# MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

# Review

- Last time, we introduced Nonparametric test for two populations comparison.
- Sign test, Wilcoxon signed-rank test and Wilcoxon rank-sum test.
- Know the model assumptions and how to conduct using R.

- Today we start on a theory of permutation test for which Wilcoxon rank-sum test is a special case. Then we go on to Module 9 Inferences on Proportions

# Rank-sum Test Statistic's distribution

- Example Data: Two groups A: 11,32 and B: 28,52,42
- Together {11,28,32,42,52}

| | | | | | |
|---|---|---|---|---|---|
| | A | B | A | B | B |
| Rank | 1 | 2 | 3 | 4 | 5 |

Hence $r_A = 1+3 = 4$, $r_B = 2+4+5 = 11$. And $W = 4$. We then look up $P(W \le 4)$ from Table A.7.

**But where does the Table A.7 comes from?**

***TABLE A.7***

Distribution functions of $W$, the Wilcoxon rank sum te

| | $n_2 = 3$ | | |
|---|---|---|---|
| $W_0$ | $n_1 = 1$ | 2 | 3 |
| 1 | 0.25 | | |
| 2 | 0.50 | | |
| 3 | | 0.10 | |
| 4 | | 0.20 | |
| 5 | | 0.40 | |
| 6 | | 0.60 | 0.05 |
| 7 | | | 0.10 |
| 8 | | | 0.20 |
| 9 | | | 0.35 |
| 10 | | | 0.50 |

# Theory of Permutation Test (Rank-sum Test)

- Since the **Rank-sum test statistic** W is based on ranks, it can only take finite many possible values. We can get its exact distribution (Table A.7) by enumeration through all possible values.
- In the example data, we can permutate the 5 observations and then put the first two in the group A. Under **$H_0$**, the two groups come from the same distribution, the resulting new separated two groups have the same probability as the original data. And we calculate W on all possible permutated data.

# Theory of Permutation Test (Rank-sum Test)

- Permutation test: permutate the 5 observations and then put the first two in the group A. Equally likely to be 11, 28,|32, 42, 52 or 28, 42,|32, 11,52 or …
- Total 5! permutations, but 2!3! of them give the same grouping, thus the same W value. That is, there are only $\frac{5!}{2!3!} = \binom{5}{2}$ different possible groupings. Each of these grouping, <u>Under **$H_0$**</u>, gives $1/\binom{5}{2}$=1/10 probability for the corresponding W value. That is,

P(ranks in group A is 1 and 2) =P(ranks in group A is 1 and 3) = … =P(ranks in group A is 4 and 5) = 1/10

# Permutation Test (Rank-sum Test)

- From the permutations,

P(W≤3) = P(ranks in group A is 1 and 2) = 1/10

P(W≤4) = P(ranks in group A is 1 and 2)
+ P(ranks in group A is 1 and 3) = 2/10

P(W≤5) = P(ranks in group A is (1,2), (1,3), (1,4), (2,3))
= 4/10

P(W≤6)
= P(A ranks (1,2), (1,3), (1,4), (1,5),
(2,3),(2,4))
= 6/10

| | $n_2 = 3$ | |
|---|---|---|
| $W_0$ | $n_1 = 1$ | 2 |
| 1 | 0.25 | |
| 2 | 0.50 | |
| 3 | | 0.10 |
| 4 | | 0.20 |
| 5 | | 0.40 |
| 6 | | 0.60 |
| 7 | | |

# Permutation Test

- The Wilcoxon rank-sum test is a permutation test using the rank-sum test statistic.
- Generally we can get a nonparametric test using ***any statistic*** in the permutation test.

- Example: to test $\mathbf{H_0}$: $\mu_A = \mu_B$ versus $\mathbf{H_A}$ : $\mu_A \neq \mu_B$.

The t-test is a parametric test using T= $\bar{X}_A - \bar{X}_B$.
We can get a nonparametric test also by permutation on T= $\bar{X}_A - \bar{X}_B$.

# Permutation Test

- On the example data {11,28,32,42,52} , each of the 10 groupings are equally likely (1/10 probability):

11, 28 in group A, $T$= 19.5-42 = -22.4
11, 32 in group A, $T$= 21.5-40.667 = -19.167
11, 42 in group A, $T$= 26.5-37.333 = -10.833
…
Can built a table with these cases.

- We observe group A = {11,32}. Thus on $T_{obs}$= -19.167. From the above table, $P(T \le T_{obs})=2/10$. 2-sided p-value = 4/10.

# Permutation Test

- Notice that the permutation mean test above is similar to the Wilcoxon rank-sum Test, but these are two different tests.
- On the example data {11,28,32,42,52} ,

  $W \leq 3 \Leftrightarrow T \leq -22.4$

  $W \leq 4 \Leftrightarrow T \leq -19.167$

But $W=5 \Leftrightarrow$ ranks in A is (1,4) or (2,3)

$\Leftrightarrow$ (11, 42) in group A or (28, 32) in group A

$\Leftrightarrow T= -10.833$ or $T=30-35 = -5$

- They are not the same test.

# Permutation Test

- Generally for two samples, choose an approximate test statistic, we can get a nonparametric test through the permutation test on that test statistic.
- **H$_A$** : $\mu_A \neq \mu_B$. T can be $\bar{X}_A - \bar{X}_B$ or rank-sum.
- **H$_A$** : $\sigma_A \neq \sigma_B$. T can be $s_A / s_B$.
- The exact distribution of T: enumerate the (n+m)! permutations.
- When n+m big, can find the approximation distribution of T by resampling K times: permute and put first n observation in group A.

  This is equivalent to resample n out of n+m without replacement: $(x_1^*, \ldots, x_n^*)(x_{n+1}^*, \ldots, x_{n+m}^*)$ and calculate T*.

  Do this K times, get $T_1^*, \ldots, T_K^*$. Then, for big K (e.g. K=10,000) use the empirical distribution of $T_1^*, \ldots, T_K^*$ as to approximate distribution of T.

# Bootstrap versus Permutation Test

- Bootstrap:

Resample with replacement:$(x_1^*, \ldots, x_n^*, x_{n+1}^*, \ldots, x_{n+m}^*)$

That is, $x_1^*$ can= $x_2^*$ .

$x_1^*$ has 1/(n+m) chance being any one of $x_1$ , …, $x_{n+m}$

Then calculate T*= T$(x_1^*, \ldots, x_n^*, x_{n+1}^*, \ldots, x_{n+m}^*)$.

- Do this K times. Then, for big K (e.g. K=10,000) use the empirical distribution of $T_1^*, \ldots, T_K^*$ as to approximate distribution of T.

# Bootstrap versus Permutation Test

- (1) Bootstrap: resample with replacement.
  Permutation test: resample without replacement.
- (2) Bootstrap is usually used in the estimation setting. Permutation test is used for hypothesis test (permute under the null hypothesis assumption).
- (3) For permutation test, it is possible to iterate over all possible permutations to get exact distribution. For computational cost, permutation test may resample only K times like bootstrap. In such a case, both methods are Monto-Carlo simulations.

# Permutation Test

- Since the permutation results in $\binom{n+m}{n}$ equally likely outcomes of T, for the permutation test, the smallest possible p-value ≥ $1/\binom{n+m}{n}$

- In the above example n=2 and m=3, the smallest possible p-value is 0.10. The $\mathbf{H_0}$ can never be rejected by the permutation test at α=0.05 level no matter what observations are obtained!

# Summary

Module 8 cover the nonparametric test

- Two-sample tests: Sign test, Wilcoxon signed-rank test and Wilcoxon rank-sum test.

- Know when to use which (paired versus independent two samples).
- Can use R to do them.
- Understand the permutation test: Wilcoxon rank-sum test is a permutation test.

- Homework 6 is due in one week.

# **Chapter 14** Inferences on Proportions

- We have mostly done the inferences for population means. Here we will do similar inferences for proportions using Binomial distribution.
- Example: We want to estimate the proportion of voters who support keeping death penalty. In a survey of 329 voters, 209 supported.
- X = # of voters out of 329 who support ~Bin(n=329, p) where p is the true proportion in all voters.
- **(1) Point estimator**: $\hat{p} = \frac{X}{n} = \frac{209}{329} = 0.635$

# Chapter 14 Inferences on Proportions

- **(2) Confidence Interval:**

  When n is big, Bin(n,p) ≈N(np, np(1-p))

  (How big is needed? Rule of thumb: n≥30, np≥5, n(1-p)≥5)

Then $\hat{p} = \frac{X}{n}$ ≈N(p, p(1-p)/n) ⇒ $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ ≈N(0, 1)

But p is unknown, we use $\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$ ≈N(0, 1) to get

(1-α) C.I. as $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

# **Chapter 14** Inferences on Proportions

- **(2) Confidence Interval**:

  When n is big, a (1-α) C.I. is $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

  Notice variance estimator also used $\hat{p}$

- Example: Since $\hat{p} = \frac{209}{329} = 0.635$, a 95% C.I. for the proportion of death penalty supporters is

$$0.635 \pm 1.96\sqrt{\frac{0.635(1-0.635)}{329}} = (0.583, 0.687)$$

# **Chapter 14** Inferences on Proportions

- **(3) Hypothesis test:** **H**$_0$: $p=p_0$ versus **H**$_A$: $p>p_0$.

When $p=p_0$, $\dfrac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0, 1)$

Hence we reject **H**$_0$ at $\alpha$ level if $\dfrac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_\alpha$.

Notice we use $\dfrac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ here, not $\dfrac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$.

This is NOT exactly equivalent to the $(1-\alpha)$ C.I. formula.

# Chapter 14 Inferences on Proportions

- **(3) Hypothesis test:**

Example: Does death penalty has majority support?

**$H_0$**: p=0.5 versus **$H_A$**: p>0.5.

$$\frac{\hat{p}-0.5}{\sqrt{\frac{0.5(1-0.5)}{329}}} = \frac{0.635-0.5}{\sqrt{\frac{0.5(1-0.5)}{329}}} = 4.90.$$

From Table A.3, p-value < 0.001.

Hence we reject **$H_0$** at $\alpha=0.01$ level.

Conclusion: Death penalty does have majority support.

# Chapter 14 Inferences on Proportions

- **(2)* Corrected Confidence Interval**:

While the (1-α) C.I. formula $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ works for large sample size, in practice may not work very well. There are improved confidence interval formulas.

One way to get a better C.I. is to simply inverse the hypothesis test formula, and yield the Wilson interval:

$$\frac{\hat{p} + (z_{\alpha/2})^2/(2n)}{1 + (z_{\alpha/2})^2/n} \pm \frac{\sqrt{n} + z_{\alpha/2}}{n + (z_{\alpha/2})^2}\sqrt{\hat{p}(1-\hat{p}) + \frac{(z_{\alpha/2})^2}{4n}}$$

# Wilson Interval

Derive the (1-α) C.I. $(p_L, p_U)$ where for testing **H$_0$**: p=$p_L$ versus **H$_A$**: p>$p_L$, p= $p_L$ is rejected exactly at α/2 level.

That is, $\frac{\hat{p}-p_L}{\sqrt{\frac{p_L(1-p_L)}{n}}}$ = $z_{\alpha/2}$.

To get the Wilson interval formula, we solve this equation

$$\hat{p} - p_L = z_{\alpha/2}\sqrt{\frac{p_L(1-p_L)}{n}} \Leftrightarrow (\hat{p} - p_L)^2 = (z_{\alpha/2})^2 \frac{p_L(1-p_L)}{n}$$

# Wilson Interval

- **Wilson Interval**: shorthand notation $\kappa = z_{\alpha/2}$

$$(\hat{p} - p_L)^2 = \kappa^2 \frac{p_L(1-p_L)}{n}$$

$$\Leftrightarrow p_L{}^2 - 2\hat{p}p_L + \hat{p}^2 = -\frac{\kappa^2}{n} p_L{}^2 + \frac{\kappa^2}{n} p_L$$

$$\Leftrightarrow \left(1+\frac{\kappa^2}{n}\right) p_L{}^2 - 2\left(\hat{p} + \frac{\kappa^2}{2n}\right) p_L + \hat{p}^2 = 0$$

Recall, the solution to $ax^2+bx+c=0$ is $x = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$

So $p_L = \dfrac{2\left(\hat{p} + \frac{\kappa^2}{2n}\right) \pm \sqrt{4\left(\hat{p} + \frac{\kappa^2}{2n}\right)^2 - 4\left(1+\frac{\kappa^2}{n}\right)\hat{p}^2}}{2\left(1+\frac{\kappa^2}{n}\right)}$

# Wilson Interval

$$p_L = \frac{\cancel{2}\left(\hat{p} + \frac{\kappa^2}{2n}\right) \pm \sqrt{\cancel{4}\left(\hat{p} + \frac{\kappa^2}{2n}\right)^2 - \cancel{4}\left(1 + \frac{\kappa^2}{n}\right)\hat{p}^2}}{\cancel{2}\left(1 + \frac{\kappa^2}{n}\right)}$$

$$\Leftrightarrow \qquad p_L = \frac{\left(\hat{p} + \frac{\kappa^2/2}{n}\right) \pm \sqrt{\hat{p}^2 + 2\frac{\kappa^2}{2n}\hat{p} + \left(\frac{\kappa^2}{n}\right)^2/4 - \hat{p}^2 - \left(\frac{\kappa^2}{n}\right)\hat{p}^2}}{\left(\frac{n+\kappa^2}{n}\right)}$$

$$= \frac{\hat{p} + \kappa^2/(2n)}{1 + \frac{\kappa^2}{n}} \pm \frac{\sqrt{\frac{\kappa^2}{n}\left[\hat{p} - \hat{p}^2 + \frac{\kappa^2}{n}/4\right]}}{\left(\frac{n+\kappa^2}{n}\right)}$$

$$= \frac{\hat{p} + \kappa^2/(2n)}{1 + \kappa^2/n} \pm \frac{\kappa\sqrt{n}}{n+\kappa^2}\sqrt{\hat{p}(1-\hat{p}) + \frac{\kappa^2}{4n}}$$

# Wilson Interval

Notice that $p_U$ solves $(\hat{p} - p_U)^2 = (-z_{\alpha/2})^2 \frac{p_U(1-p_U)}{n}$.

$p_L$ and $p_U$ are both solutions to the same equation.

$p_L$ is the solution with "-", $p_U$ is the solution with "+".

Thus **Wilson interval** formula is

$$\frac{\hat{p} + (z_{\alpha/2})^2/(2n)}{1 + (z_{\alpha/2})^2/n} \pm \frac{\sqrt{n}z_{\alpha/2}}{n + (z_{\alpha/2})^2}\sqrt{\hat{p}(1-\hat{p}) + \frac{(z_{\alpha/2})^2}{4n}}$$

The standard CI is called **Wald interval** $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

There are other CI formulas (http://projecteuclid.org/download/pdf_1/euclid.ss/1009213286)

# **Chapter 14** Inferences on Proportions

- **(3)* Exact test**: Derive the test from the Binomial distribution instead of its normal approximation.
- Example: There were 13 deaths among workers at a nuclear power plant aged between 55 and 64. Of these 5 were due to cancer. National statistics for this age group (55-64) is 20% deaths due to cancer.

  Is there reason to be concerned?
- Solution: $p$= proportion of cancer deaths among all nuclear power workers (age 55-64)

$$\mathbf{H_0}: p \leq 0.2 \text{ versus } \mathbf{H_A}: p > 0.2.$$

# **Chapter 14** Inferences on Proportions

**(3)* Exact test**: Nuclear power plant example

test $\mathbf{H_0}$: $p \le 0.2$ versus $\mathbf{H_A}$: $p > 0.2$.

$X$= # of cancer deaths out of 13 (age 55-64)

~Bin(n=13, p). We can test from here, no need for normal approximation.

p-value = $P(X \ge 5 \mid p=0.2) = 0.09913061$ from R using

```
1-pbinom(4.5, size=13, prob=0.2)
```

Hence we reject $\mathbf{H_0}$ at $\alpha=0.10$ level, but fail to reject $\mathbf{H_0}$ at $\alpha=0.05$ level. There are reasons to be concerned but not very strong evidence that something is wrong.

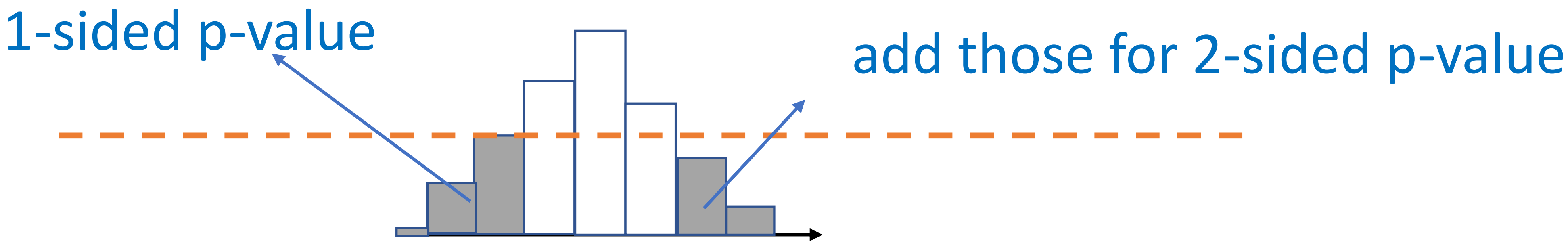
# **Chapter 14** Inferences on Proportions

## **(3)* Exact test**:

- We can also use R to directly do the binomial test

  binom.test(x=5, n=13, p=0.2, alternative="greater")

- The 2-sided p-value given by binom.test():




- Can use a simpler formula 2-sided p-value=2*(1-sided p-value)

# Summary

Today, we finished Module 8 the nonparametric test. Particularly today we discussed the permutation test, of which the Wilcoxon rank-sum test is a special case.

- Homework 6 is due in one week.

- We started Module 9 Inferences on proportions and went through the standard methods for one population proportion inferences which are based on the normal approximation.

Wilson interval and exact test were also covered.