Problem # 1

Cardiovascular Disease

Much controversy has arisen concerning the possible association between myocardial infarction (MI) and coffee drinking. Suppose the information in Table 15.1 on coffee drinking and prior MI status is obtained from 200 60–64-year-old males in the general population.

Table 15.1 Coffee drinking and prior MI status

Coffee drinking (cups/day)	MI in last 5 years	Number of people
0	Yes	3
0	No	57
1	Yes	7
1	No	43
2	Yes	8
2	No	42
3 or more	Yes	12
3 or more	No	28
	Total yes	30
	Total no	170

Test for the association between history of MI and coffee-drinking status, which is categorized as follows: 0 cups, 1 or more cups.

Problem # 1 Flowchart Analysis

- Underlying distribution normal or can the central-limit theorem be assumed to hold??
- No
- Underlying distribution is binomial?
- Yes
- Are samples independent?
- Yes

Problem # 1 Flowchart Analysis

- Are all expected values greater than or equal to 5?
- Yes
- 2 x 2 contingency table?
- Yes
- Use two-sample test for binomial proportion

Problem #1 Solution

We form the following 2×2 table:

		Coffee drinking 0	Coffee drinking 1+	
MI status	Yes	3	27	30
	No	57	113	170
		60	140	200

We use the chi-square test for 2×2 tables since all expected values are ≥ 5

$$\left(\text{the smallest expected value} = \frac{60 \times 30}{200} = 9.0\right).$$

We have the following test statistic:

$$X^2 = \frac{n\left(|ad - bc| - \frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$= \frac{200\left[|3(113) - 57(27)| - 100\right]^2}{30 \times 170 \times 60 \times 140}$$

$$= 5.65 \sim \chi_1^2 \text{ under } H_0$$

From the chi-square table (Table 6, Appendix, text), we see that $\chi^2_{1,\ .975} = 5.02$, $\chi^2_{1,\ .99} = 6.63$, and thus because

$$5.02 < 5.65 < 6.63$$

it follows that $.01 < p < .025$. Therefore, there is a significant association between prior MI status and coffee drinking, with coffee drinkers having a higher incidence of prior MI.

Problem # 1 Solution in R

```
> prop.test(xx)

        2-sample test for equality of proportions
  with continuity correction

data:  xx
X-squared = 5.6489, df = 1, p-value = 0.01747
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.38358985 -0.08699838
sample estimates:
   prop 1    prop 2
0.1000000 0.3352941
```

Problem # 2

Cardiovascular Disease

Much controversy has arisen concerning the possible association between myocardial infarction (MI) and coffee drinking. Suppose the information in Table 15.1 on coffee drinking and prior MI status is obtained from 200 60–64-year-old males in the general population.

Table 15.1 Coffee drinking and prior MI status

Coffee drinking (cups/day)	MI in last 5 years	Number of people
0	Yes	3
0	No	57
1	Yes	7
1	No	43
2	Yes	8
2	No	42
3 or more	Yes	12
3 or more	No	28
	Total yes	30
	Total no	170

15.9 Suppose coffee drinking is categorized as follows: 0 cups, 1 cup, 2 cups, 3 or more cups. Perform a test to investigate whether or not there is a "dose-response" relationship between these two variables (i.e., does the prevalence of prior MI increase or decrease as the number of cups of coffee per day increases?).

Problem # 2 Flow Chart Analysis

- Only one variable of interest?
- No
- One sample problem?
- No
- Two-sample problem?
- Yes

Problem # 2 Flowchart Analysis

- Underlying distribution normal or can the central-limit theorem be assumed to hold??
- No
- Underlying distribution is binomial?
- Yes
- Are samples independent?
- Yes

Problem # 2 Flowchart Analysis

- Are all expected values greater than or equal to 5?
- Yes
- 2 x 2 contingency table?
- No
- 2 x k contingency table?
- Yes

Problem # 2 Flowchart Analysis

- Interested in trend over k binomial proportions?

- Yes

- Use chi-square test for trend

Problem # 2 Solution

The results in Problem 15.8 would be more convincing if we were able to establish a "dose-response" relationship between coffee drinking and MI status with the risk of MI increasing as the number of cups per day of coffee consumption increases. For this purpose, we form the following 2×4 table:

		Current coffee consumption (cups per day) 0	1	2	3+	
MI status	Yes	3	7	8	12	30
	No	57	43	42	28	170
		60	50	50	40	200

We perform the chi-square test for trend in binomial proportions using the score statistic 1, 2, 3, 4 for the coffee-consumption groups 0, 1, 2, 3+ respectively. From Equation 10.24 (text, Chapter 10), we have the test statistic $X_1^2 = A^2/B$ where

$$A = \sum_{i=1}^{k} x_i S_i - x\bar{S} = 3(1) + 7(2) + 8(3) + 12(4)$$

$$-30 \times \left[\frac{60(1) + \ldots + 40(4)}{200}\right]$$

$$= 89 - \frac{30(470)}{200} = 89 - 70.5 = 18.5$$

$$B = \bar{p}\bar{q}\left[\sum_{i=1}^{k} n_i S_i^2 - \frac{\left(\sum_{i=1}^{k} n_i S_i\right)^2}{N}\right]$$

$$= \frac{30}{200} \times \frac{170}{200} \times \left[60(1)^2 + \ldots + 40(4)^2 - \frac{470^2}{200}\right]$$

$$= .1275(1350 - 1104.50) = .1275(245.5) = 31.30$$

Therefore,

$$X_1^2 = \frac{18.5^2}{31.30} = 10.93 \sim \chi_1^2 \text{ under } H_0 .$$

Since $\chi^2_{1,\,.999} = 10.83 < X_1^2$, it follows that $p < .001$. Thus, there is a significant linear trend, with the rate of prior MI increasing as the number of cups/day of coffee consumed increases.

Problem #2 in R

```
> library(stats)
> prop.trend.test(x = c(3,7,8,12), n =
  c(60,50,50,40))

        Chi-squared Test for Trend in
  Proportions

data:  c(3, 7, 8, 12) out of c(60, 50, 50,
  40) ,
 using scores: 1 2 3 4
X-squared = 10.9341, df = 1, p-value =
  0.0009441
```

Problem # 3

A study was conducted among a group of people who underwent coronary angiography at Baptist Memorial Hospital, Memphis, Tennessee, between January 1, 1972, and December 31, 1986 [2]. A group of 1493 people with coronary-artery disease were identified and were compared with a group of 707 people without coronary-artery disease (the controls). Both groups were age 35–49 years. Risk-factor information was collected on each group. Among cases, the mean serum cholesterol was 234.8 mg/dL with standard deviation = 47.3 mg/dL. Among controls, the mean serum cholesterol was 215.5 mg/dL with standard deviation = 47.3 mg/dL.

What test is appropriate to determine if the true mean serum cholesterol is different between the two groups?

Problem # 3 Flow Chart Analysis

- Only one variable of interest?
- No
- One sample problem?
- No
- Two-sample problem?
- Yes

Problem # 3 Flowchart Analysis

- Underlying distribution normal or can the central-limit theorem be assumed to hold??
- Yes
- Inference concerning means?
- Yes
- Are samples independent?
- Yes

Problem # 3 Flowchart Analysis

- Are variances of two samples significantly different? (Note – Should be tested using the F test)
- No
- Use two-sample t test with equal variances.

Problem # 3 Solution

Let x_i be the serum cholesterol for the ith case and y_j be the serum cholesterol for the jth control. We assume $x_i \sim N(\mu_1, \sigma_1^2)$, $y_j \sim N(\mu_2, \sigma_2^2)$. We wish to test the hypothesis $H_0: \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$. Since $s_1 = s_2$, we will assume equal variances, i.e., $\sigma_1^2 = \sigma_2^2$. Thus, we use the two-sample t test for independent samples with equal variances.

We have the test statistic

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{234.8 - 215.5}{\sqrt{47.3^2\left(\frac{1}{1493} + \frac{1}{707}\right)}} = \frac{19.3}{2.159}$$
$$= 8.94 \sim t_{1493+707-2} = t_{2198}$$

Since $t > t_{120,\ .9995} = 3.373 > t_{2198,\ .9995}$, it follows that $p < 2 \times (1 - .9995)$ or $p < .001$.

Problem # 4

What power did the study have to detect a significant difference using a two-sided test with $\alpha = .05$ if the true mean difference is 10 mg/dL between the two groups and the true standard deviations are the same as the sample standard deviations in the study?

Problem # 4 Solution

We use the power formula

$$\text{Power} = \Phi\left(-z_{1-\alpha/2} + \frac{|\Delta|}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right)$$

In this case $\alpha = .05$, $z_{1-\alpha/2} = z_{.975} = 1.96$, $\Delta = 10$, $n_1 = 1493$, $n_2 = 707$, $\sigma_1^2 = \sigma_2^2 = 47.3^2$. Thus, we have

$$\text{Power} = \Phi\left(-1.96 + \frac{10}{\sqrt{47.3^2/1493 + 47.3^2/707}}\right)$$

$$= \Phi(-1.96 + 4.63) = \Phi(2.671) = .996$$

Thus, there is a 99.6% chance of finding a significant difference.

Problem # 5a

A new drug therapy is proposed for the prevention of low-birthweight deliveries. A pilot study undertaken, using the drug on 20 pregnant women, found that the mean birthweight in this group is 3500 g with a standard deviation of 500 g.

What is the standard error of the mean in this case?

$$\text{sem} = 500/\sqrt{20} = 111.8$$

Problem # 5b

What is the difference in interpretation between the standard deviation and standard error in this case (in words)?

The standard deviation is a measure of variability for the birthweight of *one* infant. The standard error of the mean is a measure of variability for the *mean* birthweight of a group of n infants (in this case $n = 20$). The standard error will always be smaller than the standard deviation because a mean of more than one birthweight will be less variable in repeated samples than an individual birthweight.

Problem # 5c

Suppose $(\bar{x} - \mu_0)/(s/\sqrt{n}) = 2.73$ and a one-sample t test is performed based on 20 subjects. What is the two-tailed p-value?

$p = 2 \times Pr(t_{19} > 2.73)$. We refer to Table 5 (Appendix, text) and note that $t_{19,\ .99} = 2.539$, $t_{19,\ .995} = 2.861$. Since $2.539 < 2.73 < 2.861$, it follows that

$$2 \times (1 - .995) < p < 2 \times (1 - .99)$$

or $.01 < p < .02$. The exact p-value obtained by computer is $p = .013$.

Problem #6

A comparison is made between demographic characteristics of patients using fee-for-service practices and prepaid group health plans. Suppose the data presented in Table 15.2 are found.

Table 15.2 Characteristics of patients using fee-for-service practices and prepaid group health plans

	Fee-for-service			Prepaid group health plans		
Characteristic	Mean	sd	n	Mean	sd	n
Age (years)	58.1	6.2	57	52.6	4.3	48
Education (years)	11.8	0.7	57	12.7	0.8	48

Problem #6a

Test for a significant difference in the variance of age between the two groups.

We assume that the distribution of age is normally distributed in each group. We have the test statistic $F = s_1^2/s_2^2 = 6.2^2/4.3^2 = 2.08 \sim F_{56,\,47}$ under H_0. Since $F = 2.08 > F_{24,\,40,\,.975} = 2.01 > F_{56,\,47,\,.975}$, it follows that $p < .05$ and there are significant differences between the variances.

```
> 2 * pf(q=2.08, df1 = 24, df2 = 40,lower.tail=FALSE)
[1] 0.03939543
```

Problem #6b

What is the appropriate test to compare the mean ages of the two groups?

Problem # 6 Flow Chart Analysis

- Only one variable of interest?
- No
- One sample problem?
- No
- Two-sample problem?
- Yes

Problem # 6 Flowchart Analysis

- Underlying distribution normal or can the central-limit theorem be assumed to hold??
- Yes
- Inference concerning means?
- Yes
- Are samples independent?
- Yes

Problem # 6 Flowchart Analysis

- Are variances of two samples significantly different? (Note – Should be tested using the F test)
- Yes
- Use two-sample t test with unequal variances.

Problem #6c

Perform the test and report a p-value.

We have the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{58.1 - 52.6}{\sqrt{\frac{6.2^2}{57} + \frac{4.3^2}{48}}} = \frac{5.5}{1.029} = 5.34$$

We determine the effective df from Equation 8.21 (Chapter 8, text) as follows:

$$d' = \frac{\left(\frac{6.2^2}{57} + \frac{4.3^2}{48}\right)^2}{\frac{(6.2^2/57)^2}{56} + \frac{(4.3^2/48)^2}{47}} = 99.5$$

Therefore, $t = 5.34 \sim t_{99}$ under H_0. Since

$$t > t_{60,\ .9995} = 3.460 > t_{99,\ .9995},$$

it follows that $p < .001$ and there are significant differences in mean age between the two groups.

```
> 2 * pt(q=5.34, df = 99,lower.tail=FALSE)
[1] 5.929004e-07
```

Problem # 7

Hypertension
An investigator wishes to determine if sitting upright in a chair versus lying down on a bed will affect a person's blood pressure. The investigator decides to use each of 10 patients as his or her own control and collects systolic blood-pressure (SBP) data in both the sitting and lying positions, as given in Table 15.3.

Table 15.3 Effect of position on SBP level (mm Hg)

Patient	Sitting upright	Lying down
1	142	154
2	100	106
3	112	110
4	92	100
5	104	112
6	100	100
7	108	120
8	94	90
9	104	104
10	98	114

Problem # 7a

What is the distinction between a one-sided and a two-sided hypothesis test in this problem?

The distinction between a one-sided and two-sided test in this case is that for a one-sided test we would test the hypothesis $H_0: \mu_1 = \mu_2$ versus $H_1: \mu_1 > \mu_2$ or, alternatively, $H_0: \mu_1 = \mu_2$ versus $H_1: \mu_1 < \mu_2$, where μ_1 represents mean SBP (systolic blood pressure) sitting upright and μ_2 represents mean SBP lying down. For a two-sided test we would test the hypothesis $H_0: \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$.

Problem # 7b

Which hypothesis test is appropriate here? Why?

A two-sided test is appropriate here, since we have no preconceived notions as to the relative orderings of μ_1 and μ_2 and would be equally interested in the outcomes $\mu_1 < \mu_2$ and $\mu_1 > \mu_2$, (i.e., we don't know if SBP is higher while sitting upright or lying down in a bed).

Problem # 7c

- Which hypothesis test is appropriate here?

Problem # 7 Flow Chart Analysis

- Only one variable of interest?
- No
- One sample problem?
- No
- Two-sample problem?
- Yes

Problem # 7 Flowchart Analysis

- Underlying distribution normal or can the central-limit theorem be assumed to hold??
- Yes
- Inference concerning means?
- Yes
- Are samples independent?
- No
- Use paired t test.

Problem # 7d

- Conduct the hypothesis test and report a p-value.

Because each person is serving as his or her own control, we are dealing with highly dependent samples and must use the paired t test. We test the hypothesis H_0 : $\mu_d = 0$ versus H_1 : $\mu_d \neq 0$, where d_i = sitting SBP – lying SBP for the ith person and

$$d_i \sim N(\mu_d, \sigma_d^2).$$

We have the following set of within-pair differences: –12, –6, +2, –8, –8, 0, –12, +4, 0, –16. Compute the test statistic

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{-5.60}{\frac{6.786}{\sqrt{10}}} = \frac{-5.60}{2.146} = -2.61$$

under H_0, $t \sim t_9$ and we have from Table 5 (Appendix, text) that $t_{9,\ .975} = 2.262 < |t|$.

Therefore, H_0 would be rejected at the 5% level and the hypothesis that position affects level of SBP, with the sitting upright position having the lower blood pressure, would be accepted.

Problem # 7d in R

```
x = c(142, 100, 112, 92, 104, 100, 108, 94, 104, 98)

y = c(154, 106, 110, 100, 112, 100, 120, 90, 104, 114)
t.test(x=x, y = y, paired=TRUE)
 Paired t-test

data:  x and y
t = -2.6098, df = 9, p-value = 0.02828
alternative hypothesis: true difference in means is not
   equal to 0
95 percent confidence interval:
 -10.4541299  -0.7458701
sample estimates:
mean of the differences
                   -5.6
```