Comments for MEDB 5501, Week 10, part 2

Chi-square tests

- Variants: Chi squared, chisquare? χ^2, X^2
  - Goodness of fit
  - Independence
  - Variance (not covered today)
  - Other uses (also not covered today)

Speaker notes

The chi-square test is a very common test used in a variety of settings. You may see it with or without a dash, and some researchers will put a “d” at the end (Chi squared). Sometimes the C is capitalized and sometimes not. Sometimes you will see the Greek letter chi or a capital X.

The chi-square test is useful for a broad class of tests, known as goodness of fit. It is a also useful test of independence between two categorical variables. There’s a third test, comparing a variance or standard deviation to a fixed quantity. This third test will not be covered today.

There are other uses of the chi-square test, I can’t recall any simple uses, but the test appears all over the place.

Formula

- $\sum \frac{(O-E)^2}{E}$
  - O = Observed, E= Expected
- Cell contribution to Ch-square: $\frac{(O-E)^2}{E}$
- Standardized Residual: $\frac{O-E}{\sqrt{E}}$

Speaker notes

The general formula for the chi-square test that works both for the goodness of fit and independence tests is the sum of the observed minus the expected squared divided by the expected. You'll see precise formulas for Observed and Expected in a bit.

Sometimes software will show the individual components to the sum. This is the cell contribution to chi-square. You might also see the standardized residual, which is Observed - Expected divided by the square root of Expected. Either of these quantities will identify important deviations from goodness of fit or independence.

Chi-square goodness of fit test, 1 of 2

- Single categorical variable,
  - n_1 is frequency of first category,
  - n_2 is frequency of second category,
  - $\ldots$
  - n_k is frequency of last category.
  - $N = n_1 + n_2 + \ldots + n_k$
- Most often used for k > 2
  - Works for k = 2, but simpler test is available.

Speaker notes

The goodness of fit test uses a single categorical variable with k levels. You count the frequencies of each level, $n_1, n_2, \ldots, n_k$. The sum of these values, N, is the sample size.

Chi-square goodness of fit test, 2 of 2

- Are all categories equally likely?
  - $H_0: \ \pi_1 = \pi_2 = \ldots = \pi_k$
  - $H_1: \ \pi_i \neq \ \pi_j$, for some i,j
    - π_i is population proportion for category i.
- $O_i = n_i$
- $E_i = N/k$
- $T = \Sigma \frac{(O_i - E_i)^2}{E_i}$
- Reject H0 if T > $\chi^2(0.05, k-1)$
  - Only reject for large positive values

Speaker notes

The chi-square goodness of fit test answers the question, are all the categories equally likely. In mathematical notation, you are testing the hypothesis that all the pi_i values are equal where pi_i represents the hypothetical probability of each category, if you were able to measure the categorical variable in the entire population. Many researchers will call the π_i values population probabilities instead of population proportions, but the concept is the same.

The observed values are the counts for each category. The expected counts distribute the N values equally across all categories.

Example, clinic recruitment, 1 of 2

```
Clinic                  A    B    C    D    E Total
Patients recruited 17   29   37   15   27    125
```

Arrange the data as follows for importing into SPSS

```
"clinic","patients"
"A",17
"B",29
"C",37
"D",15
"E",27
```

Speaker notes

In a hypothetical example, five clinics of roughly equal sizes participated in a clinical trial. The number of participants recruited at each site are listed here. Is the probability of getting patients from each clinic the same?

Example, clinic recruitment, 2 of 2

Clinic	A	B	C	D	E	
Observed	17	29	37	15	27	
Expected	25	25	25	25	25	
O-E	-8	4	12	-10	2	
(O-E)/sqrt E	-1.6	0.8	2.4	-2.0	0.4	
(O-E)^2/E	2.56	0.64	5.76	4.00	0.16	Sum=13.12

Speaker notes

With 125 total observations, you would expect 25 in each of the 5 categories if the probabilities were the same.

Chi-square test of independence

- Two events are independent if
  - $P[A \cap B] = P[A] \times P[B]$
- Two categorical variables are independent if
  - $P[A = i \cap B = j] = P[A = i] \times P[B = j]$

Passenger class and mortality counts

		Survived		
		No	Yes	Total
PClass	1st	129	193	322
	2nd	161	119	280
	3rd	573	138	711
Total		863	450	1,313

Passenger class and mortality probabilities

		Survived		
		No	Yes	Total
PClass	1st	9.8%	14.7%	24.5%
	2nd	12.3%	9.1%	21.3%
	3rd	43.6%	10.5%	54.2%
Total		65.7%	34.3%	100.0%

Passenger class and mortality expected counts

		Survived No	Survived Yes	Total
PClass	1st	211.6	110.4	322.0
	2nd	184.0	96.0	280.0
	3rd	467.3	243.7	711.0
Total		863.0	450.0	1,313.0

Passenger class and mortality standardized residuals

```
              Survived
              No   Yes
PClass  1st  -5.7  7.9
        2nd  -1.7  2.4
        3rd   4.9 -6.8

T = 172.3, p-value < 0.001
```

```
```