2019ADS2 Week10 T test

by Wanlu Liu, wanluliu@intl.zju.edu.cn

2019-11-18

1. Introduction

This R Markdown file contains a tutorial of how to do different form of t test manually and in R. For those problems that need to workout manually, please include the formula and process of how you get the final number.

2. one sample t test

Let's say a gene called SOX17, from the published dataset, we found that its average expression in different lines (more than 30) of human embryonic stem cells is 8.9 (unit in RPKM). From your experiment, you used a new human embryonic stem cell line that is generated in ZJE, and you did RNA-seq of three biological replicates to check its expression in this cell line and found out its expression is 15, 30,50 (RPKM). Is there enough evidence to show our embryonic stem cell's SOX17 expression is very different from others under significance level of 0.05?

Please write out the formula you used to calculte the statistics.

1. step 1 : State the hypotheses and identify the claim.

   $$H_0 : \mu = 8.9$$

   SOX17 expression in our cells is same as public average.

   $$H_a : \mu \neq 8.9$$

   SOX17 expression in our cells is different from public average.

2. Step 2: what distribution to use? One sample two tailed t test

3. Step 3: Find the critical value. Since

   $$\alpha = 0.05$$

   and the test is a two tailed test, and the df=3-1=6, the critical value is t = 3.182.

4. Step 4: Compute the test value.

   $$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

   $$t = \frac{\frac{15+30+50}{3} - 8.9}{\frac{\sqrt{\frac{(15-31.7)^2+(30-31.7)^2+(50-31.7)^2}{2}}}{\sqrt{3}}} = 2.245686$$

5. Step 5: Make the decision. Since the test value, $2.2457 < 3.182$, falls in the noncritical region, fail to reject the null hypothesis.

6. Step 6:Summarize the results. There is no enough evidence to support the claim that our embryonic stem cells express SOX17 differently with published dataset.

Please write out the R code calculte the statistics.

```
onesample=t.test(c(15,30,50),mu=8.9)
print(onesample)
```

```
##
##  One Sample t-test
##
## data:  c(15, 30, 50)
## t = 2.2457, df = 2, p-value = 0.1538
## alternative hypothesis: true mean is not equal to 8.9
## 95 percent confidence interval:
##  -11.95336  75.28669
## sample estimates:
## mean of x
##  31.66667
```

```
names(onesample)
```

```
##  [1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
##  [6] "null.value"  "stderr"      "alternative" "method"      "data.name"
```

```
onesample$statistic #what is this?
```

```
##       t
## 2.24569
```

```
onesample$parameter #what is this?
```

```
## df
##  2
```

```
onesample$p.value #what is this?
```

```
## [1] 0.1538114
```

```
onesample$stderr #what is this?
```

```
## [1] 10.13794
```

2. two sample t test

2.1 paired two sample t test

We followed a set of 5 paitents with acute myeloid leukemia. We want to investigate whether the oncogene AML1 expression is repressed after a new treatment. Thus we tested their AML1 expression before and after the therapy. The gene expression level fo AML1 before the treatment is:

$$x_1, x_2, x_3, x_4, x_5 = c(102, 340, 234, 332, 129)$$

. And the gene expression level fo AML1 after the treatment is:

$$y_1, y_2, y_3, y_4, y_5 = c(74, 56, 70, 104, 11)$$

. Is there enough evidence to support the claim that the new treatment significantly reduce the AML1 expression level in acute myeloid leukemia paitents under significance level of 0.05?

1. step 1 : State the hypotheses and identify the claim.

   $$H_0 : \mu_x < \mu_y$$

   expression level of AML1 after treatment is higher than before treatment

   $$H_a : \mu_x \geq \mu_y$$

   expression level of AML1 after treatment is less than before treatment

2. Step 2: what distribution to use? Two sample paired one tailed t test

3. Step 3: Find the critical value. Since

   $$\alpha = 0.05$$

   and the test is a one tailed test, and the df=5-1=4, the critical value is t = 2.132.

4. Step 4: Compute the test value.

   $$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

   since this is a paired t test, we can generate a new variable called d, where

   $d_1, d_2, d_3, d_4, d_5 = x_i - y_i = c(102-74, 340-56, 234-70, 332-104, 129-11) = c(28, 284, 164, 228, 118)$

   thus

   $$\bar{d} = \bar{x} - \bar{y} = (\frac{102+340+234+332+129}{5}) - (\frac{74+56+180+104+111}{5}) = 164.4$$

   $$s_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{(28-164.4)^2 + (284-164.4)^2 + (164-164.4)^2 + (228-164.4)^2 + (118-164.4)^2}{4}} = 98.8777$$

   Thus,

   $$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{164.4}{\frac{98.8777}{\sqrt{5}}} = 3.717821$$

5. Step 5: Make the decision. Since the test value, 3.717821 > 2.132, falls in the critical region, reject the null hypothesis.

6. Step 6:Summarize the results. There is enough evidence to support the claim that the new treatment can significantly reduce the expression of AML1 gene in acute myeloid leukemia paitents.

Please write out the R code calculte the statistics.

```r
x=c(102,340,234,332,129)
y=c(74,56,70,104,11)
twosamplepaired=t.test(x,y,paired=TRUE,alternavie="greater")
print(twosamplepaired)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = 3.7178, df = 4, p-value = 0.02051
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   41.62712 287.17288
## sample estimates:
## mean of the differences
##                   164.4
```

```r
names(twosamplepaired)
```

```
##  [1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
##  [6] "null.value"  "stderr"      "alternative" "method"      "data.name"
```

```r
twosamplepaired$statistic #what is this?
```

```
##        t
## 3.717821
```

```r
twosamplepaired$parameter #what is this?
```

```
## df
##  4
```

```r
twosamplepaired$p.value #what is this?
```

```
## [1] 0.02051348
```

```r
twosamplepaired$stderr #what is this?
```

```
## [1] 44.21945
```

2.2 unpaired two sample t test

There are two types of human embryonic stem cells (naive vs primed). We have the RNAseq data for naive hESC (4 biological replicate replicates) and primed hESCs (4 biological replicates). In each RNAseq dataset, there are 23368 genes identified. We want to find out those genes that are significantly differential expressed (either up regulated or down regulated) under significance level of 0.05. Hint, use unpaired t test to find out genes with p-value less than 0.05.

```r
geneexp=read.csv("week10_t_test_problemset_testdata.csv")
head(geneexp)
```

```
##         gname naive_hESC_r1 naive_hESC_r2 naive_hESC_r3 naive_hESC_r4
## 1 1/2-SBSRNA4         3.657         3.808         7.239         4.607
## 2        A1BG         0.038         0.035         0.146         0.028
## 3    A1BG-AS1         0.032         0.348         0.361         0.299
## 4        A1CF         0.004         0.000         0.006         0.003
## 5       A2LD1         0.490         0.404         0.192         0.137
## 6         A2M         0.087         0.067         0.089         0.063
##   primed_hESC_r1 primed_hESC_r2 primed_hESC_r3 primed_hESC_r4
## 1          4.429          8.190          3.364          6.431
## 2          0.120          0.096          0.000          0.034
## 3          0.331          0.356          0.201          0.527
## 4          0.026          0.036          0.009          0.006
## 5          0.264          0.315          0.120          0.221
## 6          0.549          0.801          0.521          0.728
```

```r
tail(geneexp)
```

```
##        gname naive_hESC_r1 naive_hESC_r2 naive_hESC_r3 naive_hESC_r4
## 23363   ZXDC         5.472         7.170         6.914         6.781
## 23364 ZYG11A        38.996        40.783        43.214        34.863
## 23365 ZYG11B         9.271         9.741         7.855         9.625
## 23366    ZYX        11.731         7.882         7.685         8.729
## 23367  ZZEF1         9.343         9.973         7.814         7.089
## 23368   ZZZ3        20.996        14.505         9.251        11.654
##       primed_hESC_r1 primed_hESC_r2 primed_hESC_r3 primed_hESC_r4
## 23363          5.866          6.931          4.513          5.958
## 23364          4.714          7.006          3.506          4.975
## 23365          8.533          9.096          6.318          7.359
## 23366          3.450          2.543          1.304          3.481
## 23367          2.608          3.334          1.208          2.705
## 23368         18.732         16.479         16.563         17.609
```

```r
dim(geneexp)
```

```
## [1] 23368     9
```

```r
pvalue=rep(0,nrow(geneexp)) #initiaze the pvalue vector
for (i in 1:nrow(geneexp)){
  pvalue[i]=t.test(geneexp[i,2:5],geneexp[i,6:9],paired=FALSE,alternavie="two.sided")$p.value
}
geneexp=data.frame(geneexp,pvalue)
geneexp.sig=geneexp[which(geneexp$pvalue<=0.05),]
head(geneexp.sig)
```

```
##      gname naive_hESC_r1 naive_hESC_r2 naive_hESC_r3 naive_hESC_r4
## 6      A2M         0.087         0.067         0.089         0.063
## 8    A2MP1         0.000         0.103         0.000         0.000
## 9   A4GALT        18.052        20.011        15.624        18.936
```

```
## 13    AAAS        31.372        33.736        24.251        35.966
## 14    AACS         4.916         5.331         3.980         4.007
## 18 AADACL3         0.043         0.023         0.006         0.006
##    primed_hESC_r1 primed_hESC_r2 primed_hESC_r3 primed_hESC_r4
## 6           0.549          0.801          0.521          0.728
## 8           0.236          0.564          0.494          0.297
## 9           0.169          0.201          0.051          0.141
## 13          5.663          5.642          2.596          7.424
## 14          1.706          1.831          0.891          2.500
## 18          0.179          0.128          0.251          0.247
##          pvalue
## 6  0.0033247865
## 8  0.0131604307
## 9  0.0003005896
## 13 0.0007544200
## 14 0.0009736323
## 18 0.0060345674
```

```
dim(geneexp.sig)
```

```
## [1] 9601   10
```

2.3 One step further - Multiple testing correction (Advanced thinking, optional)

2.3.1 Why Multiple Testing Matters?

Genomics usually have Lots of Data which means there will be lots of Hypothesis Tests in one experiment. For example, a typical RNAseq experiment might result in performing 20000 separate hypothesis tests (like what we did before). If we use a standard p-value cut-off of 0.05, we'd expect 1000 (20000*0.05) genes to be deemed 'significant' by chance (not reasonable). Thus we usually will perform multiple testing correction after we calculate p-value for genomics. You can refer to this coursera online course video if you are interested in https://www.coursera.org/lecture/statistical-genomics/multiple-testing-8-25-NsJfs

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.