

## Chi-Squared

2. **Act on climate change.** The table below summarizes results from a Pew Research poll which asked respondents whether they have personally taken action to help address climate change within the last year and their generation. The differences in each generational group may be due to chance. Complete the following computations under the null hypothesis of independence between an individual's generation and whether they have personally taken action to help address climate change within the last year. [pewclimatechange2021]

```
climate_change_action <- tribble(
  ~Generation,      ~n,    ~p,
  "Gen Z",          912, 0.32,
  "Millennial",     3160, 0.28,
  "Gen X",          3518, 0.23,
  "Boomer & older", 6074, 0.21
) |>
mutate(
  Generation = fct_relevel(Generation, "Gen Z", "Millennial", "Gen X", "Boomer & older"),
  `Took action` = round(n * p, 0),
  `Didn't take action` = n - `Took action`
) |>
select(-n, -p) |>
pivot_longer(cols = c(`Took action`, `Didn't take action`),
  names_to = "response",
  values_to = "n") |>
uncount(weights = n)

climate_change_action |>
count(Generation, response) |>
pivot_wider(names_from = response, values_from = n) |>
select(Generation, `Took action`, `Didn't take action`) |>
adorn_totals(where = c("row", "col")) |>
kbl(linesep = "", booktabs = TRUE, align = "lrrr", format.args = list(big.mark = ",")) |>
kable_styling(bootstrap_options = c("striped", "condensed"),
  latex_options = "HOLD_position",
  full_width = FALSE) |>
column_spec(1, width = "7em") |>
column_spec(2:4, width = "8em") |>
add_header_above(c(" " = 1, "Response" = 2, " " = 1))
```

Generation	Response		Total
	Took action	Didn't take action	
Gen Z	292	620	912
Millennial	885	2,275	3,160
Gen X	809	2,709	3,518
Boomer & older	1,276	4,798	6,074
Total	3,262	10,402	13,664

If there is no relationship between age and action,

- a. how many Gen Z'ers would you expect to have personally taken action to help address climate change within the last year?
- b. how many Millennials would you expect to have personally taken action to help address climate change within the last year?
- c. how many Gen X'ers would you expect to have personally taken action to help address climate change within the last year?
- d. how many Boomers and older would you expect to have personally taken action to help address climate change within the last year?

4. **Disaggregating Asian American tobacco use, data.** Understanding cultural differences in tobacco use across different demographic groups can lead to improved health care education and treatment. A recent study disaggregated tobacco use across Asian American ethnic groups including Asian-Indian (n = 4,373), Chinese (n = 4,736), and Filipino (n = 4,912), in comparison to non-Hispanic Whites (n = 275,025). The number of current smokers in each group was reported as Asian-Indian (n = 223), Chinese (n = 279), Filipino (n = 609), and non-Hispanic Whites (n = 50,880). [Rao:2021]

In order to assess whether there is a difference in current smoking rates across three Asian American ethnic groups, the observed data is compared to the data that would be expected if there were no association between the variables.

```
asian_smoke <- tibble(
  ethnicity = c(
    rep("Asian-Indian", 4373),
    rep("Chinese", 4736),
    rep("Filipino", 4912)
  ),
  outcome = c(
    rep("smoke", 223), rep("don't smoke", 4150),
    rep("smoke", 279), rep("don't smoke", 4457),
    rep("smoke", 609), rep("don't smoke", 4303)
  )
)

asian_smoke |>
  count(ethnicity, outcome) |>
  pivot_wider(names_from = outcome, values_from = n) |>
  adorn_totals(where = c("row", "col")) |>
  kbl(align = "lrrrr", format.args = list(big.mark = ","), booktabs = TRUE) |>
  kable_styling(bootstrap_options = c("striped", "condensed"),
    latex_options = "HOLD_position",
    full_width = FALSE) |>
  add_header_above(c(" " = 1, "Smoking" = 2, " " = 1)) |>
  column_spec(1, width = "7em") |>
  column_spec(2:4, width = "7em")
```

ethnicity	Smoking		Total
	don't smoke	smoke	
Asian-Indian	4,150	223	4,373
Chinese	4,457	279	4,736
Filipino	4,303	609	4,912
Total	12,910	1,111	14,021

- If the variables on ethnicity and smoking status are independent, estimate the proportion of individuals (total) who smoke?
- Given the overall proportion who smoke, how many of each Asian American ethnicity would you expect to smoke?
- Compare the observed and expected counts. From a first glance, does it seem as though the Asian American ethnicity and choice of smoking may be associated?
- Regardless of your answer to part (c), is it possible to tell from looking only at the expected and observed counts whether the two variables are associated?

6. **Disaggregating Asian American tobacco use, randomize once.** In a study that aims to disaggregate tobacco use across Asian American ethnic groups (Asian-Indian, Chinese, and Filipino, in comparison to non-Hispanic Whites), respondents were asked whether they smoke tobacco or not. [Rao:2021] Then, the data were randomized once, where smoking status was randomly assigned to the participants across different ethnicities. The original data are shown on the left and the results of the randomization is shown on the right.

```
asian_smoke <- tibble(
  ethnicity = c(
    rep("Asian-Indian", 4373),
    rep("Chinese", 4736),
    rep("Filipino", 4912)
  ),
  outcome = c(
    rep("smoke", 223), rep("don't smoke", 4150),
    rep("smoke", 279), rep("don't smoke", 4457),
    rep("smoke", 609), rep("don't smoke", 4303)
  )
)

asian_smoke |>
  count(ethnicity, outcome) |>
  pivot_wider(names_from = outcome, values_from = n) |>
  adorn_totals(where = c("row", "col")) |>
  kbl(align = "lrrrr", format.args = list(big.mark = ","), booktabs = TRUE) |>
  kable_styling(bootstrap_options = c("striped", "condensed"),
    latex_options = "HOLD_position",
    full_width = FALSE) |>
  add_header_above(c(" " = 1, "Smoking" = 2, " " = 1)) |>
  column_spec(1, width = "7em") |>
  column_spec(2:4, width = "3em") |>
  add_header_above(c("Original data" = 4))
```

Original data			
ethnicity	Smoking		Total
	don't smoke	smoke	
Asian-Indian	4,150	223	4,373
Chinese	4,457	279	4,736
Filipino	4,303	609	4,912
Total	12,910	1,111	14,021

```
set.seed(47)
asian_smoke |>
  specify(outcome ~ ethnicity) |>
  hypothesize(null = "independence") |>
  generate(1, type = "permute") |>
  ungroup() |>
  count(ethnicity, outcome) |>
  pivot_wider(names_from = outcome, values_from = n) |>
  adorn_totals(where = c("row", "col")) |>
```

```

kbl(align = "lrrrr", format.args = list(big.mark = ","), booktabs = TRUE) |>
kable_styling(bootstrap_options = c("striped", "condensed"),
              latex_options = "HOLD_position",
              full_width = FALSE) |>
add_header_above(c(" " = 1, "Smoking" = 2, " " = 1)) |>
column_spec(1, width = "7em") |>
column_spec(2:4, width = "3em") |>
add_header_above(c("Randomized data" = 4))

```

Randomized data			
ethnicity	Smoking		Total
	don't smoke	smoke	
Asian-Indian	4,015	358	4,373
Chinese	4,385	351	4,736
Filipino	4,510	402	4,912
Total	12,910	1,111	14,021

Recall that the Chi-squared statistic ( $X^2$ ) measures the difference between the expected and observed counts. Without calculating the actual statistic, report on whether the original data or the randomized data will have a larger Chi-squared statistic. Explain your choice.

16. **Coffee and depression.** Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician- diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption. [Lucas:2011]

```
tribble(
  ~`Clinical depression`, ~`1 cup / week or fewer`, ~`2-6 cups / week`, ~`1 cups / day`, ~`2-3 cups / day`,
  "Yes", 670, "___", 905, 564, 95, 2607,
  "No", 11545, "6,244", 16329, 11726, 2288, 48132,
  "Total", 12215, "6,617", 17234, 12290, 2383, 50739
) |>
kbl(linesep = "", booktabs = TRUE, align = "lrrrrrr", format.args = list(big.mark = ",")) |>
kable_styling(bootstrap_options = c("striped", "condensed"),
  latex_options = "HOLD_position",
  full_width = FALSE) |>
column_spec(1, width = "5em") |>
column_spec(2:6, width = "4em") |>
column_spec(7, width = "4em") |>
add_header_above(c(" " = 1, "Caffeinated coffee consumption" = 5, " " = 1))
```

Clinical depression	Caffeinated coffee consumption					Total
	1 cup / week or fewer	2-6 cups / week	1 cups / day	2-3 cups / day	4 cups / day or more	
Yes	670	___	905	564	95	2,607
No	11,545	6,244	16,329	11,726	2,288	48,132
Total	12,215	6,617	17,234	12,290	2,383	50,739

- What type of test is appropriate for evaluating if there is an association between coffee intake and depression?
- Write the hypotheses for the test you identified in part (a).
- Calculate the overall proportion of women who do and do not suffer from depression.
- Identify the expected count for the empty cell, and calculate the contribution of this cell to the test statistic.
- The test statistic is  $\chi^2 = 20.93$ . What is the p-value?
- What is the conclusion of the hypothesis test?
- One of the authors of this study was quoted on the New York Times as saying it was “too early to recommend that women load up on extra coffee” based on just this study. [news:coffeeDepression] Do you agree with this statement? Explain your reasoning.