The Basic Practice of Statistics
Ninth Edition

David S. Moore William I. Notz

Chapter 6
Two-Way Tables

Lecture Slides

In Chapter 6, we cover …

- Marginal distributions
- Conditional distributions
- Relative risk and odds ratio for 2x2 contingency tables
- Simpson’s paradox

Categorical variables

- **Review:** Categorical variables place individuals into one of several groups or categories.
- The values of a categorical variable are labels for the different categories.
- The distribution of a categorical variable lists the count or percent of individuals who fall into each category.
- When a data set involves two categorical variables, we begin by examining the counts or percents in various categories for one of the variables.

A two-way table describes two categorical variables, organizing counts according to a row variable and a column variable.

Two-way contingency table

In 2017, the National Center for Education Statistics projected the number of academic degrees to be awarded in 2020–2021 for men and women (Table 6.1 of the textbook)

Sex	Degrees Conferred (thousands): Associate	Degrees Conferred (thousands): Bachelor's	Degrees Conferred (thousands): Master's	Degrees Conferred (thousands): Professional/ Doctorate	Total
Women	639	1087	460	97	2283
Men	402	804	329	87	1622
Total	1041	1891	789	184	3905

- What are the variables described by this two-way table?
- How many degrees were conferred (to the nearest thousand)?

This is how data are collected

	Degree	Sex
0	Associate	women
1	Associate	women
2	Associate	women
3	Associate	women
4	Associate	women
...	...	...
3900	Professional or Doctor	men
3901	Professional or Doctor	men
3902	Professional or Doctor	men
3903	Professional or Doctor	men
3904	Professional or Doctor	men

3905 rows × 2 columns

Tabulated data

Degree Sex	Associate	Bachelor	Master	Professional or Doctor	All
men	402	804	329	87	1622
women	639	1087	460	97	2283
All	1041	1891	789	184	3905

Marginal distribution (1 of 3)

- The **marginal distribution** of one of the categorical variables, in a two-way table of counts, is the distribution of values of that variable among all individuals described by the table.
- *Note:* Percents are often more informative than counts, especially when one is comparing groups of different sizes.

To examine a marginal distribution:

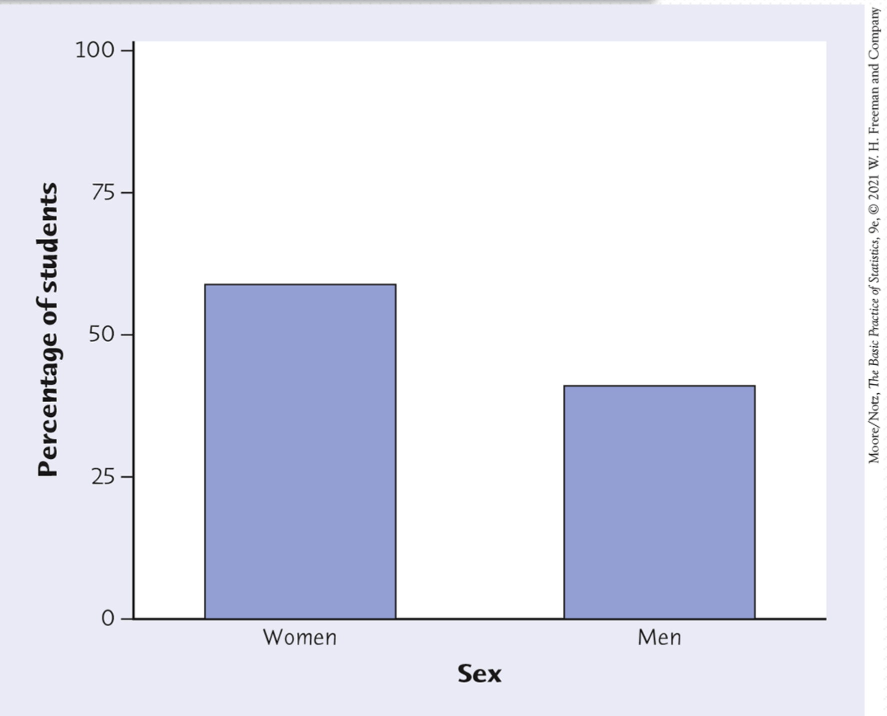
1. Use the data in the table to calculate the marginal distribution (in percents) of the row or column totals.
2. Make a graph to display the marginal distribution.

Marginal distribution (2 of 3)

Sex	Degrees Conferred (thousands): Associate	Degrees Conferred (thousands): Bachelor's	Degrees Conferred (thousands): Master's	Degrees Conferred (thousands): Professional Doctorate	Total
Women	639	1087	460	97	2283
Men	402	804	329	87	1622
Total	1041	1891	789	184	3905

Examine the **marginal distribution** of sex.

Response	Percent
Women	2283/3905= 58.5%
Men	1622/3905 = 41.5%



Marginal distribution (3 of 3)

Sex	Degrees Conferred (thousands): Associate	Degrees Conferred (thousands): Bachelor's	Degrees Conferred (thousands): Master's	Degrees Conferred (thousands): Professional/ Doctorate	Total
Women	639	1087	460	97	2283
Men	402	804	329	87	1622
Total	1041	1891	789	184	3905

Examine the **marginal distribution** of degree conferred.

Response	Percent
Associate	1041/3905= 26.7%
Bachelor's	1891/3905 = 48.4%
Master's	789/3905 = 20.2%
Doctorate	184/3905 = 4.7%

Conditional distribution (1 of 6)

- Marginal distributions tell us nothing about the relationship between two variables.

- A conditional distribution of a variable describes the values of that variable among individuals who have a given value of another variable.

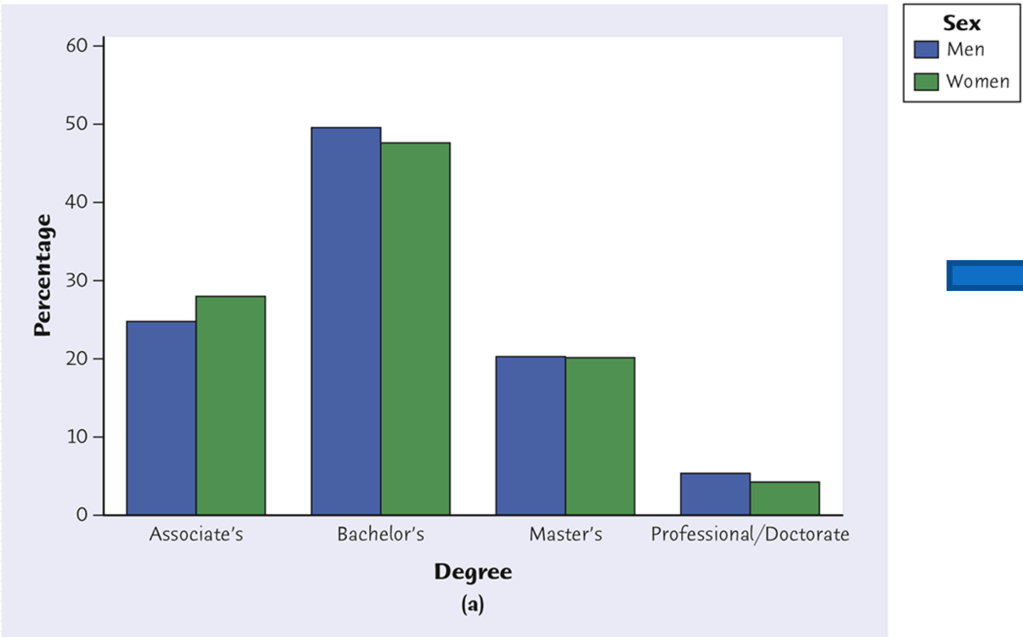
- Use software to generate a **side-by-side bar graph**, a **segmented bar graph**, or a **mosaic plot** to compare distributions.

Conditional distribution (2 of 6)

A segmented bar graph is a bar graph for presenting data about two categorical variables in which each bar is divided into parts. Each bar represents the observations that take a particular value of one variable, and the length of each part of the bar represents the proportion of those observations that take a specific value of the second variable.

A mosaic plot is a segmented bar graph in which the width of each bar represents the proportion of all observations that fall into the category that the bar represents.

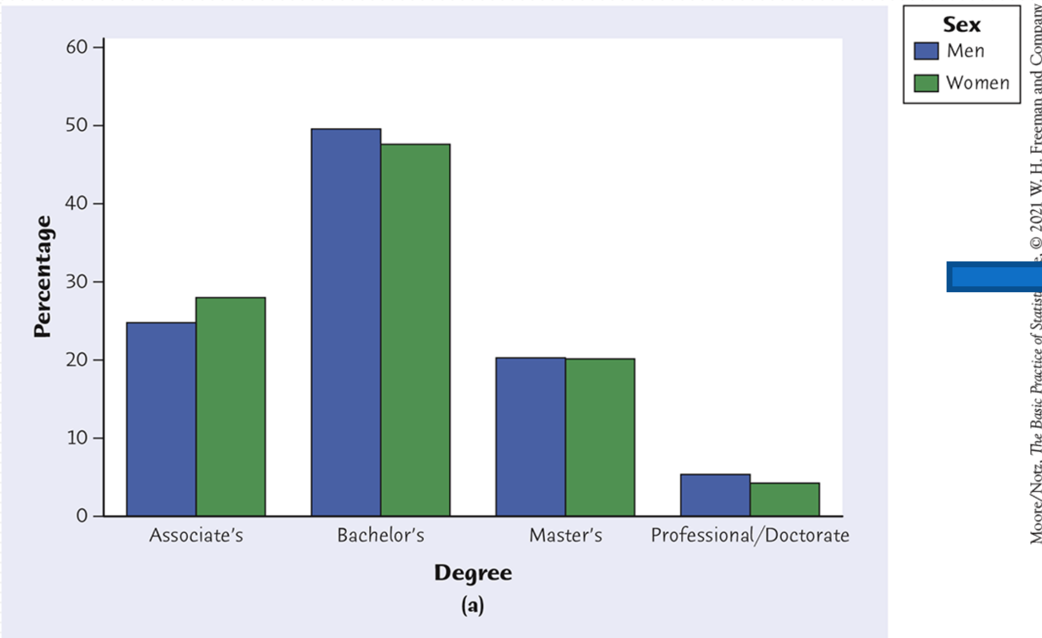
Conditional distribution (3 of 6)



Conditional distribution of **degree received given sex**

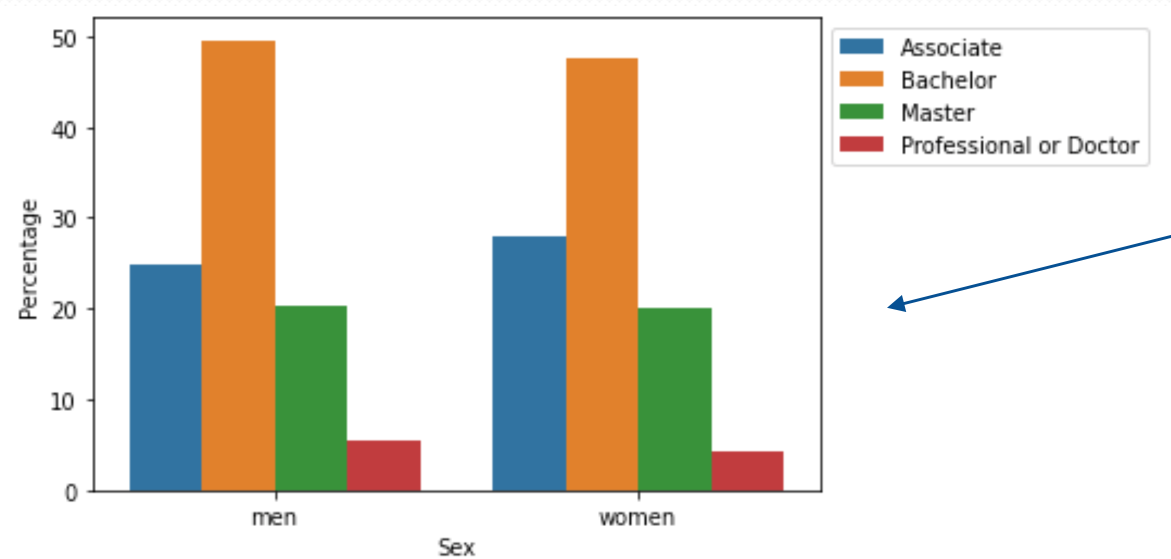
- Sum of the heights of the bars for mean should be 100%
- Sum of the heights of the bars for women should be 100%

Conditional distribution (4 of 6)



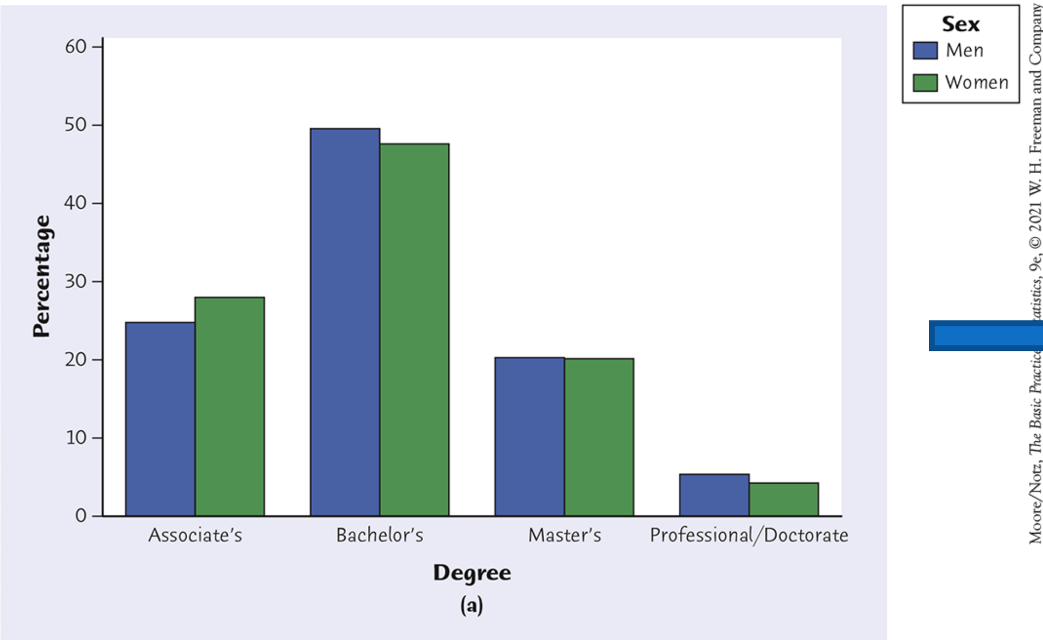
Conditional distribution of **degree received given sex**

- Sum of the heights of the bars for mean should be 100%
- Sum of the heights of the bars for women should be 100%

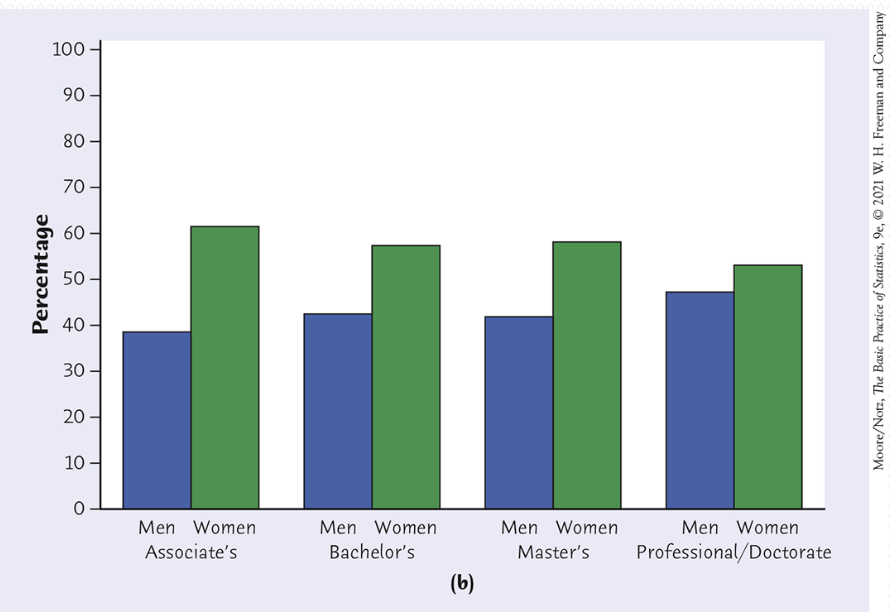


Another way we could plot the conditional distribution of **degree received given sex**

Conditional distribution (5 of 6)



Conditional distribution of degree received given sex

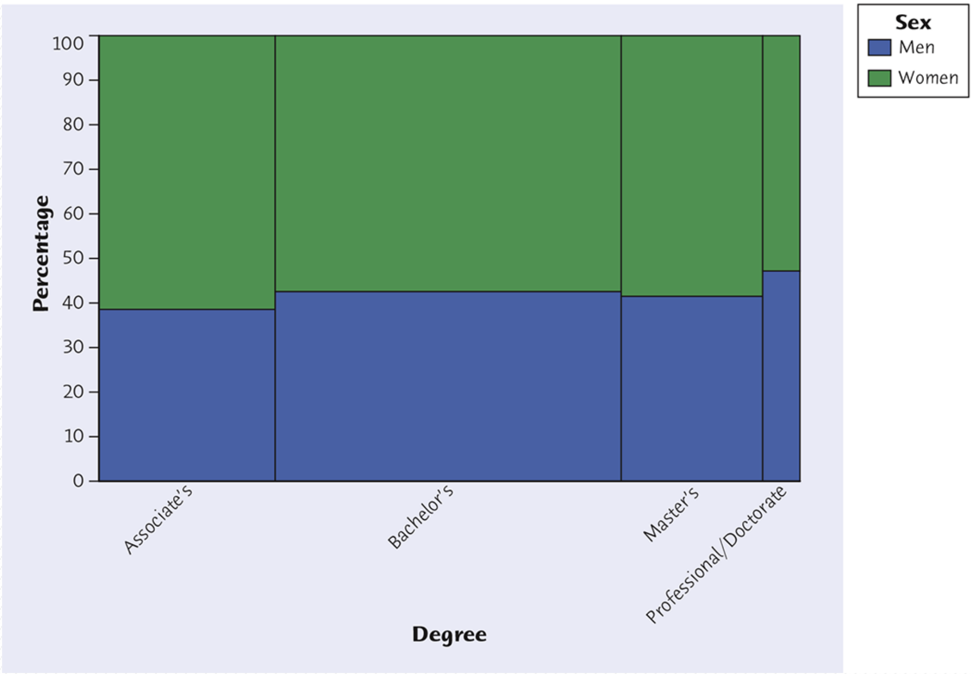


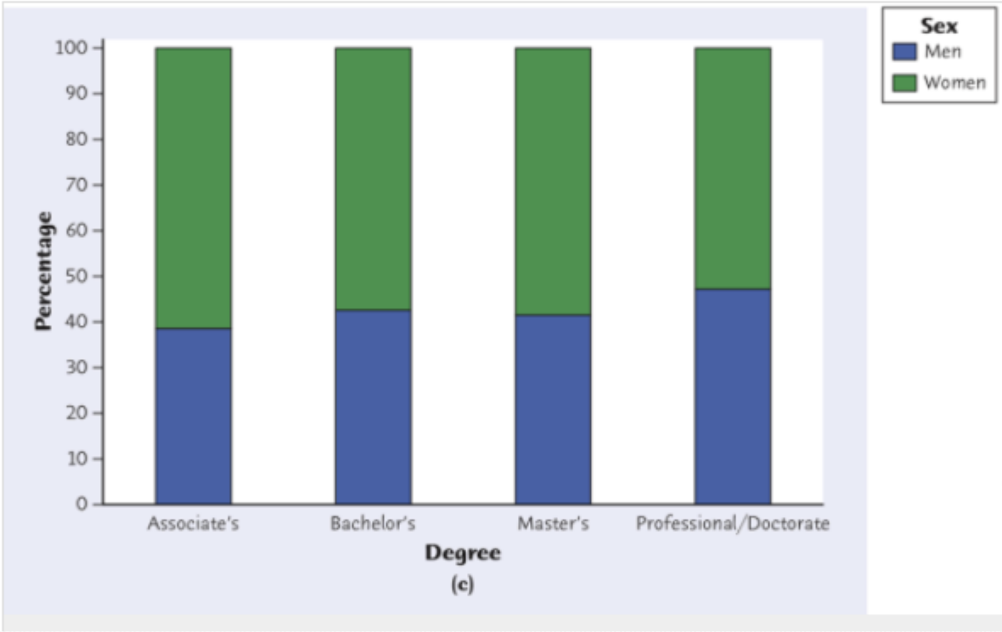
Conditional distribution of **sex given degree received**

- Sum of the heights of the bars for men and women for each degree type should sum to 100%

Conditional distribution (6 of 6)

Other plots to show conditional distributions





Mosaic plot and segmented bar graph comparing the proportions of women (green) and men (blue) among those in each degree-conferred category.

Example 6.3 in Python (1 of 7)
Libraries and importing the data

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from statsmodels.graphics.mosaicplot import mosaic
```

In [2]:
```
# Read .csv data
df = pd.read_csv("eg06-01degrees.csv")
df
```

Out[2]:

	Degree	Sex	Count
0	Associate	women	639
1	Bachelor	women	1087
2	Master	women	460
3	Professional or Doctor	women	97
4	Associate	men	402
5	Bachelor	men	804
6	Master	men	329
7	Professional or Doctor	men	87

Example 6.3 in Python (2 of 7)
Tabulating the data (creating the contingency table)

In [3]:
```
df_new = pd.DataFrame(np.repeat(df[['Degree','Sex']].values, df.Count, axis = 0),
                      columns = df[['Degree','Sex']].columns)
df_new
```

Out[3]:

	Degree	Sex
0	Associate	women
1	Associate	women
2	Associate	women
3	Associate	women
4	Associate	women
...	...	...
3900	Professional or Doctor	men
3901	Professional or Doctor	men
3902	Professional or Doctor	men
3903	Professional or Doctor	men
3904	Professional or Doctor	men

3905 rows × 2 columns

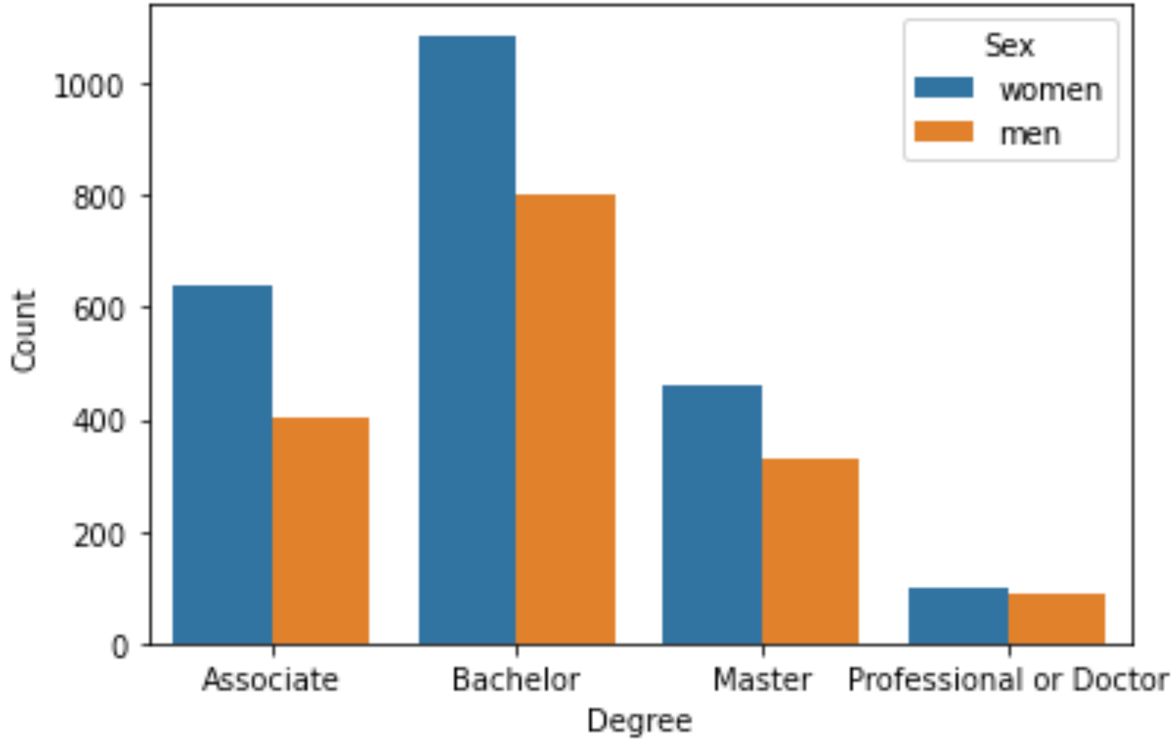
In [4]:
```
ct = pd.crosstab(index = df_new["Sex"], columns = df_new["Degree"], margins = True)
ct
```

Out[4]:

Degree	Associate	Bachelor	Master	Professional or Doctor	All
Sex					
men	402	804	329	87	1622
women	639	1087	460	97	2283
All	1041	1891	789	184	3905

Example 6.3 in Python (3 of 7)
Bar plot of table counts

In [5]:
```
sns.barplot(x = "Degree", hue = "Sex", y = "Count", data = df)
plt.show()
```



Example 6.3 in Python (4 of 7)
Computing the conditional proportions

```
In [6]: # calculating the proportions of types of degrees conferred conditional on sex
        conditional_sex = pd.crosstab(index = df_new["Sex"],
                                      columns = df_new["Degree"], normalize = 'index')
        conditional_sex
        #help(pd.crosstab)
```

Out[6]:

Degree Sex	Associate	Bachelor	Master	Professional or Doctor
men	0.247842	0.495684	0.202836	0.053637
women	0.279895	0.476128	0.201489	0.042488

Conditional distribution of degree received given sex

```
In [7]: # calculating the proportions of men and women conditional on degree type
        conditional_degree = pd.crosstab(index = df_new["Sex"],
                                         columns = df_new["Degree"], normalize = 'columns')
        conditional_degree
```

Out[7]:

Degree Sex	Associate	Bachelor	Master	Professional or Doctor
men	0.386167	0.425172	0.416984	0.472826
women	0.613833	0.574828	0.583016	0.527174

Conditional distribution of sex given degree received

Example 6.3 in Python (5 of 7)

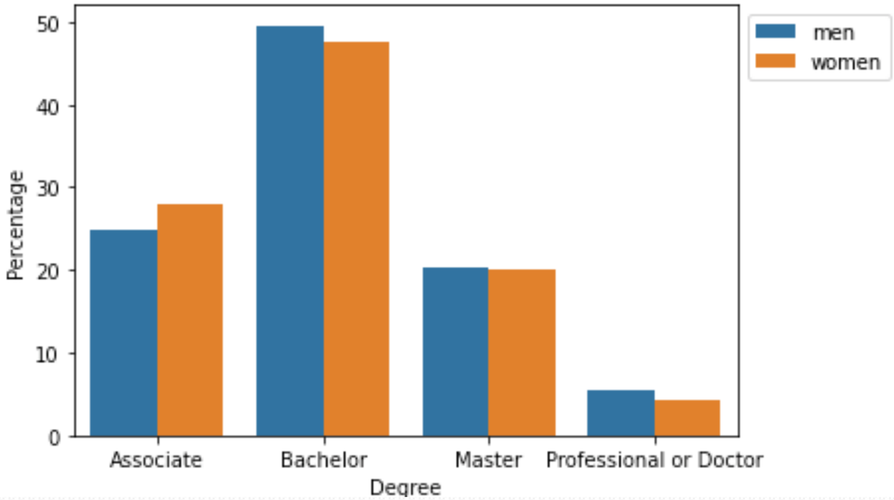
Bar plot of conditional distribution of degree received given sex

In [14]:
```
stacked = conditional_sex.stack().reset_index().rename(columns = {0:'Percentage'})
stacked['Percentage'] = stacked['Percentage']*100
stacked
```

Out[14]:

	Sex	Degree	Percentage
0	men	Associate	24.784217
1	men	Bachelor	49.568434
2	men	Master	20.283600
3	men	Professional or Doctor	5.363748
4	women	Associate	27.989488
5	women	Bachelor	47.612790
6	women	Master	20.148927
7	women	Professional or Doctor	4.248795

In [15]:
```
sns.barplot(x = "Degree", hue = "Sex", y = "Percentage", data = stacked)
plt.legend(bbox_to_anchor = (1, 1), loc = 2)
plt.show()
```



Conditional distribution of degree received given sex

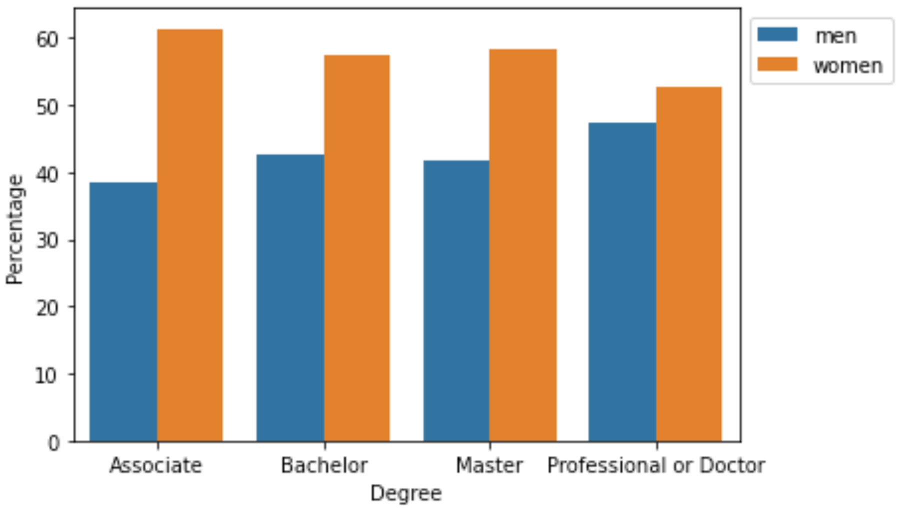
Example 6.3 in Python (6 of 7)
Bar plot of conditional distribution of sex given degree received

In [16]:
```
stacked_degree = conditional_degree.stack().reset_index().rename(columns = {0:'Percentage'})
stacked_degree['Percentage'] = stacked_degree['Percentage']*100
stacked
```

Out[16]:

	Sex	Degree	Percentage
0	men	Associate	24.784217
1	men	Bachelor	49.568434
2	men	Master	20.283600
3	men	Professional or Doctor	5.363748
4	women	Associate	27.989488
5	women	Bachelor	47.612790
6	women	Master	20.148927
7	women	Professional or Doctor	4.248795

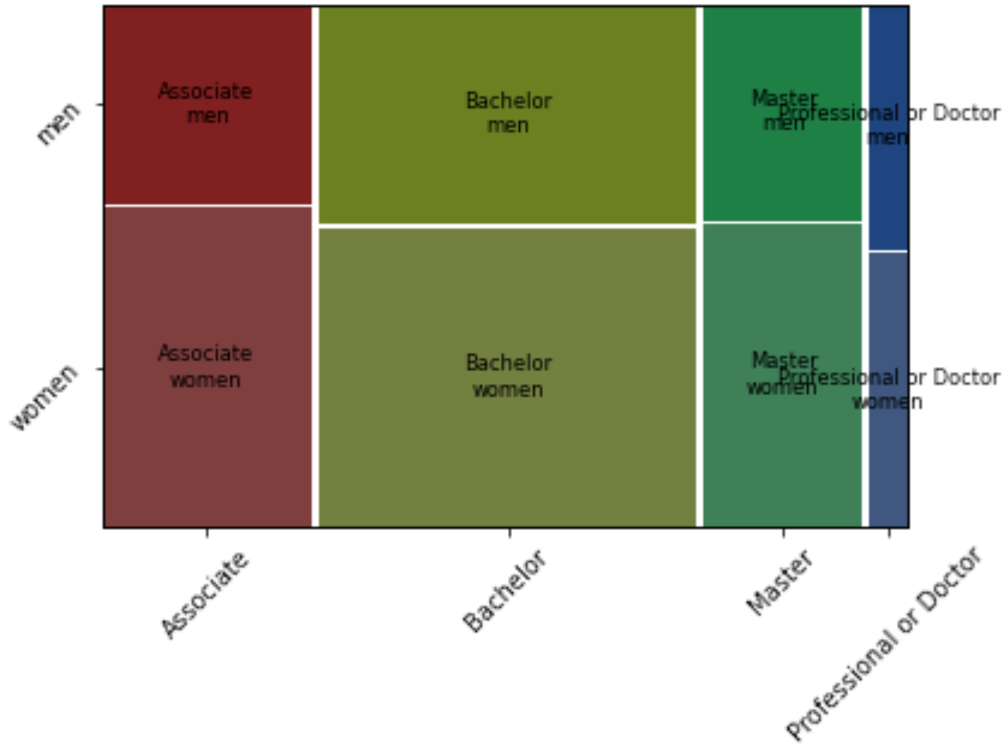
In [17]:
```
sns.barplot(x = "Degree", hue = "Sex", y = "Percentage", data = stacked_degree)
plt.legend(bbox_to_anchor = (1, 1), loc = 2)
plt.show()
```



Conditional distribution of sex given degree received

Example 6.3 in Python (7 of 7)
Mosaic plot

In [12]:
```
mosaic(df_new, ['Degree', 'Sex'], label_rotation = 45, gap = 0.01)
plt.show()
```



Relative risk and *odds ratio* for 2x2 contingency tables

Note: *a 2x2 table is a two-way table where each categorical variable take only 2 categories*

Example of 2x2 contingency table

First Child at Age 25 or Older?	Breast Cancer	No Breast Cancer	Total
Yes	31	1597	1628
No	65	4475	4540
Total	96	6072	6168

Source: Pagano and Gauvreau (1988, p. 133).

Risk, Probability, and Odds

A population contains 1000 individuals,
of which 400 carry the gene for a disease.
Equivalent ways to express this proportion:

- Forty ***percent*** (40%) of all individuals carry the gene.
- The ***proportion*** who carry the gene is 0.40.
- The ***probability*** that someone carries the gene is .40.
- The ***risk*** of carrying the gene is 0.40.
- The ***odds*** of carrying the gene are 4 to 6 (or 2 to 3, or 2/3 to 1).

Risk, Probability, and Odds

Percentage with trait =
(number with trait/total)×100%

Proportion with trait = number with trait/total

Probability of having trait = number with trait/total

Risk of having trait = number with trait/total

Odds of having trait =
(number with trait/number without trait) to 1

Baseline Risk and Relative Risk

Baseline Risk: risk without treatment or behavior

- Can be difficult to find.
- If placebo included,
  baseline risk = risk for placebo group.

Relative Risk: of outcome for two categories of explanatory variable is ratio of risks for each category.

- *Relative risk of 3*: risk of developing disease for one group is 3 times what it is for another group.
- *Relative risk of 1*: risk is same for both categories of the explanatory variable (or both groups).

Example: Relative Risk of Developing Breast Cancer

First Child at Age 25 or Older?	Breast Cancer	No Breast Cancer	Total
Yes	31	1597	1628
No	65	4475	4540
Total	96	6072	6168

- Risk for women having first child at 25 or older
  = 31/1628 = 0.0190
- Risk for women having first child before 25
  = 65/4540 = 0.0143
- Relative risk = 0.0190/0.0143 = **1.33**

Risk of developing breast cancer is 1.33 times greater for women who had their first child at 25 or older.

Source: Pagano and Gauvreau (1988, p. 133).

Odds Ratio

First Child at Age 25 or Older?	Breast Cancer	No Breast Cancer	Total
Yes	31	1597	1628
No	65	4475	4540
Total	96	6072	6168

<u>Odds Ratio:</u> ratio of the odds of getting the disease to the odds of not getting the disease.

Example: Odds Ratio for Breast Cancer

- Odds for women having first child at age 25 or older = $\boxed{31/1597 = 0.0194}$
- Odds for women having first child before age 25 = 65/4475 = 0.0145
- Odds ratio = 0.0194/0.0145 = 1.34

Alternative formula: $\text{odds ratio} = \dfrac{31 \times 4475}{1597 \times 65} = 1.34$

For Those Who Like Formulas

To represent the *observed numbers* in a 2×2 contingency table, we use the notation:

	Variable 2		
Variable 1	Yes	No	Total
Yes	a	b	$a+b$
No	c	d	$c+d$
Total	$a+c$	$b+d$	n

Relative Risk and Odds Ratio

Using the notation for the observed numbers, if variable 1 is the explanatory variable and variable 2 is the response variable, then we can compute

$$\text{relative risk} = \frac{a(c+d)}{c(a+b)}$$

$$\text{odds ratio} = \frac{ad}{bc}$$

Simpson's paradox (1 of 4)

- Affecting the relationship between two variables, there may exist a **lurking variable.** Ignoring a lurking variable creates a reversal in the direction of the relationship that exists when the lurking variable is considered.
- The lurking variable creates subgroups, and failure to take these subgroups into consideration can lead to misleading conclusions regarding the association between the two variables.

An association that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox**.

Simpson's paradox (2 of 4)

- Consider the survival rates for the following groups of victims, who were taken to the hospital either by helicopter or by road.

Counts	Helicopter	Road
Victim died	64	260
Victim survived	136	840
Total	200	1100

Percents	Died	Survived
Helicopter	**64/200=32%**	68%
Road	260/1100=24%	76%

- A higher percent of those transported by helicopter died. Does this mean that this (more costly) mode of transportation isn't helping?

Simpson’s paradox (3 of 4)

Consider the survival rates when broken down by type of accident.

Serious accidents

Counts	Helicopter	Road
Died	48	60
Survived	52	40
Total	100	100

Percents	Died	Survived
Helicopter	48%	52%
Road	**60%**	40%

Less serious accidents

Counts	Helicopter	Road
Died	16	200
Survived	84	800
Total	100	1000

Percents	Died	Survived
Helicopter	16%	84%
Road	**20%**	80%

Simpson’s paradox (4 of 4)

- Lurking variable: Accidents were of two sorts—serious (100) and less serious (1000).
- Helicopter evacuations had a higher survival rate within both types of accidents than did road evacuations.
- This is not evidence of the inefficacy of helicopter evacuation!
- This is an example of **Simpson’s paradox**.
- When the lurking variable (type of accident: serious or less serious) is ignored, the data seem to suggest that road evacuations are safer than helicopter evacuations.
- However, when the type of accident is considered, the association is reversed and suggests that helicopter evacuations are, in fact, safer.
- **Can be dangerous to summarize information over groups!**

Extra example Simpson’s paradox

Simpson's Paradox: The Missing Third Variable

- **Relationship appears to be in one direction if third variable is *not* considered and in other direction if it is.**
- **Can be dangerous to summarize information over groups.**

Example: Simpson's Paradox for Hospital Patients

Survival Rates for Standard and New Treatments

	Hospital A			Hospital B		
	Survive	Die	Total	Survive	Die	Total
Standard	5	95	100	500	500	1000
New	100	900	1000	95	5	100
Total	105	995	1100	595	505	1100

Risk Compared for Standard and New Treatments

	Hospital A	Hospital B
Risk of dying with the standard treatment	95/100 = 0.95	500/1000 = 0.50
Risk of dying with the new treatment	900/1000 = 0.90	5/100 = 0.05
Relative risk	0.95/0.90 = 1.06	0.50/0.05 = 10.0

Looks like ***new treatment is a success*** at both hospitals, especially at Hospital B.

Example: Simpson's Paradox for Hospital Patients

Estimating the Overall Reduction in Risk

	Survive	Die	Total	Risk of Death
Standard	505	595	1100	595/1100 = 0.54
New	195	905	1100	905/1100 = 0.82
Total	700	1500	2200	

What has gone wrong? With combined data it looks like the ***standard treatment is superior!*** Death rate for standard treatment is only 66% of what it is for the new treatment.

HOW?

More serious cases were treated at Hospital A (famous research hospital); more serious cases were also more likely to die, no matter what. ***And*** a higher proportion of patients at Hospital A received the new treatment.