Instructions for final data analysis project, Stat 200H

Work in groups. Prepare a presentation that will last between 10 and 15 minutes. The group members should take turns in presenting various parts of their analysis, and be ready to answer questions on all aspects of what the group produced (in other words, don’t split the work, do it together). You should prepare transparencies (I suggest you don’t exceed 7), as well as a print out to be handed in.

You can use any of the tools we have learned, such as numerical summaries, graphs, confidence intervals, tests, regression techniques. Whenever you use an inferential tool (confidence intervals, tests) mention the assumptions it involves and when possible comment on whether the data is consistent with these assumptions. Make sure the presentation provides a good summary of the line of thought you developed, and the tools you employed – in particular, use plots and appropriately edited output from Minitab.

Mon Dec 8: Faculty Salary

Data set: faculty_salary.mtv  Presenting groups: 3, 9, 8.

This data set contains information on a sample of 171 faculty members of a private college, collected in 1991. “1991$” is the current yearly salary in dollars, “seniority” is a variable that combines length of employment at the college and a measure of the experience of the individual at the time of hiring (there is no obvious measurement unit for this composite variable), “FullProf” is a categorical variable coded as 0 = not full professor (i.e. instructor, assistant or associate professor), and 1 = full professor, and “Gender” is a categorical variable coded as Female and Male. Your main aim in analyzing this data is to understand whether there is evidence of discrimination against female faculty in this college. Make sure you address the following questions:

1. What is the proportion of females in this sample? Do we have evidence that the college employs fewer females than males?
2. How does the proportion of females among full professors compare to that of females among non-full professors in this sample? Do we have evidence that females are predominantly employed in lower ranks in the college?
3. How does the average salary of females compare to that of males in this sample? Do we have evidence that female salaries are on average lower than male salaries at the college?
4. How does the average seniority of females compare to that of males in this sample? Do we have evidence that females’ seniority is on average lower than males’ seniority at the college?
5. What is the relationship between salary and seniority? Is it the same or different for males and females? Does seniority explain, at least in part, the differential in salary between males and females?
This data set contains information on a sample of 148 houses in city A. “Price” is the price in dollars, “Area” is the square footage of the house, “2Bath+” is a categorical variable coded as 0 = the house has fewer than 2 bathrooms, and 1 = the house has two or more bathrooms, and “Rooms” is the number of rooms in the house. Your main aim in analyzing this data is to understand the determinants of prices. Make sure you address the following questions:

1. What is the proportion of houses with two or more bathrooms in this sample? Do we have evidence that the majority of the houses in city A have two or more bathrooms?
2. How does the average price of homes with two or more bathrooms compare to that of homes with fewer than two bathrooms in this sample? Do we have evidence that homes with two or more bathrooms are on average more expensive than those with fewer than two bathrooms in city A?
3. How does the average area of homes with two or more bathrooms compare to that of homes with fewer than two bathrooms in this sample? Do we have evidence that homes with two or more bathrooms are on average larger than those with fewer than two bathrooms in city A?
4. What is the relationship between price and area? Is it the same or different for houses with fewer than, or two or more bathrooms? Does area explain, at least in part, the differential in price between houses with fewer than, or two or more bathrooms?
5. Predict the price of a home of 2800 sq. feet, with two+ bathrooms.
6. Looking at the regression of price on area (overall, and/or separately for the two types of houses), there is a striking feature that contradicts one of the standard assumptions for regression modeling – it can be noticed both from the scatter plot and from the residual plot, what is it? Try transforming the response to Log(Price), and repeating the fit of the regression line (overall, and/or separately for the two types of homes). Does the situation improve?

(I did not mention the variable “Rooms” in the questions above. You can perform analyses involving it if you want).
Fri Dec 12: Parenting behavior of fish

Data set: fish_parenting.mtw

This data set contains information on a sample of 80 fish of species A. “Gender” is a categorical variable coded as M (male) and F (female). “Guarding” and “Fanning” are the times (in minutes) spent guarding the nest and fanning the eggs, respectively. “pH” is the pH level of the water when each observation was collected. Your first aim in this analysis is to investigate the relationship between the two types of parenting behavior, and between parenting behavior and gender. Make sure you address the following questions:

1. How does the average time spent guarding the nest compare between male and female fish in this sample? Do we have evidence that in species A males spend on average more time guarding the nest than do females?
2. How does the average time spent fanning eggs compare between male and female fish in this sample? Do we have evidence that in species A females spend on average more time fanning eggs than do males?
3. What is the relationship between time spent guarding the nest and time spent fanning the eggs? Is it the same or different for males and females?

It is claimed that water pH level has an adverse effect on parenting behavior, in the sense that pH decreasing below the “natural level” pH=7 induce a decrease in the time spent in parenting activities such as guarding the nest and fanning eggs. Your second aim is to ascertain whether the data supports this claim. I suggest you produce the smooths for “Guarding” on “pH”, by gender, and for “Fanning” on “pH”, again by gender. These are quite informative.