

TABLE OF CONTENTS - VOLUME 1

INTRODUCTORY COMMENTS

MODELING

SECTION 1 - PROBABILITY REVIEW	LM-1
PROBLEM SET 1	LM-9
SECTION 2 - REVIEW OF RANDOM VARIABLES - PART I	LM-19
PROBLEM SET 2	LM-29
SECTION 3 - REVIEW OF RANDOM VARIABLES - PART II	LM-35
PROBLEM SET 3	LM-43
SECTION 4 - REVIEW OF RANDOM VARIABLES - PART III	LM-51
PROBLEM SET 4	LM-59
SECTION 5 - PARAMETRIC DISTRIBUTIONS AND TRANSFORMATIONS	LM-63
PROBLEM SET 5	LM-69
SECTION 6 - DISTRIBUTION TAIL BEHAVIOR	LM-73
PROBLEM SET 6	LM-77
SECTION 7 - MIXTURE OF TWO DISTRIBUTIONS	LM-79
PROBLEM SET 7	LM-85
SECTION 8 - MIXTURE OF n DISTRIBUTIONS	LM-91
PROBLEM SET 8	LM-97
SECTION 9 - CONTINUOUS MIXTURES	LM-105
PROBLEM SET 9	LM-111
SECTION 10 - POLICY LIMITS AND THE LIMITED LOSS RANDOM VARIABLE	LM-119
PROBLEM SET 10	LM-123
SECTION 11 - POLICY DEDUCTIBLE (1), THE COST PER LOSS RANDOM VARIABLE	LM-125
PROBLEM SET 11	LM-131

MODELING

SECTION 12 - POLICY DEDUCTIBLE (2), THE COST PER PAYMENT RANDOM VARIABLE	LM-141
PROBLEM SET 12	LM-147
SECTION 13 - POLICY DEDUCTIBLES APPLIED TO THE UNIFORM, EXPONENTIAL AND PARETO DISTRIBUTIONS	LM-157
PROBLEM SET 13	LM-165
SECTION 14 - COMBINED LIMIT AND DEDUCTIBLE	LM-169
PROBLEM SET 14	LM-175
SECTION 15 - ADDITIONAL POLICY ADJUSTMENTS	LM-187
PROBLEM SET 15	LM-191
SECTION 16 - MODELS FOR THE NUMBER OF CLAIMS AND THE $(a, b, 0)$ CLASS	LM-195
PROBLEM SET 16	LM-203
SECTION 17 - MODELS FOR THE AGGREGATE LOSS, COMPOUND DISTRIBUTIONS (1)	LM-215
PROBLEM SET 17	LM-219
SECTION 18 - COMPOUND DISTRIBUTIONS (2)	LM-241
PROBLEM SET 18	LM-247
SECTION 19 - MORE PROPERTIES OF THE AGGREGATE LOSS RANDOM VARIABLE	LM-259
PROBLEM SET 19	LM-263
SECTION 20 - STOP LOSS INSURANCE	LM-275
PROBLEM SET 20	LM-281
SECTION 21 - DISCRETE TIME RUIN MODELS	LM-287
PROBLEM SET 21	LM-291

MODEL ESTIMATION

SECTION 1 - REVIEW OF MATHEMATICAL STATISTICS (1)	
ESTIMATORS	ME-1
PROBLEM SET 1	ME-7
SECTION 2 - REVIEW OF MATHEMATICAL STATISTICS (2)	
CONFIDENCE INTERVALS AND HYPOTHESIS TESTS	ME-11
PROBLEM SET 2	ME-17
SECTION 3 - NON-PARAMETRIC EMPIRICAL POINT ESTIMATION	ME-23
PROBLEM SET 3	ME-31
SECTION 4 - KERNEL SMOOTHING ESTIMATORS	ME-37
PROBLEM SET 4	ME-51
SECTION 5 - EMPIRICAL ESTIMATION FROM GROUPED DATA	ME-61
PROBLEM SET 5	ME-67
SECTION 6 - ESTIMATION FROM CENSORED AND TRUNCATED DATA	ME-75
PROBLEM SET 6	ME-85
SECTION 7 - PROPERTIES OF SURVIVAL PROBABILITY ESTIMATORS	ME-99
PROBLEM SET 7	ME-107
SECTION 8 - MOMENT AND PERCENTILE MATCHING	ME-119
PROBLEM SET 8	ME-131
SECTION 9 - MAXIMUM LIKELIHOOD ESTIMATION	ME-145
PROBLEM SET 9	ME-155
SECTION 10 - MAXIMUM LIKELIHOOD ESTIMATION FOR THE EXPONENTIAL DISTRIBUTION	ME-167
PROBLEM SET 10	ME-173
SECTION 11 - MAXIMUM LIKELIHOOD ESTIMATION FOR PARETO AND WEIBULL DISTRIBUTIONS	ME-179
PROBLEM SET 11	ME-189

MODEL ESTIMATION

SECTION 12 - MAXIMUM LIKELIHOOD ESTIMATION	
FOR DISTRIBUTIONS IN THE EXAM C TABLE	ME-195
PROBLEM SET 12	ME-205
SECTION 13 - PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS	ME-213
PROBLEM SET 13	ME-217
SECTION 14 - GRAPHICAL EVALUATION OF ESTIMATED MODELS	ME-227
PROBLEM SET 14	ME-231
SECTION 15 - HYPOTHESIS TESTS FOR FITTED MODELS	ME-237
PROBLEM SET 15	ME-249
SECTION 16 - THE COX PROPORTIONAL HAZARDS MODELS	ME-267
PROBLEM SET 16	ME-277

INTRODUCTORY COMMENTS

This study guide is designed to help in the preparation for the Society of Actuaries Exam C and Casualty Actuarial Society Exam 4. The exam covers the topics of modeling, model estimation, construction and selection, credibility, simulation and risk measures.

The study manual is divided into two volumes. The first volume consists of a summary of notes, illustrative examples and problem sets with detailed solutions on the modeling and model estimation topics. The second volume consists of notes examples and problem sets on the credibility, simulation and risk measures topics, as well as 13 practice exams and the May 07 exam with detailed solutions.

The practice exams all have 40 questions. The level of difficulty of the practice exams has been designed to be similar to that of the past 4-hour exams. Some of the questions in the problem sets are taken from the relevant topics on SOA/CAS exams that have been released prior to 2007 but the practice exam questions are not from old SOA exams.

I have attempted to be thorough in the coverage of the topics upon which the exam is based. I have been, perhaps, more thorough than necessary on a couple of topics, such as maximum likelihood estimation, Bayesian credibility and applying simulation to hypothesis testing.

Because of the time constraint on the exam, a crucial aspect of exam taking is the ability to work quickly. I believe that working through many problems and examples is a good way to build up the speed at which you work. It can also be worthwhile to work through problems that have been done before, as this helps to reinforce familiarity, understanding and confidence. Working many problems will also help in being able to more quickly identify topic and question types. I have attempted, wherever possible, to emphasize shortcuts and efficient and systematic ways of setting up solutions. There are also occasional comments on interpretation of the language used in some exam questions. While the focus of the study guide is on exam preparation, from time to time there will be comments on underlying theory in places that I feel those comments may provide useful insight into a topic.

The notes and examples are divided into sections anywhere from 4 to 14 pages, with suggested time frames for covering the material. There are over 325 examples in the notes and about 850 exercises in the problem sets, all with detailed solutions. The 13 practice exams have 40 questions each, also with detailed solutions. Some of the examples and exercises are taken from previous SOA/CAS exams. Questions in the problem sets that have come from previous SOA/CAS exams are identified as such. Some of the problem set exercises are more in depth than actual exam questions, but the practice exam questions have been created in an attempt to replicate the level of depth and difficulty of actual exam questions. In total there are over 1700 examples/problems/sample exam questions with detailed solutions. ACTEX gratefully acknowledges the SOA and CAS for allowing the use of their exam problems in this study guide.

I suggest that you work through the study guide by studying a section of notes and then attempting the exercises in the problem set that follows that section. My suggested order for covering topics is

- (1) modeling , (2) model estimation , (Volume 1) ,
- (3) credibility theory , and (4) simulation (includes stock price models and risk measures) (Volume 2).

It has been my intention to make this study guide self-contained and comprehensive for all Exam C topics, but there are occasional references to the Loss Models reference book listed in the SOA/CAS catalog. While the ability to derive formulas used on the exam is usually not the focus of an exam question, it is useful in enhancing the understanding of the material and may be helpful in memorizing formulas. There may be an occasional reference in the review notes to a derivation, but you are encouraged to review the official reference material for more detail on formula derivations. In order for the review notes in this study guide to be most effective, you should have some background at the junior or senior college level in probability and statistics. It will be assumed that you are reasonably familiar with differential and integral calculus. The prerequisite concepts to modeling and model estimation are reviewed in this study guide. The study guide begins with a detailed review of probability distribution concepts such as distribution function, hazard rate, expectation and variance.

Of the various calculators that are allowed for use on the exam, I am most familiar with the BA II PLUS. It has several easily accessible memories. The TI-30X IIS has the advantage of a multi-line display. Both have the functionality needed for the exam.

There is a set of tables that has been provided with the exam in past sittings. These tables consist of some detailed description of a number of probability distributions along with tables for the standard normal and chi-squared distributions. The tables can be downloaded from the SOA website www.soa.org.

If you have any questions, comments, criticisms or compliments regarding this study guide, please contact the publisher ACTEX, or you may contact me directly at the address below. I apologize in advance for any errors, typographical or otherwise, that you might find, and it would be greatly appreciated if you would bring them to my attention. ACTEX will be maintaining a website for errata that can be accessed from www.actexamdriver.com.

It is my sincere hope that you find this study guide helpful and useful in your preparation for the exam. I wish you the best of luck on the exam.

Samuel A. Broverman
Department of Statistics
University of Toronto

October, 2008
www.sambroverman.com
E-mail: sam@utstat.toronto.edu or 2brove@rogers.com

MODEL ESTIMATION - SECTION 4 - KERNEL SMOOTHING

The material in this section relates to Section 14.3 of "Loss Models". The suggested time frame for this section is 3 hours.

ME-4.1 Definition of Kernel Density Estimator

We continue to assume that data is in the form of complete individual data. This means that we have a random sample of observations (of loss amounts, or of times of death) and we know the value of each observation (and there may be some repeated values) with no censoring or truncation of data.

Our objective with kernel smoothing is to create a density function that will in some way approximate the (discrete) empirical distribution. We are trying to create a continuous random variable (whose density function will be called the kernel smoothed density estimator that we are finding) that is an approximation to the discrete empirical distribution. The method simultaneously constructs an estimate of the density function called the **kernel density estimator of the density function** and an estimate of the distribution function called the **kernel density estimator of the distribution function**.

There are a variety of **kernels** that can be used to construct the estimator. Each kernel results in its own kernel density estimator. The kernel is itself a density function that is used in the smoothing procedure. The "Loss Models" book mentions three possible kernels (uniform, triangle and gamma), but the density function of any random variable can be used as a kernel. Essentially what is being done when kernel smoothing is applied to estimate a density function is that at each point y_i in the empirical distribution, a density function corresponding to that point is created, and this density function is denoted $k_{y_i}(x)$. For each y_i , $k_{y_i}(x)$ is an actual pdf, and satisfies the requirements of a pdf. The kernel smoothed density estimator is then a finite mixture (or weighted average) of these separate density functions. The mixing "weight" applied to $k_{y_i}(x)$ is the empirical probability $p(y_i)$, and the kernel smoothed estimate of the density function is

$$\hat{f}(x) = \sum_{\text{All } y_j} p(y_j) \cdot k_{y_j}(x) . \quad (4.1)$$

Once we have identified the empirical distribution points (the sample value y_i 's) and their empirical probabilities ($p(y_i)$ for each y_i), we choose which kernel pdf $k_{y_i}(x)$ we will use. Each kernel density function $k_{y_i}(x)$ has a corresponding distribution function $K_{y_i}(x)$. The kernel smoothed estimate of the distribution function is $\hat{F}(x) = \sum_{\text{All } y_j} p(y_j) \cdot K_{y_j}(x)$,

the same sort of "weighted average" mixture formulation that we have for the density estimator $\hat{f}(x)$.

The simple example we will first consider has the following 4-point random sample: $y_1 = 1$, $y_2 = 2$, $y_3 = 4$, $y_4 = 7$. The empirical distribution assigns a probability of .25 to each of these points, so that $p(1) = p(2) = p(4) = p(7) = \frac{1}{4}$. We will apply uniform kernel, triangle kernel and gamma kernel to this data set to show the construction and properties of the kernel density and distribution function estimator.

According to the definition of $\hat{f}(x)$, the way in which we calculate $\hat{f}(x)$, is to apply the formula

$$\hat{f}(x) = \sum_{\text{All } y_j} p(y_j) \cdot k_{y_j}(x) . \text{ For our example, we have } y_1 = 1, y_2 = 2, y_3 = 4, y_4 = 7,$$

and $p(1) = p(2) = p(4) = p(7) = .25$. Then $\hat{f}(x) = (.25)[k_1(x) + k_2(x) + k_4(x) + k_7(x)]$, where the $k_y(x)$ functions are the kernel density functions, and

$$\hat{F}(x) = (.25)[K_1(x) + K_2(x) + K_4(x) + K_7(x)] .$$

Note that the subscript of k and K is the y_i -value. For instance, $k_4(x)$ is the kernel function associated with the 3rd y -value, $y_3 = 4$ ($k_4(x)$ is not the 4-th kernel pdf).

ME-4.2 Uniform Kernel Estimator

Uniform kernel density estimator $\hat{f}(x)$ with bandwidth b

One of the kernels introduced in the "Loss Models" book is the **uniform kernel with bandwidth b** . The uniform kernel is based on the continuous uniform distribution. Recall that

the uniform distribution on the interval $[c, d]$ has pdf $k(x) = \begin{cases} \frac{1}{d-c} & \text{for } c \leq x \leq d \\ 0 & \text{otherwise} \end{cases}$. (4.3)

For the uniform kernel with bandwidth b , at each sample point y_i , the kernel density $k_{y_i}(x)$ is the density function for the uniform distribution on the interval $[y_i - b, y_i + b]$, so that

$$k_{y_i}(x) = \begin{cases} \frac{1}{2b} & \text{for } y_i - b \leq x \leq y_i + b \\ 0 & \text{otherwise} \end{cases} . \quad (4.4)$$

The graph of $k_{y_i}(x)$ is a horizontal line of height $\frac{1}{2b}$ on the interval $[y_i - b, y_i + b]$, and it is 0 outside that interval; the graph is a rectangle with an area of 1 (since $k_{y_i}(x)$ is a pdf, total area must be 1). Each y_i has its own associated kernel function $k_{y_i}(x)$, which is a uniform density on an interval centered at y_i and the interval length is $2b$.

We will illustrate this method by applying the uniform kernel with a bandwidth of .4 (a somewhat arbitrary choice). At each of the original data points we create a rectangle, with the sample data point value at the center of the base of the rectangle, and with the area of the rectangle being 1.

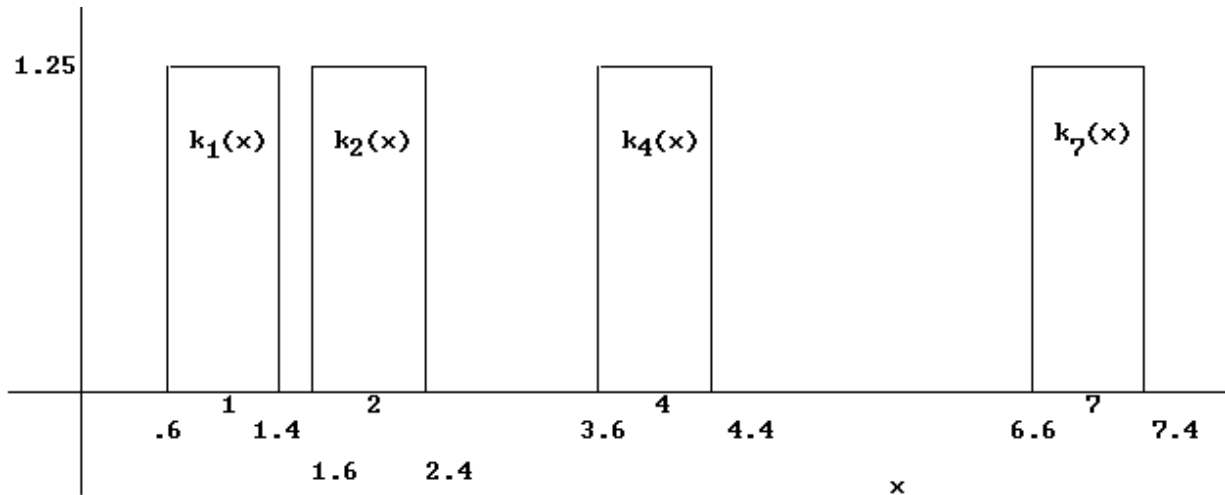
For sample data point y_i in the original random sample, we create a rectangle whose base is from $y_i - b$ to $y_i + b$, and whose height is chosen so that the area of the rectangle is 1. Since the base is $2b$, the height must be $\frac{1}{2b}$. With our chosen value of $b = .4$, the rectangles will all have height

$\frac{1}{2(.4)} = 1.25$, and there will be four rectangles with the following bases,

$$[.6, 1.4], [1.6, 2.4], [3.6, 4.4], [6.6, 7.4] .$$

The notation $k_1(x)$, $k_2(x)$, $k_4(x)$ and $k_7(x)$ describes the four "horizontal-line" functions represented in the graph below (note that the subscript of k is the y -value for the k -th interval). For instance, $k_1(x) = 1.25$ for $.6 \leq x \leq 1.4$, and $k_1(x) = 0$ for any x outside the interval $[.6, 1.4]$. Similar conditions apply to the other three rectangles. The subscript to k is the value of the data point from the original random sample. This identifies which rectangle is being considered. Note that for each sample data point y , $k_y(x)$ is the pdf of the uniform distribution on the interval from $y - b$ to $y + b$.

What we have created is four separate uniform distributions, one for each interval. The graph of these four rectangles is as follows.



The way in which we apply the formula $\hat{f}(x) = \sum_{\text{All } y_j} p(y_j) \cdot k_{y_j}(x)$ is as follows.

Given a value of x , in order to find $\hat{f}(x)$, we first determine which rectangle bases contain x . We only need to know the rectangles for which x is in the base because $k_y(x) = 0$ for values of x that are outside of the base rectangle around y . We then find the $k(x)$ value for each rectangle and multiply by the empirical probability for that rectangle's base center point.

For instance, suppose we wish to find the kernel density estimator at $x = 1.1$, i.e., we wish to find $\hat{f}(1.1)$. $\hat{f}(1.1)$ is found by first identifying which rectangle bases contain the value 1.1. We see that 1.1 is in the interval $[.6, 1.4]$, so only the kernel function $k_1(x)$ will be non-zero in calculating $\hat{f}(1.1)$ ($k_2(1.1) = 0$ since 1.1 is not in the interval $[1.6, 2.4]$ centered at $y_2 = 2$, and the same applies to $y_3 = 4$ and $y_4 = 7$).

We find $\hat{f}(1.1)$ from $\hat{f}(1.1) = \sum_{\text{All } y_j} p(y_j) \cdot k_{y_j}(1.1)$, which is
 $p(1) \cdot k_1(1.1) + p(2) \cdot k_2(1.1) + p(4) \cdot k_4(1.1) + p(7) \cdot k_7(1.1)$
 $= (.25)(1.25) + (.25)(0) + (.25)(0) + (.25)(0) = .3125$.

Again, it is important to note that we only used $k_1(1.1)$ since the value $x = 1.1$ was only in the first of the four intervals for the rectangle bases, so $k_1(1.1) = 1.25$, but $k_2(1.1) = 0$ and $k_4(1.1) = 0$ and $k_7(1.1) = 0$.

Suppose we now consider the x -value $x = 3.5$ and we wish to find $\hat{f}(3.5)$, the kernel density estimator at $x = 3.5$. We see that $x = 3.5$ is not in any of the four intervals formed by the bases of the four rectangles. Therefore, $\hat{f}(3.5) = 0$, since $k_{y_i}(3.5) = 0$ for each sample data point y_i .

Note that in the simple example we are now considering, since any x is in either one rectangle base or none, $\hat{f}(x)$ will be .3125 if x is in one of the four rectangle bases, and $\hat{f}(x) = 0$ if x is not in any of the four rectangle bases. If we were to draw the graph of this $\hat{f}(x)$ it would look the same as the four rectangles in the graph above, but the heights would be .3125 instead of 1 for each rectangle.

If the rectangle bases are wider (if the bandwidth is increased), some bases may overlap and some x 's will be in two or more rectangle bases. If that is the case, then for that x , in the relationship

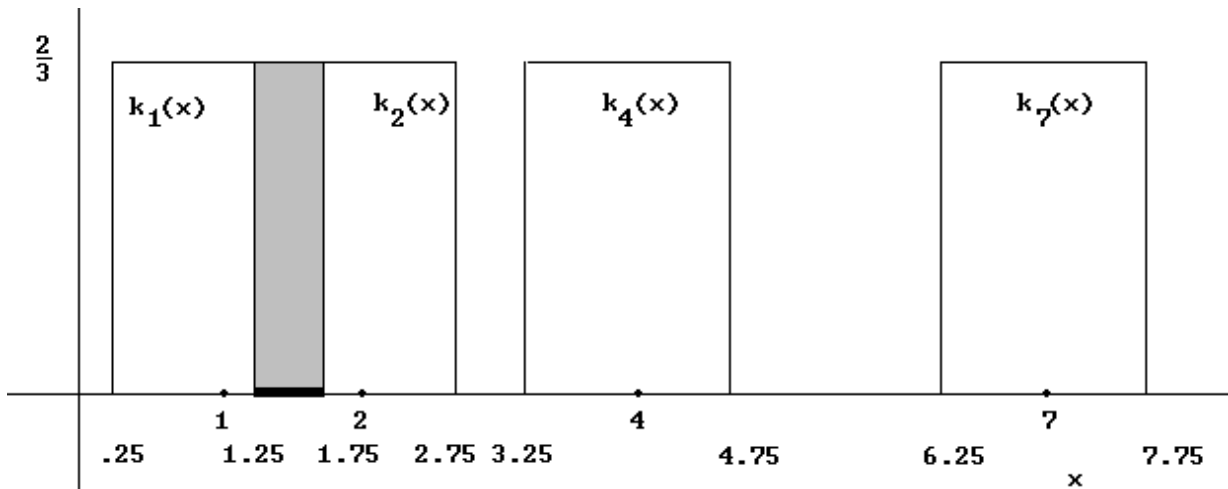
$$\hat{f}(x) = \sum_{\text{All } y_j} p(y_j) \cdot k_{y_j}(x), \text{ more than one } k_{y_j}(x) \text{ will be non-zero.}$$

The following variation on the example considers this.

Suppose we start the example over with a bandwidth of $b = .75$. There will still be four rectangles, but the bases will now be $[.25, 1.75] , [1.25, 2.75] , [3.25, 4.75] , [6.25, 7.75]$. The rectangle height for each rectangle will be $\frac{1}{2b} = \frac{1}{2(.75)} = \frac{2}{3}$. Therefore,

$$k_1(x) = \frac{2}{3} \text{ for } .25 \leq x \leq 1.75, \text{ and } k_1(x) = 0 \text{ if } x \text{ is not in the interval } [.25, 1.75].$$

Similar relationships apply for the other three rectangles. The graph of the rectangles is shown below. The darkened horizontal line segment represents the region where two rectangle bases intersect. Any x between 1.25 and 1.75 is in both of the first two rectangles on the left. The intersection of the two rectangles is also lightly shaded. The vertical scale of the graph has been changed from that of the previous graph.



As before, in order to find $\hat{f}(x)$ for a particular value of x we must calculate

$$\hat{f}(x) = \sum_{\text{All } y_j} p(y_j) \cdot k_{y_j}(x) = (.25)[k_1(x) + k_2(x) + k_4(x) + k_7(x)], \text{ using the kernel functions}$$

defined for our new bandwidth $b = .75$. Again, given a value of x , we identify in which rectangle bases x lies. For instance, for $x = 5.2$, we see that x doesn't lie in any rectangle base. Therefore, $\hat{f}(5.2) = 0$ is the kernel density estimator at $x = 5.2$, because $k_{y_j}(5.2) = 0$ for each y_j . The result will be the same for any x not contained in any rectangle base.

Suppose that $x = 1.1$. This x is only in the interval $[.25, 1.75]$, and not in any other rectangle bases. Then $\hat{f}(1.1) = (.25)(\frac{2}{3}) = \frac{1}{6}$; the rectangle containing $x = 1.1$ corresponds to the rectangle centered at the original sample value of 1, so we multiply the empirical probability value at 1 (this is .25) by $k_1(1.1)$ (which is $\frac{2}{3}$).

Note that $\hat{f}(x) = \frac{1}{6}$ for x in any of the following intervals:

$$.25 \leq x < 1.25, \quad 1.75 < x \leq 2.75, \quad 3.25 \leq x \leq 4.75 \text{ or } 6.25 \leq x \leq 7.75.$$

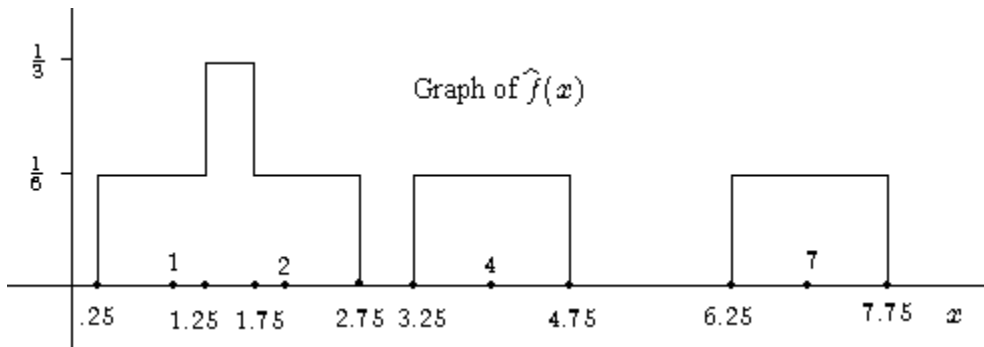
This is because for x in any of those regions, x is in exactly one rectangle base.

Suppose that $x = 1.4$. Then x is in two rectangle bases, those being $[.25, 1.75]$ and $[1.25, 2.75]$. In order to find $\hat{f}(1.4)$ we must include a factor for each rectangle base that contains x .

$$\hat{f}(1.4) = (.25)[k_1(1.4) + k_2(1.4) + k_4(1.4) + k_7(1.4)] = (.25)\left[\frac{2}{3} + \frac{2}{3} + 0 + 0\right] = \frac{1}{3}.$$

$\hat{f}(x) = \frac{1}{3}$ for any x in the interval $1.25 \leq x \leq 1.75$ because those x 's are in two intervals.

The complete graph of the kernel smoothed density estimator based on bandwidth .75 is illustrated below. It is found by combining the heights of the rectangles in any intervals for which they bases overlap. The following is the graph of the uniform kernel smoothed density estimator with bandwidth .75. As the bandwidth gets wider there will be more intersection regions and some x 's may be in several rectangle bases.



One other point to note about the uniform kernel is that if x is an interval endpoint, either $y - b$ or $y + b$, then $k_y(x) = \frac{1}{2b}$. $k_y(x) = 0$ for x values outside the closed interval $[y - b, y + b]$.

Uniform kernel estimator of the distribution function, $\hat{F}(x)$, with bandwidth b

We can apply kernel density estimation to estimate the distribution function $F(x) = P[X \leq x]$.

The algebraic expression for the kernel estimator of the distribution function is

$$\hat{F}(x) = \sum_{\text{All } y_j} p(y_j) \cdot K_{y_j}(x).$$

For specific values of x and y_j , $K_{y_j}(x)$ is the cdf for the kernel pdf $k_{y_j}(x)$; $K_{y_j}(x)$ is the probability to the left of x for the kernel distribution centered at y_j .

For the uniform kernel with bandwidth b , the formal definition of $K_{y_j}(x)$ is

$$K_y(x) = \begin{cases} 0 & x < y - b \\ \frac{x - y + b}{2b} & y - b \leq x \leq y + b \\ 1 & x > y + b \end{cases} . \tag{4.5}$$

Note that $x > y + b$ means that the rectangle base interval around y is completely to the left of x (less than x), so the full rectangle area of 1 is used ($K_y(x) = 1$), and $x < y - b$ means that the rectangle base area around y is completely to the right of x so that the interval can be ignored ($K_y(x) = 0$); see the graphs below illustrating these points.

In order to find $\hat{F}(x)$ for a particular value of x , we must determine which rectangle base intervals are completely to the left of x , which are completely to the right of x , and which contain x . $\hat{F}(x)$ will be a sum of (possibly) several $p(y_j) \cdot K_{y_j}(x)$ factors. What we trying to do is add up the probability in the kernel density that is to the left of x .

For any rectangle base interval completely to the right of x , we have $K_{y_j}(x) = 0$ and that term in $\hat{F}(x)$ can be ignored. This will occur if $x \leq y_j - b$, or equivalently, if $x + b \leq y_j$.

If a rectangle base interval is completely to the left of x , and if that rectangle base is centered at the random sample point y_j , then $K_{y_j}(x) = 1$. This will occur if $y_j + b \leq x$, or equivalently, if $y_j \geq x - b$.

If x is inside the rectangle base interval for the random sample point y_j then

$$K_{y_j}(x) = [x - (y_j - b)]\left(\frac{1}{2b}\right) = [(x + b) - y_j]\left(\frac{1}{2b}\right). \quad (4.6)$$

Note that $x - (y_j - b) = (x + b) - y_j$ is the part of the base of the rectangle centered at y_j that is to the left of x . This is illustrated in the graphs below.

Once we have identified the value of $K_{y_j}(x)$ for each y_j , we can calculate

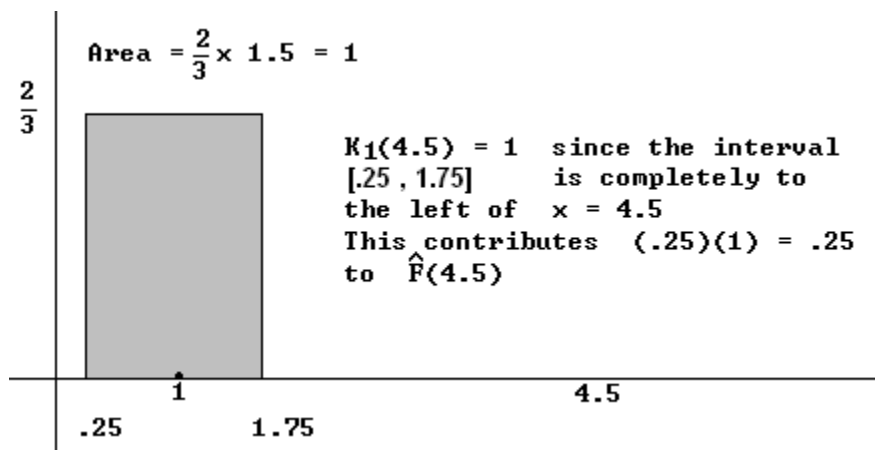
$$\hat{F}(x) = \sum_{\text{All } y_j} p(y_j) \cdot K_{y_j}(x).$$

For instance, suppose that we consider the bandwidth $b = .75$ example above, and we wish to find $\hat{F}(4.5)$. We see (in the graphs below) that the rectangle base intervals $[.25, 1.75]$ and $[1.25, 2.75]$ are both completely to the left of $x = 4.5$, so $K_1(4.5) = 1$ and $K_2(4.5) = 1$. The point $x = 4.5$ is inside the rectangle base interval $[3.25, 4.75]$, and the area in that rectangle to the left of $x = 4.5$ is $K_4(4.5) = [x - (y_j - b)]\left(\frac{1}{2b}\right) = (4.5 - 3.25)\left(\frac{2}{3}\right) = .8333$; note that we can also write this as $[(x + b) - y_j]\left(\frac{1}{2b}\right) = [4.5 + .75 - 4]\left(\frac{2}{3}\right)$.

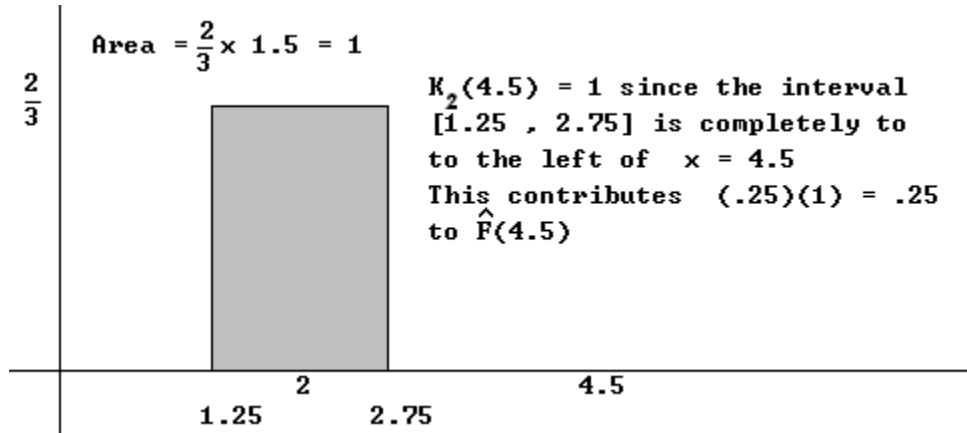
Finally, $x = 4.5$ is completely to the left of the rectangle base interval $[6.25, 7.25]$, so that $K_7(4.5) = 0$. Therefore, $\hat{F}(4.5) = (.25)(1) + (.25)(1) + (.25)(.8333) + (.25)(0) = .7083$.

The following graphs indicate the contribution of each rectangle to the total in $\hat{F}(4.5)$.

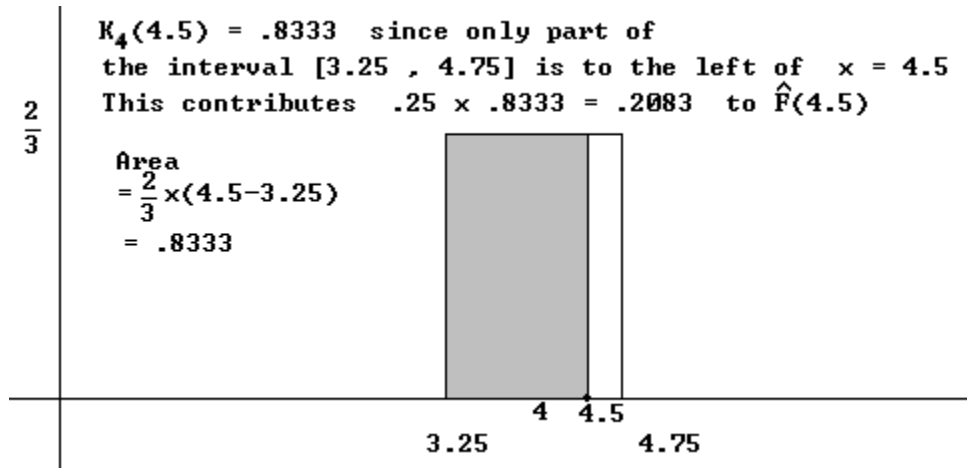
Contribution to $\hat{F}(4.5)$ from $y_1 = 1$.



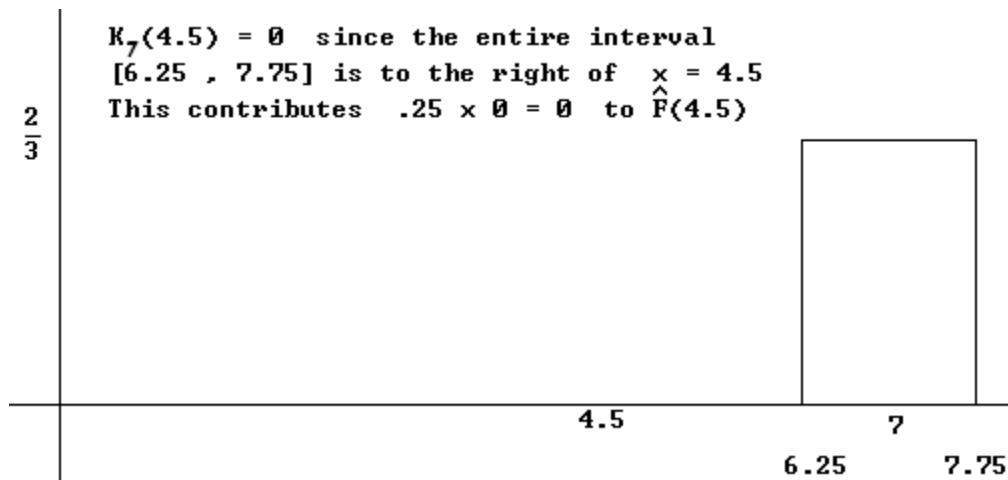
Contribution to $\hat{F}(4.5)$ from $y_2 = 2$.



Contribution to $\hat{F}(4.5)$ from $y_3 = 4$.



Contribution to $\hat{F}(4.5)$ from $y_4 = 7$.



As another example, suppose that we wish to find $\hat{F}(1.4)$.

To find $\hat{F}(1.4)$ we note that intervals $[3.25, 4.75]$ and $[6.25, 7.75]$ are both completely to the right, so they contribute nothing. We also see that $x = 1.4$ is in the interval $[-.25, 1.75]$, so that

$$K_1(1.4) = (1.4 - .25)\left(\frac{2}{3}\right) = (1.4 + .75 - 1)\left(\frac{2}{3}\right) = .7667.$$

We see that $x = 1.4$ is in the interval $[1.25, 2.75]$, so that

$$K_2(1.4) = (1.4 - 1.25)\left(\frac{2}{3}\right) = (1.4 + .75 - 2)\left(\frac{2}{3}\right) = .100.$$

Therefore,

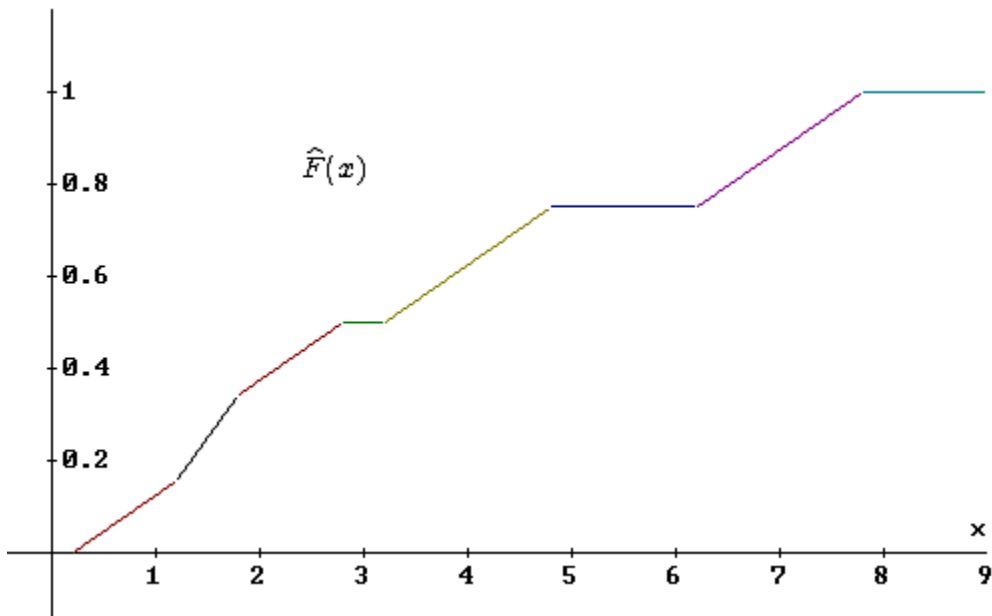
$$\begin{aligned}\hat{F}(1.4) &= p(1)K_1(1.4) + p(2)K_2(1.4) + p(4)K_4(1.4) + p(7)K_7(1.4) \\ &= (.25)(.7667) + (.25)(.100) + (.25)(0) + (.25)(0) = .2167.\end{aligned}$$

For the interval $y_j - b \leq x \leq y_j + b$, $K_{y_j}(x) = \frac{x - y_j + b}{2b}$, a straight line rising from 0 to 1 on the interval. If we carefully identify $K_{y_j}(x)$ on each interval, we can plot the graph of $\hat{F}(x)$.

Using the example with bandwidth $b = .8$, we have

$$\hat{F}(x) = \begin{cases} 0 & x < .2 \\ (.25)\left(\frac{x-1+.8}{1.6}\right) = \frac{x-.2}{6.4} & .2 \leq x \leq 1.2 \\ (.25)\left(\frac{x-1+.8}{1.6} + \frac{x-2+.8}{1.6}\right) = \frac{x-.7}{3.2} & 1.2 \leq x \leq 1.8 \\ (.25)\left(1 + \frac{x-2+.8}{1.6}\right) = \frac{x+.4}{6.4} & 1.8 \leq x \leq 2.8 \\ (.25)(1 + 1) = .5 & 2.8 \leq x \leq 3.2 \\ (.25)\left(1 + 1 + \frac{x-4+.8}{1.6}\right) = \frac{x}{6.4} & 3.2 \leq x \leq 4.8 \\ (.25)(1 + 1 + 1) = .75 & 4.8 \leq x \leq 6.2 \\ (.25)\left(1 + 1 + 1 + \frac{x-7+.8}{1.6}\right) = \frac{x-1.4}{6.4} & 6.2 \leq x \leq 7.8 \\ (.25)(1 + 1 + 1 + 1) = 1 & x \geq 7.8 \end{cases}$$

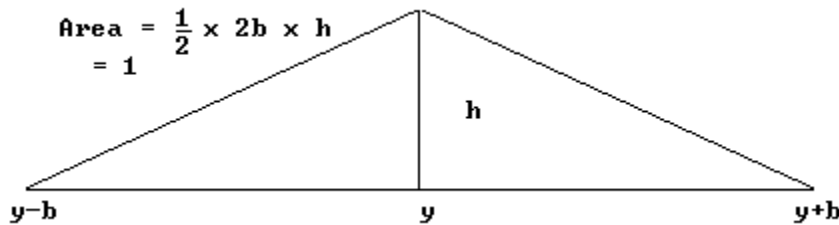
The graph of $\hat{F}(x)$ is a series of line segments. The slope changes whenever x crosses over an interval endpoint, such as .2, 1.2, 1.8, etc.



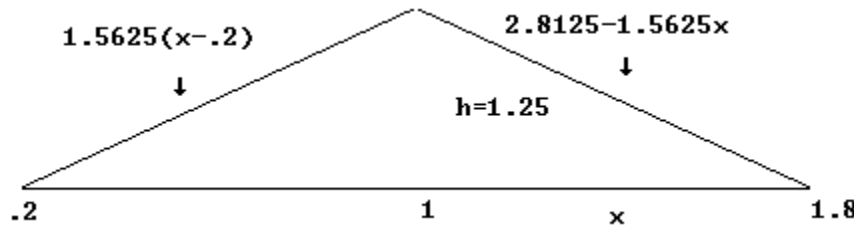
ME-4.3 The Triangle Kernel Estimator

Triangle kernel density estimator $\hat{f}(x)$ with bandwidth b

The textbook mentions kernels other than the uniform. One of them is the **triangle kernel**. The procedure is similar to that of the uniform kernel, the difference being that the rectangle constructed around each y_j is replaced by an isosceles triangle which has y_j at the center of the base. In the uniform kernel approach each rectangle had area 1. We also want each triangle to have area 1 (we are creating a density function for each y_i , and the total probability must be 1 for any density). For the triangle related to random sample point y , with bandwidth b , the triangle base will be from $y - b$ to $y + b$. In order for the triangle to have area 1, the triangle peak at the base midpoint y has height $h = \frac{1}{b}$. This can be seen in the following diagram.



We apply the triangle kernel to the random sample 1, 2, 4, 7 used above, but this time with bandwidth $b = .8$. Each triangle will peak at the middle with a height of $\frac{1}{.8} = 1.25$. The triangle for the random sample point $y = 1$ will be of the form

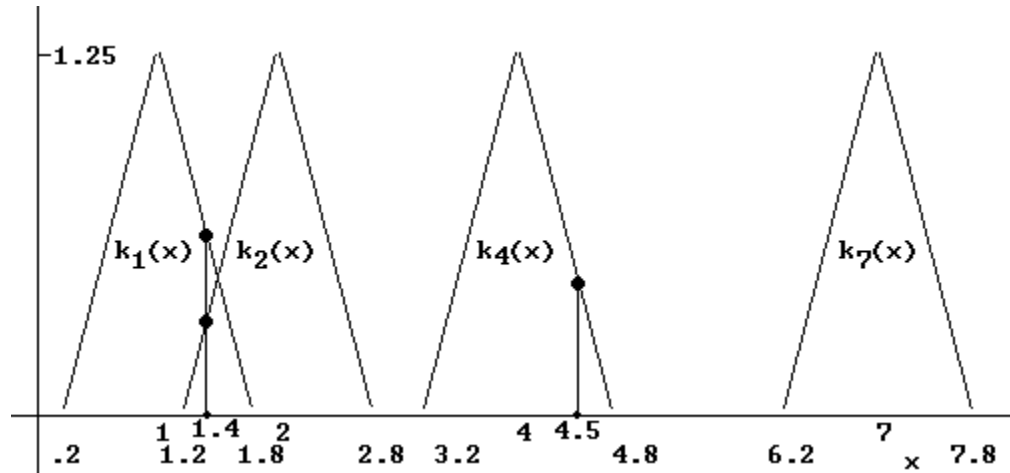


The equations of the line segments on the two sides of the triangle were found using the two-point method of finding the equation of a straight line. It is not necessary to actually write out the explicit equation form of the line for the two upper sides of the triangle. As will be seen, for actual calculations, the proportionality of the triangle can be used.

The kernel function for this triangle is made up of two components (line segments), $k_1(x) = 1.5625(x - .2)$ for $.2 \leq x \leq 1$, and $k_1(x) = 2.8125 - 1.5625x$ for $1 \leq x \leq 1.8$. For instance, $k_1(.7) = 1.5625(.7 - .2) = .78125$, and $k_1(1.4) = 2.8125 - 1.5625(1.4) = .625$.

These values could also be calculated by using a "similar triangles" approach. For instance, $x = 1.4$ is half-way from $x = 1$ to $x = 1.8$, so $k_1(1.4)$ is half-way from the triangle peak height of 1.25 to 0, i.e., .625. Also, $k_1(x) = 0$ for any x outside the interval $[.2, 1.8]$.

We construct triangle kernel functions for each of the original y -values in the random sample. The graphs of all the triangle kernel functions is as follows.



Suppose we wish to find $\hat{f}(4.5)$. As in the uniform kernel approach, we first find which triangle base intervals contain the point $x = 4.5$. We see that only the base interval for $k_4(x)$, which is $[3.2, 4.8]$ contains $x = 4.5$. To calculate the smoothed estimate, we have the general relationship defining the kernel density estimate, $\hat{f}(x) = \sum_{\text{All } y_j} p(y_j) \cdot k_{y_j}(x)$.

In this example $k_1(4.5) = k_2(4.5) = k_7(4.5) = 0$ (since $x = 4.5$ is not in the triangle base interval for those kernel functions). Therefore, $\hat{f}(4.5) = p(4) \cdot k_4(4.5)$. From the empirical distribution we know that $p(4) = .25$. Since $x = 4.5$ is $\frac{5}{8}$ of the way from $x = 4$ to $x = 4.8$, there is $\frac{3}{8}$ of the way left to go, and since the triangle drops from a height of 1.25 to 0 as x goes from 4 to 4.8, we see that $k_4(4.5) = \frac{3}{8} \times 1.25 = .46875$. Then $\hat{f}(4.5) = (.25)(.46875) = .1171875$.

Suppose that we wish to find $\hat{f}(1.4)$. We see that $x = 1.4$ lies in two triangle base intervals, for $k_1(x)$ on $[.2, 1.8]$ and for $k_2(x)$ on $[1.2, 2.8]$. Then $\hat{f}(1.4) = .25 \times k_1(1.4) + .25 \times k_2(1.4)$ (there is no contribution from k_4 or k_7 since $x = 1.4$ is not in the corresponding intervals).

Since $x = 1.4$ is half-way between $x = 1$ and $x = 1.8$ (the base interval for k_1), we see from the geometry of the diagram above that $k_1(1.4) = \frac{1}{2} \times 1.25 = .625$ (the upper "dot" in the diagram above). Since $x = 1.4$ is one-quarter of the way from $x = 1.2$ to $x = 2$ (the base interval for k_2), we see that $k_2(1.4) = \frac{1}{4} \times 1.25 = .3125$ (the lower "dot" above). Then $\hat{f}(1.4) = (.25)(.625) + (.25)(.3125) = .234375$.

We could have set up the algebraic form of the line segments in the various triangles.

For instance, $k_1(x) = 2.8125 - 1.5625x$ for $1 \leq x \leq 1.8$, so that

$k_1(1.4) = 2.8125 - 1.5625(1.4) = .625$ (as we have already seen).

Also, $k_2(x) = 1.5625(x - 1.2)$ for $1.2 \leq x \leq 2$, so that $k_2(1.4) = 1.5625(.2) = .3125$.

The algebraic form of the triangle kernel is $k_y(x) = \begin{cases} 0 & x < y - b \\ \frac{b+x-y}{b^2} & y - b \leq x \leq y \\ \frac{b+y-x}{b^2} & y \leq x \leq y + b \\ 0 & x > y + b \end{cases}$. (4.7)

For instance, in the example just considered,

$$k_1(x) = \begin{cases} 0 & x < 1 - .8 = .2 \\ \frac{.8+x-1}{.8^2} = 1.5625(x - .2) & 1 - .8 = .2 \leq x \leq 1 \\ \frac{.8+1-x}{.8^2} = 2.8125 - 1.5625x & 1 \leq x \leq 1.8 = 1 + .8 \\ 0 & x > 1.8 = 1 + .8 \end{cases}$$
. (4.8)

This gives the algebraic form of the two sides of the triangle for $k_1(x)$ in the graph above. The other kernel functions can be formulated in a similar way. It is usually unnecessary to formulate the algebraic line equations an exam question.

Triangle kernel estimator of the distribution function, $\widehat{F}(x)$, with bandwidth b

It is also possible to find the kernel density estimator of the distribution function using the

triangle kernel. We can use the form $K_y(x) = \begin{cases} 0 & x < y - b \\ \frac{(b+x-y)^2}{2b^2} & y - b \leq x \leq y \\ 1 - \frac{(b+y-x)^2}{2b^2} & y \leq x \leq y + b \\ 1 & x > y + b \end{cases}$, (4.9)

and, as before, $\widehat{F}(x) = \sum_{\text{All } y_j} p(y_j) \cdot K_{y_j}(x)$.

Again, what we are really doing is finding the area to left of x in each triangle, and we multiply this area by $p(y)$ for that rectangle.

For instance, the kernel smoothed estimate of $F(4.5)$ would be

$$\widehat{F}(4.5) = p(1)K_1(4.5) + p(2)K_2(4.5) + p(4)K_4(4.5) + p(7)K_7(4.5).$$

Since the triangle base centered at $y_1 = 1$ is completely to the left of 4.5, we have $K_1(4.5) = 1$, and the same is true for $y_2 = 2$, so that $K_2(4.5) = 1$. Since the triangle base centered at $y_4 = 7$ is completely to the right of 4.5, we have $K_7(4.5) = 0$. $x = 4.5$ is inside the interval centered at $y_3 = 4$. The area to the left of 4.5 in the triangle centered at $y_3 = 4$ is

$$K_4(4.5) = 1 - (\text{area to the right of 4.5 in the triangle centered at } y_3 = 4).$$

But [area to the right of 4.5 in the triangle centered at $y_3 = 4$] is equal to

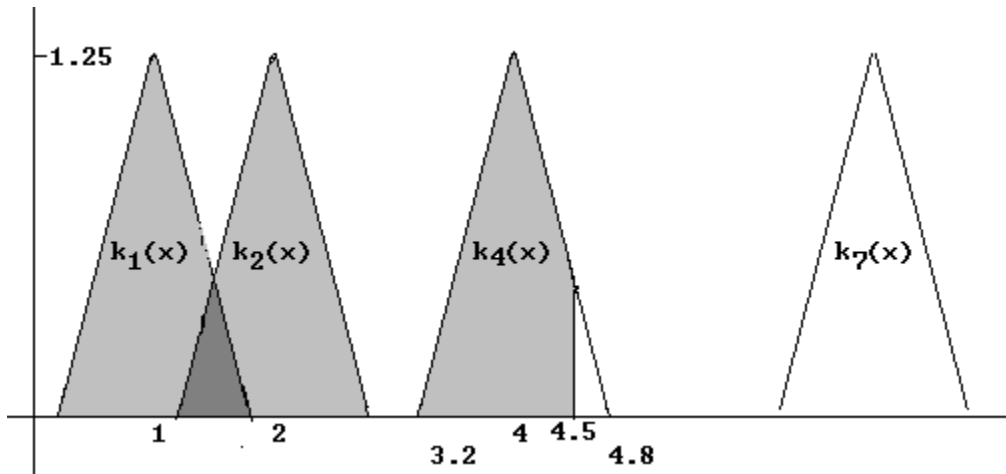
$$\frac{1}{2} \times k_4(4.5) \times (4.8 - 4.5) = \frac{1}{2} \times (.46875) \times (4.8 - 4.5) = .0703125.$$

Therefore, $K_4(4.5) = 1 - .0703125 = .9296875$.

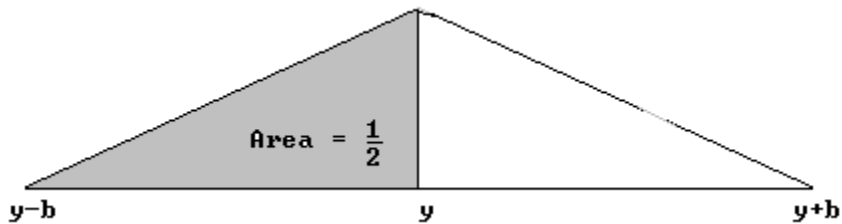
Alternatively, using Equation 4.9, we have $K_4(4.5) = 1 - \frac{(.8+4-4.5)^2}{2(.8)^2} = .9296875$.

Finally, $\widehat{F}(4.5) = p(1)K_1(4.5) + p(2)K_2(4.5) + p(4)K_4(4.5) + p(7)K_7(4.5)$
 $= (.25)(1) + (.25)(1) + (.25)(.9296875) + (.25)(0) = .7324$.

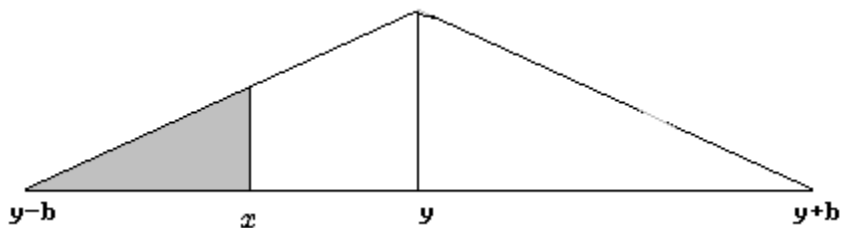
The shaded regions below are the areas that represent $K_y(4.5)$. The extra shading indicates that this region is in both triangles.



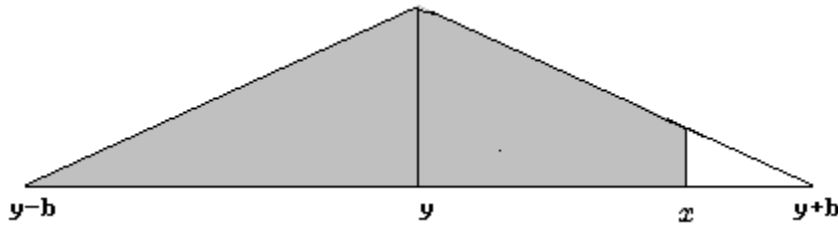
We can find $K_y(x)$ in the triangular kernel case by looking at areas of "sub-triangles". For the triangle centered at data point y with bandwidth b , $K_y(x)$ is the area in the triangle to the left of x . Since y is the midpoint of the bandwidth interval, we have $K_y(y) = \frac{1}{2}$, as shown in the diagram below.



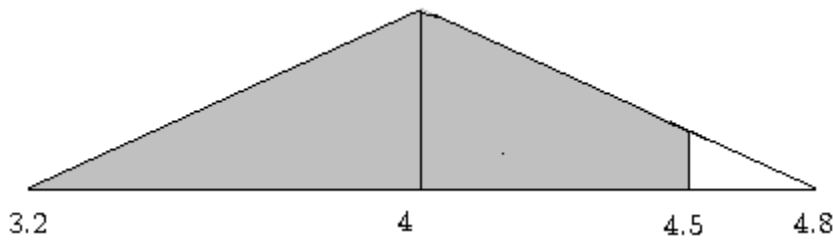
If x is in the left half of the bandwidth interval, $y - b \leq x \leq y$, then the area of the triangle to the left of x is $\frac{(b+x-y)^2}{2b^2}$, shown in the diagram below (from the general formulation of $K_y(x)$ given above). Because of the geometry related to the area of a triangle, we can also describe the area of the triangle to the left of x in the following way. The distance from $y - b$ to x is the fraction $\frac{x-(y-b)}{y}$ of the distance from $y - b$ to y , so the area of the triangle whose base is from $y - b$ to x is the fraction $\left(\frac{x-(y-b)}{b}\right)^2$ of $\frac{1}{2}$ (this is the area of the triangle whose base is from $y - b$ to y). $\frac{1}{2}$ of $\left(\frac{x-(y-b)}{b}\right)^2$ is $\frac{(b+x-y)^2}{2b^2}$ as noted above.



If x is in the right half of the bandwidth interval, then the area to the left of x is the complement of the area to the right of x , from x to the right side of the bandwidth interval $y + b$. The base of the triangle from x to $y + b$ is the fraction $\frac{y+b-x}{b}$ of the base of the triangle from y to $y + b$, so the area of the triangle whose base is from x to $y + b$ is $(\frac{y+b-x}{b})^2 \times \frac{1}{2}$. The area of the shaded region in the triangle below is $K_y(x)$, which is $1 - (\frac{y+b-x}{b})^2 \times \frac{1}{2} = 1 - \frac{(b+y-x)^2}{2b^2}$, described in the general formulation of $K_y(x)$ given above.



Applying this to the numerical example above, we can find $K_4(4.5)$. The bandwidth is $b = .8$, so the left side of the bandwidth interval is $4 - .8 = 3.2$, and the right side is 4.8. The interval from 4.5 to 4.8 is $\frac{.3}{.8} = \frac{3}{8}$ of the interval from 4 to 4.8, so the area of the unshaded triangle is $(\frac{3}{8})^2 \times \frac{1}{2} = .0703125$. The area of the shaded region is $K_4(4.5) = 1 - .0703125 = .9296875$, as noted above.



A kernel smoothing method can be created using any continuous random variable pdf as a kernel function. In the textbook, the gamma distribution is also presented as a possible kernel (the textbook also has an example with a Pareto kernel).

ME-4.4 The Gamma Kernel Estimator

The Gamma kernel with shape parameter α and $\theta = \frac{y}{\alpha}$ has kernel density function

$$k_y(x) = \frac{(\frac{\alpha x}{y})^\alpha e^{-\alpha x/y}}{x\Gamma(\alpha)} \text{ for } x > 0. \tag{4.10}$$

If $\alpha = 1$, then $k_y(x) = \frac{1}{y}e^{-x/y}$, $x > 0$, which is an exponential distribution with mean y .

The Gamma kernel does not require choosing a bandwidth b , but instead requires choosing the shape parameter α . The kernel density estimator of the pdf of X would still be

$$\hat{f}(x) = \sum_{\text{All } y_j} p(y_j) \cdot k_{y_j}(x) \text{ , where the } y_j\text{'s are the original random sample values.}$$

Note that with the gamma kernel $k_{y_j}(x)$ is never 0 for $x > 0$. Also, $\hat{f}(x)$ is a finite mixture of gamma distributions, where the mixing weights are the empirical probabilities $p(y_j)$.

The motivation behind the kernel density estimator is to create a continuous density function that approximates the probabilities assigned in the empirical distribution of a random sample. The graphs on pages 386 to 388 of the textbook illustrate some density functions that result when applying kernel smoothing. The examples given in the textbook also give graphs of some kernel smoothed density functions.

Example ME4-1: For the data of Example ME3-1, apply each of the following three kernels to obtain the kernel smoothed density estimates $\hat{f}(10)$ and $\hat{f}(20)$.

- (1) Uniform kernel with bandwidth 2.
- (2) Triangular kernel with bandwidth 2.
- (3) Gamma kernel with shape parameter $\alpha = 1$.

Solution: The 8 data points are 3, 4, 8, 10, 12, 18, 22, 35, each with empirical probability $\frac{1}{8}$.

(1) Uniform kernel. $\hat{f}(10) = \sum_{j=1}^8 (\frac{1}{8})k_{y_j}(10)$. Since the bandwidth is 2,

$k_{y_j}(10) = 0$ for $y_j = 3, 4, 18, 22$ and 35, and $k_{y_j}(10) = \frac{1}{2b} = \frac{1}{4}$ for $y_j = 8, 10$ and 12.

$\hat{f}(10) = (\frac{1}{8})(\frac{1}{4} + \frac{1}{4} + \frac{1}{4}) = .09375$. Note that when x is the endpoint of a base interval, it is included as part of that interval (so, for instance, for $y_j = 8$ the base interval is $[6, 10]$ and $x = 10$ is regarded as being in the interval).

$\hat{f}(20) = \sum_{j=1}^8 (\frac{1}{8})k_{y_j}(20)$, and $k_{y_j}(20) = 0$ for $y_j = 3, 4, 8, 10, 12$ and 35, and $k_{y_j}(20) = \frac{1}{4}$ for $y_j = 18$ and 22. $\hat{f}(20) = (\frac{1}{8})(\frac{1}{4} + \frac{1}{4}) = .0625$.

(2) Triangular kernel. Same bandwidth as (1) so the 0's from part (1) are still 0's.

$k_8(10) = \frac{2+8-10}{2^2} = 0$ (for the interval centered at 8, the triangle height is 0 at $x = 10$, the right end of the interval), $k_{10}(10) = \frac{2+10-10}{2^2} = \frac{1}{b} = \frac{1}{2}$, $k_{12}(10) = \frac{2+10-12}{2^2} = 0$.

$\hat{f}(10) = (\frac{1}{8})(\frac{1}{2}) = .0625$. $k_{18}(20) = \frac{2+18-20}{2^2} = 0$, $k_{22}(20) = \frac{2+20-22}{2^2} = 0$. $\hat{f}(20) = 0$.

(3) Gamma kernel, $\alpha = 1$, $k_y(x) = \frac{1}{y}e^{-x/y}$ (an exponential distribution).

$k_3(x) = \frac{1}{3}e^{-x/3}$, ..., $k_{35}(x) = \frac{1}{35}e^{-x/35}$.

$\hat{f}(10) = (\frac{1}{8})[\frac{1}{3}e^{-10/3} + \frac{1}{4}e^{-10/4} + \dots + \frac{1}{35}e^{-10/35}] = .0279$.

$\hat{f}(20) = (\frac{1}{8})[\frac{1}{3}e^{-20/3} + \frac{1}{4}e^{-20/4} + \dots + \frac{1}{35}e^{-20/35}] = .0118$. □

MODEL ESTIMATION - PROBLEM SET 4**Kernel Density Estimators**

Questions 1 to 3 are based on the following random sample of 12 data points from a population distribution X : 7, 12, 15, 19, 26, 27, 29, 29, 30, 33, 38, 53

Find the following kernel density estimators.

- Using the uniform kernel with bandwidth 5, find $\hat{f}(20)$, $\hat{F}(20)$, and $\hat{F}(30)$. Plot the graph of $\hat{f}(x)$.
- Using the triangle kernel with bandwidth 3, find $\hat{f}(20)$.
- Using the gamma kernel with $\alpha = 1$, find $\hat{f}(20)$, $\hat{F}(20)$ and $\hat{f}(30)$.
- (SOA) You study five lives to estimate the time from the onset of a disease to death. The times to death are: 2, 3, 3, 3, 7. Using a triangular kernel with bandwidth 2, estimate the density function at 2.5.
A) 8/40 B) 12/40 C) 14/40 D) 16/40 E) 17/40
- (SOA) From a population having distribution function F , you are given the following sample: 2.0, 3.3, 3.3, 4.0, 4.0, 4.7, 4.7, 4.7. Calculate the kernel density estimate of $F(4)$, using the uniform kernel with bandwidth 1.4.
A) .31 B) .41 C) .50 D) .53 E) .63
- (SOA) You are given the following ages at time of death of 10 individuals: 25, 30, 35, 35, 37, 39, 45, 47, 49, 55. Using a uniform kernel with bandwidth $b = 10$, determine the kernel density estimate of the probability of survival to age 40.
A) 0.377 B) 0.400 C) 0.417 D) 0.439 E) 0.485

7. (SOA) You are given the kernel $k_y(x) = \begin{cases} \frac{2}{\pi} \sqrt{1 - (x - y)^2}, & y - 1 \leq x \leq y + 1 \\ 0, & \text{otherwise} \end{cases}$

You are given the following random sample

1 3 3 5

Determine which of the following graphs determines the shape of the kernel density estimator.

(A)



(B)



(C)



(D)



(E)



8. (SOA) You are given:

(i) The sample: 1 2 3 3 3 3 3 3 3 3

(ii) $\hat{F}_1(x)$ is the kernel density estimator of the distribution function using a uniform kernel with bandwidth 1.

(iii) $\hat{F}_2(x)$ is the kernel density estimator of the distribution function using a triangular kernel with bandwidth 1.

Determine which of the following intervals has $\hat{F}_1(x) = \hat{F}_2(x)$ for all x in the interval.

A) $0 < x < 1$ B) $1 < x < 2$ C) $2 < x < 3$ D) $3 < x < 4$

E) None of A), B), C) or D)

MODEL ESTIMATION - PROBLEM SET 4 SOLUTIONS

1. Uniform kernel with bandwidth $b = 5$.

For the point $x = 20$, there are two y_j data values within the band from $20 - 5 = 15$ to $20 + 5 = 25$. These two data values are $y_3 = 15$ and $y_4 = 19$.

Therefore, $k_{y_3}(20) = k_{15}(20) = \frac{1}{2b} = \frac{1}{10}$ and $k_{y_4}(20) = k_{19}(20) = \frac{1}{10}$ and $k_{y_j}(20) = 0$ for all other y_j 's since $x = 20$ is outside the interval $y_j - 5, y_j + 5$ for the other y_j 's. Since there are 12 points in the data set, each point has empirical density of $p(y_j) = \frac{1}{12}$ except for $y_j = 29$, which has empirical density $p(29) = \frac{2}{12}$ since that value occurs twice (there are 11 y -values).

$$\begin{aligned} \text{Then } \hat{f}(20) &= \sum_{j=1}^{11} p(y_j) \cdot k_{y_j}(20) = p(y_3) \cdot k_{y_3}(20) + p(y_4) \cdot k_{y_4}(20) \\ &= p(15) \cdot k_{15}(20) + p(19) \cdot k_{19}(20) = \left(\frac{1}{12}\right)\left(\frac{1}{10}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{10}\right) = \frac{1}{60}. \end{aligned}$$

$$\hat{F}(20) = \sum_{j=1}^{11} p(y_j) \cdot K_{y_j}(20). \text{ The intervals centered at the points } y_1 = 7, y_2 = 12,$$

and $y_3 = 15$ all lie to the left of $x = 20$ so that $K_7(20) = K_{12}(20) = K_{15}(20) = 1$.

The interval centered at $y_4 = 19$ is $[14, 24]$, and $x = 20$ is $\frac{6}{10}$ of the way to the right side, so $K_{19}(20) = \frac{6}{10}$. The intervals centered at the points y_5 to y_{11} (26 to 53) are all completely to the right of $x = 20$, so $K_{y_j}(20) = 0$ for each of those y_j 's. Then,

$$\hat{F}(20) = \sum_{j=1}^{11} p(y_j) \cdot K_{y_j}(20) = \left(\frac{1}{12}\right)(1) + \left(\frac{1}{12}\right)(1) + \left(\frac{1}{12}\right)(1) + \left(\frac{1}{12}\right)\left(\frac{6}{10}\right) = .30.$$

$$\begin{aligned} \hat{F}(30) &= \sum_{j=1}^{11} p(y_j) \cdot K_{y_j}(30) = \left(\frac{1}{12}\right)(1) + \left(\frac{1}{12}\right)(1) + \left(\frac{1}{12}\right)(1) + \left(\frac{1}{12}\right)(1) \\ &+ \left(\frac{1}{12}\right)\left(\frac{9}{10}\right) + \left(\frac{1}{12}\right)\left(\frac{8}{10}\right) + \left(\frac{2}{12}\right)\left(\frac{6}{10}\right) + \left(\frac{1}{12}\right)\left(\frac{5}{10}\right) + \left(\frac{1}{12}\right)\left(\frac{2}{10}\right) + \left(\frac{1}{12}\right)(0) + \left(\frac{1}{12}\right)(0) = .6333. \end{aligned}$$

To plot the graph of $\hat{f}(x)$ we identify the successive interval endpoints of all intervals. The endpoints are

2, 7, 10, 12, 14, 17, 20, 21, 22, 24, 25, 28, 31, 32, 34, 35, 38, 43, 48, 58.

For instance 21 is the left endpoint and 31 is the right endpoint of the interval centered at 26.

For x in successive intervals, we can count the number of y_j -intervals x is in.

For $x < 2$, x is not in any intervals. For $2 \leq x < 7$, x is in 1 interval (the interval $[2, 12]$).

For $7 \leq x < 10$, x is in 2 intervals, etc. Since the sample point 29 occurs twice, its probability is doubled in the empirical distribution. Therefore, for instance, for $25 \leq x < 28$, x is in 5 intervals; those intervals are $[21, 31]$, $[22, 32]$, $[24, 34]$ (twice), and $[25, 35]$.

To plot $\hat{f}(x)$, we apply an empirical probability of $\frac{1}{12}$ to each sample point.

$$\hat{f}(x) = 0 \text{ for } x < 2, \hat{f}(x) = 1 \times \frac{1}{12} = \frac{1}{12} \text{ for } 2 \leq x < 7,$$

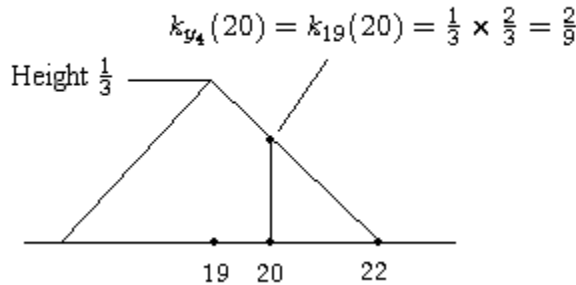
$$\hat{f}(x) = 2 \times \frac{1}{12} = \frac{1}{6} \text{ for } 7 < x < 10, \dots, \hat{f}(x) = 5 \times \frac{1}{12} = \frac{5}{12} \text{ for } 25 \leq x < 33, \dots$$

2. Triangle kernel with bandwidth $b = 3$.

For the point $x = 20$ there is one y_j value within the band from $20 - 3 = 17$ to $20 + 3 = 23$; this is the data value $y_4 = 19$. Similar to the situation in part (a), we have

$k_{y_j}(20) = 0$ for all y_j except for $y_4 = 19$. Using the triangle kernel with $b = 3$, and since $y_4 = 19 \leq 20 \leq 22 = y_4 + b$, we have $k_{y_4}(20) = \frac{b+y-x}{b^2} = \frac{3+19-20}{9} = \frac{2}{9}$

Alternatively, the height of each triangle is $\frac{1}{b} = \frac{1}{3}$, and since $x = 20$ is $\frac{1}{3}$ of the way from the triangle base midpoint at 19 to the right of the triangle base at 22, $k_{y_4}(20)$ is $\frac{2}{3}$ of the height of the triangle, so $k_{y_4}(20) = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9}$. This is illustrated in the graph below.



Then $\hat{f}(20) = \sum_{j=1}^{11} p(y_j) \cdot k_{y_j}(20) = p(y_4) \cdot k_{y_4}(20) = \left(\frac{1}{12}\right)\left(\frac{2}{9}\right) = \frac{1}{54}$.

3. We use the definition $k_y(x) = \frac{(\frac{\alpha x}{y})^\alpha e^{-\alpha x/y}}{x\Gamma(\alpha)}$ for the gamma kernel, and with $\alpha = 1$ we have $k_y(x) = \frac{1}{y} \cdot e^{-x/y}$, which has an exponential distribution with mean y .

Then $\hat{f}(20) = \sum_{j=1}^{11} p(y_j) \cdot k_{y_j}(20)$

$$= \left(\frac{1}{12}\right)\left(\frac{1}{7} \cdot e^{-20/7}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{12} \cdot e^{-20/12}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{15} \cdot e^{-20/15}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{19} \cdot e^{-20/19}\right)$$

$$+ \left(\frac{1}{12}\right)\left(\frac{1}{26} \cdot e^{-20/26}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{27} \cdot e^{-20/27}\right) + \left(\frac{2}{12}\right)\left(\frac{1}{29} \cdot e^{-20/29}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{30} \cdot e^{-20/30}\right)$$

$$+ \left(\frac{1}{12}\right)\left(\frac{1}{33} \cdot e^{-20/33}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{38} \cdot e^{-20/38}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{53} \cdot e^{-20/53}\right) = .016.$$

For the gamma kernel at y , $K_y(x) = \int_0^x k_y(t) dt = \int_0^x \frac{1}{y} \cdot e^{-t/y} dt = 1 - e^{-x/y}$.

$\hat{F}(20) = \sum_{j=1}^{11} p(y_j) \cdot K_{y_j}(20)$

$$= \left(\frac{1}{12}\right)(1 - e^{-20/7}) + \left(\frac{1}{12}\right)(1 - e^{-20/12}) + \left(\frac{1}{12}\right)(1 - e^{-20/15}) + \left(\frac{1}{12}\right)(1 - e^{-20/19})$$

$$+ \left(\frac{1}{12}\right)(1 - e^{-20/26}) + \left(\frac{1}{12}\right)(1 - e^{-20/27}) + \left(\frac{2}{12}\right)(1 - e^{-20/29}) + \left(\frac{1}{12}\right)(1 - e^{-20/30})$$

$$+ \left(\frac{1}{12}\right)(1 - e^{-20/33}) + \left(\frac{1}{12}\right)(1 - e^{-20/38}) + \left(\frac{1}{12}\right)(1 - e^{-20/53}) = .572.$$

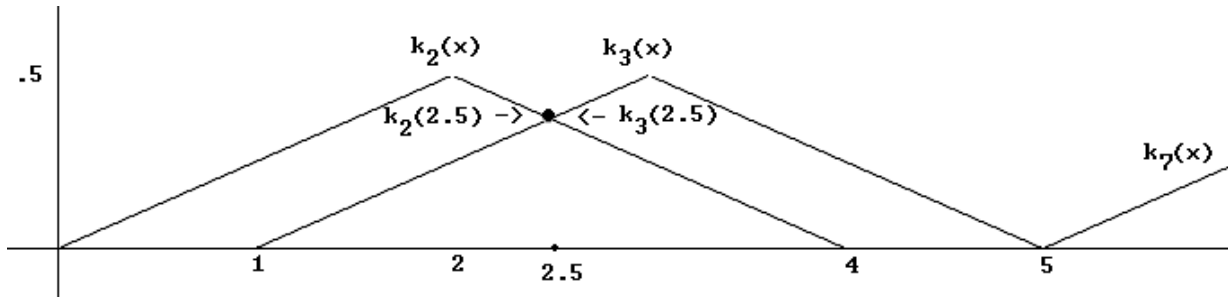
$\hat{f}(30) = \sum_{j=1}^{11} p(y_j) \cdot k_{y_j}(30)$

$$= \left(\frac{1}{12}\right)\left(\frac{1}{7} \cdot e^{-30/7}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{12} \cdot e^{-30/12}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{15} \cdot e^{-30/15}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{19} \cdot e^{-30/19}\right)$$

$$+ \left(\frac{1}{12}\right)\left(\frac{1}{26} \cdot e^{-30/26}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{27} \cdot e^{-30/27}\right) + \left(\frac{2}{12}\right)\left(\frac{1}{29} \cdot e^{-30/29}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{30} \cdot e^{-30/30}\right)$$

$$+ \left(\frac{1}{12}\right)\left(\frac{1}{33} \cdot e^{-30/33}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{38} \cdot e^{-30/38}\right) + \left(\frac{1}{12}\right)\left(\frac{1}{53} \cdot e^{-30/53}\right) = .010.$$

4. The empirical probabilities of the three sample points are $p(2) = \frac{1}{5}$, $p(3) = \frac{3}{5}$, $p(7) = \frac{1}{5}$. For each of the three y -values ($y = 2, 3, 7$) there will be a triangle kernel function whose base runs from $y - b$ to $y + b$, where b is the bandwidth. In this case, $b = 2$. The triangle kernel based on bandwidth 2 has a base of length 4 and height .5 (so that the triangle area is 1). The following diagram indicates the three kernel functions (only the left part of the kernel function for $y = 7$ is shown).



We are estimating the density at $x = 2.5$. According to the kernel smoothing technique, we must find $k_y(2.5)$ for each sample point y and then the estimated density is

$$\hat{f}(2.5) = p(2) \cdot k_2(2.5) + p(3) \cdot k_3(2.5) + p(7) \cdot k_7(2.5).$$

$x = 2.5$ lies outside of the base of the triangle kernel for $y = 7$, so that $k_7(2.5) = 0$. To find $k_2(2.5)$ we see from the diagram that the peak of the triangle kernel for $y = 2$ is $k_2(2) = .5$ and the triangle kernel drops to $k_2(4) = 0$. Since $x = 2.5$ is 25% of the way from $x = 2$ to $x = 4$, it follows from the linearity of the right side of the triangle kernel that $k_2(2.5) = .375$. Similar reasoning shows that $k_3(2.5) = .375$. Then, $\hat{f}(2.5) = (.2)(.375) + (.6)(.375) = .3 = 12/40$.

Note that the formal definition of the triangle kernel at sample point y is

$$k_y(x) = \begin{cases} 0 & x < y - b \\ \frac{b+x-y}{b^2} & y - b \leq x \leq y \\ \frac{b+y-x}{b^2} & y \leq x \leq y + b \\ 0 & x > y + b \end{cases}.$$

For $y = 2$ with bandwidth $b = 2$, and $x = 2.5$ we have $k_2(2.5) = \frac{2+2-2.5}{2^2} = .375$ (since $y = 2 \leq 2.5 = x \leq 4 = y + b$). We could find $k_3(2.5)$ in a similar way. Answer: B

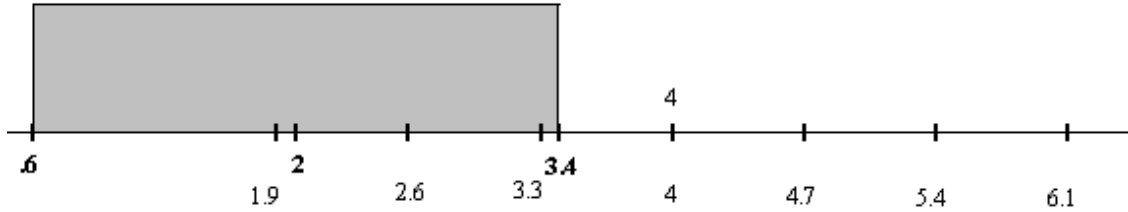
5. The data set has 8 data points. The empirical distribution assigns a probability of $p(2) = \frac{1}{8}$ to $y = 2$, $p(3.3) = \frac{2}{8}$ to $y = 3.3$, $p(4) = \frac{2}{8}$ to $y = 4.0$, and $p(4.7) = \frac{3}{8}$ to $y = 4.7$. With a bandwidth of 1.4, the point $x = 4$ will be to the right of the interval at $y = 2$, and it will be inside the intervals around 3.3, 4.0 and 4.7. The kernel smoothed estimate of $F(4)$ is equal to $\hat{F}(4) = \sum_{\text{All } y_j} p(y_j) \cdot K_{y_j}(4) = p(2)K_2(4) + p(3.3)K_{3.3}(4) + p(4)K_4(4) + p(4.7)K_{4.7}(4)$.

The uniform kernel sets up rectangles above each band interval at each sample y -value. For $y = 2.0$, the band interval is from .6 to 3.4, so $x = 4$ is to the right of the interval, and therefore $K_2(4) = 0$ (indicating the full rectangle centered at $y = 2$ is to the left of $x = 4$).

5. continued

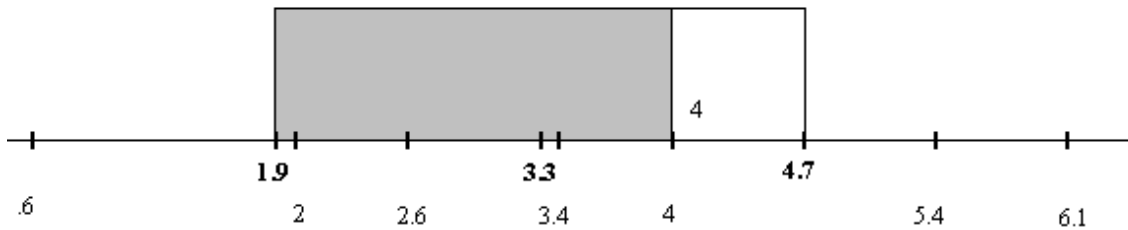
This is illustrated in the graph below.

$$K_2(4) = 1 = \text{area of rectangle centered at 2 that is to the left of } x = 4$$



For $y = 3.3$, the band interval is from 1.9 to 4.7, so the fraction to the left of $x = 4$ is $\frac{4-1.9}{2.8} = .75$ and therefore $K_{3.3}(4) = .75$.

$$K_{3.3}(4) = .75 = \text{area of rectangle centered at 3.3 that is to the left of } x = 4$$



For the point $y = 4$, the point $x = 4$ is in the middle of the band interval (from 2.6 to 5.4), so that $K_4(4) = .5$. For the point $y = 4.7$, the point $x = 4$ is .25 the band interval is from 3.3 to 6.1, so .25 of the interval is to the left of the point $x = 4$, and therefore $K_{4.7}(4) = .25$. Graphs similar to those above can be created for rectangles centered at $y = 4$ and $y = 4.7$.

$$\text{Then, } \hat{F}(4) = \sum_{\text{All } y_j} p(y_j) \cdot K_{y_j}(4) = \left(\frac{1}{8}\right)(1) + \left(\frac{2}{8}\right)(.75) + \left(\frac{2}{8}\right)(.5) + \left(\frac{3}{8}\right)(.25) = .53125.$$

Answer: D

6. There are 9 distinct observed values, The empirical probabilities are $p(y) = .1$ for all y -values except $p(35) = .2$. We wish to estimate $P(T > 40) = 1 - F(40)$.

With bandwidth $b = 10$, the interval for the uniform kernel at observed value y is from $y - 10$ to $y + 10$.

The kernel density estimate of $F(40)$ is $\hat{F}(40) = \sum_{i=1}^9 p(y_i) \cdot K_{y_i}(40)$.

For any y_i that is ≤ 30 , the interval around that y_i is totally to the left of $x = 40$, so

$F_{y_i}(40) = 1$ for that y_i ; this applies to $y = 25$ and 30 . For any $y_i \geq 50$, the interval around that y_i is totally to the right of $x = 40$, so $F_{y_i}(40) = 0$ for that y_i ; this applies to $y = 55$.

For any y_i for which $y_i - 10 \leq 40 \leq y_i + 10$, $F_{y_i}(40) = [40 - (y_i - 10)](.05)$

(this is the area of the rectangle whose base is from $y_i - 10$ to 40 with height $\frac{1}{2b} = \frac{1}{20} = .05$).

The kernel density estimate of $F(40)$ is

$$\begin{aligned} \hat{F}(40) &= (.1)(1) + (.1)(1) + (.2)[40 - (35 - 10)](.05) \\ &+ (.1)[40 - (37 - 10) + 40 - (39 - 10) + 40 - (45 - 10) \\ &+ 40 - (47 - 10) + 40 - (49 - 10)](.05) + 0 = .515. \end{aligned}$$

The estimate of $S(40)$ is $1 - .515 = .485$. Answer: E

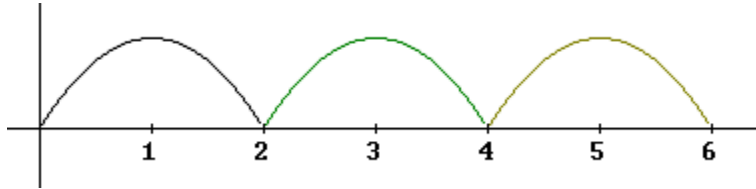
7. The graph below illustrates the three kernel functions. The first curve on the left is the kernel density function for the sample point $y_1 = 1$; $k_1(x) = \frac{2}{\pi} \sqrt{1 - (x - 1)^2}$ $0 \leq x \leq 2$.

The middle curve is the kernel density function for the sample point $y_2 = 3$;

$$k_3(x) = \frac{2}{\pi} \sqrt{1 - (x - 3)^2} \quad 2 \leq x \leq 4.$$

The curve on the right is the kernel density function for the sample point $y_3 = 5$;

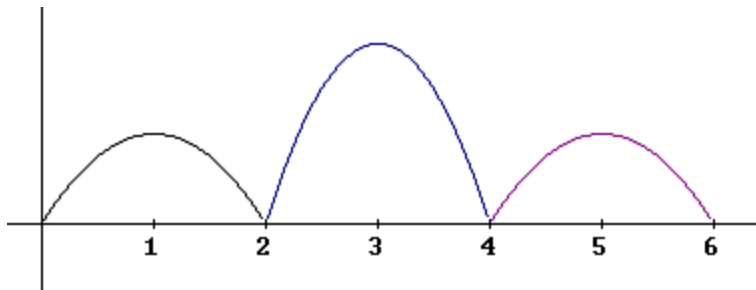
$$k_5(x) = \frac{2}{\pi} \sqrt{1 - (x - 5)^2} \quad 4 \leq x \leq 6.$$



The empirical probabilities are $p(1) = .25$, $p(3) = .5$ (2 sample values at 3), and $p(5) = .25$.

The kernel density estimator is $\sum_{j=1}^3 p(y_j)k_{y_j}(x) = (.25)k_1(x) + (.5)k_3(x) + (.25)k_5(x)$.

We see that the middle curve has double the coefficient as the curves on the left and right, so the middle curve is doubled, with resulting graph



Answer: D

8. The empirical distribution for the data set is $p(1) = .1$, $p(2) = .1$ and $p(3) = .8$.

For the uniform kernel, we have

$$\begin{aligned} \hat{F}_1(x) &= p(1) \cdot K_1^u(x) + p(2) \cdot K_2^u(x) + p(3) \cdot K_3^u(x) \\ &= (.1) \cdot K_1^u(x) + (.1) \cdot K_2^u(x) + (.8) \cdot K_3^u(x), \end{aligned}$$

and for the triangular kernel, we have

$$\begin{aligned} \hat{F}_2(x) &= p(1) \cdot K_1^t(x) + p(2) \cdot K_2^t(x) + p(3) \cdot K_3^t(x) \\ &= (.1) \cdot K_1^t(x) + (.1) \cdot K_2^t(x) + (.8) \cdot K_3^t(x), \end{aligned}$$

$K_y(x)$ is the cdf of the pdf $k_y(x)$. For the uniform kernel, $k_y^u(x)$ is constant on its bandwidth interval (so $K_y^u(x)$ is linearly increasing on its bandwidth interval). For the triangular kernel, $k_y^t(x)$ is linearly increasing on the first half of the bandwidth interval and linearly decreasing on the second half of the bandwidth interval (so $K_y^t(x)$ is a quadratic function on its bandwidth interval).

The three data points of 1, 2 and 3 have intervals with bandwidth 1 of $[0, 2]$, $[1, 3]$ and $[2, 4]$.

8. continued

On the interval $[0, 1]$, $k_1^u(x)$ is constant but $k_1^t(x)$ is linear, so

$$\widehat{F}_1(x) = p(1) \cdot K_1^u(x) \neq p(1) \cdot K_1^t(x) = \widehat{F}_2(x)$$

(note that $K_2(x)$ and $K_3(x)$ are 0 on the interval $[0, 1]$). It

A similar argument applies to the interval $[3, 4]$.

On the interval $[1, 2]$, $k_1^u(x)$ is constant at $\frac{1}{2}$, and $k_2^u(x)$ is constant at $\frac{1}{2}$, so that $k_1^u(x) + k_2^u(x)$ is constant at 1 on the interval.

On the interval $[1, 2]$, $k_1^t(x)$ is linearly decreasing from $\frac{1}{b} = 1$ to 0, and $k_2^t(x)$ is linearly increasing from 0 to 1, so that $k_1^t(x) + k_2^t(x)$ is constant at 1 on the interval.

It follows that $\widehat{f}_1(x) = .1k_1^u(x) + .1k_2^u(x)$ is constant at .2, and

$\widehat{f}_2(x) = .1k_1^t(x) + .1k_2^t(x)$ is constant at .2 on the interval $[1, 2]$, and therefore

$\widehat{F}_1(x)$ and $\widehat{F}_2(x)$ are equal on that interval. The reason this works out is that $p(1)$ and $p(2)$ are equal at .1.

On the interval $[2, 3]$ $.1k_2^u(x) + .8k_3^u(x)$ will be constant, but $.1k_2^t(x) + .8k_3^t(x)$ will not be constant, so \widehat{F}_1 and \widehat{F}_2 will not be the same. On the interval $[3, 4]$, $K_3^u(x)$ is linear and $K_3^t(x)$ is quadratic, so, $\widehat{F}_1(x)$ and $\widehat{F}_2(x)$ are not equal.

The only interval on which \widehat{F}_1 and \widehat{F}_2 are equal is $[1, 2]$.

This problem can also be viewed by considering the graphs of $\widehat{f}_1(x)$ and $\widehat{f}_2(x)$ and noting that the combination of the decreasing triangular function $k_1(x)$ and the increasing $k_2(x)$ on $[2, 3]$ results on a constant function, the same as the uniform kernel. This is valid only because $p(1) = p(2)$. This argument doesn't apply on $[2, 3]$ because $p(2) \neq p(3)$.

We can also solve the problem from a more formal algebraic point of view using the algebraic form of the kernel smoothed distribution functions. The kernel cdf function $K_y(x)$ has the following form for the uniform and triangular kernels.

$$\text{For the uniform kernel with bandwidth } b, \quad K_y^u(x) = \begin{cases} 0 & x < y - b \\ \frac{x-y+b}{2b} & y - b \leq x \leq y + b \\ 1 & x > y + b \end{cases}$$

and for the triangular kernel with bandwidth b , this is

$$K_y(x) = \begin{cases} 0 & x < y - b \\ \frac{(b+x-y)^2}{2b^2} & y - b \leq x \leq y \\ 1 - \frac{(b+y-x)^2}{2b^2} & y \leq x \leq y + b \\ 1 & x > y + b \end{cases}.$$

For the uniform kernel with bandwidth 1, we have the kernel cdf functions

$$K_1^u(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{2} & 0 \leq x \leq 2 \\ 1 & x > 2 \end{cases}, \quad K_2^u(x) = \begin{cases} 0 & x < 1 \\ \frac{x-1}{2} & 1 \leq x \leq 3 \\ 1 & x > 3 \end{cases}, \quad K_3^u(x) = \begin{cases} 0 & x < 2 \\ \frac{x-2}{2} & 2 \leq x \leq 4 \\ 1 & x > 4 \end{cases}$$

8. continued

For the triangular kernel with bandwidth 1, we have the kernel cdf functions

$$K_1^t(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{2} & 0 \leq x \leq 1 \\ 1 - \frac{(2-x)^2}{2} & 1 \leq x \leq 2 \\ 1 & x > 2 \end{cases}, \quad K_2^t(x) = \begin{cases} 0 & x < 1 \\ \frac{(x-1)^2}{2} & 1 \leq x \leq 2 \\ 1 - \frac{(3-x)^2}{2} & 2 \leq x \leq 3 \\ 1 & x > 3 \end{cases},$$

$$\text{and } K_3^t(x) = \begin{cases} 0 & x < 2 \\ \frac{(x-2)^2}{2} & 2 \leq x \leq 3 \\ 1 - \frac{(4-x)^2}{2} & 3 \leq x \leq 4 \\ 1 & x > 4 \end{cases}.$$

For $0 \leq x \leq 1$, we have $\widehat{F}_1(x) = p(1) \cdot K_1^u(x) + 0 + 0 = (.1)\left(\frac{x}{2}\right)$,

and $\widehat{F}_2(x) = p(1) \cdot K_1^t(x) + 0 + 0 = (.1)\left(\frac{x^2}{2}\right)$, so $\widehat{F}_1(x) \neq \widehat{F}_2(x)$ on $[0, 1]$.

For $1 \leq x \leq 2$, we have $\widehat{F}_1(x) = (.1) \cdot K_1^u(x) + (.1) \cdot K_2^u(x) = (.1)\left(\frac{x}{2}\right)$

$= (.1)\left(\frac{x}{2}\right) + (.1)\left(\frac{x-1}{2}\right) = (.1)\left(\frac{2x-1}{2}\right)$, and

$\widehat{F}_2(x) = (.1)\left(1 - \frac{(2-x)^2}{2}\right) + (.1)\left(\frac{(x-1)^2}{2}\right) = (.1)\left(\frac{2x-1}{2}\right)$,

so $\widehat{F}_1(x) = \widehat{F}_2(x)$ on $[1, 2]$.

Because of the nature of the possible answers, once we see that $\widehat{F}_1(x) = \widehat{F}_2(x)$ on $[1, 2]$, it follows that B must be correct. Answer: B

TABLE OF CONTENTS - VOLUME 2

CREDIBILITY

SECTION 1 - LIMITED FLUCTUATION CREDIBILITY	CR-1
PROBLEM SET 1	CR-17
SECTION 2 - BAYESIAN ESTIMATION, DISCRETE PRIOR	CR-31
PROBLEM SET 2	CR-41
SECTION 3 - BAYESIAN CREDIBILITY, DISCRETE PRIOR	CR-53
PROBLEM SET 3	CR-65
SECTION 4 - BAYESIAN CREDIBILITY, CONTINUOUS PRIOR	CR-91
PROBLEM SET 4	CR-101
SECTION 5 - BAYESIAN CREDIBILITY APPLIED TO THE EXAM C TABLE DISTRIBUTIONS	CR-115
PROBLEM SET 5	CR-127
SECTION 6 - BUHLMANN CREDIBILITY	CR-147
PROBLEM SET 6	CR-157
SECTION 7 - EMPIRICAL BAYES CREDIBILITY METHODS	CR-189
PROBLEM SET 7	CR-199

SIMULATION

SECTION 1 - THE INVERSE TRANSFORMATION METHOD	SI-1
PROBLEM SET 1	SI-9
SECTION 2 - THE BOOTSTRAP METHOD	SI-25
PROBLEM SET 2	SI-37
SECTION 3 - THE LOGNORMAL DISTRIBUTION AND ASSET PRICES	SI-45
PROBLEM SET 3	SI-53
SECTION 4 - MONTE CARLO SIMULATION	SI-59
PROBLEM SET 4	SI-67
SECTION 5 - RISK MEASURES STUDY NOTE	SI-69
PROBLEM SET 5	SI-77

PRACTICE EXAMS AND SOLUTIONS

PRACTICE EXAM 1	PE-1
PRACTICE EXAM 2	PE-25
PRACTICE EXAM 3	PE-45
PRACTICE EXAM 4	PE-69
PRACTICE EXAM 5	PE-93
PRACTICE EXAM 6	PE-115
PRACTICE EXAM 7	PE-137
PRACTICE EXAM 8	PE-159
PRACTICE EXAM 9	PE-183
PRACTICE EXAM 10	PE-207
PRACTICE EXAM 11	PE-229
PRACTICE EXAM 12	PE-251
PRACTICE EXAM 13	PE-273

MAY 2007 C/4 EXAM AND SOLUTIONS

EXAM QUESTION-TOPIC REFERENCE LIST

MAY 2007 EXAM AND SOLUTIONS

39. A sample of size n is used to estimate the parameters in two possible models for the data. The maximized log-likelihood for the 3-parameter generalized Pareto model is ℓ_A , and the maximized log-likelihood for the exponential model is ℓ_B . You are given that according to the Schwarz Bayesian Criterion, model A is preferred to model B. You are also given that according to the likelihood ratio test, in which the null hypothesis is that model B is acceptable, and the alternative hypothesis is that model A is preferable to model B, the null hypothesis is rejected at the 5% level of significance but not at the 1% level of significance. Find the maximum value of n that is compatible with these results.

A) 99 B) 101 C) 103 D) 105 E) 107

40. A stock based on the lognormal model has a current price of \$100. The expected price of the stock in one year is \$105. The stock pays no dividends and the volatility is 25% per year. Use the following uniform $(0, 1)$ numbers to simulate the stock price at time 2 using the inverse transformation method.

.2 .3 .8

Use the simulated values to estimate the average payoff at time 2 of a European call option with a strike price of 125 expiring at the end of 2 years.

- A) Less than 5 B) At least 5, but less than 6 C) At least 6, but less than 7
D) At least 7, but less than 8 E) At least 8

39. The Schwarz Bayesian Criterion compares $\ell_A - \frac{3}{2}\ln(n)$ and $\ell_B - \frac{1}{2}\ln(n)$, and since model A is preferable to model B this means that $\ell_A - \frac{3}{2}\ln(n) - [\ell_B - \frac{1}{2}\ln(n)] > 0$, which can be rewritten as $\ell_A - \ell_B > \ln(n)$.

The likelihood ratio test has test statistic $2(\ell_A - \ell_B)$. Since model A has 3 parameters and model B has 1 parameter, the number of degrees of freedom in the chi-square statistic is $3 - 1 = 2$. The critical value for a test with significance level 5% is $\chi_{.05}^2(2) = 5.991$ and the critical value for a test with significance level 1% is $\chi_{.01}^2(2) = 9.210$. Since the null hypothesis is not rejected at the 1% level, it must be true that $2(\ell_A - \ell_B) < 9.21$, so that $\ell_A - \ell_B < 4.605$. From the Schwartz Bayesian Criterion, we had $\ln(n) < \ell_A - \ell_B$, and therefore, $\ln(n) < 4.605$. It then follows that $n < e^{4.605} = 99.98$. The maximum (integer) value for n is 99. Answer: A

40. The simulated standard normal values are: $-.842, -.524, .824$.

The stock price at time 2 is $S_2 = 100 \cdot e^{(\alpha - \frac{1}{2}\sigma^2)(2)} \cdot e^{\sigma\sqrt{2}\cdot z}$, where $100e^\alpha = 105$, and $\sigma = .25$, so $S_2 = 101.77e^{.25\sqrt{2}z}$. Using the three simulated values of z , the simulated stock prices are $75.57, 84.56, 136.19$. The option values at time 2 based on the simulated stock prices are $0, 0, 16.19$. The average of these is 5.40 . Answer: B