

MATH 1051H - Lecture #13

Case Study: Gender Discrimination

Gender Discrimination

- In 1972, as a part of a study on gender discrimination, 48 male bank supervisors were each given the same personnel file and asked to judge whether the person should be promoted to a branch manager job that was described as “routine”.
- The files were identical except that half of the supervisors had files showing the person was male while the other half had files showing the person was female.
- It was randomly determined which supervisors got “male” applications and which got “female” applications.
- Of the 48 files reviewed, 35 were promoted.
- The study is testing whether females are unfairly discriminated against.
- Is this an observational study or an experiment?

B.Rosen and T. Jerdee (1974), “Influence of sex role stereotypes on personnel decisions”, J.Applied Psychology, 59:9-14.

Data

At a first glance, does there appear to be a relationship between promotion and gender?

	Promoted		
	Yes	No	Total
Gender			
Male	21	3	24
Female	14	10	24
Total	35	13	48

% of males promoted: $21 / 24 = 0.875$

% of females promoted: $14 / 24 = 0.583$

Practice

We saw a difference of almost 30% (29.2% to be exact) between the proportion of male and female files that are promoted. Based on this information, which of the below is true?

1. If we were to repeat the experiment we will definitely see that more female files get promoted. This was a fluke.
2. Promotion is dependent on gender, males are more likely to be promoted, and hence there is gender discrimination against women in promotion decisions.
3. The difference in the proportions of promoted male and female files is due to chance, this is not evidence of gender discrimination against women in promotion decisions.
4. Women are less qualified than men, and this is why fewer females get promoted.

Practice

We saw a difference of almost 30% (29.2% to be exact) between the proportion of male and female files that are promoted. Based on this information, which of the below is true?

1. If we were to repeat the experiment we will definitely see that more female files get promoted. This was a fluke.
2. *Promotion is dependent on gender, males are more likely to be promoted, and hence there is gender discrimination against women in promotion decisions.* Maybe!
3. *The difference in the proportions of promoted male and female files is due to chance, this is not evidence of gender discrimination against women in promotion decisions.* Maybe!
4. Women are less qualified than men, and this is why fewer females get promoted.

Two Competing Claims

“There is nothing going on.” (**Null Hypothesis**)

Promotion and gender are **independent**.

No gender discrimination.

Observed difference in proportions is simply due to chance.

versus

There is something going on.” (**Alternative Hypothesis**)

Promotion and gender are **dependent**.

There is gender discrimination.

Observed difference in proportions is not due to chance.

We Will Return!

We will continue with the concept of hypothesis testing later (Lecture #14), and again over the following weeks. First, let's develop some other ideas about sampling distributions.

Variability in estimates

Young, Underemployed and Optimistic

Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy. An analysis of government economic data suggests that this perception is correct. The recent indicators on the nation's labor market show a decline in the

Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience. And more than one-third (35%) say that, as a result of the poor economy, they have gone back to school. Their personal lives have also been affected: 31% have postponed either getting married or having a baby (22% say they have postponed having a baby and 20% have put off getting married). One-in-four (24%) say they have moved back in with their parents after living on their own.

<http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession>

Margin of error

The general public survey is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States, including an oversample of 346 adults ages 18 to 34. A total of 769 interviews were completed with respondents contacted by landline telephone and 1,279 with those contacted on their cellular phone. Data are weighted to produce a final sample that is representative of the general population of adults in the continental United States. Survey interviews were conducted under the direction of Princeton Survey Research Associates International, in English and Spanish. Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

- 41% \pm 2.9%: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.
- 49% \pm 4.4%: We are 95% confident that 44.6% to 53.4% of 18-34 years olds have taken a job they didn't want just to pay the bills.

Parameter estimation

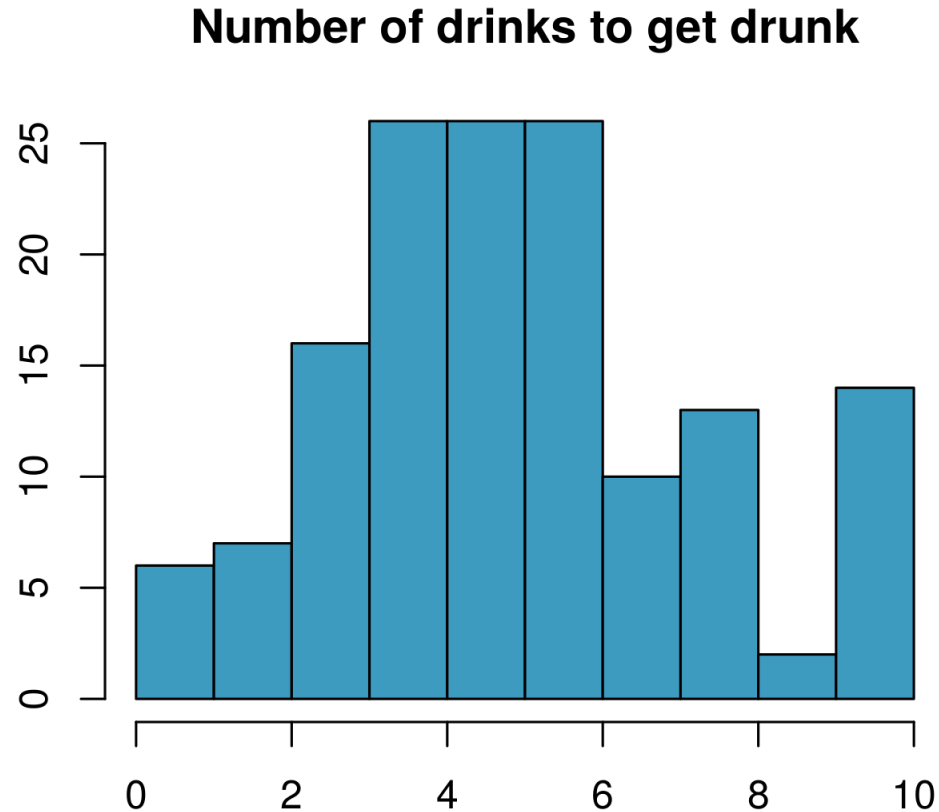
- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the **margin of error** associated with our point estimate.
- But before we get to quantifying the variability among samples, let's try to understand how and why point estimates vary from sample to sample.

Parameter estimation

Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Not the same, but only somewhat different.

The following histogram shows the distribution of number of drinks it takes a group of college students to get drunk. We will assume that this is our population of interest. If we randomly select observations from this data set, which values are most likely to be selected (which are least likely)?



Suppose that you don't have access to the population data. In order to estimate the average number of drinks it takes these college students to get drunk, you might sample from the population and use your sample mean as the best guess for the unknown population mean.

- Sample, with replacement, ten students from the population, and record the number of drinks it takes them to get drunk.
- Find the sample mean.
- Plot the distribution of the sample averages obtained by members of the class.

1	7	16	3	31	5	46	4	61	10	76	6	91	4	106	6	121	6	136	6
2	5	17	10	32	9	47	3	62	7	77	6	92	0.5	107	2	122	5	137	7
3	4	18	8	33	7	48	3	63	4	78	5	93	3	108	5	123	3	138	3
4	4	19	5	34	5	49	6	64	5	79	4	94	3	109	1	124	2	139	10
5	6	20	10	35	5	50	8	65	6	80	5	95	5	110	5	125	2	140	4
6	2	21	6	36	7	51	8	66	6	81	6	96	6	111	5	126	5	141	4
7	3	22	2	37	4	52	8	67	6	82	5	97	4	112	4	127	10	142	6
8	5	23	6	38	0	53	2	68	7	83	6	98	4	113	4	128	4	143	6
9	5	24	7	39	4	54	4	69	7	84	8	99	2	114	9	129	1	144	4
10	6	25	3	40	3	55	8	70	5	85	4	100	5	115	4	130	4	145	5
11	1	26	6	41	6	56	3	71	10	86	10	101	4	116	3	131	10	146	5
12	10	27	5	42	10	57	5	72	3	87	5	102	7	117	3	132	8		
13	4	28	8	43	3	58	5	73	5.5	88	10	103	6	118	4	133	10		
14	4	29	0	44	6	59	8	74	7	89	8	104	8	119	4	134	6		
15	6	30	8	45	10	60	4	75	10	90	5	105	3	120	8	135	6		

Example: List of random numbers: 59, 121, 88, 46, 58, 72, 82, 81, 5, 10

1	7	16	3	31	5	46	4	61	10	76	6	91	4	106	6	121	6	136	6
2	5	17	10	32	9	47	3	62	7	77	6	92	0.5	107	2	122	5	137	7
3	4	18	8	33	7	48	3	63	4	78	5	93	3	108	5	123	3	138	3
4	4	19	5	34	5	49	6	64	5	79	4	94	3	109	1	124	2	139	10
5	6	20	10	35	5	50	8	65	6	80	5	95	5	110	5	125	2	140	4
6	2	21	6	36	7	51	8	66	6	81	6	96	6	111	5	126	5	141	4
7	3	22	2	37	4	52	8	67	6	82	5	97	4	112	4	127	10	142	6
8	5	23	6	38	0	53	2	68	7	83	6	98	4	113	4	128	4	143	6
9	5	24	7	39	4	54	4	69	7	84	8	99	2	114	9	129	1	144	4
10	6	25	3	40	3	55	8	70	5	85	4	100	5	115	4	130	4	145	5
11	1	26	6	41	6	56	3	71	10	86	10	101	4	116	3	131	10	146	5
12	10	27	5	42	10	57	5	72	3	87	5	102	7	117	3	132	8		
13	4	28	8	43	3	58	5	73	5.5	88	10	103	6	118	4	133	10		
14	4	29	0	44	6	59	8	74	7	89	8	104	8	119	4	134	6		
15	6	30	8	45	10	60	4	75	10	90	5	105	3	120	8	135	6		

Sample mean: $\frac{8+6+10+4+5+3+5+6+6+6}{10} = 5.9$

Sampling distribution

What you just constructed is called a *sampling distribution*.

What is the shape and center of this distribution? Based on this distribution, what do you think is the true population average?

Sampling distribution

What you just constructed is called a *sampling distribution*.

What is the shape and center of this distribution? Based on this distribution, what do you think is the true population average?

Approximately 5.39, the true population mean.

Sampling distributions - via CLT

Central limit theorem

Central limit theorem The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim \mathcal{N} \left(\text{mean} = \mu, \text{SE} = \frac{\sigma}{\sqrt{n}} \right),$$

where SE represents **standard error**, which is defined as the standard deviation of the sampling distribution. If σ is unknown, use s (recall: standard deviation of sample).

Central limit theorem

- It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population mean.
- We won't go through a detailed proof of why $SE = \frac{\sigma}{\sqrt{n}}$, but note that as n increases SE decreases.
 - As the sample size increases we would expect samples to yield more consistent sample means, hence the variability among the sample means would be lower.

CLT - conditions

Certain conditions must be met for the CLT to apply:

- **Independence:** Sampled observations must be independent. This is difficult to verify, but is more likely if
 - random sampling/assignment is used, and
 - if sampling without replacement, $n < 10\%$ of the population.

CLT - conditions

Certain conditions must be met for the CLT to apply:

- **Independence:** Sampled observations must be independent. This is difficult to verify, but is more likely if
 - random sampling/assignment is used, and
 - if sampling without replacement, $n < 10\%$ of the population.
- **Sample size/skew:** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large. This is also difficult to verify for the population, but we can check it using the sample data, and assume that the sample mirrors the population.
 - the more skewed the population distribution, the larger sample size we need for the CLT to apply
 - for moderately skewed distributions $n > 30$ is a widely used rule of thumb

Enjoy your Break!!

Enjoy the Reading Week, and make sure to take a little bit of time for yourself. This semester (and school year) is a marathon, not a sprint - pace yourselves, stay hydrated, and remember to eat good calories.