COMP 585 **Spring 2025**
Probabilistic Toolkit for Learning and Computing *Instructor: Maryam Aliakbarpour*
January 28, 2025 *Scriber: Maryam Aliakbarpour*

Lecture 4

Property Testing of Distributions: Uniformity testing

In statistical inference, we often have limited access to the complete population and must rely on samples. We model the population as a probability distribution, which provides a mathematical framework for representing outcome probabilities. In distribution testing, we aim to efficiently determine, based on observed samples, whether this distribution has a certain property or is far from having it. This aligns with the core objective of statistical inference: inferring properties of an unknown population distribution from data.

Problem Definition

What is a property? A property $\mathcal{P}$ is a set of distributions. If a distribution $p \in \mathcal{P}$, then p has the property; otherwise, it does not. For instance, $\mathcal{P}$ could be the set of unimodal distributions or the set containing only the uniform distribution on $[n]$, denoted as U_n.

Distance measure: We need to define what it means for a distribution to be far from having a property. First, consider the distance between two distributions. Common examples include:

- ℓ_1-distance: $\|p-q\|_1 = \sum_{x\in\Omega} |p_x - q_x|$
- ℓ_2-distance: $\|p-q\|_2 = \sqrt{\sum_{x\in\Omega} (p_x - q_x)^2}$
- Total Variation distance: $\|p-q\|_{\mathrm{TV}} = \max_{E\subseteq\Omega} \left|\mathbf{Pr}_{x\sim p}[x \in E] - \mathbf{Pr}_{x\sim q}[x\in E]\right|$

where p and q are two distributions over a discrete domain Ω. A key relationship between distances is:

$$\|p-q\|_1 = 2\,\|p-q\|_{\mathrm{TV}}\,.$$

Exercise: Prove the above identity. The distance between a distribution p and a property $\mathcal{P}$ is the distance between p and its closest distribution in $\mathcal{P}$:

$$\mathrm{dist}(p,\mathcal{P}) \coloneqq \inf_{q\in\mathcal{P}} \mathrm{dist}(p,q)\,.$$

We can use various distance notions instead of dist here. We say p is *ϵ-close* to $\mathcal{P}$ if $\text{dist}(p, \mathcal{P}) \leq \epsilon$, and *$\epsilon$-far* if $\text{dist}(p, \mathcal{P}) > \epsilon$.

What do we want to distinguish? Our goal is to distinguish distributions that have a property from those that are *far* from having it. For example, consider a distribution almost perfectly uniform over $[n]$, except for a tiny probability of outputting the element 1. Finding such negligible differences can require many samples. Therefore, we aim to distinguish whether $p \in \mathcal{P}$ or p is ϵ-far from $\mathcal{P}$.

For distributions that are neither in $\mathcal{P}$ nor ϵ-far from $\mathcal{P}$ (i.e., ϵ-close but not in $\mathcal{P}$), either answer is considered valid. We accept the compromise of misclassifying distributions that are close to having the property, even if they don't have it exactly.

Tester: Consider three parameters $\epsilon \in (0, 1), \delta \in (0, 1)$, and $n \in \mathbb{N}$. Suppose an algorithm $\mathcal{A}$ receives these parameters and m-samples from a discrete distribution p over $[n]$ as its input and produce an output in $\{\textsf{accept}, \textsf{reject}\}$. We say $\mathcal{A}$ is an *(ϵ, δ)-tester for property $\mathcal{P}$*, if the following holds with probability $1 - \delta$:

1. If $p \in \mathcal{P}$, then $\mathcal{A}$ outputs accept.
2. If p is ϵ-far from $\mathcal{P}$ (in ℓ_1-distance), then $\mathcal{A}$ outputs reject.

m is considered as the sample complexity of the algorithm.

Uniformity Testing

The goal of uniformity testing is to design an algorithm that takes samples from a distribution and determines whether the distribution is uniform or ϵ-far from uniform. Consider the following sets of samples:

Scenario 1: 9, 2, 8, 5, 1, 5, 3
Scenario 2: 6, 4, 6, 4, 1, 1, 4

Could you guess one of this set of samples is drawn from a uniform distribution? Can you find out what gives it away?

The intuition that we would like to highlight from this example is that If we draw samples from a uniform distribution, we expect to see fewer repetitions compared to samples drawn from a non-uniform distribution. Based on this intuition, we design an algorithm that counts the number of repeated pair of samples or what we refer to as collisions. Consider two samples drawn from p as X_i and X_j. We consider this pair as a collision if $X_i = X_j$. We use $\sigma_{i,j}$ as the indicator variable for this event:

$$\sigma_{i,j} := \begin{cases} 1 & \text{if } X_i = X_j, \\ 0 & \text{otherwise.} \end{cases}$$

Now, in a sample set $\{X_1, \ldots, X_m\}$, the total number of collision can be a good indicator of uniformity. We formalize this argument in Algorithm 1. For some number of samples, for now denoted by m, and some sufficiently large threshold t, we count the number of collisions. We normalize this number by dividing it by the total number of pairs of samples, $\binom{m}{2}$ to obtain a value between 0 and 1. Then, we compare it with the threshold t. If the number of collisions are high we infer that the distribution is not uniform. Otherwise, it is uniform. Our goal is to determine what m and t should be so that we get an $(\epsilon, \delta = 0.1)$-tester for uniformity $(\mathcal{P} = \{U_n\})$.

Algorithm 1 Collision-Based Uniformity Tester

```
procedure COLLISION-TESTER($\epsilon, \delta, n$, sample access to $p$)
    Draw $m$ samples from the distribution $p$: $X_1, X_2, ..., X_m$.
    for $i = 1$ to $m - 1$ do
        for $j = i + 1$ to $m$ do
            if $X_i = X_j$ then
                $\sigma_{i,j} \leftarrow 1$
            else
                $\sigma_{i,j} \leftarrow 0$
    $Y \leftarrow \frac{\sum_{i<j} \sigma_{i,j}}{\binom{m}{2}}$
    if $Y < t$ then
        return accept
    else
        return reject
```

Here, we consider two cases: when the underlying distribution is uniform, denoted by p_1, and when it is far from uniform, denoted by p_2. Figure 1 illustrates the probability density function (PDF) of Y under these two scenarios. When p_1 is uniform, Y has a low expectation, $\mathbf{E}_{p_1}[Y]$. Conversely, when p_2 is far from uniform, Y has a high expectation, $\mathbf{E}_{p_2}[Y]$. We set the threshold t at the midpoint between these expectations. The concentration of Y around its respective expectation ensures that deviations beyond t are improbable. Specifically, the blue shaded area shows the unlikely event of $Y > t$ when the underlying distribution is uniform, and the red shaded area shows the unlikely event of $Y < t$ when it is far from uniform.

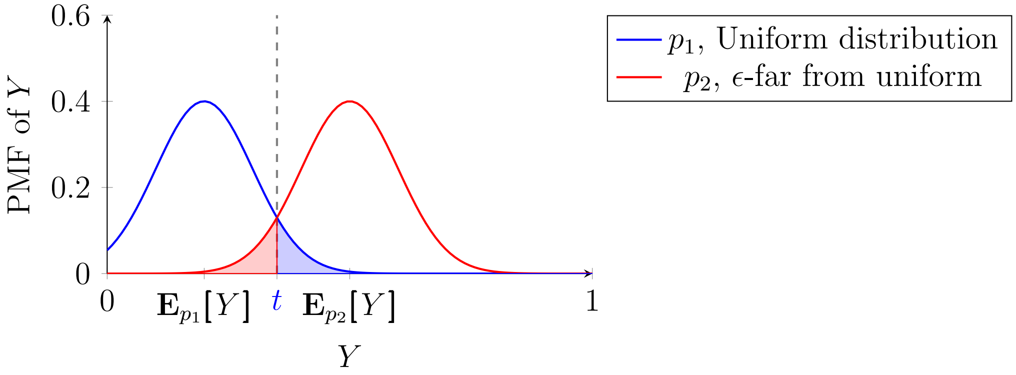

Figure 1: Separation of the PMFs of Y when the underlying distribution is uniform versus far from uniform.

Expected values

Let's first analyze the expected value of Y. The expected value of $\sigma_{i,j}$ is:

$$\begin{aligned}\mathbf{E}[\sigma_{i,j}] &= \sum_{\ell=1}^{n} \mathbf{Pr}[X_i = X_j = \ell] = \sum_{\ell=1}^{n} \mathbf{Pr}[X_i = \ell] \cdot \mathbf{Pr}[X_j = \ell] \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{(Since samples are independent.)} \\ &= \sum_{\ell=1}^{n} p_\ell^2 = \|p\|_2^2 \ .\end{aligned}$$

Therefore, the expected value of Y is:

$$\mathbf{E}[Y] = \frac{1}{\binom{m}{2}} \sum_{i<j} \mathbf{E}[\sigma_{i,j}] = \|p\|_2^2 \,. \tag{1}$$

We can relate this to the ℓ_2-distance of p from the uniform distribution:

$$\begin{aligned}\mathbf{E}[Y] = \|p\|_2^2 = \sum_{\ell=1}^{n} p_\ell^2 = \sum_{\ell=1}^{n} \left(p_\ell - \frac{1}{n} + \frac{1}{n}\right)^2 &= \sum_{\ell=1}^{n} \left(p_\ell - \frac{1}{n}\right)^2 + 2\left(p_\ell - \frac{1}{n}\right) \cdot \frac{1}{n} + \frac{1}{n^2} \\ = \|p - U_n\|_2^2 + \frac{2}{n}\left(\sum_{\ell=1}^{n} p_\ell - \frac{1}{n}\right) + \frac{1}{n} & \qquad (2) \\ = \frac{1}{n} + \|p - U_n\|_2^2 \ .\end{aligned}$$

Note that in the last equality we used the fact that $\sum_{\ell=1}^{n} p_\ell$ is one. This shows that the expected value of Y is directly related to the squared ℓ_2 distance between the distribution p and the uniform distribution. The farther p is from uniform, the larger the expected value of Y.

Now, when the underlying distribution is the uniform distribution, we have:

$$\mathbf{E}_{p_1}[Y] = \frac{1}{n}.$$

On the other hand, when p is ϵ-far from the uniform distribution (i.e., the total variation distance between p and the uniform distribution is at least ϵ), we have:

$$\mathbf{E}_{p_2}[Y] = \frac{1}{n} + \|p_2 - U_n\|_2^2 \geq \frac{1}{n} + \frac{\|p_2 - U_n\|_1^2}{n} \geq \frac{1+\epsilon^2}{n}\,.$$

The first inequality is due to the Cauchy-Schwarz inequality:

$$\left(\sum_{\ell=1}^{n}\left((p_2)_\ell - \frac{1}{n}\right)^2\right)\cdot\left(\sum_{\ell=1}^{n}(1)^2\right) \geq \sum_{\ell=1}^{n}\left|(p_2)_\ell - \frac{1}{n}\right|\cdot 1$$

$$\Rightarrow \qquad \|p_2 - U_n\|_2^2 \cdot n \geq \|p_2 - U_n\|_1^2\,.$$

The second inequality is due to the fact that p_2 is ϵ-far from being a uniform distribution.

The above analysis demonstrates a separation between the expected value of Y based on the uniformity of the underlying distribution. We set our threshold to be between these two bounds to account for the potential deviation of Y from its expectation:

$$t \coloneqq \frac{1+\epsilon^2/2}{n}\,.$$

Concentration bounds

To establish concentration bounds for Y, we can employ concentration bounds. Note that Y is not a sum of independent random variables, making it unsuitable for a direct application of the Chernoff bound. Our plan is to use Chebyshev's inequality, which requires us to compute the variance of Y. We state the following lemma:

Lemma 1.

$$\mathbf{Var}[Y] = \frac{1}{\binom{m}{2}^2}\cdot\left(\binom{m}{2}\|p\|_2^2 + 6\binom{m}{3}\|p\|_3^3\right)\,.$$

We will defer the proof of this lemma for later. In the following, we bound the probability that Y goes beyond threshold t.

Case 1: the uniform distribution. In this case, we have:

$$\begin{aligned}
\mathbf{Pr}_{p_1}[Y \geq t] \leq \mathbf{Pr}\left[|Y - \mathbf{E}_{p_1}[Y]| \geq \frac{\epsilon^2}{2n}\right] &\leq \frac{\mathbf{Var}[Y]}{\left(\frac{\epsilon^2}{2n}\right)^2} && \text{(By Chebyshev's inequality)} \\
&= \frac{1}{\binom{m}{2}^2} \cdot \left(\binom{m}{2} \|p_1\|_2^2 + 6\binom{m}{3} \|p_1\|_3^3\right) \cdot \frac{4n^2}{\epsilon^4} \\
&= \Theta\left(\frac{n^2}{m^4\epsilon^4} \cdot \left(m^2 \cdot \frac{1}{n} + \frac{m^3}{n^2}\right)\right) && \text{(Since } p_1 \text{ is uniform.)} \\
&= \Theta\left(\frac{n}{m^2\epsilon^4} + \frac{1}{m\epsilon^4}\right) \leq 0.1,
\end{aligned}$$

if $m \geq c \cdot \left(\frac{\sqrt{n}}{\epsilon^2} + \frac{1}{\epsilon^4}\right)$ for a sufficiently large constant c.

Case 2: a distribution far from uniform. Consider p_2 for which we have $\|p_2 - U_n\|_1 > \epsilon$. In this case, the bound on the variance can be large. Specifically, the term $\binom{m}{2} \|p\|_2^2 + 6\binom{m}{3} \|p\|_3^3$ could be problematic if we require $|Y - \mathbf{E}[Y]| \leq \frac{\epsilon^2}{2n}$. To address this, we adjust the bound accordingly:

$$\mathbf{Pr}_{p_2}[Y < t] = \mathbf{Pr}\left[\mathbf{E}[Y] - Y > \mathbf{E}[Y] - \frac{1 + \epsilon^2/2}{n}\right]$$

Note that using Equation 2, we have:

$$\begin{aligned}
\left(1 + \frac{\epsilon^2}{2}\right) \cdot \mathbf{E}[Y] &\geq \left(1 + \frac{\epsilon^2}{2}\right) \cdot \frac{1}{n} \\
\Rightarrow \qquad \mathbf{E}[Y] - \frac{1 + \epsilon^2/2}{n} &\geq \frac{\epsilon^2}{2} \cdot \mathbf{E}[Y]
\end{aligned}$$

Thus, we obtain:

$$\begin{aligned}
\mathbf{Pr}_{p_2}[Y < t] &= \mathbf{Pr}\left[\mathbf{E}[Y] - Y > \mathbf{E}[Y] - \frac{1+\epsilon^2/2}{n}\right] \\
&\le \mathbf{Pr}\left[\mathbf{E}[Y] - Y > \frac{\epsilon^2}{2} \cdot \mathbf{E}[Y]\right] \le \frac{4\,\mathbf{Var}[Y]}{\epsilon^4\,\mathbf{E}[Y]^2} \qquad \text{(By Chebyshev's inequality)} \\
&\le \frac{4}{\binom{m}{2}^2} \cdot \frac{\binom{m}{2}\|p\|_2^2 + 6\binom{m}{3}\|p\|_3^3}{\epsilon^4\,\|p\|_2^4} \\
&= \Theta\left(\frac{1}{m^2 \cdot \epsilon^4\,\|p\|_2^2} + \frac{\|p\|_3^3}{m \cdot \epsilon^4\,\|p\|_2^4}\right) \\
&\le \Theta\left(\frac{n}{m^2\epsilon^4} + \frac{\sqrt{n}}{m\epsilon^4}\right) \le 0.1, \\
&\text{(Using } \|p\|_2^2 \ge \tfrac{1}{n} \text{ (implied by Eq. 2), and } \|p\|_3^3 \le \|p\|_2^3 \text{ (implied by Fact 2))}
\end{aligned}$$

if $m \ge c \cdot \frac{\sqrt{n}}{\epsilon^4}$ for a sufficiently large constant c. Here, we used the following inequality that is known as ℓ_p-norm inequality:

Fact 2 (ℓ_p-norm Inequality for Distributions)**.** *For any probability distribution $d = (d_1, d_2, ..., d_n)$ over $[n]$ and $1 \le q \le p \le \infty$, the following inequality holds:*

$$\|d\|_p \le \|d\|_q$$

where $\|d\|_p$ is the ℓ_p-norm of the distribution p, defined as:

$$\|d\|_p = \left(\sum_{i=1}^{n} |d_i|^p\right)^{1/p}$$

Putting all these pieces together, we have presented a collision-based uniformity tester and analyzed its sample complexity. We have shown that the tester (Algorithm 1) requires $O\left(\frac{\sqrt{n}}{\epsilon^4}\right)$ samples to distinguish between uniform and ϵ-far from uniform distributions with probability at least $1 - \delta = 0.9$.

Variance bound: Proof of Lemma 1

Lemma 1.

$$\mathbf{Var}[Y] = \frac{1}{\binom{m}{2}^2} \cdot \left(\binom{m}{2}\|p\|_2^2 + 6\binom{m}{3}\|p\|_3^3\right).$$

Proof.

$$\begin{aligned}
\mathbf{Var}[Y] &= \mathrm{Var}\left(\frac{1}{\binom{m}{2}}\sum_{i<j}\sigma_{i,j}\right)\\
&= \frac{1}{\binom{m}{2}^2}\mathrm{Var}\left(\sum_{i<j}\sigma_{i,j}\right) && \text{(Since } \mathrm{Var}(aX)=a^2\mathrm{Var}(X) \text{ for any constant } a\text{)}\\
&= \frac{1}{\binom{m}{2}^2}\left(\mathbf{E}\left[\left(\sum_{i<j}\sigma_{i,j}\right)^2\right]-\left(\sum_{i<j}\mathbf{E}[\sigma_{i,j}]\right)^2\right)\\
&= \frac{1}{\binom{m}{2}^2}\left(\sum_{i<j}\sum_{\ell<k}\mathbf{E}[\sigma_{i,j}\cdot\sigma_{\ell,k}]-\|p\|_2^4\right) && \text{(By Eq. 1)}
\end{aligned}$$

Let's focus on the terms $\mathbf{E}[\sigma_{i,j}\,\sigma_{\ell,k}]$. We have the following cases:

Case 1: $i=\ell$ **and** $j=k$ Note that there are $\binom{m}{2}$ many of such terms in the variance bound. In this case via Equation 1, we have:

$$\mathbf{E}[\sigma_{i,j}\,\sigma_{\ell,k}]=\mathbf{E}\left[\sigma_{i,j}^2\right]=\mathbf{E}[\sigma_{i,j}]=\|p\|_2^2\ .$$

Case 2: $\{i,j,\ell,k\}$ **has three distinct elements.** First, observe that there are $6\binom{m}{3}$ many of such terms in the variance bound. Now, let's compute the expectation. Without loss of generality, assume $\ell\in\{i,j\}$, and $k\notin\{i,j\}$. Then, since X_i, X_j, and X_k are independent and identically distributed, we have:

$$\mathbf{E}[\sigma_{i,j}\,\sigma_{\ell,k}]=\mathbf{Pr}[X_i=X_j=X_k]=\sum_{r=1}^{n}p_r^3=\|p\|_3^3\ .$$

Case 3: $\{i,j,\ell,k\}$ **has four distinct elements.** There are $\binom{m}{2}\cdot\binom{m-2}{2}$ many of such terms in the variance bound. And, since all the indices are distinct, $\sigma_{i,j}$ and $\sigma_{\ell,k}$ are independent of each other. Hence, we have:

$$\mathbf{E}[\sigma_{i,j}\,\sigma_{\ell,k}]=\mathbf{E}[\sigma_{i,j}]\cdot\mathbf{E}[\sigma_{\ell,k}]=\|p\|_2^4\ ,$$

where we use Equation 1.

Exercise: Verify that $\binom{m}{2}+6\binom{m}{3}+\binom{m}{2}\binom{m-2}{2}=\binom{m}{2}^2$.

Putting these cases together, we obtain:

$$\begin{aligned}\mathbf{Var}[Y] &= \frac{1}{\binom{m}{2}^2}\left(\binom{m}{2}\|p\|_2^2 + 6\binom{m}{3}\|p\|_3^3 + \binom{m}{2}\binom{m-2}{2}\|p\|_2^4 - \binom{m}{2}^2\|p\|_2^4\right)\\ &\leq \frac{1}{\binom{m}{2}^2}\left(\binom{m}{2}\|p\|_2^2 + 6\binom{m}{3}\|p\|_3^3\right)\end{aligned}$$

□

Bibliographic Note

The content of this lecture was based on the collision-based tester that can be traced back to Goldreich and Ron's work on testing graph expansion [GR00, GR11]. Batu et al. later formalized the problem of uniformity testing for distributions [BFR+00] and provided a lower bound of $\Omega(\sqrt{n})$. Through a series of subsequent work, the sample complexity of the problem settled to be $O(\sqrt{n}/\epsilon^2)$ [Pan08, ADJ+12, VV17, CDVV14, DKN15]. Diakonikolas et al. [DGPP19] later provided a new analysis of the collision-based uniformity tester, demonstrating that it achieves the optimal sample complexity of $O(\sqrt{n}/\epsilon^2)$.

References

[ADJ+12] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Theertha Suresh. Competitive classification and closeness testing. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, volume 23 of *JMLR Proceedings*, pages 22.1–22.18. JMLR.org, 2012.

[BFR+00] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science, FOCS 2000, 12-14 November 2000, Redondo Beach, California, USA*, pages 259–269, 2000.

[CDVV14] Siu-on Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

[DGPP19] Ilias Diakonikolas, Themis Gouleakis, J. Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. *Chicago Journal of Theoretical Computer Science*, 2019(1), MAY 2019.

[DKN15] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *FOCS*, pages 1183–1202, 2015.

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electron. Colloquium Comput. Complex.*, TR00-020, 2000.

[GR11] Oded Goldreich and Dana Ron. *On testing expansion in bounded-degree graphs*, pages 68–75. Springer-Verlag, Berlin, Heidelberg, 2011.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inf. Theory*, 54(10):4750–4755, 2008.

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SICOMP*, 46(1):429–455, 2017.