Inferences for Proportions and Count Data

Corresponds to Chapter 9 of

Tamhane and Dunlop

Slides prepared by Elizabeth Newton (MIT),
with some slides by Ramón V. León
(University of Tennessee)

Inference for Proportions

- Data = {0,1,1,10,0…..1,0}, Bernoulli(p)
- Goal – estimate p, probability of success (or proportion of population with a certain attribute)
- $\hat{p} = \bar{X}$ = number of successes in n trials
- Var($\hat{p}$) = p(1-p)/n = pq/n
- Variance depends on the mean.

Large Sample Confidence Interval for Proportion

Recall that $\dfrac{(\hat{p}-p)}{\sqrt{pq/n}} \approx N(0,1)$ if n is large

$(q = 1 - p, \quad n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10)$

<u>It follows that:</u>

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

<u>Confidence interval for *p*:</u>

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

A Better Confidence Interval for Proportion

Use this probability statement

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

Solve for p using quadratic equation

<u>CI for p:</u>

$$\frac{\hat{p} + \frac{z^2}{2n} - \sqrt{\frac{\hat{p}\hat{q}z^2}{n} + \frac{z^4}{4n^2}}}{\left(1+\frac{z^2}{n}\right)} \le p \le \frac{\hat{p} + \frac{z^2}{2n} + \sqrt{\frac{\hat{p}\hat{q}z^2}{n} + \frac{z^4}{4n^2}}}{\left(1+\frac{z^2}{n}\right)}$$

$$where\ z = z_{\alpha/2}$$

Example

See Example 9.1 on page 301 of the course textbook.

Binomial CI

In S-Plus:

```
>qbinom(.975,800,0.45)
[1] 388
> qbinom(.025,800,0.45)
[1] 332
```

95% CI for proportion of gun owners is:

$332/800 \leq p \leq 388/800$

$0.415 \leq p \leq 0.485$

Sample Size Determination for a Confidence Interval for Proportion

Want (1-α)-level two-sided CI:

$\hat{p} \pm E$ where E is the margin of error. Then $E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$.

Solving for n gives $n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}\hat{q}$

Largest value of $pq = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}$ so conservative sample size is:

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \frac{1}{4} \quad \text{(Formula 9.5)}$$

Example 9.2: Presidential Poll

See Example 9.2 on page 302 of the course textbook.

Threefold increase in precision requires ninefold increase in sample size

Largest Sample Hypothesis Test on Proportion

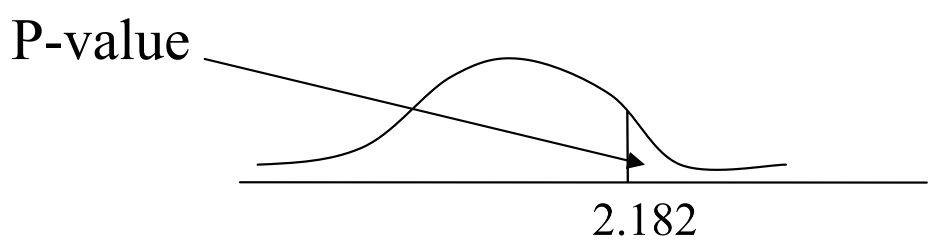
$$H_0 : p = p_0 \text{ vs. } H_1 : p \neq p_0$$

Best test statistics: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$

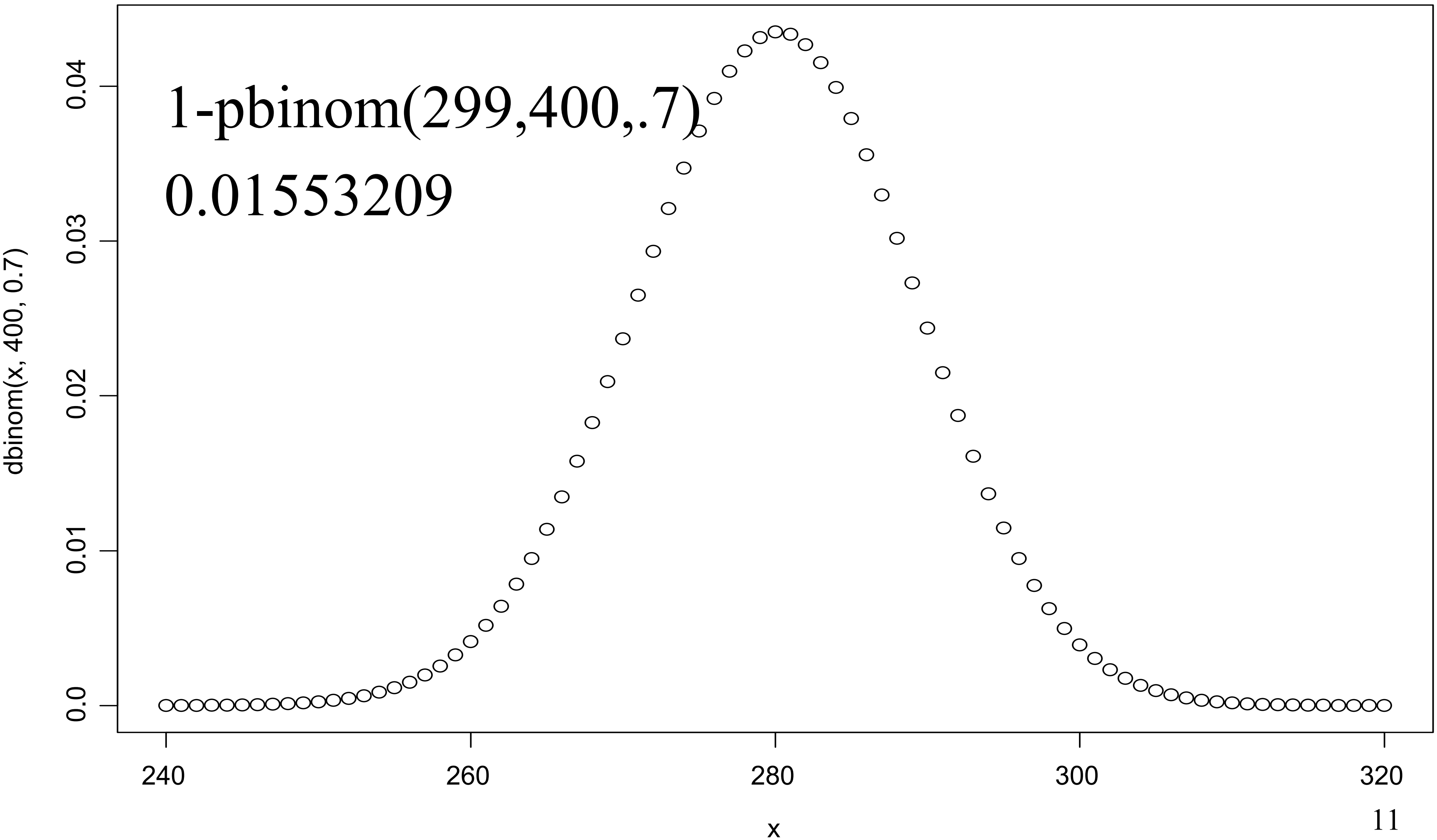
Acceptance Region: $p_0 \pm cd$, where $c = z_{a/2}$ and $d = (p_0 q_0/n)^{0.5}$

Basketball Problem: z-test

See Example 9.3 on page 303 of the course textbook.



Exact Binomial Test in S-Plus



Sample Size for Z-Test of Proportion

$$H_o: p \le p_0 \text{ vs. } H_1: p > p_0$$

Suppose that the power for rejecting H_0 must be at least 1- β when the true proportion is $p = p_1 > p_0$.

Let $\delta = p_1 - p_0$. Then

$$n = \left[\frac{z_\alpha \sqrt{p_0 q_0} + z_\beta \sqrt{p_1 q_1}}{\delta} \right]^2$$

Test based on:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$$

Replace z_α by $z_{\alpha/2}$ for two-sided test sample size.

Example 9.4: Pizza Testing

See Example 9.4 on page 305 of the course textbook.

$$n = \left[\frac{z_{\alpha/2}\sqrt{p_0 q_0} + z_{\beta}\sqrt{p_1 q_1}}{\delta} \right]^2$$

Comparing Two Proportions: Independent Sample Design

If $n_1 p_1,\ n_1 q_1,\ n_2 p_2,\ n_2 q_2 \geq 10,$ then

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

<u>Confidence Interval:</u>

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1 q_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Test for Equality of Proportions (Large n)
Independent Sample Design – pooled estimate of p

$$H_0: p_1 = p_2 \text{ vs. } H_1: p_1 \neq p_2$$

$$\text{Test statitics: } z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{where } \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x + y}{n_1 + n_2}$$

Example 9.6 – Comparing Two Leukemia Therapies

See Example 9.6 on page 310 of the course textbook.

Inference for Small Samples
Fisher's Exact Test

- Calculates the probability of obtaining observed 2x2 table or any more extreme with margins fixed.
- Uses hypergeometric distribution

$$P(X = x \mid N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$$

Inference for Count Data

Data = cell counts = number of observations in each of sevaral (>2) categories, $n_{i,}$ $i=1..c$, $\Sigma n_i=n$

Joint distribution of corresponding r.v.'s is multinomial.

Goal – determine if the probabilities of belonging to each of the categories are equal to hypothesized values, p_{i0}.

Test statistic, $\chi^2 = \Sigma(\text{observed-expected})^2/\text{expected}$, where observed=$n_i$, expected=$np_{i0}$

χ^2 has chi-square distribution when sample size is large

Multinomial Test of Proportions

See Example 9.10 on page 316 of the course textbook.

Inferences for Two-Way Count Data

	y: Job Satisfaction				
x: Annual Salary	Very Dissatisfied	Slightly Dissatisfied	Slightly Satisfied	Very Satisfied	**Row Sum**
Less than $10,000	81	64	29	10	184
$10,000-25,000	73	79	35	24	211
$25,000-50,000	47	59	75	58	239
More than $50,000	14	23	84	69	190
Column Sum	215	225	223	161	**824**

Sampling Model 1: Multinomial Model (Total Sample Size Fixed)
Sample of 824 from a single population that is then cross-classified

The null hypothesis is that X and Y are **independent**:

$$H_0 : p_{ij} = P(X = i, Y = j) = P(X = i)P(Y = j) = p_{i.}p_{.j} \text{ for all i, j}$$

Sampling Model 1 (Total Sample Size Fixed)
Based on Table 9.10 in the course textbook

	y: Job Satisfaction				
x: Annual Salary	Very Dissatisfied	Slightly Dissatisfied	Slightly Satisfied	Very Satisfied	**Row Sum**
Less than $10,000	81	64	29	10	184
$10,000-25,000	73	79	35	24	211
$25,000-50,000	47	59	75	58	239
More than $50,000	14	23	84	69	190
Column Sum	215	225	223	161	**824**

$$\text{Estimated Expected Frequency} = 824\left(\frac{215}{824}\right)\left(\frac{184}{824}\right) = \frac{215\times 184}{824} = 48.01$$

(Cell 1,1)

$$= np_{1\bullet}p_{\bullet 1}$$

Chi-Square Statistics

See Example 9.13, page 324 for instructions on calculating the chi-square statistic.

$$\chi^2 = \sum_{i=1}^{c} \frac{(n_i - e_i)^2}{e_i}$$

Chi-Square Test Critical Value

Based on Table A.5, critical values $\chi^2_{\upsilon,\alpha}$ for the Chi-square Distribution, in the course textbook:

	α						
v	.995	.99	.975	.95	.90	.10	.05
1							
2							
3							
4							
5							
6							
7							
8							
9							16.919
10							
11							

The d.f. for this χ^2- statistics is (4-1)(4-1) = 9. Since $\chi^2_{9,.05} = 16.919$ the calculated $\chi^2 = 11.989$ is not sufficiently large to reject the hypothesis of independence at $\alpha = .05$ level

S-Plus – job satisfaction example

```
Call:
crosstabs(formula = c(jobsat) ~ c(row(jobsat)) + c(col(jobsat)))
901 cases in table
+----------+
|N         |
|N/RowTotal|
|N/ColTotal|
|N/Total   |
+----------+
c(row(jobsat))|c(col(jobsat))
       |1      |2      |3      |4      |RowTotl|
-------+-------+-------+-------+-------+-------+
1      | 20    | 24    | 80    | 82    |206    |
       |0.097  |0.12   |0.39   |0.4    |0.23   |
       |0.32   |0.22   |0.25   |0.2    |       |
       |0.022  |0.027  |0.089  |0.091  |       |
-------+-------+-------+-------+-------+-------+
2      | 22    | 38    |104    |125    |289    |
       |0.076  |0.13   |0.36   |0.43   |0.32   |
       |0.35   |0.35   |0.33   |0.3    |       |
       |0.024  |0.042  |0.12   |0.14   |       |
-------+-------+-------+-------+-------+-------+
3      | 13    | 28    | 81    |113    |235    |
       |0.055  |0.12   |0.34   |0.48   |0.26   |
       |0.21   |0.26   |0.25   |0.27   |       |
       |0.014  |0.031  |0.09   |0.13   |       |
-------+-------+-------+-------+-------+-------+
4      |  7    | 18    | 54    | 92    |171    |
       |0.041  |0.11   |0.32   |0.54   |0.19   |
       |0.11   |0.17   |0.17   |0.22   |       |
       |0.0078 |0.02   |0.06   |0.1    |       |
-------+-------+-------+-------+-------+-------+
ColTotl|62     |108    |319    |412    |901    |
       |0.069  |0.12   |0.35   |0.46   |       |
-------+-------+-------+-------+-------+-------+
Test for independence of all factors
       Chi^2 = 11.98857 d.f.= 9 (p=0.2139542)
       Yates' correction not used
>
```

Product Multinomial Model: Row Totals Fixed

(See Table 9.2 in the course textbook.)

Sampling Model 2: Product Multinomial
Total number of patients in each drug group is fixed.

- The null hypothesis is that the probability of column response (success or failure) is the same, regardless of the row population:

$$H_0 : P(Y = j \mid X = i) = p_j$$

S-Plus – leukemia trial

```
Call:
crosstabs(formula = c(leuk) ~ c(row(leuk)) + c(col(leuk)))
63 cases in table
+----------+
|N         |
|N/RowTotal|
|N/ColTotal|
|N/Total   |
+----------+
c(row(leuk))|c(col(leuk))
       |1      |2      |RowTotl|
-------+-------+-------+-------+
1      |14     | 7     |21     |
       |0.67   |0.33   |0.33   |
       |0.27   |0.64   |       |
       |0.22   |0.11   |       |
-------+-------+-------+-------+
2      |38     | 4     |42     |
       |0.9    |0.095  |0.67   |
       |0.73   |0.36   |       |
       |0.6    |0.063  |       |
-------+-------+-------+-------+
ColTotl|52     |11     |63     |
       |0.83   |0.17   |       |
-------+-------+-------+-------+
Test for independence of all factors
      Chi^2 = 5.506993 d.f.= 1 (p=0.01894058)
      Yates' correction not used
      Some expected values are less than 5, don't trust stated p-value
>
```

Remarks About Chi-Square Test

- The distribution of the chi-square statistics under the null hypothesis is approximately chi-square only when the sample sizes are large
  - The rule of thumb is that all expected cell counts should be greater than 1 and
  - No more than 1/5th of the expected cell counts should be less than 5.
- Combine sparse cell (having small expected cell counts) with adjacent cells. Unfortunately, this has the drawback of losing some information.
- Never stop with the chi-square test. Look at cells with large values of (O-E), as in job satisfaction example.

Odds Ratio as a Measure of Association for a 2x2 Table

Sampling Model I: Multinomial

$$\psi = \frac{p_{11}/p_{12}}{p_{21}/p_{22}}$$

The numerator is the odds of the column 1 outcome vs. the column 2 outcome for row 1, and the denominator is the same odds for row 2, hence the name "odds ratio"

Odds Ratio as a Measure of Association for a 2x2 Table

Sampling Model II: Product Multinomial

$$\psi = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)}$$

The two column outcomes are labeled as "success" and "failure," then ψ is the odds of success for the row 1 population vs. the odds of success for the row 2 population