📘Complete Statistics Notes (Descriptive + Inferential)

This comprehensive note combines **Descriptive Statistics** and **Inferential Statistics** — with clear concepts, step-by-step formulas, visual understanding, and real-world examples.

🪠 1. Introduction to Statistics

📚What is Statistics?

Statistics is the **science of collecting, organizing, analyzing, and interpreting data** to make informed decisions.

There are two broad branches:

Type	Description	Example
Descriptive Statistics	Summarizes and describes features of data	Mean, Median, Mode, Graphs
Inferential Statistics	Makes conclusions about a population from sample data	Hypothesis tests, Confidence Intervals

---

🧰2. Descriptive Statistics

2.1 Data Types

Type	Description	Example
Numerical (Quantitative)	Represent measurable quantities	Age, Income, Weight
Categorical (Qualitative)	Represent labels or names	Gender, City, Brand
Ordinal	Ordered categories	Satisfaction level: Low < Medium < High

---

2.2 Measures of Central Tendency

Measure	Description	Formula	Example
Mean (Average)	Sum of all observations ÷ total count	$\bar{X} = \frac{\Sigma X}{n}$	(2+4+6)/3 = 4
Median	Middle value when data sorted	-	[10, 15, 20] → 15

Measure	Description	Formula	Example
Mode	Most frequent value	-	[2, 4, 4, 5] → 4

2.3 Measures of Dispersion

Measure	Description	Formula	Interpretation
Range	Max − Min	$R = X_{max} - X_{min}$	Larger → more spread
Variance	Avg. of squared deviation	$\sigma^2 = \Sigma(X-\mu)^2 / N$	Higher variance = more spread
Standard Deviation (SD)	√Variance	$\sigma = \sqrt{\sigma^2}$	Expresses spread in same units
IQR	Q3 − Q1	-	Measures middle 50% spread

Example:
Heights = [150, 160, 170, 180, 190] → Mean = 170, SD ≈ 15.8

2.4 Shape of Distribution

Shape	Description	Mean vs Median
Normal	Bell curve, symmetric	Mean = Median
Positive Skew (Right)	Tail to right	Mean > Median
Negative Skew (Left)	Tail to left	Mean < Median

2.5 Outliers and Boxplot

Outliers: Values lying far from most data points.
Detection (IQR method):

$$Lower = Q1 - 1.5 \times IQR, \quad Upper = Q3 + 1.5 \times IQR$$

2.6 Visualization with Matplotlib and Seaborn

```
import matplotlib.pyplot as plt
import seaborn as sns
sns.boxplot(x=df['Height'])
sns.histplot(df['Height'], kde=True)
```

🧮3. Inferential Statistics

Inferential Statistics allows drawing conclusions about a **population** using a **sample**.

3.1 Population vs Sample

Term	Symbol	Definition
Population	N	Entire group (all people, items)
Sample	n	Subset of the population

3.2 Sampling Methods

- **Simple Random Sampling:** Equal chance for all members.
- **Stratified Sampling:** Divide into strata (age, gender).
- **Systematic Sampling:** Pick every k^{th} element.
- **Convenience Sampling:** Based on availability.

🧪4. Hypothesis Testing

4.1 Key Terms

Term	Symbol	Meaning
Null Hypothesis	H_0	No difference/effect
Alternative Hypothesis	H_1	There is a difference/effect
α (Alpha)	Significance level = Type I Error	Usually 0.05
β (Beta)	Type II Error	Probability of missing real effect
Power (1−β)	Probability of detecting real effect	Usually 0.80

4.2 Decision Rules

p-value	Interpretation	Decision
$p \le \alpha$	Unlikely under H_0	Reject H_0
$p > \alpha$	Likely under H_0	Fail to reject H_0

4.3 Example: Tire Lifespan (Z-Test)

A company claims the average tire life = 50,000 km.
Sample of n=40 tires → mean = 48,500 km, σ=3,000 km (known).

Step 1: Hypotheses
H_0: μ = 50,000
H_1: μ < 50,000 (left-tailed test)

Step 2: Compute Z

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{48500 - 50000}{3000/\sqrt{40}} = -3.16$$

Step 3: Compare
Zcritical (α=0.05, left-tail) = -1.645 → -3.16 < -1.645 → Reject H_0

✅**Conclusion:** Tires last **less than 50,000 km** → claim is false.

4.4 p-Value Explained

> The **p-value** is the probability of getting results at least as extreme as the observed ones **if H_0 were true**.

- **Small p-value (≤ α):** strong evidence against H_0 → reject it.
- **Large p-value (> α):** weak evidence → fail to reject H_0.

Example: p = 0.0027 < 0.05 → reject H_0 → strong evidence the claim is false.

---

5. Type I and Type II Errors

Error	Description	Symbol	Analogy
Type I Error	Reject true H_0 (False Positive)	α	Convicting an innocent person
Type II Error	Fail to reject false H_0 (False Negative)	β	Freeing a guilty person
Power (1−β)	Correctly detecting true effect	—	Accuracy of the test

Trade-off: Lower α → more strict → higher β (and vice versa).
Fix: Increase sample size → lowers both α and β.

🧮6. Common Hypothesis Tests

Test	When to Use	Formula	Data Type	Python Function
Z-Test	Compare mean with known σ, $n>30$	$(\bar{x}-\mu)/(\sigma/\sqrt{n})$	Continuous	`norm.cdf()`
t-Test	Compare means, σ unknown, $n<30$	$(\bar{x}-\mu)/(s/\sqrt{n})$	Continuous	`ttest_1samp()`, `ttest_ind()`
Chi-Square Test	Test independence (categorical data)	$\Sigma(O-E)^2/E$	Categorical	`chi2_contingency()`

🧰7. Chi-Square Example: Gender vs Product Preference

Product	Male	Female	Total
A	30	10	40
B	20	30	50
Total	50	40	90

Expected Frequencies (E):

$$E = \frac{(Row\ Total) \times (Column\ Total)}{Grand\ Total}$$

Product	Male (E)	Female (E)
A	22.22	17.78
B	27.78	22.22

Chi-Square:

$$\chi^2 = \Sigma(O-E)^2/E = 9.8$$

df = 1, α=0.05 → $\chi^2_{(critical)}$ = 3.84 → 9.8 > 3.84 → Reject H_0

✅**Conclusion:** Gender and product preference **are related** (not independent).

📈8. Confidence Intervals

A **confidence interval (CI)** gives a range within which the true population parameter likely lies.

Formula:

$$CI = \bar{X} \pm Z_{\alpha/2}(\sigma/\sqrt{n})$$

Example: If $\bar{X} = 48,500$, σ = 3,000, n = 40, α = 0.05 (Z = 1.96)

$$CI = 48,500 \pm 1.96 \times (3000/\sqrt{40}) = 48,500 \pm 929 \rightarrow (47,571, 49,429)$$

Interpretation:
We are 95% confident the true mean lifespan is between **47,571 km and 49,429 km**.

🧮9. Correlation and Covariance

Concept	Formula	Meaning
Covariance	$\Sigma(X-\bar{X})(Y-\bar{Y})/(n-1)$	Measures direction of relationship
Correlation (r)	$Cov(X,Y)/(\sigma x \sigma y)$	Measures strength ($-1 \leq r \leq 1$)

Example:
If r = +0.85 → strong positive relation between height and weight.

🧰10. Probability & Distribution Summary

Distribution	Type	Use
Normal	Continuous	Natural phenomena (height, weight)
Binomial	Discrete	Success/failure trials
Poisson	Discrete	Number of events per interval
Chi-Square	Continuous	Variance & categorical testing

🕗11. Quick Summary Table

Concept	Description
Descriptive Statistics	Summarize & visualize data
Inferential Statistics	Make predictions or decisions from data
Z-test	Compare sample mean to population (σ known)

Concept	Description
t-test	Compare means when σ unknown
Chi-Square Test	Check relationship between categorical variables
p-value	Probability of observed data under H_0
α, β	Type I & II errors
Power (1−β)	Test's ability to detect true effect
CI	Range estimate for population mean
Correlation	Strength of linear relation

🔗 Final Takeaway

- Use **Descriptive Statistics** to **summarize** what the data shows.
- Use **Inferential Statistics** to **test**, **predict**, or **validate** your insights.
- Always define hypotheses clearly, interpret p-values correctly, and visualize distributions for understanding.

End of Complete Statistics Notes