LDP two-sample chi-squared test

1 Setting

- $\boldsymbol{Y}_i \overset{iid}{\sim} multi(n_1, \boldsymbol{p_Y}), \boldsymbol{Z}_i \overset{iid}{\sim} multi(n_2, \boldsymbol{p_Z})$ with k categories
- One-hot vector form i.e. random vectors with dependent Bernoulli random variable entries
- Allow for $n_1 \neq n_2$

2 Generalized Randomized Response and two sample Pearson chi-square statistic

2.1 Privacy mechanism: Generalized Randomized Response

Definition 2.1 (Generalized Randomized Response (Theorem 5.4. of Gaboardi and Rogers [1]))**.** *For a multinomial random vector $\boldsymbol{Y}_i \overset{iid}{\sim} multi(n_1, \boldsymbol{p_Y})$, we define*

$$\mathbb{P}\big(\mathcal{M}_{\texttt{GenRR}}(\boldsymbol{Y}_i) = \boldsymbol{y}' | \boldsymbol{Y}_i = \boldsymbol{y}\big) := \begin{cases} \dfrac{\exp(\alpha)}{\exp(\alpha) + k - 1} & \text{if } \boldsymbol{y}' = \boldsymbol{y} \\ \dfrac{1}{\exp(\alpha) + k - 1} & \text{if } \boldsymbol{y}' \neq \boldsymbol{y}. \end{cases}$$

Then $\tilde{\boldsymbol{Y}}_i := \mathcal{M}_{\texttt{GenRR}}(\boldsymbol{Y}_i)$ is a multinomial random vector with probability vector

$$\tilde{\boldsymbol{p}}_{\boldsymbol{Y}} := \boldsymbol{p_Y} \frac{\exp(\alpha)}{\exp(\alpha) + k - 1} + (1 - \boldsymbol{p_Y}) \frac{1}{\exp(\alpha) + k - 1}.$$

Since $e^{\alpha} > 1$ for $\alpha > 0$, the probability of sending the original category is a little bit higher than sending the other category. Gaboardi and Rogers [1] constructs a private goodness-of-fit test based on a chi-square statistic evaluated on $\tilde{\boldsymbol{Y}}_i$'s. They demonstrate that the limiting distribution is chi-square distribution both under the null and alternative.

2.2 Two sample chi-square statistic

We extend the goodness-of-fit test by Gaboardi and Rogers [1] into two-sample testing by privatizing the raw samples $\boldsymbol{Z}_i \overset{iid}{\sim} multi(n_2, \boldsymbol{p_Z})$ into $\tilde{\boldsymbol{Z}}_j := \mathcal{M}_{\texttt{GenRR}}(\boldsymbol{Z}_j)$. Under the null, $\mathcal{M}_{\texttt{GenRR}}(\boldsymbol{Y}_i)$ and $\mathcal{M}_{\texttt{GenRR}}(\boldsymbol{Z}_j)$ follow multinomial distributions with the same probability vector. Therefore, the usual two-sample chi-square test statistic

$$T_{\chi} := \sum_{\ell=1}^{k} \frac{\big(n_2 \sum_{i=1}^{n_1} \tilde{\boldsymbol{Y}}_i(\ell) - n_1 \sum_{j=1}^{n_1} \tilde{\boldsymbol{Z}}_j(\ell)\big)^2}{n_1 n_2 (n_1 + n_2) \sum_{j=1}^{n_1} \big(\tilde{\boldsymbol{Y}}_j(\ell) + \tilde{\boldsymbol{Z}}_j(\ell)\big)}$$

converges to a chi-square distribution with degree of freedom $k - 1$ and yields a valid test with size γ. This test statistic is from Van der Vaart's book Asymptotic Statistics, pp. 253.

3 Bit flip privatization and related test statisitc

3.1 Bit flip privatization

3.2 test statistic

3.2.1 Review of one-sample statistic

We first review how [1] builds goodness-of-fit chi-square statistic for histogram of bit-flipped observations. We start with applying CLT to the bit-flipped obsevations.

Lemma 3.1 (Applying CLT to bit-flipped observations, Lemma 5.7 of [1])**.** *When $Y_i \overset{iid}{\sim} multinomial(\boldsymbol{p}, 1)$, denote the histogram of flipped observations as*

$$\tilde{\boldsymbol{H}} := \sum_{i=1}^{n_1} \mathcal{M}_{bit}(Y_i). \tag{1}$$

The mean vector and covariance matrices are computed as follows:

$$\tilde{\boldsymbol{p}} := \mathbb{E}(\mathcal{M}_{bit}(Y_1)) = \frac{(\exp(\alpha/2) - 1)\boldsymbol{p} + 1}{\exp(\alpha/2) + 1}, \text{ and} \tag{2}$$

$$\Sigma_{\boldsymbol{p}} := Var(\mathcal{M}_{bit}(Y_1)) = \left(\frac{\exp(\alpha/2) - 1}{\exp(\alpha/2) + 1}\right)^2 \left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right) + \frac{\exp(\alpha/2)}{(\exp(\alpha/2) + 1)^2} I_d, \tag{3}$$

For any $\alpha > 0$ and $\boldsymbol{p} > 0$, $\Sigma_{\boldsymbol{p}}$ is full-rank and one of its eigenvector is one-vector. By the CLT for i.i.d random vectors, we get the following asymptotic distribution:

$$\sqrt{n}(\tilde{\boldsymbol{H}}/n - \tilde{\boldsymbol{p}}) \xrightarrow{d} N\big(0, \Sigma_{\boldsymbol{p}}\big) \tag{4}$$

In non-private chi-square test, we apply CLT and multiply by $diag(\boldsymbol{p})^{-1/2}$ to turn the covariance matrix on the RHS into a projector matrix. Here in the private setting, we also need a scaling matrix to turn the covariance matrix into a projector matrix. Gaboardi and Rogers [1] proposes $\tilde{\boldsymbol{p}}^{-1/2}\Pi$, where $\Pi := I_k - \frac{1}{k}\mathbf{1}\mathbf{1}^\top$. The properties of Π are as follows:

1. It is symmetric idempotent (a projecter matrix).
2. Its null space is $span\{\mathbf{1}\}$, so when multiplied to a symmetric matrix, it deletes an eigenvector $\mathbf{1}$.

$$\begin{aligned} \Pi x = 0 &\iff x = (1/k)\mathbf{1}\mathbf{1}^\top x \\ &\iff x = (1/k)\mathbf{1}(\mathbf{1}^\top x) = ((\mathbf{1}^\top x)/k)\mathbf{1} = c\mathbf{1} \end{aligned}$$

By multiplying $\tilde{\boldsymbol{p}}^{-1/2}\Pi$ to the LHS vector of (4), we get

$$\sqrt{n}\tilde{\boldsymbol{p}}^{-1/2}\Pi(\tilde{\boldsymbol{H}}/n - \tilde{\boldsymbol{p}}) \xrightarrow{d} N\big(0, \tilde{\boldsymbol{p}}^{-1/2}\Pi\Sigma_{\boldsymbol{p}}\Pi\tilde{\boldsymbol{p}}^{-1/2}\big), \tag{5}$$

where $\tilde{\boldsymbol{p}}^{-1/2}\Pi\Sigma_{\boldsymbol{p}}\Pi\tilde{\boldsymbol{p}}^{-1/2}$ is an identity matrix except one diagonal entry is zero. Therefore, the covariance matrix is idempotent and rank $k-1$. Now we invoke the following classical theorem to derive an asymptotic chi-square distribution with degree of freedom $k-1$.

Theorem 3.1 (Ferguson (1996))**.** *If $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ and Σ is a projection matrix of rank ν an $\Sigma\boldsymbol{\mu} = \boldsymbol{\mu}$ then $\boldsymbol{X}^\top\boldsymbol{X} \sim \chi^2_\nu(\boldsymbol{\mu}^\top\boldsymbol{\mu})$.*

3.3 Extension to two sample test statistic

We can extend this lemma to two-sample setting. Suppose $Y_i \overset{iid}{\sim} multinomial(\boldsymbol{p}_1, 1)$ and $Z_j \overset{iid}{\sim} multinomial(\boldsymbol{p}_2, 1)$. We follow Lemma 3.1 to denote $\tilde{\boldsymbol{p}}_Y = \mathbb{E}(\mathcal{M}_{bit}(Y_1)), \Sigma_{\boldsymbol{p}_Y} := Var(\mathcal{M}_{bit}(Y_1))$ and $\tilde{\boldsymbol{p}}_Z = \mathbb{E}(\mathcal{M}_{bit}(Z_1)), \Sigma_{\boldsymbol{p}_Z} := Var(\mathcal{M}_{bit}(Z_1))$. Denote $\tilde{Y}_i := \mathcal{M}_{bit}(Y_i) - \tilde{\boldsymbol{p}}_Y$ and $\tilde{Z}_j := \mathcal{M}_{bit}(Z_j) - \tilde{\boldsymbol{p}}_Z$. Then denote $T_n := \sum_{i=1}^n \tilde{Y}_i - \sum_{j=1}^n \tilde{Z}_j$ and $\Sigma_n := Var(T_n) = n\big(\Sigma_{\boldsymbol{p}_Y} + \Sigma_{\boldsymbol{p}_Z}\big)$.

Under the null hypothesis of $\boldsymbol{p}_Y = \boldsymbol{p}_Z = \boldsymbol{p}$, we have $T_n = \sum_{i=1}^n \mathcal{M}_{bit}(Y_i) - \sum_{j=1}^n \mathcal{M}_{bit}(Z_j) = \tilde{\boldsymbol{H}}_Y - \tilde{\boldsymbol{H}}_Z$ and $\Sigma_n = 2n\Sigma_{\boldsymbol{p}}$. So we have

$$\sqrt{n/2}(\tilde{\boldsymbol{H}}_Y/n - \tilde{\boldsymbol{H}}_Z/n) \xrightarrow{d} N\big(0, \Sigma_{\boldsymbol{p}}\big) \tag{6}$$

$$\sqrt{n/2}(\tilde{\boldsymbol{H}}_Y/n - \tilde{\boldsymbol{H}}_Z/n) \xrightarrow{d} N\big(0, \Sigma_{\boldsymbol{p}}\big) \tag{7}$$

Since $\Sigma_{\boldsymbol{p}}$ is symmetric and one of its eigenvector is one-vector, we can diagonalize it as $\Sigma_{\boldsymbol{p}} = BDB^\top$, where D is a diagonal matrix and B has orthogonal columns with one of them being $k^{-1}\mathbf{1}$.

We introduce $\Pi := I_d - \frac{1}{k}\mathbf{1}\mathbf{1}^T$. First, this is an orthogornal projection matrix, since it is symmetric and idempotent:

$$\begin{aligned}\Pi^2 = \left(I_d - \frac{1}{k}\mathbf{1}\mathbf{1}^T\right)\left(I_d - \frac{1}{k}\mathbf{1}\mathbf{1}^T\right) &= I_d - \frac{1}{k}\mathbf{1}\mathbf{1}^T - \frac{1}{k}\mathbf{1}\mathbf{1}^T + \frac{1}{k^2}\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T \\ &= I_d - 2\frac{1}{k}\mathbf{1}\mathbf{1}^T + \frac{1}{k^2}\mathbf{1}(\mathbf{1}^T\mathbf{1})\mathbf{1}^T \\ &= I_d - 2\frac{1}{k}\mathbf{1}\mathbf{1}^T + \frac{1}{k^2}\mathbf{1}(k\mathbf{1}^T) \\ &= I_d - \frac{1}{k}\mathbf{1}\mathbf{1}^T \\ &= \Pi.\end{aligned}$$

$\mathbf{1}\mathbf{1}^T$ Since Π is symmetric, its column space is the orthogonal complement of $span\{\mathbf{1}\}$. So multiplying by Π means under the null, it suffices to use the CLT for i.i.d. random vectors, but under the alternative, we would need to use Lindeburg or Lyapunov.

$$\frac{\tilde{\boldsymbol{H}}_1}{n_1} - \frac{\tilde{\boldsymbol{H}}_2}{n_2}$$

References

[1] Gaboardi, M. and Rogers, R. (2018). Local private hypothesis testing: Chi-square tests. *Proceedings of the 35th International Conference on Machine Learning*, 80:1626–1635.