

## Organization of data files

<http://marekrychlik.com/node/28>

{Kuehl discusses an example of a study in which beef steaks are packaged in controlled atmosphere to prevent the growth of meat-spoiling bacteria on the surface of the steak. The data are organized in a table. Each row of the table represents a single observation. `table cellpadding="1" cellspacing="1" border="1" align="center" summary="In summary, this happens"` `caption="Beaf steak packaging data"` `tbody` `tr` `th`Steak `th` Treatment `th` `th`

`Log(count/cm2)` `i` `tr` `td` `td`Commercial `td` `td`7.66 `tr` `td` `td`Commercial `td` `td`6.98 `tr` `td` `td`Commercial `td` `td`7.80 `tr` `td` `td`12 `td` `td`Vacuum `td` `td`5.26 `tr` `td` `td`Vacuum `td` `td`5.44 `tr` `td` `td`3 `td` `td`Vacuum `td` `td`5.80 `tr` `td` `td`10 `td` `td`Mixed Gas `td` `td`7.41 `tr` `td` `td`9 `td` `td`Mixed Gas `td` `td`7.33 `tr` `td` `td`2 `td` `td`Mixed Gas `td` `td`7.04 `tr` `td` `td`8 `td` `td`CO<sub>2</sub> `td` `td`3.51 `tr` `td` `td`4 `td` `td`CO<sub>2</sub> `td` `td`2.91 `tr` `td` `td`11 `td` `td`CO<sub>2</sub> `td` `td`3.66 `tr` `tbody` `table`

{The first column is a random permutation resulting in a random assignment of each of the 12 steaks to a treatment group. The second column identifies the treatment (quot;Commercialquot;, quot;Mixed Gasquot;, and quot;CO<sub>2</sub>quot;), and finally the last column is the (treatment) response, which in this case is the logarithmic count of the bacterium found on the surface of each steak.

{The first column is primarily for bookkeeping reasons, and could be discarded in performing the statistical analysis. We note that discarding the first column breaks the connection between individual steaks and the data. In medical studies, typically we would discard the unique identifier of a human subject for privacy reasons, before publishing data, if required by the law.

{The table can be stored in a variety of file formats, facilitating interaction with statistical software:

- {CSV (comma-separated values)
- {Databases (MySQL, Postgress, DB2, etc)
- {Plain text files
- {Custom binary files

{The choice of the format is not important for small data sets. For large datasets, a plain text file will be inefficient because of the extra storage required to store binary numbers and the necessary conversion between the decimal and binary formats while reading and writing numbers. The same comment applies to CSV. Databases and binary files with fixed-width fields will utilize efficient I/O operations of the hard drive, and will occupy less storage. Also, the access speed to data will be reduced.