homework 02 solutions

Exercise 1

This is a typical bivariate situation in which the two marginals r.v.s are both Bernoulli.

- X represents the health situation of a randomly chosen person w.r.t. the disease caused by the unknown virus. X can take as values "1" in case the virus is present, "0" otherwise.
- Y represents the outcome of the medical test developed to verify the presence of this virus. The values for Y are "+" in case of positive outcome of the test and "$-$" otherwise.

So, $\mathbb{P}(X = 1) = 0.003$ and therefore $\mathbb{P}(X = 0) = 0.997$.
$\mathbb{P}(Y = +|X = 0) = 0.1$ (false positive rate) and therefore $\mathbb{P}(Y = -|X = 0) = 0.9$.
$\mathbb{P}(Y = -|X = 1) = 0.05$ (false negative rate) and therefore $\mathbb{P}(Y = +|X = 1) = 0.95$.

1. Using the *total probability formula*:

$$\begin{aligned}\mathbb{P}(Y = +) &= \sum_{x\in\{0,1\}} \mathbb{P}(X = x, Y = +) = \\ &= \mathbb{P}(X = 0, Y = +) + \mathbb{P}(X = 1, Y = +) = \\ &= \mathbb{P}(Y = +|X = 0)\,\mathbb{P}(X = 0) + \mathbb{P}(Y = +|X = 1)\,\mathbb{P}(X = 1) = \\ &= \frac{10}{100}\frac{99.7}{100} + \frac{95}{100}\frac{0.3}{100} = 0.10255 = 10.255\%\end{aligned}$$

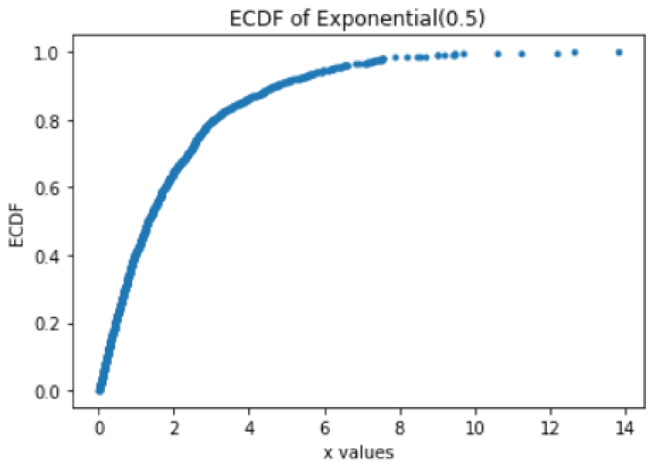
2. Using the *Bayes theorem*:

$$\begin{aligned}\mathbb{P}(X = 1|Y = +) := \frac{\mathbb{P}(X = 1, Y = +)}{\mathbb{P}(Y = +)} &\overset{\text{Bayes th.}}{=} \frac{\mathbb{P}(Y = +|X = 1)\,\mathbb{P}(X = 1)}{\mathbb{P}(Y = +)} = \\ &= \frac{0.95 \cdot 0.003}{0.10255} \simeq 0.0278 = 2.78\%\end{aligned}$$

Exercise 2

```python
# ECDF
def ecdf(dist, x, n_samples):
    # draw my sample
    samples = dist.sample((n_samples,))

    # sort the sample
    samples_sorted, indices = samples.sort()

    # compute the ECDF: for each element, the value of the ECDF is the fraction of samples <= that value
    # hence its index (+1) in the sorted list of samples divided by the number of samples
    ecdf = []
    for i in range(1, n_samples+1):
        ecdf.append(i/n_samples)

    # plot the ECDF
    plt.plot(samples_sorted, ecdf, ".")
    plt.xlabel('x values')
    plt.ylabel('ECDF')
    plt.title('ECDF of Exponential(0.5)')
    plt.show()

    # compute ecdf(x) of the passed x
    for i in range(1, n_samples+1):
        if samples_sorted[i-1]>x:   # checks previous value (so from index 0 to index n_samples)
            print(f"ECDF({x}) = {ecdf[i-1]}")
            return ecdf[i-1]
    print(f"ECDF({x}) = 1") #case in which the passed x is greater than the largest sample
    return 1
```

```python
# set up my distribution object
exp = dist.Exponential(0.5)
# my value
x = 2;
# number of samples
n_samples=1000;

# run the function
ecdf(exp, x, n_samples)
```

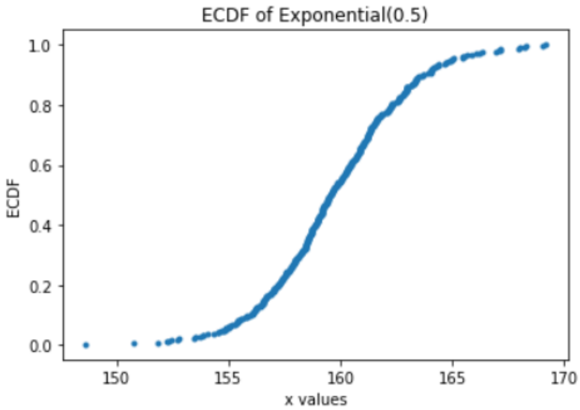



```
ECDF(2) = 0.645
0.645
```

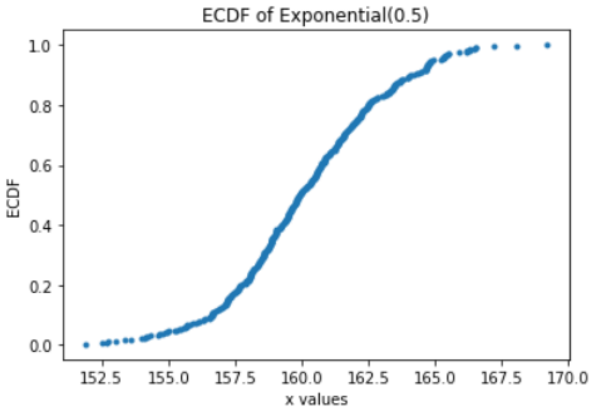
Exercise 3

1.

```
# set up the distribution object
normal = dist.Normal(160, 3)

# compute the area between [155,175] under the pdf curve = integral of the pdf is the cdf!
print(f"Prior probability of mu in [155,175] = {ecdf(normal, 175, 400)-ecdf(normal, 155, 400)}")
```




```
ECDF(175) = 1
```




```
ECDF(155) = 0.045
Prior probability of mu in [155,175] = 0.955
```

2.

The prior distribution is

$$p(\mu) = \frac{1}{(2\pi\sigma_0^2)^{1/2}} \cdot \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

The likelihood is

$$p(x|\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right\}$$

The posterior is proportional to their product and, dropping the constants given by the variance's first factors, we get

$$p(\mu|x) \propto \exp\left\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \propto$$

$$\propto \exp\left\{-\frac{\mu^2(\sigma^2+n\sigma_0^2) - 2\mu(\mu_0\sigma^2+\sigma_0^2\sum_i x_i))}{2\sigma^2\sigma_0^2}\right\}$$

where the remaining terms in the numerator are not being considered since they do not depend on μ.

$$p(\mu|x) \propto \exp\left\{-\left(\mu^2 - 2\mu\frac{(\mu_0\sigma^2+\sigma_0^2\sum_i x_i)}{\sigma^2+n\sigma_0^2}\right)\cdot\frac{1}{2\frac{\sigma_0^2\sigma^2}{\sigma^2+n\sigma_0^2}}\right\}$$

To reconstruct the binomial we want, in the numerator of the exponential we sum and subtract the term

$$\frac{(\mu_0\sigma^2+\sigma_0^2\sum_i x_i)^2}{(\sigma^2+n\sigma_0^2)^2}$$

and then by dropping the subtracted one (we don't need it and it's not a function of μ), we finally obtain

$$p(\mu|x) \propto \exp\left\{\frac{-\left(\mu - \frac{\mu_0\sigma^2+\sigma_0^2\sum_i x_i}{\sigma^2+n\sigma_0^2}\right)^2}{2\frac{\sigma_0^2\sigma^2}{\sigma^2+n\sigma_0^2}}\right\}$$

and so $p(\mu|x) \propto \mathcal{N}(\mu_1, \sigma_1^2)$, where

$$\mu_1 = \frac{\mu_0\sigma^2+\sigma_0^2\sum_i x_i}{\sigma^2+n\sigma_0^2} = \sigma_1^2\cdot\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_i x_i}{\sigma^2}\right)$$

and

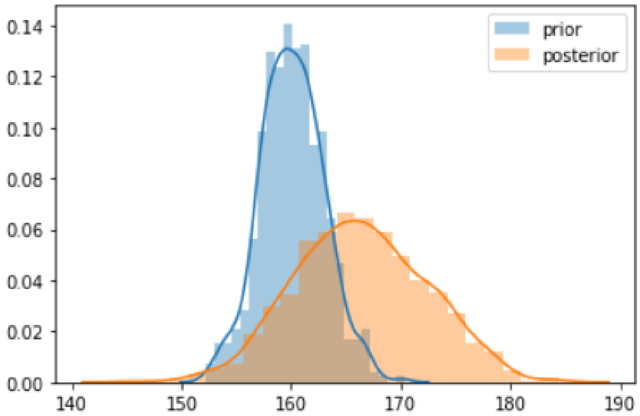
$$\sigma_1^2 = \frac{\sigma_0^2\sigma^2}{\sigma^2+n\sigma_0^2} = \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$$

3.

Plot the posterior distribution corresponding to the data x = [174, 158, 194, 167] together with the prior distribution.

```
# prior distribution object
mu0 = 160
sd0 = 3
prior = dist.Normal(mu0, sd0)

# observations
x = [174, 158, 194, 167]

# posterior distribution object set up
sd = 6
new_mean = (sum(x)/sd**2 + mu0/sd0**2) / (1/sd0**2 + len(x)/sd**2)
posterior = dist.Normal(new_mean, sd)

# drawn some samples to plot
sample_prior = prior.sample((900,))
samples_posterior = posterior.sample((900,))

# plot
sns.distplot(sample_prior, label='prior')
sns.distplot(samples_posterior, label='posterior')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x1fbaa4e1f60>
```



Exercise 4

Proof:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

is the beta prior, while

$$p(x|\theta) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i-1} = \theta^n(1-\theta)^{\sum_{i=1}^{n} x_i-1}$$

is the geometric likelihood, so as a result

$$p(\theta|x) \propto \theta^{\alpha-1+n}(1-\theta)^{\beta-1+\sum_{i=1}^{n} x_i-n}$$

which is a Beta distribution with $\alpha' = \alpha - 1 + n$ and $\beta' = \beta + \sum_{i=1}^{n} x_i - n$