**Statistics 406 Lab 5 Notes**

**Tuesday, October 10**

**GSI:** Adam Rothman

1. **Jensen's Inequality**
   For a random variable $X$ and **convex** function $f$ (eg. $f(x) = x^2$),

$$f(EX) \leq Ef(X). \tag{1}$$

For a random variable $X$ and **concave** function $f$ (eg. log, square-root),

$$Ef(X) \leq f(EX). \tag{2}$$

You will want to use (2) for the first problem on your homework. Consider estimating $\mu^2$ using $\overline{X}^2$ for data $X_1, \ldots, X_n$ iid $N(1,1)$. Write R code to demonstrate the bias that occurs and relate this bias back to Jensen's inequality.

**Solution:**

```
reps <- 1e4

for (n in c(5,10,20))
{
  X <- array(rnorm(n*reps,1,1), c(n,reps))
  sampMeans <- apply(X, 2, mean)
  sampMeanSq <- mean((sampMeans)^2)
  print(sampMeanSq)
}
```

You should notice that $\overline{X}^2$ seems to have a slight positive bias, which decreases with increasing $n$. Here is a mathematical explaination:

$$\begin{aligned}
f(E\overline{X}) &\leq Ef(\overline{X}) \\
(E\overline{X})^2 &\leq E[(\overline{X})^2] \\
\mu^2 &\leq E[(\overline{X})^2] \\
\text{population mean squared} &\leq E[\text{squared sample mean}] \\
E[\text{squared sample mean}] - \text{population mean squared} &\geq 0 \\
\text{Bias}(\overline{X}^2) &\geq 0
\end{aligned}$$

2. Consider data $X_1, \ldots, X_n$ iid, write R code to generate `nresamp` bootstrapped data sets from the list $X$ placing them in the columns of a matrix $\mathrm{B}$.

**Solution:**

```
nresamp <- 1000

X <- rnorm(n)
p <- length(X)
ii <- ceiling(p*runif(p*nresamp))
B <- X[ii]
B <- array(B, c(p,nresamp))
```

3. The 'bootstrap t-method' for confidence intervals for EX operates somewhat differently from the percentile method given in the course notes. To construct a t-method CI for EX, first calculate the mean $\overline{X}$ and standard deviation $\hat{\sigma}$ for the actual data. Then generate K = 1000 non-parametric bootstrap data sets. For each bootstrap data set, calculate the sample mean $\overline{X}_k$ and sample standard deviation $\hat{\sigma}_k$. Then let F denote the empirical 95th percentile of

$$\frac{|\overline{X}_k - \overline{X}|}{\hat{\sigma}_k}$$

The 95 percent CI for $EX$ is taken to be

$$\overline{X} \pm F\hat{\sigma}$$

Use R to construct a 95 percent CI for $EX$ using the above method for $X_1, \ldots, X_n$ iid $N(1,1)$ and $n = 20$.

**Solution:**

```
n <- 20
X <- rnorm(n,1,1)
xbar <- mean(X)
sigmahat <- sd(X)

## Bootstrap samples.
ii <- ceiling(n*runif(1000*n))
B <- array(X[ii], c(n,1000))

xbar_k <- apply(B, 2, mean)
sigmahat_k <- apply(B, 2, sd)
MD <- abs(xbar_k-xbar) / sigmahat_k
```

```
MDsorted <- sort(MD)
F <- MDsorted[950]

c1 <- xbar - F*sigmahat
c2 <- xbar + F*sigmahat
print(c(c1,c2))
```