lab4 上机题目

要求：能详细介绍解题过程，用到的函数及相关知识点等

An experiment is designed to study the pathogenesis of breast cancer. Scientists collected tissues from breast cancer patients and healthy persons (here healthy just means no breast cancer). Totally, 982 tumor samples and 110 control samples were collected, and their corresponding RNA were extracted. After performing RNA-seq and bioinformatic analysis, they got the normalized gene expression levels (FPKM) of all samples. One table showed the FPKM value of tumor samples is named 'brca-rsem-fpkm-tcga-t-lab.txt', and another table is of control samples named 'brca-rsem-fpkm-tcga-lab.txt'.

1. (1) Measure the average FPKM of KIF1C gene in tumor samples. And use other descriptive statistics in R to describe the FPKM of KIF1C gene.
(2) Randomly take 200 tumor samples, compute the mean of KIF1C genes.
(3) Is there a difference between the mean of these 200 tumor samples and the mean of of all tumor samples for KIF1C gene, is the difference significant? Use alpha = 0.05
(4) Repeat steps 2 and 3 above two times, and compare the three results, use graph to show the differences if possible.
#Z-test (single population mean)

2. tp53 is an important tumor suppressor gene. Normally, it is expressed lower in normal tissue, and higher in cancer tissue. A survey claims that 60 out of 100 cancer patients has high expression of tp53. (The expression value above 1000 was regarded as high expression in this experiment).
(1) Try to extract tp53 expression in breast cancer patients. And save it in another file.
(2) Use descriptive statistics in R to describe tp53 expression in breast cancer patients.
(3) Use boxplot add dot plot to show the tp53 expression value in high expression group and low high group.
(4) For tp53 expression of breast cancer patients in this experiment, try to judge the survey claim above is accurate or not? Use alpha = 0.05
#Z-test (One-sample test of proportions)

3. One report claim obesity may increase the risk of breast cancer. Suppose 400 breast cancer patients and 35 healthy persons are obese in this experiment.

(1) Please check whether obese is related to breast cancer, use graph to show the result if possible. Use alpha = 0.05

(2) MC4R gene is regarded to related with obesity. Try to extract MC4R expression in breast cancer patients and healthy persons. And save them in one file.

(3) Use descriptive statistics in R to describe the MC4R expression value in two groups.

(4) Use violin plot to show the MC4R expression value in two groups.

#Z-test (Two-sample test of proportions)

4. (1) Measure the average FPKM of SWAP70 gene in control samples. And use other descriptive statistics in R to describe the FPKM of SWAP70 gene.

(2) Randomly take 50 control samples, compute the mean of SWAP70 genes.

(3) Is there a difference between the mean of these 50 control samples and the mean of of all control samples for SWAP70 gene, is the difference significant? Use alpha = 0.05

(4) Repeat steps 2 and 3 if sample number in the step 2 is 70, 80, compare the three results, use graph to show the differences if possible.

One sample t-test

5. (1) Measure the average FPKM of each gene in control samples and tumor samples. Sort the gene according to the mean of FPKM from big to small in control samples and tumor samples separately.

(2) Extract the same genes in control samples and tumor samples from top 500 genes according to the order in step 1. Use descriptive statistics in R to describe the expression value of extracted genes in two groups.

(3) For those genes in step 2, is there a difference of their mean of FPKM in two different conditions (control vs tumor). Use alpha = 0.05

(4) Draw the error bar plot of the mean of FPKM in two different conditions (control vs tumor) and label pvalue in the graph.

Paired t-test

6. The protein encoded by PDGFD gene is a member of the platelet-derived growth factor family. it is reported that PDGFD expression is associate with breast cancer and cervical cancer. Now scientists want to check whether the function of PDGFD is same in the two kinds of cancers. So an experiment was designed to study the difference of PDGFD expression in the two forms of cancer. One table showed the FPKM value of cervical cancer samples is named ‘cesc-rsem-fpkm-tcga-t-lab.txt’,

(1) Please measure the mean and variance of PDGFD FPKM in breast cancer and cervical cancer separately. And use other descriptive statistics in R to describe the FPKM of PDGFD gene.

(2) Try to judge whether the value of PDGFD FPKM is equal variance in the two forms of cancers.

(3) Try to judge whether there is difference of PDGFD expression in the two forms of cancer. Use alpha = 0.05

(4) Draw the error bar plot of PDGFD expression in the two forms of cancer, and label pvalue in the graph.

Independent samples