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 CATEGORICAL 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)

Use the table below to select appropriate statistical tool(s).

STATS24x7.com© 2010 ADI-NV, INC.

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	If $np_i \geq 5$, $n(1-p_i) \geq 5$, $i = 1, 2$ use normal approximation, else exact binomial distribution. **(page 8)**
		Case 2: Paired sample If $D = X_1 - X_2$ is normally distributed, use 1-sample t on D	
	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 $\{(x_1, y_1), ..., (x_n, y_n)\}$	y = binary (0/1) Fit a logistic equation to data $\{(x_1, y_1), ..., (x_n, y_n)\}$
	ONE RESPONSE VARIABLE, SEVERAL PREDICTORS	Fit a MLR equation to data $\{(x_{1j}, x_2, ..., x_k, y_j), j = 1, 2, ..., n\}$	y = binary (0/1) Fit a LOGISTIC regression equation to data $\{(x_{1j}, x_2, ..., x_k, y_j), j = 1, 2, ..., n\}$

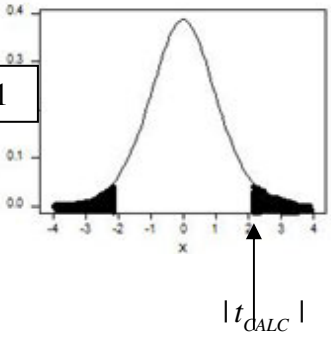
STATS24x7.com© 2010 ADI-NV, INC.

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

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

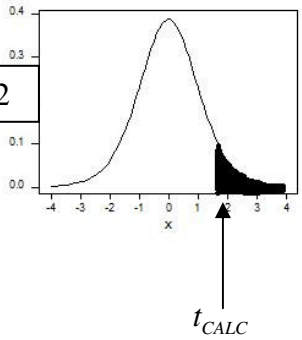
1) Test of Test $H_0 : \mu = \mu_0$ *vs.* H_1 $\quad t_{CALC} = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

Figure 1



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

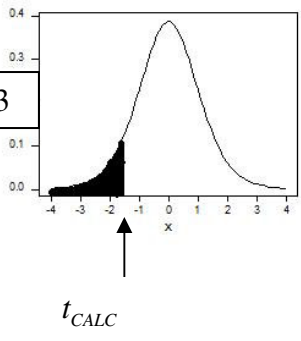
Figure 2



Reject

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

STATS24x7.com© 2010 ADI-NV, INC.

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 $\left(\sigma_1^2 = \sigma_2^2\right)$ – 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: $\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{n_1+n_2-2}\, s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

- 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 $n_1 + n_2 - 2$.

STATS24x7.com© 2010 ADI-NV, INC.

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

Case 2: Paired Samples

$\left(x_1, y_1\right), \left(x_2, y_2\right), ..., \left(x_n, y_n\right)$, $\bar{d} = \sum_{i=1}^{n} \frac{d_i}{n} =$ 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 $d_i = x_i - y_i$

STATS24x7.com© 2010 ADI-NV, INC.

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-α).

Example: For 95% confidence, 100(1-α)=95,

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

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

STATS24x7.com© 2010 ADI-NV, INC.

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.

STATS24x7.com© 2010 ADI-NV, INC.

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

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

Test of $H_0 : p_1 = p_2$

Calculate

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

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

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

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

STATS24x7.com© 2010 ADI-NV, INC.

Details of ANOVA

Given: k independent samples

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

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

Treatment 1 Treatment 2 Treatment k

Assume:

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

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 gramd 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:

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

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

We can show that

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

or

$TSS=SS\_Treatment+SS\_Error$

degrees of freedom are:

df(TSS) = N - 1, N = $n_1+n_2+...+n_k$

df(Treatment) = k -1

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

$$MS\_Treatment=\frac{SS\_Treatment}{k-1}$$

$$MSE=\frac{SS\_Error}{N-k}$$

$$F_{calc}=\frac{MS\_Treatment}{MSE}$$

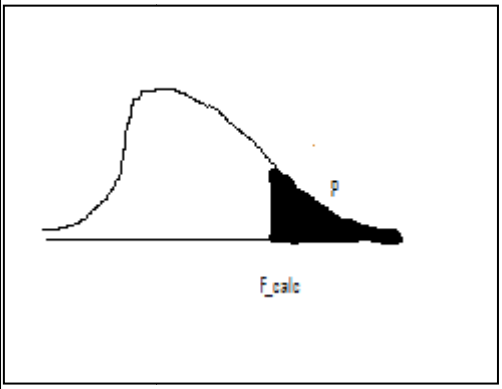
Distribution of F_{calc} when null is true is F with degrees of fredom

df(Num)=k -1

$df(den)=N-k$

Reject if F_{calc} is larger that 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).

STATS24x7.com© 2010 ADI-NV, INC.