Statistics

Friday, 11 July 2025 12:26 PM

STATISTICS
│
├── A. DESCRIPTIVE STATISTICS (Summarizing Data)
│ │
│ ├── 1. Measures of Central Tendency (Center)
│ │ ├── Mean (Arithmetic, Geometric, Harmonic)
│ │ ├── Median
│ │ └── Mode
│ │
│ ├── 2. Measures of Dispersion (Spread)
│ │ ├── Range
│ │ ├── Variance
│ │ ├── Standard Deviation
│ │ └── Interquartile Range (IQR) & Percentiles
│ │
│ └── 3. Shape of Distribution
│ ├── Skewness (Asymmetry)
│ └── Kurtosis (Tail heaviness)
│
└── B. INFERENTIAL STATISTICS (Making Predictions)
│
├── 1. Estimation
│ ├── Point Estimates
│ └── Confidence Intervals
│
└── 2. Hypothesis Testing
├── Parametric Tests (Assume Normal Distribution)
│ ├── Z-test (Population variance known)
│ ├── T-test (Population variance unknown)
│ └── ANOVA (Comparing 3+ groups)
│
└── Non-Parametric Tests (No assumption of Normal Dist.)
├── Chi-Square (Categorical data)
├── Mann-Whitney U (Median comparison)
└── Kruskal-Wallis (ANOVA alternative)

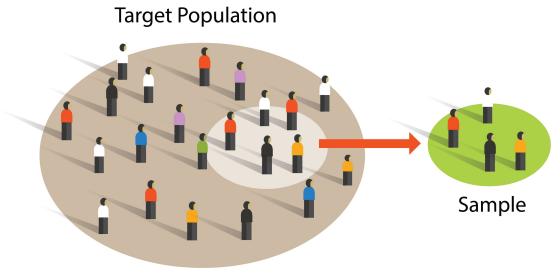
1.1 Descriptive vs. Inferential Statistics

- **Descriptive Statistics:** Techniques used to summarize, organize, and visualize data. It tells you *what happened* in your specific dataset.
  - *Example:* Calculating the average age of students in a specific classroom.

- **Inferential Statistics:** Techniques that use sample data to make generalizations (inferences) about a larger population. It helps you predict *what might happen*.
  - *Example:* Surveying 100 voters to predict who will win the national election.

1.2 Population vs. Sample

- **Population (N):** The entire group that you want to draw conclusions about.
- **Sample (n):** A specific group that you will collect data from. The sample size is always less than the total population.
  - *Analogy:* The population is the entire pot of soup; the sample is the spoonful you taste to check the seasoning.



1.3 Sampling Techniques

How we select our sample matters to avoid bias.

1. **Simple Random Sampling:** Every member has an equal chance of being selected (e.g., names out of a hat).
2. **Stratified Sampling:** Dividing the population into subgroups (strata) and sampling from each (e.g., ensuring a survey has equal representation of men and women).
3. **Systematic Sampling:** Selecting every k^{th} individual (e.g., every 10th person in a line).
4. **Cluster Sampling:** Dividing the population into clusters and randomly selecting entire clusters.

2.1 Numerical (Quantitative) Data

Data that represents amounts or quantities.

- **Continuous:** Can take any value within a range (infinite possibilities).
  - *Examples:* Height (175.5 cm), Temperature, Time.
- **Discrete:** Can only take specific, countable values (integers).
  - *Examples:* Number of students in a class, shoe size, number of cars.

2.2 Categorical (Qualitative) Data

Data that represents groups or characteristics.

- **Nominal:** Categories with no intrinsic order.
  - *Examples:* Colors (Red, Blue), Gender, Zip Codes.
- **Ordinal:** Categories with a clear rank or order.
  - *Examples:* Education level (High School < Bachelor's < Master's), Customer satisfaction rating (Satisfied > Neutral > Dissatisfied).

Module 3: Descriptive Statistics (Central Tendency)

Goal: Find the single value that best represents the center of the data.

3.1 The "Big Three"

1. **Mean (μ or $\bar{x}$):** The arithmetic average.
   - *Formula:* $\frac{\sum x_i}{n}$
   - *Pros/Cons:* Uses all data but is highly sensitive to outliers.
2. **Median:** The middle value when data is ordered.
   - *Pros/Cons:* Robust against outliers; better for skewed distributions (like salaries).
3. **Mode:** The most frequent value.
   - *Use case:* The only measure usable for Nominal data.

```
import numpy as np
from scipy import stats

data = [10, 20, 20, 40, 50, 100] # 100 is an outlier
mean = np.mean(data)      # 40.0 (skewed by outlier)
median = np.median(data)  # 30.0 (more representative)
mode = stats.mode(data)   # 20
```

3.2 Deep Dive: Central Tendency & Skewness

3.2.1 When to use which measure?

Measure	Best Used When...	Pros	Cons
Mean	Data is numerical, symmetric, and has no outliers.	Uses every data point; required for many advanced formulas (variance).	Highly sensitive to outliers (one billionaire in a room skews the average income).
Median	Data is skewed (salaries, home prices) or has outliers.	Robust; ignores extreme values.	Mathematical operations are harder to perform on it compared to mean.
Mode	Data is Categorical (Nominal) or you need the "most popular" item.	The only measure for non-numeric data.	Can be unstable (sometimes there is no mode, sometimes there are two).

3.2.2 Skewness and Central Tendency

Skewness tells us where the "tail" of the data is pulling the mean. The relative position of the Mean, Median, and Mode tells you the shape of your distribution.

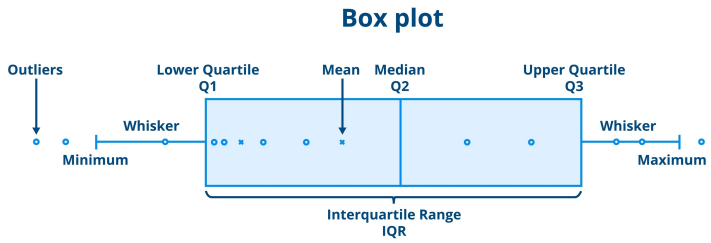
1. **Symmetrical (Normal Distribution):**
   - **Rule:** Mean = Median = Mode
   - *Visual:* A perfect bell curve.
2. **Positive Skew (Right Skewed):**
   - **Rule:** Mean > Median > Mode
   - *Visual:* The tail is on the right side.
   - *Example:* Wealth distribution. (The super-rich pull the Mean to the right, but the Median person is still middle-class).
3. **Negative Skew (Left Skewed):**
   - **Rule:** Mean < Median < Mode
   - *Visual:* The tail is on the left side.
   - *Example:* Age at death (Most people live to old age, but a few young deaths pull the mean down/left).

Module 4: Measures of Spread (Dispersion)

Goal: Understand how "spread out" or varied the data is.

4.1 Basic Measures

- **Range:** Max value - Min value. Very sensitive to outliers.
- **Percentiles:** The value below which a percentage of data falls.
  - *Code:* np.percentile(data, 50) is the same as the Median.
- **Quartiles:** Splits data into four equal parts.
  - $Q1$: 25th percentile.
  - $Q2$: 50th percentile (Median).
  - $Q3$: 75th percentile.
- **Interquartile Range (IQR):** The range of the middle 50% of the data. Used to detect outliers.
  - *Formula:* $IQR = Q3 - Q1$

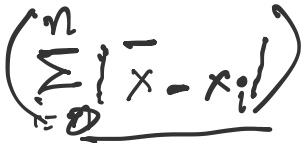


4.2 Advanced Measures of Spread

Mean Absolute Deviation (Mean AD)

The average distance between each data point and the mean.

- *Formula:* $\frac{1}{n} \sum |x_i - \bar{x}|$
- *Interpretation:* How far, on average, is a point from the mean?



Variance (σ^2)

The average of the *squared* differences from the mean

The average of the *squared* differences from the mean.

- *Formula (Population):* $\sigma^2 = \frac{1}{N} \sum (x_i - \mu)^2$
- *Formula (Sample):* $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$
- *Code:* np.var(data)
- *Concept:* Squaring punishes outliers more heavily than Mean AD.

Standard Deviation (σ)

The square root of the variance. It returns the measure to the original unit of the data (e.g., from "cm^2" back to "cm").

- *Formula:* $\sigma = \sqrt{\text{Variance}}$
- *Code:* np.std(data)

Median Absolute Deviation (Median AD)

A robust measure of variability. It uses the median instead of the mean.

- *Concept:* It is to the standard deviation what the median is to the mean (resistant to outliers).
- *Code:*
  Python

  ```
  from statsmodels import robust
  mad_val = robust.mad(data)
  ```

4.3 Why Variance over Mean AD? (The Smoothness Property)

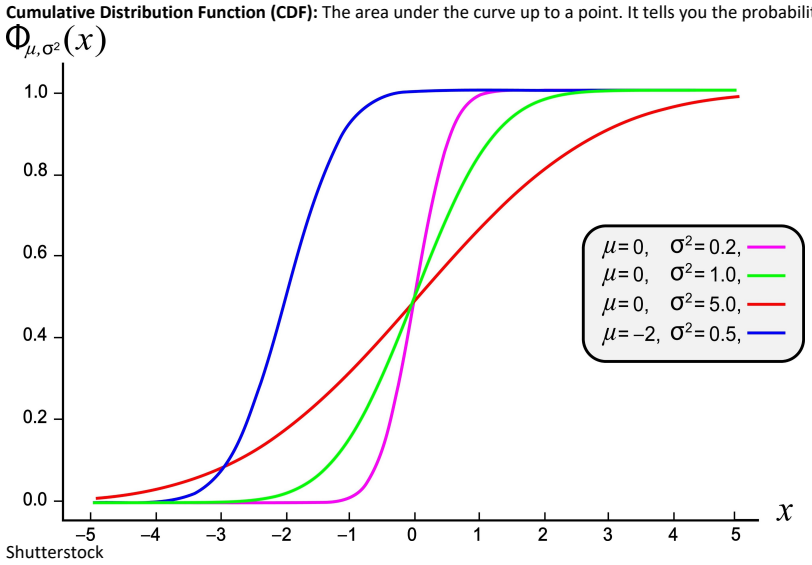
You asked about the "smooth function" difference.

- **Mean AD** uses an absolute value function ($|x|$). The graph of $|x|$ has a sharp "V" shape at the bottom. In calculus, you cannot take the derivative at that sharp point (it is non-differentiable at 0).
- **Variance** uses a square function (x^2). The graph is a parabola (smooth U-shape). It is differentiable everywhere.
- **Why it matters:** Many advanced machine learning algorithms (like Gradient Descent in Neural Networks) rely on derivatives to find the minimum error. Variance is mathematically "easier" to work with for optimization than Mean AD.

Module 5: Probability Distributions

5.1 PDF vs. CDF

- **Probability Density Function (PDF):** The height of the curve. For continuous data, it shows the relative likelihood of a value.
- **Cumulative Distribution Function (CDF):** The area under the curve up to a point. It tells you the probability that X will take a value *less than or equal to* x.



Shutterstock

5.2 The Normal Distribution (Gaussian)

The famous "Bell Curve."

- Symmetric about the mean.
- Mean = Median = Mode.
- **68-95-99.7 Rule:** 68% of data is within 1σ, 95% within 2σ, 99.7% within 3σ.

5.3 Z-Score & Z-Table

Standardization allows us to compare apples to oranges by converting distributions to a "Standard Normal Distribution" (Mean=0, Std Dev=1).

- *Formula:* $Z = \frac{x - \mu}{\sigma}$
- *Interpretation:* A Z-score tells you how many standard deviations a data point is away from the mean.

5.4 Central Limit Theorem (CLT)

The "Magic" of statistics.

- *Definition:* If you take sufficiently large random samples from *any* population (even if it's not normal) and calculate the means of those samples, the distribution of those sample means will form a Normal Distribution.
- *Application:* This is why we can use hypothesis testing on weirdly shaped real-world data.

Module 6: Inferential Statistics

Hypothesis Testing

A method to test a claim about a population parameter.

1. **Null Hypothesis (H_0):** The status quo; no effect exists (e.g., "The drug has no effect").
2. **Alternative Hypothesis (H_1):** The claim you want to prove (e.g., "The drug lowers blood pressure").
3. **P-value:** The probability of seeing your results if H_0 were true.
   - *Rule of Thumb:* If P-value < 0.05, we reject H_0 (The result is statistically significant).

6.1 Covariance and Correlation

Descriptive stats often look at one variable (Univariate). These look at two variables (Bivariate).

- **Covariance:** Tells you the *direction* of the relationship.
  - Positive: Both grow together.
  - Negative: One goes up, the other goes down

    - Negative: One goes up, the other goes down.
    - *Problem:* The number isn't standardized (it could be 5 or 5000).
- **Correlation (Pearson's r):** Standardizes Covariance between -1 and +1.
    - +1: Perfect positive linear relationship.
    - -1: Perfect negative linear relationship.
    - 0: No linear relationship.

6.1.2 Covariance ($\text{Cov}(X, Y)$)

- **Definition:** A measure of the joint variability of two random variables.
- **Formula:**
  $$\text{Cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$
- **Interpretation:**
    - **Positive (+):** As X increases, Y tends to increase.
    - **Negative (-):** As X increases, Y tends to decrease.
    - **Limitation:** The value depends on the units (e.g., covariance of height in meters vs. weight in kg will be tiny; in cm vs. grams, it will be huge). It's hard to interpret the *strength* of the relationship.

$$cov(X,Y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

6.1.3 Correlation (Pearson's r)

- **Definition:** The standardized version of covariance. It divides covariance by the standard deviations of X and Y to remove unit bias.
- **Formula:**
  $$r = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$
- **Range:** Always between -1 and $+1$.
    - **$+1$:** Perfect positive linear relationship.
    - **-1:** Perfect negative linear relationship.
    - **0:** No linear relationship.

$$r = \frac{cov(X,Y)}{\delta_x \ \delta_y}$$

Correlation tells you the strength of the relationship between the variables
Covariance just tells you the direction