

How to select an appropriate statistical method

Read the problem carefully to determine:

- ❖ the population of interest,
- ❖ variable(s) of interest, and scale of data (quantitative if MEASUREMENT, qualitative if COUNT DATA)
- ❖ if the problem involves estimating or testing the MEAN or PROPORTION, if you are given 1 sample, 2 samples, or more than 2 samples,
- ❖ if the problem involves building a model for a RESPONSE VARIABLE in terms of PREDICTOR VARIABLES(s)

PROBLEM	ONE VARIABLE	VARIABLE QUANTITATIVE PARAMETER = MEAN	VARIABLE QUALITATIVE PARAMETER = PROPORTION
ESTIMATION OR TESTING HYPOTHESIS	1 - SAMPLE PROBLEM	If sample is approximately normal or $n \geq 25$, use 1-sample t (page 3)	$np \geq 5$, $n(1-p) \geq 5$ use normal approximation, else exact binomial distribution. (p. 6 - 7)
	2-SAMPLE PROBLEM	Case 1: 2 independent samples If both samples are approximately normal, or $n_1, n_2 \geq 25$, use 2-sample t (pages 4 - 5) CASE 1 (A) - Equal Variances Calculate s^2_{pooled} , use 2-sample t based on s^2_{pooled} , with $df = n_1 + n_2 - 2$ CASE 1 (B) - Unequal variances Use Cochran's or Satterthwaite's approximate t Case 2: Paired sample If $D = X_1 - X_2$ is normally distributed, use 1-sample t on D	If $np_i \geq 5$, $n(1-p_i) \geq 5$, $i = 1, 2$ use normal approximation, else exact binomial distribution. (page 8)
	MORE THAN 2 SAMPLES	1-WAY ANOVA if subjects are homogeneous (pages 9 - 10) 2-WAY ANOVA if BLOCKS of subjects are present Store Residuals, Test Normality of Residuals.	To test equality of several proportions, use the chi-square test. (pages 12 - 18)
FITTING A REGRESSION MODEL (page 11)	ONE RESPONSE VARIABLE, ONE PREDICTOR	Fit a straight line to data $\{(x_1, y_1), \dots, (x_n, y_n)\}$	$y = \text{binary (0/1)}$ Fit a logistic equation to data $\{(x_1, y_1), \dots, (x_n, y_n)\}$
	ONE RESPONSE VARIABLE, SEVERAL PREDICTORS	Fit a MLR equation to data $\{(x_{1j}, x_{2j}, \dots, x_{kj}, y_j), j = 1, 2, \dots, n\}$	$y = \text{binary (0/1)}$ Fit a LOGISTIC regression equation to data $\{(x_{1j}, x_{2j}, \dots, x_{kj}, y_j), j = 1, 2, \dots, n\}$

1- sample t (quantitative variable only) : Point Estimates/95% Confidence Interval:

$\hat{\mu} = \bar{x}$, sample mean, $\hat{\sigma} = s$, sample sd, 95% CI for μ : $\bar{x} \pm t_{n-1, .975} \frac{s}{\sqrt{n}}$

1) Test of Test $H_0 : \mu = \mu_0$ vs. H_1 $t_{CALC} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

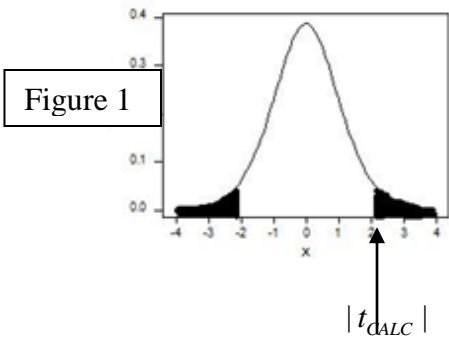


Figure 1

$H_0 = \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ (2-sided) if $P < \alpha = .05$, where $P = \text{shaded area}$

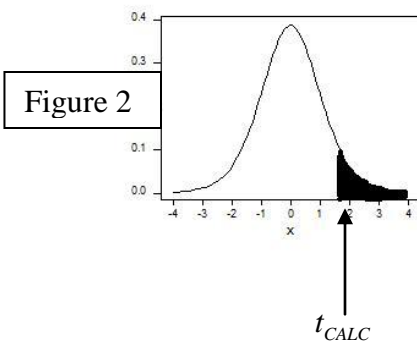


Figure 2

Reject $H_0 = \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$ (1-sided, to right) if $P < \alpha = .05$, where $P = \text{shaded area}$

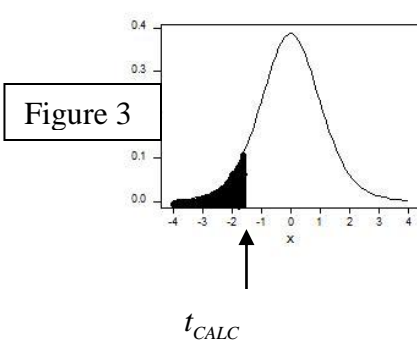


Figure 3

Reject $H_0 = \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$ (1-sided, to left) if $P < \alpha = .05$, where $P = \text{shaded area}$

2- sample t (quantitative variable only)

Case 1A: Two independent samples

	Sample size	Mean	sd
Sample 1	n_1	\bar{x}_1	s_1
Sample 2	n_2	\bar{x}_2	s_2

If the two variances are equal ($\sigma_1^2 = \sigma_2^2$) – calculate

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Confidence Interval for Difference $\mu_1 - \mu_2$

95% CI for $\mu_1 - \mu_2$ is: $(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2} s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

- If 0 falls inside this CI, declare $\mu_1 = \mu_2$

Testing Equality of Two Means

Test of $H_0 : \mu_1 = \mu_2$ is same as $H_0 : \mu_1 - \mu_2 = 0$.

$$t_{CALC} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

2-sided alternative is handled in a manner similar to the 1-sample problem (see Figures 1-3 on page 3); degrees of freedom for the t-table equals $n_1 + n_2 - 2$.

Case 1B: Variances are NOT equal ($\sigma_1^2 \neq \sigma_2^2$) - many approximate t-tests (e.g., Cochran's or Satterthwaite's) exist. Use software such as SPSS or MINITAB.

Case 2: Paired Samples

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, $\bar{d} = \sum_{i=1}^n \frac{d_i}{n} =$ sample mean of differences $d_i = x_i - y_i$,

$s_d =$ sd of differences d_i

$$t_{CALC} = \frac{\bar{d} - 0}{s_d / \sqrt{n}}$$

Confidence Interval

95% CI for $\mu_1 - \mu_2$ is: $\bar{d} \pm t_{n-1} \frac{s_d}{\sqrt{n}}$

Declare the two population means equal if 0 is inside the above confidence interval.

Testing Equality of the Two Population Means

Use 1-sample t-test on $d_i = x_i - y_i$

Variable Qualitative (Typically question will involve the population proportion p)

Estimation/testing of one proportion

Given $X = \#$ of successes in n independent and identical trials of a random experiment that only has 2 outcomes (SUCCESS and FAILURE)

X follows a binomial distribution with $\#$ of trials n and SUCCESS PROBABILITY or proportion of successes in the population p .

Confidence Interval for p

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} sd(\hat{p}) = \hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $z_{1-\frac{\alpha}{2}}$ is the z-value from the standard normal table

corresponding to the confidence coefficient $100(1-\alpha)$.

Example: For 95% confidence, $100(1-\alpha)=95$,

$$(1-\alpha) = .95, \alpha = .05, 1-\frac{\alpha}{2} = 1-.025 = .975$$

$$z_{1-\frac{\alpha}{2}} = z_{.975} = 1.96$$

Testing Hypothesis

$$H_0 : p = p_0$$

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

$$z_{calc} = \frac{\hat{p} - p_0}{sd(\hat{p})} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Reject H_0 vs. $H_1 : p \neq p_0$ if

$$P = P(Z > |z_{calc}|) < .05$$

Reject H_0 vs. $H_1 : p > p_0$ if

$$P = P(Z > z_{calc}) < .05$$

Reject H_0 vs. $H_1 : p < p_0$ if

$$P = P(Z < z_{calc}) < .05$$

Note that these tests are similar to the tests involving one population mean (page 2 of this handout), the only difference being that we use the normal table instead of the t-table.

Estimation/testing for two proportions

Given $x_1 = \#$ of successes in n_1 trials from binomial distribution with success probability p_1

$x_2 = \#$ of successes in n_2 trials from binomial distribution with success probability p_2

95% confidence interval for $p_1 - p_2$

$$\hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$$

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times sd(\hat{p}_1 - \hat{p}_2)$$

where

$$sd(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Test of $H_0 : p_1 = p_2$

Calculate

$$z_{calc} = \frac{\hat{p}_1 - \hat{p}_2}{sd(\hat{p}_1 - \hat{p}_2 | H_0 \text{ is true})}$$

$$sd(\hat{p}_1 - \hat{p}_2 | H_0 \text{ is true}) = \sqrt{\frac{\bar{p}(1 - \bar{p})}{n_1} + \frac{\bar{p}(1 - \bar{p})}{n_2}}$$

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

H_0 is rejected exactly as in the 1 proportion case (page 7).

Details of ANOVA

Given: k independent samples

$$\{x_{11}, x_{12}, \dots, x_{1n_1}; x_{21}, x_{22}, \dots, x_{2n_2}; \dots; x_{k1}, x_{k2}, \dots, x_{kn_k}\}$$

1-st sample 2-nd sample k-th sample

Treatment 1 Treatment 2 Treatment k

Assume:

$$x_{ij} = \mu + \alpha_i + e_{ij}; i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$$

e_{ij} are independent errors, assumed to be normally distributed

with common variance σ^2

μ = overall population mean of variable x

α_i = effect (above or below grand mean μ) of Treatment i

Population mean of x for Treatment $i = \mu + \alpha_i$

H_0 : all means are equal is same as H_0 : all $\alpha_i = 0$

Notations:

$$\bar{x}_{..} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}}{n_1 + n_2 + \dots + n_k} = \text{grand sample mean,}$$

$$\bar{x}_{.i} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} = \text{mean of sample } i$$

We can show that

$$\text{Total Sum of Squares} = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_{ij} - \bar{x}_{.i})^2 + \sum_{i=1}^k n_i (\bar{x}_{.i} - \bar{x}_{..})^2$$

or

$$TSS = SS_Treatment + SS_Error$$

degrees of freedom are:

$$df(TSS) = N - 1, \quad N = n_1 + n_2 + \dots + n_k$$

$$df(Treatment) = k - 1$$

$$df(Error) = N - 1 - (k - 1) = N - k$$

$$MS_Treatment = \frac{SS_Treatment}{k - 1}$$

$$MSE = \frac{SS_Error}{N - k}$$

$$F_{calc} = \frac{MS_Treatment}{MSE}$$

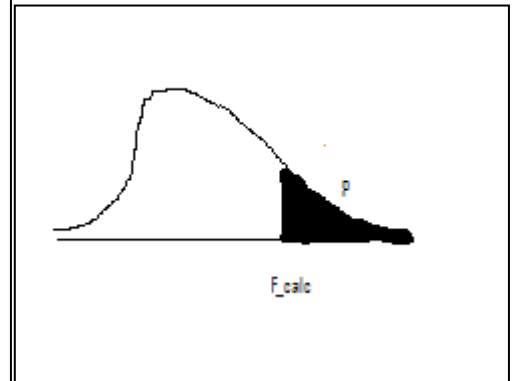
Distribution of F_{calc} when null is true is F with degrees of freedom

$$df(Num) = k - 1$$

$$df(den) = N - k$$

Reject if F_{calc} is larger than F-table value, or if

$$P = P(F_{k-1, N-k} > F_{calc}) < .05$$



=

ANOVA TABLE

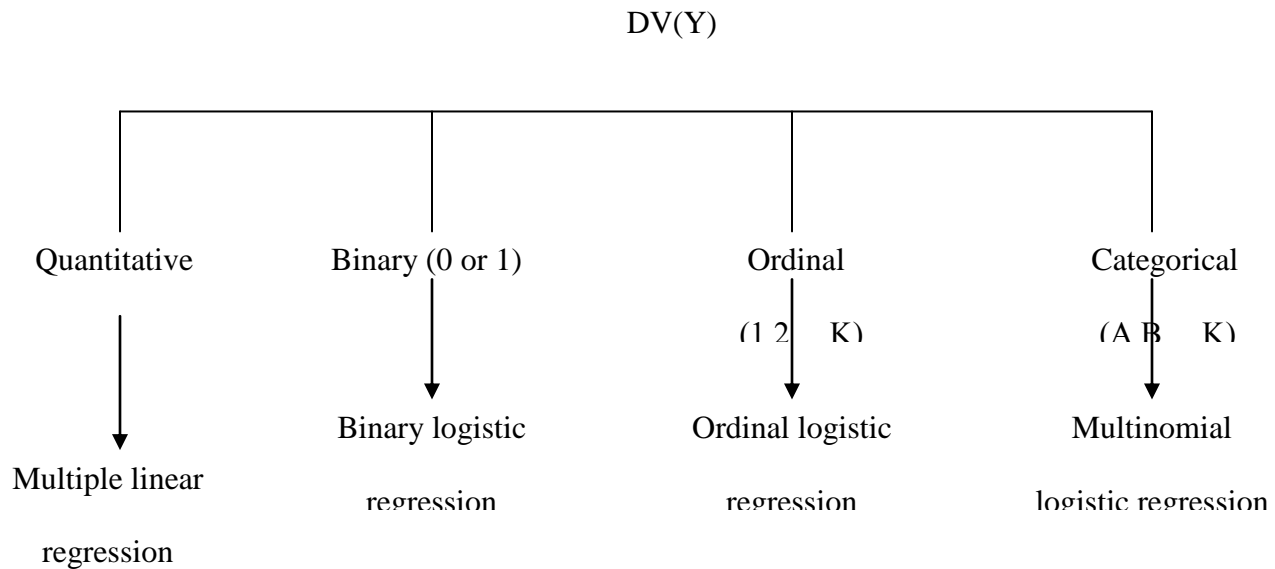
1-WAY ANOVA TABLE

Source	df	SS	MS	F	P
Treatment	k-1	SS_Trt	$\frac{SS_Trt}{(k-1)}$	$\frac{MS_Trt}{MSE}$	Shaded area
Error	N-k	SSE	$\frac{SSE}{(N-k)}$		
Total	N-1	TSS			

Reject null hypothesis of equal means (or no treatment effect) if $P < .05$.

If the null hypothesis of equal means is rejected, then run a multiple comparison test (e.g., Tukey's test).

Question involves finding a relationship between a dependent variable (DV) Y and 1 or more independent variables X_1, X_2, \dots, X_k .
(see handout on MLR)



- If some of the IV's (X 's) are binary or categorical, use DUMMY VARIABLE(S).
- If X_1 is binary (0 or 1), create one dummy DX_1 column.
- If X_2 is categorical with 4 categories (say), then create 4-1=3 dummy column DX_A , DX_B , DX_C
 - $DX_A=1$ if $X_2=A$, 0 otherwise
 - $DX_B=1$ if $X_2=B$, 0 otherwise
 - $DX_C=1$ if $X_2=C$, 0 otherwise

Chi-Square Tests for goodness of Fit, Independence and Homogeneity of

Proportions

The chi-square test is used for categorical (qualitative) data. In each application of the chi-square method, the following steps are used:
(input data = observed frequencies of various events)

- 1) Write down the null hypothesis for the problem.
- 2) Calculate the frequencies of each cell, assuming the null hypothesis to be true.
- 3) For each cell, calculate $\frac{(O-E)^2}{E}$, which is the chi-square distance between the observed and expected frequencies.
- 4) Sum all the chi-square distances for all of the cells, χ^2_{obs} .
- 5) Calculate the degree of freedom df of the chi-square test (depends on the problem at hand).
- 6) Calculate P-value as the probability that the chi-square random variable exceeds χ^2_{obs} ; if P-value $> .05$, do not reject the null and if $P < .05$, then reject the null hypothesis.

An example of each application is given here.

Example 1 (Testing goodness of fit, discrete or integer-valued data)

A die rolled 360 time, with following results:

x	1	2	3	4	5	6
O_x	76	59	58	47	50	70

O_x = observed frequency of the number x (x = 1, 2, ..., 6)

H_0 : Die is fair

Assuming the die is fair (null true), each expected frequency is

$$E_x = 360 \times (1/6) = 60, x = 1, 2, \dots, 6$$

$$\chi_{\text{obs}}^2 = \frac{(76-60)^2}{60} + \frac{(59-60)^2}{60} + \frac{(58-60)^2}{60} + \frac{(47-60)^2}{60} + \frac{(50-60)^2}{60} + \frac{(70-60)^2}{60} = 10.5$$

$$df = \#(\text{cells}) - \#(\text{parameters estimated}) - 1 = 6 - 0 - 1 = 5$$

$$P = P(\text{chi-square with 5 df} > 10.5) = 1 - .94 = 0.06 > 0.05 \text{ (Die is fair).}$$

Example 2: (Testing independence) For the following data, test if Age and GAME PREFERENCE are independent.

GAME	21-25	26-50	Over 50
Multi-Line Slots	15	37	16
Video Poker	25	25	17
Wheel of Fortune	14	40	27
Sports Book	11	4	1
Blackjack	9	23	14
Megabucks	3	8	1

Assuming H_0 : row and column attributes are independent,

$$P(\text{Multi-line slot AND 21-25 Age}) = P(\text{Multi-line slot}) \times P(\text{21-25 Age}) \\ = (68/200) \times (77/200)$$

Hence the expected value of this cell = $200 \times (68/200) \times (77/200)$

In general, $E_{ij} = (\text{Row } i \text{ total}) \times (\text{Row } j \text{ total}) / \text{Grand total}$

χ^2_{obs} is calculated by adding the individual chi-square values (calculated similarly for each of 18 cells). H_0 : GAME TYPE and AGE GROUP are independent is rejected since $P < .001$ (see next page).

Expected counts are printed below observed counts

	21 – 25	26 – 50	over 50	Total
1	15 18.06	37 32.12	16 17.82	68
2	25 17.79	25 31.65	17 17.56	67
3	14 21.51	40 38.27	27 21.23	81
4	11 4.25	4 7.56	1 4.19	16
5	9 12.21	23 21.73	14 12.06	46
6	3 3.19	8 5.67	1 3.14	12
Total	77	137	76	290

$$\begin{aligned} \text{Chi-Sq} = & 0.517 + 0.740 + 0.186 + \\ & 2.922 + 1.398 + 0.018 + \\ & 2.620 + 0.079 + 1.570 + \\ & 10.730 + 1.675 + 2.432 + \\ & 0.846 + 0.074 + 0.314 + \\ & 0.011 + 0.959 + 1.463 = 28.553 \end{aligned}$$

DF = 10, P-Value = 0.001

4 cells with expected counts less than 5.0

Example 3 (homogeneity of proportions)

In a telephone survey, respondents were asked to indicate their level of agreement with the statement “Cigarette smoking should be banned in public places”. The results are shown in the table below: SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree.

	SA	A	N	D	SD	TOTAL
F	40	38	16	37	5	136
M	16	25	11	25	11	88
TOTAL	56	63	27	62	16	224

Test if there is no difference in Males and Females with respect to their levels of agreement on the banning of smoking in public places. Let $\alpha = 0.05$.

$$H_0: p_{Fi} = p_{Mi}, \quad i = 1, 2, 3, 4, 5$$

$$H_1: H_0 \text{ is false}$$

Ch-square statistics is calculated from the formula of Example 2 above:

$$E_{ij} = (\text{Row } i \text{ total}) \times (\text{Row } j \text{ total}) / \text{Grand total}$$

Chi-Square Test:

Expected counts are printed below observed counts

	SA	A	N	D	SD	Total
1	40	38	16	37	5	136
	34.00	38.25	16.39	37.64	9.71	
2	16	25	11	25	11	88
	22.00	24.75	10.61	24.36	6.29	
Total	56	63	27	62	16	224

$$\text{Chi-Sq} = 1.059 + 0.002 + 0.009 + 0.011 + 2.288 + 1.636 + 0.003 + 0.015 + 0.017 + 3.536 = 8.575$$

$$\text{DF} = 4, \text{ P-Value} = 0.073$$

Since P-value = .073 > .05, the null hypothesis of equal proportions in the two genders is not rejected.

Examples 1 – 3 can be done with R as follows.

Example 1:

In R:

```
chisq.test(c(76, 59, 58, 47, 50, 70), p=c(1/6,1/6,1/6,1/6,1/6,1/6))
```

will give the following output:

Chi-squared test for given probabilities

data: c(76, 59, 58, 47, 50, 70)

X-squared = 10.5, df = 5, p-value = 0.06225

Example 2:

In R:

```
# read data
```

```
xx <- read.csv("G:/TEACH/DataMining_Fall2009/Data/chisqr_example2.csv", header=TRUE)
```

```
# run chi-square test
```

```
chisq.test(xx)
```

will produce the following output in R.

Pearson's Chi-squared test

data: xx

X-squared = 28.5528, df = 10, p-value = 0.001471

Warning message:

In chisq.test(xx) : Chi-squared approximation may be incorrect

Example 3:

```
xx <- read.csv("G:/TEACH/DataMining_Fall2009/Data/chisqr_example3.csv", header=TRUE)
```

OUTPUT with R: X-squared = 8.5748, df = 4, p-value = 0.07265