Statistics 2 Unit 1

Group ?

March 13, 2018

Contents

1 Task 1

By using the following equation $\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = (x+y)^n$ we can easily find the characteristic function of the binominal distribution.
$E\left(e^{itk}\right) = \sum_{k=0}^{n} \binom{n}{k} e^{itk} p^k (1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (e^{it}p)^k (1-p)^{n-k} = \left(1 + p\left(e^{it}-1\right)\right)^n$
Now, By Levy's continuity theorem we can show that binomial distribution tends to Poisson distribution when: $n \to \infty \wedge p \to 0$ s.t. $np \to \lambda$.
$\left(1 + p\left(e^{it}-1\right)\right)^n = \left(1 + \frac{np\left(e^{it}-1\right)}{n}\right)^n$
$\lim_{n\to\infty} \left(1 + \frac{np\left(e^{it}-1\right)}{n}\right)^n = e^{\lambda\left(e^{it}-1\right)}$
We can notice that $e^{\lambda\left(e^{it}-1\right)}$ is exactly the charateristic function of Poisson distribution with parameter λ. By continuity theorem the result follows.

2 Task 2

Let $X \sim \Gamma(\alpha, \lambda)$. Also note that $E(X) = \frac{\alpha}{\lambda} \wedge Var(X) = \frac{\alpha}{\lambda^2}$.
Fisrt lets apply standartization to our gamma distribution:
$Z = \frac{x - \frac{\alpha}{\lambda}}{\frac{\sqrt{\alpha}}{\lambda}} \sim N(0,1)$ Thus, the correspoinding characteristic function is:
$\varphi_z(t) = E\left(e^{itz}\right) = e^{-it\sqrt{\alpha}} \left(\frac{1}{1 - it/\sqrt{\alpha}}\right)^\alpha$
Now let's transform a bit the left part of the above equation:
$\left(\frac{1}{1 - it/\sqrt{\alpha}}\right)^\alpha = (1 + it\sqrt{\alpha}/-\alpha)^{-\alpha} = e^{-\alpha \ln\left(\left(1+it\sqrt{\alpha}/-\alpha\right)\right)}$
Now let's implement Taylor series expansion to $\ln\left((1 + it\sqrt{\alpha}/-\alpha)\right)$
$e^{-\alpha \ln\left(\left(1+it\sqrt{\alpha}/-\alpha\right)\right)} = e^{it\sqrt{\alpha} - \frac{t^2}{2} + O\left(t^3\right)}$
The, going back to our char. function, the first term of the expansion cancels with the first term of the characteristic function, and the term $O(t^3)$ gets neglictible as $\alpha \to \infty$ Thus, we get $e^{-\frac{t^2}{2}}$ which is the characteristic function of the standard normal distribution. And by continuity theorem the result follows.

3 Task 3

Let Round-off error be written as a R.V. $Y = \sum_{i=1}^{100} X_i$
Note that Expecatation and Variance of unif. distribution (from a to b) are:
$E(X) = \dfrac{b+a}{2}$ and $Var(X) = \dfrac{(b-a)^2}{12}$
Then $E(Y) = \sum_{i=1}^{100} E(X_i) = 0 \wedge Var(Y) = \sum_{i=1}^{100} Var(X_i) = \frac{25}{3}$
Now let's apply CLT:
$Z = \dfrac{S_n - n\mu}{\sigma\sqrt{n}} = \dfrac{\sum_{i=1}^{100}(X_i)}{\frac{5\sqrt{3}}{3}}$
Let's approximate the probability that the round-off error exceeds 1, 2 and 5
$P(|S_n| > 1) = P\left(\left|\sum_{i=1}^{100}(X_i)\right| > 1\right) = P\left(|Z| > \frac{5\sqrt{3}}{3}\right) = 2P(Z > 0.3464) = 2(1 - \phi(0.3464)) = 0.7290421$ 2*(1-pnorm(0.3464))
$P(|S_n| > 2) = P\left(\left|\sum_{i=1}^{100}(X_i)\right| > 2\right) = P\left(|Z| > 2\frac{5\sqrt{3}}{3}\right) = 2P(Z > 0.693) = 2(1 - \phi(0.693)) = 0.4883096$ 2*(1-pnorm(0.693))
$P(|S_n| > 5) = P\left(\left|\sum_{i=1}^{100}(X_i)\right| > 5\right) = P\left(|Z| > 5\frac{5\sqrt{3}}{3}\right) = 2P(Z > 1.732) = 2(1 - \phi(1.732)) = 0.08327356$ 2*(1-pnorm(1.732))

4 Task 4

Let's use CLT for the approximation.
$E(X) = \int_0^1 x \cdot 2x dx = \dfrac{2}{3}$
$E(X^2) = \int_0^1 x^2 \cdot 2x dx = \dfrac{1}{2}$
$Var(X) = \dfrac{1}{18}$
$\Rightarrow E(S) = 20 \cdot \dfrac{2}{3} = \dfrac{40}{3} \wedge Var(S) = \dfrac{10}{9}$

$$P(S \le 10) = P\left(S - \frac{40}{3} \le 10 - \frac{40}{3}\right) = P\left(\frac{S - \frac{40}{3}}{\sqrt{\frac{10}{9}}} \le \frac{10 - \frac{40}{3}}{\sqrt{\frac{10}{9}}}\right) = \Phi\left(\frac{10 - \frac{40}{3}}{\sqrt{\frac{10}{9}}}\right)$$

```r
pnorm((10 - 40/3)/sqrt(10/9))

## [1] 0.0007827011
```

5 Task 5

5.1 a)

1.
$n = 100$:

```
U1 <- runif(100,0,1)
mean(cos(2*pi*U1))

## [1] 0.04369909
```

2.
$n = 1000$:

```
U2 <- runif(1000,0,1)
mean(cos(2*pi*U2))

## [1] -0.0105909
```

The correct solution to the integral is zero (easy to find by substitution where $u = 2\pi x$).
The second Monte Carlo Integration with $n = 1000$ is closer to the true value, what supports the Law of Large Numbers.

5.2 b)

```
U3 <- runif(10000,0,1)
mean(cos(2*pi*U3^2))

## [1] 0.2510589
```

This integral does not have a closed form solution. One has to evaluate it with the Fresnel-integral. The Fresnel-Integral is defined as $C(x) = \int_0^x cos(t^2)dt = \sum_0^\infty(-1)^n \frac{x^{4n+1}}{(2n)!(4n+1)}$. So by substitution where $u = \sqrt(2)\sqrt(\pi)x$ we can solve $\frac{1}{\sqrt(2)\sqrt(\pi)} \int_0^x cos(u^2)du$. After resubstituion it reduces to $\frac{C(2)}{2} \approx 0.244$.

We can find the exact solution also with R:

```r
f <- function(x) cos(2 *pi*x^2)
integrate(f,0,1)

## 0.2441267 with absolute error < 2.3e-10
```

6 Task 6

The idea of MC estimation: to evaluate the integral $\Theta = \int g(x)f(x)dx$ we can compute $\hat{\Theta} = \frac{1}{n}\sum_{n=1}^{n} g(x_i)$.

If the X_i are drawn independently, the $Var(\hat{\Theta}) = \frac{Var(g(x))}{n}$, consequently the standard deviation is equal to $sd(\hat{\Theta}) = \frac{sd(g(x))}{\sqrt{n}}$.

Now let's compute the integral $\int_0^1 Cos(2\pi x)dx$ by using the MC estimation:

```r
n<- 10000
x<-runif(n)
theta_hat<-mean(cos(2 * pi * x))
theta_hat

## [1] -0.004271987

sd<-sqrt(var(cos(2 * pi * x))/n)
```

The actual value of this integral should be 0.

```r
actual_error<- theta_hat - 0
c(sd, actual_error)

## [1]  0.007025364 -0.004271987
```

Here we see that standard deviation of estimation and actual error differ a lot. We can also notice that our accuracy is depend on the size of the sample. So, now we will try to increase our n.

```r
n<- 100000
x<-runif(n)
theta_hat<-mean(cos(2 * pi * x))
theta_hat
```

```
## [1] -0.0001710685

sd<-sqrt(var(cos(2 * pi * x))/n)
actual_error<- theta_hat - 0
c(sd, actual_error)

## [1]  0.0022381366 -0.0001710685
```

We can conclude that by increasing n we increase the accuracy of estimation.

7 Task 7

We know that:
$\Theta = \int_a^b g(x)dx$

$\hat{\Theta} = \frac{1}{n}\sum \frac{g(X_i)}{f(X_i)}$
Where f is is a density function on$[a, b]$ from which we have generated the X.

7.1 a)

$\mathbf{E}(\hat{\Theta}) = \mathbf{E}(\frac{1}{n}\sum \frac{g(X_i)}{f(X_i)}) = \frac{1}{n} * n * \mathbf{E}(\frac{g(X_i)}{f(X_i)}) = \int_a^b \frac{g(X_i)}{f(X_i)} * f(X_i)dx = \int_a^b g(x)dx = \Theta$

7.2 b)

$\mathbf{Var}(\hat{\Theta}) = \mathbf{Var}(\frac{1}{n}\sum \frac{g(X_i)}{f(X_i)}) = \frac{n}{n^2} * \mathbf{Var}(\frac{g(X_i)}{f(X_i)}) = \frac{1}{n} * (\mathbf{E}(\frac{g(X_i)^2}{f(X_i)^2}) - \Theta^2) = \frac{1}{n} * (\int_a^b \frac{g(X_i)^2}{f(X_i)^2} * f(X_i)dx - \Theta^2) = \frac{1}{n} * (\int_a^b \frac{g(X_i)^2}{f(X_i)}dx - \Theta^2)$

Let's assume X is distributed uniformly on interval $[0;\ 2]$ such that $f(X) = 0.5$.
Example of finite variance:
Let $g(X) = X$. Then $\int_0^2 2X^2 dx$ is finite and therefore variance is finite.
Example of infinite variance:
Let $g(x) = X^{-0.5}$. Then $(g(X))^2 = X^{-1}$. Therefore $(\int_0^2 2X^{-1}dx = 2(ln(2) - ln(0))$, so variance is infinite.

7.3 c)

we can use R:

```
f<-function(x) {
  exp(-x^2/2) * 1/(sqrt(2*pi))
}
theta<-integrate(f,0,1)
n<-100
r<-runif(100,0,1)
thetahat<-(1/n)*(1/sqrt(2*pi)*sum(exp(-r^2/2)))
difference <- theta[[1]][1] - thetahat
difference

## [1] -0.002713075
```

we can use a normal distribution, instead

```
r1<-rnorm(100,0,1)
thetahat1<-(1/n)*(1/sqrt(2*pi)*sum(exp(-r1^2/2)))
difference1<-theta[[1]][1] - thetahat1
difference1

## [1] 0.08003201
```

we can notice that using a normal distribution gets our estimate worse. This could be explained by the fact that the Monte Carlo estimation relies on the law of large number; therefore the expected value of the uniform distribution is exactly $\frac{1}{n} * \sum(f(X))$, which is why the uniform provides a better result, by getting closer to the actual expected value.

8 Task 8

To find such δ we need to know σ, thus, we first estimate as MC with $n = 1000$. After finding $E(\theta)$ and $E(\theta^2)$, we can compute $\sigma = E(\theta^2) - E(\theta)^2$ and follow the formula to compute our δ:

$$
\begin{aligned}
&P(-\delta \leq \hat{\theta} - \theta \leq \delta) = \Phi(\frac{\delta\sqrt(n)}{\sigma}) - (1 - \Phi(\frac{\delta\sqrt(n)}{\sigma})) \\
&= 2\Phi(\frac{\delta\sqrt(n)}{\sigma}) - 1 = 0.05
\end{aligned} \tag{1}
$$

```
n= 1000

MYESTIMATE <- function(x,n){
x <- runif(n)
thetah <- mean(cos(2*pi*x))
return(thetah)
}

MYESTIMATE2 <- function(x,n){
x <- runif(n)
thetah <- mean(x*cos(2*pi*x))
return(thetah)
}

mu <- MYESTIMATE(,1000)
var <- MYESTIMATE2(,1000) - mu^2
answer8 <- pnorm(1.05/2)* var /sqrt(n)
```

Thus, $\delta =$

```
answer8

## [1] 8.315388e-05
```

9 Task9:

$U_1, ..., U_n$ independently uniformly distributed RVs on $[0, 1]$
$U_{(n)}$ is maximum

Let' consider CDF of uniformly distributed RV on $[0, 1]$:

$$F_{U_i}(x) = xI_{[0,1)} + I_{[1,\infty]}$$

Let's find the CDF of $U_{(n)}$:
$\mathrm{F}_{U_n}(x) = P(U_{(n)} \leq x) = P(max[U_1, ..., U_n] \leq x) = P(U_1 \leq x, ..., U_n \leq x) = P(U_1 \leq x) \cdot P(U_2 \leq x) \cdot ... \cdot P(U_n \leq x) = \prod_{i=1}^{n} P(U_i \leq x) = x^n I_{[0,1)} + I_{[1,\infty]}$
The density $f(x)$ is just the derivative of $F(x)$. So $f(x) = nx^{n-1}$.

$$E(U_n) = \int_0^1 xf(x)dx = \int_0^1 nx^n dx = \frac{n}{n+1}.$$

$$E(U_n^2) = \int_0^1 x^2 f(x)dx = \int_0^1 nx^{n+1}dx = \frac{n}{n+2}.$$

Now variance:

$$Var(U_n) = E(U_n^2) - E(U_n)^2 = \frac{n}{n+2} - \frac{n^2}{(n+1)^2} = \frac{n}{(n+2)(n+1)^2}.$$

And standard deviation:

$$\sigma_{U_n} = \frac{\sqrt{n}}{\sqrt{(n+2)}(n+1)}.$$

Standardization gives us

$$Z_n = \frac{U_n - \frac{n}{n+1}}{\frac{\sqrt{n}}{\sqrt{n+2}(n+1)}}$$

Since we need to find a limit at $n \to \infty$:

$$Z_n \to \frac{U_n - 1}{\frac{1}{n}}$$

$$F_{Z_n}(x) = P(Z_n \leq x) = P(U_n \leq 1 + \frac{x}{n}) =$$

$$F_{U_n}\left(1 + \frac{x}{n}\right) = (1 + \frac{x}{n})^n I_{[0,1)}(1 + \frac{x}{n}) + I_{[1,\infty]}(1 + \frac{x}{n})$$

Analyzing Indicator function and taking into account $\left(1 + \frac{x}{n}\right)^n \to e^x$ we can conclude:

$$F_{Z_n}(x) = P(Z_n \leq x) \to \begin{cases} e^x, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

10 Task11:

10.1 a)

The moment generating function:

$$m(t) = E(e^{tX})$$

Taylor series gives us the following:

$$m(t) = E(e^{tX}) = 1 + t\mu_1 + \frac{t^2\mu_2}{2!} + \frac{t^3\mu_3}{3!} + \dots + \frac{t^n\mu_n}{n!} + \dots$$

Here μ_n are not central moments.
Cumulant is defined as natural log of moment generating function:

$$k(t) = log(m(t))$$

Now let's calculate the k_1. We need to take the derivative of $log(1 + t\mu_1)$ with respect to t:

$$(log(1 + t\mu_1))' = \frac{\mu_1}{1 + t\mu_1}$$

At $t = 0$ we get :

$$k_1 = \mu_1$$

Now to calculate the k_2 we need the second derivative:

$$(log(1+t\mu_1+\frac{t^2\mu_2}{2}))'' = (\frac{\mu_1 + t\mu_2}{1 + t\mu_1 + \frac{t^2\mu_2}{2}})' = \frac{\mu_2(1 + t\mu_1 + \frac{t^2\mu_2}{2}) - (\mu_1 + t\mu_2)(\mu_1 + t\mu_2)}{(1 + t\mu_1 + \frac{t^2\mu_2}{2})^2}$$

At $t = 0$ we get:

$$k_2 = \mu_2 - (\mu_1)^2$$

3rd and 4th derivative can be calculated using the formula:

$$k_n = \mu_n - \sum_{m=1}^{n-1} \binom{n-1}{m-1} k_m \mu_{n-m}$$

.

From this formula we get:

$$k_3 = \mu_3 - 3\mu_2\mu_1 + 2(\mu_1)^3$$

$$k_4 = \mu_4 - 4\mu_3\mu_1 - 3{\mu_2}^2 + 12\mu_2{\mu_1}^2 - 6{\mu_1}^4$$

10.2 b)

Here we show the k_2, k_3, k_4 through the central moments. Central moments μ_i can be calculated using the next formula:

$$\mu_i = E((x - E(x))^i)$$

$$k_2 = E(x^2) - E(x)^2 = E((x - E(x))^2) = \mu_2$$

It is clear that k_i can be represented by μ_i, so for k_3 and k_4 we firstly calculated μ_3 and μ_4 and then compare with k_3 and k_4 sides.

$$\mu_3 = E((x - E(x))^3) = E(x^3 - 3x^2E(x) + 3xE(x)^2 - E(x)^3) = k_3$$

$$\mu_4 = E((x - E(x))^4) = E(x^4 - 4x^3E(x) + 6x^2E(x)^2 - 4xE(x)^3 + E(x)^4)$$

$$k_4 = \mu_4 - 3E(x^2)^2 + 6E(x^2)E(x)^2 - 3E(x)^4 = \mu_4 - 3\mu_2^2$$

10.3 c)

The skewness of a random variable X is the third standardized moment:

$$Skew(X) = \frac{\mu_3}{\mu_2^{3/2}} = \frac{k_3}{k_2^{3/2}}$$

The kurtosis is the fourth standardized moment, defined as

$$Kurt(X) = \frac{\mu_4}{\mu_2^2} = \frac{k_4 + 3k_2^2}{k_2^2}$$

11 Task12:

The cumulants k_n of a random variable X are defined via the cumulant-generating function $K(t)$, which is the natural logarithm of the moment-generating function.

POISSON DISTRIBUTION

The moment-generating function $m(t) = e^{\lambda(e^t - 1)}$, therefore cumulant generating function $k(t) = log(m(t)) = \lambda(e^t - 1) = E(x)(e^t - 1) = \mu(e^t - 1)$.

$$k_i = k^i(0) = \mu e^t(0) = \mu$$

So: $k_1 = k_2 = k_3 = k_4 = \mu$

NORMAL DISTRIBUTION

The moment-generating function $m(t) = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$, therefore cumulant generating function: $k(t) = \mu t + \sigma^2 t^2/2$.
$k_1 = k^{(1)}(0) = \mu$
$k_2 = k^{(2)}(0) = \sigma^2$
$k_3 = k_4 = 0$

12 Task15:

The standard Gumbel distribution has distributution function

$$F(x) = e^{-e^{-x}}$$

Mode

To find the mode, we maximize the density function setting the derivative equal to 0.

$$f(x) = e^{-x} e^{-e^{-x}}$$

$$f'(x) = -e^{-x} * e^{-e^{-x}} + e^{-2x} e^{-e^{-x}}$$

Setting $f'(x) = 0$ and solving the equation we have: $x_{mode} = 0$.

Median

For median point $F(x_{median}) = 0.5$
Solving the equation

$$F(x) = e^{-e^{-x}} = 0.5$$

we get $x_{median} = -ln[-ln(2)]$

Moment generating function

$$m(t) = E(e^{xt}) = \int_{-\infty}^{\infty} e^{xt} * e^{-x} * e^{-e^{-x}} dx =$$

$$= [e^{-x} = v] = -\int_{\infty}^{0} v^{-t} v e^{-v} \frac{1}{v} dv = \int_{0}^{\infty} v^{-t} e^{-v} dv = \Gamma(1-t)$$

Expectation

As we have seen in exercise 11, the $E(X) = \mu_1$ irrespective to log-form.

$$E(x) = \mu_1 = m'(t)|_{t=0} = (\Gamma(1-t))'\,|_{t=0} =$$

$$= -\frac{\Gamma'(1-t)}{\Gamma(1-t)} = -\psi(1) = \gamma$$

$$E(e^{-x}) = \int_{-\infty}^{\infty} e^{-x} * e^{-e^{-x}} * e^{-x} dx =$$

$$= [e^{-x} = v] = \int_{0}^{\infty} ve^{-v} dv = \Gamma(2) = 1$$

$$\begin{aligned}
E(Xe^{-X}) &= \Gamma(1-t)' \\
&= (-t\Gamma(-t))' \\
&= (-1)(\Gamma(-t)) + (-t)\Gamma'(-t)(-1) \\
&= -1 + \Gamma'(1) = \gamma - 1.
\end{aligned}$$

$$\begin{aligned}
E(X^2e^{-x}) &= ((-1)(\Gamma(-t)) + t\Gamma'(-t))' \\
&= \Gamma'(-t) + \Gamma'(-t) + t\Gamma''(-t)(-1) \\
&= 2\Gamma'(-t) - t\Gamma''(-t) \\
&= -2\gamma + \frac{\pi^2}{6} + \gamma^2
\end{aligned}$$

13 Task16:

We have $F(t) = F_0\left(\frac{t-\mu}{\sigma}\right)$, where $F_0(t) = e^{-e^{-t}}$.

Then we can represent $F(t)$ as

$$F(t) = e^{-e^{-\frac{t-\mu}{\sigma}}}.$$

To get $F^{-1}(t)$ we should solve the equation with respect to t:

$$p = F(t)$$

$$p = e^{-e^{-\frac{t-\mu}{\sigma}}}$$

$$log(p) = -e^{-\frac{t-\mu}{\sigma}}$$

$$-log(p) = e^{-\frac{t-\mu}{\sigma}}$$
$$log(-log(p)) = -\frac{t-\mu}{\sigma}$$
$$\sigma log(-log(p)) = -t + \mu$$
$$\mu - \sigma log(-log(p)) = t = F^{-1}(p)$$

14 Task17:

To find the cumulative distribution function we integrate the density function:

$$F(x|\alpha, \lambda) = \int_0^x \frac{\alpha\lambda^\alpha}{(\lambda + x)^{\alpha+1}} dx =$$
$$= \lambda^\alpha \alpha \int_0^x (\lambda + x)^{-\alpha-1} d(\lambda + x) =$$
$$= \lambda^\alpha \alpha \left[\frac{1}{-\alpha} (\lambda + x)^{-\alpha} \right]_0^x =$$
$$= 1 - (1 + \frac{x}{\lambda})^{-\alpha}$$

For the Lomax distribution we know that:

$$E(X^n) = \frac{\lambda^n \Gamma(\alpha - n)\Gamma(n+1)}{\Gamma(\alpha)}$$

Then

$$E(X) = \frac{\lambda * \Gamma(\alpha - 1)\Gamma(2)}{\Gamma(\alpha)} = \frac{\lambda}{\alpha - 1}$$

$$Var(X) = E(X^2) - E(X)^2 = \frac{\lambda^2 * \Gamma(\alpha - 2)\Gamma(3)}{\Gamma(\alpha)} - (\frac{\lambda}{\alpha - 1})^2 = \frac{\lambda^2 \alpha}{(\alpha - 1)^2(\alpha - 2)}$$