Introduction to Statistics
Lecture 23

Tuesday, November 29, 2005

- I will have your projects graded by the end of our lectures.
- We will start today’s lecture with some review and practice.

Practice: finding t critical values, n=8

- Find t* for 90% confidence interval
- For 5% significance level (as one-tail probability) find t*.
  \[ H_0 : \mu = 10 \]
  \[ H_a : \mu < 10, \text{ reject } H_0 \text{ if } t < -t^* \]
  \[ \text{or } H_a : \mu > 10, \text{ reject } H_0 \text{ if } t > t^* \]
- What about two-sided alternative.
  \[ \text{Reject } H_0 \text{ if } |t| > t^* \]
  \[ t^* \text{ is for significance level as two-tailed probability} \]

Try a Test

\[ n_1 = 8, \bar{X}_1 = 78.5, S_1 = 10 \]
\[ n_2 = 8, \bar{X}_2 = 83.5, S_2 = 10 \]

\[ H_0 : \mu_1 = \mu_2 ? \]

The actual data

<table>
<thead>
<tr>
<th>Students</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67</td>
<td>72</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>74</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>77</td>
</tr>
<tr>
<td>4</td>
<td>74</td>
<td>79</td>
</tr>
<tr>
<td>5</td>
<td>77</td>
<td>82</td>
</tr>
<tr>
<td>6</td>
<td>85</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>91</td>
<td>96</td>
</tr>
<tr>
<td>8</td>
<td>93</td>
<td>98</td>
</tr>
</tbody>
</table>

What’s wrong here?

- What’s the problem here?
- Let’s go back to the assumptions:
  - Two populations?
  - Independent samples?

Paired samples

- Two sets of observations are in the form of one set of pairs.
  - Twin studies
  - Matched-pair experiments
  - Before-and-after studies
- It is the difference within each pair that we are interested in.
  - Whether the IQ of the twins will be identical?
  - Whether in one household, the wife makes (on average) the same amount as the husband?
  - Whether today’s practice problems are helpful?!
Paired Data

<table>
<thead>
<tr>
<th>Observations</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67</td>
<td>72</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>71</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>73</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>74</td>
<td>69</td>
<td>-5</td>
</tr>
<tr>
<td>n</td>
<td>93</td>
<td>91</td>
<td>-2</td>
</tr>
</tbody>
</table>

So, it is really a one-sample problem!!

Inference on within-pair difference

- Population model
  - $d$: difference
  - Population Mean: $\mu_d$, Standard deviation: $\sigma_d$
- Sampling distribution model
  - $d$-bar is the sample statistic that estimates $\mu_d$
  - $d$-bar's sampling distribution model is
    \[
    \frac{d - \mu_d}{\sigma_d/\sqrt{n}} 
    \]

Inference procedures are the same as the inference about means!

- A paired t interval
  \[
  \bar{d} \pm t^* \frac{s}{\sqrt{n}}
  \]
- A paired t test
  \[
  t = \frac{\bar{d} - \mu_d0}{s/\sqrt{n}}
  \]

Review on categorical variables

- Categorical variable: the value of a categorical variable put an individual into one of several categories.
  - Example:
    - Gender: female, male
    - Marital status: single, married, widowed, divorced
    - Employment status: unemployed, employed
  - If you want to study political science, sociology, psychology, you should consider taking a course in ‘categorical data analysis’
  - Displaying the distribution of a categorical variable: bar graph and pie chart.

Review:
Table for categorical variables

- Frequency table
- Two-way table: study the distribution of the two categorical variables simultaneously.
Model and Sampling Variation of a table

- Flip a coin—Outcomes: head or tail
  - Probability model: 50% chance getting a head and 50% tail
  - Data: samples of random flips
    - Sample 1: 12 out of 20 flips are heads
    - Sample 2: 9 out of 20 flips are heads
- Why the samples do not reflect the 50-50 model perfectly?
- Question: if data is just a random sample from the population (where the model is claimed to apply), how do we know whether the data is consistent with the model?
- Definition of the expected values: expected pattern in the data that agrees with the model.
- We then can compare the model (represented by the expected values) with the data (the observed values).

Goodness-of-fit test for frequency table of one categorical variable

- Define $r$ as the number of possible outcome values of the categorical variable.
- The frequency table has $r$ cells (except for the TOTAL cell).
- Probability model decides the probability values for each outcome value (each cell).
- For each cell, calculate the expected value = (total # obs.) $\times$ (model prob. of the cell).
- For each cell, there is an observed value (count).
- The differences between the observed values and expected values of individual cells can be used as evidence against the model.

Goodness-of-fit test (cont’d)

- Null hypothesis: a probability model for the categorical variable
- Test statistic: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
- Statisticians already identified the distribution of $\chi^2$ as a chi-square distribution with $(r-1)$ d.f.
- P-value is defined as observing $\chi^2$ equal to or greater than the observed under $H_0$.

Relation between two categorical variables

- Say we are looking at variable X and variable Y, which are both categorical.
  - Let X be the column variable, Y the row variable.
- Association: some given values of X occur more frequently with some given values of Y.
- One can compare one column with another—conditional distributions.

Independence test in a two-way table

- Looking for evidence on association?
  - the null hypothesis should be... independence!
- Chi-square test for independence in two-way table
  - the null hypothesis is a special pattern in a two-way table
- What (pattern) does the null hypothesis imply?
  - No association: Given any event A decided by the values of X, and any event B decided by the values of Y, A and B are independent, $P(A)$ $\times$ $P(B)$.
  - If there is no association: conditional distribution of Y given X should be the same as the marginal distribution of Y.
  - Model probability of each cell $\times$ product of the marginal relative frequencies of the outcome values of X and Y that define this cell.
Chi-square test for independence

- Define \( r \) as the number of possible values of the row variable.
- Define \( c \) as the number of possible values of the column variable.
- Totally the two-way table has \( r \times c \) cells.
- A "independence" Model decides the probability (pattern) for each cell.
- For each cell, calculate the expected value \( \text{Expected} = \frac{\text{row total}}{} \times \frac{\text{column total}}{} \times \frac{\text{table total}}{} \).
- For each cell, there is an observed value.
- The differences between the observed values and expected values of individual cells can be used as evidence against the model.

Chi-square test (continued)

- Test statistic
  \[ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \]
- Statisticians already identified the distribution of \( \chi^2 \) as a chi-square distribution with \((r-1)(c-1)\) d.f.
- P-value is defined as observing \( \chi^2 \) equal to or greater than the observed under \( H_0 \).

Chi-square test for homogeneity

- Not required.

Practice: Chi-square Test

<table>
<thead>
<tr>
<th>Observed versus</th>
<th>Color of M&amp;M</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>Red</td>
<td>Yellow</td>
<td>Green</td>
<td>Total</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>20</td>
<td>17.5</td>
<td>10</td>
<td>12.5</td>
</tr>
<tr>
<td>Female</td>
<td>15</td>
<td>17.5</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>20</td>
<td>25</td>
<td>17.5</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 1.714 \]
\[ = (20 - 17.5)^2 + (15 - 17.5)^2 + (10 - 10)^2 + (10 - 10)^2 + (10 - 12.5)^2 + (15 - 12.5)^2 \]

Reading

- Chapters 25, 26