Applied Statistics
Problem Set

Andras Csepreghy
xgj708@alumni.ku.dk

University of Copenhagen — January 5, 2020

I Distributions and probabilities:

I.1

(6points) Little Peter goes to the casino and puts money on black ($p = 18/37$)

Question: In 50 games, what are the chances that he will win exactly 25 times?

Answer: Playing the game of roulette has discrete trials and the outcome consists of two mutually exclusive possible outcomes that are winning or losing with probability p and $1-p$ (Bernoulli trials). Since one trial is independent of any other, the probability of Little Peter winning n times follows a **binomial distribution**.

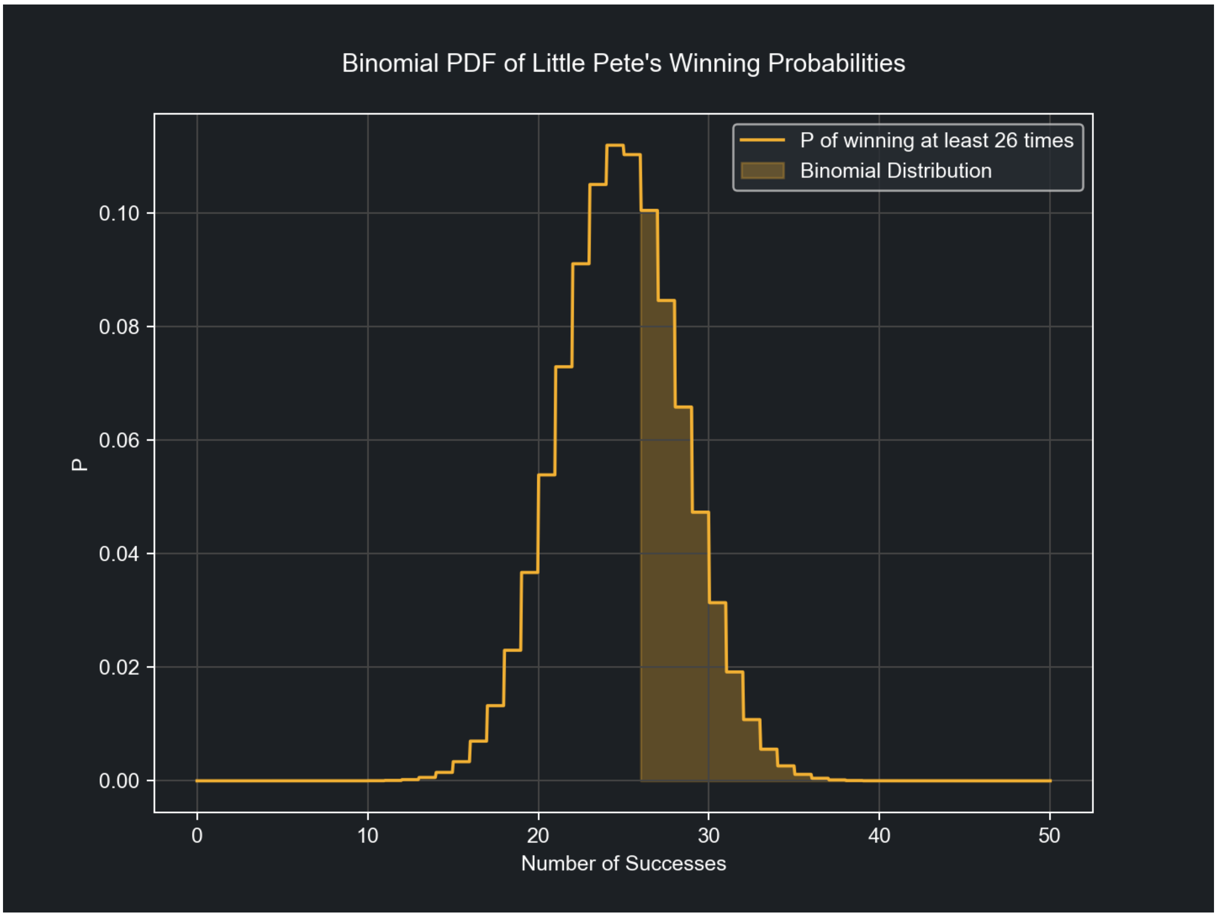

Figure 1: $B(50, \frac{18}{37})$ The binomial distribution given by probability of Little Pete winning n times in 50 trials with each trial being $p = 18/37$

The probability of winning exactly 25 times is:

$$\binom{50}{25} = \frac{50!}{25!(50-25)!}\left(\frac{18}{37}\right)^{25}\left(\frac{19}{37}\right)^{25} = 0,11024$$

Which is the same as using the binomial distribution in Python and plugging in the numbers for $x = 25$:

```
def func_binomial_pmf(x, n, p):
    return binom.pmf(np.floor(x), n, p)

x = 25
n = 50
p = 18/37

pete_wins_25 = func_binomial_pmf(x, n, p) # 0.11024273088617097
```

The probability for winning at least 26 times can be computed by calculating each exact success probability from 26 to 50 and summing them, or in the case of the binomial distribution by integrating over the distribution function from 26 to 50. Both methods give the result with $p = \frac{18}{37}$:

$$I(50, 26) = \int_{26}^{50} \binom{50}{26} p^26(1-p)^{50-26} dp = 0.32449$$

Question: How many times does he have to play in order to be 95% sure of winning at least 20 times?

Answer: To see how many trials it takes to have 20 successes with at least $p = 0.95$ it can be calculated by computing the integral from 20 to the number of trials n and increasing n until $p \geq 0.95$. Since it's relatively easy to compute the integral of a binomial, I wrote my own function, which gives the result $n = 53$ trials are required to have 20 or more successes each of which has a $p = 18/37$ The exact answer is a faction but it only makes sense to talk about an integer number of trials.

1.2

Gaussian Distribution

Question: What is the probability of a Gaussian value to lie between 1.25σ and 2.5σ away from the mean?

Answer: 0.198880
Which can be derived by integrating the Gaussian distribution function from 1.25 and 2.5 and multiplying the result by 2 since we are interested in both sides as the plot shows.

1.3

(6 points) The number of S-train delays is counted daily. Assume the following, that delays are uncorrelated, and that the number of departures is the same every day.

Question: What distribution should the number of daily delays follow?

Answer: It should follow a **Poisson Distribution** because a poisson distribution is the discrete probability distribution of the number of events occurring in a given time period, given the average number of times the event occurs over that time period.
In our case we know the given time period in which we are observing n number of events happening that range from 0 to theoretically infinity and although we don't know from the exercise the average frequency of the events, we have enough information to calculate it.

Question: Days with more than 7 delays are considered "delay days". if there were 19 "delay days" in a normal year, what is your estimate for the average number of daily delays?

Answer: Since we know the distribution follows a Poisson distribution this question can be rephrased as: What is the lambda that will produce the Poisson distribution for which values between 8 and infinity

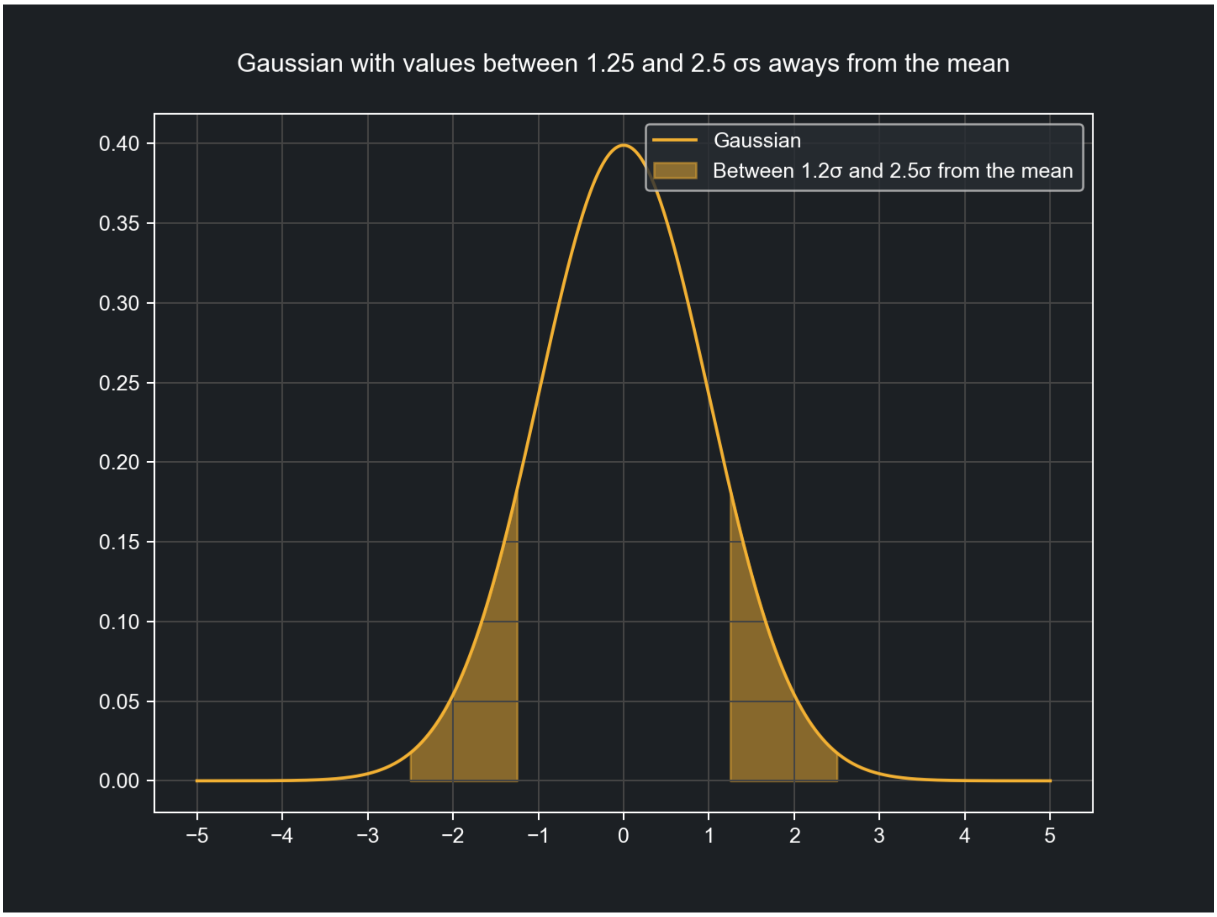

Figure 2: The Gaussian distribution function and values that lie between

would integrate to 19/365? Or when we normalize the distribution to have its integral be 365 instead of 1, the resulting integral between 8 and infinity would result in 19? Which may sound complicated, but Figure 3 shows the answer to this question.
For this solution I made a function that minimizes the absolute difference between 19 and the integral from 8 to infinity which resulted in $\lambda = 4.01538$. **My estimate for average delays is 4.01538**

II Error Propagation:

II.1 2.1

A measurement of a tumor depth (in cm) was done using two methods. The first gave 4 measurements with uncertainty while the second gave 12 without.

Question: Do the measurements with uncertainty agree with each other?

Answer: The chi-square value for the four measurements with uncertainties is: **15.19142**
Which gives a p value of: **0.00166**

From this we can conclude that the measurements are inconsistent with each other. Which we can confirm by eye looking at Figure 4 one could
Looking at Figure 3 one can conclude that 3 measurements agree with each other, however one stands out and is several sigmas away from the others which makes the 4 measurements inconsistent with each other. However since 3 out of the 4 agree, one could make the case that it is a good solution to just remove the outlier.
As Figure 4 clearly shows measurements without uncertainties seem to be

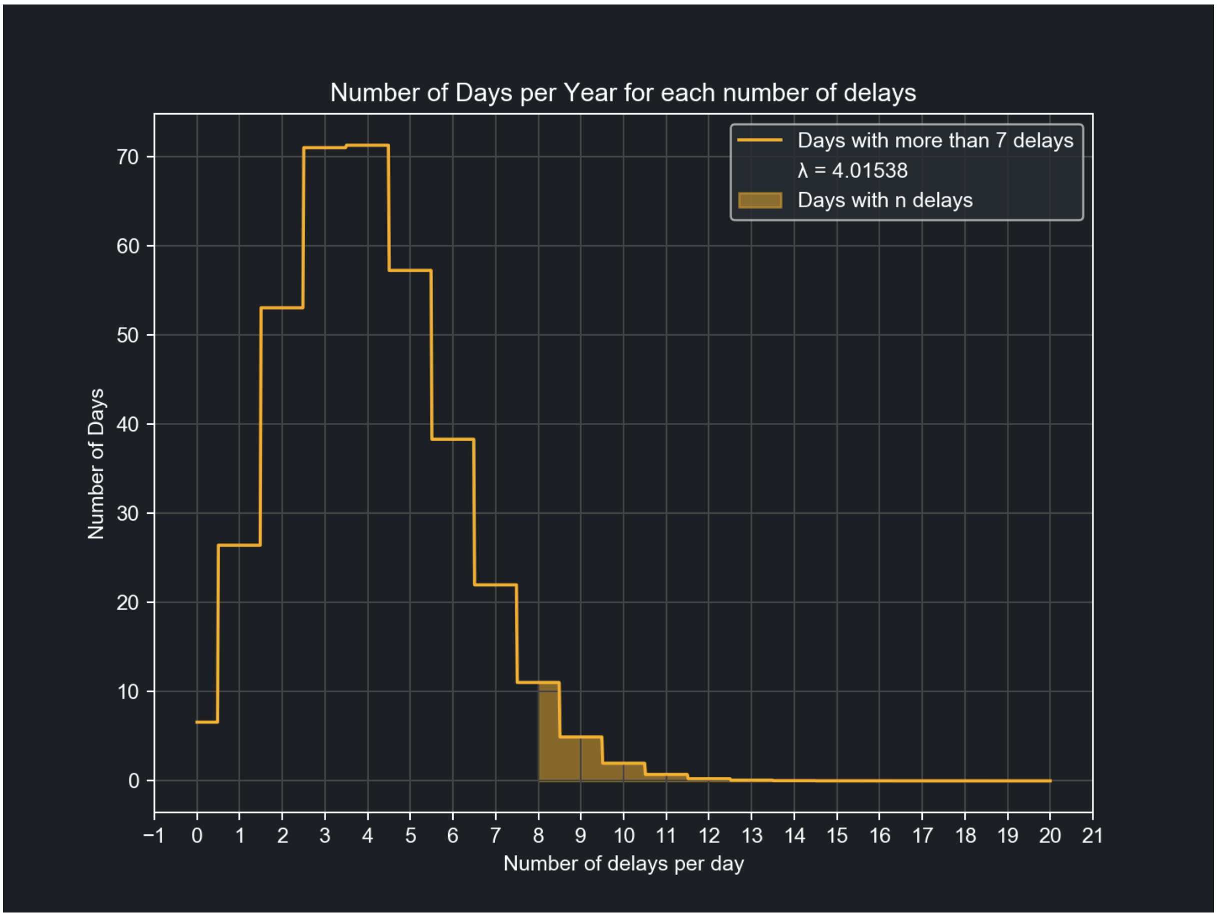

Figure 3: The poisson distribution with lambda 4.01538 and its integral (which equals 19) filled in with orange from values bigger than 7

B mean = 1.9895742317420576e-08 with no correlation
B error = 5.269953525145987e-09 with no correlation

B mean = 1.969873335231671e-08 with 0.87 correlation
B error = 3.814983178006397e-09 with 0.87 correlation

2.2

III Monte Carlo:

III.1 3.1

C = 0.66262122
rms = 0.5276481799510685 Hellooka **Questions:**

1. What is the mean and RMS of $f(x)$? Also, what is the value of C?
2. What method(s) can be used to produce random numbers according to $f(x)$? Why?

Answers:

1. RMS
2. The two methods that were mentioned in class are the *Transformation Method or Inverse Transform Sampling* and *Accept-Reject Method or Rejection Sampling* also known as *Von Neumann Method*.

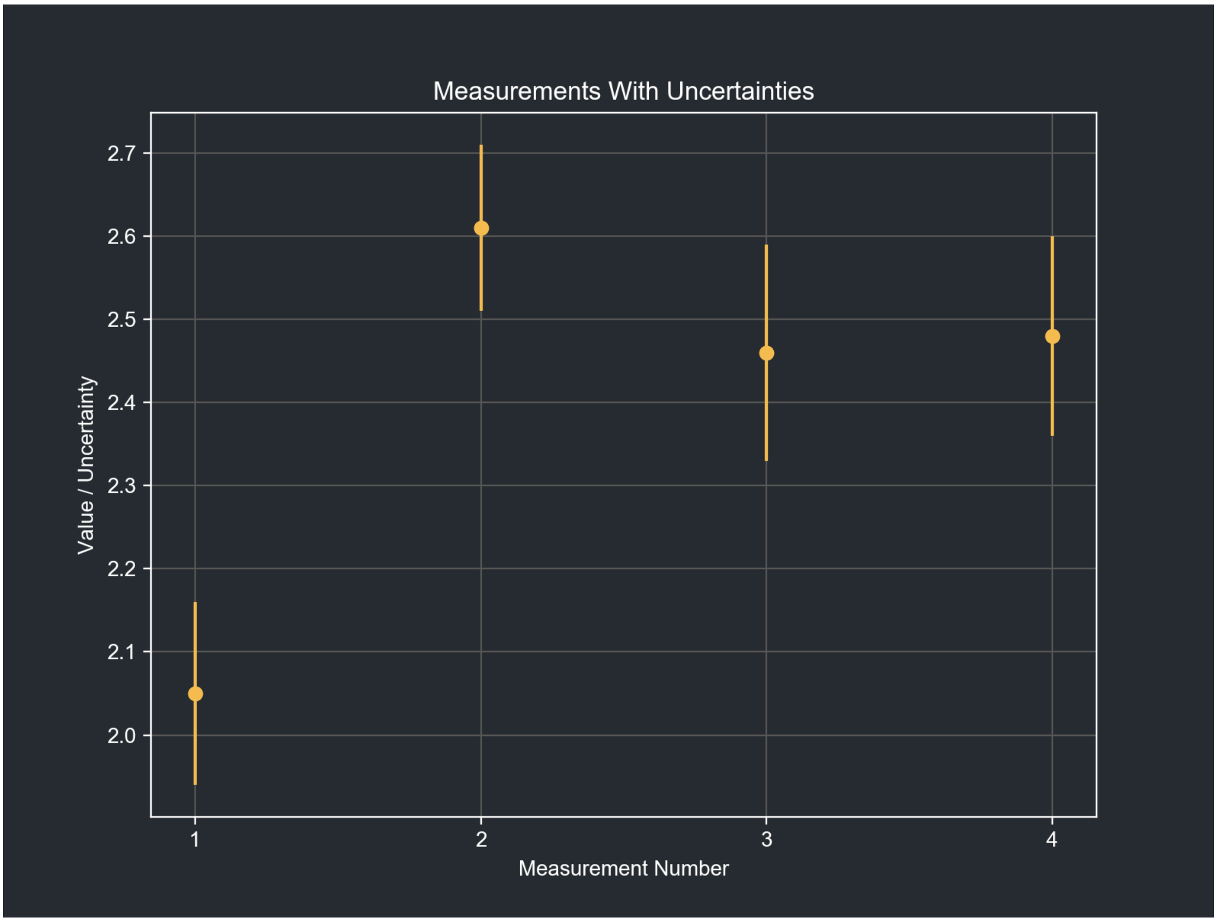

Figure 4: The 4 measurements with their corresponding uncertainties

For the transformation method it is necessary that the $f(x)$ can be integrated, and then its integral can be inverted. The integral of a PDF is known as a Cumulative Distribution Function (CDF). So we are looking for the inverse of the PDF's corresponding CDF. Then we can sample u from a uniform distribution $u_i \sim U(0, 1)$ then $x_i = CDF^-1(u_i)$, so x will be sampled according to our original $f(x)$. In our case it looks like $f(x)$ can be integrated and its integral can be inversed. When it is possible to use this method it is more computationally effective since we don't throw away any sampled numbers where as in Von Neumann method we reject some portion of our sampled numbers.

Rejection sampling requires a the function to be finite in x and y, which works well in our case. However even with a function that is infinite in x and y can be approximated by choosing large enough upper bounds. Using this method we can generate random numbers according to $f(x)$ by first $x_i \sim U(0, 2)$ and then $y_i \sim U(0, 0.65048)$. Reason for choosing 0.65048 as the upper bound is because it is the largest value the function takes. Then we decide to accept or reject the value for x if the sampled y value falls under the curve of $f(x)$, otherwise we reject it and sample again. As the number of samplings go to infinity the sampled random numbers approximate the distribution more and more.