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.

# Statistical inference/inferential statistics

Statisticians study populations. They measure/count/classify population characteristics, treating them as random variables that come from probability distributions; find probabilities and proportions; and estimate population parameters.

Typically, we cannot collect data on the entire population. Instead, we collect and study a sample of data and generalize our findings to the population. This is statistical inference.

Because we only have a sample, we need to account for variability in sample statistics across different samples when generalizing/estimating population parameters (margin of error, confidence intervals, hypothesis testing).

How does this apply to ML?

In ML, we also have a sample – it is our dataset. Each attribute in the dataset is a random variable (features and target alike), and each data point is a sample taken from a joint distribution of attributes.

In ML, we also want to estimate parameters from our sample, but instead of estimating population parameters, we estimate model parameters. We want our model to generalize to the “population” (that is, we want the model to generalize well to new samples taken from the underlying data-generation process), so this is still statistical inference but with the additional goal of building a predictive model.

In ML, we also want to account for variability across datasets. The model parameters and predictions will vary when we train it on different datasets (this is the variance part of the bias-variance tradeoff). We can apply statistical inference methods like confidence intervals, cross validation, and bootstrapping to quantify the variance of our model.

# Distributions and the central limit theorem

## Random variables

Numerical random variables represent counts and measurements. Discrete RVs (finite or countably infinite) typically represent counts while continuous RVs (uncountably infinite) typically represent measurements.

A RV is random, but that doesn’t mean that every outcome is equally likely. RV is defined by its distribution. Distributions for discrete RVs are called discrete probability density functions or probability mass functions; distributions for continuous RVs are (continuous) probably density functions.

* Mean of RV is average of all outcomes over the long term and is denoted by for . is sample mean.
* Variance of RV is interpreted as average squared distance from the mean for all outcomes over the long term and is denoted by for . Standard deviation is and represents the average distance from the mean. and are sample variance and standard deviation.
* Both mean and standard deviation are affected by outliers, while median and interquartile range are not (or less) affected.

is necessary for to be an unbiased estimate of the population variance (dividing by consistently underestimates ).

### Binomial distribution

This is the most well-known discrete RV. Related distributions: Bernoulli, multinomial, categorical.

The RV has a binomial distribution defined by and when

1. Fixed number of trials
2. Each trial has two outcomes: success and failure
3. The probability of success for each trial is the same,
4. The trials are independent

is the number of successes in trials. For example, the number of heads in 10 flips of a fair coin is a binomial RV with and .

For successes,

When and , you can approximate the binomial distribution with the normal distribution for easier calculations (CLT). The approximate normal distribution is .

### Normal distribution and standardizing

The standard normal distribution, , is also known as the Z-distribution.

Values on the Z-distribution are called z-values, z-scores, or standard scores. They represent the number of standard deviations above or below the mean, e.g. +1 means one standard deviation above the mean.

To standardize from to ,

z is both the standard score and the standardized RV with and . This is true for any distribution, not just the normal distribution, which means standard scores don’t necessarily come from a normal distribution (although the Z-distribution is specifically the standard normal distribution).

99.7% of the standard scores on the Z-distribution lie within -3 to +3.

### t-distribution

The t-distribution generalizes the Z-distribution:

* It has zero mean, but its standard deviation depends on , the degrees of freedom. , where is the sample size.
* For small , the t-distribution has fatter tails (more probability in the tails). As , the t-distribution approaches the Z-distribution.
* t-distribution is typically used to model a sampling distribution rather than a distribution of individuals. If the individual distribution is normal, or if you know the population standard deviation and sample size is large (CLT), then you can use a Z-distribution. Otherwise, use t-distribution (more on this in confidence intervals/hypothesis testing).

## Sampling distributions and the central limit theorem

### Sampling distribution of sample mean

Let be a numerical random variable that describes a characteristic of the population.

Let us take a sample of the population and calculate the mean of for the sample. Call this . If we keep collecting samples of the population and calculating , we will see that is different for each sample because itself is a random variable.

represents all individuals in the population. represents all sample means from the population.

Random variables are described by distributions, and because is a sample statistic, we say that it is described by a sampling distribution.

If has mean and standard deviation , then has mean and standard error .

* Standard error = standard deviation of a sampling distribution. Standard error decreases as sample size increases; in other words, sample mean becomes a more precise estimate of population mean.
* The mean of the sample means is equal to the population mean (that is, is an unbiased estimator for )

If is normal, then is normal.

If is not normal or if its distribution is unknown, then we can’t say that is normal. HOWEVER, if is large enough, we can say that is approximately normal.

This is the central limit theorem:

* If is not normal, the shape of the sampling distribution of is approximately normal as long as is large enough.
* As a rule of thumb, but it depends on the distribution of . The larger the , the better the approximation.

In other words, even if is not normal, as long as is large enough,

In the common form of the CLT, the random variables in each sample, must be iid.

Example.

Let represent the roll of a fair die. It has a uniform distribution from 1 to 6.

Let represent the average of 50 rolls. Its distribution is approximately normal.

A graph of a number of rolls

AI-generated content may be incorrect.

A graph of a number of rolls

AI-generated content may be incorrect.

Note that will still be approximately normal even if is not a fair die.

### Sampling distribution of sample proportion

Let be a categorical random variable. is the proportion of the population that falls into the category of interest.

Take a sample of size from the population and calculate the sample proportion, .

You can think of as the number of trials. For each trial, the probability of success (data point is in the category of interest) is , the population proportion.

Therefore, the number of data points in the sample that falls into the category is a binomial random variable defined by the parameters and .

The sample proportion, , is simply the number of data points in the category divided by the sample size, . Therefore, the sampling distribution of is also binomial with the following mean and standard error:

The mean of is the population proportion (again, an unbiased estimator). Standard error shrinks as sample size increases (again).

By CLT, the sampling distribution of is approximately normal if is large enough.

If

Then

# Confidence intervals

# Hypothesis

I think the gist of it is – take a sample, calculate the sample mean and standard deviation/error. Or calculate the sample proportion, and use the sample proportion to calculate the sample standard error. Use t-distribution to calculate confidence intervals for population mean/proportion based on the sample mean/proportion.

For sampling distribution of sample mean, you can use Z-distribution if sample size, n, is >= 30 (rule of thumb). This is central limit theorem.

For sampling distribution of sample proportion, you can approximate Z-distribution if np >= 10 and n(1-p) >= 10. This is central limit theorem.

# Data bias (look this up)

I assume that sample mean and proportion are only unbiased estimates if the samples are representative of the overall population.

In sampling distributions, what if are not iid? What if are not independent? How to measure? Correlation/covariance/orthogonality vs. independence. Probability.

I think in all discussions, we assume the RVs are iid. But I should understand what happens if they are not.

Training, validation, test, estimate of generalization error; these need to be representative. Cross-validation, bootstrapping.

# 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>