WMS5 Exercises

5.6 , p198
5.9, 5.11 p199
5.18, 5.21, p208
5.36, 5.37, p 214
5.42, p 215
5.68, p235
5.94, 5.95, p 247

Homegrown Exercises

Q1 See on the class web site Galton's 1886 data about the heights of parents and their adult children (more details are given under Resources on Session 1 of course 678)

A • Add axes labels and a chart title to the partly completed figure. *(Click on each axis, and type in the text. Or print out the graph and add the text by hand -- the point is not the aesthetics, but the understanding! and use English or French, rather than mathematical shorthand)*

• Add onto the graph the "marginal" distribution of the heights of children *[to make it more visible,. and if you have time, you can use the 3-D View options under the Chart Menu]*

B Look at the formulae behind in the cells in the last table in the spreadsheet.

• Make a 3-D diagram of these entries, and label the axes appropriately (again, in English or French, rather than in mathematical shorthand).

• Once you have printed them out, add in a second set of labels -- using "mathematical-statistical" notation -- as per Chapter 5)

Q2 See on the class web site Galton's 1886 data on the heights of 100 sets of parents [randomly selected from the 205 sets who contributed to the data used for Q1]

A In Galton's 1889 book, *Natural Inheritance*, as a preliminary step before doing his main analyses (where he averaged the heights of the two parents), he analyzed several factors that might influence marriage selection. Pages 84-89 are reproduced on the web page. One such analysis is discussed on his page 87, where he argues that if men and women married "at random as far as stature was concerned," the variability Q† in a group of couples, each couple consisting of a pair of *summed* statures, would be $\sqrt{2}$ × the (inter-individual) variability Q of the individuals used to construct the pairs.

†In 1886, standard deviations, or functions of them, were messy to work with, so Galton worked with what he called Q, which is half the distance between the 25th and 75th percentile, as a measure of variability [since the distribution of heights was symmetrical, Q is a simple linear scale factor of the standard deviation -- see Q3 below]. Today, using the standard deviation (so christened by Pearson in 1893), or its square, the variance, we can translate Galton's argument to say that the *variance* in a group of couples, each couple consisting of a pair of *summed* statures, would, under random mating, be approx 2 × the (inter-individual) variance of the individuals used to construct the pairs. i.e.,

Var(Father's height + Mother's height)
= Var(Father's height) + Variance(Mother's height) .

[this is theorem 5.12 **b** with the covariance set to zero]

• Compute the variance of the sums, and compare it with the sum of the variances.

Galton "tried the question in another but ruder way" [see bottom of p 87, and top of p 88].

• Instead of dividing them as he did, divide the men into just two: taller or shorter than the median for men; and likewise for women. Now count how many couples were (a) both shorter, or both taller, than the median and (b) the reminder [one taller and one shorter]. If indeed marriage was "at random" with respect to stature, what should the proportion in (a) be? How far are the data from this proportion?

• from your two analyses, do you agree with Galton that "I am therefore content to .. regard the Statures of married folk just as if their choice in marriage had been wholly independent of stature" (p 88) ?

B Today we measure the degree of (linear) "related-ness" of two random variables by the correlation coefficient, i.e., the "covariance" scaled so it falls between -1 and 1. (WMS5, p224).

• Calculate the covariance from (a) 1st principles, using definition 5.10 -- and using "averages" in the data as proxies for "expected" values (b) the "shortcut" in Theorem 5.10. From the covariance, and the scaling formula given in the middle of page 224, calculate the (sample) correlation coefficient [i.e. use sample statistics to estimate population parameters].

• Repeat the calculations using the "covar" and "correl" functions in Excel or other statistical calculator.

C Maybe more interesting to the 100 couples is the difference in (rather than the sum of) their heights.

• create a new variable (column) for the difference, draw a histogram showing its distribution, and calculate its mean and standard deviation.

• Derive a theoretical expression for 9i) the expected value (ii) [from rule **b** in Theorem 5.12 (p 228)] the variance, and (iii) [from its definition] the standard deviation of the difference in heights.

• Substitute the observed means and variances, and a zero value for the correlation, and -- assuming a Gaussian distribution of differences -- calculate the probability that a randomly chosen man would be shorter than a randomly chosen female.

• Repeat the calculation, but with the observed (non-zero) covariance/correlation found in B.

• Compare these 2 theoretical probabilities with the actual frequency of "man is shorter" couples among the 100.

Q3 As said in Q2, Galton used Q, or 1/2 the inter-quartile range, to measure variability.

• In a Gaussian distribution, what is the relationship between Q and the standard deviation: i.e., what multiple of the standard deviation is Q? ***[q. is good for getting used to navigating the Gaussian distribution via tables or functions]***