# Descriptive Statistics (contd.)

## Frequency Distribution Table

- We can summarize the raw data into a frequency distribution table.
- We can make some classes and count the frequency for each class.
- Each class has an upper boundary and a lower boundary.
- Conventionally the upper boundary is excluded from a class interval. So, the class interval 10 - 15 means at least 10 but strictly less than 15.

## Histogram

- Histogram is a graphical summarization of data.
- It is drawn based on a distribution table.
- Histogram is made up of a set of rectangular blocks.
- There is no gap between two consecutive blocks.
- Class intervals are drawn on the horizontal axis.
- Each block is drawn in such a way that the **area** of the block is proportional to the **percentage** (or frequency or relative frequency) in the corresponding class interval. We will deal only with percentage type histograms.

- In other words, height of the blocks are proportional to the frequency densities where

$$\text{frequency density of a class interval} = \frac{\text{percentage in that class interval}}{\text{length of the class interval}}$$

- On the vertical axis one has the density scale, i.e., the units on the vertical axis are % (or frequency or relative frequency)/units on the horizontal axis.

- In a histogram, the height of a block represents crowding – percentage (or frequency or relative frequency) per horizontal unit.

- Total area of the blocks is 100% for a percentage type histogram, total frequency for a frequency type histogram, and 1 for relative frequency type histogram.

## Numerical Summarization

- Measures of central tendency
  - "Where is the middle of the data?"
  - Two major ones are mean and median
- Measures of dispersion
  - "How much variation is there in the data?"
  - Major ones are variance, standard deviation and inter-quartile range (IQR)

## Mean or Average

- Sum data points and divide the sum by number of data point
- Provides "balance point" of data
- Easily influenced by outliers
- Notation for population mean: $\mu$, notation for sample mean $\overline{x}$.

## Median

- Sort data points. Meadian is the middle data point. If the number of data points is even, interpolate between the middle two data points.
- Usually not influenced by outliers

## Percentiles

- $x$th percentile separates the bottom $x\%$ from the top $(100 - x)\%$
- First quartile (Q1): 25th percentile, marks upper boundary for lower fourth
- Median: second quartile (Q2) or 50th percentile, separates the population into top and bottom halves
- Third quartile (Q3): 75th percentile, marks lower boundary for upper fourth

## Variance

- Might want to measure dispersion as average distance of data points from the mean:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)$$

- But this is always 0, so look at the average squared distance:

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

- Standard deviation is square root of variance
- Advantages: Mathematically useful, frequently used
- Drawbacks: Easily infulenced by outliers,difficult to calculate by hand

## Polupation Variance

- Requires knowing population mean $\mu$:

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

- This is fine if we know all the members of a population (so we can know the mean)

## Sample Variance

- Use sample mean $\overline{x}$ instead of population mean:

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

- Calculated from samples, used as estimator of $\sigma^2$
- Use $n-1$ instead of $n$ in denominator
- As $n$ grows, the two calculations become closer

## Inter-quartile Range (IQR)

- IQR = Q3 – Q1
- Measures how far the median of the top half from the median of the bottom half is
- Unlike variance not so influenced by outliers
- Not very commonly used

## Symmetry and Skewness

- If the left half of a histogram is the mirror image of the right half it is symmetric.
- If the hisogram has a long right tail, it is right-skewed
- If it has a long left tail, it is left-skewed
- If histogram is symmetric, mean = median, if right-skewed, mean > median and if left-skewed, mean < median

## Empirical Rule

- About 68% of the data fall within 1 SD from the mean
- About 95% of the data fall within 2 SD from the mean
- Almost all of the data fall within 3 SD from the mean