

DSC 410/510
Multivariate Statistical Methods
Examining Your Data

1

Why Examine Your Data?

- Applying multivariate techniques indiscriminately can:
 - ◆ further separate researcher from understanding the data
 - ◆ lead to exaggerated claims of “quick and easy” solutions
 - ◆ produce invalid results due to violated assumptions
- Researcher should endeavor to recognize potential problems as they occur, and apply appropriate remedies

2

Benefits of Examining Data

- Enables basic understanding of data and relationships amongst variables, particularly graphically
- Insurance against catastrophic errors
- Prompts consideration of missing data
- Alerts researcher to presence of potential outliers
- Allows assessment of statistical assumptions underlying multivariate techniques
- Leads to careful thought about representing nonmetric data

3

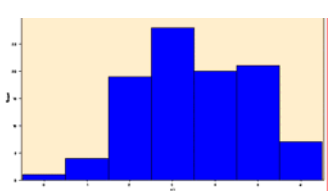
Reading Data Into SAS

- All class datasets are available in Excel spreadsheet format
 - ◆ Save hatco.xls from the class website to a folder in your home directory
- Start SAS
- Select File > Import Data to open the spreadsheet as a table in SAS
 - ◆ Select Browse to locate the Excel file you saved
 - ◆ Type “Hatco” (for example) as the Member name in the Work Library

4

Distribution Shapes: Histograms

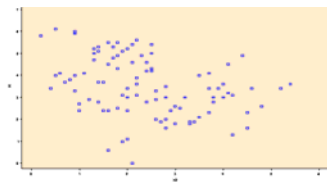
Analyzing Hatco data in SAS:
Solutions > Analysis > Analyst
File > Open By SAS Name (open Hatco in the Work library)
Graphs > Histogram



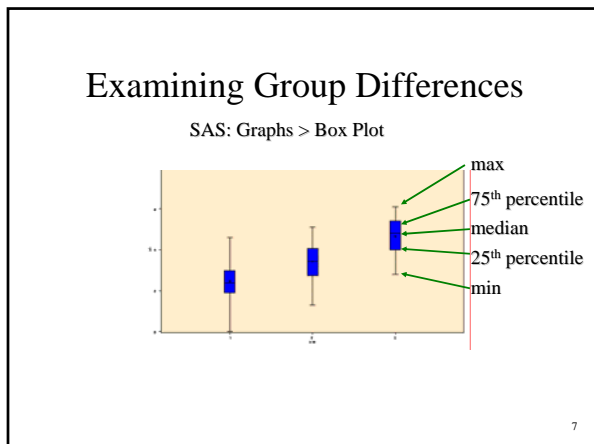
5

Relationships Between Variables

SAS: Graphs > Scatter Plot > Two-Dimensional



6



- ### Missing Data
- **Missing Data Process:** systematic cause of missing values in sample
 - If not accounted for, can lead to biased results (individuals with missing values may be different to those with complete data)
 - If all cases with any missing values are excluded, sample size can be drastically reduced
 - Need to either gather additional data or find remedy for missing data in original sample
- 8

- ### Missing Data Processes
- **Ignorable.** Consider variables X (no missing values) and Y (some missing values):
 - ◆ explicitly accommodated in analysis, e.g. censored data in survival analysis (Y = survival time)
 - ◆ missing completely at random (**MCAR**), i.e. values of Y are a random sample of all Y values
 - ◆ missing at random (**MAR**), i.e. missing values of Y depend on X but not on Y
 - **Non-ignorable**, i.e. missing data process related to missing Y values
- 9

- ### Diagnosing MCAR Processes
- Compare X -values in two groups – those with missing Y -values, those with valid Y -values
 - Consider bivariate correlations between missing value indicators for each variable
 - Overall MCAR test available in specialized software
- 10

- ### Dealing With Missing Data
- If available, use a technique that explicitly accounts for missing data, e.g. survival analysis
 - If process is MAR or non-ignorable, consult an expert!
 - If process is MCAR, use one of the following approaches
- 11

- ### Approaches for MCAR data
- Complete case approach (or listwise)
 - Delete cases and/or variables
 - All-available approach (e.g. pairwise correlation)
 - Replace missing data with imputed values
 - ◆ Case substitution
 - ◆ Mean substitution
 - ◆ Cold deck imputation
 - ◆ Regression imputation
 - ◆ Multiple imputation
 - Model-based (e.g. EM algorithm)
- 12

Outliers

- Observations with distinctly different combinations of characteristics
- Can be beneficial or problematic
- Identify and ascertain type of influence
- Types of outlier:
 - ◆ procedural error
 - ◆ result of an extraordinary event
 - ◆ no known explanation
 - ◆ values on each variable not unusual but combination of values is unique

13

Detecting Outliers

- Univariate: consider values > 2.5 standard deviations from the mean
- Bivariate: consider isolated points in scatterplots
- Multivariate: consider observations which have a very small p -value on Mahalanobis D^2 test
 - ◆ $D^2 = (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$
 - ◆ $D^2 \sim \chi_p^2$, where $p = \#$ variables in \mathbf{x}

14

Dealing With Outliers

- Once detected, profile to determine distinguishing features and classify
- Decide to retain or delete:
 - ◆ if they represent a segment of the population, retain to ensure generalizability
 - ◆ if truly aberrant, delete to avoid distorting analysis

15

Testing Assumptions

- Assessing statistical assumptions particularly important in multivariate analysis:
 - ◆ Complex relationships compound effects of assumption violations
 - ◆ Multivariate nature of the analysis can make assumption violations hard to spot
- Test assumptions twice:
 - ◆ Pre-model, for each variable separately
 - ◆ Post-model, for the multivariate model variate

16

Normality

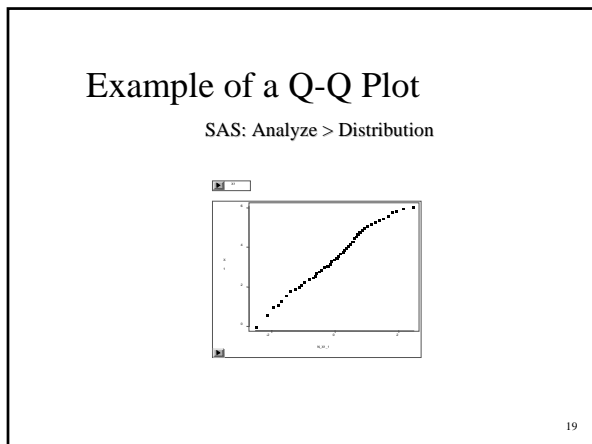
- Univariate normality for each variable separately – easily tested and corrected
- Multivariate normality for all variables simultaneously – more difficult to assess
- If MN then UN , but can be UN but not MN
- Large samples can mitigate adverse effects of violating normality assumptions

17

Graphical Analyses of Normality

- Histogram with superimposed normal curve
- Q-Q plot (or normal probability plot)
 - ◆ Consider a standard normal distribution with same number of observations as sample data
 - ◆ Draw a scatterplot of observed data values (y) versus those for expected normal values (x)
 - ◆ Do the points lie along a 45° line?

18



- ### Statistical Tests of Normality
- Rules of thumb:
 - ◆ skewness / $(\sqrt{6/N}) \sim \text{normal}$
 - ◆ kurtosis / $(\sqrt{24/N}) \sim \text{normal}$
 - Kolmogorov-Smirnov
 - Shapiro-Wilks (< 50 in sample)
 - Beware indiscriminate use in small samples (< 30) and large samples (>1000)
 - Data transformations can improve normality
- 20

- ### Homoscedasticity
- When a metric variable displays equal levels of variance across the range of another variable (metric or nonmetric)
 - More generally concerns equal variance-covariance matrices, e.g. discriminant analysis
 - Unequal matrices results in heteroscedasticity, for example:
 - ◆ larger dispersion in x_1 for larger values of x_2
 - ◆ skewness in x_2 leads to varying dispersion in x_1
- 21

- ### Assessing Homoscedasticity
- Boxplots of metric variable for each value of a nonmetric variable
 - Levene test to assess equal variances of metric variable across values of nonmetric variable
 - Box's M test to assess equality of variance-covariance matrices in discriminant analysis
 - Data transformations can also correct heteroscedasticity
- 22

- ### Linearity
- Pearson's correlation measures the *linear* association between two variables
 - Techniques based on correlations (e.g. logistic regression) can miss any nonlinear effects in the data
 - Identify nonlinear patterns in scatterplots of variables and residuals
 - Data transformations can linearize nonlinear relationships
 - Remaining nonlinear relationships can be modeled
- 23

- ### Uncorrelated Errors
- Prediction errors should not reveal systematic patterns, e.g. +, -, +, -, ...
 - Data collection process can result in correlated errors, for example:
 - ◆ Two groups analyzed together, but each group differs on a variable omitted from the model
 - ◆ Any results are biased because an unspecified cause impacts the groups differently
 - ◆ Remedy: identify the omitted factor and include in the analysis
- 24

Data Transformations

- Used to correct assumption violations
 - ◆ non-normality
 - ◆ heteroscedasticity
 - ◆ linearity
- Take the textbook guidelines with a pinch of salt, e.g. negative skew can be turned into positive skew by negating, then logarithm *or* square root can normalize
- Be careful with interpreting results after transforming original data!

25

Using Nonmetric Data

- Create dummy/indicator variables (to represent categories of nonmetric variable):
 - ◆ indicator coding to represent deviations from a comparison group (use 1's and 0's only)
 - ◆ effects coding to represent deviations from mean over all groups (use 1's, 0's, and -1's)
- Require one fewer dummy variables than there are levels of the nonmetric variable

26

Indicator Coding

- **Model:**
 $E(y) = \beta_0 + \beta_1 D_1 + \beta_2 D_2$

	D ₁	D ₂
Low:	0	0
Medium:	1	0
High:	0	1
- Estimates:
 $\beta_0 = 7, \beta_1 = -5, \beta_2 = -4$
 - Low: $\beta_0 = 7$
 - Medium: $\beta_0 + \beta_1 = 2$
 - High: $\beta_0 + \beta_2 = 3$
- β_1 and β_2 represent "deviations from the Low category"
- $\beta_1 = -5$ means we predict the response for the Medium category to be 5 units lower than the Low category ($7 - 5 = 2$)
- $\beta_2 = -4$ means we predict the response for the High category to be 4 units lower than the Low category ($7 - 4 = 3$)

27

Effects Coding

- **Model:**
 $E(y) = \eta_0 + \eta_1 D_1 + \eta_2 D_2$

	D ₁	D ₂
Low:	1	0
Medium:	0	1
High:	-1	-1
- Estimates:
 $\eta_0 = 4, \eta_1 = 3, \eta_2 = -2$
 - Low: $\eta_0 + \eta_1 = 7$
 - Medium: $\eta_0 + \eta_2 = 2$
 - High: $\eta_0 - \eta_1 - \eta_2 = 3$
 - "average": $\eta_0 = 4$
- η_1, η_2 , and $-\eta_1 - \eta_2$ represent "deviations from the average"
- $\eta_1 = 3$ means we predict the response for Low to be 3 units higher than average ($4 + 3 = 7$)
- $\eta_2 = -2$ means we predict the response for Medium to be 2 units lower than average ($4 - 2 = 2$)
- $-\eta_1 - \eta_2 = -1$ means we predict the response for High to be 1 unit lower than average ($4 - 1 = 3$)

28

Effects Coding for Jobs Example

- **Location:**
 - Eugene = 1 if Eugene, -1 if Denver, 0 otherwise
 - Portland = 1 if Portland, -1 if Denver, 0 otherwise
 - Seattle = 1 if Seattle, -1 if Denver, 0 otherwise
 - i.e. stimuli with Denver are -1 on these three dummy variables
- **Salary:**
 - 50K = 1 if 50K, -1 if 80K, 0 otherwise
 - 60K = 1 if 60K, -1 if 80K, 0 otherwise
 - 70K = 1 if 70K, -1 if 80K, 0 otherwise
 - i.e. stimuli with 80K are -1 on these three dummy variables

29