Basic Statistical Testing

Amit Sethi, EE, IITB

Learning Objectives

- Compare distributions that can explain a given sample
- Test if two samples are from two different distributions
- Test the strength of relation between two variables

The IID Assumption

- Independence of x_i and x_j for $i \neq j$.
- $p(x_i, x_j) = p(x_i)p(x_j)$

- All samples drawn from the same (identical) distributed
- $x_i, x_j \sim p_X$

Implication of the IID assumption

- Likelihood of the whole data factorizes:

- $p(X) = p(x_1, \ldots, x_N) = p_X(x_1) \times \ldots \times p_X(x_N)$
  $= \prod_{i=1}^{N} p_X(x_i)$

- $L(X) = \log p(X) = \sum_i \log p_X(x_i)$

MLE of parameterized distribution

- Between two distributions p_A is a better explanation than p_B of the entire data X if $\prod_i p_A(x_i) > \prod_i p_B(x_i)$

- By extension, if a family of distributions is parameterized by θ, then we are interested in

$$\begin{aligned} \arg\max_\theta \prod_i p_\theta(x_i) &= \arg\max_\theta \sum_i \log p_\theta(x_i) \\ &= \arg\max_\theta L_\theta(X) \end{aligned}$$

Example 1: Exponential distribution

- $p_\lambda(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0 \quad\quad\;\; , & x < 0 \end{cases}$
- Log likelihood $L_\lambda(X) = \sum_i \log p_\lambda(x_i)$
- Differntiating $\frac{\partial L_\lambda(X)}{\partial \lambda} = \sum_{x_{i \geq 0}} \left(\frac{\partial \log \lambda}{\partial \lambda} - \frac{\partial \lambda x_i}{\partial \lambda} \right) = 0$
- (assuming all samples are non-negative)
- Implies $\lambda = \frac{N}{\sum_i x_i}$ = inverse of the sample mean

Example 2: Uniform distribution

- $p_{a,b}(x) = \begin{cases} \frac{1}{b-a}, a \leq x \leq b \\ 0 \quad , \text{otherwise} \end{cases}$
- Log likelihood $L_{a,b}(X) = \sum_i \log p_{a,b}(x_i)$
- $= K \log(0) \; - (\mathrm{N} - \mathrm{K}) \log(b - a)$
- Reduce the K (terms outside $[a, b]$) to zero, and minimize $(b - a)$ by differentiating wrt a, b
- So, $a = \min x_i \, , b = \max x_i$

Example 3: Gaussian distribution

- $p_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- MLE by differentiating log likelihood wrt μ, σ
- gives $\mu = \bar{x} = \frac{\sum_i x_i}{N}$; i.e., sample mean
- and, $\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N}}$

Sufficient statistics

- Statistic is a function of the sample
$$T(X) = T(x_1, \ldots, x_N)$$
- For some distributions, computing a few statistics is sufficient for MLE estimate
- Gives complete information about the distribution
- Examples:
  - Sample mean and variance for Gaussian distribution
  - Sample mean for exponential distribution
  - Max and min for uniform distribution

Comparing two parametric distributions

- Let there be two candidate families of distributions $p_\theta(X)$ and $p_\phi(X)$ to explain the data
- Can we compare $\max L_\theta(X)$ and $\max L_\phi(X)$?
- Yes, we can, but we might overfit
- Narrow down the family of distributions based on domain knowledge (e.g. physical phenomenon)
- E.g. “Can the random variable take negative values?”

MLE vs. Bayesian estimate

- MLE finds θ that maximizes $L_\theta(X)$
- Bayesian estimate takes the expected value of θ w.r.t. $L_\theta(X)$
- Bayesian estimate: $\int \theta L_\theta(X) d\theta \;/\; \int L_\theta(X) d\theta$
- We can also incorporate a prior over θ
- $\int \theta p_\Theta(\theta) L_\theta(X) d\theta \;/\; \int p_\Theta(\theta) L_\theta(X) d\theta$

Recipe for statistical testing

1. **Explore** reasonable assumptions about the data, e.g. distribution type (including "cannot be assumed"), mean, variance, etc. and ask what do we want to verify
2. **Form null hypothesis $\boldsymbol{H_0}$** that we want to reject, e.g. "The two means are NOT different"
3. **Form alternative hypothesis** that we hope is true, e.g. "The two means are different"
4. **Decide on a significance level** $(1 - \text{confidence})$ to reject the null hypothesis BEFORE performing a test, e.g. $p < 0.05$ or $p < 0.01$
5. **Perform the test** by performing the calculations
6. **Check if the result was significant** enough to reject the null hypothesis and accept the alternative hypothesis, i.e., the alternative hypothesis was not just a chance outcome, but we are 95% or 99% confident that it is more likely than the null hypothesis

Confidence interval

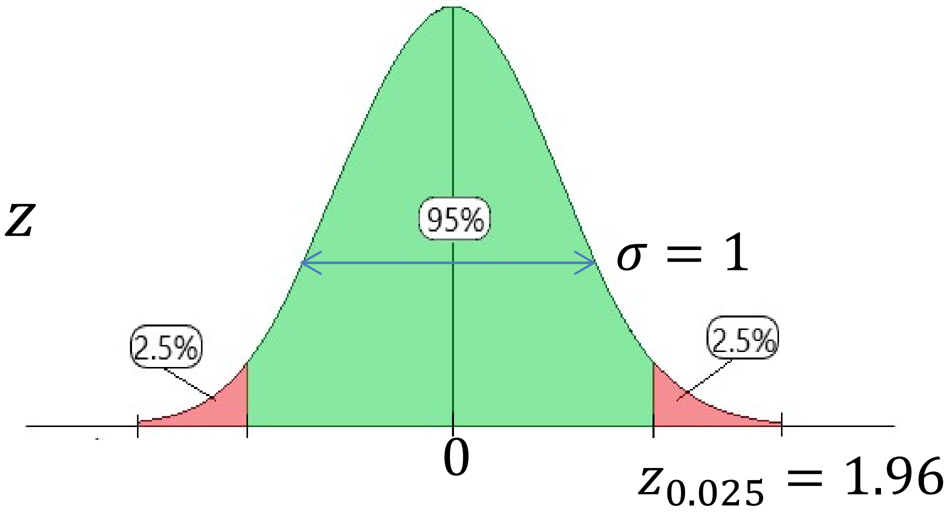
- Given sample $x_1, \ldots, x_N$ and sample mean $\bar{x}$
- What is the interval $\bar{x} \pm \varepsilon$ within which the true mean will lie with confidence $1 - \alpha$ (e.g. 95%)

$$Pr(|\bar{x} - \mu| > \varepsilon) < \alpha$$

- For Gaussian distribution

$$\varepsilon = z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$$

- Standard Gaussian is used to define z



- Replace σ by $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$

  for unknown σ

Image source: https://www.geeksforgeeks.org/confidence-interval/

Comparing means of two independent set of samples

- Given samples from two distributions, can we say with confidence that their means differ?
- $\mu_X = \frac{1}{n_X}\sum_{i=1}^{n_X} x_i, \qquad \sigma_X = \sqrt{\frac{1}{n_X}\sum_{i=1}^{n_X} (x_i - \mu_X)^2}$
- $\mu_Y = \frac{1}{n_Y}\sum_{i=1}^{n_Y} y_i, \qquad \sigma_Y = \sqrt{\frac{1}{n_Y}\sum_{i=1}^{n_Y} (y_i - \mu_Y)^2}$
- Welch's t-test: $t = \frac{\mu_X - \mu_Y}{\sqrt{\frac{{\sigma_X}^2}{n_X} + \frac{{\sigma_Y}^2}{n_Y}}}$ is matched to a table for the appropriate degrees of freedom (DoF): $\frac{\left(\frac{{\sigma_X}^2}{n_X} + \frac{{\sigma_Y}^2}{n_Y}\right)^2}{\frac{{\sigma_X}^4}{{n_X}^2(n_X-1)} + \frac{{\sigma_Y}^4}{{n_Y}^2(n_Y-1)}}$

Comparing means of paired-samples

Pre-yoga blood sugar	Post-yoga blood sugar	Δ
$\boldsymbol{x_1}$	$\boldsymbol{y_1}$	$\boldsymbol{d_1 = x_1 - y_1}$
…	…	…
$\boldsymbol{x_N}$	$\boldsymbol{y_N}$	$\boldsymbol{d_N = x_N - y_N}$

- Is there a post-event effect in a variable?
- E.g., "Does yoga lower blood sugar?"
- Mean of the difference $\overline{d} > 0$ with 95% confidence?
- Treat d as a random variable
- Is $\overline{d} - 1.8\frac{\sigma_d}{\sqrt{N}} > 0,$ where $1.8 = z_{0.05}$?

Comparing paired variables without assuming a distribution

- Let there be two paired continuous variables
- We can compare their medians, if we do not want to assume a distribution, using Wilcoxon signed rank test
- Add all the ranks of positive Δ and negative Δ separately, and pick the smaller sum or ranks as test stat w_{test}
- Test stat should be smaller than $w_{critical}$ from a table for the given *N*

Pre-yoga blood sugar	Post-yoga blood sugar	Δ	|Δ|	Rank of |Δ|
x_1	y_1	$d_1 = x_1 - y_1$	$\lvert d_1 \rvert$	r_1
…	…	…	…	…
x_N	y_N	$d_N = x_N - y_N$	$\lvert d_N \rvert$	r_N

Are two variables linearly related?

- Let there be two paired continuous variables
- Pearson's correlation coefficient
- $\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[X,Y] - \mathrm{E}[X]\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - \mathrm{E}[X]^2}\sqrt{\mathrm{E}[Y^2] - \mathrm{E}[Y]^2}}$
- For a sample $r_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$
- Ranges from −1 to +1
- Does not imply causation, nor models nonlinear relations

Are two variables monotonically related?

- Find Spearman's correlation, which is Pearson's correlation between ranks of X and Y

$$r_S = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{Cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}} = 1 - \frac{6\sum(\mathrm{R}(x_i) - \mathrm{R}(y_i))^2}{n(n^2 - 1)}$$

Some common statistical tests

Predictor	Outcome	Example	Parameteric test	Non-parameteric test
Categorical binary	Numerical unpaired	Do joggers have lower pulse rate than non-joggers	Independent t-test	Wilcoxon rank-sum test
Categorical binary	Numerical paired	Does blood sugar reduce after yoga	Paired t-test	Sign test, Wilcoxon signed-rank test
Numerical	Numerical	Are height and weight related	Pearson's correlation	Spearman's correlation
Categorical	Categorical	Does species predict color		Chi-square

How to choose a stat test

- Frame your problem
  - Predictor and outcome variable types
  - Decision to be expected
- Check if a widely accepted test is already available
- Check if the assumptions behind the test are applicable to your scenario
- Else, make your own test by using an existing test as a base for approach