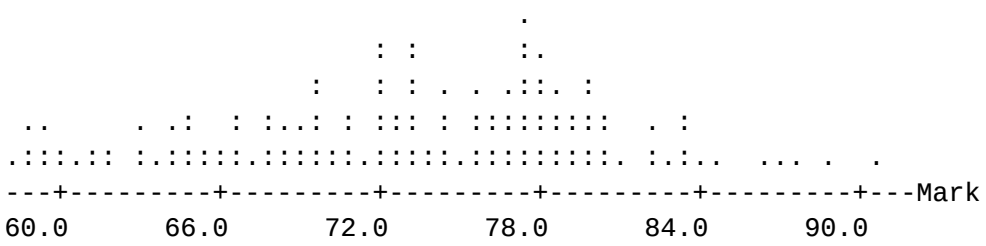
STAT 100 Chance and Data Analysis October 18, 2002

Today: More Content Feedback from Midterm
Binomial Model (Tanur pp 126-131)
Normal Model (Intro Oct 16 but more today)
Sampling (an introduction - see "Accounts" article in Tanur pp 151-160).

Next Week: More about Sampling
Sampling Surveys (Jurors article in Tanur pp 87-92. CPI article Tanur pp 198-207). Census article Tanur pp 208-217. Randomized Response Technique (no ref). Please give these articles an initial read. (No other assignment this week).

More Content Feedback on the Midterm (from the Markers):

Q1. (Mean, SD, and 10th percentile of marks distribution)



Mean – use balance point – 73.0 was mean (71-75 got full marks)
SD – calculation with 150 points not an option! – 7.5 was correct (5.5.-9.5 got full marks)
10th percentile – 150 points so just count 10% of the way through – in other words the 15th one from the bottom. Turns out to be 63.0
Mean & 10th percentile done well, SD not done well.

Q2. (random walk and stock market)
Two points – lack of predictable drift
exhibits apparent (but not persistent) drift

Done reasonably well.

Q3. (insurance – variability of averages)
(still waiting to hear …)

Q4. (reference to Tanur article contexts)
Those who had read the articles earned some easy marks. Generous marking.

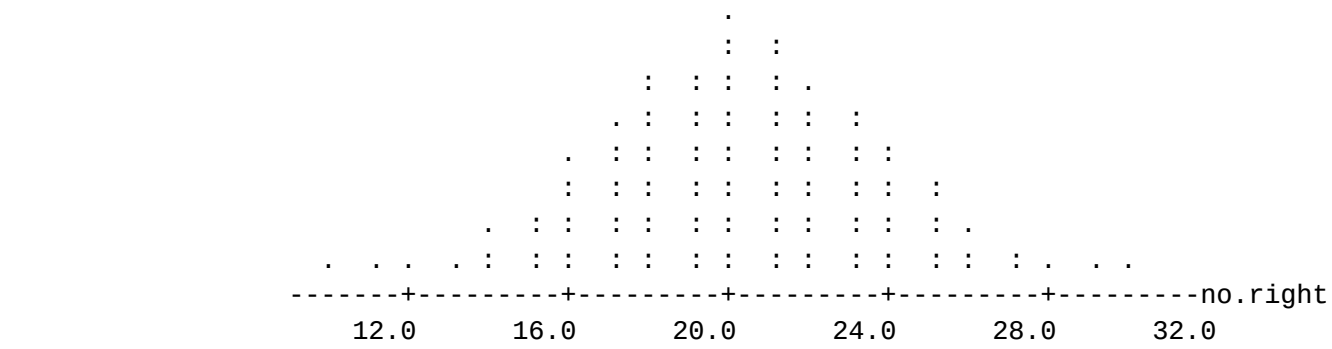
Binomial Model (Children's Recall of Pictorial Information):

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 children had it).

The method is to show the child 40 landscape pictures, for a short time each picture. The investigator collects 40 other pictures that the child has not seen. Then the child is Presented with pairs of pictures, one from each list, with the order of the two pictures is randomized (toss of a fair coin say), and asked which picture has been seen before. The child is given this test with each of the 40 pairs of pictures, and so has anywhere from 0 to 40 correct. The investigator wants to know if the child has any real recollection of the pictures previously viewed (or is the child just guessing).

The method of testing this skill required analysis of data via use of the binomial model. (Model for probabilities in a special situation). We could simulate this task assuming that the child had no recollection, for in this case the child would have a 50% chance of guessing the correct one of a pair. By tossing a fir coin 40 times, and repeating this whole process many times, we could see what kind of "success" a child would have under this "no-recollection" assumption. Here is a MINITAB simulation of 100 repetitions of this simulation.

```
Each dot represents 9 points
```



no.right	Count				
10	2		20	130
11	2		21	122
12	2		22	94
13	7		23	86
14	21		24	70
15	29		25	47
16	56		26	23
17	79		27	14
18	100		28	8
19	105		29	1
			30	2

(Note the limitation of dotplots to display distributions with large total frequencies).

Note the conditions under which the binomial distribution applies:

1. a number of trials is performed in which one of two outcomes occurs
   (e.g. 40 trials to get "right" or "wrong")
2. the outcomes of one trial does not affect the outcome of any other trial
   (e.g. the performance on one pair of pics does not affect the perfomance on another pair – it would be important that the child receive no result until the end of the 40 pairs.)
3. the chances of the two outcomes is constant over all the trials
   (e.g. Under our assumption of no recollection, we assumed chance of right =1/2)

For every n=number of trials and p=probability of outcome 1, it is possible to compute the chances for each possible number of "right" outcomes. Here is what MINITAB gives for this particular case (n=40, p=0.5):
Binomial with n = 40 and p = 0.500000. X is the number of "right" outcomes.

x	P(X = x)
7	0.0000
8	0.0001
9	0.0002
10	0.0008
11	0.0021
12	0.0051
13	0.0109
14	0.0211
15	0.0366
16	0.0572
17	0.0807
18	0.1031
19	0.1194
20	0.1254
21	0.1194
22	0.1031
23	0.0807
24	0.0572
25	0.0366
26	0.0211
27	0.0109
28	0.0051
29	0.0021
30	0.0008
31	0.0002
32	0.0001
33	0.0000

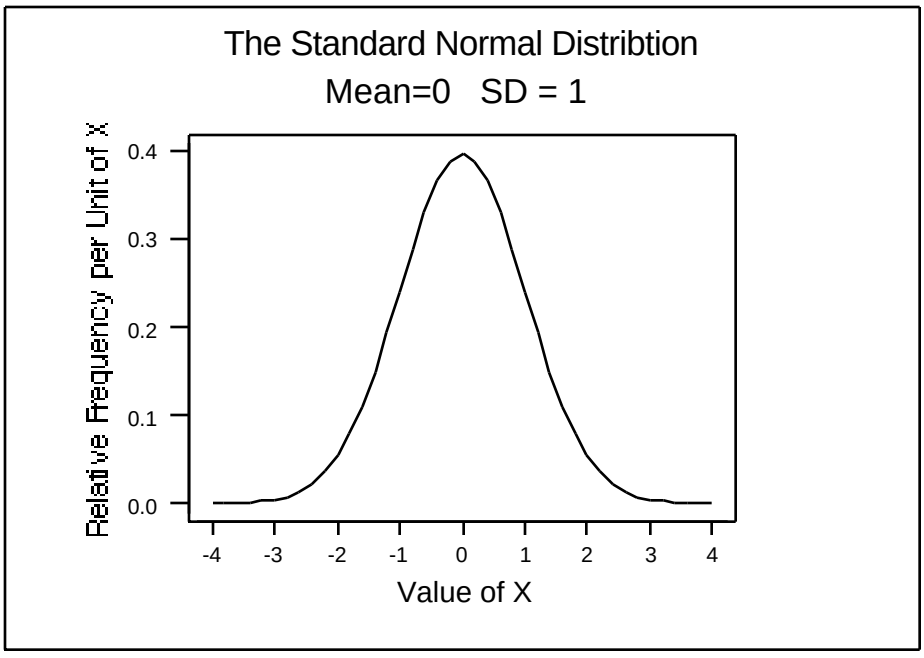
What this means is that, for example, the probability of 20 right and 20 wrong in 40 trials is .1254. That is 12.54 % of the time this experiment is done, the child would get exactly 20 right – under the assumption of no-recollection.

Note that these probabilities add to 1. They can be thought of as the long run relative frequencies of the outcomes shown. One can see from either the simulation or the binomial formula that outcomes of more than 30 right would be rare if the child had no recollection. Does this imply, if the child has 33 right, that the child does actually recall some of the images?

"If something happens that is unusual under ordinary circumstances, then this is evidence that the circumstances are not ordinary".

Note also that of we multiply the theoretical probabilities by 1000, we get frequencies similar to the ones simulated. Since we simulated the results for 40 pairs of pics, 1000 times, and since the probabilities are really the long run proportions of times that the various outcomes occur, this similarity should not be surprising. (If it is, get help to understand it!).

Next: More about Normal Model: Recall we had



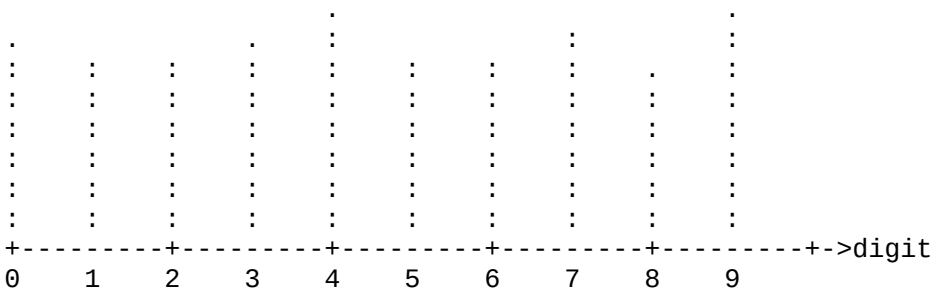
Normal Distribution:

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.

The main reason the Normal Distribution is important is that averages tend to have Normal distributions. The tendency is stronger as the number of things averaged is greater. But the number does not have to be very large. Let's do the dollar bill exercise again, where we are averaging 7 digits.

Note that the relative frequency of each digit is approx equal.
Distribution might look like:

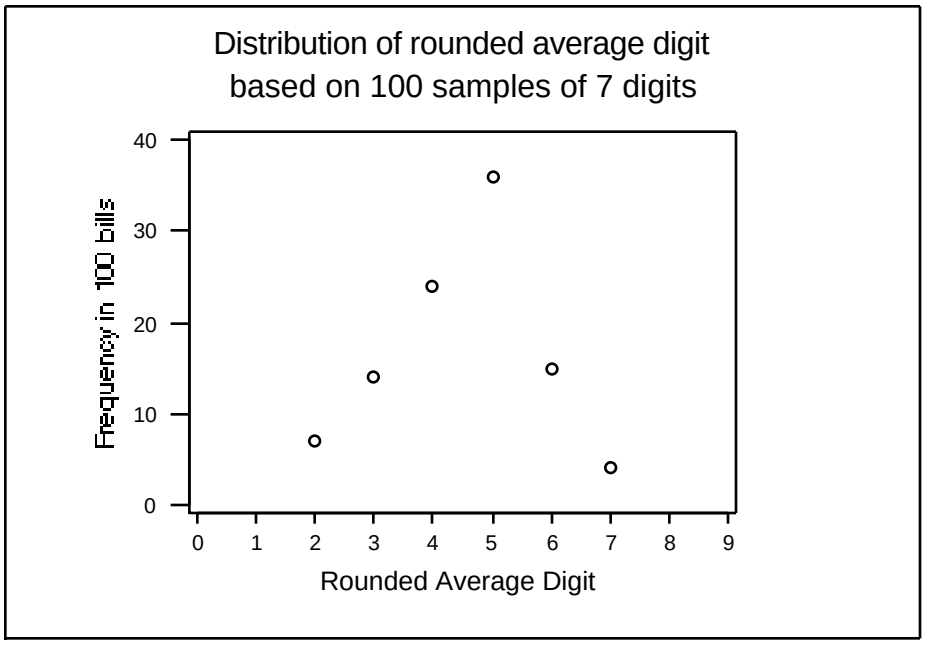


If we select 7 digits from such a distribution, at random, the average could be anything from 0 to 9, but we know it will usually be close to 4.5 (square root law of variability of averages).

What does the distribution look like?

Here is a simulation of what might happen in a class of 100 students (with one bill, seven digit serial number, and one average digit, from each student). Suppose we round the average to nearest digit ..

In the above simulation we get



Without the rounding, we would have had:

```
                                    :
                  .       .       .   :
                  :       :  .:   : : : .
        .         :.  .   :  ::   : ::: :   .  .::
     .. ::  . .:  ::  : ::. :: :: ::: :: :. :::    .. ..
  ---+---------+---------+---------+---------+---------+---------+---avg. digit
     2.0       3.0       4.0       5.0       6.0       7.0
```

How does this compare with the normal distribution?

Try mean ± 1SD. mean 4.55   SD 1.13 so mean ± 1SD is 3.42 to 5.68

It turns out that 66/100 or 66% of the means are in this range.

This simulation seems to be confirming that averages are normal!

(We need more tests but I can assure you they would be satisfied too).

Keep in mind for Binomial and Normal Models:

Binomial: proportion of outcomes of n trials that are of a type
Normal: describes proportions of outcomes for averages and totals (average = total/n)

A first look at “Sampling”:

Group sampled = population
Selected subset = sample
Number selected is called the “sample size”
Method of selection = random sampling (usually, for this course)
Method of summary: dotplots, or means and SDs usually.

Example: Population of digits in which each digit is equally represented, sample of size 7, summarize by average digit and SD of digits, or dotplot (as we did).

A first article to read for an introduction to Sampling is the one about Accounts pp 151-160.