A. Fourier Theory

Fourier Series - harmonic analysis (more general)

$f(x) = \frac{1}{2}a_0 + \sum a_n \cos(nx) + \sum b_n \sin(nx)$

$f(x) \longrightarrow [a_0, a_1, b_1, a_2, b_2, \ldots]$ feature vector

Gibbs phenomenon - boosting / noise in edge of signal

Fourier Transform / Laplace Transform | Z Transform - go from time domain to frequency domain

Laplace Transform is generalized version of FT for continuous signal, ZT is for discrete signal

B. Cross-validation

Verify the results on an independent dataset, ideally

Method 1: holdout method - simplest cross-validation

divide data into < trainval / testing

Advantage: easy

Disadvantage: evaluation has a high variance heavily depends on how data is divided

Method 2: K-fold cross validation - improved version

data is divided into K subsets, then repeat holdout method K times. Each time, one of the K subsets is used for testing (avg. over K trials)

Adv: variance of result is reduced as K ↑, can independently choose how large each test set is

Method 3: Leave-one-out cross validation

One sample is used for testing, repeat n times (where n is number of samples)

M2 vs M3 depends on computing time / sample size

C. Bootstrapping

Statistical test relying on random sampling with replacement

Coin flip experiment

$x = x_1, x_2, \ldots, x_{10}$ be 10 observations from exp.

$x_i = \begin{cases} 1 & \text{heads} \\ 0 & \text{otherwise} \end{cases}$

Use t-statistics, the dist. of sample mean

$\bar{x} = \frac{1}{10}(x_1 + x_2 + \ldots + x_{10})$

Use bootstrapping to derive $\bar{x}$'s dist.

First resample: $x_1^* = x_3, x_8, x_1, x_4, x_4, x_8, x_1, x_{10}, x_2, x_4$ (with replacement)

The number of data points in a bootstrap resample is equal to the number of data points in original obs.

Then, compute mean of x_1^* to get first bootstrap mean μ_1^*

Repeat process N times (N is large) $[\mu_1^*, \mu_2^*, \ldots, \mu_N^*]$ represents empirical dist. of sample mean; from this distribution, we can derive a bootstrap CI for hypothesis testing

D. Student's T-Test

Statistical method of testing hypotheses about mean of a sample drawn from a normally distributed pop. when pop's SD is unknown

First, formulate a null hypothesis

e.g. there is no significant difference b/t μ_1 and μ_2

then conduct t-test (one-sided or two-sided)

Two-sided: are they equal?

One-sided: is one larger?

If the observed t-statistic is more extreme than the critical value determined by appropriate reference dist., H_0 is rejected

Critical value depends on significance level (α)

α is probability of erroneously rejecting the null hypothesis

Ex: sample of size $n=25$ with $\bar{x}=79$ and $s=10$ → pop. mean $=75$

mean of sample $= 75$?

Compute $t = \left(\frac{\bar{x}-\mu}{s/\sqrt{n}}\right) = \left(\frac{79-75}{10/5}\right) = 2$ Check t-table (two-sided at $\alpha=0.05$) crit. value is 2.064

So we can not reject the null hypothesis ($2 < 2.064$)

Multiple comparison correction:

Bonferroni correction

FDR

E. Non-parametric method – permutation test

Ex: group $A = \{A_1, A_2, \dots, A_n\}$ test whether

group $B = \{B_1, B_2, \dots, B_m\}$ $\bar{A} > \bar{B}$

1. Calculate original difference

$\frac{1}{N}\sum_{i}^{N} A_i - \frac{1}{M}\sum_{i}^{M} B_i$

2. Concatenate A and B

$[A_1, A_2, \dots, A_N, B_1, \dots, B_M]$

then randomize the order, then select the first N numbers as new A^* and last M numbers as new B^*

Get perm1 $= \frac{1}{N}\sum_{i=1}^{n} A_i^* - \frac{1}{M}\sum_{i=1}^{M} B_i^*$

3. Repeat step 2 for 1000, 10000, etc. times and get [perm1, perm2, ..., perm10k]
4. Rank result - check whether original difference is within top 5% of

If yes, we can conclude that $\bar{A} > \bar{B}$ is significant.

F. Generalized Linear Regression

Sometimes we cannot compare results directly

e.g.						
age	86	78	83	68	63	75
brain volume	260	320	300	430	500	370

of course younger group has more brain volume than older group

$Y = \beta_0 + \beta_1 X_1$

$$\begin{bmatrix} 260 \\ 320 \\ 300 \\ 430 \\ 500 \\ 370 \end{bmatrix} = \beta_0 + \beta_1 \begin{bmatrix} 86 \\ 78 \\ 83 \\ 68 \\ 63 \\ 75 \end{bmatrix}$$

residual = brain region volume without age effect → might find that there is not a difference when doing a t-test now

GLM: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_N X_N + e$ The model is linear in the parameters

Model selection: which x to include

Model fit: R^2 - residual analysis

G. AIC / BIC

Akaike Information Criteria

Bayesian Information Criteria

Model with smallest AIC/BIC is best model

Both penalize models for complexity but impose different pens.

$AIC = 2k - 2\log(\ell)$ 　　 k: # parameters

$BIC = k\log(n) - 2\log(\ell)$ 　　 n: # subjects

ℓ: model likelihood

H. Cohen's d Effect size

> 0.3 means effective

measure the distance between means of groups

$d = \frac{|\bar{x}_1 - \bar{x}_2|}{S}$ 　 if $S_1 = S_2$, $S_1 = S_2 = S$

if $S_1 \neq S_2$, $S = \frac{S_1 + S_2}{2}$

$d: 0 \sim 0.2 \rightarrow$ small effect

$d: 0.2 \sim 0.5 \rightarrow$ medium effect

$d: 0.5 \rightarrow$ large effect

Sample size is number of technical replicates

$$\text{Effective Sample size} = \frac{\text{\# of samples}}{1 + (\text{\# samples} - 1) \cdot \text{correlation}}$$