



Biostatistics

Collaboration Center

# Basic Biostatistics in Medical Research

## Lecture 2: Two Group Comparisons

*Leah J. Welty, PhD*

*Biostatistics Collaboration Center (BCC)*

*Department of Preventive Medicine*

*NU Feinberg School of Medicine*

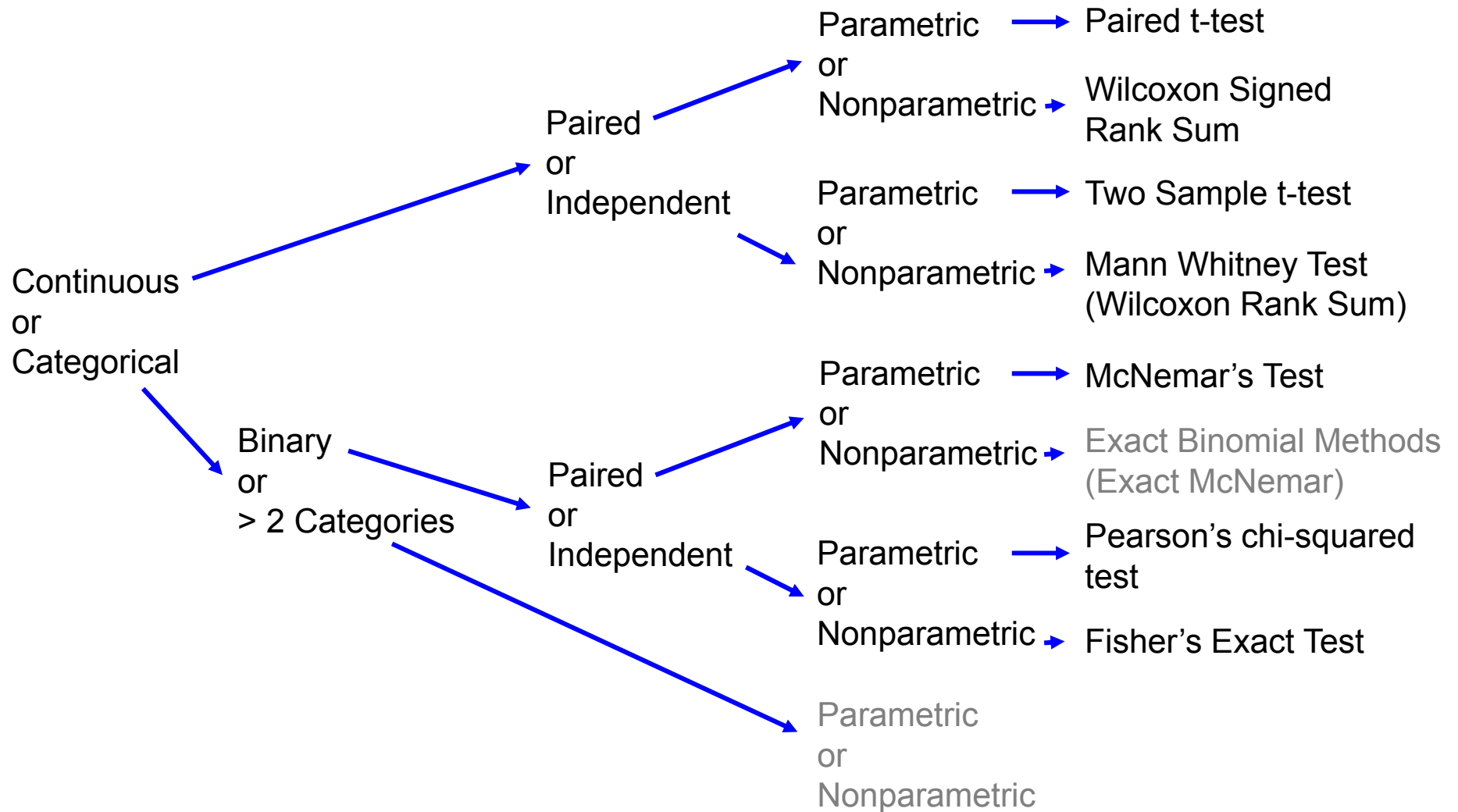
# Objectives

- Assist participants in interpreting statistics published in medical literature
- *Highlight* different statistical methodology for investigators conducting own research
- Facilitate communication between medical investigators and biostatisticians

# Two Group Comparisons

- Compare on continuous or categorical?
  - Compare SBP between two groups
  - Compare rate of Hodgkin's between two groups
- Samples paired or independent?
  - Measure on right and left hand each subject
  - Measure right hand one group, left hand other group
- Parametric or nonparametric?
  - Data satisfy necessary assumptions or not
- Looking for association or agreement?
  - Tonsillectomy a risk factor for Hodgkin's (association)
  - Two radiologists evaluating mammograms for cancer (agreement)

# Two Group Comp (Association)



# Examples of Paired Data

- Each subject measured before and after tx
- Subjects recruited in matched pairs
  - match age, postal code, diagnosis
  - one of pair gets tx A, the other gets tx B
- Twins or child/parent pairs
- Lab experiment with control and treatment samples handled in parallel

# Paired Procedures: When to Use

- If experimental design is paired, should use stats that account for pairing.
- Pairing arranged *before* data collected.
- If use stats that don't account for pairing on *paired* data, you're throwing away good information!

# Independent Groups

- Most common analyses for comparing two independent groups of observations
- Observations in one group independent of observations in other group
- Many clinical trials
  - treatment group independent of control group
- Observational Studies
  - sample individuals independently of each other

# SBP and OC Examples: Paired & Independent

- Longitudinal Study
  - Recruit nonpregnant, premenopausal women age 16-49 who are not currently OC users
  - Measure their baseline BP
  - Re-screen 1 year later. Measure BP on women who have become OC users.
  - Examine mean *difference* between baseline and follow-up SBP.
  - Values are *paired*
- Cross-Sectional Study
  - Recruit OC users and non-OC users among study population
  - Compare SBP of OC group to SBP of non-OC group
  - Groups are *independent*

# ***Comparing a Continuous Variable Between***

## ***Two Paired Groups***

### ***Parametric Methods (Comparing Means)***

# OC and SBP Example I (Paired)

- Does OC use change SBP?

<u>Subj</u>	<u>SBP pre OC</u>	<u>SBP while OC</u>	<u>Diff</u>
1	115	128	13
2	112	115	3
3	107	106	-1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	-2
10	115	117	2

# Paired t-test

- Mean difference = 4.8, sd difference = 4.6
- Estimated SE =  $SD/\sqrt{n} = 1.45$
- *Assuming differences in SBP are normally distributed*, the mean differences for a sample of 10 has  $t_9$  distribution
- 95% CI:  $4.8 \pm 2.25^* SE = (1.5, 8.1)$  mm Hg  
From  $t_9$  distribution

# Paired t-test Con't

- Hypothesis Test

$H_0$ : mean difference = 0 (no change in SBP)

$H_a$ : mean difference  $\neq$  0 (some change in SBP)

- If mean difference is 0, what is probability of observing a sample of 10 with a mean difference as or more extreme as 4.8 mm Hg?

p-value = 0.01

- Reject  $H_0$  in favor of  $H_a$ .
- Note: 0 not in 95% CI (1.5, 8.1) and p-value < 0.05

# Assumptions Paired t-test

- Assume *differences* are normally distributed (or  $n$  is large) for t-test
- Measurements themselves do not need to be normally distributed
- What if we can't assume differences are normally distributed and/or  $n$  is not large?

***Comparing Continuous Variable Between  
Two Paired Groups***

***Nonparametric Methods  
(Comparing Medians, Ranks)***

# Wilcoxon Signed Rank Sum

- Nonparametric analog to paired t-test
- For each pair, compute difference
- Rank (small to large) the *absolute* differences
  - Also ignore any differences = 0
- Add up ranks of positive differences, negative differences
- Compare the sums

# Wilcoxon Signed Rank Sum

- SBP and OC use

<u>Diff</u>	<u>Absolute Diff</u>	<u>Sign</u>	<u>Ranks</u>	
13	13	+	10	
3	3	+	4	
-1	1	-	1	
9	9	+	9	sum "+" ranks = 51.5
7	7	+	7.5	
7	7	+	7.5	
6	6	+	6	
4	4	+	5	sum "-" ranks = 3.5
-2	2	-	2.5	
2	2	+	2.5	

# Wilcoxon Signed Rank Sum

- If effect of OC on SBP, would we expect to see a difference in the rank sums as or more extreme?
  - 3.5 vs. 51.5
  - p-value = 0.02
- Caveats for Wilcoxon Signed Rank Sum
  - p-value changes if transform data (e.g. take logs)
  - assumes differences *symmetric* about zero
  - approximate for ties in ranks
  - sometimes used for comparison of ordinal data

***Comparing a Continuous Variable Between  
Two Independent Groups***

***Parametric Methods  
(Comparing Means)***

# OC and SBP Example II

- Does OC use change SBP?
- Sample 1
  - 8 35-39 year old women nonpregnant OC users
- Sample 2
  - 21 35-39 year old women nonpregnant not OC users

	<u>n</u>	<u>mean</u>	<u>sd</u>
<b>OC</b>	<b>8</b>	132.86 mm Hg	15.34 mm Hg
<b>non-OC</b>	<b>21</b>	127.44 mm Hg	18.23 mm Hg

# T-test for Independent Groups

- The difference in sample means will follow a  $t_{n_1+n_2-2}$  distribution IF
  - observations in pop 1 normally distributed
  - observations in pop 2 normally distributed
  - populations have same variance (or true std dev)
- $n_1 + n_2 - 2$  degrees of freedom
- Estimated SE of difference in sample means:

$$\sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}} \bullet \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Pooled SD

Analogous to  $\sqrt{n}$  for one sample

# T-test for Independent Groups

- For OC and non-OC groups

Diff in means =  $132.86 - 127.44 = 5.42$  mm Hg

SE = 7.282 mm Hg

$n_1 + n_2 - 2 = 8 + 21 - 2 = 27$

$t_{27}$  distribution

– 95% CI for mean diff:

$5.42 \pm 2.052 * SE = (-9.5, 20.4)$  mm Hg

From  $t_{27}$  distribution

# Paired vs Independent Comparisons of SBP

- Paired
  - Eliminate heterogeneity in baseline SBP levels by subtracting before and after values
  - Drift over time?
- Independent
  - Heterogeneity in baseline SBP levels implies additional variation to account for
  - Two samples have different SBP levels because
    - they're different subjects
    - possibly because they differ by OC/non-OC use

# T-test Assumptions

- Ways to test equal variance assumption
  - F-test with ratio of standard deviations
- t-procedures exist to compare means for populations with unequal variances
  - debate over usefulness

# ***Comparing a Continuous Variable Between***

## ***Two Independent Groups***

### ***Nonparametric Methods (Comparing Ranks)***

# Mann-Whitney Test

- What if we can't assume the underlying distributions are normal and/or  $n$  is small?
- Mann-Whitney U Test
  - Compares ranks between groups
  - Analogous to Wilcoxon Rank Sum
    - Same result, slightly different method
    - Wilcoxon Rank Sum  $\neq$  Wilcoxon Signed Rank Sum
  - Complicated corrections if lots of ties

# Receptors on Lymphocytes

- Independent (unmatched) samples

<u>Control</u>	<u>Drug</u>
1162	892
1095	903
1327	1164
1261	1002
1103	961
1235	875

- Number of receptors on lymphocytes
- small sample sizes
- Underlying distributions may not be normal

# Mann Whitney Test

- Rank all values (small to large) ignore grouping
- Sum ranks in each group
- Test compares sums of ranks
- Software or tables for p-values

# Mann Whitney Test

<u>Control</u>	<u>Drug</u>	
1162 (8)	892 (2)	
1095 (6)	903 (3)	
1327 (12)	1164 (9)	p = 0.015
1261 (11)	1002 (5)	
1103 (7)	961 (4)	
1235 (10)	875 (1)	
Sum of ranks = 56	Sum of ranks = 24	

# Side Note on Mann Whitney Test

- Also useful for some ordinal or rank data
  - e.g. visual acuity

<u>Treatment Group</u>	<u>Control Group</u>
20-30	20-40
20-20	20-240
...	
20-100	20-60

- Can assign ranks, but not take means

# ***Comparing a Binary Categorical Variable Between Two Groups***

## ***A Few More Summary Statistics***

# Two Group Binary Comparison

- Typical data format

<u>Subject</u>	<u>Exposure</u>	<u>Disease</u>
1	yes	yes
2	yes	no
3	no	yes
...		
n	no	no

- Often coded by 0s and 1s

<u>Subject</u>	<u>Exposure</u>	<u>Disease</u>
1	1	1
2	1	0
3	0	1
...		
n	0	0

# 2 x 2 Contingency Tables

- Classic Setup

	Exposed	Unexposed	
Disease	A	B	A + B
Not Diseased	C	D	C + D
	A + C	B + D	A + B + C + D

- Is proportion diseased (binary variable) different between exposed and unexposed groups?

# Additional Summaries: Relative Risk

- Cohort Study: Vitamin A & Blindness
  - Sommer & colleagues followed >25,000 children in Indonesia

	Vitamin A	Placebo
Died	101	130
Survived	12,890	12,079
	12,991	12,209

- Prop died among Vit A =  $101/12,991 \sim 0.8\%$
- Prop died among Plcbo =  $130/12,209 \sim 1.1\%$

# Relative Risk

	Vitamin A	Placebo
Died	101	130
Survived	12,890	12,079
	12,991	12,209

- Relative Risk (RR)

$$RR = \frac{\text{risk of disease in exposed group}}{\text{risk of disease in unexposed group}}$$

- $RR = 0.8/1.1 = 0.73$
- Children taking Vitamin A were about 27% less likely to die than those not taking Vitamin A.
- Difference due to sampling variability, or is Vitamin A protective for mortality?

# Additional Summaries: Odds Ratio

- Case-Control Study: Cat Scratch Disease
  - Swollen lymph nodes
  - Associated with cat having fleas?
  - 56 cases; 56 controls

	Fleas	No Fleas
Disease	32	24
No Disease	4	52
	36	76

- Prop CSD among Fleas =  $32/36 = 89\%$
- Prop CSD among No Fleas =  $24/76 = 32\%$

# Problem With Relative Risk

	Fleas	No Fleas
Disease	32	24
No Disease	4	52
	36	76

- Relative Risk (RR)

$$RR = \frac{\text{risk of disease in exposed group}}{\text{risk of disease in unexposed group}}$$

- RR = ???
- We selected controls, so no info on *risk*
- Risk of CSD among owners of cats w/ fleas  $\neq 32/36 = 89\%$ !
- Alternative: odds ratio (OR)

# Odds Ratio

	Fleas	No Fleas
Disease	32	24
No Disease	4	52
	36	76

- Odds Ratio (OR)

$$\text{OR} = \frac{\text{odds of disease in exposed group}}{\text{odds of disease in unexposed group}}$$

- $\text{OR} = (32/4)/(24/52) = 17.3$
- The odds of disease in the exposed group are about 17 times those for the unexposed group.

# Notes on the Odds Ratio

- $0 < OR < \infty$ 
  - $OR > 1$  exposure has promoting effect for disease
  - $OR < 1$  exposure has protective effect for disease
  - $OR = 1$  no effect of exposure on disease
- If a disease is rare,  $OR \sim RR$
- OR works for any study you can compute an RR for
- For mortality and Vitamin A =  
 $OR = (101/12,890)/(130/12,079) = 0.73 (= RR!)$

# ***Comparing a Binary Categorical Variable Between***

## ***Two Independent Groups***

### ***Parametric Methods (Comparing Proportions, or Testing $OR=1$ )***

# Hodgkin's Example I

- Compare proportion Hodgkin's for groups w/ and w/o tonsillectomies
- Suspect that tonsils are protective
- Case-control study (controls unmatched)
- Vianna, Greenwald, and Davies (1971)

<u>Tonsillectomy</u>	<u>Hodgkin's</u>
yes	yes
yes	no
...	
no	no

# Hodgkin's Example I Con't

	<u>Tonsillectomy</u>	<u>No Tonsillectomy</u>
Hodgkins	67	34
Control	43	64

- prop of tonsillectomy among cases =  $67/101 = 66\%$
- prop of tonsillectomy among controls =  $43/107 = 40\%$
- OR of Hodgkin's comparing tonsillectomy group to group w/ tonsils  
 $= (67/43) / (34/64) = 2.93$

# Null Hypotheses

- Two hypothesis tests

$H_0$ : no association

$H_a$ : assoc btw Hodgkin's & tonsillectomy

Or (in some cases) equivalently

$H_0$ : prop of tonsillectomy in cases ( $p_1$ ) =  
prop of tonsillectomy in controls ( $p_2$ )

$H_a$ : proportion of tonsillectomy in cases ( $p_1$ )  $\neq$   
prop of tonsillectomy in controls ( $p_2$ )

- Pearson's chi-squared for association is equivalent to normal theory method for testing difference in proportions (pooled version)

# Null Hypotheses

- Two hypothesis tests

$H_0$ : no association

$H_a$ : assoc btw Hodgkin's & tonsillectomy

(e.g. OR = 1)

(e.g. OR  $\neq$  1)

Or (in some cases) equivalently

$H_0$ :      prop of tonsillectomy in cases ( $p_1$ ) =  
              prop of tonsillectomy in controls ( $p_2$ )

$H_a$ :      proportion of tonsillectomy in cases ( $p_1$ )  $\neq$   
              prop of tonsillectomy in controls ( $p_2$ )

- Pearson's chi-squared for association is equivalent to normal theory method for testing difference in proportions (pooled version)

# Pearson's Chi-squared test

- Compute expected cell counts if independent
- Observed:

	Tonsillectomy	No Tonsillectomy	
Hodgkins	67	34	101
Control	43	64	107
	110	98	208

- Expected:

	Tonsillectomy	No Tonsillectomy	
Hodgkins	53.4	47.6	101
Control	56.6	50.4	107
	110	98	208

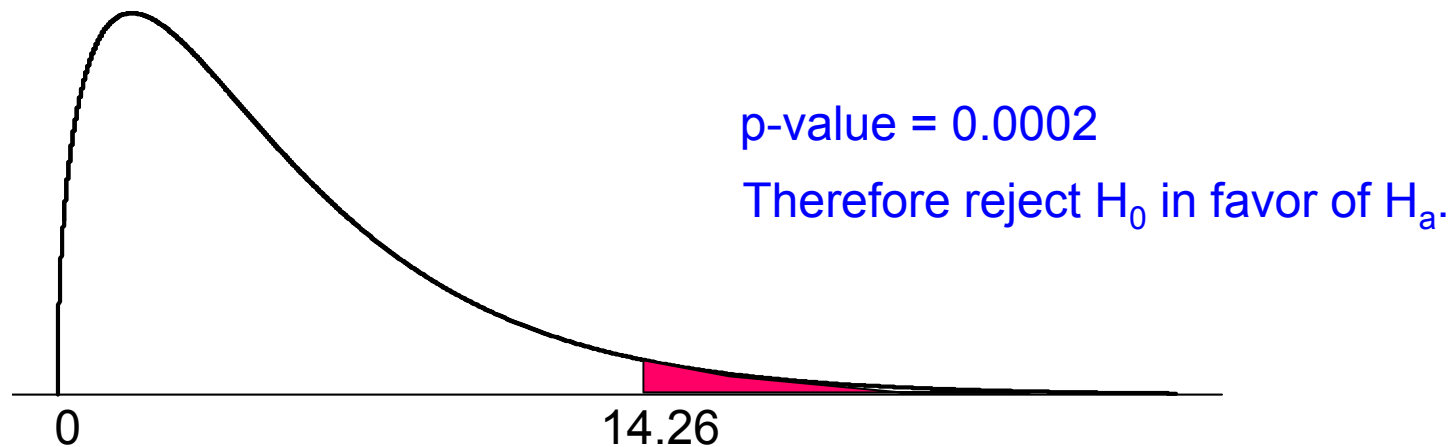
$$\text{Expected count} = \frac{\text{row total} \times \text{column total}}{\text{total}}$$

# Chi-squared statistic

- Compute Chi-squared statistic, based on weighted squared differences  
$$\sum (\text{expected} - \text{observed})^2 / \text{expected}$$
- Our chi-squared statistic will have a Chi-squared distribution with 1 degree of freedom IF
  - *expected* counts all > 5
  - AND
  - total number of subjects > 20 or 40

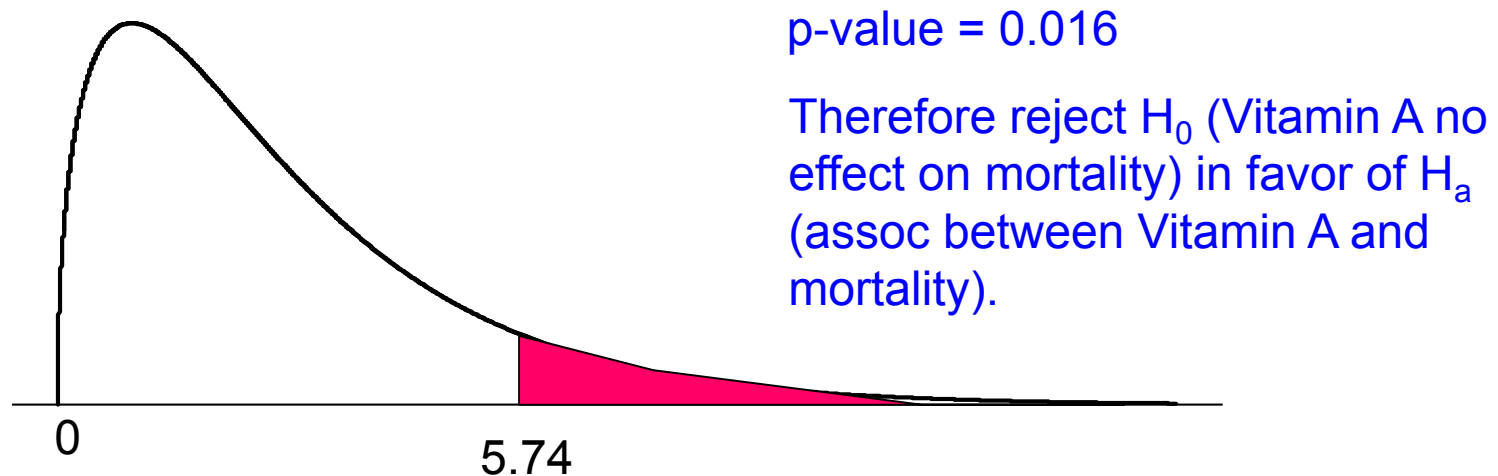
# Chi-squared results for Hodgkin's I

- Chi-squared statistic = 14.46
- Smallest expected cell is ~ 47
- Have more than 20 subjects
- Therefore our statistic has Chi-squared 1 distribution



# Chi-squared results for Vitamin A

- Chi-squared statistic = 5.74
- Smallest expected cell is ~ 112
- Have (way) more than 20 subjects
- Therefore our statistic has Chi-squared 1 distribution



# Notes on Chi-Squared Distribution

- Many different applications of Chi-squared
  - Testing categorical vars between two groups
    - 2 x k tables
    - Similar to Chi-squared procedures for 2 x 2 tables
- Degrees of freedom depend on application
  - 2 x 2 contingency tables always 1 df
- Yates continuity correction (not recommended)

# ***Comparing a Binary Categorical Variable Between***

## ***Two Independent Groups***

### ***Nonparametric Methods (Comparing Proportions)***

# Alternative to Chi-squared

- What if
  - Expected cell count(s)  $\leq 5$ ?
  - Total subjects  $\leq 20$ ?
- Traditional advice: Fisher's Exact Test
  - Nonparametric Alternative to Chi-squared

# Fisher's Exact Test

- Without changing row or column totals, create all other possible tables

- Observed

	Tonsillectomy	No Tonsillectomy	
Hodgkins	67	34	101
Control	43	64	107
	110	98	208

- Another possibility

	Tonsillectomy	No Tonsillectomy	
Hodgkins	30	71	101
Control	80	27	107
	110	98	208

# Fisher's Exact Test

- What percent of tables have association as strong or stronger than ours?

Generated table:

	Tonsillectomy	No Tonsillectomy	
Hodgkins	30	71	101
Control	80	27	107
	110	98	208

OR = 0.14

p-value = 0.0002

# Chi-squared vs. Fisher's Exact

- Fisher's Exact Test computationally intensive
- Fisher's Exact Test and the Chi-squared test will give (approximately) the same p-values *if* Chi-squared test valid
- OLD ADVICE:
  - For small-medium data sets use Fisher's Exact
  - For large samples (1000s) use Chi-squared

# Chi-squared vs. Fisher's Exact

- NEW ADVICE
  - Many times, Pearson's chi-squared and Fisher's exact test will give “the right” value
  - But, they condition on marginal counts that may not be fixed by design
    - e.g. numbers of Hodgkin's and controls were (approximately) fixed, but the tonsillectomy counts were not
  - Exact unconditional tests (e.g. Fisher-Boschloo) are optimal – see your statistician

# ***Comparing a Binary Categorical Variable Between***

## ***Two Paired Groups***

### ***Parametric Methods***

***(Comparing proportions of discordant pairs)***

# Hodgkin's Example II

- Again Hodgkin's and tonsillectomies
- Case-control study (controls matched)
  - 85 Hodgkin's who had sibling w/in 5 yrs age and same sex
  - sibling was matched control
- Johnson & Johnson (1972)

<u>Tonsillectomy(Hodgkin's)</u>	<u>Tonsillectomy(Sibling)</u>
yes	yes
yes	no
...	
no	no

# WRONG Analysis Hodgkin's II

- Create 2 x 2 contingency table

	Tonsillectomy	No Tonsillectomy
Hodgkins	41	44
Control	33	52

- Pearson's chi-squared test
- Chi-squared statistic = 1.53
- p-value = 0.22
- No evidence tonsillectomy assoc w/ Hodgkin's
- Contradicts earlier study!

# What went wrong?

- Contingency table ignored pairings!

	Tonsillectomy	No Tonsillectomy
Hodgkins	41	44
Control	33	52

- Better contingency table shows pairings

		Sibling	
		Tonsillectomy	No Tonsillectomy
Hodgkin's	Tonsillectomy	37	7
	No Tonsillectomy	15	26

# McNemar's Test Hodgkin's II

- The *concordant pairs* have the same exposure (tonsillectomy, tonsillectomy) or (no, no) & don't tell you anything about the association between exposure & disease.
- Need to look at the *discordant pairs*.

		Sibling	
		No Tonsillectomy	Tonsillectomy
Hodgkin's	No Tonsillectomy	37	7
	Tonsillectomy	15	26

**CORRECTED VERSION**

# McNemar's Test Hodgkin's II

- Compare
  - prop of *pairs* in which sibling ~~did not have~~ had tonsillectomy but patient did not  $7/85 = 8\%$
  - TO
  - prop of *pairs* in which sibling ~~had~~ did not have tonsillectomy but patient did ~~not~~  $15/85 = 17\%$

		Sibling	
		No Tonsillectomy	Tonsillectomy
Hodgkin's	No Tonsillectomy	37	7
	Tonsillectomy	15	26

# McNemar's Test Hodgkin's II

- Compare
  - prop of *pairs* in which sibling did not have tonsillectomy but patient did  $7/85 = 8\%$
  - TO
  - prop of *pairs* in which sibling had tonsillectomy but patient did not  $15/85 = 17\%$

		Sibling	
		No Tonsillectomy	Tonsillectomy
Hodgkin's	No Tonsillectomy	37	7
	Tonsillectomy	15	26

# McNemar's Test Hodgkin's II

- If there was no association between tonsillectomy, we would expect these 7 + 15 pairs to be equally divided between the two cells (so expect 11 in each).
- Is the distribution of the discordant pairs extreme enough for us to reject the null?
- $p = 0.09$
- Less doubt about previous results?

		Sibling	
		No Tonsillectomy	Tonsillectomy
Hodgkin's	No Tonsillectomy	37	7
	Tonsillectomy	15	26

# McNemar's Test

- Confession: McNemar's also uses a chi-squared distribution, but different than previous calcs
- What to do if # discordant pairs  $\leq 10$ ?
- Nonparametric versions of McNemar
  - Using exact binomial procedures
  - Sometimes called "Exact McNemar"
  - Use computer program, talk to statistician

# ***Testing for Agreement***

***Categorical Variable (Kappa)***

***Continuous Variable (Intraclass Correlation)***

# Agreement on Categorical Variable

- Two radiologists examine 100 mammograms
- Rate by 'normal' or 'not normal'
- Want measure of *agreement*, not *association*
- Chi-squared and McNemar for *association*
- Use *Kappa* statistic

# Agreement on Categorical Variable

		Radiologist 2	
		Normal	Not Normal
Radiologist 1	Normal	35	10
	Not normal	20	35

- Observed concordance =  $(35 + 35)/100$   
= 70%
- Expected concordance (if independent) =  
 $45/100 * 55/100 + 55/100 * 45/100 = 49.5\%$
- Kappa =  $(\text{obs conc} - \text{exp conc}) / (1 - \text{exp conc})$   
= 0.406

# Agreement on Categorical Variable

- $0 \leq \text{Kappa} \leq 1$
- Reproducibility Guidelines

$\text{Kappa} > 0.75$	Excellent
$0.40 \leq \text{Kappa} \leq 0.75$	Good
$0 \leq \text{Kappa} < 0.40$	Marginal/Poor

- Also for variables with  $> 2$  categories

# Agreement on Continuous Variable

- Blood pressure highly variable w/in person
- How does measurement of BP at a single doctor's visit relate to 'true' BP (i.e. average BP over a period of time)?
- Compute intraclass correlation, or reliability coefficient ( $\rho$ )

# Agreement on Continuous Variable

- (Random effects ANOVA)
- Reproducibility interpretation

$\rho < 0.40$       Poor

$0.40 \leq \rho < 0.75$       Fair/Good

$\rho \geq 0.75$       Excellent

- $\rho$  for BP at single visit and “true” BP = 0.89
- $\rho$  for average BP at 3 visits and “true” = 0.92

# Agreement on Continuous Variable

- Cannot use `regular'/Pearson correlation!

Person	1	2	3	4	5	6	7	8	9	10
Single BP	80	81	82	83	84	85	86	87	88	89
`True' BP	90	91	92	93	94	95	96	97	98	99

- Pearson correlation = 1
- But agreement is poor; rho = 0.15

# Next Time

***Linear Regression***  
***Logistic Regression***

# Useful References

- *Intuitive Biostatistics*, Harvey Motulsky, Oxford University Press, 1995
  - Highly readable, minimally technical
- *Practical Statistics for Medical Research*, Douglas G. Altman, Chapman & Hall, 1991
  - Readable, not too technical
- *Fundamentals of Biostatistics*, Bernard Rosner, Duxbury, 2000
  - Useful reference, somewhat technical

# Contact BCC

- Please go to our web site at:

<http://www.medschool.northwestern.edu/depts/bcc/index.htm>

- Fill out our online request form for further collaboration!