# Contents

# Statistics For Data Science

Akash Tesla

July 2025

## 1 Basic Terminologies

### 1.1 Population

An entire set of items you want to study

### 1.2 Sample

A subset of population used to estimate statistical behavior of the whole population

### 1.3 Histogram

A histogram is a graphical representation of numerical data that groups the data into bins and displays the frequency of data points within each bin as bars

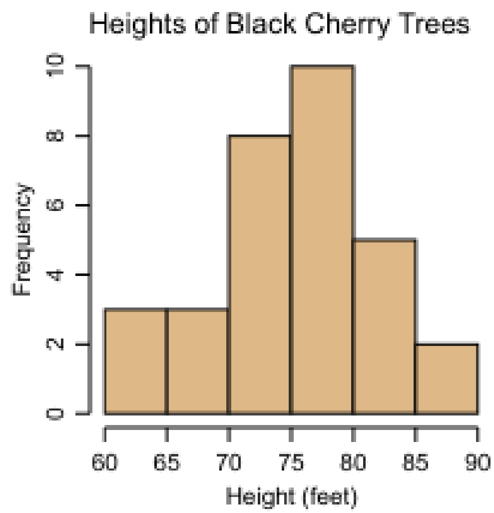


Figure 1: Example of a Histogram

# 2 Law of large numbers

As the number of trials (or samples) increases, the sample average (or empirical mean) will converge to the expected value (or population mean).

### 2.0.1 Weak Law of large numbers

The weak law states that the sample average of a sequence of independent identically distributed(i.i.d.) random variables converges in probability to the expected value as the number of samples goes to infinity

$$\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu \quad as\ n \to \infty$$

which means,

$$\forall \varepsilon > 0, \lim_{n\to\infty}^{n} \mathbf{p}(|\bar{X}_n - \mu| > \varepsilon) = 0$$

### 2.0.2 Strong Law of large numbers

The strong law states that the sample average of a sequence of i.i.d. random variables converges almost surely to the expected value as the number of samples goes to infinity

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu \quad as\ n \to \infty$$

Which means,

$$\mathrm{P}(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$$

# 3 Central-Limit Theorem

# 4 Measure of Central Tendency

## 4.1 Mean/Expected Value

Average of all data points, sensitive to outliers since a single large outlier could easily skew mean

$$\mu = \frac{\sum x_i}{n}$$

## 4.2 Median

The middle data point when data are stored, robust to outliers

## 4.3 Mode

The most frequent data point of the dataset

# 5 Measure of Spread

Range: Difference between minimum value and maximum value

$$Range = x_{max} - x_{min}$$

## 5.1 Variance

Average squared deviation, Variance represents Expected variance between mean and data points, It's basically MSE of a model that just predicts mean, that kinda gives an intutitive understanding of how it measures spread

$$\sigma^2 = E[(X-\mu)^2]$$
$$\sigma^2 = E[(X-E[X])^2]$$
$$\sigma^2 = E[X^2] - (E[X])^2$$
$$\sigma^2 = \frac{\sum(x_i-\mu)^2}{n}(Population)$$
$$s^2 = \frac{\sum(\bar{x_i}-\mu)^2}{n-1}(Sample)$$

### 5.2 Standard Deviation

Root of Variance, RMSE of a model that just predicts mean, standard deviation gives in intrepretable terms like RMSE

$$\sigma = \sqrt{\sigma^2}$$

### 5.3 Inter Quartile Range(IQR)

Difference between 75th Percentile/3rd Quartile and 25th Percentile/1st Quartile, it is used for outlier detection

$$IQR = Q_3 - Q_1$$

We calculate lower bounds and upper bounds to detect the outliers

$$\text{lower bound} = Q1 - 1.5 \times IQR$$

$$\text{upper bound} = Q3 + 1.5 \times IQR$$

the data points which values outside of the bounds is considered to be outliers, for more extreme detection $3 \times IQR$ is also used

## 6 Probability Distributions

### 6.1 Discrete Distributions

A discrete probability distribution describes the probability of occurrence of each value of a discrete random variable

- Discrete random variable: Countable values like 1,2,3
- Each individual value has an associated probability
- The sum of probabilities for all possible values is 1

$$\sum_i \mathrm{P}(X = x_i) = 1$$

## 6.2 Continuous Distributions

# 7 Discrete Probability Distributions

## 7.1 Benoulli Distribution

The benouli distribution is a discrete probability distribution for a random variable which takes only two possibilities, Sucess or a failure

### 7.1.1 Probability Mass Function(PMF)

$$P(X = x) = \begin{cases} p & \text{if x=1} \\ 1 - p & \text{if x=0} \\ 0 & \text{Otherwise} \end{cases}$$

Also written as

$$\mathrm{P}(X = x) = p^x (1 - p)^{1-x}, \quad \text{for x } \epsilon \ \{0, 1\}$$

### 7.1.2 Statistical Parameters

**Mean**
Mean is the expected value over many repetitions of the same single-trial experiment, thus it would be p since, p is probability of 1 appearing and (1-p) is probability of 0 appearing

$$\mu = 1 \times (p) + 0 \times (1 - p)$$

$$\mu = p$$

**Variance**
Variance can be defined as $\sigma^2 = E(X^2) - (E(x))^2$, Refer Variance chapter. For Bernoulli distribution, $E(X^2) = p, E(X) = p$, substituting we get

$$\sigma^2 = p - p^2$$

$$\sigma^2 = p(1-p)$$

**Mode**
Mode for Bernoulli would what ever the outcome which is more favored, which can be defined as

$$Mode = \begin{cases} 1 & \text{If } p > 0.5 \\ 0 & \text{If } p < 0.5 \end{cases}$$

### 7.1.3 Examples

- Will it rain tomorrow?
- Will this patient recover?
- Will this product be defective?

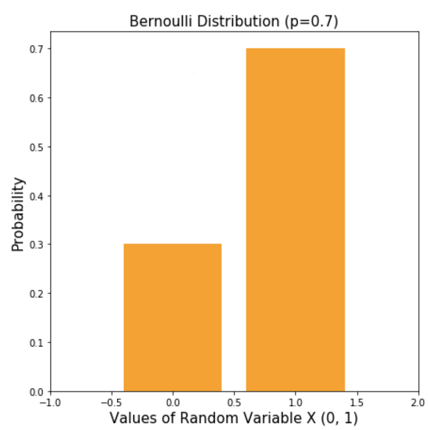


Figure 2: Example of a Bernoulli Distribution

## 7.2 Binomial

Binomial Distribution is a discrete probability distribution that models the probability of obtaining a specific number of successes in a fixed number of independent trials(n), these independent trials are just Bernoulli trials, you could see the similarity between them in statistical parameters

### 7.2.1 Probability Mass Function(PMF)

$$P(X = x) = nCx \times p^x \times (1 - p)^{(n-x)}$$

where,
n - no of trials,
p - probability of success
x - number of success

### 7.2.2 Statistical Parameters

**Mean**
Mean represents Average number of success from your trails which would be number of trials (n) times probability of success (p)

$$\mu = n \times p$$

**Variance**
Variance represents Expected variance between mean and data points,

$$\sigma^2 = n \times p \times (1 - p)$$

**Mode**

$$Mode = \begin{cases} \text{floor(n+1)p} & \text{if (n+1)p is not an Integer} \\ \text{floor((n+1)p), floor((n+1)(1-p))} & \text{if (n+1)p is an Integer} \end{cases}$$

$$\text{Mode(if p = 0.5)} = \begin{cases} \frac{n}{2} & \text{if (n+1)p is not an Integer} \\ \frac{(n-1)}{2}, \frac{(n+1)}{2} & \text{if (n+1)p is an Integer} \end{cases}$$

#### 7.2.3 Examples

- How many patients will recover out of 50?
- How many rainy days this month?
- How many defective products in a batch of 1000?

### 7.3 Negative-Binomial

### 7.4 Multinomial

### 7.5 Geometric

### 7.6 Hypergeometric

### 7.7 Poisson

### 7.8 Discrete Uniform

### 7.9 Normal Distribution

#### 7.9.1 Use cases

- When there is only one trial
- When the outcome is binary True/False Yes/No

## 8 Shape fo the Distributions

### 8.1 Skewness

Measure of Asymmetry

#### 8.1.1 Right-skewed

tail on the right ($mean > median$)

#### 8.1.2 Left-skewed

tail on the left ($mean < median$)

# 9 Hypothesis Testing

## 9.1 Null Hypothesis($H_0$)

Null Hypothesis is the default claim basically means no effect/ no differnce

## 9.2 Alternate Hypothesis($H_a$)

Alternate Hypothesis is the hypothesis that u want to prove

## 9.3 One sided Test

When you have to test if the parameter is greater than or less than the hypothesised value, but not both
Null:$H_0 : \mu = \mu_0$
Alternate:$H_a : \mu > \mu_0$ or $\mu < \mu_0$

## 9.4 Two sided Test

When you have to test if the parameter is different from the hypothesised value in either direction Null:$H_0 : \mu = \mu_0$
Alternate:$H_a : \mu \neq \mu_0$

## 9.5 Test Statistic

A test statistic is a function of the sample data that is used to decide whether to accept/reject the null hypothesis

$$\text{Test Statistic} = \frac{\text{How surprised we are}}{\text{How surprised we can be}}$$

$$\text{Test Statistic} = \frac{\text{Observed Value - Expected value under } H_0}{\text{Standard Error of observed value}}$$

## 9.6 Sampling distribution under $H_0$

If null hypothesis is true, what would the distribution of my test statistic look like across repeated samples. the sample mean/ test statistic follows normal distribution thanks to CLT (central limit theorem), we use that to calculate p-value.

1. Calculate sample statistic($\bar{x}$,$s^2$)
2. Compute test statisctic(t-test,z-test...)
3. Compare the computed test statistic to the corresponding distribution to get the p-value

### 9.7 p-value

The p-value is the probability of observing your data assuming the null hypothesis is true
if p is small($p<\alpha$), we reject null's hypothesis if p is high($p>\alpha$), we reject alternate hypothesis

## 10 Machine Learning

Machine learning(ML) is a way of teaching computers to learn patterns from data and make prediction.

### 10.1 Type of Machine learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi-Supervised Learning

## 11 Types of Data

### 11.1 Qualitative Data

Describes Qualities, Charesteristics , or categories

#### 11.1.1 Nominal

Pure categories without order, Example: blood type(A,B,AB,O), brand names

#### 11.1.2 Ordinal

Categories with meaningfull order, Examples: Rank, Survey rating

### 11.2 Quantitative Data

Measureable quantities, Number have meaningfull terms in terms of magnitude

#### 11.2.1 Discrete

Countable values, no in-betweens. Examples: number of cars

#### 11.2.2 Continuous

Countinous measurements; can take any value within a range, Examples: Height, weight, temperature

## 12 Data Clearning

## 13 Feature engineering and selection

## 14 Supervised Learning

## 15 Unsupervised Learning

## 16 Evaluation Metrics for Regression

### 16.1 Mean Absolute Error

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

- Robust to outliers, treats all errors equally doesn't square the errors like RMSE,MSE..etc
- It's used when your model can tolerate moderate outliers

- Interpretability - Has same unit as the thing you are predicting/easy to understand
- Gives out constant gradient (bad for gradient based loss function)

### 16.1.1 Gradient of MAE

$$\frac{d}{d\hat{y}}|y-\hat{y}| = \begin{cases} +1 & \text{if } \hat{y} < y \\ -1 & \text{if } \hat{y} > y \\ \text{undefined} & \text{if } \hat{y} = y \end{cases}$$

As you can see no matter how far the error is from true value it always gives a constant gradient as it treats every error as same stics

## 16.2 Mean Squared Error

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

- Penalizes large errors/outliers
- Gives out strong gradient signals

### 16.2.1 Gradient of MSE

$$\frac{dMSE}{d\hat{y}} = -\frac{2}{n}(y-\hat{y})$$

It points in the direction of the error, and it grows linearly with size of the error Larger the gradient, when prediction are more wrong $\longrightarrow$ model adjusts faster

## 16.3 Root Mean Squared Error(RMSE)

$$RMSE = \sqrt{MSE}$$

- It combines interpretability of MAE and sensitive to errors of MSE
- It has smooth gradient curves just like MSE, and it's preferred for gradient descent

### 16.3.1 Gradient of RMSE

$$\frac{dRMSE}{\hat{y}_i} = \frac{1}{n \times RMSE}(\hat{y}_i - y_i)$$

1. The gradient strength changes with RMSE, if your RMSE is very large the gradient becomes small, and if your RMSE is very small the gradient becomes large.

2. It makes RMSE a Non-constantly scaled loss

3. MSE is preferred over RMSE in training, but RMSE is preferred while reporting for interpretability

## 16.4 R-Square($R^2$)

$R^2$ is the coefficient of determination. it tells how well your regression model explains the variation in the dependent variable(Y) using independent variables(X)

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where,

- $SS_{res} = \sum(y_i - \hat{y}_i)^2 \rightarrow$ Residual sum of squares(error)/MSE

- $SS_{tot} = \sum(y_i - \bar{y}_i)^2 \rightarrow$ Total sum of squares (total variability)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

Or, it can also be written more intuitively as

$$R^2 = 1 - \frac{MSE}{\sigma^2}$$

- Let us understand the formula (1-) operator just switches from maximizing to minimizing so you can ignore that.

- $\frac{MSE}{\sigma^2}$ Explains how well our model performs to a model that just predicts mean everytime, so if the ratio is 1, then our model is same as the dumb model, we have to reduce the ratio but the world likes ”more the better” approach add (1-) operator we have to maximize the error and it’s called as $R^2$

- $R^2$ ranges from $(-\infty, 1]$

## 16.5 Adjusted $R^2$

$$R^2_{adj} = 1 - \left( \frac{(1 - R^2)(n-1)}{n-k-1} \right)$$

The above mentioned is textbook formula but we use our simplified representation for $R^2$

$$R^2 = 1 - \frac{MSE}{\sigma^2}$$

so, $R^2_{adj}$ would be

$$R^2_{adj} = 1 - \frac{MSE_{adj}}{\sigma^2_{adj}}$$

MSE adjusted accounts for the number of freedoms used up to predict the data, which is K,represents number of parameters like number of predictors, number of bias

$$MSE_{adj} = \frac{\sum (y_i - \hat{y}_i)^2}{n-k}$$

Variance adjusted for number of freedoms used up which is 1 (mean), thus it'd be n-1 insted of n

$$\sigma^2_{adj} = \frac{\sum (y_i - \mu)^2}{n-1}$$

Substituting we get,

$$R^2_{adj} = 1 - \left( \frac{MSE}{\sigma^2} \times \frac{n-1}{n-k} \right)$$

where

- n - number of samples/ training samples
- k - number of parameters

## 16.6 Mean Absolute Percentage Error(MAPE)

MAPE is a metric used to measure accuracy of a predictive model. It expresses the prediction error as the percentage of actual values

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

MAPE is just like MAE but it gives out the error in percentage thus it's easier to intrepret

### 16.7 Huber Loss

Huber loss is a robust loss function/evaluation metric that has both strengths of MAE and MSE

$$L_\delta = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{for } |y - \hat{y}| \leq \delta \\ \delta \cdot (|y - \hat{y}| - 1/2\delta) & \text{otherwise} \end{cases}$$

where,

- $\delta$ is a hyperparameter that controls the behavior between MSE and MAE behavior

#### 16.7.1 Gradient of Huberloss

$$\frac{\partial L}{\partial \hat{y}} = \begin{cases} -(y - \hat{y}) & \text{if } |y - \hat{y}| \leq \delta \\ -\delta \cdot \text{sign}(y - \hat{y}) & \text{otherwise} \end{cases}$$

example graph goes here

## 17 Evaluation Metrics for Classification

### 17.1 Basic Terminologies

#### 17.1.1 True Positive

Correctly predicted positive class

#### 17.1.2 False Positive

Falsely Predicted Positive class, actually negative

#### 17.1.3 True Negative

Correctly predicted negative class

### 17.1.4 False Negative

Falsely predicted negative class, actually positive

## 17.2 Accuracy

Out of all predictions how many are correct, ratio between correct predictions and total predictions is accuracy

$$Accuracy = \frac{\text{Correct predictions}}{\text{Total predictions}}$$

can also be written as,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

## 17.3 Precision

Out of all my positive class prediction how much did I get correctly, ratio between correct positive class prediction and total positive class predicted

$$\text{Precision} = \frac{\text{No of correct positive class predicted}}{\text{No of positive class predicted}}$$

can also be written as,

$$\text{Precision} = \frac{TP}{TP + FP}$$

Use when FP is costly, when you want every positive prediction to be trust worthy, measures reliability

## 17.4 Negative Predicted Value (NPV)

Out of all my negative predictions how much did I get correctly, ratio between

$$\text{NPV} = \frac{\text{No of correct negative class predicted}}{\text{No of negative class predicted}}$$

can also be written as,

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

Use when FN is costly, when you want every negative class to be trust worthy

## 17.5 Recall(TPR)

Out of all positive cases how many did I predict correctly, also known as True Positive Rate(TPR) ratio of correctly predicted positive class and total positive cases

$$\text{Recall} = \frac{\text{No of correctly predicted positive class}}{\text{No of actual positive class}}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Use when FN is costly/ wrongly classifying , did it recall everything, predict all positive cases

## 17.6 Specificity

Out of all negative cases how many did I predict correctly, Basically Recall for negative class

$$\text{Specificity} = \frac{\text{No of correctly predicted negative class}}{\text{No of actual negative class}}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Use when FP is costly, when predicting negative class is important

## 17.7 False Positive Rate(FPR)

Out of all negative cases how many did I fail to predict,

$$\text{FPR} = \frac{Noofwronglypredictednegativeclass}{NoofactualNegativeclass}$$

$$\text{FPR} = \frac{FP}{TN + FP}$$

$$\text{FPR} = 1 - \text{specificity}$$

## 17.8 F-score

It's a harmonic mean between precision and recall

$$F_\beta = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}}$$

#### 17.8.1 Derivation

$$\frac{1}{F_\beta} = \frac{w_p}{\text{Precision}} + \frac{w_r}{Recall}$$

We want harmonic ratio of precision and recall,

$$\frac{w_r}{w_p} = \beta^2$$

We want the weights to add up to one

$$w_r + w_p = 1$$

By solving we get

$$w_r = \frac{\beta^2}{1+\beta^2}$$

$$w_p = \frac{1}{1+\beta^2}$$

# 18 AUC-ROC curve

## 18.1 Receiver Operating Characteristic (ROC)

The ROC curve plots the True Positive Rate(TPR) vs False Positive Rate(FPR) at various threshold settings, it let's you decide which one is best for you based on your requirement

## 18.2 Area Under the Curve (AUC)

The AUC tells how much of the curve is under the line, usually compared with other models
Higher AUC = Better model performance

| Auc Score | Intrepretation |
|---|---|
| 0.5 | Random guessing |
| 0.7 - 0.8 | Acceptable |
| 0.8 - 0.9 | Excellent |
| >0.9 | Outstanding |

# 19 Regularization

Regularization adds a penalty for complexity to prevent overfitting, make model simpler

$$\text{Regularized Loss} = \text{Loss} + \text{Regularization parameter}$$

## 19.1 L1

L1 adds the sum of absolute value of coefficients to the loss function, it is used for feature selection since it encourages the optimizer to shirnk unwanted features to zero

$$\text{L1} = \lambda \sum |w_i|$$

## 19.2 L2

L2 adds teh sum of coefficient squares to the loss function this makes the gradient strength smooth, thus it won't reduce everything to zero, but towards zero, use it when you think all features contribute to ur model

$$\text{L2} = \lambda \sum w_i^2$$

## 19.3 Elastic net

It's a combination of L1 and L2 where you want to have both sparsity(l2) and stability(l2)

$$\text{Elastic Net} = \alpha \text{L1} + (1 - \alpha)\text{L2}$$

# 20 Time Series

# 21 TODOO

1. pca
2. complete the Distributions, some intro to probability
3. Time series, ML, other stuff