An Introduction to Classification

- Classification vs. Prediction
- Classification & ANOVA
- Classification Cutoffs, Errors, etc.
- Multivariate Classification & Linear Discriminant Function

Let's start by reviewing what "prediction" is…

- Using a person's scores on one or more variables to make a "best guess" of the that person's score on another variable (the value of which isn't known)

Classification is very similar …

- Using a person's scores on one or more variables to make a "best guess" of the category to which that person belongs (when the category type isn't known).

- The difference -- a language "convention"
  - if the "unknown variable" is quantitative -- its called prediction
  - if the "unknown variable" is qualitative -- its called classification

How does classification work???

Let's start with an "old friend" -- ANOVA

In its usual form…

- There are two qualitatively different IV groups
  - **naturally occurring or "created" by manipulation**
- A quantitative DV
- H0: $\text{Mean}_{G1} = \text{Mean}_{G2}$
- Rejecting H0: tells us
  - There is a relationship between the grouping and DV
  - Groups represent populations with different means on the DV
  - Knowing what group a person in allows us to guess their DV score -- mean of that group

Let's review in a little more detail…

Remember the formula for the ANOVA F-test

$$F = \frac{\text{variation between groups}}{\text{variation within groups}} = \frac{\text{size of the mean difference}}{\text{variation within groups}}$$

In words -- F compares the mean difference to the variability
around each of those means

Which of the following will produce the larger F-test ? Why ?

Data #1 (@ n = 50)

group 1 mean = 30
std dev = 5

group 2 mean = 50
std dev = 5

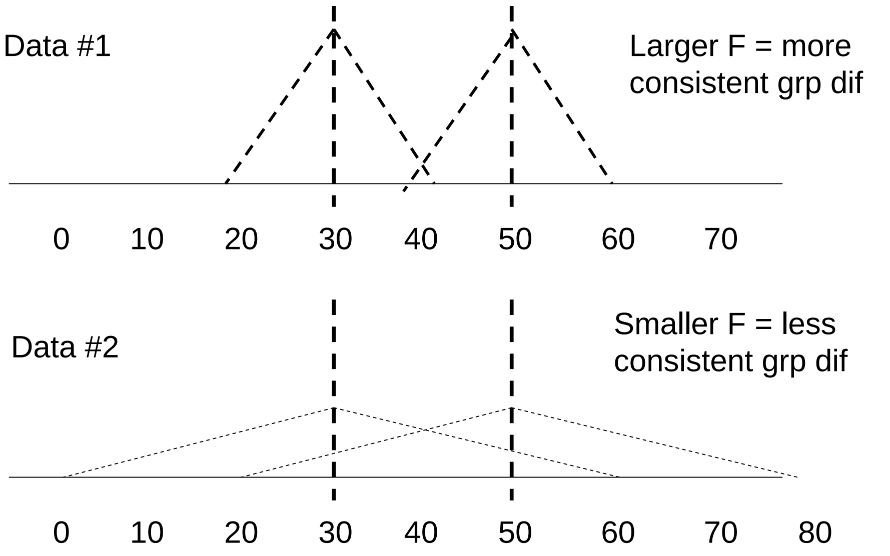
Data #2 (@ n = 50)

group 1 mean = 30
std dev = 15

group 2 mean = 50
std dev = 15

Remember -- about 96% of scores are within 2 std dev of mean

Graphical depictions of these data show that the size of F relates to the amount of overlap between the groups



Larger F = more consistent grp dif

Smaller F = less consistent grp dif

Notice: Since all the distributions have n=50, those with more variability are not as tall -- all 4 distributions have the same area

Let's consider that last one "in reverse"…

Could knowing the person's score help tell us what qualitative group they are in? …to "assign" them to the proper group?

an Example…

Research has revealed a statistical relationship between the number of times a person laughs out loud each day (quant variable) and whether they are depressed or schizophrenic (qual grouping variable).

Mean laughs$_{\textbf{Depressed}}$ = 4.0 Mean laughs$_{\textbf{Schizophrenic}}$ = 7.0

F(1,34) = 7.00, p < .05

A new (as yet undiagnosed) patient laughs 11 times the first day

what's your "assignment" depressed or schizophrenic?

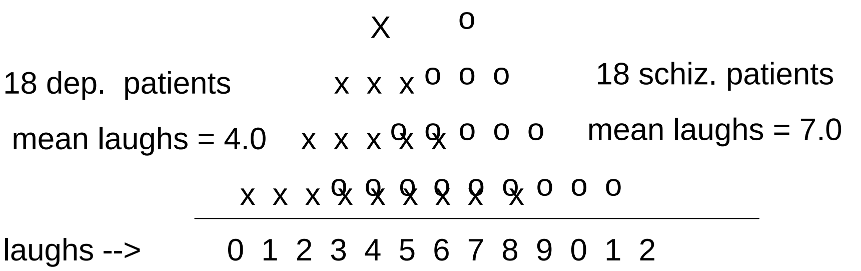
Another patient laughs 1 time -- your "assignment"?

A third new patient laughs 5 times -- your "assignment"?

Why were the first two “gimmies” and the last one not?

- When the groups have a mean difference, a score beyond one of the group means is more likely to belong to that group than to belong to the other group (unless stds are huge)
  - someone who laughs more than the mean for the schizophrenic group is more likely to be schizohrenic than to be depressed
  - someone who laughs less than the mean of the depressive group is more likely to be depressed than to be schizophrenic
- Even when the groups have a mean difference, a score between the group means is harder to correctly assign (unless stds are miniscule)
  - someone with 5-6 laughs are hardest to classify, because several depressed and schizophrenic folks have this score

Here’s a graphical depiction of the clinical data...



Looking at this, its easy to see why we would be ...

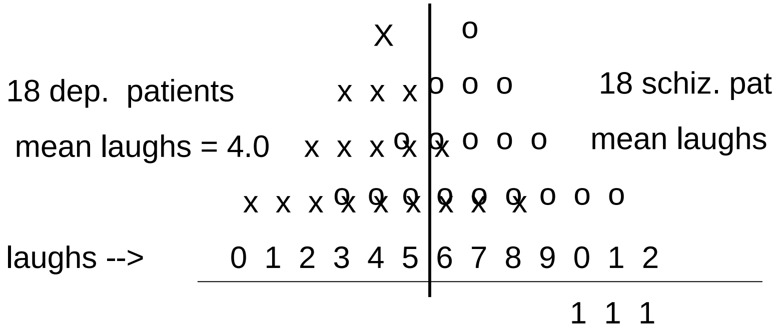
- confidant in an assignment based on 11 laughs
  - no depressed patients had a score that high
- confident in an assignment based on 1 laugh
  - no schizophrenic patients had a score that low
- lacking confidence in an assignment based on 5 or 6 laughs
  - several depressed & schizophrenic patients had 5 or 6

The process of prediction required two things…

- that there be a linear relationship between the predictor and the criterion (reject H0: r = 0)
- a formula (y’ = bx + a) to “translate” a predictor score into an estimate of a criterion variable score

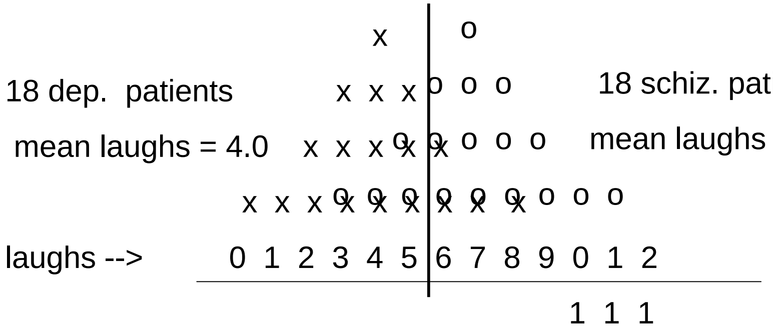
Similarly, the process of classification requires two things …

- a statistical relationship between the predictor (DV) & criterion (reject H0: $M_1 = M_2$)
- a cutoff to “translate” a person’s score on the predictor (DV) into an assignment to one group or the other
- where should be place the cutoff???
  - Wherever gives us the most accurate classification !!



When your groups are the same size and your group score distributions are symmetrical, things are pretty easy…

- place the cutoff at a position equidistant from the group means
- here, the cutoff would be 5.5 -- equidistant between 4.0 and 7.0
- anyone who laughs more than 5.5 times would be “assigned” as schizophrenic
- anyone who laughs fewer than 5.5 times would be “assigned” as depressed



We can assess the accuracy of the assignments by building a “reclassification table”

	Actual Diagnosis	
Assignment	Depressed	Schizophrenic
Depressed	14	4
Schizophrenic	4	14

reclassification accuracy would be 28/36 = 77.78%

Getting ready for ldf…

- multiple regression works better than simple regression because a y’ based on multiple predictors is a better estimate of y than a y’ based on a single predictor
- similarly, classification based on multiple predictors will do better than classification based on a single predictor
- but, how to incorporate multiple predictors into a classification ??
- Like with multiple regression, multiple variables (Xs) are each given a weighting and a constant is added
  - ldf = b_1* X_1 + b_2* X_2 + b_3* X_3 + a
  - the composite variable is called a linear discriminant function
    - function -- constructed from another variables
    - linear -- linear combination of linearly weighted vars
    - discriminant -- weights are chosen so that the resulting has the maximum possible F-test between the groups

So, how does this all work ???

- We start with a grouping variable and a set of quantitative (or binary) predictors (what would be DVs if doing ANOVAs)
- using an algorithm much like multiple regression, the bivariate relationship of predictor to the grouping variable & the collinearities among the predictors are all taken into account and the weights for the ldf formula are derived
  - remember this ldf will have the largest possible F value between the groups
- a cutoff value for the ldf is chosen the cutoff is chosen (more fancy computation) to maximize % correct reclassification
- to “use” the formula
  - a person’s values on the variables are put into the formula & their ldf score is computed
  - their score is compared to the cutoff, and they are assigned to one group or the other