On Monday, Oct 15, we were working with continuous random variables: I was showing graphically how to simulate a RV when given its CDF.

There are three ways to simulate values of a continuous random variable for which the CDF is given:

1. If the rv has a named distribution and the software includes it, just use the software.

e.g. Random 100 c4;
normal 15 4.

2. If the cdf is given as a non-decreasing continuous function, invert the function and apply the inverted function to 100 simulated values of U(0,1).

e.g. If F(x) = 1-exp(-x) then the inverse function is x=-ln(1-u) so just plug in the simulated values of U(0,1) for u in this formula for x to simulate X with cdf F(x)

3. If the cdf is given in tabulated form (as it might be from a sampling survey), then the inversion of the function described by the table must be done by a table-look-up program:

e.g. I generated a cdf table to start with (actually I cheated a bit and used a KNOWN distribution N(15,4) and MINITAB to generate the table. This is useful for checking the program below). The table has values of x from 1 to 30 in C1 and corresponding values of F(x) in C2. In C3 I generate 100 U(0,1) variates using MINITABs RANDOM command. Then I apply the "invert" program to use the table (and interpolation) to go from the 100 Us to the 100 Xs.

Here is the MINITAB description of this third situation:

#These MINITAB commands do the following:
#1. Define the cdf of a N(15,4) RV
#2. Compute the cdf as a table
#3. Generate 100 U(0,1) variates
#4. Pretend the table in step 2 is the only info about the cdf
and use the "invert" program to simulate the RV that has
this tabular cdf.
#5. Check the invert process by using the built-in invert in MINITAB, using
fact that the cdf is N(15,4)
Compare the two inversions to show they are the same, by dotplots.

```
set c1
1:30
end
cdf c1 c2;
normal 15 4.
rand 100 c3;
uniform 0 1.
%invert
dotp c4
invcdf c3 c5;
normal 15 4.
dotp c4 c5;
same.
```

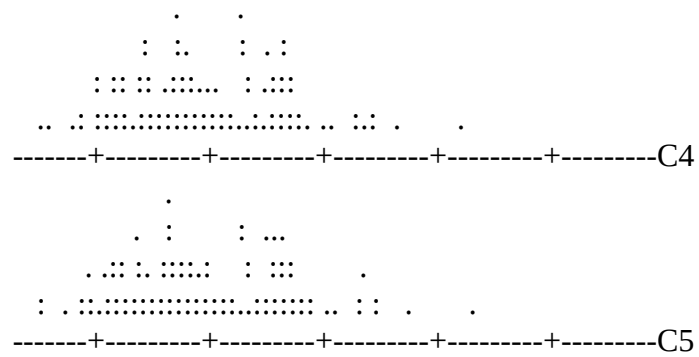
The "invert" macro is:

```
Gmacro
invert
do k1=1:100
let k2=1
while c3(k1) GT c2(k2)
let k2=k2+1
endwhile
let k2=k2-1
let c4(k1)=k2-1 + (c3(k1)-c2(k2-1))/(c2(k2)-c2(k2-1))
enddo
endmacro
```

The output for my one run is:

```
MTB > dotp c4 c5;
SUBC> same.
```

Character Dotplot



8.0 12.0 16.0 20.0 24.0 28.0

Small differences are due to the interpolation (between tabulated values of the cdf).

This ends the discussion of simulating a rv with a given CDF. [Note that the examples I used for the three methods were N(15,4), exponential(1), N(15,4) respectively. Why did I not use N(15,4) for the algebraic inversion method? Because the normal CDF does not have a sinple explicit formula (that's why there are tables.), and thus inversion of the cdf algebraically is not available for normal – but it is for expo.]

Joint Distributions:

A discrete example:

X: number of cars in household
Y: number of adults in household

Suppose that X ranges in 0 to 3
And Y ranges in 1 to 4
(to keep things simple)

A table of observations from 100 households might look like this:

	X=0	X=1	X=2	X=3
Y=1	3	15	3	0
Y=2	1	20	15	2
Y=3	0	10	17	3
Y=4	0	5	4	2

From this we might estimate the corresponding probabilities as

	X=0	X=1	X=2	X=3
Y=1	.03	.15	.03	0
Y=2	.01	.20	.15	.02
Y=3	0	.10	.17	.03
Y=4	0	.05	.04	.02

P(X=2 and Y=3) = .17 for example.

These probabilities are JOINT probabilities. Note that they sum to 1.

We can write this as $\sum_{x=0}^{3}\sum_{y=1}^{4} P(X=x, Y=y) = 1$

Suppose we focus on families with three adults (Y=3), ie. CONDITION on Y=3

P(X=2|Y=3) = ?

If we go back to the original data, we would estimate this probability to be 17/(10+17+3) = 17/30

Or, directly from the definition of conditional probability, we can compute P(X=2|Y=3) = P(X=2 AND Y=3)/P(Y=3) = .17/P(Y=3)

And P(Y=3)= $\sum_{x=1}^{4} P(X=k, Y=3)$=0+.10+.17+.03 = .30 (A MARGINAL probability)

So, P(X=2|Y=3) = .17/.30 = 17/30 as before.

This example illustrates the relationship of JOINT, MARGINAL and CONDITIONAL probabilities.
Also, it shows why the general formula for discrete rvs

P(X=i|Y=j) = P(X=i AND Y=j) / P(Y=j)

is a reasonable definition of CONDITIONAL probability.

Keep these in mind while moving into analogous formulas for continuous rvs.

Refer to text:

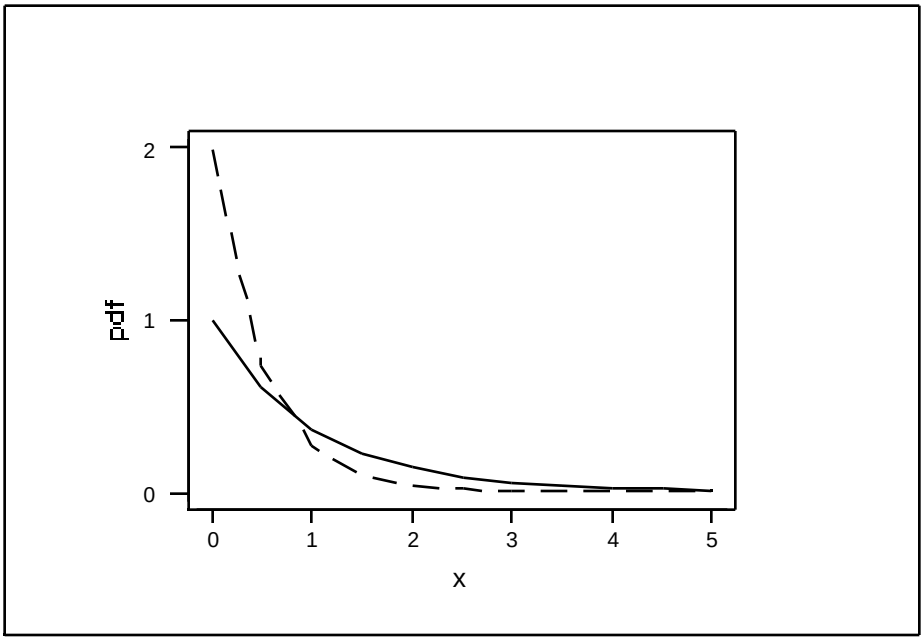
Definition 5.4-1 p 220 shows that the double integral of the joint density = 1
Definition 5.4-2 p 223 shows that the marginal density is the integral of the joint density over other variables.
Definition 5.4-3 p 224 shows that the conditional density is the ratio of the joint density to the marginal density,

By comparing these with the similar formulas for the discrete case, these definitions should seem intuitive and understandable.

Ch 6: Special Continuous Random Variables

Exponential Distribution: Models certain time durations such as interarrival times
lifetimes of things that don't age (electronic parts)
survival under constant hazard

density $f_X(x) = \lambda e^{-\lambda x}$ $x>0$ and cdf $F_X(x)=1-e^{-\lambda x}$ for $x \geq 0$



Which exponential pdf has the larger λ?
Is the integral 1 in both cases?

The exponential is a heavy tailed distribution. Here are 1000 simulated ….

Each dot represents 12 points
:

```
        : .
        :::
       :::::..
       :::::::.
       ::::::::.:....
       :::::::::::::::.:..... ..   . .          .
       +---------+---------+---------+---------+---------+-------x
     0.0       2.0       4.0       6.0       8.0      10.0
```

a few of the larger values:

4.00789 4.04294 4.09804 4.11207 4.22699 4.27774 4.37271
4.50019 4.50145 4.50855 4.51587 4.55938 4.92599 5.01175
5.29882 5.92928 5.98954 6.42500 6.43863 9.01679

Mean and SD of the exponential: both are 1/λ

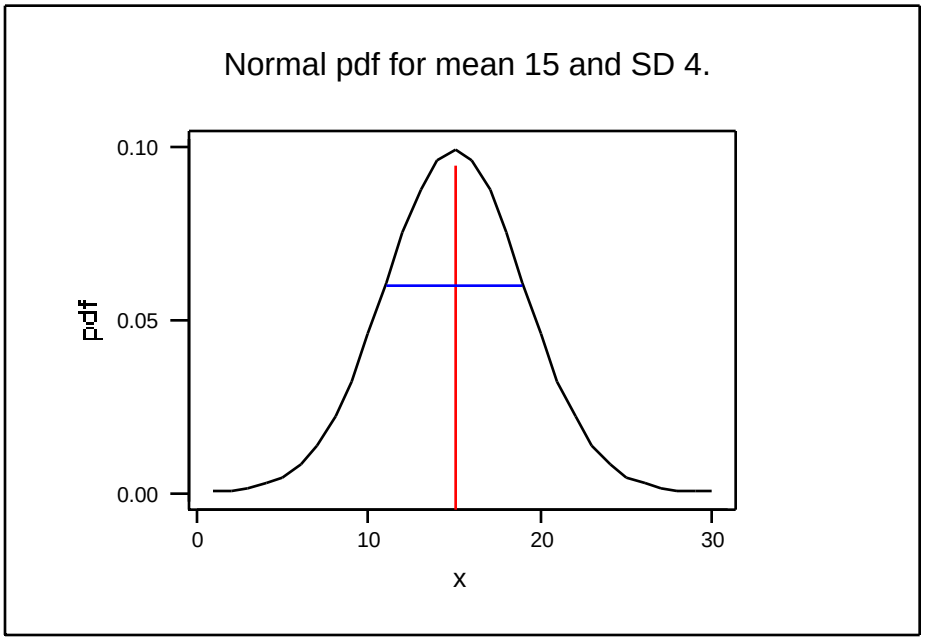
Very important property of the exponential: Lack-of-memory

P(X>t+x|X>x)=P(X>t) see p 239

Relate to survival of electronic components, or to people?

Normal Random Variable:

pdf is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$ -inf <x<+inf p 241



Note symmetry about mean, point of inflection at +- 1SD.

Mean SD?

Probability within 1 SD, 2 SDs, 3 SDs?

Relationship of any normal to N(0,1).
(Express quanitities in terms of SDs from mean).
CDF ?

What does normal model? Averages or sums or linear combos.
Gamma RV: See p 251 $G(\alpha,\beta)$ α is shape parameter, β is scale parameter.

Models durations. Especially waiting times until $\alpha-$th event occurs.
Like sum of α exponentials each with mean β.

Relation to Chi-Square

Relation to Normal