```r
## R demo for Oct 19
## Plotting functions and histograms, F distribution,
## ANOVA tables, F tests, MLR with categorical variables
```

Evaluate the function at many x values, then plot it.

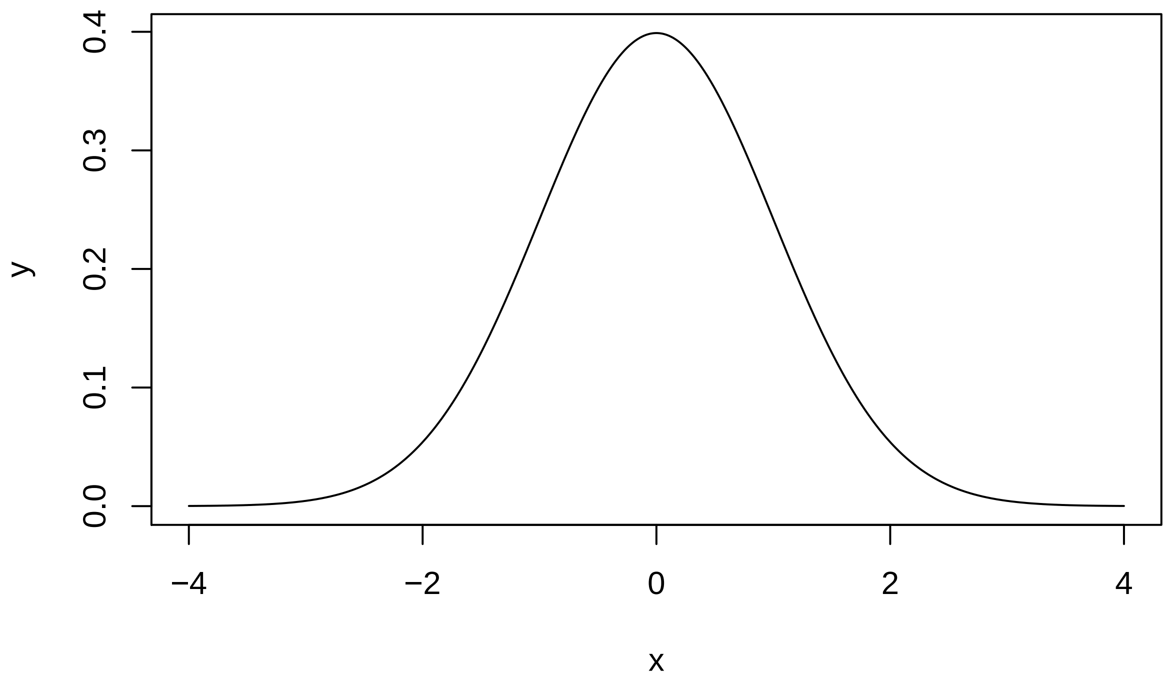
```r
# Plotting functions (e.g., probability density functions)
# Create sequence from -4 to 4 increasing 0.01 each time.
x <- seq(-4, 4, 0.01)
head(x)
```

```
## [1] -4.00 -3.99 -3.98 -3.97 -3.96 -3.95
```

```r
# Normal probability density function with mean 0, and standard deviation 1.
y <- dnorm(x, 0, 1)
```

`dnorm` is for density normal.

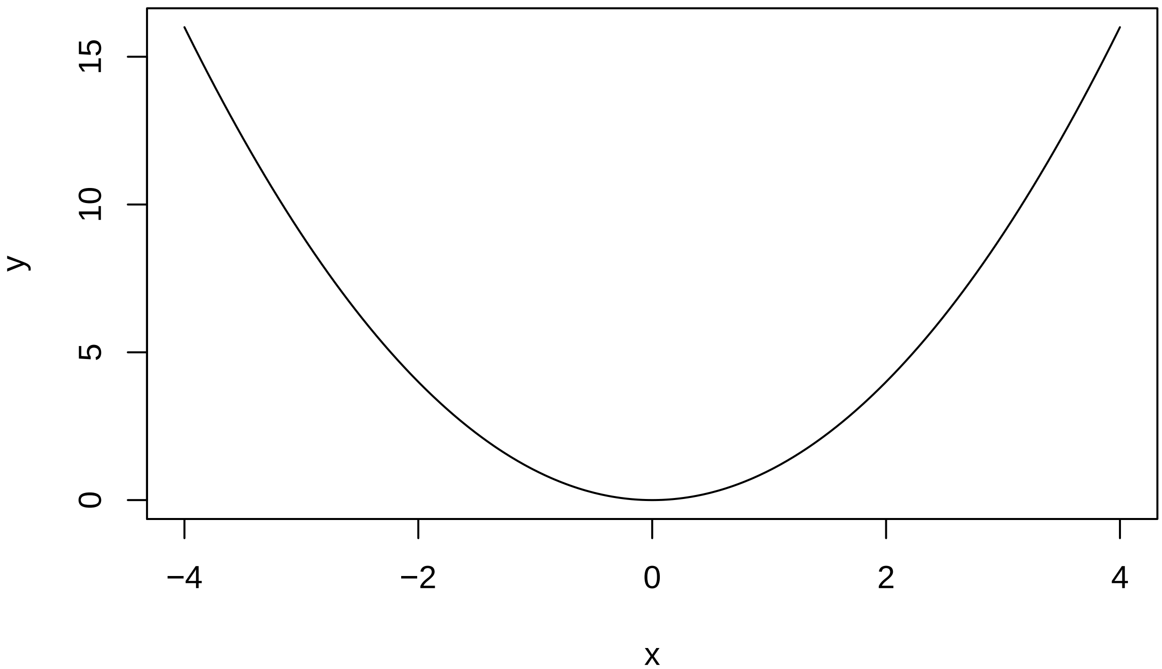
```r
plot(x, y, type = "l")
```



`type = "l"` is for a smooth line (instead of dots).

We can also plot $y = x^2$ for example.

```r
y <- x ^ 2
plot(x, y, type = "l")
```

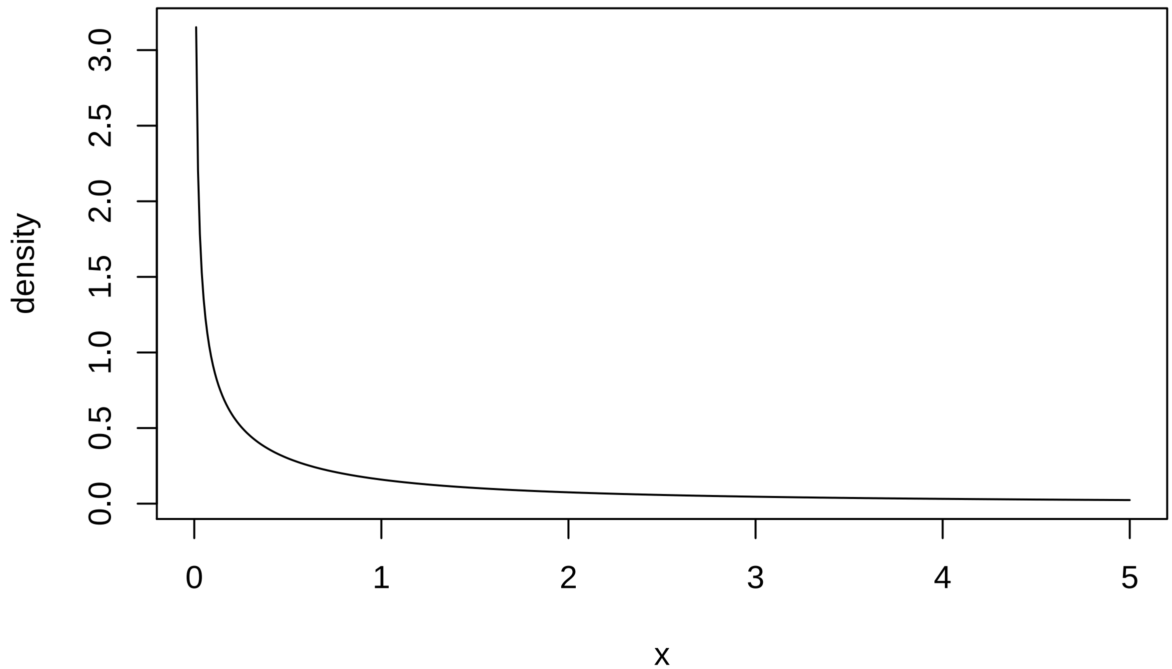


F-distribution Examples

```
x <- seq(0,5,0.01)
head(x)
```

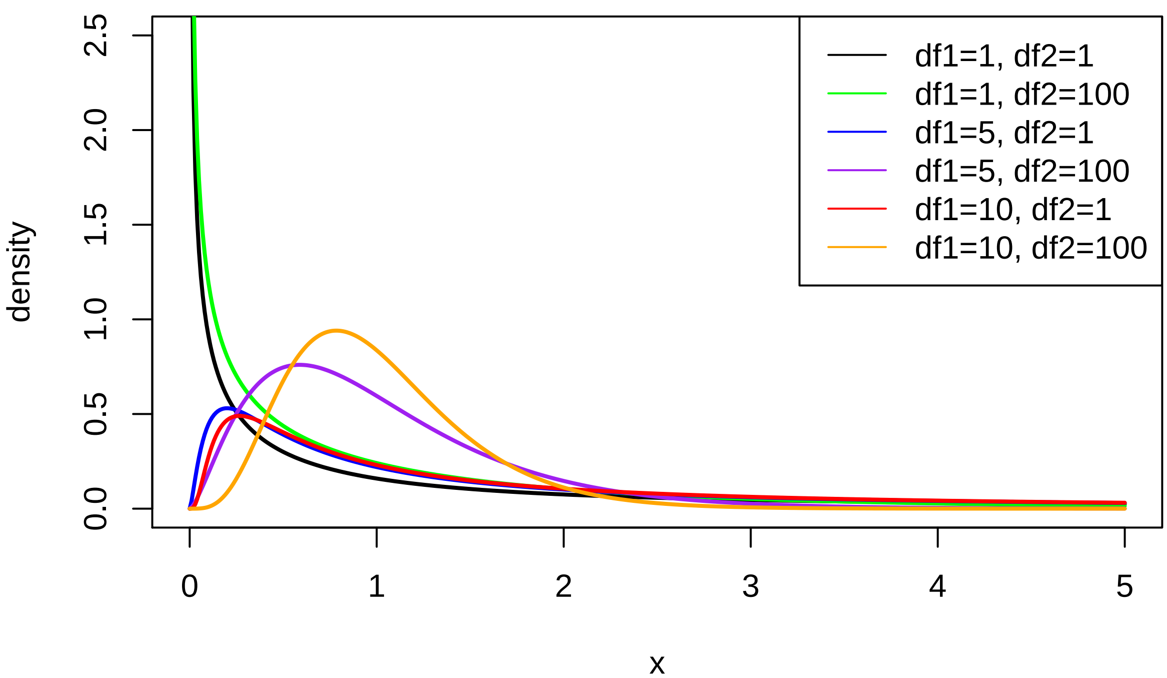
```
## [1] 0.00 0.01 0.02 0.03 0.04 0.05
```

```
# df is degrees of freedom.
# type = "l" is for a smooth curve
plot(
  x,
  y = df(x, df1 = 1, df2 = 1),
  type = "l",
  xlab = "x",
  ylab = "density"
)
```



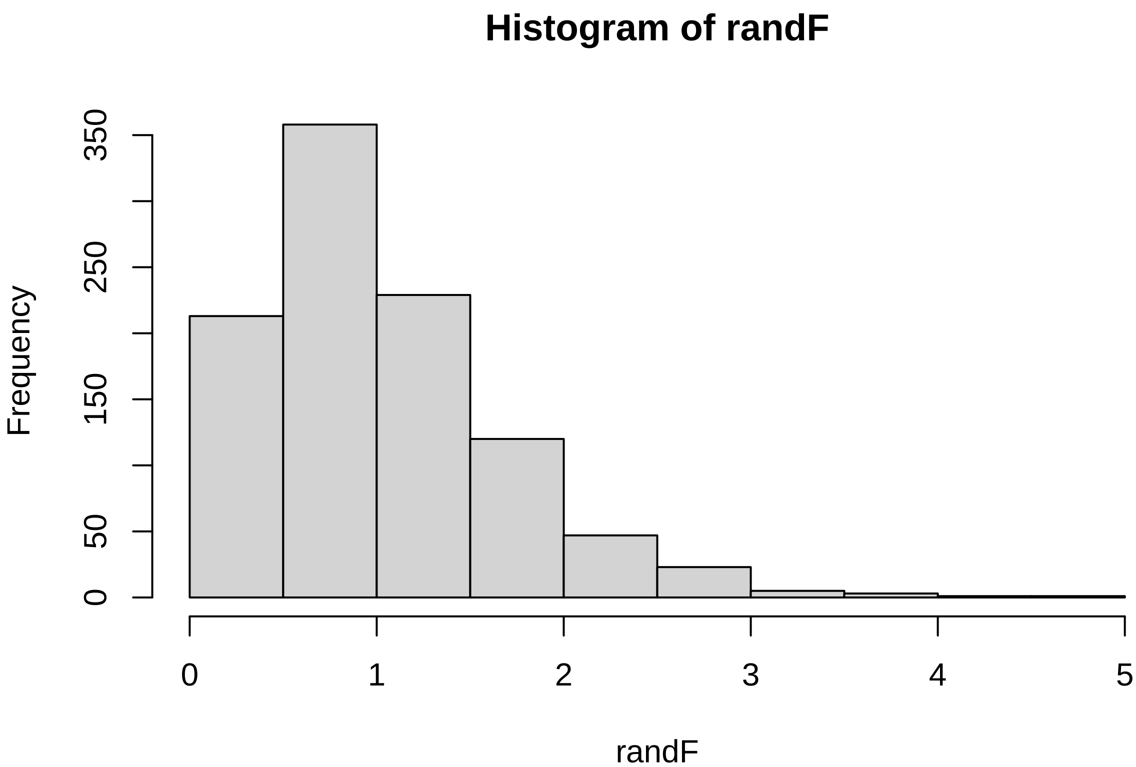

```r
# ylim is for the y-axis limits
# lwd is for line width
plot(
  x,
  y = df(x, df1 = 1, df2 = 1),
  type = "l",
  col = "black",
  xlab = "x",
  ylab = "density",
  ylim = c(0, 2.5),
  lwd = 2
)
# Add lines to the existing plot.
lines(
  x,
  y = df(x, df1 = 1, df2 = 100),
  type = "l",
  col = "green",
  lwd = 2
)
lines(
  x,
  y = df(x, df1 = 5, df2 = 1),
  type = "l",
  col = "blue",
  lwd = 2
)
lines(
  x,
  y = df(x, df1 = 5, df2 = 100),
  type = "l",
  col = "purple",
```

```
  lwd = 2
)
lines(
  x,
  y = df(x, df1 = 10, df2 = 1),
  type = "l",
  col = "red",
  lwd = 2
)
lines(
  x,
  y = df(x, df1 = 10, df2 = 100),
  type = "l",
  col = "orange",
  lwd = 2
)
# Add a legend to the top-right.
# lty = 1 is for a straight solid line.
legend(
  "topright",
  legend = c(
    "df1=1, df2=1",
    "df1=1, df2=100",
    "df1=5, df2=1",
    "df1=5, df2=100",
    "df1=10, df2=1",
    "df1=10, df2=100"
  ),
  lty = 1,
  col = c("black", "green", "blue", "purple", "red", "orange")
)
```

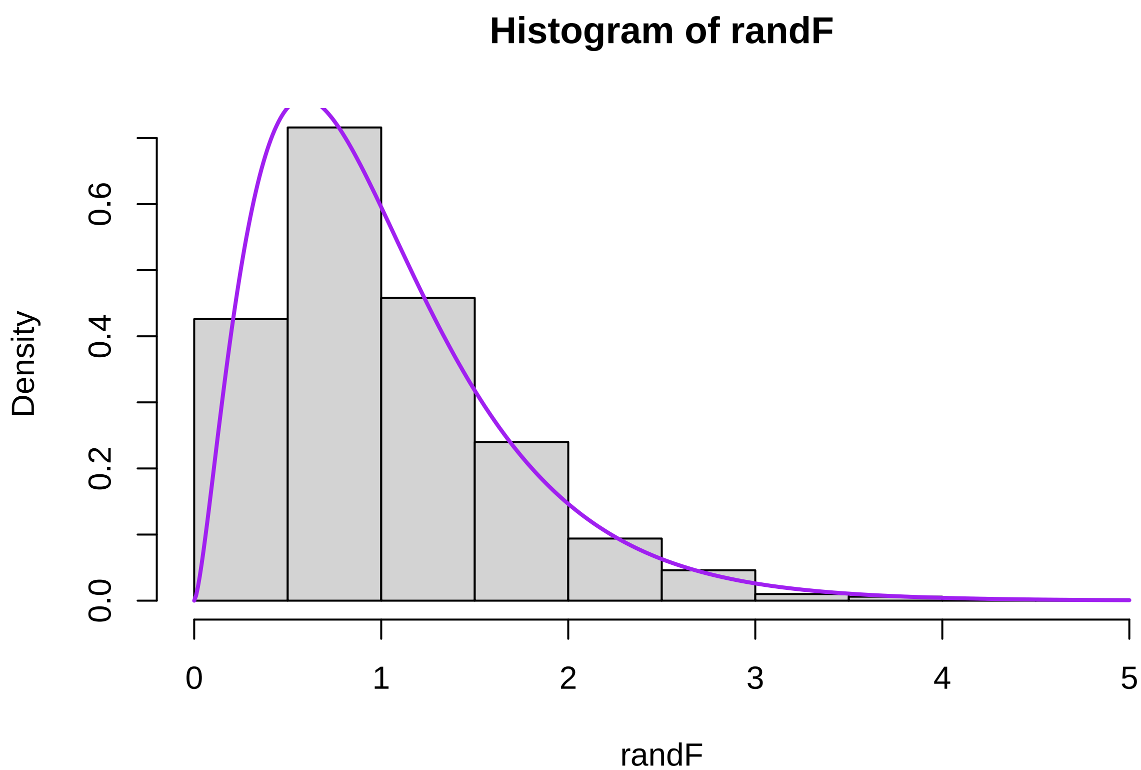


Random numbers for the F-distribution

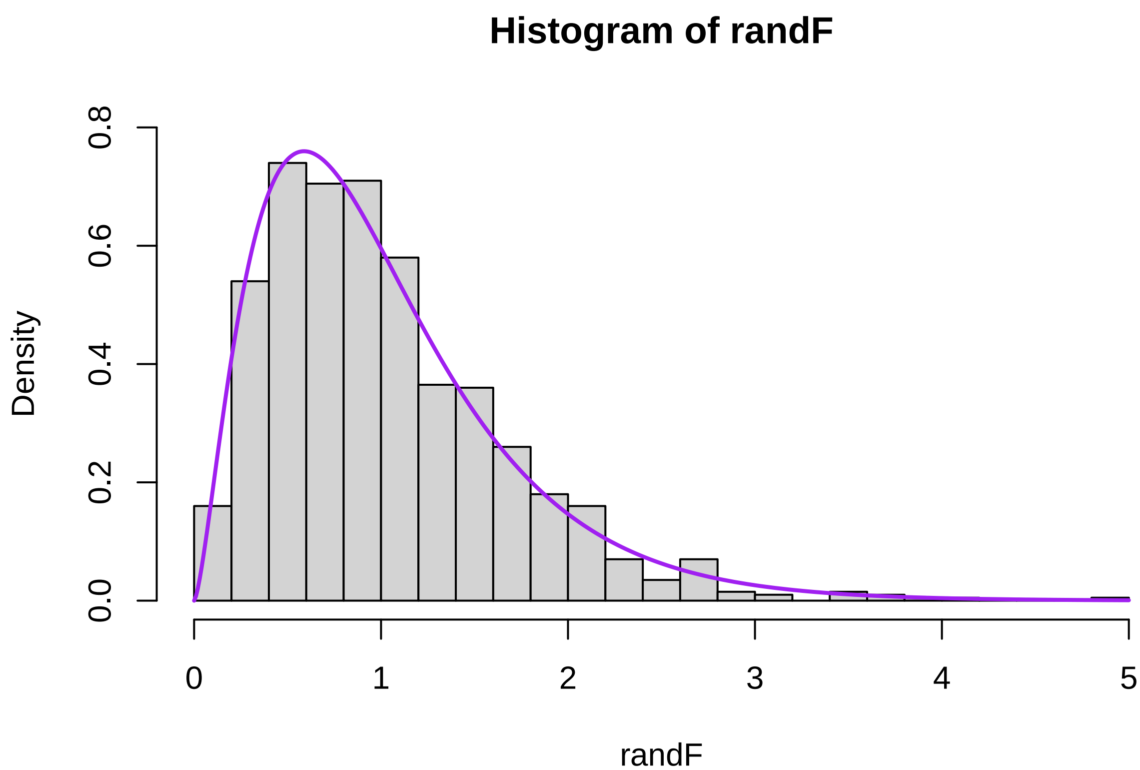
```r
# set.seed allows for exact reproduction.
set.seed(12345678)
randF <- rf(1000, 5, 100)
# Generate histogram for the random numbers with exact.
hist(randF)
```



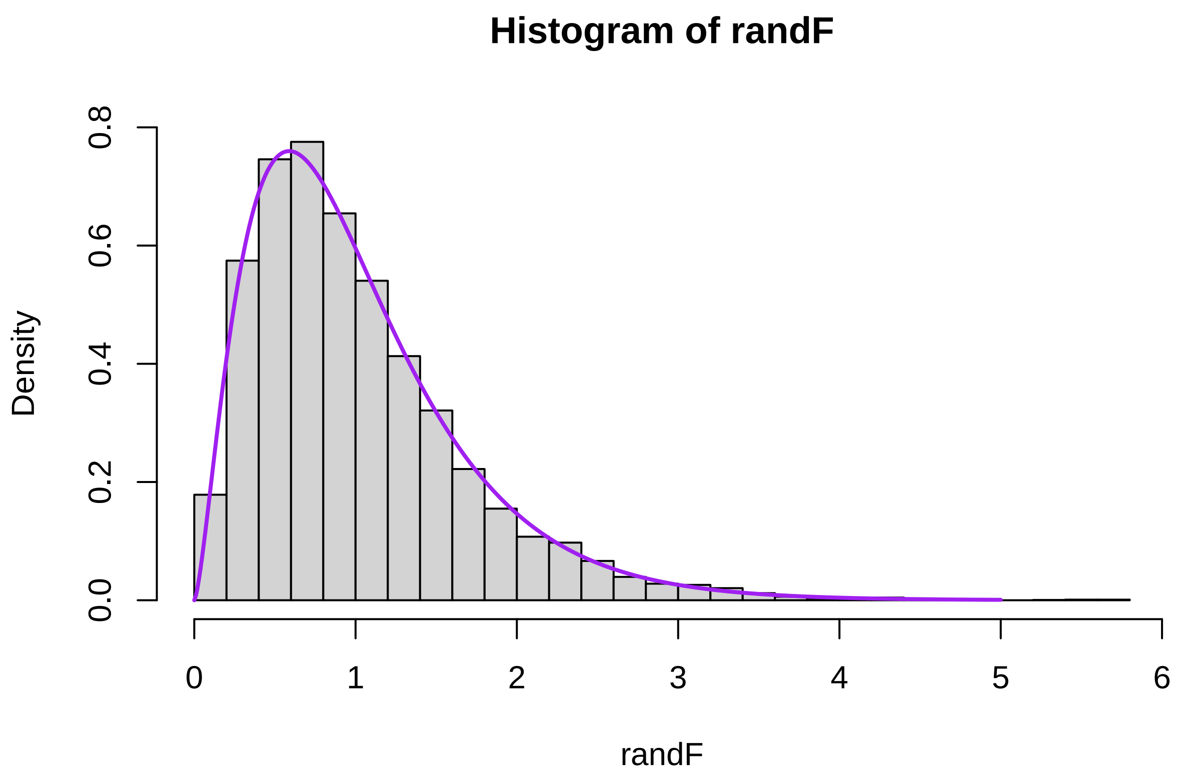

```r
# Generate histogram for the random numbers with relative frequency.
# This is normalized, so we can superimpose an F-distribution to it.
hist(randF, freq = FALSE)
# Superimpose an F-distribution on the histogram.
lines(
  x,
  y = df(x, df1 = 5, df2 = 100),
  type = "l",
  col = "purple",
  lwd = 2
)
```




```r
# Set y-axis limits and more detailed histogram bins using 'breaks = 25'
hist(randF,
     freq = FALSE,
     ylim = c(0, 0.8),
     breaks = 25)
lines(
  x,
  y = df(x, df1 = 5, df2 = 100),
  type = "l",
  col = "purple",
  lwd = 2
)
```




```r
# Generate more random F-distributions to get closer to the 'true' density.
randF <- rf(10000, 5, 100)
hist(randF,
     freq = FALSE,
     ylim = c(0, 0.8),
     breaks = 25)
lines(
  x,
  y = df(x, df1 = 5, df2 = 100),
  type = "l",
  col = "purple",
  lwd = 2
)
```



Revisit Rocket Example

```
rocket <- read.csv("csv/rocket.csv")
m1 <- lm(thrust ~ nozzle + propratio, data = rocket)
summary(m1)
```

```
##
## Call:
## lm(formula = thrust ~ nozzle + propratio, data = rocket)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8459 -1.7555  0.5934  1.2906  3.3008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 473.6039     4.7158 100.430 4.88e-15 ***
## nozzle       16.7383     1.5329  10.919 1.71e-06 ***
## propratio    -1.0948     0.9414  -1.163    0.275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.655 on 9 degrees of freedom
## Multiple R-squared:  0.9303, Adjusted R-squared:  0.9148
## F-statistic: 60.05 on 2 and 9 DF,  p-value: 6.238e-06
```

```
# Compare summary with ANOVA table on board from Oct. 5.
anova(m1)
```

```
## Analysis of Variance Table
##
```

```
## Response: thrust
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## nozzle     1 836.67  836.67 118.7377 1.743e-06 ***
## propratio  1   9.53    9.53   1.3524    0.2748
## Residuals  9  63.42    7.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m1)$`Sum Sq`
```

```
## [1] 836.670000   9.529332  63.417335
```

```r
sum(anova(m1)$`Sum Sq`[1:2])
```

```
## [1] 846.1993
```

```r
SSRes <- anova(m1)$`Sum Sq`[3]

# Test of overall significance.
m_red <- lm(thrust ~ 1, data = rocket)
summary(m_red)
```

```
##
## Call:
## lm(formula = thrust ~ 1, data = rocket)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4167  -7.1167  -0.2167   8.2333  11.3833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  476.617      2.625   181.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.094 on 11 degrees of freedom
```

```r
anova(m_red)
```

```
## Analysis of Variance Table
##
## Response: thrust
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 11 909.62  82.692
```

```r
SSRes_A <- anova(m_red)$`Sum Sq`[1]

# Manually calculate F-statistic.
l <- 2
n <- nrow(rocket)
p <- 2
Fstat <- ((SSRes_A - SSRes) / l) / (SSRes / (n - p - 1))
Fstat
```

```
## [1] 60.04505
```

```r
pval <- 1 - pf(Fstat, df1 = 1, df2 = n - p - 1)
pval
```

```
## [1] 6.238398e-06
```

```r
# Automatically calculate F-statistic.
anova(m1, m_red)$F[2]
```

```
## [1] 60.04505
```

Revist Coffee Example (Coffee Quality Institute, 2018)

```r
coffee <- read.csv("csv/coffee_arabica.csv")

mfull <-
  lm(
    Flavor ~ factor(Processing.Method) + Aroma + Aftertaste +
      Body + Acidity + Balance + Sweetness + Uniformity + Moisture,
    dat = coffee
  )
summary(mfull)
```

```
##
## Call:
## lm(formula = Flavor ~ factor(Processing.Method) + Aroma + Aftertaste +
##     Body + Acidity + Balance + Sweetness + Uniformity + Moisture,
##     data = coffee)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68587 -0.08465  0.00079  0.08910  0.63633
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                   -0.728757   0.168516  -4.325
## factor(Processing.Method)Semi-washed / Semi-pulped -0.001396   0.022021  -0.063
## factor(Processing.Method)Washed / Wet         -0.033061   0.011024  -2.999
## Aroma                                          0.220302   0.020447  10.774
## Aftertaste                                     0.468759   0.023912  19.603
## Body                                           0.096140   0.024334   3.951
## Acidity                                        0.216751   0.021194  10.227
## Balance                                        0.046806   0.022558   2.075
## Sweetness                                      0.025507   0.010150   2.513
## Uniformity                                     0.016297   0.009803   1.663
## Moisture                                       0.169012   0.102480   1.649
##                                               Pr(>|t|)
## (Intercept)                                   1.67e-05 ***
## factor(Processing.Method)Semi-washed / Semi-pulped  0.94947
## factor(Processing.Method)Washed / Wet          0.00277 **
## Aroma                                          < 2e-16 ***
## Aftertaste                                     < 2e-16 ***
## Body                                          8.28e-05 ***
## Acidity                                        < 2e-16 ***
## Balance                                        0.03823 *
## Sweetness                                      0.01211 *
## Uniformity                                     0.09669 .
```

```
## Moisture                                            0.09938 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.148 on 1108 degrees of freedom
## Multiple R-squared:  0.8091, Adjusted R-squared:  0.8073
## F-statistic: 469.5 on 10 and 1108 DF,  p-value: < 2.2e-16
```

```
anova(mfull)
```

```
## Analysis of Variance Table
##
## Response: Flavor
##                              Df Sum Sq Mean Sq   F value    Pr(>F)
## factor(Processing.Method)     2  2.313   1.156   52.8096 < 2.2e-16 ***
## Aroma                         1 67.258  67.258 3071.2889 < 2.2e-16 ***
## Aftertaste                    1 29.097  29.097 1328.6722 < 2.2e-16 ***
## Body                          1  1.129   1.129   51.5460  1.28e-12 ***
## Acidity                       1  2.522   2.522  115.1618 < 2.2e-16 ***
## Balance                       1  0.116   0.116    5.2963 0.0215553 *
## Sweetness                     1  0.251   0.251   11.4392 0.0007442 ***
## Uniformity                    1  0.064   0.064    2.9154 0.0880167 .
## Moisture                      1  0.060   0.060    2.7200 0.0993839 .
## Residuals                  1108 24.264   0.022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSRes <- anova(mfull)$`Sum Sq`[10]

# Reduced model without Uniformity and Moisture (beta9=beta10=0):
m_red <-
  lm(
    Flavor ~ factor(Processing.Method) + Aroma + Aftertaste +
      Body + Acidity + Balance + Sweetness,
    dat = coffee
  )
summary(m_red)
```

```
##
## Call:
## lm(formula = Flavor ~ factor(Processing.Method) + Aroma + Aftertaste +
##     Body + Acidity + Balance + Sweetness, data = coffee)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67907 -0.08487  0.00054  0.08490  0.64763
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                   -0.606791   0.159741  -3.799
## factor(Processing.Method)Semi-washed / Semi-pulped  0.002275   0.021969   0.104
## factor(Processing.Method)Washed / Wet         -0.031115   0.011009  -2.826
## Aroma                                          0.221362   0.020472  10.813
## Aftertaste                                     0.470849   0.023858  19.735
## Body                                           0.087671   0.024102   3.637
```

```
## Acidity                                          0.219257   0.021182  10.351
## Balance                                          0.047526   0.022283   2.133
## Sweetness                                        0.032406   0.009597   3.377
##                                                 Pr(>|t|)
## (Intercept)                                     0.000153 ***
## factor(Processing.Method)Semi-washed / Semi-pulped 0.917539
## factor(Processing.Method)Washed / Wet           0.004795 **
## Aroma                                            < 2e-16 ***
## Aftertaste                                       < 2e-16 ***
## Body                                            0.000288 ***
## Acidity                                          < 2e-16 ***
## Balance                                         0.033160 *
## Sweetness                                       0.000759 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1482 on 1110 degrees of freedom
## Multiple R-squared:  0.8081, Adjusted R-squared:  0.8067
## F-statistic: 584.2 on 8 and 1110 DF,  p-value: < 2.2e-16
```

```r
anova(m_red)
```

```
## Analysis of Variance Table
##
## Response: Flavor
##                             Df Sum Sq Mean Sq  F value    Pr(>F)
## factor(Processing.Method)    2  2.313   1.156   52.637 < 2.2e-16 ***
## Aroma                        1 67.258  67.258 3061.263 < 2.2e-16 ***
## Aftertaste                   1 29.097  29.097 1324.335 < 2.2e-16 ***
## Body                         1  1.129   1.129   51.378 1.387e-12 ***
## Acidity                      1  2.522   2.522  114.786 < 2.2e-16 ***
## Balance                      1  0.116   0.116    5.279 0.0217690 *
## Sweetness                    1  0.251   0.251   11.402 0.0007591 ***
## Residuals                 1110 24.387   0.022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
SSRes_A <- anova(m_red)$`Sum Sq`[8]

# Manually calculate F-statistic.
l <- 2
n <- nrow(coffee)
p <- 10
Fstat <- ((SSRes_A - SSRes) / l) / (SSRes / (n - p - 1))
Fstat
```

```
## [1] 2.81769
```

```r
pval <- 1 - pf(Fstat, df1 = l, df2 = n - p - 1)
pval
```

```
## [1] 0.06017197
```

```r
# Automatically calculate F-statistic.
anova(mfull, m_red)$F[2]
```

```
## [1] 2.81769
```

```r
# Reduced model without Uniformity and Moisture and
# setting effect of Dry = Semi (beta1=beta9=beta10=0)
# 1 = wet, 0 otherwise
coffee$method2 <- ifelse(coffee$Processing.Method %in%
                           c('Natural / Dry', 'Semi-washed / Semi-pulped'),
                         0,
                         1)
# 1 = semi/dry, 0 o.w
coffee$wet <-
  ifelse(coffee$Processing.Method == 'Washed / Wet', 0, 1)

m_red2 <- lm(Flavor ~ method2 + Aroma + Aftertaste +
               Body + Acidity + Balance + Sweetness,
             dat = coffee)
summary(m_red2)
```

```
##
## Call:
## lm(formula = Flavor ~ method2 + Aroma + Aftertaste + Body + Acidity +
##     Balance + Sweetness, data = coffee)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67906 -0.08508  0.00052  0.08490  0.64722
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.606597   0.159659  -3.799 0.000153 ***
## method2     -0.031543   0.010200  -3.092 0.002036 **
## Aroma        0.221408   0.020458  10.823  < 2e-16 ***
## Aftertaste   0.470861   0.023847  19.745  < 2e-16 ***
## Body         0.087561   0.024068   3.638 0.000287 ***
## Acidity      0.219266   0.021173  10.356  < 2e-16 ***
## Balance      0.047527   0.022273   2.134 0.033077 *
## Sweetness    0.032462   0.009577   3.389 0.000725 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1482 on 1111 degrees of freedom
## Multiple R-squared:  0.8081, Adjusted R-squared:  0.8069
## F-statistic: 668.3 on 7 and 1111 DF,  p-value: < 2.2e-16
```

```r
anova(m_red2)
```

```
## Analysis of Variance Table
##
## Response: Flavor
##              Df Sum Sq Mean Sq   F value    Pr(>F)
## method2       1  2.313   2.313  105.3648 < 2.2e-16 ***
## Aroma         1 67.255  67.255 3063.8526 < 2.2e-16 ***
## Aftertaste    1 29.100  29.100 1325.6571 < 2.2e-16 ***
## Body          1  1.126   1.126   51.3088 1.434e-12 ***
## Acidity       1  2.522   2.522  114.9115 < 2.2e-16 ***
## Balance       1  0.116   0.116    5.2882 0.0216552 *
```

```
## Sweetness     1  0.252   0.252   11.4883 0.0007249 ***
## Residuals  1111 24.388   0.022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
SSRes_A <- anova(m_red2)$`Sum Sq`[8]

## Manually calculate F-statistic.
l <- 3
n <- nrow(coffee)
p <- 10
Fstat <- ((SSRes_A - SSRes) / l) / (SSRes / (n - p - 1))
Fstat
```

```
## [1] 1.882046
```

```r
pval <- 1 - pf(Fstat, df1 = l, df2 = n - p - 1)
pval
```

```
## [1] 0.1308207
```

```r
# Automatically calculate F-statistic.
anova(mfull, m_red2)$F[2]
```

```
## [1] 1.882046
```