1. **Lottery data.**

   a) The frequencies appear to vary greatly, with 0's least preferred. The digits 1 through 4 appear excessively popular.

   b) Ho: Players do not favor some digits over others
   Ha: Players favor some digits over others.
   $\chi 2$=15.6, P=0.0757 (G=15.956, P=0.0678), df=9
   Since P>0.05, do not reject Ho.
   This does not mean that null hypothesis is true! Only that we haven't got enough evidence to reject it.

   c) They are different <u>approximations</u> to the theoretical chi-square distribution.

   d) The results of the chi-square test (quantitative) don't fit the visual (qualitative) impression in the frequency table. One measure of uncertainty is the probability of a Type I error, which is given by the P-value for the test (e.g., P=0.757). Another measure of uncertainty is given by the probability of a Type II error. However, this probability is unknown.

   e) Df=9. We lose a degree of freedom because of an imposed constraint: the total of the expected frequencies must equal that of the observed frequencies.

   f) We need to ensure that none of the expected frequencies are <1 and no more than 20% are <5. If this rule is violated then we need to combine categories.

2. **Physical Map of the Human Genome 1998**

   a) Chromosomes 1, 2, 3 have the most genes, 18, 21, 22 have the most genes.

   b) Obtain the expected frequencies by multiplying the "Proportion of total genome" variable by 30075 (the sum of all values in "Observed No. genes" column). Larger chromosomes indeed have more genes, as you might expect.

   c) The X chromosome has the greatest apparent deficiency, followed by chromosomes 4 and 18. Chromosomes 13, 5, and 8 are also strongly deficient. Chromosome 19 has the greatest excess, followed by 17, 1, 11, and 22. [Results are the same, more or less, if we compute the ratio of observed number of genes to the expected number for each chromosome.]

   d) Ho: Gene number on chromosomes is proportional to chromosome size
   Ha: Gene number on chromosomes is proportional to chromosome size
   $\chi 2$= 1682.42, df=23−1=22, P=0.0 (G= 1661.88, df=22, P=0.0)
   The P-value is so small that it is effectively 0.
   Since P<0.05, reject Ho.

3. Mount Everest mountaineer survival and supplemental oxygen

a) Survival is slightly higher in mountaineers that used supplemental oxygen.

b) Ho: mountaineer survival is not associated with use of oxygen
Ha: mountaineer survival is associated with use of oxygen
$\chi 2$=7.694, P=0.0055 (G=5.747, P=0.0165), df=1
Since P<0.05, reject Ho. Mountaineer survival **is** associated with use of supplemental oxygen.
This is also indicated by the results of the Fisher exact test: P=0.0128. Mountaineers using supplemental oxygen have a higher probability of survival during descent from Everest.

c) The Fisher's exact test provides an exact P-value, whereas the Pearson $\chi 2$ and G tests yield an approximate P-value.

d) You should obtain the same expected frequencies as JMP IN (to get these you will need to select the "Expected" option from the contingency table output window). However, in class you have been instructed to include the Yates' correction when df=1, but JMP IN bypasses this step. Hence you will not obtain the same values for $\chi 2$ and G as JMP IN.

e) One of the four expected frequencies is less than 5, which violates our rule of thumb. One cannot combine categories in a 2x2 table (there would be none left for a test), so the best thing to do is to verify the results with the Fisher exact test. In this case the conclusion (reject Ho) is the same.

f) It is likely that mountaineers on the same team are not independent, because mountaineers on the same team may have more similar chances of survival than mountaineers chosen at random. Mountaineers on the same team descended on the same or nearby dates under similar weather and equipment conditions. They may have helped each other (although an "every man for himself" ethic usually applies at the top of Everest). In this case the <u>team</u> data set is more appropriate.

g) A contingency test on the team data (details not shown) does not allow us to reject the null hypothesis that there is no association between mountaineer survival and use of supplemental oxygen (the P-value from the Fisher exact test is 0.4698). This does NOT mean that supplemental oxygen has no influence on the probability of survival. It just means that the existing data provide no evidence of such an association (never accept the null hypothesis).

h) The team data from K2 allow us to reject the null hypothesis (details not shown). The P-value from the Fisher exact test is 0.0227.

4. student_data.jmp

a) Use the chi-square goodness of fit test:

Ho: men and women occur with equal frequency in the Bio300 Jan 2001 class.
Ha: one sex is more common that the other in the class.
$\chi 2=23.603$, $P<0.001$ ($G=24.151$, $P<0.001$), df=1 (note that JMP IN does not use the Yates correction).
Since $P<0.05$, reject Ho. The sexes do not occur with equal frequency (female students predominate).

b) Use a contingency test. The result (not shown) is that the null hypothesis of no association between handedness of students and their mothers is not rejected.

c) -

d) Use a contingency test (results not shown). The answer is: indeed it does!