Permutation

Outline

- Review of hypothesis testing
- Permutation tests
- Two-sample problem
- Tests of independence

Test of Significance

Example: Testing

- There are two possibilities: Version A or version B
- Which version is better?

Experiment

- Subjects are randomly assigned to group A or B
- Group A gets version A
- Group B gets version B

Results

A (n=100)	B (m=150)
4	2
8	6
6	4
4	6
...	...
Average: 4.2	Average: 5.1

Is the observed effect due to chance?

- If we repeat the experiment, the observed effect could be smaller or larger due to random variation.
- Is it likely that we would observe an effect this large if there were no effect in the population?
- This question can be answered by hypothesis tests.

Neyman-Pearson

- Null hypothesis H_0 -“effect not present”
- Alternative hypothesis H_1- “effect present”
- Test statistic T- measures the size of the effect
- p-value: probability that T would be as large as observed under H_0

Permutation Tests

Permutation tests are an interesting and conceptually simple alternative to traditional tests when the required distributional assumptions (typically, Gaussian assumptions) are likely to be violated.

References

- Why Permutation Tests are Superior to t and F Tests in Biomedical Research, Ludbrook and Dudley, The American Statistician, 52 (2) 1998, 127-132

Two-sample Problems

Independent samples

X (size n)	Y (size m)
4	2
8	6
6	4
4	6
...	...

Goal: Compare the distributions of X and Y

Permutation Null

- X and Y are i.i.d. with the sample distribution. So the labeling (X or Y) of the observations does not matter.
- Shuffling procedure

1. Input: Vectors X and Y of lengths n and m, respectively.
2. Form long vector $Z = (X, Y)$ of length $n + m$
3. Take a simple random sample of size n from Z and assign to X^*; assign remaining m to Y^*

Let's look at an example. Suppose we have two samples of data $x_1 = 1$, $x_2 = 0$, $x_3 = -1$; $y_1 = 4$, $y_2 = 7$.

Then $n = 3$, $m = 2$, $Z = \{1, 0, -1, 4, 7\}$, and $\nu = \{1, 2, 3, 4, 5\}$.

There are the number of possible ways we can permute ν that lead to distinct partitions $Z^* = (X^*, Y^*)$.

To help clarify, lets look at some.

1. $X^* = \{-1, 0, 1\}$, $Y^* = \{4, 7\}$
2. $X^* = \{-1, 0, 4\}$, $Y^* = \{1, 7\}$
3. $X^* = \{-1, 0, 7\}$, $Y^* = \{4, 1\}$
4. $X^* = \{-1, 4, 1\}$, $Y^* = \{0, 7\}$
5. $X^* = \{-1, 7, 1\}$, $Y^* = \{4, 0\}$

$$\vdots$$

The key feature of permuting the data from their original values is that if $F_x = F_y$, then the distribution of (X, Y) is the same as the distribution of (X^*, Y^*) no matter how we repartition the data.

Approximate permutation test procedure

Step 1. Compute the observed test statistic $\hat{\theta} = \hat{\theta}(Z, \nu)$

Step 2. For each replicated, indexed $b = 1, 2, \ldots, B$:

- Generate a statistic $\hat{\theta}^{(b)} = \hat{\theta}^{*}(Z^{*}, \pi_{b})$

Empirical p-value

Step 3. If large values of $\hat{\theta}$ support the alternative hypothesis, then compute the empirical p-value by

$$\hat{p} = \frac{1}{B+1}\left[1 + \sum_{b=1}^{B} 1\{\hat{\theta}^{(b)} \geq \hat{\theta}\}\right]$$

Step 4. Reject H_0 at significance level α if $\hat{p} \leq \alpha$.

Proportion of observed and shuffled $\{\hat{\theta}, \hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(b)}\}$ that exceed observed $\hat{\theta}$

Small p-value is evidence against H_0

Let's see an example using the `chickwts` dataset in `R`. Here we we study the weights of newly hatched chicks fed different supplements. The sample we'll label `X` was fed soybean, and the sample we'll label `Y` was fed linseed.

We will let $\hat{\theta}$ be the two-sample t-statistic, which we know is sensitive to difference in the means of the two distributions.

It will be interesting to compare the *p*-value obtained by referring to the t-distribution with the *p*-value from the permutation distribution.

```r
x=c(158,171,193,199,230,243,248,248,250,267,271,316,327,32
y=c(141,148,169,181,203,213,229,244,257,260,271,309)
## First we find the p-value for the
## two-sample t statistic by referring to the t(n+m-2)
t.test(x,y,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = 1.3208, df = 24, p-value = 0.199
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -15.57282  70.92996
## sample estimates:
## mean of x mean of y
##  246.4286  218.7500
```

Now we'll let $\hat{\theta} = |t|$ to test null hypothesis of equal mean against a two sided alternative, and use the randomization distribution for the p-value.

```r
B=10000
z=c(x,y)
nu=1:26
reps=numeric(B)
t0=t.test(x,y,var.equal=TRUE)$statistic
t0=abs(t0)
for(i in 1:B){
  perm=sample(nu,size=14,replace=FALSE)
  x1=z[perm]
  y1=z[-perm]
  reps[i]=t.test(x1,y1,var.equal=TRUE)$statistic
  reps[i]=abs(reps[i])
  }
p=mean(c(t0,reps)>=t0)
p
```

```
## [1] 0.1976802
```

So, the two-*p*-values are quite similar.

Nonparametric two-sample Problem

$$X_1, \ldots, X_n \sim F_x$$

$$Y_1, \ldots, Y_m \sim F_y$$

$$H_0 : F_x = F_y$$
$$H_1 : F_x \neq F_y$$

Do X and Y have different distributions?

Now let's study the same data, but with a test statistic that looks for any departure in the distributions of X and Y.

Let F be the cdf for weights of chicks supplemented with soybean, and let G be the cdf for weights of chicks supplemented with linseed.

Kolmogorov-Smirnov statistic

$$D = \max_{1 \leq i \leq N} |F_n(z_i) - G_m(z_i)|$$

where F_n is the empirical cdf estimate of F computed from the x-sample, and G_m is the empirical cdf estimate of G computed from the y-sample.

- Maximum difference between empirical CDFs of X and Y
- Measures "distance" between empirical distributions $= 0$ iff empirical distributions are equal

Permutation KS-test

Are the distributions of the two groups different?

- H_0 : soybean & linseed have same distribution
- H_1 : soybean & linseed have different distributions
- T: Kolmogorov-Smirnov statistic

Same as before, except T

```r
z=c(x,y)
B=10000
nu=1:26
D=numeric(B)
options(warn=-1) #or ks.test will warn of ties
D0=ks.test(x,y,exact=FALSE)$statistic

for( i in 1:B){
perm=sample(nu,size=14,replace=FALSE)
x1=z[perm]
y1=z[-perm]
D[i]=ks.test(x1,y1,exact=FALSE)$statistic
}

p=mean(c(D0,D)>=D0);p
```

```
## [1] 0.459854
```