32 2025-04-21 | Week 14 | Lecture 32

Chapter 14.1 through 14.3

32.1 Tests of Association

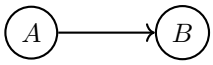
Suppose we want to know if two things A and B are related to each other. We will discuss only the case where A and B are discrete, categorical events. For example

$$A = \text{[a person vapes]}$$

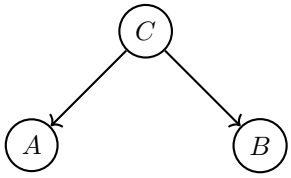
$$B = \text{[person gets coronary artery disease]}$$

There are many possible relationships:

- **Causality:** Vaping causes arterial disease.

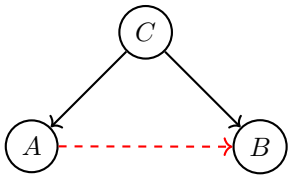


- **Common Response:** Perhaps the effect of vaping on your ateries is perfectly benign, but there is some other variable C that causes both to occur simultaneously:



  For example we might have $C = \text{[person smokes cigarettes]}$. Maybe vapers are disproportionately likely to be cigarette smokers, and it's this subset of vapers that drives the association between vaping and arterial disease. If this is true, then quitting vaping woudn't decrease your chance of arterial disease.

- **Confounding:** A third variable jointly influences or causes both A and B. For example, maybe vapers have lower rates of coronary artery disease than the general population – that doesn't mean vaping necessarily have a protective effect. For example, maybe the vapers are on average younger than the population, and CAD is something that typically hapens to old people. In that case, we might observe a spurious relationship between A and B due to the effects of some other third variable, $C = \text{[youth]}$:



What we can say, however, is that if A and B do not have any relationship, then they are probabilistically independent. In what follows, we will describe how the hypothesis testing framework can be used to test whether two things A and B are independent.

This is what tests of association do. They ask the question **"Are A and B independent?"** The answer to such tests is always either "no" or "maybe". They don't give us more information than that.

We need some theory to describe how to do this.

32.2 Ingredient #1: The Multinomial Distribution

Recall that to construct a *binomial* random variable, you flip coins n times, and count the number of heads. The coin flip only has two outimes. What if, instead of flipping coin, we rolled a dice with k sides, and counted the number of times we get $1, 2, \ldots, k$? In that case, we get an example of what is called *multinomial* random variable. To be precise:

Definition 92 (Multinomial Distribution)**.** Consider an experiment with k possible outcomes, labeled 1 through k.. These outcomes have probabilities $p_1, \ldots, p_k$, respectively. (So $p_1 + \cdots + p_k = 1$). We repeat this experiment n times, and define

$$X_i = \#\text{ of times outcome } i \text{ occurs in the } n \text{ trials}$$

for each $i = 1, \ldots, k$. Then the random vector

$$X = (X_1, \ldots, X_k)$$

is called a *multinomial random variable* with parameters $(p_1, \ldots, p_k)$ and number of trials n.

This is an example of a categorical random variable (k categories).
Let's do an example with $k = 3$.

Example 93 (One armed bandit)**.** A single play of a slot machine has three outcomes:

- outcome 1: win nothing
- outcome 2: win back your bet
- outcome 3: win big

There is a sticker on the machine which says that these outcomes have probabiliites $(p_1, p_2, p_3) = (.5, .3, .2)$. A gambler decides to play the game $n = 100$ times. For each $i = 1, 2, 3$ let N_i be a random variable indicating the number of times that outcome i occurred.

If the sticker is accurate, then the random vector $N = (N_1, N_2, N_3)$ is a multinomial random variable with parameters $(p_1, p_2, p_3) = (.5, .3, .2)$ and $n = 100$. Clearly

$$\mathbb{E}[N_i] = np_i$$

So

$$\mathbb{E}[N_1] = 50 \quad \text{and} \quad \mathbb{E}[N_2] = 30 \quad \text{and} \quad \mathbb{E}[N_3] = 20$$

After playing the game, suppose gambler gets $N = (43, 35, 22)$. It is convenient to arrange this in a table

Category	1	2	3
Observed	43	35	22
Expected	50	30	20

We expect the observed values to be pretty close to the expected values. But if they **differ substantially**, then we conclude that maybe the sticker isn't accurate. In the hypothesis testing framework, sticker represents the null hypothesis

$$H_0 : (p_1, p_2, p_3) = (.5, .3, .2).$$

Question: How do we quantify "differ substantially"?

Answer: One sensible way is to consider the following quantity:

$$\begin{aligned}\chi^2 &:= \frac{(43-50)^2}{50} + \frac{(35-30)^2}{30} + \frac{(22-20)^2}{20} \\ &= \frac{49}{50} + \frac{25}{30} + \frac{4}{20} \\ &\approx 2.01\end{aligned}$$

each term in this sum takes the form

$$\frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$

The sum in **??** is always nonnegative. If the observed values are close to the expected values, then it is small. Otherwise it is large. So we will look at the size of χ^2, and make the following determination

- if χ^2 is large: we reject the null hypothesis H_0
- if χ^2 is small: we "fail to reject" the null hypothesis

End of Example 93. □