WHOI Math Review

Lecture notes: Probability & Statistics

Dan Amrhein

24 July 2015

Outline

1. Solving systems of eqns in MATLAB using matrices
2. Probability
3. Characterizing random variables
4. The Gaussian distribution
5. Tests of hypotheses & significance

$$\underset{\text{obs}}{y} = \underset{\text{model}}{E}\,\underset{\text{unknowns}}{x} + \underset{\text{noise}}{n}$$

$$J = \|n\| = n^T n = (y - Ex)^T (y - Ex)$$

We want to minimize J.

$$dJ = \sum_i \frac{\partial J}{\partial x_i}\, dx_i = \left(\frac{\partial J}{\partial \underline{x}}\right)^T d\underline{x}$$

$$= 2\, d\underline{x}^T (E^T y - E^T E x) = 0$$

For this to be true,

$$E^T \underline{y} - E^T E \underline{x} = 0 \qquad \text{"normal eqns"}$$

$$\underline{\tilde{X}} = (E^T E)^{-1} E^T \underline{y}$$

```
>> inv(E'*E)*E'*[0,2]'
```

Underdetermined

$$1 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$ ← can always fit the data! But if it is noisy, $\exists$ a danger of overfitting

Rank deficient

$$\begin{matrix} 1 \\ 1 \end{matrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

a singular matrix!

row rank = # of linearly independent rows

column rank

rank $= K$. If $K < \min(M, N)$, $\Rightarrow$ rank deficient. SVD!

Probability

If something happens h times out of n trials, and n is very large, $p = \frac{h}{n}$

$0 \le p \le 1$

What is the probability of rolling a total of 7 in two die casts?

$$\left.\begin{matrix} 6 & 5 & 4 & 3 & 2 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{matrix}\right\} 6$$

$\frac{1}{36}$ $\frac{1}{36}$ ... $\frac{6}{36} = \frac{1}{6}$

2 principles:

1. Probabilities of mutually exclusive events <u>add</u>
2. Probabilities of multiple independent events occurring <u>multiply</u>

What does it mean to be independent?

$$P(B|A) = P(B)$$

More generally,

"and"

$$P(B|A) \equiv \frac{P(A \cap B)}{P(A)}$$

A = I have a cat

B = " dog

$$P(A \cap B) = P(B|A)\,P(A)$$

probability I have a cat

prob I have a cat and a dog

prob I have a dog given that I have a cat

Bayes' Theorem

$$P(B \cap A) = P(A|B)\,P(B)$$

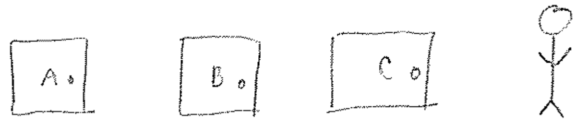
$$P(A \cap B) = P(B \cap A)$$

$$\Rightarrow \quad P(B|A)\,P(A) = P(A|B)\,P(A)$$

$$\Rightarrow \quad P(B|A) = \frac{P(A|B)\,P(A)}{P(B)} \qquad \text{BAYES' THM}$$

Monty Hall Problem

3 doors:



2 goats, 1 car

1. You guess a door
2. Monty opens one of the other two and reveals a goat.
3. Stay or switch?

Assume we chose A. Label the door the host chose as B.

@A = car behind A $\quad 1/3$
@C = car behind C $\quad 1/3$
oB = host opens B $\quad 1/2$

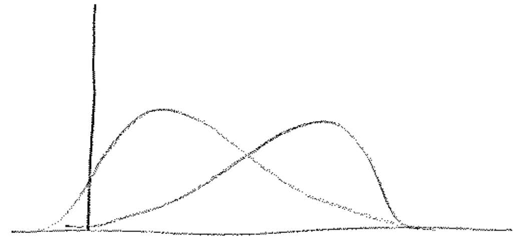
$$P(@A \mid oB) = \frac{P(oB \mid @A)\,P(@A)}{P(oB)} = \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

$$P(@C \mid oB) = \frac{P(oB \mid @C)\,P(@C)}{P(oB)} = \frac{1\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

3 cases.

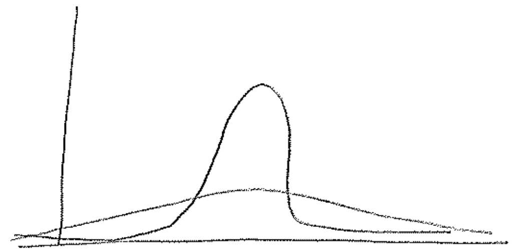
Skewness (third moment)

$$\alpha_3 = \frac{E\left[(X-\mu)^3\right]}{\sigma^3}$$



Kurtosis (Fourth moment) - peakedness

$$\alpha_4 = \frac{E\left[(X-\mu)^4\right]}{\sigma^4}$$



"long tailed" "multimodal"

Covariance & correlation

$$\sigma_{xy} = Cov(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right]$$

empirical est: >> cov

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad -1 \le \rho \le 1$$

$\rho = 0 \Rightarrow$ uncorrelated (not necessarily independent!)

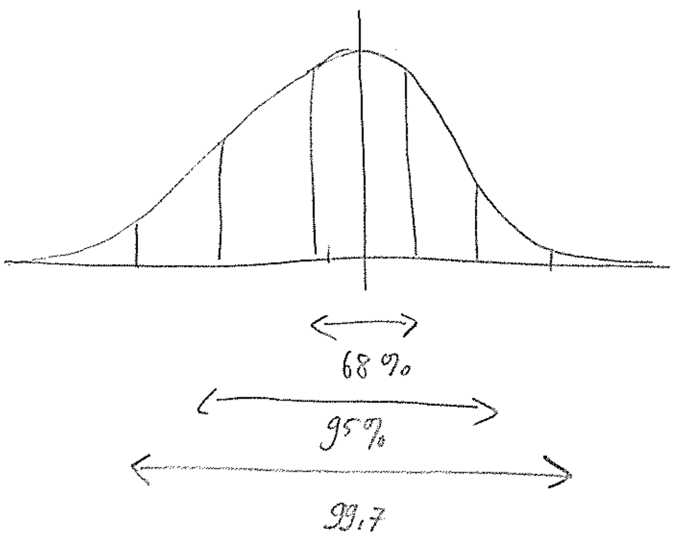
Stationarity - A stochastic process (e.g. temperature at a point) is stationary if its statistics (mean, variance, covariance, etc) do not change in time

The Gaussian distribution

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

CLT: Under certain conditions, the arithmetic mean of a large number of independent RVs will be approximately normally distributed, regardless of the underlying distribution

Characterized completely by μ, σ



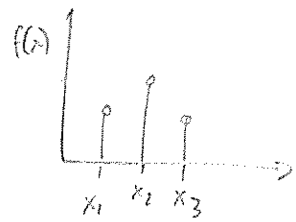
Random Variables

RV = a var whose value is subject to chance. Denote as uppercase

Discrete, cts RVs exist.

Probability density fns

Discrete PDF :

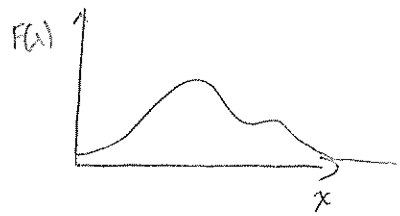


$P(X=x) = f(x)$

$f(x) \geq 0$

$\sum_x f(x) = 1$

Cts PDF :



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

Independent RVs: if $X=x$ and $Y=y$ are indep $\forall\, x, y$. Then

$$P(X=x, Y=y) = P(X=x)\,P(Y=y)$$

$$f(x,y) = f_1(x) f_2(y)$$

joint prob.

Expectation (first moment)

Discrete: $\mu = E(X) = \langle x \rangle = x_1 P(X=x_1) + x_2 P(X=x_2) + \dots$

$$= \sum_{j=1}^{n} x_j P(X=x_j)$$

When all probs =,

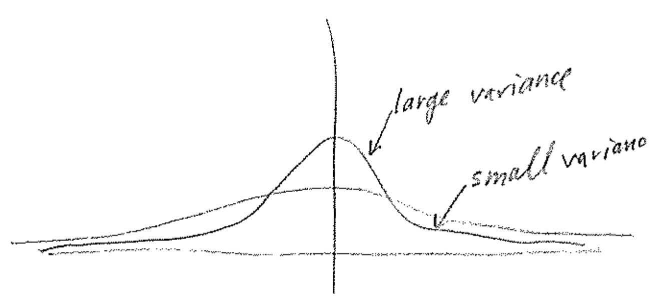
$$E(X) = \frac{x_1 + x_2 + \dots x_n}{n} = \text{mean}$$

cts: $E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$

Variance (second moment)

$\mathrm{Var}(X) = E\left[(X - E(X))^2\right]$ (units of X^2)

Standard deviation: $\sigma^2 = \mathrm{Var}(X)$



Tests of hypotheses & significance

How can we use observations to make decisions?

Often in science we want to choose between hypotheses, e.g.

(null) H_0: MOC during ice ages was the same as today's

H_1: MOC transports were reduced

Type I error: reject a hyp when it is true

Type II error: accept a hyp when it should be rejected

Level of significance: maximum probability w/ which we are willing to risk a Type I error

p value: the probability of observing a test statistic at least as extreme as the one observed given that Ho is true.

Statistical tests

Ex.: Pearson's χ^2 test

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

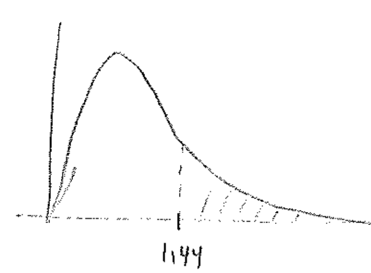
O_i = obs freq for i^{th} bin

E_i = expected freq for i^{th} bin

Ho: A random sample of 100 pets has been drawn from an equal dist of cats & dogs.

Say 44 cats, 56 dogs.

$$\chi^2 = \frac{(44-50)^2}{50} + \frac{(56-50)^2}{50} = 1.44$$




```
>> chi2cdf(1.44,1)
ans =
    0.7699
```

$\Rightarrow p = 0.23$

Would not reject at 0.05 sig level

chi2gof - useful for testing Gaussianity

CI