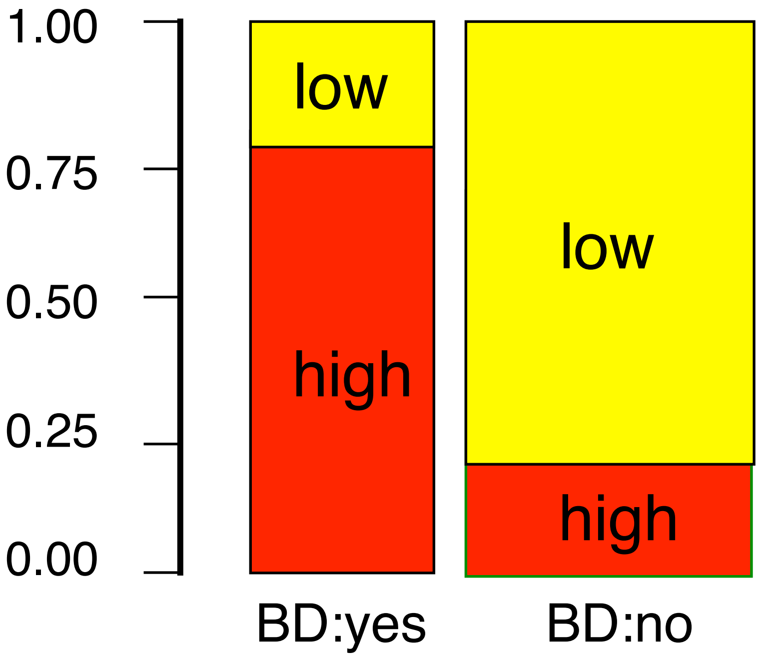
10.2 Hypothesis Testing with Two-Way Tables

How do we describe the relationship between two categorical variables?

How do we test to see if the variables are related to each other or not?

Recall earlier example…



Two categorical variables

BachelorDegree: yes or no

Salary: high or low

Here, the relationship is shown graphically in a **mosaic plot.**

Recall earlier example…

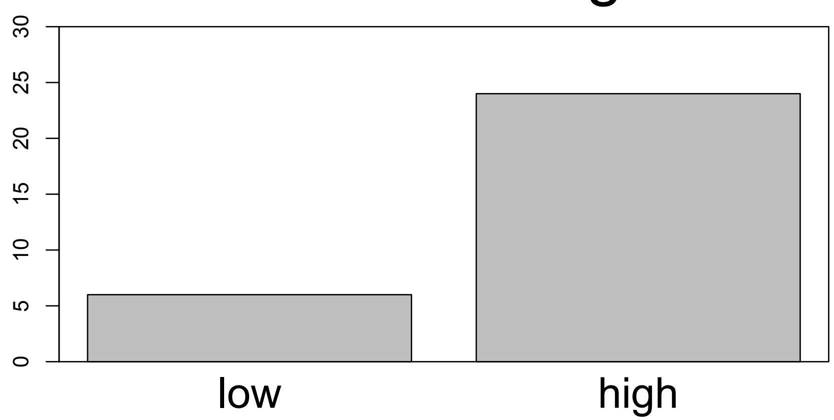
Two categorical variables

BachelorDegree: yes or no

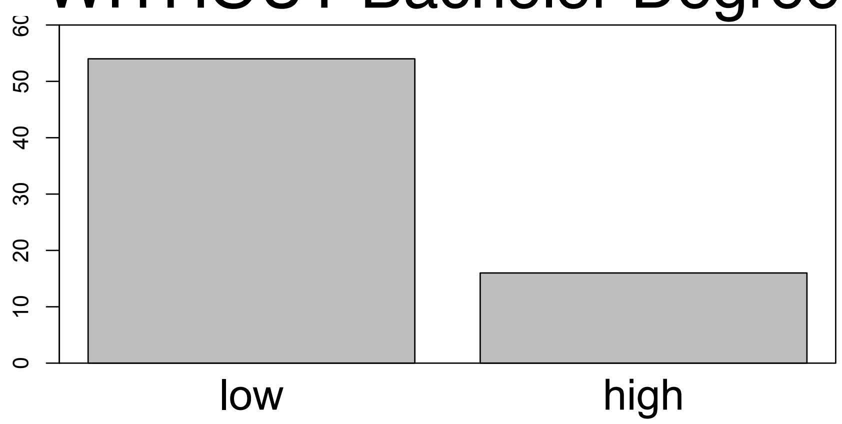
Salary: high or low

Here, the relationship is shown graphically with separate **bar graphs**.

WITH Bachelor Degree



WITHOUT Bachelor Degree



Recall earlier example…

Two categorical variables

BachelorDegree: yes or no

Salary: high or low

		Salary		
		low	high	total
Degree	yes	6	24	30
	no	54	16	70
	total	60	40	100

Here, the relationship is shown using a **two-way table** containing the count of individuals (out of 100) in each respective cell of the table.

Two-Way Tables

A **two-way table** shows the relationship between two variables by listing one variable in the rows and the other variable in the columns.

The entries in the table's cells are called *frequencies* (or *counts*).

A two-way table is also called a **contingency table**.

Recall earlier example…

		Salary		
		low	high	total
Degree	yes	6	24	30
	no	54	16	70
	total	60	40	100

Two categorical variables

- **BachelorDegree**: yes or no
- **Salary**: high or low

If there IS a relationship between **BachelorDegree** and **Salary**, then whether someone does or does not have a degree impacts whether they will or will not have a high salary.

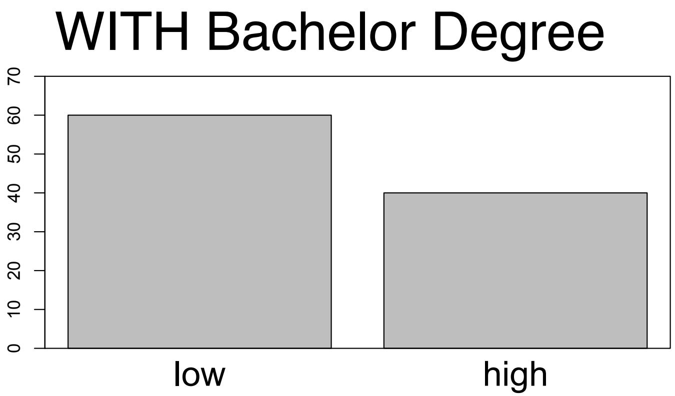
If there IS NOT a relationship, then having a degree does not impact the chance of having a high salary.

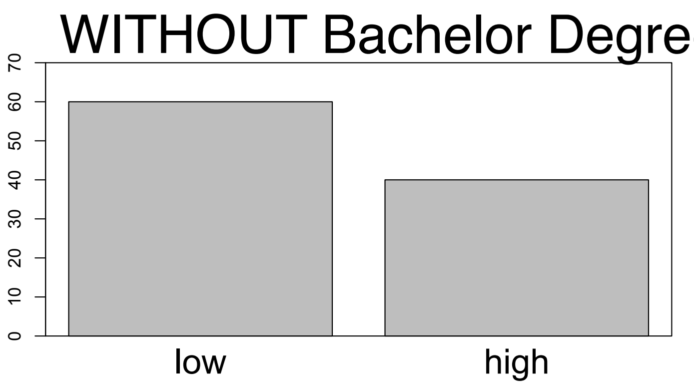
Hypothesis test for two-way tables

- If there IS NOT a relationship, then the categorical variables do not impact each other.
  - H_0: the variables are independent (no relationship exists)
- If there IS a relationship, then the categorical variables DO impact each other.
  - H_a: there is a relationship between the two variables

Performing the hypothesis test

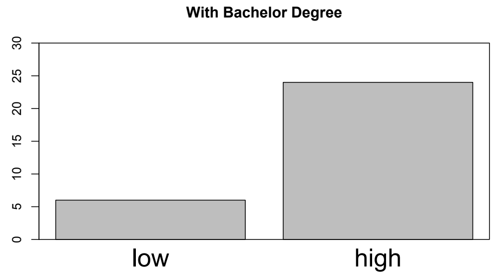
- If the null were true (i.e. there was no relationship) what would we have expected to see in this table?
- Perhaps we would've expected similar bar graphs for each **BachelorDegree** group?

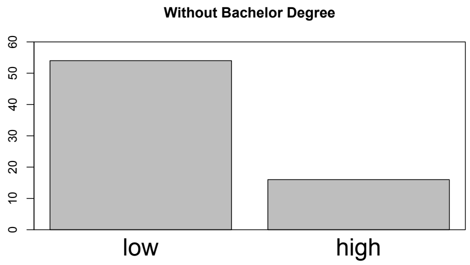




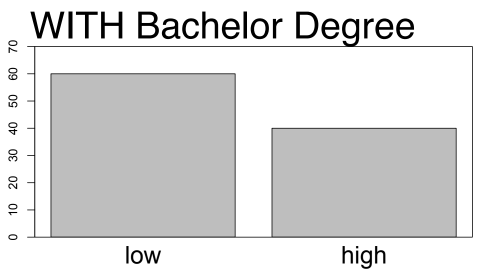
Performing the hypothesis test

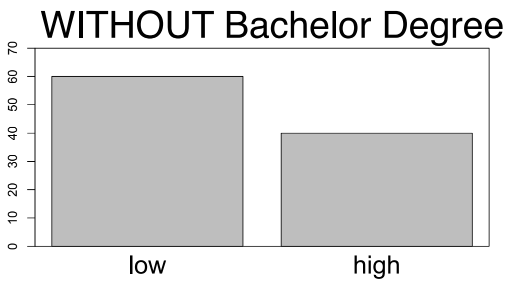
- Are the actual observed bar graphs (below)





different enough from ‘equal’ bar graphs (below) to say the variables ARE related?





Performing the hypothesis test

- Let's calculate the frequencies (counts) we would have expected if the null were true (i.e. there was no relationship).
- For this calculation, we remove the cell counts, but leave the row and column totals as is…

		Salary		
		low	high	total
Degree	yes	?	?	30
	no	?	?	70
	total	60	40	100

Performing the hypothesis test

- We convert the row and columns to **relative frequencies**…

		Salary		
		low	high	total
Degree	yes	?	?	30
	no	?	?	70
	total	60	40	100

		Salary		
		low	high	total
Degree	yes	?	?	30/100
	no	?	?	70/100
	total	60/100	40/100	100/100

Performing the hypothesis test

- We convert the row and columns to **relative frequencies**…

		Salary		
		low	high	total
Degree	yes	?	?	30
	no	?	?	70
	total	60	40	100

		Salary		
		low	high	total
Degree	yes	?	?	0.30
	no	?	?	0.70
	total	0.60	0.40	1

Performing the hypothesis test

- If the variables are independent (i.e. H_0 is true), then…

$$
\begin{aligned}
P(\text{yes and low}) &= P(\text{yes}) \times P(\text{low}) \\
&= 0.30 \times 0.60 \\
&= 0.18
\end{aligned}
$$

Relative frequency table

		Salary		
		low	high	total
Degree	yes	0.18	?	0.30
	no	?	?	0.70
	total	0.60	0.40	1

Performing the hypothesis test

- If the variables are independent (i.e. H_0 is true), then…

$$\begin{aligned} P(\text{yes and high}) &= P(\text{yes}) \times P(\text{high}) \\ &= 0.30 \times 0.40 \\ &= 0.12 \end{aligned}$$

Relative frequency table

		Salary		
		low	high	total
Degree	yes	0.18	0.12	0.30
	no	?	?	0.70
	total	0.60	0.40	1

Performing the hypothesis test

- If the variables are independent (i.e. H_0 is true), then…

$$\begin{aligned} P(\text{no and low}) &= P(\text{no}) \times P(\text{low}) \\ &= 0.70 \times 0.60 \\ &= 0.42 \end{aligned}$$

Relative frequency table

		Salary		
		low	high	total
Degree	yes	0.18	0.12	0.30
	no	0.42	?	0.70
	total	0.60	0.40	1

Performing the hypothesis test

- If the variables are independent (i.e. H_0 is true), then…

$$\begin{aligned} P(\text{no and high}) &= P(\text{no}) \times P(\text{high}) \\ &= 0.70 \times 0.40 \\ &= 0.28 \end{aligned}$$

Relative frequency table

		Salary		
		low	high	total
Degree	yes	0.18	0.12	0.30
	no	0.42	0.28	0.70
	total	0.60	0.40	1

Performing the hypothesis test

- Convert the relative frequencies back to counts by multiplying by the total count of individuals (100 in this case)

		Salary		
		low	high	total
Degree	yes	18	12	30
	no	42	28	70
	total	60	40	100

If the variables were not related, I would have expected that 28 of the 100 individuals had no degree and had a high salary, for example.

Performing the hypothesis test

- I can now compare the **expected counts** under H_0 true to the **observed counts.**

observed counts

		Salary	
		low	high
Degree	yes	6	24
	no	54	16

expected counts

		Salary	
		low	high
Degree	yes	18	12
	no	42	28

- For each cell (there are 4 in this case), we will compare the observed and expected counts to create a test statistic for our hypothesis test.

Performing the hypothesis test

- The Chi-Square Statistic:

$$\chi^2 = \text{sum of all values } \frac{(\textit{Observed - Expected})^2}{\textit{Expected}}$$

observed counts

		Salary	
		low	high
Degree	yes	6	24
	no	54	16

expected counts

		Salary	
		low	high
Degree	yes	18	12
	no	42	28

Performing the hypothesis test

- The Chi-Square Statistic:

$$\chi^2 = \frac{(6-18)^2}{18} + \frac{(24-12)^2}{12} + \frac{(54-42)^2}{42} + \frac{(16-28)^2}{28} = 28.57$$

observed counts

		Salary	
		low	high
Degree	yes	6	24
	no	54	16

expected counts

		Salary	
		low	high
Degree	yes	18	12
	no	42	28

Performing the hypothesis test

- Making the decision: $\chi^2 = 28.57$
- Table 10.7 gives the critical values of χ^2 for two significance levels, 0.05 and 0.01.

Table 10.7 Critical Values of χ^2: Reject H_0 Only If $\chi^2 >$ Critical Value

Table size (rows × columns)	Significance level 0.05	Significance level 0.01
2 × 2	3.841	6.635
2 × 3 or 3 × 2	5.991	9.210
3 × 3	9.488	13.277
2 × 4 or 4 × 2	7.815	11.345
2 × 5 or 5 × 2	9.488	13.277

Our test is significant at the 0.01 level because $\chi^2 = 28.57$ is larger than the 0.01 critical value of 6.635.

Performing the hypothesis test

- Conclusion: There is statistically significant evidence that the attainment of a bachelor's degree is related to one's salary.

NOTE: Making the decision

- A larger calculated χ^2 value gives stronger evidence against the null (in the same way that a larger absolute value of a z-value does).

- Notice that the **critical values differ for different table sizes**, so you must make sure you read the critical values for a data set from the appropriate table size row.