Lab 5.B

Bivariate Analysis of Categorical Data

KOMAL BHOSLE

RStudio Link

https://posit.cloud/spaces/603138/content/10173878

Library Calls

```
library(tidyverse)

library(gt)

library(gtsummary)

library(tidymodels)

library(openintro)

library(easystats)

library(gtExtras)

library(DescTools)

library(webshot2)
```

Dichotomous-Dichotomous Association

Data Preparation

Question 1 — In the `malaria` dataset, which is the explanatory variable (EV), and which is the response variable (RV)?

The treatment group is explanatory variable and Outcome of wether the person got infection or not is the response variable.

Question 2 — In the `malaria` dataset, which variable(s) could be considered dichotomous? Explain your rationale.

Both the treatment and outcome groups are dichotomous variables. Dichotomous variables are groups which have binary variables or two values. Treatment group has Vaccine/Placebo, and Outcomes group has Infection and no infection variables. This explains the dichotomous variables.

Question 3 — Why must dichotomous-dichotomous contingency tables have the same level names between the variables?

The contingency table variable names have same level names as it helps in clear interpretation of relationship between two variables.

Question 4 — What must you do to prepare `malaria` for a 2 x 2 contingency table?

First to prepare malaria for contingency table - Organize the columns as EV and then RV, and identify their orientation with contingency table.

```r
malaria_clean <-
  malaria |>
  rename(
    "vaccinated" = treatment,
    "infected" = outcome
  ) |>
  mutate(
    vaccinated = recode(
      vaccinated,
      "vaccine" = "Yes",
      "placebo" = "No"
    ),
    vaccinated = factor(
      vaccinated,
      levels = c("Yes", "No")
    ),
    infected = recode(
      infected,
      "infection" = "Yes",
      "no infection" = "No"
    ),
    infected = factor(
      infected,
      levels = c("Yes", "No")
    )
  ) |>
  select(
    vaccinated,
    infected
  )
```

Quick Crosstab

Question 5 — Before producing a quick crosstabulation, what characteristics would indicate that `malaria_clean` was properly prepared for a dichotomous-dichotomous contingency table?

The malaria_clean is now in a EV, RV format, which has two columns one for EV and one for RV.

```
table(
  malaria_clean
)
```

```
          infected
vaccinated Yes No
       Yes   5  9
       No    6  0
```

Contingency Table

Question 6 — Why use the `label` argument in `tbl_cross()`?

label argument allows to change variable names in the output without changing it in the dataset.

Question 7 — What does the `md()` function do?

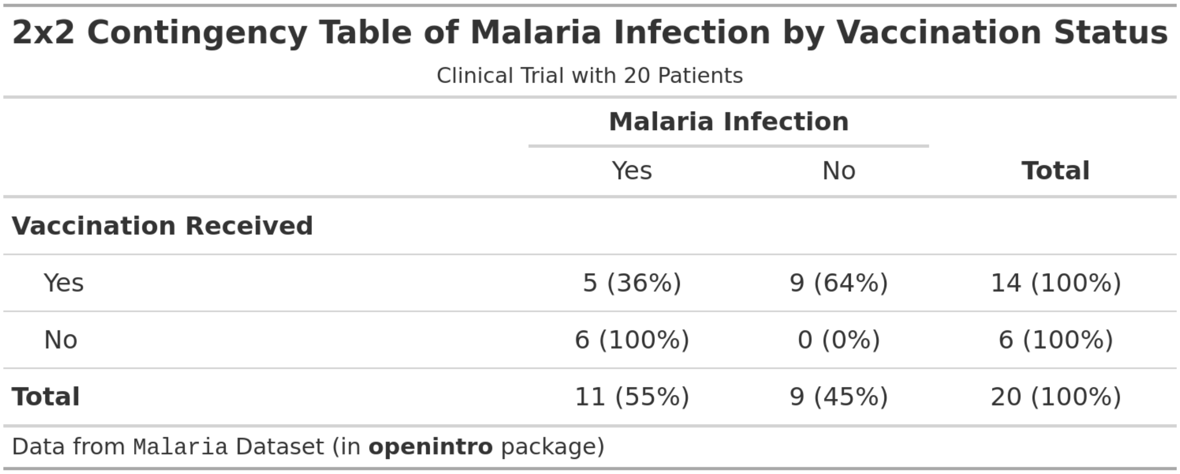
md() - markdown function allows to change the format of the text in tables to make it visually different.

Question 8 — Why must an object created with `gt_summary` be converted to a `gt` object before use of functions from the `gt` package? What function converts output to a `gt` object?

Objects created by gt_summary need to be converted to gt in order to be able to use gt function for customization. This can be possible only when gt_summary is converted to gt object. The function that converts gt_summary to gt is as_gt().

```
malaria_clean_tbl_cross <-
  malaria_clean |>
  tbl_cross(
    row = vaccinated,
    col = infected,
    percent = "row",
    label = list(
      vaccinated ~ "Vaccination Received",
      infected ~ "Malaria Infection"
    )
  ) |>
  bold_labels() |>
  as_gt() |>
  tab_source_note(
    source_note = md(
      "Data from `Malaria` Dataset (in **openintro** package)"
    )
  ) |>
  tab_header(
    title = md(
      "**2x2 Contingency Table of Malaria Infection by Vaccination Status**"
    ),
    subtitle = md(
      "Clinical Trial with 20 Patients"
    )
  )
```

```
gtsave(
  malaria_clean_tbl_cross,
  filename = "malaria_clean_tbl_cross.png")
```



2x2 Contingency Table of Malaria Infection by Vaccination Status

Clinical Trial with 20 Patients

	Malaria Infection		
	Yes	No	**Total**
Vaccination Received			
Yes	5 (36%)	9 (64%)	14 (100%)
No	6 (100%)	0 (0%)	6 (100%)
Total	11 (55%)	9 (45%)	20 (100%)

Data from `Malaria` Dataset (in **openintro** package)

Question 9 — Which argument makes it possible to analyze the previous contingency table using Percent Maximum Difference (PMD)?

the percentage = "row" argument analyzes the contingency table by calculating the percentages in each row of the table of explanatory varible.

Question 10 — Conduct PMD in the console (or by eyesight). Summarize your findings. Are these findings conclusive (why or why not)?

The PMD is 64% between the two groups. Which shows that vaccination is associated with decreased malaria infection. This can be seen where people who did not receive vaccination has 100% malaria infection where as majority population who received vaccination were not infected by malaria and that is 64%.

Odds Ratio (OR)

Question 11 — How is the odds ratio (OR) a "ratio of ratios"? What ratios are being compared?

Odds is a ratio and odds ratio is the ratio of odds. Thus odds ratio is the ratio of ratios. The ratios being compared are likelihood of concordance and discordance.

Question 12 — How do likelihood ratios (LR) reflect concordance and discordance between the EV and RV? Define the two LRs used to calculate OR.

LRs reflect the occurance of outcome between two groups. OR is the ratio of two LRs. Together they reveal the strength between EV and RV.

Question 13 — Explain what it means for OR to be the ratio of LR$_+$ to LR$_-$. How do the inherent identities of LR$_+$ and LR$_-$ determine the interpretation of OR?

Replace this text with your response.

Question 14 — Why should continuity correction be applied before calculating the OR with rare events? Would this adjustment be needed for the `malaria_clean` joint frequencies? If so, how would this be done using the `oddsratio()` function from the `effectsize` package?

Replace this text with your response.

```r
oddsratio_result <-
  oddsratio(
    x = malaria_clean$vaccinated,
    y = malaria_clean$infected
  ) + 0.5
```

Question 15 — How can you determine the existence of an association using the odds ratio (OR)? For `malaria_clean`, what does the OR suggest about the existence of an association between EV and RV?

Replace this text with your response.

Question 16 — How can you determine the directionality of an association using the odds ratio (OR)? For `malaria_clean`, what does the OR suggest about the directionality of the association between EV and RV?

Replace this text with your response.

Question 17 — How can you determine the strength of an association using the odds ratio (OR)?

Replace this text with your response.

Question 18 — How should you interpret and report the following OR values: 1.5, 2.5, 0.5, and 0.0?

Replace this text with your response.

Question 19 — For `malaria_clean`, what does the OR suggest about the strength of the association between EV and RV? Phrase your response in a conventional way for reporting an odds ratio?

Replace this text with your response.

OR Variants

Yule’s Q

Question 20 — Why is Yule’s Q considered a variant of the odds ratio (OR)?

Replace this text with your response.

```r
YuleQ(
  x = malaria_clean$vaccinated,
  y = malaria_clean$infected
) |>
  round(
    digits = 2
  )
```

```
[1] -1
```

Question 21 — What does Yule’s Q suggest about the relationship between malaria vaccination and malaria infection?

Replace this text with your response.

Yule’s Y

Question 22 — How does Yule’s Y adjust Yule’s Q?

Replace this text with your response.

```r
YuleY(
  x = malaria_clean$vaccinated,
  y = malaria_clean$infected
) |>
  round(
    digit = 2
  )
```

```
[1] -1
```

Question 23 — How does Yule’s Y insights differ from Yule’s Q insights? Does this difference reasonably suggest the presence of extreme values?

Replace this text with your response.

Ordinal-Ordinal Association

Data Preparation

```r
airquality_clean <-
  airquality |>
  as_tibble() |>
  select(
    Solar.R,
    Temp
  ) |>
  rename(
    solar_radiation = Solar.R,
    temperature = Temp
  ) |>
  drop_na() |>
  mutate(
    solar_radiation = case_when(
      solar_radiation < 115 ~ "Low Solar Radiation",
      between(solar_radiation, 115, 258) ~ "Moderate Solar Radiation",
      solar_radiation > 258 ~ "High Solar Radiation"
    ),
    solar_radiation =
      factor(
        x = solar_radiation,
        levels = c(
          "Low Solar Radiation",
          "Moderate Solar Radiation",
          "High Solar Radiation"
        )
      ),
    temperature = case_when(
      temperature < 72 ~ "Low Temperature",
      between(temperature, 72, 85) ~ "Moderate Temperature",
      temperature > 85 ~ "High Temperature"
    ),
    temperature =
      factor(
        x = temperature,
        levels = c(
          "Low Temperature",
          "Moderate Temperature",
          "High Temperature"
        )
      )
  )
```

Quick Crosstab

```
table(
  airquality_clean
)
```

```
                          temperature
solar_radiation            Low Temperature Moderate Temperature
  Low Solar Radiation                   16                   18
  Moderate Solar Radiation               9                   44
  High Solar Radiation                   8                   19
                          temperature
solar_radiation            High Temperature
  Low Solar Radiation                     2
  Moderate Solar Radiation               20
  High Solar Radiation                   10
```

R x C Contingency Table

```
## Air Quality

#| label : airquality
#| fig-cap :  "Two-Way Frequency Table of Temperature Level by Solar Radiation Level"
#| warning : False

airquality_clean_tabel <-
  airquality_clean |>
  tbl_cross(
    row = solar_radiation,
    col = temperature,
    percent = "row",
    label = list(
      solar_radiation ~ "Solar Radiation Level (by Langleys)",
      temperature ~ "Temperature Level (by Degrees Fahrenheit)"
    )
  ) |>
  bold_labels()

airquality_clean_tabel |>
  as_gt() |>
    tab_caption(
      "Two-Way Frequency Table of Temperature Level by Solar Radiation Level"
    )
```

	Temperature Level (by Degrees Fahrenheit)		
	Low Temperature	Moderate Temperature	High Temperature
Solar Radiation Level (by Langleys)			
Low Solar Radiation	16 (44%)	18 (50%)	2 (5.6%)
Moderate Solar Radiation	9 (12%)	44 (60%)	20 (27%)
High Solar Radiation	8 (22%)	19 (51%)	10 (27%)
Total	33 (23%)	81 (55%)	32 (22%)

Percent Maximum Difference (PMD)

Question 24 — Using the console for calculations, what are the percent differences between the maximum and minimum percentage across each row (i.e., EV category)? What is the percent maximum difference (PMD)?

Replace this text with your response.

Question 25 — Interpret the PMD (and PDs) given the context of the data.

Replace this text with your response.

Heatmap

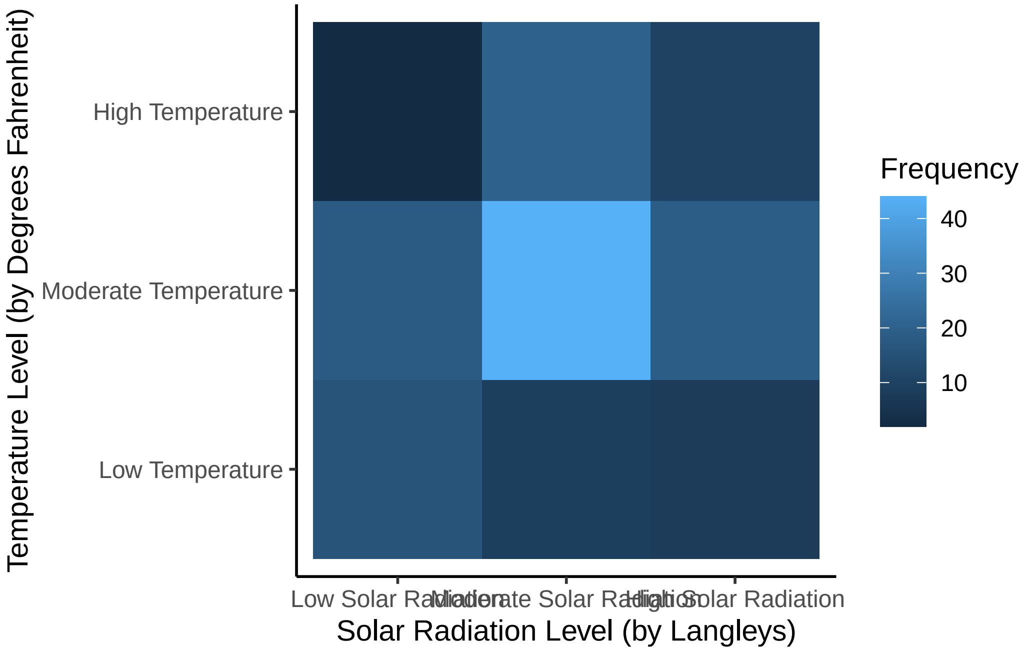
Question 26 — What insights could a heatmap provide that may not be evident within a calculation?

Replace this text with your response.

```r
#|label : airquality heatmap
#|fig-cap : "Heatmap of Temperature Level by Solar Radiation Level"
#|warning : false

airquality_frequency <-
  airquality_clean |>
  group_by(
    solar_radiation,
    temperature
  ) |>
  summarise(
    n = n(),
    .groups = "drop"
  )

airquality_frequency |>
  ggplot(
    mapping = aes(
      x = solar_radiation,
      y = temperature,
      fill = n,
      group = interaction(
        solar_radiation,
        temperature
      )
    )
  )+
    geom_tile() +
    labs(
      x = "Solar Radiation Level (by Langleys)",
      y = "Temperature Level (by Degrees Fahrenheit)",
      fill = "Frequency"
    ) +
  theme_classic()
```



Question 27 — What insights can you obtain from the heatmap depicting intersecting frequencies of solar radiation and temperature?

Replace this text with your response.

Goodman and Kruskal's Gamma

Question 28 — What types of contingency tables can be used with Goodman and Kruskal's Gamma?

Replace this text with your response.

```r
airquality_clean_gamma <-
  GoodmanKruskalGamma(
    x = table(airquality_clean),
    conf.level = 0.95
  ) |>
  round(
    digits = 2
  )
```

Question 29 — Interpret your output from using the `GoodmanKruskalGamma()` function.

Replace this text with your response.

Question 30 — What would you need to look for to determine whether or not the association detected by Gamma is statistically significant?

Replace this text with your response.

Nominal-Nominal Association

Data Preparation

```
#|label : acs12-employment
#|fig-cap : "Two-Way Frequency Table of Employment Status by Race"
#|warning : false

acs12_clean <-
  acs12 |>
  select(
    race,
    employment
  ) |>
  drop_na() |>
  mutate(
    race = str_to_title(race),
    race = factor(
      race
    ),
    employment = recode(
      employment,
      "employed" = "Employed",
      "unemployed" = "Unemployed",
      "not in labor force" = "Not Employed"
    ),
    employment = factor(
      employment
    ),
  )
```

Quick Crosstab

```r
table(
  acs12$race,
  acs12$employment
)
```

```
        not in labor force unemployed employed
  white                520         72      670
  black                 66         20       76
  asian                 31          3       39
  other                 39         11       58
```

Contingency Table

```r
acs12_clean |>
  tbl_cross(
    row = race,
    col = employment,
    percent = "row",
    label = list(
      race ~ "Race",
      employment ~ "Employment Status"
    )
  ) |>
  bold_labels() |>
  italicize_levels()
```

	Not Employed	*Unemployed*	*Employed*	**Total**
Race				
Asian	31 (42%)	3 (4.1%)	39 (53%)	73 (100%)
Black	66 (41%)	20 (12%)	76 (47%)	162 (100%)
Other	39 (36%)	11 (10%)	58 (54%)	108 (100%)
White	520 (41%)	72 (5.7%)	670 (53%)	1,262 (100%)
Total	656 (41%)	106 (6.6%)	843 (53%)	1,605 (100%)

Cramer's V

Question 31 — What types of contingency tables can be used with Cramer's V?

Replace this text with your response.

```
acs12_cramer_v <-
  acs12_clean |>
  table() |>
  CramerV(
    conf.level = 0.95
  ) |>
  round(
    digits = 2
  )
```

Question 32 — Interpret your output from using the `CramerV()` function.

Replace this text with your response.

Question 33 — What would you need to look for to determine whether or not the association detected by Cramer's V is statistically significant?

Replace this text with your response.

Grouped Bar Plot

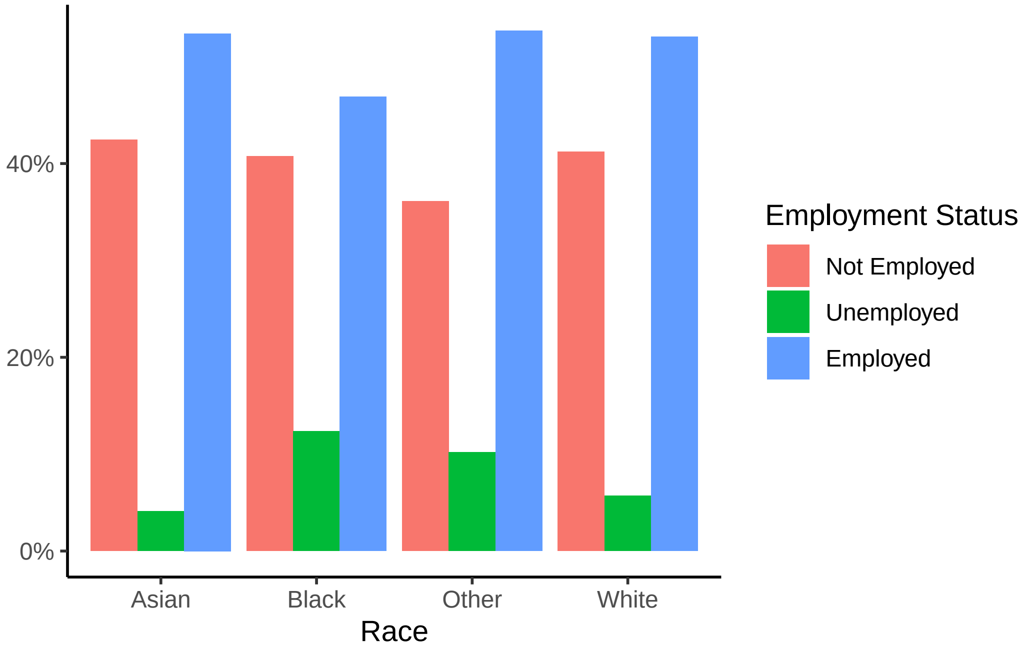
Question 34 — If analyzing employment status based upon race, why would plotting frequencies within a grouped bar plot lack key insights?

Replace this text with your response.

Question 35 — How does a grouped bar plot constructed with percentages offer better insights than the same plot built upon frequencies?

Replace this text with your response.

```
acs12_clean |>
  group_by(
    race,
    employment
  ) |>
  summarise(
    n = n(),
    .groups = "drop"
  ) |>
  group_by(
    race
  ) |>
  mutate(
    percentage = n / sum(n) * 100
  ) |>
  ggplot(
    mapping = aes(
      x = race,
      y = percentage,
      fill = employment
    )
  ) +
  geom_col(
    position = "dodge"
  ) +
  labs(
    x = "Race",
    y = NULL,
    fill = "Employment Status"
  ) +
  scale_y_continuous(
    labels = function(x) paste0(x, "%")
  ) +
  theme_classic()
```



Question 36 — Interpret the insights gained from the grouped bar plot.

Replace this text with your response.