Tutorial 3: Comparisons and Inference

Jan Vogler (jan.vogler@duke.edu)

September 11, 2015

Question: sometimes R gives you output like this:

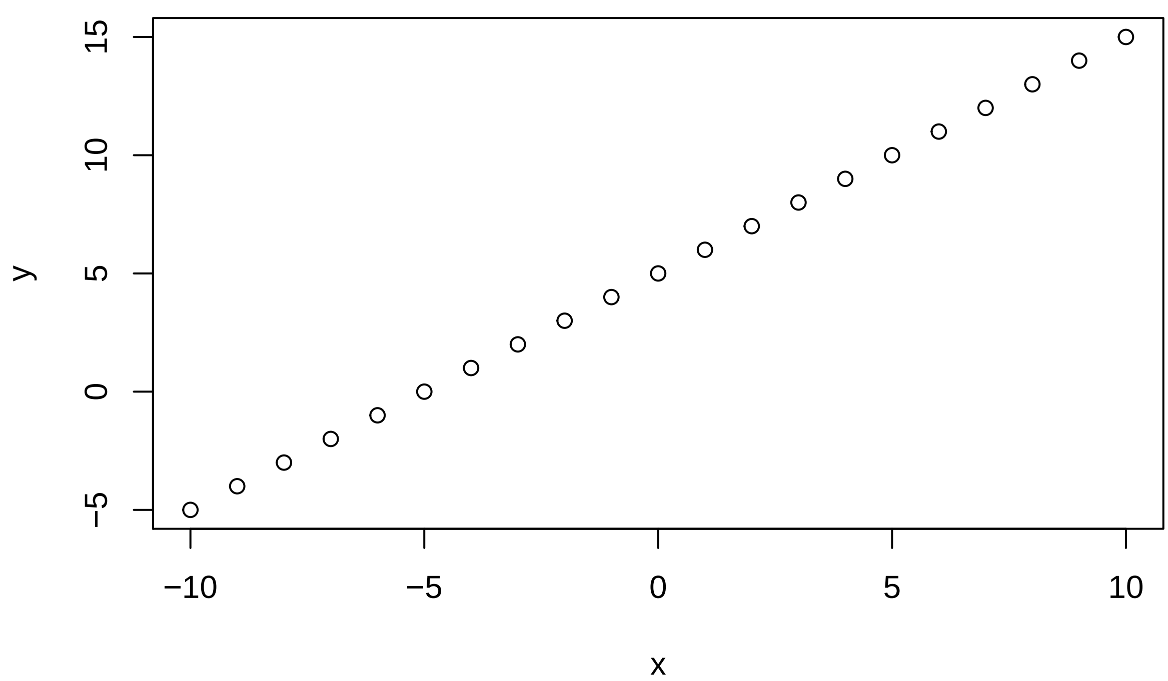
2.43e-05

What does this mean?

Topic 1: Covariance

Let us create two variables that are clearly dependent on each other.

```
x=seq(-10,10)
y=(x+5)
plot(x,y)
```



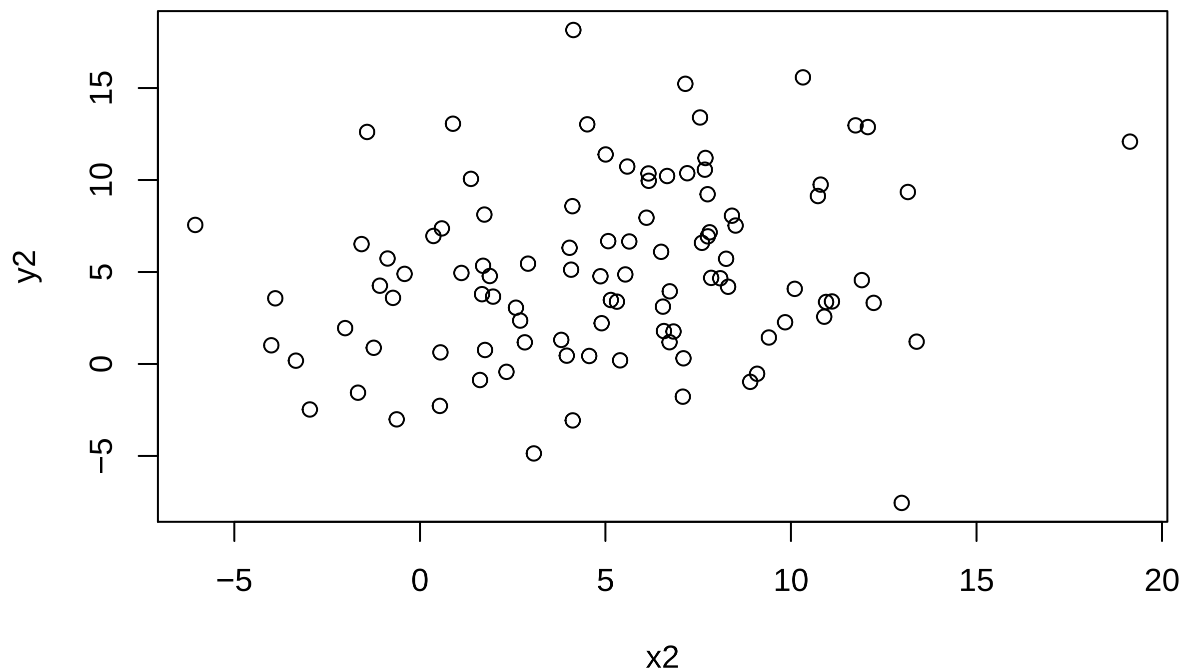
The covariance of these two variables is positive - as x increases so does y

```
cov(x,y)
```

```
## [1] 38.5
```

Let us create two variables that have no relationship to each other:

```
x2=rnorm(100,mean=5,sd=5)
y2=rnorm(100,mean=5,sd=5) # Both variables are just random drows from the normal distribution
plot(x2,y2)
```



The covariance should be close to zero - due to the randomness of the data it is most likely not exactly zero though.

```
cov(x2,y2)
```

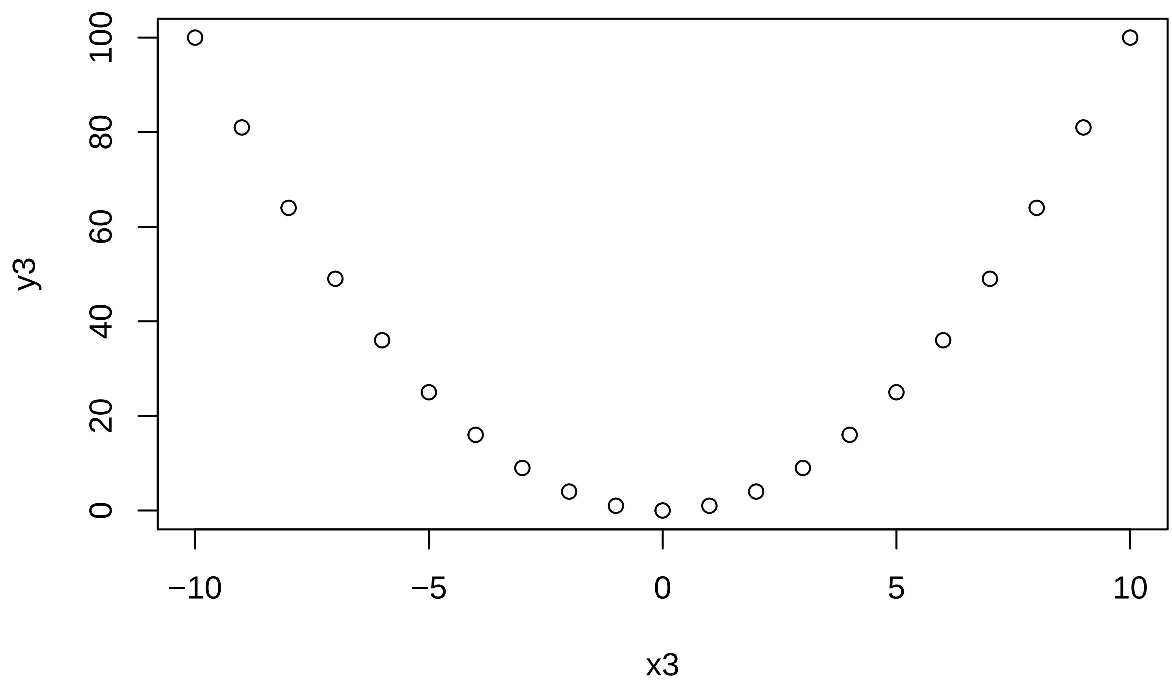
```
## [1] 4.773943
```

Note that this means even variables that are completely random and not related to each other may produce a non-zero covariance. However, the E(Cov(x2,y2)) = 0, so the distribution of the covariance is centered on the value 0.

Independence implies that the expected value of the covariance is zero.

Does a covariance of zero imply independence?

```
x3=seq(-10,10)
y3=x3^2
plot(x3,y3)
```



As we can clearly see from the plot, there is a curvilinear relationship of the two variables - they are not independent.

```r
cov(x3,y3)
```

```
## [1] 0
```

The formula tells us that the covariance is zero. Why?

Covariance captures linear relationships.

When x3 is below its mean, the values of y3 vary in the exact same way as when x3 is above its mean.

The formula can't capture the curvilinear relationship because it looks at the variation of y3 relative to x3's deviation from its mean. y3 varies in the exact same way when x3 moves above and below its mean. Thus, there is no linear relationship that could be captured.

Topic 2: Correlation

As you've learned in the lecture, the problem with covariance is that it is not to scale. It doesn't really tell us that much about how much variables vary with each other because it doesn't account for their individual variation.

```r
cor(x,y) ### Why does this produce "1". What is the meaning of this value?
```

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

How about x2 and y2 that are completely random?

```
cor(x2,y2) # The correlation is extremely close to zero, indicating that there is no systematic linear
```

```
## [1] 0.2148706
```

Does correlation capture non-linear relationships equally well?

Correlation is a mathematical concept. We cannot find the correlation between a numeric and a character vector.

```
y4=rep(c("a","b","c"),7)
y4
```

```
##  [1] "a" "b" "c" "a" "b" "c" "a" "b" "c" "a" "b" "c" "a" "b" "c" "a" "b"
## [18] "c" "a" "b" "c"
```

```
is.numeric(y4) # Checks whether y4 is numeric and returns the argument FALSE.
```

```
## [1] FALSE
```

```
cor(x,y4) # Gives us the error message "y must be numeric".
```

```
## Error in cor(x, y4): 'y' must be numeric
```

Interestingly, R allows us to find the correlation between a numeric and a logical vector, although a logical vector is not numeric.

```
y5=rep(c(T,F,F),7)
y5
```

```
##  [1]  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE
## [12] FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE
```

```
is.numeric(y5) # Checks whether y5 is numeric and returns the argument FALSE.
```

```
## [1] FALSE
```

```
class(y5) # Returns the class of the vector.
```

```
## [1] "logical"
```

```
cor(x,y5) # Returns a value.
```

```
## [1] -0.1167748
```

How do we have to think about this?

Assume that T=1 and F=0.

```r
y6=rep(c(1,0,0),7)
cor(x,y6) # Returns the same value as above, meaning that R views T=1 and F=0
```

```
## [1] -0.1167748
```

Topic 3: Cross-tabs

R has several built-in datasets, let's have a look at them.

```r
library(datasets)
data(occupationalStatus)
occupationalStatus
```

```
##       destination
## origin   1   2   3   4   5   6   7   8
##      1  50  19  26   8   7  11   6   2
##      2  16  40  34  18  11  20   8   3
##      3  12  35  65  66  35  88  23  21
##      4  11  20  58 110  40 183  64  32
##      5   2   8  12  23  25  46  28  12
##      6  12  28 102 162  90 554 230 177
##      7   0   6  19  40  21 158 143  71
##      8   0   3  14  32  15 126  91 106
```

According to the documentation this is "Cross-classification of a sample of British males according to each subject's occupational status and his father's occupational status."

The source is a journal article from 1979: "Goodman, L. A. (1979) Simple Models for the Analysis of Association in Cross-Classifications having Ordered Categories."

Let us assume that 1 is a low occupational status and 8 is a high occupational status. Is there a relationship between the status of the father and the son?

Before using the command below, use install.packages("gmodels").

```r
library(gmodels)
CrossTable(occupationalStatus)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  3498
##
##
```

```
##              | destination
##       origin |         1 |         2 |         3 |         4 |         5 |         6 |         7 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---
##            1 |        50 |        19 |        26 |         8 |         7 |        11 |         6 |
##              |   561.961 |    29.430 |    15.717 |     4.708 |     0.444 |    24.504 |    11.515 |
##              |     0.388 |     0.147 |     0.202 |     0.062 |     0.054 |     0.085 |     0.047 |
##              |     0.485 |     0.119 |     0.079 |     0.017 |     0.029 |     0.009 |     0.010 |
##              |     0.014 |     0.005 |     0.007 |     0.002 |     0.002 |     0.003 |     0.002 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---
##            2 |        16 |        40 |        34 |        18 |        11 |        20 |         8 |
##              |    30.377 |   161.485 |    27.842 |     0.144 |     0.028 |    18.723 |    11.946 |
##              |     0.107 |     0.267 |     0.227 |     0.120 |     0.073 |     0.133 |     0.053 |
##              |     0.155 |     0.252 |     0.103 |     0.039 |     0.045 |     0.017 |     0.013 |
##              |     0.005 |     0.011 |     0.010 |     0.005 |     0.003 |     0.006 |     0.002 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---
##            3 |        12 |        35 |        65 |        66 |        35 |        88 |        23 |
##              |     0.334 |    23.798 |    32.359 |     9.492 |     4.969 |     7.176 |    21.531 |
##              |     0.035 |     0.101 |     0.188 |     0.191 |     0.101 |     0.255 |     0.067 |
##              |     0.117 |     0.220 |     0.197 |     0.144 |     0.143 |     0.074 |     0.039 |
##              |     0.003 |     0.010 |     0.019 |     0.019 |     0.010 |     0.025 |     0.007 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---
##            4 |        11 |        20 |        58 |       110 |        40 |       183 |        64 |
##              |     1.186 |     0.534 |     1.707 |    25.988 |     0.414 |     0.309 |     6.458 |
##              |     0.021 |     0.039 |     0.112 |     0.212 |     0.077 |     0.353 |     0.124 |
##              |     0.107 |     0.126 |     0.176 |     0.240 |     0.164 |     0.154 |     0.108 |
##              |     0.003 |     0.006 |     0.017 |     0.031 |     0.011 |     0.052 |     0.018 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---
##            5 |         2 |         8 |        12 |        23 |        25 |        46 |        28 |
##              |     1.464 |     0.117 |     0.502 |     0.313 |    18.318 |     0.898 |     0.091 |
##              |     0.013 |     0.051 |     0.077 |     0.147 |     0.160 |     0.295 |     0.179 |
##              |     0.019 |     0.050 |     0.036 |     0.050 |     0.102 |     0.039 |     0.047 |
##              |     0.001 |     0.002 |     0.003 |     0.007 |     0.007 |     0.013 |     0.008 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---
##            6 |        12 |        28 |       102 |       162 |        90 |       554 |       230 |
##              |    19.508 |    18.320 |     5.219 |     1.404 |     0.216 |    19.474 |     0.000 |
##              |     0.009 |     0.021 |     0.075 |     0.120 |     0.066 |     0.409 |     0.170 |
##              |     0.117 |     0.176 |     0.309 |     0.353 |     0.369 |     0.467 |     0.388 |
##              |     0.003 |     0.008 |     0.029 |     0.046 |     0.026 |     0.158 |     0.066 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---
##            7 |         0 |         6 |        19 |        40 |        21 |       158 |       143 |
##              |    13.486 |    10.547 |    13.563 |     6.721 |     3.751 |     0.047 |    55.016 |
##              |     0.000 |     0.013 |     0.041 |     0.087 |     0.046 |     0.345 |     0.312 |
##              |     0.000 |     0.038 |     0.058 |     0.087 |     0.086 |     0.133 |     0.241 |
##              |     0.000 |     0.002 |     0.005 |     0.011 |     0.006 |     0.045 |     0.041 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---
##            8 |         0 |         3 |        14 |        32 |        15 |       126 |        91 |
##              |    11.395 |    12.103 |    13.878 |     6.946 |     5.330 |     0.207 |     9.829 |
##              |     0.000 |     0.008 |     0.036 |     0.083 |     0.039 |     0.326 |     0.235 |
##              |     0.000 |     0.019 |     0.042 |     0.070 |     0.061 |     0.106 |     0.153 |
##              |     0.000 |     0.001 |     0.004 |     0.009 |     0.004 |     0.036 |     0.026 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---
## Column Total |       103 |       159 |       330 |       459 |       244 |      1186 |       593 |
##              |     0.029 |     0.045 |     0.094 |     0.131 |     0.070 |     0.339 |     0.170 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---
```

```
##
##
```

How can we interpret this table?

Topic 4: Central Limit Theorem

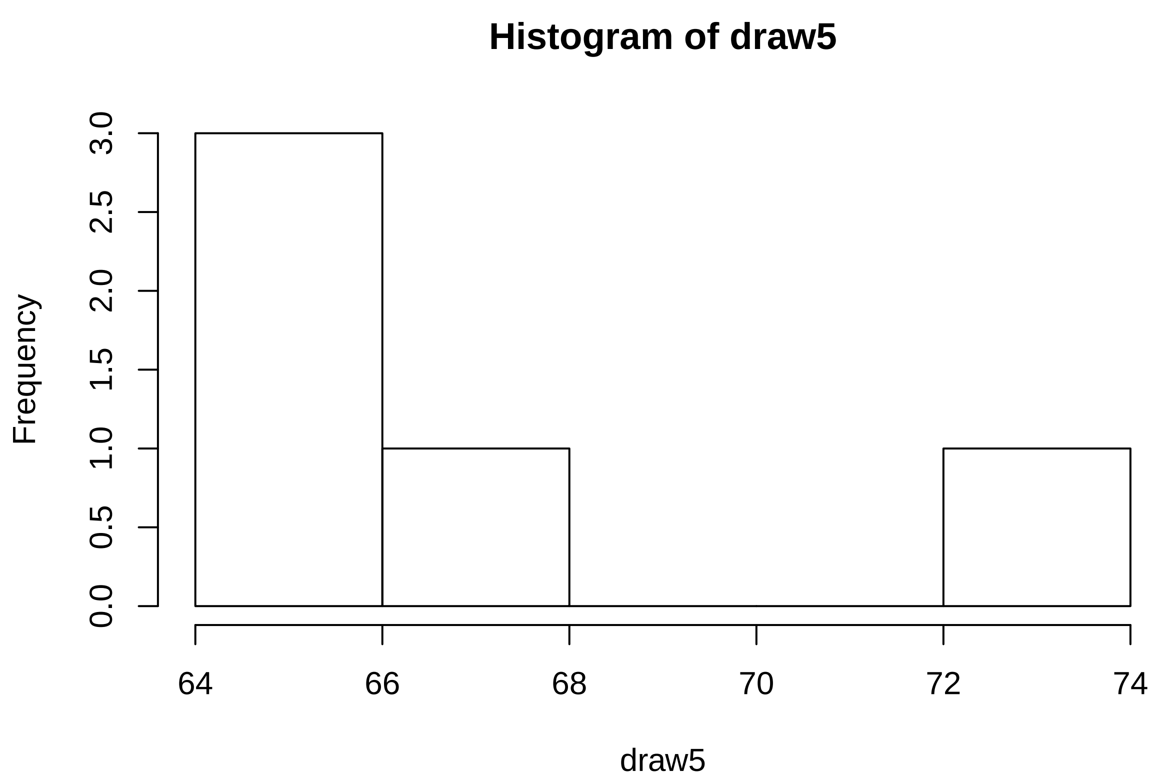
The Central Limit Theorem says that if we have infinitely many draws of the same size from a specific distribution, the mean of this distribution will be approximately normally distributed.

Let us illustrate this with a simple example of the binomial distribution.

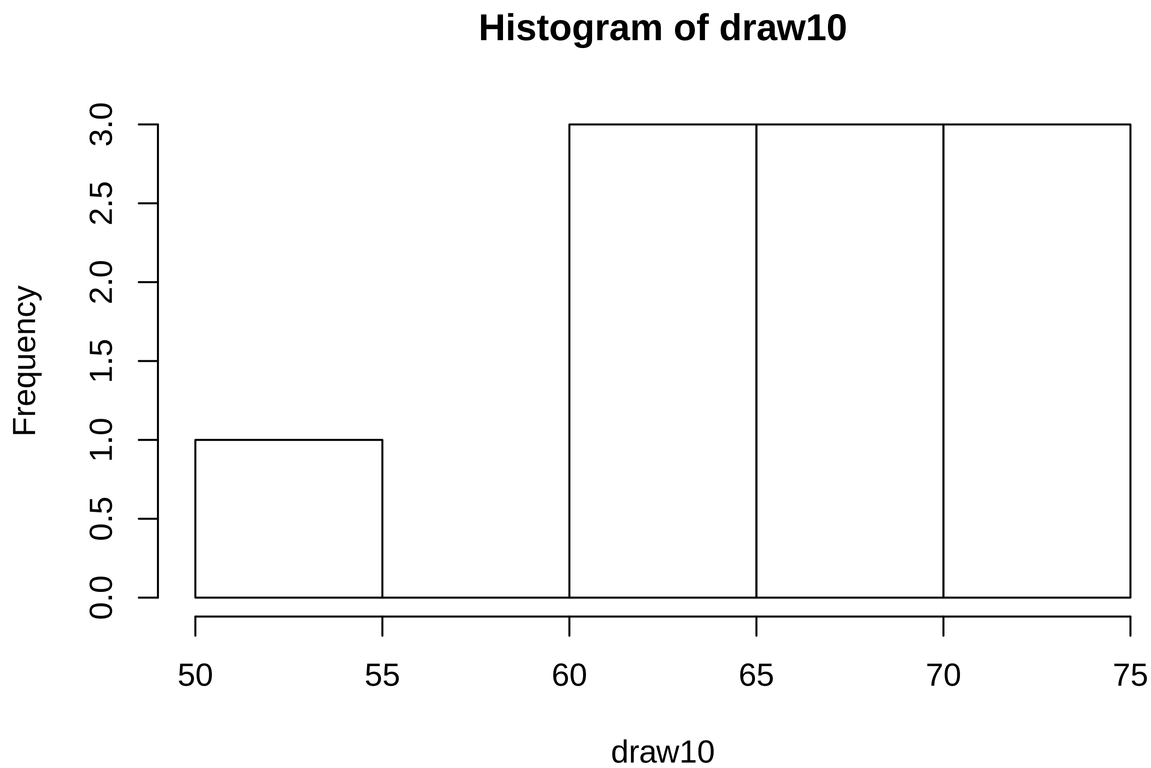
```
draw1=rbinom(1,size=100,p=0.67)
draw1
```

```
## [1] 60
```

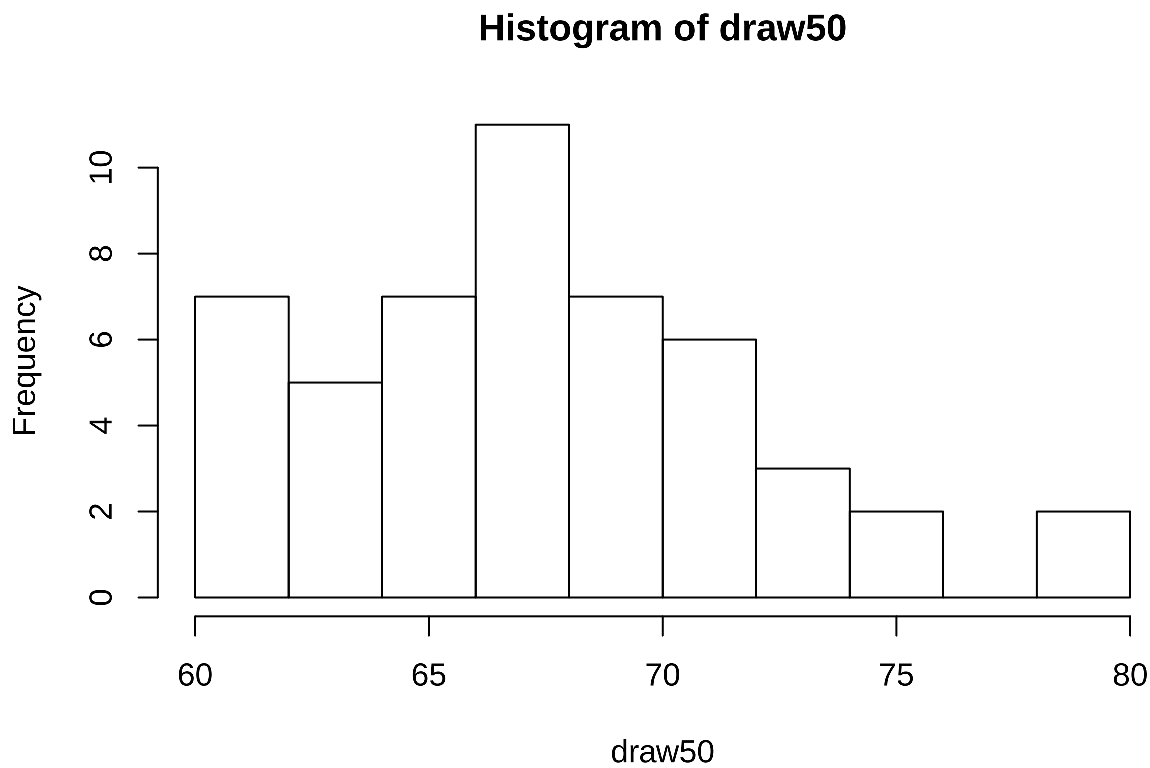
```
draw5=rbinom(5,size=100,p=0.67)
hist(draw5)
```



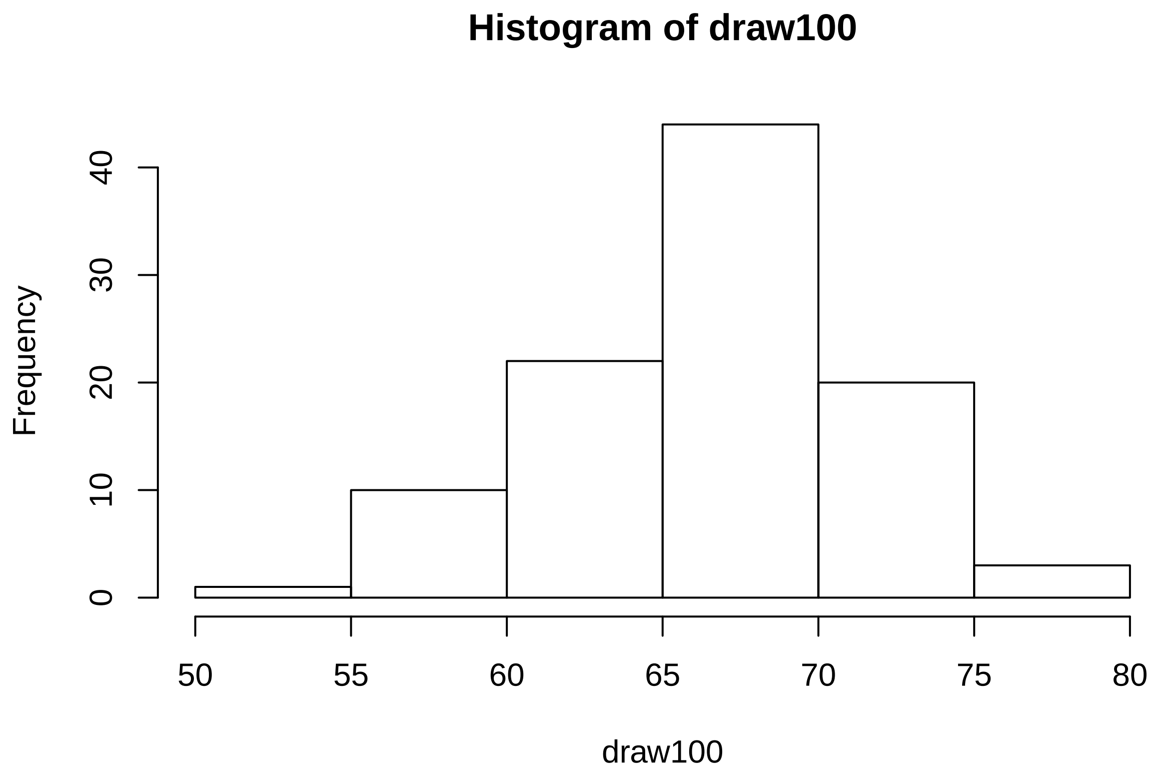

```
draw10=rbinom(10,size=100,p=0.67)
hist(draw10)
```



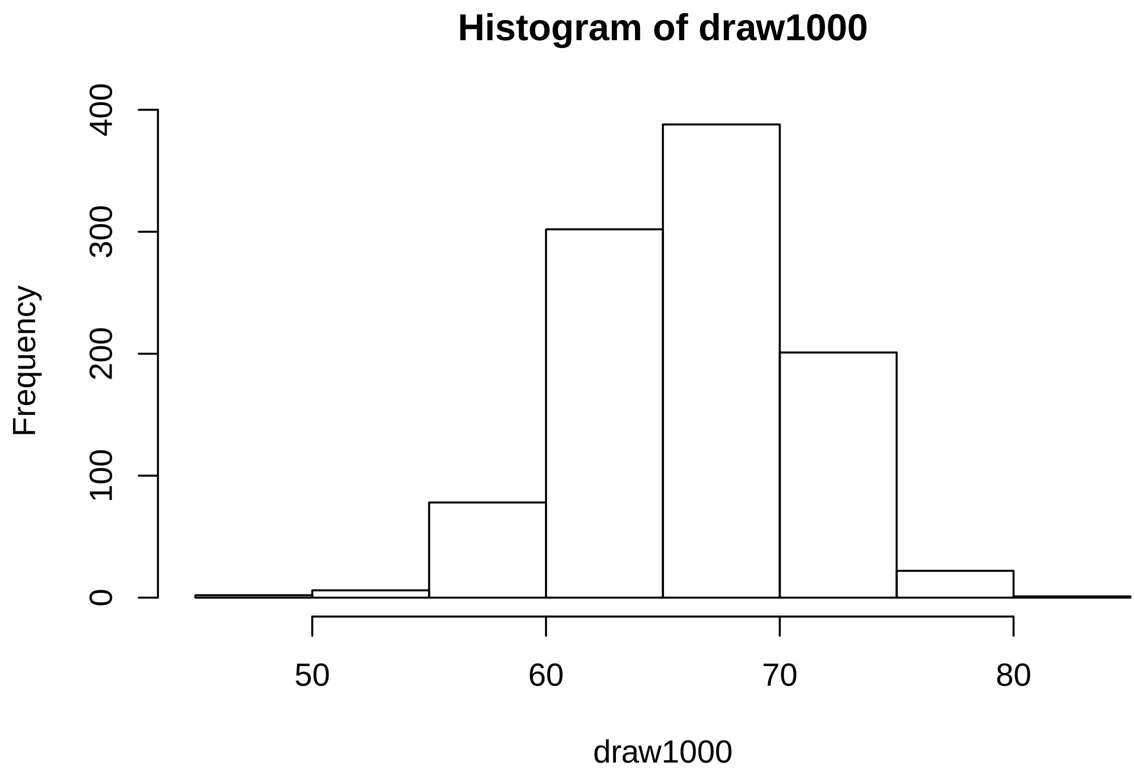

```
draw50=rbinom(50,size=100,p=0.67)
hist(draw50)
```



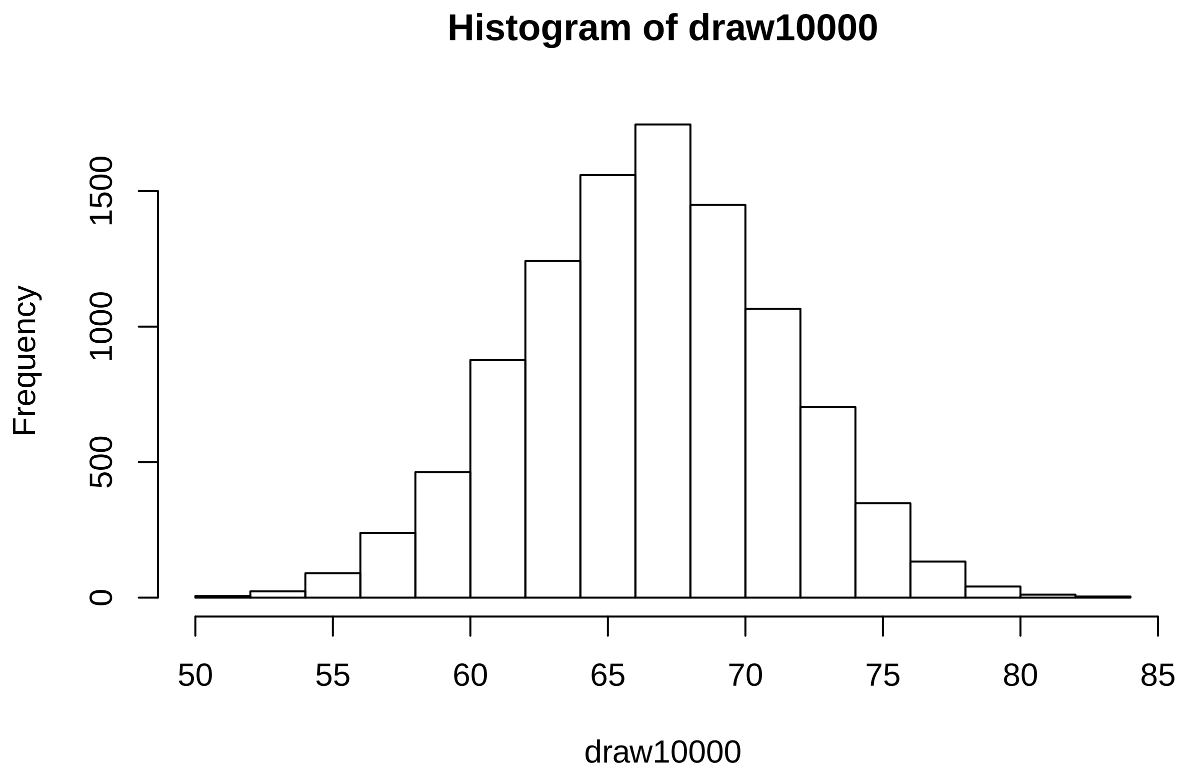

```
draw100=rbinom(100,size=100,p=0.67)
hist(draw100)
```




```
draw1000=rbinom(1000,size=100,p=0.67)
hist(draw1000)
```




```
draw10000=rbinom(10000,size=100,p=0.67)
hist(draw10000) # This really looks like a normal distribution
```

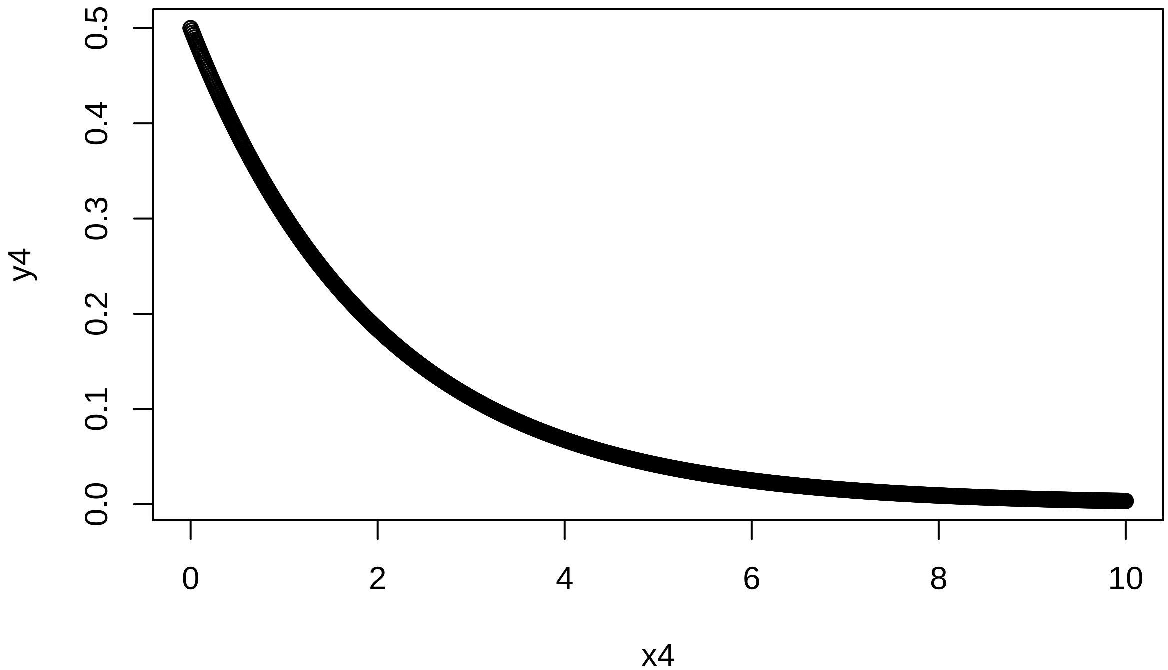


Note that this is true for any distribution, even those that are NOT normally distributed themselves (the distribution of a binomial looks similar to a normal distribution for large N).

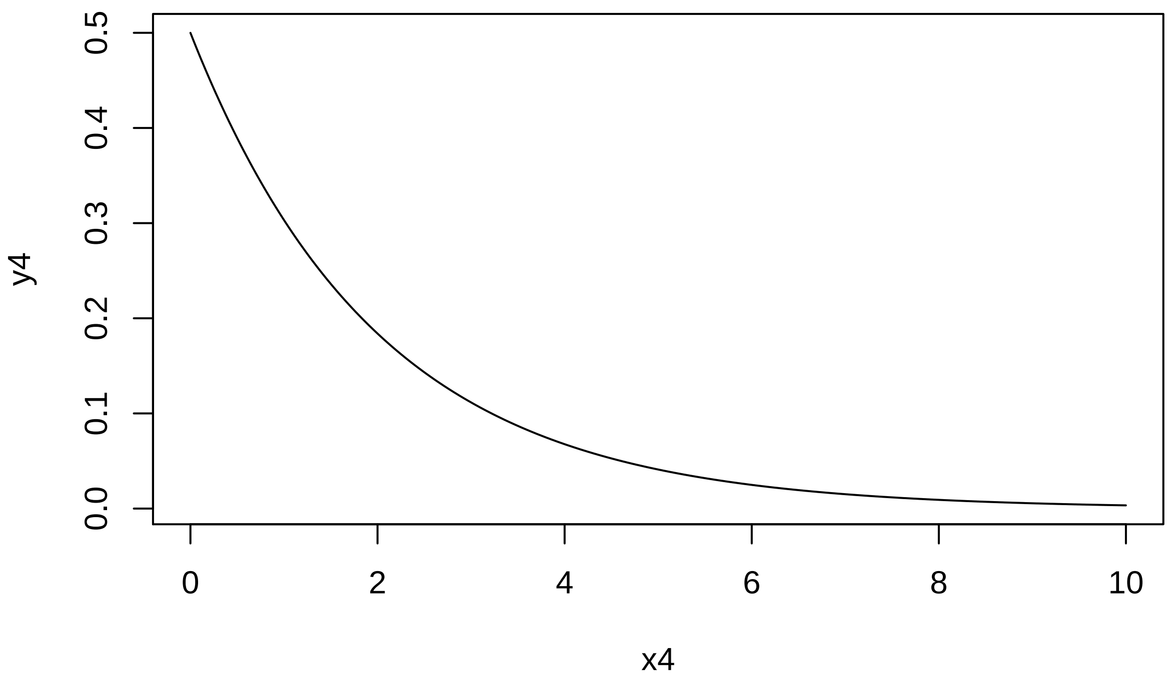
Let's try a similar example with an exponential distribution. The exponential distribution doesn't look like a normal distribution.

How does an exponential distribution look like?

```
x4=seq(0,10,by=0.01)
y4=dexp(x4, rate=0.5) # Returns the density
plot(x4,y4) # This doesn't look nice
```




```r
plot(x4,y4, type="l") # Use type="l" for a line plot
```



We can also use the ggplot2 package to make it look even nicer.

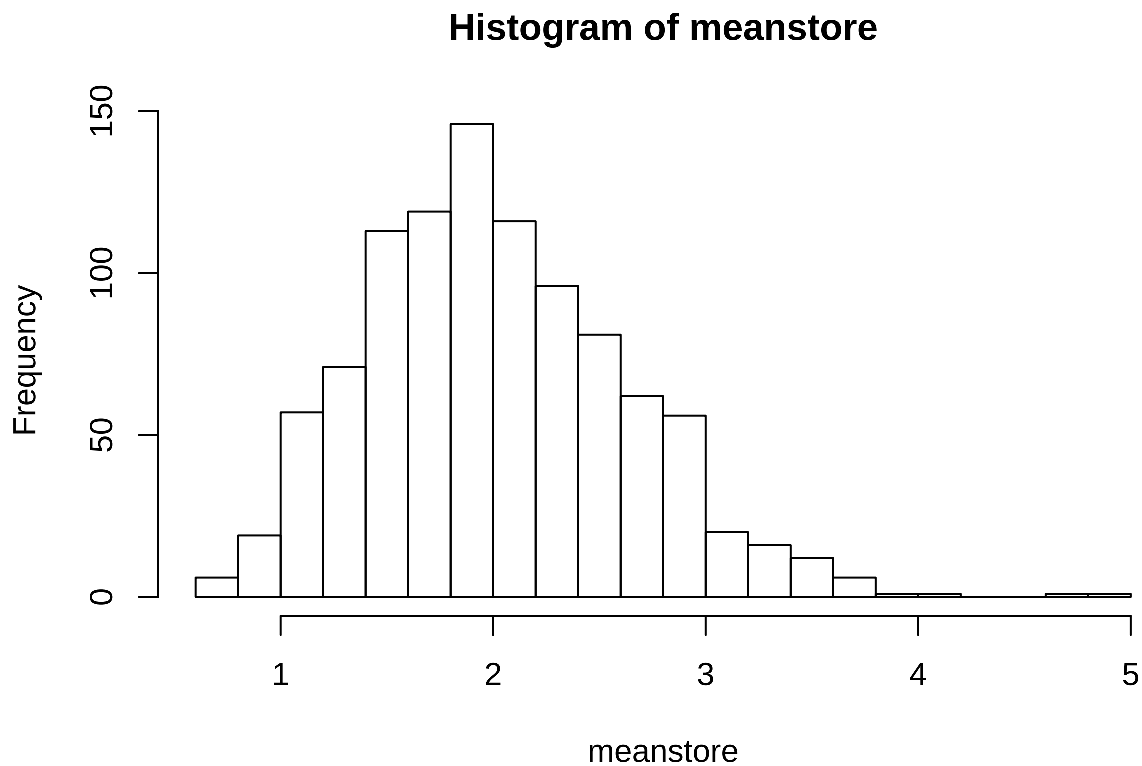
Use install.packages("ggplot2") before you run this code.

```
library(ggplot2)
plot1=qplot(x4,y4) # Now that looks even nicer
```

Recall: The Central Limit Theorem states that if we have multiple samples of the same size, their mean will be approximately normally distributed.

So, what happens if we draw 1000 times 10 samples from this distribution, how will their mean be distributed?

```
meanstore=rep(0,1000)
for (i in 1:1000){
  expdraw=rexp(10, rate=0.5)
  meanstore[i]=mean(expdraw)
}
hist(meanstore, breaks=20) # It is approximately normally distributed, as predicted by the CLT.
```



Topic 5: t-tests and p-values

Typically, a p-value is related to a type-1 error rate (alpha) that we define in advance. Often we want alpha to be smaller than 0.05.

Typically, our null hypothesis (H0) is that there is no relationship between two variables.

The p-value is the probability that we get data with evidence that is such strong AGAINST H0 if H0 was true. Think about what this means.

Let's load another R dataset that can illustrate this. The "airquality" dataset. According to the documentation, this is "Daily air quality measurements in New York, May to September 1973."

More details can be found here:

```
data(airquality)
airquality
```

```
##     Ozone Solar.R Wind Temp Month Day
## 1      41     190  7.4   67     5   1
## 2      36     118  8.0   72     5   2
## 3      12     149 12.6   74     5   3
## 4      18     313 11.5   62     5   4
## 5      NA      NA 14.3   56     5   5
## 6      28      NA 14.9   66     5   6
## 7      23     299  8.6   65     5   7
## 8      19      99 13.8   59     5   8
## 9       8      19 20.1   61     5   9
```

```
## 10    NA    194  8.6  69     5  10
## 11     7     NA  6.9  74     5  11
## 12    16    256  9.7  69     5  12
## 13    11    290  9.2  66     5  13
## 14    14    274 10.9  68     5  14
## 15    18     65 13.2  58     5  15
## 16    14    334 11.5  64     5  16
## 17    34    307 12.0  66     5  17
## 18     6     78 18.4  57     5  18
## 19    30    322 11.5  68     5  19
## 20    11     44  9.7  62     5  20
## 21     1      8  9.7  59     5  21
## 22    11    320 16.6  73     5  22
## 23     4     25  9.7  61     5  23
## 24    32     92 12.0  61     5  24
## 25    NA     66 16.6  57     5  25
## 26    NA    266 14.9  58     5  26
## 27    NA     NA  8.0  57     5  27
## 28    23     13 12.0  67     5  28
## 29    45    252 14.9  81     5  29
## 30   115    223  5.7  79     5  30
## 31    37    279  7.4  76     5  31
## 32    NA    286  8.6  78     6   1
## 33    NA    287  9.7  74     6   2
## 34    NA    242 16.1  67     6   3
## 35    NA    186  9.2  84     6   4
## 36    NA    220  8.6  85     6   5
## 37    NA    264 14.3  79     6   6
## 38    29    127  9.7  82     6   7
## 39    NA    273  6.9  87     6   8
## 40    71    291 13.8  90     6   9
## 41    39    323 11.5  87     6  10
## 42    NA    259 10.9  93     6  11
## 43    NA    250  9.2  92     6  12
## 44    23    148  8.0  82     6  13
## 45    NA    332 13.8  80     6  14
## 46    NA    322 11.5  79     6  15
## 47    21    191 14.9  77     6  16
## 48    37    284 20.7  72     6  17
## 49    20     37  9.2  65     6  18
## 50    12    120 11.5  73     6  19
## 51    13    137 10.3  76     6  20
## 52    NA    150  6.3  77     6  21
## 53    NA     59  1.7  76     6  22
## 54    NA     91  4.6  76     6  23
## 55    NA    250  6.3  76     6  24
## 56    NA    135  8.0  75     6  25
## 57    NA    127  8.0  78     6  26
## 58    NA     47 10.3  73     6  27
## 59    NA     98 11.5  80     6  28
## 60    NA     31 14.9  77     6  29
## 61    NA    138  8.0  83     6  30
## 62   135    269  4.1  84     7   1
## 63    49    248  9.2  85     7   2
```

```
## 64     32     236  9.2   81     7   3
## 65     NA     101 10.9   84     7   4
## 66     64     175  4.6   83     7   5
## 67     40     314 10.9   83     7   6
## 68     77     276  5.1   88     7   7
## 69     97     267  6.3   92     7   8
## 70     97     272  5.7   92     7   9
## 71     85     175  7.4   89     7  10
## 72     NA     139  8.6   82     7  11
## 73     10     264 14.3   73     7  12
## 74     27     175 14.9   81     7  13
## 75     NA     291 14.9   91     7  14
## 76      7      48 14.3   80     7  15
## 77     48     260  6.9   81     7  16
## 78     35     274 10.3   82     7  17
## 79     61     285  6.3   84     7  18
## 80     79     187  5.1   87     7  19
## 81     63     220 11.5   85     7  20
## 82     16       7  6.9   74     7  21
## 83     NA     258  9.7   81     7  22
## 84     NA     295 11.5   82     7  23
## 85     80     294  8.6   86     7  24
## 86    108     223  8.0   85     7  25
## 87     20      81  8.6   82     7  26
## 88     52      82 12.0   86     7  27
## 89     82     213  7.4   88     7  28
## 90     50     275  7.4   86     7  29
## 91     64     253  7.4   83     7  30
## 92     59     254  9.2   81     7  31
## 93     39      83  6.9   81     8   1
## 94      9      24 13.8   81     8   2
## 95     16      77  7.4   82     8   3
## 96     78      NA  6.9   86     8   4
## 97     35      NA  7.4   85     8   5
## 98     66      NA  4.6   87     8   6
## 99    122     255  4.0   89     8   7
## 100    89     229 10.3   90     8   8
## 101   110     207  8.0   90     8   9
## 102    NA     222  8.6   92     8  10
## 103    NA     137 11.5   86     8  11
## 104    44     192 11.5   86     8  12
## 105    28     273 11.5   82     8  13
## 106    65     157  9.7   80     8  14
## 107    NA      64 11.5   79     8  15
## 108    22      71 10.3   77     8  16
## 109    59      51  6.3   79     8  17
## 110    23     115  7.4   76     8  18
## 111    31     244 10.9   78     8  19
## 112    44     190 10.3   78     8  20
## 113    21     259 15.5   77     8  21
## 114     9      36 14.3   72     8  22
## 115    NA     255 12.6   75     8  23
## 116    45     212  9.7   79     8  24
## 117   168     238  3.4   81     8  25
```

```
## 118    73     215  8.0   86     8  26
## 119    NA     153  5.7   88     8  27
## 120    76     203  9.7   97     8  28
## 121   118     225  2.3   94     8  29
## 122    84     237  6.3   96     8  30
## 123    85     188  6.3   94     8  31
## 124    96     167  6.9   91     9   1
## 125    78     197  5.1   92     9   2
## 126    73     183  2.8   93     9   3
## 127    91     189  4.6   93     9   4
## 128    47      95  7.4   87     9   5
## 129    32      92 15.5   84     9   6
## 130    20     252 10.9   80     9   7
## 131    23     220 10.3   78     9   8
## 132    21     230 10.9   75     9   9
## 133    24     259  9.7   73     9  10
## 134    44     236 14.9   81     9  11
## 135    21     259 15.5   76     9  12
## 136    28     238  6.3   77     9  13
## 137     9      24 10.9   71     9  14
## 138    13     112 11.5   71     9  15
## 139    46     237  6.9   78     9  16
## 140    18     224 13.8   67     9  17
## 141    13      27 10.3   76     9  18
## 142    24     238 10.3   68     9  19
## 143    16     201  8.0   82     9  20
## 144    13     238 12.6   64     9  21
## 145    23      14  9.2   71     9  22
## 146    36     139 10.3   81     9  23
## 147     7      49 10.3   69     9  24
## 148    14      20 16.6   63     9  25
## 149    30     193  6.9   70     9  26
## 150    NA     145 13.2   77     9  27
## 151    14     191 14.3   75     9  28
## 152    18     131  8.0   76     9  29
## 153    20     223 11.5   68     9  30
```

```r
summary(airquality) # Use this command if you don't want to see the whole dataset but just a summary of
```

```
##      Ozone           Solar.R           Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
##  NA's   :37       NA's   :7                                       
##      Month            Day      
##  Min.   :5.000   Min.   : 1.0  
##  1st Qu.:6.000   1st Qu.: 8.0  
##  Median :7.000   Median :16.0  
##  Mean   :6.993   Mean   :15.8  
##  3rd Qu.:8.000   3rd Qu.:23.0  
##  Max.   :9.000   Max.   :31.0
```

```
##
```

Our question is: is there a linear relationship between the Ozone measures and the Solar.R measures?

Let us use linear regression to answer this question:

```
lm1=lm(Ozone ~ Solar.R, data=airquality)
```

The summary of this linear regression will return a t-value and a p-value for the inercept all coefficients.

```
summary(lm1)
```

```
##
## Call:
## lm(formula = Ozone ~ Solar.R, data = airquality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.292 -21.361  -8.864  16.373 119.136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.59873    6.74790   2.756 0.006856 **
## Solar.R      0.12717    0.03278   3.880 0.000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.33 on 109 degrees of freedom
##   (42 observations deleted due to missingness)
## Multiple R-squared:  0.1213, Adjusted R-squared:  0.1133
## F-statistic: 15.05 on 1 and 109 DF,  p-value: 0.0001793
```

The last column Pr(>|t|) is the p-value of the t-test. How do we interpret this finding?

The probability to find this data if H0 is true (i.e. there is no relationship between Ozone and Solar.R) is 0.1%. This is strong evidence against H0. Therefore we reject H0 under $p < 0.001$. Please pay attention to the difference between percent (0-100) and proportions (0-1).

For a 1-point increase in Solar.R, we would expect a 0.13 increase in Ozone (in a multivariate model we would have to add: "holding all other covariates constant"). The associated t-value is 3.880, which imples a p-value of 0.0002. This p-value is $p < 0.001$. Therefore, the relationship is statistically significant at all common levels of statistical significance.

Note that there are four important thresholds of statistical significane:

$p = 0.001$, equals a type-1 error rate (alpha) of 0.001 $p = 0.01$, equals a tye-1 error rate (alpha) of 0.01 $p = 0.05$, equals a type-1 error rate (alpha) of 0.05 $p = 0.1$, equals a type-1 error rate (alpha) of 0.1

If an indicator has a p-value of $p < 0.001$, it is significant at all common levels of statistical significance.

This indicates that there is a significan t linear relationship between the two variables.

If you have a linear regression with multiple independent variables, then the code you need looks like the following:

lm(y ~ x1 + x2 + x3)