MOTION Lab Summer 2024 Code Club
Session IV: Intro to Statistical Analysis in R

September 06, 2024

Presenter:

Leigh Cressman (crel@pennmedicine.upenn.edu)

Agenda

1) Descriptive analysis
2) Bivariable analysis
3) Multivariable analysis

IMPALA Study

- Conducted at 6 different hospitals and clinics in Botswana in 2020
- Outcome is colonization with extended-spectrum cephalosporin-resistant enterobacterales (ESCrE)
- The dataset included here is not actual data from the study. All values have been randomly generated. The data used in the in-person session were data from the actual study.

Descriptive analysis

Objective: characterize the population
1) Measures of central tendency
2) Measures of variability
3) Measures of frequency distribution

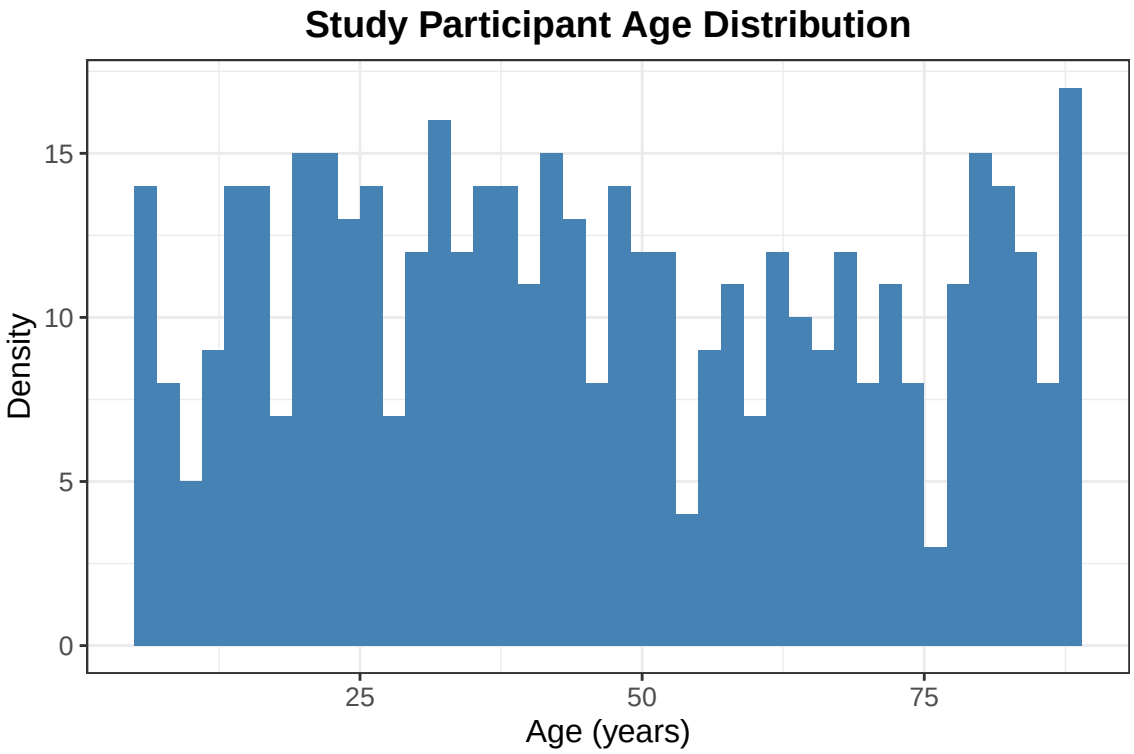
Use psych package to generate descriptive statistics of numeric variables.

Read more about psych::describe.

Summary statistics for numeric variables
Can add more formatting.

	n	min	Q0.25	median	Q0.75	max	mean	sd
participant_age	469	5	26	44	68	89	46.759062	24.282075
total_abx_days	469	0	3	7	10	13	6.477612	4.110393
days_to_sample	469	0	4	8	12	15	7.520256	4.593999

Check distribution of age



Run Shapiro-Wilk Normality Test
W statistic close to 1 indicates data are normally distributed.
P-value < 0.05 indicates data are significantly different from a normal distribution.

```
##
##  Shapiro-Wilk normality test
##
## data:  inpatient_data$participant_age
## W = 0.95445, p-value = 7.573e-11
```

Based on distribution, we will report median (Q1 - Q3) patient age for this dataset.

See results section of IMPALA manuscript to view all descriptive statistics reported.
The IMPALA manuscript linked here is for a different subset of patients than the example dataset we are using, so the numbers will not align.

Bivariable analysis

Objective: compare characteristics, medical history, etc., of cases and controls.
Identify statistically significant differences between cases and controls.
Cases: ESCrE colonized participants (n = 219)
Controls: Non-colonized participants (n = 250)

Statistical significance
"In research, statistical significance measures the probability of the null hypothesis being true compared to the acceptable level of uncertainty regarding the true answer."

Significance level: the probability the researcher is willing to be incorrect
Typically set alpha to 0.05: we are willing to be incorrect 5% of the time.
Tests for comparing groups generate a *P* value.
P values < 0.05 indicate statistical significance.
Variables identified as statistically significant between cases and controls will be examined in later analyses.

Age example
Is there a statistically significant difference in age between cases and controls?

```
## $`0`
##  participant_age    is_escre
##  Min.   : 5.00   Min.   :0
##  1st Qu.:26.00   1st Qu.:0
##  Median :44.50   Median :0
##  Mean   :46.87   Mean   :0
##  3rd Qu.:68.75   3rd Qu.:0
##  Max.   :89.00   Max.   :0
##
## $`1`
##  participant_age    is_escre
##  Min.   : 5.00   Min.   :1
##  1st Qu.:27.00   1st Qu.:1
##  Median :44.00   Median :1
##  Mean   :46.63   Mean   :1
##  3rd Qu.:67.00   3rd Qu.:1
##  Max.   :89.00   Max.   :1
```

Based on descriptive analysis, we know that the age variable is not normally distributed.
If our numeric variable were normally distributed, we could use Independent Samples *t* Test.
Instead, we will use Wilcoxon Rank Sum test.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  participant_age by is_escre
## W = 27473, p-value = 0.9469
## alternative hypothesis: true location shift is not equal to 0
```

P value > 0.05, so we do not reject the null hypothesis.
Even though this was not significant, we included participant age in our multivariable analysis for IMPALA.

Checking for independence between categorical variables

Chi-square test of Independence:
- compare observed and expected frequencies to determine if observed data deviate from expected frequencies
- should not use with small expected frequencies (if more than 20% of cells have expected frequencies below 5)

Fisher's Exact Test:
- use with small *n* or small expected frequencies

Example: count of participants with diabetes

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(inpatient_data$diabetes_mellitus, inpatient_data$is_escre)
## X-squared = 0, df = 1, p-value = 1
```

```
##
##            0        1
##   0 127.9318 112.0682
##   1 122.0682 106.9318
```

```
##
##       0   1
##   0 128 112
##   1 122 107
```

Okay to use chi-square test

```
##
##  Fisher's Exact Test for Count Data
##
## data:  inpatient_data$diabetes_mellitus and inpatient_data$is_escre
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.6858975 1.4646888
## sample estimates:
## odds ratio
##   1.002337
```

N is small enough that Fisher's test could work here, too.

Example: count of participants with 3 or more hospital visits to receive care in the past 6 months

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(inpatient_data$hosp_visit3, inpatient_data$is_escre)
## X-squared = 0.084315, df = 1, p-value = 0.7715
```

```
##
##            0        1
##   0 122.0682 106.9318
##   1 127.9318 112.0682
```

```
##
##       0   1
##   0 120 109
##   1 130 110
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  inpatient_data$hosp_visit3 and inpatient_data$is_escre
## p-value = 0.7121
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.6374541 1.3613396
## sample estimates:
## odds ratio
##  0.9316868
```

Significant difference in hospital visits between cases and controls. We will include this variable in our multivariable model.

Example: count of participants who tended swine at home at least once per week in the past 6 months

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(inpatient_data$tend_swine_home, inpatient_data$is_escre)
## X-squared = 0.54203, df = 1, p-value = 0.4616
```

```
##
##            0         1
##   0 136.4606 119.53945
##   1 113.5394  99.46055
```

```
##
##       0   1
##   0 132 124
##   1 118  95
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  inpatient_data$tend_swine_home and inpatient_data$is_escre
## p-value = 0.4572
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.585160 1.254755
## sample estimates:
## odds ratio
##  0.8573105
```

One can also run a series of logistic regression models with one independent variable in each model to assess relationships and identify which independent variables should be included in multivariable model.

broom::tidy can also be used in combination with other stats functions (e.g., fisher.test) to perform statistical test on all columns of df and generate tidy output.

```
## # A tibble: 15 x 7
##    Variable              `Odds Ratio`      SE Statistic `P-value` `95% CI Lower`
##    <chr>                        <dbl>   <dbl>     <dbl>     <dbl>          <dbl>
##  1 hosp_visit                   1.38  0.186      1.75      0.0798          0.963
##  2 tend_livestock_home          0.729 0.186     -1.70      0.0889          0.506
##  3 total_abx_days               1.04  0.0227     1.63      0.103           0.993
##  4 clin_visit3                  0.761 0.186     -1.47      0.142           0.529
##  5 participant_sex              1.24  0.185      1.16      0.246           0.862
##  6 clin_visit                   0.846 0.185     -0.902     0.367           0.588
##  7 tend_swine_home              0.857 0.186     -0.829     0.407           0.595
##  8 days_to_sample               0.984 0.0202    -0.805     0.421           0.946
##  9 chronic_kidney_disea~        1.15  0.185      0.740     0.459           0.798
## 10 travel_6m                    0.896 0.185     -0.592     0.554           0.623
## 11 tend_poultry_home            0.923 0.185     -0.433     0.665           0.641
## 12 hosp_visit3                  0.932 0.185     -0.383     0.702           0.648
## 13 high_bloodpressure           0.953 0.185     -0.259     0.796           0.663
## 14 participant_age              1.00  0.00382   -0.104     0.917           0.992
## 15 diabetes_mellitus            1.00  0.185      0.0126    0.990           0.697
## # i 1 more variable: `95% CI Upper` <dbl>
```

Can do additional formatting.

Multivariable analysis

Our dependent variable is dichotomous (participant had or did not have an ESCrE). We will use logistic regression to examine the association of the independent variables (identified in bivariable analysis) with having an ESCrE.

If we were trying to predict a continuous outcome, we might use linear regression.

In reality, there were many more independent variables which we tested in our multivariable models. The example below only shows our final (or mostly final) model.

As previously noted, the dataset used here was randomly generated and does not represent the actual study data. In our "real" final model, most of the predictors were significant or approaching significance.

Characteristic	OR[1]	95% CI[1]	p-value
Participant age	1.00	0.99, 1.01	>0.9
Days from enrollment to sampling	0.98	0.95, 1.02	0.4
Total days of antibiotic therapy	1.04	0.99, 1.08	0.11
Visited hospital 3+ times	0.93	0.64, 1.34	0.7
Tended swine at home	0.84	0.58, 1.22	0.4
Traveled outside the country	0.91	0.63, 1.31	0.6

[1]OR = Odds Ratio, CI = Confidence Interval

Read more about gtsummary.

check for multicollinearity (when independent variables are correlated)

```
## participant_age  days_to_sample  total_abx_days     hosp_visit3 tend_swine_home
##        1.005675        1.004333        1.002252        1.006567        1.004245
##       travel_6m
##        1.008799
```

Variance inflation factors are low, so there appear to be no issues.

Other useful packages and functions:

- table1 package for generating tables of descriptive statistics
- GGally::ggpairs - allows you to plot many independent variables against the outcome variable simultaneously