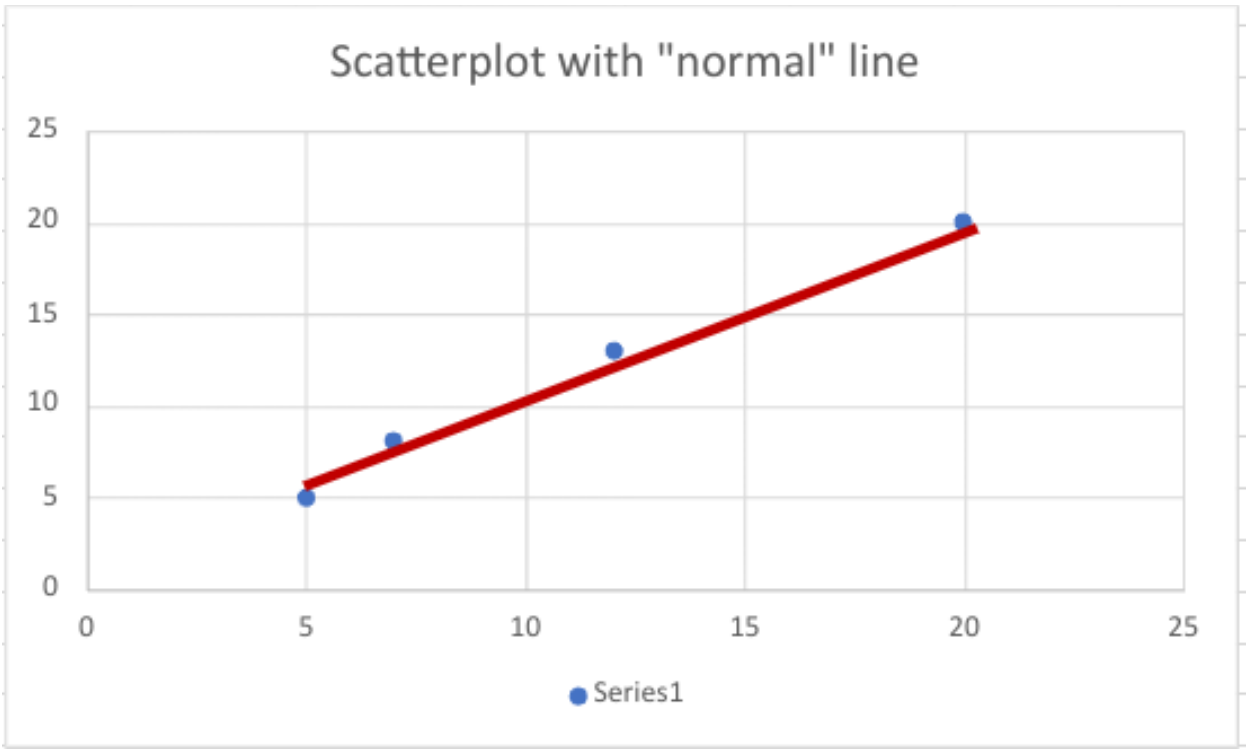
Testing normality in random variables with QQ-plots and estimating the parameter λ in the Poisson Distribution

Dylan Bolger

Introduction

The quantile-quantile plot is a good measure to review how much a distribution compares to a normal distribution. Simply put, we can draw a line starting at either the bottom or the top of the left-most point on the graph. We choose the bottom or the top depending on how the right-most point is placed: if the right-most point is higher than the left-most point, the line should start above the left-most point, otherwise. For the right-most point, it should be the opposite vertical location compared to the left-most point.



In this example, the left-most point is lower than the right-most point. In this case, we should start the line on top of the left-most point and end it below the right-most point. This red line against a QQ-plot gives us an idea of if a distribution is normal or not with a large enough n (n > 30).

With this in mind, we can begin plotting some distributions and seeing how they compare to normal distributions.

Methodology

The way I will be testing the random variables is by using the following steps:

1. Create 100 random numbers (with specific parameters that will be described in each test).
2. Sort the numbers in ascending order.

3. Assign a rank value in ascending order, starting at 1 and ending at 100.
4. Calculate the rank proportion:
    a. Subtract 0.5 from the rank value of the random number, then divide by the total number of random numbers, which is 100.
5. Obtain the z-score that corresponds to the rank proportion.

3 * RAND() + 5	RAND(5, 8) Sorted	Rank	Rank Proportion	Rank-based z-scores	Random value between 5 and 8
7.061096344	5.011279801	1	0.005	-2.575829304	5.011279801
6.771750761	5.032923643	2	0.015	-2.170090378	5.032923643
7.344502318	5.057089572	3	0.025	-1.959963985	5.057089572
7.813451599	5.064436342	4	0.035	-1.811910673	5.064436342
5.810540006	5.107325463	5	0.045	-1.69539771	5.107325463
5.958837553	5.183067057	6	0.055	-1.59819314	5.183067057
7.128173822	5.196012346	7	0.065	-1.514101888	5.196012346
6.430176555	5.205135012	8	0.075	-1.439531471	5.205135012
7.346208834	5.252478178	9	0.085	-1.372203809	5.252478178
5.663929727	5.262185612	10	0.095	-1.310579112	5.262185612
7.906631616	5.274561593	11	0.105	-1.253565438	5.274561593
6.813136608	5.277653025	12	0.115	-1.200358858	5.277653025
5.06884648	5.286924501	13	0.125	-1.15034938	5.286924501
7.539919886	5.304186228	14	0.135	-1.103062556	5.304186228
6.864928796	5.330183059	15	0.145	-1.058121618	5.330183059
5.86740084	5.354077273	16	0.155	-1.015222033	5.354077273
5.35125824	5.374566616	17	0.165	-0.974113877	5.374566616
5.62943038	5.396159949	18	0.175	-0.934589291	5.396159949
7.758222833	5.463892941	19	0.185	-0.896473364	5.463892941
5.106305998	5.509889766	20	0.195	-0.859617364	5.509889766
6.745209351	5.521118727	21	0.205	-0.82389363	5.521118727
6.459924759	5.564705624	22	0.215	-0.789191653	5.564705624
6.17425063	5.565277092	23	0.225	-0.755415026	5.565277092
7.034373678	5.593892966	24	0.235	-0.722479052	5.593892966
6.540397619	5.626191114	25	0.245	-0.690308824	5.626191114
7.34052592	5.626292404	26	0.255	-0.658837693	5.626292404
7.206875367	5.638736336	27	0.265	-0.628006014	5.638736336
5.288577477	5.642508935	28	0.275	-0.597760126	5.642508935
6.883833771	5.671240975	29	0.285	-0.568051498	5.671240975
5.607396156	5.687106837	30	0.295	-0.53883603	5.687106837
5.297313058	5.700772451	31	0.305	-0.510073457	5.700772451
5.921360649	5.768165475	32	0.315	-0.48172685	5.768165475
6.171018753	5.823422882	33	0.325	-0.45376219	5.823422882
5.637512469	5.83900125	34	0.335	-0.426148008	5.83900125
7.605069044	5.851789404	35	0.345	-0.398855066	5.851789404
6.237891036	5.911140531	36	0.355	-0.371856089	5.911140531
6.889649336	5.952547948	37	0.365	-0.345125531	5.952547948
6.122441529	5.975483364	38	0.375	-0.318639364	5.975483364
6.886140857	6.00282663	39	0.385	-0.292374896	6.00282663
5.02499005	6.010425092	40	0.395	-0.266310613	6.010425092
5.019234949	6.015670711	41	0.405	-0.240426031	6.015670711
5.272879808	6.065390549	42	0.415	-0.214701568	6.065390549
5.400809017	6.123210242	43	0.425	-0.189118426	6.123210242
7.078966261	6.155258053	44	0.435	-0.163658486	6.155258053

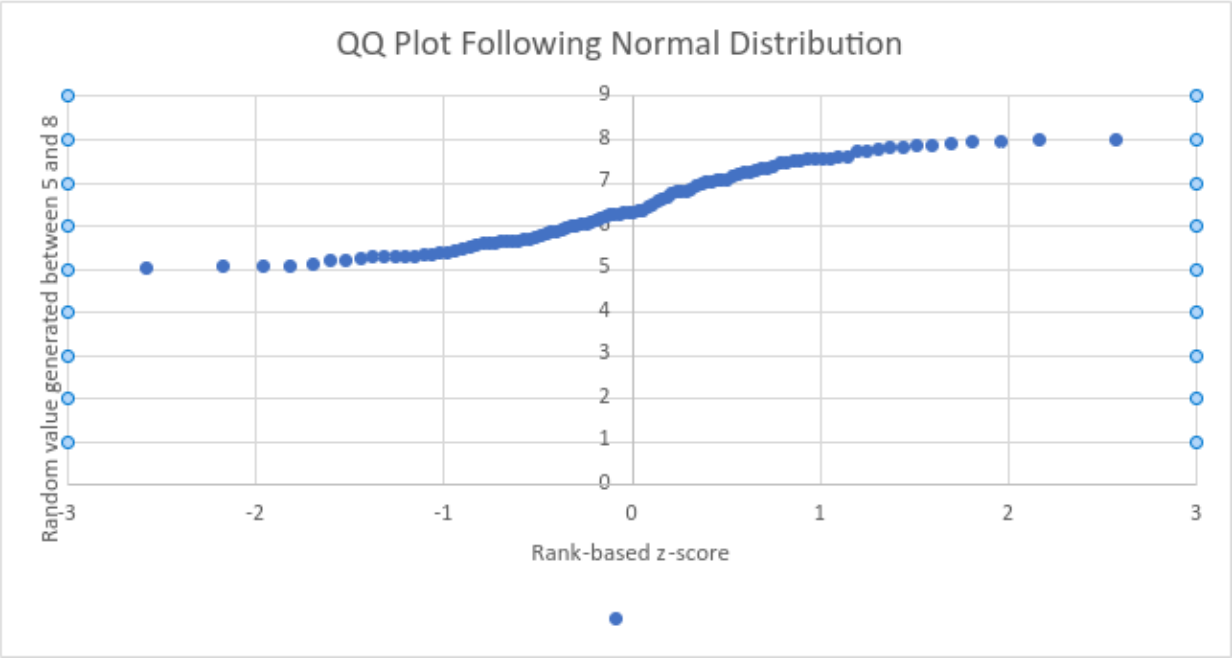
The first 45 uniform random numbers generated with their respective rank value, rank proportion, and z-score.

Plotting Uniform Random Variables

By creating some uniform random variables, we can create a QQ plot of the data and determine how similar it is to a normal distribution.

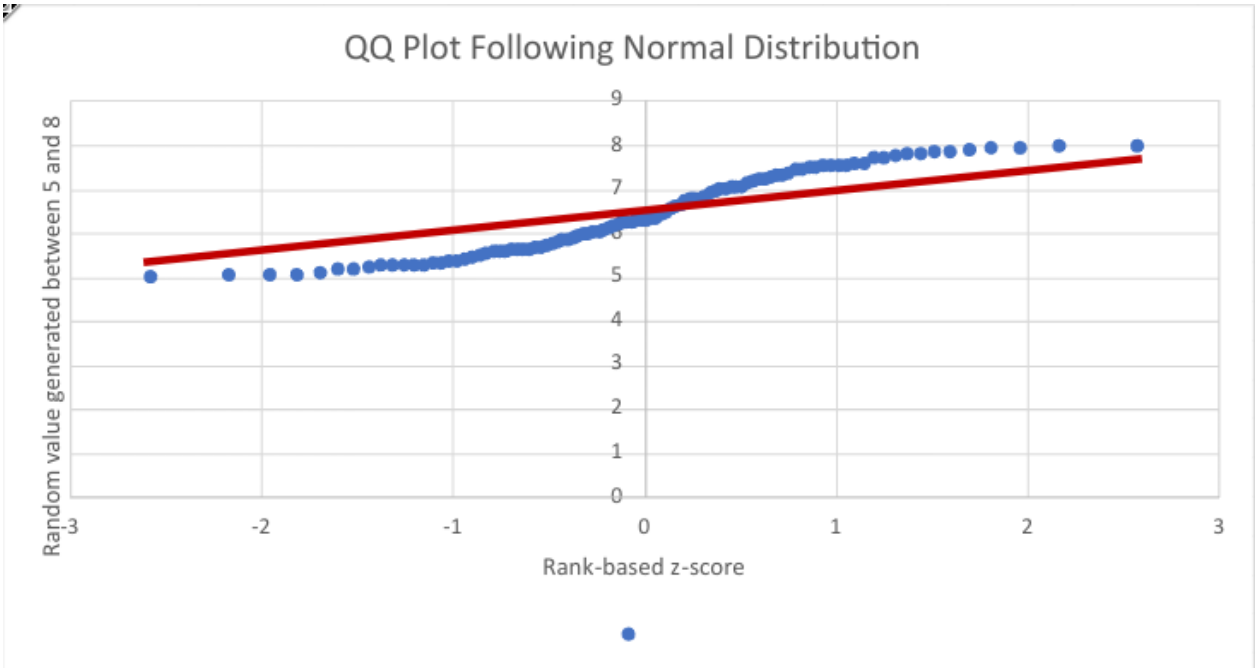
First, I created 100 random numbers between 5 and 8 inclusive using the **`3 * RAND() + 5`** function in Excel. After this step, I followed steps 2 through 5 in the methodology section.

Finally, I plot the z-scores with respect to the random value that was generated.



The uniform random variables plotted against their z-scores.

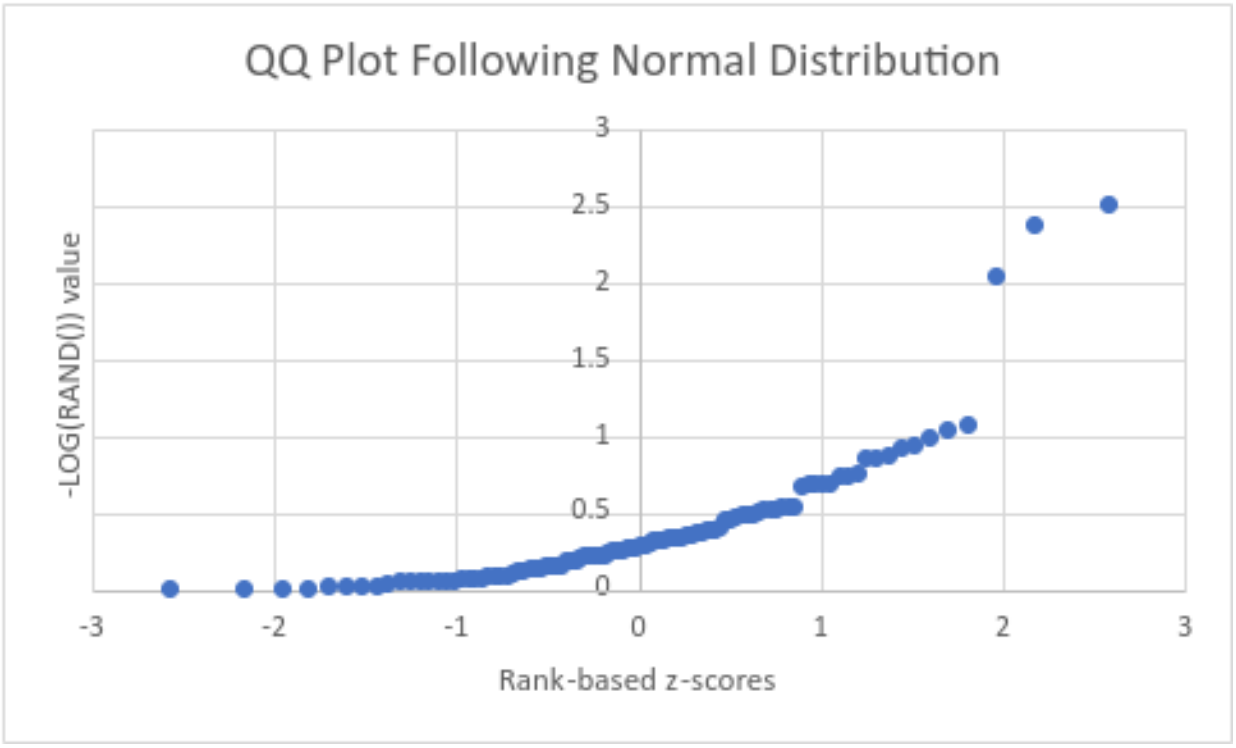
When the red line is drawn on the graph, I observed that many points do not lie on the line.



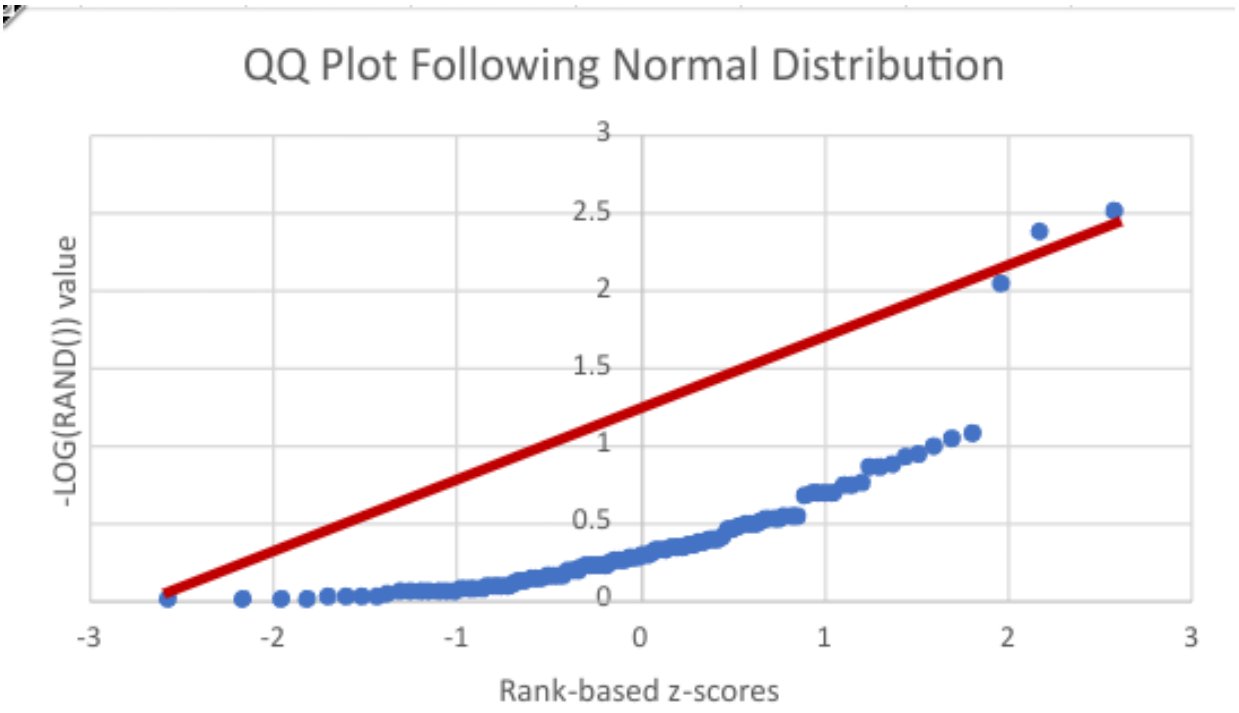
From this observation, I can conclude that the uniform random variables **do not follow a normal distribution**.

Plotting Exponential Random Variables

Like the previous section, I generated 100 random numbers; this time, the numbers are exponential. I used the function **`-LOG(RAND())`** in Excel to produce the numbers. Following the same steps as last time, I create a QQ plot to represent the data.

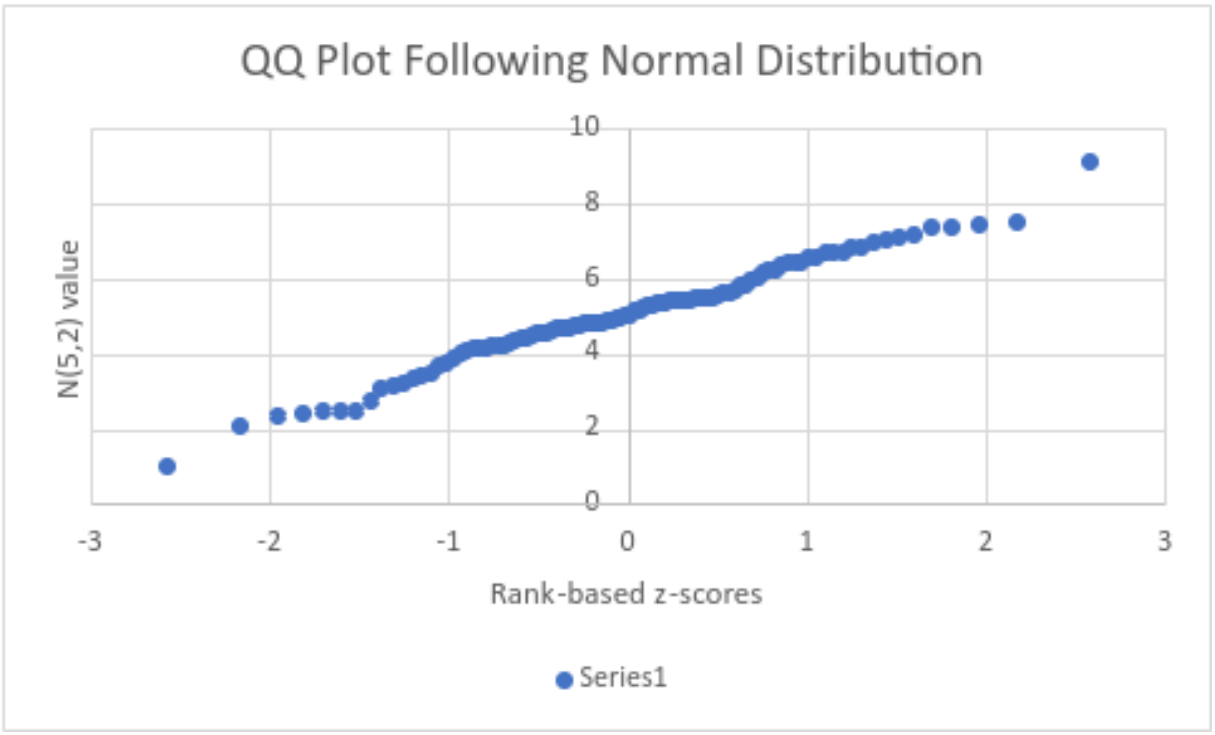


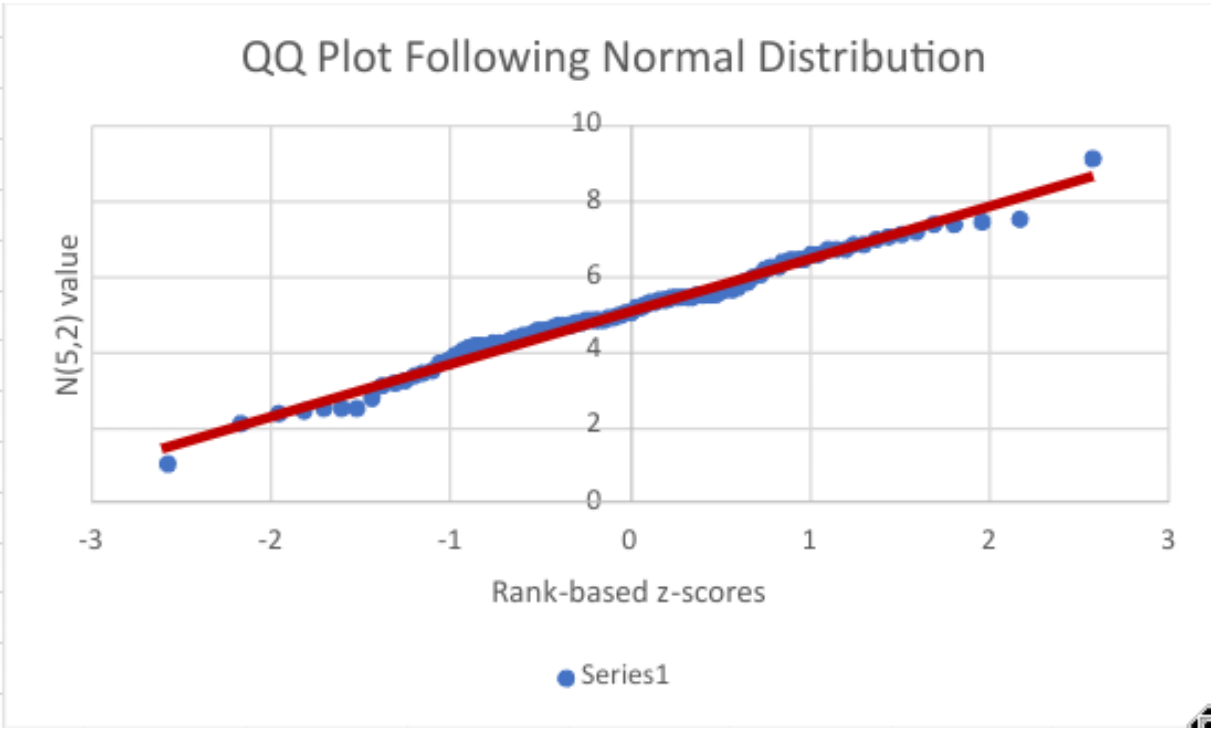
Using the same "normal" line comparison, I observed that the exponential random variables **do not follow a normal distribution**.



Plotting Normal Random Variables

Using the **`NORMINV(RAND(), 5, SQRT(2))`** function in Excel, I produced 100 normal random variables following the same methodology of the previous two types of random variables. These values each have a random probability on the normal distribution curve. All of the values have a mean of 5 and a standard deviation of the square root of 2.





I conclude that the datapoints generated from N(5, sqrt(2)) **follow a normal distribution**. Most of the datapoints in the random normal variables do lie within the line drawn.

Confidence Intervals

Now that I have a normal distribution, I can find the 90%, 95% and 99% confidence intervals of this dataset. Each of the confidence intervals contains the true population mean. Since we will later see that all confidence intervals include the number 5 (the true population mean), I can conclude that each of these confidence intervals include the true mean value.

To calculate the length of the intervals, I recalled the formula for getting the confidence interval's upper and lower bounds:

$$\bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

90% Confidence Interval

I plugged in 5 for X_{bar}. I use 1.645 as our $Z_{alpha/2}$ since that is the standard z-score of the 90% CI. For sigma, I plugged in the standard deviation, which is the square root of 2, and for n, the number of values produced, which is 100.

$$5 \pm 1.645\frac{\sqrt{2}}{\sqrt{100}}$$

I calculate the right side of the plus/minus to get the standard error:

$$1.645 * \frac{\sqrt{2}}{\sqrt{100}} \approx 0.233$$

To get the lower bound of the 90% interval, we subtract the standard error from the sample mean:

$$5 - 0.233 = 4.767$$

The upper bound is the additive side.

$$5 + 0.233 = 5.233$$

The resulting 90% confidence interval:

$$[4.767, 5.233]$$

To calculate the length of the confidence interval, we subtract the lower bound from the upper bound.

$$5.233 - 4.767 = 0.466$$

As previously mentioned, the number 5 lies within this interval.

95% Confidence Interval

$Z_{alpha/2}$ in this case is 1.96. All other values are the same.

$$5 \pm 1.96 \frac{\sqrt{2}}{\sqrt{100}}$$

Following the same steps as before to get the standard error:

$$1.96 * \frac{\sqrt{2}}{\sqrt{100}} \approx 0.277$$

The lower and upper bounds:

$$5 - 0.277 = 4.723$$

$$5 + 0.277 = 5.277$$

The 95% confidence interval:

$$[4.723, 5.277]$$

The length of the 95% confidence interval:

$$5.277 - 4.723 = 0.554$$

99% Confidence Interval

For the sake of brevity and making this document shorter, I will use the **CONFIDENCE** Excel function to calculate the last confidence interval. The confidence function will calculate the standard error. After we have the standard error, like in the previous sections, we will subtract/add that to the sample mean.

=CONFIDENCE(1-0.99, SQRT(2), 100)

For the first value in the function, we want the alpha of the confidence interval. The alpha refers to the difference in area under the set and the percentage of confidence we want. The next parameter is the standard deviation. The last value is the number of samples.

$$= 0.364277274$$

The 99% confidence interval:

$$5 \pm 0.364 = [4.636, 5.364]$$

The length of the 99% confidence interval:

$$5.364 - 4.636 = 0.728$$

Conclusion

Below is a summary of the confidence intervals with their respective lengths:

90% Confidence Interval

- Interval: [4.767, 5.233]
- Length: 0.466

95% Confidence Interval

- Interval: [4.723, 5.277]
- Length: 0.554

99% Confidence Interval

- Interval: [4.636, 5.364]
- Length: 0.728

I can tell by looking at this data that the shortest interval is the 90% confidence interval, and the longest interval is the 99% confidence interval. There emerges a pattern here: the larger the confidence interval, the larger the standard error, and as a result, the length of the interval is larger on both sides of the sample mean.

Estimating λ in the Poisson Distribution

To estimate λ in the Poisson distribution, first look at the definition of the Poisson distribution.

$$Poisson(\lambda) = f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$$

This can be reworded to the probabilities of each x given λ:

$$Poisson(\lambda) = P(x_1, x_2, \ldots, x_n|\lambda)$$

For each x passed in, we want to multiply that by each other. We can use the product notation to denote every x passed in.

$$P(x_1, x_2, \ldots, x_n|\lambda) = \prod_{i=1}^{n} P(x_i|\lambda)$$

Next, we need to simplify the product on the right. We can begin by plugging in the probability mass function for P (x_i| λ)

$$\prod_{i=1}^{n} P(x_i|\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^x}{x!}$$

To further simplify this result, we can split the product notation into multiple product notations:

$$\prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^x}{x!} = \frac{\prod_{i=1}^{n} e^{-\lambda} \prod_{i=1}^{n} \lambda^{x_i}}{\prod_{i=1}^{n} x_i!}$$

By the properties of the product notation, we can simplify the numerator further. The product of e to the negative λ can be changed to e to the negative n times λ since we take this to the power of n. The right product of the numerator can be simplified to a summation in the power since the power contains the values being summed. The denominator will remain the same as it cannot be reduced any further.

$$\frac{\prod_{i=1}^{n} e^{-\lambda} \prod_{i=1}^{n} \lambda^{x_i}}{\prod_{i=1}^{n} x_i!} = \frac{e^{-\lambda n} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

Next, we want to take the log of both sides of the equation. This will reduce our exponent and powers.

$$ln(P(x_1, x_2, \ldots, x_n|\lambda)) = ln(\frac{e^{-\lambda n} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!})$$

Using natural log rules, we can simplify. The following are the rules used:

- $ln(a^b) = b * ln(a)$
- $ln(ab) = ln(a) + ln(b)$
- $ln(\frac{a}{b}) = ln(a) - ln(b)$

$$ln(\frac{e^{-\lambda n} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}) = -\lambda n + \sum_{i=1}^{n} x_i * ln(\lambda) - ln(\prod_{i=1}^{n} x_i!)$$

We can now take the derivative with respect to λ of our simplified equation. The derivative provides us with the estimator λ hat.

$$\frac{d}{d\lambda}(P(x_1, x_2, \ldots, x_n|\lambda)) = \frac{d}{d\lambda}(-\lambda n + \sum_{i=1}^{n} x_i * ln(\lambda) - ln(\prod_{i=1}^{n} x_i!))$$

To note: The summation on the right side of the equation is treated as a constant. We keep it in the result of the derivative since it is being multiplied by a λ.

Also, to note: The product on the right side of the equation has no λ and is also a constant – in this case, this product will result in a 0 when the derivative of it is taken.

We use the following derivative rules to solve for the derivative:

- $\frac{d}{da}(ab) = b$
- $\frac{d}{da}ln(a) = \frac{1}{a}$

$$\frac{d}{d\lambda}(-\lambda n + \sum_{i=1}^{n} x_i * ln(\lambda) - ln(\prod_{i=1}^{n} x_i!)) = -n + \sum_{i=1}^{n} x_i * \frac{1}{\lambda} - 0$$

Now that we have a simplified equation, we need to solve it for λ. We can set the previous result equal to zero.

$$-n + \sum_{i=1}^{n} x_i * \frac{1}{\lambda} = 0$$

Add positive n on both sides of the equation:

$$\sum_{i=1}^{n} x_i * \frac{1}{\lambda} = n$$

Multiply by λ on both sides:

$$\sum_{i=1}^{n} x_i = n\lambda$$

Divide by n on both sides to solve for λ:

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{n\lambda}{n}$$

Variable n will cancel on the right side, leaving us with:

$$\frac{\sum_{i=1}^{n} x_i}{n} = \lambda$$

Recall that the definition of $\bar{x}$ (the sample mean) is as follows:

$$\frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

We can say that the two are equal:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Proof

We need to prove that the natural log of λ is maximized when λ is equal to λ hat. We want to show that if we take the second derivative of the natural log of the equation, we produce a value less than 0. When the value is less than 0, the value is not valid meaning the previous derivative was the maximum likelihood estimator.

We take the second derivative with respect to λ of the equation by taking the derivative of the previous derivative:

The previous resulting derivative:

$$-n + \sum_{i=1}^{n} x_i * \frac{1}{\lambda}$$

The second derivative:

$$\frac{d}{d\lambda}(-n + \sum_{i=1}^{n} x_i * \frac{1}{\lambda})$$

The derivative of the constant negative n is 0. We once again take the summation as a constant with respect to λ, so the summation stays around again. Recall that the derivative of a positive fraction will produce a negative fraction. The result of the second derivative is:

$$= \sum_{i=1}^{n} x_i * -\frac{1}{\lambda^2}$$

Looking at these two values, we can put some thought into what values will be produced out of them. For the summation, we know that since there is no negative x or n, this will always produce a positive number:

$$= \sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n = \text{positive result}$$

Ignoring the negative sign in front of the right side, we know that the denominator will always produce a positive result for any positive or negative λ, since a negative or positive value squared will produce a positive number. Let's assume λ is 10:

$$\frac{1}{\lambda^2} = \frac{1}{10^2} = \text{positive result}$$

Now we can use these ideas and replace them in the actual result:

$$\sum_{i=1}^{n} x_i * -\frac{1}{\lambda^2} = (\text{positive result}) * -(\text{positive result})$$

Plugging in the number 1 as a 'positive result' (even though this value technically isn't possible due to the fraction – this is just to show what will result out of the right side).

$$\sum_{i=1}^{n} x_i * -\frac{1}{\lambda^2} = (1) * -(1) = -1$$

This is a negative number, which is not possible. This proves that the maximum likelihood estimator for the λ parameter must be λ hat, which is also known as the sample mean of n.

$$\text{Since } \frac{d^2}{d^2\lambda}(-\lambda n + \sum_{i=1}^{n} x_i * ln(\lambda) - ln(\prod_{i=1}^{n} x_i!) \text{ results in a negative value}$$

$$\text{our MLE for } \lambda \text{ must be } \frac{\sum_{i=1}^{n} x_i}{n}, \text{ which is } \bar{x}$$