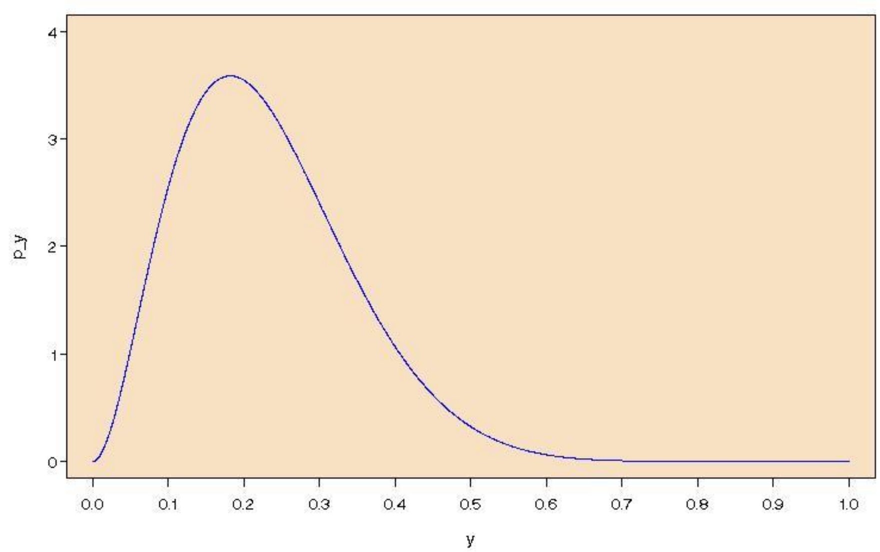
ISQS 5347 Midterm 2, Spring 2010

Instructions: Closed book and notes, no electronic devices. Put answers on blank paper. Points out of 100 are in parentheses.

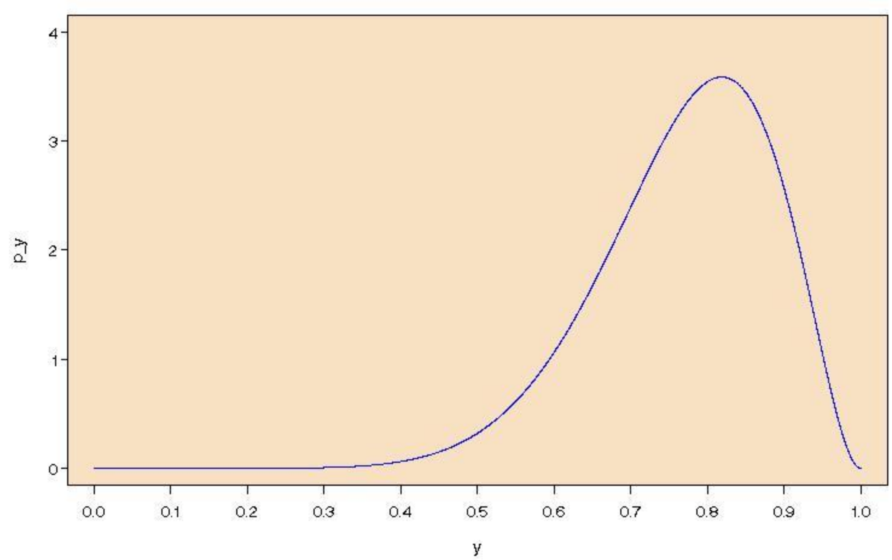
1.(20) Random variables X and Y are dependent. What does this tell you about the conditional distributions p(y | X=10) and p(y | X=20)? Draw two graphs that explain it clearly.

Solution:

When X=10, the pdf of Y might look like this:

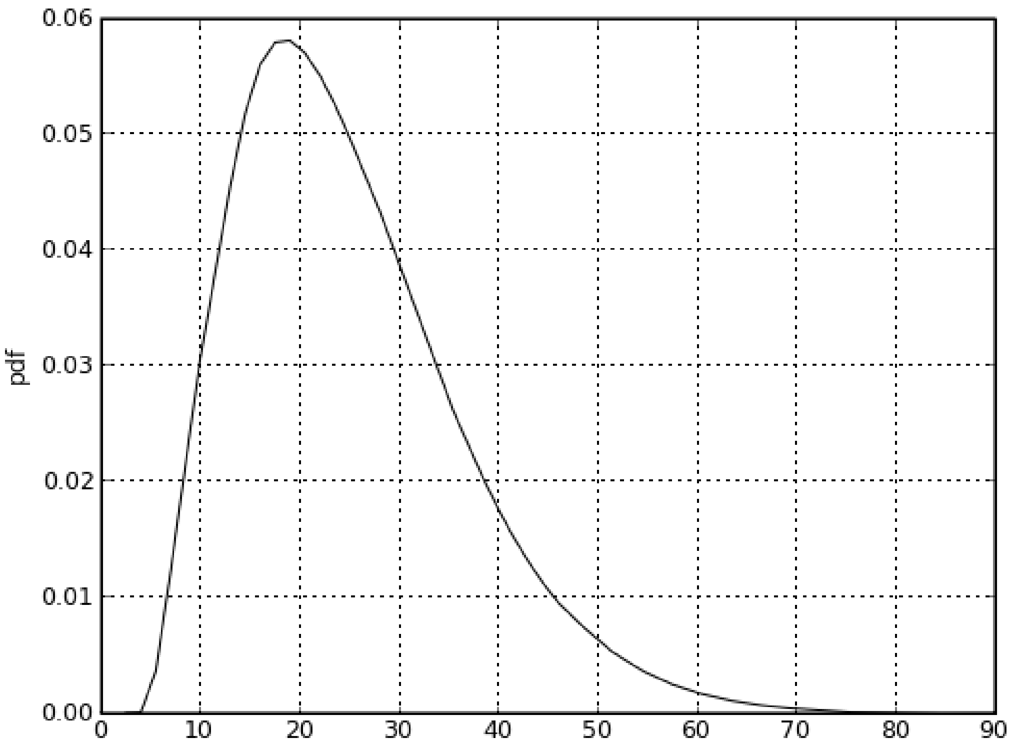


And when X=20, the pdf of Y might look like this:



The distribution of Y is clearly different when X=10 versus when X=20. This shows that Y and X are dependent. If they were independent, then these pdfs would have to be exactly the same.

2.(20) Here is a pdf:



Write down ten numbers that could have been produced as an iid sample from this pdf.

Solution: 40.3, 10.3, 25.7, 28.9, 50.9, 7.1, 19.7, 20.8, 22.0, 21.9.

Notes:
1. The distribution is continuous, so at least one decimal should be shown.
2. The data should not be arranged systematically, eg from smallest to largest.
3. The range of possible values is from 5 to 75, with values from 10-30 more frequent.

3.A.(15) In a survey of n=25 students, the average desire to buy a car was $\bar{x} = 2.0$ on a five point scale, with standard deviation s =1.0. Give an approximate 95% confidence interval for μ.

Solution: The formula is $\bar{x} \pm 2s/\sqrt{n}$. Here we have $2.0 \pm 2(1.0)/\sqrt{25}$, or 2.0 ± 0.4, or $1.6<\mu<2.4$.

Notes: The interval is approximately a 95% interval for two reasons: (i) The distribution of $\bar{X}$ is only approximately normal (by the central limit theorem). (ii) The estimate s is not equal to the estimand σ.

3.B.(15) In the context of problem 3.A., identify (i) an estimate, (ii) an estimand, and (iii) an estimator. (Note: (iii) will require explanation.)

Solution. (i) The number $\bar{x} = 2.0$ is an estimate. (ii) The number μ is the estimand corresponding to the estimate $\bar{x} = 2.0$. (iii) The number $\bar{X}$ is the estimator. To understand $\bar{X}$, we have to go back before the data collection, where $\bar{X}$ was unknown, and therefore random. We can visualize many different values of $\bar{X}$ as arising from different samples from the same process, which in this example is the process that puts students in this particular sample. Different students arise from different samples, giving different values of $\bar{X}$.

4.A.(5) An iid sample is obtained from the following distribution.

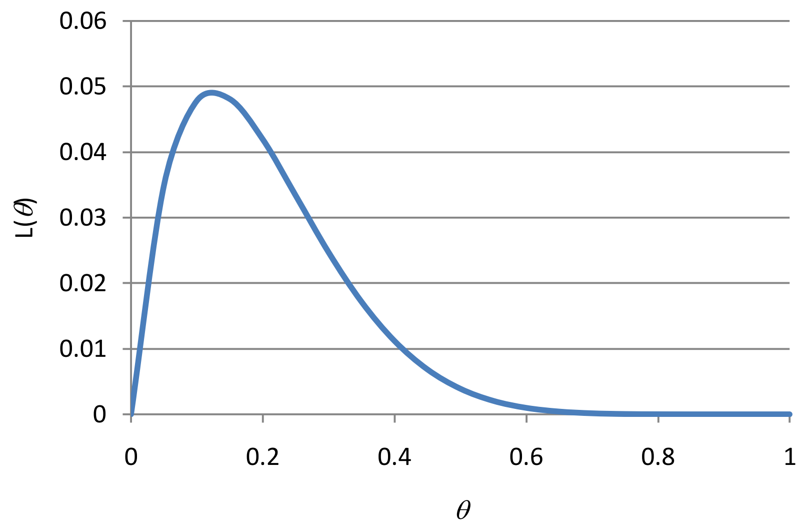
y	$p(y \mid \theta)$
F	$1-\theta$
S	θ

The sample is F,F,F,S,F,F,F,F.
Give the likelihood function for θ.

Solution: The likelihood of the sample is the product of the likelihoods for the individual observations. The likelihood of getting an "F" is $1-\theta$, and the likelihood of getting a "S" is θ. Hence the likelihood of the sample is

$L(\theta) = (1-\theta)\times(1-\theta)\times(1-\theta)\times\theta\times(1-\theta)\times(1-\theta)\times(1-\theta)\times(1-\theta)$, or
$L(\theta) = \theta(1-\theta)^7$.

4.B.(10) The graph of the likelihood function of 4.A. is shown below. What does it tell you about θ?



Solution: The graph shows that the range of likely values of θ (which in this example is the probability of a "S" occurring), is between around .01 and .6. The more likely value of θ are between .1 and .2, which is sensible considering there were 1 out of 8 "S" values in the sample, corresponding to a best guess of 1/8 = .125 for θ. But the likelihood function shows that we are very uncertain as to the actual value of θ based on these data – it might be anything from .01 to .6.

5. Explain what the given lines labeled A, B and C in the following code do. (5 points each)

```
data time;
   input y;
datalines;
4
5
6
9
;
proc nlp data=time;     /* line A */
   max logf;            /* line B */
   parms lambda;        /* line C */
   pdf = lambda*exp(-lambda*y);
   logf =log(pdf);
run;
```

Solution:
Line A tells SAS to use the "NLP" procedure, using the data set "work.time". The NLP procedure is used to optimize functions, like the "Solver" in EXCEL.

Line B tells the NLP procedure to maximize the value of the variable “logf” that is defined later.

Line C tells the NLP procedure to change the value of the variable “lambda” in order to maximize the logf value.