COMP 680
Statistics for Computing and Data Science
Week 7: Hypothesis Testing II

Su Chen, Assistant Teaching Professor,
Rice D2K Lab

Outline

Z Test

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \sigma^2)$, σ^2 **known**, and test on $H_0 : \mu = \mu_0$

Z Test

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \sigma^2)$, σ^2 **known**, and test on $H_0 : \mu = \mu_0$
- Do Rice students have higher IQ than general population?
  - IQ in general population $\sim \mathrm{N}(\mu = 100, \sigma^2 = 15^2)$
  - test a sample of n Rice students

Z Test

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \sigma^2)$, σ^2 **known**, and test on $H_0 : \mu = \mu_0$
- Do Rice students have higher IQ than general population?
  - IQ in general population $\sim \mathrm{N}(\mu = 100, \sigma^2 = 15^2)$
  - test a sample of n Rice students
- test statistic under the null follows standard normal distribution

$$z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1)$$

Z Test

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \sigma^2)$, σ^2 **known**, and test on $H_0 : \mu = \mu_0$
- Do Rice students have higher IQ than general population?
  - IQ in general population $\sim \mathrm{N}(\mu = 100, \sigma^2 = 15^2)$
  - test a sample of n Rice students
- test statistic under the null follows standard normal distribution

$$z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1)$$

  - p-value $= \mathbb{P}(Z \geq z) = 1 - \Phi(z)$ one-sided, why?
  - reject if $p < 5\%$, equivalent to reject when $z > Z_{0.95}$, why?

T Test

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \sigma^2)$, σ^2 **unknown**, and test on $H_0 : \mu = \mu_0$

T Test

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \sigma^2)$, σ^2 **unknown**, and test on $H_0 : \mu = \mu_0$
- Do Rice students have an average IQ score of 100?
  - H_1 is two-sided this time
  - n Rice students with sample mean $\bar{X}_n$ and sample variance s^2

T Test

- $X_1, X_2, \cdots X_n \stackrel{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \sigma^2)$, σ^2 **unknown**, and test on $H_0 : \mu = \mu_0$
- Do Rice students have an average IQ score of 100?
    - H_1 is two-sided this time
    - n Rice students with sample mean $\bar{X}_n$ and sample variance s^2
- test statistic under the null follows t distribution

$$t = \frac{\bar{X}_n - \mu_0}{s/\sqrt{n}} \sim \mathrm{t}_{df=n-1}$$

T Test

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \sigma^2)$, σ^2 **unknown**, and test on $H_0 : \mu = \mu_0$
- Do Rice students have an average IQ score of 100?
  - H_1 is two-sided this time
  - n Rice students with sample mean $\bar{X}_n$ and sample variance s^2
- test statistic under the null follows t distribution

$$t = \frac{\bar{X}_n - \mu_0}{s/\sqrt{n}} \sim \mathrm{t}_{df=n-1}$$

  - p-value $= \mathbb{P}(T \leq -|t| \text{or} \, T \geq |t|)$ two-sided, why?

The Wald Test

- Both z-test and t-test require population distribution to be normal.

The Wald Test

- Both z-test and t-test require population distribution to be normal.
- You flip a coin 100 times and get 65 heads, is the coin fair?

The Wald Test

- Both z-test and t-test require population distribution to be normal.
- You flip a coin 100 times and get 65 heads, is the coin fair?

The Wald Test

- Both z-test and t-test require population distribution to be normal.
- You flip a coin 100 times and get 65 heads, is the coin fair?
- Wald test applies to any test statistic that is asymptotically normal:
  - $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(x|\theta)$
  - $H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$
  - let $\hat{\theta}$ be a statistic to estimate θ, and under H_0:

$$\frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})} \to \text{N}(0, 1)$$

χ^2 Test of Goodness of Fit

- Data from multinomial distribution
- Example: jury selection Harris county demographics: 70% white, 20% black, 7% asian, 3% others.

χ^2 Test of Goodness of Fit

- Data from multinomial distribution
- Example: jury selection Harris county demographics: 70% white, 20% black, 7% asian, 3% others.
  - χ^2 test statistic:
  $$T = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}$$
  - X_j: observed count in each category
  - E_j: expected count in each category under H_0

Outline

Two Sample T Test - Paired

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_1, \sigma^2)$, and $Y_1, Y_2, \cdots Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_2, \sigma^2)$, with σ^2 **unknown**, and test on $H_0 : \mu_1 - \mu_2 = 0$

Two Sample T Test - Paired

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_1, \sigma^2)$, and $Y_1, Y_2, \cdots Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_2, \sigma^2)$, with σ^2 **unknown**, and test on $H_0 : \mu_1 - \mu_2 = 0$
- Can meditation change your Serotonin level?
  - H_1 is two-sided this time
  - n pair of measurements: often repeated measure
  - Serontonin level before and after meditation for the same individual

Two Sample T Test - Paired

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_1, \sigma^2)$, and $Y_1, Y_2, \cdots Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_2, \sigma^2)$, with σ^2 **unknown**, and test on $H_0 : \mu_1 - \mu_2 = 0$
- Can meditation change your Serotonin level?
  - H_1 is two-sided this time
  - n pair of measurements: often repeated measure
  - Serontonin level before and after meditation for the same individual
- take difference of measurement $d_i = X_i - Y_i$,
- treat d_i as the new data and apply one-sample t-test
- test statistic under the null follows t distribution

$$t = \frac{\bar{d}_n}{s_d/\sqrt{n}} \sim \mathrm{t}_{df=n-1}$$

Two Sample T Test - Independent and Equal Variance

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_1, \sigma^2)$, and $Y_1, Y_2, \cdots Y_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_2, \sigma^2)$ with σ^2 **unknown**, and test on $H_0 : \mu_1 = \mu_2$

Two Sample T Test - Independent and Equal Variance

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_1, \sigma^2)$, and $Y_1, Y_2, \cdots Y_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_2, \sigma^2)$ with σ^2 **unknown**, and test on $H_0 : \mu_1 = \mu_2$
- Do men and women have the same Serotonin level?
    - H_1 is two-sided this time
    - two groups of independent measurements

Two Sample T Test - Independent and Equal Variance

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_1, \sigma^2)$, and $Y_1, Y_2, \cdots Y_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_2, \sigma^2)$ with σ^2 **unknown**, and test on $H_0 : \mu_1 = \mu_2$
- Do men and women have the same Serotonin level?
  - H_1 is two-sided this time
  - two groups of independent measurements
- test statistic under the null follows t distribution

$$t = \frac{\bar{X}_n - \bar{Y}_m}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathrm{t}_{df=n+m-2}$$

Two Sample T Test - Independent and Equal Variance

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_1, \sigma^2)$, and $Y_1, Y_2, \cdots Y_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_2, \sigma^2)$ with σ^2 **unknown**, and test on $H_0 : \mu_1 = \mu_2$
- Do men and women have the same Serotonin level?
  - H_1 is two-sided this time
  - two groups of independent measurements
- test statistic under the null follows t distribution

$$t = \frac{\bar{X}_n - \bar{Y}_m}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathrm{t}_{df=n+m-2}$$

- pooled variance formula

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

Two Sample T Test - Independent

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \text{N}(\mu_1, \sigma_1^2)$, and $Y_1, Y_2, \cdots Y_m \overset{\text{i.i.d.}}{\sim} \text{N}(\mu_2, \sigma_2^2)$ with σ_1^2 and σ_2^2 **unknown**, and test on $H_0 : \mu_1 = \mu_2$

Two Sample T Test - Independent

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_1, \sigma_1^2)$, and $Y_1, Y_2, \cdots Y_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_2, \sigma_2^2)$ with σ_1^2 and σ_2^2 **unknown**, and test on $H_0 : \mu_1 = \mu_2$
- **equal variances not assumed**

Two Sample T Test - Independent

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_1, \sigma_1^2)$, and $Y_1, Y_2, \cdots Y_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_2, \sigma_2^2)$ with σ_1^2 and σ_2^2 **unknown**, and test on $H_0 : \mu_1 = \mu_2$
- **equal variances not assumed**
- test statistic under the null follows t distribution

$$t = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \sim \mathrm{t}_{df}$$

Two Sample T Test - Independent

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_1, \sigma_1^2)$, and $Y_1, Y_2, \cdots Y_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_2, \sigma_2^2)$ with σ_1^2 and σ_2^2 **unknown**, and test on $H_0 : \mu_1 = \mu_2$
- **equal variances not assumed**
- test statistic under the null follows t distribution

$$t = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \sim \mathrm{t}_{df}$$

- degree of freedom formula

$$df = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_Y^2}{m}\right)^2}$$

χ^2 Test of Independence

- Data from contingency tables of two categorical variables
- Example:

Type of Movie	Snacks	No Snacks
Action	50	75
Comedy	125	175
Family	90	30
Horror	45	10

χ^2 Test of Independence

- Data from contingency tables of two categorical variables
- Example:

Type of Movie	Snacks	No Snacks
Action	50	75
Comedy	125	175
Family	90	30
Horror	45	10

  - χ^2 test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{i,j} - E_{i,j})^2}{E_{i,j}}$$

  - $X_{i,j}$: observed count in each category
  - $E_{i,j}$: expected count in each category under H_0

χ^2 Test of Independence

- Data from contingency tables of two categorical variables
- Example:

Type of Movie	Snacks	No Snacks
Action	50	75
Comedy	125	175
Family	90	30
Horror	45	10

  - χ^2 test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{i,j} - E_{i,j})^2}{E_{i,j}}$$

  - $X_{i,j}$: observed count in each category
  - $E_{i,j}$: expected count in each category under H_0

More than Two Samples - ANOVA

- Data from one numerical measure and one categorical variable with more than two categories.

More than Two Samples - ANOVA

- Data from one numerical measure and one categorical variable with more than two categories.
- Do Rice students from 11 residential colleges have the same average GPA?
  - numerical variable: GPA
  - categorical variable: residential college

More than Two Samples - ANOVA

- Data from one numerical measure and one categorical variable with more than two categories.
- Do Rice students from 11 residential colleges have the same average GPA?
  - numerical variable: GPA
  - categorical variable: residential college
- $X_{k1}, X_{k2}, \cdots X_{n_k} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_k, \sigma^2)$, where $k = 1, 2, \cdots K$ of total K groups, with σ^2 **unknown but assumed equal**

More than Two Samples - ANOVA

- Data from one numerical measure and one categorical variable with more than two categories.
- Do Rice students from 11 residential colleges have the same average GPA?
  - numerical variable: GPA
  - categorical variable: residential college
- $X_{k1}, X_{k2}, \cdots X_{n_k} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_k, \sigma^2)$, where $k = 1, 2, \cdots K$ of total K groups, with σ^2 **unknown but assumed equal**
  - H_0: all the μ_k are the same
  - H_1: at least one group mean is different
  - you do not know which group is different even if H_0 is rejected

ANOVA Table

Source of Variation	SS	df	MS = SS / df
Between Groups	SSB	$K - 1$	MSB = SSB / $(K - 1)$
Within Groups	SSE	$N - K$	MSE = SSE / $(N - K)$
Total	SST	$N - 1$	

ANOVA Table

Source of Variation	SS	df	MS = SS / df
Between Groups	SSB	$K-1$	MSB = SSB / $(K-1)$
Within Groups	SSE	$N-K$	MSE = SSE / $(N-K)$
Total	SST	$N-1$	

- $\text{SSB} = \sum_{k=1}^{K} n_k(\bar{X}_k - \bar{X})^2$
- $\text{SSE} = \sum_{k=1}^{K} \sum_{i=1}^{n_k}(X_{ki} - \bar{X}_k)^2$
- $\text{SST} = \sum_{k=1}^{K} \sum_{i=1}^{n_k}(X_{ki} - \bar{X})^2 = \text{SSB} + \text{SSE}$

ANOVA Table

Source of Variation	SS	df	MS = SS / df
Between Groups	SSB	$K - 1$	MSB = SSB / $(K - 1)$
Within Groups	SSE	$N - K$	MSE = SSE / $(N - K)$
Total	SST	$N - 1$	

- $\text{SSB} = \sum_{k=1}^{K} n_k(\bar{X}_k - \bar{X})^2$
- $\text{SSE} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_{ki} - \bar{X}_k)^2$
- $\text{SST} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_{ki} - \bar{X})^2 = \text{SSB} + \text{SSE}$

$$F = \frac{\text{MSB}}{\text{MSE}} \sim F_{df_1 = K-1, df_2 = N-K}$$

Outline

Parametric v.s Nonparametric Tests

Parametric Test	Nonparametric Counterpart
1-sample t-test	Wilcoxon signed-rank test
2-sample t-test	Wilcoxon 2-sample rank-sum test
k-sample ANOVA	Kruskal-Wallis test

When to use nonparametric tests

- With correct assumptions (e.g., normal distribution), parametric methods will be more efficient than nonparametric ones but not so much more.

When to use nonparametric tests

- With correct assumptions (e.g., normal distribution), parametric methods will be more efficient than nonparametric ones but not so much more.
  - large-sample efficiency of Wilcoxon test compared to t-test ≈ 0.95

When to use nonparametric tests

- With correct assumptions (e.g., normal distribution), parametric methods will be more efficient than nonparametric ones but not so much more.
  - large-sample efficiency of Wilcoxon test compared to t-test ≈ 0.95
- If the normality assumption grossly violated, nonparametric tests can be much more efficient and powerful.

When to use nonparametric tests

- With correct assumptions (e.g., normal distribution), parametric methods will be more efficient than nonparametric ones but not so much more.
  - large-sample efficiency of Wilcoxon test compared to t-test ≈ 0.95
- If the normality assumption grossly violated, nonparametric tests can be much more efficient and powerful.
- Circumstances in which parametric methods perform poorly.
  - extreme outliers

Wilcoxon signed-rank test

- Nonparametric analogue to the 1-sample t-test
- Almost always used on paired data to test for the median difference being 0 or not

Wilcoxon signed-rank test

- Nonparametric analogue to the 1-sample t-test
- Almost always used on paired data to test for the median difference being 0 or not
  - $D = Y_{post} - Y_{pre}$
  - $H_0 : P(D > 0) = \frac{1}{2}$ v.s $H_a : P(D > 0) \neq \frac{1}{2}$
  - discard all $D = 0$, work with signed-rank SR
  - $SR =$ Sign of D $\times$ Rank of $|D|$

Wilcoxon signed-rank test

- Nonparametric analogue to the 1-sample t-test
- Almost always used on paired data to test for the median difference being 0 or not
  - $D = Y_{post} - Y_{pre}$
  - $H_0 : P(D > 0) = \frac{1}{2}$ v.s $H_a : P(D > 0) \neq \frac{1}{2}$
  - discard all $D = 0$, work with signed-rank SR
  - $SR =$ Sign of D $\times$ Rank of $|D|$
- Approx z-score

$$z = \frac{\sum SR_i}{\sqrt{\sum SR_i^2}}$$

Wilcoxon signed-rank test: Example

Subject	Drug 1	Drug 2	Diff (2-1)	Sign	Rank
1	1.9	0.7	−1.2	-	3
2	−1.6	0.8	2.4	+	8
3	−0.2	1.1	1.3	+	4.5
4	−1.2	0.1	1.3	+	4.5
5	−0.1	−0.1	0.0	NA	NA
6	3.4	4.4	1.0	+	2
7	3.7	5.5	1.8	+	7
8	0.8	1.6	0.8	+	1
9	0.0	4.6	4.6	+	9
10	2.0	3.4	1.4	+	6

Table: Hours of extra sleep on drugs 1 and 2, differences, signs and ranks of sleep study data

Wilcoxon signed-rank test: Example

Subject	Drug 1	Drug 2	Diff (2-1)	Sign	Rank
1	1.9	0.7	−1.2	-	3
2	−1.6	0.8	2.4	+	8
3	−0.2	1.1	1.3	+	4.5
4	−1.2	0.1	1.3	+	4.5
5	−0.1	−0.1	0.0	NA	NA
6	3.4	4.4	1.0	+	2
7	3.7	5.5	1.8	+	7
8	0.8	1.6	0.8	+	1
9	0.0	4.6	4.6	+	9
10	2.0	3.4	1.4	+	6

Table: Hours of extra sleep on drugs 1 and 2, differences, signs and ranks of sleep study data

$$z = \frac{\sum SR_i}{\sqrt{\sum SR_i^2}} = 2.31, \text{two-tailed p-value} = 0.021$$

Wilcoxon (WMW) 2-sample rank-sum test

- Testing for equality of central tendency of two distributions with unpaired data
- Ranking is done by combining two samples and ignoring group labels

Wilcoxon (WMW) 2-sample rank-sum test

- Testing for equality of central tendency of two distributions with unpaired data
- Ranking is done by combining two samples and ignoring group labels
  - Wilcoxon rank sum test statistic

$$W = \sum_{i \in n_1} R_i - \frac{n_1(n_1 + 1)}{2}$$

  - where R_i is sum of ranks in Group 1 with sample size n_1
  - Under H_0, $\mu_W = \frac{n_1 n_2}{2}$ and $\sigma_W = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$

$$z = \frac{W - \mu_W}{\sigma_W}$$

WMW test: Example

Female	120	118	121	119
Male	124	120	133	
Ranks for Female	3.5	1	5	2
Ranks for Male	6	3.5	7	

WMW test: Example

Female	120	118	121	119
Male	124	120	133	
Ranks for Female	3.5	1	5	2
Ranks for Male	6	3.5	7	

- $W = 1.5, z = -1.59$, p-value $= 0.056$
- The concordance probability (C index) $C = \frac{\bar{R} - \frac{n_1+1}{2}}{n_2} = 0.125$

WMW test: Example

Female	120	118	121	119
Male	124	120	133	
Ranks for Female	3.5	1	5	2
Ranks for Male	6	3.5	7	

- $W = 1.5, z = -1.59$, p-value $= 0.056$
- The concordance probability (C index) $C = \frac{\bar{R} - \frac{n_1+1}{2}}{n_2} = 0.125$
- Interpretation of C index:
  - probability that a randomly chosen female has a value greater than a randomly chosen male is 0.125

Kruskal-Wallies test

- Compare medians among k groups ($k > 2$) (like ANOVA with data replaced by their ranks)

Kruskal-Wallies test

- Compare medians among k groups (k $>$ 2) (like ANOVA with data replaced by their ranks)
- Combine $\sum_{i=1}^{G} n_i = N$ samples from $i = 1, \cdots G$ groups and rank them.

Kruskal-Wallies test

- Compare medians among k groups (k $>$ 2) (like ANOVA with data replaced by their ranks)
- Combine $\sum_{i=1}^{G} n_i = N$ samples from $i = 1, \cdots G$ groups and rank them.
- Test statistic

$$H = (N-1)\frac{\sum_{i=1}^{G} n_i(\bar{R}_i - \bar{R})^2}{\sum_{i=1}^{G}\sum_{j=1}^{n_i}(R_{ij} - \bar{R})^2}$$

Kruskal-Wallies test

- Compare medians among k groups (k $>$ 2) (like ANOVA with data replaced by their ranks)
- Combine $\sum_{i=1}^{G} n_i = N$ samples from $i = 1, \cdots G$ groups and rank them.
- Test statistic

$$H = (N-1)\frac{\sum_{i=1}^{G} n_i(\bar{R}_i - \bar{R})^2}{\sum_{i=1}^{G}\sum_{j=1}^{n_i}(R_{ij} - \bar{R})^2}$$

- Look up critical value of H and p-value approx by χ^2 with d.f.$=G-1$

Permutation test

- Compare 2-samples with simulation and re-sampling technique
- Also known as A/B testing

Permutation test

- Compare 2-samples with simulation and re-sampling technique
- Also known as A/B testing
  - a measure collected for Group A and B
  - test whether this measure is different for the two Groups
- Rational: under the null the two groups are the same, therefore the group labels should not matter
- Implement: repeatedly permute the group labels (or the measures), calculate difference between two "group" means

Kolmogorov - Smirnov test

- A nonparametric test of the equality one-dimensional probability distributions.
- Testing the entire sample, not just mean or median!
- Compare a sample with a reference probability distribution: one-sample KS test.
- Compare two samples: two-sample KS test.

One-sample K-S test

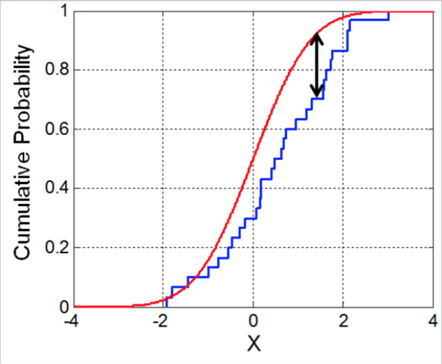

Illustration of the Kolmogorov–Smirnov statistic. Red line is CDF, blue line is an ECDF, and the black arrow is the K–S statistic.

Two-sample K-S test

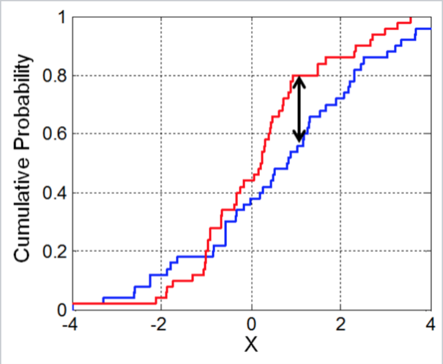

Illustration of the two-sample Kolmogorov–Smirnov statistic. Red and blue lines each correspond to an empirical distribution function, and the black arrow is the two-sample KS statistic.

Outline

Motivating Example

- "I tested how quickly different monkeys can complete a certain task. I have trial scores for 267 species, and I want to see if there is a statistically significant species effect on average trial scores."
- What test do you run here?

Motivating Example

- "I tested how quickly different monkeys can complete a certain task. I have trial scores for 267 species, and I want to see if there is a statistically significant species effect on average trial scores."
- What test do you run here?
- ANOVA.
- Assume ANOVA test determines significant.

Motivating Example

- "I tested how quickly different monkeys can complete a certain task. I have trial scores for 267 species, and I want to see if there is a statistically significant species effect on average trial scores."
- What test do you run here?
- ANOVA.
- Assume ANOVA test determines significant.
- Post-hoc test: which pairwise differences are statistically significant?

Hypothetical

- We want run a bunch of hypothesis tests **on the same data set**.

Hypothetical

- We want run a bunch of hypothesis tests **on the same data set**.
  - Significance of individual features.
  - Pairwise testing between categories.
  - Testing multiple different research questions.

Hypothetical

- We want run a bunch of hypothesis tests **on the same data set**.
  - Significance of individual features.
  - Pairwise testing between categories.
  - Testing multiple different research questions.
- What if we use the typical procedure with $\alpha = 0.05$?

Inflated Type I Error

- Each individual hypothesis test has a Type I error rate of 0.05.
- On average, expect to make a Type I for every 20 null hypothesis rejections.

Inflated Type I Error

- Each individual hypothesis test has a Type I error rate of 0.05.
- On average, expect to make a Type I for every 20 null hypothesis rejections.
- $1 - 0.95^n$ chance to make a Type I error in n null hypothesis rejections.
  - when $n = 10$, about 40% "Type I" error rate

Inflated Type I Error

- Each individual hypothesis test has a Type I error rate of 0.05.
- On average, expect to make a Type I for every 20 null hypothesis rejections.
- $1 - 0.95^n$ chance to make a Type I error in n null hypothesis rejections.
  - when $n = 10$, about 40% "Type I" error rate
- We might want to adjust our procedure due to running multiple tests.

Other Error Rate

- **Familywise Error Rate (FWER)**: probability of making at least one type I error out of all of our hypothesis tests, i.e.

$$FWER = P(\# \text{ of Type I errors} > 0).$$

Other Error Rate

- **Familywise Error Rate (FWER)**: probability of making at least one type I error out of all of our hypothesis tests, i.e.

$$FWER = P(\# \text{ of Type I errors} > 0).$$

- **False Discovery Rate (FDR)**: experted proportion of false positives out of all tests that are declared significant, i.e.

$$FDR = E\left[\frac{\#(H_0 \text{ rejected } \cap H_0 \text{ is true})}{\#(H_0 \text{ rejected })}\right].$$

Bonferroni Correction

- n tests, FWER of $\alpha \rightarrow$ critical value of α/n for all individual tests.
- Simplest and most widely-known correction.

Bonferroni Correction

- n tests, FWER of $\alpha \rightarrow$ critical value of α/n for all individual tests.
- Simplest and most widely-known correction.
- Mathematically guaranteed to work for any set of valid hypothesis tests.
    - why?

Bonferroni Correction

- n tests, FWER of $\alpha \rightarrow$ critical value of α/n for all individual tests.
- Simplest and most widely-known correction.
- Mathematically guaranteed to work for any set of valid hypothesis tests.
  - why?
- Downsides?

Benjamini-Hochberg Procedure

- Idea: order p-values and compare to different threshold
- Reject all smaller p-values once one falls below its specified threshold

Benjamini-Hochberg Procedure

- Idea: order p-values and compare to different threshold
- Reject all smaller p-values once one falls below its specified threshold
- **Controls FDR**.
- Assumes independent test statistics.

Benjamini-Hochberg Procedure

For a **desired FDR level** α:

1. Order p-values of all tests from smallest to largest (i.e., $p_{(1)}, p_{(2)} \ldots, p_{(n)}$).
2. Calculate $\alpha_k^* = \frac{\alpha k}{n}$ for $k \in 1, \ldots, n$.
3. Find the largest k such that $p_{(k)} \leq \alpha_k^*$.
4. Reject all the null hypotheses corresponding to $p_{(1)}, p_{(2)} \ldots, p_{(k)}$.

Benjamini-Hochberg Procedure

Adjusted critical values for $\alpha = 0.05$:

Test	P-Value	α_k^*
Test 1	0.0028857	0.01
Test 2	0.0096879	0.02
Test 3	0.0233847	0.03
Test 4	0.0241055	0.04
Test 5	0.0609072	0.05

Benjamini-Hochberg Procedure

Adjusted critical values for $\alpha = 0.05$:

Test	P-Value	α_k^*
Test 1	0.0028857	0.01
Test 2	0.0096879	0.02
Test 3	0.0233847	0.03
Test 4	0.0241055	0.04
Test 5	0.0609072	0.05

Test 4 is the first from the bottom such that $p_{(k)} < \alpha_k^*$. Thus, reject H_0 for Tests 1, 2, 3, and 4.

Note

- Understand the framework of Hypothesis Testing is key!

Note

- Understand the framework of Hypothesis Testing is key!
- Which test to use?
  - example of some general guideline and implementation in Python
  - always check assumptions

Note

- Understand the framework of Hypothesis Testing is key!
- Which test to use?
  - example of some general guideline and implementation in Python
  - always check assumptions
- 0.05 is not a magic number....
  - everything is significant with infinite many data
  - effect SIZE matters

Note

- Understand the framework of Hypothesis Testing is key!
- Which test to use?
  - example of some general guideline and implementation in Python
  - always check assumptions
- 0.05 is not a magic number....
  - everything is significant with infinite many data
  - effect SIZE matters
- No p-hacking! Peeking is cheating!!
  - provide evidence and leave decision to domain experts

Outline

1. Common Parametric Tests - One Sample
2. Common Parametric Tests - Two Sample
3. Common Non-Parametric Tests
4. Multiple Testing
5. Code Demo