

BSCMA1004

STATISTICS II NOTES



WEEK 0 NOTES

IITM B.S Degree

PREPARED BY

Vehaan Handa, IIT Madras

Statistics II → Week 0 Notes

Theory: Consistent explanation for observed data patterns.)

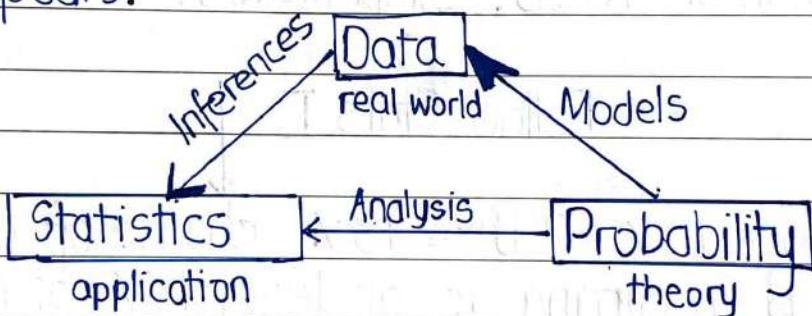
* Deterministic Patterns vs Random Patterns

Deterministic: current position = initial position + vel. × time.

But what about more complex situations? One in which there is no deterministic equation or processing ability? There are still many such useful observed patterns.

Non deterministic situations. Flipping a coin, pair of dice, predicting stock value.

When phenomena are unpredictably complex, we cannot expect exact models. Intuitively, we expect patterns when a random phenomenon repeats.



* Experiment and Outcome

An experiment is a process that we wish to study statistically.

Outcome is the result of the experiment in as much detail as necessary.

★ Sample Space (S)

A sample space is a set containing all outcomes of an experiment.

★ Event

Event is a subset of the sample space.
(but there is a restriction).

Die: $S = \{1, 2, 3, 4, 5, 6\}$ Subsets = $2^n = 2^6 = 64$

An event is said to have occurred if the actual outcome of experiment belongs to the event.

Events are sets. One event can be contained in another $A \subseteq B$, complement $A^c = (S \setminus A)$

Follow Stats I

If $P(A) = 10\%$, $P(B) = 13\%$. chance of either A or B occurring is at least 13%, at most 23%.

"Pick at random" \Rightarrow Uniform distribution

Probability that permutation of n objects is a derangement = $\frac{1}{e}$

If A_1, A_2 are disjoint, $P\left[\frac{A_1 \cup A_2}{B}\right] = \frac{P(A_1)}{B} + \frac{P(A_2)}{B}$

* Conditional probability examples

Consider 15 students \rightarrow 4 from S1, 8 from S2, 3 from S3. Probability that selected 3 are from S1, S3, S1 in that order?

$A_1 = 1^{\text{st}}$ student from S1

$A_2 = 2^{\text{nd}}$ student from S3

$A_3 = 3^{\text{rd}}$ student from S1

$$\frac{P(A_1 \cap (A_2 \cap A_3))}{P(B)} = P(B) \cdot P\left(\frac{A}{B}\right) = P(A_1) \cdot P\left(\frac{A_2 \cap A_3}{A_1}\right)$$

(4/15)

Rem. students given A_1 .

3 \rightarrow state 1

8 \rightarrow state 2

3 \rightarrow state 3

(3/14)

$$P(A_2/A_1) \cdot P(A_3/A_1 \cap A_2)$$

Conditioning

(3/13)

Nice example

Given $A_1 \cap A_2$

of conditioning

3 \rightarrow state 1

8 \rightarrow state 2

2 \rightarrow state 3

⚠ Family has 2 children. Probability that both are girls, given at least 1 is girl. ("No conditional here")

$$S = \{(G,G), (B,G), (G,B), (B,B)\}$$

$$B = \{(G,G), (B,G), (G,B)\}$$

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

$$P(\{(G,G)\} | B) = \frac{1/4}{3/4} = \frac{1}{3}$$

classmate

Disjoint non empty events are never independent.

* Single Bernoulli Trial:

Setting: Occurrence of event A is a success.
Non occurrence of A \rightarrow failure. Let $P(A) = p$.

Sample Space $S = \{\text{success, failure}\}$
 $P(\text{success}) = p$.

or $\{0,1\}$. $P(1) = p$ $P(0) = 1-p$.

Denoted Bernoulli(p)

Each trial is independent.

In general, $P(b_1 b_2 \dots b_n) = p^w (1-p)^{n-w}$ $w = \text{no of 1's}$.

Now for a biased coin. $P(H) = 1/3$ $P(T) = 2/3$
 $\therefore P(HHHHH) = (1/3)^5$ $P(TTTTT) = (2/3)^5$

so on....

Binomial (n, p):

$$\begin{aligned}
 P(B=3) &= P(\text{trials} \geq 3 \text{ 1's}) = (\text{no. of favourable outcomes}) \\
 &\quad \times \frac{p^3 (1-p)^2}{5!} \\
 &= {}^5 C_3 p^3 (1-p)^2 \\
 &= 10 p^3 (1-p)^2
 \end{aligned}$$

General: $P(B(n,p) = k)$? ($k = 0, 1, 2, \dots, n$)

$$P(B(n,p) = k) = {}^n C_k p^k (1-p)^{n-k}$$

The binomial distribution:

- Starts at $(1-p)^n \rightarrow$ increases, reaches a peak \rightarrow falls to p^n
- Peak is near np (exactly floor($p(n+1)$))
- $P(B=0 \text{ or } B=1 \text{ or } B=2 \text{ or } \dots \text{ or } B=n) = 1$
- $P(B=0) + P(B=1) + P(B=2) + \dots + P(B=n) = 1$
- $(1-p)^n + {}^nC_1 p(1-p)^{n-1} + {}^nC_2 p^2(1-p)^{n-2} + \dots + p^n = 1$

Suppose we toss a fair coin until we get a head.
How many times will we toss the coin?

$$P(1) = P(H_1) = 1/2$$

$$P(2) = P(T_1, H_2) = 1/2 \times 1/2 = 1/4$$

:

:

$$P(k) = (1/2)^k$$



Geometric Distribution

Perform independent Bernoulli(p) trials indefinitely.

$$S = \{1, 2, 3, 4, 5, 6, \dots\}$$

$$P(G_1=1) = p$$

$$P(G_1=2) = p(1-p)$$

$$P(G_1=3) = (1-p)^2 p$$

Geometric Distribution

- Starts at p , keeps falling
- Keeps on decreasing, but if $p < 1$, never goes fully to 0.

$$\begin{aligned}
 P(G \leq k) &= P(G=1 \text{ or } G=2 \text{ or } \dots \text{ or } G=k) \\
 &= P(G=1) + P(G=2) + \dots + P(G=k) \\
 &= p + (1-p)p + (1-p)^2 p + \dots + (1-p)^{k-1} p \\
 &= 1 - (1-p)^k
 \end{aligned}$$

It follows that $P(G > k) = (1-p)^k$

★ Probability Estimation by Monte Carlo Simulation

Probability of event A can be ~~simu~~ estimated as follows. We simulate experiment repeatedly and independently, say N times, count no. of times the event occurred. That is:→

$$P(A) = \frac{N_A}{N}$$

As $N \rightarrow \infty$, estimate becomes better and better.
This is called Monte Carlo simulation.

★ Birthday Problem:

In a group of n, what is the chance that two of them share the same birthday?

Event A = 2 have same birthday

A^c = No 2 have same birthday.

$A^c = (B_1 \text{ on any date } B_1) \text{ and } (B_2 \text{ on any date other than } B_1)$
 and $(B_3 \text{ on any date other than } B_1, B_2) \dots$

$$P(A^c) = 1 \cdot \left(1 - \frac{1}{365}\right) \left(1 - \frac{2}{365}\right) \dots \left(1 - \frac{n-1}{365}\right)$$

* Gambler's Ruin

Gambler starting with k units of money :→

- If he has ≥ 1 money units, coin toss. $H=+1$, $T=-1$
- If bankrupt, stops playing
- If win = N units, stops.

If $p = \text{prob. of heads}$ $q = 1-p$

$$\Pr(\text{bankruptcy}) = \begin{cases} 1 - \frac{k}{N} & \text{if } p=q=1/2 \\ \frac{(q/p)^k - (q/p)^N}{1 - (q/p)^N}, & \text{if } p \neq q \end{cases}$$

Condition on the first toss :→

$$X_k = P(\text{Bankruptcy} | \text{first toss is head}) p + P(\text{Bankruptcy} | \text{first tail}) q$$

$$= X_{k+1} p + X_{k-1} q$$

* Range of Random Variable

The set of values taken by it, which is a subset of \mathbb{R} .

First associate range with the RV.

$P(X=t) = P(\text{all outcomes resulting in } X=t)$
~ = distributed as.

DATE

--	--	--	--	--	--	--

* PMF: Discrete RV

PMF of drv X with range set T is function
 $f_X: T \rightarrow [0, 1]$ defined as $f_X(t) = P(X=t) \forall t \in T$.

If B is a subset of T , consider $X \in B$.

$$P(X \in B) = \sum_{t \in B} P(X=t) = \sum_{x \in B} f_X(t)$$

* Some distributions

1. Uniform Random Variable

$X \sim \text{Uniform}(T)$, where T is some finite set.

Range: Finite set T

$$\text{PMF: } f_X(t) = 1/|T| \text{ for all } t \in T$$

2. Bernoulli Random Variable

$X \sim \text{Bernoulli}(p)$ where $0 \leq p \leq 1$

Range: $\{0, 1\}$

$$\text{PMF: } f_X(0) = 1-p \quad f_X(1) = p$$

3. Binomial Random Variable

$X \sim \text{Binomial}(n, p)$ (n : +ve int, $0 < p \leq 1$)

Range: $\{0, 1, 2, \dots, n\}$

$$\text{PMF: } f_X(k) = {}^n C_k p^k (1-p)^{n-k}$$

4. Geometric Random Variable

$X \sim \text{Geometric}(p), 0 < p \leq 1$

Range: $\{1, 2, 3, \dots\}$

$$\text{PMF: } f_X(k) = (1-p)^{k-1} p$$

OR Range: $\{0, 1, 2, 3, \dots\}$

$$\text{PMF: } f_X(k) = (1-p)^k p$$

- No of trials for first success in repeated independent Bernoulli(p) trials.

5. Negative Binomial Random Variable

$X \sim \text{Negative Binomial}(r, p), r = \text{+ve int}, 0 < p \leq 1$

Range: $\{r, r+1, r+2, \dots\}$

$$\text{PMF: } f_X(k) = (k-1)C_{r-1} (1-p)^{k-r} p^r$$

- No of trials for r successes in repeated independent Bernoulli(p) trials.

6. Poisson Random Variable

$X \sim \text{Poisson}(\lambda), \lambda > 0, \lambda \in \mathbb{R}$

Range: $\{0, 1, 2, 3, \dots\}$

$$\text{PMF: } f_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

$$\text{f}_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

k	0	1	2	3	4	...
$f_X(k)$	$e^{-\lambda}$	$e^{-\lambda} \lambda$	$e^{-\lambda} \frac{\lambda^2}{2!}$	$e^{-\lambda} \frac{\lambda^3}{3!}$	$e^{-\lambda} \frac{\lambda^4}{4!}$	

$$\text{Note } e^\lambda = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots$$

7. Hypergeometric Random Variable

$X \sim \text{HyperGeo}(N, r, m)$ where $N, r, m \in \mathbb{R}^+$

- Consider population of 'N' persons
 - with 'r' of Type 1
 - and ' $N-r$ ' of Type 2
- Select 'm' persons uniformly at random randomly without replacement.
- $X = \text{no. of persons of Type 1 selected.}$

Range of X.

- $N = 100, r = 50, m = 20 \Rightarrow X \in 0, 1, 2, \dots, 20$
- $N = 100, r = 10, m = 20 \Rightarrow X \in 0, 1, 2, \dots, 10$
- $N = 100, r = 90, m = 20 \Rightarrow X \in 10, 11, 12, \dots, 20$

$$X \in \max(0, m-(N-r)), \dots, \min(r, m)$$

PMF

■ $f_X(k) = \frac{rC_k(N-r)C_{m-k}}{NC_m}$

i.e

$$\frac{rC_k(N-r)C_{m-k}}{NC_m}$$

★ Events over a period of time

Eg. arrival to a queue
 arrival of a visitor to website
 emission of particle by radioactive decay
 meteorite entering atmosphere.

Arrival rate can be assumed constant.
 Given one arrival, time for next arrival independent.

Under these assumptions, no. of arrivals in a fixed period of time becomes a Poisson Random Variable.

Eg. in 2608 time intervals of 7.5 seconds each, emission of alpha particles is as follows :-

Particles	0	1	2	3	4	5	6	7	8	9	10
Times	57	203	383	525	532	408	273	139	45	27	16
Fraction	0.022	0.078	0.147	0.201	0.204	0.156	0.105	0.053	0.017		0.0061

$$\text{Emission Rate: No. of particles (total)} = 3.8673 (\lambda_{\text{poisson}})$$

$$\frac{2608}{2608}$$

Time of next emission, independent of past.

$$P(\text{part} = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

It is a good Poisson Fit.

* Functions of a Random Variable

X : Random variable, range T , PMF $f_X(t)$

X : function from sample space to T .

$T \in \mathbb{R}$

Let $f(x)$ be a fn from \mathbb{R} to \mathbb{R}

Then $f(X)$ can be seen as a composition of 2 fns.
 $f(X)$ is also random variable in some probability space.

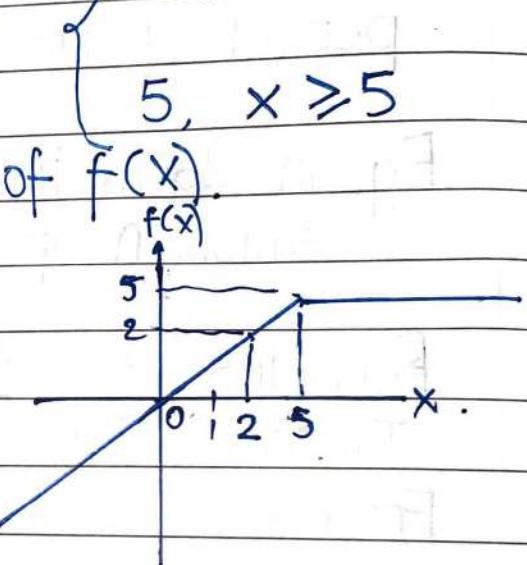
Let $X \sim \text{Geometric}(0.5)$, $f(x) = \begin{cases} x, & x < 5 \\ 5, & x \geq 5 \end{cases}$

Find range and distribution of $f(X)$.

$$f_X(k) = \left(\frac{1}{2}\right)^{k-1} \left(\frac{1}{2}\right) = \frac{1}{2^k}$$

$$k = 1, 2, 3, 4, 5, 6, 7, \dots$$

$$f(k) = 1, 2, 3, 4, 5, 5, 5, \dots$$



$$\text{Range of } f(x) = \{1, 2, 3, 4, 5\}$$

$$\downarrow \frac{1}{2} \downarrow \frac{1}{4} \downarrow \frac{1}{8} \downarrow \frac{1}{16} \downarrow \frac{1}{16}.$$

$$\Pr(f(X) = 1) = \Pr(X = 1) = \frac{1}{2}$$

$$\Pr(f(X) = 4) = \Pr(X = 4) = \frac{1}{2^4}$$

$$\begin{aligned} \Pr(f(X) = 5) &= \Pr(X \geq 5) = \frac{1}{2^5} + \frac{1}{2^6} + \frac{1}{2^7} \dots \\ &= \frac{\frac{1}{2^5}}{1 - \frac{1}{2}} = \boxed{\frac{1}{2^4}} \end{aligned}$$

X : Random Var. with PMF $f_X(t)$

$f(X)$ = random var whose PMF is \rightarrow

$$\begin{aligned}f_{f(X)}(a) &= P(f(X) = a) = P(X \in \{t : f(t) = a\}) \\&= \sum_{t : f(t) = a} f_X(t)\end{aligned}$$