Lecture 5: Summary Statistics, Normal Distribution, Hypothesis Testing

What will be covered in this lecture?

1. **Methods for Visualise Data Distributions:**

   Histogram, Probability Density Function (**pdf**), and Cumulative Density Function (**cdf**)

2. **Summary Statistics:** Mean, Variance, Skewness, and Kurtosis

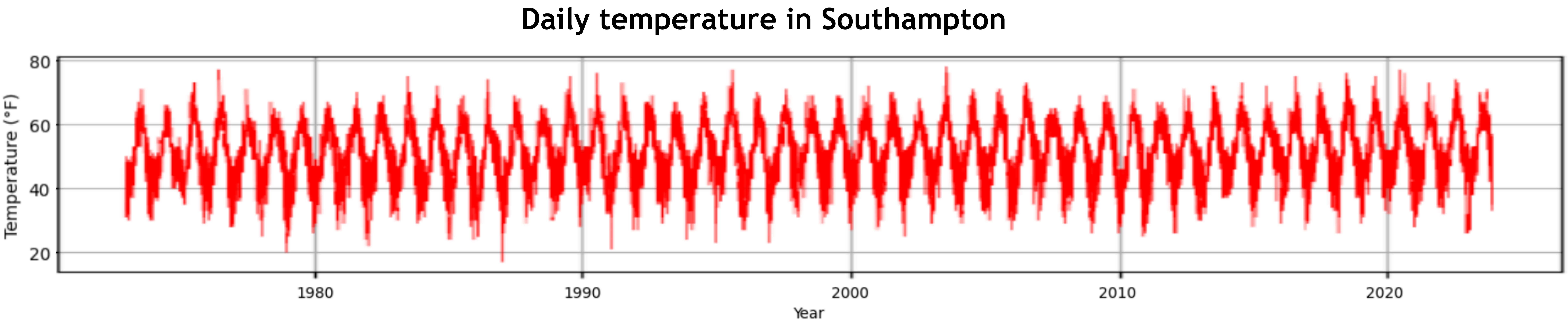
3. **Key Statistical Distributions**

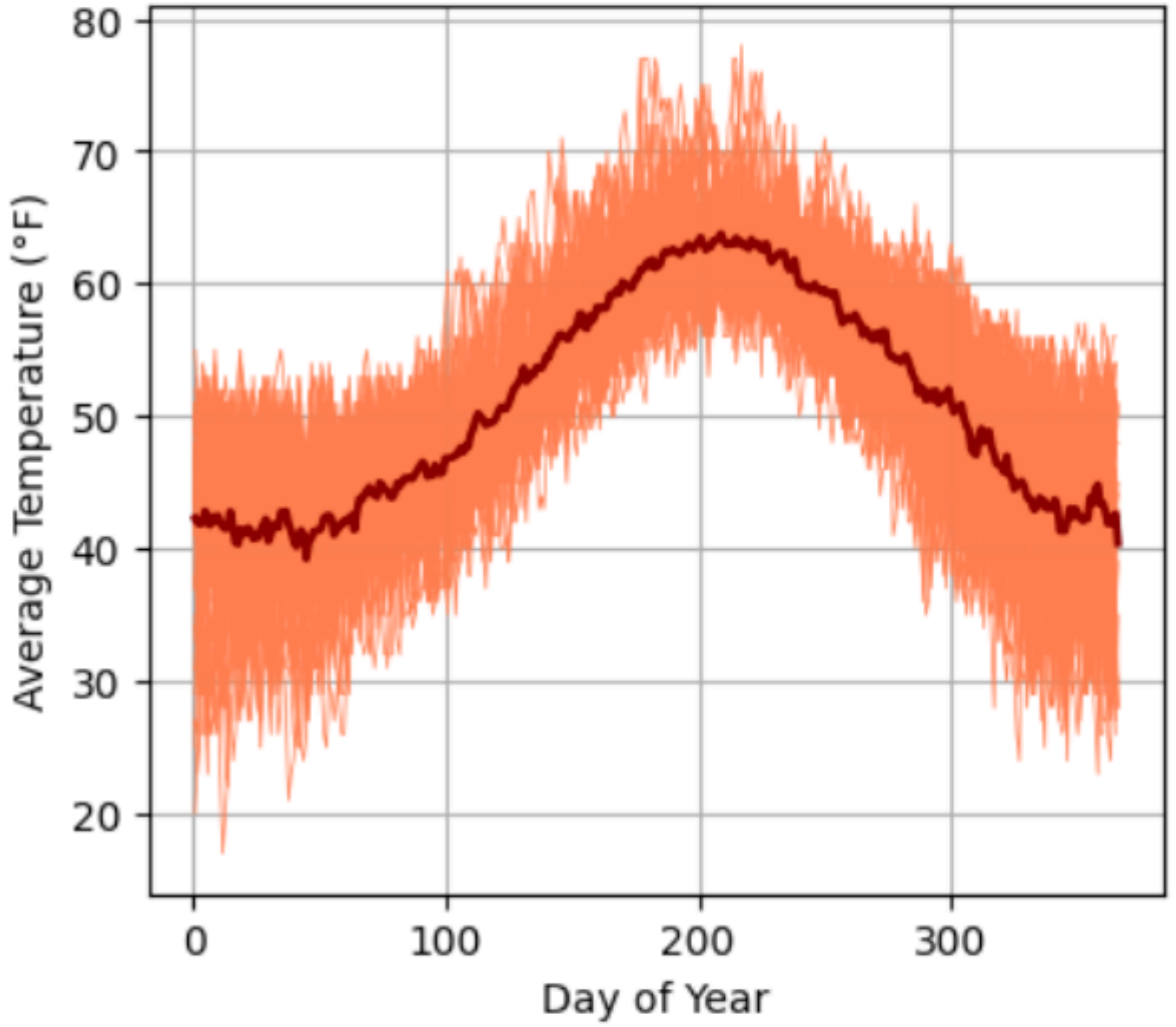
   3.1 Normal Distribution

   3.2 Chi-square (x2) distribution

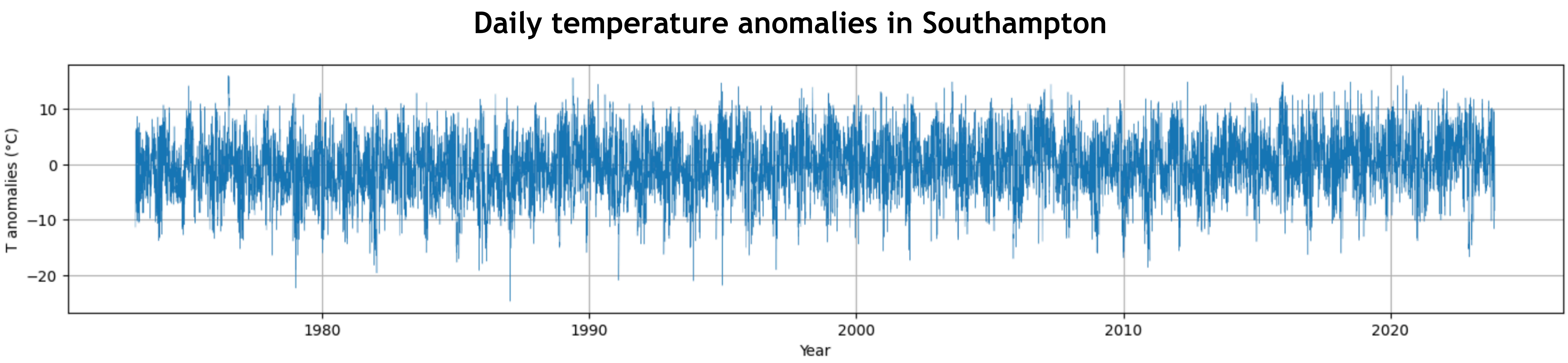
4. **Hypothesis Testing:** Testing if Your Data Fits a Normal Distribution

A dataset for Southampton's daily temperature





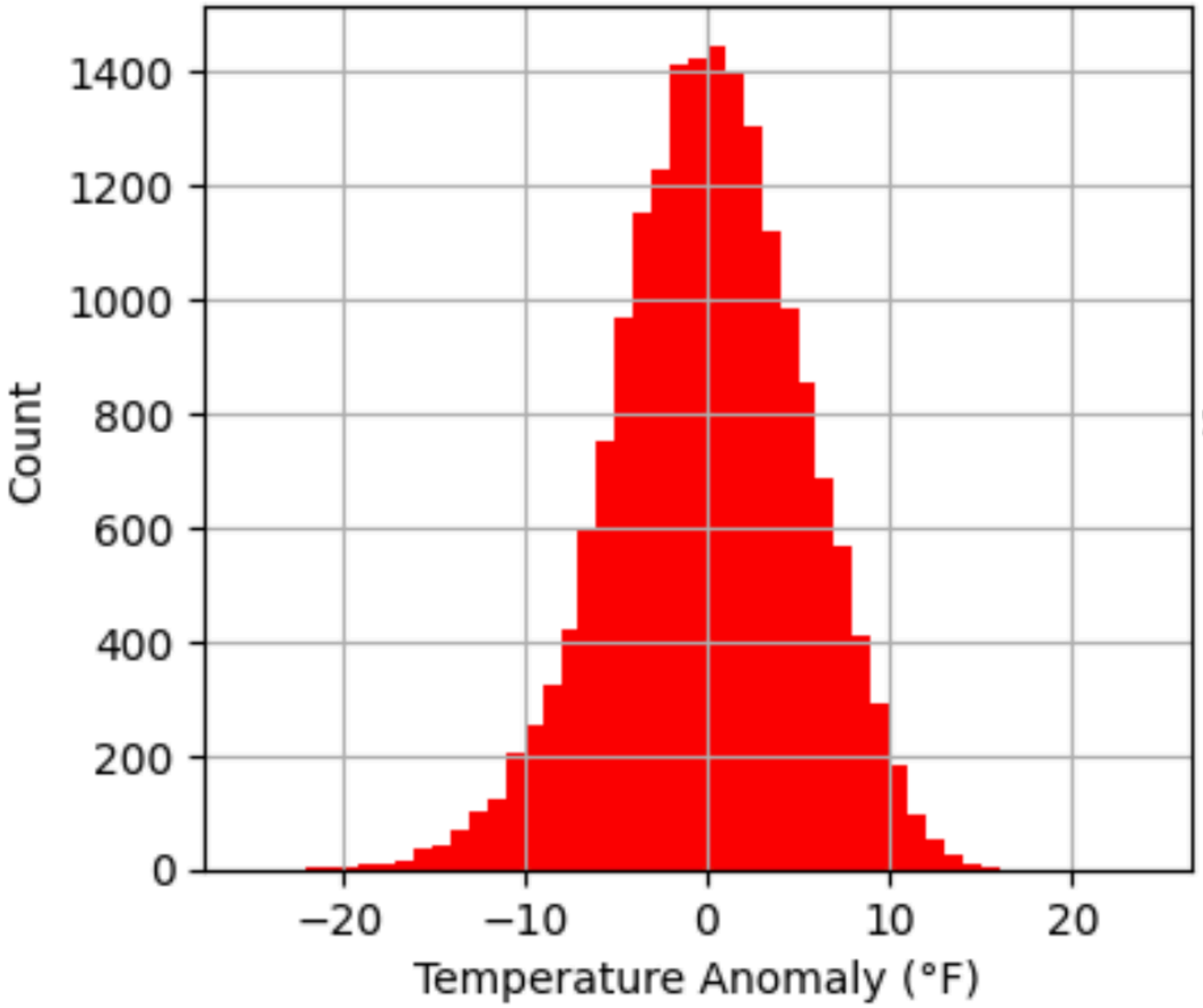
How will you describe this data set?



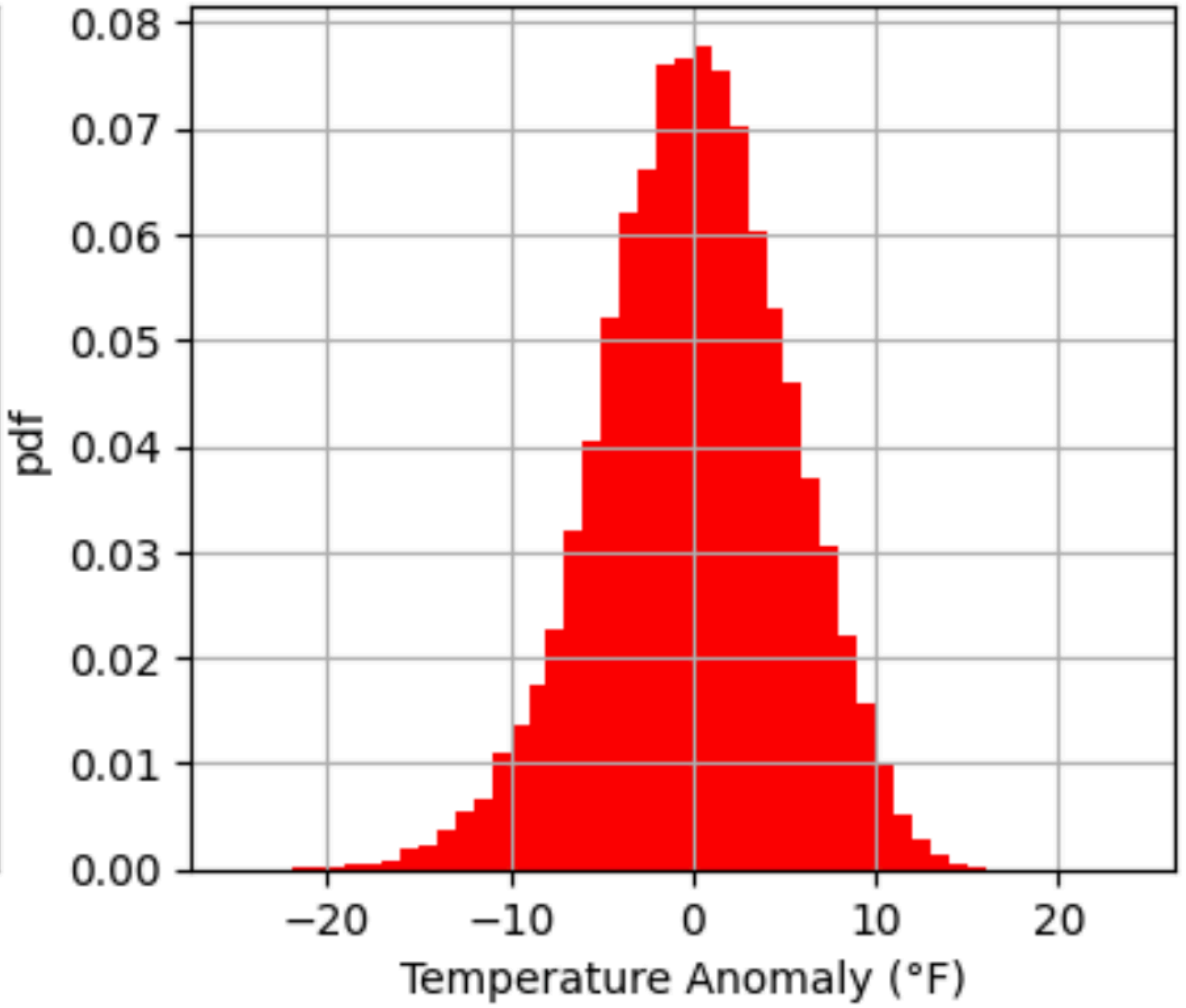
Methods for Visualise Data Distributions

Daily temperature anomalies in Southampton

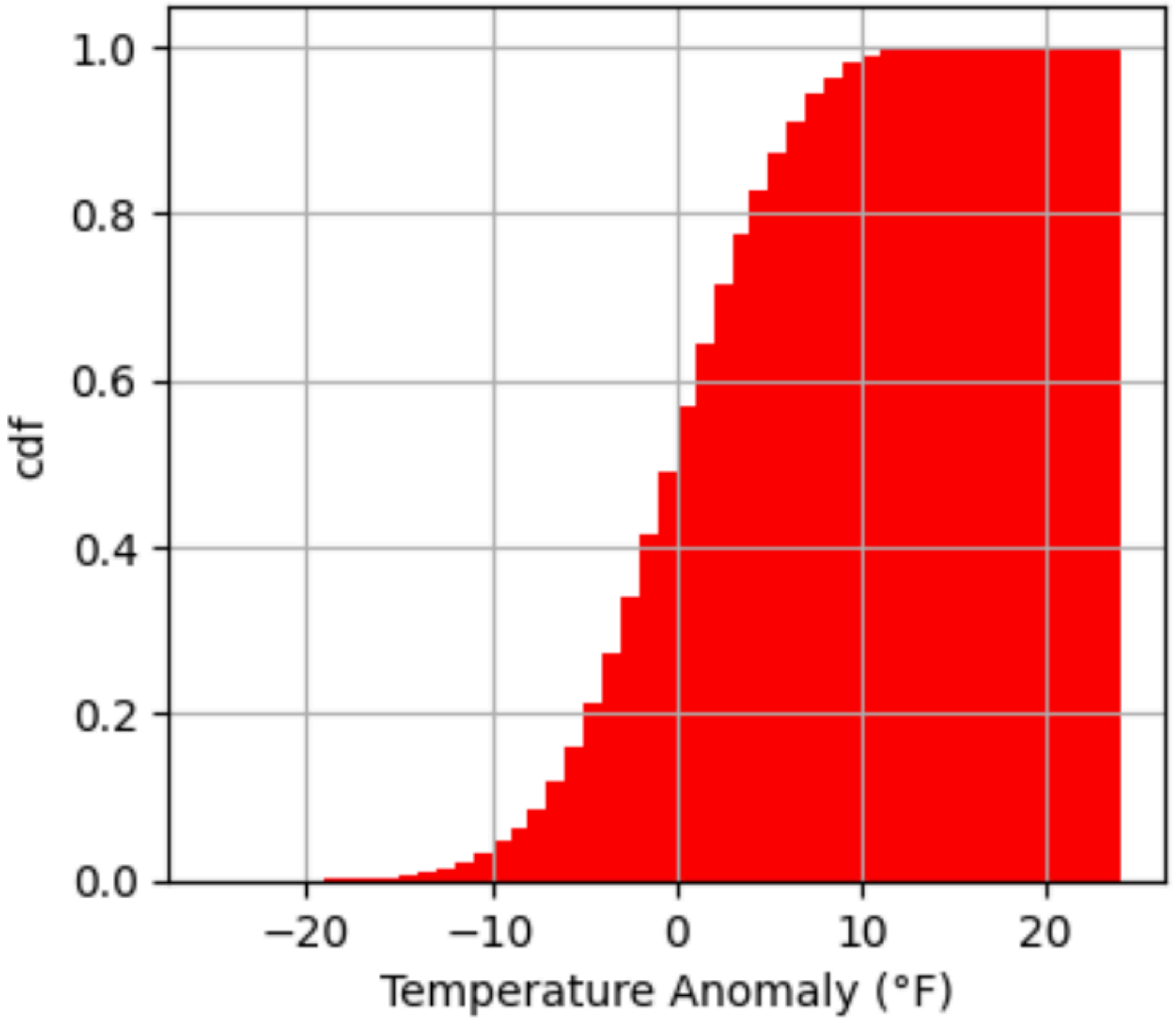
Histogram



Probability Density Function



Cumulative Density Function



Area under the curve is 1

The area under pdf before a value

plt.hist()

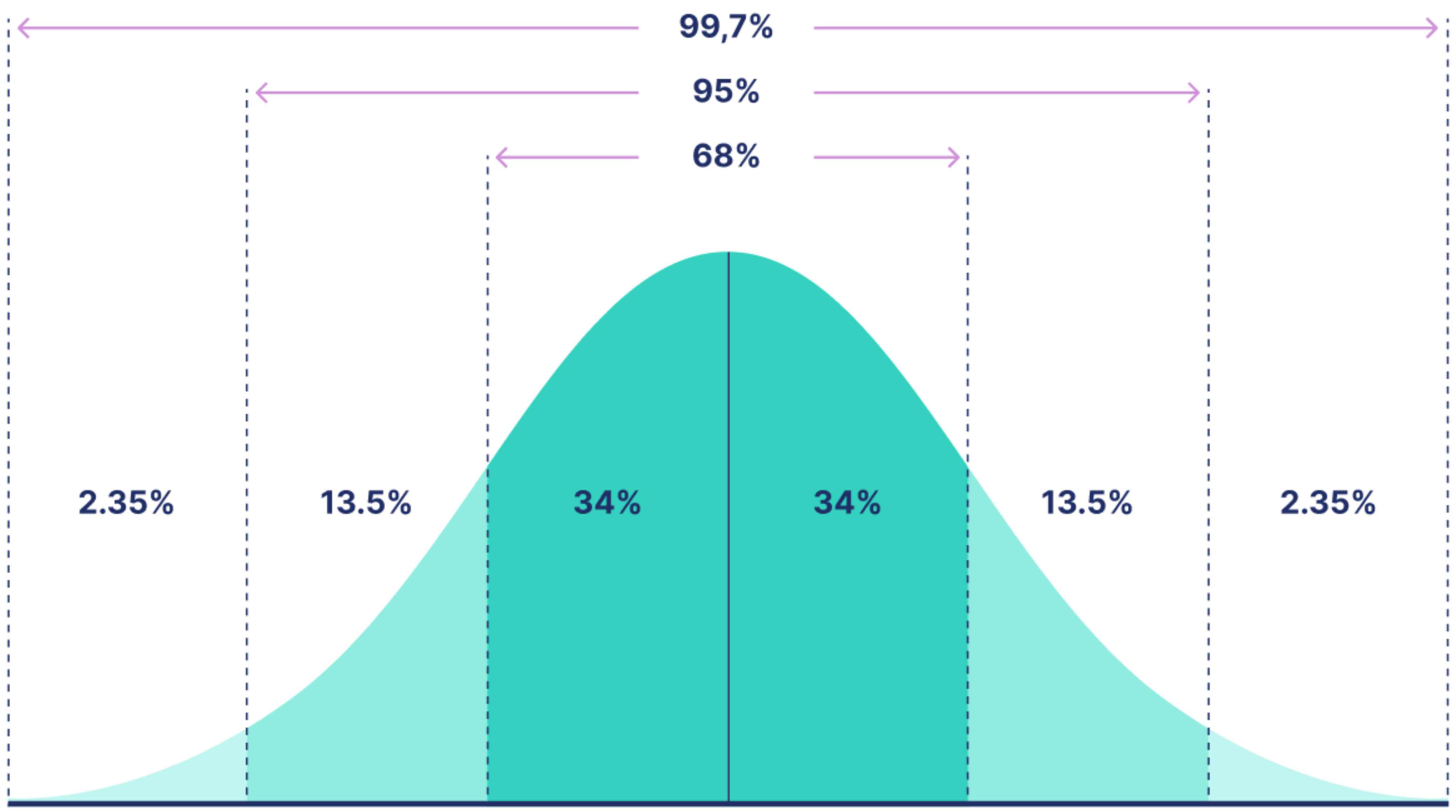
Summary Statistics: Understanding Your Data Better

Statistics	What it is	How to calculate it from data	Functions to use
Mean	The mean is just the average. It gives you an idea of the central point of your data	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$	numpy.mean(data) **or** dataframe['column'].mean()
Variance	Variance measures how spread out your data is around the mean. A bigger variance means more spread out data.	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$	numpy.var(data) **or** dataframe['column'].var()
Skewness	Skewness indicates whether your data leans more to the left or the right of the mean.	$S = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{X}}{s}\right)^3$	numpy.skew(data) **or** dataframe['column'].skew()
Kurtosis	Kurtosis tells you about the "tailedness" of your data distribution. High kurtosis means more data in the tails.	$K = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{X}}{s}\right)^4 - 3$	numpy.kurt(data) **or** dataframe['column'].kurtosis()

Standard deviation (*s*), defined as the square root of variance, also charecterizes variability around the mean. **It's in the same units as your data**, making it easier to understand the spread.

np.std()

Normal distribution



$$P(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}]$$

1. A theoretical distribution, pdf is "bell like"

2. Determined by only two parameters:

   μ: mean

   σ: standard error

3. Skewness and Excess Kurtosis are both zero

4. Standard form with zero mean and unit variance

$$Z = \frac{X-\mu}{\sigma} = \frac{X-\bar{X}}{s}$$

Z ~ N(0, 1)

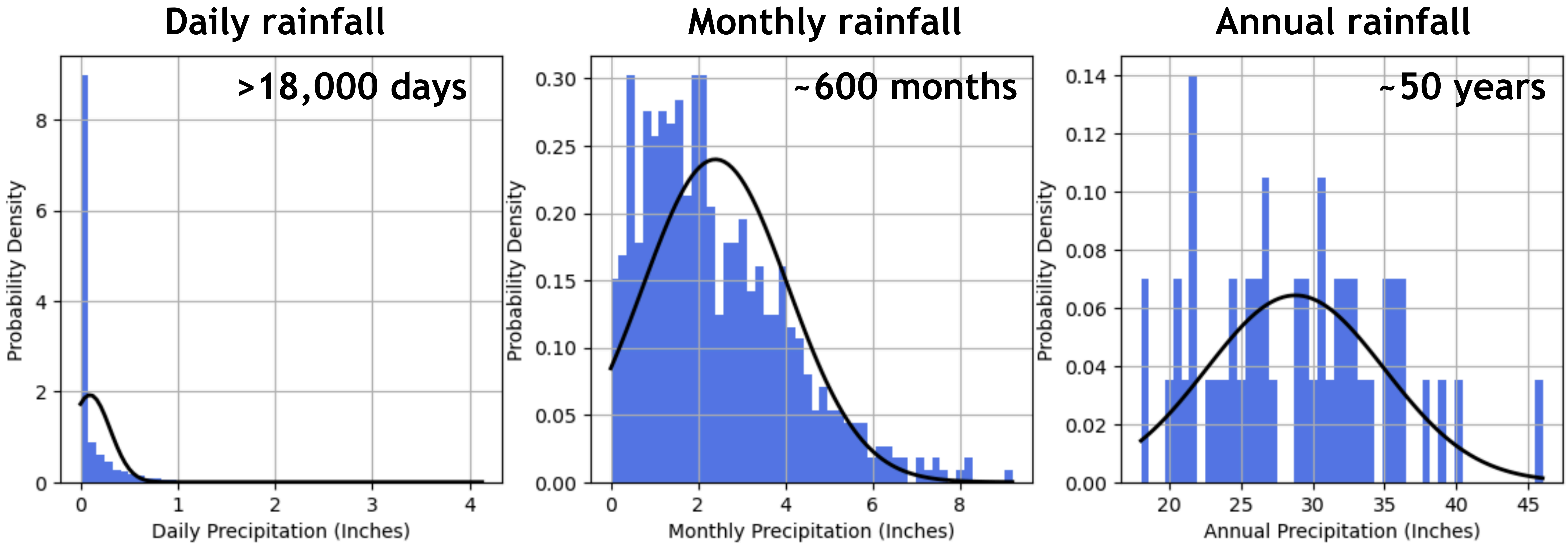
Central Limit Theorem (CLT)

scipy.stats.norm.pdf()

If you take lots of independent samples from any same distribution, the averages of those samples tend to form a Normal distribution, even if the original data was not normally distributed.

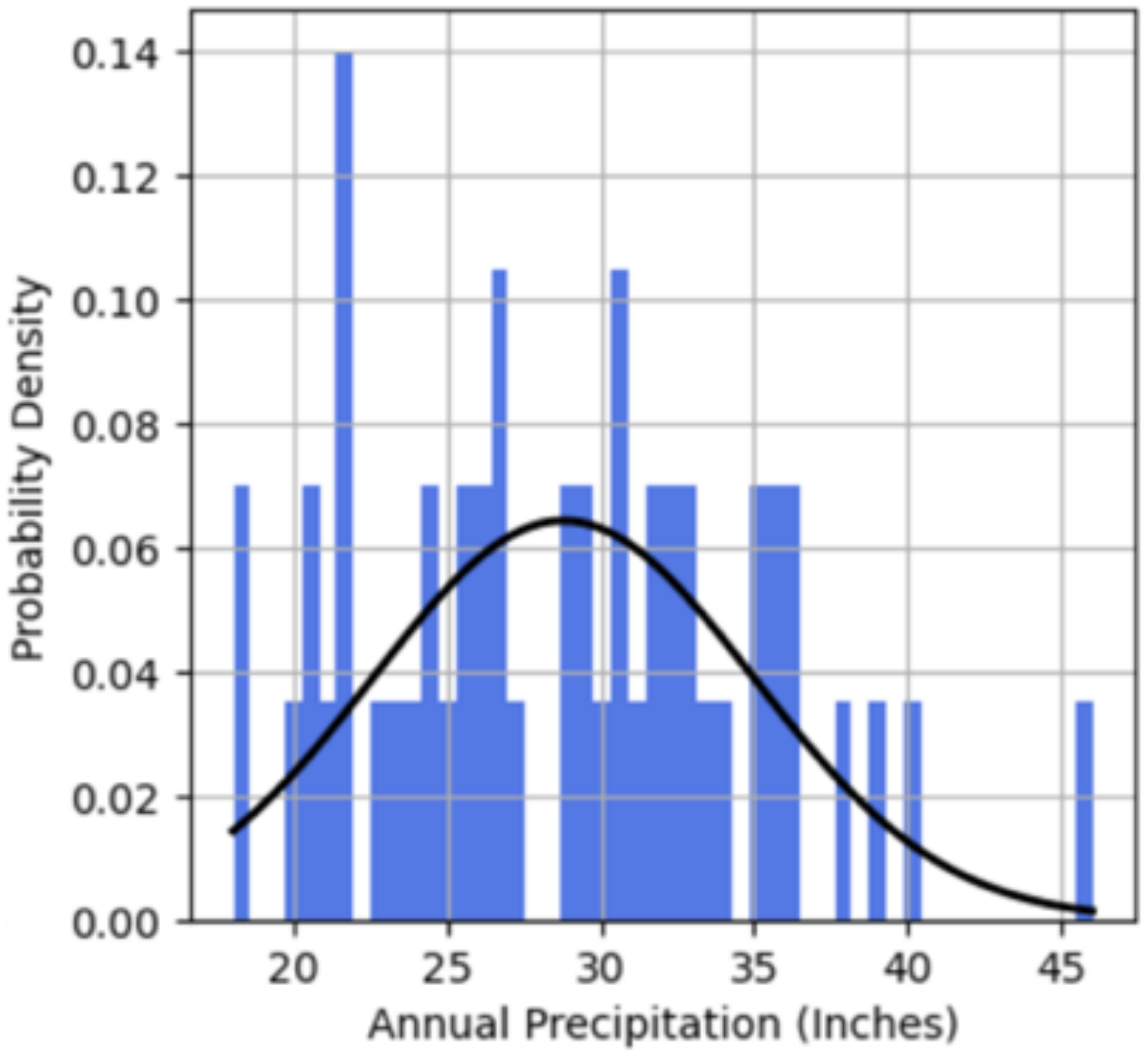
Normal distribution is the distribution of mean estimate.

iid :: Independent and identical

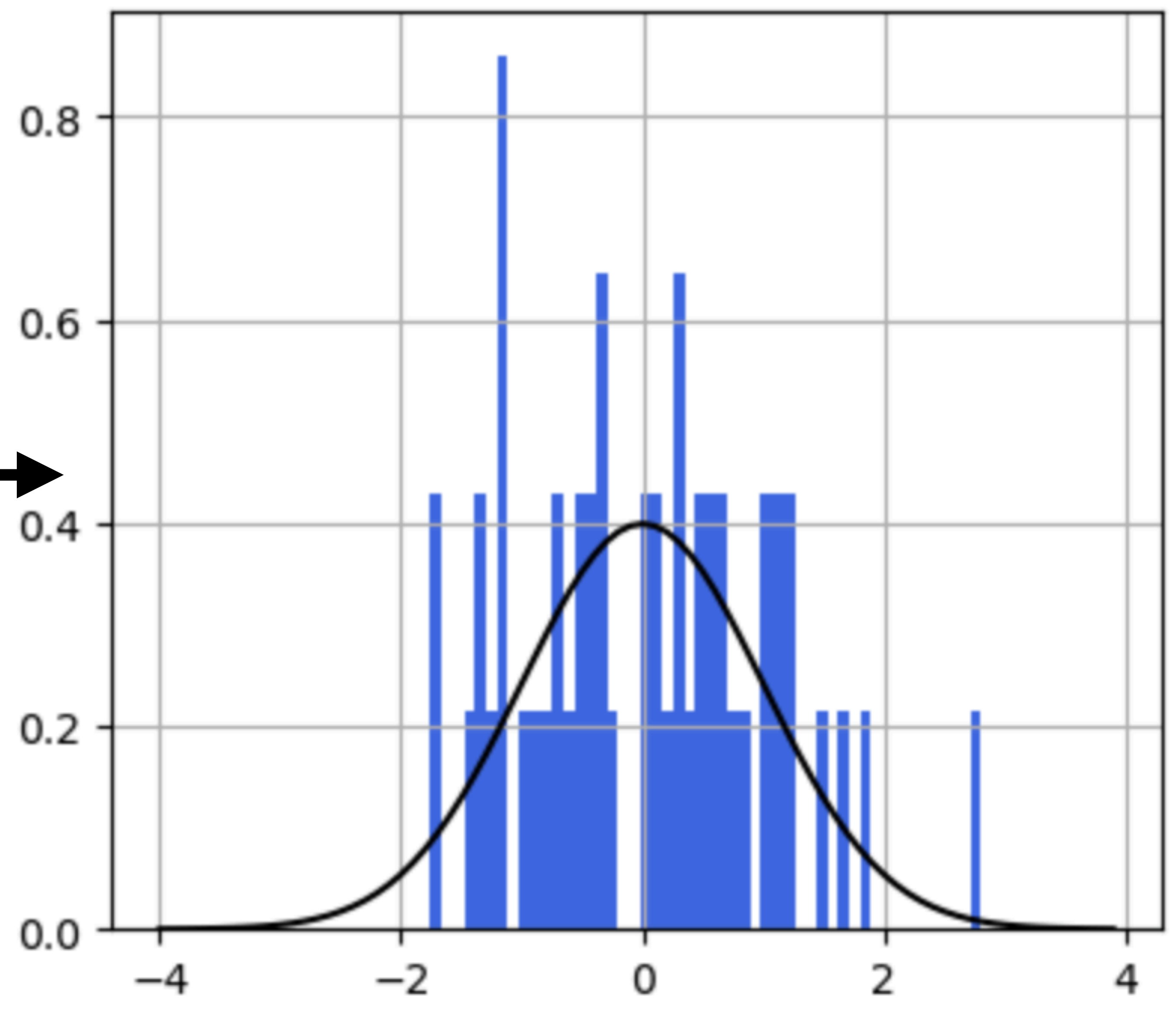


Does annual rainfall in Southampton follow a Gaussian distribution?

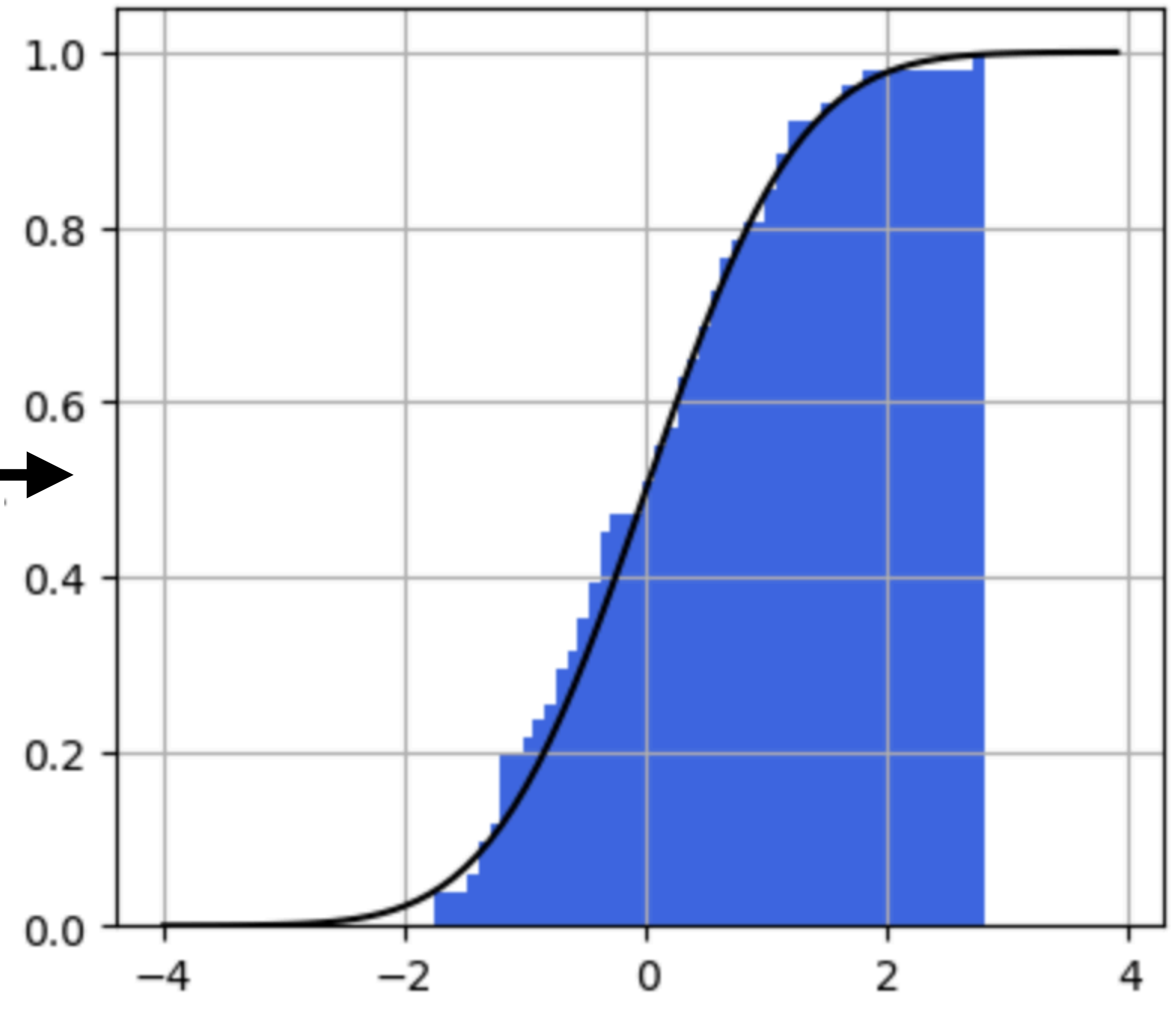
Raw Data



Standardised Data



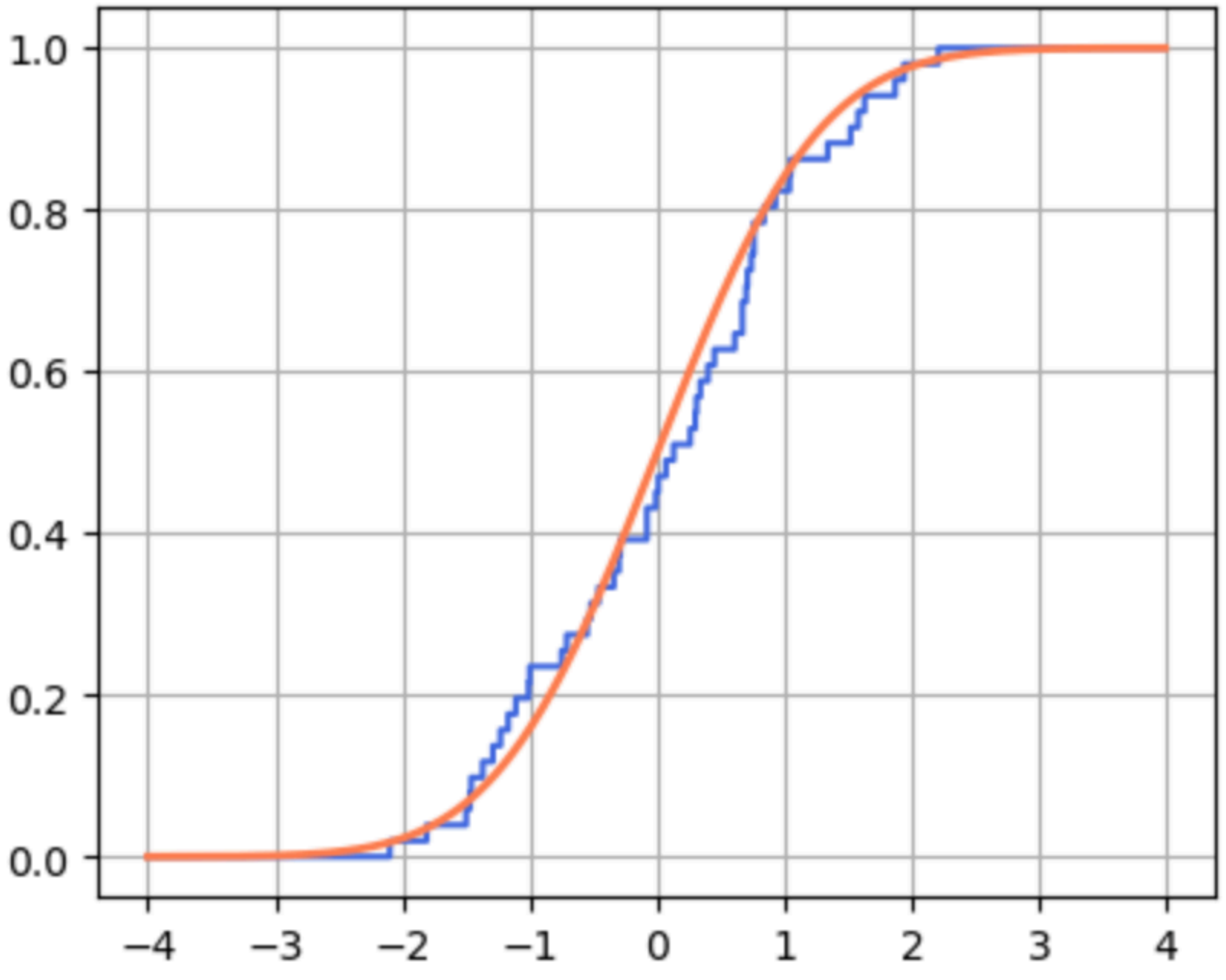
Cumulative Density Function



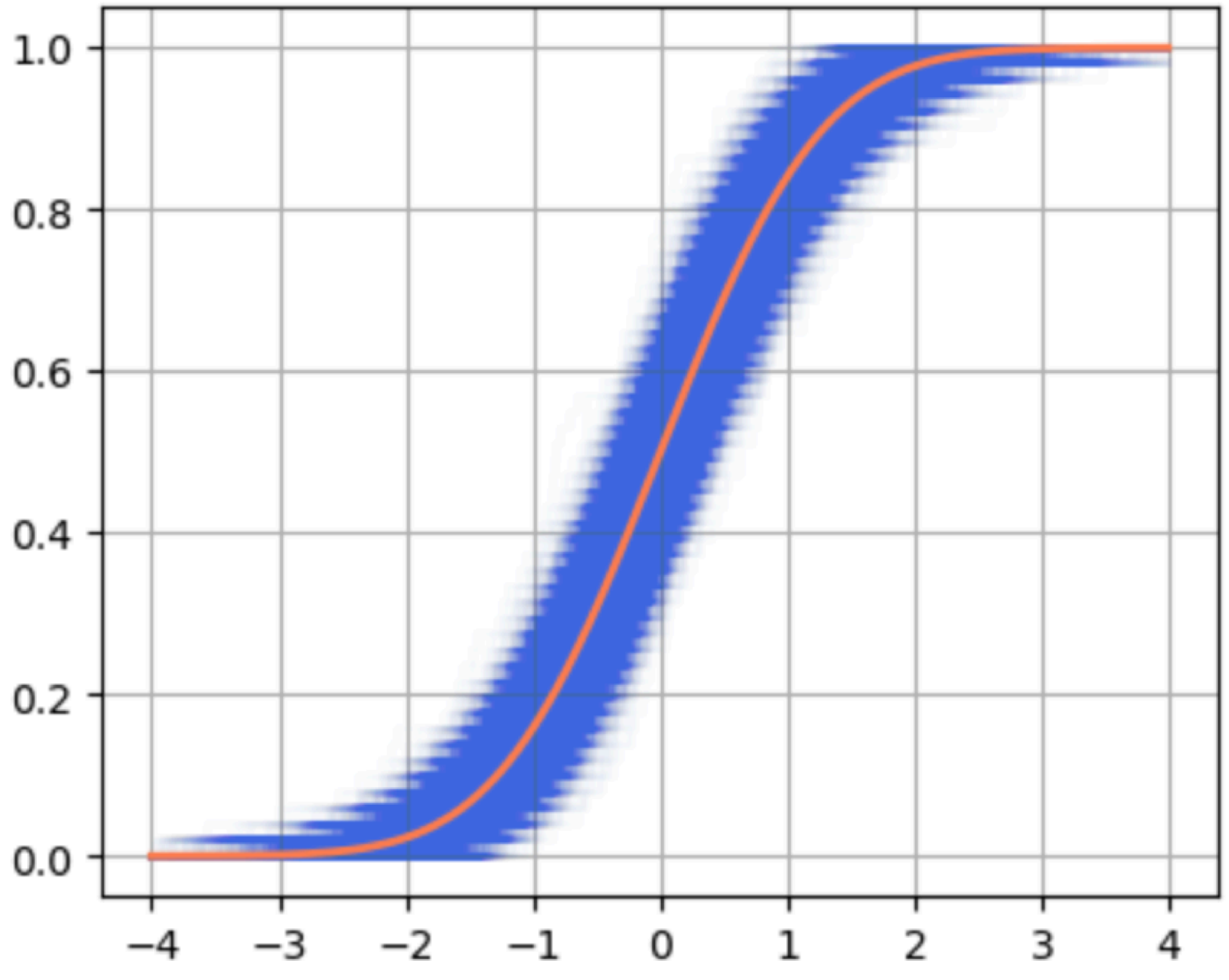
The CDF seem to align quite closely with the theoretical curve of the Standard Normal Distribution, although some discrepancies exist.

Realisations and Theoretical Curves

Realise from a standard Normal



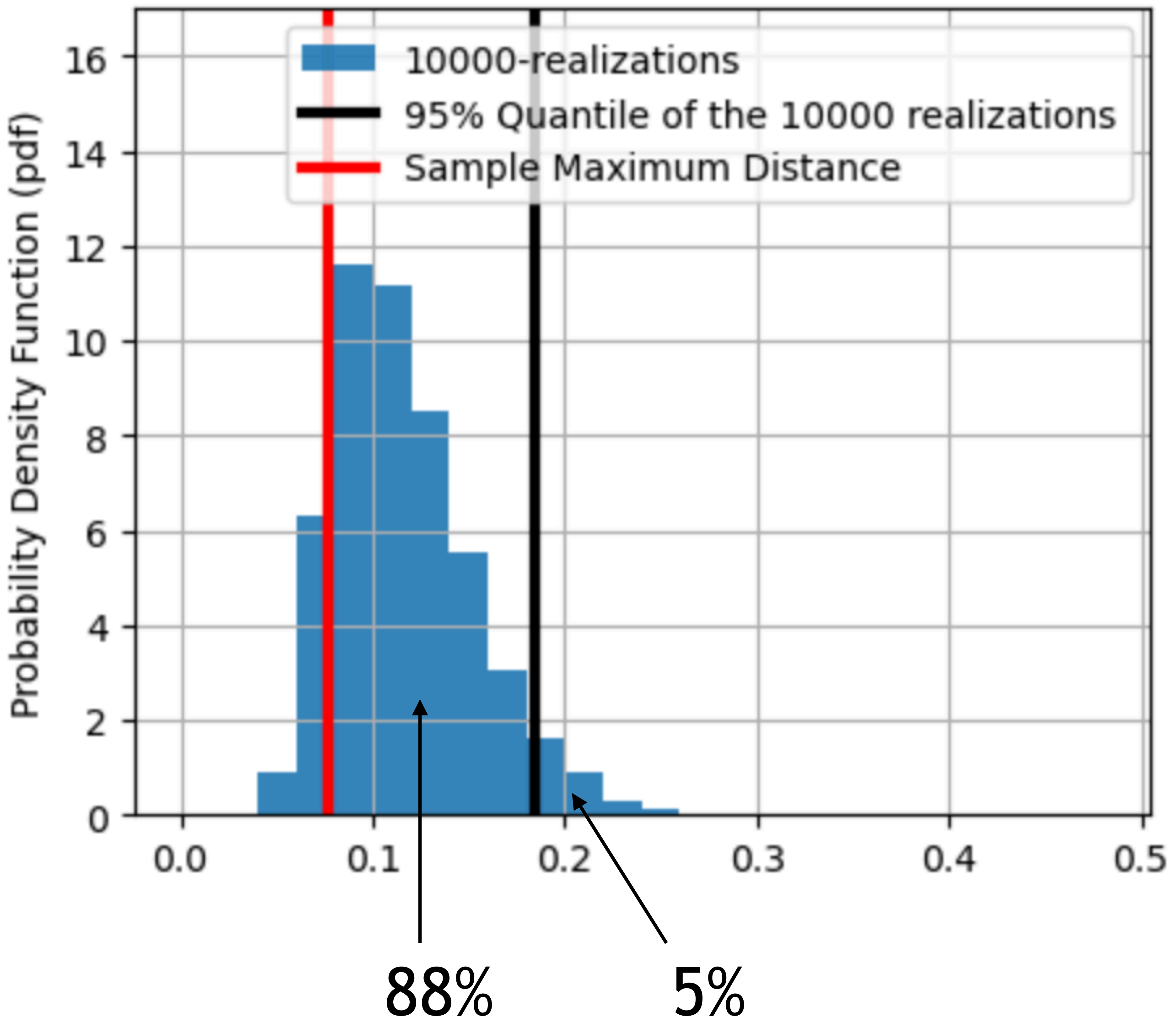
10,000 realisations



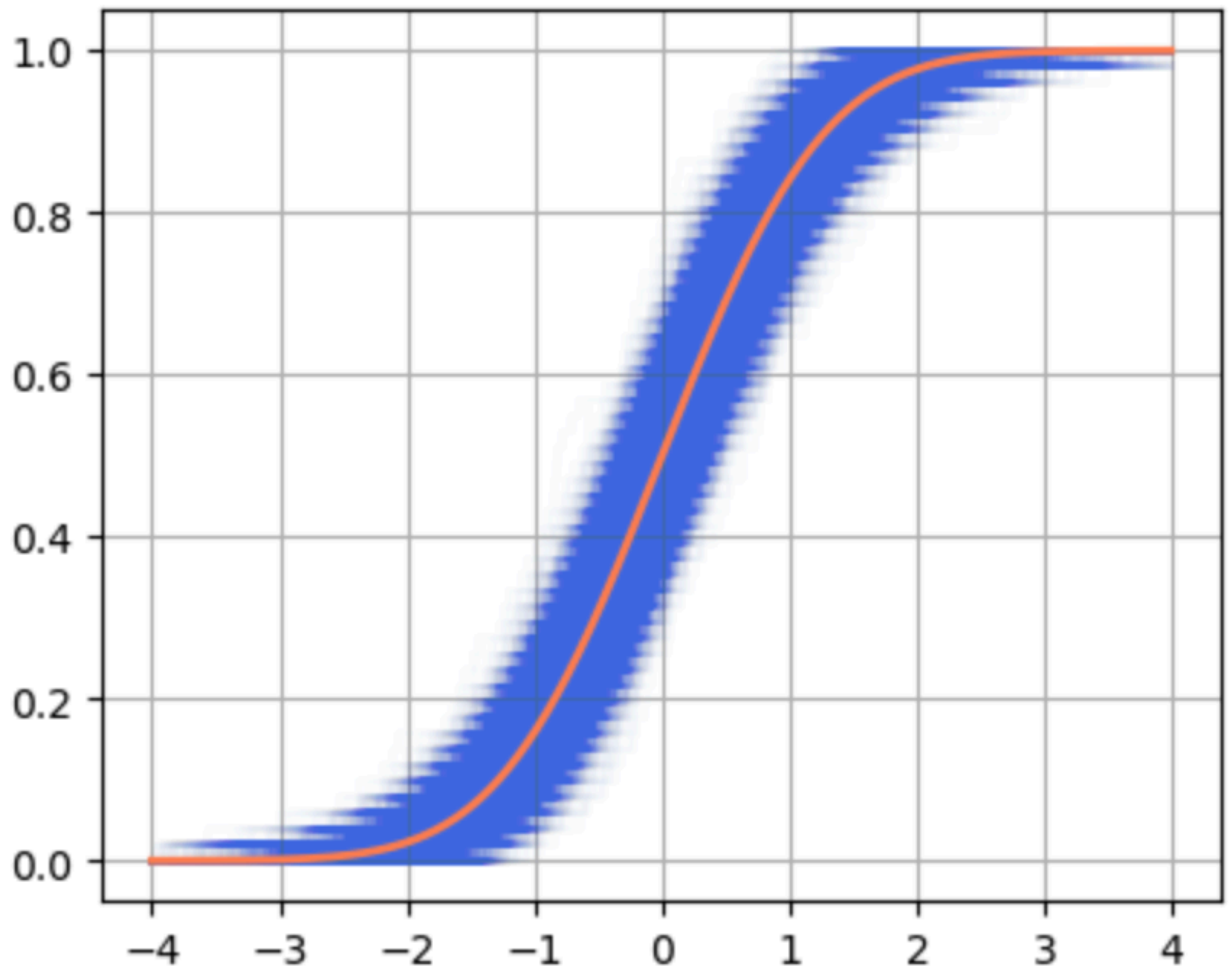
Deviation can arise due to randomness.

np.random.normal()

Comparing data against a null distribution

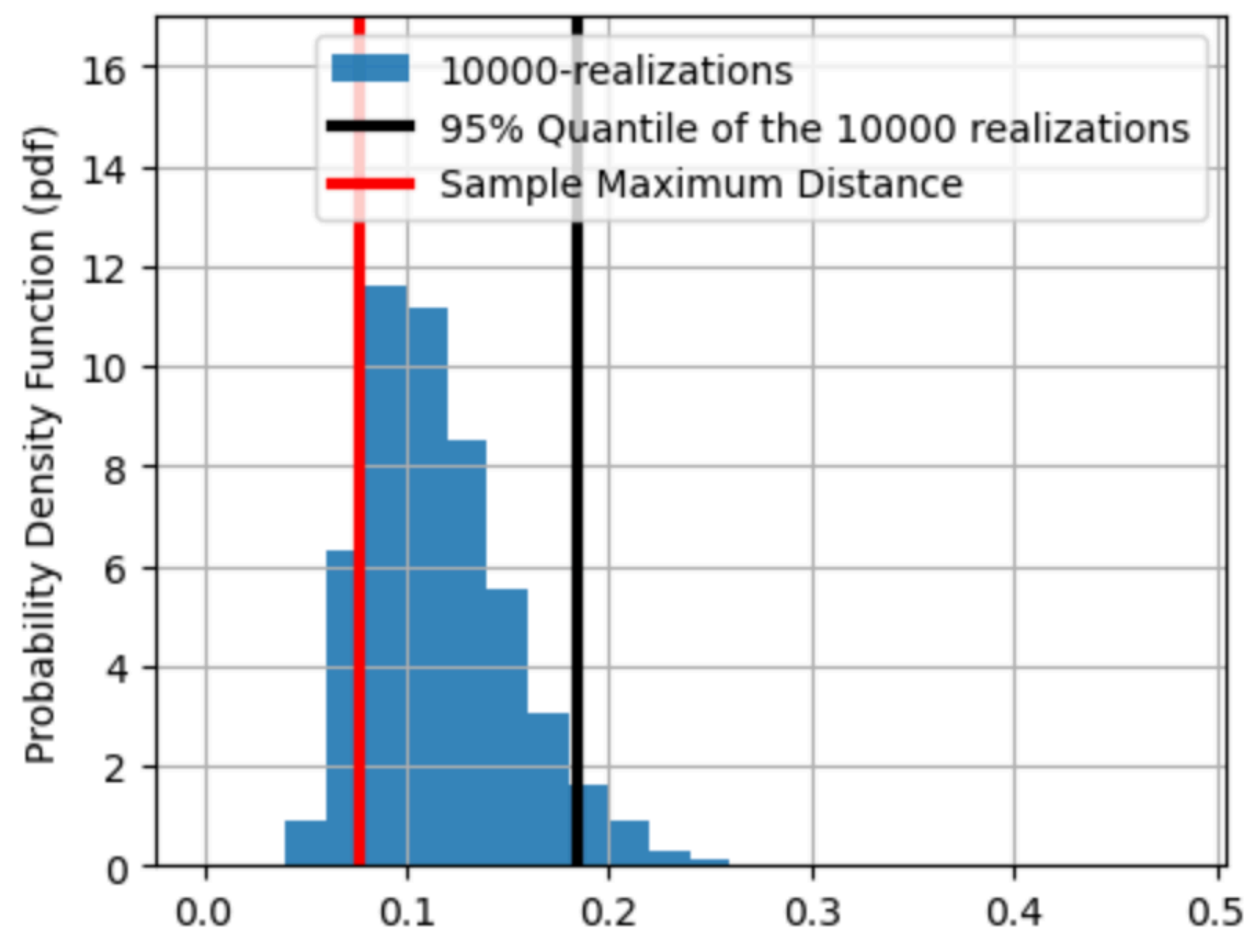


10,000 realisations



Hypothesis testing

Terminology	Meaning	When testing if annual precipitation follows a standard normal distribution
Null hypothesis (H_0)	This is our starting assumption that the effect being studied does not exist	Data follow a standard Normal distribution.
Alternative hypothesis (H_1)	This is what we might believe to be true if we find sufficient evidence against the null hypothesis.	Data do not follow a standard Normal distribution.
Test statistics	This is a calculated value from our data that we use to test our hypothesis.	Maximum distance from the theoratical CDF curve.
Null distribution	This represents what we would expect to see from our test statistic purely by chance if the null hypothesis were true.	We visualized this through the distribution of maximum distances from 10,000 random realizations.
significance level (α)	This is a threshold we set to decide when to reject the null hypothesis.	A common choice is $\alpha = 0.05$.
p-value	The probability of obtaining our data, or something more extreme, if the null hypothesis is true.	In our analysis, a p-value of 0.88 indicates that 88% of the null distribution is more extreme than what we observed in our data.



When $p < \alpha$, we can reject H_0, and accept H_1.

What we just went through is exactly a classic statistical test, called **Kolmogorov-Smirnov test**, we can call functions to implement this test.

scipy.stats.kstest()

Road Map of the Statistics Part

Lecture 5

Quantification Technique	Mean, variance, skewness, & kurtosis
Uncertainty & Significance	Gaussian distribution Chi-2 distribution
Assumptions	Data is Gaussian or follows specific types of distribution Independent Sampling
Test assumptions	K-S test
Treatment	

Lecture/Practical 13

Remove mean / seasonal cycle...

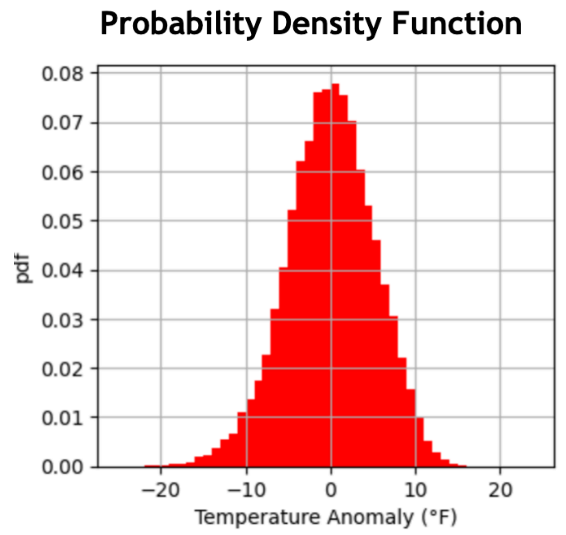
Plot and visualise your data

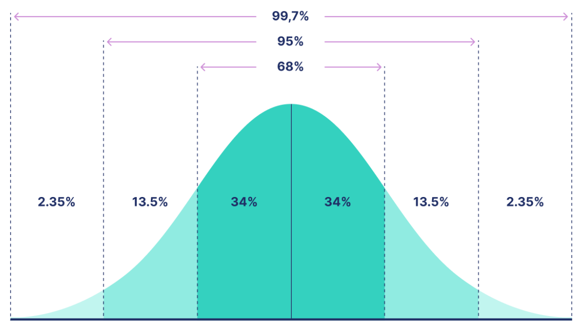
Calculate Summary Statistics

Is data Gaussian or Normally distributed?

K-S test

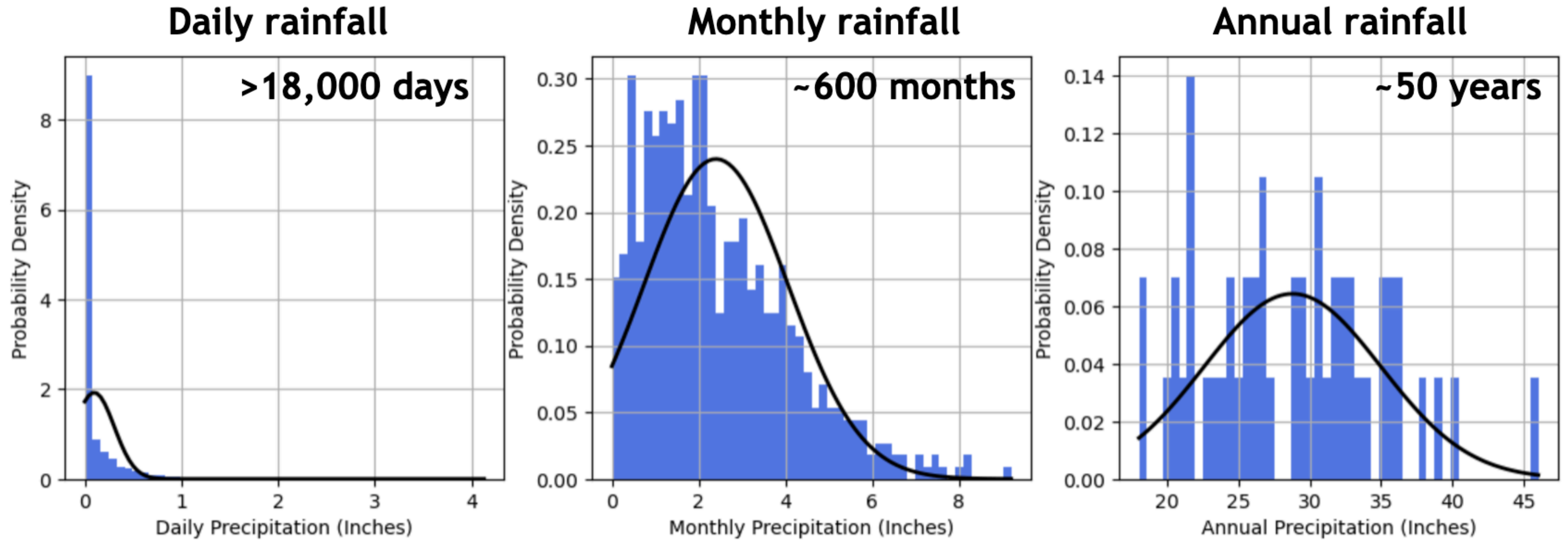
p-value?

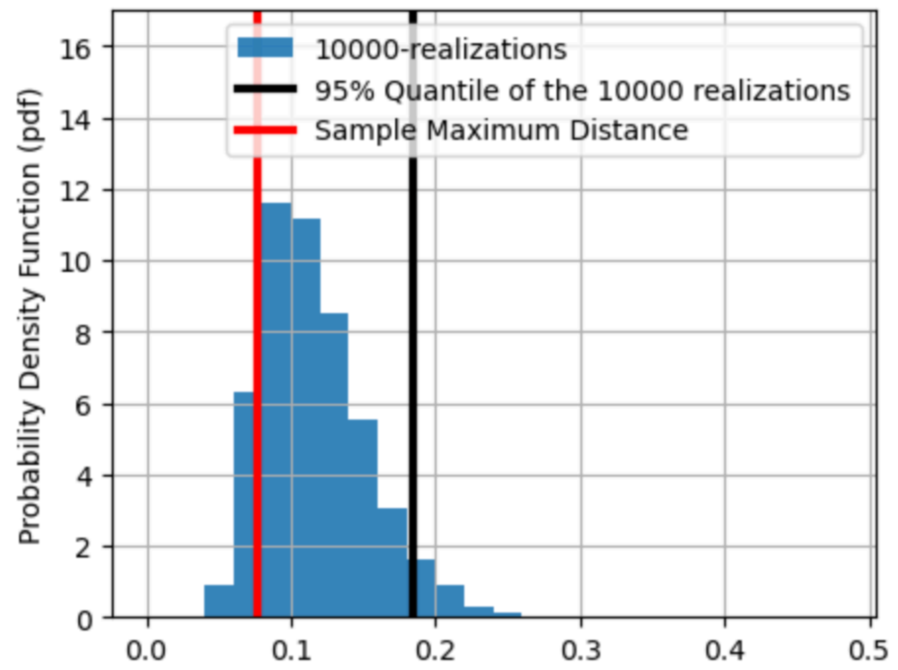




Skewness of 0
Kurtosis of 3

Central Limit Theorem:
Likely Gaussian when average or sum over >30 independent samples (such as monthly or annual mean values).





Terminology	Meaning
Null hypothesis (H_0)	This is our starting assumption that the effect being studied does not exist
Alternative hypothesis (H_1)	This is what we might believe to be true if we find sufficient evidence against the null hypothesis.
Test statistics	This is a calculated value from our data that we use to test our hypothesis.
Null distribution	This represents what we would expect to see from our test statistic purely by chance if the null hypothesis were true.
significance level (α)	This is a threshold we set to decide when to reject the null hypothesis.
p-value	The probability of obtaining our data, or something more extreme, if the null hypothesis is true.