Chapter 7
Additional Topics

7.1 Statistics

Compared to other branches of mathematics, statistics is a young discipline. Take Student’s t-test, which assesses the hypothesis that two small samples are drawn from the same population. It was published just a century ago by William S. Gosset (1876–1937), who used “Student” as a pseudonym for his work in statistics [45]. Gosset trained as a chemist and worked all his life in the management of the Guinness brewery, first in Dublin and later in London. A central aim of the company leadership at the time was to make brewing “scientific”. This required above all experimentation on such things as the effect of the resin content of hops on beer quality. However, once the relevant measurements had been made, a structured approach to their interpretation was also needed; how large a difference in hop resin content made a significant difference to the lifetime of stout? Today we call the investigation of such questions “statistics” and Gosset was one of its pioneers. Rather than the resin content of hops, we take as our example an investigation of the effect of fatty food on gene expression in mice [10].

New Concepts

Name	Comment
false discovery rate	false-positive rate among rejected hypotheses
gene ontology	functional categories for all genes
mouse transcriptome	all transcripts of the mouse genome
type I error (false-positive rate)	reject true null hypothesis
type II error (false-negative rate)	don’t reject false null hypothesis

New Programs

Name	Source	Help
`simNorm`	book website	`simNorm -h`
`testMeans`	book website	`testMeans -h`

© Springer International Publishing AG 2017

B. Haubold and A. Börsch-Haubold, *Bioinformatics for Evolutionary Biologists*, https://doi.org/10.1007/978-3-319-67395-0_7

7.1.1 The Significance of Single Experiments

Problem 444 Eight mice were given standard food, and are called sample A. Eight mice were given fatty food, sample B. RNA was extracted from liver and quantified on hybridization chips. The files `all_a.txt` and `all_b.txt` contain the results for experiments A and B, respectively. We start by investigating a single gene, *Plin5*. Use `grep` to extract its expression levels and save them in files `plin5_a.txt` and `plin5_b.txt`.

Problem 445 Compute the average expression values of *Plin5* in both experiments.

Problem 446 Next we investigate the difference between the two averages using the program `testMeans` with default parameters. Is the difference between the estimated means significant?

Problem 447 The default method for calculating significance in `testMeans` uses a formula from Gosset's original work. However, this is based on the assumption that the two samples compared were drawn from a normal distribution. As this assumption may not hold, `testMeans` also provides a Monte Carlo test for computing P-values. Like gambling at the Monte Carlo Casino in Monaco, a Monte Carlo method in statistics is based on chance: Consider two samples, S_1 and S_2, and their means, m_1 and m_2. Then calculate the difference between the two means: $\Delta_0 = |m_1 - m_2|$. Now shuffle the elements of S_1 and S_2 between the sets and repeat the computation of their means and the difference between them. Repeat this n times to get $\Delta_1, \Delta_2, ...\Delta_n$. The significance of the difference between the two samples is the frequency with a $\Delta_i \geq \Delta_0, i = 1, 2, ..., n$. One implication of this method is that P cannot fall below $1/n$; in other words, the theoretical minimum of P depends on the number of shufflings we carry out. For this reason the user of `testMeans` can vary n. Compare the P-value obtained by `testMeans` using Monte Carlo and the P-value obtained using the default method.

Before comparing all the genes in experiments A and B, we investigate the statistics of multiple tests using simulated data.

7.1.2 The Significance of Multiple Experiments

Problem 448 The program `simNorm` generates samples drawn from a normal distribution. Use the program to generate 100 samples of size 8 with mean 12. Save the results in the file `experiment1.txt`. Repeat the simulation and save the results in `experiment2.txt`. Look at the first rows of the two files (`head`): They contain the values for sample 1, S1, followed by the values for sample 2, S2, and so on. We can interpret these samples as control/experiment for genes S1, S2, and so on. What is the number of false-positive results, the false-positive rate, if $\alpha = 0.05$ (`testMeans`)?

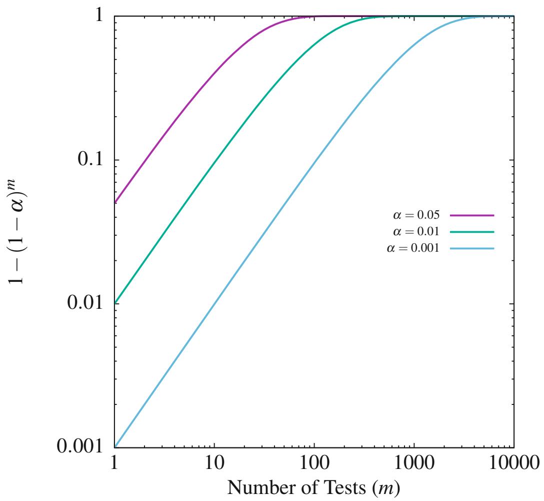

Fig. 7.1 A different view of the false-positive rate: The probability of obtaining at least one false-positive result as a function of the number of hypothesis tests

Problem 449 Repeat the estimation of the false-positive rate for 10^4 experiments.

Problem 450 As illustrated in Fig. 7.1, the more tests we carry out, the greater the probability that we obtain at least one false-positive. What is the probability of obtaining at least one false-positive with $\alpha = 0.05$ when carrying out 100 tests?

Problem 451 In statistics the false-positive rate is also known as the type I error. So far we set this to $\alpha = 0.05$ per experiment. However, we can also regard the 10^4 experiments as a single unit. Then we need to divide the original α by the number of tests carried out to assess the null hypothesis that the ensemble contains no sample with a significant difference. This division of α by the number of hypothesis tests is called the Bonferroni correction. What is the false-positive rate if we analyze our 10^4 experiments using the Bonferroni correction?

Problem 452 Now we simulate samples with different means. Run 10^4 experiments with $\mu = 6$ and $\sigma = 2.5$ and save them in `experiment1.txt`. Repeat the simulation with $\mu = 8$ and $\sigma = 2.5$, and save the result in `experiment2.txt`. What is the false-negative rate, β, if we leave $\alpha = 0.05$?

Problem 453 Repeat the simulation of samples with different means. Like before, run 10^4 experiments with $\mu = 6$ and $\sigma = 2.5$ and save them in `experiment1.txt`. For the second simulation, use $\mu = 12$ and $\sigma = 2.5$, and save the result in `experiment2.txt`. What is the false-negative rate this time?

Problem 454 Repeat the simulation of `exeriment2.txt` with $\mu = 12$ and the larger standard deviation $\sigma = 3.5$. How does β change?

Problem 455 Analyze `experiment1.txt` and `experiment2.txt` again, but this time use Bonferroni correction [40]. What is the false-negative rate, also known as type II error, now?

Problem 456 In genomics, we are often not primarily interested in eliminating the type I error, as this can lead to a large type II error. Hence the concept of false discovery rate, fdr, has been developed. The fdr is the fraction of false-positive results among the *rejected* hypotheses, rather than among all hypotheses. In order to set the fdr to some level δ, the original P-values are sorted $P_1 \leq P_2 \leq ... \leq P_m$; then P_j is significant if $P_j \leq \delta j/m$. This method is due to Benjamini–Hochberg [9] and hence also known as the Benjamini–Hochberg correction. Repeat the analysis using this correction. What is the type II error now?

Problem 457 Simulate two sets of 10^4 experiments with identical means $\mu_1 = \mu_2 = 6$ and standard deviations $\sigma_1 = \sigma_2 = 2.5$. Analyze the results using the Benjamini–Hochberg correction. How large is the type I error?

Problem 458 Our ability to detect an effect in an experiment depends on two quantities: effect size and sample size. To investigate sample size, simulate again 10^4 pairs of experiments with $\mu_1 = 6, \sigma_1 = 2.5$ and $\mu_2 = 12, \sigma_2 = 3.5$ for sample sizes of $n = 2, 5, 10, 20, 50$. Analyze the results using the Benjamini–Hochberg correction and plot the type II error, β, as a function of sample size, n.

Problem 459 Repeat the sample size simulation with a smaller effect size: $\mu_1 = 6, \sigma_1 = 2.5$, and $\mu_2 = 7, \sigma_2 = 2.5$. Since the effect size is small, carry out the simulation for a larger range of sample sizes: $n = 2, 5, 10, 20, 50, 100, 200, 500$. What sample size is necessary to drive β below 0.05? Hint: Plot a horizontal line in `gnuplot` using the syntax

```
f(x)=0.05
plot ..., f(x) t "" w l
```

7.1.3 Mouse Transcriptome Data

Problem 460 The files `all_a.txt` and `all_b.txt` contain the data for all the mouse transcriptome probes assayed in experiments A and B. How many probes were investigated (`wc -l`)? Some genes were assayed more than once. How many distinct genes were assayed (`cut`, `sort`, `uniq`, `wc -l`)?

Problem 461 Analyze the data and filter them using the Benjamini–Hochberg correction with the relatively permissive threshold of $\delta = 0.1$. Save the genes deemed significant in the file `genes.txt`. How many distinct genes does it contain?

Problem 462 To finish our analysis of mouse transcriptome data, we investigate whether the genes in in `genes.txt` are enriched for a particular function. Biological functions are codified *gene ontologies* (GO), which are hierarchical, for example genes involved in eye development are a subset of genes involved in development. Figure 7.2 shows the result of the GO analysis for our genes. Look for the highest node in the graph that is significant (red). Can you relate it to the underlying study?

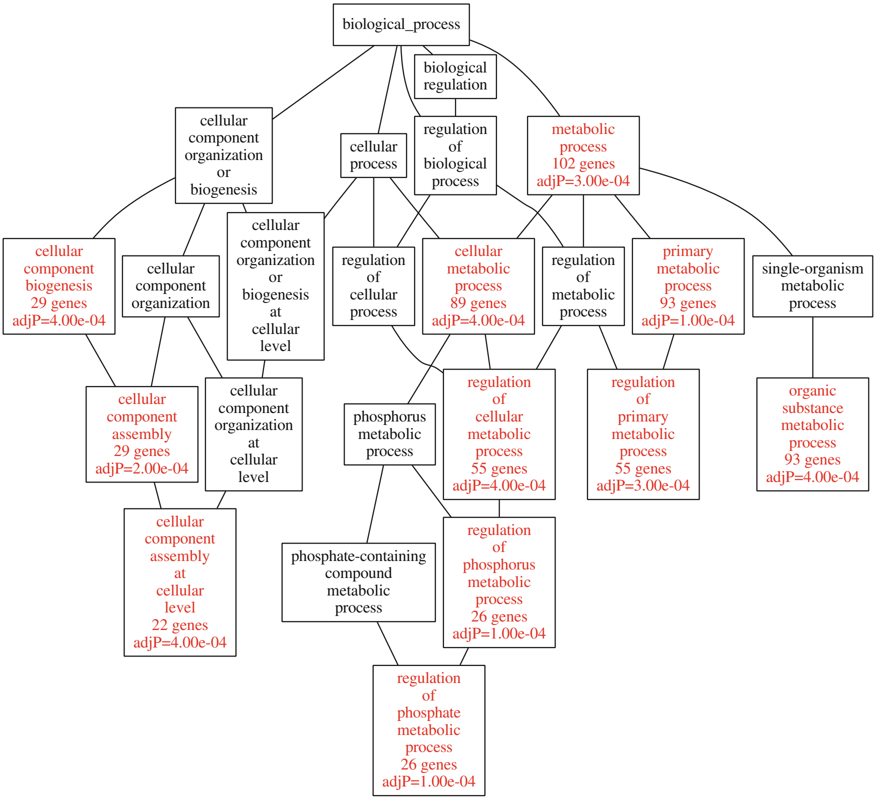
biological_process
biological regulation
cellular component organization or biogenesis
cellular process
regulation of biological process
metabolic process 102 genes adjP=3.00e-04
cellular component biogenesis 29 genes adjP=4.00e-04
cellular component organization
cellular component organization or biogenesis at cellular level
regulation of cellular process
cellular metabolic process 89 genes adjP=4.00e-04
regulation of metabolic process
primary metabolic process 93 genes adjP=1.00e-04
single-organism metabolic process
cellular component assembly 29 genes adjP=2.00e-04
cellular component organization at cellular level
phosphorus metabolic process
regulation of cellular metabolic process 55 genes adjP=4.00e-04
regulation of primary metabolic process 55 genes adjP=3.00e-04
organic substance metabolic process 93 genes adjP=4.00e-04
cellular component assembly at cellular level 22 genes adjP=4.00e-04
phosphate-containing compound metabolic process
regulation of phosphorus metabolic process 26 genes adjP=1.00e-04
regulation of phosphate metabolic process 26 genes adjP=1.00e-04

Fig. 7.2 Result of enrichment analysis using the gene ontology (GO) [47]. Computed using www.webgestalt.org

7.2 Relational Databases

In the late 1960s, the British mathematician Edgar F. Codd (1923–2003) proposed a new model for storing and accessing data. This model, called the relational data model, has become the standard way of dealing with large data sets. Originally used mainly in government and business, relational databases are now also ubiquitous in genomics. In this section we learn how to construct and query relational databases.

There are a number of software systems available for doing this and they all implement the query language SQL. However, there are differences, and the most important distinction is between systems with a client–server structure and those without. Systems with a client–server structure, such as Oracle, Mysql, and Postgresql, are usually centered on a server hosting one or more databases (Fig. 7.3). A potentially large number of clients connects via the internet to this server. As an example, we introduce and query the ENSEMBL database in this section. It contains genome data on vertebrates and is hosted on a public server under Mysql [5].

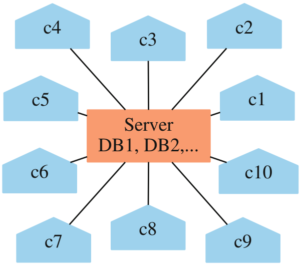

Fig. 7.3 A database server hosting several databases connected to ten clients

Server-client systems are powerful and correspondingly challenging to construct and administer, as opposed to mere querying. Fortunately, there are also simpler systems, where the database is just a local file. As an example for this kind of system we experiment with the Sqlite database.

New Concepts

Name	Comment
database client	program for accessing a database
database server	program for hosting a database
ENSEMBL	collection of vertebrate genome databases
Java	higher level programming language
relational databases	collections of data in tabular format
SQL	programming language for querying databases

New Programs

Name	Source	Help
`javac/java`	package manager	`man java`
`mysql`	package manager	`man mysql`
`sqlite3`	package manager	`man sqlite3`

7.2.1 Mouse Expression Data

Problem 463 The files `fatty_food.txt` and `normal_food.txt` contain a subset of the mouse transcriptome data we already used in Chap. 7.1. Figure 7.4 shows an Entity-relation (ER) model of the database we wish to construct. Boxes are *entities*, ellipses *attributes*, and the diamond denotes a *relationship*. It's a one-to-one relationship, where each entry in `fatty_food` has a corresponding entry in `normal_food`. The underlined attribute is a unique *primary key*. Here is the code for constructing table `normal_food`

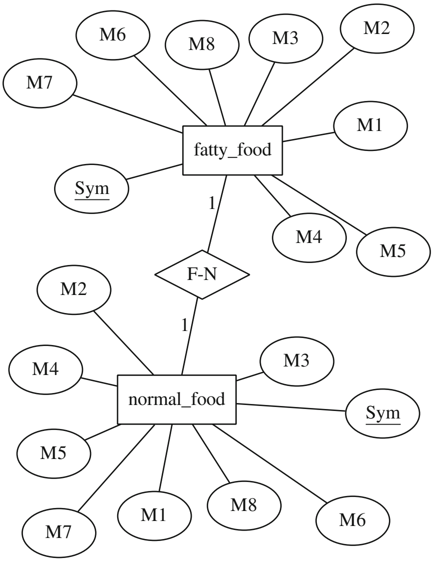

Fig. 7.4 Entity-relation (ER) model for `mouseExpressDb`

```
create table normal_food(
   Sym varchar(18),
   M1 float,
   M2 float,
   M3 float,
   M4 float,
   M5 float,
   M6 float,
   M7 float,
   M8 float,
   primary key(Sym)
);
```

Create the directory `RelationalDb`, and copy the data files into it. Write the code for constructing table `normal_food` into the the file `normal_food.sql`, and then write the corresponding file `fatty_food.sql`.

Problem 464 There is still one element of the ER-diagram in Fig. 7.4 missing from our SQL-code, the relationship. This is modeled by a *foreign key*, which is declared as

```
  foreign key(x) references some_table(y)
```

where attribute `x` refers to attribute `y` in `some_table`. We wish to model the fact that for every entry in `fatty_food` there is also an entry in `normal_food`, but not necessarily vice versa. Add code for the corresponding foreign key to `fatty_food.sql`.

Problem 465 Next, we construct the actual database mouseExpress.db using sqlite3. The following commands should get you there:

- Log on to a database: sqlite3 dbname; if dbname does not yet exist, it is created
- Switch on foreign keys: PRAGMA foreign_keys = ON;
- Read in a file containing SQL commands: .read <foo.sql>
- Show existing tables: .tables
- We use tabs to delineate columns instead of the default |: .separator "\t"
- Import data contained in file table.txt into table table: .import table.txt table
- Show all attributes of the first ten entries in a table: select * from table limit 10;
- Quit sqlite: .quit

Problem 466 Instead of adding many rows to a table using .import, we can add individual rows using the insert command:

```
insert into normal_food
values ('toy_gene1', 1.0, 2.0, 3.0, 4.0, 5.0, 6.0,
    7.0, 8.0);
```

Notice the semi-colon that closes every SQL command. Unless an SQL command is closed, sqlite prints the prompt

```
...>
```

indicating that it awaits further input. This will also occur if, for example, the single quote surrounding toy_gene is not closed again. If you are stuck with ...> and cannot get back to sqlite>, check to see what is keeping the command open. In most cases it will be a missing semi-colon. To retrieve the entry we just generated, enter

```
select * from normal_food where sym like 'toy_gene1';
```

Delete our toy entry

```
delete from normal_food where sym like 'toy_gene1';
```

and check the result

```
select * from normal_food where sym like 'toy_gene1';
```

Enter the following data to normal_food:

sym	m1	m2	m3	m4	m5	m6	m7	m8
toy_gene2	17.1	9.5	27.7	6.5	24.1	30.2	30.6	14.3

and delete them again.

Problem 467 What happens if we make a second entry for an existing sym like *Plin5*?

Problem 468 What happens if we try to enter '

sym	m1	m2	m3	m4	m5	m6	m7	m8
toy_gene3	3.4	8.0	4.4	26.7	8.6	26.6	4.8	20.5

into fatty_food?

7.2.2 SQL Queries

Problem 469 Recall form the database construction in Problem 465 that instead of entering commands interactively, they can also be read from files using .read. To experiment with this feature, write the four commands constructed in Problem 466 into the file insert.sql and enter them with .read. What happens if the insert command contains one value too many or too few?

Problem 470 Use the SQL-command count to determine the number of entries in normal_food and fatty_food.

Problem 471 We need one expression value per gene, the average of m1, m2, ..., m8. Construct a table with two columns, sym and the average expression; restrict the output to the first three genes.

Problem 472 Which gene has the largest average expression value in normal_food (max)? The smallest (min)? What is the average over the per gene average expression values (avg)?

Problem 473 Repeat these computations for fatty_food, that is, which genes have the largest and smallest average expression value? And what is the grand average expression value in fatty_food?

Problem 474 Next we compare normal_food and fatty_food. For this we need to join the two tables. Look at the first ten entries of the joined table:

```
select * from normal_food join fatty_food using (sym)
   limit 10;
```

Convert the join into a table with three columns: sym, the per gene average of normal_food, and the per gene average of fatty_food. Save the command in the file join.sql. Hint: In order to refer to attribute M1 from table normal_food as opposed to fatty_food, use the dot-notation as in

```
select normal_food.m1, fatty_food.m1
from ...
```

Problem 475 Our next aim is to compute the fold change between two average expression values. It is not straight forward to do this in SQL. So we leave `sqlite` (`.quit`) and enter on the command line

```
sqlite3 mouseExpress.db < join.sql |
tr '|' '\t'                          |
head
```

Problem 476 Instead of regenerating the joined table using `join.sql` and piping the result through `tr`, save this data in `avg.txt` for future reference.

7.2.3 *Java*

Problem 477 SQL is often embedded in a host language. Here is an example in Java for our database:

```
import java.sql.*;
public class MouseExpressDb{
    public static void main( String args[] ){
        String query = "select * from normal_food join
            fatty_food using(sym)";
        try{
             Class.forName("org.sqlite.JDBC");
             Connection c = DriverManager.getConnection
                 ("jdbc:sqlite:mouseExpress.db");
             Statement stmt = c.createStatement();
             ResultSet rs = stmt.executeQuery(query);
             while(rs.next())
                    // Access column #1
                    System.out.println(rs.getString(1));
        }catch(Exception e){
             System.err.println(e.getClass().getName() +
                    ": " + e.getMessage());
             System.exit(0);
        }
    }
}
```

Type this code into the file `MouseExpressDb.java` and compile it with

```
javac MouseExpressDb.java
```

and run the program

```
java -cp sqlite-jdbc-3.15.1.jar:. MouseExpressDb | head
```

where the `jar` file contains the JDBC-driver for `sqlite3`, which can be downloaded from the book web site. Next, copy `MouseExpressDb.java` to `MouseExpressDb2.java` and edit the code to search for the gene with the greatest difference between normal and fatty food.

7.2.4 ENSEMBL

Problem 478 To get a list of all databases hosted on the public ENSEMBL server, enter

```
mysql -h ensembldb.ensembl.org -u anonymous -e "show
   databases"
```

How many databases make up ENSEMBL?

Problem 479 We are interested in the databases for mouse (*Mus musculus*) and among those the `core` databases in particular. What is the version number of the latest `core` database for mouse?

Problem 480 The command

```
mysql -h ensembldb.ensembl.org -u anonymous -D
  someDatabase -e "show tables"
```

lists the tables of a particular database. How many tables make up the latest mouse `core` database?

Problem 481 To find out the attributes of a particular table, use

```
mysql ... -e "describe someTable"
```

Which attributes make up `seq_region`?

Problem 482 The following code returns the lengths of all mouse chromosomes

```
for a in $(seq 19) X Y
do
   mysql ... -e "select name,length from seq_region
      where coord_system_id = 3 and name = '$a'" |
   tail -n +2
done
```

What is the total length of the mouse genome?

Problem 483 What are the attributes of table `exon`?

Problem 484 Which fraction of the mouse genome is covered by exons?

Problem 485 Exons contained in all splice variants of a particular transcript are called "constitutive". Which fraction of the mouse genome is covered by constitutive exons?

Problem 486 Table `xref` contains the attribute `display_label`, which corresponds to the gene names we used in the expression analysis (`sym`). For each gene, `xref` also contains a `description` of its function. What is the function of the gene we found in Problem 475 with the greatest fold change in expressions?

Problem 487 Important entities in ENSEMBL such as genes or transcripts are labeled with a `stable_id`. A `stable_id` can, for example, be entered on the ENSEMBL web site (`www.ensembl.org`) to quickly and unambiguously look up information on a particular gene. We search for the `stable_id` that corresponds to *Hsd3b5* by joining `gene` and `xref` via `gene.display_xref_id` and `xref.xref_id`. What is the `gene.stable_id` of *Hsd3b5*?

Problem 488 Every gene has at least one, but possibly more, transcripts. They are listed in table `transcript`, which is connected to `gene` via the attribute `gene_id`. How many transcripts are known for *Hsd3b5*?