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 ⟨ 1, 0

$P[H] = \theta$

$P[T] = 1 - \theta$

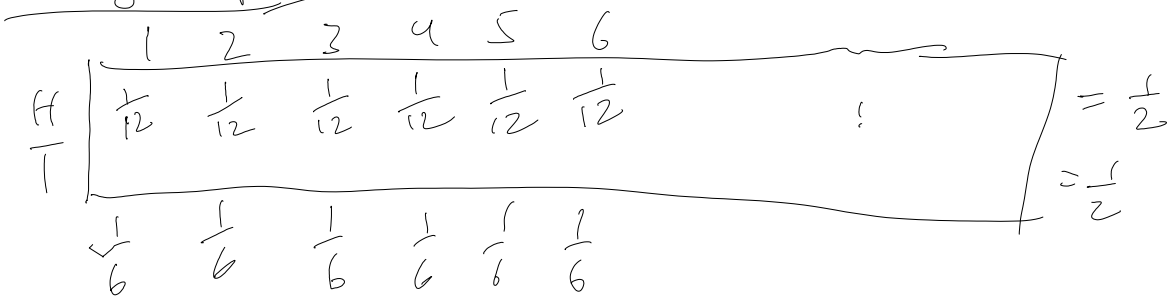
poisson mean and variance is exactly the same $\Rightarrow \lambda$

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

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

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

$P(A \cap B) = P(A)P(B) \therefore \frac{1}{4}$

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

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

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

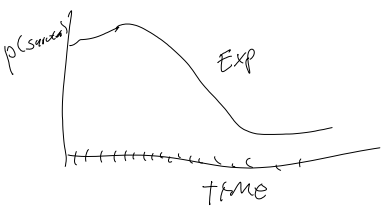
P52 - wait time 5 days → Survived 43 days - alive

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.

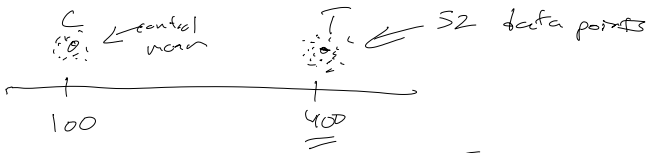
mean(T) > mean(c) are samples, interpretation we want to make is about the entire population.

assume

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

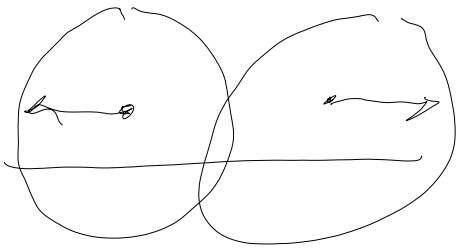
In real life you never know the distribution

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

$x \sim Exp(\lambda)$

$x_1, x_2, \ldots x_n$

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

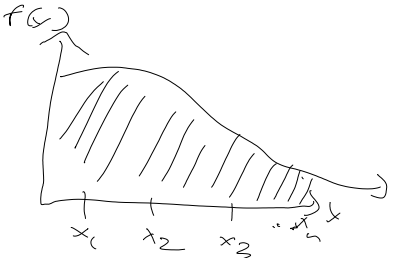
$= \int_0^\infty x \lambda e^{-\lambda x}\, dx$

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

difference between x and x_i is

Eg $\lambda = 2$

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



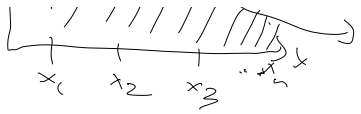
$$= \lambda \int_0^{\infty} x e^{-\lambda x} dx$$

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

$\Gamma(2) \leftarrow$ 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

$\frac{1}{\lambda}$

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

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

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

↑ 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] 5 7 9 10

log(x) => log

5
7

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

1.6698 1.74510...

$\log_{10}(x)$

$\sin(x)$

$\cos(x)$

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 $E(x)$

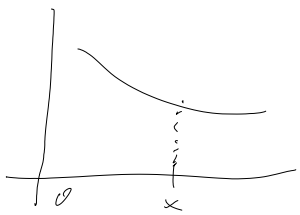
$Exp(\lambda) = \frac{1}{\lambda} \Rightarrow Exp(1) = \frac{1}{1} = 1$

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

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

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

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



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

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

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

$$-1$$

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

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

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) $0 \le P(X=x) \le 1$
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]	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

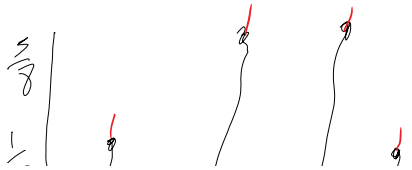
Then 0 isn't an option

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

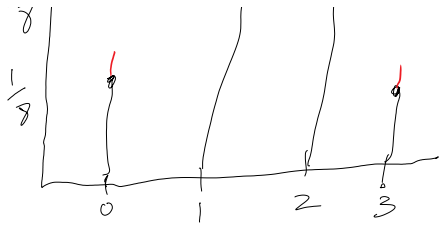
First make it a valid Probability distribution

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

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



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.

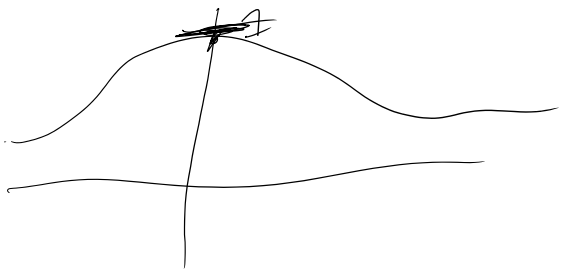
calculate cdf, equal it to 0.95* or value given and solve it.

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 $x \sim Pa(\alpha)$

* Predict the future value? < mean / mode

mode

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

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

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

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

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

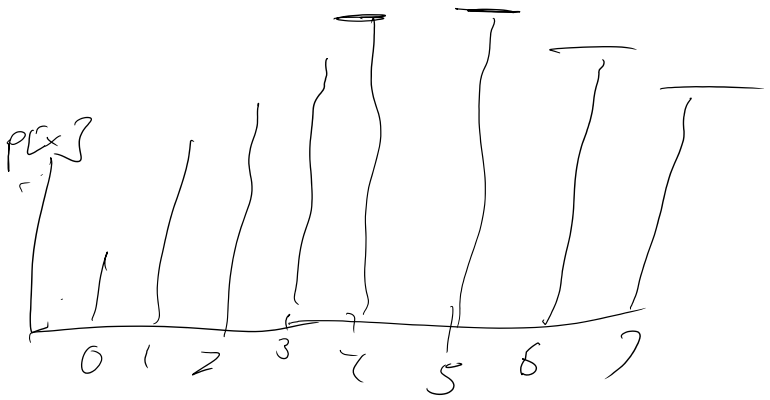
$x=0 \Rightarrow \frac{P[X=1]}{P[X=0]} = \lambda$ → if $\lambda = 5$ then its 5x ratio

$x=1 \Rightarrow \frac{P[X=2]}{P[X=1]} = \frac{\lambda}{2}$ → 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 $\tilde{E}xp(\underbrace{1}_{\text{parameter}})$, $P_\theta(x)$

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

Statistical model

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

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

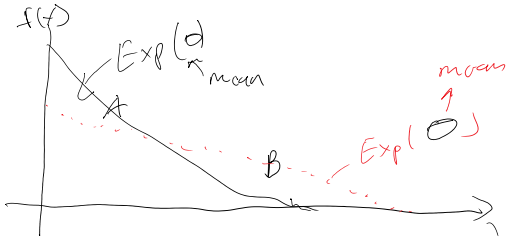
Ex 5.3.2

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

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(.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

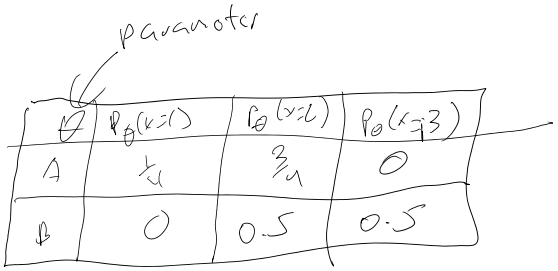
5.3.7

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

$x=2$

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

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 I break it down

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

I 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

$x \sim Bin(\theta)$

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

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

$x_1, x_2, \ldots x_n$

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

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

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

Ex 5.3.4

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

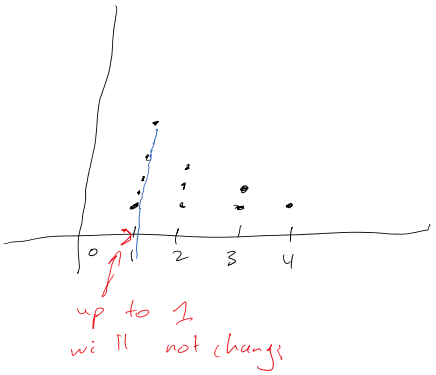
If a population is finite, do you need money / time to study them?

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	2	2	2	3	3	4	



$P[X \le 0] = 0$

$P[X \le 0.999] = 0$

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

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

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

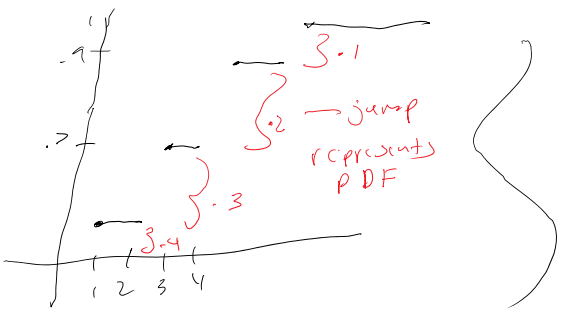
$P[X \le 2.99] = \frac{7}{10} = 0.7$

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

$P[X \le 3.99] = 0.9$

$P[X \le 4] = 1$

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

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

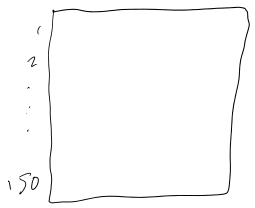
Same calculation but if its sampled its called Emperical distribution function

Emperical 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 0.00065 $\approx$ 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

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

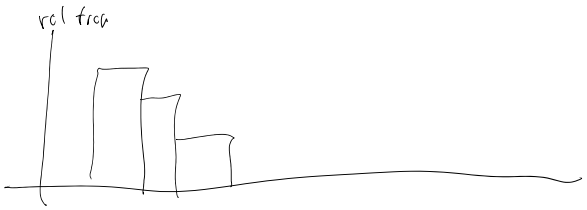
$\hat{F}_X(x)$: cdf calculated based on sample

Histogram

Ex height

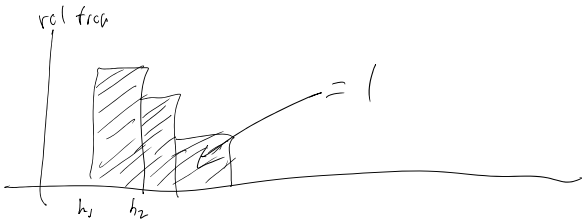
$(h_1, h_2]$ $(h_2, h_3]$ $(h_3, h_4]$ ....

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



relative frequency = proportion

Density histogram



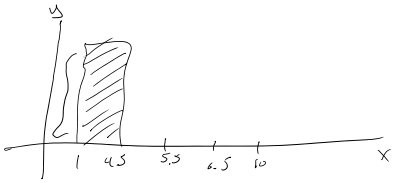
$h_X(x) = \frac{\text{proportion}}{\text{length}}$

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 catogories

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

$f_x(x)$

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

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

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

$\vdots$

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

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 emperical distribution function of x

↳ means sample.

Calculating Population Quantiles

Given value calculate percentile. $\xrightarrow{\times 100}$

p-quantile Note .75 quantile = = 75th percentile

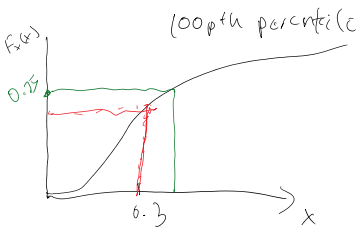
p-quantile

100pth percentile

Note .75 quantile = 75th percentile

1/100

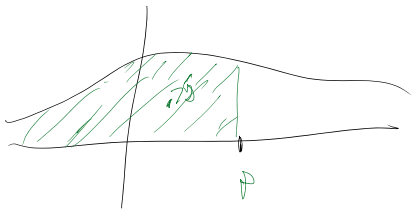
80th percentile = 0.8 quantile

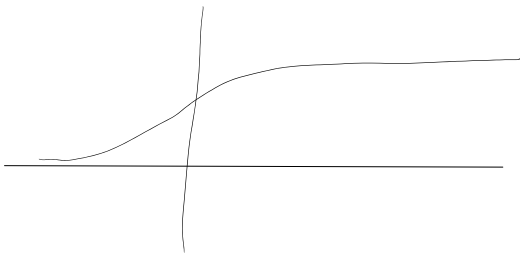


$f_x(0.3) = P[x \le 0.3]$

Percentile is opposite

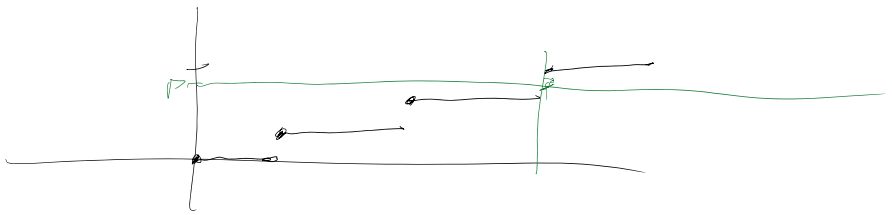
$f_x(x) = P[x \le x] = 0.75$





$F(x_p) = p$

$x_p = ?$ 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}

* if you don't have equal 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. $\frac{1}{2}$ 75%) 3.3

2.2 (10% of data) 3.3

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

← interpolation

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

$P[x \le 2.2] = 0.7$

$\frac{i-1}{n} \le p \le \frac{i}{n}$

$\frac{7}{10} \le p \le \frac{8}{10}$

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

$$\hat{x} = x_{(i-1)} + (x_i - x_{i-1})\, n \left(p - \frac{i-1}{n}\right) \xrightarrow{\text{in this case}} 2.2 + (3.3-2.2)\,10\,(0.25-0.1)$$

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

$$= 2.75$$

25th percentile ⟸ midterm / final question

$$\frac{i-1}{n} < p \leq \frac{i}{n}, \quad i = 3$$

$$0.2 < p \leq 0.3$$

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

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

$$= 0.05$$

~~*~~ unconfirmed

if calculate percentile use this formula on midterm and final

25th percentile = Q_1 = 1st quartile

75th 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)

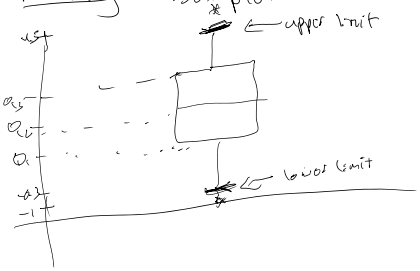
$IQR = Q_3 - Q_1$

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 $= Q_3 + 1.5 \times (Q_3 - Q_1)$

$= 4.5$

stops at 4.5

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

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.

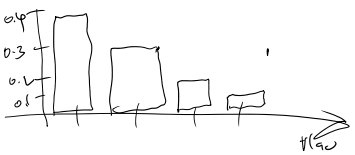
if categorical stuff above does not apply.

Do this instead

>-fcp>

put shit in order

determine region



determine regim

calculate pressure

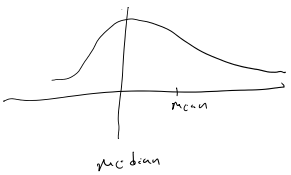
Recap

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

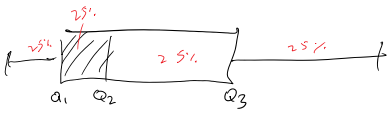
odd $\left(\frac{n+1}{2}\right)$th term

even $\frac{\frac{n}{2}\text{th} + \left(\frac{n}{2}+1\right)\text{th}}{2}$

Skewness:



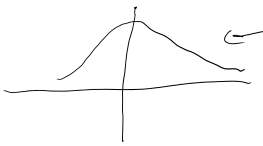
median



← right skewed distribution
↳ skew part is in the right

← left skewed distribution

if skewed use median and <u>not</u> the mean



← mean = median = mode in standard normal table.

5.2 $X \sim Exp(1)$

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 $Exp(\theta) \rightarrow$ estimate θ

chi

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

$\psi(\theta) = \frac{1}{\theta^2}$ 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 ex/ Exp(1)

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

$F_\theta(x) = \int_0^x e^{-x}\,dx = 1 - e^{-x}$

$F_\theta(m) = 1 - e^{-m} = 0.5$

$e^{-m} = 0.5$

$-m = \ln(0.5)$

$\Rightarrow m = -\ln(0.5)$

$\Rightarrow F_\theta^{-1}(0.5)$

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

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

$\psi(\theta) = \mu$

$x_1, x_2 \ldots x_n \leftarrow$ 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: $\psi(\theta) = \sigma^2$

$T(S) = \frac{\sum (x_i - \bar{x})^2}{n-1}$ is an estimate of σ^2 $\psi(\theta)$

$T(S) = S^2$

$\bar{x} \longrightarrow E(x)$

$n \longrightarrow$

5.5.8

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

μ_0 mean not; σ_0 sigma not squared

$$\psi(\theta) = E(x^3) = E[x - \mu + \mu]^3$$

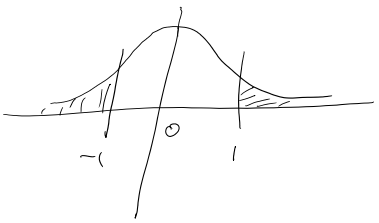
$$= E[(x-\mu)^3 + 3(x-\mu)^2\mu + 3(x-\mu)\mu^2 + \mu^3]$$

$$= E[(x-\mu)^3] + E[3(x-\mu)^2\mu] + E[3(x-\mu)\mu^2] + E[\mu^3]$$

$$= E[(x-\mu)^3] + 3\mu E[(x-\mu)^2] + 3\mu^2 E[(x-\mu)] + \mu^3$$

$E[(x-\mu)^3]$: 3rd central moment

$E[(x-\mu)^2]$: $\text{var}(y) = \sigma^2$



$$\int_{-\infty}^{0} xf(x) + \int_{0}^{\infty} xf(x)$$

$x \leq 1 = x \geq 1$

if the distribution is symmetric around (μ)

$$E[x-\mu]^1 = 0$$

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

$$E[x-\mu]^5 = 0$$

$$E[(x-\mu)^3] + 3\mu E[(x-\mu)^2] + 3\mu^2 E[(x-\mu)] + \mu^3$$

$E[(x-\mu)^3]$: 3rd central moment $= 0$

$E[(x-\mu)^2]$: $\text{var}(y) = \sigma^2$

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

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

if $N(\mu, \sigma^2)$, what's θ? $\theta = \{\mu, \sigma^2\}$

↑ many call list of parameter

$G(\alpha, \beta) \Longrightarrow \theta = \{\alpha, \beta\}$

Section: 5.2

- Predicting future value
- $X \sim (\lambda)$
  - mean
  - mode: predicting x
  - $P[Y>5]$, $P[x>5 | x>1]$
  - $E[X | Y>1]$

Section 5.3

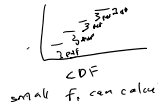
only if $\vec{S} = x_1, x_2, x_3, \ldots x_n$ (Lets say you have a sample, what is $f_\theta(s)$ ← probability function of the sample)

$f_\theta(s) = f_\theta(x_1) f_\theta(x_2) \ldots f_\theta(x_n)$

- reparameterization: $\theta \to \theta$, one to one then we can reparameterize

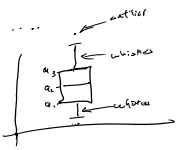
Section: 5.4

$F_x(x)$ (population CDF) → $f_x(x)$



CDF

- Once you know the small f_x can calculate everything.
- $\hat{F}_x(x)$ ← empirical distribution function
- histograms
- $\hat{f}_x(x)$
  - mean
  - median
  - $p = \ldots$
- Q_1, Q_3
- IQR
- Q-1.5(IQR)
- box plot



- $\theta = \mu, \sigma^2$

$E(\bar{x}) = \ldots$ (refer to last class notes)

6.1 Likelihood Function

$X \sim Bern(\theta)$ ⊥ $\theta \in [0,1]$

$S = \{x_1, x_2, \ldots x_n\}$

$f_\theta(S) = f_\theta(x_1) * f_\theta(x_2) \ldots f_\theta(x_n)$

Toss coins 5 times:

$[x_1, x_2, x_3, x_4, x_5] = S$

$[1, 0, 1, 1, 0]$

$f_\theta(S) = \theta(1-\theta)\theta\theta(1-\theta)$

$L(\theta | s) = \theta^3(1-\theta)^2$ } likelihood of θ

likelihood function of theta given the sample.

* Not the probability of observing θ. It is the probability of observing the sample for a given true θ.

θ → fixed true value, $\theta \in \Omega$

θ_2 any other member from Ω

believe $f_{\theta_1}(s) > f_{\theta_2}(s)$

The more data the smaller the likelihood.

Ex/ 6.1.1

Suppose $S = \{1, 2, \ldots\}$ and that the statistical model is $[P_\theta : \theta \in \{1,2\}]$, where P_1 is the uniform distribution on the integers $\{1, \ldots, 10^3\}$ and P_2 is the uniform distribution on $\{1, \ldots, 10^6\}$

Dist 2 → f_{θ_1} → Unif $\{1, 2, 3, \ldots 1000\}$

→ Unif $\{1, 2, 3, \ldots 100000\}$

Dist 1 $\rightarrow f_{\theta_1}$

Dist 2 $\rightarrow f_{\theta_2} \rightarrow$ Unif $\{1, 2, 3, \ldots 1000000\}$

$S = \{1001\}$ → comes from distribution 2. Since distribution 1 stops at 1000.

$S = \{10\}$ → # calculate likelihood for both and compare

$L(\theta_1|S) = \frac{1}{1000}$

$L(\theta_2|S) = \frac{1}{1000000}$ ⇒ sample 10 is at 1000 x more likely to come from θ_1

Example 6.1.4 $X \sim N(\theta, \sigma_0^2)$ / σ_0^2 known?

1 unknown parameter θ, mean

$S = x_1, x_2, \ldots x_n$

$$f_\theta(s) = \prod_{i=1}^{n} (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_0^2}(x_i-\theta)^2\right]$$

(constant)

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\theta)^2\right]$$

recognize this & expand it, Do it yourself.

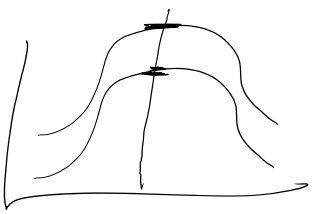
$$= (2\pi\sigma_0^2)^{-\frac{n}{2}} \exp\left[-\frac{n}{2\sigma_0^2}(\bar{x}-\theta)^2\right] \exp\left[-\frac{1}{2\sigma_0^2}(n-1)s^2\right]$$

constant — constant — sample variance

Since constant we write

$$L(\theta|s) \propto \exp\left[\frac{-n}{2\sigma_0^2}(\bar{x}-\theta)^2\right] \Rightarrow$$

proportional

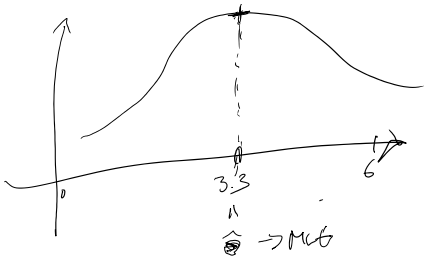


same graph but just stretched as a result you can make inference about mean

$n = 25$, $\sigma_0^2 = 1$, $\bar{x} = 3.3$

$$L(\theta|s) \propto \exp\left[\frac{-25}{2}(3.3-\theta)^2\right]$$

recall
$y = 5x$
$y \propto x$ } proportional to x



$\hat{\theta} \rightarrow$ MLE

① Point of estimate of θ

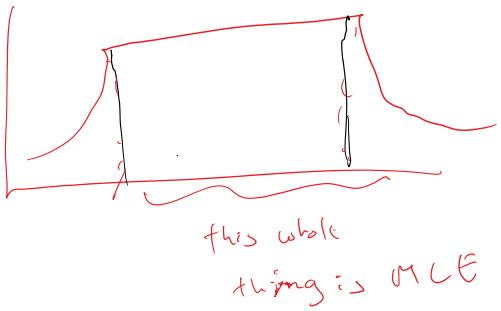
$\hat{\theta} \rightarrow \theta$

estimate

MLE → maximum likelihood estimator of θ

$L(\hat{\theta}|s) \geq L(\theta|s)$

Note you can have multiple MLE for θ



Ex 6.2.1

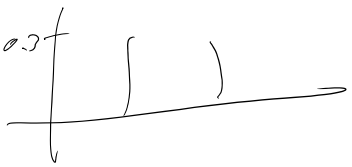
	s=1	s=2	s=3
f_1	0.3	0.4	0.3
f_2	0.1	0.7	0.2

$\{s=1\}$ $\Rightarrow$ probably came from 2st one

$\{s=2\}$ $\Rightarrow$ probably came from 2nd one

$\{s=3\}$ $\Rightarrow$ probably came from 2st one

	s=1	s=2	s=3
f_1	0.3	0.4	0.3
f_2	0	0.7	0.3



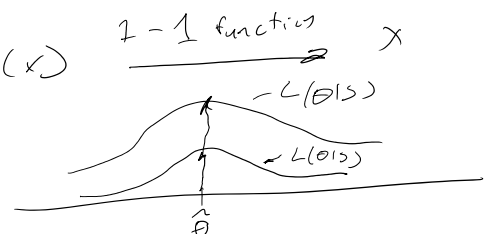
$\{s=3\}$ $\Rightarrow$ MLE is probably Θ_1 or Θ_2

$$L(\theta|s) = \prod_{i=1}^{n} f_\theta(x_i)$$

$\Downarrow$ taking the ln allows us to change to summation

$$\ln(L(\theta|s))$$

$\ln(x)$ 1-1 function



$\ln(L(\theta|s))$ summation

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

Now maximize

$L(\theta|s) \propto \exp\left[\frac{-n}{2\sigma_0^2}(\bar{x}-\theta)^2\right]$

proportional

(1) $\ell(\theta|s) = \frac{-n}{2\sigma_0^2}(\bar{x}-\theta)^2$ ← ln likelihood of θ

Differentiate

(2) Score Function $= \frac{\partial \ell(\theta|s)}{\partial \theta}$

(3) score equation $= \frac{\partial \ell(\theta|s)}{\partial \theta} = 0$

$\hat{\theta} =$ something

To check if max or min, second derivative.

$\frac{\partial^2 \ell(\theta|s)}{\partial^2 \theta}\Big|_{\theta=\hat{\theta}} < 0$

$\ell(\theta|s) = \frac{-n}{2\sigma_0^2}(\bar{x}-\theta)^2$

$\frac{\partial \ell(\theta|s)}{\partial \theta} = -\frac{2n}{2\sigma_0^2}(\bar{x}-\theta)(-1)$

$= \frac{n}{\sigma_0^2}(\bar{x}-\theta)$ ← score

$\Rightarrow \frac{n}{\sigma_0^2}(\bar{x}-\theta) = 0$

$\theta = \bar{x}$

check MLE

$\text{score} = \frac{n}{\sigma_0^2}(\bar{x}-\theta)$

$\frac{\partial^2 \ell(\theta|s)}{\partial^2(\theta)} = -\frac{n}{\sigma_0^2}$ ← n sample size; σ^2 positive

$n > 0$

$\ell'(\theta)$

$\therefore \frac{-n}{\theta_0^2} \leq 0$

$\therefore \bar{X}$ is the MLE of θ

Do it for exp distribution

Question midterm or final or both, learn to calculate MLE of everything distribution we known.

Invariance

$f_\theta(x)$ $\theta \rightarrow \hat{\theta}$ is MLE

$\psi(\theta)$ function of θ

one to one

how do you find MLE?

$\psi(\theta) \rightarrow \psi(\hat{\theta})$

6.2.2 If $(x_1, \ldots, x_n)$ is a sample from $Ber(\theta)$ distribution where $\theta \in [0,1]$ is unknown, then determine the MLE of θ

$x \sim Bir(\theta)$ ⟵ mle ?

$S = x_1, x_2, \ldots x_n$

write the distribution plug $x_1, x_2 \ldots x_n$

likelihood $L(\theta|s) = \theta^{x_1}(1-\theta)^{1-x_1} \cdot \theta^{x_2}(1-\theta)^{1-x_2} \cdots \theta^{x_n}(1-\theta)^{1-x_n}$

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

get in likelihood

ln likelihood $\ell(\theta|s) = \ln\left(\theta^{\sum_{i=1}^{n} x_i}\right) + \ln(1-\theta)^{\cdots}$

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

get first derivative

$$\frac{\partial \ell(\theta|s)}{\partial \theta} = \sum_{i=1}^{n} x_i \cdot \frac{1}{\theta} - \left(n - \sum_{i=1}^{n} x_i\right) \cdot \frac{1}{1-\theta}$$ } score

set it to zero to get score equation

$$\frac{\partial \ell(\theta|s)}{\partial \theta} = \sum_{i=1}^{n} x_i \cdot \frac{1}{\theta} - \left(n - \sum_{i=1}^{n} x_i\right) \cdot \frac{1}{1-\theta} = 0$$

$$\Rightarrow \frac{\sum x_i}{\theta} = \frac{n - \sum x_i}{1-\theta}$$

$$\Rightarrow \frac{1-\theta}{\theta} = \frac{n - \sum x_i}{\sum x_i}$$

$$\Rightarrow \frac{1}{\theta} - 1 = \frac{n}{\sum x_i} - 1$$

$$\Rightarrow \frac{1}{\theta} = \frac{n}{\sum x_i}$$

$$\sum x_i = n\theta$$

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

check second derivative then you can say it's MLE

Ex/ $\theta \to \bar{x}$ MLE } since one to one or $Bir(\theta), [0,1]$

$\theta^2 \to \bar{x}^2$

$\theta^2 \rightarrow \bar{X}^2$

6.2.5 Suppose $(x_1, \ldots x_n)$

$x \sim \text{Uniform}[0, \theta]$

$f_\theta(x) = \frac{1}{\theta} I[x \in [0, \theta]]$

$x_1, x_2 \ldots x_n$

$L(\theta | S) = \frac{1}{\theta} \cdot \frac{1}{\theta} \cdots \frac{1}{\theta}$

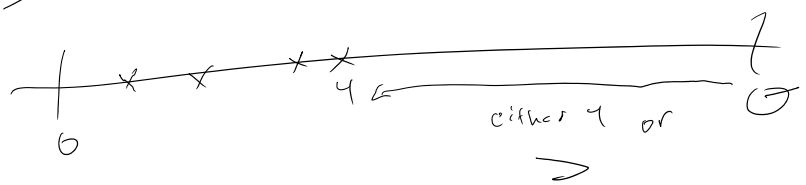
$= \frac{1}{\theta^n}$

Differentiation part doesn't work

$\frac{1}{\theta^n}$ → this is maximized when bottom is minimized.

∴ when θ is minimum $L(\theta)$ is maximum

Ex assume 1, 1.2, 2.3, 4



Minimum θ = max value in sample

$\hat{\theta} = \max(S)$

Midterm: November 3rd
1-4 pm
IC #130

Recap:

Likelihood
↓ take ln of likelihood to get MLE
MLE $\frac{\partial}{\partial\theta} = 0 \Rightarrow \hat{\theta}$ $\frac{\partial^2 \ell}{\partial\theta^2} < 0$

Ex $X \sim Bern(\theta)$

$s_1 = (1,0,1,1,0,0,1)$ $\quad 1 \Rightarrow H$
$s_2 = (0,1,0,1,0,0,1)$

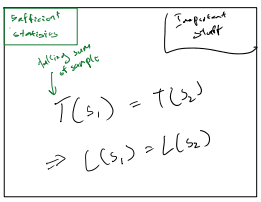
$L(\theta \mid s_1) = \theta(1-\theta)\theta\theta(1-\theta)(1-\theta)\theta$
$= (1-\theta)^2\theta^4$

$L(\theta, s_2) = (1-\theta)\theta(1-\theta)\theta(1-\theta)\theta\theta$
$= (1-\theta)^3\theta^4$

Ex/ $n = 7$
# of heads = 4

sol $\binom{7}{4}\theta^4(1-\theta)^3$
$L(\theta|s)$

All 3 give you the same likelihood functions. because $\sum x_i = 4$, sum of random variables is 4.



may or may not have constant in the front. constant doesn't matter when maximizing it.

T = sum of all your sample.

$f_\theta(s) = h(s) * g_\theta(T(s))$
↓ free of θ only function of sample
$\theta, T(s)$ ↓ sufficient stat

Ex
$\binom{7}{4}\theta^4(1-\theta)^3$
$h(s)$

In general
$\theta^{\sum x}(1-\theta)^{n-\sum x}$
$g_\theta(\ )$

Ex $X \sim Exp(\theta)$

$x_1, x_2, \ldots x_n$

Sol
$f_\theta(s) = \theta e^{-\theta x_1} \cdot \theta e^{-\theta x_2} \cdots \theta e^{-\theta x_n}$
$= \theta^n e^{-\theta \sum x_i}$

$\Rightarrow h(s) = 1$ because both terms are of θ

$= 1 \cdot \theta^n e^{-\theta \sum x_i}$ ← sufficient statistic
$h(s)$ $g_\theta(\ )$

Ex $x \sim Pois(\theta)$

$s = x_1, x_2, \ldots x_n$

likelihood function

$$f_\theta(s) = \frac{e^{-\theta}\theta^{x_1}}{x_1!} * \frac{e^{-\theta}\theta^{x_2}}{x_2!} \times \cdots \frac{e^{-\theta}(\theta)^{x_n}}{x_n!}$$

$$= \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod_{i=1}^{n} x_i!} \qquad h(s) = \frac{1}{\prod_{i=1}^{n} x_i!}$$

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

$$T(s) = \sum x_i$$

if $[2, 3, 4, 2, 1]$ is # of accidents on 401 (Hwy) and follows $f_\theta(\theta)$ then

$$\frac{e^{-n\theta}\theta^{\sum x_i}}{\prod_{i=1}^{n} x_i} = e^{-5\theta}\theta^{12}$$

take ln

$\frac{12}{5}$

, don't need to know 5 different #, you just need the sum

, sum is sufficient stat of data

from likelihood you can get function of x but not sample.

$T(s)$ can be retain back from your likelihood is called a sufficient statistic.

sufficient means data reduction.

T(x)

↑

minimal sufficient

statistic means you can't reduce further.

if $s_1 = [x_1, x_2, \ldots x_n]$

$s_2 = [y_1, y_2, \ldots y_n]$

$$\frac{\theta^n e^{-\theta \Sigma x}}{\theta^n e^{-\theta \Sigma y}} = e^{-\theta(\Sigma x - \Sigma y)}$$

if $\Sigma x = \Sigma y$

$\Rightarrow$ minimal statistic

Likelihood Review

For some data $S = (x_1, \ldots, x_n)$ observed which follows a join distribution $f_\theta(s)$, write likelihood

$$L(\theta|s) = f_\theta(s) = \prod_{i=1}^{n} f_\theta(x) \quad \text{when } x_i \sim iid$$

One of the things we can do with likelihood is sufficient statistic.

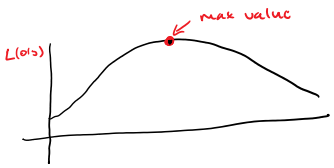
Sufficient statistic of θ ($T(s)$)

↳ To find use factorization theorem.
- This helps identify $T(s)$ ← sufficient statistic of θ

$$f_\theta(s) = h(s)\, g_\theta(T(s))$$

Maximum Likelihood Estimation (MLE)

- MLE ($\hat{\theta}$) value that maximizes $L(\theta|s)$ namely $L(\hat{\theta}|s) \geq L(\theta|s) \quad \forall \theta \in \Omega$



Steps for finding MLE

1. write out $L(\theta|s)$ ← likelihood function
2. take ln of $L(\theta|s)$ ← ln likelihood function
3. take the derivative of $\ln(L(\theta|s))$ with respect to θ
4. set to 0 and solve for θ

Review chapter 4.6 for background

Inference based on MLE

- for $T(s)$ an estimate for $\psi(\theta)$, measure "closeness" using

$$MSE_\theta(T(s)) = E_\theta[T(s) - (\psi(\theta))^2]$$

mean squared error; sufficient statistic of θ; Expectation

= average squared distance between T and unknown $\psi(\theta)$

Recall, if $T(s) = \bar{x}$ then sampling distribution of $T(s)$ when S is $N(\mu, \sigma^2)$ is $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$

- Easier MSE formula

$$MSE_\theta(T(s)) = Var_\theta(T(s)) + (E_\theta(T(s)) - \psi(\theta))^2$$

Proof

$$E_\theta[(T(s) - \psi(\theta))^2]$$

$$= E_\theta\left[[T(s) - E_\theta(T(s))] + [E_\theta(T(s)) - \psi(\theta)]^2\right],$$ cheap trick prof gonna use, add and subtract same

* update prof said look at the

value. Figure out why before the midterm!

ending answer and used only for proofs

$$= E[(T(\underline{x}) - E[T(\underline{x})])^2] + \underbrace{2E[(T(\underline{x}) - E[T(\underline{x})])(E[T(\underline{x})] - \psi(\theta))]}_{0} + (E[T(\underline{x})] - \psi(\theta))^2$$

, expand it $E(k) = 0$ where k is a constant

$$= \mathrm{Var}(T(\underline{x})) + \underbrace{(E[T(\underline{x})] - \psi(\theta))^2}_{\text{bias}} + 0$$

Bias of T when $E_\theta[T]$ exists

$$\mathrm{Bias}(T) = E_\theta[T(\underline{x})] - \psi(\theta) = \begin{cases} 0 & \text{then } T \text{ is unbiased} \\ \text{otherwise} & T \text{ is biased} \end{cases}$$

or equivalently, if $E[T(\underline{x})] = \psi(\theta)$ then T is unbiased.

Ex 6.3.1

$(x_1, \ldots, x_n) \sim N(\mu, \sigma_0^2)$ with unknown μ and known σ_0^2. Want to calculate the MSE

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + (E[\hat{\theta}] - \mu)^2$$

$$= \mathrm{Var}\left(\frac{1}{n}\sum x_i\right) + (E[\bar{x}] - \mu)^2$$

$$= \frac{1}{n^2}\sum \mathrm{Var}(x_i) + \left(\frac{1}{n}\sum E(x_i) - \mu\right)^2$$

$$= \frac{1}{n^2}(n\sigma_0^2) + \left(\frac{1}{n}n\mu - \mu\right)^2$$

$$= \frac{\sigma_0^2}{n} + (0)^2$$

This implies its unbiased

Previous section

MLE of μ (μ = mean): $\hat{\theta} = T(\underline{x}) = \bar{x}$

Recall from 4.6

$\bar{x} \sim N(\mu, \frac{\sigma_0^2}{n})$

Example 6.3.4

Now $(x_1, \ldots x_n) \sim N(\mu, \frac{\sigma^2}{n})$, (μ, σ^2) are unknown

We have MLE for (μ, σ^2) is $(\bar{x}, \frac{n-1}{n}s^2)$

Still $\bar{x} \sim N(\mu, \frac{\hat{\sigma}^2}{n})$ so $E[\bar{x}] = \mu$

$$\mathrm{MSE}(\bar{x}) = \mathrm{Var}(\bar{x}) + \mathrm{Bias}(\bar{x})$$

$$= \mathrm{Var}(\bar{x}) + 0$$

$$= \frac{\hat{\sigma}^2}{n}$$

$$= \frac{1}{n^2}(n-1)s^2 \approx \frac{s^2}{n}$$ ← for large values of n, we can approximate it to $\frac{s^2}{n}$ because as $n \uparrow$ $\frac{n-1}{n^2}$ becomes more insignificant

Bias

Variance of statistic is often called standard error $= \sqrt{\sigma^2}$

Confidence Intervals

* Copy slide 7, 8, 9 → she read off the slide really fast.

- "γ" is referred to as the confidence level of the interval
  ↑ gamma

Example 6.3.6 $(x_1, \ldots, x_n) \sim N(\mu, \sigma_0^2)$, μ unknown (mean) and σ_0^2 known (variance)

want to develop an interval for μ using our data that isn't too wide but has high confidence.

Start from likelihood

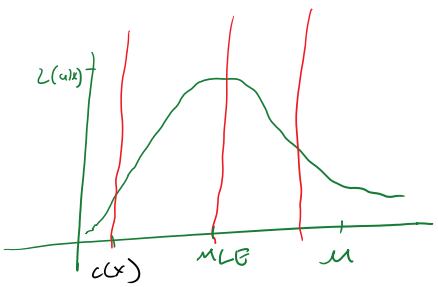
$$L(\mu | \underline{x}) \propto \exp\left\{ \frac{-n}{2\sigma_0^2} (\bar{x} - \mu)^2 \right\}$$

↑ proportional

IR – high confidence
– very wide

Define $C(x_1, \ldots, x_n)$ such that $\mu_1 \in C(x)$ and
↑ This is a "C" ↑ This is a "C"

$L(\mu_2 | \underline{x}) \geq L(\mu_1 | \underline{x})$ then $\mu_2 \in C(x)$



Thus, we can write

$$C(\underline{x}) = \{\mu : L(\mu | x_1, \ldots, x_n) \geq K(x_1, \ldots, x_n)\}$$

$$= \{\mu : \exp\left(\frac{-n}{2\sigma_0^2} (\bar{x} - \mu)^2 \right) \geq K(x_1, \ldots, x_n)\}$$

$$= \{\mu : \frac{-n}{2\sigma_0^2} (\bar{x} - \mu)^2 \geq \ln K(x_1, \ldots x_n)\}$$

$$= \{\mu \; (x - \mu)^2 \leq \frac{-2\sigma_0^2}{n} \ln K(x_1, \ldots, x_n)\}$$

$$= \{\mu \; \bar{x} - K^*(x) \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{x} + K^*(x) \frac{\sigma_0}{\sqrt{n}}\}$$

where $K^*(x) = \sqrt{-2\ln(K(x))}$

Now we need to choose $K(x)$ or equivalently $K^*(x)$

we know

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma_0}{\sqrt{n}}} \sim N(0,1)$$

can rewrite $C(\underline{x})$ as a probabalistic statement by

$$\gamma = P(\mu \in C(x)) = P\left(\bar{x} - K^* \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{x} + K^* \frac{\sigma_0}{\sqrt{n}} \right)$$

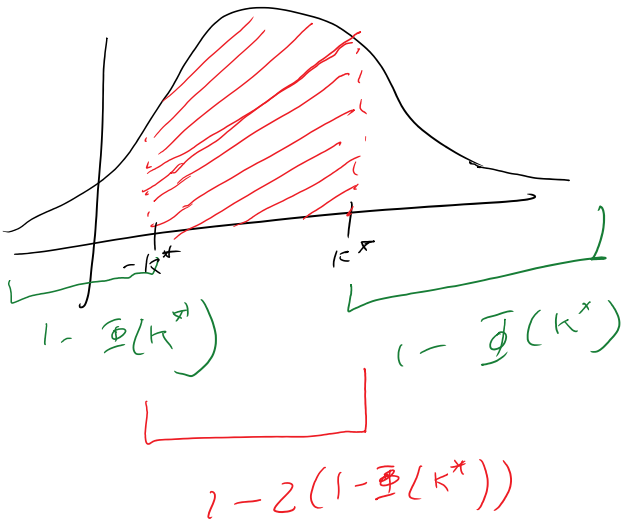
$$= P(-K^* \leq \bar{x} - \mu \ldots$$

- Visually

$\leq \mu \leq \bar{x} + k^* \frac{\sigma_0}{\sqrt{n}}$

$= P\left(-k^* \leq \frac{\bar{x} - \mu}{\frac{\sigma_0}{\sqrt{n}}} \leq k^*\right)$

$= P\left(\left|\frac{\bar{x} - \mu}{\frac{\sigma_0}{\sqrt{n}}}\right| \leq k^*\right)$



$\Rightarrow P(|Z| \leq k^*) \geq \gamma$

$\uparrow N(0,1)$

$\Rightarrow 1 - 2(1 - \Phi(k^*)) \geq \gamma$

$\Rightarrow \Phi(k^*) = \frac{1+\gamma}{2}$

$\Rightarrow \boxed{k^* = Z\left(\frac{1+\gamma}{2}\right)}$ from $N(0,1)$ table

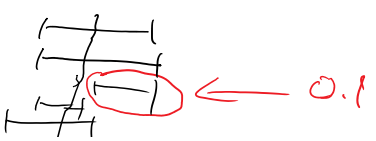
$$C(\gamma) = \left[\bar{x} - Z\left(\frac{1+\gamma}{2}\right)\frac{\sigma_0}{\sqrt{n}},\ \bar{x} + Z\left(\frac{1+\gamma}{2}\right)\frac{\sigma_0}{\sqrt{n}}\right]$$

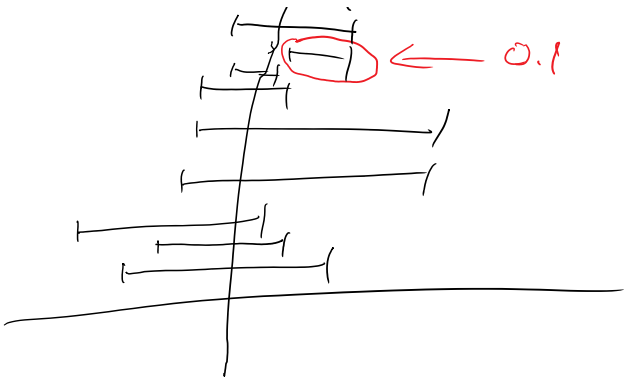
margin of error

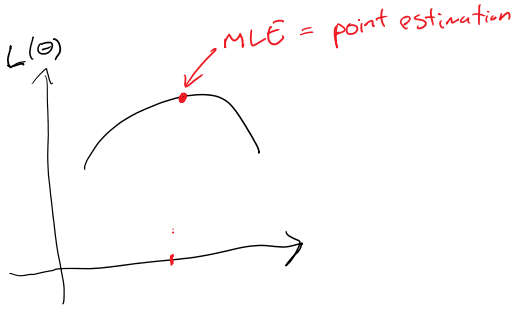
* based on the likelihood construction, proper CI interpretation is that for repeated sampling with CI constructed each time, then $100\gamma\%$ of these CI contain true $\psi(\theta)$

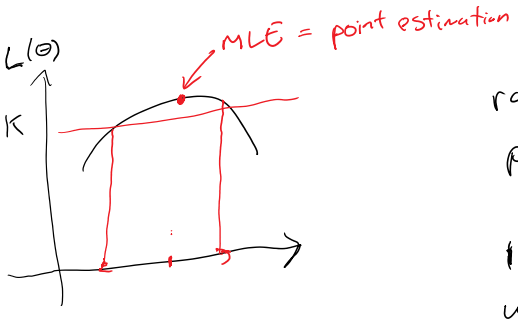
(!) write this down on exam if you want marks

Ex $\gamma = 0.9$









range of θ

$P[\text{likelihood} \geq k] \geq \gamma$

looking for a range of thetas where there is a gamma chance it will be above some threshold.

$P[\text{likelihood} \geq k] \geq \gamma$

$P[\log \text{likelihood} \geq k^*] \geq \gamma$

likelihood interval for any distribution

Ex) $N(\mu, \underbrace{\sigma_0^2}_{\text{known}})$, is $-\left(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$ the log likelihood? yes.

5) $P\left[-\left(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \geq \underbrace{k^*}_{\text{constant}}\right] \geq \gamma$

$P\left[\left(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \leq \underbrace{k^{**}}_{\text{constant with neg}}\right] \geq \gamma$

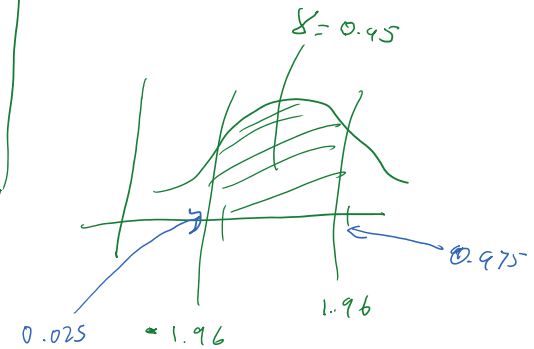
$P\left[-\sqrt{k^{**}} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \sqrt{k^{**}}\right] \geq \gamma$

Only cause of CLT

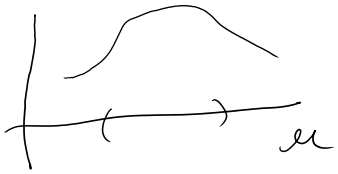
$P\left[-z_{\frac{1+\gamma}{2}} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{1+\gamma}{2}}\right] \geq \gamma$

Central limit thrm

$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$



qnorm(0.025) = -1.96



interval for μ since μ is unknown

qnorm(0.025) = −1.96

qnorm(0.975) = 1.96

if $\gamma = 0.9$.

$\left(\frac{1+\gamma}{2}\right) = \frac{1.9}{2} = 0.95$

Transform the equation to have μ on one side.

$$P\left[\bar{Y} - z_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}\right] \Rightarrow \left(\bar{x} \pm z_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Ex $Y \sim N(\mu, \sigma^2 = 6), n = 9, \bar{x} = 5$, calc 95% confidence interval

σ^2 is known ✓

What's gamma? 0.95

What quantile should I look at? $\frac{1+\gamma}{2} = 0.975 = \Phi(1.96)$

$$\Rightarrow 5 \pm 1.96 * \frac{4}{\sqrt{9}} \Rightarrow \begin{cases} 7.61 \\ 2.38 \end{cases}$$

→ interpolation lower and upper bound contain the the μ

$P[2.38 \le \mu \le 7.61] = 0.95$ ✗

likelihood is not a pdf, just a function.

$\bar{X}$ is a variable, because

lets say

Sample 1

28 − 32

| 30 | 31 | 29 | 36 |

Sample 2

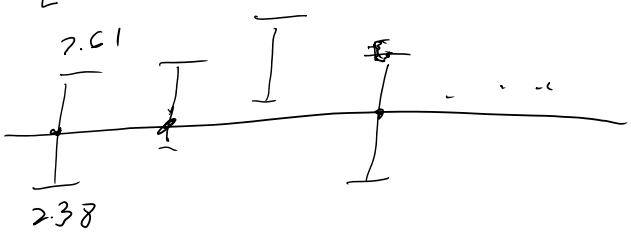
$x_2 - x_3$

| 32 | 28 | 25 | μ_2 |

it changes every sample so its a variable

Confidence interval is a variable.

$P[2.38 < \mu < 7.61] = 0.95$ ✗



* look at the whole thing not just one case

correct

$P[l_b \le \mu \le u_b] = 0.95$ ✓

Ex $N(\mu, \sigma^2 \rightarrow \text{unknown})$

$$P\left[-k^{**} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq k^{**}\right] \geq \gamma$$

, we can't use C/T since σ^2 is unknown

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$\rightarrow \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$ (chi-square distribution) ← don't worry you will learn this later from

parameter

degree of freedom (df)

no parameter in standard normal

$N(0,1) \rightarrow$ parameter Free $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

$$\frac{\text{Std Normal}}{\sqrt{\frac{\chi^2}{df}}} \sim t$$ ← will not be tested

$$\frac{\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{(n-1)}}} = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \sim t\text{-distribution}$$

magic

When you know sigma use standard normal

When you don't know, use t-distribution.

Midterm Questions

Short answers
derive
intuition regarding classifier } 4 questions

no question on midterm
on final there is gaussian covariance diagonalization/on midterm doesn't

Class Conditional model

Where can it go wrong? - If gaussian is a bad parametric model.

- These model can be overfitted easily

In covariance matrix there are 4 #s but 3 unique numbers
100 Dim space => 100 parameters to specify your cov...
# of parameter grows really quickly.

Naive Bayes Avoids the problem.

Naive Bayes

Assume: Diff measurements are $\perp$ random variables the distributions define are $\perp$ | Covariance matrix is diagonal

if $C = \begin{pmatrix} \diagdown & 0 \\ 0 & \diagdown \end{pmatrix}$ then $p(x_1, x_2) = p(x_1)p(x_2)$

Naive Bayes sorts tells suppose fit a gaussian distribution with an diagonal covariance matrix.

if 2-d space => Cov(X) has 2 parameter
D space => Cov(X) has D parameter

Downside: Can rep.  you can't do diagonal ellipses.

$Cov = E((x - \mu_x)(x - \mu_y))$

feature vector of documents are feature vectors of words

$F_{1:D} = (F_1, F_2, \ldots, F_D)$

F denotes occurrence of a word in a hypothetical of document

where $F_1, F_2, \ldots F_D$ would be words that appear in the document

F_1 = "cryptolocker" F_2 = "cryptolocker" ...

$f_1 = 1$: if contains "cryptolocker"
$= 0$, o/w

Model Prob given document class

$P(F_{1:D} | C_1)$ — Class Conditional Random

$P(F_{1:D} | C_1)$
$P(f_1 F_2 F_3 | C_1) = \prod_{i=1}^{D} P(f_i | C_1)$

Think about
$P(F_{1:3} | C_1)$
$= P(F_1 | C_1 F_2 F_3) P(F_2, F_3 | C_1)$, Bayes rule
4 words that have to be consistent and. — $P(F_2 | C_1 F_3) P(F_3 | C_1)$ — parameter 0, or 1 based on the word or not
predict F_2 and F_3 not $P(F_2 | C_1 F_3 F_3)$ $P(F_3 | C_1 F_2 F_3)$
$\therefore 2$

We want to compute posterior prob.

$P(C=j | F_{1:D}) = \frac{P(F_{1:D} | C=j) P(C=j)}{P(F_{1:D})}$

As dimension grows, # of condition grows exponentially. # of #'s needed grows exponentially.

$P(C=j | F_{1:D}) = \frac{\prod_{i=1}^{D} P(F_i | C=j) P(C=j)}{P(F_{1:D})}$

$P(C=j | F_{1:D}) = \frac{\prod_{i=1}^{D} P(F_i | C=j) P(C=j)}{\sum_{k=1}^{K} P(F_{1:D} | C=k) P(C=k)}$

$b_j \equiv P(C=j)$
$a_{ij} \equiv P(F_i = 1 | C=j)$

$P(C=j | F_{1:D}) \propto \underbrace{\prod_{i:F_i=1} a_{ij}}_{P(F_i=1)} \underbrace{\prod_{i:F_i=0} (1-a_{ij})}_{P(F_i=0)} b_j$

$\frac{\prod_{i:F_i=1} a_{ij} \prod_{i:F_i=0} (1-a_{ij}) b_j}{\sum_{k=1}^{K} \prod_{i:F_i=1} a_{ik} \prod_{i:F_i=0} (1-a_{ik}) b_k}$

To estimate to ML estimate or MAP or Bayes

$\prod_{i:F_i=1} a_{ij} \prod_{i:F_i=0} (1-a_{ij}) b_j = \prod_{i=1}^{D} a_{ij}^{F_i} (1-a_{ij})^{(1-F_i)}$ } easier to deal with where doing in treatment

$\log \prod_{i=1}^{D} a_{ij}^{F_i} (1-a_{ij})^{(1-F_i)}$

$= \sum_{i=1}^{D} \log(a_{ij}^{F_i}) + \log(1-a_{ij})^{(1-F_i)}$

$= \sum_{i=1}^{D} F_i \log(a_{ij}) + (1-F_i) \log((1-a_{ij})$ ← once you have log likelihood one you can fully estimate the log posterior

MC estimator is

$b_j = \frac{N_j}{N}$, where N is total # of training
N_j class of the total documents
$\approx \frac{N_j + B}{N + KB}$ ← where that has class j.
of Bayes estimate know assumptions

$a_{ij} = \frac{N_{ij}}{N_j}$, N_j total # of documents from class j
N_{ij} total # of documents from class j where $i = 1$.

$\approx \frac{N_{ij} + \alpha K}{N_j + 2\alpha K}$

Likelihood can ogive you a low estimate if # of sample is low

$$p(c_1|\bar{x}) = \frac{p(x|c_1)p(c_1)}{p(x)}$$

$p(x) = p(x,c_1) + p(x,c_2)$

$p(y) = p(x|c_1)p(c_1) + p(x|c_2)p(c_2)$

$$= \frac{p(x|c_1)p(c_1)}{p(x|c_1)p(c_1) + p(x|c_2)p(c_2)}$$, divide top and bottom by the top

$$= \frac{1}{1 + \frac{p(x|c_2)p(c_2)}{p(x|c_1)p(c_1)}}$$ important form

$= p(c_1|\bar{x})$ posterior prob distribution of c_1 given x

=> This form tells us that the relative prob matters

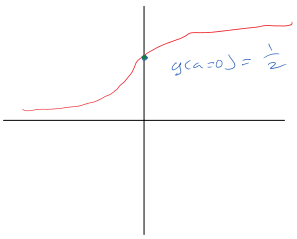
Recall

$$a(x) = \log\left(\frac{p(x|c_1)p(c_1)}{p(x|c_2)p(c_2)}\right)$$ => decision boundary $a(x) = 0$

$$p(c_1|x) = \frac{1}{1 + \frac{p(x|c_2)p(c_2)}{p(x|c_1)p(c_1)}}$$

$$= \left(\frac{1}{1+e^{-a(x)}}\right) = \underbrace{g(a(x))}_{\text{sigmoid function}}$$

$$g(a) = \frac{1}{1+e^{-a}}$$



Logistic Regression

classification technique not a classification.

Suppose $a(x) = \bar{x}^T\bar{w} + b$ ← suppose linear

$= \bar{x}^T\bar{w}$ ← ignore bias by assuming you collapse in your weight vector.

$$p(c_1|x) = \frac{1}{1+e^{-\bar{w}^T x}} = g(\bar{w}^T\bar{x})$$

Learning

Optimization on likelihood

data $\{(x_i, y_i)\}_{i=1}^{N}$ $y \in \{0,1\}$ ($0 \to c_2$, $1 \to c_1$)

$X = \{x_i\}$ $Y = \{y_i\}$

$p(X, y|\bar{w}) = p(y|\bar{w}x)\,p(x|\bar{w})$

$p(x)$ because data doesn't depend on w, y depends on us though.

$\Rightarrow x \perp w$

$\propto p(y|wx)\,p(x)$

$p(y|x,\bar{w})$

$p_i \equiv p(y_i=c_1|x_i,\bar{w})$

$1-p_i \equiv p(y_i=c_2|x_i,\bar{w})$

$$p(y|X,\bar{w}) = \prod_{i:y_i=c_1} p_i \prod_{i:y_i=c_2} (1-p_i)$$

$$= \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{(1-y_i)} \Leftarrow \text{cross entropy loss function}$$

$$= \prod_{i=1}^{N} g(\bar{x}_i^T\bar{w})^{y_i}(1-g(\bar{x}_i^T\bar{w}))^{(1-y_i)}$$

$$E(w) = -\log p(y|X,\bar{w}) = -\sum_{i=1}^{N} y_i \log g(\bar{x}^T\bar{w}) + (1-y_i)\log(1-g(\bar{x}_i^T\bar{w})) \leftarrow \text{energy function}$$

$p_i \equiv p(y_i=c_1|x_i,\bar{w}) = g(\bar{x}_i^T\bar{w})$

$1-p_i \equiv p(y_i=c_2|x_i,\bar{w}) = 1-g(\bar{x}_i^T\bar{w})$

iterative optimization technique, take gradient with respect to, step downhill re-calculate slope of the hill - Gradient decent.

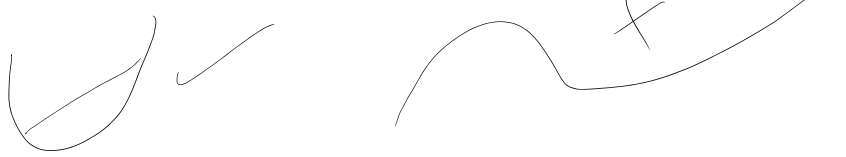
A2

get the gradient

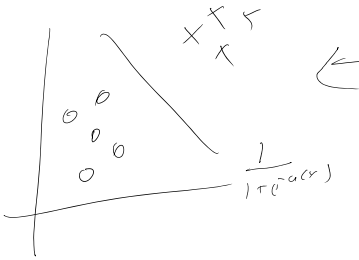
Gradient decent code is provided

find derivative of the energy functions.

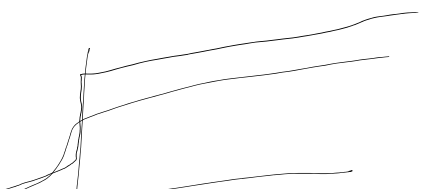
convex function:



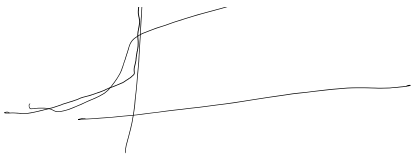
Unique minimum.



← linearly seperable data. Can draw a linear decision boundry that perfectly seperates classes.



← if weights in a become critical if data is ganna be seperable the function becomes discontinous. If won't if data becomes mixed.



function becomes discontinuous
if data becomes mixed.

∴ add a regularizer

$$E(w) = -\log p(xy|\bar{w}) = -\sum_{i=1}^{N} y_i \log g(\bar{x}^T \bar{w}) + (1-y_i)\log(1-g(\bar{x}_i^T \bar{w})) + \underbrace{\lambda \bar{w}^T \bar{w}}_{\text{regularizer}}$$

Review

Bayesian setting: θ = parameter $\in \Theta$ is random ✓

data $\in X_1, X_2, \ldots, X_n \sim \text{Bern}(\theta)$

$\theta \sim \text{Beta}(\alpha, \beta)$

↑ prior

$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

$\pi(\theta|s) \propto$ likelihood * prior

↓ sample data

$\propto \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}$

$\propto \underbrace{\theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}}$

Beta distribution, missing the constant.

$\theta|s \sim \text{Beta}(\sum x_i+\alpha, n-\sum x_i+\beta)$

Ex if $n=3$, $s=\{1,0,1\}$ and $\theta \sim \text{Beta}(\alpha=2, \beta=2)$

then what is $\theta|s \sim \text{Beta}(\sum x_i+\alpha, n-\sum x_i+\beta)$?

$\text{Beta}(\sum x_i+\alpha, n-\sum x_i+\beta)$

$= \text{Beta}(2+2, 3-2+2)$

$= \text{Beta}(4, 3)$ ← posterior, you know the parameters

mean of $\text{Beta}(4,3)$ is

$E(\theta|s) = \frac{a}{a+b} = \frac{4}{4+3} = \frac{4}{7}$

Calculate mode for beta ← possible midterm question.

posterior is a combination prior and theta.

Ex $\theta|s \sim \text{Beta}(\sum x_i+\alpha, n-\sum x_i+\beta)$

$f_{\theta|s}(\theta) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)} \cdot \theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}$

$\ln f_{\theta|s}(\theta) \propto (\sum x_i+\alpha-1)\ln\theta + (n-\sum x_i+\beta-1)\ln(1-\theta)$

$\frac{\partial \ln f_{\theta|s}(\theta)}{\partial\theta} \propto \frac{\sum x_i+\alpha-1}{\theta} - \frac{(n-\sum x_i+\beta-1)}{1-\theta}$

$\Rightarrow \frac{\sum x_i+\alpha-1}{\theta} - \frac{n-\sum x_i+\beta-1}{1-\theta} = 0$

$\Rightarrow \frac{\sum x_i+\alpha-1}{\theta} = \frac{n-\sum x_i+\beta-1}{1-\theta}$

$\frac{1-\theta}{\theta} = \frac{n-\sum x_i+\beta-1}{\sum x_i+\alpha-1}$

$\Rightarrow \frac{1}{\theta} = \frac{n-\sum x_i+\beta-1+\sum x_i+\alpha-1}{\sum x_i+\alpha-1}$

$= \frac{n+\alpha+\beta-2}{\sum x_i+\alpha-1}$

$\theta = \frac{\sum x_i+\alpha-1}{n+\alpha+\beta-2}$

For this example

$Y \sim \text{Bern}(\theta)$

$\pi(\theta) = \text{Beta}$ ← start with some distribution

$\text{post} = \text{Beta}$ ← same distribution

$\Rightarrow$ "conjugate"

if prior distribution is post distribution then it is a conjugate.

Ex $Y \sim \text{Pois}(\lambda)$ / $\lambda > 0$

$\lambda \sim \text{Gamma}(\alpha, \beta)$

$\lambda|s \sim \text{Gamma}(\alpha^*, \beta^*)$

$\lambda \sim \text{Uniform}[0, \infty)$

is a non informative prior because $\lambda > 0$. Its also a improper because integration does not $= 1$.

if you have conjugate prior, we don't have to do $m(s)$, makes calculations easier.

$\pi(\theta) \approx \text{Uniform}[0,1]$ ← non-informative prior.

if prior is not a valid pdf, it is an improper prior

Ex/ $\pi(\theta) \sim \text{Beta}(\alpha=1, \beta=1)$

$\text{Beta}(\alpha, \beta)$ then

$$\text{pdf } \frac{\Gamma(a+\beta)}{\Gamma(a)\Gamma(\beta)} x^{a-1}(1-x)^{\beta-1}$$

$$\Rightarrow \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} x^{1-1}(1-x)^{1-1}$$

$$= \Gamma(2)$$

$$= (1-1)!$$

$$= 1$$

$\Gamma(3.5) = 2.5 * 1.5 * .5$

$\Gamma(x+1) = x!$

$\Gamma(0.5) = \sqrt{\pi}$

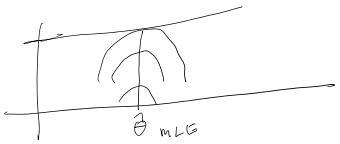
Note: $\text{Beta}(1,1) = \text{Uniform}[0,1]$

mind blown

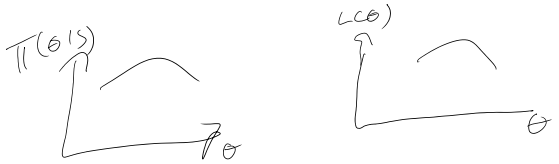
mode: $\frac{\sum x_i}{n} = \bar{x}$ | To know this MLE of $\text{Bern}(\theta)$ ← $\hat{\theta} = \bar{x}$

post $\propto$ Likelihood

$\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$



if non informative prior, whatever MLE is the mode



Both curves look the same but will be shifted up and down. But the maximum is the same.
This only the case when the prior is uninformative.

only if non informative dawg!

Ex $x_1, x_2, \ldots x_n \sim N(\mu, \sigma_0^2)$ σ_0 known

$\mu \sim N(\mu_0, \tau_0^2)$

$(\mu|s) \sim N(\mu^*, \sigma^{*2})$

posterior

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

$$\mu^* = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma_0^2}\bar{x}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}} = \frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}}\mu_0 + \frac{\frac{n}{\sigma_0^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}}\bar{x}$$

w_1 w_2

$\Rightarrow w_1 + w_2 = 1$

$= w_1 x + w_2 y$

↳ weighted average

eg $\frac{3+5}{2}$

$\frac{1}{2}(3) + \frac{1}{2}(5)$

w_1 w_2

3 | 5
$(\frac{1}{2})$ $(\frac{1}{2})$

3, 3 | 5

$\text{avg} = \frac{3+3+5}{3}$

$= \frac{2}{3}(3) + \frac{1}{3}(5)$

↓ in this case 3 is pulling the mean by $\frac{2}{3}$ and 5 is pulling by $\frac{1}{3}$

posterior mean → weighted average of the prior mean (μ_0) and sample mean ($\bar{x}$)

• ⇐ red is posterior fight between prior and sample mean, end result.

if $\sigma_0^2 = \tau_0^2 = 1$, then

$$\frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}}\mu_0 + \frac{\frac{n}{\sigma_0^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}}\bar{x} = \frac{1}{1+n}\mu_0 + \frac{n}{1+n}\bar{x}$$

$(x_1, x_2, \dots x_n)$ mean $= \frac{\sum x_i}{n}$

$(x_1, x_2, \dots x_n, \mu_0)$ mean $= \frac{\sum x_i + \mu_0}{n+1}$

if $\tau_0^2 \to \infty$

$$\frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}}\mu_0 + \frac{\frac{n}{\sigma_0^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}}\bar{x} = \frac{0}{0+\frac{n}{\sigma_0^2}} + \frac{\frac{n}{\sigma_0^2}}{0+\frac{n}{\sigma_0^2}}\bar{x}$$

$$= \bar{x}$$

⇒ if variance of Normal is really big then if I'm not a informative prior. ∴ wherever data is posterior is, posterior will be the same one.

Can't you should do: something μ is known but σ_0^2 is unknown

$x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2)$ → unknown; → known

prior, $\frac{1}{\sigma^2} \sim \text{Gamma}(\alpha, \beta)$

$(\frac{1}{\sigma^2} | s) \subset$ post

$$\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha-1} e^{-\beta\cdot\frac{1}{\sigma^2}}$$

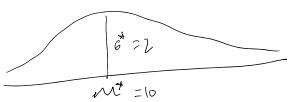
normal pdf σ^2 is always on the bottom makes calculation easier

Recall

$G(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

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

what is post? ⊂ excercise. pls do this fhem.



"Credible region" instead of a Confidence interval.

normal distribution mean = mode = median

MSE = Bias^2 + var ⇐ know this fhem its on the midterm

imagine archery

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

Bias = shoot repeatedly, you are averaging the target.