Chapter 6: Monte Carlo Methods in Inference

Lecturer: Zhao Jianhua

Department of Statistics
Yunnan University of Finance and Economics

Outline

Introduction

Monte Carlo (MC) methods may refer to any method in statistical inference or numerical analysis where simulation is used. MC methods encompass a vast set of computational tools in modern applied statistics. MC methods can be applied to

- estimate parameters of the sampling dist. of a statistic, mean squared error(MSE), percentiles, or other quantities of interest.
- assess the coverage probability for confidence intervals, an empirical Type I error rate of a test procedure;
- estimate the power of a test;
- compare the performance of different procedures for a given problem.

The methods covered in this chapter use repeated sampling from a given probability model, sometimes called parametric bootstrap, to investigate this uncertainty.

MC Methods for Estimation

Suppose $X_1, \ldots, X_n$ is a random sample from the X dist.

- An estimator $\hat{\theta}$ for a parameter θ is an n variate function of the sample.

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

- Functions of the estimator $\hat{\theta}$ are therefore n-variate functions of the data.

For simplicity, let $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, and let $x_{(1)}, x_{(2)}, \ldots,$ denote a sequence of independent random samples generated from the dist. of X.

- Random variates from the sampling dist. of $\hat{\theta}$ can be generated by repeatedly drawing independent random samples $x^{(j)}$ and computing $\hat{\theta}^{(j)} = \hat{\theta}(x_1^{(j)}, \ldots, x_n^{(j)})$ for each sample.

Example 6.1(Basic MC estimation)

Suppose thatX_1, X_2 are iid from a standard normal dist. Estimate the mean difference $E|x_1 - x_2|$.

To obtain a MC estimate of $\theta = E[g(X_1, X_2)] = E|X_1 - X_2|$ based on m replicates, generate random samples $x^{(j)} = (x_1^{(j)}, x_2^{(j)})$ of size 2 from the standard normal dist., $j = 1, \ldots, m$. Then compute the replicates $\hat{\theta}^{(j)} = g_j(x_1, x_2) = |x_1^{(j)} - x_2^{(j)}|$, $j = 1, \ldots, m$, and the mean of the replicates

$$\theta = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}^{(j)} = \overline{g(X_1, X_2)} = \frac{1}{m} \sum_{i=1}^{m} |x_1^{(j)} - x_2^{(j)}|$$

```
m < - 1000; g < - numeric(m)
for (i in 1 : m) {x<-rnorm(2); g[i]<-abs(x[1]-x[2])}
est < - mean(g); est
[1]1.128402
```

By integration that $E|X_1 - X_2| = 2/\sqrt{\pi} = 1.128379$ and $Var(|X_1 - X_2|) = 2 - 4/\pi$. Thus the standard error of $\hat{\theta}$ $\sqrt{(2 - 4/\pi)/m} = 0.02695850$.

The standard error of a mean $\overline{x}$ of a sample size n is $\sqrt{var(x)/n}$. When the dist. of X is unknown we can substitute for F the empirical dist. F_n of the sample $x_1, \ldots, x_n$. The plug-in estimate of the variance of X is

$$\widehat{var}(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$\widehat{var}(x)$ is the population variance of the finite pseudo population $x_1, \ldots, x_n$ with cdf F_n. The corresponding estimate of the standard error of $\overline{x}$ is

$$\widehat{se}(\overline{x}) = \frac{1}{\sqrt{n}} \{ \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \}^{1/2} = \frac{1}{n} \{ \sum_{i=1}^{n} (x_i - \overline{x})^2 \}^{1/2}$$

Using the unbiased estimator of $Var(X)$ we have

$$\widehat{se}(\overline{x}) = \frac{1}{\sqrt{n}} \{ \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \}^{1/2}$$

In Example 6.1 (sample size m), estimate of standard error of $\hat{\theta}$ is

```
sqrt(sum((g-mean(g))^2))/m
#sd(g)/sqrt(m) #for unbiased estimator
```

Estimation of MSE

MC methods can be applied to estimate the MSE of an estimator. Recall that the MSE of an estimator $\hat{\theta}$ for a parameter θ is defined by MSE $(\hat{\theta}) = \mathrm{E}[(\hat{\theta}-\theta)^2]$. If m (pseudo) random samples$x^{(1)}, \ldots, x^{(m)}$ are generated from the dist. of X, then a MC estimate of the MSE of $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ is

$$\widehat{MSE} = \frac{1}{m} \sum_{j=1}^{m} (\hat{\theta}^{(j)} - \theta)^2$$

where $\hat{\theta}^{(j)} = \hat{\theta}(x^{(j)}) = \hat{\theta}(x_1^{(j)}, \ldots, x_2^{(j)})$.

Example 6.2 (Estimating the MSE of a trimmed mean)

A trimmed mean is sometimes applied to estimate the center of a continuous symmetric dist. that is not necessarily normal. In this example, we compute an estimate of the MSE of a trimmed mean. Suppose that $X_1, \ldots, X_n$ is a random sample and $X_{(1)}, \ldots,$ $X_{(n)}$ is the corresponding ordered sample.

The trimmed sample mean is computed by averaging all but the largest and smallest sample observations.

Generally, the kth level trimmed sample mean is defined by

$$\widehat{X}_{|(k-1)|} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}.$$

Obtain a MC estimate of the MSE ($\overline{X}_{[-1]}$) of the first level trimmed mean assuming that the sampled dist. is standard normal.

The center of the dist. is 0 and the target is $\theta = E[\overline{X}] = E[\overline{X}_{[-1]}] = 0$. Denote the first level trimmed sample mean by T. A MC estimate of $MSE(T)$ based on m replicates:

1. Generate the replicates $T^{(j)}$, $j = 1 \ldots, m$ by repeating:

(a) Generate ${x_1}^{(j)}, \ldots, {x_n}^{(j)}$, iid from the dist. of X.

(b) Sort ${x_1}^{(j)}, \ldots, {x_n}^{(j)}$ in increasing order, to obtain ${x_{(1)}}^{(j)} \ldots {x_{(n)}}^{(j)}$.

(c) Compute $T^{(j)} = \frac{1}{n-2} \sum_{i=2}^{n-1} x_{(i)}^{(j)}$.

2. Compute $\widehat{MSE}(T) = \frac{1}{m} \sum_{j=1}^{m} (T^{(j)} - \theta)^2 = \frac{1}{m} \sum_{j=1}^{m} (T^{(j)})^2$

Then $T_{(1)}, \ldots, T_{(m)}$ are i.i.d. , and we are computing the sample mean estimate $\widehat{MSE}(T)$ of $MSE(T)$.

```
n <- 20
m <- 1000
tmean <- numeric(m)
for (i in 1:m) {
x <- sort(rnorm(n))
tmean[i] <- sum(x[2:(n-1)]) / (n-2)
}
mse <- mean(tmean^2)
> mse
[1] 0.05176437
> sqrt(sum((tmean - mean(tmean))^2)) / m #se
[1] 0.007193428
```

The estimate of MSE in this run is approximately 0.052 ($\hat{se} = 0.007$). For comparison, the MSE of the sample mean $\overline{X}$ is $Var(X)/n$, which is $1/20 = 0.05$ in this example.

Note that the median is actually a trimmed mean; it trims all but one or two of the observations.

```
n <- 20
m <- 1000
tmean <- numeric(m)
for (i in 1 : m ) {
x <- sort (rnorm(n))
tmean[i] <- median(x)
}
mse <- mean (tmean ^ 2 )
> mse
[1] 0.07483438
> sqrt(sum((tmen-mean(tmean))^2))/m#se
[1] 0.008649554
```

The estimate of MSE for the sample median is approximately 0.075 and $\hat{se}(\widehat{MSE}) = 0.0086$.

Example 6.3 (MSE of a trimmed mean, cont.)

Compare the MSE of level-k trimmed means for the standard normal and a 'contaminated' normal dist. The contaminated normal dist. is a mixture

$$pN(0, \sigma^2 = 1) + (1 - p)N(0, \sigma^2 = 100)$$

Write a function to estimate $MSE(\overline{X}_{[-k]})$ for different k and p.

Estimates of MSE for the kth Level Trimmed Mean in Exp6.3 ($n = 20$)

	Normal		p = 0.95		p = 0.90	
k	$n\widehat{MSE}$	$n\hat{se}$	$n\widehat{MSE}$	$n\hat{se}$	$n\widehat{MSE}$	$n\hat{se}$
0	0.976	0.140	6.229	0.140	11.485	0.140
1	1.019	0.143	1.954	0.143	4.126	0.143
2	1.009	0.142	1.304	0.142	1.956	0.142
3	1.081	0.147	1.168	0.147	1.578	0.147
4	1.048	0.145	1.280	0.145	1.453	0.145
5	1.103	0.149	1.395	0.149	1.423	0.149
6	1.316	0.162	1.349	0.162	1.574	0.162
7	1.377	0.166	1.503	0.166	1.734	0.166
8	1.382	0.166	1.525	0.166	1.694	0.166
9	1.491	0.172	1.646	0.172	1.843	0.172

To generate the contaminated normal samples, first randomly select σ according to the probability dist. $P(\sigma = 1) = p; P(\sigma = 10) = 1 - p$.

`rnorm` can accept a vector of parameters for standard deviation. After generating n values for σ, pass this vector as `sd` argument to `rnorm`.

```
n<-20; K <-n/2-1; m<-1000; mse<-matrix(0,n/2,6)
trimmed.mse <- function ( n,m,k,p ) {
# Mc est of mse for k-level trimmed mean of
# contaminated normal pN (0,1) + (1-p) N (0,100)
t mean < - numeric( m )
for (i in1 : m) {
sigma <-sample(prob=c(p,1-p)); x<-sort(rnorm (n,0,sigma))
t mean [i] < - sum (x[(k+1):(n-k)])/(n-2*k)
mse.est < - mean(tmeanˆ2)
se.mse < - sqrt(mean((tmean-mean(tmean))ˆ2)) /sqrt(m)
return(c(mse.est,se.mse))
}
for ( k in 0 : K ) {
mse [k+1,1:2 ] < - trimmed.mse( n=n,m=m,k=k,p=1.0 )
mse [K+1,3:4 ] < - trimmed.mse( n=n,m=m,k=k,p=.95 )
mse [k+1,5:6 ] < - trimmed.mse( n=n,m=m,k=k,p=.9 )}
```

Estimating a confidence level

One type of problem that arises frequently in statistical applications is the need to evaluate the cdf of the sampling dist. of a statistic, when the density function of the statistic is unknown or intractable.

Many commonly used estimation procedures are derived under the assumption that the sampled population is normally distributed. In practice, it is often the case that the population is non-normal and the true dist. may be unknown or intractable.

Example: If (U, V) is a confidence interval (CI) estimate for an unknown parameter θ, then U and V are statistics with dist. that depend on the dist. F_X of the sampled population X. The confidence level (CL) is the probability that the interval (U, V) covers the true value of the parameter θ. Evaluating the CL is therefore an integration problem.

- Sample-mean MC approaches to evaluating an integral $\int g(x)dx$ do not require that $g(x)$ is specified. It is only necessary that the sample from the dist. $g(X)$ can be generated.

Example 6.4 (Confidence interval for variance)

Use MC methods to estimate the true level when the normal theory CI for variance is applied to non-normal data. If $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma_2)$ dist., $n \geq 2$, and S^2 is the sample variance, then

$$V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1). \tag{6.1}$$

A one side $100(1-\alpha)\%$ CI is $(0, (n-1)S^2/\chi^2_\alpha(n-1))$. If the sampled population is normal with variance σ^2, then the probability that the CI contains σ^2 is $1-\alpha$.

The calculation of 95% upper confidence limit (UCL) for a random sample size $n = 20$ from $N(0, \sigma^2 = 4)$:

```
n <- 20; alpha <- .05
x <- rnorm (n,mean=0,sd = 2 )
UCL <- (n-1)*var(x)/qchisq(alpha,df=n-1)
```

Several runs produce UCL $= 6.628$, 7.348, 9.621, etc. All contain $\sigma^2 = 4$. If the sampling and estimation is repeated a large number of times, approximately 95% of the intervals should contain σ^2.

Monte Carlo experiment to estimate a confidence level

- Empirical CL is an estimate of the CL obtained by simulation.
- Repeat the steps above a large number of times, and compute the proportion of intervals that contain the target parameter.

Suppose that $X \sim F_x$: r.v. of interest and θ: the target parameter.

1. For each replicate, indexed j = 1,. . . , m:
   - **(a)** Generate the j^{th} random sample, ${X_1}^{(j)}, \ldots, {X_n}^{(j)}$.
   - **(b)** Compute the CI C_j for the j^{th} sample.
   - **(c)** Compute $y_j = I(\theta \in C_j)$ for the j^{th} sample.
2. Compute the empirical CL $\bar{y}=\frac{1}{m}\sum_{j=1}^{m} y_j$.

The estimator $\bar{y}$ is a sample proportion estimating the true CL $1-\alpha^*$, so $Var(\bar{y}) = (1-\alpha^*)\alpha^*/m$ and an estimate of standard error is $\hat{se}(\bar{y}) = \sqrt{(1-\bar{y})\bar{y}/m}$.

Example 6.5 (MC estimate of confidence level)

Refer to Example 6.4. Here, $\mu = 0, \sigma = 2, n = 20, m = 1000$ replicates, and $\alpha = 0.05$. The sample proportion of intervals that contain $\sigma^2 = 4$ is a MC estimate. This type of simulation can be conveniently implemented by using `replicate` function.

```
> round(rbind(table(x)/n, p, se),3)
n <- 20; alpha <- .05
UCL <- replicate(1000, expr = {
x <- rnorm(n, mean = 0, sd = 2)
(n-1) * var(x) / qchisq(alpha, df = n-1)} )
sum(UCL > 4) # count the number of intervals that contain s
> mean (UCL > 4 ) # or compute the mean to get the CL
[1] 0.956
```

The result is that 956 intervals satisfied (UCL > 4), so the empirical confidence level is 95.6%. The result will vary but should be close to the theoretical value, 95%. The standard error of the estimate is $(0.95(1 - 0.95)/1000)^{1/2} \doteq 0.00689$.

R note 6.1

In `replicate` function, the lines to be repeatedly executed are enclosed in braces { }. Alternately, the expression argument (expr) can be a function call:

```
> round(rbind(table(x)/n, p, se),3)
calcCI <- function(n, alpha) {
y <- rnorm(n, mean = 0, sd = 2)
return((n-1) * var(y) / qchisq(alpha, df = n-1))}
UCL<-replicate(1000,expr=calcCI(n=20,alpha=.05))
```

The interval estimation procedure based on (6.1) for estimating variance is sensitive to departures from normality, so the true CL may be different from the stated CL when data are non-normal.

The true CL depends on the cdf of S^2. The CL is the probability that the interval $(0, (n-1)S^2/\chi^2_\alpha)$ contains the true value of σ^2,

$$P(\frac{(n-1)S^2}{{\chi_\alpha}^2} > \sigma^2) = P(S^2 > \frac{\sigma^2{\chi_\alpha}^2}{n-1}) = 1 - G(\frac{\sigma^2{\chi_\alpha}^2}{n-1})$$

where $G(\cdot)$ is the cdf of S^2.

If the sampled population is non-normal, we have the problem of estimating the cdf

$$G(t) = P(S^2 \leq c_\alpha) = \int_0^{c_\alpha} g(x)dx,$$

where $g(x)$ is the (unknown) density of S^2 and $c_\alpha = \sigma^2 {\chi_\alpha}^2/(n-1)$. An approximate solution can be computed empirically using Monte Carlo integration to estimate $G(c_\alpha)$. The estimate of $G(t) = P(S^2 \leq t) = \int_0^t g(x)dx$, is computed by MC integration. It is not necessary to have an explicit formula for $g(x)$, provided that we can sample from the dist. of $g(X)$.

Example 6.6 (Empirical confidence level)

In Example 6.4, what happens if the sampled population is non-normal? For example, suppose that the sampled population is $\chi^2_{(2)}$, which has variance 4, but is distinctly no n-normal. We repeat the simulation, replacing the $N(0,4)$ samples with $\chi^2_{(2)}$ samples.

```
n <- 20; alpha <- .05
UCL <- replicate (1000,expr= {
x <- rchisq(n,df=2 )
(n-1)*var(x)/qchisq(alpha,df=n-1)})
> sum (UCL > 4); mean( UCL > 4)
[1] 773
[1] 0.773
```

In this experiment, only 773 or 77.3% of the intervals contained the population variance, which is far from the 95% coverage under normality.

Remark 6.1

The MC approach here is sometimes called parametric bootstrap. In 'parametric' bootstrap, the pseudo random samples are generated from a given probability dist. In the 'ordinary' bootstrap, the samples are generated by resampling from an observed sample. methods.

Monto Carlo Methods for Hypothesis Tests

Suppose that we wish to test a hypothesis concerning a parameter hat lies in a parameter space Θ. The hypotheses of interest are

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta \in \Theta_1$$

where $\Theta = \Theta_0 \cup \Theta_1$. Two types of error can occur:

- Type I error: if the null hypothesis is rejected when in fact the null hypothesis is true.
- Type II error: if the null hypothesis is not rejected when in fact the null hypothesis is false.

The prob. of rejecting the null hypothesis depends on the true value of θ, denoted by $\pi(\theta)$. The significance level of a test is α, which is an upper bound on the prob. of Type I error, i.e.,

$$\alpha = \sup_{\theta \in \Theta_0} \pi(\theta)$$

Prob. of Type I error is the conditional prob. that the null hypothesis is rejected given that H_0 is true. If the test procedure is replicated a large number of times under the null hypothesis, the observed Type I error rate should be at most (approximately) α.

6.3.1 Empirical Type I error rate

An empirical Type I error rate can be computed by a MC experiment. The empirical Type I error rate is the sample proportion of significant test statistics among the replicates.

1. For each replicate, indexed by $j = 1, \ldots, m$:
   - **(a)** Generate the j^{th} random sample ${x_1}^{(j)}, \ldots, {x_n}^{(j)}$ from the null dist.
   - **(b)** Compute the test statistic T_j from the j^{th} sample.
   - **(c)** Record the test decision $I_j = 1$ if H_0 is rejected at significance level α and otherwise$I_j = 0$.
2. Compute the proportion of significant tests $\frac{1}{m}\sum_{j=1}^{m} I_j$. This proportion is the observed Type I error rate.

The observed Type I error rate, is a sample proportion. If we denote the observed Type I error rate by $\hat{p}$, then an estimate of $se(\hat{p})$ is

$$\hat{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{m}} \leq \frac{0.5}{\sqrt{m}}$$

Example 6.7 (Empirical Type I error rate)

Suppose that $X_1, \ldots, X_{20}$ is a random sample from a $N(\mu, \sigma^2)$. Test $H_0 : \mu = 500$ $H_1 : \mu > 500$ at $\alpha = 0.05$. Under H_0,

$$T^* = \frac{\overline{X} - 500}{S/\sqrt{20}} \sim t(19),$$

Use MC method to compute an empirical probability of Type I error when $\sigma = 100$, and check that it is approximately equal to $\alpha = 0.05$, basing the test decisions on the p-values returned by `t.test`.

```
n <- 20; alpha <- .05; mu0 <- 500
sigma <- 100; m <- 10000 #number of replicates
p <- numeric(m) #storage for p-values
for (j in 1:m) {
  x <- rnorm(n, mu0, sigma)
  ttest <- t.test(x, alternative = "greater", mu = mu0)
  p[j] <- ttest$p.value }
p.hat<-mean(p <alpha); se.hat<-sqrt(p.hat*(1-p.hat)/m)
print(c(p.hat, se.hat))
[1] 0.050600000 0.002191795
```

The observed Type I error rate is 0.0506, and the standard error of the estimate is approximately $\sqrt{0.05 \times 0.95/m} = .0.0022$.

Example 6.8 (Skewness test of normality)

Investigate whether a test based on the asymptotic dist. of the skewness statistic achieves the nominal significance level α under the null hypothesis of normality. The skewness $\sqrt{\beta_1}$ of a r.v. X is

$$\sqrt{\beta_1} = \frac{E[(x - \mu_x)]^3}{{\sigma_x}^3}$$

where $\mu_x = E[X]$ and ${\sigma_X}^2 = Var(X)$. A dist. is symmetric if $\sqrt{\beta_1} = 0$, positively skewed if $\sqrt{\beta_1} > 0$, and negatively skewed if $\sqrt{\beta_1} < 0$. The sample coefficient of skewness $\sqrt{b_1}$ defined as

$$\sqrt{b_1} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2)^{3/2}}$$

If the dist. of X is normal, then is $\sqrt{b_1}$ asymptotically normal with mean 0 and variance $6/n$. Normal dist.s are symmetric, and the hypothesis of normality is rejected for large values of $|\sqrt{b_1}|$.

The hypotheses are

$$H_0 : \sqrt{\beta_1} = 0; \; H_1 : \sqrt{\beta_1} \neq 0.$$

where the sampling dist. of the skewness statistic is derived under the assumption of normality.

- However, the convergence of $\sqrt{\beta_1}$ to its limit dist. is rather slow and the asymptotic dist. is not a good approximation for small to moderate sample sizes.

Assess the Type I error rate for a skewness test of normality at $\alpha = 0.05$ based on the asymptotic dist. of $\sqrt{\beta_1}$ for sample sizes n $= 10$, 20, 30, 50, 100, and 500. The vector of critical values cv for each n are computed under normal:

```
n <- c(10, 20, 30, 50, 100, 500) #sample sizes
cv <- qnorm(.975, 0, sqrt(6/n)) #crit. values for each n
asymptotic critical values:
n   10      20      30      50      100     500
cv 1.5182 1.0735 0.8765 0.6790 0.4801 0.2147
```

The asymptotic dist. of $\sqrt{b_1}$ does not depend on the mean and variance of the sampled normal dist., so the samples can be generated from the standard normal dist. If the sample size is `n[i]` then H_0 is rejected if $\sqrt{b1} >$ cv[i].

First write a function to compute the sample skewness statistic.

```
sk <- function(x) {
  #computes the sample skewness coeff.
  xbar <- mean(x); m3 <- mean((x - xbar)^3)
  m2 <- mean((x - xbar)^2); return( m3 / m2^1.5 )}
```

```
#n is a vector of sample sizes
#we are doing length(n) different simulations
p.reject <- numeric(length(n)) #to store sim. results
m <- 10000 #num. repl. each sim.
for (i in 1:length(n)) {sktests<-numeric(m) #test decisions
  for (j in 1:m) { x <- rnorm(n[i])
    #test decision is 1 (reject) or 0
    sktests[j] <- as.integer(abs(sk(x))>= cv[i] )}
    p.reject[i] <- mean(sktests) #proportion rejected}
> p.reject
[1] 0.0129 0.0272 0.0339 0.0415 0.0464 0.0539
```

With $m = 10000$ replicates the standard error of the estimate is approximately $\sqrt{0.05 \times 0.95/m} = 0.0022$. The results of the simulation suggest that the asymptotic normal approximation for of $\sqrt{b_1}$ is not adequate for $n \leq 50$, and questionable as large as $n = 500$. For finite samples one should use

$$Var(\sqrt{b_1}) = \frac{6(n-2)}{(n+1)(n+3)}$$

Repeating the simulation with

```
cv < - qnorm (.975,0,sqrt(6*(n-2)/((n+1)*(n+3))))
> round (cv,4)
[1]1.1355 0.9268 0.7943 0.6398 0.4660 0.2134
n              10     20      30      50     100    500
estimate 0.0548 0.0515 0.0543 0.0514 0.0511 0.0479
```

These estimates are closer to the nominal level $\alpha = 0.05$.

6.3.2 Power of a Test

The power of a test is the power function $\pi : \Theta \to [0, 1]$, which is the prob. $\pi(\theta)$ of rejecting H_0 given that the true value of the parameter is θ. Thus, for a given $\theta \in \Theta_1$, the prob. of Type II error is $1 - \pi(\theta_1)$. Ideally, we would prefer a test with low prob. of error. Type I error is controlled by the significance level α. Thus,

- when comparing tests for the same hypotheses at same significance level, we are interested in comparing power of the tests.

In general the comparison is not one problem but many; the power $\pi(\theta_1)$ of a test under the alternative hypothesis depends on the particular value of the alternative θ_1.

If the power function of a test cannot be derived analytically, the power of a test against a fixed alternative $\theta \in \Theta_1$ can be estimated by MC methods. The power function is defined for all $\theta \in \Theta$, but the significance level α controls $\pi(\theta) \leq \alpha$ for all $\theta \in \Theta_0$.

MC estimate of the power of a test against a fixed alternative

1. Select a particular value of the parameter $\theta \in \Theta$.
2. For each replicate, indexed by $j = 1, \ldots, m$:
   - **(a)** Generate the j^{th} random sample ${x_1}^j, \ldots, {x_n}^j$ under the conditions of the alternative $\theta = \theta_1$.
   - **(b)** Compute the test statistic T_j from the j^{th} sample.
   - **(c)** Record the test decision: set $I_j = 1$ if H_0 is rejected at significance levelα, and otherwise set $I_j = 0$.
3. Compute the proportion of significant tests $\hat{\pi}(\theta_1) = \frac{1}{m} \sum_{j=1}^{m} I_j$.

Example 6.9 (Empirical power)

Use simulation to estimate power and plot an empirical power curve for the t-test in Example 6.7. (For a numerical approach that does not involve simulation, see the remark 6.2.)

To plot the curve, we need the empirical power for a sequence of alternatives θ along the horizontal axis. The outer for loop varies the points $\theta(\mu)$ and the inner `replicate` loop estimates the power at the current θ.

```
n <- 20; m <- 1000; mu0 <- 500
sigma <- 100; mu <- c(seq(450, 650, 10))  #alternatives
M <- length(mu); power <- numeric(M)
for (i in 1:M) {
  mu1 <- mu[i]; pvalues <- replicate(m, expr = {
    #simulate under alternative mu1
    x <- rnorm(n, mean = mu1, sd = sigma)
    ttest <- t.test(x,alternative = "greater", mu = mu0)
    ttest$p.value  } )
  power[i] <- mean(pvalues <= .05)}
```

The estimated power $\hat{\pi}(\theta)$ values are now stored in the vector `power`. Next, plot the empirical power curve, adding vertical error bars at $\hat{\pi}(\theta) \pm \hat{se}(\pi(\theta))$ using the errbar function in `Hmisc` package.

```
library(Hmisc)  #for errbar
plot(mu, power); abline(v = mu0, lty = 1)
abline(h = .05, lty = 1)
#add standard errors
se <- sqrt(power * (1-power) / m)
errbar(mu, power, yplus = power+se, yminus = power-se,
       xlab = bquote(theta))
lines(mu, power, lty=3); detach(package:Hmisc)
```

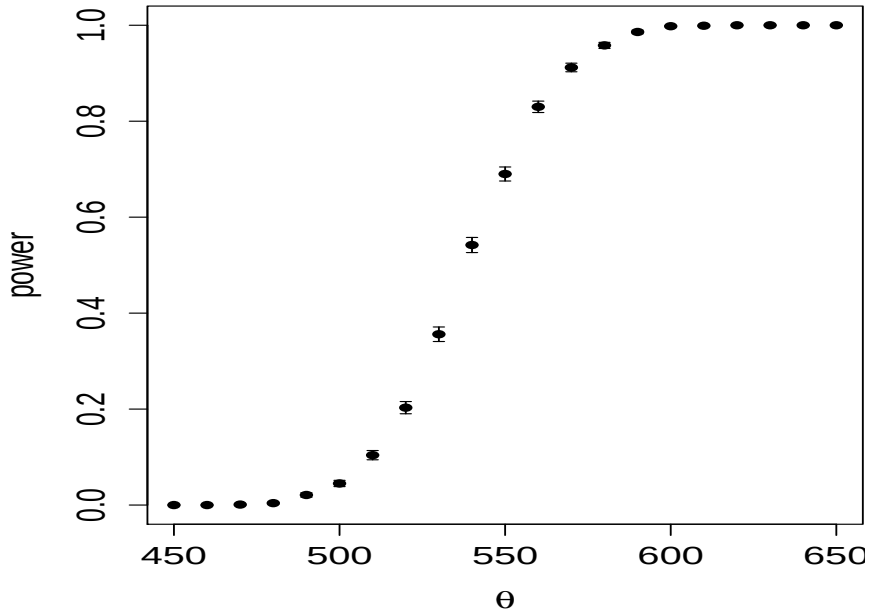

Fig.6.1: Empirical power $\hat{\pi}(\theta) \pm \hat{se}(\pi(\theta))$ for the t-test of $H_0 : \theta = 500$ vs $H_1 : \theta > 500$ in Example 6.9.

Remark 6.2

The non-central t dist. arises in power calculations for t-tests. The general non-central t with parameters (ν, δ) is $T(\nu, \delta) = (Z + \delta)/\sqrt{V/\nu}$ where $Z \sim N(0, 1)$ and $V \sim \chi^2(\nu)$ are independent.

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, and the t-statistic $T = (\overline{X} - \mu_0)/(S/\sqrt{n})$ is applied to test $H_0 : \mu = \mu_0$. Under H_0, T follows the central $t(n-1)$, but if $\mu = \mu_0$, T follows the non-central $t(n-1, \delta)$, where non-centrality parameter $\delta = (\mu - \mu_0)\sqrt{n}/\sigma$. A numerical approach to evaluating the cdf of $t(n-1, \delta)$, is R function `pt`. Also see `power.t.test`.

Example 6.10 (Power of the skewness test of normality)

The skewness test of normality was described in Example 6.8. Here, we estimate it by simulation against a contaminated normal dist. alternative in Example 6.3, which is

$$(1-\epsilon)N(\mu=0,\sigma^2=1)+\epsilon N(\mu=0,\sigma^2=100), 0\le\epsilon\le 1$$

When $\epsilon=0$ or $\varepsilon=1$, the dist. is normal, but it is non-normal for $0<\epsilon<1$. We can estimate the power of the skewness test for a sequence of alternatives indexed by ϵ and plot a power curve. We use $\alpha=0.1$ and the sample size $n=30$.

```
alpha <- .1; n <- 30; m <- 2500
epsilon <- c(seq(0, .15, .01), seq(.15, 1, .05))
N <- length(epsilon); pwr <- numeric(N)
#critical value for the skewness test
cv <- qnorm(1-alpha/2, 0, sqrt(6*(n-2) / ((n+1)*(n+3))))
for (j in 1:N) {              #for each epsilon
  e <- epsilon[j];  sktests <- numeric(m)
```

```
for (i in 1:m) {          #for each replicate
    sigma <- sample(c(1, 10), replace = TRUE,
                    size = n, prob = c(1-e, e))
    x <- rnorm(n, 0, sigma)
    sktests[i] <- as.integer(abs(sk(x)) >= cv)
  }
  pwr[j] <- mean(sktests)
}
#plot power vs epsilon
plot(epsilon, pwr, type = "b",
     xlab = bquote(epsilon), ylim = c(0,1))
abline(h = .1, lty = 3)
se <- sqrt(pwr * (1-pwr) / m)  #add standard errors
lines(epsilon, pwr+se, lty = 3)
lines(epsilon, pwr-se, lty = 3)
```

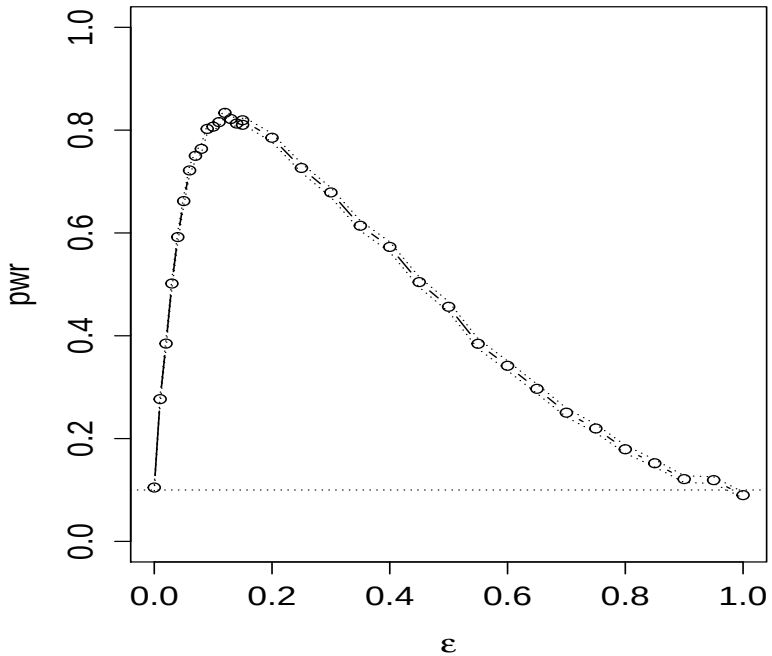

Fig.6.2: Empirical power $\hat{\pi}(\epsilon) \pm \hat{se}(\pi(\epsilon))$ for the skewness test of normality against ϵ-contaminated normal scale mixture alternative in Example 6.10.

Note that the power curve crosses the horizontal line corresponding to $\alpha = 0.10$ at both endpoints, $\varepsilon = 0$ and $\epsilon = 1$ where the alternative is normally distributed. For $0 < \epsilon < 1$ the empirical power of the test is greater than 0.10 and highest when ϵ is about 0.15.

6.3.3 Power comparisons

MC methods are often applied to compare the performance of different test procedures. Below we compare three tests of univariate normality.

Example 6.11 (Power comparison of tests of normality)

Compare the empirical power of the skewness test of univariate normality with the Shapiro-Wilk test and the energy test.

Let $\mathcal{N}$ denote the family of univariate normal dist.s. Then the test hypotheses are

$$H_0 : F_x \in \mathcal{N} \quad H_1 : F_x \notin \mathcal{N}.$$

The Shapiro-Wilk test is based on the regression of the sample order statistics on their expected values under normality, so it falls in the general category of tests based on regression and correlation.

The approximate critical values of the statistic are determined by a transformation of the statistic W to normality for sample sizes $7 \leq n \leq 2000$. The Shapiro-Wilk test is implemented by the R function shapiro.test.

The energy test is based on an energy distance between the sampled dist. and normal dist., so large values of the statistic are significant. The energy test for univariate and multivariate normality is implemented in `mvnorm.etest` in `energy` package.

For this comparison we set significance level $\alpha = 0.1$. The example below compares the power of the tests against the contaminated normal alternatives described in Example 6.3. The alternative is the normal mixture denoted by

$$(1-\epsilon)N(\mu = 0, \sigma^2 = 1) + \epsilon N(\mu = 0, \sigma^2 = 100), 0 \leq \epsilon \leq 1$$

```
# initialize input and output
library(energy)
alpha<-.1; n<-30; m<-500 #try small m for a trial run
test1<-test2<-test3<-numeric(m)
#critical value for the skewness test
cv <- qnorm(1-alpha/2, 0, sqrt(6*(n-2) / ((n+1)*(n+3))))
sim <- matrix(0, 11, 4)
# estimate power
for (i in 0:10) {
  epsilon <- i * .1
  for (j in 1:m) {
    e <- epsilon; sigma <- sample(c(1, 10), replace = TRUE,
                    size = n, prob = c(1-e, e))
    x <- rnorm(n, 0, sigma)
    test1[j]<-as.integer(abs(sk(x))>=cv)
    test2[j]<-as.integer(shapiro.test(x)$p.value<=alpha)
test3[j]<-as.integer(mvnorm.etest(x,R=200)$p.value<=alpha)}
  print(c(epsilon, mean(test1), mean(test2), mean(test3)))
  sim[i+1, ]<-c(epsilon,mean(test1),mean(test2),mean(test3))
detach(package:energy)
```

Standard error of the estimates is at most $0.5/\sqrt{m} = 0.01$. Estimates for empirical Type I error rate correspond to $\epsilon = 0$ and $\epsilon = 1$. All tests achieve approximately $\alpha = 0.1$ within one standard error, so it is meaningful to compare the results for power.

```
# plot the empirical estimates of power
plot(sim[,1], sim[,2], ylim = c(0, 1), type = "l",
     xlab = bquote(epsilon), ylab = "power")
lines(sim[,1], sim[,3], lty = 2)
lines(sim[,1], sim[,4], lty = 4)
abline(h = alpha, lty = 3)
legend("topright", 1, c("skewness", "S-W", "energy"),
       lty = c(1,2,4), inset = .02)
```

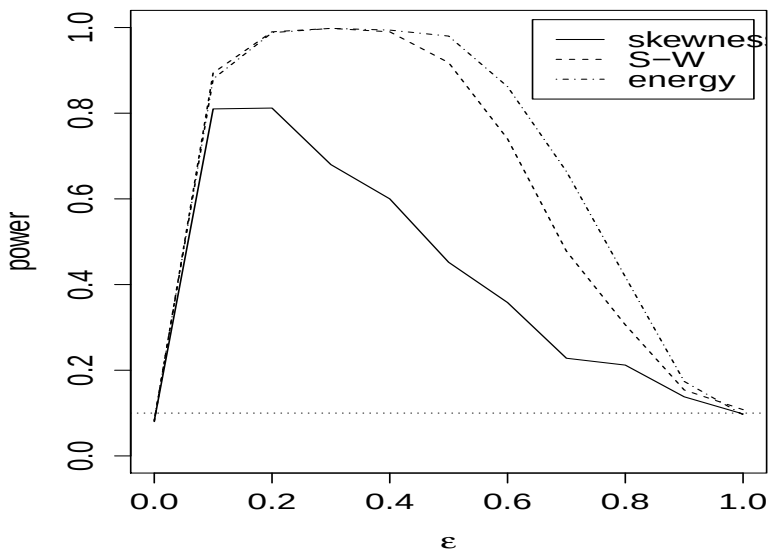

Fig.6.3: Empirical power of three tests of normality against a contaminated normal alternative in Example 6.11 ($n = 30, \alpha = 0.1, se \leq 0.01$)

The simulation results suggest that the Shapiro-Wilk and energy tests are about equally powerful against this type of alternative when $n = 30$ and $\epsilon < 0.5$. Both have higher power than the skewness test overall and energy appears to have highest power for $0.5 \leqslant \epsilon \leqslant 0.8$.

Empirical Power of Three Tests of Normality against a Contaminated Normal Alternative in Example 6.11 ($n = 30, \alpha = 0.1, se \leq 0.01$)

ϵ	skewness test	Shapiro-Wilk	energy test
0.00	0.0984	0.1076	0.1064
0.05	0.6484	0.6704	0.6560
0.10	0.8172	0.9008	0.8896
0.15	0.8236	0.9644	0.9624
0.20	0.7816	0.9816	0.9800
0.25	0.7444	0.9940	0.9924
0.30	0.6724	0.9960	0.9980
0.40	0.5672	0.9828	0.9964
0.50	0.4424	0.9112	0.9724
0.60	0.3368	0.7380	0.8868
0.70	0.2532	0.4900	0.6596
0.80	0.1980	0.2856	0.3932
0.90	0.1296	0.1416	0.1724
1.00	0.0992	0.0964	0.0980

6.4 Application: Count Five Test for Equal Variance

The examples in this section illustrate the MC method for a simple two sample test of equal variance. The two sample 'Count Five' test for equality of variance counts the number of extreme points of each sample relative to the range of the other sample.

- Suppose the means of the two samples are equal and the sample sizes are equal.
- An observation in one sample is considered extreme if it is not within the range of the other sample.
- If either sample has five or more extreme points, the hypothesis of equal variance is rejected.

Example 6.12 (Count Five test statistic)

The computation of the test statistic is illustrated with a numerical example. Compare the side-by-side boxplots in Fig.6.4 and observe that there are some extreme points in each sample with respect to the other sample.

```
x1 < - rnorm (20,0,sd=1); x2 < - rnorm (20,0,sd=1.5)
y < - c (x1,x2); group < -rep(1:2,each=length(x1))
boxplot(y~group,boxwex =.3,xlim =c(.5,2.5),main=" ")
points(group,y)
# now identify the extreme points
> range(x1); range(x2)
[1] -2.7825761.728505
[1]-1.5989173.710319
> i < - which (x1<min(x2)); j < -which (x2>max(x1))
> x1[i]; x2[j]
[1]-2.782576
[1]2.035521 1.809902 3.710319
```

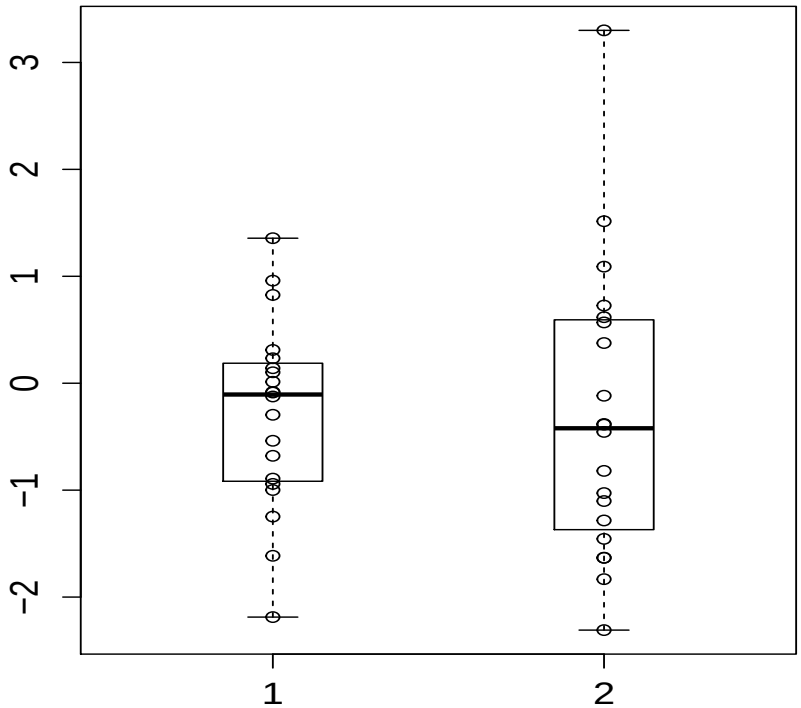

Fig.6.4: Boxplots showing extreme points for the Count Five statistic in Example 6.12.

The Count Five statistic is the maximum number of extreme points, `max(1, 3)`, so the Count Five test will not reject the hypothesis of equal variance. We only need the number of extreme points, and the extreme count can be determined without reference to boxplot.

```
out1 <- sum(x1 > max(x2)) + sum(x1 < min(x2))
out2 <- sum(x2 > max(x1)) + sum(x2 < min(x1))
> max(c(out1, out2))
[1] 3
```

Example 6.13 (Count Five test statistic, cont.)

Consider the case of two independent random samples from the same normal dist. Estimate the sampling dist. of the maximum number of extreme points, and find the 0.80, 0.90, and 0.95 quantiles of the sampling dist.

```
maxout <- function(x, y) {
  X <- x - mean(x)
  Y <- y - mean(y)
  outx <- sum(X > max(Y)) + sum(X < min(Y))
  outy <- sum(Y > max(X)) + sum(Y < min(X))
  return(max(c(outx, outy)))}
n1 <- n2 <- 20; mu1 <- mu2 <- 0
sigma1 <- sigma2 <- 1; m <- 1000
# generate samples under H0
stat <- replicate(m, expr={
  x <- rnorm(n1, mu1, sigma1)
  y <- rnorm(n2, mu2, sigma2)
  maxout(x, y)})
print(cumsum(table(stat)) / m)
print(quantile(stat, c(.8, .9, .95)))
```

The empirical cdf and quantiles are

```
1  2  3  4  5  6  7  8  9  10  11
0.149  0.512  0.748  0.871  0.945  0.974  0.986  0.990  0.996  0.999  1.000
80%  90%  95%
4  5  6
```

The `quantile` function gives 6 as the 0.95 quantile. However, if $\alpha = 0.05$, 5 appears to be the best choice. The `quantile` function is not always the best way to estimate a critical value. If quantile is used, compare the result to the empirical cdf.

The 'Count Five' test criterion can be applied for independent random samples when the r.v. are similarly distributed and sample sizes are equal. (r.v. X and Y are called similarly distributed if Y has the same distribution as $(X-a)/b$ where a and $b > 0$ are constants.) When the data are centered by their respective population means, the Count Five test has significance level at most 0.0625.

In practice, the populations means are generally unknown and each sample would be centered by subtracting its sample mean. Also, the sample sizes may be unequal.

Example 6.14 (Count Five test)

Use MC methods to estimate the significance level of the test when each sample is centered by subtracting its sample mean. Here again we consider normal dist. The function `count5test` returns the value 1 (reject H_0) or 0 (do not reject H_0).

```
count5test <- function(x, y) {
  X <- x - mean(x);    Y <- y - mean(y)
  outx <- sum(X > max(Y)) + sum(X < min(Y))
  outy <- sum(Y > max(X)) + sum(Y < min(X))
  # return 1 (reject) or 0 (do not reject H0)
  return(as.integer(max(c(outx, outy)) > 5))}
n1 <- n2 <- 20; mu1 <- mu2 <- 0
sigma1 <-  sigma2 <- 1; m <- 10000
tests <- replicate(m, expr = {
  x <- rnorm(n1, mu1, sigma1)
  y <- rnorm(n2, mu2, sigma2)
  x <- x - mean(x)  #centered by sample mean
  y <- y - mean(y);    count5test(x, y)} )
alphahat <- mean(tests); print(alphahat)
> print(alphahat)
[1] 0.0565
```

If the samples are centered by the population mean, we should expect an empirical Type I error rate of about 0.055, from our previous simulation to estimate the quantiles of the `maxout` statistic. In the simulation, each sample was centered by subtracting the sample mean, and the empirical Type I error rate was $0.0565 (se = 0.0022)$.

Example 6.15 (Count Five test, cont.)

Repeating the previous example, estimate the empirical Type I error rate when sample sizes differ and the 'Count Five' test criterion is applied. Each sample is centered by subtracting the sample mean

```
n1 <- 20; n2 <- 30; mu1 <- mu2 <- 0
sigma1 <- sigma2 <- 1; m <- 10000
alphahat <- mean(replicate(m, expr={
  x <- rnorm(n1, mu1, sigma1); y <- rnorm(n2, mu2, sigma2)
  x <- x - mean(x)  #centered by sample mean
  y <- y - mean(y); count5test(x, y)})); print(alphahat)
[1] 0.1064
```

The 'Count Five' criterion does not control Type I error at $\alpha \leq 0.0625$ when the sample sizes are unequal. Repeating the simulation with $n_1 = 20$ and $n_2 = 50$, the empirical Type I error rate was 0.2934.

Example 6.16 (Count Five, cont.)

Use MC methods to estimate the power of the Count Five test, where the sampled dist. are $N(\mu_1 = 0, \sigma_1^2 = 1)$, $N(\mu_2 = 0, \sigma^2 = 1.52)$, and the sample sizes are $n_1 = n_2 = 20$.

```
# generate samples under H 1 to estimate power
sigma1 < - 1
sigma2 < - 1.5
power < -mean (replicate(m,expr = {
x < - rnorm (20,0,sigma1 )
y < - rnorm (20,0,sigma2 )
count 5 test (x,y)
} ) )
> print (power)
[1]0.3129
```

The empirical power of the test is $0.3129 (se \leq 0.005)$ against the alternative $(\sigma_1 = 1, \sigma_2 = 1.5)$ with $n_1 = n_2 = 20$.