Looking at Data

Quantitative variables:

Maximum wind speeds of 2011 tropical storms
65, 70, 70, 50, 50, 45, 65, 65, 120, 35, 45, 140,
45, 60, 80, 75, 140, 90, 115, 65

Stem and Leaf Plot

leaves are sorted

can see: distribution
max/min
clustering

Histogram

Frequency

max wind speed

Frequency Distribution - summarizes the histogram in a table

<table>
<thead>
<tr>
<th>Speed</th>
<th>Frequency</th>
<th>Relative Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-59</td>
<td>4</td>
<td>20%</td>
</tr>
<tr>
<td>50-69</td>
<td>7</td>
<td>35%</td>
</tr>
<tr>
<td>70-89</td>
<td>4</td>
<td>20%</td>
</tr>
<tr>
<td>90-109</td>
<td>1</td>
<td>5%</td>
</tr>
<tr>
<td>110-129</td>
<td>2</td>
<td>10%</td>
</tr>
<tr>
<td>130-149</td>
<td>2</td>
<td>10%</td>
</tr>
</tbody>
</table>

Relative Frequency - table

Counts divided by total counts

Useful when comparing
samples or populations
of different sizes

Qualitative Data

Pareto Chart - sorted bar chart

histogram for categorical data

Pie Chart - show relative frequencies

White: 73.3%
Latino: 8.1%
Asian: 14.7%
Native Am: 1.5%

UCSC regular faculty ethnicity 2011

Frequency

N=546

Whites: \( \frac{400}{546} = 0.733 \)
Descriptive Statistics

Measures of Center

What is a “typical” value of a dataset?

\[ \text{Mean} = \text{average} \quad \bar{x} \quad \text{sample mean} \quad \mu \quad \text{population mean} \]

\[ \bar{x} = \frac{\sum x_i}{n} = \frac{1}{5} \left( 65 + 105 + 50 + 175 + 40 \right) = \frac{435}{5} = 87 \]

The mean is often used because it has good theoretical and practical behavior.

\[ \bar{x} \] is usually the best possible estimator of \( \mu \)

The mean can be unreliable if there are outliers (extremely unusual observations).

When there are extreme values, the \underline{median} is a more stable measure of center. The median is the middle value.

Ex: Aug 2005 storms: median is 65

If there is an even number of observations, then the median is the average of the two middle values.

Ex: 2011 storms: \( \frac{65 + 65}{2} = 65 \)

Aug 2004: 49, 45, 65, 75, 120, 135, 145 median \( \frac{70 + 105}{2} = 87.5 \)

When the data are symmetric and without outliers, the mean and median will be very similar.

(Note: 1 datum, 2 data)

The \underline{mode} is the most common value.

Ex: 2011 storms, mode = 65

Aug 2005 storms, no mode

If only one value is most common \( \rightarrow \) unimodal

Two values equally most common \( \rightarrow \) bimodal

More than two \( \rightarrow \) multimodal

Skewness - If the distribution of the data is the same on both sides of the mean, it is called symmetric. Otherwise it is skewed.

\[ \begin{align*}
\text{Income} & \quad \text{Years at USC} \\
\hline
\text{Exam Scores} & \end{align*} \]