

# **Fundamentals of Machine Learning**

## **Appendix A - Descriptive Statistics and Data Visualization for Machine learning**

## 1 Descriptive Statistics

- Descriptive Statistics for Continuous Features
- Descriptive Statistics for Categorical Features
- Populations & Samples

## 2 Data Visualization

- Bar Plots
- Histograms
- Box Plots

## 3 Summary

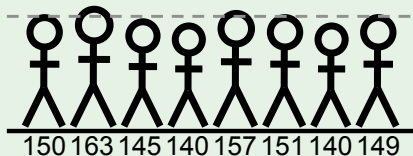
# Descriptive Statistics

- The **arithmetic mean** (or **sample mean** or just **mean**) of a set of  $n$  values for a feature  $a$ ,  $a_1, a_2 \dots a_n$ , is denoted by the symbol  $\bar{a}$ , and is calculated as:

$$\bar{a} = \frac{1}{n} \sum_{i=1}^n a_i$$

## Example

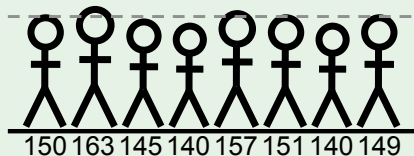
ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149



**Figure:** The members of a school basketball squad. The dashed grey line shows the arithmetic mean of the players' heights.

## Example

ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149



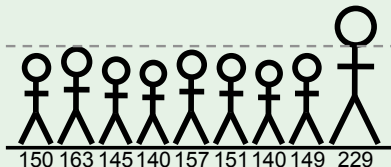
**Figure:** The members of a school basketball squad. The dashed grey line shows the arithmetic mean of the players' heights.

$$\begin{aligned}\overline{\text{HEIGHT}} &= \frac{1}{8} \times (150 + 163 + 145 + 140 + 157 + 151 + 140 + 149) \\ &= 149.375\end{aligned}$$

- The **arithmetic mean** is one measure of the **central tendency** of a **sample** (for our purposes a sample is just a set of values for a feature in an ABT).
- Any measure of **central tendency** is, however, just an approximation.

## Example

- Suppose our basketball squad manage to sign a *ringer* measuring in at 229cm

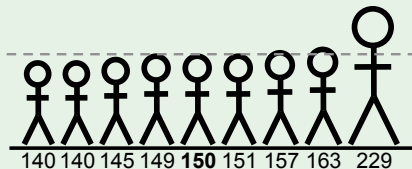


- The arithmetic mean for the full group is 158.235cm and no longer represents the central tendency of the group.
- An unusually large or small value like this is referred to as an **outlier** - the arithmetic mean is very sensitive to outliers.



- The **median** of a set of values can be calculated by ordering the values from lowest to highest and selecting the middle value.
- If there is an even number of values in the sample then the median is obtained by calculating the arithmetic mean of the middle two values.

## Example



**Figure:** The members of the school basketball squad ordered by height, the dashed grey line shows the **median**.

ID	4	7	3	8	1	6	5	2	9
Height	140	140	145	149	<u>150</u>	151	157	163	229

- We also measure the **variation** in our data.
- In essence, most of statistics, and in turn analytics, is about describing and understanding variation.

- The simplest measure of variation is the **range**:

$$\text{range} = \max(a) - \min(a)$$

## Example

What is the range of the heights of the two basketball squads?

ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149

ID	1	2	3	4	5	6	7	8
Height	192	102	145	165	126	154	123	188

### Example

What is the range of the heights of the two basketball squads?

$$\text{range} = 163 - 140 = 23$$

$$\text{range} = 192 - 102 = 90$$

- The **variance** of a sample measures the average difference between each value in a sample and the mean of that sample.
- The **variance** of the  $n$  values of a feature  $a, a_1, a_2 \dots a_n$ , is denoted  $var(a)$  and is calculated as:

$$var(a) = \frac{\sum_{i=1}^n (a_i - \bar{a})^2}{n - 1}$$

## Example

What is the variance of the heights of the two basketball squads?

ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149

ID	1	2	3	4	5	6	7	8
Height	192	102	145	165	126	154	123	188



## Example

$$\begin{aligned} \text{var}(\text{HEIGHT}) &= \frac{(150 - 149.375)^2 + (163 - 149.375)^2 + \dots + (149 - 149.375)^2}{8 - 1} \\ &= 63.125 \end{aligned}$$

$$\begin{aligned} \text{var}(\text{HEIGHT}) &= \frac{(192 - 149.375)^2 + (102 - 149.375)^2 + \dots + (188 - 149.375)^2}{8 - 1} \\ &= 1,011.41071 \end{aligned}$$

- The **standard deviation**,  $sd$ , of a sample is calculated by taking the square root of the **variance** of the sample:

$$sd(a) = \sqrt{var(a)} \quad (1)$$

$$= \sqrt{\frac{\sum_{i=1}^n (a_i - \bar{a})^2}{n - 1}} \quad (2)$$

## Example

What is the standard deviation of the heights of the two basketball squads?

ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149

ID	1	2	3	4	5	6	7	8
Height	192	102	145	165	126	154	123	188

## Example

$$\begin{aligned}sd(\text{HEIGHT}) &= \sqrt{63.125} \\ &= 7.9451 \dots\end{aligned}$$

$$\begin{aligned}sd(\text{HEIGHT}) &= \sqrt{1,011.41071} \\ &= 31.8026 \dots\end{aligned}$$

- **Percentiles** are another useful measure of the variation of the values for a feature: a proportion of  $\frac{i}{100}$  of the values in a sample take values equal to or lower than the  $i^{th}$  percentile of that sample.

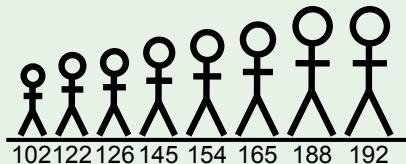
- To calculate the  $i^{th}$  percentile of the  $n$  values of a feature  $a$ ,  $a_1, a_2 \dots a_n$ :
  - ▶ First order the values in ascending order and then multiply  $n$  by  $\frac{i}{100}$  to determine the *index*.
  - ▶ If the *index* is a whole number we take the value at that position in the ordered list of values as the  $i^{th}$  percentile.
  - ▶ If *index* is not a whole number then we **interpolate** the value for the  $i^{th}$  percentile as:

$$i^{th} \text{ percentile} = (1 - \text{index\_f}) \times a_{\text{index\_w}} + \text{index\_f} \times a_{\text{index\_w}+1}$$

where *index\_w* is the whole part of *index*, *index\_f* is the fractional part of *index* and  $a_{\text{index\_w}}$  is the value in the ordered list at position *index\_w*.

## Example

ID	2	7	5	3	6	4	8	1
Height	102	123	126	145	154	165	188	192



- What is the 25<sup>th</sup> percentile of the heights of the basketball squad?
- What is the 80<sup>th</sup> percentile of the heights of the basketball squad?

## Example

- To calculate the 25<sup>th</sup> percentile we first calculate *index* as  $\frac{25}{100} \times 8 = 2$ . So, the 25<sup>th</sup> percentile is the second value in the ordered list which is 123.
- To calculate the 80<sup>th</sup> percentile we first calculate *index* as  $\frac{80}{100} \times 8 = 6.4$ . Because *index* is not a whole number we set *index\_w* to the whole part of *index*, 6, and *index\_f* to the fractional part, 0.4. Then we can calculate the 80<sup>th</sup> percentile as:

$$(1 - 0.4) \times 165 + 0.4 \times 188 = 174.2$$



- We can use percentiles to describe another measure of variation known as the **inter-quartile range**.
- The inter-quartile range is calculated as the difference between the 25<sup>th</sup> percentile and the 75<sup>th</sup> percentile.<sup>1</sup>

---

<sup>1</sup>These percentiles are also known as the **lower quartile** (or 1<sup>st</sup> quartile) and **upper quartile** (or 3<sup>rd</sup> quartile) hence the name inter-quartile range.

### Example

For the heights of the first basketball team the inter-quartile range is  $151 - 140 = 11$ , while for the second team it is  $165 - 123 = 42$ .

- For categorical features we are interested primarily in **frequency counts** and **proportions**.
  - ▶ The frequency count of each level of a categorical feature is calculated by counting the number of times that level appears in the sample.
  - ▶ The proportion for each level is calculated by dividing the frequency count for that level by the total sample size.
  - ▶ Frequencies and proportions are typically presented in a **frequency table**.
- The **mode** is a measure of the central tendency of a categorical feature and is simply the most frequent level.
- We often also calculate a **second mode** which is just the second most common level of a feature.

**Table:** A dataset showing the positions and weekly training expenses of a school basketball squad.

Training			Training		
ID	Position	Expenses	ID	Position	Expenses
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41

**Table:** A frequency table for the POSITION feature from the professional basketball squad dataset in Table 4 <sup>[34]</sup>.

Level	Count	Proportion
guard	8	40%
forward	7	35%
center	5	25%

- In statistics it is very important to understand the difference between a **population** and a **sample**.
- The term population is used in statistics to represent all possible measurements or outcomes that are of interest to us in a particular study or piece of analysis.
- The term sample refers to the subset of the population that is selected for analysis.
- The **margin of error** reported in poll results takes into account the fact that the result is based on a sample from a much larger population.

**Table:** A number of poll results from the run up to the 2012 US Presidential election.

<b>Poll</b>	<b>Obama</b>	<b>Romney</b>	<b>Other</b>	<b>Date</b>	<b>Margin of Error</b>	<b>Sample Size</b>
Pew Research	50	47	3	04-Nov	$\pm 2.2$	2,709
Gallup	49	50	1	04-Nov	$\pm 2.0$	2,700
ABC News/Wash Pos	50	47	3	04-Nov	$\pm 2.5$	2,345
CNN/Opinion Research	49	49	2	04-Nov	$\pm 3.5$	963
Pew Research	50	47	3	03-Nov	$\pm 2.2$	2,709
ABC News/Wash Post	49	48	3	03-Nov	$\pm 2.5$	2,069
ABC News/Wash Post	49	49	2	30-Oct	$\pm 3.0$	1,288

# Data Visualization

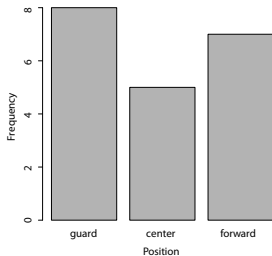


- When performing data exploration **data visualization** can help enormously.
- In this section we will describe three important data visualization techniques that can be used to visualize the values in a single feature:
  - ▶ the **bar plot**
  - ▶ the **histogram**
  - ▶ the **box plot**

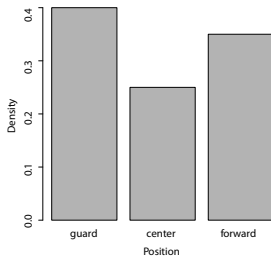
**Table:** A dataset showing the positions and weekly training expenses of a school basketball squad.

Training			Training		
ID	Position	Expenses	ID	Position	Expenses
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41

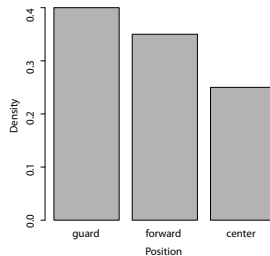
## Bar plots are great for categorical features



(a) Frequency

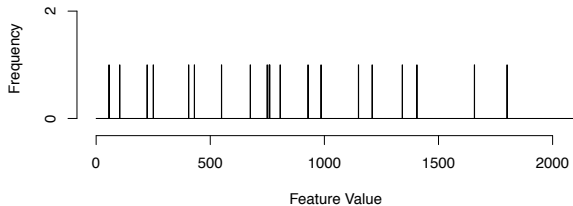


(b) Proportion



(c) Ordered

## Bar plots don't work for continuous features



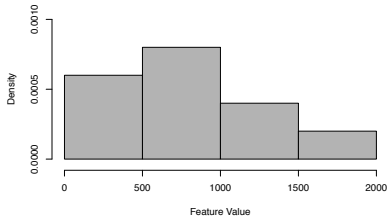
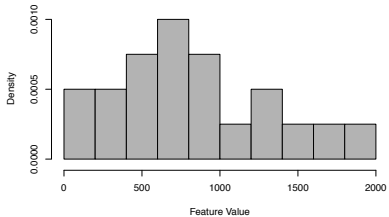
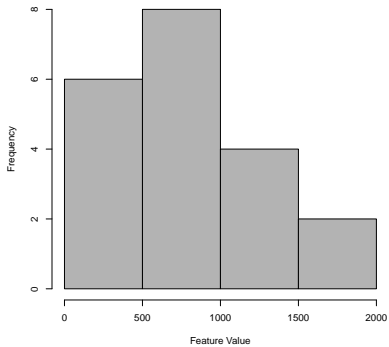
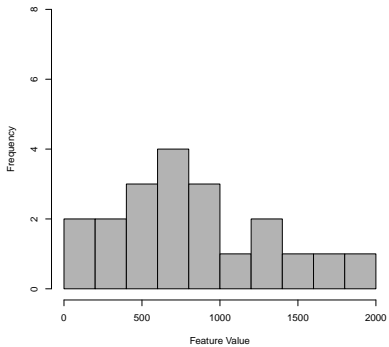
By dividing the range of a variable into intervals, or bins, we can generate **histograms**

(a) 200 unit intervals

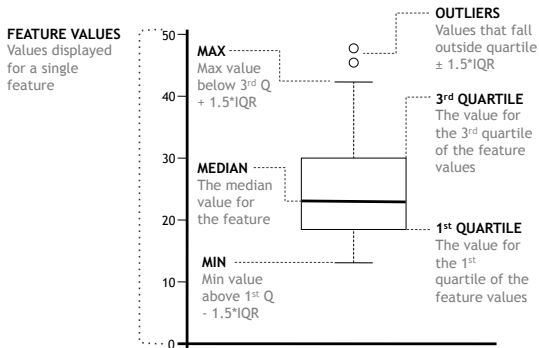
Interval	Count	Density	Prob
[0, 200)	2	0.0005	0.1
[200, 400)	2	0.0005	0.1
[400, 600)	3	0.00075	0.15
[600, 800)	4	0.001	0.2
[800, 1000)	3	0.00075	0.15
[1000, 1200)	1	0.00025	0.05
[1200, 1400)	2	0.0005	0.1
[1400, 1600)	1	0.00025	0.05
[1600, 1800)	1	0.00025	0.05
[1800, 2000)	1	0.00025	0.02

(b) 500 unit intervals

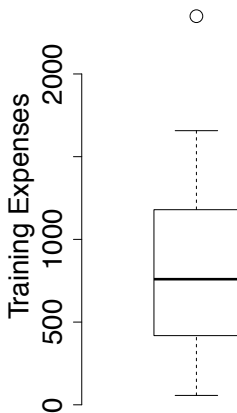
Interval	Count	Density	Prob
[0, 500)	6	0.0006	0.3
[500, 1000)	8	0.0008	0.4
[1000, 1500)	4	0.0004	0.2
[1500, 2000)	2	0.0002	0.1



**Box plots** are another useful way of visualising continuous variables



**Figure:** The structure of a box plot.



**Figure:** A box plot for the TRAINING EXPENSES feature from the basketball squad dataset in Table 4 <sup>[34]</sup>.



# Summary

## 1 Descriptive Statistics

- Descriptive Statistics for Continuous Features
- Descriptive Statistics for Categorical Features
- Populations & Samples

## 2 Data Visualization

- Bar Plots
- Histograms
- Box Plots

## 3 Summary