SUMMARY

- **SAMPLE SPACE AND EVENTS:** (~> BRIEF SUMMARY, ALREADY KNOWN STUFF)

- "SAMPLE SPACE": ALL POSSIBLE OUTCOMES OF AN EXPERIMENT ~> "S"
- "EVENT": ANY SUBSET OF S

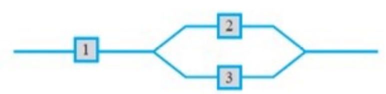
es. COIN TOSS, A = EVEN n° OF DOTS ON UPPER FACE

$\rightarrow S = \{1, 2, 3, 4, 5, 6\}, \quad A = \{2, 4, 6\}$

- UNION OF EVENTS: $A \cup B$, A OR B
- INTERSECTION OF EVENTS: $A \cap B$, A AND B
- COMPLEMENT OF A: $\bar{A}$
- MUTUALLY EXCLUSIVE EVENTS: A & B, IF NO OUTCOMES IN COMMON ~> $A \cap B = \emptyset$

es. PAG. 17

- IF COMPONENT WORKS: S
- " " NOT " : F



$\rightarrow \# S = 2^3 = \{(\overset{1}{F}, \overset{2}{F}, \overset{3}{F}), \ldots, (S, S, S)\}$

~> SYSTEM WORKS $\Leftrightarrow A = \{(S, S, F), (S, F, S), (S, S, S)\}$

es. PAG. 19

An academic department has just completed voting by secret ballot for a department head. The ballot box contains four slips with votes for candidate *A* and three slips with votes for candidate *B*. Suppose these slips are removed from the box one by one.

a. List all possible outcomes.

A B B A A ...
A 1 1 1 2 3
B 0 1 2 2 2

b. Suppose a running tally is kept as slips are removed. For what outcomes does *A* remain ahead of *B* throughout the tally?

a. $A \times 4, \quad B \times 3$

$\rightarrow \# A = \frac{7!}{4!\,3!} = \binom{7}{3} = \binom{7}{4}$

"n chose k"

$\rightarrow \binom{n}{k} = \binom{n}{n-k}$

- **PROBABILITY**:

"QUANTITATIVE MEASURE OF HOW LIKELY IS AN EVENT TO OCCUR"

- 3 DEFINITIONS:
  - CLASSICAL: $P(A) = \#A / \#S$
  - FREQUENTIST: $P(A)$ IS THE PORTION OF TIMES THAT EVENT A WOULD OCCUR IN THE LONG RUN, IF THE EXPERIMENT WERE TO BE REPEATED OVER AN OVER AGAIN
  - SUBJECTIVE (DE FINETTI, 1933): $P(A)$ IS THE AMOUNT OF MONEY I'M WILLING TO BET FOR WINNING 1 UNIT OF MONEY

PROBABILITY AXIOMS:

- $0 \leq P(A) \leq 1$
- $P(S) = 1$
- IF $A_1, \dots, A_n$ ARE DISJOINT EVENTS $\rightarrow P(A_1 \cup \dots \cup A_n) = P(A_1) + \dots + P(A_n)$
- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

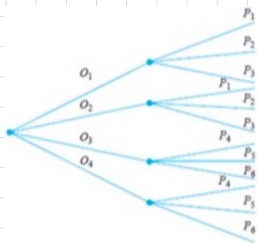
DETERMINING PROBABILITY:

LET $A = E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n \;/\; \underset{i \neq j}{E_i \cap E_j = \emptyset}$

$$\rightarrow \sum P(E_i) = 1 = P(A)$$

- FOR EQUALLY LIKELY OUTCOMES: $P(A) = \sum_{E_i \text{ in } A} P(E_i) = \sum_{E_i \text{ in } A} \frac{1}{N}$

COUNTING TECHNIQUES:



LET $O = \{\underbrace{O_1, \dots, O_4}_{n_1 = 4}\}$, IF FOR EACH O_i EXISTS $P = \{\underbrace{P_1, \dots, P_3}_{n_2 = 3}\}$

$\rightarrow$ # OF PAIRS $= n_1 \cdot n_2 = 12$

- COMBINATIONS: $\binom{n}{k}$
- PERMUTATION: $\frac{n!}{(n-k)!}$

Disposizioni:

Semplici: $D_{n,k} = \frac{n!}{(n-k)!}$

Ripetzione: $D'_{n,k} = n^k$

Permutazioni:

Semplici: $P_n = n!$

Ripetzione: $P_n^{(\dots,\dots)} = \frac{n!}{\dots!\dots!}$

Combinazioni:

Semplici: $C_{n,k} = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$

Ripetzione: $C'_{n,k} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!\,(n-1)!}$

ex. BIRTHDAYS PARADOX

$\rightarrow$ WHAT IS THE p THAT AT LEAST 2 PEOPLE CELEBRATE BIRTHDAY ON THE SAME DAY?

LET n = n° PEOPLE CONSIDERED IN TOTAL $\rightarrow 2 \leq n < 365$

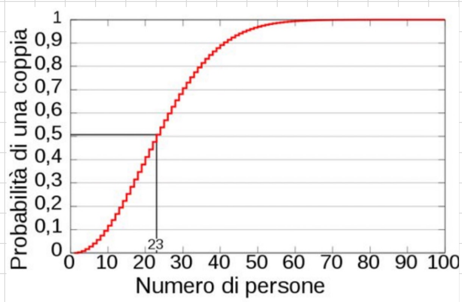
- IF $n = 2$: $\#S = 365^2 \rightarrow P(E_2) = \frac{\#E_2}{\#S} = \frac{365}{365^2} = \frac{1}{365}$
- IF $n = 3$: $\#S = 365^3 \rightarrow \#E_3 = ?$ $\rightarrow$ IT'S EASIER TO COUNT $\bar{E}_3$

$\rightarrow \#\bar{E}_3 = 365 \cdot 364 \cdot 363$

$\vdots$

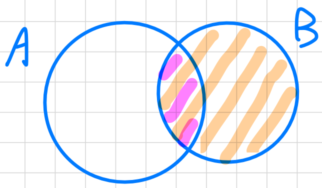
$$\rightarrow P(E_n) = 1 - \frac{\left(\frac{365!}{(365-n)!}\right)}{365^n}$$

$\rightarrow$ IF $n \geq 23 \rightarrow P(E_n) \geq 50\%$!



- **CONDITIONAL PROBABILITY AND INDEPENDENCE:**

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



- **TOTAL LAW PROBABILITY:**

IF $A_1, \dots, A_n$ ARE MUTUALLY EXCLUSIVE EVENTS :

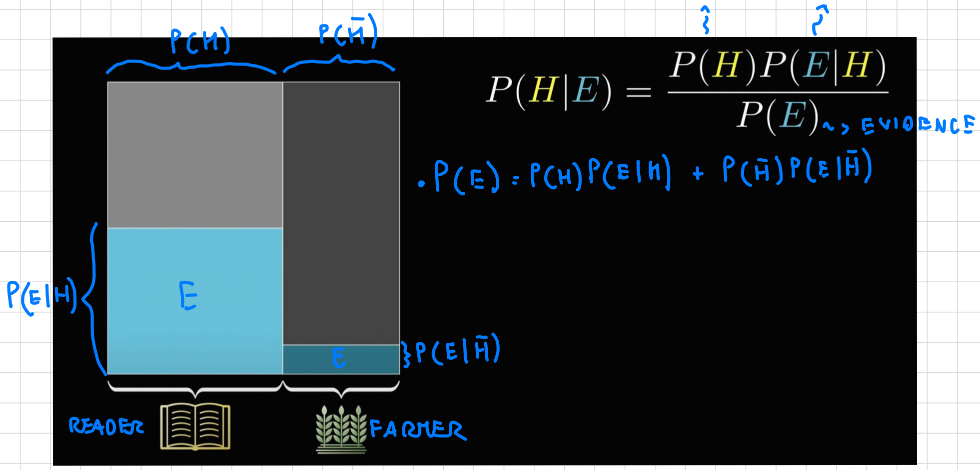
$$\rightarrow P(B) = P(B|A_1)P(A_1) + \dots + P(B|A_n)P(A_n)$$

- **BAYES RULE:**

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B)} \quad / \quad P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

"HE IS A READER"

"REAL EVIDENCE, KNOWING HE IS A READER"



RANDOM VARIABLES.

THE EXPECTED VALUE OF X AND A FUNCTION :

LET X BE A DISCRETE R.V., SET OF POSSIBLE VALUES D AND PMF $p_X(x)$:

$$\rightarrow E[X] = \mu_x = \sum_{x \in D} x \cdot p_X(x)$$

$$\rightarrow E[h(X)] = \sum_D h(x) \cdot p(x)$$

- PROPERTIES :
  - IF $h(X) = aX + b \rightarrow E[h(X)] = a E[X] + b$
  - $E[y_1(X) + \ldots + y_n(X)] = E[y_1(X)] + \ldots + E[y_n(X)]$

THE VARIANCE OF X ; $\rightsquigarrow$ ANALOG. FOR $h(X)$

LET X HAVE A PMF $p_X(x)$ AND EXPECTED VALUE μ :

$$\rightarrow V[X] = \sigma_X^2 = \sum_D (x-\mu)^2 \cdot p(x) = E[(x-\mu)^2]$$

- $\sigma_X = \sqrt{V[X]}$: STANDARD DEVIATION
- $V[X] = E[X^2] - E^2[X] = \mu_2 - \mu^2$

es. 24

- A library has an upper limit of 6 on the number of videos that can be checked out to an individual at one time. Consider only those who check out videos, and let X denote the number of videos checked out to a randomly selected individual. The pmf of X is as follows:

x	1	2	3	4	5	6
$p(x)$	.30	.25	.15	.05	.10	.15

- $E[X] = \sum_1^6 x \cdot p(x) = 1 \cdot 0{,}30 + \ldots + 6 \cdot 0{,}15 =$
  $= 2{,}85$

$\rightarrow V[X] = E[(X-\mu)^2] =$

$= \sum_1^6 (x - 2{,}85)^2 \cdot p(x) = 3{,}2275$

$\rightarrow \sigma = \sqrt{V[X]} = 3{,}2275$

- **PROPERTIES:**

  - LET $h(X) = aX + b \rightarrow V[h(X)] = a^2 V[X]$

    $\rightarrow \sigma_{h(X)} = |a| \sigma_X$

es. 26

- $\mu = 2$, $\mu_2 = 5$

  $\rightarrow V[X] = 5 - 2^2 = 1$

- $h(X) = 800X - 900$

  $\rightarrow V[h(X)] = 800^2 \cdot V[X]$

- **CHEBYSEV'S INEQUALITY**. ($\rightsquigarrow$ SEE PART. 2 FOR BETTER CLARITY)

LET X BE A R.V. / $E[X] = \mu$, $V[X] = \sigma^2$

$$\longrightarrow P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad k \geq 1$$

COROLLARY

$$\longrightarrow P(|X - \mu| \leq k) \geq 1 - \frac{V^{\alpha}[X]}{k^{\alpha}}$$

(DIM.)

- $P(|X| > a) \leq \dfrac{E[|X|^k]}{a^k}$

  - IF $X = X - \mu$, $a = k\sigma$

    $\alpha = 2$

    $$\rightarrow P(|X - \mu| > k\sigma) \leq \frac{E[|X-\mu|^2]}{k^2 \sigma^2} = \frac{V[X]}{k^2 \sigma^2} = \frac{1}{k^2}$$

• DISCRETE RANDOM VARIABLES, PART 1:

LET X BE A R.V., S = 'SUCCESS', F = 'FAILURE'

• BERNOULLI DISTRIBUTION:

LET Y HAVE 1 OF 2 OUTCOMES: $\overset{1}{S}$ (PROBABILITY p) OR $\overset{0}{F}$ (PROBABILITY $1-p$):

$\rightarrow p(Y) = \begin{cases} p &, Y=1 \\ 1-p &, Y=0 \end{cases}$ ⇝ CAN ALSO BE WRITTEN AS $p(x, \alpha)$, WHERE $\alpha = p(Y=1)$

• $E[Y] = p$, $V[Y] = p(1-p)$

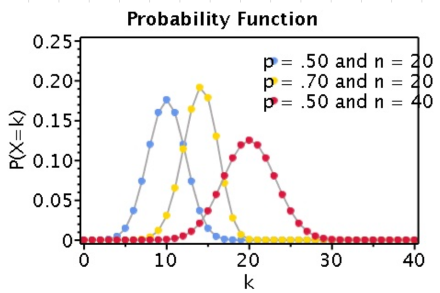
• BINOMIAL DISTRIBUTION:

⇝ $\overset{\text{FIXED}}{n}$ TRIALS → #S? (WITH REPLACEMENT)

EXPERIMENT: SERIES OF n TRIALS / EACH TRIAL $X \sim$ BERNOULLI (p) $\begin{matrix} \nearrow S \\ \searrow F \end{matrix}$

• X: n° OF S IN n TRIALS

• $X \sim Bin(n, p)$, PMF: $b(x; n, p)$



→

LET $X \sim Bin(n, p)$:

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

• $E[X] = np$, $V[X] = np(1-p)$

es. $X \sim Bin(n=3, p)$

• $P(X=2) = P(SSF \cup SFS \cup FSS) = \binom{3}{2} p^2 (1-p)$

• $P(X=1) = P(SFF \cup FSF \cup FFS) = \binom{3}{1} p (1-p)^2$

es. PAG. 17 ⇝ WE CAN USE TABLES : 'BINOMIAL cdf'

Suppose that 20% of all copies of a particular textbook fail a certain binding strength test. Let X denote the number among 15 randomly selected copies that fail the test.

Then X has a binomial distribution with n=15 and p=0.2

1. The probability that at most 8 fail the test is…
2. The probability that exactly 8 fail is…
3. The probability that at least 8 fail is…
4. The probability that between 4 and 7, inclusive, fail is…

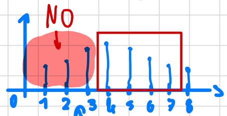
$p = 0{,}2 \quad , \quad n = 15$

→

1. $P(X \leq 8) = \overset{*}{\dots} = 0{,}9992$

2. $P(X = 8) = \dots = F_X(8) - F_X(7) = 0{,}0034$ *

3. $P(X \geq 8) = 1 - P(X < 8) = 1 - P(X \leq 7) =$



$= 1 - F_X(7) = 0{,}0042$

4. $P(4 \leq X \leq 7) = F_X(7) - F_X(3) = 0{,}3476$

n	x	p = 0,05	0,10	0,15	0,20	0,25
15	0	0.4633	0.2059	0.0874	0.0352	0.0134
	1	0.8290	0.5490	0.3186	0.1671	0.0802
	2	0.9638	0.8159	0.6042	0.3980	0.2361
	3	0.9945	0.9444	0.8227	0.6482	0.4613
	4	0.9994	0.9873	0.9383	0.8358	0.6865
	5	0.9999	0.9978	0.9832	0.9389	0.8516
	6	1.0000	0.9997	0.9964	0.9819	0.9434
	7	1.0000	1.0000	0.9994	0.9958	0.9827
	8	1.0000	1.0000	0.9999	0.9992	0.9958
	9	1.0000	1.0000	1.0000	0.9999	0.9992
	10	1.0000	1.0000	1.0000	1.0000	0.9999
	11	1.0000	1.0000	1.0000	1.0000	1.0000
	12	1.0000	1.0000	1.0000	1.0000	1.0000
	13	1.0000	1.0000	1.0000	1.0000	1.0000
	14	1.0000	1.0000	1.0000	1.0000	1.0000

- **HYPERGEOMETRIC DISTRIBUTION:** ⇝ GIVEN POPULATION OF N ELEMENT, I KNOW M SUCCESSES
  ↳ SELECTED $n < N$ ELEMENTS : $X = k$ SUCCESSES ?

POPULATION OF N ELEMENTS / EACH ONE S OR F

M SUCCESS IN THE POPULATION , SET OF n ELEMENTS SELECTED (WITH NO REPLACEMENTS)

- X : n° OF S IN SAMPLE , PMF : $h(x; n; M; N)$

→

LET $X \sim$ HYPERGEOMETRIC (n, M, N) :

$$\to \quad P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

- $E[X] = n \cdot \dfrac{M}{N} \quad , \quad V[X] = \left(\dfrac{N-n}{N-1}\right) \cdot n \cdot \dfrac{M}{N}\left(1 - \dfrac{M}{N}\right)$

→ M/N IS THE PROPORTION OF S IN THE POPULATION, IF $p = \dfrac{M}{N}$

$$\to \quad E[X] = \overbrace{np}^{E[X]\text{ BINOMIAL}} \quad , \quad V[X] = \underbrace{\left(\frac{N-n}{N-1}\right)}_{\text{FINITE POPULATION CORRECTION FACTOR}} \cdot \overbrace{np(1-p)}^{V[X]\text{ BINOMIAL}}$$

- **N.B.:** IN THE BINOMIAL THERE'S NOT AN INITIAL POPULATION OF N
  / $n < M + (N - M) = N$

es. 35

- During a particular period a university's information technology office received 20 service orders for problems with printers, of which 8 were laser printers and 12 were inkjet models. A sample of 5 of these service orders is to be selected for inclusion in a customer satisfaction survey.

- Suppose that the 5 are selected in a completely random fashion, so that any particular subset of size 5 has the same chance of being selected as does any other subset. What then is the probability that exactly x ($x = 0, 1, 2, 3, 4$, or 5) of the selected service orders were for inkjet printers?

$N = 20$, $n = 5$

$\underset{S}{M = 12} \rightarrow N - M = 8 \; F$

$$\cdot\ P(X=2) = h(\overset{K}{2}; 5; 12; 20) = \frac{\binom{M}{K}\binom{N-M}{n-K}}{\binom{n}{N}} = \frac{\binom{12}{2}\binom{20-12}{5-2}}{\binom{20}{5}} = \frac{77}{323} = 0{,}238$$

- IF:

  - $n > \underbrace{N-M}_{\text{FAILURES IN POPULATION}} \rightarrow$ SMALLEST $\overset{n^\circ S}{X} = n-(N-M)$

  - $n > M \rightarrow$ HIGHEST $X = M$

$\rightarrow$ VALUES OF X : $\max(0;\ n-(N-M)) \leq X \leq \min(n;M)$

- SUPPOSE WE DON'T KNOW N, WE WISH TO ESTIMATE IT :

$\rightarrow$ APPROXIMATING $\dfrac{K}{n} \approx \dfrac{M}{N} \rightarrow \hat{N} = \dfrac{n \cdot M}{K}$

- IF $\dfrac{n}{N} < 5\% \rightarrow h() \approx Bin()$

- NEGATIVE BINOMIAL DISTRIBUTION: ~> r FIXED SUCCESSES -> # F ?
↳ DIFFERENT FROM BINOMIAL

EXPERIMENT CONTINUES UNTIL A TOTAL OF r SUCCESSES HAVE BEEN OBSERVED

- X : n° F THAT PRECEDES r^{TH} S

-> LET $X \sim nb(r, p)$

$$\rightarrow P(X=k) = \binom{k+r-1}{r-1} p^r (1-p)^k \quad , \text{PMF} = nb(k; r, p)$$

- $E[X] = \dfrac{r(1-p)}{p}$, $V[X] = \dfrac{r(1-p)}{p^2}$

- IF $r=1$ —> GEOMETRIC DISTRIBUTION: $P(X=k) = p(1-p)^k$
↳ AT LEAST 1 SUCCESS

es. 38

- A pediatrician wishes to recruit 5 couples, each of whom is expecting their first child, to participate in a new natural childbirth regimen. Let $p = P$(a randomly selected couple agrees to participate).
- If $p = .2$, what is the probability that 15 couples must be asked before 5 are found who agree to participate? That is, with $S = \{\text{agrees to participate}\}$, what is the probability that 10 F's occur before the fifth S?

- n° OF SUCCESSES: $r = 5$
- n° OF F BEFORE r^{TH} SUCCESS: $k = 10$

$$\rightarrow P(X=10) = nb(k=10, r=5, p=0{,}2) =$$

$$= \binom{14}{4} (0{,}2)^5 (1-0{,}2)^{10} = 0{,}034$$

- SUMMARY:
  - $X \sim Bin(n, p)$, $P(X=k)$: k SUCCESSES (S with p) IN n TRIALS
  - $X \sim h(n, M, N)$, $P(X=k)$: k SUCCESSES IN SUBSET OF LENGTH n IN A POPULATION OF N ELEMENTS / M SUCCESSES
    ↳ IF $\frac{n}{N} < 5\%$ -> $h() \approx Bin()$
  - $X \sim nb(r, p)$, $P(X=k)$: k FAILURES BEFORE r^{TH} SUCCESS
    ↳ $r=1$ -> $X \sim \text{GEOMETRIC}(p)$, $P(X=k)$: k FAILURE BEFORE 1 SUCCESS

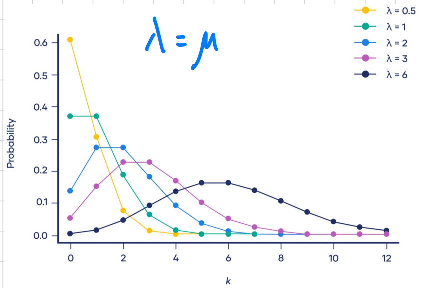
DISCRETE RANDOM VARIABLES, PART 2 :

• DISCRETE UNIFORM DISTRIBUTION.

$$f_X(x) = f_X(x;N) = \begin{cases} 1/N & , \ x = 1, 2, \dots \\ 0 & , \ \text{OTHERWISE} \end{cases}$$

- $E[X] = \dfrac{N+1}{2}$, $V[X] = \dfrac{N^2-1}{12}$

• POISSON DISTRIBUTION :



$X \sim Poiss(\mu)$ ~> μ ∵ IT'S THE $E[X]$

$$\to p(x;\mu) = \frac{\mu^x}{x!} e^{-\mu} , \quad x \in [0;\infty)$$

- DERIVED FROM McLAURIN SERIES EXPANSION ; $e^{\mu} = \sum_{x=0}^{\infty} \dfrac{\mu^x}{x!} = 1 + \mu + \dfrac{\mu^2}{2} + \dots$

ex. 34

- Let X denote the number of creatures of a particular type captured in a trap during a given time period.
- Suppose that X has a Poisson distribution with $\mu = 4.5$, so on average traps will contain 4.5 creatures.

- $P(X = 5) = \dfrac{4{,}5^5}{5!} e^{-4{,}5} = 0{,}178$

  X = n° OF CREATURES CAPTURED

- $P(X \le 5) = \sum_{x=0}^{5} \dfrac{4{,}5^x}{x!} e^{-4{,}5} = 0{,}7029$

• POISSON DISTRIBUTION AS A LIMIT :

SUPPOSE A BINOMIAL PMF $b(x;n,p) \ / \ n \mapsto \infty, \ p \mapsto 0$, SUCH THAT $np \mapsto \mu$

$\to b(x;n,p) \longmapsto p(x;\mu)$, FOR ANY BINOMIAL EXPERIMENT WITH $n >> 0$, $p << n$

- ↳ THIS APPROXIMATION IS SAFE FOR $(n > 50)$ AND $(np < 5)$

es. 40

- If a publisher of nontechnical books takes great pains to ensure that its books are free of typographical errors, so that the probability of any given page containing at least one such error is .005 and errors are independent from page to page, what is the probability that one of its 400-page novels will contain exactly one page with errors? At most three pages with errors?

S: PAGE WITH AT LEAST 1 ERROR

$\rightarrow X \sim Bin\,(n = 400,\ p = 0{,}005)$

- $np = 2\ ,\ n > 50$

$\rightarrow P(X=1) = b(1;\ 400,\ 0{,}005) \approx p(1;2) = \frac{2^1}{1!} e^{-2} = 0{,}27$

$\rightarrow P(X \leq 3) \approx \sum_{x=0}^{3} p(x,2) = \sum_{x=0}^{3} \frac{2^x}{x!} e^{-2} = 0{,}857$

- $\uparrow n \longrightarrow \uparrow$ ACCURACY OF THE APPROXIMATION

- MEAN AND VARIANCE:

  - $E[X]_{,\,BINOMIAL} = np$
    $V[X]_{,\,BINOMIAL} = np(1-p)$
    $\rightarrow$ IF $Bin \approx p \rightarrow \begin{cases} np \mapsto \mu \\ np(1-p) \mapsto \mu \end{cases}$

$\rightarrow E[X] = V[X] = \mu$

- THE POISSON PROCESS:

1. $\exists\, \alpha > 0 \,/\, \forall\, \Delta t$ ~> TIME INTERVAL $\rightarrow P_{1\ \text{EVENT OCCURRES}} = \alpha \cdot \Delta t + o(\Delta t)$
2. $P_{\text{MORE }1\text{ EVENT}}$ IN $\Delta t = o(\Delta t)$ ~> TRASCURABILE

$\rightarrow p$ 1 EVENT OCCURRES IN Δt IS PROPORTIONAL TO Δt : $p = \alpha \Delta t$

- LET $P_k(t) = p$ THAT k EVENTS IN t TIME

$\rightarrow P_k(t) = \frac{(\alpha t)^k}{k!} e^{-\alpha t} \quad / \quad \mu = \alpha t$

es. 42

- Suppose pulses arrive at a counter at an average rate of six per minute, so that $\alpha = 6$.

$\hookrightarrow P_{1 \text{ PULSE}}$ IN $\Delta t = 0,5$ min ?

$\rightarrow \mu = \alpha t = 6 \cdot 0,5 = 3$

- LET X = n° OF PULSES IN Δt :

$$\rightarrow P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - \frac{3^0}{0!} e^{-3} = 0,95$$

- CONTINOUS RANDOM VARIABLES, PART 1 :

- PROBABILITY DISTRIBUTION :

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

SUBSET

$$\to P(X \in \tilde{B}) = \int_B f(x)$$

PROBABILITY DISTRIBUTION FUNCTION

- $f(x)$ IS A PDF $\iff \begin{cases} f(x) \ge 0, \ \forall x \\ \int_{-\infty}^{\infty} f(x)\,dx = 1 \end{cases}$

- UNIFORM DISTRIBUTION :

R.V. X, ON INTERVAL $[A, B]$:

$$f(x; A, B) = \begin{cases} \frac{1}{B-A} & , A \le x \le B \\ 0 & , \text{OTHERWISE} \end{cases}$$

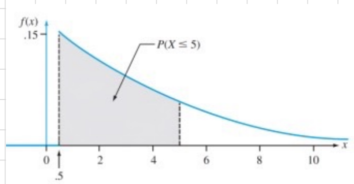
- N.B. : $P(X = c) = \int_c^c f(x)\,dx = \lim_{\varepsilon \to 0} \int_{c-\varepsilon}^{c+\varepsilon} f(x)\,dx = 0$

$\hookrightarrow$ IF X CONTINUOUS

$$\to P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b)$$

es, 5

$$f(x) = \begin{cases} 0.15 \cdot e^{-0.15(x-0.5)} & , x \ge 0.5 \\ 0 & , 0 \end{cases}$$



$$\to P(X \le 5) = \int_{-\infty}^{5} f(x)\,dx = \int_{0.5}^{5} 0.15\, e^{-0.15(x-0.5)}\,dx =$$

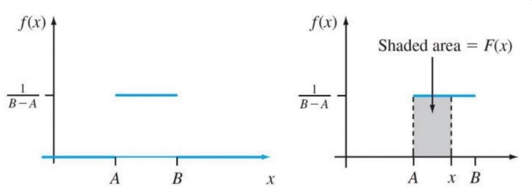
$$= 0.15\, e^{0.075} \left(-\frac{1}{0.15} e^{-0.15x} \right) \Bigg|_{x=0.5}^{x=5} = 0.491$$

• CUMULATIVE DISTRIBUTION FUNCTION :

CDF $F(x)$, FOR A CONTINUOUS R.V. X :

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

es. 6



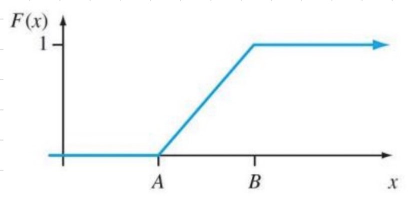
FROM THE FIGURE :

- $x < A \rightarrow F(x) = 0$
- $x \geq B \rightarrow F(x) = 1$
- $A \leq x \leq B$:

$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \int_{A}^{x} \frac{1}{B-A}\,dy = \frac{1}{B-A}\, y \Big|_{y=A}^{y=x} = \frac{x-A}{B-A}$$

$\rightarrow$ CDF $F(x)$:

$$\begin{cases} 0 & , \; x < A \\ \dfrac{x-A}{B-A} & , \; A \leq x < B \\ 1 & , \; x \geq B \end{cases}$$



- **USING $F(X)$ TO COMPUTE PROBABILITIES:**

  - PROPERTIES:

    - $P(X > a) = 1 - F(a)$
    - $P(a \leq X \leq b) = F(b) - F(a)$

es. 7

$$f(x) = \begin{cases} \frac{1}{8} + \frac{3}{8}x & , 0 \leq x \leq 2 \\ 0 & , \text{OTHERWISE} \end{cases}$$

$\rightarrow$

$$F(x) = \int_{-\infty}^{x} f(r)\, dr = \int_{0}^{x} \frac{1}{8} + \frac{3}{8}r \; dr = \frac{1}{8}r + \frac{3}{8}r^{2} \Big|_{0}^{x} = \frac{1}{8}x + \frac{3}{8}x^{2}$$

$$\rightarrow F(X) = \begin{cases} 0 & , X < 0 \\ \frac{x}{8} + \frac{3}{8}x^{2} & , 0 \leq x \leq 2 \\ 1 & , 2 < X \end{cases}$$

- $P(1 \leq X \leq 1.5) = F(1.5) - F(1)$
- $P(X > 1) = 1 - P(X \leq 1) = 1 - F(1)$

- **$f(x)$ FROM $F(x)$:**

$$\rightarrow f(x) = \frac{d}{dx} F(x)$$

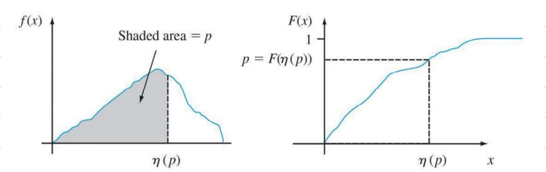
es. 6^{*}

$$f(x) = \frac{d}{dx} F(x) = \frac{d}{dx}\left(\frac{x - A}{B - A}\right) = \frac{1}{B - A}$$

- **PERCENTILES OF CONTINUOUS DISTRIBUTION:**

- LET $p \,/\, 0 < p < 1$, $\eta(p) = (100p)^{TH}$ PERCENTILE OF R.V. X:

$$\to \; p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y)\, dy$$



e.g.

$$f(x) = \begin{cases} \frac{3}{2}(1-x^2) &, 0 \le x \le 1 \\ 0 &, \text{OTHERWISE} \end{cases} \;\to\; F(x) = \dots = \frac{3}{2}\left(x - \frac{1}{3}x^3\right)$$

$$\to \quad p = F(\eta(p)) = \frac{3}{2}\left[\eta(p) - \frac{1}{3}\eta^3(p)\right] = \eta^3(p) - 3\eta(p) + 2p = 0$$

- 50% : $p = 0.5 \to \eta^3(p) - 3\eta(p) + 2 \cdot 0.5 \to \eta(p = 0.5) = 0.347$

- **MEDIAN** : $\tilde{\mu} \,/\, F(\tilde{\mu}) = 50\%$

- **EXPECTED VALUES:**

LET X BE A R.V. CONTINUOUS, WITH PDF $f(x)$:

$$\to \quad \mu_X = E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

$$\to \quad E[h(X)] = \mu_{h(X)} = \int_{-\infty}^{\infty} h(x) \cdot f(x)\, dx$$

- **VARIANCE:**

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx = E[(X-\mu)^2]$$

$$\hookrightarrow \sigma_X = \sqrt{V(X)}$$

- **N.B.**: SAME PROPERTIES OF DISCRETE X HOLD FOR CONTINUOUS X

- CONTINOUS RANDOM VARIABLES, PART 2.

- THE NORMAL DISTRIBUTION:

LET X BE A CONTINUOUS R.V.:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty$$

-> WHERE $E(X) = \mu$, $V(X) = \sigma^2$

- IF X HAS A NORMAL DISTRIBUTION: $X \sim N(\mu, \sigma^2)$

- STANDARD NORMAL DISTRIBUTION:

- IF $\mu = 0$, $\sigma = 1$ -> $f(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

-> $X \sim Z(0,1) = X \sim N(\mu = 0, \sigma^2 = 1)$

- $P(Z \leq z) = \phi(z)$ ~> SEE TABLES FOR VALUES
- $P(Z \geq z) = P(Z \leq -z) = \phi(-z)$
- $P(a \leq Z \leq b) = \phi(b) - \phi(a)$

- PERCENTILES OF THE STANDARD NORMAL DISTRIBUTION:

- $\eta(p) = p$

- Z_α NOTATION FOR Z CRITICAL VALUES:

Z_α IS THE VALUE OF z / AREA α, FOR $z \geq z_\alpha$

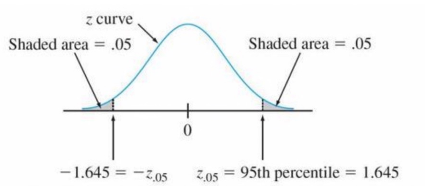
-> $P(Z \geq z_\alpha) = \alpha$

-> $z_\alpha = 100(1-\alpha)^{TH}$ PERCENTILE OF STANDARD NORMAL DISTRIBUTION

ES. 4.15

- $z_{0,05} = 1.645$

↳ 95% OF THE DISTRIBUTION



- **NON-STANDARD NORMAL DISTRIBUTIONS:**

IF $X \sim N(\mu, \sigma^2)$, PROBABILITIES ARE COMPUTING STANDARDIZING X:

$$\rightarrow X \xrightarrow{\text{STANDARDIZATION}} \left(\frac{X-\mu}{\sigma}\right) \rightarrow \mu = 0,\ \sigma^2 = 1$$

$$\begin{cases} \cdot\ N \mapsto Z:\ Z = \dfrac{X-\mu}{\sigma} \\ \cdot\ Z \mapsto N:\ X = \sigma \cdot Z + \mu \end{cases}$$

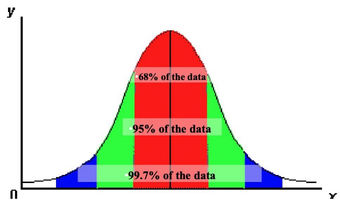
$$\cdot\ P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq X \leq \frac{b-\mu}{\sigma}\right) =$$

$$= \phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)$$

$$\rightarrow P(Z \leq z) = P(X \leq \sigma \cdot z + \mu)$$

Useful Information about the Normal Distribution:

~68% of a normal population is within $\pm 1\sigma$ of μ
~95% of a normal population is within $\pm 2\sigma$ of μ
~99.7% of a normal population is within $\pm 3\sigma$ of μ



es. 2.16

$$X \sim N(\mu = 1.25,\ \sigma^2 = 0.46^2)$$

$$\rightarrow \cdot\ P(1 \leq X \leq 1.75) = P\left(\frac{1-1.25}{0.46} \leq Z \leq \frac{1.75-1.25}{0.46}\right) =$$

$$= \phi(1.09) - \phi(-0.54)$$

$$\cdot\ P(X > 2) = P\left(Z > \frac{2-1.25}{0.46}\right) = P(Z > 1.63) = 1 - \phi(1.63)$$

- **PERCENTILES OF AN ARBITRARY NORMAL DISTRIBUTION:**

$$\rightarrow (100p)^{TH} \text{ PERCENTILE FOR } N(\mu, \sigma^2) = \mu + \left[(100p)^{TH} \text{ PERCENTILE FOR } Z\right] \sigma$$

es.

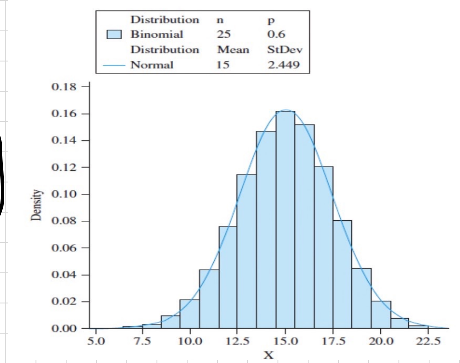
IF $\eta_Z(99.5\%) = 2.58$, AND FOR THE NORMAL DISTRIB.: $\mu = 5.496,\ \sigma = 0.067$

$$\rightarrow c = \eta(99.5\%) = 5.496 + 2.58 \cdot 0.067 = 5.669$$

- APPROXIMATING Bin DISTRIB. WITH N:

LET $X \sim Bin(x; n, p)$ / $\mu_X = np$, $\sigma_X = \sqrt{np(1-p)}$

IF $np \geq 10$ AND $n(1-p) \geq 10$ -> VALID APPROX.



- -> $P(X \leq x) = Bin(x; n, p) \approx \phi\left(\frac{x + 0.5 - \mu_x}{\sigma_x}\right)$

es. 6.20

- 25 % STUDENT -> FINANCIAL AID
- $X \sim Bin(x; n = 50, p = 0{,}25)$ -> $\mu = 12.5$, $\sigma = 3{,}06$

->

$np = 12.5$, $np(1-p) = 37.5$ -> APPROXIMATION IS VALID

-> $P(X \leq 10) = Bin(10; 50, 0.25) \approx \phi\left(\frac{10 + 0.5 - 12.5}{3.06}\right) =$

$= \phi(-0.65) = 0.2578$

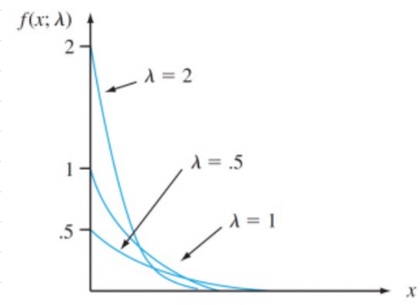
-> $P(5 \leq X \leq 15) = Bin(15; 50, 0.25) - Bin(5-1; 50, 0.25)$

• CONTINOUS RANDOM VARIABLES, PART 3:

• THE EXPONENTIAL DISTRIBUTION:

GIVEN A CONTINOUS R.V. X, PARAMETER λ, PDF:

$$\to f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{OTHERWISE} \end{cases}$$



- $E(X) = \int_0^\infty x \lambda e^{-\lambda x} \, dx = \frac{1}{\lambda}$
- $V(X) = E(X^2) - E^2(X) = \frac{1}{\lambda^2}$
- CDF: $F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$

• EXPONENTIAL DISTRIBUTION IS RELATED TO POISSON DISTRIBUTION: $\lambda = \alpha$

$$\to P(X_1 \leq t) = 1 - \underbrace{P(X_1 > t)}_{\text{P NO EVENTS IN } (0,t)} = 1 - \underbrace{\frac{e^{-\alpha t}(\alpha t)^0}{0!}}_{\text{POISSON PROCESS, } k=0} = \underbrace{1 - e^{-\alpha t}}_{\text{CDF EXPONENTIAL DISTRIB, } \lambda = \alpha}$$

es. 4.22

$\alpha = 0.5$

$$\to P(X > 2) = 1 - P(X \leq 2) = 1 - F(2; 0.5) = e^{-\lambda x} = e^{-0.5 \cdot 2}$$

• MEMORYLESS PROPERTY:

SUPPOSE A COMPONENT WITH AGE t_0, WE WANT TO KNOW THE PROBABILITY THAT IT LASTS AT LEAST t:

$$P(X > t + t_0 \mid X > t_0) \overset{\text{COND. PROB.}}{=} \frac{P[(X \geq t + t_0) \cap (X > t_0)]}{P(X \geq t_0)} \overset{X > t_0 \text{ REDUNDANT}}{=} \frac{P[(X \geq t + t_0)]}{P(X \geq t_0)} = \frac{1 - F(t + t_0, \lambda)}{1 - F(t_0, \lambda)} = e^{-\lambda t}$$

- DISTRIBUTION OF REMAINING LIFETIME IS INDIPENDENT FROM AGE
- MORE GENERAL LIFETIME MODELS: WEIBULL, GAMMA

$\to$ MEMORYLESS PROPERTY: $P(X > t + t_0 \mid X > t_0) = P(X > t)$

• GAMMA DISTRIBUTION: (NO EXAM)

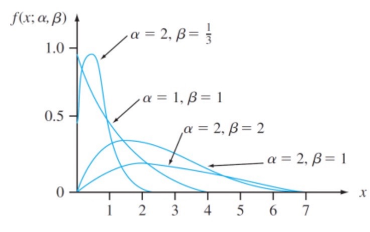
• GAMMA FUNCTION:

LET $\alpha > 0$: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$

• PROPERTIES:

- $\forall \alpha > 1$: $\Gamma(\alpha) = (\alpha-1)\cdot\Gamma(\alpha-1)$
- $\forall n > 0, n \in \mathbb{Z}$: $\Gamma(n) = (n-1)!$
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

• PDF: $f(x;\alpha) = \begin{cases} \dfrac{x^{\alpha-1} e^{-x}}{\Gamma(\alpha)} & , x \geq 0 \\ 0 & , \text{OTHERWISE} \end{cases}$



Gamma density curves

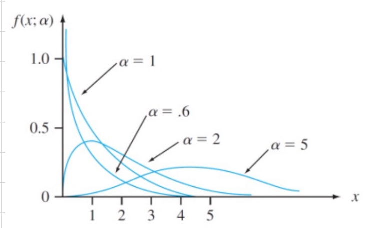
$\rightarrow$

LET X BE A CONTINUOUS R.V.:

$\rightarrow f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} & , x \geq 0 \\ 0 & , \text{OTHERWISE} \end{cases}$

↳ β: SCALE PARAMETER

↳ STRETCHES OR COMPRESS



standard gamma density curves

- IF $\beta = 1 \rightarrow$ STANDARD GAMMA DISTRIBUTION
- IF $\alpha = 1, \beta = \frac{1}{\lambda} \rightarrow$ EXPONENTIAL DISTRIBUTION

• $E(X) = \mu = \alpha\beta$, $V(X) = \sigma^2 = \alpha\beta^2$

• CDF (X: STANDARD GAMMA R.V.): $F(x;\alpha) = \int_0^x \dfrac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}\, dy$, $x > 0$

• IF NON-STANDARD DISTRIBUTION:

$$P(X \leq x) = F(x;\alpha,\beta) = F\left(\frac{x}{\beta};\alpha\right)$$

es. 4.24

X : SURVIVAL TIME IN WEEKS OF A RANDOMLY SELECTED MOUSE EXPOSED TO RADIATIONS

- $\alpha = 8$, $\beta = 15$

->

- $E(X) = \alpha\beta = 120$ WEEKS .
- $V(X) = \alpha\beta^2 = 1800$
- $P(60 \leq X \leq 120) = P(X \leq 120) - P(X \leq 60) =$

$\exists \beta$ -> NON-STANDARD DISTRIBUTION (~) $= F\left(\frac{120}{15}; 8\right) - F\left(\frac{60}{15}; 8\right) = \dots = 0.496$

- $P(X \geq 30) = 1 - P(X < 30) = 1 - F\left(\frac{30}{15}; 8\right) = 0.999$

- **CHI-SQUARED DISTRIBUTION:** (NO EXAM, ONLY FROM TABLES)

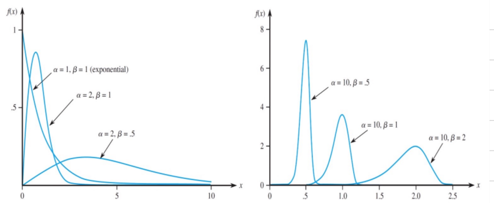
LET $X \sim \gamma(x; \alpha; \beta)$, $v \in \mathbb{N}, v > 0 \ / \ \alpha = \frac{1}{2} v, \ \beta = 2$

$\rightarrow \chi^2$ DISTRIBUTION:
$$f(x; v) = \begin{cases} \frac{1}{2^{v/2}\, \Gamma\left(\frac{v}{2}\right)} x^{\left(\frac{v}{2}\right)-1} e^{-\frac{x}{2}} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$

- v : n° OF DEGREE OF FREEDOM

- **WEIBULL DISTRIBUTION:** (NO EXAM)

LET X BE A CONTINUOUS R.V., PARAMETERS α, β (SCALE PARAMETER):
$$f(x; \alpha, \beta) = \begin{cases} \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$



- IF $\alpha = 1 \rightarrow$ EXPONENTIAL DISTRIBUTION, $\lambda = 1/\beta$

- $E(X) = \mu = \beta\, \Gamma\left(1 + \frac{1}{\alpha}\right)$

- $V(X) = \sigma^2 = \beta^2 \left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right) \right]$

- CDF: $$F(x; \alpha, \beta) = \begin{cases} 1 - e^{-\left(\frac{x}{\beta}\right)^{\alpha}} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$

es. 4.25

X : AMOUNT OF NO_x EMISSION FROM ENGINE

$X \sim \text{WEIBULL}(x; \alpha = 2, \beta = 10)$

$\rightarrow$
- $P(X \leq 10) = F(10; 2, 10) = 1 - e^{-\left(\frac{10}{10}\right)^2}$

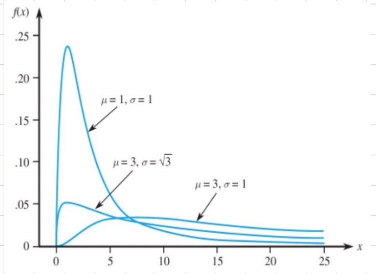
- c / 5% OF ENGINES WITH LARGEST NO_x EMISSIONS:

$$0.95 = 1 - e^{-\left(\frac{c}{10}\right)^2} \rightarrow c \approx 17.3$$

• LOGNORMAL DISTRIBUTION: (NO EXAM)

LET $Y = \ln(X) \,/\, Y \sim N(\mu, \sigma^2) \to X \sim$ LOGNORMAL

$$\to f(x;\mu,\sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\,\sigma x}\, e^{-[\ln(x)-\mu]^2 \cdot \frac{1}{(2\sigma)^2}} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$



• N.B.: $E(\ln(X)) = \mu \neq E(X)$, $V(\ln(X)) = \sigma^2 \neq V(X)$

• $E(X) = e^{\mu + \frac{\sigma^2}{2}}$, $V(X) = e^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right)$

• CDF: $F(x;\mu,\sigma) = P(X \leq x) = P\left(\ln(X) \leq \ln(x)\right) =$

$$= P\left(Z \leq \frac{\ln(x) - \mu}{\sigma}\right) = \phi\left(\frac{\ln(x) - \mu}{\sigma}\right)$$

ex.

$\mu = 0.353$, $\sigma = 0.754$

$\to$

• $E(X) = e^{0.353 + 0.754^2 \cdot \frac{1}{2}} = 1.891$

• $V(X) = e^{2(0.353) + 0.754^2}\left(e^{0.754^2} - 1\right) = 2.7387$

• $P(1 \leq X \leq 2) = P\left(\ln(1) \leq \ln(X) \leq \ln(2)\right) = P\left(0 \leq \ln(X) \leq 0.693\right) =$

$$= P\left(\frac{0 - 0.353}{0.754} \leq Z \leq \frac{0.693 - 0.353}{0.754}\right) = \phi(0.47) - \phi(-0.45) = 0.354$$

• $99\% = P(X \leq c) = P\Bigg(Z \leq \underbrace{\frac{\ln(c) - 0.353}{0.754}}_{z_{0.1}}\Bigg) \,/\, z_{0.1} = 2.33$

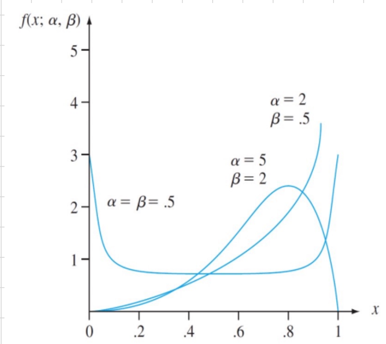
$\to \ln(c) = 2.1098$, $c = 8.247$

• BETA DISTRIBUTION : (NO EXAM, ONLY FOR BAYESIAN STATISTICS PART)

LET $\alpha > 0$, $\beta > 0$, A, B, PDF OF X:

$$\rightarrow f(x;\alpha,\beta,A,B) = \begin{cases} \dfrac{1}{B-A} \cdot \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\dfrac{x-A}{B-A}\right)^{\alpha-1} \left(\dfrac{B-x}{B-A}\right)^{\beta-1} & , A \le x \le B \\ 0 & , \text{OTHERWISE} \end{cases}$$

- IF $A = 0$, $B = 1$ $\rightarrow$ STANDARD BETA DISTRIBUTION



- $E(X) = \mu = A + (B-A)\dfrac{\alpha}{\alpha+\beta}$
- $V(X) = \dfrac{(B-A)^2 \alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

es. 4.28

- X : TIME IN DAYS
- $A = 2$, $B = 5$, $\alpha = 2$, $\beta = 3$

$\rightarrow$

- $E(X) = 2 + 3 \cdot \dfrac{\alpha}{\alpha+\beta} = 2 + 3 \cdot 0.4 = 3.2$
- $P(X \le 3) = \displaystyle\int_2^3 \frac{1}{3} \cdot \frac{4!}{1!\,2!} \left(\frac{x-2}{3}\right)\left(\frac{5-x}{3}\right)^2 dx = \frac{4}{27}\int_2^3 (x-2)(5-x)^2\,dx$

$$= \frac{4}{27} \cdot \frac{11}{4} = \frac{11}{27} = 0.407$$

JOINT R.V., PART. 1:

2 DISCRETE RV:

LET X, Y BE 2 R.V. DEFINED ON THE SAME SPACE, THE JOINT PROBABILITY MASS FUNCTION $p(x,y)$:

$$p(x,y) = P(X=x, Y=y) \quad / \quad p(x,y) \geq 0 \qquad \sum_x \sum_y p(x,y) = 1$$

- LET $A = \{(x,y) : f(x,y)\}$ → $P[(X,Y) \in A] = \sum_{(x,y) \in A} \sum p(x,y)$

es. $x+y=5$

- MARGINAL PROBABILITY MASS FUNCTION:
  - X: $p_X(x) = \sum_y p(x,y)$
  - Y: $p_Y(y) = \sum_x p(x,y)$

es.

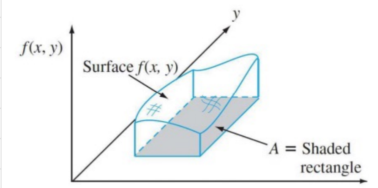
- $p_X(100) = p(100,0) + p(100,100) + p(100,200) =$
  $= 0.2 + 0.1 + 0.2$
- $p_Y(200) = 0.2 + 0.3$

$p(x, y)$		y: 0	100	200	$p_X(x)$
x	100	.20	.10	.20	0.5
	250	.05	.15	.30	0.5
$p_Y(y)$		0.25	0.25	0.5	1

2 CONTINUOUS RV:

LET X, Y BE 2 R.V. DEFINED ON THE SAME SPACE, THE JOINT PROBABILITY DENSITY FUNCTION $f(x,y)$:

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



- LET A BE A SET → $P[(X,Y) \in A] = \iint_A f(x,y)\, dx\, dy$

es. $A = \{a \leq X \leq b, c \leq Y \leq d\}$

$$\to \int_a^b \int_c^d f(x,y)\, dx\, dy$$

es.

$$\cdot f(x,y) = \begin{cases} \frac{6}{5}(x+y^2) & , 0 \le x \le 1, \quad 0 \le y \le 1 \\ 0 & , \text{otherwise} \end{cases}$$

$$\rightarrow P\left(0 \le X \le \frac{1}{4}, 0 \le Y \le \frac{1}{4}\right) = \int_0^{1/4} \int_0^{1/4} \frac{6}{5}(x+y^2)\, dx\, dy =$$

$$= \frac{6}{20}\left[\frac{x^2}{2} \Big|_{x=0}^{x=\frac{1}{4}} + \frac{y^3}{3} \Big|_{y=0}^{y=\frac{1}{4}} \right] = 0.0109$$

· MARGINAL PROBABILITY DENSITY FUNCTIONS:

$$\cdot X: \quad f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy \quad , \quad -\infty < x < \infty$$

$$\cdot Y: \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx \quad , \quad -\infty < y < \infty$$

· INDEPENDENT RV:

· IN GENERAL, 2 EVENTS A, B ARE INDEPENDENT $\Longleftrightarrow P(A \cap B) = P(A) \cdot P(B)$

· 2 RV X, Y ARE INDEPENDENT $\Longleftrightarrow \forall (x,y) \begin{cases} p(x,y) = p_X(x)\, p_Y(y) & \text{DISCRETE} \\ f(x,y) = f_X(x)\, f_Y(y) & \text{CONTINUOUS} \end{cases}$

es. (TABLE BEFORE) $\quad X, Y$ NOT INDEPENDENT

$$p(100,100) = 0.1 \neq p_X(100) \cdot p_Y(100) = 0.5 \cdot 0.25 = 0.125$$

• MORE THAN 2 RV:

• JOINT PMF: $p(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n)$

• JOINT PDF: $P(a_1 \le X_1 \le b_1, \dots, a_n \le X_n \le b_n) = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} f(x_1, \dots, x_n)\, dx_1, \dots, x_n$

• CONSIDER AN EXPERIMENT OF n INDEPENDENT AND IDENTICAL TRIALS / $\forall$ TRIAL $\rightarrow$ r POSSIBLE OUTCOMES

LET $p_i = P(\text{outcome } i \text{ on any particular trial})$, X_i = n° TRIALS RESULTING IN OUTCOME i, $i = 1, \dots, r$

↳ THIS IS A **MULTINOMIAL EXPERIMENT**

• JOINT PMF OF $X_1, \dots, X_r$: MULTINOMIAL DISTRIBUTION

$$p(x_1, \dots, x_r) = \begin{cases} \dfrac{n!}{(x_1!) \cdot \ldots \cdot (x_r!)} \cdot p_1^{x_1} \cdots p_r^{x_r}, & x_i = 0, 1, \dots; \ x_1 + \dots + x_r = n \\ 0, & \text{otherwise} \end{cases}$$

es. 4

If the allele of each of ten independently obtained pea sections is determined and $p_1 = P(\text{AA})$, $p_2 = P(\text{Aa})$, $p_3 = P(\text{aa})$, X_1 = number of AAs, X_2 = number of Aas, and X_3 = number of aa's, then the multinomial pmf for these X_i's is

$n = 10$

$$p(x_1, x_2, x_3) = \frac{10!}{(x_1!)(x_2!)(x_3!)} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

$, \ x_i = 0, 1, 2, \dots; \ x_1 + x_2 + x_3 = 10$

• IF $p_1 = p_3 = 0.25$, $p_2 = 0.5$, $P(X_1 = 2, X_2 = 5, X_3 = 3) \rightarrow \dfrac{10!}{2!\,3!\,5!} \cdot 0.25^2 \cdot 0.5^5\, 0.25^3 = 0.0769$

• $X_1, \dots, X_n$ INDEPENDENT $\iff$ $\forall$ SUBSET $\rightarrow$ JOINT PMF/PDF $= \prod$ MARGINAL PMF/PDF

- CONDITIONAL DISTRIBUTIONS :

LET X, Y BE 2 RV / MARGINAL $X = f_X(x)$.

CONDITIONAL PROBABILITY DENSITY FUNCTION OF $Y | X = x$:

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} \quad , \quad -\infty < y < \infty$$

⇝ ANALOG FOR PMF AND $X | Y = y$

es.

- $f(x,y) = \begin{cases} \frac{6}{5}(x+y^2) & , 0 \le x \le 1, \quad 0 \le y \le 1 \\ 0 & , \text{otherwise} \end{cases}$

→

- $Y | X = 0.8$:

$$f_{Y|X}(y|0.8) = \frac{f(0.8,y)}{f_X(0.8)} = \frac{1.2(0.8+y^2)}{\left.\int_0^1 f(x,y)\,dy\right|_{x=0.8}} = \frac{1}{34}(24+30y^2) \quad , \quad 0 < y < 1$$

$$\int_0^1 \frac{6}{5}(x+y^2)\,dy = \frac{6}{5}x + \frac{6}{5}\int_0^1 y^2\,dy = \frac{6}{5}x + \frac{1}{3}\cdot\frac{6}{5} \overset{*}{=} 1.2x + 0.4$$

- $$P(Y \le 0.5 | X = 0.8) = \int_{-\infty}^{0.5} f_{Y|X}(y|0.8)\,dy =$$

$$= \int_0^{0.5} \frac{1}{34}(24+30y^2)\,dy = 0.390$$

- $$E(Y|X=0.8) = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y|0.8)\,dy =$$

$$= \int_0^1 y \cdot \frac{1}{34}(24+30y^2)\,dy = 0.574$$

• JOINT R.V., PART. 2 :

• EXPECTED VALUES :

LET X, Y BE 2 R.V. WITH PMF $p(x,y)$ / PDF $f(x,y)$:

$$E[h(X,Y)] = \mu_{h(x,y)} = \begin{cases} \sum_x \sum_y h(x,y)\, p(x,y) & \text{DISCRETE} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f(x,y)\, dx\, dy & \text{CONTINUOUS} \end{cases}$$

ES. 13

• $p(x,y) = \begin{cases} 1/20 & , x = 1, \dots, 5 ,\ y = 1, \dots, 5 ,\ x \neq y \\ 0 & , \text{OTHERWISE} \end{cases}$

• $h(x,y) = |X - Y| - 1$ $\longrightarrow$

$h(x, y)$		1	2	3	4	5
	1	—	0	1	2	3
	2	0	—	0	1	2
y	3	1	0	—	0	1
	4	2	1	0	—	0
	5	3	2	1	0	—

$\rightarrow$ • $E[h(X,Y)] = \sum_{i=1}^{5} \sum_{j=1}^{5} (|X - Y| - 1) \frac{1}{20} = 1$

• IF X, Y INDEPENDENT $\longrightarrow \forall h, g : E[g(X) h(Y)] = E[g(X)] \cdot E[h(Y)]$

• COVARIANCE :

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y)\, p(x,y) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x,y)\, dx\, dy \end{cases}$$

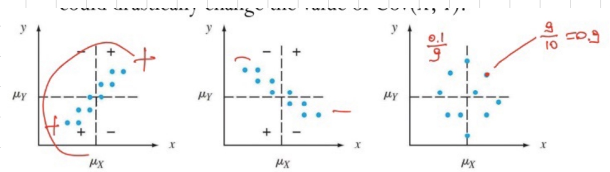
• PROPERTIES :

• $Cov(X,Y) = E(XY) - \mu_X \mu_Y$

• X, Y INDEPENDENT $\rightarrow Cov(X,Y) = 0$

• $Cov(X,X) = V(X) = E[(X - \mu_X)^2]$

• $Cov(aX + b, Y) = a \cdot Cov(X,Y)$



$\rightarrow$ IT IS A MEASURE OF "RELATION" BETWEEN X AND Y :

IF THEY HAVE $Cov(X,Y) >> 0 \rightarrow$ IF X "HIGH" $\rightarrow Y$ WILL POSSIBLY BE "HIGH" TOO

CORRELATION:

$$\rho_{X,Y} = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

- PROPERTIES:
  - IF $a \cdot c > 0 \rightarrow \rho(aX+b, cY+d) = \rho(X,Y)$
  - $\forall X, Y: \quad -1 \leq \rho_{X,Y} \leq 1$
  - IF X, Y INDEPENDENT $\overset{\not\longleftarrow}{\longrightarrow} \rho_{X,Y} = 0 \rightarrow X, Y$ UNCORRELATED
  - $\rho_{X,Y} = \pm 1 \iff Y = aX + b, \quad a \neq 0, \quad a, b \in \mathbb{R}$

$\rightarrow \rho_{X,Y} = 0 \not\rightarrow X, Y$ INDEPENDENT : IT ONLY SAYS THERE IS NOT A LINEAR DEPENDENCE

$\rightarrow$ 2 R.V. CAN BE UNCORRELATED BUT HIGHLY DEPENDENT : STRONG NON-LINEAR RELATION

$\rightarrow$ ASSOCIATION $\neq$ CAUSATION :

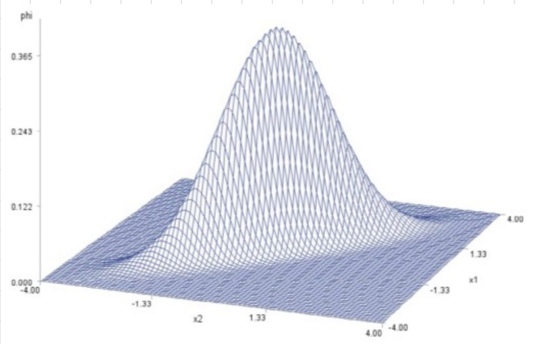
IF $\rho_{X,Y} >> 0 \rightarrow \quad \uparrow X \not\rightarrow \uparrow Y$

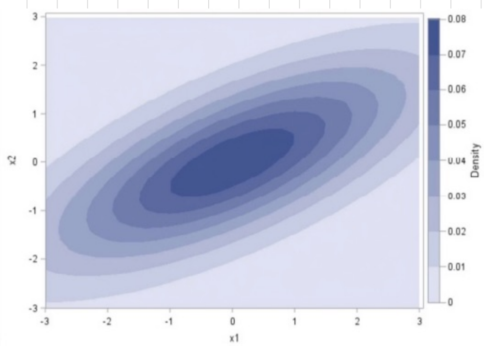
BIVARIATE NORMAL DISTRIBUTION: $\rightarrow$ MOST IMPORTANT JOINT DISTRIBUTION

One of the most important joint distributions is the bivariate normal distribution. We say that the random variables X, Y have a bivariate normal distribution if, for constants $\mu_x, \mu_y, \sigma_x > 0, \sigma_y > 0, -1 < \rho < 1$, their joint density function is given, for all $-\infty < x, y < \infty$, by

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\}$$

$$\phi(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\}$$





JOINT R.V., PART. 3:

STATISTICS AND THEIR DISTRIBUTIONS:

TWO SAMPLES OF THE SAME POPULATION OF n ELEMENTS WILL BE DIFFERENT

→ EACH OBSERVATION CAN BE A R.V. X_i

STATISTIC:

DEF.: ANY QUANTITY WHOSE VALUE CAN BE CALCULATED FROM SAMPLE DATA

- SAMPLE MEAN: $\underbrace{\bar{X}_m}_{\text{R.V.}} = \dfrac{X_1 + \dots + X_m}{n}$ → R.V. / $\bar{X} = \bar{x}$ → VALUE OF THE STATISTICS

- SAMPLE ST. DEV.: $S = \sqrt{\dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_m)^2}$

RANDOM SAMPLE:

DEF.: THE RV'S $X_1, \dots, X_m$ ARE SAID TO FORM A RANDOM SAMPLE OF SIZE n IF:

- X_i ARE INDEPENDENT R.V.
- $\forall X_i$ → X_i HAVE SAME PROBABILITY DISTRIBUTION

} → IID: INDEPENDENT AND IDENTICALLY DISTRIBUTED

- LET n: SAMPLE SIZE, N: POPULATION SIZE;

→ IN PRACTICE, IF $\dfrac{n}{N} < 0.05 = 5\%$ → X_i's FORM A RANDOM SAMPLE

DERIVING A SAMPLING DISTRIBUTION:

es. 5.27

x	80	100	120
$p(x)$	.2	.3	.5

with $\mu = 106$, $\sigma^2 = 244$

- IF X_1 = RENEWED 1ST SALE, X_2 = RENEWED 2ND SALE:

x_1	x_2	$p(x_1, x_2)$	$\bar{x}$ $\left(\frac{x_1+x_2}{2}\right)$	s^2 $\left((x_1-\bar{x})^2 + (x_2-\bar{x})^2\right)$
80	80	.04	80	0
80	100	.06	90	200
80	120	.10	100	800
100	80	.06	90	200
100	100	.09	100	0
100	120	.15	110	200
120	80	.10	100	800
120	100	.15	110	200
120	120	.25	120	0

- es. $p_{\bar{X}}(100) = P(\bar{X} = 100) = 0.1 + 0.09 + 0.1 = 0.29$
- es. $P(s^2 = 800) = 0.1 + 0.1 = 0.2$

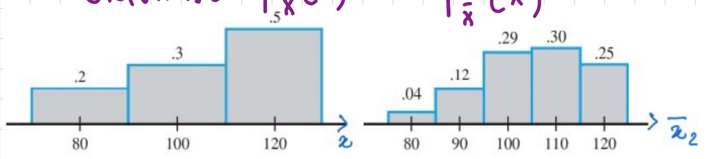
→

$\bar{x}$	80	90	100	110	120
$p_{\bar{X}}(\bar{x})$	.04	.12	.29	.30	.25

s^2	0	200	800
$p_{S^2}(s^2)$	.38	.42	.20

SAMPLING DISTRIBUTIONS

→ ORIGINAL $P_X(x)$ $\quad P_{\bar{x}}(\bar{x})$



- $\mu_{\bar{x}} = \sum \bar{x}\, p_{\bar{x}}(\bar{x}) = 106 = \mu_x$
- $\sigma^2_{\bar{x}} = \sum \bar{x}^2\, p_{\bar{x}}(\bar{x}) - \mu^2_{\bar{x}} = 122 = \frac{\sigma_x^2}{2} = \frac{244}{2}$
- IF $n = 4$, $\mu_{\bar{x}} = 106$, $\sigma^2_{\bar{x}} = 61 = \frac{\sigma^2}{4}$ ⤳ $n = 4$

- SIMULATION EXPERIMENTS:

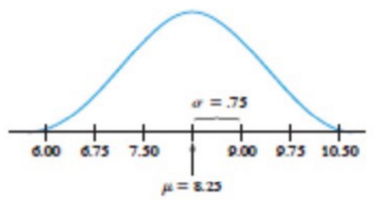
  - THESE CHARACTERISTICS MUST BE SPECIFIED FOR THIS EXPERIMENT:
    - STATISTIC OF INTEREST
    - POPULATION DISTRIBUTION
    - n : SAMPLE SIZE
    - k : n^o OF REPLICATIONS (IF $k \mapsto \infty$ → BETTER APPROXIMATION)

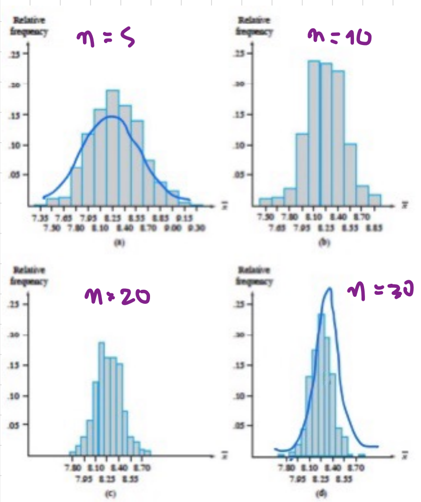
→ $\forall$ SAMPLE → CALCULATE THE STATISTIC OF INTEREST

→ CREATE AN HISTOGRAM $\forall k$: APPROX. SAMPLING DISTRIB.

es. 5.22

- $\mu = 8.25$, $\sigma^2 = 0.75$





⤳ $k = 500$

→ $\uparrow n$ → $\downarrow \sigma^2_{\bar{x}}$

- $\bar{X} \xrightarrow{n \mapsto \infty} \mu$

PART 2:

CONVERGENCE OF RV:

GIVEN THE SEQUENCE $\{X_n, n \in \mathbb{N}\}$ OF RVs:

$\rightarrow \lim_{n \to \infty} P[X_n \in A] = P[X \in A], \forall A \in \mathcal{B}$ → ALGEBRA DI BOREL

OR

$\rightarrow \lim_{n \to \infty} F_n(t) = F(t), \forall t \in \mathbb{R}$

↳ BUT THESE REQUIREMENTS ARE TOO RESTRICTIVE

CONVERGENCE IN DISTRIBUTION:

GIVEN X_n, $X_n \sim F_n$, WE SAY THAT:

$$\begin{cases} X_n \xrightarrow{d} X \\ F_n \xrightarrow{d} F \end{cases}$$

→ DISTRIBUTIONS

OR, FOR $X \sim F$ IF $\lim_{n \to \infty} F_n(t) = F(t), \forall t \,/\, F$ IS CONTINOUS IN t

PROPERTIES:

- $X_n \xrightarrow{d} X \iff E[g(X_n)] \xrightarrow{n \to \infty} E[g(X)], \forall g$ BOUNDED AND CONTINOUS
- $X_n \xrightarrow{d} X \not\Rightarrow X_n - X \xrightarrow{d} 0$

  ↳ TRUE $\iff X = a =$ CONST., WITH $p = 1$

es.

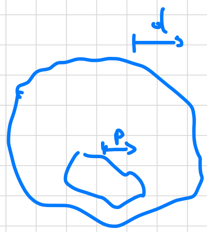
$$X_1, X_2 = \begin{cases} 0 & p = 1/2 \\ 1 & p = 1/2 \end{cases} \rightarrow X_1 - X_2 = \left.\begin{cases} -1 & p = \frac{1}{4} \\ 0 & p = 1/2 \\ 1 & p = 1/4 \end{cases}\right\} \neq 0$$

• CONVERGENCE IN PROBABILITY:

GIVEN $\{X_n, \ n \in \mathbb{N}\}$, WE SAY:

$$X_n \xrightarrow{P} X \quad \text{IF} \quad \forall \varepsilon > 0, \ \lim_{n \to \infty} P[\,|X_n - X| < \varepsilon\,] = 1$$

• PROPERTIES:



- $X_n \xrightarrow{P} X \ \underset{\longrightarrow}{\not\longleftarrow} \ X_n \xrightarrow{d} X$
- $X_n \xrightarrow{P} X \longrightarrow (X_n - X) \xrightarrow{P} 0$

- LIMIT THEOREMS:

SEE AFTER

- LLN: LAW OF LARGE NUMBERS $\rightarrow$ $\overline{X}_n \xrightarrow{P} \mu$

- CTL: CENTRAL LIMIT THEOREM $\rightarrow$ $U_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} Z \sim N(0,1)$

$\overline{X}_n = \frac{X_1 + \dots + X_n}{n} \xrightarrow{d} N \sim \left(\mu, \frac{\sigma^2}{n}\right) \xrightarrow{n \to \infty} N \sim (\mu, 0)$

- MARKOV INEQUALITY:

$\forall$ R.V. X HAVING FINITE MOMENTS $\left(E[X^k] < \infty \right)$:

$$P[|X| \geq a] \leq \frac{E[|X|^k]}{a^k}, \quad \forall a \in \mathbb{R}^+, \forall k \in \mathbb{N}^+$$

- COROLLARY:

$$P[|X| < a] \geq 1 - \frac{E[|X|^k]}{a^k}$$

- COROLLARY (CHEBYCHEV'S INEQUALITY):

$\forall$ R.V. X / $|E[X]| \cdot V[X] < \infty$:

$$P[|X - E[X]| < a] \geq 1 - \frac{V[X]}{a^2}$$

AND $$P[|X - E[X]| \geq a] < \frac{V[X]}{a^2}$$

- LAW OF LARGE NUMBERS (LLN):

LET $X_1, \dots, X_n$ BE IID RVs (IDENTICALLY DISTRIBUTED AND INDEPENDENT) $/ \; E[X_i] = \mu, \; V[X_i] = \sigma^2$

IF WE DEFINE $\overline{X}_n = \dfrac{X_1 + \dots + X_n}{n}$ AS THE SAMPLE MEAN

IF WE CONSIDER THE SEQUENCE $\{\overline{X}_n, n \in \mathbb{N}^+\}$, IT IS SUCH THAT:

$$(\overline{X}_n - \mu) \xrightarrow{P} 0, \quad \overline{X}_n \xrightarrow{P} \mu$$

- PROOF:

$$\overline{X}_n - \mu = \frac{X_1 + \dots + X_n}{n} - \frac{n\mu}{n} = \frac{(X_1 - \mu) + \dots + (X_n - \mu)}{n} =$$

$$= \frac{Y_1 + \dots + Y_n}{n} \quad / \; Y_i = X_i - \mu, \; E[Y_i] = \mu, V[Y_i] = \sigma^2$$

FIX $\varepsilon > 0$:

$$\to P[|\overline{X}_n - \mu| < \varepsilon] = P\left[\left|\frac{Y_1 + \dots + Y_n}{n}\right| < \varepsilon\right] =$$

MARKOV INEQUALITY, $k = 2$

$$= P[|Y_1 + \dots + Y_n| < n\varepsilon] \geq 1 - \frac{E[|Y_1 + \dots + Y_n|^2]}{(n\varepsilon)^2} =$$

$|\cdot|^2 = (\cdot)^2$

$$= 1 - \frac{V[Y_1 + \dots + Y_n]}{n^2 \varepsilon^2} = 1 - \frac{V[Y_1] + \dots + V[Y_n]}{n^2 \varepsilon^2} =$$

$$= 1 - \frac{n\sigma^2}{n^2 \varepsilon^2} \xrightarrow{n \to \infty} 1$$

$$\to \overline{X}_n \xrightarrow{P} \mu \longrightarrow \overline{X}_n \xrightarrow{d} \mu$$

es. DICE

$$X_i = \{1, \dots, 6\} \quad \to \overline{X}_n = \frac{X_1 + \dots + X_n}{n} \xrightarrow{n \to \infty} \mu = 3,5$$

- CENTRAL LIMIT THEOREM (CLT):

LET $X_1, \dots, X_n$ BE IID RVs / $E[X_i] = \mu$, $V[X_i] = \sigma^2$

IF WE DEFINE $S_n = X_1 + \dots + X_n \rightarrow E[S_n] = n\mu$, $V[S_n] = n\sigma^2$

$\rightarrow$

- $$U_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xmapsto{d} Z \sim N(0,1)$$
- $$\bar{X}_n = \frac{X_1 + \dots + X_n}{n} \xmapsto{d} N \sim \left(\mu, \frac{\sigma^2}{n}\right) \xmapsto{n \to \infty} N \sim (\mu, 0)$$

- FOR PROOF, CONCEPT OF MOMENT GENERATING FUNCTION IS NEEDED

- MOMENT GENERATING FUNCTION:

GIVEN A RV X, MGF OF X IS $\phi_X : \mathbb{R} \mapsto \mathbb{R}$:

$$\phi_X(s) = E[e^{sX}] = \int_{\mathbb{R}} e^{st} \cdot f_X(t)\, dt$$

- PROPERTY: IF $\exists \phi_X \rightarrow E[X^n] = \phi_X^{(n)}(0)$ ~> DERIVATIVE

ex.

- $\phi_X'(s) = E[X e^{sX}] \longrightarrow \phi_X'(0) = E[X \cdot 1] = E[X]$
- $\phi_X''(s) = E[X^2 e^{sX}] \longrightarrow \phi_X''(0) = E[X^2 \cdot 1] = E[X^2]$

$\rightarrow V[X] = E[X^2] - E^2[X] = \phi_X''(0) - (\phi_X'(0))^2$

- PROPERTY: IF IT $\exists$, MGF OF X DETERMINES UNIQUELY F_X

- PROPERTY:

LET $Y = X_1 + \dots + X_n \longrightarrow \phi_Y(s) = E[e^{sY}] = \prod_{i=1}^{n} \phi_{X_i}(s)$

- PROOF OF CLT:

$$U_n = \frac{(X_1 + \dots + X_n) - n\mu}{\sigma\sqrt{n}} = \frac{(X_1 - n\mu)}{\sigma} \cdot \frac{1}{\sqrt{n}} + \dots + \underbrace{\frac{(X_1 - n\mu)}{\sigma}}_{Y} \cdot \frac{1}{\sqrt{n}} =$$

$$= \frac{Y_1}{\sqrt{n}} + \dots + \frac{Y_n}{\sqrt{n}} \quad / \quad E[Y_i] = 0, \ V[Y_i] = 1, \ E[Y_i^2] = 1$$

- $$\phi_{U_n}(s) = \phi_{\frac{Y_1}{\sqrt{n}}}(s) \cdot \dots \ \phi_{\frac{Y_n}{\sqrt{n}}}(s) = \left[\phi_{\frac{Y_i}{\sqrt{n}}}(s)\right]^n =$$

$$= \left[E\left[e^{s\frac{Y_i}{\sqrt{n}}}\right]\right]^n \overset{\substack{\text{SERIES} \\ \text{EXPANSION} \\ \uparrow}}{=} \left\{E\left[1 + \frac{Y_i}{\sqrt{n}} s + \frac{1}{2}\frac{Y_i}{n} s^2 + \dots + o()\right]\right\}^n =$$

$$= \left\{1 + \frac{s}{\sqrt{n}} E[Y_1] + \frac{s^2}{2n} E[Y_i^2] + \dots + o()\right\}^n =$$

$$= \left(1 + \frac{1}{2}\frac{s^2}{n} + o\left(\frac{1}{n}\right)\right)^n \approx$$

$$\approx \left(1 + \frac{1}{2}\frac{s^2}{n}\right)^n \xrightarrow[\text{LIMIT OF } e]{n \mapsto \infty} e^{\frac{1}{2}s^2} = \text{MGF OF } Z \sim N(0,1)$$

- FOR $n >> 0$. $\bar{X}_n = \frac{S_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \xrightarrow{n \mapsto \infty} N(\mu, 0)$

- $S_n \sim N(E[S_n], V[S_n]) = N(n\mu, n\sigma^2)$

en.

LET $S_{100} = X_1 + \dots + X_{100}$ / $X_i = \begin{cases} 0 & p = 1/3 \\ 1 & p = 1/3 \\ 2 & p = 1/3 \end{cases}$

FIND APPROXIMATELY $P[S_{100} > 110]$

$\rightarrow$

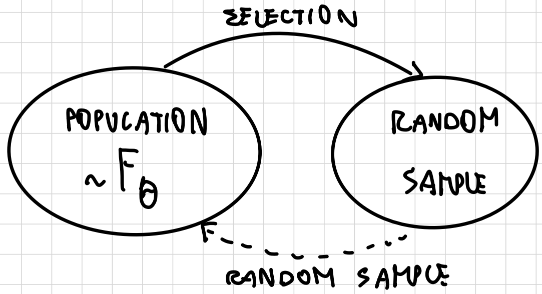
SINCE $n >> 0 \rightarrow S_n \sim N(n\mu, n\sigma^2)$

- $E[X_i] = \mu = 1$
- $V[X_i] = \sigma^2 = \left(0^2 \cdot \frac{1}{3} + 1^2 \cdot \frac{1}{3} + 2^2 \frac{1}{3}\right) - 1^2 = \frac{2}{3}$

$\rightarrow S_n \sim N(100 \cdot \mu, 100 \cdot \sigma^2) = N(100, \underset{\sim}{66})$ $200/3$

$\rightarrow P[S_{100} > 110] \cong P\left[Z \geq \frac{110 - \overset{\mu}{100}}{\underset{\sigma}{\sqrt{66}}}\right] = 1 - P[Z < 1.25] = 0.11$

- INFERENTIAL STATISTICS:



GIVEN A POPULATION, FIND THE ANALYTICAL EXPRESSION OF

$f_X,\ F_X,\ \mu_X,\ \sigma^2_X$

-> IT CAN BE DONE MEASURING A SAMPLE OF THE POPULATION

- SAMPLE: $(X_1, \dots, X_m) \,/\, X_i \sim f_X,\ X_i$ IID
- DATASET: $(x_1, \dots, x_m)$ REALIZATIONS OF $(X_1, \dots, X_m) \,/\, X_i = x_i$
- $F_{(X_1,\dots,X_m)}(t_1, \dots, t_m) = \prod_{i=1}^{m} F_X(t_i)$
- $f_{(X_1,\dots,X_m)}(t_1, \dots, t_m) = \prod_{i=1}^{m} f_X(t_i)$

- ESTIMATOR: STATISTIC USED TO ESTIMATE A PARAMETER

es.

WE WANT TO FIND μ_X OF A POPULATION, GIVEN THE SAMPLE $(X_1, \dots, X_m)$

- SAMPLE MEAN: $\bar{X}_m = \dfrac{X_1 + \dots + X_m}{m}$ ->R.V. $\left(/ E[\bar{X}_m] = \mu_m \right)$

-> ESTIMATE OF μ_X: $\hat{\mu}_X = \dfrac{x_1, \dots, x_m}{m}$ -> REALIZATION OF RVs

TOPICS COVERED NEXT ~>

- POINT ESTIMATION
  - Methods of finding estimators
    - Method of moments
    - Maximum likelihood
    - Bayes estimators
  - Methods of evaluating estimators
    - Bias
    - Variance
    - Mean square erro
- INTERVAL ESTIMATION
  - Methods of finding confidence intervals
  - Confidence intervals for μ
  - Confidence intervals for σ^2
  - Other confidence intervals
  - ~~Effects of the sample size~~
- Hypothesis Testing
  - Characteristics of a test
  - Tests for mean and variance
  - Test for goodness of fit
  - Other useful test

- POINT ESTIMATION:

- METHODS OF MOMENTS (MME):

  - ASSUME THAT $X \sim f_{\bar{\theta}}$ / $\bar{\theta}$ IS A VECTOR OF PARAMETERS

    es IF $X \sim N(\mu, \sigma^2) \rightarrow \bar{\theta} = (\mu, \sigma^2)$

  - ASSUME WE HAVE A SAMPLE $\bar{X} = (X_1, \dots, X_n)$

→ EMPIRICAL MOMENTS:

- $M_1 = \dfrac{X_1 + \dots + X_n}{n} = \bar{X}_n$: ESTIMATOR FOR $\mu_1(\bar{\theta}) = E[X^1]$
- $M_2 = \dfrac{X_1^2 + \dots + X_n^2}{n}$: ESTIMATOR FOR $\mu_2(\bar{\theta}) = E[X^2]$
- $M_k = \dfrac{X_1^k + \dots + X_n^k}{n}$: ESTIMATOR FOR $\mu_k(\bar{\theta}) = E[X^k]$

→ IN ORDER TO FIND THE ESTIMATES (REAL VALUES OF THE ESTIMATORS), SOLVE THE SYSTEM:

$$\begin{cases} \mu_1(\bar{\theta}) = M_1 \\ \mu_2(\bar{\theta}) = M_2 \\ \vdots \\ \mu_k(\bar{\theta}) = M_k \end{cases}$$

/ K: n° OF PARAMETERS TO ESTIMATE

es.

$X \sim Exp(\lambda)$, λ UNKNOWN → $\bar{\theta} = (\lambda)$ → 1 PARAMETER

→ $\mu_1(\bar{\theta}) = E[X] = \dfrac{1}{\lambda}$; $M_1 = \dfrac{X_1 + \dots + X_n}{n}$

$\{ \mu_1(\bar{\theta}) = M_1 \rightarrow \dfrac{1}{\lambda} = M_1 \rightarrow \lambda = \dfrac{1}{\bar{X}_n}$

- GIVEN SAMPLE $(X_1, \dots, X_5) = (1, 3, 5, 8, 15)$: $\hat{\lambda} = \dfrac{1}{\left(\frac{1+3+5+8+15}{5}\right)} = 0.15$

es.

$$X \sim N(\mu, \sigma^2) \rightarrow \bar{\theta} = (\mu, \sigma^2)$$

$\rightarrow$

- $\mu_1(\bar{\theta}) = E[X] = \mu$
- $\mu_2(\bar{\theta}) = E[X^2] = V[X] + E^2[X] = \sigma^2 + \mu^2$

$\rightarrow$

$$\begin{cases} \mu_1(\bar{\theta}) = M_1 \\ \mu_2(\bar{\theta}) = M_2 \end{cases} \rightarrow \begin{cases} \mu = M_1 \\ \sigma^2 + \mu^2 = M_2 \end{cases}$$

$\rightarrow$

- $\hat{\mu} = \dfrac{X_1 + \dots + X_n}{n} = \bar{X}_n$
- $\sigma^2 = \dfrac{X_1^2 + \dots + X_n^2}{n} - (\bar{X}_n)^2 = \dfrac{1}{n} \sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\right)^2$

- IF SAMPLE : $(10, 12, 17)$:

  - $\hat{\mu} = \dfrac{10 + 12 + 17}{3} = 13$
  - $\hat{\sigma}^2 = \dfrac{1}{3}\left(10^2 + 12^2 + 17^2\right) - 13^2 = 8.66$

es.

$X \sim Bin(k, p) \rightarrow \bar{\theta} = (k, p)$

- $\mu_1(\bar{\theta}) = k \cdot p$; $\mu_2(\bar{\theta}) = E[X^2] = V^2[X] + E^2[X] =$

$= kp(1-p) + (kp)^2 = kp(1-p+kp)$

$\rightarrow$

$$\begin{cases} \mu_1(\bar{\theta}) = M_1 \\ \mu_2(\bar{\theta}) = M_2 \end{cases} \rightarrow \begin{cases} kp = M_1 \\ kp(1-p+kp) = M_2 \end{cases} \begin{cases} \hat{p} = \dfrac{k}{M_1} \\ \dots \end{cases}$$

$$\longrightarrow 2^{ND} \text{ EQ.} \Big|_{\hat{p} = \frac{k}{M_1}} = \dots = \hat{k} = \frac{M_1^2}{-M_2 + M_1 + M_1^2}$$

- CORRECTED SAMPLE VARIANCE:

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

$\rightsquigarrow$ INSTEAD OF 'n'

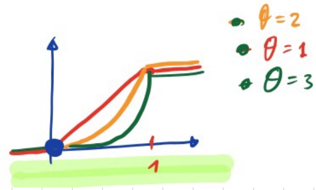
$\rightarrow$ IN THIS WAY $E[\hat{\sigma}_n^2] = \sigma^2$

- IF 'n' INSTEAD OF 'n-1':

GIVEN n SAMPLES: $E[\hat{\sigma}^2] = \dfrac{n-1}{n} \sigma^2$

es.

$n = 4 \; : \; (0.42, \; 0.50, \; 0.65, \; 0.23)$



$$\cdot \; X_\theta \sim f_\theta(t) = \begin{cases} \theta t^{\theta-1} & , \text{ IF } t \in [0,1] \\ 0 & , \text{ OTHERWISE} \end{cases}$$

$$\hookrightarrow F_\theta(t) = t^\theta \; , \; t \in [0,1]$$

->

$$\cdot \; \mu_1(\theta) = E[X_\theta] = \int_0^1 t \, f_{X_\theta}(t) \, dt = \int_0^1 t \cdot \theta t^{\theta-1} \, dt =$$

$$= \theta \int_0^1 t^\theta \, dt = \theta \left[\frac{t^{\theta+1}}{\theta+1} \right]_0^1 = \frac{\theta}{\theta+1}$$

$$\to \left\{ \mu_1(\theta) = M_1 \to \frac{\theta}{\theta+1} = M_1 \to \hat{\theta} = \frac{M_1}{1-M_1} \simeq 0.69 \right.$$

$$/ \; M_1 = \frac{0.42 + 0.50 + 0.65 + 0.23}{4}$$

- METHOD OF MAX LIKELIHOOD (MLE):

LET $X_{\bar{\theta}} = f_{\bar{\theta}}$ / $\bar{\theta}$: VECTOR OF PARAMETERS

ASSUME SAMPLE $\bar{X} = (X_1, \dots, X_n)$: ~> RVs

$$f_{\bar{X}, \bar{\theta}}(t_1, \dots, t_n) = \prod_{i=1}^{n} f_{\bar{\theta}}(t_i)$$

- WE OBSERVE THE EMPIRICAL SAMPLE $(x_1, \dots, x_n)$:

$$L(\bar{\theta} \mid \bar{x}) = f_{\bar{x}, \bar{\theta}}(x_1, \dots, x_n) = \prod_{i=1}^{n} f_{\bar{\theta}}(x_i)$$

- $\hat{\bar{\theta}}$: ESTIMATE OF $\bar{\theta}$

↳ IT MAXIMIZES THE LIKELIHOOD FUNCTION $L(\bar{\theta} \mid \bar{x})$, WHICH REPRESENTS THE PROBABILITY OF OBSERVING WHAT WE OBSERVED

es.

$X_\theta \sim Exp(\theta)$

SAMPLE $(X_1, \dots, X_n) \longmapsto \bar{x} = (x_1, \dots, x_n)$

->

$$L(\theta \mid \bar{x}) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \theta e^{-\theta x_i} =$$

$$= \theta^n e^{-\theta(x_1 + \dots + x_n)} = \theta^n e^{-\theta \sum x_i}$$

- TO MAXIMIZE $L(\theta \mid \bar{x})$:

$$\to \frac{d L(\theta \mid \bar{x})}{d\theta} = 0 \to n\theta^{n-1} e^{-\theta \sum x_i} + \theta^n \left(-\sum x_i\right) e^{-\theta \sum x_i} =$$

$$= \theta^{n-1} \left(n - \theta \sum x_i\right) e^{-\theta \sum x_i} = 0$$

- IF $n - \theta \sum x_i = 0 \to \hat{\theta} \sum x_i = n \to \hat{\theta} = \frac{n}{\sum x_i} = \frac{1}{\bar{X}_n}$

- IT IS EASIER TO CONSIDER THE log - LIKELIHOOD:

$\rightarrow \log\left(L(\Theta \mid \bar{x})\right)$

es.

$X_{\bar{\Theta}} \sim N(\mu, \sigma^2)$, SAMPLE : $(x_1, \dots, x_n)$

$\rightarrow$

$$L(\Theta \mid \bar{x}) = \prod_{i=1}^{n} f_{\bar{\Theta}}(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} =$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)}$$

- WE MUST FIND $\hat{\mu}, \hat{\sigma}$ THAT MAXIMIZE:

$$\log\left[L(\bar{\Theta} \mid \bar{x})\right] = \log\left[(2\pi\sigma^2)^{-\frac{n}{2}}\right] - \frac{1}{2}\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 =$$

$$= \log(2\pi)^{2/n} + \log\left((\sigma^2)^{-\frac{n}{2}}\right) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

- $$\frac{\delta \log\left(L(\bar{\Theta} \mid \bar{x})\right)}{\delta \mu} = 0 \rightarrow -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu) = 0$$

$$\rightarrow \sum x_i - n\mu = 0 \rightarrow n\mu = \sum x_i \rightarrow \hat{\mu} = \frac{\sum x_i}{n} = \bar{X}_n$$

- $$\frac{\delta \log\left(L(\bar{\Theta} \mid \bar{x})\right)}{\delta \sigma} = 0 \rightarrow -\frac{n}{\sigma} + \frac{\sum(x_i-\mu)^2}{\sigma^3} = 0$$

$$\rightarrow -\sigma^2 n + \sum(x_i-\mu)^2 = 0 \rightarrow \hat{\sigma}^2 = \frac{\sum(x_i-\mu)^2}{n} \quad / \mu = \bar{X}_n$$

es.

$$X_\theta \sim f_\theta(t) = \begin{cases} \theta t^{\theta-1} & , t \in [0,1] \\ 0 & , \text{OTHERWISE} \end{cases} \quad , \text{SAMPLE}: (x_1, \dots, x_n)$$

$\rightarrow$

$$L(\bar{\theta} \mid \bar{x}) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \theta x_i^{\theta-1} =$$

$$= \theta^n (x_1 \cdot \dots \cdot x_n)^{n-1} = \theta^n \left(\prod x_i\right)^{\theta-1}$$

$$\cdot \log\left(L(\bar{\theta} \mid \bar{x})\right) = \log \theta^n + \log\left(\left(\prod x_i\right)^{\theta-1}\right) =$$

$$= n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log x_i$$

$$\cdot \frac{d L(\theta \mid \bar{x})}{d\theta} = 0 \rightarrow n \cdot \frac{1}{\theta} + \sum_{i=1}^{n} \log(x_i) = 0$$

$$\rightarrow \hat{\theta} = - \frac{n}{\sum \log(x_i)}$$

es.

$$X_p \sim \text{Bernoulli}(p)$$

$\rightarrow$

$$L(p \mid \bar{x}) = \prod_{i=1}^{n} f_p(x_i) = p^{\sum x_i} (1-p)^{n - \sum x_i}$$

$$\cdot \log\left(L(p \mid \bar{x})\right) = \sum x_i \log p + \left(n - \sum x_i\right) \log(1-p)$$

$$\cdot \frac{d \log\left(L(p \mid \bar{x})\right)}{dp} = 0 \rightarrow \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0$$

$$\rightarrow (1-p) \sum x_i = p\left(n - \sum x_i\right) \rightarrow \sum x_i - p \sum x_i = np - p \sum x_i$$

$$\rightarrow \hat{p} = \frac{\sum x_i}{n} = \bar{x}$$

BAYESIAN ESTIMATION :

SUBJECTIVE BELIEF ABOUT π

IN BAYESIAN SETTING, θ IS CONSIDERED A RV ITSELF

ITS VARIABILITY (BEFORE SEEING ANY DATA) IS ENCODED IN A PRIOR DISTRIBUTION $\pi(\theta)$

- THE PRIOR IS THEN UPDATED WITH INFORMATION BY THE SAMPLE, ACCORDING TO BAYES RULE :

POSTERIOR DISTRIBUTION : $\pi(\theta|\bar{x}) = \dfrac{f(\bar{x}|\theta)\,\pi(\theta)}{m(\bar{x})} \propto f(\bar{x}|\theta)\,\pi(\theta)$

- $f(\bar{x}|\theta)$: JOINT DENSITY OF THE SAMPLE GIVEN THE PARAMETER
- $m(\bar{x}) = \int f(\bar{x}|\theta)\,\pi(\theta)\,d\theta$: NORMALIZING FACTOR

  ↳ DIFFICULT TO COMPUTE IN GENERAL, EASIER FOR SOME FAMILY OF PRIORS

  ↳ CONJUGATE

SO,

$X_\theta \sim \text{Bernoulli}(p)$ / p NOT CONSTANT, p IS A R.V.

LET $\Gamma(x)$ BE THE GAMMA FUNCTION $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\,dt$

$$\left[\ \Gamma(x+1) = x\,\Gamma(x),\quad \Gamma(0) = 1,\quad \Gamma(n) = n!\ \right]$$

- IF THE PRIOR p IS IN THE BETA FAMILY : $\underbrace{B}_{\text{PRIOR}} = A_{\alpha,\beta} \cdot \underbrace{\text{Bernoulli}(p,\alpha,\beta)}_{\text{KERNEL}}$ POSTERIOR

$$f(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, p^{\alpha-1}(1-p)^{\beta-1} = A_{\alpha,\beta}\, K(p,\alpha,\beta)$$

-> THEN THE POSTERIOR IS IN THE BETA FAMILY AS WELL

- IN THIS CASE : THE BETA FAMILY IS A CONJUGATE TO THE BERNOULLI FAMILY

-> CONJUGATE : A CLASS OF PRIOR $\tilde{\pi}(\theta) \in \Pi$ FORM A CONJUGATE FAMILY WITH RESPECT TO $f(X|\theta)$, IF THE POSTERIOR $\tilde{\pi}(\theta|X) \in \Pi$, $\forall\, \tilde{\pi}(\theta) \in \Pi$

$$f(p \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} = A_{\alpha,\beta}\, K(p, \alpha, \beta)$$

- $K(p, \alpha, \beta)$: KERNEL OF $f(p \mid \alpha, \beta)$
- $A_{\alpha,\beta}$: NORMALIZING CONSTANT
  - ↳ c/ x

GIVEN a $Beta(\alpha, \beta)$
- ↳ $E[\cdot] = \frac{\alpha}{\alpha+\beta}$
- ↳ $V[\cdot] = \dots$

- IN ORDER TO NOT CALCULATE $m(\bar{x})$, WE NEED TO SHOW THAT POSTERIOR IS STILL A BETA, CHECKING THAT IT HAS A BETA KERNEL

$$\rightarrow \pi(\theta \mid \bar{x}) \sim f(x \mid \theta)\,\pi(\theta) = \binom{n}{r} p^{r}(1-p)^{n-r}\left[A_{\alpha,\beta}\, p^{\alpha-1}(1-p)^{\beta-1}\right]$$

$$\sim p^{r+\alpha-1}(1-p)^{n-r+\beta-1} = K(p, r+\alpha, n-r+\beta)$$

↳ BETA KERNEL

- POINT ESTIMATION OF $p = E[X]$:

$$\hat{p} = E[Beta(r+\alpha, n-r+\beta)] = \frac{r+\alpha}{(r+\alpha)+(n-r+\beta)} = \frac{r+\alpha}{n+\alpha+\beta}$$

- $\hat{p}$ CAN BE DECOMPOSE AS A WEIGHTED AVERAGE:

$$\hat{p} = \frac{n}{n+\alpha+\beta} \cdot \frac{r}{n} + \frac{\alpha+\beta}{n+\alpha+\beta} \cdot \frac{\alpha}{\alpha+\beta}$$

- WHERE:
  - SAMPLE MEAN $= \frac{\sum x_i}{n} = \frac{r}{n}$
  - PRIOR MEAN $= \frac{\alpha}{\alpha+\beta}$

∞ AMOUNT OF DATA

$\rightarrow$ IF $n \mapsto \infty$: PRIOR BELIEF LOSES IMPORTANCE $\rightarrow \hat{p} \longmapsto \frac{n}{n+\alpha+\beta} \cdot \frac{r}{n}$

es.

LET $X \sim N(\mu, \overset{\overset{\text{KNOWN}}{\uparrow}}{\sigma^2})$

- PRIOR : $\sim N(\nu, \tau^2)$

$\rightarrow \pi(\mu \mid X, \sigma^2, \nu, \tau^2)$ IS STILL NORMAL WITH

- $E[\mu \mid X] = \dfrac{n\tau^2}{n\tau^2 + \sigma^2} \bar{X}_n + \dfrac{\sigma^2}{n\tau^2 + \sigma^2} \nu$

SAMPLE DATA — PRIOR MEAN

- $V[\mu \mid X] = \dfrac{\sigma^2 \tau^2}{n\tau^2 + \sigma^2}$

$\rightarrow$ POINT BAYES ESTIMATOR FOR μ : $\hat{\mu} = \dfrac{n\tau^2}{n\tau^2 + \sigma^2} \bar{X}_n + \dfrac{\sigma^2}{n\tau^2 + \sigma^2} \nu$

- $\pi(x \mid \theta) \sim L(\theta \mid x)\, \pi(\theta) =$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(X_i - \theta)^2}{\sigma^2}\right) \cdot \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2}\frac{(\nu - \theta)^2}{\tau^2}\right) =$$

$$= \dots = \overbrace{\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \frac{1}{\sqrt{2\pi\tau^2}}\right)}^{A_{\alpha,\beta}} \overbrace{\exp\left(-\frac{1}{2}\left(\frac{\sum (X_i - \theta)^2}{\sigma^2} + \frac{(\nu - \theta)^2}{\tau^2}\right)\right)}^{K(\,)}$$

$\rightarrow \pi(X \mid \theta) \sim \exp(\dots)$.

We know that

$$\pi(\mathbf{X}|\theta) \sim L(\theta|\mathbf{X})\pi(\theta) \qquad [\theta \equiv \mu]$$

but the RHS is now thought as a density in θ, while $\mathbf{X}$ is considered as a parameter, and therefore factor in the posterior that depends only on $\mathbf{X}$ and σ (known) can be included in the normalizing constant. We have:

$$\pi(\mathbf{X}|\theta) \sim \exp\left[-\frac{\sum_i (X_i - \theta)^2}{2\sigma^2} - \frac{(\theta - \nu)^2}{2\tau^2}\right]$$

$$\sim \exp\left[-\frac{\sum_i X_i^2 - 2\theta \sum_i X_i + n\theta^2}{2\sigma^2} - \frac{\theta^2 - 2\nu\theta + \nu^2}{2\tau^2}\right]$$

$$\sim \exp\left[-\frac{n\tau^2 + \sigma^2}{2\sigma^2\tau^2}\theta^2 + 2\frac{\nu\sigma^2 + n\tau^2 \bar{X}}{2\sigma^2\tau^2}\theta\right]$$

$$\sim \exp\left[-\frac{\theta^2 - 2\frac{\nu\sigma^2 + n\tau^2\bar{X}}{n\tau^2+\sigma^2}\theta}{2\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}}\right] \sim \exp\left[-\frac{\left[\theta - \frac{\nu\sigma^2 + n\tau^2\bar{X}}{n\tau^2+\sigma^2}\right]^2}{2\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}}\right]$$

for the exponent, in fact, we have

$$- \frac{1}{2\sigma^2\tau^2}\left[\tau^2 \sum X_i^2 - 2\theta\tau^2 \sum X_i + n\theta^2\tau^2 + \theta^2\sigma^2 - 2\nu\theta\sigma^2 + \nu^2\sigma^2\right]$$

$$= -\frac{1}{2\sigma^2\tau^2}\left[\theta^2(n\tau^2 + \sigma^2) - 2\theta(\nu\sigma^2 + n\tau^2\bar{X}) + \underbrace{\tau^2 n \cdot \frac{\sum X_i^2}{n} + \nu^2\sigma^2}_{\text{constant in } \theta}\right]$$

$$= -\frac{(n\tau^2 + \sigma^2)}{2\sigma^2\tau^2}\left[\theta^2 - 2\theta\frac{(\nu\sigma^2 + n\tau^2\bar{X})}{(n\tau^2 + \sigma^2)} + \text{const.}\right]$$

$$= -\frac{1}{2}\frac{(n\tau^2 + \sigma^2)}{\sigma^2\tau^2}\left[\left(\theta - \frac{\nu\sigma^2 + n\tau^2\bar{X}}{n\tau^2 + \sigma^2}\right)^2\right]$$

- **HOW TO EVALUATE ESTIMATORS?**

GIVEN 2 ESTIMATORS M_1, M_2 WE CHOSE THE ONE WITH LESS VARIANCE:

$$V[M_1] < V[M_2] \rightarrow V[M_1]$$

$\rightarrow$ IT HAS AN HIGHER PROBABILITY OF OBTAINING ESTIMATES CLOSER TO μ

- **BIAS**:

GIVEN AN ESTIMATOR $\hat{\theta}$, WE DEFINE:

$$BIAS(\hat{\theta}) = E[\hat{\theta}] - \theta$$

- IF $BIAS(\hat{\theta}) = 0 \rightarrow E[\hat{\theta}] = \theta \rightarrow \hat{\theta}$ IS UNBIASED

es. $\hat{\sigma}^2 = \dfrac{\sum (X - \bar{X}_n)^2}{n} \rightarrow$ BIASED, $\because E[\hat{\sigma}] = \dfrac{n-1}{n} \sigma^2$

es, $f_\theta(t) = \dfrac{2t}{\theta^2} \cdot \mathbb{1}_{[0,\theta]}(t) \longrightarrow$ 2 ESTIMATORS: $\begin{cases} \theta_1 = \max\{X_i\} \\ \theta_2 = \frac{3}{2}\bar{X}_n \end{cases}$

- $E[\theta_2] = \frac{3}{2} \cdot \frac{1}{n} E[X_1 + \dots + X_n] = \frac{3}{2n}(E[X_1] + \dots + E[X_n]) =$

$$= \frac{3 \cdot n \cdot \mu}{2n} = \theta \;/\; \mu = \frac{2}{3}\theta \rightarrow \theta_2 \text{ UNBIASED}$$

- $E[\theta_1] = E[\underbrace{\max\{X_i\}}_{Y}]$ BIASED

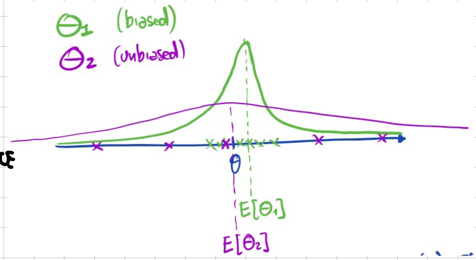
$\dots, f_Y(t) = \frac{1}{\theta^{2n}} \cdot 2n t^{2n-1}, t \in [0,\theta] \rightarrow E[\theta_1] = E[\max\{X_i\}] = \frac{2n}{2n+1}\theta$

- **EFFICENCY**:

GIVEN 2 ESTIMATORS θ_1, θ_2:

IF $V[\theta_1] < V[\theta_2] \rightarrow \theta_1$ IS MORE EFFICIENT THAN θ_2

- SOMETIMES IT IS BETTER TO USE A BIASED ESTIMATOR WITH A SMALLER VARIANCE RATHER THAN AN UNBIASED ESTIMATOR WITH HIGH VARIANCE



$\rightarrow$ BETTER TO USE MSE

- MSE:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = V[\hat{\theta}] + BIAS^2(\hat{\theta})$$

$\rightarrow$ BETTER TO CHOSE THE ONE WITH SMALLER MSE

es.

$$X \sim U[0, \theta]$$

SINCE $E[X] = \frac{\theta}{2}$ $\rightarrow$ ESTIMATOR $\hat{\theta}_1$ CAN BE: $\hat{\theta}_1 = 2\bar{X}_n$ (UNBIASED)

OR USING MLE: $\hat{\theta}_2 = \max\{X_1, \dots, X_n\}$ (BIASED)

$\rightarrow$

- $MSE(\hat{\theta}_1) = V[\hat{\theta}_1] + 0 = 2^2 V[\bar{X}_n] = 2^2 \cdot \frac{1}{n} \frac{\theta^2}{12} = \frac{\theta^2}{3n}$
- $MSE(\hat{\theta}_2) = V[\hat{\theta}_2] + (E[\hat{X}_2] - \theta)^2 = \frac{n\theta^2}{(n+2)(n+1)^2} + \left(\frac{n}{n+1}\theta - \theta\right)^2 =$

$$= \frac{2\theta^2}{(n+1)(n+2)} < MSE(\hat{\theta}_1)$$

$\rightarrow$ $\hat{\theta}_2$ BETTER THAN $\hat{\theta}_1$

- THEOREM (CRAMER - RAO LOWER BOUND):

LET $X \sim f_\theta(t, \theta)$, WITH CONTINOUS 1^{ST}, 2^{ND} DERIVATIVE,

FOR ANY UNBIASED ESTIMATOR $\hat{\theta}$:

$$V[\hat{\theta}] \geq \left\{ n \cdot E\left[\left(\frac{\delta \log(f_x(t,\theta))}{\delta \theta} \right)^2 \right] \right\}^{-1} = n\, E\left[\left(\frac{d \ln f}{d\theta}\right)^2\right]$$

$\log \rightsquigarrow \ln$

$$= \left\{ -n \cdot E\left[\frac{\delta^2 \log(f_x(t,\theta))}{\delta^2 \theta^2} \right] \right\}^{-1} \qquad -n\, E\left[\frac{d^2 \ln f}{d^2 \theta}\right]$$

es

$$X_p \sim \text{Bernoulli}(\underset{\text{UNKNOWN}}{\hat{p}}), \quad E[X_p] = p$$

$\to \theta = \bar{X}_n$ IS A SUITABLE ESTIMATOR FOR p

- $E[\theta] = E[\bar{X}_n] = p \to$ UNBIASED
- $V[\theta] = V[\bar{X}_n] = \frac{1}{n} V[X_p] = \frac{1}{n}(p(1-p))$

$\to$

- THERE EXISTS OTHERS $\hat{\theta}$ WITH SMALLER $V[\cdot]$? $\to$ CHECK WITH CRAMER-RAO

- $f_p(t) = \begin{cases} p, & t=1 \\ 1-p, & t=0 \end{cases} \to \log f_p(t) = \begin{cases} \log(p), & t=1 \\ \log(1-p), & t=0 \end{cases}$

- $\frac{\delta \log f_p(t)}{\delta p} = \begin{cases} 1/p, & t=1 \\ -1/(1-p), & t=0 \end{cases} \to \underbrace{\left(\frac{\delta \log f_p(t)}{\delta p}\right)^2}_{h(X_p)} = \begin{cases} 1/p^2, & t=1 \\ 1/(1-p)^2, & t=0 \end{cases}$

- $E[h(X_p)] = h(0) \cdot P[X_p = 0] + h(1) \cdot P[X_p = 1] =$

$$\frac{1}{(1-p)^2}(1-p) + \frac{1}{p^2} \cdot p = \frac{1}{p(1-p)}$$

$\to$ LOWER BOUND: $\{n \cdot E[h(X_p)]\}^{-1} = \frac{p(1-p)}{n} \to \hat{\theta} = \bar{X}_n$ IS THE BEST POSSIBLE

- **CONSISTENCY**:

AN ESTIMATOR $\hat{\theta}_n = W(X_1, \dots, X_n)$ IS SAID TO BE CONSISTENT $\iff$ $\hat{\theta}_n \xrightarrow{P} \theta$, $\lim_{n\to\infty} P[|\hat{\theta}_n - \theta| < \varepsilon] = 1, \forall \varepsilon > 0$

es.

LET $X_\theta \sim U[0, \theta]$

- MLE : $\hat{\theta}_n = \max\{X_i, \ i = 1, \dots, n\}$
- WE NEED TO PROVE $\lim_{n\to\infty} P[|\hat{\theta}_n - \theta| < \varepsilon] = 1$

->

$$X_\theta \sim f_\theta(t) = \begin{cases} 1/\theta, & t \in [0, \theta] \\ 0, & \text{OTHERWISE} \end{cases}$$

$\to F_\theta(t) = \frac{t}{\theta}, \ t \in [0, \theta]$ AND (0 IF $t < 0$) AND (1 IF $t > \theta$)

- $\hat{\theta}_n = \max\{X_i\}$, $F_{\hat{\theta}_n}(t) = [F_\theta(t)]^n = \left(\frac{t}{\theta}\right)^n, \ t \in [0, \theta]$

$\to f_{\hat{\theta}_n}(t) = \frac{n t^{n-1}}{\theta^n}, \ t \in [0, \theta]$; 0 OTHERWISE

$$\to P[|\hat{\theta}_n - \theta| < \varepsilon] = P[\theta - \varepsilon < \hat{\theta}_n < \theta + \varepsilon] = P[\theta - \varepsilon < \hat{\theta}_n < \theta]$$

$$= \int_{\theta-\varepsilon}^{\theta} f_{\hat{\theta}_n}(t)\,dt = \int_{\theta-\varepsilon}^{\theta} \frac{n t^{n-1}}{\theta^n}\,dt = \left(\frac{t}{\theta}\right)^n \Bigg|_{\theta-\varepsilon}^{\theta} = 1 - \frac{(\theta-\varepsilon)^n}{\theta^n}$$

$$= 1 - \frac{(\theta-\varepsilon)^n}{\theta^n} \xrightarrow{n\to\infty} 1 \to \hat{\theta}_n \xrightarrow{P} \theta \to \text{IT IS CONSISTENT}$$

es. 12/15 ...

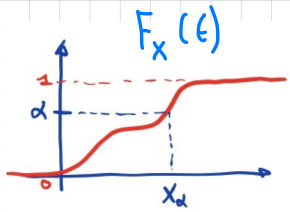
- # INTERVAL ESTIMATION (CONFIDENCE INTERVAL):

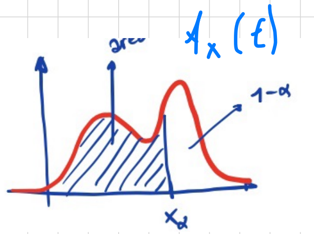
PRELIMINARY

- QUANTILES:

QUANTILE OF ORDER α;

$$P[X \leq x_\alpha] = \alpha$$





THEY ARE ESTIMATES OF PARAMETERS EXPRESSED IN TERMS OF INTERVALS, SUCH THAT THE TRUE VALUE IS INSIDE THAT INTERVAL WITH A FIXED CONFIDENCE ("PROBABILITY" IS NOT APPROPRIATE)

↳ THERE IS NO RANDOMNESS

es.

$$X \sim N(\mu, \sigma^2) \to \bar{X}_n \in \left(\mu - 3\cdot\sqrt{\frac{\sigma^2}{n}},\ \mu + 3\cdot\sqrt{\frac{\sigma^2}{n}}\right) \text{ WITH PROBABILITY } 99\%$$

$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

→ LET $X_\theta \sim F_\theta$ / θ: TO BE ESTIMATED

LET $\bar{X} = (X_1, \dots, X_n)$ BE A SAMPLE

- AN INTERVAL ESTIMATE OF θ IS A PAIR OF STATISTICS

$$L(\bar{X}), U(\bar{X}) \;/\; \theta \in [L(\bar{X}), U(\bar{X})] \text{ WITH A CONFIDENCE INTERVAL}$$

- COVERAGE PROBABILITY OF $I(\bar{X}) = (L(\bar{X}), U(\bar{X}))$: $p = P[\theta \in I(\bar{X})]$

- WHEN $\bar{X}$ IS OBSERVED: $(X_1, \dots, X_n) \mapsto (x_1, \dots, x_n)$

$$L(\bar{X}), U(\bar{X}) \mapsto \ell, u$$

$$\to I = (\ell, u) \subseteq \mathbb{R}$$

- **INTERVAL ESTIMATE FOR μ** (WHEN σ^2 KNOWN, n LARGE $(n \geq 30)$):

$\hookrightarrow \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$$X \sim F_\mu \quad , \quad E[X] = \mu$$

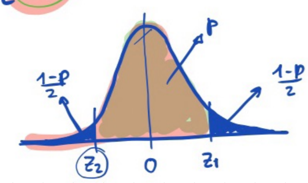
$$\cdot \; p = P\left[\mu \in \left(L(\bar{X}), U(\bar{X})\right)\right] = P\left[L(\bar{X}) \leq \mu \leq U(\bar{X})\right] =$$

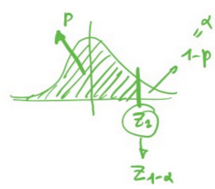
$$= P\left[-L(\bar{X}) \geq -\mu \geq -U(\bar{X})\right] =$$

$$= P\left[\underbrace{\frac{\bar{X}_n - L(\bar{X})}{\sqrt{\frac{\sigma^2}{n}}}}_{z_1} \geq \underbrace{\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}}_{Z} \geq \underbrace{\frac{\bar{X}_n - U(\bar{X})}{\sqrt{\frac{\sigma^2}{n}}}}_{z_2}\right]$$

$\rightarrow z_{1,2}$: QUANTILE OF Z OF ORDER $\frac{1-p}{2}$

$P[z_1 \geq Z \geq z_2] = p$





$$\cdot \; z_1 = \frac{\bar{X}_n - L(\bar{X})}{\sqrt{\frac{\sigma^2}{n}}}$$

ANALOG FOR z_2

$$\rightarrow L(\bar{X}) = \bar{X}_n - z_1 \sqrt{\frac{\sigma^2}{n}}$$

$$\rightarrow U(\bar{X}) = \bar{X}_n + z_2 \sqrt{\frac{\sigma^2}{n}}$$

$$\rightarrow I(\bar{X}) = \left(\bar{X}_n - z_1 \sqrt{\frac{\sigma^2}{n}} \quad , \quad \bar{X}_n + z_1 \sqrt{\frac{\sigma^2}{n}}\right)$$

- LET $\alpha = 1 - p$: PROB. THAT $\mu \notin I$

$$\rightarrow I(\bar{X}) = \left(\bar{X}_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \quad , \quad \bar{X}_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{\mu}}\right)$$

es.

LET $X \sim N(\mu, 0.04)$, $n = 100$, $p = 0.95 \to \alpha = 0.05$

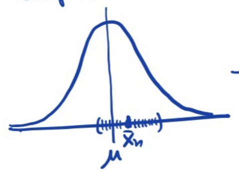
FROM SAMPLE : $\bar{X}_{100} = 19.52$

→

CI WITH $\overset{p}{\uparrow}$ 95 % : $z_{1-\frac{\alpha}{2}} = z_{0.995} \simeq 2.58$

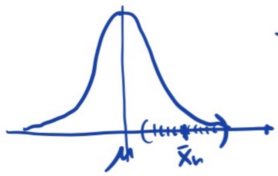
→ $I = \left(19.52 - 2.58\sqrt{\frac{0.04}{100}} \,;\, 19.52 + 2.58\sqrt{\frac{0.08}{100}}\right) \approx (19.47 \,;\, 19.57)$

About "Confidence"



→ here $\mu \in I$

this happens in 95% of cases



→ here $\mu \notin I$

this happens in 5% of cases

- I IS NOT ALWAYS BILATERAL : $L(\bar{X}), U(\bar{X})$ CAN BE $\pm\infty$ IN SOME CASES

  ↳ $z_{1-\frac{\alpha}{2}} \longmapsto z_{1-\alpha}$

- $I(\bar{X}) = \left(\bar{X}_n - z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}} \,,\, +\infty\right)$
- $I(\bar{X}) = \left(-\infty \,,\, \bar{X}_n + z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}\right)$

- σ^2 UNKNOWN:

IF σ^2 IS UNKNOWN, IT CAN BE REPLACED WITH THE POINT ESTIMATE:

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

$$\rightarrow I = \left(\bar{X}_n - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2_n}{n}} \, , \, \bar{X}_n + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2_n}{n}} \right)$$

ES.

POPULATION OF ITEMS, WE ARE INTERESTED IN EXPECTED LIFETIME, CI = 90%

- FAILURE TIMES OF 50 ITEMS: $\bar{X} = (18, 20, \dots, 15)$

$\hookrightarrow \hat{\mu} = \bar{X}_{50} = 12.68 \rightarrow \hat{\sigma}^2_n = \frac{(18-12.68)^2 + \dots}{50-1} = 36$

$$\rightarrow \alpha = 0.1 \rightarrow I = \left(\bar{X}_{50} - \overbrace{Z_{0.9}}^{\text{FROM TABLE: } 1.28} \sqrt{\frac{\hat{\sigma}^2_n}{50}} \, , \, +\infty \right) =$$

$$\approx (11.7 \, , \, +\infty)$$

ES. (AS BEFORE: $\bar{X}_{50} = 12.68$, $\hat{\sigma}^2_n = 36$)

- HYPOTHESIS: $I = (11.09, 14.27) \rightarrow CI = ?$

$$\rightarrow \bar{X}_n - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2_n}{n}} = 11.09 \rightarrow Z_{1-\frac{\alpha}{2}} = 1.85$$

$\rightarrow$ (FROM TABLE) $1 - \frac{\alpha}{2} = 0.97 \rightarrow \alpha = 0.06 \rightarrow CI = 94\%$

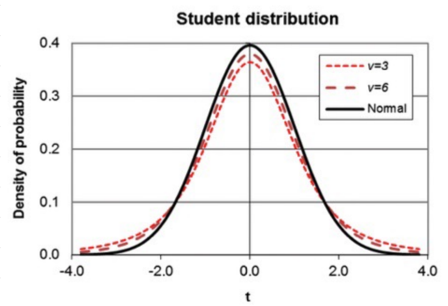
· t-STUDENT DISTRIBUTION:

WHAT IF n IS SMALL $(n < 30)$?

$\rightarrow$ THE t-STUDENT DISTRIBUTION CAN BE USED $\rightsquigarrow \Leftrightarrow X \sim N(\cdot)$

$\hookrightarrow \exists$ TABLE

· IT DEPENDS ON "ν": DEGREE $\overset{\text{DOF}}{\text{OF}}$ FREEDOM



· IF $\nu \mapsto \infty \rightarrow t_s \approx Z \sim N(0,1)$

· READ TABLE: $t_{\overset{n}{5},\ \underset{\text{ORDER OF QUANTILES}}{0.95}}$

· PROPERTY:

LET $n \in \mathbb{N}^+$, $X \sim N(\mu, \sigma^2)$, SAMPLE: $\bar{X} = (X_1, \ldots, X_n)$

$$\rightarrow T = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\hat{\sigma}^2_n}{n}}} \sim t_{n-1} \ / \ \nu = n-1 \ \text{DOF}$$

$$\rightarrow I(\bar{X}) \overset{\text{USABLE} \Leftrightarrow X \sim N(\mu, \sigma^2)}{=} \left(\bar{X}_n - t_{n-1, 1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2_n}{n}} \ ; \ \bar{X}_n + t_{n-1, 1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2_n}{n}} \right)$$

ex.

$X \sim N(\mu, \sigma^2)$, μ, σ^2 UNKNOWN

SAMPLE $(X_1, \ldots, X_n)$, $\underset{n \text{ SMALL}}{n = 4} \rightarrow (18, 20, 30, 24)$, $CI = 95\%$ ($\alpha = 0.05$)

$$\rightarrow \bar{X}_n = \frac{18 + 24 + 30 + 24}{4} = 24, \quad \hat{\sigma}^2_n = \frac{(18-24)^2 + \ldots}{4-1} = 24$$

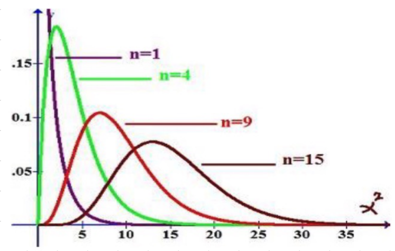
$$\cdot \ t_{n-1, 1-\frac{\alpha}{2}} = t_{3, 0.975} = 3.18$$

$$\rightarrow I(\bar{X}) = \left(24 - 3.18 \sqrt{\frac{24}{4}} \ ; \ 24 + 3.18 \sqrt{\frac{24}{4}} \right) \approx (18, 30)$$

CONFIDENCE INTERVAL FOR VARIANCE:

CHI-SQUARE DISTRIBUTION:

LET $X_1, \dots, X_n \,/\, X_i \sim N(0,1)$, X_i IID



- $X = X_1^2 + \dots + X_n^2$

$\rightarrow X \sim \chi^2_n$ $\leadsto$ CHI-SQUARED DISTRIBUTION, WITH n DOF

$\hookrightarrow$ $\exists$ TABLES

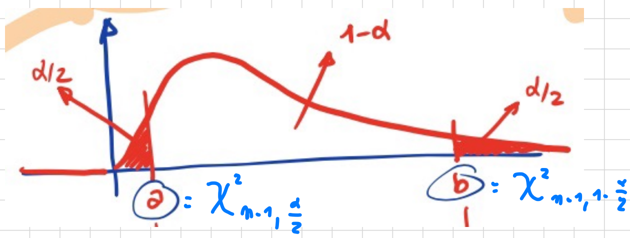
- t-STUDENT WITH n DOF: $T_n = \dfrac{Z}{\sqrt{\frac{X}{n}}} \Big/ \begin{matrix} Z \sim N(0,1) \\ X \sim \chi^2_n \end{matrix}$

- PROPERTY: MUST BE $\sim N(\cdot)$

LET $X \sim N(\mu, \sigma^2)$

$$\rightarrow \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2}{\sigma^2} \left(\cdot \frac{(n-1)}{(n-1)} \right) = \frac{(n-1)}{\sigma^2} \hat{\sigma}_n^2 \sim \chi^2_{n-1}$$

- CI: $1-\alpha$ FOR σ^2



$$I = [L(\bar{X}), U(\bar{X})]$$

- $1-\alpha = P[\sigma^2 \in (L(\bar{X}), U(\bar{X}))] = P[L(\bar{X}) \le \sigma^2 \le U(\bar{X})] =$

$\sigma^2 \hookrightarrow \dots$

$$= P\left[\frac{L(\bar{X})}{(n-1)\hat{\sigma}_n^2} \le \frac{\sigma^2}{(n-1)\hat{\sigma}_n^2} \le \frac{U(\bar{X})}{(n-1)\hat{\sigma}_n^2} \right] =$$

$$= P\left[\frac{L(\bar{X})}{(n-1)\hat{\sigma}_n^2} \le \chi^2_{n-1} \le \frac{U(\bar{X})}{(n-1)\hat{\sigma}_n^2} \right]$$

$\chi^2_{n-1, \frac{\alpha}{2}}$ $\quad$ $\chi^2_{n-1, 1-\frac{\alpha}{2}}$

$$\rightarrow U(\bar{X}) = \frac{(n-1)\hat{\sigma}_n^2}{\chi^2_{n-1, \frac{\alpha}{2}}}, \quad L(\bar{X}) = \frac{(n-1)\hat{\sigma}_n^2}{\chi^2_{n-1, 1-\frac{\alpha}{2}}}$$

- $$I = \left[\frac{(n-1)\hat{\sigma}_n^2}{\chi^2_{n-1, \frac{\alpha}{2}}}, \frac{(n-1)\hat{\sigma}_n^2}{\chi^2_{n-1, 1-\frac{\alpha}{2}}} \right]$$

es.

$X \sim N(\mu, \sigma^2)$, $n=4$: $(10.1,\ 9.9,\ 12.8,\ 7.2)$

- CI : 99% FOR σ^2

->

$$\bar{X}_4 = \frac{1}{n} \sum x_i = 10$$

$$\hat{\sigma}^2_n = \frac{1}{n-1} \sum (X_i - \bar{X}_n)^2 = \frac{(10.1-10)^2 + \dots + (7.2-10)^2}{3} = 5.23$$

- $1-\alpha = 99\% \rightarrow \alpha = 0.01$

$$\rightarrow I = \left[\frac{(n-1)\hat{\sigma}^2_n}{\chi^2_{n-1,\frac{\alpha}{2}}} ; \frac{(n-1)\hat{\sigma}^2_n}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \right] = [1.3 ; 224]$$

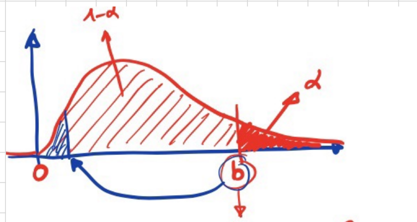
$\chi^2_{4-1,\ 0.005} = 0.07$ ($0.005 = \frac{\alpha}{2}$, $4-1 = n-1$)

$\chi^2_{4-1,\ 0.995} = 12.8$ ($0.995 = 1-\frac{\alpha}{2}$)

- WHAT IF CI : 99% FOR σ^2 / $I = (-\infty, b)$? ($-\infty \rightarrow 0$)

- $b = \dfrac{(n-1)\hat{\sigma}^2_n}{\chi^2_{n-1,\alpha}}$

$$\rightarrow I = \left[0 , \frac{(n-1)\hat{\sigma}^2_n}{\chi^2_{3,0.01}} \right] = [0, 143]$$



$\chi^2_{3,\alpha} = \chi^2_{3,0.01}$

- CONFIDENCE INTERVAL FOR DIFFERENCE OF MEANS:

GIVEN $X : \mu_X = E[X]$, $Y : \mu_Y = E[Y]$

WE ARE INTERESTED IN $d = \mu_X - \mu_Y$, $\hat{d} = \bar{X}_{n_1} - \bar{Y}_{n_2}$

- ASSUME:
  - $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$
    - $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
    - $X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$
  - SAMPLE SIZE n_X $(n_X \geq 30)$ FROM X ; $\bar{X}_{n_X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n_X}\right)$
  - SAMPLE SIZE n_Y $(n_Y \geq 30)$ FROM Y. $\bar{Y}_{n_Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_Y}\right)$

$$\rightarrow \hat{d} = \bar{X}_{n_X} - \bar{Y}_{n_Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_{n_X}^2}{n_X} + \frac{\sigma_{n_Y}^2}{n_Y}\right)$$

$$\rightarrow I = \left[\hat{d} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_{n_X}^2}{n_X} + \frac{\sigma_{n_Y}^2}{n_Y}}, \hat{d} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right]$$

- IT CAN BE USED FOR LARGE SAMPLES $(n_X, n_Y \geq 30)$
- IT CAN BE USED FOR SMALL SAMPLES IF $X, Y \sim N(\cdot)$ AND σ_X^2, σ_Y^2 ARE KNOWN

- **CONFIDENCE INTERVAL FOR $Exp(p)$:**

LET $X_i \sim Exp(\lambda)$ / $Y = \sum X_i \sim \Gamma(n, \lambda)$

- PROPERTY: $2\lambda Y \sim \chi^2_{2n}$

$$\to 1-\alpha = P\left[\chi^2_{2n,\frac{\alpha}{2}} \le 2\lambda Y \le \chi^2_{2n,1-\frac{\alpha}{2}}\right]$$

$$\to \underset{\text{FOR } \lambda}{I} = \left[\frac{\chi^2_{2n,\frac{\alpha}{2}}}{2\cdot\sum X_i} ; \frac{\chi^2_{2n,1-\frac{\alpha}{2}}}{2\sum X_i}\right]$$

- **CONFIDENCE INTERVAL FOR $Bernoulli(p)$:**

LET $X_i \sim Bernoulli(p)$ / $Y = \sum X_i \sim Bin(n,p)$

$$\to 1-\alpha = P\left[-Z_{1-\alpha/2} \le \frac{Y-np}{\sqrt{np(1-p)}} \le Z_{1-\frac{\alpha}{2}}\right] \quad / \quad \hat{p} = E[X_i] = \frac{\bar{X}_n}{n}$$

$$\to I = \left[\hat{p} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} ; \hat{p} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

- **DELTA METHOD:**

USED TO ESTIMATE A f OF A PARAMETER / $Y = \sum X_i$

- LET $g : \mathbb{R} \mapsto \mathbb{R}$ / CONTINOUS $g'(\theta)$, $g'(\theta) \neq 0$

$$\to \sqrt{n}\left(g(Y) - g(\overset{\mu}{\overset{\shortparallel}{\theta}})\right) \xrightarrow{d} N\left(0, \sigma^2 [g'(\theta)]^2\right)$$

$$\xrightarrow{CLT} g(\bar{X}_n) \xrightarrow{d} N\left(g(\mu), \frac{\sigma^2 [g'(\theta)]^2}{n}\right)$$

$$\to 1-\alpha = P\left[-Z_{1-\frac{\alpha}{2}} \le \frac{g(\bar{X}_n) - g(\mu)}{\sqrt{\frac{\sigma^2[g'(\theta)]^2}{n}}} \le Z_{1-\frac{\alpha}{2}}\right]$$

es.

A ; SAMPLE $n_X = 50$ MESURAMENTS -> $\bar{X}_A = 10.5$, $\hat{\sigma}^2_A = 25$

B ; SAMPLE $n_Y = 100$ MESURAMENTS -> $\bar{X}_B = 11.2$, $\hat{\sigma}_B = 36$

- FIND CI : 90% OF $\mu_A - \mu_B$

->

$$\hat{d} = \bar{X}_A - \bar{X}_B = 10.5 - 11.2 = -0.7$$

- $1 - \alpha = 90\% \rightarrow \alpha = 0.1$

$$\rightarrow I = \left[-0.7 - Z_{0.95} \sqrt{\frac{25}{50} + \frac{36}{100}} \; ; \; -0.7 + Z_{0.95} \sqrt{\frac{25}{50} + \frac{36}{100}} \right] =$$

$$= [-2.22 \, , \; 0.87]$$

es.

SAMPLE $n = 9$ FROM $X \sim N(\mu, \sigma^2)$

ASSUME $\bar{X}_9 = 0.5$, $\hat{\sigma}^2_n = 2.16$

a) FIND C.I. FOR μ , $\sigma^2 = 4$

b) " σ^2 UNKNOWN

c) " FOR σ^2

a) SMALL SAMPLE, σ^2 KNOWN -> CI FOR μ

$$I = \left[\bar{X}_n - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \, , \; \bar{X}_n + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right]$$

- CI ; 90% -> $\alpha = 0.1$ -> $1 - \frac{\alpha}{2} = 0.95$ -> $Z_{0.95} = 1.65$

$$\rightarrow I = \left[0.5 - 1.65 \sqrt{\frac{4}{9}} \, , \; 0.5 + 1.65 \sqrt{\frac{4}{9}} \right] = [-0.6 \; ; \; 1.6]$$

b) SMALL SAMPLE, σ^2 UNKNOWN -> t- STUDENT

$$I = \left[\bar{X}_n - \overbrace{t_{9-1,\,1-\frac{\alpha}{2}}}^{1.86} \sqrt{\frac{\hat{\sigma}^2_n}{9}} \; ; \; \bar{X}_n + \overbrace{t_{9-1,\,1-\frac{\alpha}{2}}}^{1.86} \sqrt{\frac{\hat{\sigma}^2_n}{9}} \right] =$$

$$= \left[0.5 - 1.86 \sqrt{\frac{2.16}{9}} \, , \; 0.5 + 1.86 \sqrt{\frac{2.16}{9}} \right] = \left[-0.4 \, , \; 1.4 \right]$$

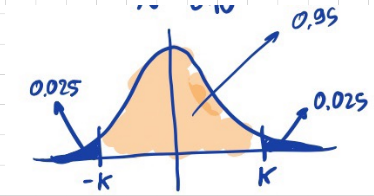
c) ... -> CI FOR σ^2 -> χ^2 DISTRIB. , $\alpha = 0.1$

$$I = \left[\frac{\overbrace{(n-1)}^{8} \overbrace{\hat{\sigma}^2_n}^{2.16}}{\chi^2_{8,\,1-\frac{\alpha}{2}}} \; ; \; \frac{(n-1)\,\hat{\sigma}^2_n}{\chi^2_{8,\,\frac{\alpha}{2}}} \right] \qquad \Big/ \qquad \chi^2_{8,\,0.05} = 2.7 \qquad \chi^2_{8,\,0.95} = 15.5$$

es.

SAMPLE $n = 11$ FROM $X \sim N(15, \sigma^2)$

WHAT IS k / $P\left[\left| \frac{\bar{X}_{11} - 15}{\sqrt{\frac{\hat{\sigma}^2_n}{11}}} \right| \le k \right] = 0.95$



-> FROM PROPERTY : $\frac{\bar{X}_{11} - 15}{\sqrt{\frac{\hat{\sigma}^2_n}{11}}} \sim t_{n-1} = t_{10}$

- $1 - \alpha = 0.95$ -> $\alpha = 0.05$

-> $k = t_{10,\,1-\frac{\alpha}{2}} = t_{10,\,0.975} = 2.23$

• HYPOTHESIS TESTING.

GIVEN AN HYPOTHESIS H_0, WE WANT TO DETERMINE IF IT CAN BE ACCEPTED OR REJECTED.

- H_0 : NULL HYPOTHESIS -> CLAIM TO BE TESTED
- H_1 : ALTERNATIVE HYPOTHESIS -> THE OPPOSITE OF H_0

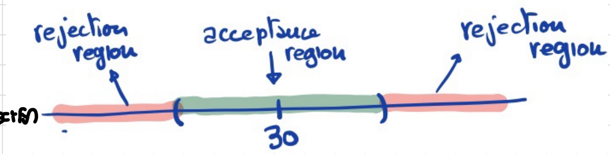
$\neq$ ACCEPTED

-> AFTER STATISTICAL TEST : H_0 REJECTED / NOT REJECTED

es.

GIVEN X, $H_0 : \mu = E[X] = 30$ $(\rightarrow H_1 : \mu \neq 30)$

- TO TEST H_0, WE CONSIDER A SAMPLE $(X_1, \dots, X_{10})$ -> $\bar{X}_{10} = 35.5$
- WE DEFINE A REJECTION REGION



  - $\bar{X}_n \in$ ACCEPTANCE REGION -> H_0 NOT REJECTED
  - $\bar{X}_n \in$ REJECTION REGION -> H_0 REJECTED

- KIND OF ERRORS:

  - I TYPE. WE REJECT H_0, BUT IT IS TRUE

MOST IMPORTANT

    - PROB. ERROR I TYPE : $\alpha = P[\text{REJECT } H_0 \mid H_0 \text{ TRUE}]$

  - II TYPE: WE ACCEPT H_0, BUT IT IS FALSE

    - PROB. ERROR II TYPE : $\beta = P[\text{ACCEPT } H_0 \mid H_0 \text{ FALSE}]$

es.

$$X \sim N(\mu, 144), \quad \sigma^2 = 144$$

WE WANT TO TEST $H_0 : \mu = 30$, $H_1 : \mu \neq 30$

- SAMPLE : $n = 10$, $\bar{X}_{10} = 35.5$

ASSUME : ACCEPTANCE REGION : $[25, 35]$, REJECTION REGION : $\notin [25, 35]$

->

- USUALLY α IS A PRE-FIXED VALUE, THEN FIND ACCEPTANCE REGION
- BUT :

$$\alpha = P[\text{REJECT } H_0 \mid H_0 \text{ TRUE}] = 1 - P[\text{ACCEPT } H_0 \mid H_0 \text{ TRUE}] =$$

$$= 1 - P[\bar{X}_{10} \in [25, 35] \mid X \sim N(30, 144)] =$$

$$= 1 - P[25 \leq \bar{X}_{10} \leq 35 \mid X \sim N(30, 144)] = \quad \bar{X}_n \sim N\left(30, \frac{144}{n}\right)$$

$$= 1 - P\left[\frac{25-30}{\sqrt{\frac{144}{10}}} \leq \frac{\bar{X}_{10} - 30}{\sqrt{\frac{144}{10}}} \leq \frac{35-30}{\sqrt{\frac{144}{10}}}\right] =$$

$$= 1 - P\left[\frac{-5}{\sqrt{14}} \leq Z \leq \frac{5}{\sqrt{14}}\right] = 0.18$$

- β :
  - IN GENERAL HARD TO COMPUTE

es. $H_0 : \mu = 30$, $H_1 : \mu = 32$

$$\to \beta = P[\text{ACCEPT } H_0 \mid H_1 \text{ TRUE}]$$

- HERE H_1 TRUE $\cong \mu \neq 30 = 32$ -> ONLY 1 VALUE

-> BUT IN GENERAL $H_1 : \bar{H}_0 = \mu \neq 30$ -> H_1 TRUE DIFFICULT...

↳ $\beta = \sup_{\mu \neq 30} \beta(\mu)$ -> POWER OF TEST (NOT COVERED)

es.

$$X \sim N(\overset{\text{UNKNOWN}}{\mu, \sigma^2})$$

$H_0 : \mu = 50$, $H_1 : \mu \neq 50$, $\alpha = 0.05$, SAMPLE : $n = 100$, $\bar{X}_{100}$

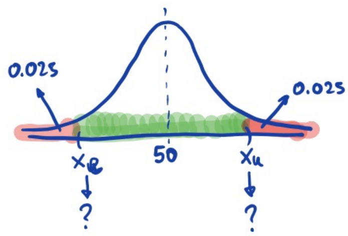
-> $\bar{X}_{100} \sim N\left(50, \frac{\hat{\sigma}^2}{100}\right)$; $\sigma^2 \mapsto \hat{\sigma}^2$, $n \geq 30$

• REJECTION AN ACCEPTANCE REGIONS:

$$\alpha = P[\text{REJECT } H_0 \mid H_0 \text{ IS TRUE}] = P[\text{REJECT } H_0 \mid X \sim N(50, \sigma^2)] =$$

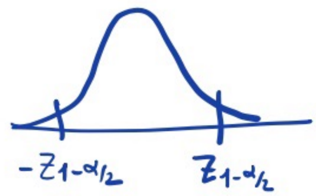
$$= 1 - P\left[\text{ACCEPT } H_0 \mid \bar{X}_{100} \sim N\left(50, \frac{\hat{\sigma}^2}{100}\right)\right] =$$

$$= 0.05$$



• $$P\left[x_l \leq \bar{X}_{100} \leq x_u \mid \bar{X}_{100} \sim N\left(50, \frac{\hat{\sigma}^2}{100}\right)\right] = 0.95$$

$$\to P\left[\frac{x_l - 50}{\sqrt{\frac{\hat{\sigma}^2}{100}}} \leq \overbrace{\frac{\bar{X}_{100} - 50}{\sqrt{\frac{\hat{\sigma}^2}{100}}}}^{Z} \leq \frac{x_u - 50}{\sqrt{\frac{\hat{\sigma}^2}{100}}}\right] = 0.95$$



• $$z_{1-\frac{\alpha}{2}} = \frac{\bar{X}_n - 50}{\sqrt{\frac{\hat{\sigma}^2}{100}}} \to z_{0.975} = 1.96$$

-> ACCEPTANCE REJON : $\left[\overset{\mu}{50} - 1.96\sqrt{\frac{\hat{\sigma}^2}{100}} , \overset{\mu}{50} + 1.96\sqrt{\frac{\hat{\sigma}^2}{100}}\right]$

• FROM SAMPLE : $\bar{X}_{100} = 55.2$, $\hat{\sigma}^2_{100} = 400$

-> ACCEPTANCE REJON $\Big|_{\hat{\sigma}^2 = 400} = [45.84, 54.16]$

-> $H_0 \notin$ ACCEPTANCE REGION -> H_0 REJECTED

• **N.B.:**

ACCEPTANCE REGION $\equiv \overset{CI}{I}\Big|_{\bar{X}_n \mapsto \mu_0}$ μ_{H_0}

- IT IS EASIER TO CONSIDER ;

$$\frac{\bar{X}_{100} - \mu_0}{\sqrt{\frac{\hat{\sigma}_n^2}{n}}} \sim Z = N(0,1) \quad \text{IF} \quad H_0 : \mu = \mu_0 \text{ IS TRUE}$$

→ IF IT IS CLOSE TO 0 → H_0 ACCEPTED , REJECTED OTHERWISE

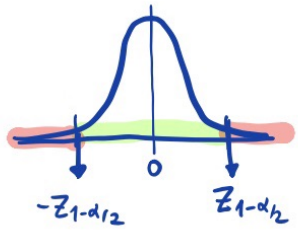
- TEST FOR MEAN :

  - $H_0 : \mu = \mu_0 \quad , \quad H_1 : \mu \neq \mu_0$

  - LARGE SAMPLE, σ^2 KNOWN :

    USE $$Z_n = \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \quad \text{UNDER } H_0$$

    - ACCEPTANCE REGION . $\left[-Z_{1-\frac{\alpha}{2}} , Z_{1-\frac{\alpha}{2}} \right]$



  - LARGE SAMPLE, σ^2 UNKNOWN :

    USE $$Z_n = \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{\hat{\sigma}_n^2}{n}}} \sim N(0,1) \quad \text{UNDER } H_0$$

  - SMALL SAMPLE, σ^2 KNOWN , $X \sim N(\mu, \sigma^2)$:

    USE $$Z_n = \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \quad \text{UNDER } H_0$$

  - SMALL SAMPLE , σ^2 UNKNOWN , $X \sim N(\mu, \sigma^2)$:

    USE $$T_{n-1} = \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{\hat{\sigma}_n^2}{n}}} \sim t_{n-1} \quad \text{UNDER } H_0$$

    - ACCEPTANCE REGION . $\left[-t_{n-1, 1-\frac{\alpha}{2}} \; ; \; t_{n-1, 1-\frac{\alpha}{2}} \right]$

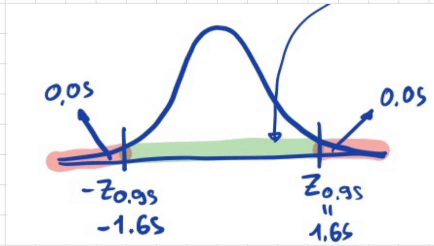
es.

$H_0: \mu = 1.5$, $H_1: \mu \neq 1.5$, $\alpha = 0.1$

- SAMPLE: $n = 40$, $\bar{X}_n = 1.85$, $\hat{\sigma}^2_n = 4$

→ LARGE SAMPLE, UNKNOWN VARIANCE

- $$Z_{40} = \frac{\bar{X}_n - 1.5}{\sqrt{\frac{\hat{\sigma}^2_n}{40}}} = \frac{1.85 - 1.5}{\sqrt{\frac{4}{40}}} = 1.1$$



- $1.1 \in$ ACCEPTANCE REGION → H_0 NOT REJECTED

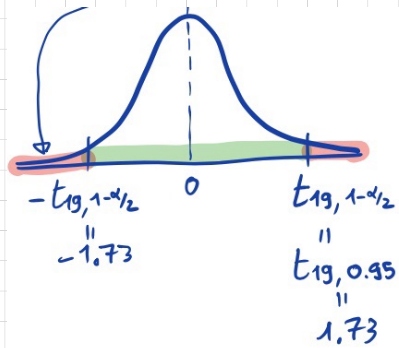
es.

$X \sim N(\mu, \sigma^2)$, $H_0: \mu = 50$, $\alpha = 0.1$

- SAMPLE: $n = 20$, $\bar{X}_n = 1.85$, $\hat{\sigma}^2_n = 4.2$

→ SMALL SAMPLE, σ^2 UNKNOWN

- $$T_{19} = \frac{\bar{X}_n - 50}{\sqrt{\frac{\hat{\sigma}^2_n}{20}}} = -2.7$$



- $-2.7 \notin$ ACCEPTANCE REGION → H_0 REJECTED

- **TEST FOR VARIANCE** (FOR $X \sim N(\mu, \sigma^2)$):

  - $H_0: \sigma^2 = \sigma_0^2$, $H_1: \sigma^2 \neq \sigma_0^2$
  - PROPERTY: $Q_{n-1} = \dfrac{(n-1)\hat{\sigma}_n^2}{\sigma_0^2} \sim \chi^2_{n-1}$
  - ACCEPTANCE REGION = $\left[\chi^2_{n-1, \frac{\alpha}{2}} \; ; \; \chi^2_{n-1, 1-\frac{\alpha}{2}}\right]$

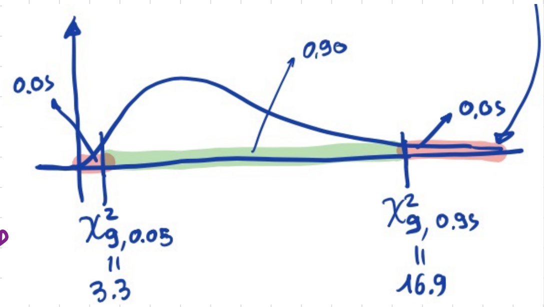
es.

$X \sim N(\mu, \sigma^2)$, $H_0.\ \sigma^2 = 1$, $H_1: \sigma^2 \neq 1$, $\alpha = 0.1$

- SAMPLE: $n = 10$, $\bar{X}_{10} = 11$, $\hat{\sigma}_n^2 = 2.55$

->

$$Q_{n-1} = \frac{(n-1)\cdot 2.55}{1} = 22.95$$



- $22.95 \notin$ ACCEPTANCE REGION -> H_0 REJECTED

- **TEST FOR DIFFERENCE BETWEEN MEANS:**

$$X: \quad E[X] = \mu_x \quad , \quad Y: \quad E[Y] = \mu_Y$$

- $H_0 : \mu_X - \mu_Y = d_0 \quad , \quad H_1 : \mu_X - \mu_Y \neq d_0$
- $\bar{X}_{n_1} - \bar{Y}_{n_2} \sim N\left(\overbrace{\mu_x - \mu_Y}^{d_0} , \frac{\sigma_x^2}{n_1} + \frac{\sigma_Y^2}{n_2} \right)$, IF n_1, n_2 LARGE

-> COMPUTE $D_{n_1, n_2} = \dfrac{(\bar{X}_{n_1} - \bar{Y}_{n_2}) - d_0}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n}}} \sim N(0,1)$ IF H_0 IS TRUE

- ACCEPTANCE REGION $= \left[-z_{1-\frac{\alpha}{2}} ; z_{1-\frac{\alpha}{2}} \right]$

es.

$X \sim N(\mu_X, 200) : n_1 = 100, \bar{X}_{100} = 12$

$Y \sim N(\mu_Y, 200) : n_2 = 100, \bar{Y}_{100} = 10$

- $H_0 : \mu_X - \mu_Y = 5 \quad , \quad \alpha = 0.05$

-> $D_{n_1, n_2} = \dfrac{(12-10) - 5}{\sqrt{\frac{200}{100} + \frac{200}{100}}} = -1.5$

- $-1.5 \in$ ACCEPTANCE REGION -> H_0 NOT REJECTED

· TEST DIFFERENCE OF MEANS, SMALL SAMPLES (σ^2 UNKNOWN):

$$X \sim N(\mu_X, \sigma_X^2),\ Y \sim N(\mu_Y, \sigma_Y^2) \ \big/ \ \sigma_X^2 = \sigma_Y^2 = \sigma^2$$

· $H_0 : \mu_X - \mu_Y = d_0$

· ESTIMATOR OF σ^2 : $\hat{\sigma}_m^2 = \dfrac{(n_X - 1)\hat{\sigma}_X^2 + (n_Y - 1)\hat{\sigma}_Y^2}{n_X + n_Y - 2}$

· PROPERTY : $T = \dfrac{(\bar{X}_{n_X} - \bar{Y}_{n_Y}) - d_0}{\sqrt{\dfrac{(n_X + n_Y)}{n_X n_Y} \cdot \hat{\sigma}_m^2}} \sim t_{n_X + n_Y - 2}$ UNDER H_0

· ACCEPTANCE REGION . $\left[-t_{n_X+n_Y-2,1-\frac{\alpha}{2}} ; t_{n_X+n_Y-2,1-\frac{\alpha}{2}} \right]$

ex.

X : LIFETIME OF COMPONENTS FROM A

Y : " B

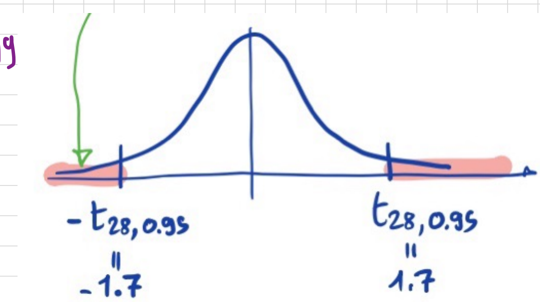
· $H_0 : \mu_X - \mu_Y = 0$, $\alpha = 0.1$

· SAMPLE X : $n_X = 10$, $\bar{X}_{10} = 10$, $\hat{\sigma}_X^2 = 1$

· SAMPLE Y : $n_Y = 20$, $\bar{Y}_{20} = 11$, $\hat{\sigma}_Y^2 = 2$ ~> SMALL SAMPLE

->

$$T = \frac{(10 - 11) - 0}{\sqrt{\dfrac{(10+20)}{200} \cdot \dfrac{(10-1)1 + (20-1)2}{(10+20-2)}}} = -1,99$$



· $-1,99 \notin$ ACCEPTANCE REGION -> H_0 REJECTED

- UNILATERAL TESTS :

$$H_0 : \mu \le \mu_0 \quad , \quad H_1 : \mu > \mu_0$$

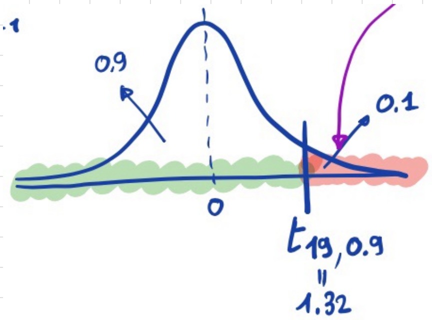
es.

LET $X \sim N(\mu, \sigma^2)$ — UNKNOWN

- $H_0 : \mu \le 12$, $H_1 : \mu > 12$, $\alpha = 0.1$
- SAMPLE : $n = 20$ (SMALL) , $\bar{X}_{20} = 12.5$, $\hat{\sigma}^2_n = 2$

->

$$T_{n-1} = \frac{\bar{X}_n - 12}{\sqrt{\frac{\hat{\sigma}^2_n}{n}}} = \frac{12.5 - 12}{\sqrt{\frac{2}{20}}} = 1.58$$

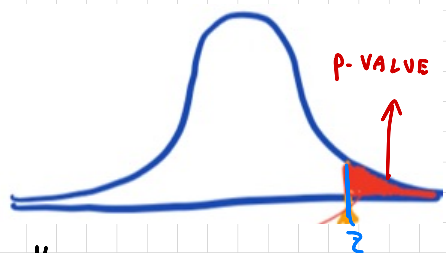


- $1.58 > 1.32$ -> H_0 REJECTED

es. INFECTION BY DISEASE (13/17)

...

- FOR STATISTICAL SOFTWARE :



- ACCEPT / REJECT $\longmapsto$ p-VALUE

- p-VALUE : AREA ON THE RIGHT OF $z = \frac{x - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$ / $x = \bar{X}_n$

$$\to \begin{cases} \text{p-VALUE} < \frac{\alpha}{2} \ (\text{OR } \alpha) \to H_0 \text{ REJECTED} \\ \text{p-VALUE} > \frac{\alpha}{2} \ (\text{OR } \alpha) \to H_0 \text{ ACCEPTED} \end{cases}$$

-> p-VALUE (FOR BILATERAL) $= 2 \cdot z$

(z IF UNILATERAL)

- **LIKELIHOOD RATIO TEST:**

GIVEN $H_0: \hat{\theta} \in \Theta_0$, $H_1: \hat{\theta} \in \bar{\Theta}_0$ $\rightarrow$ $\lambda(\bar{x}) = \dfrac{\sup_{\Theta_0} L(\theta|\bar{x})}{\sup_{\bar{\Theta}_0} L(\theta|\bar{x})}$ ↓ MAX LIKELIHOOD ESTIMATOR

- $\alpha = P[\underbrace{\lambda(\bar{x}) < \lambda^*}_{\text{REJECTION REGION}} \mid H_0 \text{ IS TRUE}] \rightarrow \lambda^* = \dots$

  $\rightarrow$ FIND $\bar{x} \lessgtr \dots$

$\rightarrow$ H_0 REJECTED IF $\lambda(\bar{x}) < \lambda^*$

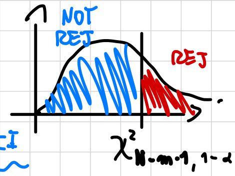
- **BAYESIAN TESTS:**

H_0 REJECT $\iff$ $\dfrac{P(\hat{\theta} \in \Theta_0 \mid \bar{x})}{P(\hat{\theta} \in \Theta_1 \mid \bar{x})} < c$ / IF $c = 1$ $\rightarrow$ MAP TEST

- **GOODNESS OF FIT:** $\rightarrow$ TEST FOR A DISTRIBUTION

  - $H_0: X \sim f = f_0$, $H_1: X \sim f \neq f_0$

  - **χ^2 GOF:**



  - THEOREM, $W = \sum_{i=1}^{N} \dfrac{(f_i - np_i)^2}{np_i} \sim \chi^2_{N-m-1,\ 1-\alpha}$ / $p_i = P[X \in I_i \mid f_0 \text{ TRUE DENSITY}]$

    f_i: # OUTCOME OCCURRED I_i; IF $np_i \geq 5,\ \forall i$; CI $1-\alpha$

  $\rightarrow$ IF $W < \chi^2_{m-1,\ 1-\alpha}$ $\rightarrow$ H_0 NOT REJECTED

  - SAME CAN BE APPLIED IF f_i FROM A SAMPLE $X = (X_1, \dots, X_n)$

    $\rightarrow$ CHECK $W < \chi^2_{N-m-1,\ 1-\alpha}$ / N: n° INTERVALS

    m: # PARAM. ESTIMATED THROUGH THE SAMPLE

- **TEST FOR INDEPENDENCE:**

TEST H_0: X, Y INDEPENDENT IF $n\hat{p}_i\hat{q}_j \geq 5,\ \forall i,j$

- $W = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \dfrac{(f_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j} \sim \chi^2_{(N_X-1)(N_Y-1),\ 1-\alpha}$

- CHECK H_0 AS IN χ^2 GOF

/ $\hat{p}_i$ = OBSERVED $P[X \in I_i^X]$, $\hat{q}_j$ ANALOG...