The final is here on Thursday, May 10 from 12:50-02:50 PM. **You are allowed 6 sheets of notes and a calculator.** The exam is cumulative. See reviews for exams 1, 2 and 3. The old practice final should be useful. Then the remaining quizzes, then the homework. Memorize $\overline{x}$, sd s, zscore, and how to use tables A and C.

```
Response      = Y
Coefficient Estimates
Label      Estimate Std. Error    t-value    p-value
Constant     a       not         important    for final
x            b       SE(b)     to = b/SE(b)  pvalue for Ho beta = 0
-----------------------------------------------------------------
Response      = brnweight
Terms         = (size)
Coefficient Estimates
Label       Coef           St. Dev         T           P
Constant   305.945         35.1814         8.696     0.0000
size       0.271373        0.00986642     27.505     0.0000
```

The following three problems will be on the final.

55) Be able to find the least squares line $\hat{y} = a + bx$ from Minitab output. A typical table and a table with numbers are shown above. Then predict y for a given x.
See Q11; E2 7; HW3 Eb, H; HW13 C.

56) Using a t-table, be able to find the 100 $(1 - \delta)$ % CI for β is $b \pm t^* \; SE(b)$. If $df = n - 2 \leq 30$, get t^* from table C. If $df > 30$ use $t^* = z^*$ in table C (as usual).
See Q11; HW13 Aa.

57) Be able to perform the 4 step t–test of hypotheses:
i) State the hypotheses Ho: $\beta = 0$ Ha: $\beta \neq 0$.
ii) Find the test statistic $to = b/SE(b)$ from output (usually).
iii) Find the p–value from output (usually).
iv) If p–value $< \delta$, reject Ho and conclude that x is a useful linear predictor of y. If p–value $\geq \delta$, fail to reject Ho and conclude that x is not a useful linear predictor of y. Get x and y from the story problem and use $\delta = 0.05$ if δ is not given.

See Q11, HW13 Ab, B, I.

The p–value can also be obtained from table C (with the "Two Sided P" line) if $df = n - 2 \leq 30$: p–value $= 2P(t_{n-2} > |to|)$. Use table A if $df = n - 2 > 30$: p–value $= 2P(Z < -|to|)$. Note that "linear" is crucial. It could be that x is a very useful nonlinear predictor for y, but not a good linear predictor.

58) A 100 $(1 - \delta)$% confidence interval (CI) for $\mu_y = \alpha + \beta x^*$ when $x = x^*$ is for the parameter (mean) μ_y while a 100 $(1 - \delta)$% prediction interval (PI) for a new observation y_{new} when $x = x^*$ is for the random variable y_{new}. If both intervals are given by output, know which is which. See HW13 C.

The following problem will be on the final.

59) Suppose that there are two categorical variables: the row variable with r categories and the column variable with c categories. Know how to perform the 4 step test:
i) Ho: there is no relationship between the two categorical variables
Ha: there is a relationship.
ii) test statistic = X^2
iii) p-value = $P(\chi^2_{(r-1)(c-1)} > X^2)$.
iv) Reject Ho if the p–value $\leq \delta$, and conclude that there is a relationship between the two categorical variables. If the p-value $> \delta$, fail to reject Ho and conclude that there is no relationship between the two variables.

See Q11, HW12 A, B, C, D. Sometimes X^2 is given by output but sometimes you need to compute the expected count and the chisquare contribution. Recall that the expected cell count = (row total)(column total)/(table total). The chisquare cell contribution = $(O-E)^2/E$ where O and E are the observed and expected cell counts. The expected cell count and the cell chisquare contribution need to be computed for each of the rc cells. Finally, X^2 is the sum of all rc cell chisquare contributions. See Q11, HW12 B.

Sometimes the p–value is given by output but sometimes it needs to be obtained from table E. The df = (r-1)(c-1). Since this test is always a right tail test, find the two values in the df row of table E that are closest to X^2. Then the p–value is between the values on the top row of the table. For example, if df = 5 and $X^2 = 13.00$ then 12.83 and 13.39 bracket X^2 and $0.02 < pvalue < 0.025$. If X^2 is big and way off table E, then p-value < 0.0005. For example, if df = 5 and $X^2 = 57$, then p–value = 0. If X^2 is small and way off table E, then p-value > 0.25. For example, if df = 5 and $X^2 = 4.33$, then p–value > 0.25.

Know what a lurking variable is.

Know the difference between an observational study and an experiment.

In this class, double blinded completely randomized controlled (comparative) experiments are best. Next best are single blinded completely randomized controlled (comparative) experiments and completely randomized controlled (comparative) experiments are still very good. Observational studies are ok.

Experiments that are controlled (comparative) but not randomized and experiments that (are not comparative) have a treatment group but no control group are bad (analogous to voluntary response samples and samples of convenience).

Know that randomization is the most important step in an experiment. Randomization washes out the effects of lurking variables, makes the treatment group like the control group except for the treatment, and allows one to find valid confidence intervals and two sample tests of hypotheses.

The following two problems will probably be on the final.

60) If you are given the results of observational studies and completely randomized experiments and the two results differ, then conclude that the results from the completely randomized experiment are correct. See HW13 F and maybe Q11.

61) Know how to use the random numbers to divide n individuals into a treatment group and a control group. See HW 13 Gc.

DO NOT GIVE OUT

7) Know how to do a **forwards calculation using table A** for $\overline{x}$, $\hat{p}$ or X where X is normal with mean μ and SD σ. See E1 9, 10, 11; E2 4, 11; E3 10; Q3 2, Q5 2; Q10 4; HW2 1.58, 1.60a, 1.67bc; HW6 43a, 51a, 63a; HW 11 8.4ac.

9) Know the difference between an individual and a population. HW3 3.5.

10) Know that association does not imply causation.

11) Know what a lurking variable is.

The following problem is very important for the midterm and the final.

12) Know how to get a SRS using table B. See Q4 1, HW4 3.8, 3.10.

13) Know that voluntary response samples and samples of convenience are bad while probability samples are good. Q5 3, HW4 3.6.

14) Know that SRS's are too expensive so multistage samples are used when interviewers are sent out. Random digit dialing is used for many opinion polls.

15) Know that the accuracy of the sample depends on the size of the sample. Two SRS's of the same size have the same accuracy if the sample size is small compared to both population sizes. Bigger samples have greater accuracy. Some times you will be given the sample size, sometimes a percentage of the population sampled (then you need to figure out which sample is larger. See Q5 5, HW4 3.31.

The following problem is very important for the midterm and the final.

16) Know that the normal approx for $\overline{x}$ holds for $n \geq 1$ if the population of x is (approximately) normal. Know that the CLT **does not apply** if $n \leq 30$ and x comes from a highly skewed population. See Q5 1,2.

17) Know that for the CLT to apply, the data needs to be a SRS or observations from a randomized experiment (eg coin tossing). If the data comes from a sample of convenience or a voluntary response sample, you can not find probabilities such as $P(\overline{x} < a)$.

18) Know that $\mu_{\overline{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

The following problem is very important for both the midterm and the final.

19) Know how to do a forwards calculation involving $\overline{x}$. See Q5 2, HW6 4.43bc, 4.51b and 4.63bcd.

20) Law of large numbers. Figure out the mean μ. If μ is favorable (eg stock market, number of questions likely to get right if you are a good student) larger sample sizes n are better than smaller. If μ is not favorable (eg casino gambling or guessing on a multiple choice exam) smaller sample sizes are better. See HW6 4.46.

21) Know that for any event A, $0 \leq P(A) \leq 1$.

22) **Probability rules:** i) $P(S) = 1$
ii) **Complement rule:** P(not A) = 1− P(A).
iii) A and B are disjoint events if A and B have no outcomes in common. Hence if A occurs, B did not occur and vice verca. If A and B are disjoint, then the **addition rule for disjoint events** is P(A or B) = P(A) + P(B).

iv) Finite S. If $S = \{e_1, ..., e_k\}$ then $0 \leq P(e_i) \leq 1$, $\sum_{i=1}^{k} P(e_i) = 1$. If e_i is a sample point, then $P(A) = \sum_{i:e_i \in A} P(e_i)$. That is, $P(A)$ is the sum of the probabilities of the sample points in A. If all of the outcomes e_i are *equally likely,* then $P(e_i) = 1/k$ and $P(A) =$ (number of outcomes in A)$/k$ if S contains k outcomes.

v) Two events A and B are **independent** if knowing that one occurs does not change the probability that the other occurs. If events are not independent, then they are dependent. Two events A and B are independent if P(A and B) = P(A) P(B). The events $A_1, ..., A_n$ are independent if knowing any subset of one to n - 1 events occurred does not change the probabilities of the other events.

vi) **Multiplication rule:** If A and B are independent, then
P(A and B) = P(A) P(B).

vii) **General Addition Rule:** P(A or B) = P(A) + P(B) − P(A and B).
Notice that if A and B are disjoint, then P(A or B) = P(A) + P(B).
Notice that if A and B are independent, then P(A or B) = P(A) + P(B) − P(A) P(B).

viii) **Addition rule for n disjoint events:** If $A_1, ..., A_n$ are disjoint, then $P(A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$. This is the probability that at least one of the n events occurs.

ix) **Multiplication rule for n independent events:** If $A_1, ..., A_n$ are independent, then $P(A_1 \text{ and } A_2 \text{ and } \cdots \text{ and } A_n) = P(A_1)P(A_2)\cdots P(A_n)$. This is the probability that all n events occur.

23) Leave the probabilities of some outcomes blank. See Q5 6,7, HW 5 4.18, and HW6 4.28.

24) Given a story problem, list the outcomes that make up an event (especially for die problems). Often you can use order to find S. Using a table to find S if two die are tossed or if a die is tossed twice and to find S if a coin is flipped 2, 3, or 4 times are typical examples. After listing all outcomes in S, use these outcomes to find $P(A)$.

25) **Toss two die** (eg red or green) (or toss a die twice with a 1st die, 2nd die). Find the probability that the sum of the two die = k. Solution: fill a table with 36 entries and find the number of entries where the sum is equal to k. These entries lie on a diagonal. Let E_k = "sum of the dice is k". Then $P(e_k)$ = P(sum of the dice is equal to k) = (number of table entries where the sum is k)/(number of table entries). Frequently a 4, 5, or 6-sided die will be used. For a 6–sided die the number of table entries is (6)(6) = 36 and

k	2	3	4	5	6	7	8	9	10	11	12

```
P(sum of two dice = k)|1/36 2/36 3/36 4/26 5/36 6/36 5/36 4/36 3/36 2/36 1/36
```

26) Given P(A), find P(not A). Given P(not A), find P(A). Use the complement rule: P(not A) = 1 − P(A).

27) **General Addition Rule:** P(A or B) = P(A) + P(B) − P(A and B).
Notice that if A and B are disjoint, then P(A or B) = P(A) + P(B).
Notice that if A and B are independent, then P(A or B) = P(A) + P(B) − P(A) P(B).
Given any three of the above probabilities, use the general additive rule to find the fourth probability. USING A PROBABILITY VENN DIAGRAM CAN BE USEFUL. The probabilities in 4 regions are P(A and not B), P(A and B), P(not A and B) and P(not A and not B). The four regions are disjoint. See HW6 5.7.

28) Given P(A), P(B), and that A and B are disjoint, find P(A and B) or find P(A and B). If A and B are disjoint, P(A and B) = 0 while P(A or B) = $P(A) + P(B)$.

29) Given P(A), P(B), and that A and B are independent, find P(A and B) or find P(A or B). If A and B are independent, P(A and B) = P(A) P(B) while P(A or B) = $P(A) + P(B) - P(A)P(B)$.

30) Know P(x was at least k) = $P(x \geq k)$ and P(x at most k) = $P(x \leq k)$.
See Q5 7.

31) Suppose there are n independent identical trials and x counts the number of successes and the p = probability of success for any given trial. Then
i) P(x=0) = P(none of the n trials were successes) = $(1-p)^n$.
ii) $P(x \geq 1)$ = P(at least one of the trials was a success) = $1 - P(x = 0) = 1 - (1-p)^n$.
iii) P(x=n) = P(all n trials were successes) = p^n.
iv) $P(x < n)$ = P(not all n trials were successes) = $1 - P(x = n) = 1 - p^n$.
See HW6 5.4, 5.10.