Summary measures

Preliminaries

- Σ is shorthand for addition
- Suppose x_i is the *ith* observation.

$$x_1 + x_2 + x_3 = \sum_{i=1}^{3} x_i$$

- If a is a constant

$$\sum_{i=1}^{n} a = na$$

- Mixed example:

$$\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i$$

Summary measures

Preliminaries

- b a constant
- n observations
- Simplify $\sum_{i=1}^{n}(a + bx_i)$
- Group terms:

$$
\begin{aligned}
\sum_{i=1}^{n}(a + bx_i) &= a + bx_1 + a + bx_2 + \ldots + a + bx_n \\
&= a + a + a \ldots . bx_1 + bx_2 + \ldots bx_n \\
&= an + b\sum_{i=1}^{n} x_i
\end{aligned}
$$

Summary measures

Preliminaries

- Let $\{x_i, y_i\}$ be paired observations
- k another constant.
- Simplify:

$$\sum_{i=1}^{n} \left(ak + bx_i + c\left(x_i y_i\right)\right) = ???$$

- Treat $x_i y_i$ as any other variable that varies in i.

$$\sum_{i=1}^{n} \left(ak + bx_i + c\left(x_i y_i\right)\right) = nak + \sum_{i=1}^{n} bx_i + \sum_{i=1}^{n} \left(cx_i y_i\right)$$

- Pull out the constants:

$$nak + \sum_{i=1}^{n} bx_i + \sum_{i=1}^{n} \left(cx_i y_i\right) = nak + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} \left(x_i y_i\right)$$

Summary measures

Statistics that describe distributions

- Measures of central tendency
  1. Mean
  2. Median
  3. Mode
- Measures of variation
  1. Range
  2. Variance
  3. Standard Deviation
- Measures of shape
  1. "Skew"

Summary measures

Measures of central tendency

- **Mean**
  - Most common
  - Central tendency

$$\widehat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Median**
  - Mid-point of the data.
  - To calculate:
    1. Order the data.
    2. Calculate $(n+1)/2$
    3. *Take the value at $(n+1)/2$, or the average of the two closest (if $(n+1)/2$ is not whole)*

Summary measures

Measures of central tendency (cont.)

- **Mode**
  - The value that occurs most often
- Question: which measure(s) of central tendency are not affected by extreme values?

$\Rightarrow$ *Median and Mode*

Summary measures

Measures of variation

- **Range**
  - A measure of data dispersion, though not used for many applications
  - To calculate:
    1. Identify the largest observation
    2. Identify the smallest observation
    3. Take the difference

Summary measures

Measures of variation (cont.)

- **Sample Variance**
  - The most commonly used measure of dispersion.
  - Summarizes the how far a typical observation is from the mean

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

  - Why do we divide by $n - 1$ instead of n?

$\Rightarrow$ *We used one "piece" of information to calculate* $\widehat{\mu}_x$

Summary measures

Measures of variation (cont.)

- **Sample Standard deviation**

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

- This is more desirable than the sample variance. Why?
  - Same scale as the mean.

Covariance

Relationships

- Covariance describes the relationship between two random variables

$$\widehat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \widehat{\mu}_x)(y_i - \widehat{\mu}_y)$$

- When will covariance be positive/negative?
  - $\widehat{\sigma}_{xy} > 0 \Rightarrow$ tends to have $x_i > \widehat{\mu}_x$ when $y_i > \widehat{\mu}_y$ (and vice versa)
  - $\widehat{\sigma}_{xy} < 0 \Rightarrow$ tends to have $x_i > \widehat{\mu}_x$ when $y_i < \widehat{\mu}_y$ (and vice versa)
- Covariance describes a "linear" relationship
- Any non-linear relationships with zero covariance?

Correlation

Basic

- Correlation describes a linear relationship
- $\widehat{\rho}_{xy} = \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x \widehat{\sigma}_y}$
- $\widehat{\rho}_{xy} \in [-1, 1]$
- Can you prove that correlation is between -1 and 1?

Data Scaling

Mean

- What happens when we scale variables by either adding a constant or multiplying by a constant?
- Sample: $X = \{2, 3, 4\}$
- Calculate the mean of X:

$$\begin{aligned}
\widehat{\mu}_x &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{3}(2+3+4) \\
&= 3
\end{aligned}$$

- Calculate the variance of X:

$$\begin{aligned}
\widehat{\sigma}_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \widehat{\mu}_x\right)^2 \\
\widehat{\sigma}_x^2 &= \frac{1}{2}\left((2-3)^2 + (3-3)^2 + (4-3)^2\right) = \frac{1}{2}(1+0+1) = 1
\end{aligned}$$

Scaling

Adding a constant

- What if we define a new variable, $Z = X + 3$
- Calculate the mean of $Z = \{5, 6, 7\}$
- What will happen to the mean, variance?

$$\begin{aligned}
\widehat{\mu}_z &= \frac{1}{3}(5+6+7) = 6 \\
\widehat{\sigma}_z^2 &= \frac{1}{2}\left((5-6)^2 + (6-6)^2 + (7-6)^2\right) \\
&= \frac{1}{2}(1+0+1) = 1
\end{aligned}$$

- Adding a constant **does** affect central tendency.
- Adding a constant **does not** affect dispersion.
- $1 \quad \underline{\mathbf{2} \quad \mathbf{3} \quad \mathbf{4}} \quad 5 \quad 6 \quad 7 \quad 8 \quad 9$
- $1 \quad 2 \quad 3 \quad 4 \quad \underline{\mathbf{5} \quad \mathbf{6} \quad \mathbf{7}} \quad 8 \quad 9$

Scaling

Multiplying by a constant

- What if we define a new variable, $J = 3X$
- Calculate the mean/variance of $J = \{6, 9, 12\}$
- What will happen to the mean, variance?

$$
\begin{aligned}
\widehat{\mu}_z &= \frac{1}{3}(6+9+12) = 9 \\
\widehat{\sigma}_z^2 &= \widehat{\sigma}_j^2 = \frac{1}{2}\left((6-9)^2 + (9-9)^2 + (12-9)^2\right) \\
&= \frac{1}{2}(9+0+9) = 9
\end{aligned}
$$

- Multiplying by a constant affects both mean and variance.

1	**2**	**3**	**4**	5	6	7	8	9	10	11	12
1	2	3	4	5	**6**	7	8	**9**	10	11	**12**

- Generally, if $J = aX$, then $\widehat{\sigma}_j^2 = a^2 \widehat{\sigma}_x^2$, $\widehat{\mu}_j = a\widehat{\mu}_x$

Covariance

Scaling

- Suppose $Z = aX$. What is $\widehat{\sigma}_{zy}$?
- Write $\widehat{\sigma}_{zy}$

$$\widehat{\sigma}_{zy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \widehat{\mu}_z\right)\left(y_i - \widehat{\mu}_y\right)$$

- Substitute for $\widehat{\mu}_z$ and z_i

$$\widehat{\sigma}_{zy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(ax_i - a\widehat{\mu}_x\right)\left(y_i - \widehat{\mu}_y\right)$$

- Factor out a and simplify:

$$\widehat{\sigma}_{zy} = a\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \widehat{\mu}_x\right)\left(y_i - \widehat{\mu}_y\right) = a\widehat{\sigma}_{xy}$$

- Covariance is sensitive to scale!! Is this a problem?

Correlation

Scaling

- Is correlation sensitive to scale?
- Suppose $Z = aX$, $a > 0$.
- $\widehat{\rho}_{zy} = \frac{\widehat{\sigma}_{zy}}{\widehat{\sigma}_z \widehat{\sigma}_y} = \frac{a\widehat{\sigma}_{xy}}{a\widehat{\sigma}_x \widehat{\sigma}_y} = \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x \widehat{\sigma}_y} = \widehat{\rho}_{xy}$