SS

Wednesday, May 25, 2022 7:53 AM

Basics (pop, sample, perm & comb, variable types)

Sample space

sample event

sample point

Collecting data

- survey
  - targeted audience
- census
  - whole pop (pop of interest)
- cohort
  - share event
  - time span

perms & comps

And = *

OR = +

at most / at least } count all

Permutations

Order matters? — No → **Combinations**

$^nC_r = \frac{n!}{(n-r)!\,r!}$

rep allowed $\Rightarrow \binom{n+r-1}{r}$

yes ↓

repetition allowed $\Rightarrow n^r$

" not allowed $\Rightarrow \frac{n!}{(n-r)!}$

$\frac{n!}{r!}$ ← repeated chrs

circula

$(n-1)!$

bracelet ⇓ $\frac{(n-1)!}{2}$

---

data

- quantitative 123
  - ordinal: ranked/order ? ✓
  - nominal: order ✗
  - no arithmetic
  - ordinal → eg: rating, user satisfaction
  - nominal → eg: names, colors
- qualitative ♡
  - interval: true zero ✗, ratios ✗
  - ratio: true zero ✓, ratios ✓
  - arithmetic ✓
  - interval → eg: temp, time
  - ratio → eg: age, weight, money/salary

discrete

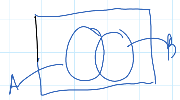
cts

look about the output

Independent

$P(A) = P(A|B) = P(B|A)$

$P(A \cap B) = P(A)\,P(B)$ check



$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

can use demorgan's law

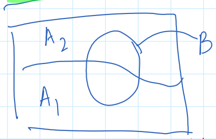
mutualy exclusive VS mutually enclusive

$P(A \cap B) = 0$

Baye's Thm

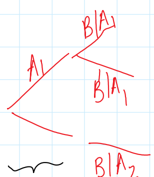
$$\boxed{P(A_k|B) = \frac{P(B|A_k)\,P(A_k)}{P(B)}}$$

$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \dots$



- $A_1, A_2, \dots$ must add up to 100%
- if $A_1, A_2 \dots$ given always $P(B|A_k)$ is given

  or $P(A_k \cap B)$ is given (rarely)
- tree method



each 2 add up to 100%

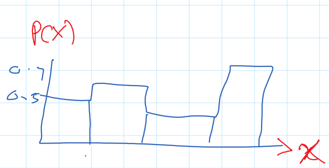
Prop of cts & discrete vars

vars

discrete

PMF each pts

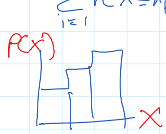
$P(X=x) \Rightarrow \sum_x P(X=x) = 1$



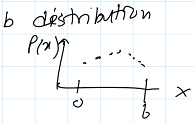
CDF (upto a point)

$P(X \le x_k)$

$\Rightarrow \sum_{i=1}^{x_k} P(X \le x_k)$



Prob distribution



cts

PDF (Prob density fun) whole region

$P(a \le X \le b) = \int_a^b f\,dx$

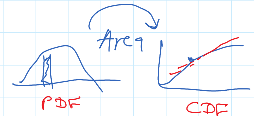
$$\boxed{\Rightarrow \underbrace{\int_{-\infty}^{\infty} f\,dx = 1}_{\text{to show it's a pdf}}}$$

CDF upto a pt

$F(x) = P(X \le x)$

$= \int_{-\infty}^{x} f\,dx$

Area



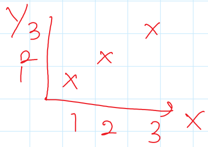
$$\boxed{F(x) = \int_{-\infty}^{x} f(x)} \Rightarrow f(x) = F'(x)$$

Subst $F(LL) = 0$ check

$-\infty$

subst $F(LL) = 0$ } check*
$F(UL) = 1$ }

① find what is r.v (X?)
② all outcomes $X = x$
$0, 1, 2$?
③ Sample space all posibilities $\Rightarrow S = \{(1,1), \dots\}$
④ Prob distribution



write out what's been asked
like $P(X \geq 3)$
$P(X \leq 3 \cup Y \leq 3)$
$\hookrightarrow 1 - P'$

$f(x)$ PMF — Joint — CDF $F(x)$

$p(x,y) = P(X = x, Y = y)$

$P(a \leq X, y \leq Y) = \sum_y \sum_x p(x,y)$

Ivar joint: $P(X \cap Y) = P(X) P(Y)$

$\boxed{f(x,y) = g(x)\, g(y)}$

joint pdf — marginal

convert to $P(X > 1 \cap Y > 2) \Rightarrow \int_{-\infty} \int_{-\infty}^{2} f(x,y)$ ✓

$f(x)$ PDF — Joint — $F(x)$ CDF

$P[(X,Y) \in A] = \underbrace{\iint_A f(x,y)\,dx\,dy}_{vol}$

$F(a,b) = P(X \leq a, Y \leq b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)\, dx\,dy$

valid joint pdf

$\boxed{\iint f(x,y)\,dy\,dx = 1}$

marginal

$P(x)$ marginal PDF of X $= \int_{-\infty}^{\infty} f(x,y)\, dy = g(x)$

$P(Y)$ " $Y = \int_{-\infty}^{\infty} f(x,y)\, dx = g(y)$

$\boxed{\text{conditional prob} = \frac{\text{joint prob}}{\text{marginal prob}}}$ $\equiv P(B|A) = \frac{P(B \cap A)}{P(A)}$

$E(f(x)g(y)) = \iint f(x)\, g(x)\, dy\, dx$

---

mode {most occuring} ocurs
median → ocurs, (Sorted)

discrete prob dist

mean

$\boxed{E(X) = \sum_{i=1}^{n} x P(x)} = \mu$

x: 3 4 5
$P(A)$: 1 2 3

expected val — observation — prob

population

Variance

$\boxed{V(X) = E(x^2) - [E(x)]^2} = \underbrace{E[(X - \mu)^2]}_{\text{expected val of squared diff btw } X \,\&\, \mu}$

Sample (irrespective of size)

$\bar{x}$, $S_x^2 \longrightarrow \boxed{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

BLUE of μ

mean

$\boxed{E(X) = \int_{-\infty}^{\infty} x\, p(x)\, dx}$

$E(x^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx$

Variance

$V(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$

$\sigma = \sqrt{V(x)}$

Standard deviation

$\boxed{\sigma = \sqrt{E(x^2) - \mu^2}}$

* expected val → compare
risk → use σ SD *
$\uparrow$ risk , $\uparrow$ SD

* cmp → discrete using $\boxed{\frac{\sigma}{\mu} \times 100\%}$ risk

coefficient of variation

---

prop of $E(x)$

* $E(x+c) = E(x) + c$
* $E(cx) = c\,E(x)$
* $E[g(x)] = g[E(x)]$
  linear
* $E(X+Y) = E(X) + E(Y)$

$V(X)$

$V(X+c) = V(X)$
$V(cX) = c^2 V(X)$
$\boxed{V(X+Y) =} V(X) + V(Y) + 2\,\mathrm{COV}(X, Y)$
0 if independent

independence ⇒ $\mathrm{COV}(X,Y) = 0$
$\not\Leftarrow$

---

covarience (Joint) 2 rv

$\boxed{\mathrm{COV}(X,Y) = E(XY) - E(X)E(Y)}$
μ_x μ_y

discrete ⇒ $\sum x_i y_i P(x_i, y_i)$
cts ⇒ $\iint_A x_i y_i\, P(x,y)\,dx\,dy$

prop of cov

* $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
* $\mathrm{Cov}(X,a) = 0$
* $\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X,Y)$
* $\mathrm{Cov}(X+a, Y+b) = \mathrm{Cov}(X,Y)$
* $\mathrm{Cov}(aX+bY, cW+dZ) = ac\,\mathrm{Cov}(X,W) + ad\,\mathrm{Cov}(X,Z) + bc\,\mathrm{Cov}(Y,W) + bd\,\mathrm{Cov}(Y,Z)$

Correlation pop → $\rho_{x,y} = \frac{\mathrm{COV}(X,Y)}{\sqrt{V(X)V(Y)}}$

pt estimation

sample → $$r_{xy} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2}\,\sqrt{\sum_i y_i^2 - n\bar{y}^2}}.$$

Pearson's

relationship $-1 \le r \le 1$

doesn't depend on gradient

relationship $-1 \le r \le 1$

$$r_{xy} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2}\sqrt{\sum_i y_i^2 - n\bar{y}^2}}$$

① linear

1	Perfect
0.9 – 1	strong / very high
0.7 – 0.9	high
0.5 – 0.7	moderate
0.3 – 0.5	low
0 – 0.3	negligible

② (+)ve or (−)ve } r sign

doesn't depend on gradient

$r = 0.92$, $r = -0.92$ } strong

monotonic

non-monotonic

---

skewness $\boxed{S_k = \frac{3(\text{mean} - \text{median})}{\sigma}}$ ← S_x if sample

$S_x = 0$ symmetry μ $\sim N(\mu, \sigma^2)$

$S_x > 0$ (+)ve μ direction → tail → right

(−)ve μ

---

pop kurtosis $K_p = 3$ $\sim N$

Peakness

① flat / high

② compare $\sim N$

③ tails

$K_p > 3$ distinct high μ heavy tail

$K_p < 3$ flat

$$\cancel{K_p = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{(n-1) s^4}}$$

pop excess kurtosis $\boxed{EK_p = K_p - 3}$

- $= 0$ symmetrical $\sim N$
- > 0 high
- < 0 flat

---

distributions

cts

$\boxed{\text{① uniform}}$

$\frac{1}{b-a}$ a b

$\mu = \frac{b+a}{2}$

$\sigma^2 = \frac{(b-a)^2}{12}$

discrete

a b

1 2 3 4

1	2	3	4	n
1/n	1/n	.	..	

$E(X) = \sum x\, p(x)$ ✓

$V(X) = \sum x^2 p(x)$ ← ✓

---

② $X \sim Bin(n, p)$ each time n trials independent

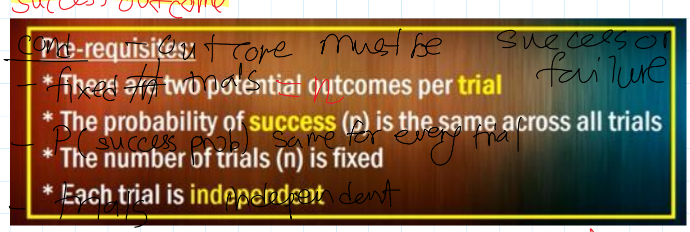
\# sample have the success outcome $= X = 0, 1, 2, \ldots$

$$\boxed{P(X) = {}^nC_x\, p^x (1-p)^{n-x}}$$

$$\boxed{\bar{x} = np}$$

$$\boxed{\sigma = \sqrt{npq}}$$

$q = 1 - p$



cont. outcome must be success or failure

fixed # trials

P(success prob) same for every trial

trials independent

$np \ge 5$ check first if we can approx to $\sim N$ (cts r.v)

---

③ $X \sim N(\mu, \sigma^2)$

or $p = \frac{X}{n} \sim N\left(\frac{np}{n}, \frac{npq}{n^2}\right)$

proportion $N\left(p, \frac{pq}{n}\right)$

if $\sim N(\mu, \sigma)$ $\boxed{\text{Normal = Bell}}$

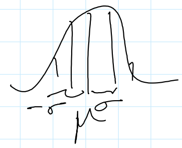
emperical rule (68, 95, 99.7) } prob

calc → σ use

$\hat{p}$ n n n^2

proportion of success

$N\left(p, \frac{pq}{n}\right)$



$\boxed{\text{calc} \to \sigma \quad \text{use } \sqrt{\sigma} \; ✱}$

---

④ Poisson

$X \sim Pois(\lambda)$

$$\boxed{P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}} \; ; \; x = 0, 1, 2$$

$np < 10$

$\lambda = np$

$p \to 0$

$n \to \infty$

✱ λ = mean × time period

$\boxed{X \sim \# \text{ trials until success}}$

① count

② Independent (random)

③ time interval
 ↳ area
 ↳ vol
 ↳ time

+ λ given ✱

④ p same over trials

$\boxed{\text{Discrete}}$

Basis for comparison	Covariance	Correlation
Definition	Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.	Correlation is a statistical measure that indicates how strongly two variables are related.
Values	The value of covariance lies in the range of -∞ and +∞.	Correlation is limited to values between the range -1 and +1
Change in scale	Affects covariance	Does not affect the correlation
Unit-free measure	No	Yes

height and weight

height scale changes covariance changes (not useful)