Part II: Probability Models

Motivating example: (Meerschaert 7.2)

Two dice are rolled and the bank pays the player the sum of both rolls in dollars. How much would you pay to play this game?

Definition:

A random variable X on a sample space S is a function $X: S \to \mathbb{R}$ associated with the random outcome of some experiment. (or subset of $\mathbb{R}$)

There are two types of random variables:

1) Discrete:

X can attain any of a discrete set of values,

$$X \in \{x_1, x_2, x_3, \ldots, x_N\} = \{x_i\}, \quad i = 1, 2, \ldots, N$$

X attaining each of these values can be thought of as an event E_i, and the probability of attaining these events is denoted by

$$Pr\{E_i\} = Pr\{X = x_i\} = p_i > 0$$

where

$$\sum_{i=1}^{N} p_i = 1 \qquad \text{and} \qquad 0 \le p_i \le 1 \quad \forall i$$

2) Continuous:

X can attain values on the real line, $\mathbb{R}$. The probability associated with outcomes of an experiment can be described using a function,

$$F(x) = Pr\{\mathbb{X} \leq x\}$$

This is called a (cumulative) distribution function. Note that if $F(x)$ is differentiable, we call $f(x) = F'(x)$ the probability density function of $\mathbb{X}$, so that

$$P\{a < \mathbb{X} \leq b\} = F(b) - F(a) = \int_a^b f(x)\,dx$$

Note: So $P\{\mathbb{X} = x\} = 0$, but nonzero if we consider a range of values

· Definition:

The average or expected value (aka the expectation) of $\mathbb{X}$ is denoted $E[\mathbb{X}]$, and is defined as

$\leftarrow$ $E\mathbb{X}$ in Meerschaert

1) Discrete:

$$E[\mathbb{X}] = \sum_{i=1}^{N} x_i p_i$$

2) Continuous:

$$E[\mathbb{X}] = \int_{-\infty}^{\infty} x f(x)\,dx$$

Back to example:

One die: $\mathbb{X} = \{1, 2, 3, 4, 5, 6\}$, with uniform probability, $p = \frac{1}{6}$

Then

$$Pr\{\mathbb{X} = 4\} = \frac{1}{6}$$

$$Pr\{\mathbb{X} = 4 \text{ or } \mathbb{X} = 5\} = \frac{1}{6} + \frac{1}{6}$$

$$Pr\{\mathbb{X} \leq 3\} = \frac{1}{2}$$

etc.

Two dice: $\overset{6\times 6}{36}$ possible outcomes, each equally likely

$$\Pr\{\mathbb{X}=2\}=\frac{1}{36}$$

$$\Pr\{\mathbb{X}=3\}=\frac{1}{36}+\frac{1}{36}=\frac{2}{36}$$

roll a 1 and a 2

or 2 and a 1

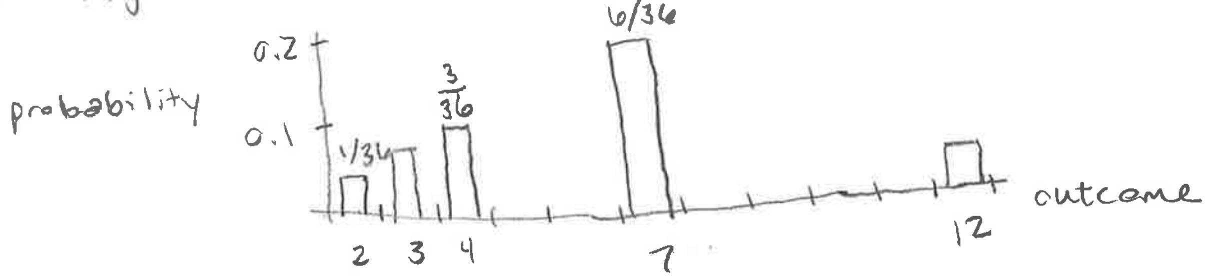
$$\vdots$$

$$\Pr\{\mathbb{X}=7\}=\frac{6}{36}$$

$$\vdots$$

$$\Pr\{\mathbb{X}=12\}=\frac{1}{36}$$

A histogram is a good way to visualize these probabilities: see Figure 7.2 in Meerschaert.



The expected value of $\mathbb{X}$ is

$$E[\mathbb{X}]=2\left(\frac{1}{36}\right)+3\left(\frac{2}{36}\right)+\cdots+12\left(\frac{1}{36}\right)=7$$

So it appears that the average, or expected, payout is \$7. You wouldn't want to pay more than \$7 to play.

We could also model this problem by considering each die as an <u>independent</u> random variable.

Definition:

Let Y and Z be two random variables, with
$$Y \in \{y_1, y_2, \dots, y_N\} \text{ and } Z = \{z_1, \dots, z_M\}$$ ← discrete

we say Y and Z are independent if
$$Pr\{Y = y_i \text{ and } Z = z_j\} = Pr\{Y = y_i\} Pr\{Z = z_j\} \quad \forall i, j$$

i.e., each out come Y_i does not influence Z_j

Theorem: (Law of Large Numbers)

Let $\{X_1, \dots, X_n\}$ be a random sample; that is, a repetition of n independent and identically distributed random variables. Then
$$\frac{X_1 + \cdots + X_n}{n} \to E[X] \quad \text{as } n \to \infty$$

Definition:

Random variables $X_1, \dots, X_n$ are identically distributed if each X_j has the same underlying set $\{x_1, \dots, x_N\}$, with associated probabilities p_i, $i = 1, \dots, N$

Note: Independent, identically distributed is often abbreviated as 'i.i.d.'

29/1

Last time: two random variables $\mathbb{Y}$ and $\mathbb{Z}$ are independent if

$$\Pr\{\mathbb{Y}=y_i \text{ and } \mathbb{Z}=z_j\} = \Pr\{\mathbb{Y}=y_i\}\Pr\{\mathbb{Z}=z_j\} \quad \forall i,j$$

Example: $\mathbb{X}\in\{x_1,x_2]$ with $p_1=0.8$ and $p_2=0.2$

$\mathbb{Y}\in\{y_1,y_2\}$ with $p_1=0.5$ and $p_2=0.5$

Suppose you are given

$\Pr\{\mathbb{X}=x_1 \text{ and } \mathbb{Y}=y_1\}=0.4$

$\Pr\{\mathbb{X}=x_2 \text{ and } \mathbb{Y}=y_1\}=0.1$

$\Pr\{\mathbb{X}=x_1 \text{ and } \mathbb{Y}=y_2\}=0.4$

$\Pr\{\mathbb{X}=x_2 \text{ and } \mathbb{Y}=y_2\}=0.1$

Then $\mathbb{X}$ and $\mathbb{Y}$ are independent because

$$\Pr\{\mathbb{X}=x_i \text{ and } \mathbb{Y}=y_j\} = \Pr\{\mathbb{X}=x_i\}\Pr\{\mathbb{Y}=y_j\} \quad \forall i,j$$

$$0.4 = 0.8\cdot 0.5$$

$$0.1 = 0.2\cdot 0.5$$

Example:

An electronics manufacturer produces a variety of diodes. Need to 0.3% of the diodes produced will be faulty. (estimated) do a quality analysis to test diodes, either individually or in a group. Options:

electrical component that primarily allows current in 1 direction while blocking it in the other direction

① Test $n>1$ diodes in a group (cost: $4+n$ cents)

② If a group fails, each diode must be tested (cost: 5 cents/diode)

What is the most effective quality control procedure?

Variables: n = # diodes per group

C = testing cost for one group ← random variable

A = average testing cost per diode

$q = 0.003$ = probability that a diode is faulty

Goal: Choose n such that A is minimized.

Model formulation:

$C \in \{c_1, c_2\} = \{$success/no faulty diodes, failure/≥ 1 faulty diode$\}$

If sucess:

$c_1 = 4+n \quad , \quad p_1 = \underline{(1-q)^n}$

Note: we've assumed that diode failures are independent

If failure:

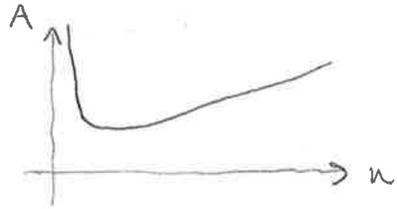
$c_2 = 4+n+5n \quad , \quad p_2 = 1-p_1 = 1-(1-q)^n$

Then the average cost is

$$E[C] = c_1 p_1 + c_2 p_2$$
$$= (4+n)(1-q)^n + \left[(4+n)+5n\right](1-(1-q)^n)$$
$$= 4+6n-5n(1-q)^n$$

So the average cost per diode is

$$A = \frac{E[C]}{n} = \frac{4}{n} + 6 - 5(1-q)^n$$



Minimize A: find $A'(n) = 0$ where $A''(n) > 0$

29/3

$$A'(n) = -\frac{4}{n^2} - 5\ln(1-q)(1-q)^n \qquad \leftarrow \left(\frac{d}{dx}(a^x) = a^x \ln a\right)$$

$$A''(n) = \frac{8}{n^3} - 5(\ln(1-q))^2(1-q)^n$$

Using Newton's Method:

$$n^* \approx 16.7331 \rightarrow n^* = 16 \quad \text{or} \quad n^* = 17$$

$$\underbrace{A(16) \approx 1.4847 \qquad A(17) \approx 1.4843}$$

Note: as in optimization section, need to test both values for optimal A

$\Rightarrow$ The optimal # of diodes per group is $n = 17$.

The sensitivity of A to q is

$$S(A,q) = \frac{dA}{dq} \cdot \frac{q}{A} \bigg|_{n=n^*} = \frac{5n(1-q)^{n-1}q}{\frac{4}{n} + 6 - 5(1-q)^n} \bigg|_{n=n^*}$$

$$\approx 0.163$$

We also may want to investigate the robustness of our conclusion to the assumption that faulty diodes are independently distributed.

↳ Assume the faulty diodes always occur in clusters of k diodes. How does this affect the optimal value of n?

Example: Diodes as dependent random variables by batch

Let C = testing cost for one group $= \{C_1, C_2\}$

29/4

Then if testing results in a success,
$C_1 = 4 + n$, $p_1 = ?$

Idea:

Let $q = 0.003$ denote the frequency of faulty diodes.
The probability of a diode to be first in a batch of $K < n$ faulty diodes is $\frac{q}{K}$
why? ①

For the test to be sucessful, $\underbrace{(n+k-1)}_{②}$ diodes must not be a 'first found faulty' diode.

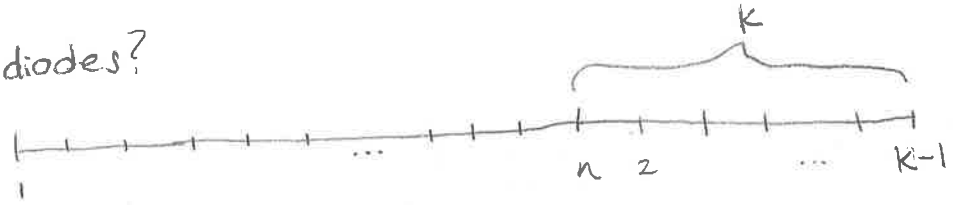
① The probability of a diode to be first in a batch of k faulty diodes:

$$\Pr\{\text{faulty AND first in a set of } k \text{ diodes}\}$$

$$= \Pr\{\text{faulty}\} \Pr\{\text{first} \mid \text{faulty}\}$$

$$= q\left(\frac{1}{k}\right)$$

② How many diodes?



↑
if first in a batch of faulty diodes

We need not only our batch of n diodes, but the next k-1 diodes to not be faulty as well. → (n+k-1) diodes

Updated model:

$$E[c] = \underbrace{c_1 p_1}_{\text{sucess}} + \underbrace{c_2 p_2}_{\text{failure}}$$

$$= (4+n)\left(1-\frac{q}{k}\right)^{n+k-1} + (4+n+5n)\left[1-\left(1-\frac{q}{k}\right)^{n+k-1}\right]$$

$$= 4 + 6n - 5n\left(1-\frac{q}{k}\right)^{n+k-1}$$

$$\Rightarrow \tilde{A} = \frac{4}{n} + 6 - 5\left(1-\frac{q}{k}\right)^{n+k-1}$$

For

$k=1$:	$n^* \approx 16.7$	$\tilde{A}(17) \approx 1.484$	
$k=3$:	$n^* \approx 28.7$	$\tilde{A}(17) \approx 1.329$	$\tilde{A}(29) \approx 1.291$
$k=10$:	$n^* \approx 52.1$	$\tilde{A}(17) \approx 1.274$	$\tilde{A}(52) \approx 1.168$

In this example we used the idea of conditional probability.

Definition:

The conditional probability of E given F is

$$Pr\{E \mid F\} = \frac{Pr\{E \cap F\}}{Pr\{F\}}, \quad Pr\{F\} \neq 0$$

Conversely, we can write

$$Pr\{E \cap F\} = Pr\{F\} Pr\{E \mid F\}$$

The conditional probability $Pr\{E|F\}$ can be thought of as the relative likelihood of event E occuring among all possible events once F has occurred.

Section 7.2: Continuous Probability Models

Recall: For a continuous random variable, $\mathbb{X} \in \mathbb{R}$,

(cumulative) distribution function: $F(x) = Pr\{\mathbb{X} \leq x\}$ (cdf)

(probability) density function: $f(x) = F'(x)$ (pdf)

In general,

$$Pr\{\mathbb{X} \in B\} = \int_B f(x)\,dx$$

where $f(x)$ must satisfy

$$1 = Pr\{\mathbb{X} \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)\,dx$$

$$f(x) \geqslant 0 \quad \forall x$$

Note:
If $B = [a,b]$,

$$Pr\{a \leq \mathbb{X} \leq b\} = \int_a^b f(x)\,dx = F(b) - F(a)$$

and

$$Pr\{\mathbb{X} \in (-\infty, x)\} = \int_{-\infty}^{x} f(z)\,dz = F(x)$$

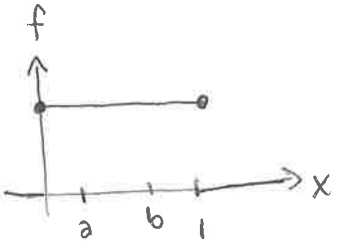
$\Rightarrow \lim_{x \to \infty} F(x) = 1$, $F(x)$ is a monotonic function

30/3

Example: Uniform distribution

A random variable X is said to be uniformly distributed over $x \in (0,1)$ if the density function is given by

$$f(x) = \begin{cases} 1 & , x \in [0,1] \\ 0 & , \text{otherwise} \end{cases}$$



Notice, for any $0<a<b<1$,

$$\int_a^b f(x)dx = F(b) - F(a)$$

$$= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} F(x)dx$$

$$= b - a$$

So the probability that X is any subinterval (a,b) is equal to the length of that interval. In general, we can write the pdf for a uniform random variable over any interval as

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & , x \in [\alpha, \beta] \\ 0 & , \text{otherwise.} \end{cases}$$

Example: Exponential distribution

A continuous random variable whose density function is given, for some $\lambda > 0$, by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$

is called an exponential random variable.

Then the distribution function is given by

$$F(x) = \int_0^x \lambda e^{-\lambda \xi} d\xi = 1 - e^{-\lambda x} \quad , \quad x \geq 0$$

The exponential distribution has the important property of a 'lack of memory': consider $\Pr\{\underline{X} > s+t \mid \underline{X} > s\}$, the cond'l probability of an event occuring after $s+t$ seconds, given that the event has not occured in the (for example) previous s seconds.

By definition,

$$\Pr\{\underline{X} > s+t \mid \underline{X} > s\} = \frac{\Pr\{\underline{X} > s+t\}}{\Pr\{\underline{X} > s\}}$$

$$= \frac{\int_{s+t}^{\infty} f(x)dx}{\int_s^{\infty} f(x)dx} = \frac{\int_{s+t}^{\infty} \lambda e^{-\lambda x} dx}{\int_s^{\infty} \lambda e^{-\lambda x} dx}$$

$$= \frac{\lim_{x \to \infty} (-e^{-\lambda x}) + e^{-\lambda(s+t)}}{\lim_{x \to \infty} (-e^{-\lambda x}) + e^{-\lambda(s)}}$$

$$= \frac{e^{-\lambda s} e^{-\lambda t}}{e^{-\lambda s}}$$

$$= e^{-\lambda t}$$

$$= \Pr\{\underline{X} > t\}$$

Example:

A Geiger counter is used to measure the activity of a radioactive source: a decay event triggers a discharge that can be registered, but the events occur at random, with an unknown rate.

After each decay event, the Geiger counter is insensitive for a 'dead time' of $a = 3\times10^{-9}$ sec.

In T seconds, n events have been counted. What is the decay rate, λ?

Variables: λ = decay rate (seconds)

T_n = time at the n^{th} observed decay event.

$a = 3\times10^{-9}$ sec. = dead time for Geiger counter

Model:

Time between observations of decay event:

$$T_n - T_{n-1} \equiv X_n \geq 3\times10^{-9} = a$$

After a decay event and dead time, need to wait until next event: let

$$X_n \equiv a + \underbrace{Y_n}_{\text{time after dead time until next event}}$$

Assumptions:

because radioactive decay occurs without memory

- Y_n is expenentially distributed with rate λ
- X_n is not exponentially distributed: need $X_n > 3\times10^{-9}$, which we can't guarantee with the exponential distribution
- Times between decays are independent

Goal: Find λ, in terms of T and n we can use the expected value of Y_n,

$$E[Y_n] = \int_0^\infty t\lambda e^{-\lambda t}\,dt \underset{\substack{\uparrow \\ \text{IBP}}}{=} \frac{1}{\lambda}$$

since $X_n = a + Y_n$, then $E[X_n] = a + E[Y_n] = a + \frac{1}{\lambda}$

Now, consider $X_n = T_n - T_{n-1}$. By the law of large numbers,

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = a + \frac{1}{\lambda}$$

$$= \lim_{n\to\infty} \frac{(\cancel{T_1} - \overset{0}{T_0}) + (\cancel{T_2} - \cancel{T_1}) + \cdots + (T_n - \cancel{T_{n-1}})}{n}$$

$$= \lim_{n\to\infty} \frac{T_n}{n}$$

so for large n, $\frac{T_n}{n} \approx a + \frac{1}{\lambda} \rightarrow \lambda\left(\frac{T_n}{n} - a\right) \approx 1$

$$\Rightarrow \lambda \approx \frac{1}{\frac{T_n}{n} - a} = \frac{n}{T_n - an}$$

So with the assumption that Y_n is expenentially distributed, the distribution of the decays in the observation window itself does not need to be calculated.

This type of problem, where the times between decay events are independent and exponentially distributed, can be classified as a type of arrival process called a Poisson process.

In general, a stochastic process $\{X(t), t \in T\}$ is a collection of random variables, where $X(t)$ is the state of the process at time t. The set T is called the index set of the process.

Definition:
A Poisson process $\{N(t), t \geq 0\}$ is a stochastic process where

1) $N(0) = 0$

2) The process has independent increments

3) The number of events in any time interval follows a Poisson distribution,

$$\Pr\{N(t+s) - N(s) = n\} = e^{-\lambda t}\frac{(\lambda t)^n}{n!} \quad , n = 0, 1, 2, \ldots$$

the parameter λ is called the Poisson rate.

Note: A Poisson process is commonly used to represent random arrivals, i.e. an 'arrival process'.

We also assumed that the number of events was large, which is not exactly realistic. We expect empirical observations to vary slightly from this mean.

31/4

Section 7.3: Statistics

Recall the definition of an expected value,

$$E[X] = \sum x_k p_k \qquad \text{or} \qquad E[X] = \int x f(x)\,dx$$

(discrete r.v.) (continuous r.v.)

The expectation gives us the mean of a distribution. We can also quantify the 'spread' of a distribution, or the extent to which X tends to deviate from the mean $E[X]$, as the variance of X:

$$V[X] = \sum (x_k - E[X])^2 p_k \qquad \text{or} \qquad V[X] = \int (x - E[X])^2 f(x)\,dx$$

(discrete)

Theorem: Central Limit Theorem
As $n \to \infty$. The distribution of $X_1 + \cdots + X_n$, $\forall$ r.v. X_i, approaches the normal distribution. I.e., if $\mu = E[X]$ and $\sigma^2 = V[X]$, then $\forall t \in \mathbb{R}$,

$$\lim_{n\to\infty} \Pr\left\{ \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le t \right\} \to \Phi(t)$$

where $\Phi(t)$ is the normal distribution.

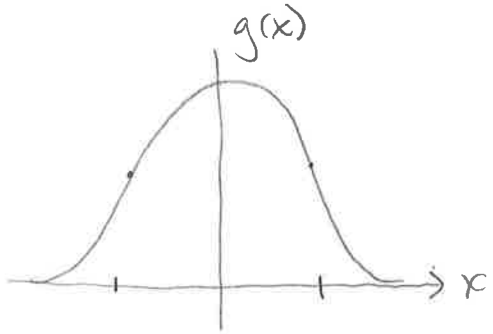
Definition: (Normal Distribution) (aka Gaussian distribution)
A random variable X follows the normal distribution if the density function (pdf) is given by

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

where $\mu = E[X]$ and $\sigma^2 = V[X]$, so that

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx$$

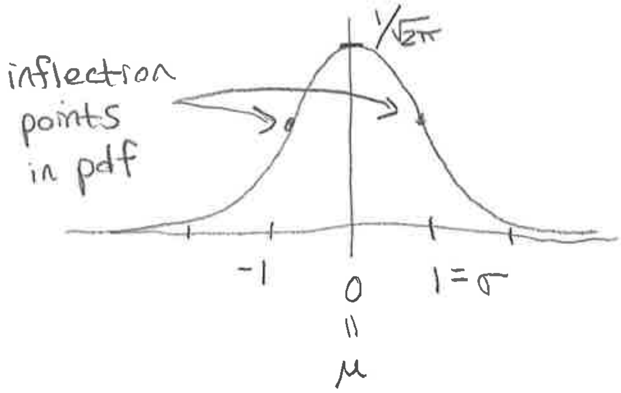
Note: The <u>standard</u> normal distribution. assumes a mean $\mu = 0$ and variance $\sigma^2 = 1$



In that case

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

<u>Properties of the normal distribution</u>



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

$\int_{-1}^{1} g(x)dx \approx 0.68$, $\int_{-2}^{2} g(x) \approx 0.95$, $\int_{-3}^{3} g(x) \approx 0.997$

↳ So $Pr\{-1 \leq X \leq 1\} \approx 68\%$

In general, for a normal distribution with μ and σ,

$$\Pr\{-\sigma \le X - \mu \le \sigma\} \approx 68\%$$

i.e., a result within one standard deviation occurs about 68% of the time.

Note: In particle physics, a discovery needs to have a 'certainty' of 5σ, or 99.99994%.

If you have n iid draws, for n sufficiently large, we get
(independent, identically-distributed)

$$n\mu - \sigma\sqrt{n} \le X_1 + X_2 + \cdots + X_n \le n\mu + \sigma\sqrt{n}$$

about 68% of the time.

$$n\mu - 2\sigma\sqrt{n} \le X_1 + X_2 + \cdots + X_n \le n\mu + \sigma\sqrt{n}$$

about 95% of the time. This is usually what we refer to when we say a result is 'statistically significant'

Markov Chains

Definition:

A Markov chain is a discrete-time stochastic model that consists of a sequence of random variables $\{X_n\}$ that

(a) have a discrete set of states:

$$X_n \in \{x_1, x_2, \dots, x_m\}$$

(b) The probability that $X_{n+1} = x_j$ only depends on X_n, and is given by

$$p_{ij} = \Pr\{X_{n+1} = x_j \mid X_n = x_i\}$$

Note: The sequence of the process $\{X_n\}$ is determined by each p_{ij} and the probability distribution for the initial value, X_0.

Example: $X_n \in \{1, 2, 3\}$

Rules:

1) If $X_n = 1$, then $X_{n+1} = 1, 2,$ or 3 are equally probable

2) If $X_n = 2$, then $X_{n+1} = 1$ with $p = 0.7$

$X_{n+1} = 2$ with $p = 0.3$

3) If $X_n = 3$, then $X_{n+1} = 1$ with $p = 1$.

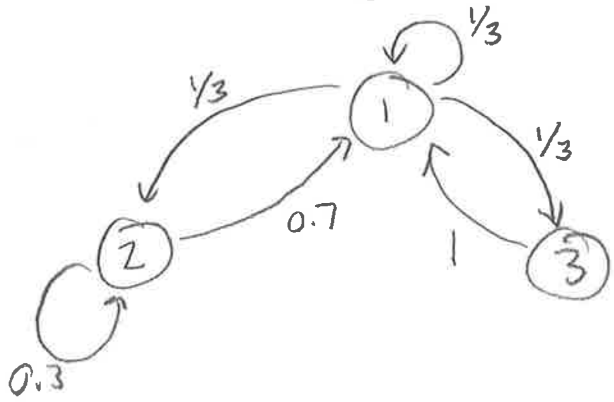
We can summarize these probabilities with a transition matrix,

$$\underline{P} = [p_{ij}] = \begin{bmatrix} p_{11} & \cdots & p_{1m} \\ \vdots & & \vdots \\ p_{m1} & & p_{mm} \end{bmatrix} \qquad (*)$$

Here,

$$\underline{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0.7 & 0.3 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

We can also visualize the transition probabilities with a state transition diagram:



Question:
What is $\Pr\{X=j\}$ for large n?

Notice, $X_0 = 1 \Rightarrow \Pr\{X_1 = 1\} = \frac{1}{3}$

$$\Pr\{X_2 = 1\} = p_{11}\Pr\{X_1 = 1\} + p_{21}\Pr\{X_1 = 2\} + p_{31}\Pr\{X_1 = 3\}$$
$$= \tfrac{1}{3}\left(\tfrac{1}{3}\right) + 0.7\left(\tfrac{1}{3}\right) + 1\left(\tfrac{1}{3}\right)$$
$$\approx 0.678$$

In general,

$$Pr\{X_{n+1} = j\} = \sum_i p_{ij} Pr\{X_n = i\}$$

If we define $\pi_n(i) = Pr\{X_n = i)$, then we can write this as

$$\pi_{n+1}(j) = \sum_i p_{ij} \pi_n(i)$$

If we let $\vec{\pi}_n = (\pi_n(1), \pi_n(2), \pi_n(3), \ldots)$ and let $\underline{P}$ be the transition matrix (*), we can also write this as

$$\pi_{n+1} = \pi_n \underline{P}$$

If we iterate this map, as $n \to \infty$, we say that the limiting distribution (if it exists) is called the <u>steady-state distribution</u>,

$$\vec{\pi}_n \to \vec{\pi}$$

<u>Note:</u>

We can calculate the steady-state distribution $\vec{x}$ the state vector $\vec{x}_n$ as $n \to \infty$ by letting

$$\vec{x}_n \to \vec{x} \text{ and } \vec{x}_{n+1} \to \vec{x}$$

so

$$\vec{x} = \vec{x}\underline{P} \quad \to \quad \vec{x}(I - P) = \vec{0}$$

$$\to \sum_{i=1}^{m} \pi_i = 1$$

Return to example:

$$\underline{P} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 7/10 & 3/10 & 0 \\ 1 & 0 & 0 \end{bmatrix} \qquad \pi_0 = ?$$

↑

we had $\pi_0 = (1,0,0)$

$\rightarrow \quad \pi_1 = (1,0,0)\underline{P}$

$$= \left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right)$$

Recall:

$$(x_1, x_2)\begin{bmatrix} a & b \\ c & d \end{bmatrix} = (x_1 a + x_2 c,\ x_1 b + x_2 d)$$

Note: $\sum_j \pi_{n+1}(j) = \sum_i \sum_j p_{ij}\pi_n(i) = \sum_i \pi_n(i) \underbrace{\sum_j p_{ij}}_{=1} = 1$

$\rightarrow$ π_{n+1} is again a probability distribution.

Question: what would $\pi_0 = (0, \tfrac{1}{2}, \tfrac{1}{2})$ mean?

Definition:

$\{\pi_n\}$ approaches a steady-state distribution π if $\pi_n \rightarrow \pi$ for $n \rightarrow \infty$. If it exists, then π is the solution of

$$\pi = \pi \underline{P}, \quad \sum_i \pi(i) = 1, \quad \pi(i) \geq 0$$

Example: $\pi_0 = (0, \frac{1}{2}, \frac{1}{2})$, $P = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 7/10 & 3/10 & 0 \\ 1 & 0 & 0 \end{bmatrix}$

$\rightarrow \pi_1 = (17/20, 3/20, 0)$

$\pi_2 \approx (0.38, 0.32, 0.28)$

$\pi_3 \approx (0.64, 0.22, 0.12)$

$\vdots$

$\pi_{11} \approx (0.553, 0.262, 0.183)$

$\vdots$

$\pi \approx (0.553, 0.263, 0.184)$

Question: In what instances might such a limit not exist?

1) It could be a periodic Markov chain

Ex: $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $\pi_0 = (1, 0)$

$\rightarrow \pi_1 = (0, 1)$, $\pi_2 = (1, 0)$, ... $\Rightarrow$ state i is periodic with period $\delta = 2$

2) It could be an aperiodic Markov chain

↳ every state has period $\delta = 1$

Question: If $\{\pi_n\}$ tends to a limit, is that limit always independent of π_0? $\rightarrow$ No: $P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. $\pi_0 = (1, 0) = \pi_1 = \pi_2 = \cdots$

$\pi_0 = (0, 1) = \pi_1 = \pi_2 = \cdots$

Definition:

A Markov chain is called an ergodic chain if it is possible to go from every state to every state (not necessarily in one move), and it is aperiodic.

Monte Carlo Simulation

Probability models with no time dynamics can be solved analytically, and steady-state results are possible for simple stochastic models, but the majority of probability models need to be solved numerically, especially if interested in transient behavior (before steady state).

Example:

The weather service forecasts 50% chance of rain every day this week. What are the chances of 3 consecutive rainy days?

Variables:

$$X_t = \begin{cases} 0, & \text{no rain on day } t \\ 1, & \text{rain on day } t \end{cases}$$

Assumptions:

$X_n,\ n=1,\dots,7$ are iid random variables

$Pr\{X_t = 0\} = Pr\{X_t = 1\} = \frac{1}{2}$

Objective:

Determine the probability that $\underbrace{X_t = X_{t+1} = X_{t+2} = 1 \text{ for some } t = 1,2,3,4,5}$

$:= Y \rightarrow$ want $E[Y]$

We will use Monte Carlo simulation. In a Monte Carlo simulation, we will randomly assign values to each random variable according to its probability distribution.

Subsequent repetitions of this produce different results, since the r.v. is assigned randomly, so we typically perform a Monte Carlo simulation a number of times to determine an average or expected outcome. These repeated simulations are considered independent trials.

Suppose we want to simulate a random variable Y. The steps in a Monte Carlo simulation are:

1) Draw a random number from a given distribution using a pseudorandom number generator. In Matlab some helpful commands are:

rand	: uniformly distributed r.v. $X \in [0,1]$
randi	: discrete unif. dist. r.v. $X \in [1, i_{max}]$
randn	: r.v. drawn from the standard normal distribution
exprnd	: r.v. from the exponential distribution

2) Assign output Y_n based on a deterministic rule

3) Repeat; combine repeated trials

If we do 'enough' trials to get Y_i, $i=1,\dots,n$, we know

$$\frac{Y_1 + \cdots + Y_n}{n} \to E[Y] \quad \text{as} \quad n \to \infty$$

by the (strong) law of large numbers. Furthermore, if we define

$$S_n = Y_1 + \cdots + Y_n$$

the Central Limit Theorem implies that

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \quad , \; \mu = E[Y] \, , \; \sigma^2 = V[Y]$$

is approximately distributed according to the standard normal distribution for large n.

↖ for practical purposes, $n \geq 10$ is often 'good enough'

Note: we don't know μ or σ! But the difference between the observed average $\frac{S_n}{n}$ and the true mean of Y, $\mu = E[Y]$, is

$$\frac{S_n}{n} - \sigma = \frac{\sigma}{\sqrt{n}} \left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \right) \sim \underbrace{\frac{1}{\sqrt{n}}}_{\text{tends to zero as fast as } \frac{1}{\sqrt{n}}}$$

Back to example:

Monte Carlo simulation:

1) Draw a uniformly distributed r.v. from $[0,1]$: r

↖ not given an indication any other dist. is better

2) If $r < p = 0.5$, let it be a rainy day

$r > p$, let there be no rain

3) Keep track of consecutive rainy days. If 3 occur in one week, consider it a 'rainy' week, y

3) Repeat!

See Matlab code

Back to Markov chains:

Example: Pet store (Example 8.1)

A pet store orders 3 20-gal aquariums at the end of a week if all aquariums have been sold. If 1 aquarium remains in stock, no new ones are ordered. What is the probability of not having an aquarium in stock when someone wants to buy one?

Variables: S_n = supply of aquariums at the beginning of week n

D_n = demand for aquariums during week n

Assumptions: $D_{n-1} < S_{n-1} \Rightarrow S_n = S_{n-1} - D_{n-1}$

$D_{n-1} \geq S_{n-1} \Rightarrow S_n = 3$

$\Pr\{D_n = k\} = \frac{e^{-1}}{k!}$ ← potential buyers arrive at random at a rate of $1/\text{wk} = \lambda$

Why Poisson:

We've been working with waiting times in a Poisson process, which have no memory - these r.v.'s are exponential. If we consider the probability of events occuring during a particular time interval, which have memory, the distribution we should consider is Poisson,

$$\Pr\{D_t = k\} = \frac{e^{-\lambda t}(\lambda t)^k}{k!},$$

for some interval of length t, rate λ, and number of events in the interval k.

↑ average # of events per interval

so with $\lambda = 1/\text{wk}$ $\longrightarrow$ $Pr\{D_n = k\} = \dfrac{e^{-1}}{k!}$

$t = 1 \text{ wk}$

Note: For X_t, the time it takes for another arrival after time t and N_t, # arrivals during time period t

$Pr\{X_t \le x\} = 1 - Pr\{X_t > x\}$

$$\Rightarrow Pr\{X_t \le x\} = 1 - Pr\{N_{t+x} - N_t = 0\} \quad \leftarrow (X_t > x) \equiv (N_t = N_{t+x})$$

$$= 1 - Pr\{N_x = 0\}$$

$$= 1 - \frac{(\lambda x)^0}{0!} e^{-\lambda x}$$

Poisson probability mass function

(density for discrete rv's)

$$= 1 - e^{-\lambda x}$$

Goal: Calculate $Pr\{D_n > S_n\}$

Model:

$\{S_n\}$ is a Markov chain with $S_n = \{1, 2, 3\}$ and

$$\underline{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

To determine the transition probabilities, we want

$$Pr\{S_{n+1} = 1 \mid S_n = 1\}$$

$$Pr\{S_{n+1} = 2 \mid S_n = 1\}$$

$$\vdots$$

$$Pr\{S_{n+1} = 3 \mid S_n = 3\}$$

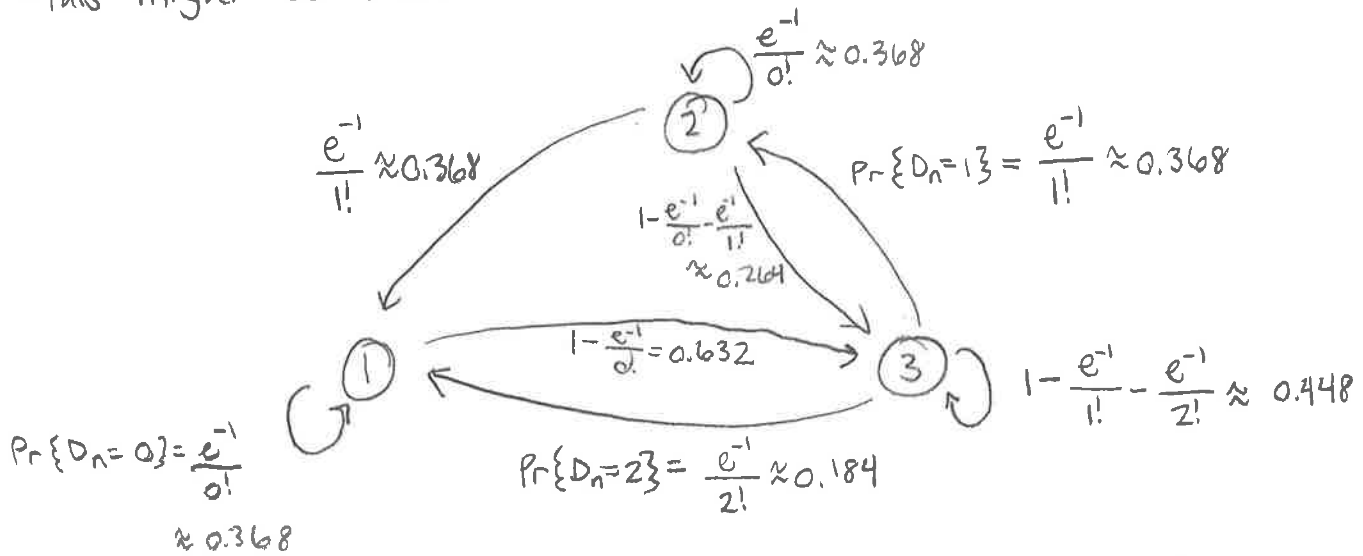
For example, if $S_n = 3$, then

$$Pr\{S_{n+1}=1\} = Pr\{D_n=2\}$$

$$Pr\{S_{n+1}=2\} = Pr\{D_n=1\}$$

$$Pr\{S_{n+1}=3\} = 1 - Pr\{D_n=1\} - Pr\{D_n=2\}$$

This might be easier to summarize in a state transition diagram:


$\frac{e^{-1}}{0!} \approx 0.368$
2
$\frac{e^{-1}}{1!} \approx 0.368$
$Pr\{D_n=1\} = \frac{e^{-1}}{1!} \approx 0.368$
$1-\frac{e^{-1}}{0!}-\frac{e^{-1}}{1!} \approx 0.264$
1
$1-\frac{e^{-1}}{0!} = 0.632$
3
$1-\frac{e^{-1}}{1!}-\frac{e^{-1}}{2!} \approx 0.448$
$Pr\{D_n=0\} = \frac{e^{-1}}{0!} \approx 0.368$
$Pr\{D_n=2\} = \frac{e^{-1}}{2!} \approx 0.184$

In summary,

$$P \approx \begin{bmatrix} 0.368 & 0 & 0.632 \\ 0.368 & 0.368 & 0.264 \\ 0.184 & 0.368 & 0.448 \end{bmatrix} \qquad (*)$$

We want $Pr\{D_n > S_n\}$. To find this, we can look at the steady-state distribution of $\{S_n\}$. Some observations:

1) The system is <u>irreducible</u>

↳ we can reach 2 from 1 by $1 \to 3 \to 2$

2) and <u>aperiodic</u>

↳ every state has $P_{ii} > 0$

So this Markov chain is ergodic.

Theorem:

Every ergodic Markov chain tends to a steady state, and the steady-state distribution is independent of the initial state of the system.

Steady state: $\vec{\pi} = \vec{\pi} \underline{P}$, where $\underline{P}$ is (*).

$$\hookrightarrow \pi_1 = 0.368\pi_1 + 0.368\pi_2 + 0.184\pi_3 \quad (1)$$

$$\pi_2 = 0.368\pi_2 + 0.368\pi_3 \quad (2)$$

$$\pi_3 = 0.632\pi_1 + 0.264\pi_2 + 0.448\pi_3 \quad (3)$$

Need to solve this with the condition

$$\pi_1 + \pi_2 + \pi_3 = 1 \quad (4)$$

This system has 3 unknowns and 4 equations

$\hookrightarrow$ the solution to (2), (3), (4) is

$$\vec{\pi} = (\pi_1, \pi_2, \pi_3) \approx (\underbrace{0.285}_{Pr\{S_n=1\}}, \underbrace{0.263}_{Pr\{S_n=2\}}, \underbrace{0.452}_{Pr\{S_n=3\}})$$

So

$$Pr\{D_n > S_n\} = \sum_{i=1}^{3} \underbrace{Pr\{D_n > S_n \mid S_n = i\}}_{= Pr\{D_n > i\}} Pr\{S_n = i\}$$

$$= \left(1 - \frac{e^{-1}}{0!} - \frac{e^{-1}}{1!}\right)\pi_1 + \left(1 - \frac{e^{-1}}{0!} - \frac{e^{-1}}{1!} - \frac{e^{-1}}{2!}\right)\pi_2$$

$$+ \left(1 - \frac{e^{-1}}{0!} - \frac{e^{-1}}{1!} - \frac{e^{-1}}{2!} - \frac{e^{-1}}{3!}\right)\pi_3$$

$$\approx 0.264(0.285) + 0.080(0.263) + 0.019(0.452)$$

$$\approx 0.105$$ $\Rightarrow$ Demand exceeds supply ~10% of the time

Markov Processes

Definition:

A Markov Process is a continuous-time stochastic system consisting of $\{X_t\}_{t\geq 0}$, with

$$X_t \in \{1, 2, 3, \ldots, m\}$$

obeying the Markov property:

$$Pr\{X_{t+s}=j \mid X_u : u \leq t\} = Pr\{X_{t+s}=j \mid X_t\}$$

i.e., the future solely depends on the current state.

Note: The distribution of T_i, the time spent in state i, is memoryless - the time to the next transition does not depend on the time the process has spent in the current state:

$$Pr\{T_i > t+s \mid T_i > s\} = Pr\{T_i > t\}$$

and T_i is exponential

the exponential distribution is the only continuous-time distribution with the memoryless property

Thus, we can describe all transitions in a Markov process by

1) Fixing λ_i (rate) of the exponential distribution describing time spent in i

2) Fixing a state transition probability matrix $P=[p_{ij}]$

Example:

$$\underline{P} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 7/10 & 3/10 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \lambda_1 = 1, \lambda_2 = \tfrac{1}{3}, \lambda_3 = 1/2$$

We already know $\vec{\pi} = \vec{\pi}\underline{P}$, $\Sigma \pi_i = 1 \Rightarrow \vec{\pi} \approx (0.553, 0.263, 0.184)$

$\hookrightarrow \pi$ describes how often we transition into a state

The proportion of time spent in state i is

$$P_i = \frac{\pi_i/\lambda_i}{(\pi_1/\lambda_1) + \cdots + (\pi_m/\lambda_m)}, \quad \vec{\pi} = (\pi_1, \ldots, \pi_m), \ \lambda_1, \ldots, \lambda_m$$

$$\Rightarrow P_1 = \frac{0.553(1)}{1.71} \approx 0.32$$

$P_2 \approx 0.46$

$P_3 \approx 0.22$

This $\{P_i\}$ is called the steady-state distribution for the Markov process (contrasting with the steady-state distribution for the embedded Markov chain).

Summary:

A Markov process is a Markov chain where the time between transitions has an exponential distribution.

Connection to dynamical systems:

Given $X_t = i$, let T_i be the time until the next jump, where

$T_i = \min\{T_{i1}, \dots, T_{im}\}$, T_{ij} is exponential with parameter $a_{ij} = \lambda_i p_{ij}$

The parameter $a_{ij} = \lambda_i p_{ij}$ denotes the rate at which the process goes from state i to state j

Note: $a_{ii} = -\lambda_i$ is the rate at which the process tends to leave state i.

Theorem:

The probability functions $P_i(t) = Pr\{X_t = i\}$ in a Markov process, with transition-rate matrix $A = [a_{ij}]$, $a_{ij} = \lambda_i p_{ij}$, $a_{ii} = -\lambda_i$, must satify the differential equations

$$\frac{d\vec{P}}{dt} = A\vec{P} \quad , \quad \vec{P} = (P_1, P_2, \dots, P_m)$$

i.e.,

$$\dot{P}_1 = a_{11} P_1 + \dots + a_{m1} P_m$$

$$\dot{P}_2 = a_{21} P_1 + \dots$$

$$\vdots$$

$$\dot{P}_m = a_{m1} P_1 + \dots + a_{mm} P_m$$

The steady-state distribution for the Markov process corresponds to the steady-state solution to <u>this system</u>!!!

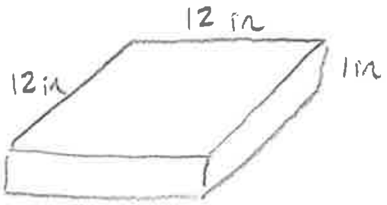
Linear Regression

Example: Predict the number of board-feet produced from a felled Ponderosa pine tree, based on the diameter of the tree at its base.

Data given in Matlab file.

Note: Board-feet is a unit of volume of cut timber

$$1 \text{ board-foot} = 144 \text{ in}^3$$



We want to create a model that best fits our data. As an initial step, let's assume there are no dynamics to the model. What type of functional relationship might there be between diameter and board-feet?

Variables: d = diameter

V = volume, in board-feet

Assumptions: 1) Ponderosa pines are shaped like right circular cylinders

2) The height of a tree is proportional to its diameter

Model: There are several options:

1) Assuming geometric similarity

$$V \propto d^3 \to V = ad^3 + b$$

2) Assuming pines have constant height

$$V \propto d^2 \to V = ad^2 + b$$

(Matlab interlude)

The simplist linear regression model to find a, b, and the quality of these models' fit to the data, is assumed to be a linear function,

$$X_i = ax + b + \varepsilon_i$$

where a and b are real constants and ε_i is a random variable that represents the effect of random fluctuations. Each $\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots$ are typically assumed iid normal with mean zero. Since ε_n has mean zero,

$$E[X_i] = ax + b$$

The overall goodness of fit is given by

$$F(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \qquad (*)$$

This is also called the 'error sum of squares', or SSE. The best-fitting line will occur at a global minimum of $(*)$. Setting the partial derivatives

$$\frac{\partial F}{\partial a} = 0 \quad \text{and} \quad \frac{\partial F}{\partial b} = 0$$

we obtain

$$\left.\begin{aligned} \sum_{i=1}^{n} y_i &= nb + a\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i &= b\sum_{i=1}^{n} x_i + a\sum_{i=1}^{n} x_i^2 \end{aligned}\right\}$$

Note: a & b are switched in book

solving these for a and b yields

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i}{m\sum x_i^2 - (\sum x_i)^2}$$

The 'total corrected sum of squares', SST, is given by

$$SST = \sum_{i=1}^{n} (\underbrace{y_i - \bar{y}})^2 \quad , \quad \bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

$\forall i$, this gives the total variation between any one data point and the mean

Then the R^2 value gives a measure of the fit for the regression line, and is defined by

$$\begin{aligned} R^2 &= 1 - \frac{SSE}{SST} \\ &= 1 - \frac{\sum_{i=1}^{n} (y_i - (ax_i + b))^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \\ &\leq 1 \end{aligned}$$

← variation about regression line

← variation about mean

It expresses the proportion of total variation in the model that can be accounted for by the model given by the line $y=ax+b$, calcuated from our model, when compared with the line $y=\bar{y}$, where $\bar{y}$ is the average of the data y-values.

It may be helpful to visualize goodness of fit graphically. One thing we can do is to plot the <u>residuals</u>, or the errors between the actual and predicted values,

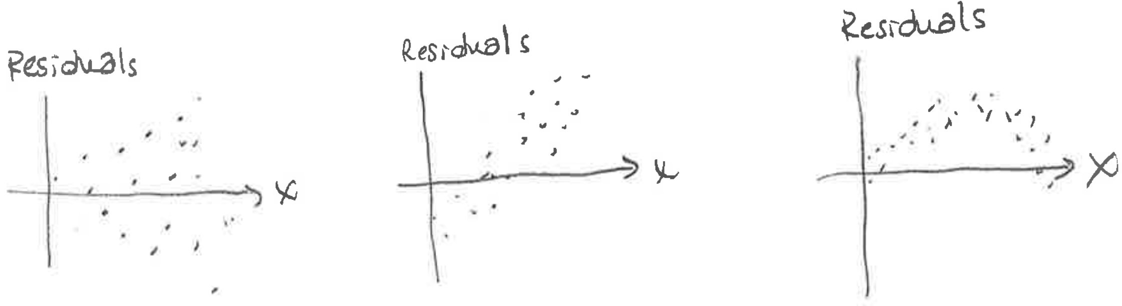
$$r_i = \underbrace{y_i}_{\text{actual}} - \underbrace{f(x_i)}_{\text{predicted}} = \underbrace{y_i} - \underline{(ax_i - b)}$$

If:

1) <u>The residuals are randomly distributed</u> and are in a reasonably small band commensurate with the accuracy of the data, then your model appears to be adequate to explain the variation in the data.

2) <u>There is a pattern or trend</u> in the residuals, then a predictable effect remains to be modeled.

Sometimes the pattern or trend in a residual indicates ways to improve the model:



(a) Fanning (b) Linear (c) Curvilinear

Other methods of model fitting

Visual fitting with non-'linear' functions:

If the data appears nonpolynomial (e.g. exponential), it's common practice to transform the data so that it looks linear, and compare that with your nonlinear model

Example:

Suppose we have the data for some sub-model:

x	1	2	3	4
y	8.1	22.1	60.1	165

this looks exponential, so we want to see if we can fit this to $y = Ce^x$. Since

$$\ln y = \ln(Ce^x) = \ln C + x,$$

it looks like we should take the natural log of our y-values, then fit to the line

$$\tilde{y} = x + a \quad , \qquad a = \ln C$$

From Matlab, we get a fit of

$$\tilde{y} = x + 1.1 \quad \to \quad y = (e^{1.1})e^x \approx 3e^x$$

Methods from time series analysis:
(ref. Section 8.4 in Meerschaert)

A time series is a stochastic process with r.v.'s X_i which vary over time, typically with fixed time steps.

With time series, we can create a time series model directly, or we can use time series ideas to compare other model output with data. In either case, we can compare two time series using various indicators.

Covariance:

A typical assumption is that time series a r.v. that evolves over time involving a trend in addition to a stationary time series, which is a time series that has mean and variance that do not change over time. (Many of these time series definitions have technical versions involving distribution properties - this is just the general idea). For two stationary time series, $(X_1)_t$, $(X_2)_t$ we can measure how dependent their r.v.'s are using covariance:

$$Cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

$$\mu_1 = E[X_1]$$

$$\mu_2 = E[X_2]$$

If X_1, X_2 are independent, then $Cov(X_1, X_2) = 0$

Note: $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

Correlation:

The correlation is a dimensionless version of the covariance,

$$\rho = \mathrm{corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1 \sigma_2}$$

$$= E\left[\frac{(X_1 - \mu_1)(X_2 - \mu_2)}{\sigma_1 \sigma_2}\right]$$

If $\rho > 0$, X_1, X_2 are positively correlated

$\rho < 0$, X_1, X_2 are negatively correlated

Correlation is useful measure of dependence. Autocorrelation, or

$$\rho(t, h) = \mathrm{corr}(X_t, X_{t+h})$$

is a measure of depencence from one time point to another.