Chapter 9 Notes, Part 1 - Inference for Proportion and Count Data

We want to estimate the proportion $p$ of a population that have a specific attribute, like “what percent of houses in Cambridge have a mouse in the house?”

We are given $X_1, \ldots, X_p$ where $X_i$’s are Bernoulli, and $P(X_i = 1) = p$.

$X_i$ is 1 if house $i$ has a mouse.

Let $Y = \sum_i X_i$ so $Y \sim Bin(n, p)$.

An estimator for $p$ is:

$$\hat{p} = \frac{Y}{n} = \frac{1}{n} \sum_i X_i.$$  

$\hat{p}$ is a random variable. For large $n$ (rule of thumb, $n\hat{p} \geq 10, n(1 - \hat{p}) \geq 10$) the CLT says that approximately:

$$\hat{p} \sim N\left(p, \frac{pq}{n}\right) \text{ where } q = 1 - p.$$  

Questions: What’s up with that rule of thumb? Where did the $pq/n$ come from?

---

Confidence Intervals

The CI can be computed in 2 ways (here for 2-sided case):

- CI first try:

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{pq/n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$  

We could solve it for $p$ but the expression is quite large...

- CI second try:

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\hat{p}\hat{q}/n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$  

which yields

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$  

So that’s the approximate CI for $p$.

---

Sample size calculation for CI
Want a CI of width $2E$:

$$\hat{p} - E \leq p \leq \hat{p} + E$$

so

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

which means

$$n = \left( \frac{z_{\alpha/2}}{E} \right)^2 \frac{1}{4} \hat{p}\hat{q}.$$

**Question:** We’ll take $\hat{p}\hat{q}$ to be its largest possible value, $(1/2) \times (1/2)$. Why do we do this? Why don’t we just use the $\hat{p}$ and $\hat{q}$ that we measure from the data?

So, we need:

$$n = \left( \frac{z_{\alpha/2}}{E} \right)^2 \frac{1}{4} \text{ observations.}$$

**Hypothesis testing on proportion for large $n$**

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0.$$  

Large $n$ and $H_0$ imply $\hat{p} \approx N(p_0, \frac{p_0q_0}{n})$ (where $q_0 = 1 - p_0$) so we use z-test with test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/n}}$$

**Example**

For small $n$ one can use the binomial distribution to compute probabilities directly (rather than approximating by normal). (Not covered here.)

**Chapter 9.2 Comparing 2 proportions**

**Example:** The Salk polio vaccine trial: compare rate of polio in control and treatment (vaccinated) group. Is this independent samples design or matched pairs?

Sample 1: number of successes $X \sim Bin(n_1, p_1)$, observe $X = x$.

Sample 2: number of successes $Y \sim Bin(n_2, p_2)$, observe $Y = y$.

We could compare rates in several ways:

$$p_1 - p_2 \quad \rightarrow \quad \text{we’ll use this one}$$

$$\frac{p_1}{p_2} \quad \text{“relative risk”}$$

$$\left( \frac{p_1}{1-p_1} \right) / \left( \frac{p_2}{1-p_2} \right) \quad \text{“odds ratio”}$$
For large samples, we’ll use the CLT:

\[
Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} \approx N(0, 1)
\]

where \( \hat{p}_1 = X/n_1 \) and \( \hat{p}_2 = Y/n_2 \).

To test

\[ H_0 : p_1 - p_2 = \delta_0, \]
\[ H_1 : p_1 - p_2 = \delta_0, \]

we can just compute z-scores, p-values, and CI. The test statistic is:

\[
z = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}
\]

This is a little weird because it really should have terms like “\( p_{1,0} q_{1,0}/n_1 \)” in the denominator, but we don’t have those values under the null hypothesis. So we get an approximation by using \( \hat{p}_1 \) and \( \hat{q}_1 \) in the denominator.

Example

For an independent samples design with small samples, use Fisher’s Exact Test which uses the Hypergeometric distribution. For a matched pairs design, use McNemar’s Test which uses the binomial distribution (both beyond the scope.)

Two Challenges