STT200 Recitation assignment for 2-12-08

You will receive a fresh copy in recitation. It is to be submitted before leaving.

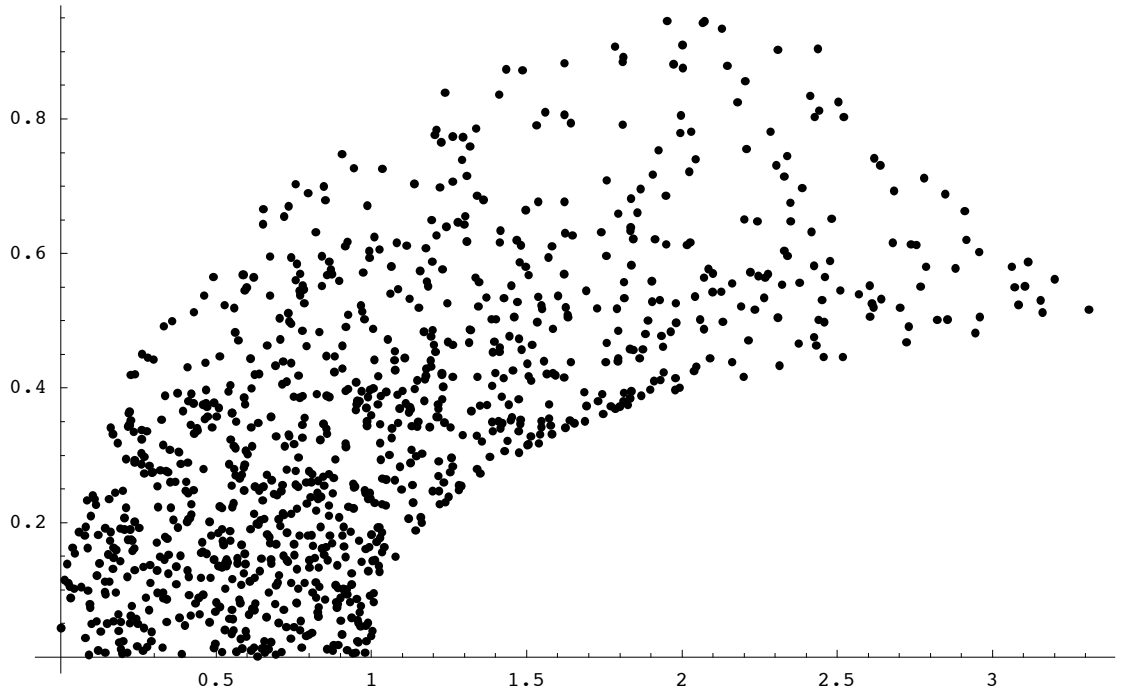
This material below covers:

a. Chapter 8 and the supplement to chapter 8 from lecture 2-6-08 (posted).
b. Smoothing data (alternative to histogram) to be covered in lecture 2-11-08.
c. Bootstrap re-sampling principle to be covered in both lectures next week.

It will be tested on exam 2.

1. Regression. Below are the statistics calculated for the pictured data.

regrstats[x,y] = {1.05, 0.34, 0.68, 0.21, 0.651, 0.21}
= {xavg, yavg, sx , sy , r , slope}



1a. Place a **plus sign** at the point of averages in the plot and label the axes x, y.
point of averages =

1b. Place **another plus sign** at (xavg + sx, yavg + r sy) and draw in the regression line.
(xavg + sx, yavg + r sy) =

1c. Determine the numerical values of the y-intercept and slope of the regression line.

y-intercept = yavg - xavg times slope =

slope =

1d. From (1c) determine the predicted value y for x = 1.7 (express numerically but do not reduce).

1e. Determine the residual y - (b0 + b1 x) for the point (x, y) = (1.5, 0.7).

1f. Overall, about what percentage of $(sy)^2$ is accounted for by regression of y on x?

1g. In the plot above, draw in an ellipse representing the general appearance of a list of pairs $(\overline{x}, \overline{y})$, each of which is obtained from a sample of n = 30. These samples of 30 are independently gathered from the population above. You have not been shown how to obtain the exact scale for this ellipse but your sketch should be properly placed in relation to the point of averages of the population, and the population regression line, and suggest that $(\overline{x}, \overline{y})$ has relatively small variation relative to the individual population (x, y) pairs.

2. Approximate sampling distribution of the sample mean. This material is not in Chapter 8. We introduce it now to help prepare you for what is to come. The list of all $\overline{x}$ produced by a random with-replacement sample of large enough n is **approximately normal** with

a. mean of all $\overline{x}$ (over all samples of n) = population x-average = μ_x
b. standard deviation of $\overline{x}$ (over all samples of n = 30) =

$$\frac{\text{population standard deviation of } x}{\sqrt{n}} = \frac{\sigma_x}{\sqrt{n}}$$

A population of patients is scored with x = satisfaction level with hospital stay measured six months following their release from the hospital. Reasoning that it will be extremely costly to survey all patients in this way it is decided to select a random sample of n = 100 patients. As an estimate of the population average satisfaction score μ_x we choose the sample mean $\overline{x}$ of our sample of 100 patients. If the population mean and standard deviation are μ_x = 7.2 (scale of 10) with σ_x = 2.1 sketch the approximate distribution of the

sample mean $\overline{x}$ (i.e. normal with mean μ_x and standard deviation $\frac{\sigma_x}{\sqrt{n}}$).

In your sketch above around 95% of the time $\overline{x}$ is within the distance $\pm 1.96 \frac{\sigma_x}{\sqrt{n}}$ of the population mean μ_x (shade the relevant area!).

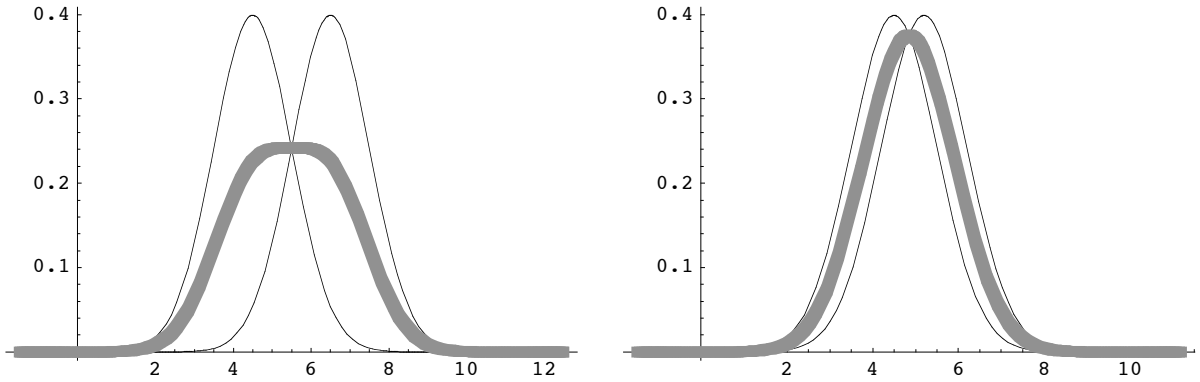
When we process **our sample of n = 100** and report

$$\overline{x} \pm 1.96 \pm \frac{s_x}{\sqrt{n}}$$

we are effectively estimating the true Margin of Error $\boxed{1.96 \frac{\sigma_x}{\sqrt{n}}}$ by the sample estimated Margin of Error $\boxed{1.96 \frac{s_x}{\sqrt{n}}}$. Sometimes in press reports they will say margin of error when they really are reporting estimated margin of error.

A sample of n = 100 is selected at random from the population with $\mu_x = 7.2$ (scale of 10) and $\sigma_x = 2.1$. I've actually selected such a sample from a particular population having mean 7.2 and standard deviation 2.1. My sample of 100 has $\overline{x} = 6.94$ and $s_x = 2.23$. Report the standard 95% confidence interval $\overline{x} \pm 1.96 \pm \frac{s_x}{\sqrt{n}}$ for my sample. Around 95% of the time such an interval prepared from the sample should cover the unknown population mean μ_x. Does this example cover μ_x?

3. Smoothing data. The idea is to place a bell curve over each data value and plot the average height of these curves. For two data values you simply pick a point on the horizontal axis and mark midway between the two curves above that point. Repeat this for several points on the horizontal axis and join those midpoint values in to a smooth curve.



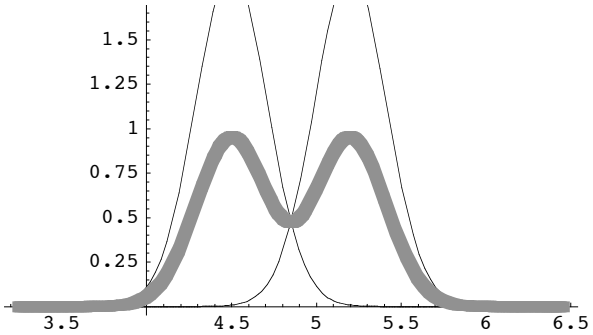
The standard deviation of the bell curves used is called "bandwidth." Taking a larger bandwidth flattens-widens out the bell curves. Generally, we want small bandwidth to reveal detail but not so small as to pursue illusory detail that may not be reproduced for a different sample of data. This is similar to the situation with histograms where choosing the bins very narrow may result in very small bin counts. Why smooth data rather than use a histogram? Reasons:

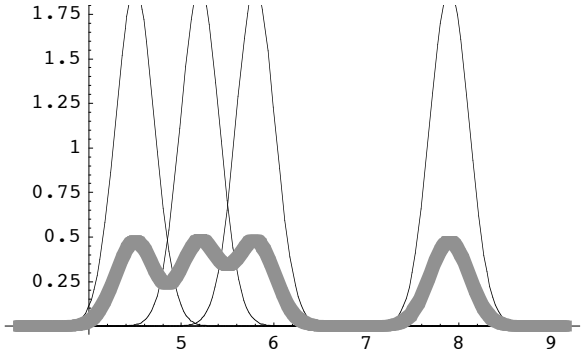
a. A smooth curve prepared from a random sample of a population tends to more accurately represent the analogous smoothing of the population than does a sample histogram represent the analogous histogram prepared on the population.

b. Small changes in the locations or widths of bin intervals tend to produce greater changes in a histogram than do small changes in bandwidth affect the curve.

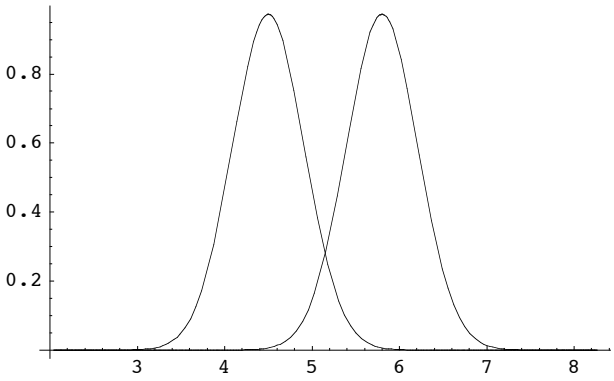
Points (a) and (b) of course require assumptions but (a) (b) are the general motivations for using smoothing. The issues of which bandwidth to choose will not be discussed at this time.

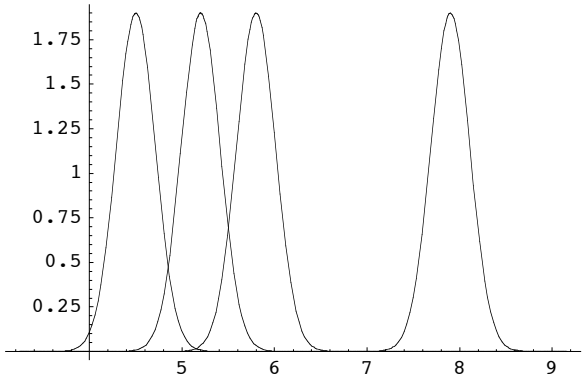
If there are four data values you may average each of two pairs of curves then average those two average curves. Here are examples of 2 and 4 curves.





Try your hand at these:





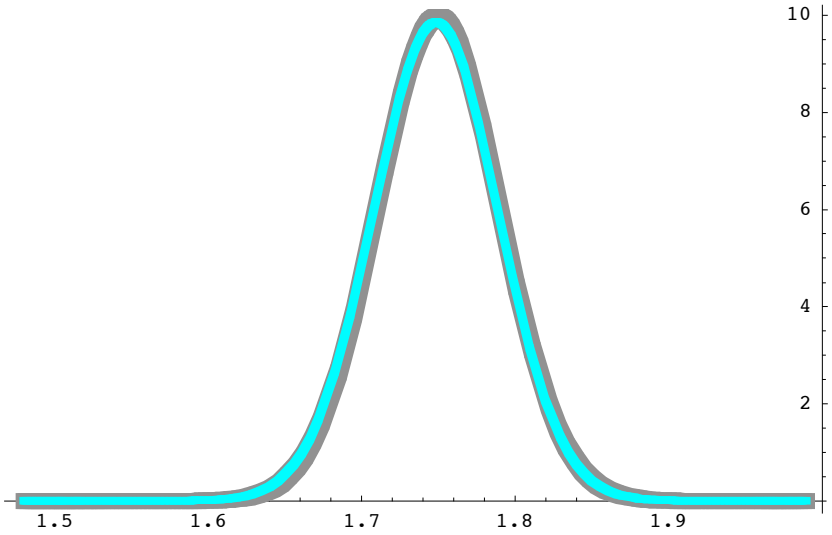
3. Bootstrap resampling. This topic is not seen in Chapter 8. We will go over it Monday 2-11-08. As seen in (2), Statisticians are interested in the sampling variation of estimators of population quantities. For instance, an economist may estimate the average spending per household using $\overline{x}$ from a random sampling of households. This estimate will vary subject to the randomness of the sample. May we use the data from our sample of households to estimate this variation as well?

One example of this is the Margin of Error $\boxed{1.96\ \frac{s_x}{\sqrt{n}}}$ for the estimate $\overline{x}$ of the population mean. But what if we wish to estimate the population median or Q1, or some other feature of the population? Classical solutions to this problem treated each estimation as an individual problem with its own formula for calculating the relevant margin of error. You will see many such specialized formulas in your textbook, each appealing to appropriate tables.

In 1978-79 Brad Efron of Stanford University initiated a general principle applicable to a large range of estimators and substituting large computer calculations for the specialized formulas. The idea is that variations of a sample from the population are rather well

approximated by variations of bootstrap samples from the sample. What is a bootstrap sample? It is just a with replacement sample of n from our sample of n. As usual, the sample size n must not be small.

To illustrate, below is a smoothed plot of the means $\overline{x}$ of each of many independently gathered samples of n = 1000 (shown in grey). It is centered on the population mean μ_x. Slightly displaced from the grey plot is an almost identical plot but centered on the sample mean $\overline{x}$ (shown in blue). Efron's remarkable insight is that our sample of 1000 can show us this blue picture. All we do is produce many samples of 1000 FROM OUR SAMPLE OF 1000 and smooth the plot of these "bootstrap" sample averages. Since our sample of 1000 is in the computer we can draw as many samples of n = 1000 (from our original sample of 1000) plot their sample means (blue curve). It is like having the grey curve but centered on our sample mean $\overline{x}$ instead of the population mean μ_x. The key idea is that the margin of error seen in the blue curve can tell us the margin of error of the grey curve! For this example, the grey curve would essentially be the normal with mean μ_x and standard deviation $\frac{\sigma_x}{\sqrt{n}}$ (determined from the population). The blue curve would essentially be the normal with mean $\overline{x}$ and standard deviation $\frac{s_x}{\sqrt{n}}$ (determined from the sample). Efron's bootstrap gets it without any formulas and by a method that works in the same way for very many other estimators as well. Out with the formulas? Well to some degree, but every idea has its caveats and overhead. For example, a formula is easily communicated and evaluated. Bootstrap requires a computer running many thousands of bootstrap samples of n from our original sample of n.



distribution of means (samples of n = 100)
samples of n = 100 from population (grey)
samples of n = 100 from our original sample of n = 100 (blue)

Suppose you have a sample of n = 100 and from this sample you find
sample mean $\overline{x}$ = 22.79 (estimate of population mean μ_x)
sample standard deviation s_x = 4.68 (estimate of population standard deviation σ_x

Sketch the normal curve with mean 22.79 and standard deviation 4.68 / $\sqrt{100}$. It is known that this will also be very close to the blue curve produced by Efron's method. This blue curve tells us the margin of error for $\overline{x}$ based on our sample of 100.