Hypothesis Testing

STA 198: Introduction to Health Data Science

Yue Jiang

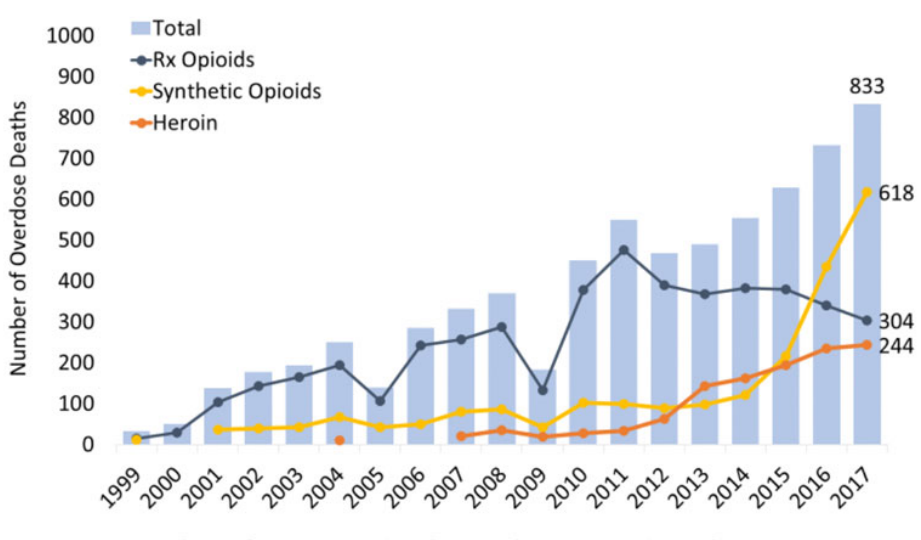
June 15, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

The opioid crisis

National Institute on Drug Abuse
Advancing Addiction Science

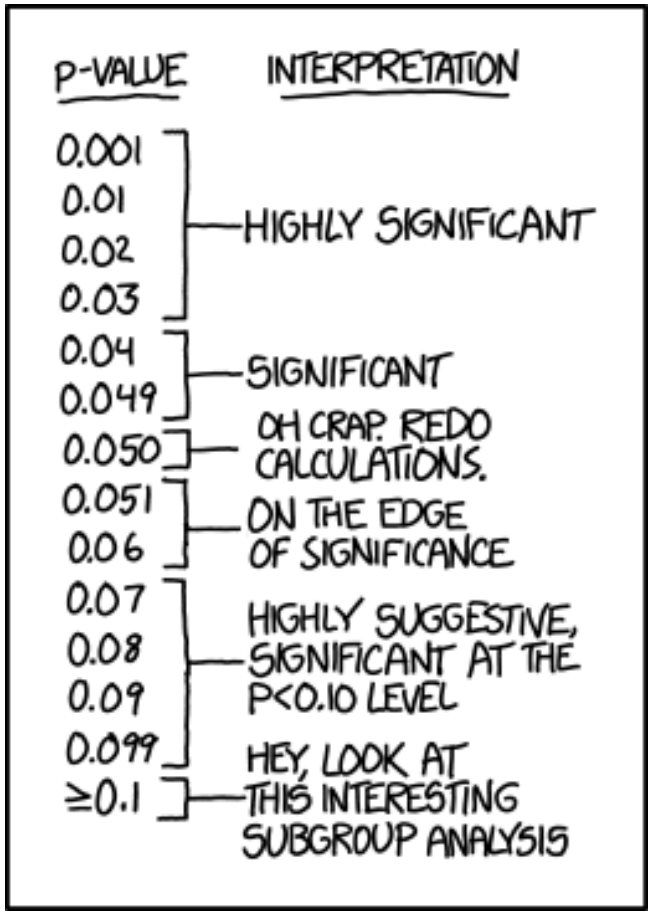


West Virginia has the highest age-adjusted rate of drug overdose deaths involving opioids

Statistical inference

- Point estimation: estimating an unknown parameter using a single number calculated from the sample
  - We estimate the one-year death rate due to opioid-related overdoses in WV to be 49.6 per 100,000
- Interval estimation: estimating an unknown parameter using a range of values that is likely to contain the true parameter
  - We estimate that the one-year death rate due to opioid-related overdoses in WV is between 45 and 55 per 100,000
- Hypothesis testing: evaluating whether our observed sample data provides evidence against some populuation claim
  - We evaluate the hypothesis that the opioid overdose death rate is the same in WV and NC. In a random sample of death certificates from the two states, the rate was considerably higher in WV, providing evidence against this hypothesis

Why should we care about hypothesis testing?



Randall Munroe, xkcd

Emperor Antonius Pius

Ei incumbit probatio qui dicit, non qui negat

How can we answer research questions using statistics?

Statistical hypothesis testing is the procedure that assesses evidence provided by the data in favor of or against some claim about the population (often about a population parameter or potential associations).

The hypothesis testing framework

1. Start with two hypotheses about the population: the null hypothesis and the alternative hypothesis
2. Choose a sample, collect data, and analyze the data
3. Figure out how likely it is to see data like what we got/observed, IF the null hypothesis were true
4. If our data would have been extremely unlikely if the null claim were true, then we reject it and deem the alternative claim worthy of further study. Otherwise, we cannot reject the null claim

Ultra-low dose contraception

Oral contraceptive pills work well, but must have a precise dose of estrogen.

If a pill has too high a dose, then women may risk side effects such as headaches, nausea, and rare but potentially fatal blood clots.

If a pill has too low a dose, then women may get pregnant.

Ultra-low dose contraception

A certain contraceptive pill is supposed to contain precisely 0.020 μg of estrogen. During QC, 50 randomly selected pills are tested, with sample mean dose 0.017 μg and sample SD 0.008 μg.

Do you think this is cause for concern? Why or why not?

(don't worry about calculations yet)

Two competing hypotheses

The null hypothesis (H_0) states that "nothing unusual is happening" / there is no change from the status quo / there is no relationship / etc.

The alternative hypothesis (H_A or H_1) states the opposite: that there is some sort of relationship (usually this is what we want to check or really think is happening)

Remember, in statistical hypothesis testing we *always first assume the null hypothesis is true*, and see whether we reject or fail to reject this claim

Defining the null and alternative hypotheses

Stated in words:

- H_0: The pills are consistent with a population that has a mean of 0.020 μg estrogen
- H_1: The pills are not consistent with a population that has a mean of 0.020 μg estrogen

Stated in symbols:

- $H_0 : \mu = 0.020$
- $H_1 : \mu \neq 0.020$,

where μ is the mean estrogen level of the manufactured pills, in μg

Collecting and summarizing the data

With these two hypotheses, we now take a sample and summarize the data

The choice of summary statistic calculated depends on the type of data as well as its distribution

In our example, quality control technicians randomly selected a sample of 50 pills and calculated the sample mean $\bar{x} = 0.017$ μg and sample standard deviation $s = 0.008$ μg

Assessing the evidence observed

Next, we calculate the probability of getting data like ours, or more extreme, if H_0 were actually true

This is a conditional probability: "if H_0 were true (i.e., if μ were truly 0.020), what would be the probability of observing $\bar{x} = 0.017$ and $s = 0.008$?"

This probability is the p-value

Some philosophical details

The obtained p-value relates to the test specific itself, so use of the same data can result in different p-values or confidence intervals depending on which test is used

Importantly, we have assumed from the start that the null hypothesis is true, and the p-value calculates conditioned on that event

p-values do NOT provide information on the probability that the null hypothesis is true given our observed data

Making a conclusion

We reject the null hypothesis if the conditional probability of obtaining our test statistic, or more extreme, given it is true, is very small

What is "very small"? We often consider a cutpoint (the significance level or α level) defined prior to conducting the analysis

Many analyses use $\alpha = 0.05$: if H_0 were in fact true, we would expect to make the wrong decision only 5% of the time (why?)

If the p-value is less than α, we say the results are statistically significant and we reject the null hypothesis. On the other hand, if the $p - value$ is α or greater, we say the results are not statistically significant and fail to reject H_0.

What do we actually conclude when $p > \alpha$?

We never "accept" the null hypothesis – we assumed that H_0 was true to begin with and assessed the probability of obtaining our test statistic (or more extreme) under this assumption

When we fail to reject the null hypothesis, we are stating that there is *insufficient evidence* to assert that it is false

Don't forget to use common sense

Common sense should also be used in drawing conclusions. For example, if there is a very strong body of evidence in favor of the alternative hypothesis and you see $p = 0.057$, you wouldn't necessarily use your results to refute all the prior work in an area. This is one reason confidence intervals are often used in place of hypothesis tests.

Back to the oral contraceptives

As it turns out, the probability of observing a sample mean of 0.017 and sample SD of 0.008 in 50 pills if H_0 were actually true is approximately 0.01.

What might we conclude?

What could go wrong?

Suppose we test the null hypothesis $H_0 : \mu = \mu_0$. We could potentially make two types of errors:

	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Fail to Reject H_0	✓	Type II Error
Reject H_0	Type I Error	✓

Type I Error: rejecting H_0 when it is actually true (falsely rejecting the null hypothesis)

Type II Error: not rejecting H_0 when it is false (falsely failing to reject the null hypothesis)

Type I vs. Type II errors

HIV Test

Pregnancy Test

Review: steps in hypothesis testing about the mean

1. Hypothesis a value (μ_0) and set up H_0 and H_1
2. Take a random sample of size n and calculate summary statistics (e.g., sample mean and sample variance)
3. Determine whether it is likely or unlikely that the sample, or one even more extreme, came from a population with mean μ_0 with $\alpha = 0.05$ or some other pre-specified value
4. Draw conclusions

Different sets of hypothesses

We set up the hypotheses to cover *all* possibilities for μ and consider three possibilities:

- Two-sided: $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$
- One-sided: $H_0 : \mu \geq \mu_0$; $H_1 : \mu < \mu_0$
- One-sided: $H_0 : \mu \leq \mu_0$; $H_1 : \mu > \mu_0$

One-sided tests are pretty rare (why?).

Two-sided tests of hypotheses

To conduct the hypothesis test, we use what we learned about the sampling distribution of the sample mean $\bar{X}$. If the underlying population is normally distributed (or n is pretty large), then the random variable

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

has a t_{n-1} distribution

Breaking down the test statistic

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

- $\bar{X} - \mu_0$ tells us how far our sample mean is from the hypothesized population mean
- Whether $\bar{X} - \mu_0$ is big depends on the variance (and standard deviation): a difference of $\bar{X} - \mu_0 = 1$ is a small difference if we are looking at weight in grams, but large for height in meters. This is why we standardize the difference by the estimated SD of the mean ($s/\sqrt{n}$)

Thus, the test statistic t is an estimate of how many SDs apart μ_0 and $\bar{X}$ are from each other

Getting the p-value graphically

(We'll walk through this during the live session)

Why not use a one-sided test?

The New England
Journal of Medicine

Volume 324 MARCH 21, 1991 Number 12

MORTALITY AND MORBIDITY IN PATIENTS RECEIVING ENCAINIDE, FLECAINIDE, OR PLACEBO

The Cardiac Arrhythmia Suppression Trial

DEBRA S. ECHT, M.D., PHILIP R. LIEBSON, M.D., L. BRENT MITCHELL, M.D., ROBERT W. PETERS, M.D., DULCE OBIAS-MANNO, R.N., ALLAN H. BARKER, M.D., DANIEL ARENSBERG, M.D., ANDREA BAKER, R.N., LAWRENCE FRIEDMAN, M.D., H. LEON GREENE, M.D., MELISSA L. HUTHER, DAVID W. RICHARDSON, M.D., AND THE CAST INVESTIGATORS*

Ok, so what is a p-value?

The p-value is the probability, under the assumption that the null hypothesis is true, of obtaining a test statistic equal to or more extreme than what was actually observed.

What isn't a p-value?

"A p-value of 0.05 means the null hypothesis has a probability of only 5% of being true." – This is wrong!

"$p = 0.05$ means that that there is a 95% chance or greater that the null hypothesis is incorrect." – This is wrong!

A p-value is calculated *assuming* that H_0 is true. It cannot be used to tell us how likely it is that that assumption is correct

Even more bad news

While we of course want to know if any one study is showing us something real or a Type I or Type II error, hypothesis testing does NOT give us the tools to determine this

What DO you do with a p-value, then?

Ideally, you evaluate a p-value in light of other information, such a proposed biological mechanism, supporting evidence in the literature, size and quality of the study, and size of the purported effect.

In addition, when you interpret a p-value, be sure you are aware of other important factors that can inflate the false positive rate, e.g. multiple tests, "hidden" multiple tests (e.g., testing most appropriate form of covariate or for an interaction or other model selection procedures), changes to sample size, etc.

I'm nervous about p-values. What alternatives are there?

- Many researchers prefer confidence intervals, which represent the range of effects that are "comparible with the data"
- They are sometimes used as a hypothesis test (i.e., reporting results as significant when confidence interval does not include null value) and share many of the properties of p-values we have discussed
- Like p-values, confidence intervals do not offer a mechanism to unite external evidence with that provided by the study at hand
- *Bayesian methods* can be used to incorporate current study results with prior knowledge in order to provide a probability a hypothesis is true conditional on the current data