Chapter 1

Assessing a test size

Prompted by a speaker in S13 and some of my own work with categorical tables, let's discuss Fisher's exact test and whether it is too conservative or not.

1. How can we assess this?
2. What experimental designs can we use to help us?
3. What tests might perform better?

1.1 Tests to compare

1.1.1 Fisher's exact test

Fisher's exact test is a statistical significance test used in the analysis of contingency tables. Fisher is said to have devised the test following a comment from Dr Muriel Bristol, who claimed to be able to detect whether the tea or the milk was added first to her cup (it turns out that she could).

The test is useful for categorical data that result from classifying objects in two different ways; it is used to examine the significance of the association (contingency) between the two kinds of classification. So in Fisher's original example, one criterion of classification could be whether milk or tea was put in the cup first; the other could be whether Dr Bristol thinks that the milk or tea was put in first. We want to know whether these two classifications are associated — that is, whether Dr Bristol really can tell whether milk or tea was poured in first. Most uses of the Fisher test involve, like this example, a 2-by-2 contingency table. The p-value from the test is computed as if the margins of the table are fixed, i.e., as if, in the tea-tasting example, Dr Bristol knows the number of cups with each treatment (milk or tea first) and will therefore provide guesses with the correct number in each category. As pointed out by Fisher, this leads under a null hypothesis of independence to a hypergeometric distribution of the numbers in the cells of the table.

We represent the cell frequencies by the letters a, b, c, and d, call the totals across rows and columns marginal totals, and represent the grand total by n. Such a table looks like this.

	Condition 1		
Cond 2	W	X	Row total
Y	a	b	$a+b$
Z	c	d	$c+d$
Col tot	$a+c$	$b+c$	$a+b+c+d=n$

Fisher showed that the probability of obtaining any such set of values (conditional on the marginal frequencies) was given by the hypergeometric distribution:

$$\begin{aligned} p &= \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \\ &= \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ b!\ c!\ d!\ n!} \end{aligned}$$

The formula above gives the exact hypergeometric probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that W and X are equally likely to be Y.

To put it another way, if we assume that the probability that a W is a Y is p, the probability that a X is a Y is p, and we assume that both W and X enter our sample independently of whether or not they are Y, then this hypergeometric formula gives the conditional probability of observing the values a, b, c, and d in the four cells, conditionally on the observed marginals (i.e., assuming the row and column totals shown in the margins of the table are given). This remains true even if W enters our sample with different probabilities than X. The requirement is merely that the two classification characteristics, Y (or Z), are not associated.

Here is an example in R using criminal convictions of like-sex twins (Fisher 1962, 1970). Note that "Dizygotic" (two eggs) is for fraternal twins and "Monozygotic" is for identical twins.

```
Convictions <- matrix(c(2, 10, 15, 3)
                    , nrow = 2
                    , dimnames = list(Twins = c('Dizygotic', 'Monozygotic')
                                    , Status = c('Convicted', 'Not convicted'))
                     )
Convictions
```

```
##              Status
## Twins         Convicted Not convicted
##   Dizygotic           2            15
##   Monozygotic        10             3
```

```
fisher.test(Convictions)
```

```
## 
##  Fisher's Exact Test for Count Data
## 
## data:  Convictions
## p-value = 0.0005367
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.003325764 0.363182271
## sample estimates:
## odds ratio 
## 0.04693661
```

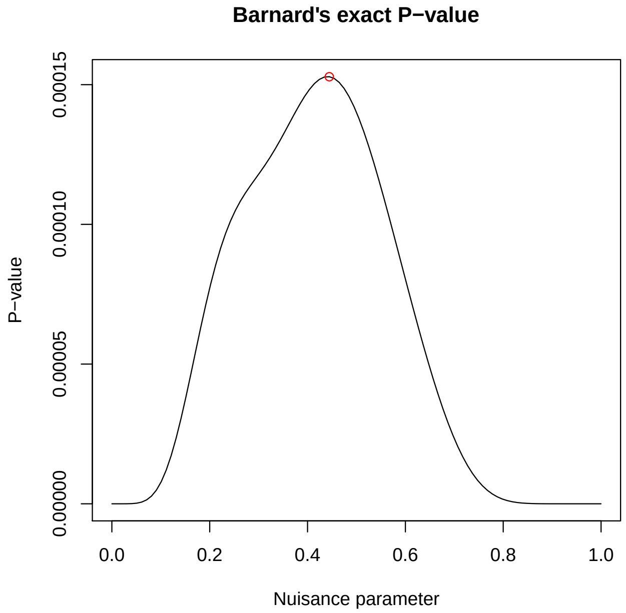
1.1.2 Barnard's test

In statistics, Barnard's test is an exact test of the null hypothesis of independence of rows and columns in a contingency table. It is an alternative to Fisher's exact test but is more time-consuming to compute. The test was first published by George Alfred Barnard (1945, 1947) who claimed this test for 2-by-2 contingency tables is more powerful than Fisher's exact test.

Mehta and Senchaudhuri (2003) explain why Barnard's test can be more powerful than Fisher's under certain conditions: "When comparing Fisher's and Barnard's exact tests, the loss of power due to the greater discreteness of the Fisher statistic is somewhat offset by the requirement that Barnard's exact test must maximize over all possible p-values, by choice of the nuisance parameter p. For 2-by-2 tables the loss of power due to the discreteness dominates over the loss of power due to the maximization, resulting in greater power for Barnard's exact test. But as the number of rows and columns of the observed table increase, the maximizing factor will tend to dominate, and Fisher's exact test will achieve greater power than Barnard's."

```
# Function available from:
# https://raw.github.com/talgalili/R-code-snippets/master/Barnard.R
# (for function, see R code of notes, it's long and not commented)
barnard.test(Convictions)
```

```
##
##  2x2 matrix Barnard's exact test: 100 13x19 tables were evaluated
##  -----------------------------------------------------------
##  Wald statistic =  3.6099
##  Nuisance parameter =  0.44446
##  p-values:  1-tailed =  0.00015285 2-tailed =  0.00030569
##  -----------------------------------------------------------
##
## [1] 0.0003056916
```



1.2 Comparison of tests, Type-I error

As a starting point, let's consider the following table, where the probability of Y for both W and X is 0.5 with sample sizes of 10 for each W and X.

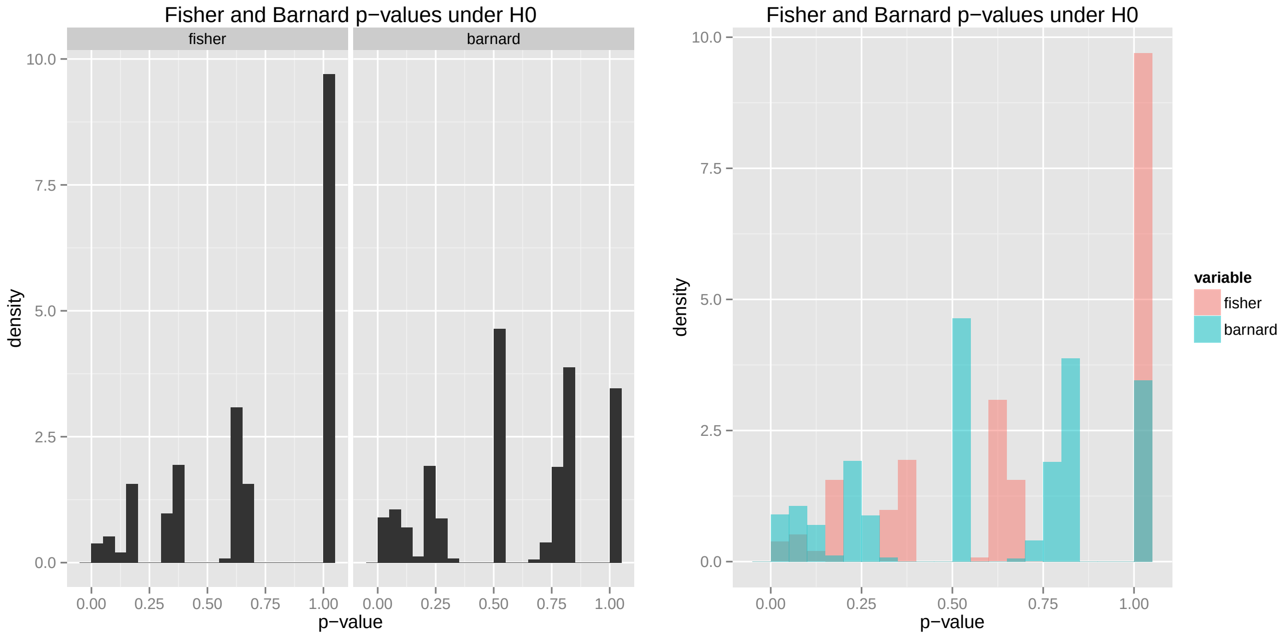
	Condition 1		
Cond 2	W	X	Row total
Y	a	b	$a+b$
Z	c	d	$c+d$
Col tot	10	10	$10+10=20$

Using Monte Carlo, we can draw a large number (R) of random samples under the null hypothesis of "no association" and compare the observed size of the test to the expected size.

```
# number of repetitions
R <- 1e3  # 1e3
# column totals
col.n <- c(10, 10)
# first row probabilities
row.p <- c(0.5, 0.5)
```

```
# draw independent samples of Y|W and Y|X
freq.Y <- data.frame(W = rbinom(R, col.n[1], row.p[1])
                   , X = rbinom(R, col.n[2], row.p[2])
                    )
head(freq.Y)

##   W X
## 1 4 5
## 2 4 6
## 3 5 5
## 4 7 8
## 5 4 3
## 6 7 2

p.values <- data.frame(fisher  = rep(NA, R)
                     , barnard = rep(NA, R)
                      )

for (i.R in 1:R) {
  tab <- matrix(c(freq.Y[i.R, 1], col.n[1] - freq.Y[i.R, 1]
                , freq.Y[i.R, 2], col.n[2] - freq.Y[i.R, 2])
              , nrow = 2)
  p.values$fisher[i.R]  <- fisher.test(tab)$p.value
  p.values$barnard[i.R] <- barnard.test(tab, to.print = FALSE, to.plot = FALSE)
}


library(reshape2)
p.values.long <- melt(p.values)

## No id variables; using all as measure variables
library(ggplot2)
p <- ggplot(p.values.long, aes(x = value)) #, fill = variable))
p <- p + geom_histogram(aes(y = ..density..), binwidth = 0.05, alpha = 1, position="identity")
p <- p + labs(title = "Fisher and Barnard p-values under H0")
p <- p + facet_wrap( ~ variable)
p <- p + xlab("p-value")
p <- p + ylab("density")
print(p)

library(ggplot2)
p <- ggplot(p.values.long, aes(x = value, fill = variable))
p <- p + geom_histogram(aes(y = ..density..), binwidth = 0.05, alpha = 0.5, position="identity
p <- p + labs(title = "Fisher and Barnard p-values under H0")
p <- p + xlab("p-value")
p <- p + ylab("density")
```

```r
print(p)
```



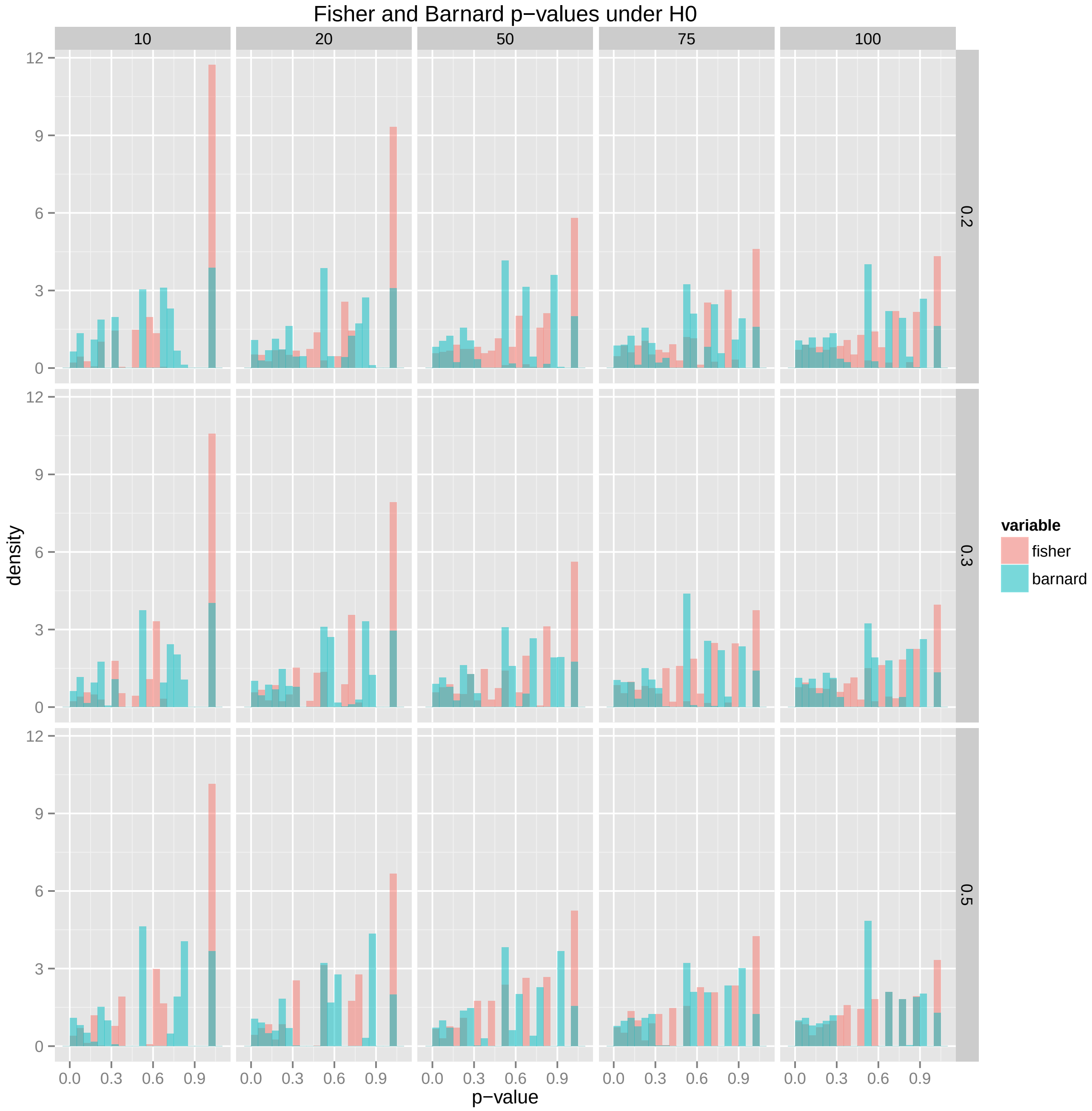
Let's compare this for a variety of sample sizes and probabilities.

Note, that some samples may not work for Barnard's test, since it requires at least one observation in each row or column. Below I place a "1" in the first column of a row with 0 counts for the sake of computation with the expectation it will not greatly distort the results since it is a rare event where both columns have the same characteristic (thus a p-value close to 1).

```r
# number of repetitions
R <- 1e3  # 1e3
n.set <- c(10, 20, 50, 75, 100)
#p.set <- c(0.05, 0.1, 0.2, 0.3, 0.5)
p.set <- c(0.2, 0.3, 0.5)

total.set <- R * length(n.set) * length(p.set)
p.values2 <- data.frame(n       = rep(NA, total.set)
                      , p       = rep(NA, total.set)
                      , fisher  = rep(NA, total.set)
                      , barnard = rep(NA, total.set)
                       )

ii.count <- 0
```

```
for (i.n in n.set) {
  for (i.p in p.set) {

    # column totals
    col.n <- c(i.n, i.n)
    # first row probabilities
    row.p <- c(i.p, i.p)

    # draw samples of Y|W and Y|X
    freq.Y <- data.frame(W = rbinom(R, col.n[1], row.p[1])
                       , X = rbinom(R, col.n[2], row.p[2])
                        )

    # if there are 0's for both columns, then replace one with a 1 so
    #   Barnard's test works
    ind.0 <- which(apply(freq.Y, 1, sum) == 0)
    freq.Y[ind.0, 1] <- 1

    for (i.R in 1:R) {
      ii.count <- ii.count + 1
      tab <- matrix(c(freq.Y[i.R, 1], col.n[1] - freq.Y[i.R, 1]
                    , freq.Y[i.R, 2], col.n[2] - freq.Y[i.R, 2])
                  , nrow = 2)
      # save values
      p.values2$n[ii.count]       <- i.n
      p.values2$p[ii.count]       <- i.p
      p.values2$fisher[ii.count]  <- fisher.test(tab)$p.value
      p.values2$barnard[ii.count] <- barnard.test(tab, to.print = FALSE, to.plot = FALSE)
    }
  }
}
#£

library(reshape2)
p.values2.long <- melt(p.values2, c("n","p"))

library(ggplot2)
p <- ggplot(p.values2.long, aes(x = value, fill = variable))
p <- p + geom_histogram(aes(y = ..density..), binwidth = 0.05, alpha = 0.5, position="identity
p <- p + facet_grid(p ~ n)
p <- p + labs(title = "Fisher and Barnard p-values under H0")
p <- p + xlab("p-value")
p <- p + ylab("density")
print(p)
```

```r
# calculate estimated Type-I error
library(plyr)
df.type1 <- ddply(p.values2.long
            , c("n", "p", "variable")
            , function(.X) {
                out <- .X[1,-4]
                out$Type_1_Error <- sum(.X$value < 0.05) / length(.X$value)
                return(out)
              }
              )

library(reshape2)
tab.type1 <- dcast(df.type1, variable + p ~ n, value.var = "Type_1_Error")
```



This table of Type 1 errors should all be about $\alpha = 0.05$.

	variable	p	10	20	50	75	100
1	fisher	0.200	0.010	0.026	0.028	0.023	0.035
2	fisher	0.300	0.011	0.028	0.028	0.042	0.038
3	fisher	0.500	0.020	0.022	0.033	0.037	0.047
4	barnard	0.200	0.032	0.054	0.041	0.043	0.053
5	barnard	0.300	0.031	0.051	0.045	0.052	0.056
6	barnard	0.500	0.055	0.053	0.036	0.039	0.050

1.3 Next steps

To calculate power for each test by performing a similar set of simulations for differences in probability.

1.3.1 Why is power important?

Consider Harry Khamis's consulting story about a (unnamed for these notes) hotel near Dayton, OH. In brief: A black woman made a reservation, arrived on the day of the reservation, and filled out the paperwork for her room. The clerk noted her address and said the hotel does not rent rooms to people who live within 25 miles of the hotel. Thinking this strange, and possibly discriminatory, she brought this case to a lawyer, who conducted a "sting" operation. Five more people went through the same sequence of events with addresses within 25 miles of the hotel, and the 3 black people were refused rooms and the 2 white people were not. Given all the observations, this is our table with significance tests of no assication between race and room rental.

```
hotel <- matrix(c(4, 0, 0, 2)
              , nrow = 2
              , dimnames = list(Hotel = c('Denied', 'Rented')
                              , Race = c('Black', 'White'))
               )
hotel
```

```
##         Race
## Hotel    Black White
##   Denied     4     0
##   Rented     0     2
fisher.test(hotel)
##
##  Fisher's Exact Test for Count Data
##
## data:  hotel
## p-value = 0.06667
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5079839       Inf
## sample estimates:
## odds ratio
##        Inf
barnard.test(hotel, to.plot = FALSE)
##
##  2x2 matrix Barnard's exact test: 100 5x3 tables were evaluated
##  -----------------------------------------------------------
##  Wald statistic =  2.4495
##  Nuisance parameter =  0.66663
##  p-values:  1-tailed =  0.021948 2-tailed =  0.043896
##  -----------------------------------------------------------
##
## [1] 0.04389575
```

Using the standard 0.05 significance level, the Fisher's test fails to reject the null while Barnard's test rejects the null. Given that the size (or level) of these tests are correct (see previous section), then we will prefer the test that has the greater probability of rejecting the null hypothesis when the null is false (that is, has greater power).

Note that Fisher's is a significance test of the null hypothesis (not intended with respect to an alternative), but the power can still be computed under a range of alternatives (analytically or via simulation).

The same strategy in the previous section to assess test size can be used to calculate test power.

The concept of this lesson is that Monte Carlo may be used to assess test size and power, and such an assessment may be critical to understand an choose among tests in particular research situations.