March 7, 2018

CS 361: Probability & Statistics

Hypothesis testing

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

When we suppose a hypothesis is true, that allows us to reason about the dataset. Given the hypothesis as true, the p-value allows us to quantify how unusual our sample is. A low p-value means our sample is unusual.

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 (or two)

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 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 with D=mean(x)-mean(y), 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 standard normal 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 gives a p-value 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

Summary

The way we've done these last two examples is similar to how a lot of hypothesis testing is done. We derive some test statistic of the dataset and hypothesis and then query a distribution to see how extreme our sample is, supposing the hypothesis is true. If it is an extremely unlikely sample, we may reject our hypothesis

If our test statistic follows a standard normal distribution as in the last two examples we've looked at, then a test statistic that's far from zero means we have seen a very unlikely sample if our hypothesis is true, which means the f-value will be high and the p-value will be low

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

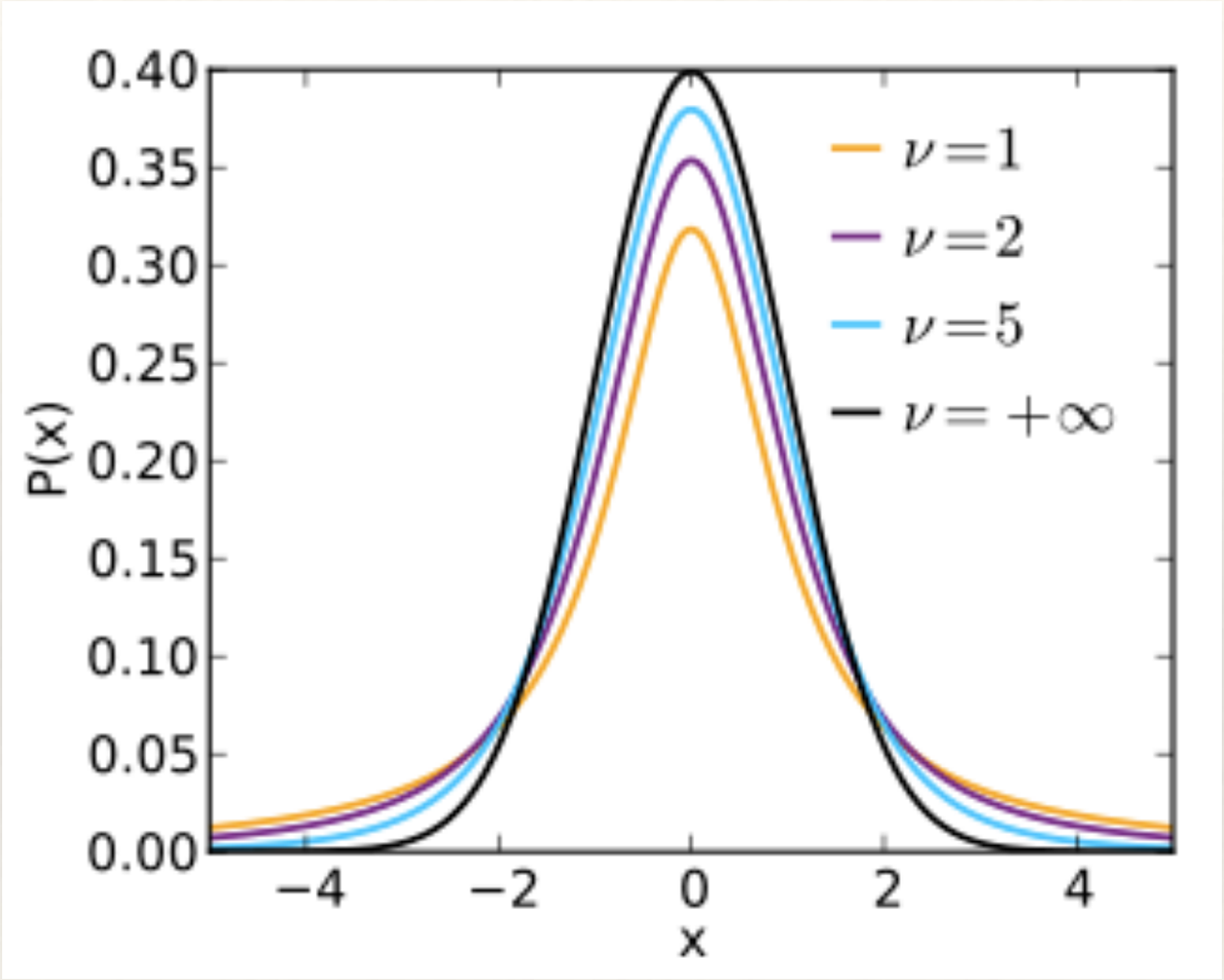
Z-tests and t-tests

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 N-*1* where *N* 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

Example: does this data follow a normal distribution?

To tell if data follows a normal distribution with a hypothesized mean and variance, we would use the chi squared test again.

To make the chi squared test applicable we need some notion of theoretical and observed frequencies, but since we don't have discrete values for the random variable in question, we will achieve this by breaking the range of values into some intervals

We want intervals to be the same size, except for the largest and smallest interval which run to infinity and negative infinity. We want as many intervals as possible, but no interval should have less than 5 data items in it. Our dataset will thus determine our interval strategy

We then compare the observed frequencies to the theoretical frequencies as before and compute the chi squared statistic

Inferring probability models from data

Binomial likelihood: Example

Suppose we have a coin with an unknown probability of heads. We flip the coin 10 times and observe 2 heads. What can we say about the probability of heads?

Estimating model parameters

Suppose we have a dataset $D=\{x_i\}$

Furthermore, suppose we know the type of distribution that models the data. E.g. Bernoulli, Binomial, Poisson, Normal, etc.

Suppose that we do not know the parameters of the model

We will use the letter theta θ to represent the model parameters

We will be trying to answer the following question:
What is a good value for θ, given the data?

Examples

If our data is given by a Bernoulli, Geometric, or Binomial random variable we are trying to find a good $\theta = p$

For the case of multinomial data we are looking for

$$\theta = (p_1, p_2, \ldots, p_k)$$

For a Poisson or exponential distribution, we are looking for a good estimate of

$$\theta = \lambda$$

For a normal distribution we are looking to estimate a good

$$\theta = (\mu, \sigma)$$

Overall, theta is an abbreviated way to refer to the parameters of a model we are trying to figure out

Estimating model parameters via maximum likelihood

Which theta to choose?

When we know θ, and we have a set of data, it is possible to get a numerical answer to the question "what was the probability of seeing that data, given theta"

$$P(\mathcal{D}|\theta)$$

When we have a dataset but don't know theta we can still write this quantity, but it will be a function of theta. We call this expression the **likelihood function**

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta)$$

One procedure for estimating the parameters of the model is to choose the theta that maximizes this expression. We call this the **max likelihood** estimate

Likelihood

Notice that the likelihood function is not a probability distribution over theta

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta)$$

For instance, if we know our data is a single sequence of coin flips, we know that the number of heads may be well-modeled by a binomial distribution, with parameters

$$\theta = (N, p)$$

If the sequence we observe is HHHHH, we know that N=5, so there's no inference to do to find out what N is. But evaluating the likelihood function for $\theta = p$

$$\mathcal{L}(\theta) = \binom{5}{5}\theta^5(1-\theta)^0$$ and we have for example $\mathcal{L}(1) = 1$ and $\mathcal{L}(.9) \approx 0.59$

L(1) + L(.9) > 1, so L is definitely not a probability distribution

Likelihood

We will assume that the data that we are observing are IID — independent and identically distributed. So we will be able to write

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta) = \prod_{i \in \text{dataset}} P(d_i|\theta).$$

As a notational convention, we will write $\hat{\theta}$ to indicate our actual, calculated estimate of θ

Thus our maximum likelihood estimate can be written as

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta)$$

Binomial likelihood

In N independent coin flips, we observe k heads. What is the maximum likelihood estimate for θ ?

We want to calculate

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta)$$

What is the likelihood function in this setup?

$$\mathcal{L}(\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

Binomial likelihood

We need to take a derivative and set equal to 0

$$\mathcal{L}(\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \binom{N}{k} \left[k\theta^{k-1}(1-\theta)^{N-k} - (N-k)\theta^k (1-\theta)^{N-k-1} \right]$$

Set $\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$ and solve for theta

$$k\theta^{k-1}(1-\theta)^{N-k} = (N-k)\theta^k (1-\theta)^{N-k-1}$$