COMPARE PROPORTIONS

- *packages*: **magrittr, pacman, survival, tidyverse**
- *dataset:* **survival::lung**
- chi-square test
  - chi-sq test for sex & status
- proportions test
  - get survival proportions by sex
  - proportions test **`propr.test()`**
  - alternate hypothesis
- many proportions
  - method for comparing many proportions)

OVERVIEW

- basic analysis = looking at proportions
  - specifically: what percentage of people say ‘yes’ in this group
- depending on your data (frequencies, 2x2 table, several proportions all at once)
  - bw the chi-square test & prop.test function - you can get what you need out of it

INSTALL AND LOAD PACKAGES

```
pacman::p_load (magrittr, pacman, survival, tidyverse)
```

- **survival** - sample dataset

LOAD AND PREPARE DATA

```
?lung   # info on “NCCTG” Lung Cancer Data” from survival
```

- about people with lung cancer / how long they survived / their sex / severity of cancer
- > look at status: 1 = censored (alive), 2 = dead / sex: male = 1, female = 2

Select 2 variables & Save data as df (data frame):

```
df %>% lung %>%
    select(status, sex) %>%
    as_tibble() %>%
    print()
```

- result: tibble 228 x 2
  status <*dbl*> / sex <*dbl*>

```
# A tibble: 228 × 2
   status   sex
    <dbl> <dbl>
 1      2     1
 2      2     1
 3      1     1
 4      2     1
 5      2     1
 6      1     1
 7      2     2
 8      2     2
 9      2     1
10      2     1
# ℹ 218 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Recode ‘sex’ and ‘status’ from numeric to names:

```
df %>%
    mutate(
       status = ifelse(status  == 1, “alive”, “dead”),
       sex = ifelse(sex == 1, “male”, “female”)
    ) %>%
```

- result: status *<chr>* / sex *<chr>*

```
# A tibble: 228 × 2
   status sex
   <chr>  <chr>
 1 dead   male
 2 dead   male
 3 alive  male
 4 dead   male
 5 dead   male
 6 alive  male
 7 dead   female
 8 dead   female
 9 dead   male
10 dead   male
# ℹ 218 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Create frequency table, save for reuse:

```
ptable <- df  %>%              # save table for reuse
  select(sex, status) %>%      # variables in table
  table() %>%                  # create 2 x 2 table
   print()                     # show table
```

- result: environment - values: `ptable` ‘table’ int[1:2, 1”2] 37 26 …
    - these are frequencies

```
        status
sex      alive dead
  female    37   53
  male      26  112
```

CHI-SQUARED TEST

- inferential test

Get chi-squared test for sex and status:

```
ptable %>% chisq.test()
```

- result:

Pearson’s Chi-squared test with Yates’ continuity correction
data:
X-squared = 12.42 | df = 1 (bc 2 x 2 table)
p-value = 0.0004247 *(definitely below standard cutoff of .05)*

```
> ptable %>% chisq.test()

	Pearson's Chi-squared test with Yates' continuity correction

data:  .
X-squared = 12.42, df = 1, p-value = 0.0004247
```

  - > this is a statistically significant difference - Lets us know that survival and sex , in this particular dataset, operate together - there is a connection between the two. (they are NOT independent)

PROPORTIONS TEST

Get survival proportions by sex:

```
df %>%
    group_by(sex, status) %>%   # variables to group by
    summarize(n = n()) %>%          # calculate n for each group
    mutate(freq = n / sum(n))       # proportions by sex
```

- results: tibble 4 x 4

groups: sex [2]

sex <chr>	status <chr>	n <int>	freq <dbl>
female	alive	37	0.411
female	dead	53	0.589
male	alive	26	0.188
male	dead	112	0.812

```
`summarise()` has grouped output by 'sex'. You can override using the `.groups`
argument.
# A tibble: 4 × 4
# Groups:   sex [2]
  sex    status     n  freq
  <chr>  <chr>  <int> <dbl>
1 female alive     37 0.411
2 female dead      53 0.589
3 male   alive     26 0.188
4 male   dead     112 0.812
```

- Of the female observations, 41% are still alive, and 58.9% had died; where as for the male observations, 18.8% and 81.2% had died. Looks like dramatic differences in proportions between males and females.

Proportions test: `propr.test()`

- `prop.test()` – quick insight into the differences , and proportions/ percentages bw 2 groups - prop.test() give us that info

```
ptable %>% prop.test()
```

- takes table that consists of the frequencies - use the prop.test function
- results:

-gives proportions: prop1 (females) 0.4111 / prop2 (male) 0.884 still alive
-does chi-square test - get same probability values
-95% CI on difference bw the 2 groups is between 9% and 35%, based on the variability of the data: 0.092 0.352

```
> # Proportions test
> ptable %>% prop.test()

	2-sample test for equality of proportions with continuity correction

data:  .
X-squared = 12.42, df = 1, p-value = 0.0004247
alternative hypothesis: two.sided
95 percent confidence interval:
 0.09273771 0.35267292
sample estimates:
   prop 1    prop 2 
0.4111111 0.1884058 
```

Alternative Hypothesis:

- Is survival greater for female patients than for male patients? (with 80% CI)

```
ptable %>%
    prop.test(
        alt = “greater”,          # specify directional hypothesis
       conf.level = .80   # specify 80% confidence interval
   )
```

- Results:

p-value half of what was before (p-value = 00002123)
CI changes 0.161 1.000 (ranges from 16% up to 100%) bc one-sided test
proportions = still the same

```
	2-sample test for equality of proportions with continuity correction

data:  .
X-squared = 12.42, df = 1, p-value = 0.0002123
alternative hypothesis: greater
80 percent confidence interval:
 0.1616591 1.0000000
sample estimates:
   prop 1    prop 2 
0.4111111 0.1884058 
```

MANY PROPORTIONS

Method for comparing many proportions:

```
tibble(                           # create new tibble
  n = c(rep(100,5)),              # 100 trials 5 times
 #n = c(100, 100, 100, 100, 100)  # or this way
  x - seq(65, 45, by = -5)        # number of successes
 #x = c(65, 60, 55, 50, 45)      # or this way
 ) %$%                            # exposition pipe
  prop.test(x, n)                 # proportion test
```

- create data: make tibble:
    - make number of trials `(n)`: (the denominator in the proportion) > make 5 groups, each with denominator of 100) n = c(rep(100,5))
    - `(x)` number of successes: proportion that actually looking at
    - exposition pipe - to turn tibble into vector (prop.test needs vector)
    - `prop.test(successes, number of trials)`
- results: 5 dift proportions that correspond

5-sample test for equality of proportions without continuity correction
data: x out of n
X-squared = 10.101, **df** = 4, p-value = 0.03876 (less than standard cutoff of 0.5 - tf is statistically significance difference in this table of proportions)
alternative hypothesis: two.sided
sample estimates:

prop1	prop2	prop3	prop4	prop5
0.65	0.60	0.55	0.50	0.45

```
	5-sample test for equality of proportions without continuity correction

data:  x out of n
X-squared = 10.101, df = 4, p-value = 0.03876
alternative hypothesis: two.sided
sample estimates:
prop 1 prop 2 prop 3 prop 4 prop 5 
  0.65   0.60   0.55   0.50   0.45 
```