Samples and sampling distribution

Population

Each individual value in the population is the value of a random variable. Thus a population corresponds to a random variable. The density (PDF) and number characteristic of the population is just the PDF and number of characteristic of a random variable X.

Random samples

In mathematical terms, given a random variable with distribution F, a random sample of length n is a set of n independent, identically distributed (iid) random variables $X_1, X_2, \ldots . X_n$ with distribution F. By definition, a random sample is already IID, so multiplication rule of PDF applies. The realizations of these samples $x_1, x_2, \ldots x_n$ are not different n features of an observation in machine learning. Instead x_i here indicates a row in a data set.

How to obtain random and iid samples

For finite population, sampling with replacement gives the random sample (with iid feature). While sampling with replacement is simple in simulation, it is not convenient in practice. Therefore, we take a sample as random sample when the sample size n is much smaller than the number of elements in population (10% rule).

Examples of random sample

- Within a huge population, a real population of a country, we want to do sampling on the proportion of people voting for a specific candidate. We take a small sample, e.g. 100, as as to satisfy the iid requirement, and obtain a sample proportion of $\hat{p} = 0.55$. Note this is not the unknown population proportion p. The sample proportion $\hat{p}$ is a so-called statistics.
- We are measuring the length of an object. It is impossible to measuring infinite number of times. So we measuring n times to have a sample, and then calculate the mean of these measurements: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. $\bar{X}$, called sample mean, is a statistic on the sample.

Sampling distribution

As stated earlier, a population corresponds to a random variable X. Population PDF, population variance, population STD are therefore defined for the random variable X. For a sample of size n, the population variance, or STD, will approach constant as n increases, but not decreases.

Now consider the distribution of a statistic, sample mean $\bar{X}$. Assuming it is a normal distribution (center limit theorem), then it is easy to obtain its variance as:

$$Var(\bar{X}) = Var(\frac{\sum_{i=1}^{n} X_i}{n}) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Thus the STD of the sampling distribution, i.e. the distribution of $\bar{X}$, is $\frac{\sigma}{\sqrt{n}}$. **The STD of sampling distribution of a statistic (sample mean here) is called the standard error (SE).** Note SE decreases as n increases, which is different from the population STD. Once SE of a sampling statistic is calculated, we then can calculate confidence intervals, and do hypothesis testing for our samplings. From above, we know SE is just a special STD. When we talk about STD, we refer to the distribution of population or random variable X. When we talk about SE, we refer to the distribution of a statistic on a sample, e.g. the sample mean $\bar{x}$. In practice, the population variance σ^2 is not known, and thus is replaced by the sample variance $\hat{\sigma}^2$. However, be careful that sample variance is not the STD of sampling distribution. It is just the approximate version of population STD. Moreover, it is not that any STDs will approach constant when n approaches infinity. Only population STD (the STD of random variable X), or sample variance, approaches constant when n becomes very large.

In the voting example discussed earlier, the SE is $\sqrt{\frac{p(1-p)}{n}}$, where p is the population proportion voting for a specific candidate. As it is not known, it is replaced by the sample proportion.

The above results about sample mean assume that samples are drawn from an identically, independent distribution (iid). If we drop the independent assumption but consider an averaged correlation p among samples, then we have https://en.wikipedia.org/wiki/Variance (https://en.wikipedia.org/wiki/Variance), https://en.wikipedia.org/wiki/Covariance (https://en.wikipedia.org/wiki/Covariance)

$$Var(\bar{X}) = p\sigma^2 + \frac{1-p}{n}\sigma^2$$

Although the variance is now related to correlation, the idea of using more samples to reduce variance is still valid. First, we can make sure the samples from a bootstraping process are as independent as possible. Second the variance is decreasing as the number of samples increases.

Examples of sampling distributions

Assuming $X_1, X_2, \ldots X_n$ is a sample from population $N(\mu, \sigma^2)$, and $\bar{X}$ is the sample mean and S^2 is sample variance, then we have:

- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, $\bar{X}$ and S^2 are independent.
- $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$ Note t-distribution is for small-sized sample where we cannot use sample variance to replace population variance. Thus we cannot use z-testing but have to use t-testing. Because t-testing is for small size and we know that STD of sample mean get smaller as size increases, we know that t-distribution is just a fatter normal distribution.
- $\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$

Statistics estimation

Point and interval estimates

An point estimator is a statistic defined on random samples. Maximum likelihood estimation (MLE) is a typical way for point estimation. Normally the population distribution is known but only parameters are unknown. Unlike the sample mean, the estimator using MLE is sometimes not an explicit function of $X_1, X_2 \ldots$. Interval estimates are related to estimating standard error of sampling distribution described in previous chapter. Two fundamental problems in statistical inference: estimation and hypothesis testing. The way to estimate can be understood as calculating the expectation value of a statistic, usually in forms of calculating maximum likelihood or minimum cost function. Hypothesis testing is just the application of the point and interval estimates. For example, in Z-testing, only when we know the interval, then we can do hypothesis testing with p-value.

Biased and unbiased estimate

In statistics, the bias (or bias function) of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated. An estimator or decision rule with zero bias is called unbiased. Otherwise the estimator is said to be biased. General MLE estimate of population variance from a sample is biased. However, we can make this unbiased by multiplying a factor.

MLE in machine learning

The applications of many machine learning algorithms are essentially doing statistics estimation. Linear regression is a point estimate (we estimate a function point in function space). Interval estimate can also be done with linear regression. The typical parameter estimating approach, MLE, can be used to derive many supervised learning algorithms such as linear or logistic regressions. When combined with Bayes rules, unsupervised learning algorithms such as mixtures of Gaussians/naive Bayes can be derived with MLE. In fact, a two-step iterative approach called expectation maximization is used, although the key is still MLE. See details on these applications of MLE in the notes of machine learning.

Hypothesis Testing -- significance testing

An intuitive example for hypothesis or significant testing.

- A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to a neurological stimulus, and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats' response times is 1.05 seconds with a sample standard deviation of 0.5 seconds. Do you think the drug has an effect on response time?
- Be careful the sample variance is not the standard deviation of the sampling distribution. There is a $\frac{1}{\sqrt{N}}$ difference. Assuming null hypothesis that the drug has no effect, i.e., we assume that the observation of small response time 1.05s OR LESS is just from measuring volatility. In other words, if we measure a lot of average mean response time we can easily obtain a 1.05s or less response time even without the drug effect. This is however, just our assumption. If we find that probability to obtain small response time $<= 1.05$ is not that appreciable but very tiny, then we should reject our null hypothesis.

The p value.

- It is the **probability** of **finding the observed, or more extreme**, results **when the null hypothesis of a study question is true** – the definition of 'extreme' depends on how the hypothesis is being tested. Three conditions for the definition of p value are bold.
- The significance level for a given hypothesis test is a value for which a p value less than or equal to is considered statistically significant. Typical values for are 0.1, 0.05, and 0.01. These values correspond to the probability of observing such an extreme value by chance.

Hypothesis testing with numerical simulation

- The following is a summary of hypothesis testing with numerical simulations. Specific exercises are in the notes "Statistical Thinking in Python_Part 2".
- Numerical approach is very flexible in doing hypothesis testing. We can do many types of testing without knowing any information of the closed form probability densities. Moreover, this can be done with almost a standard steps.

General steps of hypothesis testing with numerical approach

- Find a reasonable statistics, which could be the observable, or one of several observables, or function of observables. For example, it could be mean, mean difference, variance, correlation coefficients, and anything else observed from data.
- Once statistics is fixed, we need numerically create many REPLICATES of the predetermined statistics. Depending on specific situation, permutation and bootstrap approaches are often used to generate statistics replicates under the (null) hypothesis. All the generated statistics replicates essentially provide a histogram or PDF for a random variable corresponding to the statistics. With this PDF or histogram we can calculate the p value in order to do the hypothesis testing.
- Calculation of p value. If the observed statistics is less than most of the generated statistics replicates, then we have $p = P(x \leq x_0)$, where x_0 is the observed statistic. Otherwise $p = P(x \geq x_0)$. If p is very tiny, then it indicates that observed x_0 is not from observing volatility. In other words, this observation is statistically significant, and thus the original null hypothesis should be rejected.

Key points in generating statistics replicates

- The statistics replicates are generated under the assumption that (null) hypothesis is true. Only after we clearly state the null hypothesis, we can then generate replicates under this hypothesis.
- When testing the same distribution of two samples, the better way is to join the two samples and do permutation. It is more accurate than bootstraping. It is not always necessary to joint the samples and then perform permutation. If we can assume one same is independent from the other, then permutation on one sample is enough.
- Bootstraping, though less accurate, is more flexible. When we cannot assume same distributions, but only assume the same other quantities such as means, etc., then it is a better way.

A general A/B testing framework

- A/B testing applies to typical problems such as when we examining the effect of the upgrading of a website. For example, we may examine the whether the spending time of visitors on the website has changed before and after the upgrading.
- Some other examples, though not with a before and after features, can also be solved with the A/B testing approach. For example, when we want to check whether a congress voting results has strong effect of party affiliation, whether the strike forces of two frogs has same distributions...
- The key idea to solve this type of problems is: we assume 'there is no effect', 'there is no party affiliation', 'the strike force is with same distribution',.... then we shuffle/permute, or bootstrap to obtain statistics replicates, and then calculate p value to check whether the observed results are statistically significant.

Hypothesis testing using closed-form PDF

For distributions with closed-form PDF, p value can be calculated with integration over the PDF. Although this approach is not as flexible as the way using numerical simulation, it can provide a lot of insights. This can be found in standard statistics books.

Test for the mean of normally distributed population

Test for the variance of normally distributed population

- SINGLE normal distributed population
  - χ^2 testing. Check an example in https://www.khanacademy.org/math/ap-statistics#tests-significance-ap (https://www.khanacademy.org/math/ap-statistics#tests-significance-ap) for χ^2 tests for categorical data.
- Two normal distributed population
  - F-testing.

HT types include non-normally-distributed population

Cross entropy

https://en.wikipedia.org/wiki/Cross_entropy (https://en.wikipedia.org/wiki/Cross_entropy)

Cross-entropy error function and logistic regression

$$H(p,q) = \sum_i p_i \log \frac{1}{q_i} = -\sum_i p_i \log q_i = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

where p_i and p_i are respectively the true labels and predicted labels. We also have $p_{y=1} = y, p_{y=0} = 1-y$ and $q_{y=1} = \hat{y}, q_{y=0} = 1-\hat{y}$.

Log loss provides a steep penalty for predictions that are both wrong and confident, i.e., a high probability is assigned to the incorrect class.

Relation to log-likelihood

In classification problems we want to estimate the probability of different outcomes. If the estimated probability of outcome i is q_i, while the frequency (empirical probability) of outcome i in the training set is p_i, and there are N samples, then the likelihood of the training set is

$$\Pi_i q_i^{N p_i}$$

so the log-likelihood, divided by N is

$$\frac{1}{N} \log \Pi_i q_i^{Np_i} = \sum_i p_i \log q_i = -H(p, q)$$

so that maximizing the likelihood is the same as minimizing the cross entropy.

Relation of surprisal, entropy, cross-entropy and cross-entropy loss

https://medium.com/@vijendra1125/understanding-entropy-cross-entropy-and-softmax-3b79d9b23c8a (https://medium.com/@vijendra1125/understanding-entropy-cross-entropy-and-softmax-3b79d9b23c8a)

Surprisal

Degree to which you are surprised to see the result. Now, if y_i is the probability of ith outcome then we could represent surprisal as:

$$s = \log \frac{1}{y_i}$$

Entropy:

After knowing the surprisal for individual outcomes, we would like to know surprisal for the event. It would be intuitive to take a weighted average of surprisals. Taking the probability of each outcome as weight makes sense because this is how likely each outcome is supposed to occur. This weighted average of surprisal is nothing but Entropy. If there are n outcomes then it could be written as:

$$e = \sum_i y_i \log \frac{1}{y_i}$$

Cross-Entropy:

What if each outcome's actual probability is p_i but someone is estimating probability as q_i. In this case, each event will occur with the probability of p_i but surprisal will be given by q_i in its formula (since that person will be surprised thinking that probability of the outcome is q_i). Now, weighted average surprisal, in this case, is nothing but cross entropy and it could be scribbled as:

$$c = \sum_i p_i \log \frac{1}{q_i}$$

Check the link below for an animation showing how cross entropy is bigger when p_i is away from q_i.
https://www.desmos.com/calculator/zytm2sf56e (https://www.desmos.com/calculator/zytm2sf56e)