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 $1 \times 2$ 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:**
Note: Because observations are independent,

- Of interest: Estimating or testing about the parameter $\pi = \text{probability of “success” for any random observation}$.

- Also, $\pi = \text{the proportion of “successes” in the population}$.

- The least-squares estimator of $\pi$ is:

Proof:

Note:

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

$E(T) = \quad \text{and } \text{var}(T) =$

so
Inference about \( \pi \)

Note:

Since \( \hat{\pi} \) is a consistent estimator of \( \pi \), by Slutsky’s theorem,

Hence

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

\[ z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \]

\( z \)-test about \( \pi \)

• Consider testing \( H_0: \pi = \pi_0 \) where \( \pi_0 \) is some specified number between 0 and 1.

• If \( H_0 \) is true,

So \( z^* \) is our test statistic:
Rule of thumb: The large-sample methods are appropriate if:

R example (Driver’s exam data):

• The 95% “score” CI consists of all values $\pi_0$ that are not 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 $\pi$: the binomial test / CI.

• The p-value is obtained by adding the exact probabilities, from the Binom$(n, \pi_0)$ distribution, of observing data at least as favorable to $H_a$ as the data we did observe.

R Example (diseased tree data):
Analysis of $1 \times c$ 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 $\pi_i = \text{the probability a random observation falls in category } i$, for $i = 1, \ldots, c$.

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

$\chi^2$ 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

Example: 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:

<table>
<thead>
<tr>
<th>Blood Type</th>
<th>O</th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>465</td>
<td>394</td>
<td>96</td>
<td>45</td>
<td>1000</td>
</tr>
</tbody>
</table>

Test:

Expected counts:
Rule of thumb: $n$ is large enough for the $\chi^2$ test to be valid if all expected cell counts are at least 5.

- The $\chi^2$ 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 $\chi^2$ distribution for each such estimated parameter.