2015

Statistics Notes

Saverio Perugini
University of Dayton, sperugini1@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
WWW: http://academic.udayton.edu/SaverioPerugini

Contents

1 Fundamentals

1.1 Foundational Statistical Terms, Definitions, and Formulae

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

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 (α) 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 η

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 $n \to \infty$, 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 (σ) of the *population*.
- Use the fact of whether or not you know the mean (η) of the *population* to determine whether to use $\mathbf{s^2}$ or $\dot{\mathbf{s}}^2$.
- Use $\dot{\mathbf{s}}^2$ when you **know** the mean (η) of the *population*.
- Use $\mathbf{s^2}$ when you **do NOT know** the mean (η) of the *population*.
- If you use $\mathbf{s}$, then $t = \frac{y-\bar{y}}{s}$.

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

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

For a population with mean η and variance σ^2, you take samples of size n. The averages y of all possible samples from this population have mean η 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
One-Sided Hypothesis Test	$H_0 : \eta = \#$ $H_A : \eta > \#$	$z = \frac{y-\eta}{\sigma}$ $t = \frac{y-\overline{y}}{s}$	z table: normal distribution / t table: t-distribution
Two-Sided Hypothesis Test	$H_0 : \eta = \#$ $H_A : \eta \neq \#$	$z = \frac{y-\eta}{\sigma}$ $t = \frac{y-\overline{y}}{s}$	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
One-Sided Hypothesis Test	$H_0 : \eta_B - \eta_A = 0 \Rightarrow \eta_B = \eta_A$ $H_A : \eta_B - \eta_A > 0 \Rightarrow \eta_B > \eta_A$	$z = \frac{(\overline{y_B}-\overline{y}_A)-(\eta_B-\eta_B)}{\sigma\sqrt{\frac{1}{n_B}+\frac{1}{n_A}}}$ $t = \frac{(\overline{y}_B-\overline{y}_A)-(\eta_B-\eta_B)}{s\sqrt{\frac{1}{n_B}+\frac{1}{n_A}}}$	z table: normal distribution / t table: t-distribution
Two-Sided Hypothesis Test	$H_0 : \eta_B - \eta_A = 0 \Rightarrow \eta_B = \eta_A$ $H_A : \eta_B - \eta_A \neq 0 \Rightarrow \eta_B \neq \eta_A$	$z = \frac{(\overline{y}_B-\overline{y}_A)-(\eta_B-\eta_B)}{\sigma\sqrt{\frac{1}{n_B}+\frac{1}{n_A}}}$ $t = \frac{(\overline{y}_B-\overline{y}_A)-(\eta_B-\eta_B)}{s\sqrt{\frac{1}{n_B}+\frac{1}{n_A}}}$	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	$H_0 : \delta = 0$ $H_A : \delta > 0$	$\overline{d} = \sum_{i=1}^{n} \frac{(x_i - y_i)}{n}$ $t = \frac{\overline{d} - \delta}{s_{\overline{d}}} = \frac{\overline{d} - \delta}{\frac{s}{\sqrt{n}}}$	t table: t-distribution
Two-Sided Hypothesis Test	$H_0 : \delta = 0$ $H_A : \delta \neq 0$	$\overline{d} = \sum_{i=1}^{n} \frac{(x_i - y_i)}{n}$ $t = \frac{\overline{d} - \delta}{s_{\overline{d}}} = \frac{\overline{d} - \delta}{\frac{s}{\sqrt{n}}}$	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	$H_0 : \sigma^2 = \#$ $H_A : \sigma^2 > \#$	$\dot{s}^2 = \frac{\sum_{i=1}^{n}(y-\eta)^2}{N} \sim \frac{\sigma^2\chi^2}{N} \sim \frac{\sigma^2}{N}\chi_N^2 / s^2 = \frac{\sum_{i=1}^{n}(y-\overline{y})^2}{n-1} \sim \frac{\sigma^2}{n-1}\chi_{n-1}^2$	χ^2 distribution
Two-Sided Hypothesis Test			χ^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	$H_0 : \sigma_A^2 = \sigma_B^2 \Rightarrow \frac{\sigma_A^2}{\sigma_B^2} = 1$ $H_A : \sigma_A^2 > \sigma_B^2 \Rightarrow \frac{\sigma_A^2}{\sigma_B^2} > 1$	$\frac{s_A^2}{s_B^2}$	F distribution

Y matrix (data) = **A** matrix (grand average) + **T** matrix ($\overline{y}$ - 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
Between (T)	$S_T =$ Sum of Squares	$\nu_T = \#$ of columns - 1	$S_T^2 = \frac{S_T}{\nu_T}$
Within (R)	$S_R =$ Sum of Squares	$\nu_R = \Sigma(n-1)$ for each column	$S_R^2 = \frac{S_R}{\nu_R}$
Total (D)	$S_D =$ 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: η

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

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

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

Mean of Paired Differences: δ

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

3 Experimental Design

3.1 Completely Randomized Design

1 Treatment, NO Blocks

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

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