7th Class **6/15/10**

"The race may not always go to the swift, nor the battle to the strong, but that's the way to bet it."--Damon Runyon

another take on correlation and causation [show cartoon from webpage]

last word on probability
[overhead on probability]
Jacob Bernouilli, posthumously published Ars Conjectandi in 1713; this is the beginning of the mathematical theory of probability--he developed the first law of large numbers

Ch. 6

random sampling
[overheads on sampling]

important insight: each individual observation in a random sample has the population probability distribution p(x).

> A sample is called a simple random sample (SRS) if each member of the population is equally likely to be chosen every time an observation is drawn.

A very simple random sample (VSRS) is a sample whose n observations $X_1, X_2, \ldots, X_n$ are independent. The distribution of each X is the population distribution p(x); that is, $p(x_1) = p(x_2) = \ldots = p(x_n) = p(x)$. Then each observation has the mean μ and standard deviation σ of the population.

This does not hold when the population is small and sampling is done without replacement. Note that nonreplacement is actually more efficient because each observation brings fresh information, but this gain in efficiency has to be weighed against the cost of having to use a more complicated procedure for estimating the mean

and standard deviation of the population, and in general is not worth doing/assuming unless the population is very small.

another important insight: Because of averaging, the sample mean $\overline{X}$ is not as extreme (doesn't vary so widely) as the individuals in the population.

We can build up a sampling distribution for $\overline{X}$, denoted $p(\overline{x})$, by taking repeated sets of observations and calculating their mean.

note that $E(\overline{X}) = \mu$

this is because $\overline{X} = \frac{1}{n}[X_1 + X_2 + \ldots + X_n]$

so $E(\overline{X}) = \frac{1}{n}[E(X_1) + E(X_2) + \ldots + E(X_n)]$

$= \frac{1}{n}[\mu + \mu + \ldots + \mu]$

$= \frac{1}{n}[n\mu]$

$= \mu$

also, $Var(\overline{X}) = \frac{\sigma^2}{n}$; since all the observations $X_1, X_2, \ldots, X_n$ are independent, we can use the formula from the last chapter to calculate the variance of a linear function of random variables (because of independence, no covariance terms to mess things up in extending this from 2 random variables to n random variables):

$\overline{X} = \frac{1}{n} X_1 + \frac{1}{n} X_2 + \ldots + \frac{1}{n} X_n$

so $Var(\overline{X}) = (\frac{1}{n})^2 Var(X_1) + (\frac{1}{n})^2 Var(X_2) + \ldots + (\frac{1}{n})^2 Var(X_n)$

$= (\frac{1}{n})^2 \sigma^2 + (\frac{1}{n})^2 \sigma^2 + \ldots + (\frac{1}{n})^2 \sigma^2$

$$= \left(\frac{1}{n}\right)^2 (n\sigma^2)$$

$$= \frac{\sigma^2}{n}$$

and the standard deviation of $\overline{X} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

the standard deviation of $\overline{X}$ is called the <u>standard error</u> of $\overline{X}$, or SE

note this goes to 0 as n goes to infinity

[note in the book they use the term <u>moments</u> of the sample mean]

<u>Normal Approximation Rule</u>, a.k.a. <u>Central Limit Theorem</u>

as n increases, the sampling distribution of $\overline{X}$ becomes concentrated around μ and becomes more and more of a normal distribution, <u>regardless</u> of the underlying shape of the population distribution

[draw the distinction between:

the distribution of $\overline{X}$ (with parameters $\mu, \frac{\sigma}{\sqrt{n}}$)

and the population distribution (with parameters μ, σ)]

1810 Pierre Simon Laplace came up with the Central Limit Theorem, where Central is meant as "fundamental"

repercussion of this theorem: can answer questions about how close the sample mean $\overline{X}$ is to the population mean μ by using the standard normal table after transforming $\overline{X}$:

$$Z = \frac{\overline{X} - \mu}{SE}$$

Normal Approximation Rule for Proportions:

as n increases, the sampling distribution of P, the sample proportion, becomes concentrated around π and becomes more and more of a normal distribution, regardless of the underlying shape of the population distribution

the standard error of P is $\sqrt{\frac{\pi (1 - \pi)}{n}}$, which goes to 0 as n gets large

so P is distributed with parameters (π, $\sqrt{\frac{\pi (1 - \pi)}{n}}$)

therefore, we can again use standard normal tables to describe how close the sample mean P is to the population proportion π by using the standard normal table after transforming P:

$$Z = \frac{P - \pi}{SE}$$

note that proportions are simply means for (0,1) (dummy) variables

do a couple of Ch. 6 type problems:

do problem 6-12:
mean 150, sd 25
capacity 7800
n 50
a. what is the probability you end up overloaded?

can translate into a question about what is the average passenger weight

= Pr (X-bar > 156) = Pr (Z > (156-150) / (25 / SQRT(50))) = Pr(Z > 1.70) = .045

b. possible dependence in passenger weights (e.g. plane booked by a football team)

problem 6-36:
what is Pr (P > .50)?
pi = 43.2 / (34.9+43.2) = 0.55

$SE(P) = SQRT[(.55)(.45)/1500] = .0128$

so $Pr(P > .50) = Pr\ (\ Z > (.50-.55)/.0128)) = Pr(Z > -.05/.0128)$

$= Pr(Z > -4) = 1 - Pr(Z>4)$ (by symmetry) = basically zero

QAC students can leave; a few review problems for Econ 300 students from old tests:

Person A tells the truth with probability $\frac{2}{3}$, person B with probability $\frac{3}{4}$, and person C with probability $\frac{4}{5}$.

Persons A and B make a statement which person C denies. What is the probability that the statement is true?

Pr(A tells the truth)*Pr(B tells the truth)*Pr(C lies) $= \frac{2}{3} * \frac{3}{4} * \frac{1}{5} = \frac{6}{60}$

Pr(A lies)*Pr(B lies)*Pr(C tells the truth) $= \frac{1}{3} * \frac{1}{4} * \frac{4}{5} = \frac{4}{60}$

then Pr(statement is true | A and B agree and C disagrees) $= \frac{\frac{6}{60}}{\frac{6}{60} + \frac{4}{60}} = 0.6$, or 60%

Suppose that 3% of interviewees for a position with the National Security Agency lie on a certain question during a lie detector test. Suppose that 5% of the time the lie detector labels as a truthteller someone who is actually lying, and that 10% of the time the lie detector labels as a liar someone who is actually telling the truth.

a. For a randomly picked person, what is the probability that the lie detector will label him or her a liar?

b. If someone passes the lie test, what is the probability that the person is actually telling the truth?

This is an application of Bayes' Theorem:

.97 truth	.90 "truth"	= .873
	.10 "lie"	= .097
.03 lie	.95 "lie"	= .0285
	.05 "truth"	= .0015
		100% of possible states of the world

a. Pr("liar") = .097 + .0285 = 0.1255, or 12.55%

b. Pr(truth | "truth") $= \frac{.873}{.873 + .0015} = 0.9982$, or 99.8%

Suppose that a drawing is to be done for four prizes. There are 100,000 ballots to choose from, but 30,000 of them contain a single person's name. What is the probability that the person will win:

a. All four prizes?

b. No prizes?
c. At least one prize?

These can all be solved using the individual binormial probability table with n = 4 and $\pi = .30$

a. s = 4, so Pr(win all prizes) = 0.008, or 0.8%

b. s = 0, so Pr(no prizes) = 0.24, or 24%

c. Pr(at least one prize) = 1 - Pr(no prizes) = 1 - 0.24 = 0.76, or 76%

A used car dealership has found that the length of time before a major repair is required for the cars it sells is normally distributed with a mean of 10 months and a standard deviation of 3 months.

a. How many cars will still be running perfectly after a year?

b. If the dealer wants no more than 10 percent of the cars to fail before the guarantee time, for how many months should the cars be guaranteed?

These are both solved using the standard normal probability table.

a. $Pr(X > 12) = Pr(Z > \frac{12 - 10}{3}) = Pr(Z > 0.67) = 0.251$, or 25.1%

b. $0.10 = Pr(Z > 1.28) = Pr(Z < -1.28) = Pr(\frac{X - 10}{3} < -1.28) = Pr(X < 6.16)$, so 6 months

At Daniel's Deli you can buy an 8 ounce or 12 ounce hamburger. You can also order a 16 or 24 ounce soft drink. Daniel has noticed that burger and soft drink combined orders have the following distribution:

	soft drink size	
hamburger size	16	24
8	.60	.15
12	.15	.10

a. Calculate the mean and variance for hamburger size.

b. Show that hamburger and soft drink sizes are dependent.

c. What is the correlation between hamburger and soft drink sizes?

a. E(hamburger) = 0.75*8 + 0.25*12 = **9 ounces**

Var(hamburger) = 0.75 + .25*9 = **3 ounces^2**

b. can show this several ways;

e.g., Pr(hamburger = 8 & drink = 16) = .6
Pr(hamburger = 8)*Pr(drink = 16) = .75*.75 = .5625; .6 ≠ .5625

c. calculate E(drink) = 18; Var(drink) = 12, and Covariance (hamburger, drink) = 1.2

so $\rho = \frac{1.2}{\sqrt{3 * 12}} =$ **0.20**