

Measure? ← why?

$$\text{Eg: } P[2^{\text{nd}} \text{ year}] = 10\% = \frac{10}{100} \leftarrow \begin{array}{l} \# \text{ of second year student} \\ \text{← all people sitting in the room} \end{array}$$

You are measuring something.

Axioms

- ① $P[A]$ has to be non-negative
- ② $P[\emptyset] = 0$
- ③ $P[S] = 1$

$$\text{Ex } P[S] = P[\text{uni students}] = 1$$

$$\text{④ } P[\text{Additive}] = P[A_1 \cup A_2 \cup \dots] = P[A_1] + P[A_2] + \dots$$

Eg $A_1 = \text{first year}$ } disjoint because
 $A_2 = \text{second year}$ } nothing in common

$$\begin{aligned} 3 &= HHH \\ 2 &= HHT \\ 2 &= HTH \\ 1 &= HTT \\ 2 &= THH \\ 1 &= THT \\ 1 &= TTH \\ 0 &= TTT \end{aligned} \quad \left. \begin{array}{l} \\ \\ \\ \\ \\ \\ \\ \end{array} \right\} \text{sample space} \quad \begin{aligned} P[H] &= \frac{1}{2} \\ P[T] &= \frac{1}{2} \\ P[3H] &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} \end{aligned}$$

Random variable

Maps sample space

$$S \rightarrow \mathbb{R} \quad , \quad X = \# \text{ of heads}$$

X_i :	0	1	2	3
$P[X_i]$:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

/ or add probabilities

↑

table called pmf or probability mass functions.

$$\text{To find } E(X) = \sum x p[x] \rightarrow \text{discrete}$$

$$\int x p[x] dx \rightarrow \text{continuous}$$

$$\text{To find } E[X^2] = \sum x^2 p[x] \rightarrow \text{discrete}$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 p[x] dx \rightarrow \text{continuous.}$$

Bernoulli:

$$p_x(x) = \begin{cases} \theta & \\ 1 - \theta & \end{cases}$$

$$\left| \begin{array}{l} E_x / x = \# \text{ of H} \leq 0 \\ P[H] = \theta \end{array} \right.$$

$$P[T] = 1 - \varnothing$$

Poisson mean and variance is exactly the same \Rightarrow ✓

Continuous &

Uniform, Exp, gamma, normal, Beta.

Next class review function, mean, cdf, pdf etc...

		Marginal prob example						
		1	2	3	4	5	6	
A	B	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	\vdots
		$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

$A = H$;
 $B = I$ is $A \perp B$? Yes

Proof $P[\text{Joint } (A \cap B)] = \frac{1}{12}$

$$P(A) = \frac{1}{2}$$

$$P(B) = \frac{1}{6}$$

$$P(A \cap B) = P(A) P(B) \quad \text{if } A \perp B$$

$$P[1|H] = \frac{P[1 \cap H]}{P(H)} = \frac{\frac{1}{12}}{\frac{1}{2}} = \frac{1}{6}$$

x:	0	1	2	3
P[x]:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Treatment Heart Transplant

Control group

P1 - survived 42 days after died

P28 - survived 118 days - but he is alive

P - indicator

P1 - wait time for transplant is 0 days \rightarrow survived 15 days - dead

P52 - wait time 5 days \rightarrow survived 43 days - alive

$$30 \text{ ppl} \rightarrow \text{mean}(t)$$

Treatment

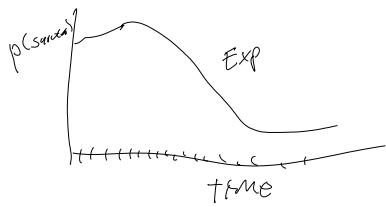
$$\text{Control}$$

P

$$52 \text{ ppl} \rightarrow \text{mean}(c)$$

compare the two values.

Assume SD is coming from a distribution
e.g.



if mean of treatment group is bigger \Rightarrow mean of control group

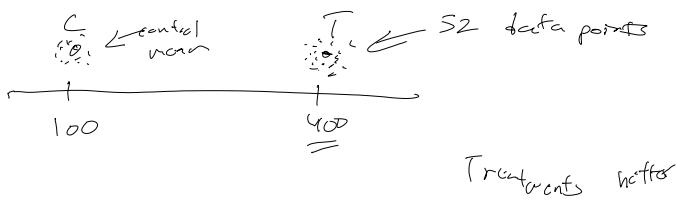
Interpretation about the population:

\Rightarrow if you go through the treatment you are expected to live longer.

$\text{mean}(T) > \text{mean}(c)$ are samples, interpretation we want to make is about the entire population.

assume

- ① $X \sim \text{Exp}(\lambda) \rightarrow$
② estimate of λ from these
fits (parameters)

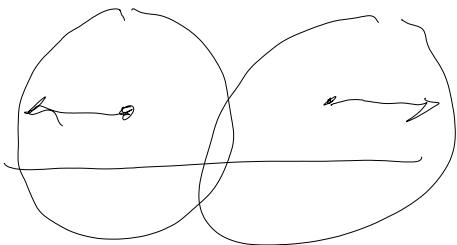


In real life you never know the distribution

$$\left. \begin{array}{l} \text{Recall} \\ X \sim \text{exp}(x) \end{array} \right\}$$

$$f_x(x) = \lambda e^{-\lambda x}$$

mean is the center of the distribution



Data - Circle

mean - center of data

standard deviation - radius

look if they are overlapping or far apart.

$$\begin{aligned} 5.1.1 \quad \text{mean}(c) &= 93.2 \\ &= \frac{\text{sum (30 numbers)}}{30} \end{aligned}$$

$$\text{mean}(t) = 356.2$$

you could just compare the mean, look at SD and then mean.

Ex 5.1.8

$$X \sim \text{Exp}(\lambda)$$

$$x_1, x_2, \dots, x_n$$

$$E(x) = \int_0^\infty x f(x) dx$$

$$= \int_0^\infty x \lambda e^{-\lambda x} dx$$

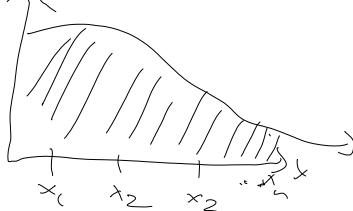
$$= \lambda \int_0^\infty x e^{-\lambda x} dx$$

difference between x and x_i is

$$E(x) - x = 2$$

$$f_{x|>2} = 2e^{-2x}$$

$$f(x)$$



$$= \lambda \int_0^{\infty} x e^{-\lambda x} dx$$

$$= \frac{\lambda}{\Gamma(2)} \int_0^{\infty} x^{2-1} e^{-\lambda x} dx \cdot \frac{\lambda^2}{\Gamma(2)} \leftarrow \text{divide by } \Gamma(2)$$

gamma density = 1

$$= \frac{1}{\lambda} (1) = \frac{1}{\lambda}$$

weak law of large #'s says

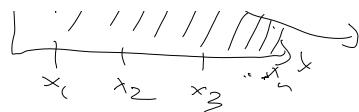
$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \xrightarrow{\lambda} \frac{x_1 + x_2 + \dots + x_n}{n} \xrightarrow{\text{converge}} \frac{1}{\lambda} \xrightarrow{\lambda \approx 0} \frac{1}{\lambda}$$

\Rightarrow

$$= \bar{x} \xrightarrow{\lambda} \frac{1}{\lambda} \leftarrow \text{estimate of mean}$$

$$= \frac{1}{\bar{x}} \xrightarrow{\lambda} \lambda \quad \uparrow$$

Estimate of lambda



can take any value x or y . x is the random variable that can take any value

x_i is sample value
 $x_1 = 100$ days
 $x_2 = 200$ days

Recall: Gamma

$$y \sim G(\alpha, \beta)$$

$$f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, y \geq 0$$

problems

$$\bar{x} = 0.05 \quad \hat{\lambda} = 20 \quad \text{Estimate}$$

$$\bar{x} = 0.6 \quad \hat{\lambda} = 16.66$$

if λ is really big eg 200 to contain it you need a large sample size

R-vector: $c(1, 3, 5) \Rightarrow$

1
3
5

$$c(1, 3, 5) + 2 = \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix}$$

variable: $x = c(5, 7, 9, 10)$

printing: $>x$
 $[1] 5 > 9 10$

$$\log(x) \Rightarrow \log \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

1 2 3

$$\log(x) \Rightarrow \log \begin{pmatrix} 5 \\ 7 \\ 9 \\ 10 \end{pmatrix}$$

1.60903 1.10510...

$\log_{10}(x)$

$\sin(x)$

$\cos(x)$

x^2

does it for all others

$y =$

5.1.1 / $x \sim N(0, 1)$

$$y = x + 2x^3 - 3$$

$P(y \in (1, 2)) \leftarrow$ probability between (1, 2)

$$x^2 \rightarrow \text{hi - sq}$$

$$-2\pi \quad \quad \quad 2\pi$$

Ex 5.2.1

x = life length of a machine

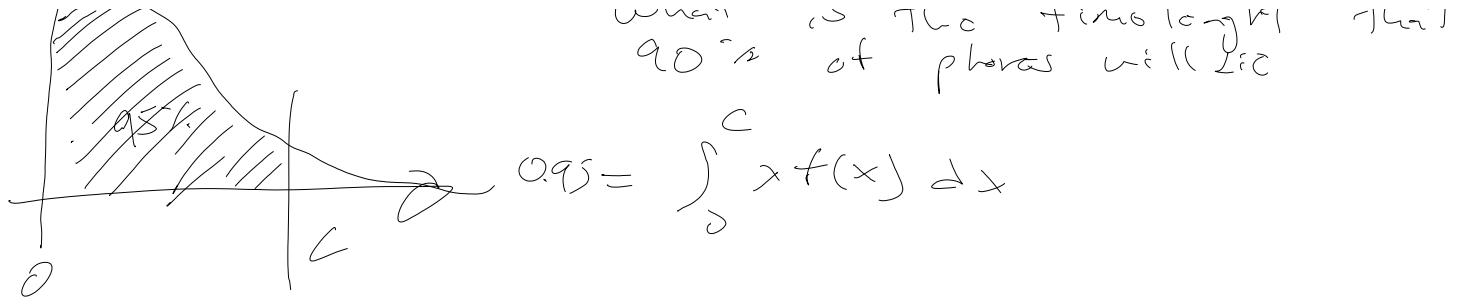
$$x \sim Ex_p(1)$$

mean life length \Rightarrow asking for $E(x)$

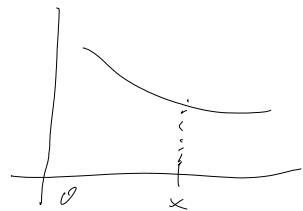
$$Ex_p(\lambda) = \frac{1}{\lambda} \Rightarrow Ex(1) = \frac{1}{1} = 1$$



what is the time length plant 90% of phones will last



$$F(x) = \int_0^x f(x) dx \leftarrow cdf$$



$$1 - e^{-c} = 0.95$$

$$e^{-c} = \frac{0.95}{1}$$

$$\ln(e^{-c}) = \ln(0.05)$$

$$c = -\ln(0.05)$$

Conditional Distribution

X:	0	1	2	3
$P[X=x]$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$\text{mean} = (0)\left(\frac{1}{8}\right) + 1\left(\frac{3}{8}\right) + 2\left(\frac{3}{8}\right) + 3\left(\frac{1}{8}\right) \\ = 1.5$$

Is this valid Probability Distribution?

Yes because

- 1) $0 \leq P(X=x) \leq 1$
- 2) $\sum P[X=x] = 1$

Condition: cell phone survived 6 month then you won't have the same distribution

X:	0	1	2	3
$P[X=x]$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

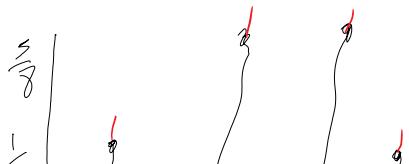
Then 0 isn't an option

This is not a valid probability distribution since it does not add up to 1.

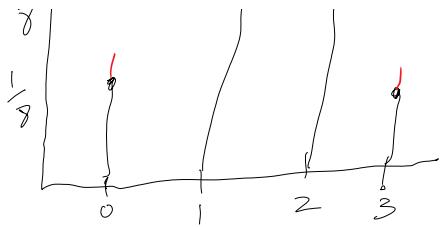
First make it a valid probability distribution

$$P(X=1 | X \geq 0)$$

$$\frac{P(X=1 \cap X \geq 0)}{P(X \geq 0)} = \frac{P(X=1)}{P(X \geq 0)} = \frac{\frac{3}{8}}{\frac{7}{8}} = \frac{3}{7}$$



make it a bit bigger
in. down -



makes it a hit bigger
for $\text{PPF} = 1$

if $x \sim \text{Exp}(1)$

PDF: $f_x(x) = e^{-x}$

$$E[x | x > 1] = \int x, \text{ conditional function}$$

$$f_{x|x>1}(x) = \frac{f_x(x)}{P[x > 1]} \leftarrow \begin{matrix} \text{condition needed} \\ \text{for } \sum \text{ to } = 1 \end{matrix}$$

$$= \frac{e^{-x}}{\int_1^\infty e^{-x} dx}$$

$$\Rightarrow = \int_1^\infty x e^{-(x-1)} dx$$

Review S.1

Example \rightarrow Samples

\hookrightarrow Distribution

\hookrightarrow Inference \rightarrow commenting on the calculation

* If $x \sim \text{life length}$

what is mean/expected life length?

\hookrightarrow asking for expected value $E(x)$

\hookrightarrow by what time 95% of the machines

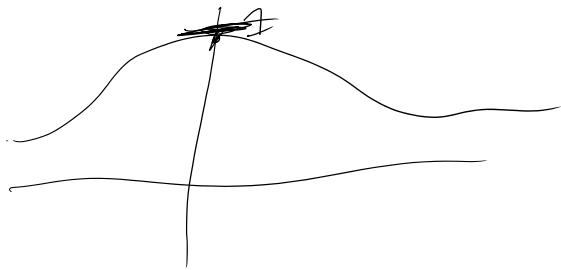
↳ by what time 95% of the machines will fail.

calculate cdf, equal it to 0.95 or value given and solve it.

2nd question: conditional, given $x > 1$
↳ mean / expected life?

$$E(x | x > 1)$$

product life time distribution; mean or mode



mode = highest point

To find # failure derivative
set it = 0

max min

5.2.7 example where you have to calculate the mode.

5.2.8 $x \sim P(X)$

* Predict the future values? $\begin{cases} \text{mean} \\ \text{mode} \end{cases}$

mode

$$P[X=x] = \frac{\lambda^x e^{-\lambda}}{x!}, \text{ if discrete no derivative}$$

$$P[X=x+1] = \frac{\lambda^{x+1} e^{-\lambda}}{(x+1)!}$$

$$\frac{P[Y=x+1]}{P[X=x]} = \frac{\cancel{\lambda}^{x+1} \cancel{e^{-\lambda}}}{(x+1)!} \div \frac{\cancel{\lambda} \cancel{e^{-\lambda}}}{x!}$$

$$= \lambda \frac{x!}{(x+1)!} = \frac{\lambda}{x+1}$$

$$\Rightarrow \frac{P[Y=x+1]}{P[X=x]} = \frac{1}{x+1}$$

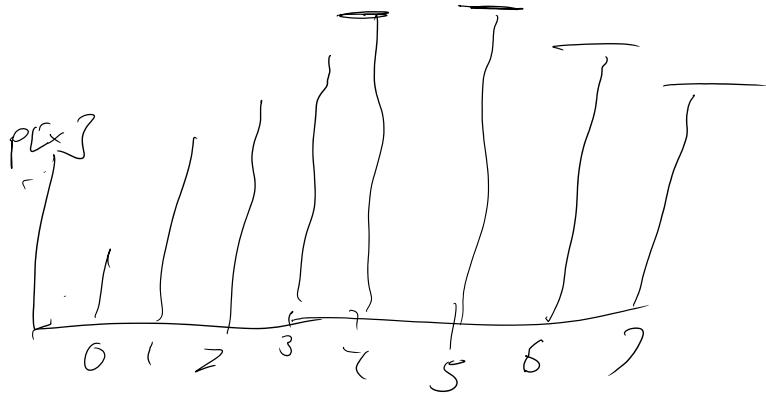
$$X=0 \Rightarrow \frac{P[X=1]}{P[X=0]} = \lambda \rightarrow \text{if } \lambda = 5 \text{ then its Sx ratio}$$

$$X=1 \Rightarrow \frac{P[X=2]}{P[X=1]} = \frac{\lambda}{2} \rightarrow \text{if } \lambda = 5 \text{ then its Sx ratio}$$

$$\frac{P[X=3]}{P[X=2]} = \frac{5}{3} = 1.666$$

$$\frac{P[X=4]}{P[X=3]} = \frac{5}{4} = 1.25$$

$$\frac{P[x=5]}{P[x=4]} = \frac{5}{5}$$



5.3 Statistical Models

① Population: a combination of all the subjects in your subspace. All the outcome.

Sample: A small subset of the population
 $\text{Sample} \subseteq \text{Population}$

Parameter: Any characteristic of a population is a parameter

populations are too big samples are easier to work with.

Study sample \Rightarrow make inference about population. That is statistics

$$\text{Eg } f_{\lambda}(x) \quad , \quad p_{\theta}(x)$$

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x \text{ is r.v., } \lambda \text{ is the parameter.}$$

Statistical model

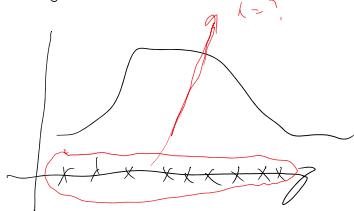
$\left\{ \begin{array}{l} p_{\theta} : \theta \in \Omega \\ \text{function} \\ \text{e.g. poisson} \\ \text{normal} \end{array} \right.$	θ can take any value in the sample space
--	---

parameter space

$$\frac{e^{-5} 5^x}{x!} \sim p_5(x)$$

$$\frac{e^{-100} 100^x}{x!} \sim p_{100}(x)$$

$$\text{Eg } x \sim p_{\lambda}(\lambda = ?)$$



$$\bar{x} \rightarrow \frac{1}{\lambda}$$

we use \bar{x} to estimate $\frac{1}{\lambda}$. This is an example of point estimation.

Interval estimation - Review

Example: Pg 264

Ex 5.3.2

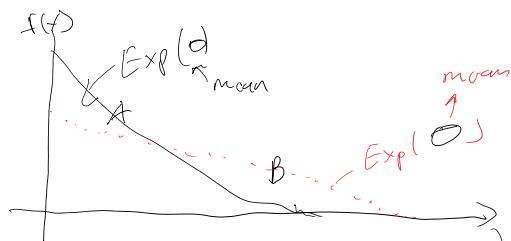


$$P_\theta(x) = f_\theta(x)$$

\curvearrowleft discrete continuous

Sample (x_1, x_2, \dots, x_5)

In this case, we know the parameters as the mean is already there.



\curvearrowleft we should be able to distinguish between if a function has a mean of 2 or 1.5.

$$\textcircled{1} (x_1, \dots, x_5) = (5.0, 3.5, 3.3, 4.1, 2.8)$$

$$\textcircled{2} (x_1, \dots, x_5) = (2.0, 2.5, 3.0, 3.1, 3.0)$$

\curvearrowleft which one belongs with which sample?

$$\textcircled{1} \Rightarrow \text{Exp}(1.5)$$

$$\textcircled{2} \Rightarrow \text{Exp}(1)$$

Parameter space in this case is

$$\Pi = \{1, 1.5\}$$

$$\Pi = \{A, B\}$$

Ex 5.3.1

- \curvearrowleft
- ① Exp \curvearrowleft know function
 - ② $\Pi = \{1, 1.5\}$ \curvearrowleft know parameter space
 - ③ Sample \curvearrowleft select sample and pick

Real life

- ① observe sample
- ② find the distribution
- known or assumed

② Sample \leftarrow select sample and pick θ - known or assumed

③ Parameter - Use sample to make inference about parameter

5.3.7 parameter

θ	$P_\theta(x=1)$	$P_\theta(x=2)$	$P_\theta(x=3)$
A	0.5	0.5	0
B	0	0.5	0.5

\rightarrow space $\{A, B\}$

b) if we observe $x=1$ then x is coming from A

if we observe $x=3$ then x is coming from B

if we observe $x=2$ then either A or B.

parameter

θ	$P_\theta(x=1)$	$P_\theta(x=2)$	$P_\theta(x=3)$
A	λ_1	λ_2	0
B	0	0.5	0.5

what if $x=2$?

probably A

Notation

$$\{f_\theta : \theta \in \Omega\}$$

for one ^{sample} value what is the statistic model

$$\text{Ex: } \hat{P}_{\text{obs}}(x)$$

$$\text{Ex: } x \sim \text{Exp}(\theta)$$

$$x = L$$

$$\sum \frac{\hat{P}_{\text{obs}}(x)}{2!}$$

$$f_\theta(x) = \theta e^{-\theta x}$$

$$x=5 = \theta e^{-\theta(5)} \leftarrow \text{statistical model of one sample.}$$

Joint density of sample

$$\text{Ex: } p[x_1, x_2, x_3, \dots, x_n] \text{ if I break it down}$$

$$\Rightarrow f_\theta(x_1) f_\theta(x_2) \dots f_\theta(x_n)$$

I is an assumption

$$x = \{2, 4, 9\}$$

$$x = \{2, 4, 9\}$$

$$\theta e^{-2\theta} \times \theta e^{-k\theta} \times \theta e^{-q\theta}$$

Ex 5.3.3

$$x \sim B(n, \theta)$$

$$p_X(x) = \begin{cases} 1 & , \theta \\ 0 & , 1-\theta \end{cases}$$

$$p_X(x) = \theta^x (1-\theta)^{n-x}$$

$$\theta^x (1-\theta)^{n-x} = \theta^{x_1 + x_2 + \dots + x_n} (1-\theta)^{n - (x_1 + x_2 + \dots + x_n)}$$

$$= \theta^{x_1 + x_2 + \dots + x_n} (1-\theta)^{n - (x_1 + x_2 + \dots + x_n)}$$

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

Ex 5.3.4

$$(x_1, \dots, x_n) \sim N(\mu, \sigma^2)$$

$$\theta = (\mu, \sigma^2) \in \mathbb{R}^1 \times \mathbb{R}^+, \quad \mathbb{R}^+ = (0, \infty)$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_1-\mu)^2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_2-\mu)^2} \cdot \dots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}$$

$$\begin{aligned}
 &= \sum_{i=1}^n (x_i - \mu)^2 \quad \begin{matrix} a-b \\ a-c+c-b \end{matrix} \\
 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2 \\
 &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - \mu) \sum_{i=1}^n (x_i - \bar{x}) + \sum_{i=1}^n (\bar{x} - \mu)^2 \\
 &\quad \begin{matrix} \cancel{\sum_{i=1}^n (x_i - \bar{x})} \\ = \sum x_i - \sum \bar{x} \end{matrix} \quad \text{sum of deviations from mean} = 0 \\
 &= n\bar{x} - n\bar{x} = 0
 \end{aligned}$$

$$\sum_{i=1}^n (\bar{x} - \mu)^2 = n(\bar{x} - \mu)^2$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \sim \text{sample variance}$$

x_i	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	-2	4	16
4	0	4	16
6	2	0	0
	0	8	8

$$\begin{aligned}
 s^2 &= \frac{\sum (x_i - \bar{x})^2}{n-1} \\
 &= \frac{8}{2} = 4
 \end{aligned}$$

$\frac{1}{\text{Var}} = \text{precision}$

$$\lambda = \frac{1}{6^2}$$

precision

$$\begin{aligned}
 &(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} (x-\mu)^2\right] \\
 &= (2\pi)^{-\frac{1}{2}} (\lambda)^{\frac{1}{2}} \exp\left[-\frac{1}{2} \lambda (x-\mu)^2\right] \quad \leftarrow \text{still normal but in precision instead of variance}
 \end{aligned}$$

Reparameterization - only do it if it's a one to one function. Precision is \perp variance. The change old parameter to new parameter has to be one to one.

old param $\xrightarrow{\text{func}}$ new param.

Tues Oct 3. 3.5.

Population

Population CDF

$$F_X(x) = \frac{\text{Count of } X \leq x}{N}$$

Ex 5.4.1

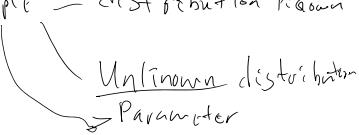
$$N = 20$$

$$\underline{\text{min}} = 3$$

$$P(X \leq 3) = 0$$

$$P(X \leq 4) = \frac{3}{20}$$

$$P(X \leq 4) - P(X \leq 3) = P(X=3)$$

5.4 = Population = finite
= Sample — distribution known

Parameter

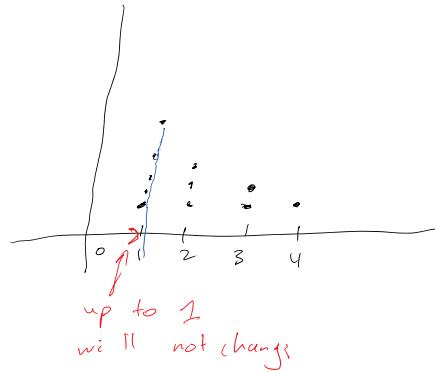
If a population is finite, do you need money/time to study them?

Ex

i	1	2	3	4	5	6	7	8	9	10
X(n _i)	1	1	2	1	2	3	3	1	2	4

Sort

i	1	1	2	1	2	3	3	1	2	4
X(n _i)	1	1	1	1	2	2	3	3	4	



$$P[X \leq 0] = 0$$

$$P[X \leq 0.999] = 0$$

$$P[X \leq 1] = \frac{4}{10} = 0.4$$

$$P[X \leq 1.999] = \frac{4}{10} = 0.4$$

$$P[X \leq 2] = \frac{7}{10} = 0.7$$

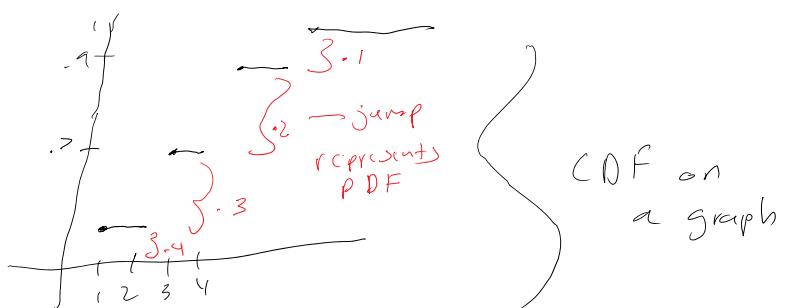
$$P[X \leq 2.99] = \frac{7}{10} = 0.7$$

$$P[X \leq 3] = \frac{9}{10} = 0.9$$

$$P[X \leq 3.99] = 0.9$$

$$P[X \leq 4] = 1$$

$$CDF = F_X(x) = \begin{cases} 0 & x < 1 \\ 0.4 & 1 \leq x \\ 0.7 & 2 \leq x < 3, 4 \\ 0.9 & 3 \leq x \leq 4 \\ 1 & x \leq 4 \end{cases}$$



PMF:

$$f_X(x) = \begin{cases} 0.4, & x=1 \\ 0.3, & x=2 \\ 0.2, & x=3 \\ 0.1, & x=4 \\ 0, & otherwise \end{cases} \Rightarrow$$

x	1	2	3	4
p[x]	0.4	0.3	0.2	0.1

To calculate small f, pdf just calculate the proportion.

$$\text{CDF} = F_x(x) = \begin{cases} 0 & x < 1 \\ .4 & 1 \leq x \\ .7 & 2 \leq x < 3 \\ .9 & 3 \leq x \leq 4 \\ 1 & x \geq 4 \end{cases}$$

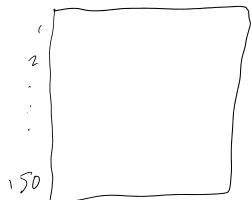
Same calculation but if it's sampled it's called Empirical distribution function

Empirical distribution: Same calculation but put $\hat{F}_x(x)$

g.f.l ex/ do it on your own drug.

Simple random Sampling:

Ex, in a class, blindly picking #'s



To do it in R: `sample(1:150, size=1)`

for a random sample between 1-150

A, B, C, D, E, Z

$N=5$ - draw 2 samples $n=2$

$$P(A) = \frac{1}{5}$$

$$P(A \text{ being selected}) = \frac{1}{N} = \frac{1}{5}$$

$$P(A \text{ being selected} | b \text{ is selected}) = \frac{1}{N-1} = \frac{1}{4} = 0.25$$

Samples are not \perp because one being in the sample changes prob of the other one.

After picking with replacement — pick and put it back \rightarrow samples are \perp

$$\frac{1}{N} \quad \frac{1}{N-1}$$

$$N=100,000$$

0.2

0.25

0.00000 | \approx next numbers
 are closer.
 something
 like that

large sample even though they are dependent because N is large change is insignificant, so we treat as 1

Conditions

$N \rightarrow$ large

$n \rightarrow$ small relative to N

$$\therefore \hat{F}_x(x) \rightarrow F_x(x)$$

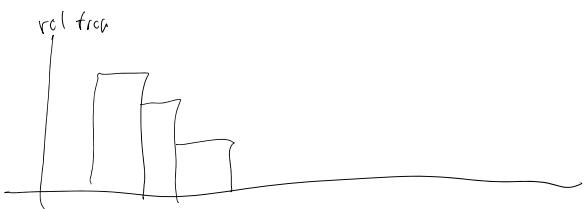
$\hat{F}_x(x)$
 cdf
 calculated
 based on
 sample

Histogram

Ex height

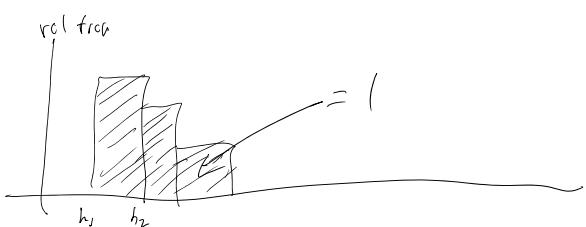
$$(h_1, h_2] \quad (h_2, h_3] \quad (h_3, h_4] \quad \dots \dots$$

5 5:6 5:6 6 6:5



relative frequency = proportion

Density histogram



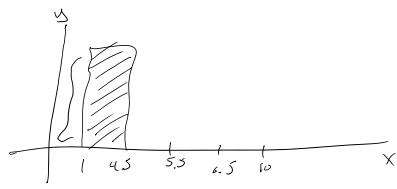
$$h_x(x) = \frac{\text{proportion}}{\text{length}}$$

5.4.5

$$n \geq [1.2 \ 1.8 \ 2.3 \ 2.5 \ 3.1 \ 3.4 \ 3.7 \ 3.2 \ 3.9 \ 4.3 \ 4.4 \ 4.5 \ 4.5]$$

$$[4.8 \ 4.8] \ [5.6 \ 5.8] \ [6.9 \ 7.2 \ 8.5]$$

$$h_x(x) = \frac{\frac{13}{20}}{(4.5 - 1)} \quad (1, 4.5]$$



$$h_x(x) = \frac{\frac{13}{20}}{(4.5 - 1)} \quad , (1, 4.5)$$

$$= \frac{13}{20}$$

Question on
Midterm on this.

$$(4.5, 5.5] \quad , (4.5, 5.5]$$

$$h_x(x) = \frac{\frac{2}{20}}{(5.5 - 4.5)}$$

Ex

$$f_x(x) = \begin{cases} .4 & / x = 1 \\ .3 & / x = 2 \\ .2 & / x = 3 \\ .1 & / x = 4 \\ 0 & / \text{o/w} \end{cases}$$

from this list make the table

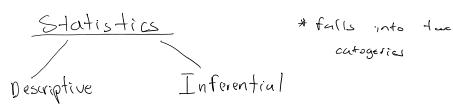
look for unique numbers

① 2, 3, 4

count pop by total + get proportion

loop

Prof uploaded w code. look at the table.



Descriptive: describes ^{any summary}
 e.g. mean of sample
 standard deviation
 median

Inferential: moment you use these #'s to make a inference about the data.

Recall: $f(x)$ is the proportion of the population members whose X measurements equal x .

$$\Rightarrow f_x(x) = P[X=x]$$

$F_x(x)$ is the proportion of population members whose X measurements is less than or equal to x

$$Ex \quad \{1.2, -2.1, 0.4, 3.3, -2.1, 4.0, -0.3, 2.2, 1.5, 5.0\}$$

put in ascending order

$$\{-2.1, -0.3, 0.4, 1.2, 1.5, 2.1, 2.2, 3.3, 4.0, 5.0\}$$

$$\begin{aligned} &\# flag in each data point \quad | \quad x = -3 \\ &\text{to get pdf} \\ f_x(x) &= P[x = -3] = 0 \\ f_x(-2.1) &= P[x = -2.1] = \frac{1}{10} \\ f_x(-0.3) &= P[x = -0.3] = \frac{1}{10} \\ &\vdots \\ f_x(5) &= P[x = 5] = \frac{1}{10} \end{aligned}$$

CDF same thing but everything up to the point

$$F_x(-2.1) = P[X \leq -2.1] = \frac{1}{10}$$

$$F_x(-0.3) = P[X \leq 0.3] = \frac{2}{10} = \frac{1}{5}$$

A natural estimate of $F_x(x)$ is given by $\hat{F}_x(x)$

$$\hat{F}_x(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(x_i) \quad \text{that's the formal definition for what we did above}$$

This is also called empirical distribution function of x
 ↳ means sample.

Calculating Population Quantiles

Given value calculate percentile.

$x(60)$

p-quantile

Note .75 quantile = 75th percentile

↳ $F(x) = \Pr[X \leq x]$

p-quantile

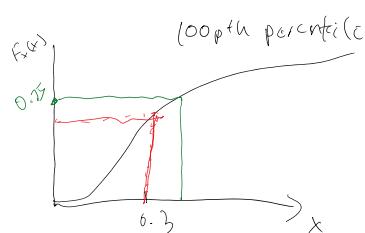
$N_{\text{th}} \cdot 75 \text{ quantile} = 75^{\text{th}} \text{ percentile}$

100th percentile

↗

$\frac{1}{100}$

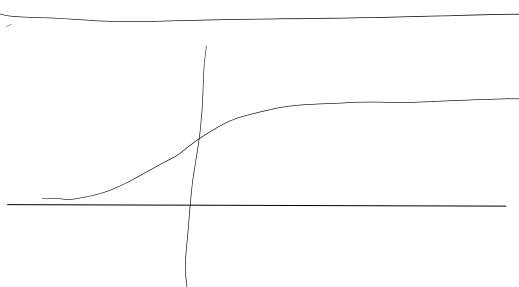
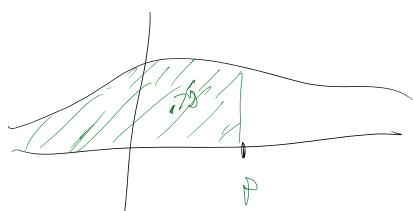
80th percentile = 0.8 quantile



$$f_x(0.3) = \Pr[x \leq 0.3]$$

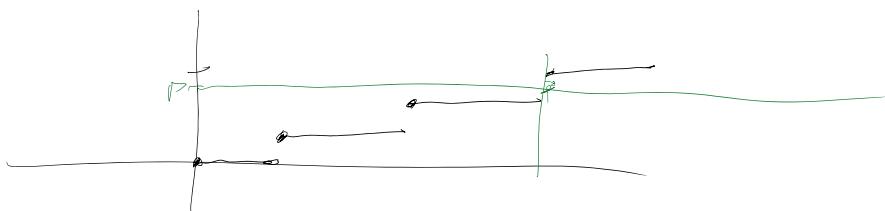
Percentile is opposite

$$f_x(x) = \Pr[x \leq x] = 0.75$$



$$F(x_p) = p$$

$x_p = ?$ if continuous
cdf



$x_1, x_2, x_3, x_4, \dots, x_n$

$\{1.2, 2.1, 0.4, 3.3, -2.1, 4.0, -0.3, 2.2, 1.5, 5.0\}$

put in ascending order

$x(1), x(2), \dots, x(n)$ ↴ (x)^{bucket} means ordered

$\{-2.1, -0.3, 0.4, 1.2, 1.5, 2.1, 2.2, 3.3, 4.0, 5.0\}$

$\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}$

70th percentile? 2.2

, $P = 0.75 ?$

* if you don't have equal then go to the next one

its a # between 2.2 - 3.3

2.2 (70% of data) 3.3

of total $\frac{1}{2}$

← interpolation

$$2.2 (5\% \text{ of data}) = 2.2 + \frac{(3.3 - 2.2)}{2}$$

$$\therefore \frac{i-1}{n} \leq P \leq \frac{i}{n}$$
$$\therefore 1 < i < 8$$

$$\bar{x} = \frac{r}{10}$$

Continuous one always a solution, if discrete no solution sometimes!

$$x = x_{(i-1)} + (x_i - x_{i-1}) n \left(p - \frac{i-1}{n} \right) \stackrel{\text{in this case}}{\Rightarrow} 2.2 + (3.3 - 2.2) 10 (0.75 - 0.7) \\ = 2.2 + (1.1)(0.5) \\ = 2.75$$

25th percentile ← midterm / final question

$$\frac{i-1}{n} < p \leq \frac{i}{n} \\ 0.2 < p \leq 0.3, i=2$$

$$x = x_{(i-1)} + (x_i - x_{i-1}) n \left(p - \frac{i-1}{n} \right) \\ = -0.3 + (0.4 - 0.3) 10 (0.25 - 0.2) \\ = 0.05$$

* uncorrected

if calculate percentile use this formula on midterm and final

25th percentile = $Q_1 = 1^{\text{st}}$ quartile

75th percentile = $Q_3 = 3^{\text{rd}}$ quartile

$$P[x \leq 1.5] = 0.5$$

$$P[x \geq 1.5] = 0.6$$

$$\{1.2, 2.1, 0.4, 3.3, -2.1, 4.0, -0.3, 2.2, 1.5, 5.0\}$$

put in ascending order

$$\{-2.1, -0.3, 0.4, 1.2, 1.5, 2.1, 2.2, 3.3, 4.0, 5.0\}$$

\uparrow
median

In one definition

$P[x \leq 1.5] = 0.5$ is enough to call it a median but in another definition

$P[x \geq 1.5] = 0.6$ will not suffice if $i = 0.5$.

Result if n is odd $\Rightarrow \frac{n+1}{2}$ th term

even $\Rightarrow \frac{n}{2}$ th + $\frac{n+1}{2}$ th term

$$P(x \geq 1.8) = 0.5 \quad \checkmark$$

$$1.5 \quad | \quad 2.1 \\ \hline 1.8$$

Two definitions of median, also easiest one.

1, 3, 5 median is $\frac{n+1}{2} = \frac{3+1}{2} = 2^{\text{nd}}$ term

7

un

Interquartile range: (width of data)

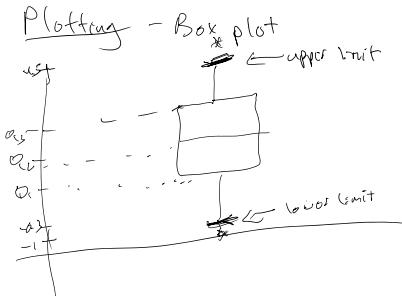
$$IQR = Q_3 - Q_1$$

replacement of SD.



extreme outliers median doesn't suffer

Skew \rightarrow median
symmetric \rightarrow mean.



$$\text{lower limit} = Q_1 - 1.5 \times IQR$$

$$\text{upper limit} = Q_3 + 1.5 \times IQR$$

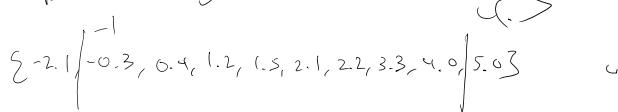
* Box plot
also for continuous
cases.

$$\text{lower limit} = Q_1 - 1.5 \times (Q_3 - Q_1) = 1 \leftarrow \text{means line goes down to } -1$$

stop at -0.3 * — outliers

$$\text{upper limit} = Q_3 + 1.5 \times (Q_3 - Q_1) \\ = 4.5$$

stop at 4.5.



4.5 is also an outlier.

whiskers = *

adjacent values: 4.5, -1

Ex/

- (1) car 0.42
 - (2) Van 0.28
 - (3) BS 0.22
 - (4) St 0.08
- } categorical variable has no order.

if categorical stuff above does not apply

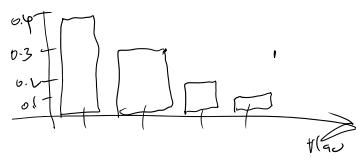
Do this instead



freq

put shift in order

determine region



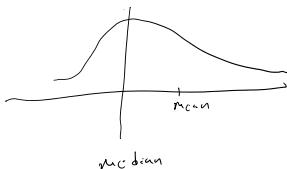
cover

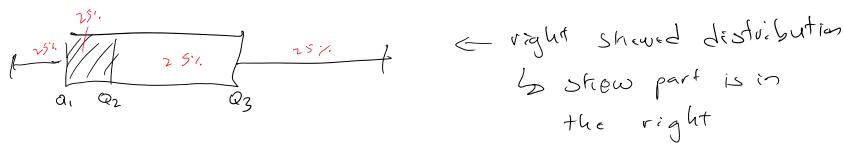
determine region

calculate percentile

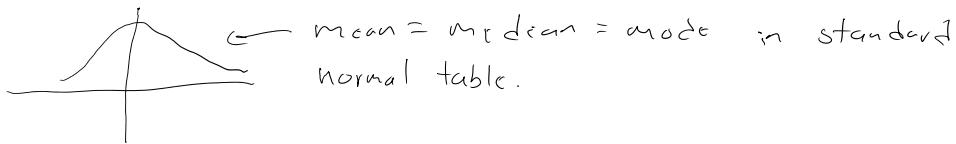
Recap

$$\tilde{x}_p = x_{(i-1)} + (x_{(i)} - x_{(i-1)}) n \left(p - \frac{i-1}{n}\right)$$

<u>skewness:</u> 	odd $\left(\frac{n+1}{2}\right)$ th term even $\frac{\frac{n}{2} + \left(\frac{n}{2} + 1\right)}{2}$ th
--	--



if skewed use median and not the mean



5.2 $X \sim \text{Exp}(1)$

- ① Predict future value
 - ② $P[X > 5]$
 - ③ $P[X \leq a] = 0.95$
- These are inferences about the future X . (observation)

5.3 inference is about distribution

In this chapter $\text{Exp}(\theta) \rightarrow$ estimate θ

Chi
 $\psi(\theta) = \frac{1}{\theta}$ mean = $\frac{1}{\theta}$ var(θ) = $\frac{1}{\theta^2}$

$$\psi(\theta) = \frac{1}{\theta^2} \quad \text{median} = \int_0^M f_\theta(x) dx = 0.5$$

$$\Rightarrow F(m) = 0.5 \\ \Rightarrow m = F^{-1}(0.5)$$

$\boxed{\psi(\theta) = F_\theta^{-1}(0.5)}$

median

25th percentile: $\psi(\theta) = F_\theta^{-1}(0.25)$

Inverse or Exp(1)

$$f_\theta(x) = e^{-x}$$

$$F_\theta(x) = \int_0^x e^{-r} dr = 1 - e^{-x}$$

$$\begin{aligned} F_\theta(m) &= 1 - e^{-m} = 0.5 \\ e^{-m} &= 0.5 \\ -m &= \ln(0.5) \\ \Rightarrow m &= -\ln(0.5) \\ \Rightarrow F_\theta^{-1}(0.5) & \end{aligned}$$

$$f_\theta(x) \rightarrow \text{Normal}(\mu, \sigma^2)$$

$\psi(\theta) \leftarrow$ Parameter of interest

$$\psi(\theta) = \mu$$

$x_1, x_2, \dots, x_n \leftarrow$ can you use this to get mean?

yes $\rightarrow \bar{x}$ use sample mean

$$T(s) = \frac{1}{n} \sum x_i$$

$T(s)$ is an estimate of $\psi(\theta)$

Variance: $\psi(\theta) = \sigma^2$

$$T(s) = \frac{\sum (x_i - \bar{x})^2}{n-1} \rightarrow \text{an estimate of } \frac{\sigma^2}{\psi(\theta)}$$

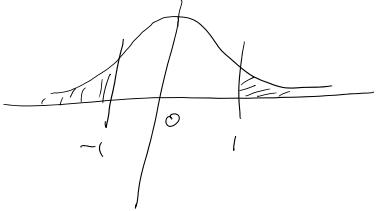
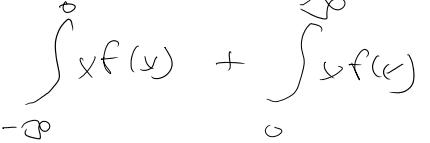
$$\bar{x} \rightarrow E(x)$$
$$n \rightarrow$$

5.5.8

3rd moment $N(\mu_0, \sigma_0^2)$

$$\begin{aligned}
 \text{Var}(x) &= E(x^2) - \mu^2 = E[(x-\mu)^2] \\
 &= E[(x-\mu)^3 + 3(x-\mu)^2\mu + 3(x-\mu)\mu^2 + \mu^3] \\
 &= E[(x-\mu)^3] + E[3(x-\mu)^2\mu] + E[3(x-\mu)\mu^2] + E[\mu^3] \\
 &= E[(x-\mu)^3] + 3\mu E[(x-\mu)^2] + 3\mu^2 E[(x-\mu)] + \mu^3
 \end{aligned}$$

↓ ↓ ↓
 3rd central var(x) ||
 moment σ²

$x \leq 1 = x \geq 1$

if the distribution is symmetric around (μ)

$$E[(x-\mu)^3] = 0$$

$$E[(x-\mu)^3] = 0$$

$$E[(x-\mu)^3] = 0$$

$$\begin{aligned}
 E[(x-\mu)^3] + 3\mu E[(x-\mu)^2] + 3\mu^2 E[(x-\mu)] + \mu^3
 \end{aligned}$$

↓ ↓ ↓
 3rd central var(x) ||
 moment σ² 0

$$= 3\mu\sigma^2 + \mu^3$$

if $N(\mu, \sigma^2)$, what's Θ ? $\Theta = \{\mu, \sigma^2\}$
 ↑
 means all first of parameter

$$G(a, b) \Rightarrow \Theta = \{a, b\}$$