SDS 384-11 PS #4, Spring 2021

Instructor: Purnamrita Sarkar

Solutions by: Anonymous

Due: Friday, May 7, 2021 at 11:59pm
Last modified: May 12, 2021

Exercise 0.1. *Consider an ii.d sample of size n from a discrete distribution parametrized by $p_1, \ldots, p_m$ on m atoms. A common test for uniformity of the distribution is to look at the fraction of pairs that collide, or are equal. Call this statistic U.*

1. *Is U a U-statistic? When is it degenerate?*

   Solution.
   Let us denote the samples as $\{X_i\}_{i=1}^n$. We may write U as

$$U = \frac{1}{\binom{n}{2}} \sum_{i>j\in[n]} \mathbf{1}\{X_i = X_j\}$$

   Hence, by definition, U is a U -statistic of order 2 with (symmetric) kernel $h(x,y) = \mathbf{1}\{x = y\}$. Note that

$$\begin{aligned}\theta &= \mathbb{E}\left[\mathbf{1}\{X_1 = X_2\}\right] \\ &= \sum_{a\in[m]} p_a^2 \\ &= \mathbb{E}[U]\end{aligned}$$

   U is called degenerate when the limiting variance of the U -statistic, $4\xi_1 = 0$. We have that Thus, $\xi_1 = 0$ iff $p_a = p_b$ for all $a, b \in \{a \in [m] : p_a > 0\}$. That is, U is degenerate exactly when the distribution is uniform over some subset of the atoms in $[m]$ (and assigns 0 mass.to the rest of the atoms) □

2. *What is the variance of U ? Plense give the exact answer, without approximation.*

   Solution.
   Recall that, from the previous part,

$$\xi_1 = \sum_{a,b\in[m], p_a>p_b} (p_a - p_b)^2 p_a p_b$$

Additionally, we may compute

$$\begin{aligned}
\xi_2 &= \operatorname{Var}\left(\mathbb{E}\left[h\left(X_1, X_2\right) \mid X_1, X_2\right]\right) \\
&= \operatorname{Var}\left(h\left(X_1, X_2\right)\right) \\
&= \operatorname{Var}\left(\mathbf{1}\left\{X_1 = X_2\right\}\right) \\
&= \mathrm{E}\left(\mathbf{1}\left\{X_1 = X_2\right\}\right) - \left(\mathbb{E}\left[\mathbf{1}\left\{X_1 = X_2\right\}\right]\right)^2 \\
&= \sum_{a\in[m]} p_a^2 - \left(\sum_{a\in|m]} p_a^2\right)^2 \\
&= \sum_{a\in|m]} p_a^2\left(1 - p_a^2\right) - \sum_{a\neq b\in[m]} p_a^2 p_b^2
\end{aligned}$$

Therefore, the variance of U is given by

$$\begin{aligned}
\operatorname{Var}(U) &= \frac{1}{\binom{n}{2}^2} \sum_{c=0}^{2} \binom{n}{2}\binom{2}{c}\binom{n-2}{2-c}\xi_c \\
&= \frac{1}{\binom{n}{2}^2}\left(\binom{n}{2}\binom{2}{1}\binom{n-2}{1}\xi_1 + \binom{n}{2}\binom{2}{2}\binom{n-2}{0}\xi_2\right) \\
&= \frac{1}{\binom{n}{2}^2}\left(\binom{n}{2}\binom{2}{1}\binom{n-2}{1}\xi_1 + \binom{n}{2}\binom{2}{2}\binom{n-2}{0}\xi_2\right) \\
&= \frac{1}{\binom{n}{2}^2}\left(\binom{n}{2}\binom{2}{1}\binom{n-2}{1}\xi_1 + \binom{n}{2}\binom{2}{2}\binom{n-2}{0}\xi_2\right) \\
&= \frac{1}{\binom{n}{2}}\left(2\binom{n-2}{1}\xi_1 + \xi_2\right) \\
&= \frac{1}{\binom{n}{2}}\left(2(n-2)\sum_{\substack{a,k\in|m| \\ pa>pb}} \left(p_a - p_b\right)^2 p_a p_b + \sum_{a\in[m]} p_a^2\left(1 - p_a^2\right) - \sum_{a\neq b\in[m]} p_a^2 p_b^2\right)
\end{aligned}$$

□

3. *For a hypothesis test, we will consider alternative distributions which have* $p_i = \frac{1+a}{m}$ *for half of the atoms of the distribution and* $\frac{1-a}{m}$ *for the other half* $(0 \leq a \leq 1)$*, for some* $a > 0$*. Assume that there are an even number of atoms.*

   (a) What are the mean and variance of this statistic under the null?

   Solution.

Erom the previous parts, we note that

$$\begin{aligned}\mathbb{E}\left[U \mid H_0\right] &= \frac{1}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \sum_{i>j\in[n]} \sum_{a\in[m]} p_a^2 \\ &= m\frac{1}{m^2} \\ &= \frac{1}{m}\end{aligned}$$

and additionally,

$$\begin{aligned}\operatorname{Var}\left(U \mid H_0\right) &= \frac{1}{\begin{pmatrix} n \\ 2 \end{pmatrix}}\xi_2 \\ &= \frac{1}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \left(\sum_{a\in[m]} \frac{1}{m^2}\left(1-\frac{1}{m^2}\right) - 2\begin{pmatrix} m \\ 2 \end{pmatrix}\frac{1}{m^4}\right) \\ &= \frac{1}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \left(\frac{1}{m} - \frac{1}{m^3} - \frac{m-1}{m^3}\right) \\ &= \frac{1}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \left(\frac{1}{m} - \frac{1}{m^2}\right) \\ &= \frac{p_1\left(1-p_1\right)}{\begin{pmatrix} n \\ 2 \end{pmatrix}}\end{aligned}$$

□

(b) What are the mean and variance of this under the alternative?

Solution.
Similarly, plugging into our equations,

$$\begin{aligned}\mathbb{E}\left[U \mid H_a\right] &= \frac{1}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \sum_{i>j} \sum_{a\in[m]} p_a^2 \\ &= \frac{m}{2}\left(\frac{(1-a)^2}{m^2} + \frac{(1+a)^2}{m^2}\right) \\ &= \frac{1+a^2}{m}\end{aligned}$$

and, noting that

$$\begin{aligned}\xi_1 &= \left(\frac{m}{2}\right)\left(\frac{2a}{m}\right)^2\left(\frac{(1-a)(1+a)}{m^2}\right) \\ &= \frac{a^2\left(1-a^2\right)}{m^2}\end{aligned}$$

and

$$\begin{aligned}
\xi_2 =& \frac{m}{2}\left(\frac{(1-a)^2}{m^2}\left(1-\frac{(1-a)^2}{m^2}\right)+\frac{(1+a)^2}{m^2}\left(1-\frac{(1+a)^2}{m^2}\right)\right) \\
&-2\left(\binom{m/2}{2}\left(\frac{(1-a)^4}{m^4}+\frac{(1+a)^4}{m^4}\right)+\frac{m^2}{4}\frac{(1-a)^2(1+a)^2}{m^4}\right) \\
=& \frac{m}{2}\left(\frac{2+2a^2}{m^2}-\frac{2+12a^2+2a^4}{m^4}\right)-2\left(\frac{m}{4}\left(\frac{m}{2}-1\right)\frac{2+12a^2+2a^4}{m^4}+\frac{m^4}{4}\frac{1-2a^2+a^4}{m^4}\right) \\
=& \frac{1+a^2}{m}\left(1-\frac{1+a^2}{m}\right)
\end{aligned}$$

Therefore, we have that

$$\operatorname{Var}\left(U \mid H_a\right)=\frac{2}{n(n-1)}\left(2(n-2)\frac{a^2\left(1-a^2\right)}{m^2}+\frac{1+a^2}{m}\left(1-\frac{1+a^2}{m}\right)\right)$$

□

(c) What is the asymptotic distribution of U under the null hypothesis that $p_i=\frac{1}{m}$?
Hint: you can use the fact that for $X_1,\ldots,X_N \overset{iid}{\sim} \text{multinomial}\left(q_1,\ldots,q_k\right), \sum_{i=1}^{k}\frac{(N_i-Nq_i)^2}{Nq_i}\xrightarrow{d}\chi^2_{k-1}$

Solution.

Observe that

$$
\begin{aligned}
U &= \frac{1}{\left(\begin{array}{c} n \\ 2 \end{array}\right)} \sum_{i>j} \mathbf{1}\left\{X_i = X_j\right\} \\
&= \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{1}\left\{X_i = X_j\right\} \\
&= \frac{1}{n(n-1)} \sum_{a \in [m]} \sum_{i \neq j} \mathbf{1}\left\{X_i = a\right\} \mathbf{1}\left\{X_j = a\right\} \\
&= \frac{1}{n(n-1)} \sum_{a \in [m]} \left[\underbrace{(\sum_{i=1}^{n} \mathbf{1}\left\{X_i = a\right\}}_{=N_a})^2 - \sum_{i=1}^{n} \mathbf{1}\left\{X_i = a\right\} \right] \\
&= \frac{1}{n(n-1)} \sum_{a \in [m]} N_a^2 - N_a \\
&= \frac{1}{n(n-1)} \left(\sum_{a \in [m]} N_a^2 \right) - \frac{1}{n-1} \qquad \text{since } \sum_{a} N_a = n \text{ a.s} \\
&= \frac{1}{(n-1)} \left(\sum_{a \in [m]} \left(\frac{\sqrt{n} N_a}{n} \right)^2 \right) - \frac{1}{n-1} \\
&= \frac{1}{(n-1)} \left(\sum_{a \in [m]} \left(\sqrt{n} F_a \right)^2 \right) - \frac{1}{n-1} \\
&= \frac{n}{(n-1)} \left(\sum_{a \in [m]} F_a^2 \right) - \frac{1}{n-1}
\end{aligned}
$$

last step taking $F_a = \frac{N_a}{n}$

Additionally, under the null hypothesis, by the hint,

$$\begin{aligned}
V &= \sum_{a\in[m]} \frac{(N_i - np_a)^2}{np_a} \\
&= \sum_{a\in[m]} \frac{1}{p_a}\left(\frac{\sqrt{n}N_a}{n} - \sqrt{n}p_a\right)^2 \\
&= \sum_{a\in[m]} \frac{n}{p_a}(F_a - p_a)^2 \\
&= \sum_{a\in[m]} \frac{n}{p_a}\left(F_a^2 + p_a^2 - 2p_aF_a\right) \\
&= \sum_{a\in[m]} \frac{n}{p_a}F_a^2 + np_a - 2nF_a \\
&= \sum_{a\in[m]} \frac{n}{p_a}F_a^2 + n\underbrace{\sum_{a\in[m]} p_a}_{=1} -2n\underbrace{\sum_{a\in[m]} F_a}_{=1} \\
&= \left(\sum_{a\in[m]} \frac{n}{p_a}F_a^2\right) - n \\
&= nm\left(\sum_{a\in[m]} F_a^2\right) - n \\
&\overset{d}{\to} \chi^2_{m-1}
\end{aligned}$$

Now, observe that, by the above calculations, we have the following:

$$\begin{aligned}
m(n-m)\left(U - \mathbb{E}\left[U \mid H_0\right]\right) &= m(n-m)\left(U - \frac{1}{m}\right) \\
&= m[(n-1)U + 1] - n \\
&= nm\left(\sum_{a\in[m]} F_a^2\right) - n \\
&= V \\
&\overset{d}{\to} \chi^2_{m-1}
\end{aligned}$$

□

(d) Under the alternative hypothesis, is it always the case that U has a limiting normal distribution? Can you give a sufficient condition on the sample size n so that this is true?

Solution.

By normal convergence of U -statistics, since $\mathbb{E}\left[h^2\right] = \mathbb{E}\left[\nVdash\{X_1 = X_2\}\right] = \dfrac{1+a^2}{m} < \infty,$ we have that, for $a \in (0,1)$, and treating m as a constant that does not scale with n,

$$\sqrt{n}\left(U-\mathbb{E}\left[U \mid H_a\right]\right) \xrightarrow{d} \mathcal{N}\left(0, 4 \xi_1^2\right)$$

where

$$\xi_1^2=\frac{a^2\left(1-a^2\right)}{m^2}$$

Note that, when $a \in \{0,1\}$, the $U - statisticisdegenerate$. In these cases, we note that the distribution is uniform over $\frac{m}{2}$ atoms when $a=1$ and uniform over m atoms when $a=0$. Thus, in both of these cases, the convergence result from the previous section is applicable (with m replaced by $\frac{m}{2}$ for $a=1$). Thus, in these cases, the limiting distribution is not normal. Additionally, let us refer back to the variance under H_a :

$$\operatorname{Var}\left(U \mid H_a\right)=\frac{2}{n(n-1)}\left(2(n-2) \frac{a^2\left(1-a^2\right)}{m^2}+\frac{1+a^2}{m}\left(1-\frac{1+a^2}{m}\right)\right)$$

Note that we can break this term into two terms, the first term (to which ξ_1 contributes) scales as $\frac{1}{nm^2}$, and the second term (to which ξ_2 contributes), scales as $\frac{1}{n^2 m}$ Consider a regime in which $m = n^c$ for some positive constant c. Then, in order for the ξ_1 term to dominate, we need that $c < 1$. Indeed, this ensures that $1 + 2c < 2 + c$, and thus, that the first term will dominate. As noted in the updated homework document, this implies normal convergence in the case where $m = o(n)$. □

Exercise 0.2. *Compute the VC dimension of the following function classes*

1. *Circles in* $\mathcal{R}^2$

   Solution.
   For any three points which are not collinear , we can easily draw a circle that includes all three of them, any two of them, any one of them, or none of them, so VC dimension is at least 3.

   However, for any set of four points, they are not shattered. We show this by constructing a counterexample in several cases:

   (a) Collinear: the labeling +-+- (going along the line) is impossible, among numerous others.

   (b) Convex hull is a triangle: then the labeling with +(the three points of the triangle) and -(the interior point) is not possible.

   (c) Convex hull is a quadrilateral: let (x_1, x_2) be the points separated by the long diagonal and (y_1, y_2) be the points separated by the short diagonal. At least one of the labelings $\{+x_1, +x_2, -y_1, -y_2\}$, $\{-x_1, -x_2, +y_1, +y_2\}$ will not be achieved. If they were both possible, then this would mean that the circles satisfying the two labelings can have four non-overlapping regions, which is not possible. (is it possible with ellipses?)Since some set of 3 points is shattered by the class of circles, and no set of 4 points is, the VC dimension of the class of circles is 3.

   □

2. *Axis aligned rectangles in $\mathcal{R}^2$*

   Solution.
   It is easy to see that one can shatter four points. Consider 5 points and the following cases.

   (a) They are all collinear. In which case, a trivial alternative labels cannot be shattered by axis aligned rectangles.

   (b) If they are not all collinear, then draw the largest rectangle through the largest and smallest x and y coordinates. Either all five points are on this rectangle, or one is inside. In the first case, do an alternative labeling, this cannot be shattered by a axis aligned rectangle. In the second case, label everyone on the rectangle as one label, and the one inside with the opposite label.

   □

3. *Axis aligned squares in $\mathcal{R}^2$*

   Solution.
   It is easy to construct 3 points which are shattered. Let us take 4 points. Let the leftmost point be L, rightmost R, top one T and bottom one B. Draw a rectangle like last question through these points. If there are ties, then it is easy to label them so that they cannot be shattered. So let us think about the case where there are no ties. In this case, if the rectangle is a square, i.e. the distance between x coordinates of L and R (d_x) is equal to the distance between y coordinates of T and B (d_y) are such that $d_x \geq d_y$, then (L+, R+, T-, B-) cannot be shattered. If $d_x < d_y$ then (L-, R-, T+, B+) cannot be shattered. So VC dimension is 4. □

Exercise 0.3. *We will find the covering number of ellipses in this problem. Given a collection of positive numbers $\{\mu_j\}_{j=1}^d$, consider the ellipse*

$$\mathcal{E} = \{\theta \in \mathbb{RR}^d : \sum_i \theta_i^2/\mu_i^2 \leq 1\}$$

1. *Show that*

$$\log N(\epsilon; \mathcal{E}, \|\cdot\|_2) \geq d \log \frac{1}{\epsilon} + \sum_{j=1}^d \log \mu_j$$

   Solution.
   Suppose that $\{\theta_1, \ldots, \theta_N\}$ is an ϵ-cover of $\mathcal{E}$. Then, by definition, $\mathcal{E} \subset \cup_{i=1}^N \mathcal{B}_\epsilon(\theta_i)$, where $\mathcal{B}_\epsilon(\theta_i) = \{\|\theta - \theta_i\|_2 \leq \epsilon : \theta \in \mathbb{RR}^d\}$. Thus, we have that

$$\begin{aligned}\text{Vol}(\mathcal{E}) &\leq \sum_{i=1}^N \text{Vol}(\mathcal{B}_\epsilon(\theta_i)) \\ &= N\text{Vol}(\mathcal{B}_\epsilon(\mathbf{0}))\end{aligned}$$

Now, let us consider the change of coordinates from points in the ellipsoid to points in the ball. Given coordinates $\{u_i\}_{i=1}^d$ from the ϵ-ball, we may map these coordinates in a one-to-one manner to points $\{x_i\}_{i=1}^d$ in $\mathcal{E}$ by the formula:

$$x_i = \frac{\mu_i}{\epsilon} u_i$$

Indeed, since by definition $\sum_i u_i^2 \leq \epsilon^2$, and so

$$\begin{aligned} \epsilon^2 &\geq \sum_i u_i^2 = \sum_i \frac{\epsilon^2}{\mu_i^2} x_i^2 \\ \implies & \sum_i \frac{x_i^2}{\mu_i^2} \leq 1 \end{aligned}$$

as desired. Therefore, we may compute the volume of $\mathcal{E}$ using the change of variable formula

$$\begin{aligned} \mathrm{Vol}(\mathcal{E}) &= \int_{\mathcal{E}} dx_1, \dots, x_n \\ &= \int_{\mathcal{B}_\epsilon(\mathbf{0})} \left| \frac{\partial(x_1, \dots, x_n)}{\partial(u_1, \dots, u_n)} \right| du_1, \dots, u_n \\ &= \int_{\mathcal{B}_\epsilon(\mathbf{0})} \left(\prod_{i=1}^d \frac{\mu_i}{\epsilon} \right) du_1, \dots, u_n \\ &= \left(\prod_{i=1}^d \frac{\mu_i}{\epsilon} \right) \mathrm{Vol}(\mathcal{B}_\epsilon(\mathbf{0})) \end{aligned}$$

Hence,

$$\begin{aligned} \left(\prod_{i=1}^d \frac{\mu_i}{\epsilon} \right) \mathrm{Vol}(B_\epsilon(\mathbf{0})) &= \mathrm{Vol}(\mathcal{E}) \\ &\leq N \mathrm{Vol}(\mathcal{B}_\epsilon(\mathbf{0})) \end{aligned}$$

and thus,

$$\begin{aligned} N &\geq \prod_{i=1}^d \frac{\mu_i}{\epsilon} \\ \implies \log N &\geq d \log \frac{1}{\epsilon} + \sum_{i=1}^d \log \mu_i \end{aligned}$$

as desired. □

2. *Now consider the infinite-dimensional ellipse, specified by the sequence $\mu_j = j^{-2\beta}$ for some parameter $\beta > \frac{1}{2}$. Show that*

$$\log N(\epsilon; \mathcal{E}, \|\cdot\|_2) \geq C \left(\frac{1}{\epsilon} \right)^{1/2\beta}$$

where $\|\theta - \theta'\|_{\ell_2}^2 = \sum_{i=1}^\infty (\theta(i) - \theta(i)')^2$.

Solution.
Let us denote the ellipse truncated to d dimensions as:

$$\mathcal{E}_d = \{\tilde{\theta} \in \mathbb{RR}^d : \theta \in \mathcal{E}, \tilde{\theta}(i) = \theta(i) \forall i \in [d]\}$$

Let $S = \{\theta_1, \ldots, \theta_N\}$ be an ϵ-covering of $\mathcal{E}$. Define S_d as the elements of S truncated to d dimensions, that is, the set of N elements $\tilde{\theta}_i$ such that $\tilde{\theta}_i(j) = \theta_i(j)$ for $j \in [d]$.

Now, we will show that S_d is an ϵ-covering of $\mathcal{E}_d$. Indeed, fix any $\tilde{\theta} \in \mathcal{E}_d$. By definition, there is some θ such that $\tilde{\theta}(j) = \theta(j)$ for every $j \in [d]$. By definition of S, there exists some θ_i satisfying $\|\theta - \theta_i\|_{\ell_2} \leq \epsilon$. Therefore,

$$\begin{aligned}
\epsilon^2 &\geq \|\theta - \theta_i\|_{\ell_2}^2 \\
&= \sum_{j=1}^{d} (\theta(i) - \theta_i(j))^2 + \sum_{j=d+1}^{\infty} (\theta(i) - \theta_i(j))^2 \\
&= \sum_{j=1}^{d} (\tilde{\theta}(i) - \tilde{\theta}_i(j))^2 + \sum_{j=d+1}^{\infty} (\theta(i) - \theta_i(j))^2 \\
&\geq \sum_{j=1}^{d} (\tilde{\theta}(i) - \tilde{\theta}_i(j))^2 + \sum_{j=d+1}^{\infty} (0 - 0)^2 \\
&= \|\tilde{\theta} - \tilde{\theta}_i\|_2^2
\end{aligned}$$

and thus S_d is also an ϵ-cover of $\mathcal{E}_d$. Therefore, we have that

$$\begin{aligned}
\log N(\epsilon; \mathcal{E}, \|\cdot\|_2) &\geq \log N(\epsilon, \mathcal{E}_d, \|\cdot\|_2) \\
&\geq d \log \frac{1}{\epsilon} + \sum_{i=1}^{d} \log \mu_i && \text{by the previous problem} \\
&\geq d \log \frac{1}{\epsilon} - 2\beta \log d! \\
&\geq d \log \frac{1}{\epsilon} - 2\beta \log(d^{d+1/2} e^{-d+1}) && \text{by Sterling's approximation} \\
&= d \log \frac{1}{\epsilon} - 2\beta d \log d + 2\beta \left(d - 1 + \frac{1}{2} \log d \right)
\end{aligned}$$

Now, choose $d = \left\lceil \left(\frac{1}{\epsilon}\right)^{1/2\beta} \right\rceil$. Then the above inequality becomes

$$\begin{aligned}
\log N(\epsilon; \mathcal{E}, \|\cdot\|_2) &\geq d \log \frac{1}{\epsilon} - 2\beta d \underbrace{\log \left(\left(\frac{1}{\epsilon} \right)^{1/2\beta} + 1 \right)}_{\leq \frac{1}{2\beta} \log\left(\frac{1}{\epsilon}\right) + \frac{1}{2}} + 2\beta \left(d - 1 + \frac{1}{2} \underbrace{\log d}_{\geq 0} \right) \\
&\geq \beta\,(d - 2) \\
&\geq C\beta d && \text{for } C < 1 \text{ small enough} \\
&\geq C\beta \left(\frac{1}{\epsilon} \right)^{1/2\beta}
\end{aligned}$$

as desired. □