PHYS 509C Assignment 1

Callum McCracken, 20334298

September 22, 2022

Code for this assignment is here:
`https://github.com/callum-mccracken/PHYS-509C-A2`

1 Three easy applications of Bayes's theorem:

A. Suppose supernovae follow a poisson distribution with an unknown rate R. Calculate and plot the posterior distribution for R, given an observation of 4 supernovae in 10 centuries, using (a) a prior uniform in R and (b) a prior uniform in $\log_{10}(R)$.

Remember Bayes's theorem:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{probability of data}}$$

The priors:

- Uniform in R, but we're not given bounds... Suppose it goes from some n to m.
$$P(R) = \frac{1}{m-n}.$$
- Uniform in $\log_{10}(R)$: use $P(x)dx = P(f(x))df(x)$ to get
$$\begin{aligned} P(R) &= P(\log_{10}(R))\frac{d\log_{10}(R)}{dR} \\ &= \frac{\ln(10)}{\ln(m)-\ln(n)}\frac{1}{\ln(10)R} \\ &= \frac{1}{\ln(m)-\ln(n)}\frac{1}{R} \end{aligned}$$

Likelihood: Poisson distribution ($T = 1000$ years)

$$P(k|R) = \frac{e^{-RT}(RT)^k}{k!}$$

Probability of Data:

$$P(k) = \int P(R)P(k|R)dR$$

for uniform R,

$$
\begin{aligned}
P(k) &= \int_0^\infty \frac{1}{m-n}\frac{e^{-RT}(RT)^k}{k!}dR \\
&= \frac{1}{m-n}\frac{1}{k!}\int_0^\infty e^{-RT}(RT)^k dR \\
&= \frac{1}{m-n}\frac{1}{k!}\frac{1}{T}\int_0^\infty e^{-\lambda}(\lambda)^k d\lambda \\
&= \frac{1}{m-n}\frac{1}{k!}\frac{1}{T}k! \\
&= \frac{1}{m-n}\frac{1}{T}
\end{aligned}
$$

or for uniform $\log_{10}(R)$

$$
\begin{aligned}
P(k) &= \int_0^\infty \frac{1}{\ln(m)-\ln(n)}\frac{1}{R}\frac{e^{-RT}(RT)^k}{k!}dR \\
&= \frac{1}{\ln(m)-\ln(n)}\int_0^\infty T\frac{1}{RT}\frac{e^{-RT}(RT)^k}{k!}dR \\
&= \frac{1}{\ln(m)-\ln(n)}\int_0^\infty \frac{e^{-\lambda}(\lambda)^{k-1}}{k!}d\lambda && [\lambda = RT] \\
&= \frac{1}{\ln(m)-\ln(n)}\frac{1}{k}
\end{aligned}
$$

Now calculate posteriors:

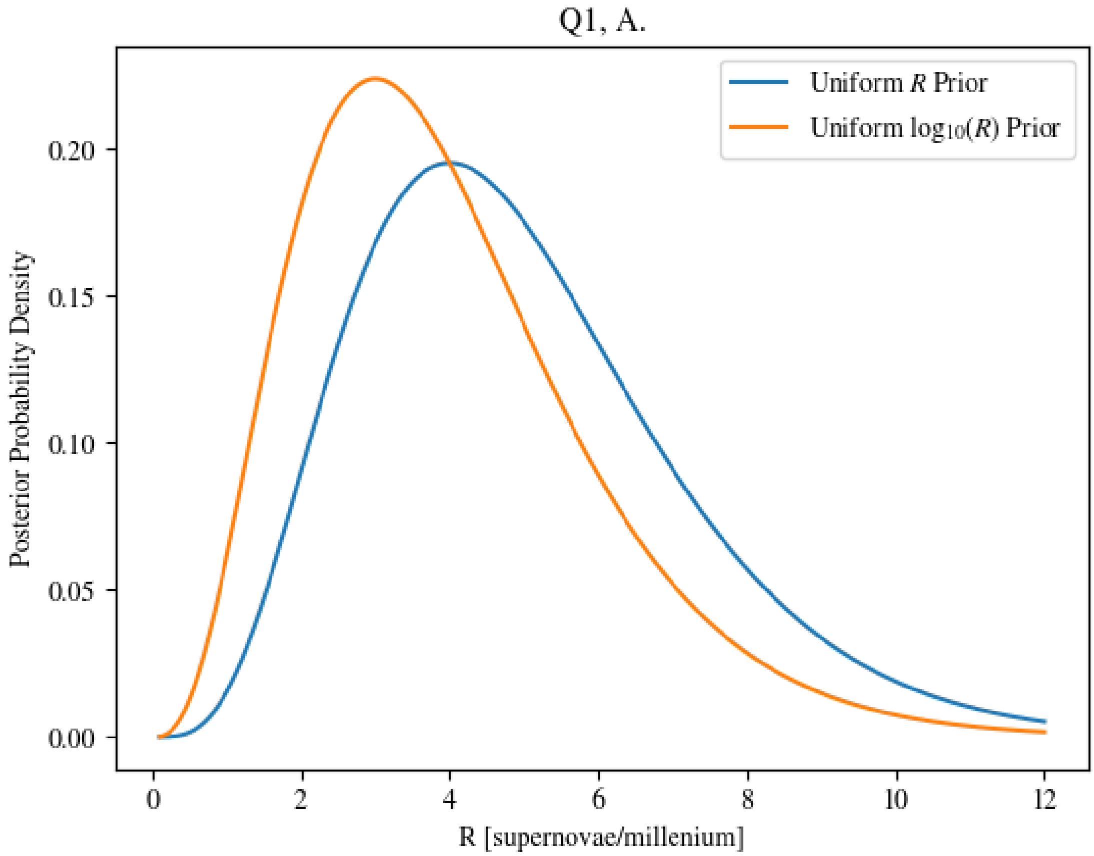
Uniform R:

$$
\begin{aligned}
P(R|k) &= \frac{P(k|R)P(R)}{P(k)} \\
&= \frac{\frac{e^{-RT}(RT)^k}{k!}\frac{1}{m-n}}{\frac{1}{m-n}\frac{1}{T}} \\
&= \frac{\frac{e^{-RT}(RT)^k}{k!}}{\frac{1}{T}} \\
&= \frac{Te^{-RT}(RT)^k}{k!}
\end{aligned}
$$

Uniform $\log_{10}(R)$

$$
\begin{aligned}
P(R|k) &= \frac{P(k|R)P(R)}{P(k)} \\
&= \frac{\frac{e^{-RT}(RT)^k}{k!}\frac{1}{\ln(m)-\ln(n)}\frac{1}{R}}{\frac{1}{\ln(m)-\ln(n)}\frac{1}{k}} \\
&= \frac{Te^{-RT}(RT)^{k-1}}{(k-1)!}
\end{aligned}
$$

Plot of these distributions:



B. Measurements are drawn from a uniform distribution spanning the interval (0, m). The probability of getting a measurement outside of this range is zero. The endpoint m is not well-known, but a prior experiment yields a Gaussian prior of m = 3 +/- 1. You take three measurements, getting values of 2.5, 3.1, and 2.9. Use Bayes' theorem to calculate and plot the new probability distribution for m.

Likelihood: if our model is a uniform distribution on $[0, m]$ and we have independent measurements (which I think we can assume for this question, right?), then $P(D|m) = \frac{1}{m^3}$ if $m \geq 3.1$, $= 0$ otherwise. That's the product of three uniform probabilities.

Prior (Gaussian, $\mu = 3, \sigma = 1$): $P(m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(m-3)^2}{2}}$.

Probability of Data:

$$P(D) = \int_0^\infty P(m)P(D|m)dm \tag{1}$$

$$= \int_{3.1}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{(m-3)^2}{2}} \frac{1}{m^3} dm \tag{2}$$

$$\tag{3}$$

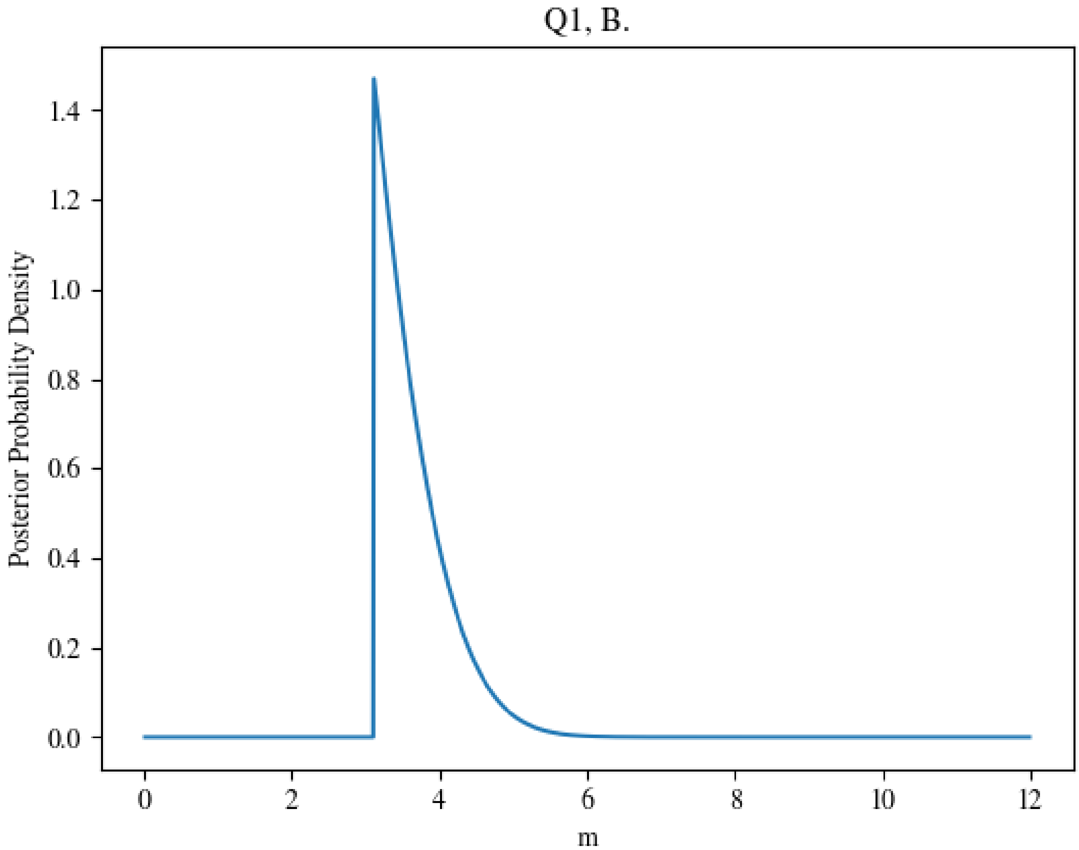
Posterior:

$$P(m|D) = \frac{P(D|m)P(m)}{P(D)}$$

$$= \frac{\frac{1}{m^3}\frac{1}{\sqrt{2\pi}}e^{-\frac{(m-3)^2}{2}}}{\int_{3.1}^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{(m'-3)^2}{2}}\frac{1}{m'^3}dm'} \qquad [m \geq 3.1,\ \text{else } 0]$$

$$= \frac{\frac{1}{m^3}e^{-\frac{(m-3)^2}{2}}}{\int_{3.1}^\infty e^{-\frac{(m'-3)^2}{2}}\frac{1}{m'^3}dm'} \qquad [m \geq 3.1,\ \text{else } 0]$$

Plot:



C. Consider a person considering whether or not to launch a rocket with a possible malfunctioning component. In the control centre there is a warning light that is not completely reliable. During launch the warning light doesn't go on. From a costs standpoint, should she abort the mission or not? Compute and compare the expected cost of launching to the expected cost of aborting, given that the light didn't go on.

- $P(\text{light on}|\text{malfunction}) = 1/2$, $P(\text{light on}|\text{no malfunction}) = 1/3$
- $C(\text{no launch, no malfunction}) = 2M$, $C(\text{launch, malfunction}) = 5M$
- $C(\text{no launch, malfunction}) = C(\text{launch, no malfunction}) = 0$
- $\text{Prior}(\text{malfunction}) = 2/5$

Let F denote component failure, L denote the light being on, and A denote a launch (since L was taken).

Priors: $P(F) = \frac{2}{5}, P(\neg F) = \frac{3}{5}$.
Likelihood of observing data (no light): $P(\neg L|F) = \frac{1}{2}, P(\neg L|\neg F) = \frac{2}{3}$.
Probability of seeing our data (no light):

$$P(\neg L) = P(\neg L|F)P(F) + P(\neg L|\neg F)P(\neg F) = \frac{3}{5}$$

Posteriors:

$$\begin{aligned} P(F|\neg L) &= \frac{P(F)P(\neg L|F)}{P(\neg L)} = \frac{\frac{2}{5}\frac{1}{2}}{\frac{3}{5}} = \frac{1}{3} \\ P(\neg F|\neg L) &= 1 - P(F|\neg L) = \frac{2}{3} \end{aligned}$$

Costs:

$$\begin{aligned} C(A) &= P(F|\neg L)C(F|A) + P(\neg F|\neg L)C(\neg F|A) \\ &= \frac{1}{3}(5M) + \frac{2}{3}(0) \\ &= \frac{5}{3}M \end{aligned}$$

$$\begin{aligned} C(\neg A) &= P(F|\neg L)C(F|\neg A) + P(\neg F|\neg L)C(\neg F|\neg A) \\ &= \frac{1}{3}(0) + \frac{2}{3}(2M) \\ &= \frac{4}{3}M \end{aligned}$$

The expected cost of not launching is less, so from a cost standpoint, the mission should be aborted, and the rocket makers should probably invest in a better warning light for next time.

2 COVID-19 Study

- Sample size: 3330 people in Santa Clara County, California
- $N_{+,t} = 50$ positive test results in test group
- Control group 1: 3324 people, definitely negative
- $N_{+,c1} = 16$ positives in control 1
- Control group 2: 157 people, definitely negative
- $N_{+,c2} = 130$ positives in control 2 (27 false negatives)

Calculate the Bayesian 95% central interval on the fraction of people in Santa Clara County who actually had antibodies for COVID-19, marginalizing over the false positive and false negative rates. Assume flat priors on all parameters. Submit a plot of the posterior distribution for the true incidence rate as well as your code or calculation.

We want to find the distribution of the true positive rate in the population (and then get the 95% interval), $P(\text{real pos})$.

Relevant variables, $R_{TP}, R_{FP}, R_{TN}, R_{FN}, R_P, R_N$ (P=Positive, N=Negative, T=True, F=False).

Use Bayes to get probability distributions for 3 of these (just three since $R_P = 1 - R_N, R_{TP} = 1 - R_{FP}, R_{TN} = 1 - R_{FN}$).

Each one of these 3 follows a binomial distribution, so

Distribution for $P(R_{FP}|$data from control group 1$)$:

Prior: $P(R_{FP}) = 1$

Data: in our control group 1, $N_1 = 3324, N_{FP,1} = 16$.

Likelihood (binomial): $P(N_1 = 3324, N_{FP,1} = 16|R_{FP}) = \binom{N_1}{N_{FP,1}} R_{FP}^{N_{FP,1}}(1 - R_{FP,1})^{N_1 - N_{FP,1}}$

Normalization:

$$P(N_1 = 3324, N_{FP,1} = 16) = \int_0^1 dR_{FP} \binom{N_1}{N_{FP,1}} R_{FP}^{N_{FP,1}}(1 - R_{FP})^{N_1 - N_{FP,1}}$$

Posterior:

$$P(R_{FP}|N_1 = 3324, N_{FP,1} = 16) = \frac{P(R_{FP})P(N_1 = 3324, N_{FP,1} = 16|R_{FP})}{P(N_1 = 3324, N_{FP,1} = 16)}$$
$$= \frac{R_{FP}^{N_{FP,1}}(1 - R_{FP,1})^{N_1 - N_{FP,1}}}{\int_0^1 dR_{FP} R_{FP}^{N_{FP,1}}(1 - R_{FP})^{N_1 - N_{FP,1}}}$$

Similarly, the other distributions:

$$P(R_{FN}|N_2 = 157, N_{FN,2} = 27) = \frac{R_{FN}^{N_{FN,2}}(1 - R_{FN,2})^{N_2 - N_{FN,2}}}{\int_0^1 dR_{FN} R_{FN}^{N_{FN,2}}(1 - R_{FN})^{N_2 - N_{FN,2}}}$$

$$P(R_P|N_t = 3330, N_{P,t} = 50) = \frac{R_P^{N_{P,t}}(1 - R_{P,t})^{N_t - N_{P,t}}}{\int_0^1 dR_P R_P^{N_{P,t}}(1 - R_P)^{N_t - N_{P,t}}}$$

What we really want is the PDF for the probability of having antibodies $P(A)$. Let's find that by relating the things we already have, make it a variable, say a.

A positive test could have come from a true positive or a false positive, so

$$R_P = R_{TP}a + R_{FP}(1 - a)$$
$$\frac{R_P - R_{FP}}{R_{TP} - R_{FP}} = a$$
$$\frac{R_P - R_{FP}}{(1 - R_{FP}) - R_{FP}} = a$$

We have PDFs for all those variables, use transformations to get the prior PDF for a.

How to get that? I can't use convolutions like on A1... Do I do two convolutions for numerator and denominator then something else for the ratio?

Finally marginalize over R_{FP}, R_{FN}, and plot

3 CO_2 Meter

If we assume that the data follows an exponential plateau, approaching some steady state value C, then that means for any t our expected y (concentration in ppm) is

$$y = C - Be^{-At}$$

To include the error as a fit parameter, let's say (since we have no other reason to pick a different distribution) that for every t we have a Gaussian distribution with mean $\mu = C - Be^{-At}$ and standard deviation σ_y.
With that, we can calculate the likelihood:

$$\begin{aligned} L &= \prod P(x_i, y_i, A, B, C, \sigma_y) \\ -\ln(L) &= -\sum P_i \\ -\ln(L) &= -\sum \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{\left(y_i - C - Be^{-At}\right)^2}{2\sigma^2}\right) \end{aligned}$$

We only care about C and σ though, marginalize the others:

$$\begin{aligned} P(x_i, y_i, A, C, \sigma_y) &= \int_B P(B)P(x_i, y_i, A, B, C, \sigma_y)dB \\ &= \int_{-\infty}^{\infty} 1P(x_i, y_i, A, B, C, \sigma_y)dB \\ &= \int_{-\infty}^{\infty} 1P(x_i, y_i, A, B, C, \sigma_y)dB \end{aligned}$$

To minimize this, set all the derivatives with respect to parameters equal to zero.

$$\frac{\partial}{\partial C} -\sum \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{\left(y_i - C - Be^{-At_i}\right)^2}{2\sigma^2}\right)$$
$$= -\sum \frac{1}{\sqrt{2\pi}\sigma_y} \frac{\partial}{\partial C} e^{-\left(y_i - C - Be^{-At_i}\right)^2} e^{-2\sigma^2}$$
$$= \sum \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\left(y_i - C - Be^{-At_i}\right)^2} e^{-2\sigma^2} \frac{\partial}{\partial C} \left(y_i - C - Be^{-At_i}\right)^2$$
$$= \sum \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\left(y_i - C - Be^{-At_i}\right)^2} e^{-2\sigma^2} 2\left(y_i - C - Be^{-At_i}\right)(-1)$$
$$= -\sum \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{\left(y_i - C - Be^{-At_i}\right)^2}{2\sigma^2}\right) 2\left(y_i - C - Be^{-At_i}\right)$$

$$\frac{\partial}{\partial B} -\sum \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{\left(y_i - C - Be^{-At_i}\right)^2}{2\sigma^2}\right)$$
$$= -\sum \frac{1}{\sqrt{2\pi}\sigma_y} \frac{\partial}{\partial C} e^{-\left(y_i - C - Be^{-At_i}\right)^2} e^{-2\sigma^2}$$
$$= \sum \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\left(y_i - C - Be^{-At_i}\right)^2} e^{-2\sigma^2} \frac{\partial}{\partial C} \left(y_i - C - Be^{-At_i}\right)^2$$
$$= \sum \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\left(y_i - C - Be^{-At_i}\right)^2} e^{-2\sigma^2} 2\left(y_i - C - Be^{-At_i}\right)(-1)$$
$$= -\sum \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{\left(y_i - C - Be^{-At_i}\right)^2}{2\sigma^2}\right) 2\left(y_i - C - Be^{-At_i}\right)$$

How much of that uncertainty can be attributed to the time binning (measurements are only re-

4 Retirement investments.

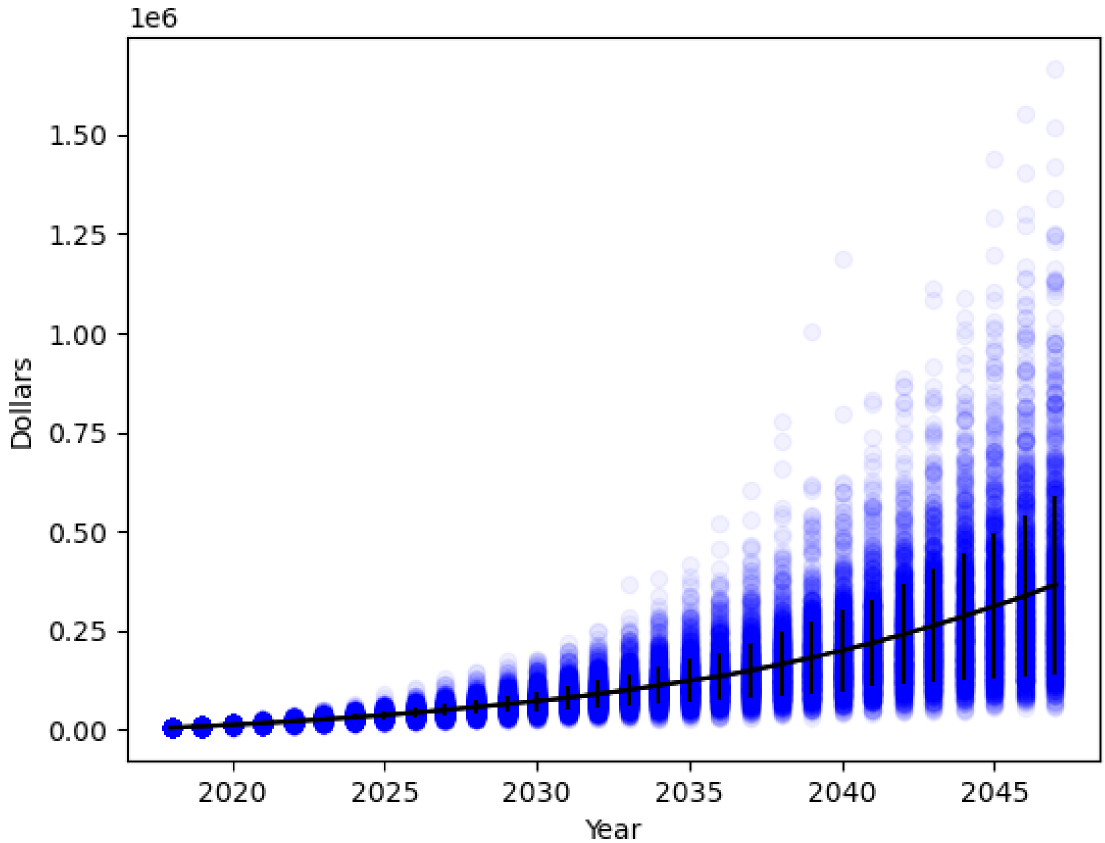
A. The percentage yield on an investment has a Gaussian distribution with mean of 8% and standard deviation (SD) of 15%. (A yield of 8% would mean the amount of money increases by a factor of 1.08 in a year. A yield of -8% would mean multiplying by 0.92 instead.) Suppose that you put \$3000 into a retirement account investing in this item on January 1st of every year, starting in 2018. What is the mean amount of money you will have in the account on Dec 31, 2047? Show a plot of the distribution of the amount of money on that date for 1000 trials of the "experiment". What is the SD? Hand in your code or equivalent documentation.

I did this computationally, here are the results:

Value on Dec 31, 2047: 370000

Standard deviation: 228000

how to deal with dec 31st vs jan 1?



B. Suppose now that the retirement account contains three classes of investments: Canadian stocks, foreign stocks, and bonds. The yields on these three investments each vary randomly but with some correlation. Here is the yield information for each investment: $\mu_C = 8\%, \sigma_C = 15\%, \mu_F = 8\%, \sigma_F = 15\%, \mu_B = 5\%, \sigma_B = 7\%, \rho_{CF} = 0.50, \rho_{CB} = 0.20, \rho_{FB} = 0.05$. On January 1 of each year you put \$1000 into each class of investment. Show the distribution of the total amount of money in your account on Dec 31, 2047. What are the mean and SD?

Let's do this computationally again, but first let's explain some mystery numbers that will appear in the code, using the hint in the question.

$$C = \mathcal{N}(\mu = 0.08, \sigma = 0.15)$$

Then define a linear combination of C and F in terms of another Gaussian:

$$Y = C - F$$

$Y = \mathcal{N}(\mu = 0, \sigma_Y = \sigma_C)$.
Check this works out:

$$\begin{aligned}
\operatorname{Var}(Y) &= 1^2 \operatorname{Var}(C) + 1^2 \operatorname{Var}(F) - 2 \operatorname{Cov}(C, F) \\
\operatorname{Var}(Y) &= \sigma_C^2 + \sigma_F^2 - 2\rho_{CF}\sigma_C\sigma_F \\
\sigma_C^2 &= \sigma_C^2 + \sigma_C^2 - 2\frac{1}{2}\sigma_C^2 \\
\sigma_C^2 &= \sigma_C^2
\end{aligned}$$

$$\mu_Y = \mu_C - \mu_F = 0$$

So to generate F we can use this:

$$F = C - Y, Y = \mathcal{N}(\mu = 0, \sigma = \sigma_C)$$

Then to generate B, consider another combination: $Q = aY + bB$

$$\begin{aligned}
\mu_Q &= a\mu_Y + b\mu_B \\
\mu_Q &= a(0) + b(0.05) \\
\mu_Q &= b(0.05)
\end{aligned}$$

Let $b = 1$ for simplicity.

$$
\begin{aligned}
\operatorname{Var}(Q) &= a^2 \operatorname{Var}(Y) + \operatorname{Var}(B) + 2ab \operatorname{Cov}(Y, B) \\
&= a^2 (0.15)^2 + (0.07)^2 + 2a(\operatorname{Cov}(C, B) - \operatorname{Cov}(F, B)) \\
&= a^2 (0.15)^2 + (0.07)^2 + 2a(\rho_{C,B}\sigma_C\sigma_B - \rho_{F,B}\sigma_F\sigma_B) \\
&= a^2 (0.15)^2 + (0.07)^2 + 2a((0.20)(0.15)(0.07) - (0.05)(0.15)(0.07)) \\
&= a^2 (0.15)^2 + (0.07)^2 + 0.00315a
\end{aligned}
$$

We still have a free parameter, let $a = 1$

$$
\begin{aligned}
\sigma_Q^2 &= (0.15)^2 + 0.00315 + (0.07)^2 \\
\sigma_Q &= 0.174785583
\end{aligned}
$$

(there's some round-off error here but I assume it's small enough we can ignore it)

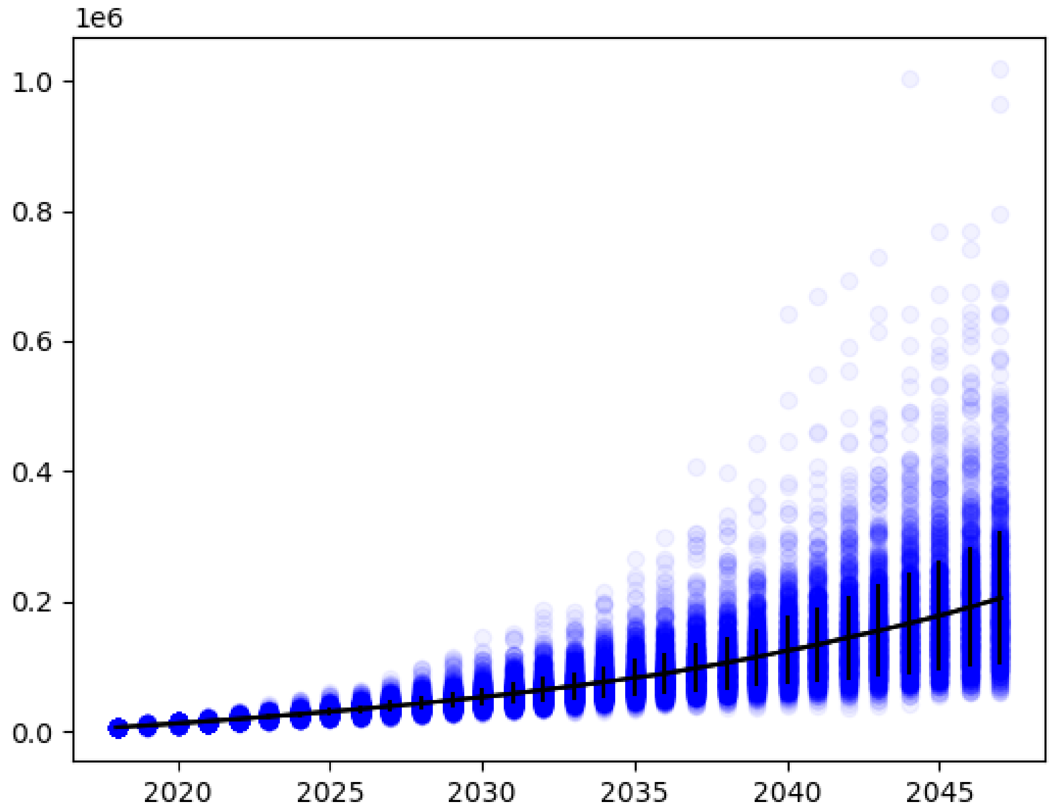
So $Q = Y + B, Q = \mathcal{N}(\mu = 0.05, \sigma = 0.174785583)$.

Results:

Value Dec 31, 2047: 204918.0565428472

Standard deviation: 102948.69611667504

Plot of our results:

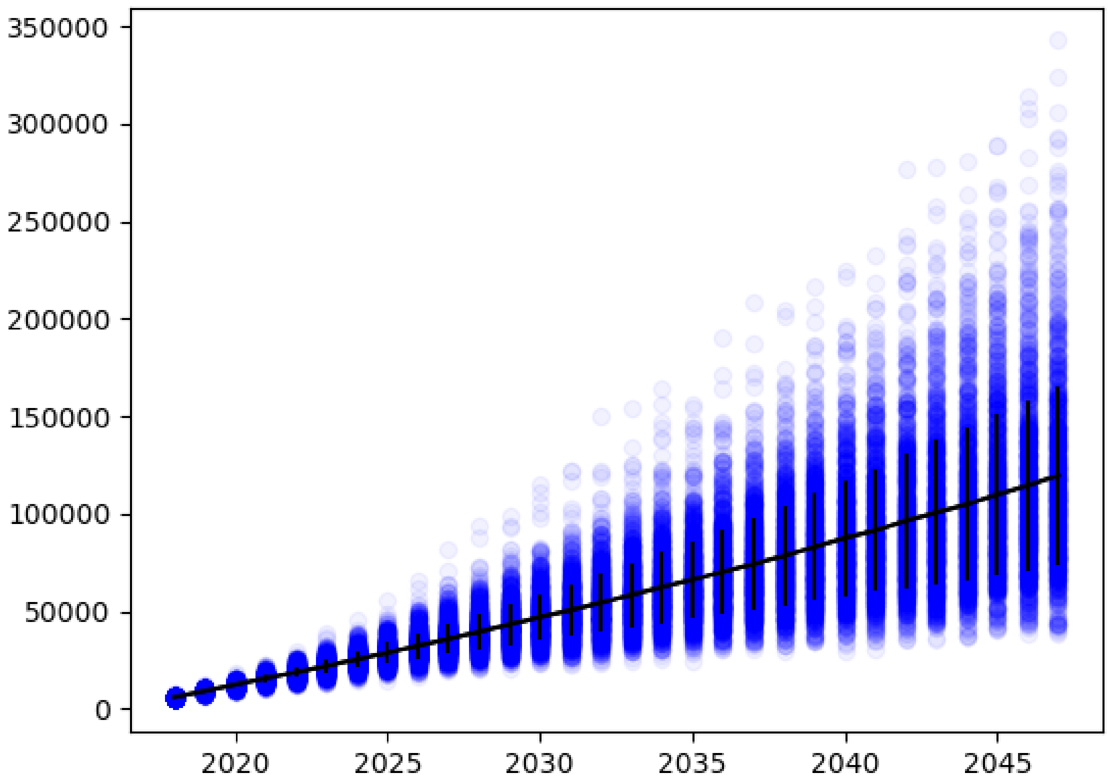


C. Now suppose we add a procedure called "rebalancing". On January 1 of each year we contribute a total of $3000 to the account, but at the same time we redistribute the total amount of money in the account evenly between the three investments. How does this change the total amount on Dec 31, 2047? Show a plot of the distribution, and report the mean and SD as well.

Value Dec 31, 2047: 119628.72535169181

Standard deviation: 45756.874877228125

Plot of results:



This feels kinda sussy, why is everything so much lower? Was there another constraint we could have used? Also this looks linear rather than exponential...