Measurement Math
DeShon - 2007

Univariate Descriptives
- Mean
- Variance, standard deviation
  \[ s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \]
- Skew & Kurtosis
- If normal distribution, mean and SD are sufficient statistics

Normal Distribution

Univariate Probability Functions

Bivariate Descriptives
- Mean and SD of each variable and the correlation (\( \rho \)) between them are sufficient statistics for a bivariate normal distribution
- Distributions are abstractions or models
  - Used to simplify
  - Useful to the extent the assumptions of the model are met

Readings?
- Dawes
- Grove et al
- Wang & Stanley
Covariance

- Covariance is the extent to which two variables co-vary from their respective means

\[
\text{Cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}
\]

\[
\text{Cov}(X, Y) = \frac{17}{(4-1)} = 5.667
\]

Covariance ranges from negative to positive infinity.

- Variance - Covariance matrix
  - Variance is the covariance of a variable with itself

\[
\Sigma = \begin{bmatrix}
\text{Var}(X) & \text{Cov}(XY) & \text{Cov}(XZ) \\
\text{Cov}(XY) & \text{Var}(Y) & \text{Cov}(YZ) \\
\text{Cov}(XZ) & \text{Cov}(YZ) & \text{Var}(Z)
\end{bmatrix}
\]

Correlation

- Covariance is an unbounded statistic
- Standardize the covariance with the standard deviations
- \(-1 \leq r \leq 1\)

\[
\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}
\]

Correlation Matrix

Table 1. Descriptive Statistics for the Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>SD</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual cog ability</td>
<td>1.60</td>
<td>0.71</td>
<td>-1.01 -0.71 -0.02 0.07 0.03 0.11 0.08 -0.02</td>
</tr>
<tr>
<td>Academic seniority</td>
<td>1.70</td>
<td>0.01</td>
<td>-0.12 0.45 0.06 0.14 0.23 0.22 0.07 0.17 1.01</td>
</tr>
<tr>
<td>Gender</td>
<td>2.17</td>
<td>0.00</td>
<td>-0.01 -0.21 -0.11 -0.06 -0.16 0.49 1.63</td>
</tr>
<tr>
<td>Age</td>
<td>2.74</td>
<td>0.05</td>
<td>-0.72 0.06 0.06 0.08 0.11 0.21 0.74 5.05</td>
</tr>
<tr>
<td>Collectivism</td>
<td>4.65</td>
<td>0.11</td>
<td>-0.72 0.37 0.84 0.42 0.25 1.11 4.65</td>
</tr>
<tr>
<td>Vert individualism</td>
<td>5.19</td>
<td>0.25</td>
<td>-0.80 0.82 0.25 0.41 1.05 5.19</td>
</tr>
<tr>
<td>Horiz individualism</td>
<td>5.92</td>
<td>0.40</td>
<td>-0.78 0.41 0.40 0.89 4.92</td>
</tr>
<tr>
<td>Individualism</td>
<td>4.89</td>
<td>0.34</td>
<td>-0.79 0.34 0.85 4.03</td>
</tr>
<tr>
<td>Self-enhancement</td>
<td>4.89</td>
<td>0.34</td>
<td>-0.81 0.86 4.89</td>
</tr>
<tr>
<td>Self-rated cog ability</td>
<td>2.17</td>
<td>0.00</td>
<td>-0.12 0.45 0.06 0.14 0.23 0.22 0.07 0.17 1.01</td>
</tr>
</tbody>
</table>

Notes: \(N = 608\); gender was coded 1 for male and 2 for female. Reliabilities (Coefficient alpha) are on the diagonal.
Coefficient of Determination
- $r^2 =$ percentage of variance in Y accounted for by X
- Ranges from 0 to 1 (positive only)
- This number is a meaningful proportion
- Used for criterion-related validity studies

Other measures of association
- Point Biserial Correlation
- Biserial Correlation
- Tetrachoric Correlation (binary variables)
- Polychoric Correlation (ordinal variables)
- Odds Ratio (binary variables)

Point Biserial Correlation
- Used when one variable is a natural (real) dichotomy (two categories) and the other variable is interval or continuous
- Just a normal correlation between a continuous and a dichotomous variable

Biserial Correlation
- When one variable is an artificial dichotomy (two categories) and the criterion variable is interval or continuous

Tetrachoric Correlation
- Estimates what the correlation between two binary variables would be if you could measure variables on a continuous scale.
- Example: difficulty walking up 10 steps and difficulty lifting 10 lbs.
- Assumes that both "traits" are normally distributed
- Correlation, $\rho$, measures how narrow the ellipse is.
- $a$, $b$, $c$, $d$ are the proportions in each quadrant
Tetrachoric Correlation

For $\alpha = \frac{ad}{bc}$,
Approximation 1:
$$Q = \frac{1}{1 + \alpha}$$
Approximation 2 (Digby):
$$Q = \frac{1}{\alpha} - \frac{1}{\alpha^2 + 1}$$

Example:
Tetrachoric correlation = 0.61
Pearson correlation = 0.41

Odds Ratio

- Measure of association between two binary variables
- Risk associated with $x$ given $y$.
- Example: odds of difficulty walking up 10 steps to the odds of difficulty lifting 10 lb.

$$OR = \frac{p1(1-p2)}{p2(1-p1)}$$
$$= \frac{40 \times 30}{20 \times 10} = 6$$

Pros and Cons

Tetrachoric correlation
- same interpretation as Spearman and Pearson correlations
- "difficult" to calculate exactly
- Makes assumptions

Odds Ratio
- easy to understand, but no "perfect" association that is manageable (i.e. $\infty$, -)
- easy to calculate
- not comparable to correlations
- May give you different results/inference!

Dichotomized Data: A Bad Habit of Psychologists

- Sometimes perfectly good quantitative data is made binary because it seems easier to talk about "High" vs. "Low"
- The worst habit is median split
- Usually, the high and low groups are mixtures of the continua
- Rarely is the median interpreted rationally
- See references


Simple Regression

- The simple linear regression MODEL is:
$$y = \beta_0 + \beta_1 x + \varepsilon$$
- describes how $y$ is related to $x$
- $\beta_0$ and $\beta_1$ are called parameters of the model
- $\varepsilon$ is a random variable called the error term.
Simple Regression

\[ E(y) = \beta_0 + \beta_1 x \]

- Graph of the regression equation is a straight line.
- \( \beta_0 \) is the population \( y \)-intercept of the regression line.
- \( \beta_1 \) is the population slope of the regression line.
- \( E(y) \) is the expected value of \( y \) for a given \( x \) value.

Estimated Simple Regression

The estimated simple linear regression equation is:

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]

- The graph is called the estimated regression line.
- \( \hat{\beta}_0 \) is the \( y \)-intercept of the line.
- \( \hat{\beta}_1 \) is the slope of the line.
- \( \hat{y} \) is the estimated/predicted value of \( y \) for a given \( x \) value.

Estimation process

- Least Squares Criterion

\[ \min \sum (y_i - \hat{y})^2 \]

where:

- \( y_i \) = observed value of the dependent variable for the \( i \)th observation
- \( \hat{y}_i \) = predicted/estimated value of the dependent variable for the \( i \)th observation

\[ \hat{\beta}_0 \text{ and } \hat{\beta}_1 \text{ provide estimates of } \beta_0 \text{ and } \beta_1 \]
Least Squares Estimation

- Estimated Slope
  \[ b_1 = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2/n} - \frac{\text{Cov}(X,Y)}{\text{Var}(X)} \]

- Estimated y-Intercept
  \[ b_0 = \bar{y} - b_1 \bar{x} \]

Model Assumptions

1. X is measured without error.
2. X and \( \varepsilon \) are independent.
3. The error \( \varepsilon \) is a random variable with mean of zero.
4. The variance of \( \varepsilon_i \), denoted by \( \sigma^2 \), is the same for all values of the independent variable (homogeneity of error variance).
5. The values of \( \varepsilon_i \) are independent.
6. The error \( \varepsilon_i \) is a normally distributed random variable.

Example: Consumer Warfare

<table>
<thead>
<tr>
<th>Number of Ads (X)</th>
<th>Purchases (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
</tr>
</tbody>
</table>

Example

- Slope for the Estimated Regression Equation
  \[ b_1 = \frac{220 - (10)(100)/5}{24 - (10)/5} = 5 \]
- y-Intercept for the Estimated Regression Equation
  \[ b_0 = 20 - 5(2) = 10 \]
- Estimated Regression Equation
  \[ \hat{y} = 10 + 5x \]

Example

- Scatter plot with regression line

Evaluating Fit

- Coefficient of Determination
  \[ r^2 = \frac{SST - SSE}{SST} \]
  where:
  \[ \text{SST} = \text{total sum of squares} \]
  \[ \text{SSR} = \text{sum of squares due to regression} \]
  \[ \text{SSE} = \text{sum of squares due to error} \]
Evaluating Fit

- **Coefficient of Determination**
  \[ r^2 = \frac{SSR}{SST} = \frac{100}{114} = 0.8772 \]

The regression relationship is very strong because 88% of the variation in number of purchases can be explained by the linear relationship with the number of TV ads.

Mean Square Error

- **An Estimate of \( \sigma^2 \)**
  The mean square error (MSE) provides the estimate of \( \sigma^2 \),
  \[ S^2 = MSE = \frac{SSE}{n-2} \]

where:
  \[ SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_0 - b_1X_i)^2 \]

Standard Error of Estimate

- **An Estimate of \( \sigma \)**
  To estimate \( \sigma \) we take the square root of \( \sigma^2 \).
  The resulting \( S \) is called the standard error of the estimate.
  Also called "Root Mean Squared Error"

\[ S = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} \]

Linear Composites

- **Linear composites** are fundamental to behavioral measurement
  - Prediction & Multiple Regression
  - Principle Component Analysis
  - Factor Analysis
  - Confirmatory Factor Analysis
  - Scale Development

Ex: Unit-weighting of items in a test
  \[ Test = 1*X_1 + 1*X_2 + 1*X_3 + \ldots + 1*X_n \]

Linear Composites

- **Sum Scale**
  \[ Scale_a = X_1 + X_2 + X_3 + \ldots + X_n \]

- **Unit-weighted linear composite**
  \[ Scale_{a1} = 1*X_1 + 1*X_2 + 1*X_3 + \ldots + 1*X_n \]

- **Weighted linear composite**
  \[ Scale_{aw} = b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n \]

Variance of a weighted Composite

\[ Var(Y) = \alpha_{22}^*Var(X) + \alpha_{22}^*Var(X) + 2\alpha_{22}^*Cov(X,Y) \]
Effective vs. Nominal Weights

- Nominal weights
  - The desired weight assigned to each component
- Effective weights
  - The actual contribution of each component to the composite
  - Function of the desired weights, standard deviations, and covariances of the components

Principles of Composite Formation

- Standardize before combining!!!!!
- Weighting doesn’t matter much when the correlations among the components are moderate to large
- As the number of components increases, the importance of weighting decreases
- Differential weights are difficult to replicate/cross-validate

Decision Accuracy

Polygraph Example

- Sensitivity, etc...