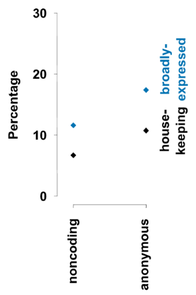
重点：Bootstrapping

2022年3月22日 星期二　09:13

The frequent evolutionary birth and death of functional promoters in mouse and human

GENOME RESEARCH

Robert S. Young,[1] Yoshihide Hayashizaki,[2] Robin Andersson,[3] Albin Sandelin,[3] Hideya Kawaji,[2,4] Masayoshi Itoh,[2,4] Timo Lassmann,[4] Piero Carninci,[4] The FANTOM Consortium, Wendy A. Bickmore,[1] Alistair R. Forrest,[4,5] and Martin S. Taylor[1]



- 21/313 noncoding promoters are housekeeping (6.7%).

```
> house_samples<-vector()
> for (rep in 1:1000){
+     sample1<-sample(noncoding, size = length(noncoding), replace = T)
+     house_length<-length(subset(sample1, sample1=="house"))
+     house_samples<-c(house_samples, house_length)
+ }
> summary(house_samples)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    8.0    18.0    21.0    20.9    24.0    36.0
> quantile(house_samples, probs = c(0.025,0.975))
 2.5% 97.5%
   13    30
> lower<-13/313
> lower
[1] 0.04153355
> upper<-30/313
> upper
[1] 0.09584665
```

※ sampling with replacement (T)

将所有值混合，重新抽样（抽相同个数的组成新样本）

解题思路（2类）

Ⓐ 看两样本中位数差别：（不符合正态）

permutation test

replace = F

sample 1 = c(2,2,3,3,5)　(len=5)

sample 2 = c(3,4,5,6)　(len=4)

(1) 计算 real median difference

(2) 将两样本混合，重取样，再计算 median difference

boot1 = sample (total, 5, F)

boot2 = sample1 + sample2 − boot1

median(boot1) − median(boot2)

(3) 将第2步重复多次 ⟶ results

p-value = mean (results ≥ real)

若 H0 为两样本 median 无差异 ⇒ abs(median(boot1) − median(boot2))

H0 为 sample1 > sample2 ⇒ median(boot1) − median(boot2)

Ⓑ 看置信区间，用于比较同一样本下两种数据有无差异

例：non-coding 中有21个 house（占比5%），求置信区间

case resampling

replace = T

```
> house_samples<-vector()
> for (rep in 1:1000){
+     sample1<-sample(noncoding, size = length(noncoding), replace = T)
+     house_length<-length(subset(sample1, sample1=="house"))
+     house_samples<-c(house_samples, house_length)
+ }
> summary(house_samples)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    8.0    18.0    21.0    20.9    24.0    36.0
> quantile(house_samples, probs = c(0.025,0.975))
 2.5% 97.5%
   13    30
> lower<-13/313
> lower
[1] 0.04153355
> upper<-30/313
> upper
[1] 0.09584665
```

95% CI ⇒ 95% 的确信度，真实的占比在此区间内

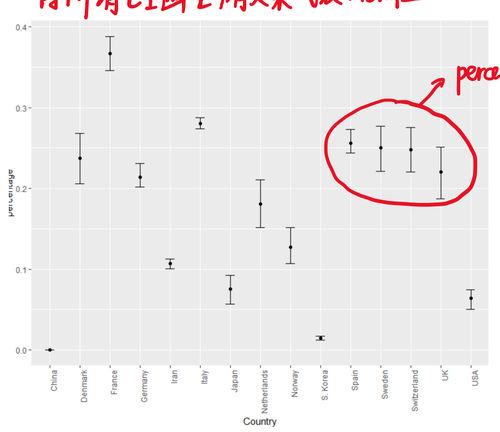
```
ggplot(data2, aes(x = Country, y = percentage)) +
  geom_point() +
  geom_errorbar(aes(ymin = lower, ymax = upper), width = 0.3) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

误差线

例：不同 country 的 percentage 是否不同（one sample）

	Country	Total	New	percentage	upper	lower
1	China	80796	18	0.0002227833	0.0002970444	0.0001172707
2	Italy	12462	3497	0.2806130637	0.2878811587	0.2734934200
3	Iran	10075	1075	0.1066997519	0.1128684864	0.1005781638
4	S. Korea	7869	114	0.0144872284	0.0170542636	0.0121330538
5	Spain	3059	782	0.2556390977	0.2729813665	0.2438542007
6	Germany	2502	536	0.2142286171	0.2306155076	0.2013988809
7	France	2281	838	0.3673827269	0.3882398071	0.3456488382
8	USA	1390	89	0.0640287770	0.0748201439	0.0507014388
9	Switzerland	867	215	0.2479815456	0.2757785467	0.2201845444
10	Norway	721	92	0.1276005548	0.1511789182	0.1067961165
11	Japan	691	52	0.0752532562	0.0926917511	0.0571273517
12	Denmark	674	160	0.2373887240	0.2678412463	0.2061572700
13	Sweden	667	167	0.2503748126	0.2773613193	0.2211019490
14	Netherlands	614	111	0.1807817590	0.2109527687	0.1514657980
15	UK	590	130	0.2203389831	0.2508474576	0.1872457627

将所有 CI 画出，有交集代表 H0 成立



percentage is the same

Bootstrapping -> investigate one sample -> replace = T

Permutation -> investigate 1+ samples -> replace = F (but combine the samples when performing the sampling).