Caveat Emptor: On the Use, Mis-use, and Re-use of Statistics in Computational Biology

CMP BIO 201

Sandrine Dudoit
Division of Biostatistics and Department of Statistics
University of California, Berkeley
www.stat.berkeley.edu/~sandrine

Fall 2013
What/Why/How (Bio)Statistics?
1. What is Statistics?

2. What is Biostatistics?

3. Why (Bio)Statistics?

4. How (Bio)Statistics?

5. Caveat Emptor

6. Reproducible Research
   - What/Why/How?
   - R Software
   - Resources
What is Statistics?

Many views …

- We muddle through life making choices based on incomplete information. (Gonick and Smith, 1993).

- What makes Statistics unique is its ability to quantify uncertainty, to make it precise. This allows statisticians to make categorical statements, with complete assurance — about their level of uncertainty. (Gonick and Smith, 1993).
What is Statistics?

Figure 1: Gonick and Smith (1993). Cartoon Guide to Statistics.
What is Statistics?

Figure 2: Gonick and Smith (1993). Cartoon Guide to Statistics.
What is Statistics?

Figure 3: Gonick and Smith (1993). Cartoon Guide to Statistics.
• Statistics is the art of making numerical conjectures about puzzling questions. (Freedman et al., 1978).

• The raw material of a statistical investigation is a set of observations; these are the values taken on by random variables $X$ whose distribution $P_{\theta}$ is at least partly unknown. ... Statistical inference is concerned with methods of using this observational material to obtain information concerning the distribution of $X$ or of the parameter $\theta$ with which it is labeled. (Lehmann, 1986).
What is Statistics?

• **The objective of statistics is to make inferences** *(predictions, decisions)* **about a population** based on information contained **in a sample**. (Mendenhall, 1987).

• **Statistics is an area of science concerned with the extraction of information from numerical data and its use in making inferences about a population** from which the data are obtained. (Mendenhall, 1987).

• **The statistician studies various inferential procedures**, looking for the **best predictor or decision-making process** for a given situation. Even more important, the statistician provides information concerning the **goodness of an inferential procedure**. (Mendenhall, 1987).
There is a long history of joint development between statistics and biology.

- Mendel (1866), www.mendelweb.org. Mendel’s laws of heredity were entirely based on statistical inference.
- Statistical analyses allowed Mendel to make the bold leap from experimental results to theoretical conclusions.
- We now know that Mendel’s hypothetical factors are genes, i.e., segments of DNA that code for proteins.
- Experimental confirmation came much later: Genes lie on chromosomes (Sutton, 1903; Morgan, 1910); DNA is the genetic material (Avery et al., 1944; Hershey and Chase, 1952); double helical structure of DNA (Watson and Crick, 1953); genetic code (Nirenberg, 1961).

- *Genetics and Statistics* have in common that they are both characteristic products of the twentieth century.
- *... connection between our two subjects ... the “factorial” method of experimentation ...* derives its structure, and its name, from the simultaneous inheritance of Mendelian factors.
- *Its [Genetics’s] characteristic frequencies are a constant stimulus to statistical inquiry.*
- *... beautifully randomized by the meiotic process.*
- *Quite suddenly in the intellectual history of mankind it has become possible to think coherently and confidently about variation ...*
- *Experimental design has become an intelligible subject for discussion ...*
• *It is not, I believe, sufficiently realized that this need for absolute realism is particularly required in statistical work when applied to genetic purposes. It is in general the statisticians task to bring theory into a truly organic coherence with objective and verifiable observations.*

• *Direct contact with what is actually done in experimentation helps the statistician in another very essential way, by leading him to consider variations in procedure, and the reasons why one method is to be preferred to others. The whole wide subject of experimental design is opened out by this consideration.*
What is Biostatistics?

2013 ... Today’s statistical inference problems in biology and medicine are truly multivariate and involve the joint analysis of multiple, diverse, and high-dimensional datasets.

- Multi-species nucleotide and protein sequences.
- High-throughput microarray and sequencing genome-wide measures for
  - identity-by-descent (IBD) states (GMS);
  - single nucleotide polymorphisms (SNP);
  - DNA copy numbers (CGH DNA-Chip/DNA-Seq);
  - transcript (mRNA) levels (RNA-Chip/RNA-Seq);
  - protein-nucleic acid interactions, e.g., transcription factor binding, histone modification (ChIP-Chip/ChIP-Seq);
  - DNA methylation status (methyl-Chip/methyl-Seq).
- Phenotypes: Biological and clinical outcomes, e.g., cell type/state, affectedness/unaffectedness, quantitative trait, (censored) survival time, response to treatment.
What is Biostatistics?

- **Covariates**: E.g. Age, sex, environmental exposure, treatment, dose, time.
- **Biological annotation metadata**: In-house or WWW, e.g., Gene Ontology (GO), pathway (KEGG), protein structure (PDB), literature (PubMed).
What is Biostatistics?

Figure 4: Biomedical and genomic data.
Why (Bio)Statistics?

• Design of experiment/study/sampling procedure.
  ▶ Ensure that the question of interest can actually be answered with the experiment or study. Cf. Bias, confounding.
  ▶ Answer the question as accurately as possible, given available resources (e.g., $$, biological samples, reagents, time). Cf. Sample size, variance.

• Inference, i.e., parameter estimation and hypothesis testing.
  ▶ Provide answers to scientific questions, i.e., decisions/estimates/predictions.
  ▶ Provide measures of confidence/goodness/reliability in these answers, i.e., bias, standard errors, risk, false positive (Type I) and false negative (Type II) error rates.
Why (Bio)Statistics?

Data complexity.

- **Dimensionality.**
  - “Large \( n \)” e.g., astronomy, marketing, social networks.
  - “Small \( n \), large \( p \)” e.g., microarray and sequencing data.

- **Multiple types.** Quantitative (continuous, discrete), qualitative, text, graph, image, sound.

- **Censored, erroneous, missing.**

- **Various levels of processing.** E.g. Microarray and sequencing data.

- **Dynamic and evolving.** E.g. DNA sequence (GenBank), Gene Ontology (GO), literature (PubMed).

- **Multiple sources and locations.** In-house, WWW.

  🔄 No longer just numerical data or \( X_{n \times p} \)!
Complexity: Microarray low-level data.

- **Experimental variables.** A variety of experimental variables are relevant in terms of bias, variability, and confounding: Clone origin, PCR amplification, print-run, target sample preparation, hybridization conditions, laser detection, etc. Cf. **MIAME**: Minimum Information About a Microarray Experiment.

- **Image analysis.** The “raw” data consist of one or two TIFF files for each hybridization. One must perform addressing, background correction, and segmentation to obtain probe-level hybridization measures.

- **Normalization.** Adjust expression measures to ensure that observed differences in expression measures between genomic regions and/or samples are truly due to differential expression and not experimental/technical artifacts.
• **Expression quantitation.** Infer/deconvolve exon/gene/isoform-level expression measures from probe-level expression measures.
Complexity: High-throughput sequencing (HTS) low-level data.
E.g. Illumina sequencing.

- **Experimental variables.** A variety of experimental variables are relevant in terms of bias, variability, and confounding: Flow-cell/lane/tile, library preparation, size selection, adapter ligation, laser detection, etc.
  Cf. **MINSEQE:** Minimal Information about a high-throughput SEQuencing Experiment.

- **Image analysis.** The “raw” data consist of a collection of TIFF files for each flow-cell. One must perform registration, addressing, background correction, and segmentation to obtain fluorescence intensities for the four nucleotides at each cycle.
Why (Bio)Statistics?

- **Scale.** One flow-cell
  \( \sim 8 \text{ lanes} \times 100 \text{ tiles} \times 4 \text{ nucleotides} \times 32 \text{ cycles} \)
  \( \sim 100,000 \text{ images} \)
  \( \sim 1 \text{ terabyte (TB) of data.} \)

- **Complexity.** Segmentation: Unlike microarray probe sequences, sequence clusters are not located on a grid. Registration: 4 images per cycle, multiple tiles.

- **Base-calling.** For each cluster, deconvolve base sequence from measured fluorescence intensities of the four nucleotides at each cycle. Obtain base-level and read-level quality scores. Cf. Cross-talk, cycle effect.

- **Read-alignment/mapping.** Assign reads to positions in the genome, transcriptome, or other reference sequence. Cf. Multiple hits, mismatches, paired-end reads.
Why (Bio)Statistics?

- **Normalization.** Adjust expression measures to ensure that observed differences in expression measures between genomic regions and/or samples are truly due to differential expression and not experimental/technical artifacts.
- **Expression quantitation.** Infer/deconvolve exon/gene/isoform-level expression measures from base-level read counts.
Why (Bio)Statistics?

Complexity: High-throughput gene expression higher-level data. Joint analysis of multiple, diverse, and high-dimensional datasets.

- **Genome-wide expression measures** from microarray or HTS.
- **Phenotypes.** Biological and clinical outcomes, e.g., cell type/state, affectedness/unaffectedness, quantitative trait, (censored) survival time, response to treatment.
- **Covariates.** E.g. Age, sex, environmental exposure, treatment, dose, time.
- **Biological annotation metadata.** In-house or WWW, e.g., Gene Ontology (GO), pathway (KEGG), protein structure (PDB), literature (PubMed).
Data complexity in other areas of application of statistics.

- **Astronomy.** Classification of celestial objects (e.g., comets, quasars).

- **Brain imaging.** Analysis of 3D brain images from magnetic resonance imaging (MRI), positron emission tomography (PET), and functional MRI (fMRI), to study brain structure, function, and neurochemistry (e.g., in studies of Alzheimer’s disease).

- **Business.** Analysis of telephone call records and web logs for the purposes of fraud detection, market research, usage management.

- **Telecommunications.** Classify e-mail messages as spam or not, based on characters in subject header, sender’s address, and other features of message.
Aspects of biostatistical practice.

- **Formulate biological question.**
- **Translate** biological question into a **statistical question**, i.e., define population, data generating distribution, and parameter of interest.
  - **Essential and perhaps hardest step.**
- **Design** experiment/study/sampling procedure.
- **Collect, store, access, and pre-process data** from a sample drawn from the population of interest (and other sources, e.g., WWW database).
- **Summarize and visualize** data – Beyond histograms and pie charts!
- **Derive estimator** of parameter of interest and associated measures of goodness, e.g., bias, standard error, risk.
• **Test hypotheses** concerning parameter of interest and control suitably defined Type I and II error rates.

• **Answer scientific question**, i.e., **make inference about population/data generating distribution** from which sample was drawn.

Current questions in biological and medical research raise challenges in each of these areas.
The development and application of sound statistical methodology involves constant feedback between

- subject matter (Fisher, 1952);
- software design and implementation;
- computational methods;
- mathematics.
Mathematics.

- **Mathematical language and methodology** (e.g., algebra, analysis, graph theory, probability theory) are key for establishing the **theoretical foundations and general properties** of a statistical method (e.g., optimality and asymptotic properties of estimators).

- **Statistical methods** motivated by a very specific biological question are typically applicable to a much broader class of problems, e.g., in astronomy and marketing.

- **Maintaining a certain level of mathematical abstraction** is important for
  - a **clean, precise, rigorous, and general formulation** of a statistical question and method,
  - not getting lost in irrelevant details,
  - comprehending the **full potential** of a method.
Computational methods.

• Current statistical questions in biology and medicine involve inference for high-dimensional multivariate distributions, with complex and unknown dependence structures among variables. E.g. Genetic mapping with SNPs; nucleotide and protein sequence analysis; high-throughput microarray and sequencing gene expression experiments; biological annotation metadata analysis.

• The application of statistical inference methods poses challenging computational and numerical analysis questions, e.g., matrix algebra, optimization, resampling, simulation. E.g. Loss-based estimation involves risk minimization over large parameter spaces; multiple testing involves bootstrap estimation of high-dimensional test statistics joint null distribution.
Software design and implementation. Statistical computing is an essential link for reliable and timely development and application of statistical methodology.

- Collect, store, access, and manipulate multiple and diverse datasets.
- Summarize, visualize, and report data and results, at various stages of analysis.
- Explore and implement new statistical methodology.
- Disseminate statistical methodology.
- Extensibility, interoperability, scalability.
- Reproducible research: Integrated, dynamic, and reproducible statistical documents (e.g., Sweave).

R Project (www.r-project.org) and Bioconductor Project (www.bioconductor.org).
Important notions raised in previous definitions of statistics:

- puzzling questions;
- population, sample;
- experimental design, randomization;
- data, observations;
- incomplete information, uncertainty, variation;
- random variable, distribution, parameter;
- choice, conjecture, decision, estimation, inference, information, prediction, test;
- accuracy, confidence, goodness, precision, reliability;
- an art, ...

... but no longer just numerical data or $X_{n \times p}$. 

...
What/Why/How (Bio)Statistics?

Attempt at synthesis.

- Given a sample from a population of interest, estimate and test hypotheses concerning parameters, i.e., functions, of the unknown data generating distribution.
- Understand, characterize, and model randomness/uncertainty, i.e., distribution of estimator.
- Extract signal (= biological and medical knowledge) from noise (= experimental artifacts, sampling variability), based on multiple and diverse data sources.
- Use statistical and computational skills to learn about the world.
Caveat emptor. Statistics is pervasive in biology and medicine but ...

- *Reinventing the wheel.*
  It is worth revisiting STAT 2 and going back to basics.
  E.g. Multiple rediscoveries, reincarnations, and rebrandings of \( t \)-statistics or Fisher (1936) linear discriminant analysis in the computation biology literature.

- “*Pourquoi faire simple quand on peut faire compliqué?!*”
  Do not fall for the hottest trends and buzzwords.
  Do not lose the forest for the trees.
  **Look at the data:** Basic numerical and graphical summaries.
  E.g. A good old-fashioned and simple prediction method such as Fisher (1936) linear discriminant analysis is more intuitive and often performs better than the latest neural network or support vector machine.
Caveat Emptor

- **Of hammers and nails.**
  
  *Horses for courses.*

  Make sure that the statistical methods answer the biological question of interest.

  It is all too common to look for a nail for a favorite hammer or to provide orphan answers without posing a question. E.g. Omnipresence of HMM in computational biology.

- **In model land.**

  Check that modeling assumptions are reasonable and that answers are data-driven rather than model-driven.

• *Whence statistics?*
  Is there a statistical inference problem?
  Does it make sense to compute estimators and test hypotheses?
  E.g. Provide standard errors/perform a *t*-test for an entire population.

• “*Quand on ne sait pas où l’on va, il faut y aller!! ... et le plus vite possible.*”
  Pause and think 😊.

These basic, if not trivial, *common sense* principles are often lost when faced with complex data structures.
Figure 5: *Les devises Shadok*: “*Pourquoi faire simple quand on peut faire compliqué?*!”
Figure 6: Les devises Shadok: “Quand on ne sait pas où l’on va, il faut y aller!! ... et le plus vite possible.”
Reproducible Research: What/Why/How?

- **What?** Reproducible research refers to the ability to regenerate, given the same input data, all of the computational results presented in a given publication, e.g., tables, figures.

- **Why?** Good practice, cf. the scientific method.

- **How?** Then, the lab book. Now, compendia with documentation text, “raw” data, code, and software.

- **Raises a variety of scientific, editorial, and legal issues.**

- **Fairly narrow definition of computational reproducibility, to be distinguished from biological reproducibility, i.e., whether the biological findings from one study hold in another. Cf. Steve Horvath’s “reproducible network module”**.
• There is a long history of advocacy for reproducible research.

• Interest in reproducible research is gaining momentum among the biological and statistical communities, with a few controversial studies that made headlines.

• Duke’s irreproducible research. Potti et al. (2006) propose an approach for predicting patient sensitivity to chemotherapeutic drugs based on in vitro drug sensitivity and gene expression measures. Coombes et al. (2007) and Baggerly and Coombes (2009) discuss their failure to reproduce the results in Potti et al. (2006), despite using the same data and software. We do not believe that any of the errors we found were intentional. We believe that the paper demonstrates a breakdown that results from the complexity of many bioinformatics analyses. This complexity requires extensive double-checking and documentation to ensure both data
validity and analysis reproducibility. We believe that this situation may be improved by an approach that allows a complete, auditable trail of data handling and statistical analysis. We use Sweave, a package that allows analysts to combine source code (in R) and documentation (in \LaTeX) in the same file. Our Sweave files are available at (bioinformatics.mdanderson.org/Supplements/ReproRsch-Chemo). Running them reproduces our results and generates figures, tables and a complete PDF manuscript.

- **Climategate.** McShane and Wyner (2011) and discussion examine method for inferring surface temperatures over the last 1,000 years.
- **Hothorn and Leisch (2011).** *Case studies in reproducibility.*
• The scientific method. Aristotle (384 BC–322 BC), Descartes (1596–1650), Newton (1643–1727), etc.

• Roger Bacon (c. 1214–1294). A repeating cycle of observation, hypothesis, experimentation, and the need for independent verification. Record the manner in which experiments are conducted in precise detail so that others can reproduce and independently test results.

• Claerbout. An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.
Reproducible Research: Approaches

- Knuth (1992). Literate Programming. *Literate programming is an idea that was introduced by Knuth (1992) and implemented in a variety of software tools such as noweb* (Ramsey, 1994). A literate program is a document that is a mixture of code segments and text segments. It is written to be read by humans rather than a computer and is organized as such. The text segments provide descriptions and details of what the code is supposed to do. The code itself must be syntactically correct but need not be organized in a fashion that can be directly compiled or evaluated. A literate program should support two types of transformation: weaving and tangling. (Gentleman and Temple Lang, 2004)

- Buckheit and Donoho (1995). Published figures should be accompanied by the complete software environment necessary for generating those figures.
Reproducible Research: Approaches


- Matti Pastell. *Pweave*. Literate programming interface for Python inspired from Sweave, but with Python replacing R.

- Gentleman and Temple Lang (2004). *Compendia*. A software framework for authoring and distributing these integrated, dynamic documents that contain text, code, data, and any auxiliary content needed to recreate the computations. The documents are dynamic in that the contents, including figures, tables, etc., can be recalculated each time a view of the document is generated. Our model treats a dynamic document as a master or “source” document from which one can generate different views in the form of traditional, derived documents for different audiences.
Reproducible Research: Approaches

- **Mesirov (2010).** *Accessible reproducible research.* This reproducible research system (RRS) is an adaptation of Microsoft Word that links to the Broad Institute’s GenePattern platform.

- **Donoho.** Permanently register each computational result with a unique universal result identifier (URI). The package formed by a URI, its associated content, and server behaviors yields a verifiable computational result (VCR).

- **Stodden.** The Reproducible Research Standard. Policies, copyright, and open licensing solutions.

- **Diggle and Zeger (2010).** *Biostatistics.* Associate Editor for Reproducibility. Articles are kite-marked: **D** if data are freely available, **C** if code is freely available, and **R** if both data and code are available.
• Scharpf et al. (2010). Compendium for *Using the R package crlmm for genotyping and copy number estimation* – uses Sweave (Ruczinski).
Reproducible Research: Recent Events

The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer, Symposium, AAAS Annual Meeting, Washington, DC, February 19, 2011.

www.stanford.edu/ vcs/AAAS2011

- Keith A. Baggerly. The Importance of Reproducibility in High-Throughput Biology: Case Studies.
- Fernando Perez. Reproducible Software versus Reproducible Research.
- Michael Reich. GenePattern.
- David Donoho. A Universal Identifier for Computational Results.
• Mark Liberman. Lessons for Reproducible Science from the DARPA Speech and Language Program.
Panelists: Keith Baggerly, Larry Kessler, and Roger Peng.
Chair: David Banks.

- Question 0: Why is reproducibility emerging as an ethical issue?
- Question 1: What is an investigator’s personal responsibility with respect to research ethics (including reproducibility) is his/her own lab?
- Question 2: What role, if any, should journals serve in ensuring reproducibility?
- Question 3: What is the institutional role?
• Question 4: What is the responsibility of federal agencies with respect to reproducibility and research ethics? How does this role differ when the agency is in the role of the grantor (e.g., NIH) versus the role of a regulator (e.g., FDA)?
Resources for reproducible research using R are listed at cran.r-project.org/web/views/ReproducibleResearch.html.

- We focus on the Sweave system.
- A more recent and general system than Sweave is provided by the package knitr.
The Sweave system allows the generation of integrated, dynamic, and reproducible statistical documents, intermixing text, code, and code output (textual and graphical).

The source file is an executable document consisting of a collection of code chunks and documentation text chunks.

Sweave is applicable to R and \texttt{\LaTeX}.

Functions are provided in the R \texttt{utils} package.

Please consult the documentation for the functions \texttt{Stangle} and \texttt{Sweave} and the Sweave manual (www.statistik.lmu.de/~leisch/Sweave).

Note that other more general frameworks for reproducible research (not limited to R and \texttt{\LaTeX}) have been proposed (Gentleman and Temple Lang, 2004).
Sweave input. A text file which consists of a sequence of code chunks and documentation text chunks (noweb file).

- **Documentation text chunks**
  start with `@`;
  text in a mark-up language like \LaTeX.

- **Code chunks**
  start with `<<name>>=` and options;
  R or S-PLUS code.

- **File extension**: `.rnw`, `.Rnw`, `.snw`, `.Snw`.
Reproducible Research: Sweave

Sweave functions.

- **Stangle.** Given an input file (.Rnw), the *Stangle* function concatenates the code chunks into a .R script file.

- **Sweave.** Given an input file (.Rnw), the *Sweave* function executes the code chunks and includes their (textual and graphical) output, along with the documentation text chunks, in a *LaTeX* file (.tex) and postscript and/or PDF files, which can then be processed as usual, e.g., using `latex`, `pdflatex`. 
Sweave output. A single document, e.g., .tex file or .pdf file, containing the documentation text, the code, and the code output (text and graphs). The document can be automatically regenerated whenever the data, code, or documentation text change.
Figure 7: Sweave system. Main functions (Stangle, Sweave) and file formats.
This example illustrates how one may embed R code and code output into a \LaTeX{} document.

\begin{verbatim}
<<summary>>=
data(Titanic)
class(Titanic)
dim(Titanic)
T
@ 

\begin{verbatim}
\begin{figure}
\begin{center}
<<mosaicPlot,fig=TRUE,echo=FALSE>>=
mosaicplot(Titanic, main = "Survival on the Titanic", color = TRUE)
@ 
\end{center}
\caption{{\em Mosaic plot of the Titanic survival data.}}
\end{figure}
\end{verbatim}
\end{verbatim}

\end{document}

Figure 8: \emph{Sweave system}. Sample \texttt{.Rnw} input file, \texttt{Sweave1.Rnw}.
### chunk number 1: summary

data(Titanic)
class(Titanic)
dim(Titanic)
Titanic

### chunk number 2: mosaicPlot

mosaicplot(Titanic, main = "Survival on the Titanic", color = TRUE)

Figure 9: Sweave system. Sample Stangle .R output file, Sweave1.R.
This example illustrates how one may embed R code and code output into a \LaTeX{} document.

\begin{Schunk}
\begin{Sinput}
> data(Titanic)
\end{Sinput}
\begin{Soutput}
[1] "table"
\end{Soutput}
\end{Schunk}

\begin{Schunk}
\begin{Sinput}
> dim(Titanic)
\end{Sinput}
\begin{Soutput}
[1] 4 2 2 2
\end{Soutput}
\end{Schunk}

\begin{Schunk}
\begin{Sinput}
> Titanic
\end{Sinput}
\begin{Soutput}
, , Age = Child, Survived = No
Sex
Class Male Female
1st  0      0
2nd  0      0
3rd 35     17
Crew 0      0
, , Age = Adult, Survived = No
Sex
Class Male Female
1st 118     4
2nd 154    13
3rd 287    89
Crew 670    1
, , Age = Child, Survived = Yes
Sex
Class Male Female
1st  5      1
2nd 11     13
3rd 13     14
Crew 0      0
\end{Schunk}

\begin{figure}
\centering
\includegraphics{mosaicPlot}
\caption{Mosaic plot of the Titanic survival data.}
\end{figure}

Figure 10: \textit{Sweave system}. Sample \texttt{Sweave .tex} output file, \texttt{Sweave1.tex}.
This example illustrates how one may embed R code and code output into a \LaTeX\ document.

```r
> data(Titanic)
> class(Titanic)
[1] "table"
> dim(Titanic)
[1] 4 2 2 2
> Titanic

, , Age = Child, Survived = No

<table>
<thead>
<tr>
<th>Sex</th>
<th>Class</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
<td>35</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Crew</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

, , Age = Adult, Survived = No

<table>
<thead>
<tr>
<th>Sex</th>
<th>Class</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
<td>118</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>154</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
<td>387</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Crew</td>
<td>670</td>
<td>3</td>
</tr>
</tbody>
</table>

, , Age = Child, Survived = Yes

<table>
<thead>
<tr>
<th>Sex</th>
<th>Class</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Crew</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

, , Age = Adult, Survived = Yes

<table>
<thead>
<tr>
<th>Sex</th>
<th>Class</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
<td>57</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>14</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
<td>75</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Crew</td>
<td>192</td>
<td>20</td>
</tr>
</tbody>
</table>
```

Figure 1: Mosaic plot of the Titanic survival data.

Figure 11: Sweave system. Sample \texttt{pdflatex .pdf} output file, \texttt{Sweave1.pdf}.
The Bioconductor Project has adopted the Sweave system for its vignettes, i.e., interactive task-oriented tutorials describing a package’s functionality.

Vignettes are located in the doc subdirectory of an installed package and are accessible from the on-line documentation browser, via the help.start function.

Vignettes can be used interactively.

Vignettes are also available separately and statically on the Bioconductor Project website.
Software tools are being developed for managing and using this new type of documentation.

- **vignette function (utils package):** View a specified vignette or list the available ones.

```r
<<vignettes, eval=FALSE>>=
vignette(all = TRUE)
vignette("grid")
v1 <- vignette("grid")
edit(v1)
Stangle(v1$file)
```

- **browseVignettes function (utils package):** List available vignettes in an HTML browser with links to PDF, LaTeX/noweb source, and (tangled) R code (if available).
The knitr package generalizes Sweave in a number of respects. It enables integration of R code into other types of documents than \LaTeX, e.g., Asciidoc, HTML, LyX, Markdown (text-to-HTML conversion tool), and reStructuredText. It provides support for other languages than R, e.g., C++, Python. It provides better modularization than Sweave. It allows caching (cache results from computations, so that computations are skipped next time). It yields more control over formatting of R code chunks via the formatR package and improved graphics (e.g., range of graphics devices, control of output). The knitr functions corresponding to the Sweave functions Sweave and Stangle are knit and purl, respectively.
• If accustomed to Sweave, a useful function is `Sweave2knitr`, which converts an Sweave document to a knitr-compatible document.
Reproducible Research: File Conversion

- **Markdown**: A text-to-HTML conversion tool.
- **pandoc**: A “universal document converter”, that can convert files from one mark-up format into another, e.g., DocBook, HTML, \LaTeX, Markdown, MediaWiki, reStructuredText, and textile.
Reproducible Research: Version Control

- **CVS**: Concurrent Versions System.
- **Git**: Initially designed for Linux kernel development.
- **GitHub**: Web-based hosting service for software development projects that use the Git revision control system.
- **svn**: Subversion.
• Scientists for Reproducible Research Google group: groups.google.com/group/reproducible-research?hl=en


