CS2080: Applied Privacy for Data Science
Membership Attacks

School of Engineering & Applied Sciences
Harvard University

February 5, 2025

Null Distributions

If t is a function of the data, it has a sampling distribution. The distribution that t would obtain if the null hypothesis were true is called the *null distribution*.

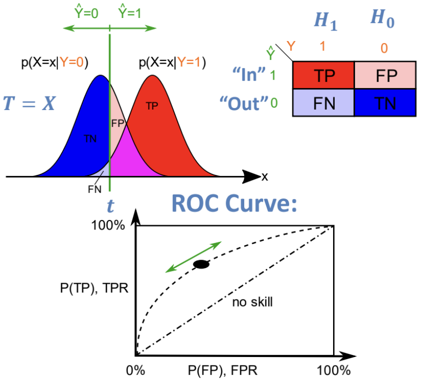
- If we use the value of t to draw an inference about the null hypothesis, we call t a **test statistic**.
- We observe t^* in some observed dataset $\mathbf{X}^*$ and reason whether it could have been a draw from the null distribution.
- If t^* is unlikely to have come from the null distribution, we **reject the null** hypothesis.
- If t^* could have been obtained from the null distribution, we **fail to reject the null**.
- Failing to reject the null, does not prove the null to be true.

How to Design MIAs

- Design Test Statistic $T = T$(everything given to attacker) that you expect to be larger under H_1 than H_0.

- Declare "In" if $T \geq t$
  "Out" otherwise
  for a threshold t
  carefully selected
  to tune FPR and TPR.

- **Q:** Why is the "Area Under the ROC Curve" (AUC) not so informative for privacy?



[By Sharpr for svg version. original work by kakau in a png - Own work, CC BY-SA 3.0]

Inferential Errors

Reasoning from known data to about an unknown hypothesis is called inference. Inferential errors are commonly labelled by type:

	H_0 True	H_1 True
Fail to Reject H_0	Correct **TNR** Specificity	Error **FNR** (Type II) Sensitivity
Reject H_0	Error **FPR** (Type I) Specificity	Correct **TPR** Sensitivity

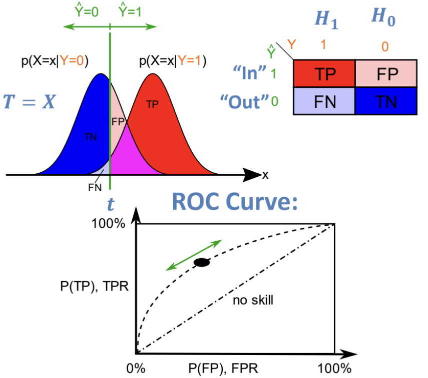
Sensitivity = TPR / (TPR + FNR)
Specificity = TNR/ (TNR + FPR)

How to Design MIAs

- Design Test Statistic $T = T(\text{everything given to attacker})$ that you expect to be larger under H_1 than H_0.

- Declare "In" if $T \geq t$
  "Out" otherwise
  for a threshold t
  carefully selected
  to tune FPR and TPR.

- **Q:** Why is the "Area Under the ROC Curve" (AUC) not so informative for privacy?



Example

$H_0 : K$-dimensional random variables $\mathbf{x}$ and $\mathbf{z}$ are both drawn from a standard Normal distribution with the same mean, $\mathcal{N}(\vec{\mu}, 1)$.

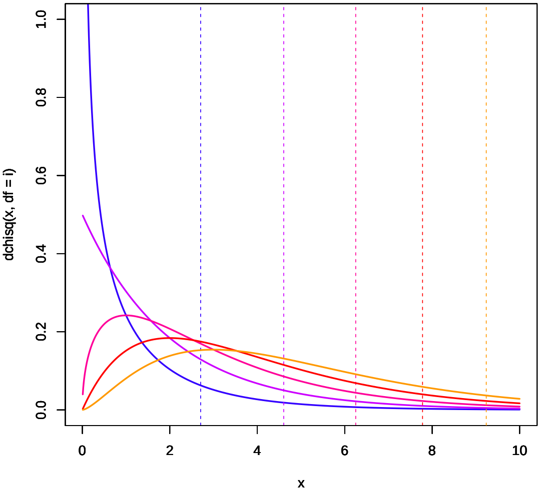
Then one test statistic is:

$$t(\mathbf{x}, \mathbf{z}) = ||\mathbf{x} - \mathbf{z}||_2 = \sqrt{\sum_{i=1}^{K} (x_i - z_i)^2}$$

Which has null distribution $\chi^2(K)$.

Example

$\chi^2(K)$ Distribution with critical values for $\delta = 0.1$, for i in 1 to 5:



Netflix Challenge (from last week)

Narayanan-Shmatikov Algorithm

1. Calculate $\mathrm{score}(aux, r')$ for each $r' \in \hat{x}$, as well as the standard deviation σ of the calculated scores.
2. Let $r_1{}'$ and $r_2{}'$ be the records with the largest and second-largest scores.
3. If $\mathrm{score}(aux, r_1{}') - \mathrm{score}(aux, r_2{}') > \phi \cdot \sigma$, output $r_1{}'$, else output $\perp$.

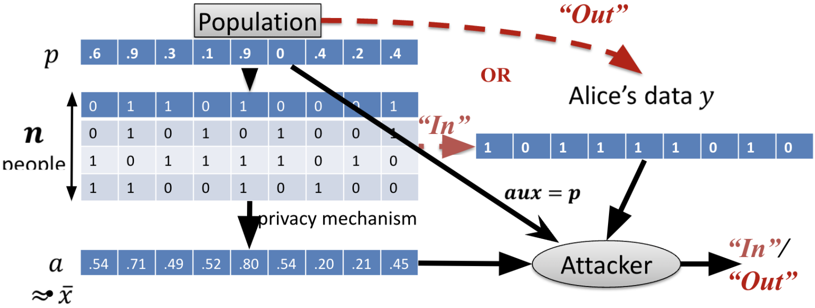
An instantiation:

$$\mathrm{score}(aux, r') = \overbrace{\sum_{a \in \mathrm{supp}(aux)}}^{\text{IMDB movies rated by user}} \overbrace{\frac{1}{\log |\{r' \in \hat{x} : a \in \mathrm{supp}(r')\}|}}^{\text{Downweight movies watched by many Netflix users}} \cdot \overbrace{\mathrm{sim}(aux_a, r_a')}^{\text{Similarity of rating \& date}}$$

eccentricity $\phi = 1.5$

Dwork, Smith, Steinke, Ullman, Vadhan (2015)

A Test Statistic for Means



$$A(y, a, p) = \begin{cases} \text{IN} & \text{if } \langle y - p, a - p \rangle \geq t \\ \text{OUT} & \text{if } \langle y - p, a - p \rangle < t \end{cases}$$

Thm [Dwork et al. `15]: under natural distributional assumptions, if mechanism outputs have error smaller than $\gamma < ½$, can achieve

- $\text{FPR} = \exp(-\Omega(d/(\gamma n)^2))$ [very small when $d \gg (\gamma n)^2$]
- $\text{TPR} = \Omega(1/(\gamma^2 n))$ [declare "In" for $k = \Omega(1/\gamma^2)$ members of dataset]