STAT 403 HW2

Chongyi Xu

April 12, 2018

1. Let $X_1, \cdots, X_n$ be IID random points from Beta distribution $\texttt{Beta}(\alpha = 2, \beta = 2)$. The PDF of $\texttt{Beta}(\alpha = 2, \beta = 2)$ is

$$p(x) = 6 \cdot x \cdot (1 - x)$$

for $x \in [0, 1]$ and $p(x) = 0$ outside $[0, 1]$. Let $F(x)$ be the CDF of the $\texttt{Beta}(\alpha = 2, \beta = 2)$. Let $\hat{F_n}(x)$ be the EDF using $X_1, \cdots, X_n$.

(a) What is the CDF of $\texttt{Beta}(\alpha = 2, \beta = 2)$?

With given PDF of $\texttt{Beta}(\alpha = 2, \beta = 2)$, we could found that $B(\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{p(x)} = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{1}{6}$

Therefore, the CDF of the beta distribution is

$$F(x) = \frac{\int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt}{B(2,2)} = 6\int_0^x t(1-t)dt$$

(b) What is the mean and variance of the EDF for a given $x \in [0, 1]$

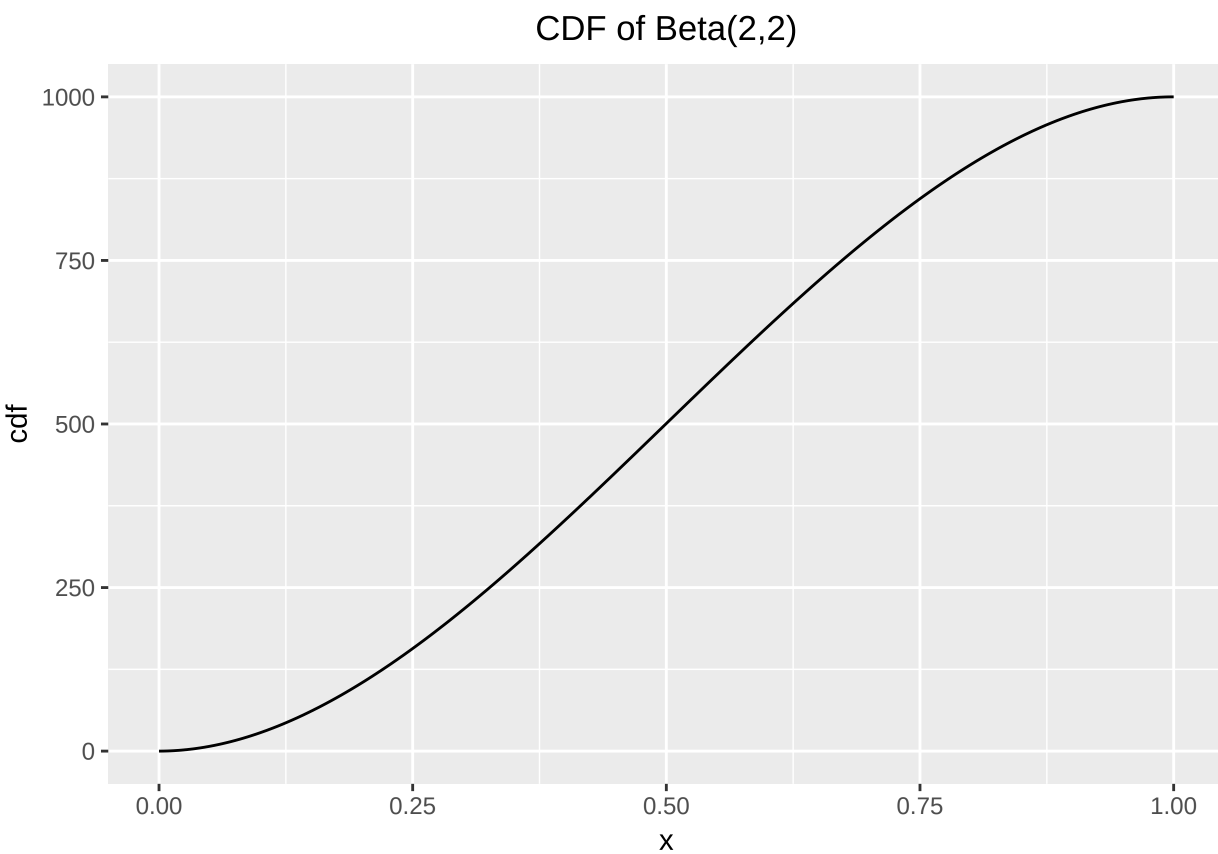
$$\begin{aligned}
\mathbb{E}(\hat{F_n}(x)) &= \mathbb{E}(I(X_i < x)) = F(x) \\
&= 6\int_0^x t(1-t)dt \\
Var(\hat{F_n}(x)) &= \frac{\sum_{i=1}^n Var(I(X_i < x))}{n^2} = \frac{F(x)(1-F(x))}{n} \\
&= \frac{6}{n}(\int_0^x t(1-t)dt - 6(\int_0^x t(1-t)dt)^2)
\end{aligned}$$

(c) Plot the CDF of $\texttt{Beta}(\alpha = 2, \beta = 2)$ within the range $[0, 1]$.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
f <- function(x) {
  x * (1-x)
}

x <- seq(from=0, to=1, by=0.001)
cdf <- rep(0, length(x))
for (i in 1:length(x)) {
  cdf[i] <- 6 * sum(x[1:i]*(1-x[1:i]))
}

ggplot() + geom_line(aes(x, cdf)) + ggtitle('CDF of Beta(2,2)') +
      theme(plot.title=element_text(hjust=0.5))
```



2. Let U be a uniform random variable over $[0, 1]$. We define another random variable $W = -2logU$.

(a) Show that W has the same distribution as $Exp(0.5)$

Consider for the culmulative distribution function for W

$$\begin{aligned} P(W < w) &= P(-2ln(U) < w) \\ &= P(U > e^{-1/2w}) \end{aligned}$$

With given condition U is uniformly distributed between $[0, 1]$

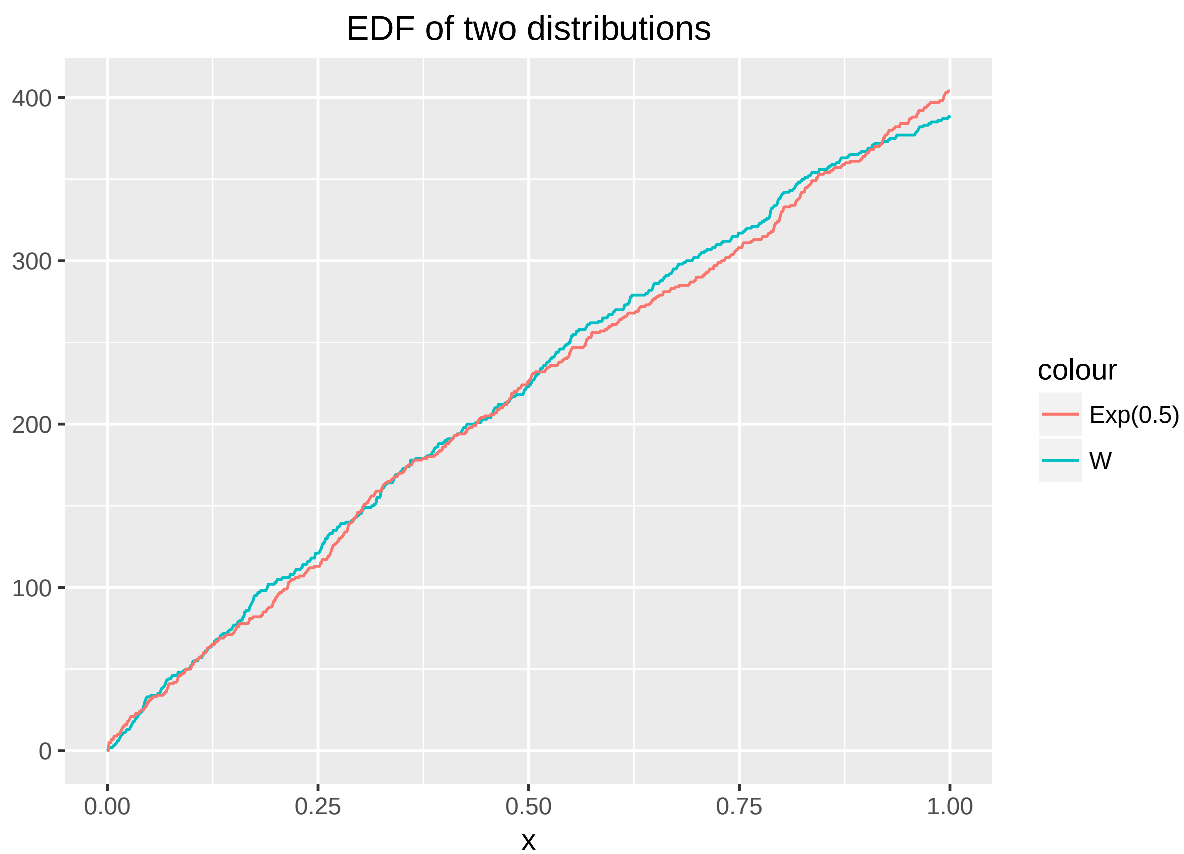
$$\begin{aligned} P(W < w) &= P(u > e^{-1/2w}) \\ &= F(1 - e^{-1/2w}) \end{aligned}$$

which is the CDF of $Exp(1/2)$.

(b) Show that W has the same distribution as $Exp(0.5)$ by simulating realizations.

```
n <- 1000
set.seed(99)

U <- runif(n, min=0, max=1)
W <- -2 * log(U)
E <- rexp(n, rate=0.5)
x <- seq(from=0, to=1, by=0.001)
fn <- matrix(0, 2, length(x))
for (i in 1:length(x)) {
```

```
  fn[1, i] <- sum(W <= x[i])
  fn[2, i] <- sum(E <= x[i])
}

ggplot() + geom_line(aes(x, fn[1,], color='W')) +
  geom_line(aes(x, fn[2,], color='Exp(0.5)')) +
  ylab('') + ggtitle('EDF of two distributions') +
  theme(plot.title=element_text(hjust=0.5))
```



From the plot we can see that two EDFs are generally the same with accpetable variance. So we can conclude that W and $exp(0.5)$ are the same distribution.

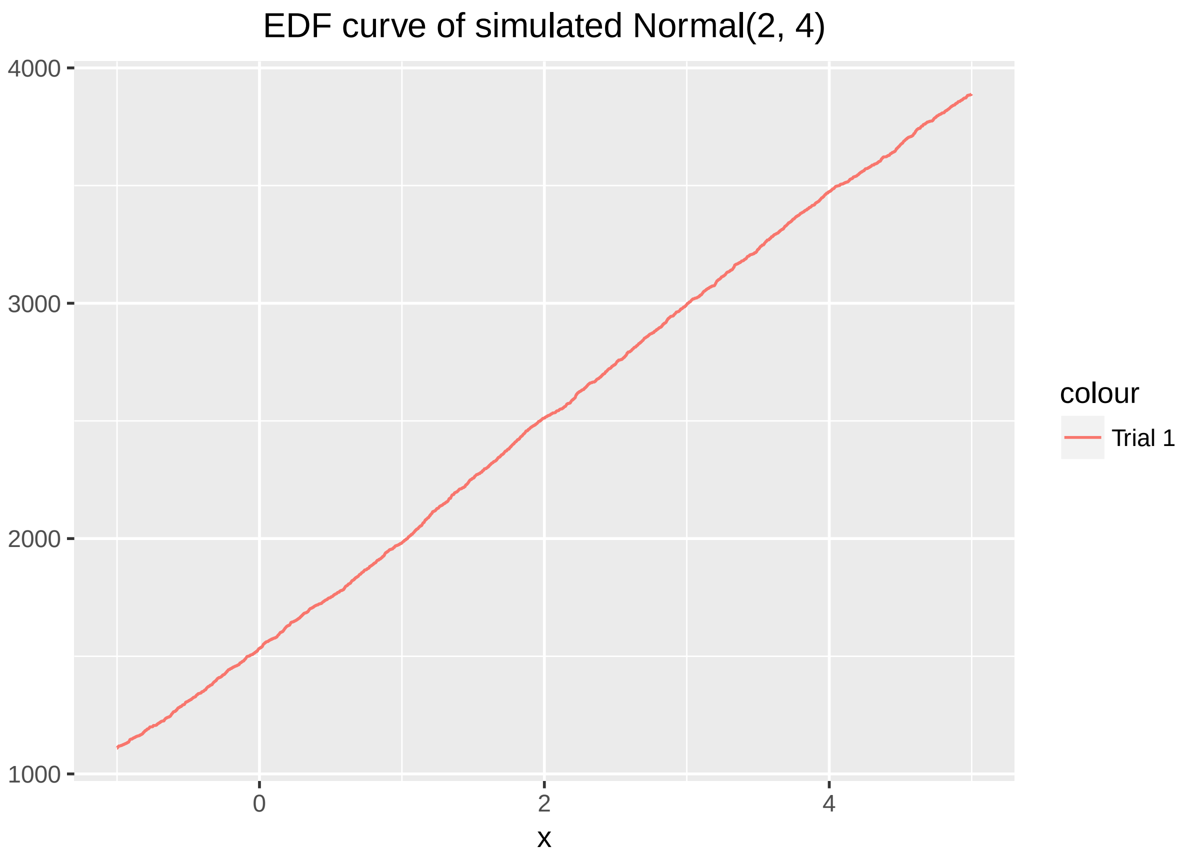
3. Use R to generate 5000 data points from $N(2, 2^2)$.

```
set.seed(123)
n <- 5000

dat <- rnorm(n, mean=2, sd=4)
```

(a) Plot the EDF curve of these data points within [-1, 5]

```
x <- seq(from=-1, to=5, by=0.001)
edf <- matrix(0, 11, length(x))

for (i in 1:length(x)) {
  edf[1, i] = sum(dat <= x[i])
}
```

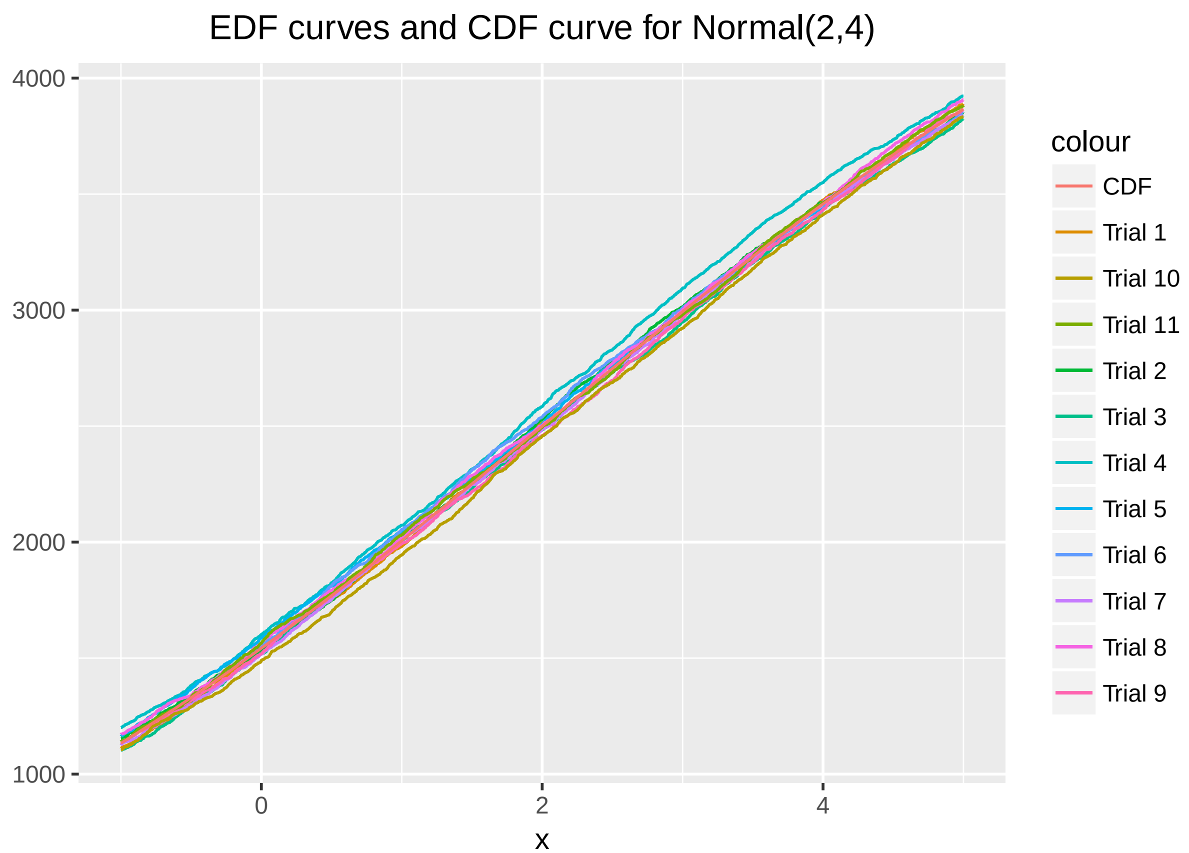
```r
p <- ggplot() + geom_line(aes(x, edf[1,], color='Trial 1')) +
  ylab('') + ggtitle('EDF curve of simulated Normal(2, 4)') +
  theme(plot.title=element_text(hjust=0.5))
p
```



(b) Repeat the above procedure 10 times to generate another 10 EDF curves from the same distribution and same sample size, and attach the actual CDF curve.

```r
for (k in 2:11) {
  dat <- rnorm(n, mean=2, sd=4)
  for (i in 1:length(x)) {
    edf[k, i] = sum(dat <= x[i])
  }
}

cdf <- pnorm(x, mean=2, sd=4) * n

p + geom_line(aes(x, edf[2,], color='Trial 2')) +
  geom_line(aes(x, edf[3,], color='Trial 3')) +
  geom_line(aes(x, edf[4,], color='Trial 4')) +
  geom_line(aes(x, edf[5,], color='Trial 5')) +
  geom_line(aes(x, edf[6,], color='Trial 6')) +
  geom_line(aes(x, edf[7,], color='Trial 7')) +
  geom_line(aes(x, edf[8,], color='Trial 8')) +
  geom_line(aes(x, edf[9,], color='Trial 9')) +
  geom_line(aes(x, edf[10,], color='Trial 10')) +
```

```
geom_line(aes(x, edf[11,], color='Trial 11')) +
geom_line(aes(x, cdf, color='CDF')) +
ggtitle('EDF curves and CDF curve for Normal(2,4)')
```



From the plot above, we can see that the EDF agrees pretty well with the actual CDF curve.

4. Let $X_1, \cdots, X_n$ be an IID random sample from an unknown CDF $F(x)$. Let $\hat{F_n}(x)$ be the EDF. For a fixed point x_0, explain why the following can be used as a $1-\alpha$ confidence interval of $F(x_0)$

$$\hat{F_n}(x_0) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{F_n}(x_0)(1-\hat{F_n}(x_0))}{n}}$$

where z_γ is the γ quantile of $N(0,1)$.

For a random sample, $\hat{F_n}(x_0) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x_0) = \bar{x}$, which could be treated as sample mean. And the term $\sqrt{\frac{\hat{F_n}(x_0)(1-\hat{F_n}(x_0))}{n}} = \sqrt{Var(F(x))} = \sigma$. Therefore,

$$\hat{F_n}(x_0) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{F_n}(x_0)(1-\hat{F_n}(x_0))}{n}} = sample\ mean \pm t\ multiplier\ at\ (1-\alpha) * standard\ error$$

, which is the confidence interval at $(1-\alpha)$ for $F(x_0)$.