Applied Statistics

Problem Set

Andras Csepreghy
xgj708@alumni.ku.dk

University of Copenhagen — January 5, 2020

I Distributions and probabilities:

I.1

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**.

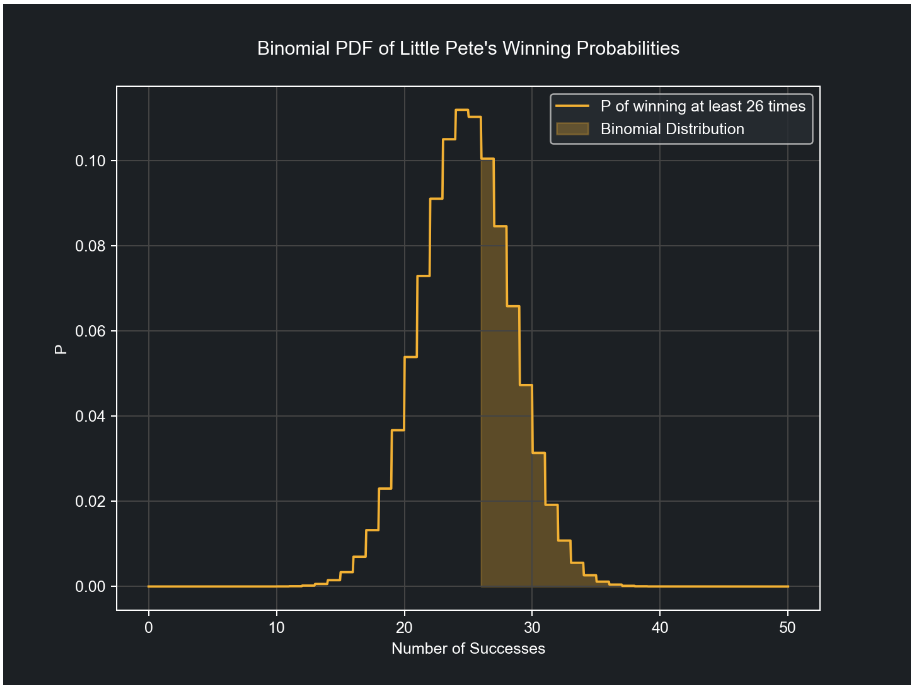

Figure 1: $B(50, \frac{18}{37})$ The binomial distribution given by probability of Little Pete winning n times in 50 trials with each trial being $p = 18/37$

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.

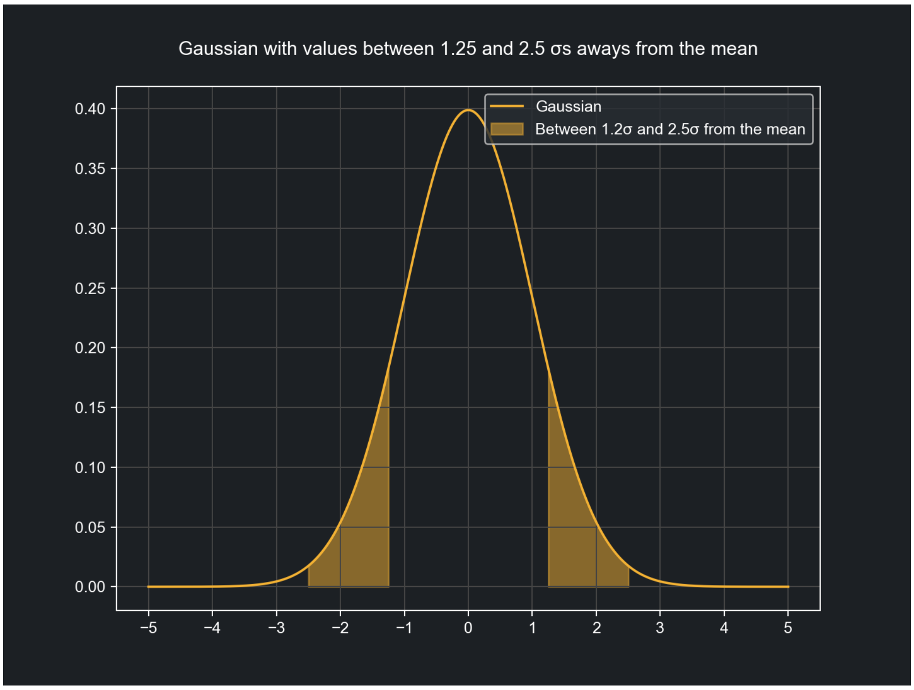

Figure 2: The Gaussian distribution function and values that lie between

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 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**

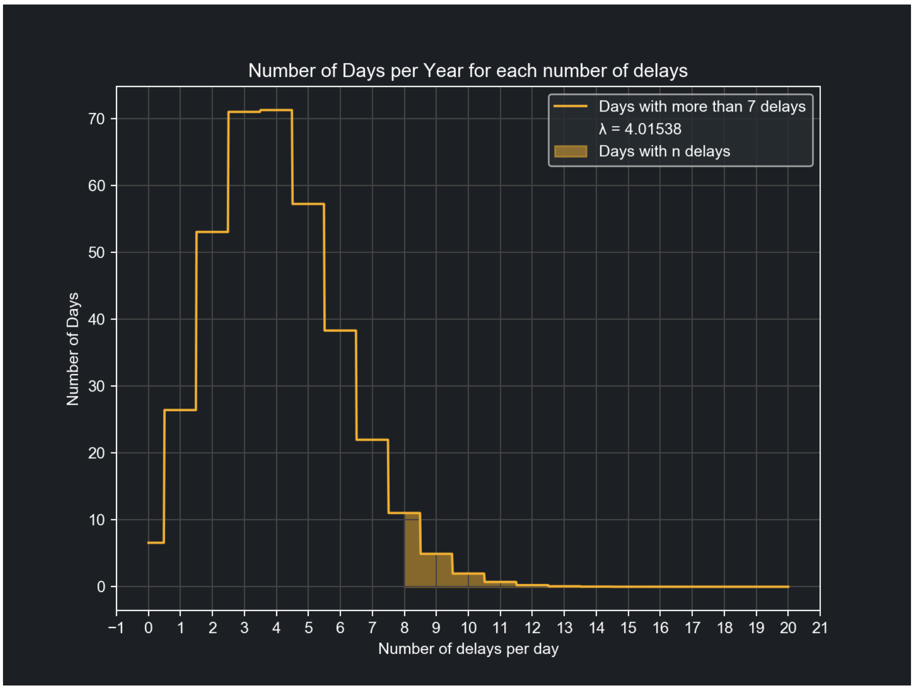

Figure 3: The poisson distribution with lambda 4.01538 and its integral (which equals 19) filled in with orange from values bigger than 7

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: $\chi^2 =$ **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 by looking at Figure 4. It appears that 3 out of the 4 measurements agree with each other and one is several sigmas away from the mean. So one could argue that we can exclude the "faulty" measurement and keep the 3 that agree with each other.

Question: Do the measurements without uncertainty agree with each other?

Answer: Doing a normality test, which checks whether a set of datapoints are normally distributed gave a p value of 0.17395, which means we can conclude that the 12 measurements are consistent and normally distributed. Looking at **Figure 5**, we can confirm this by eye.

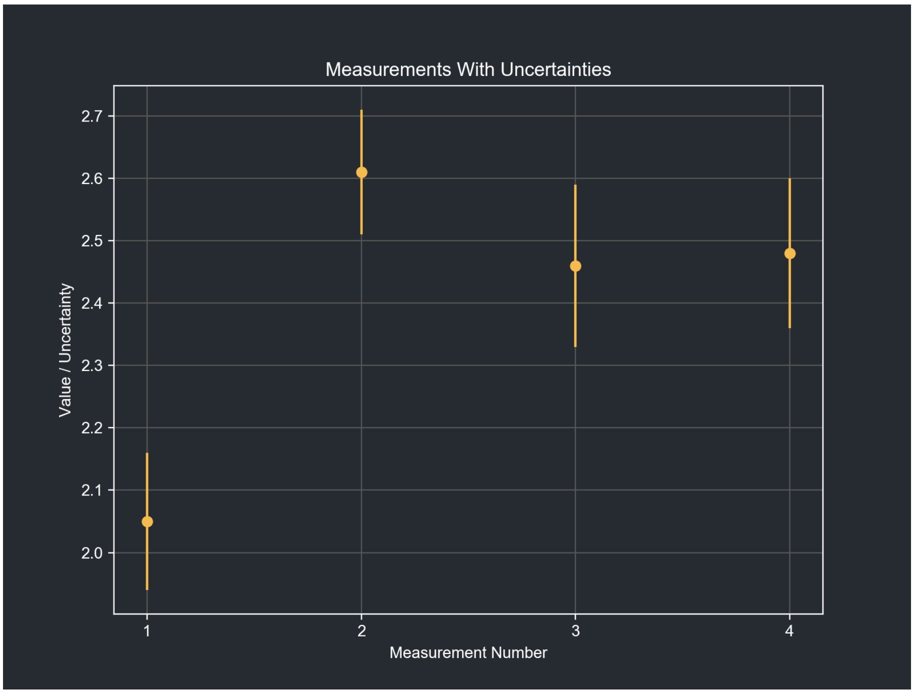

Figure 4: The 4 measurements with their corresponding uncertainties

Question: What is your best estimate of the tumor position? And with what uncertainty?

Answer: For this solution I used the 3 measurements from the ones with uncertainties that were consistent and all 12 measurements from the ones without uncertainties. I set the uncertainty for the 12 measurements to be their standard deviation, then took the weighted mean of the all measurements taking their uncertainties into account. Best estimate for tumor position: **2.59 ± 0.12**

2.2

Planck's Law

Question: Given values of $v = (0.566 \pm 0.025) \times 10^15$ Hz and $T = (5.50 \pm 0.29) \times 10^3$ K (uncorrelated)

Answer: For this solution I simulated v and T 10,000 times as normally distributed random variables with their mean and uncertainty (standard deviation) given in the exercise text. For each v T value pair I computed $B(v,T)$ according to Planck's Law equation. From the new standard deviation given by the 10,000 resulting values I computed the uncertainty on the result:
$B(v,T) = 1.97 \pm 0.53 \times 10^8$

Question: How does the uncertainty change, if there is a correlation of $\rho(v,T) = 0.87$

Answer: Based on the method from Barlow (page 43-44) I introduced correlation between the two

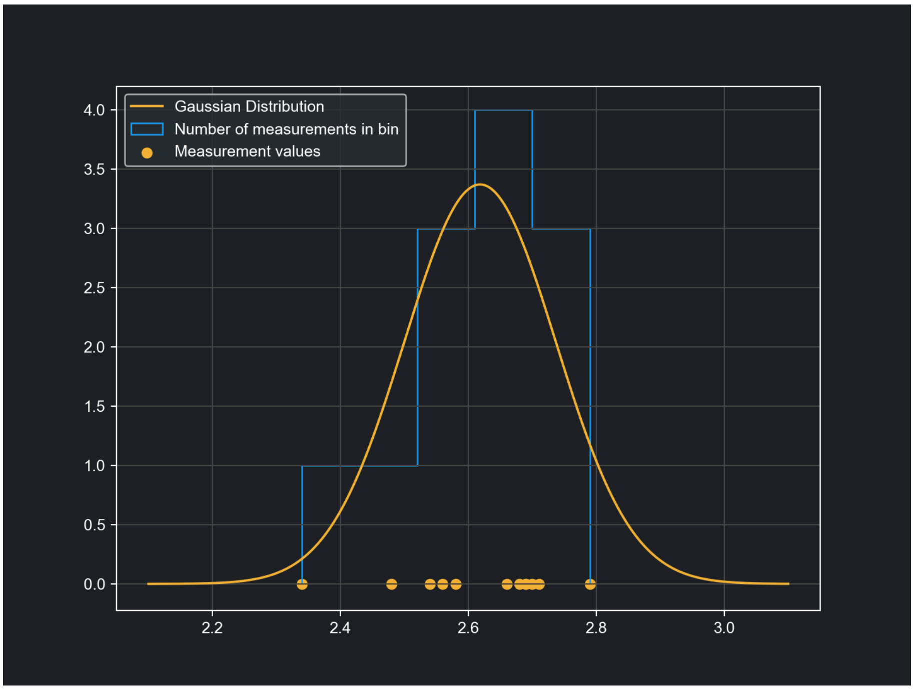

Figure 5: The 12 measurements without uncertainties and the bins that approximate a normal distribution

normally distributed random variables and computed another 10,000 simulated value pairs from which I computed 10,000 $B(v,T)$ values and took their standard deviation. The result:
$B(v,T) = 1.95 \pm 0.37 \times 10^8$

III Monte Carlo:

3.1

Let $f(x)$ be a PDF defined as $f(x) = C(1 - e^{-ax})$ for $x \in [0,2]$ and $a = 2$.

Question: What is the mean and RMS of $f(x)$? Also what is the value of C?

Answer:

- Mean: 0.49982
- RMS: 0.52759
- C: 0.66262

I calculated C by minimising the absolute distance between the integral of the function and 1 so as to get the integral of the function to equal 1.

Question: What method(s) can be used to produce random numbers according to $f(x)$? Why?

Answer:
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*.
Inverse Transform Sampling:
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.
Von Neumann Method:

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.

Question: Produce 500 random numbers distributed according to $f(x)$ and plot these.

Answer: To create this plot (Figure 6) I used the Von Neumann method that I described earlier.

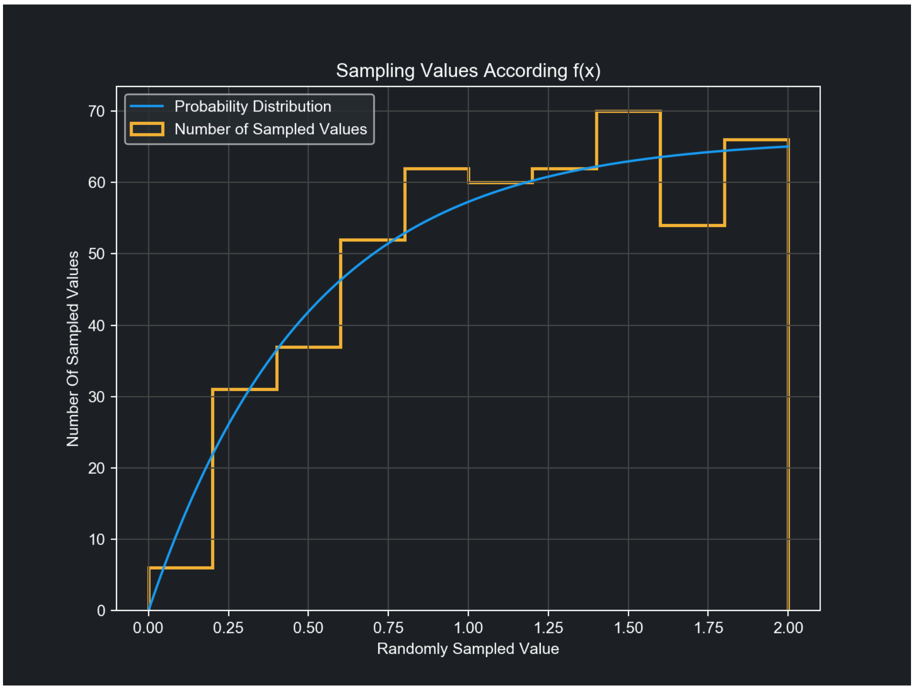

Figure 6: Randomly sampled values according to $f(x)$ based on Von Neumann method

Question: Fit the numbers you produced above leaving a as a floating parameter.

Answer: The original sampling with $a = 2$ produces a $\chi^2 = 412.320$. After fitting and leaving a as a floating parameter I got a $\chi^2 = 364.540$ where $a = 4.6558$

IV Statistical tests:

4.1

National UFO Reporting Center

Question: Plot the distribution of duration of observation, and calculate both mean and median.

Answer: The plot is Figure 7

- Mean: **471.33**
- Median: **180.0**

Question: Do these durations follow the same distribution on the East and West coast?

Answer: Looking at Figure 8, it appears that the two distributions are very similar, both following an exponential distribution. I have done the Kolmogorov–Smirnov test which gives a p value of 0.09220 which is rather small, however if we set the threshold to be 0.05, we can't reject the null hypothesis that says the two sets of values (East coast durations and West cosat durations) come from the same distributions.

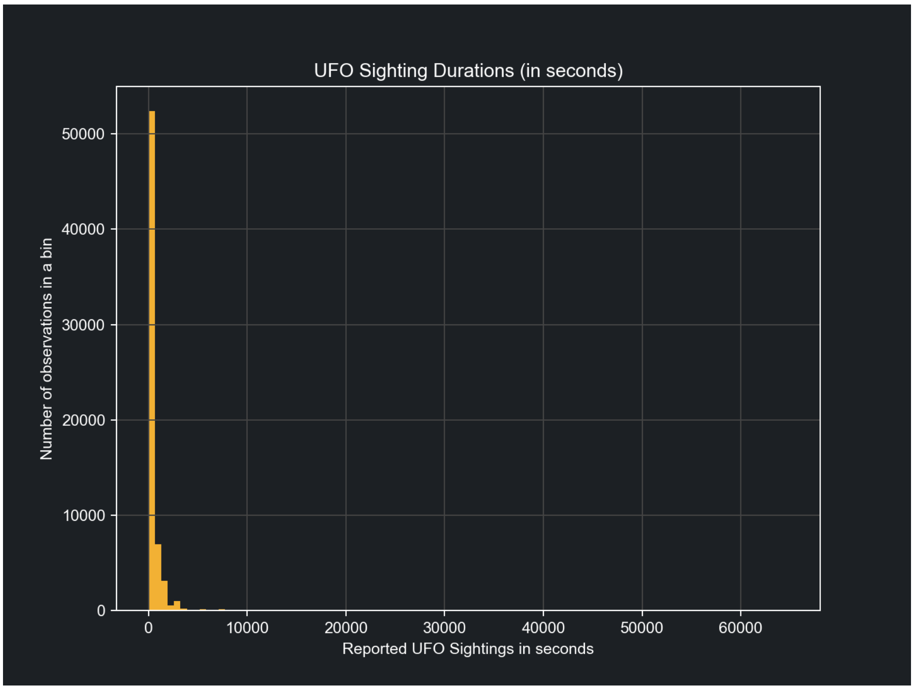

Figure 7: UFO Sighting Durations (in second)

Question: What is the correlation between day in the year and and time of the day of observation?

Answer:
Pearson correlation: 0.024
Spearman correlation: -0.010

From these two types of correlations we can safely say there is no correlation between the time of year and the hour of day.

Question: Considering only the West Coast, is the distribution of number of observations uniform over the seven week days?

Answer: Comparing the number of observations for each day to a uniform distribution we obtain $\chi^2 = 250.022$ which gives a $p = 4.057 \times 10^-51$. We can safely say that this is not a uniform distribution. This is shown in Figure 8

Question: How about when considering only Monday to Thursday?

Answer: Comparing the number of observations for each day from Monday to Thursday to a uniform distribution we obtain $\chi^2 = 9.8132$ which gives a $p = 0.02022$. This distribution is much closer to a uniform distribution and has a much better χ^2 however the p value could still be considered too low. This is shown in Figure 9

4.2

Fairness of dice

Question: What distribution should the number of 5s and 6s follow?

Answer: There are a given number of total Bernoulli trials (26306) and each have a certain probability of success or each trial, which means the results should follow a binomial distribution. The probability of success is 1/3 because there are 6 possible outcomes for each die to have and we consider an outcome success if it's either 5 or 6, that gives 2/6 = 1/3.

Question: Compare the data with the expected distribution. Does this hypothesis match the data well?

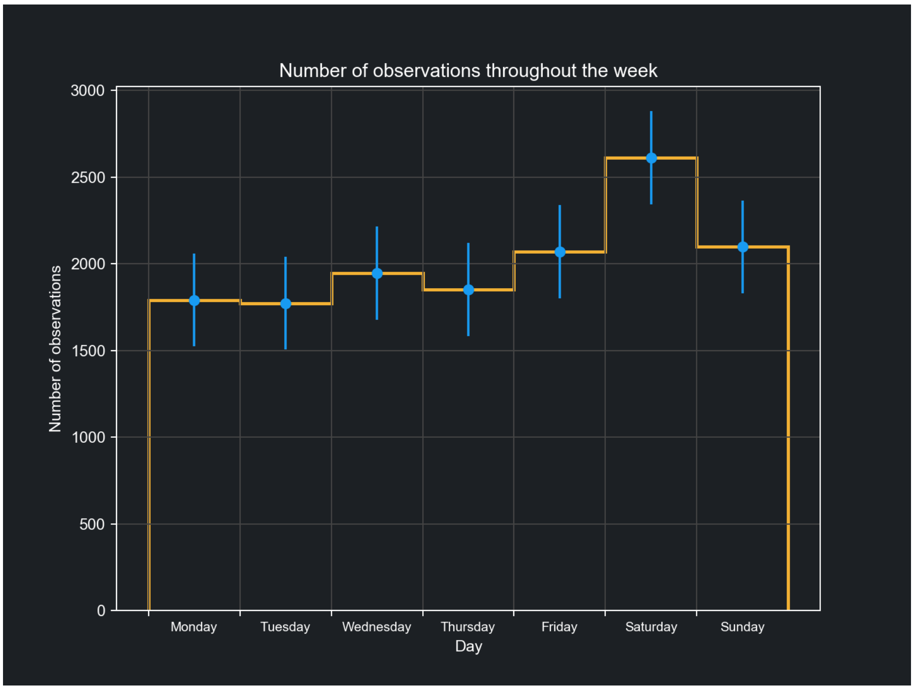

Figure 8: Number of observations throughout the week

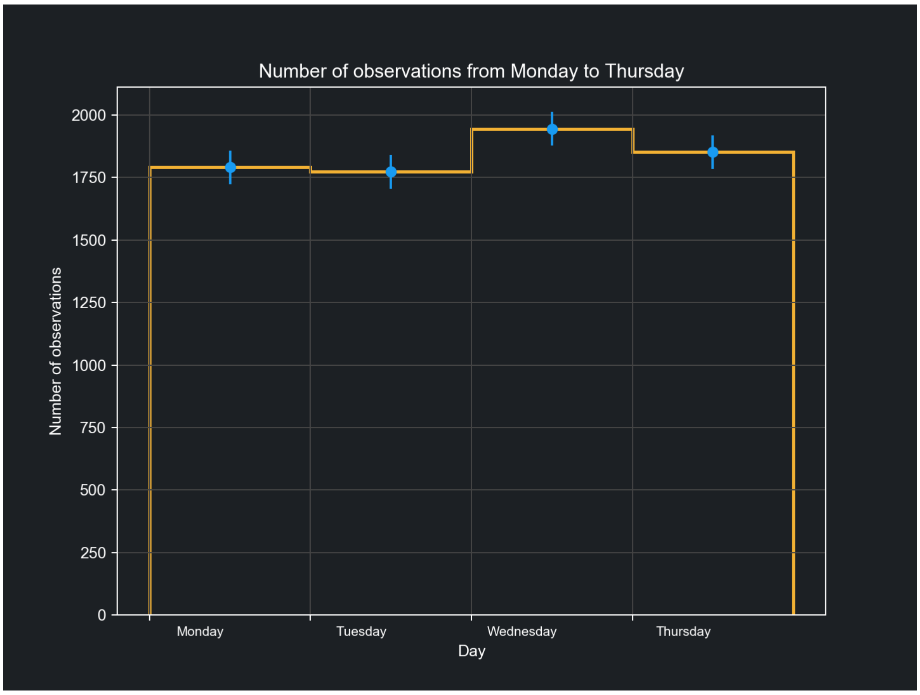

Figure 9: Number of observations from Monday to Thursday

Answer: By looking at the plot of the data it could be suggested that the binomial distribution with $n = 12$ and $p = 1/3$ matches the data well, however the χ^2 paints a different picture.

For this the binomial distribution with $p = 1/3$
$\chi^2 = 41.312$
$p = 4.3448 \times 10^-05$

From the χ^2 and the resulting p value one can conclude that the observed values significantly differ from the distribution. See Figure 10

Question: Fit the data and test if alternative hypothesis match the data better. Also, determine the

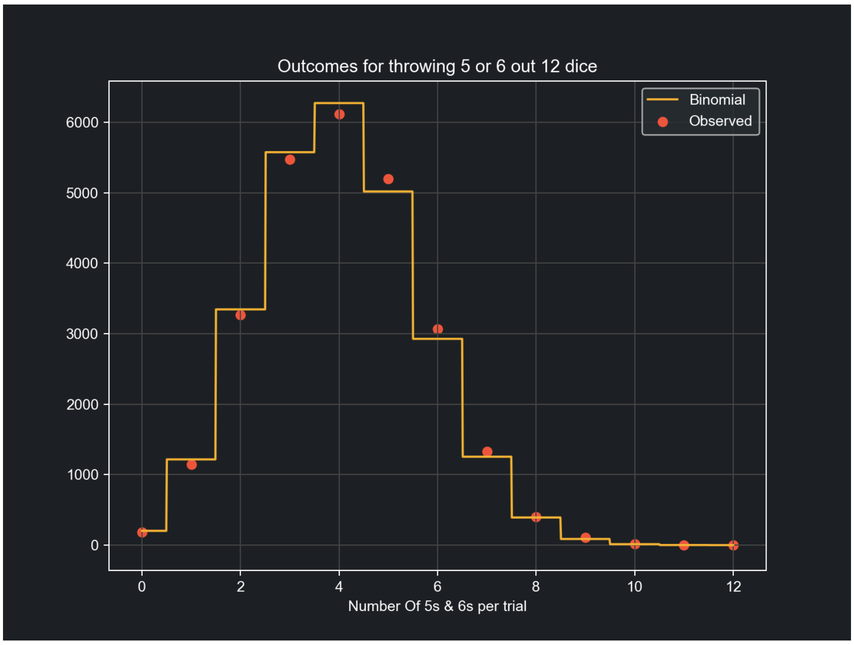

Figure 10: Observed dice outcomes and the binomial distribution with p = 1/3

probability for a 5 or a 6, and decide if the dice are consistent with being fair.

Answer: I tested 3 distributions: Gaussian, Poisson and Binomial. I fitted each to find the parameter that minimizes their χ^2. These are the results:
- **Gaussian**: $\chi^2 = 155.869$, $p = 3.640 \times 10 - 27$, Optimal $\sigma = 1.6783$
- **Poisson**: $\chi^2 = 1628.647$, $p = 0$, Optimal $\lambda = 4.09910851$
- **Binomial**: $\chi^2 = 13.152$, $p = 0.35807$, Optimal $p = 0.33776$

Because the probability of success for the binomial gives the optimal χ^2 for 0.3377 and fair dice would give 0.3333 we can conclude that these dice are fair. Also, see the fitted distributions in Figure 11.

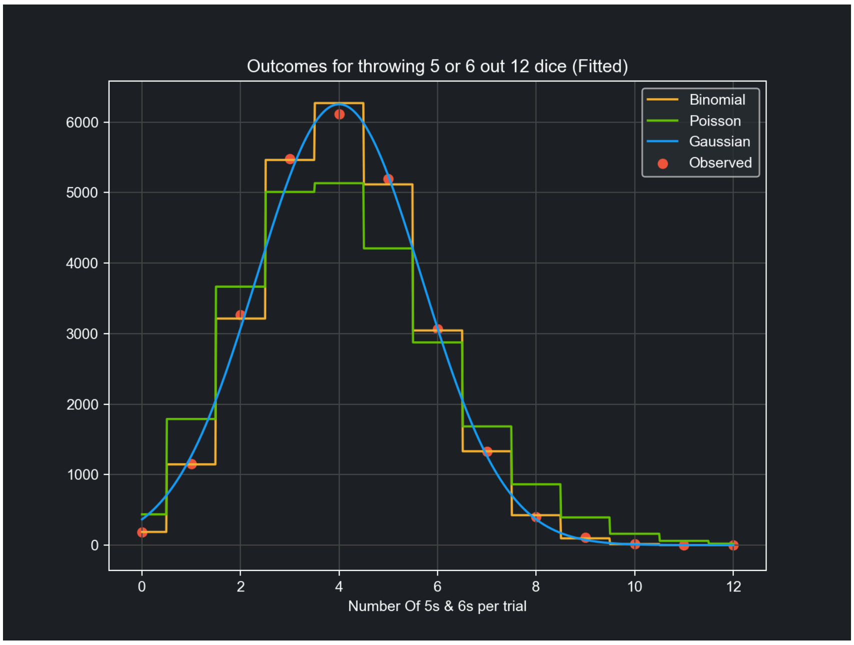

Figure 11: Observed dice outcomes and 3 distributions fitted