2028: Basic Statistical Methods
Sample Final Review Problems

1. **Effective marketing.** To devise effective marketing strategies it is helpful to know the characteristics of your customers. A study compared demographic characteristics of people who use the internet for travel arrangements and of people who do not. Of 1132 internet users, 643 had completed college. Among the 852 nonusers, 349 had completed college.

   (a) Compute a 95% confidence interval for the difference in the proportions. Do users and nonusers differ significantly in the proportion of college graduates?

   (b) Among the internet users, 493 reported income of less than $ 50,000 and 378 reported income of $ 50,000 or more. The corresponding numbers for nonusers were 477 and 200. (Note that not all participants responded to this survey.) Perform a hypothesis test to compare the proportion of users with nonusers with income of $ 50,000 or more.

2. **No-credit Card Fee.** A bank would like to investigate whether omitting annual credit card fee for customers who charge at least $3000 in a year would increase the amount charged on its credit card. The bank makes this offer to a random sample of 500 existing credit card costumers. It then compares how much these customers charge this year to the amount that they charged last year. Based on these data, the mean increase is $265, and the standard deviation is $167.

   a Is there statistically significant evidence at the $\alpha = 0.1$ significance level that the mean amount charged increases under the no-fee offer? State H_0 and H_a and carry out the test.

   b Give a 95% confidence interval for the mean increase.

   c Suppose that the bank wanted to be quite certain of detecting a mean increase of $\mu = 100$ in the amount charged at the level of significance $\alpha = 0.01$. Perhaps a sample of n = 100 customers would accomplish this. Find the approximate power of the test with n = 100 against the alternative $\mu = 100$. (*Hint:* The rejection region is provided by the sample of customers in part (a) with the correction of the sample size.)

3. **Blood sample tests.** Researchers into the genetic disease sikle cell anemia are interested in how red blood cells adhere to endothelial cells, which form the innermost lining of blood vessels. A set of 14 blood samples are obtained, and each sample is split in half. One half of the blood sample is profused over an endothelial monolayer of type A, and the other half of the blood sample is profused over an endothelial monolayer of type B. The data reported below are the number of adherent red blood cells per mm^2. Is there any evidence that the different stimulation conditions affect the adhesion of red blood cells?

Blood Sample	endothelial monolayer of type A	endothelial monolayer of type B
1	127	129
2	133	133
3	127	127
4	116	122
5	132	131
6	126	125
7	132	128
8	144	136
9	127	133
10	139	133
11	117	119
12	135	139
13	121	123
14	121	138

4. **Dwarf Plants**

   A genetic model suggests that 3/4 of the plants grown from a cross between two given strains of seeds will be of the dwarf variety. After breeding 200 of these plants, 131 were of the dwarf variety.

   (a) Does this observation strongly contradict the genetic model? Formulate the hypotheses and perform the test. State the decision.

   (b) Construct a 95% confidence interval for the true proportion of dwarf plants obtained from the given cross.

5. **Java against Perl.**

   To compare different programming languages the following task is performed on a Pentium II machine running Windows NT.

   Task: Read a text in a human language and generate random nonsense text in that language.[1]

   The following table shows the results:

[1] From *The Practice of Programming,* by B. W. Kernighan and R. Pike, 1999.

Language	CPU-time	Lines of Code
Java	9.2	105
C++ (STL-deque)	7.2	70
C++ (STL-list)	2.1	70
Python	2.4	25
Awk	2.1	20
Perl	1.0	18

The following output was obtained

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07386    1.52335  -0.048   0.9637
lines        0.07936    0.02506   3.167   0.0340 *
---
Residual standard error: 1.999 on 4 degrees of freedom
```

(a) The estimates of the intercept β_0 and of the slope β_1 have magnitudes comparable in absolute value (0.07-0.08). Yet, only one of them is statistically significant. Which one and why?

(b) There is statistical evidence that longer codes need more CPU time, suggesting that slope is positive. Test the hypothesis $H_0 : \beta_1 = 1.3$ versus the alternative $H_1 : \beta_1 < 1.3$ at the level $\alpha = 0.05$.

(c) Your friend, a professional programmer claims that he developed a new programming language "$D--$" which takes 50 lines of code for the aforementioned task. According to the regression equation your friend's code will take 3.894 CPU time, but you are ready to test his software. Find 90% prediction interval for the result of this test.

6. **Compression Strength.**

The article "Compression of single-Wall Corrugated Shipping Containers Using Fixed and Floating Test Platens" (*J. Testing and Evaluation*, 1992) describes an experiment in which several different types of boxes were compared with respect to compression strength (lb). Table below presents the results of an experiment involving four types of boxes.

		Compression Strength					
Type of box	1	655.5	788.3	734.3	721.4	679.1	699.4
	2	789.2	772.5	786.9	686.1	732.1	774.8
	3	737.1	639.0	696.3	671.7	717.2	727.1
	4	535.1	628.7	542.4	559.0	586.9	520.0

(a) Fill in the ANOVA table below:

	Df	Sum Sq	Mean Sq
Treatments	--	127375	---
Residuals	--	33839	---

(b) With μ_i denoting the true average compression strength for boxes of type i ($i = 1, 2, 3, 4$), test the hypothesis that the four means are equal: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$.

(c) Compute 95% pairwise confidence intervals of the means. Comment on these estimated confidence intervals. (You will need the critical point $q_{0.05,k,\nu} = 3.96$, where $k = 4$ the number of treatments, and $\nu = n - k = 20$)