Homework 8

Problem 1

According to the Stanford blood center's 2011 data, the proportions of blood types in the US population were as follows.

Blood Type	O+	A+	B+	AB+	O-	A-	B-	AB-
Proportion in US	0.374	0.357	0.085	0.034	0.066	0.063	0.015	0.006

A random sample of 1000 US residents produced the following counts of blood types.

Blood Type	O+	A+	B+	AB+	O-	A-	B-	AB-
Sample Count	331	398	76	40	61	67	17	10

Use a hypothesis test to check if there the sample provides significant evidence that the proportions of blood types have changed since 2011.

(a) State the hypotheses.

(b) Compute the test statistic.

(c) Are the conditions for a one-way test satisfied?

(d) What is the null distribution of the statistic?

(e) Approximate the p-value.

(f) Make a conclusion at $\alpha = 0.01$.

(g) What would a Type I error be in this setting? What is the probability of a Type I error?

Problem 2

A diner serves milkshakes in three flavors: chocolate, strawberry, and vanilla. The owner hypothesizes that 40% of all milkshake orders are chocolate, another 40% are strawberry, and 20% are vanilla. A random sample of 357 customers who bought milkshakes revealed the following counts.

Flavor	chocolate	strawberry	vanilla
Sample Count	54	135	168

Use a hypothesis test to check if there the owner was correct in guessing the proportions.

(a) State the hypotheses.

(b) Compute the test statistic.

(c) Are the conditions for a one-way test satisfied?

(d) What is the null distribution of the statistic?

(e) Approximate the p-value.

(f) Make a conclusion at $\alpha = 0.1$.

(g) What would a Type II error be in this setting?

Problem 3
The table below shows the counts of different degrees in foreign languages earned by men and women in 1992. (Moore, D. S. (1996). *Statistics: concepts and controversies*, p 296)

	Bachelors	Masters	Doctorate
Men	3990	971	378
Women	9913	1955	472

Assume the data represents a random sample. Use a hypothesis test to check if gender and type of degree are independent.

(a) State the hypotheses.

(b) Compute the test statistic.

(c) Are the conditions for a test for independence satisfied?

(d) What is the null distribution of the statistic?

(e) Approximate the p-value.

(f) Make a conclusion at $\alpha = 0.05$.

(g) What would a Type I error be in this setting?

Problem 4
Here is a morbid contingency table that shows the methods of US suicides from 1992 separated by gender. (Moore, D. S. (1996). *Statistics: concepts and controversies*, p 296)

	Firearms	Poison	Hanging	Other
Men	15802	3262	3822	1571
Women	2367	2233	856	571

Assume the data represents a random sample. Use a hypothesis test to check if gender and method of suicide are independent.

(a) State the hypotheses.

(b) Compute the test statistic.

(c) Are the conditions for a test for independence satisfied?

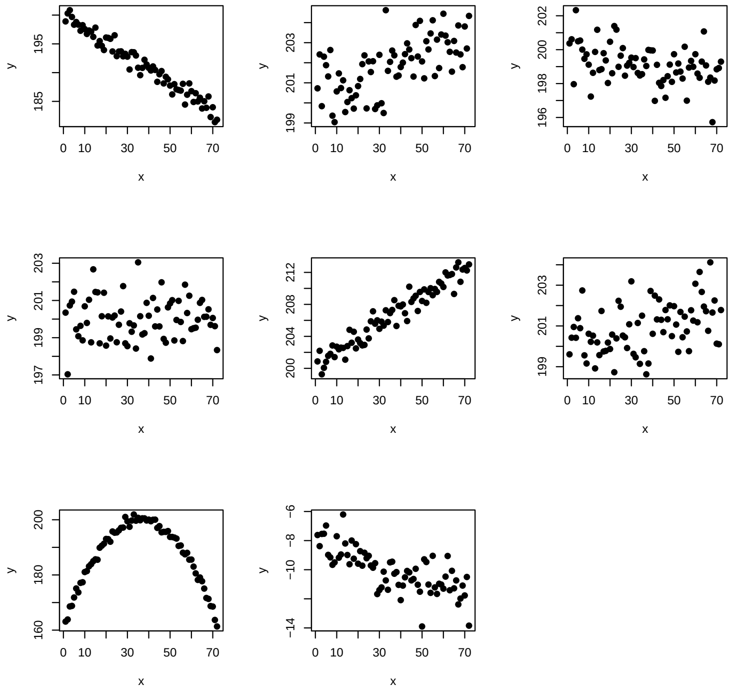
(d) What is the null distribution of the statistic?

(e) Approximate the p-value.

(f) Make a conclusion at $\alpha = 0.05$.

(g) What would a Type II error be in this setting?

Problem 5
Use the scatterplots below to answer the questions.



(a) For each plot, state if it seems to show any linear relationship at all.

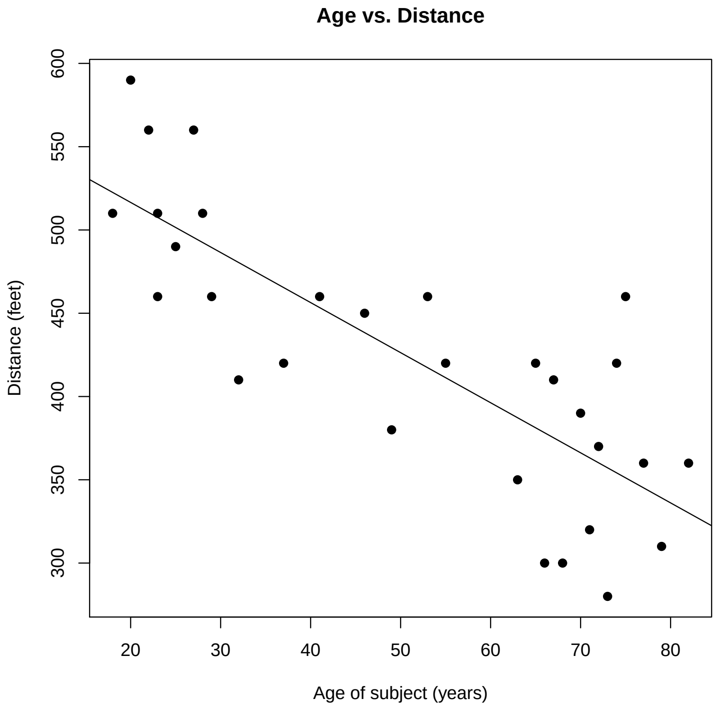
(b) For the plots that show linear relationships, state the sign (positive or negative), and the strength (weak, moderate, strong) of the relationship.

(c) For the plots that do not show linear relationships, state if there seems to be some other relationship, or none at all.

(d) Assign each sample correlation value to the scaterplot that matches it best. Two of the values/plots may be hard to classify. Do your best. Here are the sample correlation values: 0.016, -0.98, -0.73, 0.967, -0.006, 0.624, -0.39, 0.408.

Problem 6

On average, it is typical for the distance a driver can see clearly to decrease with the driver's age. The scatterplot below shows the age (X) and the distance a driver can see (Y) clearly for a random sample of 30 drivers. The estimated regression line is shown on the plot, and its equation is $\hat{Y} = 576.682 - 3.007X$.



(a) Comment on the strength and sign of the linear relationship between a driver's age, and the distance they can see clearly.

(b) Does it seem like the sample correlation coefficient is positive or negative?

(c) Interpret the value 576.682 from the estimated regression line. State your answer in terms of the problem.

(d) Interpret the value -3.007 from the estimated regression line. State your answer in terms of the problem.

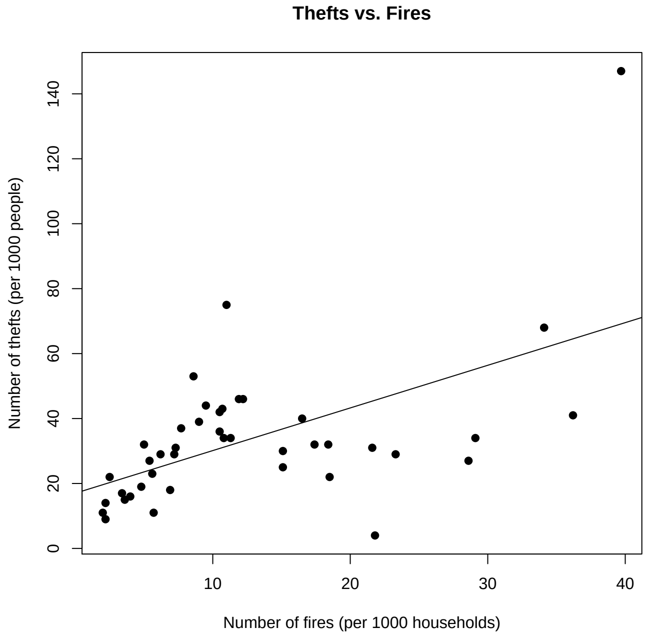
(e) If possible, use the estimated line to predict the average distance that drivers of the following ages can see clearly:

- 32 years old
- 90 years old
- 67 years old

(f) The simple linear regression model has the form $Y = \beta_0 + \beta_1 + \epsilon$, where ϵ represents the random error. State the typical assumptions about ϵ.

Problem 7

The following scatterplot shows the relationship between the number of fires (X) and the number of thefts (Y) within the same zip codes in the Chicago metropolitan area (Reference: U.S. Commission on Civil Rights).The estimated regression line is shown on the plot, and its equation is $\hat{Y} = 16.995 + 1.313X$.



(a) Comment on the strength and sign of the linear relationship between the number of thefts, and the number of fires.

(b) Interpret the value 16.995 from the estimated regression line. State your answer in terms of the problem.

(c) Interpret the value 1.313 from the estimated regression line. State your answer in terms of the problem.

(d) If possible, use the estimated line to predict the average number of thefts (per 1000 people) in areas that have the following numbers of fires:

- 50 fires per 1000 households
- 25 fires per 1000 households
- 12 fires per 1000 households

(e) The estimated coefficients (16.995 and 1.313) were computed using the least squares approach. Explain what that means. You do not need to use any formulas or calculations.

Textbook Problems

Lecture 21: 8.31, 8.32, 8.33, 8.34, 8.36, 8.38, 8.50, 8.52, 8.55, 8.57
Lecture 22: none