1A; Hypotheses Testing Single Population Samples

Topic 1A | Week 1 | Material Drawn From Week 1 Lecture

When comparing population means standard normal distributions are used when the population standard deviation is known and t-distributions are used when only the sample standard deviation is known.

Contents

Steps to Hypothesis Testing

1. Hypothesis
2. Test Statistic
3. Rejection Region
4. P-value
5. Conclusion

1. Hypothesis
    a. Describe the Problem
    b. List the Data
    c. State the null hypothesis and alternative hypothesis
2. Test Statistic
    a. Decide upon the correct test statistic
        i. Relative to population size, known variance etc.
    b. Find the Test statistic
3. Rejection Region
    a. Find the Rejection region
    b. State whether or not the null hypothesis is rejected
4. P-value
    a. Find the value of α required to reject the null hypothesis
5. Conclusion
    a. Write a statement answering the initial question relative to the confidence level

Estimate Sample Size for statistical Test[1]

The Confidence Interval for the population mean μ is given by

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

To find μ within B of the sample mean it must be true that

$$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq B$$

Rearranging to solve for the sample size n gives

$$n \geq \frac{\sigma\, z_{\alpha/2}}{B^2}$$

Thus the sample size of a sample that will give a confidence interval of μ at a confidence level of $1 - \alpha$ to within B of the sample mean is n:

$$n \geq \frac{\sigma\, z_{\alpha/2}}{B^2}$$

[1] Biometry Lecture 7 Page 10

Hypothesis Testing and Confidence Intervals

Where the population variance σ is known[2]

If a sample is from a normal population or is practically large (30 or greater) it can be modelled using a ***Standard Normal Distribution.***

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 $|Z| > |Z_{\alpha/2}|$ => $Z > Z_{\alpha/2}$ & $Z < -Z_{\alpha/2}$

Confidence Interval

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

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

[2] Refer to Biometry Week 8, Biometry Formula Sheet P. 9, P. 31 Applied Statistics Week 1 Lecture.

Where σ is estimated by s[3]

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 $|t| > |t_{\alpha/2, d.f.}|$ => $t > t_{\alpha/2, d.f.}$ & $t < -t_{\alpha/2, d.f.}$

Confidence Interval

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

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

[3] Refer to Wk. 8 of Biometry, Page 10 of Biometry Formula Sheet, P. 32 of Lecture Notes for Applied Statistics

Hypothesis Testing and Confidence Intervals for Proportions

Hypothesis Testing

Data

Let:

$p = \frac{x}{n}$, where x is the number of successes in n trials in a measured sample

$\pi = \frac{x}{n}$, where x is the number of successes in n trials for a population

Test Statistic

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Rejection Region

$H_A: \pi < \pi_0$, reject H_0 when $Z < -Z_\alpha$

$H_A: \pi > \pi_0$, reject H_0 when $Z > Z_\alpha$

$H_A: \pi = \pi_0$, reject H_0 when $Z > Z_{\frac{\alpha}{2}}$ or $Z < -Z_{\frac{\alpha}{2}}$, which is same as $|Z| > Z_{\frac{\alpha}{2}}$

Confidence Interval

A $(1-\alpha)$ Confidence Interval for π can be found by:

$$p \pm Z_{\alpha/2} \left(\sqrt{\frac{\pi_0(1-\pi_0)}{n}} \right.$$

Hypothesis Testing and Confidence Intervals for Population Variance

Hypothesis Testing

Hypothesis

Any of the three hypothesis may occur:

1. H_0: σ^2 = g, H_a: σ^2 < g
2. H_0: σ^2 = g, H_a: σ^2 > g
3. H_0: σ^2 = g, H_a: σ^2 =/= g

Test Statistic

$$t = \frac{(n-1)s^2}{\sigma^2}$$

Rejection Region

1. H_0: σ^2 = g, H_a: σ^2 < g
   a. H_0 fails for t < X_α
2. H_0: σ^2 = g, H_a: σ^2 > g
   a. H_0 fails for t > $X_{1-\alpha}$
3. H_0: σ^2 = g, H_a: σ^2 =/= g
   a. H_0 fails for t < $X_{1-\frac{\alpha}{2}}$
   b. H_0 fails for t > $X_{1-\frac{\alpha}{2}}$

Confidence Interval

If the original population of data that a sample is taken from is normally distributed then the expression$\frac{(n-1)s^2}{\sigma^2}$ has a chi square distribution with $n-1$ degrees of freedom.

Thus it can be shown:

For a confidence level of $1-\alpha$

$$\frac{(n-1)s^2}{\chi_{\frac{\alpha}{2},d.f.}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2},d.f.}}$$

$$\sqrt{\frac{(n-1)s^2}{\chi_{\frac{\alpha}{2},d.f.}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2},d.f.}}}$$

Chi Distribution

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

This is found by using a computer to measure the distribution of categorical data in samples.

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)

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

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*-1, α

α: Is the probability that H_0 is rejected incorrectly

The larger x^2, *n*-1, α, 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

Calculating Power

Statistical Power is the Probability that an incorrect null hypothesis will be rejected, it is best shown by way of example.

Actual Situation	***Decision***	
	Do Not Reject H_0	Reject H_0
H_0 is True	Correct Decision(1-α)	Type 1 Error (α)
H_0 is False	Type 2 Error(β)	Correct Decision ($1-\beta$)

Example (1.1 From Week 1 Lecture Notes)

An ISP stated that users average 10 hours a week of internet usage, it is already known that the standard deviation of this population is 5.2 hours. A sample of n=100 was taken to verify this claim.

A worldwide census determined that the average is in fact 12 hours a week not 10 hours.

Data

$n = 100$
$\sigma = 5.2$
$\mu = 10$
$\bar{x} = 11$

$\mu_{true} = 12$

$\alpha = 0.05$
β = ???

H_0: Users Average ten hours a week of internet usage

H_a: Users Average over ten hours a week of internet usage

Step 1: Find the Critical Sample Mean ($\bar{x}_{Critical}$)

Find the Critical Value of $\bar{x}$ (the sample mean) that determines when H_0 would be rejected

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

$$\bar{x} = \mu + z\,(\frac{\sigma}{\sqrt{n}})$$

For α = 0.05 H_0 is rejected for Z> $Z_{0.05}$ which translates to Z > 1.645, thus substitute z for 1.645 and solve for the sample mean ($\bar{x}$

$$\bar{x} = (10) + (1.645)\left(\frac{(5.2)}{\sqrt{(100)}}\right) = 10.8554$$

Thus H_0 Is rejected for a sample mean of 10.8554 hours per week at a confidence level of α = 0.005.

Step 2: Find the Difference between the Critical Sample Mean and the True Mean as a z-value.

Find the difference between the true population mean and the sample mean that would reject the hypotheses, then find that small section of the distribution relative to a standard normal distribution.

$$Z = \frac{\bar{x}_{Critical} - \mu_{True}}{(\frac{\sigma}{\sqrt{n}})}$$

$$Z = \frac{10.8554 - 12}{(\frac{5.2}{10})} = -2.20115$$

Step 3: Find β

$\beta = P(Type\ 2\ Error)$

$\beta = P(H_0\ is\ not\ rejected\ |H_0\ Is\ wrong)$

$\beta = P(\bar{X} > \bar{x}_{Critical}\ |\mu = 12)$

$= P(Z > -2.20115)$

$= 0.0136$

$$\therefore The\ probability\ of\ a\ type\ 2\ error\ in\ this\ test\ is\ 1.36\%$$

Step 4: Find the Power (1-β)

$Power = (H_0\ Is\ rejected\ |\ H_0\ is\ false)$

$= P(\bar{X} < \bar{x}_{Critical}\ |\mu = 12)$

$= 1 - \beta$

$= 1 - 0.0136$

$= 0.9864 = 98.6\%$

1B; Hypothesis Testing Two Population Samples

Topic 1B | Week 2 | Material Drawn From Week 2 Lecture

When comparing two population means t-distributions and standard normal distributions can be used, when comparing multiple population means an ANOVA table must be used (Topic 3 of this Unit).

Contents

Difference between Two Population Means (Independent Populations)

Hypotheses[4]

The hypothesis in all scenarios of comparing independent population means is the same.

$H_0: \mu_1 - \mu_2 = D$ (D can be anything e.g. zero)

1. $H_a: \mu_1 - \mu_2 < D,\ or$
2. $H_a: \mu_1 - \mu_2 > D,\ or$
3. $H_a: \mu_1 - \mu_2 \neq D$

Where σ is known

If the population Variance is known the *Standard Normal Distribution* is used.

Where Variance between populations are Equal

If the population variance is equal it doesn't matter, it is the same equation as if they were different, just sub in identical values for σ. (Refer below)

Where Variance between populations are NOT Equal

Hypotheses

As Above

Test Statistic

The test statistic used is essentially a combination of both test statistics given by:

$$Z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\ , where\ (\mu_1 - \mu_2) = D$$

If the test was of a single population mean it would simplify back to:

$$Z = \frac{(\bar{x}_1 - \mu_1)}{\sigma/\sqrt{n}}$$

Rejection Region

H_0 is rejected if:

1. $H_a: \mu_1 - \mu_2 < D$
   a. $Z < -Z_\alpha$
2. $H_a: \mu_1 - \mu_2 > D$
   a. $Z > Z\alpha$
3. $H_a: \mu_1 - \mu_2 \neq D$
   a. $Z < -Z_{\frac{\alpha}{2}}\ or\ Z > Z_\alpha\ Which\ is\ the\ Same\ as\ |Z| > |Z_{\frac{\alpha}{2}}|$

Confidence Interval

A $(1-\alpha)\%$ Confidence interval of $(\mu_1 - \mu_2)$ is given by

$$(\overline{x_1} - \overline{x_2}) \pm Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

[4] Applied statistics (200030), Week 2 Lecture Slides, p. 2

Where σ is unknown and s is used

If the population variance is not known the *Student's t Statistic* is used[5] because the sample variance is being used as an estimate and the *t-distribution* compensates for that.

Small Sample Sizes[6]

If sample sizes are considered small then the degrees of freedom does not equal (n_1-1) + (n_2-1), instead *Welch's Approximation* must be used.

$$d.f. = (n_1 - 1) + (n_2 - 1)\,, for\; n\; \geq 30$$

$$d.f. = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}, for\; n\; < 30$$

It should be noted that d.f.$_{Welch}$ < $(n_1 - 1) + (n_2 - 1)$

Where Variance between populations are NOT Equal[7]

Unless there is evidence to the contrary variances are usually considered to be equal, this method is sensitive to sample sizes due to the effect they have on degrees of freedom.

Hypothesis

As above

Test statistic

The test statistic is essentially a combination of both the statistics, the *t-distribution* is used because σ is unknown.

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\,, where\; (\mu_1 - \mu_2) = D$$

Rejection Region

H_0 is rejected if:

Where degrees of freedom (d.f.) is given by (d.f. = $n_1 + n_2 - 2$) (for $n_1 \geq 30$ and n2 $\geq$ 30)

1. $H_a: \mu_1 - \mu_2 < D$
   a. $t < -t_{\alpha, d.f.}$
2. $H_a: \mu_1 - \mu_2 > D$
   a. $t > t_{\alpha, d.f.}$
3. $H_a: \mu_1 - \mu_2 \neq D$
   a. $t < -t\; or\; t > t_{\alpha, d.f.}\; Which\; is\; the\; Same\; as\; |t| > |t_{\frac{\alpha}{2}, d.f.}|$

[5] *Ibid.* p. 2
[6] *Ibid.* p. 14
[7] *Ibid.* p. 9

Confidence Interval

$$(\overline{x_1} - \overline{x_2}) \pm t_{\frac{\alpha}{2},d.f.} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where Variance between populations are Equal (N needn't be > 30)

For this method sample sizes needn't necessarily be large.

Hypothesis

As above

Test Statistic

The test statistic essentially averages the variances to get a more accurate result, the average variance is referred to as the pooled variance (s_p^2)

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\ , where\ (\mu_1 - \mu_2) = D$$

Where pooled Variance is given by:

$$s_p^2 = \frac{s_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

Rejection Region

H_0 is rejected if:

Where degrees of freedom (d.f.) is given by (d.f. = $n_1 + n_2 - 2$) (for $n_1+n_2 \geq 30$)

1. $H_a: \mu_1 - \mu_2 < D$
   a. $t < -t_{\alpha,d.f.}$
2. $H_a: \mu_1 - \mu_2 > D$
   a. $t > t_{\alpha,d.f.}$
3. $H_a: \mu_1 - \mu_2 \neq D$
   a. $t < -t\ or\ t > t_{\alpha,d.f.}\ Which\ is\ the\ Same\ as\ |t| > |t_{\frac{\alpha}{2},d.f.}|$

Confidence Interval

A $(1 - \alpha)\%$ Confidence Interval of $(\mu_1 - \mu_2)$ is given by:

$$(\overline{x_1} - \overline{x_2}) \pm t_{\frac{\alpha}{2},d.f.} \times \mathrm{s_p}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Difference between Two Population Means (Dependent Populations or Paired Samples)

Where two populations are dependent on one another (e.g. height with shoes, height without shoes) the sample should first find the difference between the samples then deal with one set of sample data.

This helps reduce the variance in the sample and makes the differences in data more statistically pronounced. There is a good explanation of this on (http://vassarstats.net/textbook/)

Difference between Two Independent Population Proportions ($n\pi \geq 10\ and\ n(1-\pi) \geq 10$)

This method must utilise practically large samples, because that way the difference of proportions can be assumed to be normally distributed and the sample proportion can be used in place of π.

Data

Given two independent samples:

	Sample 1	***Sample 2***
Sample Size	n_1	n_2
Population Proportion	π_1	π_1
Sample Proprtion	p_1	p_2
Sample Successes	x_1	X_2

Two independent samples, with sample proportions of p and (usually unknown) population proportions of π, where:

$$p = \frac{x}{n}$$

Such that x is the number of successes in n trials.

Essential $\mu\ becomes\ \pi$ and $\bar{x}\ becomes\ p$ and the mean values are actually proportions of successes in trials

Hypothesis

The hypothesis is similar to ordinary comparisons of population means:

$$H_0: \pi_1 - \pi_2 = 0$$

1. $H_a: \pi_1 - \pi_2 < D,\ or$
2. $H_a: \pi_1 - \pi_2 > D,\ or$
3. $H_a: \pi_1 - \pi_2 \neq D$

Test Statistic

Test Statistic for $H_0: \pi_1 - \pi_2 = D$

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\}}}\ , only\ for\ practically\ large\ samples, where\ \pi_1 - \pi_2 = D$$

Pooled Sample Proportion for $H_0: \pi_1 = \pi_2$

If the null hypothesis is true the sample proportions should represent samples of the same population proportion, thus the proportions can be averaged (or pooled) to give a better estimate of the overall sample population proportion.

Pooled sample proportion is represented by $\bar{p}$.

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 p_1 + n_1 p_2}{n_1 + n_2}$$

This method can only be used if $\pi_1 - \pi_2 = 0$

Test Statistic for Pooled Variance (where $H_0: \pi_1 = \pi_2$)

$$Z = \frac{(p_1 - p_2)}{\sqrt{\bar{p}\,(1-\bar{p}\ )\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}}, only\ for\ practically\ large\ samples\ where\ H_0: \pi_1 - \pi_2 = 0$$

Rejection Region

H_0 is rejected for:

Where D can represent whatever discrete difference including 0.

1. $H_\alpha: \pi_1 - \pi_2 < D$
   a. $Z < -Z_\alpha$
2. $H_\alpha: \pi_1 - \pi_2 > D$
   a. $Z > Z\alpha$
3. $H_\alpha: \pi_1 - \pi_2 \neq D$
   a. $Z < -Z_{\frac{\alpha}{2}}\ or\ Z > Z_\alpha\ Which\ is\ the\ Same\ as\ |Z| > |Z_{\frac{\alpha}{2}}|$

Confidence Interval

A (1-α)% Confidence Interval for $(\pi_1 - \pi_2)$ is Given by:

$$(p_1 - p_2)\ \pm\ Z_{\frac{\alpha}{2}}\ \left(\sqrt{\left\{\frac{p_1\,(1-p_1)}{n_1} + \frac{p_2\,(1-p_2)}{n_2}\right\}}\ \right)$$

Comparing Population variances

In certain circumstances it can be necessary to compare population variances, it isn't as theoretical as it seems, e.g.

- Comparing the output of two Computers is equally steady and stable
  - Stability as opposed to raw computer power for instance
- Results in treatments with drugs
  - How consistently the drug offers results (variance) as opposed to the number of cures (mean)

Hypothesis

Typically variation is compared as a ratio (e.g. in ANOVA tables) thus it is the correct null hypotheses to employ.

$$H_0: \sigma_1 = \sigma_2, same\ as$$
$$H_0: \sigma_1 - \sigma_2 = 0, \text{ same as}$$
$$H_0: \frac{\sigma_1}{\sigma_2} = 1$$

1. $H_a: \frac{\sigma_1}{\sigma_2} > 1\ Meaning\ \sigma_1 > \sigma_2$
2. $H_a: \frac{\sigma_1}{\sigma_2} < 1\ Meaning\ \sigma_1 < \sigma_2$
3. $H_a: \frac{\sigma_1}{\sigma_2} \neq\ 1\ Meaning\ \sigma_1 \neq\ \sigma_2$

Test Statistic and Rejection Region

The Test statistic used is the F-statistic, it measures the ratio of Treatment Variance and Error Variance ($\frac{MST}{MSE}$)to against the probability of such an occurrence.

It is important to note that the F-statistic is F_{α,n_1-1,n_2-1} and the order of the degrees of freedom and the α value are important and are not interchangeable.

If the Null Hypothesis is	**1.** $H_a: \frac{\sigma_1}{\sigma_2} > 1$	**2.** $H_a: \frac{\sigma_1}{\sigma_2} < 1$	**3.** $H_a: \frac{\sigma_1}{\sigma_2} \neq\ 1$
The Test Statistic Will be	$F_{Test} = \frac{s_1^2}{s_2^2}$	$F_{Test} = \frac{s_2^2}{s_1^2}$	$F_{Test} = max\{\frac{s_1^2}{s_2^2}, \frac{s_2^2}{s_1^2}\}$
And the Null hypothesis will fail for:	$H_a: \frac{\sigma_1}{\sigma_2} \neq\ 1$	$F > F_{\alpha,n_2-1,n_1-1}$	$F > F_{\frac{\alpha}{2},n_1-1,n_2-1}$ or $F >\ F > F_{\frac{\alpha}{2},n_2-1,n_1}$

2A; Explaining the Chi-Distribution

Chi Square Procedures[8, 9, 10]

Chi-square statistical distributions extends the logic of binomial procedures to cover situations where there are more than two categories of possible outcome.

A *Chi-square* distribution can be used for more than one dimension of category, e.g. a person's political persuasion relative to their level of education, these are two dimensions of categories, each with many distinct sub-categories within e.g. conservative, liberal, right wing, communist faction and tertiary, secondary, trade level etc.

The Chi Square Distribution

The chi-square values are not generated via a mathematical equation, instead they are measured, much like measuring other constants of life like gravity and pi chi square values can also be measured.

Given a Computer generates an indefinite stream of the letters a, b and c, such that each number has a probability of being produced equal to one third and the stream of numbers is produced purely randomly with no order.

An extraction of the numbers produced might look like this

… abcbabccbbaabcbabcbbcabcacbbbaccbbaaaaccbabcbabcbacbacbacbacbacbaccbbaabcca …

[8] Richard Lowry, *Concepts & Applications of Inferential Statistics* (29 March 2012) VassarStats: Web Site for Statistical Computation Ch 7 <http://vassarstats.net/textbook/>
[9] *Biometry Formulas* (Bound Reference for Biometry, Spring 2013) p 12
[10] *Tables and Distributions* (Bound Reference for Biometry, Spring 2013) p 8

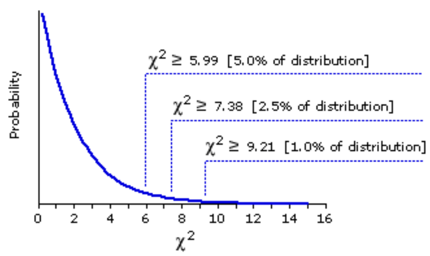
If a random sample of this data was taken (say n = 300) the distribution of this sample might look like this:

	Letter *'a'*	Letter *'b'*	Letter *'c'*	Totals
Observed frequency	96 (32%)	103 (34.33%)	101 (33.66%)	300
Expected frequency	100 (33.3%)	100 (33.3%)	100 (33.3%)	300
Difference between Observed Value and Expected Value (o-e)	-4	3	1	0
Squared Difference between observed Value and Expected Value $(o\text{-}e)^2$	16	9	1	26
Squared Difference relative to Expected Value $(o\text{-}e)^2/e$	0.16	0.09	0.01	0.26

The reason the Differences must be squared is so that they can be summed to give an overall value.

If a very large number of samples was taken, say 1 000, 000, 000, of sample sizes of say n=300 then a histogram was created comparing the *Squared Differences relative to the Expected Value =* $(o\text{-}e)^2/e$ for every sample, the histogram might look something like this.

It is important to remember that there can be no negative values due to the squared nature of the differences hence the one sided shape



That distribution of numbers is how the Chi-Distribution is created and why it is relevant to such circumstances relating to the probability of random distribution of classes.

This particular distribution is a chi distribution with 2 *degrees of freedom* (*df)* the distribution however does change relative to the *degrees of freedom (df)*.

2B; Chi-Squared Test

Goodness-of-Fit Tests for Statistical Distributions.

Contents

No table of contents entries found.

Goodness-of-fit Tests

It is always necessary to understand the behaviour of data, such as how it is distributed (e.g. Normal, Uniform, Poisson, Chi-square, F-stat, student's t curve etc.)

The goodness of fit test provides a tool for which the difference between how data behaves and how it is thought to behave can be understood.

As long as observations are independent, data can be tested against any type of distribution.

Rule of Five

The test statistic used to test data is only approximately chi-squared distributed, for the approximation to apply expected cell frequencies must be 5 or greater, otherwise rows and columns must be combined until the expected cell frequency is greater than 5.

Chi-Squared Test for Normality

Although the Chi-squared test is often used for discrete categorical data, e.g. 'counts' it is possible to summarise continuous data using intervals.

Most of the time continuous data is assumed to be normally distributed.

A Chi-Square test can be used to determine whether continuous data is normally distributed by separating the data into intervals and then using a statistical distribution to solve for an expected value.

Example of Chi-Squared Test for Normality

Is the Following Data Normally Distributed?

$\bar{x} = 80.86$ and $s = 12$

Group	**Observed Frequency**
<50	57
50-60	330
60-70	2132
70-80	4584
80-90	4604
90-100	2119
100-110	659
>110	251
Total	14736

In order to relate the data to a *Standard Normal Distribution* it will be necessary to find the sample mean and the sample variance of the data.

Descriptive Statistics from Calculator

These values can be found using a calculator by first enabling frequency with:

Then typing in Data and solving as standard.

In scenarios where Interval Classes are used a midpoint must be estimated, the quality of the estimate of the sample mean will be affected by how closely the sample data falls to the midpoints.

Hypothesis

H_0: The cells match a *Normal Distribution*

H_a: The cells do not match a *Normal Distribution*.Test statistic

Test Statistic

Expected Frequencies

Group	Observed Frequency	Probability	Expected Frequencies
<50	57	$P(X < 50) = P\left(Z < \frac{50-80)}{12}\right)$ $= P(Z < -2.55)$ $= 0.0054$	$n = 0.0054 \times 14\,736$ $= 79.57$
50-60	330	P(50<X<60) = P($\frac{50-80)}{12} < Z < \frac{60-80)}{12}$) $= P(-2.55 < Z < -1.72)$ =0.0427-0.0054 =0.0373	$n = 0.0373 \times 14\,736$ $= 550$
"	"	"	"
"	"	"	"
"	"	"	"
"	"	"	"

Thus solving for all normal probabilities and expected frequencies:

Group	Observed Frequency	Upper Z Limit	Normal Probability	Expected Frequency
<50	57	-2.56	0.0052	76.6272
50-60	330	-1.72	0.0384	565.8624
60-70	2132	-0.89	0.1431	2108.7216
70-80	4584	-0.06	0.2894	4264.5984
80-90	4604	0.78	0.3062	4512.1632
90-100	2119	1.61	0.164	2416.704
100-110	659	2.44	0.0464	683.7504
>110	251	2.44 (Minum)	0.0073	107.5728
Total	14736		1	14736

$$X^2 = \sum \frac{(e-0)^2}{e} = \frac{(76.6-52)^2}{76.6} + \frac{(565.9-330)^2}{565.9} \ldots + \frac{(251-107.57)^2}{107.57}$$

Expanding the Table for a solution

Group	Observed Frequency (O)	Expected Frequency (e)	Difference (o-e)	Squared Difference (o-e)^2	Standard Squared Difference (0-e)^2/e
<50	57	76.63	19.63	385.23	5.0272877
50-60	330	565.86	235.86	55631.07	98.312013
60-70	2132	2108.72	-23.28	541.88	0.2569727
70-80	4584	4264.60	-319.40	102017.38	23.92192
80-90	4604	4512.16	-91.84	8434.00	1.8691695
90-100	2119	2416.70	297.70	88627.67	36.672953
100-110	659	683.75	24.75	612.58	0.8959151
>110	251	107.57	-143.43	20571.36	191.232
Total	***14736.00***	***14736.00***	***0.00***	***276821.18***	***358.19***

Thus $X^2 = 358.19$

Rejection Region

Degrees of Freedom is = 8 -1 -2 = 5 (the extra 2 being an extra 2 parameters from the normal distribution)

H_0 Fails for $X^2 > X^2_{0.01,5}$

$358.19 > 15.086$

Thus H_0 Fails

Conclusion

At a significance level of 99% the data can be assumed to NOT be *Normally Distributed*.

Example of Chi-Squared Test for Poisson distribution

Is the following data distributed by way of a *Posson Distribution*?

Number of Plants	0	1	2	3	4	5	6
Observeed Frequency	9	9	10	14	2	2	2

Data

Directly from a calculator:

$\bar{x} = 2.10$

$s = 1.57$

$n = 48$

Hypothesis

H_0: The data follows a *Poisson Distribution*

H_a: The data DOES NOT follow a *Poisson Distribution*

Rejection Region

Degrees of Freedom = $k - 1 - 1 = 7 - 1 - 1 = 5$

The extra -1 comes from the one parameter of the *Poisson Distribution*

H_0 fails for $X^2 > X^2_{d.f,\alpha}$

$X^2 > 11.07$

$$p = \frac{e^{-\mu} \times \mu^k}{k!}$$

Test Statistic

Expected Frequencies

Number of Plants (k)	0	1	2	3	4	5	6	TOTAL *(n)*
Observed Frequency (f)	9	9	10	14	2	2	2	48
Poisson Probablity $p = \frac{e^{-\mu} \times \mu^k}{k!}$	0.12	0.26	0.27	0.19	0.1	0.04	0.01	1.0
Expected Frequency $(p \times n)$	5.9	12.3	13.0	9.1	4.8	2.0	0.7	47.71862

Solving for X^2

$$X^2 = \sum\frac{(e-o)^2}{e} = \frac{(5.9-9)^2}{5.9} + \frac{(12.3-9)^2}{12.3} \dots \frac{(0.7-2)^2}{0.7}$$

No. of Plants	*Expected Frequency*	*Observed Frequency*	*Squared Difference*	*Standard Square Difference*
0	5.9	9	9.61	1.62881356
1	12.3	9	10.89	0.88536585
2	13	10	9	0.69230769
3	9.1	14	24.01	2.63846154
4	4.8	2	7.84	1.63333333
5	2	2	0	0
6	0.7	2	1.69	2.41428571
SUM	47.8	48	63.04	9.89256769

And thus $X^2 = 9.89$

H_0 is not rejected because $X^2 > X^{2}_{4,0.05}$, 9.89 > 11.07

Conclusion

At a significance level of 95% there is not enough evidence to assume the data is not distrusted in accordance with a *Poisson Distribution*.

That is to say that the growth of the seedlings in pots (after one week from planting) follows a *Poisson Distribution*.

2B(a); Finding Descriptive Statistics from a Frequency Table

Example of Chi-Squared Test for Normality

Is the Following Data Normally Distributed?

$\bar{x} = 80.86$ and $s = 12$

Group	**Observed Frequency**
<50	57
50-60	330
60-70	2132
70-80	4584
80-90	4604
90-100	2119
100-110	659
>110	251
Total	14736

In order to relate the data to a *Standard Normal Distribution* it will be necessary to find the sample mean and the sample variance of the data.

Sample Mean

In using class intervals the sample mean is only an estimate, the accuracy of the estimate depends on how close the data is to the various midpoints.

Where the Class or Group Intervals are open ended the midpoint must be decided on some arbitrary rational basis.[11] $\bar{x}$

Group	Observed Frequency (f)	Midpoint(x)	Total amount of Midpoint (xf)
<50	57	40	2280.00
50-60	330	55	18150.00
60-70	2132	65	138580.00
70-80	4584	75	343800.00
80-90	4604	85	391340.00
90-100	2119	95	201305.00
100-110	659	110	72490.00
>110	251	120	30120.00
Total	14736		1198065.00

$$\bar{x} = \frac{1198065}{14736} = 81$$

Sample Standard Deviation

The sample standard deviation must be calculated as the squared difference from data point to mean, multiplied by its frequency then averaged.

Group	Observed Frequency (f)	Midpoint(x)	Total amount of Midpoint (xf)	Difference $(\bar{x} - x)$	Squared Difference $(\bar{x} - x)^2$	Total amount of Squared Difference $f(\bar{x} - x)^2$
<50	57	40	2280.00	-41.3	1705.69	97224.33
50-60	330	55	18150.00	-26.3	691.69	228257.7
60-70	2132	65	138580.00	-16.3	265.69	566451.08
70-80	4584	75	343800.00	-6.3	39.69	181938.96
80-90	4604	85	391340.00	3.7	13.69	63028.76
90-100	2119	95	201305.00	13.7	187.69	397715.11
100-110	659	110	72490.00	28.7	823.69	542811.71
>110	251	120	30120.00	38.7	1497.69	375920.19

[11] Transtutors, <http://www.transtutors.com/homework-help/statistics/central-tendency/open-end-class-intervals-series.aspx>

Total	14736		1198065.00	-5.40	5225.52	2453347.84

Thus the Total amount of Squared Deviation in the data is: 2, 453, 347.84

The total number of data points is 14, 736

Finally the average amount of squared deviation is $s^2 = \frac{2453347}{14736} = 166$

$$s = \sqrt{166} = 12.902$$

Once again differences between the lecture notes and the standard deviation found is related to the somewhat arbitrary midpoints chosen for the interval classes.

Calculator

These values can be found using a calculator by first enabling frequency with:

Then typing in Data and solving as standard.

3A; Analysis of Variance (One-Way)

Topic 3A | Analysis of Variance: One-Way | Week 5 Material

Contents

Analysis of Variance

When more than two independent population means are compared a *t-distribution* will no longer cut it, instead a technique known as *Analysis of Variance* must be used.

This compares different causes of variation between groups and provides a test statistic to determine whether such differentiation is significant.

One Way Analysis of Variance

This is a comparison of the variance caused by random error and the variance caused by differing treatments being tested in order to yield some result.

In essence there is only on block of treatments however a two-way or greater might have many blocks of treatments, that is to say, a comparison of the results yielded by different treatments under a variety of circumstances.

For an ANOVA table to be used observations between treatments must:

1. Be Normally distributed
2. Have Equal variance

ANOVA Table

One-Way ANOVA Table

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

Where:

- x_{ij} is the i^{th} response of the j^{th} treatment
- $\bar{x}_j$ is the mean of the j^{th} treatment
- $\bar{x}_{Grand}$ is the mean value of every single data point, which is equal to the mean of each group mean
- n_j is the no. of data points in the j^{th} treatment
- *n*=n_{total} is the total no. of data points
- *k* is the number of treatments
- *i* represents data point and *j* represents treatment

- SST_{Total} is the total variation in all samples, SS = SSE + SST
- $SSA_{Treatment}$ is the total variance present within all of the groups or treatmens

- SSE_{Error} is the variance between the groups
- Also the pooled sample variance (S_p) = MSE = $\frac{\text{SSE}}{n-k}$

Sum of Square Deviants

$SSA_{Treatment}$ – Variance between groups

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

$$\text{SSA}_{\text{Treatment}} = \sum_{j=1}^{k}[n_j(\bar{x}_j - \bar{x}_{Grand})^2]$$

However the definition or explanatory equation is almost useless for doing calculations thus by way of some algebra a computational equation can be derived like so:

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

$$\sum_{j=1}^{k}\left[\frac{(\sum_{i=1}^{n}[x_{ij}])\text{^}2}{n_j}\right] - \left[\frac{\{\sum_{j=1}^{k}[\ \sum_{i=1}^{n_j}(x_{ij})\,]\}\text{^}2}{n_{total}}\right]$$

SSE_{Error} – Variance within groups

SSE_{Error} is the total amount of variance in data within each group or treatment but not the variance in between data from different groups.:

$$SSE_{Error} = \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2\right] = \sum_{j=1}^{k}[s_j^2(n_j - 1)]$$

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

$$SSE_{Error} = \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j}[x_{ij}^2]\right] - \sum_{j=1}^{k}\left[\frac{\{\sum_{i=1}^{n_j}(x_{ij})\}\text{^}2}{n_j}\right]$$

SST_{Total} – Total Variance

SST_{Total} is the total squared deviation of the data, that is the sum of the squared difference from data point to grand mean.

$$SST_{Total} = \sum_{ij=1,1}^{n_{total}}\left[(x_{ij} - \bar{\bar{x}})^2\right] = s_{Grand}^2\,(n_{total} - 1)$$

The total Variance can also be found by summing the total variation from data point to group mean respectively thus:

$$SST_{Total} = \sum_{ij=1,1}^{n_{total}}\left[(x_{ij} - \bar{\bar{x}})^2\right] = \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j}(x_{ij} - \bar{x})\right]$$

And using algebra a computational sum that can be used to find the value (with more ease) can be derived such that:

$$SST_{Total} = \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j}[x_{ij}^2]\right] - \left[\frac{\{\sum_{j=1}^{k}[\ \sum_{i=1}^{n_j}(x_{ij})\,]\}\text{^}2}{n_{total}}\right]$$

However all the variation can also be defined like so:

$$\mathrm{SST_{Total}} = \mathrm{SSE_{Error}} + \mathrm{SST_{Total}}$$
$$\mathrm{SST_{Total}} = \sum_{i=1}^{k}[s_i^2(n_i - 1)^2] + \sum_{i=1}^{k}[\bar{x}_i(\bar{x}_i - \bar{x}_{Grand})^2]$$

Hypotheses Testing

Hypotheses

H_0: $\mu_1 = \mu_2 = \mu_{3\ldots} = \mu_k$

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*~~
>
> Work your way down the ANOVA table whilst writing the F-Stat subtext and in the table the *df* numbers left to write as 1 and 2 respectively

P-Value

Find α such that $F = F_{k-1,n-k,\alpha}$

Conclusion

Are the population means equal or different?

Example

The tensile strength resulting from mixing method used when creating cement is being analysed by builders.

Four different techniques are being compared such that.

	Mixing Technique			
	1	2	3	4
Tensile Strength	3129	3200	2800	2600
	3000	3300	2900	2700
	2865	2975	2985	2600
	2890	3150	3050	2765
Mean =	2971	3156.25	2933.75	2666.25

One Way ANOVA Table

Descriptive Statistics

$$\bar{\bar{x}}_{Grand} = Average\ of\ Means = \frac{2971 + 3156.25 + 2933.75 + 2666.25}{4} = 2931.813$$

$$s_1 = \frac{(3129 - 3000)^2 + (3000 - 2971)^2 \ldots}{4} = 14\,534$$

	Mix	***Stir***	***Shake***	***Pour***
1	3129	3200	2800	2600
2	3000	3300	2900	2700
3	2865	2975	2985	2600
4	2890	3150	3050	2765
Average	2971	3156.25	2933.75	2666.25
Variance	14534	18489.58	11722.92	6556.25

Granc Variance	Grand Average
42909.89583	2931.8125

Sum of Square Deviants

$SSA_{Treatment}$ (489 740.2)

The sum of square differences between treatments is the squared difference between group mean and grand mean for every data point:

$$SSA_{Treatment} = \sum_{j=1}^{k}[n_j(\bar{x}_j - \bar{x}_{Grand})^2] = \sum_{j=1}^{k}\left[\frac{(\sum_{i=1}^{n}[x_{ij}])^{\wedge}2}{n_j}\right] - \left[\frac{\{\sum_{j=1}^{k}[\sum_{i=1}^{n_j}(x_{ij})]\}^{\wedge}2}{n_{total}}\right]$$

$$SSA_{Treatment} = \sum_{j=1}^{k}\left[n_j(\bar{x}_j - \bar{x}_{Grand})^2\right]$$

$$= 4(2971 - 2931.813)^2 + 4(3156.25 - 2931.813)^2 + 4(2933.75 - 2931.813)^2 + 4(2666.25 - 2931.813)^2$$

$$SSA_{Treatment} = 489\,740.2$$

SSE_{Error}

The sum of square differences caused by random error is the variance from every data point to the group mean and then summed with the other groups thus:

$$SSE_{Error} = \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2\right] = \sum_{j=1}^{k}[s_j^2(n-1)]$$

$$SSE_{Error} = s_1^2(4-1) + S_2^2(4-1)\ldots$$

$$SSE_{Error} = 43\,602(4-1) + 55\,468.75(4-1) + 35\,168\,(4-1) + 19\,668.75(4-1) = 153\,908$$

$$SSE_{Error} = 153\,908.25$$

SST_{Total}(643 648.44)

This is the total squared difference from data point to mean for the data.

$$SST_{Total} = \sum_{ij=1,1}^{n_{total}}\left[(x_{ij} - \bar{\bar{x}})^2\right]$$

$$= s_{Grand}^2\,(n_{total} - 1)$$

$$= 42\,909.9(16-1)$$

$$= 643\,648.438$$

ANOVA Table

Source of Variation	***Sum of Squares***	***Degrees of Freedom***	***Mean Squares***	***F-Stat***
Between Treatments	489 740	*4-1=3*	$MST = \frac{SST}{k-1} = \frac{489\,740}{4-1} = 160580$	$\frac{MST}{MSE} = \frac{160\,580}{12\,825} = 12.52$
Within Treatments (Error)	153 908	*16-4=12*	$MSE = \frac{SSE}{n-k}$	/

			$= 12\ 825$	
Total	643 648	*16-1=15*	/	/

Hypothesis Test

Hypotheses

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a:: at least one μ is different

Test Statistic

$F = \frac{\text{MST}}{\text{MSE}} = 12.52$

Rejection Region

H_0 rejected at a significance level of α if:

$$F > F_{k-1,\ n-1,\ \alpha}$$
$$F > F_{3,12,0.05}$$
$$F > 3.49$$
$$12.52 > 3.49$$

Thus the Null Hypothesis is rejected

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*~~, to get the order correct enter the *d.f.* top to bottom from the ANOVA, in the distribution the *d.f.* are numbered as 1 and 2 left to right respectively.

P-Value

Find α such that $F = F_{k-1,\ n-1,\ \alpha}$

$$F = F_{3,12,\alpha}$$

$$12.52 = F_{3,12,\alpha}$$

$$\alpha < 0.005$$

Thus at a probability of incorrectly rejecting equality between means (H_0) of less than 0.5% the population treatment means can be found to be different

Conclusion

At a significance level of over 99.5% (p-value) the method used for mixing concrete does create some difference in tensile strength.

:

3B; Analysis of Variance (Two-Way)

Topic 3B | Analysis of Variance: Two-Way | Week 6 Material

Contents

Two-Way ANOVA

Block Design

One-way Anova deals with only one source of explained variation, treatment to treatment variation, whereas Two-Way Anova deals with different sources of variation, such as the distribution of fish in different streams at different depths for each respective stream.

For a two-way ANOVA the total variation is partitioned in to three sources:

1. Between Treatments (A)
    a. E.g. Fish numbers between streams
2. Between Blocks (B)
    a. E.g. Fish numbers at different depths of the respective streams
3. Within Treatments, i.e. random error (E)
    a. The variation of the data taken at said streams at said depth

Sum of Square Deviants

$SSA_{Factor\ A\ or\ Treatment}$

This is the variation of the data between all the treatments of data (e.g. the streams)

$$SSA = \sum_{i=1}^{r}[n_i(\overline{x_\imath} - \bar{\bar{x}})^2]$$

$SSB_{Factor\ B\ or\ Block}$

This is the variation of the data between the blocks of said treatments (e.g. the depths)

$$SSB = \sum_{j=1}^{r}\left[n_j\left(\overline{x_\jmath} - \bar{\bar{x}}\right)^2\right]$$

SSE_{Error}

The variation caused by error is the amount of variation that occurs by way of random errors within all the treatments throughout the various blocks and throughout the various treatments.

The error must be calculated by way of finding the remaining variation of the data:

$$SSE = SST - SSA - SSB$$

SST_{Total}

The total variance is the total squared deviation from data point to mean, thus:

$$\begin{aligned} SST &= \sum_{j=1}^{r}\left[\sum_{i=1}^{c}\left[\left(x_{ij} - \bar{\bar{x}}\right)^2\right]\right] \\ &= s_{grand}^2 \times (n_{total} - 1) \end{aligned}$$

Also be definition the total variance is the sum of all sources of variation:

$$SST = SSA + SSB + SSE$$

Where

i = The i^{th} response of a treatment or block

n_i = the number of responses in the i^{th} treatment

n_j = the number of responses in the j^{th} block

$r = j$= The number of treatments (Factor A)

$c = k$ = The number of Blocks (Factor B) {in the j^{th} treatment}?

n = The total number of data points (i.e. responses)

s_p = is the pooled variance $s_p = MSE = \frac{SSE}{n-r}$

ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F-Statistic
Between Treatments ***(Include name)***	$SSA_{Treatments}$	$df_A = c - 1$ $= j - 1$	$\mathrm{MSA} = \frac{\mathrm{SSA}}{\mathrm{r} - 1}$	$\frac{\mathrm{MSA}}{\mathrm{MSE}}$
Between Blocks ***(Include name)***	SSB_{Blocks}	$df_B = r - 1$ $= k - 1$	$\mathrm{MSB} = \frac{\mathrm{SSB}}{c - 1}$	$\frac{\mathrm{MSB}}{\mathrm{MSE}}$
Within Treatments (Error)	SSE_{Error}	$df_E = (r-1)(c-1)$	$MSE = \frac{SSE}{(k-1)(b-1)}$	
Total	SST_{Total}	n-1 = rc - 1		

Hypothesis Testing Treatment Means

Testing for the treatment, that is Factor A or the columns

Hypothesis

$H_0: A_1 = A_2 = A_3 \dots = A_j$ (That is to say that all of the treatment means, the row or Factor A, are the same in the population)

H_A: At least two of the means are different.

Test Statistic

The test statistic should be in line with the source of variation that is being tested

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

Rejection Region

H_0 fails for:

$$F > F_{c-1,(r-1)(c-1),\alpha}$$

$$F > F_{df_A, df_E, \alpha}$$

(Take the degrees of freedom from the corresponding row and from the error)

Testing for the blocks, that is Factor B or the rows

$H_0: B_1 = B_2 = B_3 \dots = B_k$ (Population means for the blocks are equal, that is the columns or Factor B have no significant difference)

H_A: At least two of the means are different.

Test Statistic

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

Rejection Region

The null hypothesis fails for:

$$F > F_{r-1,(r-1)(c-1),\alpha}$$

$$F > F_{df_B, df_E, \alpha}$$

(Take the degrees of freedom from the corresponding row and from the error)

Tukey's Test for Multiple Comparisons (One-Way ANOVA Only)

Tukey's test is essentially a t-test except that it corrects for experiment-wise error rate, as multiple comparisons are bing made the probability of making a type 1 error increases – *Tukey's* test corrects for that

Tukey's Test assumes that the data under scrutiny is:

1. The observations are being tested are independent of one another
2. The means are from *Normally Distributed* Populations
3. There is equal variation across observations

Hypotheses

The hypotheses to compare some group j with some other group k are:

$$H_0 : \mu_j = \mu_k$$

$$H_a : \mu_j \neq \mu_k$$

Test Statistic

$$T_{calc} = \frac{\left|\bar{x}_j - \bar{x}_k\right|}{\sqrt{MSE\left(\frac{1}{n_j} + \frac{1}{n_k}\right)}}$$

Rejection Region

The null hypothesis fails for:

$$T_{calc} > T_{c,\ n-j}$$

Where:

- $j = c =$ the number of treatments
- $n =$ the number of responses
- $T_{c,n-c}$ is the critical value of the *Tukey* Test Statistic T_{calc} for the desired level of significance. This can be found at Table 11.4 of Page 452 of the prescribed Test.

Statistical Distribution Table

A 95 % Statistical Distribution of H Values is given:

	Number of Groups (c)								
$n - c$	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
5	2.57	3.26	3.69	4.01	4.27	4.48	4.66	4.81	4.95
6	2.45	3.07	3.46	3.75	3.98	4.17	4.33	4.47	4.59
7	2.37	2.95	3.31	3.58	3.79	3.96	4.11	4.24	4.36
8	2.31	2.86	3.20	3.46	3.66	3.82	3.96	4.08	4.19
9	2.26	2.79	3.12	3.36	3.55	3.71	3.84	3.96	4.06
10	2.23	2.74	3.06	3.29	3.47	3.62	3.75	3.86	3.96
15	2.13	2.60	2.88	3.09	3.25	3.38	3.49	3.59	3.68
20	2.09	2.53	2.80	2.99	3.14	3.27	3.37	3.46	3.54
30	2.04	2.47	2.72	2.90	3.04	3.16	3.25	3.34	3.41
40	2.02	2.43	2.68	2.86	2.99	3.10	3.20	3.28	3.35
60	2.00	2.40	2.64	2.81	2.94	3.05	3.14	3.22	3.29
120	1.98	2.37	2.61	2.77	2.90	3.00	3.09	3.16	3.22
∞	1.96	2.34	2.57	2.73	2.85	2.95	3.03	3.10	3.16

Test for Homogeneity of Variances

ANOVA methods assume that observations between treatments are:

1. Normally Distributed
2. Have Equal Variance

Thus it is necessary to test the assumption of homogenous variance, this test assumes equal group sizes.

Hypothesis

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_c^2$$

$$H_a: Not\ all\ variance\ is\ equal$$

Test Statistic

The test statistic is the ratio of the largest sample variance to the smallest sample variance:

$$H_{calc} = \frac{s_{max}^2}{s_{min}^2}$$

Rejection Region

The null hypothesis is rejected if:

$$H_{calc} > H_{df1,df2}$$

Where:

- The Critical values of $H_{Critical}$ can be found in Table 11.5 of page 453 of the prescribed text
- c is the number of treatments
- n is the number of observations
- Degrees of Freedom can be calculated like so:
  - Numerator = c = df_1
  - Denominator = $\frac{n}{c} - 1$ = df_2

Statistical Distribution

A 95 % Statistical Distribution of H Values is given:

	Numerator df_1								
Denominator df_2	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	39.0	87.5	142	202	266	333	403	475	550
3	15.4	27.8	39.2	50.7	62.0	72.9	83.5	93.9	104
4	9.60	15.5	20.6	25.2	29.5	33.6	37.5	41.1	44.6
5	7.15	10.8	13.7	16.3	18.7	20.8	22.9	24.7	26.5
6	5.82	8.38	10.4	12.1	13.7	15.0	16.3	17.5	18.6
7	4.99	6.94	8.44	9.7	10.8	11.8	12.7	13.5	14.3
8	4.43	6.00	7.18	8.12	9.03	9.78	10.5	11.1	11.7
9	4.03	5.34	6.31	7.11	7.80	8.41	8.95	9.45	9.91
10	3.72	4.85	5.67	6.34	6.92	7.42	7.87	8.28	8.66
12	3.28	4.16	4.79	5.30	5.72	6.09	6.42	6.72	7.00
15	2.86	3.54	4.01	4.37	4.68	4.95	5.19	5.40	5.59
20	2.46	2.95	3.29	3.54	3.76	3.94	4.10	4.24	4.37
30	2.07	2.40	2.61	2.78	2.91	3.02	3.12	3.21	3.29
60	1.67	1.85	1.96	2.04	2.11	2.17	2.22	2.26	2.30
∞	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Model of Simple Linear Regression

Linear Regression is about determining the best model that describes the association between y and x, this is done by finding an equation that has the least amount of error between all the points of recorded data.

If we have a sample of n pairs of observations from a population such that (x_i, y_i) represents the I'th pair a linear model can be created like so:

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \epsilon \qquad => \qquad E[y_i] = \hat{y} = \beta_0 + \beta_1 x_i$$

Intercept

Slope

Where ϵ is the error between the recorded y_i value and the expected value$\hat{y}$. (i.e. $\epsilon = y_i - \hat{y}_i$)

Estimation of Slope and Intercept Values

The slope and coefficient can be estimated by way of using the method of least squares

Estimation of Slope	*Estimation of Intercept*
$\beta_1 = \frac{SS_{xy}}{SS_x} = \frac{\sum_{i-1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}[(x_i - \bar{x})^2]}$	$\beta_0 = \bar{y} - \beta_1 \bar{x}$

Least Squares Calculation

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

$$SS_x = \sum_{i=1}^{n}[(x_i - \bar{x})^2] = \sum_{i=1}^{n}[x_i^2] - n(\bar{x})^2$$

$$\beta_1 = \frac{SS_{xy}}{SS_x} \qquad \beta_0 = \hat{y} - \beta_1 \bar{x}$$

$$SS_y = \sum_{i=1}^{n}[(y_i - \bar{y})^2] = \sum_{i=1}^{n}[y_i^2] - n(y)^2$$

Examining the Regression equation

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \epsilon$$

The error ϵ is a random variable with a mean value of 0 and an unknown variance σ^2

$$E(\epsilon_i) = 0 \text{ and } \mathrm{Var}(\epsilon) = \sigma^2 - \textit{homoscedastic error}$$

Given that that ϵ_i is random with mean zero, it is also true that ϵ is normally distributed such that:

$$\epsilon_i \sim N(0, \sigma^2)$$

This is an important assumption when performing tests

ϵ_i and ϵ_j are uncorrelated $i \neq j$

$\mathrm{Cov}(\epsilon_i, \epsilon_j)$

Hypothesis Test for the Regression Model

$$SST_{Total} = SS(Reg)_{Regression} + SSE_{Error}$$

$$\sum_{i=1}^{n} [(y_i - \bar{y})^2] = \sum_{i=1}^{n} [(\hat{y}_i - \bar{y})^2] + \sum_{i=1}^{n} [(y_i - \hat{y}_i)^2]$$

Total variation of y values

Variation between predicted value and mean value

Variation in the error

Notes on Sum of Squares

$$SSE = SS(Reg)_{Regression} + SSE$$

$$SSE = \sum_{i=1}^{n}[(y_i - \hat{y}_i)^2] = \sum_{i=1}^{n}[(\epsilon_i)^2]$$

$$SS(Reg) = \sum_{i=1}^{n}[(\hat{y}_i - \bar{y})^2] = \beta_{1-Slope} \times SS_{xy} = \beta_{1-slope}^2 \times SS_x$$

$$SS(Reg) = \sum_{i=1}^{n}[(\hat{y}_i - \bar{y})^2] = \beta_1 SS_{xy} = \beta_1^2 SS_x$$

A regression model that is poor at predicting the y value will lead to $SS(Reg) = 0$. This will also occur when $\beta_1 = 0$, thus a hypothesis test can test whether the slope (β_1) is zero, which will test the appropriateness of the regression model.

Degrees of Freedom

$$d.f. = (n - 2)$$

Variance of the Regression Line

$$S^2 = \frac{\sum_{i=1}^{n}[(y_i - \hat{y}_i)^2]}{n-2} = \frac{\sum_{i=1}^{n}[(\epsilon_i)^2]}{n-2} = MSE$$

Proof

$$\sum_{i=1}^{n}[(y_i - \bar{y})^2] = \sum_{i=1}^{n}[(\hat{y}_i - \bar{y})^2] + \sum_{i=1}^{n}[(y_i - \hat{y}_i)^2]$$

$$\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n}\{(y_i - \bar{y}) - (\hat{y}_i - \bar{y})\}^2$$

$$= \sum_{i=1}^{n}\left\{(y_i - \bar{y})^2 - 2(y_i - \bar{y})(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\right\}$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 - 2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares (SS)	Mean Sum of Squares (MSE)	F-Stat
Regression	*1*	*SS(reg)*	$MSR = \frac{SS(Reg)}{1}$	$F = \frac{MSR}{MSE}$
Error	*n-2*	*SSE*	$MSE = \frac{SSE}{n-2} = s^2$	\
Total	*n-1*	*SST*	\	\

Hypothesis Test

Hypothesis

H_0: The linear model does not model the two factors

H_a: The linear model is useable to describe the relationship between the two values

Test Statistic

$F_{Calc} = \frac{MSR}{MSE}$

Rejection Region

The null hypothesis (H_0) is rejected for:

$$F_{Calc} > F_{\alpha,df1,df2}$$

Where the Degrees of Freedom are:

1. Regression, which is 1
2. Error, Which is n-2

Conclusion

At a 95% Confidence Interval…

Example 7.1 – Linear Regression Model & Hypothesis Test

A company wishes to create a model to describe the correlation between work hours and the lot size of manufactured parts:

Example 7.1			
	Toluca Company		
Run	*Lot Size (x)*	*Work Hours (y)*	*Lot Size Work Hours (x × y)*
1	80	399	31920
2	30	121	3630
3	50	221	11050
4	90	376	33840
5	70	361	25270
6	60	224	13440
7	120	546	65520
8	80	352	28160
9	100	353	35300
10	50	157	7850
11	40	160	6400
12	70	252	17640
13	90	389	35010
14	20	113	2260
15	110	435	47850
16	100	420	42000
17	30	212	6360
18	50	268	13400
19	90	377	33930
20	110	421	46310
21	30	273	8190
22	90	468	42120
23	40	244	9760
24	80	342	27360
25	70	323	22610
Sum	1750	7807	617180
Mean Values	70	312.28	N/A
Sum of Squared Values	142300	2745173	N/A

Linear Regression Model

$$n = 25$$

$$\bar{x} = 70 \quad \bar{y} = 312.28$$

$$\begin{aligned} ss_x &= \sum_{i=1}^{n}[x_i^2] - n(\bar{x})^2 \\ &= 142,300 - 25\,(70)^2 \\ &= \mathbf{19,800} \end{aligned}$$

$$\begin{aligned} SS_y &= \sum_{i=1}^{n}[y_i^2] - n(y)^2 \\ &= 2,745,173 - 25\,(312.28)^2 \\ &= \mathbf{307,203} \end{aligned}$$

$$\begin{aligned} SS_{xy} &= \sum_{i=1}^{n}[x_i y_i] - n\bar{x}\bar{y} \\ &= 617,180 - 25 \times 70 \times 312.28 \\ &= \mathbf{706,90} \end{aligned}$$

$$\beta_{1-Slope} = \frac{SS_{xy}}{SS_x} = \frac{706,90}{19,800} = 3.57$$

$$\begin{aligned} \beta_{0-Intercept} &= \bar{y} - b_{1-Slope} \times \bar{x} \\ &= 312.28 - 3.57 \times 70 \\ &= 62.37 \end{aligned}$$

Thus a linear model would be:

$$\hat{y}_i = b_0 + b_1 x_i$$

$$\hat{y}_i = 62.37 + 3.57 x_i$$

Construction of ANOVA Table

$$
\begin{aligned}
SS(Reg) &= \beta_1\, SS_{xy} \\
&= 3.5702 \times 70690 \\
&= 252,378
\end{aligned}
$$

$$SST_{Total} = SS_y = 307,203$$

$$
\begin{aligned}
SSE_{Error} &= SST_{Total} - SS(Reg) \\
&= 307,203 - 252,378 \\
&= 54,825
\end{aligned}
$$

Source of Variation	Degrees of Freedom	Sum of Squares (SS)	Mean Sum of Squares (MSE)	F-Stat
Regression	*1*	*SS(reg)=252, 378*	$MSR = 252,378$	$F = \frac{252,378}{2,383.7}$ $= \mathbf{105.9}$
Error	*25-2=23*	*SSE=54, 825*	$MSE = \frac{54,825}{23}$ $= \mathbf{2,383.7}$ $= s^2$	\
Total	*25-1=24*	*SST=307, 203*	\	\

Hypothesis

H_0: The regression model does not describe the linear relationship between lot size and hours worked

H_a: The regression model provides a useable model

Test Statistic

$$F = 105.9$$

Rejection Region

The Critical value, at the 0.05 Significance level is:

$F_{0.05}(1, 23) = 4.28$

$$F_{Calc} = 105.9 > F_{0.05} = 4.28$$

Thus the Null hypothesis is rejected

Conclusion

At a 95% significance level the regression model describes the relationship between lot size and hours worked.

Hypothesis Testing for the Slope and Intercept Values (Parameters)

Assuming the model is correct, if σ^2 is unknown, we may use s^2 in place of σ^2.

The significance of the relationship between x and y can also be tested by:

$$t_{Calc} = r\sqrt{\frac{n-2}{1-r^2}}$$

<table>
<tr>
<td>

Hypothesis Test for the Intercept Value

Hypothesis

$$H_0: \beta_{0-Intercept} = 0$$
$$H_a: \beta_{0-Intercept} \neq 0$$

Test Statistic

$$t = \frac{b_0 - \beta_0}{s.e(b_0)} = \frac{b_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}}$$

$$d.f. = (n-2)$$

Variance of the Intercept Value

$$Var(b_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right]$$

Rejection Region

The null hypothesis fails if:

$$|t_{calc}| > t_{\alpha,d.f.}$$

Conclusion

At whatever significance level the value intercept value appears to be a non-zero value (or whatever).

Confidence Interval of the Intercept Value

$$\beta_0 = b_0 \pm t_{\frac{\alpha}{2}}(n-2) \times s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}$$

$$std.Error\ (b_0) = S.E.(b_0) = s \times \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}$$

$$\beta_0 = b_0 \pm t_{\frac{\alpha}{2}} \times (n-2) \times\ S.E.(b_0)$$

</td>
<td>

Hypothesis Test for the Slope Value

Hypothesis

$$H_0: \beta_{0-Intercept} = 0$$
$$H_a: \beta_{0-Intercept} \neq 0$$

Test Statistic

$$t = \frac{b_1 - \beta_1}{s.e(b_1)} = \frac{b_1 - \beta_1}{\frac{s}{\sqrt{SS_x}}}$$

$$d.f. = (n-2)$$

Variance of the Slope Value

$$Var(b_1) = \frac{\sigma^2}{SS_x}$$

Rejection Region

The null hypothesis fails if:

$$|t_{calc}| > t_{\alpha,d.f.}$$

Conclusion

At whatever significance level the value intercept value appears to be a non-zero value (or whatever).

Confidence Interval of the Slope Value

$$\beta_1 = b_1 \pm t_{\frac{\alpha}{2}}(n-2)\frac{s}{\sqrt{SS_x}}$$

$$std.Error\ (b_1) = S.E.(b_1) = \frac{s}{\sqrt{SS_x}}$$

$$\beta_1 = b_1 \pm t_{\frac{\alpha}{2}} \times (n-2) \times\ S.E.(b_1)$$

</td>
</tr>
</table>

Correlation between Variables

Sample Correlation Coefficient

A regression analysis implies that there exists some linear relationship between the variables x and y.

A way to measure this is by using the value:

$$r = \frac{SS_{xy}}{\sqrt{SS_x \times SS_y}}$$

$$b_1 = \frac{SS_{xy}}{SS_x} = r \times \sqrt{\frac{SS_y}{SS_x}}$$

This measures the linear association between the two variables for $|r| < 1$.

If the value of r is negative 1 there is a perfect negative linear relationship between the variables and likewise for positive values of r.

If r is 0 there is no LINEAR relationship between the variables, they are not correlated. However they may be related in some other way (e.g. quadratically, cubicly, logarithmically, exponentially etc.)

Coefficient of Determination

The squared r value describes the proportion of variation in the true y values that are explained by the regression line:

$$r^2 = R^2 = \frac{\sum_{i=1}^{n}[(\widehat{y}_i - \bar{y})^\wedge 2\,]}{\sum_{i=1}^{n}[(y_i - \bar{y})^2]} = \frac{SS(Reg)}{SST_{Total}} = 1 - \frac{SSE_{Error}}{SST_{Total}}$$

Adjusted Coefficient of Determination

The value of R^2 can be inflated by additional predictor variables thus the value of rR^2 can be adjusted relative to the number of parameters in the regression model.

$$R^2_{Adjusted} = 1 - \left(\frac{n-1}{n-k-1}\right) \times \frac{SSE}{SST}$$

Proof

$$\begin{aligned} R^2_{adj} &= \frac{\mathrm{SS(Reg)}/df_{reg}}{SST/df_T} \\ &= 1 - \frac{SSE/df_E}{SST/df_T} \\ &= 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} \\ &= 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSE}{SST} \end{aligned}$$

Forecasting

Let x_p be a given particular value of x. The *forecast* of y for $x = x_p$ can be given by:

Confidence Interval for y at x_p

$$\hat{y} \pm t_{\frac{\alpha}{2},n-2} \times \hat{\sigma}_\epsilon \sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

Prediction Interval for value of y at x_p

$$\hat{y} \pm t_{\frac{\alpha}{2},n-2} \times \hat{\sigma}_\epsilon \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

4B; Wk. 8 Material, Regression Analysis

Topic 4B | Lecture of Wk. 8 | Tutorial held Week 10

Contents

Introduction to Multiple Regression

Where a response value y is *linearly* related to multiple independent values then:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots + \epsilon$$

Where ϵ represents the residual or error that occurs disjoint from the model.

Solving for Coefficient Values via Matrix values

The statistical model must satisfy the following n equations:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \varepsilon_2$$
$$\ldots \ldots$$
$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \varepsilon_n$$

Thus writing it in matrix form:

$$y_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix}, x_{n\times 4} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{bmatrix}, \beta_{4\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \varepsilon_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}.$$

The coefficient values can be estimated by Minitab or Excel and the use of matrix mathematics isn't within the scope of this unit.

Assessing Overall Significance of Regresion[12]

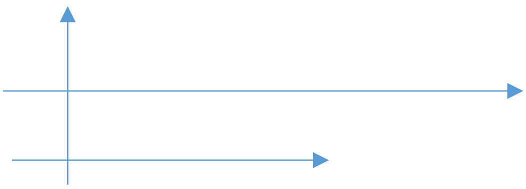
Coefficient of Zero

If the coefficient (β_k) of any x term in the regression equation:

$$y = \beta_0 + \beta_1\, x_1 + \ \beta_2 x_2 + \beta_3\, x_3 \ldots \beta_k\, x_{k+1}$$

Is insignificant it will be zero, that is to say a coefficient of zero will represent a linear regression with no real relationship between x & y.

(This doesn't include the intercept value of)



Hypothesis Test of overall significance (F Test for Significance)

Before determining whether or not specific coefficients are significant to the linear regression it is best to perform an overall test for overall fit

The *Test Statistic* and degrees of freedom can all be found by way of the ANOVA table.

Hypothesis

$$H_0 \colon \beta_1 = \beta_2 = \beta_3 = \beta_3 = \beta_4 = \cdots \beta_k = 0$$

All coefficient values are insignificant and equal to zero

$$H_a \colon At\ least\ one\ of\ the\ coefficients\ (\beta_k) is\ not\ zero.$$

Rejection Region

H_0 fails for:

$$F_{Calc} > F_{df_R,\ df_E,\ \alpha}$$

Where:

$$df_R = df_{Regression} = k$$
$$df_E = df_{Error} = n - k - 1\ (Error\ AKA\ Residual)$$

Test Statistic

$$F_{Calc} = \frac{MSR}{MSE} = \frac{SSR/df_R}{SSE/df_E}$$

Conclusion

At least on coefficient (β_k) is nonzero and the regression is overall significant

All coefficients are zero and the regression has no significance.

[12] David Doane & Lori Seward, *Applied Statistics In Business & Economics*, (McGraw Hill Publishing, 2013, 4th ed.) ch. 13.2, p. 553

ANOVA Table for F-Test of Significance

Source of Variation	Sum of Squares	df	Mean Square	F-stat
Regression (Explained variation of Data as per the Model)	$SSR = \sum_{i=1}^{n}[(\hat{y}_i - \bar{y})^2]$	k	$MSR = \frac{SSR}{k}$	$F_{Calc} = \frac{MSR}{MSE}$
Residual or Error (Random Variation that occurs distinct of the model)	$SSE = \sum_{i=1}^{n}[(y_i - \hat{y})^2]$	$n-k-1$	$MSE = \frac{SSE}{n-k-1}$	
Total)	$SST = SSE + SSR$ $= s^2(n-1)$ $= \sum_{i=1}^{n}[(y_i - \bar{y})^2]$	$n-1$		

The Excel p-value is given by:

$$= F.DIST.RT(F_{calc}, k, (n-k-1)$$

The variance of the regression is given by:

$$s^2 = \frac{SSE}{n-k-1}$$

The standard error of the regression is:

$$s = \sqrt{\frac{SSE}{n-k-1}}$$

Significance of Individual Coefficients[13]

In order to test whether an individual coefficient is significant it is tested as being equal to zero, by default the tests are two tailed because if the null hypothesis can be rejected by a two tailed test then it can also by rejected in a one-tailed test at the same α.

Hypothesis Test

Hypothesis

$H_0: \beta_j = 0 \; (x_j \text{ is not related to } y)$

$H_a: \beta_j \neq 0 \; (x_j \text{ is related to } y)$

Test Statistic

$$t_{calc} = \frac{b_j - \beta_j}{s_j} = \frac{b_{j-0}}{s_j} = \frac{b_j}{s_j}$$

Where:

$s_j = s\sqrt{c_{jj}} = \sqrt{\frac{MSE}{SS_{x_j}(1-R_j^2)}}$, s_j represents the standard error of b_i, (the standard error being the standard deviation)

The value of s_j is not calculated because the calculation is tedious, instead it is usually taken from a Minitab or excel output.

Rejection Region

The Null Hypothesis fails for:

$|t_{Calc}| > t_{dfE,\frac{\alpha}{2}}$

$|t_{Calc}| > t_{n-k-1,\frac{\alpha}{2}}$

Where:

$df_R = df_{Regression} = k$
$df_E = df_{Error} = n - k - 1$

Confidence Interval

A $(1-\alpha)\%$ Confidence Interval for the Coefficient β_j is given by:

$$\beta_j = b_j \pm \frac{s_j}{\sqrt{n}} \; t_{\frac{\alpha}{2}, df_E}$$

[13] Ibid, Ch. 13.3, p. 557

Excel Output

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	9694299.568	4847149.784	50.269	0.001
Residual	4	385700.432	96425.108		
Total	6	10080000.000			

	Coefficients	*Std Error*	*t Stat*	*P-values*	*Lower 95%*	*Upper 95%*
Intercept	8536.214	386.912	22.062	0.000	7461.975	9610.453
Price	-835.722	99.653	-8.386	0.001	-1112.404	-559.041
Advertising	0.592	0.104	5.676	0.005	0.303	0.882

In this layout the upper table represents a typical ANOVA table and the bottom layout represents data regarding coefficient values.

- *Coefficients:* This is the corresponding value of b_j
- *StdError:* This is the value of s_j
- *t Stat*: This is the value of t_{Calc} for the significance of β_j, it is given by $t_{calc} = \frac{b_j}{s_j}$

Coefficient of Determination

More predictor values can inflate the value of R^2, which represents the proportion of data explained by the model, the adjusted R^2 value makes it relative to the number of parameters

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SST}{n-1}} = 1 - \frac{\frac{SSE}{df_E}}{\frac{SST}{df_T}} = 1 - (1 - R^2)(\frac{n-1}{n-k-1})$$

Multicollinearity

This occurs when other variables are related to one another instead of just x to y, while to some degree this is always going to be somewhat the case, the depth of such concern would depend upon the degree of multicollinearity.

Klein's Rule

Klein's Rule states that multicollinearity is only an issue where the correlation coefficient matrix demonstrats correlations higher that the overall multiple correlation coefficient, i.e. R.

Forward Selection and Backward Elimination Method

Where a coefficient has a statistically insignificant coefficient (that is the *t-stat* of that coefficient is less that t_{Crit})

The smallest t-value's can be removed one by one, (Backward elimination)

OR

The coefficients with the largest *t-stat's* can be selected to be used only, (Forward Selection)

Where: $t_{calc}(\beta_j\ test) < \frac{b_j}{s_j}$

This is all in an effort to simplify the equation

Confidence Interval of $\hat{y}$

Confidence Interval

A $(1-\alpha)$ % Confidence Interval is given by:

$$\hat{y} = \pm t_{dfE,\frac{\alpha}{2}} \times \frac{s_j}{\sqrt{n}}$$

This is a confidence Interval for the expected value of y.

This is used where

Prediction Interval

A $(1-\alpha)$ % Prediction Interval is given by:

$$\hat{y} = \pm t_{dfE,\frac{\alpha}{2}} \times s_j$$

This is a prediction interval for a single value of y

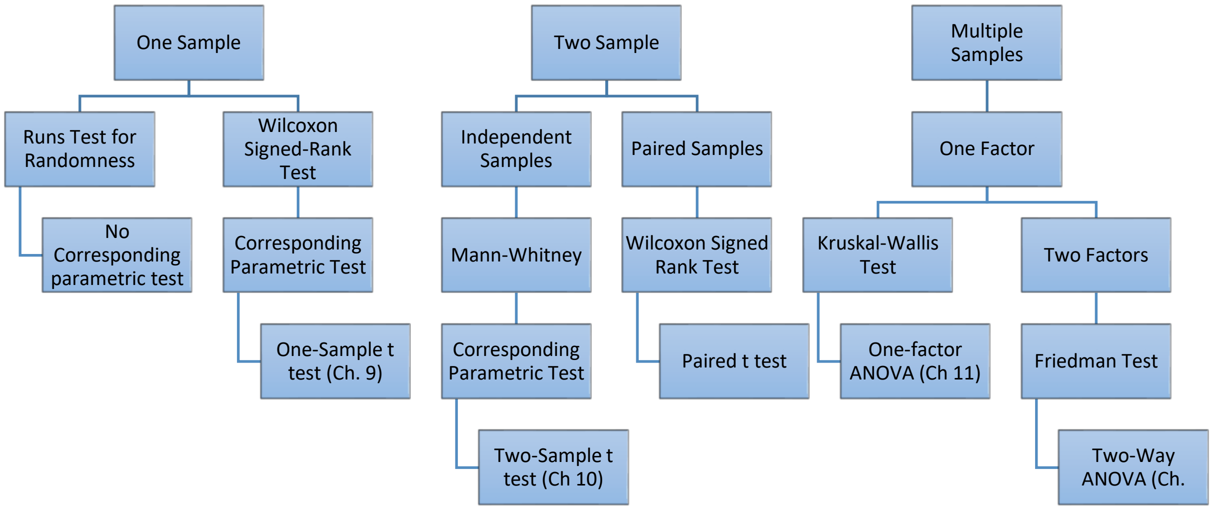
5A; Wk. 10, Non-Parametric Tests

Wk. 10 Material | Tutorial of Wk. 11 | Topic 5A

Contents

Non Parametric Hypothesis Testing

There are statistical tests which can be applied to data to avoid the assumption of the Normal Distribution. The most common are called *Non-Parametric-Tests* and are based on ranking data.



One-Sample Runs Test

The Runs test is for testing the randomness of a sample.

A *run* is a series of consecutive outcomes of the same type, surrounded by a sequence of outcomes of the other type.

$n_1 = Number\ of\ outcomes\ of\ the\ first\ type$

$n_2 = Number\ of\ outcomes\ of\ the\ second\ type$

$n = total\ sample\ size = n_1 + n_2$

Hypothesis

$$H_0: Events\ follow\ a\ random\ pattern$$

$$H_1: Events\ do\ not\ follow\ a\ random\ pattern$$

Test Statistic

For Large Samples (n>10)

$$Z_{Calc} = \frac{R - \mu_R}{\sigma_R}$$

$$\mu_R = \frac{2n_1n_2}{n} + 1 \qquad \sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}}$$

For Small Samples (small or big)

The test statistic is given by the number of runs, e.g.

DAAAADDDDAAADDAAAADDAAAA ⟶ D AAAA DDDD AAA DD AAAA DD AAAA

So there are eight distinct groups of data (runs.

R=8

P-value

As with all hypothesis tests, the smaller the *p-value*, the stronger the case is to reject H_0.

Rejection Region

For Large Samples (n>10)

H_0 is rejected for:

$$|Z| > Z_\alpha$$

For Small Samples (small or big)

For small samples H_0 is rejected where:

The number of runs fall outside the interval specified by the *Runs Test Table*

Conclusion

At a $(1-\alpha)\%$ significance level the events do/don't follow a random pattern.

Run's Test Table

TABLE VIII Critical values of u[1]

Values of $u_{0.025}$

$n_1 \backslash n_2$	4	5	6	7	8	9	10	11	12	13	14	15
4		9	9									
5	9	10	10	11	11							
6	9	10	11	12	12	13	13	13	13			
7		11	12	13	13	14	14	14	14	15	15	15
8		11	12	13	14	14	15	15	16	16	16	16
9			13	14	14	15	16	16	16	17	17	18
10			13	14	15	16	16	17	17	18	18	18
11			13	14	15	16	17	17	18	19	19	19
12			13	14	16	16	17	18	19	19	20	20
13				15	16	17	18	19	19	20	20	21
14				15	16	17	18	19	20	20	21	22
15				15	16	18	18	19	20	21	22	22

Values of $u'_{0.025}$

$n_1 \backslash n_2$	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2											2	2	2	2
3					2	2	2	2	2	2	2	2	2	3
4				2	2	2	3	3	3	3	3	3	3	3
5			2	2	3	3	3	3	3	4	4	4	4	4
6		2	2	3	3	3	3	4	4	4	4	5	5	5
7		2	2	3	3	3	4	4	5	5	5	5	5	6
8		2	3	3	3	4	4	5	5	5	6	6	6	6
9		2	3	3	4	4	5	5	5	6	6	6	7	7
10		2	3	3	4	5	5	5	6	6	7	7	7	7
11		2	3	4	4	5	5	6	6	7	7	7	8	8
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10

[1] This table is adapted, by permission, from F. S. Swed and C. Eisenhart, "Tables for testing randomness of grouping in a sequence of alternatives," *Annals of Mathematical Statistics*, Vol. 14.

TABLE VIII Critical Values of u (*Continued*)

Values of $u_{0.005}$

$n_1 \backslash n_2$	5	6	7	8	9	10	11	12	13	14	15
5		11									
6	11	12	13	13							
7		13	13	14	15	15	15				
8		13	14	15	15	16	16	17	17	17	
9			15	15	16	17	17	18	18	18	19
10			15	16	17	17	18	19	19	19	20
11			15	16	17	18	19	19	20	20	21
12				17	18	19	19	20	21	21	22
13				17	18	19	20	21	21	22	22
14				17	18	19	20	21	22	23	23
15					19	20	21	22	22	23	24

Values of $u'_{0.005}$

$n_1 \backslash n_2$	3	4	5	6	7	8	9	10	11	12	13	14	15
3										2	2	2	2
4						2	2	2	2	2	2	2	3
5				2	2	2	2	3	3	3	3	3	3
6			2	2	2	3	3	3	3	3	3	4	4
7			2	2	3	3	3	3	4	4	4	4	4
8		2	2	3	3	3	3	4	4	4	5	5	5
9		2	2	3	3	3	4	4	5	5	5	5	6
10		2	3	3	3	4	4	5	5	5	5	6	6
11		2	3	3	4	4	5	5	5	6	6	6	7
12	2	2	3	3	4	4	5	5	6	6	6	7	7
13	2	2	3	3	4	5	5	5	6	6	7	7	7
14	2	2	3	4	4	5	5	6	6	7	7	7	8
15	2	3	3	4	4	5	6	6	7	7	7	8	8

Mann Whitney U Test

This test involves pooling the data and ranking the values from 1 to $(n_1\ to\ n_2)$, if one of the populations significantly differ it will be to the left or the right of the other.

Calculating Ranks

1. Combine all the samples and sort the values from lowest to highset
2. Assign an ascending rank to each value
3. Give repeating values the same rank number
    a. The rank is the average of all of the previous ranks.
    b. All other numbers keep their original rank
4. Sum the value of the ranks from each column as T_1, T_2
    a. The sum of the ranks will satisfy:
        i. $T_1 + T_2 = \frac{n(n+1)}{2}$
5. Find the Average Values of T_1 and T_2, divide T by the number of values in that sample.

Hypothesis

$$H_0: \mu_1 = \mu_2\ ;\ both\ population\ means\ are\ equal.$$

1. $H_a: \mu_1 < \mu_2$ 2. $H_a: \mu_1 > \mu_2$ 3. $H_a: \mu_1 \neq \mu_2$

Test Statistic

$$Z_{Calc} = \frac{\overline{T_1} - \overline{T_2}}{(n_1 + n_2)\sqrt{\frac{n_1 + n_2 + 1}{12 n_1 n_2}}}$$

$$\overline{T_k} = \frac{\overline{T}_k}{n_k}$$

Rejection Region

$\alpha = ???$

Reject H_0 for:

1. $H_a: Z_{Calc} < Z_\alpha$ 2. $H_a: Z_{Calc} > Z_\alpha$ 3. $H_a: |Z_{Calc}| > \left|Z_{\frac{\alpha}{2}}\right|$

Conclusion

At some significance level the population means do/don't differ from one another.

Mann Whitney Example

2004 Times	23 22	18 28	27 18	14 15	17 21	26 25	19 18
2009 Times	17 18	21 10	16 12	20 15	14 11	13	

Combine all the data and rank it:

Rank of United Values	***Values***
1	10
2	11
3	12
4	13
5	14
6	14
7	15
8	15
9	16
10	17
11	17
12	18
13	18
14	18
15	18
16	19
17	20
18	21
19	21
20	22
21	23
22	25
23	26
24	27
25	28

If two or more numbers are the same they must be given the same rank, this rank is the average of all of the corresponding united ranks (e.g. for 18, the rank would be $\frac{12+13+14+15}{4} = 13.5$).

Otherwise the Ranks stay the same.

Rank of United Values	***Values***	***Rank***
1	10	1
2	11	2
3	12	3
4	13	4
5	14	5.5
6	14	5.5
7	15	7.5
8	15	7.5
9	16	9
10	17	10.5
11	17	10.5
12	18	13.5
13	18	13.5
14	18	13.5
15	18	13.5
16	19	16
17	20	17
18	21	18
19	21	19
20	22	20
21	23	21
22	25	22
23	26	23
24	27	24
25	28	25

Now the data needs to be reseparated

2004	***Rank04***
23	21
18	13.5
27	24
14	5.5
17	10.5
26	23
19	16
22	20
28	25
18	13.5
15	7.5
21	18.5
25	22
18	13.5
Total	233.5
Average	16.67857

2009	***Rank09***
17	10.5
21	18.5
16	9
20	17
14	5.5
13	4
18	13.5
10	1
12	3
15	7.5
10	2
Total	91.5
Average	8.318182

Hypothesis

$$H_0: \mu_{04} = \mu_{09}\ ; both\ population\ means\ are\ equal.$$

$$H_a: \mu_{04} \neq \mu_{09}\ ; The\ population\ means\ are\ different.$$

Test Statistic

$$Z_{Calc} = \frac{\overline{T_{04}} - \overline{T_{09}}}{(n_{04} + n_{09})\sqrt{\frac{n_{04} + n_{09} + 1}{12 n_{04} n_{09}}}} = \frac{16.67857 - 8.318182}{(14 + 11)\sqrt{\frac{14 + 11 + 1}{12 \times 14 \times 11}}} = 2.832$$

$$Z_{Calc} = 2.832$$

Rejection Region

$\alpha = 0.05$

Reject H_0 for:

$$Z_{Calc} < Z_{0.05}$$

$$2.832 \nless 1.6$$

H_0 is not rejected

Conclusion

At a 95% significance level there is not enough data to conclude that the population means are not equal.

The Sign Test – Single Sample

Hypothesis

This is really a test that the median equals μ_0 but the mean and median are equal if it is assumed that the population is symmetric.

$$H_0: \mu = \mu_0$$

$1. H_a: \mu < \mu_0$ $\quad$ $2. H_a: \mu > \mu_0$ $\quad$ $3. H_a: \mu \neq \mu_0$

Test Statistic

The data can be ranked and the number of sample values that exceed μ_0 can be counted

$x =$ The number of positive signs among n values that do not equal the population mean

A number is a plus sign if it is greater that the population mean in the null hypothesis.

Ignore ties and reduce the sample size by 1 for each tie.

Assuming $p = 0.5$ if H_0 is correct:

$$Z_{Calc} = \frac{x - np}{\sqrt{npq}} = \frac{x - 0.5n}{0.5\sqrt{n}}$$

$$Z_{Calc} = \frac{x - 0.5n}{0.5\sqrt{n}}$$

Rejection Region

H_0 is rejected for :

$1. H_a: Z_{Calc} < Z_\alpha$ $\quad$ $2. H_a: Z_{Calc} > Z_\alpha$ $\quad$ $3.\ H_a: |Z_{Calc}| > \left|Z_{\frac{\alpha}{2}}\right|$

Single Sample Sign Test Example

To determine the effectiveness of a new traffic control system the numbers of accidents that occurred at a random sample of 1 dangerous intersections during the 4 weeks before and after the installation of the new system were observed with the following results:

Before	9	7	3	16	12	12	5	6
After	5	3	4	11	7	5	5	1
Difference	4	4	-1	5	5	7	0	5

Sign Value	+	+	-	+	+	+	0	+

Revised Sign Value	+	+	-	+	+	+	+

Thus one tie is removed and the population size is reduced by that one value:

- $n = 7$
- - = 1
- X = + = 6

Hypothesis

$$H_0: \mu_b - \mu_a = 0$$
$$H_a: \mu_b - \mu_a = D \qquad , D > 0$$

Test Statistic

Population mean of the null hypothesis: 0

$$Z_{Calc} = \frac{x - 0.5n}{0.5\sqrt{n}}$$
$$Z_{Calc} = \frac{6 - 0.5 \times 7}{0.5\sqrt{7}} = 1.8898$$

Rejection Region

H_0 fails for:

$$Z_{Calc} > Z_\alpha$$
$$1.8898 > 1.645$$

Thus the null hypothesis is rejected

Conclusion

At a 95% significance level the number of accidents prior to the new traffic control system was more than afterwards.

5B; Wk. 11, Non-Parametric Tests 2

Wk. 11 Material | Tutorial of Wk. 12 | Topic 5B

Contents

Wilcoxon Signed-Rank Test[14]

The *Wilcoxon Signed-rank* test is the non-parametric version of a student's t-test for paired samples, it compares the differences of paired samples without the assumption of normality. The population should be roughly symmetric.

Data

Obs. 1	6	4	8	6	8	4
Obs. 2	3	2	9	5	2	5
Difference	3	2	-1	4	6	-1

Hypothesis

Right-Tailed	*Left-Tailed*	*Two-Tailed*
$H_0: M \leq M_0$	$H_0: M \geq M_0$	$H_0: M = M_0$
$H_a: M > M_0$	$H_a: M < M_0$	$H_0: M \neq M_0$

This can also evaluate the median difference between paired observations (M_d)

Right-Tailed	*Left-Tailed*	*Two-Tailed*
$H_0: M_d \leq 0$	$H_0: M_d \geq 0$	$H_0: M_d = 0$
$H_a: M_d > 0$	$H_a: M_d < 0_0$	$H_0: M_d \neq 0$

Test Statistic

Calculate the difference between paired ovservations, eliminate samples where d=0

Rank the differences from smallest to largest by absolute value

Add the ranks of the (all positive) differences to obtain the rank sum (W)

$$W = \sum_{i=1}^{n} [|R|] \text{, the sum of all positive ranks}$$

$$\mu_W = \frac{n(n+1)}{4}, \textit{the expected value of the W statistic}$$

$$\sigma_W = \sqrt{\left[\frac{n(n+1)(2n+1)}{24}\right]}$$

For large samples (n≥20), the test statistic is approximately normal:

$$Z_{Calc} = \frac{W - \mu_W}{\sigma_W} = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\left[\frac{n(n+1)(2n+1)}{24}\right]}}$$

Where data is a small sample (n<20), a special table is required to obtain critical values (that is the test statistic).

Rejection Region

The null hypothesis is rejected where:

$$p - value \ < \alpha$$

The null hypothesis is rejected also where

[14] P. 698 of David P. Doane & Lori E. Seward, Applied Statistics in Business & Economics, (McGraw Hill, 4th Ed., 2013)

Right-Tailed	*Left-Tailed*	*Two-Tailed*
$Z_{Calc} < -Z_\alpha$	$Z_{Calc} > Z_\alpha$	$\lvert Z_{Calc}\rvert < \left\lvert Z_{\frac{\alpha}{2}}\right\rvert$

Kruskal-Wallis[15]

The Kruskal Wallis is the non-parametric version of a one factor ANOVA (Ch. 11).

Data

Data Group 1	***Rank of 1***	***Data Group 2***	***Rank of 2***	***Data Group 3***	***Rank of 3***
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

For ranking procedures refer below to the Friedman Test for Related Samples.

Hypothesis

$$H_0: All\ c\ Population\ medians\ are\ the\ same, there\ exists\ no\ significant\ differene$$

$$H_a: Atleast\ one\ population\ median\ differs$$

Test Statistic

The Test Statistic follows a Chi-Square Distribution:

$$X^2_{Calc} = H_{Calc} = \frac{12}{n(n+1)}\sum_{j=1}^{c}\left[\frac{T_j^2}{n_j}\right] - 3(n+1)$$

Where:

- j : is the number of the Data Group (above is 1, 2 & 3)
- c : is the total number of Data Groups (There are 3 above)
- T_j : is the sum of the corresponding ranks of group j
- n : is the total number of observations
- n_j: is the total number of observations in group j

Rejection Region

The null hypothesis is rejected where:

$$\chi^2_{Calc} > \chi^2_{d.f.,\alpha}$$

$$H_{Calc} > \chi^2_{,d.f.,\alpha}$$

$$\chi^2_{Calc} = H_{Calc}$$

Where:

$$d.f. = degrees\ of\ freedom = c - 1$$

Conclusion

Bigger Test Stat

Null hypothesis is rejected; At least one population median significantly differs between the others.

[15] P. 701 of David P. Doane & Lori E. Seward, Applied Statistics in Business & Economics, (McGraw Hill, 4th Ed., 2013)

Smaller Test Stat

Null hypothesis is **NOT** rejected; There is not enough evidence to suggest that population medians significantly differ between groups (AKA Treatments).

Friedman Test for Related Samples[16]

This is the non-parametric version of a two-factor ANOVA

The Friedman Test resembles the *Kruskal-Wallis* test except it also specifies r block factor levels.

The number of columns or the number of rows (at least one of the two) must be greater than 5.

Data

An example of what the data might look like is:

	Shiny	Satin	Pebbled	Pattern	Embossed
Youth Under 21	6.7	6.6	5.5	4.3	4.4
Adult (21 to 39)	5.5	5.3	6.2	5.9	6.2
Middle−Age (40−61)	4.5	5.1	6.7	5.5	5.4
Senior (62 and over)	3.9	4.5	6.1	4.1	4.9

This has:

- $c = 5$
- $r = 4$

Data must be ranked by blocks like so:

[In the case of the Kruskal-Wallis Test there are no blocks and as such the data can be ranked all together]

Youth Under 21	
Observed Value	***Relative Rank***
4.30	1
4.40	2
5.50	3
6.60	4
6.70	5

Middle−Age (40−61)	
Observed Value	***Relative Rank***
4.50	1
5.10	2
5.40	3
5.50	4
6.70	5

Adult (21 to 39)	
Observed Value	***Relative Rank***
5.30	1
5.50	2
5.90	3
6.20	4

Senior (62 and over)	
Observed Value	***Relative Rank***
3.90	1
4.10	2
4.50	3
4.90	4

[16] P. 706 of David P. Doane & Lori E. Seward, Applied Statistics in Business & Economics, (McGraw Hill, 4th Ed., 2013)

6.20	5			6.10	5

And then the data must be recombined (Excel's Vlookup helps) to give a final table like so:

	Shiny		Satin		Pebbled		Pattern		Embossed	
	Consumer Rating	Relative Rank of Data	Consumer Rating	Relative Rank of Data	Consumer Rating	Relative Rank of Data	Consumer Rating	Relative Rank of Data	Consumer Rating	Relative Rank of Data
Youth Under 21	6.7	5	6.6	4	5.5	3	4.3	1	4.4	2
Adult (21 to 39)	5.5	2	5.3	1	6.2	4	5.9	3	6.2	4
Middle-Age (40-61)	4.5	1	5.1	2	6.7	5	5.5	4	5.4	3
Senior (62 and over)	3.9	1	4.5	3	6.1	5	4.1	2	4.9	4
Rank Total (T)		9		10		17		10		13

Hypothesis

$$H_0: All\ c\ populations\ have\ the\ same\ median$$

$$H_a: Not\ all\ the\ populations\ have\ the\ same\ median$$

Test Statistic

The test statistic follows the chi distribution:

$$F_{Calc} = \chi^2_{Calc} = \frac{12}{rc(c+1)} \sum_{j=1}^{c} [T_j^2] - 3r(c+1)$$

Where:

- $r =$ the number of blocks (rows)
- $c =$ the number of treatments (columns)
- $T_j =$ the sum of ranks for treatment j
- As a check of arithmetic it must be true that:
  - $\sum_{j=1}^{c} [T_j] = \frac{rc(c+1)}{2}$

Rejection Region

The null hypothesis is rejected where:

$$\chi^2_{Calc} > \chi^2_{d.f.,\alpha}$$

$$F_{Calc} > \chi^2_{,d.f.,\alpha}$$

$$\chi^2_{Calc} = F_{Calc}$$

Where:

$$d.f. = degrees\ of\ freedom = c - 1$$

6A; Wk. 12, Time Series Analysis

Wk. 12 Material | Tutorial of Wk. 13 | Topic 6A

Contents

Time Series and Trends

A time series (y_t), is data or response variables plotted against time, where time is on the x-axis

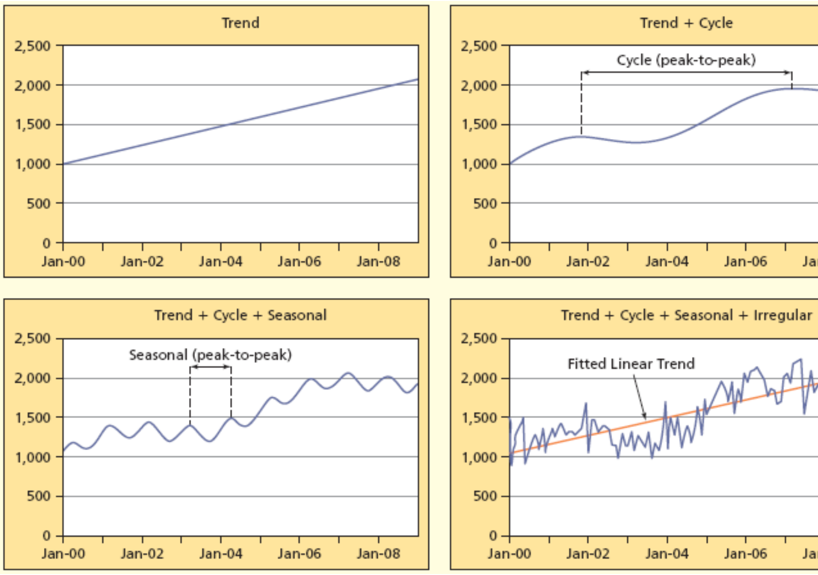
Periodicity is the time interval over which the data is collected.

Additive and Multiplicative Models

Time series can be broken up into four types of components:

- Trend (T)
- Cycle (C)
- Seasonal (S)
- Irregular (I)

These are best illustrated by way of a diagram:



It is assumed that these components interact in either an additive or multiplicative fashion

Additive Model	Multiplicative Model
$$Y = T + C + S + I$$ This is used where data is of similar magnitude (trend-free or over a short run) with constant absolute growth or decline	$$Y = T \times C \times S \times I$$ This is used where data is of increasing or decreasing magnitude (long run or trended data) with constant percent growth or decline.

The multiplicative model becomes additive as logarithms are taken (of non-negative data).

Parts of time series

Time series have two major parts, a **deterministic** part and a **stochastic** part.

Deterministic

The **deterministic** part may consist of various effects, such as,

- T_t Long-term Trend-The general movement over all years;
- C_t Cyclical effect – repetitive up and down movements about a trend that covers several years;
- S_t Seasonal effect – repetitive cyclical pattern within a year (or a week or other smaller time period)

Stochastic

The **stochastic** component is the random variation I_t.

- I_t Essentially this is just random error and fluctuation, irregular difficult to explain movement of data.

Trend Analysis Regression Techniques

Obviously data could fit a myriad of different regression models, three useful models in business are:

1. Linear Model
   a. $y_t = b_0 + t\, b_1$
2. Quadratic Model
   a. $y_t = b_0 + b_1 t + b_2 t^2$
   b. $y_t = a + b_t + c_t^2$
      i. {subscripts intentional}
3. Exponential Model
   a. $y_t = ae^{bt}$

Linear Model

This is the simplest applicable model and might suffice for short-run forecasting or as a baseline model

The linear model can be solved by the method of least squares

Quadratic Model

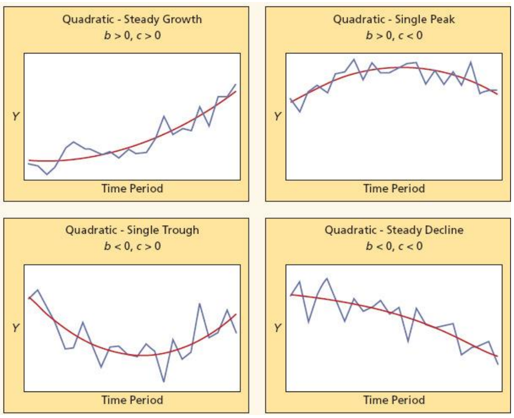
$y_t = b_0 + b_1 t + b_2 t^2$ is the quadratic model, if $b_2 = 0$ the model becomes linear (i.e. the linear model can be considered a special case of the quadratic model).

Fitting a quadratic model is one way to check for nonlinearity.

Where b_2 does not significantly differ from zero the linear model would suffice.

Shapes of the Quadratic Model

The quadratic Model can assume four shapes:



Exponential Model

The exponential model has the form $y_t = ae^{bt}$

It is used where a time series grows or declines at the same rate (b) in each time period.

This model is preferred for financial data over a longer period of time.

Linear models and exponential models may not differ much over small time periods.

Solving for an exponential model

To solve an exponential trend let:

$$z_t = \ln(y_t)\,,\text{then } z_t = \ln(a) + bt$$

1. Using the values of $\ln(y)$ solve for a linear regression.
2. The linear regression will have the form $z = \ln(a) + bt$
3. Where the values of a and b correspond to the exponential equation
4. $y_t = ae^{bt}$

Firstly we need to determine z_t

Year	1985	1986	1987	1988	...	2001	2002	2003	2004
t	1	2	3	4	...	17	18	19	20
Interest Rate (%)	14.7	16.3	14.5	13.4	...	5.3	5.3	5.1	5.8
z	2.69	2.79	2.67	2.60	...	1.67	1.67	1.63	1.76

Using Minitab we can determine the exponential model.

```
The regression equation is
z = 2.7056 - 0.0594 t

R-Sq = 79.7%
```

We now need to transform back into y

$\ln(a) = 2.7056$ $\quad a = 14.9633$

Thus $\quad y_t = 14.9633e^{-0.594t}$

Smoothing Techniques

Where data is erratic fitting a trend line could just be a waste of time, another approach is to just smooth the data out by running an average through it.

Trailing Moving Average (TMA)

This averages the data over the last m periods:

$$\hat{y}_t = \frac{y_t + t_{t-1} + \cdots + y_{t-m+1}}{m}$$

Choosing more periods i.e. a larger m yields a ‘smoother’ average but requires more data.

The value of $\hat{y}_t$ can be usesd as a forecast for period $+1$, that is it is a noe-period-ahead forecast.

Centred Moving Average (CMA)

This method looks forward and backward in time to express the current ‘forecast’ as a mean of the current observation and observations on both sides of the current data.

For example, where $m = 3$ periods are used, the CM is:

$$\hat{y}_t = \frac{y_{t-1} + y_t + y_{t+1}}{3}$$

When m is odd like 3, the CMA is easy to calculate, but when it is even, the mean of an even number of data points would lie between two data points and would be incorrectly centred, in this case a double moving average is taken to get the resulting CMA centred properly.

Example

Quarter	2001	2002	2003	2004	2005	2006
Qtr1	4,330	5,101	5,530	6,131	6,585	7,205
Qtr2	5,713	6,178	6,538	7,070	7,697	8,599
Qtr3	5,906	6,376	6,830	7,257	8,184	8,950
Qtr4	6,986	7,457	8,073	8,803	10,096	10,383

Smoothing the time series by using a centered moving average. We would use m = 4. Hence we have to calculate a double moving average. Looking at the first 6 data points.

The 1st MA is the average of data points 1 to 4

$$MA_1 = \frac{4330+5713+5906+6986}{4} = 5734$$

The 2nd MA is the average of data points 2 to 5

$$MA_2 = \frac{5713+5906+6986+5101}{4} = 59265$$

Then the first CMA is the average of the first 2 MA

$$CMA_1 = \frac{5733.75+5926.5}{2} = 5830.1$$

Exponential Smoothing

The exponential smoothing model is a special kind of moving average.

It is used where data has up=and=down movements with no consistent trend.

The updating formula is:

$$F_{t+1} = \alpha y_t + (1-\alpha)F_t$$

Where:

- F_{t+1} = the forecast for the next period
- α = the 'smoothing' constant, $0 \leq \alpha \leq 1$
- y_t = the actual data value in period t
- F_t = the previous forecast for period t

The next forecast (F_{t+1}) is a weighted average of the current data (y_t) and the previous forecast (F_t)

The value of α (the smoothing constant) is the weight given to the latest data, where a low α value would give little weight to the most recent observation, a larger α value means the forecast quickly adapts to recent data.

$\alpha = 1$ means no smoothing at all, (i.e. the forecast value for the nesxt period is the same as the latest data point.)

It should be noted that F_{t-1} is always dependent on the value of F_t

Initialising the Process

Because the forecast values all run off of one-another it is necessary to pick an initial forecast value, there are two ways to go about this

Method A – No Forecast

Just set the first forecast as the first data value:

$$F_1 = y_1$$

An unusual y_1 value may mean it could take a few iterations before the forecasts stabilize.

Method B – Average Value

Average the first 6 data values.

Set $F_1 = \frac{1}{6}(y_1 + y_2 + \cdots + y_6)$

This however consumes a bunch of data and is still somewhat vulnerable to an unusual y value.

6B; Wk. 13, Time Series Analysis

Seasonal Effect and Forecasting

Wk. 13 Material | Tutorial of Wk. 14 | Topic 6B

Contents

Calculating Seasonal Indexes

When data periodicity is monthly or quarterly, a seasonal index can be used to remove seasonal effects of a time series.

This is known as a deseasonalised or seasonal adjusted time series.

For the multiplicative model, a seasonal index is a ratio

$$Y = T \times C \times S \times I, since\ MA = T \times C, then \frac{Y}{MA} = S \times I$$

S

1. Calculate a centred moving average (CMA) for each month (or quarter or whatever)

2. Divide each observed value (y_t) by the MA to obtain seasonal Ratios

3. Average the seasonal ratios by the month (or quarter or whatever) to get raw seasonal indexes
4. Adjust the raw seasonal indexes so they sum to 12 (12 in the case of monthly or 4 quarterly and so on)
5. Divide each y_t by its seasonal index to get deseasonalized data.

s

Example 13.1

Quarterly PepsiCo Revenues (millions), 2001-2006

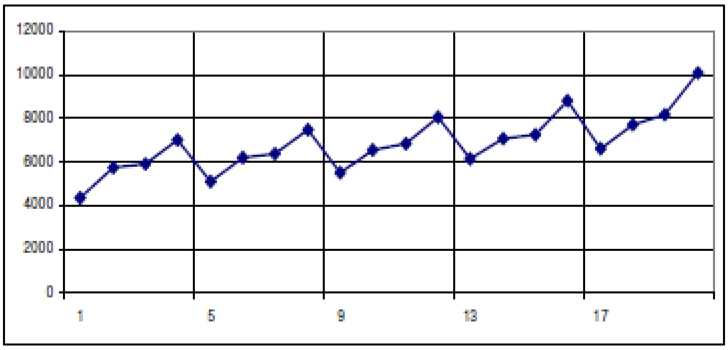
We will use 2001 - 2005 to forecast 2006 sales

Quarter	2001	2002	2003	2004	2005	2006
Qtr1	4,330	5,101	5,530	6,131	6,585	7,205
Qtr2	5,713	6,178	6,538	7,070	7,697	8,599
Qtr3	5,906	6,376	6,830	7,257	8,184	8,950
Qtr4	6,986	7,457	8,073	8,803	10,096	10,383

4

Example 13.1

Quarterly PepsiCo Revenues (millions), 2001-2005



5

Example 13.1

Firstly, we assume just a linear trend remembering to replace the year with *t*.

t	Sales
1	4330
2	5713
3	5906
4	6986
5	5,101
6	6,178
7	6,376
8	7,457
9	5,530
10	6,538
11	6,830
12	8,073
13	6,131
14	7,070
15	7,257
16	8,803
17	6,585
18	7,697
19	8,184
20	10,096

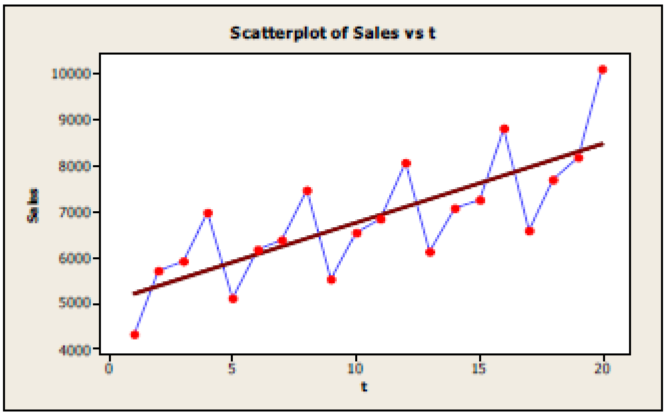
Determining the trend line

```
The regression equation is

Sales = 5027.88 + 172.78 t

R-Sq = 59.0%
```

6

Example 13.1

Quarterly PepsiCo Revenues (millions), 2001-2005



7

Example 13.1

Using this equation to forecast the next 4 quarters

```
Sales = 5027.88 + 172.78 t
```

		Actual	t	Forecast
2006	Qtr 1	7205	21	8656
2006	Qtr 2	8599	22	8829
2006	Qtr 3	8950	23	9002
2006	Qtr 4	10383	24	9175

8

Example 13.1

Our forecast is not the best. We should remove the seasonsal effects.

Step 1 Since the data is in quarters we take m = 4 and hence since m is even; to calculate the centered moving average we have to take a double moving average

We did this in last week's lecture.

9

Example 13.1

Quarter	Month	Sales		CMA
2001	Qtr 1	4330		
2001	Qtr 2	5713	5733.75	
2001	Qtr 3	5906	5926.5	5830.125
2001	Qtr 4	6986	6042.75	5984.625
2002	Qtr 1	5101	6160.25	6101.5
2002	Qtr 2	6178	6278	6219.125
2002	Qtr 3	6376	6385.25	6331.625
2002	Qtr 4	7457	6475.25	6430.25
2003	Qtr 1	5530	6588.75	6532
2003	Qtr 2	6538	6742.75	6665.75
2003	Qtr 3	6830	6893	6817.875
2003	Qtr 4	8073	7026	6959.5
2004	Qtr 1	6131	7132.75	7079.375
2004	Qtr 2	7070	7315.25	7224
2004	Qtr 3	7257	7428.75	7372
2004	Qtr 4	8803	7585.5	7507.125
2005	Qtr 1	6585	7817.25	7701.375
2005	Qtr 2	7697	8140.5	7978.875
2005	Qtr 3	8184		
2005	Qtr 4	10096		

10

Example 13.1

Step 2: Divide each observed y_t value by the MA to obtain seasonal ratios.

Year	Quarter	Sales		CMA	Ratios
2001	Qtr 1	4330			
2001	Qtr 2	5713	5733.75		
2001	Qtr 3	5906	5926.5	5830.125	1.0130
2001	Qtr 4	6986	6042.75	5984.625	1.1673
2002	Qtr 1	5101	6160.25	6101.5	0.8360
2002	Qtr 2	6178	6278	6219.125	0.9934
2002	Qtr 3	6376	6385.25	6331.625	1.0070
2002	Qtr 4	7457	6475.25	6430.25	1.1597
2003	Qtr 1	5530	6588.75	6532	0.8466
2003	Qtr 2	6538	6742.75	6665.75	0.9808
2003	Qtr 3	6830	6893	6817.875	1.0018
2003	Qtr 4	8073	7026	6959.5	1.1600
2004	Qtr 1	6131	7132.75	7079.375	0.8660
2004	Qtr 2	7070	7315.25	7224	0.9787
2004	Qtr 3	7257	7428.75	7372	0.9844
2004	Qtr 4	8803	7585.5	7507.125	1.1726
2005	Qtr 1	6585	7817.25	7701.375	0.8550
2005	Qtr 2	7697	8140.5	7978.875	0.9647
2005	Qtr 3	8184			
2005	Qtr 4	10096			

Example 13.1

Step 3: Average the seasonal ratios by the month (quarter) to get raw seasonal indexes.

Quarter	2001	2002	2003	2004	2005	Average
1		0.8360	0.8466	0.8660	0.8550	0.8509
2		0.9934	0.9808	0.9787	0.9647	0.9794
3	1.0130	1.0070	1.0018	0.9844		1.0016
4	1.1673	1.1597	1.1600	1.1726		1.1649

The sum of the averages = 3.9968.

12

Example 13.1

Step 4: Adjust the raw seasonal indexes so they sum to 4 (quarterly).

The sum of the averages = 3.9968. Since we are working with quarters, we need to scale the averages so they add to 4. This is done by the following $\times \frac{4}{3.9968}$

Seasonal Ratio	Multiply by	Adjusted Seasonal Ratio
0.8509	1.0008	0.8516
0.9794		0.9802
1.0016		1.0024
1.1649		1.1658
3.9968		4.0000

13

Example 13.1

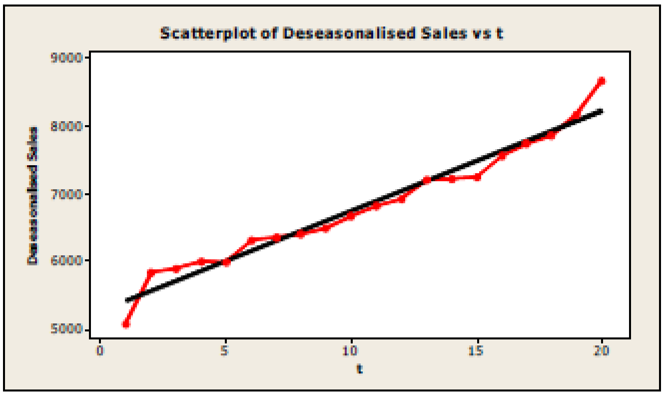
Step 5: Divide each y_t by its seasonal index to get deseasonalized data.

Year	Quarter	t	Sales	Adjusted Seasonal Ratio	Deseasonalised Sales
2001	Qtr 1	1	4330	0.8516	5084.50
2001	Qtr 2	2	5713	0.9802	5828.53
2001	Qtr 3	3	5906	1.0024	5892.14
2001	Qtr 4	4	6986	1.1658	5992.26
2002	Qtr 1	5	5101	0.8516	5989.85
2002	Qtr 2	6	6178	0.9802	6302.94
2002	Qtr 3	7	6376	1.0024	6361.04
2002	Qtr 4	8	7457	1.1658	6396.26
2003	Qtr 1	9	5530	0.8516	6493.60
2003	Qtr 2	10	6538	0.9802	6670.22
2003	Qtr 3	11	6830	1.0024	6813.97
2003	Qtr 4	12	8073	1.1658	6924.64
2004	Qtr 1	13	6131	0.8516	7199.33
2004	Qtr 2	14	7070	0.9802	7212.97
2004	Qtr 3	15	7257	1.0024	7239.97
2004	Qtr 4	16	8803	1.1658	7550.80
2005	Qtr 1	17	6585	0.8516	7732.44
2005	Qtr 2	18	7697	0.9802	7852.65
2005	Qtr 3	19	8184	1.0024	8164.80
2005	Qtr 4	20	10096	1.1658	8659.88

14

Example 13.1

Plotting the Deseasonlised data with a linear trend



15

Example 13.1

Once we have the seasonally adjusted data, we can then apply a linear trend to the data. Using Minitab

```
The regression equation is

Deseasonalised Sales = 5265.26 + 147.893 t

R-Sq = 96.1%
```

16

Example 13.1

Use this equation to forecast the next 4 quarters. This formula will just find the trend effect T_t and then we will have to multiply it by the seasonal effect S_t.

		Actual	t	Seasonal effect	Trend Effect = 5265.23 + 147.892t	Forecasted Sales
2006	Qtr 1	7205	21	0.8516	8370.96	7129
2006	Qtr 2	8599	22	0.9802	8518.85	8350
2006	Qtr 3	8950	23	1.0024	8666.75	8687
2006	Qtr 4	10383	24	1.1658	8814.64	10276

Our forecast is much better after accounting for the seasonal effects.

17