Descriptive Statistics

Statistical analysis learns from data.
Consider the following dataset obtained after testing 15 beams at the lab:

<table>
<thead>
<tr>
<th>Load in lb.</th>
<th>First crack load</th>
<th>Failure load</th>
</tr>
</thead>
<tbody>
<tr>
<td>10350</td>
<td>10350</td>
<td>10350</td>
</tr>
<tr>
<td>8450</td>
<td>9300</td>
<td>9600</td>
</tr>
<tr>
<td>7200</td>
<td>9600</td>
<td>10300</td>
</tr>
<tr>
<td>5100</td>
<td>10300</td>
<td>9400</td>
</tr>
<tr>
<td>6500</td>
<td>9400</td>
<td>10600</td>
</tr>
<tr>
<td>10600</td>
<td>10600</td>
<td>6000</td>
</tr>
<tr>
<td>6000</td>
<td>10100</td>
<td>9900</td>
</tr>
<tr>
<td>6000</td>
<td>9900</td>
<td>5100</td>
</tr>
<tr>
<td>10600</td>
<td>10100</td>
<td>9500</td>
</tr>
<tr>
<td>6500</td>
<td>10200</td>
<td>9300</td>
</tr>
<tr>
<td>9300</td>
<td>9500</td>
<td>6500</td>
</tr>
<tr>
<td>6000</td>
<td>9500</td>
<td>6000</td>
</tr>
<tr>
<td>5800</td>
<td>9500</td>
<td>6000</td>
</tr>
<tr>
<td>6500</td>
<td>10200</td>
<td>6000</td>
</tr>
<tr>
<td>6000</td>
<td>9550</td>
<td>6500</td>
</tr>
<tr>
<td>5800</td>
<td>10500</td>
<td>6500</td>
</tr>
<tr>
<td>6000</td>
<td>9550</td>
<td>9300</td>
</tr>
<tr>
<td>10500</td>
<td>9550</td>
<td>6500</td>
</tr>
<tr>
<td>6500</td>
<td>10200</td>
<td>9300</td>
</tr>
<tr>
<td>6500</td>
<td>9550</td>
<td>6000</td>
</tr>
<tr>
<td>5800</td>
<td>10500</td>
<td>6000</td>
</tr>
</tbody>
</table>

Numerical Summaries

$n$ observed values are $x_1, x_2, \ldots, x_n$.

- The sample mean
  \[
  \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
  \]
  - The sample mean of the first crack load is $\frac{1}{15} \times 109800 = 7320$ lb.
  - The sample mean of the failure load is $\frac{1}{15} \times 148350 = 9890$ lb.

- The sample median
  Order the observed values $x_i$. If $n$ is odd then the median is $(n + 1)/2$th value. If $n$ is even the median is the average of values at $n/2$ and $n/2 + 1$th places.
- The sample median of the first crack load is 6500 lb.
- The sample median of the failure load is 9900 lb.

- The sample mode
  most frequently occurring value(s)
  - The sample mode of the first crack load is 6000 lb.
  - The sample modal values of the failure load are 9300, 9550, 10200 lb.

- The sample variance
  \[ s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

- Unbiased estimator of variance
  \[ s^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

  Note: \[ \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i)^2 - n \bar{x}^2 \]

- The sample standard deviation
  \[ s = \sqrt{\frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \]

  - The sample variance of the failure load is 193067 lb^2.
  - The sample standard deviation of the failure load is 439.4 lb.

- The sample coefficient of variation (COV)
  \[ v = \frac{s}{\bar{x}} \]

  - The sample coefficient of variation (COV) of failure load is 0.0444.

**Data observed in pairs**

Two sets of data \( \{x_i\}_{i=1}^{n} \) and \( \{y_i\}_{i=1}^{n} \).

- The sample covariance
  \[ s_{XY} = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \]

- The sample correlation coefficient
  \[ r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{1}{n - 1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_X} \right) \left( \frac{y_i - \bar{y}}{s_Y} \right) \]

  \[ -1 \leq r_{XY} \leq 1 \]

  - The sample correlation coefficient between first crack and failure loads is –0.0229.

Subhayan De
Sample Percentile

Sample 100p percentile:
  The data value such that 100p% of the data are less than or equal to it.
25 percentile = first quantile
50 percentile = second quantile
75 percentile = third quantile

Chebyshev’s Inequality

Data set: \(x_1, x_2, \ldots, x_n\)
  Sample mean \(\bar{x}\)
  Sample standard deviation \(s > 0\)
Define: \(S_k = \{i, 1 \leq i \leq n : |x_i - \bar{x}| < ks\}\)
\(N(S_k) = \text{Number of elements in the set } S_k\) (i.e., No. of \(i\) such that \(|x_i - \bar{X}| < ks\))
For \(k \geq 1\)

\[
\frac{N(S_k)}{n} \geq 1 - \frac{n - 1}{nk^2} > 1 - \frac{1}{k^2}
\]

One sided version: \(N(k) = \text{No. of } i\) such that \(x_i - \bar{x} \geq ks\)
Then for \(k \geq 1\):

\[
\frac{N(k)}{n} \leq \frac{1}{1 + k^2}
\]

Graphical Displays

- Histograms
- Cumulative frequency plot
- Box plots

Figure 1: Box plot.