**STA 250/MTH 342 Intro to Mathematical Statistics**
**Lab Session 4 / Feb 02, 2015 / Handout**

In this session we study how to use **R** to generate random variables from some existing distributions. We then use these random variables to verify the law of large numbers, and the central limit theorem. If time permits, I will also present the way of making gif animations to visualize some pdf/pmf's.

See: `https://stat.duke.edu/courses/Spring15/sta250/labs/` for links to source code and data. Submit lab solutions via email to: `sta250@stat.duke.edu`. Any plots should be included in postscript form as attachments. The email subject must be "STA250 ..." with "..." replaced by your name.

**1: Generating random variables from existing distributions.** To generate $N$ independent $\mathsf{Po}(\lambda)$ random variables, use the command: `rpois(`$N$`, lambda = `$\lambda$`)`.

```
> rpois(10, lambda = 12)
 [1] 10  6 13  6 11  9 17 12 16 13
```

We may check the sample mean and variance of the generated random variables:

```
> mean(rpois(100000, lambda = 12))
[1] 11.99181
> var(rpois(100000, lambda = 12))
[1] 11.99201
```

Note that according to the **R** documentation, if $\mathbf{x} = (x_1, \cdots, x_N)$,

$$\texttt{var}(\mathbf{x}) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2,$$

which is an unbiased estimator of the true variance (here $\lambda$).

**TASK 1** Draw $N = 10,000$ independent random variables from the $\mathsf{Ga}(\alpha, \beta)$ distribution with $\alpha = 3$ and $\beta = 5$, and evaluate their sample mean and variance. Recall that the mean and variance of the distribution is $\alpha/\beta$ and $\alpha/\beta^2$ respectively. The **R** command to use is `rgamma(`$N$`, shape = `$\alpha$`, rate = `$\beta$`)`.

**2: Distribution of order statistics.** This section is motivated by Problem 6, in HW4. Let $X_1, \cdots, X_N$ be independent random variables from Uniform([0, 1]). Let $Y_N = \max\{X_1, \cdots, X_N\}$. Now we generate $m$ independent copies of $Y_N$ and plot their histogram. We compare the histogram with the density function of $Y_N$ obtained by theoretical derivation.

Recall that the CDF and pdf of $Y_N$ at $x \in [0, 1]$ are

$$F_N(x) := \Pr(Y_N \le x) = \Pr(X_1, \cdots, X_N \le x) = \prod_{i=1}^{N}\Pr(X_i \le x) = x^N,$$

$$f_N(x) := F_N'(x) = N x^{N-1}.$$

Now set $N = 5$. On each call, the following function `umax(N)` will generate one independent copy of $Y_N$.

```
umax <- function(N) max(runif(N));
```

Now we generate $m = 2000$ independent copies of $Y_N$, plot the histogram and the density function.

```
a <- Vectorize(umax)(rep(5, 2000));
x <- seq(0, 1, len = 1000);
hist(a, breaks = 30, prob = T, ylim = c(0,5));
lines(x, 5 * x^4, type = "l", lwd = 3, col = "red");
```

You will get a picture as shown in Fig 1. The above code is available in file `h1.R`.

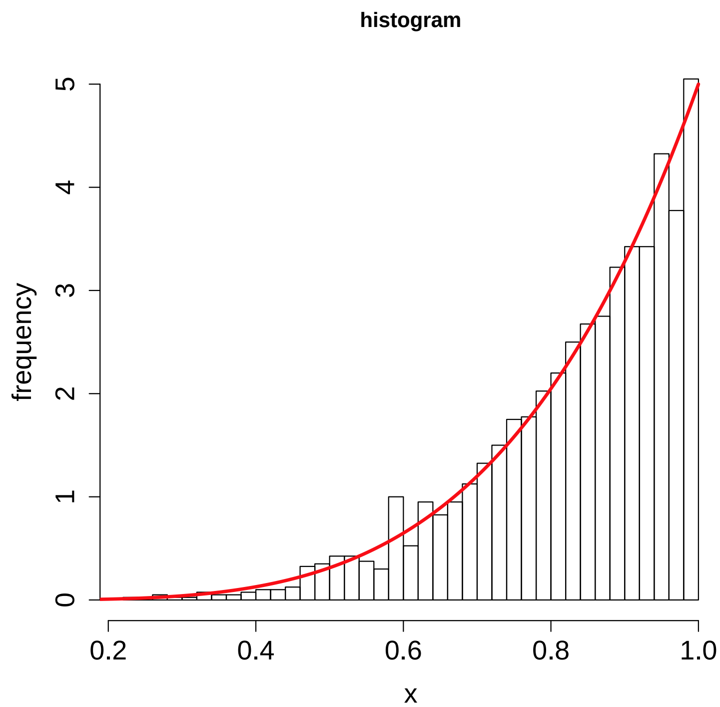


Figure 1: Histogram of 2000 independent copies of $Y_N$.

**TASK 2** Recall that $Z_N := \min\{X_1, \cdots, X_N\}$ has CDF $F_N(x) = 1 - (1-x)^N$ and pdf $f_N(x) = N(1-x)^{N-1}$ for $x \in [0,1]$. Please make a picture similar to Figure 1 but for $Z_N$ with $N = 6$. Evaluate the sample mean and sample variance. Compare them with the true values $E[Z_6] = 1/7 = 0.1428571$ and $\mathsf{Var}(Z_6) = 3/196 = 0.01530612$. Paste your code to the email, and attach the picture you obtain.

**3: Verify the law of large numbers numerically.** Recall that with independent observations $X_1, \cdots, X_n$ from $\mathsf{Ex}(1/\theta)$, the MLE of $\theta$ is $\widehat{\theta}_N = \bar{X}$. We now study that as $N \to \infty$, how does the absolute error loss $|\widehat{\theta}_N - \theta|$ decay. We let $\theta = 1.5$. Recall that $f(x \mid \theta) = \frac{1}{\theta} e^{-x/\theta} \mathbf{1}_{\{x>0\}}$ and when $X \sim \mathsf{Ex}(1/\theta)$, $\mathsf{E}[X] = \theta$ and $\mathsf{Var}[X] = \theta^2$.

The following function `exp.ael` generate $N$ independent samples and gives the absolute error loss $|\widehat{\theta}_N - \theta|$.

```
exp.ael <- function(N) abs(mean(rexp(N, rate = 1 / 1.5)) - 1.5);
```

We call `exp.ael` 30 times with $N = 10, 100, 1000, 10,000, 100,000$ each. We average the results.

```
> mean(replicate(30, exp.ael(10)));
[1] 0.3307285
> mean(replicate(30, exp.ael(100)));
[1] 0.1019838
> mean(replicate(30, exp.ael(1000)));
[1] 0.04417023
> mean(replicate(30, exp.ael(10000)));
[1] 0.0141305
> mean(replicate(30, exp.ael(100000)));
[1] 0.004082873
```

The decay of the error is well demonstrated.

**TASK 3** Recall that by its invariance, the MLE of $\theta^2$ is $\bar{X}^2$. For the same problem, examine the decay of the estimator of variance with increasing sample size.

**4: Verify the central limit theorem numerically.** Recall the Central Limit Theorem. If the random variables $X_1, \cdots, X_n$ are drawn independently from a distribution with mean $\mu$ and variance $\sigma^2 < \infty$, then as $n \to \infty$, the statistic

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

will converge in distribution to the standard normal distribution $\mathsf{No}(0, 1)$. We now do a simple simulation with Poisson(1). The following function gives the above statistic.

```
pos.sta <- function(n) (mean(rpois(n, lambda = 1)) - 1) * sqrt(n);
```

We now do the simulation 10,000 times with $n = 1000$ and plot the histogram of the values of the statistic. We also plot the $\mathsf{No}(0, 1)$ density curve for comparison.

```
hist(Vectorize(pos.sta)(rep(1000,10000)), breaks = 30, prob = T,
  xlim = c(-4, 4), ylim = c(0, 0.5));
x <- seq(-4, 4, len = 1000);
lines(x, dnorm(x), type = "l", lwd = 3, col = "red");
```

One may get a plot like Figure 2. The above code is available on `h2.R`. One may get some feeling about how fast **R** is! On my aging desktop machine,

```
> system.time(source("h2.R"));
   user  system elapsed
   user  system elapsed
  0.807   0.001   0.819
```

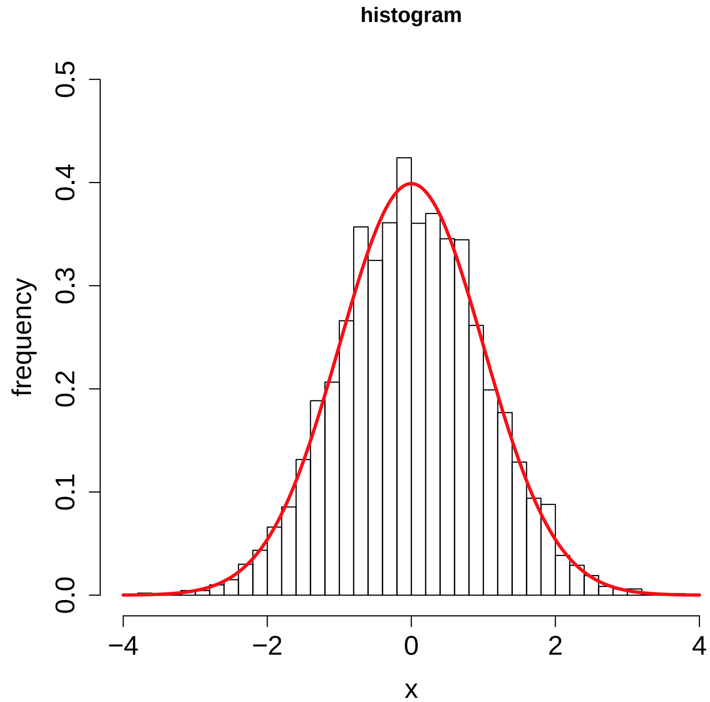


Figure 2: Histogram of 10,000 independent copies of $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ with $n = 1000$.

**TASK 4** Recall that the uniform distribution $\mathtt{Un}([0,1])$ has mean 1/2 and variance 1/12. Please repeat the above simulation by replacing Poisson distribution with Uniform on the unit interval.

**5: Animation.** If time permits I will talk about the code to generate animations. They are `anim.****.R`. The obtained `.gif` animations help one establish feelings about some distributions. The code calls a free image manipulation application called ImageMagick. If it is not yet on your system, just install it from `www.imagemagick.org`.

~~END~~