

# Chapter 5

---

## *Nonparametric Effect Size Indexes*

Destiny is no matter of chance. It is a matter of choice.  
It is not a thing to be waited for, it is a thing to be achieved.

—William Jennings Bryan

### Overview

- Categorical outcomes
- Effect size indexes for 2×2 tables
- Effect size estimation for larger tables
- Sensitivity, specificity, and predictive value

# Categorical outcomes

---

- Levels of a categorical variable are either unordered or ordered
- *Unordered categories* do not imply a rank order (e.g., geographical regions, marital status)
- *Ordered categories*—also called *multilevel ordinal categories*—imply a rank order (e.g., the Likert scale *agree, uncertain, disagree*)

# Categorical outcomes

---

- Darlington (1996) describes specialized methods for ordered categories, but they are not as well developed or known as for unordered categories
- One alternative for ordered categories is to rescale them to interval data and then apply methods for continuous outcomes
- Another is to collapse multilevel categories into two meaningful, mutually exclusive outcomes and then apply methods for two unordered categories (i.e., a dichotomous outcome)
- Only methods for unordered categories are covered here

## Effect size indexes for 2×2 tables

---

- Risk difference:

- ✓  $p_T$  = proportion of treated cases with a negative outcome
- ✓  $p_C$  = proportion of untreated (control) cases with a negative outcome
- ✓ Sample risk difference,  $RD = p_C - p_T$ , estimates  $\pi_C - \pi_T$ , the population difference in risk between treated and untreated cases
- ✓ Easy to interpret—for example,  $RD = .20$  indicates a 20% higher risk for the negative outcome among untreated cases compared with treated cases

## Effect size indexes for 2×2 tables

---

- Risk difference:

- ✓ Problem with RD: range depends on values of  $\pi_C$  and  $\pi_T$ , so RD may not be comparable across samples where population risk rates are different
- ✓ Traditional confidence intervals for  $\pi_C - \pi_T$  are based on the following approximate standard error in large samples:

$$s_{RD} = \sqrt{\frac{p_C (1 - p_C)}{n_C} + \frac{p_T (1 - p_T)}{n_T}}$$

## Effect size indexes for 2×2 tables

---

- Risk ratio:
  - ✓ Sample *risk ratio* is  $RR = p_C/p_T$
  - ✓ RR measures the difference in the proportionate risk for the negative outcome among for untreated versus treated cases
  - ✓ RR estimates the parameter  $\pi_C/\pi_T$
  - ✓ Also easy to interpret—for example,  $RR = 1.50$  indicates a risk for the negative outcome 1½ times higher among untreated cases

## Effect size indexes for 2×2 tables

---

- Risk ratio:

- ✓ Problem with RR: range depends on the value of its denominator
- ✓ Another is that only the finite interval 0 to 1.00 indicates lower risk in the group represented in the numerator, but the interval from 1.00 to infinity is theoretically available for describing higher risk in the other group
- ✓ Can be handled by analyzing logarithm transformations of RR
- ✓ Approximate standard error in logarithm units:

$$s_{\ln(RR)} = \sqrt{\frac{1 - p_C}{n_C p_C} + \frac{1 - p_T}{n_T p_T}}$$

## Effect size indexes for 2×2 tables

---

- Odds ratio:
  - ✓ Sample odds ratio (OR) is the ratio of the within-groups odds
  - ✓  $p_T/(1 - p_T)$  = odds for negative outcome in the treatment group
  - ✓  $p_C/(1 - p_C)$  = odds for negative outcome in the control group
  - ✓ For example, if  $p_C = .80$ , the odds of the negative event among untreated cases are  $.80/.20$ , or 4 to 1
  - ✓ OR measures the proportionate difference in odds for untreated versus treated cases

## Effect size indexes for 2×2 tables

---

- Odds ratio:
  - ✓ OR estimates the parameter  $\omega = \Omega_C/\Omega_T$
  - ✓  $\Omega_C = \pi_C/(1 - \pi_C)$  and  $\Omega_T = \pi_T/(1 - \pi_T)$ , which are the odds for negative outcomes in, respectively, the untreated and treated populations
  - ✓ Less intuitive to interpret, but OR has the best overall statistical properties

## Effect size indexes for 2×2 tables

---

- Odds ratio:

- ✓ OR can be estimated in

1. prospective studies
2. studies that randomly sample from exposed and unexposed populations
3. retrospective studies where groups are first formed based on the presence or absence of a disease before their exposure to a supposed risk factor (Fleiss, 1994)

## Effect size indexes for 2×2 tables

---

- Odds ratio:
  - ✓ OR shares with RR the property that the finite interval 0 to 1.00 indicates lower risk in the group represented in the numerator, but the interval from 1.00 to infinity describes higher risk for other group
  - ✓ Also dealt with by analyzing logarithm transformations
  - ✓ Approximate standard error in logarithm units:

$$S_{\ln(\text{OR})} = \sqrt{\frac{1}{n_C p_C (1 - p_C)} + \frac{1}{n_T p_T (1 - p_T)}}$$

## Effect size indexes for 2×2 tables

---

- Odds ratio:

- ✓ OR can be converted to a standardized mean difference for dichotomous outcomes known as a *logit d*
- ✓ The logistic distribution is approximately normal with a standard deviation that equals  $\pi/\sqrt{3}$ , or about 1.8138:

$$\text{logit } d = \frac{\ln(\text{OR})}{1.8138}$$

- ✓ Logit *d* can be compared with *d* for continuous outcomes for the same two groups
- ✓ See Haddock, Rindskopf, and Shadish (1998) for more information about analyzing OR

## Effect size indexes for 2×2 tables

---

- Phi coefficient:
  - ✓ Pearson correlation between two dichotomous variables is the phi coefficient,  $\hat{\phi}$
  - ✓  $\hat{\phi}$  can be calculated using the standard equation for the Pearson  $r$  if the levels of both dichotomies are coded as 0 or 1
  - ✓ There is also an equation for  $\hat{\phi}$  based on the cell frequencies in a 2×2 table (see Table 5.2)

## Effect size indexes for 2×2 tables

---

- Phi coefficient:
  - ✓  $\hat{\phi}$  may be the least adequate effect size index for a 2×2 table
  - ✓  $\hat{\phi}$  is *margin bound*—its value will change if cell frequencies in any row or column are multiplied by an arbitrary constant (other than 1)
  - ✓ Requires random sampling for correct interpretation
  - ✓ Formula for approximate standard error in large samples is quite complicated—see Fleiss (1994, p. 249)

## Effect size estimation for larger tables

---

- Cramér's  $V$ :

- ✓ Best known measure of association for tables larger than 2×2 is Cramér's  $V$ , an extension of  $\hat{\phi}$
- ✓ The formula is

$$V = \sqrt{\frac{\chi^2((r-1) \times (c-1))}{\min(r-1, c-1) \times N}}$$

The numerator under the radical is the chi-square statistic for the contingency table with degrees of freedom equal to the number of rows ( $r$ ) minus 1 times the number of columns ( $c$ ) minus 1

## Effect size estimation for larger tables

---

- Cramér's  $V$ :

- ✓ For a  $2 \times 2$  table,  $V = |\hat{\phi}|$  (i.e.,  $V$  is an unsigned correlation in this case)
- ✓ For larger tables,  $V$  is not generally a correlation coefficient, so its square cannot be interpreted as a proportion of explained variance
- ✓  $V$  is also a margin-bound measure of association

# Sensitivity, specificity, and predictive value

---

- Sensitivity, specificity, and predictive value (SSPV) framework is better known in medicine
- But it has been applied in clinical psychology and education (e.g., Glaros & Kline, 1988; Kennedy, Willis, & Faust, 1997)
- SSPV deals with situations where a (cheaper) screening test can be evaluated against a (more expensive) diagnostic “gold standard”

## Sensitivity, specificity, and predictive value

---

- “Effect sizes” analyzed in the SSPV framework for the 2×2 table of screening test results (positive–negative) and true status (disorder–no disorder) concern the estimated accuracies of positive and negative test results as a function of disorder base rate
- The base rate of a disorder plays an important but often neglected role in evaluating screening tests (e.g., Medin & Edelson, 1988)

## Sensitivity, specificity, and predictive value

---

- *Sensitivity* is the proportion of screening results from cases with the disorder that are correct (i.e., positive)
- *Specificity* is the proportion of screening results from cases *without* the disorder that are correct (i.e., negative)
- Sensitivity and specificity are determined by the cutting point on the screening test, that is, the score that differentiates a negative (normal) versus a positive (clinical) test result

## Sensitivity, specificity, and predictive value

---

- The ideal screening test is 100% sensitive and 100% specific, but this ideal cannot be achieved when distributions on the screening test from groups with and without the disorder overlap (e.g., Figure 5.1)
- Predictive value is the proportion of correct screening test results:
  - ✓ *Positive predictive value (+PV)*: the proportion of all positive test results that are correct
  - ✓ *Negative predictive value (–PV)*: the proportion of all negative test results that are correct
- Predictive value is affected by sensitivity, specificity, and the base rate of the disorder

# Sensitivity, specificity, and predictive value

---

- Definitions for a 2×2 table (Table 5.4):

		True status	
		Disorder	No disorder
Screening test result	Prediction		
+	Disorder	<i>A</i>	<i>B</i>
-	No disorder	<i>C</i>	<i>D</i>

Statistic	Definition
Sensitivity	$A / (A + C)$
Specificity	$D / (B + D)$
Predictive value	
Positive (+PV)	$A / (A + B)$
Negative (-PV)	$D / (C + D)$
Base rate	$(A + C) / (A + B + C + D)$

## Sensitivity, specificity, and predictive value

---

- Assume constant values of sensitivity and specificity
- In general, +PV decreases as the base rate decreases but –PV increases

*Implication:* For rare conditions, screening tests tend to accurately rule out the disorder but are not very good in detecting it (i.e., there are many false positives)

*This is one reason why it is so hard to accurately predict rare events, such as suicide or violent behavior!*

## Sensitivity, specificity, and predictive value

---

- Again assume constant values of sensitivity and specificity
- In general, +PV increases as the base rate increases but –PV decreases

*Implication:* For common conditions, screening tests tend to accurately detect the disorder but are not very good at ruling it out (i.e., there are many false negative results)

## Sensitivity, specificity, and predictive value

---

- +PV and –PV at two different base rates for a screening test that is 80% sensitive and 70% specific (Table 5.5):

Screening test result	True status			Predictive value	
	Disorder	No disorder	Total	+PV	–PV
Base rate = .10					
+	80	270	350	.23	.97
–	20	630	650		
Total	100	900	1,000		
Base rate = .75					
+	600	75	675	.89	.54
–	150	175	325		
Total	750	250	1,000		

# References

---

- Darlington, R. B. (1996). *Measures for ordered categories*. Retrieved January 9, 2002, from <http://comp9.psyc.cornell.edu/Darlington/crosstab/table5.htm>
- Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245-260). New York: Russell Sage Foundation.
- Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical Psychology, 44*, 1013-1023.
- Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3*, 339-353.
- Kennedy, M. L., Willis, W. G., & Faust, D. (1997). The base-rate fallacy in school psychology. *Journal of Psychoeducational Assessment, 15*, 292-307.
- Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117*, 68-85.