Section 1.1 - Graphical Summaries

Statistics 104

Autumn 2004

Introduction

Definitions

- **Individuals**

  The objects described by a set of data. Individuals may be people, but also include animals or things.

  Also referred to as cases or subjects (usually only used for people).

- **Variables**

  Any characteristic of an individual. Can take different values for different individuals.

- **Quantitative variable**

  A variable that takes numerical values.

  e.g. height, weight, annual income, number of winners in Saturday's lottery.

- **Categorical variable**

  A variable that places an individual into one of serval groups

  e.g. gender, education level (elementary, high school, college, post graduate)

- **Distribution**

  The distribution of a variable tells us what values it takes and how often it takes these values.

  Example: Graduate school admissions data for the University of California, Berkeley in Autumn 1973.

Major	A	B	C	D	E	F
# Applied	933	585	918	792	584	714

1 variable: Major
6 possible values for the variable (A through F). Actual names of majors can't be released for privacy reasons.

Displaying Distributions Graphically

Part of **Exploratory Data Analysis (EDA)**

Usual approach to EDA:

1. Look at each variable separately
2. Then move onto examining for relationships between the variables
3. Start with graphical summaries and then move to numerical summaries
4. Approach may depend on the questions of interest in the study

Graphs for categorical variables

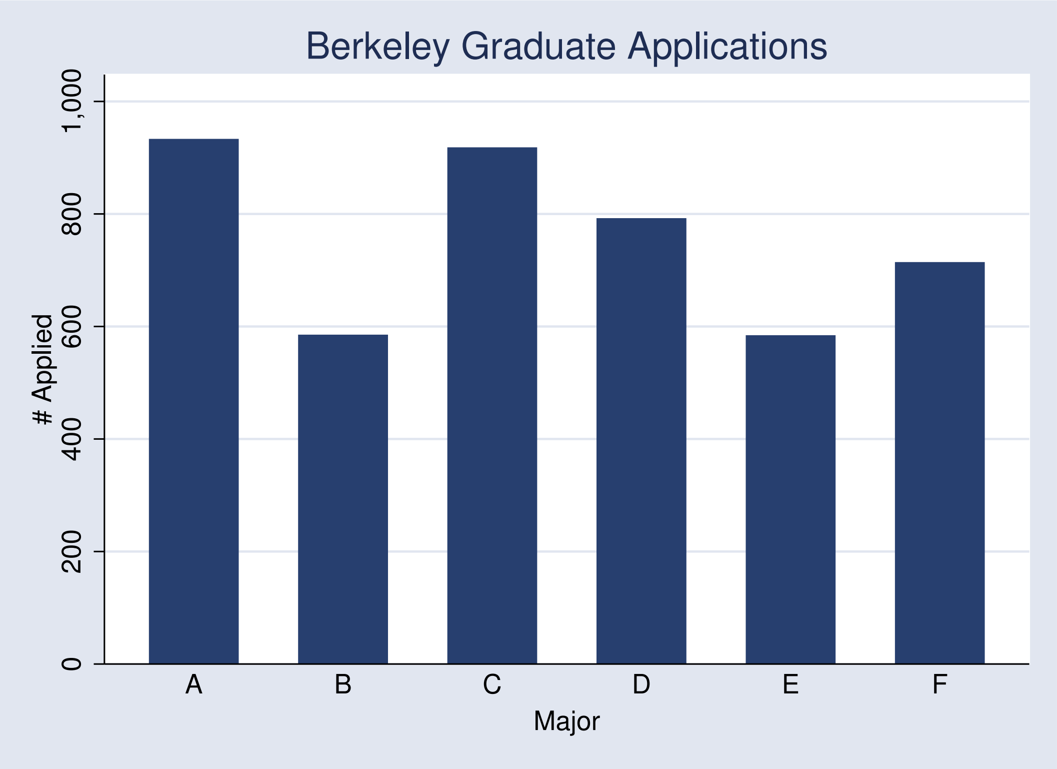
Based either on count or percent of individuals falling in each category

Major	A	B	C	D	E	F
# Applied	933	585	918	792	584	714
% Applied	20.6	12.9	20.3	17.5	12.9	15.8

1. Bar graph

   1 bar for each category

   Height of bar is count or percentage for category

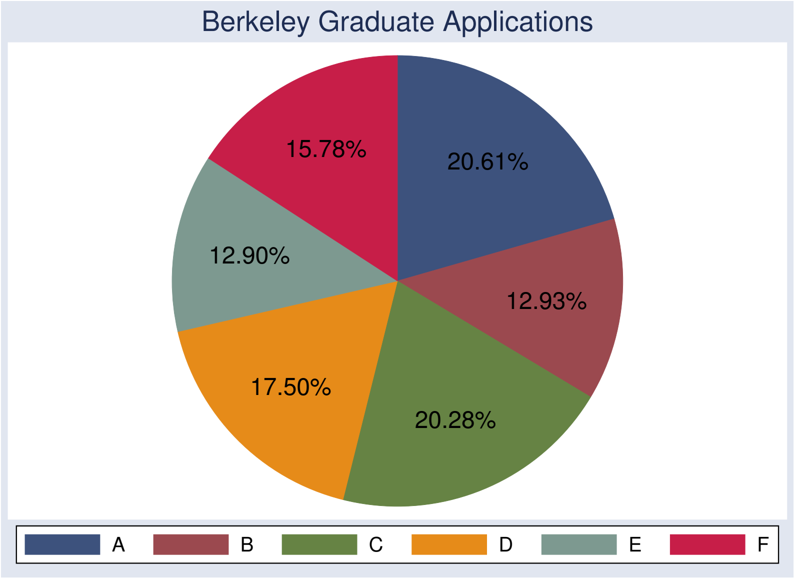

Berkeley Graduate Applications
# Applied
0
200
400
600
800
1,000
A
B
C
D
E
F
Major

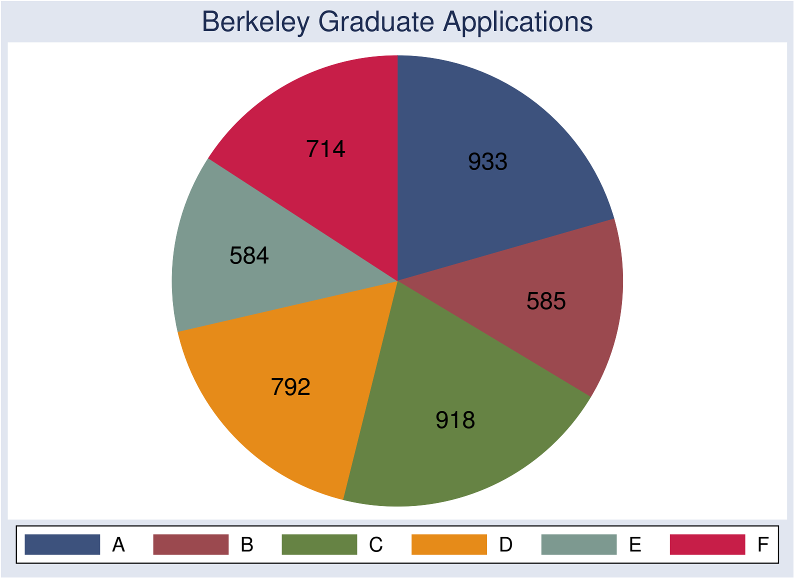
2. Pie chart

   1 segment for each category

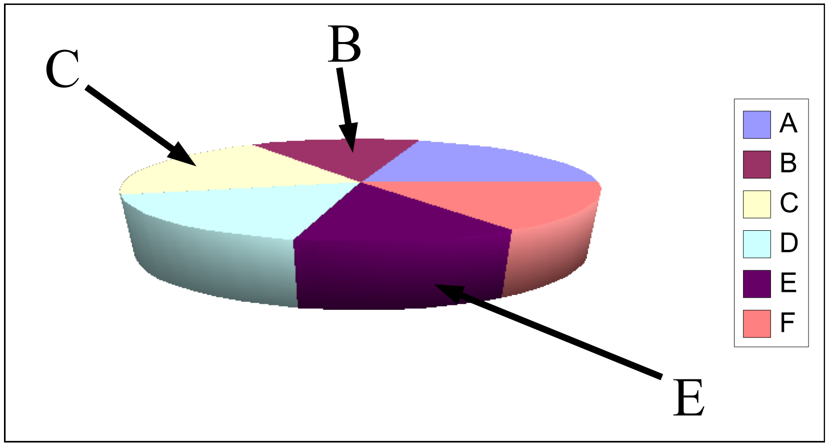
   The area of each segment is proportional to the percentage (or count) of each category

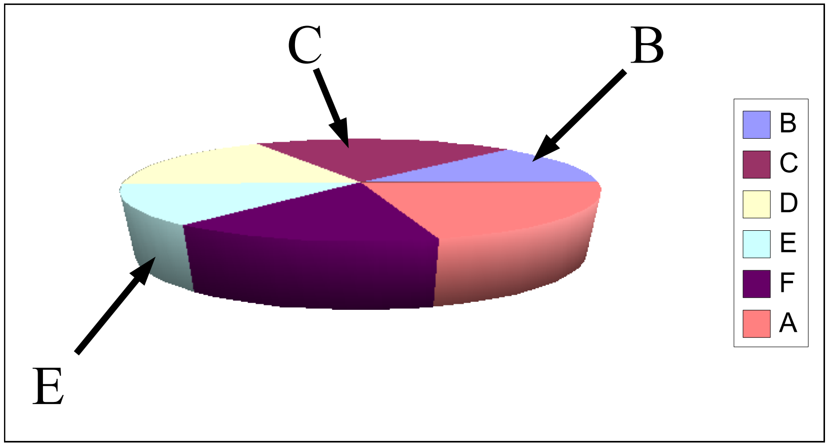
$$\text{Angle} = 360 \times \frac{\text{Percentage}}{100} = 360 \times \frac{\text{Count}}{\text{Total}}$$





Other programs, such as Excel, will give different other options for pie charts, such as the following perspective version. You shouldn't use these as they can be misleading. (Similarly, these sort of effects shouldn't be used for other types of plots either.)





The apparent size of each category depends on how the pie is rotated.

(This example was created in Open Office)

Generally bar graphs are preferred to pie charts

In particular, bar charts can be used when looking at a subset of the categories much easier than for pie charts.

Quantitative variables:

Number of options for graphical summaries of quantitative variables.

Need to consider **what** is being measured and **how** it is being measured.

Does what is being measured answer the question of interest?

Don't want to measure the size of an atom with a tape measure.

What units are the measurements in?

Feet vs inches, °F vs °C

Are the the various values for the variables comparable?

Counts vs Rates/Percentages

Example: Berkeley Admissions Data

Major	Admitted	Rejected	Applied	% Admit	% Reject
A	600	333	933	64.31	35.69
B	370	215	585	63.25	36.75
C	322	596	918	35.08	64.92
D	269	523	792	33.96	66.04
E	148	436	584	25.34	74.66
F	46	668	714	6.44	93.56
Total	1755	2771	4526	38.78	61.22

In this case, which is more informative depends on the question of interest

- Counts might be more useful if interested in tuition collected for each program
- Percentages would be more useful if interested which are the difficult programs to get into

Variation:

Data varies (see Newcomb example on pages 7 & 8)

For example, if I gave each of you a very precise thermometer (measure to 0.01°F) and asked each of you to measure the temperature of the room, there would be different answers. These differences could be possibly due to:

- Differences between the different thermometers
- Where you are in the room
- Exactly when you make the measurement
- and so on

Want to describe this variation (one part of the distribution)

Example: Chicago Civil Service Exam

Wanted to hire 15 operating engineers in 1966. Had 233 applicants. As part of the hiring procedure, these applicants had to write an exam.

Sample of 22 scores from the 233 applicants

41	49	58	75	69	53	46	32
54	35	48	91	45	60	52	67
43	56	95	27	37	62		

Want to see if we can find interesting features in this data.

(Full and reduced data sets available on the web site)

Stemplots (Stem and Leaf plots)

To make a stemplot

1. Separate each observation into a *stem* and a *leaf*.
   - Stem - leading digits of the number
   - Leaf - next digit of the number, usually the last

2. List the stems vertically in increasing order from top to bottom and draw a vertical line to the right of the stems.

3. Write each leaf in the row to the right of its stem. (Don't bother to arrange the leaves in increasing order as the book suggests.)

Stemplots are tend to be useful in describing small datasets. Easy to construct by hand.

Look at the reduced version of the Chicago Civil service data (22 observations)

Stem: 10's digit

Leaf: 1's digit

```
. stem CivilSm

Stem-and-leaf plot for CivilSm

  2* | 7
  3* | 257
  4* | 135689
  5* | 23468
  6* | 0279
  7* | 5
  8* |
  9* | 15
```

When using stem and leaf plots, we are looking for basic patterns in the data, so ordering of the leaves doesn't make much difference. Ordering can be helpful if you are trying to determine some of the summary statistics (quartiles and median) directly from the stem and leaf plot. However usually you'll want to use the computer to do that.

```
Stem-and-leaf plot for CivilSm

  2* | 7                            2* | 7
  3* | 257                          3* | 257
  4* | 135689                       4* | 196853
  5* | 23468                        5* | 83426
  6* | 0279                         6* | 9072
  7* | 5                            7* | 5
  8* |                              8* |
  9* | 15                           9* | 15
```

Notice the basic shape is the same in both plots. The left plot has the leaves ordered, whereas the right plot orders the leaves based on the order the data was listed in the earlier table.

```
Stem-and-leaf plot for CivilSm

  2* | 7
  3* | 257
  4* | 135689
  5* | 23468
  6* | 0279
  7* | 5
  8* |
  9* | 15
```

Lets assume that the were only 22 applicants for 2 positions.

There are two extreme values on the high end (outliers)

Is this evidence of a rigged exam?

- This evidence is consistent with the rigged exam theory
- However it is also consistent with having two “stars” writing the exam.

Coupling the exam scores with other evidence led to charges and convictions of exam rigging (with real data). There are 15 scores above 90, exactly the number of positions open.

```
2. | 677779
3* | 00001112233333444
3. | 5566677777779999999
4* | 012222233333333444444
4. | 5555555666666777777888888889999
5* | 0011111222223333344444
5. | 555666667777888888889999
6* | 000000111111222334
6. | 566677778899999999
7* | 1123444
7. | 55668
8* | 00001112233334444444
8. |
9* | 0001112223333
9. | 55
```

Common mistakes with stem and leaf plots

- Omitting a stem because it has no leaves

  Skipping the 80's stem might have led to missing the cheating in the exam

- Not lining leaves up in columns

  Example: Students heights (in cm)

  Data:

160	165	178	175	188	195
173	190	184	199	182	187

```
16* | 05               16* |   0  5
17* | 853              17* | 8 53
18* | 8427             18* | 8427
19* | 509              19* | 5  0  9
```

- Poor choice of stem size
  - Option 1: Stem - 100's digit, Leaf - 10's digit

```
1* | 66777899798988
```

  - Option 2: Stem - 10 & 100's digits, Leaf - 1's digit

```
16* | 05
17* | 853
18* | 8427
19* | 509
```

- Putting the leaf with the wrong stem

Back to Back Stemplots

An approach to comparing **two** groups.

It is what is says. Two stemplots are put back to back.

Stems run down the middle. Leaves for one group go to the right of the stems, leaves for the otherr group go to the left.

Example: Marina Rental Rates in Los Angeles area (Data on web site)

Rental rates (per foot) for boat storage

Marina del Rey:

$6.37 $6.60 $6.27 $6.49 $6.64 $6.82 $7.16 $6.45 $5.60
$5.95 $4.50 $6.60 $6.00 $6.82 $7.04 $5.30 $7.05 $7.05
$6.96

Long Beach Harbor:

$4.60 $4.75 $4.70 $8.75 $4.50 $5.40 $6.00 $6.00 $6.50
$6.00 $5.00 $5.00 $5.50 $4.35 $4.50 $5.20 $4.95

```
        Stems 0-4 / 5-9

              |4|3
             5|4|677559
             3|5|4002
            96|5|5
         04423|6|000
        986866|6|5
          0001|7|
              |7|
              |8|
              |8|7
Marina del Rey   Long Beach
```

```
          Stems 0-9

             5|4|3677559
           963|5|40025
   98686604423|6|0005
          0001|7|
              |8|7

Marina del Rey   Long Beach
```

The left hand version is an example of **splitting stems**. In this case, each stem is split into 2 (0-4 and 5-9). You can also split into 5 (0-1, 2-3, etc). In this example, I think splitting the stems in 2 gives a better picture.

Note: I do not know of software that will create back to back stem and leaf plots.

Histograms

- Another graphical display of data

- Useful when
  - No natural of good stem
  - Large data sets

To make a histogram

1. Split range of data into classes of equal width

2. Count number of observations in each class

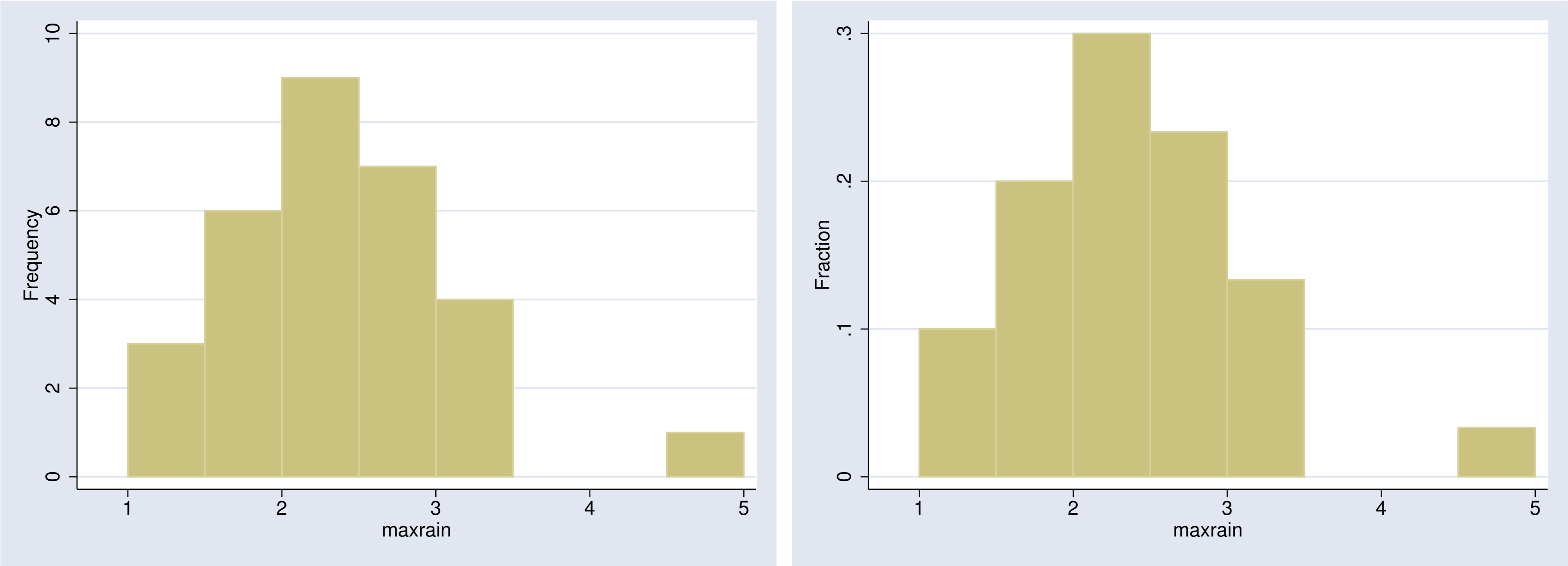
3. Draw histogram

   For each class (or bin), draw a bar. The base covers the width of the class and the height is the class frequency

Note: Instead of using the frequency as the height of the bar, can also use the relative frequency (proportion) instead.

Example: South Bend rainfall (available on web site)

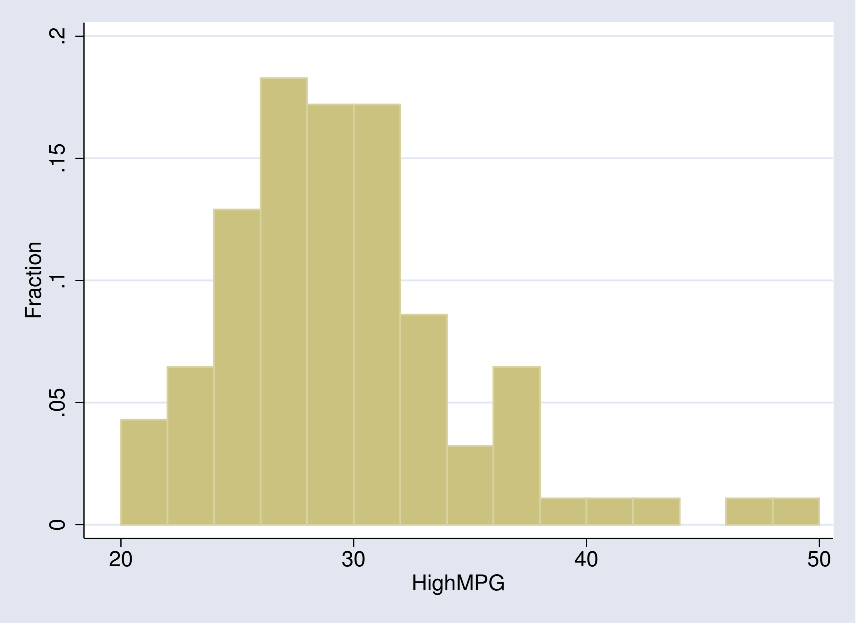
Maximum daily rainfall for 1941 to 1970 in South Bend Indiana



Example: 1993 Model Cars (available on web site)

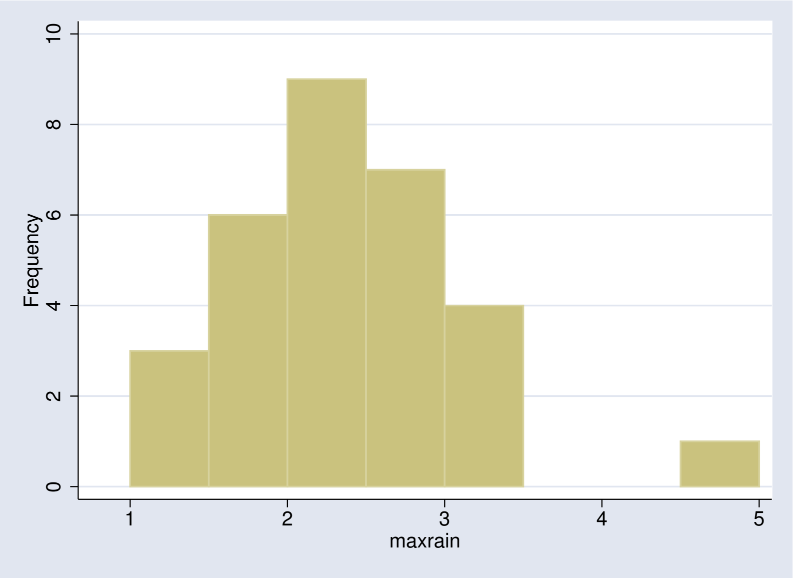
Large number of measurements on 93 different cars.

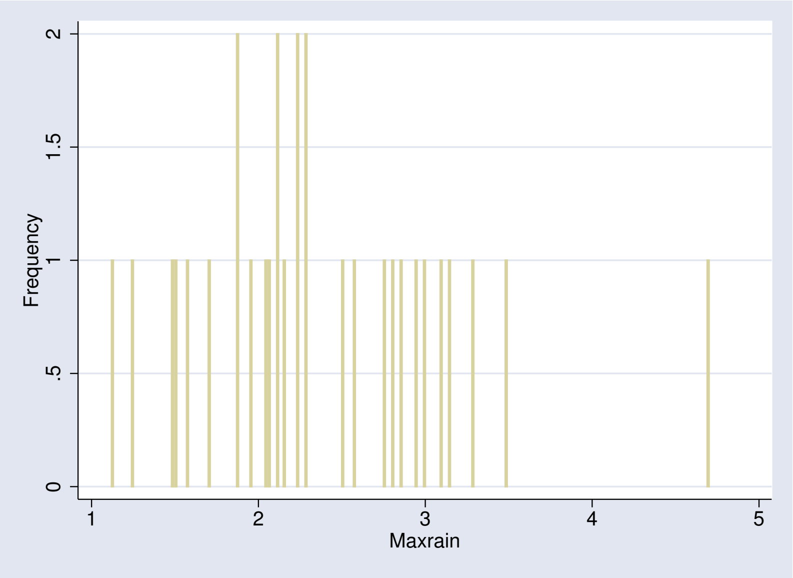
Look at EPA Highway MPG ratings

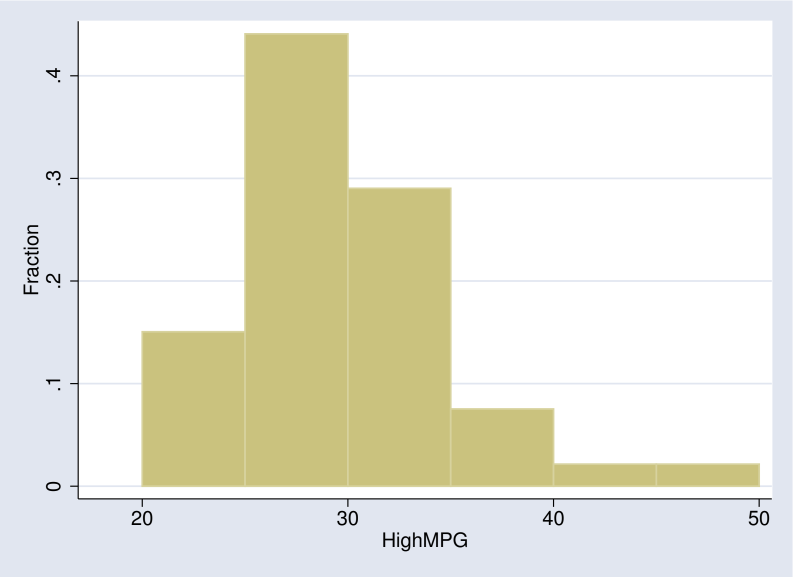


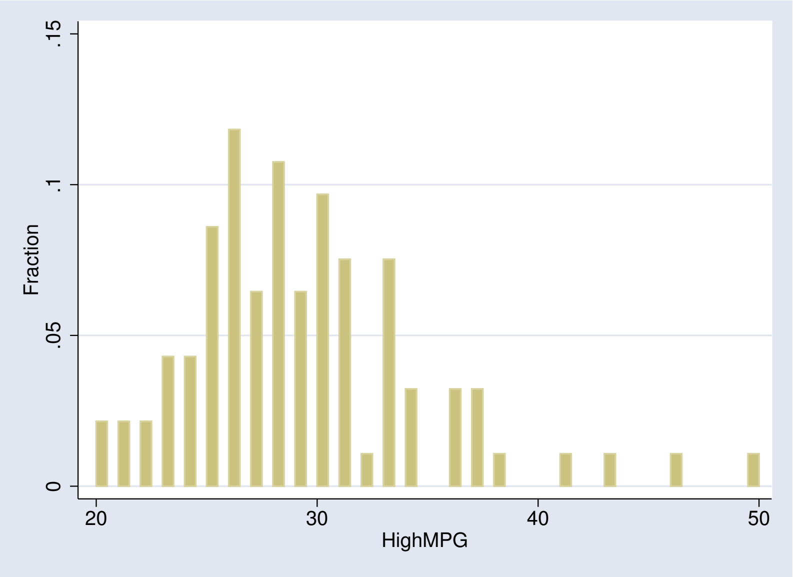
Choosing classes:

- Don't want too many or too few classes. Stat packages have defaults for the number of classes based on the number of observations.


Frequency
0
2
4
6
8
10
1
2
3
4
5
maxrain


Frequency
0
.5
1
1.5
2
1
2
3
4
5
Maxrain


Fraction
0
.1
.2
.3
.4
20
30
40
50
HighMPG


Fraction
0
.05
.1
.15
20
30
40
50
HighMPG

- Pick nice numbers for class boundaries

  e.g. 5, 10, 15, 20, ... or 1, 2, 3, 4, ...

  Most packages will do this, though Stata doesn't. However its easy to change the bins in Stata to something nice.

- Any possible observation can only go into one class

  e.g. Don't have classes 5-15 & 10-20

- Suppose you have classes 5-10 & 10-15. Where does 10 belong?
  - It up to you
  - The problem many suggest which to do
  - Stata, Minitab and Data Desk will put it in the 10-15 bin, whereas `R` will put it in the 5-10 bin. (These programs might deviate from this if you are dealing with the first or last bins).
  - If doing it by hand, just be consistent.

- Equal bin widths?

  You do not have to have to use bins of equal width. However you need to make an adjustment in the bin height to adjust for the different widths. Need to have the area of each bar equal to the proportion (fraction) of observations in that bin

  $$\text{Height} = \frac{\text{Proportion}}{\text{Width}}$$

  This can be done in Stata by choosing the density option. Note that if the bin widths are the same, switching to density just relabels the y-axis from what you would get with frequency, fraction, or percentage.

  (Note: in Stata it is not possible to have different bin widths to the best of my knowledge. Some other packages (e.g. R) do allow this.)

What to look for in stem & leaf plot or a histogram

- Center - “typical” value
- Spread - around “typical” value
- Shape - Symmetric vs skewed, unimodal vs multimodal
  - Symmetric: mirror image around center
  - Skewed: one “tail” longer than the other
    Skewed right: right tail (larger values) is longer than the left
- Modes: # of peaks

  Unimodal = one peak, Bimodal = two peaks

  Why can modes occur?

  - Mixtures of groups (e.g. Chicago civil service exam - Cheaters vs non-cheaters)

- Special points - “Outliers”

  Where can outliers come from

  – Errors - recording problem, equipment failure (delete observation)
  – Extraordinary data
  – Misunderstanding of the model

  In these last two cases, you may or may not want to keep the observation in the data set. Often you can learn things from these outlying observations. Try to learn why they are different.

  In the Chicago Civil Service exam, the outliers were found to be due to cheating. For the marina rental rates, it is not clear why there is an outlier marina. It could be due to the location, the owner ripping people off, etc.

Time series data

Data recorded over time

Examples:

- Daily closing stock prices
- Hourly average ozone
- Monthly natural gas use
- Annual number of births
- Annual Dow Jones closing price
- Hurricanes per year
- Annual sunspot activity

Consider equally spaced observations over time, though there might be missing data

Common summary: Time series plot

- x-axis: time
- y-axis: variable of interest
- Points often joined by lines, but they don't have to be.

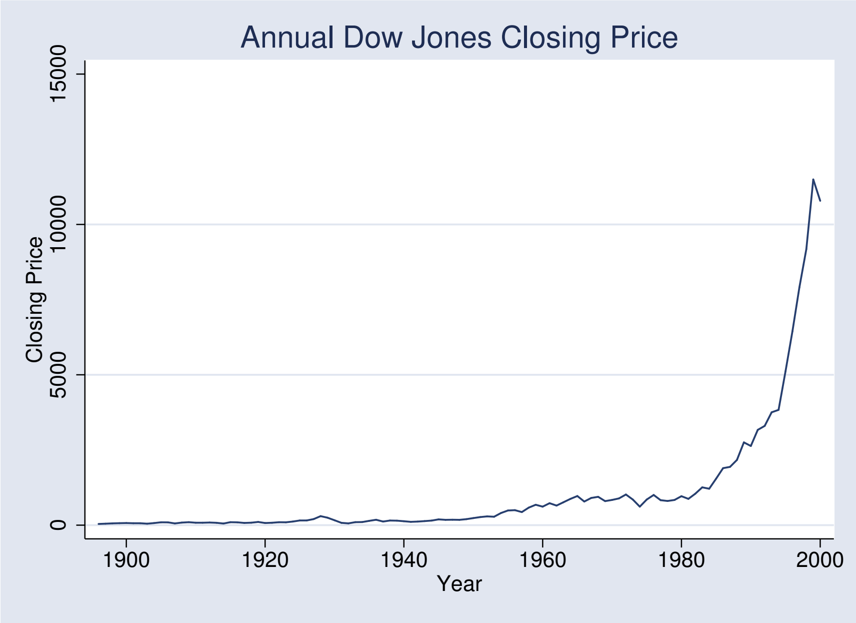
Patterns of interest:

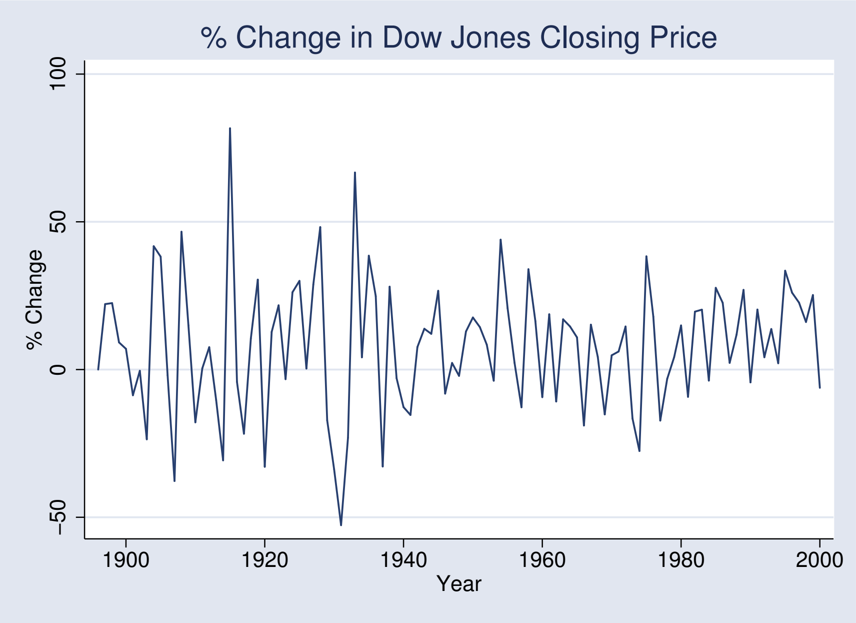
- Trend - long term, persistent rise or fall
- Seasonal variation - pattern that repeats itself at known regular intervals of time

Example: Dow Jones Industrial Average (on web site)

Year end closing values for 1896 to 2000

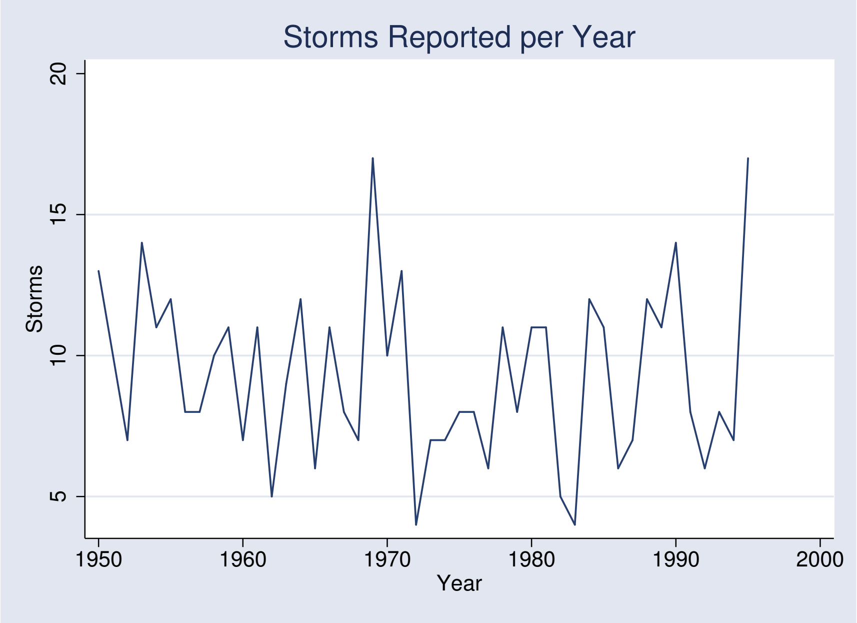
$$\%\text{Change} = 100 \times \frac{DJ_t - DJ_{t-1}}{DJ_{t-1}}$$

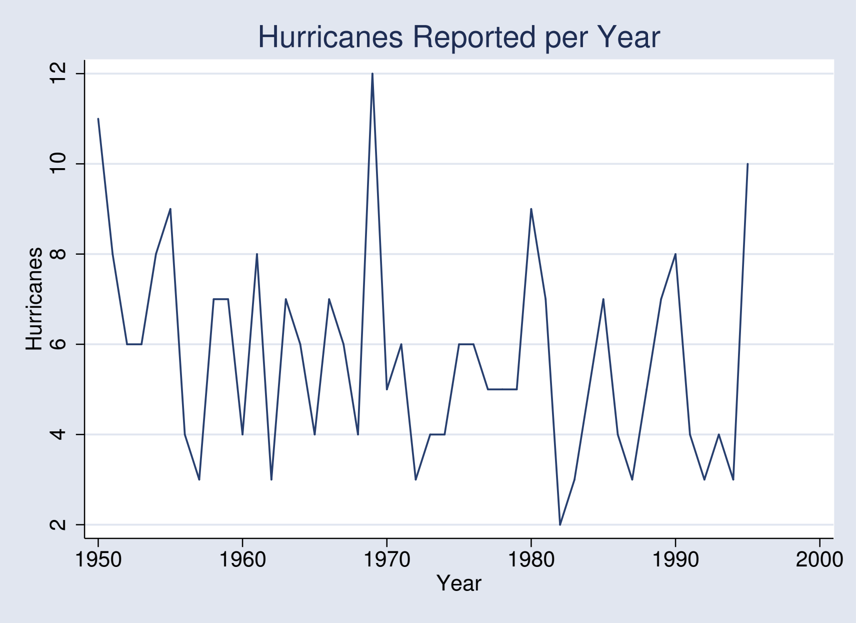




Example: Atlantic Storms (on web site)

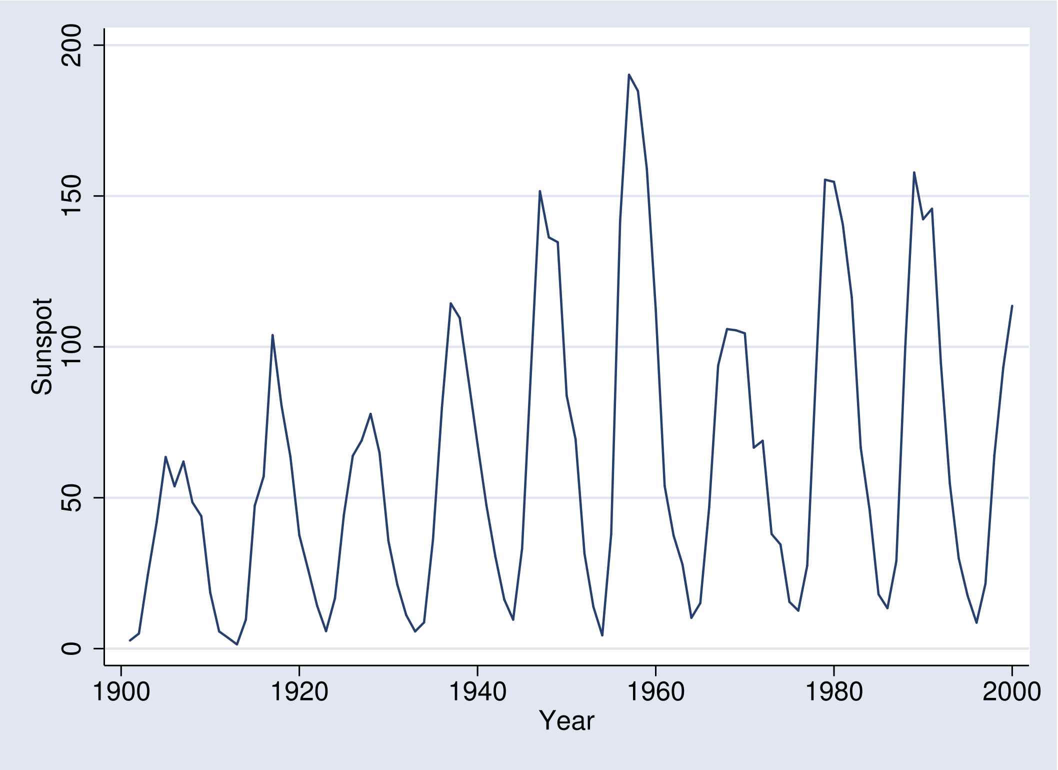
Major storms and hurricanes from 1950 to 1995





Example: Sunspots (on web site)

Sunspot activity from 1900 to 2000

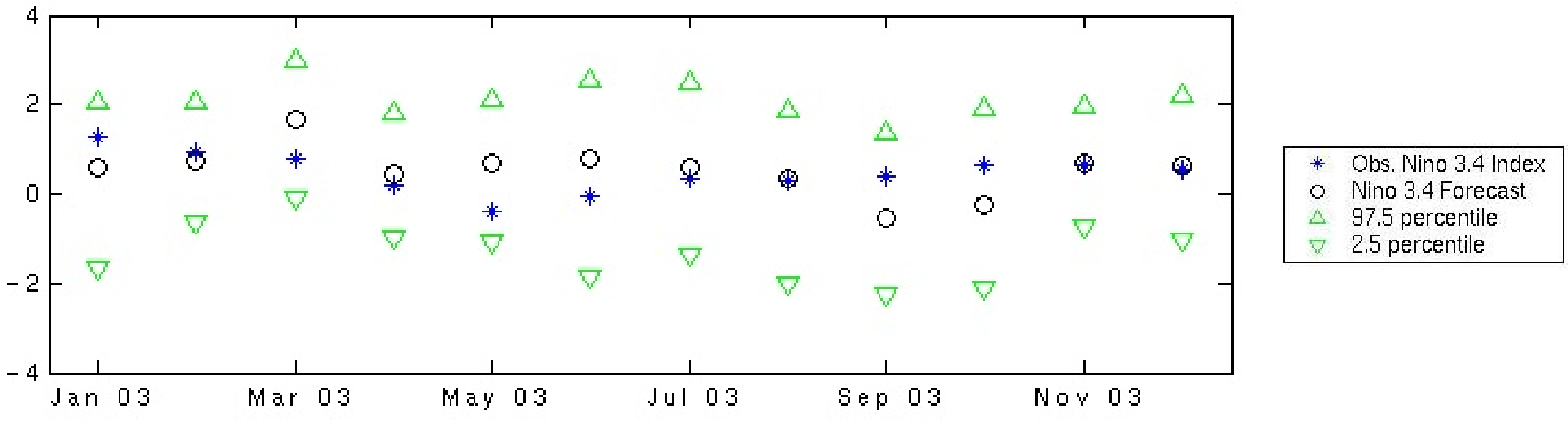


Seasonally adjusted time series

Often government reported data, such as unemployment or inflation rates are often seasonally adjusted. In these data sets, the “typical” seasonal trend is used to adjust the actual observed values to give what is reported.

For example, a reported increase in the inflation rate, means that the change in the inflation rate was larger than expected due the model describing the seasonal variation.

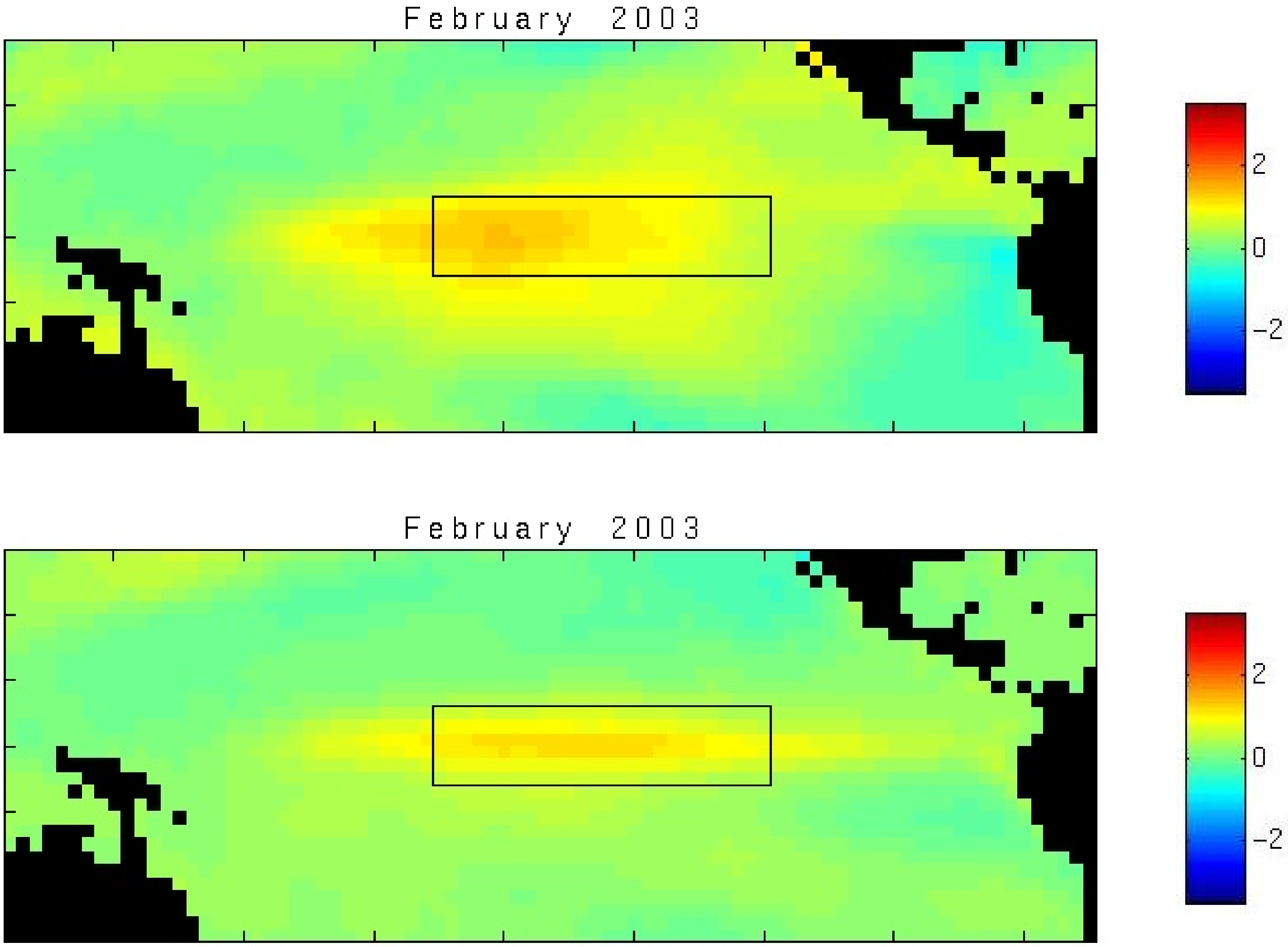
Example: Nino 3.4 values for 2003



This is an example of an adjusted time series. Reported are the actual monthly temperature for the Nino 3.4 region minus the monthly average for each month from 1971 to 2000.

The forecast Nino 3.4 values are based on data from January 1970 to 7 months before the forecast date.

<http://www.stat.ohio-state.edu/∼sses/collab_enso.php>.



February 2003 observed (top) and forecast (bottom) temperature anomaly based on January 1970 to July 2002 data.