Statistical Methodology for Software Engineering

Hadas Lapid, PhD

Contents

- Pairwise Comparison of Means
- Correlation

References:
https://en.wikipedia.org/wiki/Tukey%27s_range_test
https://en.wikipedia.org/wiki/Studentized_range_distribution
On Lyman & Longnecker: Section 9.2, P 454. Section 9.5, P. 468-471

https://en.wikipedia.org/wiki/Correlation_and_dependence

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Pairwise Comparison of Means

Tested Hypothesis:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$

H_A: otherwise

- If H_0 is retained, we're done.
- Otherwise, we wish to perform pairwise comparisons to find which of the pairs is different:

$$H_0: \mu_i = \mu_j$$

$$H_A: \mu_i \neq \mu_j$$

$$(i, j) \in \{1, \dots k\}$$

→ **Multiple hypothesis testing problem arises**

Contrast t-test

Hypothesis Testing Assumptions

- (i, j) are two independent samples
- $\sigma_i = \sigma_j$
- Mutual variance can be estimated from the overall ANOVA variance, MSE
- MSE has n-k degrees of freedom
- The statistical estimate:

$$t_{i,j} = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

$\boxed{\text{Decide to reject/retain } H_0 \text{ based on } t_{(n\text{-}k)} \text{ distribution}}$

Numerical Example of Contrast t-test

$$H_0: \mu_B = \mu_C$$
$$H_A: \mu_B \neq \mu_C$$

Treatment	Sample Mean	Sample Size
B	168.1	8
C	193.8	8

equation's reminder:

$$MSE = \frac{\sum_{i=1}^{k}(n_i - 1) \cdot S_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

$$t_{i,j} = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

$$MSE = \frac{6*657.14+7*592.41+7*426.79}{6+7+7} = 553.86$$

df_E = 23-3 = 20, $t_{(20,0.975)}$=2.085

$$\mathbf{t_{C,B}} = \frac{193.8 - 168.1}{\sqrt{553.86 \cdot \left(\frac{1}{8} + \frac{1}{8}\right)}} = 2.184$$

$$\left|\mathbf{t_{C,B}}\right| > t_{(20,0.975)}$$

Reject H_0 at 95% confidence

B and C treatments are significantly different

Multiple Comparison tests
Controlling the FWER

- Bonferroni & Holmes-applicable but conservative
- Tukey HSD – Honestly Significant Difference test (Tukey-Kramer, Tukey-HSD) is less conservative but reliable method to control the FWER at significance level α
- Critical values reflect **the distribution of maximal difference between pairwise means in k groups**
- Also called the **Tukey Distribution**

Multiple Comparison Studentized (Tukey) distribution test

Suppose N samples from k populations

$Y_i \sim N(\mu i, \sigma^2)$

$\bar{Y}_{min}$- smallest population mean

$\bar{Y}_{max}$- largest population mean

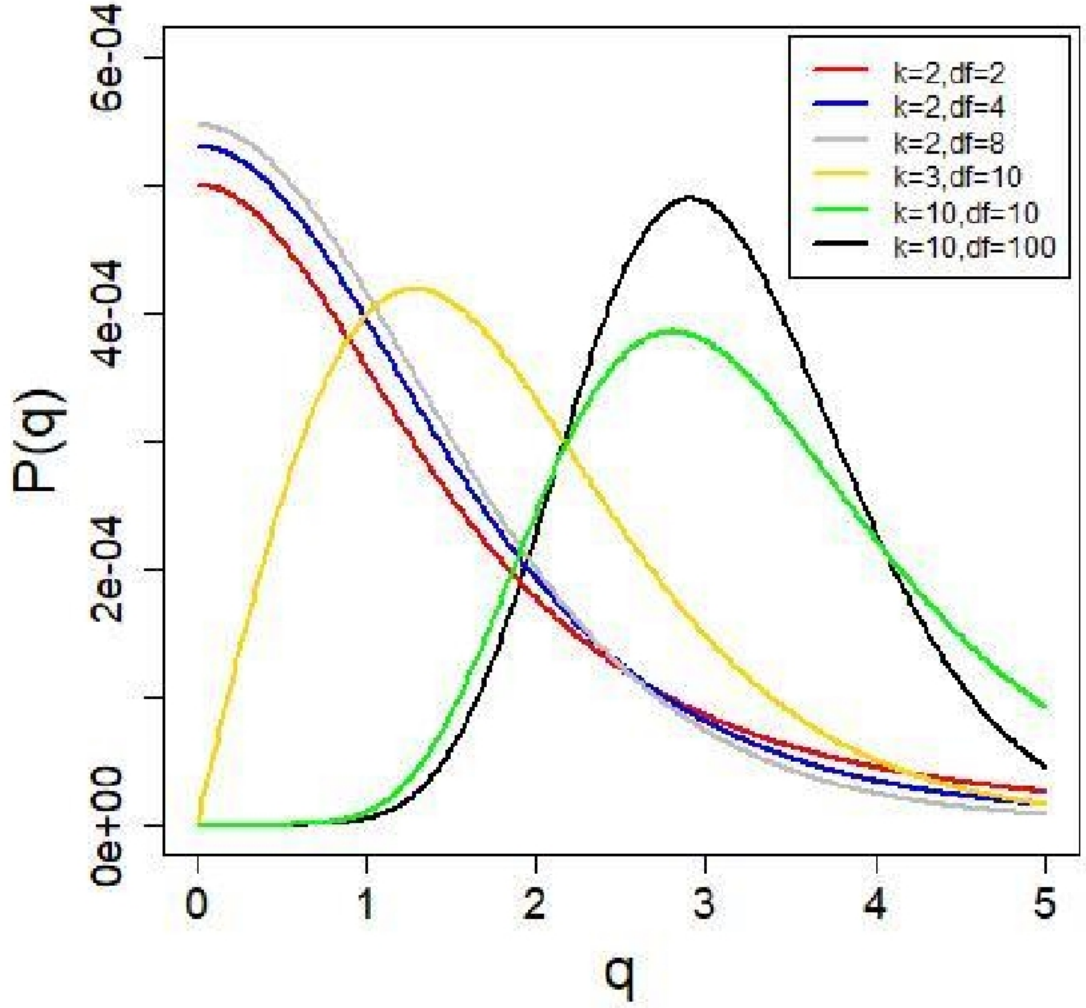
σ^2 is the pooled sample variance (the MSE)

→ q follows studentized range distribution

$$\mathbf{q_{i,j}} = \frac{\mathbf{\bar{Y}_i - \bar{Y}_j}}{\sqrt{\mathbf{\frac{1}{2} \cdot MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}} \qquad \mathbf{q^*} = \frac{\mathbf{\bar{Y}_{max} - \bar{Y}_{min}}}{\boldsymbol{\sigma}/\sqrt{\mathbf{n}}}$$

Multiple contrast t-tests

Studentized range distribution



The Tukey Correction

Calculation of studentized critical value in R:

$q^*_{(3,20,0.05)}$ = qtukey(k,n-k,α) = 2.53

$$q_{C,B} = \frac{193.8 - 168.1}{\sqrt{\frac{1}{2} \cdot 553.86 \cdot \left(\frac{1}{8} + \frac{1}{8}\right)}} = 3.089$$

Decide to reject H_0 if $|q_{ij}| > q^*$

3.089 > 2.53 → reject H_0

Studentized Range q Tables

q tables can be found here:
http://www.real-statistics.com/statistics-tables/studentized-range-q-table/

Multiple contrast t-tests
R output

Given k groups and n samples
Provides **Adjusted $\mathbf{P_{values}}$** to a given set of pairwise comparisons in ANOVA output model

TukeyHSD(aov(vals~ind,data=y))

	diff	lwr	upr	p adj
B-A	10.26786	-20.547673	41.08339	0.68131716
C-A	35.89286	5.077327	66.70839	0.02082549
C-B	25.62500	-4.145631	55.39563	0.09970565

Multiple Hypothesis Testing Summary

- Y is normally distributed in each sub-population.
- Equality of Variances: $\sigma_i = \sigma_j \forall i, j \in \{1..k\}$ (homoscedasticity)
- If variances are not equal:

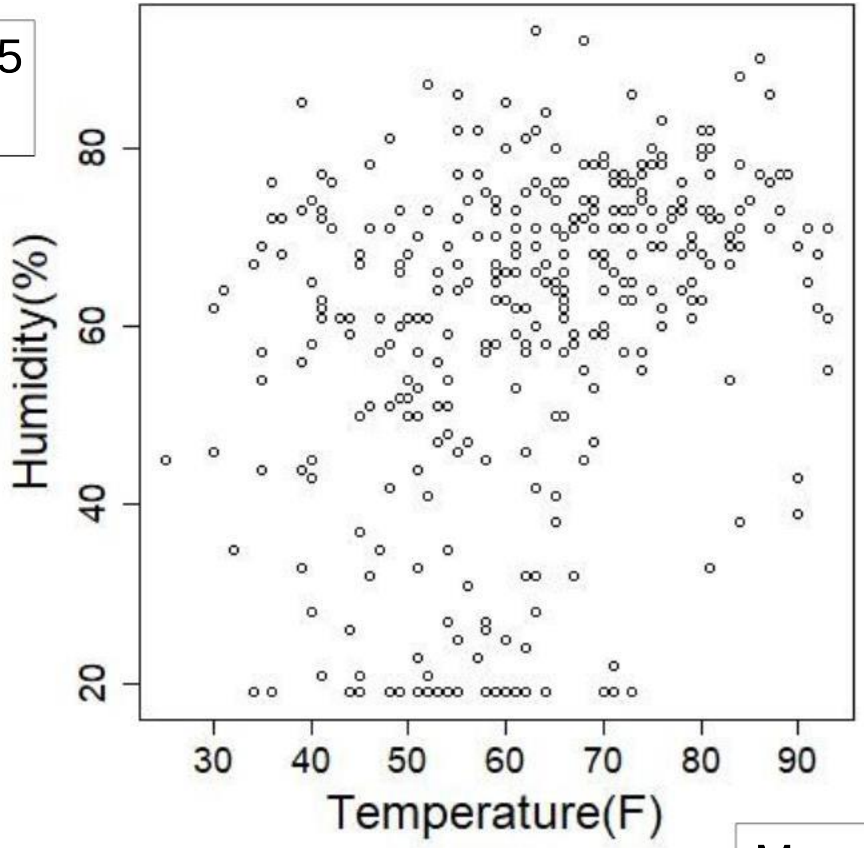
- **Discard** small samples which has deviated variances
- Use **Kruskal-Wallis** non-parametric test (extension of Wilcoxon Rank-Sum test)
- **Transform** Y (e.g., use log(Y))

Correlation

Definition:

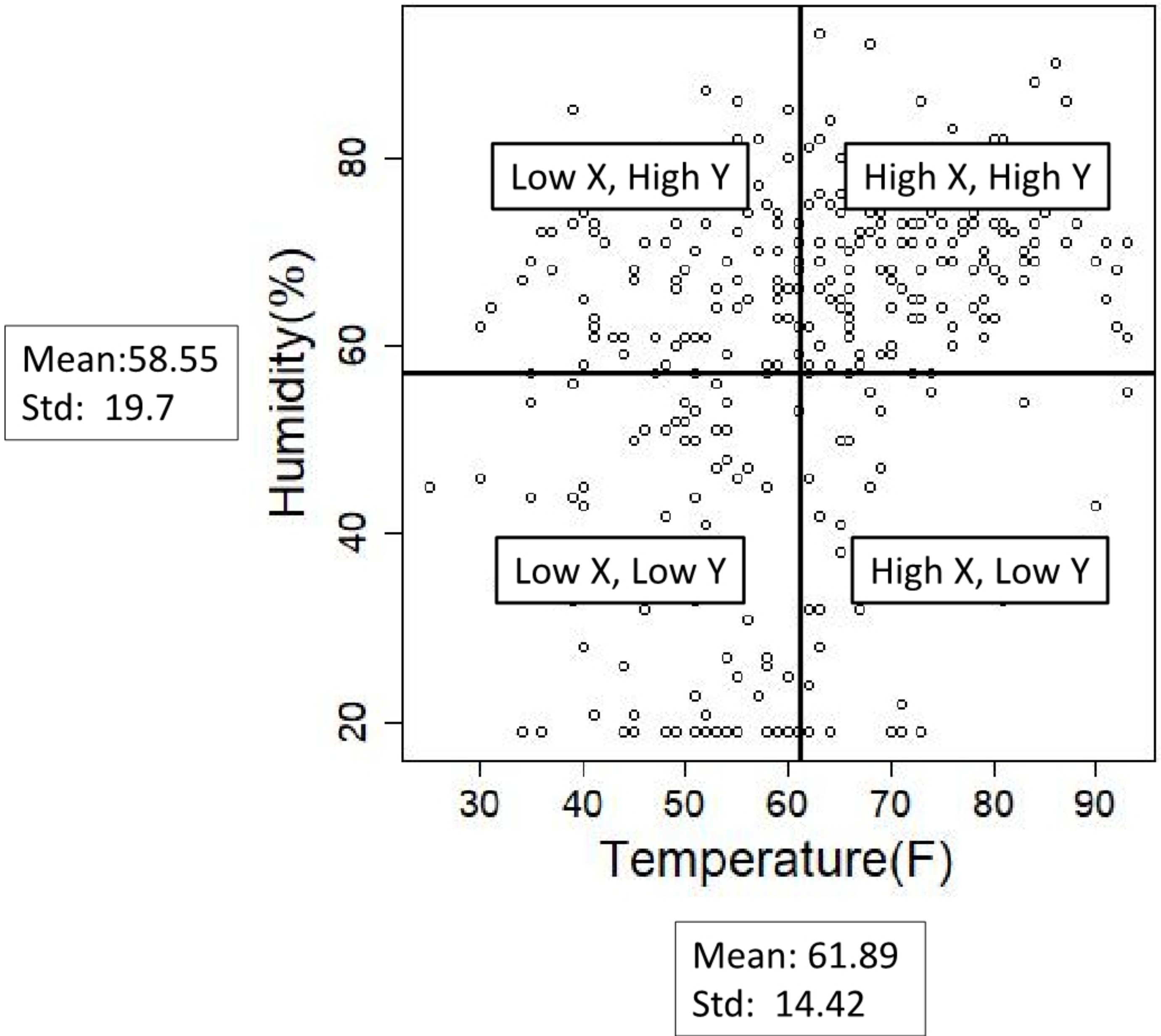
the strength of association between two random variables

Mean:58.55
Std: 19.7



Mean: 61.89
Std: 14.42

Correlation



Positive Correlation Coefficient:

%[High X, High Y] | [Low X, Low Y] ⬆

Negative Correlation Coefficient:

%[Low X, High Y] | [High X, Low Y] ⬆

Calculation of Correlation Coefficient

$$z_{xi} = \frac{(x_i - \bar{x})}{s_x}$$

$$r_i = z_{xi} \cdot z_{yi}$$

	Humidity	Temperature	Z_humidity	Z_Temperature	R_Humid_Temp
3	28	40	-1.55103502	-1.518813032	2.355732208
4	37	45	-1.09414662	-1.172047529	1.282391847
5	51	54	-0.38343133	-0.547869623	0.210070381
6	69	35	0.53034547	-1.865578536	-0.989401117
7	19	45	-2.00792342	-1.172047529	2.353381688
8	25	55	-1.70333116	-0.478516523	0.815072102
9	73	41	0.73340698	-1.449459932	-1.063044026
10	59	44	0.02269169	-1.241400630	-0.028169475

$$\boxed{r = \frac{1}{n-1} \cdot \sum_{i=1}^{n} Z_{x_i} \cdot Z_{y_i}} = \frac{1}{n-1} \sum_{i=1}^{n} r_i = 0.349$$

Properties of Correlation Coefficient r

- In the range of r is [-1,1]
  (Negative and Positive Association)
- Values of exactly ±1 are highly suspicious (indicates synthetic correlation)
- Correlation is constant with measurement scale
- r measures only monotonous association

The Significance of Correlation Coefficient

- Sample correlation coefficient, r, is an estimate of population coefficient, ρ
- Test the Hypothesis:

$$H_0:\ \rho = 0$$
$$H_1:\ \rho \neq 0$$

- Define the Fisher Transform of r, **F(r) = r*** :

$$r^* \sim N\left(\mu = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right), \sigma^2 = \frac{1}{n-3}\right)$$

- To test H_0, calculate the **statistical estimate Z** such that:

$$z = r^* \cdot \sqrt{n-3} = r^*/\sigma$$

- Reject H_0 if $|z| > z_{0.975} = 1.96$ (for 2-sided hypothesis)

Alternative calculation for The Significance of Correlation Coefficient

- Sample correlation coefficient, r, is an estimate of population coefficient, ρ
- Test the Hypothesis:

$$H_0:\ \rho = 0$$
$$H_1:\ \rho \neq 0$$

- Use ***bivariate normal distribution*** with unknown σ and df = n-2
- Define **t-Statistic** that is normally distributed around 0:

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

- Reject H_0 if $|t| > t_{0.975,n-2}$ (2-sided hypothesis)

The Significance of Correlation Coefficient

Numerical Example

- $r^* = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) = 0.5 \cdot \ln\left(\frac{1.349}{0.651}\right) = 0.3645$
- $z = r^* \cdot \sqrt{n-3} = 0.3645 \cdot \sqrt{17} = 6.780$
- $Z > Z_{0.975}$: 6.78>1.96 → reject H_0
- **Conclusion**:

The association between Humidity and Temperature of 0.349 is statistically significant at 95% confidence