Days missed data

Numerical summaries
Mean: value not useful - overly influenced by one large value
Range: a little useful
- outlier

Median, trimmed mean: somewhat useful

50\% "Trimmed mean" - computed by removing the top and bottom 5\% of the data and then calculating the mean of what remains

Bar chart tells the story.

Distribution is right-skewed.

Example 3 - 20 people, number of issues of a monthly magazine read in the past year

Distribution bimodal (U-shaped)

Bar chart tells story.

Recall: distribution function (cdf)
for a r.v. X is
\[ F(x) = P(X \leq x) \]

For data \( x_1, x_2, \ldots, x_n \)
the cumulative frequency is number of observations \( \leq x \)
the relative cumulative frequency is cumulative frequency
\[
\frac{\text{cumulative frequency}}{n}
\]

The empirical cdf (cdfs) is
\[ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x) \]
(proportion of observations \( \leq x \))
where \( I(x_i \leq x) \) = 1 if \( x_i \leq x \)
0 otherwise.

Ordered data: \( x_1, x_2, \ldots, x_n \)

\[ \xi \] smallest data point \[ \eta \] largest data point

Mean & other numerical summaries
not so useful
\( \bar{x} \) or 25\% regular readers

For right-skewed distributions
- sqrt / log or inverse transformation (can be done)

Can result in an approximately symmetric distribution

Stem-and-leaf plot and histograms are approximations to density functions/probability mass functions for r.v.

To make analogous to density function, rather than counts plot as relative frequency
(count / n) on vertical axis

\[ \text{hist}(\text{data}, \text{freq} = F) \]

Approximation to distribution function

Recall: distribution function (cdf)
for a r.v. X is
\[ F(x) = P(X \leq x) \]

For data \( x_1, x_2, \ldots, x_n \)
the cumulative frequency is number of observations \( \leq x \)
the relative cumulative frequency is cumulative frequency
\[
\frac{\text{cumulative frequency}}{n}
\]

The empirical cdf (cdfs) is
\[ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x) \]
(proportion of observations \( \leq x \))
where \( I(x_i \leq x) \) = 1 if \( x_i \leq x \)
0 otherwise.

Ordered data: \( x_1, x_2, \ldots, x_n \)

\[ \xi \] smallest data point \[ \eta \] largest data point

Mean & other numerical summaries
not so useful
\( \bar{x} \) or 25\% regular readers

For right-skewed distributions
- sqrt / log or inverse transformation (can be done)

Can result in an approximately symmetric distribution

Stem-and-leaf plot and histograms are approximations to density functions/probability mass functions for r.v.

To make analogous to density function, rather than counts plot as relative frequency
(count / n) on vertical axis

\[ \text{hist}(\text{data}, \text{freq} = F) \]

Approximation to distribution function

Example

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Rel. Frequency</th>
<th>Cum. Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-14</td>
<td>4</td>
<td>.2</td>
<td>4</td>
</tr>
<tr>
<td>15-19</td>
<td>8</td>
<td>.4</td>
<td>12</td>
</tr>
<tr>
<td>20-24</td>
<td>5</td>
<td>.25</td>
<td>17</td>
</tr>
<tr>
<td>25-29</td>
<td>2</td>
<td>.1</td>
<td>19</td>
</tr>
<tr>
<td>30-34</td>
<td>1</td>
<td>.05</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Eligible
- approx. to cdf from histogram
- plot: horizontal axis - upper bound of each category
  - vertical axis - (relative) cumulative frequency
    - connected by piecewise straight lines

Example
Useful for:
- approx. median or any other percentile
- What % of data is \leq 22?