Biostatistics and Epidemiology - REPORT3

By Gihozo Christian (223013295) & Iradukunda Claudine (223010775)

20-02-2026

Contents

Question 1: Prevalence of Caries in Belo Horizonte (at the beginning of study)

```
library(VGAMdata)
```

```
data("belcap")
View(belcap)
```

Overall

```
belcap$caries_b <- ifelse(belcap$dmftb > 0, 1, 0)
```

```r
n_total <- nrow(belcap)
n_cases <- sum(belcap$caries_b)

prev <- n_cases / n_total
prev
```

```
## [1] 0.7841907
```

Confidence Interval for Overall

```r
se <- sqrt(prev * (1 - prev) / n_total)

lower <- prev - 2 * se
upper <- prev + 2 * se

c(lower, upper)
```

```
## [1] 0.7550469 0.8133346
```

interpretation

At baseline, the prevalence of dental caries was 78.4% (95% CI: 75.5%–81.3%).

Epidemiologically, this indicates that nearly four out of five 7-year-old children already had evidence of caries at the start of the study, reflecting a very high burden of disease in this population.

Stratified by Gender

```r
library(dplyr)
```

```r
belcap %>%
  group_by(gender) %>%
  summarise(
    n = n(),
    cases = sum(caries_b),
    prevalence = cases/n,
    se = sqrt(prevalence*(1-prevalence)/n),
    lower = prevalence - 2*se,
    upper = prevalence + 2*se
  )
```

```
## # A tibble: 2 x 7
##   gender     n cases prevalence     se lower upper
##   <fct>  <int> <dbl>      <dbl>  <dbl> <dbl> <dbl>
## 1 0        389   301      0.774 0.0212 0.731 0.816
## 2 1        408   324      0.794 0.0200 0.754 0.834
```

interpretation

The prevalence of caries among females was 77.4% (95% CI: 73.1%–81.6%), while among males it was 79.4% (95% CI: 75.4%–83.4%).

The prevalence was slightly higher in males; however, the confidence intervals overlap substantially, suggesting no strong epidemiological evidence of a meaningful difference in baseline caries prevalence between these boys and girls.

stratified by School

```
belcap %>%
  group_by(school) %>%
  summarise(
    n = n(),
    cases = sum(caries_b),
    prevalence = cases/n,
    se = sqrt(prevalence*(1-prevalence)/n),
    lower = prevalence - 2*se,
    upper = prevalence + 2*se
  )
```

```
## # A tibble: 6 x 7
##   school     n cases prevalence     se lower upper
##   <fct>  <int> <dbl>      <dbl>  <dbl> <dbl> <dbl>
## 1 1        124   108      0.871 0.0301 0.811 0.931
## 2 2        127    87      0.685 0.0412 0.603 0.767
## 3 3        136   118      0.868 0.0291 0.810 0.926
## 4 4        132   101      0.765 0.0369 0.691 0.839
## 5 5        155   116      0.748 0.0349 0.679 0.818
## 6 6        123    95      0.772 0.0378 0.697 0.848
```

interpretation

Baseline caries prevalence varied considerably across schools, ranging from 68.5% in School 2 to 87.1% in School 1.

This variation suggests potential differences in environmental, behavioral, or preventive factors between schools, indicating that school-level factors may influence caries distribution in this population.

statified by gender and school

```
belcap %>%
  group_by(gender, school) %>%
  summarise(
    n = n(),
    cases = sum(caries_b),
    prevalence = cases/n,
    se = sqrt(prevalence*(1-prevalence)/n),
    lower = prevalence - 2*se,
    upper = prevalence + 2*se
  )
```

```
## `summarise()` has grouped output by 'gender'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 12 x 8
## # Groups:   gender [2]
##    gender school     n cases prevalence     se lower upper
##    <fct>  <fct>  <int> <dbl>      <dbl>  <dbl> <dbl> <dbl>
##  1 0      1         62    49      0.790 0.0517 0.687 0.894
##  2 0      2         60    47      0.783 0.0532 0.677 0.890
##  3 0      3         58    48      0.828 0.0496 0.728 0.927
##  4 0      4         65    46      0.708 0.0564 0.595 0.821
##  5 0      5         86    67      0.779 0.0447 0.690 0.869
##  6 0      6         58    44      0.759 0.0562 0.646 0.871
##  7 1      1         62    59      0.952 0.0273 0.897 1.01
##  8 1      2         67    40      0.597 0.0599 0.477 0.717
##  9 1      3         78    70      0.897 0.0344 0.829 0.966
## 10 1      4         67    55      0.821 0.0468 0.727 0.915
## 11 1      5         69    49      0.710 0.0546 0.601 0.819
## 12 1      6         65    51      0.785 0.0510 0.683 0.887
```

interpretation

When stratified by both gender and school, substantial variability was observed across subgroups. For example, prevalence among males in School 1 reached 95.2%, while males in School 2 had a much lower prevalence (59.7%).

These findings suggest that both gender and school context may jointly influence caries distribution. However, some confidence intervals are wide, indicating smaller subgroup sample sizes and less precise estimates.

Question 2: Prevalence of Caries in Belo Horizonte(at the beginning to end of study)

overall

```
belcap <- belcap %>%
  mutate(
    at_risk = ifelse(dmftb == 0, 1, 0),
    incident = ifelse(dmftb == 0 & dmfte > 0, 1, 0)
  )


risk_set <- belcap %>% filter(at_risk == 1)


risk_set %>%
  summarise(
    incidence = mean(incident),
    lower_95CI = incidence - 2 * sqrt(incidence*(1-incidence)/n()),
    upper_95CI = incidence + 2 * sqrt(incidence*(1-incidence)/n())
  )
```

```
##   incidence lower_95CI upper_95CI
## 1 0.2732558  0.2052977  0.3412139
```

interpretation

The overall incidence of caries during the follow-up period was 27.3% (95% CI: 20.5%–34.1%).

This indicates that approximately one in four children who were initially caries-free developed new caries over the two-year follow-up, demonstrating continued disease occurrence despite preventive interventions.

stratified by Gender

```
risk_set %>%
  group_by(gender) %>%
  summarise(
    incidence = mean(incident),
    lower_95CI = incidence - 2 * sqrt(incidence*(1-incidence)/n()),
    upper_95CI = incidence + 2 * sqrt(incidence*(1-incidence)/n())
  )
```

```
## # A tibble: 2 x 4
##   gender incidence lower_95CI upper_95CI
##   <fct>      <dbl>      <dbl>      <dbl>
## 1 0          0.295      0.198      0.393
## 2 1          0.25       0.156      0.344
```

interpretation

Incidence was 29.5% among females and 25.0% among males.

Although females showed a slightly higher incidence, the overlapping confidence intervals suggest no strong epidemiological evidence of a significant gender difference in new caries development.

stratified by School

```
risk_set %>%
  group_by(school) %>%
  summarise(
    incidence = mean(incident),
    lower_95CI = incidence - 2 * sqrt(incidence*(1-incidence)/n()),
    upper_95CI = incidence + 2 * sqrt(incidence*(1-incidence)/n())
  )
```

```
## # A tibble: 6 x 4
##   school incidence lower_95CI upper_95CI
##   <fct>      <dbl>      <dbl>      <dbl>
## 1 1          0.312    0.0807       0.544
## 2 2          0.275    0.134        0.416
## 3 3          0.278    0.0666       0.489
## 4 4          0.419    0.242        0.597
## 5 5          0.103    0.00540      0.200
## 6 6          0.321    0.145        0.498
```

interpretation

Incidence varied notably by school, ranging from 10.3% in School 5 to 41.9% in School 4.

This suggests that the risk of developing new caries differed across schools, potentially reflecting differences in intervention effectiveness, oral hygiene practices, dietary habits, or other contextual factors.

Stratified by Gender and School

```
risk_set %>%
  group_by(gender, school ) %>%
  summarise(
    incidence = mean(incident),
    lower_95CI = incidence - 2 * sqrt(incidence*(1-incidence)/n()),
    upper_95CI = incidence + 2 * sqrt(incidence*(1-incidence)/n())
  )
```

```
## `summarise()` has grouped output by 'gender'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 12 x 5
## # Groups:   gender [2]
##    gender school incidence lower_95CI upper_95CI
##    <fct>  <fct>      <dbl>      <dbl>      <dbl>
##  1 0      1          0.308    0.0517       0.564
##  2 0      2          0.231   -0.00294      0.464
##  3 0      3          0.2     -0.0530       0.453
##  4 0      4          0.526    0.297        0.755
##  5 0      5          0.158   -0.00941      0.325
##  6 0      6          0.286    0.0442       0.527
##  7 1      1          0.333   -0.211        0.878
##  8 1      2          0.296    0.121        0.472
##  9 1      3          0.375    0.0327       0.717
## 10 1      4          0.25     0            0.5
## 11 1      5          0.05    -0.0475       0.147
## 12 1      6          0.357    0.101        0.613
```

interpretation

When stratified by both gender and school, incidence estimates showed substantial variability, and several confidence intervals were wide or included negative lower bounds (due to normal approximation).

This indicates reduced precision in smaller subgroups and suggests that some subgroup estimates should be interpreted cautiously. Nonetheless, differences across schools appear more pronounced than differences between genders.

Why Is It Useless to Stratify by Age?

Because:

All children were 7 years old at baseline.

There is no variation in age.

Stratification requires variability.

So stratifying by age would give identical groups.