AMS 572 Data Analysis I
Inference on one sample proportion

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

Review

- $X_1, \ldots, X_n \sim$ Bernoulli(p)
- $Y = \sum X_i \sim B(n, p)$
- Four key conditions
  1. Binary response (0/1)
  2. Observed a known number of times n
  3. Success probability (p) same
  4. Independence between trials
- MGF of binomial $M(t) = (pe^t + 1 - p)^n$
- If $X \sim B(n_1, p)$ and $Y \sim B(n_2, p)$, X and Y are independent. Then $X + Y \sim$
- If $X_1, \ldots, X_n \sim$ Bernoulli(p), by CLT

Inference on the population proportion p

Setting:

- Suppose we observe $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$
- The statistic $Y = \sum X_i \sim B(n, p)$ is the count of successes
- A point estimator for p is

$$\hat{p} =$$

Inference on the population proportion p

- Hypothesis testing

$$H_0 : p = p_0 \text{ vs } H_a : p \neq p_0$$

- Under the null $H_0 : Y \sim \text{Binomial}(n, p_0)$
- Need to find $y_{\alpha/2}$ and $y_{1-\alpha/2}$ such that

Exact Test for Binomial Proportion

- For small samples, compute exact rejection region using

Exact Test for Binomial Proportion

Alternatively, one may also compute the p-values for the exact tests.

1. One tailed: $H_0 : p = p_0$ vs $H_1 : p > p_0$
   p-value $= P(Y \geq y|H_0) = \sum_{i=y}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}$
2. One tailed: $H_0 : p = p_0$ vs $H_1 : p < p_0$
   p-value $= P(Y \leq y|H_0) = \sum_{i=0}^{y} \binom{n}{i} p_0^i (1-p_0)^{n-i}$
3. Two tailed: $H_0 : p = p_0$ vs $H_1 : p \neq p_0$
   p-value $= 2\min(P(Y \leq y|H_0)P(Y \geq y|H_0))$
   (NOTE: This is the formula from your textbook. A pitfall of this approach is that the value can exceed 1.)

Alternative approach: (a) Mid p-value, (b) Sum all probabilities $\leq P(Y = y|H_0)$. In R, use `binom.test()`

Example: Suppose it is known that the 1-year death rate for a particular form of cancer is 30%. A new therapy designed to decrease death rate is to be tried on 15 patients

$$H_0 : p = 0.3 \text{ vs } H_a : p < 0.3$$

R Code

```
> pbinom(0:15,15,0.3)
 [1] 0.004747562 0.035267600 0.126827715 0.296867928 0.515491059
 [6] 0.721621440 0.868857427 0.949987460 0.984757474 0.996347479
[11] 0.999327766 0.999908341 0.999991281 0.999999483 0.999999986
[16] 1.000000000
```

Inference on the population proportion p for large sample

- Normal approximation to binomial
- If $Y \sim \text{Binomial}(n, p)$, then for large n the distribution of

$$Z = \frac{Y - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

  is approximately $N(0, 1)$ by CLT
-

  is approximately $N(0, 1)$ by CLT and Slutsky's. Z^* will be pivotal quantity for inference on p for large sample.
- Approximation improves as $n \to \infty$
- Rule of thumb: $np(1-p) \geq 10$

Confidence interval for p for large sample

$$P(-z_{\alpha/2} \leq Z^* \leq z_{\alpha/2}) = 1 - \alpha$$
$$P(-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq z_{\alpha/2}) = 1 - \alpha$$
$$P(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 1 - \alpha$$

100(1-α)% large sample CI for p is

Hypothesis test on p for large sample

Suppose $H_0 : p = p_0$

- Test statistic

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$$

under H_0

An alternative test statistic is

$$\tilde{Z}_0 = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1)$$

Both test statistics are asymptotically standard normal under H_0, and either statistic may be used. Z_0 is more commonly used, although some prefer $\tilde{Z}_0$ which has a dual relationship between CI and hypothesis tests.

Hypothesis test on p for large sample

1. One tailed: $H_0 : p = p_0$ vs $H_1 : p > p_0$
   - At the significance level α, reject H_0 in favor of H_a if $Z_0 \geq z_\alpha$
   - Alternatively, if p-value$= P(Z_0 \geq z_0|H_0) \leq \alpha$, reject H_0
2. One tailed: $H_0 : p = p_0$ vs $H_1 : p < p_0$
   - At the significance level α, reject H_0 in favor of H_a if $Z_0 \leq -z_\alpha$
   - Alternatively, if p-value$= P(Z_0 \leq z_0|H_0) \leq \alpha$, reject H_0
3. Two tailed: $H_0 : p = p_0$ vs $H_1 : p \neq p_0$
   - At the significance level α, reject H_0 in favor of H_a if $|Z_0| \geq z_{\alpha/2}$
   - Alternatively, if p-value$= P(|Z_0| \geq |z_0||H_0) \leq \alpha$, reject H_0
   - Note that p-value$= P(|Z_0| \geq |z_0||H_0) = 2\min(P(Z_0 \leq z_0|H_0)P(Z_0 \geq z_0|H_0))$

Example: A college has 500 women students and 1,000 men students. The introductory zoology course has 90 students, 50 of whom are women. It is suspected that more women tend to take zoology than men. Test this suspicion at $\alpha = 0.05$.

Power calculation of $H_0 : p = p_0$ vs $H_a : p = p_1 > p_0$ for large sample

Suppose $H_0 : p = p_0$ vs $H_a : p = p_1 > p_0$.
At the significance level α, we reject H_0 if $Z_0 \geq z_\alpha$.

$\beta = P(\text{fail to reject } H_0 | H_a)$
Power $= 1 - \beta = P(\text{reject } H_0 | H_a) = P(Z_0 \geq z_\alpha | H_a : p = p_1 > p_0)$

$$= P\left(\frac{\hat{p}}{\sqrt{\frac{p_0(1-p_0)}{n}}} - \frac{p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_\alpha | H_a : p = p_1 > p_0\right)$$

$$= P\left(\hat{p} - p_1 \geq z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} + p_0 - p_1 | p_1\right)$$

$$= P\left(\frac{\hat{p}-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \geq \frac{p_0-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} + z_\alpha \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} | p_1\right)$$

$$=\mathrm{P}\left(Z \geq \frac{p_0-p_1}{\sqrt{\frac{p_1(1-p_1)}{\mathrm{n}}}}+z_\alpha\sqrt{\frac{p_0\left(1-p_0\right)}{p_1\left(1-p_1\right)}}\right)$$

where $Z \sim N\left(0,1\right)$

$$=1-\mathrm{P}\left(Z \leq \frac{p_0-p_1}{\sqrt{\frac{p_1(1-p_1)}{\mathrm{n}}}}+z_\alpha\sqrt{\frac{p_0\left(1-p_0\right)}{p_1\left(1-p_1\right)}}\right)$$

$$=1-\Phi\left(\frac{p_0-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}}+z_\alpha\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}\right)$$

$$=\Phi\left(\frac{p_1-p_0}{\sqrt{\frac{p_1(1-p_1)}{n}}}-z_\alpha\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}\right)$$

Power calculation of $H_0 : p = p_0$ vs $H_a : p = p_1 < p_0$ for large sample

Suppose $H_0 : p = p_0$ vs $H_a : p = p_1 < p_0$.
At the significance level α, we reject H_0 if $Z_0 \leq -z_\alpha$.

$$\text{Power} = 1-\beta = \text{P}\left(Z_0 \leq -z_\alpha | H_\text{a}\text{: p} = p_1\right)$$

$$= \text{P}\left(Z \leq \frac{p_0 - p_1}{\sqrt{\frac{p_1(1-p_1)}{\text{n}}}} - z_\alpha \sqrt{\frac{p_0\,(1-p_0)}{p_1\,(1-p_1)}} \right)$$

$$= \Phi\left(\frac{p_0 - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} - z_\alpha \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \right)$$

Power calculation of $H_0 : p = p_0$ vs $H_a : p = p_1 \neq p_0$ for large sample

Suppose $H_0 : p = p_0$ vs $H_a : p = p_1 \neq p_0$.
At the significance level α, we reject H_0 if $|Z_0| \geq z_{\alpha/2}$.

$$\text{Power} = 1-\beta = \text{P}\left(|Z_0| \geq z_{\alpha/2} | H_\text{a}: p = p_1\right)$$

$$= \text{P}\left(Z_0 \geq z_{\alpha/2} | H_\text{a}: p = p_1\right) + \text{P}\left(Z_0 \leq -z_{\alpha/2} | H_\text{a}: \text{p} = p_1\right)$$

$$= \text{P}\left(Z \geq \frac{p_0 - p_1}{\sqrt{\frac{p_1(1-p_1)}{\text{n}}}} + z_{\alpha/2}\sqrt{\frac{p_0\,(1-p_0)}{p_1\,(1-p_1)}}\right)$$

$$+\text{P}\left(Z \leq \frac{p_0 - p_1}{\sqrt{\frac{p_1(1-p_1)}{\text{n}}}} - z_{\alpha/2}\sqrt{\frac{p_0\,(1-p_0)}{p_1\,(1-p_1)}}\right)$$

Power

$$= \Phi\left(\frac{p_1 - p_0}{\sqrt{\frac{p_1(1-p_1)}{n}}} - z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}\right)$$

$$+\Phi\left(\frac{p_0 - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} - z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}\right)$$

Sample size determination

Sample size determination for large sample for a given margin of error E or CI length

For a given CI length L

$$L = \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} - (\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 2 \cdot z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$n = \frac{4z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{L^2} \leq$$

For a given margin of error E

$$P(|\hat{p} - p| \leq E) = 1 - \alpha \rightarrow E = L/2$$

$$n = \frac{4z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{(2E)^2} = \frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{E^2} \leq$$

Example: Thanksgiving was coming up and Harvey's Turkey Farm was doing a land-office business. Harvey sold 100 turkeys to Nedicks for their famous Turkey-dogs. Nedicks found that 90 of Harvey's turkeys were in reality peacocks.

(a) Estimate the proportion of peacocks at Harvey's Turkey Farm and find a 95% confidence interval for the true proportion of turkeys that Harvey owns using large sample approximation.

(b) How large a random sample should we select from Harvey's Farm to guarantee the length of the 95% confidence interval to be no more than 0.06? (Note: First derive the general formula for sample size calculation based on the length of the CI for inference on one population proportion, large sample case. Provide the formula for the two cases:

   (i) we have an estimate of the proportion
   (ii) we do not have an estimate of the proportion
   (iii) finally, plug in the numerical values and obtain the sample size for this particular problem.

Sample size determination for large sample for a given power and $H_0 : p = p_0$ vs $H_1 : p = p_1 > p_0$

Suppose $H_0 : p = p_0$ vs $H_1 : p = p_1 > p_0$.
At significance level α and a given power $1 - \beta$

$$-z_\beta = \frac{p_0 - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} + z_\alpha \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}$$

$$n = \left[\frac{z_\alpha \sqrt{p_0\,(1-p_0)} + z_\beta \sqrt{p_1\,(1-p_1)}}{p_1 - p_0} \right]^2$$

Note $\Phi(z_\beta) = 1 - \beta$

Sample size determination for large sample for a given power and $H_0 : p = p_0$ vs $H_1 : p = p_1 < p_0$

Suppose $H_0 : p = p_0$ vs $H_1 : p = p_1 < p_0$.
At significance level α and a given power $1 - \beta$

$$n = \left[\frac{z_\alpha \sqrt{p_0 (1-p_0)} + z_\beta \sqrt{p_1 (1-p_1)}}{p_1 - p_0} \right]^2$$

Sample size determination for large sample for a given power and $H_0 : p = p_0$ vs $H_1 : p = p_1 \neq p_0$

Suppose $H_0 : p = p_0$ vs $H_1 : p = p_1 \neq p_0$.
At significance level α and a given power $1 - \beta$

$$n \approx \left[\frac{z_{\alpha/2}\sqrt{p_0\,(1-p_0)} + z_\beta \sqrt{p_1\,(1-p_1)}}{p_1 - p_0} \right]^2$$

Example: A pre-election poll is to be planned for a senatorial election between two candidates. Previous polls have shown that the election is hanging in delicate balance. If there is a shift (in either direction) by more than 2 percentage points since the last poll, then the polling agency would like to detect it with probability of at least 0.80 using a 0.05-level test. Determine how many voters should be polled. If actually 2500 voters are polled, what is the power?

Example: Treatment of Kidney Cancer

Historically, one in five kidney cancer patients (i.e. 20%) survive 5 years past diagnosis. An oncologist using an experimental therapy treats $n = 40$ kidney cancer patients and 16 of them survive at least 5 years. Is there evidence that patients receiving the experimental therapy have a higher 5-year survival rate? Test at the significance level of $\alpha = 0.05$.