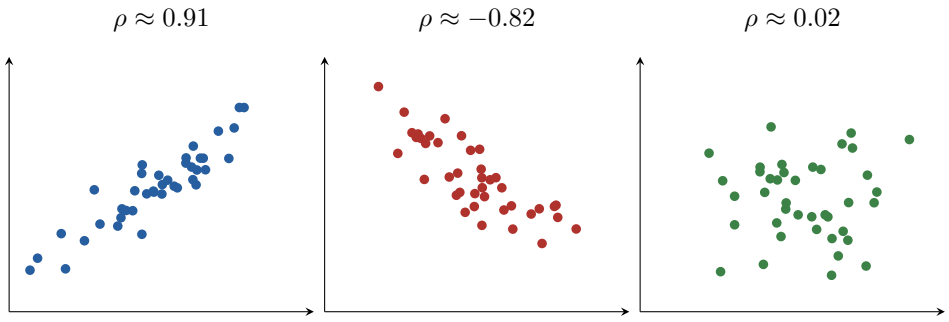
Data Analytics Workshop, JHU Pre-College Program
Day 5: Questions, Hypotheses, and Correlation (optional notes)

TL;DR Today is about the project: how to ask good questions of a dataset and how to tell whether the answer is real or just luck.

- A meaningful question is **observable** (our data can see it), **measurable** (with numbers or categories), and **repeatable** (someone could redo the study). Data is **numerical** (salary, age) or **categorical** (college attended, favorite flavor).
- The **null hypothesis** H_0 says any effect we see is due to chance. The **alternative hypothesis** H_1 says there is a real relationship. We collect data and ask whether we have enough evidence to **reject the null**.
- The ***p*-value** is the probability, *assuming the null hypothesis is true*, of seeing data at least as extreme as ours. If it is below a threshold (often 0.05), the result is **statistically significant**.
- Data analysis does not show *causation*. It measures **correlation**: a number ρ between -1 and 1 describing how one variable tends to change as another changes.

(Covariance and correlation)

$$\operatorname{cov}(x,y) = \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y}), \qquad \rho_{x,y} = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}.$$



Strong positive, strong negative, and no correlation

1 Meaningful Questions

By now you may have a large dataset for your project. Suppose ours is a sample of 1000 employees at a company, with each person's salary, years at the company, age, race, and gender. Before testing anything, we need questions worth asking. A meaningful question should be

1. **Observable**: we can actually observe it with the data we have. We cannot ask about favorite ice cream flavors if the dataset does not record them.
2. **Measurable**: we can measure it, though not necessarily with numbers.
3. **Repeatable**: the study could be done again, for example at a different company.

Definition 1.1 (Numerical and categorical data)

Numerical *data can be assigned a number, such as salary, years worked, or age.* ***Categorical*** *data sorts each data point into a group, such as race, favorite ice cream flavor, or college attended.*

Categorical data is still measurable. To compare salary across races, we can group people by race, take the average salary of each group, and plot the averages as a bar chart. The kind of data also tells you the kind of visual to make: categories go on a bar chart, while two numerical variables go on a scatter plot (more on this on Day 7).

Meaning is subjective (ask ten philosophers and you'll get ten thousand answers), but a question that is observable, measurable, and repeatable is on the right track. Examples are whether age relates to salary, or whether years at the company relate to salary. Notice the word *relate*. Data analysis rarely tells us that one thing *causes* another, so we try not to phrase questions that way.

2 Hypothesis Testing and p-Values

Even among people who have worked at a company the exact same number of years, salaries vary. They come from some distribution with some mean and some variance. Because of this randomness, we might see that people who have worked longer earn more in our sample, even if no real relationship exists in the company as a whole. Hypothesis testing is how we guard against being fooled by luck.

Definition 2.1 (Null and alternative hypotheses)

The ***null hypothesis*** H_0 *states that any effect we see in the data is due to chance. The* ***alternative hypothesis*** H_1 *states that there is a relationship that is not due to chance alone.*

A null hypothesis can also compare two groups. For two coins that may not be fair, H_0 could state that both coins land heads with the same probability, and H_1 that coin 1 lands heads more often.

Example 2.1

We flip each coin 100 *times.*

	Heads	*Tails*
Coin 1	*89*	*11*
Coin 2	*78*	*22*

Coin 1 came up heads more often, but this could be luck. Coin 2 could secretly be the more unfair one and we just got unlucky with it. The question is whether the data gives us enough evidence to ***reject the null hypothesis****. Notably, with* 1000 *flips each, a gap this large would be much harder to blame on luck.*

So the process is:

1. Ask a meaningful question.
2. State a null hypothesis (the effect is due to chance) and an alternative hypothesis.
3. Collect data.
4. Use a measure of statistical significance, such as error bars or a p-value, to decide whether we can reject the null hypothesis.

Definition 2.2 (p-value)

The ***p*-value** *is the probability, assuming the null hypothesis is true, of observing data at least as extreme as the data we observed.*

A p-value is always between 0 and 1. Since it is computed from the observed data, it depends on the sample size. Here is the key idea: we *pretend* the null hypothesis is true, and ask how surprising our data would be. A small p-value means our data would be very unlikely if only chance were at play. If the p-value is below a chosen threshold, often 0.05, we reject the null hypothesis and call the result **statistically significant**. For example, a p-value of 0.003 for the two coins would give strong evidence that coin 1 really does land heads more often than coin 2.

Remark 1

A very common mistake (even in published papers!) is to say the p-value is "the probability that the result is due to chance," or "the probability the null hypothesis is true." It is neither. It is computed assuming the null hypothesis holds, so it only measures how surprising the data would be in that world. After Day 6 you will have the language for this: the p-value is about $\mathbb{P}(\text{data} \mid H_0)$*, not* $\mathbb{P}(H_0 \mid \text{data})$*, and those can be very different. In Discussion 4 you read an article on why p-values, though one of the most used tests, are not always the best one.*

3 Correlation versus Causation

You may have heard that "correlation does not imply causation." In data analysis we try to avoid the word causation altogether, because it is just too strong. Causation says one thing definitely causes another. Even dropping your phone does not always crack the screen, and we have all had the opposite happen too.

Correlation, on the other hand, is exactly the kind of thing we like: it is observable, repeatable, and most importantly measurable. It is a number describing how one variable tends to change as another changes. If we plot age against salary, we might see salary increase with age, decrease with age, or show no pattern at all, like the three plots on the first page. In the first two we could easily draw a line through the data. In the third, we could not even say which direction the line should go.

Example 3.1

Take $x = (1, 2, 3, 4, 5)$ and $y = (1, 2, 3, 4, 5)$, where the lists are paired by position. As x increases, y increases by exactly the same amount. This is perfect correlation, with correlation 1. If instead $z = (5, 4, 3, 2, 1)$, then z decreases by the same amount each time x increases, and the correlation of x and z is -1.

Definition 3.1 (Covariance and correlation)

*For paired data $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$, the **covariance** is*

$$\operatorname{cov}(x, y) = \frac{1}{N} \sum_{k=1}^{N} (x_k - \bar{x})\,(y_k - \bar{y})\,,$$

*and the **correlation** is*

$$\rho_{x,y} = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}.$$

The correlation always satisfies $-1 \le \rho_{x,y} \le 1$.

A correlation near 1 means that increasing one variable tends to increase the other. A correlation near -1 means that increasing one variable tends to decrease the other. A correlation near 0 means there is no clear linear relationship.

Why does the sign come out right? Suppose x and y are positively correlated. When x_k is above $\bar{x}$, y_k tends to be above $\bar{y}$, so both factors in the sum are positive. When x_k is below $\bar{x}$, y_k tends to be below $\bar{y}$, so both factors are negative, and their product is positive again! Either way we mostly add positive numbers. When the variables are negatively correlated, the factors tend to have opposite signs and the sum is negative. Dividing by $\sigma_x \sigma_y$ is what keeps the answer between -1 and 1 (the reason why takes more advanced math).

Example 3.2

Let $x = (1, 2, 3, 4, 5)$ and $y = (2, 1, 4, 3, 5)$. Both have mean 3.

$x_k - \bar{x}$	-2	-1	0	1	2
$y_k - \bar{y}$	-1	-2	1	0	2
product	2	2	0	0	4

The products add to 8, so $\operatorname{cov}(x, y) = \frac{8}{5} = 1.6$. Both variances are $\frac{4+1+0+1+4}{5} = 2$, so $\sigma_x = \sigma_y = \sqrt{2}$ and

$$\rho_{x,y} = \frac{1.6}{\sqrt{2}\sqrt{2}} = \frac{1.6}{2} = 0.8.$$

That is, y mostly goes up with x, but not perfectly. I invite you to make up some data of your own, compute its correlation by hand, and then check it with `CORREL` *in Excel.*

Remark 1

The covariance of x with itself is the variance:

$$\operatorname{cov}(x,x) = \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x})^2 = \sigma_x^2, \qquad \text{so} \qquad \rho_{x,x} = \frac{\sigma_x^2}{\sigma_x \sigma_x} = 1.$$

Any data is perfectly correlated with itself, which is exactly what we should expect.

Two warnings to keep in mind for your project.

Remark 2

Correlation only sees lines. *Take $x = (-2,-1,0,1,2)$ and $y = x^2 = (4,1,0,1,4)$. Here y is completely determined by x, but you can check that $\rho_{x,y} = 0$! As x increases, y first goes down and then goes up, and these cancel. Always plot your data, don't just trust one number.*

Remark 3

Hidden variables. *Across a year, ice cream sales and sunburns are strongly positively correlated. Ice cream does not cause sunburns (I hope). Both go up in the summer, when it is hot and sunny. A third variable driving both is called a confounding variable, and it is one of the main reasons correlation does not imply causation.*

4 Practice

Problem 4.1 *1) Is each question observable, measurable, and repeatable with a dataset of students' GPA, hours of sleep, and favorite subject? (a) Do students who sleep more have higher GPAs? (b) Does sleep cause better grades? (c) Do students who like math more have higher GPAs?*

2) Compute $\rho_{x,z}$ for $x = (1,2,3,4,5)$ and $z = (5,4,3,2,1)$ using the formula.

Solution:

1) (a) Yes: both are numerical and recorded. (b) Not as stated: data can show a relationship, but not causation. Rephrase it as (a). (c) Not quite: we know each student's favorite subject, not how much they like math. Comparing the average GPA of students whose favorite subject is math against the others would work.
2) Both means are 3. The deviations are $(-2,-1,0,1,2)$ and $(2,1,0,-1,-2)$, with products $-4,-1,0,-1,-4$ adding to -10. So cov $= -2$, both variances are 2, and $\rho_{x,z} = \frac{-2}{2} = -1$.