Homework 11

Brown University

DATA 1010

Fall 2020

Problem 1

A weather station in Providence classifies each day's weather as "good", "fair", or "poor" according to meteorological data. The following table shows the probabilistic relationship between weather on the current day and the probability of the weather expected on next day conditioned on the type of the current day.

current\next	good	fair	poor
good	0.60	0.30	0.10
fair	0.50	0.25	0.25
poor	0.20	0.40	0.40

(a) Determine the probability that the weather will be "poor" exactly 3 days after a "good" weather day.

(b) Over a long period of time, what percentage of days can we expect to have "good" weather?

SOL

(a) This is equal to the first row, third column entry of P3, where P is the transition matrix:

In [1]:
```
P = [0.6 0.3 0.1
     0.5 0.25 0.25
     0.2 0.4 0.4]
(P^3)[1,3]
```

Out[1]: 0.19675

(b) The stationary distribution of P is the 1-eigenvector of P′:

In [8]:
```
using LinearAlgebra
eigen(P')
```

Out[8]:
```
Eigen{Float64,Float64,Array{Float64,2},Array{Float64,1}}
values:
3-element Array{Float64,1}:
 -0.0886000936329383
  0.33860009363293836
  0.9999999999999993
vectors:
3×3 Array{Float64,2}:
  0.49704   -0.772796  -0.795866
 -0.809513   0.158165  -0.500258
  0.312473   0.61463   -0.341085
```

The first entry of the eigenvector corresponding to the eigen value 1, 79.5%, means the percentage of the "good" weather

Problem 2

Consider the state space $X = \{0, 1\}^n$ of binary strings having length n. Define $p(y, x) = 1/n$ if y differs from x in exactly one bit, and $p(y, x) = 0$ otherwise.

Suppose we desire an equilibrium distribution π for which $\pi(x)$ is proportional to the number of ones that occur in vector $\mathbf{x}$. For example, in the long run, a random walk should visit a string having five 1's five times as often as it visits a string having only a single 1.

(a) Provide a general formula for the acceptance ratio $\alpha(x, y)$ that would be used if we were to obtain the desired equilibrium distribution using the Metropolis-Hastings algorithm.

(b) Implement this algorithm for $n = 100$. Test it out to see how many Markov chain steps appear to be necessary before we get approximately the desired statinoary distribution.

SOL

(a) According to Metropolis-Hastings, the acceptance ratio is 1 if the number of 1's increases between the current and proposed state, and (k−1)/k if the number of 1's goes down by 1, where k is the number of 1's at current state.

(b):

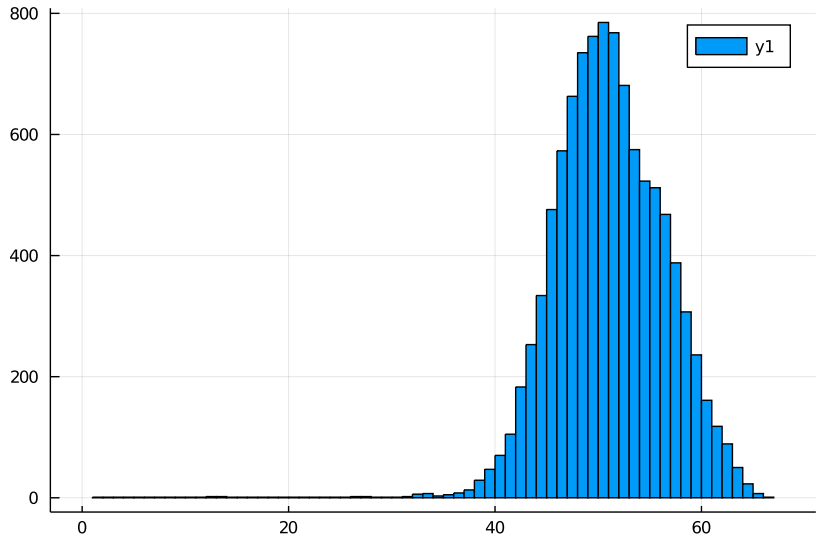
In [12]:
```
using Random
function simulate(cur)
    idx = rand(1:100)
    if cur[idx] == 0
        cur[idx] = 1
    else
        if rand() < (sum(cur)-1)/sum(cur)
            cur[idx] = 0
        end
    end
    return cur
end
```

Out[12]:
```
simulate (generic function with 1 method)
```

In [16]:
```
next = zeros(100)
n_ones = sum(next)
results = []
for i = 1:10000
    next = simulate(next)
    push!(results, sum(next))
end
five = count(k->(k==5), results)
one = count(k->(k==1), results)
println(five, " ", one)
histogram(results, bins=100)
```

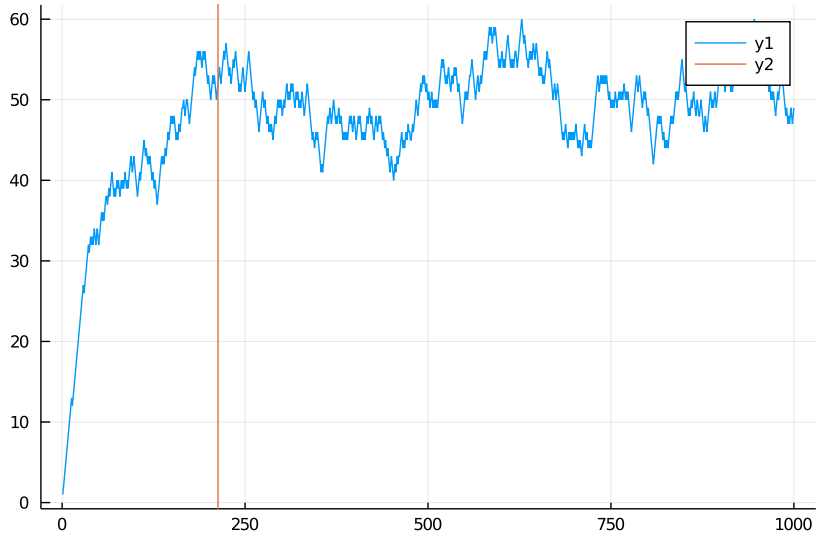
```
1 1
```

Out[16]:



In [25]:
```
using Plots
xs = [i for i = 1:1000]
ys = [results[i] for i = 1:1000]
plot(xs, ys)
vline!([213])
```

Out[25]:



SOL b

The above code sample from the sample space by fliping a bit in the current string since only 1-bit difference is allowed in the proposed state. The final stationary distribution is reached as a normal distribution with mean at 50.

Problem 3

Consider a dataset $\{(x_i, y_i), i = 1, 2, \ldots, 1000\}$ generated by the following procedure:

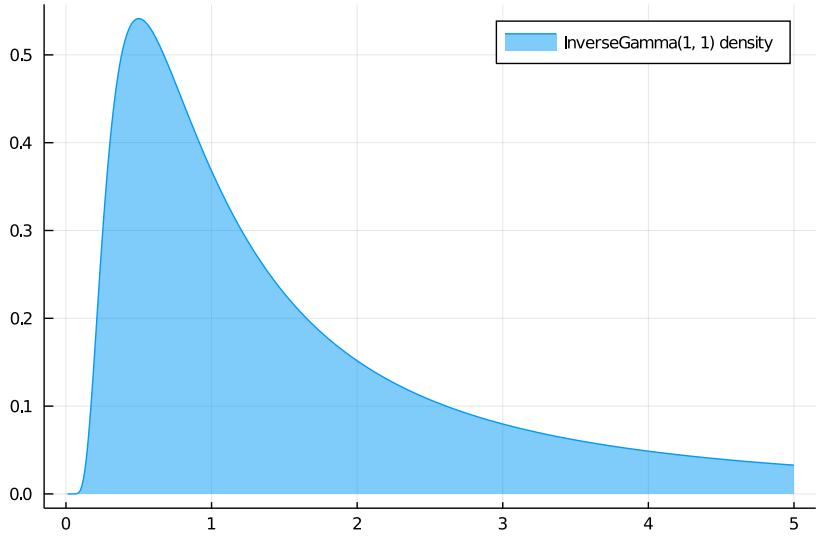
We choose an intercept value b from $\mathcal{N}(0, 1)$ a slope value m independently from $\mathcal{N}(0, 1)$, a σ value from $\mathrm{InverseGamma}(1, 1)$ (see below). Then we generate 1000 independent random pairs (X, Y) by sampling each X uniformly from $[0, 1]$ and setting

$$Y = mX + b + \epsilon,$$

where ϵ is a mean-zero Gaussian with standard deviation ϵ.

In [1]:
```
using Plots, Distributions
plot(0:0.01:5, x -> pdf(InverseGamma(1, 1), x),
     fillrange = 0, fillopacity = 0.5, label = "InverseGamma(1, 1) density")
```

Out[1]:



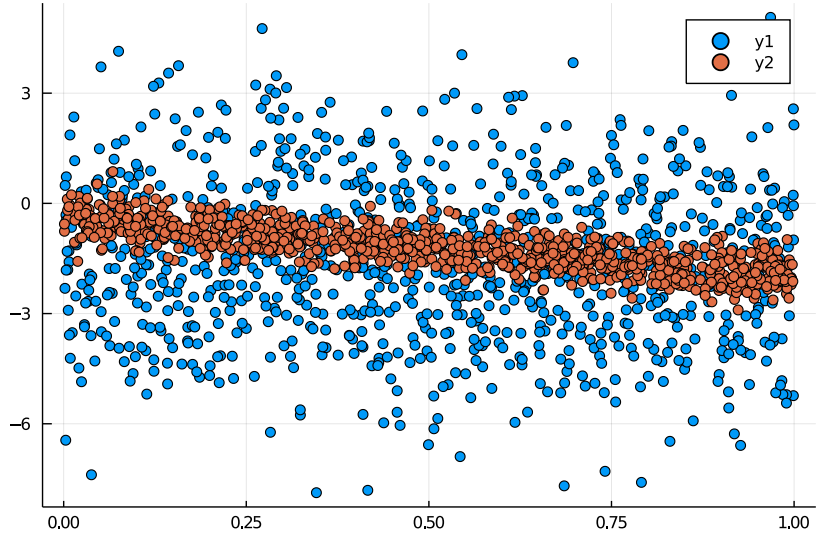
(a) The values of the observations (the one thousand (x, y) pairs) induce a conditional distribution on the model parameters (m, b, and σ).

You can recognize some of this with your eye. Create scatter plots of two samples from this probability measure (on the parameters plus the data points), and see if you can recognize purely from the plots which of the two has most of its conditional probability mass in a region with a larger m value, and which has most of its mass in a region with a larger σ value.

In [12]:
```
using Random
m1 = rand(Normal(0,1))
m2 = rand(Normal(0,1))
b1 = rand(Normal(0,1))
b2 = rand(Normal(0,1))
sig1 = rand(InverseGamma(1,1))
sig2 = rand(InverseGamma(1,1))
ys = []
xs = []
for _ in 1:1000
    x = rand(Uniform(0,1))
    e = rand(Normal(0, sig1))
    push!(xs, x)
    push!(ys, m1*x+b1+e)
end
println(m1, "  ", sig1)
println(m2, "  ", sig2)
scatter(xs, ys)
ys = []
xs = []
for _ in 1:1000
    x = rand(Uniform(0,1))
    e = rand(Normal(0, sig2))
    push!(xs, x)
    push!(ys, m2*x+b2+e)
end
scatter!(xs, ys)
```

```
-0.3958639910532646  2.204294002675396
-1.635129967660837  0.36398932881824975
```

Out[12]:



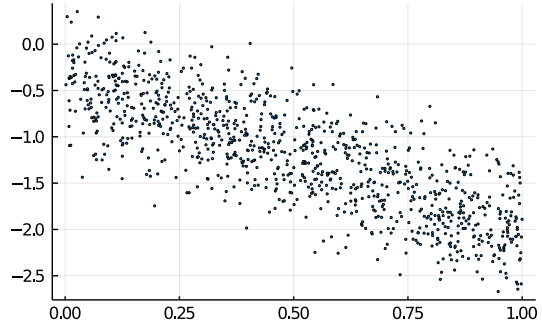
SOL a

From the plot above, we can infer that the blue points has a larger sigma value as the standard deviateion is larger than that of red points. However, it is hard to recognize which sample has a larger m value since the noise is big and two m values are close to each other.

(b) Fix the data (the (x, y) pairs) for particular simulation run from part (a). Use Markov Chain Monte Carlo to sample from the posterior distribution of m and σ to get a sense of just how concentrated the probability mass is around the mean values. Are m and σ correlated (with respect to the posterior distribution)?

In [183]:
```
x = rand(Uniform(0, 1), 1000)
y_mean = m2*x .+ b2
y = y_mean + rand(Normal(0, sig2), 1000)
function observations()
    scatter(x, y, ms = 1, msw = 0.2,
                    size = (400, 250), legend = false)
end
observations()
```

Out[183]:



In [184]:
```
N = Normal(0, 10)
δ(x,y,m,b) = sum((yᵢ - m*xᵢ - b)^2/2 for (xᵢ, yᵢ) in zip(x,y))
function α(x, y, m, b, m_prop, b_prop)
    min(1.0, exp(-δ(x,y,m_prop,b_prop) + δ(x,y,m,b)) *
        pdf(N, m_prop)/pdf(N, m) * pdf(N, b_prop)/pdf(N, b))
end
```

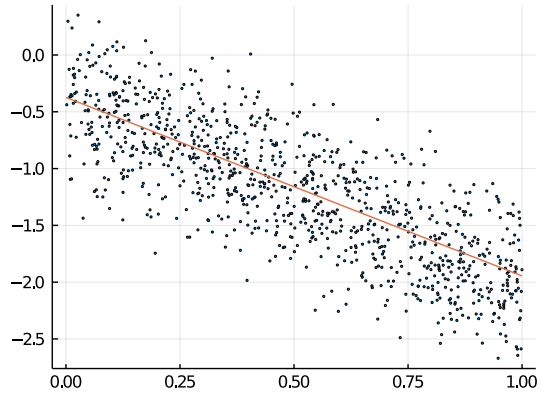
Out[184]:
```
α (generic function with 1 method)
```

In [185]:
```
function mcmc(n_iterations)
    m, b, σ = 0.0, 0.0, 1.0
    θs = [(m, b)]
    for i in 1:n_iterations
        m_prop, b_prop = m + rand(Normal(0,0.005)), b + rand(Normal(0,0.005))
        if rand() < α(x, y, m, b, m_prop, b_prop)
            m, b = m_prop, b_prop
            push!(θs, (m, b))
        end
    end
    θs
end
```

Out[185]:
```
mcmc (generic function with 1 method)
```

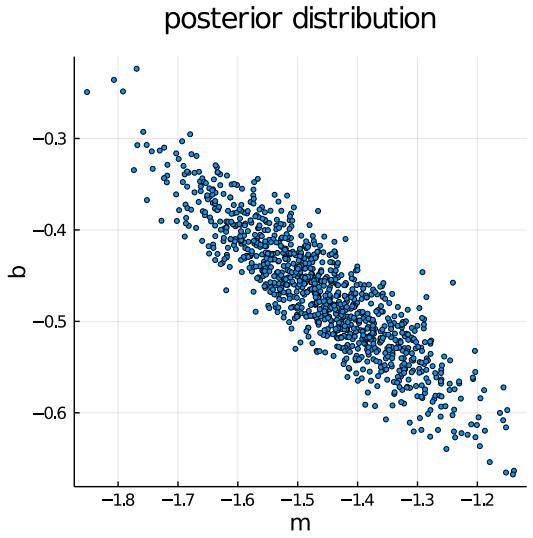
In [194]:
```
θs = mcmc(3_000)
m, b = θs[end]
observations()
plot!(0:1, x-> m*x + b, size = (400, 300))
```

Out[194]:



In [181]:
```
scatter([mcmc(3000)[end] for _ in 1:1000], size = (400, 400),
        ms = 2, msw = 0.2, title = "posterior distribution",
        xlabel = "m", ylabel = "b", legend = false)
```

Out[181]:



SOL b

There is a negative correlation between m and b

Problem 4

Consider the following statement from an advertisement.

'In this town we observe that those who take soy isoflavone supplements have improved cognitive performance. Thus, for anyone looking to improve their cognitive performance, they should consider taking soy isoflavone supplementation.'

(a) Identify the problematic causal inference in this statement.

(b) For each of the following statements, indicate whether it weaks or strengthens the assertion in the ad (or neither).

```
i) The alleged cause and effect are both effects of some common cause. (Note: give a specific example one some such common cause.)

ii) The cause and effect are flipped; the alleged cause is the effect and vice versa.

iii) The presence of the cause in a relatively large population does not coincide with the detection of the effect.

iv) The presence of the assoication has been identified in a larger population.

v) A possible confounder for this effect has been considered and ruled out.
```

SOL

(a) The variable is not controlled and there is no comparison in this case. The people who do not take the soy isoflavone supplements might also have improved cognitive performance. And the soy isoflavone supplements might not be the reason, but something that is used afterwards by the group of people who experience in the improvement in mental health. Therefore, taking the soy isoflavone does not direcly bring the effect of improving the cognitive performance.

(b):

i) Weaks. People who need much brain work and experience pressure. They need oy isoflavone supplements to help relax. The improvement in cognitive performance comes from the daily training of their brain and pressure in their work.

ii) Weaks. People who experience cognitive performance might need to absorb more soy isoflavone to support their mental health. They might just recover from disease.

iii) Weaks. If there is a casuality between taking soy isoflavone supplements and improvement of cognitive performace, we should expect it always brings the improvement of cognitive supplements in this population.

iv) No impact. Whether the population is large or not, it does not specify the attributes and qualities of the population. The casual inference should be indicated with constraints and comparisions. A large population who do not use this drug might also experience cognitive performance improvement.

v) Strengthens. Because there is a limited number of causes that bring the improvement in the cognitive performance, ruling out one of the causes in this pool will increase the possibility that soy isoflavone is the actual reason behind this phenomenon.

Problem 5

Consider the following classic example given by Edward Simpson in 1951.

700 patients were given access to a drug the effectiveness of which we would like to study. A total of 350 patients chose to take the drug and 350 didn't. The result is given below:

Categories	Drug	No Drug
Men	81 out of 87 (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

a) Not controlling gender, describe the overall trend of taking the drug.

b) Conditioning on each sex, describe the overall trend of taking the drug.

c) Is there a discrepancy? If yes, what is the cause of this discrepancy?

SOL

(a) Since 273 is smaller than 289, taking the drug is negatively correlated with recovering from the disease.

(b) Since 0.93 > 0.87 for men and 0.73 > 0.69 for women, taking the drug is positively correlated with recovering from the disease.

(c) Even though the men who take the drug has a higher recover rate the that of women who do not take the drug, its base number is small. But the women who take the drug has a lower recover rate than that of the men who do not take the drug. This base is large and overflip the effect of the previous one. Therfore, controlling for local variablesshows a overall trend that is opposite in the "men vs women" categories.

In [ ]: