Fun with Data

Statistics is the science of data. The ability to find patterns and anomalies in data is a desirable skill for a statistician. Here are some exercises to help you practice these skills.

It Fits You Perfectly

The data set sasdata.regpara contains x-y data. Regress y on x. How do you like the fit?

Trio

At a postal distribution center, a first class package is selected every 15 minutes at random from a specific point on the shipping queue and weighed. The data in the set sasdata.atrio consist of weights of 100 consecutive packages taken on each of three separate days. Characterize and compare the distributions obtained on the three days.

Homer’s Paradox (DOH!)

Consider results of a survey relating marital status, gender and age. The resulting three-way table and the combined table formed by collapsing over age are shown in Table 1. What the heck is going on here?

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Marital Status</th>
<th>Married</th>
<th>Not Married</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young</td>
<td>Male</td>
<td></td>
<td>1623</td>
<td>809</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
<td>4481</td>
<td>3195</td>
</tr>
<tr>
<td>Old</td>
<td>Male</td>
<td></td>
<td>1412</td>
<td>4271</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
<td>811</td>
<td>3538</td>
</tr>
<tr>
<td>Combined</td>
<td>Male</td>
<td></td>
<td>3035</td>
<td>5080</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
<td>5292</td>
<td>6733</td>
</tr>
</tbody>
</table>

Table 1: Results of a survey relating marital status, gender and age.

John Barleycorn Must Die

In his 1993 book Visualizing Data, William S. Cleveland presents a data set consisting of barley yields. He describes it as follows:

“In the early 1930s, agronomists in Minnesota ran a field trial to study the crop barley. At six sites in Minnesota, ten varieties of barley were grown in each of two years. The data are the yields for all combinations of site, variety, and year, so there are $6 \times 10 \times 2 = 120$ observations ... The barley data have been analyzed and re-analyzed for decades. Their first analysis appeared in a 1934 report published by the experimenters. The statistician and geneticist R. A. Fisher, who established the modern foundations of statistics, presented the data for five of the sites in his book, The Design of Experiments. Francis J. Anscombe and Cuthbert Daniel, pioneers of diagnostic methods for determining when statistical models fit data, also analyzed them.

... an anomaly ... was missed in these previous analyses ... ”

The data are found in sasdata.barley and sasdata.barley2 (two different formats). Your task is to find the anomaly in the data.
The Third Kind Is The Charm

Kimball (1957) defines an error of the third kind (or Type III error) as providing the right answer to the wrong question. To illustrate, consider the following:

The goal of the project was to develop an on-site test to measure the quality of asphalt mix used in road surfacing. The test would result in a numerical measure which the researchers wanted to compare with a range of values obtained from known “good” mixes. The researchers proposed comparing the measure with 95% confidence limits developed from a large sample of good mixes.

Explain why this is an error of the third kind. What is the right question, and its answer?

1Stolen from D. J. Hand.