Categorical Data / Contingency Table Analysis

• A loose classification of statistical methods:

• A contingency table is a convenient way to summarize data on one or more categorical variables.

One-way Tables

• The simplest contingency table is a 1×2 table.

• This summarizes data on one variable that classifies observations into two categories.

Example: A sample of 50 driver's license applicants are classified according to success ($Y = 1$) or failure ($Y = 0$) on the driver's exam.

Data:

Model:

<u>Note</u>: Because observations are independent,

• Of interest: Estimating or testing about the parameter π = probability of "success" for any random observation.

• Also, π = the proportion of "successes" in the population.

• The least-squares estimator of π is:

<u>Proof</u>:

Note:

• Clearly, $\hat{\pi}$ is a sample ____________, so if *n* is large, then $\hat{\pi}$ is approximately

E(*T*) = and var(*T*) =

so

Inference about π

Note:

Since $\hat{\pi}$ is a consistent estimator of π, by Slutsky's theorem,

Hence

is a $100(1 - \alpha)\%$ (Wald) CI for π (for large *n*).

z-test about π

- **Consider testing H_0: $\pi = \pi_0$ where π_0 is some specified number between 0 and 1.**

- **If H_0 is true,**

So z^* is our test statistic:

<u>Rule of thumb</u>: The large-sample methods are appropriate if:

<u>R example</u> (Driver's exam data):

• The 95% "score" CI consists of all values π_0 that are <u>not</u> rejected (at $\alpha = 0.05$) using the z-test of H_0: $\pi = \pi_0$ vs. H_a: $\pi \neq \pi_0$.

• Do we have evidence that the proportion passing among all those in the population is greater than 0.6?

• If our sample is small, we can use nonparametric inference about π: the <u>binomial test / CI</u>.

• The p-value is obtained by adding the exact probabilities, from the Binom(n, π_0) distribution, of observing data at least as favorable to H_a as the data we did observe.

<u>R Example</u> (diseased tree data):

Analysis of 1 × *c* Tables

• Now suppose the categorical variable we observe has *c* possible categories.

• For $i = 1, \ldots, c$ and $j = 1, \ldots, n$,

Then

represent the observed counts for each category.

• If the observations are independent, the vector

where π_i = the probability a random observation falls in category *i*, for $i = 1, \ldots, c$.

• Note only $c - 1$ of these probabilities must be estimated, since

χ^2 goodness-of-fit test

• This tests whether the category probabilities are equal to some specified values

• Under H_0, we would expect observations to fall in category *i* ($i = 1, \ldots, c$).

• Let denote the *i*-th "expected cell count".

• Let denote the *i*-th "observed cell count".

When *n* is large, under H_0,

has a

• Large discrepancies between Obs_i and Exp_i are evidence ____ H_0 and lead to __________ values of

• Therefore we reject H_0 when

<u>Example</u>: It is believed that the blood types of students in a college are distributed as: 45% = type O, 40% = type A, 10% = type B, 5% = type AB. A random sample of 1000 students revealed the sample counts:

<u>Blood Type</u>

O	A	B	AB	Total
465	394	96	45	1000

Test:

Expected counts:

χ^{2*} =

<u>Rule of thumb</u>: *n* is large enough for the χ^2 test to be valid if <u>all expected cell counts</u> are at least 5.

• The χ^2 test can be used as a general goodness-of-fit test for any discrete (or even continuous) distribution.

• We must calculate the expected counts for each category based on the distribution in question.

• If any parameter values are estimated from the sample data rather than being specified by the null hypothesis, then we subtract one d.f. from the χ^2 distribution for each such estimated parameter.