Measure? ← why?

Eg: $P[2^{nd}\ year] = 10\% = \frac{10}{|S|}$ ← # of second year student; |S| ← all people sitting in the room

you are measuring something.

Axioms

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

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

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

Eg A_1 = first year, A_2 = second year } disjoint because nothing in common

3 = HHH
2 = HHT
2 = HTH
1 = HTT
2 = THH
1 = THT
1 = TTH
0 = TTT

} sample space

$P[H] = \frac{1}{2}$

$P[T] = \frac{1}{2}$

$P[3H] = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$

Random variable

Maps sample space

$S \longrightarrow \mathbb{R}$, X = # of heads

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

or add probabilities

table called pmf or probability mass function.

To find $E(X) = \sum x\, p[x]$ → discrete

$\int x\, p[x]\, dx$ → continuous

To find $E[x^2] = \sum x^2 p[x]$ → discrete

$E(x^2] = \int_{-\infty}^{\infty} x^2 p[x]$ → continuous.

Bernoulli:

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

Ex / x = # of H $\begin{matrix} 1 \\ 0 \end{matrix}$

$P[H] = \theta$

$P[T] = 1 - \Theta$

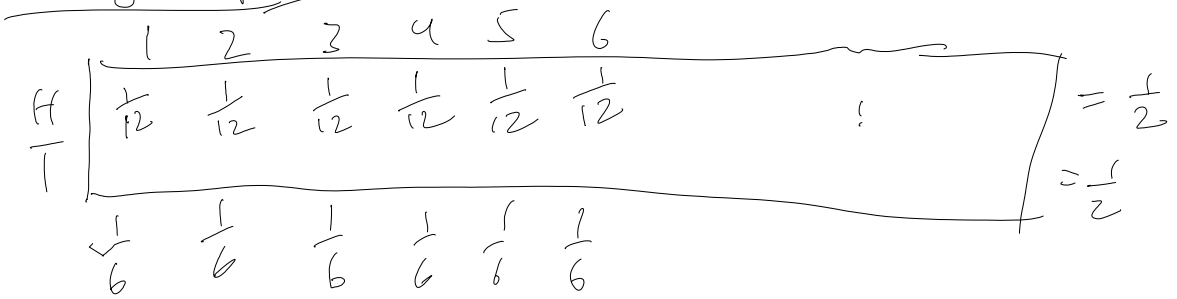
poisson mean and variance is exactly the same $\Rightarrow \lambda$

Continuous &

Uniform, Exp, gamma, normal, Beta.

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

Marginal prob example



$A = H$
$B = 1$; is $A \perp B$? Yes

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

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

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

$P(A \cap B) = P(A)\,P(B) \quad \therefore \quad \frac{1}{12}$

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

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

Treatment Heart Transplant

Control group

P1 - survived 49 days after died
P28 - survived 118 days - but he is alive

P - indicator

P1 - wait time for trasplant is 0 days → Survived 15 days - dead

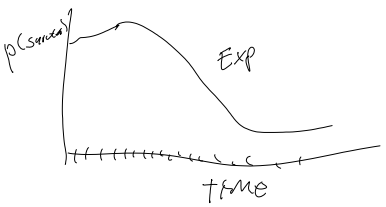
P52 - wait time 5 days → Survived 43 days - alive

30 ppl → mean(t)
Treatment

Control
52 ppl → mean(c)

compare the two values.

Assume 30 is coming from a distribution
eg.



if mean of treatment group is bigger > mean of control group

Interpretation about the population:

=> if you go through the treatment you are expected to live longer.

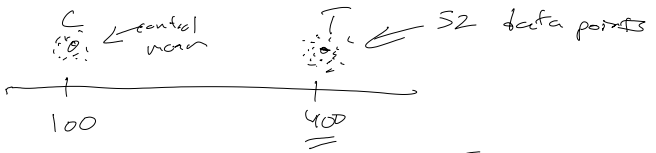
mean(T) > mean(c) are samples, interpretation we want to make is about the entire population.

assume

① $x \sim Exp(\lambda) \rightarrow$
② estimate of λ from these #s (parameter)

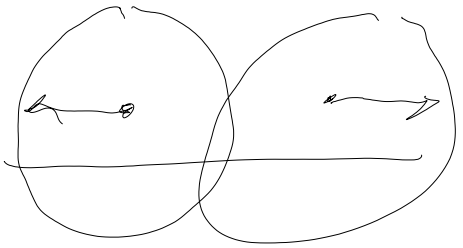
In real life you never know the distribution

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



mean is the center of the distribution

Treatments better



radius also matters.

Data – Circle

mean – center of data

Standard deviation – radius

look if they are overlapping or far apart.

5.1.1 $mean(c) = 93.2$

$= \frac{sum(30\ numbers)}{30}$

$mean(t) = 356.2$

you can't just compare the mean, look at SD and then mean.

Ex 5.1.8

$x \sim Exp(\lambda)$

$x_1, x_2, \ldots x_n$

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

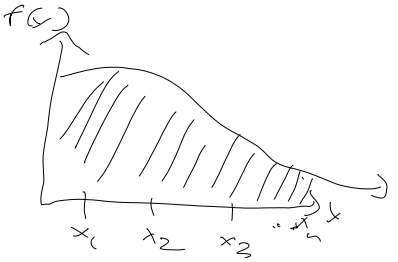
$= \int_0^\infty x \lambda e^{-\lambda x}\, dx$

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

difference between $\bar{x}$ and x_i is

Eg $\lambda = 2$

$f_x(x) = 2e^{-2x}$



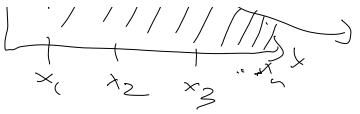
$$= \lambda \int_0^{\infty} x e^{-\lambda x} dx$$

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

$\Gamma(2) \leftarrow$ this is 1

gamma density = 1

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



Can take any value X or Y. is the random variable that can take any value

x_1 is sample value eg x_1: 100 days

$x_2 = 200$ days

weak law of large #'s says

$x_1, x_2, \ldots, x_n$ with $\frac{1}{\lambda}$

$$\Rightarrow \frac{x_1 + x_2 + \ldots + x_n}{n} \xrightarrow{\text{converge}} \frac{1}{\lambda} = E(x)$$

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

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

$\uparrow$ Estimate of lambda

Recall: Gamma

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

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

problems

$\bar{x} = 0.05 \quad \hat{\lambda} = 20$ ← estimate

$\bar{x} = 0.06 \quad \hat{\lambda} = 16.66$

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

$c(1,3,5) + 2 =$

3
5
7

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

printing: $>x$

$[1] \quad 5 \quad 7 \quad 9 \quad 10$

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

$\log(x) \Rightarrow \log\lfloor 5 \rfloor$, $\log\lfloor 7 \rfloor$, $\log\lfloor 9 \rfloor$, $\log\lfloor 10 \rfloor$

1.6094379124341...

$\log_{10}(x)$

$\sin(x)$

$\cos(x)$

x^{n^2}

} does it for all classes

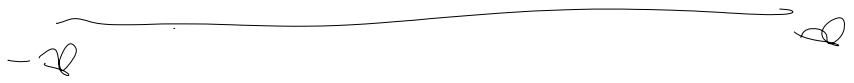
$y =$

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

$y = x^4 + 2y^3 - 3$

$P(y \in (1,2))$ ← prob of between $(1,2)$

$x^2 \Rightarrow$ Chi-sq



Ex 5.2.1

x = life lenght of a machine

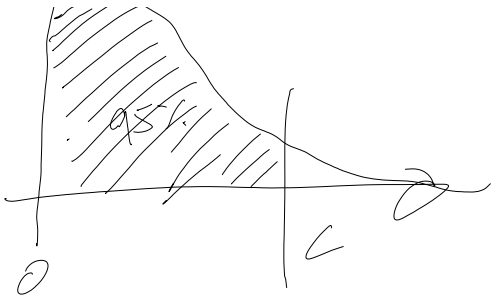
$x \sim Exp(1)$

mean life lenght ? => asking for $E(x)$

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

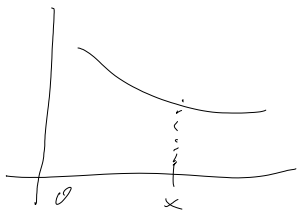
what is the time lenght that 90% of phones will lie

What is the time length that 90% of phones will last



$$0.95 = \int_0^c x f(x)\,dx$$

$$F(x) = \int_0^x f(x)\,dx \quad \text{— cdf}$$



$$\int_0^x e^{-x}\,dx$$

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

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

$$-c$$

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

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

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

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

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

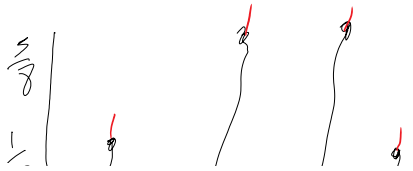
Then 0 isn't an option

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

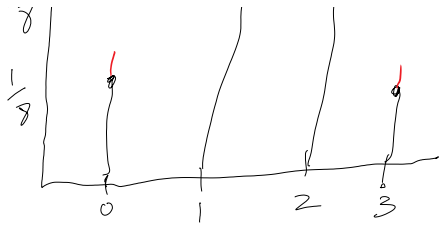
First make it a valid Probability distribution

$P(X=1 \mid X>0)$?

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



make it a bit bigger



make it a bit bigger for PDF = 1

if $x \sim Exp(1)$

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

$E[x \mid x>1] = \int x \cdot$ conditional function

$f_{x|x>1}(x) = \frac{f_x(x)}{P[x>1]}$ ← condition needed for $\sum$ to = 1

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

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

Review 5.1

Example → Samples
- → Distribution
- → Inference → commenting on the calculation

* If x = life length
what is mean/expected life length?
- → asking for expected value $E(x)$
- → by what time 95% of the machines

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

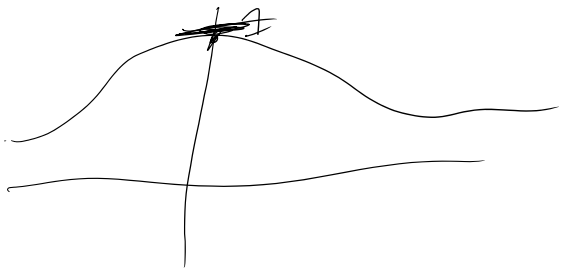
calculate cdf, equal it to 0.95* or value given and solve it.

2nd question: conditional, given $x > 1$

↳ mean / expected life?

$E(x \mid x > 1)$

predict life time distribution; mean or mode



mode = highest point

To find # take derivative set it = 0

max

min

5.2.7 example where you have to calculate the mode.

5.2.8 $x \sim P_0(x)$

* Predict the future value? mean / mode

mode

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

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

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

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

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

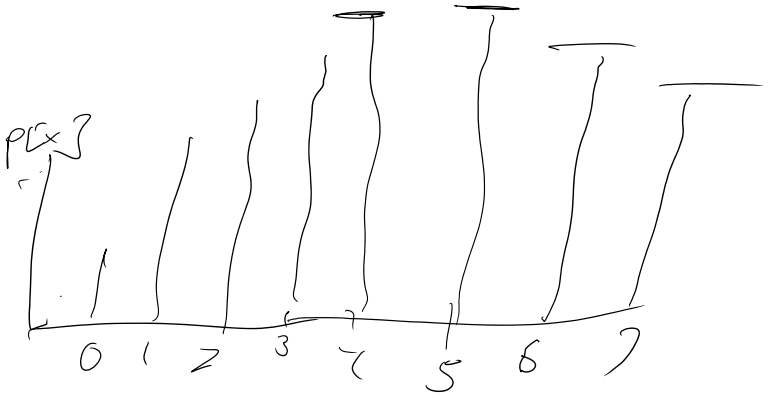
$x=0 \Rightarrow \frac{P[X=1]}{P[X=0]} = \lambda$ → if $\lambda = 5$ then its 5x ratio

$x=1 \Rightarrow \frac{P[X=2]}{P[X=1]} = \frac{\lambda}{2}$ → if $\lambda = 5$ then its 2x 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
Sample ⊆ Population

Parameter: Any characteristics of a population is a parameter

Statistic: summary of sample

populations are too big, samples are easier to work with.

Study sample => make inference about population. That is statistics

Eg $\tilde{E}xp(\frac{1}{\lambda})$, $P_\theta(x)$

(λ: parameter)

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

Statistical model

$\{P_\theta : \theta \in \Omega\}$ (Ω: parameter space)

(P_θ: function, e.g. poisson, normal, etc.)

θ can take any value in the sample space

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

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

Eg $x \sim P_o(\lambda = ?)$

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

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

Interval estimation – review

Example: Pg 264

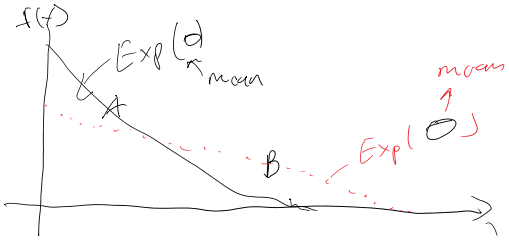
Ex 5.3.2

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

discrete — continuous

Sample $(X_1, X_2, \ldots, X_5)$

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



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

① $(X_1, \ldots, X_5) = (5.0, 3.5, 3.3, 4.1, 2.8)$

② $(X_1, \ldots, X_5) = (2.0, 2.5, 3.0, 3.1, 3.1)$

} which one belongs with which sample?

① ⇒ Exp(1.5)

② ⇒ Exp(1)

Parameter space in this case is $\Omega = \{1, 1.5\}$

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

Ex 5.3.1

① Exp ← know function

② $\Omega = \{1, 1.5\}$ ← know parameter space

③ sample ← see sample and pick

Real life

① observe sample

② find the distribution – known or assumed

② sample ~~=~~ set's sample and pick — known or assumed

③ Parameter - use sample to make inference about parameter

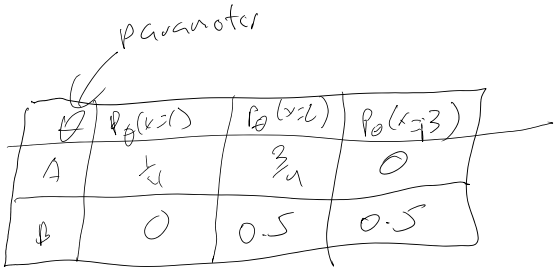
5.3.7

a) Ω 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.



what if $x=2$?

probably A

Notation

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

for one sample value what is the statistic model

Ex $\hat{P}_{Pois}(x)$

$x=1$

$\left\{ \frac{e^{-\lambda} \cdot \lambda^2}{2!} \right\}$

Ex/ $x \sim Exp(\theta)$

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

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

Joint density of sample

Ex/ $p[x_1, x_2, x_3 \ldots x_n]$ if 1 break it down

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

1 is an assumption

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

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

$\theta e^{-2\theta} \times \theta e^{-4\theta} \times \theta e^{-9\theta}$

Ex 5.3.3

$x \sim Bin(\theta)$

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

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

$x_1, x_2, \ldots, x_n$

$\theta^{x_1}(1-\theta)^{1-x_1} * \theta^{x_2}(1-\theta)^{1-x_2} * \cdots * \theta^{x_n}(1-\theta)^{1-x_n}$

$= \theta^{x_1+x_2+x_3\ldots x_n}(1-\theta)^{n-(x_1+x_2\ldots x_n)}$

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

Ex 5.3.4

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

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

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

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

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

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

$$\sum_{i=1}^{n}(x_i-\mu)^2 \qquad a-b \quad a-c+c-b$$

$$= \sum(x_i-\bar{x}+\bar{x}-\mu)^2$$

$$= \sum\left[(x_i-\bar{x})^2+2(x_i-\bar{x})(\bar{x}-\mu)+(\bar{x}-\mu)^2\right]$$

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

$\sum(x_i-\bar{x})$ $\quad n(\bar{x}-\mu)^2$

$= \sum x_i - \sum\bar{x}$ $\qquad$ sum of deviation of mean $=0$

$= n\bar{x} - n\bar{x} = 0$

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

$s^2 = \dfrac{\sum(x_i-\bar{x})^2}{n-1}$ ~ sample variance

sample mean

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

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

$\dfrac{1}{\text{var}}$ = precision

$\lambda = \dfrac{1}{\sigma^2}$ precision

$$(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

$$= (2\pi)^{-\frac{1}{2}}(\lambda)^{\frac{1}{2}}\exp\left[-\frac{1}{2}*\lambda*(x-\mu)^2\right]$$ ← still normal but in precision instead of variance

Reparameterization: only do it if it's a one to one function. precision is $\frac{1}{\text{variance}}$) then change old parameter to new parameter yes to no one to one

old param — func — new param.

Try ex 3.3.5.

Population

Population CDF

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

Ex 5.4.1

$N = 20$

min = 3

$P(X < 3] = 0$

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

$P(X < 4) - P(X < 3) = P(X = 3)$

5.4 = Population = finite

= Sample — distribution known

Unknown distribution

Parameter

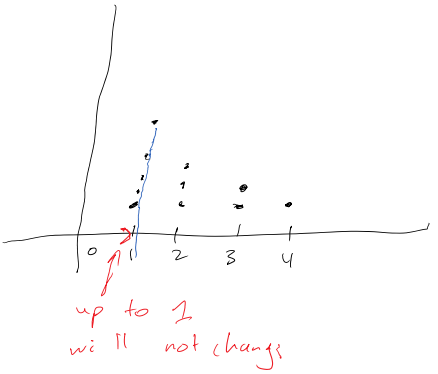
If a population is finite, do you need money / time to study them?

Ex

i	1	2	3	4	5	6	7	8	9	10
$x(\pi_i)$	1	1	2	1	2	3	3	1	2	4

Sort

i	1	2	3	4	5	6	7	8	9	10
$x(\pi_i)$	1	1	1	1	2	2	2	3	3	4



$P[X \le 0] = 0$

$P[X \le 0.999] = 0$

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

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

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

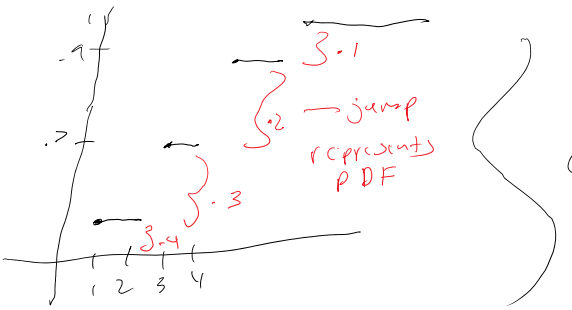
$P[X \le 2.99] = \frac{7}{10} = 0.7$

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

$P[X \le 3.99] = 0.9$

$P[X \le 4] = 1$

CDF = $F_X(x) = \begin{cases} 0 & x < 1 \\ .4 & 1 \le x < 2 \\ .7 & 2 \le x < 3 \\ .9 & 3 \le x < 4 \\ 1 & x \ge 4 \end{cases}$



CDF on a graph

PMF:

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

X	1	2	3	4
$P[X]$	.4	.3	.2	.1

To calculate small f, pdf just calculate the proportion.

CDF = $F_X(x)$ = population

$$\begin{cases} 0 & x < 1 \\ .4 & 1 \le x \\ .7 & 2 \le x < 3 \\ .9 & 3 \le x \le 4 \\ 1 & x \le 4 \end{cases}$$

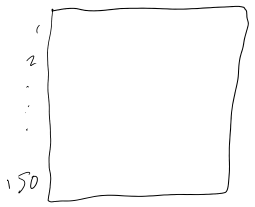
Same calculation but if its sampled its called Empirical distribution function

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

9.4.1 ex/ ← do it on ya own chung.

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}

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

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

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

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

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

After picking < with replacement — pick and put it back → samples are not ⊥
without replacement. — pick # or name → samples are ⊥ don't replace

$\frac{1}{N}$ $\frac{1}{N-1}$ $N = 100,000$

0.2 0.25 0.00060 $\approx$ next numbers are closer.

something like that

large samples even though they are dependent because N is large change is insignificant, so we treat as $\perp$

Conditions

$N \rightarrow$ large

$n \rightarrow$ small relative to N

$\therefore \hat{F}_X(x) \rightarrow F_X(x)$

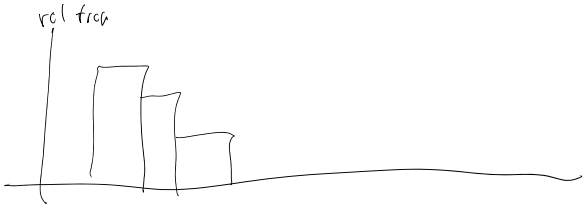
$\hat{F}_X(x)$: cdf calculated based on sample

Histogram

Ex height

$(h_1, h_2]$ $(h_2, h_3]$ $(h_3, h_4]$ . . . .

5 5:6 5:6 6 6 6:5



relative frequency = proportion

Density histogram

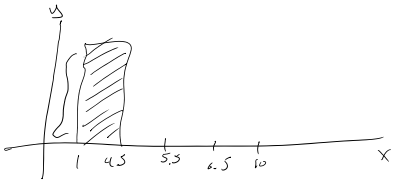
$h_X(x) = \frac{\text{proportion}}{\text{length}}$

5.4.5

a) [1.2 1.8 2.3 2.5 3.1 3.4 3.7 3.7 3.9 4.3 4.4 4.5 4.5]

(4.8 4.8] (5.6 5.8] (6.9 7.2 8.5]

$h_x(x) = \dfrac{\frac{13}{20}}{(4.5-0)}$ $(1, 4.5]$



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

$= \dfrac{13}{20}$

Question on Midterm on this.

$(4.5, 5.5]$, $(4.5, 5.5]$

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

Ex

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

from this list make the table

look for unique numbers

(1) 2, 3, 4

count pop by total to get proportion

loop

Prof uploaded ✓ code. look at the table.

Statistics * falls into two categories

Descriptive Inferential

Descriptive: describes any summary eg mean of sample, standard deviation, median

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

Recall: $f_x(x)$ 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 $\{1.2, 2.1, 0.4, 3.3, -2.1, 4.0, -0.3, 2.2, 1.5, 5.0\}$

Put in ascending order

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

Plug in each data point to get pdf

$x = -3$

$f_x(x)$

$f_x(-3) = P[x = -3] = 0$

$f_x(-2.1) = P[x = -2.1] = \frac{1}{10}$

$f_x(-0.3) = P[x = -0.3] = \frac{1}{10}$

$\vdots$

$f_x(5) = P[x = 5] = \frac{1}{10}$

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 $\hat{F}_x(x)$

$\hat{F}_x(x) = n^{-1} \sum_{i=1}^{n} I_{(-\infty, x]}(x_i)$ thats 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.

p-quantile Note .75 quantile $\xrightarrow{\times 100}$ = 75th percentile

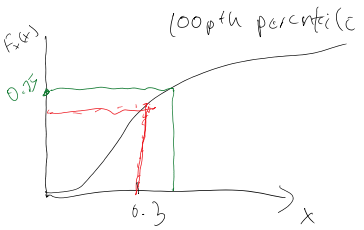
p-quantile

100pth percentile

Note .75 quantile $\hat{=}$ 75th percentile

$\frac{1}{100}$

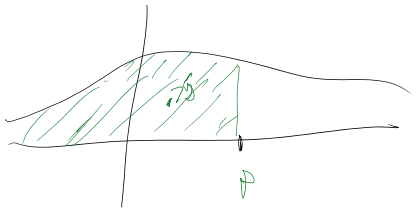
80th percentile = 0.8 quantile

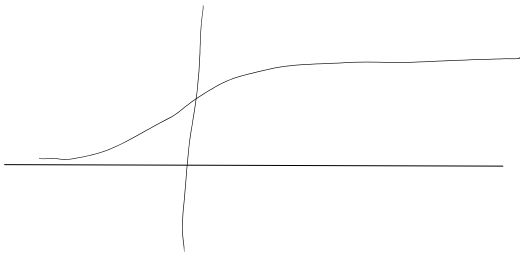


$f_x(0.3) = P[x \leq 0.3]$

Percentile is opposite

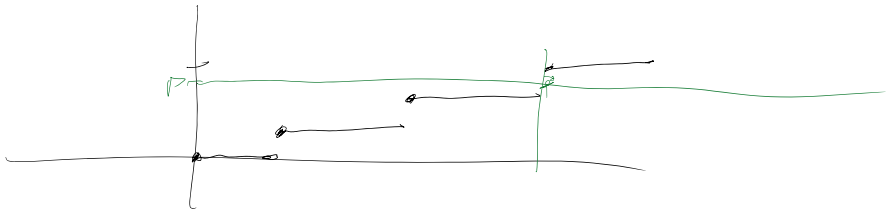
$f_x(x) = P[x \leq x] = 0.75$





$F(x_p) = p$

$x_p = ?$ if continuous cdf



x_1 x_2 x_3 x_{u_1} , , , , x_n

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

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

Put in ascending order

$x_{(1)}$ $x_{(2)}$ ... $x_{(n)}$ ← (x) bracket means ordered

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

$\frac{1}{10}$ $\frac{1}{10}$ $\frac{1}{10}$ $\frac{1}{10}$ $\frac{1}{10}$ $\frac{1}{10}$ $\frac{1}{10}$ $= \frac{7}{10}$

70th percentile? 2.2

$P = 0.75$?

its a # between 2.2 ~ 3.3

2.2 (per. $\frac{1}{2.5\%}$) 80% 3.3

2.2 (10% of data) 3.3

& take $\frac{1}{2}$

← interpolation

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

$P[x \leq 2.2] = 0.7$

$\therefore \quad \frac{i-.5}{n} \leq p \leq \frac{i}{n}$

$2 < n < 8$

$\overline{10} = 1 - \overline{10}$

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

$$\hat{x} = x_{(i-1)} + (x_i - x_{i-1})\, m \left(p - \frac{i-1}{n}\right) \xrightarrow{\text{in this case}} 2.2 + (3.3 - 2.2)\, 10\,(0.25 - 0.2)$$

$$= 2.2 + (1.1)(0.5)$$
$$= 2.75$$

25th percentile ← midterm / final question

$\frac{i-1}{n} < p \leq \frac{i}{n}$

$0.2 < p \leq .3$, $i = 3$

$$\hat{x} = x_{(i-1)} + (x_i - x_{i-1})\, m \left(p - \frac{i-1}{n}\right)$$

$$= -0.3 + (0.4 + 0.3)\,10\,(.25 - .2)$$

$$= 0.05$$

~~*~~ unconfirmed

if calculate percentile use this formula.

on midterm and final

25 th percentile = Q_1 = 1st quartile

75 th percentile = Q_3 = 3rd quartile

$P[X \leq 1.5] = 0.5$ $\quad P[X \geq 1.5] = 0.6$

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

Put in ascending order

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

↑ median

In one definition $P[X \leq 1.5] = 0.5$ is enough to call a median but in another definiti $P[X \geq 1.5) = 0.6$ will not since it $!= 0.5$.

Recall if n is odd → $\frac{n+1}{2}$ th term

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

$P(x \geq 1.8) = 0.5$ ✓

$P(x \leq 1.8) = 0.5$ ✓

1.5 | 2.1

1.8

Two definition of median, also easiest one.

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

Interquartile range: (width of data)

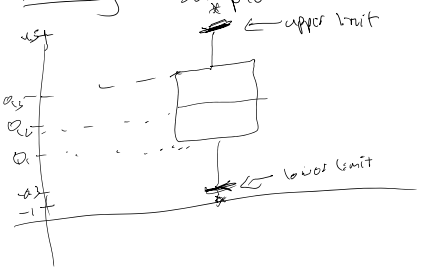
$IQR = Q_3 - Q_1$

replacement of SD.

extreme outliers median doesn't suffer

Skew → median
symmetric → mean.

Plotting - Box plot



lower limit $= Q_1 - 1.5 * IQR$

upper limit $= Q_3 + 1.5 * IQR$

* Box plot also for continuous cases.

lower-limit $= Q_1 - 1.5 * (Q_3 - Q_1) = (-$ ← means line goes down to -(

stop at -0.3 * — outliers

upper limit $= Q_{33} + 1.5 \times (Q_3 - Q_1)$

$= 4.5$

stops at 4.5.

$\{-2.1, -1 | -0.3, 0.4, 1.2, 1.5, 2.1, 2.2, 3.3, 4.0, 5.0\}$ 4.5

4.5 is also an outlier.

whiskers = *

adjacent value: 4.5, -(

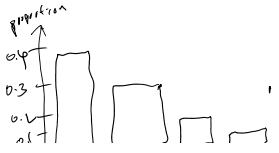
Ex/

(1) Chac 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.

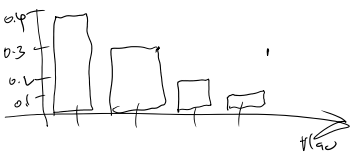
Do this instead



>for>

put shit in order

determine region



determine region

calculate percentile