1 Premature Death in Jazz Musicians: Fact or Fiction?

letter from retired prof. to Amer. J Public Health 1991 June; 81(6): 804-805

Statistical study of 86 jazz musicians listed in a university syllabus refutes these tenets,[4] the second and third of which were made by two of America's most respected critics, and all of which foster the commonly held view that jazz players die prematurely. Dates of birth, and of death when it had occurred, were tabulated, and longevity matched with that expected in the United States by year of birth, race, and sex.(5–7) One musician who had not reached the age of his life expectancy was excluded from the list; the musicians were born in the US.

Birth years ranged from 1862 to 1938; 16 births occurred before 1900, 23 between 1900 and 1909, 19 between 1910 and 1919, 22 between 1920 and 1929, and five between 1930 and 1939. Comparison with national values showed that 70 (82%) of the musicians exceeded their life expectancy; four-fifths of the Black men, three fourths of the White men, and all the women lived longer than expected as shown in this frequency distribution.

	Male			Female		
	Total	n	%	Total	n	%
White	19	14	74	-	-	-
Black	59	49	83	7	7	100

i. "Comparison with national values showed that 70 (82%) of the musicians exceeded their life expectancy". In the Canadian lifetable used in last week's exercise, what percentage of 100000 newborns would be expected to exceed the life expectancy at birth? Comment![1]

ii. What is the shape of the distribution of the ages-at-death in that lifetable? Does this explain how, as in Lake Wobegon, `http://en.wikipedia.org/wiki/Lake_Wobegon_effect`, more than 50% can genuinely be "above average"?[2]

[1] JH suspects that the author, if asked, would have argued: 'under the null hypothesis, the expected percentage that exceeded the life expectancy at birth should be 50%.'

[2] The editor of Amer. Journal of Roentgenology missed this point in JH's 1994 article, as did the British newspaper *The Independent* when it wrote "The usually wonderful Jeremy Paxman, introducing a Newsnight discussion last Friday on the teaching of reading skills, expressed dismay that 'a third of our primary schoolchildren have below-average reading ability'. Had he paid more attention in his 'rithmetic lessons, perhaps Paxman would have realised that half our schoolchildren are below average in everything. As, indeed. are half our Newsnight presenters."

2 Clinical Research in General Medical Journals: A 30-Year Perspective

Table from Fletcher et al., NEJM 301: 180-183, 1979

Fletcher et al. studied the features of 612 randomly selected articles published in the NEJM, JAMA and Lancet since 1946. Two features were the number of authors per article and the number of subjects studied in each article.

Year	No. articles examined	No. authors Mean	(SD)	No. Subjects Median
1946	151	2.0	(1.4)	25
1956	149	2.3	(1.6)	36
1966	157	2.8	(1.2)	16
1976	155	4.9	(7.3)	30

i. From the mean and SD, *roughly* reconstruct the actual frequency distribution of the no. of authors per article for 1946 [Excel or `R` can help]

ii. Can the 1976 SD of 7.3 really be larger than the mean of 4.9? Explain.

3 Statistics in the U.K. Parliament

Hansard 29 November 1991

Mr. Arbuthnot: the Labour party's suggestion of a minimum wage is in itself rather obscure and bizarre. As I understand it, it is tied to the average and would therefore not only be relatively high at £3.40 but would increase as the average wage itself increased. With each increase in the average rate of pay, the minimum wage itself would have to go up and it would be forever chasing its own tail.

Mr. Tony Lloyd: Perhaps I can help the hon. Gentleman. It will be tied to the median, which is not the same as the average. It is simply a mid-point on the range and would not be affected by changes in the minimum wage.

Mr. Arbuthnot: From what I understand, even an amount tied to the median would be affected because if the lowest wage were increased to £3.40 per hour, the median would have to rise.

Mr. Tony Lloyd: I shall put the matter in simple terms. The median, the mid-point in a series of numbers such as 2.2, 5.6 and 7, is defined as being the difference between 2 and 7, which is 3.5. If we alter the figures 2 and 2 to 3.5, the middle figure of 5 would remain unaltered because it is independent of the bottom figures.

Mr. Arbuthnot: I do not understand the hon. Gentleman's mathematics and I slightly doubt whether he does.

Mr. Matthew Carrington (Fulham): I am extremely confused. I studied mathematics for some years at school and I have not totally forgotten all of them. The median is not the mid-point between the first number and the last. It is where the largest number of items in a sample comes to, whereas the average is obviously the sample multiplied by the number of items. The hon. Member for Stretford (Mr Lloyd) is obviously extremely confused. The median has a precise mathematical definition which is absolutely right, and my hon. Friend is correct in saying that the median is bound to alter if the number at the bottom on the scale is changed. That will alter the average as well in a different way, but it is bound to alter the median. Perhaps the hon. Member for Stretford wishes to define median in a non mathematical sense.

Mr. Arbuthnot: I am grateful to my hon. Friend for sorting out at least the hon. Gentleman's mathematics with obvious skill and knowledge.

Exercise: Correct the honourable Gentlemen

4 The mean absolute deviation of a set of values, or of a random variable, is minimized when these deviations are measured from the median.

For the questions below, the 'elevator' applet – and article "Visualizing the Median as the Minimum-Deviation Location" – under Articles in the resources page might provide some insights.

Denote the random variable by Y, its p.d.f. by $f(y)$, the median (or in the indeterminate case, *any*[3] median value) by $Y_{0.5}$, and another proposed "central" location by Y_P, located at the P-*th* percentile of Y ($P \neq 0.5$).

i. Consider first the central location $Y_{0.5}$. Denote the Mean Absolute Deviation from this location as $M.A.D_{Y_{0.5}}$. Now re-locate the proposed 'center' to the right of $Y_{0.5}$, at the P-*th* percentile ($P > 0.5$) of Y. By doing so, one has moved a certain distance *further away from* a proportion of the values, and this same distance *closer to* the remaining proportion of the values. Use this argument to show that by moving the proposed central location to the right [or to the left] of $Y_{0.5}$, one is – overall, or on average – further away from the values, i.e., that

$$M.A.D_{Y_P} > M.A.D_{Y_{0.5}} \text{ for } P \neq 0.5.$$

ii.
- Suppose that Y takes on discrete values, and that $Y_{0.5}^{<}$ is the largest discrete Y_P value such that $P < 0.5$, and $Y_{0.5}^{>}$ is the smallest discrete Y_P value such that $P > 0.5$. Show that any real value between $Y_{0.5}^{<}$ and $Y_{0.5}^{>}$ serves as a 'M.A.D' location.
- When a sample consists of an even number ($n = 2m$) of observations, the usual definition of the sample median is $(y_{[m]} + y_{[m+1]})/2$. In light of your result, suggest another definition.

iii. For $P > 0.5$, Cramér (1946, pp. 178-179) used the relation

$$E(\mid Y - Y_P \mid) = E(\mid Y - Y_{0.5} \mid) + 2\int_{Y_{0.5}}^{Y_P} (Y_P - y) f(y) dy$$

to show that the first absolute moment $E(\mid Y - Y_P \mid)$ becomes a minimum when $P = 0.5$. Fill in the details in Cramér's proof.

[3] As will be seen in (iii), .

5 Inference concerning median – and other quantiles

http://en.wikipedia.org/wiki/Quantile

As in the previous question, denote the random variable by Y, its p.d.f. by $f(y)$, and the $(100 \times P)$-*th* percentile by Y_P. Let $y_1, \ldots, y_n$ be n realizations (sample values) of Y, and let $y_P = y_{[100nP]}$ denote the corresponding sample quantile (assume that n is large enough that $100nP$ is close to an integer). Using the (asymptotic) distribution theory for order statistics, one can show that

$$E(y_P) = Y_P; \qquad Var(y_P) = \frac{P(1-P)}{n \times \{f(Y_P)\}^2} \ .$$

i. Calculate the sampling variance of the sample median ($P = 0.5$) for the situation where
(a) $Y \sim N(\mu,\ \sigma = 1)$,
(b) $Y \sim \Gamma(\mu = 20,\ \sigma = 10)$.

ii. Calculate the sampling variance of the first and third quartiles of the sample ($P = 0.25, 0.75$) when $Y \sim Beta(\alpha = 1/3,\ \beta = 2/3)$

iii. How could one use these results to form an approximate (Gaussian-based) CI for Y_P?

iv. Perform a Monte Carlo simulation to investigate how 'Gaussian' are the sampling distributions of the (a) $50th$ and (b) $90th$ order statistics from a sample of $n = 100$ from (1) $Y \sim N(\mu,\ \sigma = 1)$ (2) $Y \sim \Gamma(\mu = 20,\ \sigma = 10)$. `R` code for these is available under Resources.

v. For the $N(\mu, \sigma)$ case, $\bar{y}$, the sample mean, is the most efficient estimator of μ (in the sense that no other unbiased statistic for estimating it can have smaller variance). Consider the sample *median* as an estimator of μ. Calculate the efficiency of the sample median, measured as the ratio of the variance of the sample mean to the variance of the sample median.

6 Who is more variable (relatively speaking)?

i. a younger child who consumes 887, 672, 757, 867, 899 and 872 calories on 6 observed days or

ii. an older child who consumes 1155, 1193, 1167, 1315, 1401 and 1133?

7 Describe the students in EPIB606/EPIB607 in 2001

In 2001 JH and DM carried out a web-based questionnaire survey to help the course instructors know better the background of the students taking this course, and their potential instructional needs. The data from this survey are available under (Web-based) Survey of students in McGill intro Epi and Biostat courses, 2001, URL `http://www.biostat.mcgill.ca/hanley/bios601/Surveys`.

i. Prepare[4] (in LateX or a word processor or in your own handwriting) a Table entitled "Demographic and Educational Background of Students in EPIB606/EPIB607 in 2001" that gives a 'statistical picture' of these students, separated by sex, using 8-10 variables.[5]

ii. Prepare[6] (in `R` or `SAS` or in your own handwriting) a Figure entitled "Relationship between .. " that shows the relationship (or lack of one) between two variables in the dataset.[6]

[4] You may hand in 1 set of answers for a pair of you.

[5] To find examples to follow, I suggest you Google with search words such as Table 1 characteristics subjects/patients. Or you might want to browse through a specific medical journal such as CMAJ (`http://www.cmaj.ca`) or BMJ (`http://www.bmj.com/`, or look at websites of Statistics Canada or the Public Health Agency of Canada, or ISQ or RAMQ, or NCHS or CDC. I will be asking you separately to bring to the Lab session one example of a clear and well-made Table, and a clear Figure, as well as a table and a figure at the other end of the scale!

[6] the CDC site `http://www.cdc.gov/mmwr/mguide_qs.html` has some nice simple graphs.