9th Class **6/18/10**

"When it is not in our power to follow what is true, we ought to follow what is most probable."--Descartes

show results from M&M experiment
[overhead]
sampling distribution of P in attempting to find π, the true proportion of blue M&Ms
(24% regular, 23% of peanut, 20% of almond, 20% of peanut butter, 17% of crispy)
[http://us.mms.com/us/about/products/--not on here anymore!]
note this has risen over time [show comparison of spring 2003 results to now]
(then blue % was 10% regular, 20% peanut)

and also to give us a confidence interval around π--in this case of 40 samples, can drop the two outlier values to get
a 95% confidence interval
median: 25; mean: 25; 95% CI: 16-37; 90% CI: 18-33
[mark on overheads; also π]

note that n fluctuated since the samples were based on min. weight instead of n
[mean of 56, range of 50 to 60 per packet, range of 6 to 24 blue ones per packet;
if I pool all the data and treat as one large sample I get 556B/2228= 25% blue]
[range has gone down; in 2003 sample it was 35 to 63 per packet with same mean]

review ideas from last class re Ch. 7: concepts of population and sample. A random sample is a random subset of the population, in which relative frequencies $\frac{f}{n}$ are used to compute $\overline{X}$ and s^2. These random variables are examples of <u>statistics</u>, or <u>estimators</u>. In a population, probabilities p(x) are used to compute μ and σ^2. These fixed constants are examples of <u>parameters</u>, or <u>targets</u>.

criteria to consider about estimators:
bias? in any size sample or only if n large, in which case asymptotically unbiased
precision? (low variance)
accuracy? (low MSE, which is a linear combination of bias and variance)

consistent? so as n increases, MSE tends to zero (so must be asymptotically unbiased to be consistent)

[handout/piece on website on pedometers and another piece on the website contrasting to one of the ones showed to be inaccurate in the handout]

spend some time mentioning some uses of probability and statistics

application of probability:
--game theory: literally is the theory of games—any game can be described by a game tree
--sometimes a branch indicates another card (or domino)
--but sometimes each branch indicates a move by one or another player
can include moving simultaneously--e.g., rock, paper, scissors, so you treat it as if it were random

it is interesting to analyze the structure of games--obvious application to gambling
[overhead on gambling]
e.g. poker
[show overhead with probability of different outcomes for a given hand]
note chance of getting a royal flush in 2 consecutive hands of poker is somewhat better than your chance of being killed in a plane crash
as a social scientist, we might want to move beyond simply working out these probabilities to ask questions about the structure of the game itself:
--how did people know to work these out before the techniques we use were invented? Did they derive them on the basis of observed frequencies in many games, or were primitive counting techniques sufficient?
--why is five-card poker so popular? consider why the number of cards is chosen four-card poker less interesting (no full houses, for one thing) and six-card and seven-card poker are more colorful, but lead to junk beating one-pair hands (also two-pair hands in seven-card poker)--in other words becoming less likely.

for odds and gaming strategies for a whole bunch of games see:
http://wizardofodds.com/
e.g. here you can also see the hand probabilities for five-suite poker

--distinguish between deduction and induction [overhead]

--we already saw deduction at work above in analyzing a game structure (and obviously there is a monetary aspect to games), but there are also less "frivolous" uses of probability, or deduction, like detection of noncompetitive behavior and "cheating" (e.g., discrimination, monopolistic price-setting):
--bias in jury selection [overhead on jury selection] -- note reference to poker
[also show link to recent article re Alito]

--bridge between deduction and induction: monte carlo example of evaluating test methods for detecting banking system race discrimination

--use deduction to tell us what the outcomes might be in the simulation setting, then use the techniques to aid in induction in the real world

The paper contrasts two techniques, one which costs less but does not have full information available to the loan granters in the credit files
1) generate a pool of loan applicants to simulate the actual population of both nonminorities and minorities in terms of income, net worth, debt payments, and credit history
2) these generated loan applicants then "apply" for loans in a credit approval model that is representative of actual approval processes used by financial institutions
3) the credit approval model allows for the possibility of bias against minorities and the bias parameter can be adjusted from 0 to 1

(allows for a number of history "blemishes" to be forgiven for whites but not for nonwhites)
4) a sample is then extracted from the loan "files" (which include whether or not the loan was granted) and tests for discrimination by seeing if the variable representing race of the applicant has a significant impact on the probability of the loan being granted, controlling for relevant factors (income, net worth, etc.)
5) the sample, or "bank," is then graded as being a discriminator or not depending on the outcome of this test.
Note that any time the bias parameter is greater than zero, the bank is in fact a discriminator. A good test avoids both false positives and false negatives.
[reference: "A Monte Carlo Examination of Bias Tests in Mortgage Lending," Paul W. Bauer and Brian A. Cromwell, Economic Review, Federal Reserve Bank of Cleveland, V. 30 no. 3 (1994 3rd quarter): 27-40]

--there are applications of both deduction and induction in economics, but there are many more applications of statistics, or inference, in economics (and business; use more business examples, but segueing into mostly economics examples as the course goes on--of course I have a very expansive view of what constitutes an economics problem): we rarely know what the parameters are, except in artificial frameworks such as the Monte Carlo testing case
[overhead on granola example]

now let's move carefully towards being able to do problems like the one I just showed, first by deriving confidence intervals and then creating a framework of formal hypothesis testing.

Ch. 8

confidence interval for a single mean:

assume interested in a 95% confidence interval

show the algebra behind turning the confidence interval around:

look for place in Z-table where there is .025 of probability in each table:

Z-value is 1.96

so $Pr(-1.96 < Z < 1.96) = .95$

going back into X-terms:

$Pr(\mu - 1.96SE < \overline{X} < \mu + 1.96SE) = .95$

$\mu - \overline{X} - 1.96SE < 0 < \mu - \overline{X} + 1.96SE$
$-\overline{X} - 1.96SE < -\mu < -\overline{X} + 1.96SE$

$Pr(\overline{X} + 1.96SE > \mu > \overline{X} - 1.96SE) = .95$

remember SE $= \frac{\sigma}{\sqrt{n}}$

so can write the 95% confidence interval for the population mean when σ is known (theory):

$$\mu = \overline{X} \pm z_{.025}\frac{\sigma}{\sqrt{n}}$$

now improve on this—if do not know the population variance and have a small sample size, use the sample variance but read the critical value from the t-table instead of the standard normal table to adjust the interval wider to account for the additional unreliability introduced by estimating the variance:

95% confidence interval for the population mean when σ is unknown (in practice):

$$\mu = \overline{X} \pm t_{.025}\frac{s}{\sqrt{n}}$$

t-table is tabulated by degrees of freedom, where d.f. = n-1 (which is also the divisor used in calculating s^2). The random variable t indexes a family of distributions that are more spread out than Z but approach Z as the d.f. goes to infinity. Its distribution is called Student's t because its inventor, William Gosset, published under the pseudonym "student." He was employed by the Guinness Brewery, which for some reason required him to publish under a pseudonym.

what if want to test the difference between two means?

95% confidence interval for the difference between population means if the population variances are known (theory):

$$(\mu_1 - \mu_2) = (\overline{X_1} - \overline{X_2}) \pm z_{.025}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

95% confidence interval for the difference between population means if the population variances are known and equal (theory):

$$(\mu_1 - \mu_2) = (\overline{X_1} - \overline{X_2}) \pm z_{.025}\,\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

95% confidence interval for the difference between population means if the population variances are unknown but assumed to be equal (in practice):

$$(\mu_1 - \mu_2) = (\overline{X_1} - \overline{X_2}) \pm t_{.025}\,s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{where } s_p^2 = \frac{\sum_{n_1}(X_1 - \overline{X_1})^2 + \sum_{n_2}(X_2 - \overline{X_2})^2}{(n_1 - 1) + (n_2 - 1)}$$

and d.f. for use in the t-table $= (n_1 - 1) + (n_2 - 1)$, same as the divisor in s_p^2

the rest of the book makes the assumption of equal population variances in the different samples--think about how defensible this is for many cases (e.g., pay for women and men)

we also assumed above that the two samples are considered to be independent. What if we instead use dependent, or <u>matched/paired</u> samples?

95% confidence interval for the mean population difference:

$$\Delta = \overline{D} \pm t_{.025}\frac{s_D}{\sqrt{n}}$$

where $\overline{D} = \frac{\sum D}{n} = \frac{\sum(X_1 - X_2)}{n}$, the average difference in the sample, where n = number of paired measurements, so the set of differences is treated as a single sample in this case

$$\text{so } s_D^2 = \frac{\sum_n (D - \bar{D})^2}{n - 1}$$

note that the mean of the differences Δ equals the difference of the means $(\mu_1 - \mu_2)$

95% confidence interval for a proportion, for large n:

$$\pi = P \pm 1.96 \sqrt{\frac{P(1 - P)}{n}}$$

if n is small, would use the graphical/geometrical method shown in Figure 8-4
--shows that confidence intervals are not always symmetric
--and not always a simple formula

95% confidence interval for the difference between two population proportions, for large n_1 and n_2, assuming independent samples:

$$(\pi_1 - \pi_2) = (P_1 - P_2) \pm 1.96 \sqrt{\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}}$$

note: can easily adjust level of confidence for interval in <u>all</u> the above formulas by using standard normal or t-tables and picking different critical value; .90 level makes the interval narrower, .99 level makes the interval wider

problem: confidence intervals should not always be symmetric
--and not always a simple formula
classical method can give values outside the range (e.g., for π, <0 or >1--see the graph 8-4 for another way to adjust for the range and allow for asymmetry)

next class we'll see another way of calculating confidence intervals, the bootstrap method, that avoids these two problems

next class: review and extend Ch. 8 material, start Ch. 9