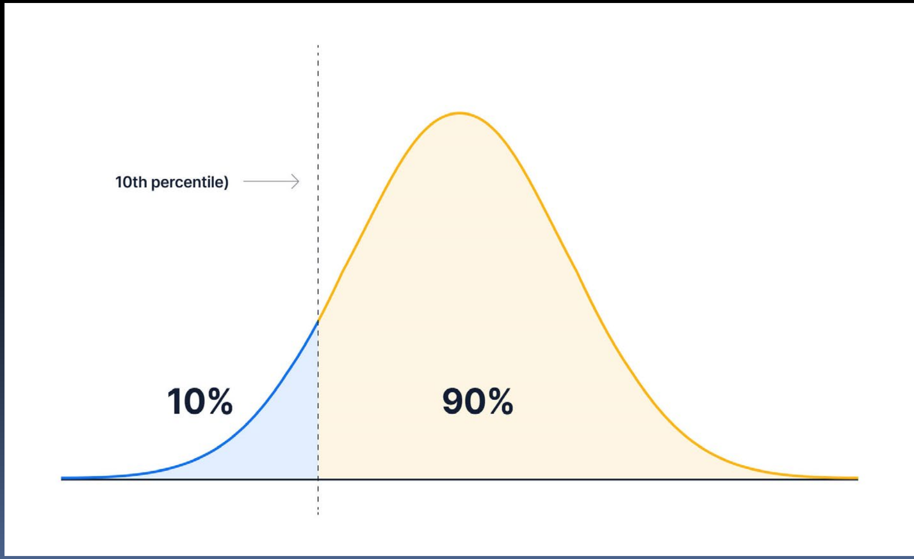
MAS250 Probability and Statistics

CHAPTER 2.

# DESCRIPTIVE STATISTICS

# Population Percentiles

- 100p^th percentile: location that indicates the percent of a distribution that is equal to or below 100p
- p=0.25: first quartile (Q1)
- p=0.5: median (Q2)
- p=0.75: third quartile (Q3)




IQR=Q3-Q1

Range = Max – min

More robust to outliers.

- The sample median
  - Order the values of a data set of size $n$ from smallest to largest.
  - If $n$ is odd, the sample median is the value in position $\frac{n+1}{2}$.
  - If $n$ is even, the sample median is the average of the values in positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

# The sample $\mathbf{100}\boldsymbol{p}$ percentile of a data set

- The sample $100p$ percentile $x$ of a data set
  - at least $np$ of the values are less than or equal to $x$
  - at least $n(1-p)$ of the values are greater than and equal to $x$
  - If there are more than one, take their average.

# The sample $100p$ percentile

- Calculating the 100pth Percentile:
- Step 1: Arrange the data values in ascending order.
- Step 2: Compute an index i as follows: i = (pn) where 100p is the percentile of interest and n is the number of data values.
- Step 3:

(a) If i is not an integer, the next integer greater than i denotes the position of the 100pth percentile.

(b) If i is an integer, the 100pth percentile is the average of the data values in positions i and i+1.

Sum of Deviation $= 0$. $S = 0 \Rightarrow \forall x \in X,\ x = 0$.

# Shape of Distribution

Range, $S^2$, $S$ affected a lot by outliers

$\Rightarrow$ IQR might be good choice

bell-shaped
(uni-model, sym.)

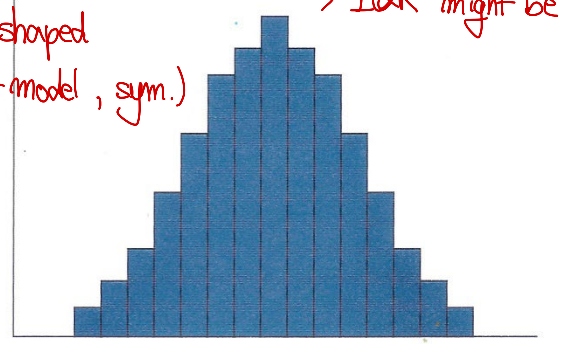


FIGURE 2.8 *Histogram of a normal data set.* mean $\approx$ median

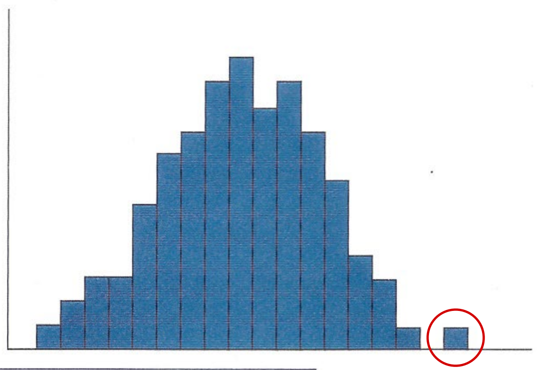


FIGURE 2.9 *Histogram of an approximately normal data set.*

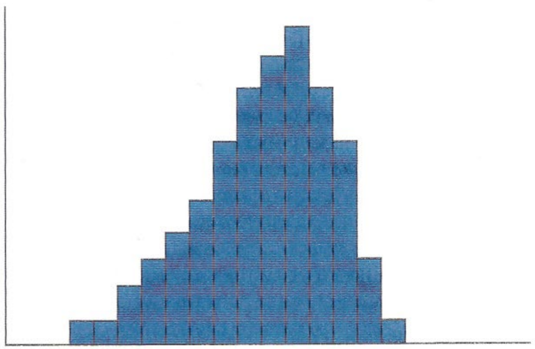


FIGURE 2.10 *Histogram of a data set skewed to the left.* mean $<$ median

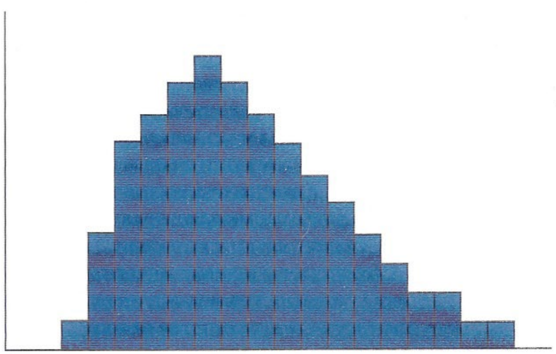


FIGURE 2.11 *Histogram of a data set skewed to the right.* mean $>$ median

# Bimodal Shape

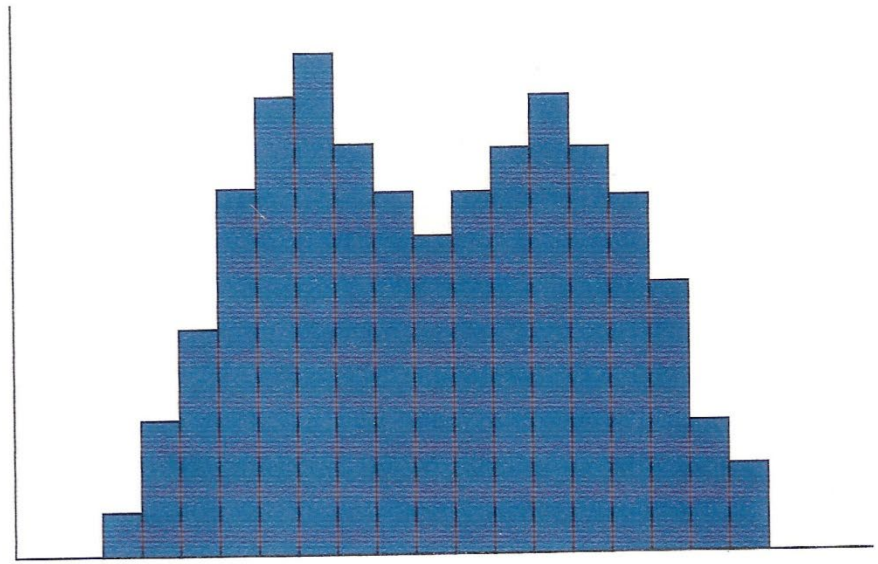


FIGURE 2.12 *Histogram of a bimodal data set.*

# The empirical rule Bell-Shaped

- Approximately 68 % of the observations lie within $\bar{x} \pm s$
- Approximately 95 % of the observations lie within $\bar{x} \pm 2s$
- Approximately 99.7 % of the observations lie within $\bar{x} \pm 3s$

General-Shaped

# Chebyshev's Inequality

- For a data set $\{x_1, x_2, \cdots, x_n\}$

  The sample mean: $\bar{x}$

  The sample standard deviation: $s$

$k \geq 2.$

Let $S_k = \{i, 1 \leq i \leq n : |x_i - \bar{x}| < ks\}$.

$N(S_k)$ : the number of elements in the set $S_k$

$$\frac{N(S_k)}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}$$

# One-sided Chebyshev Inequality

- For $k > 0$,
  let $N(k)$ be the number of $i$'s such that
  $$\{i, 1 \leq i \leq n : x_i - \bar{x} \geq ks\}.$$

$$\frac{N(k)}{n} \leq \frac{1}{1 + k^2}$$

# Paired data

- For $(x_i, y_i), i = 1,2, \dots, n$
- The sample correlation coefficient $r$ $= \hat{\rho}$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

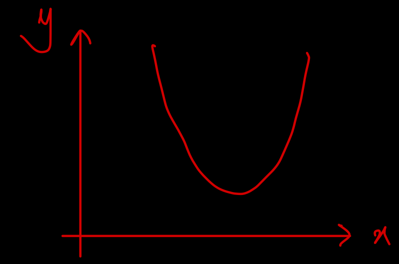
$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$x = (x_1, \cdots, x_n)$
$y = (y_1, \cdots, y_n)$

$$\cos\theta = \frac{x \cdot y}{\|x\| \|y\|} = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = r. \quad \text{if } \bar{x} = \bar{y} = 0.$$

- If $r > 0$, then the sample data pairs are positively correlated.
- If $r < 0$, then the sample data pairs are negatively correlated.

# Properties of $r$




$\Rightarrow r \approx 0$

since r is linear —.

- $-1 \leq r \leq 1$
- If for constants $a$ and $b$, with $b > 0$,
$$y_i = a + bx_i \,, i = 1,2, \ldots, n,$$
then $r = 1$.
- If for constants $a$ and $b$, with $b < 0$,
$$y_i = a + bx_i \,, i = 1,2, \ldots, n,$$
then $r = -1$.
- $r$ is also the sample correlation coefficient for the data pairs $(a + bx_i, c + dy_i), 1 \leq i \leq n$, provided that $b$ and $d$ are both positive or both negative.