

Biostatistics (MATH11230), 2022/2023

Vanda Inácio

We will start by reproducing the plots we have in slides 7 and 9 of the lectures (sampling distributions of \widehat{OR} and $\log \widehat{OR}$). We will also make the corresponding plots for the sampling distributions of \widehat{RR} and $\log \widehat{RR}$. The key is to remember that, under a cohort design, the entry a in the 2×2 contingency table (slide 3) follows a binomial distribution whose parameters are the number of exposed individuals and probability of disease given exposure. Similarly, the entry c (in the same contingency table) also follows a binomial distribution whose parameters are, in turn, the number of individuals who are not exposed and the probability of disease given no exposure. We will be assuming that $p_1 = \Pr(D | E) = 0.2$ and that $p_2 = \Pr(D | \text{not } E) = 0.2$ and, further, that we have 50 individuals in the exposed group and another 50 individuals in the unexposed group. To make the results reproducible, I will be fixing the seed. The code is as follows.

```
n_exp <- 50
n_unexp <- 50
p1 <- 0.2
p2 <- 0.2

nsim <- 1000
OR <- RR <- numeric(nsim)

set.seed(123)
for(i in 1:nsim){
  a <- rbinom(1, n_exp, p1)
  c <- rbinom(1, n_unexp, p2)
  b <- n_exp - a
  d <- n_unexp - c
  OR[i] <- (a*d)/(b*c)
  RR[i] <- (a/(a+b))/(c/(c+d))
}

df_OR <- data.frame("Mean" = c(mean(OR), mean(log(OR))),
                    "Median" = c(median(OR), median(log(OR))),
                    "Min" = c(min(OR), min(log(OR))),
                    "Max" = c(max(OR), max(log(OR)))
                    )

rownames(df_OR) <- c("OR", "log OR")
knitr::kable(df_OR, escape = FALSE, digits = 3,
              caption = "Summary statistics of the sampling distributions of the OR and log OR")
```

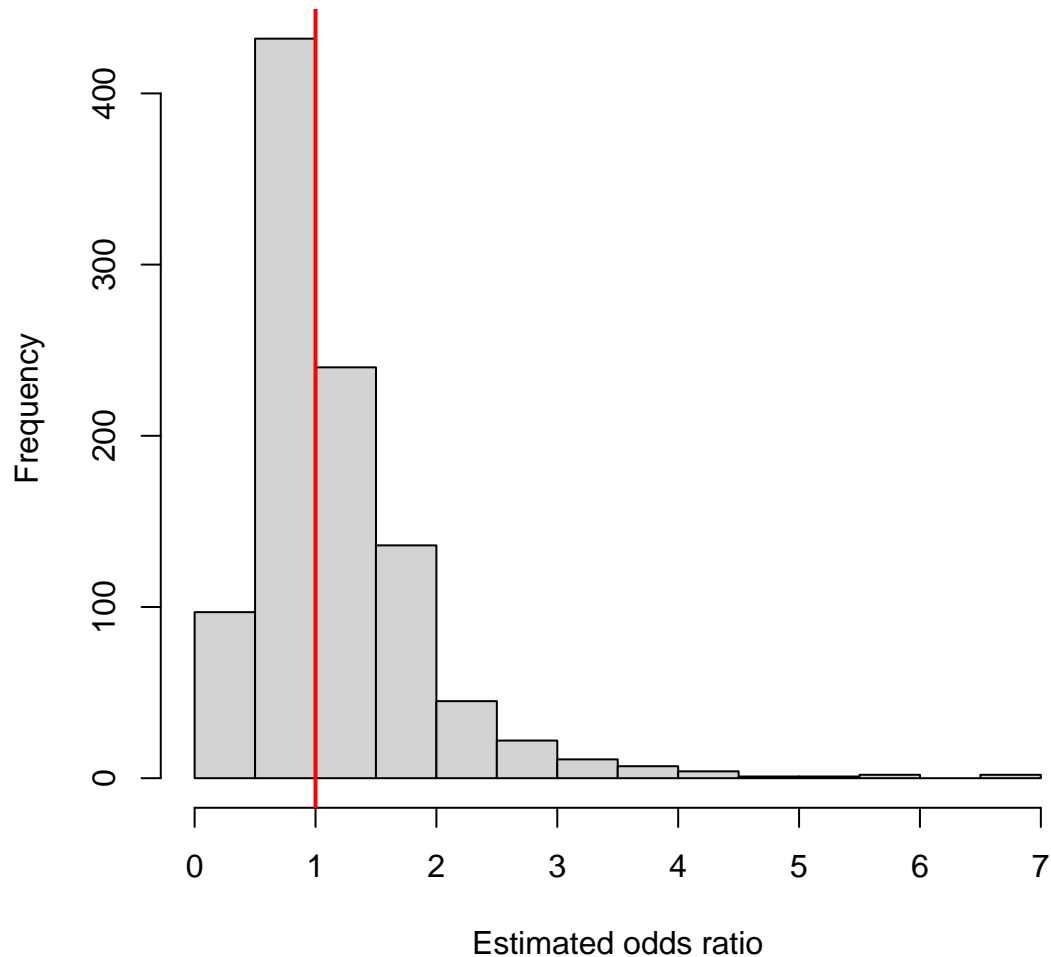
Table 1: Summary statistics of the sampling distributions of the OR and log OR

	Mean	Median	Min	Max
OR	1.193	1	0.202	6.769

	Mean	Median	Min	Max
log OR	0.023	0	-1.599	1.912

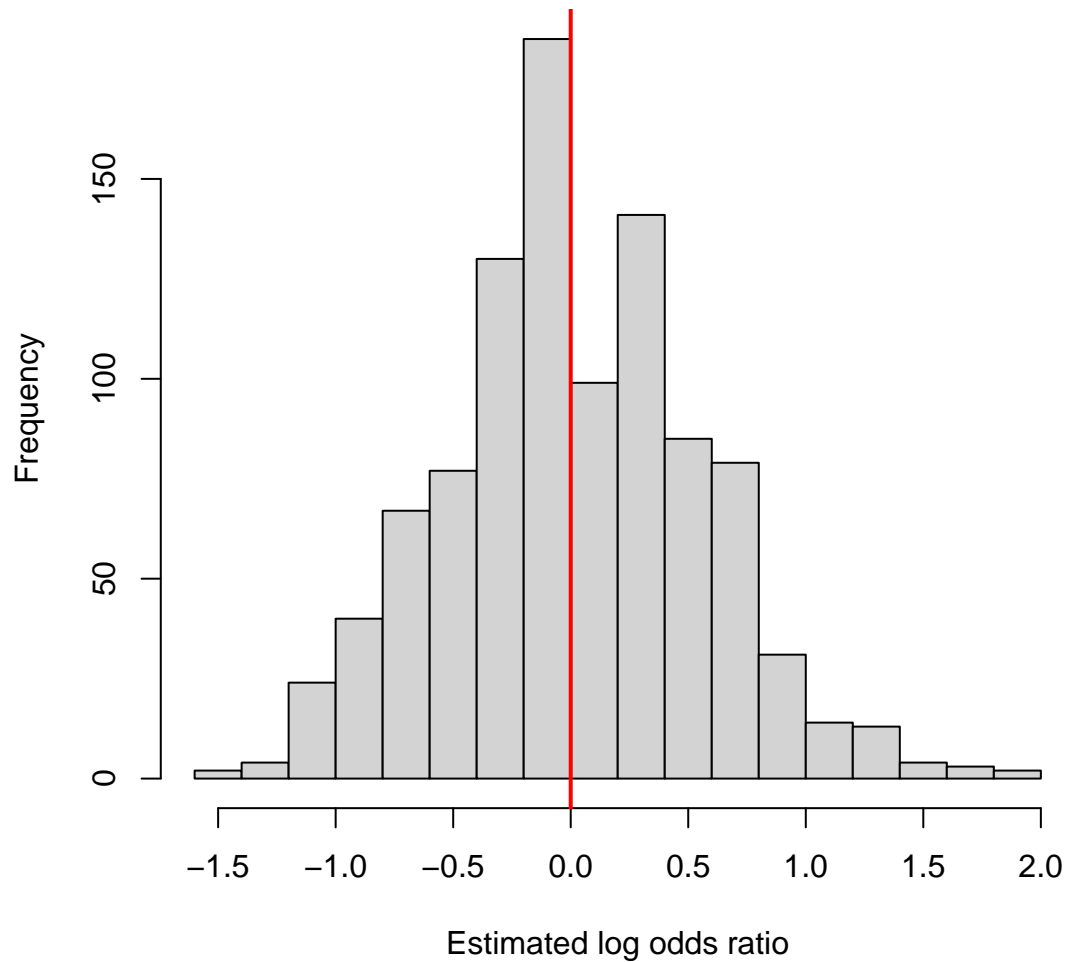
```
hist(OR, nclass = 20, xlab = "Estimated odds ratio",
     ylab = "Frequency", main = "Sampling distribution estimated OR")
abline(v = 1, lwd = 2, col = "red")
```

Sampling distribution estimated OR



```
hist(log(OR), nclass = 20, xlab = "Estimated log odds ratio",
     ylab = "Frequency", main = "Sampling distribution estimated log OR")
abline(v = 0, lwd = 2, col = "red")
```

Sampling distribution estimated log OR



```
df_RR <- data.frame("Mean" = c(mean(RR), mean(log(RR))),
                    "Median" = c(median(RR), median(log(RR))),
                    "Min" = c(min(RR), min(log(RR))),
                    "Max" = c(max(RR), max(log(RR)))
                    )

rownames(df_RR) <- c("RR", "log RR")
knitr::kable(df_RR, escape = FALSE, digits = 3,
             caption = "Summary statistics of the sampling distributions of the RR and log RR")
```

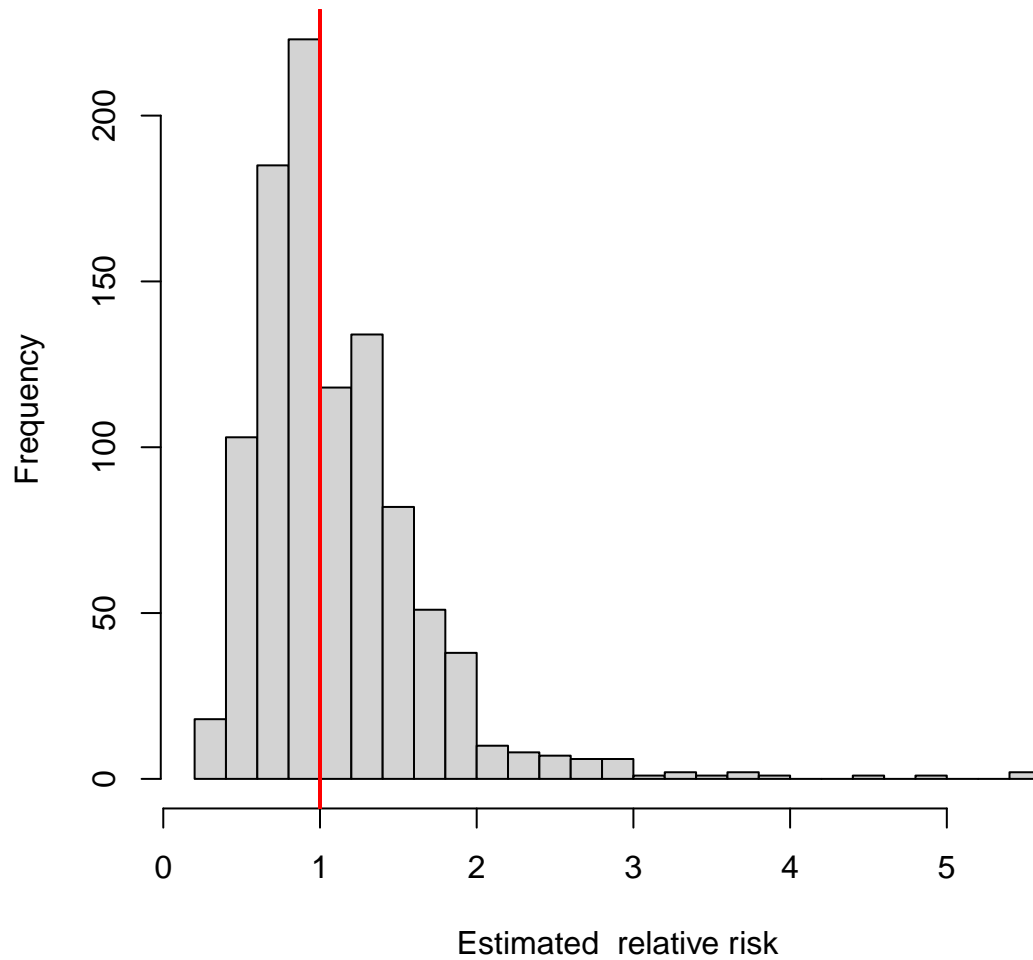
Table 2: Summary statistics of the sampling distributions of the RR and log RR

	Mean	Median	Min	Max
RR	1.129	1	0.250	5.500
log RR	0.019	0	-1.386	1.705

```
hist(RR, nclass = 20, xlab = "Estimated relative risk",
     ylab = "Frequency", main = "Sampling distribution estimated RR")
```

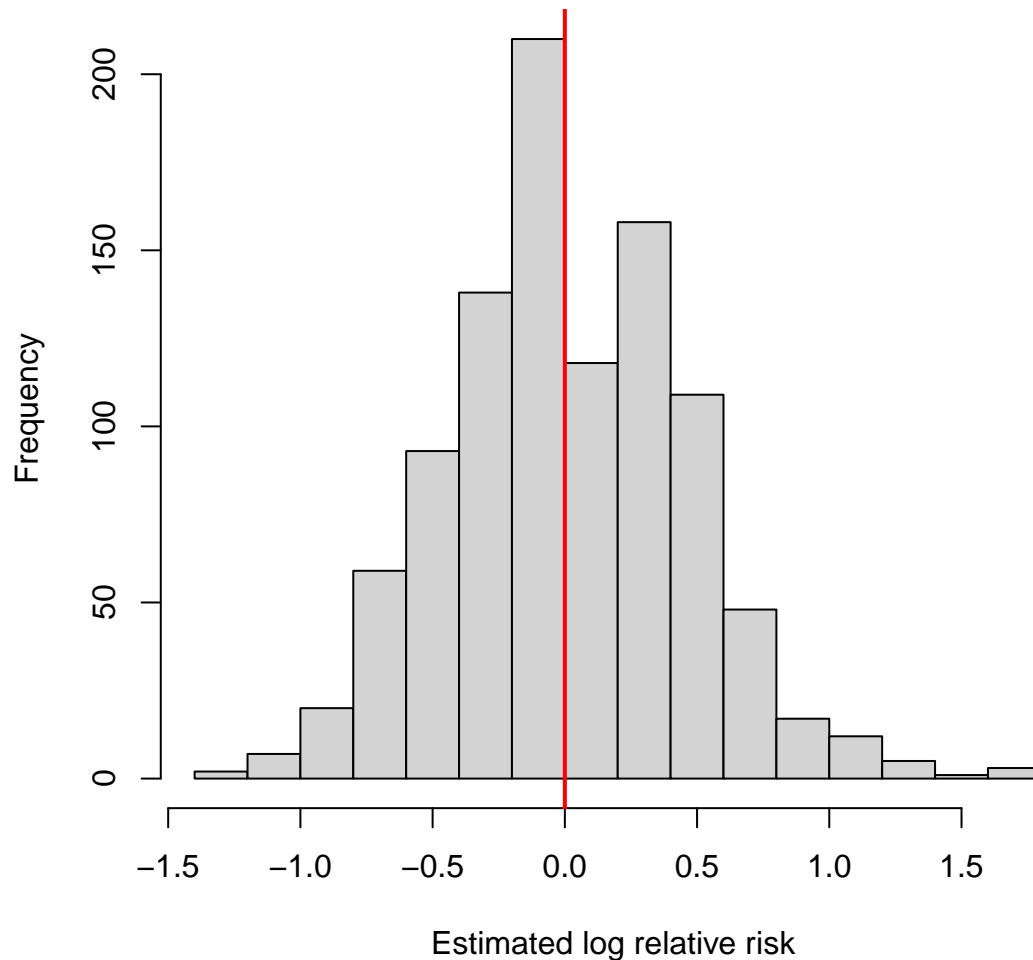
```
abline(v = 1, lwd = 2, col = "red")
```

Sampling distribution estimated RR



```
hist(log(RR), nclass = 20, xlab = "Estimated log relative risk ",  
      ylab = "Frequency", main = "Sampling distribution estimated log RR")  
abline(v = 0, lwd = 2, col = "red")
```

Sampling distribution estimated log RR



I now illustrate the usage of the `epiR` and `epitools` packages which, among many things, have functions to compute estimates and CIs for different measures of association. We will use the data from the example in slide 17.

```
require(epiR)
data <- c(62, 76, 5, 55)
epi.2by2(data, method = "case.control", conf.level = 0.95, units = 100,
          interpret = TRUE, outcome = "as.columns")
```

##	Outcome +	Outcome -	Total	Prevalence *	Odds
## Exposed +	62	76	138	44.93	0.8158
## Exposed -	5	55	60	8.33	0.0909
## Total	67	131	198	33.84	0.5115

##

Point estimates and 95% CIs:

Odds ratio 8.97 (3.38, 23.79)

Attrib fraction (est) in the exposed (%) 88.75 (69.69, 96.69)

Attrib fraction (est) in the population (%) 82.23 (57.71, 92.53)

Uncorrected chi2 test that OR = 1: chi2(1) = 25.013 Pr>chi2 = <0.001

```

## Fisher exact test that OR = 1: Pr>chi2 = <0.001
## Wald confidence limits
## CI: confidence interval
## * Outcomes per 100 population units
##
## Measures of association strength:
## The exposure odds among cases was 8.97 (95% CI 3.38 to 23.79) times greater than exposure odds among
##
## Measures of effect in the exposed:
## 88.7% of outcomes in the exposed were attributable to exposure (95% CI 69.7% to 96.7%).
##
## Measures of effect in the population:
## 82.2% of outcomes in the population were attributable to exposure (95% CI 57.7% to 92.5%).

```

```

require(epitools)
oddsratio(data, method = "wald", conf = 0.95, correct = FALSE)

```

```

## $data
##           Outcome
## Predictor Disease1 Disease2 Total
##   Exposed1      62      76   138
##   Exposed2       5      55    60
##   Total        67     131   198
##
## $measure
##           odds ratio with 95% C.I.
## Predictor estimate      lower      upper
##   Exposed1 1.000000         NA         NA
##   Exposed2 8.973684 3.384777 23.79093
##
## $p.value
##           two-sided
## Predictor      midp.exact fisher.exact   chi.square
##   Exposed1           NA           NA           NA
##   Exposed2 1.225344e-07 1.967011e-07 5.693107e-07
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"

```