3.6) poisson pmf $poi(x/\lambda) = e^{-\lambda}\frac{\lambda^{x}}{x!}$

To estimate the parameter λ, we do MLE by solving optimization problem

$$\cancel{d}\,\lambda_{MLE} = \arg\max_{\lambda} poi(D/\lambda)$$

$$\approx \arg\max_{\lambda} \log(poi(D/\lambda))$$

To solve this, we'll do differentiation ~~on both sides~~ on log-likelihood.

$$\boxed{poi(D/\lambda) = \prod_{i=1}^{N} \cancel{\log}\, e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}}$$

d $$\frac{d}{d\lambda}(\log(poi\, D/\lambda)) = 0$$

$$\frac{d}{d\lambda}\left(\sum_{i=1}^{N} \log e^{-\lambda}\cdot\frac{\lambda^{x_i}}{x_i!}\right) = 0$$

$$\frac{d}{d\lambda}\left(\sum_{i=1}^{N} \log e^{-\lambda} + \sum_{i=1}^{N} \log \lambda^{x_i} + \sum_{i=1}^{N} \log x_i!\right) = 0$$

$$\frac{d}{d\lambda}\left(-N\lambda + (\log\lambda)\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} \log x_i!\right) = 0$$

$$-N + \frac{1}{\lambda}\sum_{i=1}^{N} x_i = 0 \Rightarrow \lambda_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$$

So, the MLE estimate for parameter is empirical mean. When N is large, this coincides with the expected value of distribution

for N large, $E(x \sim poi(x/\lambda)) = \lambda = \lambda_{MLE}$

3.8)
a) Uniform distribution Centered on 0 with width 2a.

$$x \sim Unif(0, 2a) = \frac{1}{2a} I(x \in [-a, a])$$

To estimate the parameter a, we ~~do~~ calculate MLE by solving optimization problem for likelihood.

$$\hat{a}_{MLE} = \arg\max_{a} P_{Unif}(D/a) \quad D = dataset$$

$$= \arg\max_{a} \prod_{i=1}^{N} \frac{1}{2a} I(x \in [-a, a])$$

$$= \frac{1}{\cancel{2a}} \arg\max_{a} \left(\frac{1}{2a}\right)^{N} \underbrace{\prod_{i=1}^{N} I(x \in [-a, a])}$$

Inorder to maximize this, 'a' value should be minimum and in between $[-a, a]$

$$So, \hat{a}_{MLE} = \min(x_i \in [-a, a])$$

3.11)
a) MLE estimation for exponential distribution

$$x \sim \exp(x/\theta) = \theta e^{-\theta x}$$

To estimate parameter θ, we ~~do~~ calculate MLE ~~on~~ by solving optimization problem for likelihood.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \exp(D/\theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{N} \theta \cdot e^{-\theta x_i}$$

$$\sim \arg\max_{\theta} \sum_{i=1}^{N} \log(\theta \cdot e^{-\theta x_i})$$

Scanned by CamScanner

To get maximum for θ, we do differentiation

$$\frac{d}{d\theta}\left(\sum_{i=1}^{N} \log \theta \cdot e^{-\theta x_i}\right) d\theta = 0$$

$$\sum_{i=1}^{N} \cdot 1/\theta - \sum_{i=1}^{N} x_i = 0$$

$$\frac{N}{\theta} = \sum_{i=1}^{N} x_i \Rightarrow \theta = \frac{1}{\sum_{i=1}^{N} x_i / N}$$

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

If we say $\bar{x}$ is empirical estimate for mean

then MLE ~~for~~ parameter θ is

$$\hat{\theta}_{MLE} = \frac{1}{\bar{x}}$$

and when N is large $\bar{x} = E(X)$

So, MLE substantiates that exponential factor is $1/E(x)$.

3.11 b) According to above derivation

$$\hat{\theta}_{MLE} = \frac{1}{\bar{x}}$$

for $\{x_1 = 5, x_2 = 6, x_3 = 4\}$ dataset

$$\bar{x} = \frac{5+6+4}{3} = 5$$

$$\hat{\theta}_{MLE} = \frac{1}{\bar{x}} = 1/5 = 0.2$$

4.1) Its given that $X \sim U(-1,1)$

$$Y = X^2$$

$$E(X) = \int_{-1}^{1} x \cdot p(x)\, dx = \int_{-1}^{1} x \cdot \frac{1}{2}\, dx = \left(\frac{x^2}{4}\right)_{-1}^{1} = 0$$

$$E(Y) = \int_{-1}^{1} x^2 \cdot p(x)\, dx = \int_{-1}^{1} x^2 \cdot \frac{1}{2}\, dx = \frac{1}{2}\left(\frac{x^3}{3}\right)_{-1}^{1}$$

$$= \frac{1}{3}$$

Correlation Coefficient

$$\rho(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)\, Var(y)}}$$

$$Cov(x,y) = E\left[(x - E(x))(y - E(y))\right]$$

$$= E(xy) - E(x)E(y)$$

$$E(xy) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \underbrace{}_{}^{p(x)\,p(y)|} dx\, dy$$

$$= \int_{-\infty}^{\infty} x \cdot \left(\int_{-\infty}^{\infty} y \underset{p(y)}{\cdot} dy\right) d\, p(x)\, dy$$

$$= \int_{-\infty}^{\infty} x \cdot \left[\int_{-1}^{1} x^2 p(x)\, dx\right] dx \quad \left[\begin{array}{l}\text{change of} \\ \text{variables} \\ \text{formula}\end{array}\right]$$

$$= \int_{-1}^{1} x \cdot \left[\frac{1}{2} \cdot \frac{x^3}{3}\right]_{-1}^{1} dx = \int_{-1}^{1} \frac{1}{3} \cdot x\, dx$$

$$= \left(\frac{x^2}{6}\right)_{-1}^{1} = 0$$

So, $Cov(x,y) = 0 - 0 = 0$.

From this we can say, even though x, y are not independent, their $Cov(x,y)$ ~~is 0~~. can be 0.

4.2) a $X \sim N(0,1)$

and $Y = WX$, W is ~~dis~~ r.v $\begin{cases} 1 \\ -1 \end{cases}$

Here W is discrete random variable with support $\{-1, 1\}$ and

$$p(W) = 0.5$$

Now, we can see random variable

$$Y = \begin{cases} -X & \text{When } W = -1 \\ X & \text{When } W = 1 \end{cases}$$

As x fallows Gaussian $N(0,1)$

$$P(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

If we put, $-x$ in place x we get same equation

So, $P(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$, so Y also fallows Gaussian with same parameters 0, 1.

$$Y \sim N(0,1)$$

b) $\cdot \operatorname{cov}(X,Y) = E(XY) - E(X)E(Y)$

Since X, Y are both Gaussian $E(X) = 0$
$E(Y) = 0$

$$E(XY) = E[E(XY/W)]$$

$$= 0.5\, E[XY/W=-1] + 0.5\, E[XY/W=1]$$

$$= 0.5\, E[-X^2] + 0.5\, E[X^2]$$

$$= 0.5\, E[X^2 - X^2] = 0$$

4.3) we can prove $\rho(x,y)$ (correlation coefficient) is always between -1 and 1. by using Cauchy - Schwarz inequality

$$[\operatorname{cov}(x,y)]^2 \le \operatorname{Var}(x)\operatorname{Var}(y)$$

So, $|\operatorname{cov}(x,y)| \le \sqrt{\operatorname{Var}(x)\operatorname{Var}(y)}$

$$|\rho(x,y)| \le 1.$$

$$-1 \le \rho(x,y) \le 1$$

4.4) Correlation coefficient for linearly related variables is ± 1.

let y is linearly dependent on x

$$y = ax + b$$

$\operatorname{Var}(x) = E(x^2) - E^2(x)$ $\quad \operatorname{Var}(y) = a^2[E(x^2) - E^2(x)]$

$$\operatorname{cov}(x,y) = E(xy) - E(x)E(y)$$

$$= E[x(ax+b)] - [E(x)(aE(x)+b)]$$

$$= aE(x^2) + b\cancel{E}(x) - aE^2(x) - b\cancel{E}(x)$$

$$= aE(x^2) - aE^2(x)$$

$$\rho(x,y) = \frac{\operatorname{Cov}(x,y)}{\sqrt{\operatorname{Var}(x)\operatorname{Var}(y)}} = \frac{a(E(x^2) - E(x))}{\sqrt{a^2 \cdot [E(x^2) - E^2(x)]^2}} = \frac{a}{|a|} = \pm 1.$$

so, when $a > 0 \Rightarrow \frac{a}{|a|} = 1$

and $a < 0 \Rightarrow \frac{a}{|a|} = -1$

* We know that posterior

$$p(\theta/D) \propto P(D/\theta)P(\theta)$$

θ - parameters to be estimated from data

D - Dataset

$P(D/\theta)$ - likelihood

$p(\theta)$ - prior for the parameters.

the Maximum A posteriori (MAP) estimate for θ

$$\hat{\theta}_{MAP} \Rightarrow \underset{\theta}{\operatorname{argmax}}\, p(D/\theta)\, p(\theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \left[\log p(D/\theta) + \log p(\theta)\right]$$

In this equation prior is always constant, so we can write it

$$\hat{\theta}_{MAP} \approx \underset{\theta}{\operatorname{argmax}} \log P(D/\theta) = \hat{\theta}_{MLE}$$

This is MLE (Maximum likelihood estimate) for θ and as the dataset size (N) increases $P(D/\theta)$ converges to single θ. This is nothing when data size increases it overwhelms the prior

This is analogous to central limit theorm, where when we have enough sample data, parameters calculated on samples best fit the true original underlying distribution. Dat...

In same way, as we are trying to optimize based on liklihood data [$\arg\max_{\theta} p(D/\theta)$], when we get enough data, this estimated parameters after solving optimization problem converges to true parameters..