Maximum Likelihood Estimation and Model Fitting, CMEE MSc. Tin-Yu Hui

Practical 1 (Monday)

<u>Question 1</u>

i. Let X be a Bernoulli random variable with success probability p. Show that $E(X) = p$ and $E(X^2) = p$.

ii. Calculate the variance of X from $E(X)$ and $E(X^2)$.

<u>Question 2</u>

i. Given a r.v. X with pdf $f_X(x) = \frac{2}{x^3}, 1 \leq x < \infty$, show that $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

ii. Find $E(X)$.

Question 3

i. Let $X \sim Poisson(\lambda)$. Calculate (by hand) its expected value $E(X)$

ii. Let $X \sim Poisson(\lambda = 2)$. Remember that a Poisson distribution models the number of events that had occurred within a time interval with an average rate of occurrence λ. Let us visualise its pmf in `R` using `dpois()` for x between 0 and 10:

```
x<-0:10
y<-dpois(x, lambda=2)
plot(x, y, pch=16, ylab='pmf', xlab='outcome')
```

Note that the Poisson pmf is still defined for $x > 10$, but the associated probabilities are too small that the pmf at that range becomes less interesting.

iii. Calculate $Pr(X = 4)$ and $Pr(X \leq 3)$, and verify your answers with R:

```
# Pr(X=4)
dpois(4, lambda=2)
# Pr(x<=3)
sum(dpois(0:3, lambda=2)
# OR EQUIVALENTLY
ppois(3, lambda=2)
```

iv. We can generate Poisson random numbers using `rpois()`. Let us sample 1000 independent Poisson numbers with $\lambda = 2$:

```
x<-rpois(1000, lambda=2)
x
hist(x) # PLOT A HISTOGRAM
```

A different set of random numbers will be generated every time when you run `rpois()`. In some circumstances (e.g. for debugging) we would like to generate a “fixed” set of random numbers, which can be achieved by setting a random seed:

```
set.seed(123)
x<-rpois(10, lambda=2)
y<-rpois(10, lambda=2)
x==y # NOT THE SAME
x
```

```
# RESET RANDOM SEED
set.seed(123)
z<-rpois(10, lambda=2)
x==z # SHOULD BE IDENTICAL
z
```

Note that the random seeds in R-4.x.x are not compatible to those from R-3.x.x and below.

Question 4

i. Suppose X is Exponentially distributed with rate $\lambda > 0$. Clearly X is a continuous r.v.. The pdf is X is $f_X(x) = \lambda e^{-\lambda x}$, and its support is $[0, \infty)$. Show that $f_X(x)$ is a valid pdf (i.e. area under pdf=1).
$\int_0^\infty f_X(x)dx =$

ii. Calculate $E(X)$.

iii. Show that the cumulative density function (cdf) of X is $F_X(x) = 1 - e^{-\lambda x}$.

iv. Let $X \sim Exponential(\lambda = 2)$. Plot the pdf of X in R for $0 \leq x \leq 5$ with interval 0.01. The function `dexp()` may be useful.

v. Calculating probabilities from pdfs. For a continuous r.v. X it is pointless to calculate the probability of X takes on any certain value, as $\Pr(X = x) = 0$. This is because there are infinitely many outcomes for a continuous r.v. and that you can never exactly hit a particular number (think of a number with infinitely many decimal places). We can only calculate the probability that X falls within an interval, say, $\Pr(a \leq X \leq b)$. And such probability is the area under the pdf between the interval: $\Pr(a \leq X \leq b) = \int_a^b f_X(x)dx$. Let us calculate the probability of $\Pr(0 \leq X \leq 1)$ for $X \sim Exponential(\lambda = 2)$. Use `integrate()` for numerical integration.

```
integrate(dexp, lower=0, upper=1, rate=2)
```

Question 5

i. Let $X \sim N(\mu, \sigma^2)$. We call X the “standard normal” when $\mu = 0$ and $\sigma^2 = 1$. Plot the pdf of the standard normal distribution from $x = -3$ to $x = 3$.

ii. What is $\Pr(2 \leq X \leq 3)$? And what is $\Pr(-1.96 \leq X \leq 1.96)$? Use `integrate()` to help you with this.

iii. Verify your answer with the following commands:

```
pnorm(3)-prnorm(2)
pnorm(1.96)-pnorm(-1.96)
```

From the examples above it is safe to say that R knows a lot of distributions. For each distribution, we use the commands with prefix `r` (e.g. `rnorm()`, `rpois()`, `rbinom()`) to generate random samples, and use those with prefix `d` to evaluate pmf/pdf. Those with prefix `p` (e.g. `pnorm()`) return the cdf of a distribution, and `q` for quantiles.

Question 6 [Central Limit Theorem, adopted from Mick Crawley's GLM course]

The central limit theorem (CLT) states that for any distribution with finite expected value and variance, the sample mean of the random samples from that distribution tends to be normally distributed.

Take the negative binomial distribution as an example. Let us plot the histogram of 1000 negative binomial random numbers with $r = 1$ and $p = 0.2$:

```
y<-rnbinom(1000, 1, 0.2)

hist(y)
```

It is definitely not "normal" as it skews to a side. The CLT says the mean of samples will follow a normal distribution even for a badly behaved distribution like this. To visualise this effect let us consider the following codes:

```
# GENERATE 30000 NETAGIVE BINOMIAL RANDOM NUMBERS

# AND PUT THEM INTO A 1000-BY-30 MATRIX

y<-matrix(rnbinom(30*1000, 1, 0.2), nr=1000, nc=30)


# CALCULATE ROW MEAN

y.row.mean<-apply(y, 1, mean)


# PLOT THE HISTOGRAM OF THESE 1000 ROW MEANS

hist(y.row.mean)
```

Does the histogram look more "normal" now? This is why the normal distribution is so famous and widely used. The CLT holds for larger sample sizes, say, 30 or above.

Question 7 [Moment-generating function, as an example]

Let $X \sim N(\mu, \sigma^2)$. Find $E(X)$ and $E(X^2)$ from its mgf, $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$.