Econ 300/QAC 201: Quantitative Methods in Economics/Applied Data Analysis

5th Class **6/11/10**

"No one can possibly win at roulette unless he steals money from the table while the croupier isn't looking."
Albert Einstein

on Mon. will hand out review sheet (practice test) for test

couple of follow-up points:
show cartoons re weather and monty hall problem ; discuss weather probability calculations

Probability distributions
review our new terminology
random variable X can take on values x with probabilities p(x)

$E(X) = \sum xp(x)$

Expectations carry over into functions of X:

if $G = g(X)$, then $E(G) = E[g(X)] = \sum g(x)p(x)$

note one possible function $G = g(X) = (X - \mu)^2$

so $E[(X - \mu)^2] = \sum (x - \mu)^2 p(x)$

$= \sigma^2_x$

If G is a linear function of X:

$G = a + bX$

then the formulas for μ_G and σ_G are simple:

$\mu_G = a + b\mu_X$

$$\sigma_G = |b|\sigma_X$$

If G is nonlinear, have to use the general formulas to calculate μ_G and σ_G:

$$\mu_G = \sum g(x)p(x)$$

$$\sigma^2_G = \sum (g(x) - \mu_G)^2 p(x)$$

Now continue discussing discrete and continuous distributions from last class:

Probabilities and moments under common discrete and continuous probability distributions—in general all concepts translate between the two:

1) When showing a continuous probability distribution, need to use an integral sign--can't use a sum since all values are possible (i.e., continuous). Integrating over the range of the density function always gives you a value of 1. E.g., if p(x) is probability density function for a possible grade of 0 to 4:

$$\sum_{i=0}^{4} p(i) = 1 \qquad\qquad \int_0^4 p(x)\partial x = 1$$

2) can use integrals to get the moments of continuous distributions:

	Discrete	Continuous
μ:	$\sum xp(x)$	$\int xp(x)\partial x$
σ^2:	$\sum (x-\mu)^2 p(x)$	$\int (x-\mu)^2 p(x)\partial x$

[show website] monopoly squares: discrete distribution
[reference: Ian Stewart, "Monopoly Revisited," Scientific American October 1996, pp. 116-119]

3) the binomial distribution: this is a well-behaved discrete distribution
classic example: number of heads in several coin tosses (coin can be fair or unfair)

Need to know number of trials: n; probability of getting a head on each trial: π (assumed that trials are statistically independent); and number of heads that we are interested in: s (can also be interested in getting at least s heads). Then:

$$p(s) = \binom{n}{s} \pi^s (1 - \pi)^{n-s} = \frac{n!}{s!(n-s)!} \pi^s (1 - \pi)^{n-s}$$

[note that $\binom{n}{n}$ and $\binom{n}{0}$ are defined independently of the above statement to each equal 1]

To motivate this formula, it may be helpful to read the appendix to Section 4-3 in the back of the book.

example of use:

imagine flipping a biased coin 5 times, where there is a 60% chance it will come up heads. What is the probability that the number of heads will equal 3?

Note there are numerous ways in which you could get 3 heads in 5 tosses, all of which are equally likely: HHHTT, HTHTH, etc. -- 10 in fact But how does $\binom{5}{3}$ yield the right number of ways?

i) how many ways can you arrange/permute 5 distinct objects, H_1, H_2, H_3, T_1, T_2? answer: 5*4*3*2*1 or 5! = 120 ways/permutations

ii) given that we cannot distinguish between Hs (or between Ts) this would yield an overcount of the ways to get 3 heads in 5 tosses. So this figure for 5 distinct objects must be deflated to take account of this.

a. there are 3*2*1 = 3! = 6 ways to arrange the 3 Hs among themselves, none of which we can distinguish

b. and there are 2*1 = 2! = 2 ways to arrange the 2 Ts

c. so the total number of combinations, in which the order of the individual Hs and Ts among themselves does not matter, is

$$\frac{5!}{3!2!} = \frac{120}{6*2} = \frac{120}{12} = 10$$

then the event (number of heads = 3) includes $\binom{5}{3}$ = 10 outcomes, each with probability $(.60)^3(.40)^2$ = .035 of occurrence, so the probability of the event is:

$$\binom{5}{3}(.60)^3(.40)^2 = 10*.035 = .35 = 35\%$$

As a shortcut, can look up these probabilities in a binomial table (IIIb) and can look up cumulative binomial probabilities (to answer the questions $Pr(X > s)$ or $Pr(X \geq s)$ or their complements) in a table as well (IIIc) – e.g. can do Example 4-2 by just looking it up in the table

note that in problems of sampling without replacement, if n is very large we can use the binomial formula as a very good approximation rather than having to take into account the nonindependence of the trials as items are removed from the population (which normally would affect the probability that subsequent items are defective/flawless)

note also that $\mu = n\pi$ and $\sigma = \sqrt{n\pi(1 - \pi)}$

The binomial formula is useful, but note that there are still lots of questions for which we have to go back to the more involved event space:
Question: Do basketball players shoot in streaks? Is there such a thing as a "hot hand?"
Background: The average NBA player shoots around 50 percent from the field and offensive stars often take 20 or more shots in a game.
Question: what is the chance of flipping a coin 4 times and coming up with heads every time? What is the chance of getting 4 heads in a row on a series of 6 flips? What is the chance of getting 5 in a row? What is the chance of getting 6 in a row?

Let's answer the first question first: [switch to powerpoint presentation]

answer: 4 heads out of 4 shots: $(.5)^4 = \frac{1}{16} = 6\%$

note this expression is a reduction of the standard binomial formula:

$$p(4) = \frac{n!}{s!(n-s)!}\pi^s(1-\pi)^{(n-s)} = \frac{4!}{4!}(.5)^4(.5)^0$$

but it is more time-consuming to construct the change of getting at least 4 in a <u>row</u>, let alone exactly 4 in a row, as one cannot apply the binomial formula where order matters.

Need to count up all the different strings, just as with the 3-child family when we wanted to know the number of runs (problem 4-1), in a tree with 20 sets of branches: e.g., with 6 shots, there are $2^6 = 64$ equally likely possible outcomes:

HHHHHH

HHHHHT, … , TTTTTT — of which 8 contain a string of at least 4 Hs

(5 w/4 Hs in a row, 2 w/5 Hs in a row, 1 w/6 Hs)

so pr(4 in a row) = $\frac{5}{64}$ = 7.8%; pr(at least 4 in a row) = $\frac{8}{64} = \frac{1}{8}$ = 12.5%

clearly there are more chances as n increases of a particular run occurring. And with 20 shots, a string can contain more than one streak of 4 or more hits.

Note. Players are actually more likely to score after a miss than after a hit. Why is this? May result from belief in the hot hand myth: A shooter who thinks he is hot may take riskier shots, or the opposition, thinking him hot, might guard him more closely.
[full references: T. Gilovich and A. Tversky, "The Cold Facts about the Hot Hand in Basketball," Chance, Vol. 2, no. 1, pp. 16-21 (1989) and "The 'Hot Hand': Statistical Reality or Cognitive Illusion," Chance, Vol. 3, no. 4, pp. 31-34.]

Note. Streaks are extremely likely if n is large
[full reference: Mark Schilling, "Long Run Predictions," Math Horizons Spring 1994, pp. 10-12.]

This technique has been used on other sports questions since Gilovich and Tversky's seminal paper:

baseball:

1) team winning streaks in MLB: conclusion: random
[reference: Mitchell Laks, letter to the New York Times, 7/24/97]

2) hitting streaks in MLB: conclusion: random
[reference: S. Christian Albright, with comments by Jim Albert, Hal Stern, and Carl Morris, "A statistical analysis of hitting streaks in baseball,"
Journal of the American Statistical Association, Vol 88, No 424, pp 1175-1194]

[go on in the powerpoint to show other cases of calculations]

[also reference websites on hot hand in general and hitting streaks in particular]
[give handout with the material on it]

this doesn't mean that all streaks are random:

3) Cal Ripken's playing streak: conclusion: nonrandom
[references: Buster Olney, "Statistically speaking, it's unbreakable,"
Baltimore Sun, 5 Sept. 1995, 8C
and David Leonhardt, "The last Ripken; After 2,131 articles, the statistical truth, he's impossibly good." Washington Post, 10 Sept. 1995, C5]

tennis: success-breeds-success model holds (i.e., nonrandom)
[reference: David Jackson and Krzysztof Mosurski, "Heavy defeats in tennis: psychological momentum or random effect?" Chance Magazine, Spring 1997, pp. 27-34]

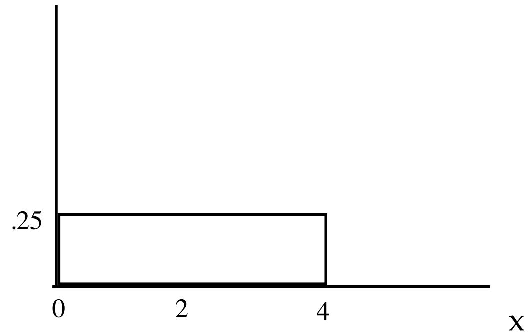
4) the uniform distribution: this can be modeled as a discrete or a continuous distribution

--in discrete form, it is any distribution where the n possible outcomes all occur with equal probability. E.g., for one throw of the fair die/toss of the fair coin, all outcomes are equally likely. [but for multiple throws, one has to use the binomial distribution to consider the varying likelihood of different outcomes]

--in continuous form, any particular value is extremely unlikely, but can find the probability of an event like getting a 3 or better (in the continuous grade scale 0-4) by finding the area under the density function between the values of interest

[draw sketch of probability density function for the uniform function from 0 to 4]

p(x)



In general the probability density function of the uniform distribution is just $\frac{1}{\text{range}}$, so here $p(x) = \frac{1}{4}$, and the total area under the function = 1:

5) the normal distribution: this is a continuous distribution

$$\text{general normal distribution: } p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The normal is centered at the mean μ and has a symmetric bell shape [draw graph]. Note that π is the constant pi (about 3.14) here, not a probability, and e is also a constant (about 2.72). The normal distribution is not analytically integrable, so one has to use tables to look up its values--table not just for convenience, but necessary, as with logarithms. In order to evaluate the probability of any range of values in a normal distribution for which we know the mean and standard deviation, it turns out we can just use the table for the standard normal, where $\mu = 0$ and $\sigma = 1$, for all other normal distributions are a scaled function of this one. This is often known as the standardized score Z, or a Z-score, where Z is a linear function of X:

$$Z = \frac{X - \mu}{\sigma}$$, so Z has mean 0 and s.d. 1.

Z is distributed by the standard normal $p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2}$

Using the table from the standard normal, we can calculate the answers to questions such as $\Pr(Z > 1.6)$, $\Pr(-1 < Z < 1)$—and $\Pr(X > 3)$ if we know the mean and s.d. for X. Note to answer many of these questions, we exploit the symmetry of the normal distribution.

Use of expected values and probability distributions in decisionmaking situations.
[overhead and handout]

Question: Imagine you are an oil company executive trying to decide at which one of three sites your company should drill for oil. At site 1, there is a 20% probability of finding oil, thereby making $30,000,000; otherwise the loss will be $3,000,000. At site 2, there is a 10% probability of making $70,000,000; otherwise the loss will be $4,000,000. At site 3, there is a 10% probability of making $30,000,000; otherwise the loss will be $2,000,000. Where should you drill?

answer: it depends on your philosophy. In comparing courses of action, rational decisions can be made at least 4 different ways:

1) if your philosophy is to maximize the probability of success, drill at site 1

2) if you have a go-for-broke, or max-max philosophy, drill at site 2 to maximize the maximum profit

3) if you have a reserved, or max-min philosophy, drill at site 3 to maximize the minimum profit

4) if you are risk-neutral, choose the site which maximizes the expected profit:

site 1: E(profit) = .2*($30M) + .8*(-$3M) = $3.6M
site 2: E(profit) = .1*($70M) + .9*(-$4M) = $3.4M
site 3: E(profit) = .1*($30M) + .9*(-$2M) = $1.2M; so drill at site 1

Thus each of the 3 sites is "best" in some sense. Maximizing the expected value is usually the best when one is in the position of needing to make many decisions like it. But if one is going to make only one decision, "average outcome" may not be important and some other philosophy may govern.

for Monday, read Ch. 5