Biost 517 / Biost 514 Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 10:

Inference using Exact Binomial Distribution; Comparing Two Proportions; Risk Difference: Confidence intervals and Hypothesis Testing

Review: Normal Approximation to the Binomial

- Consider a random variable Y that has a Binomial distribution, $Y \sim B(n, p)$, where

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, 2, \ldots, n$$

with mean $E[Y] = np$ and standard deviation $\mathrm{SD}(Y) = \sqrt{np(1-p)}$

- Since Y is a sum of independent Bernoulli random variables, from the **central limit theorem** we know that for sufficiently large n, the binomial distribution can be approximated by a normal distribution where

$$Y \dot{\sim} N\left(\mu_Y = np, \sigma_Y = \sqrt{np(1-p)}\right)$$

Exact Binomial Distribution for inference on p

- The normal approximation for the binomial distribution will generally be appropriate if

$$np \geq 10 \qquad \text{and} \qquad n(1-p) \geq 10$$

- With this approximation, we were able to provide inference on p, using confidence intervals and hypothesis testing with a normal distribution.
- What about the settings when $np < 10$. This often occurs in studies when outcomes of interest are rare, e.g., p is close to 0.
- When the normal distribution is not a good approximation to the binomial, we will need to use the binomial distribution directly for inference on p.

Exact Binomial Confidence Intervals

- Use the binomial distribution
  - (But let a statistical software do it for you)

An exact $100(1-\alpha)\%$ confidence interval for p based on observation $Y = k$ is $(\hat{p}_L, \hat{p}_U)$ where an iterative search is used to find

$$\Pr[Y \le k; \hat{p}_U] = \sum_{i=0}^{k} \frac{n!}{i!(n-i)!} \hat{p}_U^i (1-\hat{p}_U)^{n-i} = \alpha/2$$

$$\Pr[Y \ge k; \hat{p}_L] = \sum_{i=k}^{n} \frac{n!}{i!(n-i)!} \hat{p}_L^i (1-\hat{p}_L)^{n-i} = \alpha/2$$

R: Exact Binomial Calculations

- Can obtain exact binomial calculations using the **binom.test()** function in R
- You provide the **binom.test()** function with the number of successes “x” and number of trials “n”.
- The function provides exact binomial confidence intervals for the proportion p.
- Also performs an exact binomial test for a simple null hypothesis. The default setting is “p=0.5”. To test a different value, you must specify it in the call to the function, e.g. “p=0.01”.

Example: Binomial Exact CIs

- We randomly sample 40 children and test for peanut allergy. Two of the children have a peanut allergy.
- Estimate the proportion of children who have a peanut allergy.
- Provide a 95% confidence interval for the proportion of children with a peanut allergy.

Example: Binomial Exact CIs

```
> binom.test(x=2,n=40)

        Exact binomial test

data:  2 and 40
number of successes = 2, number of trials = 40, p-value =
1.493e-09
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.006113647 0.169196864
sample estimates:
probability of success
                  0.05
```

Example: Binomial Exact CIs

- The point estimate for the proportion of children who have a peanut allergy is 0.05. Using an exact binomial calculation, with 95% confidence, the data are consistent with the true proportion of children who have a peanut allergy being anywhere between 0.006 and 0.17.

Example: Binomial Exact CIs

- Perform an hypothesis test to assess the evidence that true proportion of children with a peanut allergy is 0.2.

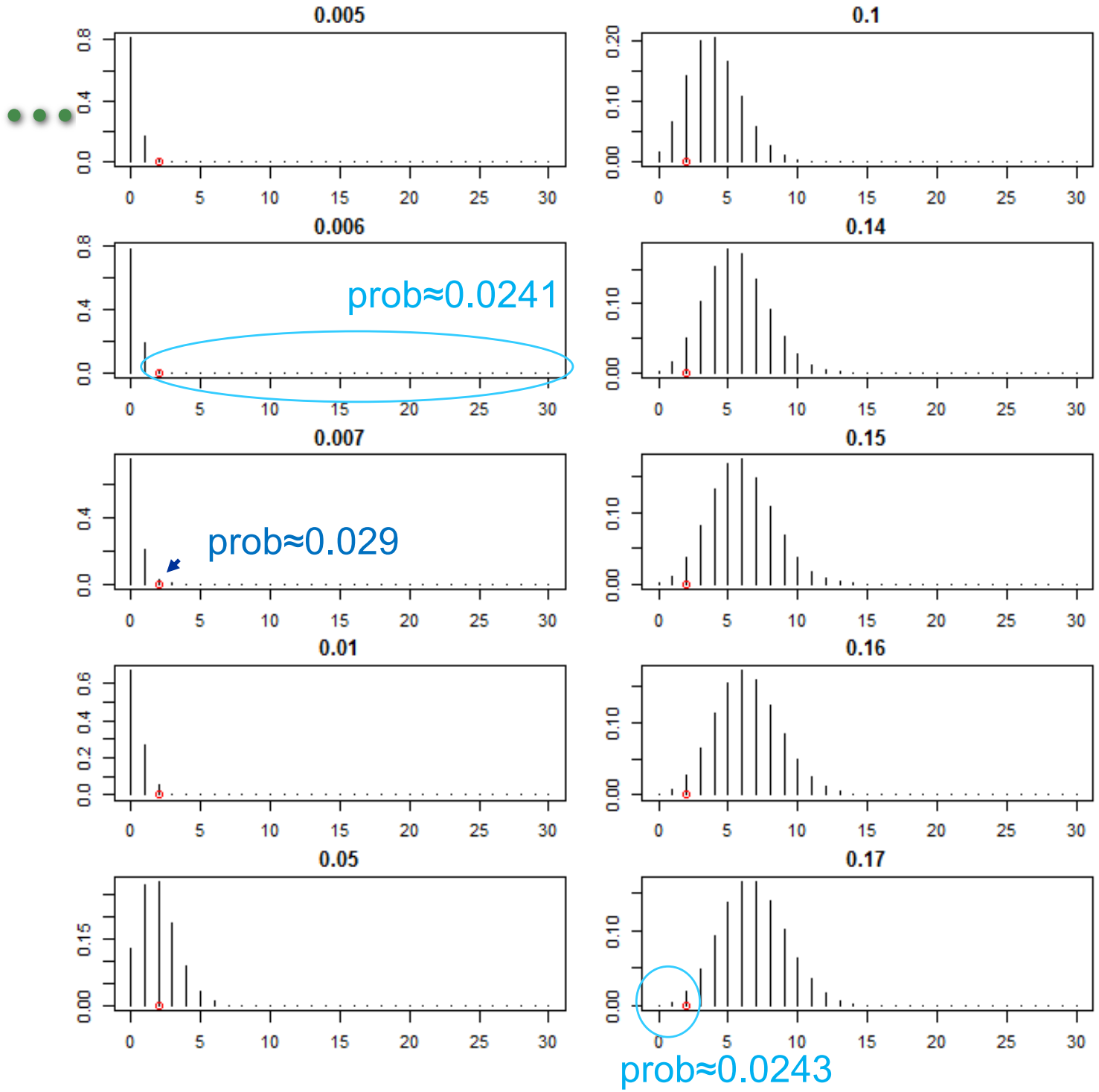
```
> binom.test(x=2,n=40,p=0.2)

        Exact binomial test

data:  2 and 40
number of successes = 2, number of trials = 40, p-value =
0.01586
alternative hypothesis: true probability of success is not equal to 0.2
95 percent confidence interval:
 0.006113647 0.169196864
sample estimates:
probability of success
                  0.05
```

Example: Binomial Exact CIs

- A two-sided exact binomial test was used to test the null hypothesis that the true proportion of children with a peanut allergy is 0.2. With a p-value is 0.016, we reject the null hypothesis at the 0.05 significance level. The data are not consistent with a proportion of 0.2 for children with a peanut allergy.

Binomial(40,p) distribution for different values of p.

Which values of p are in the binomial exact 95% confidence interval when we observe 2 out of 40?



Proportions: 0 events in n trials

- 2-sided confidence intervals fail in case where there are either 0 or n events observed in n Bernoulli trials
- However, we can derive one-sided confidence bounds in these cases.

Proportions: 0 events in n trials: Upper Confidence Bound

- Exact upper confidence bound when there are 0 “successes” or “events” in n trials

Suppose $Y \sim B(n, p)$ and $Y = 0$ is observed

Exact $100(1-\alpha)\%$ upper confidence bound for p is $\hat{p}_U$

$$\Pr[Y = 0; \hat{p}_U] = (1 - \hat{p}_U)^n = \alpha$$

$$\Downarrow$$

$$\hat{p}_U = 1 - \alpha^{1/n}$$

Large sample approximation

$$\left(1-\hat{p}_U\right)^n = \alpha \quad \Rightarrow \quad n\log\left(1-\hat{p}_U\right) = \log\left(\alpha\right)$$

For small $\hat{p}_U$ $\qquad\qquad$ $\log\left(1-\hat{p}_U\right) \approx -\hat{p}_U$

so for large n $\quad \Rightarrow \quad$ $\hat{p}_U \approx -\dfrac{\log\left(\alpha\right)}{n}$

Large sample approximation

- "Three over n rule"
  - log(.05) = -2.9957
  - So for 0 events in n trials upper confidence bound is approximately 3/n

- 99% upper confidence bound
  - log(0.01)= -4.605
  - Use 4.6/n as 99% upper confidence bound

Approximation vs Exact

- When Y=0 events observed in n Bernoulli trials

	95% bound		99% bound	
n	Exact	3/n	Exact	4.6/n
2	.7764	1.50	.9000	2.3000
5	.4507	.60	.6019	.9200
10	.2589	.30	.3690	.4600
20	.1391	.15	.2057	.2300
30	.0950	.10	.1423	.1533
50	.0582	.06	.0880	.0920
100	.0295	.03	.0450	.0460

- Impress your friends! Compute confidence intervals during an elevator ride!

n events in n trials

- We can also use the “three over n rule” to find the lower confidence bound for p when every trial of n trials has an event
  - Lower 95% confidence bound is 1 - 3/n

Comparing two proportions

- It is often of interest to comparing rates or proportions across two groups.
- For example, it may be of interest to assess if there is a difference in the probability of death within 5 years among elderly subjects who smoke as compared to elderly individuals who do not smoke in the Cardiovascular Health Study.
- To do this we would first calculate the probability (proportion) of death within 5 years for smokers and the probability (proportion of death) within 5 years for non-smokers, and the obtain a point estimate for the difference of the two proportions

Comparing two proportions: risk difference

- The difference in proportions is often referred to as the **risk difference.**
- How can we obtain inference on the risk difference for two populations using the estimated **risk difference** from two samples?
- If the true risk difference for the populations is 0, then there is no association between the risk (or probability) of the event, which is the outcome of interest, and the predictor variable used to define the two populations that are being compared.

Confidence Intervals for Two Proportions: Risk Difference

Confidence Intervals for Risk Difference

Suppose we have two populations A and B with unknown proportions p_1 and p_2 respectively. A SRS of size n_1 from A yields $\hat{p}_1$, and an independent SRS of size n_2 from B yields $\hat{p}_2$. Then,

$$(\hat{p}_1 - \hat{p}_2) \dot{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

when n_1 and n_2 are large.

An approximate $(1-\alpha)$ CI for $p_1 - p_2$ is then given by

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where as before, z^* is the critical value that is selected such that $P(Z > z^*) = \alpha/2$ for $Z \sim N(0,1)$.

Hypothesis Testing with Two Proportions: Risk Difference

A Test for Two Population Proportions

To test $H_0 : p_1 = p_2$, we compute the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p}$ is the *combined* proportion of successes in both samples

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

with X_1, X_2 denoting the number of successes in each sample.

Under H_0, the Z-statistic has approximately a standard normal distribution (using the normal approximation to the binomial), and p-values are then calculated similarly to the one-sample case.