Review for the Final Exam (STA248)

1. (a) Review the assignment questions, the midterm test, in class examples, homework problems;

   (b) Go over the review for the midterm.

One-Sample Case:

CI for proportion p: $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$

CI for μ when σ is known: $\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

CI for μ when σ is unknown: $\bar{X} \pm t_{\alpha/2} \cdot s/\sqrt{n}$

CI for σ^2: $\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}$

Tests for p: $H_0: p = p_0$ use statistic $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, where $\hat{p}$ is the sample proportion of successes

Test for μ (σ is known): $H_0: \mu = \mu_0$ use statistic $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

Test for μ (σ is unknown): $H_0: \mu = \mu_0$ use statistic $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$

(review: p-value, power, type I and II errors)

Test for σ^2: $H_0: \sigma^2 = \sigma_0^2$ use statistic $X = \frac{(n-1)s^2}{\sigma_0^2}$

Paired t-test: one-sample t-test performed on the differences

Non-parametric tests: sign test, the Wilcoxon signed-rank test

Two-Sample Case:

CI for $p_1 - p_2$: $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

CI for $\mu_1 - \mu_2$ (σ_1 and σ_2 are unknown and unequal): $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

CI for $\mu_1 - \mu_2$ (σ_1 and σ_2 are unknown and equal):
$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}} * s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, where $s_p = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

Test for proportions: $H_0: p_1 = p_2$ use statistic $z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}_{pooled}(1-\hat{p}_{pooled})(\frac{1}{n_1}+\frac{1}{n_2})}}$, where $\hat{p}_{pooled} = \frac{X_1+X_2}{n_1+n_2}$

Test for $\mu_1 - \mu_2$ (σ_1 and σ_2 are known): $H_0: \mu_1 = \mu_2$ use statistic $z = \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$

Test for $\mu_1 - \mu_2$ (σ_1 and σ_2 are unknown and unequal): $H_0: \mu_1 = \mu_2$ use statistic $t = \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ (use t_k distribution where k is either approximated by software or $\min(n_1 - 1, n_2 - 1)$

Test for $\mu_1 - \mu_2$ (σ_1 and σ_2 are unknown and equal σ): $H_0: \mu_1 = \mu_2$ use statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Test for variances: $H_0: \sigma_1^2 = \sigma_2^2$ use statistic $F = \frac{s_1^2}{s_2^2}$

Non-parametric tests: the Wilcoxon rank sum test

Also review: One-way ANOVA
Two-way ANOVA
Simple Linear Regression

2. Given the following partial software output:

Source	df	SS
Between	6	20.2
Within	*	*
Total	88	110.5

$SS_E = SS_T - SS_G$

$= 110.5 - 20.2$

$= 90.3$

(a) What is the value of F statistic?

$$F = \frac{MS_G}{MS_E} = \frac{SS_G/df_G}{SS_E/df_E} = \frac{20.2/6}{90.3/82} = 3.71$$

(b) What decision would be made regarding H_0: population means are equal?

$$P\text{-value} = P(X \geq 3.71) < 0.01$$

$$X \sim F(6, 82)$$

Software $= 0.00259$

$\Rightarrow$ reject H_0.

3. **Jury Selection.** One study of grand juries in Alameda County, California, compared the demographic characteristics of jurors with the general population, to see if jury panels were representative. The results for age are shown below. The investigators wanted to know if the 66 jurors were selected at random from the population of Alameda County. (Only persons over 21 and over are considered; the county age distribution is known from Public Health Department data.) The study was published in the UCLA Law Review.

Age	County-wide %	# of jurors observed	# of jurors expected	(O-E)	$(O-E)^2/E$
21-40	42%	5	27.72	-22.72	
41-50	23%	9	15.18	-6.18	
51-60	16%	19	10.56	8.44	
over 60	19%	33	12.54	20.46	
Total	100%	66			

Do we have evidence that grand juries are selected at random for the population of Alameda County?

H_0: For each age group, the proportion of jurors is consistent with the county proportion.

H_a: at least one is not

$$\chi^2 = \frac{(-22.72)^2}{27.72} + \frac{(-6.18)^2}{15.18} + \cdots$$

$$= 61.27$$

P-value $P(\chi \geq 61.27) < 0.001$

$\chi \sim \chi^2_3$ $\Rightarrow$ reject H_0

4. **Decay of polyester fabrics in landfills.** How quickly do synthetics such as polyester decay in landfills? A researcher buried polyester strips in the soil for different lengths of time, then dug up the strips and measured the force required to break them. Breaking strength is easy to measure and is a good indicator of decay. Lower strength means the fabric has decayed. Part of the study involved burying 10 polyester strips in well drained soil in the summer. Five of the strips, chosen at random, were dug up after 2 weeks; the other 5 were dug up after 16 weeks. Here are the breaking strengths in pounds:

2 weeks	118	126	128	120	129
16 weeks	124	98	110	140	109

Is there evidence that breaking strengths are lower for strips buried longer?

1 2 3 4 5 6 7 8 9 10

98 109 110 118 120 124 126 128 129 140

16 weeks

$W = 1+2+3+6+10 = 22$

$\mu_W = \frac{n_1(N+1)}{2} = \frac{5(10+1)}{2} = 27.5$

$\sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}} = \sqrt{\frac{5 \cdot 5 \cdot 11}{12}} = 4.8$

$P\text{-value} = P(W \le 22) = P\left(Z \le \frac{22.5 - 27.5}{4.8}\right)$

↑ continuity correction

$= P(Z \le -1.04) = 0.1492$

$\Rightarrow$ fail to reject H_0

5. Below are scores for 24 students who took the same final examination, but who are from the groups in which were used different teaching techniques

Group I	Group II	Group III	Group IV
81	55	63	51
90	60	53	67
69	61	64	80
66	72	67	70
81	39	56	68
75	85	70	67
462	372	373	403

Suppose SS(total) = 700, and SS(between groups) = 200.

(a) What is the observed value of the statistic one computes to test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ against H_a: not all 4 means are equal?

(b) If α = .01, what is the critical value of the statistic?

(c) What your conclusion would be?

(a) $F = \frac{MS_G}{MS_E} = \frac{SS_G/df_G}{SS_E/df_E} = \frac{200/3}{500/20} = 2.66$ $\quad df_G = k-1$, $df_E = N-k$

(b) $F_{3,20,0.01} = 4.94$

(c) $2.66 < 4.94 \Rightarrow$ fail to reject H_0.

6. Every few years, the National Assessment of Educational Progress asks a national sample of eighth-graders to perform the same math tasks. The goal is to get an honest picture of progress in math. Here are the last few national mean scores:

Year	1990	1992	1996	2000	2003	2005
Score	263	268	272	273	278	279

Find the regression line of mean score on time. What percent of the year-to-year variation in scores is explained by the linear trend?

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = r \frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = 0.9739$$

$\bar{x} = 1997.6$ $\quad s_x = 6.0222$

$\bar{y} = 272.16$ $\quad s_y = 6.0470$

$$b_1 = 0.9739 \cdot \frac{6.0470}{6.0222} \simeq 0.98$$

$$b_0 = 272.16 - 0.98 \cdot 1997.6 \simeq -1681$$

$$\hat{y} = -1681 + 0.98x$$

$\rightarrow$ explains $r^2 = 95\%$ of the variation in scores.