JSC370 - Class 1

Today's Class

- Course overview
- Causal inference
- Confounding
- Contingency Tables
- Logistic Regression

Course overview

The course website is https://jsc370.github.io (https://jsc370.github.io)

Simpson's Paradox

	Drug	No Drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

- In male and female patients, drug takers had a better recovery rate than those who went without the drug.
- In the combined population, those who did not take the drug had a better recovery rate than those who did (83% vs 78%).
- What's going on?

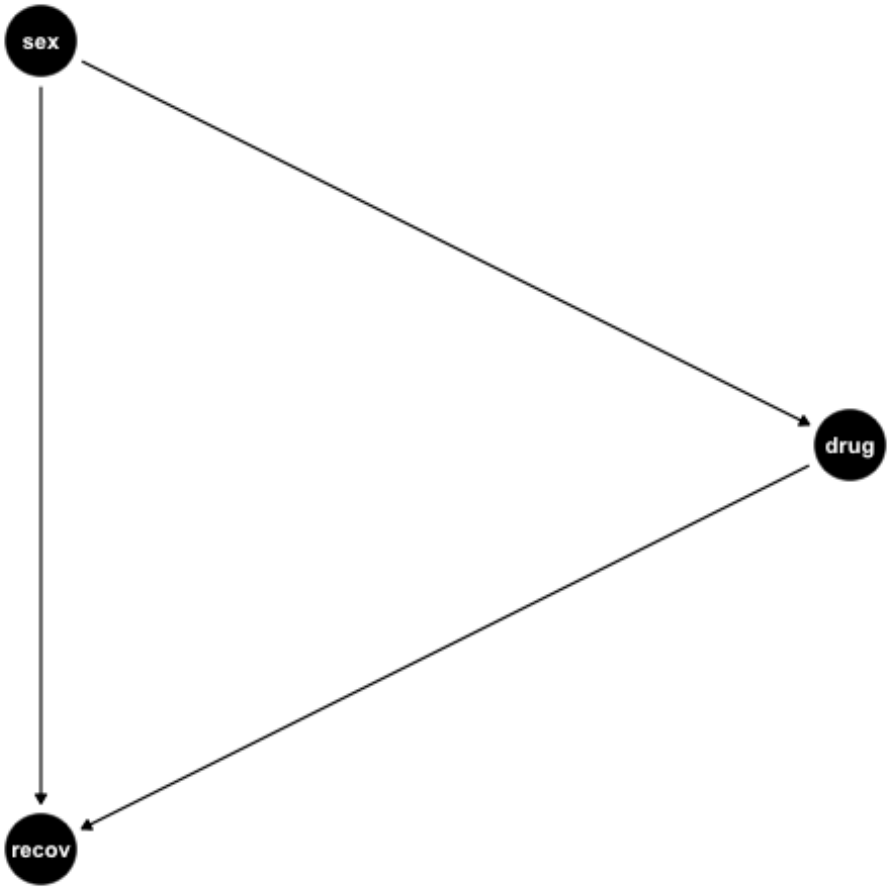
- Estrogen has a negative effect on recovery, so women are less likely to recover than men, regardless of the drug.
- Women are significantly more likely to take the drug than men are.
- So, the reason the drug appears to be harmful overall is that, if we select a drug user at random, that person is more likely to be a woman and hence less likely to recover than a random person who does not take the drug.
- Being a woman is a **common cause** of both drug taking and failure to recover.
- Therefore, to assess the effectiveness, we need to compare subjects of the same gender, thereby ensuring that any difference in recovery rates between those who take the drug and those who do not is not ascribable to estrogen.

In [19]:
```
%load_ext rpy2.ipython
```

```
The rpy2.ipython extension is already loaded. To reload it, use:
  %reload_ext rpy2.ipython
```

In [20]:

```
%%R
# Run R code in Python kernel
options(warn=-1)
library(dagitty)
library(ggdag)
d <- dagify(drug~sex,recov~sex, recov~drug)
ggdag(d, layout = "circle") + theme_void()
```



Male

	Drug	No Drug
Recovery	81	234
No Recovery	6	36
Total	87	270

Female

	Drug	No Drug
Recovery	192	55
No Recovery	71	25
Total	263	80

Use Python or R to generate the data in the male and female tables.

Logistic Regression

Let $Y = 1, 0$. The Logistic regression model is:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

In [24]:
```
df_mf = pd.concat([df_male,df_female])
# copy data frame into R namespace
%R -i df_mf
%R -i df_female
%R -i df_male
```

Use `%%R` magic to run two logistic regression models to predict `recov` in R:

`mod1` has `drug` as a covariate (feature);

`mod2` has `drug` and `sex` as covariates.

Confounding

- The coefficient for drug changes sign when sex is included in the model.
- This indicates that sex confounds the relationship between drug and recovery.

Causal inference

- Susie is a student that decided not to get a flu shot this year. One month later she develops influenza.

- Imagine that we can somehow know (e.g., by divine revelation), that had Susie got a flu shot she would not have developed the flu one month later.

- Another student, Lei, also decided not to get the flu shot at the same time as Susie, but one month later she did not develop the flu.

- The flu shot had a causal effect on Susie and Lei becoming infected with the flu in one month.

Individual Causal Effects

These two vignettes illustrate how humans reason about causal effects:

- We compare (usually only mentally) the outcome when an action A is taken with the outcome when the action A is withheld.
- If the two outcomes differ, we say that the action A has an *individual* causal effect, causative or preventive, on the outcome.

- Otherwise, we say that the action A has no causal effect on the outcome.

- Epidemiologists, statisticians, economists, and other social scientists often refer to the action A as an intervention, an exposure, or a treatment.

Treatment and Outcome

$$A = \begin{cases} 1 & \text{if treated} \\ 0 & \text{if untreated} \end{cases}$$

$$Y = \begin{cases} 1 & \text{if flu in 4 weeks} \\ 0 & \text{if no flu in 4 weeks} \end{cases}$$

- $Y^{a=1}$ (read Y under treatment $a = 1$) be the outcome variable that would have been observed under the treatment value $a = 1$, and $Y^{a=0}$ (read Y under treatment $a = 0$) be the outcome variable that would have been observed under the treatment value $a = 0$.
- A has a causal effect on an individual's outcome Y if $Y^{a=1} \neq Y^{a=0}$.
- The variables $Y^{a=1}$ and $Y^{a=0}$ are referred to as counterfactual outcomes or potential outcomes.
- For each individual, one of the counterfactual outcomes – the one that corresponds to the treatment value that the individual actually received – is actually factual.

Let's simulate the flu data for 20 subjects, where subjects have a 50:50 chance of developing flu in both treated and untreated groups.

In [28]:
```
print(' Frequency (proportion) of flu when Flu shot recieved:', sum(df['A1']),
      '(', sum(df['A1'])/len(df['A1']),')', '\n',
     'Frequency (proportion) of flu when Flu shot not recieved:', sum(df['A0']),
      '(', sum(df['A0'])/len(df['A0']),')',)
```

```
Frequency (proportion) of flu when Flu shot recieved: 10 ( 0.5 )
Frequency (proportion) of flu when Flu shot not recieved: 10 ( 0.5 )
```

- The average causal effect in this population is present if
$$P(Y^{a=1}) \neq P(Y^{a=0}).$$
- In this case $P(Y^{a=1}) = P(Y^{a=0}) = 0.5$, so the treatment flu shot (A) does not have an *average* causal effect on developing flu in one month (Y).
- The average causal effect can often be identified from data even if the individual causal effects cannot.

Observed Treatment

- The data from an actual study will look different since we can usually only observe the outcome under treatment or no treatment.
- Consider a study where a subject is treated, if their potential outcome under treatment is not developing flu; otherwise the subject is not treated. What would the observed data look like?

In [29]:
```
# A1 A0  treated
# 1   1   no
# 1   0   no
# 0   1   yes
# 0   0   yes

#subject is treated if flu shot is effective (i.e., A1 = 0)
df['A1_obs'] = df[(df['A1']==0) & (df['A0']==0)|(df['A1']==0) & (df['A0']==1)]['
A1']

#subject is not treated if flu shot is effective (i.e., A1 = 0)
df['A0_obs'] = df[(df['A0']==0) & (df['A1']==1)|(df['A0']==1) & (df['A1']==1)]['
A0']
df.head()
```

Out[29]:

	A0	A1	A1_obs	A0_obs
0	0	0	0.0	NaN
1	1	0	0.0	NaN
2	0	0	0.0	NaN
3	0	1	NaN	0.0
4	0	1	NaN	0.0

In [30]:
```
f1 = sum(df['A1_obs'].dropna())/len(df['A1_obs'].dropna())
f0 = sum(df['A0_obs'].dropna())/len(df['A0_obs'].dropna())

print(' Proportion that developed flu in treated is:', np.round(f1,2),'\n',
      'Proportion that developed flu in untreated is:', np.round(f0,2))
```

```
 Proportion that developed flu in treated is: 0.0
 Proportion that developed flu in untreated is: 0.4
```

- In this case there is an observed *association* between treatment and outcome, but it's different than the true causal effect.
- What happened?

- The treatment A is dependent on the potential outcome. In this case subjects were treated only if they would not develop flu.
- The probability of developing flu in the treated group is not the same as the probability of developing flu in the untreated group had these individuals received treatment.

In [31]:
```
df['obs_treat'] = np.where(df['A1']==0,1,0) # define treatment indicator
df['Y'] = np.where(df['obs_treat']==1, df['A1'], df['A0']) # define observed out
come
y11 = df[df['obs_treat']==1]['A1'] # num of flu if treated in obs. treatment gro
up
y10 = df[df['obs_treat']==0]['A1'] # num of flu if not treated in obs. untreated
group
sum(y11)/len(y11), sum(y10)/len(y10) # define probabilities
```

Out[31]:
```
(0.0, 1.0)
```

- The treatment assignment is not independent of the potential outcomes: $P(Y^a = 1|A = 1) \neq P(Y^a = 1|A = 0)$, for $a = 0, 1$.
- This is called lack of exchangeability or non-ignorable treatment assignment.
- Randomization is so valued since it's expected to produce exchangeability/ignorable treatment assignment.

Treatment Randomization

What if treatment is randomly assigned?

In [33]:
```
# this function assigns treatment randomly
# and returns the proportion that would develop flu
# in both groups
def rand_samp(a):
    np.random.seed(a)
    df['rand_treat'] = np.random.binomial(1,0.5,20) # prob = 0.5 of being treate
d
    y11 = df[df['rand_treat']==1]['A1']
    y01 = df[df['rand_treat']==0]['A1']
    p11 = sum(y11)/len(y11)
    p01 = sum(y01)/len(y01)
    return(p11, p01)

print(' proportion with flu in treated',rand_samp(20)[0], '\n',
      'proportion with flu in untreated',rand_samp(20)[1], '\n')

print(' proportion with flu in treated',rand_samp(200)[0], '\n',
      'proportion with flu in untreated',rand_samp(200)[1], '\n')

print(' proportion with flu in treated',rand_samp(80)[0], '\n',
      'proportion with flu in untreated',rand_samp(80)[1], '\n')
```

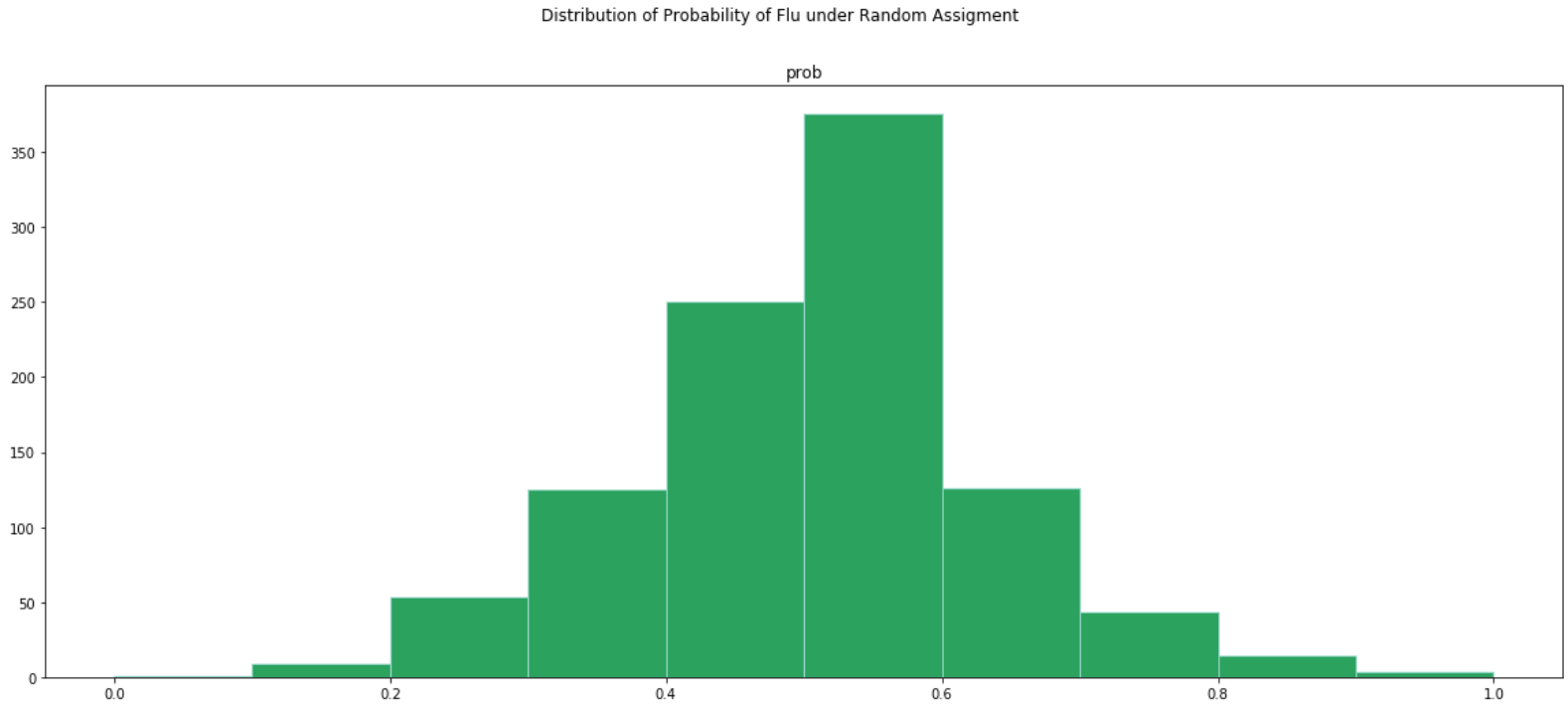
```
 proportion with flu in treated 0.5
 proportion with flu in untreated 0.5

 proportion with flu in treated 0.45454545454545453
 proportion with flu in untreated 0.5555555555555556

 proportion with flu in treated 0.6
 proportion with flu in untreated 0.4
```

In [34]:
```
import pylab as pl
probs = [rand_samp(a) for a in range(1000)]
p_treat = [probs[a][0] for a in range(1000)] #create a list of probs in flu grou
p
df = pd.DataFrame({'prob':p_treat})
df.hist(figsize=(20,8), grid = False, edgecolor = '#99d8c9', color = '#2ca25f')
pl.suptitle("Distribution of Probability of Flu under Random Assigment")
```

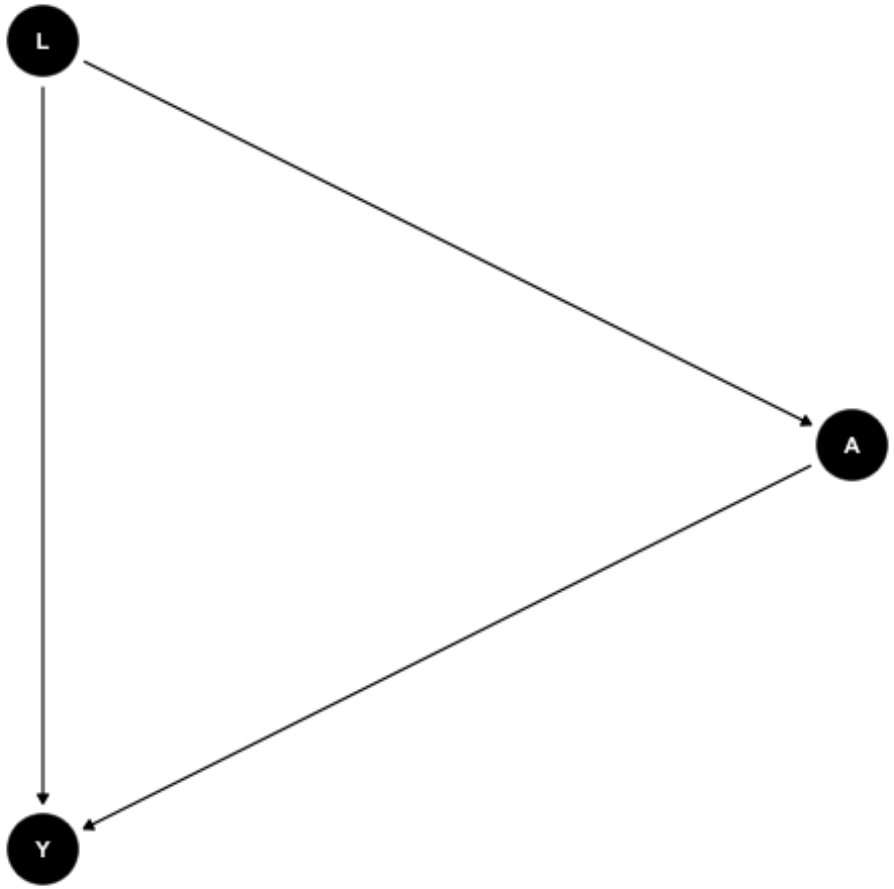
Out[34]:
```
Text(0.5, 0.98, 'Distribution of Probability of Flu under Random Assigment')
```



- The correct inference, on average, would be made if treatment were randomly assigned.
- The prob of flu from the rule where treatment was based on the potential outcome (prob = 0) is atypical of the probabilities observed.

In [35]:

```
%%R
# Run R code in Python kernel
options(warn=-1)
library(dagitty)
library(ggdag)
d <- dagify(A~L,Y~L, Y~A)
ggdag(d, layout = "circle") + theme_void()
```



An example of a directed acyclic graph (DAG) showing the causal relationships between outcome Y, treatment A, and common cause L.

Study Description

A study published in March, 1986 in the British Medical Journal examined the success of three different procedures for removing kidney stones.

The three procedures were:

1. Open surgery
2. Percutanous nephrolithotomy
3. ESWL

Methods

350 cases of open stone removal, 350 cases of percutaneous nephrolithotomy, and 352 cases of ESWL. All patients were treated by the same team of surgeons under the direct supervision of one consultant. (Carig et al.)

TABLE I—*Details of patients in each treatment group*

	No (%) of patients with stones				
	<2 cm (group 1)	≥2 cm (group 2)	Total	Mean age (range)	M:F (%)
Nephrolithotomy/pyelolithotomy	13 (6)	218 (94)	231	45 (12-78)	45:55
Pyelolithotomy	31 (41)	45 (59)	76	47 (16-72)	51:49
Ureterolithotomy	43 (100)		43	46 (20-68)	69:31
Percutaneous nephrolithotomy	270 (77)	80 (23)	350	52 (23-72)	68:32
ESWL	204 (62)	124 (38)	328	48 (22-83)	70:30
Percutaneous nephrolithotomy and ESWL		24 (100)	24		

- group 1: patients with a stone < 2cm; and group 2 patients with a stone $\geq$ 2cm.

Data

TABLE II—*Success rate of treatment* (figures are numbers (%) of patients)*

	Group 1	Group 2	Overall
Nephrolithotomy/pyelolithotomy	12 (92)	154 (71)	166 (72)
Pyelolithotomy	26 (84)	38 (84)	64 (84)
Ureterolithotomy	43 (100)		43 (100)
All open procedures	81 (93)	192 (73)	273 (78)
Percutaneous nephrolithotomy†	234 (87)	55 (69)	289 (83)
ESWL	200 (98)	101 (82)	301 (92)
Percutaneous nephrolithotomy and ESWL		15 (62)	15 (62)

*Success defined as no stones at three months or stone reduced to particles <2 mm in size.
†52 with electrohydraulic lithotripsy, 69 with ultrasound.

Questions

1. Use Python or R to write a **function** that can create this study's data set by returning a Pandas data frame/R data frame. It should contain information on patients that received open stone removal and percutaneous removal. In particular the data set should contain the type of procedure, the outcome of the procedure, and the size of the kidney stone.

2. Briefly explain why kidney stone size is a confounding variable.

3. Calculate two-way contingency tables of success versus procedure among all patients; patients with a stone size $\leq 2cm$; and patients with a stone size $> 2cm$. What do you observe?

4. Fit an appropriate logistic regression model to the data. What is the odds ratio of success in open versus percutaneous? Is it appropriate to adjust for kidney stone size?

5. Which procedure is more effective? Explain your reasoning.