


2015

## Statistics Notes

Saverio Perugini

*University of Dayton*, [sperugini1@udayton.edu](mailto:sperugini1@udayton.edu)

Follow this and additional works at: [http://ecommons.udayton.edu/cps\\_wk\\_papers](http://ecommons.udayton.edu/cps_wk_papers)

 Part of the [Artificial Intelligence and Robotics Commons](#), [Computer Security Commons](#), [Databases and Information Systems Commons](#), [Graphics and Human Computer Interfaces Commons](#), [Other Computer Sciences Commons](#), [Programming Languages and Compilers Commons](#), [Systems Architecture Commons](#), and the [Theory and Algorithms Commons](#)

---

### eCommons Citation

Perugini, Saverio, "Statistics Notes" (2015). *Computer Science Working Papers*. Paper 2.  
[http://ecommons.udayton.edu/cps\\_wk\\_papers/2](http://ecommons.udayton.edu/cps_wk_papers/2)

This Working Paper is brought to you for free and open access by the Department of Computer Science at eCommons. It has been accepted for inclusion in Computer Science Working Papers by an authorized administrator of eCommons. For more information, please contact [frice1@udayton.edu](mailto:frice1@udayton.edu).

# Statistics Notes

**Saverio Perugini**

Department of Computer Science  
University of Dayton  
300 College Park  
Dayton, Ohio 45469-2160 USA

Tel: +001 (937) 229-4079

Fax: +001 (937) 229-2193

E-mail: [saverio@udayton.edu](mailto:saverio@udayton.edu)

www: <http://academic.udayton.edu/SaverioPerugini>

# Contents

<b>1</b>	<b>Fundamentals</b>	<b>3</b>
1.1	Foundational Statistical Terms, Definitions, and Formulae . . . . .	3
1.2	External Reference Distribution . . . . .	4
1.3	Normal Distribution . . . . .	4
1.3.1	Properties of the Normal Distribution . . . . .	4
1.3.2	Empirical Rule for the Normal Distribution . . . . .	4
1.4	Student's $t$ Distribution . . . . .	4
<b>2</b>	<b>Means, Variances, and Analysis of Variance</b>	<b>5</b>
2.1	Means . . . . .	5
2.1.1	One Population Mean . . . . .	5
2.1.2	Two Population Means (unpaired) . . . . .	5
2.1.3	Two Population Means (paired) . . . . .	6
2.2	Variances . . . . .	6
2.2.1	One Population Variance . . . . .	6
2.2.2	Two Population Variances . . . . .	6
2.3	Analysis of Variance (ANOVA) Table . . . . .	7
2.4	Confidence Interval Formulae . . . . .	7
<b>3</b>	<b>Experimental Design</b>	<b>7</b>
3.1	Completely Randomized Design . . . . .	7
3.2	Randomized Block Design . . . . .	8
3.3	Two-Way Factorial Design . . . . .	8
3.4	Latin Square Design . . . . .	8
3.5	Factorial Design . . . . .	9

# 1 Fundamentals

## 1.1 Foundational Statistical Terms, Definitions, and Formulae

Term	Definition	Population	Sample
<b>Mean</b>	a measure of location	$\eta = \frac{\sum_{i=1}^n y_i}{N} \Rightarrow$ parameter	$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \Rightarrow$ statistic
<b>Variance</b>	a measure of spread	$\sigma^2 = \frac{\sum_{i=1}^n (y_i - \eta)^2}{N}$	$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}$
<b>Other</b>	you use this when you <b>know</b> $\eta$		$\dot{s}^2 = \frac{\sum_{i=1}^n (y_i - \eta)^2}{n}$
<b>Standard Deviation</b>	a measure of spread	$\sigma = \sqrt{\frac{\sum_{i=1}^n (y_i - \eta)^2}{N}}$	$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$
<b>N (0, 1)</b>	Unit (Standard) Normal Distribution		$z = \frac{y - \eta}{\sigma}$ and $z = \frac{\bar{y} - \eta}{\frac{s}{\sqrt{n}}}$ if $\bar{y}$ is known
<b>Covariance</b>	a measure of linear dependence between two random variables $X$ and $Y$	$cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \eta_x)(y_i - \eta_y)}{N}$	$cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$
<b>Correlation Coefficient</b>		$corr(X, Y) = \frac{cov(X, Y)}{\sigma_x \sigma_y}$	$corr(X, Y) = \frac{cov(X, Y)}{s_x s_y}$

**Mode:** most frequently occurring

**Frequency Histogram:**  $f$  versus  $y$

**Probability Density Histogram:**  $p(y)$  (probability density) versus  $y$

## 1.2 External Reference Distribution

$H_0 : \eta_B - \eta_A = 0$   $y_B - y_A$  significant evidence  $\Rightarrow$  reject  $H_0$  (null hypothesis) and conclude  $H_A$   
(alternate hypothesis)

$H_A : \eta_B - \eta_A > 0$  insignificant evidence  $\Rightarrow$  fail to reject  $H_0$  (you never accepts  $H_0$ )  
use significance level ( $\alpha$ ) to decide (typically  $\alpha = 0.05$  or  $\alpha = 0.01$ )

## 1.3 Normal Distribution

### 1.3.1 Properties of the Normal Distribution

- i) an equation
- ii) symmetric about  $\eta$
- iii) infinite tails
- iv) the area under the curve is 1

### 1.3.2 Empirical Rule for the Normal Distribution

- $\eta \pm 3\sigma$  : 99.7% of data
- $\eta \pm 2\sigma$  : 95.0% of data
- $\eta \pm 1\sigma$  : 68.0% of data

**Central Limit Theorem:** “*Averages*” (i.e.,  $\bar{y}$ 's) tend to be bell shaped and normally distributed; as  $n \rightarrow \infty$ , the distribution of the sample means approaches normality.

## 1.4 Student's $t$ Distribution

- Use the student's  $t$  distribution when you **do NOT know** the standard deviation ( $\sigma$ ) of the *population*.
- Use the fact of whether or not you know the mean ( $\eta$ ) of the *population* to determine whether to use  $s^2$  or  $\hat{s}^2$ .
- Use  $\hat{s}^2$  when you **know** the mean ( $\eta$ ) of the *population*.
- Use  $s^2$  when you **do NOT know** the mean ( $\eta$ ) of the *population*.
- If you use  $s$ , then  $t = \frac{y - \bar{y}}{s}$ .

- If you use  $\hat{s}$ , then  $t = \frac{y-\eta}{\hat{s}}$ .

As the sample gets larger, the t-distribution approaches the normal distribution.

For a population with mean  $\eta$  and variance  $\sigma^2$ , you take samples of size  $n$ . The averages  $y$  of all possible samples from this population have mean  $\eta$  and variance  $\frac{\sigma^2}{n}$ .

**Standard Error of Mean**  $\frac{s}{\sqrt{n}}, \frac{\sigma}{\sqrt{n}}$

## 2 Means, Variances, and Analysis of Variance

### 2.1 Means

#### 2.1.1 One Population Mean

Test Type	$H_0$ and $H_A$	Test Statistic	Table
<b>One-Sided Hypothesis Test</b>	$H_0 : \eta = \#$ $H_A : \eta > \#$	$z = \frac{y-\eta}{\sigma}$ $t = \frac{y-\bar{y}}{s}$	z table: normal distribution / t table: t-distribution
<b>Two-Sided Hypothesis Test</b>	$H_0 : \eta = \#$ $H_A : \eta \neq \#$	$z = \frac{y-\eta}{\sigma}$ $t = \frac{y-\bar{y}}{s}$	z table: normal distribution / t table: t-distribution

**Note: do not forget to double  $p$ .**

#### 2.1.2 Two Population Means (unpaired)

Test Type	$H_0$ and $H_A$	Test Statistic	Table
<b>One-Sided Hypothesis Test</b>	$H_0 : \eta_B - \eta_A = 0 \Rightarrow \eta_B = \eta_A$ $H_A : \eta_B - \eta_A > 0 \Rightarrow \eta_B > \eta_A$	$z = \frac{(\bar{y}_B - \bar{y}_A) - (\eta_B - \eta_A)}{\sigma \sqrt{\frac{1}{n_B} + \frac{1}{n_A}}}$ $t = \frac{(\bar{y}_B - \bar{y}_A) - (\eta_B - \eta_A)}{s \sqrt{\frac{1}{n_B} + \frac{1}{n_A}}}$	z table: normal distribution / t table: t-distribution
<b>Two-Sided Hypothesis Test</b>	$H_0 : \eta_B - \eta_A = 0 \Rightarrow \eta_B = \eta_A$ $H_A : \eta_B - \eta_A \neq 0 \Rightarrow \eta_B \neq \eta_A$	$z = \frac{(\bar{y}_B - \bar{y}_A) - (\eta_B - \eta_A)}{\sigma \sqrt{\frac{1}{n_B} + \frac{1}{n_A}}}$ $t = \frac{(\bar{y}_B - \bar{y}_A) - (\eta_B - \eta_A)}{s \sqrt{\frac{1}{n_B} + \frac{1}{n_A}}}$	z table: normal distribution / t table: t-distribution

**Note: do not forget to double  $p$ .**

**Notes:**

- We must assume that the two populations are independent of each other.
- We must assume that the variances of the populations are equal.

### 2.1.3 Two Population Means (paired)

Test Type	$H_0$ and $H_A$	Test Statistic	Table
One-Sided Hypothesis Test	$H_0 : \delta = 0$ $H_A : \delta > 0$	$\bar{d} = \sum_{i=1}^n \frac{(x_i - y_i)}{n}$ $t = \frac{\bar{d} - \delta}{\frac{s_{\bar{d}}}{\sqrt{n}}} = \frac{\bar{d} - \delta}{\frac{s}{\sqrt{n}}}$	t table: t-distribution
Two-Sided Hypothesis Test	$H_0 : \delta = 0$ $H_A : \delta \neq 0$	$\bar{d} = \sum_{i=1}^n \frac{(x_i - y_i)}{n}$ $t = \frac{\bar{d} - \delta}{\frac{s_{\bar{d}}}{\sqrt{n}}} = \frac{\bar{d} - \delta}{\frac{s}{\sqrt{n}}}$	t table: t-distribution

Note: do not forget to double  $p$ .

## 2.2 Variances

### 2.2.1 One Population Variance

Test Type	$H_0$ and $H_A$	Test Statistic	Table
One-Sided Hypothesis Test	$H_0 : \sigma^2 = \#$ $H_A : \sigma^2 > \#$	$\frac{s^2}{\sigma^2} \chi^2_N =$ $\frac{\sum_{i=1}^n (y - \eta)^2}{N} \sim$ $\frac{\sigma^2 \chi^2_N}{N} \sim$ $\frac{\sigma^2}{N} \chi^2_N / s^2 =$ $\frac{\sum_{i=1}^n (y - \bar{y})^2}{n-1} \sim$ $\frac{\sigma^2}{n-1} \chi^2_{n-1}$	$\chi^2$ distribution
Two-Sided Hypothesis Test			$\chi^2$ distribution

Note: do not forget to double  $p$ .

### 2.2.2 Two Population Variances

Test Type	$H_0$ and $H_A$	Test Statistic	Table
One-Sided Hypothesis Test	$H_0 : \sigma_A^2 = \sigma_B^2$ $\sigma_B^2 \Rightarrow \frac{\sigma_A^2}{\sigma_B^2} = 1$ $H_A : \sigma_A^2 > \sigma_B^2$ $\sigma_B^2 \Rightarrow \frac{\sigma_A^2}{\sigma_B^2} > 1$	$\frac{s_A^2}{s_B^2}$	F distribution

**Y** matrix (data) = **A** matrix (grand average) + **T** matrix ( $\bar{y}$  - grand average) + **R** matrix (residue)

**D** matrix = **Y** matrix - **A** matrix = **T** matrix + **R** matrix

### 2.3 Analysis of Variance (ANOVA) Table

Source	Sum of Squares	Degrees of Freedom	Mean Square
<b>Between (T)</b>	$S_T = \text{Sum of Squares}$	$\nu_T = \# \text{ of columns} - 1$	$S_T^2 = \frac{S_T}{\nu_T}$
<b>Within (R)</b>	$S_R = \text{Sum of Squares}$	$\nu_R = \Sigma(n - 1) \text{ for each column}$	$S_R^2 = \frac{S_R}{\nu_R}$
<b>Total (D)</b>	$S_D = \text{Sum of Squares}$ $= S_T + S_R$	$\nu_D = \nu_T + \nu_R$	$X$

F	P	Analysis
$\frac{S_T^2}{S_R^2}$		

### 2.4 Confidence Interval Formulae

Population Mean:  $\eta$

$\bar{y} \pm z_{\frac{\infty}{2}} \frac{\sigma}{\sqrt{n}}$	$\sigma$ known
$\bar{y} \pm t_{\frac{\infty}{2}} \frac{s}{\sqrt{n}}$	$\sigma$ unknown

Difference in Population Means:  $\eta_B - \eta_A$

$(\bar{y}_B - \bar{y}_A) \pm z_{\frac{\infty}{2}} \sigma \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$	$\sigma$ known
$(\bar{y}_B - \bar{y}_A) \pm t_{\frac{\infty}{2}} s \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$	$\sigma$ unknown

Mean of Paired Differences:  $\delta$

$\bar{d} \pm z_{\frac{\infty}{2}} \frac{\sigma}{\sqrt{n}}$	$\sigma$ known
$\bar{d} \pm t_{\frac{\infty}{2}} \frac{S_d}{\sqrt{n}}$	$\sigma$ known

## 3 Experimental Design

### 3.1 Completely Randomized Design

1 Treatment, no Blocks

$H_0 : \eta_A = \eta_B = \eta_C$

$H_A : \text{not all population means are equal}$

Matrix:  $Y = A + T + R \Rightarrow D = T + R$



## 3.2 Randomized Block Design

### 1 Treatment, 1 Block

$$H_0 : \eta_A = \eta_B = \eta_C = \eta_D$$

$H_A$  : not all population means are equal

$$H_0 : \tau_A = \tau_B = \tau_C = \tau_D = 0$$

$H_A$  :  $\tau_t \neq 0$  for some t

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$H_A$  :  $\beta_i \neq 0$  for some i

**Matrix:**  $Y = A + B + T + R \Rightarrow D = B + T + R$

## 3.3 Two-Way Factorial Design

### 2 Treatments at once, no Blocks, MUST replicate at least twice

$$H_0 : \tau_A = \tau_B = \tau_C = \tau_D = 0$$

$H_A$  :  $\tau_i \neq 0$  for some i

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = 0$$

$H_A$  :  $\tau_j \neq 0$  for some j

$$H_0 : \forall \omega, \omega = 0$$

$H_A$  :  $\omega_{ij} \neq 0$  for some ij

**Matrix:**  $Y = A + T_1 + T_2 + I + R \Rightarrow D = T_1 + T_2 + I + R$

## 3.4 Latin Square Design

### 1 Treatment, 2 Blocks, no Interactions

$$H_0 : \eta_A = \eta_B = \eta_C = \eta_D$$

$H_A$  : not all population means are equal

$$H_0 : \tau_A = \tau_B = \tau_C = \tau_D = 0$$

$H_A$  :  $\tau_t \neq 0$  for some t

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$H_A$  :  $\beta_i \neq 0$  for some i

$$H_0 : \beta_I = \beta_{II} = \beta_{III} = \beta_{IV} = 0$$

$H_A$  :  $\beta_j \neq 0$  for some j

**Matrix:**  $Y = A + B_1 + B_2 + T + R \Rightarrow D = B_1 + B_2 + T + R$

### 3.5 Factorial Design

**Any # of Treatments, no Blocks, Must Replicate at least twice**

$$H_0 : \eta_A = \eta_B = \eta_C$$

$H_A$  : not all population means are equal

#### **Treatment A**

$$H_0 : \tau_1 = \tau_2 = 0$$

$H_A$  :  $\tau_t \neq 0$  for some t

#### **Treatment B**

$$H_0 : \tau_1 = \tau_2 = 0$$

$H_A$  :  $\tau_t \neq 0$  for some t

#### **Treatment C**

$$H_0 : \tau_1 = \tau_2 = 0$$

$H_A$  :  $\tau_t \neq 0$  for some t

**Matrix:**  $Y = A + T_1 + T_2 + T_3 + R \Rightarrow D = T_1 + T_2 + T_3 + R$

**Tukey Test** is to test which pairs of population means are different. You only perform tukey tests if you rejected the null hypothesis.

$$q_{4,6,0.05} = \frac{4.90}{\sqrt{2}} = 3.46 \text{ abs } (\#) > 3.46$$