Quantitative Sociological Analysis

Inferential Statistics
Hypothesis Testing and Bivariate Statistics

Part 7

April 15, 2025

Analysis of Variance (ANOVA)

- Thus far, in terms of continuous dependent variables we've only considered hypothesis tests to determine statistically significant differences in the mean between two groups
  - What if we're interested in three or more groups?

- ANOVA decomposes total variance of a continuous variable into
  - between-group variance: How much do group means differ?
  - within-group variance: How much do cases that compose a group differ?
- to determine if differences between group means are large after accounting for differences within groups that may lead to highly variable group means from sample to sample

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Y_{ij}: observed value for the *j*th subject in group I

μ: overall mean (grand mean)

τ_i: effect of group *i* (how much group *i*'s mean differs from the overall mean)

ε_{ij}: random error (individual deviation from the group mean)

ANOVA: example

See Rscript_4 in R + RStudio instructions module on Canvas

- Does mean years of education differ by political party affiliation?

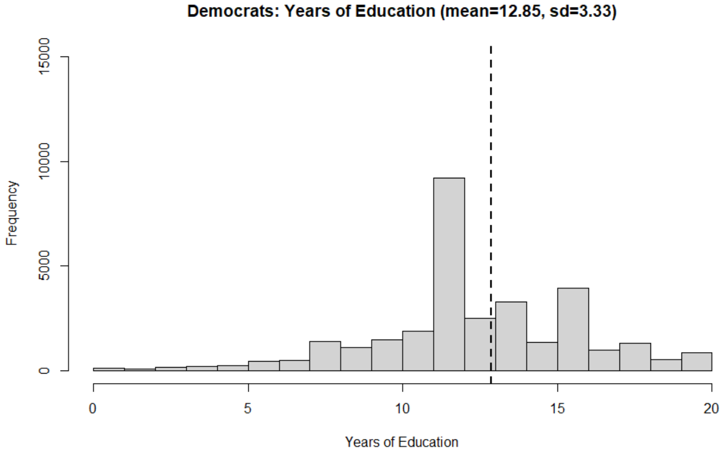
H_0: $Democrat\ \bar{X}_{edu} = IndepOther\ \bar{X}_{edu} = Republican\ \bar{X}_{edu}$

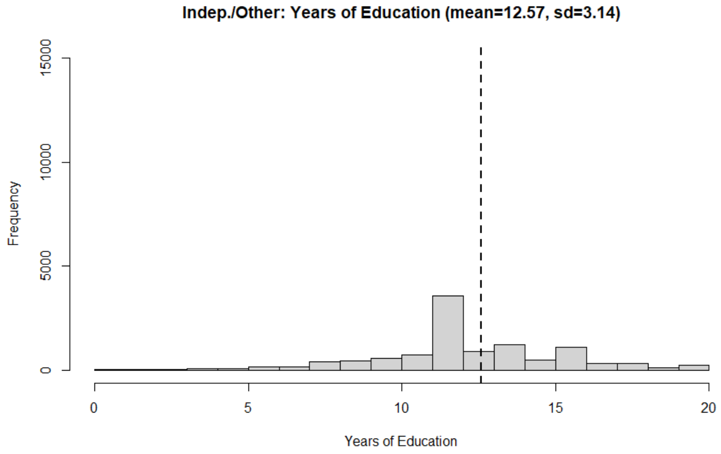
H_a: At least one group's $\bar{X}_{edu} \neq$ another group's $\bar{X}_{edu}$

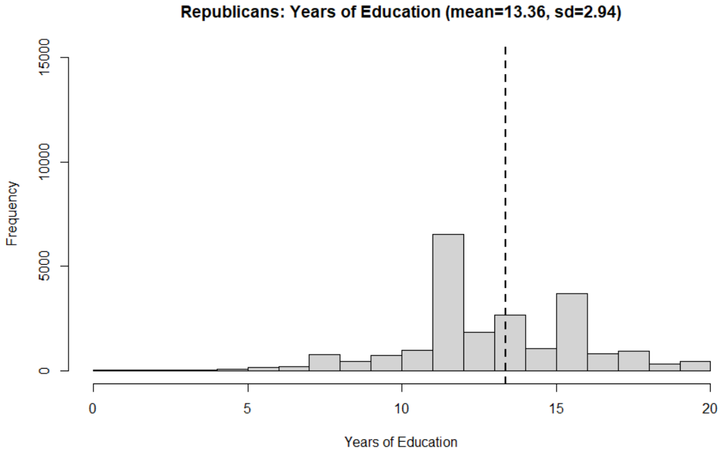
The distribution of education by political party affiliation looks like this

```
505   table(polit_party,educ_yrs)
```

polit_party	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	77	25	83	154	190	248	456	489	1413	1093	1485	1897	9189	2504	3282	1372	3948	975	1326	532	877
2	42	7	25	51	61	78	176	153	428	444	569	743	3570	899	1242	476	1099	308	326	128	225
3	28	10	30	47	52	62	162	205	790	458	724	970	6548	1865	2653	1071	3691	819	924	318	463







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

ANOVA: example continued

- Does mean years of education differ by political party affiliation?

$H_0: Democrat\ \bar{X}_{edu} = IndepOther\ \bar{X}_{edu} = Republican\ \bar{X}_{edu}$

H_a: At least one group's $\bar{X}_{edu} \neq$ another group's $\bar{X}_{edu}$

The distribution of education by political party affiliation looks like this

```
536  prop.table(table(polit_party,educ_yrs))
```

polit_party	0	1	2	3	4	5	6	7	8	9	10
1	0.00119	0.00039	0.00129	0.00239	0.00294	0.00384	0.00706	0.00757	0.02189	0.01693	0.02300
2	0.00065	0.00011	0.00039	0.00079	0.00094	0.00121	0.00273	0.00237	0.00663	0.00688	0.00881
3	0.00043	0.00015	0.00046	0.00073	0.00081	0.00096	0.00251	0.00318	0.01224	0.00709	0.01122

polit_party	11	12	13	14	15	16	17	18	19	20
1	0.02939	0.14234	0.03879	0.05084	0.02125	0.06116	0.01510	0.02054	0.00824	0.01359
2	0.01151	0.05530	0.01393	0.01924	0.00737	0.01702	0.00477	0.00505	0.00198	0.00349
3	0.01503	0.10143	0.02889	0.04110	0.01659	0.05718	0.01269	0.01431	0.00493	0.00717

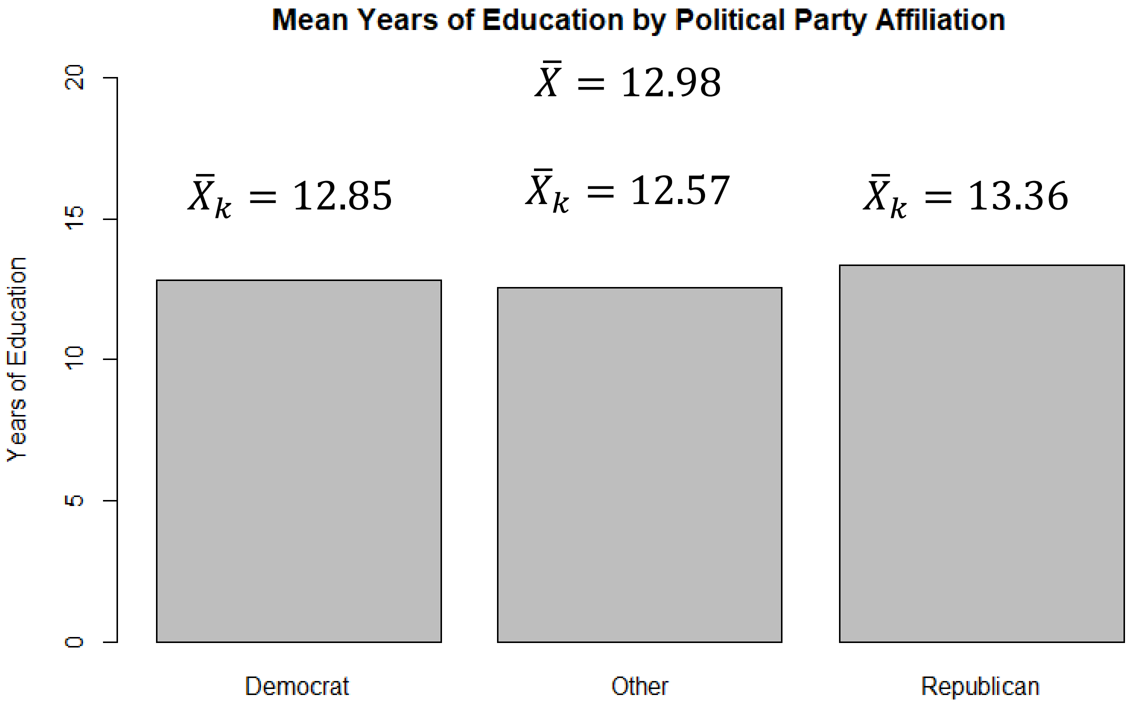
Let's see if there are any statistically significant differences between group means…

ANOVA: example continued

- Does mean years of education differ by political party affiliation?

H_0: $Democrat\ \bar{X}_{edu} = IndepOther\ \bar{X}_{edu} = Republican\ \bar{X}_{edu}$

H_a: At least one group's $\bar{X}_{edu} \neq$ another group's $\bar{X}_{edu}$



One-way ANOVA

1. Find total sum of squares (SST)
2. Find sum of square between (SSB)
3. Find sum of squares within (SSW)
4. Find degrees of freedom
5. Find mean square estimates
6. Find the F ratio

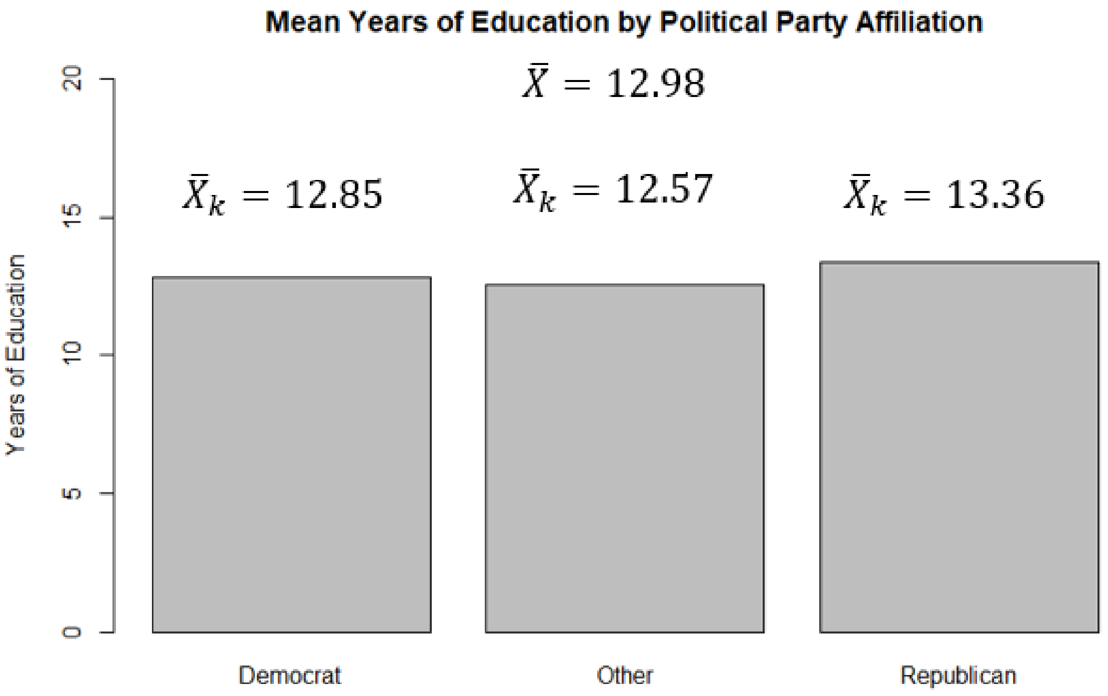
Let's break this down to see how the total variance is decomposed…

ANOVA: example (SST)

- Does mean years of education differ by political party affiliation?

H_0: $Democrat\ \bar{X}_{edu} = IndepOther\ \bar{X}_{edu} = Republican\ \bar{X}_{edu}$

H_a: At least one group's $\bar{X}_{edu} \neq$ another group's $\bar{X}_{edu}$



$\sum X_i^2 = 5{,}571{,}665$ $\quad \sum X_i^2 = 1{,}856{,}215$ $\quad \sum X_i^2 = 4{,}096{,}261$

$$SST = \sum X_i^2 - N\bar{X}^2$$

SST = (5,571,665 + 1,856,215 + 4,096,261) – (64,555)$(12.98)^2$

SST = 11,524,141 – (64,555)(168.3642)

SST = 11,524,141 – (64,555)(168.3642)

SST = 11,524,141 – 10,868,753.7

SST = 655,387.3

```
598  SST_equation<-sum((educ_yrs-mean(educ_yrs))^2)

> SST_equation
[1] 655387.3
```

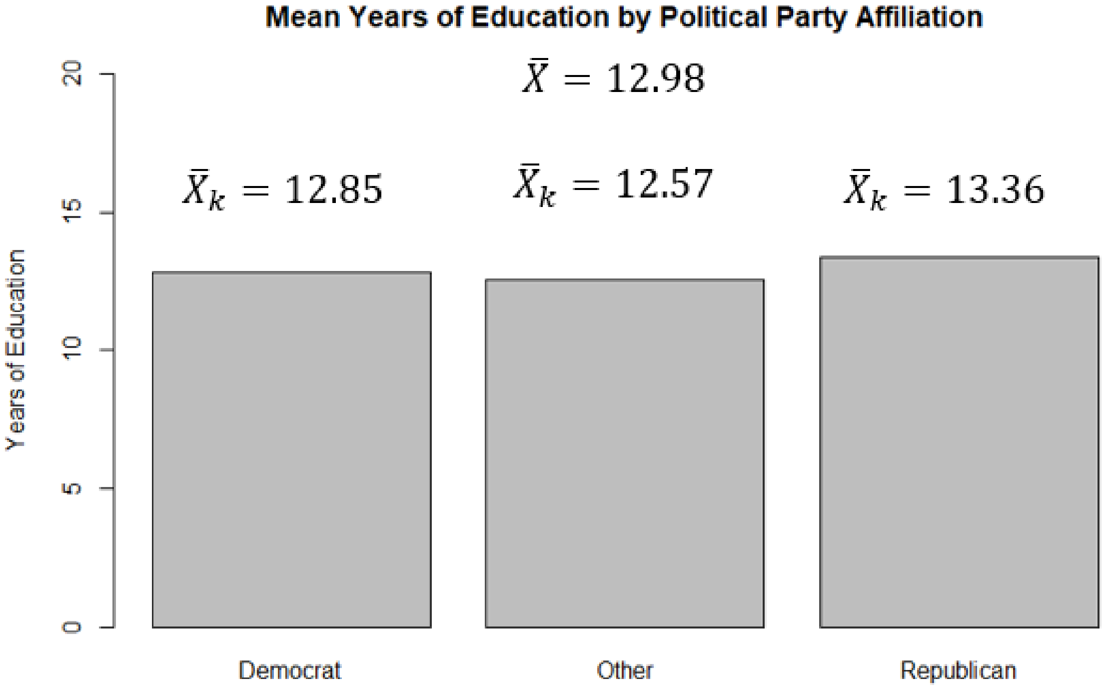
Now let's find the SSB…

ANOVA: example (SSB)

- Does mean years of education differ by political party affiliation?

H_0: $Democrat\ \bar{X}_{edu} = IndepOther\ \bar{X}_{edu} = Republican\ \bar{X}_{edu}$

H_a: At least one group's $\bar{X}_{edu}$ $\neq$ another group's $\bar{X}_{edu}$



$\bar{X} = 12.98$

$\bar{X}_k = 12.85$ $\quad \bar{X}_k = 12.57$ $\quad \bar{X}_k = 13.36$

$\sum X_i^2 = 5{,}571{,}665$ $\quad \sum X_i^2 = 1{,}856{,}215$ $\quad \sum X_i^2 = 4{,}096{,}261$

$$SSB = \sum N_k(\bar{X}_k - \bar{X})^2$$

SSB = 31,615(12.85-12.98)^2 + 11,050(12.5712.98)^2 + 21,890(13.36-12.98)^2

SSB = 31,615(0.016) + 11,050(0.162) + 21,890(0.147)

SSB = 492.07 + 1,792.52 + 3,219.37

SSB = 5,503.966

```
637  SSB_equation<-sum(group_counts*(group_means-mean_grand)^2)
> SSB_equation
[1] 5503.966
```

N_k: number of cases in each group

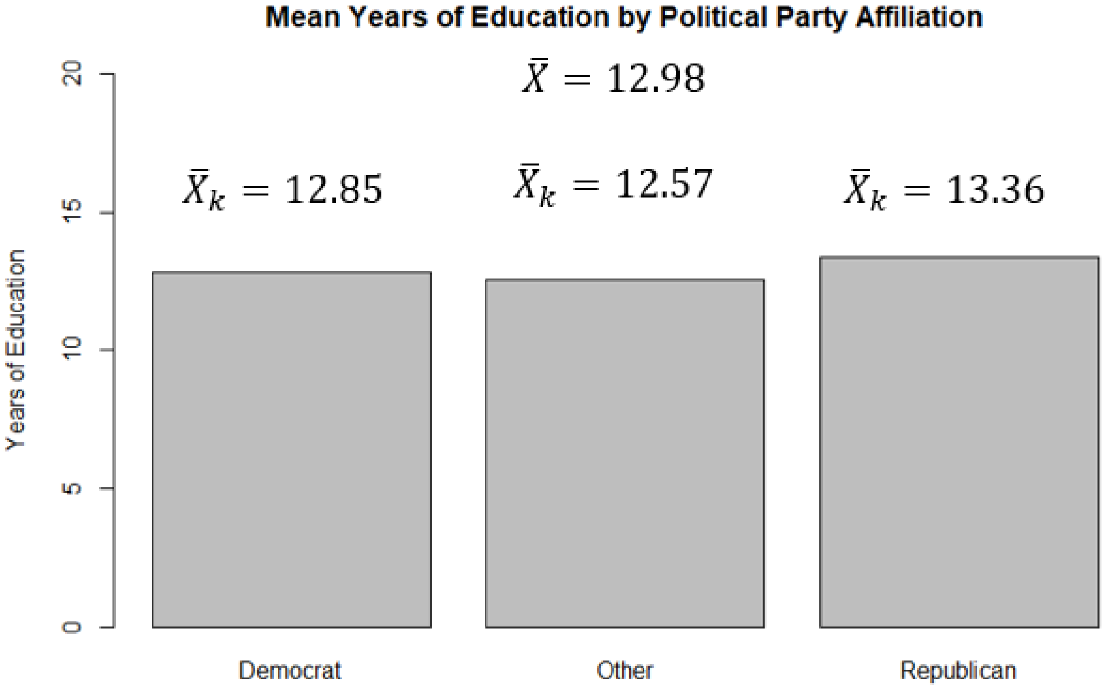
Now let's find the SSW…

ANOVA: example (SSW)

- Does mean years of education differ by political party affiliation?

H_0: $Democrat\ \bar{X}_{edu} = IndepOther\ \bar{X}_{edu} = Republican\ \bar{X}_{edu}$

H_a: At least one group's $\bar{X}_{edu} \neq$ another group's $\bar{X}_{edu}$



$\sum X_i^2 = 5{,}571{,}665$ $\quad \sum X_i^2 = 1{,}856{,}215$ $\quad \sum X_i^2 = 4{,}096{,}261$

$$SSW = \sum_{i=1}^{k} \left(\sum X_{ij}^2 - \frac{(\sum X_{ij})^2}{n_i} \right)$$

For each group…

(3a) Square each value and sum: $\sum X_{ij}^2$

(3b) Square the group total and divide by group size: $\frac{(\sum X_{ij})^2}{n_i}$

(3c) Subtract the second from the first, and repeat for all groups

(3d) Sum them all

X_{ij}: the j-th observation in group i

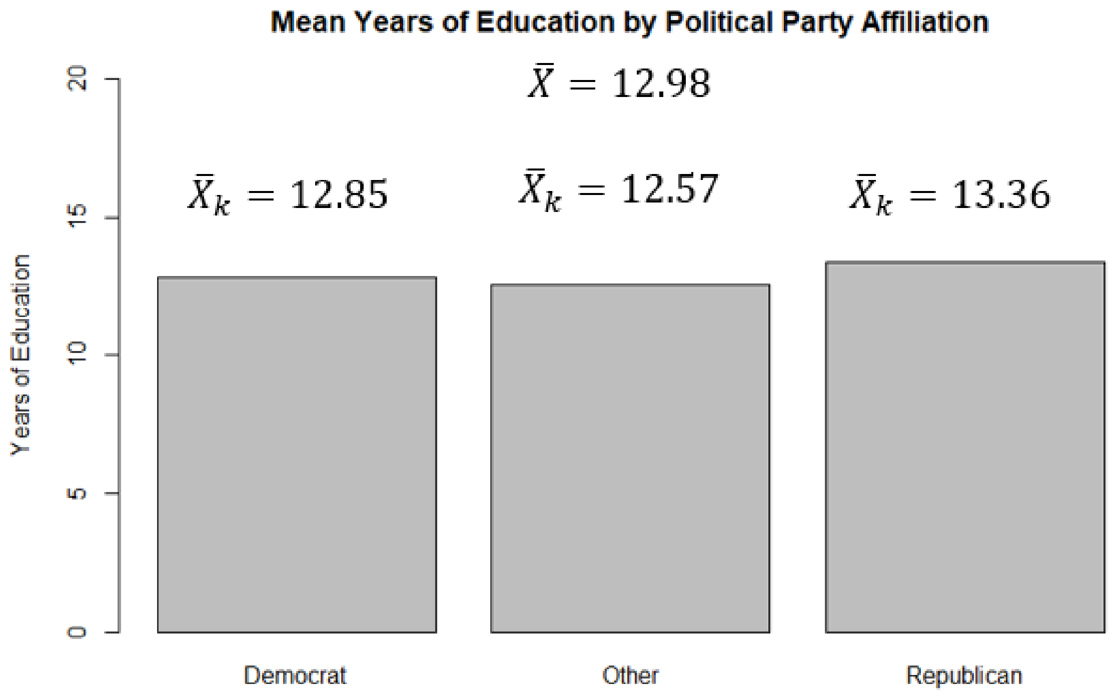
Let's work this out…

ANOVA: example (SSW) continued

- Does mean years of education differ by political party affiliation?

H_0: $Democrat\ \bar{X}_{edu} = IndepOther\ \bar{X}_{edu} = Republican\ \bar{X}_{edu}$

H_a: At least one group's $\bar{X}_{edu} \neq$ another group's $\bar{X}_{edu}$



$\sum X_i^2 = 5{,}571{,}665$ $\sum X_i^2 = 1{,}856{,}215$ $\sum X_i^2 = 4{,}096{,}261$

$$SSW = \sum_{i=1}^{k}\left(\sum X_{ij}^2 - \frac{(\sum X_{ij})^2}{n_i}\right)$$

SSW = (5,571,665 – 5,220,971.08) + (1,856,215 – 1,746,721.00) + (4,096,261 – 3,906,565.56)

SSW = 350,693.92 + 109,494.00 + 189,695.44

SSW = 649,883.36

```
673  SSW_equation<-sum((educ_yrs-group_means[polit_party])^2)
```

```
> SSW_equation
[1] 649883.4
```

X_{ij}: the j-th observation in group i

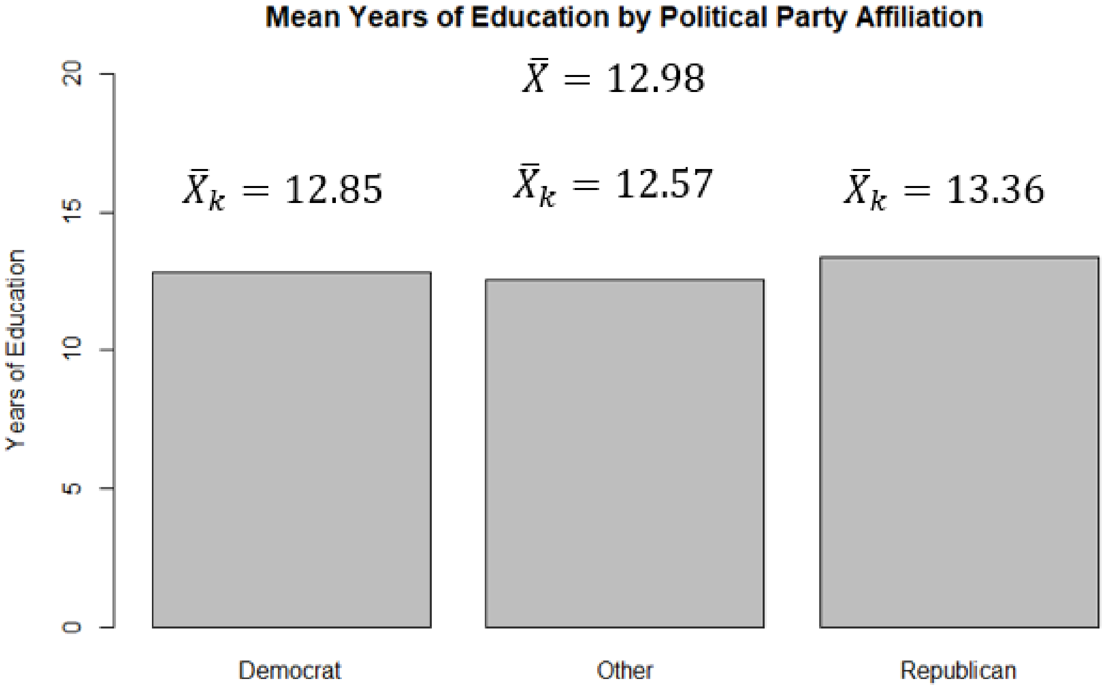
Now let's find the degrees of freedom (df)…

ANOVA: example (df)

- Does mean years of education differ by political party affiliation?

H_0: $Democrat\ \bar{X}_{edu} = IndepOther\ \bar{X}_{edu} = Republican\ \bar{X}_{edu}$

H_a: At least one group's $\bar{X}_{edu}$ $\neq$ another group's $\bar{X}_{edu}$



$\sum X_i^2 = 5{,}571{,}665$ $\sum X_i^2 = 1{,}856{,}215$ $\sum X_i^2 = 4{,}096{,}261$

$$dfw = N - k$$

dfw = 64,555 - 3

dfw = 64,552

$$dfb = k - 1$$

dfw = 3 - 1

dfw = 2

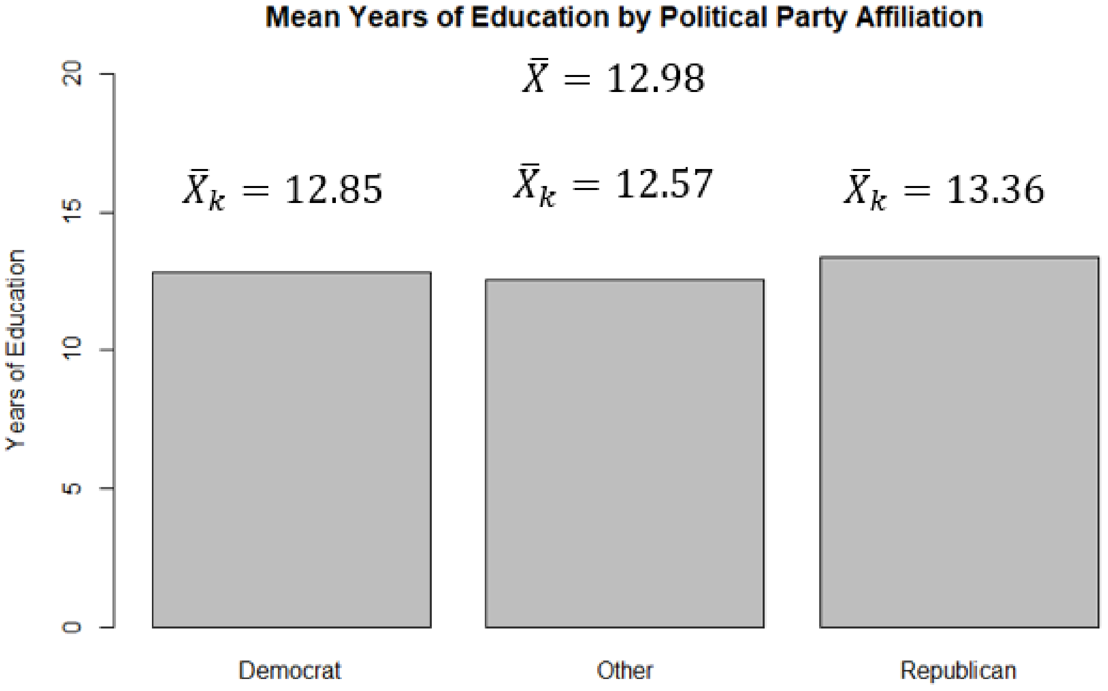
Now let's find the mean square estimates and then the F ratio…

ANOVA: example means squares and F

- Does mean years of education differ by political party affiliation?

$H_0: Democrat\ \bar{X}_{edu} = IndepOther\ \bar{X}_{edu} = Republican\ \bar{X}_{edu}$

H_a: At least one group's $\bar{X}_{edu}$ $\neq$ another group's $\bar{X}_{edu}$



$\bar{X} = 12.98$

$\bar{X}_k = 12.85$ $\bar{X}_k = 12.57$ $\bar{X}_k = 13.36$

$\sum X_i^2 = 5{,}571{,}665$ $\sum X_i^2 = 1{,}856{,}215$ $\sum X_i^2 = 4{,}096{,}261$

$$Mean\ Square\ Within\ (MSW) = \frac{SSW}{dfw}$$

MSW = $\frac{649{,}883.4}{64{,}552}$ = 10.07

$$Mean\ Square\ Between\ (MSB) = \frac{SSB}{dfb}$$

MSW = $\frac{5503.966}{2}$ = 2,751.983

$$F = \frac{MSB}{MSW}$$

F = $\frac{2{,}751.983}{10.07}$ = 273.3506

Let's interpret the test results…

ANOVA: example interpretation

- Does mean years of education differ by political party affiliation?

H_0: $Democrat\ \bar{X}_{edu} = IndepOther\ \bar{X}_{edu} = Republican\ \bar{X}_{edu}$

H_a: At least one group's $\bar{X}_{edu} \neq$ another group's $\bar{X}_{edu}$
$\boldsymbol{F = 273.35}$

- Use the F probability distribution table to identify critical values
  - like the Z, t, and x^2 tables from previous examples
- Find the critical value that corresponds with the degrees of freedom (df), and
  - your selected alpha (α), significance level
- If $F > critical\ value$ then reject H_0

/	df_1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
df_2=1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	243.9060	245.9499	248.0131	249.0518	250.0951	251.1432	252.1957	253.2529	254.3144
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4125	19.4291	19.4458	19.4541	19.4624	19.4707	19.4791	19.4874	19.4957

- 273.35 > 19.4957 at the 0.05 level, so at least one group mean is not equal to another

Let's use the ANOVA function in R, which provides us with a p-value…

5503.966

ANOVA: results

- Does mean years of education differ by political party affiliation?

H_0: $Democrat\ \bar{X}_{edu} = IndepOther\ \bar{X}_{edu} = Republican\ \bar{X}_{edu}$

H_a: At least one group's $\bar{X}_{edu}$ $\neq$ another group's $\bar{X}_{edu}$

```
709  aov(educ_yrs ~ polit_party)
                polit_party Residuals
Sum of Squares       2925.6  652461.7
Deg. of Freedom           1     64553

Residual standard error: 3.179211
Estimated effects may be unbalanced
```

ANOVA results from R look like this

```
              Df  Sum Sq Mean Sq F value Pr(>F)
polit_party    1    2926  2925.6   289.5 <2e-16 ***
Residuals  64553  652462    10.1
```

- ANOVA works best when the number of observations across groups is equal
- Group size in our example is unbalanced, so the F ratio is likely biased

```
polit_party
    1     2     3
31615 11050 21890
```

- Not exactly like the results we computed step by step because of different methods used by R under the hood to address unbalanced design, but close

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

Recall how…

- ANOVA decomposes total variance of a continuous variable into
  - between-group variance: How much do group means differ?
  - within-group variance: How much do cases that compose a group differ?
- to determine if differences between group means are large after accounting for differences within groups that may lead to highly variable group means from sample to sample, thus

SST = SSW + SSB

SSB = SST – SSW

SSW = SST – SSB

- R-squared (R^2): proportion of total variance in the DV that is explained by group differences
  - i.e., the model

$$R^2 = \frac{SSB}{SST} = 1 - \frac{SSW}{SST}$$

We'll talk more about R^2 later. Let's first consider limitations of ANOVA…

ANOVA: limitations

- Requires relatively equal number of observations across categories of the IV
  - balanced designed

- The alternative hypothesis (H_a) does not specify differences between groups,
  - but post hoc techniques are needed to do this

Regression is often a reasonable alternative to ANOVA, but let's first…

ANOVA: practice with Netflix survey data

- Select an interval-ratio variable and a categorial variable with three or more categories

```
aov(ContVarName ~ CatVarName)
```

- If R does not return results because unbalanced, then try this:

```
model<-aov(ContVarName ~ CatVarName)
summary(model)
```

- Share results and practice interpretation

So far, we’ve considered…

- whether the means of an interval/ratio measure vary across two or more groups measured at the nominal level
  - z-tests, t-tests, ANOVA
- whether two nominal, or possibly ordinal, variables are associated
  - Chi-square test of independence
- How do we determine whether two interval/ratio variables are associated?

Covariance

- Recall the variance of an interval/ratio measure: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$

- When considering two interval/ratio measures simultaneously:
  - covariance: $cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

- If no relationship between x and y, then covariance = 0
  - However, there is no inherent scale to assess values < or > 0

Thus, the covariance can be standardized via Pearson correlation coefficient…

Pearson correlation (r)

$$r = \frac{cov(x, y)}{sd(x) \times sd(y)}$$

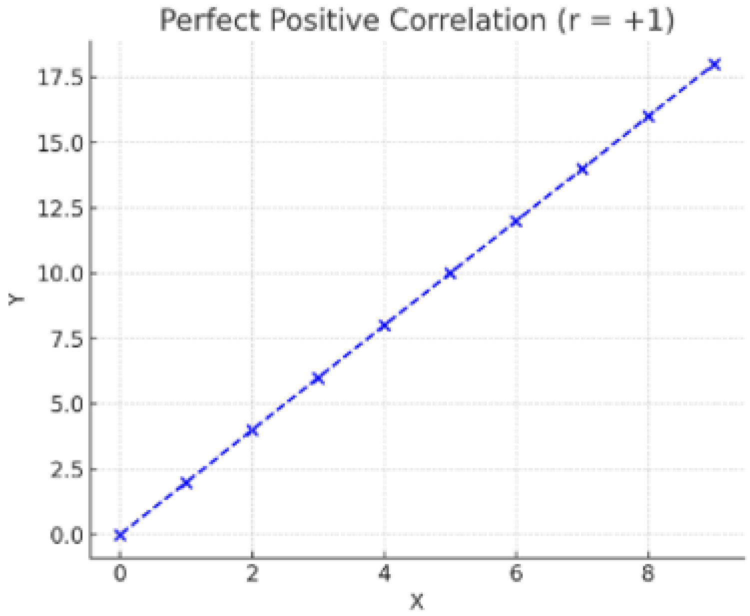
- Bounded between -1 and 1
  - 0 means no linear relationship

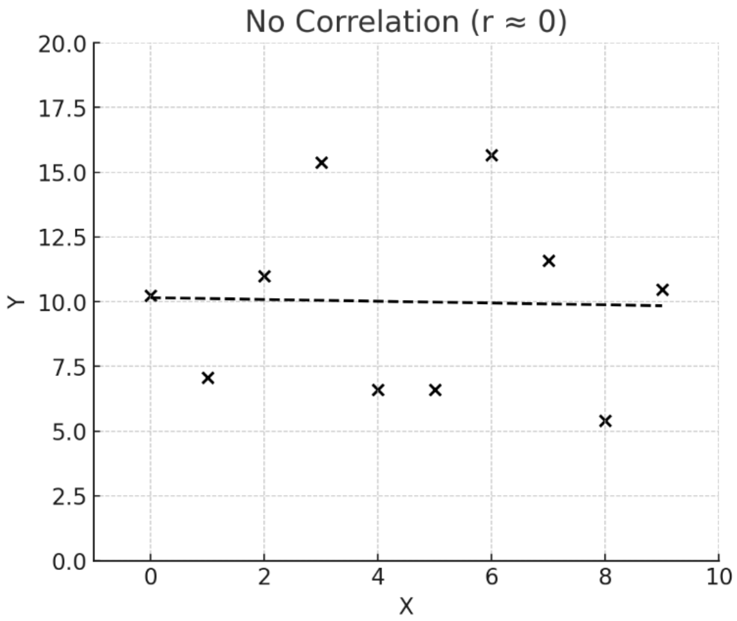
- r < 0.3 = weak; 0.3 < r < 0.6 = moderate; r > .6 = strong
  - social science rule of thumb

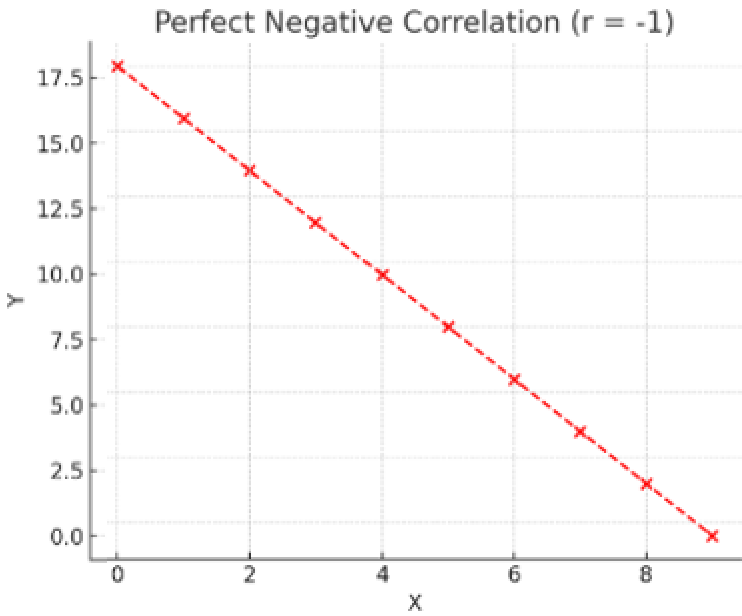
Let's first check out some scattergrams that simultaneously depict continuous-continuous associations…

Scattergram: continuous-continuous plot

- each hashmark reflects a case
  - independent observation
- lines were added for effect
  - more on this later



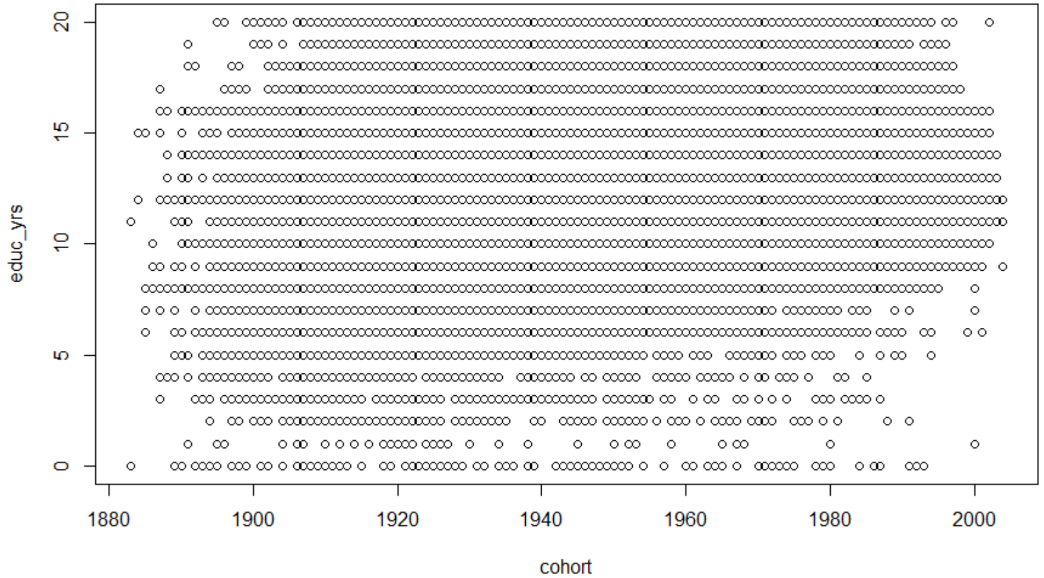




Let’s revisit our education-cohort example...

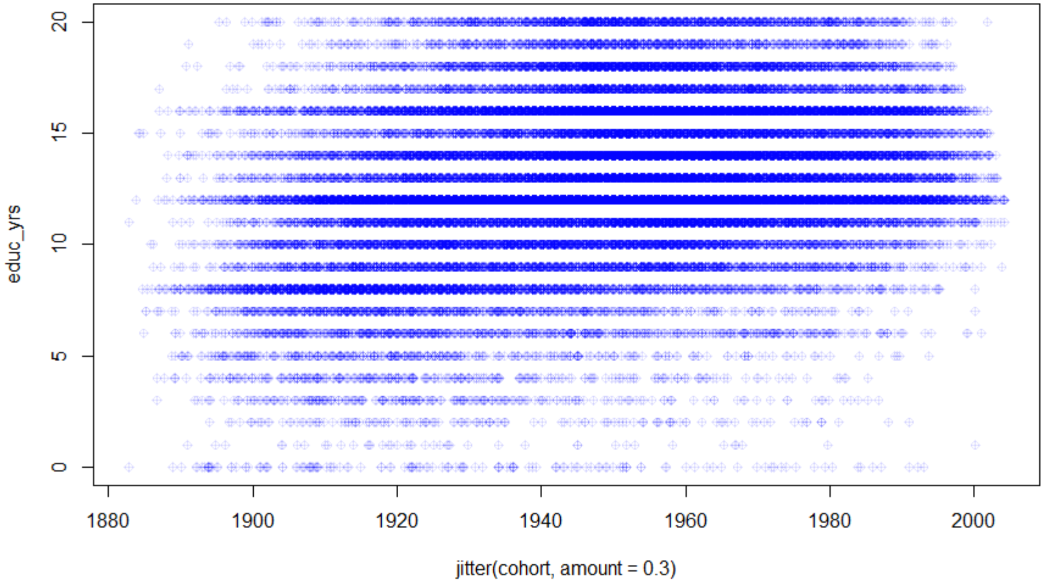
Scattergram: example

too busy to see if a linear pattern is present




```
743  plot(cohort,educ_yrs,ylim=c(0,20))
```

some enhanced visualization techniques




```
748  plot(jitter(cohort,amount=0.3),educ_yrs,pch = 10, col = rgb(0, 0, 1, 0.1))
```

- consider installing ggplot2 data visualization package
  - for those who may move forward with using R down the road

Let's estimate a Pearson's r correlation coefficient to test for a linear relationship…

Pearson correlation (r): example

$$r = \frac{cov(x,y)}{sd(x) \times sd(y)}$$

$$cov(cohort, education) = \frac{\sum_{i=1}^{n}(cohort_i - 1950)(education_i - 12.98)}{64555 - 1} = 20.75$$

$$r = \frac{20.75}{21.93 \times 3.19} = \frac{20.75}{69.96} = 0.297$$

```
790  cor(cohort,educ_yrs,method="pearson")
```

```
> cor(cohort,educ_yrs,method="pearson")
[1] 0.2970501
```

Given the rule of thumb in the social sciences, this appears to be a weakly moderate positive association

Let’s see if this estimate of a linear association is statistically significant…

Pearson correlation (r): hypothesis test

H_0: *no linear relationship between X and Y*
$\rho = 0$
H_a: *association between X and Y*
$\rho \neq 0$

Test requirements, assumptions

- continuous-continuous
- linearity
- homoscedasticity (equal variance)
  - spread of Y values relatively same across all values of X, and vice versa
- normality

small sample size

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$df = N - 2$

large sample size

Fisher's Z-transformation

$$Z = \frac{Z_f - H_0}{\hat{\sigma}_{Z_f}}$$

$Z_f = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$: approximates normal probability distribution

$\hat{\sigma}_{Z_f} = \frac{1}{\sqrt{N-3}}$: estimated standard error of transformed coefficient

Pearson r hypothesis test: example

$H_0: no\ linear\ relationship\ between\ cohort\ and\ education$
$H_a: association\ betweencohort\ and\ education$

$$Z = \frac{Z_f - H_0}{\hat{\sigma}_{Z_f}}$$

$$Z_f = \frac{1}{2}\ln\left(\frac{1+0.297}{1-0.297}\right) = 0.30628$$

$$\hat{\sigma}_{Z_f} = \frac{1}{\sqrt{64555-3}} = 0.00393$$

$$Z = \frac{0.30628 - 0}{0.00393} = 77.82$$

Find the corresponding p-value in the Z-table or just use the cor.test command in R

```
829  cor.test(cohort,educ_yrs,method="pearson")
```

```
t = 79.04, df = 64553, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2900005 0.3040674
sample estimates:
      cor
0.2970501
```

Pearson correlation (r) continued

Let's revisit R-squared (R^2), sometimes called the coefficient of determination

- Conceptually similar to R^2 with ANOVA: $R^2 = \frac{SSB}{SST}$, where the variance in Y was
  - decomposed to obtain the proportion explained by X

- r^2: proportion of variance in Y explained by X
  - tells us about shared variance, not causation

$r^2 = 0.297^2 = 0.0882$, in other words

8.82% of the variation in education is explained by cohort

Pearson correlation (r): limitations

- Useful first step in examining relationships between variables, but limited in that:
- assumes both variables are continuous
- only measures linear association
- doesn't specify direction of causality

Correlation matrix

Correlation matrix: summary of bivariate correlations across three or more variables

Let’s practice some interpretation…

```
cor(GSS) # pearson's r is default
```

	year	id_	age	cohort	female	white	marital	married	hhsize	educ_yrs	educ_deg	polit_party	happy
year	1.000	0.358	0.11	0.60	-0.001	-0.109	0.162	-0.15	-0.165	0.26	0.26	0.036	-0.072
id_	0.358	1.000	0.06	0.20	-0.002	-0.007	0.024	-0.03	-0.059	0.04	0.03	0.044	-0.022
age	0.107	0.057	1.00	-0.73	0.029	0.098	-0.266	0.04	-0.354	-0.15	-0.10	0.013	0.020
cohort	0.601	0.200	-0.73	1.00	-0.024	-0.154	0.325	-0.13	0.171	0.30	0.25	0.014	-0.066
female	-0.001	-0.002	0.03	-0.02	1.000	-0.035	0.002	-0.08	0.006	-0.04	-0.05	-0.058	0.005
white	-0.109	-0.007	0.10	-0.15	-0.035	1.000	-0.170	0.15	-0.056	0.09	0.07	0.234	0.100
marital	0.162	0.024	-0.27	0.33	0.002	-0.170	1.000	-0.91	-0.254	0.04	0.03	-0.084	-0.215
married	-0.147	-0.031	0.04	-0.13	-0.077	0.149	-0.906	1.00	0.352	0.03	0.03	0.087	0.241
hhsize	-0.165	-0.059	-0.35	0.17	0.006	-0.056	-0.254	0.35	1.000	-0.05	-0.07	0.007	0.059
educ_yrs	0.255	0.038	-0.15	0.30	-0.037	0.089	0.040	0.03	-0.050	1.00	0.93	0.067	0.078
educ_deg	0.255	0.033	-0.10	0.25	-0.046	0.073	0.028	0.03	-0.072	0.93	1.00	0.062	0.081
polit_party	0.036	0.044	0.01	0.01	-0.058	0.234	-0.084	0.09	0.007	0.07	0.06	1.000	0.083
happy	-0.072	-0.022	0.02	-0.07	0.005	0.100	-0.215	0.24	0.059	0.08	0.08	0.083	1.000

Pearson’s r is just one of many types of correlation, which differ based on level of measurement

Let’s briefly entertain some other correlation types…

Bivariate correlation : practice with Netflix survey data

- Select an interval-ratio variable and a categorial variable with three or more categories

cor(netflix_survey)

- Let's select some variables, what are you interested in?

cor(cbind(VarName1,Varname2…,$VarName_i$))

Correlation types by level of measurement

Variable 1	Variable 2	Recommended Correlation	Notes
Interval/Ratio	Interval/Ratio	**Pearson's r**	Assumes linear relationship and normal distribution.
Ordinal	Ordinal	**Spearman's ρ** (rho)	Non-parametric; uses rank-order.
Ordinal	Interval/Ratio	**Spearman's ρ**	Non-parametric; when one variable is ordinal, this is still suitable.
Binary	Interval/Ratio	**Point-biserial r**	Special case of Pearson's r; binary must be coded as 0 and 1.
Binary	Binary	**Phi coefficient**	Equivalent to Pearson's r for two binary variables.
Nominal	Nominal	**Cramér's V / Chi-square**	Measures association strength; use Cramér's V for symmetric tables.
Ordinal	Nominal	Not appropriate	Consider recoding or using a nonparametric association test (e.g., Kendall's τ).
Interval/Ratio	Nominal	Not appropriate directly	Consider ANOVA or use dummy coding and regression.