Probability and Statistics: Data Science and ML Refresher

Shubhankar Agrawal

Abstract

This document serves as a quick refresher for Data Science and Machine Learning interviews. It covers mathematical concepts across a range of Probability and Statistical topics. This requires the reader to have a foundational level knowledge with tertiary education in the field. This PDF contains material for revision over key concepts that are tested in interviews.

■ Contents

1. Mathematics

1.1. Basic Formulae

Basic mathematical formulae to be known.

Series Progressions

$$S_n = a + (a + d) + ... + [a + (n-1)d] = \frac{n}{2}[2a + (n-1)d]$$
$$S_n = a + ar + + ... + ar^{n-1} = a\frac{1 - r^n}{1 - r} \tag{1}$$

Euler's Number (e) [2.718]

$$e^x = \sum \frac{x^k}{k!}$$
$$e = \left(1 + \frac{1}{n}\right)^n \tag{2}$$

Fibonacci Series

$$f(n) = f(n-1) + f(n-2) \tag{3}$$

Taylor Series

$$f(a) + \frac{f'(a)(x-a)}{1!} + \frac{f''(a)(x-a)^2}{2!} + ... + \frac{f'^{(n)}(a)(x-a)^n}{n!} + ... \tag{4}$$

1.2. Combinatorics

Basic formulae to generate permutations and combinations.

Permutations (Arrange n) : $n!$

$$\text{Combinations (Choose r from n)} : \binom{n}{r} = \frac{\prod n..n - r}{r!} \tag{5}$$

Coupon Collector

Number of draws (n) to get k items when each draw is uniform:

$$n = k * \left(\frac{1}{1} + \frac{1}{2} + ... + \frac{1}{k}\right) \tag{6}$$

Circular Seating

Number of people: $n - 1$ since one position is fixed.

Stars and Bars

Partitioning a set of n items into k boxes is given by $\binom{n+k-1}{k-1}$

1.3. Linear Algebra

Eigenvalues and Eigenvectors are given by calculating determinant of this equation.

$$A \cdot x = \lambda \cdot x \tag{7}$$

$$A^{-1} = \frac{1}{\text{mod } A} \cdot adj(A) \tag{8}$$

Determinant non-zero => Non-singular and invertible.

2. Probability

2.1. Basic Concepts

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A, B) = P(A/B) \cdot P(B)$$
$$P(A) = \sum_B P(A, B) \tag{9}$$

Here $P(A \cap B) = P(A, B)$

A and B are independent

$$P(A \cap B) = P(A) \cdot P(B)$$
$$P(A/B) = P(A) \tag{10}$$

Independence $\neq$ Mutual Exclusivity. Independence means they don't depend on each other. Mutually exclusive would mean they cannot occur together.

Bayes Theorem

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{\sum_A P(B/A) \cdot P(A)}$$
$$\text{where } P(A/B) = Posterior$$
$$P(B/A) = Likelihood$$
$$P(A) = Prior$$
$$P(B) = Evidence \tag{11}$$

2.2. Random Variables

PMF: Probability Mass Function
PDF: Probability Density Function
CDF: Cumulative Density Function

PMF and CDF sum to 1 over the entire range, formulae in Table 1

Table 1. Random Variables

Variable	Discrete	Continuous
Point	PMF: $P(X = x)$	PDF: $f(x)$
Cumulative	Sum of PMF: $\sum P(X)$	CDF: $\int f(x)$

2.3. Moments and Functions

Moment-generating functions across random variables form the basis of distributions and their capabilities.

Table 2. Moments

Moment	Discrete	Continuous
$E[X]$	$\sum X \cdot P(X)$	$\int x \cdot f(x)$
$E[X^2]$	$\sum X^2 \cdot P(X)$	$\int x^2 \cdot f(x)$

Mean: The average value.
Variance/Covariance: Strength of variation with itself/another.
Skewness: Left (Negative) skewed right (Positive) distribution weights.
Kurtosis: Normal (Mesokurtic) vs Negative (Platykurtic) is flatter vs Positive (Leptokurtic) higher peak.

$$
\begin{aligned}
Mean(\mu) &= E[X] \\
Variance(\sigma^2) &= E[X^2] - (E[X])^2 \\
StandardDeviation(\sigma) &= \sqrt{\sigma} \\
Skewness &= E[X^3] \\
Kurtosis &= E[X^4] - 3
\end{aligned}
\tag{12}
$$

Correlation: Measures both direction and strength of variation (between -1 and 1)

$$
\begin{aligned}
Covariance &= E[X, Y] - E[X] \cdot E[Y] \\
Correlation(\rho) &= \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y} \\
(Continuous) &= \frac{\sum_i (x - \mu_x) \cdot (y - \mu_y)}{\sigma_x \cdot \sigma_y}
\end{aligned}
\tag{13}
$$

2.4. Distributions

The most common continuous and discrete distributions are listed in Table 8. Apart from these, some CDFs of continuous distributions are listed in Table 3.

2.4.1. Estimation

Parameter estimation can be done with Maximum Likelihood Estimation (MLE) or Maximum A Posteriori (MAP) algorithms.

$$
\begin{aligned}
\ell(\theta) &= \log \prod_{i=1}^{n} P(x_i \mid \theta) = \sum_{i=1}^{n} \log P(x_i \mid \theta) \\
\hat{\theta}_{\text{MLE}} &= \arg\max_{\theta} \ell(\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log P(x_i \mid \theta)
\end{aligned}
\tag{14}
$$

MAP incorporates the prior distribution as well..

$$
\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \log P(\theta \mid \mathbf{x}) = \arg\max_{\theta} \left[\log P(\mathbf{x} \mid \theta) + \log P(\theta)\right]
$$

$$
\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \left[\sum_{i=1}^{n} \log P(x_i \mid \theta) + \log P(\theta)\right]
\tag{15}
$$

2.4.2. Functions

Table 3. Continuous Distribution CDFs

Distribution	PDF	CDF
Uniform	$\dfrac{1}{b-a}$	$\dfrac{x-a}{b-a}$
Normal	$\dfrac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	$\Phi\left(\dfrac{x-\mu}{\sigma}\right)$
Exponential	$\lambda \cdot e^{-\lambda x}$	$1 - e^{-\lambda x}$
Gamma	$\dfrac{1}{\Gamma(a)}(\lambda x)^a e^{-\lambda x} \dfrac{1}{x}$	$\int_0^x \dfrac{\lambda^k t^{k-1} e^{-\lambda t}}{\Gamma(k)} \, dt$

It is also helpful to know how to derive the PMF, PDF and CDF of the relevant probability distributions.

3. Statistics

3.1. Concepts

Central Tendency Functions: Mean, Median, Mode, Quartiles.

$$
Mode = 3 \cdot Median - 2 \cdot Mean
\tag{16}
$$

Law of Large Numbers: Mean value converges over trials.
Central Limit Theorem: Distribution converges to Normal

Chebyshev's Inequality

$$
P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}
\tag{17}
$$

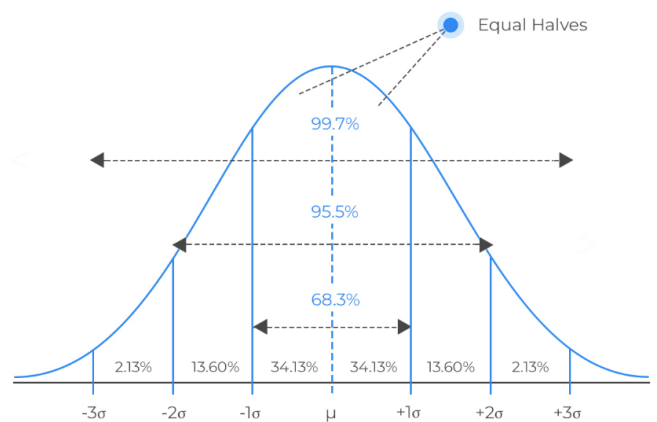**Normal Distribution** Key concept of statistics.

Figure 1. Normal Distribution
[3]

3.2. Hypothesis Testing

Testing variables to analyse performance
Population: The entire group

Sample: A subset of evaluation

$$\text{Population } \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$\text{Sample } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{18}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard Error (SE): The SE exists only for the sample since it is a subset. The standard deviation is divided by $\sqrt{n}$ to penalize for the smaller sample size. As the sample size increases, the SE vanishes.

$$\text{(One Sample) } SE = \frac{s}{\sqrt{n}}$$

$$\text{(Two Sample Eq Var) } SE = \sqrt{s^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{19}$$

$$\text{(Two Sample Uneq Var) } SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Types of Hypothesis:

Null H_0: The variable is as expected (values are equal).

Alternate H_1: The variable is strange (values are high/low).

Table 4. T Test or Z Test

Test	σ^2 **Variance unknown**	σ^2 **Variance Known**
$n < 30$	T - Test	T - Test
$n > 30$	T - Test	Z - Test

Test Statistic

$$\text{z (or) t} = \frac{x - \mu_0}{\sigma/\sqrt{n}} \tag{20}$$

3.2.1. Terminology

Important terms used in Hypothesis Testing

p-value: Under the null hypothesis, probability a seeing a value more extreme than test statistic.

Significance Level (α): Threshold at which test is conducted.

Critical Value: The point on the distribution corresponding to the significance level.

Power (β): Probability of accepting null hypothesis when it is true.

Confidence Interval: Estimated range of values containing the population mean.

Minimum Detectable Effect: Smallest difference that can be detected.

Cohen's D: Effect size related to standard deviation

Based on the statistic formula, the variables can be estimated with equations 21.

$$\text{Confidence Interval} = \mu \pm t_{\alpha/2,n} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\text{Minimum detectable effect (MDE)} = t_{\alpha/2,n} \cdot \frac{\sigma}{\sqrt{n}} \tag{21}$$

$$\text{Cohen's D} = \frac{\text{MDE}}{\text{Pooled SE}}$$

There are several types of T-tests.

Sample Greater/Lesser Than: One Tail [Significance $= \alpha$]

Sample Not Equal To: Two Tail [Significance $= \alpha/2$]

Table 5. Types of T Test

Sample	DF	Notes
1	$n - 1$	
2 (Paired)	$n - 1$	
2 (Unpaired)	$n_1 + n_2 - 2$	Same Population: Pool variance

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{22}$$

Pooled variance is given by Equation 22. Use the appropriate standard error formula based on the samples.

NOTE: T test uses degrees of freedom (DF), the Z test does not.

3.2.2. Test of Proportions

A similar formula as the T test when comparing proportions is used as in Equation 23. This can be used while conducting binomial trials across samples.

$$\text{z (or) t} = \frac{p_2 - p_1}{SE}$$

$$\text{Independent SE} = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}$$

$$\text{Pooled SE} = \frac{p(1 - p)}{n_1 + n_2} \tag{23}$$

$$\text{where } p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$$

Super Helpful X and T Test Table (for common values): T Test Table [5]

3.2.3. Errors

The different types of outcomes for a test are summarized in Table 6.

Table 6. Hypothesis Testing Outcomes

Real / Pred	Accept H_0	Reject H_0
$H_0 = $ **True**	Confidence Level $(1 - \alpha)$	**Type I** Error Significance Level (α)
$H_0 = $ **False**	**Type II** Error Fail to Reject (β)	Power $(1 - \beta)$

3.2.4. Power Analysis

Power Analysis is used to calculate the sample size needed to observe the Minimum detectable effect, as seen in Figure 2.
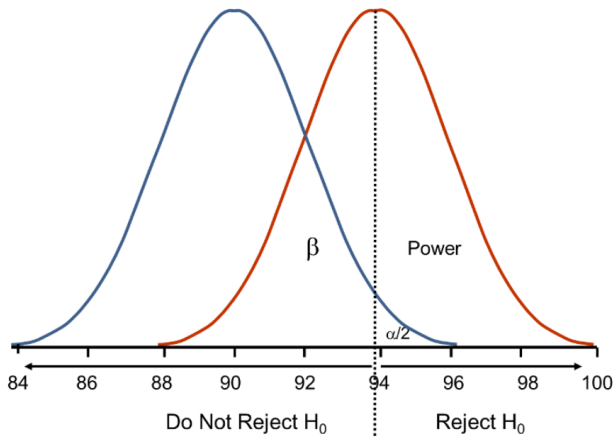
Figure 2. Power over Distributions
[4]

Deriving from the formula for minimum detectable effect with a small tweak, the minimum sample size needed is calculated. Usually a power of 0.8 is used in hypothesis testing.

$$\text{Sample Size} \, n = \left(\frac{Z_{\alpha/2} + Z_\beta}{\delta/\sigma} \right)^2 \tag{24}$$

where δ = Minimum Detectable Effect

NOTE: Power Analysis uses the Z test to calculate the sample size. If two independent samples are sized, the size calculated is doubled.

3.3. Analysis of Variance

ANOVA is used to compare two variables and if the means of their are statistically different from each other. This is used for comparing the groups within a continuous variable when values can be assumed to follow a normal distribution. These steps are needed to calculate the ANOVA statistic.

$$
\begin{aligned}
\text{Sum Squares Between Groups} \ SSB &= \sum_{j=1}^{k} n_j (\bar{Y}_j - \bar{Y})^2 \\
\text{Sum Squares Within Groups} \ SSW &= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ji} - \bar{Y}_j)^2 \\
\text{Mean Squares Between} \ MSB &= \frac{SSB}{k-1} \\
\text{Mean Squares Within} \ MSW &= \frac{SSW}{N-k} \\
\text{F Statistic} &= \frac{MSB}{MSW}
\end{aligned} \tag{25}
$$

Useful link for F Test: F Test Table [2]

3.4. Non Parametric Methods

3.4.1. Chi Square

Chi Square is used to evaluate on categorical variables. It is a special case of the Gamma distribution. It requires building the frequency table.

Goodness of Fit: Check if a sample fits a population.

$$chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$
$$\text{where } df = k - 1 \tag{26}$$
$$O = \text{Observed Frequency}$$
$$E = \text{Expected Frequency}$$

Independence: Check if two categorical variables are independent.

$$chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
$$df = (r-1) \times (c-1) \tag{27}$$
$$\text{r, c} = \text{Number of rows and columns}$$

Variance: Hypothesis Test on Sample Variance

$$chi^2 = \frac{(n-1)s^2}{\sigma^2}$$
$$df = n - 1 \tag{28}$$

Useful table for Chi Square: Chi Square Table [1]

3.4.2. Mann Whitney U Test

Also known as Wilcoxon Rank Sum test to determine significant difference between two ordinal samples. All samples from both groups need to be ranked together for this. It uses the Z score table for evaluation.

$$
\begin{aligned}
U_1 &= \sum_{\text{Group 1}} \text{rank} - \frac{n_1(n_1 + 1)}{2} \\
U_2 &= \sum_{\text{Group 2}} \text{rank} - \frac{n_2(n_2 + 1)}{2} \\
U &= \min(U_1, U_2) \\
\mu_U &= \frac{n_1 n_2}{2} \\
\sigma_U^2 &= \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \\
Z &= \frac{U - \mu_U}{\sqrt{\sigma_U^2}}
\end{aligned} \tag{29}
$$

A summary of all tests is provided in Table 9.

3.5. Stochastic and Temporal Models

3.5.1. Markov Chains

Markov Chain problems can be solved by building the transition matrix P.

To calculate the probabilities to get to a certain end state, extract two matrices **Q**: sub-matrix of transient states (no absorbing state) and **R**: sub-matrix of transient states to absorbing state.

$$\text{Fundamental Matrix } N = (I - Q)^{-1}$$
$$\text{Absorbing Probability } B = N \cdot R \tag{30}$$

Irreducible: No state is unreachable **Aperiodic**: No periodic self loops **Ergodic**: Irreducible (and) Aperiodic. In this case, the steady state equations can be solved.

$$\pi P = \pi$$

$$\sum_{i=1}^{n} \pi_i = 1 \qquad (31)$$

3.5.2. ARIMA

Family of models for time series forecasting.
Stationarity: No changing components (trends)
Heteroskedasticity: Non constant variance

Table 7. SARIMAX Components

Component	Variable	Notes
Auto-Regressive	p	Partial Autocorrelation (PACF)
Differencing	d	Augmented Dickey Fuller (ADF)
Moving Average	q	Autocorrelation (ACF)
Seasonality	s	Seasonal Decomposition

4. Contact me

You can contact me through these methods:

🔗 Personal Website - astronights.github.io
✉ shubhankar.31@gmail.com
in linkedin.com/in/shubhankar-agrawal

■ References

[1] *Chi Square Table*. [Online]. Available: https://math.arizona.edu/~jwatkins/chi-square-table.pdf.

[2] *F Test Table*. [Online]. Available: https://www.stat.purdue.edu/~lfindsen/stat503/F_alpha_05.pdf.

[3] *Key Properties of the Normal Distribution*. [Online]. Available: https://analystprep.com/cfa-level-1-exam/wp-content/uploads/2019/10/page-123.jpg.

[4] *Power and Sample Size Determination*. [Online]. Available: https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_power/bs704_power_print.html.

[5] *T Test Table*. [Online]. Available: https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf.

Table 8. Common Probability Distributions

Type	Name	PMF/PDF	Mean	Variance	Summary
Discrete	Uniform	$\dfrac{1}{n}$	$\dfrac{n+1}{2}$	$\dfrac{n^2-1}{12}$	Single uniform outcome
	Bernoulli	$p^{\lambda} \cdot (1-p)^{(1-\lambda)}$	p	$p \cdot (1-p)$	Single binary outcome
	Binomial	$\binom{n}{k} \cdot p^k \cdot (1-p)^{(n-k)}$	$n \cdot p$	$n \cdot p \cdot (1-p)$	Pick k from n (with replacement)
	Geometric	$(1-p)^{(n-1)} \cdot p$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$	# of events till first success
	Negative Binomial	$\binom{n+r-1}{r-1} \cdot (1-p)^n \cdot p^r$	$\dfrac{r}{p}$	$\dfrac{r \cdot (1-p)}{p^2}$	# of events till r successes
	Hyper-geometric	$\dfrac{\binom{K}{k} \cdot \binom{N-n}{K-k}}{\binom{N}{n}}$	$\dfrac{n \cdot K}{N}$	$n \cdot \dfrac{K}{N} \cdot \dfrac{N-K}{N} \cdot \dfrac{N-n}{N-1}$	Pick k from n picks of K within N items (no replacement)
	Poisson	$\dfrac{e^{-\lambda} \cdot \lambda^x}{x!}$	λ	λ	# of events in λ time
Continuous	Uniform	$\dfrac{1}{b-a}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$	Single uniform outcome
	Normal	$\dfrac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	Gaussian Event
	Exponential	$\lambda \cdot e^{-\lambda x}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$	Time to witness x events (Inverse of Poisson)
	Gamma	$\dfrac{1}{\Gamma(a)}(\lambda x)^a e^{-\lambda x}\dfrac{1}{x}$	$\dfrac{a}{\lambda}$	$\dfrac{a}{\lambda^2}$	Gamma Event (a > 0) $\Gamma(a) = (a-1)!$

Table 9. Statistical Tests

Test	Data	Comparing	Statistic	Notes
Z Test	Numerical	Means	Z	Known σ^2
T Test	Numerical	Means	T	Unknown σ^2
ANOVA	Numerical	Means	F	Groups
Chi Square	Categorical	Frequency / Variance	Chi2	Requires Building Frequency Table
Mann Whitney U	Ordinal	Medians	Z	Non normal data, Requires Ranking