Mean, Median and Mode from a Decision Perspective

Melinda Miller Holt
Stephen M. Scariano
Sam Houston State University

Key Words: Probability distribution; Return function; Prediction.

Abstract

The classroom activity described here allows mathematically mature students to explore the role of mean, median and mode in a decision-making environment. While students discover the importance of choosing a measure of central tendency, their understanding of probability distributions, maximization, and prediction is reinforced through active learning. The lesson incorporates the GAISE recommendations by actively engaging students in the process of statistical problem-solving in a realistic situation.

1. Introduction

Over the last two decades, the statistics reform movement has transformed teaching and learning in elementary statistics courses. The guiding principles of this movement are found in the Cobb Report (Cobb, 1992) and the Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report (Franklin and Garfield, 2006). Both documents encourage development of statistical reasoning, use of real (or at least realistic) data, an emphasis on conceptual knowledge and active learning in the classroom. In recent years, attention has turned to introductory statistics courses designed for students with a higher level of mathematical maturity than those in the typical algebra-based elementary course, thanks in large part to the efforts of Allan Rossman and Beth Chance (Chance and Rossman, 2006; Rossman, Medina, and Chance, 2006). In this article, we offer a useful framework for studying mean, median and mode in the context of prediction and decision-making. It is designed to emphasize the importance of selecting an appropriate measure of center when faced with real-world decision problems where monetary gains (and losses) are incurred. This approach, as presented here, is developmentally suitable for post-calculus students but may be easily adapted to a variety of ability levels. Our excursion is intended to answer the rhetorical question "Which measure of center should I use in a particular application?" It simultaneously strengthens student algebraic, probabilistic, and data analytic capabilities. Throughout, we highlight the constant interplay between reasoning/proof and problem-solving.

2. The Problem

Sky Bound Airlines would like to avoid empty seating on its Flight 123, which is flown daily between Houston and Dallas. In order to maximize revenue, Sky Bound would like to "overbook" these flights; that is, sell more seats than the jet can actually accommodate to compensate for passengers who do not show up for the flight. As a statistical consultant, you have been contracted to predict the number of "No Shows" on a given flight so the airline can sell the appropriate number of extra tickets. Historical records reveal that each of these daily flights has had at least one vacant seat. Note that we assume here that "No Shows" occur independently and at random. That is, we assume that each passenger's "No Show" status is independent of that of all other passengers. We also assume that there are no "No Show" patterns based on time (of day, week, month, or year). Formulate a solution for Sky Bound Airlines.

2.1 Step 1: Organize the data!

Based on a random selection of two hundred Sky Bound Flight 123 records between Houston and Dallas, we begin by tabulating and arranging the number of "No Shows" for these flights into the frequency distribution shown in Table 1.

Table 1: Frequency distribution for the number of "No Shows" from a sample of 200 flights.

"No Shows"	One	Two	Three	Four	Five	Six	*Total*
Frequency	**70**	**40**	**10**	**20**	**20**	**40**	***200***

Letting Y
represent the number of "No Shows" observed in our sample, the frequency distribution in Table 1 permits us to write its companion relative frequency distribution, or empirical probability mass function, $pr(Y)$, shown in Table 2.

Table 2: Companion empirical probability mass function for the number of

"No Shows" on Sky Bound Flight 123.

Y	1	2	3	4	5	6
$pr(Y)$	0.35	0.20	0.05	0.10	0.10	0.20

As qualified consultants, our goal is to predict the number of "No Show" passengers for Sky Bound Flight 123. We would, of course, like our prediction to exactly match the number of "No Shows" each day, but the natural variability seen in the frequency distribution suggests that such a hope is unrealistic. Since Y is a random variable having associated probability mass function $pr(Y)$, errors in the prediction are expected. Given this scenario, the key question here is: What is the "best" strategy for making a prediction? As we shall see, the answer actually depends not only on the probability mass function $pr(Y)$, but also on what is to be gained (or lost).

2.2 Step 2: Describe the distribution

We begin by determining the number of "No Shows" to expect. How many "No Shows" would YOU expect? Some students will answer $y = 1$, because it occurs most often. This is, of course, the ***mode***. Most courses cover the mode as a measure of central tendency during the introduction of descriptive statistics. Here, it is seen in a

different context.

Other students will incorrectly answer $y = 3.5$, which is concluded by averaging the possible values of Y, with no regard for the associated probabilities. This mistake provides an ideal opportunity to discuss the significance of weighted averages and introduce the definition of **expected value**. In general, the expected value of a discrete random variable Y having probability mass function $pr(Y)$ is its **mean**, defined as

$$\mu_Y = \sum_{all\ y} y \times pr(y)$$ 			(1)

So, for our problem, we expect the number of "No Shows" to be

$$m_Y = 1(0.35) + 2(0.20) + 3(0.05) + 4(0.10) + 5(0.10) + 6(0.20) = 3.$$

Now, recall another measure of central tendency – the **median**. Although somewhat less intuitive than the mode and mean, Table 2 shows that $y = 2$ basically splits the probability mass function roughly in half because $pr(Y < 2) \leq 0.50 \leq pr(Y \leq 2)$.

Here, each of these three measures of central tendency could be used as a legitimate prediction for the number of "No Shows," but they are all different! Which should we provide to Sky Bound Airlines? For that, we need a reasoned strategy.

2.3 Step 3: A general strategy for prediction

Our client needs a reasonable and practical solution for its daily operations. Let us begin by mathematically abstracting the information at hand. In general, let x denote our predicted number of "No Shows", and let $g(x, y)$ denote the monetary gain (or loss) to be realized when our prediction is x and then later $Y = y$ "No Shows" actually occur.

Now since we must make a prediction, x, before knowing which seemingly random outcome of Y ultimately occurs, let us define the **return function**, written $R(x)$, as the "average gain" derived from the probability distribution of the constellation of Y values for that particular prediction, x. That is, once a prediction x is made, the return function for that value of x is calculated as the weighted average

$$R(x) = \sum_{all\ y} g(x, y) pr(y)$$ 			(2)

Notice that the return function in equation (2) incorporates both the prediction, x, we make and the random nature of the "No Shows" using a weighted average of gains times probabilities. Adapting equation 1 appropriately, it is easy to see that $R(x)$ is simply the mean, or expected value of $g(x, y)$ for each fixed prediction x.

Next, we focus on how to make a "best" prediction. Of course, the adjective "best" is ambiguous and must be made mathematically precise. As hired consultants, it is quite reasonable for us to conclude that the "best" prediction, say x^*, is one that will yield the largest return for our consulting services. We call this decision criterion the **return maximum criterion.**
While we consider OUR maximum return, it is safe to assume that return functions discussed here reflect the best interests of the company. We leave the construction of potential exact return functions for the company as a student exercise.

Return Maximum Criterion: A prediction, $x*$, is said to be "best" provided it maximizes $R(x)$. That is, $x*$

satisfies $R(x*) \geq R(x)$ over all possible predictions x.

3. What's in it for us?

3.1 Scenario 1: The MODE is the "best" prediction!

Suppose Sky Bound Airlines contracts to pay us \$200 for our consulting services, with a \$600 bonus if our prediction exactly matches the true number of "No Shows". Here, it is easy to see that we are confronted with a "Hit or Miss" situation with resulting gain function

$$g(x,y) = \begin{cases} \$800, & \text{if } x = y \\ \$200, & \text{if } x \neq y \end{cases} \qquad (3)$$

Approach 1: If students are allowed to intuit the answer, some will arrive naturally at the **mode**, the most probable value of **Y**, but they typically have a difficult time defending their answer. The mode is, in fact, the correct choice but we need to confirm it mathematically.

Approach 2:
Other students may construct a table for the gain function similar to Table 3 and enumerate the possibilities.

Table 3: The gain function, $g(x,y)$, for Scenario 1.

	g(x,y)	"No Shows" y					
		1	2	3	4	5	6
Prediction	1	\$800	\$200	\$200	\$200	\$200	\$200
	2	\$200	\$800	\$200	\$200	\$200	\$200
	3	\$200	\$200	\$800	\$200	\$200	\$200
x	4	\$200	\$200	\$200	\$800	\$200	\$200
	5	\$200	\$200	\$200	\$200	\$800	\$200
	6	\$200	\$200	\$200	\$200	\$200	\$800

A common mistake here is to simply average the gains across each row without also considering the weighted effect of each individual probability associated with each possible y value. This mistake leads a student to incorrectly conclude that each prediction x
yields the same return, but the error opens an avenue for discussing the importance of weighted averages.

When gains are properly weighted with their corresponding probabilities and the return function in equation (2) is computed for $x \in \{1, 2, 3, 4, 5, 6\}$, we obtain the returns given in [Table 4](#).

Table 4: The return function, $R(x)$, for Scenario 1

$R(1) = \$800(0.35) + \$200(0.20) + \$200(0.05) + \$200(0.10) + \$200(0.10) + \$200(0.20) = \$410$
$R(2) = \$200(0.35) + \$800(0.20) + \$200(0.05) + \$200(0.10) + \$200(0.10) + \$200(0.20) = \$320$
$R(3) = \$200(0.35) + \$200(0.20) + \$800(0.05) + \$200(0.10) + \$200(0.10) + \$200(0.20) = \$230$
$R(4) = \$200(0.35) + \$200(0.20) + \$200(0.05) + \$800(0.10) + \$200(0.10) + \$200(0.20) = \$260$
$R(5) = \$200(0.35) + \$200(0.20) + \$200(0.05) + \$200(0.10) + \$800(0.10) + \$200(0.20) = \$260$
$R(6) = \$200(0.35) + \$200(0.20) + \$200(0.05) + \$200(0.10) + \$200(0.10) + \$800(0.20) = \$320$

Therefore, our predicted number of "No Shows" is $x^* = 1$ because $\$410 = R(1) = max\{ R(1), R(2), R(3), R(4), R(5), R(6)\}$. Moreover, Table 4 suggests a general pattern for the return function: $R(x) = \$800pr(x = y) + \$200pr(x \neq y)$. Because $pr(x = y) + pr(x \neq y) = 1$, it is easy to see that we should choose x^* to be the most probable, or modal, value from the empirical distribution of Y shown in <u>Table 2</u>. Moreover, the intuitive solution suggested in *Approach 1*
is easily defended when the return function is explicitly written in this particular form.

Approach 3: A slightly more sophisticated development allows for discussion of the **indicator** function. The function $I_x(y)$ defined as

$$I_x(y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{if } x \neq y \end{cases}.$$

is called the **indicator function of x** and it takes only two values: 1 if and only if $x = y$ and zero otherwise. In this case, $g(x, y)$ in equation (3) can easily be rewritten as the single rule $g(x, y) = \$200 + \$600I_x(y),$ which is more amenable to computation. In turn, this more compact functional representation simplifies the return function since

$$R(x) = \sum_{all\ y} g(x,y)pr(y)$$
$$R(x) = \sum_{all\ y} [\$200 + \$600I_x(y)]pr(y)$$
$$R(x) = \$200\sum_{all\ y} pr(y) + \$600\sum_{all\ y} I_x(y)pr(y)$$
$$R(x) = \$200(1) + \$600pr(Y = x)$$
$$R(x) = \$200 + \$600pr(Y = x).$$

It is now easy to see from <u>Table 5</u>
that this return function will be maximized only at a value yielding the largest probability in the distribution of Y, and we recognize this value as the **mode**.

Table 5: Computed returns from the alternate return function formulation for Scenario 1

$R(1) = \$200 + \$600\,pr(1) = \$200 + \$600(0.35) = \$410$
$R(2) = \$200 + \$600\,pr(2) = \$200 + \$600(0.20) = \$320$
$R(3) = \$200 + \$600\,pr(3) = \$200 + \$600(0.05) = \$230$
$R(4) = \$200 + \$600\,pr(4) = \$200 + \$600(0.10) = \$260$
$R(5) = \$200 + \$600\,pr(5) = \$200 + \$600(0.10) = \$260$
$R(6) = \$200 + \$600\,pr(6) = \$200 + \$600(.20) = \$320$

As before, $\$410 = R(1) = max\{R(1), R(2), R(3), R(4), R(5), R(6)\}$, so that \$410 is the maximum return we can expect and it occurs if our predicted number of "No Shows" is $x^* = 1$. Note that our actual compensation will be either \$200 or \$800, not \$410 or any of the other returns shown in Table 5. It is critically important to remember that the <u>criterion</u> used to arrive at the prediction $x^* = 1$ is designed to maximize the return (<u>average</u> <u>gain</u>) function, and no other. The general conclusion here is that the Return Maximum Criterion leads naturally to the *mode* as the "best" prediction when the gain function is specified as "Hit or Miss".

3.2 Scenario 2: The MEDIAN is the "best" prediction!

Suppose Sky Bound Airlines contracts to pay us \$800 minus \$50 for each seat that we over or under book. Then, our compensation is \$800 minus \$50 times *the absolute magnitude of our prediction error*. In that event, ever greater prediction errors (either too large or too small) are accompanied by ever lower compensation levels.

Approach 1: Some students choose the brute force method and compute the individual gains found in <u>Table 6</u>.

Table 6: The gain function, $g(x,y)$, for Scenario 2

		"No Shows"					
	$g(x,y)$			y			
		1	2	3	4	5	6
	1	\$800	\$750	\$700	\$650	\$600	\$550
	2	\$750	\$800	\$750	\$700	\$650	\$600
Prediction	3	\$700	\$750	\$800	\$750	\$700	\$650
	4	\$650	\$700	\$750	\$800	\$750	\$700
x	5	\$600	\$650	\$700	\$750	\$800	\$750
	6	\$550	\$600	\$650	\$700	\$750	\$800

In turn, the entries in <u>Table 6</u> lead directly to the returns given in <u>Table 7</u>, from which $x* = 2$ is seen to satisfy the Return Maximum Criterion.

Table 7: The return function, $R(x)$, for Scenario 2

$R(1) = \$800(0.35) + \$750(0.20) + \$700(0.05) + \$650(0.10) + \$600(0.10) + \$550(0.20) = \$700$
$R(2) = \$750(0.35) + \$800(0.20) + \$750(0.05) + \$700(0.10) + \$650(0.10) + \$600(0.20) = \$715$
$R(3) = \$700(0.35) + \$750(0.20) + \$800(0.05) + \$750(0.10) + \$700(0.10) + \$650(0.20) = \$710$
$R(4) = \$650(0.35) + \$700(0.20) + \$750(0.05) + \$800(0.10) + \$750(0.10) + \$700(0.20) = \$700$
$R(5) = \$600(0.35) + \$650(0.20) + \$700(0.05) + \$750(0.10) + \$800(0.10) + \$750(0.20) = \$680$
$R(6) = \$550(0.35) + \$600(0.20) + \$650(0.05) + \$700(0.10) + \$750(0.10) + \$800(0.20) = \$650$

Allow students to attempt a justification for this answer by asking what is "special" about this particular prediction. It is, in fact, a *median*
of the empirical distribution of "No Shows". Many students mistakenly believe that $x = 3.5$ is a median in this context. If so, they have again discounted the importance of the empirical probability distribution of *Y*. Table 7 shows that $x^* = 2$ is a median (or 50th percentile) precisely because $pr(x < 2) \leq 0.50 \leq pr(x \leq 2)$. [Recall that a number m^* is called a *median* of the probability distribution of a random variable *X* provided it satisfies the compound inequality $pr(X < m^*) \leq 0.5 \leq pr(X \leq m^*).$]

Approach 2:
A mathematically rigorous development that incorporates the absolute value function along with limits can be undertaken with advanced students. Careful scrutiny of Table 6 in tandem with knowledge of the nature of our compensation in this scenario allows us to write the gain function as

$$g(x,y) = \$800 - \$50|x - y|, \qquad (4)$$

where the absolute value function is used to compute the magnitude of the prediction error. This compact expression for the gain function leads to the following return function:

$$R(x) = \sum_{all\ y} g(x,y)pr(y)$$
$$R(x) = \sum_{all\ y} [\$800 - \$50|x - y|]pr(y)$$
$$R(x) = \$800 \sum_{all\ y} pr(y) - \$50 \sum_{all\ y} |x - y|pr(y)$$
$$R(x) = \$800 \times 1 - \$50 \sum_{all\ y} |x - y|pr(y)$$
$$R(x) = \$800 - \$50 \sum_{all\ y} |x - y|pr(y).$$

The absolute value function is generally formidable to work with. So, let $M(x) = \sum_{all\ y} |x - y|pr(y)$ and notice that $R(x)$ will be *maximized* when the nonnegative function $M(x)$ is *minimized*. and of $M(x)$ is a probability times the absolute value of a difference between prediction, x, and "No Shows", y. A useful

alternate expression for $|x - y|$ as given above in equation (4) is

$$|x - y| = \begin{cases} x - y, & \text{if } x \geq y \\ y - x, & \text{if } x < y \end{cases} \qquad (5)$$

which can be exploited to find the minimum of $M(x)$, and, thereafter, the maximum of the return function, $R(x)$.

In this scenario, function $M(x)$ is most easily computed by partitioning its domain, which is the entire real number line, into the seven subintervals,

$$x < 1, \ 1 \le x < 2, \ 2 \le x < 3, \ 3 \le x < 4, \ 4 \le x < 5, \ 5 \le x < 6, \ 6 \le x,$$

and writing equivalent expressions, using equation (5) repeatedly, that are free of the absolute sign. Table 8 exhausts the domain and shows how $M(x)$ can be simplified.

Table 8: Rewriting and simplifying $M(x)$ across its domain for easier analysis

x	$M(x)$
$x < 1$	$(1-x)(.35) + (2-x)(.20) + (3-x)(.05) + (4-x)(.10) + (5-x)(.10) + (6-x)(.20) = 3 - x$
$1 \le x < 2$	$(x-1)(.35) + (2-x)(.20) + (3-x)(.05) + (4-x)(.10) + (5-x)(.10) + (6-x)(.20) = 2.3 - .3x$
$2 \le x < 3$	$(x-1)(.35) + (x-2)(.20) + (3-x)(.05) + (4-x)(.10) + (5-x)(.10) + (6-x)(.20) = 1.5 + .1x$
$3 \le x < 4$	$(x-1)(.35) + (x-2)(.20) + (x-3)(.05) + (4-x)(.10) + (5-x)(.10) + (6-x)(.20) = 1.2 + .2x$
$4 \le x < 5$	$(x-1)(.35) + (x-2)(.20) + (x-3)(.05) + (x-4)(.10) + (5-x)(.10) + (6-x)(.20) = .4 + .4x$
$5 \le x < 6$	$(x-1)(.35) + (x-2)(.20) + (x-3)(.05) + (x-4)(.10) + (x-5)(.10) + (6-x)(.20) = -.6 + .6x$
$6 \le x$	$(x-1)(.35) + (x-2)(.20) + (x-3)(.05) + (x-4)(.10) + (x-5)(.10) + (x-6)(.20) = -3 + x$

Although $M(x)$ is expressed as a different linear
function on each of the seven subintervals defining its domain (the entire real number line), it is, nevertheless, a continuous function of x. More importantly, since

$$\lim_{x \to -\infty} M(x) = \lim_{x \to -\infty} (\mu - x) = \lim_{x \to -\infty} (3 - x) = \infty \ \text{ and } \ \lim_{x \to \infty} M(x) = \lim_{x \to \infty} (x - \mu) = \lim_{x \to \infty} (x - 3) = \infty,$$

the piecewise linear nature of $M(x)$ guarantees that its minimum value must occur either at the left endpoint of one of its domain subintervals or at all values between the endpoints of one of its domain subintervals. In fact, it is not hard to show (See Appendix 1) that the minimum value of $M(x)$ must occur when the slope coefficient changes from a negative value to either a positive value or zero. In the present context, Table 9 confirms the entries in Table 7.

Table 9: Comparing returns using $M(x)$ with Table 7 entries for Scenario 2

x	$M(x)$	$R(x)$
$1 \le x < 2$	$M(1) = -0.3 \times 1 + 2.3 = 2$	$R(1) = \$800 - \$50 \times 2.0 = \$700$
$2 \le x < 3$	$M(2) = 0.1 \times 2 + 1.5 = 1.7$	$R(2) = \$800 - \$50 \times 1.7 = \$715$
$3 \le x < 4$	$M(3) = 0.2 \times 3 + 1.2 = 1.8$	$R(3) = \$800 - \$50 \times 1.8 = \$710$
$4 \le x < 5$	$M(4) = 0.4 \times 4 + 0.4 = 2.0$	$R(4) = \$800 - \$50 \times 2.0 = \$700$
$5 \le x < 6$	$M(5) = 0.6 \times 5 - 0.6 = 2.4$	$R(5) = \$800 - \$50 \times 2.4 = \$680$
$6 \le x$	$M(6) = 6 - 3 = 3$	$R(6) = \$800 - \$50 \times 3.0 = \$650$

Since $715 = R(2) = max\{R(1), R(2), R(3), R(4), R(5), R(6)\}$, the maximum return we can expect is $715 and it occurs if our predicted number of "No Shows" is x^*
= 2, a median of the empirical distribution of "No Shows". As in the previous scenario, note that our actual compensation will not be $715; but, practically speaking, will be a multiple of $50 (up to $800) or, in the worst case, $0. The <u>criterion</u> used to arrive at the prediction x^*
= 2 is designed to maximize the return function, and the overall conclusion here is that the Return Maximum Criterion leads naturally to the *median*
as the "best" prediction when the gain function is specified in terms of the magnitude of the absolute prediction error.

3.3 Scenario 3: The MEAN is the "best" prediction!

Now suppose that Sky Bound Airlines contracts to pay us $1,000 minus $30 times the *square of the magnitude of the prediction error*. Then, it is easy to see that the gain function is $g(x,y) = \$1,000 - \$30(x-y)^2$. Notice that this gain function punishes us more severely for large prediction errors than the one based on the absolute values found in Scenario 2.

Approach 1: Some students choose to compute the individual gains seen in Table 10.

Table 10: The gain function, $g(x,y)$, for Scenario 3.

		"No Shows"					
	$g(x,y)$			y			
		1	2	3	4	5	6
Prediction **x**	1	$1000	$970	$880	$730	$520	$250
	2	$970	$1000	$970	$880	$730	$520
	3	$880	$970	$1000	$970	$880	$730
	4	$730	$880	$970	$1000	$970	$880
	5	$520	$730	$880	$970	$1000	$970
	6	$250	$520	$730	$880	$970	$1000

The gain function in Table 10 yields the selected returns shown in Table 11.

Table 11: The return function, $R(x)$, For Scenario 3.

$R(1) = \$1,000(0.35) + \$970(0.20) + \$880(0.05) + \$730(0.10) + \$520(0.10) + \$250(0.20) = \$763$
$R(2) = \$970(0.35) + \$1,000(0.20) + \$970(0.05) + \$880(0.10) + \$730(0.10) + \$520(0.20) = \$853$
$R(3) = \$880(0.35) + \$970(0.20) + \$1,000(0.05) + \$970(0.10) + \$880(0.10) + \$730(0.20) = \$883$
$R(4) = \$730(0.35) + \$880(0.20) + \$970(0.05) + \$1,000(0.10) + \$970(0.10) + \$880(0.20) = \$853$
$R(5) = \$520(0.35) + \$730(0.20) + \$880(0.05) + \$970(0.10) + \$1,000(0.10) + \$970(0.20) = \$763$
$R(6) = \$250(0.35) + \$520(0.20) + \$730(0.05) + \$880(0.10) + \$970(0.10) + \$1,000(0.20) = \$613$

Here we have yet another "best" prediction, namely $x^* = 3$. This time we have arrived at the *mean* as the value that maximizes the Return Maximum Criterion. As before, students may incorrectly surmise that the mean is 3.5, but weighting the values of *Y* with its associated empirical probabilities yields the correct value of 3.

Approach 2:
A more advanced mathematical confirmation of the conclusions reached above requires knowledge of

differential calculus. Note that the return function can be rewritten as

$$R(x) = \sum_{all\ y} g(x,y)pr(y)$$

$$R(x) = \sum_{all\ y} [\$1000 - \$30(x-y)^2]pr(y)$$

$$R(x) = \$1000 \sum_{all\ y} pr(y) - \$30 \sum_{all\ y} (x-y)^2 pr(y)$$

$$R(x) = \$1000 - \$30 \sum_{all\ y} (x-y)^2 pr(y)$$

Momentarily, let $L(x) = \sum_{all\ y} (x-y)^2 pr(y)$
(This is, of cours　　　　　　　　east-squares concept.). Clearly, $R(x)$ will be maximized when $L(x)$ is minimized. Since $L(x)$ is continuously differentiable,

$$L'(x) = \sum_{all\ y} (+2)(x-y)^1 pr(y).$$

Setting

$$0 = L'(x) = (+2)\sum_{all\ y} (x-y)^1 pr(y)$$

and solving for all critical numbers,

$$0 = (+2)\sum_{all\ y} (x-y)^1 pr(y)$$

$$0 = x\sum_{all\ y} p(y) - \sum_{all\ y} (y \times pr(y))$$

$$0 = x - \mu$$

$$x = \mu$$

we get $\mu = \sum_{all\ y} (y \times pr(y)) =$ the mean of the empirical distribution of *Y*, the "No Shows".

Additionally, since the second derivative evaluated at μ is $L''(\mu) = 2$, $L(x)$ is, indeed, minimized at $x = \mu$ by the result of the Second Derivative Test. In our scenario,

$$\mu = (1 \times .35) + (2 \times .20) + (3 \times .05) + (4 \times .10) + (5 \times .10) + (6 \times .20) = 3,$$

so that

$$R(3) = \$1,000 - \$30 \times L(3)$$
$$R(3) = \$1,000 - \$30 \times \left[(1-3)^2(.35) + (2-3)^2(.20) + (3-3)^2(.05) + (4-3)^2(.10) + (5-3)^2(.10) + (6-3)^2(.20) \right]$$
$$R(3) = \$1,000 - \$30 \times [1.40 + .20 + 0 + .10 + .40 + 1.80]$$
$$R(3) = \$1,000 - \$30 \times 3.9$$
$$R(3) = \$1,000 - \$117$$
$$R(3) = \$883,$$

which concurs with its corresponding Table 11 entry.

As in the previous two scenarios, notice that our actual compensation will not be \$883, which is simply the maximum value of the return function at the "best" prediction, $x^* = 3$. In fact, we can expect our actual compensation to be one of the values seen in the gain matrix of Table 10.

4. Conclusions

The three scenarios studied here provide guidance for answering the question, "Which measure of center should I use in a particular application?" In Scenario 1, the gain function is specified as "Hit or Miss" and the return function, where A and B are constants with $A > 0$ and $B > 0$, has the general form: $R(x) = A + (B \times pr(x))$, so choose the **mode** of **Y** as the "best" prediction. In Scenario 2, the gain is specified as a function of the *absolute magnitude of the prediction error*, $|x-y|$; its general form is $R(x) = A - B \sum_{all\ y} |x-y| pr(y)$, so choose a **median** of **Y** as the "best" prediction. Finally, Scenario 3 has gain spe   uare of the magnitude of the prediction error $(x-y)^2$; its general form is $R(x) = A - B \sum_{all\ y} (x-y)^2 pr(y)$, so choose the **mean** of **Y** as the "best" prediction. Table 12 summarizes these c

Table 12: "Which Measure of Center Should I Use?" for the Return Maximum Criterion.

ESSENCE OF GAIN FUNCTION	RETURN FUNCTION (general form) $(A > 0$ and $B > 0)$	MEASURE OF CENTER TO USE
"Hit or Miss"	$R(x) = A + (B \times pr(x))$	Mode of Y
Absolute prediction error, $|x-y|$	$R(x) = A - B \sum_{all\ y} |x-y| pr(y),$	Median of Y
Squared prediction error, $(x-y)^2$	$R(x) = A - B \sum_{all\ y} (x-y)^2 pr(y),$	Mean of Y

The gain and return functions studied here are predicated on making a "best" decision from the consultant's viewpoint for simplicity of assumptions and function descriptions. Based on our experiences, we recommend starting from the consultant's viewpoint; however, you and your students may decide to pursue different decision scenarios. For example, as mentioned above, you may choose to investigate what the gain and return functions might look like from the company's perspective, and, following that, what the "best" predictions are from that viewpoint. That exploration can be done individually, in small groups, or with the whole class. Table 12 provides a good starting point for that discussion, which we leave for self-discovery. When modeling directly from the company's perspective, absolute prediction error is the most reasonable and realistic of the three types discussed here. Advanced students may choose to study functions that model greater loss for flights with empty seats than for flights that are overbooked. These do not necessarily lead to the mode, mean or median as the best decision.

Appendix 2
contains additional instructor resources for alternative probability distributions for the number of "No Shows", Y, that involve the possibility Y
$= 0$ as well as a bimodal distribution that might represent passengers traveling in pairs. It also includes references we have found useful for active learning of probability distributions. These resources provide discussions, examples and assignments that produce empirical distributions through simulations. This would allow for rich discussions of probability distributions and measures of center for computer savvy but less mathematically inclined students.

We have taught this lesson to students in sophomore/junior level courses. To do so takes a full 80 minute class. To cover the full content in lower level courses, we recommend allowing up to two class meetings. Students are typically anxious during the introduction of the problem and with development of the Return Maximum Criterion objective function. They become more comfortable with the topics and lesson goals once they see the complete development of the decision scenario involving the mode as the "best prediction" (Section 3.1). At this point, students see the role of an indicator function and the interrelationships among gain, risk, probability and money. Section 3.2 and introduction of the median, reinforces their ability to work with absolute values. For those with a calculus background, coverage of the mean in Section 3.3 is an excellent opportunity to apply their knowledge of maximization using derivatives. For students with less mathematical maturity, the activity can be limited to Approaches 1 and 2 when motivating the mode, and to Approach 1 when motivating the median and mean. In addition to the reinforcing knowledge of the indicator function, absolute value, maximization, and weighted averages, this lesson stresses the interconnectedness among probability, prediction, decision-making, and money management issues.

Prediction and inference supporting real-world decision-making certainly involves data collection and analysis, but often much more. Reason demands of us not only a sound decision strategy that incorporates both the probabilistic and deterministic aspects of any decision problem, but also careful consideration of the consequences associated with the decisions that are possible. The decision strategy studied here is the Return Maximum Criterion, but it is just one of many alternatives that could be chosen. The consequences associated with decision-making are quantified through the gain function, which is intimately linked to the decision strategy adopted. In general, studying the mean, median, and mode from a decision perspective is a rich mathematical experience, and it affords the teacher a superb opportunity to connect a variety of statistical concepts in a single pedagogical stroke.

APPENDIX 1

Let $y_1 < y_2 < \ldots < y_m$ be the **ordered**, **distinct** values of the discrete random variable Y having probability mass function $pr(y) = pr(Y = y)$ and expected value $E(Y) = \mu$. Note that

For $x \in (-\infty, y_1)$,	$M(x) = (y_1 - x)pr(y_1) + (y_2 - x)pr(y_2) + \ldots + (y_m - x)pr(y_m)$
For $x \in [y_1, y_2)$,	$M(x) = (x - y_1)pr(y_1) + (y_2 - x)pr(y_2) + \ldots + (y_m - x)pr(y_m)$
For $x \in [y_2, y_3)$,	$M(x) = (x - y_1)pr(y_1) + (x - y_2)pr(y_2) + \ldots + (y_m - x)pr(y_m)$
....	....
For $x \in [y_j, y_{j+1})$,	$\sum_{i=1}^{j}(x - y_i)pr(y_i) + \sum_{i=j+1}^{m}(y_i - x)pr(y_i)$
....	....
For $x \in [y_{m-1}, y_m)$,	$M(x) = (x - y_1)pr(y_1) + \ldots + (x - y_{m-1})pr(y_{m-1}) + (y_m - x)pr(y_m)$
For $x \in [y_m, \infty)$,	$M(x) = (x - y_1)pr(y_1) + (x - y_2)pr(y_2) + \ldots + (x - y_m)pr(y_m)$

On collecting like terms and setting $\mu = \sum_{i=1}^{m} y_i pr(y_i)$, this expression for $M(x)$ can be written more compactly as

$$M(x) = \begin{cases} \mu - x, & x < y_1 \\ x\left[\sum_{i=1}^{j} pr(y_i) - \sum_{i=j+1}^{m} pr(y_i)\right] - \left[\sum_{i=1}^{j} y_i pr(y_i) - \sum_{i=j+1}^{m} y_i pr(y_i)\right], & y_j \leq x < y_{j+1} \\ x - \mu, & x \geq y_m \end{cases}$$

for $j = 1, 2$

$\mu_j = \sum_{i=1}^{j} y_i pr(y_i)$ Employing the *cumulative distribution function* of Y, commonly written as $F(x) = \sum_{y_i \leq x} pr(y_i)$, ... en more compactly as

$$M(x) = \begin{cases} \mu - x, & x < y_i \\ x\left[2F(y_j) - 1\right] - \left[2\mu_j - \mu\right], & y_j \leq x < y_{j+1} \\ x - \mu & x \geq y_m \end{cases}$$

It is now straightforward to check that

a) $M(x)$ is <u>linear</u> on each of $(-\infty, y_1), [y_1, y_2),\ldots [y_{m-1}, y_m), [y_m, \infty)$.

b) $M(x)$ is continuous throughout $\mathbf{R}$, the entire real line.

c)   Assuming that the probabilities $pr(Y = y)$ are all nonzero, $M(x)$ will achieve its minimum value at the smallest y_i such that $2F(y_i) - 1 \geq 0$. That is, $M(x)$ will achieve its minimum value at, say y_k , where k ($k \in \{1, 2, \ldots, m\}$) is the first index such that $F(y_k) = pr(Y \leq y_k) \geq \frac{1}{2}$ and yet $pr(Y < y_k) \leq \frac{1}{2}$. This is, however, the defining criterion for y_k to be a *median* of the probability distribution of Y. Let $\tilde{\mu} = y_k$ denote this *median*.

d)   Therefore, the minimum value of $M(x)$ is computed as $M(\tilde{\mu}) = \tilde{\mu}[2F(\tilde{\mu})-1]-[2\mu_k - \mu]$.

e)   Because the family of return functions discussed here are all of the form $R(x) = A - (B \times M(x))$, where A and B are suitable positive constants, $R(\tilde{\mu})$ is the maximum value of the return function, $R(x)$.

Aside:

The median is not necessarily unique. It is possible for $F(y_j) = \frac{1}{2}$ for some index j. In that event, $2F(y_j) - 1 = 0$ and $M(x)$ achieves a constant, minimum value of $(\mu - 2\mu_j)$ throughout $[y_j, y_{j+1})$.   In that case, **any** $x \in [y_j, y_{j+1})$ could be called a *median* for Y. Many textbooks choose $(y_j + y_{j+1})/2$, but this is only a convenience.

APPENDIX 2

Here are three alternative empirical probability mass functions for the number of "No Shows", Y, that include the possibility Y
$= 0$. The distribution used in the lesson above was right-skewed. Here we include bimodal, symmetric, and left-skewed distributions. Using these can lead to a discussion of the relationship between the shape of a distribution and the measures of center.

     Bimodal:

Y	0	1	2	3
$pr(Y)$	0.30	0.20	0.30	0.20

Symmetric:

Y	0	1	2	3	4
$pr(Y)$	0.10	0.20	0.40	0.20	0.10

Left-Skewed:

Y	0	1	2	3
$pr(Y)$	0.10	0.20	0.30	0.40

For instructors who wish to emphasize the development of probability distributions and the discussion of empirical distributions through active learning, we recommend Topic 11 of Rossman and Chance (2008). In addition, Chapters 19 and 20 of Moore and Notz (2009) and Sections 6.2 and 6.3 of Peck and Devore (2007) provide nice explanations of simulating probability distributions and associated activities. Also, Freund (2004), Chapter 7, Sections 7.2 and 7.3 provide concise, elementary presentations of Bayesian decision analysis and statistical decision theory. These additional resources can be used to emphasize the calculation of central tendency measures, develop additional objective functions and nuance our development.

Acknowledgment

References

Chance, B. and Rossman, A. (2006), *Investigating Statistical Concepts, Applications, and Methods*, Pacific Grove CA: Duxbury Press.

Cobb, G. (1992), "Teaching Statistics," in *Heeding the Call for Change: Suggestions for Curricular Action*, ed. L. Steen, MAA Notes No. 22, Washington, D.C.: Mathematical Association of America, 3-23.

Franklin, C.A. and Garfield, J.G. (2006), "The GAISE Project: Developing statistics education guidelines for grades pre-K-12 and college courses," In G.F. Burrill alnd P.C. Elliott (eds) *Thinking and Reasoning with Data and Chance: 2006 NCTM Yearbook*, pp. 345-376. Reston, VA: National Council of Teachers of Mathematics. Available from http://www.amstat.org/education/gaise/GAISECollege.htm.

Freund, J. E. (2004), *Modern Elementary Statistics*, *Eleventh Edition*, New Jersey: Pearson Education, Inc.

Moore, D. S. and Notz, W. I. (2009), *Statistics: Concepts and Controversies*, *Seventh Edition*, New York: W. H. Freeman and Company.

Peck, R. and Devore, J. L. (2007), *Statistics: The Exploration & Analysis of Data, Sixth Edition*, Pacific Grove,

CA: Duxbury Press.

Rossman, A. J. and Chance, B. L. (2008), *Workshop Statistics: Discovery with Data, Third Edition*, Emeryville, CA: Key College Publishing.

Rossman, A., Medina, E. and Chance, B. (2006), "A Post-Calculus Introduction to Statistics for Future Secondary Teachers," *Proceedings of the Seventh International Conference on Teaching of Statistics*. Available from http://www.stat.auckland.ac.nz/~iase/publications/17/2E2_ROSS.pdf.

Melinda Miller Holt
Department of Mathematics and Statistics
Sam Houston State University
P.O. Box 2206
Huntsville, TX 77341
936-294-4859
E-mail: mholt@shsu.edu

Stephen M. Scariano
Department of Mathematics and Statistics
Sam Houston State University
P.O. Box 2206
Huntsville, TX 77341
936-294-1506
E-mail: sms049@shsu.edu