CSIP5403: Research Methods and Applications

Lecture 6: Statistics in Research

Shengxiang Yang[1]

[1]Institute of Artificial Intelligence
School of Computer Science and Informatics
De Montfort Universty – UK

Outline

Statistics: Descriptive and Inferential

- Collection, processing, analysis, interpretation, and presentation of numerical data belongs to the domain of statistics
  - Descriptive: Organise, summarise or describe important features of a set of data without going any further
  - Inferential: Interpret data to make generalizations which go beyond the data

Descriptive Statistics

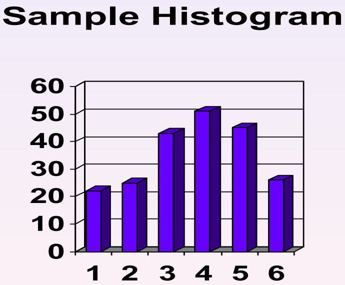
- Grouping, classifying and describing measurements and observations is a basic in statistics
- Many types of descriptive statistics
  - Frequency counts and distributions
  - Summary measures such as measures of central tendency, variability, and relationship
  - Graphical representations of the data
- A way to visualize the data and the first step in any statistical analysis

Frequency Distributions

- First step in organization of data
    - Can see how the scores are distributed
- Used with all types of data
- Illustrate relationships between variables in a cross-tabulation
- Simplify distributions with a large range by using a grouped frequency distribution (5-20)

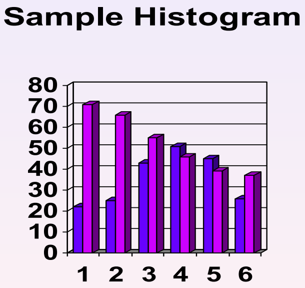
Histograms

- A bar graph can be used to graph either
  - data from discrete groups
  - data from individual scores of a continuous variable



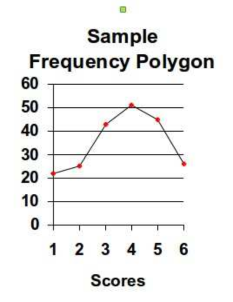
Histograms (2 Distributions)

- Possible to graph two or more distributions on the same histogram to see how they compare



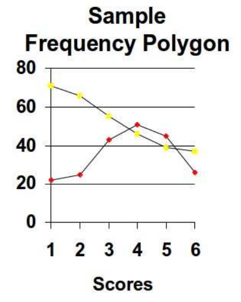
Frequency Polygon (1 Group)

- Like a histogram except that, instead of a bar representing the mean score or frequency, a dot is used, with the dots connected as shown
- Data could be either discrete or continuous



Frequency Polygon (2 Groups)

- Can compare two or more frequency polygons on the same scale as shown
- Easier to compare frequency polygons because the graph appears less cluttered than multiple histograms



Measures of Central Tendency

- One simple way to summarise the results of an experiment is to calculate the value of the typical score
- There are 3 common measures of central tendency
  - Mean
  - Median
  - Mode

Measures of Central Tendency: Mean

- Mean: the arithmetic average score
  - Used to represent data by means of a single number
  - Centre of gravity
  - Familiar to most people
  - Always exists
  - Unique
  - Take into account each individual score
  - Sensible to single extreme values
  - Used in later inferential statistics

Measures of Central Tendency: Median and Mode

- Median: the middle score in a data set when arranged in increasing or decreasing order of magnitude
  - Always exists
  - Unique
  - Unlike mean, not affected by a few deviant scores
- Mode: the most frequently occurring score
  - Can be used for qualitative data
  - May not be unique

Measures of Variation

- They will tell us something about the extent to which data are disperse, spread out or bunched
- The concept of variability is of special importance in statistical inference
- Variance: average squared distance from the mean
  - Used in later inferential statistics
- Standard Deviation: square root of variance
  - Expressed on the same scale as the mean

Formulas

- Mean: $\bar{X} = \frac{\sum X}{N}$
- Variance: $s^2 = \frac{SS}{df} = \frac{\sum (X - \bar{X})^2}{N-1}$
  - *df* — "degree of freedom"
  - Computational Formula: $SS = \sum X^2 - \frac{(\sum X)^2}{N}$
- Standard Deviation: $s = \sqrt{s^2}$

Computational Example

- Compute mean, variance, and standard deviation of the following: 2, 5, 7, 4, 6, 5
- Compute mean: $\bar{X} = \frac{\sum X}{N} = \frac{29}{6} = 4.83$
- Next, compute sum of X and sum of X^2:

$$\sum X = 29, \quad \sum X^2 = 155$$

- Compute variance, and standard deviation:

$$SS = \sum X^2 - \frac{(\sum X)^2}{N} = 155 - \frac{29^2}{6} = 14.83$$

$$s^2 = \frac{SS}{df} = \frac{14.83}{6-1} = 2.97, \quad s = \sqrt{s^2} = \sqrt{2.97} = 1.72$$

Measures of Relationship

- Pearson product-moment correlation
  - Used with interval or ratio data
- Spearman rank-order correlation
  - Used when one variable is ordinal and the second is at least ordinal
- Scatter plots
  - Visual representation of a correlation
  - Helps to identify possible nonlinear relationships

Correlations: Formulas

- Pearson product-moment correlation

$$r_{xy} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{[\sum X^2 - \frac{(\sum X)^2}{N}][\sum Y^2 - \frac{(\sum Y)^2}{N}]}}$$

- Spearman rank-order correlation

$$r_s = 1 - \frac{6 \sum d_{XY}^2}{N(N^2 - 1)}$$

where d_{XY} is the difference in ranks between X and Y

Computational Example

- Compute the product-moment correlation for the data below

	X	Y	X^2	Y^2	XY
	4	5	16	25	20
	3	5	9	25	15
	5	6	25	36	30
	2	4	4	16	8
	6	7	36	49	42
	5	7	25	49	35
$\sum$	25	34	115	200	150

- Compute sums, sums of squares, and cross products as shown
- Calculate the product-moment correlation

$$
\begin{aligned}
r_{xy} &= \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}} \\
&= \frac{150 - \frac{(25)(34)}{6}}{\sqrt{\left[115 - \frac{(25)^2}{6}\right]\left[200 - \frac{(34)^2}{6}\right]}} = \frac{8.33}{\sqrt{\left[10.83\right]\left[7.33\right]}} = 0.93
\end{aligned}
$$

Computational Example

- Compute a rank-order correlation for the ordinal data below

X	Y	d_{XY}	d^2_{XY}
1	3	-2	4
2	2	0	0
3	1	2	4
4	5	-1	1
5	4	1	1
			10

- Compute the difference in ranks (d_{XY}) and d^2_{XY} as shown
- Calculate the rank-order correlation

$$r_s = 1 - \frac{6 \sum d^2_{XY}}{N(N^2 - 1)} = 1 - \frac{6(10)}{5(5^2 - 1)} = 1 - \frac{60}{120} = 0.5$$

Regression

- Using a correlation (relationship between variables) to predict one variable from knowing the score on the other variable
- Usually a linear regression (finding the best fitting straight line for the data)
- Best illustrated in a scatter plot with the regression line also plotted
- Correlation coefficients close to $+1$ or -1 linear regression

Regression

Go to Excel for Exercise ......

Inferential Statistics

- Used to draw inferences about populations on the basis of samples from the populations
- The "statistical tests" that we perform on our data are inferential statistics
- Provide an objective way of quantifying the strength of the evidence for our hypothesis

Populations and Samples

- Population: the larger group of all subjects of interest to the researcher
- Sample: a subset of the population
- Samples almost never represent populations perfectly (termed "sampling error")
  - Not really an error; just the natural variability that you can expect from one sample to another
- Population parameters and sample statistics
  - Population parameter is a descriptive statistic computed from everyone in the population
  - Sample statistics is a descriptive statistic computed from everyone in your sample

Hypothesis Testing

- So far all problems of estimation
- Sample mean estimates Population mean
- Is population mean equal to sample mean?
- In this case, we must test a hypothesis
  - There is NO difference between the population mean and the sample mean

The Null Hypothesis

- Example: Suppose two populations
- We want to test if populations means are equal
- Null Hypothesis: States that there is NO difference between the population means
- We have to test the Null Hypothesis:
    - Compare sample means to test the null hypothesis

Statistical Decisions

- We can either Reject or Fail to Reject the null hypothesis
  - Rejecting the null hypothesis suggests that there is a difference in the populations sampled
  - Failing to reject suggests that no difference exists
  - Decision is based on probability (reject if it is unlikely that the null hypothesis is true)
    - Alpha (α): the statistical decision criteria used
    - Traditionally α is set to small values (.05 or .01)
- Always a chance for error in our decision

Statistical Decision Process

	Reject Null Hypothesis	**Retain Null Hypothesis**
Null Hypothesis is True	Type I Error	Correct Decision
Null Hypothesis is False	Correct Decision	Type II Error

Test Concerning Mean

- Null Hypothesis H_0
- Alternative Hypothesis H_1
  - Standard deviation known
    - Normal distribution
  - Standard deviation not known
    - Normal distribution if sample size $>=$ 30
    - Student distribution (t-test) if sample size $<$ 30

Testing Differences Between Means

- Simple t-test: tests mean difference of two independent groups
- Correlated t-test: tests mean difference of two correlated groups
- Analysis of Variance: tests mean differences in two or more groups

The t-test

- For independent samples

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left[\frac{SS_1 + SS_2}{N_1 + N_2 - 2}\right]\left[\frac{1}{N_1} + \frac{1}{N_2}\right]}}$$

- For correlated samples

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} - 2r\left(\frac{s_1}{\sqrt{N_1}}\right)\left(\frac{s_2}{\sqrt{N_2}}\right)}}$$

where r is the Correlation Coefficient between X_1 and X_2

Computational Example: Independent t-test

- Calculate mean and SS:

	Group 1	Group 2
	5	3
	8	5
	7	2
	8	3
	7	
Mean	7.00	3.25
SS	6.00	4.75
N	5	4

- Calculate t-test value:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left[\frac{SS_1 + SS_2}{N_1 + N_2 - 2}\right]\left[\frac{1}{N_1} + \frac{1}{N_2}\right]}} = \frac{(7.00 - 3.25)}{\sqrt{\left[\frac{6.00+4.75}{5+4-2}\right]\left[\frac{1}{5} + \frac{1}{4}\right]}} = \frac{3.75}{0.83} = 4.52$$

- Make decision:

  $t = 4.52 > t_{crit} = 2.365 \Rightarrow$ Reject Null Hypothesis

Computational Example: Correlated t-test

- Statistical data:

	Before	After
Mean	3.17	2.96
Standard Deviation	0.893	0.975
N=91	$r = 0.84$	

- Calculate t-test value:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} - 2r\left(\frac{s_1}{\sqrt{N_1}}\right)\left(\frac{s_2}{\sqrt{N_2}}\right)}} = \frac{(3.17 - 2.96)}{\sqrt{\frac{0.893^2}{91} + \frac{0.975^2}{91} - 2(0.84)\left(\frac{0.893}{\sqrt{91}}\right)\left(\frac{0.975}{\sqrt{91}}\right)}} = \frac{0.21}{0.056} = 3.75$$

- Make decision:
  $t = 3.75 > t_{crit} = 1.99 \Rightarrow$ Reject Null Hypothesis

Power of a Statistical Test

- Sensitivity of the procedure to detect real differences between the populations
- Not just a function of the statistical test, but also a function of the precision of the research design and execution
- Increasing the sample size increases the power because larger samples estimate the population parameters more precisely

Summary

- Statistics allow us to detect and evaluate group differences that are small compared to individual differences
- Descriptive vs. inferential statistics
  - Descriptive statistics describe the data
  - Inferential statistics are used to draw inferences about population parameters on the basis of sample statistics
- Statistics help objective evaluation, but do not guarantee correct decision every time