **all members** of a population	The average of a **subset** (sample) of the population
Nature	A **fixed** value	A **variable** value that depends on the sample chosen
Purpose	Describes the **true central value** of the population	**Estimates** the population mean (μ)

Recap: sample mean vs mean

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

Recap: variance and standard deviation

Variance and Standard Deviation Formula

cuemath THE MATH EXPERT

	Population	Sample
Variance	$\sigma^2 = \dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}$
Standard Deviation	$\sigma = \sqrt{\dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}}$

```
sd = np.std(original_data, ddof=1)
```

Recap: Standard deviation vs standard error

Aspect	Standard Deviation (SD)	Standard Error (SE)
What it measures	The variability of **individual data points** in a sample/population	The variability of a **sample statistic** (e.g., sample mean)
Used for	Describing the spread of a dataset	Describing the accuracy of a sample statistic as an estimate of a population parameter
Formula	Measures the deviation of data points from the mean	Measures the deviation of sample means from the population mean
Effect of sample size	Unaffected by sample size	Decreases as sample size increases

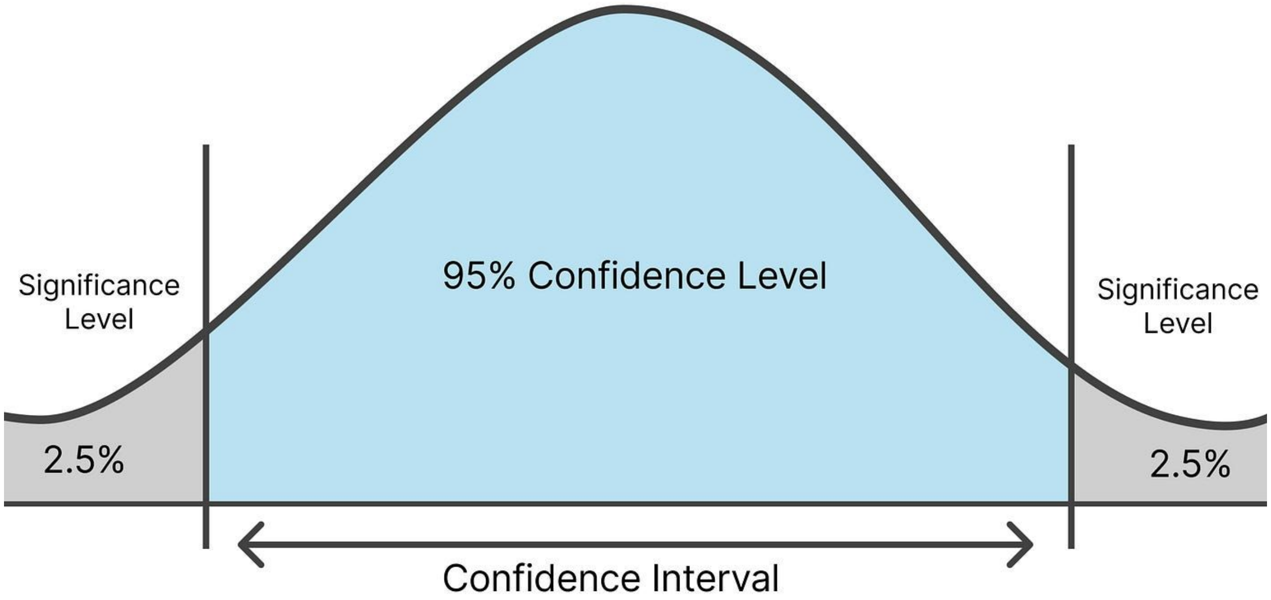
Recap: Standard deviation vs standard error

$$SD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$SE = \frac{SD}{\sqrt{n}}$$

```
sd = np.std(original_data, ddof=1)
```

Recap: confidence interval



A **confidence interval** gives us a range of values within which we are confident the true population parameter lies. For example, a **95% confidence interval** means that we are 95% confident that the interval contains the population mean.

Recap: confidence interval

```
lower_bound, upper_bound
=
np.percentile(bootstrapped
_means, [2.5, 97.5])
```

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval

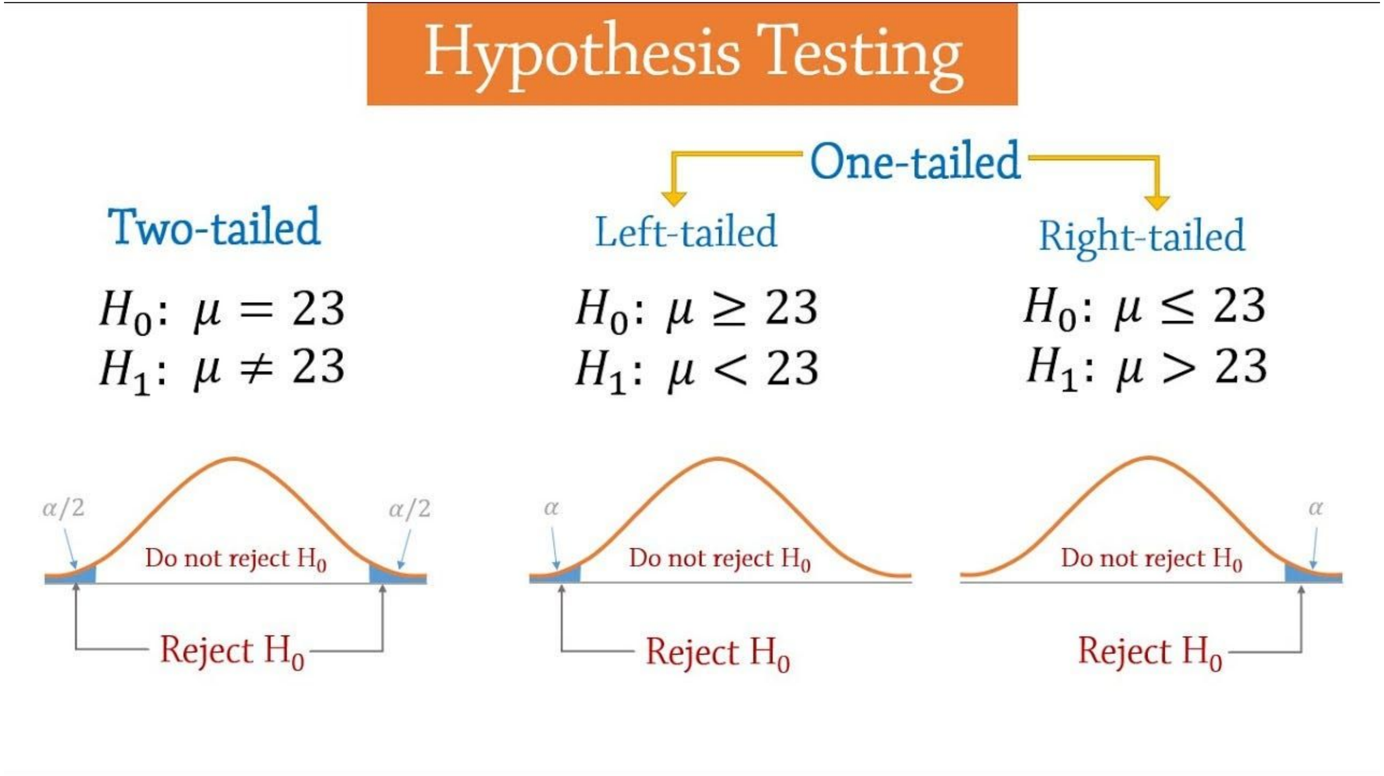
$\bar{x}$ = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

Recap: Formal Hypothesis Testing



Recap: Formal Hypothesis Testing