Topic 1

Displaying Data

Quantitative Data

Outline

Histograms

- Empirical Cumulative Distribution Function
- Survival Function

Scatterplots

- Time Plots

Histogams

Histograms are a common visual representation of a quantitative variable. Histograms visual the data using rectangles to display frequencies and proportions as normalized frequencies. In making a histogram, we

- Divide the range of data into bins of equal width (usually, but not always)
- Count the number of observations in each class.
- Draw the histogram rectangles representing frequencies or percents by *area*.

Histogams

Interpret the histogram by giving

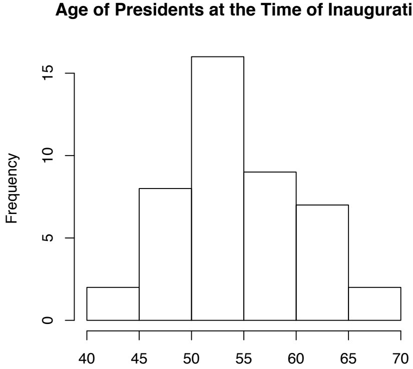
- the overall pattern
  - the center
  - the spread
  - the shape (symmetry, skewness, peaks)
- and deviations from the pattern
  - outliers
  - gaps

The direction of the skewness is the direction of the longer of the two tails (left or right) of the distribution.

Histogams

Taking the age of the presidents of the United States at the time of their inauguration and creating its histogram in R is accomplished as follows.

```
> age<-c(57,61,57,57,58,57,61,54,68,51,49,64,50,48,65,52,56,46,54,49,51,47,55,
55,54,42,51,56,55,51,54,51,60,61,43,55,56,61,52,69,64,46,54,47)
> hist(age, main = c("Age of Presidents at the Time of Inauguration"))
```



Empirical Cumulative Distribution Function

The empirical cumulative distribution function $F_n(x)$ gives, for each value x, the fraction of the data less than or equal to x. If the number of observations is n, then

$$F_n(x) = \frac{1}{n}\#(\text{observations less than or equal to } x).$$

- $F_n(x) = 0$ for any value of x less than all of the observed values.
- $F_n(x) = 1$ for any x greater than all of the observed values.
- In between, we will see steps that are multiples of the $1/n$.

Empirical Cumulative Distribution Function

In order to create a graph of the empirical cumulative distribution function,

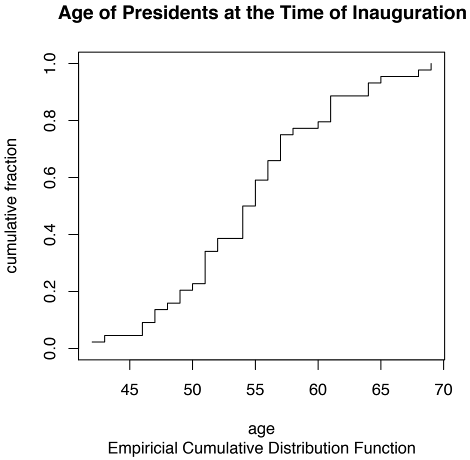
- Place the observations in order from smallest to largest by writing `sort(age)`.
- Next match these up with the integral multiples of the 1 over the number of observations using `1:length(age)/length(age)`.
- Finally, `type="s"` to give us the steps described above.

```
> plot(sort(age),1:length(age)/length(age),type="s",ylim=c(0,1),
main = c("Age of Presidents at the Time of Inauguration"),
sub=("Empiricial Cumulative Distribution Function"),
xlab=c("age"),ylab=c("cumulative fraction"))
```

Empirical Cumulative Distribution Function

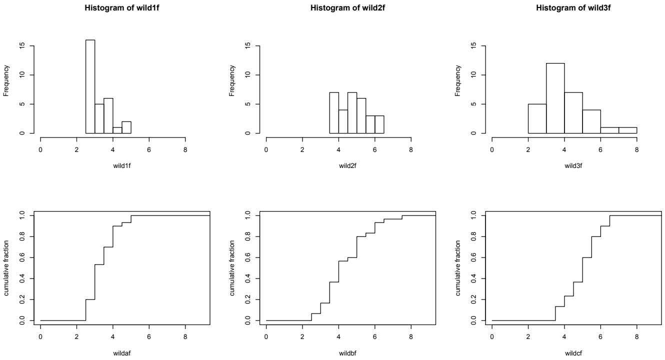
Exercise.

1. What fraction of presidents were 50 years old or younger at the time of their inauguration?
2. What fraction were older than 60?
3. What age place a president on the lower 1/3 of ages at the time of their inauguration?



Empirical Cumulative Distribution Function

Exercise.The histogram for data on the length of three bacterial strains is shown below. Lengths are given in microns. Below the histograms (but not necessarily directly below) are the corresponding empirical cumulative distribution functions.



Match the histograms to their respective empirical cumulative distribution functions.

Survival Function

In looking at life span data, the natural question is "What fraction of the individuals have survived a given length of time?" The survival function, $S_n(x)$, gives, for each value x, the fraction of the data greater than or equal to x. If the number of observations is n, then

$$\begin{aligned} S_n(x) &= \frac{1}{n}\#(\text{observations greater than } x) \\ &= \frac{1}{n}(n - \#(\text{observations less than or equal to } x)) \\ &= 1 - \frac{1}{n}\#(\text{observations less than or equal to } x) \\ &= 1 - F_n(x) \end{aligned}$$

Scatterplots

Scatterplots show the relationship for pairs of observations.

- The values of the first variable $x_1, x_2, \ldots, x_n$ are often assumed known, for example, when they are set by an experimenter.
  - Called explanatory, predictor, or discriptor variable
  - Displayed on the horizontal axis.
- The values of $y_1, y_2 \ldots, y_n$ are taken from observations with input $x_1, x_2, \ldots, x_n$.
  - Called the response variable
  - Displayed on the vertical axis.

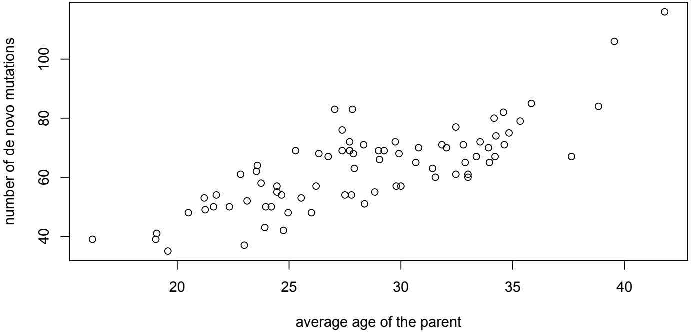
Scatterplots

In describing a scatterplot, take into consideration

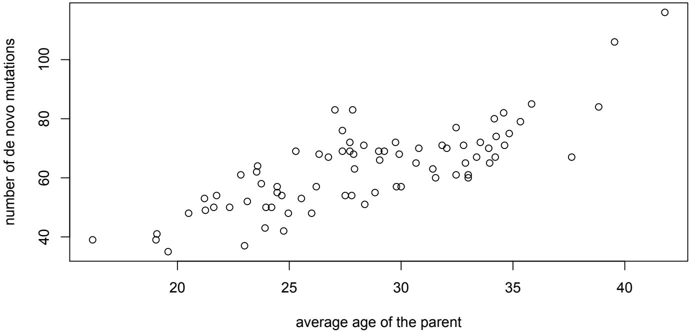
- the form, for example,
  - linear
  - curved relationships
  - clusters
- the direction,
  - a positive or negative association
- and the strength of the aspects of the scatterplot.

Scatterplots

Example. Genetic evolution is based on mutation. Consequently, one fundamental question in evolutionary biology is the rate of *de novo* mutations. To investigate this question in humans, Kong et al, sequenced the entire genomes of 78 Icelandic trios and recorded the age of the parents and the number of *de novo* mutations in the offspring.



Scatterplots



The plot shows a moderate positive linear association, children of older parent have, on average, more mutations. The number of mutations range from ~ 40 for children of younger parents to ~ 100 for children of older parents.

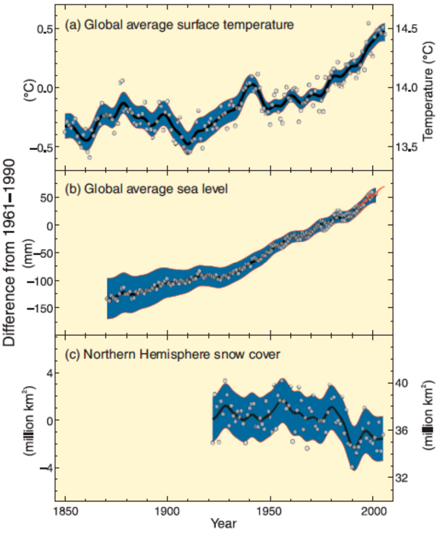
Time Plots

Time plots have time as the explanatory variable and some measurement over time as the response.

The Intergovernmental Panel on Climate Change (IPCC) is a scientific body tasked with evaluating the risk of climate change caused by human activity. The IPCC does not perform original research but rather synthesizes research and prepare a report. In addition, the IPCC prepares a summary report. The Fourth Assessment Report was completed in early 2007. The fifth is scheduled for release in 2014.

Exercise. On the next slide is the first graph from the 2007 *Climate Change Synthesis Report: Summary for Policymakers*. Describe the graphs and give a short explanation why these scientists would give these time plots such prominence.

Time Plots



Notes. Snow cover is measured in the Northern Hemisphere for March and April. Differences are relative to the averages for the period 1961-1990. Smoothed curves are based on decadal average. Shaded area are uncertainty intervals.