October 31, 2017

CS 361: Probability & Statistics

Hypothesis testing

Does the population have this mean?

Suppose we hypothesize the average human body temperature is 95 degrees. Let $\overline{T}$ be the random variable evaluated by collecting a random sample of people, measuring their temperatures, and computing the average of these temperatures. This is the sample mean, then. Which means the random variable is normal. Its expected value is the population mean and its standard error is *s*

If our hypothesis is true, then the following is a standard normal random variable

$$G = \frac{(\overline{T} - 95^o)}{s}$$

We can now tell whether the evidence contradicts our hypothesis because we can calculate how unusual the sample we actually observed would be if the hypothesis were true

Does the population have this mean?

Denote the actual value we observe for the random variable $\overline{T}$ as $\overline{t}$ and use it to calculate g

$$g = \frac{(\overline{t} - 95^o)}{s}$$

Now if our hypothesis is true G is a standard normal random variable which means that for 68% of the temperature samples we could have drawn, this **test statistic** above, g, would have been between -1 and 1

If we calculate

$$f = \frac{1}{\sqrt{2\pi}} \int_{-|g|}^{|g|} \exp\left(\frac{-u^2}{2}\right) du$$

f is the fraction of samples, assuming our hypothesis is true, which would have had values less extreme than the one we observed

Does the population have this mean?

So a value of g with a large absolute value gives an f very close to 1. Which means that if our hypothesis was true, we have observed a very unusual sample —most samples would have given a value of g closer to 0. If assuming our hypothesis is true makes our data highly unusual, we can say that the data fails to support the hypothesis

Traditionally we look at p=1-f which is called the **p-value**

We reject the hypothesis we are evaluating if the p-value is sufficiently small. The cutoff used in many scientific disciplines is p=0.05 or smaller. A p-value of 0.05 means that you would see evidence this unusual in about one experiment out of twenty if the hypothesis were true

Example: outliers

Assess the hypothesis that the average BMI of humans is 27

We have a dataset of 252 heights/weights. Two are outliers which we discard. We get a mean BMI of 25.3 and a standard deviation of 3.35

Giving a standard error of 0.21 and a test statistic of -8.1 which implies a p-value of almost 0. Which means this is an extremely unusual dataset if the hypothesis is true, so we should reject the hypothesis

If we include the outliers, the mean BMI is 25.9 and the standard deviation is 9.56. This causes us to have a p-value of 0.08 which might make us accept the hypothesis

Summary

If we want to hypothesize about the mean of a population based on a sample, rejecting that hypothesis will occur if the dataset we observe would have needed to be quite unlikely, given the hypothesis

The p-value tells us what proportion of possible samples would have had a less unusual value than the one we observed if the hypothesis were true

Outliers can blow up the standard deviation dramatically and can therefore lower our threshold for non-rejection of a hypothesis

Note the connection to confidence intervals: there we knew the distribution of the sample mean (approximately) and reported an interval where the sample mean would lie in 95% of possible samples. Hypothesis testing asks how large of a confidence interval we would have to draw to enclose the test statistic

Do two populations have the same mean?

Suppose we have two different samples and we want to know if they came from the same or different populations

That is we have a dataset {x} with size k_x and a dataset {y} with size k_y, not necessarily the same size and each is a dataset drawn with replacement from a population

The sample means are likely to be different no matter what since they are random samples, but can we tell whether they are different because the underlying populations are different?

Using some tricks about normal random variables we can answer this kind of question

Sums and differences of normal RVs

Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$. Suppose X_1 and X_2 are independent.

For any constant $c_1 \neq 0$, we have $c_1 X_1 \sim \mathcal{N}(c_1\mu_1, |c_1\sigma_1|)$

For any constant c_2, we have $X_1 + c_2 \sim \mathcal{N}(\mu_1 + c_2, \sigma_1)$

And we have $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$

Two samples

Let $X^{(k_x)}$ be the random variable corresponding to calculating the sample mean of our first sample of size k_x. And let $Y^{(k_y)}$ correspond to the sample mean of the other sample of size k_y

We know from our work on sample means that these random variables are normally distributed. Now if we hypothesize that if we hypothesize that these two samples come from populations with the same mean we must have that $D = X^{(k_x)} - Y^{(k_y)}$ is normally distributed with

$$\mathbb{E}[D] = 0.$$ and $$\text{std}\,(D) = \sqrt{\text{std}\left(X^{(k_x)}\right)^2 + \text{std}\left(X^{(k_y)}\right)^2}.$$

We approximate the standard deviation of the random variable with

$$\text{std}\,(D) \approx \sqrt{\left(\frac{\text{std}\,(x)}{\sqrt{k_x}}\right)^2 + \left(\frac{\text{std}\,(y)}{\sqrt{k_y}}\right)^2}.$$

Two samples

We want our test statistic to be a standard normal random variable, we already have an expected value of 0, so we just need to normalize

If we write

$$s_{ed} = \sqrt{\left(\frac{\text{std}(x)}{\sqrt{k_x}}\right)^2 + \left(\frac{\text{std}(y)}{\sqrt{k_y}}\right)^2}$$

Then our test statistic *s* is given by

$$s = \frac{\text{mean}(x) - \text{mean}(y)}{s_{ed}}$$

And our p-value is as before

$$p = (1 - f) = (1 - \int_{-|s|}^{|s|} \exp\left(\frac{-u^2}{2}\right) du)$$

Example

Assess the evidence that Japanese and US cars have the same MPG

We have a dataset with 249 Japanese cars and 79 US cars. The mean for Japanese cars is 20.1446 MPG and for US is 30.4810. The standard error for Japanese cars is 0.4065 and for US is 0.6872

The test statistic comes out to 12.94 which is so close to zero that you might not be able to get sensible numbers out of your software when computing it. Thus we can fairly comfortably reject the hypothesis that the two kinds of cars have the same MPG

Variations

We have been doing so-called **two-sided tests**. Because we have been computing

$$P(\{X > |s|\}) \cup P(\{X < -|s|\})$$

Which is the fraction of samples that would have produced a value of the test statistic greater than |s| or less than -|s|

If we have set up a test statistic that can only be positive or negative it may make sense to only look at $P(\{X > s\})$ or $P(\{X < s\})$. Doing this is called a **one-sided test**

Very often authors will use a one-sided test because it produces a smaller p-value and small p-values are a requirement for publication. Be on the lookout for this when reading the literature

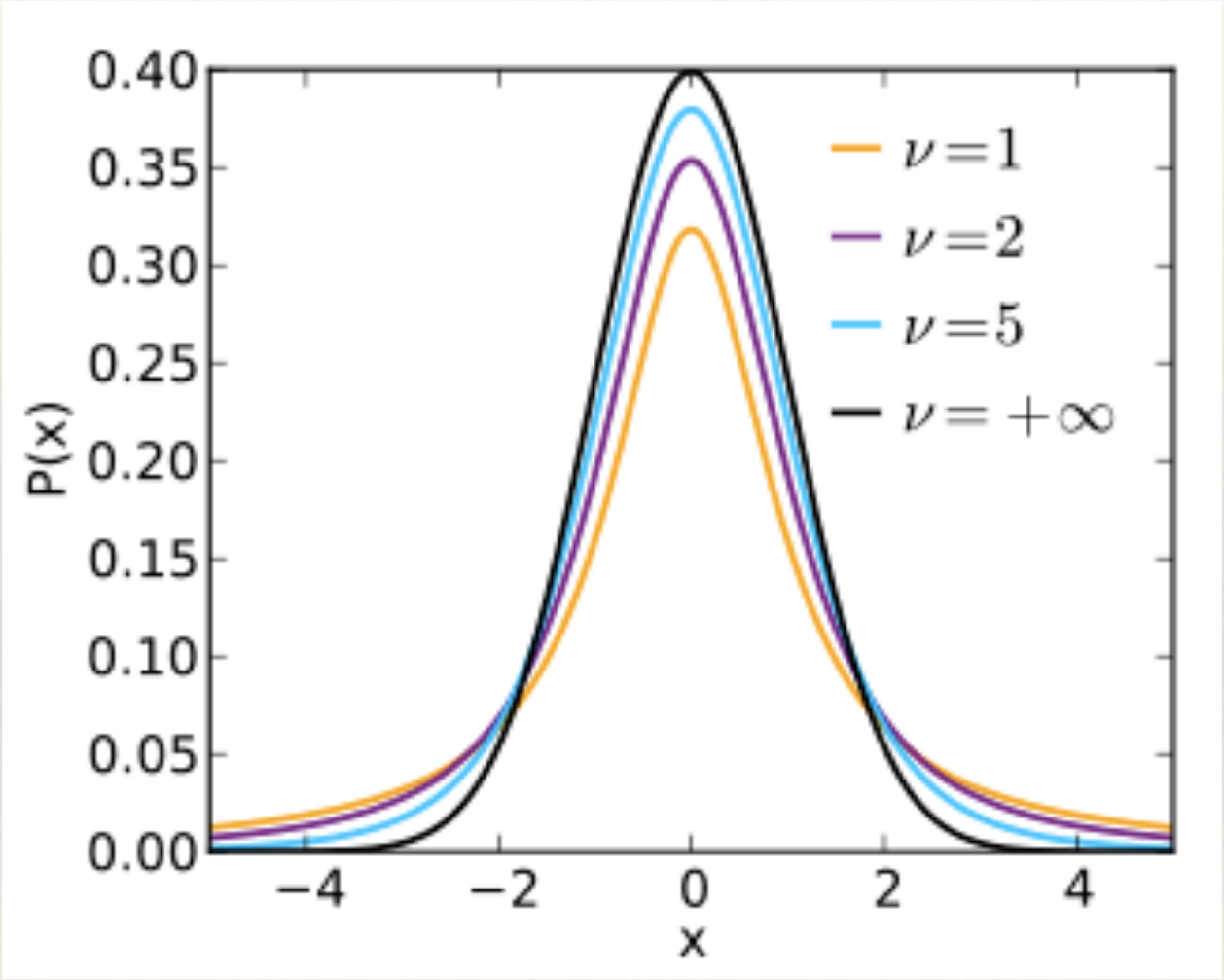
Z-tests and t-tests

We showed before that the sample mean follows a normal distribution with expected value equal to the population mean and standard deviation equal to the standard error. Using these assumptions, the tests we conduct are known as a **z-test**. The problem is that for a small sample size, our estimate of the standard error might not be great. 30 seems to be the magic number after which we can use a z-test. If the sample size is smaller however, we use what's known as a **t-test**

The distribution of the sample mean doesn't follow a normal distribution for small samples (less than 30). It follows what's known as a Student's t-distribution. The t-distribution has a parameter corresponding to the number of "degrees of freedom" and for our purposes it will be *k-1* where *k* is the sample size

When the number of degrees of freedom is small, the t-distribution has fatter tails than a normal distribution and when it's large it is very similar to the normal

t-tests


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
P(x)
−4
−2
0
2
4
x
$\nu=1$
$\nu=2$
$\nu=5$
$\nu=+\infty$

Chi-squared tests

Sometimes we want to test the hypothesis that a given set of data is well-described by a given model.

We could observe a bunch of rolls of a die and ask whether it is a fair die

We could observe a politician who swears during their speeches and ask whether or not a Poisson model with a given parameter is a good model for the swearing

We could ask if a given dataset appears to follow a normal distribution

In all cases, our model can be used to predict the theoretical frequencies of events

Chi-squared tests

If we have a set of *k* events which cover the whole space of outcomes $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$

We will have the observed frequencies $f_o(\varepsilon_i)$ and theoretical frequencies $f_t(\varepsilon_i)$ for each event ε_i

We form the following statistic

$$\sum_i \frac{(f_o(\mathcal{E}_i) - f_t(\mathcal{E}_i)^2}{f_t(\mathcal{E}_i)}$$

It turns out this statistic follows a so-called **chi-squared distribution** as long as the count of each event is 5 or more which means we can use it for hypothesis testing, namely the hypothesis that the model is a good fit for the data

Chi-square distribution

The chi-squared distribution has a parameter, the number of degrees of freedom which for our cases will be *k-1* or where *k* is the number of events, as in $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$

The probability that we get from the chi-squared distribution with *k-1* degrees of freedom for a test statistic *s* is interpreted as the fraction of samples which would have produced a dataset with a test statistic at least that large, denoted by *f*

As before if we take *p=1-f* we have a p-value which we can use to reject hypotheses

Example

We throw a 6 sided die 100 times and record the following frequencies of events. Is the die fair?

face	count
1	46
2	13
3	12
4	11
5	9
6	9

The theoretical frequency for a fair die is 100/6 for each face. Our chi-square statistic turns out to be 62.7 and there are 5 degrees of freedom. Our p-value is 3e-12, which means we would have to run this experiment on average 3e12 times to see a table this skewed by chance. So we reject the hypothesis that the die is fair

Dealing with high dimensional data

High dimensional data

We will be supposing that we have a dataset $\{\mathbf{x}\}$ of N data items, with each point being d numerical dimensions. We will thus think of our data items as vectors, which can be added, subtracted, and multiplied by constants

When we want to indicate the vector corresponding to the i-th data item, we will write $\mathbf{x}_i$

If we want to refer to the j-th component of the i-th data item, we will write $\mathbf{x}_i^{(j)}$

The mean

Just as in one dimension, we can have a mean for a dataset. In this context it makes sense to talk about the mean vector and calculate it as

$$\text{mean}\left(\{\mathbf{x}\}\right) = \frac{\sum_i \mathbf{x}_i}{N}$$

Where the j-th component of the mean vector is given by

$$\text{mean}(\{x\})^{(j)} = \frac{\sum_i x_i^{(j)}}{N}$$

Just as with the 1 dimensional mean we have $\text{mean}\left(\{\mathbf{x} - \text{mean}\left(\{\mathbf{x}\}\right)\}\right) = 0$ which is to say, we can produce a 0 mean dataset by subtracting the mean from each data item