last time, we introduced the STANDARD DEVIATION
a measure of how whacko particular points are.

it relativizes particular samples wrt the amount of variability in the whole sample

-> you can use that to compare students on different kinds of tests, as in your HW

Now, lets take control over the origin of these "scores"
using R's rbinom function we can generate simulated data subject to our specifications

one specification is that the raindrop fall EITHER on the left stone, stone zero
or on the right stone, stone one.

If we just want to know how many drops hit a particular stone, then we don't care
about the order. The two samples

    LLLRL
    and
    RLLLL

are identical for our purposes. They should both count as an observation of 1/5 right-stone hits.

Equation 2.7, the definition of the binomial, gives this exactly

VARIABILITY IN SAMPLES

chunk #11 shows us that as we observe more and more 40-raindrop sequences,
the number of right-stone hits tends more sharply cluster around 20 - i.e. exactly half, exactly prob=0.5

visually, the mode is 20. and it gets sharper as we have more data.

    [show proportionStones.R]

chunks 12 and 13 show that as the sample size increases the stddev is going up in
absolute terms. That is, there is more uncertainty in a bigger sample.
    [ this is what I alluded to last week with degrees of freedom ]

chunk 15: if we divide by the sample size we see that in fact a bigger sample size actually lowers SD.
in other words, a bigger sample is helping us home-in more accurately on the proportion p that is generating all this
(simulated) data. [remember, you are gods]

CONFIDENCE INTERVALS

2.4.3: as the sample gets bigger, the odds of sampling "exactly" your desired proportion from the binomial goes down
radically.

    [ plot the probs from code chunk 21 ]

the question is, how big does the interval surrounding the mean have to be
such that we have a 95% probability of a sample containing the mean?

his point with the simulations in section 2.6 is that with samples of size 40,
sample of size 40                          95% of sample means are within 6 of the true mean, p
sample of size 400                       95% of sample means are within 20 of the true mean, p
since 6/40 is 0.15 and 20/400 is 0.05, relatively speaking we have a much smaller margin of error with the larger sample if we want to be 95% confident we have guessed the parameter p based on the observed mean of the sample

Q-Q and CHECKING THE MODEL

n<-1000
p<- mean(havelaar$Frequency / n)
qnts <- seq(0.005,0.995,by=0.01)
# what we'd expect
expected <- qbinom(qnts,n,p)

#quantile is the inverse of the ECDF
actual <- quantile(havelaar$Frequency,qnts)

plot(expected,actual,xlab=paste("quantiles of (",n","round(p,4),")-binomial",sep=""),ylab="frequencies")

this suggests the appearance of "het" in this Dutch novel is not well-modeled by a binomial
this isn't surprising since we don't think of stretches of text as unordered bags of words.

try the Lewis Carroll exercise at the end of chapter 3 of Baayen