1 Intro: What is Statistics?

It is easy to lie with statistics. It is hard to tell the truth without it.

Andrejs Dunkels
What is Statistics?

Nature is full of Variability
How to make sense of variable data?
Use mathematical theory of randomness: Probability.

Statistics = Data Analysis based on Probabilistic Models

Descriptive Statistics

Descriptive Statistics is the first look at the data.

Statistics Software R

http://www.r-project.org

2 Data Visualization

Data Example
Data from a biology diploma thesis, 2001, Forschungsinstitut Senckenberg, Frankfurt am Main

Crustacea section

Advisor: Prof. Dr. Michael Türkay

Charybdis acutidens TÜRKay 1985

The Squat Lobster

*Galathea intermedia*


Helgoländer Tiefe Rinne, North Sea

Carapace Lengths (mm): Females, not egg-carrying ($n = 215$)

<table>
<thead>
<tr>
<th>Length (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.9</td>
</tr>
<tr>
<td>3.0</td>
</tr>
<tr>
<td>2.8</td>
</tr>
<tr>
<td>2.7</td>
</tr>
<tr>
<td>2.6</td>
</tr>
<tr>
<td>2.6</td>
</tr>
</tbody>
</table>
2.1 Histograms und Density Polygons

How many have Carapace Length between 2.0 and 2.2?

6. Sept. '88, n=215

Female Galathea, not egg-carrying, caught 6. Sept. '88, n=215

Female Galathea, not egg-carrying, caught 6. Sept. '88, n=215

How many have Carapace Length between 2.0 and 2.2?

Female Galathea, not carrying eggs, caught 6. Sept. '88, n=215
Two Months Later (3. Nov '88)

Comparing the two Distributions

Problem: different sample sizes 6.9.1988 : \( n = 215 \) 3.11.1988 : \( n = 57 \) Idea: scale y-axis such that each distribution has total area 1.
Female Crabs, not egg-carrying, caught 6. Sept. ’88, n=215

Density

Female Crabs, not egg-carrying, caught 6. Sept. ’88, n=215

Carapace Length [mm]

Density

Which proportion had a length between 2.8 and 3.0 mm?

\[(3.0 - 2.8) \times 0.5 = 0.1\]
How to compare the two distributions?

My Advice

If you are a commercial artist:

* Impress everybody with cool 3D graphics!

If you are a scientist:

* Visualize your data in clear and simple 2D plots.

(As long as you print on 2D paper and project your slides on 2D screens)

Simple and Clear: Density Polygons
Convenient to show two or more Density Polygons in one plot

Biological Interpretation: What may be the reason for this shift?
2.2 Stripcharts and Boxplots

Simplify to understand

Histograms and density polygons allow a comprehensive view on the data.

*Sometimes too comprehensive.*
Comparison of four groups

The Boxplot

Boxplot, simple type
2.3 Example: Darwin Finches

Charles Robert Darwin (1809-1882)
Darwin Finches

http://darwin-online.org.uk/graphics/Zoology_Illustrations.html

Darwin’s collection of Finches

References


Wing Sizes of Darwin’s Finches
Beak Sizes of Darwin’s Finches

2.4 Conclusions

Conclusions
• Histograms give detailed information.
• Density Polygons allow multiple comparisons.
• Boxplots can simplify large datasets.
• Stripcharts more appropriate for small datasets.
• Sophisticated graphics with 3D or semi-transparent colors do not always improve clarity.

3 Summarizing Data Numerically

Idea

It is often possible to summarize essential information about a sample numerically.

   e.g.:

• How large? Location Parameters
• How variable? Dispersion Parameters

Already known from Boxplots

Location (How large?)

\( Median \)

Dispersion (How variable?)

\( Inter \ quartile \ range \ (Q_3 - Q_1) \)

3.1 Median and other Quartiles

The median is the 50% quantile of the data.
i.e.: half of the data are smaller or equal to the median, the other half are larger or equal.
The Quartiles

*The first Quartile, $Q_1$: A quarter of the observations are smaller than or equal to $Q_1$. Three quarters are larger or equal.*

i.e. $Q_1$ is the 25%-Quantile

*The third Quartile, $Q_3$: Tree quarters of the observations are smaller than or equal to $Q_3$. One quarter are larger or equal.*

i.e. $Q_3$ is the 75%-Quantile

### 3.2 Mean, Standard Deviation and Variance

Most frequently used

**Location Parameter:** *The Mean* $\overline{x}$

**Dispersion Parameter:** *The Standard Deviation* $s$

**NOTATION:**

Given data named $x_1, x_2, x_3, \ldots, x_n$

it is common to write $\overline{x}$ for the mean.

**DEFINITION:**

\[
\text{Mean} = \frac{\text{Sum of observed values}}{\text{Number of Observations}}
\]

\[
\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

The formula for the mean of $x_1, x_2, \ldots, x_n$:

\[
\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}
\]

\[
\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]
Example for computing the mean

\[ x_1 = 3, \ x_2 = 0, \ x_3 = 2, \ x_4 = 3, \ x_5 = 1 \]
\[ \bar{x} = \text{Sum/Number} \]
\[ \bar{x} = (3 + 0 + 2 + 3 + 1)/5 \]
\[ \bar{x} = 9/5 \]
\[ \bar{x} = 1.8 \]

Geometric Interpretation of the Mean
Center of Gravity

\textbf{Mean = Center of Gravity}

Where is the center of gravity?

\[ \odot \odot \odot \odot \odot \]

\[ 0 \ 1 \ 2 \ 3 \]

\[ x \]

\[ m = 1.5 \ ? \]
\[ m = 2 \ ? \]
\[ m = 1.8 \ ? \]
Conjecture: Ratio higher for sexually mature female crabs.

Example:

3.11.88
The Standard Deviation

How far do typical observations deviate from the mean?
Die Standard Deviation $\sigma$ ("sigma") ist a slightly weired weighted mean of the deviations:

$$
\sigma = \sqrt{\text{Sum}(\text{Deviations}^2)/n}
$$

The formula for the Standard Deviation of $x_1, x_2, \ldots, x_n$:

$$
\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}
$$

$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is the Variance.

Rule of Thumb for the Standard Deviation

In more or less bell-shaped (i.e. single peak, symmetric) distributions:
ca. 2/3 are located between $\bar{x} - \sigma$ und $\bar{x} + \sigma$.

**Standard Deviation of Carapace lengths from 6.9.88**

In this case 72% are between $\bar{x} - \sigma$ and $\bar{x} + \sigma$

**Variance of Carapace lengths from 6.9.88**

All Carace Lengths in North Sea: $\mathcal{X} = (X_1, X_2, \ldots, X_N)$.

Carapace Length in our Sample: $\mathcal{S} = (S_1, S_2, \ldots, S_{n=215})$

Sample Variance:

$$\sigma^2_{\mathcal{S}} = \frac{1}{n} \sum_{i=1}^{215} (S_i - \bar{S})^2 \approx 0.0768$$

Can we use 0.0768 as estimation for $\sigma^2_{\mathcal{X}}$, the variance in the whole population? Yes, we can! However, $\sigma^2_{\mathcal{S}}$ is on average by a factor of $\frac{n-1}{n}$ ($= 214/215 \approx 0.995$) smaller than $\sigma^2_{\mathcal{X}}$. 22
Variances

Variance in the Population: $\sigma^2_X = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2$

Sample Variance: $\sigma^2_S = \frac{1}{n} \sum_{i=1}^{n} (S_i - \bar{S})^2$

(Corrected) Sample Variance:

$$s^2 = \frac{n}{n-1} \sigma^2_S$$
$$= \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} (S_i - \bar{S})^2$$
$$= \frac{1}{n-1} \cdot \sum_{i=1}^{n} (S_i - \bar{S})^2$$

Usually, “Standard Deviation (SD) of $S$” refers to the corrected $s$.

Example: Computing SD

Given Data $\bar{x} = \frac{10}{5} = 2$  Sum

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>3</th>
<th>0</th>
<th>5</th>
<th>1</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x - \bar{x}$</td>
<td>-1</td>
<td>1</td>
<td>-2</td>
<td>3</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>$(x - \bar{x})^2$</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>1</td>
<td>16</td>
</tr>
</tbody>
</table>

$$s^2 = \frac{\text{Sum}((x - \bar{x})^2)}{(n - 1)}$$
$$= 16/(5 - 1) = 4$$

$$s = 2$$