Lesson 17

Statistics

So far, we have seen that we can calculate probabilities using densities of r.v.'s.

However, in real life, we usually have a set of data, whose density is unknown. In order to calculate probabilities, We have to know the density

off associated with the data.

Let's start from the simplest case, where we know that the data $X_1, X_2, \cdots, X_n$ is i.i.d, and we know the

type of distribution of X_i (e.g. we know that it is Pois(λ)), but we don't know its mean.

We learned that

$$\bar{X} =$$

can be viewed as an "estimate" of the mean $E[X]$. In particular, the LLNs guarantees that for large n's.

When one wishes to calculate the mean of a distribution, $\mathbb{E}[X]$ one can generate a sample $X_1, \dots, X_n$ and take $\frac{\sum X_i}{n}$ as an estimate of the mean.

We also learned that $\bar{X}$ is distributed normally, according to the CLT.

We are interested in constructing the so-called Confidence Intervals (C.I.'s) for $\bar{X}$, the estimate of $\mu = \mathbb{E}[X]$.

Roughly speaking, this is an interval that contains $\mu = \mathbb{E}[X]$ with a high Probability

More precisely, we fix a level of confidence, $1-\alpha$, where α, the level of doubt, is typically a small number.

We replace the estimate $\bar{X}$ with two estimators L, U, s.t.

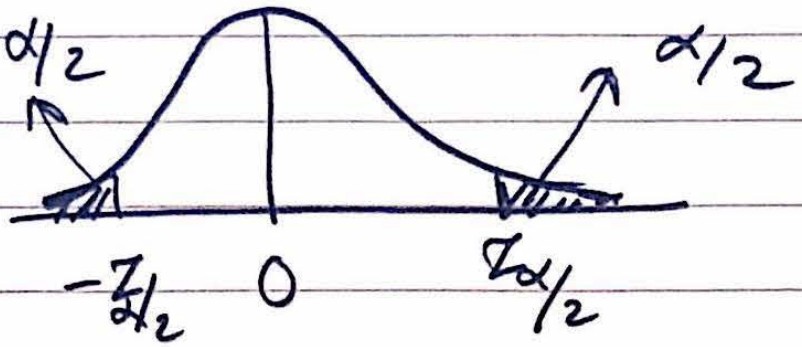
$$\mathbb{P}(L \leq \mu \leq U) \geq 1-\alpha$$

Remark: Since we model data $X_1, \ldots, X_n$ as $\overset{i.i.d}{\text{r.v.'s}}$, and estimates of parameters of F_X are regarded as functions of $X_1, \ldots, X_n$, estimates are also r.v.'s themselves. Therefore, $\bar{X}$, L, and U are r.v.'s whose distributions are dependent on μ. $[L, U]$ is called

the $(1-\alpha)$ confidence interval.

Because $\bar{X}$ is approximately distributed according to a Gaussian, and because we want a symmetric interval around $\bar{X}$, we follow the following method:

Observe that for a standard normal Z:



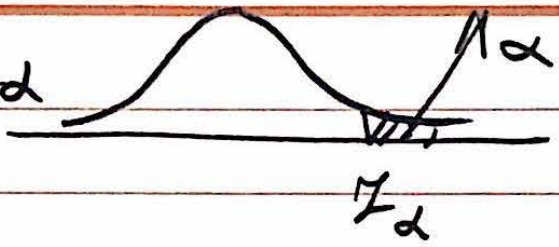
$$1-\alpha = P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right)$$

$$= P$$

Exercise: Build a "one-sided" $1-\alpha$ Confidence Interval $[L, +\infty)$ or $(-\infty, U]$, using the same methodology.

Hint: Use the following

$$\mathbb{P}(Z \leq z_\alpha) = 1-\alpha$$



$$\mathbb{P}(Z \geq -z_\alpha) = 1-\alpha$$

Interpretation of C.I's is subtle. We are tempted to interprete the Concept of a C.I by "the true parameter lies in the C.I with probability $1-\alpha$ (e.g. 95%)".

Here, μ is deterministic and L, U are r.v.'s, so based on different experiments

L, U may have different realizations. In other words, randomness is not in μ, but is in L, U!

The precise meaning of a C.I. is the following:

Confidence Interval for The mean Using Sample Mean & Sample Standard Deviation

We assumed to have σ_X, which is NOT ALWAYS a realistic assumption.

If X_i's are i.i.d $N(\mu, \sigma^2)$, even if we don't have a large sample, the quantity

$T =$

is distributed according to a Student t distribution, with $n-1$ degrees of freedom

The pdf of T is

$$f_T(x) = \frac{\left(1+\frac{x^2}{n-1}\right)^{-n/2}}{\sqrt{n-1}\ B\left(\frac{1}{2}, \frac{n-1}{2}\right)}$$

$$-\infty < x < \infty$$

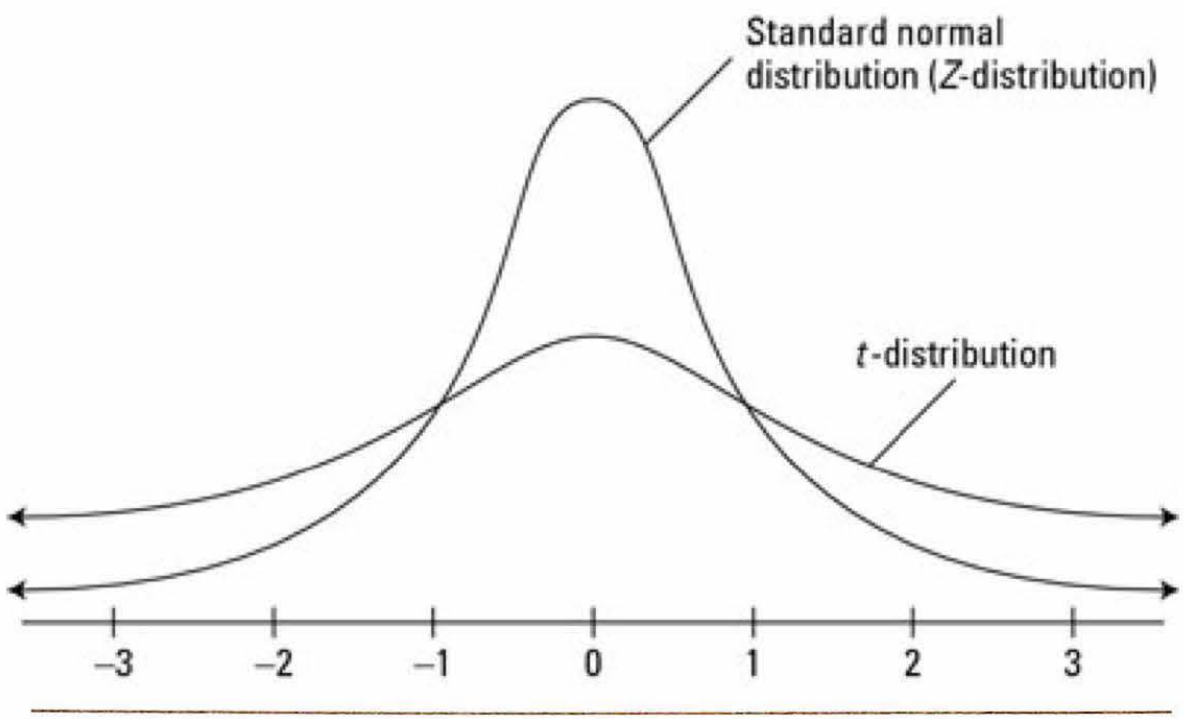
The $1-\alpha$ C.I. is

where

S_X: Sample Standard Deviation

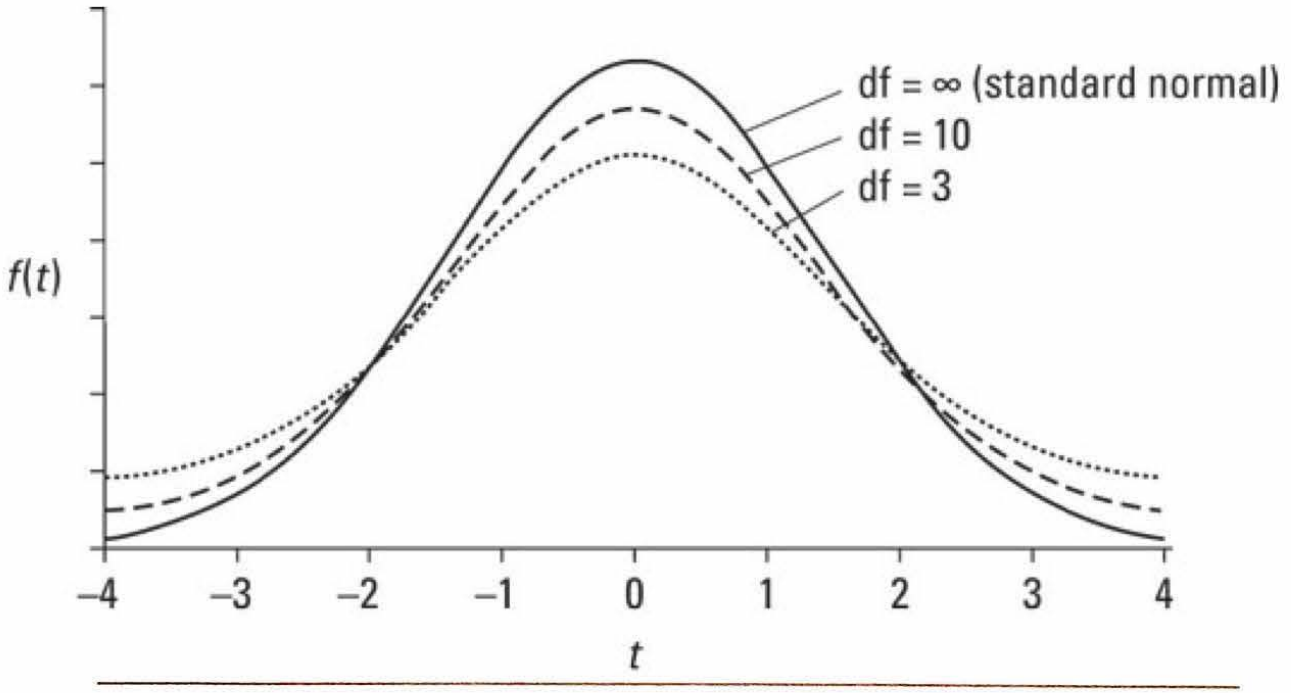
$$S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

The t-distribution looks like a normal distribution $N(0,1)$ i.e. it is bell-shaped & symmetric but has heavier tails.



t-distributions are members of a family of distributions parameterized by their degrees of freedom.

$t_{n-1,\alpha/2}$ is a cut off point that signifies a tail probability of $\alpha/2$. We usually use tables to determine $t_{n-1,\alpha/2}$



Example: Using the CLT to estimate proportions / Pollster

Confidence Intervals

Assume that we are interested to estimate the proportion of a population that has some characteristic, e.g.,

- proportion of USC students who study engineering
- proportion of Californians who support measure X

This is an attempt to estimate p for $X_i \sim Ber(p)$ ($i = 1, 2, \dots, N \rightarrow$ population size) by using a sample mean

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

where

n: sample size

$\sum_{i=1}^{n} X_i$: number of samples that have the desired characteristic

For large n $\hat{p} \sim N(\mu_X, \frac{\sigma_X^2}{n})$

where $\mu_X =$

$\sigma_X^2 =$

The $(1-\alpha)$ C.I for $\hat{p}$ is

The "Margin of Error" is defined as:

For $\alpha = 0.05$, determine the sample size n such that the Margin of Error is less than or equal to 3%

Confidence Intervals for Frequentist Hypothesis Testing

C.I.'s provide a basis for frequentist test of hypotheses. The result of a hypothesis testing procedure depends on whether or not a test statistic falls inside a $(1-\alpha)$ C.I. A hypothesis is usually a claim about a parameter of

a population (e.g., mean, variance, median, etc) to which we don't have access, using a statistic (a function of n i.i.d. samples from that population) with some level of confidence.

The null hypothesis is a statement about a population parameter that is sought to be rejected by using

the statistic as evidence.
If the statistic is too unlikely to have come from a population with hypothesized parameter, it is considered as evidence that rejects the null hypothesis.

Procedure for testing two-sided hypotheses about the mean of a population:

(i) Formulate competing hypotheses about μ_X

$$\begin{cases} H_0 : \mu_X = \mu_0 & \text{(Null hypothesis)} \\ H_1 : \mu_X \neq \mu_0 & \text{(Alternative hypothesis)} \end{cases}$$

(ii) Specify the two-sided $(1-\alpha)$ C.I for $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

which is

$$[-z_{\alpha/2}, z_{\alpha/2}]$$

(iii) obtain the test statistic $\bar{x}$ from the samples

(iv) Reject H_0 (= support H_1) if

$$\frac{\bar{x}-\mu_0}{\sigma_x/\sqrt{n}} \notin [-z_{\alpha/2}, z_{\alpha/2}]$$

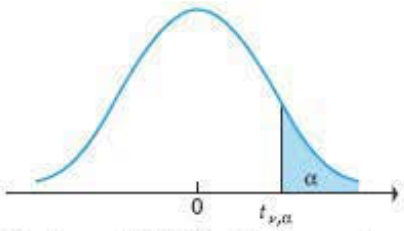
Fail to reject H_0

if $$\frac{\bar{x}-\mu_0}{\sigma_x/\sqrt{n}} \in [-z_{\alpha/2}, z_{\alpha/2}]$$

Exercise: Based on the one-sided C.I.'s you found in the previous exercise, Construct a procedure for testing the following one-sided hypotheses:

$$\begin{cases} H_0: \mu \leq \mu_0 \\ H_1: \mu > \mu_0 \end{cases} \quad \text{or} \quad \begin{cases} H_0: \mu \geq \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$$

Exercise: The average cost of a hotel room in Chicago is said to be $168 per night. A random sample of 25 hotels resulted in $\bar{x} = \$172.5$ and $s = \$15.4$. Test at $\alpha = 0.05$.



For selected probabilities, α, the table shows the values $t_{\nu,\alpha}$ such that $P(t_\nu > t_{\nu,\alpha}) = \alpha$, where t_ν is a Student's t random variable with ν degress of freedom. For example, the probability is .10 that a Student's t random variable with 10 degrees of freedom exceeds 1.372.

	Probability of Exceeding the Critical Value					
ν	0.10	0.05	0.025	0.01	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.313
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.782
8	1.397	1.860	2.306	2.896	3.355	4.499
9	1.383	1.833	2.262	2.821	3.250	4.296
10	1.372	1.812	2.228	2.764	3.169	4.143
11	1.363	1.796	2.201	2.718	3.106	4.024
12	1.356	1.782	2.179	2.681	3.055	3.929
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
60	1.296	1.671	2.000	2.390	2.660	3.232
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.090
ν	0.10	0.05	0.025	0.01	0.005	0.001

Upper Critical Values of Student's t Distribution with ν Degrees of Freedom

Parameter Estimation

Assume that $X_1, \dots, X_n$ are i.i.d. r.v.'s and $X \sim f(x|\theta)$ where θ is an unknown parameter (or parameter vector). The conditional notation here does not necessarily mean that θ is a r.v. It means that the distribution (pdf/pmf) depends on θ.

We wish to use $X_1, X_2, \dots, X_n$ to estimate θ.

More precisely, we want to find a "statistic" T

$$T = T(X_1, X_2, \dots, X_n)$$

which is a deterministic function of $X_1, \dots, X_n$ such that $\theta \approx T$.

Usually, we denote this T as $\hat{\theta}_n$.

We have two goals:

(i) Find a "good" $\hat{\theta}$

(ii) Understand what "good" means

There are three measures for determining what "good" means:

(a) Bias

$\text{Bias}(\hat{\theta}_n, \theta) =$

If $Bias(\hat{\theta}_n, \theta) = 0$, we say $\hat{\theta}_n$ is an unbiased estimate for θ.

If $\lim_{n \to \infty} Bias(\hat{\theta}_n, \theta) = 0$, then $\hat{\theta}_n$ is called "asymptotically unbiased"

(b) Mean-Squared Error

$MSE(\hat{\theta}_n, \theta) =$

Theorem (The Bias-Variance Decomposition): For a single estimate $\hat{\theta}_n$,

$$MSE(\hat{\theta}_n, \theta) = Var(\hat{\theta}_n) + Bias^2(\hat{\theta}_n, \theta)$$

Proof)

Remark: For the same mean-squared error, there is a compromise between the variance of the estimator and its bias. One cannot decrease the

variance and the bias simultaneously. In Machine Learning, this is called the bias-variance trade-off.

(c) Consistency

$\hat{\theta}_n$ is called consistent iff

$$\hat{\theta}_n \xrightarrow{i.p.} \theta$$

$\Leftrightarrow$

(Population) Median:

For a random variable X, m is a median iff

$$\mathbb{P}(X \leq m) \leq 1/2 \text{ and } \mathbb{P}(X > m) > 1/2$$

Sample Median:

Assume that $X_1, X_2, \ldots, X_n$ are i.i.d r.v.'s. The sample median is $m_n = \text{median}(X_1, \ldots, X_n)$.

Exercise: Show that m_n is an unbiased estimate of $\mathbb{E}[X]$ if f_X is symmetric about $\mu_X = \mathbb{E}[X]$

The Bias-Variance Decomposition for Convergence of Random Variables (Estimates)

Convergence in Mean-Squared Sense is important in science and engineering.

For an estimator, one can prove consistency (convergence in probability) using the bias-variance decomposition.

$$MSE(\hat{\theta}_n, \theta) = Var(\hat{\theta}_n) + Bias^2(\hat{\theta}_n, \theta)$$

If

$$\lim_{n\to\infty} Bias(\hat{\theta}_n, \theta) = 0$$

(i.e. $\hat{\theta}_n$ is asymptotically unbiased)

and

$$\lim_{n\to\infty} Var(\hat{\theta}_n) = 0$$

then $\lim_{n\to\infty} MSE(\hat{\theta}_n, \theta)$

$$= \lim_{n\to\infty} E[(\hat{\theta}_n - \theta)^2] = 0$$

i.e. $\hat{\theta}_n \xrightarrow{m.s.} \theta$, which

implies $\hat{\theta}_n \xrightarrow{i.p.} \theta$

i.e., $\hat{\theta}_n$ is a <u>consistent</u> estimator of θ.

Mean-Squared LLN

(Theorem) Assume that $X_1, X_2, \cdots, X_n$ are i.i.d and $\mathbb{E}[X^2] < \infty$. Then

$$\overline{X}_n \xrightarrow{\text{m.s.}} \mathbb{E}[X]$$

Proof) The bias variance decomposition.

(Important) Exercise: Assume that i.i.d r.v.'s $X_1, X_2, \cdots, X_n$ are similarly distributed and $X_i \sim N(0, 1/i^2)$. Define $Y_n = \sqrt{n} X_n$. Does Y_n converge in probability to 0?

We use multiple methods to solve this problem:

Method of Moments Estimation (MME)

The basic idea of MME is to use the law of large numbers to estimate parameters of a distribution

$$\bar{X}_n \xrightarrow{a.s.} \mu$$

- Assume $\underline{\theta} = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$ as parameters
- Denote $\mu_j = \mathbb{E}[X^j | \underline{\theta}]$ $j = 1, \dots, k$

(Assuming that all moments exist)

- Obviously $\mu_j = \mu_j(\theta_1, \dots, \theta_k)$

- If the inverse functions exist

$$\theta_j = h_j(\mu_1, \dots, \mu_k)$$

Denote $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$

By SLNN $\hat{\mu}_j \xrightarrow{a.s.} \mu_j$

as $n \to \infty$

The MME of $\underline{\theta}$ is then:

$$\hat{\theta}_j = h_j(\hat{\mu}_1, \dots, \hat{\mu}_k)$$

Theorem: If h_j is continuous, then $\hat{\theta}_j \xrightarrow{a.s.} \theta_j$ as $n \to \infty$

Example: Assume that $X \sim U(0, \theta)$. Find the MME of θ.

Lemma: For i.i.d r.v.'s that have finite first and second moments estimated as:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

the following holds:

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{1}{n}\left(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) + \bar{X}_n^2$$

Proof:

Remark: We use the following notation

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

and

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

Therefore, the results of the previous lemma can be written as:

$$\hat{\mu}_2 = \frac{n-1}{n} S^2 + \bar{X}_n^2 = \hat{\sigma}^2 + \bar{X}_n^2$$

In fact S^2 and $\hat{\sigma}^2$ are estimates of $Var(X)$.

Exercise: Assume that X_i's are i.i.d r.v.'s with $E[X] = \mu$ and $Var(X) = \sigma^2$. Calculate $E[S^2]$ and $E[\hat{\sigma}^2]$. Which one is an unbiased

estimate of σ^2?

Example: Assume that $X \sim N(\mu, \sigma^2)$. Find the MME of μ and σ^2.

Exercise: Are S_X^2 and $\hat{\sigma}^2$ consistent estimators of σ^2?

Assume that $X_1, X_2, \ldots$ are i.i.d and $\sigma_X^2 < \infty$ and $\mathbb{E}[X^4] < \infty$

Maximum Likelihood Estimation (MLE)

Assume that $X_1, \dots, X_n$ are i.i.d r.v.'s with $X \sim f(x|\theta)$.

Then the joint pdf/pmf of

X_i's is

$$f_{X_1, X_2, \cdots X_n}(x_1, x_2, \cdots, x_n | \theta) =$$

We also call the joint distribution the "likelihood function"

$$L(\theta | \underline{x}) = L(\theta | x_1, \cdots, x_n)$$

observations

$$=$$

likelihood function evaluated at $x_1, \cdots, x_n$, the observations

Example:

Intuition: Toss a biased coin with $p = \mathbb{P}(\{H\})$. We know that $p \in \{.2, .5\}$. If the coin is tossed 10 times and 9 heads are observed, what is your estimate of p? $\hat{p} = .2$ or $\hat{p} = .5$? Why?

$L(p = .5 \mid \text{observation})$

$L(p=.2 \mid \text{observation})$

The above example presents a motivation for MLE.

Def) $\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta \mid \underline{x})$

That is $L(\hat{\theta}_{MLE} \mid \underline{x})$

$= \max_{\theta \in \Theta} L(\theta \mid \underline{x})$

In practice, it is usually more efficient to consider the "log-likelihood" function

$$l(\theta|\underline{x}) = \log_e L(\theta|\underline{x})$$

$$=$$

because $\log_e$ is an increasing function, we can also say

$$\hat{\theta}_{MLE} = \text{argmax } l(\theta|\underline{x})$$

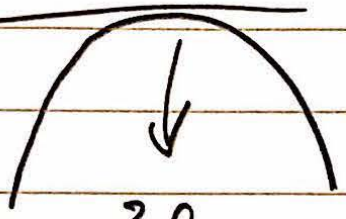
Remark: The log likelihood contains sum of individual likelihoods, and it is often easier to work with.

When l is differentiable with respect to θ we should have

$$\frac{\partial l}{\partial \theta}\Big|_{\theta=\hat{\theta}_{MLE}} =$$

And to make sure $\hat{\theta}_{MLE}$ maximizes the log likelihood function we should have

$$\left.\frac{\partial l}{\partial \theta}\right|_{\hat{\theta}_{MLE}}$$



$$\left.\frac{\partial^2 l}{\partial \theta^2}\right|_{\theta=\hat{\theta}_{MLE}}$$

Example: Assume that $X \sim Ber(p)$. Find $\hat{p}_{MLE}$

Example: Assume that $X \sim N(0, \sigma^2)$. Find the MLE of σ^2.

Example: Assume that $X \sim U(0, \theta)$. Find the Maximum Likelihood Estimate of θ

Properties of MLEs:

(1) Consistency

$$\hat{\theta}_{MLE} \xrightarrow{i.p.} \theta$$

2) Asymptotic Normality

$$\hat{\theta}_{MLE} \xrightarrow{d} N\left(\theta, \frac{1}{nI(\theta)}\right)$$

where $I(\theta)$ is called the Fisher Information for θ.

(3) $\widehat{g(\theta)}_{MLE} = g(\hat{\theta}_{MLE})$

In general, when $\underline{\theta}$ is a parameter vector eg. $\underline{\theta} = [\theta_1 \cdots \theta_k]^T$ the gradient vector should be zero at $\hat{\theta}_{MLE}$:

$$\nabla_{\underline{\theta}} l(\underline{\theta})\Big|_{\theta=\hat{\theta}_{MLE}} = 0$$

where $\nabla_{\underline{\theta}} l(\underline{\theta}) = \begin{bmatrix} \frac{\partial l}{\partial \theta_1} \\ \vdots \\ \frac{\partial l}{\partial \theta_k} \end{bmatrix} = \underline{0}$

Also, the second order derivative conditions are stated in terms of the Hessian Matrix:

$$\nabla^2_{\underline{\theta}} l(\underline{\theta}) = [h_{ij}]$$

where $h_{ij} = \dfrac{\partial^2 l(\underline{\theta})}{\partial \theta_i \partial \theta_j}$

For $\hat{\underline{\theta}}_{MLE}$ to maximize $l(\underline{\theta})$, the Hessian Matrix has to be ~~positive~~ negative definite, i.e.

$$\nabla^2_{\underline{\theta}} l(\underline{\theta}) \Big|_{\underline{\theta} = \hat{\underline{\theta}}_{MLE}} < 0$$

Recall that a ~~matrix~~ symmetric matrix

is negative definite if all of its eigenvalues are strictly negative

Definition (Fisher Information)

If $X \sim f(x|\theta)$, and $X_1, \ldots, X_n$ are i.i.d r.v.'s whose distribution is $f(x|\theta)$, the Fisher Information for θ is defined as

$$I(\theta) = \mathbb{E}_{X|\theta}\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right]$$

$$= -\mathbb{E}_{X|\theta}\left[\frac{\partial^2}{\partial \theta^2} \log f(X|\theta)\right]$$

Example: Assume that $X \sim \text{Ber}(p)$. Calculate the Fisher Information $I(p)$.

Example: We saw that the MLE is asymptotically normal, i.e.

$$\hat{\theta}_n \overset{d}{\approx} \mathcal{N}\left(\theta, \frac{1}{nI(\theta)}\right).$$

Using this result, find the asymptotic distribution of the MLE of p for $\text{Ber}(p)$ and compare the results to those of the CLT.

Question : If $\hat{\theta}_n$ is an unbiased estimator of θ, is it possible to improve the precision ($=$ variance) of $\hat{\theta}$ with adding more samples?

Theorem: (The Cramér-Rao Bound) Assume that $\hat{\theta}_n$ is an unbiased estimator of θ. Then

$$Var(\hat{\theta}_n) \geq \frac{1}{nI(\theta)}$$

Remark: The Cramér-Rao bound is a lower bound for the variance of any unbiased estimater $\hat{\theta}_n$.

Maximum A-Posteriori Estimate (MAP)

This estimator makes the "Bayesian Assumption" that the unknown parameter

θ is also a r.v. with pdf $h(\theta)$, which is called the "prior".

Consequently, the best estimate for θ is the one that maximizes the "posterior pdf" $f(\theta|\underline{x})$, i.e.

$$\hat{\theta}_{MAP} = \text{argmax}\left(f(\theta|\underline{x})\right)$$

However

$$g(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\, h(\theta)}{\int f(\underline{x}|\theta')\, h(\theta')\, d\theta'}$$

the denominator is a constant,

therefor $g(\theta|\underline{x}) \propto f(\underline{x}|\theta)\, h(\theta)$

and

$$\hat{\theta}_{MAP} = \text{argmax}$$

Note that

(i) $\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$ when

$h(\theta) =$ constant $\forall \theta$

which means MLE is the same as MAP when the prior pdf of θ is uniform.

(ii) The prior pdf encodes subjective knowledge about θ.

(iii) MLE and MAP heavily rely on the parametric form of $f(x|\theta)$.

In practice, $f(\underline{x}|\theta)$ may be difficult to find or too complicated to optimize. This often happens when X is a complicated/corrupted

function/version of a simple r.v. Z, i.e. $X = T(Z)$.

To calculate MLEs in such situations, the celebrated Expectation

Maximization (EM) Algorithm is used.

Example: Assum that $X_1, \dots, X_n$ are i.i.d $N(\mu, 1)$ RVs and $\mu \sim N(\mu_0, 1)$, where μ_0 is a constant. What is the MAP estimate of μ?

Exercise: Show that the MLE of μ when $X \sim \mathcal{N}(\mu, I)$ is

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Methods that assume a particular form for the distribution of a r.v. and try to estimate the parameters of that particular distribution from a "random sample," i.e. i.i.d. r.v.'s $X_1, \ldots, X_n$ can be called parametric methods.

Here, we focus on the so-called non-parametric methods, that are generic and do not assume parametric forms for probability distributions.

Histograms

Given i.i.d $X_1, \ldots, X_n$, assume that we select m intervals called bins, $[e_j, e_{j+1})$ where

$$e_1 < \ldots < e_{m+1} \quad \text{(edge sequence)}$$

where $e_1 \leq \min_i X_i$

$$e_{m+1} \geq \max_i X_i$$

When $e_{m+1} = \max_i X_i$, we use the bin $[e_m, e_{m+1}]$ instead of $[e_m, e_{m+1})$ so that no data is lost.

The histogram count for the j^{th} bin $[e_j, e_{j+1})$ is

$$H_j = \sum_{i=1}^{n} I_{[e_j, e_{j+1})}(X_i)$$

which is the number of samples that reside in $[e_j, e_{j+1})$.

Observe that $I_{[e_j, e_{j+1})}(X_i)$ is a Bernoulli r.v. The probability of success is calculated as

$$E\left[I_{[e_j, e_{j+1})}(X_i)\right] = P(e_j \leq X_i < e_{j+1})$$

Since X_i are i.i.d., $I_{[e_j, e_{j+1})}(X_i)$ are also i.i.d for each j.

The sample mean of $I_{[e_j, e_{j+1})}(X_i)$ is

$$\frac{1}{n} \sum_{i=1}^{n} I_{[e_j, e_{j+1})}(x_i) = H_j / n$$

For large n, $\frac{H_j}{n} \approx E[I_{[e_j, e_{j+1})}(X_i)]$ according to the LLN, so

$$P(e_j \leq X_i < e_{j+1}) \approx \frac{H_j}{n}$$

which is a straightforward way to estimate the distribution of X in $[e_j, e_{j+1})$ by a constant

If X_i's are integer valued, it is convenient to use $e_j = j - 1/2$ and $e_{j+1} = j + 1/2$, i.e. to center the bins on the integers,

Then $$\frac{H_j}{n} \approx \mathbb{P}(X_i = j)$$

If we are planning to draw a density function over the histogram;

$$\frac{H_j}{n} \approx \mathbb{P}(e_j \leq X_i \leq e_{j+1}) = \int_{e_j}^{e_{j+1}} f_X(x)\,dx$$

$$\approx f_X(c_j)\,\Delta x_j \quad \text{where } c_j = \frac{e_j + e_{j+1}}{2}$$

$$\Rightarrow f_X(c_j) = \frac{H_j}{n\,\Delta x_j}$$

Kernel Density Estimation

The empirical distribution places the probability mass $\frac{1}{n}$ to each $X_i = x_i$ that is observed.

In order to have a smooth estimate of the pdf of X, one can imagine distributing it in a neighborhood of $X_i = x_i$ according to some pdf. This approach is called Kernel Density Estimation (KDE).

KDE:

1. Choose a pdf K, the kernel. Typically, K is a symmetric pdf centered at origin. $N(0,1)$ and $U[-0.5, 0.5]$, Triangle $[-1, 1]$ are common choices.

2. At each x_i, center a copy of the kernel

$$\frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

This will control the distribution

of the $\frac{1}{n}$ probability assigned to x_i. h is called the smoothing parameter, the window width, or the bandwidth

3. The KDE estimate of f_X is

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

$$= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Remark: The R function density performs KDE.

Remark: The difficult choice in KDE is choosing h.

Empirical Distribution

Definition: Assume that $X_1, X_2, \cdots, X_n$ are i.i.d. samples with common CDF F_X. The empirical distribution function of $X_1, X_2, \cdots, X_n$ is defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i, +\infty)}(x)$$

where I_A is the indicator function of set A.

Informally, the empirical cdf

is increased by $\frac{1}{n}$ at each X_i.

Goodness-of-fit

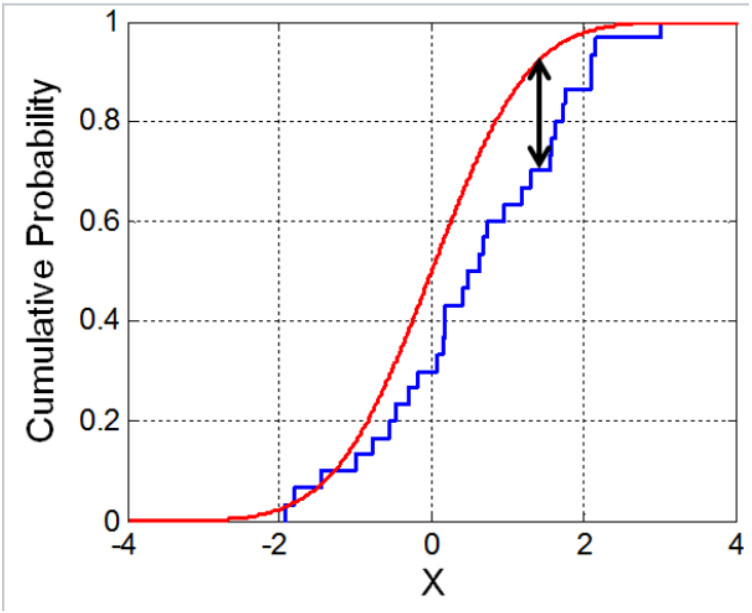
To measure how well a fitted distribution resembles the sample data, we can use the Kolmogorov-Smirnov (K-S) test statistic:

$$D_n = \sup_x |\hat{F}_n(x) - F(x)|$$

$\hat{F}_n(x)$: empirical distribution

$F(x)$: fitted distribution

$$D_n = \sup_x |F_n(x) - F(x)|$$



The Glivenko–Cantelli Theorem

Theorem: Assume that $X_1, X_2, \dots, X_n$ are i.i.d. samples with common cdf $F_X(x)$ and the empirical cdf $\hat{F}_n(x)$. Then the

Kolmogorov-Smirnov Statistic

$$D_n = \sup_x |\hat{F}_n(x) - F(x)|$$

converges to zero almost surely for all x, i.e.

$$\mathbb{P}\left(\lim_{n\to\infty} D_n \neq 0\right) = 0.$$

Remark: $\hat{F}_n(x)$ is a random variable at each x. D_n is also a r.v.

Remark: The G-K Theorem is sometimes called the fundamental theorem of statistics, and is a basis for the Kolmogorov-Smirnov test.

Assume that multiple distributions are estimated from i.i.d samples using different (parametric or non-parametric) methods. Which distribution is "the best?"

The Kolmogorov–Smirnov Test Criterion says that the KS statistic has to be computed for each distribution and the distribution with the minimum D_n has to be selected.

Goodness of fit as hypothesis Testing

H_0: All samples $X_1, \cdots, X_n$ come from the cdf F

H_1: At least one of the samples $X_1, \cdots, X_n$ <u>does not</u> come from the cdf F

Under the null hypothesis (assuming H_0 is true)

the K-S test statistic D_n follows a K-S distribution, which has a complicated form.

Therefore, to reject the null, we have to observe a sufficiently large D_n so that we are in the rejection region.

K-S distribution has a complicated form.

$$P\left(D_n < \frac{1}{2n} + v\right) = n! \int_{\frac{1}{2n}-v}^{\frac{1}{2n}+v} \int_{\frac{3}{2n}-v}^{\frac{3}{2n}+v} \dots \int_{\frac{2n-1}{2n}-v}^{\frac{2n-1}{2n}+v} g(u_1, u_2, \dots, u_n) du_n \dots du_2 du_1 \qquad 0 \le v \le \frac{2n-1}{2n}$$

$$g(u_1, u_2, \dots, u_n) = 1 \text{ over } 0 \le u_1 \le u_2 \le \dots \le u_n \le 1,$$

And 0 otherwise

Birnbaum, Z.W. (1952), "Numerical Tabulation of the Distribution of Kolomogorov's Statistic for Finite Sample Size," Journal of the American Statistical Association, 47, 425–441

$$F_{D_6}(t)=\begin{cases}0 & t<\frac{1}{12}\\ 46080t^6-23040t^5+4800t^4-\frac{1600}{3}t^3+\frac{100}{3}t^2-\frac{10}{9}t+\frac{5}{324} & \frac{1}{12}\le t<\frac{1}{6}\\ 2880t^6-4800t^5+2360t^4-\frac{1280}{3}t^3+\frac{235}{9}t^2+\frac{10}{27}t-\frac{5}{81} & \frac{1}{6}\le t<\frac{1}{4}\\ 320t^6+320t^5-\frac{2600}{3}t^4+\frac{4240}{9}t^3-\frac{785}{9}t^2+\frac{145}{27}t-\frac{35}{1296} & \frac{1}{4}\le t<\frac{1}{3}\\ -280t^6+560t^5-\frac{1115}{3}t^4+\frac{515}{9}t^3+\frac{1525}{54}t^2-\frac{565}{81}t+\frac{5}{16} & \frac{1}{3}\le t<\frac{5}{12}\\ 104t^6-240t^5+295t^4-\frac{1985}{9}t^3+\frac{775}{9}t^2-\frac{7645}{648}t+\frac{5}{16} & \frac{5}{12}\le t<\frac{1}{2}\\ -20t^6+32t^5-\frac{185}{9}t^3+\frac{175}{36}t^2+\frac{3371}{648}t-1 & \frac{1}{2}\le t<\frac{2}{3}\\ 10t^6-38t^5+\frac{160}{3}t^4-\frac{265}{9}t^3-\frac{115}{108}t^2+\frac{4651}{648}t-1 & \frac{2}{3}\le t<\frac{5}{6}\\ -2t^6+12t^5-30t^4+40t^3-30t^2+12t-1 & \frac{5}{6}\le t<1\\ 1 & t\ge 1.\end{cases}$$

Fortunately, there are tables for the critical values of the distribution of K–S test.

n\α	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	$\frac{1.94947}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.22385}{\sqrt{n}}$	$\frac{1.13795}{\sqrt{n}}$	$\frac{1.07275}{\sqrt{n}}$

K-S is distribution free, i.e. it does not assume any particular distribution for the data

Drawback: Only fore one dimensional distributions

There are other goodness of fit tests including the Anderson-Darling test.

The Chi-Squared Goodness of Fit Test

This test is based on the histogram of data. The idea is that if the pmf or pdf estimated for a sample is too different from its histogram, it probably is not a good fit for the data.

In particular, if $\mathbb{P}(e_j \leq X_i \leq e_{j+1})$ calculated according to the estimated pmf/pdf is not close to the histogram estimate of $\mathbb{P}(e_j \leq X_i \leq e_{j+1})$, which is $\frac{H_j}{n}$, the estimated pdf/pmf is not a good fit.

It can be shown that the quantity

$$\chi^2 = \sum_{j=1}^{m} \left(\frac{|H_j - np_j|^2}{np_j} \right)$$

has a Chi-squared distributia

with $m-1$ degrees of freedom when n is large.

This is because

$$\frac{H_j - np_j}{\sqrt{np_j}}$$

has a standard normal distributia

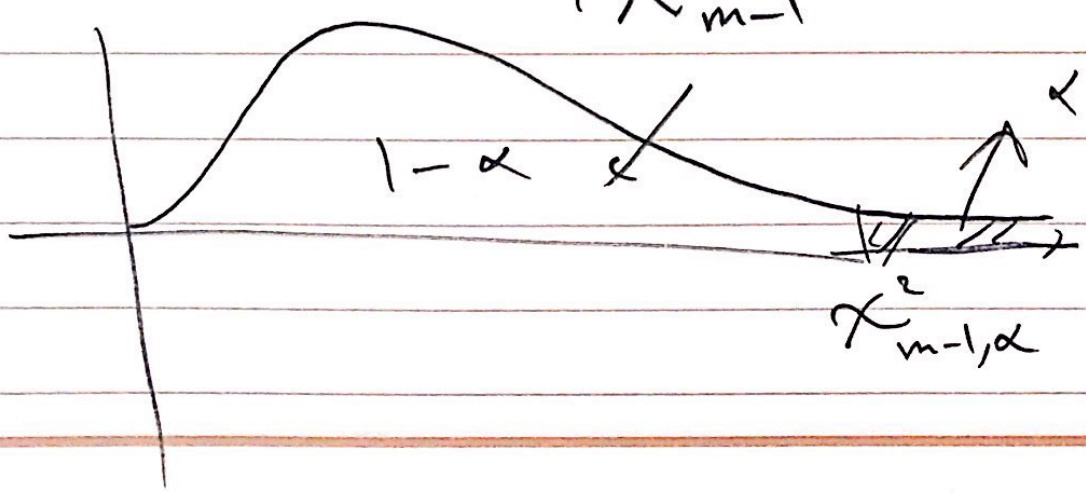
when n is large.

Therefore, if

$$\chi^2 > \chi^2_{m-1,\alpha}$$

(i.e. if the discrepancy is too large), we reject the null

that the pdf dpmt we found is a good fit for the data.



The Bootstrap

Assume that we have an i.i.d sample $X_1, X_2, \ldots, X_n$ and a statistic $T(X_1, X_2, \ldots, X_n)$. If we have access to the common distribution of X_i's, we can derive the distribution of T. However, it is not usually the case. If we have the luxury of having multiple samples from the common

distribution, we can still compute $T(\cdot)$ from each sample to have a sample of $T(\cdot)$'s and somehow estimate the distribution of $T(\cdot)$, e.g. by KDE.

Sample 1: $X_1^{(1)}$ $X_2^{(1)}$ ... $X_n^{(1)} \to T^{(1)}$

Sample 2: $X_1^{(2)}$ $X_2^{(2)}$ .... $X_n^{(2)} \to T^{(2)}$

⋮

Sample m: $X_1^{(m)}$ $X_2^{(m)}$ ... $X_n^{(m)} \to T^{(m)}$

Estimate F_T from $T^{(1)} \dots T^{(m)}$

Unfortunately, having multiple samples is not always possible.

Therefore, we use a "resampling" method called The Bootstrap.

The basic idea is to randomly draw datasets from the sample we have <u>with replacement.</u>

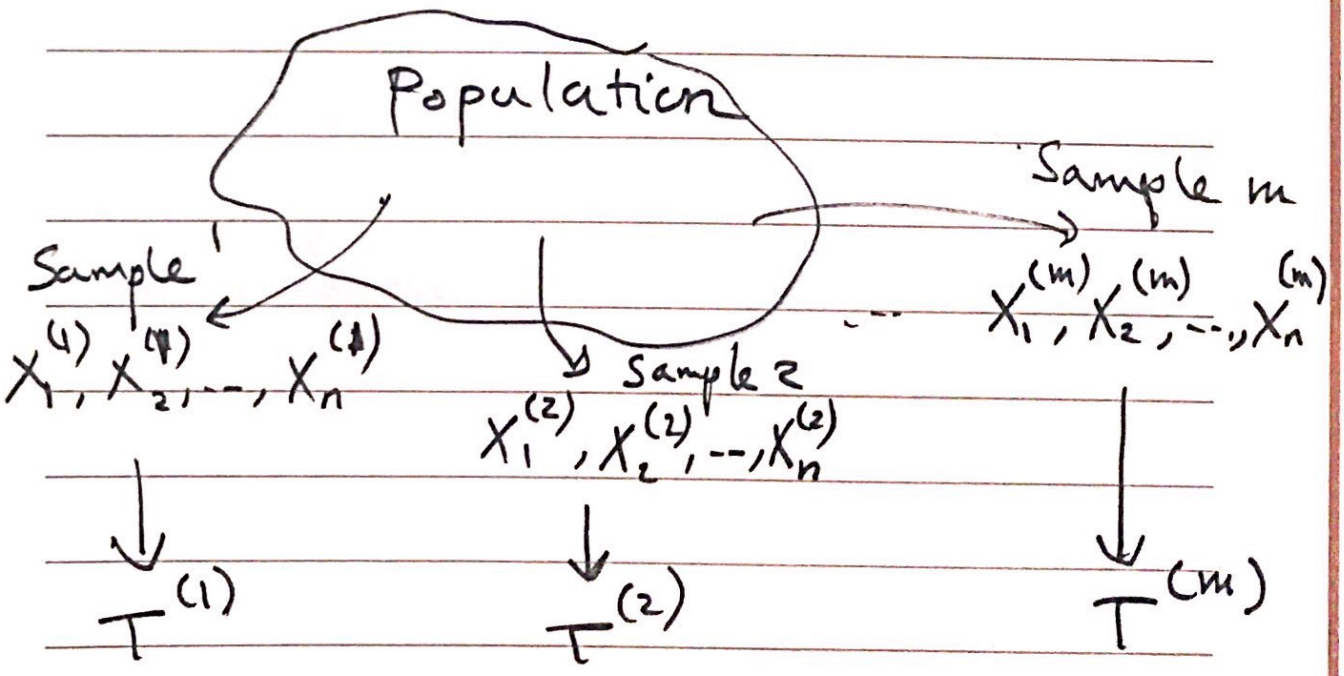
This is done B times, producing B bootstrap datasets. Then we calculate $T(\cdot)$ for each of

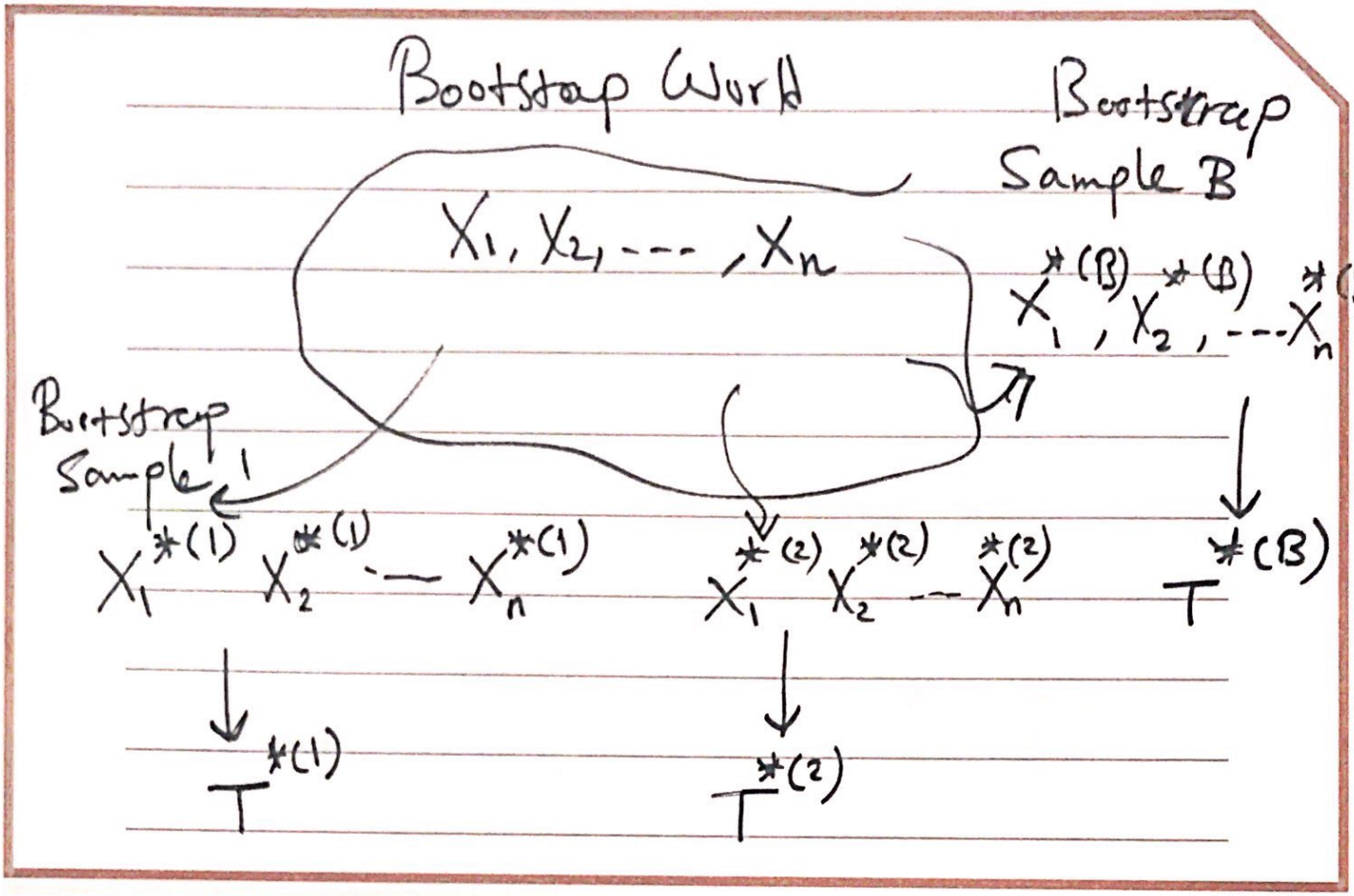
the B data sets and obtain a bootstrap sample of $T(\cdot)$. Then we can estimate the distribution or statistical properties of $T(\cdot)$ (such as mean, variance etc) from the bootstrap sample.

Remark: Sampling with replacement from the i.i.d sample $X_1, \cdots, X_n$ is equivalent to sampling from the empirical distribution function of $X_1, X_2, \cdots, X_n$.

Remark: One can view the original sample, from which bootstrap samples are drawn, the new population, and bootstrap samples as the samples drawn from that population

Real World:





In bootstrap world, the statistics of the original sample are considered as population parameters.

Example: Calculating the estimated variance of a statistic using bootstrap.

$$\widehat{\mathrm{Var}}\left(T(X_1, X_2, \ldots, X_n)\right) = \frac{1}{B-1} \sum_{b=1}^{B} \left(T(X_1^{*(b)}, X_2^{*(b)}, \ldots, X_n^{*(b)}) - \bar{T}^*\right)^2$$

where

$$\bar{T}^* = \frac{1}{B} \sum_{b=1}^{B} T(X_1^{*(b)}, X_2^{*(b)}, \ldots, X_n^{*(b)})$$

Bootstrap Percentile Confidence Intervals

Assume the sample $X_1, \dots, X_n$ was drawn from a population.

We wish to construct a $(1-\alpha)$

Confidence Interval for $T(X_1, \dots, X_n)$ (say $\bar{X}_n$ or S_n) using bootstrap:

1. Resample $R_1, R_2, \dots, R_B$ from $X_1, X_2, \dots, X_n$ with replacement

$$R_i = \left\{ X_1^{*(i)}, X_2^{*(i)}, \dots, X_n^{*(i)} \right\} \quad i = 1, 2, \dots, B$$

Note: $X_j^{*(i)}$ is may be repetetive

2. Calculate

$$T^{*(i)} = T(X_1^{*(i)}, X_2^{*(i)}, \cdots, X_n^{*(i)})$$

$i = 1, 2, \cdots, B$

for each bootstrap sample R_i.

3. Order $T^{*(i)}$'s and call the ordered version $U_{(1)}, U_{(2)}, \cdots, U_{(B)}$

$$U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(B)}$$

The $1-\alpha$ C.I is

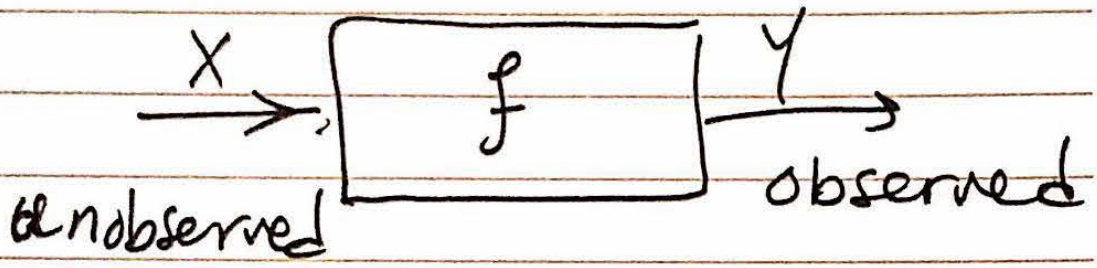
$$[U_{(a)}, U_{(b)}]$$

where $a = L \frac{\alpha}{2} B \rfloor$

and $b = [(1-\frac{\alpha}{2})B]$

$B = 1000$, $\alpha = 5\%$ $\rightarrow$ $a = 25$ $b = 975$

Minimum Mean-Squared Error Estimation (MMSE)



Find an estimate for X using Y, $\hat{X} = g(Y)$, such that $g(Y)$ minimizes the average mean-square estimation error $E[(X-g(Y))^2]$

You may recall that such an estimate is the conditional expectation

$$\hat{X} =$$

The estimation error is orthogonal to Y, i.e. no further information about Y is in the error.

Linear Estimation of Random Vectors

Question: Given two random vectors $\underline{X}, \underline{Y}$, where $\underline{X}$ is not observed but $\underline{Y}$ is observed,

what is the best estimate of $\underline{X}$ in terms of a linear (affine) function of $\underline{Y}$?

This needs clarification.

What does best mean?

The estimate with minimum mean-squared error,

$$MSE = \mathbb{E}\left[\|\underline{X} - A\underline{Y} - \underline{b}\|_2^2 \right]$$

Let $\mathbb{E}[\underline{X}] = \underline{m}_X$ and $\mathbb{E}[\underline{Y}] = \underline{m}_Y$

Then

$$MSE = \mathbb{E}\Big[\|\underbrace{(\underline{X} - \underline{m}_X) - A(\underline{Y} - \underline{m}_Y)}_{\underline{Z}} + \underbrace{\underline{m}_X - A\underline{m}_Y - \underline{b}}_{\underline{g}}\|^2 \Big]$$

$$MSE = \mathbb{E}\left[(\underline{Z} + \underline{g})^T (\underline{Z} + \underline{g}) \right]$$

$$= \mathbb{E}\left[\underline{Z}^T\underline{Z} + \underline{Z}^T\underline{g} + \underline{g}^T\underline{Z} + \underline{g}^T\underline{g} \right]$$

$$= E\left[\|(\underline{X} - m_X) - A(\underline{Y} - \underline{m}_Y)\|^2\right]$$

$$+ E\left[\|\underline{m}_X - A\underline{m}_Y - \underline{b}\|^2\right]$$

Obviously, to minimize MSE, $\underline{b}$ must be

Now, let's estimate A that minimizes

$$\mathbb{E}\left[\|(\underline{X}-\underline{m}_X)-A(\underline{Y}-\underline{m}_Y)\|^2\right]$$

$$=\mathbb{E}\left[\|\tilde{\underline{X}}-A\tilde{\underline{Y}}\|^2\right]$$

This looks like the MMSE problem, except $g(\tilde{\underline{Y}})$ is constrained to be $A\tilde{\underline{Y}}$.

We use the Principle of Orthogonality to solve this problem.

The error $\tilde{X} - A\tilde{Y}$ has to be orthogonal to all linear functions of $\tilde{Y}$, $g(\tilde{Y}) = B\tilde{Y}$:

$$\mathbb{E}[\qquad\qquad] =$$

To find the A that ~~maximize~~ minimizes the error, let's apply the function tr to both sides

$$tr\left(\mathbb{E}\left[(B\tilde{Y})^T(\tilde{X} - A\tilde{Y})\right]\right) = 0$$

$tr(.)$ commutes with $\mathbb{E}[.]$ (why?)

So,

$$\mathbb{E}\left[tr\left((B\tilde{Y})^T(\tilde{X}-A\tilde{Y})\right)\right]=0$$

On the other hand $tr(CD)=tr(DC)$

So

$$\mathbb{E}\left[tr\left((\tilde{X}-A\tilde{Y})(B\tilde{Y})^T\right)\right]=0$$

changing the order of $\mathbb{E}[.]$ and $tr(.)$ again

which means A solves

$$AC_Y = C_{XY}$$

and if C_Y is non-singular

$$A = C_{XY} C_Y^{-1}$$

Therefore the LMMSE of X in

terms of $\underline{Y}$ is

$$C_{XY} C_Y^{-1} (\underline{Y} - \underline{m}_Y) + \underline{m}_X$$

Note that when $\underline{X}$ and $\underline{Y}$ are jointly Gaussian, LMMSE estimate and MMSE coincide!

KNN Estimation of Conditional Expectation

We learned that $\mathbb{E}[Y|X]$ is the MMSE estimate of Y given X. How to estimate it from data?

Assume that we have n samples of Y, given $X=x$, and they are (conditionally) ~~independent~~ i.i.d. Then $\mathbb{E}[Y|X=x]$ can be

approximated using LLNs:

$$\mathbb{E}[Y|X=x] \approx \frac{1}{n}\sum_{i=1}^{n} Y_i|_{X=x}$$

However, in real situations, we may not have enough samples for estimating Y at $X=x$.

Therefore, we relax the above approximation to k other x's that are "close" to x:

$$\mathbb{E}[Y|X=x] \approx \frac{1}{n}\sum_{i=1}^{n} Y_i|_{X=x_1 \cup X=x_2 \cup \dots \cup X=x_k}$$

Graphically