Week 04 Questionable research practices

Open and reproducible science: dependable computations and statistics

Homework Solutions

We will work on the data collected for the fivethirtyeight article "You Can't Trust What You Read About Nutrition" https://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition/

The complete data that were produced are available here along with analysis code: https://www.kaggle.com/fivethirtyeight/fivethirtyeight-nutrition-studies-dataset

Read in the p-values that have been calculated for the article. Note that there is an error in one name of an FFQ variable.

```
# load the results of the original analysis
pvals_orig <- read.csv("data/p_values_analysis.csv")
pvals_orig$food[ is.na(pvals_orig$food)] <- "BREAKFASTSANDWICHFREQ"
```

Task 1

Assuming that 10% of the performed regressions are quantifying "true" associations and that the sample size of 54 corresponds to a power of 60% (which is extremely wishful thinking) calculate the (hypothetical) positive predictive value in this situation. Assume a significance level of 5%.

```
siglevel <- 0.05
# suppose 10 % of the hypothesis are indeed true
perctrue <- 0.1
# No. false positives on average
nofalsepos <- dim(pvals_orig)[1]*(1- perctrue) * siglevel
# notrueneg <- dim(pvals_orig)[1]*(1- perctrue) * (1-siglevel)

# suppose 54 respondents correspond to a power of 60%
suppower <- 0.6
# No. true positives on average
# nofalseneg <- dim(pvals_orig)[1]*perctrue * (1 - suppower)
notruepos <- dim(pvals_orig)[1]*perctrue * suppower

ppv <- notruepos/(notruepos + nofalsepos)
ppv
```

```
## [1] 0.5714286
```

Task 2

How many positive decisions have been taken at the 5% significance level? How many are hypothetically really true under the parameters of Task 1?

```r
noposdec <- sum(pvals_orig$p_values < 0.05, na.rm =TRUE)
trueposdec <- noposdec*ppv
noposdec
```

```
## [1] 1081
```

```r
trueposdec
```

```
## [1] 617.7143
```

Task 3

Take the association between cat ownership `cat` and coffee consumption `COFFEEDRINKSFREQ` and fit the corresponding linear model (see the results table in the article). What is the estimated coefficient for the association and its confidence interval? How do you interpret these values correctly? What do you think about the associated p-value?

```r
rawData <- read.csv("data/raw_anonymized_data.csv")
modelcatcoffee <- lm(COFFEEDRINKSFREQ ~ cat, data=rawData)
coefcatcoffee <- coef(modelcatcoffee)[2]
confintcatcoffee <- confint(modelcatcoffee)[2,]
coefcatcoffee
```

```
##   catYes
## 1.855882
```

```r
confintcatcoffee
```

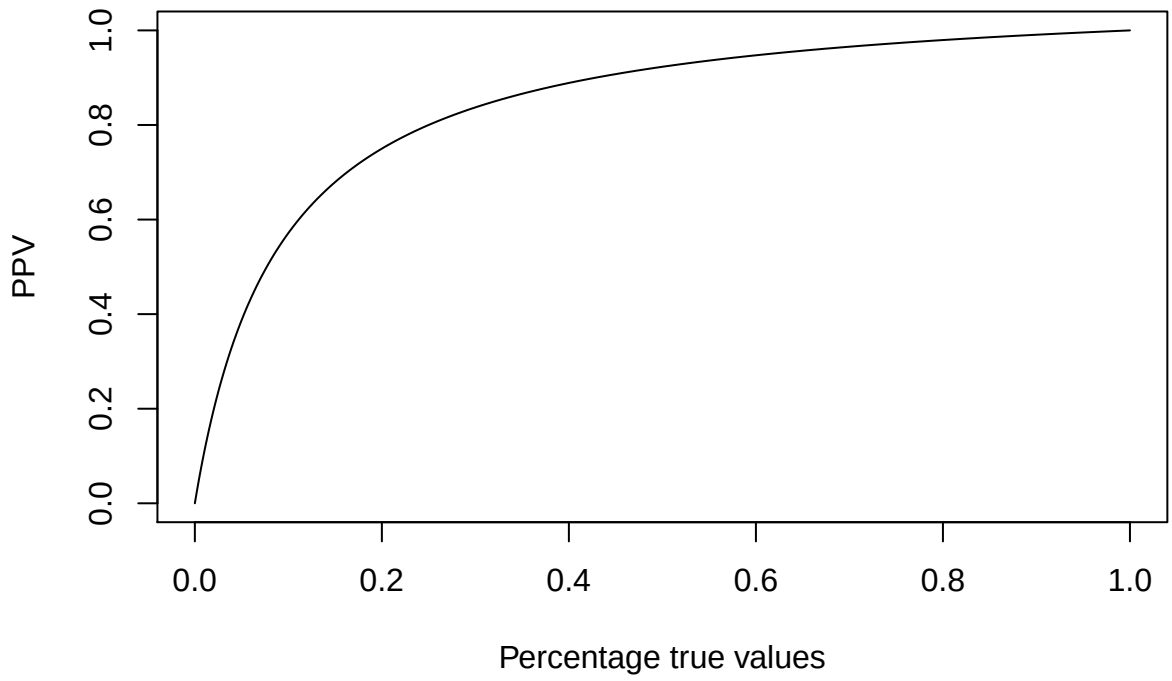
```
##     2.5 %    97.5 %
## 0.7355611 2.9762036
```

Task 4

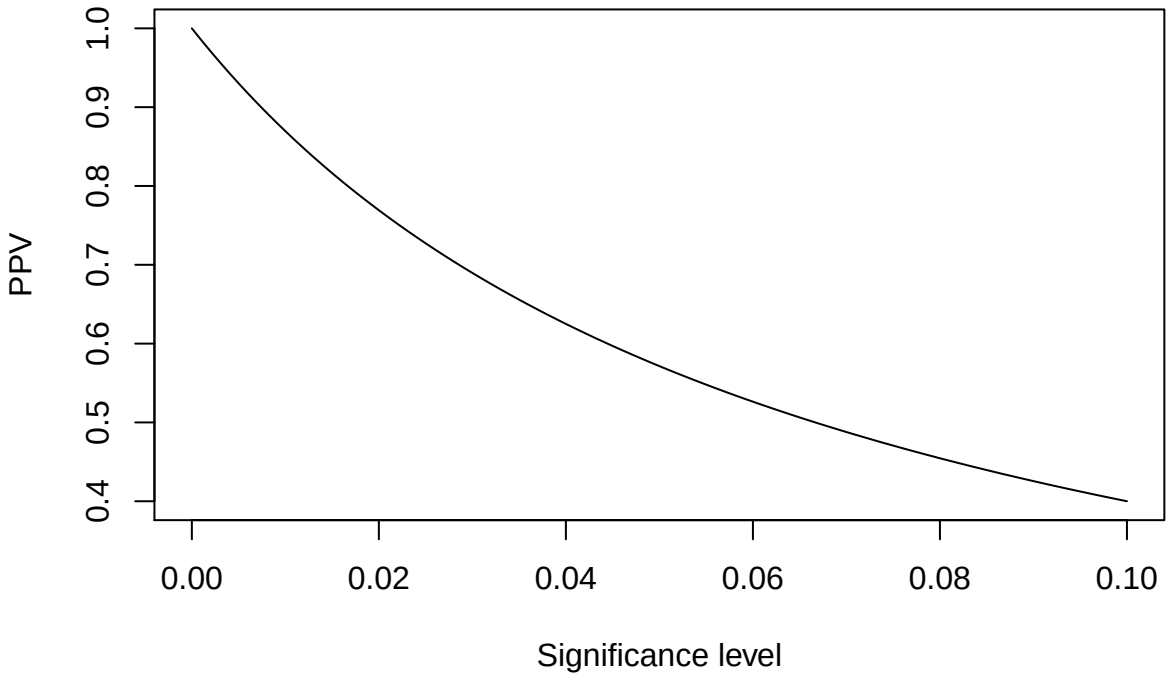
Going back to the PPV calculation write a function that parameterizes the percentage of true associations, the significance level and the power; and illustrate how the PPV depends on these three characteristics.

```r
#' Positive Predictive Value
#'
#' @param perctrue numeric, percentage of true results
#' @param siglevel numeric, significance level, default=0.05
#' @param power numeric, power, default=0.6
#'
#' @return numeric
#' @export
#'
#' @examples
#'  ppv(0.1)
ppv <- function(perctrue, siglevel = 0.05, power = 0.6){
  truepos <- perctrue * power
```

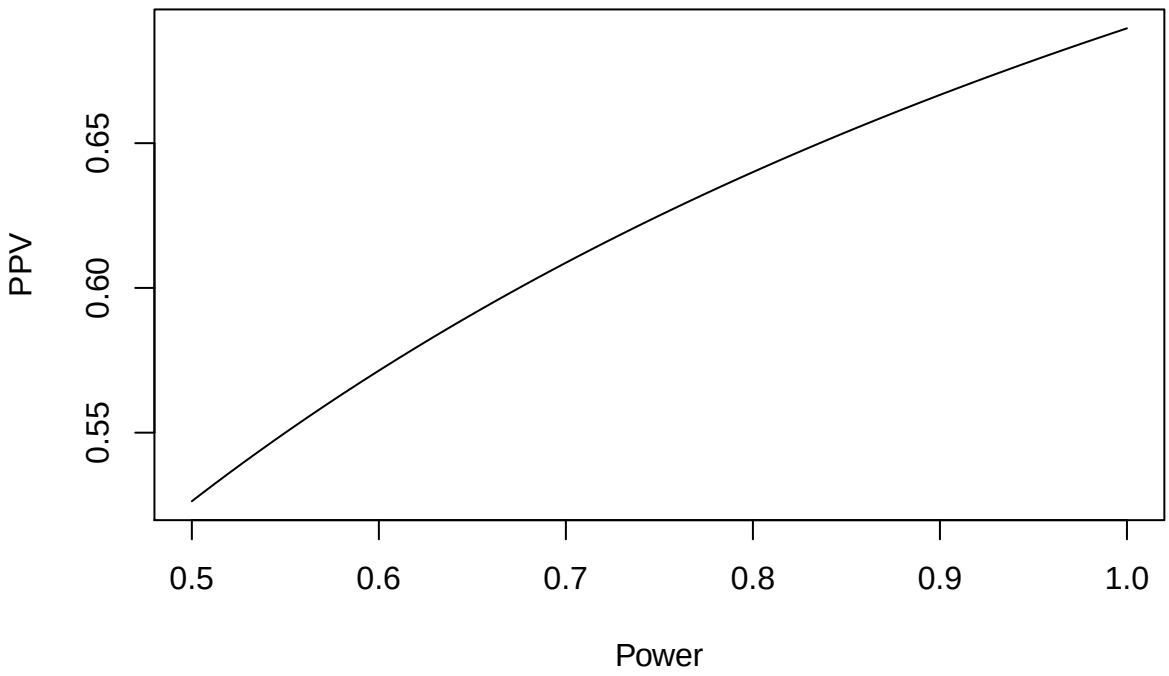
```
  falsepos <- (1-perctrue) * siglevel
  truepos/(truepos + falsepos)
}

v <- seq(0,1,by=0.001)
ppvs <- sapply(v, function(p) ppv(p))
plot(v,ppvs, xlab = "Percentage true values", ylab = "PPV", type="l")
```




```
v <- seq(0,.1,by=0.001)
ppvs <- sapply(v, function(p) ppv(0.1,siglevel=p))
plot(v,ppvs, xlab = "Significance level", ylab = "PPV", type="l")
```




```
v <- seq(.5,1,by=0.001)
ppvs <- sapply(v, function(p) ppv(0.1,power=p))
plot(v,ppvs, xlab = "Power", ylab = "PPV", type="l")
```


PPV
0.55
0.60
0.65
0.5
0.6
0.7
0.8
0.9
1.0
Power