STA501. Flowchart with formulae

1		Is the data?	
	Quantitative	(NOTE: If σ given, must use **z**, not **t**)	go to 2
	Qualitative		go to 18

2	Quantitative	How many samples are there ?	
		Population mean – One	goto 3
		Population mean – Two	goto 8
		Population mean – Many	goto 14
		Relationship between two quantitative variables	goto 24
3	Quantitative One Sample	Is n >= 30?	
		Yes	goto 4
		No	goto 5
4	Quantitative One Sample N >= 30	Confidence Interval $\bar{x} \pm Z_{\alpha/2}$ x $\frac{s}{\sqrt{n}}$ For interpretation of Confidence Interval (CI)	goto 29
		Hypothesis Test $Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim N(0,1)$ $X \sim N(\mu, \sigma^2)$	

5	Quantitative One Sample	Is data normally distributed?	(Use QQ Plot)
	N < 30	Yes	goto 7
		No	goto 6
6	Quantitative One Sample N < 30 Data not normally distributed	Assumption not met.	
		state that assumptions not met and therefore results not reliable /accurate)	goto 7
7	Quantitative One Sample N < 30 Data normally distributed	Confidence Interval $x \pm t_{n-1,\frac{\alpha}{2}}$ x $\frac{s}{\sqrt{n}}$ Hypothesis Test $t = \frac{x-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$	

8	Quantitative Two samples	Is data paired or independent?	
		Paired Calculate difference of means, and treat as single sample	goto 3
		Note : paired data : data is related / dependent *Different sample size indicates independent samples.*	
		Independent	goto 9
9	Quantitative Two samples Independent	Are both samples >= 30	
		Yes	goto 10
		No	goto 11
10	Quantitative Two samples Independent Both samples >= 30	Confidence Interval $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ Hypothesis Test $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	

11	Quantitative Two samples Independent One / Both Samples n < 30	Test if Variances are equal $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 = \sigma_2^2$ $F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}$ $(biggest\ on\ top, smallest\ on\ bottom)$ $Reject\ H_0\ if\ F_{obs} > F_{n_1-1,n_2-1,\alpha/2}$ Are variances equal?	
		Yes	goto 12
		No	goto 13
12	Quantitative Two samples Independent One / Both samples n < 30 Variances equal	Confidence Interval $(\bar{x}_1 - \bar{x}_2) \pm t_{df,\frac{\alpha}{2}} \sqrt{s_p^2(\frac{1}{n} + \frac{1}{n})}$ df = (n1 + n2 - 2) Hypothesis Test $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2(\frac{1}{n} + \frac{1}{n})}} \sim t_{n_1+n_2-2}$ $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$	
13	Quantitative Two samples Independent One / Both Samples n < 30 Variances NOT equal	Confidence Interval $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ df = min (n1 – 1 , n2 – 1) Hypothesis Test $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2(\frac{1}{n} + \frac{1}{n})}} \sim t_{\min(df)}$	

14	Quantitative Many Samples	ANOVA Test Test if ANOVA Test assumptions met:- Independent Simple Random Sample (SRS) Parent populations are normally distributed All populations have same standard deviation Results are approximately correct if largest sample std dev is no more than twice as large as smallest sample std dev (or side by side boxplots) Variability Between Samples $F = \frac{MSG}{MSE} \sim F_{dfG,dfE}$ Reject H0 if F_{obs} □ F_{dfG, dfE,} □ Hypothesis Testing $H_0 : \mu_1 = \mu_2 = \mu_3$ $H_1 : not\ all\ \mu' sare\ equal$ Is the NULL Hypothesis Rejected?	use QQ Plots / boxplots practical rule
		Yes	goto 15
		No	*no further test required*
15	Quantitative Many Samples not all μ's equal	Use Bonferroni Method valid only if Ho rejected not all μ's are equal Post-hoc procedure.	

18	Qualitative	Is there one or two samples?	
		Population proportion - One sample	goto 19
		Population proportion - Two samples	goto 21
		Relationship between 2 categorical variables	goto 23a
		One qualitative variable with many levels – Goodness of Fit	goto 23b
19	Qualitative One sample	$np > 5$ $n(1-p) > 5$	
		Yes	goto 20
		No	got 20, *NOTE: results not reliable*
20	Qualitative One sample np and nq > 5	Confidence Interval $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ Hypothesis Test $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$	
21	Qualitative Two samples	Both Samples $np > 5$ $n(1-p) > 5$	
		Yes	goto 22
		No	goto 22. *NOTE results not reliable*
22	Qualitative Two samples np and nq > 5	Confidence Interval $(\hat{p}_1 - \hat{p}_2) \pm Z^* \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}$ $\hat{p} = count\ of\ success\ in\ \ both\ samples / total\ count\ in\ both\ samples$ Hypothesis Test $Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$	

23a — Chi square test of independence

$H_0 : variables\ are\ NOT\ related\ and\ are\ independent$
$H_1 : variables\ are\ related\ and\ dependent$

$$X^2 = \sum \frac{(Observed\ Value)^2}{Expected\ Value} - N \sim X^2_{df}$$

$$df = (rows - 1)\ \text{x}\ (columns - 1)$$

$$Reject\ H_0\ if\ X^2_{obs} > X^2_{df,\alpha}$$

Steps For Chi-Square Test

Table 1 : Observed Values
Table 2 : Expected Values (E)

$$Expected\ Value = \frac{(Row\ Total)\ x\ (Column\ Total)}{Grand\ Total}$$

$$Expected\ Value = \frac{Row\ Total}{Grand\ Total}\ x\ \frac{Column\ Total}{Grand\ Total}\ x\ Grand\ Total$$

Assumptions

- No table cell has an Expected Value less than 1
- No more than 20% of cells have expected values less than 5
- Outcome of each trial can only fall into one of the categories

23b — Chi square test goodness of fit

Does data match a pattern?

$H_0 : observed\ value\ DOES\ match\ expected\ value$
$H_1 : observed\ value\ DOES\ NOT\ match\ expected\ value$

$$X^2 = \sum \frac{(Observed\ Value)^2}{Expected\ Value} - N \sim X^2_{df}$$

$$df = n - number\ of\ assumptions$$
$$df = n - 1$$

$$Reject\ H_0\ if\ X^2_{obs} > X^2_{df,\alpha}$$

Steps For Chi-Square Test

Table 1 : Observed Values
Table 2 : Expected Values (E) apply pattern to N to determine this???

Assumptions

- No table cell has an Expected Value less than 1
- No more than 20% of cells have expected values less than 5
- Outcome of each trial can only fall into one of the categories

24	Correlation	goto 25
	Regression	goto 26
25	Correlation	$H_0 : \rho = 0 \ (linear\ relationship - does\ not\ exist)$ $H_1 : \rho \neq 0 \ (linear\ relarionship - does\ exist)$ $t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} \sim t_{n-2}$ $df = n - 2 \ (n = total\ number\ of\ points \mid pairs \mid y\ values)$ Coefficient of Determination (CoD) $R = r^2$ Coefficient of Determination (CoD) tells us that • x% of the variation in y is • explained by the linear regression on x.
26	Regression	$\hat{y} - a + bx$ • b : the slope of the estimated regression line ○ The addition of one unit of x is associated with a change of b units of y on average. • A : the intercept of the estimated regression line ○ When x is zero, y is a. /gives the minimum y value. Using regression equation for estimation – should be within range of existing data. Confidence Interval $b \pm t_{n-2,\alpha/2}\, SE_b$ Hypothesis Test $H_0 : \beta = 0 \ (slope\ is\ zero)$ $H_1 : \beta \neq 0 \ (slope\ is\ NOT\ zero)$ $(linear\ relationship\ between\ the\ two\ variables)$ $t_{obs} = \frac{b}{SE_b} \sim t_{n-2}$ SEb - does not need to be calculated This value will always be given (prob as part of S-Plus output) $tolerance = 1 - r^2$ $tolerance < 0.1, \quad suggests\ variables\ are\ colinear$ *Used for determining best model* $Adjusted\ R^2 = 1 - (1 - Multiple\ R^2)\frac{n - 1}{n - p - 1}$

27	Hypothesis Template	1. Write null and alternative hypothesis ○ H_0: $does\ NOT\ have\ any\ affect$ ○ H_1 : $DOES\ have\ an\ affect$ 2. State level of significance ○ $\alpha = Level\ of\ significance$ ○ $CI = (100 - \alpha)\ \ Confidence\ Interval$ 3. State the sampling distribution / test statistic ○ ***Write down Actual Formula / Equation*** ○ ***Write down Name of Formula / Equation*** 4. Decision Rule: Find Critical region (use diagram) *Examples* ○ t_{cv} *OR* ○ Reject H0 if p-value < alpha 5. Calculate ○ Test statistic (*results from using formula*) ○ Observed value 6. Decision $(H_0 : Reject\ or\ Accept)$ ○ Explicitly compare observed value with critical value *OR* ○ Explicitly compare *p-value* and *alpha* 7. Conclusion ○ Interpretation of test result ○ Answer in terms of original problem.
28	ANOVA	**1** For each sample sample size n — also N Total T — also Grand Total Total^2 — T^2 T^2 / n — also Total **2** Calculated sum of al X^2 — (ΣX^2) **3** $SST = (\sum X^2)$ — (Grand Total)^2 / N $SSG = \sum T^2/n$ — (Grand Total)^2 / N $SSE = SST - SSG$ **4** $df\ treatment = number\ of\ samples - 1$ $df\ total = N - 1$ $df\ error = df\ total - df\ treatment$
29	CI	The true population mean will be contained in the xx% confidence interval, xx%of the time.