Research Design - - Topic1
Fundamentals of Statistical Inference

- Course Outline
- Approaches to Data Analysis
- Basic Definitions (Gardner & Tremblay, 2007, Ch. 1)
- Basic Arithmetic
- Basic Distributions (Kirk, 1995, Ch. 3)
- Basic Statistical Inference (Kirk, 1995, Chs. 1-2)

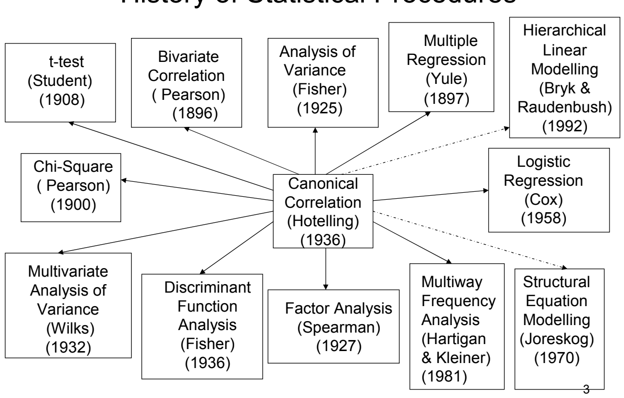
1

Approaches to Data Analysis

1. Traditional
   Sample values (statistics) are calculated and these values are used to make inferences about population values (parameters).
2. Modelling
   Population values (parameters) are estimated based on assumptions about the underlying population.
3. Overview and History of Statistical Procedures
   Knapp (1948) discussed the commonality and relationships among a number of statistical procedures. The following slide expands on that presentation.

2

History of Statistical Procedures



3

Basic Definitions

- Population: a collection of distinguishable measurements
- Parameter: a value that describes some aspect of the population
- Sample: any subset of a population singled out in some way for the purposes of study
- Statistic: a value that describes some aspect of the sample

4

Types of Statistics and Parameters:

- Location: A number that locates a sample or population in space (e.g., mean, median, mode, lowest value, 75th percentile)
- Scale: A number that indicates the variation in the values (e.g., range, interquartile range, standard deviation, variance)
- Shape: A number that describes the nature of the distribution (e.g., skewness, kurtosis)
- Association: A number that describes the relationship between two sets of values (e.g., correlation, regression)

Basic Arithmetic

- Variance of a sum:

$$S^2_{\mathbf{A+B}} = S^2_A + S^2_B + 2r_{AB}S_A S_B$$

- Standard error of the mean: standard deviation of the sampling distribution of the means

$$\sigma_\mu = \frac{\sigma}{\sqrt{n}}$$

- Biased estimates (e.g. variance) $\sigma^2 = \frac{\sum (X-\mu)^2}{N}$

$$when \quad s^2 = \frac{\sum (X-\overline{X})^2}{n} \qquad \overline{s^2} = \sigma^2 - \sigma^2_\mu$$

$$when \quad s^2 = \frac{\sum (X-\overline{X})^2}{n-1} \qquad \overline{s^2} = \sigma^2$$

Basic Distributions

$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \qquad \overline{Z} = 0 \qquad \sigma^2 = 1$$

$$t = \frac{\overline{X}-\mu}{S/\sqrt{n}} \qquad \bar{t} = 0 \qquad \sigma_t^{\,2} = \frac{df}{df-2} \qquad df > 2$$

$$\chi^2 = \frac{\Sigma(X-\overline{X})^2}{\sigma^2} \qquad \overline{\chi}^2 = df \qquad \sigma^2 = 2df$$

$$F = \frac{S_1^{\,2}}{S_2^{\,2}} \qquad \overline{F} = \frac{df_2}{df_2-2} \qquad \sigma_F^{\,2} = \frac{2df_2^{\,2}(df_1+df_2-2)}{df_1(df_2-2)^2(df_2-4)} \qquad df_2 > 4$$

Density Functions

Standard Normal Distribution $\quad y = C\left(\frac{1}{\sqrt{2\pi}}e^{-Z^2/2}\right)$

t - Distribution $\quad y = C\left(1+\frac{t^2}{df}\right)^{-\frac{df+1}{2}}$

Chi-Square Distribution $\quad y = C\left(e^{\frac{-\chi^2}{2}}\chi^{2\,\frac{df-2}{2}}\right)$

F- Distribution $\quad y = C\left(\frac{F^{\frac{df_1-2}{2}}}{(df_1F+df_2)^{\frac{df_1+df_2}{2}}}\right)$

Relations among the Distributions

	Z	t	χ^2	F
Z	*	Z = t(∞)	$Z^2 = \chi^2(1)$	$Z^2 = F(1,\infty)$
T		*	$t^2(\infty) = \chi^2(1)$	$t^2(df) = F(1,df)$
χ^2			*	$\chi^2/df = F(df, \infty)$
F				*

Basic Statistical Inference

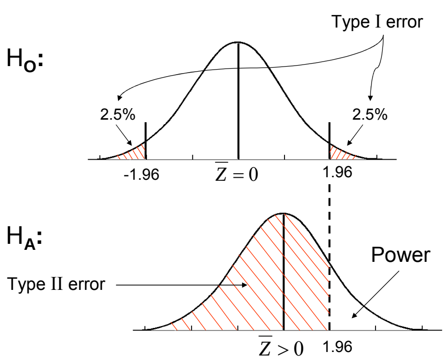
Rationale:

- Establish the Null Hypothesis (H_0)
- Determine the Alternate Hypothesis (H_A)
- Generate the Test Statistic
- Reject or Fail to Reject the Null Hypothesis (H_0)

Definitions:

- **Type I Error:** The probability of rejecting the null hypothesis when the null hypothesis is true.
- **Type II Error:** The probability of failing to reject the null hypothesis when the null hypothesis is false.
- **Power:** The probability of rejecting the null hypothesis when the null hypothesis is false.

Given: $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \dfrac{\text{statistic - parameter}}{\text{S.E.}}$



Other Standard Errors

(Applicable for large sample sizes)

Median	$\sigma_{mdn} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{\pi}{2}}$
Standard Deviation	$\sigma_S = \dfrac{\sigma}{\sqrt{2n}}$
Skewness	$\sigma_{g_1} = \sqrt{\dfrac{6}{n}}$
Kurtosis	$\sigma_{g_2} = 2\sqrt{\dfrac{6}{n}}$
Proportion	$\sigma_P = \sqrt{\dfrac{pq}{n}}$
General Test of Significance	$Z = \dfrac{\text{statistic - parameter}}{\text{S.E.}}$

References

Gardner, R.C. & Tremblay, P.F. (2007). *Essentials of Data Analysis: Statistics and Computer Applications*. London, ON: UWO Graphic Services.

Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*. (3rd ed.) Pacific Grove, CA: Brooks/Cole.

Knapp, T. (1978). Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin, 85,* 1-40.

13

A History of Major Data Analytic Procedures

Analysis of Variance
Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Bivariate Regression and Correlation
Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society, A, 186*, 343-414.

Canonical Correlation
Hotelling H. (1936) Relations between two sets of variates. *Biometrika* 28: 321-377

Chi-square
Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society, A, 195*, 1-47.

Discriminant Function Analysis
Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen. 7,* 179-88.

Factor Analysis
Spearman, C. (1927). *The abilities of man*. New York: Macmillan.

14

Hierarchical Linear Modeling
Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models: applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Logistic Regression
Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society,* Series B (Methodological), 20(2) 215-242.

Multiple Regression and Correlation
Yule, G. U. (1897). On the theory of correlation. *Journal of the Royal Statistical Society, 60,* 812-854.

Multivariate analysis of Variance
Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika, 24,* 471-94.

Multiway Frequency Analysis
Hartigan, J. A., & Kleiner, B. (1981). Mosaics for contingency tables. In W. F. Eddy (Ed.), *Proceedings of the 13th symposium on the interface between computer science and statistics* (pp. 268-273). New York: Springer-Verlag.

Structural Equation Modelling
Jöreskog (1970). A general method for analysis of covariance structures. *Biometrika, 57, 239-251*.

t-test
Student (1908). The probable error of a mean. VI. *Biometrika*, 6:1-25.

15