Math 373   Lecture 2  
Mean, variance, std. dev., Chebyshev, z-score, quartiles  
The mean, median, and mode measure the center of a data set. The range, variance, and standard deviation measure how dispersed or spread out it is.  
When applied to a population, they are parameters; when applied to a sample, they are statistics.  
Definition. For a data set \( x_1, x_2, \ldots, x_n \) of \( n \) numbers:  
\[
\Sigma x_i = x_1 + x_2 + \ldots + x_n.
\]
\( \bar{x} = \mu = \Sigma x_i / n \) is the mean, average or arithmetic mean. 
Use \( \bar{x} \) for samples, \( \mu \) for populations. 

After listing the numbers \( x_i \) in ascending order, the median is the middle number or the average of the two middle numbers. 
The mode (if any) is the most frequently occurring measurement. 
The range \( R \) is the largest number minus the smallest. 
\( \sigma^2 = \Sigma (x_i - \bar{x})^2 / n \) is the variance of a population; 
\( s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1) \) is the variance of a sample. 
The standard deviation or std. dev. of the sample or population is the (positive) square root of the variance: 
\[
s = \sqrt{s^2}, \quad \sigma = \sqrt{\sigma^2}.
\]

For the 9-element set of sample data: \{1, 2, 3, 6, 6, 6, 9, 10, 11\}:  
The arithmetic mean = _____, mode = ____. 
.5(\( n+1 \)) = ____. The median = _____. 
The variance = ____. The standard deviation = _____. 
The range = _____. Learn to do these with your calculator. 

In a group of 21 people, 11 are unemployed and receive $0/month and 10 are employed and receive $2100/month. Does the mean or the median best describe the “average” wage for this group? 

In a group of 11 people, 10 receive $0/month, 1 receives $1100/month. Does the mean or the median best describe the wage earned by the “typical” member of this group? 

The following graphs are smoothed out line graphs of the frequency distributions of IQ scores of four different groups of people: 
groups A, B, C, and D. 
Which group has the most people? 
Which group has the highest average IQ? 
Which group has the largest standard deviation? 

We rarely use the mode. Use the median if the distribution is highly skewed; otherwise use the mean. 

\((x_i - \bar{x})\) measures how far \( x_i \) is from the mean \( \bar{x} \). 
\( \Sigma (x_i - \bar{x})^2 = 0 \), the negative terms cancel the positive. 
Squaring \( \Sigma (x_i - \bar{x})^2 \) the terms prevents this. We divide by \( n \) to get the average, then take the square root to compensate for the squaring. 
The standard deviation measures how far the average data item is from the mean. The variance is easier to work with since it doesn’t involve a square root. But the std. dev. has the same units as the data. 

Empirical Bell Rule. In normal distributions pictured above (bell or mound-shaped curves): About 
- 68% of the data are within 1 std. dev. of the mean. 
- About 95% are within 2 std. dev. 
- About 99% are within 3 std. dev. 

For distributions not known to be normal, we have: 
Chebyshev’s Theorem. At least \( 1 - \frac{1}{k^2} \) of all measurements are within \( k \) std. dev. of the mean. 
Thus for \( k = 2 \), at least \( 1 - \frac{1}{4} = 1 - \frac{1}{4} = \frac{3}{4} \) of all items are within 2 std. dev. of the mean. Other events are unusual. 
For \( k = 3 \), at least \( 1 - \frac{1}{9} = 1 - \frac{1}{9} = \frac{8}{9} \) items are within 3 std. dev. of the mean. Other events are rare. 

Range Approximation. Quite often, \( R \approx 4s \) and \( s \approx R/4 \). 
The z-score of \( x \) measures how many std. devs. \( x \) is from the mean \( \bar{x} \). Thus \( x = \bar{x} + zs \). Solving for \( z \) gives, the z-score of \( x \):  
\[
z = (x - \bar{x}) / s.
\]

The mean is 50, the std. dev. is 10. Find the z-scores: 
\( z \)-score of 30 = ____. \( z \)-score of 55 = ____. \( z \)-score of 50 = _____. 

If you measure the heights of 5 people in inches, you get larger numbers than if you measure them in feet. But the \( z \)-scores would be the same. They are independent of the original units of measurement. 

- The median, the “middle item” is the \( 50^{th} \) percentile \( Q_2 \). Roughly 50% of the data are below it, 50% are above it. 
- The first quartile \( Q_1 \) lies between the lowest 25% and the upper 75%. 
- The third quartile \( Q_3 \) lies between the lowest 75% and the upper 25%. 

If there are \( n \) data items: 
The median, \( Q_2 \), is the value at position \( .5(n+1) \) where interpolation is used when \( .5(n+1) \) is fractional. 
\( Q_1 \) is the value (interpolated) at the position \( .25(n+1) \). 
\( Q_3 \) is the value (interpolated) at the position \( .75(n+1) \). 
IQR, the interquartile range, = \( Q_3 - Q_1 \). 

Six data items: 0, 4, 5, 8, 9, 9. Find the median, \( Q_1, Q_2 \). 

\[\begin{array}{c|cccccccc}
\text{position} & 1 & 2 & 3 & 4 & 5 & 6 \\
\text{data} & 0 & 4 & 5 & 8 & 9 & 9 \\
\end{array}\] 
At \( .5(6+1) = 3.5 \), the median 
\( Q_2 = 5 + (.5)(8 - 5) = 6.5 \). 
At \( .25(6+1) = 1.75 \), \( Q_1 = 0 + (.75)(4 - 0) = 3 \). 
At \( .75(6+1) = 5.25 \), \( Q_3 = 9 + (.25)(9 - 9) = 9 \). IQR = 6.