Discussion review 3

Math 181B

1 Review:

1.1 Bernoulli distribution

$X \sim Bernoulli(p)$: $P(X=1) = p = 1 - P(X=0)$

1.2 Binomial distribution

$X \sim Binomial(n,p)$: $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$

Remark 1.1. *Some facts:*

- *If* $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} Bernoulli(p)$, $X = Y_1 + Y_2 + \cdots + Y_n \sim Binomial(n,p)$
- *If* $Y \sim Binomial(n,p)$ *and* $Z \sim Binomial(m,p)$, Y *and* Z *are independent, then* $X = Y + Z \sim Binomial(m+n, p)$.

1.3 Multinomial distribution

$X := (X_1, \ldots, X_t) \sim Multinomial\,(n, p_1, p_2, \ldots, p_t)$:

$$P\left((X_1, \ldots, X_t) = (k_1, \ldots, k_t)\right) = \binom{n}{k_1, \ldots, k_t} p_1^{k_1} \ldots p_t^{k_t};$$

where $\sum_{i=1}^t p_i = 1, \sum_{i=1}^t k_i = n$, and $\binom{n}{k_1,\ldots,k_t} = \frac{n!}{k_1! k_2! \cdots k_t!}$.

Remark 1.2. *Some facts:*

- *If* $Y_1, \cdots, Y_n \overset{i.i.d.}{\sim} Multinomial\,(1, p_1, p_2, \ldots, p_t)$, $X = Y_1 + \cdots + Y_n \sim Multinomial\,(n, p_1, p_2, \ldots, p_t)$.
- *The moment generating function of* $X \sim Multinomial\,(n, p_1, p_2, \ldots, p_t)$ *is*

$$M_X(s) = \mathbb{E}\left[e^{s^T X}\right] = \left(\mathbb{E}\left[e^{s^T Y_1}\right]\right)^n = \left(\sum_{j=1}^t p_j e^{s_j}\right)^n$$

- *If* $Y \sim Multinomial\,(n, p_1, p_2, \ldots, p_t)$ *and* $Z \sim Multinomial\,(m, p_1, p_2, \ldots, p_t)$, *then* $Y + Z \sim Multinomial\,(n+m, p_1, p_2, \ldots, p_t)$

- *The marginal distribution of X_j for any $j = 1, \ldots, t$*

$$X_j \sim Binomial\,(n, p_j)\,;$$

note that $X_1, \cdots, X_t$ are not independent due to the constraint that $X_1 + X_2 + \cdots + X_t = n$.

1.4 Goodness-of-fit test (known parameters)

1.4.1 Test setting

For continuous distribution:

$$\begin{aligned} H_0 &: f_X(x) = f_0(x) \\ H_1 &: f_X(x) \neq f_0(x) \end{aligned}$$

For discrete models with t classes:

$$\begin{aligned} H_0 &: p_1 = p_{10}, \ldots, p_t = p_{t0} \\ H_1 &: p_i \neq p_{i0} \text{ for at least one } i \end{aligned}$$

Remark 1.3. *To test the density of continuous distribution, we need to reduce data to a set of classes, i.e., separate the whole domain of the density function to several non-overlapping intervals.*

1.4.2 Test statistics

Let $r_1, r_2, \ldots, r_t$ be the set of possible outcomes (or ranges of outcomes) associated with each of n independent trials, where $P\,(r_i) = p_i, i = 1, 2, \ldots, t$. Let X_i = number of times r_i occurs, $i = 1, 2, \ldots, t$. Then, the random variable

$$D = \sum_{i=1}^{t} \frac{(X_i - np_{i0})^2}{np_{i0}}$$

has approximately a χ^2 distribution with $t-1$ degrees of freedom. For the approximation to be adequate, the t classes should be defined so that $np_i \geq 5$, for all i.

We reject the null hypothesis if

$$d = \sum_{i=1}^{t} \frac{(k_i - np_{i0})^2}{np_{i0}} \geq \chi^2_{1-\alpha, t-1};$$

where $k_1, k_2, \ldots, k_t$ be the observed frequencies for the outcomes $r_1, r_2, \ldots, r_t$.

1.5 Goodness-of-fit test (unknown parameters)

Suppose that a random sample of n independent observations is taken from $f_Y(y)$ or $p_X(k)$, a pdf having s unknown parameters. Let $r_1, r_2, \ldots, r_t$ be a set of mutually exclusive ranges (or outcomes) associated with each of the n observations. Let $\hat{p}_i$ = estimated probability of $r_i, i = 1, 2, \ldots, t$ (as calculated from $f_Y(y)$ or $p_X(k)$ after the s unknown parameters have been replaced by their maximum likelihood estimates). Let X_i denote the number of times that r_i occurs, $i = 1, 2, \ldots, t$. Then, the random variable

$$D_1 = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

has approximately a χ^2 distribution with $t-1-s$ degrees of freedom. For the approximation to be fully adequate, the r_i 's should be defined so that $n\hat{p}_i \geq 5$ for all i.

Remark 1.4. *We pay a price for having to rely on the data to fill in details about the presumed model, i.e., replacing unknown parameters with their maximum likelihood estimators. More specifically, the power of the test will decrease since the distribution of our test statistics has a fatter tail so it is harder to detect a significant effect.*