Exercise for week 4, StatB/E

Kang Li, kang@math.ku.dk

Group presentation tasks: Exercises 4.4 - 4.7

Elementary problems

Exercise 4.1. Plot the probability density function of the t-distribution and the chi-squared distribution for different choices of the degrees of freedom. How does the distribution change with different degrees of freedom?

Exercise 4.2. Load two experimental data sets by

```
load(url("http://www.math.ku.dk/~tfb525/teaching/statbe/Mir101.RData"))
```

The `RData` file contains two vectors, `liver.miR101` and `HCC.miR101`, representing the measurements of the Micro-RNA 101 expression in normal liver cells and cancer liver cells (Hepatocellular carcinoma).

1. Compute the maximum likelihood estimate for the mean of the Micro-RNA 101 expression in normal liver cells (which is the sample mean). Show that the 99%-confidence interval for the mean in normal liver given the observed expression data is
$$[-0.3593, -0.1654].$$
2. Compute the maximum likelihood estimate for the mean of the Micro-RNA 101 expression in cancer liver cells (which is the sample mean). Show that the 99%-confidence interval for the mean in cancer liver given the observed expression data is
$$[-0.0942, 0.2370].$$
3. Perform a two-sample t-test for the expression levels in normal and cancer liver cells. What conclusion can you draw from the result?

Exercise 4.3. Perform a χ^2 test for the dice experiment shown in one of the lecture slides. Can we say the two loaded dice have different frequency patterns?

The likelihood ratio test and the two-sample t-test

Exercise 4.4. Given two data sets $x = \{x_1, \dots, x_n\}$ and $y = \{y_1, \dots, y_m\}$, we want to test if x and y have equal mean. We assume both data sets are normally distributed and have equal variance. This is a case for the two-sample t-test with equal variance. Here we attempt to perform a likelihood ratio test and compare its p-value with the t-test p-value.

To describe the data sets x and y, we consider n independent and identically distributed (IID) random variables $X = \{X_1, \dots, X_n\}$ following a normal distribution $N(\mu_1, \sigma)$, and m IID random variables $Y = \{Y_1, \dots, Y_m\}$ also following a normal distribution $N(\mu_2, \sigma)$. Our null hypothesis is that the two distributions are identical, i.e. $H_0 : \mu_1 = \mu_2$.

For the likelihood ratio test, the full statistical model states that X and Y follow two normal distributions with unequal mean but equal variance, $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, with three parameters. The nested model states X and Y follow the same normal distribution $N(\mu, \sigma)$, with two parameters.

1. Simulate two groups of IID data following two normal distributions with different means but equal variance. For example, `x = rnorm(20, 2, 4)` and `y = rnorm(40, 2.5, 4)`.
2. Perform the maximum likelihood estimation for the full model and the nested model, and compute the maximized likelihood values (or the minimized minus-log-likelihood values).
3. Compute the likelihood ratio test statistic and the p-value.
4. Perform a two-sample t-test for x and y with equal variance. Do you get the same p-value?

R simulations

Exercise 4.5. Consider n IID random variables

$$X = \{X_1, \dots, X_n\}$$

following a normal distribution $N(\mu, \sigma)$ with unknown mean and variance. The 95%-confidence interval for μ is given by

$$I(X) = [\overline{X} - z_{0.975}\widehat{SEM}, \quad \overline{X} + z_{0.975}\widehat{SEM}].$$

Also see the lecture slides for details.

Here we verify by R simulations that the probability of $I(X)$ covering μ is 0.95:

$$P(\mu \in I(X)) = 0.95.$$

Example procedure (different approaches are encouraged):

1. Set $n = 20$, $\mu = 2$, $\sigma = 3$, $N = 10000$
2. Simulate an observation x of length n following $N(\mu, \sigma)$ with mean μ and standard deviation σ.
3. Compute the 95%-confidence interval for μ using data x.
4. Check if the confidence interval contains μ and output a True or False.
5. Repeat steps 2 - 4 for N times. How many repetitions give a True?

Note: This simulation is essentially the same as what we did in the exercise "Distribution of empirical mean" in the second week. If the statistic $t(X)$ follows a t-distribution, then of course the $I(X)$ constructed in the above way will have a 0.95 probability of covering μ.

Power of a statistical test

Exercise 4.6. We consider a simplified example to study the properties of statistical tests. Suppose a measurement, e.g. a blood test, that can be conducted on people to diagnose a certain syndrome. Denote the measurement result by the random variable $X \in \mathbb{R}$. The measurement conducted on healthy people follows a standard normal distribution $N(0, 1)$. The same measurement conducted on people having the syndrome follows a different normal distribution, $N(2, 1)$. To diagnose this syndrome, we employ a simple statistical test with the null hypothesis being that the person is healthy, i.e. $H_0 : E(X) = 0$. We use the measurement result as the test statistic, $t(X) = X$, and reject the null hypothesis if X is greater than a threshold value τ. This is a one-sided test and large values are extreme.

1. Suppose we use a threshold value $\tau = 2$. What is the probability of type I error, the probability of type II error, and the power of the test?
2. Calculate the probability of type I error for different τ values, and make a curve of the probability of type I error as a function of τ.
3. Make a curve for the probability of type II error as a function of τ.
4. Make a curve for the power of the test as a function of τ.

What conclusions can you make from the plots? This is a simplified toy example to study statistical testing, but the conclusions are quite general.

A case study of neural data (still continued)

We will study the neural ISI data for the last time. Load the data by

```
neuron = read.table("http://math.ku.dk/~tfb525/teaching/statbe/neuronspikes.txt",
    col.names="isi")
```

Exercise 4.7. Previously we have fitted the exponential model and the gamma model to the neural data. Since the exponential distribution is a special case of the gamma distribution when $\alpha = 1$, the exponential model is a nested model. Check by the likelihood ratio test whether the exponential model is an adequate replacement of the gamma model for the neural ISI data.