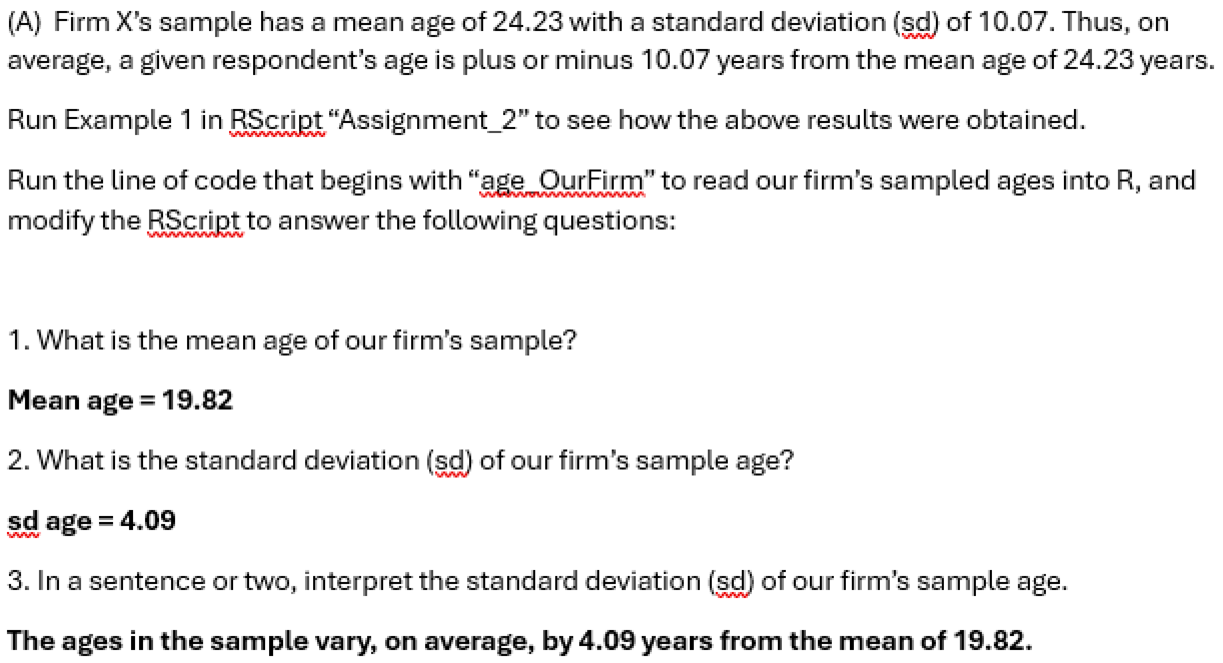
Quantitative Sociological Analysis

Inferential Statistics
Hypothesis Testing and Bivariate Statistics

Part 7

April 3, 2025

Assignment 2: Task A


(A) Firm X's sample has a mean age of 24.23 with a standard deviation (sd) of 10.07. Thus, on average, a given respondent's age is plus or minus 10.07 years from the mean age of 24.23 years.

Run Example 1 in RScript "Assignment_2" to see how the above results were obtained.

Run the line of code that begins with "age_OurFirm" to read our firm's sampled ages into R, and modify the RScript to answer the following questions:

1. What is the mean age of our firm's sample?

Mean age = 19.82

2. What is the standard deviation (sd) of our firm's sample age?

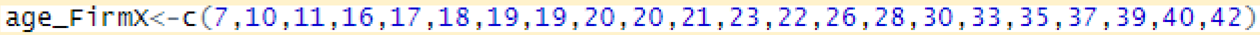
sd age = 4.09

3. In a sentence or two, interpret the standard deviation (sd) of our firm's sample age.

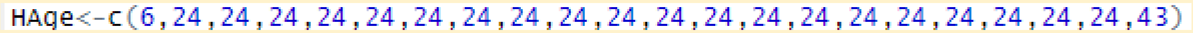
The ages in the sample vary, on average, by 4.09 years from the mean of 19.82.

- Some suggested sd had to do with range
- While range is also a measure of dispersion, it does not directly determine the sd
- Consider the following example:

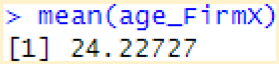
FirmX: min 7, max 42, range = 42-7 = 35

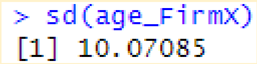

```
age_FirmX<-c(7,10,11,16,17,18,19,19,20,20,21,23,22,26,28,30,33,35,37,39,40,42)
```

HAge: min 6, max 43, range = 43-6 = 37

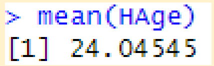

```
HAge<-c(6,24,24,24,24,24,24,24,24,24,24,24,24,24,24,24,24,24,24,24,24,43)
```

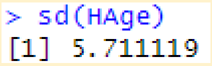
FirmX: mean age = 24.23, sd = 10.07


```
> mean(age_FirmX)
[1] 24.22727
```


```
> sd(age_FirmX)
[1] 10.07085
```

HAge: mean age = 24.05, sd =   5.71


```
> mean(HAge)
[1] 24.04545
```

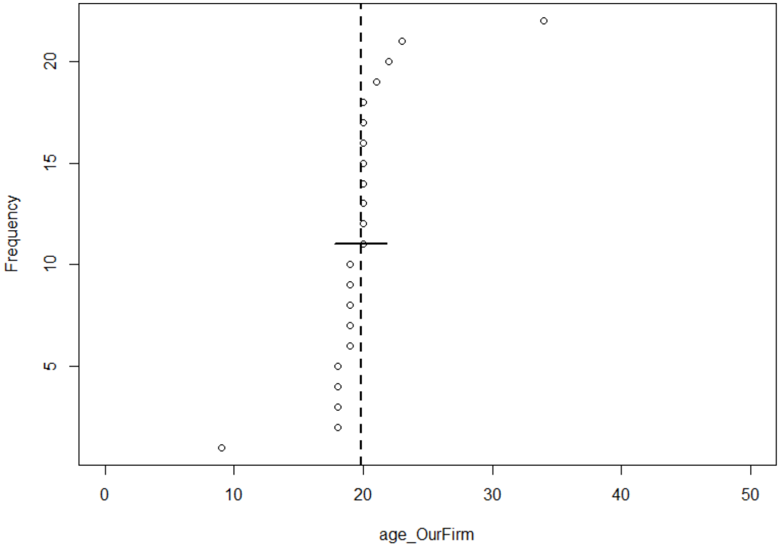

```
> sd(HAge)
[1] 5.711119
```

Let's see how sd measures avg. spread around a mean…

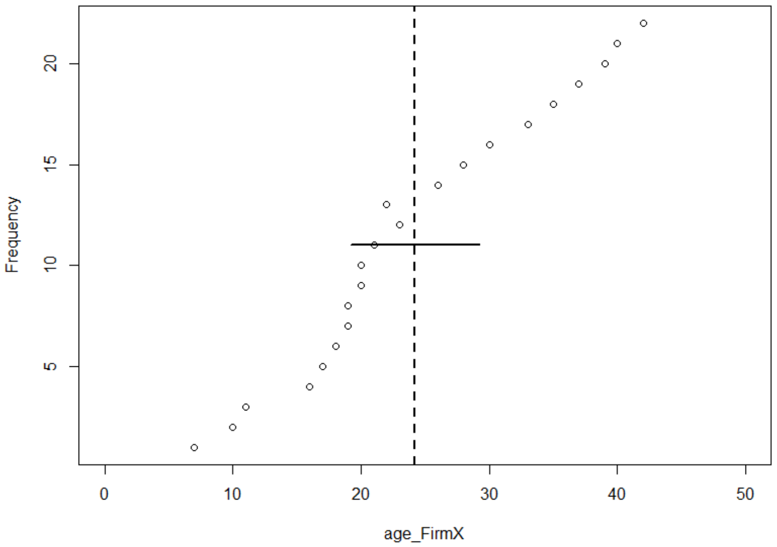
Assignment 2: Task A cont.

$$sd = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

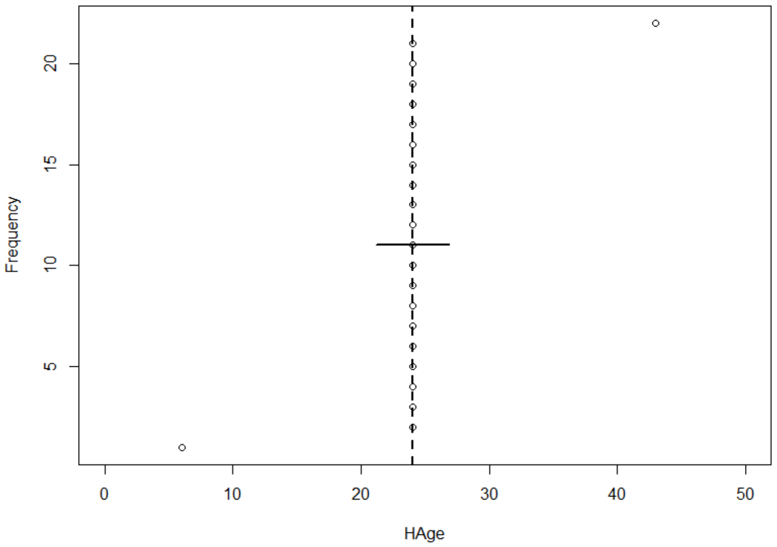
$\bar{X} = 19.82$ $sd = 4.09$



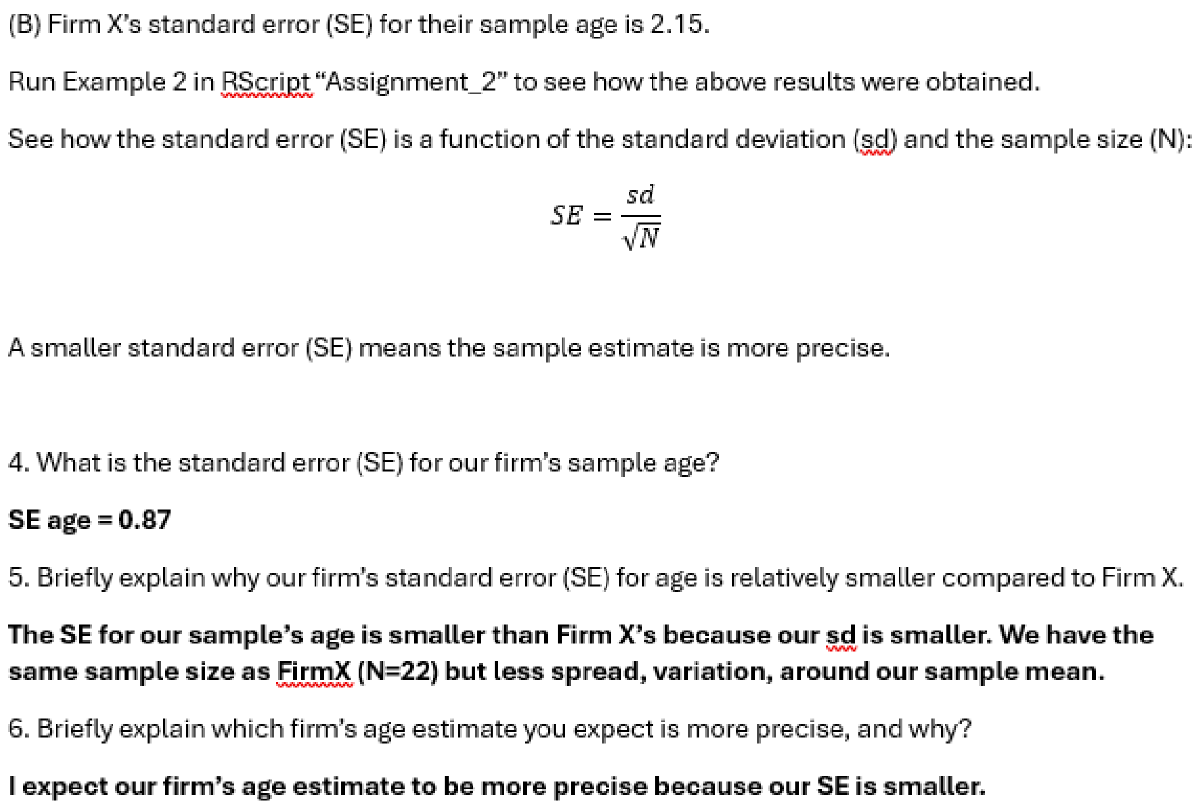
$\bar{X} = 24.23$ $sd = 10.07$



$\bar{X} = 24.05$ $sd = 5.71$



Assignment 2: Task B



(B) Firm X's standard error (SE) for their sample age is 2.15.

Run Example 2 in RScript "Assignment_2" to see how the above results were obtained.

See how the standard error (SE) is a function of the standard deviation (sd) and the sample size (N):

$$SE = \frac{sd}{\sqrt{N}}$$

A smaller standard error (SE) means the sample estimate is more precise.

4. What is the standard error (SE) for our firm's sample age?

SE age = 0.87

5. Briefly explain why our firm's standard error (SE) for age is relatively smaller compared to Firm X.

The SE for our sample's age is smaller than Firm X's because our sd is smaller. We have the same sample size as FirmX (N=22) but less spread, variation, around our sample mean.

6. Briefly explain which firm's age estimate you expect is more precise, and why?

I expect our firm's age estimate to be more precise because our SE is smaller.

- Some suggested SE had to do with accuracy, implying how well a sample statistic estimates the population

- precision ≠ accuracy

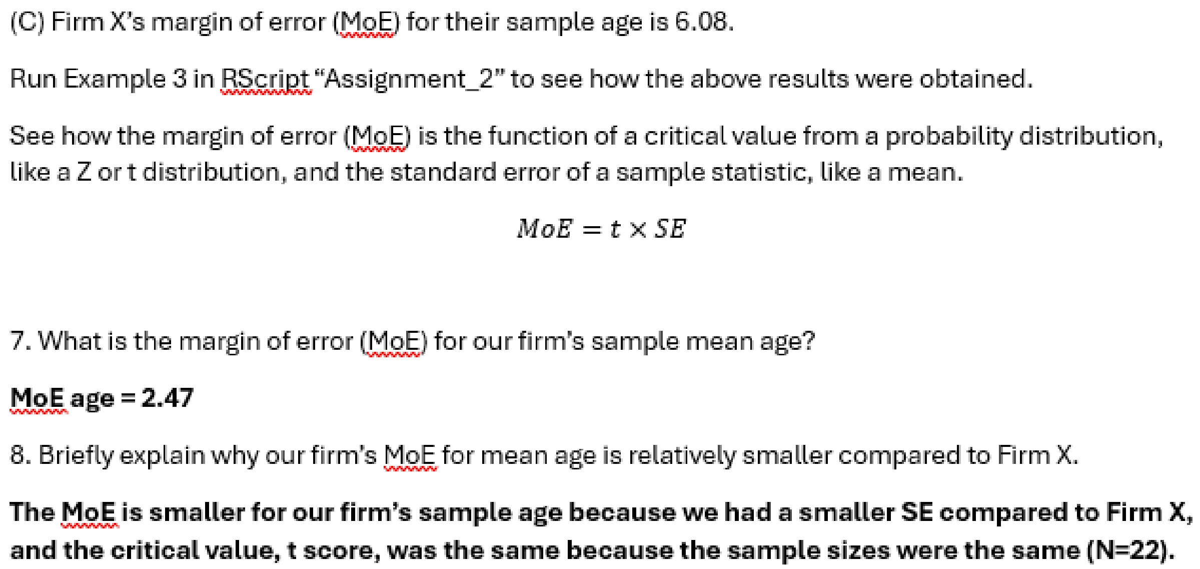
When population parameters are unknown, we can never be certain how well a sample statistic estimates the population (accuracy).

When population parameters are known then features like age, sex, and race in survey data can be used to assess how well a sample estimates the population.

In general, relatively larger random samples are theoretically more generalizable, lead to more accurate estimates of population parameters.

This type of "accuracy" in terms of representativeness, generalizability, is distinct from the reliability and validity of a variable in terms of measurement.

Assignment 2: Task C



(C) Firm X's margin of error (MoE) for their sample age is 6.08.

Run Example 3 in RScript "Assignment_2" to see how the above results were obtained.

See how the margin of error (MoE) is the function of a critical value from a probability distribution, like a Z or t distribution, and the standard error of a sample statistic, like a mean.

$$MoE = t \times SE$$

7. What is the margin of error (MoE) for our firm's sample mean age?

MoE age = 2.47

8. Briefly explain why our firm's MoE for mean age is relatively smaller compared to Firm X.

The MoE is smaller for our firm's sample age because we had a smaller SE compared to Firm X, and the critical value, t score, was the same because the sample sizes were the same (N=22).

- MoE expands on SE by incorporating a critical value from a probability distribution, which accounts for random sampling variability at a specific confidence, alpha, level.

- The MoE quantifies the expected range within which the true population parameter is likely to fall, given the sampling variability.

Again, when the population variability is unknown, we can never be certain how well a sample statistic estimates the population (accuracy).

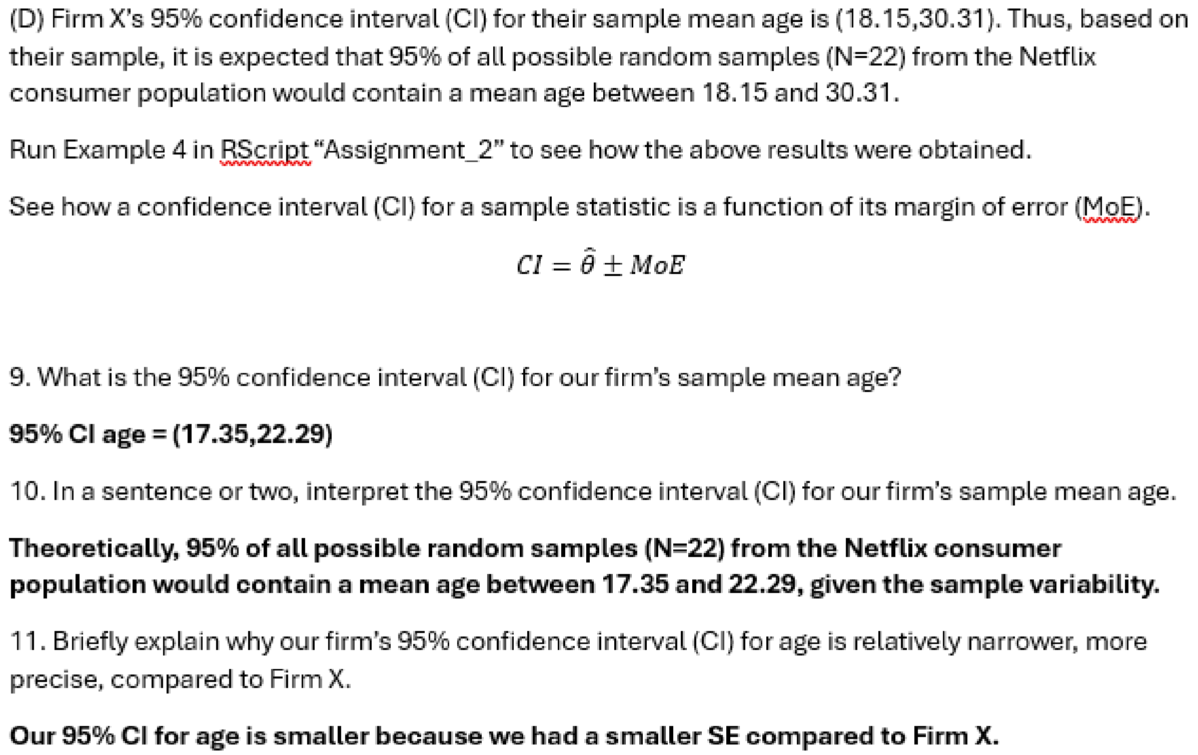
Consider that a sample mean ($\bar{X}$) that matches the population (μ) can often be obtained in ways by which the sample standard deviation does not match the population standard deviation (σ)

For example:

FirmX's mean age = 24.23, sd = 10.07

HAge mean age = 24.05, sd = 5.71

Assignment 2: Task D



(D) Firm X's 95% confidence interval (CI) for their sample mean age is (18.15,30.31). Thus, based on their sample, it is expected that 95% of all possible random samples (N=22) from the Netflix consumer population would contain a mean age between 18.15 and 30.31.

Run Example 4 in RScript "Assignment_2" to see how the above results were obtained.

See how a confidence interval (CI) for a sample statistic is a function of its margin of error (MoE).

$$CI = \hat{\theta} \pm MoE$$

9. What is the 95% confidence interval (CI) for our firm's sample mean age?

95% CI age = (17.35,22.29)

10. In a sentence or two, interpret the 95% confidence interval (CI) for our firm's sample mean age.

Theoretically, 95% of all possible random samples (N=22) from the Netflix consumer population would contain a mean age between 17.35 and 22.29, given the sample variability.

11. Briefly explain why our firm's 95% confidence interval (CI) for age is relatively narrower, more precise, compared to Firm X.

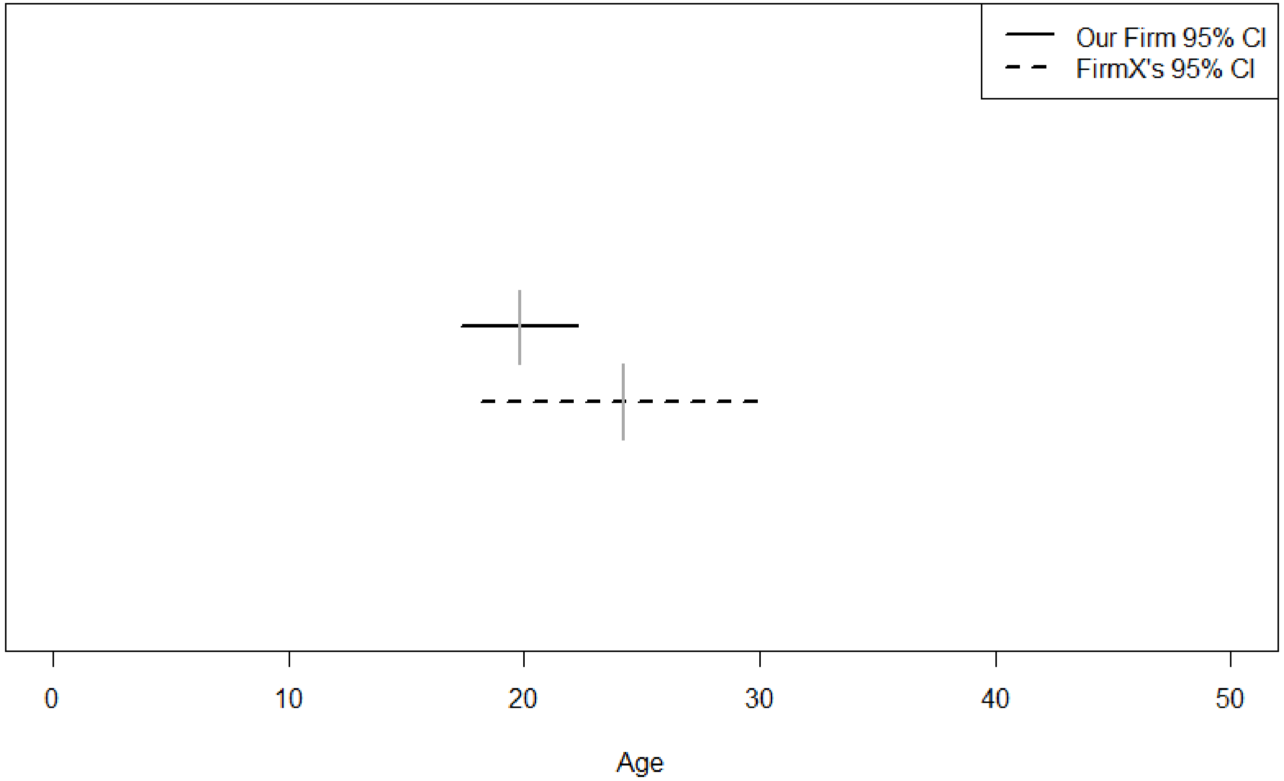
Our 95% CI for age is smaller because we had a smaller SE compared to Firm X.

- In view of the previous tasks, recognize how a confidence interval is a function of sampling variability and sample size.

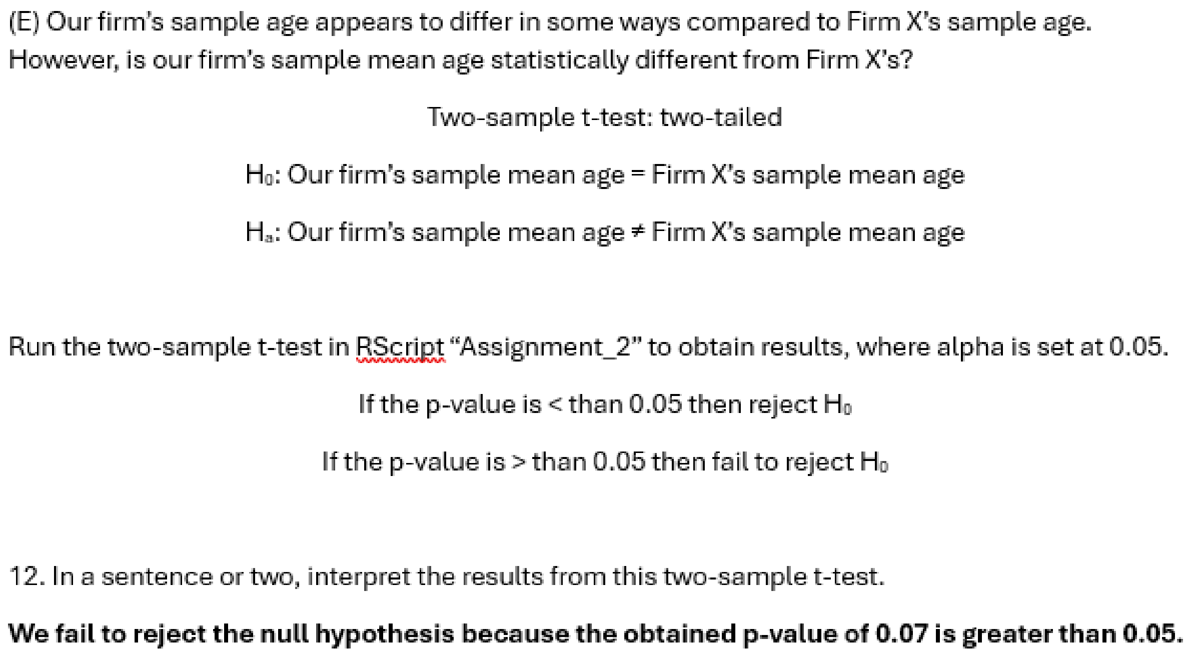
- This tells us about precision, but not accuracy

Recall that very different age distributions were obtained from random samples of the same size (N=22) from the Netflix consumer population.

However, the 95% CIs for each sample's statistic overlap.



Assignment 2: Task E



(E) Our firm's sample age appears to differ in some ways compared to Firm X's sample age. However, is our firm's sample mean age statistically different from Firm X's?

Two-sample t-test: two-tailed

H_0: Our firm's sample mean age = Firm X's sample mean age

H_a: Our firm's sample mean age ≠ Firm X's sample mean age

Run the two-sample t-test in RScript "Assignment_2" to obtain results, where alpha is set at 0.05.

If the p-value is < than 0.05 then reject H_0

If the p-value is > than 0.05 then fail to reject H_0

12. In a sentence or two, interpret the results from this two-sample t-test.

We fail to reject the null hypothesis because the obtained p-value of 0.07 is greater than 0.05.

- Our sample's mean age is not statistically different from FirmX's at the alpha=0.05 level

- Our sample statistic is more precise, but no way to tell which sample statistic is more accurate

If the hypothesis test obtained a p-value ≤ 0.05, then all we would know is that these two sample means are statistically different from one another at the 0.05 level.

No way to determine accuracy unless population parameters are known, or informed hypotheticals.

Theoretically, as sample size increases so does a sample statistic's precision and accuracy.

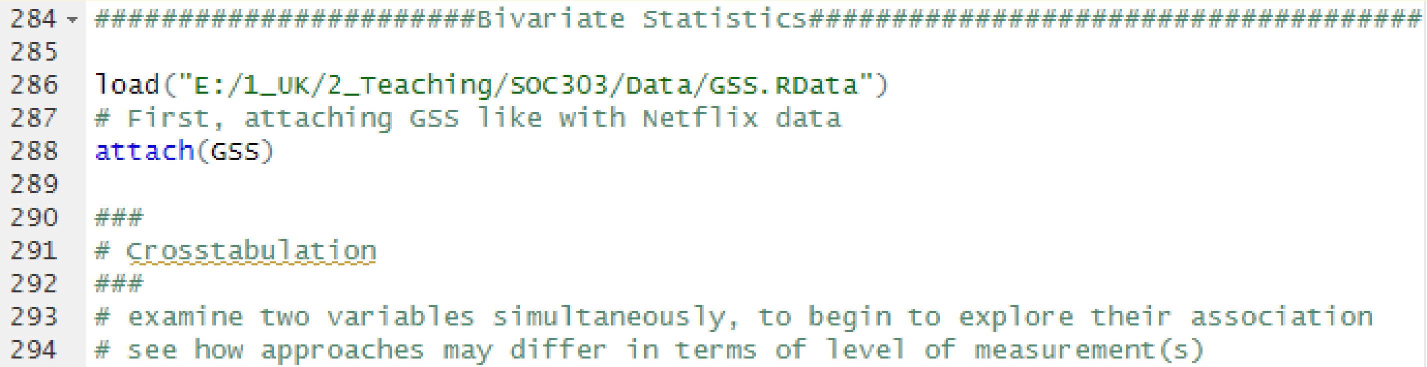
Univariate vs bivariate statistics

- Thus far, we've only worked with one single variable at a time
  - univariate statistics

- This is useful for summarizing sample data in terms of central tendency and dispersion, or
  - accounting for uncertainty by estimating a sample statistic's precision and
    - its accuracy if population parameters are known

- Bivariate statistics: analysis of two variables simultaneously to determine their association
  - see how quantitative analytic techniques and statistical tests differ by level of measurement

See RScript_3 in R + RStudio instructions module on Canvas



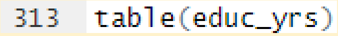
```
#######################Bivariate Statistics#####################################

load("E:/1_UK/2_Teaching/SOC303/Data/GSS.RData")
# First, attaching GSS like with Netflix data
attach(GSS)

###
# Crosstabulation
###
# examine two variables simultaneously, to begin to explore their association
# see how approaches may differ in terms of level of measurement(s)
```

Continuous-continuous

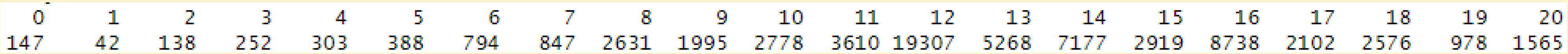
- Is mean years of education greater among relatively more recent birth cohorts?

- GSS includes 34 random samples with surveys administered from 1972-2004 (N=64,555)
  - birth cohort ranges from 1883-2004

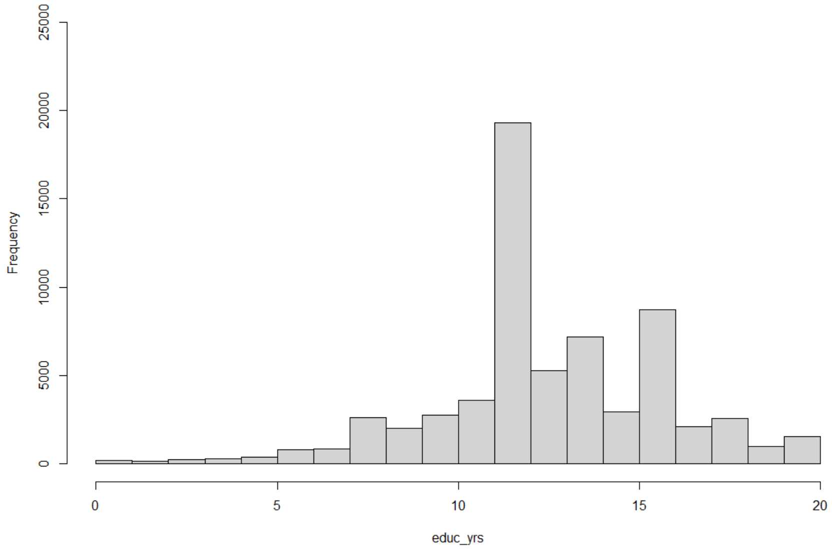
The distribution of education ($\bar{X} = 12.98\ (3.19)$) looks like this,



```
313  table(educ_yrs)
```

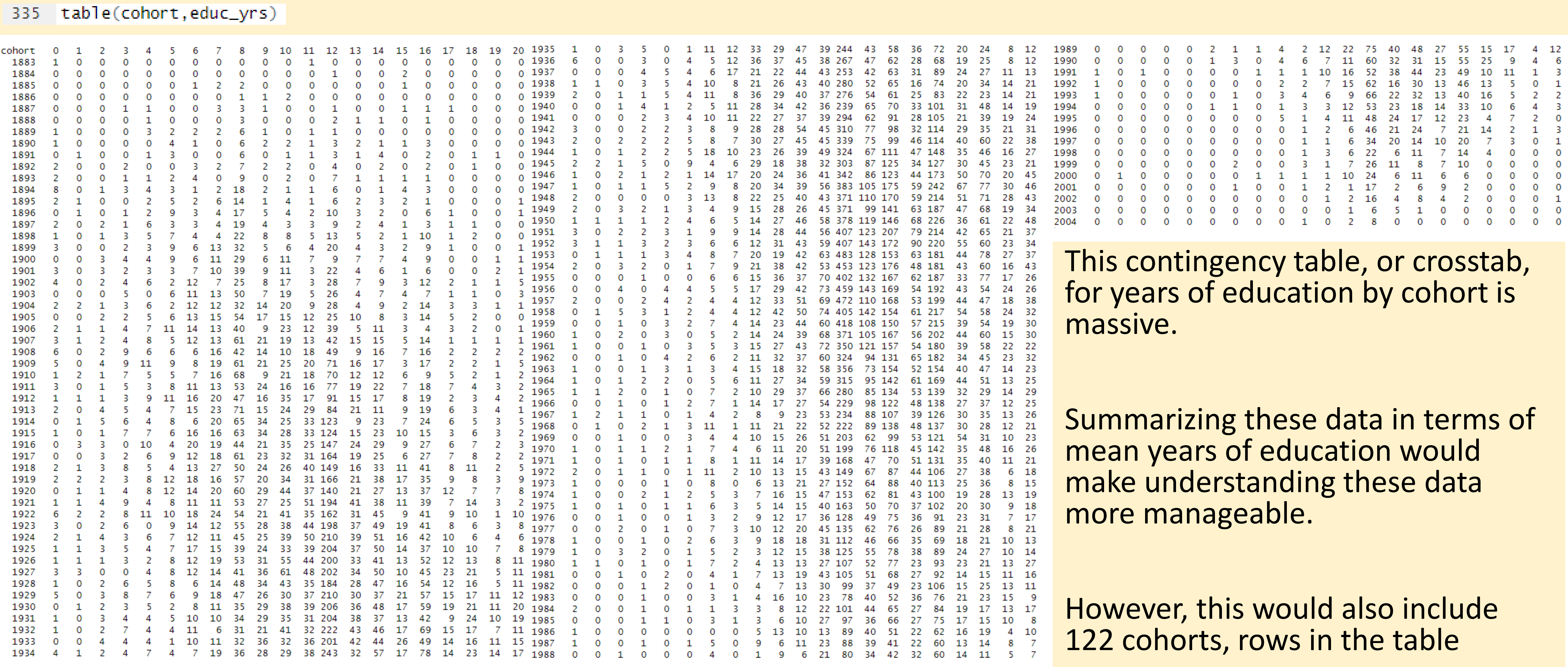


```
  0     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19    20
147    42   138   252   303   388   794   847  2631  1995  2778  3610 19307  5268  7177  2919  8738  2102  2576   978  1565
```



but we are interested in education by cohort….

Continuous-continuous



```
335  table(cohort,educ_yrs)
```

```
cohort   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
  1883   1   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
  1884   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   2   0   0   0   0   0
  1885   0   0   0   0   0   0   1   2   2   0   0   0   0   0   0   1   0   0   0   0   0
  1886   0   0   0   0   0   0   0   0   1   1   2   0   0   0   0   0   0   0   0   0   0
  1887   0   0   0   1   1   0   0   3   3   1   0   0   1   0   0   1   1   1   0   0   0
  1888   0   0   0   0   1   0   0   0   3   0   0   0   2   1   1   0   1   0   0   0   0
  1889   1   0   0   0   3   2   2   2   6   1   0   1   1   0   0   0   0   0   0   0   0
  1890   1   0   0   0   0   4   1   0   6   2   2   1   3   2   1   1   3   0   0   0   0
  1891   0   1   0   0   1   3   0   0   6   0   1   1   3   1   4   0   2   0   1   1   0
  1892   2   0   0   2   0   0   3   2   7   2   2   0   4   0   2   0   2   0   1   0   0
  1893   2   0   0   1   1   2   4   0   9   0   2   0   7   1   1   1   1   0   0   0   0
  1894   8   0   1   3   4   3   1   2  18   2   1   1   6   0   1   4   3   0   0   0   0
  1895   2   1   0   0   2   5   2   6  14   1   4   1   6   2   3   2   1   0   0   0   1
  1896   0   1   0   1   2   9   3   4  17   5   4   2  10   3   2   0   6   1   0   0   1
  1897   2   0   2   1   6   3   3   4  19   4   3   3   9   2   4   1   3   1   1   0   0
  1898   1   0   1   3   5   7   4   4  22   8   8   5  13   5   2   1  10   1   2   0   0
  1899   3   0   0   2   3   9   6  13  32   5   6   4  20   4   3   2   9   1   0   0   1
  1900   0   0   3   4   4   9   6  11  29   6  11   7   9   7   7   4   9   0   0   1   1
  1901   3   0   3   2   3   3   7  10  39   9  11   3  22   4   6   1   6   0   0   2   1
  1902   4   0   2   4   6   2  12   7  25   8  17   3  28   7   9   3  12   2   1   1   5
  1903   0   0   0   5   0   6  11  13  50   7  19   5  26   4   7   4   7   1   1   0   3
  1904   2   2   1   3   6   2  12  12  32  14  20   9  28   4   9   2  14   3   3   1   1
  1905   0   0   2   2   5   6  13  15  54  17  15  12  25  10   8   3  14   5   2   0   0
  1906   2   1   1   4   7  11  14  13  40   9  23  12  39   5  11   3   4   3   2   0   1
  1907   3   1   2   4   8   5  12  13  61  21  19  13  42  15  15   5  14   1   1   1   1
  1908   6   0   2   9   6   6   6  16  42  14  10  18  49   9  16   7  16   2   2   2   2
  1909   5   0   4   9  11   9   8  19  61  21  25  20  71  16  17   3  17   2   2   1   5
  1910   1   2   1   7   5   5   7  16  68   9  21  18  70  12  12   6   9   5   2   1   2
  1911   3   0   1   5   3   8  11  13  53  24  16  16  77  19  22   7  18   7   4   3   2
  1912   1   1   1   3   9  11  16  20  47  16  35  17  91  15  17   8  19   2   3   4   2
  1913   2   0   4   5   4   7  15  23  71  15  24  29  84  21  11   9  19   6   3   4   1
  1914   0   1   5   6   4   8   6  20  65  34  25  33 123   9  23   7  24   6   5   3   5
  1915   1   0   1   7   7   6  16  16  63  34  28  33 124  15  23  10  15   3   6   3   2
  1916   0   3   3   0  10   4  20  19  44  21  35  25 147  24  29   9  27   6   7   2   3
  1917   0   0   3   2   6   9  12  18  61  23  32  31 164  19  25   6  27   7   8   2   2
  1918   2   1   3   8   5   4  13  27  50  24  26  40 149  16  33  11  41   8  11   2   5
  1919   2   2   2   3   8  12  18  16  57  20  34  31 166  21  38  17  35   9   8   3   9
  1920   0   1   1   4   8  12  14  20  60  29  44  37 140  21  27  13  37  12   7   7   8
  1921   1   1   4   9   4   8  11  11  53  27  25  51 194  41  38  11  39   7  14   3   2
  1922   6   2   2   8  11  10  18  24  54  21  41  35 162  31  45   9  41   9  10   1  10
  1923   3   0   2   6   0   9  14  12  55  28  38  44 198  37  49  19  41   8   6   3   8
  1924   2   1   4   3   6   7  12  11  45  25  39  50 210  39  51  16  42  10   6   4   6
  1925   1   1   3   5   4   7  17  15  39  24  33  39 204  37  50  14  37  10  10   7   8
  1926   1   1   1   3   2   8  12  19  53  31  55  44 200  33  41  13  52  12  13   8  11
  1927   3   3   0   0   4   8  12  14  41  36  61  48 202  34  50  10  45  23  21   5  11
  1928   1   0   2   6   5   8   6  14  48  34  43  35 184  28  47  16  54  12  16   5  11
  1929   5   0   3   8   7   6   9  18  47  26  30  37 210  30  37  21  57  15  17  11  12
  1930   0   1   2   3   5   2   8  11  35  29  38  39 206  36  48  17  59  19  21  11  20
  1931   1   0   3   4   4   5  10  10  34  29  35  31 204  38  37  13  42   9  24  10  19
  1932   1   0   2   7   4   4  11   6  31  21  41  32 222  43  46  17  69  15  17   7  11
  1933   0   0   4   4   4   1  10  11  32  36  32  36 201  42  44  26  49  14  16  11  15
  1934   4   1   2   4   7   4   7  19  36  28  29  38 243  32  57  17  78  14  23  14  17
  1935   1   0   3   5   0   1  11  12  33  29  47  39 244  43  58  36  72  20  24   8  12
  1936   6   0   0   3   0   4   5  12  36  37  45  38 267  47  62  28  68  19  25   8  12
  1937   0   0   0   4   5   4   6  17  21  22  44  43 253  42  63  31  89  24  27  11  13
  1938   1   1   0   3   5   4  10   8  21  26  43  40 280  52  65  16  74  20  34  14  21
  1939   2   0   1   1   5   4  11   8  36  29  40  37 276  54  61  25  83  22  23  14  21
  1940   0   0   1   4   1   2   5  11  28  34  42  36 239  65  70  33 101  31  48  14  19
  1941   0   0   0   2   3   4  10  11  22  27  37  39 294  62  91  28 105  21  39  19  24
  1942   3   0   0   2   2   3   8   9  28  28  54  45 310  77  98  32 114  29  35  21  31
  1943   2   0   2   2   2   5   8   7  30  27  45  45 339  75  99  46 114  40  60  22  38
  1944   1   0   1   2   2   5  18  10  23  26  39  49 324  67 111  47 148  35  46  16  27
  1945   2   2   1   5   0   9   4   6  29  18  38  32 303  87 125  34 127  30  45  23  21
  1946   1   0   2   1   2   1  14  17  20  24  36  41 342  86 123  44 173  50  70  20  45
  1947   1   0   1   1   5   2   9   8  20  34  39  56 383 105 175  59 242  67  77  30  46
  1948   2   0   0   0   0   3  13   8  22  25  40  43 371 110 170  59 214  51  71  28  43
  1949   2   0   3   2   1   3   4   9  15  28  26  45 371  99 141  63 187  47  68  19  34
  1950   1   1   1   1   2   4   6   5  14  27  46  58 378 119 146  68 226  36  61  22  48
  1951   3   0   2   2   3   1   9   9  14  28  44  56 407 123 207  79 214  42  65  21  37
  1952   3   1   1   3   2   3   6   6  12  31  43  59 407 143 172  90 220  55  60  23  34
  1953   0   1   1   1   3   4   8   7  20  19  42  63 483 128 153  63 181  44  78  27  37
  1954   2   0   3   2   0   1   7   9  21  38  42  53 453 123 176  48 181  43  60  16  43
  1955   0   0   0   1   0   0   6   6  15  36  37  70 402 132 167  62 187  33  77  17  26
  1956   0   0   4   0   4   4   5   5  17  29  42  73 459 143 169  54 192  43  54  24  26
  1957   2   0   0   2   4   2   4   4  12  33  51  69 472 110 168  53 199  44  47  18  38
  1958   0   1   5   3   1   2   4   4  12  42  50  74 405 142 154  61 217  54  58  24  32
  1959   0   0   1   0   3   2   7   4  14  23  44  60 418 108 150  57 215  39  54  19  30
  1960   1   0   2   0   3   0   5   2  14  24  39  68 371 105 167  56 202  44  60  15  30
  1961   1   0   0   1   0   3   5   3  15  27  43  72 350 121 157  54 180  39  58  22  22
  1962   0   0   1   0   4   2   6   2  11  32  37  60 324  94 131  65 182  34  45  23  32
  1963   1   0   0   1   3   1   3   4  15  18  32  58 356  73 154  52 154  40  47  14  23
  1964   1   0   1   2   2   0   5   6  11  27  34  59 315  95 142  61 169  44  51  13  25
  1965   1   1   2   0   1   0   7   2  10  29  37  66 280  85 134  53 139  32  29  14  29
  1966   0   0   1   0   1   2   7   1  14  17  27  54 229  98 122  48 138  27  37  12  25
  1967   1   2   1   1   0   1   4   2   8   9  23  53 234  88 107  39 126  30  35  13  26
  1968   0   1   0   2   1   3  11   1  11  21  22  52 222  89 138  48 137  30  28  12  21
  1969   0   0   1   0   0   3   4   4  10  15  26  51 203  62  99  53 121  54  31  10  23
  1970   1   0   1   1   2   1   7   4   6  11  20  51 199  76 118  45 142  35  48  16  26
  1971   1   0   1   0   1   1   8   1  11  14  17  39 168  47  70  51 131  35  40  11  21
  1972   2   0   1   1   0   1  11   2  10  13  15  43 149  67  87  44 106  27  38   6  18
  1973   1   0   0   0   1   0   8   0   6  13  21  27 152  64  88  40 113  25  36   8  15
  1974   1   0   0   2   1   2   5   3   7  16  15  47 153  62  81  43 100  19  28  13  19
  1975   1   0   1   0   1   1   6   3   5  14  15  40 163  50  70  37 102  20  30   9  18
  1976   0   0   1   0   0   1   3   2   9  12  17  36 128  49  75  36  91  23  31   7  17
  1977   0   0   2   0   1   0   7   3  10  12  20  45 135  62  76  26  89  21  28   8  21
  1978   1   0   0   1   0   2   6   3   9  18  18  31 112  46  66  35  69  18  21  10  13
  1979   1   0   3   2   0   1   5   2   3  12  15  38 125  55  78  38  89  24  27  10  14
  1980   1   1   0   1   0   1   7   2   4  13  13  27 107  52  77  23  93  23  21  13  27
  1981   0   0   1   0   2   0   4   1   7  13  19  43 105  51  68  27  92  14  15  11  16
  1982   0   0   0   1   2   0   1   0   4   7  13  30  99  37  49  23 106  15  25  13  11
  1983   0   0   0   1   0   0   3   1   4  16  10  23  78  40  52  36  76  21  23  15   9
  1984   2   0   0   1   0   1   1   3   3   8  12  22 101  44  65  27  84  19  17  13  17
  1985   0   0   0   1   1   0   3   1   3   6  10  27  97  36  66  27  75  17  15  10   8
  1986   1   0   0   0   0   0   0   0   5  13  10  13  89  40  51  22  62  16  19   4  10
  1987   1   0   0   1   0   1   5   0   9   6  11  23  88  39  41  22  60  13  14   8   7
  1988   0   0   1   0   0   0   4   0   1   9   6  21  80  34  42  32  60  14  11   5   7
  1989   0   0   0   0   0   2   1   1   4   2  12  22  75  40  48  27  55  15  17   4  12
  1990   0   0   0   0   0   1   3   0   4   6   7  11  60  32  31  15  55  25   9   4   6
  1991   1   0   1   0   0   0   0   1   1   1  10  16  52  38  44  23  49  10  11   1   3
  1992   1   0   0   0   0   0   0   0   2   2   7  15  62  16  30  13  46  13   5   0   1
  1993   1   0   0   0   0   0   1   0   3   4   6   9  66  22  32  13  40  16   5   2   2
  1994   0   0   0   0   0   1   1   0   1   3   3  12  53  23  18  14  33  10   6   4   3
  1995   0   0   0   0   0   0   0   0   5   1   4  11  48  24  17  12  23   4   7   2   0
  1996   0   0   0   0   0   0   0   0   0   1   2   6  46  21  24   7  21  14   2   1   3
  1997   0   0   0   0   0   0   0   0   0   1   1   6  34  20  14  10  20   7   3   0   1
  1998   0   0   0   0   0   0   0   0   0   1   3   6  22   6  11   7  14   4   0   0   0
  1999   0   0   0   0   0   0   2   0   0   3   1   7  26  11   8   7  10   0   0   0   0
  2000   0   1   0   0   0   0   0   1   1   1   1  10  24   6  11   6   6   0   0   0   0
  2001   0   0   0   0   0   0   1   0   0   1   2   1  17   2   6   9   2   0   0   0   0
  2002   0   0   0   0   0   0   0   0   0   0   1   2  16   4   8   4   2   0   0   0   1
  2003   0   0   0   0   0   0   0   0   0   0   0   1   6   5   1   0   0   0   0   0   0
  2004   0   0   0   0   0   0   0   0   0   1   0   2   8   0   0   0   0   0   0   0   0
```

This contingency table, or crosstab, for years of education by cohort is massive.

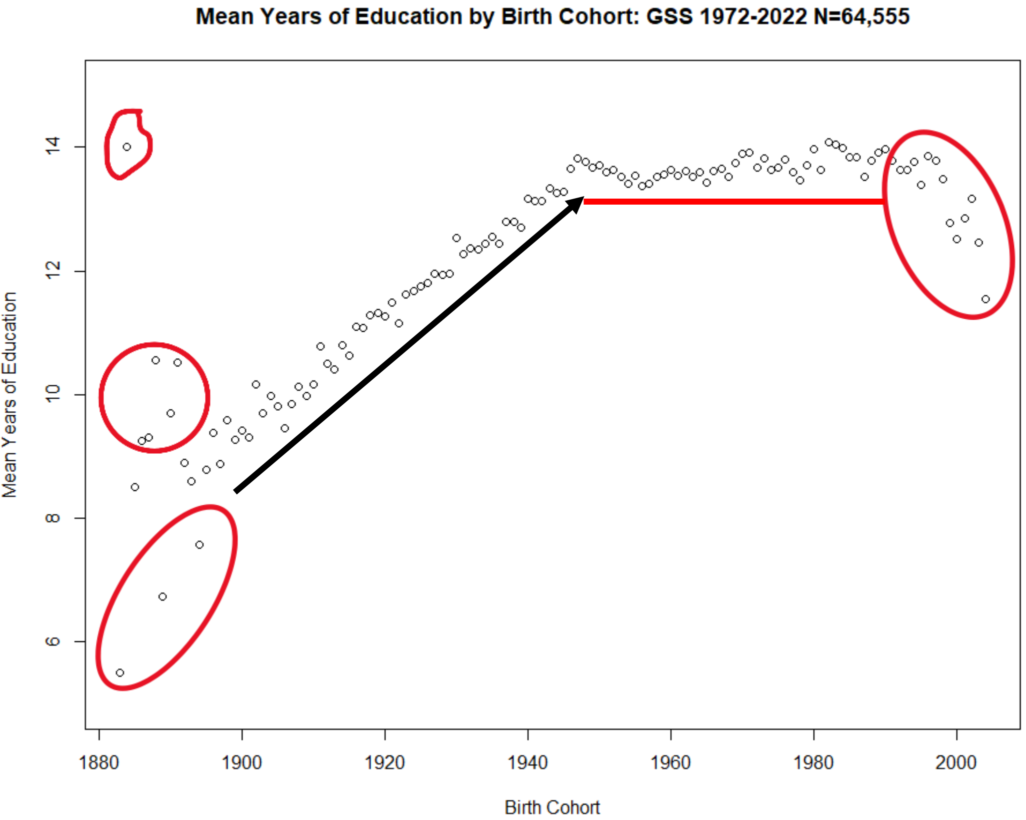
Summarizing these data in terms of mean years of education would make understanding these data more manageable.

However, this would also include 122 cohorts, rows in the table

Let’s consider mean years of education by cohort depicted in a plot…

Continuous-continuous

- Is mean years of education greater among relatively more recent birth cohorts?
    - It looks so, but what’s up with…



H_0: $mean\ years\ of\ edu \leq among\ more\ recent\ cohorts$
H_a: $mean\ years\ of\ edu > among\ more\ recent\ cohorts$

Last week we determined whether one mean was statistically different from another mean.

The present hypothesis test involves multiple means, which requires a different type of hypothesis test – more complex than we are ready for now, but we will build up to this.

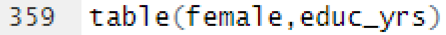
Let’s first explore some different bivariate combinations by level of measurement …

Continuous-nominal

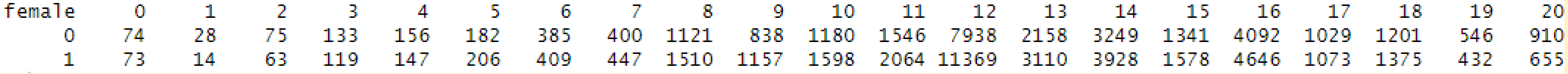
- Is mean years of education greater among males compared females?

$H_0: male\ \bar{X}_{edu} \leq female\ \bar{X}_{edu}$
$H_a: male\ \bar{X}_{edu} > female\ \bar{X}_{edu}$

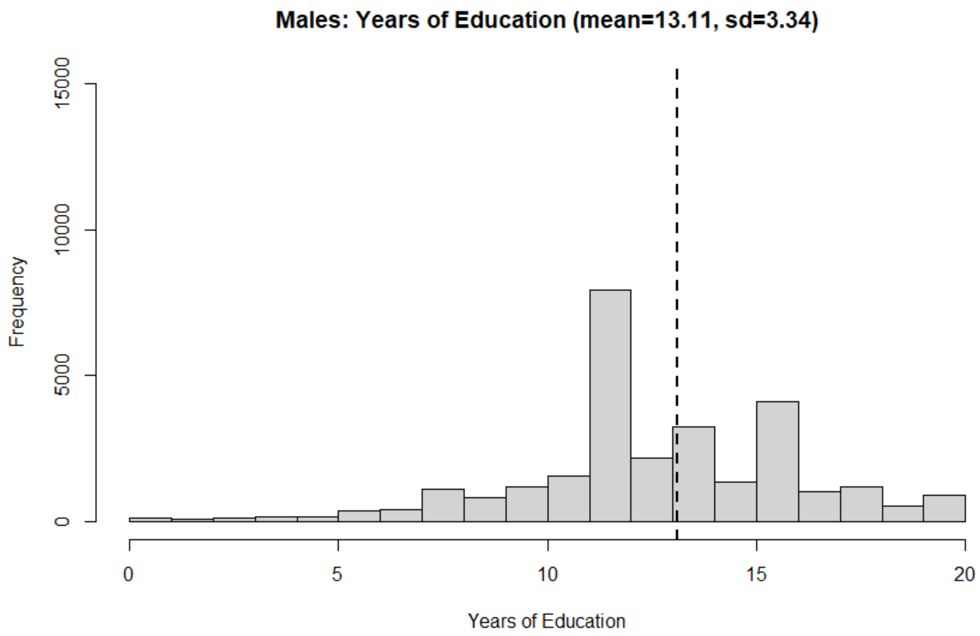
The distribution of education by sex looks like this

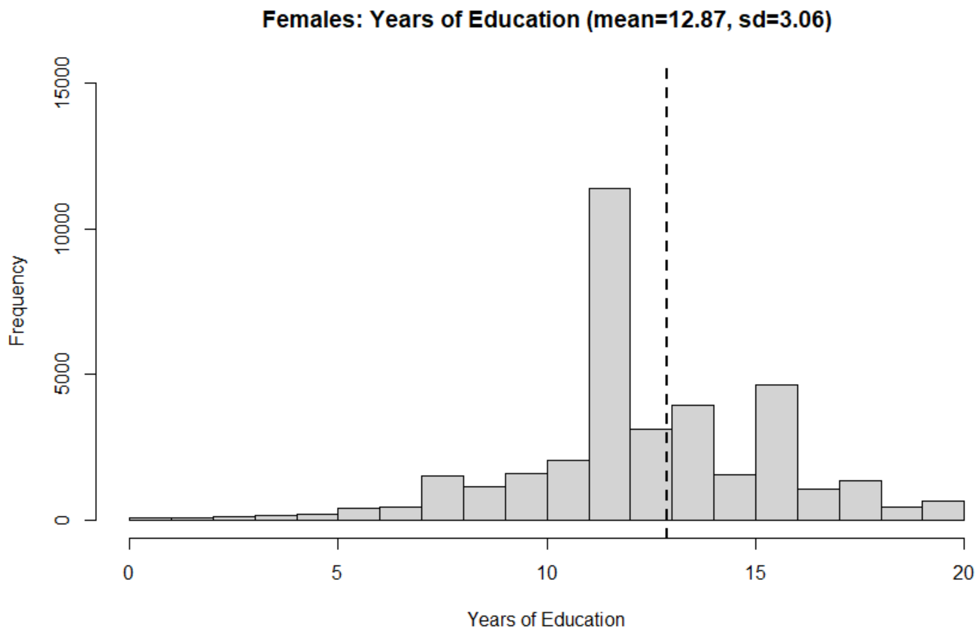


```
table(female,educ_yrs)
```



```
female     0     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19    20
     0    74    28    75   133   156   182   385   400  1121   838  1180  1546  7938  2158  3249  1341  4092  1029  1201   546   910
     1    73    14    63   119   147   206   409   447  1510  1157  1598  2064 11369  3110  3928  1578  4646  1073  1375   432   655
```





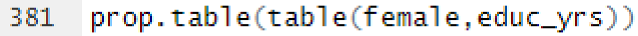
See how considering within-group proportions can be helpful in bivariate statistics…

Continuous-nominal

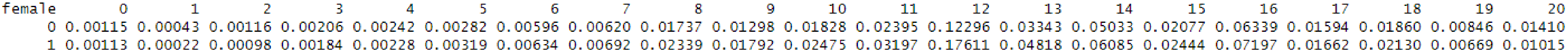
- Is mean years of education greater among males compared females?

$H_0: male\ \bar{X}_{edu} \leq female\ \bar{X}_{edu}$
$H_a: male\ \bar{X}_{edu} > female\ \bar{X}_{edu}$

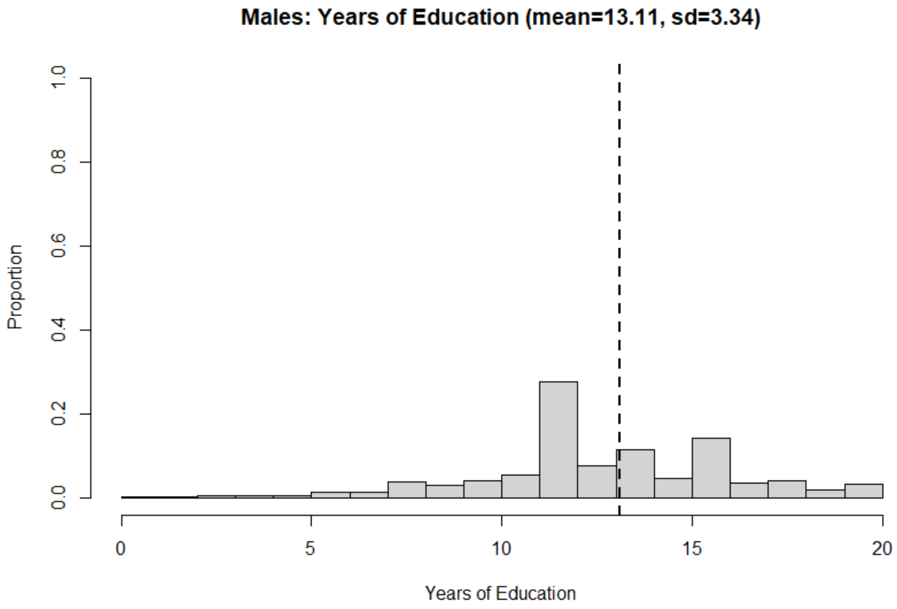
The distribution of education by sex looks like this

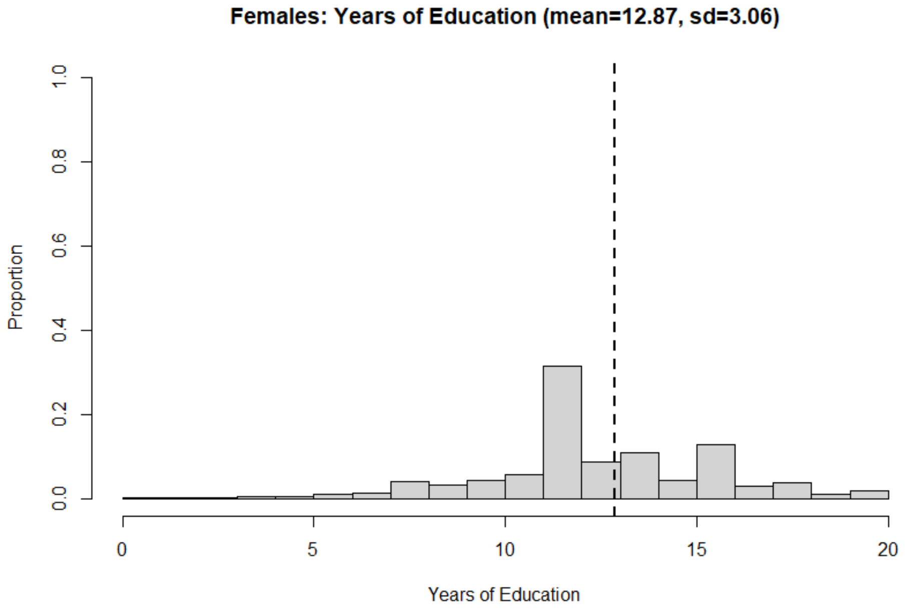


```
prop.table(table(female,educ_yrs))
```



```
female        0       1       2       3       4       5       6       7       8       9      10      11      12      13      14      15      16      17      18      19      20
     0 0.00115 0.00043 0.00116 0.00206 0.00242 0.00282 0.00596 0.00620 0.01737 0.01298 0.01828 0.02395 0.12296 0.03343 0.05033 0.02077 0.06339 0.01594 0.01860 0.00846 0.01410
     1 0.00113 0.00022 0.00098 0.00184 0.00228 0.00319 0.00634 0.00692 0.02339 0.01792 0.02475 0.03197 0.17611 0.04818 0.06085 0.02444 0.07197 0.01662 0.02130 0.00669 0.01015
```





See how considering within-group proportions can be helpful in bivariate statistics…

Continuous-nominal

- Is mean years of education greater among males compared females?

$H_0: male\ \bar{X}_{edu} \le female\ \bar{X}_{edu}$
$H_a: male\ \bar{X}_{edu} > female\ \bar{X}_{edu}$

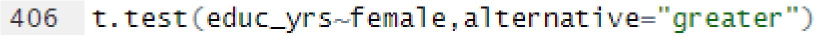
How many tails in this example, where's the critical region?

two-sample t-test

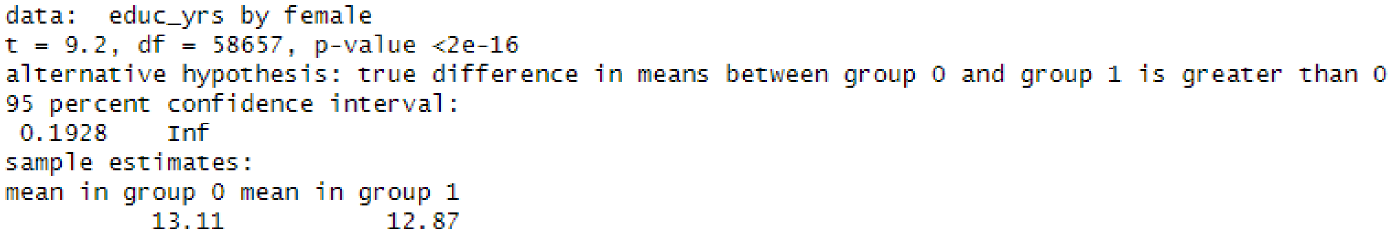
$$t = \frac{\bar{X}_1 - \bar{X}_2}{sd_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

sd_p: pooled standard deviation

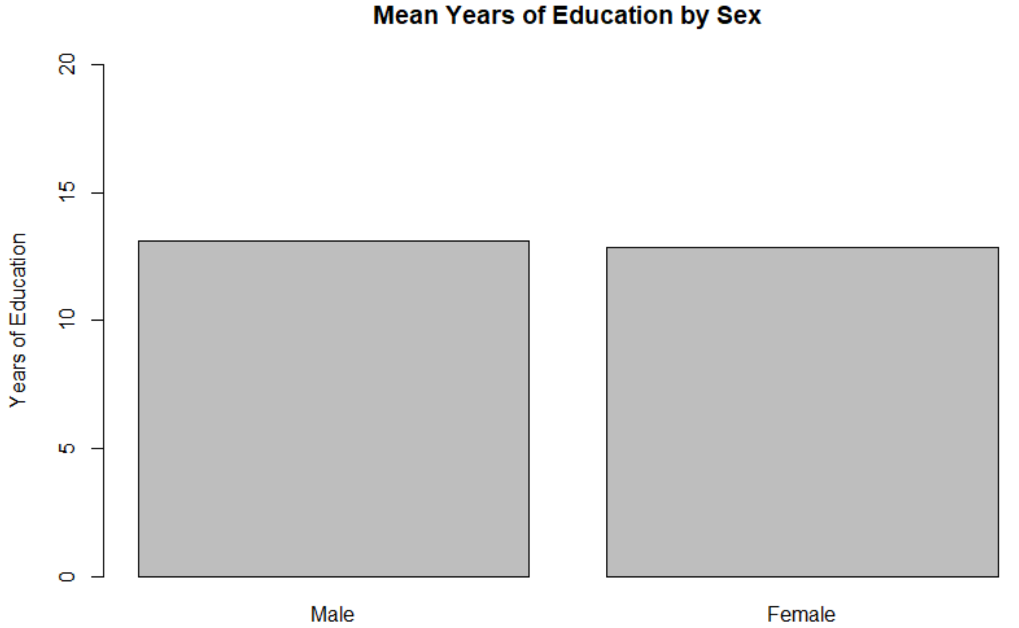
$$sd_p = \sqrt{\frac{(n_1-1)sd_1^2 + (n_2-1)sd_2^2}{n_1+n_2-2}}$$



```
t.test(educ_yrs~female,alternative="greater")
```



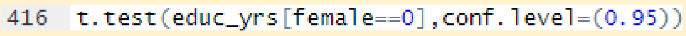
```
data:  educ_yrs by female
t = 9.2, df = 58657, p-value <2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is greater than 0
95 percent confidence interval:
 0.1928    Inf
sample estimates:
mean in group 0 mean in group 1
          13.11           12.87
```



The difference in means is small, but statistically significant.

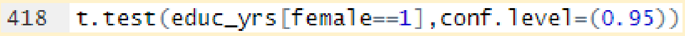
Why might that be? Hint: consider n, SE, and MoE

Male 95% CI (13.07,13.15)



```
t.test(educ_yrs[female==0],conf.level=(0.95))
```

Female 95% CI (12.84,12.90)



```
t.test(educ_yrs[female==1],conf.level=(0.95))
```

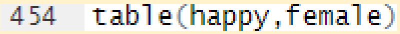
Let's further explore different bivariate combinations by level of measurement …

Ordinal-nominal

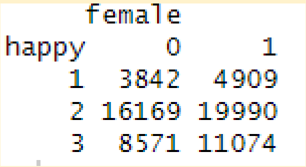
- Unlike continuous variables, the mean is not very meaningful for categorical variables

- Let’s consider whether happiness differs by sex
  - 1=not very happy, 2=pretty happy, 3=very happy

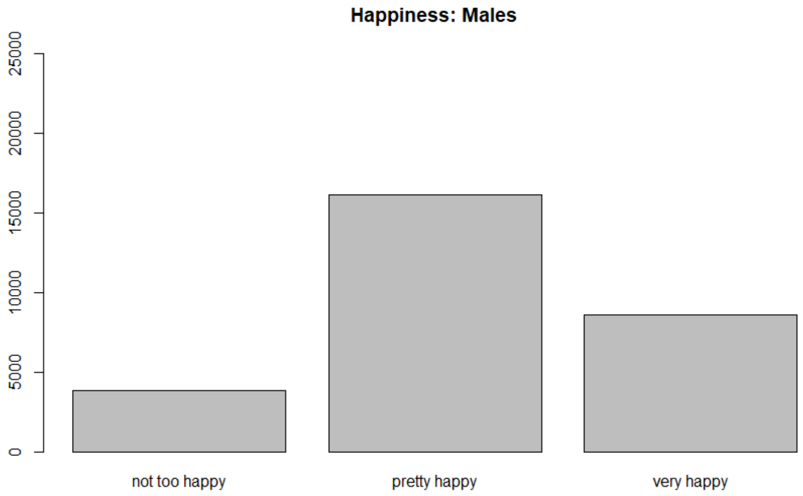
The distribution of happiness by sex looks like this



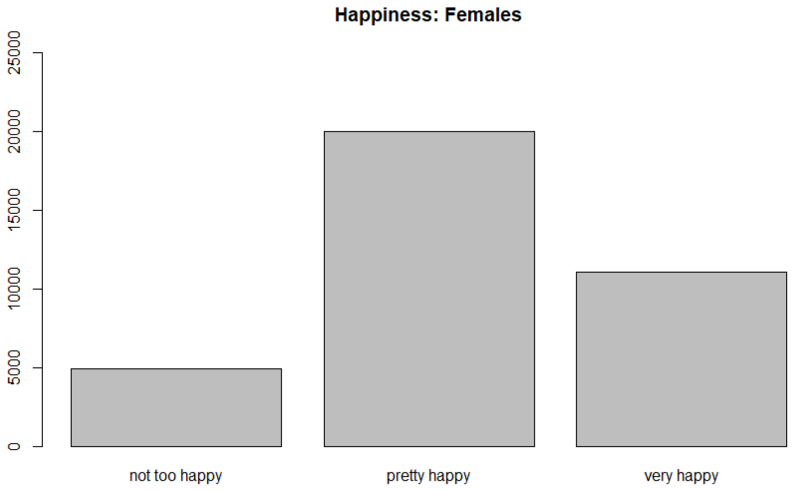
```
table(happy,female)
```



```
       female
happy       0     1
    1    3842  4909
    2   16169 19990
    3    8571 11074
```



Happiness: Males



Happiness: Females

See how considering within-group proportions can be helpful in bivariate statistics…

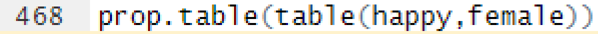
Ordinal-nominal

- Does happiness differ by sex?

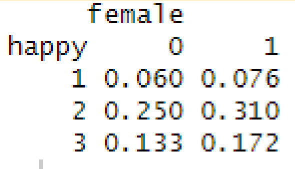
H_0: happiness is independent of sex

H_a: happiness is dependent on sex

The distribution of happiness by sex looks like this

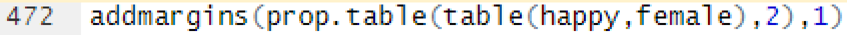


```
prop.table(table(happy,female))
```

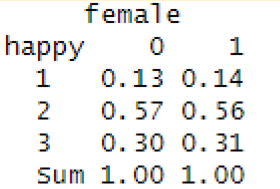


```
      female
happy      0     1
    1 0.060 0.076
    2 0.250 0.310
    3 0.133 0.172
```

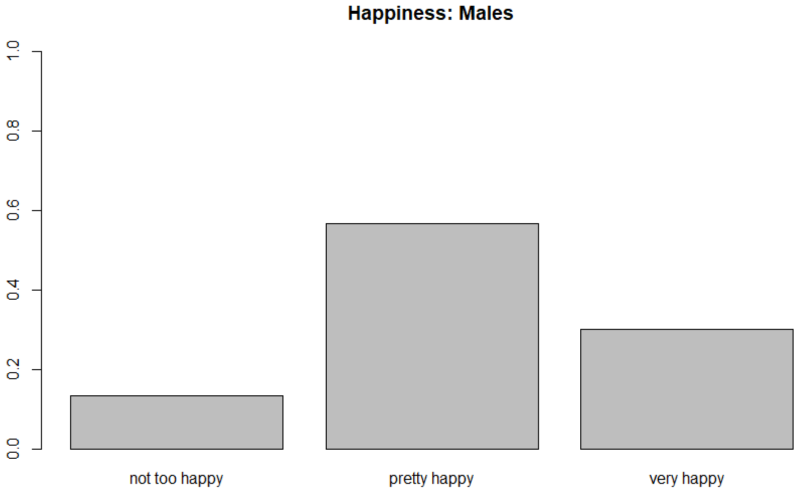
Margins can be an incredibly helpful tool

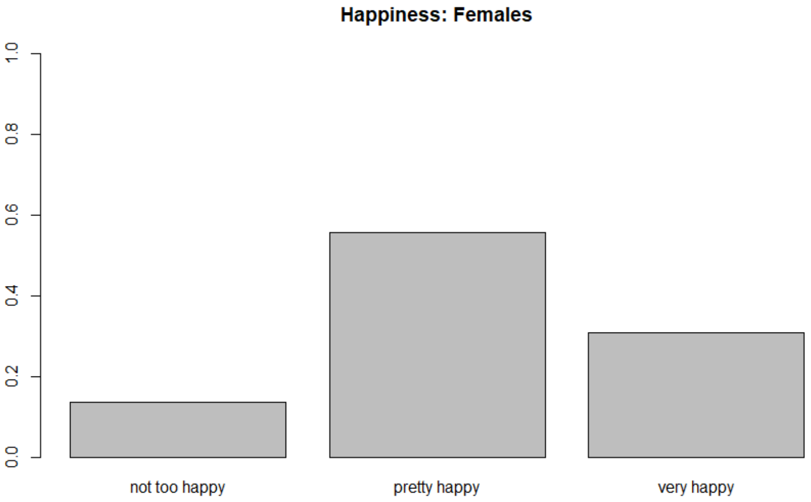


```
addmargins(prop.table(table(happy,female),2),1)
```



```
      female
happy     0    1
  1    0.13 0.14
  2    0.57 0.56
  3    0.30 0.31
  Sum 1.00 1.00
```





We need a different type of hypothesis test to determine if happiness differs by sex…

Hypothesis testing: chi-squared

Same steps as with any hypothesis test,

1. Make assumptions and meet test requirements
   - random sample; level of measurement; sample size; are parameters un/known
2. State the null (H_0) and alternative (H_a) hypothesis
   - H_0 no difference; $H_a$there is a difference
3. Choose a significance level (critical value)
   - e.g., $\alpha = 0.05$
4. Compute the test statistic
5. Draw a conclusion and interpret the test results
   - If $x^2 > critical\ value$ then reject H_0

but different procedures for computing the test statistic…

Chi-squared test of independence

4a. Create a bivariate contingency table

4b. Find the discrepancies between observed (f_o) and expected (f_e) counts

4c. Sum the discrepancies between f_o and f_e

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

4d. Determine statistical significance

$$df = (rows - 1) \times (\text{columns} - 1)$$

but different procedures for computing the test statistic…

Chi-squared test of independence: step 4a

4a. Create a bivariate contingency table, sometimes called a crosstab

	Columns (IV)		
Rows (DV)	Column 2	Column 1	
Row 1	cell a	cell b	Row Marginal 3
Row 2	cell c	cell d	Row Marginal 2
	Column Marginal 1	Column Marginal 2	N

Let’s check out our example for happiness by sex…

Chi-squared test of independence: step 4a

4a. Create a bivariate contingency table, sometimes called a crosstab

	Columns (IV)		
	male	female	
not too happy	3842	4909	8751
pretty happy	16169	19990	36159
very happy	8571	11074	19645
	28582	35973	64555

Probability theory: if two events are independent then their joint probability equals the product of their respective marginal probabilities. So, ….

Chi-squared test of independence: step 4b

4b. Find the discrepancies between frequency observed (f_o) and frequency expected (f_e)

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

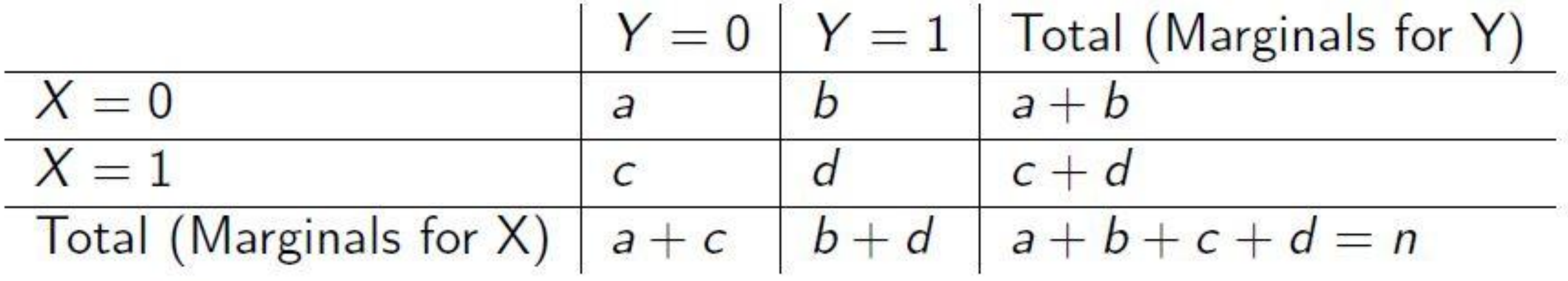
$f_o =$ the cell frequencies observed in the bivariate table
$f_e =$ the cell frequencies that would be expected if the variables were independent

$$f_e = \frac{row\ marginal\ \times column\ marginal}{N}$$

Let’s look at this in another way…

Chi-squared test of independence: step 4b

4b. Find the discrepancies between frequency observed (f_o) and frequency expected (f_e)



	$Y = 0$	$Y = 1$	Total (Marginals for Y)
$X = 0$	a	b	$a + b$
$X = 1$	c	d	$c + d$
Total (Marginals for X)	$a + c$	$b + d$	$a + b + c + d = n$

- If X and Y are independent, then $p(x, y) = p(x) \times p(y)$
- So, $p(x = 0, y = 0) = p(x = 0) \times p(y = 0) = \frac{a+c}{n} \times \frac{a+b}{n}$
- Expected cell count is: $\frac{(O_{x=0})(O_{y=0})}{n}$

Let’s check out our example for happiness by sex…

Chi-squared test of independence: step 4b-c

4b. Find the discrepancies between frequency observed (f_o) and frequency expected (f_e)

	Columns (IV)		
	male	female	Total
not too happy	3842	4909	8751
Expected	$\frac{(8751)(28582)}{64555} = 3875$	$\frac{(8751)(35973)}{64555} = 4876$	
$\frac{(O-E)^2}{E}$	0.27	0.22	0.49
pretty happy	16169	19990	36159
Expected	$\frac{(36159)(28582)}{64555} = 16009$	$\frac{(36159)(35973)}{64555} = 20149$	
$\frac{(O-E)^2}{E}$	1.59	1.26	2.85
very happy	8571	11074	19645
Expected	$\frac{(19645)(28582)}{64555} = 8698$	$\frac{(19645)(35973)}{64555} = 10947$	
$\frac{(O-E)^2}{E}$	1.85	1.47	3.32
Total	28582	35973	64555

4c.

$$\chi^2 = 0.27 + 0.22 + 1.59 + 1.26 + 1.85 + 1.47 = 6.66$$

Chi-squared test of independence: step 4d

4d. Determine statistical significance

- Use the x^2 probability distribution table to identify critical values
  - like the Z and t tables for a sample mean

- Find the critical value that corresponds with the degrees of freedom (df), and
  - your selected alpha (α), significance level

$$df = (rows - 1) \times (\mathrm{columns} - 1)$$

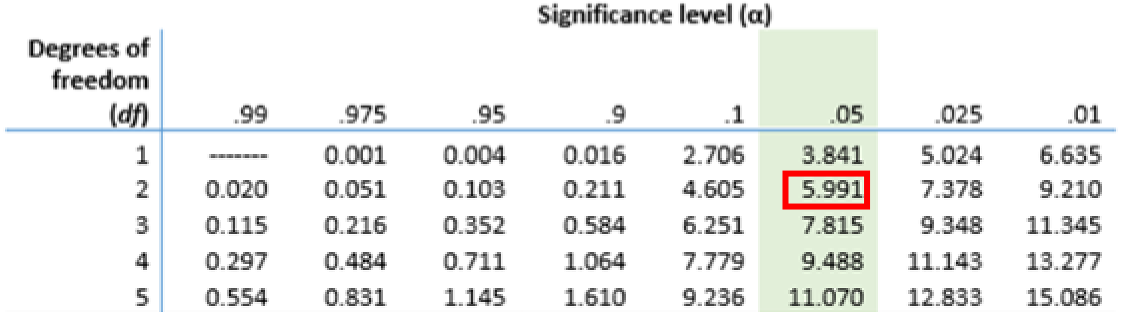
Like with the Z- and t-tests, I won’t have you do this, but let’s check it out…

Chi-squared test of independence: step 5

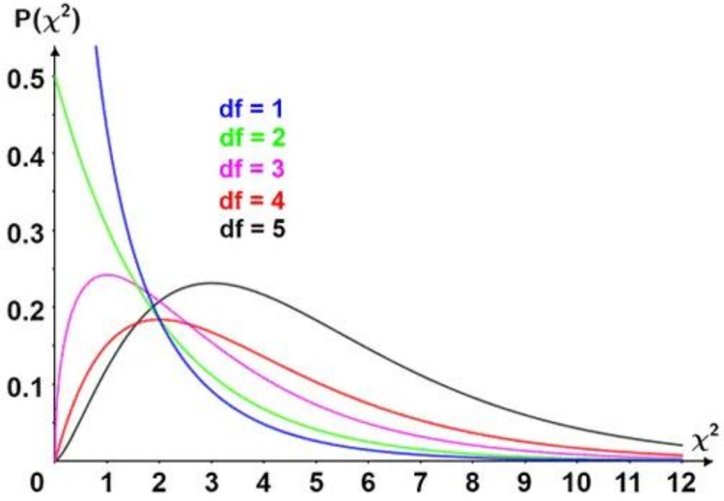
5. Draw a conclusion and interpret the test results

- If $x^2 > critical\ value$ then reject H_0
  - $df = (3-1) \times (2-1) = 2 \times 1 = 2$
  - $\alpha = 0.05$, or 95% confidence level

$$df = (rows - 1) \times (\text{columns} - 1)$$

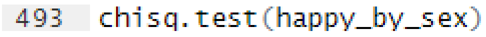


	Significance level (α)							
Degrees of freedom (*df*)	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086

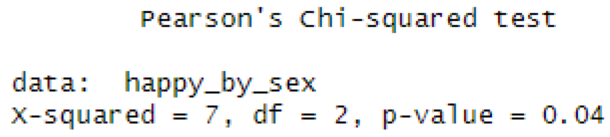


As the df increases the chi square distribution tends to normal distribution.

- 6.66 > 5.99, so reject H_0
  - there is a statistically significant association between happiness and biologically assigned sex
- Or, just use the chisq.test command in RStudio



```
493  chisq.test(happy_by_sex)
```



```
        Pearson's Chi-squared test

data:  happy_by_sex
X-squared = 7, df = 2, p-value = 0.04
```

Again, this is a large sample, so hypothesis tests are likely to be statistically significant

Chi-squared test of independence: limitations

- x^2 test statistics are unstable with small cell counts
  - Rule of thumb: don't use when expected cell counts <5
    - Instead, could use Fisher's exact test

- x^2 tests are more sensitive to sample size than other test statistics
  - e.g., in $\frac{(O-E)^2}{E}$ numerator grows faster than the denominator as *n* increase