Data Types

Nominal: Categories, no order (e.g., colors).
Ordinal: Ranked categories (e.g., movie ratings).
Discrete: Countable (e.g., books).
Continuous: Measurable (e.g., temperature).

Descriptive Stats

Mean: $\bar{X} = \frac{\sum X}{N}$,
Variance: $\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$,
Std Dev: $\sigma = \sqrt{\sigma^2}$,
Z-score: $Z = \frac{X-\mu}{\sigma}$.

Confidence Intervals

A range of values likely to contain the true population parameter.
Large n: $CI = \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.
Small n: $CI = \bar{X} \pm t_{\alpha/2} SE$, where $SE = \frac{s}{\sqrt{n}}$.

Hypothesis Testing

H_0: No effect
H_a: Effect exists.
p-value: $p < 0.05 \rightarrow$ Reject H_0, else Fail to reject H_0.

Correlation & Regression

Pearson Correlation:
$r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}}$.
Linear Regression: $Y = b_0 + b_1 X$, where b_1 is the slope.

Z-Scores Table

Z	Prob
1.64	0.9495
1.96	0.9750
2.00	0.9772
2.33	0.9901
2.58	0.9949

t-Distribution (95% CI)

n	t-Value
5	2.776
10	2.262
15	2.145
20	2.093
30	2.042

Percentiles (Interpolation)

$i = 1 + (n-1) \times p$.
If i is not an integer, interpolate:
$P = X_i + (X_{i+1} - X_i) \times (i - \lfloor i \rfloor)$.
Example (75th percentile): $i = 1 + 7 \times 0.75 = 6.25 \rightarrow$ Interpolate between 6th and 7th value.

Diminishing Returns

More study time → Higher scores, but improvements decrease at high study hours.

Spark & Data Engineering

Why Spark? - Scales better than Pandas for large datasets. - Distributed, in-memory processing. - Handles structured & unstructured data.

RDD vs. DataFrames vs. Datasets

Feature	RDD	DataFrame/Dataset
Optimization	None	Query Optimized (Catalyst)
Storage	Distributed objects	Columnar format (Parquet, ORC)
Schema	Unstructured	Schema-aware
Use Case	Low-level ops	SQL-like queries, ML Pipelines

Key Concepts:
- Lazy Evaluation: Computation occurs only when an action is triggered (e.g., '.collect()', '.count()').
- Transformations (lazy) vs. **Actions** (triggers execution).
- Shuffling Impact: Moving data between partitions slows performance.
- Partitioning: Avoids unnecessary shuffling; optimize using '.repartition()' or '.coalesce()'.

Performance Optimization:
- Broadcast Variables: Share small data efficiently across nodes.
- Cache/Persist: Avoid recomputation, speeds up repeated queries.
- Avoid Collect(): Prevents bringing too much data to the driver.
- Optimize Joins: Prefer **broadcast joins** for small tables.

Lazy Evaluation: Actions only compute when triggered. Shuffling Impact: Slows down performance.

Sampling Distributions

Definition: The distribution of a sample statistic (e.g., mean).
Central Limit Theorem: For large n, sample means follow a normal distribution.

Common Statistical Tests

Test	Use Case
Z-Test	$n > 30$, known variance
t-Test	$n < 30$, unknown variance
Chi-Square	Categorical data
ANOVA	Compare multiple groups

Probability Distributions

Normal: Symmetric bell curve.
Exponential: Skewed right, models time until event.
Poisson: Models rare events over time.

Machine Learning

Overfitting: Model too complex, captures noise.
Underfitting: Model too simple, poor predictions.
Bias-Variance Tradeoff:
- High Bias: Too simple, underfitting.
- High Variance: Too complex, overfitting.

Key Formulas Recap

Mean: $\bar{X} = \frac{\sum X}{N}$,
Variance: $\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$,
Z-score: $Z = \frac{X-\mu}{\sigma}$,
Correlation: $r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}}$,
Regression: $Y = b_0 + b_1 X$,
CI: $\bar{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$.

Sample Covariance

$\mathrm{Cov}(X,Y) = \frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y})$
Example: $X = \{10, 20, 30, 40\}, Y = \{15, 30, 60, 90\}$
$\bar{X} = 25, \quad \bar{Y} = 48.75$
$\mathrm{Cov}(X,Y) = \frac{1}{3}[(-15)(-33.75) + (-5)(-18.75) + (5)(11.25) + (15)(41.25)] = 425$

Law of Large Numbers (LLN)

As n increases, the sample mean $\bar{X}$ converges to the population mean μ:
$\lim_{n\to\infty} P(|\bar{X} - \mu| < \epsilon) = 1$
Larger samples provide more reliable estimates.

Standard Error (SE)

Measures accuracy of the sample mean:
$SE = \frac{\sigma}{\sqrt{n}}$
For a sample:
$SE = \frac{s}{\sqrt{n}}$
Smaller SE means more precise estimates.

Bayes' Theorem

$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$
Example: Disease Testing - $P(H)$: Probability of having the disease - $P(D \mid H)$: Probability of a positive test if diseased - $P(D)$: Probability of any positive test
$P(H \mid D) = \frac{P(D|H)P(H)}{P(D)}$

Linear Interpolation Example

Given the sample set: $S = \{-1, 3, -4, 2, 6\}$
(a) **Range:** Range $= \max(S) - \min(S) = 6 - (-4) = 10$
(b) **Median:** Sorted data: $S = \{-4, -1, 2, 3, 6\}$ Median $= 2$
(c) **Mean:** Mean $= \frac{-1+3+(-4)+2+6}{5} = \frac{6}{5} = 1.2$
(d) **20th Percentile using Linear Interpolation:**
Find position: $P = (n-1) \times p + 1 = (5-1) \times 0.2 + 1 = 1.8$
Since $P = 1.8$, interpolate between the 1st and 2nd values in the sorted set:
20th Percentile $= S_1 + (P-1) \times (S_2 - S_1)$
$= -4 + 0.8 \times (-1 - (-4))$
$= -4 + 0.8 \times 3$
$= -4 + 2.4 = -1.6$
Thus, the 20th percentile is -1.6.

Effects of Outliers

Given: $S = [3, 8, 6, 9, -1, 10, 1000, 7, 7, 0]$
(a) Mean of S: Mean $= \frac{3+8+6+9+(-1)+10+1000+7+7+0}{10} = \frac{1049}{10} = 104.9$
(b) Median of S: Sorted: $S_{\text{sorted}} = [-1, 0, 3, 6, 7, 7, 8, 9, 10, 1000]$ Median $= \frac{7+7}{2} = 7$ Difference $= 104.9 - 7 = 97.9$
(c) Skewness: Right-skewed since Mean > Median.
(d) 10% Trimmed Mean: Trimmed set: $S_{\text{trimmed}} = [0, 3, 6, 7, 7, 8, 9, 10]$
Trimmed Mean $= \frac{0+3+6+7+7+8+9+10}{8} = 6.25$
(e) 80th Percentile (Interpolation): $P = (10-1) \times 0.8 + 1 = 8.2$
80th Percentile $= 9 + (0.2 \times (10-9)) = 9.2$
(g) 90th Percentile (Interpolation): $P = (10-1) \times 0.9 + 1 = 9.1$
90th Percentile $= 10 + (0.1 \times (1000-10)) = 109$

Handling Outliers

Replace $1000 \to 100$, recompute:
Sorted: $S'_{\text{sorted}} = [-1, 0, 3, 6, 7, 7, 8, 9, 10, 100]$
Mean $= \frac{-1+0+3+6+7+7+8+9+10+100}{10} = 14.9$
Median $= \frac{7+7}{2} = 7$
80th Percentile $= 9 + (0.2 \times (10-9)) = 9.2$
90th Percentile $= 10 + (0.1 \times (100-10)) = 19$

Short Answer: Show Your Work

29. **Percentiles**
Suppose you have a data set of exam scores: 65, 70, 72, 78, 80, 85, 90, 95.
Find the 75th percentile using linear interpolation if needed.

Use interpolation method

$i = 1 + (n-1) \cdot p$
$1 + 7 \cdot .75$
$i = 6.25$

$i = 0.75 \cdot (8+1)$
$i = 6.75$

30. **Z-Score Calculation**
A population of exam scores is approximately normal with mean \(\mu = 80\) and standard deviation \(\sigma = 5\).
What is the Z-score for a student who scores 90 on the exam?

$Z = \frac{x - \mu}{\sigma}$
$= \frac{90-80}{5}$
$= 2$

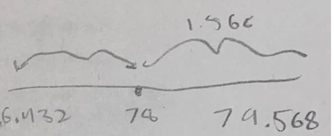
. **Confidence Interval (Large n)**
You collect a sample of **n=100** exam scores, with a sample mean \(\overline{x} = 78\) and **known** population standard deviation \(\sigma = 8\).
Construct a **95% confidence interval** for the population mean.
(Use the Z-interval since n is large and \(\sigma\) is known. See the table on the next page)

$Z^{*} = 1.96$ 0.975

$\bar{x} \pm Z^{*} \cdot SE$
* critical z-score

$SE = \frac{8}{10} = \frac{8}{\sqrt{n}}$ 100 $= .8$

$\bar{x} \pm 1.96 \cdot .8$
78 ± 1.568



. **Confidence Interval (Small n)**
Now consider a smaller sample, **n=15**, with a sample mean of 78 and a **sample standard deviation** of 8.
Construct a **95% confidence interval** for the population mean.
(Use the **t-distribution** here because n is small and \(\sigma\) is not known. See the table on the next page)

30
$i = 2.145$

$\bar{x} \pm z^{*} \cdot SE$

$\bar{x} \pm 2.145 \cdot SE$

$SE = \frac{5}{\sqrt{n}} = \frac{5}{\sqrt{15}}$

$\bar{x} \pm 4.431 = 78 \pm 4.431$