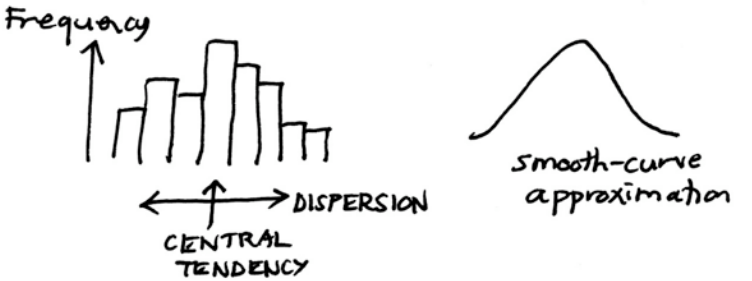
STATISTICS: A CONCEPTUAL OVERVIEW

I. Single variable, x, observed to vary in some defined population
Observations recorded for a sample of (members of) the population.

Q: What can be done on the basis of this knowledge? What more do we need to know?

A. Summarize observations



→ How would summaries change if sample were larger (perhaps, the whole population)?

How accurately was x measured?
How replicably was x defined?

Assume that members of population similar in x share underlying causes
Hypothesize about: the type of population ← Implied comparison with other populations
the causes of this type
the causes of spread away from type } ← Additional knowledge needed to hypothesize about causes
the causes of the variation

B. Summarize observations in ~~four~~ mixture of Gaussians groups ← Could the population be a mix of populations? (Extreme case: each point a separate type.)

Compare whether one group or two-group (or etc) is a more economical summary

Actual comparison with other groups

Assume & Hypothesize as in A. (But hypothesizing utilizes comparisons among groups)

Could the causes for the observations be heterogeneous within the population or the groups mixed into the population?

C. Don't assume similar x means similar causes
Seek additional information
through: measuring more variables, making interventions, eliciting insights from observed population members ...

STATISTICS 2/

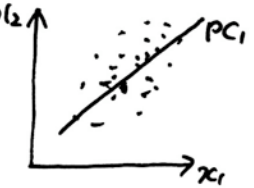
II. Many variables $x_1, x_2, x_3, \ldots$

A, B, C as for I, single variable

A. Summarize observations

Multidimensional space reduced to a smaller number of dimensions within which most variation occurs. ("principal components", "catena")

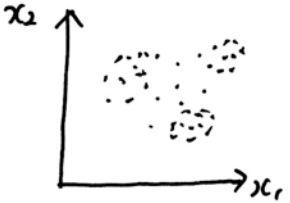
e.g.



Assume that members of a population similar in $PC_1, PC_2 \ldots$ share underlying causes

Hypothesize about: [see I]

B. Summarize observations as a mixture of groups (using cluster analysis)



Assume; hypothesize as in I.B.

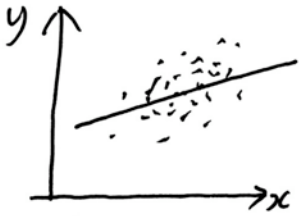
C. Seek additional information ~~about~~ re: underlying heterogeneity as in I.C.

(Additional measured variables influences our imagination about what the underlying factors could be.)

STATISTICS 3/

III. Many variables separated into dependent $y_1, y_2, y_3 \ldots$ and independent $x_1, x_2, \ldots$

A. Summarize observations
by minimizing residuals ("least squares")
or other criterion ("maximum likelihood")



regression line
(minimizes average (vertical distance)^2)

(± tricks to deal with binary y & categorical x)

Assume & hypothesize (as before)
about: causes that produce association summarized by regression line(s)

Treat population as if independent variable is causal (a "policy lever")*
Summarize observations by associating dependent variables with principal component variables

Assume & hypothesize (as before),
but hypothesizing more difficult for y's vs PCs

How many independent variables to use?

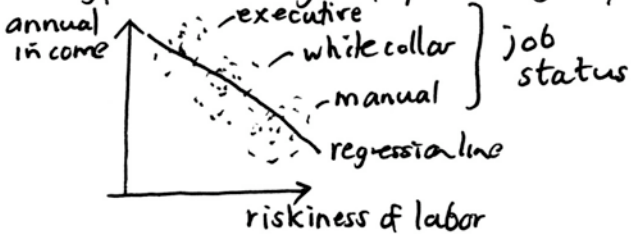
How to choose among correlated independent variables?

*e.g. association of health outcome in current data extended to situations where families are given $1000 check.

difficulty exposes → Hidden assumption that independent variables are causes*

Path diagram ("causal" or structural equation model) used to limit the range of possible associations

B. Revisit hypotheses in light of possible groups



← Knowledge about additional variable(s)
(here: job status)

STATISTICS 4/

IV. Two+ groups defined by experimental intervention, prior categorization (in contrast to emerging out of data analysis type B)

Examine ... difference ... between means in relation to average ~~spread~~ dispersion within the groups

(± extension to multiple groups "ANOVA")

Distinctions among groups make a difference in observed variable(s)?

Could division into groups be made differently — how are cut off points established?

Assume members of each group similar in ~~x~~ share underlying causes

Hypothesize about the causes of difference separately from causes of dispersion.

(But recall C re: possible underlying heterogeneity)

Treat groups differently

← Assumption that groups can be dealt with as separate types

↖ Assumption that underlying causes for current observations persist when data analysis result is turned into policy.