Final exam prep 8 - 16 - 10 Try these. Not handed in. Will go over in class. Ask questions. Report errors (especially since so much calculation and formatting is involved).

1. Chi-square basic method (for classified independent samples having completely specified expected counts). Sales of sizes of particular athletic shirts have, from past experience, the probabilities

	XS	**S**	**M**	**L**	**XL**	**XXL**
	.02	**.04**	**.09**	**.31**	**.40**	**.14**

This season we have changed the fabric to one having a lighter, silky, breathable feel. Here are sales figures from our test run of the new shirts at an event where 1100 shirts were sold:

	XS	S	M	L	XL	XXL
observed sales	**30**	**67**	**106**	**319**	**430**	**148**
expected counts	**22**	**44**	**99**	**341**	**440**	**154**

a. Fill in the above *expected counts* if 1100 sales follow the past model.

See above. For example, 22 = 1100 .02.

b. Chi-square statistic

For example, XS contributes $\frac{(o-e)^2}{e} = \frac{(30-22)^2}{22}$ **= 2.90909 to the total chi-square of 17.3072.**

c. df

6-1 = 5

e. P-value

0.00395257

f. If the past model continues to apply to present sales what is the probability of a P-value as small or smaller than we have seen?

0.00395257

g. What, if anything, appears to have happened to size preferences due to the new fabric?

It appears that sales have shifted towards smaller sizes, perhaps due to a perceived comfort of the new fabric in close proximity to the body?

2. Another chi-square for classified independent samples having completely specified expected counts. This exercise will illustrate the fact that the basic method is not restricted to any particular shape for the table. A with-replacement random sample of **100** parts from production has been sorted below by weight of material, weight of scrap, and time of assembly. The usual cell probabilities (from past production) are given in parentheses and are used to determine the expected counts:

On-time parts

	over weight	not over weight
allowable scrap	3 (.04)	57 (.32)
excessive scrap	1 (.03)	4 (.06)

Late parts (includes machine down time)

	over weight	not over weight
allowable scrap	3 (.06)	19 (.30)
excessive scrap	3 (.05)	10 (.14)

The system has just returned to production following maintenance. We question whether there is evidence that the current sample differs materially from past experience. If so, we need to know if there seems to have been improvement.

a. Are all expected counts at least three (a rule of thumb sometimes used)?

Since the sample size is 100, expected counts in the eight cells are just 100 times the parenthesized probabilities. For example, we expect 4 in the upper left cell and 32 in the cell just to its right. The smallest expected count is in fact 3.

b. Chi-square statistic **For example, the upper left cell contributes $\frac{(3-4)^2}{4}$ = 0.25 to the chi-square statistic 29.257.**

c. df **8 - 1 = 7**

d. P-value (use your calculator and check against table entries)

0.0001298

e. Is P-value small enough to convince you that production is different from past experience? If so, what changes seem to have occurred and do these appear to be favorable or mixed? **Either production has changed or an event of probability 0.0001298 has occurred (the chi-square has produced a small P-value without cause, just by "luck of the draw"). It seems there could well have been a change. For on-time parts there has been a great improvement in the number of parts having allowable scrap and not being over-weight. Other improvements are noted while there seem to be no important degradations of product.**

Suppose the cell probabilities had been changed (in parentheses) as below.

On-time parts

	over weight	not over weight
allowable scrap	3 (.04)	57 (.33)
excessive scrap	1 (.02)	4 (.06)

Late parts (includes machine down time)

	over weight	not over weight
allowable scrap	3 (.06)	19 (.33)
excessive scrap	3 (.02)	10 (.14)

f. Merge cells, as makes most sense to you, bringing all expected counts up to at least 3.

Two cells have expected counts below 3, on-time excessive scrap over weight and its late counterpart. If we merge these into a single cell the expected count for that cell becomes 2 + 2 = 4. The only possible downside to such a merge would be if it were crucial to see if these two cells contribute importantly to the issue of whether there is departure from expected counts. Particularly in view of the rarity of these cells it seems they do not, and anyway there is a much larger departure from what is expected happening elsewhere in the table.

g. Chi-square statistic after merge

For example, the merged cell contributes $\frac{((1+3)-(2+2))^2}{2+2}$ = 0 to the after-merge total chi-square of 26.9535.

h. df after merge

There are now only 7 cells so df = 7 - 1 = 6.

i. P-value after merge

0.000147741

j. Is P-value after merge small enough to convince you that production is different from past experience? If so, what changes seem to have occurred and do these appear to be favorable or mixed?

The P-value does seem convincingly small.

k. By comparison with your findings (e) has the need to merge interfered with any important conclusions you were able to make when no merge was needed?

The same comments apply as when we had no need to merge. This is because we merged two small count-cells that were not a factor in the improvements seen after the equipment underwent maintenance. The slight loss of detail due to the merge was irrelevant to the improvements noted.

3. A gene for flower color has the following visible outcomes

AA	**red flower**
aA or Aa	**pink flower**
aa	**white flower**

If there is random mating of flowers there will be a p in [0, 1] for which the population distribution takes the form

AA	aA or Aa	aa
p^2	2 p (1-p)	$(1-p)^2$

Suppose a random sample of 100 flowers finds

AA	aA or Aa	aa
41	38	21

a. Estimate p = $\frac{\#\text{ letters } A \text{ in the population}}{\#\text{ letters in the population}}$ by

$$\hat{p} = \frac{\#\text{ letters } A \text{ in the sample of 100 flowers}}{\#\text{ letters in 100 flowers}} = \frac{2\times 41 + 38}{2\times 100} = 0.6$$

b. From (a) and the distribution determine the expected counts for AA, aA or Aa, aa.

	AA	**aA or Aa**	**aa**
expected	**36**	**48**	**16**

for example, the expected count for cell AA is 100 $.6^2$ = 36.

c. Chi-square statistic

For example, cell AA contributes $\frac{(41-36)^2}{36}$ = 0.694444 to the total chi-square 4.34028.

d. df

Were all expected counts known it would be df = 3-1 = 2. However, in order to estimate expected counts we estimated $\hat{p}$. The price for doing so is the loss of one df. So the df = 1.

e. P-value
0.0372208

f. Does there seem to be strong evidence against what is expected in random mating?
While .0372 is on the small side it cannot be considered truly rare and might have happened by pure chance even if random mating was taking place. I would be cautious but open to further experiments to more confidently decide the matter. Notice that in this example there is no need for genetic analysis of the flowers in order to determine gene-types. Gene types are directly deduced from flower color. So we can directly get at the matter of whether the flowers are consistent with random mating.

4. Chi-square test of independence. A random sample of 113 customers of a large resort is classified according to number of nights and number of beds.

	1 night	2 nights	longer
1 bed	29	31	9
2 beds	12	19	5
more	0	6	2

a. Determine the marginal counts and from them the *expected counts* under the model that the number of beds is statistically independent of the number of nights.

	1 night	2 nights	longer	
1 bed	**25.0354**	**34.1947**	**9.76991**	69
2 beds	**13.0619**	**17.8407**	**5.09735**	36
more	**2.90265**	**3.9646**	**1.13274**	8
	41	56	16	**113**

b. Chi-square statistic

For example, cell "1 bed and 1 night" has expected count $\frac{41 \times 69}{113}$ = 25.0354. It contributes **$\frac{(29 - 25.0354)^2}{25.0354}$ = 0.627833 to the total chi-square 5.76211.**

c. df **(r-1)(c-1) = (3-1)(3-1) = 4**

d. P-value **0.217632**

e. Is there much reason to question independence of the number of beds from the nights stayed?
Not at all. Around 22% of the time the observed pattern or one in worse agreement with the hypothesis of independence would occur just by chance alone, if number of beds were independent of number of nights stayed.

1

```
1 - CDF[ChiSquareDistribution[5], 17.3072]
```

```
0.00395257
```

```
oshirt = {30, 67 , 106, 319, 430 , 148}
```

```
{30, 67, 106, 319, 430, 148}
```

```
oshirt.{1, 1, 1, 1, 1, 1}
```

```
1100
```

```
pshirt = {.02 , .04 , .09, .31, .40 , .14}
```

```
{0.02, 0.04, 0.09, 0.31, 0.4, 0.14}
```

```
pshirt.{1, 1, 1, 1, 1, 1}
```

```
1.
```

```
eshirt = pshirt 1100
```

```
{22., 44., 99., 341., 440., 154.}
```

```
Apply[Plus, (oshirt - eshirt)^2 / eshirt]
```

```
17.3072
```

2

```
eparts = {4, 32, 3, 6, 6, 30, 5, 14}
```

```
{4, 32, 3, 6, 6, 30, 5, 14}
```

```
eparts.{1, 1, 1, 1, 1, 1, 1, 1}
```

```
oparts = {3, 57, 1, 4, 3, 19, 3, 10}
```

```
{3, 57, 1, 4, 3, 19, 3, 10}
```

```
oparts.{1, 1, 1, 1, 1, 1, 1, 1}
```

```
100
```

```
Apply[Plus, (oparts - eparts)^2 / eparts] 1.
```

```
29.2574
```

```
1 - CDF[ChiSquareDistribution[7], 29.257440476190474`]
```

```
0.000129852
```

2 (merge)

```
epartsR = {4, 33, 4, 6, 6, 33, 14}
```

```
{4, 33, 4, 6, 6, 33, 14}
```

```
epartsR.{1, 1, 1, 1, 1, 1, 1}
```

```
100
```

```
opartsR = {3, 57, 4, 4, 3, 19, 10}
```

```
{3, 57, 4, 4, 3, 19, 10}
```

```
opartsR.{1, 1, 1, 1, 1, 1, 1}
```

```
100
```

```
Apply[Plus, (opartsR - epartsR)^2 / epartsR] 1.
```

```
26.9535
```

```
1 - CDF[ChiSquareDistribution[6], 26.953463203463205`]
```

```
0.000147743
```

3

```
oflwr = {41, 38, 21}
```

```
{41, 38, 21}
```

```
oflwr.{1, 1, 1}
```

```
100
```

```
pHAT = (2 × 41 + 38) / 200.
```

```
0.6
```

```
pflwr = {.6^2, 2 × .6 × .4, .4^2}
```

```
{0.36, 0.48, 0.16}
```

```
eflwr = pflwr 100
```

```
{36., 48., 16.}
```

```
Apply[Plus, (oflwr - eflwr)^2 / eflwr] 1.
```

```
4.34028
```

```
1 - CDF[ChiSquareDistribution[1], 4.340277777777775`]
```

```
0.0372209
```

4

```
oresort = {29, 31, 9, 12, 19, 5, 0, 6, 2}
```

```
{29, 31, 9, 12, 19, 5, 0, 6, 2}
```

```
oresort.{1, 1, 1, 1, 1, 1, 1, 1, 1}
```

```
113
```

```
MatrixForm[Outer[Times, {69, 36, 8}, {41, 56, 16}] / 113.]
```

$$\begin{pmatrix} 25.0354 & 34.1947 & 9.76991 \\ 13.0619 & 17.8407 & 5.09735 \\ 2.90265 & 3.9646 & 1.13274 \end{pmatrix}$$

```
eresort = Flatten[%]
```

```
{25.0354, 34.1947, 9.76991, 13.0619, 17.8407, 5.09735, 2.90265, 3.9646, 1.13274}
```

```
Apply[Plus, (oresort - eresort)^2 / eresort] 1.
```

```
5.76211
```

```
1 - CDF[ChiSquareDistribution[4], 5.762109118567894`]
```

```
0.217632
```