Winter Institute in Data Science and Big Data

Inference and Prediction

JEFF GILL
Distinguished Professor
Departments of Government, and Mathematics & Statistics
Center for Data Science
American University

Example: Fluoridated Water Supplies

- Research question: is fluorine in tap water related to cancer mortality?

Example: Fluoridated Water Supplies

- Research question: is fluorine in tap water related to cancer mortality?
- Burke & Yiamouyannis (1975):
  - 10 fluoridated and 10 non-fluoridated US towns
  - between 1950 and 1970
  - cancer deaths increased 20% in fluoridated and 10% in non-fluoridated.

Example: Fluoridated Water Supplies

- Research question: is fluorine in tap water related to cancer mortality?
- Burke & Yiamouyannis (1975):
  - 10 fluoridated and 10 non-fluoridated US towns
  - between 1950 and 1970
  - cancer deaths increased 20% in fluoridated and 10% in non-fluoridated.
- Oldham & Newell (1977):
  - controlled for changes in age and ethnicity in these same towns
  - cancer deaths increased 1% in fluoridated and 4% in non-fluoridated.

Example: Fluoridated Water Supplies

- Research question: is fluorine in tap water related to cancer mortality?
- Burke & Yiamouyannis (1975):
  - 10 fluoridated and 10 non-fluoridated US towns
  - between 1950 and 1970
  - cancer deaths increased 20% in fluoridated and 10% in non-fluoridated.
- Oldham & Newell (1977):
  - controlled for changes in age and ethnicity in these same towns
  - cancer deaths increased 1% in fluoridated and 4% in non-fluoridated.
- Real story: in fluoridated towns there was more out-migration of young and white residents (confounding variables).

Example: Fluoridated Water Supplies

- Research question: is fluorine in tap water related to cancer mortality?
- Burke & Yiamouyannis (1975):
  - 10 fluoridated and 10 non-fluoridated US towns
  - between 1950 and 1970
  - cancer deaths increased 20% in fluoridated and 10% in non-fluoridated.
- Oldham & Newell (1977):
  - controlled for changes in age and ethnicity in these same towns
  - cancer deaths increased 1% in fluoridated and 4% in non-fluoridated.
- Real story: in fluoridated towns there was more out-migration of young and white residents (confounding variables).
- The point: choice of model really matters.

Thoughts On Science

- It turns out that we are the "hard" science (Bob Keohane is wrong).
- Simon DeDeo, a research fellow in applied mathematics and complex systems at the Santa Fe Institute:

  > "In physics, you typically have one kind of data and you know the system really well," said DeDeo. "Now we have this new multimodal data [gleaned] from biological systems and human social systems, and the data is gathered before we even have a hypothesis." The data is there in all its messy, multi-dimensional glory, waiting to be queried, but how does one know which questions to ask when the scientific method has been turned on its head?

  `http://www.wired.com/wiredscience/2013/10/topology-data-sets/`

- New trends: big data, data engineering, biometric measurement, machine learning, sophisticated marketing research, google research, non-random samples from online surveys, text as data.

General Points About the Research Process

- Description always comes first.

- Generalization is the process of going from the particular to the general.

- Interpretation is placing events in context, e.g.
  –empathize with subject motives,
  –consider culture before research question.

- Another approach: study systematic patterns in similar and parallel events.

Characteristics of Science vs. Interpretation

- Preparing careful descriptions.
- Gaining deep understandings of the social world.
- Asking good questions.
- Formulating falsifiable hypotheses.
- Collecting evidence.
- Testing hypotheses.

The Necessity of Simplification

- We do not learn without simplification of natural phenomenon.

- Every model is a simplification/approximation and thus actually *wrong*. Therefore models are never "true," but good ones extract important features.

- KKV (p.43):

> ...the difference between the amount of complexity in the world and that in the thickest of descriptions is till vastly larger than the difference between this thickest of descriptions and the most abstract quantitative or formal analysis.

Models, Models, Models

- Advantages of restrictive models:
  - clear
  - parsimonious, easy to understand and explain
  - abstract
- Advantages of non-restrictive models:
  - detailed
  - contextual
  - realistic

Characteristics

- Characteristics of quantitative models:
  - looking at underlying trends and principles
  - usually symbolic and abstract
  - note: the quantification process produces precision but not necessarily accuracy since there is always measurement error.
- Characteristics of qualitative models:
  - good at seeing causality, but often not generalizable
  - complements description
  - provides nuance and detail otherwise unobservable.

Some Definitions

- Descriptive Model: a narrative simplification describing key causal factors.
- Statistical Model: has a systematic component (replicative) and a non-systematic component (varying).
- Formal Model: a purely mathematical representation of reality with no non-systematic component.
- A key distinction: do we think that it is a probabilistic world (there always exists variation), or a deterministic world (variation is just attributable to what we have not yet measured).

Types of Data Collection

- observation
- participant observation
- large scale sample surveys
- historical recordings from secondary sources
- randomized experiments
- ethnography
- content analysis

Terminology

Inference: using sample data and a model to make claims (estimates) about population parameters.

Prediction: using sample data and a model to make claims about future observations.

Statistic: a descriptive measure based on a population or sample data that does not depend on a parameter.

Model: a necessarily unrealistic picture of nature, a formal representation and simplification using symbology and assumptions.

On Models

- Formal/Mathematical Model: a mathematical and logical construct.

- Statistical Model: a probabilistic construct (has an error term).

$$Y_i = X_i\beta + e_i \qquad e \sim f(\sigma^2)$$

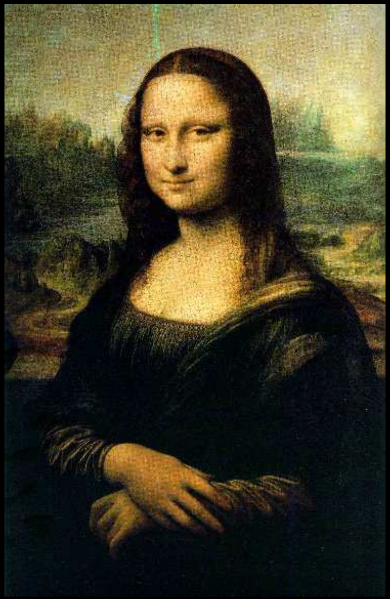
- Two models of humans...

On Models

- Formal/Mathematical Model: a mathematical and logical construct.
- Statistical Model: a probabilistic construct (has an error term).

$$Y_i = X_i\beta + e_i \qquad e \sim f(\sigma^2)$$

- Two models of humans...

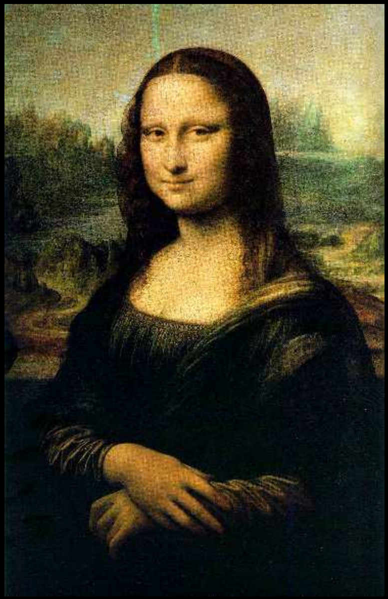


On Models

- Formal/Mathematical Model: a mathematical and logical construct.
- Statistical Model: a probabilistic construct (has an error term).

$$Y_i = X_i\beta + e_i \qquad e \sim f(\sigma^2)$$

- Two models of humans...



The Arbitrariness of Alpha

1. There is absolutely no theoretical justification for the standard significance thresholds.

2. *Level of significance* is therefore arbitrary and subject to custom by disciplines.

3. These values and the subsequent tables come from Fisher, whose work emphasized experimentation.

4. Fisher (1933):

    > "...the evidence would have reached a point which may be called the verge of significance; for it is convenient to draw the line at about the level at which we can say 'Either there is something in the treatment or a coincidence has occurred such as does not occur more than once in twenty trials.' This level, which we may call the 5 percent level point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials."

5. And:

    > "If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point) or one in a hundred (the 1 percent point). Personally, the writer prefers to set the low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach this level."

Sometimes People Just Really Want It

From `https://mchankins.wordpress.com`...

(barely) not statistically significant (p=0.052)

a barely detectable statistically significant difference (p=0.073) a borderline significant trend (p=0.09)

a certain trend toward significance (p=0.08)

a clear tendency to significance (p=0.052)

a clear trend (p<0.09)

a clear, strong trend (p=0.09)

a considerable trend toward significance (p=0.069)

a decreasing trend (p=0.09)

a definite trend (p=0.08)

a distinct trend toward significance (p=0.07)

a favorable trend (p=0.09)

a favourable statistical trend (p=0.09)

a little significant (p<0.1)

a margin at the edge of significance (p=0.0608)

a marginal trend (p=0.09)

a marginal trend toward significance ($p=0.052$)

a marked trend ($p=0.07$)

a mild trend ($p<0.09$)

a moderate trend toward significance ($p=0.068$)

a near-significant trend ($p=0.07$)

a negative trend ($p=0.09$)

a nonsignificant trend ($p<0.1$)

a nonsignificant trend toward significance ($p=0.1$)

a notable trend ($p<0.1$)

a numerical increasing trend ($p=0.09$)

a numerical trend ($p=0.09$)

a positive trend ($p=0.09$)

a possible trend ($p=0.09$)

a possible trend toward significance ($p=0.052$)

a pronounced trend ($p=0.09$)

a reliable trend ($p=0.058$)

a robust trend toward significance ($p=0.0503$)

a significant trend ($p=0.09$)

a slight slide towards significance ($p<0.20$)

a slight tendency toward significance($p<0.08$) a slight trend ($p<0.09$)

a slight trend toward significance ($p=0.098$)

a slightly increasing trend ($p=0.09$)

a small trend ($p=0.09$)

a statistical trend ($p=0.09$)

a statistical trend toward significance ($p=0.09$)

a strong tendency towards statistical significance ($p=0.051$) a strong trend ($p=0.077$)

a strong trend toward significance ($p=0.08$)

a substantial trend toward significance ($p=0.068$)

a suggestive trend ($p=0.06$)

a trend close to significance ($p=0.08$)

a trend significance level ($p=0.08$)

a trend that approached significance ($p<0.06$)

a very slight trend toward significance ($p=0.20$)

a weak trend ($p=0.09$)

a weak trend toward significance ($p=0.12$)

a worrying trend ($p=0.07$)

all but significant (p=0.055)

almost achieved significance (p=0-065)

almost approached significance (p=0.065)

almost attained significance ($p<0.06$)

almost became significant (p=0.06)

almost but not quite significant (p=0.06)

almost clinically significant ($p<0.10$)

almost insignificant ($p>0.065$)

almost marginally significant ($p>0.05$)

almost non-significant (p=0.083)

almost reached statistical significance (p=0.06)

almost significant (p=0.06)

almost significant tendency (p=0.06)

almost statistically significant (p=0.06)

an adverse trend (p=0.10)

an apparent trend (p=0.286)

an associative trend (p=0.09)

an elevated trend ($p<0.05$)

an encouraging trend ($p<0.1$)

an established trend ($p<0.10$)

an evident trend ($p=0.13$)

an expected trend ($p=0.08$)

an important trend ($p=0.066$)

an increasing trend ($p<0.09$)

an interesting trend ($p=0.1$)

an inverse trend toward significance ($p=0.06$)

an observed trend ($p=0.06$)

an obvious trend ($p=0.06$)

an overall trend ($p=0.2$)

an unexpected trend ($p=0.09$)

an unexplained trend ($p=0.09$)

an unfavorable trend ($p<0.10$)

appeared to be marginally significant ($p<0.10$)

approached acceptable levels of statistical significance ($p=0.054$)

approached but did not quite achieve significance ($p>0.05$)

approached but fell short of significance ($p=0.07$)

approached conventional levels of significance (p<0.10)

approached near significance (p=0.06)

approached our criterion of significance (p>0.08)

approached significant (p=0.11)

approached the borderline of significance (p=0.07)

approached the level of significance (p=0.09)

approached trend levels of significance (p0.05)

approached, but did reach, significance (p=0.065)

approaches but fails to achieve a customary level of statistical significance (p=0.154) approaches statistical significance (p>0.06)

approaching a level of significance (p=0.089)

approaching an acceptable significance level (p=0.056)

approaching borderline significance (p=0.08)

approaching borderline statistical significance (p=0.07)

approaching but not reaching significance (p=0.53)

approaching clinical significance (p=0.07)

approaching close to significance (p<0.1)

approaching conventional significance levels (p=0.06)

approaching conventional statistical significance (p=0.06) approaching formal significance ($p=0.1052$)

approaching independent prognostic significance ($p=0.08$)

approaching marginal levels of significance $p<0.107$)

approaching marginal significance ($p=0.064$)

approaching more closely significance ($p=0.06$)

approaching our preset significance level ($p=0.076$)

approaching prognostic significance ($p=0.052$)

approaching significance ($p=0.09$)

approaching the traditional significance level ($p=0.06$)

approaching to statistical significance ($p=0.075$)

approaching, although not reaching, significance ($p=0.08$) approaching, but not reaching, significance ($p<0.09$) approximately

significant ($p=0.053$)

approximating significance ($p=0.09$)

arguably significant ($p=0.07$)

as good as significant ($p=0.0502$)

at the brink of significance ($p=0.06$)

at the cusp of significance ($p=0.06$)

at the edge of significance ($p=0.055$)

at the limit of significance ($p=0.054$)

at the limits of significance ($p=0.053$)

at the margin of significance ($p=0.056$)

at the margin of statistical significance ($p<0.07$)

at the verge of significance ($p=0.058$)

at the very edge of significance ($p=0.053$)

barely below the level of significance ($p=0.06$)

barely escaped statistical significance ($p=0.07$)

barely escapes being statistically significant at the 5% risk level (0.1¿p>0.05) barely failed to attain statistical significance

($p=0.067$)

barely fails to attain statistical significance at conventional levels ($p<0.10$ barely insignificant ($p=0.075$)

barely missed statistical significance ($p=0.051$)

barely missed the commonly acceptable significance level ($p<0.053$)

barely outside the range of significance ($p=0.06$)

barely significant ($p=0.07$)

below (but verging on) the statistical significant level ($p>0.05$)

better trends of improvement ($p=0.056$)

bordered on a statistically significant value ($p=0.06$)

bordered on being significant ($p>0.07$)

bordered on being statistically significant ($p=0.0502$)

bordered on but was not less than the accepted level of significance ($p>0.05$) bordered on significant ($p=0.09$)

borderline conventional significance ($p=0.051$)

borderline level of statistical significance ($p=0.053$)

borderline significant ($p=0.09$)

borderline significant trends ($p=0.099$)

close to a marginally significant level ($p=0.06$)

close to being significant ($p=0.06$)

close to being statistically significant ($p=0.055$)

close to borderline significance ($p=0.072$)

close to the boundary of significance ($p=0.06$)

close to the level of significance ($p=0.07$)

close to the limit of significance ($p=0.17$)

close to the margin of significance ($p=0.055$)

close to the margin of statistical significance ($p=0.075$)

closely approaches the brink of significance ($p=0.07$)

closely approaches the statistical significance ($p=0.0669$)

closely approximating significance ($p>0.05$)

closely not significant ($p=0.06$)

closely significant ($p=0.058$)

close-to-significant ($p=0.09$)

did not achieve conventional threshold levels of statistical significance ($p=0.08$) did not exceed the conventional level of statistical significance ($p<0.08$)

did not quite achieve acceptable levels of statistical significance ($p=0.054$)

did not quite achieve significance ($p=0.076$)

did not quite achieve the conventional levels of significance ($p=0.052$)

did not quite achieve the threshold for statistical significance ($p=0.08$)

did not quite attain conventional levels of significance ($p=0.07$)

did not quite reach a statistically significant level ($p=0.108$)

did not quite reach conventional levels of statistical significance ($p=0.079$)

did not quite reach statistical significance ($p=0.063$)

did not reach the traditional level of significance ($p=0.10$)

did not reach the usually accepted level of clinical significance ($p=0.07$) difference was apparent ($p=0.07$)

direction heading towards significance ($p=0.10$)

does not appear to be sufficiently significant ($p>0.05$)

does not narrowly reach statistical significance (p=0.06)

does not reach the conventional significance level (p=0.098)

effectively significant (p=0.051)

equivocal significance (p=0.06)

essentially significant (p=0.10)

extremely close to significance (p=0.07)

failed to reach significance on this occasion (p=0.09)

failed to reach statistical significance (p=0.06)

fairly close to significance (p=0.065)

fairly significant (p=0.09)

falls just short of standard levels of statistical significance (p=0.06)

fell (just) short of significance (p=0.08)

fell barely short of significance (p=0.08)

fell just short of significance (p=0.07)

fell just short of statistical significance (p=0.12)

fell just short of the traditional definition of statistical significance (p=0.051) fell marginally short of significance (p=0.07)

fell narrowly short of significance (p=0.0623)

fell only marginally short of significance (p=0.0879) fell only short of significance (p=0.06) fell short of significance ($p=0.07$)

fell slightly short of significance ($p>0.0167$)

fell somewhat short of significance ($p=0.138$)

felt short of significance ($p=0.07$)

flirting with conventional levels of significance ($p>0.1$)

heading towards significance ($p=0.086$)

highly significant ($p=0.09$)

hint of significance ($p>0.05$)

hovered around significance ($p = 0.061$)

hovered at nearly a significant level ($p=0.058$)

hovering closer to statistical significance ($p=0.076$)

hovers on the brink of significance ($p=0.055$)

in the edge of significance ($p=0.059$)

in the verge of significance ($p=0.06$)

inconclusively significant ($p=0.070$)

indeterminate significance ($p=0.08$)

indicative significance ($p=0.08$)

is just outside the conventional levels of significance just about significant ($p=0.051$)

just above the arbitrary level of significance ($p=0.07$)

just above the margin of significance ($p=0.053$)

just at the conventional level of significance ($p=0.05001$)

just barely below the level of significance ($p=0.06$)

just barely failed to reach significance ($p<0.06$)

just barely insignificant ($p=0.11$)

just barely statistically significant ($p=0.054$)

just beyond significance ($p=0.06$)

just borderline significant ($p=0.058$)

just escaped significance ($p=0.07$)

just failed significance ($p=0.057$)

just failed to be significant ($p=0.072$)

just failed to reach statistical significance ($p=0.06$)

just failing to reach statistical significance ($p=0.06$)

just fails to reach conventional levels of statistical significance ($p=0.07$) just lacked significance ($p=0.053$)

just marginally significant ($p=0.0562$)

just missed being statistically significant ($p=0.06$)

just missing significance (p=0.07)

just on the verge of significance (p=0.06)

just outside accepted levels of significance (p=0.06)

just outside levels of significance (p<0.08)

just outside the bounds of significance (p=0.06)

just outside the conventional levels of significance (p=0.1076) just outside the level of significance (p=0.0683)

just outside the limits of significance (p=0.06)

just outside the traditional bounds of significance (p=0.06) just over the limits of statistical significance (p=0.06)

just short of significance (p=0.07)

just shy of significance (p=0.053)

just skirting the boundary of significance (p=0.052)

just tendentially significant (p=0.056)

just tottering on the brink of significance at the 0.05 level

just very slightly missed the significance level (p=0.086) leaning towards significance (p=0.15)

leaning towards statistical significance (p=0.06)

likely to be significant (p=0.054)

loosely significant (p=0.10)

marginal significance (p=0.07)

marginally and negatively significant (p=0.08)

marginally insignificant (p=0.08)

marginally nonsignificant (p=0.096)

marginally outside the level of significance

marginally significant (p>=0.1)

marginally significant tendency (p=0.08)

marginally statistically significant (p=0.08)

may not be significant (p=0.06)

medium level of significance (p=0.051)

mildly significant (p=0.07)

missed narrowly statistical significance (p=0.054) moderately significant (p>0.11)

modestly significant (p=0.09)

narrowly avoided significance (p=0.052)

narrowly eluded statistical significance (p=0.0789)

narrowly escaped significance (p=0.08)

narrowly evaded statistical significance (p>0.05)

narrowly failed significance (p=0.054)

narrowly missed achieving significance (p=0.055) narrowly missed overall significance ($p=0.06$)

narrowly missed significance ($p=0.051$)

narrowly missed standard significance levels ($p<0.07$) narrowly missed the significance level ($p=0.07$)

narrowly missing conventional significance ($p=0.054$)

near limit significance ($p=0.073$)

near miss of statistical significance ($p>0.1$)

near nominal significance ($p=0.064$)

near significance ($p=0.07$)

near to statistical significance ($p=0.056$)

near/possible significance($p=0.0661$)

near-borderline significance ($p=0.10$)

near-certain significance ($p=0.07$)

nearing significance ($p<0.051$)

nearly acceptable level of significance ($p=0.06$)

nearly approaches statistical significance ($p=0.079$) nearly borderline significance ($p=0.052$)

nearly negatively significant ($p<0.1$)

nearly positively significant ($p=0.063$)

nearly reached a significant level ($p=0.07$)

nearly reaching the level of significance ($p<0.06$) nearly significant ($p=0.06$)

nearly significant tendency ($p=0.06$)

nearly, but not quite significant ($p>0.06$) near-marginal significance ($p=0.18$)

near-significant ($p=0.09$)

near-to-significance ($p=0.093$)

near-trend significance ($p=0.11$)

nominally significant ($p=0.08$)

non-insignificant result ($p=0.500$)

non-significant in the statistical sense ($p>0.05$

not absolutely significant but very probably so ($p>0.05$) not as significant ($p=0.06$)

not clearly significant ($p=0.08$)

not completely significant ($p=0.07$)

not completely statistically significant ($p=0.0811$)

not conventionally significant ($p=0.089$), but..

not currently significant ($p=0.06$)

not decisively significant ($p=0.106$)

not entirely significant ($p=0.10$)

not especially significant ($p>0.05$)

not exactly significant ($p=0.052$)

not extremely significant ($p<0.06$)

not formally significant ($p=0.06$)

not fully significant ($p=0.085$)

not globally significant ($p=0.11$)

not highly significant ($p=0.089$)

not insignificant ($p=0.056$)

not markedly significant ($p=0.06$)

not moderately significant ($P>0.20$)

not non-significant ($p>0.1$)

not numerically significant ($p>0.05$)

not obviously significant ($p>0.3$)

not overly significant ($p>0.08$)

not quite borderline significance ($p>=0.089$)

not quite reach the level of significance ($p=0.07$)

not quite significant ($p=0.118$)

not quite within the conventional bounds of statistical significance ($p=0.12$) not reliably significant ($p=0.091$)

not remarkably significant ($p=0.236$)

not significant by common standards (p=0.099) not significant by conventional standards (p=0.10)

not significant by traditional standards (p<0.1)

not significant in the formal statistical sense (p=0.08)

not significant in the narrow sense of the word (p=0.29)

not significant in the normally accepted statistical sense (p=0.064)

not significantly significant but..clinically meaningful (p=0.072)

not statistically quite significant (p<0.06)

not strictly significant (p=0.06)

not strictly speaking significant (p=0.057)

not technically significant (p=0.06)

not that significant (p=0.08)

not to an extent that was fully statistically significant (p=0.06)

not too distant from statistical significance at the 10% level

not too far from significant at the 10% level

not totally significant (p=0.09)

not unequivocally significant (p=0.055)

not very definitely significant (p=0.08)

not very definitely significant from the statistical point of view (p=0.08) not very far from significance (p<0.092) not very significant ($p=0.1$)

not very statistically significant ($p=0.10$)

not wholly significant ($p>0.1$)

not yet significant ($p=0.09$)

not strongly significant ($p=0.08$)

noticeably significant ($p=0.055$)

on the border of significance ($p=0.063$)

on the borderline of significance ($p=0.0699$)

on the borderlines of significance ($p=0.08$)

on the boundaries of significance ($p=0.056$)

on the boundary of significance ($p=0.055$)

on the brink of significance ($p=0.052$)

on the cusp of conventional statistical significance ($p=0.054$) on the cusp of significance ($p=0.058$)

on the edge of significance ($p>0.08$)

on the limit to significant ($p=0.06$)

on the margin of significance ($p=0.051$)

on the threshold of significance ($p=0.059$)

on the verge of significance ($p=0.053$)

on the very borderline of significance (0.05<p<0.06)

on the very fringes of significance (p=0.099)

on the very limits of significance (0.1¿p>0.05)

only a little short of significance (p>0.05)

only just failed to meet statistical significance (p=0.051)

only just insignificant (p>0.10)

only just missed significance at the 5% level

only marginally fails to be significant at the 95% level (p=0.06)

only marginally nearly insignificant (p=0.059)

only marginally significant (p=0.9)

only slightly less than significant (p=0.08)

only slightly missed the conventional threshold of significance (p=0.062) only slightly missed the level of significance (p=0.058)

only slightly missed the significance level (p=0·0556)

only slightly non-significant (p=0.0738)

only slightly significant (p=0.08)

partial significance (p>0.09)

partially significant (p=0.08)

partly significant (p=0.08)

perceivable statistical significance (p=0.0501)

possible significance (p<0.098)

possibly marginally significant (p=0.116)

possibly significant (0.05<p>0.10)

possibly statistically significant (p=0.10)

potentially significant (p>0.1)

practically significant (p=0.06)

probably not experimentally significant (p=0.2)

probably not significant (p>0.25)

probably not statistically significant (p=0.14)

probably significant (p=0.06)

provisionally significant (p=0.073)

quasi-significant (p=0.09)

questionably significant (p=0.13)

quite close to significance at the 10% level (p=0.104)

quite significant (p=0.07)

rather marginal significance (p>0.10)

reached borderline significance (p=0.0509) reached near significance (p=0.07)

reasonably significant (p=0.07)

remarkably close to significance (p=0.05009)

resides on the edge of significance (p=0.10)

roughly significant (p>0.1)

scarcely significant (0.05<p>0.1)

significant at the .07 level

significant tendency (p=0.09)

significant to some degree (0<p>1)

significant, or close to significant effects (p=0.08, p=0.05) significantly better overall (p=0.051)

significantly significant (p=0.065)

similar but not nonsignificant trends (p>0.05)

slight evidence of significance (0.1>p>0.05)

slight non-significance (p=0.06)

slight significance (p=0.128)

slight tendency toward significance (p=0.086)

slightly above the level of significance (p=0.06)

slightly below the level of significance (p=0.068) slightly exceeded significance level (p=0.06)

slightly failed to reach statistical significance (p=0.061) slightly insignificant (p=0.07)

slightly less than needed for significance (p=0.08)

slightly marginally significant (p=0.06)

slightly missed being of statistical significance (p=0.08) slightly missed statistical significance (p=0.059)

slightly missed the conventional level of significance (p=0.061) slightly missed the level of statistical significance (p<0.10)

slightly missed the margin of significance (p=0.051)

slightly not significant (p=0.06)

slightly outside conventional statistical significance (p=0.051) slightly outside the margins of significance (p=0.08)

slightly outside the range of significance (p=0.09)

slightly outside the significance level (p=0.077)

slightly outside the statistical significance level (p=0.053) slightly significant (p=0.09)

somewhat marginally significant (p>0.055)

somewhat statistically significant (p=0.092) strong trend toward significance (p=0.08) sufficiently close to significance (p=0.07)

suggestive but not quite significant (p=0.061) suggestive of a significant trend (p=0.08) suggestive of statistical significance (p=0.06) suggestively significant (p=0.064)

tailed to insignificance (p=0.1)

tantalisingly close to significance (p=0.104) technically not significant ($p=0.06$)

teetering on the brink of significance ($p=0.06$)

tend to significant ($p>0.1$)

tended to approach significance ($p=0.09$)

tended to be significant ($p=0.06$)

tended toward significance ($p=0.13$)

tendency toward significance (p approaching 0.1)

tendency toward statistical significance ($p=0.07$)

tends to approach significance ($p=0.12$)

tentatively significant ($p=0.107$)

too far from significance ($p=0.12$)

trend bordering on statistical significance ($p=0.066$)

trend in a significant direction ($p=0.09$)

trend in the direction of significance ($p=0.089$)

trend significance level ($p=0.06$)

trend toward ($p>0.07$)

trending towards significance ($p>0.15$)

trending towards significant ($p=0.099$)

uncertain significance ($p>0.07$)

vaguely significant ($p>0.2$)

verged on being significant ($p=0.11$)

verging on significance ($p=0.056$)

verging on the statistically significant ($p<0.1$)

verging-on-significant ($p=0.06$)

very close to approaching significance ($p=0.060$)

very close to significant ($p=0.11$)

very close to the conventional level of significance ($p=0.055$)

very close to the cut-off for significance ($p=0.07$)

very close to the established statistical significance level of $p=0.05$ ($p=0.065$) very close to the threshold of significance ($p=0.07$)

very closely approaches the conventional significance level ($p=0.055$)

very closely brushed the limit of statistical significance ($p=0.051$) very narrowly missed significance ($p<0.06$)

very nearly significant ($p=0.0656$)

very slightly non-significant ($p=0.10$)

very slightly significant ($p<0.1$)

virtually significant ($p=0.059$)

weak significance ($p>0.10$) weakened..significance ($p=0.06$)

weakly non-significant (p=0.07)

weakly significant (p=0.11)

weakly statistically significant (p=0.0557)

well-nigh significant (p=0.11)

Lack of a Loss Function

- With the NHST there is no explicitly modeled consequence of making the wrong decision.
- Some decisions (policy) have direct consequences for citizens, employees, managers, and agencies in general.
- Yet hypothesis testing confuses inference and decision making since it "does not allow for the costs of possible wrong actions to be taken into account in any precise way" (Barnett 1973).
- Decision theory (Raiffa and Schlaifer 1961) is the logical adjunct to hypothesis testing that formalizes the cost of alternatives by explicitly defining the cost of making the wrong decision by specifying a loss function and associated risk for each alternative.
- Risk is just the average loss across decisions by using this decision rule for a given value of $\tilde{\beta}$.

Decision Theory

- Given a sample $\mathbf{X}$, we apply a decision rule, T, such that it dictates an action given observed $\mathbf{X}$:

$$T(X) = A, \qquad\qquad A \in \mathbb{A}$$

  where $\mathbb{A}$ is the allowable class of actions.

- This decision rule is based on a loss function, such that if $X = x$ is observed and our decision rule is T, then the loss is dictated by

$$\ell(A, T(X)),$$

  where smaller losses are preferred.

- Since loss is a *function*, it is difficult to compare two losses, so we use the average of the loss function, which is called Bayesian risk:

$$R_T(X) = E[\ell(A, T(X))] = \int_{\mathcal{A}} \ell(A, T(X))p(A|X)dA.$$

- $R_T(X)$ is admissable if no other risk is lower, otherwise it is inadmissable.

Decision Theory

- Consider this in the context of Frequentist point estimation, $\theta \in \Theta$,
  - ▷ sampling distribution: $f(\mathbf{x}|\theta)$
  - ▷ allowable actions: $A \in \mathbb{A}$
  - ▷ decision rule: $T(X) = A$ means θ_A chosen as estimate of θ
  - ▷ loss function: $\ell(\theta, \mathbf{X})$
  - ▷ Frequentist risk: $E_{\theta|\mathbf{x}}[\ell(\theta, \mathbf{X})] = \int_{\mathcal{X}} \ell(\theta, \mathbf{X}) f(\mathbf{x}|\theta) dx$.
- Examples of loss functions for true θ and estimated θ_A:
  - ▷ squared error loss: $\ell(A, T(X)) = (\theta - \theta_A)^2$
  - ▷ absolute error loss: $\ell(A, T(X)) = |\theta - \theta_A|$
  - ▷ $0 - 1$ loss for discrete state spaces: $\ell(A, T(X)) = 0$ if $\theta = \theta_A$, and 1 otherwise
  - ▷ interval loss: $\ell(A, T(X)) = 0$ if $CI_{1-\alpha}[\theta_A]$ covers θ, and 1 otherwise.

Decision Theory

- Suppose $X_1, \ldots, X_n$ are iid with true mean μ and true variance σ^2.
- Our decision rule is $T(X) = \bar{X}$, for estimating μ.
- The chosen loss function is squared error loss $\ell(A, T(X)) = (\mu - \bar{X})^2$.
- So the Frequentist risk is calculated by:

$$\begin{aligned}
R_T(X) &= E_{\mathbf{X}}[(\mu - \bar{X})^2] \\
&= E_{\mathbf{X}}[\mu^2 - 2\mu\bar{X} + \bar{X}^2] \\
&= \mu^2 - 2\mu E_{\mathbf{X}}[\bar{X}] + E_{\mathbf{X}}[\bar{X}^2] \\
&= \mu^2 - 2\mu E_{\mathbf{X}}[\bar{X}] + \mathrm{Var}[\bar{X}] + (E_{\mathbf{X}}[\bar{X}])^2 \\
&= \mathrm{Var}[\bar{X}] \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Confidence Intervals and Leamer Bounds in a Discrete-Time Survival Analysis of Unionization (Western 1995)

Table 1: Survival Analysis, Predicting Year Union Decline: 18 OECD Countries, 1973-1989

	Quasi-Likelihood:		99% Confidence	Leamer	Estimated Power
Explanatory Variable	Coef.	Std. Err.	Interval	Bounds	($\alpha = 0.01$)
Highly Reliable Estimates					
Constant	-2.42	0.78	[-4.43:-0.41]	(-5.24:-0.98)	0.70
Year	1.41	0.22	[0.84: 1.98]	(0.24: 1.41)	0.90
Economic Openness	0.12	0.04	[0.02: 0.22]	(0.04: 0.12)	0.70
Unemployment	1.18	0.39	[0.34: 2.03]	(0.89: 1.18)	0.85
Union Density (lagged)	-0.23	0.04	[-0.33:-0.13]	(-0.23:-0.11)	0.99
Decentralization	2.77	0.64	[1.11: 4.43]	(0.33: 2.77)	0.97
Left Government	-4.41	0.86	[-6.64:-2.18]	(-4.41:-1.71)	0.99
Highly Suspect Estimates					
Strike Activity($\times 10^4$)	0.11	2.75	[-6.97: 7.19]	(-1.50: 2.15)	0.01
Indeterminant Estimates					
(none)					

Setting for Causal Inference

- Causal Inference is concerned with what would have happened to case i's outcome variable, y_i, if it had received a different treatment level.

- Causal inference can also be considered as a special case of *prediction* under varying circumstances, only with much stricter assumptions than usual.

- Define T_c as the control, and T_t as the treatment, very generally defined.

- This is a highly active and highly controversial area of research.

A Critical Problem

- The manner in which the treatment is assigned is critical to experimental and observational studies.
- Suppose healthier people are more commonly given an effective treatment, then this exaggerates the actual effect.
- Suppose, alternatively, that healthier people are more commonly assigned to the control, then this mitigates the actual effect.
- In this context health is a confounding covariate.

Dealing with Confounders

- We want to compare treated and controlled units conditional on the confounding covariate.
- Simple solution: regress the outcome on two inputs: the treatment indicator and the confounding covariate (and any others).
- If the model is "correct," the coefficient on the treatment indicator estimates the average causal effect in the sample.
- This is an optimistic scenario with real data.

Omitted Variable Bias

- Suppose the "correct" model is given by:

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i,$$

  where T_i is the treatment and x_i is the confounding covariate, both for unit i.

- The model that ignores this confounding covariate is obviously:

$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i.$$

- Define a third model that attempts to measure the effect of the treatment on on the confounding effect:

$$x_i = \gamma_0 + \gamma_1 T_i + \nu_i.$$

- Insert this specification into the correct equation:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 T_i + \beta_2(\gamma_0 + \gamma_1 T_i + \nu_i) + \epsilon_i \\ &= \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1)T_i + \epsilon_i + \beta_2\nu_i. \end{aligned}$$

The Big Problem

- For a unit assigned the treatment, $T_i = 1$, the outcome y_i^1 observed and y_i^0 is the unobserved *counterfactual* outcome.
- Conversely, for a unit assigned the control, $T_i = 0$, the outcome y_i^0 observed and y_i^1 is the unobserved *counterfactual* outcome.
- The simple treatment effect is for unit $i$is:

$$TE_i = y_i^1 - y_i^0.$$

- The fundamental problem of causal inference is that only one of the right-hand-side terms can be realized, so TE_i can never be directly observed.
- Therefore we also cannot directly get the *average treatment effect*:

$$\overline{TE} = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0).$$

Dealing With the Big Problem

- Randomization: the "gold standard" available only for controlled experiments.
- Close Substitutes: using similar treatments on the same subject, splitting units, making strong assumptions about controls on units (eg. the baseline equals the control outcome).
- Model Adjustment: statistical manipulation of the data and/or the model such that a pseudo-experimental treatment results and the interpretation of the outcome variable is changed.
- The latter approach often gives better *internal validity* (supportable claims about the sample) than *external validity* (supportable claims about the population from which it is drawn).

Assumption Number 1 for Causal Regression Modeling

- Conditional on the confounding variables used in the model, $\mathbf{X}$, the distribution of the potential outcomes across treatment conditions, T is the same.

$$y^0, y^1 \perp T | X,$$

- This is called ignorability ("selection on observables"), and it means that we control for the pretreatment variables that correlate with the treatment and the outcome.

- If the probability of treatment selection is equal, conditional on modeled confounding covariates, then ignorability holds.

- Therefore if ignorability holds in this sense (and we have the "correct" model), then causal inferences are valid without considering the treatment assignment process.

Assumption Number 2 for Causal Regression Modeling

- *SUTVA* (Rubin 1980): "stable unit treatment value assumption.
- Two components:
  - For each unit there is only one form of treatment that was not received,

$$T_c \cup T_t = \Omega.$$

  - Treatment for one unit does not affect other units potential outcomes.

$$T_i \perp T_j \ \forall i \neq j.$$

Data Issue Number 1 for Causal Regression Modeling

- Overlap describes the extent to which the *range* of the data is the same across groups.
- Poor overlap between the two groups means that there are data cases for which no counterfactuals exist, and therefore the model has to make out-of-sample claims about the treatment effect.
- Fortunately overlap is easy to test.
- Perfect overlap (generally from a controlled experiment) demands much less of the specified model.

Data Issue Number 2 for Causal Regression Modeling

- Balance is the degree to which the covariate levels match-up across paired cases.

- Imbalance limits the comparison of group means: $\bar{y}_1 - \bar{y}_0$ for some outcome of interest because the covariate distribution differ in the sample.

- Consider a *true* treatment effect of interest, ζ, which is tested by the simple regression:

$$y_{it} = \beta_0 + \beta_1 x_i + \zeta + \epsilon_i, \qquad y_{ic} = \beta_0 + \beta_1 x_i + \epsilon_i.$$

- Averaging over the outcome variable means that the estimate of the treatment effect is:

$$\hat{\zeta} = \bar{y}_t - \bar{y}_c = \beta(\bar{x}_t - \bar{x}_c).$$

- The magnitude of the bias is therefore the extent to which the distributions of x_t and x_c are different (particularly in their variance).

Matching

- Matching is family of procedures that attempts to associate pairs of cases in the data as if they were identical units in the experimental sense.
- Typically this means disposing of cases in the larger of the two groups, after matching all of the cases from the smaller group.
- Sometimes with highly discrete data and a small number of covariates perfect matching is possible.
- Usually this is impossible and matches are done on multivariate nearest-neighbor criteria (such as *Mahalanobis distance*) or one number summary criteria (such as *propensity scores*).

Common Matching Strategies

- Minimize the summed distance between all i, j cases D_{ij} according to:
  - Exact: $D_{ij} = \begin{cases} 0 \text{ if } x_i = x_j \\ \infty \text{ if } x_i \neq x_j \end{cases}$.
  - Mahalanobis: $D_{ij} = (x_i - x_j)' S^{-1} (x_i - x_j)$, where S is the sample variance/covariance matrix.
  - Propensity Score: $D_{ij} = |e_i - e_j|$, where e_k is unit k's probability of receiving the treatment given the observed covariates.

A Note On Post-Treatment Covariates

- These are controversial in their use:
  - "...all post-treatment variables require careful evaluation and thought before they are used as covariates." (Greiner & Rubin 2011).
  - "...it is generally not a good idea to control for variables measured *after* the treatment. (Gelman & Hill 2007).
  - Post-treatment variables require adjustment of estimated treatment effects (Frangakis & Rubin 2002).
  - Pretreatment effects can be mitigated by post-treatment variables (Pearl 1995).

Problems with Causal Statements

- Causal vs. Correlational
- Reluctance to use causal
- Erroneous use of causal
- Other language problems
- Explanatory Variables = key causal variable(s)‘ $\bigcup$ control variables
- Ω for the key causal variable has at least two states: treatment and control

Realized Causal Effect

- Example: the Fourth Congressional District of New York, 1988
- Fraction of vote for Democratic incumbent: y_4^I
- Fraction of vote for hypothetical Democratic non-incumbent: y_4^N
- Realized Causal Effect of Incumbency: $y_4^I - y_4^N$
- Effect only defined in theory since the two quantities are not observable (counterfactual).

Fundamental Problem of Causal Inference

- No matter how perfect the research design...
- no matter how much data are collected...
- no matter how perceptive the field workers are...
- no matter how hard the research associates work...
- and no matter how much experimental control exists:
- we can *never* know causal inferences for certain.

The Systematic Component

- Random Causal Effect for unit i: $Y_i^I - Y_i^N$
- Mean Causal Effect:

$$\begin{aligned} MCE &= E(Y_i^I - Y_i^N) \\ &= E(Y_i^I) - E(Y_i^N) \\ &= \mu_i^I - \mu_i^N \end{aligned}$$

- Variance of the Causal Effect: $Var(Y_i^I - Y_i^N)$
- Qualitative Example: the fall of the Soviet Union and choice of government types.

More Definitions

- Causal Mechanisms: qualitative explainers.
- Multiple Causality: a plurality of causes, combinations of different explanatory variables, seen by many variables and fewer cases.
- Symmetric and Asymmetric Causality (direction of effect).

Two Assumptions Around the Fundamental Problem

- Unit Homogeneity:

$$\mu_I^i = \mu_I^j \forall \quad i, j$$
$$\mu_N^i = \mu_N^j \forall \quad i, j$$

- Conditional Independence: the values of the explanatory variables are not caused by levels of the outcome variables.

Rules for Constructing Causal Theories

- Construct Falsifiable Theories
- Build Theories That Are Internally Consistent
- Select Outcome Variables Carefully
- Maximize Concreteness
- State Theories in as Encompassing Ways as Feasible

Okay, maybe I buy all of this. What should I do now?

- Don't "over-scientize" the discussion.
- Discuss null results carefully.
- Don't put stars on tables.
- Use p-values only if you have to or you need them to make some point.
- Don't fixate on p-value thresholds if you use them.
- Confidence intervals are preferred to t-statistics.
- Separate statistical significance and substantive significance in your writing.
- Tables should reflect the actual number of significance digits in the result.
- Graphical displays are often better than tables or descriptions.
- Make causal statements carefully and with support.

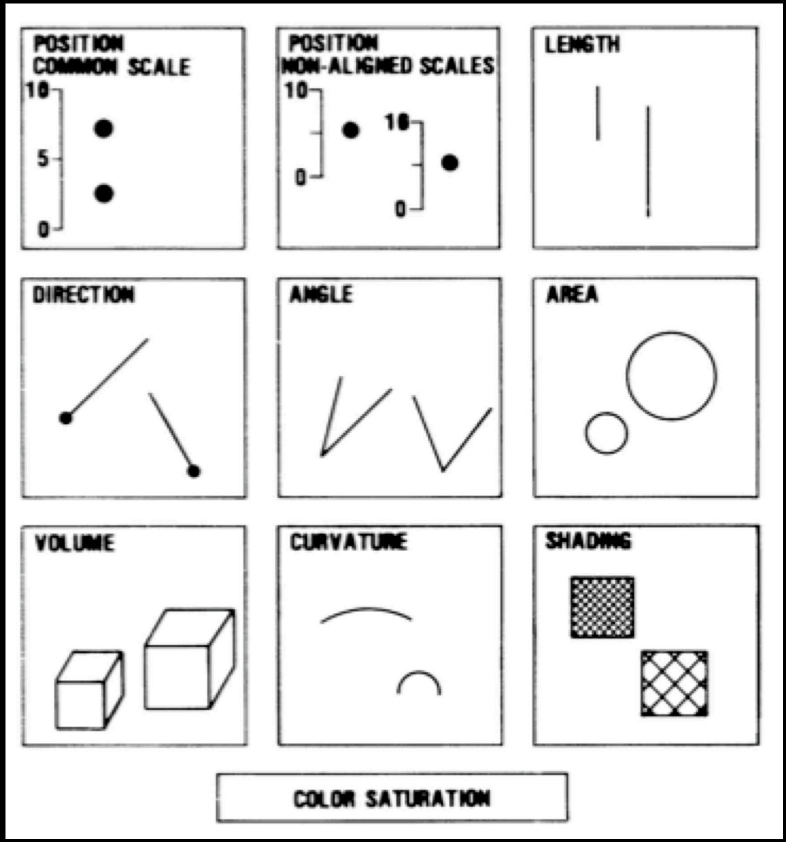
Obvious Points About Graphics

- Graphics are an important means of conveying statistical/empirical information to increasingly impatient readers and audience members.
- There is no real scientific foundation (although some attempts) to statistical graphics.
- Key technology: ocular reception, psychological perception, cultural interpretation.
- The first known graphic was on a Mesopotamian clay tablet dated 3800 BCE.
- Key historical figures: Halley Playfair, Tukey, Tufte, Cleveland (see the excellent Appendix in Beniger and Robyn, "Quantitative Graphics in Statistics: A Brief History." *The American Statistician*, 32:1-11 [1978]).

Cleveland & McGill's Elementary Perceptual Tasks

Ordered by Perceptual Accuracy:

- ▷ position along a common scale
- ▷ position along non-aligned scales
- ▷ length, direction, angle
- ▷ area
- ▷ volume, curvature
- ▷ shading, color saturation



Findings From Theoretical Psychophysics (Stevens 1975)

- Define: $p =$ perceived magnitude, $a =$ actual magnitude.
- Then for "inaccuracy" parameters k and α:

$$p = ka^{\alpha}$$

- For comparing magnitudes (assuming that $p_1 > p_2$ is a fact):

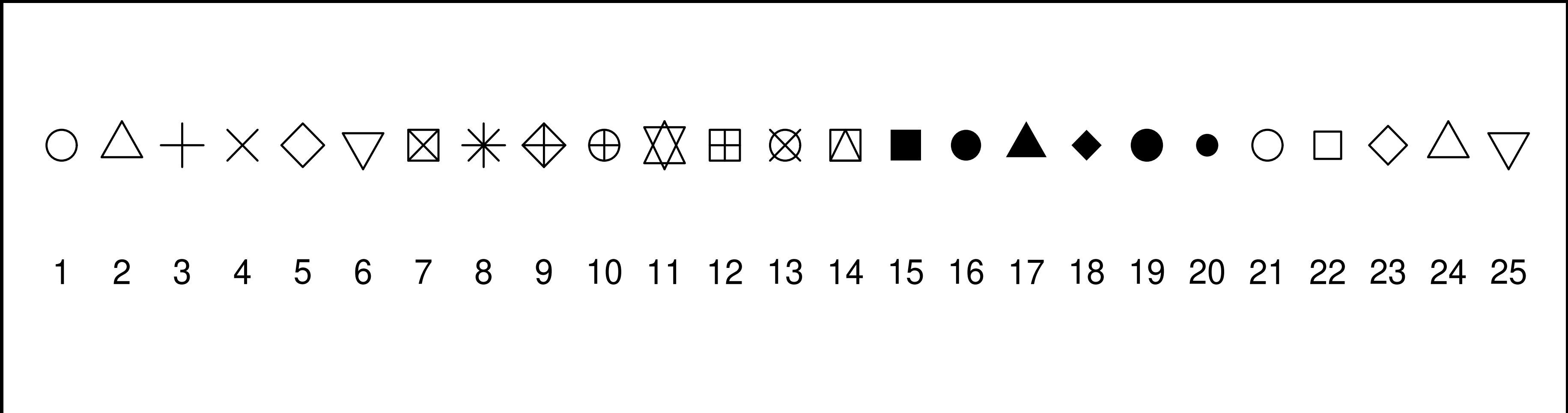
$$\frac{p_1}{p_2} = \left[\frac{a_1}{a_2}\right]^{\alpha}$$

- Findings:
  - $\alpha \approx 1$ for length judgments
  - $\alpha < 1$ for area judgments
  - $\alpha \ll 1$ for volume judgments

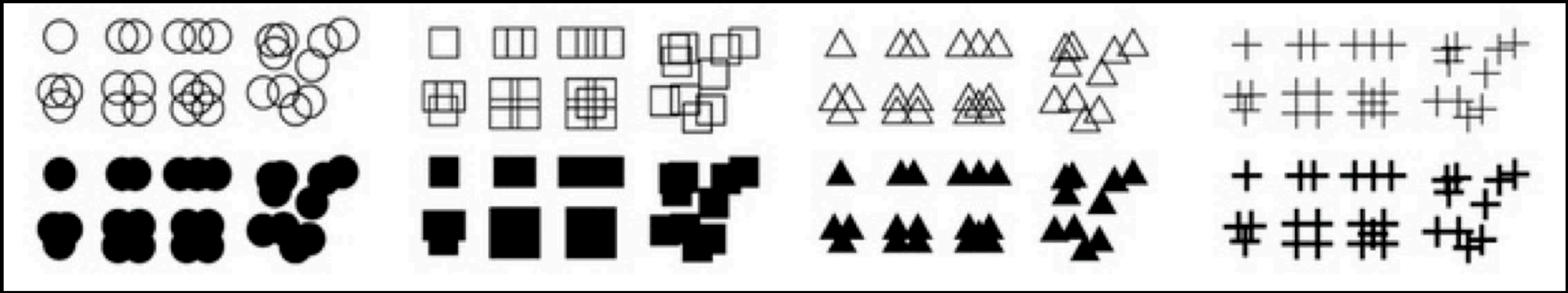
Symbols in Scatterplots

- With only one category of data, symbol choice is easy.
- Multiple categories should be encoded with distinct symbols that form strong visual boundaries.
- Symbols that have similar appearances can be easily missed on first inspection, especially in regions where symbols overlap.
- Insufficient symbol contrast can make it difficult to identify each data category.

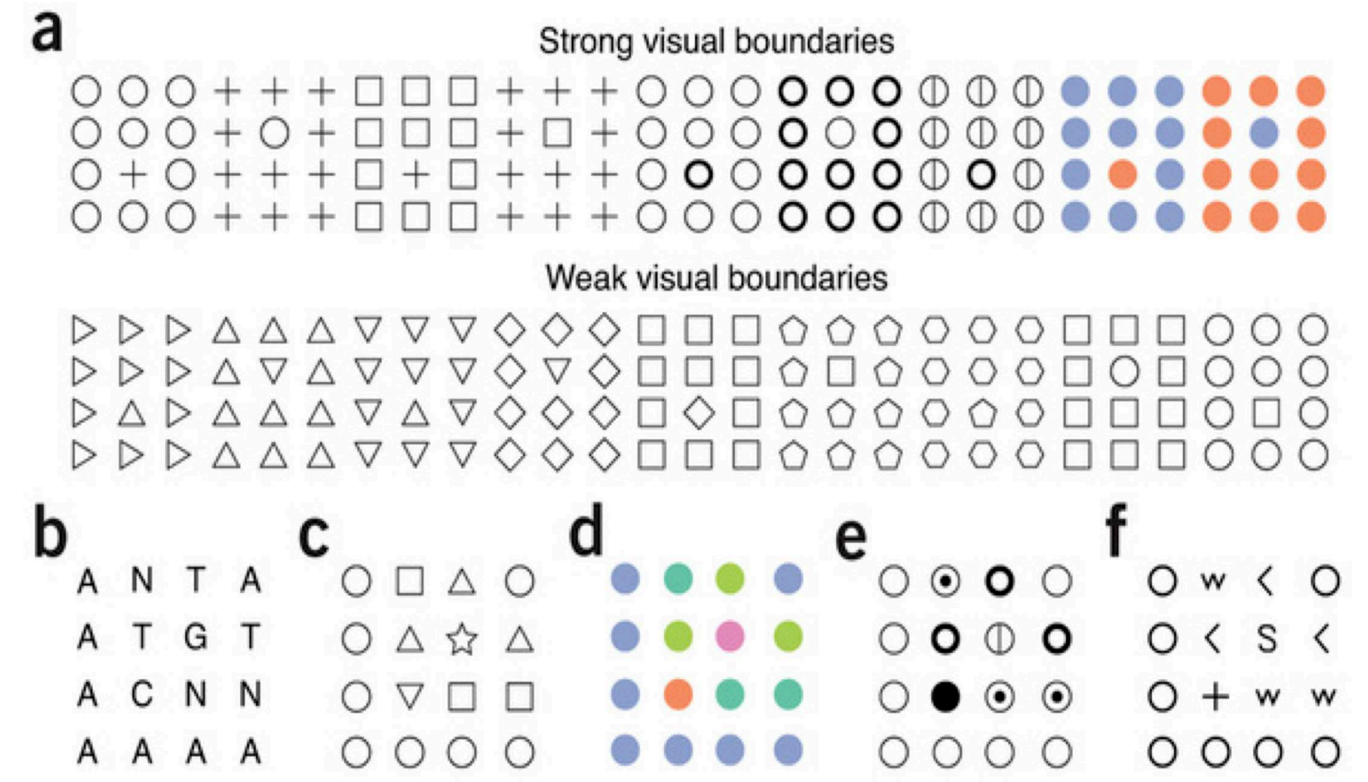
The numbered choices in `R` (`pch = 32:127` gives the ASCII characters):



Hollow Versus Filled-In Objects

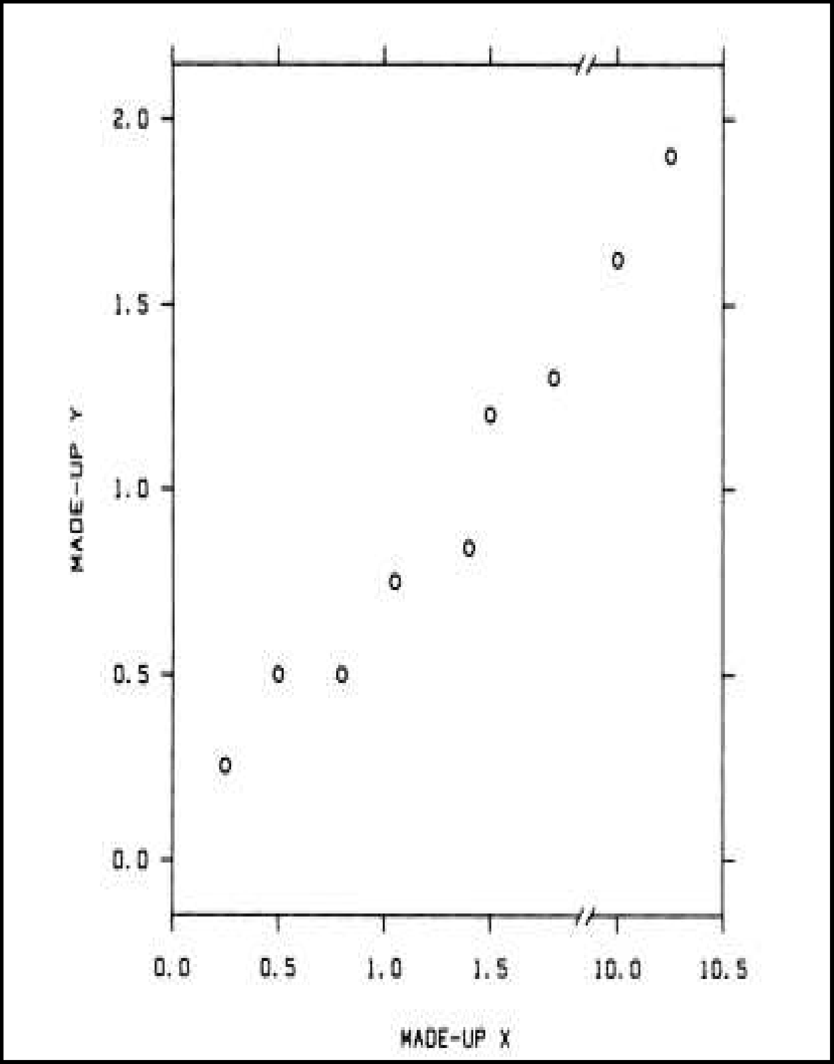


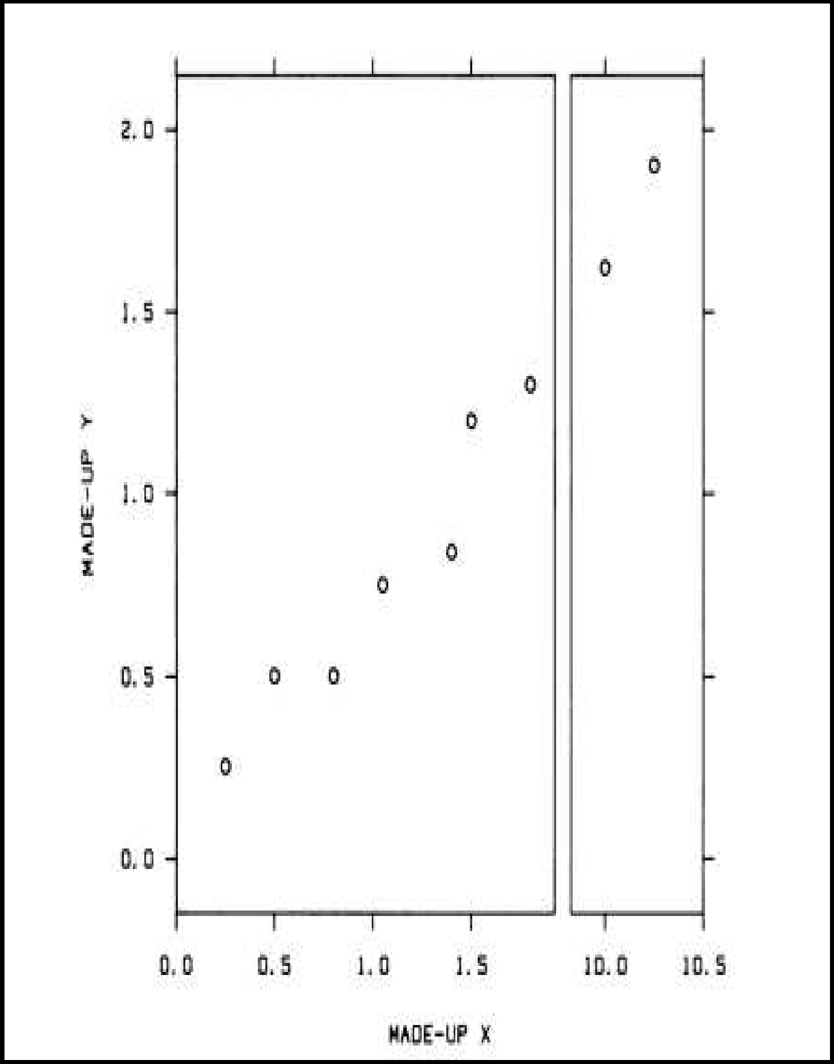
Obtaining Symbol Diversity



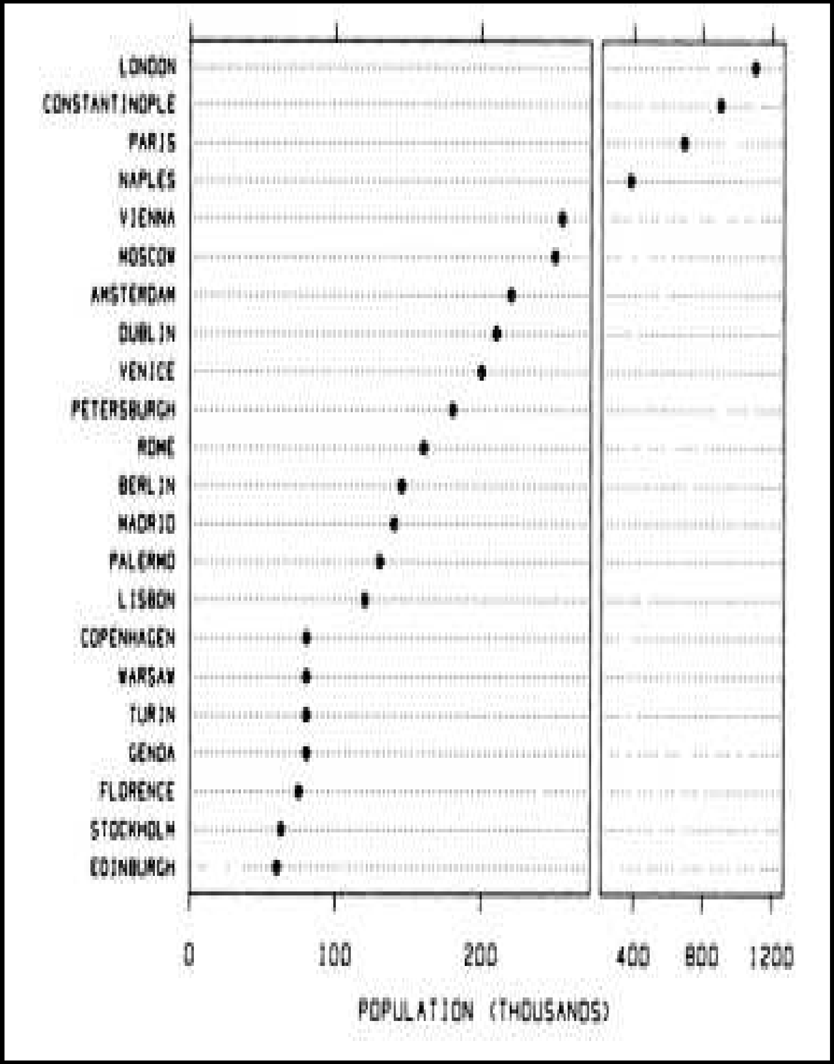
(**a**) Symbols that contrast with one another make good combinations. (**b**) Letters simplify legend lookups, but many appear the same (such as C/G, B/R/P and E/F/H). (**c**) Shapes are powerful discriminators—but beware that, for a given width, they may appear to have different sizes owing to differences in areas. (**d–f**) Color is one of the differentiators (**d**). For black-and-white applications, vary the fills for low data densities (**e**) and use texture symbols when overlap is high (**f**).

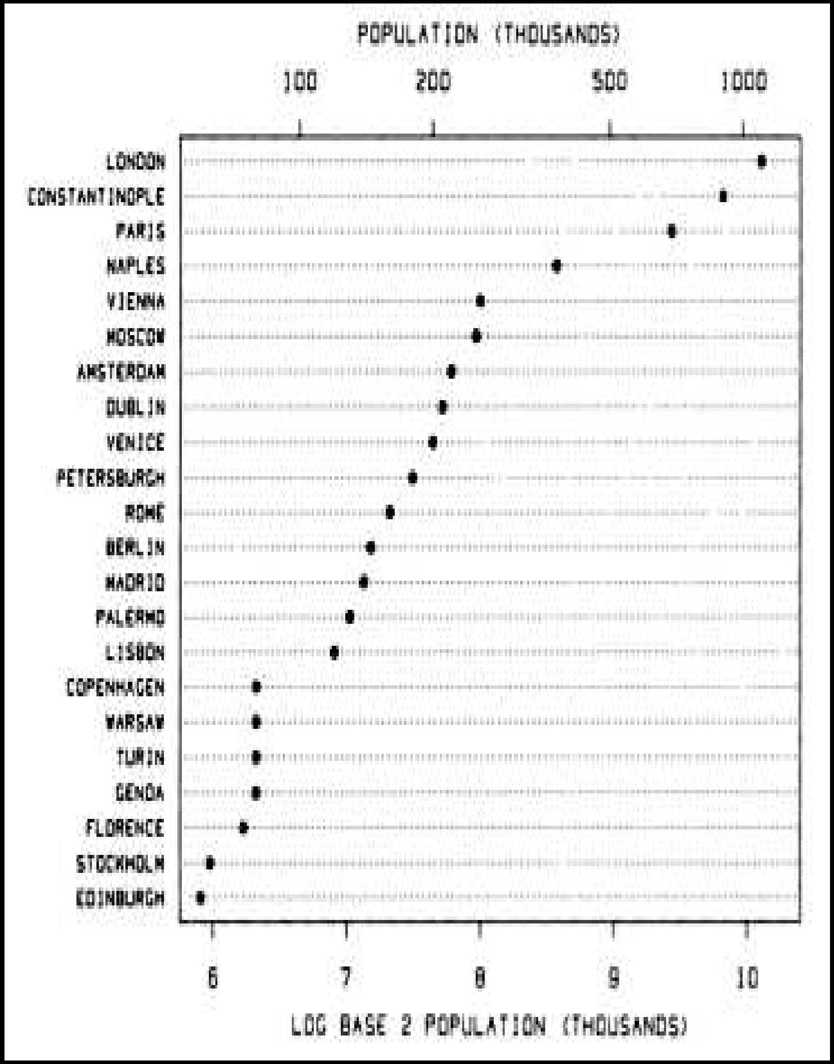
Warning #1: Breaks



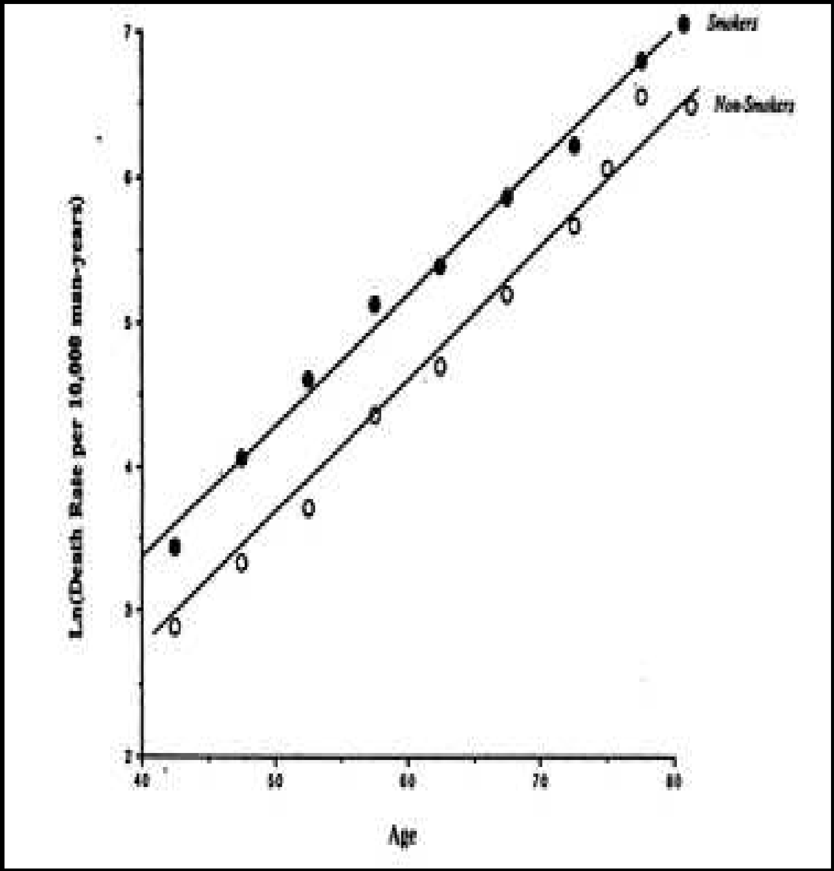


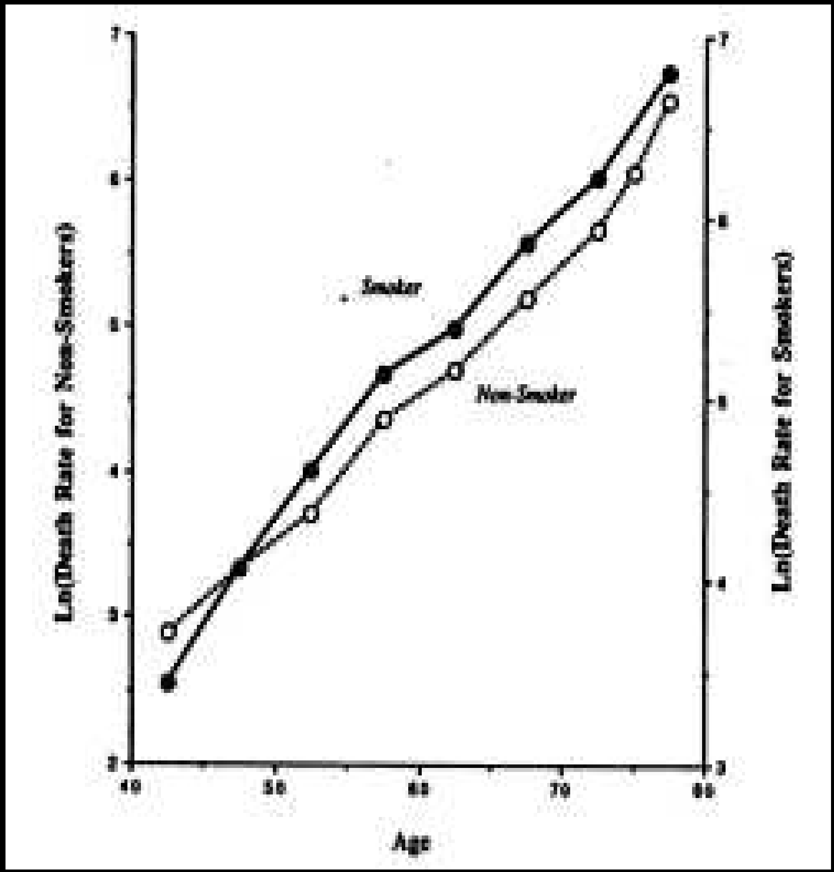
Warning #2: Scale



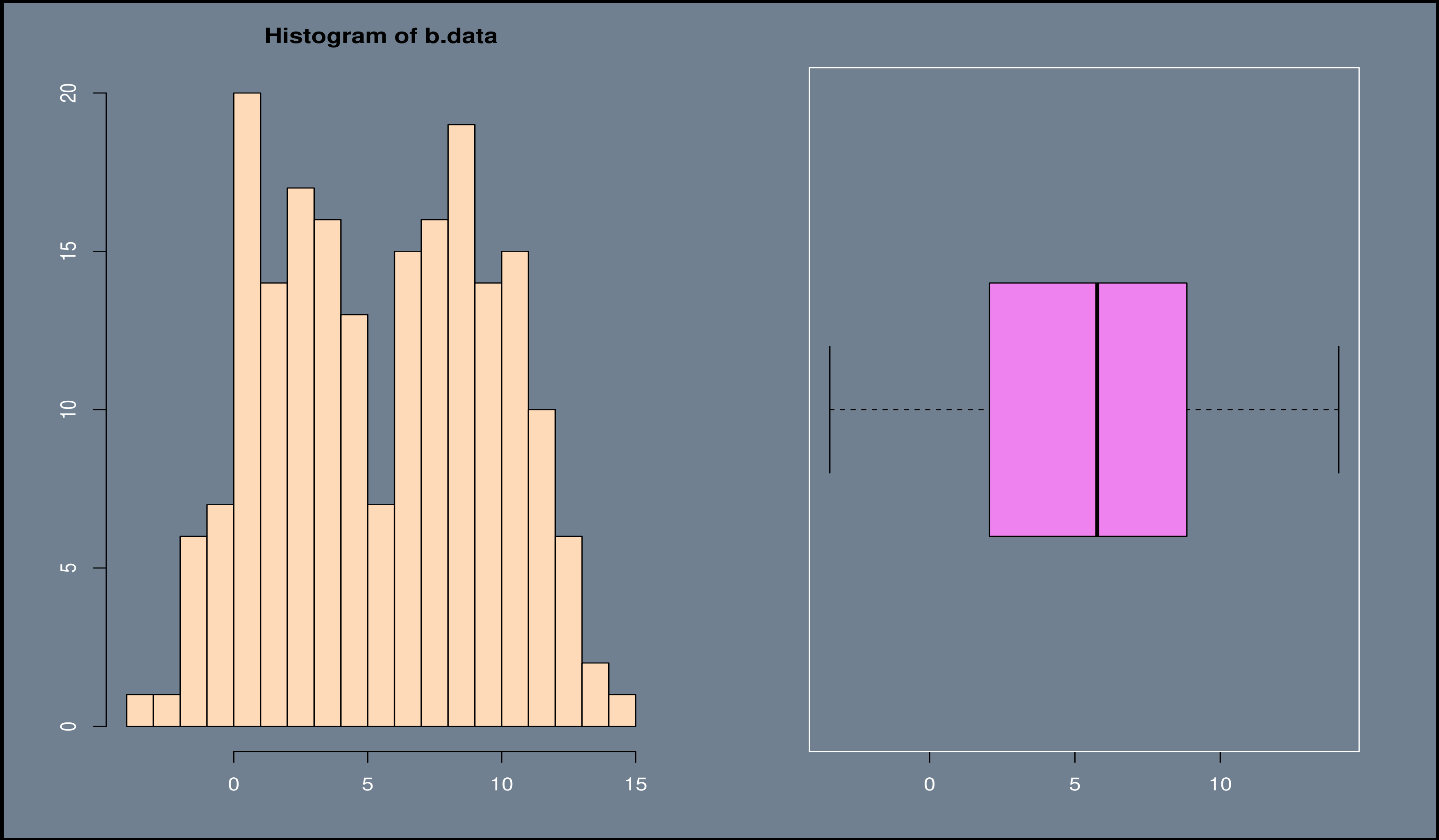


Warning #2: Interpret the Scale (Wainer 1990, mean distance 7 years)

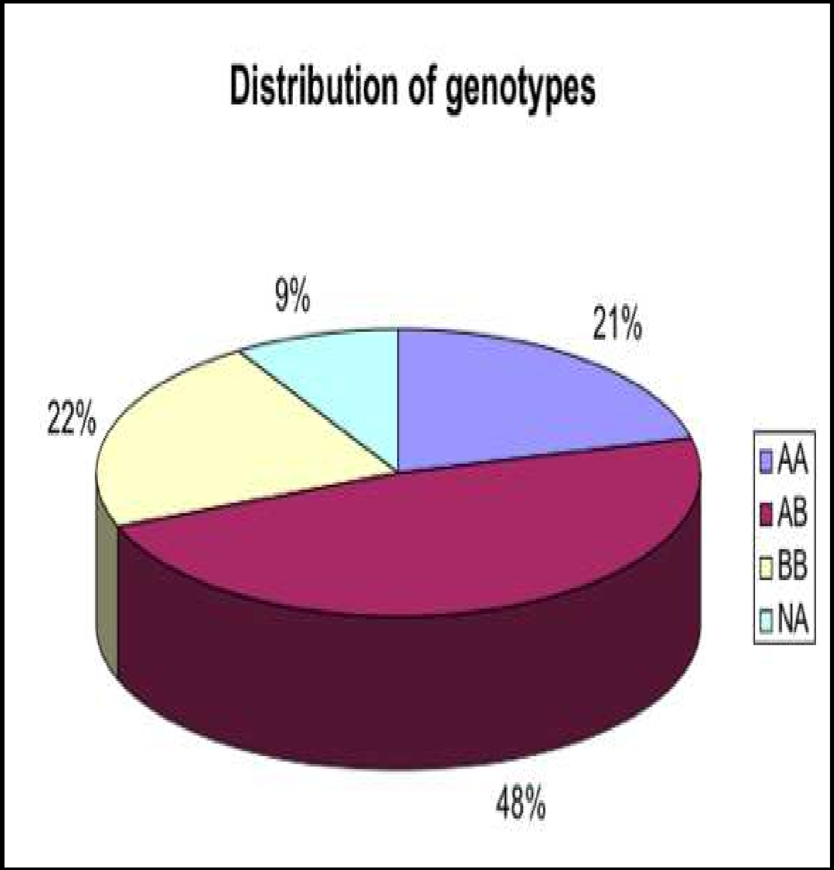


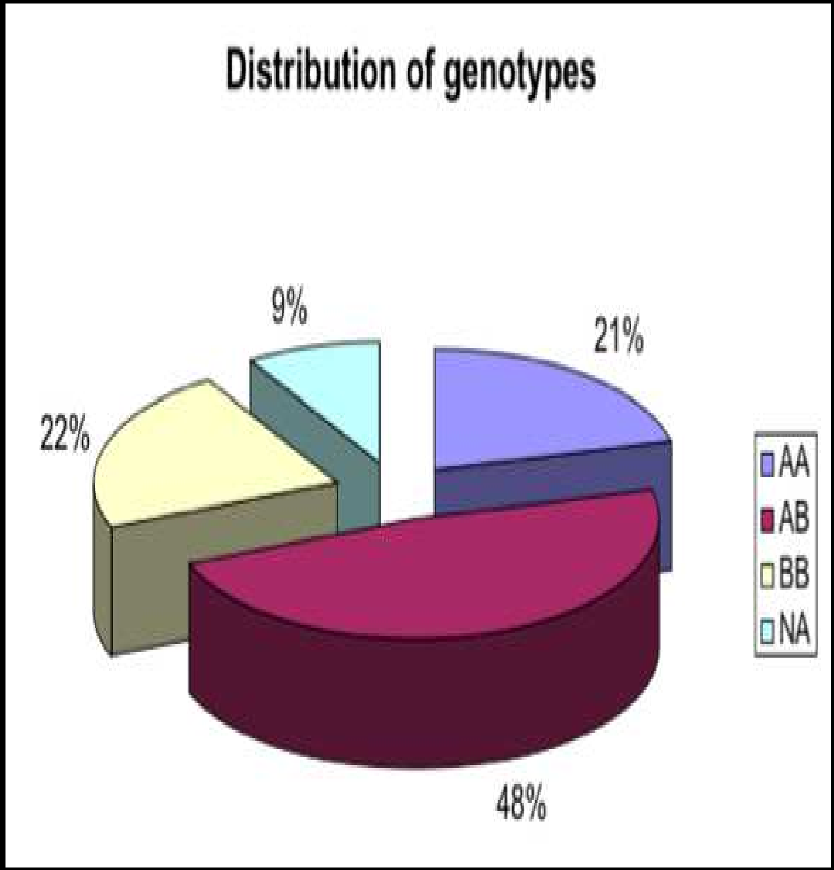


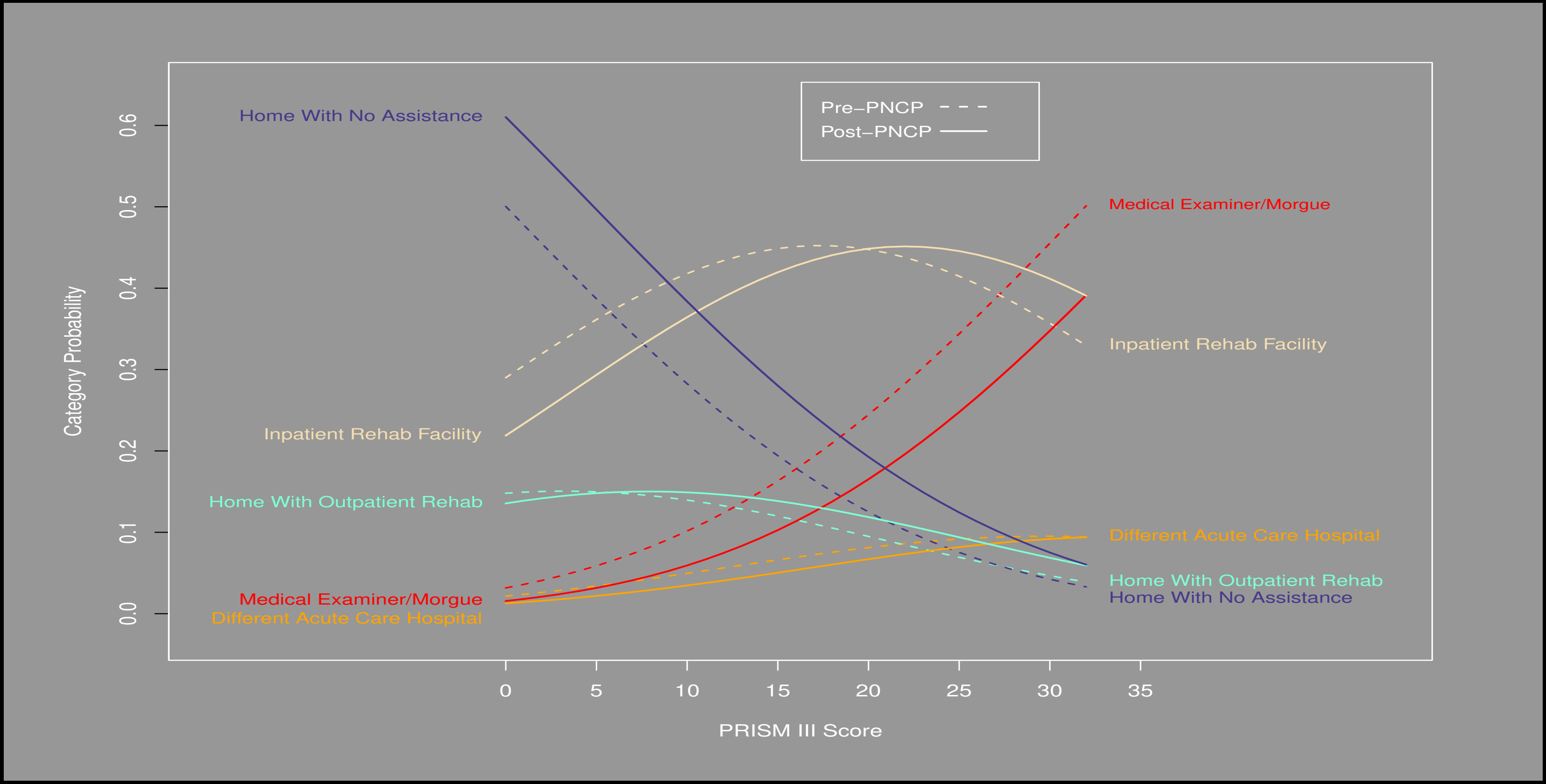
Warning #3: Boxplots Can Hide Information [whiskers=1.5(IQR)]



Warning #4: Piechart Mistakes

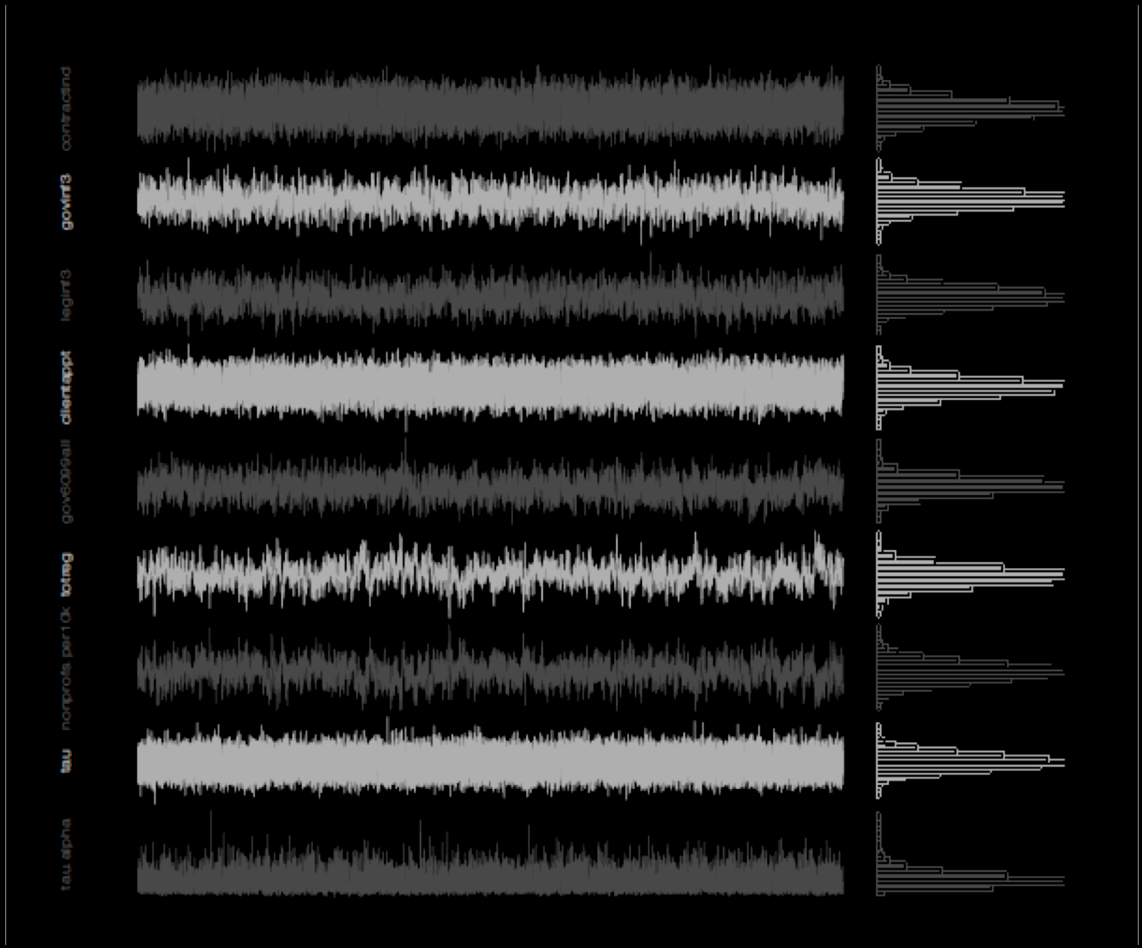




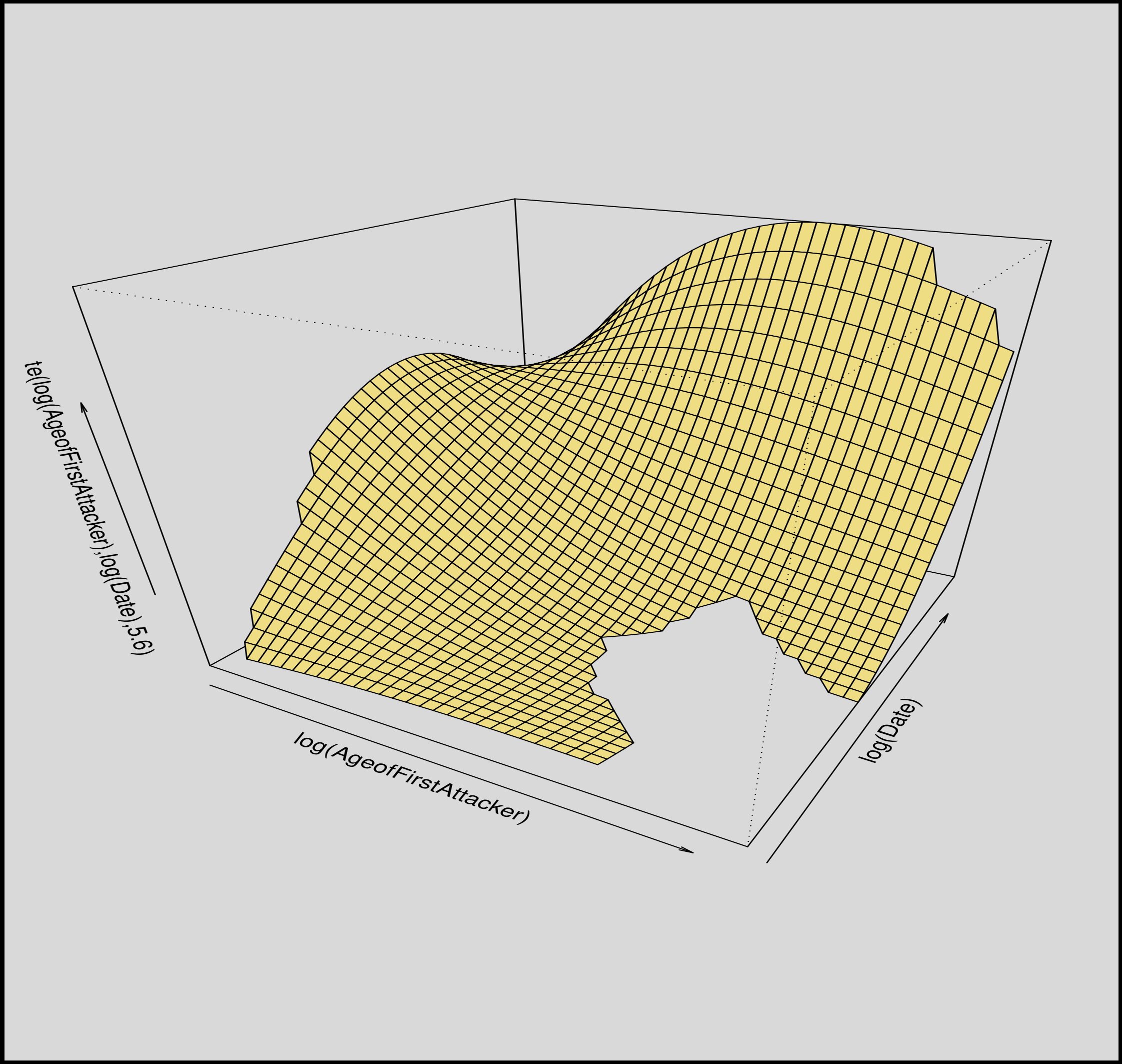


Output from Ordered Probit

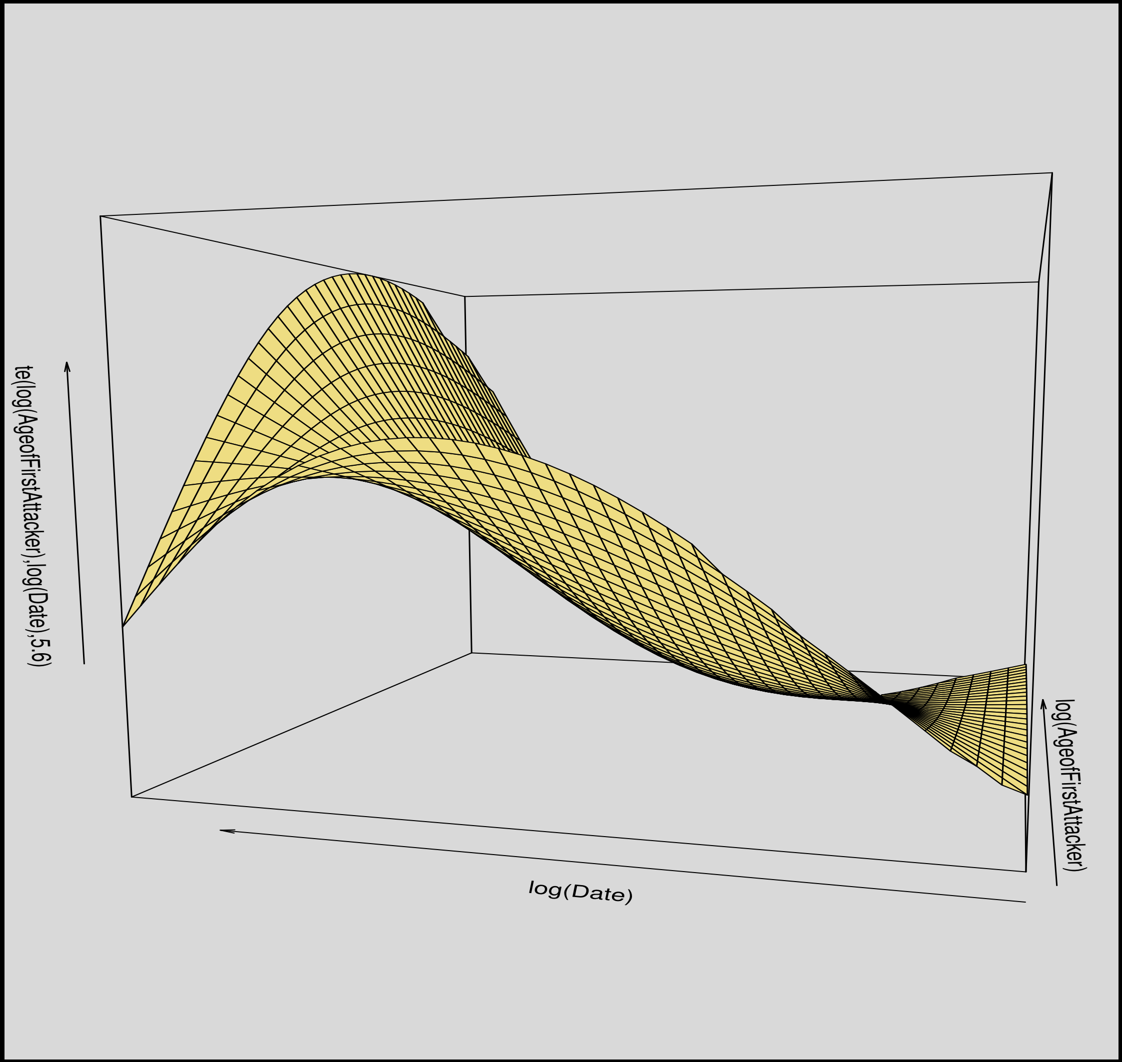
A Hierarchical Model of Lobbying Influence in the US States, Convergence Plots



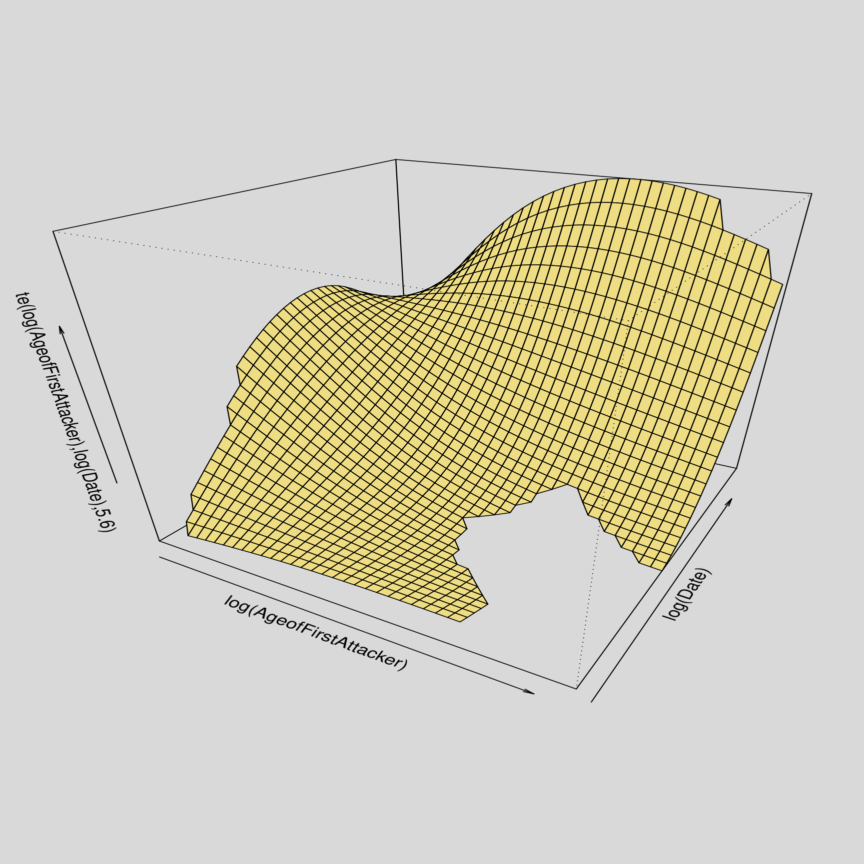
Viewing the Nonparametric Results

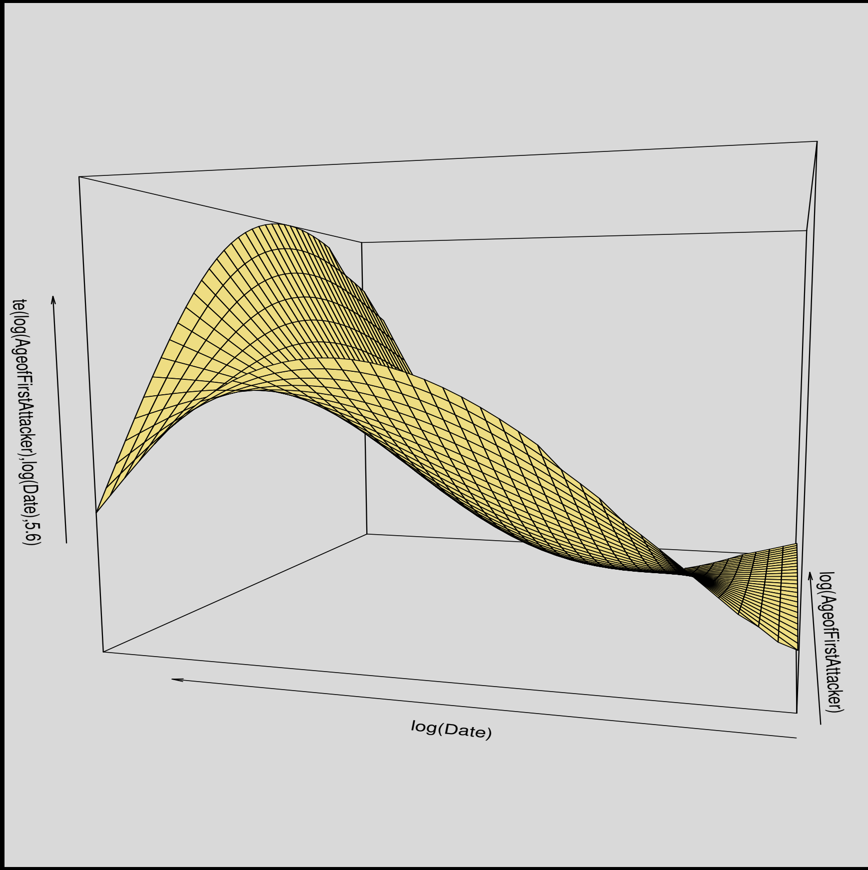


Viewing the Nonparametric Results

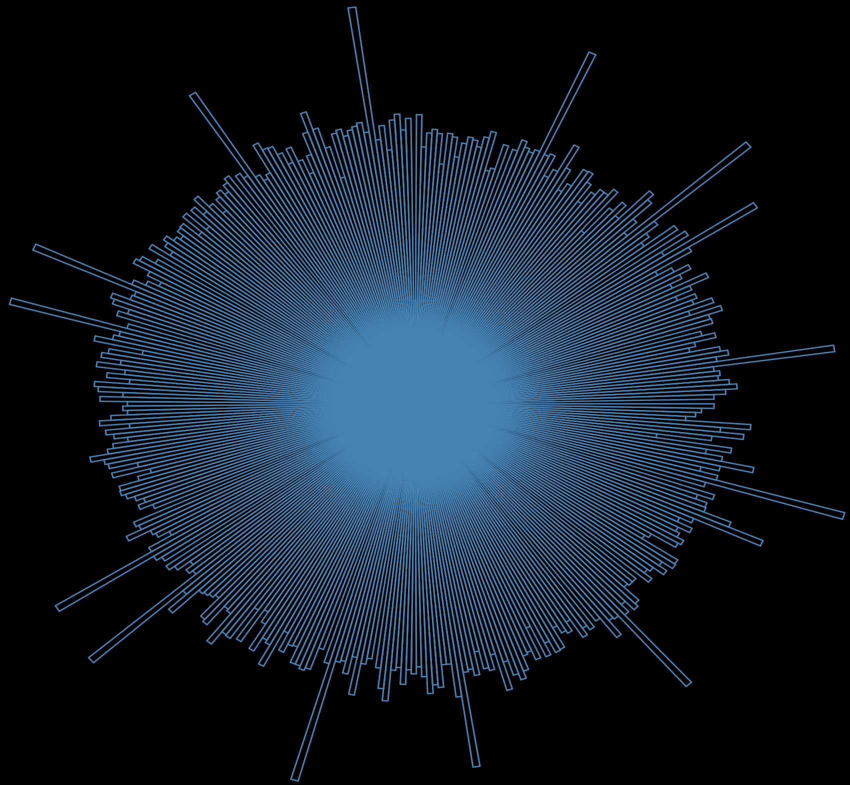


Graphing the Terrorism Data Analysis

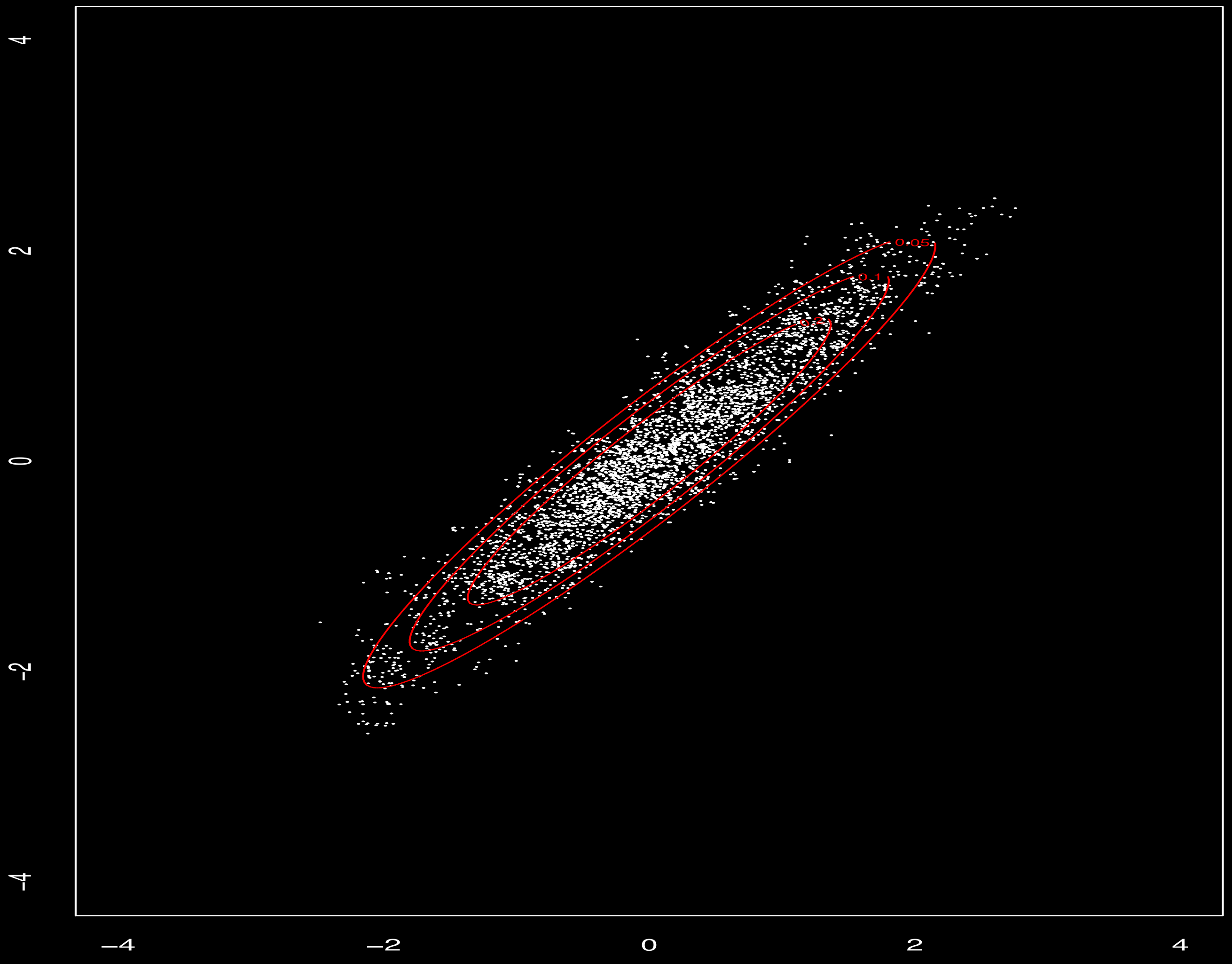




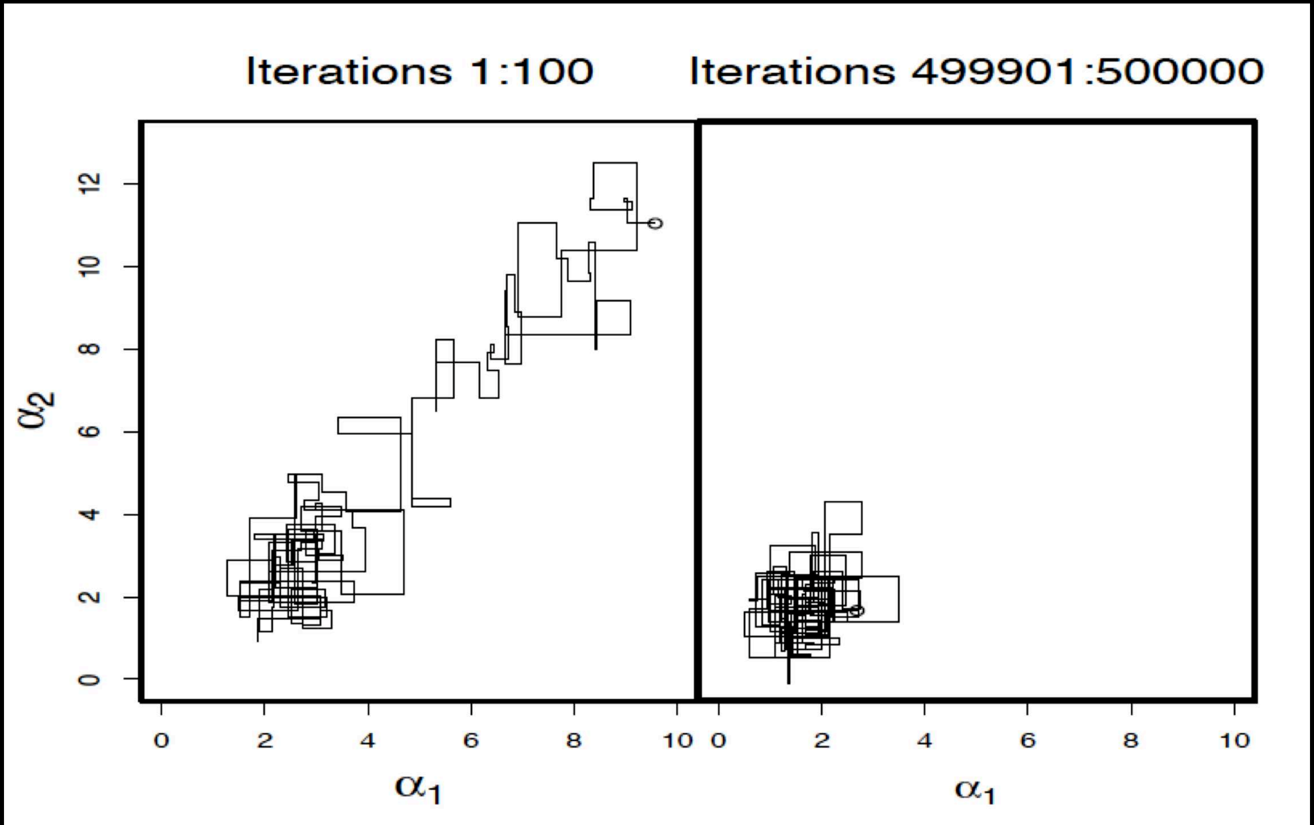
Rose Diagram: 1969-2004 Suicides in Switzerland by Numbered Date, $N = 51025$



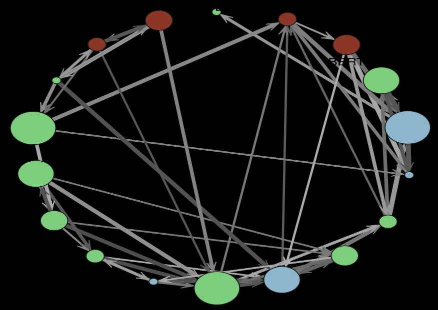
Hit-and-Run Markov Chain Example



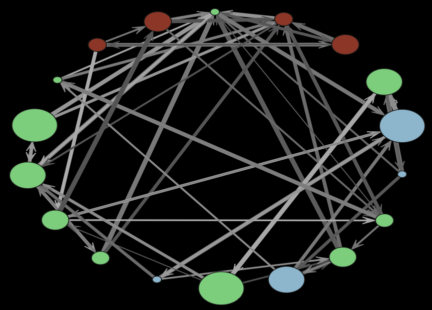
Gibbs Sampler Path



The Effect of Multiple Imputation on Monks Networking Data



Full Data



After Imputation