BIOMETRY FORMULAS

A Ready Reckoner for Biometrical Statistics

NOVEMBER 17, 2013

BIOMETRY

Contents

Question 1

Probability Rules

The probability of event A happening or event B happening or both is equal to the probability of event A happening or even B happening minus the probability of both events happening at the same time (which is already included in both the probabilities)

$$P(A \cup B) = P(A) + [P(B) - P(A \cap B)]$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The probability of event A given that event B has happened is equal to the probability of both events happening divided probability of event B

$$P(A|B) = \left(\frac{P(B \cap A)}{P(B)}\right)$$

The probability of both events A and B happening is equal to the probability of event A happening after B has happened multiplied by the probability of event B.

$$P(A \cap B) = P(A|B) \text{ x } P(B)$$

If events A and B are independent then P(A|B) just = P(A)

Independent If:

Two events A and B are considered to be independent from one another if the probability of event A happening given event B is equal to the probability of event A.

$$P(A|B) = \left(\frac{P(B \cap A)}{P(B)}\right) = P(A)$$

Mutually Exclusive if:

Two events are considered mutually exclusive i.e. cannot happen given the other has happened if mathematically they cannot intersect i.e. happen if the other has happened.

$$P(A \cap B) = \emptyset = 0$$

Nota Bene: Circular Logic

It is not possible to know one of the following three statements unless at least the other two are known:

1. P(B)
2. P(A|B)
3. P(A∩B)

Discrete Probability Distributions

x	0	1	2	3	4	5	6	7
$P(X=x)$	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X=4)$	$P(X=5)$	$P(X=6)$	$P(X=7)$

Validity

For a discrete probability distribution to be valid it must be true that:

1. $P(X=x) < 0$
2. $\sum[\, P(X = x)\,] = 1$

Mean Value

To find the mean of a Discrete Probability Distribution:

$\sum[\, P(X = x)$ x $x] = \mu$ = E(x) i.e. the expected value of x

Standard Deviation

$\sigma^2 = E(x^2) - \mu^2$

$\mu^2 = E(x^2) = \sum[\, x^2$ x $P(X = x)]$

Binomial Distributions

When is it applicable

1. Experiments consists of n trials
2. Each trial is yes/no or success/fail or similar
3. Probability of yes is consistent for each trial
4. Each trial is independent
5. The experimentser is interested in *x*, the number of successes during n trials

Function

$$P(X = x) = \binom{n}{x} \text{ x } p^x \text{ x } q^{n-x} \quad \text{, where } x \in \text{Natural Numbers}$$

n: Is the number of trials

p: Is the probability of Success

q: Is the probability of failure, q=1-p

x: Is the number of successes in n trials

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

$$\mu = E(x) = n \text{ x } p$$

$$\text{Var (x)} = \sigma^2 = n \text{ x } p \text{ x } q = n \text{ x } p \text{ x } (1-p)$$

Poisson Distribution

When is it applicable

1. Data is Discrete
2. Data is Independent
3. Mean is known
4. Sample Sizes are known

Function

$$P(X = k) = \frac{e^{-\mu} \text{ x } \mu^{k}}{k!} \text{ , where } k \in \text{Natural Numbers}$$

Question 2

Error Types

Type 1 Error

The null hypothesis is incorrectly rejected, Probability of such error is α.

Type 2 Error

The null hypothesis was not rejected when it should have been.

Normal Distributions

$$Z = \frac{x - \mu}{\sigma} \sim N(\mu, \sigma^2), \quad x = Z\sigma + \mu$$

$$P(Z \leq z) = P(Z < z)$$

$$t = \frac{x - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

Sample Size in order to estimate μ, from Confidence Intervals

Refer to Lecture 7, Page 10.

Samples Taken from Normal Distribution, Probability of Sample Mean.

Refer to Lecture 6, Pages 2 & 3 of 6

This is done by the Central Limit Theorem

Normal Confidence Intervals and Hypotheses Testing (Wk. 8)

If $X \sim N(\mu,\sigma^2)$ or $N \geq 30$

And σ is known

Confidence Interval

A (1-α) x 100% Confidence Interval for μ is given by:

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \text{ x } \left(\frac{\sigma}{\sqrt{n}}\right)$$

Hypotheses Testing

Hypotheses:

Three types of Hypotheses occur:

1. H_0: μ = *g* then Ha: μ < *g*
2. H_0: μ = *g* then Ha: μ > *g*
3. H_0: μ = *g* then Ha: μ =/= *g*

Test Statistic

$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

Rejection Region

At significance level α, H_0 is rejected and H_a accepted for the following statistics:

1. H_0: μ = *g* then Ha: μ < *g*
    a. H_0 fails and H_a is accepted if $Z < -Z_\alpha$
2. H_0: μ = *g* then Ha: μ > *g*
    a. H_0 fails and H_a is accepted if $Z > Z_\alpha$
3. H_0: μ = *g* then Ha: μ =/= *g*
    a. H_0 fails and H_a is accepted if $|Z| > |Z_{\alpha/2}| \Rightarrow Z > Z_{\alpha/2}$ & $Z < -Z_{\alpha/2}$

If X~N (μ,σ^2) or N $\geq$ 30

But σ is not known

A (1-α) x 100% confidence interval for μ is given by:

$$\bar{x} \pm t_{\alpha/2} \text{ x } (\frac{s}{\sqrt{n}})$$

Hypotheses Testing

Hypotheses:

Three types of Hypotheses occur:

1. H_0: μ = *g* then Ha: μ < *g*
2. H_0: μ = *g* then Ha: μ > *g*
3. H_0: μ = *g* then Ha: μ =/= *g*

Test Statistic

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

Rejection Region

At significance level α, H_0 is rejected and H_a accepted for the following statistics:

1. H_0: μ = *g* then Ha: μ < *g*
    a. H_0 fails and H_a is accepted if $t < -t_{\alpha, d.f.}$
2. H_0: μ = *g* then Ha: μ > *g*
    a. H_0 fails and H_a is accepted if $t > t_{\alpha, d.f.}$
3. H_0: μ = *g* then Ha: μ =/= *g*
    a. H_0 fails and H_a is accepted if $|t| > |t_{\alpha/2, d.f.}| \Rightarrow t > t_{\alpha/2, d.f.}$ & $t < -t_{\alpha/2, d.f.}$

Where

$\bar{x}$: Is the sample mean

s: Is the sample standard deviation

n-1: Is the degrees of Freedom

σ: Is the population standard deviation

μ: Is the population mean

t_{n-1}: Is the critical point on the t distribution and is equal to $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

Chi Distribution (Wk. 12)

Used for analysing categorical data, each cell should contain at least 5 counts.

Hypotheses

H_0: Cells match hypothesised value for distribution at α confidence using the sample to predict the population

H_a: Cells do not match the hypothesised value

Test Statistic (p. 2 Wk. 13 Lecture Notes)

$$x^2 = \sum_{i=1}^{k} \left(\frac{(e_i - o_i)^2}{e_i} \right) \sim \textit{Chi Distribution}$$

Where:

- e_i is the expected value
  - e_i is equal to the legitimate expected value (which is usually the average value)
  - OR, $e_i = \frac{(row\ total)(Column\ total)}{(Grand\ Total)}$
- o_i is the observed value

Rejection Region

H_0 fails for $x^2 > x^2, n\text{-}1, \alpha$

α: Is the probability that H_0 is rejected incorrectly

The larger $x^2, n\text{-}1, \alpha$, the smaller α, the harder it is to accept H_0.

n: Is the number of cells and *d.f.* is the degrees of freedom = *n*-1

Difference in Population Means, Confidence Intervals, for Normal Data (Wk. 10)

If samples are independent and N ≥ 30

A (1-α) x 100% confidence interval of (μ_1 - μ_2) is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\alpha}{2}} \text{ x } \sqrt{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})}$$

Where:

σ can be approximated by s and the standard error of ($\bar{x}_1$ - $\bar{x}_2$) = $\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}$

If samples are independent but N<30 (Variance of samples must be equal)

For this method variances from both populations must be equal, unequal variances for n<30 are not suitable for this method and ergo such samples are not required of this unit in such a way.

A (1-α) x 100% confidence interval of (μ_1 - μ_2) is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2},n1+n2-2} \text{ x } \sqrt{(s_p^2(\frac{1}{n_1} + \frac{1}{n_2}))}$$

Where:

The sample pooled variation (S^2_P) is used as an estimate of the overall population standard deviation of both samples, if otherwise not known. In such a case that it is known the standard deviation should be used.

The Pooled Variance

Is essentially the average variance between the two samples (applicable where$\sigma^2_1 = \sigma^2_2$:

$$S_p^2 = \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

Difference in Population Means, Hypotheses Testing for ($\mu_1 - \mu_2$)

If samples are random and independent and:

- Sample sizes are individually greater than 30 (in which case the population standard deviation σ can be estimated with sample standard deviation s)

OR

- Sample sizes are from a normal population and σ is known

Hypotheses

Three hypotheses may be used, with the null hypotheses implying equal population means.

4. H_0: $\mu_1 - \mu_2 = 0$ then Ha: $\mu_1 < \mu_2$
5. H_0: $\mu_1 - \mu_2 = 0$ then Ha: $\mu_1 > \mu_2$
6. H_0: $\mu_1 - \mu_2 = 0$ then Ha: $\mu_1 =/= \mu_2$

Test Statistic

$$Z = \frac{(\bar{x}_1 + x_2) - (\bar{\mu}_1 - \mu_2)}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$

$$Z = \frac{(\bar{x}_1 + \bar{x}_2) - (0)}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$

Rejection Region

At significance level α, H_0 is rejected and H_a accepted for the following statistics:

7. H_0: $\mu_1 - \mu_2 = 0$ then Ha: $\mu_1 < \mu_2$
    a. H_0 fails and H_a is accepted if $Z < -Z_\alpha$
8. H_0: $\mu_1 - \mu_2 = 0$ then Ha: $\mu_1 > \mu_2$
    a. H_0 fails and H_a is accepted if $Z > Z_\alpha$
9. H_0: $\mu_1 - \mu_2 = 0$ then Ha: $\mu_1 =/= \mu_2$
    a. H_0 fails and H_a is accepted if $|Z| > |Z_{\alpha/2}| \Rightarrow Z > Z_{\alpha/2}$ & $Z < -Z_{\alpha/2}$

Question 3

ANOVA

One-Way ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F-Stat
Between Treatments	SST	*k*-1	$\text{MST} = \frac{\text{SST}}{k-1}$	$\frac{MST}{MSE}$
Within Treatments (Error)	SSE	*n-k*	$\text{MSE} = \frac{\text{SSE}}{n-k}$	
Total	SS	*n-1*		

Where:

- SS is the total variation in all samples, SS = SSE + SST
- SSE is the total variance present within all of the groups or treatmens
- SST is the variance between the groups
- $\bar{x}_{Grand}$ is the mean value of every single data point, which is equal to the mean of each group mean
- *k* is the number of treatments
- *n* is the total no. of data points
- Also the pooled sample variance (S_p) = MSE = $\frac{\text{SSE}}{n-k}$

SST – Variance between groups

SST is the total amount of variance in data between the different treatments or groups of data

$$\text{SST} = \sum_{i=1}^{k} [\bar{x}_i(\bar{x}_i - \bar{x}_{Grand})^2]$$

$$\text{SST} = \bar{x}_1(\bar{x}_1 - \bar{x}_{Grand}) + \bar{x}_2(\bar{x}_2 - \bar{x}_{Grand}) + \bar{x}_3(\bar{x}_3 - \bar{x}_{Grand}) + \bar{x}_k(\bar{x}_k - \bar{x}_{Grand})$$

SSE – Variance within groups

SSE is the total amount of variance in data within each group or treatment but not the variance in between data from different groups.

$$\text{SSE} = \sum_{i=1}^{k} [s_i^2 (n_i - 1)^2]$$

$$\text{SSE} = s_1^2 (n_1 - 1) + s_2^2 (n_2 - 1) + s_3^2 (n_3 - 1) + s_k^2 (n_k - 1)$$

SS – Total Variance

SS is the total variance between all the data.

$$\text{SS} = \sum_{i=1}^{k} [s_i^2 (n_i - 1)^2] + \sum_{i=1}^{k} [\bar{x}_i (\bar{x}_i - \bar{x}_{Grand})^2]$$

$$\text{SS} = \text{SSE} + \text{SST}$$

Hypotheses Testing

Hypotheses

If H_0: $\mu_1 = \mu_2 = \mu_{3...} = \mu_k$ & H_a : at least one μ is different

Test Statistic

$$F = \frac{MST}{MSE}$$

Rejection Region

H_0 rejected at a significance level of α if:

$F > F_{k-1,n-k,\alpha}$

THE ORDER OF *k-1, n-k, α,* IS IMPORTANT IN THE F – STAT:

k-1, n-k, α IS NOT THE SAME AS ~~*n-k, k-1,α*~~

Two-Way ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F-Statistic
Between Treatments	SST	k-1	$\text{MST} = \frac{\text{SST}}{k-1}$	$\frac{\text{MST}}{\text{MSE}}$
Between Blocks	SSB	b-1	$\text{MSB} = \frac{\text{SSB}}{b-1}$	$\frac{\text{MSB}}{\text{MSE}}$
Within Treatments (Error)	SSE	(k-1)(b-1)	$MSE = \frac{SSE}{(k-1)(b-1)}$	
Total	SS	n-1 = bk - 1		

Where:

- SS is the total variation in all samples, SS = SSE + SST
- SSE is the total variance present within all of the groups or treatments
- SST is the variance between the groups
- SSB is the total variance (mean squared distance) between the block means and the mean value of all data points ($\bar{x}_{Grand}$).
- $\bar{x}_{Grand}$ is the mean value of every single data point, which is equal to the mean of each group mean
- $\bar{x}_{sample,i}$ is the sample mean, one for each treatment.
- k is the number of treatments
- n is the total number of data points
- b is the total number of blocks
- Also the pooled sample variance (S_p) = MSE = $\frac{\text{SSE}}{n-k}$
- s is the sample standard deviation of the
- X is an arbitrary data point.

SST – Total Variance of the Treatments

This is the variance within the treatments, the total squared distance from the mean to the grand mean for each sample

$$\sum_{i=1}^{k}[b(\bar{x}_{sample,i} - \bar{x}_{Grand})^2]$$

SSB – Total Variance of the Blocks

This is the variance within blocks, the total squared distance between the block means and the mean of every data point i.e. the grand mean ($\bar{x}_{Grand}$)

$$\sum_{i=1}^{b}[k(\bar{x}_{sample,i} - \bar{x}_{Grand})^2]$$

SS – Total Variance

SS is the total variance between all the data, calculated just like normal variation but not divided by n or n-1, this is because the variance need not be the average variance of the sample but the total variance.

$$\text{SS} = \sum_{i=1}^{n}[(X_i - \bar{x}_{Grand})^2]$$

SSE – Total Variance of the Blocks

This is the total variance of all the data within the treatments; it represents errors in the data and is calculated by subtracting variance from within the blocks and the variance from within the treatments from the total variance of the data.

$$\text{SSE} = \text{SS} - \text{SSB} - \text{SST}$$

$$\text{SSE} = \sum_{i=1}^{n}[(X_i - \bar{x}_{Grand})^2] - \sum_{i=1}^{b}[k(\bar{x}_{sample,i} - \bar{x}_{Grand})^2] - \sum_{i=1}^{k}[b(\bar{x}_{sample,i} - \bar{x}_{Grand})^2]$$

$$\text{SSE} = \sum_{i=1}^{n}[(X_i - \bar{x}_{Grand})^2 - (\bar{x}_{sample,i} - \bar{x}_{Grand})^2 - b(\bar{x}_{sample,i} - \bar{x}_{Grand})^2]$$

Hypotheses Testing, Treatment Means

Hypotheses

If H_0: $\mu_{t1} = \mu_{t2} = \mu_{t3\ldots} = \mu_k$ & Ha : at least one μ is different

Test Statistic

$$F = \frac{MST}{MSE}$$

Rejection Region

H_0 rejected at a significance α level if:

$F > F_{k-1,(k-1)(b-1),\,a}$

Hypotheses Testing, Block Means

Hypotheses

If H_0: $\mu_{b1} = \mu_{b2} = \mu_{b\,3\ldots} = \mu_{b\,k}$ & H_a : at least one μ is different

Test Statistic

$$F = \frac{MSB}{MSE}$$

Rejection Region

H_0 rejected at a significance α level if:

$F > F_{b-1,\,(k-1)(b-1),\,a}$

THE ORDER OF *k-1, n-k, α,* IS IMPORTANT IN THE F – STAT:

k-1, n-k, α IS NOT THE SAME AS ~~*n-k, k-1,α*~~

Question 5

Linear Regression and Correlation (Lecture Notes 13.)

Linear Model

A population Linear Regression Line is given by

$y = \alpha + \beta x$

A Sample Linear Regression Line is given by

$\hat{y} = a + bx$

Linear Regression Assumptions

1. Normality
   a. Y values for each X value are normally distributed around the regression line.
   b. Probability distribution of error terms is normal.
2. For each x value, the 'spread' or standard deviation around the regression line is constant.

Estimation of Coefficients, by *Method of Least Squares*

$b = \frac{S_{xy}}{S_{xx}}$ and $a = \bar{y} - b\,\bar{x}$

Where:

$$\mathbf{S_{xy}} = \sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})] = \sum_{i=1}^{n}[x_i y_i] - \frac{\sum_{i=1}^{n}[x_i] \text{x } \sum_{i=1}^{n}[y_i]}{n}$$

$$\mathbf{S_{xx}} = \sum_{i=1}^{n}[(x_i - \bar{x})^2] = \sum_{i=1}^{n}[x^2{}_i] - \frac{[\sum_{i=1}^{n}[x_i]]^2}{n}$$

$$\mathbf{S_{yy}} = \sum_{i=1}^{n}[(y_i - \bar{y})^2] = \sum_{i=1}^{n}[y^2{}_i] - \frac{[\sum_{i=1}^{n}[y_i]]^2}{n}$$

Coefficient of Correlation

The Coefficient of for population (*P)* quantifies the strength and direction of the linear relationship that may exist between two variables. The greater the magnitude of the coefficient, the greater the relation.

P usually has to be estimated by the sample correlation coefficient r.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{xy}}}$$

Sample Coefficient of Determination r^2

The sample coefficient of determination is a ratio that explains how much of the difference between x and y can be explained by the linear regression, much like a percentage.

$$r^2 = \frac{\text{SSR}}{\text{SS}}$$

$$r^2 = \frac{\left(S_{xy}\right)^2}{S_{xx}\, S_{xy}}$$

Slope Confidence Intervals and Hypotheses Testing

Refer to Pages 14 and 15 of Week 13 Lecture Notes

Significance of Linear Relation

Refer to:

1. Page 14 of week 13 lecture notes for statistical test of slope (*b)* = 0
    a. If the slope = 0 then knowledge of x cannot predict y and linear relationship is false
    b. If the question asks if there is a <u>significant</u> linear relation use this test, unless a confidence interval has been done then refer to that and use the rho test.
2. Page 18 week 13 lecture notes for ρ test
    a. ρ (AKA rho) quantifies the strength and direction of a linear relation

Interpreting Slope

To interpret the slope of a linear regression, describe what it actually means.

e.g. if a linear regression has a slope of 3 then:

for every increase in x (state what it is cm, kg, L, etc.) y increases by 3 (units again)

Residual Error and Standard Error and Variation due to Error (P. 11 of Wk. 13)

SSE = $S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$ = Residual Sum of Squares

$\sqrt{\text{MSE}} = \frac{\text{SSE}}{\text{n}-2}$ = residual Error $\approx \sigma^2$

$\sqrt{\frac{\text{MSE}}{\text{S}_{xx}}}$ = the standard error of the estimated slope, *b*

Where:

- *n* is the number of points on the number plane NOT the total amount of data, e.g. the number of points is half of the total points of data.

Refer to page 11 of week 13 for ANOVA distribution for linear regression.