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

https://en.wikipedia.org/wiki/Sampling_distribution (https://en.wikipedia.org/wiki/Sampling_distribution)
In statistics, a sampling distribution or **finite-sample distribution** is the probability distribution of a given **random-sample-based statistic**. If an arbitrarily large number of samples, each involving multiple observations (data points), were separately used in order to compute one value of a statistic (such as, for example, the sample mean or sample variance) for each sample, then the sampling distribution is the probability distribution of the values that the statistic takes on. **In many contexts, only one sample is observed, but the sampling distribution can be found theoretically.**

Sampling distributions are important in statistics because they provide a major simplification en route to statistical inference. More specifically, they allow analytical considerations to be based on the probability distribution of a statistic, rather than on the joint probability distribution of all the individual sample values.

Further comments

- Sampling distribution is a FINITE-sample distribution. Each sample value, e.g. sample mean, is obtained by taking the average of FINITE number of data points.
- If we use infinite number of data points to obtain a sample statistic, then the sampling distribution of any cases should be normal distribution according to central limit theorem.
- For the finite-sample distribution (i.e. the sampling distribution we referred to), we thus usually don't have normal distribution. Only in some special cases, the sampling distribution might have a normal distribution. For example, if the corresponding population X is normally distributed, then the sampling distribution of sample mean $\bar{X}$ is Gaussian, as detailed in the example below. If the sample size is very large, we might also approximate its distribution as Gaussian.
- Because in many cases the sampling distribution is not normal, in interval estimate we usually don't have a symmetric standard error distribution.
- **Be careful of the term 'sample size' might indicate different things. Sometimes we mean the number of data points to calculate a single statistic, and sometimes we mean the number of statistic calculated.**

Example 1: Sampling distribution of sample mean in a normal population

Distribution of sum of normally distributed and independent random variables

- If X_i for $i = 1, 2, \ldots n$ are normally distributed and **independent**, then its sum $X_1 + X_2 + \ldots + X_n$ is still normally distributed with a mean of $\mu_1 + \mu_2 + \ldots + \mu_n$ and variance of $\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2$. This actually can also extend to the linear combination of X_i. See proof in standard statistic books.
- Note that the n here can be small and we don't need the central limit theorem for $\sum_i^n X_i$ to satisfy normal distribution.
- Thus, the distribution of a statistic, sample mean $\bar{X}$ is a Gaussian, with the variance

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

Standard deviation (STD) vs standard error (SE)

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.
From previous section, we know the STD of the sampling distribution is $\frac{\sigma}{\sqrt{n}}$. The special STD for sampling distribution is also called standard error (SE). Obviously, SE decreases as n increases, which is different from the population STD. 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.

From independence to dependence

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.

Example 2 Sampling distribution of sum of sample variance in a normal population

- If X_i for $i = 1, 2, \ldots n$ are normally (**to be specific: $\in N(0, 1)$) distributed and independent**, then the sum $X^2 = X_1^2 + X_2^2 + \ldots + X_n^2$ no longer satisfies the normal distribution. We instead have $X^2 \sim \chi^2(n)$ distribution with n degree of freedom.
- Check the shape of $\chi^2(n)$ distribution elsewhere. As n increases, the χ^2 will be more like a Gaussian (central limit theorem). This can also be understood this way: If there are only two degrees of freedom, then it is highly unlikely to obtain big X^2 values and thus PDF is skewed to the lower X^2 value. As the degree of freedom becomes larger and larger, we have the sum of the square of so many

normally distributed random variables. This the probabilities to obtain either big or small X^2 become equal.

- χ^2 distribution is strongly related to γ distribution. See other notes.
- Assume $X_1^2 \in \chi^2(n_1)$ and $X_2^2 \in \chi^2(n_2)$ and are independent to each other, then $X_1^2 + X_2^2 \in \chi^2(n_1 + n_2)$.
- For $\chi^2(n)$, we have $E(\chi^2(n)) = n$ and $\text{Var}(\chi^2(n)) = 2n$.
- **As in the sampling distribution of sample mean $\bar{X}$ where the x-axis is just the $\bar{X}$, in the χ^2 distribution, the x-axis is X^2 defined earlier. However, it is not divided by number of points to calculate a single statistic.**

Example 3 t-distribution and F-distribution

- If $X \in N(0, 1)$ and $Y \in \chi^2(n)$ and are independent to each other, then $t = \frac{X}{Y/n}$ satisfies a t-distribution $t(n)$.
- If $U = \chi^2(n_1)$ and $V = \chi^2(n_2)$ and are independent to each other, $F = \frac{U/n_1}{V/n_2}$ satisfy a F-distribution $F(n_1, n_2)$ with degree of (n_1, n_2).

Example 4 Summary of sampling distribution

The examples above are about the sampling distributions **that are ALL related to a normal population**. In other words, either χ^2, t or F distribution, all have an underlying normal population. Otherwise, it is usually hard to obtain an analytical form of distribution.

Now 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 the following:

- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ or when σ^2 is not available or not accurate due to small sample size, then $\bar{X} \sim t(\mu, \frac{S^2}{n})$, where σ^2 is replaced by S^2.
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, $\bar{X}$ and S^2 are independent.
- $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$
- $\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$

The different forms here arise from the following fact: (1) First we no longer assume the $N(0, 1)$ distribution. (2) Second, definition of variance S^2 is different from that of $X^2 = X_1^2 + X_2^2 + \ldots + X_n^2$ in previous sections.

Finally, we note that for even simpler Bernoulli population, then it is easy to shown that the sampling distribution of sample mean $n\bar{X}$ (multiplied by n) follows a binomial distribution. https://en.wikipedia.org/wiki/Sampling_distribution (https://en.wikipedia.org/wiki/Sampling_distribution)

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. To do testing, we just calculate their replica and do not need know their closed form of sampling distribution, or such a closed-form sampling distribution may not be existed at all.**
- 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 sampling distributions

Z-testing and t-testing

If we know the closed form **sampling distribution** of a statistic, we can apply them in hypothesis testing by integration. Because we are handling sampling distribution in hypothesis testing, be careful of the difference of SE and STD mentioned earlier for the case of Gaussian sampling distribution. However, note that not so many statistics follow a normal sampling distribution, including the sample mean. For example, if the population corresponds to a Bernoulli variable, then the sampling distribution of sample mean will be binomial.

If a statistic, sample mean, follows a normal distribution, then we usually apply a hypothesis test called z-test, where z is just shifted and normalized sample mean. The z-test is best used for greater than 30 samples (meaning how many data to calculate a single statistic?) because, under the central limit theorem, as the number of samples gets larger, the samples are considered to be approximately normally distributed. **Comments:** Or if the sample size (the number of data to calculate a single statistic) is not large, but the underlying population is normally distributed, then we still can apply z-testing.

For small-sized sample where we cannot use sample variance to replace population variance, we can use t-testing where t-distribution is used. 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.**

χ^2 testing

See details in the notes about feature selection in machine learning.

Test for the variance of a normal population

— Page 162, Statistics in Plain English, Third Edition, 2010.

Relation of surprisal, entropy, cross-entropy,log-likelihood and KL divergence

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:

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

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:

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

Here is an example to show that **entry is an impurity error metric or an quantity to show the disorder**. In the classification problem with decision trees, we minimize the entropy in the whole region or its sub-optimal regions. If a sub-region is perfectly classified, then the entropy in that region will be 0. Take binary classification for example, the entropy will be $0\log 0 + 1\log 1 = 0$, where $0\log 0$ is treated as 0 by convention.

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)

Here is an example to show that why we introduce cross-entropy. Take the binary classification problem using logistic regression, the cross entropy is:

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

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

We see that the cross entropy provides **a steep penalty for predictions that are both wrong and confident**, i.e., a high probability is assigned to the incorrect class. So to minimize cross-entropy, we can obtain optimized classification result. If we classify perfectly, then cross-entropy will be minimized.

Now we check the relation of cross-entropy and the commonly used 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

$$\Pi_i q_i^{Np_i}$$

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.

Kullback–Leibler (KL) divergence

- KL divergence (**also called relative entropy**) is a measure of how one probability distribution is different from a second, reference probability distribution.
- (Wikipedia) For discrete probability distributions P and Q defined on the same probability space, the KL divergence between P and Q is defined to be

  $$D_{KL}(P\|Q) = -\sum_{x\in X} P(x)\log\left(\frac{Q(x)}{P(x)}\right)$$

  Sometimes the sign is changed when numerator and denominator is flipped in the above definition. The KL divergence is defined only if for all x, $Q(x) = 0$ implies $P(x) = 0$ (absolute continuity). Whenever $P(x)$ is zero the contribution of the corresponding term is interpreted as zero because

  $$\lim_{x\to 0^+} x\log(x) = 0$$

Relation between cross-entropy and KL divergence

Comparing the definition of KL divergence and cross-entropy (both in discrete case), which are respectively $D_{KL}(P\|Q) = -\sum_i P_i \log\frac{Q_i}{P_i}$ and $H(P,Q) = \sum_i p_i \log\frac{1}{Q_i}$, we can obtain the relation

$$H(P,Q) = H(P) + D_{KL}(P\|Q)$$

where $H(P)$ is the entropy of P, which is $H(P) = -\sum_i P_i \log P_i$. In classification problems, $H(P)$ is a constant and has no contribution to the gradient for parameter updating. Therefore we have the following conclusion.

- In the classification problems of machine learning, using either KL divergence or cross-entropy is equivalent.
- Both of them are a measure of the difference of two distributions. If the two distributions are same, then the KL divergence is zero.
- There is a constant between them. However, this constant will not contribute to the gradient used for updating parameters.