Progress Report on App Development (Week 03)

-by Digvijay Jadhav

An Intern in UpSkill Campus, 28/01/2026

Digvijay Jadhav

I am pleased to present you with a comprehensive report on data science and machine learning (Week 03), which provides an overview of the process, challenges, and best practices for successful data science and machine learning. This report aims to make understanding of the key aspects of data science and machine learning and making informed decisions in this domain.

Introduction to Probability and Statistics

in Data Science & Machine Learning

Probability and Statistics form the mathematical backbone of **Data Science and Machine Learning**. They help us **understand data, deal with uncertainty, make predictions, and build intelligent models** from real-world information.

In real life, data is never perfect—it is noisy, incomplete, and uncertain. Probability allows us to **measure uncertainty**, while statistics helps us **analyze, summarize, and draw conclusions from data**.

---

1. Role of Probability

Probability deals with the likelihood of events occurring.

In Data Science & ML, probability is used to:

- Handle **uncertain outcomes**
- Model **random variables**
- Predict future events
- Build **probabilistic models**

Examples:

- Predicting whether an email is **spam or not**
- Estimating the probability that a customer will **buy a product**
- Weather prediction models

Common probability concepts:

- Random variables
- Probability distributions (Normal, Binomial, Poisson)
- Conditional probability
- Bayes' Theorem

---

2. Role of Statistics

Statistics focuses on collecting, organizing, analyzing, and interpreting data.

In Data Science & ML, statistics helps to:

- **Summarize data** using mean, median, variance
- **Understand data patterns**
- Test assumptions and hypotheses
- Validate machine learning models

Examples:

- Finding average sales per month
- Comparing performance of two ML models
- Detecting outliers in datasets

Types of Statistics:

- **Descriptive Statistics** – describes data
- **Inferential Statistics** – draws conclusions about a population

---

3. Why They Are Important in Machine Learning

Machine learning algorithms are built on probability and statistics:

ML Concept	Probability / Statistics Used
Linear Regression	Mean, variance, correlation
Logistic Regression	Probability & sigmoid function
Naive Bayes	Bayes' Theorem
Decision Trees	Entropy, information gain
Model Evaluation	Mean squared error, confidence intervals

---

4. Real-World Applications

- **Healthcare**: Disease prediction
- **Finance**: Risk analysis & fraud detection
- **E-commerce**: Recommendation systems
- **Cloud & Web Systems**: User behavior analysis
- **Machine Learning Models**: Training, testing, validation

Random Variables and Their Probability Distributions

in Data Science and Machine Learning

---

1. Random Variable (RV)

A **random variable** is a numerical value that represents the outcome of a **random experiment**.

□ Instead of dealing with outcomes directly, we assign **numbers** to them.

Examples:

- Tossing a coin →
  Head = 1, Tail = 0
- Number of clicks on a website in an hour
- Daily stock price change
- Marks obtained by students

In **Data Science & ML**, features like age, salary, number of purchases, and sensor readings are treated as **random variables**.

---

2. Types of Random Variables

1□ Discrete Random Variable

- Takes **countable values**
- Usually integers

Examples:

- Number of emails received per day
- Number of defective items
- Number of customers visiting a website

2□ Continuous Random Variable

- Takes **uncountable values**
- Can take any value in a range

Examples:

- Height, weight
- Time taken to load a webpage
- Temperature
- Stock prices

---

3. Probability Distribution

A **probability distribution** describes how probabilities are assigned to the possible values of a random variable.

It answers:

“How likely is each outcome?”

---

4. Probability Distributions for Discrete RVs

□ 1. Bernoulli Distribution

- Only **two outcomes** (0 or 1)
- Used in **binary classification**

Examples:

- Spam (1) or Not Spam (0)
- Pass or Fail

Used in ML: Logistic Regression

□ 2. Binomial Distribution

- Fixed number of trials
- Each trial has two outcomes

Example:

- Number of successful predictions out of 10 models

Used in DS: A/B testing, success–failure experiments

□ 3. Poisson Distribution

- Counts number of events in a fixed interval

Example:

- Number of server requests per minute
- Number of calls in a call center

Used in DS: Traffic analysis, cloud monitoring

---

5. Probability Distributions for Continuous RVs

□ 1. Normal (Gaussian) Distribution

- Bell-shaped curve
- Defined by **mean (μ)** and **standard deviation (σ)**

Examples:

- Exam scores
- Human height
- Measurement errors

Used in ML:

- Feature scaling
- Noise modeling
- Many algorithms assume normality

▯ 2. Uniform Distribution

- All values have equal probability

Example:

- Random number generation
- Sampling techniques

▯ 3. Exponential Distribution

- Time between events

Example:

- Time between customer arrivals
- System failure time

Used in DS: Reliability analysis

Moments and Moment Generating Function

in Data Science and Machine Learning

1. Moments

Moments are numerical measures that describe the **shape and characteristics** of a probability distribution.

In Data Science, moments help us understand:

- Central tendency
- Spread of data
- Skewness
- Kurtosis

2. Types of Moments

▯ 1. Raw Moments (Moments about Origin)

The **r-th raw moment** is defined as:

μr′=E[Xr]\mu'_r = E[X^r]μr′=E[Xr]

Examples:

- 1st raw moment → Mean

- 2nd raw moment → Used in variance calculation

□ 2. Central Moments (Moments about Mean)

The **r-th central moment** is:

μr=E[(X−μ)r]\mu_r = E[(X - \mu)^r]μr=E[(X−μ)r]

where μ=E[X]\mu = E[X]μ=E[X].

. Importance of Moments in Data Science & ML

Moments are used to:

- **Summarize datasets**
- **Understand data distribution**
- **Detect skewness and outliers**
- **Assume distributions in ML models**
- **Feature engineering**

---

4. Moment Generating Function (MGF)

The **Moment Generating Function** of a random variable XXX is defined as:

MX(t)=E[etX]M_X(t) = E[e^{tX}]MX(t)=E[etX]

It is called “moment generating” because **all moments can be obtained by differentiating it**.

---

5. Extracting Moments from MGF

n-th moment=dnMX(t)dtn∣t=0\text{n-th moment} = \left.\frac{d^n M_X(t)}{dt^n}\right|_{t=0}n-th moment=dtndnMX(t)t=0

Examples:

- First derivative → Mean
- Second derivative → Variance

This makes MGF a **powerful mathematical tool**.

---

6. Why MGF Is Useful

□ Key advantages:

1. Helps in **finding moments easily**
2. Used to **identify distributions**

3. Simplifies proofs in ML theory
4. Helps analyze **sum of independent random variables**
5. Useful in probabilistic modeling

7. Common MGFs of Distributions

Distribution	**MGF**
Bernoulli	1−p+pet1 - p + pe^t1−p+pet
Binomial	(1−p+pet)n(1 - p + pe^t)^n(1−p+pet)n
Poisson	eλ(et−1)e^{\lambda (e^t - 1)}eλ(et−1)
Normal	eμt+12σ2t2e^{\mu t + \frac{1}{2}\sigma^2 t^2}eμt+21σ2t2

Used heavily in **probabilistic ML models**.

8. Role in Machine Learning

Moments and MGF are used in:

- **Naive Bayes classifier**
- **Bayesian inference**
- **Gaussian assumptions in regression**
- **Loss function analysis**
- **Uncertainty estimation**

Multiple Random Variables

in Data Science and Machine Learning

1. What Are Multiple Random Variables?

When a system involves **more than one random outcome**, we use **multiple random variables**.

Let:

- XXX, YYY, ZZZ be random variables defined on the same experiment.

Examples:

- XXX = height of a person, YYY = weight of the same person
- XXX = number of website visits, YYY = number of purchases
- XXX = feature value, YYY = label in ML dataset

In ML, **datasets naturally contain multiple random variables (features)**.

2. Joint Probability Distribution

The **joint probability distribution** describes the probability of two or more random variables **occurring together**.

Discrete Case:

P(X=x,Y=y)P(X = x, Y = y)P(X=x,Y=y)

Continuous Case:

fX,Y(x,y)f_{X,Y}(x,y)fX,Y(x,y)

Used to model **relationships between features**.

---

3. Marginal Probability Distribution

The **marginal distribution** of one variable is obtained by **summing or integrating** over the other variable.

Discrete:

P(X=x)=∑yP(X=x,Y=y)P(X=x) = \sum_y P(X=x, Y=y)P(X=x)=y∑P(X=x,Y=y)

Continuous:

fX(x)=∫fX,Y(x,y) dyf_X(x) = \int f_{X,Y}(x,y)\,dyfX(x)=∫fX,Y(x,y)dy

---

4. Conditional Probability Distribution

The **conditional distribution** tells us the probability of one variable **given another**.

P(X=x∣Y=y)P(X=x \mid Y=y)P(X=x∣Y=y)

In ML:

- Core idea behind **Naive Bayes**
- Used in **Bayesian networks**

---

5. Independence of Random Variables

Two variables XXX and YYY are **independent** if:

P(X,Y)=P(X)P(Y)P(X,Y) = P(X)P(Y)P(X,Y)=P(X)P(Y)

ML Insight:

- Naive Bayes assumes **conditional independence**
- Feature dependence can affect model accuracy

---

6. Expectation with Multiple Random Variables

Expected Value:

E[X+Y]=E[X]+E[Y]E[X+Y] = E[X] + E[Y]E[X+Y]=E[X]+E[Y]

Covariance:

Cov(X,Y)=E[(X−μX)(Y−μY)]\text{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]Cov(X,Y)=E[(X−μX)(Y−μY)]

- Measures **linear relationship**
- Zero covariance → uncorrelated (not always independent)

---

7. Correlation

ρ=Cov(X,Y)σXσY\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}ρ=σXσYCov(X,Y)

Value	**Meaning**
+1	Strong positive relation
0	No linear relation
-1	Strong negative relation

Used in:

- Feature selection
- Multicollinearity detection

---

8. Joint Distributions in ML Algorithms

ML Concept	**Use of Multiple RVs**
Linear Regression	Joint distribution of features & errors
PCA	Covariance matrix
Gaussian Mixture Models	Multivariate normal
Bayesian Networks	Conditional dependencies
Hidden Markov Models	Joint & conditional probabilities

---

9. Multivariate Distributions

Common multivariate distributions:

- Multivariate Normal (Gaussian)
- Multinomial
- Dirichlet

Used in:

- Clustering
- Topic modeling
- Recommendation systems

---

10. Real-World Example

Customer behavior model:

- XXX = time spent on website
- YYY = number of clicks
- ZZZ = purchase amount

Analyzing joint and conditional distributions helps predict **purchase probability**.

Sample Statistics and Their Distributions

in Data Science and Machine Learning

---

1. Sample and Sample Statistics

A **sample** is a subset of data taken from a population.

A **sample statistic** is a numerical measure calculated from a sample and used to **estimate population parameters**.

Examples of Sample Statistics:

- Sample mean (X¯\bar{X}X¯)
- Sample variance (S2S^2S2)
- Sample proportion (p^\hat{p}p^)
- Sample median

In Data Science, models are trained on **samples**, not full populations.

---

2. Common Sample Statistics

⍰ 1. Sample Mean

X¯=1n∑i=1nXi\bar{X} = \frac{1}{n}\sum_{i=1}^n X_iX¯=n1i=1∑nXi

Used in:

- Feature scaling
- Model evaluation
- Regression

⍰ 2. Sample Variance

S2=1n−1∑i=1n(Xi−X¯)2S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2S2=n−11i=1∑n(Xi−X¯)2

Measures data **spread**.

□ 3. Sample Proportion

p^=Xn\hat{p} = \frac{X}{n}p^=nX

Used in:

- Classification problems
- A/B testing

---

3. Sampling Distribution

A **sampling distribution** is the probability distribution of a **sample statistic** obtained from all possible samples of size nnn.

□ It describes how a statistic **varies from sample to sample**.

---

4. Sampling Distribution of Sample Mean

According to the **Central Limit Theorem (CLT)**:

- The distribution of X¯\bar{X}X¯ approaches a **normal distribution** as nnn increases.
- Mean = population mean μ\muμ
- Variance = σ2n\frac{\sigma^2}{n}nσ2

Importance in ML:

- Error estimation
- Confidence intervals
- Model performance evaluation

---

5. Sampling Distribution of Sample Proportion

- Approximated by **normal distribution** for large samples
- Mean = ppp
- Variance = p(1−p)n\frac{p(1-p)}{n}np(1−p)

Used in:

- Binary classification accuracy
- Conversion rate analysis

---

6. Distribution of Sample Variance

The sample variance follows a **Chi-square distribution**:

(n−1)S2σ2∼χn−12\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}σ2(n−1)S2∼χn−12

Used in:

- Variance estimation
- Hypothesis testing

---

7. Student's t-Distribution

Used when:

- Sample size is small
- Population variance is unknown

Common in:

- Model evaluation
- Confidence interval estimation

---

8. F-Distribution

- Ratio of two sample variances
- Used in:
  - ANOVA
  - Feature comparison
  - Model selection

---

9. Importance in Data Science & ML

Concept	Application
Sample Mean	Feature normalization
Variance	Model regularization
Sampling Distribution	Performance stability
CLT	Approximation methods
t, χ^2, F	Hypothesis testing

---

10. Real-World Example

In **machine learning model evaluation**:

- Accuracy from test data is a **sample statistic**
- Its distribution helps estimate **true model performance**
- Confidence intervals assess **model reliability**

Basic Asymptotic (Large Sample) Theory

in Data Science and Machine Learning

---

1. Meaning of Asymptotic / Large Sample Theory

Asymptotic theory studies the behavior of **sample statistics and estimators** as the **sample size n→∞n \to \inftyn→∞**.

□ It answers:

- What happens to estimators when we have **very large data**?
- Do they become accurate and stable?

This is extremely important in **Big Data, Data Science, and ML**.

---

2. Why Large Sample Theory Is Important

In real-world ML:

- Datasets are **large**
- Exact distributions are often **unknown**
- Asymptotic results give **approximate solutions**

Used for:

- Parameter estimation
- Model consistency
- Confidence intervals
- Hypothesis testing

---

3. Law of Large Numbers (LLN)

Statement:

As sample size increases, the **sample mean** converges to the **population mean**.

X¯→pμ\bar{X} \xrightarrow{p} \muX¯pμ

Meaning:

- Large data → more reliable estimates

ML Example:

- Average loss converges to true expected loss

---

4. Central Limit Theorem (CLT)

Statement:

For large nnn, the sampling distribution of the sample mean is **approximately normal**, regardless of the population distribution.

n(X‾−μ)∼N(0,σ2)\sqrt{n}(\bar{X} - \mu) \sim N(0, \sigma^2)n(X‾−μ)∼N(0,σ2)

Importance:

- Enables normal approximation
- Used in confidence intervals & testing

ML Example:

- Model accuracy distribution
- Error estimation

---

5. Consistency of Estimators

An estimator θ^n\hat{\theta}_nθ^n is **consistent** if:

θ^n→pθ\hat{\theta}_n \xrightarrow{p} \thetaθ^npθ

Meaning:

- Estimator gets **closer to true value** as data grows

ML Example:

- MLE parameters in regression

---

6. Asymptotic Bias

Bias(θ^)=E(θ^)−θ\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \thetaBias(θ^)=E(θ^)−θ

- **Asymptotically unbiased** if bias → 0 as n→∞n \to \inftyn→∞

Used in:

- Model evaluation
- Regularization analysis

---

7. Asymptotic Variance

As sample size increases:

Var(θ^)→0\text{Var}(\hat{\theta}) \to 0Var(θ^)→0

Means:

- Estimates become **more precise**

---

8. Asymptotic Normality

An estimator is **asymptotically normal** if:

n(θ^−θ)→dN(0,V)\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)n(θ^−θ)dN(0,V)

Importance:

- Enables hypothesis testing
- Confidence interval construction

ML Example:

- Logistic regression coefficients

---

9. Maximum Likelihood Estimators (MLE)

Properties of MLE (as n→∞n \to \inftyn→∞):

- Consistent
- Asymptotically unbiased
- Asymptotically normal
- Efficient

Used in:

- Linear & logistic regression
- Gaussian models

---

10. Asymptotic Efficiency

An estimator is **efficient** if it has **minimum variance** among all unbiased estimators.

Related to:

- Cramér–Rao lower bound

Parametric Point Estimation

in Data Science and Machine Learning

---

1. Meaning of Parametric Point Estimation

Parametric point estimation is the process of **estimating an unknown population parameter** (such as mean, variance, or probability) **using a single numerical value** calculated from sample data.

- The population distribution is **assumed to be known** (normal, binomial, Poisson, etc.)
- The unknown quantity is a **parameter** of that distribution

Examples:

- Estimating population mean μ\muμ
- Estimating population variance σ2\sigma^2σ2
- Estimating probability ppp

---

2. Parameter vs Estimator vs Estimate

Term	Meaning
Parameter	True (unknown) population value
Estimator	Rule/formula to estimate parameter
Estimate	Numerical value obtained from data

Example:

- Parameter → μ\muμ
- Estimator → X¯\bar{X}X¯
- Estimate → sample mean value

---

3. Common Parametric Point Estimators

⍰ 1. Estimation of Mean

μ^=X¯\hat{\mu} = \bar{X}μ^=X¯

Used in:

- Regression models
- Feature scaling
- Error analysis

⍰ 2. Estimation of Variance

σ^2=S2=1n−1∑(Xi−X¯)2\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2σ^2=S2=n−11∑(Xi−X¯)2

□ 3. Estimation of Probability

p^=Xn\hat{p} = \frac{X}{n}p^=nX

Used in:

- Classification accuracy
- Conversion rate estimation

---

4. Methods of Parametric Point Estimation

□ 1. Method of Moments (MoM)

- Equates sample moments to population moments
- Simple and intuitive

Example:

E(X)=μ⇒Xˉ=μE(X) = \mu \Rightarrow \bar{X} = \muE(X)=μ⇒Xˉ=μ

□ 2. Maximum Likelihood Estimation (MLE)

- Chooses parameter values that **maximize likelihood**
- Most widely used in ML

L(θ)=P(X∣θ)L(\theta) = P(X \mid \theta)L(θ)=P(X∣θ)

Used in:

- Linear regression
- Logistic regression
- Naive Bayes

□ 3. Bayesian Estimation (MAP)

- Uses prior information
- Produces posterior estimates

Used in:

- Probabilistic ML
- Bayesian networks