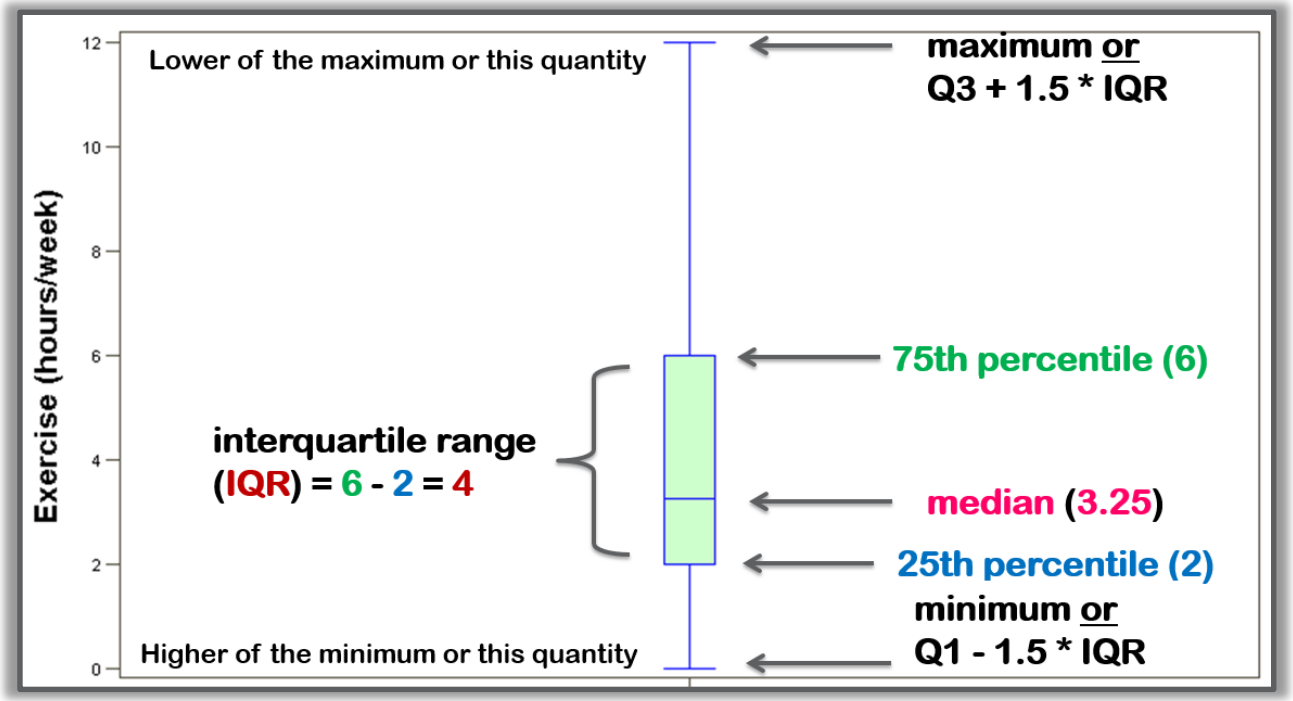
Univariate & Bivariate EDA: Variability



- Variability (spread): variance, standard deviation, mean absolute deviation, range, order statistics (ranks), percentile (quantile), interquartile range (IQR)
- Visualization: Boxplot

Prof. Nagiza F. Samatova
samatova@csc.ncsu.edu
Department of Computer Science
North Carolina State University

Measures of Variability

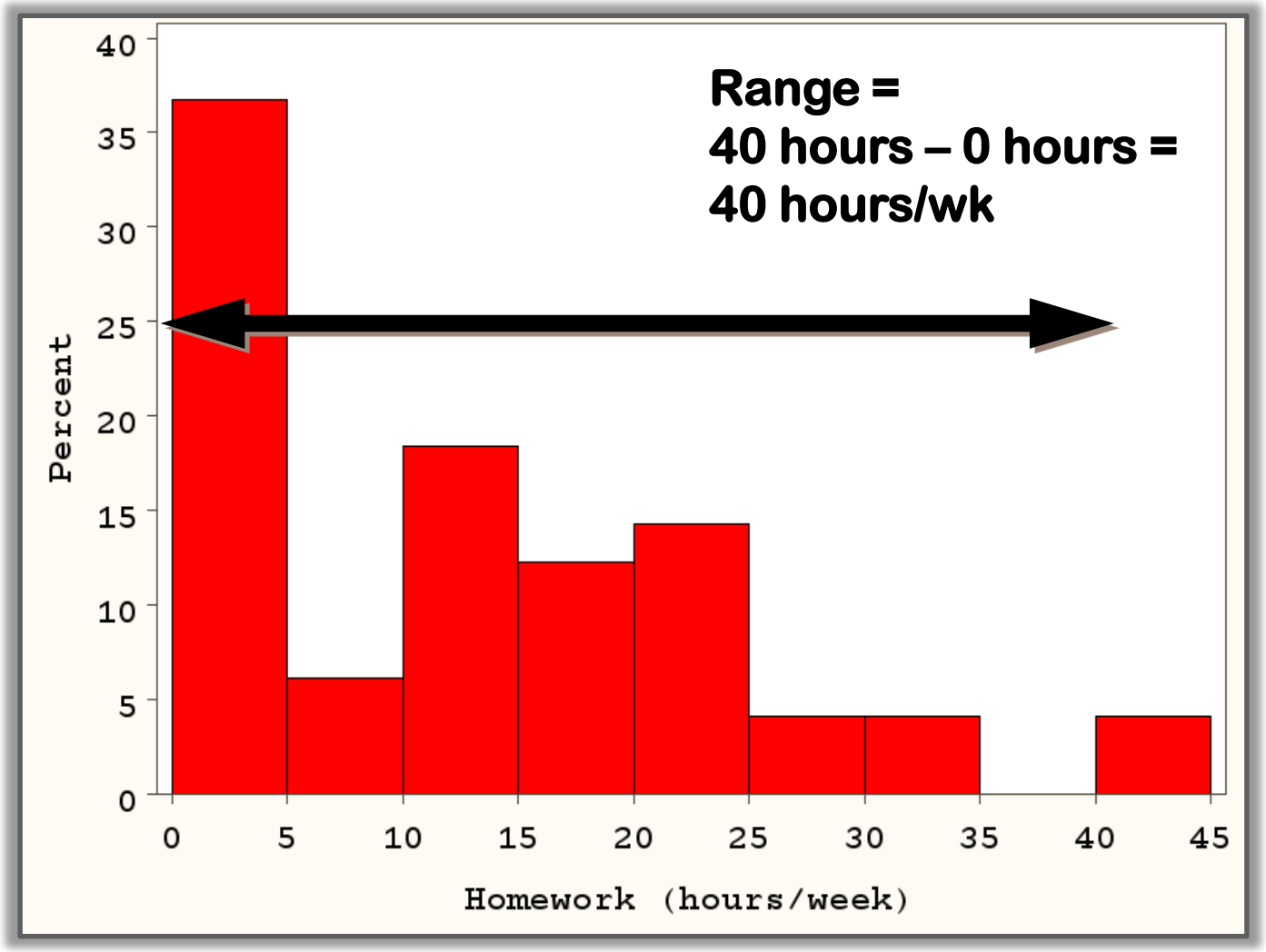
Measure	Description	Synonyms
Range	difference between the largest and the smallest observations	
Variance	average distance from the mean	mean-squared-error
Standard deviation	average spread around the mean = square root of the variance	Euclidean norm, L2-norm
Mean/median absolute deviation	mean/median of the absolute value of the deviations from the mean	Manhattan norm, L1-norm
Percentile	value such that P percent of the values take on this value or less and (100-P) percent take more	quantile
Inter-quartile range (IQR)	difference between the 75^{th} percentile and the 25^{th} percentile of the sorted data values	IQR
Deviance	difference between the observed value and the estimated location	errors, residuals
Order statistics	metrics based on the data values sorted from smallest to biggest	ranks

Measures of Variability

- Range
- Standard deviation
- Variance
- Percentiles
- Inter-quartile range (IQR)

Range

Difference between the largest and the smallest observations



Variance

Average squared distance from the mean

$$S^2 = \frac{\sum_{i}^{n} (x_i - \bar{X})^2}{n-1}$$

We lose a “degree of freedom because we have already estimated the mean.

Standard Deviation $\equiv \sqrt{Variance}$

Average spread around the mean

$$S = \sqrt{\frac{\sum_{i}^{n}(x_i - \bar{X})^2}{n-1}}$$

Age data (n=8):

17 19 21 22 23 23 23 38

n = 8 Mean = 23.25

$$S = \sqrt{\frac{(17-23.25)^2 + (19-23.25)^2 + \cdots + (38-23.25)^2}{8-1}}$$

$$= \sqrt{\frac{280}{7}} = 6.3$$

Both the variance & standard deviation are sensitive to outliers

- Because of the squaring, values farther from the mean contribute more to the standard deviation than values closer to the mean

- More robust metrics of variance:
  - mean and median absolute deviations from the mean (**MAD**)
  - percentiles (quantiles)

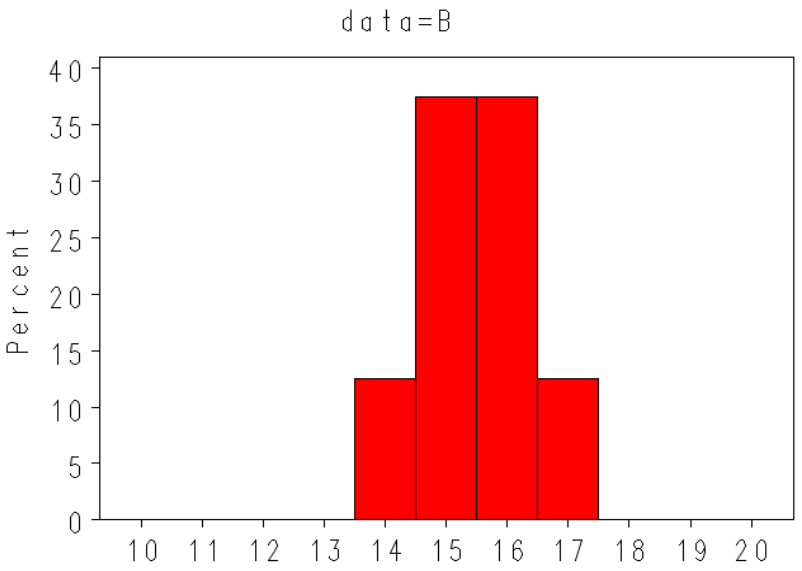
$$\overline{X} = 5$$

$$(6-5)^2 = 1$$

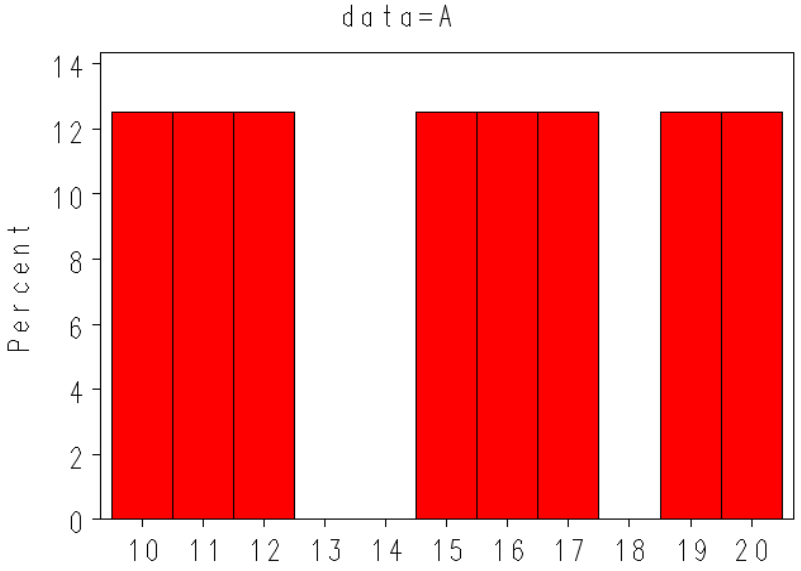
$$(10-5)^2 = 25$$

Understanding Standard Deviation

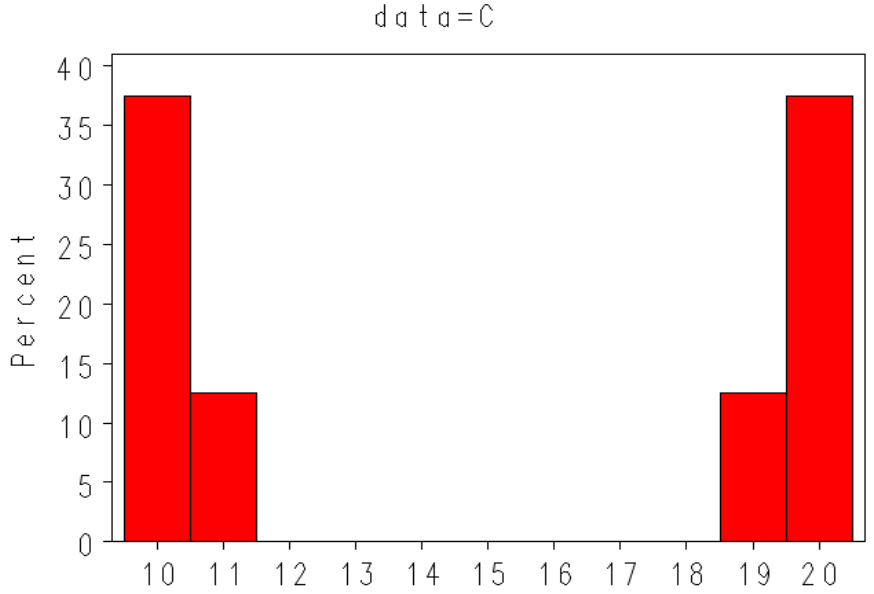
Mean = 15
S = 0.9



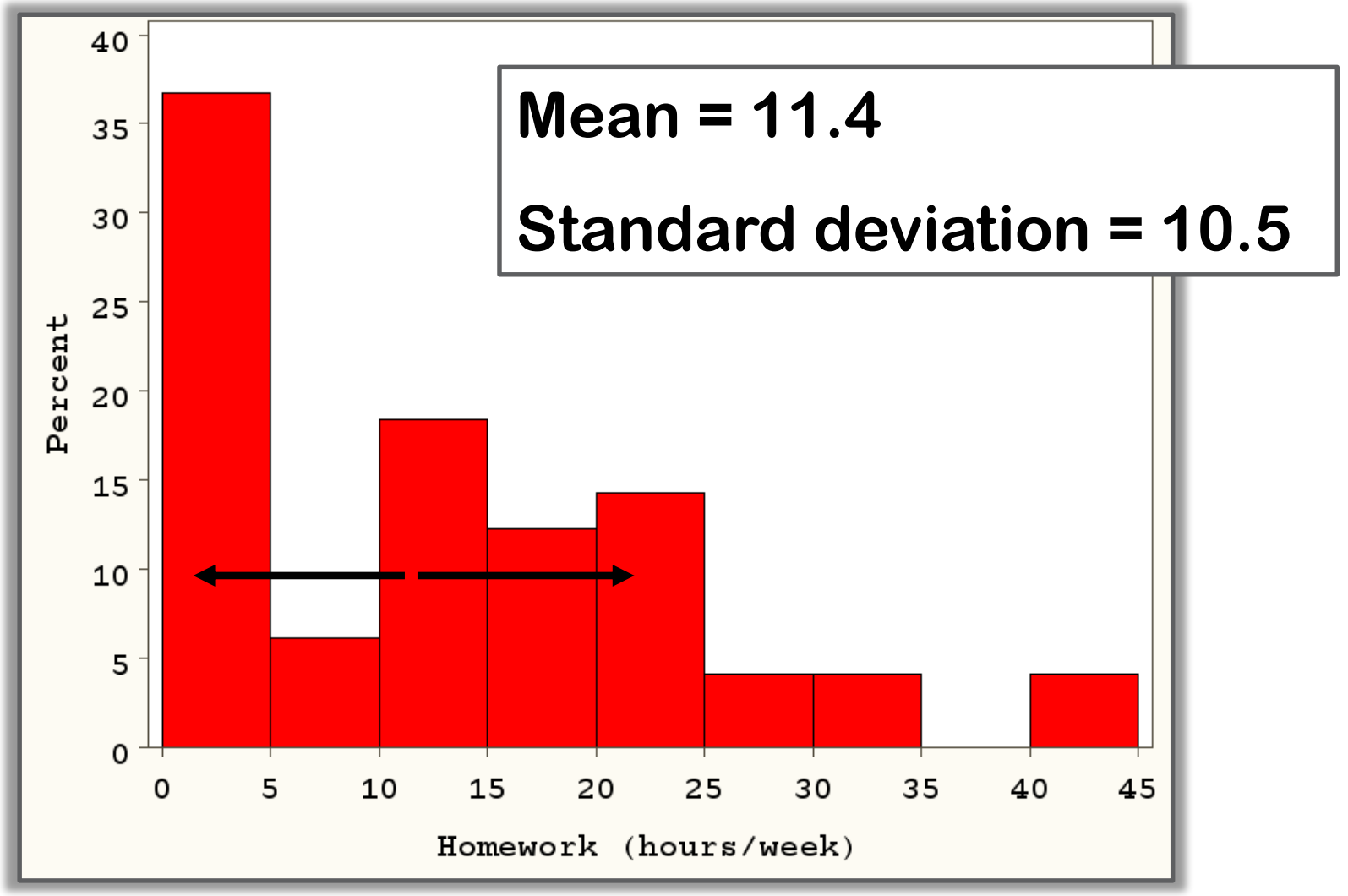
Mean = 15
S = 3.7



Mean = 15
S = 5.1



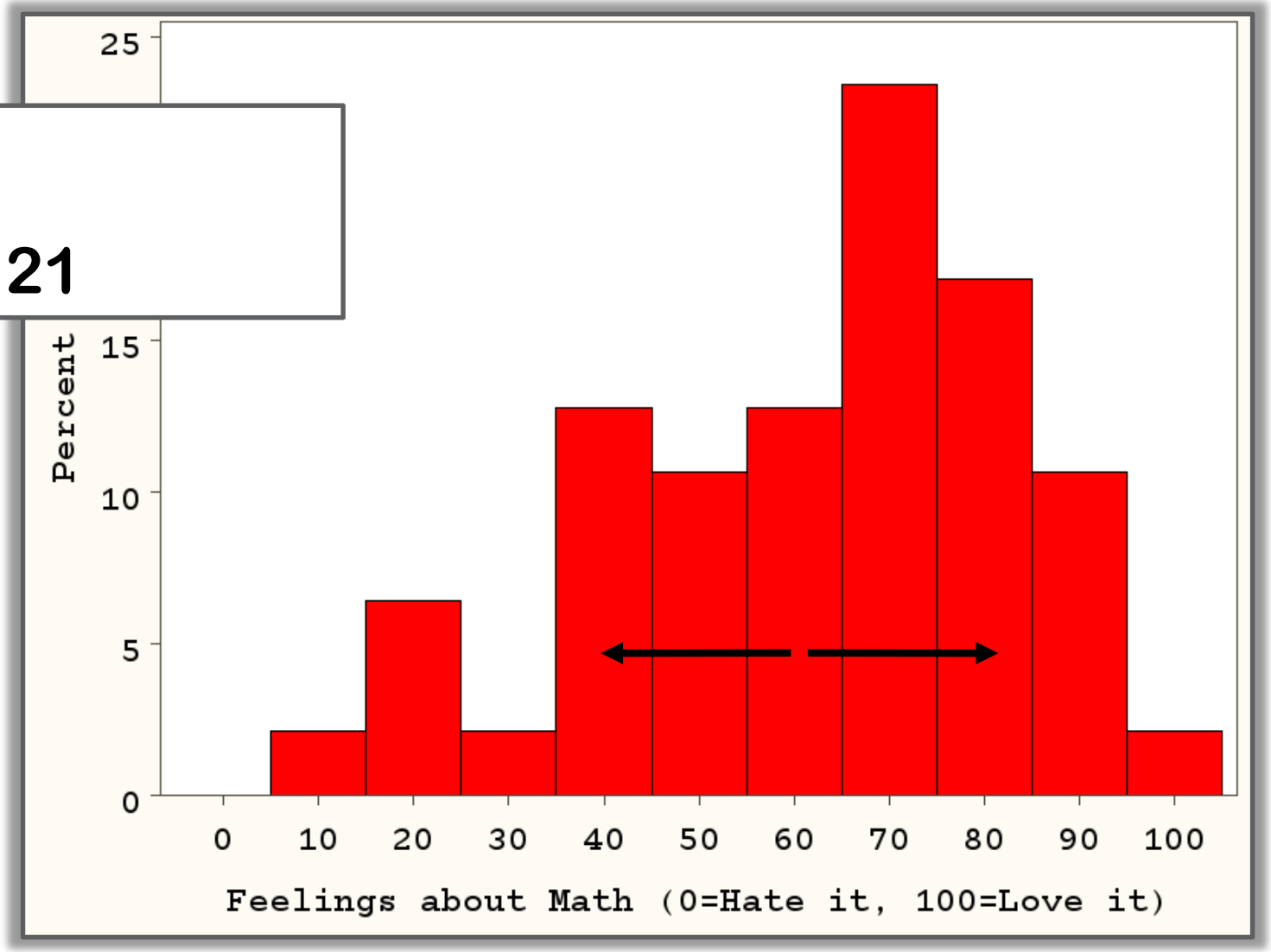
Example: Homework (hours/week)



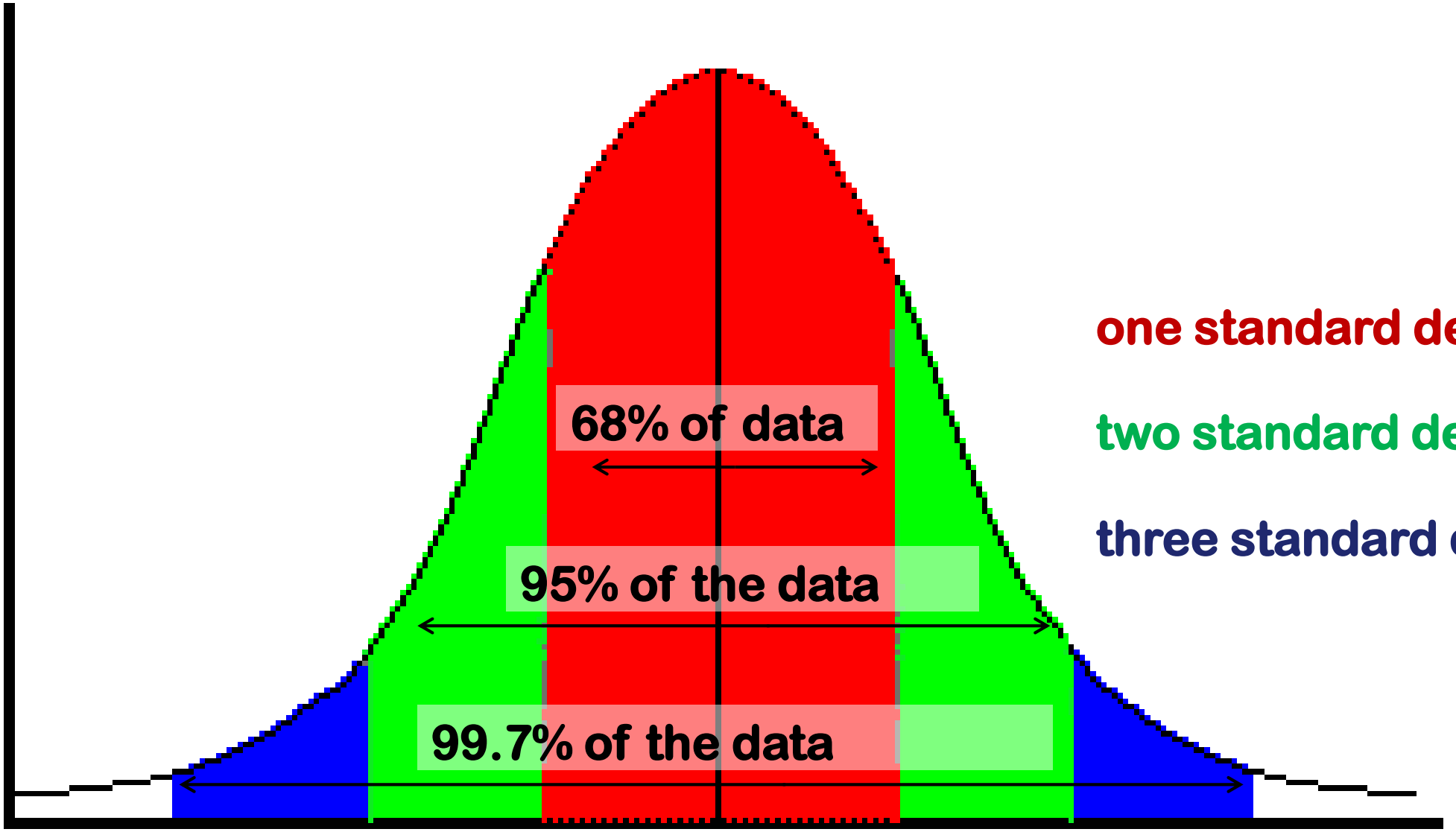
Example: Feelings about math (0=lowest, 100=highest)

Mean = 61

Standard deviation = 21



68 – 95 – 99.7 rule (for a perfect bell curve)



one standard deviation

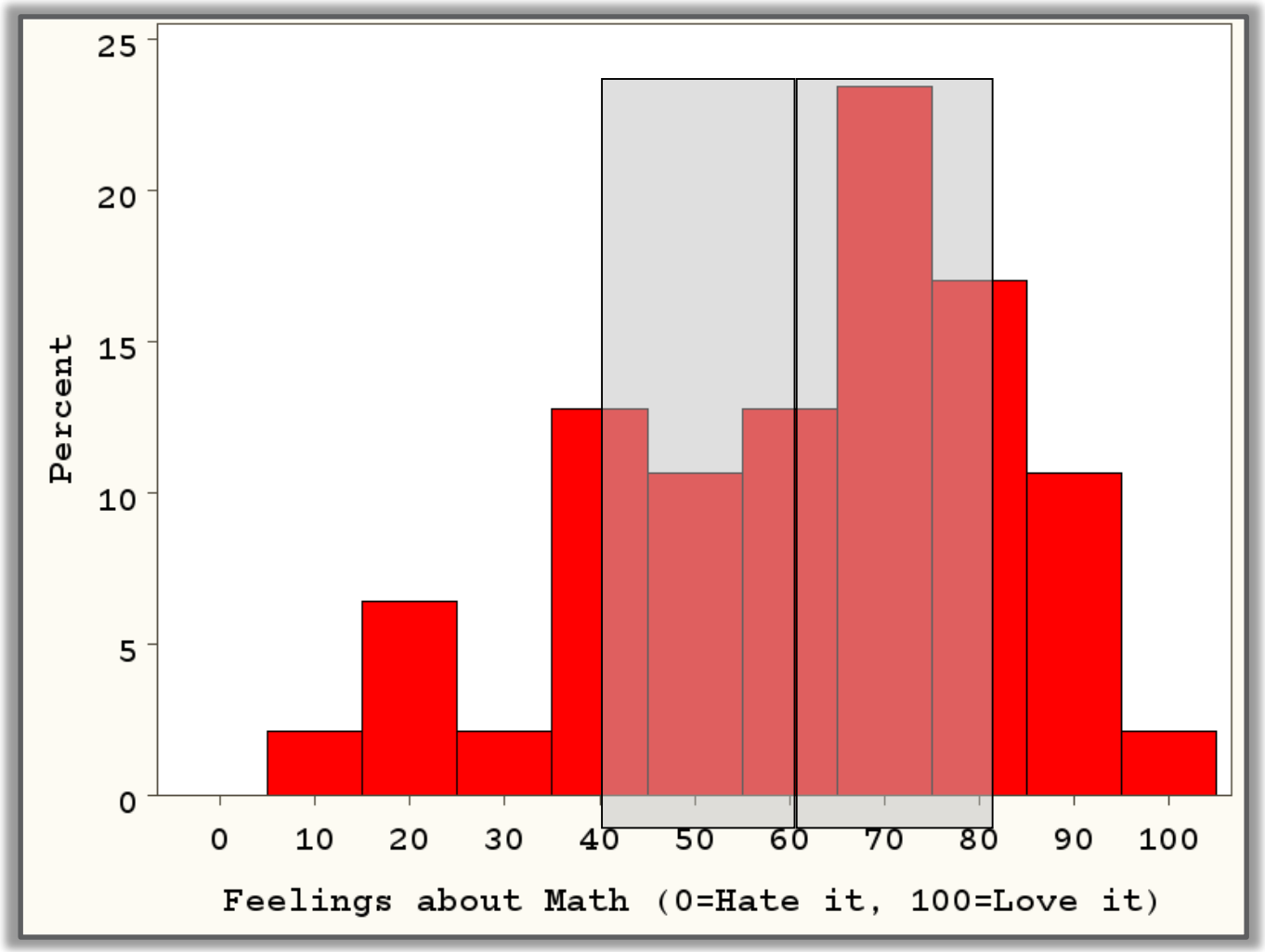
two standard deviations

three standard deviations

One std: Feelings about math (0=lowest, 100=highest)

Mean +/- 1 std = 40 – 82

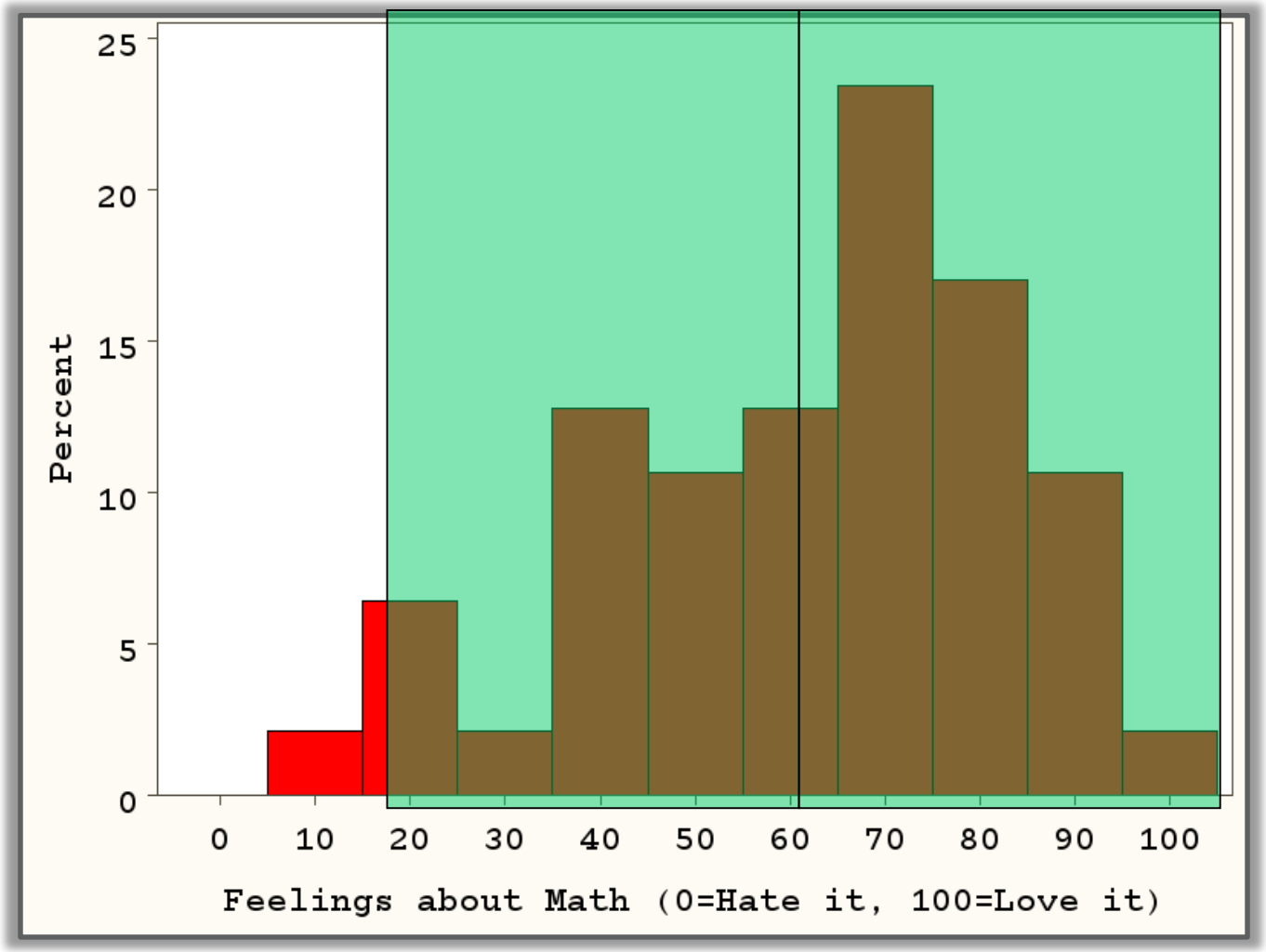
Percent between 40 and 82 = 34/47 = 72%



Two std's: Feelings about math (0=lowest, 100=highest)

Mean +/- 2 std =19 – 100

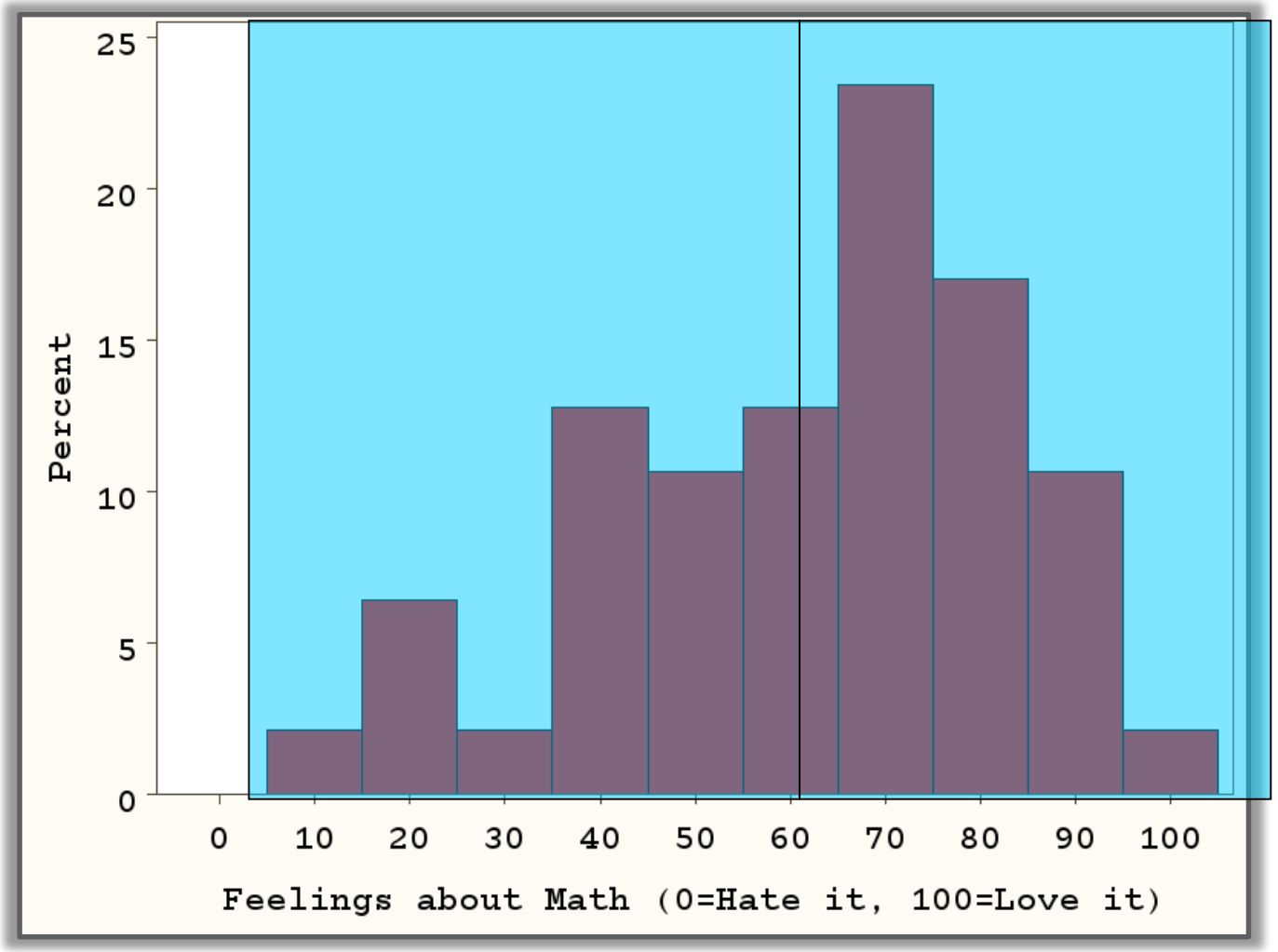
Percent between 19 and 100 = 46/47= 98%



Three std's: Feelings about math (0=lowest, 100=highest)

Mean +/- 3 std = 0 – 100

100% of the data!



Standard deviations vs. Standard errors

- **Standard deviation** measures the variability of a trait.
- **Standard error** measures the variability of a statistic:
  - which is a theoretical construct!

Percentiles

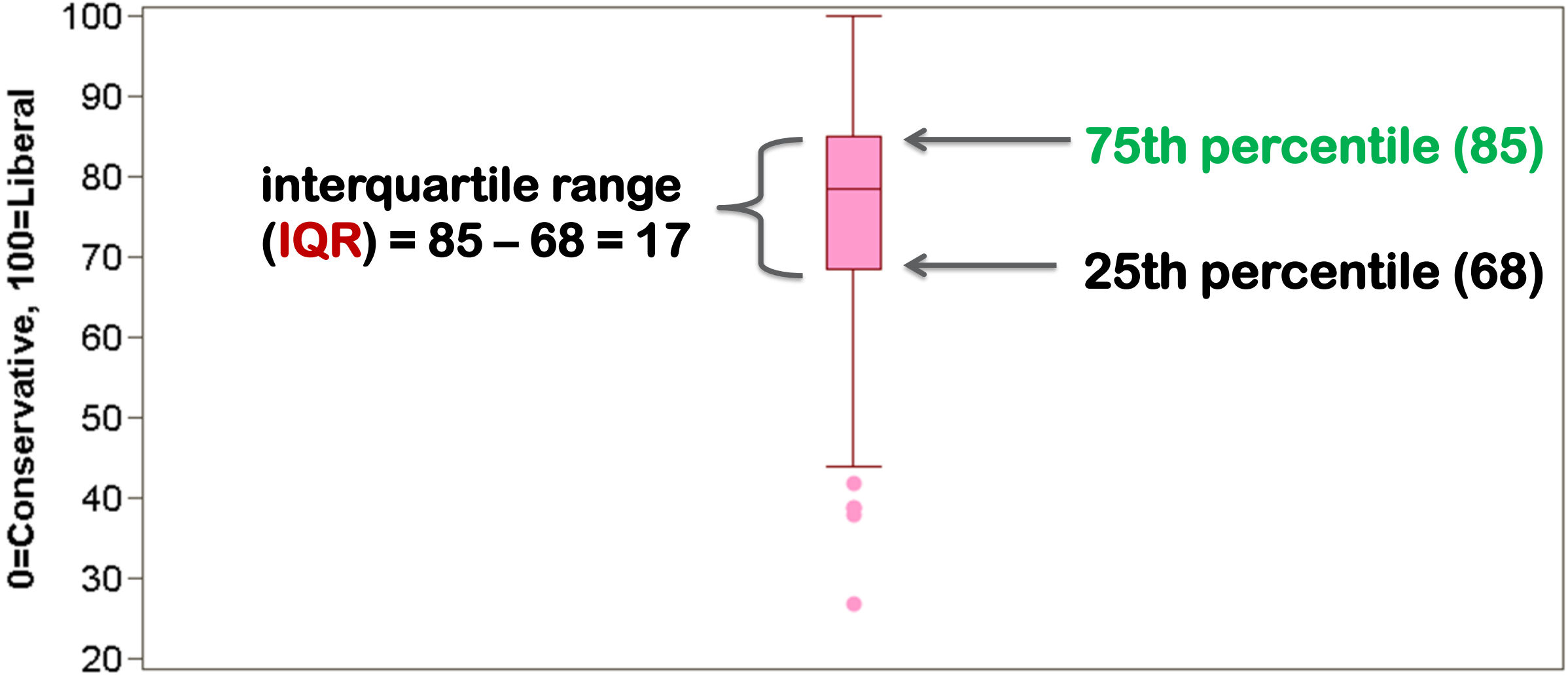
- **Based on ranking the data**
  - The 90^{th} percentile is the value for which 90% of observations are lower
  - The 50^{th} percentile is the median
  - The 10^{th} percentile is the value for which 10% of observations are lower

- **Percentiles are NOT affected by extreme values**
  - unlike standard deviations

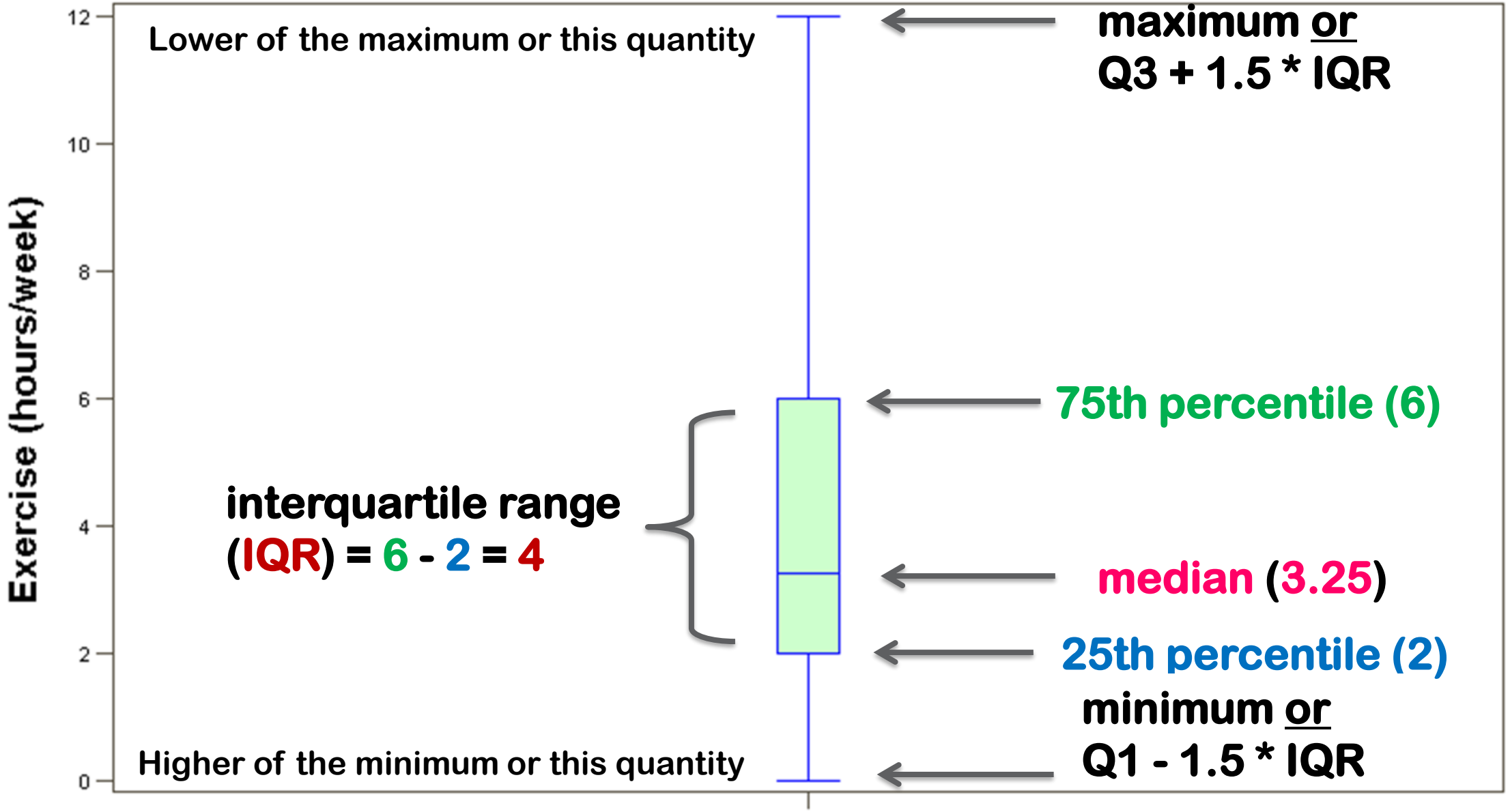
Interquartile Range (IQR)

- Interquartile range = 3rd quartile – 1st quartile
- The middle 50% of the data.
- Interquartile range is not affected by outliers.

Boxplot of Political Bent
(0=Most Conservative, 100=Most Liberal)



Boxplot of Exercise



Tests: Central Tendency and Variability

Comparison	Groups	Normal or Almost Normal	Not Normal	Binomial (Proportions)	Variances
Compare data within one group to a standard or target value	1	One sample t-test	Wilcoxon Rank-Sum test	One proportion z-test (or exact Binomial test)	Chi-square for one variance
Compare data within two unpaired groups	2	Two sample t-test	Mann Whitney Wilcoxon Rank-Sum test (or U-test)	Two proportions z-test, Chi-square test of independence (or Fisher's exact test if counts in cells <5)	F-test for homogeneity of variances
Compare two paired groups	2	Paired t-test	Wilcoxon Rank-Sum test	McNemar's test	Bonett's test
Compare data among many groups	>2	One-Way ANOVA	Kruskal-Wallis test	Chi-square test of independence (or Fisher's exact test)	Levene's test or Bartlett's test for normal data