Formulating and testing hypotheses

Scales of data

- Nominal

Objects can be classified into categories

- Ordinal

The relation of greater and less than exists, but not known by how much

- Interval

Along with greater than or less than, by how much is also known

- Ratio

Ratio is preserved

CIRCULAR SCALE

Some important terms

- **Population**
- **Sample**
- **Parameter**
- **Constant**
- **Variable (random Variable)**
- **Statistic – Statistic – Statistics**
- **Estimator**
- **Estimate**
- **Degrees of freedom**
- **The Normal Distribution**

Hypothesis

A statement based on some prior information, some data analysis, intelligent guess, etc.

NULL Hypothesis

Hypothesis of no difference (or as it exists)

e.g., mean of tree height under different conditions is same (or statistically not different)

A simple null hypothesis of testing two means is written as:

$$H_0: \mu_1 = \mu_2$$

ALTERNATIVE Hypothesis

The experimenter's hypothesis
e.g., mean of tree height under condition A is more than that under condition B

$H_1: \mu_1 < \mu_2$

$H_1: \mu_1 > \mu_2$

$H_1: \mu_1 \neq \mu_2$

Hypotheses are always written in terms of Parameters, not statistics.

Data are then collected as per a plan to test the hypotheses developed. Depending upon the outcome, the hypothesis is accepted or rejected.

Errors in testing hypotheses

There is always a chance, however minimum, to accept a wrong null hypothesis or reject a true null hypothesis – Errors

When the null hypothesis is true (who knows?) and gets rejected – Error of type I

When the null hypothesis is false (who knows?) and gets accepted - Error of type II

Probabilities

The errors may occur some of the times (not always) – thus the notion of chance or probability is introduced.

Probability of committing error of type I: α

Probability of committing error of type II: β

Probability of not committing error of type II: $1 - \beta$ (Power)

What is the probability of accepting or rejecting a true null hypothesis?

The Decision Rule

The total probability of accepting or rejecting a true null hypothesis is divided in two parts:

0 to α (CR, RR or Level of Significance) and α to 1 (AR)

The test statistic is determined by a sample of observations – the statistic used depends upon the null hypothesis along with its probability

The null hypothesis is rejected if this probability (p –value or Sig.) is less than α

The decision is always interpreted in terms of the NULL HYPOTHESIS

Alternative rule

If p-value cannot be calculated owing to non-availability of software, a table is seen for the desired degrees of freedom. Reject null hypothesis if calculated value is more than tabulated value.

Notice the reversal in the rule

What is p-value?

Test Statistic is different for different samples – it has a probability distribution. Sig. or p-value is the probability of getting a value of test statistic as the one calculated from the sample observations.

Rare events are significant

Probability of committing type II error decides the power of the test.

Some important null hypotheses and their tests

- One sample against a standard
- Two independent samples
- Two dependent samples
- More than 2 samples
- Attributes and frequency tables (contingency tables or cross tabulations)

Assumptions, Non-parametric tests

Mean of one sample against a standard

- There is only one sample
- Mean of the sample is to be tested against a hypothesized mean (say A).
- H_0: $\mu = A$ against H_1: $\mu = A$ (two-directional)
- Or H_1: $\mu > A$ or $\mu < A$ (one-directional)

Example

An experiment was conducted to see if a particular treatment of soil can increase the nitrogen content of barren soil to say 50 units per ha.

Mean of one sample against a standard

15 locations were selected for giving the treatment (replications – 15)

Barren land implies nitrogen content was zero (if not zero, the method would be different)

$$H_0: \mu = 50$$

$$H_1: \mu > 50$$

Test : 1 sample t-test

Level of significance: 0.05

Mean of one sample against a standard

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

X-bar: Sample mean, μ: Hypothesized mean (50 in this experiment), n: Number of observations

S: Sample s.d. If not given, calculate. {note: denominator is (n-1), not n}

Mean of one sample against a standard

Example

The mean of the 15 observations = 45.3

Hypothesized mean is 50

n (number of observations) = 15

d.f. = 14

Sample s^2 = 23.52

Sample s = 4.85

t_{cal} = 3.73

t_{tab} = 1.761 (1-tailed alternative)

p-value = 0.002 (from software)

$t_{cal} > t_{tab}$

Inference: Reject H_o. Same inference as p-value < 0.05

For a 2-tailed alternative, compare with 0.025

The fallacy of significance

Means of samples of two populations

To compare two treatment means: Whether two soil treatments have the same effect on the nitrogen content?

15 locations in each (the sample size may be different in each treatment)

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2 \text{ or } \mu_1 < \mu_2 \text{ or } \mu_1 > \mu_2$$

Test : Independent samples t-test

Level of significance: 0.05

Means of samples of two populations

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

The X-bars in numerator: 2 sample means
n_1, n_2: Number of observations in i^{th} sample

S: pooled sample s.d. If not given, calculate from formula above or $\{(n_1\text{-}1)s_1^2 + n_2\text{-}1)s_2^2\}/(n1+n2\text{-}2)$

Means of samples of two populations

$n_1 = n_2 = 15$

Mean of sample 1 = 45.3. Mean of sample 2 = 38.8

s1 = 4.85. s2 = 4.80 (assumption of equal variances)

d.f. = 28

Pooled Sample s^2 = 23.28

Sample s = 4.82

t_{cal} = 3.709

t_{tab} = 1.701 (1-tailed alternative), 2.048 (2-tailed, under 0.025)

p-value = 0.001 (from software)

$t_{cal} > t_{tab}$

Inference: Reject H_o. Same inference as p-value < 0.05

For a 2-tailed alternative, compare with 0.025

Mean of related populations

H_0: d = 0, where d is the mean of differences

Suppose we want to see if the treatment increases soil nitrogen as compared to the previous levels. We have 15 locations and values are observed before and after treatment.

$$t_{calc} = \frac{\bar{d}}{s_d / \sqrt{n}}$$

Take difference of each pair of observations and find mean. If mean = 7.25 and S_d = 5.833. t_{cal} = 4.306 with 14 df. ttab = 1.796 p = 0.002, reject H_o.

The F – statistic

Ratio of two variances is distributed as F distribution with n_1, n_2 degrees of freedom

$$F = \sigma_1^2 / \sigma_2^2$$

When testing equality of variances, numerator is more than denominator

In Analysis of Variance, Error variance is always the denominator

Means of samples of three of more populations
Analysis Of Variance (ANOVA)

$$H_0: \mu_1 = \mu_2 = \mu_3 = ... = \mu_n$$

H_1 may be constructed accordingly (unidirectional or two-directional, non-directional)

Procedure:
Partition variance into variance due to known causes (treatment) and unknown causes (error)
Find ratio as known / unknown
Compare with F_{tab} or see p-values.

Means of samples of three of more populations Analysis Of Variance (ANOVA)

Steps:

1. Compute Raw Sum of square
2. Compute Correction Factor = Grand sum^2/N
3. Compute TSS = RSS – CF
4. Compute SSt = $\Sigma(T_i^2/n_i)$ – CF, T_i = sum of observations under i^{th} Treatment.
5. Compute ESS = TSS – SSt
6. DF: Total df = N – 1, Trt = no. of trts – 1, Error by difference
7. Mean SS = Total SS/DF
8. F = Trt MSS/Error MSS
9. F_{tab} = F value for N^r and D^r DF
10. Reject H_0 if $F_{cal} > F_{tab}$.

Means of samples of three of more populations
Analysis Of Variance (ANOVA)

One-way ANOVA – one set of treatments – say 3 soil treatments and to test of the mean nitrogen is significantly different in each 3.

trt1	trt2	trt3
42	39	35
41	39	41
39	42	42
37	47	38
43	36	37
46	40	41
248	**243**	**234**

G = 725

CF = (725)2/18 = 29201.39

RSS = 29375

TSS = 29375 – 29201.39 = 173.61

SSt = 29218.17 – 29201.39 = 16.78

Means of samples of three of more populations
Analysis Of Variance (ANOVA)

Analysis of Variance Table

Source of variation	DF	SS	MSS	F
Treatment	2	16.78	8.39	0.749
Error	15	156.83	11.20	
Total	17	173.61		

$F_{2,14} = 3.74$

When F is less than 1, Accept null hypothesis

When ANOVA is significant, carry out post hoc tests to find out which of the pairs are significantly different

Assumptions

- Continuous
- Normal Distribution
- Homogeneous variance across groups

When assumptions are not fulfilled, transform data or use Non-parametric tests

Testing independence of attributes

Example to test If there is any association between altitude and species

H_0: The altitude and species are independent of each other

The presence of species along the altitude gradient is observed as frequencies and a contingency table is made. If there are three altitudes and 4 species, the table will look like

Testing independence of attributes

Observed frequencies

	Altitude 1	Altitude 2	Altitude 3	Total
Spp1	75	52	36	163
Spp2	64	41	58	163
Spp3	49	41	29	119
Spp4	44	62	83	189
Total	232	196	206	634

4 x 3 contingency table

Testing independence of attributes

Determine expected frequencies: Given (actual or ratios) or from data
From Data by multiplying the corresponding row and column totals and dividing by grand total.
E.g. For spp1 and alt 1, expected frequency = 232x163/634

Expected frequencies

	Alt1	Alt2	Alt3
Spp 1	59.65	50.39	52.96
Spp2	59.65	50.39	52.96
Spp3	43.55	36.79	38.67
Spp4	69.16	58.43	61.41

Testing independence of attributes

$$\chi^2 = \Sigma(O - E)^2/E, \; df = (r-1)x(c-1)$$

χ_{cal}^2 = 32.53 with 6 df
χ_{tab}^2 = 1.64
Asymmetric distribution, see tabulated values for one and two sided.

Null hypothesis is rejected. There is an association between species and altitude.

Chi-square test is also used to test goodness of fit while validating models

Testing one, or 2 samples when number of observations is large

Large means 31 or more

T-test is replaced by z-test

Standard value for 5% level of significance is 1.96

End of the session on Testing Hypothesis