251descr2 2/10/06 (Open this document in 'Outline' view!)

G. Measures of Dispersion and Asymmetry.

1. Range

$Range = highest\ number - lowest\ number$ or $highest\ midpoint - lowest\ midpoint$.

Interquartile Range: $IQR = Q3 - Q1$. See 251descr2ex2 for example.

2. The Variance and Standard Deviation of Ungrouped Data.

a. The Population Variance - Definitional and Computational Formulas.

The definition of the population variance is 'the average squared deviation of measurements from the mean.' The definitional formula just realizes this definition.

Definitional $\sigma^2 = \frac{\sum (x - \mu)^2}{N}$ Computational $\sigma^2 = \frac{\sum x^2}{N} - \mu^2$

Standard Deviation = $\sqrt{variance}$

b. The Sample Variance.

Definitional $s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$ Computational $s^2 = \frac{\sum x^2 - n\bar{x}^2}{n - 1}$

The computational formula is one of the most important formulas you will learn. **Note that $\sum x^2$ is not the same as $(\sum x)^2$.** For example, if x is $\{2,3,5\}$, $\sum x^2 = 2^2 + 3^2 + 5^2 = 4 + 9 + 25 = 38$, not $(2+3+5)^2 = 10^2 = 100$.

Example: Use $x = \{2,3,5\}$

Computational Method

x	x^2
2	4
3	9
5	25
10	38

Definitional Method

x	$(x - \bar{x})$	$(x - \bar{x})^2$
2	-1.33333	1.77778
3	-0.33333	0.11111
5	1.66667	2.77778
10	0.00001	4.66667

From this we find $\sum x = 10$, $\sum x^2 = 38$, $\bar{x} = \frac{\sum x}{n} = \frac{10}{3} = 3.33333$ and $\sum (x - \bar{x})^2 = 4.66667$ Note that $\sum (x - \bar{x})$ should be zero, but is not because of rounding. Now, if we use the computational method, we can use $s^2 = \frac{\sum x^2 - n\bar{x}^2}{n - 1} = \frac{38 - 3(3.33333)^2}{3 - 1} = \frac{4.6667}{2} = 2.3333$

(Some texts prefer $s^2 = \frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n - 1} = \frac{38 - \frac{1}{3}(10)^2}{3 - 1} = \frac{4.66666667}{2} = 2.33333$ which gives us a little more accuracy for a little more work.) If we use the definitional method $s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{4.66667}{2} = 2.33333$, but note that we had to do three subtractions instead of 1.

c. The Coefficient of Variation.

$$C = \frac{std.deviation}{mean}$$

d. Chebyshef's Inequality and the Empirical Rule

Chebyshef Inequality: $P\left(|x - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}$ or $P\left(\mu - k\sigma \leq x \leq \mu + k\sigma\right) > 1 - \frac{1}{k^2}$. A z-score $z = \frac{x - \mu}{\sigma}$ is the same as k. (See explanation below)

Empirical rule: (For Symmetrical Unimodal distributions only)
68% within one standard distribution of the mean, 95% within two and almost all (99.7%) within three.

3. The Variance and Standard Deviation of Grouped Data.

For grouped data generally substitute $\sum f$ for $\sum$.

4. Skewness and Kurtosis.

Define Population Skewness, the 3rd k-statistic, coefficients of Skewness; Population Kurtosis, the 4th k-statistic, the Coefficient of Excess; Leptokurtic, Platykurtic and Mesokurtic distributions.
The usual measurement of skewness is often called the third moment about the mean .
(The population variance is the second). The formula for population skewness is:

$$\mu_3 = \frac{\sum (x - \mu)^3}{N}.$$

The corresponding sample statistic is the third k-statistic, $k_3 = \frac{n}{(n-1)(n-2)} \sum (x - \bar{x})^3$. The corresponding computational formulas are

$\mu_3 = \frac{1}{N}\left(\sum x^3 - 3\mu \sum x^2 + 2N\mu^3\right)$ and $k_3 = \frac{n}{(n-1)(n-2)}\left[\sum x^3 - 3\bar{x} \sum x^2 + 2n\bar{x}^3\right]$. To make grouped data formulas, put an f to the right of the $\sum$ sign. Positive values of these formulas imply skewness to the right, negative values to the left. Note that multiplying all the values of x by two would multiply the values of these coefficients by eight, but would not change the shape of the distribution. If we want to compare shapes, we need measurements that will not change if we multiply all values by a constant. Such a measure would be called the coefficient of relative skewness, with the formulas $\gamma_1 = \frac{\mu_3}{\sigma^3}$ and $g_1 = \frac{k_3}{s^3}$. Note that for the Normal distribution $\gamma_1 = 0$. Other measures of skewness are Pearson's measures of skewness, $SK1 = \frac{(mean - \text{mode})}{std.deviation}$ and $SK2 = \frac{3(mean - median)}{std.deviation}$. These are roughly equivalent, since, for a moderately skewed distribution, $(mean - \text{mode}) \approx 3(mean - median)$. It seems that $-3 \leq SK1 \leq 3$ and that values between. 1 and -1 are considered to indicate moderate skewness.

251descr2 2/10/06

Example:

Profit Rate	f	X (midpoint)	fx	fx^2	fx^3
9-10.99	3	10	30	300	3000
11-12.99	3	12	36	432	5184
13-14.99	5	14	70	980	13720
15-16.99	3	16	48	768	12288
17-18.99	1	18	18	324	5832
Total	15		202	2804	40024

So $\sum f = n = 15$, $\sum fx = 202$, $\sum fx^2 = 2804$, $\sum fx^3 = 40024$, so that

$\bar{x} = \frac{\sum fx}{n} = \frac{202}{15} = 13.467$ and $s^2 = \frac{\sum fx^2 - n\bar{x}^2}{n-1} = \frac{2804 - 15(13.467)^2}{15-1} = \frac{82.733}{14} = 5.981$,

which means $s = \sqrt{5.981} = 2.446$. $C = \frac{s}{\bar{x}} = \frac{2.446}{13.467} = 0.182$. To measure skewness, use one of the following three results.

$$k_3 = \frac{n}{(n-1)(n-2)}\left[\sum fx^3 - 3\bar{x}\sum fx^2 + 2n\bar{x}^3\right] = \frac{15}{(14)(13)}\left[40024 - 3(13.467)(2804) + 2(15)(13.467)^3\right]$$

$$= \frac{15(8.249)}{(14)(13)} = 0.680, \quad \text{or} \qquad \text{Relative Skewness } g_1 = \frac{k_3}{s^3} = \frac{0.680}{(2.446)^3} = .046 \quad \text{or}$$

Pearson's Measure of Skewness $SK1 = \frac{(mean - mode)}{std.deviation} = \frac{13.467 - 14}{2.446} = -0.2179.$ Note that, in this case, Pearson's Measure 1 and Relative Skewness contradict each other as to the direction of skewness.

The measures of kurtosis are, for populations,

$\mu_4 = \frac{\sum (x-\mu)^4}{N} = \frac{1}{N}\left[\sum x^4 - 4\mu\sum x^3 + 6\mu^2\sum x^2 - 3\mu^4\right]$ and, for samples,

$k_4 = \frac{n^2}{(n-1)(n-2)(n-3)}\left[(n+1)\frac{\sum (x-\bar{x})^4}{n} - \frac{3(n-1)^3 s^4}{n^2}\right]$. k_4 can be considered an estimate of $\mu_4 - 3\sigma^4$. To get a measurement of shape use the Coefficient of Excess $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$ or $g_2 = \frac{k_4}{s^4}$. Since the Normal distribution has $\mu_4 = 3\sigma^4$, the coefficient of excess is zero for the Normal distribution. Kurtosis has traditionally been considered a measure of the peakedness of a distribution relative to the Normal distribution, though there are some exceptions to this interpretation. If the coefficient of excess is positive, we may call a distribution leptokurtic or sharp-peaked (and long-tailed). If the coefficient of excess is negative, the distribution can be called platykurtic or flat-peaked (and short-tailed). If the coefficient of excess is close to zero, we call the distribution mesokurtic, middle-peaked. A symmetric, mesokurtic distribution is essentially Normal. An alternate measure, called simply the coefficient of kurtosis is $K = \frac{.5(x_{.25} - x_{.75})}{x_{.10} - x_{.90}}$. This is dimension-free and takes values between zero and 0.5. Values above .263 (K for the Normal distribution) indicate a leptokurtic distribution. Values below .263 indicate a platykurtic distribution.