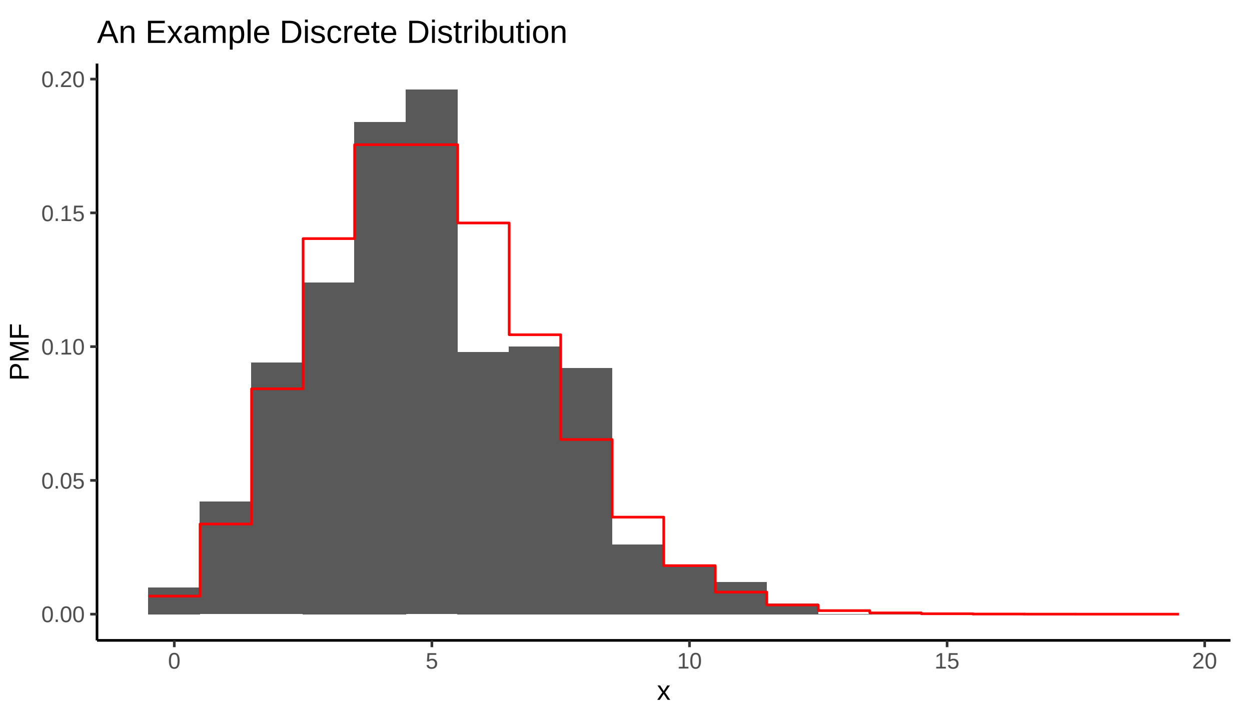
BUDS Training: Transformations Lab

Raven Ico

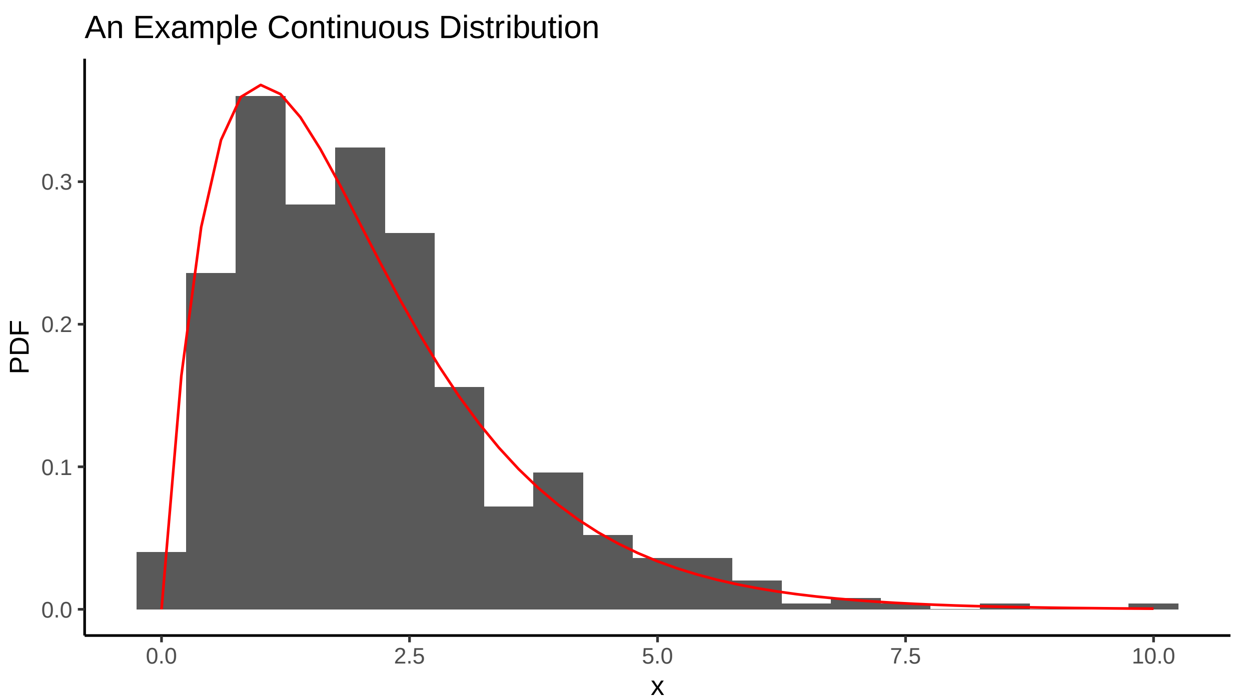
You have now spent some time learning about transformations of random variables. In this lab, we will programatically show relationships between distributions that can also be proved mathematically. To get you started, the following code compares the histogram of a randomly generated (simulated) sample dataset to the probability mass function or the probability density function that was used to generate the data. Notice that the red line generally follows the top of the histogram. The histograms here represent frequencys instead of counts.

```
# An example PMF
sample_df <- tibble(X = rpois(500,5))
pmf_df <- tibble(X = seq(0, 20, 1),
                 pmf = dpois(X, 5))
ggplot(sample_df, aes(x=X)) +
  geom_histogram(aes(y=stat(density)), binwidth = 1) +
  geom_step(data = pmf_df, aes(x=X-0.5, y=pmf), col="red")+
  theme_classic() +
  labs(x="x", y="PMF", title = "An Example Discrete Distribution")
```




```
# An example PDF
sample_df <- tibble(X = rgamma(500, 2, 1))
pdf_df <- tibble(X = seq(0, 10, .2),
                 pdf = dgamma(X, 2, 1))
ggplot(sample_df, aes(x=X)) +
  geom_histogram(aes(y=stat(density)), binwidth = .5) +
  geom_line(data = pdf_df, aes(x=X, y=pdf), col="red")+
  theme_classic() +
```

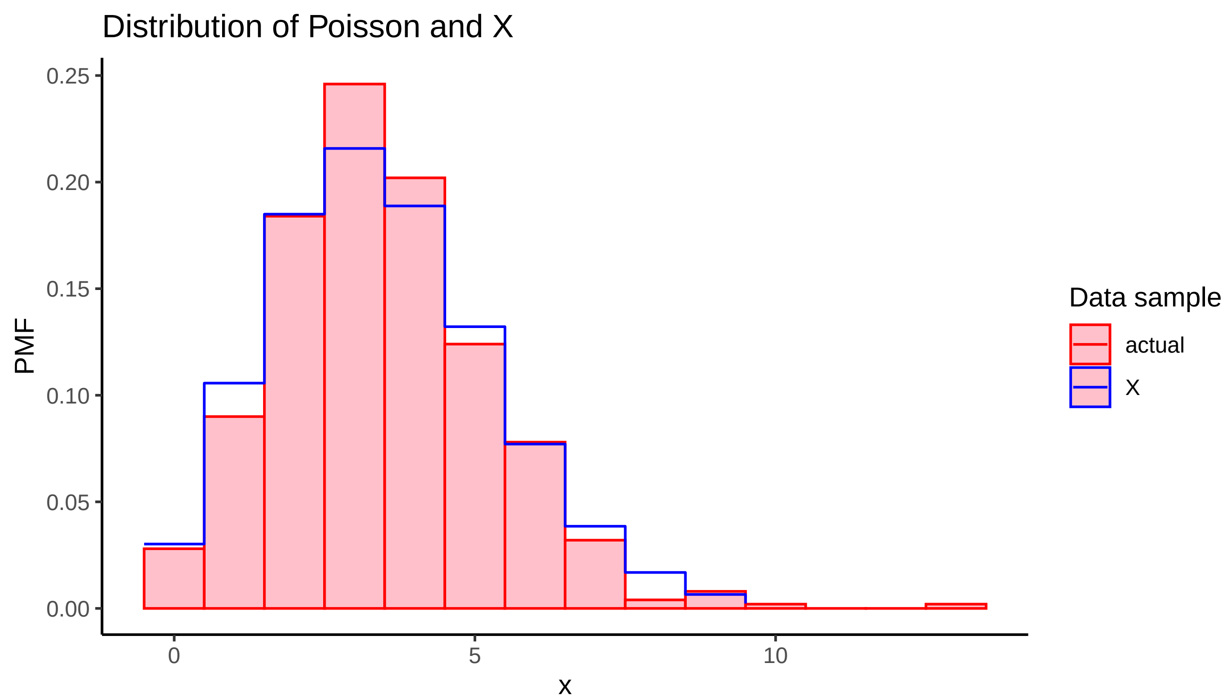
```
  labs(x="x", y="PDF", title = "An Example Continuous Distribution")
```



Problem 1

Recall if $N \sim$ Poisson(λ) and $X|N \sim$ Binomial(N, p) then $X \sim$ Poisson(λp). Pick a λ and p and generate X by first generating an N and then $X|N$. Plot a histogram of X versus the Poisson(λp) distribution.

```
lamda=5
p=0.7
poss=rpois(500,lamda)
sample_df <- tibble(X = rbinom(500,poss,p))
pmf_df <- tibble(X = seq(0, 10, 1),
                 pmf = dpois(X, lamda*p))
ggplot(sample_df, aes(x=X)) +
  geom_histogram(aes(y=stat(density), color='actual'),fill='pink',binwidth = 1) +
  geom_step(data = pmf_df, aes(x=X-0.5, y=pmf, color='X'))+
  theme_classic() +
  labs(x="x", y="PMF", title = "Distribution of Poisson and X") +
  scale_colour_manual(name="Data sample",
    values=c(actual="red", X="blue"))
```



Problem 2

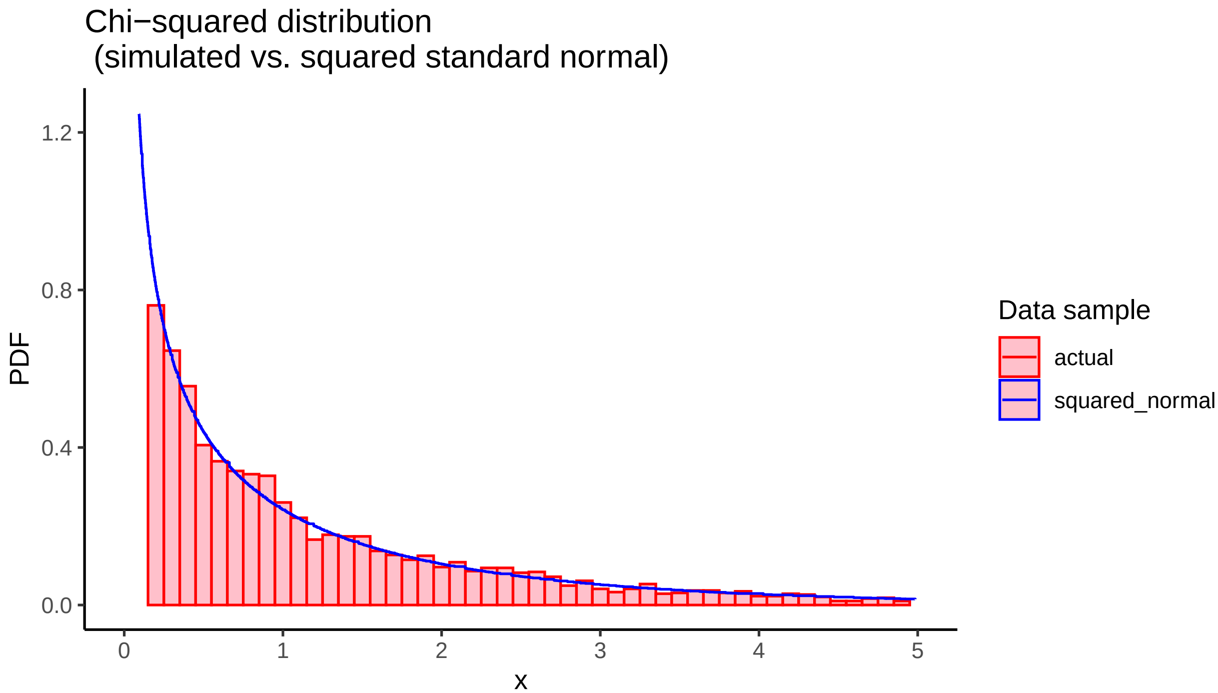
Show via a plot that $Z \sim \text{Normal}(0,1) \implies Z^2 \sim \chi^2_1$. Note: χ^2_1 is the Chi-squared distribution with 1 degree of freedom.

```
norm_df <- tibble(rnorm(1000))
chisq_df2 <- tibble(X = rnorm(1000)^2,
                 pdf = dchisq(X, df=1))
chisq_df <- tibble(X = rchisq(5000,df=1))

ggplot(chisq_df, aes(x=X)) +
  geom_histogram(aes(y=stat(density), color='actual'),binwidth = 1/10, fill='pink') +
  geom_step(data = chisq_df2,aes(x=X, y=pdf,color='squared_normal'))+
  theme_classic() +
  xlim(0, 5)+ ylim(0, 1.25)+
  labs(x="x", y="PDF", title = "Chi-squared distribution \n (simulated vs. squared standard normal)")+
  scale_colour_manual(name="Data sample",
    values=c(actual="red", squared_normal="blue"))
```

```
## Warning: Removed 124 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 3 rows containing missing values (geom_bar).
```



Problem 3

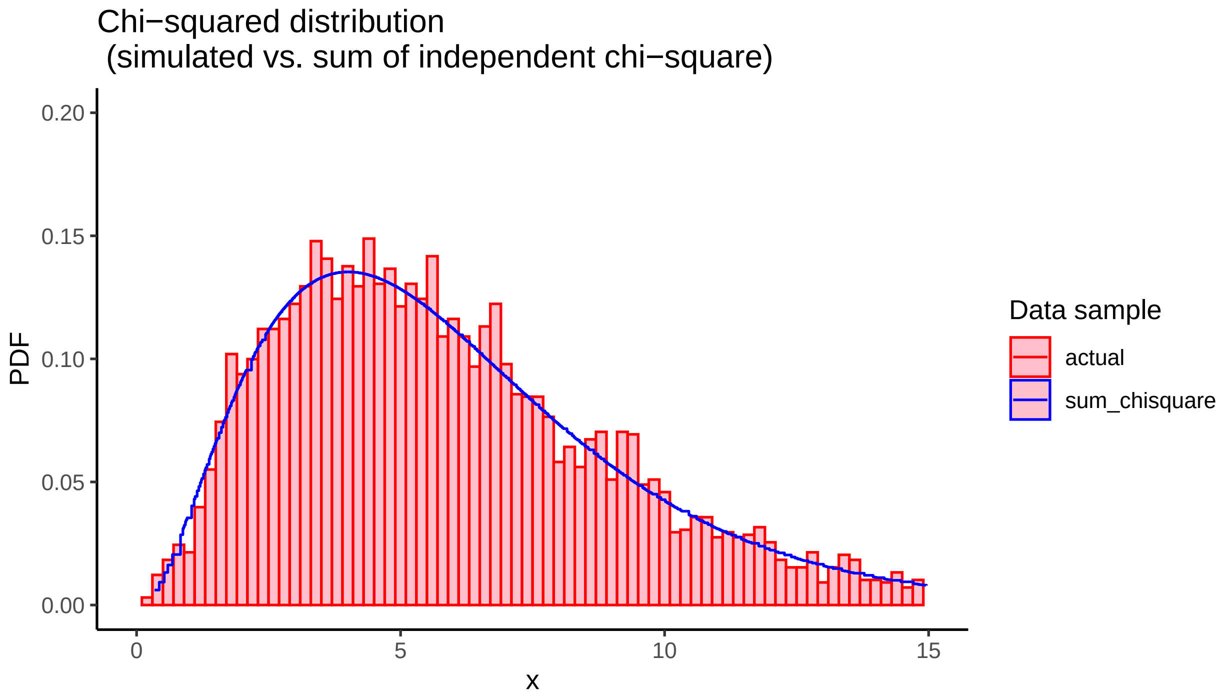
Sum of χ^2: If $U_1, \ldots, U_n$ are independent chi-square random variables with degrees of freedom d_i respectively, then the distribution of $V = U_1 + \cdots + U_n$ is χ^2_d where $d = \sum_{i=1}^{n} d_i$. Show this using 3 chi-square random variables and a plot.

```
chisq_df_sum <- tibble(X = rchisq(1000, df=1) + rchisq(1000, df=2) + rchisq(1000, df=3),
                  pdf = dchisq(X, df=6))
chisq_df <- tibble(X = rchisq(5000,df=6))

ggplot(chisq_df, aes(x=X)) +
  geom_histogram(aes(y=stat(density), color = 'actual'),binwidth = 1/5, fill='pink') +
  geom_step(data = chisq_df_sum,aes(x=X, y=pdf, color = 'sum_chisquare'))+
  theme_classic() +
  xlim(0, 15)+
  ylim(0, 0.2) +
  labs(x="x", y="PDF", title = "Chi-squared distribution \n (simulated vs. sum of independent chi-square
  scale_colour_manual(name="Data sample",
    values=c(actual="red", sum_chisquare="blue"))
```

```
## Warning: Removed 96 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Problem 4

T distribution: Show via a plot that if $Z \sim N(0,1)$ and $U \sim \chi^2_n$ and $Z \perp U$ then the distribution of

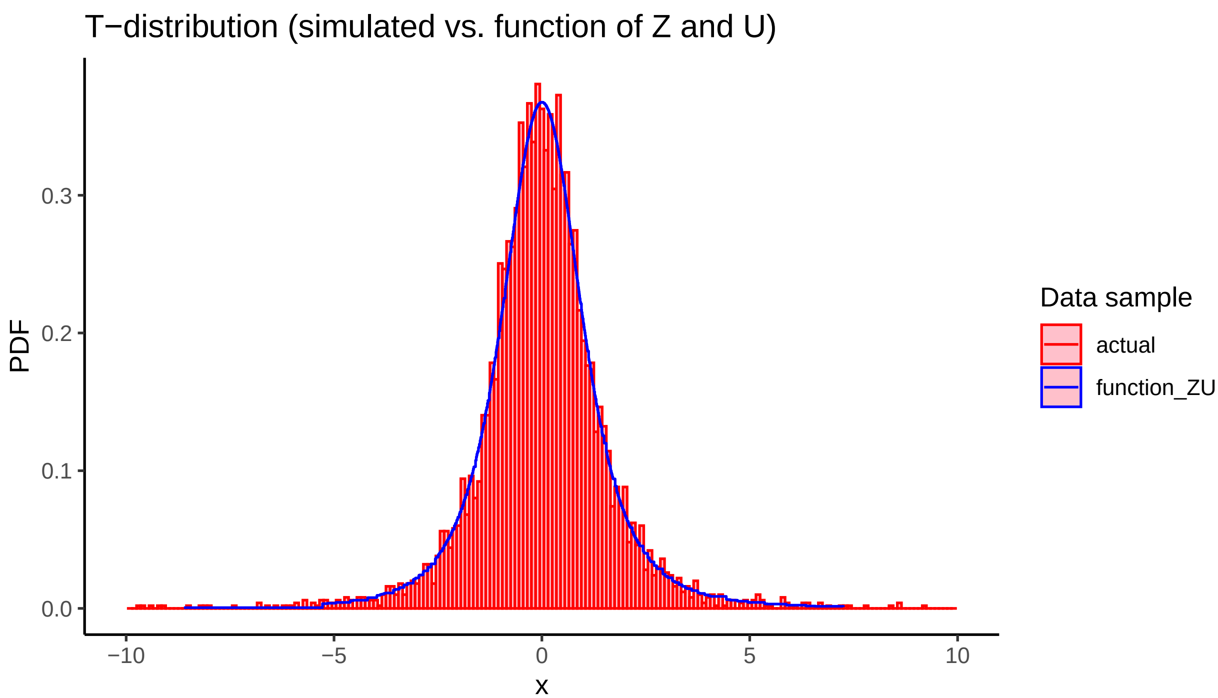
$$\frac{Z}{\sqrt{U/n}}$$

is a t-distribution with n degrees of freedom.

```
n=3
Z = rnorm(1000)
U = rchisq(1000,df=n)
var = Z/sqrt(U/n)
t_dist_func <- tibble(X = var,
                  pdf = dt(X, df=n))
t_dist <- tibble(X = rt(5000, df=n))
ggplot(t_dist, aes(x=X)) +
  geom_histogram(aes(y=stat(density), col='actual'),binwidth = 1/10, fill = 'pink') +
  geom_step(data = t_dist_func,aes(x=X, y=pdf, col = 'function_ZU'))+
  theme_classic() +
  xlim(-10,10) +
  labs(x="x", y="PDF", title = "T-distribution (simulated vs. function of Z and U)")+
  scale_colour_manual(name="Data sample",
    values=c(actual="red", function_ZU="blue"))
```

```
## Warning: Removed 10 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Problem 5

F distribution Show via a plot that if $U \perp V$ are independent chi-square random variables with m and n degrees of freedom, respectively, then the distribution of

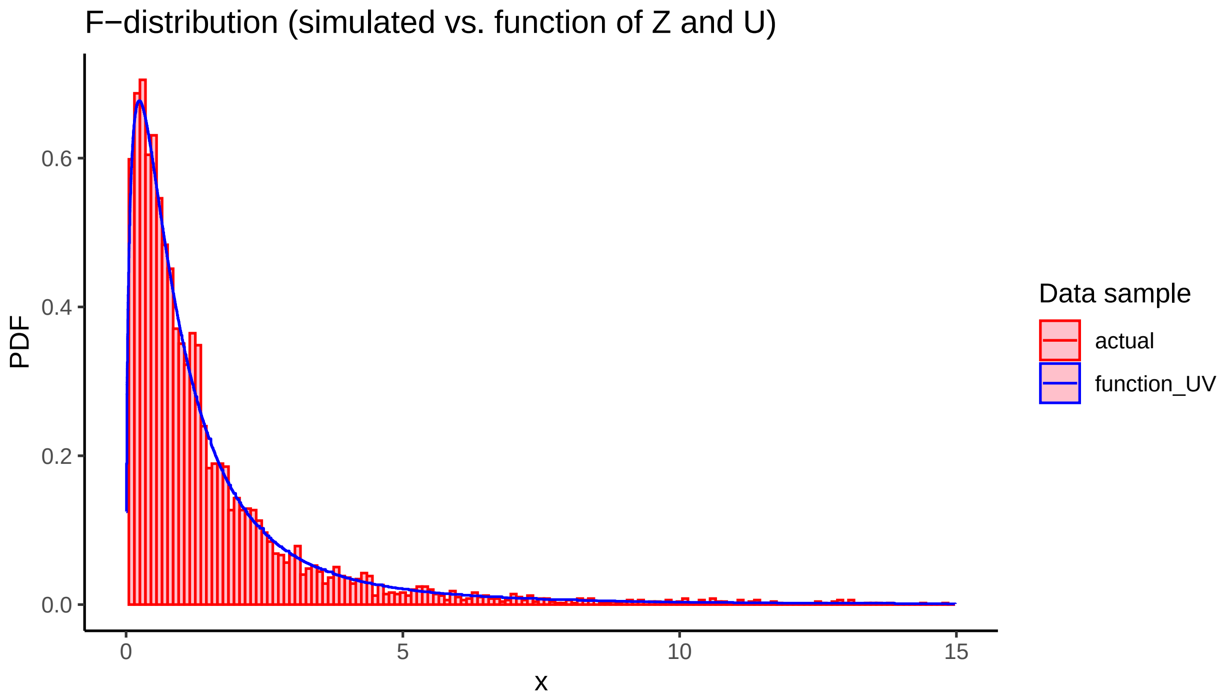
$$W = \frac{U/m}{V/n}$$

is the F distribution with m and n degrees of freedom denoted $F_{m,n}$.

```
m = 3
n = 5
U = rchisq(1000,df=m)
V = rchisq(1000,df=n)
var = (U/m)/(V/n)
f_dist_func <- tibble(X = var,
                 pmf = df(X, df1=m, df2=n))
f_dist <- tibble(X = rf(5000, df1=m, df2=n))
ggplot(f_dist, aes(x=X)) +
  geom_histogram(aes(y=stat(density),col='actual'),fill='pink',binwidth = 1/10) +
  geom_step(data = f_dist_func,aes(x=X, y=pmf, col='function_UV'))+
  theme_classic() +
  xlim(0,15)+
  labs(x="x", y="PMF", title = "An Example Discrete Distribution")+
  labs(x="x", y="PDF", title = "F-distribution (simulated vs. function of Z and U)")+
  scale_colour_manual(name="Data sample",
    values=c(actual="red", function_UV="blue"))
```

```
## Warning: Removed 37 rows containing non-finite values (stat_bin).

## Warning: Removed 2 rows containing missing values (geom_bar).
```


F−distribution (simulated vs. function of Z and U)
PDF
0.0
0.2
0.4
0.6
0
5
10
15
x
Data sample
actual
function_UV