STA 250: Statistics
Lab 9

Pearson's chi-square test for point null hypothesis on multinomial models

In this lab we'd explore Pearson's chi-square test, which is an approximate ML test for the model $X \sim \mathsf{Multinomial}(n, p)$, $p \in \Delta_k$ for hypotheses $H_0 : p = p_0$ against $H_1 : p \neq p_0$. Pearson's test rejects H_0 if $Q(x) > c$ where

$$Q(x) = \sum_{l=1}^{k} \frac{(x_l - e_l)^2}{e_l}$$

with $e_l = np_{0l}$ denoting the expected category counts under the null hypothesis. Size calculation of this test is done by the approximate description $Q(X) \sim \chi^2(k-1)$ under the null. We'll generate hypothetical data X from the null and see whether the distribution of $Q(X)$ can indeed be approximated by $\chi^2(k-1)$. The key things to look at are (i) how large n should be for the approximation to be useful and (ii) does the quality of the approximation depends on what p_0 is?

Before proceeding let's see how to sample an x from a $\mathsf{Multinomial}(n, p_0)$ pmf. `R` provides a function `rmultinom()` in its `stats` package to do just this. You must first load the package (need to do this only once for your whole session):

```
library(stats)
```

Next, write a function `getQ()` that would take n and p_0 as inputs, sample an x from $\mathsf{Multinomial}(n, p_0)$ and calculate and return $Q(x)$:

```
getQ <- function(n, p0){
  x <- c(rmultinom(1, n, p0))
  ex.count <- n * p0
  Q.x <- sum((x - ex.count)^2 / ex.count)
  return(Q.x)
}
```

The following code helps to compare $Q(X)$ with the candidate $\chi^2(k-1)$ distribution. We essentially draw a large number of samples of $Q(x)$ with x generated from $\mathsf{Multinomial}(n, p_0)$ and then compare the histogram of the sampled values of $Q(x)$ against the pdf of $\chi^2(k-1) = \mathsf{Gamma}(\frac{k-1}{2}, \frac{1}{2})$

```
M <- 1e5
Q.samp <- replicate(M, getQ(n, p0))
Qs.max <- ceiling(max(Q.samp)) + 1
hist(Q.samp, freq = FALSE, col = "gray", border = "white", breaks = 0:Qs.max)
q.grid <- seq(0, Qs.max, .1)
lines(q.grid, dgamma(q.grid, (k - 1)/2, 1/2))
Q.chi <- rgamma(M, (k - 1)/2, 1/2)
Qc.max <- ceiling(max(Q.chi)) + 1
hist(Q.chi, freq = FALSE, add = TRUE, breaks = 0:Qc.max)
```

TASK 1. Fix $n = 20$, $k = 3$ and $p_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Use the code above to compare the actual distribution of $Q(X)$ against the proposed approximation $\chi^2(k-1)$. Is the approximation satisfactory at this n (for the given p_0)?

TASK 2. Repeat Task 1 for increasing values of $n = 30, 40, \cdots$ etc. At what n do you see a satisfactory agreement between the actual and the approximating distributions?

TASK 3. Now change p_0 to $(\frac{3}{4}, \frac{1}{8}, \frac{1}{8})$. At what n do you see a good agreement between the actual distribution of $Q(X)$ and Pearson's approximation?

TASK 4. Further change p_0 to $(\frac{7}{8}, \frac{1}{16}, \frac{1}{16})$. At what n do you see a good agreement between the actual distribution of $Q(X)$ and Pearson's approximation?

TASK 5. How is the quality of the normal approximation affected by p_0?