STAT 100 Chance and Data Analysis October 16, 2002 (Rev)

Today: Feedback from Midterm I

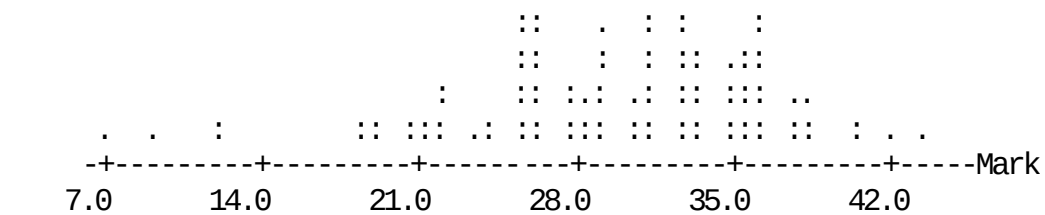
More Models

Zipf's Law (Tanur pp 142-150)

Binomial Model (Tanur pp 126-131)

Normal Distribution (any stats text)

Feedback from Midterm:



mean was 29.2 (65%)
SD was 6.7 (15%)
Q1,Q2,Q3 25,30,34 (56%,67%,76%)

Note that the interval (Q1,Q3) will contain 50% of the marks.

What percent does mean ± SD contain? Can't say in general but a good guess is 68%.

In this case mean ± SD is 29.2 ± 6.7 or (22.5,35.9) and the percentage is (based on the actual data) = (69/103) * 100 or about 67%. Good Guess!

How about mean ± 2 SD? 29.2 ± 2*6.7 = 29.2±13.4 or (15.8, 42.6).

The good guess in this case is 95%. From the actual data, the percentage is (98/103)*100 = 95%. Another good guess!

Many real data distributions have the following property:

68% of the observations are within 1 SD of the mean
95% of the observations are within 2 SD of the mean
100% of the observations are within 3 SD of the mean

(These rough estimates are based on a family of normal distributions – more later.)

These percentages could have been guessed BEFORE the midterm was written!

Zipf's Law (Tanur 142-150):

Not a mathematical "law" – more like an empirical observation – something that tends to happen.

Size times Rank = constant.

(So Log(size) + Log(Rank) = Log(constant) = constant. Equation of a line like the one in Fig 1 p 143).

Zipf for words …Explanation in terms of a model: suppose next word's probability proportional to past occurrences of words – this can be shown to imply Zipf.

Zipf for cities … Growth proportional to size.

Zipf for firms – growth proportional to size of acquiring firm.

Zipf for research papers – research funding proportional to papers published, so output proportional to past publications.

Main point of article: a model does not have to be exactly true to be useful, and the best models are the most useful ones, not the most exact ones.

Binomial Model (Children's Recall of Pictorial Information, pp126-131)

Idea pf article: Young children will vary widely with respect to the ability to recall pictorial information, and this skill is not well assessed by traditional tests of reading and math. This skill may be useful as a learning agent (if educators knew which childen had it).

The method of testing this skill required analysis of data via use of the binomial model. (Model for probabilities in a special situation). This will be discussed in class. More detail in revised notes ..

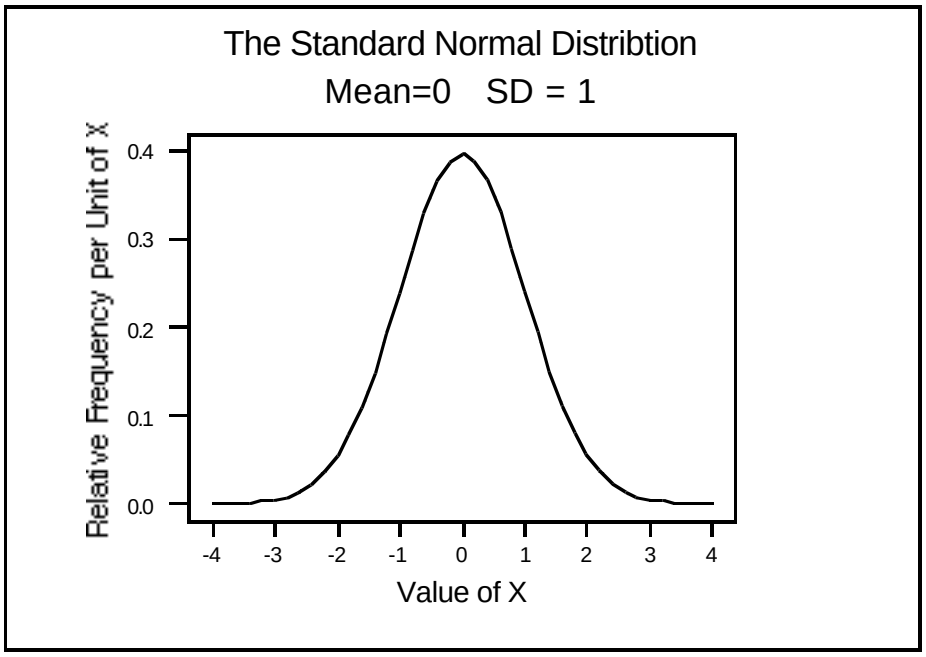
How many heads in 10 tosses of a fair coin? See Table 1.
Two ways to get it. 1. Simulation 2. Calculation with Binomial Model
25% chance to get 5 H and 5 T, etc.
How many heads do you have to get to convince you coin is biased?
See Table 1 to decide what is extreme when coin fair.
Recall: When something happens that is rare under ordinary circumstances, this is evidence the cirmcumstances are not ordinary. This is the logic of statistical testing of an hypothesis.

40 slides seen, paired in random order with 40 unseen slides.
Many children identified 31 or more correctly.

No assignment for Oct 23. Big one for Oct 30. Meanwhile, read the Tanur articles as outlined in the Course Outline for week 7.
And make sure you understand why you did not get a perfect score on the MT. The ideas will come up again. And again!

Normal Distribution:

Shape of distribution of averages.
mean±1SD 68%
mean±2SD 95%
mean±3SD 99.7%
exact for normal distribution (a good model for many data sets).



Example: I.Q.s are Normal with mean=100 SD=15
Your IQ is 130. How high is that relative to population?
130 is 2SDs above mean, so only 2.5% are higher.
130 is in the 97th percentile.