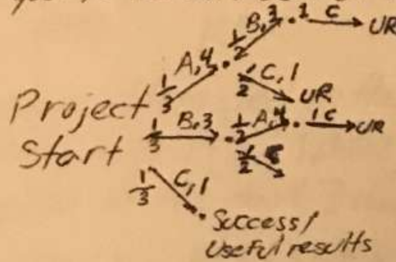
Homework 1

1. You are a data scientist and are choosing between 3 approaches A, B, and C to a problem. With approach A you will spend a total of four days coding and running an algorithm that will not produce useful results. With approach B you will spend 3 days and your algorithm will not produce useful results. With approach C you will spend 1 day and get useful results. You are starting your project and do not know which approach will work. You are equally likely to choose among options. If your selected approach does not work you will select a new approach. What is the expected time in days for you to obtain useful results? What is the variance on this time?



The probability of route $A \to B \to C$ $P(A \to B \to C) = P(A)P(B|A)P(C|A \to B)$
$= \frac{1}{3}\frac{1}{2}1 = \frac{1}{6}$. $A \to B \to C$ takes $(4 \text{ days}) + (3 \text{ days}) + (1 \text{ day}) = t_A + t_B + t_C = 8 \text{ days} = t_{A \to B \to C}$

$P(A \to C) = \frac{1}{3}\frac{1}{2} = \frac{1}{6} = P(A)P(C|A)$ | $t_{A \to B \to C}^2 = 64\, d^2$

$t_{A \to C} = t_A + t_C = (4d) + (1d) = 5d$ | $t_{A \to C}^2 = 25\, d^2$

$P(B \to A \to C) = P(B)P(A|B)P(C|B \to A) = \frac{1}{3}\frac{1}{2}1 = \frac{1}{6}$

$t_{B \to A \to C} = t_B + t_A + t_C = (3d) + (4d) + (1d) = 8d$ | $t_{B \to A \to C} = 64\, d^2$

$P(B \to C) = P(B)P(B|C) = \frac{1}{3}\frac{1}{2} = \frac{1}{6}$

$t_{B \to C} = t_B + t_C = (3d) + (1d) = 4d$ | $t_{B \to C}^2 = 16\, d^2$

$P(C) = \frac{1}{3}$

$t_C = 1d$ | $t_C^2 = 1\, d^2$

$$E[t] = \sum_{x_i}[P(x_i)x_i] = P(A \to B \to C)t_{A \to B \to C} + P(A \to C)t_{A \to C} + P(B \to A \to C)t_{B \to A \to C}$$
$$+ P(B \to C)t_{B \to C} + P(C)t_C = (\tfrac{1}{6})(8d) + (\tfrac{1}{6})(5d) + (\tfrac{1}{6})(8d) + (\tfrac{1}{6})(4d) + (\tfrac{1}{3})(1d)$$
$$= \frac{8}{6}d + \frac{5}{6}d + \frac{8}{6}d + \frac{4}{6}d + \frac{2}{6}d = \frac{27}{6}d = 4.5d$$

$$V[t] = \sum_{x_i}[[t_i - E[t_i]]^2 P(x_i)] = E[(t - E[t])^2] = (t_{A \to B \to C} - E[t])^2 P(A \to B \to C)$$
$$+ (t_{A \to C} - E[t])^2 P(A \to C) + (t_{B \to A \to C} - E[t])^2 P(B \to A \to C) + (t_{B \to C} - E[t])^2 P(B \to C)$$
$$+ (t_C - E[t])^2 P(C) = (8d - 4.5d)^2 \tfrac{1}{6} + (5d - 4.5d)^2 \tfrac{1}{6} + (8d - 4.5d)^2 \tfrac{1}{6} + (4d - 4.5d)^2 \tfrac{1}{6}$$
$$+ (1d - 4.5d)^2 \tfrac{1}{3} = (3.5d)^2 \tfrac{1}{6} + (0.5d)^2 \tfrac{1}{6} + (3.5d)^2 \tfrac{1}{6} + (-0.5d)^2 \tfrac{1}{6} + (-3.5d)^2 \tfrac{1}{3}$$
$$= 8.25\, d^2$$

$$E[t^2] = \sum[P(i)t_i^2] = 8.25\, d^2$$

2. Suppose that whether it is sunny or not in Charlottesville depends on the weather of the last 3 days. Show how this can be modeled as a Markov chain by displaying a diagram and transition matrix.

We have 2 possible weather categories: sunny and not sunny/rainy.

Considering permutations of weather for the last 3 days, each day can have 1 of 2 possible weathers. The total number of permutations of weathers is $2^3 = 8$.

$$_nP_r = \frac{n!}{(n-r)!}$$

SSS	SRR
SSR	RSR
SRS	RRS
RSS	RRR

We are interested in the probability that it is sunny today given the weather of the past 3 days $P(S \mid WWW)$, or more generally the probability of weather W given the weather of the past 3 days $P(W \mid WWW)$.

SSS/S: $P(S/SSS)$

SSS/R: $P(R/SSS)$

WWW/W: $P(W/WWW)$

$P(R/SSS) = 1 - P(S/SSS)$

$P(S/SSS)$	$P(S/SRR)$
$P(S/SSR)$	$P(S/RSR)$
$P(S/SRS)$	$P(S/RRS)$
$P(S/RSS)$	$P(S/RRR)$

According to "Predicting the Weather with Markov Chains", "the probability of it being sunny or rainy tomorrow depends on whether it is sunny or rainy today."
In our case, the probability of it being sunny depends on the weather of the past 3 days.
Alternately, the probability of it being sunny depends on the weather yesterday.

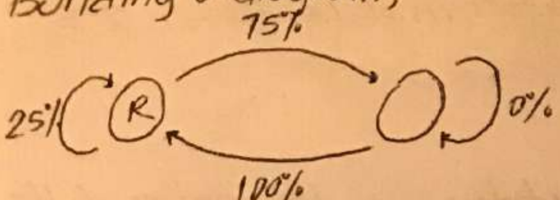
For the second interpretation, we could collect data [R, S, R, S, R, R, S].
We calculate the percentage of instances it's sunny on days directly following rainy days: 3/4. We calculate the percentage of instances it's rainy on days directly following sunny days: 2/2.

For the first interpretation, we could collect a data set of triplets of weather (e.g., SRS). We could calculate the percentage of instances of one ~~state~~ triplet being directly after another triplet, though it would be hard to model "sunny" happening after "SRS".

Building a transition matrix,

~~Today~~ Yesterday \ Today	S	R
S	0	1
R	0.75	0.25

Building a diagram,



For the second interpretation, probabilities for weather today depend on the weather yesterday, not further back. We look up transitional probabilities for today's weather given yesterday's in our transition matrix.
The past 3 days were [R, S, R]. The probability of sun today is 75%. The probability of rain today is 25%.
The past 3 days were [X, R, R]. The probability of sun today is 75%. The probability of rain today is 25%.
The past 3 days were [S, R, S]. The probability of sun today is 0%. The probability of rain today is 100%.

2 (cont.)

https://setosa.io/ev/markov-chains

A Football player can pass (P) or run (R).

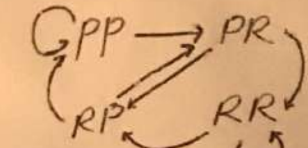
The four possible sequences of two plays in a row are PP, PR, RP, RR.

Each pair of plays consists of Play 1 and Play 2.

consider the next pair of plays after a previous pair of plays. The second play of the previous ~~play~~ pair is the first play of the next pair.

$$\_\,\_\,\_\,\overset{1\,2}{\underset{1\,2}{P\,R\,\_}}$$

Suppose we have a pair of plays PP. If the next play is a run, we have transitioned from PP to PR. If the next play is a pass, we have transitioned from PP to PP.



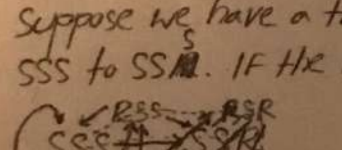
Markov Diagram For Markov Process

Suppose we have a pair of plays PR. If the next play is R, we have transitioned from PR to RR. If the next play is a pass, we have transitioned from PR to RP.
Suppose we have a pair of plays RP. If the next play is R, we transition from RP to PR. If the next play is P, we transition from RP to PP.
Suppose we have a pair of plays RR. If the next play is R, we transition from RR to RR. If the next play is P, we transition from RR to RP.
All other transitions are impossible.
Probabilities x_{ij} sum to 1 across horizontals.

	PP	PR	RP	RR
PP	x_{00}	x_{01}	0	0
PR	0	0	x_{12}	x_{13}
RP	x_{20}	x_{21}	0	0
RR	0	0	x_{32}	x_{33}

Transition Matrix

Suppose we have a triplet of weathers SSS. If the next weather is S, we transition from SSS to SSS. If the next weather is R we transition from SSS to SSR.



Markov Diagram For Markov Process

Suppose we have a triplet of weathers SSR. If the next weather is S, we transition from SSR to SRS. If the next weather is R, we transition from SSR to SRR.
Suppose we have a triplet of weathers SRS. If the next weather is S, we transition from SRS to RSS. If the next weather is R, we transition from SRS to RSR.
Suppose we have a ~~transition~~ triplet of weathers SRR. If the next weather is S, we transition from SRR to RRS. If the next weather is R, we transition from SRR to RRR.

RSS: RSS → SSS, RSS → SSR   RSR: RSR → SRS, RSR → SRR

RRS: RRS → RSS, RRS → RSR   RRR: RRR → RRS, RRR → RRR

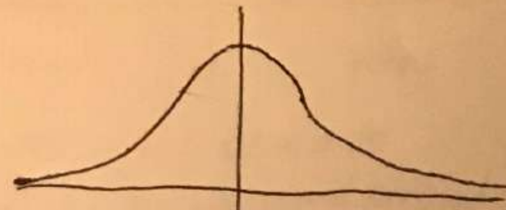
Transition Matrix For Triplets of Weathers

	SSS	SSR	SRS	SRR	RSS	RSR	RRS	RRR
SSS	$p_{SSS \to SSS}$	$p_{SSS \to SSR}$						
SSR			$p_{SSR \to SRS}$	$p_{SSR \to SRR}$				
SRS					$p_{SRS \to RSS}$	$p_{SRS \to RSR}$		
SRR							$p_{SRR \to RRS}$	$p_{SRR \to RRR}$
RSS	$p_{RSS \to SSS}$	$p_{RSS \to SSR}$						
RSR			$p_{RSR \to SRS}$	$p_{RSR \to SRR}$				
RRS					$p_{RRS \to RSS}$	$p_{RRS \to RSR}$		
RRR							$p_{RRR \to RRS}$	$p_{RRR \to RRR}$

Each cell in the matrix tells you the probability of transitioning from the cell's row's state to the cell's column's state.

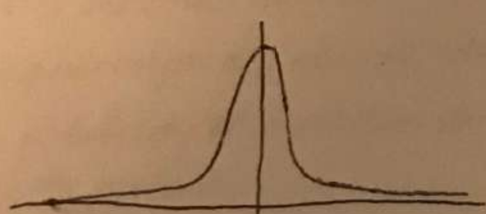
3. Assume a Gaussian distribution for observations X_i, $i=1,\ldots,N$ with unknown mean M and known variance $\underset{\text{likelihood}}{V_X}=5$. Suppose the prior for M is Gaussian with variance $\underset{\text{prior}}{V_M}=10$. The prior for M is a prior probability distribution. How large a random sample must be taken (i.e., what is the minimum value for N) to specify a ~~credible~~ confidence interval having unit length l such that the probability that M lies in this interval $P(M \in CI)$ is 0.95? Our data is normally distributed. We have N data. Each datum could be the height of a flower. The variance of our data $V_X=5$.

If I don't have much data, I have a pretty broad uncertainty, and a prob dist like



If I go fishing one day and catch 1 fish, I don't know that every day I go fishing I'll catch 1 fish. Maybe 50% of the time I catch 1 fish. Maybe 80% of the time I catch 0 fish. Maybe on other days I'll catch 8 fish each day. If I have 5 data points, I may begin to construct a probability distribution for the number of fish caught on a day.

If I have more data, I have less uncertainty, and a prob dist like



How do I get the spread of my data so that 95% of my data is in the confidence interval $[M-\underset{ME=}{0.5}, M+\underset{ME=}{0.5}]$? The width of this confidence interval is $1=2ME$.

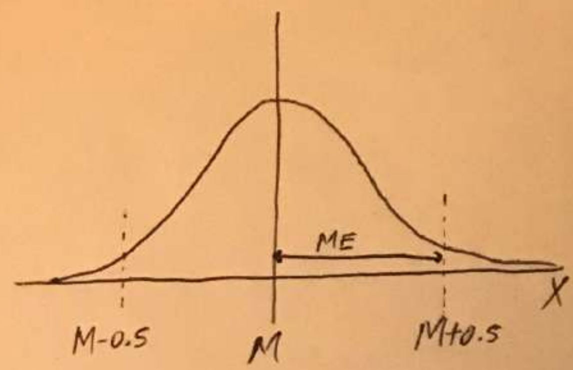
We had a prior for coin flips with a mean of 0.5 and a standard deviation of 0.03 that described what we believed before taking any data about coin flips. Specifically, we decided that a good prior probability distribution for the value of $p=P(\text{Heads})$ is normal with mean 0.5 and standard deviation 0.03.

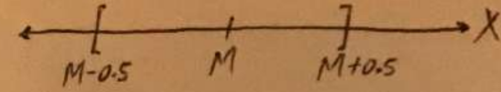
Here, we have a prior probability distribution for the mean M that is normal with variance $V_M=10$.

As you increase your data, you decrease the spread of your posterior probability distribution. You can be more confident in your "final values". How big does N have to be so that our estimate on M, a normal distribution, has a small enough standard deviation so that M lies within a confidence interval of length 1?

Posterior Probality Distribution With Mean M

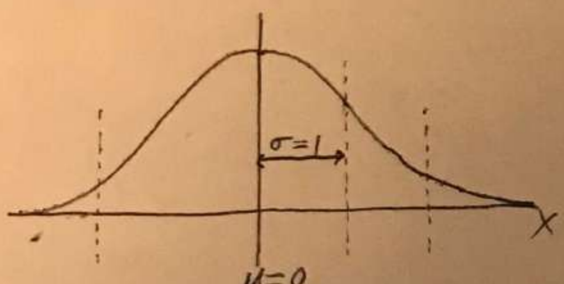
Visualization of Confidence Interval With Confidence $CL = 0.95$ And Length 1





$CL = 0.95, \alpha = 1 - CL = 0.05$

We consider the standard normal distribution.



The critical z score for which the probability under the standard normal probability density function/distribution is $CL + \frac{\alpha}{2} = 0.95 + \frac{0.05}{2} = 0.975$ is given by

$z_c = qnorm(p=0.9, mean=0, sd=1) = 1.960$

The probability of a random variable X being within $z_c =$ 1.96 standard deviations σ of mean M is 95 percent.

What is the minimum value of N such that

$$1.96\sigma = z_c \sigma_{post} \leq 0.5 = \text{Margin Of Error ME}$$

$$\sigma_{post} \leq \frac{0.5}{z_c} = \frac{ME}{z_c} = \text{Standard Error SE} = \frac{0.5}{1.960} = 0.255$$

$$\sigma_{post} = \sqrt{\frac{\sigma_{prior}^2 \sigma_{likelihood}^2}{\sigma_{likelihood}^2 + N\sigma_{prior}^2}} = \sqrt{\frac{V_{prior} V_{likelihood}}{V_{likelihood} + N V_{prior}}} \leq SE$$

$$\frac{V_{prior} V_{likelihood}}{V_{likelihood} + N V_{prior}} \leq (SE)^2$$

$$V_{likelihood} + N V_{prior} \geq \frac{V_{prior} V_{likelihood}}{(SE)^2}$$

$$N V_{prior} \geq \frac{V_{prior} V_{likelihood}}{(SE)^2} - V_{likelihood} = \frac{V_{prior} V_{likelihood} - V_{likelihood}(SE)^2}{(SE)^2} = \frac{V_{likelihood}(V_{prior} - (SE)^2)}{(SE)^2}$$

$$N \geq \frac{V_{likelihood}(V_{prior} - (SE)^2)}{(SE)^2 V_{prior}} = \frac{5(10 - 0.255^2)}{0.255^2 \cdot 10} = 76.394 = 77$$