ST2005: Applied Probability II

Computing Laboratory 2

Coverage of confidence intervals for the mean

We saw in class that the confidence interval for the mean for a Normal Distribution with known variance is in form of $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ has a coverage of $100(1-\alpha)\%$. This means that having assumed that $X_{1,} X_{2,} \ldots, X_n \sim N(\mu, \sigma^2)$, then the interval above will contain the true value of μ in $100(1-\alpha)\%$ of cases. You are now going to verify this using simulation.
The following function in R returns the CI limits as a vector.

```
CIfunc <- function(x, alpha, x.var)
  {
# find the number of elements in x
n <- length(x)
# calculate the mean of x
x.mean <- mean(x)
# appropriate z-value
z <- qnorm(1-alpha/2)
Low <- x.mean - z*sqrt(x.var/n) # lower bound
High <- x.mean + z*sqrt(x.var/n) # upper bound
# return confidence interval as a vector
return(c(Low, High))
}
```

Q1. Simulate a Normal dataset of size $n = 50$ from a $N(0,1)$ distribution and find the 90%, (95%, 99% CIs. Explain the changes in CIs different level of confidence.

Use the following function to generate m rows of n iid Normal samples.

```
m <- 100000
n <- 15
Samples <- function(m,n)
{
  matrix( rnorm( m*n ), nrow=m, ncol=n, byrow=TRUE  )
}
Samples <- Samples(m,n)
```

Q2. Simulate 100,000 normal datasets of size $n = 15$ from a $N(0,1)$ distribution. Return your datasets as a 100,000 by 15 matrix.

Use the following function to check what would be the empirical percentage of the CIs which contain the actual population mean of 0.

```
Percent <- function( X ) {
  CIs <- matrix( nrow=nrow(X), ncol=2 )
  for( k in 1:nrow(X) ) CIs[k, ] <- CIfunc( X[k,], 0.05 , 1)
  z <- ( CIs[,1] < 0 ) * ( CIs[ , 2 ] > 0)
  sum(z)/nrow(X)
}
Percent(Samples)
```

Q3. Compute a 95% confidence interval for the mean for each row of the matrix returned by **Q2** by using the function used for **Q1.** Return the proportion of the 100,000 confidence intervals which contain zero. What would you expect this to be?

If the population Variance is unknown, t-distribution is used instead of Normal distribution in CI calculation as $\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$. You can find $t_{n-1,\alpha/2}$ using the R syntax

```
qt(1-alpha/2, df)
```

where df (Degrees of Freedom) equals n-1.

Q4. Amend the functions used in Q1-Q3 for the situation that you wish to assume the population Variance is unknown when calculating CI.