STAT 100 Jan 14, 2010

Have I learned any statistics yet?
In the list below, I mention some jargon words that have been discussed so far, and I also mention one place in the text or notes where they are mentioned.

Check out these **tools and concepts** met so far:
Observational study (p xviii)
Experiment (Randomized Experiment) (p xviii)
Random Sample (p xix)
Random Walk (Jan 5)
Averages and Proportions (Jan 5) [rwalk.run() again]
Variation in Data (Jan 5)
Histogram (Jan 7) [window.demo()]
p-Value (p 7)
Bar Chart (Jan 7)
Time Series (Jan 12)
Hypothesis Test (p 5)
Probability (p 8)
Probability Updating (p 300)
Quantitative vs Qualitative Measures (Jan 12)
Reliability (p 343)
Interaction * (Jan 14 and p 381)
Covariate * (Jan 14 and p 386)
* These are important concepts mentioned in the readings so far, but I need to elaborate on them today.

Of course, understanding has many levels, so you need to keep updating your understanding as you work with these jargon terms.

More "statistics" exposed so far:

At least as important as a familiarity with these tools and concepts is the ability to recognize their role in particular applications. It is difficult to list all the ones we have discussed, but the following examples may suggest what I mean:

There is an important connection between study design (experiment vs obs. Study) and establishing causality. (Moore Intro)
A small p-value suggests that a certain hypothesis used to compute it, is false. (p 13)
Data Analysis usually begins with plotting of the data. (Moore Intro)
Everyday data can contain deceiving trends (e.g. stock market index, sports leagues) – there are many opportunities for illusions of randomness. (Jan 5)
Models can be useful even when they are wrong. Note: Models are always "wrong" when they are describing the real world. (Jan 12 – Zipf)
Averaging and Smoothing Methods can help to extract a signal from "noisy" data. (Jan 12, Fuel)
Interaction in a predictive model requires detailed summary. (Jan 14)
Covariates can be used to improve the sensitivity of a predictive model. (Jan 14)

More about *Interaction* and *Covariates*:

Interaction: On p 381 is written:

"Note that we are discussing here a difference between differences – that is, and interaction effect."

The comment is in reference to Fig 2 on the same page. Note that the click through rate on Friday (11.8 for "Planning") compared to Tuesday (8.7 for "Planning") was a larger difference than for Festive (9.1 – 8.8) or Elegance (8.8-8.5).

This seems almost too complicated to have a name! But actually, all it is saying is that if you were to ignore the SUBJECT category and look at the overall click through rates differential for Day of Week (something like 9.9 – 8.7), that difference would not apply to each of the ignored categories. So when there is an interaction between SUBJECT and DAY-OF-THE-WEEK in the prediction of Click-through rate, you have to report it separately for each level of each variable.
Like this:

Difference in click-through rate due to DAY-OF-THE WEEK:

Planning: 3.1 (11.8-8.7)
Festive: 0.3 (9.1-8.8)
Elegance: 0.3 (8.8-8.5)

Or, report it this way:

Difference in click-through rate due to SUBJECT:

mid-week: 1.65*
Friday: 0.15*

* Don't worry about calculation details for this at this stage.

It would not be enough to say that DAY-OF-THE-WEEK effect is 1.2 (9.9-8.7) or that SUBJECT effect is 1.6 *. Look how wrong this is for Festive and Elegance, and for Friday.

The bottom line: When X and Y interact in their prediction of Z, the effect of X and Y on Z must be reported for joint (=combined) values of X and Y. If there is no interaction, the reporting is simpler with each variable effect reported separately.

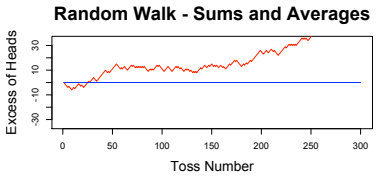
Covariates: On p 386, we see

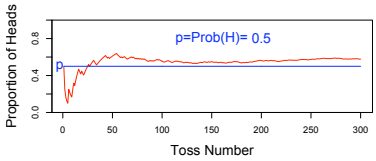
" But for the covariates, those variables that are observed but not controlled and randomized over, it is this …."
Again on p 389 "A covariate is a variable that is not included in the design of the experiment (i.e. is uncontrolled) but is measured for each subject."

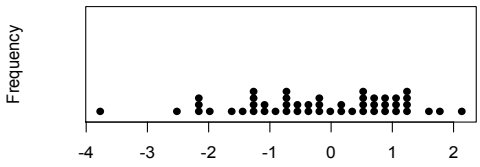
In the Kahn & Roseman study, covariates that were mentioned were AGE, TITLE, and REGISTRATION DATE. As the graphs on pp 383-385 show, these were clearly associated with the outcome variable, CLICK-THROUGH RATE. So to get a good look at the effects of the primary variables of interest, namely SUBJECT and DAY-OF-THE-WEEK, it is best if there is a preliminary adjustment for these relatively uninteresting "covariates".

The adjustment process uses "Logistic Regression" which we will not discuss in this course, but you should at least know what it does (as described above). (The term is mentioned on p 386.)

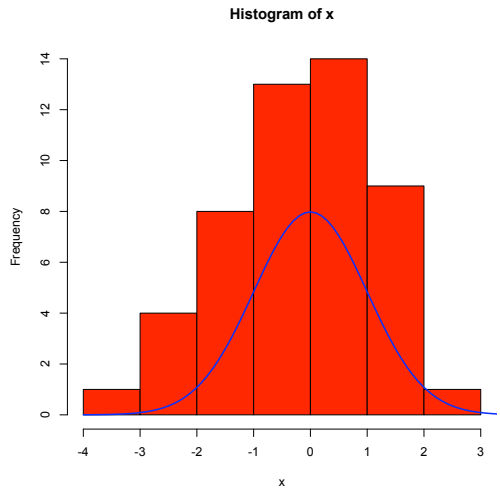
The bottom Line: You need to know what a covariate is, and how it is used.

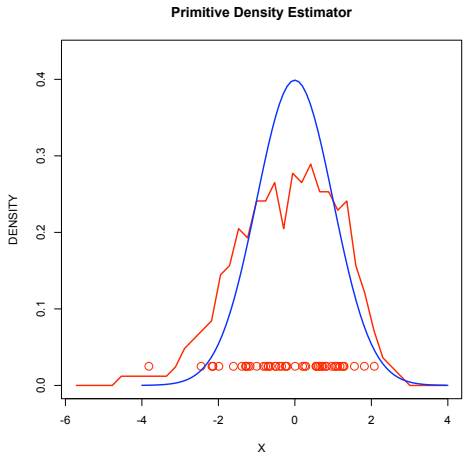


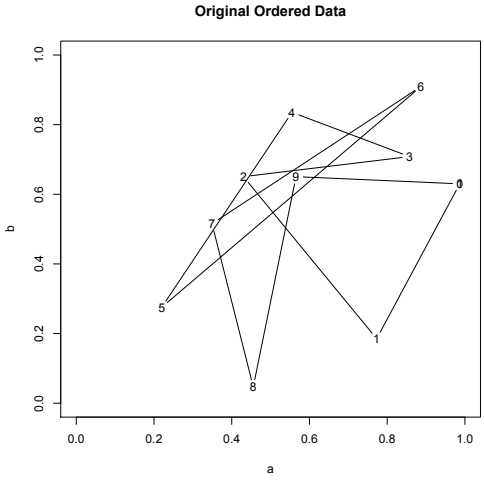


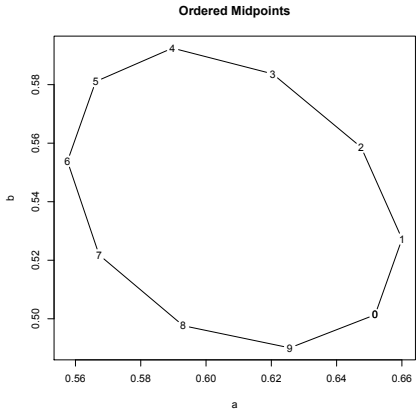


More graphs next page …









Miscellaneous Points:

Triangle Pattern is called Serpinski Graph.

Average of {0,1,1,0,0,1,0,0,1,0} is 0.4
Proportion of 1's is also 0.4
This illustrates what is now obvious: A proportion is an average.
(This will be a useful fact in a future lecture.)

Probability is Long Run Relative Frequency

A **random sample** of size n is a sample of n population items selected from the population in such a way that all possible samples of size n have the same chance of being selected.