Bonus Stats Knowledge

David Wen

08/09/2021

Statistics

Since we're doing data analysis, understanding statistics is **super** important. If you haven't taken a stats class before (take QBIO 310!) or if you want a review of basic terminology, I'd highly recommend reading over this tutorial.

Types of Variables

In statistics, variables fall under two main classes: *categorical* and *continuous*. Categorical variables take on discrete values; they describe some qualitative aspect of a population. On the other hand, continuous variables take on measurable numeric values (they are quantitative, like our major). A heuristic to tell the difference is to ask if you can take the mean value of the variable: if you can, it's probably a continuous variable. It's important that you can recognize the difference to understand what analyses you can perform and what plots you can use to visualize the data (more on that in the next section).

Here are some examples of categorical variables you might see in a cancer data set:

- *Race/ethnicity*: white, black or african american, asian, american indian or alaska native.
- *Treatment type*: pharmaceutical, radiation, none.
- *Cancer type*: breast ductal carcinoma, lung adenocarcinoma, sarcoma, etc.
- *Stage*: I, II, III, IV
- *Age group*: young, middle, old.

As you can see, for each categorical variable, a cancer patient will take on at least one of the possible variables, which are non-numerical descriptors (note that even though stage is *represented* by a number, it's not something that's actually counted – you can't really have stage 2.43 cancer).

Here are some continuous variables you're likely to see in a cancer context:

- *Counts of the gene ESR1*
- *Days to death*
- *Number of mutated genes*
- *Age*

For each of these, you'll have a range of values, where patients will fall on a spectrum, and you can measure and do math on these quantities.

Some basic terminology for continuous variables

When you're describing a continuous variable, you'll need to describe its characteristics and distribution. Here are the ones that you'll see most often:

- Mean (μ): this is the numerical average, i.e. the sum of all observations divided by the number of observations. This is the one you'll see most often in daily life, but often it's not adequate when describing gene expression because of skewness (see that section for more info).

- Median: this is the "middle" value, i.e. if you sorted the values and took the middle one. This is usually the most useful value when the data is skewed or if there are outliers.
- Variance (σ^2): this is a measure of how spread the values are. Mathematically, it is the mean squared-distance from the sample mean, i.e. $\sigma^2 = \frac{1}{n}\sum(x-\mu)^2$.
- Standard deviation (σ): this is just the square root of the variance. This is useful because the units are the same the measurements (notice that for variance the units will be squared).

Hypothesis Testing

You've probably heard of p-values before, but there's a lot of confusion about what they actually mean. To understand what a p-value is, we have to introduce two terms: the null hypothesis (H_0) and the alternative hypothesis (H_1 or H_a).

- The **null hypothesis** is the statement to be tested. More often, we are interested in seeing a difference between two populations, so generally the null hypothesis states that two populations are the same in some respect.
- The **alternative hypothesis** is the statement that is tested against the null hypothesis. Think of it as the "interesting" result, i.e. there *is* some difference between the two populations.

The **p-value** relates these two concepts: it is the probability that assuming the null hypothesis is true, that you see a results at least as extreme as the alternative hypothesis. In other words, let's say you perform an experiment. The p-value is the probability that you would see a result at least as extreme as the one you just saw, assuming the null hypothesis is true. In general, you can reject the null hypothesis if $p < 0.05$ (though there's plenty of discussion about the proper significance level).

This is all very abstract, so let's take an example. Your untrustworthy friend has a coin, and says he'll pay you one dollar for every tails he flips, but he'll take one dollar from you for every heads he flips. Naturally, he flips 16 heads and you're down 12 dollars. You suspect the coin is unfair and you got punked – can you do a statistical test?

First, you define your null hypothesis: the coin is fair, i.e. $Pr(H) = 0.5$. Your alternative hypothesis is that the coin is not fair, i.e. $Pr(H) \neq 0.5$. In other words:

$$H_0 : Pr(H) = 0.5$$
$$H_1 : Pr(H) \neq 0.5$$

Then you need to calculate the chance of seeing a result at least as extreme as the one you saw: flipping 16 or more heads, assuming the coin is fair. The probability of this result is $p = 0.0059$,[1] so you can reject your null hypothesis at the 0.05 confidence level and accuse your friend of swindling you.

Here's another example: you have a shuffled deck of 52 cards, and your same friend flips over the top half of the deck (26 cards). You see no threes of any kind, and you start to get suspicious that the deck has been tampered in some way. How do you go about testing this?

First, you define your null hypothesis: this deck is normal, i.e. there are 4 threes in the deck. Then you define your null hypothesis: there are not 4 threes in the deck. In mathematical notation, let n be the number of threes in the deck.

$$H_0 : n = 4$$
$$H_1 : n \neq 4$$

[1] For you math nerds (me), you use the binomial distribution, and $P(H_0) = \sum_{i=16}^{20}\binom{20}{i}0.5^{20}$. You can calculate in R by `sum(dbinom(16:20, 20, 0.5))`.

You need to calculate the probability of drawing 0 threes in 26 cards (since you can't get more extreme than 0), assuming there are 4 threes in the deck. It turns out the probability is $\frac{46}{833} \approx 0.055$,[2]. In other words, if you did the experiment 1000 times, you would expect to see this result to happen about 55 times. Since $p > 0.05$, you can't reject the null hypothesis and say the deck has been tampered with. Maybe your friend was just a lucksack after all.

Important Note: when you're actually performing data analysis, you're going to be using statistical tests that spit out a p-value. The math behind these tests isn't nearly as simple as the ones above, but it's a very similar principle: you choose a particular statistical test tailored to answer a specific question, and it will tell you the likelihood of the observed result under the certain assumptions. *Be sure to select the right test*, because as you can see, the p-value depends very strongly on the initial assumptions (null hypothesis) and the alternative hypothesis.

A Sidenote on P-values

When working with genomic data, you'll see adjusted p-values. The reason they exist is that p-values say that there is p chance that you see a difference assuming two populations are the same. However, when you're comparing gene expressions, you're working with over 20,000 genes. If you accept the $p = 0.05$ cutoff, you'll see a lot of genes being called significant just because of statistical chance. For example, if you're looking at 20,000 genes between two equivalent individuals (i.e. there is no "real" difference between them), you'd expect 1000 of these differences (5%) to be statistically significant just by the definition of the p-value. There's statistical modifications to remove this noise, but I'm not going to get into them (look up False Discovery Rate, Bonferroni Correction, Multiple Comparisons, Q-Value, etc.)

[2] This is the hypergeometric distribution, and the probability is $\binom{4}{0}\binom{48}{4}/\binom{52}{4}$. R will calculate this for you with `dhyper(0, 4, 48, 26)`.