Math C067 — Why Statistics?

Richard Beigel

Last revision: August 30, 2006

1. Stock Picking

Statistics can help you decide which stock is better.

- Consider Hypothetical returns on \$100 invested in Stock A and Stock B.

Year	2001	2002	2003
Return on Stock X	3	7	5
Return on Stock Y	-3	8	10

- Which Stock do you like better? Why?

2. The Mean

The mean of a list of numbers is just their arithmetic average.

- Consider Hypothetical returns on $100 invested in Stock A and Stock B.

Year	2001	2002	2003
Return on Stock X	3	7	5
Return on Stock Y	-3	8	10

- Let's start by giving names to the data.
- Call the returns on stock X in each year x_1, x_2, x_3.
- x_i is the return on stock X in year $2000 + i$
- $x_1 = 3, x_2 = 7, x_3 = 5$

3. The Mean

The mean of a list of numbers is just their arithmetic average.

- Consider Hypothetical returns on \$100 invested in Stock A and Stock B.

Year	2001	2002	2003
Return on Stock X	3	7	5
Return on Stock Y	-3	8	10

- Let's start by giving names to the data.
- Call the returns on stock X in each year x_1, x_2, x_3.
- x_i is the return on stock X in year $2000 + i$
- $x_1 = 3, x_2 = 7, x_3 = 5$
- Call the returns on stock Y in each year y_1, y_2, y_3.
- y_i is the return on stock Y in year $2000 + i$
- $y_1 = -3, y_2 = 8, y_3 = 10$

4. The Mean

The mean of a list of numbers is just their arithmetic average.

- Consider Hypothetical returns on $100 invested in Stock A and Stock B.

Year	2001	2002	2003
Return on Stock X	3	7	5
Return on Stock Y	-3	8	10

- Let's start by giving names to the data.
- The mean return on stock X is given by the formula

$$\bar{x} = (x_1 + x_2 + x_3)/3$$

  The parentheses matter!

- so $\bar{x} = (3 + 7 + 5)/3 = 15/3 = 5$
- This formula can also be written

$$\bar{x} = \frac{1}{3}\sum_{i=1}^{3} x_i$$

  It's another way of writing the same thing, but you will need to get used to it.

- How do we calculate the mean return on stock Y?
- What if we were looking at more than 3 years worth of data? If there are n data items,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = (x_1 + \cdots + x_n)/n$$

5. Variance and Standard Deviation

Variance and Standard Deviation measure the amount of dispersion in data, i.e., how far away the data are from their mean on average.

- Consider Hypothetical returns on \$100 invested in Stock A and Stock B.

Year	2001	2002	2003
Return on Stock X	3	7	5
Return on Stock Y	-3	8	10

- The variance of the returns on stock X is given by the formula

$$\text{Var}(x) = \left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2\right)/2$$

- The parentheses matter!
- Why didn't we divide by 3?
- This formula can also be written

$$\text{Var}(x) = \frac{1}{2}\sum_{i=1}^{3}(x_i - \bar{x})^2$$

  It's another way of writing the same thing, but you will need to get used to it.

- $\text{Var}(x)$ is also called s_x^2 or simply s^2 when there is only one data set.

6. Variance and Standard Deviation

Let's look only at stock X so we can understand the formula for variance better.

- Consider Hypothetical returns on \$100 invested in Stock A and Stock B.

i	1	2	3
x_i	3	7	5
$x_i - \bar{x}$	-2	2	0
$(x_i - \bar{x})^2$	4	4	0

- Remember that $\bar{x} = 5$, so $x_1 - \bar{x} = 3 - 5 = -2$. What is $x_2 - \bar{x}$? What is $x_3 - \bar{x}$?

$$\begin{aligned} \text{Var}(x) &= \left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2\right)/2 \\ &= (4 + 4 + 0)/2 \\ &= 4 \end{aligned}$$

- What if we had more data instead of just 3 items?
- What if we were looking at more than 3 years worth of data? If there are n data items,

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \left((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right)/(n-1) \end{aligned}$$

- Standard deviation is the square root of variance:

$$s = \sqrt{s^2}$$

- $s_x = \sqrt{\text{Var}(x)} = \sqrt{4} = 2$

7. Correlation

- Loosely speaking, two stocks are correlated if they move the same way every day (both go up or both go down)
- Loosely speaking, two stocks are anti-correlated if they move the opposite way
- The correlation coefficient (r) measures how correlated two data sequences are.
- If $r = 1$ they are perfectly correlated
- If $r = -1$ they are perfectly anti-correlated
- If $r = 0$ they are uncorrelated
- Fractional values indicate partial correlation (or partial anti-correlation)
- Application [Harry Markowitz]: A diverse portfolio of anti-correlated or uncorrelated stocks reduces risk.

8. Covariance

- Covariance of X and Y:

$$
\begin{aligned}
s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\
&= ((x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y}))\,/(n-1)
\end{aligned}
$$

- Note: Covariance of X and X:

$$
\begin{aligned}
s_{xx} &= ((x_1-\bar{x})(x_1-\bar{x})+\cdots+(x_n-\bar{x})(x_n-\bar{x}))\,/(n-1) \\
&= \left((x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2\right)/(n-1) \\
&= s_x^2
\end{aligned}
$$

9. Covariance and Correlation

- Covariance of X and Y:

$$
\begin{aligned}
s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \left((x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})\right)/(n-1)
\end{aligned}
$$

- Correlation (coefficient) of X and Y:

$$r = \frac{s_{xy}}{s_x s_y}$$

10. Linear Regression (Least Squares Best-Fit Line)

- Two data sequences can be plotted as ordered pairs (x,y). The *best-fit* line comes as close as possible (in a sense) to the points in the plot.
- Equation for the best-fit line: $y = a + bx$ where
- $b = rs_y/s_x$ and
- $a = \bar{y} - b\bar{x}$