Module 7

Lecture 4: Measures of symmetry and correlation
Measures of Symmetry

Coefficient of Skewness (Cs):

Denoted as \( \gamma = \frac{\mu_3}{\mu_2^{3/2}} \) Population parameter

\[ Cs = \frac{n}{(n-1)(n-2)S^3} \sum_{i=1}^{n} (x_i - \bar{x})^3 \] Sample estimate

and S is Standard Deviation.

Mode > Median > Mean
Skewed towards right, so, (Positive skew), Cs > 0

Mode = Median = Mean
Symmetrical

Mode > Median > Mean
Skewed towards left, so, (Negative skew), Cs < 0
Measures of Symmetry

Kutosis or Measure of Peakedness ($\kappa$):

$$\kappa = \frac{\mu_4}{\mu_2^2} \quad \text{Population parameter}$$

$$\kappa = \frac{n^2}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{s^4} \quad \text{Sample estimate}$$

- If $k > 3$ Leptokurtic Very Steep
- If $K = 3$ Mesokurtic Normal
- If $K < 3$ Platykurtic Flat

![Diagram showing distributions with different kurtosis values: Steep Dist. $K > 3$, Avg Dist. $K > 1$, Flat Dist. $K < 1$.]

Module 7
Covariance

- Applicable to only 2-D RVs

\[
\text{CoV} (x,y) = \sigma_{x,y} = M_{1,1} = \int_{-\alpha}^{\alpha} \int_{-\alpha}^{\alpha} (x - \mu_x)(y - \mu_y)f(x,y)dx\,dy
\]

\[
= E [(x - \mu_x)(y - \mu_y)]
\]

Sample Estimate for covariance,

\[
S_{x,y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}
\]

*When* \(x\) & \(y\) *are independent*

\[
\text{CoV} (x,y) = \int_{-\alpha}^{\alpha} \int_{-\alpha}^{\alpha} (x - \mu_x)(y - \mu_y)g(x)h(y)dx\,dy
\]

\[
= \int_{-\alpha}^{\alpha} (x - \mu_x)g(x)dx + \int_{-\alpha}^{\alpha} (y - \mu_y)h(y)dy = 0
\]

If \(x\) and \(y\) are independent, \(\text{CoV} (x,y) = 0\), but converse is not true.
Correlation Coefficient

Degree of linear association between two RVs $X$ and $Y$

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} \quad -1 \leq \rho_{x,y} \leq 1$$

where $\sigma_{x,y}$ - Covariance
$\sigma_x$ - Std. dev. of $x$
$\sigma_y$ - Std. dev. of $y$

(only indicates the degree of linear association)

If $\rho_{x,y} = 0$ we can not say that $X$ and $Y$ are independent but we can say that there is no linear dependence

Sample estimate

$$r_{x,y} = \frac{S_{x,y}}{S_x S_y} \quad -1 \leq r_{x,y} \leq 1 \text{ (unitless)}$$
Perfect Linear Relationship

\[ Y = aX + b \]

\[ \rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} \]

\[ \therefore E[(X - \mu_x)(Y - \mu_y)] = E[XY - X\mu_y - Y\mu_x + \mu_x \mu_y] = E[XY] - 2\mu_x \mu_y + \mu_x \mu_y \]

\[ = E[XY] - \mu_x \mu_y = E[XY] - E[X]E[Y] \]

\[ \rho_{x,y} = \frac{E(XY) - E(X)E(Y)}{\sigma_x \sigma_y} \]

Now, \( y = ax + b \)

\[ \therefore \rho^2 = \left[ \frac{E(XY) - E(X)E(Y)}{\sigma_x^2 \sigma_y^2} \right]^2 \]
Perfect Linear Relationship

Linear relationship, \( y = ax + b \)

\[
\rho^2 = \frac{\left[ \text{E}(XY) - \text{E}(X)\text{E}(Y) \right]^2}{\sigma_x^2 \sigma_y^2} \quad \text{var}(ax+b) = a^2 \text{var}(x)
\]

Putting that relation,

\[
\begin{align*}
\rho^2 &= \frac{\left[ \text{E}(ax^2 + xb) - \text{E}(x)\text{E}(ax + b) \right]^2}{\sigma_x^2 \sigma_y^2} \\
&= \frac{\left[ a\text{E}(X^2) + b\text{E}(x) - a\text{E}(x))^2 - b\text{E}(x) \right]^2}{\sigma_x^2 \sigma_y^2} \\
&= \frac{\left[ a\text{E}(X^2) - a\text{E}(x))^2 \right]^2}{\sigma_x^2 \sigma_y^2} = \frac{a^2(\sigma_x^2)^2}{\sigma_x^2 \sigma_y^2} \quad \text{Eq (1)}
\end{align*}
\]

\[\therefore \sigma^2 = \text{E}[X^2] - [\text{E}(x)]^2\]
\[ \sigma_y^2 = E[Y^2] - [E(Y)]^2 \]

\[ = E[(ax+b)^2] - [E(ax + b)]^2 \]

\[ = E[a^2x^2 + 2abx + b^2] - [aE(x) + b]^2 \]

\[ = a^2E(x^2) + 2abE(x) + b^2 - a^2[E(x)]^2 - 2aE(x)b - b^2 \]

\[ = a^2E[X^2] - (E(X))^2 = a^2\sigma_x^2 \]

Put this value in Eq (1)

\[ \frac{a^2(\sigma_x^2)^2}{\sigma_x^2 \sigma_y^2} = 1 \quad \therefore \rho = 1 \text{ or } \rho = \pm 1 \]

\[ \Rightarrow \text{a perfect linear relationship results in } \rho = \pm 1 \text{ (increasing or decreasing)} \]
Example Problem

• Find correlation between A & Q

<table>
<thead>
<tr>
<th>Q</th>
<th>A</th>
<th>(Q - Q̄)</th>
<th>(A - Ā)</th>
<th>(Q - Q̄)(A - Ā)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.5</td>
<td>1.25</td>
<td>-6.223</td>
<td>-0.271</td>
<td>1.686</td>
</tr>
<tr>
<td>8.5</td>
<td>0.871</td>
<td>-13.223</td>
<td>-0.650</td>
<td>8.593</td>
</tr>
<tr>
<td>85</td>
<td>5.69</td>
<td>63.277</td>
<td>4.169</td>
<td>263.811</td>
</tr>
<tr>
<td>105</td>
<td>8.27</td>
<td>83.277</td>
<td>6.749</td>
<td>562.049</td>
</tr>
<tr>
<td>24.8</td>
<td>1.62</td>
<td>3.077</td>
<td>0.099</td>
<td>0.305</td>
</tr>
<tr>
<td>3.8</td>
<td>0.175</td>
<td>-17.923</td>
<td>-1.346</td>
<td>24.122</td>
</tr>
<tr>
<td>1.76</td>
<td>0.148</td>
<td>-19.963</td>
<td>-1.373</td>
<td>27.406</td>
</tr>
<tr>
<td>18.00</td>
<td>1.40</td>
<td>-3.723</td>
<td>-0.121</td>
<td>0.450</td>
</tr>
<tr>
<td>8.75</td>
<td>0.257</td>
<td>-12.973</td>
<td>-1.264</td>
<td>16.396</td>
</tr>
<tr>
<td>8.25</td>
<td>0.322</td>
<td>-13.473</td>
<td>-1.199</td>
<td>16.152</td>
</tr>
<tr>
<td>3.56</td>
<td>0.178</td>
<td>-18.163</td>
<td>-1.343</td>
<td>24.390</td>
</tr>
<tr>
<td>1.90</td>
<td>0.148</td>
<td>-19.823</td>
<td>-1.373</td>
<td>27.214</td>
</tr>
<tr>
<td>16.50</td>
<td>0.872</td>
<td>-5.223</td>
<td>-0.649</td>
<td>3.389</td>
</tr>
<tr>
<td>2.80</td>
<td>0.091</td>
<td>-18.923</td>
<td>-1.430</td>
<td>27.057</td>
</tr>
<tr>
<td>Mean</td>
<td>21.723</td>
<td>1.521</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td></td>
<td></td>
<td>1003.020</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>32.050</td>
<td>2.423</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Spurious Correlation

- A spurious correlation arises purely by chance

- Assume that, we got some correlation between height of employees and their salary, but we are sure that there is no correlation, this type of correlation is called “spurious correlation”

- Most correlations encountered in practice are not “spurious”