Parametric Tests

Isaac Griffith

CS 6620
Department of Informatics and Computer Science
Idaho State University

Hypothesis Testing

- **Objective**: Determine if we can reject a null hypothesis, H_0, based on a sample
- The null hypothesis should be formulated negatively
  - If not rejected, nothing can be said about the outcome
  - If rejected, H_0 is said to be false with a significance of α
- When a test is carried out
  - we calculate the lowest possible significance value (p-value) with which we can reject H_0

Hypothesis Testing

- To test H_0
  - A test, t, is defined
  - A critical area, C, is given over which, partly, t varies
- Significance testing is then formulated as:
  - If $t \in C$, reject H_0
  - If $t \notin C$, fail to reject H_0
- **Example**:
  - H_0: the observed vehicle is a car
  - t is the number of wheels
  - $C = 1, 2, 3, 4, \ldots$
  - Test:
    - If $t \leq 3$ and $t \geq 5$, reject H_0
    - If $t = 4$, fail to reject H_0

Important Probabilities

- $\alpha = P(\text{type-I-error}) = P(\text{reject } H_0 | H_0 \text{ is true})$

- $\beta = P(\text{type-II-error}) = P(\text{not reject } H_0 | H_0 \text{ is false})$

- $Power = 1 - \beta = P(\text{reject } H_0 | H_0 \text{ is false})$

- Power is affected by:
  - Test efficacy
  - Sample size (larger sample = higher power)
  - Choice of one- or two-sided H_A (one-sided = higher power)

Parametric and Non-parametric Tests

- **Parametric Tests**: tests based on a model (set of parameters) involving a specific distribution.
  - Typically assumes that some of the parameters are normally distributed
  - Requires parameters be at least interval scale

- **Non-Parametric Tests**: Do not make the same assumptions, rather only very general assumptions

Selecting Tests

- Two factors to be considered when selecting between non-parametric and parametric tests:
  - **Applicability**: What are the assumptions to be made by the tests?
  - **Power**: Parametric tests tend to have higher power than non-parametric
    - Thus, require fewer data points, if the assumptions are true.
- It should be noted that several parametric tests are fairly robust to violations of their assumptions
  - Thus, they can be used as long as the deviations are not too large

Parametric Tests

Parametric Tests: Overview

- **t-Test**: Used to compare two sample means (medians)
- **Paired t-Test**: t-test for paired comparison designs
- **F-Test**: Used to compare two sample distributions
- **ANOVA**: Family of tests used for designs with more than two levels of a factor

t-Test Overview

- Compare to **independent** samples (one factor with two levels).
- **Input**: samples $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$
- **Hypotheses**:
  - H_0: $\mu_x = \mu_y$
  - Two-Sided H_A: $\mu_x \neq \mu_y$
  - One-Sided H_A: $\mu_x > \mu_y$
- **Calculations**:
  - $t_0 = \frac{\bar{x} - \bar{y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{n}}}$
  - $S_p = \sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}$
  - S_x^2 and S_y^2 are sample variances
- **Criterion**:
  - Degrees of freedom: $df = n + m - 2$
  - Two-Sided: reject H_0 if $|t_0| > t_{\alpha/2,df}$
  - One-Sided: reject H_0 if $t_0 > t_{\alpha,df}$

t-Test Example

Defect density in different programs have been compared in two projects

- Hypotheses
  - H_0: defect density is the same in both projects
  - H_A: defect density is not the same
- Data: Defect density results for project x and project y
  - $x = 3.42, 2.71, 2.84, 1.85, 3.22, 3.48, 3.68, 4.30, 2.49, 1.54$
  - $y = 3.44, 4.97, 4.76, 4.96, 4.10, 3.05, 4.09, 3.69, 4.21, 4.40, 3.49$
- Data Sizes and Means:
  - $n = 10$ - size of x
  - $m = 11$ - size of y
  - $\bar{x} = 2.853$
  - $\bar{y} = 4.1055$

t-Test Example

- Sample variances:
  - $S_x^2 = 0.6506$
  - $S_y^2 = 0.4112$
- Calculations
  - $t_0 = -3.96$
  - $S_p = 0.7243$
  - $df = n + m - 2 = 10 + 11 - 2 = 19$
- Statistic
  - $t_{0.025,19} = 2.093$
  - Since $|t_0| > t_{0.025,19}$ we can reject H_0 with a two tailed test at the 0.05 level.

Paired t-Test Overview

- Compares two samples from repeated measures
- **Input**: Paired samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- **Hypotheses**:
  - H_0: $\mu_d = 0$, where $d_i = x_i - y_i$
  - Two-Sided H_A: $\mu_d \neq 0$
  - One-Sided H_A: $\mu_d > 0$
- **Calculations**:
  - Degrees of freedom: $df = n - 1$
  - $t_0 = \frac{\bar{d}}{S_d/(\sqrt{n})}$
  - $S_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$
- **Criterion**:
  - Two-Sided: reject H_0 if $|t_0| > t_{\alpha/2,df}$
  - One-Sided: reject H_0 if $|t_0| > t_{\alpha,df}$

Paired t-Test Example

Ten programs independently developed two different programs. They measured the effort required, as shown in the table

- Hypotheses
  - H_0: required effort to develop program 1 is the same as for program 2
  - H_A: it is not

Programmer	1	2	3	4	5	6	7	8	9	10
Program 1	105	137	124	111	151	150	168	159	104	102
Program 2	86.1	115	175	94.9	174	120	153	178	71.3	110

Paired t-Test Example

- Calculation:
  - $d = 18.9, 22, -51, 16.1, 23, 30, 15, 19, 32.7, 9$
  - $S_d = 27.358$
  - $t_0 = 0.39$
  - $df = n - 1 = 10 - 1 = 9$
- Statistics
  - $t_{0.025,9} = 2.262$
- Result:
  - Since $t_0 < t_{0.025,9}$ we cannot reject H_0 at the 0.05 level

F-Test Overview

- Compares variances of two **independent** samples
- **Input**: samples $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$
- **Hypotheses**:
  - H_0: $\sigma_x^2 = \sigma_y^2$
  - Two-Sided: H_A: $\sigma_x^2 \neq \sigma_y^2$
  - One-Sided: H_A: $\sigma_x^2 > \sigma_y^2$
- **Calculations**:
  - $F_0 = \frac{\max(S_x^2, S_y^2)}{\min(S_x^2, S_y^2)}$
  - S_x^2 and S_y^2 are sample variances
- **Criterion**
  - Degress of Freedom: $df_1 = n_{min} - 1$ and $df_2 = n_{max} - 1$
  - Two-Sided: reject H_0 if $F_0 > F_{\alpha/2, df_1, df_2}$
  - One-Sided: reject H_0 if $F_0 > F_{\alpha, df_1, df_2}$ and $S_x^2 > S_y^2$

F-Test Example

Defect density in different programs have been compared in two projects

- Hypotheses
  - H_0: both project defect densities have the same variance
  - H_A: they do not
- Data: Defect density results for project x and project y
  - $x = 3.42, 2.71, 2.84, 1.85, 3.22, 3.48, 3.68, 4.30, 2.49, 1.54$
  - $y = 3.44, 4.97, 4.76, 4.96, 4.10, 3.05, 4.09, 3.69, 4.21, 4.40, 3.49$

F-Test Example

- Data Sizes and Means:
  - $n_{min} = 10$ - size of x
  - $n_{max} = 11$ - size of y
  - $df_1 = n_{min} - 1 = 9$
  - $df_2 = n_{max} - 1 = 10$
- Calculations
  - $S_x = 0.6506$
  - $S_y = 0.4112$
  - $F_0 = 1.58$
- Statistic
  - $F_{0.025,9,10} = 3.78$
- Result
  - $F_0 < F_{0.025,9,10}$, fail to reject H_0 at 0.05 level

ANOVA Overview

- Used to analyze experiments of many different designs.
- Looks at the total variability of the data as well as the variability of different components
- **Input**: a samples: $x_{11}, x_{12}, \ldots, x_{1n_1}; x_{21}, x_{22}, \ldots, x_{2n_2}; \ldots; x_{a1}, x_{a2}, \ldots, x_{an_a}$
- **Hypotheses**:
  - H_0: $\mu x_1 = \mu_{x_2} = \ldots = \mu_{x_a}$
  - H_A: $\mu x_i \neq \mu_{x_j}$ where $i \neq j$

ANOVA Overview

- **Calculations**:
  - $SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{x_{..}^2}{N}$
  - $SS_{Treatment} = \sum_{i=1}^{a} \frac{x_{i.}^2}{n_i} - \frac{x_{..}^2}{N}$
  - $SS_{Error} = SS_T - SS_{Treatment}$
  - $MS_{Treatment} = SS_{Treatment}/(a-1)$
  - $MS_{Error} = SS_{Error}/(N-a)$
  - $F_0 = MS_{Treatment}/MS_{Error}$
  - N is the total number of measurements
  - a is the number of groups
  - $x_{i.} = \sum_j x_{ij}$

ANOVA Overview

Source of variation	Sum of Squares	DF	Mean Square	F_0
Between Treatments	$SS_{Treatment}$	$df_1 = a - 1$	$MS_{Treatment}$	$F_0 = \frac{MS_{Treatment}}{MS_{Error}}$
Error	SS_{Error}	$df_2 = N - a$	MS_{Error}	
Total	SS_T	$N - 1$		

- **Criterion**:
  - reject H_0 if $F_0 > F_{\alpha, df_1, df_2}$

ANOVA Example

The module size in three different programs have been measured.

- Hypotheses:
  - H_0: mean module size is the same across programs
  - H_A: at least one program's mean module size is different
- Data:
  - Program 1: 221, 159, 191, 194, 156, 238, 220, 197, 197, 194
  - Program 2: 173, 171, 168, 286, 206, 140, 226, 248, 189, 208, 213
  - Program 3: 234, 188, 181, 207, 266, 153, 190, 195, 181, 238, 191, 260

ANOVA Example

- Calculations:

Source of variation	Sum of Squares	DF	Mean Square	F_0
Between treatments	579.0515	2	289.5258	0.24
Error	36,151	30	1,205	
Total	36,730	32		

- Error row also called "Within treatments"

- Statistic:
  - $df_1 = a - 1 = 3 - 1 = 2$ and $df_2 = N - a = 33 - 3 = 30$
  - $F_{0.025,2,30} = 4.18$
- Result
  - Since $F_0 < F_{0.025,2,30}$, fail to reject H_0

Multiple Comparison

- ANOVA: We rejected H_0, what's next?
  - Contrasts
  - Multiple Comparison
- Multiple Comparison Procedures
  - Bonferroni's MCP
  - Tukey's HSD
  - Sidak's MCP
  - Fischer's LSD
  - Dunnett's Comparison to Control

Are there any questions?