STA 114: 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. Would you say that the approximation kicks in at a lower n when p_0 has equal coordinates than when it does not?

Pearson's chi-square goodness-of-fit tests

Now suppose we have data $Y_1, \cdots, Y_n$ and we want to test whether Y_i's are distributed according to $\mathsf{Normal}(0,1)$. We can set this up as a point null hypothesis in a multinomial model. To see this, first split the range of the data $(-\infty, \infty)$ into k bins:

$$(a_0 = -\infty, a_1], (a_1, a_2], \cdots, (a_{k-2}, a_{k-1}], (a_{k-1}, a_k = \infty)$$

and let X_l count the number of Y_i values in the l-th bin. Then $X = (X_1, \cdots, X_k) \sim \mathsf{Multinomial}(n, p)$, $p = (p_1, \cdots, p_k) \in \Delta_k$ where p_l gives the probability that Y_1 belongs to the l-th bin. So accepting that $Y_i \overset{\text{IID}}{\sim} \mathsf{Normal}(0,1)$ would imply accepting $H_0 : p = p_0$ where $p_{0l} = \Phi(a_l) - \Phi(a_{l-1})$ with $\Phi(z) =$ the cdf of $\mathsf{Normal}(0,1)$. This hypothesis can be tested by using a size α Pearson's chi-square test.

From our study before, we would want to choose the bins so that p_0 has close to identical coordinate values [i.e., $p_0 = (1/k, 1/k, \cdots, 1/k)$]. This can be achieved by taking $a_1 = \Phi^{-1}(1/k)$, $a_2 = \Phi^{-1}(2/k)$ and so on. Also it is generally recommended that each bin must have an expected count of at least 5, i.e., $np_{0l} \geq 5$. Which means we should not go for more than $k = n/5$ bins.

We would like to see how much power (1 - type II error probability) a size 5% test of this kind has when Y_i's are NOT from the $\mathsf{Normal}(0,1)$ distribution. The code below does this when true distribution of Y_i's is the $t(d)$ distribution for some $d > 0$.

```
typeII.err <- function(n, d){
  y <- rt(n, df = d)
  k <- floor(n / 5)
  a <- qnorm(0:k/k)
  a[1] <- min(y) - 1      ## adjust the end points
  a[k + 1] <- max(y) + 1  ## to avoid -Inf, Inf
  x <- hist(y, breaks = a, plot = FALSE)$counts
  ex.ct <- n * rep(1/k, k)
  Q.x <- sum((x - ex.ct)^2 / ex.ct)
  return(Q.x < qgamma(1 - .05, (k -1)/2, 1/2))  ## TRUE = type II error
}
```

You can use `1 - mean(replicate(5e3, typeII.err(n, d)))` to approximate the power of the test at true distribution $t(d)$ for the given n.

TASK 5. Use the above codes to approximate the power of the test at $t(1)$. What n do you need to get power more than 50%?

TASK 6. Now do the same but with $d = 5$. Do you need a larger n to hit the 50% power mark? Why would you need more observations than in the case of $d = 1$?