Pearson's Rule for Sample Medians

by

George R. Terrell

Abstract

Pearson's rule \((\text{median-mode}) = 2(\text{mean-median})\) has long been observed to be approximately true for slightly skewed densities. Under mild conditions, the rule is asymptotically true for densities of sample medians from independently identically distributed random samples. This complements previous results that established the rule asymptotically for sample means.

Key words

median; mode; Central Limit Theorem.
Pearson’s Rule for Sample Medians

I. Introduction

In 1895 Karl Pearson noted that for what we would call gamma densities with large shape parameter, the relation (median-mode) = 2(mean-median) was approximately true. It has often been noticed since that the relation is often nearly true for slightly skewed unimodal densities. For example, the density of an $F$-statistic on $n$ and $n$ degrees of freedom is

\[ f_{n,n}(x) = \frac{1}{B\left(\frac{n}{2}, \frac{n}{2}\right)} \frac{x^{\frac{1}{2}(n-2)}}{(1+x)^n} \quad \text{for } x \geq 0. \]

Since the distributions of $z$ and $\frac{1}{z}$ are identical, we have $F(x) = 1 - F\left(\frac{1}{z}\right)$; so that for the median $m$, $\frac{1}{2} = F(m) = 1 - F\left(\frac{1}{m}\right)$. We conclude that $m = 1$. An easy calculation shows that the mean $= E(z) = \frac{n}{n-2} = 1 + \frac{2}{n} + O\left(\frac{1}{n^2}\right)$. A second calculation shows that the $F$-statistics are unimodal and the mode $= \frac{n-2}{n+2} = 1 - \frac{4}{n} + O\left(\frac{1}{n^2}\right)$. The asymptotic result median-mode - 2(mean-median) for large $n$ is immediate.

J.B.S. Haldane (1942) gave the first statistical explanation of the phenomenon. He noted that for random variables with an Edgeworth expansion through the third cumulant, independently and identically distributed samples had sample means that obeyed Pearson’s rule asymptotically for large sample size. Hall (1980) used characteristic function techniques to establish that sample means obeyed the rule so long as the underlying density possessed a third moment. It is interesting to notice that the $F$-statistic example (and many other examples of the phenomenon) is not in any obvious way connected to Hall’s result. An $F$-statistic is, after all, a ratio of chi-square random variables and not a mean of i.i.d. variables. The question of the real reason for Pearson’s rule is apparently still open.

Note that if $n$ is even (say, $= 2m$) in the $F$-statistic example, we have a density

\[ f_m(x) = \frac{1}{((m-1)!)^2} \frac{x^{m-1}}{(1+x)^{2m}}. \]

This is precisely the density of a sample median from an i.i.d. random sample from the density $f_1(x) = \frac{1}{(1+z)^2}$ for $x \geq 0$, where the sample size is $2m - 1$. In
this paper it will be proven under conditions I believe to be nearly as mild as possible that the distributions of sample medians from i.i.d. random samples follow Pearson's rule asymptotically for large sample sizes. Presumably such a result could be obtained by applying Bahadur-type (1966) projection techniques and then Hall's result. We believe that the proof in this paper is of interest because it is elementary (in contrast to Hall's use of characteristic functions) and rather direct (in contrast to projection techniques). Thus, our theorem and proof are nicely complementary to Hall's result.

II. The Theorem: Given a density \( f(x) \) such that (1) the sequence of sample medians of odd-size i.i.d. samples eventually possesses expectations; (2) the density is positive at the population median and continuous in a neighborhood of the median; and (3) the density is differentiable with nonzero derivative at the median, then the sequence of sample medians of odd-size i.i.d. samples has a local mode near its median and

\[
\lim_{n \to \infty} \frac{\text{median}(x_{(n+1)}) - \text{mode}(x_{(n+1)})}{\text{mean}(x_{(n+1)}) - \text{median}(x_{(n+1)})} = 2
\]

where \( x_{(n+1)} \) is the sample median from a sample of size \( 2n + 1 \).

We will proceed to outline the proof to emphasize its direct nature. First note that assumption (1) is necessary to guarantee that there are eventually means to check the rule for. A density may be too heavy-tailed to meet this condition, e.g., \( f(x) = \frac{1}{x \log x} \) for \( x \geq e \).

Proof:

Lemma 1: The population median of i.i.d. sample medians is just the population median of the underlying random variable.

Thus, the population medians of our sequence are equal to a constant value; call it \( m \). The rest of our argument will relate the distance of the mean and mode from the median to the quantity \( -f''(m)/f''(m) \); this is simply the second derivative of the inverse cumulative distribution function of the original random variable.
Lemma 2: Under conditions of the theorem, the expectation of the sample medians is asymptotically

\[ E(x_{(n+1)}) = m - \frac{1}{16n} \frac{f'(m)}{f^3(m)} \text{ as } n \to \infty. \]

The proof of this lemma will utilize a method due to Karl Pearson (1931); represent the expectation in terms of an integral involving the inverse cumulative distribution function, and then approximate that function by a Taylor's series.

Lemma 3: Under conditions (2) and (3) of the theorem, for sufficiently large \( n \) there exists a local maximum of the density of sample medians at asymptotically \( m + \frac{1}{8n} \frac{f'(m)}{f^3(m)} \).

The proof will involve approximately maximizing the density of a sample median by taking a single step of Newton's method with starting value \( m \).

A comparison of our three lemmas immediately establishes the theorem.

Q.E.D.

It should be noted that at no point did we require that \( n \) be an integer; thus our result applies to sample medians of fractional order statistic processes (Stigler, 1977). For example, our result applies to the case of \( F \)-statistics with equal but odd degrees of freedom.

It would be of interest to extend our results to order statistic sequences other than medians; generally, sequences \( \{x_{(i)}\} \) from samples of size \( n \) where \( \lim_{n \to \infty} \frac{i}{n} = p \), \( 0 < p < 1 \). The resulting distributions do follow Pearson's rule under conditions like the ones of this note. The analogues of our Lemma 2 and Lemma 3 are similar, but the proof of an analogue of Lemma 1 is somewhat more difficult. This will be addressed in a subsequent paper.

III. Proofs:

of Lemma 1:

Let \( F(x) \) denote the cumulative distribution function of the underlying random variable. Then the cumulative distribution function of an \( i^{th} \) order statistic from a sample of size \( m \) is
\[
F_{(i)}(x) = \text{Prob} \{ x_{(i)} \leq x \}
= \text{Prob} \{ \text{exactly } i \text{ sample values are } \leq x \} + \text{Prob} \{ \text{exactly } i + 1 \text{ samples are } \leq x \} + \cdots + \text{Prob} \{ \text{all } m \text{ samples are } \leq x \}
= \sum_{j=i}^{m} \text{binomial} (j; m, F(x)).
\]

Now let \( i = n + 1, m = 2n + 1, \) and \( F(x) = \frac{1}{2} \); so that \( x \) is the population median. Then
\[
F_{(n+1)}(x) = \sum_{j=n+1}^{2n+1} \text{binomial} (j; 2n + 1, \frac{1}{2}).
\]

But by the symmetry of the binomial,
\[
= \sum_{j=1}^{n} \text{binomial} (j; 2n + 1, \frac{1}{2}) = \frac{1}{2}
\]

since the two equal sums add to 1. Thus, \( x \) is the population median of the sample median.

Q.E.D.

of Lemma 2:

Let \( Q(y) \) be the quantile function of the underlying random variable. Since \( f \) is continuous and positive near the median, \( Q = F^{-1} \) near \( y = \frac{1}{2} \). Further, since \( f' (x) \) exists at the median, we have a Taylor’s series
\[
Q(y) = Q(\frac{1}{2}) + Q' (\frac{1}{2})(y - \frac{1}{2}) + Q'' (\frac{1}{2}) \frac{1}{2} (y - \frac{1}{2})^2 + (y - \frac{1}{2})^2 \epsilon(y - \frac{1}{2})
\]
where \( \epsilon(x) \) is continuously equal to zero at \( z = 0 \). Let \( g_n(y) = \frac{(2n + 1)!}{n! n!} [y(1-y)]^n \) be the symmetric beta densities. Our condition that the expectations of the medians exist for all \( n \geq N \) says that
\[
E(x_{(n+1)}) = \int x \frac{(2n + 1)!}{n! n!} f(x) F(x)^n (1-F(x))^n \, dx
= \int Q(y) \frac{(2n + 1)!}{n! n!} y^n (1-y)^n \, dy = \int Q(y) g_n(y) \, dy.
\]

Since \( Q \) is monotone and this integral exists, our Taylor’s series says that for all \( n \geq N \),
\[
\int_0^1 (y - \frac{1}{2})^2 |e(y - \frac{1}{2})| g_n(y) \, dy
\]
exists. Thus

\[
E(x_{(n+1)}) = \int_0^1 g_n(y) Q(y) \, dy
\]

\[
= Q(\frac{1}{2}) + \frac{1}{8(2n + 3)} + \int_0^1 g_n(y) (y - \frac{1}{2})^2 e(y - \frac{1}{2}) \, dy
\]

by standard facts about beta distributions. It is sufficient to show that for \( n \) sufficiently large, the last term is small compared to \( n \). By the continuity of \( e \) at 0, for any \( \epsilon > 0 \), there exists \( \delta > 0 \) such that \( |e(y - \frac{1}{2})| < \epsilon \) whenever \( |y - \frac{1}{2}| < \delta \). For that choice of \( \epsilon \),

\[
\int_0^1 g_n(y) (y - \frac{1}{2})^2 e(y - \frac{1}{2}) \, dy = \int_{|y - \frac{1}{2}| < \delta} g_n(y) (y - \frac{1}{2})^2 e(y - \frac{1}{2}) \, dy
\]

\[
+ \int_{\delta < |y - \frac{1}{2}| < \frac{1}{2}} g_n(y) (y - \frac{1}{2})^2 e(y - \frac{1}{2}) \, dy.
\]

Now

\[
\left| \int_{|y - \frac{1}{2}| \leq \delta} g_n(y) (y - \frac{1}{2})^2 e(y - \frac{1}{2}) \, dy \right| \leq \epsilon \int_{|y - \frac{1}{2}| \leq \delta} g_n(y) (y - \frac{1}{2})^2 \, dy
\]

\[
\leq \epsilon \int_0^1 g_n(y) (y - \frac{1}{2})^2 \, dy \leq \frac{\epsilon}{4(2n + 3)}
\]

for any \( n \). Also,

\[
\left| \int_{\delta < |y - \frac{1}{2}| < 1} g_n(y) (y - \frac{1}{2})^2 e(y - \frac{1}{2}) \, dy \right|
\]

\[
= \left| \int_{\delta < |y - \frac{1}{2}| < 1} \frac{g_n(y)}{g_N(y)} g_n(y) (y - \frac{1}{2})^2 e(y - \frac{1}{2}) \, dy \right|
\]

\[
\leq \max_{\delta < |y - \frac{1}{2}| < \frac{1}{2}} \frac{g_n(y)}{g_N(y)} \int_{\delta < |y - \frac{1}{2}| < \frac{1}{2}} g_N(y) (y - \frac{1}{2})^2 e(y - \frac{1}{2}) \, dy
\]

\[
\leq \max_{\delta < |y - \frac{1}{2}| < \frac{1}{2}} \frac{g_n(y)}{g_N(y)} \int_0^1 g_N(y) (y - \frac{1}{2})^2 e(y - \frac{1}{2}) \, dy.
\]
The integral exists and does not depend on $n$.

$$\max_{\frac{1}{2} - \frac{1}{2} < y < \frac{1}{2} - \frac{1}{2}} \frac{g_n(y)}{g_N(y)} = \max_{\frac{1}{2} - \frac{1}{2} < y < \frac{1}{2} - \frac{1}{2}} \frac{(2n + 1)!}{n! n!} \frac{(2n + 1)!}{N! N!} \left[ y(1 - y) \right]^{a-N}.$$ 

Since the polynomial piece increases below $\frac{1}{2}$ and decreases above $\frac{1}{2}$,

$$= \frac{(2n + 1)!}{n! n!} \left[ \frac{1}{4 - 2^2} \right]^{a-N}.$$

The combinatorial piece in $n$ is, by Stirling's formula, bounded above by a constant times $2^{2a} n^\frac{1}{2}$, so

$$\leq kn^\frac{1}{2} 2^{2a} \left[ 1 - 4 \left( \frac{1}{2 \sqrt{2}} \right)^a \right]^{a-N} 2^{-2a} < \frac{\xi}{n}$$

for $n$ sufficiently large, since the expression shows exponential decrease in $n$. Now

$$Q'(\frac{1}{2}) = \frac{-f'(Q(\frac{1}{2}))}{f^3(Q(\frac{1}{2}))} = -\frac{f'(m)}{f^3(m)}.$$ 

Therefore, our Taylor's series becomes

$$E(x_{x+1}) = m - \frac{f'(m)}{16nf^3(m)} + o\left(\frac{1}{n}\right)$$

which is Lemma 2.

Q.E.D.

of Lemma 3:

We are searching for a local maximum of the density $f(x)[F(x)/(1-F(x))]^n$. $F^{-1}(y) = Q(y)$ exists and is continuous near $m$, so what we have written is equivalent there to maximizing $f(Q(y))[F(Q(y))/(1-F(Q(y))]^n = f(Q(y))[y(1-y)]^n$ which is equivalent to maximizing the logarithm $\log f(Q(y)) + n \log y(1-y)$. Let us expand in a Taylor's series
\[
\log f(Q(y)) = \log f(m) + (y - \frac{1}{2}) \frac{f'(m)}{f^2(m)} + (y - \frac{1}{2}) e(y \frac{1}{2}) \\
+ n \log(y(1-y)) = n \log \frac{1}{4} + (-4n (y - \frac{1}{2})^2 + n (y - \frac{1}{2})^4 g(y - \frac{1}{2}).
\]

Since we will be investigating a neighborhood of \(m\), the \(g\) term (continuous at zero) may be incorporated in the \(e\) term (continuously zero at zero). Denote \(z = 2(y - \frac{1}{2})\) and \(a = \frac{f'(m)}{2f^2(m)}\).

We find ourselves searching for a maximizer of \(g(z) = (a + \hat{e}(z))z - nz^2\), where elementary algebra predicts an answer near \(\frac{a}{2n}\). For any \(\eta > 0\) small compared to \(a\), we will establish that there is a maximizer in the interval \(\left(\frac{a - \eta}{a_n}, \frac{a + \eta}{2n}\right)\) for sufficiently large \(n\). Choose an \(\epsilon > 0\) small compared to \(\frac{\eta^2}{a}\) and to \(\eta\). Then by the continuity of \(\hat{e}\), there exists a \(\delta > 0\) such that \(|z| < \delta\) implies \(|\hat{e}| < \epsilon\). Now make \(n\) large enough that \(\frac{a}{n} < \delta\) and \(f(Q)\) is continuous in the neighborhood \(|y - \frac{1}{2}| < \frac{a}{2n}\). Without loss of generality, consider the case \(a > 0\). Then for \(z's\) in the positive half of the neighborhood \(0 < z < \frac{a}{n}\), we have

\[
(a - \epsilon) z - nz^2 \leq g(z) \leq (a + \epsilon) z - nz^2.
\]

Calculations yield

\[
g \left(\frac{a - \eta}{2n}\right) \leq \frac{a^2 + 2a \epsilon - 2n \epsilon - \eta^2}{4n}
\]

\[
g \left(\frac{a}{2n}\right) \geq \frac{a^2 + 2a \epsilon}{4n}
\]

\[
g \left(\frac{a + \eta}{2n}\right) \leq \frac{a^2 + 2a \epsilon + 2n \epsilon - \eta^2}{4n}
\]

We have chosen \(\epsilon\) small compared to \(\frac{\eta^2}{a}\) and \(\eta\), so that for sufficiently large \(n\),

\[
g \left(\frac{a}{2n}\right) > g \left(\frac{a - \eta}{2n}\right) \quad \text{and} \quad g \left(\frac{a}{n}\right) > g \left(\frac{a + \eta}{2n}\right).
\]

Thus, by the continuity of \(g\) over \(|z| < \frac{a}{n}\), we conclude that \(g\) has a local maximum in

\(\left(\frac{a - \eta}{2n}, \frac{a + \eta}{2n}\right)\). Going back to our definitions, we conclude that \(y = \frac{1}{2} + \frac{f'(m)}{8f^2(m)n}\) is asymptot-
cally a maximum of \( f(Q(y))[y(1-y)]^n \). But then \( Q\left(\frac{1}{2} + \frac{f'(m)}{8f^3(m)n}\right) \) is a local maximizer asymptotically of \( f(x)[F(x)(1-F(x))]^n \). But \( Q \) is differentiable near \( \frac{1}{2} \), and so

\[
Q(y) = Q\left(\frac{1}{2}\right) + \frac{y - \frac{1}{2}}{f(Q(\frac{1}{2}))} + \left(\frac{1}{2} - y\right) \epsilon(y - \frac{1}{2})
\]

where \( \epsilon \) is continuously zero at zero. We conclude that there is asymptotically a mode of the sequence of sample medians at

\[
Q\left(\frac{1}{2} + \frac{f'(m)}{8f^3(m)n}\right) = m + \frac{f'(m)}{8f^3(m)n} + o\left(\frac{1}{n}\right).
\]

\text{Q.E.D.}
IV. Bibliography


