

# Measurement Error

Seattle Epidemiology and  
Biostatistics Summer Session  
June, 2004

## Introduction: measurement

- *Measurement* of exposures, outcomes, other characteristics a key part of most epidemiologic studies
- *Tests or measures* take many forms—e.g.:
  - Response on self-administered questionnaire
  - Answer to interview question
  - Lab result
  - Symptom recorded in medical record
  - Physical finding
  - Diagnosis code in a database

## Measurement error

- Nearly all measures are imperfect
- Quantifying a measure's performance helps in:
  - Choosing among alternative measures for same purpose
  - Interpreting study results

## Reliability

### Reliability

- Degree of agreement between 2+ measurements of same characteristic on same study subject
- Also called *reproducibility* or *consistency*
- Approach to quantifying reliability depends on measurement scale
  - *Categorical*: e.g., presence/absence of disease or exposure
  - *Continuous*: e.g., weight, systolic blood pressure

### Reliability of a binary measure

- Data setup:

	Obs. #2	
Obs. #1	Positive	Negative
Positive	<i>a</i>	<i>b</i>
Negative	<i>c</i>	<i>d</i>

*n*

- **Concordance** =  $\frac{a+d}{n}$
- Can also be expressed as **agreement %**

### Example of concordance (hypothetical data)

- Presence or absence of heart murmur on physical exam:

		Examiner #2:		
		Present	Absent	
Examiner #1:	Present	3	17	20
	Absent	17	63	80
		20	80	100

- Concordance =  $\frac{3+63}{100} = 0.66 = 66\%$

### Pitfall with concordance

But sometimes examiners are bound to agree just by chance, even if no association between their observations:

		Observed		Expected by chance	
		#2		#2	
#1		+	-	+	-
+		3	17	4	16
-		17	63	16	64
		20	80	20	80
		20	80	20	80
		Concordance = 0.66		Concordance = 0.68 (!)	

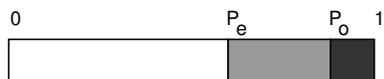
### Kappa

- Accounts for chance agreement

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

$P_o$  = observed concordance

$P_e$  = expected concordance by chance

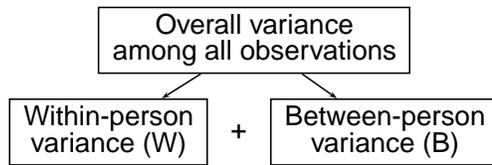


Actual improvement beyond chance:  $\longleftarrow$   $\longrightarrow$

Potential improvement beyond chance:  $\longleftarrow$   $\longrightarrow$

## Intraclass correlation coefficient (ICC)

- Measures reliability for continuous variables
- Statistically:

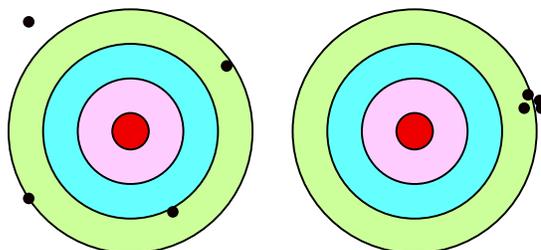


- $ICC = \frac{B}{B+W}$

## Interpretation guidelines for kappa and ICC

Kappa or ICC	Interpretation
> .80	Almost perfect
.61 – .80	Substantial
.41 – .60	Moderate
.21 – .40	Fair
.00 – .20	Slight
< .00	Poor

## Reliability and validity



Poor reliability

Good reliability  
Poor validity

## Validity

### Validity

- Degree to which a measure yields *correct* value of underlying characteristic
- Assessing validity requires error-free *criterion* or “*gold standard*” to which the measure can be compared—e.g.:

Measure	Gold standard
Self-reported body weight	Weight by scale
Clinical diagnosis	Autopsy diagnosis
Self-reported smoking	Biochemical test for smoke metabolites

### Sensitivity and specificity—1

- Many common tests or measures yield a *binary* result—e.g., positive or negative
- Validity assessed by applying *both* the measure in question and the gold standard to a set of study subjects
- Two key components of validity:
  - **Sensitivity:** when gold standard is positive, how often is the test positive?
  - **Specificity:** when gold standard is negative, how often is the test negative?

## Sensitivity and specificity—2

Test result	Gold standard	
	Positive	Negative
Positive	$a$	$b$
Negative	$c$	$d$

Sensitivity =  $\frac{a}{a+c}$

Specificity =  $\frac{d}{b+d}$

## Another common notation

Test result	Gold standard		<i>TP</i> = true positives <i>FP</i> = false positives <i>FN</i> = false negatives <i>TN</i> = true negatives
	Positive	Negative	
Positive	<i>TP</i>	<i>FP</i>	
Negative	<i>FN</i>	<i>TN</i>	

Sensitivity =  $\frac{TP}{TP+FN}$

Specificity =  $\frac{TN}{FP+TN}$

## Example: APOE for Alzheimer's disease

Ferritin	Brain pathology at autopsy ← Gold standard	
	Alzheimer's	Other dementia
Positive	1,142	133
Negative	628	285
Total	1,770	418

Sensitivity =  $\frac{1,142}{1,770} = 0.65$

Specificity =  $\frac{285}{418} = 0.68$

(Source: *N Engl J Med* 1998; 338:506–11)

## Comparing biochemical tests for smoking

Test	For true smoking status*	
	Sensitivity	Specificity
Saliva thiocyanate	81%	71%
Plasma thiocyanate	84%	92%
Saliva cotinine	96%	99%

\*True smoker = self-reported smoking or plasma cotinine >13.7 ng/ml

(Source: *Am J Public Health* 1987; 77:1435–8)

## Tests that yield a numerical result

Often underlying characteristic itself binary, but test produces a numerical result—e.g.:

Characteristic	Test
Depression	Score on depression scale
Iron deficiency anemia	Ferritin level
Congenital hypothyroidism	TSH level

## TSH test for congenital hypothyroidism–1

TSH	Congenital hypothyroidism	
	Present	Absent
>100	8	7
41–100	7	11
31–40	3	29
21–30	1	381
11–20	1	6,405
0–10	0	69,371
Total	20	76,204

- TSH values generally higher in babies with disease
- But distributions overlap

### TSH test for congenital hypothyroidism-2

TSH	Congenital hypothyroidism	
	Present	Absent
>100	8	7
41-100	7	11
31-40	3	29
21-30	1	381
11-20	1	6,405
0-10	0	69,371
Total	20	76,204

- Suppose a *cutoff* value is set at 100
- Numbers in red are *positives*, numbers in black are *negatives*
- Sensitivity =  $\frac{8}{20} = 0.40$
- Specificity =  $\frac{76,197}{76,204} = 0.9999$

### TSH cutoff value = 40

TSH	Congenital hypothyroidism	
	Present	Absent
>100	8	7
41-100	7	11
31-40	3	29
21-30	1	381
11-20	1	6,405
0-10	0	69,371
Total	20	76,204

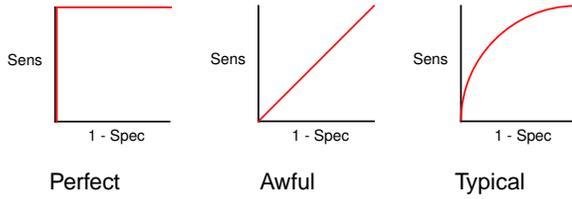
- Now move cutoff down to 40
- Sensitivity =  $\frac{15}{20} = 0.75$
- Specificity =  $\frac{76,186}{76,204} = 0.9998$

### Sensitivity and specificity by cutoff level

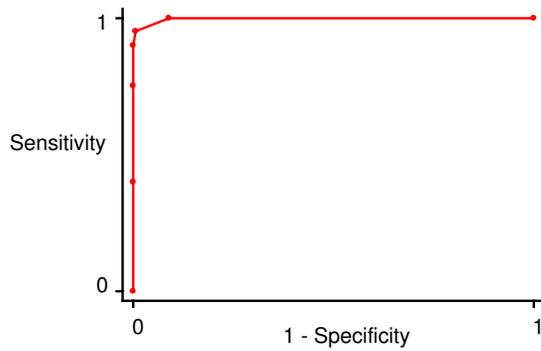
TSH cutoff	Congenital hypothyroidism		Sensitivity	Specificity
	Present	Absent		
100	8	7	.40	.9999
40	7	11	.75	.9998
30	3	29	.90	.9994
20	1	381	.95	.9944
10	1	6,405	1.00	.9103
0	0	69,371	1.00	.0000
Total	20	76,204		

### Receiver Operating Characteristic (ROC) curve

- Shows trade-off between sensitivity and specificity at *all possible* cutoff values
- For historical reasons, sensitivity is plotted against 1 – specificity
- Some noteworthy possible patterns:

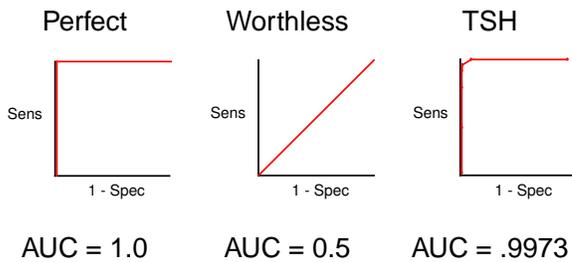


### ROC curve for TSH test



### Area Under the Curve (AUC)

Area Under the Curve (AUC) a commonly used summary measure of test accuracy:



**Introduction to Epidemiologic Methods — Summer, 2004**  
**Discussion Questions: Measurement Error**

1. Delirium is a common but often unrecognized problem among hospitalized older adults. Recognizing it can be important for treatment. A recent study compared assessment of delirium between physicians and nurses on 2,721 paired observations of older inpatients. The results were as follows:

Nurse's rating	Physician's rating		Total
	Delirium	No delirium	
Delirium	46	105	151
No delirium	193	2,377	2,570
Total	239	2,482	2,721

- (a) What percentage of the time did the physicians and nurses agree in their ratings?
  - (b) Given the overall frequency with which physicians and nurses rated the patient as having delirium, how much agreement would one expect just by chance?
  - (c) Calculate kappa. How good is this level of concordance after correcting for chance agreement?
2. Iron deficiency anemia can be diagnosed definitively by the absence of iron stores in a bone-marrow aspirate. However, bone marrow aspiration is a painful, fairly invasive procedure. Rimon, *et al.*, obtained bone-marrow aspirates on 63 older adults on a geriatric unit, 49 of whom proved to have iron deficiency anemia.  
  
On ordinary venous blood samples from the same patients, the investigators also performed three routine tests (serum iron, transferrin saturation, and serum ferritin) and a new test based on a transferrin receptor level assay, which they called the *transferrin receptor–ferritin index* (TR-F index). The correspondence between test results and bone marrow aspirate results was as follows:

Test and result	Iron deficiency anemia, by bone marrow aspiration	
	Present	Absent
Composite of routine tests		
Positive*	8	0
Negative	41	14
TR-F index		
> 1.5	43	1
≤ 1.5	6	13

\*Serum iron, transferrin saturation, and serum ferritin all abnormal

- (a) What were the sensitivity and specificity of the new TR-F index using a cutoff value of 1.5? How did they compare with the sensitivity and specificity of the combined three routine tests?
- (b) The sensitivity of the composite result of three routine tests was very low. Can you suggest a way to increase the sensitivity of the composite result by combining information from the three component tests in a different way?