This post tries to explain the difference between odds ratios and relative risk ratios; and how just a few letters in the code fitting a generalized linear model mean the difference between extracting one or the other. There are plenty of other explanations available (for example, [here](https://www.theanalysisfactor.com/the-difference-between-relative-risk-and-odds-ratios/) and [here](https://www.mdedge.com/jfponline/article/65515/relative-risks-and-odds-ratios-whats-difference)), but there is also still plenty of [confusion](https://www.researchgate.net/post/When_should_we_use_a_risk_ratio_odds_ratio_and_relative_ratio_in_the_meta-analysis_to_compare_between_2_predictors_variables) about the differences. I wanted to illustrate the issues with a concrete but simulated example and actual code that could be used as a foundation in the wild.

**Concepts**

Odds ratios and relative risk are commonly used to contrast the prevalence of some indicator (eg disease) in different categories of population. They seem to get particular emphasis in medical and epidemiological literature but are used broadly.

This diagram demonstrates with some simulated data the core concepts:

* Tigers have a 1/4 (0.25) probability of being diseased, which is “1 to 3” odds of being diseased.
* Lions have a 1/2 (0.5) probability of being diseased, which is “1 to 1” or “even” odds.
* Bears have a 1/10 (0.1) probability of being diseased, which is “1 to 9” odds.

I find the relative probabilities very intuitive – Tigers have 2.5 times the probability of being diseased as Bears, and Lions have 5 times the probability.

Odds ratios – not so much. Even though I’m a statistician with a particular interest in games of chance (which form the genealogy of odds), I don’t really intuit how to interpret the Tiger to Bears odds ratio of 3 times the odds of being diseased. In fact, there’s evidence that when scientific papers report odds ratios, they are [commonly misinterpreted as being risk ratios](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1112884/) (“People of type X are twice as likely to have the disease compared to people of type Y” being a common formulation).

When the probabilities are very close to zero, then odds ratios are similar to risk ratios. The farther they get from zero, the more they differ, and the more dangerous it is if people misinterpret an odds ratio as a risk ratio.

The best argument for reporting odds ratios is that they are symmetric – whether you are reporting on presence or absence, the odds ratio is unaffected. In our example above, if we looked at the probability of not being diseased, the Tigers:Bears odds ratio becomes 3:1 / 9:1 which is still a 3, whereas the relative probability is (3/4) / (9/10) which is 5/6 or 0.833. So while it’s nice and intuitive to say Tigers are 2.5 times as likely as Bears to be diseased, they are 0.833 times as likely as Bears to be not-diseased. This can lead to confusion. The odds ratio is three both ways (or at least, 3 in one direction and 1/3 in the other).

Most situations where we want these statistics, I think relative risk ratios are much better, more intutitive, and easier to explain. I don’t think the asymmetry matters much in most communications, whereas the misunderstanding of odds ratios often does. Frequent misunderstanding of the target audience is a high price to pay for symmetry. And I’m not sure why symmetry is necessary anyway.

A balanced (in my view) and non-paywalled article on the topic (which has generated many) is [this one by Cummings in JAMA Pediatrics](https://jamanetwork.com/journals/jamapediatrics/fullarticle/381459).

Before we go onto to getting these estimates from a fitted model, here’s the R code that created that simulated data and the graphic:

library(tidyverse)

library(scales)

library(viridis)

set.seed(765)

n <- 100000

data <- data\_frame(

animal\_ = sample(c("lion", "tiger", "bear"), size = n, replace = TRUE, prob = c(0.2, 0.3, 0.5))

) %>%

mutate(diseased = case\_when(

animal\_ == "lion" ~ sample(0:1, size = n, replace = TRUE, prob = c(0.5, 0.5)),

animal\_ == "tiger" ~ sample(0:1, size = n, replace = TRUE, prob = c(0.75, 0.25)),

animal\_ == "bear" ~ sample(0:1, size = n, replace = TRUE, prob = c(0.9, 0.1))

))

true\_props <- data %>%

group\_by(animal\_) %>%

summarise(diseased = round(mean(diseased), 2)) %>%

mutate(odds = c("1:9", "1:1", "1:3"),

lab = paste0("Probability: ", diseased, "\nOdds: ", odds))

ggplot(data, aes(fill = as.logical(diseased), x = animal\_)) +

geom\_bar(position = "fill") +

geom\_text(data = true\_props, aes(label = lab, y = diseased), colour = "white") +

labs(x = "", y = "Proportion", fill = "Diseased or not:") +

guides(fill = guide\_legend(reverse = TRUE)) +

coord\_flip() +

ggtitle("Simulated data for demonstrating odds and risk ratios",

"Risk ratio: tiger / bear = 2.5, lion / bear = 5.0\nOdds ratio: tiger / bear = 3.0, lion / bear = 9.0")

**Obtaining odds and risk ratios from a generalized linear model**

Putting that aside, how do we get these estimates from a model?

**Odds ratios**

Frankly, I suspect the more material reason for the prevalence of odds ratios is that they fall easily out of the results of a logistic regression (generalized linear model with the canonical logit link function relating the mean of the response to the linear predictor – where the logit function is the logarithm of the odds). For a categorical explanatory variable that has been represented as dummy indicator variables, *e* to the power of the coefficient for the dummy variable will give an estimate of the odds ratio.

Here’s an example of extracting estimates of odds ratios from our data:

# Odds ratio:

model\_2 <- glm(diseased ~ animal\_, family = quasibinomial, data = data)

exp(coef(model\_2))[-1]

# animal\_lion animal\_tiger

# 9.286664 2.997837

#

These odds ratios are of the given animal (Lion or Tiger) relative to the disease rate of the reference level, which in this case is Bear. So these are estimates of the ratios depicted in the original diagram.

You would get the same point estimate if you used family = binomial, or family = quasi(link = 'logit', variance = "mu(1-mu)"); it’s the logit link that’s important here.

Note that these estimates are biased; despite being based on large samples of 100,000 animals they haven’t converged on the true odds ratios of 9 and 3. The estimates of the original coefficients are unbiased; but the non-linear transformation of exp(coef(model)) inevitably introduces bias. A statistic that is unbiased on one scale will be biased if you transform it to another scale; such is life.

Using binomial versus quasibinomial does make a difference to confidence intervals, but in our current case it’s not material. The confidence intervals at least contain the correct values; although the point estimate is biased (and also not in the centre of the confidence interval, again due to the non-linear transformation), the confidence interval is still valid for the transformed scale:

exp(confint(model\_2))[-1, ]

# 2.5 % 97.5 %

# animal\_lion 8.920358 9.669207

# animal\_tiger 2.881889 3.118685

#

**Risk ratios**

At a minimum, the *only* change that needs to be done to get risk ratios is to change the link function that relates the mean value of the response variable to the linear predictor. For estimates of odds ratios, this is logit (ie the logarithm of the odds of the mean); for estimates of relative risk ratios, this becomes logarithm. We can specify this manually, or just use a built-in family for our generalized linear model for which the logarithm is the canonical link fucntion, and hence the default. The quasipoisson is a great choice here:

# Risk ratio:

model\_1 <- glm(diseased ~ animal\_, family = quasipoisson, data = data)

exp(coef(model\_1))[-1]

# animal\_lion animal\_tiger

# 5.112507 2.504792

#

Again, these point estimates are biased because of the non-linear exp() transformation of the unbiased original coefficients. But the confidence intervals contain the correct values (which recall are 5.0 and 2.5):

exp(confint(model\_1))[-1, ]

# 2.5 % 97.5 %

# animal\_lion 4.961963 5.268162

# animal\_tiger 2.426520 2.585758

#

If we’d used poisson as our family instead of quasipoisson, the dispersion parameter is forced to be 1.0, instead of 0.7 which it is estimated to be by the quasipoisson model. The net effect of that “underdispersion” in the data in this case is that the confidence intervals with family = poisson are *larger* than in the quasipoisson instance. This won’t always apply, but it often will when we are modelling data such as this which is, in fact under-dispersed compared to a pure poisson distribution (because none of the counts can exceed 1).

These examples have avoided complications of other explanatory variables, but the point of using a generalized linear model for this (rather than observing proportions directly) is that we can add in other variables and report on relative risks (or odds, if we must) “while controlling for …” those other variables.

The approach also [nicely extends to complex surveys](http://grokbase.com/t/r/r-help/109dad6km5/r-relative-risk-regression-with-survey-data); we can use svyglm() from Lumley’s survey package, with any of the family specifications above, according to our needs and get good results.

**Summary**

So, key points;

* carefully choose which (or both) of odds or risk ratios to use
* explicitly communicate what you are reporting, and try to counter in your communication the likely misunderstandings particularly with odds ratios
* exp(coef(model)) for a glm fit with family = quasibinomial will give odds ratios
* exp(coef(model)) for a glm fit with family = quasipoisson will give relative risk ratios
* remember that point estimates from this approach are biased, and like all point estimates give an undue impression of precision; always include a confidence interval or credibility interval as well.