THE CHI-SQUARE STATISTIC

Comparing observed and expected frequencies:

Meaningful v. meaningless differences

EXAMPLE PROBLEM: LIKERT-SCALE QUESTIONS

- Did the instructor dress appropriately for class?
  - (a) always
  - (b) most of the time
  - (c) seldom
  - (d) never

IS THE DATA VALID OR RANDOM?

- 100 students responds:
  - (a) always 40
  - (b) most of the time 30
  - (c) seldom 20
  - (d) never 10

THE CHI-SQUARE FORMULAS

- N = the number of data values (students responding) = 100
- k = the number of categories (choices) = 4
- f_e = the expected frequency if the data are random = 100 / 4 = 25
- f_o = the observed frequency = 40, 30, 20, and 10

$$f_e = \frac{N}{k}$$

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_e: EXPECTED FREQUENCY

$$f_e = \frac{N}{k} = \frac{100}{4} = 25$$

- C++ can't replicate subscripts
- Write subscripts with "regular" text

```
int k;
...
int N;
...
double fe = (double)N / k;
```

χ^2 AND FREQUENCY SIGNIFICANCE

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{1}{f_e} \sum (f_o - f_e)^2$$

$$\chi^2 = \frac{1}{25}\left[(40-25)^2 + (30-25)^2 + (20-25)^2 + (10-25)^2\right]$$

PROGRAMMING χ^2

```
double    sum = 0;
for (int i = 0; i < k; i++)
    sum += pow(fo[i] - fe, 2);

double chi2 = sum / fe;
```

- The Σ operator combines two behaviors
  - It loops over the elements of the observed frequencies
  - It adds a sequence of terms
- Converting Σ to C++ requires two features
  - A for-loop to iterate and index
  - += to sum the formula terms

CHI-SQUARE SOLUTION: TWO IMPLEMENTATIONS

- Single function solution
  - Focus on data input and output
  - Converting formula, especially $\sum$, to C++
- Client-supplier solution
  - Separating input and output from computation is good program design
  - A supplier defines and supplies reusable services
  - The client is an application program that uses the supplier's services