NAME:____________________

NAME:____________________

NAME:____________________

NAME:____________________

Theorem Checbyshev's theorem states that given a collection of data values, then for any number $k > 1$, the proportion of these data values that fall within k standard deviations of the mean is at least $1 - 1/k^2$

Proof Say that the standard deviation for the collection of data values is σ, N is the number of data values, and μ is the mean. Then, by the definition of variance, it is true that

$$N\sigma^2 = \sum(x - \mu)^2$$

The sum on the right is the sum of squared deviations from the mean. Split now the collection x in two disjoint subset x' and x'' such that the elements in x' fall within k standard deviations from the mean, and the elements in x'' fall further away. Suppose that the subcollection x' has $m \leq N$ elements. Then we have that

$$N\sigma^2 = \sum(x' - \mu)^2 + \sum(x'' - \mu)^2.$$

Basically the sum of the squared deviations from the mean has two contributions. The first contribution could be really small because the datavalues in x' could be even equal to the mean. The second contribution can instead be bounded from below, since we know that each element in x'' is at distance at least $k\sigma$ from the mean. Therefore

$$\sum(x' - \mu)^2 + \sum(x'' - \mu)^2 \geq \sum 0 + \sum(k\sigma)^2 = (N - m)k^2\sigma^2.$$

Putting this together with the previous equality we get

$$N\sigma^2 \geq (N - m)k^2\sigma^2 \Rightarrow N \geq Nk^2 - mk^2 \Rightarrow mk^2 \geq Nk^2 - N \Rightarrow m/N \geq 1 - 1/k^2,$$

and the last inequality is exactly the statement of the theorem.

1. Suppose that a collection of data values has mean 150 and standard deviation 15. At least what proportion of the data values lie between
   - 120 and 180?
   - 132 and 168?
   - 90 and 210?

2. Assume that a collection of 1000 data values has mean 400 and standard deviation 25.
   - At least how many data values lie between 325 and 475?
   - At least how many data values lie between 300 and 500?
   - At most how many data values are smaller than 350 or larger than 450?

3. Can you devise a dataset where the closest element to the mean of the dataset is exactly at one standard deviation from the mean?

4. Can you devise a dataset where 25% of elements fall at 2 standard deviations off of the mean?

5. What is the maximum distance of the third quartile from the mean, measured in standard deviations?

6. What is the maximum length of the interquartile range, measured in standard deviations?

7. What is the maximum distance of the median from the mean, measured in standard deviations?

8. What is the maximum distance of the highest value from the mean, measured in standard deviations?