Introduction to Clinical Biostatistics: Assessing reliability of clinical instruments

Assoc. Prof. Cameron Hurst
cphurst@gmail.com

Faculty of Medicine,
Chulalongkorn University

2nd August, 2560

Measurement reliability

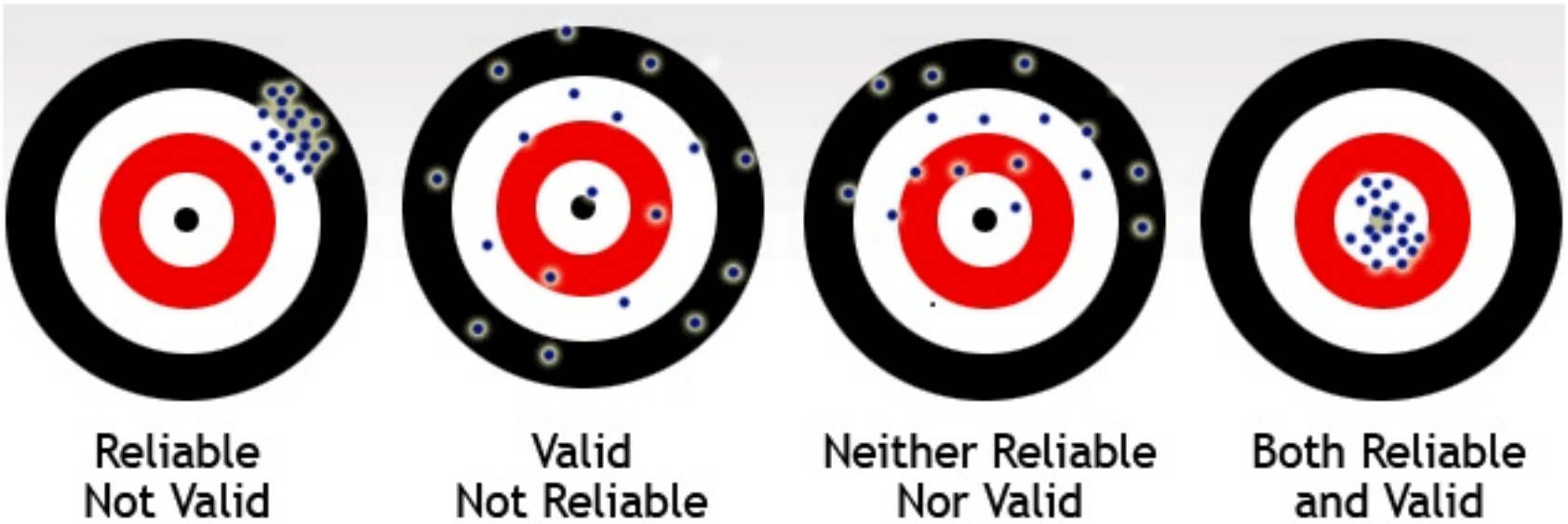
A measurement instruments **RELIABILITY** is:

> *The degree to which a measurement technique can be depended upon to secure* **consistent** *results upon repeated application*

This is not to be confused with the **VALIDITY** of a measurement instrument:

> *The degree to which any measurement approach or instrument succeeds in describing or* **accurately quantifying** *what it is designed to measure*

Reliability Vs. Validity: An illustration



Strictly speaking the second target is not really **valid**. Reliability is really a pre-condition of validity. Only the last figure (RHS) is truly valid.

Assessing reliability in clinical research

In clinical research, we often want to answer questions like:

1. Is a new method (of measurement) good enough to replace an old one?
   - Where the old one might be accurate, but expensive, burdensome, invasive or have nasty side effects.
2. Do measurements made by clinician A agree with those made by clinician B?
3. Are repeated measurements by the same clinician similar?

Analysis involving these **agreement issues** is called **reliability**, **repeatability**, **reproducibility** and/or **consistency** (depending on the exact setting in which the questions is asked)

Assessing agreement

In clinical studies, the most common situation in which we want to assess agreement are as follows:

1. **INTER**-rater reliability: A few raters measure the same characteristic (on a group of patients) with each rater measuring all patients. In this case we want to see how much the different raters agree in terms of their measurements.
2. **INTRA**-rater reliability: One rater successively measures a characteristic while the subjects' characteristic stay constant (*aka* **repeatability** or **reproducibility**)

Aside:

We should note that **rater** or **method** can refer to a person (doctor, nurse or patient), a machine or any instrument that produces a *score*.

Assessing agreement

The are two main features we need to consider in any of the above situations:

1. **Bias: The degree to which two methods/raters systematically disagree (i.e. consistently over- or under-estimation**
2. **Random variation: How much 'random' noise there is**

In terms of reliability, we will consider methods to assess continuous, nominal and ordinal measures.

Where to from here?

1. Measures of agreement: Continuous measures
   - Continuous measures: Bland-Altman plots
   - Continuous measures: Intra-class correlation
2. Measures of agreement: Categorical measures
   - Nominal outcomes
   - Ordinal measures

Caveat:

In this lecture, I will mainly focus on the two-rater problem, but I will comment on methods to gauge 'm-rater agreement' at the end of the session.

Methods for agreement of continuous measures

First we will consider measures of reliability on continuous measures. We will focus on two different approaches:

1. A graphical method: **Bland-Altman plots**; and
2. A 'statistical' measure: The **Intra-class correlation coefficient - (ICC)**

There is a lot of controversy about which of these methods represents the better approach. However, both methods have their advantages and limitations.

ICC or Bland-Altman plots

Personally, if the agreement is good, I will only use the ICC. If agreement is not so good, I will use **both** ICC and the Bland-Altman plot (more later)

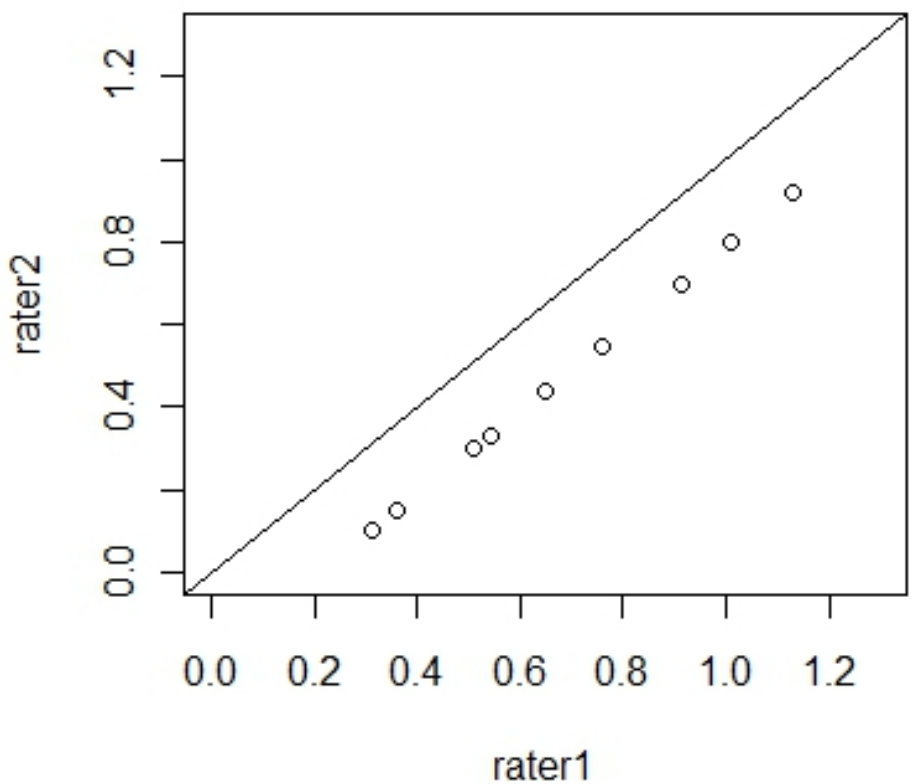
Pearson's correlation coefficient: A commonly used but poor measure of agreement

- In the past many studies have used Pearson correlation coefficient, r, to assess the agreement between two raters
- The problem with this approach is two raters can be highly correlated, but have low agreement
- This happens when one rater **consistently** over- or underestimates compared to the other rater. That is, **bias**

Problems with Pearson's correlation coefficient

Measures from two raters with a perfect correlation (r=1)

We can see that despite **perfect correlation**, *Rater 1 consistently rates lower than Rater 2. i.e.* **Bias**



Bland-Altman plots: Two raters only

Very simple idea:

1. Calculate the difference between each rater's measure for each subject
2. Calculate the average scores (of two raters) for each subject
3. Plot differences(1) on y-axis against average score(2) on x-axis
4. Include line that represent the **overall average difference**
5. Plot some **limits of agreement (LOA)** around average difference

Interpretation:

- Values falling within **LOA** show 'agreement'
- Values falling outside **LOA** demonstrate lack of agreement
- A **overall average difference** substantially different from zero indicate **bias**
- Wide **LOA** suggest high **noise** (random variation)

Bland-Altman plots: Calculating the limits of agreement (LOA)

The limits of agreement (LOA) are very similar to a 95% confidence interval, except the **Standard deviation of the differences** rather than the standard error is used:

$$\bar{\delta} \pm 1.96 S_{\delta}$$

where:
$\bar{\delta}$ is the average difference between the raters' scores
S_{δ} is the standard deviation of the difference between the raters' scores

LOA: Interpretation

The limits of agreement simply represent where we would expect 95% of differences between raters to fall (assuming a normal distribution).

Example or Bland-Altman plot: Heart rate

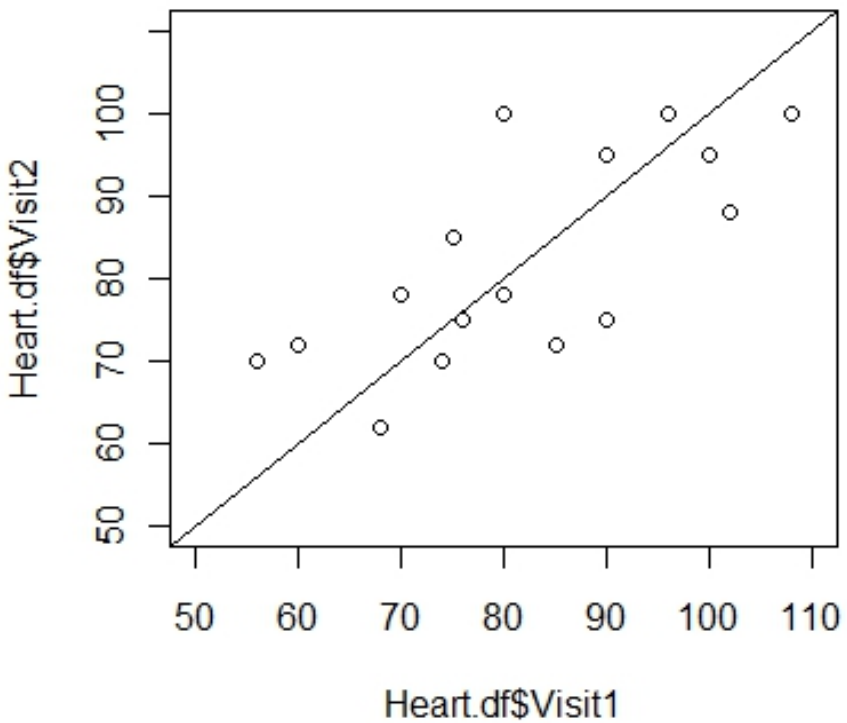
Here we will consider the measurement of heart rates of 16 patients taken by the same clinician, on two successive visits. Since we have a single clinician taking measurements on two occasions this is a **reproducibility** (or repeatability/intra-rater agreement) study:

Patient ID	Visit 1	Visit 2	**Patient ID**	Visit 1	Visit 2
1	90	75	9	85	72
2	100	95	10	108	100
3	80	72	11	75	85
4	56	70	12	74	70
5	76	75	13	70	78
6	80	100	14	80	78
7	90	95	15	68	62
8	96	100	16	102	68

We assume here that there hasn't been any disease progression (or other relevant patient changes) between visits

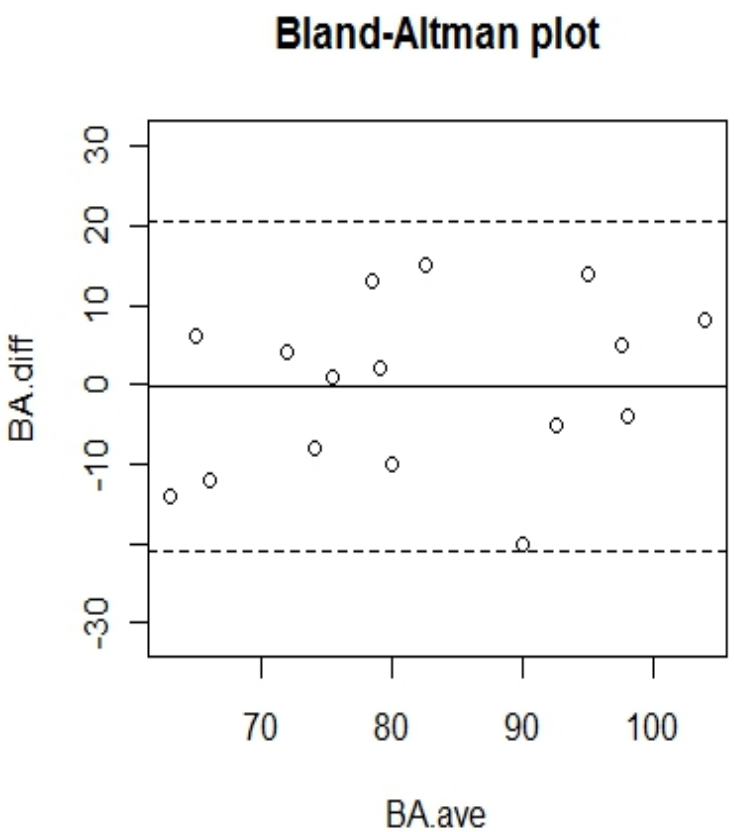
A preliminary perusal

Scatter plot of rater measures for two different visits of 16 patients



The straight diagonal line represents perfect agreement

The Bland-Altman plot



Interpretation:

- Average difference close to zero suggests low bias (solid middle line)
- Level of agreement independent of level
- All values fall within limits of agreement (suggest low variation/noise)
- No outliers (to investigate)

Conclusion: This scale demonstrated a reasonable level of reproducibility for this rater

Pros and Cons of the Bland-Altman approach

BA-plots have some major advantages, but this method has also come under a lot of criticism.

- Strengths:
  - BA-plot easy and intuitive to understand
  - Able to peruse all observations (not just get a measure of the 'average level of agreement')
  - Can visualize both **Bias** and **Noise** from the plot
  - Can also consider the instruments reliability over the entire spectrum (investigate **Spectrum-bias** and **Spectrum-noise**)
- Limitations:
  - As a graph, we don't have any *magic cutoff boundary* as to what constitutes agreement and disagreement
    - i.e. Subjective: What one researcher might think of as agreement, might represent disagreement for another
  - Does not provide a single 'quotable' statistic (measure)

Intra-Class Correlation (ICC)

An alternative way of assessing agreement is the Intra-class correlation coefficient (ICC). This approach is:

- A statistic representing the 'average' level of agreement
- Flexible: allows for greater than 2 raters and other reliability study designs
- However, this flexibility comes with a price: we have to think about and choose the right model
- ICC is not as intuitive as Bland-Altman plots, but if we understand what the ICC represents, it makes life easier
- There are a number of different models for ICC. For example:
  1. Simple m raters (random) problem
  2. m raters (random) by p methods (random)
  3. m raters (random) by p methods (fixed)
- We will only cover the first (and simplest) case here. Those interested in the more advanced models are directed to *Armitage and Berry*(1994) and *Chinn*(1990)

The Intra-Class Coefficient (ICC)

- ICCs are essentially a statistic representing the proportion of variation of an observation due to subject-to-subject variability in error free scores
- For the above reason, ICCs can be calculated using various Analysis of Variance (ANOVA) models.
- We will consider the first case from the previous slide (m raters) which can be calculated using a *Components of Variance* model; A one way ANOVA model with a **Random** effect (Unlike **fixed effect** models the raters are considered a random sample from the *population of* raters)

Aside:

ICCs are also used in a totally different area of clinical epidemiology: To represent (or measure) the level of within cluster association in multi-centre studies.

Analysis of components model for m-rater ICC

n subjects are rated by m raters (i.e. $n \times m$ observations). The Analysis of components (aka One-way random effects model) is given by:

$$Y_{ij} = \mu + s_i + \epsilon_{ij}$$

where $i = 1, 2, \cdots, n$ (number of patients); $j = 1, 2, \cdots, m$ (number of raters), μ is an unknown constant (mean of all observations), $s_i \sim N(0, \sigma_s^2)$ (random variation due to the subject) and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$

The components of variance ANOVA table:

Source of variation	df	MS	E(MS)
Between subjects	$n-1$	M_s	$\sigma_\epsilon^2 + m\sigma_s^2$
Residual	$n(m-1)$	M_r	σ_ϵ^2
Total	$nm-1$		

Analysis of components model for m-rater ICC

The quantity we are trying the estimate:

$$\rho_{ICC} = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_\epsilon^2}$$

Using the quantities given in the ANOVA table:

$$ICC = \hat{\rho}_{ICC} = \frac{M_s - M_r}{M_s + (m-1)M_r}$$

For two raters (i.e. $m = 2$)

$$ICC = \hat{\rho}_{ICC} = \frac{M_s - M_r}{M_s + M_r}$$

R code for ICC

Good news is that R makes ICC calculation simple:

```
library(psychometric) #Contains ICC function
#Have to stack data in long format
#Combine values from visit1 and visit2 into a single column
heart.rate<-c(Heart.df$Visit1, Heart.df$Visit3)
#Generate a patient ID variable
patient <- rep(c(1:16), times=2)
#Combine into a data frame
hold.df<-data.frame(heart.rate, patient )
#Caluclate one way random effects ICC
ICC1.CI(dv=heart.rate, iv=patient, data=hold.df, level = 0.95)
```

R Output

This results in:

```
heart.rate, patient
      L95       ICC1      U95
1 0.2507445 0.6462212 0.8591235
```

Our reproducability study

With an ICC of 0.65 we would conclude that the ICC (and therefore reproducibility) is *moderate*. Generally:

- $ICC < 0.4 \Rightarrow$ weak
- $0.4 \leq ICC < 0.7 \Rightarrow$ moderate
- $ICC \geq 0.7 \Rightarrow$ strong

The 95 % confidence intervals $[0.25, 0.85]$ give us an idea of the precision of our estimate. Narrower confidence intervals give a higher degree of certainty.

However, to test the hypothesis $H_0 : \rho_{ICC} = 0$ is not really meaningful. Although a 95%CI containing zero would suggest (very) poor reliability.

Bland-Altman or ICC???

- As I mentioned previously, there is a lot of contention about whether the Bland-Altman or ICC approach is best
- My advice is to use both, only presenting the full Bland-Altman plot if it adds something to the story (i.e. to demonstrate the nature of bias or noise)

In terms of writing up the results I would use something like:

> *...Both the Bland-Altman plot and Intraclass correlation coefficient(ICC) were used to evaluate the reproducibility of the XXX measure on the patients. The ICC showed moderate agreement between the measurements (ICC=0.65, 95%CI: 0.25, 0.86). The Bland-Altman plot supported this by showing a low degree of bias (average difference=-0.312) and no values falling outside the 95% lower and upper limits of agreement (-21.05, 20.43)...*

Assessing agreement for categorical measures

- Now let's consider how to assess the reliability of categorical measures
- We will consider two cases
  - Nominal measures (where there is **no** basis for ordering responses)
  - Ordinal measures (where there **is** a basis for ordering responses)
- Fortunately there is a standard approach for doing this: Cohens' Kappa.
- There are two variants of Cohens' Kappa. One to deal with he **nominal** case (**Cohens' Kappa**) and one for the **ordinal** case (**Cohens' Weighted Kappa**)

However, before we use these **appropriate** methods for assessing agreement for categorical measures, let's consider a naive (and inappropriate) approach for this purpose: the χ^2 test of independence

χ^2 test of independence based on cross-tabulation

Consider two situations (Table 1 and Table 2) where two raters are asked to diagnose a disease:

Table 1:

		Rater B	
		+	-
Rater A	+	75	25
	-	25	75

Table 2:

		Rater B	
		+	-
Rater A	+	50	50
	-	50	50

- **Table 1** Here we see reasonable agreement. It might seem logical to assess agreement using a standard χ^2 **test of Independence**. If we (inadvisedly) used the χ^2 test, we might conclude strong agreement($\chi^2 = 48.02$, $df = 1$, $p < 0.0001$)
- Problem: *Isn't it possible that a large proportion of the classifications could have been correct just by chance??*

If the two raters '**flip of the coin**' to classify people (see **Table 2**) we would expect both raters to agree about 50% of the time **just by chance**

Agreement by chance

This problem is exacerbated when we have diseases of very low prevalence. Consider a disease that has a prevalence of 0.01:

Table 3:

		Rater B	
		+	-
Rater A	+	0	1
	-	1	98

In this case, both raters failed to agree on any patients with the disease, but they still managed to agree on 98% of cases.

We need a method that can measure true agreement by accounting for **agreement just by chance**

Cohens' Kappa: Binary case

Consider a general table (similar to those above) representing the agreements and disagreements between raters on a two-point (Binary) scale:

		Rater B		
		+	-	
Rater A	+	**a**	b	$a+b$
	-	c	**d**	$c+d$
		$a+c$	$b+d$	n

One way of representing agreement would be to calculate the quantity:

$$I_o = \frac{a+d}{n}$$

Observed agreement, I_o, would be 0.75, 0.5 and 0.98 in Tables 1, 2 and 3 respectively. But as we have seen already, we need to account for the agreement we would see by chance

Cohens' Kappa

Cohens' kappa allows us to calculate the agreement we would expect just by chance. As before **observed** agreement is given by:

$$I_o = \frac{a+d}{n}$$

and **chance** agreement can be calculated:

$$I_e = \frac{E(a)+E(d)}{n}$$

where $E(a) = \frac{(a+c)(a+b)}{n}$ and $E(d) = \frac{(c+d)(b+d)}{n}$ are the **expected frequencies** (the same as in a χ^2 test of independence: $\frac{row_{total} \times col_{total}}{overall_{total}}$)

Note:

Only expected frequencies of **agreement cells** are used to calculate I_e

Cohens' Kappa

Now Cohens' Kappa represents **the difference between the observed and expected frequencies as a fraction of the maximal difference**. This maximal difference also accounts for the agreement just by chance.

$$\kappa = \frac{\mathbf{I_o} - \mathbf{I_e}}{\mathbf{1} - \mathbf{I_e}}$$

Example: *tardive dyskinesia*

Two raters are asked to administer a new test to diagnose *tardive dyskinesia*, with the following results:

		Rater B		
		+	-	
Rater A	+	**123**	10	133
	-	6	**29**	35
		129	39	168

Now:

$$I_o = \frac{123 + 29}{168} = \frac{152}{168} = 0.905$$

$$I_e = \frac{\frac{(129)(133)}{168} + \frac{(39)(35)}{168}}{168} = 0.656$$

$$\kappa = \frac{0.905 - 0.656}{1 - 0.656} = 0.72$$

How high should Cohens' κ be?

A value of 0.72 suggests a reasonably high degree of reliability. We would conclude that the raters generally agree.

So what represents a high 'enough' level of Cohens' κ?

Fleiss(1999) suggests the following guidelines:

- $\kappa \leq 0.4 \Rightarrow$ *Poor*
- $0.4 < \kappa \leq 0.75 \Rightarrow$ *Fair to Good*
- $\kappa > 0.75 \Rightarrow$ *Excellent*

Incidentally, these are similar to the cut-offs for the ICC.

Cohens' Kappa for nominal outcomes

It is quite simple to extend the binary form of Cohens' Kappa to the nominal (>2 class)problem.

		Rater B		
		absent	typeR	typeQ
Rater A	absent	**a**	*b*	*c*
	typeR	*d*	**e**	*f*
	typeQ	*h*	*i*	**j**

Gives: $I_o = \frac{a+e+j}{n}$ and $I_e = \frac{E(a)+E(e)+E(j)}{n}$

and as before: $\kappa = \frac{I_o - I_e}{1 - I_e}$

Note:

Again note that only the frequencies of the **agreement cells** are used in both I_o and I_e

Ordinal outcomes and Cohens' weighted Kappa

- Often we are presented with the case where our measurement scale is ordinal:
  - *Mild, Moderate, Severe*
  - *Absent, Benign, Suspect, Cancer*
- Cohens' Kappa can be extended to account for this ordering
- **Basic idea: Increase the amount of penalty with higher levels of disagreement**
  - For example, we might penalize disagreements in diagnoses two categories apart (e.g. Mild vs Severe) twice as highly as those adjacent (Mild vs Moderate and Moderate vs Severe) which we might term **partial agreement**
- This idea is implemented in the method: **Cohens' Weighted Kappa**

Calculation of Cohens' Weighted Kappa

- I wont go into too much details about how the statistic is calculated (those interested are directed to Armitage and Berry, 1994)
- The main innovation in this method is that both the observed and expected agreements are calculated using weights that reflect the **level** of agreement (e.g. full agreement, partial agreement and total disagreement)
- Higher levels of disagreement are penalized higher (weighed lower) than lower levels of disagreement (partial agreement)
- Now the question is how much to penalize for different levels of (dis)agreement
- A number of ways to do this. A common approach used is the **equally spaced** penalty (see below)

Agreement, partial agreement and disagreement

I will illustrate how the Weighted κ statistic works using a number of examples. Consider a 2 rater by 3 point ordinal scale of disease severity:

		Rater B		
		mild	moderate	severe
Rater A	mild	a	b	c
	moderate	d	e	f
	severe	h	i	j

The 'standard' Cohens' Kappa

We might decide to use a 'standard' κ statistic (and assume all disagreement is equally bad). That is:

Weight table: unweighted κ

		Rater B		
		mild	moderate	severe
Rater A	mild	1	0	0
	moderate	0	1	0
	severe	0	0	1

That is, in this '**weight table**' (**the unweighted case**) there is either 'full agreement' (with weight of 1), or 'disagreement' (weight of 0).

Equally-spaced Cohens' weighted kappa

Or we might use an 'equally spaced' approach where closer disagreements (partial agreement) is weighted more highly (penalized less) than full disagreements:

Weight table: equally spaced κ for 3 x 3 problem

		Rater B		
		mild	moderate	severe
Rater A	mild	1	0.5	0
	moderate	0.5	1	0.5
	severe	0	0.5	1

Aside:

For the k point ordinal scales, weights are: $w_i = 1 - \frac{i}{k-1}$

- Three-point scale: 0, 0.5, 1
- Four-point scale: 0, 0.33, 0.66, 1
- Five-point scale: 0, 0.25, 0.5, 0.75, 1
- etc

Asymmetric weight tables

- We may also want to penalize more highly for particular types of disagreements
- For example, consider a situation where patients classified as moderate or severe are triaged to further examination, whereas those classified as mild aren't
- In this case we would want to **highly** penalize a disagreement involving a mild classification:

Weight table: Example of asymmetric weighting

		Rater B		
		mild	moderate	severe
Rater A	mild	1	**0.3**	0
	moderate	**0.3**	1	**0.8**
	severe	0	**0.8**	1

Here we are penalizing highly (weighting low) disagreements on the lower end of the spectrum.

Example: Patient vs Nurse rating of cholesterol levels

40 patients self-rated themselves as having **low** (chol < 3.8), **high** (chol $\epsilon(3.8, 42]$ or **very high** (chol ≥ 4.2) cholesterol levels. Nurses were then asked to classify these patients cholesterol levels (using the same method) with the following results:

Patient vs nurse cholesterol levels

		Patients		
		low	high	very high
Nurses	low	17	0	0
	high	4	6	1
	very high	1	7	4

Eyeball: Summarize what you think is happening here
Nurses are overestimating (and\or patients are underestimating) cholesterol levels.

Now let's use R to calculate:

1. The standard (unweighted) κ
2. The equally-spaced Weighted κ
3. An example of a asymmetric weighted κ

I will use the *epicalc* library developed by Prof Virasakdi Chongsuvivatwong at PSU

Unweighted (nominal) case

```
#Assume data set with ratings read in
library(epicalc)

#Cross tabulate patient and nurse ratings
my.tab<-table(Chol.df$nurse.rate,
+ Chol.df$patient.rate, dnn=c("Nurses","Patients"))

#Run unweighted Kappa
kap(my.tab)
```

Results (Unweighted Kappa):

```
Patients
Nurses  0  1  2
0 17  0  0
1  4  6  1
2  1  7  4

Observed agreement = 67.5 %
Expected agreement = 36.06 %
Kappa = 0.492
Standard error = 0.109 , Z = 4.523 , P value = < 0.001
```

- We would say the agreement was on the lower end of moderate
- Note that here all types of disagreement are just as bad
- The hypothesis test ($H_0 : \kappa = 0$) is rather meaningless
- 95% CI is $\kappa \pm 1.96 S_\kappa = 0.492 \pm 1.96(0.109) = [0.278, 0.706]$

Equally-spaced Kappa (R code)

```
kap(my.tab, wttable = "w")
#wttable="w" means  equally-spaced
```

Results:

```
Patients
Nurses  0  1  2
0 17  0  0
1  4  6  1
2  1  7  4
Observed agreement = 82.5 %
Expected agreement = 57.12 %
Kappa = 0.592
Standard error = 0.117 , Z = 5.05 , P value = < 0.001
```

- By accounting for ordinality, we can see an improvement ($\kappa = 0.592$), which comfortably falls in the 'good range'
- In this case disagreements in close proximity are penalized less than those further apart

Asymmetric weighted Kappa:

```
my.wts <-as.table(rbind(c(1,0.8,0), c(0.3,1,0.8), c(0, 0.3, 1)))
my.wts
A   B   C
A 1.0 0.3 0.0
B 0.3 1.0 0.8
C 0.0 0.8 1.0
kap(my.tab, wttable = my.wts)
Patients
Nurses  0  1  2
0 17  0  0
1  4  6  1
2  1  7  4
Observed agreement = 77.75 %
Expected agreement = 57.32 %
Kappa = 0.479
Standard error = 0.115 , Z = 4.178 , P value = < 0.001
```

- Disagreements involving mild levels are penalized more highly
- Not as successful, but maybe a more useful measure of agreement

A last word on Cohens' κ

- The choice of if/how we should weight our disagreement is one that should be on clinical grounds and/or design considerations
- We should **NEVER** just choose the result that sells our idea better (gives the highest κ): **UNETHICAL**
- The κ-statistic is a widely used and standard measure of gauging agreement between raters using categorical instruments
- κ is quite robust, but it is susceptible to the very low (very high) prevalence problem
- Some modifications have been proposed to the Kappa the circumvent this problem (e.g. see Byrt *et al.*, 1993)
- Versions of the Kappa statistics for the *m*-rater problem have also been proposed (also available in R)

References and resources

Armitage, P. and Berry, G. (1994) *Statistical Methods in Medical Research (3ed)*

Byrt, T., Bishop, J., and Carlin, J. (1993) Bias, Prevalence, and Kappa. *Journal of Clinical Epidemiology* **46(5)**, 423-429

Chinn, S.(1990) The assessment of methods of measurement. *Statistics in Medicine* **9**, 351-362

Fleiss, J. L. (1999) *The Design and Analysis of Clinical Experiments*, John Wiley & Sons, Inc., Hoboken, NJ, USA.

THANK-YOU

Questions?????