Stat 10, Section 1A
Thursday, April 22nd, 2010

David Diez

1 Lab 3

Some common errors:

- When asked to *compare distributions*, you should compare their (1) shape, (2) center, (3) spread, and (4*) any unusual features. It is not enough to say

  > This birth weights of babies from smoking mothers had a median of 7.06 pounds and for babies from non-smoking mothers it was 7.31 pounds.

  You should compare them:

  > This birth weights of babies from smoking mothers had a median of 7.06 pounds, which was *0.25 pounds lower* than that of babies from non-smoking mothers.

  In a similar vein, compare the shape and spread as well

  > Both distributions are slightly left skewed and each have nearly the same IQR of about 1.6 and 1.7 pounds.

  If there are unusual features, you can compare those as well.

- In statistics, we never actually **prove** anything. That is the realm of mathematics (TV pundits also claim it is their realm). In statistics we *collect evidence and evaluate whether the evidence supports one position or the other.* A few common ways to describe findings are below. The first two are more intuitive and in non-statistical language, but the last is more commonly used.

  – The data provides convincing evidence that a mother's smoking habit and the birth weight of her baby are dependent.

  – Based on our analysis, it is implausible that the difference in average birth weight between the smoking and non-smoking groups is due to chance alone, therefore the mother's habit and baby's birth weight are dependent.

  – The difference in average birth weight between the smoking and non-smoking groups is statistically significant so these variables are dependent.

  This isn't a failure in our methods; we sometimes get unhelpful data just by chance and this is a reality we must face.

- Be extremely cautious about using causal language (e.g. *smoking affects birth weight* or *the effect of smoking on birth weight...*). In statistics, we can confidently infer causal conclusions if (and only if!) the data is from a randomized controlled experiments and we actually find a significant dependence between the explanatory and response.

2 Reminder: extra component for all future labs

At the top of your lab write-up, summarize the lab by completing (1)-(5) below. Provide a single sentence answer for each part, and the total answer for all parts should be no more than 150 words.

1. Summarize the question being answered or the problem being solved in the lab.
2. Briefly describe the data and study type.
3. Briefly describe the analysis techniques that you applied (in plain language).
4. Provide a conclusion in plain language.
5. Describe a real world decision that might be influenced by your conclusion.

Recommendation: Have a friend who is not in statistics read your answer, which should be in plain language.

TB Lab Sample:

- Does streptomycin help in treating tuberculosis (TB)?
- An experiment was conducted with TB patients, where 55 were assigned to receive streptomycin (4 died) and 52 were assigned to receive only bed rest (14 died).
- To see whether the difference in death rates between the streptomycin and bed rest groups provides convincing evidence that streptomycin works, we used computer simulations to see what sort of difference we would get from chance alone.
- We found that our actual result wouldn't often occur from chance alone, so the experimental data provide convincing evidence that streptomycin helps treat TB.
- Based on the data and analysis, I would prefer to treat future patients with streptomycin over simple bed rest.

3 Quiz problems

(2) Defective Tires The probability distribution of x, the number of defective tires on randomly selected automobiles at a certain inspection station, is given in the accompanying table. x is the number of defective tires.

x	0	1	2	3	4
$P(x)$	0.54	0.16	0.06	0.04	0.20

After you calculate the expected number (the mean) of defective tires, determine the probability that X exceeds the expected value.

(4) (This also covers (3).) In a recent issue of Science magazine, we read about a new computer-based test for ovarian cancer, "clinical proteomics" that exams a blood sample for the presence of certain patterns of proteins. Ovarian cancer, though dangerous, is very rare, affecting only one in 5000 women. The test is highly sensitive, able to correctly detect the present of ovarian cancer in 99.97% of women who have the disease. However, it is unlikely to be used as a screening test for the general public, because the test gives a false positive 5% of the time. Draw a tree diagram or table and determine the probability that a woman who tests positive using this method actually has ovarian cancer.

4 Probability worksheet from Moodle