The Binomial Distribution

binomial distribution is a discrete distribution.

Binomial Experiment

- A binomial experiment has the following properties:
  - experiment consists of n identical and independent trials
  - each trial results in one of two outcomes: success or failure
    - P(success) = p
    - P(failure) = q = 1 - p for all trials
  - The random variable of interest, X, is the number of successes in the n trials.
  - X has a binomial distribution with parameters n and p

What is P(x) for binomial?

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

Mean and Standard Deviation

- The mean (expected value) of a binomial random variable is

$$\mu = np$$

- The standard deviation of a binomial random variable is

$$\sigma = \sqrt{npq}$$

Example

- Random Guessing; n = 100 questions.
  - Probability of correct guess; p = 1/4
  - Probability of wrong guess; q = 3/4
  - Expected Value = $$\mu = np = 100\left(\frac{1}{4}\right) = 25$$
    - On average, you will get 25 right.
- Standard Deviation = $$\sigma = \sqrt{npq} = \sqrt{np(1-p)} = \sqrt{100\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)} = 4.33$$

- **Exposure (E)** ≡ an explanatory factor; any potential health determinant; the independent variable.
- **Disease (D)** ≡ the response; any health-related outcome; the dependent variable.
- **Measure of association (syn. measure of effect)** ≡ a statistic that quantifies the relationship between an exposure and a disease.

Risk Difference

Risk Difference (RD) ≡ absolute effect associated with exposure

$$RD = R_1 - R_0$$

where

R_1 ≡ risk in the exposed group

R_0 ≡ risk in the non-exposed group

Risk ratio

$$\frac{\textbf{Numerator}\\ \text{Risk of disease in } \textbf{exposed}}{\textbf{Denominator}\\ \text{Risk of disease in } \textbf{unexposed}} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Risk ratio interpretation

- Ratios > 1.0 indicate rate is **higher** among exposed than unexposed
- Ratios = 1.0 indicate **no** association
- Ratios < 1.0 indicate rate is **lower** among exposed than unexposed

Steps: risk ratio 95% confidence interval

1. Take natural log of risk ratio

   ln (Risk ratio)

2. Estimate standard error (SE)

$$\sqrt{\left(\frac{1}{a}\right) - \left(\frac{1}{a+b}\right) + \left(\frac{1}{c}\right) - \left(\frac{1}{c+d}\right)}$$

Steps: risk ratio 95% confidence interval

3. Estimate upper and lower bounds on log scale
   - 95% confidence interval **upper** bound

     ln(Risk ratio) + 1.96(SE[ln(Risk ratio)])
   - 95% confidence interval **lower** bound

     ln(Risk ratio) - 1.96(SE[ln(Risk ratio)])

Steps: risk ratio 95% confidence interval

4. Exponentiate upper and lower bounds
5. Report and interpret estimate and confidence interval

Example: risk ratio 95% confidence interval

- Measure association between family history of Alzheimer's disease (AD) and incidence of AD among those aged >70
- Random sample of 1,000 individuals aged >70, no symptoms of AD
- Followed for 20 years
- Measure symptoms of AD every year
- No losses to follow-up

Example: risk ratio 95% confidence interval

$$Risk\ ratio = \frac{\left(\frac{50}{350}\right)}{\left(\frac{60}{650}\right)} = 1.548$$

Example: risk ratio 95% confidence interval

1. Take natural log of risk ratio

   ln (Risk ratio) = ln(1.548) = 0.437

2. Estimate standard error (SE)

$$\sqrt{\left(\frac{1}{a}\right) - \left(\frac{1}{a+b}\right) + \left(\frac{1}{c}\right) - \left(\frac{1}{c+d}\right)}$$

$$\text{SE}(\text{Ln}[\text{Risk ratio}]) = \sqrt{\left(\left(\frac{1}{50}\right) - \left(\frac{1}{350}\right) + \left(\frac{1}{60}\right) - \left(\frac{1}{650}\right)\right)} = 0.1796$$

Example: risk ratio 95% confidence interval

3. Estimate upper and lower bounds on log scale
   - 95% confidence interval **upper** bound

     ln(Risk ratio) + 1.96(SE[ln(Risk ratio)])

     0.437 + 1.96(0.1796)
   - 95% confidence interval **lower** bound

     ln(Risk ratio) - 1.96(SE[ln(Risk ratio)])

     0.437 - 1.96(0.1796)

Steps: risk ratio 95% confidence interval

4. Exponentiate upper and lower bounds

$$e^{.789} = 2.20$$

$$e^{.085} = 1.09$$

5. Report and interpret estimate and confidence interval

Individuals >70 in Farrlandia with a family history of AD had 1.55 times the risk of developing AD over 20 years, with a 95% confidence interval for the risk ratio of 1.09 to 2.20.

Comparison of RR and RD

RR ⇒ strength of effect
RD ⇒ effect in absolute terms

Rates (per 100000) of Lung CA & CHD assoc. w/smoking

	Smoker	Nonsmoke	RR	RD
LungCA	104	10	10.40	94
CHD	565	413	1.37	152

Smoking ⇒ Stronger effect for LungCA

Smoking ⇒ Causes more CHD

Odds ratio

Numerator

- Odds of disease in **exposed**

Denominator

- Odds of disease in **unexposed**

Example A: odds ratio

- Odds of ADHD among **exposed**

$$\frac{\left(\frac{300}{5000}\right)}{1-\left[\frac{300}{5000}\right]} = 0.064$$

- Odds of ADHD among **unexposed**

$$\frac{\left(\frac{200}{5000}\right)}{1-\left[\frac{200}{5000}\right]} = 0.042$$

- Odds ratio

$$\frac{0.064}{0.042} = 1.53$$

Example A: odds ratio interpretation

The odds of developing ADHD in the first 10 years of life among those exposed are 1.53 times the odds of disease in the unexposed.

Steps: odds ratio 95% confidence interval

1. Take natural log of odds ratio

   ln (Odds ratio)

2. Estimate standard error (SE)

$$\sqrt{\left(\frac{1}{a}\right)+\left(\frac{1}{b}\right)+\left(\frac{1}{c}\right)+\left(\frac{1}{d}\right)}$$

Steps: odds ratio 95% confidence interval

3. Estimate upper and lower bounds on log scale
   - 95% confidence interval **upper** bound

     ln(Odds ratio) + 1.96(SE[ln(Odds ratio)])
   - 95% confidence interval **lower** bound

     ln(Odds ratio) - 1.96(SE[ln(Odds ratio)])

Steps: odds ratio 95% confidence interval

4. Exponentiate upper and lower bounds
5. Report and interpret estimate and confidence interval

Summary: odds ratio

- Cannot estimate the risk of disease directly when we sample people based on whether they have the disease or not (case control study)
- Can estimate proportion exposed among diseased and non-diseased
  - Estimate odds ratio for exposure
  - Odds ratio for exposure = odds ratio for disease
- If disease is rare in population, the odds ratio approximates the risk ratio from a prospective study

Terminology

For simplicity sake, the terms “risk” and “rate” will be applied to all incidence and prevalence measures.

What do you do when you have multiple levels of exposure?

Compare rates to least exposed "reference" group

	LungCA Rate (per 100,000 person-years)	*RR*
Non-smoker (0)	10	1.0 (ref.)
Light smoker (1)	52	5.2
Mod. smoker (2)	106	10.6
Heavy sm. (3)	224	22.4

$$RR_1 = \frac{R_1}{R_0} = \frac{52}{10} = 5.2$$

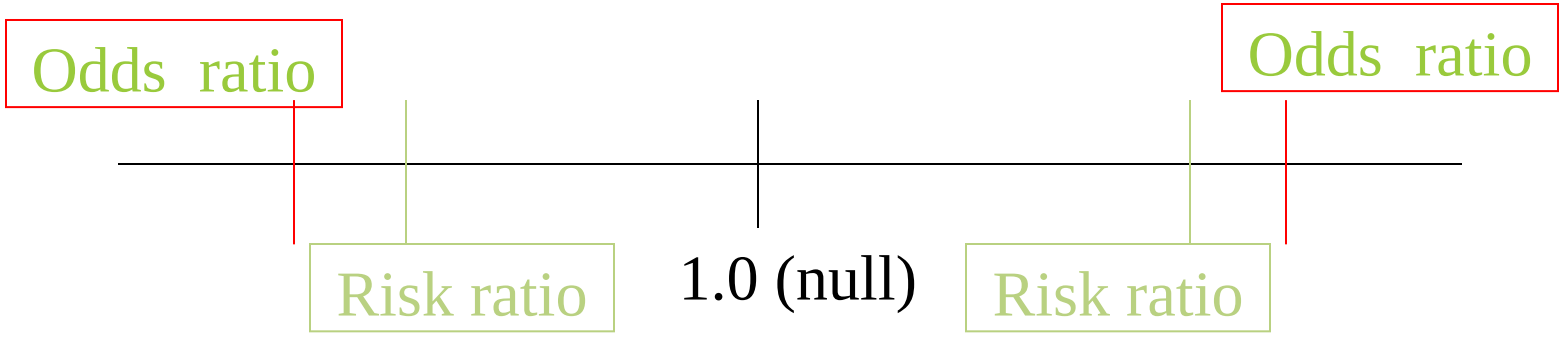
$$RR_2 = \frac{R_2}{R_0} = \frac{106}{10} = 10.6$$

OR versus RR Key Messages

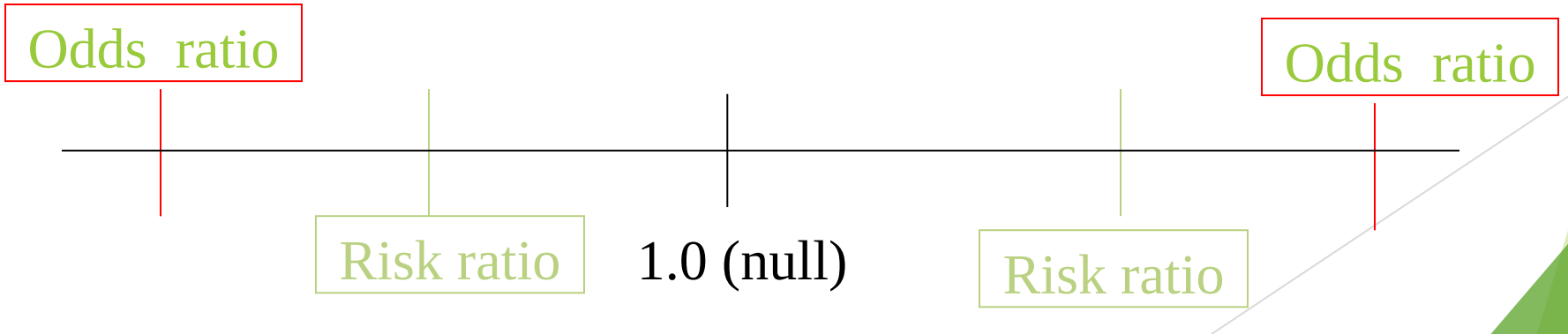
- Odds and Odds Ratios are difficult to conceptualize but statisticians prefer them in some situations because of their mathematical properties
- Odds Ratios always exaggerate the relative risk, but when baseline risk is low (e.g. <10%), the OR approximates the relative risk
- Relative Risk is a more intuitive measure and is becoming more common in medical literature

The odds ratio vs. the risk ratio

Rare Outcome

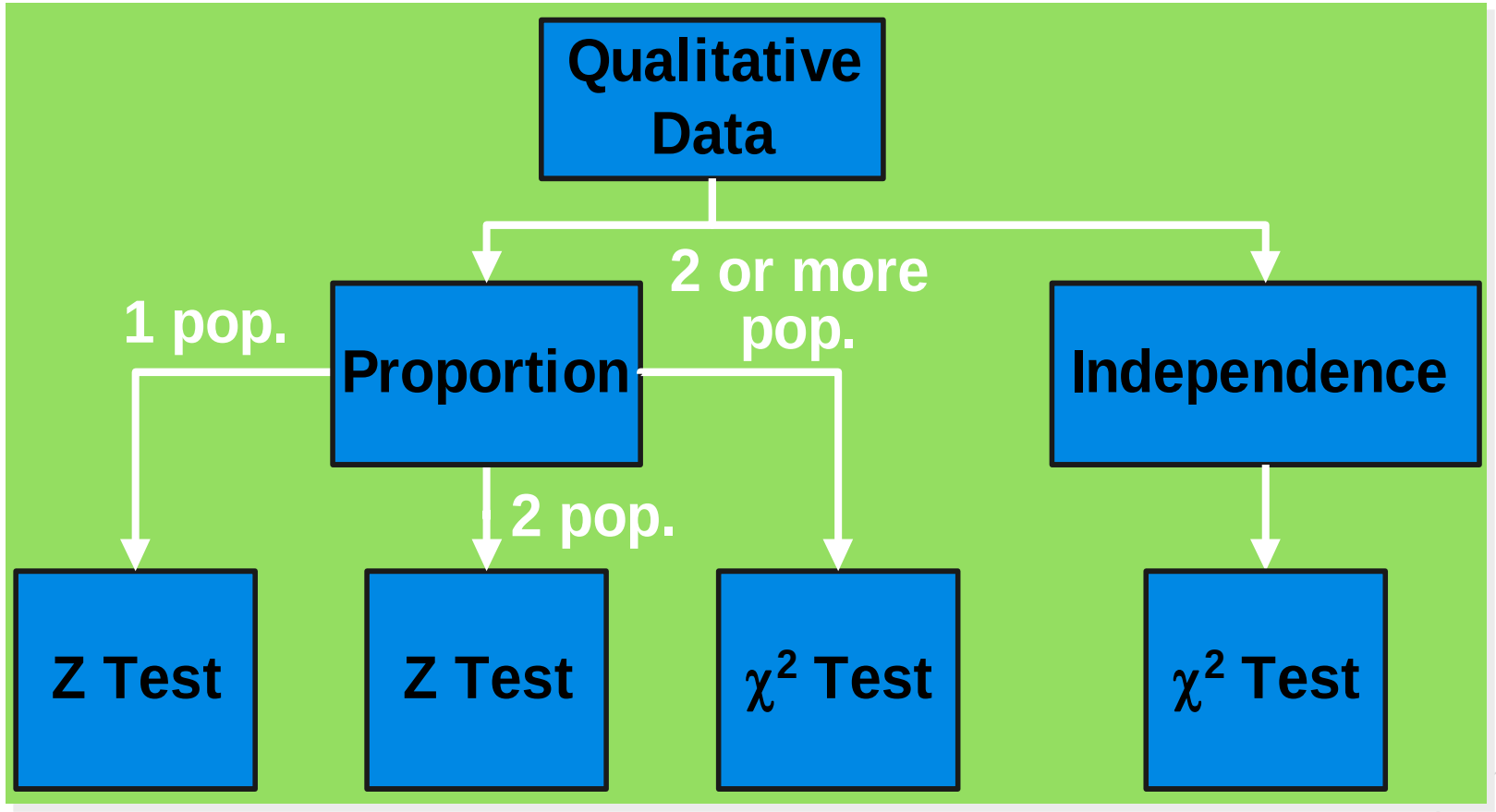


Common Outcome



Chi-Square Applications

Hypothesis Tests Qualitative Data



Z Test for Differences in Two Proportions

Hypotheses for Two Proportions

Hypothesis	Research Questions		
	No Difference Any Difference	Pop 1 ≥ Pop 2 Pop 1 < Pop 2	Pop 1 ≤ Pop 2 Pop 1 > Pop 2
H_0	$p_1 - p_2 = 0$	$p_1 - p_2 \geq 0$	$p_1 - p_2 \leq 0$
H_a	$p_1 - p_2 \neq 0$	$p_1 - p_2 < 0$	$p_1 - p_2 > 0$

Z Test for Difference in Two Proportions

1. Assumptions
   - Populations Are Independent
   - Populations Follow Binomial Distribution
   - Normal Approximation Can Be Used for large samples (All Expected Counts ≥ 5)
2. Z-Test Statistic for Two Proportions

$$Z \cong \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Goodness-of-Fit Tests

- Does the distribution of sample data resemble a specified probability distribution

- Hypotheses:
  - **H_0:** π_i = values expected **H_1:** $\pi_i \neq$ values expected

    where $\sum \pi_j = 1$.

Goodness-of-Fit Tests

- Test Statistic:

$$\chi^2 = \sum \frac{(O_j - E_j)^2}{E_j}$$

where O_j = Actual number observed in each class

E_j = Expected number, $p_j \bullet n$

Goodness-of-Fit: An Example

- In a study of vehicle ownership, it has been found that 13.5% of U.S. households do not own a vehicle, with 33.7% owning 1 vehicle, 33.5% owning 2 vehicles, and 19.3% owning 3 or more vehicles. The data for a random sample of 100 households in a resort community are summarized below. At the 0.05 level of significance, can we reject the possibility that the vehicle-ownership distribution in this community differs from that of the nation as a whole?

# Vehicles Owned	# Households
0	20
1	35
2	23
3 or more	22

Goodness-of-Fit: An Example

# Vehicles	O_i	E_i	$[O_i - E_i]^2 / E_i$
0	20	13.5	3.1296
1	35	33.7	0.0501
2	23	33.5	3.2910
3+	22	19.3	0.3777
		Sum =	6.8484

I. **H_0:** $\pi_0 = 0.135$, $\pi_1 = 0.337$, $\pi_2 = 0.335$, $\pi_{3+} = 0.193$

Vehicle-ownership distribution in this community is the same as it is in the nation as a whole.

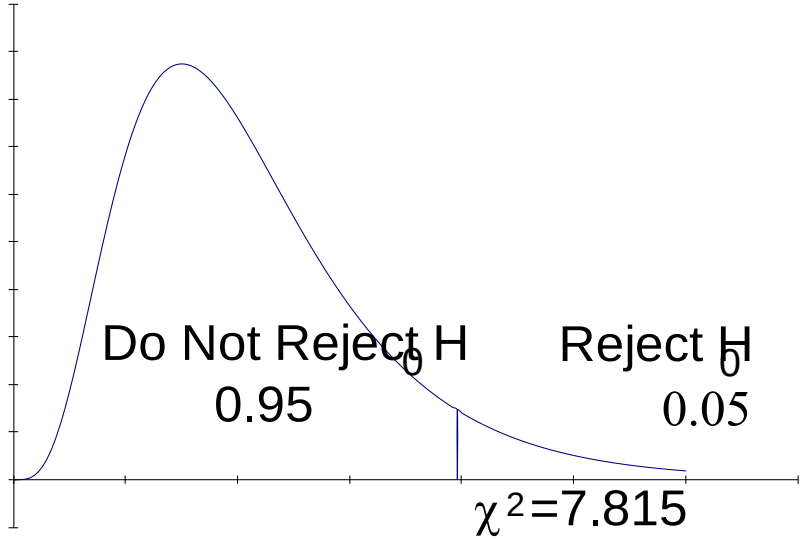
H_1: At least one of the proportions does not equal the stated value. Vehicle-ownership distribution in this community is not the same as it is in the nation as a whole.

Goodness-of-Fit: An Example

II. Rejection Region:

$\alpha = 0.05$

$df = k - 1 = 4 - 1 = 3$



III. Test Statistic:

$\chi^2 = 6.8484$

IV. Conclusion: Since the test statistic of $\chi^2 = 6.8484$ falls below the critical value of $\chi^2 = 7.815$, we do not reject H_0 with at least 95% confidence.

V. Implications: There is not enough evidence to show that vehicle ownership in this community differs from that in the nation as a whole.

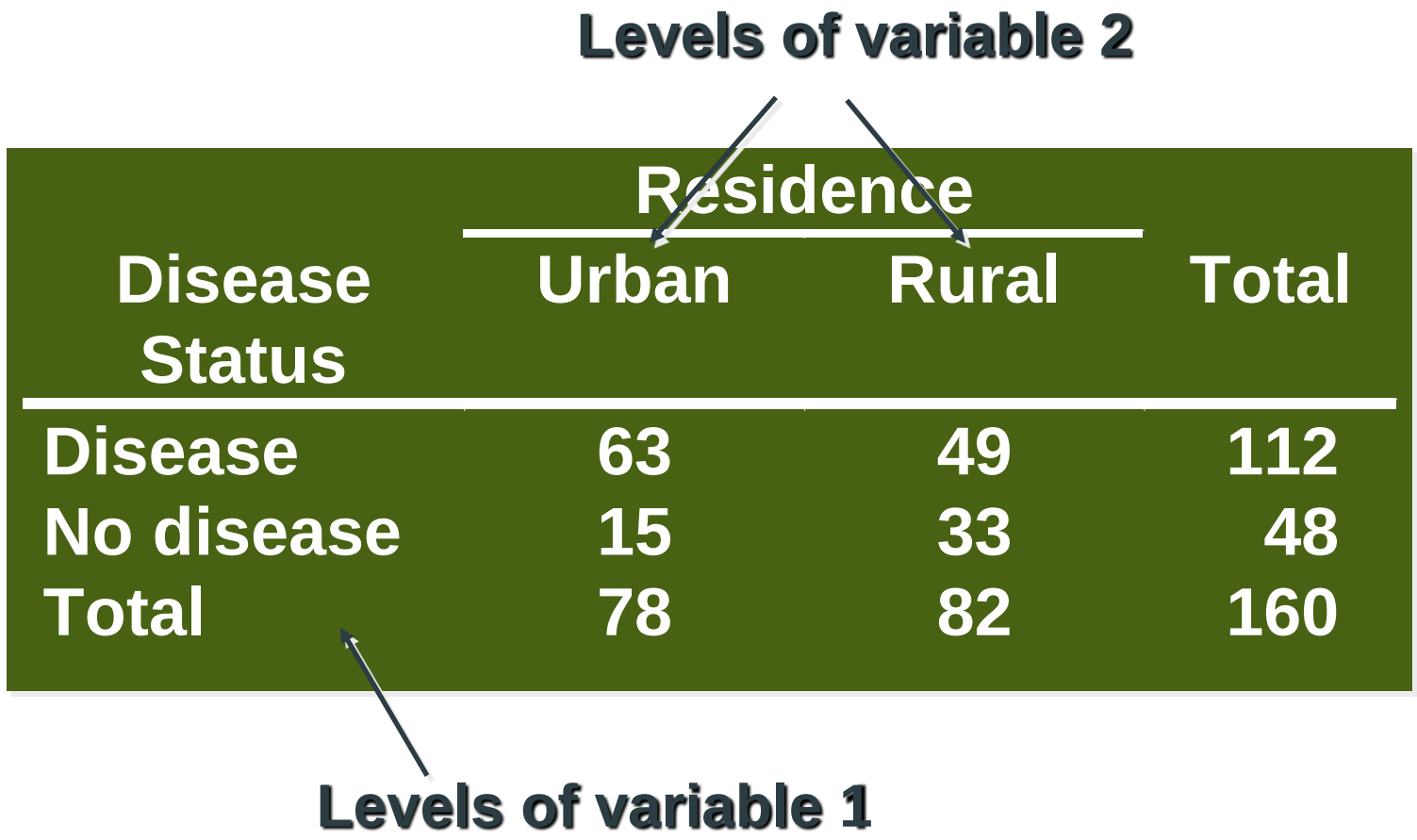
χ^2 Test of Independence

1. Shows If a Relationship Exists Between 2 Qualitative Variables, but does **Not** Show Causality
2. Assumptions

   Multinomial Experiment

   All Expected Counts ▫ 5
3. Uses Two-Way Contingency Table

χ^2 Test of Independence Contingency Table



χ^2 Test of Independence Hypotheses & Statistic

1. Hypotheses
   - H_0: Variables Are Independent
   - H_a: Variables Are Related (Dependent)

χ^2 Test of Independence Example on HIV

- You randomly sample **286** sexually active individuals and collect information on their HIV status and History of STDs. At the **.05** level, is there evidence of a **relationship**?

	HIV		
STDs Hx	No	Yes	Total
No	84	32	116
Yes	48	122	170
Total	132	154	286

χ^2 Test of Independence Solution

✓ $E(n_{ij}) \geq 5$ in all cells

$\frac{116 \times 132}{286}$

	HIV				
	No		Yes		
STDs HX	Obs.	Exp.	Obs.	Exp.	Total
No	84	53.5	32	62.5	116
Yes	48	78.5	122	91.5	170
Total	132	132	154	154	286

$\frac{154 \times 116}{286}$

$\frac{170 \times 132}{286}$

$\frac{170 \times 154}{286}$

χ^2 Test of Independence Solution

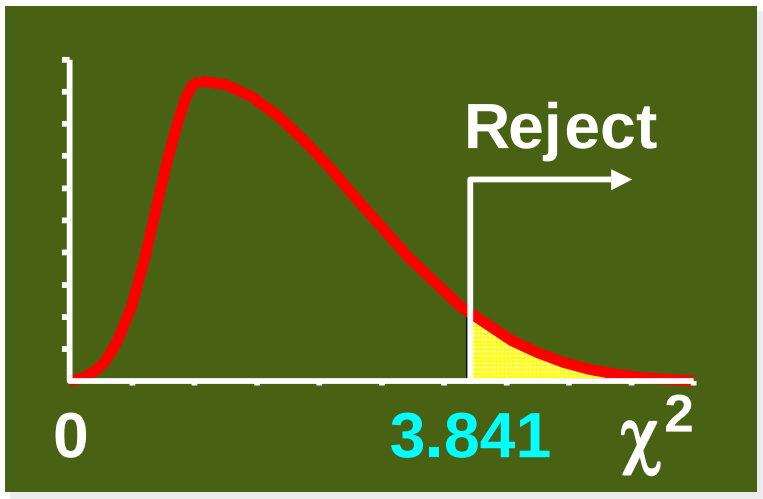
H_0: No Relationship

H_a: Relationship

$\alpha = .05$

$df = (2 - 1)(2 - 1) = 1$

Critical Value(s):



Test Statistic:

$\chi^2 = 54.29$

Decision:

Reject at $\alpha = .05$

Conclusion:

There is evidence of a relationship

Fisher's Exact Test

- **Fisher's Exact Test** is a test for independence in a 2 X 2 table. It is most useful when the total sample size and the expected values are small. The test holds the marginal totals fixed and computes the hypergeometric probability that n_{11} is at least as large as the observed value
- Useful when E(cell counts) < 5.

Fisher's Exact Test

- Example: 2x2 table with cell counts a, b, c, d. Assuming marginal totals are fixed:

  M1 = a+b, M2 = c+d, N1 = a+c, N2 = b+d.

  for convenience assume N1<N2, M1<M2.

  possible value of a are: 0, 1, …min(M1,N1).

- Probability distribution of cell count a follows a hypergeometric distribution:

  N = a + b + c + d = N1 + N2 = M1 + M2

  - Pr (x=a) = N1!N2!M1!M2! / [N!a!b!c!d!]

- Fisher exact test is based on this hypergeometric distr.

Fisher's Exact Test Example

HIV Infection

Hx of STDs

	yes	no	total
yes	3	7	10
no	5	10	15
total	8	17	

- Is HIV Infection related to Hx of STDs in Sub Saharan African Countries? Test at 5% level.

Fisher's Exact Test Example

- Probability of observing this specific table given fixed marginal totals is

  Pr (3,7, 5, 10) = 10!15!8!17!/[25!3!7!5!10!]

  = 0.3332
- Note the above is not the p-value. Why?
- Not the accumulative probability, or not the tail probability.
- Tail prob = sum of all values (a = 3, 2, 1, 0).

Fisher's Exact Test Example

Pr (2, 8, 6, 9) = 10!15!8!17!/[25!2!8!6!9!]

= 0.2082

Pr (1, 9, 7, 8) = 10!15!8!17!/[25!1!9!7!8!]

= 0.0595

Pr (0,10, 8, 7) = 10!15!8!17!/[25!0!10!8!7!]

= 0.0059

Pearson Chi-squares test Yates correction

- Pearson Chi-squares test

  $\chi^2 = \sum_i (O_i - E_i)^2/E_i$ follows a chi-squares distribution with df = (r-1)(c-1)

  if $E_i \geq 5$.

- Yates correction for more accurate p-value

  $\chi^2 = \sum_i (|O_i - E_i| - 0.5)^2/E_i$

  when O_i and E_i are close to each other.

Chi square test for trend

- 1 variable is binary and the other is ordered categorical and we want to assess whether the association between the variables follows a trend.

Chi square test for trend

$$U = \Sigma(dx) - \frac{O}{N}\Sigma(nx) \quad \text{and} \quad V = \frac{O(N-O)}{N^2(N-1)}[N\Sigma(nx^2) - (\Sigma nx)^2]$$

$$\chi^2_{\text{trend}} = \frac{U^2}{V}, \text{d.f.} = 1$$

Where
dx= the product of the observed number and the exposure group score
nx= the product of the total and the exposure group score
nx2= the product of the total and the square of exposure group score

Chi square test for trend example

	Triceps skinfold group			
Age at menarche	Small	Intermediate	Large	Total
< 12 years (D)	15 (8.8%)	29 (12.8%)	36 (19.4%)	80
12+ years (H)	156 (91.2%)	197 (87.2%)	150 (80.6%)	503
Total	171 (100%)	226 (100%)	186 (100%)	583
Exposure group score (x)	0	1	2	
Odds of early menarche	0.10 (0.06 to 0.16)	0.15 (0.10 to 0.22)	0.24 (0.17 to 0.35)	
Log odds	−2.34 (−2.87 to −1.81)	−1.92 (−2.31 to −1.53)	−1.43 (−1.79 to −1.06)	

In this example difference log odds between (small & intermediate) groups is not equal to (intermediate and large) groups. It seems there is a trend.

Chi square test for trend example

$$\Sigma(dx) = 15\times 0 + 29\times 1 + 36\times 2 = 101$$
$$\Sigma(nx) = 171\times 0 + 226\times 1 + 186\times 2 = 598$$
$$\Sigma(nx^2) = 171\times 0 + 226\times 1 + 186\times 4 = 970$$
$$O = 80,\ N = 583,\ N - O = 503$$
$$U = 101 - \left(\frac{80}{583}\times 598\right) = 18.9417$$
$$V = \left(\frac{80\times 503}{583^2\times 582}\right)\times(583\times 970 - 598^2) = 42.2927$$

$$\chi^2_{\text{trend}} = \frac{(18.9417)^2}{42.2927} = 8.483, \quad \text{d.f.} = 1, \quad P = 0.0036.$$