List of Definitions, Concepts, and Formulas for Math 125

Definitions and Concepts

Chapter 6

Measurement Error:

No matter how carefully it was made, a measurement could have come out a bit differently. If the measurement is repeated, it will come out a bit differently. By how much? The best way to answer this question is to replicate the measurement.

The SD of a series of repeated measurements estimates the likely size of the chance error in a single measurement.

Bias affects all measurements the same way, pushing them in the same direction. Chance errors change from measurement to measurement, sometimes up and sometimes down.

Chapter 7

Straight Lines:

Slope: The slope of a line equals the rise over the run: slope = rise/run.

The Equation for a Line:

The graph of the equation $y = mx + b$ is a straight line, with slope m and intercept b.

For example: $y = 7x - 24$ has slope =7 and intercept of -24 (that's a point on the line: $(0, -24)$).

Finding the Equation of a Line: Given the slope m and a point (c, d) on the line, to get the equation of the line, substitute c for x and d for y in the equation $y = mx + b$ and solve for b.

For example: The equation for the line with slope $m = -1.5$ and the point $(6, 13)$ will be $13 = -1.5(6) + b$. Then $13 = -9 + b$ and $b = 22$.

The equation is $y = -1.5x + 22$.

Chapter 8

The Scatter Diagram:

A plot of two-variable data on the x,y-plane is called a scatter diagram.

The Point of Averages:

A point on the x,y-plane showing the average of the x-values and the average of the y-values.

The Correlation Coefficient:

The correlation coeficient is a measure of linear association, or clustering around a line. The relationship between two variables can be summarized by

- the average of the x-values, the SD of the x-values.
- the average of the y-values, the SD of the y-values.
- the correlation coefficient r.

Correlations are always between -1 and 1, but can take any value in between. A positive correlation means that the cloud slopes up; as one variable increases, so does the other. A negative correlation means that the cloud slopes down; as one variable increases, the other decreases.

The SD Line:

A line through the point of averages that goes up or down (depending on whether the correlation coefficient is positive or negative) by one SD of y for each SD of x.

The slope is (SD of y)/(SD of x) or $-$(SD of y)/(SD of x).

Chapter 10

Examples of the Regression Method for Percentiles:

Example 1: A student's percentile rank on the midterm was 94%, and the correlation between midterm scores and final exam scores was 0.52. The scatter diagram is football-shaped.

Predict his percentile rank on the final exam.

Answer: The score on the midterm in standard units is 1.55. (See example 10 in Chapter 5.5 on pages 90 to 91).

There is 50% below 0 and 44% from 0 to z. That makes $-z$ to z about 88%, using the symmetry of the normal curve. So $z_x = 1.55$. Use the positive value for the 94th percentile.

Now mutliply z_x by r to get the y-value in standard units: $1.55 \times 0.52 = .806$, use 0.80.

Find the percentile rank for 0.80 by looking up -0.80 to 0.80: about 58%.

Finally, add half that area (0 to 0.80) to the area below 0. It will give the area below 0.80, and that is the definition of the percentile rank for 0.80.
The answer is $50\% + 58\%/2 = 50\% + 29\% = 79\%$.

We predict a percentile rank of 79% on the final exam for this student.

A double-check is that regression is toward the mean (50%). Indeed, 94% moved closer to 50%, as expected.

Example 2: A student's percentile rank on the midterm was 3%, and the correlation between midterm and final exam scores was 0.52. The scatter diagram is football-shaped.

Predict his percentile rank on the final exam.

Answer: The score on the midterm in standard units is -1.90.

Reason: There is 3% below $-z$ and 3% above z (by symmetry). In the middle there is $100\% - (3\% + 3\%) = 100\% - 6\% = 94\%$. For that area, z is about 1.90 and we use the negative value for the 3rd percentile. So, $z_x = -1.90$.

Now multiply z_x by r to get the y-value in standard units: $-1.90 \times 0.52 = -0.988$, use -1.

Find the percentile rank for $z = -1$ by looking up the area from -1 to 1: about 68%.

Finally, subtract 68% from 100% to get the area of both tails: 32%. We need only the lower tail for the percentile rank. By symmetry it equals 32%/2. The answer is 16%.

We predict a percentile rank of 16% on the final exam for this student.

A double-check is that regression is toward the mean (50%). Indeed, 3% moved closer to 50%, as expected.

Also see example 2 on page 166.

The Regression Effect:

In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test—and the top group will on average fall back. This is the *regression effect.*

Example: Applicants to a European prep school are required to take tests in various subjects. Applicants who score high on the mathematics test also tend to score high on the physics test. On both tests, the average score is 55; the SDs are the same too. The scatter diagram is football-shaped. Of the students who scored about 70 on the mathematics test:

(A) just about half scored over 70 on the physics test.

(B) more than half scored over 70 on the physics test.

(C) less than half scored over 70 on the physics test.

The correct answer is (C), because the average will fall back from 70 towards 55. With the average at, say 64, the students who scored over 70 will constitute a right tail starting at 70. Since the tail is less than 50% on the normal curve, (C) is the correct answer. Draw the picture.

The Regression Fallacy:

Thinking that the regresssion effect must be due to something important, not just the spread around the line, is the *regression fallacy.*

Example: Freshmen at a major university take precalculus and elementary physics together.

The average grade on the precalculus final was 70 points with an SD of 10 points, and the average grade on the elementary physics final was also 70 points with an SD of 10 points.

The correlation between final-exam grades in precalculus and in elementary physics was 0.60, and the scatter diagram turned out to be football-shaped.

For all the students who got 95 on the math final, their average grade on the physics final is estimated to be 85. And, furthermore, if one chose one of those students at random, it is reasonable to predict that his or her grade on the physics final is about 85.

This leads us to conclude that a grade of 95 on the math final is associated by the regression method with a grade of 85 on the physics final.

Now decide if the following two statements are true or false and explain.

> Since a 95 on the math final is associated with an 85 on the physics final, the best guess for the average on the math final of a student who got 85 on the physics final is 95.
>
> Also since a random student who got 95 on the math final is predicted to have an 85 on the physics final, a random student of all those who got 85 on the physics final must be predicted to have a grade of 95 on the math final. It is clear from the simple association of those two numbers.

Both statements are false. They presented the regression fallacy.

Reason: 95 on the math tended to lead to 85 on the physics, but the 95's were a much smaller group than the 85's. A small part of the 85's got 95 on the math. The rest of the 85's tended to get lower scores on the math exam. The statements confused *some* of the 85's on the physics test (the ones who got 95 on the math) with *all* of the 85's.

All of the 95's in math were associated with some of the 85's in physics, but the rest of the 85's in physics were not associated with 95 in math.

Chapter 11:

The Root-mean-square (r.m.s.) Error for Regression:

The distance of a point above (+) or below (−) the regression line is

$$\text{error} = \text{actual} - \text{predicted}.$$

The r.m.s. error for regression says how far typical points are above or below the regression line.

About 68% of the points on a scatter diagram fall inside the strip whose edges are parallel to the regression line, and one r.m.s. error away (up or down). About 95% of the points are in the wider strip whose edges are parallel to the regression line, and twice the r.m.s. error away.

Looking at Vertical Strips:

Homoscedastic Diagram: One in which all the vertical strips show similar amounts of spread.

Heteroscedastic Diagram: Vertical strips in the diagram show differing amounts of spread so the regression method is off by different amounts in different parts of the scatter diagram.

Suppose that a scatter diagram is football-shaped. Take the points in a narrow vetical strip. They will be off the regression line (up or down) by amounts similar in size to the r.m.s. error. If the diagram is heteroscedastic, the r.m.s. error should not be used for individual strips.

(See example 1 on pages 195 to 197.)

Chapter 12:

The Regression Line:

The regression line is a straight line that, for each value of x, gives the estimated average value of y (the result of the regression method).

Finding the Equation of the Regression Line:

First find the slope of the regression line using the formula

$$\frac{r \times \text{SD of } y}{\text{SD of } x}$$

Then pick a point on the regression line and substitute into the formula $y = mx+b$ the values (x, y) for that point. Since the regression line always goes through the point of averages and that point is often known, it is convenient to substitute the values of the average x and average y into that equation. **Don't substitute a point on the scatter diagram; points on the scatter diagram are generally not on the regression equation.**

Next, solve the equation for b. The equation is: $y = mx + b$ with m and b known numbers.

Do not use the suggestions on pages 202 to 207.

Do not follow the method of Example 1 on page 205. These ideas are very confusing and convoluted.

Example 1:

A group of 1,000 men had the follwing data (the scatter diagram is football-shaped):

$$\begin{aligned} \text{average height} \approx 69 \text{ in}, &\quad \text{SD} \approx 3 \text{ in} \\ \text{average weight} \approx 175 \text{ lb}, &\quad \text{SD} \approx 42 \text{ lb}, \quad r \approx 0.45 \end{aligned}$$

- Find the regression equation for predicting weight from height. Use it to predict the height of a 73-in. man.
- What is the slope of that equation? Explain how it could be intepreted in the context of predicting the weight of a man who is 8 inches heavier than his friend.
- A prediction made by this equation is likely to be off by about how many pounds?

Answer to above example:

- The slope is $\dfrac{0.45 \times 42 \text{ pounds}}{3 \text{ inches}} = 6.3$ pounds per inch.

  Plug in $(69, 175)$ into the equation $y = 6.3x+b$ to get $175 = 6.3\times 69+b$, $175 = 434.7+b$, and $b = 175 - 434.7 = -259.7$.

  Equation: $y = 6.3x - 259.7$. Prediction: $6.3 \times 73 - 259.7 = 459.9 - 259.7 - 200.2$ pounds.
- The slope with its units is 6.3 pounds per inch. It means that a person 8 pounds heavier is predicted to be $8 \times 6.3 = 50.4$ pounds heavier.
- $\sqrt{1 - 0.45^2} \times 42 = \sqrt{1 - 0.2025} \times 42 = \sqrt{0.7975} \times 42 = 0.893 \times 42 \approx 37.5$ pounds.

Example 2:

Find the regression equation for predicting final score from midterm score, based on the following information and predict the final score from a midterm score of 95.

$$\begin{aligned} \text{average midterm score} = 70, &\quad \text{SD} = 12 \\ \text{average final score} = 55, &\quad \text{SD} = 20, \quad r \approx 0.60 \end{aligned}$$

Answer: The slope is $(0.6 \times 20)/12 = 12/12 = 1$, and the equation is $y = x - 15$.

The prediction is $95 - 15 = 80$.

Formulas

Computing the Correlation Coefficient

- Convert the x- and y-values to standard units by finding the average and SD for both the x-values and y-values and then subtracting the average from each value and dividing by the SD. Put the results in two new columns of the table of data.
- For each row multiply (x in standard units) $\times$ (y in standard units). Put the resulting products in the last column of the table.
- Take the average of the products. That average is r.

A. Start with a list of the points of the scatter diagram. Make a table with two columns: the first for x's, the second for y's. Each row of the table will represent a data point.

B. Find the average and SD of x and y and then convert each x-value and each y-value to standard units. Use the formula: (value $-$ average)/SD. Make a new table in standard units.

C. Multiply across each row and average those products. Do not take the square root; that average is r.

Change of Scale

- Adding the same number to every entry on a list adds that constant to the average; the SD and the standard units do not change.
- Multiplying every entry on a list by the same positive number multiplies the average and the SD by that constant, but does not change the standard units.
- Multiplying all the numbers on a list by the same negative constant multiplies the average by that constant, multiplies the SD by the absolute value of that constant, and reverses the signs of the standard units.

Understanding the Correlation Coefficient

The correlation coefficient is a pure number, without units. It is not affected by

- Interchanging the two variables,
- adding the same number to all the values of one variable,
- multplying all the values of one variable by the same positive number.

Correlation masures association. But association is not the same as causation.

Regression

The regression line is to a scatter diagram as the average is to a list. The regression line for y on x estimates the average value for y corresponding to each value of x.

Associated with each increase of one SD in x there is an increase of only r SDs in y, on the average. (This way of using the correlation coefficient to estimate the average value of y for each value of x is called the *regression method.*)

A formula for the estimate of the average y for a given x is

$$\text{estimated average} = \text{average } y + (x \text{ in standard units} \times r \times \text{SD of } y)$$

The regression line is a smoothed version of the graph of averages. If the graph of averages follows a straight line, that line is the regression line.

To find y on the SD line for a given x, we would use the formula

$$y \text{ (on SD line)} = \text{average } y + (x \text{ in standard units} \times \text{SD of } y)$$

A formula predicting the value of y for an individual with a given x is

$$\text{predicted } y = \text{average } y + (x \text{ in standard units} \times r \times \text{SD of } y)$$

This is of course the same formula that was used to estimate the average y for that given x.

In this case a prediction rather than an estimate is in order because the value of y that turns up for the given x depends on which individual is chosen among all those with the given x. The actual average—on the other hand—is a certain value that we are trying to estimate using the regression line.

Percentiles and the Regression Method

If the two variables are perfectly correlated, it makes sense to predict that the percentile of y will be equal to the percentile of x. At the other extreme, if the correlation is zero, then the percentile of x does not help at all in predicting the percentile of y. In the absence of other information, the safest guess is to put the prediction of y at the median or the 50th percentile. In fact when the correlation is somewhere between the two extremes of $r = 1$ and $r = 0$, we have to predict a value of y somewhere between the percentile of x and the median (50th percentile). The regression method for percentiles will provide that value.

Method: Convert the percentile rank of x into standard units, multiply those standard units by r to get standard units of y, then convert the result into a percentile rank for y.

Formula

$$\boldsymbol{z_y = r \cdot z_x}$$

(where z_y and z_x represent y and x in standard units)

Suppose that we are using the technique of regression of y on x to predict the value of y (or estimate the average value of all such y) given a value of x. For such an x, there will be a single point on the regression line. This formula indicates that the y-value of that point—in standard units—is r times the x-value of that point—in standard units.

The Regression Fallacy

In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test—and the top group will on average fall back. This is the *regression effect*.

The regression fallacy consists of thinking that the regression effect must be due to something important, not just the spread around the line. The uncertainty of the estimate of the regression line of y on x will lead to more uncertainty when that estimate is subsequently used to estimate x from y using the regression line of x on y; thus the result will not be the original x.

The R.M.S. Error of the Regression Line

The distance of a point above (+) or below (−) the regression line is

$$\text{error} = \text{actual} - \text{predicted}.$$

Such prediction errors are often called residuals.

The r.m.s. error for regression says how far typical points are above or below the regression line. It is the root-mean-square error of the residuals. This measures the accuracy of the regression predictions. The predictions are off by amounts similar in size to the r.m.s. error. For many scatter diagrams, about 68% of the predictions will be right to within one r.m.s. error. About 95% will be right to within two r.m.s. errors.

Now, looking at the graph and scatter diagram: As a rule of thumb:

About 68% of the points on a scatter diagram fall inside the strip whose edges are parallel to the regression line, and one r.m.s. error away (up or down).

About 95% of the points are in the wider strip whose edges are parallel to the regression line, and twice the r.m.s. error away.

The R.M.S. Error for Regression

The r.m.s. error for the regression of y on x can be figured as

$$\sqrt{1 - r^2} \times \text{the SD of } y.$$

The units for the r.m.s. error are the same as the units for the variable being predicted.

The Regression Line: Slope and Intercept

Associated with each unit increase in x there is some average change in y. The slope of the regression line says how much this change is. The formula for the slope (of the regression line for y on x) is

$$\frac{r \times \text{SD of } y}{\text{SD of } x}$$

To find the intercept of the regression line after the slope m has been found, take any known point (x, y) on the line, substitute it into the equation $y = mx + b$, and then solve for b. The point of averages, which is always on the regression line, is a good point to select for that purpose.

(The intercept of the regression line is just the predicted value for y when x is 0. It may not be a reliable—or even relevant—prediction when 0 is not near the center of the data.)