Date: October 19, 2021

M T W T F S S
Page No.:
Date:
YOUVA

WEEK 7

Statistics from samples and Limit Theorems

STATISTICS FROM iid sAMPLES

- Where we have seen iid samples ?
  - → Bernoulli Trials
  - → Monte Carlo simulations
  - → Computing Histograms

BERNOULLI TRIALS

- Experiment and an event of interest A
  - → Occurence of an Event A is considered success
  - → What is $p = P(A)$?

- Bernoulli Trials
  - → Independent Repetitions or trials of experiment, say, n times
  - → $X_i = 1$ if A occurs in the i^{th} trial, and $X_i = 0$ otherwise

iid Bernoulli Samples : $X_1, X_2, \ldots, X_n$

GOAL : Try to estimate $P(A) = P(X_i = 1)$
→ Useful in finding

- Useful in finding prevalence of a disease in a population etc

MONTE CARLO SIMULATIONS

- Experiment and event of interest A
  → Too complex for modeling and computation
  → Can be simulated on a computer
- Repeat simulation n times independently
  → Record no. of occurrences of A as n_A, $P(A) \approx \frac{n_A}{n}$
  → Why is the above true? What should be n?

iid samples : $X_1, X_2, X_3, \ldots, X_n$

- $X_i = 1$ if A occurs in the i^{th} trial, and $X_i = 0$, otherwise
- Estimate $P(A) = P(X_i = 1)$

COMPUTING HISTOGRAMS

- n data points of some variable of interest
  → $x_1, x_2, x_3, \ldots, x_n$

- Bin : $[a, b]$
  → n_b : no. of x_i that fall inside $[a, b]$

MODEL

- X : continuous random variable with density $f_X(x)$
- Event $A = (a < X < b)$

HISTOGRAM COUNT

- Data points : iid samples $x_1, x_2, \ldots, x_n \sim f_X(x)$
- Estimate $P(a < X < b) \approx n_b / n$

---

iid Samples hold information on distribution

- What is common to all three of the previous scenarios?
  → Given : iid samples
  → Goal : get some partial information about distribution
    * Procedures to gather the info needed

- iid samples of an unknown or partially known distribution form the input for statistical procedures
  → Data : modelled as observations from iid repetitions of an experiment
  → Example : Iris Data
    * Data from every iris is considered to be iid observations from the distribution of the 4 Cogths

ANALYSIS :

- How to decide if the statistical procedure is 'good'?
- How many samples are needed for a 'goodness' guarantee?

Example : 20 iid Bernoulli (p) samples with p unknown

- **Goal** : find p from iid samples

→ Sampling 1 : 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1
→ Sampling 2 : 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1
→ Sampling 3 : 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
and so on.

- **Important** :
  - → p is the same for all samplings
  - → However, samples do not remain the same
  - → Each sample in each sampling is an observation of a random variable

- **Requirement on statistical procedure**
  - → Inspite of variations in samples, provide p with some guarantee

---

What is a typical statistical problem ?

- Model for samples : $X_1, X_2, \ldots, X_n \sim$ iid X
- Given 'data' : $x_1, x_2, \ldots, x_n$ from one sampling instance
- Distribution of X is partially known or unknown
  - → What is partially known ? ~~or unknown~~ Known distribution but parameters unknown
  - → Example : Bernoulli (p) with p unknown, Normal (μ, σ^2) with μ and σ unknown

- **Goal**: Procedures to find information about the distribution of X.

- **What information?**
  - → What is the mean of X? What is the variance of X?
  - → What is $P(X > t)$? What is $P(a < X < b)$?
  - → What is the distribution of X? What is the size of I_X?

EMPIRICAL DISTRIBUTION AND DESC-RIPTIVE STATISTICS

EMPIRICAL DISTRIBUTION

Let $X_1, X_2, \ldots, X_n \sim X$ be iid samples. Let $\#(X_i = t)$ denote the number of times t occurs in the samples. The empirical distribution is the discrete distribution with PMF.

$$p(t) = \frac{\#(X_i = t)}{n}$$

EXAMPLE: $n = 20$ → Range $= \{0, 1\}$

- 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1
  - → $p(0) = 8/20$ → $p(1) = 12/20$

Range $= \{0, 1\}$ · 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1
- → $p(0) = 7/20$ → $p(1) = 13/20$

Range $= \{0, 1, 2, 3\}$ · 1, 2, 0, 3, 0, 0, 1, 2, 0, 1, 3, 2, 1, 1, 0, 3, 0, 2, 2, 1
- → $p(0) = \frac{6}{20}$ → $p(1) = \frac{6}{20}$ $p(2) = \frac{5}{20}$ $p(3) = \frac{3}{20}$

Observations about empirical distribution

- Is the empirical distribution random ?
  - → Yes, it depends on the actual sample instances
  - → t and p(t) may change from one sampling to another
  - → Example : 20 Bernoulli (p) samples
- Descriptive Statistics : properties of empirical distribution
  - → Mean of the distribution
  - → Variance of the distribution
  - → Probability of an event
- As no. of samples increases, the properties of empirical distribution should become close to that of original distribution.

SAMPLE MEAN

Let $X_1, X_2, \ldots, X_n$ be iid samples. The sample mean, denoted $\bar{X}$, is defined to be the random variable.

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- Given a sampling $x_1, \ldots, x_n$, the value taken by the sample mean $\bar{X}$ is $\bar{x} = (x_1 + \ldots + x_n)/n$. Often, $\bar{X}$ and $\bar{x}$ are both called sample mean.

Example : $n = 20$

- 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1
  - → Value taken by $\bar{X}$: $\frac{12}{20}$

- 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1
  → Values taken by $\bar{x}$ = 13/20

- 1, 2, 0, 3, 0, 0, 1, 2, 0, 1, 3, 2, 1, 1, 0, 3, 0, 2, 2, 1
  → Sample mean : 25/20

Illustration 1 : Bernoulli (0.5) samples

$X_1, X_2, \ldots X_n \sim iid \{0, 1\}$

- $n = 5$
  - → Samples : 0, 0, 1, 1, 1 ; Sample mean : 3/5
  - → Samples : 1, 1, 1, 0, 1 ; Sample mean : 4/5
  - → Samples : 0, 1, 1, 1, 0 ; Sample mean : 3/5

- $n = 20$
  - → Sampling 1 : 12/20
  - → Sampling 2 : 13/20
  - → Sampling 3 : 13/20

- $n = 200$
  - → Sampling 1 : 95/200
  - → Sampling 2 : 102/200
  - → Sampling 3 : 98/200

- $n = 1000$
  - → Sampling 1 : 495/1000
  - → Sampling 2 : 490/1000
  - → Sampling 3 : 504/1000

Illustration 2 : Normal (0,1) samples

$X_1, \ldots, X_n \sim$ iid Normal (0,1)

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

- $n = 5$
  - → Samples : 2.17, 0.10, -0.75, -1.05, -1.72 ; Sample Mean = -0.25
  - → Samples : -0.26, 0.12, -0.31, -0.07, 1.35 : Sample Mean = 0.17
  - → Samples : -0.20, 0.37, 1.00, -0.41, -0.21 ; Sample Mean = 0.11
- $n = 20$
  - → Sampling 1 : 0.08 → Sampling 2 : -0.24
- $n = 200$
  - → Sampling 1 : -0.01 → Sampling 2 : 0.11
- $n = 1000$
  - → Sampling 1 : 0.04 → Sampling 2 : -0.04

EXPECTED VALUE AND VARIANCE OF SAMPLE MEAN

Let $X_1, X_2, \ldots, X_n$ be iid samples whose distribution has a finite mean μ and variance σ^2. The sample mean $\bar{X} = (X_1 + X_2 + \ldots + X_n)/n$ has expected value and variance given by

$$E[\bar{X}] = \mu \quad , \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

- Expected value of sample mean equals the expected value or mean of the distribution.
  - → Mean of distribution : constant real number and not random
  - → Sample Mean : random variable with mean equal to distribution mean
- Variance of sample mean decreases with n
  - → As n increases...
    - * variance of sample mean tends to zero
    - * the spread of sample mean will decrease
    - * sample mean will take values close to the distribution mean.

SAMPLE VARIANCE

Let $X_1, X_2, \ldots X_n$ be iid samples. The sample variance, denoted S^2, is defined to be the random variable

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1}$$

Where $\bar{X}$ is the sample mean.

- Given a sampling $x_1, \ldots, x_n$, the value taken by the sample variance S^2 is $s^2 = \left((x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right)/(n-1)$
  Often, s^2 and s^2 are both called sample variance
- Why $n-1$ in the denominator instead of n?
  - → Some books use n (this causes confusion)

→ Expected value of sample variance is simple in this case

EXPECTED VALUE OF SAMPLE VARIANCE

Let $X_1, X_2, \ldots, X_n$ be iid samples whose distribution has a finite variance σ^2. The sample variance $S^2 = \frac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n}$ has expected value given by,

$$E[S^2] = \sigma^2$$

- Expected value of sample variance equals the variance of the distribution
  - → Variance of distribution : constant real no. & not random
  - → Sample variance : random variable with mean equal to distribution variance

- Values of sample variance, on average, give the variance of distribution
  - → Variance of sample variance will decrease with no. of samples (in most cases)
  - → As n increases, sample variance takes values close to distribution variance

Illustration

- Bernoulli (1/2) , mean = 0.5 , variance = 0.25

→ Sample variance values : $n = 20$

★ 0.26 , 0.26 , 0.26 , 0.25 , 0.26

→ Sample variance values : $n = 200$
  * 0.2500 , 0.2487 , 0.2496 , 0.2456 , 0.2476

→ Sample variance values : $n = 1000$
  * 1.0268 , 0.9535 , 0.9781 , 0.9766 , 0.9831

· Normal (0,1) , mean = 0 variance = 1

→ Sample variance values : $n = 20$
  * 0.89 , 0.57 , 1.19 , 1.01 , 1.41

→ Sample variance values : $n = 200$
  * 0.93 , 1.07 , 0.85 , 0.83 , 1.09

→ Sample variance values : $n = 1000$
  * 1.0268 , 0.9535 , 0.9781 , 0.9766 , 0.9831

SAMPLE PROPORTION

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} X$

· iid samples from the distribution of X

· Let A be an event defined using X
  → Example : $A = (X > t)$, $A = (a < X < b)$ etc.

Definition : The sample proportion of A, denoted $s(A)$, is defined as

$$s(A) = \frac{\#(X_i \text{ for which } A \text{ is true})}{n}$$

- Samples : 0, 1, 1, 1, 0
  → $S(X=1) = 3/5$

- Samples : -0.2, 1.1, 0.3, -1.2, 0.7
  → $S(X \leq 0) = 2/5$, $S(0 < X < 1) = 2/5$, $S(X > 1) = 1/5$

EXPECTED VALUE AND VARIANCE OF SAMPLE PROPOR.

Let $X_1, X_2, \ldots, X_n$ be iid samples from the distribution of X. Let A be an event defined using X and let $P(A)$ be the probability of A. The sample proportion of A, denoted $S(A)$, has expected value and variance given by

$$E[S(A)] = P(A) \quad , \quad Var(S(A)) = \frac{P(A)(1-P(A))}{n}$$

PROOF :
- Convert samples into Bernoulli $(P(A))$ samples $Y_1, \ldots, Y_n$
  → $Y_i = 1$ if A is true for X_i, and $Y_i = 0$ otherwise
- $S(A)$ is the sample mean of $Y_1, \ldots, Y_n$
- As n increases, values of $S(A)$ will be close to $P(A)$
  → Mean of $S(A)$ equals $P(A)$
  → Variance of $S(A)$ tends to 0

Illustration

$X_1, \ldots, X_n \sim$ Normal $(0, 1)$

- $P(X \leq -1) = 0.159$
  - → Sample proportion values : $n = 20$
    - * 0.15, 0.20, 0.15, 0.15, 0.15
  - → Sample proportion values : $n = 200$
    - * 0.170, 0.140, 0.150, 0.155, 0.165
  - → Sample proportion values : $n = 1000$
    - * 0.160, 0.180, 0.162, 0.135, 0.153
- $P(-1 < X < 1) = 0.683$
  - → Sample proportion values : $n = 20$
    - * 0.75, 0.70, 0.55, 0.45, 0.70
  - → Sample proportion values : $n = 200$
    - * 0.705, 0.690, 0.705, 0.670, 0.720
  - → Sample proportion values : $n = 1000$
    - * 0.678, 0.678, 0.686, 0.679, 0.681

Where have we seen iid samples?

- Bernoulli Trials : Sample mean tends to distribution mean
  - → Bernoulli(p) samples
  - → Distribution mean = p
  - → Sample mean = fraction of successes
- Monte Carlo Simulations
  - → Sample proportion tends to actual probability
- Computing Histograms
  - → Sample proportion tends to actual probability.

Date: October 20, 2021

M T W T F S S
Page No.:
Date:
YOUVA

ILLUSTRATIONS WITH DATA

Iris Data

- 3 classes of irises : 0, 1, 2
  - → 50 instances of data for each class
  - → Each instance : [sepal length, sepal width, petal length, petal width] (cm)

- Sepal length of Class 0
  - → Model : iid samples according to some unknown distri.
  - → Data : 5.1, 4.9, 4.7, ..., 5.3, 5
  - → Sample Mean : 5.006, Sample Variance : 0.1242 = 0.3524^2
  - → $S(\text{Sepal length} > 5) = 22/50$, $S(4.8 < \text{Sepal length} < 5.2) = 20/50$

- Petal Width of Class 3
  - → Model : iid samples according to some unknown distri.
  - → Data : 2.5, 1.9, 2.1, ..., 2.3, 1.8
  - → Sample Mean : 2.026, Sample Variance : 0.0754 = 0.2746^2
  - → $S(\text{Petal Width} > 2) = 23/50$, $S(1.8 < \text{Petal length} < 2.2) = 17/50$

- Model : how good is the iid samples model ?

SUM OF INDEPENDENT R.Vs

Expected Value and Variance

Let $X_1, X_2, \ldots, X_n$ be random variables. Let $S = X_1 + \ldots + X_n$ be their sum. Then,

$$E[S] = E[X_1] + \ldots + E[X_n]$$

If $X_1, \ldots, X_n$ are pairwise uncorrelated, then

$$Var(S) = Var(X_1) + \ldots + Var(X_n)$$

- What is pairwise uncorrelated?
  $E[X_i X_j] = E[X_i]E[X_j]$ for all i, j, $i \neq j$
- Mean of sum is sum of means
- If uncorrelated, variance of sum is sum of variances
- If the X_i are independent, they are also uncorrelated
  → So, above result holds for independent R.Vs.

EXTENSIONS OF PREVIOUS RESULT

- Scaling and summing
  → Suppose $S = a_1 X_1 + \ldots + a_n X_n$ where a_i are constants
  → $E[S] = a_1 E[X_1] + \ldots + a_n E[X_n]$
  → $Var(S) = a_1^2 Var(X_1) + \ldots + a_n^2 Var(X_n)$, if uncorrelated
- iid samples : $X_1, X_2, \ldots, X_n \sim iid\ X$
  → Suppose $S = a_1 X_1 + \ldots + a_n X_n$, where a_i are constants

→ $E[s] = (a_1 + \dots + a_n)\, E[x]$
→ $Var(s) = (a_1^2 + \dots + a_n^2)\, Var(x)$

- Sample Mean : $X_1, \dots, X_n \sim X$, iid
  → $\bar{x} = (X_1 + \dots + X_n)/n$, $a_i = 1/n$
  → $E[(\bar{x})] = E[x]$
  → $Var((\bar{x})) = Var[x]/n$

Sample Mean versus distribution mean

$X_1, \dots, X_n \sim$ iid X

- Let $\mu = E[x]$, $\sigma^2 = Var(x)$
- Sample Mean : $\bar{x} = (X_1 + \dots + X_n)/n$
  → Expected value : μ , Variance : σ^2/n
  → Variance (or spread) goes to 0 as n grows

Can we say something more precise about $\bar{x}$ and μ?

- What is $P(\bar{x} > \mu + \delta)$?
- What is $P(\bar{x} < \mu - \delta)$?
- What is $P(|\bar{x} - \mu| > \delta)$?

Weak Law of Large Numbers

$X_1, X_2, \dots, X_n \sim$ iid X

- Let $\mu = E[x]$, $\sigma^2 = Var(x)$

- Sample Mean : $\bar{X} = (X_1 + \dots + X_n)/n$
  - Expected value : μ , Variance : σ^2/n

THEOREM : $P(|\bar{X} - \mu| > \delta) \leq \frac{\sigma^2}{n\delta^2} \to 0$

- With probability more than $1 - \frac{\sigma^2}{n\delta^2}$, sample mean lies in $[\mu - \delta, \mu + \delta]$.
  - What is the meaning of this probability?
- Chebyshev is usually a very "weak" bound, and we will see sharper bounds soon.

EXAMPLES : n iid samples

- Bernoulli (p) samples
  - With probability more than $1 - \frac{p(1-p)}{n\delta^2}$, sample mean lies in $[p - \delta, p + \delta]$
- Uniform $\{-M, \dots, M\}$ samples
  - With probability more than $1 - \frac{M(M+1)}{3n\delta^2}$, sample mean lies in $[-\delta, \delta]$
- Normal $(0, \sigma^2)$ samples
  - With probability more than $1 - \frac{\sigma^2}{n\delta^2}$, sample mean lies in $[-\delta, \delta]$
- Uniform $[-A, A]$ samples
  - With probability more than $1 - \frac{A^2}{3n\delta^2}$, sample mean lies in $[-\delta, \delta]$

- When distribution is known, a precise statement is possible about 'confidence' of finding sample mean within a certain precise interval
  → Improvement in bound will improve precision

EXAMPLES : Iris, Taj Mahal and IPL

- Iris Data : Sepal Length
  → $n = 50$, Sample mean : 5.006, Sample variance : 0.1242
  → With probability more than $1 - \frac{\sigma^2}{50\delta^2}$, sample mean lies in $[\mu - \delta, \mu + \delta]$
    * Works for $\delta > \sigma/\sqrt{50}$

- Taj Mahal air quality : PM2.5
  → $n = 11$, Sample mean = 65.72, Sample variance = 15.9^2
  → With probability more than $1 - \frac{\sigma^2}{11\delta^2}$, sample mean lies in $[\mu - \delta, \mu + \delta]$
    * Works for $\delta > \sigma/\sqrt{11}$

- IPL : Runs scored in Delivery 0.3
  → $n = 1598$, Sample mean : 0.9524, Sample variance : 2.0666
  → With probability more than $1 - \frac{\sigma^2}{1598\delta^2}$, sample mean lies in $[\mu - \delta, \mu + \delta]$
    * Works for $\delta > \sigma/\sqrt{1598}$

- What to do when distribution is unknown? Have to assume something.