Median Duration Estimation When Continuous Event Data Is Reported in Discrete Intervals

Donald M. Luery, Richard S. Sigman, U.S. Bureau of the Census
Donald M. Luery, U.S. Bureau of the Census, Washington, DC 20233

Key Words: Linear interpolation; quadratic interpolation; Survey of Construction

The Survey of Construction (SOC) is a monthly survey of residential new construction. SOC has a multi-stage design where the ultimate sample units are single family and multi family building permits. For single family construction, SOC collects event data such as the month of start, the month of sale if the house is for sale, and the month of completion, financial data, and characteristics of the buildings. SOC publishes statistics based on these event data such as the number of starts in a month and the median and frequency distribution of the number of months on the market (duration from start to sale).

A duration is the time between two events. A duration can be reported directly or derived as in SOC as the difference between the reported times of the two events. These estimates will be referred to as derived durations. Event data is often reported as occurring in broad intervals such as months when the actual events happen at a much finer time scale such as day and hour for the issuance of a permit. The events occur over a continuous time scale and the durations on this scale will be referred to as the true durations.

SOC computes the median by a linear interpolation into the empirical distribution for the derived duration data. SOC finds the largest duration with cumulative frequency less than or equal to half the total frequency. Call this largest duration m. SOC interpolates between durations m and $m+1$ and adds ½ to this interpolated value to obtain the estimated median duration. This is illustrated in table 1.

This procedure follows from an assumption that, for a derived duration, the true duration is distributed uniformly between $m-½$ and $m+½$. From this assumption, the cumulative frequency count for duration m is then at the mid point of the month, $m+½$, and interpolation is linear between $m+½$ and $m+{}^3/_2$.

The problem with SOC's procedure is the assumption that the true duration for a derived duration of m months is uniformly distributed over an interval of length one centered at $m+½$. The assumption that a true duration is uniformly distributed is not justified.

Table 1. Months from Start to Sale
–Single Family–

Months from Start	Frequency	Cum. Dist.
0	10,370	10,370
1	4,103	14,473
2	2,274	16,747
3	3,104	19,851
4	3,254	23,105
5	3,282	26,387
6	3,429	29,816
7+	13,912	43,728
Total	43,728	
Median number of months		4.1

Half total frequency = 21,864, m = 3

$$median = 3 + \frac{21{,}864 - 19{,}851}{23{,}105 - 19{,}851} + 1/2 = 4.118$$

For example, for a reported start in January and a reported sale in April (a derived duration of three months), the duration between a start at the beginning of January and a sale at the end of April is about four months. Similarly, the duration between a start at the end of January and a sale at the beginning of April is only two months. The true duration extends over an interval of length two months centered at m but may not be uniform.

The most realistic assumption using only the available information is that an event is uniformly distributed through the reported month. For example, if the reported start date is January, the actual start time could equally likely be any day or time in that month. From this assumption and an assumption of conditional independence described below, we show that the true duration does not have a uniform distribution but it has a triangular distribution over an interval of length two centered at m.

This paper develops an alternative estimator (called quadratic median) for the median duration and compares it to the linear median analytically and through simulation.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

Estimation for Median Duration

Define the following:

- s = the reported month of the first event,
- c = the reported month of the second event,
- $n = c - s$ = the derived duration as an integer number of months based on the reported months for the two events,
- x = the actual time of the first event measured in months and fractions of months,
- y = the actual time of the second event measured in months and fractions of months,
- $z = y - x$ = the true duration time,
- p_n = the probability that the derived duration is n months, and
- $F(z)$ is the cumulative distribution function (cdf) of the true duration (z).

We make the following assumptions.

(1) The conditional distribution of x given s is uniform over the interval $[s, s+1)$, its density is $f_X(x \mid s) = I_{[s, s+1)}(x)$.

(2) The conditional distribution of y given c is uniform over the interval $[c, c+1)$, its density is $f_Y(y \mid c) = I_{[c, c+1)}(y)$.

(3) The two conditional distributions are independent, $f_{XY}(x, y \mid s, c) = f_X(x \mid s)\, f_Y(y \mid c)$.

For example, this last condition says that, if the months of the two events are both known, the actual time of the first event provides no information on the time in the month when the second event occurred and *vice versa*.

It follows from these assumptions that the conditional distribution of z is

$$f(z \mid s, n) = \int_{-\infty}^{\infty} f_y(x+z \mid s+n)\, f_x(x \mid s)\, dx$$
$$= \int_{-\infty}^{\infty} I_{[s+n, s+n+1)}(x+z)\, I_{[s, s+1)}(x)\, dx.$$

The integrand is 1 on the interval $\max(s+n-z, s) \le x < \min(s+n-z+1, s+1)$ and 0 otherwise. Thus,

$$f(z \mid s, n) = \min(s+n-z+1, s+1) - \max(s+n-z, s)$$
$$= 1 - |n - z|$$

is the density for a triangular distribution on the interval $[n-1, n+1)$ that depends only on n, not on c or s. In the following, we will write this conditional density as $f(z \mid n)$.

Without loss of generality, let $n \ge 0$. The unconditional density z is $f(z) = \sum_{n=0}^{\infty} p_n f(z \mid n)$ and, on the interval $[n, n+1)$, $f(z) = p_n + (p_{n+1} - p_n)(z - n)$.

<u>Lemma 1</u>

The cumulative distribution function of the true duration z is continuous and, on the interval $[n, n+1)$, $F(z) = F(n) + \frac{1}{2}(p_{n+1} - p_n)(z-n)^2 + p_n(z-n)$ where, for integers n,

$$F(n) = \sum_{j=0}^{n-1} p_j + \frac{1}{2} p_n \text{ and } F(n+1) = F(n) + \frac{1}{2}(p_{n+1} + p_n).$$

<u>Proof:</u> This is shown by integrating over $f(z)$.□

The median is the value of z when the cumulative distribution function equals ½, $F(z) = ½$. Because the median is a solution to a quadratic equation, we called this method *quadratic interpolation.*

<u>Theorem</u> 1: Let m be such that $F(m) \le ½ < F(m+1)$ then the quadratic median Q_{med} is given by

(1) $$Q_{med} = m + \frac{\sqrt{p_m^2 + 2(p_{m+1} - p_m)(½ - F(m))} - p_m}{p_{m+1} - p_m}$$ for $p_{m+1} - p_m \ne 0$ and

(2) $$Q_{med} = m + \frac{½ - F(m)}{p_m}$$ for $p_{m+1} - p_m = 0$ and $p_m > 0$.

Q_{med} is continuous at $p_{m+1} - p_m$.

<u>Proof:</u> This is shown by solving the quadratic equation in Lemma 1.□

Analysis of Linear and Quadratic Medians

Before proceeding further, we need to define the cdf for the linear median. The cdf is $G(n+½) = \sum_{j=0}^{n} p_j$ where the point masses, p_n, are taken to be at the midpoint of the month, $n+½$. The notation $G(n+½)$ is suggestive of this. $G(n+½)$ is between $F(n)$ and $F(n+1)$ because

$$F(n) = \sum_{j=0}^{n-1} p_j + ½ p_n \le \sum_{j=0}^{n} p_j \le \sum_{j=0}^{n} p_j + ½ p_{n+1} = F(n+1).$$

For m, as defined for Q_{med}, $F(m) \le ½ < F(m+1)$ so that $p_{m+1} + p_m > 0$. For linear interpolation, there are two cases: when $G(m+½)$ is greater than ½ (case A), or less than or equal to ½ (case B).

A. $G(m-½) < F(m) \le ½ < G(m+½) \le F(m+1)$ (1) for $p_m > 0$.

B. $F(m) \le G(m+½) \le ½ < F(m+1) < G(m+\frac{3}{2})$ (2) for $p_{m+1} > 0$.

The value of $G(n+½)$ that is closest to ½ is $G(m+½)$ for both cases.

For case A, the linear interpolation is between $m-½$ and $m+½$. For case B, the linear interpolation is between $m+½$ and $m+3/2$. The formulas for the linear median in the two cases are

A. $$L_{med} = m - ½ + \frac{½ - G(m-½)}{G(m+½) - G(m-½)} = m + \frac{½(1 + p_m) - G(m+½)}{p_m}$$ for $G(m+½) > ½$

$$L_{med} = m + \tfrac{1}{2} + \frac{\tfrac{1}{2} - G(m+\tfrac{1}{2})}{G(m+\tfrac{3}{2}) - G(m+\tfrac{1}{2})}$$

B. for $G(m+\tfrac{1}{2}) \le \tfrac{1}{2}$.

$$= m + \frac{\tfrac{1}{2}(1+p_{m+1}) - G(m+\tfrac{1}{2})}{p_{m+1}}$$

When $G(m+\tfrac{1}{2}) = \tfrac{1}{2}$, $L_{med} = m + \tfrac{1}{2}$.

It will be more convenient in the remainder of this paper to work with the $G(\cdot)$ distribution instead of the $F(\cdot)$. In the prior formulae for the quadratic median, substitute $G(m+\tfrac{1}{2}) - \tfrac{1}{2}\,p_m$ for $F(m)$.

$$Q_{med} = m + \frac{\sqrt{p_m^2 + 2(p_{m+1} - p_m)(\tfrac{1}{2}(1+p_m) - G(m+\tfrac{1}{2}))} - p_m}{p_{m+1} - p_m}$$

The difference between the quadratic and linear medians is $\Delta_{med} = Q_{med} - L_{med}$. When $G(m+\tfrac{1}{2}) = \tfrac{1}{2}$, this simplifies to

$$\Delta_{med} = -\frac{1}{2}\,\frac{\left(\sqrt{p_{m+1}} - \sqrt{p_m}\right)^2}{p_{m+1} - p_m}. \tag{3}$$

From Theorem 1, the quadratic and the linear medians are equal, i.e., $\Delta_{med} = 0$, when $p_{m+1} - p_m = 0$.

Theorem 2 shows that the difference in the medians has a negative symmetry so that Δ_{med} has the opposite sign for corresponding points in case A, $(G(m+\tfrac{1}{2}), p_m, p_{m+1})$, and case B, $(1-G(m+\tfrac{1}{2}), p_{m+1}, p_m)$. It will be sufficient in the remainder to show results only for case A.

Define a new distribution $H(\cdot)$ such that $H(m+\tfrac{1}{2}) = 1-G(m+\tfrac{1}{2})$, $h_{m+1}=p_m$, and $h_m=p_{m+1}$. $H(\cdot)$ is a proper cdf.

Theorem 2: The median for the $H(\cdot)$ distribution, $\Delta_{med}(H)$, equals $-\Delta_{med}(G)$, the median for the $G(\cdot)$ distribution for the corresponding points.

Proof: This is shown for case A, $G(m+\tfrac{1}{2}) > \tfrac{1}{2}$. The proof for case B is very similar. First, suppose m is such that $H(m+\tfrac{1}{2}) < \tfrac{1}{2}$, case B for the $H(\cdot)$ distribution. By substitution into $\Delta_{med}(H)$ of $H(m+\tfrac{1}{2}) = 1-G(m+\tfrac{1}{2})$, $h_{m+1}=p_m$, and $h_m=p_{m+1}$, it can be shown that $\Delta_{med}(H) = -\Delta_{med}(G)$ for $G(m+\tfrac{1}{2}) > \tfrac{1}{2}$, (case A for the $G(\cdot)$ distribution). From (3), it is clear that there is negative symmetry when $G(m+\tfrac{1}{2}) = \tfrac{1}{2}$.□

Theorem 2 shows that results for case A will be the same as for case B but with the sign of Δ_{med} changed.

Feasible Region

Δ_{med} is determined only by $G(m+\tfrac{1}{2})$, p_m and p_{m+1}. For $G(m+\tfrac{1}{2}) > \tfrac{1}{2}$, p_m and p_{m+1} are contained within the feasible region given by

1. $0 < 2G(m+\tfrac{1}{2}) - 1 \le p_m \le G(m+\tfrac{1}{2})$
2. $0 \le p_{m+1} \le 1 - G(m+\tfrac{1}{2})$.

The region is square with sides equal to $1-G(m+\tfrac{1}{2})$.

For $G(m+\tfrac{1}{2}) < \tfrac{1}{2}$, p_m and p_{m+1} are contained within the feasible region given by

1. $0 < 1 - 2G(m+\tfrac{1}{2}) \le p_{m+1} \le 1 - G(m+\tfrac{1}{2})$
2. $0 \le p_m \le G(m+\tfrac{1}{2})$.

The region is square with the sides equal to $G(m+\tfrac{1}{2})$. These inequalities follow from (1) and (2).

Bounds for the Difference, Δ_{med}

We wished to determine the largest and smallest differences between the quadratic and linear medians to provide bounds on Δ_{med}. Lemma 2 in the appendix shows that there are no critical points of Δ_{med} in the interior of the feasible region so that any maxima or minima must be on its boundaries.

Figure 1 shows a contour plot of a slice through the feasible region where $G(m+\tfrac{1}{2}) = 0.55$. The numbers to the top and the right of the feasible region are the values of the contour lines. Above the dashed contour line, Δ_{med} is negative and, below the line, it is positive. On the dashed contour line, Δ_{med} is zero and p_m and p_{m+1} are equal. At the left boundary, Δ_{med} is also zero. Δ_{med} appears to increase for fixed p_m as $p_{m+1} \downarrow 0$. The plot shows that Δ_{med} also appears to increase for fixed p_{m+1} near zero as $p_m \uparrow G(m+\tfrac{1}{2})$ but, for some $p_{m+1} > 0$, Δ_{med} does not increase uniformly as $p_m \uparrow G(m+\tfrac{1}{2})$ as shown in the upper left quadrant.

Figure 1

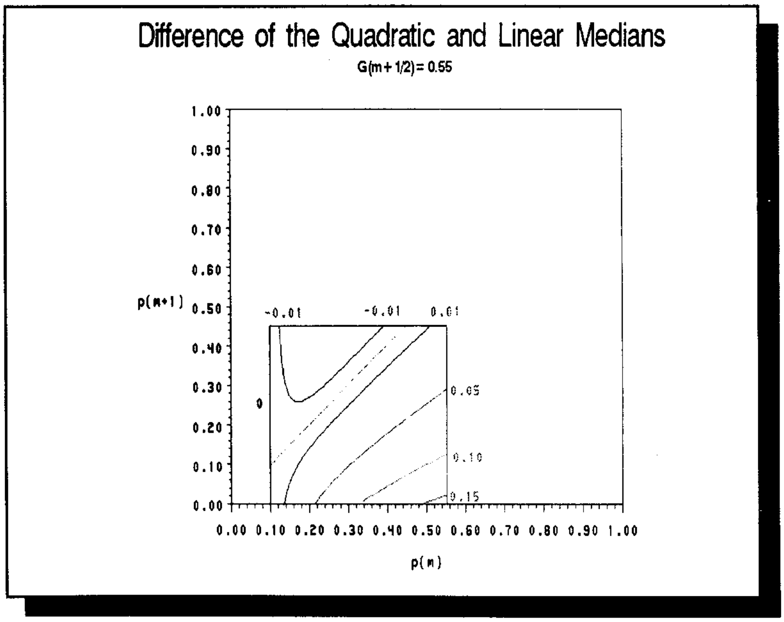


Figure 2 shows how the size of the feasible region decreases for larger G(m+½) and that Δ_{med} is bounded closer to zero.

Lemma 3 in the appendix shows, for all fixed G(m+½)>½ (Case A), that (a) Δ_{med} does increase for fixed p_m in the feasible region as $p_{m+1} \downarrow 0$ so that the maximum for each vertical line is at the bottom (p_{m+1}=0) of the slice of the feasible region, (b) for p_{m+1}=0, Δ_{med} increases uniformly as $p_m \uparrow$ G(m+½), (c) the relative maximum on this slice is at the lower left hand corner, positive, and

$$\max \Delta_{med} = \frac{3}{2} - \frac{\sqrt{G(m+½)(2\,G(m+½)-1)} + ½}{G(m+½)} \tag{4}$$

and (d) for p_{m+1}=0 the limit as G(m+½)↓½ of Δ_{med} equals ½ for all p_m where 0<p_m≤½.

Case B has comparable results for G(m+½)<½ where the signs are changed so that the relative minimum is negative and occurs at the upper right-hand corner of each slice. All the slices for Case B lie along the p_{m+1} axis.

Figure 2

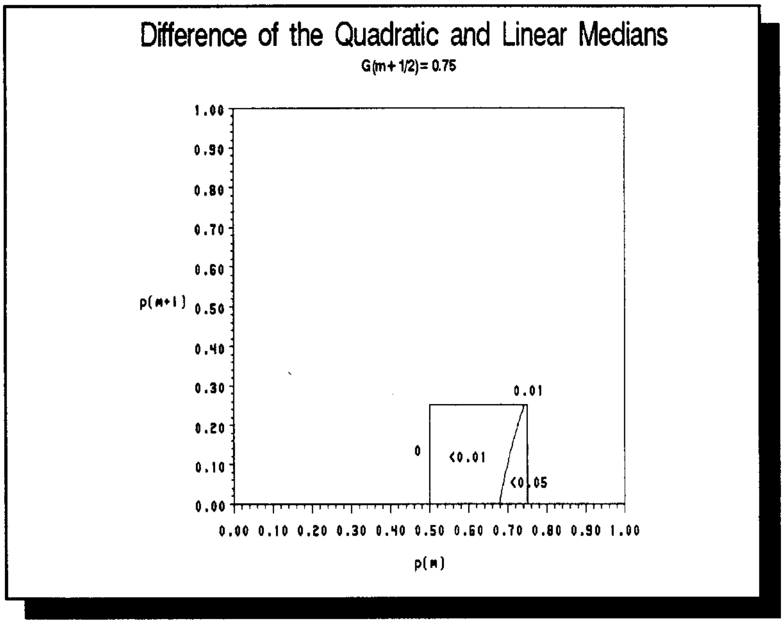
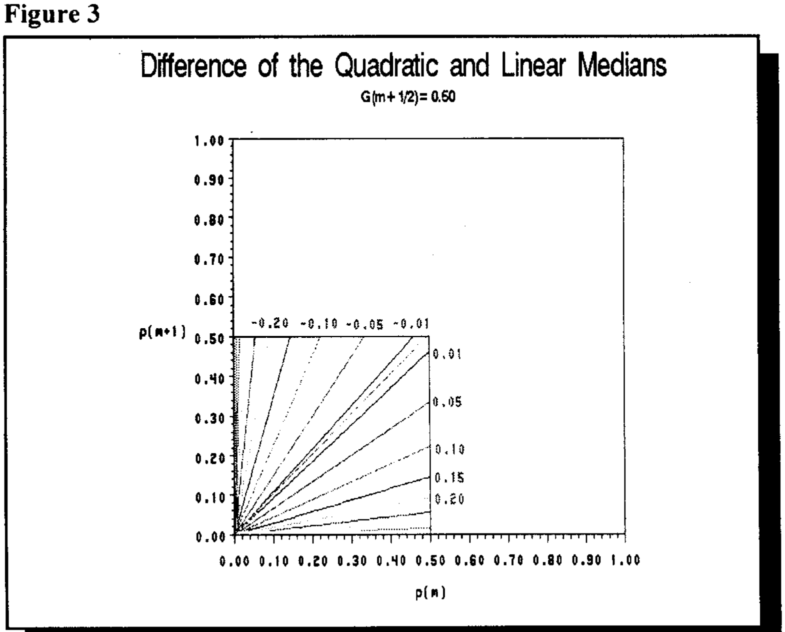


Figure 3



These minima and maxima are bounded by −½ and +½. Δ_{med}→+½ along the p_m axis as G(m+½)↓½ and Δ_{med}→−½ along the p_{m+1} axis as G(m+½)↑½. Figure 3 shows the contours for G(m+½)=½. These contours are linear arrays starting at the origin. The intervals between the unlabeled contours increase or decrease by 0.05. The contours for Δ_{med}>0.35 and Δ_{med}<−0.35 cannot be seen in this figure. It shows that the difference in the medians increases toward 0.5 as p_{m+1} approaches zero and that the difference decreases toward −0.5 as p_m approaches zero.

Simulation Analysis

We generated a sample of 100,000 points in the feasible region using SAS RANUNI. We first generated points uniformly in the unit cube and retained points in the feasible region until 100,000 points were reached.

Table 2 shows the frequency distribution for the absolute value of the difference in the medians. This table shows that most differences are near zero. Forty-four and a half percent of the differences are less than 0.01 and almost 85 percent are less than 0.05. The table also shows that very large differences are infrequent. This simulation found that out of 100,000 points only one difference was greater than 0.40 and only eight tenths of a percent larger than 0.20.

Table 3 shows examples of the difference in the medians for four distributions drawn from the Survey of Construction. The distribution in Table 1 is the first row in Table 3. This table shows that when p_m and p_{m+1} are nearly equal, examples one and three, that the difference in the medians is microscopic. When p_m and p_{m+1} are not nearly equal, the

Table 2. Absolute Difference of the Quadratic and Linear Medians

Abs. Dif. of Medians	Frequency	Percent	Cumulative Frequency	Cumulative Percent
$0 \le \lvert\Delta_{med}\rvert \le .01$	44,495	44.5	44,495	44.5
$.01 < \lvert\Delta_{med}\rvert \le .05$	39,054	39.1	83,549	83.5
$.05 < \lvert\Delta_{med}\rvert \le .10$	10,822	10.8	94,371	94.4
$.10 < \lvert\Delta_{med}\rvert \le .15$	3,518	3.5	97,889	97.9
$.15 < \lvert\Delta_{med}\rvert \le .20$	1,337	1.3	99,226	99.2
$.20 < \lvert\Delta_{med}\rvert \le .25$	494	0.5	99,720	99.7
$.25 < \lvert\Delta_{med}\rvert \le .30$	185	0.2	99,905	99.9
$.30 < \lvert\Delta_{med}\rvert \le .35$	70	0.1	99,975	100.0
$.35 < \lvert\Delta_{med}\rvert \le .40$	24	0.0	99,999	100.0
$.40 < \lvert\Delta_{med}\rvert \le .45$	1	0.0	100,000	100.0
$.45 < \lvert\Delta_{med}\rvert \le .50$	0	0.0	100,000	100.0

difference is still small. Two of the examples have $G(m+½)$ within 0.02 of 0.50 but all of the differences are still small because neither p_m nor p_{m+1} are close to zero.

The Survey of Construction publishes medians to only one decimal point. We explored the effect of the two methods of estimation on the published medians using the same precision as this survey. Table 4 shows the results from the simulation data set. About 76 percent of the time, there would be no difference in the published medians and nearly 98 percent would have a difference of either 0 or 0.1.

Discussion and Conclusions

In this study, we developed a new interpolation method for duration estimation based on sounder assumptions. We found that the difference between the new median based on quadratic interpolation and the previous median based on linear interpolation can be large and the difference is bounded by ±½. However, most of the differences are small and very few of them are large. Very large differences exist when one or both of the probabilities of the cells near the median is nearly zero.

Distributions, like those for the Survey of Construction durations, are convex with the mass of the probability in the central portion of the distribution. For these kinds of distributions, the cell probabilities near the median would very rarely be near zero so that a large difference between the quadratic and linear medians would be less likely than what was found by the simulation study. Because we expect much fewer significant differences than we found from the simulation, we did not recommend changing their procedures at this time.

Appendix

Lemma 2: There are no critical points of Δ_{med} in the interior of the feasible region.

Proof:

Define for case A and $p_m>0$, $a=(p_{m+1}-p_m)/p_m \neq 0$ and $b=(½(1+p_m)-G(m+½))/p_m$. As a function of a and b, $\Delta_{med}=a^{-1}[(1+2ab)^{½}-1]-b$.

For there to be a critical point (a, b) in the interior of the feasible region, $\partial\Delta_{med}/\partial a=0$ and $\partial\Delta_{med}/\partial b=0$. Both $\partial\Delta_{med}/\partial a=a^{-1}b(1+2ab)^{-½}-a^{-2}[(1+2ab)^{½}-1]$ and $\partial\Delta_{med}/\partial b=(1+2ab)^{-½}-1$ equal zero when $b=0$. At $G(m+½)>½$, $p_m=2G(m+½)-1$, the left boundary of the feasible region. When b is zero, Δ_{med} is zero along the left boundary. Since $\partial\Delta_{med}/\partial a \neq 0$ in the interior of the feasible region, there are no critical points in it. This shows the result for case A. By theorem 2, the result is also true for $G(m+½)<½$ (case B). For $p_m>0$, $p_{m+1}>0$ and $G(m+½)=½$, $b=½$ and both partial derivatives are not zero.□

Lemma 3: For case A, fixed $G(m+½)>½$,

(a) Δ_{med} increases uniformly for fixed $p_m>0$ as $p_{m+1}\downarrow 0$ so that the maximum for each vertical line is at the bottom ($p_{m+1}=0$) of the slice of the feasible region,

(b) for $p_{m+1}=0$, Δ_{med} increases uniformly as $p_m \uparrow G(m+½)$, (the relative maximum on this slice is at the lower left hand corner, positive and

Table 3. Examples of the Difference in the Medians

Case	m	$G(m+½)$	p_m	p_{m+1}	Quadratic Median	Linear Median	Difference of Medians
A	4	0.528426	0.074415	0.075055	4.117949	4.118009	-0.000060
B	0	0.442652	0.442652	0.167456	0.858841	0.842467	0.016373
B	7	0.471405	0.056401	0.057140	8.000431	8.000431	0.000000
B	1	0.483572	0.159352	0.091178	1.711237	1.680173	0.031154

$$\max \Delta_{med} = \frac{3}{2} - \frac{\sqrt{G(m+½)(2G(m+½)-1)} + ½}{G(m+½)},$$

(d) for $p_{m+1}=0$ the limit as $G(m+½) \downarrow ½$ of Δ_{med} is ½ for all p_m where $2G(m+½)-1 < p_m \le G(m+½)$.

Proof:

The proof will be for (a). The others are easy to show. It is sufficient to show that $\partial\Delta_{med}/\partial p_{m+1} \le 0$. The partial derivative is

$$\frac{\partial\Delta_{med}}{\partial p_{m+1}} = \frac{(p_m^2 + 2(p_{m+1}-p_m)(½(1+p_m) - G(m+½)))^{-½}}{(p_{m+1}-p_m)^2} \times \left[\begin{array}{l} p_m(p_m^2 + 2(p_{m+1}-p_m)(½(1+p_m) - G(m+½)))^{½} \\ -(p_m^2 + (p_{m+1}-p_m)(½(1+p_m) - G(m+½))) \end{array} \right]. \quad (5)$$

$$0 \le ½(1+p_m) - G(m+½) < ½(p_{m+1}+p_m) \quad (6)$$

follows from $F(m) \le ½ < F(m+1)$. For $p_{m+1}-p_m \ge 0$, it is easy to show that the discriminant,

$_{n+1} - p_m)(½(1+p_m) - ($

is positive. When $p_{m+1}-p_m<0$, multiplying (6) by 2 $(p_{m+1}-p_m)$ shows that the discriminant is positive.

$$p_{m+1}^2 < p_m^2 + 2(p_{m+1}-p_m)(½(1+p_m) - G(m+½)) \le p_m^2$$

It follows that $p_m^2 + (p_{m+1}-p_m)(½(1+p_m) - G(m+½))$ is positive.

$\partial\Delta_{med}/\partial p_{m+1} \le 0$.

It also follows easily from (6) that

$$p_m(p_m^2 + 2(p_{m+1}-p_m)(½(1+p_m) - G(m+½)))^{½} \le p_m^2 + (p_{m+1}-p_m)(½(1+p_m) - G(m+½))$$

by squaring each side. Since both sides are positive, the direction of the inequality will not change. Thus, the partial derivative is never positive.□

Table 4. Difference in Rounded Quadratic and Linear Medians (Rounded to a Tenth)

Difference in the Medians	Frequency	Cumulative Frequency	Percent	Cumulative Percent
0.0	75,940	75,940	75.9	75.9
0.1	21,643	97,583	21.6	97.6
0.2	2,157	99,740	2.2	99.7
0.3	239	99,979	0.2	100.0
0.4	21	100,000	0.0	100.0