Basic Statistics

Pramod Duvvuri

February 26, 2019

These are the basics in Statistics one must be familiar with if they aspire to become a *Data Scientist*. This list will be including a mix of both Inferential and Descriptive Statistics. This list only covers *Parametric Methods*. This list was prepared from reading the free online book Online Statistics Education by **David Lane**. Learning all of these concepts theoretically is advisable before picking up a programming language to implement these concepts. One needs to know how a particular distribution looks like: Bernoulli, Binomial, Normal, Student-t. One could go study these distributions in more details if required as a pre-requisite for Machine Learning.

1. **Univariate Data**
2. Types of Sampling
   (a) Randomized Sampling
      i. Simple Random Sampling
      ii. Stratified Sampling
      iii. Cluster Sampling
   (b) Non-Random Sample (Biased)
      i. Voluntary Sampling
      ii. Convenience Sampling
3. Bias from Sampling
   (a) Response bias
   (b) Undercoverage
   (c) Convenience Bias
   (d) Non-response Bias
   (e) Voluntary Response Bias
4. Types of Variables/Data
   (a) Qualitative (Categorical)
      i. Nominal (No Order)
      ii. Ordinal (Order Matters)
   (b) Quantitative (Numerical)
      i. Continuos (Floating)
      ii. Discrete (Integer)
   (c) Interval
   (d) Ratio
5. Quantiles
   Definition: The lines which divide data into equally sized groups
   (a) Median
   (b) q_1, q_2, q_3 (Quartiles)
   (c) Inter-Quartile Range (IQR)

6. Percentiles
   Definition: The quantiles which divide data into 100 equally sized groups
7. Frequency Distribution
   (a) Frequency Table
   (b) Dot Plot (1-Dimensional)
   (c) Histogram (Number of bins/buckets)
   (d) Range (Maximum - Minimum)
8. Statistical Distribution (Histogram/Curve)
9. **Normal Distribution** (Gaussian Distribution): We need to know at least two of three parameters below to estimate/draw the curve.
   (a) Mean (μ)
   (b) Variance (σ^2)
   (c) Standard Deviation (σ)
   (d) Z-Score Calculation: Measure of how many sd's away each datapoint is from the mean (μ)
   Formula: $Z = \frac{X-\mu}{\sigma}$
   (e) Standard Normal Distribution: A normal distribution with mean (μ) equal to 0 and standard deviation (σ) equal to 1 is called a standard normal.
10. Skewed Distributions (Shape of the Curve)
   (a) Left Skewed (Negative Skew) : Longer Tail or thicker tail on the left side and (Mean $<$ Median)
   (b) Right Skewed (Positive Skew) : Longer Tail or thicker tail on the right side (Mean $>$ Median)
   (c) Bi-Modal (Two Peaks) : Two peaks in the curve
11. Sampling a Distribution
12. Data Transformations
   (a) Linear Transformation
   (b) Logarithmic Transformation
13. Plots:
   (a) Box-Whisker Plot
   (b) Bar Charts
   (c) Line Graphs
14. Mean(s):
   (a) Arithmetic Mean (Standard Mean)
   (b) Geometric Mean
   (c) Harmonic Mean
   (d) Tri-Mean
   (e) Trimmed Mean (Mean after removing X% of data on both sides of the curve)
15. Variability Measures:
   (a) Index of Skew : $\frac{3*(Mean-Median)}{\sigma}$ (Pearson's Formula)
   (b) Kurtosis
16. QQ-Plots

17. QQ-Line (R Only)
18. Contour Plot (2D)
19. Uniform Distribution
20. **Central Limit Theorem (CLT)** (Simulation helps you better understand this concept)
21. Population vs Sample:
    (a) Point Estimate
    (b) Sample Proportion ($\bar{p}$)
    (c) Mean ($\bar{x}$)
    (d) Variance (Sample Variance = ($\frac{\sigma^2}{n}$))
    (e) Standard Deviation (s)
    (f) *Standard Error*
22. Theoretical vs Empirical Distribution
23. Degrees of Freedom (DF)
24. **Confidence Intervals (CI)**
    (a) Upper Bound
    (b) Lower Bound
    (c) 95% CI
    (d) 99% CI (Wider than the 95% CI)
    (e) Margin of Error (2 * Std Error)
25. **t-distribution (student)** (Normal Distribution with df $\to \infty$)
    (a) t-statistic (score):
    Formula: $T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$
    (b) It has a lower peak and heavier tails implying more variance than the normal distribution and area of ($> 5\%$) in the tails combined
    (c) As df $\to \infty$ the peak increases and tends toward the normal curve but the area in the tails is more than the normal curve
26. **Hypothesis Testing** (Significance Testing)
    (a) Assumptions
        i. Check Normality Assumption with qq-plot, approximately normal data is allowed but if the data is heavily skewed we cannot accept the null hypothesis (H_0)
        ii. Box-Plot to check means
    (b) Null Hypothesis (H_0)
    (c) Alternate Hypothesis ($H_1|H_a$)
    (d) Test Statistic (Z/T)
    (e) p-value
    (f) alpha (α) (Significance Level)
    (g) Rejection Region (Tails)
    (h) 1-tail test
    (i) 2-tail test
    (j) Type-I (False Positive) and Type-II (False Negative) Errors

(k) Power

i. $Power = (1 - \beta)$ (β = Probability of Type-II Error)

ii. $(\alpha + \beta = 1)$

(l) Rough Guidelines:

i. $p < 0.01$ (Very Strong evidence against H_0)

ii. $0.01 < p \leq 0.05$ (Strong evidence against H_0)

iii. $p > 0.05$ (Weak evidence against H_0)

iv. $p > 0.1$ (Very Weak evidence against H_0)

27. **Bi-Variate Data**

(a) Population (ρ)

(b) Sample (r)

(c) Fisher's Z Transform (z')
Formula: $z' = 0.5 * ln(\frac{1+r}{1-r})$
Std Error = $\frac{1}{\sqrt{N-3}}$

28. **Hypothesis Testing (2-Sample/Population):**

(a) Assumptions

(b) Types of Hypothesis Tesing:

i. Independent Sample t-test

ii. Matched Sample t-test

(c) t-test or Welch's t-test (Welch is more robust)

(d) Test Statistic Calculation

29. **Trivariate/Multi-variate Data**

30. **ANOVA**

(a) Assumptions

(b) F-distribution

(c) F-Statistic ($F = \frac{SSB}{SSW}$)

(d) ANOVA table

31. One-way ANOVA

(a) One dependent variable

(b) One independent variable

32. Factorial ANOVA (Two-way ANOVA)

(a) One dependent variable

(b) One or more independent variable

33. Effects of unequal samples

34. Goodness of Fit

(a) Chi-Squared Test

i. Likelihood Ratio Test (G-Test)

ii. Pearson's Chi-squared Test

(b) Test Statistic

(c) χ^2 Distribution

35. Association
    (a) Scatter Plot
    (b) Correlation (Here is a **fun** game to test your understanding of this concept)
    (c) Correlation Test
36. **Linear Regression**
    (a) Assumptions
    (b) Simple Regression
        i. Slope
        ii. Intercept
        iii. Regression Line
        iv. OLS
    (c) Multiple Regression

Before learning each method one must know the assumptions that are made. Most of the methods listed above are robust and can perform reasonably well on data that violate some of these assumption. However, the violation of these said assumptions can lead to poor performance and questionable results. The data in most scenarios can be approximately normal but if it is heavily skewed it is best to consider transforming this data. If transforming data is not helpful then it might be helpful to know some *Non-Parametric Methods* which can then be used to test and make inferences.