Summarizing a Data Set

MEASURES OF CENTRAL TENDENCY

Mean: Mean is a measure of central tendency of a numerical data. It is calculated by adding all the values and dividing it by the total number of observations. Consider the following readings of the fasting blood sugar that I measured on myself on 7 different days. The units are in mgDL.

78 89 94 83 91 81 93

The mean of these seven observations is

$$\frac{78+89+94+83+91+81+93}{7}=\frac{609}{7}\cong 87 \text{ mgDL}$$

The mean takes into account each observation in a data set but is sensitive to extreme values.

Median: Another measure of center that is resistant to extreme value in a data set is the median. To find the median, arrange the data set in order, the middle entry is the median. For example, for the fasting blood sugar data, to compute the median

First arrange the data in order

78 81 83 89 91 93 94

The median is the middle entry that is 89 mgDL.

In case the number of entries is even, we use the mean of the middle two entries when the data is in order.

For example if the following data shows the weights of 10 individuals in lb

164, 169, 170, 179, **183, 189**, 194, 198, 211, 318

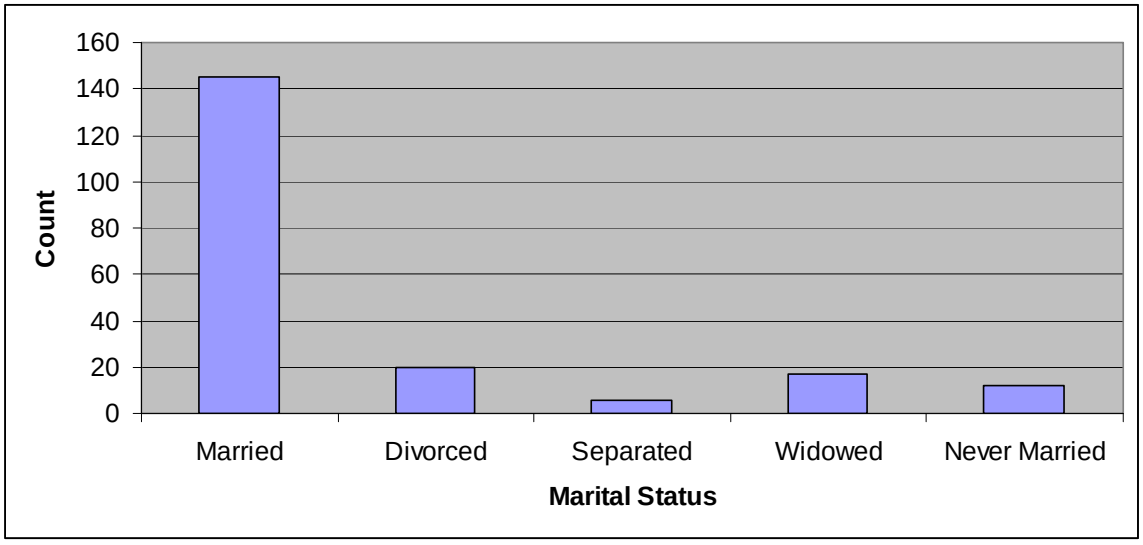
The median is $\frac{183+189}{2}=186$ lb.

The mean is $\frac{164+169+170+179+183+189+194+211+318}{10}=\frac{1975}{10}=197.5$ lb

The use of mean and median depends on the context of an application.

For example, if we are considering this data set of the weights for the people riding in an elevator that can carry at the most ten people, we should look at the mean for the capacity of the elevator. On the other hand, if we are considering this data for the clothes for these individuals, we shall consider the median.

Mode: Mode is the most frequent entry. It is often used to refer to the average of a categorical data set. For example, the mode of the marital status of the subjects in the benign breast cancer study is "Married."



MEASURES OF SPREAD

Standard Deviation: The standard deviation is a measure of spread around the mean. For a sample data with n observations, the standard deviation is computed by using the rule

$s = \sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n-1}}$ where $\bar{x}$ is the mean of the sample data under consideration.

Like the mean, the standard deviation involves each entry in the calculation as shown below.

Let us compute the standard deviation of the fasting blood sugar data that was considered above

78 81 83 89 91 93 94

We computed the mean $\bar{x}$ =87 mgDL

For the standard deviation, use the following table

$\bar{x}$	$\left(x-\bar{x}\right)$	$\left(x-\bar{x}\right)^2$
78	$78-87=-9$	81
81	$81-87=-6$	36
83	$83-87=-4$	16
89	$89-87=2$	4
91	$91-87=4$	16
93	$93-87=6$	36
94	$94-87=7$	49

$\left(x-\bar{x}\right)^2$ =81+36+16+4+16+36+49=238

therefore $s = \sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n-1}} = \sqrt{\frac{238}{7-1}} = \sqrt{\frac{238}{6}} \cong 6.30$ mgDL

Just like the mean, the standard deviation is sensitive to extreme observations.

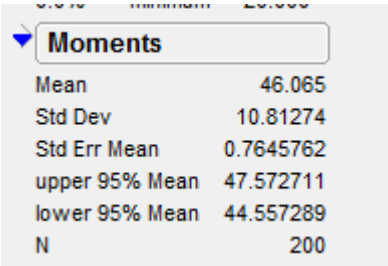
Let us look at a use of the standard deviation as a measure of spread

Recall the data set

Age_50 Sorted

20	22	23	24	25	25	25	26	26	26
26	28	28	29	29	29	30	30	30	30
30	30	30	32	32	33	33	33	34	34
34	34	34	34	34	34	35	35	35	35
36	36	36	36	36	37	37	37	38	38
38	39	39	39	40	40	40	41	41	41
42	42	42	42	42	42	42	42	43	43
43	43	43	43	44	44	44	44	44	44
45	45	45	45	45	45	45	45	46	46
46	46	46	46	46	46	47	47	47	47
47	47	47	47	47	48	48	48	48	48
48	48	48	48	49	49	49	49	49	49
49	49	50	50	50	50	50	50	51	51
51	51	51	51	51	52	52	52	52	52
53	53	53	53	53	53	53	53	53	54
54	54	54	55	55	55	55	55	56	56
56	57	57	57	57	57	57	57	57	57
57	58	58	58	59	59	59	60	60	60
60	61	61	61	61	62	62	62	63	63
63	63	64	64	64	64	65	65	68	69

from the previous lesson. We can use JMP7 to obtain the mean and the standard deviation of this data (refer to the **instructions to use JMP7**)



We round the mean to 46 years and the standard deviation to 11 years, and are going to look at the percentage of the data that lies within one standard deviation of the that is between

46—11=35 years and 46+11=57 years

20	22	23	24	25	25	25	26	26	26
26	28	28	29	29	29	30	30	30	30
30	30	30	32	32	33	33	33	34	34
34	34	34	34	34	34	35	35	35	35
36	**36**	**36**	**36**	**36**	**37**	**37**	**37**	**38**	**38**
38	**39**	**39**	**39**	**40**	**40**	**40**	**41**	**41**	**41**
42	**42**	**42**	**42**	**42**	**42**	**42**	**42**	**43**	**43**
43	**43**	**43**	**43**	**44**	**44**	**44**	**44**	**44**	**44**
45	**45**	**45**	**45**	**45**	**45**	**45**	**45**	**46**	**46**
46	**46**	**46**	**46**	**46**	**46**	**47**	**47**	**47**	**47**
47	**47**	**47**	**47**	**47**	**48**	**48**	**48**	**48**	**48**
48	**48**	**48**	**48**	**49**	**49**	**49**	**49**	**49**	**49**
49	**49**	**50**	**50**	**50**	**50**	**50**	**50**	**51**	**51**
51	**51**	**51**	**51**	**51**	**52**	**52**	**52**	**52**	**52**
53	**53**	**53**	**53**	**53**	**53**	**53**	**53**	**53**	**54**
54	**54**	**54**	55	55	55	55	55	56	56
56	57	57	57	57	57	57	57	57	57
57	58	58	58	59	59	59	60	60	60
60	61	61	61	61	62	62	62	63	63
63	63	64	64	64	64	65	65	68	69

Note that 113 out of 200 OR 56.5% of the data is within one standard deviation of the mean

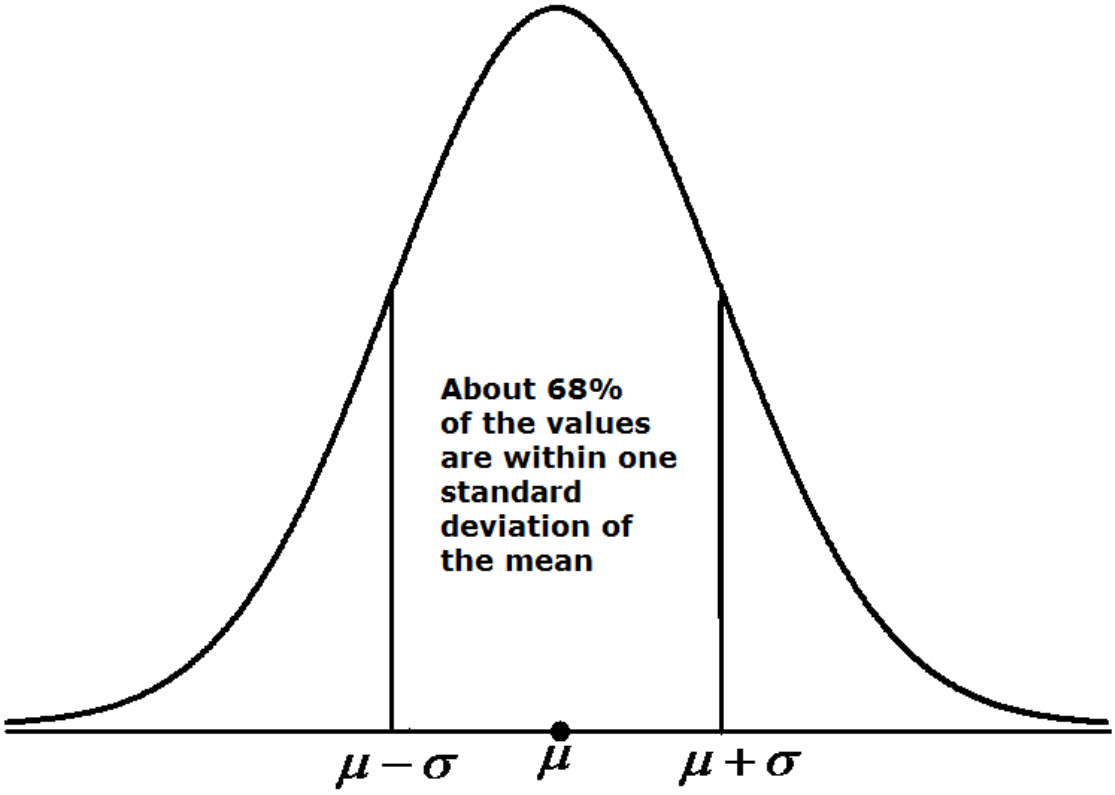
If we count the percentage within two standard deviations of the mean, i.e. between 46—2*11=24 years and 46+2*11=68 years

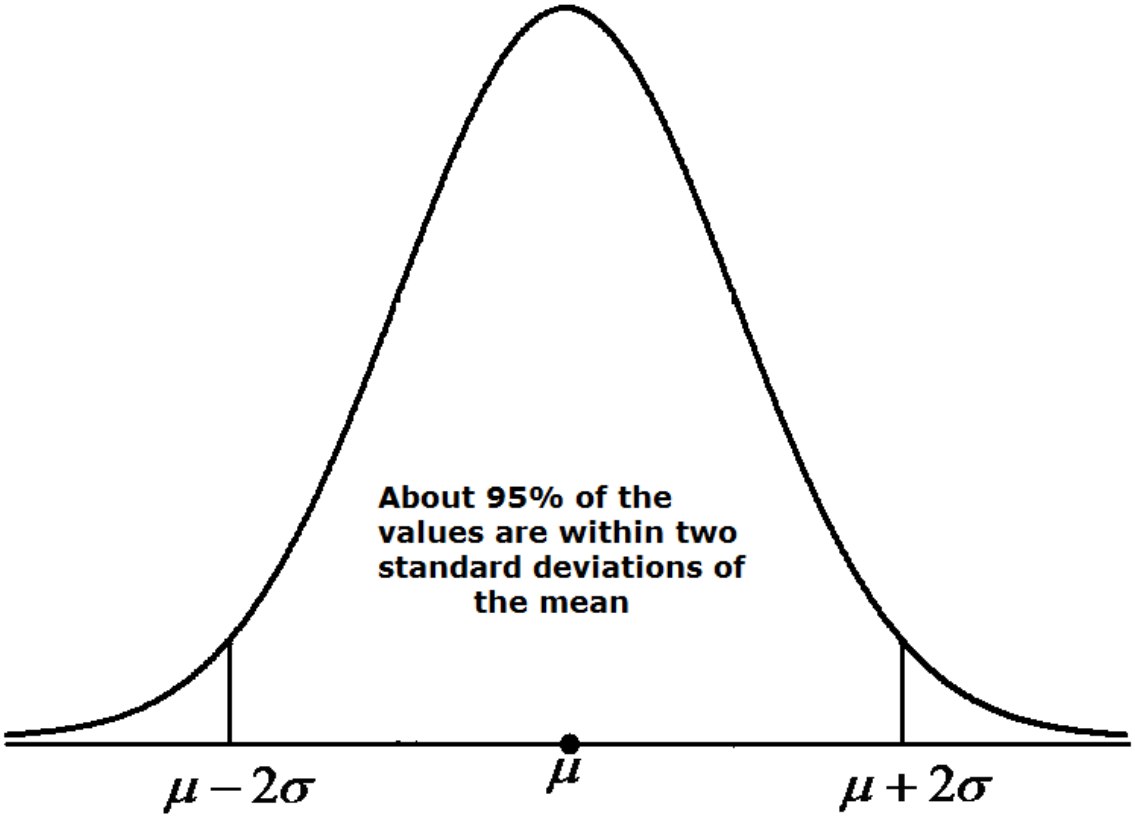
20	22	23	24	**25**	**25**	**25**	**26**	**26**	**26**
26	**28**	**28**	**29**	**29**	**29**	**30**	**30**	**30**	**30**
30	**30**	**30**	**32**	**32**	**33**	**33**	**33**	**34**	**34**
34	**34**	**34**	**34**	**34**	**34**	**35**	**35**	**35**	**35**
36	**36**	**36**	**36**	**36**	**37**	**37**	**37**	**38**	**38**
38	**39**	**39**	**39**	**40**	**40**	**40**	**41**	**41**	**41**
42	**42**	**42**	**42**	**42**	**42**	**42**	**42**	**43**	**43**
43	**43**	**43**	**43**	**44**	**44**	**44**	**44**	**44**	**44**
45	**45**	**45**	**45**	**45**	**45**	**45**	**45**	**46**	**46**
46	**46**	**46**	**46**	**46**	**46**	**47**	**47**	**47**	**47**
47	**47**	**47**	**47**	**47**	**48**	**48**	**48**	**48**	**48**
48	**48**	**48**	**48**	**49**	**49**	**49**	**49**	**49**	**49**
49	**49**	**50**	**50**	**50**	**50**	**50**	**50**	**51**	**51**
51	**51**	**51**	**51**	**51**	**52**	**52**	**52**	**52**	**52**
53	**53**	**53**	**53**	**53**	**53**	**53**	**53**	**53**	**54**
54	**54**	**54**	**55**	**55**	**55**	**55**	**55**	**56**	**56**
56	**57**	**57**	**57**	**57**	**57**	**57**	**57**	**57**	**57**
57	**58**	**58**	**58**	**59**	**59**	**59**	**60**	**60**	**60**
60	**61**	**61**	**61**	**61**	**62**	**62**	**62**	**63**	**63**
63	**63**	**64**	**64**	**64**	**64**	**65**	**65**	68	69

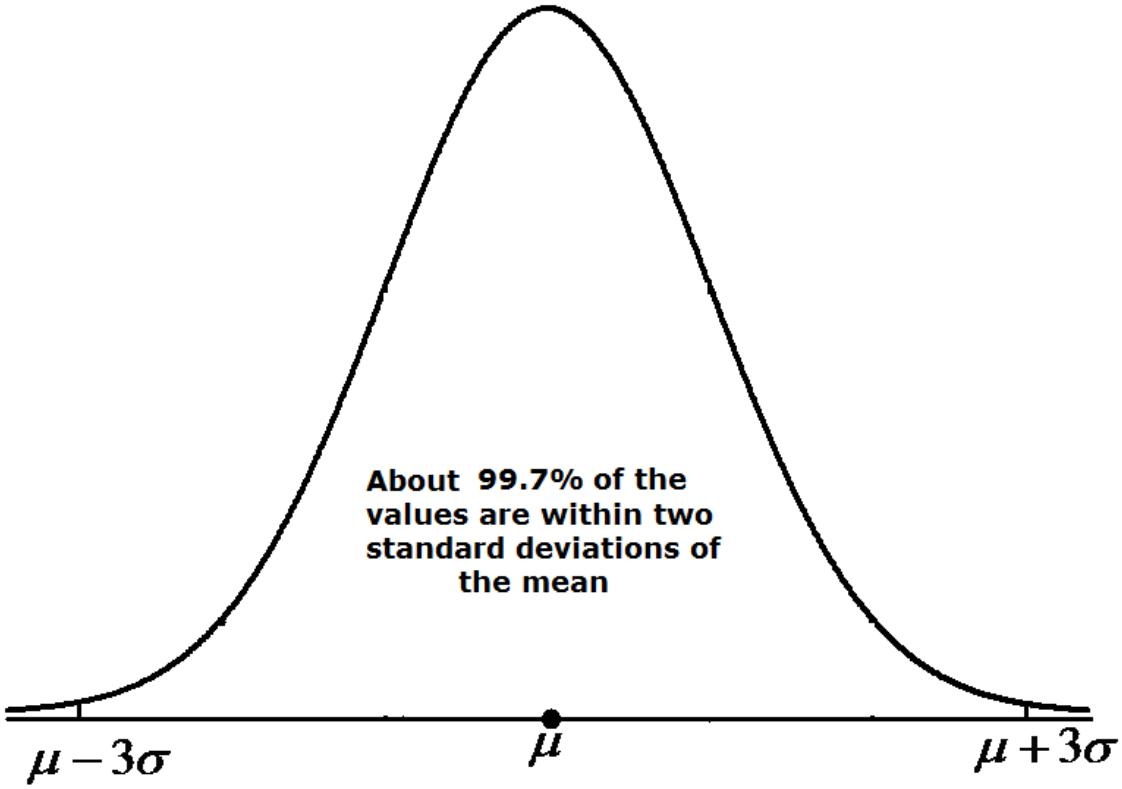
194 out of 200 that is 97% of the values are within two standard deviations of the mean.

If we count the percentage within THREE standard deviations of the mean, i.e. between 46—3*11=13 years and 46+3*11=77 years note that 100% of the values are covered in this range.

Theoretically: If a distribution is normal (traditionally known as bell shaped) with the mean μ and standard deviation σ then




About 95% of the values are within two standard deviations of the mean
$\mu - 2\sigma$
μ
$\mu + 2\sigma$


About 99.7% of the values are within two standard deviations of the mean
$\mu - 3\sigma$
μ
$\mu + 3\sigma$

Since the standard deviation is not resistant to the outliers, we can obtain other measures of spread that depend upon the quantiles. For the age data that we are discussing, they are shown below as a JMP7 output

Quantiles

100.0%	maximum	69.000
99.5%		68.995
97.5%		64.000
90.0%		60.000
75.0%	quartile	54.000
50.0%	median	47.000
25.0%	quartile	38.000
10.0%		30.000
2.5%		25.000
0.5%		20.010
0.0%	minimum	20.000

20	22	23	24	25	25	25	26	26	26
26	28	28	29	29	29	30	30	30	30
30	30	30	32	32	33	33	33	34	34
34	34	34	34	34	34	35	35	35	35
36	36	36	36	36	37	37	37	38	38

25% of the values are smaller than 38

38	39	39	39	40	40	40	41	41	41
42	42	42	42	42	42	42	42	43	43
43	43	43	43	44	44	44	44	44	44
45	45	45	45	45	45	45	45	46	46
46	46	46	46	46	46	47	47	47	47

50% of the values are less than 47

47	47	47	47	47	48	48	48	48	48
48	48	48	48	49	49	49	49	49	49
49	49	50	50	50	50	50	50	51	51
51	51	51	51	51	52	52	52	52	52
53	53	53	53	53	53	53	53	53	54

75% of the values are less than 54

54	54	54	55	55	55	55	55	56	56
56	57	57	57	57	57	57	57	57	57
57	58	58	58	59	59	59	60	60	60
60	61	61	61	61	62	62	62	63	63
63	63	64	64	64	64	65	65	68	69

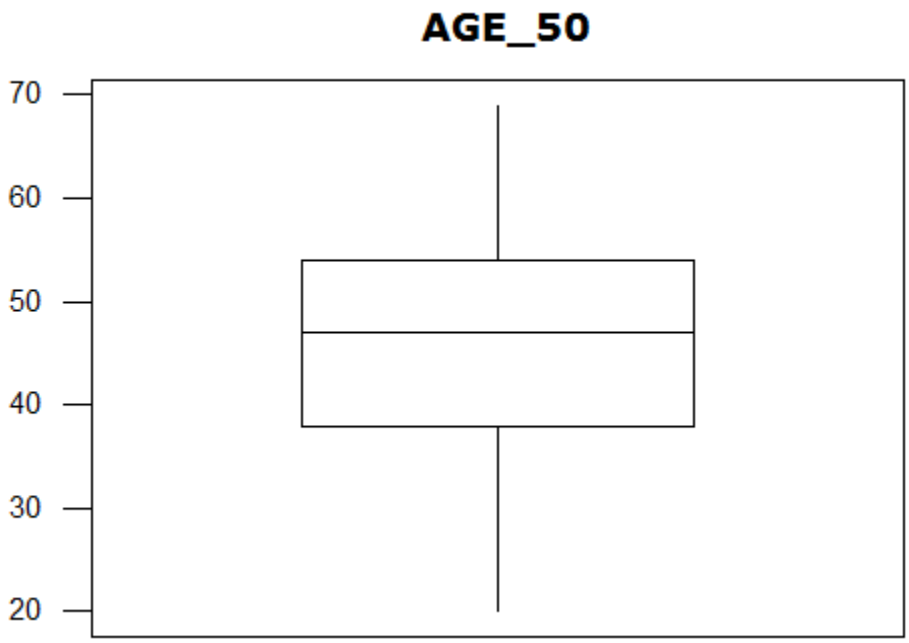
The quantiles are used to write a five number summary of a numerical data set according to the

lowest First Quartile Q_1 Median Third Quartile Q_3 Largest

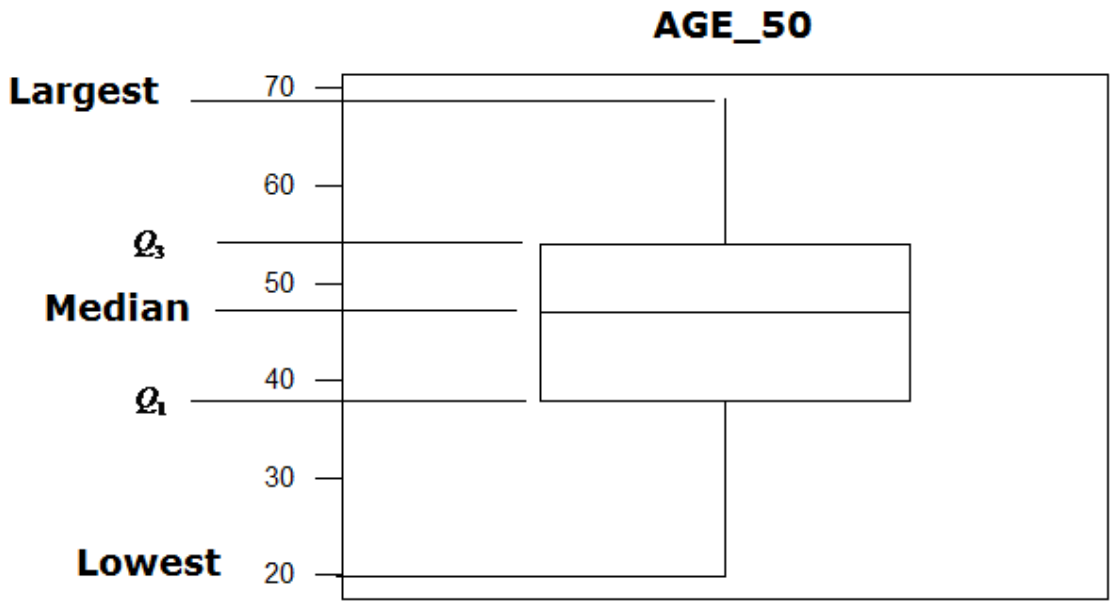
The five number summary for the age_50 data set is

20 38 47 54 69

A five number summary is displayed in a boxplot as shown below



This boxplot has been created by using the Five Number Summary as shown below.



The Inter Quartile Range that equals $Q_3 - Q_1$ gives us the spread of the middle 50% of the data.

For AGE_50 data, The Inter Quartile Range is 54—38=16 years

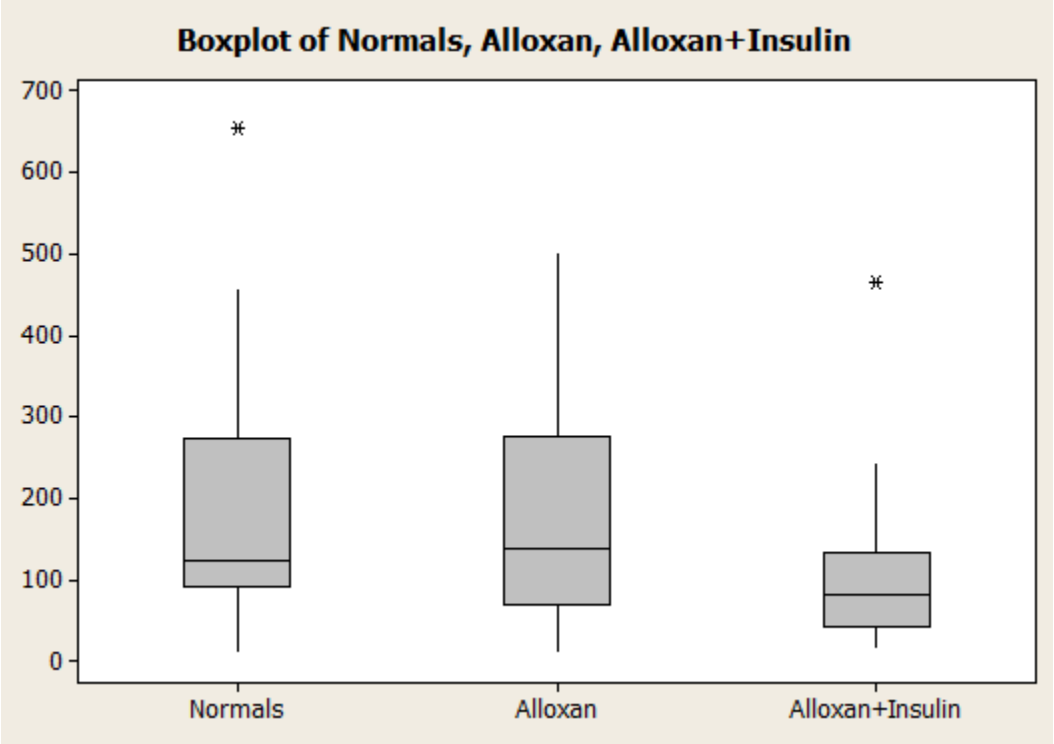
The boxplots may be used to compare different populations. To see this, consider the following data from

Dolkert R.E., Halperin, B. E., and Perlman J. (1971) Comparison of antibody responses in normal and alloxan diabetic mice, *Diabetes* , **20** , 162-167

The columns show the amounts of nitrogen-bound bovine serum albumen produced by three groups of diabetic mice, normal, alloxan diabetic, and alloxan diabetic treated with insulin

Row	Normals	Alloxan	Alloxan+Insulin
1	156	391	82
2	282	46	100
3	197	469	98
4	297	86	150
5	116	174	243
6	127	133	68
7	119	13	228
8	29	499	131
9	253	168	73
10	122	62	18
11	349	127	20
12	110	276	100
13	143	176	72
14	64	146	133
15	26	108	465
16	86	276	40
17	122	50	46
18	455	73	34
19	655		44
20	14		

A look at side by side boxplots for these measures helps us compare the three groups.



The asterisks denote outliers.

1.5*IQR criterion to determine outlers

The quartiles give us a criterion for determining outliers. We can designate a value in the data set an outliers if it is either

less than $Q_1 - 1.5(Q_3 - Q_1)$

or it is more than $Q_3 + 1.5(Q_3 - Q_1)$

For Example, consider the Five Number Summary of the Age_50 data set

Lowest=20 Q_1=38 Median=47 Q_3=54 Highest=69

$$Q_1 - 1.5(Q_3 - Q_1) = 38 - 1.5 \times 16 = 38 - 24 = 14$$
$$Q_3 + 1.5(Q_3 - Q_1) = 54 + 1.5 \times 16 = 54 + 24 = 78$$

Since there are no values in the data set below 14 or above 78, there are no outliers in the data set according to this criterion.