For each of the problems below, do the following:

a) **Read the scenario and determine the appropriate chi-square test:**
   i) Chi-Square Goodness of Fit
   ii) Chi-Square of Independence/Association

b) **Do the following steps to perform the significance test:**
   i) Write the hypotheses
   ii) Calculate the expected counts
   iii) Check the appropriate conditions
   iv) Calculate the degrees of freedom
   v) Calculate the test statistic and p-value
   vi) Interpret the p-value
   vii) Conclude the test in context.

For significance tests, we will use a significance level of $\alpha = 0.05$, unless otherwise indicated.

Solutions are at the end of the document.

The problems below are from Daren Starnes, et al, *Statistics and Probability with Applications*, 4th ed.

Problem 1

Will randomly selecting households with landlines produce a representative sample? According to the Census Bureau, of all U.S. residents aged 18 and older, 13% are 18−24 years old, 35% are 25−44 years old, 35% are 45−64 years old, and 17% are 65 years and over. The table below gives the age distribution for a sample of U.S. residents aged 18 and older that was chosen by randomly dialing landline telephone numbers. Is there convincing evidence that the sample is not representative of the population?

Age	**18−24**	**25−44**	**45−64**	**65+**	**Total**
Count	19	120	162	99	400

Problem 2

Faked numbers in tax returns, invoices, or expense account claims often display patterns that aren't present in legitimate records. Some patterns are obvious and easily avoided by a clever crook. Others are more subtle. It is a striking fact that the first digits of numbers in legitimate records often follow a model known as Benford's law. Here is the distribution of first digits for variables that follow Benford's law.

First digit	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Proportion	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

A forensic accountant who is familiar with Benford's law inspects a random sample of 250 invoices from a company that is accused of committing fraud. The table displays the sample data. Is there convincing evidence that the invoices from this company don't follow Benford's law?

First digit	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
Count	61	50	43	34	25	16	7	8	6	250

Problem 3

It's hard for smokers to quit. Perhaps prescribing a drug to fight depression will work as well as the usual nicotine patch. Perhaps combining the patch and the drug will work better than either treatment alone. Here are data from a randomized, double-blind trial that compared four treatments. A "success" means that the subject did not smoke for a year following the beginning of the study.

		Treatment				
		Nicotine patch	**Drug**	**Patch plus drug**	**Placebo**	**Total**
Response	**Success**	40	74	87	25	226
	Failure	204	170	158	135	667
	Total	244	244	245	160	893

a) Do the data provide convincing evidence of an association between treatment and response for subjects like these?

b) Based on your conclusion in part (a), could you have made a Type I or a Type II error? What is a potential consequence of this error? Explain.

Problem 4

In a long-term study of male firefighters in Indiana, researchers divided the firefighters into 5 categories based on the number of pushups they could do without stopping at the beginning of the study. At the end of 10 years, researchers recorded whether or not the firefighters had a cardiovascular disease (CVD) related event since the beginning of the study. The two-way table summarizes the data. Do the data provide convincing evidence that there is an association between the number of pushups a firefighter can do and the occurrence of a CVD event?

		Number of pushups			
		0–20	**21–40**	**40+**	**Total**
CVD-related event	**Yes**	17	19	1	37
	No	258	655	154	1067
	Total	275	674	155	1104

Solutions

Problem 1

Name the procedure: Chi-squared Goodness of Fit test

Write the hypotheses:

H_o: $P_{18\text{-}24}$= 0.13 $P_{25\text{-}44}$=0.35 $P_{45\text{-}64}$=0.35 P_{65+}=0.17
H_a: At least one of these is incorrect

OR

H_o: The distribution of age is the same as what is stated by the census bureau.
H_a: The distribution of age is **not** the same as what is stated by the census bureau.

Calculate the expected counts:

age	18-24	25-44	45-64	65+
observed counts	19	120	162	99
expected counts	0.13*400 = 52	0.35*400=140	0.35*400=140	0.17*400=68

Check the appropriate conditions

Random: Since we randomly dialed telephone numbers, the sample data is representative of the population.

Independent: We assume there are more than 4000 (400*10) households in the USA.

Expected Counts: The expected counts found above all are greater than or equal to 5. (The smallest is 52.) so our chi-squared test statistic follows a chi-square distribution.

Calculate the degrees of freedom:

DF= number of categories minus 1
4-1=3
DF=3

Calculate the test statistic and p-value:

$$\sum \frac{(observed-expected)^2}{expected} = \frac{(19-52)^2}{52} + \frac{(120-140)^2}{140} + \frac{(162-140)^2}{140} + \frac{(99-68)^2}{68} = 41.3889$$

Approach 1: Using the χ^2 table with df = 4-1=3 degrees of freedom gives a P-value smaller than 0.005 (see screenshot); the largest test statistic in the table is 12.838, and our 41.3889 is larger and off the table.

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838

Approach 2:

To calculate the P-value with R, we use the **pchisq(boundary, df, lower.tail=F)** function

pchisq(41.3889, df = 3,lower.tail = F) = 0.00000 **(it was really 5.408 x 10^{-9})**

Interpret the P-value

Assuming that the true distribution of ages is as stated by the census bureau, there is a 0.000000 probability that we would observe the distribution we did or something further from the expected distribution by sampling variability alone.

Conclude the test:

Since the P-value of 0.000000 is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that we have convincing evidence that the sample distribution is **not** the same as what is stated by the census bureau

Problem 2

Name the procedure: Chi-squared Goodness of Fit test

Write the hypotheses:

Ho: P_1= 0.301 P_2=0.176 P_3=0.125 P_4=0.097 P_5=0.079 P_6=0.067 P_7=0.058 P_8=0.051 P_9=0.046
Ha: At least one of these is incorrect

OR

Ho: The sample distribution is the same as what is stated by Benford's law
Ha: The sample distribution is **not** the same as what is stated by Benford's law

Calculate the expected counts: expected counts = percent expected *total observed counts

digits	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
observed	**61**	**50**	**43**	**34**	**25**	**16**	**7**	**8**	**6**
expected	**75.25**	**44**	**31.25**	**24.25**	**19.75**	**16.75**	**14.5**	**12.75**	**11.5**

Check the appropriate conditions

Random: Since we took a random sample of 250 invoices, the sample data is an unbiased estimator of the population.

Independent: We assume there are more than 2500 (250*10) invoices to choose from

Expected Counts: The expected counts found above all are greater than or equal to 5.

Calculate the degrees of freedom:

DF= number of categories minus 1
9-1=8
DF=8

Calculate the test statistic and p-value:

$\sum \frac{(observed-expected)^2}{expected}$ **=**

$\frac{(61-75.25)^2}{75.25} + \frac{(50-44)^2}{44} + \frac{(43-31.25)^2}{31.25} + \frac{(34-24.25)^2}{24.25} + \frac{(25-19.75)^2}{19.75} + \frac{(16-16.75)^2}{16.75} + \frac{(7-14.5)^2}{14.5} + \frac{(8-12.75)^2}{12.75} + \frac{(6-11.5)^2}{11.5}$**=**

21.563

Approach 1: Using the χ^2-table with df = 9-1=8 degrees of freedom gives a P-value between 0.01 and 0.005 (see screenshot).

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955

Approach 2:

To calculate the P-value with R, we use the **pchisq(boundary, df, lower.tail=F)** function

`pchisq(21.563, df = 8,lower.tail = F) = 0.005792`

Interpret the P-value

Assuming that the true distribution of the invoice starting digit follows Benford's law, there is a **0.005792** probability that we would observe the distribution we did or something further from the expected distribution by sampling variability alone.

Conclude the test:

Since the P-value of**0.005792** is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that we have convincing evidence that the distribution of our sample is **not** the same as Benford's law and therefore we believe there is fraud.

Problem 3

Name the procedure: Chi-Squared test of Association/Independence

Write the hypotheses:
Ho: There is no association between success of smokers quitting and the method used to quit.
Ha: There is an association between success of smokers quitting and the method used to quit

Calculate the expected counts: $\frac{(Row\ Total)*(Column\ Total)}{Table\ Total}$

		Treatment			
		Nicotine patch	**Drug**	**Patch plus drug**	**Placebo**
Response	**Success**	61.7514	61.7514	62.0045	40.4927
	Failure	182.249	182.249	182.996	119.507

Check the appropriate conditions
Random: We have a randomized, double-blind trial that compared four treatments.

Independent: Each participant is independent from the next
Expected Counts: The expected counts found above in the chart are all greater than or equal to 5.

Calculate the degrees of freedom:
DF= (number of rows -1) *(number of columns -1)
1*3=3
DF=3

Calculate the test statistic and p-value:

$\sum \frac{(observed-expected)^2}{expected}$ **=**

$$\frac{(40-61.7514)^2}{61.7514}+\frac{(74-61.7514)^2}{61.7514}+\frac{(87-62.0045)^2}{62.0045}+\frac{(25-40.4927)^2}{40.4927}+\frac{(204-182.249)^2}{182.249}+\frac{(170-182.249)^2}{182.249}+\frac{(158-182.996)^2}{182.996}+\frac{(135-119.507)^2}{119.507}$$

= 34.937

Approach 1: Using the χ^2 table with df = 1*3=3 degrees of freedom gives a P-value smaller than 0.005 (see screenshot); the largest test statistic in the table is 12.838, and our **34.937** is larger and off the table.

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838

Approach 2:
To calculate the P-value with R, we use the **pchisq(boundary, df, lower.tail=F)** function

pchisq(34.937, df = 3, lower.tail = F) = 0.00000 (it was really 1.256 x 10^{-7})

Interpret the P-value

Assuming there is no association between success of smokers quitting and the method used to quit, there is a 0.000000 probability that we would observe the results we did or something more extreme by sampling variability alone.

Conclude the test:

Since the P-value of 0.000000 is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that we have convincing evidence that there is an association between success of smokers quitting and the method used to quit

Problem 4

Name the procedure: Chi-Squared test of Association/Independence

Write the hypotheses:

Ho: There is no association between the number of pushups a firefighter can do and the occurrence of a CVD event.

.

Ha: There is an association between the number of pushups a firefighter can do and the occurrence of a CVD event.

Calculate the expected counts: $\frac{(Row\ Total)*(Column\ Total)}{Table\ Total}$ **shown in table below**

		Number of pushups		
		0–20	**21–40**	**40+**
CVD-related event	**Yes**	9.21649	22.5888	5.19475
	No	265.784	651.411	149.805

Check the appropriate conditions

Random: Since the firefighters were not randomly selected, we cannot be sure that this sample is representative of all firefighters. We will proceed with caution when discussing the scope of inference.

Independent: We will assume that each participant is independent from the next and that there are more than

Expected Counts: All expected counts are greater than 5

Calculate the test statistic and p-value:

$$\sum \frac{(observed-expected)^2}{expected} = \frac{(17-9.21649)^2}{9.21649} + \frac{(19-22.5888)^2}{22.5888} + \frac{(1-5.19475)^2}{5.19475} + \frac{(258-265.784)^2}{265.784} + \frac{(655-651.411)^2}{651.411} + \frac{(154-149.805)^2}{149.805} =$$

10.8959

Approach 1: Using the χ^2 table with df = 1*2=2 degrees of freedom gives a P-value below 0.005. 10.8959 is off of the chart therefore gives us a P-value less than 0.005(see screenshot).

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838

Approach 2:

To calculate the P-value with R, we use the **pchisq(boundary, df, lower.tail=F)** function

pchisq(10.8959, df = 2,lower.tail = F) = 0.004305

Interpret the P-value

Assuming there is no association between the number of pushups a firefighter can do and the occurrence of a CVD event, there is a **0.004305** probability that we would observe the results we did or something more extreme by sampling variability alone.

Conclude the test:

Since the P-value of **0.004305** is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that we have convincing evidence that there is an association between the number of pushups a firefighter can do and the occurrence of a CVD event.