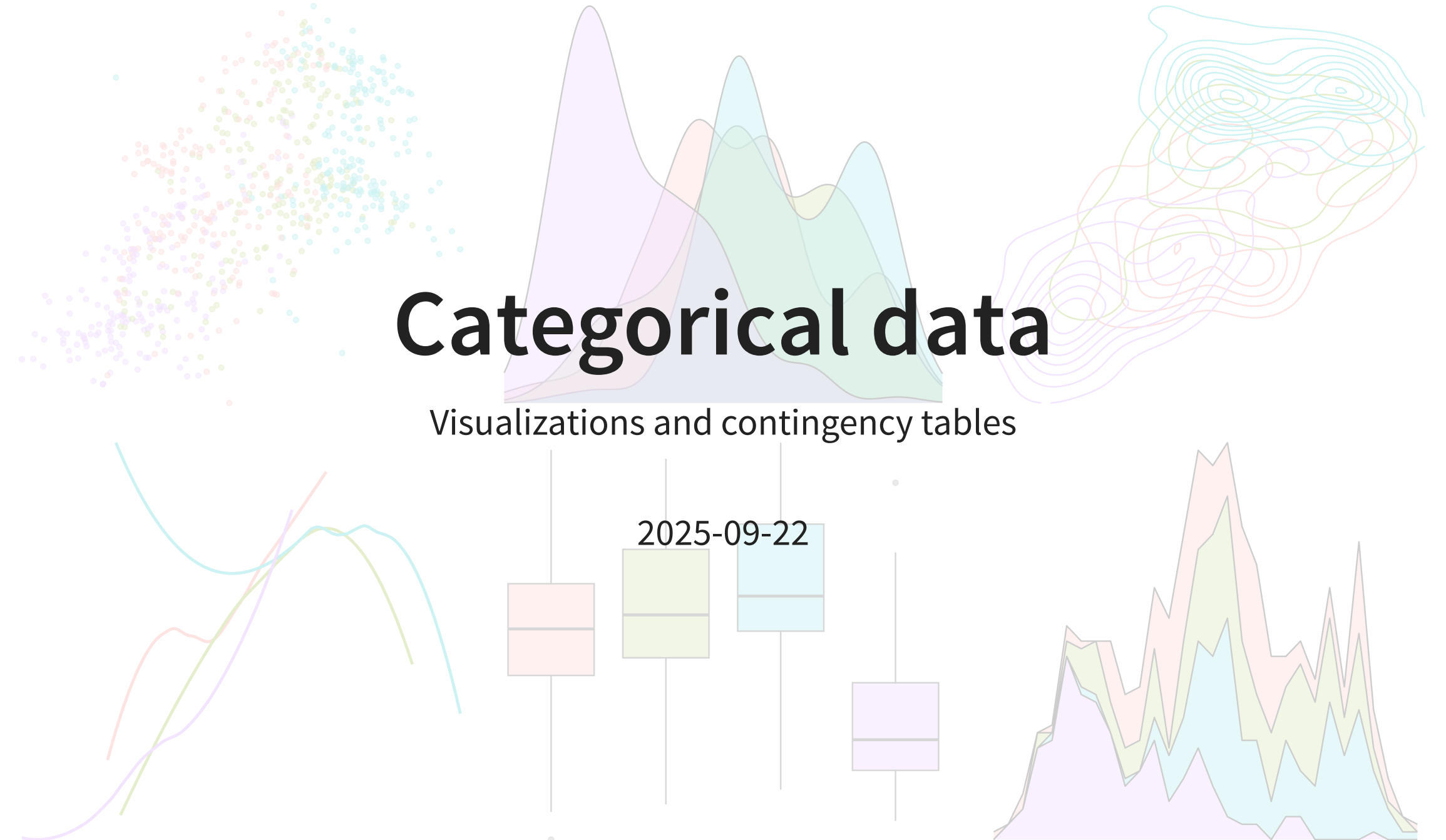


Categorical data

Visualizations and contingency tables

2025-09-22

Housekeeping

- Problem set 2 due tonight! Please be sure to submit both written and rendered parts by combining into a single PDF
- Problem set 1 graded

Categorical data

- Recall that a variable is either numerical or categorical
- **Categorical** variables are variables that can take one of a limited (usually fixed) number of possible values, known as **levels**
  - Represent data that can be divided into groups
- Two types:
  - **Ordinal**: the levels have a special ordering
  - **Nominal**: the levels don't have an ordering
    - We almost exclusively treat categorical variables as nominal in this class
- Example:
  - Blood type (A, B, AB, O)
  - Education level (high school, college, graduate degree, other)

Insurance data

Show 5 entries Search:

	age	sex	bmi	children	smoker	region	charges
1	19	female	27.9	0	yes	southwest	16884.924
2	18	male	33.77	1	no	southeast	1725.5523
3	28	male	33	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.47061
5	32	male	28.88	0	no	northwest	3866.8552

Showing 1 to 5 of 200 entries

Previous 1 2 3 4 5 … 40 Next

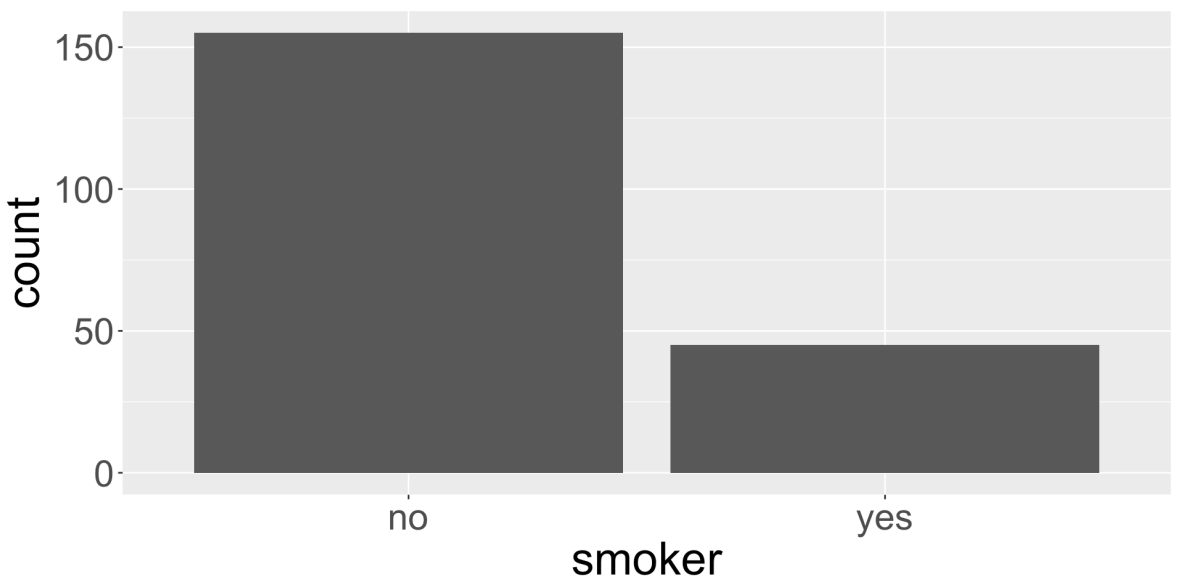
Univariate EDA

If we are interested in understanding the distribution of a single categorical variable, it is common to:

Display a **frequency table**, which is a table of counts of each level

```
# A tibble: 2 × 2
  smoker     n
  <chr>  <int>
1 no       155
2 yes       45
```

Create a **bar plot**, where different levels are displayed on one axis and the counts are portrayed on the other



Bivariate EDA

- Perhaps we are interested in examining the distribution of two categorical variables at the same time
- Summarize the distribution using a two-way table known as a **contingency table**:
  - Each value in the table counts the number of times a particular combination of variable 1 and variable 2 levels occurred in data

Contingency table

smoker	**female**	**male**
no	87	68
yes	17	28

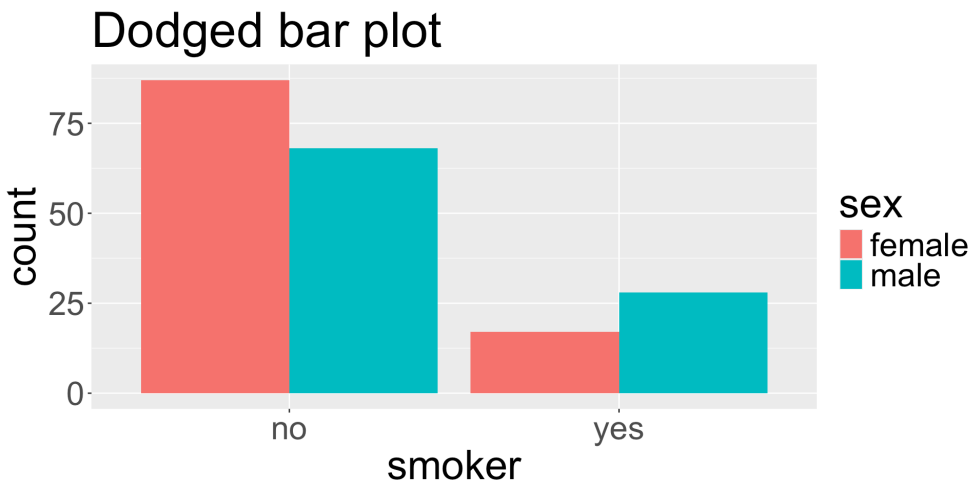
- How can we use contingency table to obtain the distribution of just one of the variables?

Dodged bar plot

The **dodged bar plot** directly converts the contingency table to a visualization.

Contingency table

smoker	female	male
no	87	68
yes	17	28

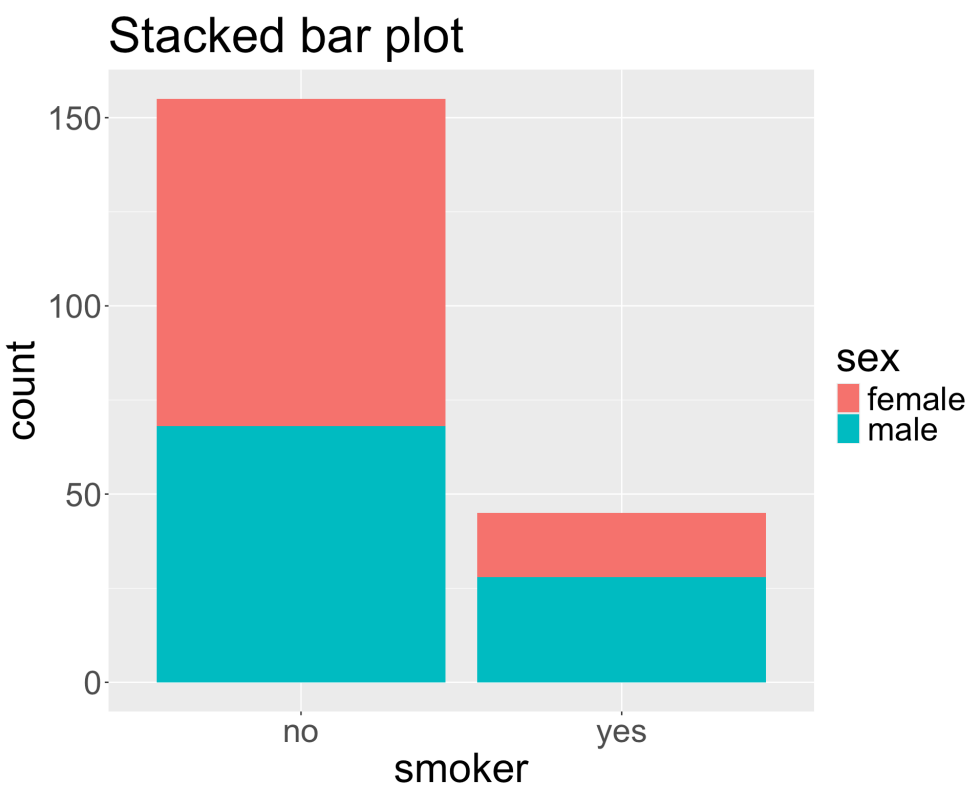


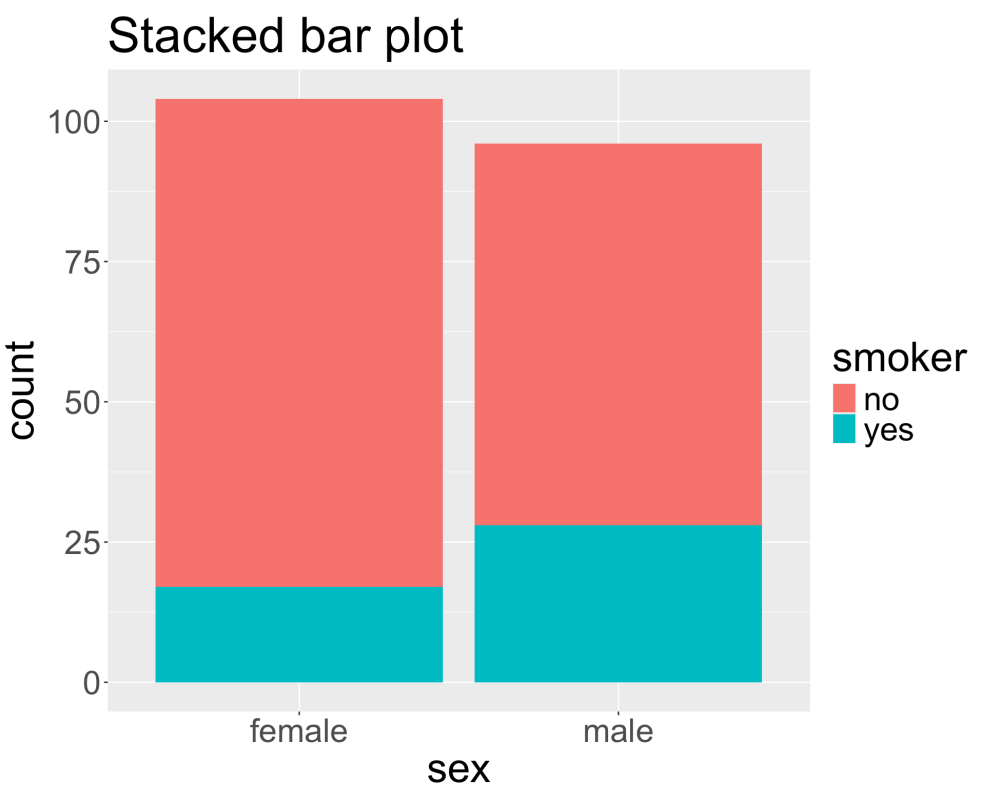
Stacked bar plot

The **stacked bar plot** looks at the counts either row-wise or column-wise.

Contingency table

smoker	female	male
no	87	68
yes	17	28





Proportions

Can convert the contingency table to proportions row-wise or column-wise to obtain the fractional breakdown of one variable in another.

Contingency table

smoker	female	male
no	87	68
yes	17	28

Row-wise proportions

smoker	female	male
no	0.561	0.439
yes	0.378	0.622

- What does the quantity 0.378 represent?
- If we take the proportions row-wise, does each row need to sum to 1?
- If we take the proportions row-wise, does each column need to sum to 1?

Proportions (cont.)

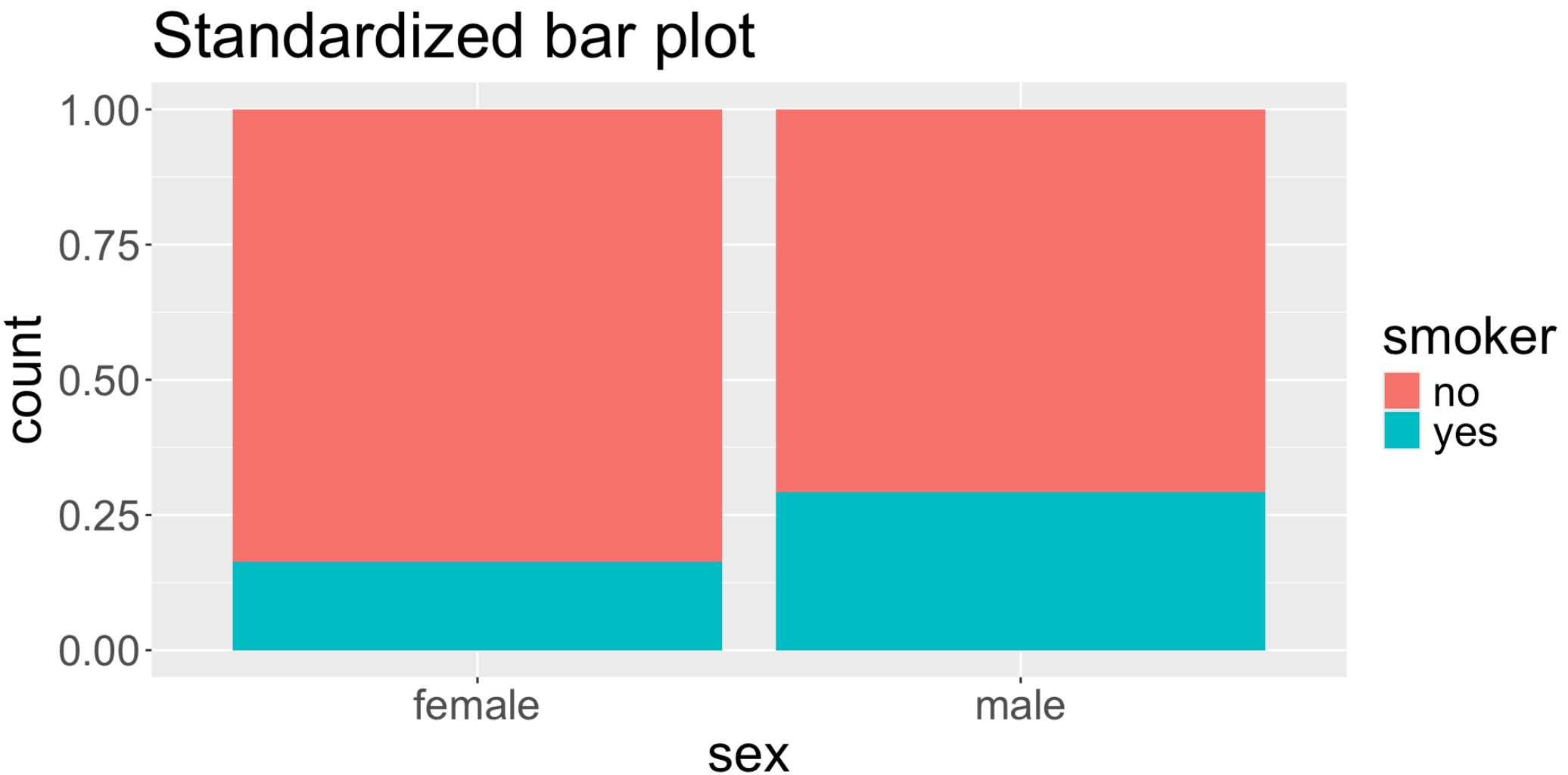
Set up how to find the column-wise proportions using our contingency table

Contingency table

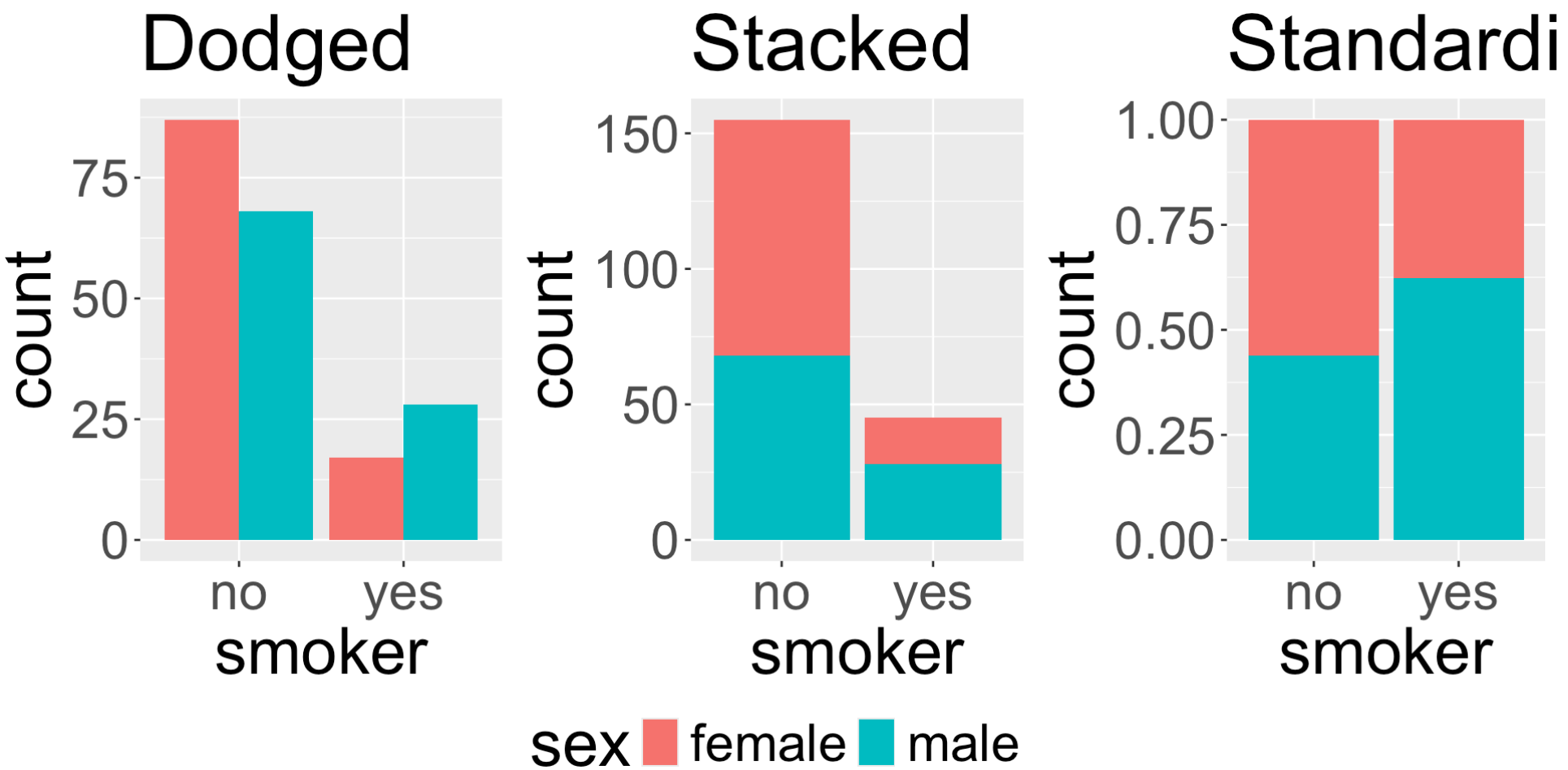
smoker	female	male
no	87	68
yes	17	28

Standardized bar plot

The **standardized bar plot** visualizes these row-wise or column-wise proportions.



Choosing a bar plot



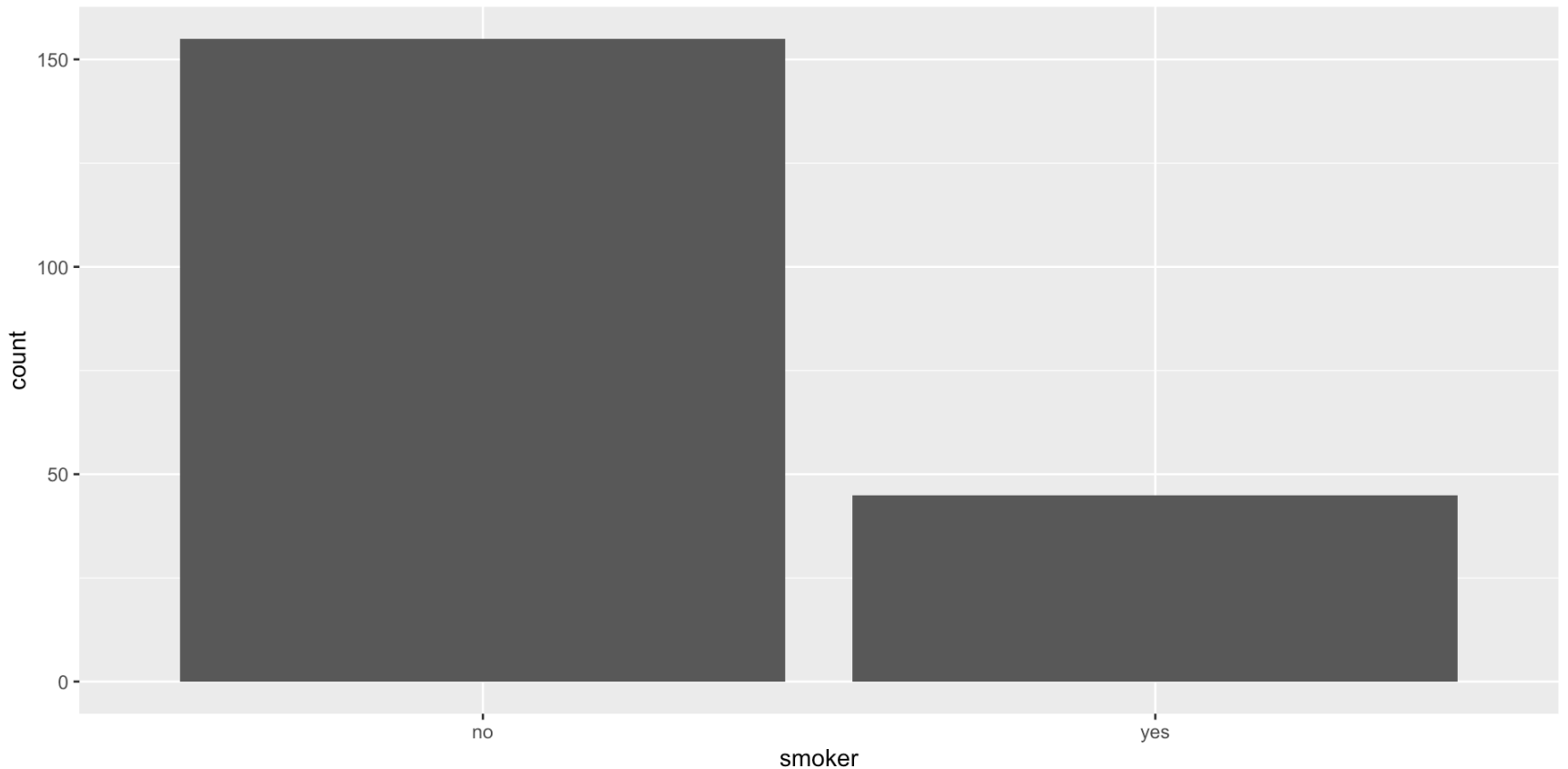
- Using any of the plots, do you believe the smoker status and sex are associated?
- When might you prefer to use the stacked, dodged, or standardized bar plot?

Live code

- Bar plots
- Aesthetics: fill, shape
- Faceting
- Plot background

Bar plot (univariate)

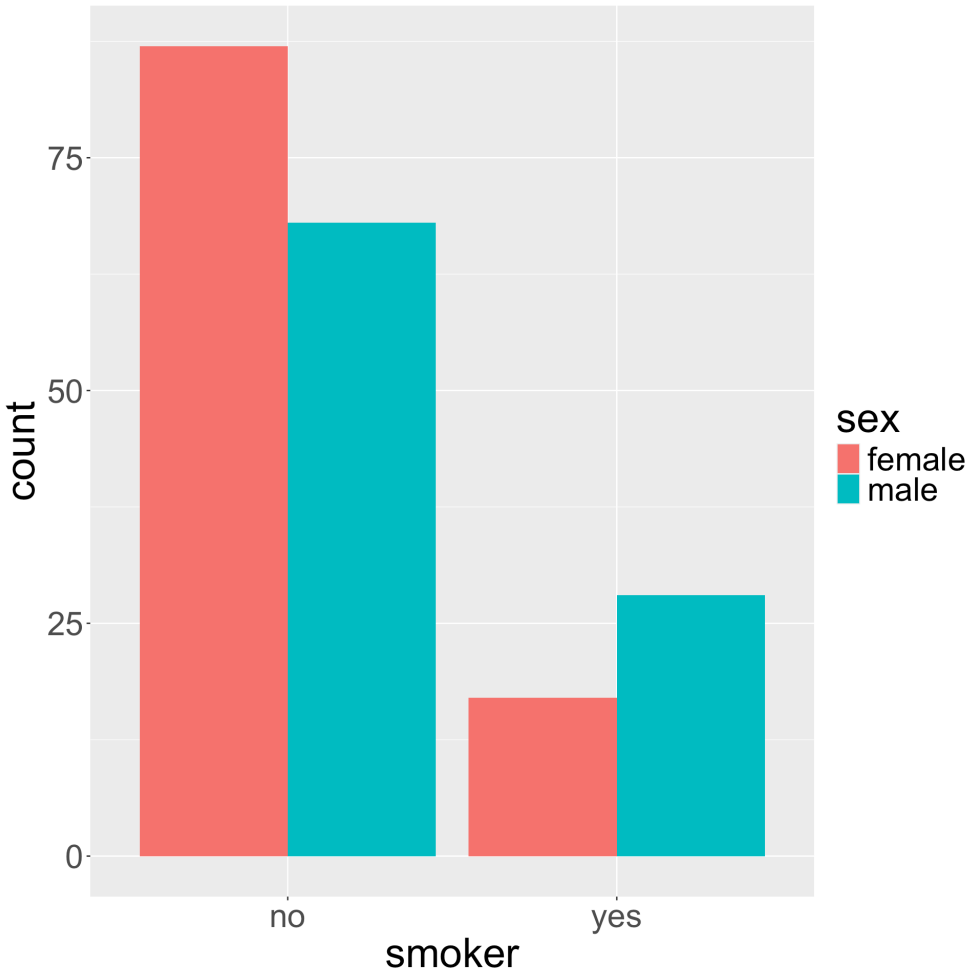
```
1 ggplot(data = insurance, mapping = aes(x = smoker)) +
2   geom_bar()
```



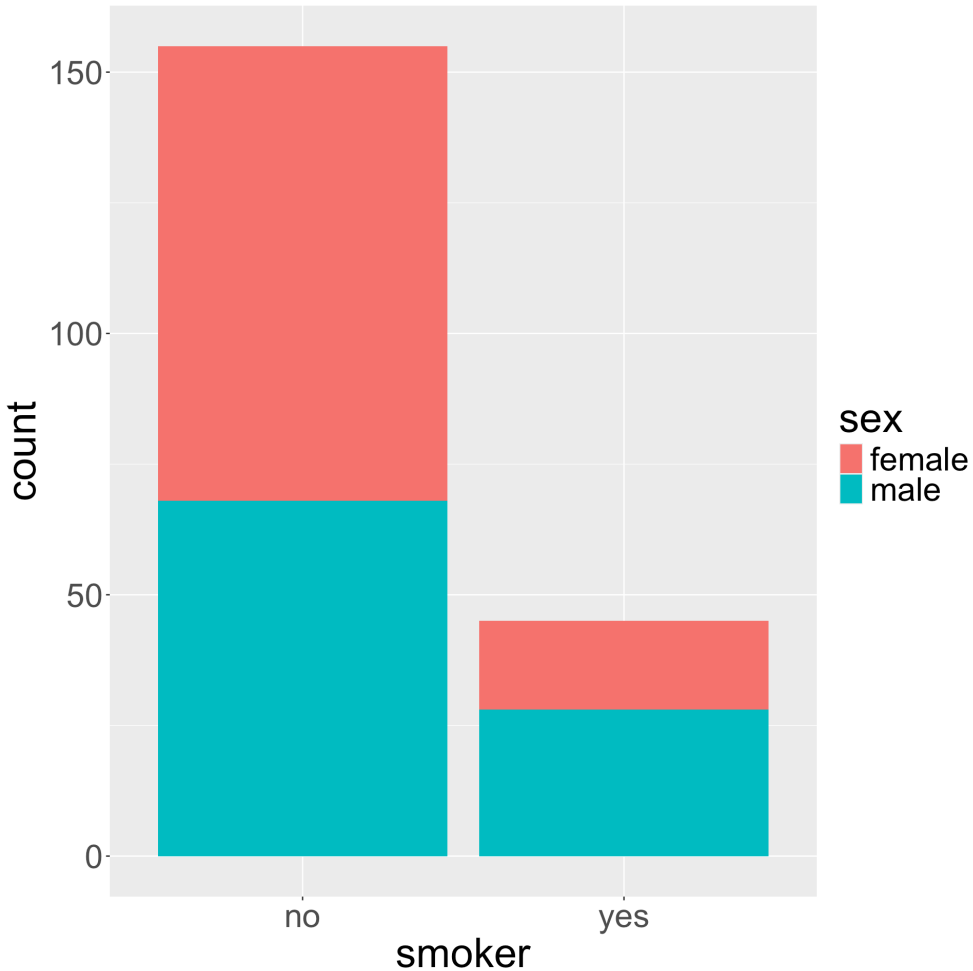
Note: if your data are already in the form of frequency table, we should use `geom_col()` instead!

Bivariate bar plots

```r
ggplot(insurance, aes(x = smoker, fill = sex))
  geom_bar(position = "dodge")
```

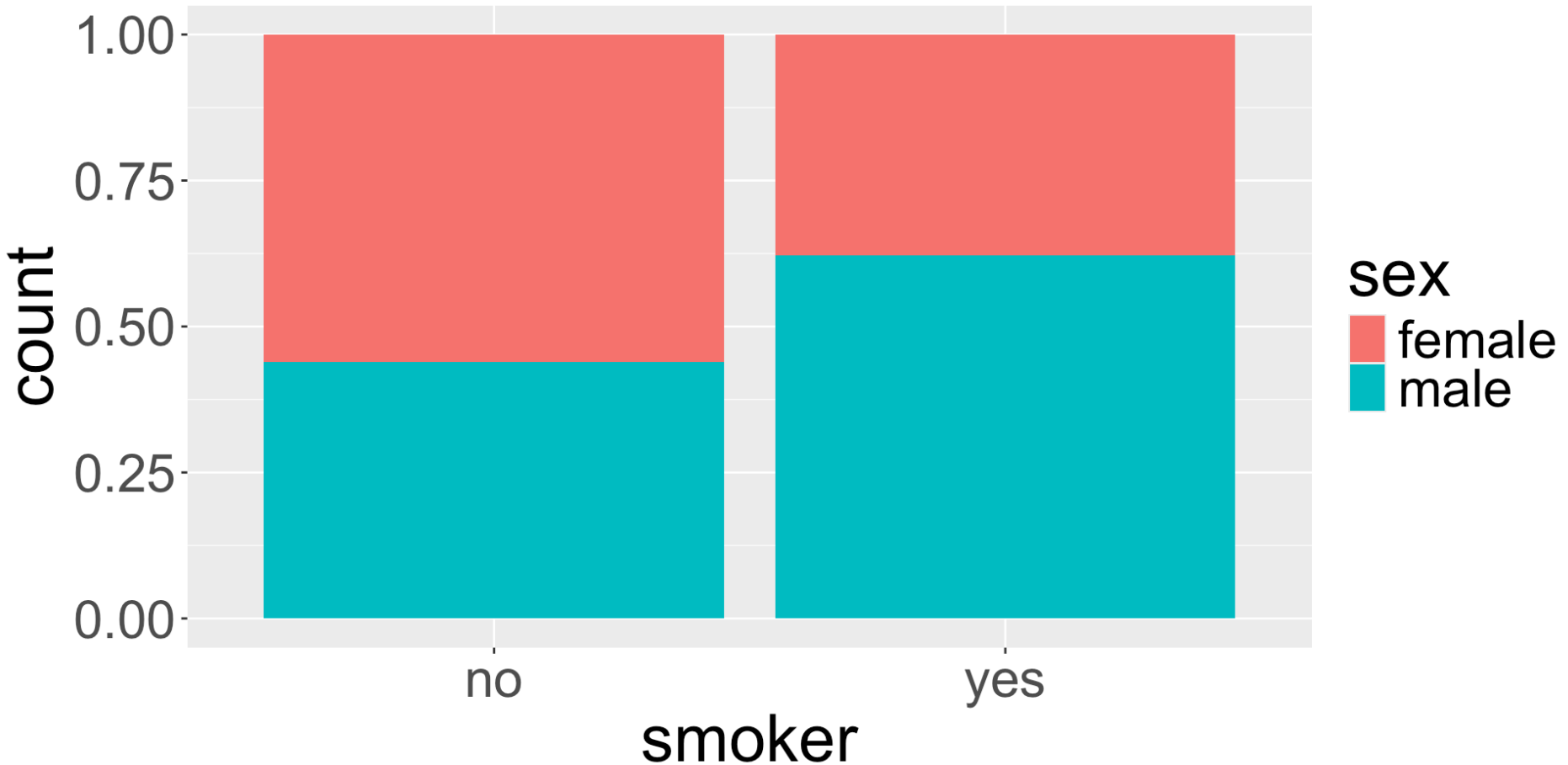



```r
ggplot(insurance, aes(x = smoker, fill = sex))
  geom_bar(position = "stack") # this is defaul
```



Bivariate bar plots (cont.)

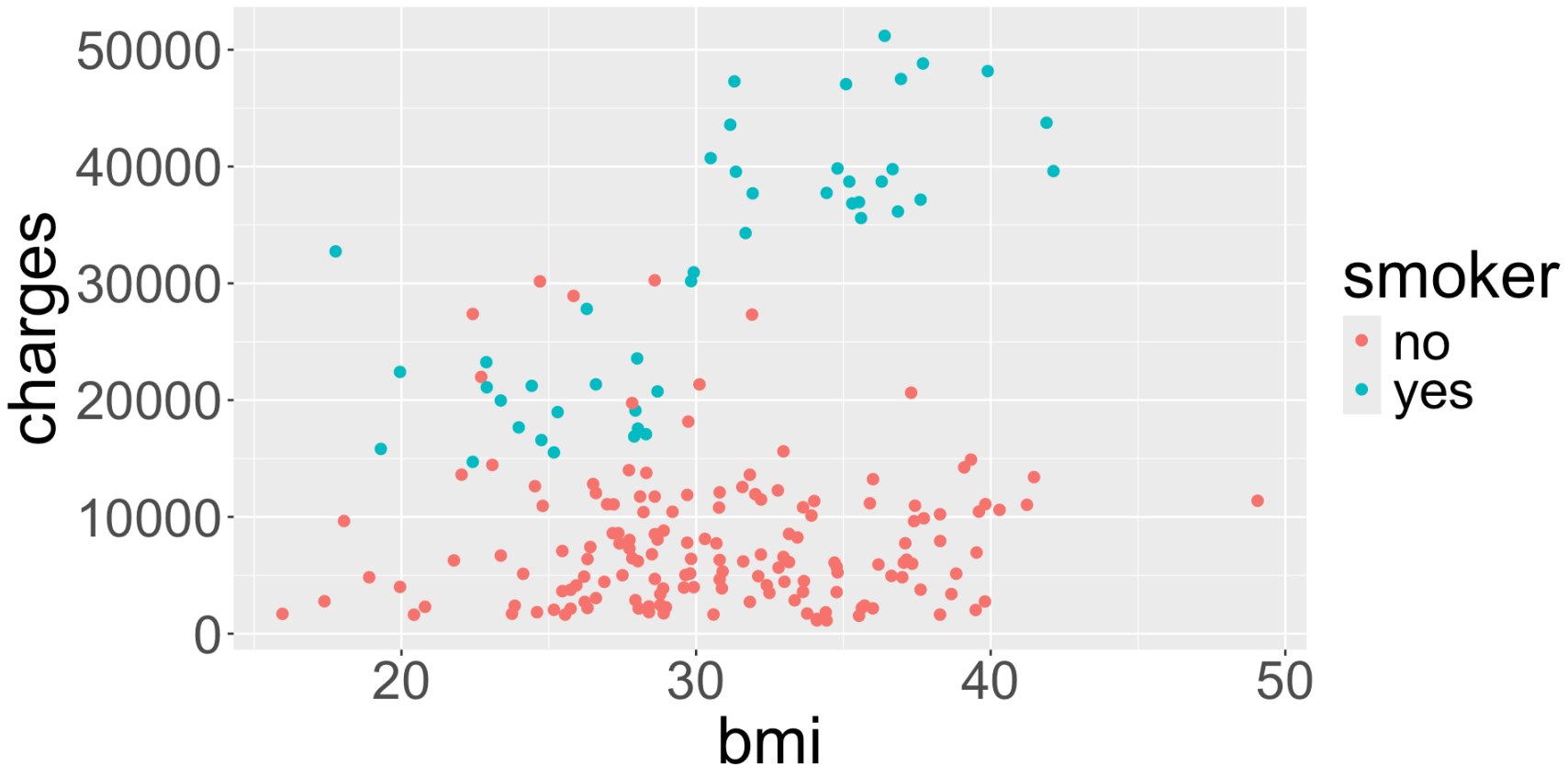
```
1 ggplot(insurance, aes(x = smoker, fill = sex)) +
2   geom_bar(position = "fill")
```



How might we make the bars horizontal instead of vertical?

Visualizing numerical and categorical

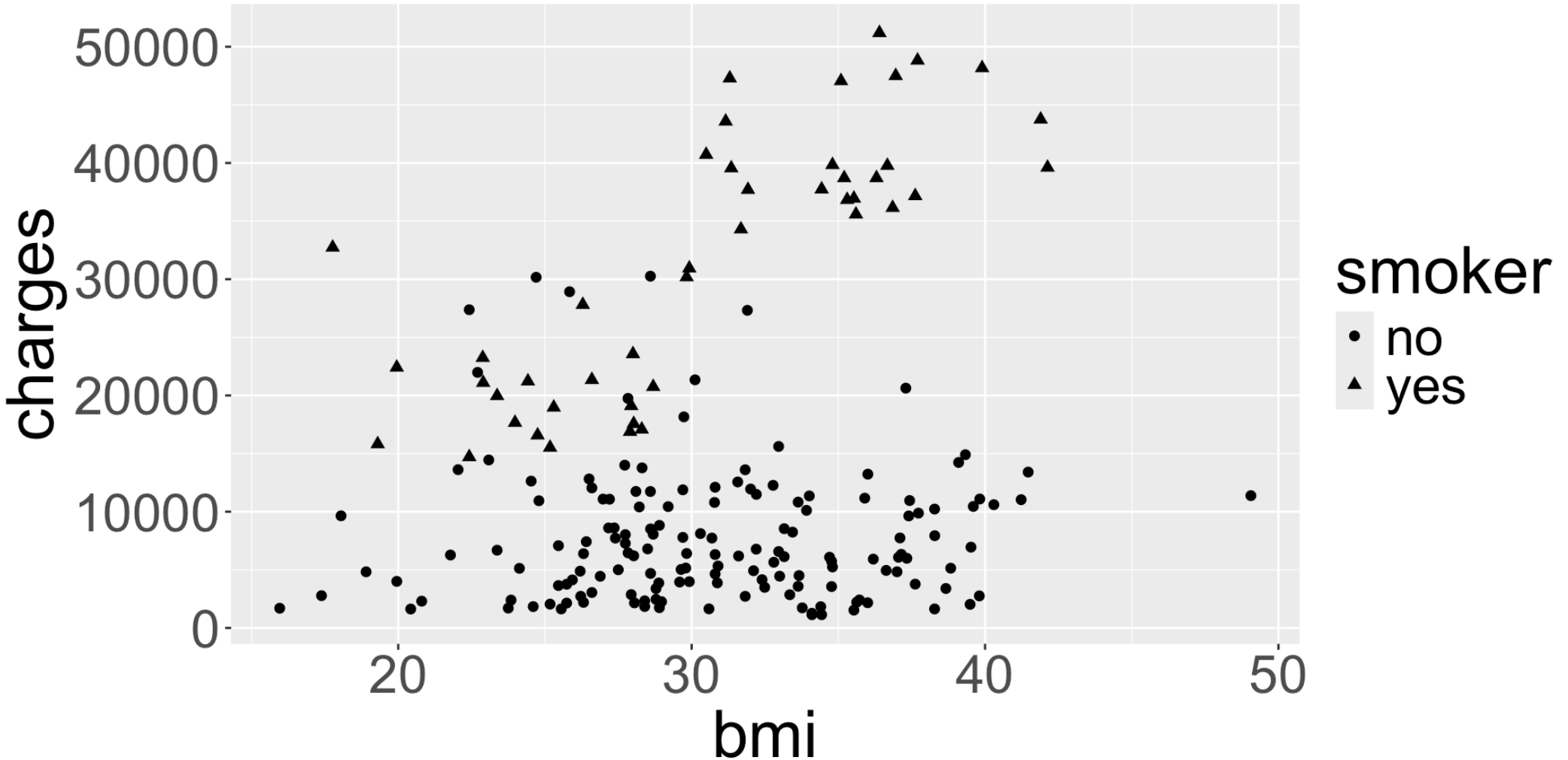
```
1 ggplot(data = insurance, mapping = aes(x = bmi, y = charges, col = smoker)) +
2   geom_point()
```



What do you notice about the legend for color compared to the legend for color from last week?

Aesthetic: shape

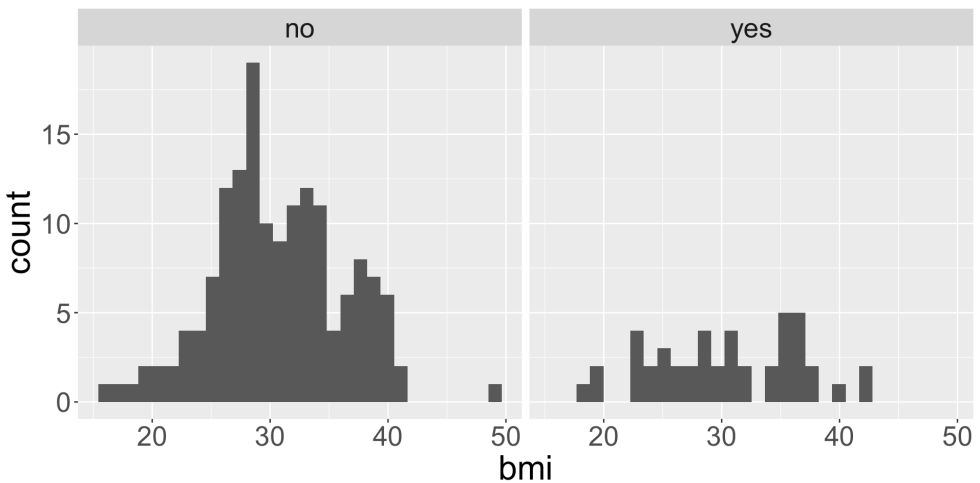
```
1 ggplot(data = insurance, mapping = aes(x = bmi, y = charges, shape = smoker)) +
2   geom_point()
```



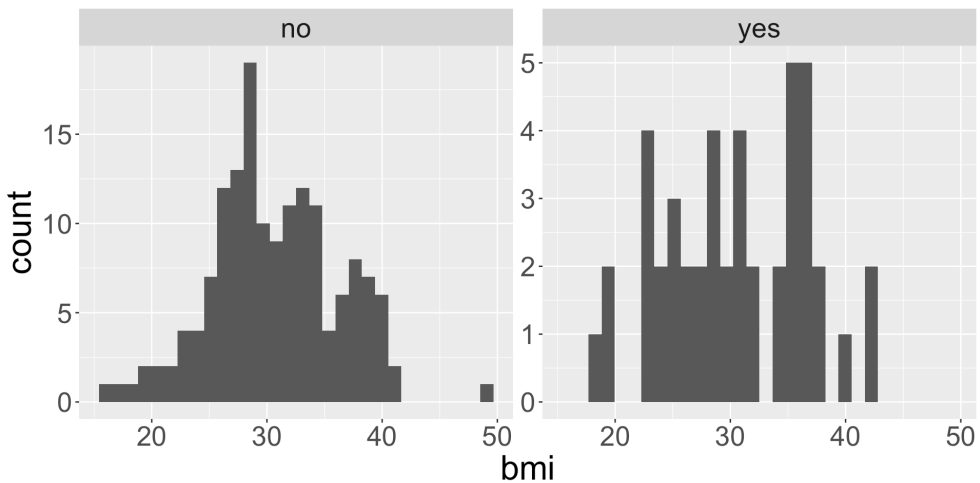
facet_wrap()

Faceting is used when we want to split a particular visualization by the values of another (categorical) variable

```
1 ggplot(data = insurance,
2        mapping = aes(x = bmi)) +
3   geom_histogram() +
4   facet_wrap(~ smoker)
```

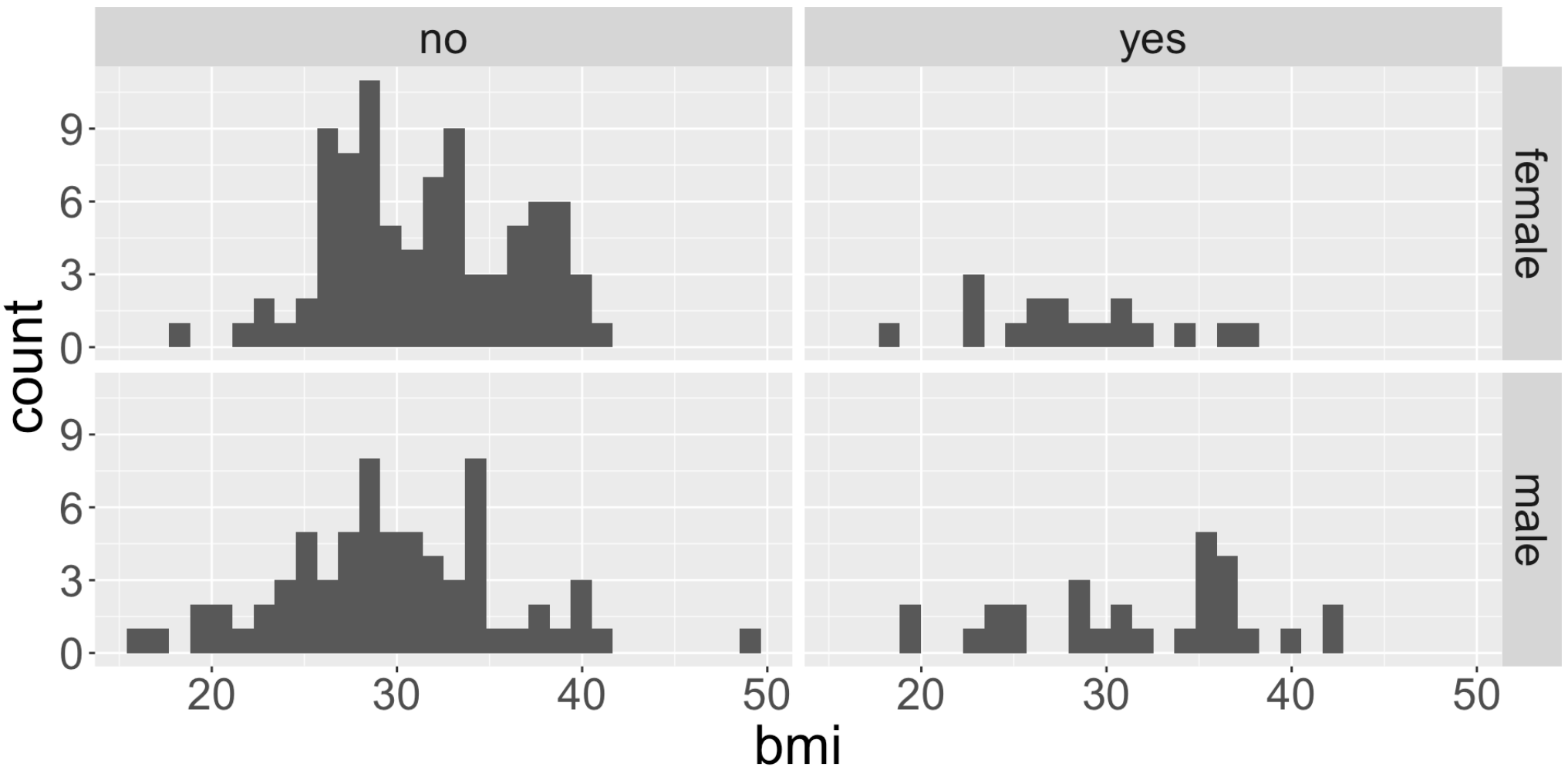



```
1 ggplot(data = insurance,
2        mapping = aes(x = bmi)) +
3   geom_histogram() +
4   facet_wrap(~ smoker, scales = "free_y")
```



facet_grid()

```
1 ggplot(data = insurance, mapping = aes(x = bmi)) +
2   geom_histogram() +
3   facet_grid(sex ~ smoker)
```



Side-by-side box plots

```
ggplot(data = insurance,
       mapping = aes(x = smoker, y = bmi)) +
  geom_boxplot()
```

Like faceting, but only for box plots. Really good for comparing a numerical variable across across a categorical!

Changing plot theme

Change the background of plots by adding on any one of the following:

- `theme_bw()`, `theme_minimal()`, `theme_gray()`, `theme_void()` and a few more (see all options by checking the help file for any one of these)

```
ggplot(data = insurance,
       mapping = aes(x = smoker, y = bmi)) +
  geom_boxplot() +
  theme_minimal()
```