AN EMPIRICAL DETERMINATION OF THE DISTRIBUTION OF MEANS, STANDARD DEVIATIONS AND CORRELATION COEFFICIENTS DRAWN FROM RECTANGULAR POPULATIONS*

By

HILDA FROST DUNLAP

Territorial Normal and Training School, Honolulu, Hawaii

Formulae for the standard errors of means, standard deviations and correlation coefficients have been derived on the assumption of a normal distribution in the sampled population. They are said to serve approximately even when the population varies considerably from the normal. This paper presents empirical evidence of their applicability in the case of means and standard deviations of samples of ten from a rectangular discontinuous population, and of correlation coefficients of samples of fifty-two from a rank distribution.

The data for the study of the distribution of means and standard deviations were secured by throwing ten dice 1600 times.

The dice were cubes four-tenths of an inch along an edge and numbered on opposite faces 1-6, 2-5, 3-4. They were constructed of bone and formed a matched set.

These were thrown from a cup whose inside diameter was 1.75 inches and whose depth was 2.5 inches. The dice were shaken in a box and then cast upon an especially prepared flat topped table covered with eight thicknesses of an army blanket.

As a guard against any possible bias in the table, the dice were thrown alternately with the right and left hands. After each throw the number of aces, deuces, treys, fours, fives, and sixes were recorded, and the mean and standard deviation calculated. In this study each throw was taken as a sample of ten drawn from a population of 16,000.

The next step was to determine whether there was any systematic bias in the dice used. The *a priori* expectation for any particular face of the die is one-sixth, here one sixth of 16,000, or 2,666⅔. This is of the nature of a point binomial of the form $(p + q)^n$ with a standard deviation equal to $\sqrt{N p q}$

TABLE I

Distribution of Observed and Theoretical Populations with a Test of the Difference of Their Standard Deviations

Die Face	Observed Frequency	Expected Frequency	Difference
1	2726	2666⅔	59⅓
2	2653	2666⅔	14⅔
3	2671	2666⅔	4⅓
4	2763	2666⅔	96⅔
5	2650	2666⅔	17⅔
6	2537	2666⅔	130⅔

$$\sigma = (1600 \cdot 1/6 \cdot 5/6)^{\frac{1}{2}} = 47.1 \qquad s = (\sum d^2 / N)^{\frac{1}{2}} = 70.8$$

$$s - \sigma = 23.7 \pm 13.76$$

Table I gives the observed and expected values of each face. The standard deviation of the differences was determined and compared with the standard deviation of the expected distribution and the probable error of this difference was found.

Small s is used here to denote a standard deviation of a sample, while σ represents the standard deviation of the theoretical or true population. The formula for the standard deviation of a difference is

$$\sigma_d = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\, r_{12}\, \sigma_1 \sigma_2}$$

and in particular

$$\sigma_{s-d} = \sqrt{\sigma_s^2}$$

The second term drops out here because it is the standard deviation of the true standard error and this is equal to zero. The third term drops out for the same reason. Table I shows that the difference between the obtained and expected standard deviations is 23.7 ± 13.76. As this is less than twice its probable error, it can be concluded that the difference is not significant and that there is no significant bias in the dice.

MEANS

Figure 1 shows the distribution of the 1600 observed means. a normal curve for $N = 1600$ is superimposed on the histogram. For this distribution

$$r_1 \left(= \sqrt{\beta_1} \right) = .0160 \pm .0413, \text{ indicating symmetry}$$

$$r_2 \left(\sqrt{\beta_2} - 3 \right) = -.1050 \pm .0826, \text{ indicating mesokurtosis}$$

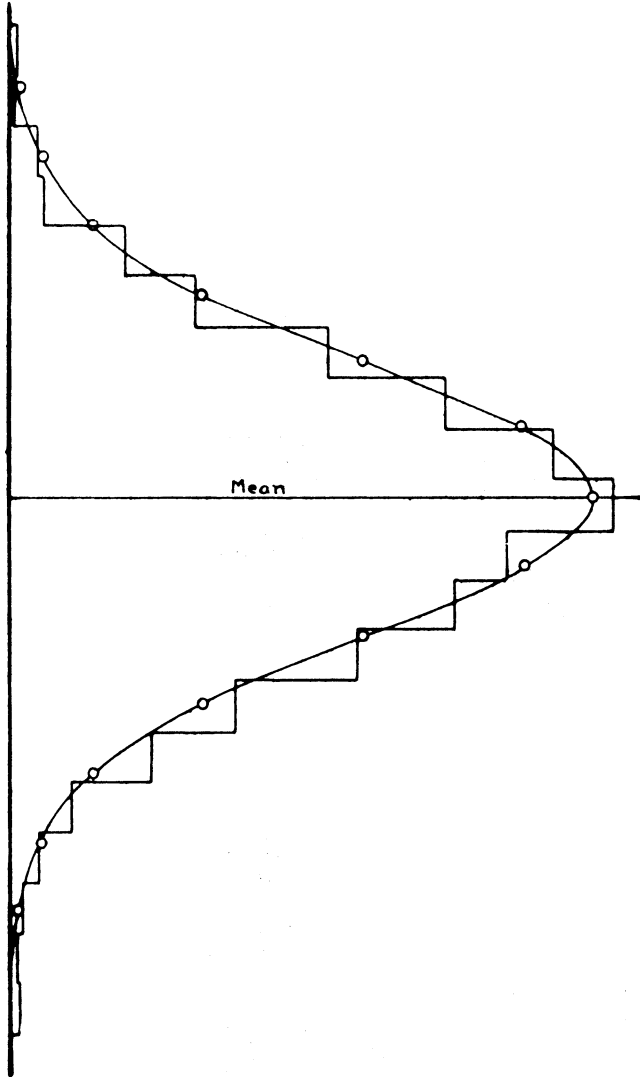whence we may conclude that the normal curve represents this

FIGURE 1

Distribution of 1600 means of samples of ten, with fitted normal curve.

distribution adequately.

The curves of this and succeeding figures were drawn through points calculated at intervals of $\frac{1}{2} \sigma$, except that in the case of Figures 2 and 3, points beyond $\pm 2 \sigma$ were calculated at intervals of 1σ.

The values of the observed means varied from 1.6 to 5.4, a range of 6.9129 standard deviations.

The basic information to be drawn from this study of the distribution of 1600 means of samples of ten is given in Table II. The table is interpreted as follows:

The mean of the sampled population (16,000) is 3.47306, while the theoretical mean of the infinite population is 3.500000. The standard deviation of the sampled population (16,000) is 1.6788, and of the theoretical population 1.7078. The standard error of the mean of the sampled population is .0133. In comparing the mean of the sampled population with the mean of the theoretical infinite population, the former is treated as an experimental value whose standard error can be estimated, while the latter, being a true value, has no error.

The standard deviation of the difference between the means M (theoretical population) and $\bar{x}$ (sampled population) is

$$\sigma_{(M \cdot \bar{x})} = \sqrt{\sigma_M^2 + \sigma_{\bar{x}}^2 - 2 r_{M\bar{x}} \, \sigma_M \, \sigma_{\bar{x}}}$$

$$\sqrt{\sigma_{\bar{x}}^2} = .0133$$

The first and third terms drop out because σ_M equals zero. The difference between the mean of the theoretical population and the sampled population is $.02694 \pm .00897$, from which it can be concluded that the mean tends to vary from the true mean.

$\bar{x}$ will hereafter refer to the mean of a sample of ten. The best estimate of the mean of a sample of ten that can be made for any sample chosen at random from the sampled population

TABLE II

Distribution of 1600 Means of Samples of 10

Description	Observed Value ($\bar{x}$)	Theoretical Value (M)
Mean of Sampled Pop.	3.47306	3.5000
σ of Sampled Pop.	1.6788	1.7078
σ_{mean} of Sampled Pop.	.0133	.0000
$\sigma_{(M-\bar{x})}$ of Sampled Pop.	.0133	.0000
$M-\bar{x}$ of Sampled Pop.	.0269 ± .00897	.0000
Mean of Means of Samples	3.47306	3.47306 or 3.5000
S. D. of Means of Samples	.5497	.5372 or .5401
S. E. of S. D. of Means of Samples	.0097	.0000 or .0000
$s_{\bar{x}} - \sigma_m$	.0125 ± .0065 or .0096 ± .0065	.0000 or .0000
γ_1 of Distri. of Means of Samples	.0160 ± .0413	.00 (normal theory)
γ_2 of Distri. of Means of Samples	-.1050 ± .0826	.00 (normal theory)

is 3.47306, and from the infinite population, 3.5000.

The standard deviation of the means of 1600 samples is .5467, while the estimated value for a sample picked at random from the sampled population is .5372 and from the theoretical infinite population .5401. These last two values are calculated by the formula

The best estimate of the standard deviation of a sample of ten picked at random from the sampled population is the σ of the sampled population, 1.6788, or of the theoretical infinite population, 1.7078, whence the values in the tables are obtained.

The standard error of the standard deviation of the means of samples is .0097. The standard error of the standard error σ_M of the mean of a sample of ten from the sampled and theoretical infinite populations is zero, as these are true values.

The difference between the standard deviation of the means and the standard error of such means of samples of ten from the sampled population or the theoretical infinite population is .0125 ±.0065. Thus there is no significant difference between the value of σ_M when calculated by the formula $\sigma_M = \frac{\sigma}{\sqrt{N}}$ and an actual distribution when samples as small as ten are used.

γ_1 indicates, as pointed out above, that the distribution is not skewed, while γ_2 shows the distribution to be slightly peaked but not significantly so.

STANDARD DEVIATIONS

Figure 2 shows a histogram and a fitted Gram-Charlier Type A curve, of the distribution of 1600 standard deviations of samples of ten calculated by the formula

$$S = \sqrt{\frac{\sum x^2}{N}}$$

X being measured from the mean, $\bar{x}$

Figure 3 shows a similar histogram and curve fitted to the
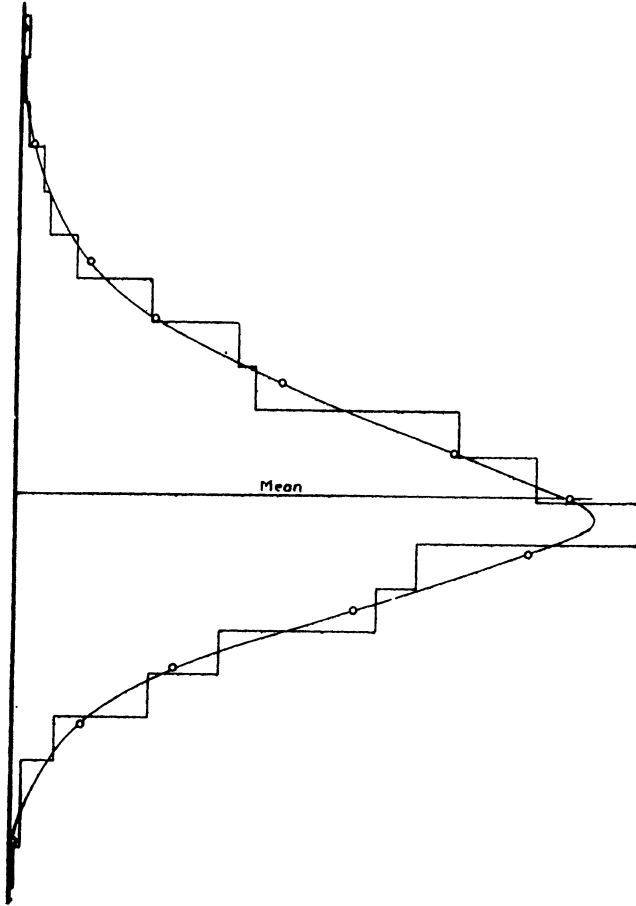
Mean

FIGURE 2

Distribution and fitted Gram-Charlier curve of 1600 standard deviations of samples of ten, calculated by the formula $s = \left(\frac{1}{N} \Sigma x^2 \right)^{\frac{1}{2}}$
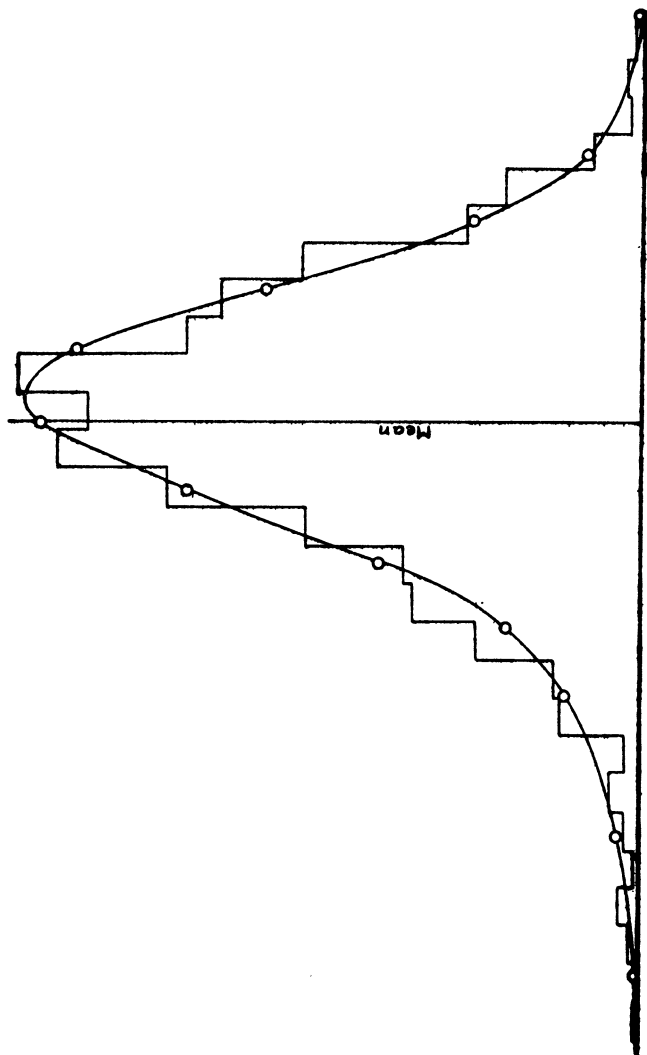
FIGURE 3

Distribution and fitted Gram-Charlier curve of 1600 standard deviations of samples of ten, calculated by the formula $s = (\frac{1}{N-1}\sum x^2)^{\frac{1}{2}}$

TABLE III

Distribution ot 1600 Standard Deviations of Samples of Ten

Description	Observed Value		Theoretical Value	
	$s^2 = \frac{\Sigma x^2}{N}$	$s^2 = \frac{\Sigma x^2}{N-1}$	Sampled Population	Infinite Population
$\bar{x}$ of s's of sam.	1.5869	2.0403	1.6988	1.7078
S. D. of s's of sam.	.2665	.2538	.3799	.3818
S. D. of $\bar{x}$ of s's of samples	.0067	.0063	.0000	.0000
S. D. of s of s's of samples	.0047	.0045	.0000	.0000
$\sigma - \bar{x}_s$	.1119	.3415	.0000	.0000
	±.0045 or	±.0042 or		
	.1209	.3325		
	±.0045	±.0042		
$\sigma_\sigma - s_s$	.1134	.1261	.0000	.0000
	±.0032 or	±.0030 or		
	.1153	.1280		
	±.0032	±.0030		
γ_1 (skewness)	−.3568	−.5026	.0000 (normal	
	±.0413	±.0413		theory)
γ_2 (kurtosis)	.5140	.6851	.0000 (normal	
	±.0826	±.0826		theory)
N	1600	1600		

same data when the standard deviations are calculated by the formula

$$S = \sqrt{\frac{\Sigma x^2}{N-1}}$$

A study of this latter formula is included here to test which is more appropriate when dealing with small samples from a rectangular population.

An interpretation of Table III is now in order. Column one is a description of the statistics involved. Column two is subdivided into two parts: First, when s equals $\sqrt{\frac{\Sigma x^2}{N}}$, and second when s equals $\sqrt{\frac{\Sigma x^2}{N-1}}$. Column three gives the theoretical values. There are two of these—one for the sampled population and one for the infinite population. In the case of the sampled population the values calculated for the standard deviation and the σ_σ become true values when a single sample is compared with them in exactly the same manner as if compared with similar values from the infinite population. The reason for this is that for a given sample the 16,000 constitutes the actual population from which the sample is drawn.

In the first line the means of the standard deviations of the samples are found to equal respectively, 1.5869 and 2.0403. The theoretical means for the sampled and infinite populations are respectively 1.6988 and 1.7078.

In the next line are the standard deviations of standard deviations of samples. These are calculated values, obtained by substituting in the formula

$$\sigma_s = \frac{\sigma}{\sqrt{2N}}$$

As the best estimate of the standard deviations of any particular sample chosen at random is the standard deviation of the sampled population, or the infinite population these values can be substituted in the above formula in obtaining the standard error of the standard deviation of such a sample of ten.

The standard error of the mean of standard deviations in

samples for both observed values is given in line three. Obviously
in the case of the sampled and infinite populations these equal
zero. It should be clearly understood by the reader that here N
equals 1600, the number of standard deviations used in deter-
mining the mean standard deviation.

Line four gives the standard error of the standard deviation
of standard deviations of samples of ten.

Line five gives the difference between each of the true stand-
ard deviations (sampled and infinite) and the two observed mean
standard deviations. The standard deviations of the sampled
population and of the infinite population are each greater than the
mean standard deviation of the observed population when calcu-
lated by the formula $s = \sqrt{\frac{\sum x^2}{N}}$. In the first case the differ-
ence is $.1119 \pm .0045$. This is approximately 25 times its prob-
able error, so it must be considered a significant difference. The
difference when compared with the theoretical infinite population
is $.1209 \pm .0045$. This is even more significant. When the
theoretical values are compared with the mean standard deviation
calculated by the formula $s = \sqrt{\frac{\sum x^2}{N-1}}$ the differences are found
to be $.3415 \pm .0042$, and $.3325 \pm .0042$. The differences here are
much greater than those found from the first formula.

Line six shows the difference between the standard errors
of the standard deviations of the true populations and the cal-
culated s_s of the samples. The difference between σ_σ and
s_s $(.3799 - .2665)$, is $.1134 \pm .0032$. This difference is approx-
imately 35 times its probable error. The difference between $.3799$
and $.2538$ is even greater. Still larger differences are found
when s_s is calculated for the $s = \sqrt{\frac{\sum x^2}{N-1}}$ formula.

γ_1 in the case of both curves is negative and more than
8 times its probable error, definitely showing a negative skewness,
γ_2 in the case of both curves is 6 times greater than its prob-
able error, indicating definite leptokurtosis. The Gram-Charlier
curves shown in Figures 2 and 3 were fitted to the first four

moments according to the equation

$$Y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left[1 - \left(\frac{\mu_3}{6\sigma^3}\right)(3x - x^3) + \left(\frac{\mu_4}{24\sigma^4} - .125\right)(x^4 - 6x^2 + 3) \right]$$

where
$$x = \frac{X - \bar{x}}{s}$$

If we compute values of s by the empirical formula $s = \sqrt{\frac{\Sigma x^2}{N - .25}}$, the mean value is 1.7039, which lies very close to the theoretical values 1.6988 and 1.7078, in fact almost exactly half-way between them.

CORRELATION COEFFICIENTS

The product-moment correlation coefficient varies between the limits plus one and minus one. Obviously, the distribution of correlation coefficients cannot be normal, although in the case where $r = 0$ their distribution should approximate a normal curve, as it can become symmetrical. Coefficients around any other point tend to be distributed asymmetrically.

It was assumed that if a deck of cards be thoroughly shuffled there should be no correlation between successive deals. Using a deck of cards gives a sample of 52. A new pack was thoroughly shuffled. The cards were then dealt one at a time, the first card dealt being recorded as number one, the second card dealt as number two, the third card as number three, etc. That is, if the seven of hearts was turned first, the value one was recorded against its place in the table. After each deal the cards were picked up in the same order and shuffled three times by the fan method and then cut twice. Sixty such deals were made and recorded. Then rank correlations were calculated be-
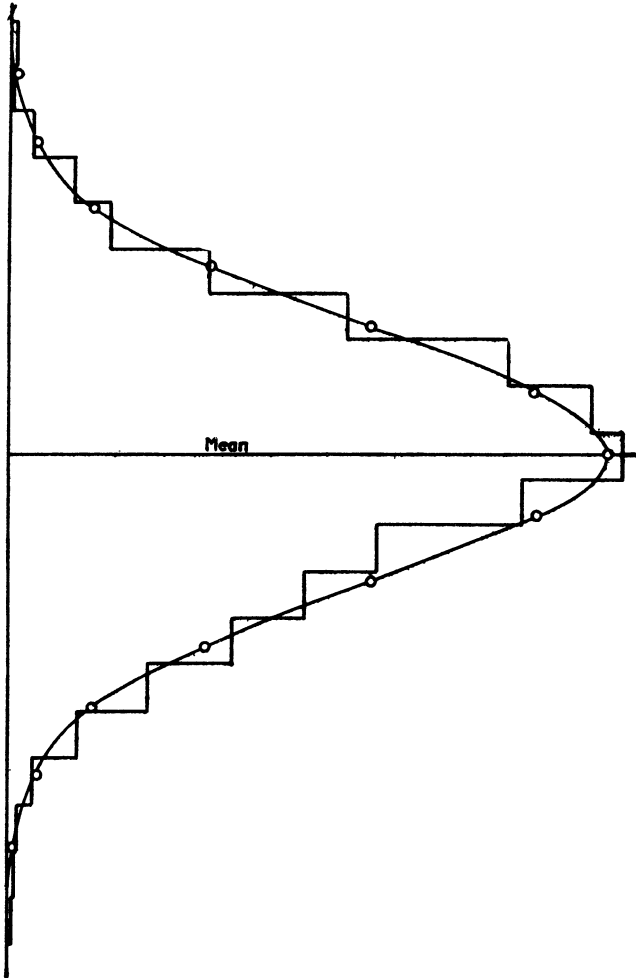
FIGURE 4

Distribution of 1770 correlation coefficients of samples of 52, with fitted normal curve.

tween each pair of deals, the total number of intercorrelations being $\frac{n(n-1)}{2}$, here 1770.

In this study, there could be no split ranks. Each card could receive one and only one rank on each deal. Thus, the rank correlation formula gave exactly the same values as would a Pearson product-moment coefficient.

Figure 4 shows a histogram with a fitted normal curve super-imposed on it. γ_1 for this curve is $.000015 \pm .0392$, indicating no skewness, and γ_2 is $.2174 \pm .0785$, indicating a slight tendency to peakedness. Both of these facts are shown by the fit of the curve to the histogram.

The formula for the standard error of a correlation coefficient from a normal population is

$$\sigma_r = \frac{1-\rho^2}{\sqrt{N}}$$

ρ being the correlation in the population. Thus when $r = .0000$ and $N = 52$, $\sigma_r = .1387$.

The mean value of the 1770 coefficients is $r = -.0012$. The expected mean is zero. The difference between these two values is $.0012 \pm .0022$. This shows that the mean correlation coefficient is not significantly different from the expected mean correlation.

The standard deviation of the observed distribution is $.1359$. This value differs from the expected value by $.0028 \pm .0091$. The formula $\sigma_r = \frac{1-\rho^2}{\sqrt{N}}$ is therefore seen to give a sufficiently close approximation in this case.

CONCLUSIONS

1. The distribution of means of samples of ten drawn from a discontinuous rectangular population is normal. The formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$ gives a reasonably close estimate of the standard error of such means.

2. The distribution of standard deviations of samples of

ten drawn from a discontinuous rectangular population is skewed and leptokurtic. The formula $\sigma_s = \dfrac{\sigma}{\sqrt{2\,N}}$ does not give a reasonably close estimate of the standard deviation of standard deviations of samples of ten, whether the latter are computed from the formula $s = \sqrt{\dfrac{\Sigma\,x^2}{N}}$ or $s = \sqrt{\dfrac{\Sigma\,x^2}{N-1}}$

3. Neither of the formulas, $s = \sqrt{\dfrac{\Sigma\,x^2}{N}}$ and $s = \sqrt{\dfrac{\Sigma\,x^2}{N-1}}$ for the standard deviation of a sample of ten gives a reasonably close estimate of the true standard deviation in a rectangular discontinuous population. The empirical formula $s = \sqrt{\dfrac{\Sigma\,x^2}{N-.25}}$ does appear to do so.

4. The distribution of correlation coefficients of samples of 52 from a rank population in which the expected correlation is zero, is symmetrical and very slightly leptokurtic. The formula $\sigma_r = \dfrac{1-\rho^2}{\sqrt{N}}$ represents adequately the standard deviation of such correlation coefficients.

Hilda F Dunlap