1. The grades in an upper division class are tabled to compare the performance of juniors and seniors:

	A	B	C	F
Juniors	16	32	23	2
Seniors	8	24	16	3

We want to test whether or not the distribution of grades was the same for the two classes of students.

(a) Is the chi-squared approximation appropriate for this table? If necessary, redraw the table in such a way that it corrects the problem.

Solution: The expected value for each cell in this table comes out to be

	A	B	C	F	
Juniors	14.1	33	23	2.9	73
Seniors	9.9	23	16	2.1	51
	24	56	39	5	124

The expected number of Fs for both Juniors and Seniors is less than 4 so the chi-squared approximation is not appropriate. The best solution for this table is to aggregate the number of Cs and Fs into one column. The new table is

	A	B	C or F
Juniors	16	32	25
Seniors	8	24	19

(b) Calculate the appropriate test statistic for the χ^2 test of independence.

Solution: The expected values for this table are

	A	B	C or F	
Juniors	14.1	33	25.9	73
Seniors	9.9	23	18.1	51
	24	56	44	124

Thus the test statistic is

$$X^2 = \frac{(16-14.1)^2}{14.1} + \frac{(32-33)^2}{33} + \ldots + \frac{(24-23)^2}{23} + \frac{(19-18.1)^2}{18.1} = 0.748$$

(c) What is the critical value for this test with size $\alpha = 0.01$?

Solution: There are 2 rows and 3 columns which leads to $df = (2-1)(3-1) = 2$. Thus, the critical value for a χ^2 distribution for $\alpha = 0.01$ is 9.21 (from the Table).

(d) What do you conclude?

Solution: We conclude that there is not a significant difference between the grades that Juniors are getting and the grades that the Seniors are getting.

2. The following table counts the number of outcomes in 100 trials for each of four events:

Event	A	B	C	D
Outcomes	34	30	12	24

Perform goodness-of-fit test of size $\alpha = 0.05$ to test the hypothesis that $P(A) = P(B)$ and $P(C) = P(D)$.

Solution: If we assume that $P(A) = P(B)$, then the expected number of observations in those two cells should be the same. Likewise for C and D. We could estimate $P(A) = \frac{1}{2}\left(\frac{34+30}{100}\right)$, then $P(B) = P(A)$ and $P(C) = P(D) = \frac{1}{2}(1 - 2P(A))$. The expected values under the null hypothesis are

Event	A	B	C	D
Observed	34	30	12	24
Expected	32	32	18	18
(Obs-Exp)^2/Exp	0.125	0.125	2	2

The chi-squared statistic is $X^2 = 0.125 + 0.125 + 2 + 2 = 4.25$. This χ^2 has two degrees of freedom because we estimated one probability and $\sum_{i=1}^{4} p_i = 1$. The value 4.15 is not significant when compare to the critical value $\chi^2.05, 2 = 5.991$. Therefore, we accept the hypothesis that the probabilities are equal.

3. The chair of the statistic department is interested in the majors of students in the 120 courses. The registrar tabulated the number of engineers, economist, and statisticians in each semester of 120.

	Engineering	Economics	Statistics
120A	112	63	35
120B	80	53	27
120C	40	23	12

Test whether or not the distribution of majors is the same for the three classes.(Use size $\alpha = 0.05$) What interpretation would you give to the results?

Solution: The table with its marginals is

	Engineering	Economics	Statistics	
120A	112	63	35	210
120B	80	53	27	160
120C	40	23	12	75
	232	139	74	445

The expected values are

	Engineering	Economics	Statistics
120A	109.25	65.6	34.9
120B	83.4	50	26.6
120C	39.1	23.4	12.5

The chi-squared statistic is

$$X^2 = \frac{(112-109.5)^2}{109.5} + \frac{(63-65.6)^2}{65.6} + \ldots + \frac{(23-23.4)^2}{23.4} + \frac{(12-12.5)^2}{12.5} = 0.5353.$$

Compare this to a chi-squared critical value with 4 degrees of freedom $\chi^2_{.05,4} = 9.488$, and accept H_0.

The conclusion is that the two categorizations are independent, and the proportion of different majors is essentially the same from semester to semester.

4. The National Traffic and Safety Administration reported the following data on fatal accidents in Los Angeles County over 3 years. In each accident two factors were recorded: whether alcohol was involved or not and whether speeding was involved or not. Here is a cross tabulation:

	2003	2004	2005
Neither Speeding nor Alcohol	241	251	232
Speeding Only	244	243	235
Alcohol Only	259	206	223
Both Speeding and Alcohol	68	53	60

(a) Suppose that instead of using the null hypothesis that the marginal distribution of the two factors is independent of the year (as we did in a homework problem), we wanted to test whether the conditional probability of speeding did not change from year to year. What would be the expected number of accidents in 2004 where speeding was involved conditional on the fact that alcohol was involved under the null hypothesis that the conditional distribution (given alcohol) is the same for each year?

Solution: In this case, we break out a separate table for the two rows which included accidents with alcohol involved

	2003	2004	2005	
Alcohol only	259	206	223	688
Both	68	53	60	181
	327	259	283	869

Therefore the expected value is $E_{\text{Both 2004}} = \frac{259(181)}{869} = 53.95$.

(b) The χ^2 test statistic for this test is $X^2 = 0.25$. Is this significant?

Solution: The number of degrees of freedom is 2 for each of the two tables for a total of 4 degrees of freedom. The critical value is therefore 9.49, and the test statistic is not significant.

(c) Interpret your conclusion.

Solution: The test is not significant so it is reasonable to assume that the proportion of traffic accidents that involve speeding is about the same from year to year if you condition on whether or not alcohol is involved.

5. A study of juror summonses recorded the race/ethnicity of 1271 people sent jury summonses, and then tracked how many of them served on juries, how many had their service deferred, and how many never showed up.

	No Show	Deferred	Served
White	438	250	39
Black	103	120	12
Hispanic	187	115	7

We want to test the null hypothesis that the outcome of their jury service is independent of race and ethnicity.

(a) Calculate the expected number of people in each of the cells in the third row of the table (i.e. how many Hispanics do we expect to be no shows, to be deferred, and to serve on a jury, respectively) assuming the null hypothesis is true.

Solution: The expectations are

	No Show	Deferred	Served	
Hispanic	187	115	7	309
Total	728 485	58	1271	
E	177	117.9	14.1	

(b) The χ^2 test statistic for this table is $X^2 = 26.9$. What do you conclude? (Use $\alpha = 0.01$)

Solution: The χ^2 test statistic for this table is $X^2 = 26.9$. The critical value is $\chi^2_{4,0.01} = 13.2767$. Reject the null hypothesis.

6. In the same survey as in question 6, the sample contained 625 men and 646 women.

Men

	No Show	Deferred	Served
White	225	133	20
Black	47	52	7
Hispanic	87	52	2

Women

	No Show	Deferred	Served
White	213	117	19
Black	56	68	5
Hispanic	100	63	5

We want to test the hypothesis that their jury service is marginally independent of gender and race.

(a) Calculate the expected number Hispanic women that served on a jury and the expected number of Hispanic men that served on a jury.

Solution: The expected number of Hispanic women that served on a jury is

$$\frac{168(58)}{1271} = 7.6664$$

The expected number of Hispanic men that served on a jury is

$$\frac{141(58)}{1271} = 6.4343$$

(b) Is the chi-squred approximation appropriate for this table? If not, then how can it be fixed?

Solution: The chi-squred approximation is appropriate because the smallest marginal values lead to the expectation

$$\frac{106(58)}{1271} = 4.837$$

which is greater than 4 which is our rule of thumb.

(c) If the test statistic is $X^2 = 28.75$, then what do we conclude at an $\alpha = 0.01$ level?

Solution: If the test statistic is $X^2 = 28.75$, then the critical value is $\chi^2_{10,0.01} = 23.2093$ and we conclude that there still is a significant relationship.

7. A model for incomes in a heterogeneous population takes n independent observations $X_1, \ldots, X_n$ and transforms them to generate $Y_i = \log X_i$.

Assume that Y_i have independent normal distributions with unknown mean θ and variance $\sigma^2 = 1$. We will impose a prior distribution on this mean parameter that is normal with mean 10 and variance 5.

(a) For 38 observations with $\bar{y} = 11.129$, what is our Bayes estimator of θ?

Solution: For 38 observations with $\bar{y} = 11.129$, the Bayes estimator of θ is posterior mean:

$$\begin{aligned} E(\theta|Y) &= \bar{y}\frac{n\tau^2}{n\tau^2+\sigma^2} + \mu\frac{\sigma^2}{n\tau^2+\sigma^2} \\ &= 11.129\frac{38(5)}{38(5)+1} + 10\frac{1}{38(5)+1} = 11.1231 \end{aligned}$$

(b) Give a 95% Bayesian Credible Interval for θ.

Solution: The posterior variance is $Var(\theta|y) = \frac{1(5)}{38(5)+1} = \frac{5}{191}$. The 95% credible interval is

$$11.1231 \pm 1.96\sqrt{\frac{5}{191}} = (10.806, 11.44)$$

(c) To generate an estimate on the original scale we need e^θ. Calculate the Bayes estimator for e^θ.

Solution: To generate an estimate on the original scale we need e^θ. The Bayes estimator is $E(e^\theta)$. From the MGF we get

$$E(e^X) = \exp(\theta + \sigma^2/2) = \exp(E(X) + Var(X)/2).$$

From the posterior distribution

$$E(e^\theta|y) = \exp(E(\theta|y) + Var(\theta|y)/2) = \exp\left(11.1231 + \frac{5}{191(2)}\right) = 6.8610$$

8. The daily returns on a portfolio managed by Fiduciary Inc. were recorded over 275 days. The average return was 0.0038 with a sample standard deviation of 0.00534. Then for the next 12 days, a new strategy was tested, and it produced an average return of 0.0067 with a sample standard deviation of 0.00832.

(a) Perform a two-sample t test to see whether there is a significant improvement in the strategy over the old. What would you conclude at a $\alpha = 0.05$ level?

Solution: The pooled estimate of the standard deviation is

$$S_p^2 = \frac{274(0.00534)^2 + 11(0.00832)^2}{285} = 3.04944 \times 10^{-4}$$

The test statistic is therefore

$$t = \frac{0.0038 - 0.0067}{\sqrt{3.04944 \times 10^{-4}\left(\frac{1}{275} + \frac{1}{12}\right)}} = -1.7928$$

This is greater than 1.65, therefore there is improvement at the $\alpha = 0.05$ level. The new strategy is more effective than the older one.

(b) What assumptions are you making in part a? How can you check those assumptions? Be specific.

Solution: We assume the random sample comes from normal distribution, i.e. we assume independence and normality. To check normality we can do goodness of fit test and to check the independence we can do runs test.

If normality condition is false, we may be better off just using a nonparametric Mann-Whitney U test for this problem.

9. An environmental engineer is studying the effects of a sand bar on the flow rate through 15 coastal lagoons. The engineer measures the flow rate in each lagoon before and after the sand bar is built, and the increase in the rate is recorded in gallons per minute (gpm). The engineer has a prior that the average difference in the flow rate is normally distributed with an expectation of 0 and a standard deviation of 75 gpm.

   (a) Find the Bayesian estimator of the average difference in the flow rate if we observe that the 15 lagoons saw an average increase of 5.86 gpm. (We can assume that the measurements come form a normal distribution with standard deviation 25 gpm.)

   Solution: The estimate of the mean is

   $$\hat{\theta} = \bar{x}\frac{n\tau^2}{n\tau^2+\sigma^2} + 0 = 5.86\frac{15(75^2)}{15(75^2)+25^2} = 5.8169$$

   (b) Give a 95% credible interval for the average flow rate.

   Solution: The standard deviation of the posterior is

   $$\sqrt{\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}} = \sqrt{\frac{25^2 75^2}{15(75^2+25^2}} = 6.43119$$

   This means the 95% credible interval is

   $$5.8169 \pm 1.96(6.43119),$$

   which is

   $$[-6.788, 18.4522].$$

   (c) In a sentence or two, explain to the engineer what this interval represents.

   Solution: The information about the mean leads us to believe that there is a 95% chance that it is between -6.788 and 18.4522. This interval includes 0 which indicates that the data does not convince us that there is no difference.

   (d) Calculate the posterior probability that the average flow rate is greater than 0.

   Solution:

   $$P(\theta > 0) = P\left(Z > \frac{0-5.8169}{6.43119}\right) = P(Z > -0.90448) = 0.8159$$

10. A survey of 356 registered voters in Carpinteria asked them their opinions on the local ballot measures (Measure K and Measure J) in todays election. The respondents were divided into three groups for each question: supporters, opposition, and those that said they dont know or wont be voting. Here is cross-tabulation of the responses:

		Measure J		
		Support	Oppose	Don't Know
Measure K	Support	86	92	15
	Oppose	40	25	17
	Don't Know	36	23	22

(a) Suppose that we are only interested in the outcome of Measure J. Would we conclude that there is significant difference in the proportions of supporters versus opposition to Measure J? Calculate the appropriate test statistic and interpret your result (use $\alpha = 0.05$).

Solution: The χ^2 statistic based on the three totals where the expected value is (151, 151, 54) is

$$X^2 = \frac{(162-151)^2}{151} + \frac{(140-151)^2}{151} + \frac{(54-54)^2}{54} = 1.6026$$

This not greater than the $\chi^2 0.05, 1 = 3.8$ critical value. Thus, there is not a statistically significant difference between the proportion of supporters and opposers.

(b) For testing whether the respondents answers to the two questions are independent, the expected number of respondents under the null hypothesis are tabled below.

		Measure J		
		Support	Oppose	Don't Know
Measure K	Support	87.83	??	??
	Oppose	37.31	32.25	12.44
	Don't Know	36.86	31.85	??

Calculate the 3 missing expectations.

Solution: The expected values are

$$\text{K Support and J Oppose} = 193 \times 140 = 75.9356$$
$$\text{K Support and J Dont Know} = 193 \times 54 = 29.28356$$
$$\text{K and J Dont Know} = 81 \times 54 = 12.29.$$

(c) For the test from part (b), Calculate a test statistic. What would we conclude at a $\alpha = 0.05$ level?

Solution: The X^2 statistic is

$$X^2 = \frac{1.83^2}{87.83} + \frac{(92-75.9)^2}{75.9} + \ldots + \frac{(9.71)^2}{12/29} = 24.07$$

This is significant. $\chi^2_{0.05,4} = 9.48$. There is sufficient evidence to conclude that the responses are not independent.

11. A study on the causes of ulcers took a random sample of patients with peptic ulcers or gastric ulcers as well as a control group of healthy patients. The blood type (O, A, or B) and gender of each patient was recorded.

Male Patients

	Peptic	Gastric	Control
O	472	190	1470
A	285	208	1356
B	62	43	250

Female Patients

	Peptic	Gastric	Control
O	511	193	1422
A	394	208	1269
B	72	41	320

The researchers want to test if the ulcer diagnosis group is independent of gender and blood type. We want to use a null hypothesis of marginal independence.

(a) Calculate the three corresponding expected values for Female patients with Type O blood.

Solution:

$$E(\text{Peptic, O}) = \frac{2126(1796)}{8766} = 435.58$$
$$E(\text{Gastric, O}) = \frac{2126(883)}{8766} = 214.15$$
$$E(\text{Control, O}) = \frac{2126(6087)}{8766} = 1476.27$$

(b) Confirm that the test statistic is $X^2 = 63.40$. What do we conclude?

Solution: I calculated the wrong X^2. It should have been 64. In either case this is large compared to the χ^2_{10} distribution. Because under the marginal hypothesis, we construct a table with 6 rows and 3 columns to get df $= (6-1)(3-1) = 10$. The critical value is 18.31. We conclude that there is a difference between the blood types in the two groups. People with O blood type seem to be more likely to have Gastric as opposed to Peptic Ulcers.

12. A union supervisor claims that applicants for jobs are selected without regard to race. The hiring records of the local – one that contains all male members – gave the following sequence of White (W) and Black (B) hirings:

WWWWBWWWBBWBB

Do these data suggest a nonrandom racial selection in the hiring of the union's members?

Solution: Here, $n_1 = 5$ (blacks hired), $n_2 = 8$ (whites hired), and $R = 6$. From Table 10,

$$\text{P-value } = 2P(R \leq 6) = 2(.347) = .694.$$

So, there is no evidence of nonrandom racial selection.