In [1]:
```
from scipy.stats import norm, binom
from scipy import stats
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Q1.

Random samples of size n were taken from populations given below. Given your knowledge of the underlying distributions, determine the means and standard deviations of the sampling distributions

a. Sample is from an unknown distribution with n = 42, μ = 15, σ^2 = 9, statistic is X

$$Statictics(\bar{X}):$$
$$\mu(\bar{X}) = 15,$$
$$\sigma(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{9}{42}} = .4629$$

In [2]:
```
np.sqrt(9/42)
```

Out[2]: 0.4629100498862757

b. Sample is from a Normal distribution with n = 11, μ = 120, σ^2 = 10, statistic is X

$$Statictics(\bar{X}):$$
$$\mu(\bar{X}) = 120,$$
$$\sigma(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{10}{11}} = 0.9534$$

In [3]:
```
np.sqrt(10/11)
```

Out[3]: 0.9534625892455924

c. Sample is from a Binomial distribution with n = 50, p = 0.50, statistic is $\hat{p}$

$$\hat{P} = P = 0.50$$

$$\sigma(\hat{P}) = \sqrt{\frac{P*(1-P)}{n}} = \sqrt{\frac{.5*.5}{50}} = \sqrt{.005} = 0.0707$$

d. Sample is from a Poisson distribution with n = 35, λ = 30, statistic is X (we didn't explicitly talk about this case, but use what we know about sampling distributions of averages to solve for the Poisson distribution).

$\bar{X}$ statistics are going to shape a normal distribution with same μ:

$$E(\bar{X}) = E(X) = \lambda = 30$$

$$\sigma(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{\lambda}{n}} = \sqrt{\frac{30}{35}} = 0.9258$$

Q2

Suppose a random sample of size 18 observations is selected from a population that is normally distributed with mean 56 and standard deviation equal to 7.

a. Give the mean and standard deviation of the sampling distribution of the sample mean

$$\mu = 56, \sigma = 7, n = 18$$

$$E(\bar{X}) = \mu = 56$$

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{7}{\sqrt{18}} = 1.6499 = 1.65$$

b. Find the probability that x-bar exceeds 58.

In [4]:
```
normal_dist = norm(56,1.65)
1 - normal_dist.cdf(58)
```

Out[4]: 0.11273299250225355

Since $\bar{X}$ has a normal distribyption we have:

$$P(\bar{X} > 58) = 1 - F(58) = .1127$$

c. Find the probability that the sample mean deviates from the population mean by no more than 2.

In [5]:
```
normal_dist.cdf(58) - normal_dist.cdf(54)
```

Out[5]: 0.7745340149954929

$$P(54 < \bar{X} < 58) = F(58) - F(54) = 0.7745$$

d. Find the probability that x-bar is less than 53.

In [6]:
```
1 normal_dist.cdf(53)
```

Out[6]: 0.03451817399720761

$$P(\bar{X} < 53) = F(53) = 0.0345$$

Q3

An advertiser claims that the average percentage of brown M&Ms candies in a package of milk chocolate M&Ms is 17%. Suppose you randomly select a package that contains 72 pieces and determine the proportion of brown candies in the package.

a. What is the approximate distribution of the sample proportion of brown candies in a package that contains 72 candies?

$P =$ Proportion of brown M&Ms as claimed by the advertiser = 0.17 $N =$ Sample size = 72

Since it is a binominal distribution we have:

$\hat{P} = P = 0.17$

$n * P = 72 * 0.17 = 12.24$

$n * (1 - P) = 72 * .83 = 59.76$

b. What is the probability that the sample percentage of brown candies is less than 20%?

If we use a distribution estimator we will have :

In [7]:
```
1 normal_dist = norm(.17,.044)
2 normal_dist.cdf(.2)
```

Out[7]: 0.7523230370047824

$$\mu_{\hat{P}} = P = 0.17$$

$$\sigma = \sqrt{\frac{p * (1 - p)}{n}} = \sqrt{\frac{.1411}{72}} = .0443$$

$$\hat{P} \sim N(0.17, .0443)$$

$$P(\hat{P} < .20) = F(.2) = .7523$$

if we use the binominal distribution itself we have almost same result

In [8]:
```
1 binom.cdf(14,72,.17)
```

Out[8]: 0.7661715507470683

c. What is the probability that the sample percentage exceeds 35%?

In [9]:
```
1 print(1 - normal_dist.cdf(.35))
```

2.1484277525463114e-05

$$P(\hat{P} > .35) = 1 - F(.35) \approx 0$$

Or you can directley calculate the result with binominal distribution and the result is going to be almost same which is 0

In [10]:
```
1 print(1 - binom.cdf(np.floor(72*.35), 72, .17))
```

7.76553378180056e-05

d. What value would you expect the sample proportion to be greater than 95% of the time?

In [11]:
```
1 normal_dist.ppf(.05)
```

Out[11]: 0.09762644041413521

$$P(\hat{P} > X) = 0.95$$
$$F(\hat{P} = X) = .05$$
$$X = .0976$$

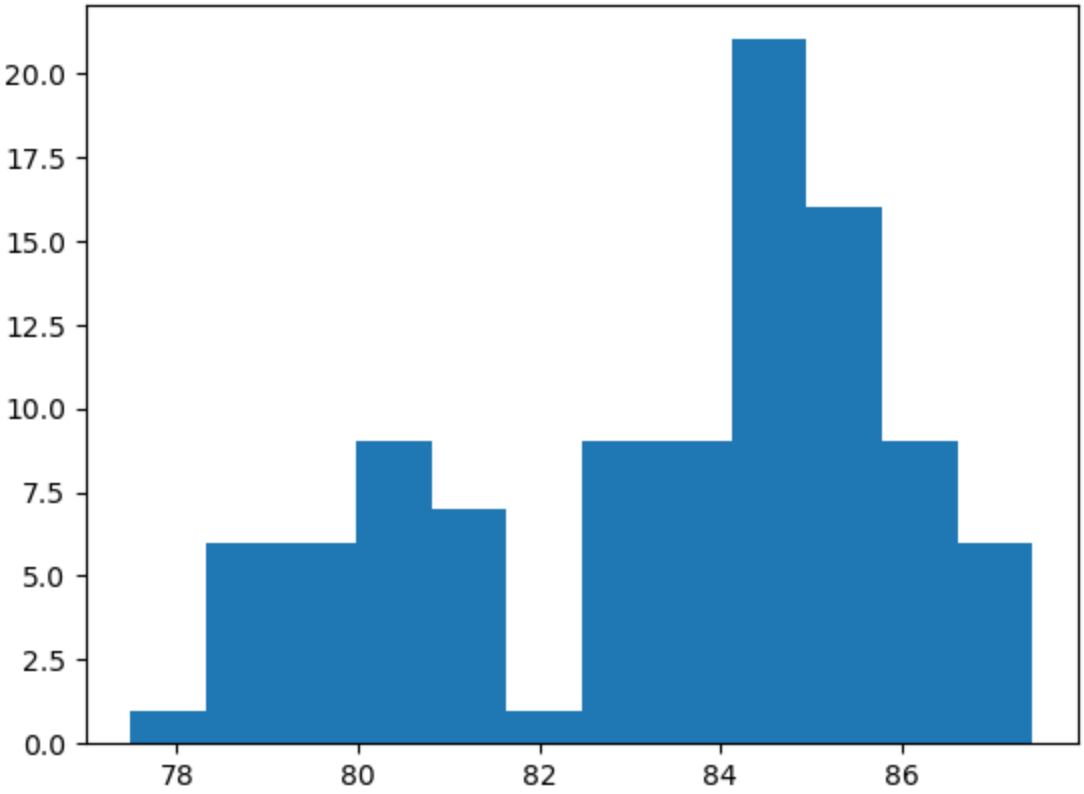
Q4

Use the viscosity data from HW1

a. Determine if the data can be considered normal. Justify your answer with a normal probability plot.

In [12]:

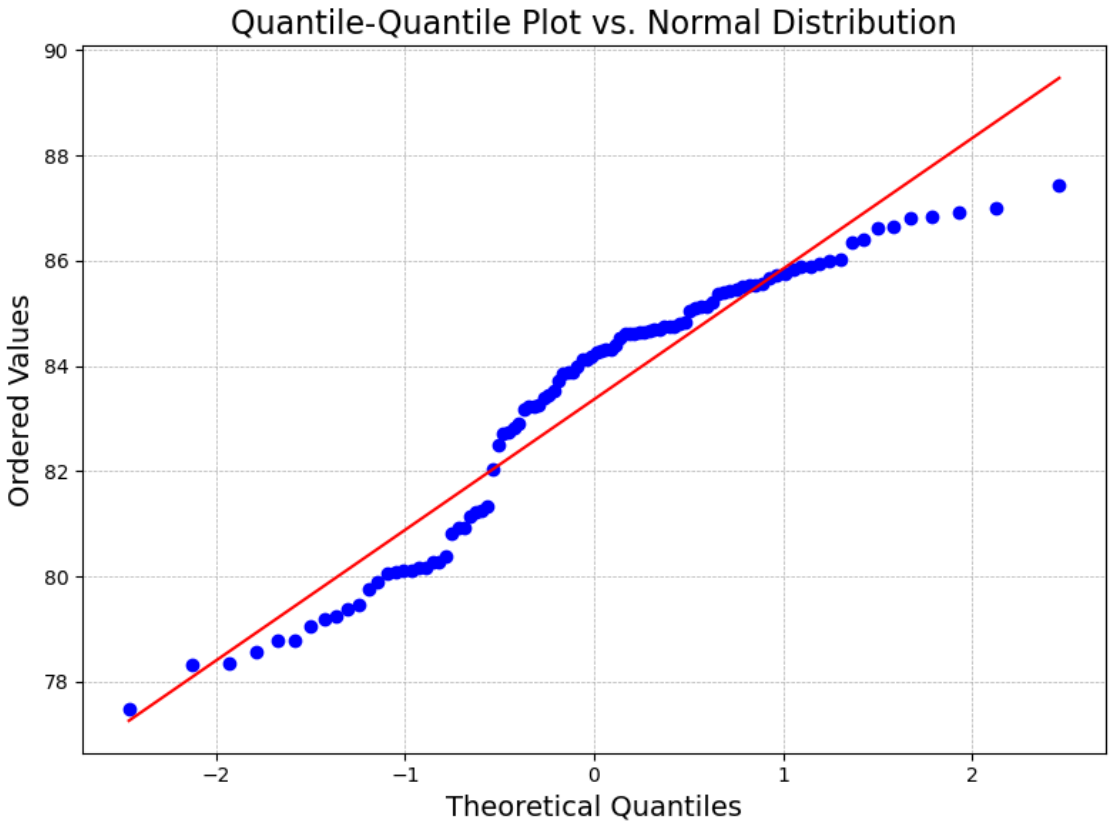
```
df = pd.read_csv("../Week1/HW1/viscosity.txt", sep="\t",index_col=0)
data = df.Viscosity
plt.hist(df.Viscosity,bins=12)
plt.show()
print(F"We have {len(data)} samples")
```




```
We have 100 samples
```

In [13]:

```
import matplotlib.pyplot as plt
import scipy.stats as stats

# Enhanced Quantile-Quantile Plot
plt.figure(figsize=(8, 6))  # Set the figure size
stats.probplot(data, dist="norm", plot=plt)

# Adding title and labels
plt.title("Quantile-Quantile Plot vs. Normal Distribution", fontsize=1
plt.xlabel("Theoretical Quantiles", fontsize=14)
plt.ylabel("Ordered Values", fontsize=14)
plt.grid(True, which="both", linestyle="--", linewidth=0.5)
plt.tight_layout()  # Adjust the layout for better appearance

# Display the plot
plt.show()

# Shapiro-Wilk Test for Normality
shapiro_test_stat, shapiro_p_value = stats.shapiro(data)

# Use f-strings for cleaner message display
message = "Data seems to be normally distributed (Shapiro-Wilk test)"
print(message)

```




```
Data does not seem to be normally distributed (Shapiro-Wilk test)
```

b. Assuming the process mean is typically 85 with a standard deviation of 2. What's the probability of a sample mean less than 83.36 (the $\bar{X}$ of the given sample)?

In [14]:
```
1 normal_dist = norm(85,.2)
2 normal_dist.cdf(83.36)
```

Out[14]: 1.2019351542735388e-16

$$\mu_{\bar{X}} = 85$$
$$\sigma_X = 2$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{10} = .2$$
$$P(\bar{X} < 83.36) = F(\bar{X} = 83.36) \approx 0$$

c. What do you think about this sample? Is this sample representative of the typical process? Maybe think about our conclusions from the last time we looked at this sample.

Upon analyzing the graphical representation of the sample data, particularly from the histogram, a key observation emerges: the data showcases a bimodal distribution. Bimodal distributions occur when the data reveals two different modes or peaks. This suggests that there might be two underlying processes or groups within the data, or there could be some external factors influencing the data to manifest in this manner.
Given that we're observing a bimodal distribution, simply using the mean and standard deviation as representative statistics may be misleading not only for data itself but also for the $\bar{X}$. These metrics, while commonly used for unimodal and normally distributed data, might not capture the nuances and underlying characteristics of a bimodal distribution. The mean might land in between the two peaks, where fewer data points actually reside, making it less representative of the typical data value. Similarly, the standard deviation might be inflated due to the spread of two distinct groups.