STAT 100 Chance and Data Analysis Sept. 27, 2002(rev)

Repeat of Notices:

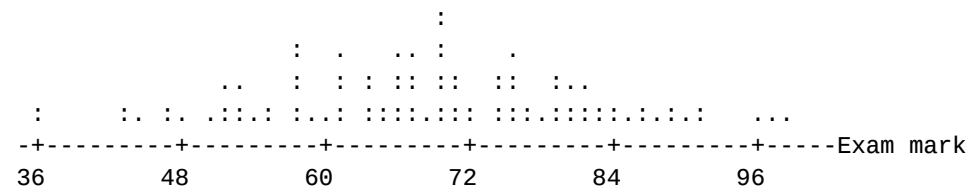
1. Office Hours Change MF1430-1520 W1130-1220
2. Midterm Dates Oct 11 and Nov 8.

Preliminary demo: Showed how a symmetric (Prob = .5 of +1) random walk, which by construction has no tendency to increase or to decrease, would appear to have trends up and down that appear to be persistent. That is, by looking at a portion of the random walk, it often looks as if a trend up (or down) is established, and so one might be tempted to predict a continuation of such a trend. But we know that the random walk is as likely to go up next step as go down, and that there is no benefit to extrapolating past trends in an attempt to predict the future. The relevance of this to the stock market is that, when one sees similar trends in stock prices, it is possible, in spite of appearances to the contrary, that such trends are not persistent, and if so are not useful for predicting the future stock prices.

Today's topic: Study Design: Test Score Article (pp 178-187)
Balancing unknown or uncontrollable biases – experimental designs.
Calibration, Percentiles.
Curving Results, Adjusting Test Scores

Educational Testing Service: often compulsory "GRE" exams for graduate school admission in US and elsewhere. Report percentile compared to reference population.

The **rth percentile** of a list of scores is the score that just exceeds r percent of the scores. For example:
The 25th percentile is usually called the first quartile. The 50th percentile is the median. If scores on and exam with 100 students looked like this:.



What would the 90th percentile be? About 86. Note that there is some ambiguity in the definition – making it exact is complicated, but the idea is simple, and exactness is unnecessary for most purposes.

The GRE would likely be in high 90s percentiles for entry to the best US grad schools.

Another kind of test administered by the Educational Testing Service is the AP test. These tests involve both multiple choice questions and free-response questions. The free-response questions involve problem solving or essay-writing.

Thousands of students write these exams, so there has to be an army of readers (markers). Moreover, the time period for marking may be several days for a given reader. This creates a problem for fairness of marking. Adjustment of scores is needed to account for different styles of marking and different trends over time. The adjustment process is called **calibration**.

But how do you discover the reader and day biases? Need to have one essay read by more than one reader, and on more than one day

The article discusses an example with 12 readers. They wanted to use a small number of essays (32 were used) to judge the "reader effect" on the marks. The problem was to have the 12 readers grade the same 32 tests, but somehow balance out the effect of the order in which the papers were read. This is a "design of experiments" problem. It is not obvious how to do this but the design shown in the article has the following characteristics:

Each reader reads each paper once only
Each reader reads eight papers each day
Each paper is read by three readers each day

A design that does this is shown in the article. But not obvious how to get it!

The aim was to try to detect the effect of reader stringency (how hard they marked) and the effect of day (e.g. whether they got more generous over the four days of reading). If these effects could be determined, then all the papers marked subsequently (apparently tens of thousands) could be adjusted to make the reported marks more fair to the students.

For example, if reader A gave 10 more points to students than was the average for the 32 papers (average for all 12 readers), then a fair correction would be to subtract 10 points from the papers that A has read, and in fact it would make sense to subtract 10 marks from all the paper that A WILL read from the many more in the same exam.

The small scale experiment has given us information that is useful for improving the fairness of marking in this large scale marking context.

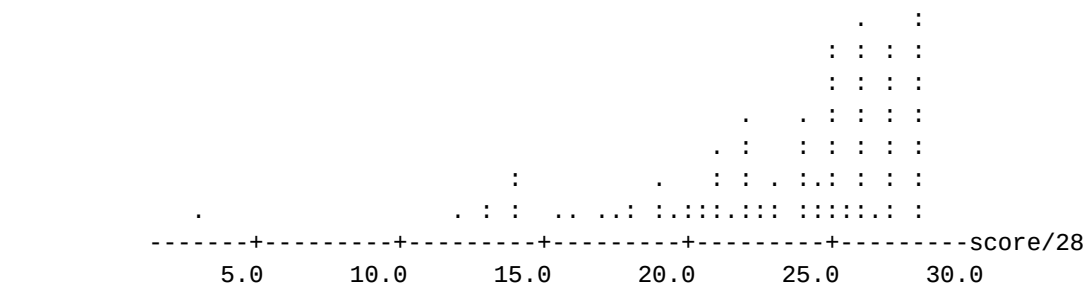
The article introduces the need for calibration by mentioning the practice of "**curving the marks**". What is that?

At SFU and many other universities, marks are assigned a letter grade – this allows instructors to maintain stable standards even when the exam difficulty varies from course to course. One method is to assign a certain percentage of grades to A, another percentage to B, etc. (and various splits to A-, A, A+ …). For example, I use 20% A, 30% B, and 50% C,D,F as a rough guide.

What this is doing in a crude way is to use the percentile of scores to determine the letter grade. Another way of saying the same thing is the letter grade is based on the rank of the papers in the group of students that are tested. Could you argue for and against this practice?

Caution: A percentile is not a percent - it is the score that corresponds to a certain percent.

Assignment #1 Marks:



50th percentile is 25 (/28).

25 is B-

80th percentile is 27

27 is A-

Readings for next two lectures:
Salk Poliomelitis Vaccine Trial pp3-14
Health Insurance Experiment pp 31-40

Preparing for Mid-term I Oct 11.

Look back over your (and my) notes. Prepare questions for concepts that seem fuzzy, or when you don't know what the point of the discussion is. Take them to the Stat Workshop (K 9510) - if that fails, bring to my office hours.

Feedback: Find a piece of paper: please tell me your opinion …

1. Most important concept since Sept 16.

2. Most confusing concept since Sept 16.
3. The topic you would like to hear more about.
4. Anything else you would like me to know.