Chapter 2

Summary statistics: Measures of Location and Dispersion

Def^n: Parameters

Numbers that describes populations characteristics are called parameters

Def^n: Estimate and Estimator

Values calculated from a sample of data that are used to estimate population parameters are called estimates. The formula used to calculate an estimate is called an estimator.

An estimator is a function, while an estimate is a specific value.

The mean, median and mode are three different numerical measures used to describe the centre of a data set. The range, the variance, the standard deviation, and the mean absolute deviation are four different measures used to measure the amount of spread, dispersion, or variability in a set of data.

2.1 The mean, the median and the mode

Def^n: Population mean

The population mean is denoted by μ and is calculated by using the following formula where N is the number of observations of the popalation

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Def^n: Sample mean
Sample mean is denoted by $\bar{x}$. Given the sample of n observations $x_1, x_2, \ldots x_n$ the sample mean $\bar{x}$ is

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Def^n: Median
The median is the middle value of data ordered from lowest to highest. If the sequence contains an odd number of observations, the median is the middle value in the ordered sequence. If the sequence contains an even number of observations, the median is the arithmetic average of the two central values.

When a population of data is available, we calculate the population median. The median calculated from a sample of data is called the sample median and is an estimate of the population median.

e.g. On a sample of 7 days, # of patients at a walk-in clinic

$x_1 = 92, x_2 = 84, x_3 = 70, x_4 = 76, x_5 = 66, x_6 = 80, x_7 = 71$

sample mean $\bar{x} = \frac{\sum x_i}{7} = \frac{539}{7} = 77$

Order the observations from lowest to highest

66 70 71 76 80 84 92

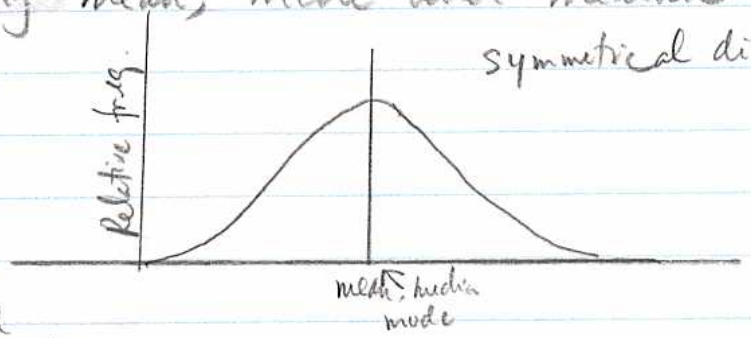
obs. = 7 which is odd, so the median is the middle or fourth value in ordered sequence.

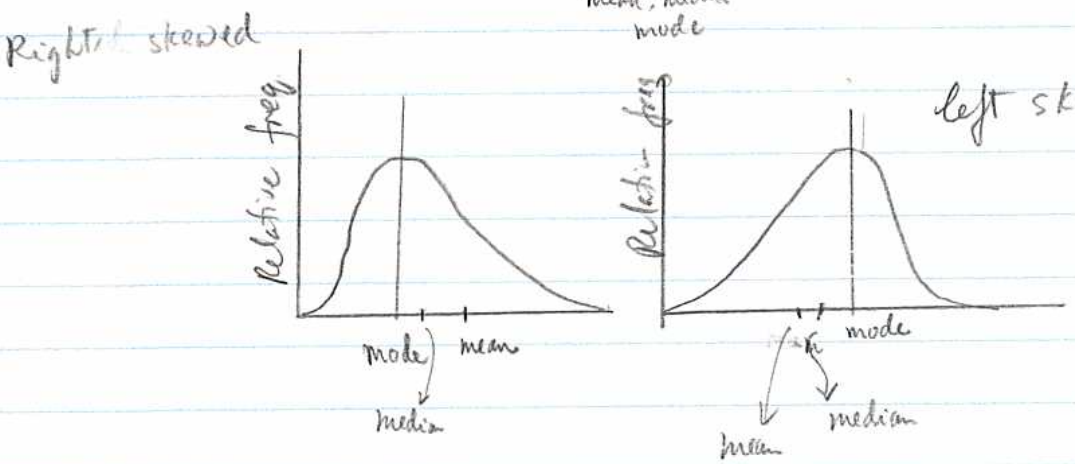
Sample median = 76

Def^{n}: Mode
The mode of a set of observations is the value that occurs with the greatest frequency. The mode is NOT necessarily unique.

e.g. Our walk-in clinic example, each of the values occurs exactly once. Because each value has the same frequency, each value could be a mode. In this problem, the mode conveys NO useful information.

Relationship among mean, mode and median





When a distribution is unimodal and symmetric like the bell-shaped normal distribution, the mean, median and mode all coincide. If the distribution is skewed to the right, the mean is larger than the mode because it is affected by the large values in the tail, whereas

the mode is influenced only by the most frequently occuring values, which are small. The median is less influenced by the right-hand tail than is the mean, so the median will be less than the mean. Furthermore, the peak of a right-skewed distribution will be left of the value that divides the area under the curve into two equal parts, and so the median will exceed the mode.

i.e. (a) skewed to the right:

mode < median < mean

(b) skewed to the left:

mean < median < mode.

Calculate the mean of Grouped Data

Let k denote # of groups or classes, let $M_1, M_2, \ldots M_k$ denote the midpoints of the k classes. Let $f_1, f_2 \ldots f_k$ denote the frequencies of the k classes. The mean is estimated using the formula

$$\bar{x} = \frac{\sum f_i M_i}{n} = \frac{f_1 M_1 + \cdots + f_k M_k}{n}$$

Calculate the median of Grouped Data

1. Determine which class contains the median
2. Denote the lower boundary of this class by A & upper boundary by B
3. Let c denote # of observations less than A, and let d denote # of observations between A and B
4. Let n denote the sample size

If the data are in frequency distribution, estimate the median using the following formula:

Estimate the Median from a Frequency Distribution

$$\text{Median} = A + \frac{n/2 - c}{d}(B - A)$$

e.g.

Monthly Salary	Frequency (f_i)	Midpoint (M_i)	Product ($f_i M_i$)
1300 - under 1400	8	1350	10,800
1400 - under 1500	11	1450	15,950
1500 - under 1600	17	1550	26,350
1600 - under 1700	19	1650	31,350
1700 - under 1800	10	1750	17,500
1800 - under 1900	6	1850	11,100
1900 - under 2000	4	1950	7,800
2000 - under 2100	5	2050	10,250
TOTAL	80	1,0000	131,100

$\Sigma f_i M_i = 131,100$,

The estimated mean $\bar{x} = \frac{\Sigma f_i M_i}{n} = \frac{131\,100}{80} = 1638.75$

$$\text{Median} = 1600 + \frac{\frac{80}{2} - 36}{19}(1700 - 1600)$$

$$= 1600 + 0.2105 \times 100 = 1621.05$$

Measure of dispersion

Summary statistics that measured the amount of variation in a data set are called measure of dispersion.

(a) Range
The range of a set of observations is the difference between the largest & the smallest value.

(b) Interquantile Range
For any population or sample of values, the interquantile range is $Q_3 - Q_1$.
Q_3 : 75th percentile (upper Quartile)
Q_1 : 25th " " (lower Quantile).

(c) Deviation from the Mean.
For any value x, the deviation from the mean is difference $(x-\mu)$ if the population mean is known or $(x-\bar{x})$ if the sample mean is used.

(d) Mean absolute deviation

$$M.A.D. = \frac{\sum |x_i - \bar{x}|}{n}$$

e.g. On a sample of 7 days, # of patients in a walk in clinic:
92, 84, 70, 76, 66, 80, 71

$\bar{x} = 77$

$$M.A.D = \frac{|92-77| + |84-77| + |70-77| + |76-77| + |66-77| + |80-77| + |71-77|}{7}$$

$$= 7.143$$

(e) Variance and Standard Deviation

Def^n: population variance is calculated using

$$\sigma^2 = \frac{\Sigma (x_i - \mu)^2}{N}$$

μ: population mean
N: population size

Def^n: sample variance is calculated using

$$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$$

where $\bar{x}$: sample mean
n: sample size

$$= \frac{\Sigma x_i^2 - n\bar{x}^2}{n-1}$$

The use of $n-1$ rather than n in the denominator of s^2 should be explained. Recall that $\bar{x}$ and s^2 are calculated from sample data and are estimates of the population parameter μ and σ^2. In general, the sample will not exactly reproduce the characteristics of the population, and it is unlikely that $\bar{x}$ will be exactly equal to μ.

When estimating a population parameter, we want to use a procedure that has no tendency to over- or underestimate the population value on the average. If an estimator equals the population value on the average (i.e overestimate & underestimate balance out), then we have an unbiased estimator.

s^2 has $n-1$ rather than n as its denominator so that it will be unbiased estimator of σ^2. If n were used as the denominator, on the average s^2 would be smaller than σ^2 and thus be a biased estimator. The reason for this bias

Quartiles and Percentiles:

Definition: Percentiles

Suppose the observations x1,x2,……xn, have been arranged in ascending order. The pth percentile is the value xp such that p percent of the observations are less than or equal to xp and (1-p) percent of the observations are greater than or equal to xp.

Lower quartile(or first quartile) is the 25^{th} percentile of the data. i.e. 25% of the data below the first quartile and thus 75% above it. The first quartile is denoted by Q1.
Upper quartile(or third quartile) is the 75^{th} percentile of the data. i.e 75% of the data below the third quartile and thus 25% above it. The third quartile is denoted by Q3.
Median(or second quartile) is the 50^{th} percentile of the data. i.e. 50% of the data below the median and 50% lie above it.

Procedure for Calculating Percentiles-Ungrouped Data:

1. Arrange the n observations in ascending order. The smallest observation is given rank 1, the second smallest observation is given rank 2, and so forth.
2. Calculate the index i=(p x n)/100, where p is the percentile of interst and n is the sample size.
3. If i is an integer, the pth percentile is the arithmetic average of the values having ranks i & (i+1).
4. If I is not an integer, the next integer value greater than I denotes the rank of the value which is the pth percentile.

For example:

1. Calculate the 70^{th} & 80^{th} percentiles for the following 15 values:

   18,14,45,32,65,43,25,41,83,51,26,36,40,55,20

   Solution:
   Arrange the 15 values in ascending order:

   14,18,20,25,26,32,36,40,41,43,45,51,55,65,83

   For the 70^{th} percentiles, the index I : (70x15)/100=10.5

   Because I is not an integer, the position of the 70^{th} percentile is the next integer greater than 10.5, which is 11, that is, the 11^{th} position. The 70^{th} percentile corresponds to the 11^{th} data value is 45.

   For the 80^{th} percentile, the index I: (80x15)/100=12.0

   Because I is an integer, the 80^{th} percentile is the average of the data values in positions 12 & 13. Thus the 80^{th} percentile is (51+55)/2=53

2. The following 24 observations have been ordered from lowest to highest:

6	8	12	17	20	22	25	26	26	28	30
32	34	37	40	50	61	63	65	67	69	80
82	86									

Find the first and third quartiles.

Solution:

For the lower quartile, calculate the index with p=25 and n=24. Using the formula I=(p x n)/100, we get I=6.0. Because I is an integer, Q1 is the average of observations having ranks 6 and 7. Because the 6^{th} observation is 22 and the 7^{th} obseration is 25, the lower quartile is Q1=(22+25)/2=23.5.

For the upper quartile, calculate the index with p=75 and n=24, we have I=18.0. Because I is an integer, Q3 is the average of observations having ranks 18 and 19. The 18^{th} observation is 63 and 19^{th} observation is 65. The upper quartile Q3 is (63+65)/2=64.0.

Procedure for Calculating Percentiles-Grouped Data:

If the data have been grouped as a frequency distribution, the pth percentile can only be approximated because the indivdual observations are not known. The pth percentile is approximated by using the following procedure:

1. Find the interval that contains the pth percentile.
2. Let A and B denote the lower and upper boundaries of this interval.
3. Let c denote the number of observations less than A.
4. Let d denote the number of observations between A and B.
5. Let n denote the sample size and let p denote the percentile desired. Let m=n x p/100.
6. The pthe percentile can be approximated as the vale

Pth percentile=A + (m-c) (B-A)/d

The formula is the result of interpolation between points A and B.

There are several methods of calculating percentiles under the Procedure-Examine of SPSS-please see my SPSS handouts. For the sake of simplicity & popularity, we are using **aempirical method**.

The Stem-and-Leaf Plot:

Whenever we are presented with a set of numerical observations, it is natural for us to want to condense or simplify the data in some way so that we can recognize general characteristics of the data. For example, most people would like to know the following things:

1. How are the data distributed? What is the general shape of the distribution? Where are most of the observations concentrated?
2. Where is the center of the data?
3. How spread out, or dispersed, are the data about this central value?
4. Are there any very large or very small values that appear to be quite different from the other observations?

The most frequently used technique for answering these questions is to construct a frequency or relative frequency distribution and to calculate some summary statistics. Another closely related way of answering these questions is to construct what is called a **stem-and-leaf** diagram, a technique proposed by John Tukey. The purpose of the **stem-and-leaf diagram** are as follows:

1. To show the range of the data, that is, to make it easy to determine the difference between the largest observation and the smallest observation;

2. To show how the data are distributed;

3. To show in a general way where the center of the data is;

4. To show in a general way how spread out the data are; and

5. To show whether there are any extremely high or low values, called outliers.

Example: Construction of a Stem-and-leaf Diagram:

The data below show the ages of 60 employees in a company:

43	34	21	26	57	64	51	30	38	60
19	26	41	47	58	50	42	32	25	18
36	27	38	37	24	49	56	61	20	39
32	51	60	52	45	44	33	25	29	27
31	35	39	19	20	30	46	47	32	24
29	22	20	37	34	29	60	55	30	41

In a stem-and-leaf diagram, the first or leading digit serves as the stem and the trailing digit serves as the leaf. The leading digit determines the row in which an observation is placed;the trailing digit is written in this row to the right of a vertical bar that separates the stems from the leaves. In this example, the stems are the first digits for each person's age; thus the stems are the digits 1 through 6. For example, for a person aged 53, the stem is 5 and the leaf is 3. To improve the appearance of the diagram, it is useful to arrange the leaves in each row from lowest to highest. The stem-and-leaf diagram for the age data above is shown:

1	8 9 9
2	0 0 0 1 2 4 4 5 5 6 6 7 7 9 9 9
3	0 0 0 1 2 2 2 3 4 4 5 6 7 7 8 8 9 9
4	1 1 2 3 4 5 6 7 7 9
5	0 1 1 2 5 6 7 8
6	0 0 0 1 4

The Boxplot:

Both the histogram and the stem-and-leaf plot provide useful information about the distribution of observed values.We can see how tightly cases cluster together. We can see if there is a single peak or several peaks. We can determine if there are extreme values.

A display that further summarizes information about the distribution of the values is the boxplot. Instead of plotting the actual values, a boxplot displays summary statistics for the distribution. It plots the median, the 25^{th} percentile, the 75^{th} percentile, and values that are far removed from the rest. Following is an example of a boxplot from the lower boundary of the box is the 25^{th} percentile, and the upper boundary is 75^{th} percentile. A line in the box represents the median. Fifty percent of the cases have values within the box. The length of the box corresponds to the interquartile range(IQR), which is the difference between the 75^{th} and 25^{th} percentile.

The next step is to decide how far away from the median a measurement must lie before it is classified as an outlier. The decision is made by constructing two sets of **fences**. The **inner fences** are located 1.5 IQR below Q1 and 1.5 IQR above Q3. The **inner fences** are defined as follows:

Lower inner fence=Q1 - 1.5(IQR)
Upper inner fence=Q3 + 1.5(IQR)

The second set of fences, the **outer fences**, are located 3 IQR below Q1 and 3 IQR above Q3:

Lower outer fence=Q1 - 3(IQR)

Upper outer fence=Q3 + 3(IQR)

Values lying between an inner fence and the corresponding outer fence are considered to be 'mild' outliers; values outside the outer fences are considered to be 'extreme' outliers.

Example:

Here is a set of data on ages of 50 patients:

39 37 32 40 40 37 39 35 42 43
40 41 31 38 38 37 24 34 42 47
40 51 33 37 36 29 39 44 49 37
34 37 39 38 35 36 43 45 38 42
41 41 56 40 42 48 37 42 39 41

2.16

```
-> get file='d:\stat601.14\p133.sav'.

-> examine variables=age/percentiles(25,50,75)=aempirical/
->      plot=boxplot stemleaf.
```

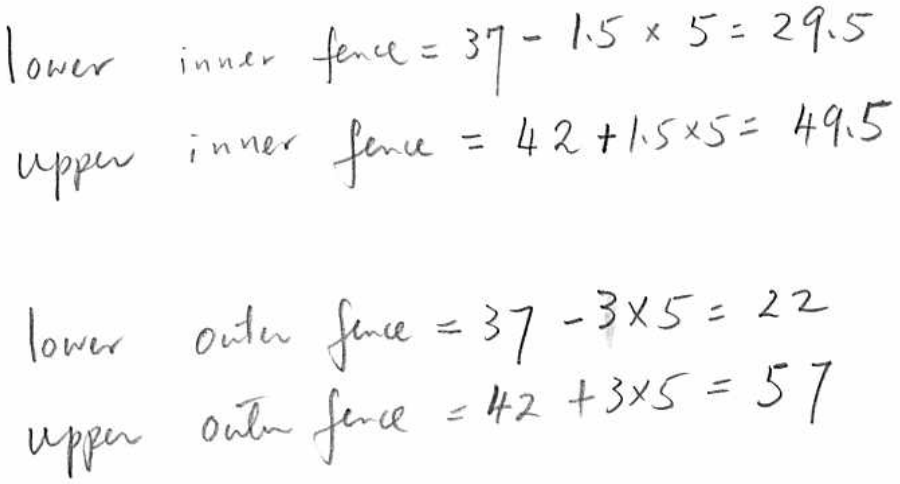
AGE

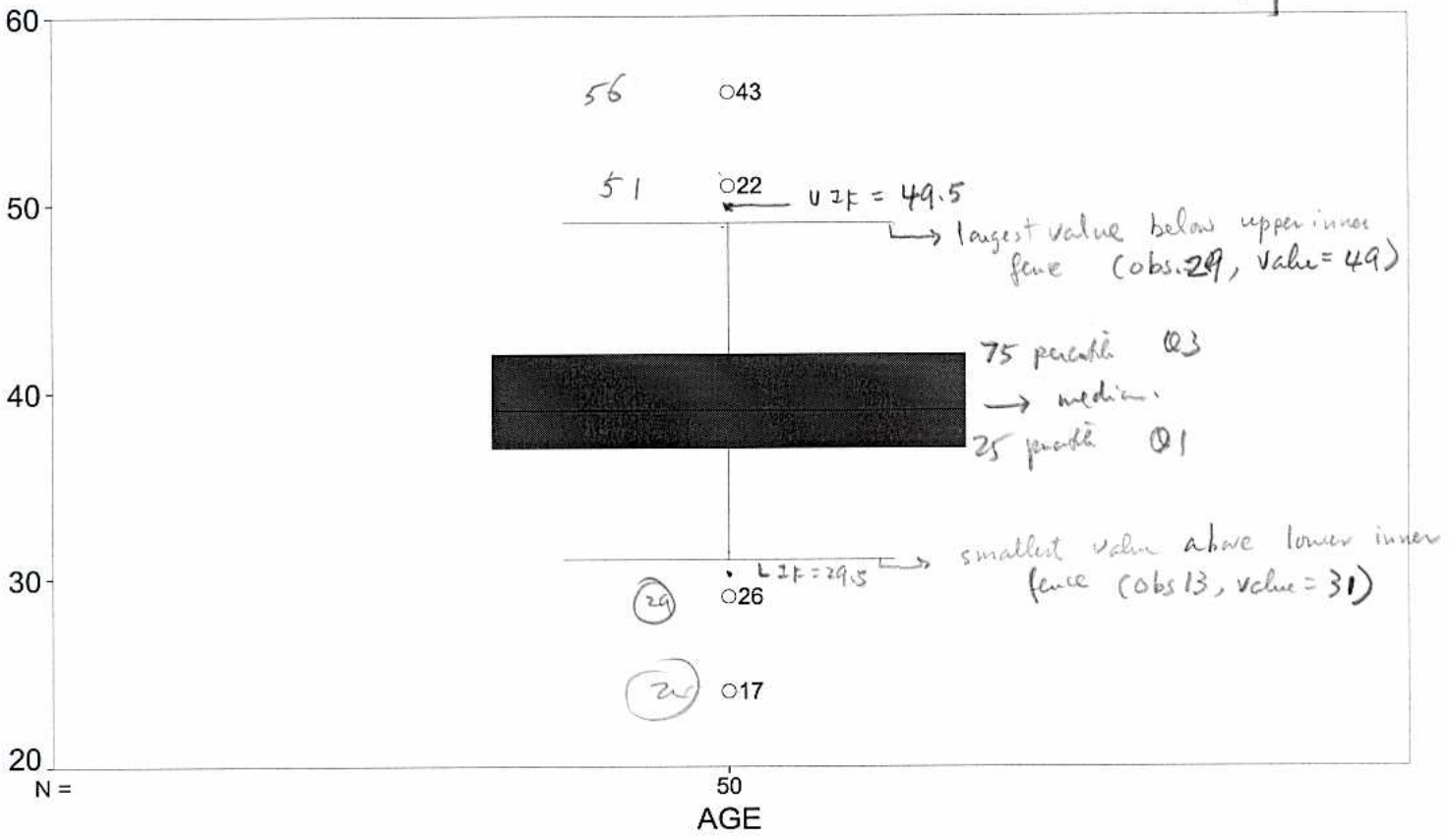
Valid cases:	50.0	Missing cases:	.0	Percent missing:	.0		
Mean	39.3000	Std Err	.7663	Min	24.0000	Skewness	.3059
Median	39.0000	Variance	29.3571	Max	56.0000	S E Skew	.3366
5% Trim	39.2222	Std Dev	5.4182	Range	32.0000	Kurtosis	2.0199
95% CI for Mean	(37.7602, 40.8398)			IQR	5.0000	S E Kurt	.6619

```
Frequency    Stem &  Leaf

     2.00 Extremes    (24), (29)
     1.00         3 *  1
     2.00         3 t  23
     4.00         3 f  4455
     9.00         3 s  667777777
     9.00         3 .  888899999
     9.00         4 *  000001111
     7.00         4 t  2222233
     2.00         4 f  45
     1.00         4 s  7
     2.00         4 .  89
     2.00 Extremes    (51), (56)

Stem width:     10.00
Each leaf:       1 case(s)
```

	Percentiles		
	25.0000	50.0000	75.0000
Percentiles			
Aempirical	37.0000	39.0000	42.0000
Tukey's Hinges	37.0000	39.0000	42.0000

Hi-Res Chart # 7:Boxplot of age




60
50
40
30
20
N =
50
AGE
56
43
51
22
26
17
29
27
UIF = 49.5
largest value below upper inner fence (obs. 29, value = 49)
75 percentile Q3
median
25 percentile Q1
LIF = 29.5
smallest value above lower inner fence (obs 13, value = 31)

Further explanation of Boxplot example(i.e. age of 50 patients):

OBS	AGE
1.00	39.00
2.00	37.00
3.00	32.00
4.00	40.00
5.00	40.00
6.00	37.00
7.00	39.00
8.00	35.00
9.00	42.00
10.00	43.00
11.00	40.00
12.00	41.00
13.00	31.00
14.00	38.00
15.00	38.00
16.00	37.00
17.00	24.00
18.00	34.00
19.00	42.00
20.00	47.00
21.00	40.00
22.00	51.00
23.00	33.00
24.00	37.00
25.00	36.00
26.00	29.00
27.00	39.00
28.00	44.00
29.00	49.00
30.00	37.00
31.00	34.00
32.00	37.00
33.00	39.00
34.00	38.00
35.00	35.00
36.00	36.00
37.00	43.00
38.00	45.00
39.00	38.00
40.00	42.00
41.00	41.00
42.00	41.00
43.00	56.00
44.00	40.00
45.00	42.00
46.00	48.00
47.00	37.00
48.00	42.00
49.00	39.00
50.00	41.00

Sort ages:

ORDER	OBS	AGE
1.00	17.00	24.00
2.00	26.00	29.00
3.00	13.00	31.00
4.00	3.00	32.00
5.00	23.00	33.00
6.00	18.00	34.00
7.00	31.00	34.00
8.00	8.00	35.00
9.00	35.00	35.00
10.00	25.00	36.00
11.00	36.00	36.00
12.00	2.00	37.00
13.00	6.00	37.00
14.00	16.00	37.00
15.00	24.00	37.00
16.00	30.00	37.00
17.00	32.00	37.00
18.00	47.00	37.00
19.00	14.00	38.00
20.00	15.00	38.00
21.00	34.00	38.00
22.00	39.00	38.00
23.00	1.00	39.00
24.00	7.00	39.00
25.00	27.00	39.00
26.00	33.00	39.00
27.00	49.00	39.00
28.00	4.00	40.00
29.00	5.00	40.00
30.00	11.00	40.00
31.00	21.00	40.00
32.00	44.00	40.00
33.00	12.00	41.00
34.00	41.00	41.00
35.00	42.00	41.00
36.00	50.00	41.00
37.00	9.00	42.00
38.00	19.00	42.00
39.00	40.00	42.00
40.00	45.00	42.00
41.00	48.00	42.00
42.00	10.00	43.00
43.00	37.00	43.00
44.00	28.00	44.00
45.00	38.00	45.00
46.00	20.00	47.00
47.00	46.00	48.00
48.00	29.00	49.00
49.00	22.00	51.00
50.00	43.00	56.00

→ 22 mild

→ 29.5 lower inner fence

→ 49.50 upper inner fence

mild

57 OF