Recitation: Introduction to statistics

Felipe Balcazar

NYU

August, 2021

NEW YORK UNIVERSITY

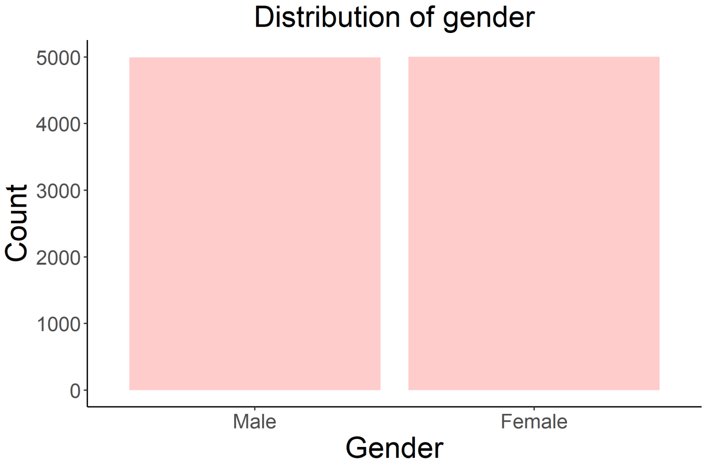
Let's start with the basics: a data set

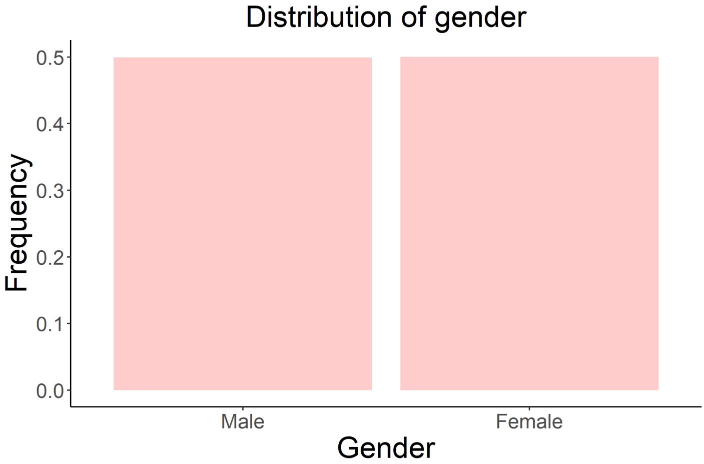
Y	X	D
1875.76	1	12.4950
1891.65	0	12.4751
1866.13	0	12.3136
1855.94	1	12.3765
1831.51	0	12.0826
1776.29	0	11.7044
⋮	⋮	⋮
1891.84	0	12.4841
1835.27	1	12.2289
1894.73	1	12.6287

Let's start with the basics: a distribution

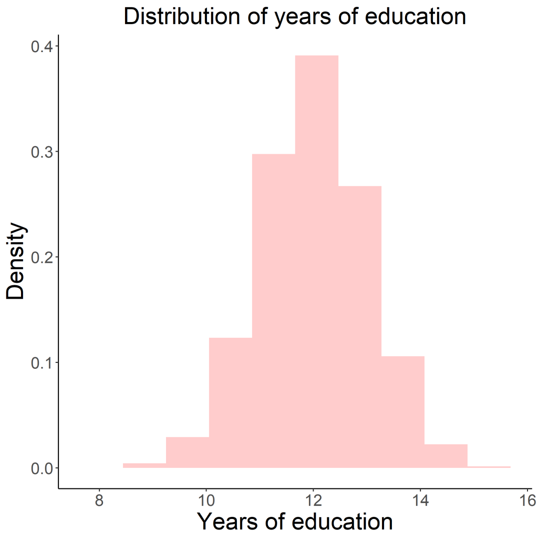
Wage	Gender	Years of Ed.
1875.76	Female	12.4950
1891.65	Male	12.4751
1866.13	Male	12.3136
1855.94	Female	12.3765
1831.51	Male	12.0826
1776.29	Male	11.7044
⋮	⋮	⋮
1891.84	Male	12.4841
1835.27	Female	12.2289
1894.73	Female	12.6287

Let's start with the basics: a distribution

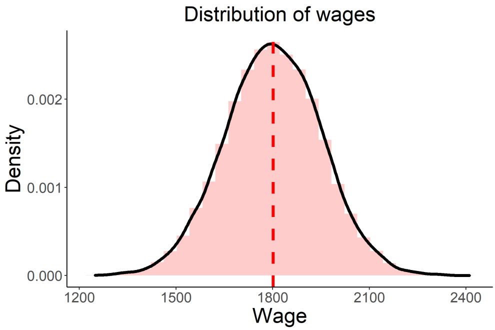



Distribution of gender
Frequency
0.5
0.4
0.3
0.2
0.1
0.0
Male
Female
Gender

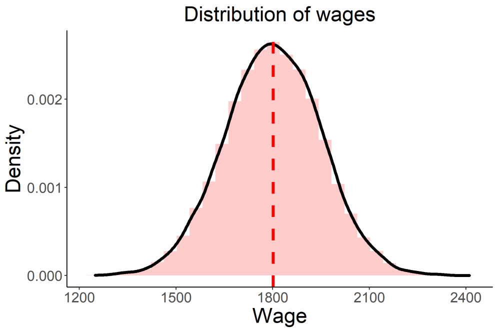
Let's start with the basics: a distribution



Let's start with the basics: statistics



Let's start with the basics: statistics



- How can we begin to analyze Y? We can compute...
  - The mean: $\mu = \frac{\sum y}{N}$.
  - The variance: $\sigma^2 = \frac{\sum (y-\mu)^2}{(N-1)}$.
  - The standard deviation: $\sigma = \frac{\sum (y-\mu)}{\sqrt{(N-1)}}$.
  - The median or 50th percentile.

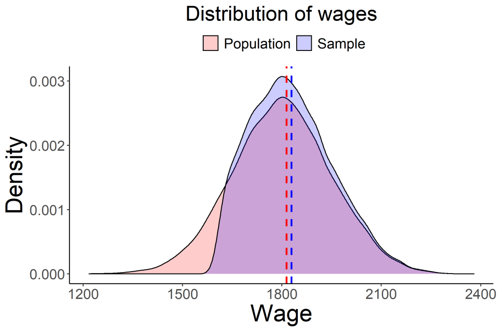
Let's start with the basics: statistics

Statistic	Value
Mean	1800
Variance	22379
Standard deivation	150
Median (or Percentile 50)	1800
Percentile 10	1608
Percentile 90	1994
Number of companies	10000

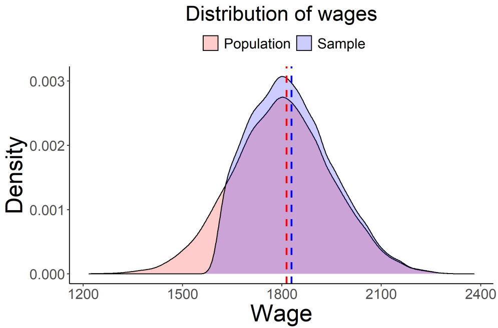
Let's start with the basics: statistics

Statistic	Population
Mean	1800
Variance	22379
Standard deivation	150
Median (or Percentile 50)	1800
Percentile 10	1608
Percentile 90	1994
Number of companies	10000

Often we observe just a sample



Often we observe just a sample



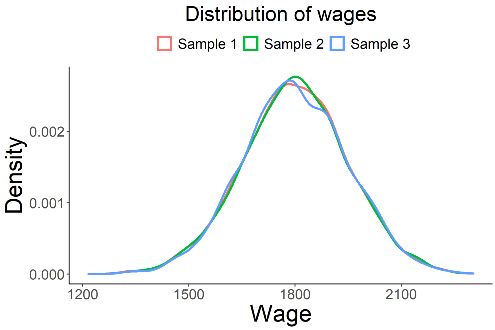
- How can we begin to analyze Y? We can compute...
  - The sample mean (average): $\bar{y} = \frac{\sum y}{N}$.
  - The sample variance: $s^2 = \frac{\sum (y - \bar{y})^2}{N-1}$.
  - The sample standard deviation: $s = \frac{\sum (y - \bar{y})}{\sqrt{N-1}}$.
  - The sample median or 50th percentile.

Often we observe just a sample: descriptive statistics

	Population	Sample
Mean	1800	1829
Variance	22379	15880
Standard deviation (SD)	150	126
Median	1800	1817
Number of obs. (N)	10000	9000

- The sample is non-random, excludes the 10% poorest.
- This happens when it is costly to survey poor individuals.
- The mean in this case is biased!
- This is addressed by collecting random samples.

Different random samples result in different estimates



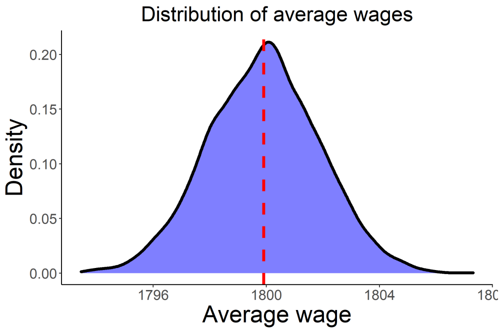
- Different samples result in different averages: 1801, 1799, 1803.

Different random samples result in different estimates

	Population	Sample 1	Sample 2	Sample 3
Mean	1800	1801	1799	1803
Variance	22379	21950	22509	22147
SD	150	148	150	149
Median	1800	1802	1798	1803
N	10000	5004	5117	5019

- They result as well in different estimates for other statistics: variance, median, etc.

We focus in the average



- Using multiple random draws we obtain the distribution of averages.
- The average has its own "standard deviation."
- We approximate this with the standard error:

$$SE \approx \frac{s}{\sqrt{N}}$$

We can make questions about the average

- Is the average income 1800USD?
  - *Null hypothesis* (H_0): $\overline{y} = 1800$.
  - The alternative (H_a): $\overline{y} \neq 1800$.
- Let's be careful not to reject the null hypothesis when it is true. This is called *type I error*.
- We want to be sure that the probability that this occurs is small, less than α. Usually less than 5% (that is $\alpha = 0.05$).

		Null hypothesis	
		TRUE	FALSE
Findings	Reject null	Type I error (α)	Correct decision
	Accept null	Correct decision	Type II error (β)

Note: β is what is called "power." This often related to the sample size because small samples have low power, but we won't worry about that here.

Is the average income 1800USD?

- To test this hypothesis we compute a t-statistic:

$$\hat{t} = \frac{\overline{y} - \overline{y}_{H_0}}{SE}.$$

  $\overline{y}$ is the value of the mean we estimate from our sample.
- Then we ask $Pr(|t| \geq |\hat{t}|) \equiv$ p-value.
- If p-value $< \alpha$, we reject the null hypothesis.
  - Don't worry, the computer does all of this for you.
  - You probably have seen stars in papers.

Is the average income 1800USD?

- To test this hypothesis we compute a t-statistic:

$$\hat{t} = \frac{\overline{y} - 1800}{SE}.$$

  $\overline{y}$ is the value of the mean we estimate from our sample.
- Then we ask $Pr(|t| \geq |\hat{t}|) \equiv$ p-value.
- If p-value $< \alpha$, we reject the null hypothesis.
  - Don't worry, the computer does all of this for you.
  - You probably have seen stars in papers.

Is the average income 1800USD?

- To test this hypothesis we compute a t-statistic:

$$\hat{t} = \frac{\overline{y} - 1800}{SE}.$$

  $\overline{y}$ is the value of the mean we estimate from our sample.
- Then we ask $Pr(|t| \geq |\hat{t}|) \equiv$ p-value.
- If p-value $< \alpha$, we reject the null hypothesis.
  - Don't worry, the computer does all of this for you.
  - You probably have seen stars in papers.

Is the average income 1800USD?

- To test this hypothesis we compute a t-statistic:

$$\hat{t} = \frac{\overline{y} - 1800}{SE}.$$

  $\overline{y}$ is the value of the mean we estimate from our sample.
- Then we ask $Pr(|t| \geq |\hat{t}|) \equiv$ p-value.
- If p-value $< \alpha$, we reject the null hypothesis.
  - Don't worry, the computer does all of this for you.
  - You probably have seen stars in papers.

Example of stars in papers

	Dependent variable: Average years of education attained			
	(1)	(2)	(3)	(4)
(a) Full sample				
Stock of democracy	0.004	0.005	0.006*	0.009**
	(0.003)	(0.003)	(0.004)	(0.004)
Constant	9.959***	9.048***	8.974***	8.893***
	(0.476)	(0.521)	(0.513)	(0.502)
Discount factor (r)	0.01	0.03	0.06	0.10
R^2	0.435	0.437	0.441	0.446
Clusters	210	210	210	210
Observations	3078	3078	3078	3078
(b) Restricted sample				
Stock of democracy	−0.007	−0.007	−0.006	−0.005
	(0.005)	(0.005)	(0.006)	(0.007)
Constant	10.998***	10.908***	10.767***	10.584***
	(0.481)	(0.495)	(0.510)	(0.518)
Discount factor (r)	0.01	0.03	0.06	0.10
R^2	0.421	0.417	0.411	0.405
Clusters	168	168	168	168
Observations	2714	2714	2714	2714

Note: Standard errors are clustered by country and birth-cohort in parentheses. * Significant at 10 %, ** significant at 5 %, *** significant at 1 %. The stock of democracy is calculated from 6 to 18 years after a

Eyeballing the answer to the same question

- Use a confidence interval!
- If $\overline{y}_{H_0}$ is in it you have $(1-\alpha) \times 100$ percent confidence of this being the case.
  - 90% confidence when $\overline{y}_{H_0} \in [\overline{y} - 1.65SE; \overline{y} + 1.65SE]$
  - 95% confidence when $\overline{y}_{H_0} \in [\overline{y} - 1.96SE; \overline{y} + 1.96SE]$
  - 99% confidence when $\overline{y}_{H_0} \in [\overline{y} - 2.56SE; \overline{y} + 2.56SE]$

Eyeballing the answer to the same question

- Use a confidence interval!
- If $\overline{y}_{H_0}$ is in it you have $(1-\alpha)\times 100$ percent confidence of this being the case.
  - 90% confidence when $\overline{y}_{H_0} \in [\overline{y} - 1.65SE; \overline{y} + 1.65SE]$
  - 95% confidence when $\overline{y}_{H_0} \in [\overline{y} - 1.96SE; \overline{y} + 1.96SE]$
  - 99% confidence when $\overline{y}_{H_0} \in [\overline{y} - 2.56SE; \overline{y} + 2.56SE]$

				95% confidence interval	
Average	**SD**	**N**	**SE**	**lower bound**	**upper bound**
1800	150	1000	5	1791	1809

Eyeballing the answer to the same question

- Use a confidence interval!
- If $\overline{y}_{H_0}$ is in it you have $(1-\alpha)\times 100$ percent confidence of this being the case.
  - 90% confidence when $\overline{y}_{H_0} \in [\overline{y} - 1.65SE; \overline{y} + 1.65SE]$
  - 95% confidence when $\overline{y}_{H_0} \in [\overline{y} - 1.96SE; \overline{y} + 1.96SE]$
  - 99% confidence when $\overline{y}_{H_0} \in [\overline{y} - 2.56SE; \overline{y} + 2.56SE]$

				95% confidence interval	
Average	**SD**	**N**	**SE**	**lower bound**	**upper bound**
1800	150	1000	5	1791	1809

- Practice (at 95% confidence):
  - Reject the null that $\overline{y}$ is zero: $0 \notin [1791, 1809]$.
  - Reject the null that $\overline{y}$ is 1750: $1750 \notin [1791, 1809]$.
  - Accept the null that $\overline{y}$ is 1808: $1808 \in [1791, 1809]$.

Eyeballing the answer to the same question

- Use a confidence interval!
- If $\overline{y}_{H_0}$ is in it you have $(1-\alpha) \times 100$ percent confidence of this being the case.
  - 90% confidence when $\overline{y}_{H_0} \in [\overline{y} - 1.65SE; \overline{y} + 1.65SE]$
  - 95% confidence when $\overline{y}_{H_0} \in [\overline{y} - 1.96SE; \overline{y} + 1.96SE]$
  - 99% confidence when $\overline{y}_{H_0} \in [\overline{y} - 2.56SE; \overline{y} + 2.56SE]$

				95% confidence interval	
Average	**SD**	**N**	**SE**	**lower bound**	**upper bound**
1800	150	1000	5	1791	1809

- Practice (at 95% confidence):
  - Reject the null that $\overline{y}$ is zero: $0 \notin [1791, 1809]$.
  - Reject the null that $\overline{y}$ is 1750: $1750 \notin [1791, 1809]$.
  - Accept the null that $\overline{y}$ is 1808: $1808 \in [1791, 1809]$.

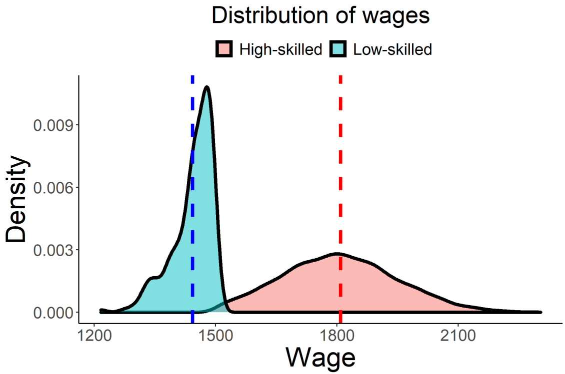
Eyeballing the answer to the same question

- Use a confidence interval!
- If $\overline{y}_{H_0}$ is in it you have $(1-\alpha) \times 100$ percent confidence of this being the case.
  - 90% confidence when $\overline{y}_{H_0} \in [\overline{y} - 1.65SE; \overline{y} + 1.65SE]$
  - 95% confidence when $\overline{y}_{H_0} \in [\overline{y} - 1.96SE; \overline{y} + 1.96SE]$
  - 99% confidence when $\overline{y}_{H_0} \in [\overline{y} - 2.56SE; \overline{y} + 2.56SE]$

				95% confidence interval	
Average	**SD**	**N**	**SE**	**lower bound**	**upper bound**
1800	150	1000	5	1791	1809

- Practice (at 95% confidence):
  - Reject the null that $\overline{y}$ is zero: $0 \notin [1791, 1809]$.
  - Reject the null that $\overline{y}$ is 1750: $1750 \notin [1791, 1809]$.
  - Accept the null that $\overline{y}$ is 1808: $1808 \in [1791, 1809]$.

Differences in means



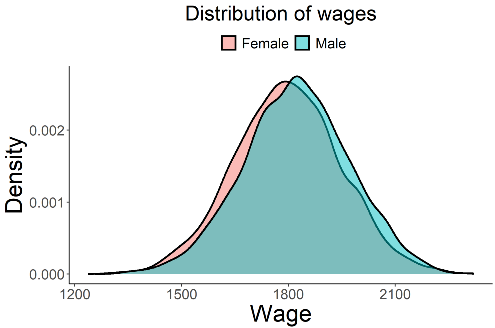
- Let's imagine there are two groups: A and B.
- Does A have a higher wage on average than B?
  1. We see that A earns on average more than B.
  2. False! unless we reject the null hypothesis that $\overline{y}_A - \overline{y}_B = 0$.

- We can extrapolate the same principle of the average, to differences in averages.
- The difference in averages then also has a standard error

$$SE_{\overline{y}_A - \overline{y}_B} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

	Skill level		
	High	**Low**	**Difference**
Average	1809	1443	366
SD	140	50.2	
SE	1.4	3.2	3.5
N	9756	244	10000

Differences in means: another example



	Gender		
	Male	**Female**	**Difference**
Average	1818	1802	16
	(5.72)	(6.05)	(8.33)
N	1234	1241	2475

Note: Standard errors in parentheses.

Differences in means: yet another example

Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
			A. Health			
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
			B. Characteristics			
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.