Methods for Describing Data Revisited

Example:

Sorted EPA Mileage Ratings on 100 Cars

30.0	33.9	35.2	36.1	36.7	37.0	37.4	38.0	39.0	40.2
31.8	33.9	35.3	36.2	36.7	37.0	37.4	38.1	39.0	40.3
32.5	34.0	35.5	36.3	36.7	37.0	37.5	38.2	39.3	40.5
32.7	34.2	35.6	36.3	36.8	37.1	37.6	38.2	39.4	40.5
32.9	34.4	35.6	36.4	36.8	37.1	37.6	38.3	39.5	40.7
32.9	34.5	35.7	36.4	36.8	37.1	37.7	38.4	39.7	41.0
33.1	34.8	35.8	36.5	36.9	37.2	37.7	38.5	39.8	41.0
33.2	34.8	35.9	36.5	36.9	37.2	37.8	38.6	39.9	41.2
33.6	35.0	35.9	36.6	36.9	37.3	37.9	38.7	40.0	42.1
33.8	35.1	36.0	36.6	37.0	37.3	37.9	38.8	40.1	44.9

Descriptive Statistics:

Variable	N	Mean	Median	StDev	Min	Max	Q1	Q3
MPG	100	36.994	37.000	2.418	30.000	44.900	35.625	38.375

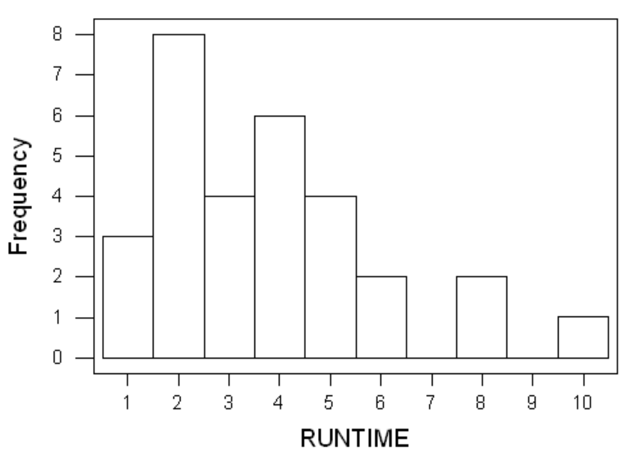
25% of the MPG values are less than 35.65; 50% of the MPG values are less than 37;
75% of the MPG values are less than 38.35.
About 68% of the MPG values are within one StDev of the mean.
About 95% of the MPG values are within two StDevs of the mean.
About 99% of the MPG values are within three StDevs of the mean.

Example of run times (in minutes) of 30 rats running through a maze:

1.97	5.36	9.70	6.06	1.93	7.60	1.74	4.02	1.71	5.63	2.02	2.06
3.77	3.81	1.15	4.25	4.55	3.65	0.60	1.06	8.29	4.44	5.15	3.16
2.75	3.20	2.47	5.21	3.37	1.65						

Descriptive Statistics

Variable	N	Mean	Median	StDev
RUNTIME	30	3.744	3.510	2.198



To be within 1 standard deviation of the mean, a run time must be in the interval (1.546, 5.942) About 77% of the run times are in that interval.

To be within two standard deviations of the mean, run time must be in the interval (-0.652, 8.14). About 93% of the run times are in that interval.

To be within three standard deviations of the mean, run time must be in the interval (-2.85, 10.338). 100% of the run times are in that interval.

Some Methods for Determining the Median and Quartiles
(Source: http://mathworld.wolfram.com/Quartile.html

The following table summarizes a number of common methods for computing the position of the first and third quartiles from a sample size ***n***. (P. Stikker, pers. comm., Jan. 24, 2005). In the table, [x] denotes the nearest integer function.)

method	1st quartile	1st quartile	3rd quartile	3rd quartile
	n odd	*n* even	*n* odd	*n* even
Minitab	$\frac{n+1}{4}$	$\frac{n+1}{4}$	$\frac{3n+3}{4}$	$\frac{3n+3}{4}$
Tukey (Hoaglin et al. 1983)	$\frac{n+3}{4}$	$\frac{n+2}{4}$	$\frac{3n+1}{4}$	$\frac{3n+2}{4}$
Moore and McCabe (2002)	$\frac{n+1}{4}$	$\frac{n+2}{4}$	$\frac{3n+3}{4}$	$\frac{3n+2}{4}$
Mendenhall and Sincich (1995)	$\left[\frac{n+1}{4}\right]$	$\left[\frac{n+1}{4}\right]$	$\left[\frac{3n+3}{4}\right]$	$\left[\frac{3n+3}{4}\right]$
Freund and Perles (1987)	$\frac{n+3}{4}$	$\frac{n+3}{4}$	$\frac{3n+1}{4}$	$\frac{3n+1}{4}$

Numerical Measures of Variability

	Sample	Population
Variance	s^2	σ^2
Standard Deviation	s	σ

Interpreting the Standard Deviation

How many observations fit within ± n standard deviations of the mean?

	Chebyshev's Rule	Empirical Rule
±1s or ±1σ	No useful info	Approximately 68%
±2s or ±2σ	At least 75%	Approximately 95%
±3s or ±3σ	At least 8/9	Approximately 99.7%

How well do the two rules describe the distribution of MPG ratings and run times?