Exactly one empty box when n balls are randomly placed in n boxes: $\frac{\binom{n}{2} n!}{n^n}$

$P(\bar{B} \mid A) = 1 - P(B \mid A)$

$P(\bar{A} \mid B) = 1 - P(A \mid B)$

GROUP 1 : SETS, COUNTING

SETS

Union : $A \cup B$

Intersection : $A \cap B$

Disjoint events : $A \cap B = AB = \emptyset$

PERMUTATIONS

permutations of n distinct object taken k at a time

$$A_n^k = \frac{n!}{(n-k)!} = n(n-1)\dots(n-k+1)$$

→ ORDERED, NO REPEAT

n- permutations of n objects

$$P_n = n!$$

→ ORDERED, NO REPEAT

COMBINATION

n- combinations, taken k

$$C_n^k = \frac{n!}{(n-k)!\,k!} = \binom{n}{k}$$

→ NO ORDER, NO REPEAT

ways to choose k from n with replacement

$$\overline{A}^k = n^k$$

The number of distinct permutations of n things, n_1 of one kind, n_2, ...

$$\frac{n!}{n_1!\,n_2!\,n_3!\dots n_k!}$$

Circle : $(n-1)!$

GROUP 2: PROBABILITY

DEFINITIONS

Sample space Ω — Set of all possible outcomes

Event E — $E \subseteq \Omega$

→ Assume Ω have equally likely elements $\quad P(E) = \frac{|E|}{|\Omega|}$

PROPERTIES

ADDITION RULES

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$ → disjoint = 0

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

CONDITIONAL PROBABILITY

$$P(A|B) = \frac{P(AB)}{P(B)} \quad \neq 0$$

→ A, B are independent if: $P(B|A) = P(B)$, $P(A|B) = P(A)$ → $P(A \cap B) = P(A) \cdot P(B)$

PRODUCT RULE

$$P(AB) = P(A)\,P(B|A)$$
$$P(BA) = P(B)\,P(A|B)$$

BOTH CAN OCCUR

$= P(A)$ if independent

NOTE: Sum of r-digit numbers, taken from n numbers, without repitition:

$$({}^{n-1}P_{r-1})\,(\text{sum of } n \text{ numbers})\,(\underbrace{111\ldots1}_{r \text{ times}})$$

BAYES' THEOREM

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

TOTAL PROBABILITY

$$P(A) = \sum_n P(A|B_n) \cdot P(B_n)$$

BERNOULLI'S TRIAL

A → k successes → p

A → $n-k$ failures → $(1-p)$

$$P(A) = \binom{n}{k} p^k (1-p)^{n-k}$$

MUTUALLY EXCLUSIVE / DISJOINT

$$\begin{cases} P(AB) = 0 \\ P(A \cup B) = P(A) + P(B) \end{cases}$$

INDEPENDENT

$$\begin{cases} P(AB) = P(A)\,P(B) \\ P(A \cup B) = P(A) + P(B) - P(A)\,P(B) \end{cases}$$

Note: $P(A) \Rightarrow P(B)$ but $P(B) \not\Rightarrow P(A)$ then $P(B) \geq P(A)$

GROUP 3: DISCRETE RANDOM VARIABLES

RANDOM VARIABLES

Suppose we conduct an exp with sample space Ω

A RV is a numeric function of the outcome: $X : \Omega \to \mathbb{R}$

$\omega \to X(\omega)$

$\to \Omega_X$ is the set of all variables of X $\equiv$ range / support

- Ω_X finite or countably infinite: DISCRETE
- Ω_X is uncountably large: CONTINUOUS

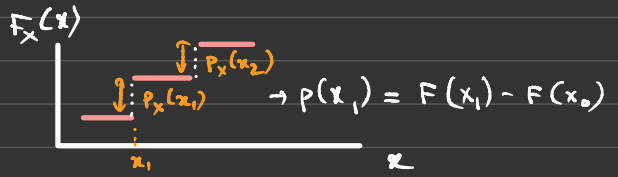
PMF — DISCRETE

$$p_X : \Omega_X \to [0,1], \quad k \to P(X = k) \longrightarrow \begin{cases} \sum_{i \in \Omega_X} p_X(i) = 1 \\ p(x) = \text{Prob}(X = x) \end{cases}$$

NOTE: Given $f(x)$ $\to$ x_{med}: $F_X(x) = \frac{1}{2}$; x_{mod}: $f_X(x) = \max$ (consider f'_X)

CDF — DISCRETE

$$F_X(x) = P[X \leq x]$$



$$P(a < X \leq b) = F_X(b) - F_X(a)$$
$$0 \leq F_X(a) \leq 1$$
$$P(X < b) = P(X < a) + P(a \leq X < b)$$
$$F(-\infty) = 0 \; ; \; F(+\infty) = 1$$

EXPECTATION — DISCRETE

$$E[X] = \sum_{k \in \Omega_X} k \cdot p_X(k)$$

Average of possible values, weighted by their prob

$$E[X^2] = \sum x^2 p_X(x)$$

VARIANCE — DISCRETE

$$Var(X) = E[X^2] - E[X]^2 \geq 0$$

How "spread out" the distribution is

STANDARD DEVIATION — DISCRETE

$$\sigma_X = \sqrt{Var(X)}$$

$$Var[aX + b] = a^2 Var(X)$$
$$\forall a, b \in \mathbb{R}$$

LINEARITY OF EXPECTATION — DISCRETE

$$E[X+Y] = E[X] + E[Y]$$

$$E[aX+b] = aE[X] + b$$

$$\rightarrow E[aX+bY+c] = aE[X] + bE[Y] + c$$

a, b, c are scalars
X, Y are RVs

LOTUS — DISCRETE

X : RV, $g : D \rightarrow \mathbb{R}$ is a function defined atleast over $\Omega_X \subseteq D$:

$$E[g(X)] = \sum_{b \in \Omega_X} g(b)\, p_X(b)$$

Note: $E[g(X)] \neq g(E[X])$

$$\rightarrow \sigma^2_{g(X)} = E[g(X) - \mu_{g(X)}]^2 = \sum_{x \in S_X} [g(X) - \mu_{g(X)}]^2 p_X(x)$$

DISCRETE RV DISTRIBUTIONS

BERNOULLI

$X \sim \text{Ber}(p) \iff \Omega_X = [0,1]$ and:

$$\text{PMF} \equiv p_X(k) = \begin{cases} p & , k=1 \\ 1-p & , k=0 \end{cases}$$

$\rightarrow$ $E[X] = p$ and $\text{Var}(X) = p(1-p)$

YES/NO QUESTIONS

BINOMIAL

$X \sim B(n,p) \iff \forall k \in \Omega_X = \{0,1,2,\dots,n\}$,

$$\text{PMF} \equiv p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$\rightarrow$ $E[X] = np$ and $\text{Var}(X) = np(1-p)$

SUCCESSES/FAILURES ?

GEOMETRIC

$X \sim \text{Geo}(p) \iff \Omega_X = \{1,2,3,\dots\}$

$$\text{PMF} \equiv p_X(k) = (1-p)^k p \qquad k = 1,2,\dots$$

$\rightarrow$ $E[X] = \frac{1}{p}$ and $\text{Var}(X) = \frac{1-p}{p^2}$

BAO LÂU NỮA MỚI ĐƯỢC 1 TỶ GÓI MÈ ?

How many flips to have a heads for the first time ?

How many (independent) flips to get the r-th heads?

NEGATIVE BINOMIAL

$X \sim \text{NegBin}(r, p) \iff \Omega_X = \{r, r+1, \dots\}$:

$$\text{PMF} \equiv p_X(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

$k = r, r-1, \dots$

SUM OF Geo(p) vars

$$\rightarrow E[X] = \frac{r}{p} \quad \text{and} \quad \text{Var}(X) = \frac{r(1-p)}{p^2}$$

UNIFORM — DISCRETE

$X \sim \text{Unif}(a, b)$, $a < b \in \mathbb{Z}$ $\iff$ PMF:

$$p_X(k) = \begin{cases} \frac{1}{b-a+1} & , k \in \{a, a+1, \dots b\} \\ 0 & , \text{ow} \end{cases}$$

$$\rightarrow E[X] = \frac{a+b}{2} \quad \text{and} \quad \text{Var}(X) = \frac{(b-a)(b-a+2)}{12}$$

KNOWN, FINITE # OUTCOMES EQUALLY LIKELY TO HAPPEN

POISSON

$X \sim \mathcal{P}(\lambda) \iff \Omega_X = \{0, 1, 2, \dots\}$

$$\text{PMF} \equiv p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

$\lambda > 0$ is the RATE OF OCCURENCE

$\#$ OUTCOMES OCCURING IN A TIME INTERVAL

$$\Rightarrow \boxed{E[X] = \text{Var}(X) = \lambda}$$

BINOMIAL APPROXIMATION BY POISSON DISTRIBUTION

$$\mathcal{B}(n, p) \to \mathcal{P}(\lambda) \quad \text{as } n \to \infty, \; p \to 0$$

ideal: $\lambda = np < 7$

GROUP 4: CONTINUOUS RVs

PDF
— CONTINUOUS

PDF: $f_X : \mathbb{R} \to \mathbb{R}$:

$$f_X(x) \geq 0 \quad , \forall x \in \mathbb{R}$$
$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$

} condition

$$P(a < X < b) = \int_a^b f_X(x)\,dx$$

CDF
— CONTINUOUS

CDF: $F_X : \mathbb{R} \to \mathbb{R}$

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt, \quad t \in \mathbb{R}$$

$$\frac{dF_X(x)}{dx} = f_X(x)$$

$$P(a \leq X \leq b) = F_X(b) - F_X(a)$$

$$F_X(-\infty) = 0, \quad F_X(+\infty) = 1$$

$$F_X(c) \leq F_X(d) \quad , \forall c \leq d$$

EXPECTATION
- CONTINUOUS

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x)\, dx$$

LOTUS
- CONTINUOUS

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx$$

$$\rightarrow \quad \sigma_g^2 = E[g(X) - \mu_{g(X)}]^2 = \int_{-\infty}^{+\infty} [g(X) - \mu_{g(X)}]^2 f_X(x)\, dx$$

CONTINUOUS RV DISTRIBUTIONS

UNIFORM — CONTINUOUS

$X \sim \text{Unif}(a,b)$, $a < b \in \mathbb{R}$ $\Leftrightarrow$ PDF:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0 & \text{ow} \end{cases}$$

< x can take ∀ values in [a,b]. >

$$P(a \le x \le b) = \int_a^b f_X(x)$$

$\rightarrow$ $E[X] = \frac{a+b}{2}$ and $\text{Var}(X) = \frac{(b-a)^2}{12}$,

$$F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$$

NORMAL — CONTINUOUS

$X \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow \Omega_X = (-\infty, +\infty)$:

$$\text{PDF} \equiv f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

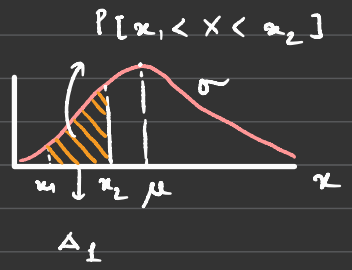
$\rightarrow$ $E[X] = \mu$, $\text{Var}(X) = \sigma^2$

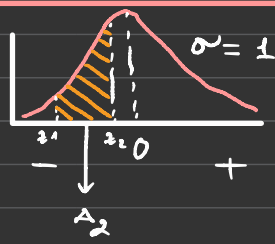
STANDARD NORMAL DISTRIBUTION

$Z \sim \mathcal{N}(0,1)$, $z = \frac{x-\mu}{\sigma}$

$$CDF(z) \equiv \varphi_z(a) = F_z(a) = P(z \leq a)$$

$$\varphi(-a) = 1 - \varphi(a)$$





$z = \frac{x-\mu}{\sigma}$

$\rightarrow$ $A_1 = A_2$

NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

$$P(x_1 < X < x_2) = \Phi\left(\frac{x_2 + 0.5 - \mu}{\sigma}\right) - \Phi\left(\frac{x_2 - 0.5 - \mu}{\sigma}\right)$$

$\sigma \rightarrow \sqrt{npq}$

Ideal: $np > 5$ and $n(1-p) > 5$

EXPONENTIAL

$X \sim \text{Exp}(\lambda) \Leftrightarrow \Omega_X = [0, +\infty)$

$$\text{~~PMF~~ PDF} \equiv f_X(x) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & , \text{ow} \end{cases}$$

X is the waiting time until the first occurrence of an event in a Poisson process with para. λ → POISSON DISTRIBUTION : # events occuring in a time interval

EXPONENTIAL DISTRIBUTION: Time taken between 2 events

→ $E[X] = \frac{1}{\lambda}$, $\text{Var}(X) = \frac{1}{\lambda^2}$

$$\text{CDF} \equiv F_X(x) = \begin{cases} 1 - e^{-\lambda x} & , x \geq 0 \\ 0 & , \text{ow} \end{cases}$$

JOINT CDF OF TWO RVs

PROPERTIES

$$F_{X,Y}(x,y) = P[X<x,\ Y<y]$$

$$0 \le F_{X,Y}(x,y) \le 1$$

$$F_X(x) = F_{X,Y}(x,\infty)$$
$$F_Y(y) = F_{X,Y}(\infty,y)$$
} MARGINAL CDF

$$F_{X,Y}(-\infty,y) = F_{X,Y}(x,-\infty) = 0$$

$$F_{X,Y}(\infty,\infty) = 1$$

If $x_1 < x_2$, $y_1 < y_2$ then:

$$P[x_1 \le X \le x_2,\ y_1 \le Y \le y_2] = F_{X,Y}(x_2,y_2) - F_{X,Y}(x_2,y_1) - F_{X,Y}(x_1,y_2) + F_{X,Y}(x_1,y_1)$$

GROUP 5 : JOINT DISCRETE RVs

JOINT PMF

Let X, Y be discrete random variables

→ Joint PMF: $\boxed{p_{X,Y}(a,b) = P(X=a, Y=b)}$

Joint range: $\Omega_{X,Y} = \{(c,d) : p_{X,Y}(c,d) > 0\} \subseteq \Omega_X \times \Omega_Y$

$$\sum_{(s,t)\in\Omega_{X,Y}} p_{X,Y}(s,t) = 1$$

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_n	$\sum_j$
x_1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1n}	$P[X=x_1]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{in}	$P[X=x_i]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	p_{m1}	$\cdots$	p_{mj}	$\cdots$	p_{mn}	$P[X=x_m]$
$\sum_i$	$P[Y=y_1]$	$\cdots$	$P[Y=y_j]$	$\cdots$	$P[Y=y_n]$	$\sum_i\sum_j = 1$

THEOREM

$0 \le p_{ij} \le 1, \quad \forall i = \overline{1,m}, \quad \forall j = \overline{1,n}$

$\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} = 1$

→ $F_{XY}(x,y) = \sum_{x_i<x}\sum_{y_j<y} p_{ij}$

MARGINAL PMF

$$P_X(a) = \sum_{b \in \Omega_Y} P_{X,Y}(a,b)$$

$$P_Y(b) = \sum_{a \in \Omega_X} P_{X,Y}(a,b)$$

LOTUS $g : \mathbb{R}^2 \to \mathbb{R}^2$

$$E[g(X,Y)] = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} g(x,y)\, P_{X,Y}(x,y)$$

We display $P_X(x)$ and $P_Y(y)$ by rewriting the matrix in Example 1 and placing the row sums and column sums in the margins.

$P_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	$P_X(x)$
$x=0$	0.01	0	0	0.01
$x=1$	0.09	0.09	0	0.18
$x=2$	0	0	0.81	0.81
$P_Y(y)$	0.10	0.09	0.81	

$$P_X(x) = \begin{cases} 0.01, & x=0, \\ 0.18, & x=1, \\ 0.81, & x=2, \\ 0, & \text{otherwise.} \end{cases} \qquad P_Y(y) = \begin{cases} 0.1, & y=0, \\ 0.09, & y=1, \\ 0.81, & y=2, \\ 0, & \text{otherwise.} \end{cases}$$

GROUP 6: JOINT CONTINUOUS RVs

JOINT PDF

Joint PDF: $f_{X,Y}(a,b) > 0 = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$

→ Joint range: $\Omega_{X,Y} = \{(c,d) : f_{X,Y}(c,d) > 0\} \subset \Omega_X \times \Omega_Y$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$$

MARGINAL PDF

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$ → domain of y

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx$$ → domain of x

LOTUS

$g: \mathbb{R}^2 \to \mathbb{R}^2$

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f_{X,Y}(x,y)\,dy\,dx$$

→ $$P(a \le X \le b,\ c \le Y \le d) = \int_a^b\int_c^d f_{X,Y}(x,y)\,dy\,dx$$

EXPECTED VALUE OF TWO RVs

For R.Vs : X and Y, the expected value of $W = g(X,Y)$ is :

DISCRETE

$$E[W] = \sum_{x \in S_X} \sum_{y \in S_Y} g(x,y) \cdot P_{X,Y}(x,y)$$

$$= \sum_i \sum_j g(x_i, y_j) \, P[X = x_i, Y = y_j]$$

CONTINUOUS

$$E[W] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) \, f_{XY}(x,y) \, dx \, dy$$

THEOREMS

$$E[X+Y] = E[X] + E[Y]$$

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2 \cdot E[(X-\mu_X)(Y-\mu_Y)]$$

GROUP 7: COVARIANCE, CORRELATION

DEFINITION

Covariance of X and Y:

$$\begin{aligned} Cov(X,Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

$$E[XY] = \begin{cases} \sum_i \sum_j x_i y_j p_{ij} \\ \iint xy f(x,y) \end{cases}$$

PROPERTIES

$$Cov(X,X) = Var(X)$$
$$Cov(X,Y) = Cov(Y,X)$$
$$Cov(aX+b, cY+d) = ac \cdot Cov(X,Y)$$
$$Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$$
$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

COVARIANCE MATRIX

$$\Gamma = \begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix} = \begin{bmatrix} Var(X) & Cov(X,Y) \\ Cov(Y,X) & Var(Y) \end{bmatrix}$$

CORRELATION

$$r_{X,Y} = E[XY]$$

$\rightarrow$ { X, Y are orthogonal: $r_{X,Y} = 0$

X, Y are uncorrelated: $\text{Cov}[X,Y] = 0$

CORRELATION COEFFICIENT

$$\rho_{X,Y} = \frac{\text{Cov}[X,Y]}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{\text{cov}[X,Y]}{\sigma_X \sigma_Y}$$

$$-1 \le \rho_{X,Y} \le 1$$

$$Y = aX + b : \quad \rho_{X,Y} = \begin{cases} -1 & , a < 0 \\ 0 & , a = 0 \end{cases}$$

CONDITIONING BY AN EVENT

CONDITIONAL JOINT PMF

$$p_{X,Y|B}(x,y) = P[X = x, Y = y \mid B]$$

B is a region

$$= \begin{cases} \dfrac{p_{X,Y}(x,y)}{P[B]} & , (x,y) \in B \\ 0 & \text{ow} \end{cases}$$

$P[B] \rightarrow \sum\sum$

CONDITIONAL JOINT PDF

$$f_{X,Y|B}(x,y) = \begin{cases} \dfrac{f_{X,Y}(x,y)}{P[B]} & , (x,y) \in B \\ 0 & \end{cases}$$

$P[B] \rightarrow \iint_B f_{XY}(x,y)$

CONDITIONING BY A RV

CONDITIONAL EXPECTED VALUE

$$E[W|B] = \sum_{x \in S_X} \sum_{y \in S_Y} g(x,y)\, p_{X,Y|B}(x,y)$$

$$E[W|B] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y)\, f_{X,Y|B}(x,y)\, dx\, dy$$

CONDITIONAL VARIANCE

$$\mathrm{Var}[W|B] = E[W^2|B] - (E[W|B])^2$$

CONDITIONAL PMF

$$\Rightarrow \quad p_{X|Y}(x|y) = P[X = x \mid Y = y]$$

$$p_{X,Y}(x,y) = p_{X|Y}(x|y)\, p_Y(y) = p_{Y|X}(y|x)\, p_X(x)$$

CONDITIONAL EXPETED VALUE

$$E[X|Y=y] = \sum_{x \in S_X} x\, p_{X|Y}(x|y)$$

CONDITIONAL PDF

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

GROUP 8: INDEPENDENT RVs

INDEPENDENCE

Discrete: $P_{X,Y}(x,y) = P_X(x)\, P_Y(y)$

Continuous: $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$

$$\Rightarrow \begin{cases} P_{X|Y}(x,y) = P_X(x), & P_{Y|X}(y,x) = P_Y(y) \\ f_{X|Y}(x,y) = f_X(x), & f_{Y|X}(y|x) = f_Y(y) \end{cases}$$

PROPERTIES

$$E[g(X)\, h(Y)] = E[g(X)]\, E[h(Y)]$$

$$r_{X,Y} = E[XY] = E[X]\, E[Y]$$

$$Cov[X,Y] = \rho_{X,Y} = 0$$

$$Var[X+Y] = Var[X] + Var[Y]$$

$$E[X \mid Y = y] = E[X]$$

$$E[Y \mid X = x] = E[Y]$$

GROUP 9: LIMIT THEOREMS

SAMPLE MEAN

Let $X_{\overline{1,n}}$ be independent identically distributed RVs, μ and σ^2

Sample mean:

$$\overline{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

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

LIMIT THEOREMS

LAW OF LARGE NUMBERS

As $n \to \infty$, $\overline{X}_n \xrightarrow{prob}$ converges to the true mean μ

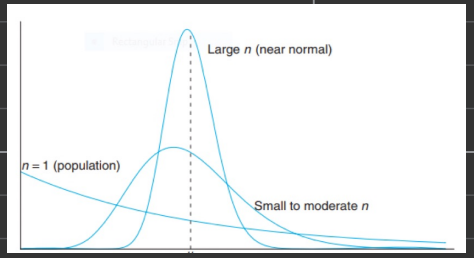
$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1$$

CENTRAL LIMIT THEOREM

The standardised mean approaches the standard Normal distribution

$$\overline{Z}_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \to N(0,1) \text{ as } n \to \infty$$

ideal: $n \geq 30$, not skewed



Sum: $Z_n = \frac{T - n\mu}{\sqrt{n}\sigma} \sim N(n\mu, \sqrt{n}\sigma)$

GROUP 10 : RANDOM SAMPLES - STATISTICS

SAMPLE MEAN, MEDIAN, MOD

SAMPLE VARIANCE, S.D, RANGE

SAMPLE MEAN

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

SAMPLE MEDIAN

$$\hat{x} = \begin{cases} x_{(n+1)/2}, & n \text{ odd} \\ \frac{1}{2}\left(x_{n/2} + x_{\frac{n}{2}+1}\right), & n \text{ even} \end{cases}$$

sorted in increasing order

SAMPLE VARIANCE

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$= \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]$$

SAMPLE S.D

$$S = \sqrt{S^2}$$

RANGE

$$R = x_{max} - x_{min}$$

Note

Sample	Population
n: number of measurements in the sample	N: number of measurements in the population
$\overline{x}$: sample mean	μ: population mean
s^2: sample variance	σ^2: population variance
s: sample standard deviation	σ: population standard deviation

GROUP 11: SAMPLING DISTRIBUTION

S.D OF MEANS

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$$

↳ has a normal distribution

$$\begin{cases} \mu_{\bar{x}} = \mu \\ \sigma^2_{\bar{x}} = \dfrac{\sigma^2}{n} \end{cases}$$

S.D OF DIFFERENCE OF MEANS

Theorem 3

If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\overline{X}_1 - \overline{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

is approximately a standard normal variable $\mathcal{N}(0, 1)$.

S.D OF S^2

$S^2 \sim \sigma^2$: random sample of size n $\quad \chi^2 = \dfrac{(n-1)^2 S^2}{\sigma^2} = \displaystyle\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2}$

SPECIAL: STUDENT t- DISTRIBUTION

DEFINITIONS

Theorem 5

Let Z be a standard normal random variable and V a chi-squared random variable with ν degrees of freedom. If Z and V are independent, then the distribution of the random variable T, where

$$T = \frac{Z}{\sqrt{V/(n-1)}}$$

is given by the density function

$$h(t) = \frac{\Gamma(v+1)/2}{\Gamma(v/2)\sqrt{\pi v}}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty < t < +\infty.$$

This is known as the *t*-distribution with ν degrees of freedom.

$$h(t) = \frac{\Gamma(\upsilon+1)/2}{\Gamma(\upsilon/2)\sqrt{\pi\upsilon}}\left(1+\frac{t^2}{\upsilon}\right)^{-(\upsilon+1)/2}$$

↳ Let $X_1, \dots, X_n$ be independent RVs, normal with μ, σ^2.

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Then RV: $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ has a t-distribution, $\upsilon = n - 1$

↳ deg. of freedom

t-dis $\xrightarrow{n\to\infty}$ Normal

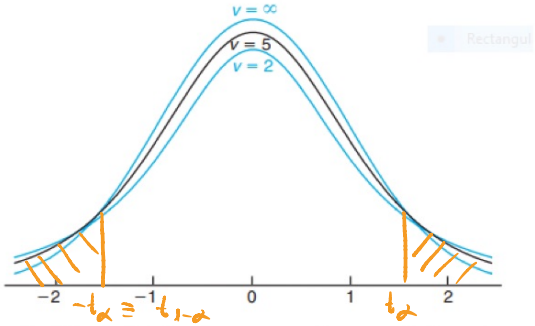


Figure 5: The t-distribution curves for $\nu = 2, 5$, and ∞

GROUP 12 : ESTIMATIONS

POINT ESTIMATE

UNBIASED ESTIMATOR

A statistic $\hat{\Theta}$ is said to be an unbiased estimator of para. θ if: $\mu_{\hat{\Theta}} = E[\hat{\Theta}] = \theta$

↳ Distribution centred at θ

→ Smallest variance = most efficient estimator of θ.

INTERVAL ESTIMATE

Interval $\hat{\theta}_L < \theta < \hat{\theta}_U$, $\hat{\theta}_{L,U}$ depend {on the value of $\hat{\theta}$ for a sample / sampling dist. of $\hat{\theta}$.

→ $P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1-\alpha$, $0 < \alpha < 1$

confidence interval → $100(1-\alpha)\%$ confidence interval

→ Shorter interval, with higher degree of confidence is BETTER!

MEAN ESTIMATING

KNOWN σ

TWO SIDED INTERVALS

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$$

$$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

CONFIDENCE for μ

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

ERROR $100(1-\alpha)\%$ confident that the error not exceed:

$$e = \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}$$

SAMPLE SIZE confident that error $\leq e$ when:

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{e}\right)^2$$

ONE-SIDED BOUNDS

CLT $\Rightarrow$ $P\left(\mu < \bar{X} + \frac{z_\alpha \sigma}{\sqrt{n}}\right) = 1-\alpha$

- upper bound: $\bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$
- lower bound: $\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}$

UNKNOWN σ

$$P\left(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}\right) = 1-\alpha$$

< similar to known σ, but $\sigma \to s$ and $N \to t$-dist. >

CONFIDENCE INTERVAL $\quad \bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

BOUNDS

- upper: $\bar{x} + t_\alpha \, s/\sqrt{n}$
- lower: $\bar{x} - t_\alpha \, s/\sqrt{n}$

$n \geq 30$, σ unknown $\Rightarrow$ $s \leftarrow \sigma$: $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

PROPORTION ESTIMATING

A point estimator of the proportion p in a binomial exp is given by stat. $\hat{P} = X/n$, X = # successes in n trials. → $\hat{p} = \frac{x}{n}$: point est.

CLT → $\mu_{\hat{p}} = p$, $\sigma^2_{\hat{p}} = \frac{pq}{n}$

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha \qquad , Z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

CONFIDENCE INTERVAL

$\hat{q} = 1 - \hat{p}$

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$n\hat{p}$ and $n\hat{q} \geq 5$

ERROR $\quad e = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

SAMPLE SIZE $\quad n = \frac{z^2_{\alpha/2}\,\hat{p}\hat{q}}{e^2}$ // % confident

$n = \frac{z^2_{\alpha/2}}{4e^2}$ // at least % confident

GROUP B: HYPOTHESIS TESTING

DEFINITIONS

Alternative hypothesis H_1 : which we will wish to support : $\mu_1 = \mu_2$

Null hypothesis H_0 : contradiction

- $\mu_1 < \mu_2$: left-tailed test
- $\mu_1 > \mu_2$: right-tailed test
- $\mu_1 \neq \mu_2$: two-tailed test

- Assume H_0 true $\xrightarrow{\text{standardised test statistic}}$
  - $\notin$ rejection region : fail to reject H_0
  - $\in$ rejection region : reject H_0

at $a\%$ level of significance

- Error

	H_0 true	H_0 false
Not reject H_0	Correct	Type II
Reject H_0	Type I	Correct

$\rightarrow P[.] = \beta$

$P[.] = \alpha$

$\rightarrow \alpha \sim \frac{1}{\beta}$

Approach to Hypothesis Testing with Fixed Probability of Type I Error

1. State the null and alternative hypotheses.
2. Choose a fixed significance level α.
3. Choose an appropriate test statistic and establish the critical region/rejection region based on α.
4. Reject H_0 if the computed test statistic is in the critical region/rejection region. Otherwise, do not reject.
5. Draw scientific or engineering conclusions.

SINGLE SAMPLE HYPOTHESIS TESTING

SINGLE MEAN, VARIANCE KNOWN

TWO - TAILED

1) $H_0 : \mu = \mu_0$ $\quad H_1 : \mu \neq \mu_0$

2) $$z = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$$

3) $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$ → Reject H_0

   $-z_{\alpha/2} < z < z_{\alpha/2}$ → Fail to reject H_0

ONE - TAILED

1) $H_0 : \mu = \mu_0$ $\quad H_1 : \mu > \mu_0$ (Right)
   $\mu < \mu_0$ (Left)

2) $$z = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$$

3) $z > z_\alpha$ → Reject
   $z < z_\alpha$ → Fail } Right

   $z < -z_\alpha$ → Reject
   $z > -z_\alpha$ → Fail } Left

SINGLE SAMPLE, VARIANCE UNKNOWN

TWO-TAILED

1) $H_0 : \mu = \mu_0$ $\qquad H_1 : \mu \neq \mu_0$

2) $$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$$

3) $t < -t_{\alpha/2}^{(n-1)}$ or $t > t_{\alpha/2}$ $\rightarrow$ Reject

$-t_{\alpha/2} < t < t_{\alpha/2}$ $\rightarrow$ Fail

ONE-TAILED

1) $H_0 : \mu = \mu_0$ $\qquad H_1 : \mu > \mu_0$ (R) $\quad \mu < \mu_0$ (L)

2) $$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$$

3) Right:
$t > t_\alpha \rightarrow$ Reject
$t < t_\alpha \rightarrow$ Fail

Left:
$t < -t_\alpha \rightarrow$ Reject
$t > -t_\alpha \rightarrow$ Fail

$n \geq 30 \rightarrow T \sim N(0,1)$ $\qquad t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$

PROPORTIONS

TWO-TAILED

1) $H_0: p = p_0$ $\quad H_1: p \neq p_0$

2) $$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$$

NOTE: $\begin{cases} np_0 \geq 5 \\ n(1-p_0) \geq 5 \end{cases}$

3) $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$ → reject

$-z_{\alpha/2} < z < z_{\alpha/2}$ → Fail

Similarly for one-tailed ~

TWO SAMPLES HYPOTHESIS TESTING

CONDITIONS:

1) Samples are randomly selected
2) Samples are independent
3) Each sample size ≥ 30; or each must have normal distribution, known standard deviation

σ_1^2, σ_1^2 KNOWN

$H_0 : \mu_1 - \mu_2 = D_0$

$H_1 : \mu_1 - \mu_2 > D_0$ or $\mu_1 - \mu_2 < D_0$

$H_1 : \mu_1 - \mu_2 \neq D_0$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Rejection region: Similar

σ_1^2, σ_2^2 UNKNOWN
$n_1, n_2 \geq 30$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

σ_1^2, σ_2^2 UNKNOWN
$n_1, n_2 < 30$

CONDITIONS

1) Samples are randomly selected
2) Independent
3) Each sample has a normal distribution, population variances are equal.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Rejection region:

- One-tailed: $t > t_{\alpha}^{(n_1+n_2-2)}$ → Reject (Right)
- One-tailed: $t < -t_{\alpha}^{(n_1+n_2-2)}$ → Reject (left)
- Two-tailed: $t > t_{\alpha/2}^{(n_1+n_2-2)}$ or $t < -t_{\alpha/2}^{(n_1+n_2-2)}$

DIFFERENCE BETWEEN PROPORTIONS

CONDITIONS

1) Samples are randomly selected
2) Independent
3) Large enough for normal sampling distributions

$$\begin{cases} n_1 p_1 \geq 5 \;,\; n_1(1-p_1) \geq 5 \\ n_2 p_2 \geq 5 \;,\; n_2(1-p_2) \geq 5 \end{cases}$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad , \quad \bar{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

H_0	H_1	Rejection region W_α
$p_1 = p_2$	$p_1 \neq p_2$	$(-\infty; -z_{\alpha/2}) \cup (z_{\alpha/2}; +\infty)$
$p_1 = p_2$	$p_1 > p_2$	$(z_\alpha; +\infty)$
$p_1 = p_2$	$p_1 < p_2$	$(-\infty; -z_\alpha)$