Res. Lett. Inf. Math. Sci. (2003) **4**, 55-65
Available online at http://iims.massey.ac.nz/research/letters/

Analysis Of The Last Page

Barry McDonald,
I.I.M.S., Massey University Albany Campus, Auckland, N.Z.
b.mcdonald@massey.ac.nz

Abstract

A sample of death notices from the New Zealand Herald was used as the basis of a Data Analysis assignment. This note explores some interesting statistical aspects of these death notices, using common data analysis techniques, and illustrates how they can be used as a resource for teaching. In particular they provide a clear example of biased sampling, a concept that is usually hard to quantify.

1. Introduction

The analysis of obituary data has a long history, dating back at least as far as Graunt, 1662 [2]. This paper examines newspaper death notices, from the perspective of a student of human behaviour. This study was motivated a pragmatic need for some general-interest data for teaching and examination purposes. Since obituary data is readily available, a study of death notices might make a suitable high-school project. Or a sociology or history student may wish to look for changes in the ways New Zealanders have expressed themselves through death notices through past decades - a simple research project since old newspapers are readily available on microfiche in public libraries. This paper raises some issues that could be considered in such studies. The data analysed in this paper are available on request.

2. Data collection

The study began as a prospective analysis of all death notices, published usually on the last page of *The New Zealand Herald* [6] for people recorded as having died during a two-week period in 1995. This yielded records of 498 separate individuals, of whom 17 (3.4%) were identified as Pacific Islanders (generally by name, place of origin, and/or church affiliation for example "PIC") and another 57 (11.4%) were identified as Maori. Usually the distinction between Maori and Pacific Islander was clear, for example by mention of the marae where the tangi would be held, but in ambiguous cases the deceased was arbitrarily classified as Maori. Although the label is not exact, in what follows Maori and Pacific Islanders may be referred to together as 'Polynesian': the remaining 424 individuals (85.1%) are thus assigned to the majority 'non-Polynesian' population - that is persons predominantly of European or Asian descent. To allow the study of ethnic differences, the sample was augmented by all identifiable Maori or Pacific Islander deaths for an extra six weeks. This brought the total to 220 Maori and 67 Pacific Islanders.

The vast majority of death notices were published within 3 days of death, but the newspaper was checked for a further two weeks. Only notices from the *Deaths* column were used, excluding *Funeral Notices* (inserted mainly by Masonic lodges), *Bereavement Notices* and *In Memorium*. Data was collected for each individual on: the name of the deceased (for data matching), the date of death (if given), the age at death (if given), the gender (if given), the ethnicity (as determined above), the number of death notices, the number of notices recording the age, and the first and last date of notification. There was also an indicator for whether the

age was given as 'aged x years' or 'in his/her $x+1$ th year'. All analyses were performed using MINITAB [3].

3. Number of death notices

All told there were 2650 death notices, for a total of 711 individuals, giving an average of 3.73 per person. However the distribution was extremely right-skewed, with 37% of deaths notified only once, 18% notified twice (usually the same notice inserted for a second day), but one person with 60 notices - see Table 1. The number of death notices probably reflects the cultural values of that segment of society to which the person belonged - in particular the extent to which open display of grief is encouraged - as well as personal characteristics of the deceased, such as their age, family position, and the suddenness of the death.

Table 1. Number of death notices

Number	Count	Percent	CumPct	Number	Count	Percent	CumPct
1	261	36.71	36.71	14	4	0.56	97.19
2	29	18.14	54.85	15	2	0.28	97.47
3	88	12.38	67.23	16	3	0.42	97.89
4	53	7.45	74.68	17	3	0.42	98.31
5	40	5.63	80.31	18	2	0.28	98.59
6	33	4.64	84.95	19	3	0.42	99.02
7	16	2.25	87.20	20	2	0.28	99.30
8	17	2.39	89.59	22	1	0.14	99.44
9	21	2.95	92.55	23	1	0.14	99.58
10	7	0.98	93.53	30	1	0.14	99.72
11	10	1.41	94.94	35	1	0.14	99.86
12	4	0.56	95.50	60	1	0.14	100.0
13	8	1.13	96.62				

N= 711

Along these lines, one can check for ethnic differences in the number of notices submitted. A nonparametric one-way ANOVA suggests that Pacific Islanders tend to submit more notices than Maori or non-Polynesians ($p=0.041$). An approximate analysis available to undergraduate students consists of first taking the logarithm of the number of notices - in the hope of obtaining approximate Normality - and then applying standard ANOVA. A normal probability plot of the log(number of notices) (Figure 1) shows a reasonably straight line, although the discrete nature of the data means the accompanying Normality test statistic is highly significant ($p=0.000$). If we nonetheless ignore the Normality issue, a one-way ANOVA again gives a significant ethnic effect ($p=0.031$). It is left as an exercise for graduate students to confirm or deny this effect using a generalized linear model. At least there seems to be some suggestion that Pacific Islanders make more use of this (essentially European!) method of expressing grief than other ethnic groups. A sociologist might speculate on reasons for this: the writer will simply comment that the notices were also frequently very long, mentioning scores of mourners, so that the effect seems to be real, not an artefact of data-dredging.

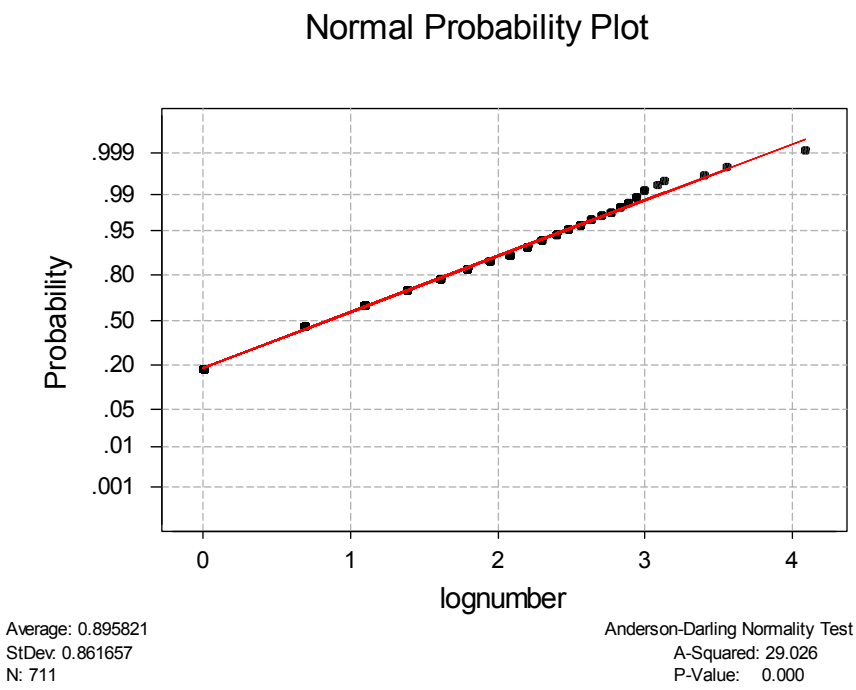


Figure 1. Normal Plot

4. The 'delay' before publication

A curious piece of statistical trivia was the small (but statistically significant) difference between the average delay (in days) before male and female deaths were reported. This might be regarded as just chance if it were not consistent across ethnic groups, as depicted in the confidence intervals for mean delay in Figure 2. A Two-Way ANOVA in log(delay) showed both main effects for Gender (p=0.007) and Ethnicity (p=0.000) of the deceased were significant. The extra day's delay for Pacific Islanders probably reflects notices placed for people in the Islands. The faster reporting of female deaths might be the result of confounding with age - the average delay decreased by about one day over the range of ages from 0 to 102.

5. The sample age distribution

5.1 Bias in reporting the age?

A more interesting issue from a statistical point of view is whether the published ages are an unbiased sample for the population. Putting it another way, if some archaeologist of the distant future were to study our newspapers, would (s)he get an accurate understanding of New Zealanders lifetimes? The problem is that we only have limited data. Of the 2650 death notices, 544 (20.5 %) mentioned the deceased's age. Usually this was just the first notice, but occasionally several notices for the same person would mention the age - up to 15 times. All told we have definite information on the age at death for 337 people out of 711 deaths, or 47% of the total. The probability of recording the age *at least once* seems to be independent of gender, but a chi-square test based on Table 2 suggests a relationship to the ethnicity of the deceased (p=0.0093). That is, Maori are less likely to mention the deceased's age than the majority non-Polynesian population, with Pacific Islanders in the middle.

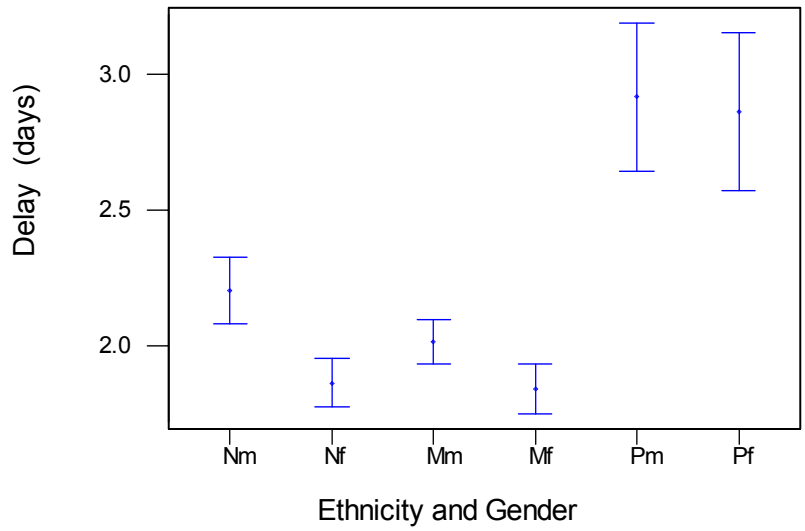


Figure 2 Analysis Of Effect Of Gender And Ethnicity On Reporting Delays
95% Confidence intervals for Mean Reporting Delay
Nm=non-Polynesian male, Nf=non-Polynesian female, Mm=Maori male, Mf=Maori female, Pm=Pacific Island male, Pf=Pacific Island female.

It may be a matter of sociological interest why Maori (and to a lesser extent Pacific Islanders) mention the person's age less frequently than non-Polynesians. Some factors to consider may be a non-western attitude to time (for example, that it is much less important to mention the age than the fact the deceased was a grandparent with many mokopuna); less importance placed on the notice as a record of family history; and also the importance of the tangi, where many aspects of the deceased's life would be discussed over several hours or days.

Table 2: Ethnic differences in whether or not deceased's age is mentioned

Observed (Expected)	Maori	Pacific Islander	non-Polynesian	Total
Age not mentioned	133 (115.7)	34 (35.2)	207 (223.0)	374
Age mentioned	87 (104.3)	33 (31.8)	217 (201.0)	337

It is interesting to consider whether the probability of recording the age *depends on* the age. For example, if the person is extremely old, is it more likely that the age will be quoted, either because it is traditional or as if to gain comfort by the thought '(s)he had a good innings'? Conversely is it less likely that the age will be quoted, out of respect for an elder, or since fewer people may know the deceased anyway, and they would tend to know the age? At the other end of the scale, could the tragic effect of young lives lost make it more likely that the deceased's age would be quoted? Of course there is no straightforward answer to these questions since we do not know the age of those for whom it was not reported. But there are two indirect ways of shedding light on the question.

5.2 Looking for evidence within the data

The first way, just using the available data, is to analyse how *often* the age is quoted, for those for whom it is quoted at least once. A rationale is that if some sentiment (such as a sense of tragedy) *causes* the notifier to mention the age once, then the same sentiment may lead to the age being quoted again in several notices. (Of course this effect will be confounded if the same notice is inserted on, say, three successive days.) Figure 3 shows the relationship between the number of times age was mentioned and the age itself (if given). There is clear heterogeneity between younger people (under 50) and older people: in particular there were 10 mentions of age for a four-year old and 15 mentions of age for a 22-year old. The jagged polygon marks the running mean of nine counts, while the dashed line is a smoothed version. The graph suggests the mean number of mentions drops by about 1.0 over the age range, as does a regression. So this approach is *suggestive* that the age is likely to be mentioned more often for younger people than for older. A student exercise might examine whether a model that corrects for heterogeneity would also remove the trend.

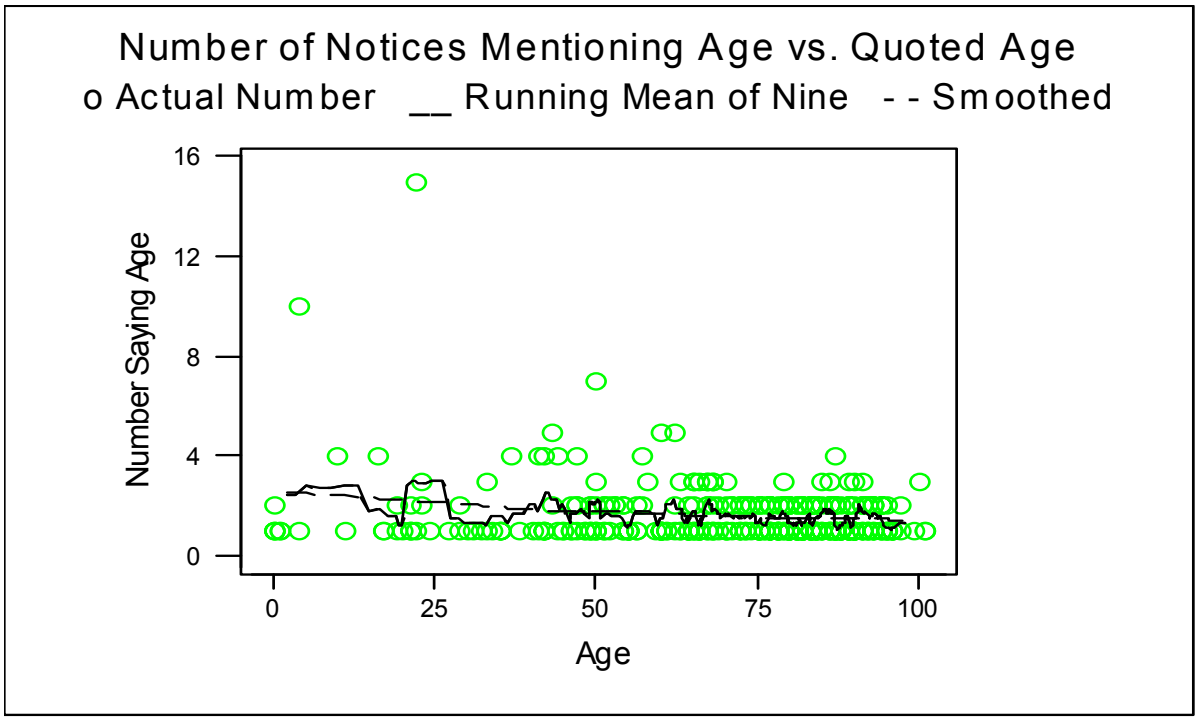


Figure 3. Mentions of Age

The MINITAB output in Figure 4 gives stem and leaf plots of the sample age distribution for males and females. To the uninitiated eye these plots do not say much except that there were quite a few deaths among young males, especially in infancy and in the 'accident hump' years of 17-24. As one who had not collected real demographic data before, the writer was at first struck by how elderly most of the deceased in the newspaper were. After all, most deaths one seems to hear about are people who died young. However the media tend to focus on the tragic, and one's memory also, which can give a biased impression of the age distribution. It seems then that little progress can be made without comparing the sample to an objective standard, which we consider next.

4(a) Males

Stem-and-leaf of Age sex = 1 N = 182
Leaf Unit = 1.0 N* = 192

5	0	00014
10	1	01779
21	2	01111233479
28	3	0335558
45	4	01223445667777889
63	5	000112234445555568
(30)	6	000001223444455556667777888899
89	7	000012222334556677777788888999999
57	8	000000011111222222233333444666667777777889999
13	9	0011233445566

4(b) Females

Stem-and-leaf of Age sex = Females N = 156
Leaf Unit = 1.0 N* = 179

3	0	014
5	1	69
7	2	29
11	3	1247
21	4	1223569999
33	5	000113466779
52	6	0001233445566667788
72	7	01133445555556777899
(51)	8	000000111222233333333344445555666677777888889999999
33	9	00000011112222233344445556679
3	10	011

Figure 4 Stem-and Leaf Plots of Recorded Age For Death Notice Data

5.3 Using information from the New Zealand Life Tables

We can compare the newspaper data with what we know about the population from the '*New Zealand Life Tables 1995-97'* [5] . These tables provide smoothed estimates of the population and death rates for each year of age, for males and females; and for Maori, non-Maori, and the combined population. So our problem becomes that of having a sample of ages, and seeing whether the sample is biased for this population. Biased sampling problems occur frequently in econometric and other settings: see, for example, [1,4] where it is referred to as choice-based sampling. In some contexts one may wish to correct for bias, but here we content ourselves with detecting and assessing it.

In the first place one can compare means. The sample mean age at death was only 63.4 years for males and 73.4 years for females. This is much younger than the (population) mean life expectancy for New Zealanders, which was estimated at 74.2 years for males and 79.6 years for females.

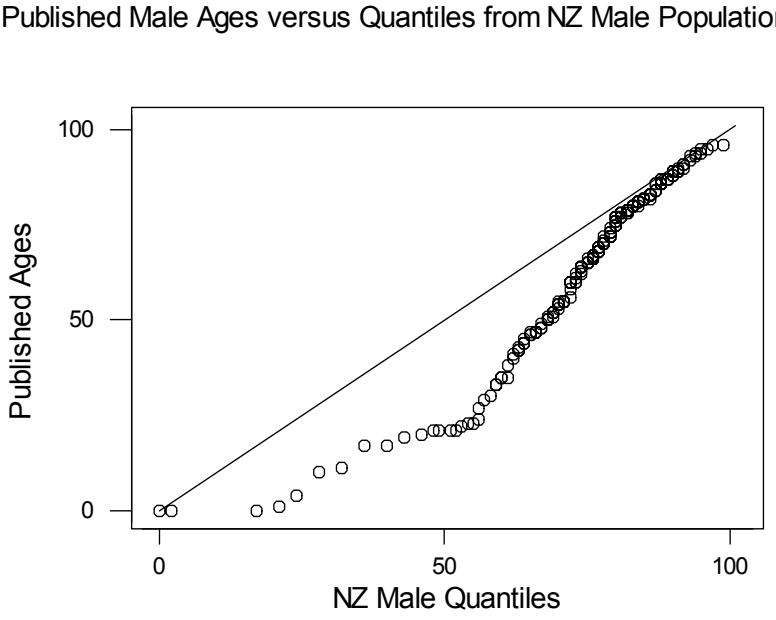
Table 3 presents a more detailed comparison of certain age percentiles for the New Zealand population and the corresponding sample percentiles. For example, only 5% of females in the New Zealand population die before the age of 52.5, whereas the sample proportion is over 15%. Thus it is clear that the *published* ages *are* a biased sample of the true age distribution. It seems that grieving relatives are much more likely to record the age of a younger person than an older one.

Table 3. Age-at-Death Percentiles for Males and Females

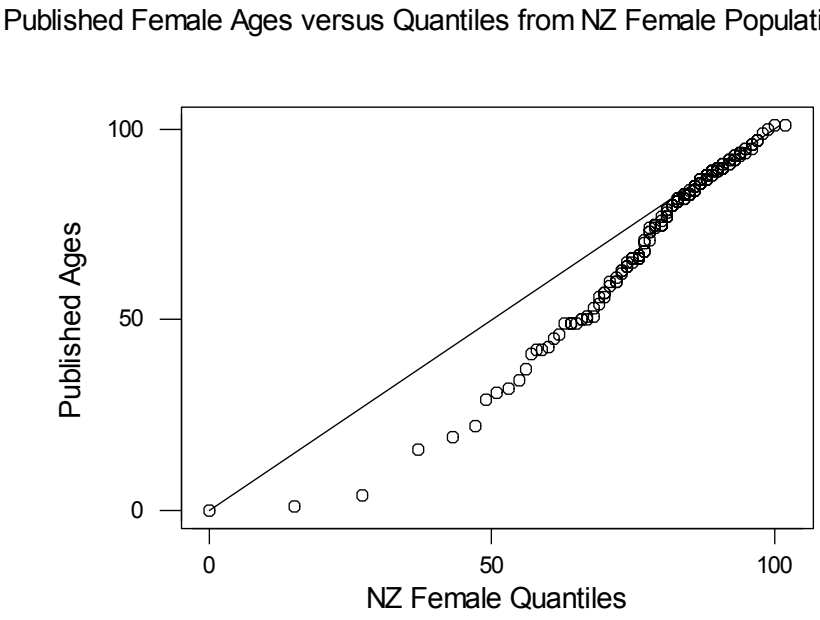
Percent	Age for Males (New Zealand)	Age of Males ('Herald' Sample)	Age for Females (New Zealand)	Age for Females ('Herald' Sample)
5	42.5	17.1	52.5	30.6
10	56.0	23.1	61.9	44.2
15	61.6	35.5	67.2	50.0
20	65.3	45.2	70.9	57.0
25	68.1	49.3	73.8	63.0
30	70.5	54.0	76.2	66.0
35	72.6	58.7	78.3	71.0
40	74.4	63.4	80.0	75.0
45	76.1	66.0	81.6	78.2
50	77.7	68.0	83.0	80.0
55	79.2	72.0	84.4	82.8
60	80.6	76.0	85.7	83.0
65	82.0	78.0	87.0	85.0
70	83.4	80.0	88.2	87.0
75	84.9	81.0	89.5	89.0
80	86.4	82.8	90.9	90.0
85	88.1	86.0	92.4	91.0
90	90.2	87.0	94.1	93.0
95	93.0	91.0	96.5	95.2

Adapted from '*New Zealand Life Tables 1995-1997'*, Tables 3.1,3.2.

We can break these figures down further. The quantile-quantile plots in Figure 5 show the sample ages and corresponding New Zealand population percentiles for males and females. The sample ages are represented by circles, and a 45 degree line shows where the points should lie if the sample matched the population. For males, in Figure 5(a), the points fall well below the line until the individual is aged around 85, and then follow the line (or just below). The major discrepancy is for published ages in the 'accident hump' years (17-24 in these data). For females, in Figure 5(b), the points again remain below the line (biased towards reporting the age) until about 85. The main difference between males and females is the virtual absence of an accident hump for females.



5(a) Males



5(b) Females

Figure 5. Quantile-Quantile Plots For Published Ages

5.4 Wording of the reported age

A curious sideline issue is the wording used in reporting the age. It is uncommon in conversational English to hear of somebody doing something, say, "in his 67th year" rather than "aged 66", and yet this idiom is frequently used in death notices. One might anticipate this usage to be more common if the death was just before a birthday, or if the mourners wish to emphasise the great age of the deceased. Conversely "aged *x*" might be more common if the mourners wish to emphasise relative youth. Figure 6 shows an indicator variable for "*x+1*th year" plotted against age, with a lowess curve used to give a smoothed average. It seems the "*x+1*th year" usage is more popular once the person exceeds 85 years, a fact which may indicate what age New Zealanders regard as being very old. Graduate students may find a challenge in modelling the number of notices, number of mentions of age, or the data on '*x+1* th year' using a generalized linear model.

5.5 Transformations of age?

An obvious student exercise using these data is to compute confidence intervals and hypothesis tests for the mean *published* age at death for males and females from different ethnic groups, and differences between the groups. Figure 7 suggests the usual normal-based confidence interval may not be appropriate since the non-Polynesian age data is very skewed. Since the samples are large, students could be instructed to ignore the skewness. Alternatively they could be required to use a nonparametric procedure or to find a transformation so that the ages are approximately symmetrical before computing the confidence intervals and hypothesis tests. It is clear from Figure 7 that no one transformation will suit all the data: indeed published Maori data need no transformation whereas non-Polynesian female ages raised to the fourth power are approximately Normally distributed. The male 'accident hump' and high infant mortality make it impossible to achieve symmetry in the sample values.

6. Discussion

The issue of biased sampling (especially non-response bias) is a crucial one in Statistics, but is very hard to demonstrate. In the present study the non-volunteering of age can be throught of as an analogue of non-response. Recent polls where non-response bias has been suggested include the 1995 fire-fighter's referendum and a 1995 survey suggesting that a majority of New Zealanders support compulsory military training. But for teaching purposes one needs examples where the bias is unequivocally demonstrated, and it is hard to find irrefutable modern examples in the literature (maybe we can blame statistical referees for that!) This dataset fills that gap. For an exercise, students may be invited to comment on whether they would anticipate any bias in the age distribution, and its possible effects, and they can then be set the task of searching for evidence of bias for themselves, using standard statistical techniques, much as was done here.

The analysis in this paper has used standard undergraduate methods such as ANOVA, so that results are accessible to students. However it is clearly less than satisfactory to model discrete data - even after transformation - using a Normal distribution, and a generalized linear model should be used instead. Finally, an interesting statistical question that arises from this data is how one could assess or model the probability that the age is missing, and what impact such missingness may have on the usual simple comparison techniques used here.

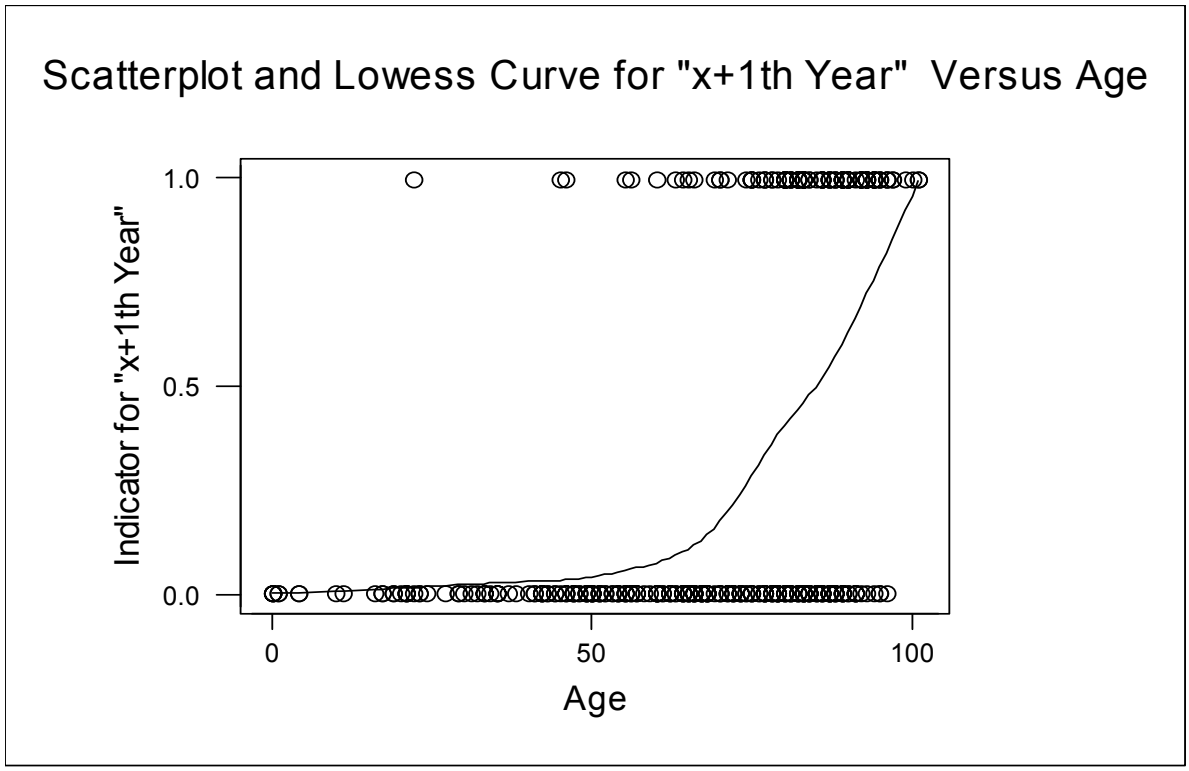


Figure 6

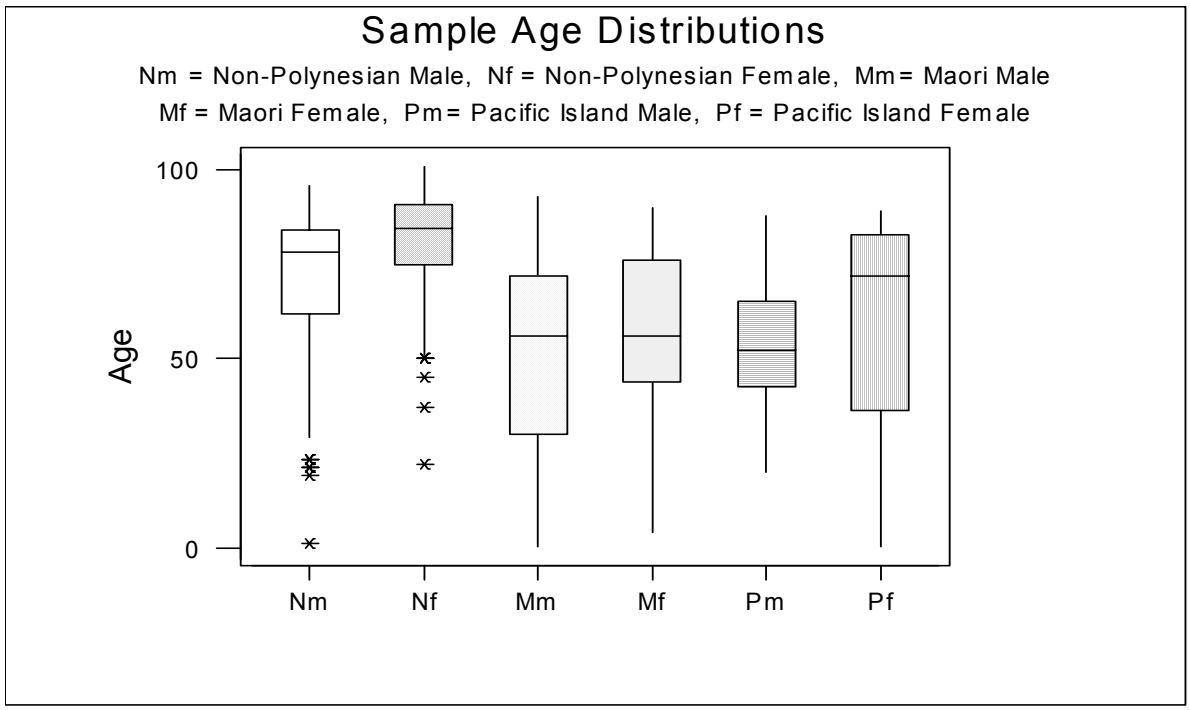


Figure 7 Sample Age Distribution Broken Down By Gender and Ethnicity

References

[1] Coslett. S.R. (1981). Efficient Estimation of Discrete-Choice Models. In *Structural Analysis of Discrete Data with Econometric Applications*. C.F. Manski and D.L. McFadden (editors). Cambridge Massachusetts: The MIT Press.

[2] Graunt, J. (1662). *Natural and Political Observations Made Upon the Bills of Mortality*.

[3] MINITAB Inc. (2000) *MINITAB release 13 for Windows*. College Park, PA: Minitab.

[4] McFadden, D. L. (1984). Econometric Analysis of Qualitative Response Models. In *Handbook of Econometrics, Volume II*. Z. Griliches and M.D. Intriligator (editors). New York: Elsevier.

[5] Statistics New Zealand (1998). *'New Zealand Life Tables 1995-1997'*. Wellington.

[6] *The New Zealand Herald.* (1995) Published by Wilson and Horton Ltd., Auckland.