Advanced Statistical Computing
Week 2: Monte Carlo Study of Statistical Procedures and Permutation Tests

Aad van der Vaart

Fall 2012

Contents

Sampling distribution

Sampling distribution

DATA X (possibly equal to $(X_1, \ldots, X_n)$)
STATISTIC $T = T(X)$

$\boxed{\text{The distribution of } T \text{ is called } \textit{sampling distribution}}$

EXAMPLE OF THEORETICAL SAMPLING DISTRIBUTION:
If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$, then $T = \sqrt{n}\bar{X}/S_X \sim t_{n-1}$

DETERMINATION BY SIMULATION:

- For $b = 1, \ldots, B$
  - generate independent replicate X^b of data X
  - compute $T^b = T(X^b)$
- Make a plot (histogram, density, ecdf) of $T^1, \ldots, T^B$.

Standard error

DATA X (possibly equal to $(X_1, \dots, X_n)$)
STATISTIC $T = T(X)$

The standard deviation of T is called *standard error* or *se*

EXAMPLE OF THEORETICAL STANDARD ERROR:
If $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$, then se of $\bar{X}$ is $\sigma/\sqrt{n}$.

DETERMINATION OF STANDARD ERROR BY SIMULATION:

- For $b = 1, \dots, B$
  - generate replicate X^b of data
  - compute $T^b = T(X^b)$
- Compute root of sample variance of $T^1, \dots, T^B$, i.e.

$$\widehat{se} = \sqrt{\frac{1}{B} \sum_{b=1}^{B} (T^b - \bar{T})^2}, \qquad \text{where } \bar{T} = \frac{1}{B} \sum_{b=1}^{B} T^b.$$

Estimators

Bias

DATA X (possibly equal to $(X_1, \ldots, X_n)$)
ESTIMATOR $T = T(X)$ OF QUANTITY θ

$\boxed{\text{The bias of } T \text{ is } \mathrm{E}T - \theta}$

EXAMPLE OF THEORETICAL BIAS:
If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$, then bias of $\bar{X}$ for θ is 0.

DETERMINATION OF BIAS BY SIMULATION:

- For $b = 1, \ldots, B$
  - generate replicate X^b of data
  - compute $T^b = T(X^b)$
- Compute

$$\widehat{bias} = \frac{1}{B} \sum_{b=1}^{B} T^b - \theta.$$

Mean square error

DATA X (possibly equal to $(X_1, \ldots, X_n)$)
ESTIMATOR $T = T(X)$ OF QUANTITY θ

The *mean square error* (MSE) of T is $\mathrm{E}(T - \theta)^2$.
(It is also the sum of the squared bias and the squared se.)

EXAMPLE OF THEORETICAL MSE:
If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$, then MSE of $\bar{X}$ for θ is σ^2/n.

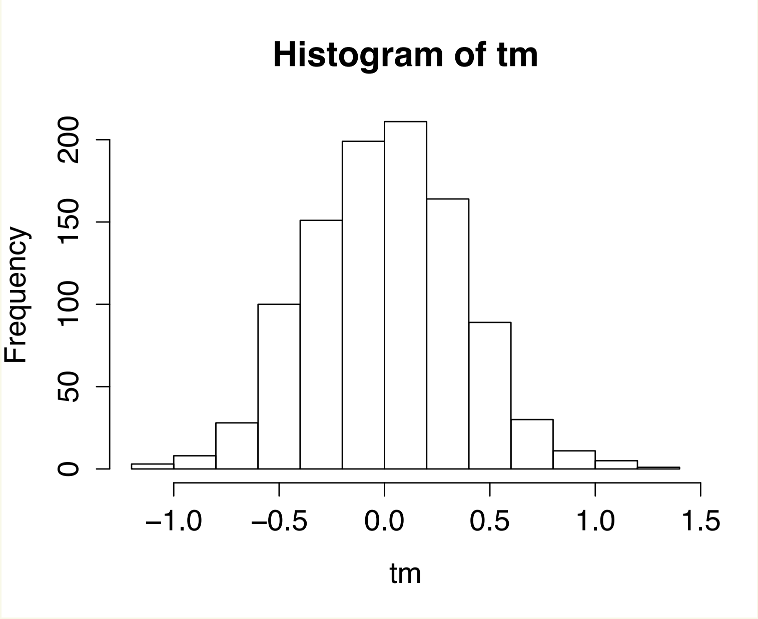
DETERMINATION OF MSE BY SIMULATION:

- For $b = 1, \ldots, B$
  - generate replicate X^b of data
  - compute $T^b = T(X^b)$
- Compute mean sum of squares

$$\widehat{MSE} = \frac{1}{B}\sum_{b=1}^{B}(T^b - \theta)^2.$$

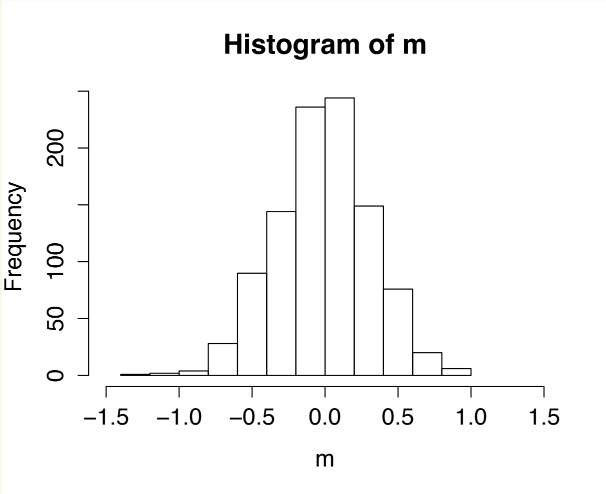
R

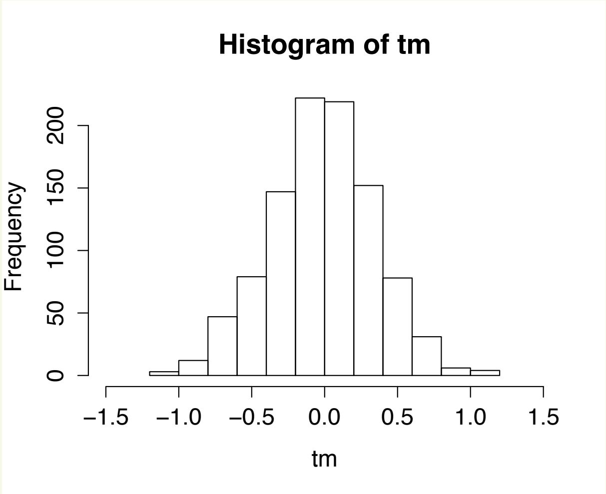
```
> B=1000
> tm = numeric(B)
> for (i in 1: B){ x=rnorm(10)
+                   tm[i]=mean(x,trim=0.3)}
> hist(tm)
> sd(tm)
[1] 0.3638536
> mean(tm)-0
[1] 0.00902773
> mean((tm-0)^2)
[1] 0.1323385
```



R

```
> B=1000
> m=tm = numeric(B)
> for (i in 1: B){ x=rnorm(10)
+                  m[i]=mean(x)
+                  tm[i]=mean(x,trim=0.3)}
> hist(m,xlim=c(-1.5,1.5)); hist(tm,xlim=c(-1.5,1.5))
> sd(m); sd(tm)
[1] 0.3222936
[1] 0.3566153
> mean(m)-0; mean(tm)-0
[1] -0.009780955
[1] -0.01743183
> mean((m-0)^2); mean((tm-0)^2)
[1] 0.103865
[1] 0.1273511
```





Simulation error

By making B larger the simulation error can be made arbitrarily small.

$B = 10000$ is desirable, but $B = 1000$ typical, and only $B = 100$ may be feasible.

> If $\mathrm{E}Z$ is estimated by $\bar{Z}_B$ for iid $Z^1, \ldots, Z^B$, then
>
> $$\bar{Z}_B - \mathrm{E}Z \sim\approx N(0, \operatorname{var} Z/B).$$
>
> Thus the *se of simulation* is $\sqrt{\operatorname{var} Z/B}$, and can be estimated by
>
> $$\sqrt{B^{-1}\sum_{b=1}^{B}(Z^b - \bar{Z})^2/B}.$$

EXAMPLE: If the MSE $\mathrm{E}(T-\theta)^2$ is estimated by $\widehat{MSE} = B^{-1}\sum_{b=1}^{B}(T^b-\theta)^2$, then the *se of simulation* can be estimated by

$$\frac{1}{B}\sqrt{\sum_{b=1}^{B}\left[(T^b-\theta)^2 - \widehat{MSE}\right]^2}.$$

R

```
> B=100
> tm = numeric(B)
> for (i in 1: B){ x=rnorm(10)
+                  tm[i]=mean(x,trim=0.3)}
> mean((tm-0)^2)
[1] 0.1603573
> sqrt(sum(((tm-0)^2-mean((tm-0)^2))^2))/B # se of simulation error in MSE
[1] 0.02689583
>
> B=1000
> tm = numeric(B)
> for (i in 1: B){ x=rnorm(10)
+                  tm[i]=mean(x,trim=0.3)}
> mean((tm-0)^2)
[1] 0.125947
> sqrt(sum(((tm-0)^2-mean((tm-0)^2))^2))/B # se of simulation error in MSE
[1] 0.005799775
```

Tests

Size

DATA X (possibly equal to $(X_1, \ldots, X_n)$)
TEST STATISTIC $T = T(X)$, REJECTS H_0 IF $T \in K$.

> The *size* of the test is $\alpha = \mathrm{P}_{H_0}(T \in K)$.

EXAMPLE OF THEORETICAL SIZE
Tests are constructed so that the size equals the *level*, e.g. 5%, in prescribed situation.

DETERMINATION OF SIZE BY SIMULATION:

- For $b = 1, \ldots, B$
  - generate replicate X^b of data *using its null distribution*
  - compute $T^b = T(X^b)$
- Compute

$$\hat{\alpha} = \frac{1}{B} \sum_{b=1}^{B} 1_{T^b \in K} = \text{ fraction rejections.}$$

[If H_0 is composite, must simulate using the worst case null distribution, or repeatedly simulate using all null distributions and take the maximum of the simulated $\hat{\alpha}$.]

Power

DATA X (possibly equal to $(X_1, \ldots, X_n)$)
TEST STATISTIC $T = T(X)$ THAT REJECTS H_0 IF IT FALLS IN CRITICAL REGION K

The *power* is $\mathrm{P}_\theta(T \in K)$ viewed as function of the alternative $\theta \in H_1$.

EXAMPLE OF THEORETICAL POWER
Power of the t-test can be expressed using non-central t-distributions.

DETERMINATION OF POWER BY SIMULATION:

- For $b = 1, \ldots, B$
  - generate replicate X^b of data *using alternative* θ
  - compute $T^b = T(X^b)$
- Compute

$$\frac{1}{B}\sum_{b=1}^{B} 1_{T^b \in K}.$$

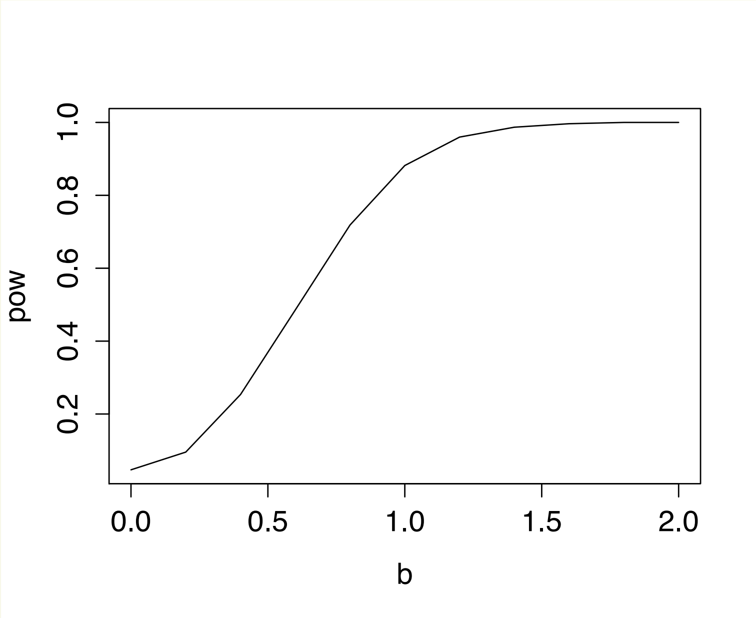
- Repeat for all alternatives.

R

```
> tmtest=function(x){n=length(x)
+                    k=trunc(0.3*n)
+                    y=sort(x)[(k+1):(n-k)]
+                    mean(y)/sd(y)}
>
> B=10000
> s=numeric(B)
> for (i in 1:B) s[i]=tmtest(rnorm(20))
> CV=quantile(s,.975)        # determine critical value by simulation
>
> a=numeric(1000)
> B=10000
> for (i in 1:B){x=rnorm(20)
+                a[i]=(abs(tmtest(x))>CV)}
> mean(a)
[1] 0.0478
```

R

```
> b=seq(0,2,by=0.2)
> pow=numeric(length(b))
> for (j in 1:length(b)){
+   t=numeric(B)
+   for (i in 1:B){x=rnorm(20,b[j],1)
+                 t[i]=tmtest(x)}
+   pow[j]=mean(abs(t)>CV)}
> plot(b,pow,type="l")
```



[Is it useful to resimulate new normal samples for every b?]

Coverage

DATA X (possibly equal to $(X_1, \ldots, X_n)$)
CONFIDENCE INTERVAL $[L(X), U(X)]$ FOR PARAMETER θ

The *coverage* is $\inf_\theta \mathrm{P}_\theta\big(L(X) \leq \theta \leq U(X)\big)$

EXAMPLE OF THEORETICAL COVERAGE
Theoretical confidence intervals have given coverage, e.g. 95%, in the prescribed situation.

DETERMINATION OF CONFIDENCE LEVEL BY SIMULATION:

- For $b = 1, \ldots, B$
  - generate replicate X^b of data *using parameter* θ
  - compute $L^b = L(X^b)$, $U^b = U(X^b)$,
- Compute
$$\hat{C}_\theta = \frac{1}{B} \sum_{b=1}^{B} 1_{L^b \leq \theta \leq U^b}.$$
- Repeat for all θ. Compute $\min_\theta \hat{C}_\theta$.

[In practice you cannot use *all* θ.]

Simulation error

By making B larger the simulation error can be made arbitrarily small.

$B = 10000$ is desirable, but $B = 1000$ typical, and only $B = 100$ may be feasible.

When estimating a proportion $p = \mathrm{P}(Z = 1)$ by a sample fraction $\bar{Z}$, for iid $Z_1, \ldots, Z_B \in \{0, 1\}$,

$$\bar{Z} - p \sim\approx N\big(0, p(1-p)/B\big).$$

The *se of simulation* can be estimated by

$$\sqrt{\bar{Z}(1-\bar{Z})/B} \leq \sqrt{1/(4B)}.$$

Permutation Tests

Permutation test

IDEA
Take any reasonable test statistic.
Compare its observed value to the set of values obtained by applying the statistic after *permuting* the data in such a way that the *distribution under H_0 does not change*.

Advantages: very flexible, no theory needed, correct level guaranteed
Disadvantage: computationally very expensive.

The *level is guaranteed*, because:

- conditionally, given any observed data, we reject with probability less than, say 5 %,
- hence unconditionally, we reject with probability less than 5 %.

Two sample test

DATA $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ independent random samples from F and G.
TEST STATISTIC Reject H_0 if $T = T(X_1, \dots, X_m, Y_1, \dots, Y_n)$ is large.

Let $Z_1, \dots, Z_N$ be the *pooled sample* $X_1, \dots, X_m, Y_1, \dots, Y_n$, i.e. $N = m + n$.

Under H_0: $F = G$ this is i.i.d. and so is every permutation $Z_{\pi(1)}, \dots, Z_{\pi(N)}$.

PERMUTATION TEST

- For every partition b of the observed values $z_1, \dots, z_N$ into 2 sets of sizes m and n compute T^b with the X's and Y's taken equal to the two sets.
- For $tobs = T(x_1, \dots, x_m, y_1, \dots, y_n)$ the observed value, compute

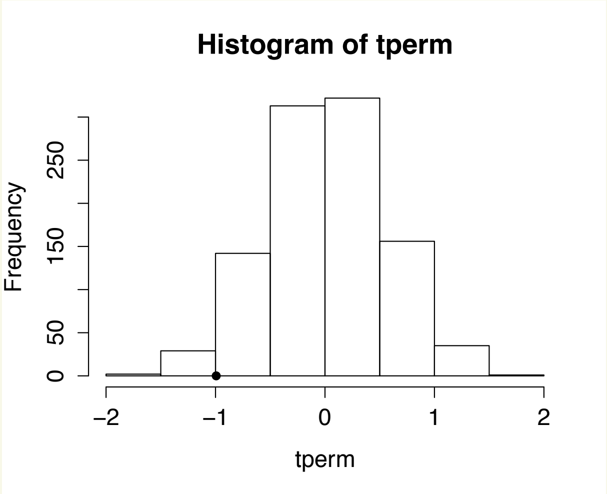
$$\hat{p} = \frac{1}{B} \sum_{b=1}^{B} 1_{T^b > tobs}.$$

IN PRACTICE
The number of partitions $\binom{N}{m}$ is too large and one takes a large number of *random partitions* instead.

[The observations may be multivariate. E.g. in genomics, of dimension 20000.]

R

```
> mu=0.5
> x=rnorm(20); y=rnorm(25,mu,2)
> tm2=function(x,y) mean(x,trim=0.2)-mean(y,trim=0.2)
> tobs=tm2(x,y)
> B=1000; tperm=numeric(B)
> z=c(x,y); N=length(z); m=length(x)
> for (i in 1:B) {xco=sample(1:N,m)
+                 tperm[i]=tm2(z[xco],z[-xco])}
> mean(tperm>tobs); mean(tperm<tobs)
[1] 0.969
[1] 0.031
>
> tobs
[1] -0.9936397
> hist(tperm); points(tobs,0,pc=19)
```



Paired two-sample test

DATA $(X_1, Y_1), \ldots, (X_n, Y_n)$ random sample from bivariate distribution.
TEST STATISTIC Reject H_0 if $T = T(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ is large.

Under H_0: (X, Y) is exchangeable, every of the 2^n possible permutations within the pairs (X_i, Y_i) gives the same distribution.

PERMUTATION TEST

- For every of the 2^n possible (re)assignments b of the X or Y -label within the pairs compute T^b.
- For $t = T(x_1, \ldots, x_m, y_1, \ldots, y_n)$ the observed value, compute

$$\hat{p} = \frac{1}{B} \sum_{b=1}^{B} 1_{T^b > t}.$$

IN PRACTICE
The number of (re)assignments 2^n is too large and one takes a large number of *random assignments* instead.

Symmetry

DATA Random sample $X_1, \ldots, X_n$ of univariate distribution.

Under the hypothesis H_0 that the distribution is symmetric about 0
$S_1 X_1, \ldots, S_n X_n$ and $X_1, \ldots, X_n$ have the same distribution for $S_1, \ldots, S_n$ random signs.

PERMUTATION TEST

- For every of the 2^n possible sign vectors compute $T^b = T(S_1 X_1, \ldots, S_n X_n)$.
- For $t = T(x_1, \ldots, x_m, y_1, \ldots, y_n)$ the observed value, compute

$$\hat{p} = \frac{1}{B} \sum_{b=1}^{B} 1_{T^b > t}.$$

IN PRACTICE
The number of sign vectors is too large and one takes a large number of *random sign vectors* instead.

When you do simulations

- Good documentation is essential
- Write separate programs for each case
- Or keep a precise record
- Save as much of the output as you can
- Do that in a structured way (multidimensional arrays)
- Summarize later
- Make programs “re-startable”
- To continue smoothly after a computer crash
- Or to divide the work over more computers