Unit 7 Formulas/Info you can use on the test

Hypothesis tests
1) Hypotheses
2) Conditions
3) Calculate p-value
4) Conclusion paragraph:
With significance of .05, p-value=0.02 is low so we reject Ho.
We do have sufficient statistical evidence to conclude (Ha).

Confidence Intervals
1) Conditions
2) Calculate confidence interval
3) Conclusion sentence:
We are 90% confident that the true difference in SAT scores (after-before) is between 25 and 36 Points, on average.

Regression LSRL:

$$\hat{y} = a + bx \qquad x: \textit{defined w/units}$$
$$y: \textit{defined w/units}$$

Regression example wording:

Slope, b: For every 1 additional inch in height, the number of steps decrease by 0.573 steps, on average.

Intercept, a: A person who is zero inches tall is predicted to take 53.8 steps, on average.

Correlation coefficient, r: There is a linear, negative, strong relationship between steps and height.

Coefficient of determination, r^2: About 76% of the variation in number of steps is explained by the LSRL which relates number of steps to height.

Standard deviation of residuals, s: The average difference between actual number of steps and predicted number of steps (for each given height) is 1.58 steps.

Standard deviation of slopes, s_b (for inference): If we took many samples and computed LSRLs for each, these LSRLs would each have a slope b. The standard deviation of these slopes would be 3.4 steps/inch.

Common z* values: 90%: z*=1.64, 95%: z*=1.96, 99%: z*=2.576

Success/Fail? Percentages? **Inference for Proportions** 1 Proportion — 2 Proportions Z-statistics Normal distributions (no df)	Means of numbers? **Inference for Means** **1 Mean**: df = n - 1 **2 Means**: **2 Sample** — **Diff. of means**: df = TI calc **Matched Pair** — **Mean of diffs.**: df = n - 1 t-statistics, t distributions or if n>25: Z-statistics, Normal distributions	Bivariate (y vs. x) data? **Inference for Regression LSRL Slope** t-distributions df = n - 2 t-statistic: $t = \frac{b - \beta}{s_b}$ (b = statistic, β = parameter) S_b = standard error of slope S = standard error of residuals $usually\ \beta_0 = 0, so\ t = \frac{b}{s_b}$	Counts? **Inference for Counts** χ^2 - statistics χ^2 distributions **1 col (or row)** (compared to expected %): **Goodness of Fit** df = #categories - 1 $\chi^2 = \sum \frac{(obs - exp)^2}{exp}$ **>1 col (or row)**: **1 population**: **Independence** **>1 population**: **Homogeneity** df = (#rows - 1)(#cols - 1) expected cell count = $\frac{(row\ total)(col\ total)}{grand\ total}$
Hypotheses: 1 proportion: 1PropZTest/Int $H_0: p = p_0$ $H_A: p > p_0 (or <, \neq)$ 2 proportions: 2PropZTest/Int $H_0: p_1 = p_2 (p_1 - p_2 = 0)$ $H_A: p_1 > p_2 (p_1 - p_2 > 0)(or <, \neq)$	1 mean: T-Test/T-Interval $H_0: \mu = \mu_0$ $H_A: \mu > \mu_0 (or <, \neq)$ 2 mean (independent): 2SampTTest/Int $H_0: \mu_1 = \mu_2 (\mu_1 - \mu_2 = 0)$ $H_A: \mu_1 > \mu_2 (\mu_1 - \mu_2 > 0)(or <, \neq)$ 2 mean (matched pairs): TTest/Int on diffs $H_0: \mu_D = 0$ $H_A: \mu_D > 0 (or <, \neq)$ $\mu_D = mean\ of\ diffs$	slope: LinRegTTest/Int $H_0: \beta = 0 (no\ association)$ $H_A: \beta \neq 0 (or <, >)(association)$ $CI: b \pm (t^*)(s_b)$	GOF: X^2GOF-test (obs in L1, exp in L2) H_0: Observed distribution of counts same as expected. H_A: Observed distribution of counts not same as expected. Independence: X^2-Test (2D data in matrix A) H_0: Row and column variables are independent. H_A: Row and column variables are not independent. Homogeneity: X^2-Test (2D data in matrix A) H_0: The distribution of ___ is the same among all populations. H_A: The distribution of ___ is not the same among all populations
Conditions: 1 proportion: SRS, n<10%pop, success/fail >10 2 proportions: For each group... SRS, n<10%pop, success/fail >10 Groups independent of each other	1 mean: SRS, n<10%pop, Nearly Normal 2 means (indep): Groups independent For each group... SRS, n<10%pop, Nearly Normal 2 means (matched): How matched? SRS, n<10%pop, diffs are Nearly Normal	Straight enough Residuals show no pattern or fanning Residuals are Nearly Normal	All cell expected counts are > 5 - or - 80% of cells' expected counts are > 5 and none of the expected counts are 0

Sampling distributions for proportions:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For one population:			
$\hat{p}$	$\mu_{\hat{p}} = p$	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$	$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
For two populations:			
$\hat{p}_1 - \hat{p}_2$	$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$	$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$s_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ *When* $p_1 = p_2$ *is assumed*: $s_{\hat{p}_1-\hat{p}_2} = \sqrt{\hat{p}_C(1-\hat{p}_C)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ *where* $\hat{p}_C = \frac{X_1 + X_2}{n_1 + n_2}$

Sampling distributions for means:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For one population:			
$\overline{X}$	$\mu_{\overline{X}} = \mu$	$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$	$s_{\overline{X}} = \frac{s}{\sqrt{n}}$
For two populations:			
$\overline{X_1} - \overline{X_2}$	$\mu_{\overline{X_1}-\overline{X_2}} = \mu_1 - \mu_2$	$\sigma_{\overline{X_1}-\overline{X_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$s_{\overline{X_1}-\overline{X_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

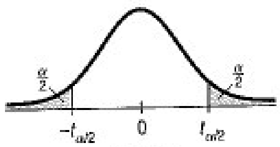
Sampling distributions for regression:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For slope:			
b	$\mu_b = \beta$	$\sigma_b = \frac{\sigma}{\sigma_x \sqrt{n}}$ where $\sigma_x = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$	$s_b = \frac{s}{s_x \sqrt{n-1}}$ where $s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$ and $s_x = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}}$

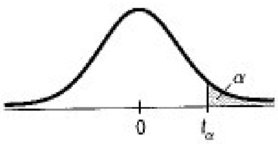
* Standard deviation is a measure of variability from the theoretical population. Standard error is the estimate of the standard deviation. If the standard deviation of the statistic is assumed to be known, then the standard deviation should be used instead of the standard error.

Table T

Values of t_α



Two tails



One tail

Two tail probability	0.20	0.10	0.05	0.02	0.01	
One tail probability	0.10	0.05	0.025	0.01	0.005	
df						df
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
30	1.310	1.697	2.042	2.457	2.750	30
32	1.309	1.694	2.037	2.449	2.738	32
35	1.306	1.690	2.030	2.438	2.725	35
40	1.303	1.684	2.021	2.423	2.704	40
45	1.301	1.679	2.014	2.412	2.690	45
50	1.299	1.676	2.009	2.403	2.678	50
60	1.296	1.671	2.000	2.390	2.660	60
75	1.293	1.665	1.992	2.377	2.643	75
100	1.290	1.660	1.984	2.364	2.626	100
120	1.289	1.658	1.980	2.358	2.617	120
140	1.288	1.656	1.977	2.353	2.611	140
180	1.286	1.653	1.973	2.347	2.603	180
250	1.285	1.651	1.969	2.341	2.596	250
400	1.284	1.649	1.966	2.336	2.588	400
1000	1.282	1.646	1.962	2.330	2.581	1000
∞	1.282	1.645	1.960	2.326	2.576	∞
Confidence levels	80%	90%	95%	98%	99%	