Two aspects: • Reliability • Validity

Reliability (Reproducibility, Precision)

Extent to which obtain the same answer/value/score if object/subject is measured repeatedly under similar situations...

Some ways to quantify Reliability:

- For one subject: average variation of individual measurements around their mean... either the square root of the average of squared deviations) i.e. standard deviation (SD); or the average absolute deviation, which will usually be quite close to the SD. Could also use range or other measures such as Inter Quartile Range)

- For one subject: average variation or SD as % of the mean of the measurements for that subject...called the {within-subject} Coefficient of Variation (CV) if calculate it as [SD/Mean] × 100.

- For several subjects: : average the the CV's calculated for the different subjects; if CV's are highly variable, may want to give some sense of this using the range or other measure of spread of the CV's.

  Unfortunately, CV gives no sense of how well the measurements of different subjects (ss) segregate from each other

  How about

  $$\frac{\text{SD of within-ss measurements}}{\text{SD of between-ss meassurements}}$$?? see last item below*

- Correlation (Pearson or Spearman) if 2 assessments of each ss. ??

- Using correlation betwen scores on random halves of a test, can estimate how 'reproducible' the full test is (helpful if cannot repeat the test)

- If the measurement in question concerns a population (eg the percentage of smokers among Canadian adults) and if it is measured (estimated) using a statistic: e.g. the proportion in a random sample of 1000 adults, it is possible from statistical laws concerning averages to quantify the reliability of the statistic without having to actually perform repeated measurements (samples). For simple random sampling, the formula

  $$\text{SE[average]} = \frac{\text{SD[individuals]}}{\sqrt{\text{number of individuals measured}}}$$

  allows us to quantify the reliability indirectly. If we didn't know this formula, we could also arrive at an answer by various re-sampling methods applied to the individuals in the sample at hand -- again without resorting to oberving any additional individuals.

- * Some function of Variance of Within-ss measurements and Variance of Between-ss values? ? Estimate these COMPONENTS OF VARIANCE USING Analysis of Variance (ANOVA)

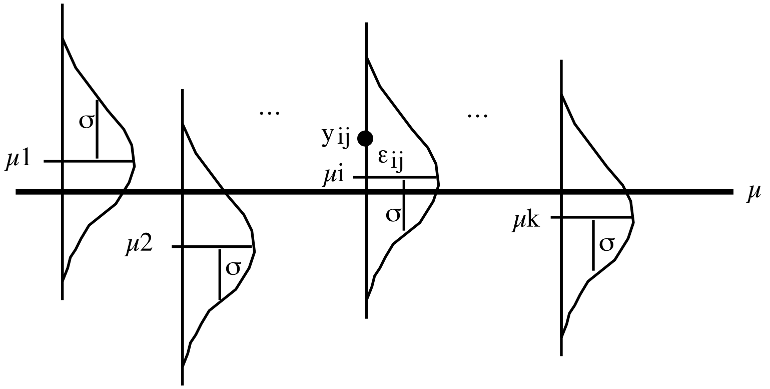
First, a General Orientation to ANOVA and its primary use, namely testing differences between *μ*'s of k (≥2) different groups.

E.g. 1-way ANOVA:

DATA:

	Group					
Subject	1	2	.	i	.	k
1	y_{11}	.	.	.	.	.
2						
.	.	.	.	.	.	.
j				y_{ij}		
.	.	.	.	.	.	.
n						y_{kn}
Mean	$\bar{y}_1$	$\bar{y}_2$		$\bar{y}_i$		$\bar{y}_k$
Variance	$s^2{}_1$	$s^2{}_2$				$s^2{}_k$

MODEL



σ refers to the variation (SD) of all possible individuals in a group;
It is an (unknowable) parameter; it can only be ESTIMATED.

Or, in symbols...

$$y_{ij} = \mu_i + e_{ij} = \mu + (\mu_i - \mu) + e_{ij}$$

DE-COMPOSITION OF OBSERVED (EMPIRICAL) VARIATION

$$\sum\sum(y_{ij} - \bar{y})^2 = \sum\sum(\bar{y}_i - \bar{y})^2 + \sum\sum(y_{ij} - \bar{y}_i)^2$$

TOTAL Sum of Squares = BETWEEN Groups Sum of Squares + WITHIN Group Sum of Squares

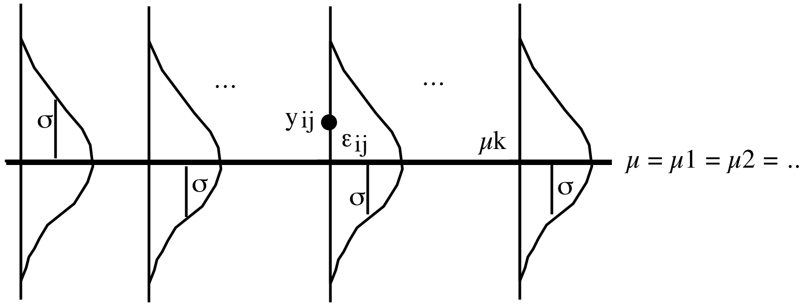
ANOVA TABLE

	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio	P-Value
SOURCE	SS	df	MS (= SS /df)	$\frac{MS_{BETWEEN}}{MS_{WITHIN}}$	Prob(>F)
BETWEEN	xx.x	k–1	xx.x	x.xx	0.xx
WITHIN	xx.x	k(n–1)	xx.x		

LOGIC FOR F-TEST (Ratio of variances) as a test of

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu_k$$

UNDER H0



Means, based on samples of n, should vary around *μ* with a variance of $\frac{\sigma^2}{n}$

Thus, if H_0 is true, and we calculate the empirical variance of the k different $\bar{y}_i$'s, it should give us an unbiased estimate of $\frac{\sigma^2}{n}$

i.e. $\frac{\sum[\bar{y}_i - \bar{y}]^2}{k-1}$ is an unbiased estimate of $\frac{\sigma^2}{n}$

i.e. $\frac{n \sum[\bar{y}_i - \bar{y}]^2}{k-1}$ is an unbiased estimate of σ^2

i.e. $\frac{\sum\sum[\bar{y}_i - \bar{y}]^2}{k-1} = MS_{BETWEEN}$ is an unbiased estimate of σ^2

Whether or not H_0 is true, the empirical variance of the n (within-group) values y_{i1} to y_{in} i.e. $\frac{\sum[\bar{y}_{ij} - \bar{y}_i]^2}{n-1}$ should give us an unbiased estimate of σ^2

i.e. $s^2{}_i = \frac{\sum[\bar{y}_{ij} - \bar{y}_i]^2}{n-1}$ is an unbiased estimate of σ^2

so the average of the k diferent estimates,

$$\frac{1}{k}\sum s^2{}_i = \frac{1}{k}\sum \frac{\sum[\bar{y}_{ij} - \bar{y}_i]^2}{n-1}$$

is also an unbiased estimate of σ^2

i.e. $\frac{\sum\sum[\bar{y}_{ij} - \bar{y}_i]^2}{k[n-1]} = MS_{WITHIN}$ is an unbiased estimate of σ^2

THUS, under H_0, both $MS_{BETWEEN}$ and MS_{WITHIN} are unbiased estimates of estimates of σ^2 and so their ratio should, apart from sampling variability, be 1. IF however, H_0 is not true, $MS_{BETWEEN}$ will tend to be larger than MS_{WITHIN}, since it contains an extra contribution that is proportional to how far the μ's are from each other.

In this "non-null" case, the $MS_{BETWEEN}$ is an unbiased estimate of

$$\sigma^2 + \frac{\sum n[\mu_i - \bar{\mu}]^2}{k-1}$$

and so we expect that, apart from sampling variability, the ratio $\frac{MS_{BETWEEN}}{MS_{WITHIN}}$ should be greater than 1. The tabulated values of the F distribution (tabulated under the assumption that the numerator and denominator of the ratio are both estimaes of the same quantity) can thus be used to assess how extreme the observed F ratio is and to assess the evidence against the H_0 that the μ's are equal.

How ANOVA can be used to estimate Components of Variance used in quantifying Reliability.

The basic ANOVA calculations are the same, but the MODEL underlying them is different. First, in the more common use of ANOVA just described, the groups can be though of as all the levels of the factor of interest. The number of levels is necessarily finite. The groups might be the two genders, all of the age groups, the 4 blood groups, etc. Moreover, when you publish the results, you explicitly identify the groups.

When we come to study subjects, and ask "How big is the intra-subject variation compared with the inter-subject varaition, we will for budget reasons only study a sample of all the possible subjects of interest. We can still number them 1 to k, and we can make n measurements on each subject, so the basic layout of the data doesn'y change. All we do is replace the word 'Group' by 'Subject' and speak of BETWEEN-**SUBJECT** and WITHIN-**SUBJECT** variation. So the data layout is...

DATA:

	Subject					
	1	2	.	i	.	k
Measurement						
1	y_{11}	.	.	.	.	.
2						
.	.	.	.	.	.	.
j				y_{ij}		
.	.	.	.	.	.	.
n						y_{kn}
Mean	$\bar{y}_1$	$\bar{y}_2$		$\bar{y}_i$		$\bar{y}_k$
Variance	$s^2{}_1$	$s^2{}_2$				$s^2{}_k$

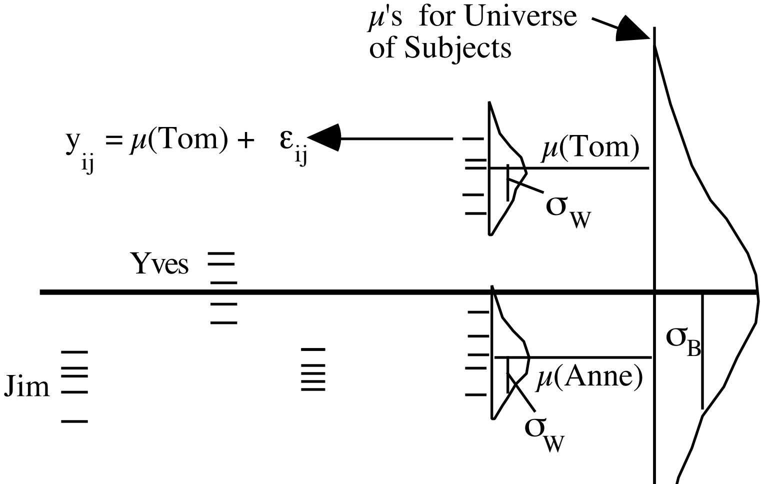
MODEL

The model is different. There is no interest in the specific subjects. Unlike the critical labels "male" anf "female", or "smokers", "nonsmokers" and "exsmokers" to identify groups of interest, we certainly are not going to identify subjects as Yves, Claire, Jean, Anne, Tom, Jim, and Harry in the publication, and nobody would be fussed if in the dataset we used arbitrary subject identifiers to keep track of which measurements were made on whom. we wouldn't even care if the research assistant lost the identities of the subjects -- as long as we know that the correct measurents go with the correct subject!

The "Random Effects" Model uses 2 stages:

(1) random sample of subjects, each with his/her own μ
(2) For each subject, series of random variations around his/her μ

Notice the diagram has considerable 'segregation' of the measurements on different individuals. There is no point in TESTING for (inter-subject) differences in the μ's. The task is rather to estimate the relative magnitudes of the two variance components σ^2_B and σ^2_W.


μ's for Universe of Subjects
$y_{ij} = \mu(\text{Tom}) + \varepsilon_{ij}$
μ(Tom)
σ_W
Yves
Jim
σ_B
μ(Anne)
σ_W

σ_B refers to the SD of the universe of μ's ; It is an unknowable parameter and can only be ESTIMATED

σ_W refers to the variation (SD) of all possible measurements on a subject It is an (unknowable) parameter; it can only be ESTIMATED.

Or, in symbols...

$$y_{ij} = \mu_i + e_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij}$$
$$= \mu + \alpha_i + \varepsilon_{ij}$$

$\alpha_i \sim N(0, \sigma^2_B)$

$\varepsilon_i \sim N(0, \sigma^2_W)$

DE-COMPOSITION OF OBSERVED (EMPIRICAL) VARIATION

$$\sum\sum(\bar{y}_{ij} - \bar{y})^2 = \sum\sum(\bar{y}_i - \bar{y})^2 + \sum\sum(\bar{y}_{ij} - \bar{y}_i)^2$$

TOTAL Sum of Squares = BETWEEN Subjects + Sum of Squares + WITHIN Subjects Sum of Squares

ANOVA TABLE *(Note absence of F and P-value Columns)*

	Sum of Squares	Degrees of Freedom	**Mean Square**	**What the Mean Square is an estimate of***
SOURCE	SS	df	MS (= SS /df)	
BETWEEN Subjects	xx.x	k–1	xx.x	$\sigma^2_W + n\,\sigma^2_B$
WITHIN Subjects	xx.x	k(n–1)	xx.x	σ^2_W

ACTUAL ESTIMATION OF 2 Variance Components

$MS_{BETWEEN}$ is an unbiased estimate of $\sigma^2_W + n\,\sigma^2_B$

MS_{WITHIN} is an unbiased estimate of σ^2_W

By subtraction...

$MS_{BETWEEN} - MS_{WITHIN}$ is an unbiased estimate of $n\,\sigma^2_B$

$\dfrac{MS_{BETWEEN} - MS_{WITHIN}}{n}$ is an unbiased estimate of σ^2_B

This is the **definitional** formula; the **computational** formula may be different.

----

* Pardon my ending with a preposition, but I find it difficult to say otherwise. These parameter combinations are also called the "Expected Mean Squares". They are the long-run expectations of the MS statistics As Winston Churchill would say, "For the sake of clarity, this one time this wording is something up which you would put".

Example....

DATA:

	Subject					
	Tom	Anne	Yves	Jean	Claire	
Measurement						
1	4.8	5.5	5.1	6.4	5.8	4.5
2	4.7	5.2	4.9	6.2	6.3	4.1
3	4.9	5.2	5.3	6.6	5.6	4.0
Mean	4.8	5.3	5.1	6.4	5.9	4.2 Variance = 0.614
Variance	0.01	0.03	0.04	0.04	0.13	0.07

ANOVA TABLE (Check... I did it by hand!)

SOURCE	Sum of Squares SS	Degrees of Freedom df	**Mean Square** MS (= SS /df)	**What the Mean Square is an estimate of...** *
BETWEEN Subjects	9.205	5	1.841	$\sigma^2_W + n\,\sigma^2_B$
WITHIN Subjects	0.640	12	0.053	σ^2_W
TOTAL	9.845	17		

ESTIMATES OF VARIANCE COMPONENTS

$\text{MS}_{\text{WITHIN}}$ **= 0.053** is an unbiased estimate of σ^2_W

$\dfrac{1.841 - 0.053}{3}$ **= 0.596** is an unbiased estimate of σ^2_B

1-Way ANOVA Calculations performed by SAS; Components estimated manually

```
PROC GLM in SAS ==> estimating components 'by hand'
```

DATA a; INPUT Subject Value; LINES;
1 4.8
1 4.7
...
6 4.5

```
proc glm; class subject; model value=subject / ss3;
  random subject ;
```

See worked example using earsize data.
If unequal numbers of measurements per subject, see formula in A&B or Fleiss

Estimating Components of Variance using "Black Box"

`PROC VARCOMP; class subject ; model Value = Subject ;`
See worked example following...

2 measurements (in mm) of earsize of 8 subjects by each of 4 observers

```
subject    1                2                3                 4
obsr 1   2   3  4      1   2  3   4     1   2  3   4      1   2  3   4
1st|67 65 65 64     74 74 74 72     67 68 66 65     65 65 65 65
2nd|67 66 66 66     74 73 71 73     68 67 68 67     64 65 65 64

subject    5                6                7                 6
obsr 1   2   3  4      1   2  3   4     1   2  3   4      1   2  3   4
1st|65 62 62 61     59 56 55 53     60 62 60 59     66 65 65 63
2nd|61 62 60 61     57 57 57 53     60 65 60 58     66 65 65 65
```

INTRA-OBSERVER VARIATION (e.g. observer #1)

e.g. observer #1

```
PROC GLM in SAS ==> estimating components 'by hand'
INPUT subject rater occasion earsize; if observer=1;
  The data set has 16 obsns & 4 variables.

proc glm; class subject; model earsize=subject / ss3;
  random subject ;

 General Linear Models Procedure: Class Level Information

Class   Levels Values
SUBJECT    8    1 2 3 4 5 6 7 8 ; # of obsns. in data set = 16

Dependent Variable: EARSIZE
                      Sum of    Mean
Source             DF  Squares  Square  F Value  Pr > F
Model               7   341.00   48.71    35.43    0.0001
Error               8    11.00    1.38
Corrected Total    15   352.00

R-Square      C.V.    Root MSE       EARSIZE Mean
0.968750     1.80    1.17260            65.0

Source             DF  Type III SS  Mean Square  F Value   Pr > F
SUBJECT             7   341.00      48.71          35.43    0.0001

Source      Type III Expected Mean Square
SUBJECT     Var(Error) + 2 Var(SUBJECT)

Var(Error) + 2 Var(SUBJECT) = 48.71
Var(Error)                  =  1.38
             2 Var(SUBJECT) = 47.33
               Var(SUBJECT) = 47.33 / 2 = 23.67
```

Estimating Variance components using PROC VARCOMP in SAS

```
proc varcomp; class subject ; model earsize = subject ;

Variance Components Estimation Procedure: Class Level Information

Class    Levels    Values

SUBJECT    8    1 2 3 4 5 6 7 8 ; # obsns in data set = 16

MIVQUE(0) Variance Component Estimation Procedure

                         Estimate
Variance Component       EARSIZE

Var(SUBJECT)              23.67
Var(Error)                 1.38
```

- **ICC** (Fleiss § 1.3)

```
      Var(SUBJECT)                23.67
ICC = ------------------------- = -------------- = 0.94
      Var(SUBJECT) + Var(Error)   23.67 + 1.38
```

1-sided 95% Confidence Interval (see Fleiss p 12)

```
df for F in CI: (8-1)= 7 and 8

so from Tables of F distribution with 7 & 8 df, F = 3.5

So lower limit of CI for ICC is

      35.43  - 3.5
    = -------------------- = 0.82
      35.43 + (2 - 1)•3.5
```

EXERCISE: Carry out the estimation procedure for one of the other 3 observers.

INTERPRETING YOUR GRE SCORES

(Blurb from Educational Testing Service)

Your test score is an estimate, not a complete and perfect measure, of your knowledge and ability in the area tested. In fact, if you had taken a different edition of the test that contained different questions but covered the same content, it is likely that your score would have been slightly different. The only way to obtain perfect assessment of your knowledge and ability in the area tested would be for you to take all possible test editions that could ever be constructed. Then assuming that your ability and knowledge did not change, the average score on all those editions, referred to as your "true score," would be a perfect measure of your knowledge and ability in the content areas covered by the test. Therefore, scores are estimates and not perfect measures of a person's knowledge and ability. Statistical indices that address the imprecision of scores in terms of standard error of measurement and reliability are discussed in the next two sections.

STANDARD ERROR OF MEASUREMENT

The difference between a person's true and obtained scores is referred to as "error of measurement."* The error of measurement for an individual person cannot be known because a person's true score can never be known. The average size of these errors, however, can be estimated for a group of examinees by the statistic called the "standard error of measurement for individual scores:" The standard error of measurement for individual scores is expressed in score points. About 95 percent of examinees will have test scores that fall within two standard errors of measurement of their true scores. For example, the standard error of measurement of the GRE Psychology Test is about 23 points. Therefore, about 95 percent of examinees obtain scores in Psychology that are within 46 points of their true scores. About 5 percent of examinees, however, obtain scores that are more than 46 points higher or lower than their true scores.

Errors of measurement also affect any comparison of the scores of two examinees. Small differences in scores may be due to measurement error and not to true differences in the abilities of the examinees. The statistic "standard error of measurement of score differences" incorporates the error of measurement in each examinee's score being compared. This statistic is about 1.4 times as large as the standard error of measurement for the individual scores themselves. Approximately 95 percent of the differences between the obtained scores of examinees who have the same true score will be less than two times the standard error of measurement of score differences. Fine distinctions should not be made when comparing the scores of two or more examinees.

RELIABILITY

The reliability of a test is an estimate of the degree to which the relative position of examinees' scores would change if the test had been administered under somewhat different conditions (for example, examinees were tested with a different test edition).

Reliability is represented by a statistical coefficient that is affected by errors of measurement. Generally, the smaller the errors of measurement in a test, the higher the reliability. Reliability coefficients may range from 0 to 1, with 1 indicating a perfectly reliable test (i.e., no measurement error) and zero reliability indicating a test that yields completely inconsistent scores. Statistical methods are used to estimate the reliability of the test from the data provided by a single test administration. Average reliabilities of the three scores on the General Test and of the total scores on the Subject Tests range from .88 to .96 on recent editions. Average reliabilities of subscores on recent editions of the Subject Test range from .82 to .90.

Data regarding standard errors of measurement and reliability of individual GRE tests may be found in the leaflet *Interpreting Your GRE General and Subject Test Scores*, which will be sent to you with your GRE Report of Scores.

VALIDITY

The validity of a test—the extent to which it measures what it is intended to measure—can be assessed in several ways. One way of addressing validity is to delineate the relevant skills and areas of knowledge for a test, and then, when building each edition of the test, make sure items are included for each area. This is usually referred to as content validity. A committee of ETS specialists defines the content of the General Test, which measures the content skills needed for graduate study. For Subject Tests, ETS specialists work with professors in that subject to define test content. In the assessment of content validity, content representativeness studies are performed to ensure that relevant content is covered by items in the test edition.

Another way to evaluate the validity of a test is to assess how well test scores forecast some criterion, such as success in grade school. This is referred to as predictive validity. Indicators of success in graduate school may include measures such as graduate school grades, attainment of a graduate degree, faculty ratings, and departmental examinations. The most commonly used measure of success in assessing the predictive validity of the GRE tests is graduate first year grade point average. Reports on content representativeness and predictive validity studies of GRE tests may be obtained through the GRE Program office.

* The term "error of measurement" does not mean that someone has made a mistake in constructing or scoring the test. It means only that a test is an imperfect measure of the ability or knowledge being tested.

Outline

Some ways to quantify Reliability:

- Reliability Coefficient

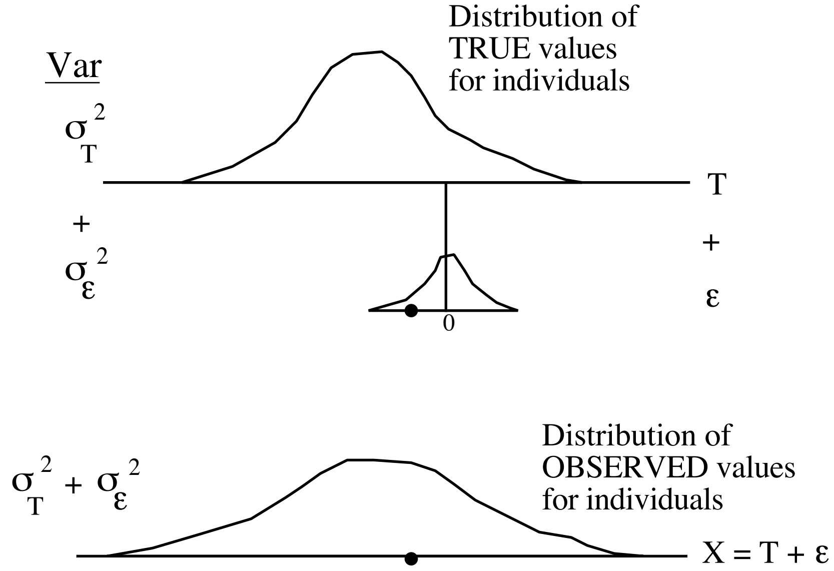
- Internal Consistency (Cronbach's α)

Implications:

-

....

Model for Reliability


Var
σ_T^2
+
σ_ε^2
Distribution of TRUE values for individuals
T
+
ε
0
$\sigma_T^2 + \sigma_\varepsilon^2$
Distribution of OBSERVED values for individuals
$X = T + \varepsilon$

"True" scores / values not knowable;

Variance calculation assumes that the distribution of errors is independent of T

Reliability Coefficient

$$r_{XX} = \frac{\sigma^2{}_T}{\sigma^2{}_T + \sigma^2{}_\varepsilon}$$ i.e. the fraction of observed variation that is 'real'

Note that one can 'manipulate' r by choosing a large or small $\sigma^2{}_T$

Effect of # of Items on Reliability Coefficient
(if all items have same variance and same intercorrelations)

SCALE 2 N Times more items than SCALE 1

$$r_{\text{SCALE 2}} = \frac{N \times r_{\text{SCALE 1}}}{1 + [N-1] \times r_{\text{SCALE 1}}}$$

e.g.

Scale	# Items	r
1	10	0.4
2	20 (× 2)	0.57
3	30 (× 3)	0.67

Cronbach's α

k items

$$\alpha = \frac{k \times \bar{r}}{1 + [k-1] \times \bar{r}}$$, where $\bar{r}$ = average of inter-item correlations

α is an estimate of the expected correlation of one test with an alternative form with the same number of items.

α is a lower bound for r_{XX} i.e $r_{XX} \geq \alpha$

$r_{XX} = \alpha$ if items are parallel.

parallel

Average [item 1] = Average [item 2] =Average [item 3] = ...

Variance[item 1] = Variance[item 2] =Variance [item 3] = ...

Correlation[item 1, item 2] = Correlation[item 1, item 3] = ...

=Correlation[item 2, item 3] = ...

INTRACLASS CORRELATIONS (ICC's)

- **Various versions**

  TEST-RETEST

  INTRA-RATER

  INTER-RATER...

- **Formed as Ratios of various Variances**

e.g. $$\frac{\sigma^2_{TRUE}}{\sigma^2_{TRUE} + \sigma^2_{ERROR}}$$

with estimates of various σ^2 's substituted for the σ^2 's .

Estimates of various components typically derived from ANOVA.

- **Note the distinction between DEFINITIONAL FORM (involving PARAMETERS) and COMPUTATIONAL FORM (involving STATISTICS)**

  Fleiss Chapter 1 good here; Norman & Streiner not so good!!)

ICC's (Portnoy and Wilkins)

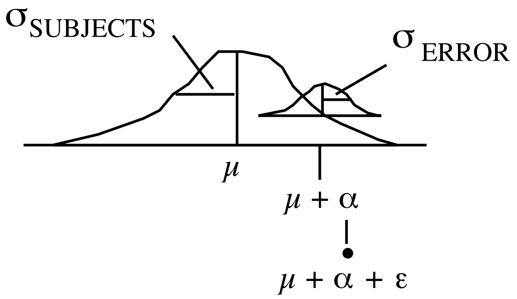
(1) multiple (unlabeled) measurements of each subject

(2) same set of raters measure each subject; raters thought of as a random sample of all possible raters.

(3) as in (2), but these raters studied are the only raters of interest

....

(1) multiple (unlabeled) measurements of each subject



$$\text{ICC} = \frac{\sigma^2_{\text{SUBJECTS}}}{\sigma^2_{\text{SUBJECTS}} + \sigma^2_{\text{ERROR}}}$$

Model for observed data:

y[subject i, measurement j] = $\mu + \alpha_i + \varepsilon_{ij}$

EXAMPLE 1

This example is in the spirit of the way the ICC was first used, as a measure of the greater similarity within families than between families: Study by Bouchard (NEJM) on weight gains of 2 members from each of 12 families: It is thought that there will be more variation between members of different families than between members of the same family: family (genes) is though to be a large source of variation; the two twins per family are thought of as 'replicates' from the family and closer to each other (than to others) in their responses. Here the "between" factor is family i.e. families are the subjects and the two twins in the family are just replicates and they don't need to be labeled (if we did label them 1 and 2, the labels would be arbitrary, since the two twins are thought to be 'interchangeable'. (weight gain in Kg over a summer)

model: weight gain for person j in family i = $\mu + \mu + \alpha_i + e_{ij}$

1-way Anova and Expected Mean Square (EMS)

Source	Sum of Sq	d.f	Mean Square	Expected Mean Square
Between (families)	99	11	9.0	$\sigma^2_{\text{"error"}} + k_0\, \sigma^2_{\text{between}}$
Error(Within families)	30	12	2.5	$\sigma^2_{\text{"error"}}$
Total	129	23		

In our example, we measure k=2 members from each family, so k_0 is simply 2

[if the k's are unequal, k_0 is somewhat less than the average k... k_0 = average k – (variance of k's) / (n times average k) ...see Fleiss page 10]

Estimation of parameters that go to make up ICC

2.5 is an estimate of $\sigma^2_{\text{"error"}}$

9.0 is an estimate of $\sigma^2_{\text{"error"}} + 2\,\sigma^2_{\text{between}}$

∴ 6.5 is an estimate of $2\,\sigma^2_{\text{between}}$

$\frac{6.5}{2}$ is an estimate of $\sigma^2_{\text{between}}$

$$\frac{\frac{6.5}{2}}{\frac{6.5}{2} + 2.5} = \frac{3.25}{3.25 + 2.5} = 0.57$$

is an estimate of ICC $= \dfrac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{error}}}$

COMPUTATIONAL Formula for "1-way" ICC

$$\frac{\dfrac{\text{MSbetween} - \text{MSwithin}}{k_0}}{\dfrac{\text{MSbetween} - \text{MSwithin}}{k_0} + \text{MSwithin}}$$

$$= \frac{\text{MSbetween} - \text{MSwithin}}{\text{MSbetween} + (k_0 - 1)\text{MSwithin}} \quad \text{[shortcut]}$$

is an estimate of the ICC

Notes:

- Streiner and Norman start on page 109 with the 2-way anova for inter-observer variation. There are mistakes in their depiction of the SSerror on p 110 [it should be $(6\text{-}6)^2+(4\text{-}4)^2+(2\text{-}1)^2 + ...(8\text{-})^2 = 10$. If one were to do the calculations by hand, one usually calculates the SStotal and then obtains the SSerror by subtraction]
- They then mention the 1-way case, which we have discussed above, as "the observer nested within subject" on page 112
- Fleiss gives methods for calculating CI's for ICC's.

EXAMPLE 2: INTRA-OBSERVER VARIATION FOR 1 OBSERVER

Computations performed on earlier handout...

Var(SUBJECT) = 23.67   Var(ERROR) = 1.38

$\widehat{\text{ICC}} = 23.67 / (23.67 + 1.38) = 0.94$

An estimated 94% of observed variation in earsize measurements by this observer is 'real' .. i.e. reflects true between-subject variability.

Note that I say 'an estimated 94% ...". I do this because the 94% is a statistic that is subject to sampling variability (94% is just a point estimate or a 0% Confidence Interval). An interval estimate is given by say a 95% confidence interval for the true ICC (lower bound of a 1-sided CI is 82% ... see previous handout)

Increasing Reliability by averaging several measurements

In 1-way model: $y_{i,j} = \mu + \alpha_i + e_{ij}$

where $\text{var}[\alpha_i] = \sigma^2_{\text{between subjects}}$; $\text{var}[e_{ij}] = \sigma^2_{\text{"error"}}$

Then if we average k measurements, i.e.,

$$\text{ybar}_i = \mu + \alpha_i + \text{ebar}_i$$

then

$$\underset{i}{\text{Var}}\,[\text{ybar}_i] = \sigma^2_{\text{between}} + \frac{\sigma^2_{\text{"error"}}}{k}$$

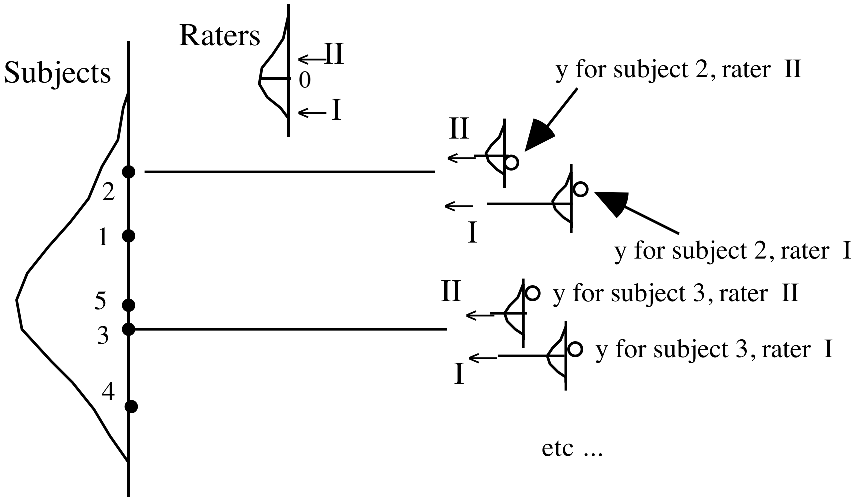
$$\text{So ICC[k]} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \dfrac{\sigma^2_{\text{"error"}}}{k}}$$

This is called **"Stepped-Up" Reliability**.

ICC's (Portnoy and Wilkins).

(2) same set of raters measure each subject; raters thought of as a random sample of all possible raters.

- **Model**



$$\mu \quad + \quad \alpha\,[\text{subject}] + \beta\,[\text{rater}] \quad + \quad \varepsilon$$

$$\sigma^2_{\text{subjects}} \qquad \sigma^2_{\text{raters}} \qquad \sigma^2_{\text{error}}$$

- **From 2- way data layout (subjects x Raters)**

estimate σ^2"subjects" , σ^2"raters" and σ^2"error" by 2-way ANOVA

- **Substitute variance estimates in appropriate ICC form**

e.g. 2 measurements (in mm) of earsize of 8 subjects by each of 4 observers

subject	1				2				3				4			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	67	65	65	64	74	74	74	72	67	68	66	65	65	65	65	65
2nd	67	66	66	66	74	73	71	73	68	67	68	67	64	65	65	64

subject	5				6				7				6			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	65	62	62	61	59	56	55	53	60	62	60	59	66	65	65	63
2nd	61	62	60	61	57	57	57	53	60	65	60	58	66	65	65	65

ESTIMATING INTER-OBSERVER VARIATION from occasion=1;

```
PROC GLM in SAS ==> estimating components 'by hand'
INPUT subject rater occasion earsize; if occasion=1; (32 obsns)

proc glm; class subject rater; model earsize=subject rater / ss3;
  random subject rater;

 General Linear Models Procedure: Class Level Information

Class     Levels    Values
SUBJECT        8    1 2 3 4 5 6 7 8
RATER          4    1 2 3 4 Number of observations in data set = 32

                         Sum of    Mean
Source              DF  Squares   Square  F Value  Pr > F
Model               10  764.500   76.45   78.80    0.0001
Error               21   20.375    0.97
Corrected Total     31  784.875

R-Square       C.V.  Root MSE   EARSIZE Mean
0.974040   1.534577   0.98501      64.1875

Source   DF  Type III SS  Mean Square  F Value  Pr > F
SUBJECT   7   734.875000   104.98      108.20  0.0001
RATER     3    29.625000     9.87       10.18  0.0002

Source      Type III Expected Mean Square
SUBJECT     Var(Error) + 4 Var(SUBJECT)
RATER       Var(Error) + 8 Var(RATER)
```

So... solving 'by hand' for the 3 components...

```
Var(Error) + 4 Var(SUBJECT) = 104.98
Var(Error)                  =   0.97
        ==>  4 Var(SUBJECT) = 104.01
        ==>    Var(SUBJECT) = 104.01 / 4 = 26.00

Var(Error) + 8 Var(RATER)   =   9.87
Var(Error)                  =   0.97
        ==>  8 Var(RATER)   =   8.90
        ==>    Var(RATER)   =   8.90 / 8 =  1.11

               Var(Error)               =  0.97
```

```
Estimating Variance components using PROC VARCOMP in SAS
proc varcomp; class subject rater; model earsize = subject rater;

                         Estimate
Variance Component        EARSIZE
Var(SUBJECT)             26.00
Var(RATER)                1.11
Var(Error)                0.97
```

- **ICC: "Raters Random"** (Fleiss § 1.5.2)

$$\underline{\mathrm{ICC}} = \frac{\mathrm{Var(SUBJECT)}}{\mathrm{Var(SUBJECT)} + \mathrm{Var(RATER)} + \mathrm{Var(Error)}} = \frac{26.00}{26.00+1.11+0.97} = 0.93$$

1-sided 95% Confidence Interval (see Fleiss p 27)

df for F in CI: (8-1)= 7 and v* , where

$$v^* = \frac{(8-1)(4-1)(4\bullet 0.93\bullet 10.18 + 8[1+(4-1)\bullet 0.93]-4\bullet 0.93)^2}{(8-1)\bullet 4^2\bullet 0.93^2\bullet 10.18^2 + (8[1+(4-1)\bullet 0.93]-4\bullet 0.93)^2} = 8.12$$

so from Tables of F distribution with 7 & 8 df, F = 3.5

So lower limit of CI for ICC is

$$= \frac{8(104.98 - 3.5\bullet 0.97)}{8\bullet 104.98 + 3.5\bullet[4\bullet 9.87 + (8\bullet 4 - 8 - 4)\bullet 0.97]} = \underline{0.78}$$

- **ICC: if use one "fixed" observer** (see Fleiss p 23, strategy 3)

$$\mathrm{ICC} = \frac{\mathrm{Var(SUBJECT)}}{\mathrm{Var(SUBJECT)} + \mathrm{Var(Error)}} = \frac{26.00}{26.00 + 0.97} = 0.96$$

lower limit of 95% 1-sided CI (eqn 1.49: F = 2.5 ; 7 & 7x3=21 df)

$$\mathrm{ICC} = \frac{104.98 - 2.5}{104.98 + (4-1)\bullet 2.5} = 0.91$$

USING ALL THE DATA SIMULTANEOUSLY

(can now estimate subject x Rater interaction .. i.e extent to which raters 'reverse themselves' with different subjects)

Components of variance when use **both** measurements (all 64 obsns)

```
proc varcomp;                          proc varcomp;
 class subject rater;                  class subject rater;
 model earsize = subject rater;        model earsize = subject rater
                                                       subject*rater;
```

```
                    Estimate
Variance Component   EARSIZE       Variance Component    EARSIZE

Var(SUBJECT)          25.52        Var(SUBJECT)            25.47
Var(RATER)             0.70        Var(RATER)               0.67
Var(Error)             1.37        Var(SUBJECT*RATER)       0.31
                                   Var(Error)               1.13
```

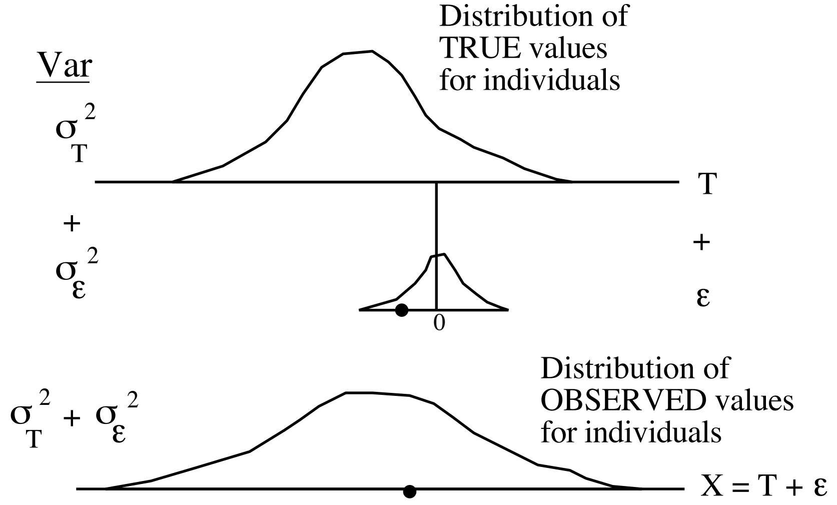
LINK between STANDARD ERROR OF MEASUREMENT and RELIABILITY COEFFICIENT

Example: GRE Tests (cf blurb from Educational Testing Service)

Standard Error of Measurement = 23 points

Reliability Coefficient: R = 0.93

recall...



$\sigma^2{}_e = 23$ ==> $\sigma^2{}_e = 529$;

$$R = \frac{\sigma^2{}_T}{\sigma^2{}_T + \sigma^2{}_e} = 0.93 \implies \sigma^2{}_T = \frac{R \times \sigma^2{}_e}{1 - R} = \frac{0.93 \times 529}{1 - 0.93} = 7028$$

$$\sigma^2{}_T + \sigma^2{}_e = 7028 + 529 = 7557 \implies \sqrt{\sigma^2{}_T + \sigma^2{}_e} = \sqrt{7557} = 87$$

So if 3 SD's on either side of the mean of 500 covers most of the observed scores, this would give a range of observed scores of 500 – 261 = 239 to 500 + 261 = 761.

Another way to say it (see Streiner and Norman, bottom of page 119) :-

$$\sigma_e = \sqrt{\sigma^2{}_T + \sigma^2{}_e} \times \sqrt{1 - R} = \text{SD[observed scores]} \times \sqrt{1 - R}$$

Confidence Intervals / Sample Sizes for ICC's

see Fleiss...

CI's based on F distribution tables;

CI's not symmetric;

More interested in 1-sided CI's i.e. (lower bound, 1) i.e. ICC ≥ 0.xx;

See also Donner and Eliasziw.

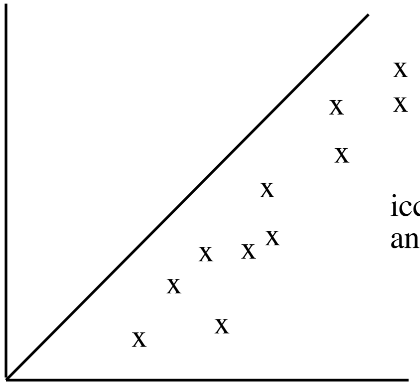
NOTE: If interested in ICC that incorporates random raters, then sample size must involve both # of raters and # of raters

CI will be very wide if use only 2 or 3 raters

Approach sample size as "n's or raters and subjects needed for a sufficiently narrow CI.

Why Pearson's r is not always a good [or practical] measure of reproducibility

1. It does not pick up "shifts"

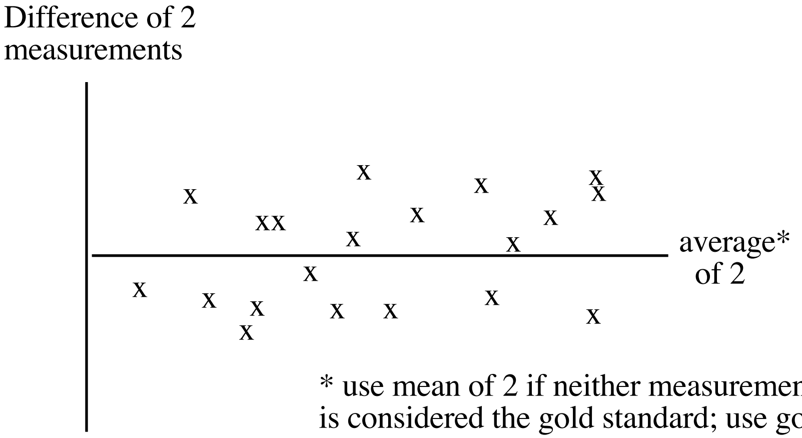


icc includes "shifts" and is lower than r

2. not practical if > 2 measurements or variable # of measurements per subject

ICC 'made for' such situations

Method of Bland & Altman [Lancet]



* use mean of 2 if neither measurements is considered the gold standard; use gold standard otherwise

+++ see biases quickly

can explain to your in-laws
(can you explain ICC to them?)

emphasises errors in measurements scale itself
(like ±23 in GRE score)

––– if don't know real range, magnitudes of standard error of measurement not helpful (see Norman & Streiner)

cannot use with > 2 measurements

doesn't generalize to raters

Assessing reproducibility of measurements made on a CATEGORICAL scale

Categorizations of n subjects by RATER 2

Categorizations of subjects by RATER 1

	C1	C2	C3	
C1				
C2				
C3				
				n

See chapter 13 in Fleiss's book on Rates and Proportions
or pp 516-523 of Chapter 26 of Portnoy and Wilkins

- Simple Measure

$$\% \text{ agreement} = \frac{\#\text{ in diagonal cells}}{n} \text{ x } 100$$

- Chance-Corrected Measure

$$\kappa = \frac{\%\text{ agreement} - \%\text{ agreement expected by chance*}}{100\%\text{ agreement} - \%\text{ agreement expected by chance}}$$

* expected proportion = $\sum$ p[row]*p[col] --- $\sum$ over the diagonals

(see Aickin's arguments against 'logic' of chance-correction:
Biometrics 199)

can give weights for 'partial' agreement

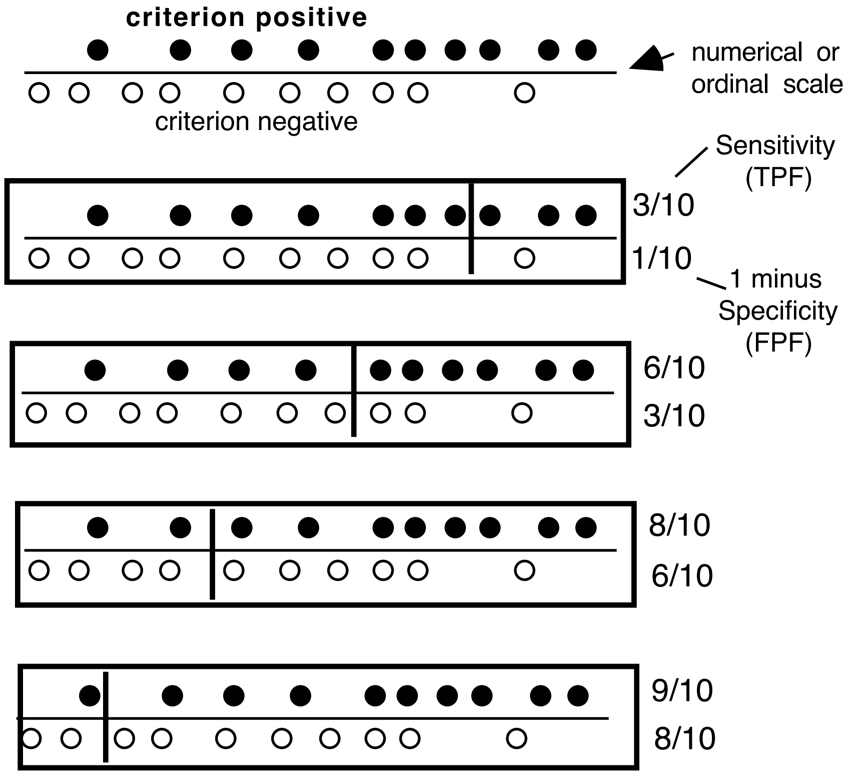
if > 2 raters, use range or average of pairwise kappas

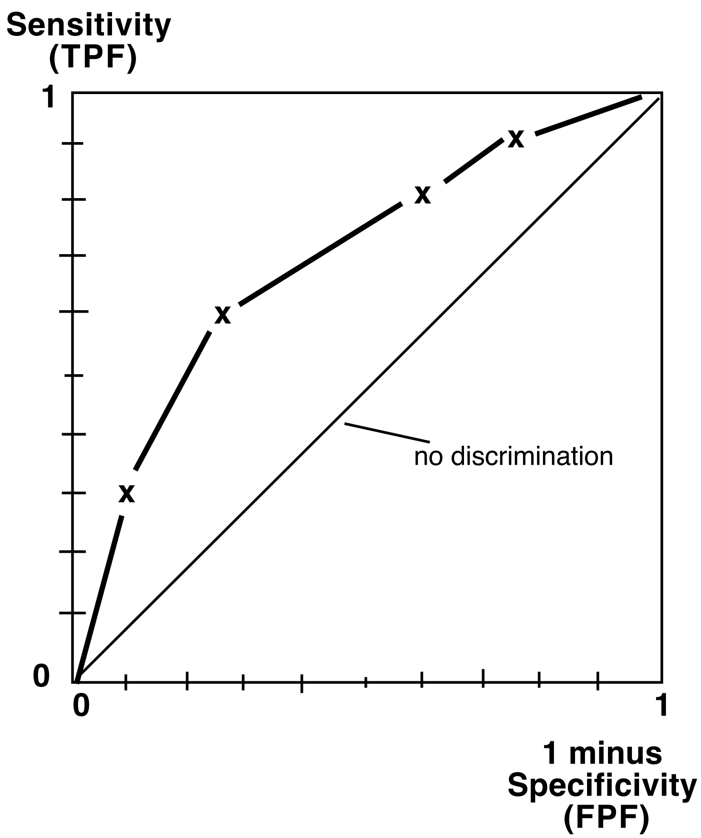
with quadratic weights, weighted kappa = icc

Receiver Operating Characteristic Curve

instrument	criterion
numerical or ordered scale	binary (• or o)

- SERIES of {sensitivity, specificity} statistics, each based on a different cut-off
- usually plotted on a graph, showing tradeoff between sensitivity and specificity
- Summary statistics (performance)
  - sensitivity at a given (specified) specificity
  - area under the ROC curve





TPF: True Positive Fraction
FPF: False Positive Fraction

Reference: section 5 chapter 13 in 2nd edition of Basic & Clinical Biostatistics by Beth Dawson-Saunders and Robert Trapp, Appleton & Lange, Norwalk (CT)