Ch13: Categorical Data

17 Nov 2011
BUSI275
Dr. Sean Ho

- **HW8 due Tue**
- **Please download: 20-ChiSq.xls**

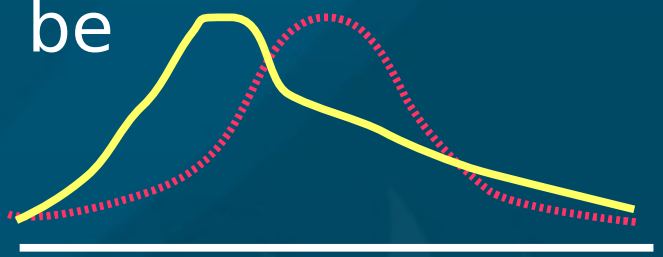
TRINITY WESTERN UNIVERSITY

Outline for today

- χ^2 (chi-squared) test of goodness-of-fit
  - Observed vs. Expected
- χ^2 on a single nominal variable
  - Test against uniform distribution
    - CHIDIST(), CHITEST()
  - Test of normality
- χ^2 on 2-way contingency tables
  - Test for independence
  - Marginal probabilities
  - Limitations of χ^2 test for independence

Goodness of fit

- The χ^2 (chi-squared) test is one way to assess goodness of fit:
  - How well an observed distribution fits a hypothesized distribution
  - Hypothesized distribution can be uniform, normal, etc.



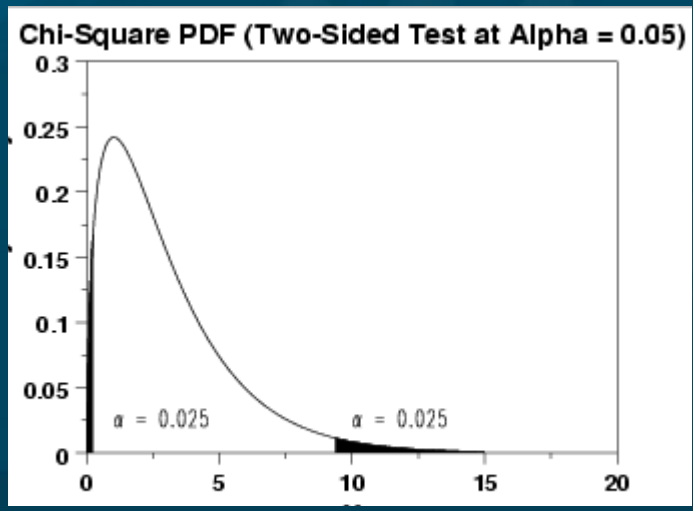
- χ^2 can also be applied to test if two nominal variables are independent
  - Compare pivot table (contingency table) with hypothesized results if vars independent
  - Analogous to correlation for continuous vars

χ^2 vs. uniform distribution

- e.g., are technical support calls evenly distributed across the weekdays?
  - H_0: evenly distributed, matches uniform dist.
- Expected # calls per day (uniform distribution):
  - Total observed calls (1300), divided by 5

Observed	**290**	**250**	**238**	**257**	**265**
Expected	**260**	**260**	**260**	**260**	**260**

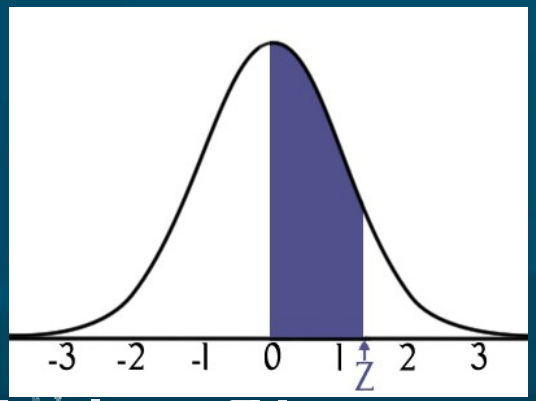
- Test statistic: $$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$
- Use CHIDIST(χ^2, *#cells* – 1)
  - Or CHITEST(*obs*, *exp*)



itl.nist.gov

χ^2 vs. normal distribution

- e.g., are student test scores normally distrib?
  - Other normality tests: Shapiro-Wilk, K-S
- Count frequency of test scores by bins
- How to find expected frequencies?
  - Find mean, SD of the data
  - Use NORMDIST() to find percentage of the data that would lie within each bin on the ideal normal:
    - NORMDIST(80, μ, σ, 1) – NORMDIST(75, μ, σ, 1)



Bin	Freq	Norm Freq
40		
45	1	1.30
50	4	2.77
55	15	5.14
60	7	8.28
65	5	11.60
70	7	14.11
75	11	14.93
80	15	13.72
85	22	10.97
90	9	7.62
95	4	4.60
100	0	2.41

Contingency tables

- Joint freq. distribs for multiple nominal variables
  - Each cell of the table holds the # (frequency) of observations that match that combo
  - Pivot tables, with Count in the Data field
- E.g., Handedness vs. Gender
  - H_0: handedness is independent of gender: the probability of being left-handed stays the same, regardless of the gender
    - P(left | M) = P(left | F) = P(left)

Observed	**Left**	**Right**
Male	17	163
Female	33	167

χ^2 on 2-way contingency

- Expected values assume independence
- Calculate marginal probabilities:
  - P(female) = 200/380 ≈ 52.6%
  - P(left) = 50/380 ≈ 13.2%
- Assuming independence,
  - P(F ∩ L) = P(F) * P(L) = (.526)(.132)
- Thus the expected count for (F ∩ L) is
  - P(F) * P(L) * (total) = (.526)(.132)(380)
- Calculate χ^2 summed over all cells
  - df = (#rows – 1) (#cols – 1)
    - = 1 in this case!

	L	R	Tot
M	17	163	180
F	33	167	200
Tot	50	330	380

Summary on χ^2

- Test of goodness-of-fit: observed vs. expected
- May apply to a single nominal variable:
  - Expected distrib. may be uniform, normal, ...
- May apply to two nominal variables:
  - Expected distrib. is that vars are independent
  - Akin to correlation on continuous variables
    - Large χ^2 ↔ $|r| \approx 1$
- But only an approximation to the true distrib:
  - Results may be invalid if cell counts are <5
  - May need to combine levels of a var

TODO

- HW8 (ch15,12): due Tues
- Projects:
  - Presentations in two weeks!
  - If you don't know what analysis to perform, or how to perform it, ask me for help