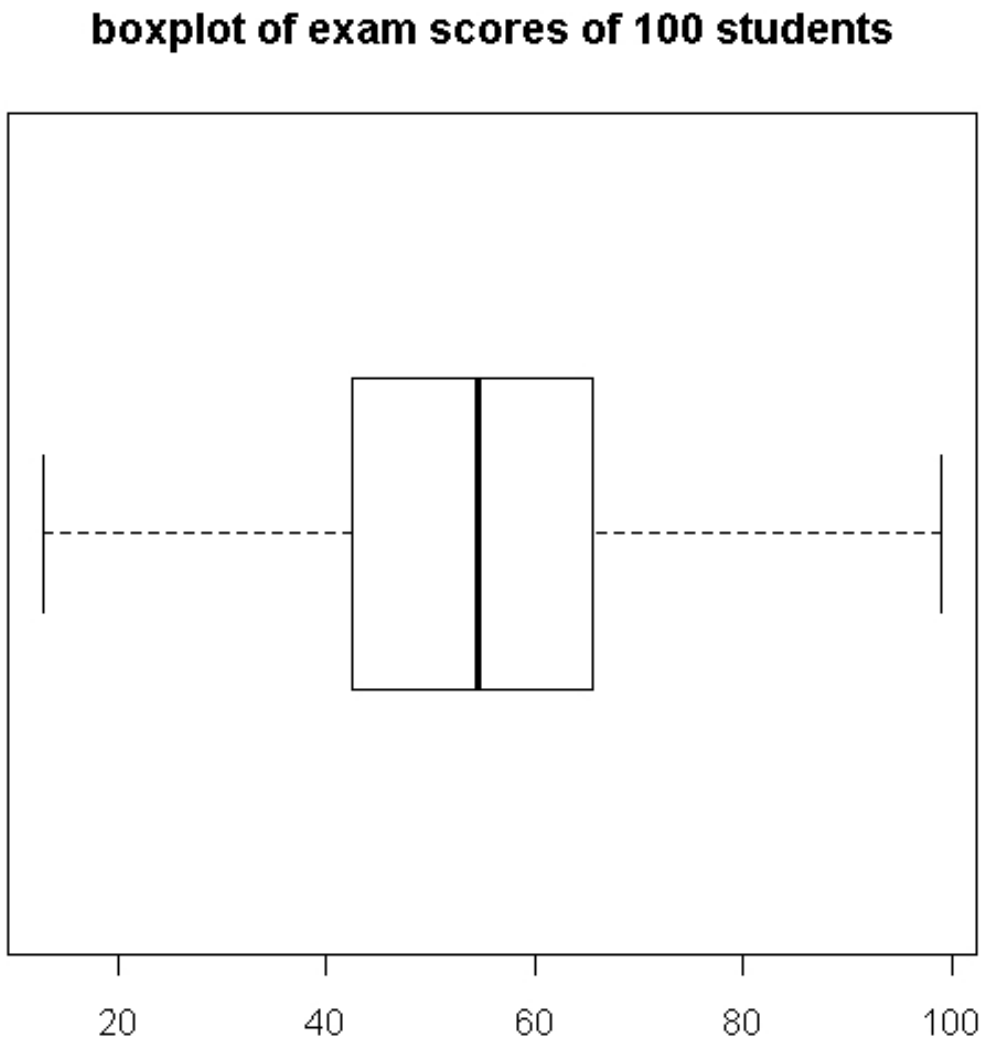
Boxplots

- A graphical method we will be looking at today is the 'box-and-whisker' plot (commonly just referred to as 'boxplots')
- The boxplots is a useful tool for assessing the distribution of a dataset, by means of a visual summary.
- consider the data set of the exam scores of 100 students (see next slide).
- The quartiles of the data set were $Q_1 = 42.5$, $Q_2 = 54.5$ (with Q_2 being the median), and $Q_3 = 65.5$ respectively.
- The interquartile range is $Q_3 - Q_1 = 23$
- The boxplot of the distribution is featured on the second next slide.

Table: Exam results of 100 students

13	21	22	23	24	25	26	28	29	30
31	32	33	34	35	36	36	36	37	38
39	41	41	41	42	43	44	44	44	45
45	46	47	49	50	51	51	52	53	53
53	53	53	54	54	54	54	54	54	54
55	55	55	56	56	56	57	57	58	59
62	63	63	63	63	64	64	64	64	64
65	65	65	65	65	66	66	66	67	69
71	71	72	72	73	74	75	76	76	76
77	82	84	85	87	88	91	91	92	99

Boxplots



Boxplots

- The boxplot is a visual summary containing important aspects of a distribution.
- The main component of the plot , the '**box**', stretches from the **lower hinge**, defined as Q_1, to the **upperhinge**, defined as Q_3 .
- (**Important:**) The median is shown as a line across the box.
- Therefore the box contains the middle half of the scores in the distribution.
- 1/4 of the distribution is between the median line and the upper hinge. Similary 1/4 of the distribution is between the median line and the lower hinge.

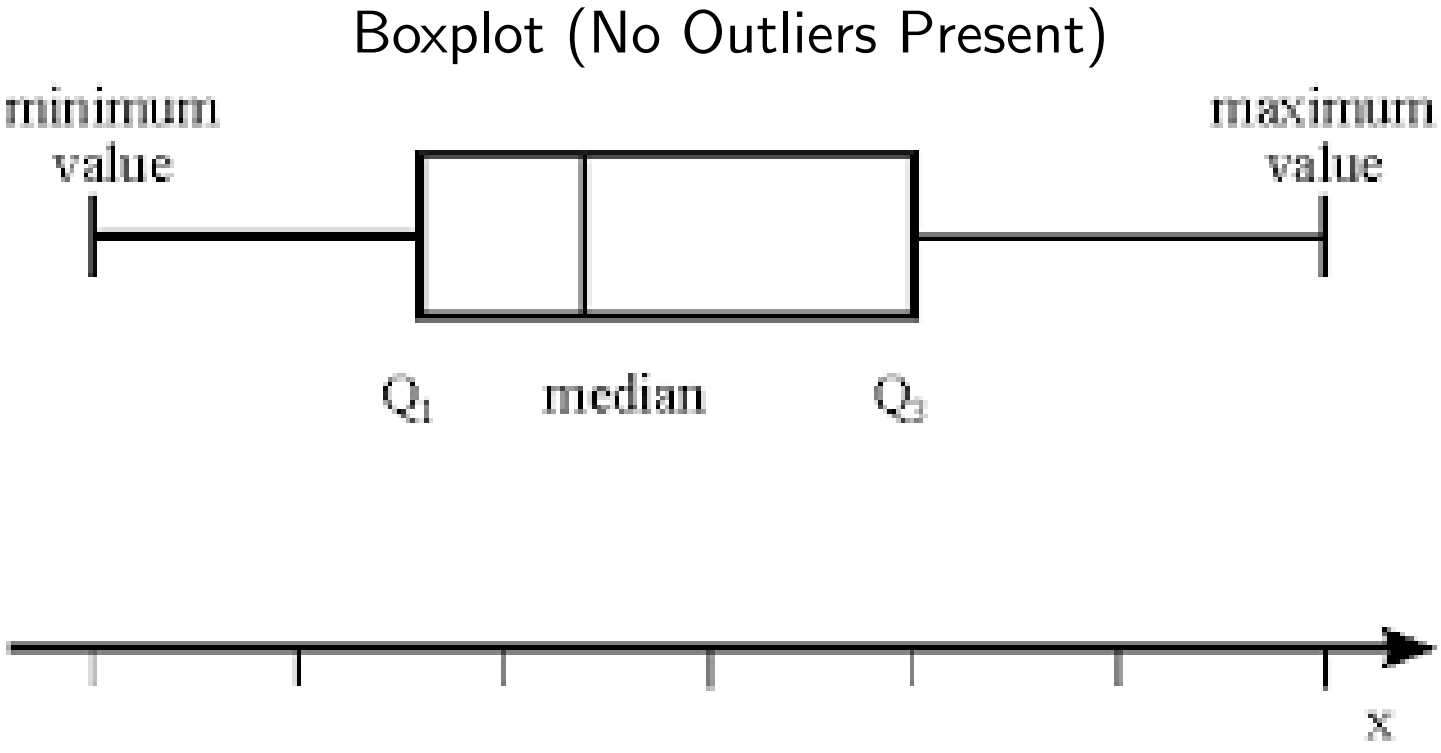
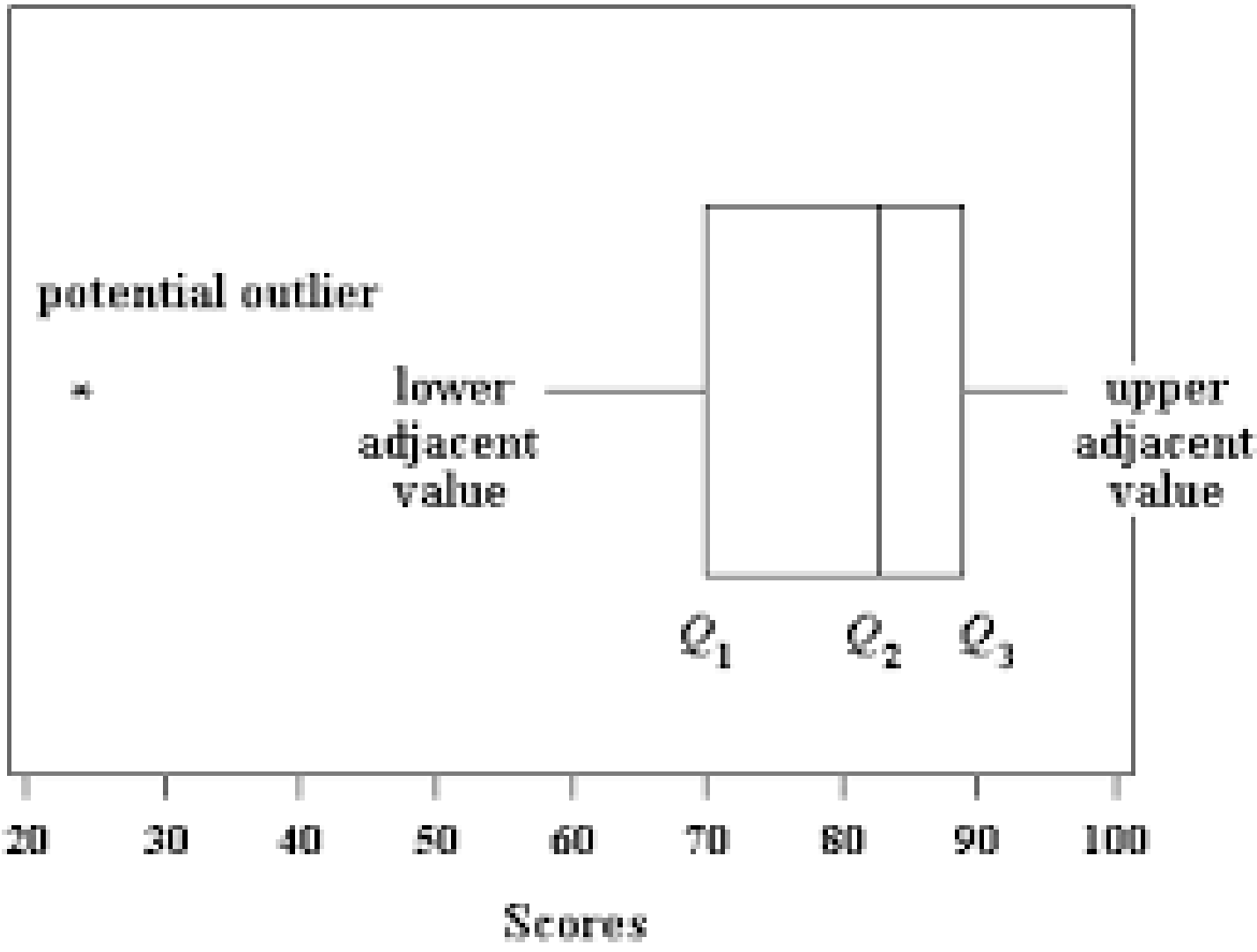


Figure: Structure of Boxplots

Boxplots

- Any value considered to be an outlier should be indicated with an asterisk or a small circle.
- We will see an example of a boxplot with outliers on the next slide.

Boxplots can be used to indicate presence of outliers.



Boxplots

- On either side of the box are the **whiskers**.
- To find where to place the whiskers, we must first compute the location of the **fences**, and determine whether or not there are any **outliers** present.
- Firstly, we must compute the location of the **lower fence**.

$$\text{Lower Fence} = Q_1 - 1.5 \times IQR$$

- For our example, the lower fence is

$$\text{Lower Fence} = 42.5 - 1.5 \times 23 = 42.5 - 34.5 = 8$$

- *(Fences were called adjacent values on diagram on previous slide)*

Boxplots

Detecting Outliers

- The lower fence is used to determine whether there are any outliers in the lower half of the data set.
- **Important:** If there is any observed value less than the lower fence, it is considered an outlier.
- **Important:** The first whisker is drawn at the location of the lowest value that is not considered an outlier.
- If no values are considered outliers, then the whisker is drawn at the location of the smallest value of the dataset.
- For our dataset, the lowest value is 13, which is not less than the lower fence.
- Therefore we draw the first whisker, a vertical line, at this location. A horizontal line is drawn connecting the location of this whisker to Q_1.

Boxplots

- Now we must compute the location of the **upper fence**.

$$\text{Upper Fence} = Q_3 + 1.5 \times IQR$$

- For our example, the upper fence is

$$\text{Upper Fence} = 65.5 + 1.5 \times 23 = 65.5 + 34.5 = 100$$

Boxplots

- The upper fence is used to determine whether there are any outliers in the upper half of the data set.
- If there is any observed value greater than the upper fence, it is considered an outlier.
- The second whisker is drawn at the location of the highest value that is not considered an outlier.
- If no values are considered outliers, then the whisker is drawn at the location of the highest value of the dataset.
- For our dataset, the highest value is 99, which is less than the upper fence.
- Therefore we draw the second whisker , a vertical line, at this location.
- A horizontal line is drawn connecting the location of this whisker to Q_3.

Boxplots

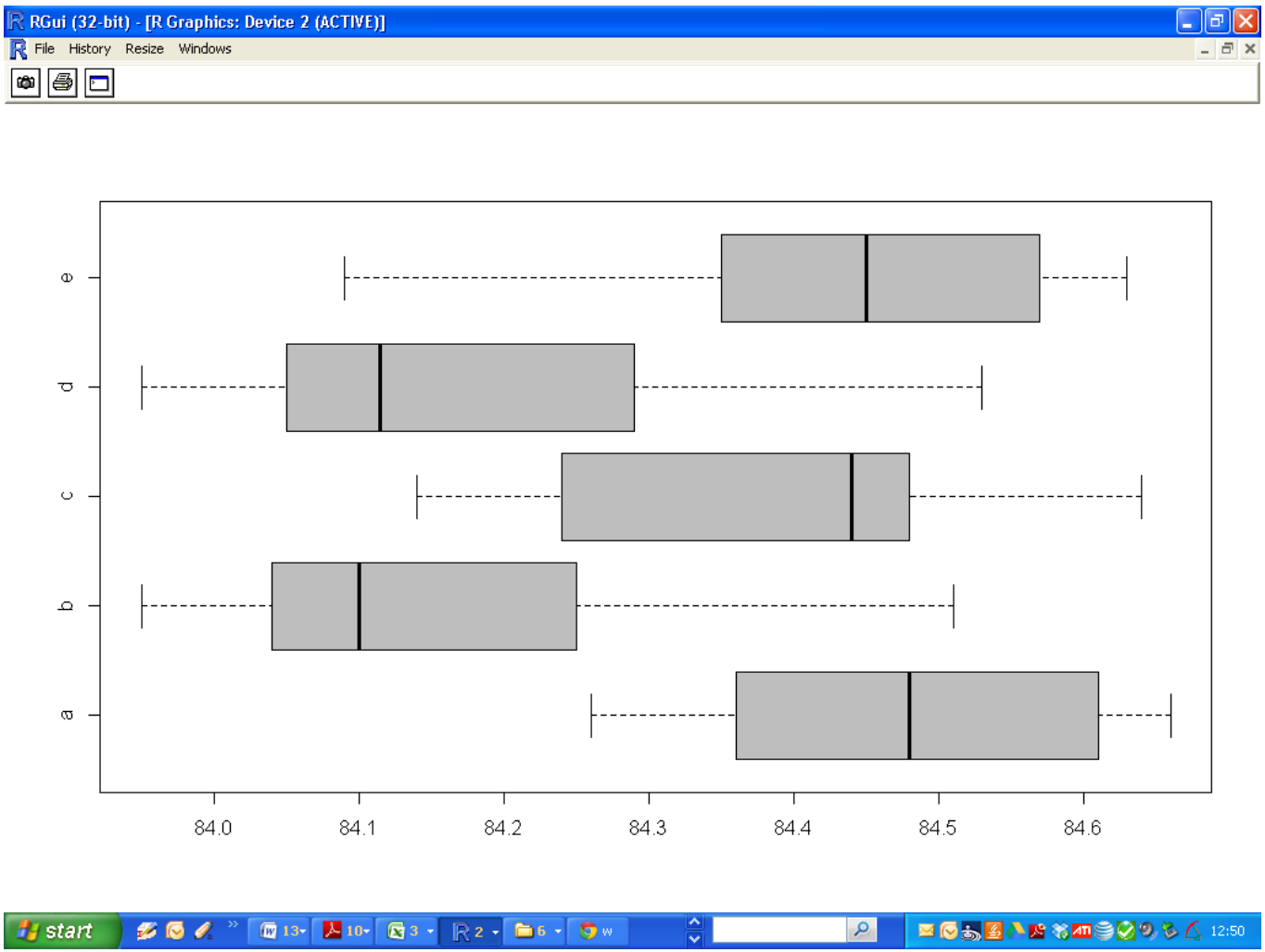
- Remark: If you do not get a sensible value for either the upper or lower fence, you can replace it with the nearest sensible value
- For example, suppose we got a negative lower fence value. It does not make sense to get a negative score in an exam.
- In this case, we could replace the value with a value of 0.
- Similarly for the upper fence: any fence value greater than 100 should be replaced with the value of 100.

Boxplots

- Boxplots are very useful in comparing the distributions of two or more groups when measured by the same variable.
- They can be used to assess how similar the centrality of the various groups. (also called “location”).
- Boxplots can use used to compare the dispersion of scores for all of these groups. (also called “scale”).

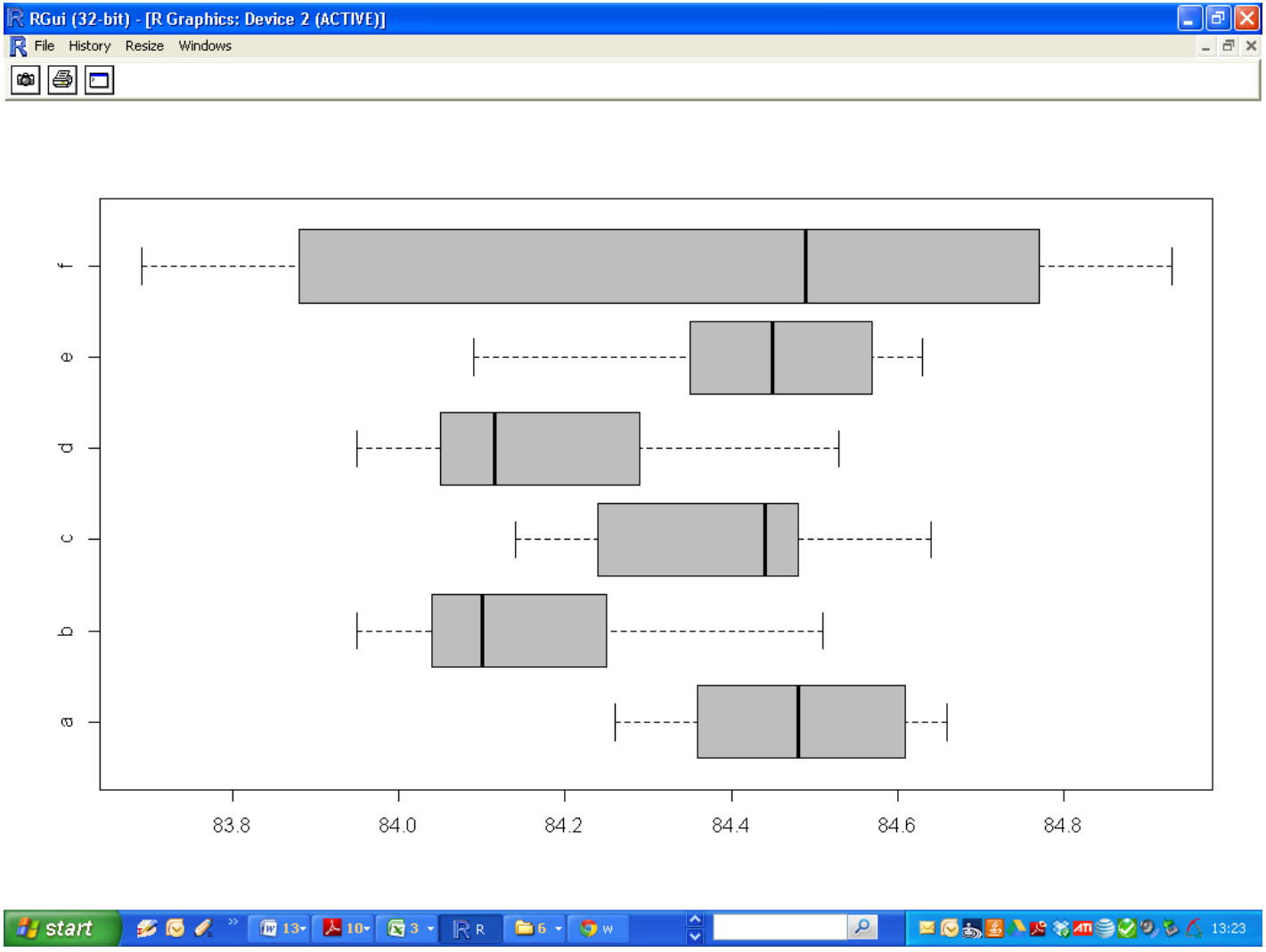
Boxplots

Groups A,C,E have similar centrality. Group B and D has similar centralities also, but different from A,C and E. The dispersion is roughly the same for each group.



Boxplots

Much higher dispersion indicated for Group F (top group).
Other groups have similar dispersion ("scale").



Review

- Be able to interpret a box-plot, particularly for indicating outliers.
- Know the procedure used to determine if a point should be considered an outlier (i.e. comparing values to Lower and Upper Hinges)
- Use boxplots to compare scale and location for different groups.