Postgraduate course
in

Evaluation and comparison of method of measurements

Day 3 (part 2)

Kappa (κ) and design considerations

Niels Trolle Andersen

Dept. of Biostatistics, Aarhus University

Introduction

Until now: continuous data

What about categorical data?

As in other situations (ex regression analysis) it is much more complicated:

- **the analysis**
- **the interpretations**
- **the requirement to sample size etc.**

Today: A little bit about kappa (κ)
(the presentation maybe biased…)

and at the end

design considerations (in general)

Kappa statistics

		observer 1		
		ill	**healthy**	total
observer 2	**ill**	**22**	**4**	26
	healthy	**8**	**45**	53
	total	30	49	79

How well do the observers agree?

The observers agree on 67 out of 79 i.e.

$P_{obs} = (22+45)/79 = 0.85 = 85\%$

The chance of ´random' agreement

$P_{chance} = (30*26+49*53)/(79*79) = 0.54 = 54\%$

Can we describe the agreement in just one number?

Kappa statistics General setup

		observer 1		
		ill	**healthy**	total
observer 2	**ill**	**a**	**b**	a+b
	healthy	**c**	**d**	c+d
	total	a+c	b+d	n

Kappa statistics

Observed agreement

$$p_{obs} = (a+d)/n$$

The chance of 'random' agreement (if for example they looked at different things):

$$p_{chance} = ((a+c)*(a+b) + (d+c)*(d+b)) / n^2$$

Kappa (κ) is the proportion of additional agreement:

$$\kappa = (p_{obs} - p_{chance})/(1 - p_{chance})$$

An easy formula for the se of κ (you can find 'better' formulas):

$$se(\kappa) = \sqrt{\frac{p_{obs}(1-p_{obs})}{n(1-p_{chance})^2}}$$

The observers agree on 67 out of 79 i.e.

$$P_{obs} = (22+45)/79 = 0.85 = 85\%$$

The chance of ´random' agreement

$$P_{chance} = (30*26+49*53)/(79*79) = 0.54 = 54\%$$

Kappa (κ) is the proportion of additional agreement:

$$\kappa = (P_{obs} - P_{chance})/(1 - P_{chance})$$
$$= (0.85–0.54)/(1–0.54)$$
$$= 0.67$$
$$se(\kappa) = 0.088$$

95%CI for κ (approximately): (0.50, 0.84)

Stata:

```
. kap obs1  obs2
                 Expected
Agreement   Agreement     Kappa    Std. Err.          Z       Prob>Z
-----------------------------------------------------------------
  84.81%      54.11%     0.6690     0.1118        5.98      0.0000

. kapci obs1  obs2
                                          N=79
------------------------------------------------
 Kappa (95% CI) = 0.669 (0.498 - 0.840)     (A)
------------------------------------------------
 A = analytical


. kapci obs1  obs2,estim( bc ) reps(20000)
This may take quite a long time. Please wait ...
                                B=20000 N=79
------------------------------------------------
 Kappa (95% CI) = 0.669 (0.486 - 0.831)     (BC)
------------------------------------------------
 BC = bias corrected
```

kapci isn't a ´default' comand in Stata (but can be downloaded)

		observer 4 ill	observer 4 healthy	total
observer 3	ill	22	0	22
	healthy	12	45	57
	total	32	45	79

Here we have κ=0.68; almost the same as before: Do we have the same agreement as before?

Do we have a systematic difference between the observers?

We have a statistically significant difference between the two observers with respect to the portions of persons judges 'ill'.
(McNemar test, day 4 basic course)

Stata:..

```
. ci obs3,bin

                                                         -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
        obs3 |         79     .278481    .0504322        .183455    .3907351
. ci obs4,bin

                                                         -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
        obs4 |         79    .4303797    .0557064       .3194235    .5467142

. mcc obs3 obs4
(table)
McNemar's chi2(1) =     12.00    Prob > chi2 = 0.0005
Exact McNemar significance probability       = 0.0005
Proportion with factor
        Cases          .278481
        Controls      .4303797     [95% Conf. Interval]
                     ---------     --------------------
        difference -.1518987     -.2437041  -.0600933
        ratio         .6470588      .504813   .8293866
        rel. diff. -.2666667     -.4364744   -.096859
        odds ratio           0            0   .3598938   (exact)
```


Stata:.

```
kap obs3 obs4
             Expected
Agreement   Agreement     Kappa   Std. Err.         Z      Prob>Z
-----------------------------------------------------------------
  84.81%      53.08%     0.6762     0.1064       6.35      0.0000

. kapci obs3 obs4
                                          N=79
------------------------------------------------
 Kappa (95% CI) = 0.676 (0.517 - 0.836)    (A)
------------------------------------------------
 A = analytical

. kapci obs3 obs4,estim( bc ) reps(20000)
This may take quite a long time. Please wait ...
                                  B=20000 N=79
------------------------------------------------
 Kappa (95% CI) = 0.676 (0.516 - 0.832)    (BC)
------------------------------------------------
 BC = bias corrected
.
```


		observer 1 ill	observer 1 healthy	total
observer 2	ill	7	4	11
	healthy	8	60	68
	total	15	64	79

In this example κ=0.45; less than as before but the same observed agreement.

How much better is 0.68 compared to 0.45????

Stata:.

```
. kap obs5 obs6
             Expected
Agreement   Agreement     Kappa   Std. Err.         Z      Prob>Z
-----------------------------------------------------------------
  84.81%      72.38%     0.4501     0.1106       4.07      0.0000

. kapci obs5 obs6
                                          N=79
------------------------------------------------
 Kappa (95% CI) = 0.450 (0.190 - 0.710)    (A)
------------------------------------------------
 A = analytical

. kapci obs5 obs6,estim( bc ) reps(20000)
This may take quite a long time. Please wait ...
                                  B=20000 N=79
------------------------------------------------
 Kappa (95% CI) = 0.450 (0.165 - 0.704)    (BC)
------------------------------------------------
 BC = bias corrected
```

An example with 4 categories:

```
                    pat2
      pat1 |         1           2           3           4 |     Total
-----------+-----------------------------------------------+----------
         1 |        22           2           2           0 |        26
         2 |         5           7          14           0 |        26
         3 |         0           2          36           0 |        38
         4 |         0           1          17          10 |        28
-----------+-----------------------------------------------+----------
     Total |        27          12          69          10 |       118
```

If we use the same definition as above we get

```
. kap pat1 pat2
             Expected
Agreement   Agreement     Kappa   Std. Err.         Z      Prob>Z
-----------------------------------------------------------------
  63.56%      28.12%     0.4930     0.0501       9.83      0.0000

. kapci  pat1 pat2,estim( bc ) reps(20000)
Kappa (95% CI) = 0.493 (0.385 - 0.606)    (BC)

It doesn't take into account the 'degree' of agreement
```



```
                    pat2
      pat1 |         1           2           3           4 |     Total
-----------+-----------------------------------------------+----------
         1 |        22           2           2           0 |        26
         2 |         5           7          14           0 |        26
         3 |         0           2          36           0 |        38
         4 |         0           1          17          10 |        28
-----------+-----------------------------------------------+----------
     Total |        27          12          69          10 |       118
```

We can 'weight' the agreement (see help or stata manual for details)

```
. kap pat1 pat2,wgt(w)

Ratings weighted by:
   1.0000   0.6667   0.3333   0.0000
   0.6667   1.0000   0.6667   0.3333
   0.3333   0.6667   1.0000   0.6667
   0.0000   0.3333   0.6667   1.0000

             Expected
Agreement   Agreement     Kappa   Std. Err.         Z      Prob>Z
-----------------------------------------------------------------
  87.01%      63.00%     0.6488     0.0631      10.29      0.0000

Compared to (from the previous slide)
  63.56%      28.12%     0.4930     0.0501       9.83      0.0000
```


Kappa statistics

```
. kap pat1 pat2,wgt(w2)

Ratings weighted by:
   1.0000   0.8889   0.5556   0.0000
   0.8889   1.0000   0.8889   0.5556
   0.5556   0.8889   1.0000   0.8889
   0.0000   0.5556   0.8889   1.0000
             Expected
Agreement   Agreement     Kappa   Std. Err.         Z      Prob>Z
-----------------------------------------------------------------
  95.10%      77.35%     0.7838     0.0910       8.61      0.0000

Compared to (from the previous slides)
 87.01%      63.00%     0.6488     0.0631      10.29      0.0000
 63.56%      28.12%     0.4930     0.0501       9.83      0.0000
```

You can also define your own weights..


```
                    pat2
      pat1 |         1           2           3           4 |     Total
-----------+-----------------------------------------------+----------
         1 |        22           2           2           0 |        26
         2 |         5           7          14           0 |        26
         3 |         0           2          36           0 |        38
         4 |         0           1          17          10 |        28
-----------+-----------------------------------------------+----------
     Total |        27          12          69          10 |       118
```

Do the observes have the same distribution:

```
 signrank pat1=pat2
Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |       25        2409      2085.5
    negative |       18        1762      2085.5
        zero |       75        2850        2850
-------------+---------------------------------
         all |      118        7021        7021

unadjusted variance   138664.75
adjustment for ties    -1333.00
adjustment for zeros  -35862.50
                     ----------
adjusted variance     101469.25

Ho: pat1 = pat2
             z =   1.016
    Prob > |z| =   0.3098
```

pat1	pat2 1	2	3	4	Total
1	22	2	2	0	26
2	5	7	14	0	26
3	0	2	36	0	38
4	0	1	17	10	28
Total	27	12	69	10	118

Alternative: A kappa-value for each category

Category 1:

	0	1	Total
0	87	5	92
1	4	22	26
Total	91	27	118

Agreement	Expected Agreement	Kappa	Std. Err.	Z	Prob>Z
92.37%	65.17%	0.7810	0.0920	8.49	0.0000

No systematic difference between the observers

Category 2	0	2	Total
0	87	5	92
2	19	7	26
Total	106	12	118

Agreement 80%
Expected agreement 72%
Kappa 27%

Category 3	0	3	Total
0	47	33	80
3	2	36	38
Total	49	69	118

Agreement 70%
Expected agreement 47%
Kappa 44%

Category 4	0	4	Total
0	90	0	90
4	18	10	28
Total	108	10	118

Agreement 85%
Expected agreement 72%
Kappa 46%

Systematic difference?

Kappa statistics

Other extensions:

Repetitions within an observer.

More observes.

Different observers. (look in the stata manual)

Other models

Describing the ‘Probability of agreement/disagreement

Models like the models we used analyzing continuous data

Kappa statistics

Remarks:

The κ doesn’t separate systematic and random variation

When does the observers have the same distribution of the answers?

The κ is related to correlations i.e. that it depends on the ’variation’ in the sample.

The sample used for estimating κ should be a random sample from the population (latent variable?)

When is κ large/good?

Knowing the truth (diagnostic test?)

Design considerations

When comparing/evaluating methods of measuring it is important:

- – to realize how the method is going to be used
- – to identify the main contributions to the variation in the data
- –to define what is acceptable/unacceptable (in advance) and how check it

Design considerations

Contribution to variation in data

or

Sources of variation

or

Variance components

Biological variation (systematic and or random):

- inter-subject variation
- intra subject variation:
  - day to day variation
  - intra day variation
  - other with-in subject variation

Design considerations

Contribution to variation in data

Technical variation (systematic and or random):

- inter-method variation
- inter-device variation
- intra method/device variation:
  - day to day variation
  - intra day variation
  - other with-in method variation

Design considerations

Which of the different variance-components do we want to estimate (may be combinations) depends on how the method is going to be used:

- On individuals or groups?
- Direct measurements or changes?
- If changes how? (directly or as a difference)
- How many repetitions (and how)?

What is acceptable/unacceptable?

The size of some or combinations of sd?

The precision of an stimated standard deviation - the 95% CI for σ

$$\hat{\sigma}\cdot\sqrt{\frac{df}{\chi^2_{df}(0.975)}}\le\sigma\le\hat{\sigma}\cdot\sqrt{\frac{df}{\chi^2_{df}(0.025)}}$$

$$\hat{\sigma}\cdot l(df)\le\sigma\le\hat{\sigma}\cdot u(df)$$

df	*l(df)*	*u(df)*
5	0.624	2.453
10	0.699	1.755
15	0.739	1.548
20	0.765	1.444
25	0.784	1.380
50	0.837	1.243
150	0.899	1.128
200	0.911	1.109

Design considerations

Showing superiority of one method compared to another method:

- **Smaller measurement error**
  - sample size calculation (α=0.05, power 0.8)
  - Ratio between sd's: = 2 df=18 in each group
  - = 1.5 df=49 in each group
  - = 1.25 df=192 in each group
  - (df=(no of measurement – number of subject) and at least 2 measurement on each subject)
- **????? (can a method with a larger measurement error be superior?)**

Design considerations

Comparing/evaluating methods of measuring it a never ending process and consist of contributions from different studies.

It is not possible to do
'the ultimative comparison/evaluation'

Where to start (or stop)?

Evaluation???