In 1910 the physicists Ernest Rutherford and Hans Geiger conducted an experiment in which they recorded the number of alpha particles emitted from a polonium source (as detected by a Geiger counter) during 2608 time intervals each of length 1/8 minute.

Number of α-particles detected: j	Observed Frequency: f_j	Expected Frequency: e_j
0	57	54.31
1	203	210.28
2	383	407.06
3	525	525.31
4	532	508.44
5	408	393.69
6	273	254.03
7	139	140.50
8	45	67.99
9	27	29.25
10	10	11.32
11	4	3.99
12	0	1.29
13	1	0.38
≥ 14	1	0.14
Total	2608	2607.9

Table 2.3 Frequency table for Rutherford/Geiger data

$$\hat{e}_j = \frac{e^{-\hat{\theta}} \cdot \hat{\theta}^j}{j!}$$

We will determine whether the Poisson distribution is appropriate for this.

Note that for the Poisson distribution, the MLE is the sample mean. Thus:

$$\hat{\theta} = \frac{1}{2608} \sum_{j=0}^{14} j \cdot f_j$$

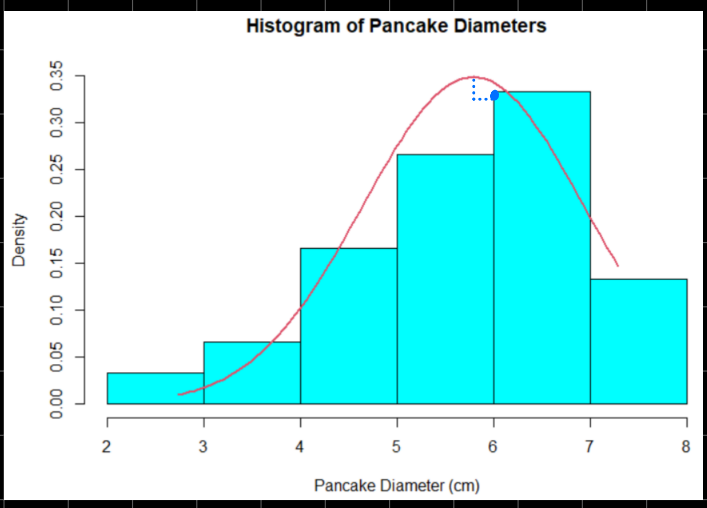
: avg. 3.87 particles per time interval

↳ # of intervals

Compute the expected number of intervals in which j particles are observed using $\hat{\theta}$:

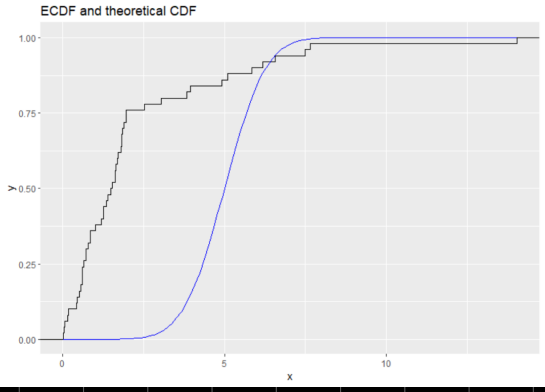
$$\hat{e}_j = \frac{1}{2608} \frac{3.8715^j e^{-3.8715}}{j!} \text{ for } j = 0, 1, \ldots$$

Graphical Checks



Here, the overlaid Gaussian/normal curve is slightly to the left of the top bin, signifying a slight left skew relative to the Gaussian distribution

Let's compare the ecdf of a sample of data generated from an exponential distribution against the cdf of a $G(5, 1)$:



Gaussian is to the right of the ecdf -> *right skew*

Q-Q plots

If the Gaussian model is appropriate for a dataset, its ECDF and the CDF of a Gaussian random variate should agree

We can determine this by calculating:

- Q(p) — theoretical quantile from (assumed) Gaussian distribution
- q(p) — sample quantile

$$Q\left(\frac{i}{n+1}\right) = q\left(\frac{i}{n+1}\right) \quad \forall\, 1 \le i \le n \rightarrow \checkmark$$

↳ $n+1$ because $q(1) = \infty$

Calculating theoretical quantiles for a Gaussian distribution requires fitting our data to the Gaussian, which requires us to know the mean and standard deviation of our data.

If these are not known, standardize to G(0,1)

- Mean is 0, standard deviation is 1

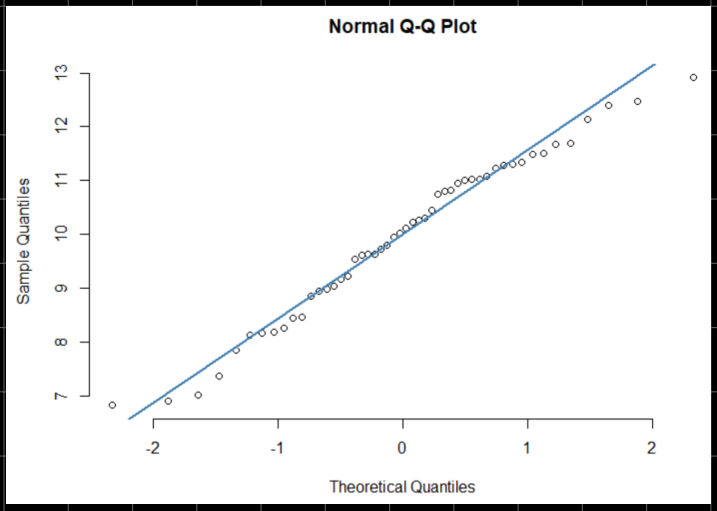
$$Q(p) = \mu + \sigma Q_z(p)$$

follows from $Y \rightarrow \frac{Y-\mu}{\sigma}$

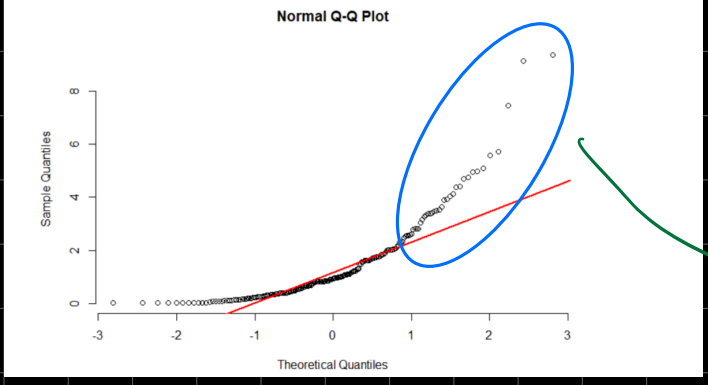
↳ quantile on $G(0, 1)$

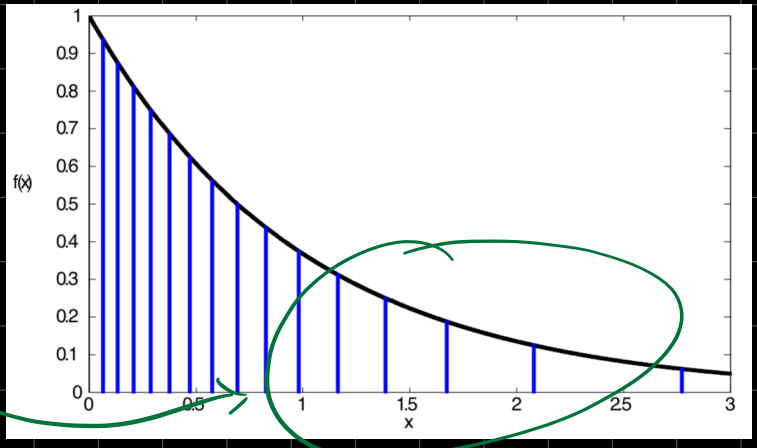
This is a *linear transformation*, so the shape is generally the same regardless of what the average and standard deviation are.

Then, the Qplot would just be plotting each (Q_z(p), q(p)):



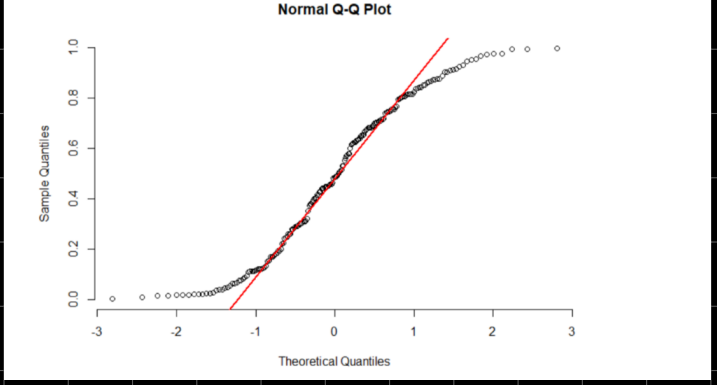
Question: what sort of distribution may show a U-shaped trend in points on a Qqplot?





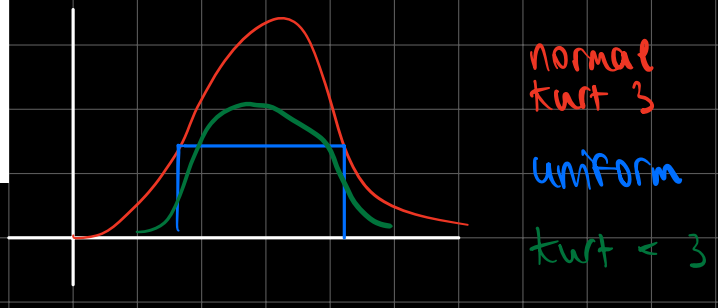
In an *exponential distribution*, events get less and less likely to happen further along in x. This is reflected in the above Q-Q plot.

Question: what sort of distribution may show an S-shaped trend in points on a Qqplot?

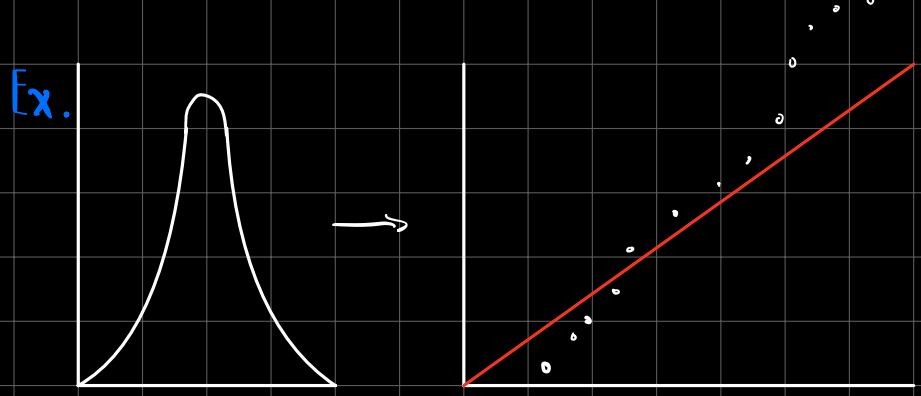


This is a uniform distribution.

- At the start and end, it deviates significantly from the Gaussian because the Gaussian has less samples near the ends
- Weaker tails -> smaller kurtosis



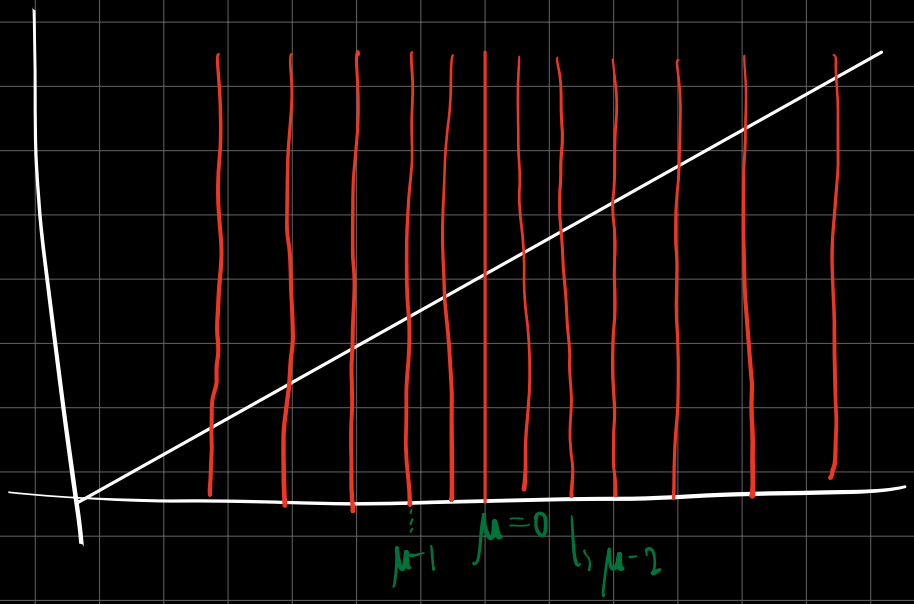
The *binomial model* cannot be checked using numerical or graphical summaries



Proper how-to:

On a Q-Q plot, each datapoint is represented as its own quantile.

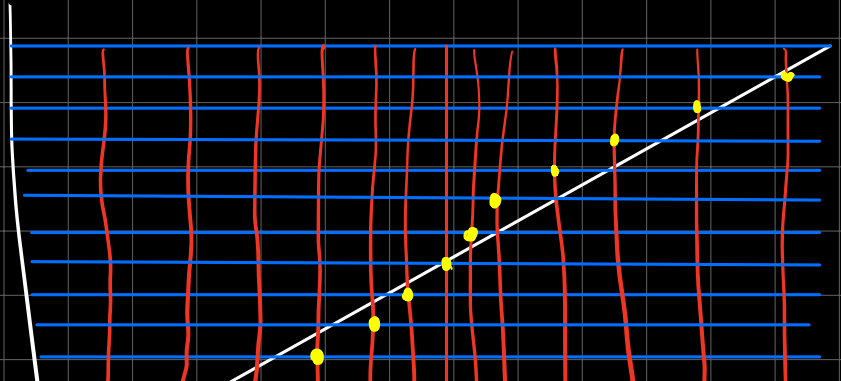
Then, suppose we want to represent each datapoint (quantile) in a *Gaussian distribution* as vertical lines around the mean. This looks something like

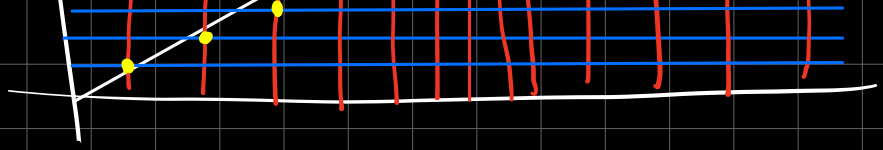


Where datapoints get more spread out the further they get from the mean.

We want each space between lines to store the same amount of data (none)

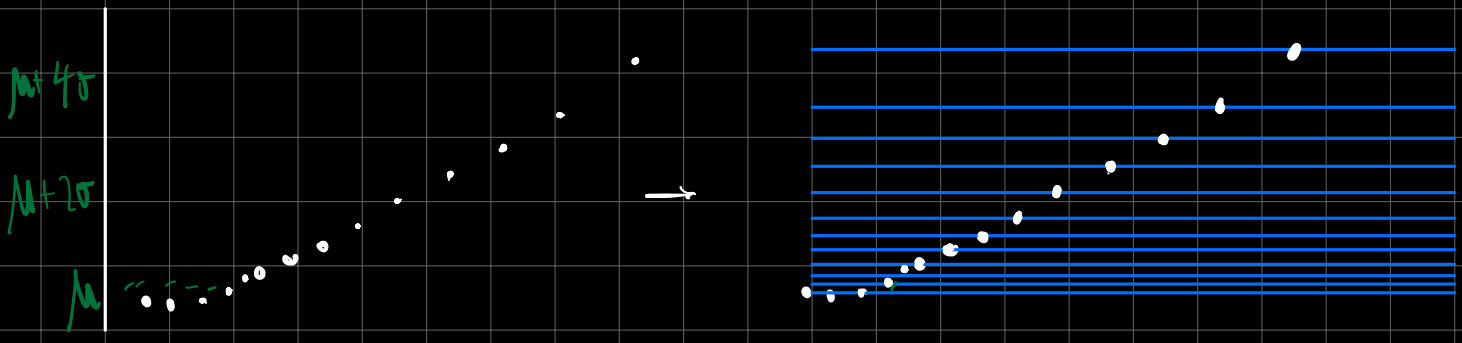
Next, consider a discrete uniform distribution (in blue), which we will represent as horizontal lines. These lines should be evenly spaced. The intersection between the red and blue lines gives us the points of a Q-Q plot.





As mentioned in the shitty notes above, this is an S-shape. Generally, S-shape = uniform.

<u>Most importantly:</u> the best way to tell what kind of distribution we are seeing is to check the *vertical spread of datapoints*. For example:



Here, datapoints occur more frequently near the (vertical) mean, and consistently become less and less common as we move further away. This signifies an *exponential* distribution.

On the other hand, for the above discrete uniform distribution, points seem distributed pretty evenly along the vertical axis.

Vertical spread is usually *the only thing you should worry about*, unless you somehow grew a brain. Everything else is too complicated