Simulación Estocástica

Métodos de Monte Carlo en problemas de inferencia estadística

Vanda Inácio de Carvalho

Primer Semestre 2015

- Las propiedades de los métodos estadísticos deben establecerse de forma que los métodos se puedan utilizar con confianza.
- Pocas o raras veces se pueden hacer derivaciones analíticas exactas de tales propiedades.
- Aproximaciones para muestras grandes son a menudo posibles, pero, sin embargo, es necesária la evaluación de la calidad de la aproximación para tamaños de muestras (finitos) encuentrados en la práctica.
- Además, los resultados analíticos pueden requerir supuestos (p.e., normalidad).
- Pero, qué sucede cuando se violan estos supuestos? Resultados analíticos, mismo los para grandes muestras, pueden no ser posibles.

- Es el estimador no sesgado para muestras finitas? Es aún consistente cuando los supuestos son violados? Cual es su varianza muestral?

- Alcanzará un test de hipótesis el nível de significación especificado?

- Como responder a estas preguntas en la ausencia de resultados analíticos?

- Un típico estudio de simulación de Monte Carlo implica el siguiente
  - Generar m conjuntos de datos independientes bajo las condiciones de interés.
  - Calcular el valor numerico del estimador/estadístico del test, llamésmole T, para cada conjunto de datos $\Rightarrow T_1, \ldots, T_m$.
  - Si m es grande, estadísticas sumárias a lo largo de $T_1, \ldots, T_m$ deberían ser buenas aproximaciones de las verdaderas propiedades del estimador.

- Comparar tres estimadores para la media μ de una determinada distribución con base en una muestra aleatoria $X_1, \ldots, X_n$.

  1. Media muestral, $T^{(1)}$.
  2. Media muestral recortada en 20%, $T^{(2)}$.
  3. Mediana muestral, $T^{(3)}$.

- Sea $X_1, \ldots, X_n$ una muestra aleatoria y sea $X_{(1)}, \ldots, X_{(n)}$ la muestra ordenada correspondiente.
- La media recortada (*trimmed mean*) es calculada descartando los valores más pequeños/grandes (en general un porcentaje especificado) de la muestra.
- Sea entonces p el porcentaje especificado y sea $k = np/100$ el número de observaciones a descartar.
- Descartamos así las k mayores y las k más pequeñas observaciones.
- Media recortada
$$\frac{X_{k+1} + X_{k+2} + \ldots + X_{n-k}}{n - 2k}$$
- La media recortada es menos sensible a outliers que la media.

- La media recortada es bastante simple de se programar.

```
trimmean=function(x,p){
n=length(x); x=sort(x); k=n*p/100
trimmean=sum(x[(k+1):(n-k)])/(n-2*k)
return(trimmean)
}

set.seed(123); x=rnorm(20)
trimmean(x,20)
0.08555956
mean(x,0.2)
0.08555956
```

- El procedimiento de simulación es el siguiente. Para cada elección particular de μ, n y verdadera distribución subyacente:
  - Generar $X_1, \ldots, X_n$ de la distribución elegida.
  - Calcular $T^{(1)}$, $T^{(2)}$ y $T^{(3)}$.
  - Repetir m veces, obteniendo

$$T_1^{(1)}, \ldots, T_m^{(1)}; \quad T_1^{(2)}, \ldots, T_m^{(2)}; T_1^{(3)}, \ldots, T_m^{(3)}.$$

- Para $k = 1, 2, 3$, calcular

$$\hat{\mu}_k = \frac{1}{m} \sum_{i=1}^{m} T_i^{(k)} = \overline{T}^{(k)}, \quad \widehat{se}(\hat{\mu}_k) = \sqrt{\frac{1}{m-1} \sum_{i=1}^{m} \left(T_i^{(k)} - \overline{T}^{(k)} \right)^2},$$

$$\widehat{\text{sesgo}}(\hat{\mu}_k) = \overline{T}^{(k)} - \mu, \quad \widehat{\text{ecm}}(\hat{\mu}_k) = \frac{1}{m} \sum_{i=1}^{m} \left(T_i^{(k)} - \mu \right)^2.$$

- Simularemos 20 observaciones de una distribución normal com media 0 y varianza 1. Repetiremos el procedimiento 1000 veces.

```
set.seed(123)
n=20; mu=0;sigma=1;m=1000
media=tmean=mediana=numeric(m)
for(i in 1:m){
x=rnorm(n,mu,sigma)
media[i]=mean(x)
tmean[i]=trimmean(x,20) #mean(x,0.2)
mediana[i]=median(x) #quantile(x,0.5)
}

mean(media); mean(tmean); mean(mediana)
ecmmedia=sum((media-mu)**2)/m
ecmtmean=sum((tmean-mu)**2)/m
ecmmediana=sum((mediana-mu)**2)/m

c(ecmmedia,ecmtmean,ecmmediana)
0.04823896 0.05376052 0.06797592
```

- Contaminemos ahora la muestra generada de la siguiente manera: $X \sim \mathrm{N}(n-3,0,1)+\mathrm{N}(3,0,100)$. Repetimos el procedimiento anterior.

```
set.seed(123)
n=20; mu=0;sigma=1;m=1000
media=tmean=mediana=numeric(m)
for(i in 1:m){
x=c(rnorm(n-3,mu,sigma),rnorm(3,0,100))
media[i]=mean(x)
tmean[i]=trimmean(x,20)
mediana[i]=median(x)
}

print(c(ecmmedia,ecmtmean,ecmmediana))
73.98882261 0.08327559 0.09341217
```

- Obviamente, la robustez de la media recortada depende del tamaño de la muestra, del número de observaciones contaminadas en la muestra (outliers) y del porcentaje de observaciones que recortamos.

- En este ejemplo vamos a hacer un estudio de simulación de Monte Carlo para evaluar la cobertura (coverage) de intervalos de confianza.

- Sabemos que un intervalo con un nivel de confianza $(1-\alpha)\%$ para un parámetro desconocido μ es de la forma

$$\Pr(\hat{\mu}_L(\mathbf{X}) < \mu < \hat{\mu}_U(\mathbf{X})) = 1-\alpha,$$

- Una vez observada una muestra concreta $\mathbf{X} = \mathbf{x}$, el intervalo deja de ser aleatorio e ya no hace sentido escribir

$$\Pr(\hat{\mu}_L(\mathbf{X}) < \mu < \hat{\mu}_U(\mathbf{X})) = 1-\alpha,$$

  una vez que, observada la muestra, esta probabilidad es cero o uno y no $1-\alpha$.

- Lo que se puede decir es que el intervalo $(\hat{\mu}_L(\mathbf{X}), \hat{\mu}_U(\mathbf{X}))$ contiene μ, $(1-\alpha)\times 100\%$ de las veces.

- Si μ representa la media poblacional, la cual estimamos usando la media muestral $\bar{x}$, entonces sabemos que un intervalo de confianza de $(1-\alpha) \times 100\%$ para μ es

$$\left(\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right),$$

 donde s es el error estándar muestral, dada por

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- Para ver si la cobertura del intervalo es de hecho $1-\alpha$, debemos generar varias (muchas!) muestras y determinar la proporción de veces que el intervalo contiene μ.

```
set.seed(123)
n=20; mu=0; sigma=1; alpha=0.05; t05=qt(1-alpha/2,n-1);nsim=1000
ci.lower=ci.upper=numeric(nsim)
for(i in 1:nsim){
x=rnorm(n,mu,sigma)
m=mean(x); s=sd(x)
ci.lower[i]=m-t05*s/sqrt(n); ci.upper[i]=m+t05*s/sqrt(n)
}
coverage=mean(ci.lower<=mu & ci.upper>=mu)
```

- Vamos ahora estudiar métodos de Monte Carlo para pruebas de hipótesis.
- Supongamos que queremos probar una hipótesis relativa a un determinado parámetro, llamésmole θ
$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0 \quad (\text{o} \quad \theta > \theta_0, \quad \text{o} \quad \theta < \theta_0).$$
- Dos tipos de errores pueden ocurrir
  - Error tipo I: rechazar H_0 cuando H_0 es verdadera.
    La probabilidade de error tipo I es
$$\alpha = \text{Pr(error tipo I)} = \text{Pr(rechazar } H_0 \mid H_0 \text{ verdadera)},$$
    y también se llama nivel de significación de la prueba.
  - Error tipo II: no rechazar H_0 cuando H_0 no es verdadera.
    La probabilidade de error tipo II es
$$\beta = \text{Pr(error tipo II)} = \text{Pr(no rechazar } H_0 \mid H_0 \text{ no es verdadera)},$$
    y a $1 - \beta$ se llama el poder de la prueba.

- Una prueba de hipótesis, con un nível de significación $\alpha = 0.05$, para un determinado parámetro, no tiene realmente una tasa de error de tipo I de 0.05 la mayoria de las veces.

- Esto se puede comprobar mediante un estudio de Monte Carlo. Los pasos del proceso de simulación son los siguientes:
    - simular m conjuntos de datos bajo las condiciones de la hipótesis nula;
    - determinar si cada p-valor es menor que el nível especificado α;
    - estimar el verdadero nível de significación por $\frac{1}{m}\sum_{i=1}^{m} I(t_i, \theta)$, donde $I(t_i, \theta) = 1$ si el valor-p es inferior a α (o sea si rechazamos H_0) para el i-ésimo conjunto de datos y $I(t_i, \theta) = 0$ en otro caso.

- Suponga que $X_1, \ldots, X_{20}$ es una muestra aleatoria de una distribución $\mathrm{N}(\mu, \sigma^2)$. Probar

$$H_0 : \mu = 500 \text{ vs } H_1 : \mu > 500, \quad \text{para} \quad \alpha = 0.05.$$

- La estadística del test es

$$T = \frac{\bar{X} - 500}{s/\sqrt{n}} \sim t_{n-1},$$

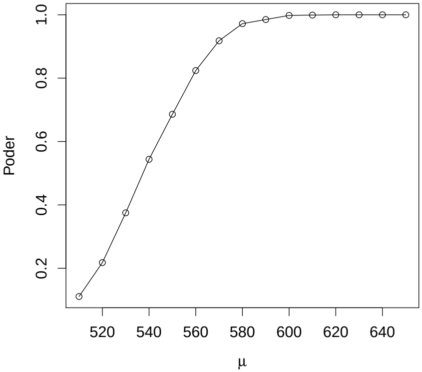
con s^2 la varianza muestral.

- Valores elevados de T apoyan la hipótesis alternativa.

```
nsim=1000; n=20; alpha=0.05; mu0=500; sigma=100
set.seed(123); p=numeric(nsim)
for(i in 1:nsim){
x=rnorm(n,mu0,sigma)
Ttest=(mean(x)-mu0)/(sd(x)/sqrt(n))
p[i]=1-pt(Ttest,n-1) #p[i]=t.test(x,alernative="greater",mu=mu0)$p.value
}
phat=mean(p<alpha)
```

- Para estudiar el poder (probabilidad de rechazar H_0 cuando H_0 no es verdadera) de una prueba de hipótesis, simplemente hay que repetir el procedimiento anterior pero ahora simulando datos bajo las condiciones de la hipótesis alternativa.

```
nsim=1000; n=20; mu0=500; sigma=100; mu=seq(510,650,by=10);
nmu=length(mu)
poder=numeric(nmu)
for(i in 1:nmu){
p=replicate(nsim,expr={x=rnorm(n,mu[i],sigma);
Ttest=t.test(x,alternative="greater",mu=mu0); Ttest$p.value})
poder[i]=mean(p<0.05)
}
```


Poder
0.2
0.4
0.6
0.8
1.0
520
540
560
580
600
620
640
μ