# Statistics for dummies

Sample should be representative of the population, otherwise your statistical inference will be skewed/not generalize well. Application to ML: algorithms “learn” from the training data by minimizing loss; if the training data is not representative, then the algorithm will learn the wrong things. Imbalanced classes, small adjuncts – less data for certain classes means the algorithm will not learn those patterns as well. Noise.

If a sample is not representative, then it’s said to have bias (sampling bias). Is the main source of bias that ML engineers need to worry about imbalance in target classes/variable? And then you can handle with resampling techniques like over/undersampling, etc. But you don’t necessarily need to resample if you don’t want to – perhaps you want the majority class to be weighted more heavily. This applies to both regression and classification. Remember DPD.

Biased data introduces bias into your model/estimator (this is bias of an estimator, not inductive bias). Check discussion on bias/variance in k-fold CV. Although I guess inductive bias could be seen as a form of bias.

For train/val split, you can use StratifiedKFold for classification and for regression if you bin the target variable. Is stratification normally only along the target? What about features? Small adjuncts?

Bias is systematic favoritism that is present in the data collection process, resulting in lopsided, misleading results.

Sample results vary from sample to sample, and this variability needs to be reported (margin of error, confidence interval).

Sampling error is the error that results from inferring a population parameter (statistical inference) based on a sample. Margin of error is an estimate of this error. CI is estimate +/- margin of error. Note that sampling error is not related to bias.

CI represents a range of likely values for the population parameter based on your sample statistic.

I think statistical inference is inferential statistics and this is different from predictive statistics.

A hypothesis test is a technique for using data (a sample) to validate or invalidate a claim about a population. A hypothesis test accounts for sampling error. The elements of a population that are most often tested are

* The population mean (like MSE of a regression model?)
* The population proportion (like accuracy of a classifier model?)
* The difference in two population means or proportions (like accuracies between two classifier models?)

A p-value quantifies the strength of the conclusion of the hypothesis test. A small p-value indicates strong evidence against the null hypothesis. This is called a statistically significant result. There is a chance that a statistically significant result is actually a fluke.

The Central Limit Theorem (CLT) is central to estimating variability across samples and for hypothesis testing. CLT tells us that the sample mean has an approximate normal distribution as long as sample size is large enough (even if the variable itself is not normally distributed). CLT is true for other sample statistics as well, like sample proportion.

## Key terms

Data can be numerical (aka quantitative) or categorical (aka qualitative/nominal). Numerical falls into discrete or continuous. Ordinal data is a mix of numerical and categorical – the data falls into numeric categories, but the numbers also have meaning. Ordinal data is often treated as categorical with order mattering.

Dataset = collection of data taken from a sample.

Variable = any characteristic or numerical value that varies from individual to individual. Can represent a count, a measurement, or a category. Data is collected on variables.

Statistic = number that summarizes data collected from a sample.

Parameter = number that summarizes a variable from a population. Oftentimes we try to estimate population parameters from sample statistics.

## Random variables

You can think of each feature and the target in your dataset as a random variable. Each sample is a measurement of these random variables taken from a joint distribution.

Estimating the parameters that define the relationship b/w features and target is statistical inference (MLE, MAP).

# Statquest

Normal distribution

Multivariate normal distribution

Population parameters – the parameters that define the statistical distribution fitted to the entire population, e.g. population mean and population standard deviation for a population that is normally distributed.

Given a sample from the population, we want to estimate the population parameters (like how we want to estimate the model parameters from our training data in machine learning). We want to estimate population parameters to make inferences about the population based on the sample. We use these estimates to generalize findings from the sample to the entire population. Population parameters are used to develop predictive models that can be used to forecast future outcomes or behaviors.

We estimate the population parameters from the sample parameters. The more data we have, the more confidence we have in the estimates.

P-values and confidence intervals quantify the confidence we have in population parameter estimates.

Given two different samples, we can use statistics to quantify our confidence in how different they are. P-values and confidence intervals can tell us if the differences in two samples are statistically significant. This is called hypothesis testing.

Let represent the observed values/measurements. Let there be measurements (can represent population or sample size).

In statistics, denotes the sample mean (aka estimated population mean) while denotes the population mean. In either case, mean is the average of all measurements.

Population variance is given by . It is the average of the squared differences b/w the measurements and the mean. It measures the variance of around the mean.

Population standard deviation is .

Sample variance is . Sample standard deviation is .

compensates for the fact that we are calculating the differences from the sample mean instead of population mean. Otherwise, we would consistently underestimate the population variance.

A model explores the relationship between different attributes. We use statistics (hypothesis testing) to determine how useful or reliable our model is.

Hypothesis testing

Example hypothesis based on a sample: People taking Drug A need, on average, 15 fewer hours to recover than people taking Drug B.

Imagine that we test this hypothesis by giving drugs to additional samples of people, and that each new sample contradicts our original hypothesis (e.g. Drug A recovery time is always higher than Drug B recovery time). Then, we can reject the original hypothesis because we are given strong contradictory evidence.

Imagine the same scenario but this time, the additional samples support our hypothesis and the differences between samples are due to random things out of our control. For example, the additional samples show a difference of 14 hours between Drug A and Drug B recovery, 16 hours, 12 hours, etc.

We picked 15 because that was the data from the first sample. Based on the additional samples, we cannot reject the hypothesis. However, because the recovery time difference in the additional samples is different, we also cannot be confident in our original hypothesis. The best we can do is fail to reject the original hypothesis because the additional data is similar but not the same.

There are many reasonable hypotheses (14, 16, 12, 15, 12.5, 13.5, etc.). We cannot possibly pick one.

Therefore, we instead test the hypothesis that there is no difference in recovery time between the drugs – this is called the null hypothesis.

If the recovery time difference was similar, then we would fail to reject the null hypothesis.

The null hypothesis does not require any preliminary data: we do not need a preliminary sample to form the null hypothesis because the null hypothesis is unique.

To decide whether to reject or fail to reject the null hypothesis, we need to run the data through a statistical test. The output of the statistical test is the decision.

A statistical test needs three things: data, a null (or primary) hypothesis, and an alternative hypothesis. In our example with two drugs, the alternative hypothesis is simply the opposite of the null hypothesis: there is a difference in recovery time between drugs.

What are the steps for the statistical test?

1. Calculate the mean value for all data from both categories
2. Calculate the distances from all data to this common mean. This represents the null hypothesis.
3. Calculate the mean for each category
4. Calculate the distances from the data in each category to its corresponding mean. This represents the alternative hypothesis.
5. If the distances around the two means are much shorter than the distances around the common mean, this suggests that using two means to summarize the data makes more sense than using one. So, we reject the null hypothesis.
6. If the distances are around the same, this suggests the differences between the two means reflects random things we can’t control for. So, we fail to reject the null hypothesis. In ML, failing to reject the null hypothesis means that using two averages is overfitting the data.

When there are only two classes, the alternative hypothesis is the opposite of the null hypothesis. What happens when there are three or more classes?

The null hypothesis easily extends to there is no difference between drugs A, B, C, and again, we measure distances from all data to the common mean.

However, there are multiple alternative hypotheses. Two examples:

1. All three drugs are different. Calculate means for each drug.
2. A and B are the same while C is different. Calculate two means, A+B and C.

The output of the statistical test depends on our alternative hypothesis – the options are the same, but the decision on the null hypothesis can be swayed by the alternative hypothesis.

If we reject the null hypothesis, then we could say we reject it in favor of the alternative hypothesis we chose. However, we could not say we accept the alternative hypothesis because other alternatives may be better. There are too many alternative hypotheses to test, so that’s why the output of the statistical test is always the decision to reject or fail to reject the null hypothesis.

p-values (for hypothesis testing)

p-value helps us decide if we should reject the null hypothesis or not.

Example: Drug A and Drug B are the categories (categorical feature), and cured/not cured are the targets (classification).

We give A and B to different samples of the population. Some people in each group are cured, and some aren’t cured.

p-values are numbers between 0 and 1 that quantify how confident we should be that A is different from B. 0 = different, 1 = same.

How small does a p-value have to be before we are sufficiently confident that A is different from B?

A common threshold for p-value is 0.05: If there is no difference between A and B, and if we run this same experiment a bunch of times, then only 5% of those experiments would result in the wrong decision.

What does this mean?

If we gave A to both groups, then most of the time, p-value will be very large (close to 1). Sometimes, p-value will be very small even though there is no difference between the groups. This is a false positive.

In 5% of experiments, p-value will be less than 0.05, that is, we have a 5% false positive rate (FPR).

We can set the threshold differently based on our tolerance for false positives, but we often set it to 5% because oftentimes, getting FPR below 5% is too costly.

If we run our experiment and p-value is less than 0.05, we decide that A and B are different.

However, p-value does not tell us how different A and B are. The difference can be tiny or huge. A small p-value does not imply that the effect size between A and B is large.

How to calculate p-values

For a hypothesis, and for a measurement, the measurement p-value is equal to the sum of 3 things:

1. The probability that random chance would result in the observation
2. The probability of observing something else that is equally rare
3. The probability of observing something rarer or more extreme

Coin flip example: We flip a coin twice and get heads both times (the observation). Null hypothesis: this is a normal coin.

We fail to reject the null hypothesis.

Height example: We have height measurements (sample). From these measurements, we assume the underlying distribution is normal, and we estimate the population mean and standard deviation (model).

Given a new height measurement, our null hypothesis is this measurement comes from the normal distribution defined by our estimated mean and standard deviation.

p-value is the area under the curve corresponding to heights equally or more extreme than the new measurement (which is equal to the probability of measuring a height equally or more extreme).

If p-value < threshold, then we reject the hypothesis and say that another distribution would make more sense.

Recall: p-values and confidence intervals quantify the confidence we have in our population parameter estimates.

Note: even if the probability of a specific measurement is small, the p-value is what matters for hypothesis testing. For example, measuring a height very close to the mean might have a relatively small probability, but the p-value would be close to 1.

It’s typically recommended to use two-sided p-values: Given a distribution, extreme values are defined as further from the mean in both directions.

In some cases, it may be tempting to use one-sided p-values. For example, let’s say we’re testing recovery time with SuperDrug. We have measurements of recovery time (distribution) without SuperDrug. We collect data that says average recovery time of SuperDrug is X days. If we use one-sided p-value, then we need to pick the direction. Since shorter recovery time is better, we say that only values <= X are extreme. All values > X are not extreme. This may be fine if SuperDrug does nothing or indeed shortens recovery time (X < original mean).

What if SuperDrug makes things worse (X > original mean)? With the same one-sided p-value, p-value will be very large, and we will make the erroneous decision that SuperDrug has no effect when it’s making things worse.

This one-sided p-value only checks if a distribution to the left of the original distribution makes more sense and does not check if a distribution to the right would make more sense.

Is p-value calculated per new measurement? Or aggregated over all measurements?

p-values for categorical vs. numerical features, p-value for multiple categories?

# Stuff to study

Biased/unbiased estimator, variance in an estimator

MLE

MAP

Confidence interval

Distributions

Descriptive stats like mean, stdev, variance, correlation, covariance

t-tests, f-tests, ANOVA, chi^2

Expected value

Integrating over a distribution

Calculating mean of a distribution (integral of x\*P(x) I think)

Central limit theorem – this explains why we often see the normal distribution in nature.

is a random variable (a random process, where each outcome is associated with a number).

Add samples of this variable,

The distribution of this sum looks more like a normal distribution as .

<https://www.youtube.com/watch?v=zeJD6dqJ5lo&t=255s>