Section F
Graphs and Hypothesis Testing
Lectures 11 and 12

Michael F. Seese

Department of Political Science
University of California San Diego

Political Science 30, Week 7

Outline

Graphs

Hypothesis Testing

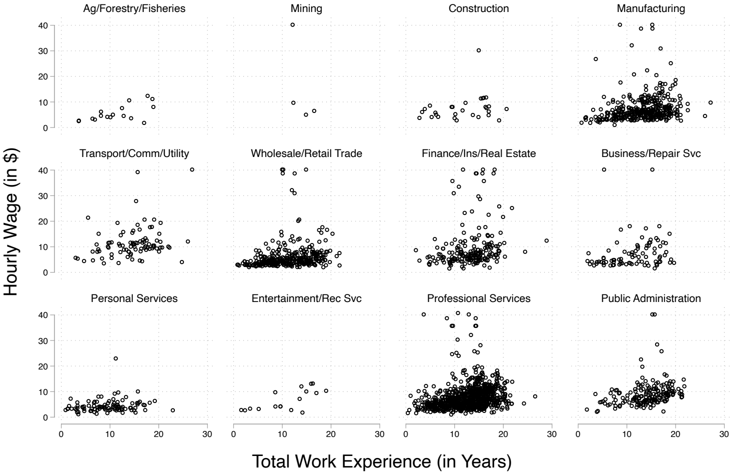
Replication code is available on GitHub

https://github.com/mfseese/Poli30_Spring2021

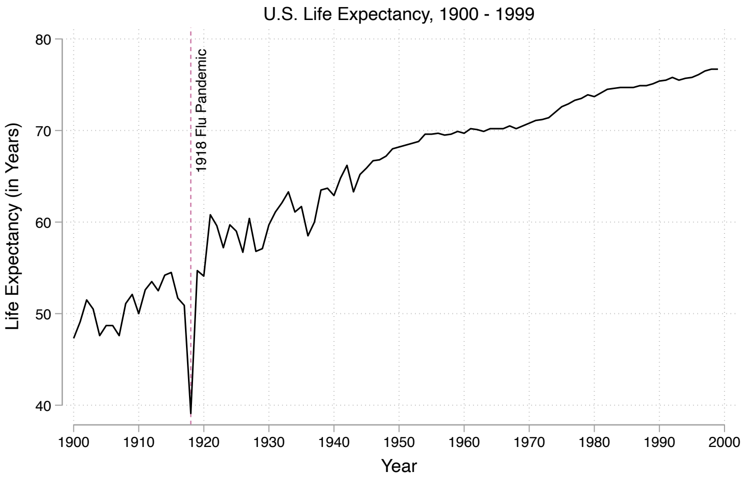
Graphs: Scatter Plots

Graphs: Scatter Plots

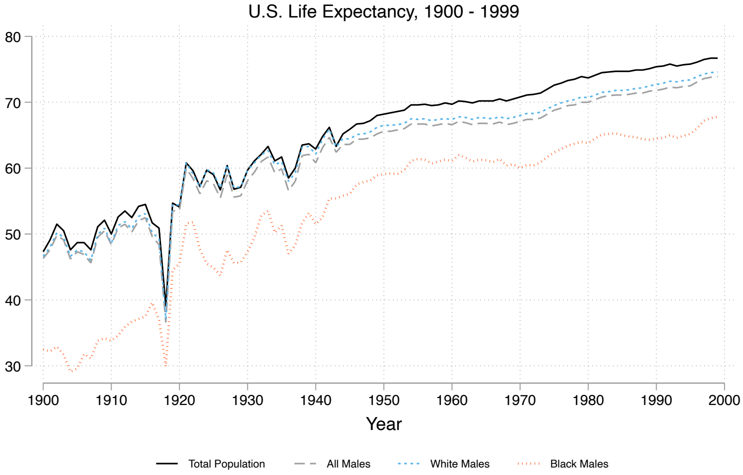
Graphs: Scatter Plots



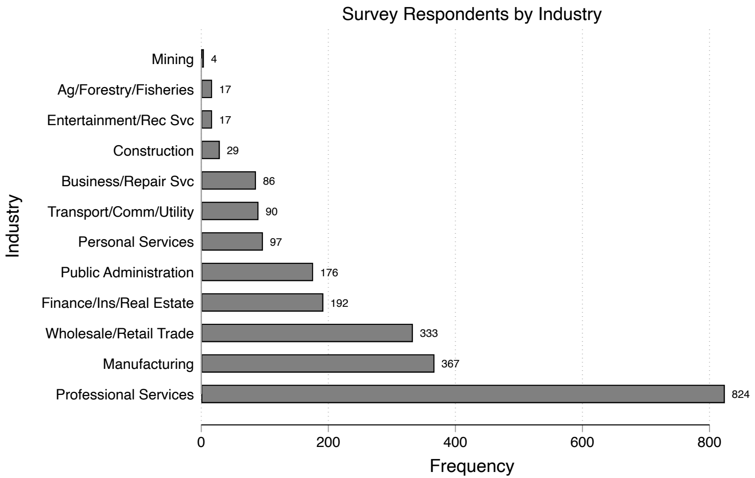
Graphs: Line Graphs



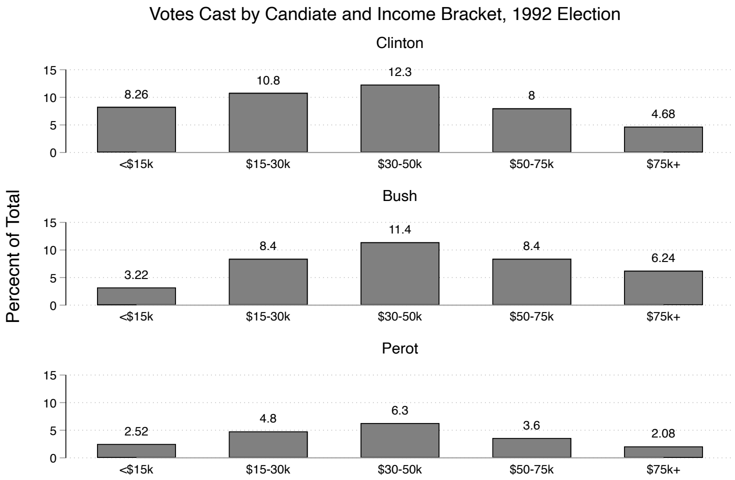
Graphs: Line Graphs



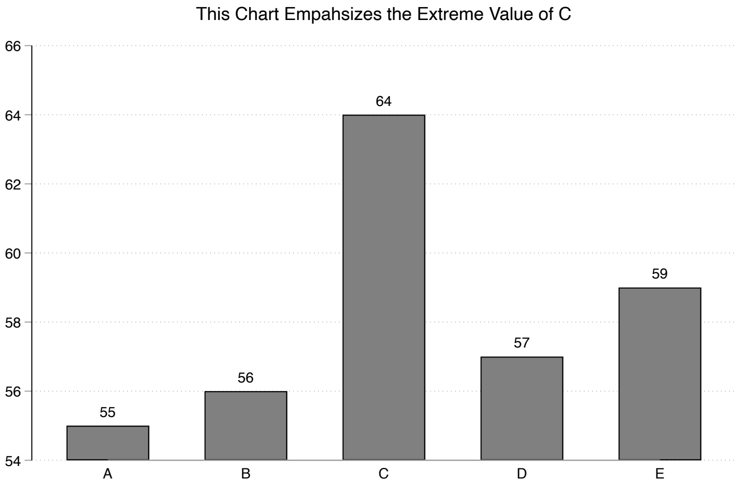
Graphs: Bar Charts



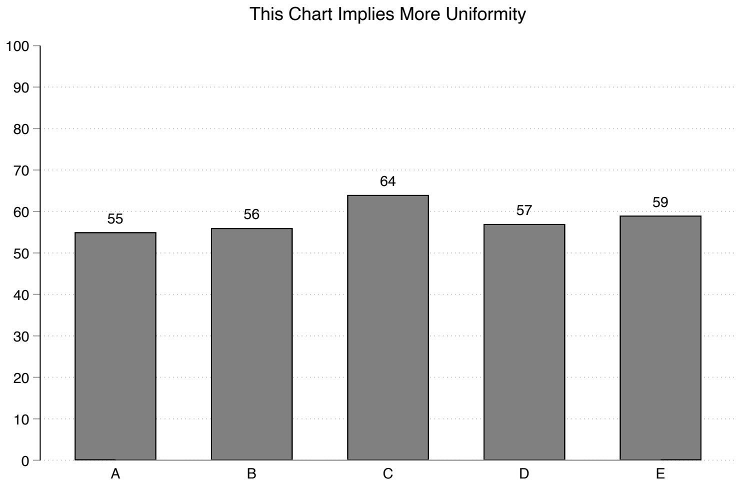
Graphs: Bar Charts



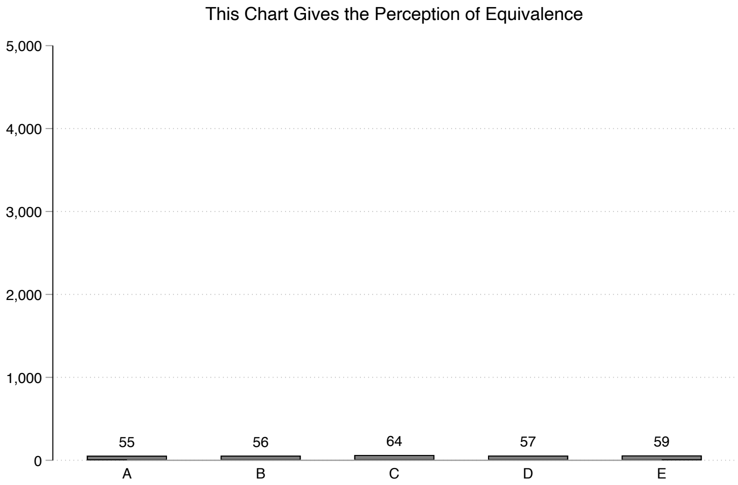
Graphs: Bar Charts



Graphs: Bar Charts



Graphs: Bar Charts



Hypothesis Testing

- Statistical method that uses [sample] data to evaluate a hypothesis [about a population]
- Used to determine whether there is a *significant* difference, effect, or relationship

Steps in a Hypothesis Test

1. State your hypotheses
   - Null, H_0, which hypothesizes no difference, effect, or relationship
     Generally takes the form H_0: $\bar{x}_2 - \bar{x}_1 = 0$
   - Alternative, H_1, which posits some relationship
     Something like H_0: $\bar{x}_2 - \bar{x}_1 \neq 0$
2. Gather your data and calculate the difference in means / proportions, slopes, etc.
3. Calculate the 95% confidence interval
4. Make a decision
   - Reject the null hypothesis
   - Fail to reject the null

Confidence Intervals

Interval and Ratio Variables

$$(\bar{X}_2 - \bar{X}_1) \pm 2 \cdot \sqrt{(SE_1)^2 + (SE_2)^2} \tag{1}$$

Where

$$SE = \frac{\hat{\sigma}}{\sqrt{N}} \tag{2}$$

Nominal and Ordinal Variables

$$\left(\hat{P}_2 - \hat{P}_1\right) \pm 2 \cdot \sqrt{(SE_1)^2 + (SE_2)^2} \tag{3}$$

Where

$$SE = \frac{\hat{\sigma}}{\sqrt{N}} = \frac{\sqrt{(\hat{P})(1 - \hat{P})}}{\sqrt{N}} \tag{4}$$

Example: Testing a Difference in Means

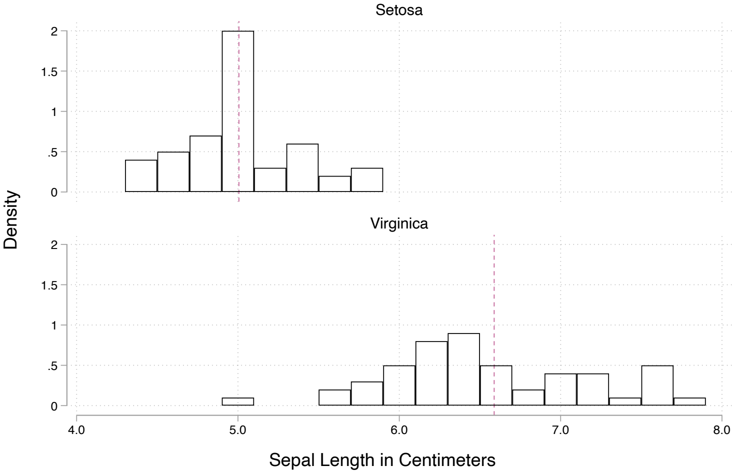
- Use Ronald Fisher's famous Iris Dataset
- Look at some Stata output (replication code posted to GitHub)

- We're going to test whether the sepal length of the Virginica Iris is significantly different from that of the Setosa Iris

  H_0 $\bar{X}_{\text{Virginica}} - \bar{X}_{\text{Setosa}} = 0$

  H_1 $\bar{X}_{\text{Virginica}} - \bar{X}_{\text{Setosa}} > 0$

Example: Testing a Difference in Means



Example: Testing a Difference in Means

Table: Iris Data Summary Statistics

Iris Species	Observations	Mean	Standard Deviation	Min	Max
Setosa	50	5.006	0.3524897	4.3	5.8
Virginica	50	6.588	0.6358796	4.9	7.9

$$(\bar{X}_2 - \bar{X}_1) \pm 2 \cdot \sqrt{(SE_1)^2 + (SE_2)^2} \tag{1}$$

$$(6.588 - 5.006) \pm 2 \cdot \sqrt{\left(\frac{0.352}{\sqrt{50}}\right)^2 + \left(\frac{0.635}{\sqrt{50}}\right)^2}$$

Example: Testing a Difference in Means

Table: Iris Data Summary Statistics

Iris Species	Observations	Mean	Standard Deviation	Min	Max
Setosa	50	5.006	0.3524897	4.3	5.8
Virginica	50	6.588	0.6358796	4.9	7.9

$$(\bar{X}_2 - \bar{X}_1) \pm 2 \cdot \sqrt{(SE_1)^2 + (SE_2)^2} \tag{1}$$

$$(6.588 - 5.006) \pm 2 \cdot \sqrt{\left(\frac{0.352}{\sqrt{50}}\right)^2 + \left(\frac{0.635}{\sqrt{50}}\right)^2}$$

Example: Testing a Difference in Means

Table: Iris Data Summary Statistics

Iris Species	Observations	Mean	Standard Deviation	Min	Max
Setosa	50	5.006	0.3524897	4.3	5.8
Virginica	50	6.588	0.6358796	4.9	7.9

$$(\bar{X}_2 - \bar{X}_1) \pm 2 \cdot \sqrt{(SE_1)^2 + (SE_2)^2} \tag{1}$$

$$(6.588 - 5.006) \pm 2 \cdot \sqrt{\left(\frac{0.352}{\sqrt{50}}\right)^2 + \left(\frac{0.635}{\sqrt{50}}\right)^2}$$

Example: Testing a Difference in Means

$$(6.588 - 5.006) \pm 2 \cdot \sqrt{\left(\frac{0.352}{\sqrt{50}}\right)^2 + \left(\frac{0.635}{\sqrt{50}}\right)^2}$$

$$= 1.582 \pm 2 \cdot \sqrt{(0.0497)^2 + (0.0898)^2}$$

$$= 1.582 \pm 2 \cdot 0.102$$

$$= 1.582 \pm 0.205$$

$$= 1.377 \text{ or } 1.787 \Longleftarrow \text{ Confidence Interval}$$

We can therefore <u>reject the null hypothesis</u>, as the CI does not contain 0

Example: Testing a Difference in Means

```
. ttest seplen, by(igroup2)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Setosa	50	5.006	.0498496	.3524897	4.905824	5.106176
Virginic	50	6.588	.089927	.6358796	6.407285	6.768715
combined	100	5.797	.0945319	.9453186	5.609428	5.984572
diff		-1.582	.1028194		-1.786042	-1.377958

```
    diff = mean(Setosa) - mean(Virginic)                          t = -15.3862
Ho: diff = 0                                     degrees of freedom =       98

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Example: Testing a Difference in Means

```
. ttest seplen, by(igroup2)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Setosa	50	5.006	.0498496	.3524897	4.905824	5.106176
Virginic	50	6.588	.089927	.6358796	6.407285	6.768715
combined	100	5.797	.0945319	.9453186	5.609428	5.984572
diff		-1.582	.1028194		-1.786042	-1.377958

```
    diff = mean(Setosa) - mean(Virginic)                      t = -15.3862
Ho: diff = 0                                 degrees of freedom =       98

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Difference in Slopes

- Slope is just rise over run, or $\frac{\Delta Y}{\Delta X}$
- You probably learned slope as the m in the equation $y = mx + b$
  - In statistics, we usually call it β, as in $y = \alpha + \beta x$
  - You might also see it as b, like: $y = a + bx$

- Slopes help us define a relationship between two variables
- For example: Does the <u>weight</u> of your car affect the <u>gas mileage</u>?

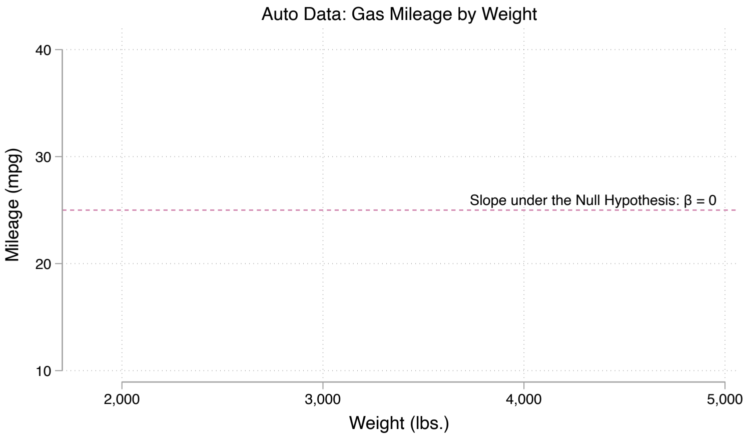
Example: Testing a Difference in Slopes

- Use some data on cars
- Look at some Stata output (replication code posted to GitHub)

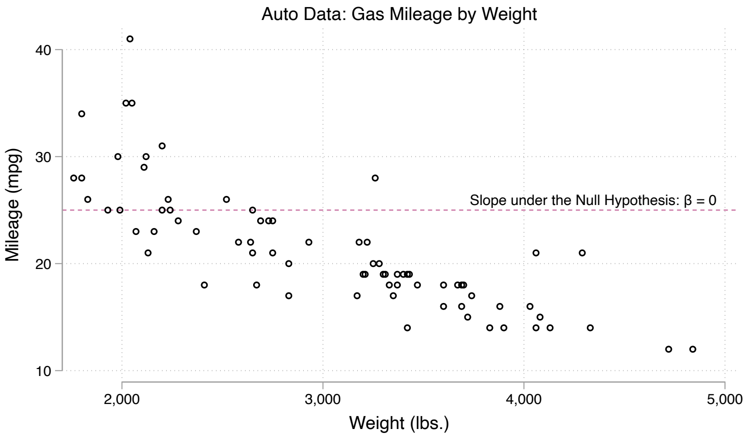
- Let's test whether a car's gas mileage decreases as the weight of the car increases

  H_0 $\beta_1 = 0$
  H_1 $\beta_1 < 0$

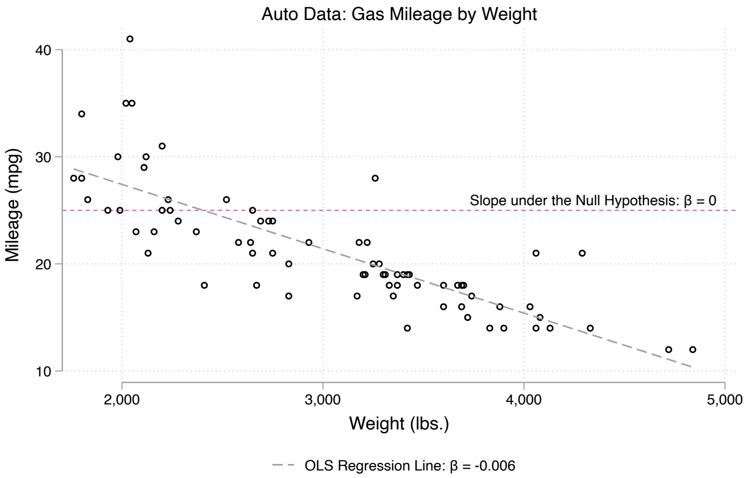
Example: Testing a Difference in Slopes



Example: Testing a Difference in Slopes



Example: Testing a Difference in Slopes



Example: Testing a Difference in Slopes

```
. reg mpg weight
```

Source	SS	df	MS
Model	1591.9902	1	1591.9902
Residual	851.469256	72	11.8259619
Total	2443.45946	73	33.4720474

Number of obs	=	74
F(1, 72)	=	134.62
Prob > F	=	0.0000
R-squared	=	0.6515
Adj R-squared	=	0.6467
Root MSE	=	3.4389

mpg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	-.0060087	.0005179	-11.60	0.000	-.0070411	-.0049763
_cons	39.44028	1.614003	24.44	0.000	36.22283	42.65774

Our regression equation is: $\hat{y} = 39.440 - 0.006x$

Example: Testing a Difference in Slopes

```
. reg mpg weight

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =    134.62
       Model |   1591.9902         1   1591.9902   Prob > F        =    0.0000
    Residual |  851.469256        72  11.8259619   R-squared       =    0.6515
-------------+----------------------------------   Adj R-squared   =    0.6467
       Total |  2443.45946        73  33.4720474   Root MSE        =    3.4389

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005179   -11.60   0.000    -.0070411   -.0049763
       _cons |   39.44028   1.614003    24.44   0.000     36.22283    42.65774
------------------------------------------------------------------------------
```

Our interval is given by: $-0.0060087 \pm (2 \cdot 0.0005179)$