Stat 371-003
Supplementary problems to help you prepare for the Final Exam
Solutions

- **2.55 (pg 50)**

  The distribution is pretty close to being symmetric, but has a slightly longer right tail. The mean should be close to the center, say around **100** or **105**.

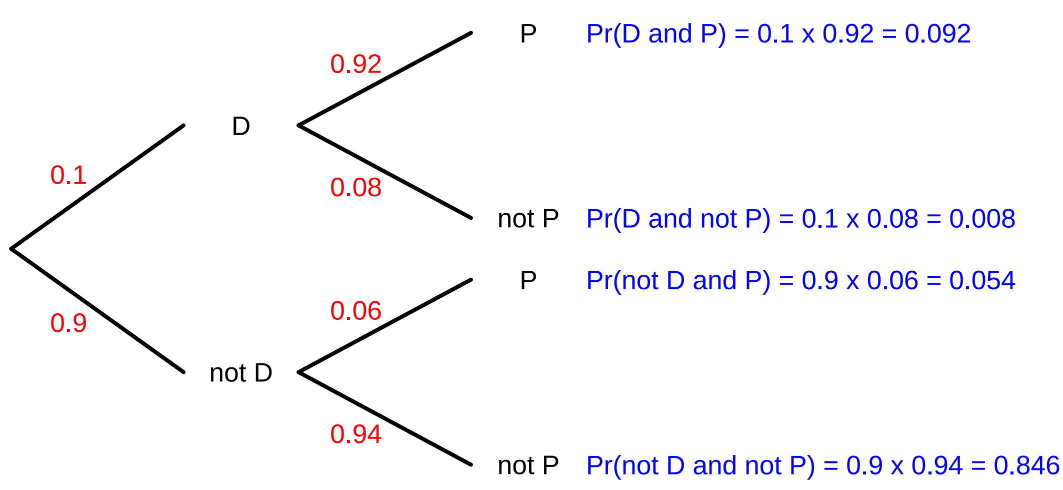
  The SD is the *typical deviation from the average*. The bulk of the data range from around 65 to around 150, and so the SD is about (150–65)/4 ≈ **20**.

- **3.11 (pg 88)**

  Let D = { person has disease } and P = { test is positive }.

  Then Pr(P | D) = 0.92, Pr(not P | not D) = 0.94, and Pr(D) = 0.1.

  We can make a probability tree, as follows.



  (a) Pr(P) = Pr(D and P) + Pr(D and not P) = 0.092 + 0.054 = 0.146

  (b) Pr(D | P) = Pr(D and P) / Pr(P) = 0.092 / 0.146 ≈ 0.63.

- **3.31 (pg 111)**

  Let X be the number of infants (out of four) that are male. The X follows a binomial distribution with n=4 and p = 105/(105+100) ≈ 0.5122.

(a) Pr(two are male and two are female) = Pr(X=2) = $\binom{4}{2}(0.5122)^2(1-0.5122)^2 = (4 \times 3/2)(0.5122)^2(0.4878)^2 \approx 0.375$

(b) Pr(all four are male) = Pr(X=4) = $(0.5122)^4 \approx 0.0688$.

(c) Pr(all four are the same sex) = Pr(X=4) + Pr(X=0) = $(0.5122)^4 + (0.4878)^4 \approx 0.125$.

- **4.13 (pg 133)**

  Let X be the amount of growth for a random plant, following a normal distribution with mean 3.18 cm and SD 0.53 cm.

  Let $Z = (X - 3.18)/0.53$

  (a) $\Pr(X \geq 4) = \Pr[Z \geq (4-3.18)/0.53] \approx \Pr(Z \geq 1.55) = 1 - \Pr(Z < 1.55) \approx 1 - 0.9394 = 0.0606$.

  (b) $\Pr(X \leq 3) = \Pr[Z \leq (3-3.18)/0.53] \approx \Pr(Z \leq -0.34) \approx 0.3669$

  (c) $\Pr(2.5 < X < 3.5) = \Pr[(2.5-3.18)/0.53 < Z < (3.5-3.18)/0.53] \approx \Pr(-1.28 < Z < 0.60) = Pr(Z < 0.60) - Pr(Z < -1.28) \approx 0.7257 - 0.1003 = 0.6254$

- **5.28 (pg 166)**

  Let $\bar{X}$ denote the sample average. It will be approximately normally distributed with mean 38 mm Hg and SD $9/\sqrt{25}$ = 1.8 mm Hg.

  Let $Z = (\bar{X} - 38)/1.8$.

  (a) $\Pr(\bar{X} > 36) = \Pr[Z > (36-38)/1.8] \approx \Pr(Z > -1.11) = 1 - \Pr(Z < -1.11) \approx 1 - 0.1335 = 0.8665$

  (b) $\Pr(\bar{X} > 41) = \Pr[Z > (41-38)/1.8] \approx \Pr(Z > 1.67) = 1 - \Pr(Z < 1.67) \approx 1 - 0.9525 = 0.0475$

- **6.12 (pg 194)**

  (a) The estimated standard error of the sample mean is $4.6/\sqrt{(6)} \approx 1.878$. The 97.5 percentile of the t distribution with 5 degrees of freedom is 2.571. Thus the 95% confidence interval for the population mean is

  $$28.7 \pm 2.571 \times 1.878 \approx 28.7 \pm 4.8 = (23.9, 33.5)$$

  (b) We are estimating the average blood serum concentration of Gentamicin in 3-year-old female Suffolk sheep 1.5 hours after injection.

  (c) For small sample sizes, this would be true, but for large sample sizes this would not be true.

- **7.12 (pg 232)**

  (a) The difference between the two sample means is $13.8 - 4.0 = 9.8$.

  The estimated standard error for the difference between the sample means is $\sqrt{(1.34)^2 + (1.30)^2} \approx 1.87$. (Note that they gave us SEs rather than SDs, to trick us.)

  The 97.5 percentile of a t distribution with 190 degrees of freedom is about 1.977 (refering to df=140 in the table).

  So the 95% confidence interval for the difference between the population means is

  $$9.8 \pm 1.977 \times 1.87 \approx 9.8 \pm 3.7 = (6.1, 13.5)$$

  (b) These are the plausible values for the average effect of eight weeks of biofeedback training on the reduction in systolic blood pressure.

- **7.13 (pg 232)**

  The confidence interval remains valid, as the sample sizes are quite large.

- **7.32 (pgs 245–246)**

  (a) The difference between the sample averages is $1.190 - 1.785 = -0.595$.

  The estimated standard error for the difference between the population averages is $\sqrt{(0.184)^2/4 + (0.241)^2/4} \approx 0.152$.

  The t statistic is $-0.595/0.152 \approx -3.92$.

  The critical value for a two-sided test is the 97.5 percentile of the t distribution with 5.6 degrees of freedom: approximately 2.447.

  Since $3.92 > 2.447$, we reject the null hypothesis.

  The P-value is between 0.001 and 0.01.

  (b) We conclude that flooding does affect ATP content.

- **10.12 (pg 401)**

  The total number of data points is $179 + 44 + 23 = 246$, and so the expected counts are as follows:

	Type		
	I	II	III
Observed	179	44	23
Expected	184.5	46.125	15.375

  The χ^2 statistic is

  $$\frac{(179 - 184.5)^2}{184.5} + \frac{(44 - 46.125)^2}{46.125} + \frac{(23 - 15.375)^2}{15.375} \approx 4.04$$

The critical value of the test is the 90th percente of the χ^2 distribution with 2 degrees of freedom: 4.61.

Since $4.04 < 4.61$, we fail to reject the null hypothesis. (The P-value is between 0.1 and 0.2.)

We conclude that the data could reasonably come from a population in the proportions 12:3:1.

- **10.24 (pg 411)**

  Note: I've never seen this "directional alternatives" for χ^2 tests outside of our textbook, and so on the final exam I'd ask you to consider only non-directional alternatives.

  The overall sample size is $653 + 1148 = 1801$. The proportions hip fractures in the hip protector and control groups are 2.0% and 5.8%, respectively. So the effect of the hip protector is in the correct direction.

  The row sums are $13 + 67 = 80$ and $640 + 1081 = 1721$.

  The expected count in the upper-left cell is $653 \times 80/1801 \approx 29.0$. The others are calculated similarly.

	Treatment	
	Hip protector	*Control*
Hip fracture	29	51
No hip fracture	624	1097

  The χ^2 statistic is

$$\frac{(13-29)^2}{29} + \frac{(67-51)^2}{51} + \frac{(640-624)^2}{624} + \frac{(1081-1097)^2}{1097} \approx 14.50$$

  We compare this to the χ^2 distribution with 1 degree of freedom. With $\alpha = 0.01$ and a directional alternative, we look at the tail probability of 0.02, and so the critical value is 5.41, and we clearly reject the null hypothesis. The p-value is between 0.00005 and 0.0005. (We divide the tail areas by two, since we are considering a directional alternative.)

  And so we conclude that hip protectors are indeed protective.

- **10.47 (pg 432)**

  (a) We compare the χ^2 statistic of 49.77 to the χ^2 distribution with 3 degrees of freedom (since there are 4 rows and 2 columns). We get a p-value of < 0.0001, and so we clearly reject the null hypothesis and conclude that the regions have different HSV-2 prevalences.

  (b) We arrange the observed data as follows.

Region	HSV-2	no HSV-2	Total
Northeast	323	1165	1488
Midwest	381	1689	2070
South	1320	4003	5323
West	712	1986	2698
Total	2736	8843	11579

The expected counts under the null hypothesis (of no difference in HSV-2 prevalences across the four regions) are calculated by taking, e.g., $1488 \times 2736/11579 \approx 351.6$ for the upper-left cell.

Region	HSV-2	no HSV-2	Total
Northeast	351.6	1136.4	1488
Midwest	489.1	1580.9	2070
South	1257.8	4065.2	5323
West	637.5	2060.5	2698
Total	2736	8843	11579

We calculate the χ^2 statistic as

$$\frac{(323-351.6)^2}{351.6}+\frac{(1165-1136.4)^2}{1136.4}+\frac{(381-489.1)^2}{489.1}+\cdots+\frac{(1986-2060.5)^2}{2060.5} \approx 49.77$$

- **11.3 (pgs 475–476)**

  (a) SS(between) = SS(total) – SS(within) = 338.769 – 116 = 222.769

  (b) The overall sample size is 13, and their are 3 groups, so df(between) = 2 and df(within) = 12 - 2 = 10.

  MS(between) = SS(between)/df(between) = 222.769/2 = 111.3845.

  MS(within) = SS(within)/df(within) = 116/10 = 11.6.

  $s_{\text{pooled}} = \sqrt{\text{MS(within)}} = \sqrt{11.6} \approx 3.41.$

- **11.12 (pg2 482–483)**

  (a) The dot plot indicates that the null hypothesis is not true; there looks to be variation in daffodil length across the five regions. In particular, the open area appears to have smaller lengths.

  (b) Let μ_N, μ_E, μ_S, μ_W and μ_O be the true average daffodil length for the five regions. The null hypothesis is that all five means are the same: $\mu_N = \mu_E = \mu_S = \mu_W = \mu_O$.

  (c) The ANOVA table is as follows.

Source	**df**	**SS**	**MS**
Between groups	4	871.41	217.85
Within groups	60	3588.54	59.81
Total	64	4459.95	

The test statistic is F = 217.85/59.81 = 3.64. Comparing this to an F distribution with 4 and 60 degrees of freedom, at $\alpha = 0.1$, the critical value is 2.04, and so we reject the null hypothesis and conclude that there are real differences in daffodil length across the five regions. The p-values is very close to 0.01.

- **11.20 (pg 498)**

  The degrees of freedom for the interaction is 1, and the "within" degrees of freedom is $53 + 57 + 55 + 58 - 4 = 219$.

  The MS(interaction) = SS(interaction) = 31.33.

  MS(within) = SS(within)/219 = 30648.81/219 = 139.95.

  The test statistic is F = 31.33/139.95 = 0.22, and we shouldn't have to even look at the table to know that we will fail to reject the null hypothesis (as we at least need F > 1): there is no evidence for an interaction.

  Looking at the F distribution with 1 and 219 degrees of freedom (we can look at 1 and 140 df, since 219 isn't in the table), we see that $P > 0.2$.

- **11.21 (pg 498)**

  (a) The ANOVA table is as follows.

Source	df	SS	MS
drug	1	69.22	69.22
dose	1	330.00	330.00
interaction	1	31.33	31.33
within	219	30648.81	139.95
Total	222	31079.36	

  (b) The wording here may confuse you. One might consider the "effect" of a drug to be the difference in the response between the high and low doses, in which case whether the drugs have the same effect is the same as looking for an interaction, as in problem 11.20.

  Alternatively, one might view the effect of a drug to be the response at a given dose, and so that the drugs having different effects means that the response (averaged across the two doses) is the different. It appears that the book meant this latter interpretation.

  Thus, we consider the "drug" row in the ANOVA table and calculate the F statistic as 69.22/139.95 = 0.49. As I mentioned in problem 11.20, with F < 1, it's not necessary to look at the table, as we will clearly fail to reject the null hypothesis.

  Nevertheless, we could look at the F table with 1 and 219 degrees of freedom (though we would have to look at 1 and 140 df, since that's what's in the table), and we would find $P > 0.2$.

- **12.9 (pgs 539–540)**

  (a) The estimated slope is $\hat{\beta}_1 = \sum(x_i - \bar{x})(y_i - \bar{y})/\sum(x_i - \bar{x})^2 = 731.36/2094.55 = 0.349$.

  The estimated y-intercept is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 103.99 - 0.349 \times 149.64 = 51.77$.

  Thus the regression line (of Y on X) is $\hat{y} = 51.77 + 0.349x$.

  (b) With each 1 mm increase in frog length, the average maximum jump increases by the amount β_1.

  (c) $s_Y = \sqrt{\sum(y - \bar{y})^2/(n-1)} = \sqrt{3218.99/10} \approx 17.9$.
  $s_{Y|X} = \sqrt{\text{SS(resid)}/(n-2)} = \sqrt{2963.61/9} \approx 18.1$

  (d) $s_{Y|X}$ is the SD of maximum jumps among frogs with a fixed length.

  Alternatively, if we think of predicting maximum jump length for a frog, based on its length, using the regression line, $s_{Y|X}$ could be viewed as the typical error in the prediction. (That $s_{Y|X} > s_Y$ indicates that the regression line is not very useful for predicting maximum jump.)

- **12.17 (pg 548)**

  We would estimate the maximum jump length for a frog who is 150 mm long as $51.77 + 0.349 \times 150 = 104.12$ mm.

- **12.24 (pg 553)**

  Referring back to the data for problem 12.8 (pg 539), the estimated slope is $\hat{\beta}_1 = \sum(x_i - \bar{x})(y_i - \bar{y})/\sum(x_i - \bar{x})^2 = 161.40/50667 = 0.0003186$.

  The residual SD is $s_{Y|X} = \sqrt{\text{SS(resid)}/(n-2)} = \sqrt{0.013986/10} = 0.03740$.

  The estimated standard error of $\hat{\beta}_1$ is $s_{Y|X}/\sqrt{\sum(x_i - \bar{x})^2} = 0.03740/\sqrt{506667} = 0.00005254$.

  Our test statistic is $\hat{\beta}_1/(\hat{\text{SE}})(\hat{\beta}_1) = 0.0003186/0.00005254 = 6.06$.

  We compare this to a t distribution with 10 degrees of freedom. For a one-sided test at $\alpha = 0.05$, we look at the 95th percentile of the $t(df = 10)$, which is 1.812. Since 6.06 $>$ 1.812, we reject the null hypothesis and conclude that there is a relationship between altitude and respiration rate. The p-value is $<$ 0.0005.