NNSE 784
Advanced Analytics Methods

Instructor: F Doyle (CESTM L210)

MW 4:30 – 5:50, NFN 203

Slide Set #7
Sampling Distributions
&
Central Limit Theorem
Introduction

Course Direction

- We have a few weeks of statistics left:
  - Finish Sample Distributions
  - Estimation and Confidence Intervals
  - Hypothesis Testing
  - Analysis of Variance (ANOVA)
  - Linear regression, which will segue into…
- Machine Learning
  - Regression, classification, clustering, PCA
  - Measures of performance (F1, specificity, etc.)
  - Cross validation
  - Hyper-parameter tuning
  - Models
    - Logistic regression
    - Decision trees/random forests
    - Support Vector Machines
    - Multi-layer perceptron (feed forward neural nets)
- Time Series Data Analysis

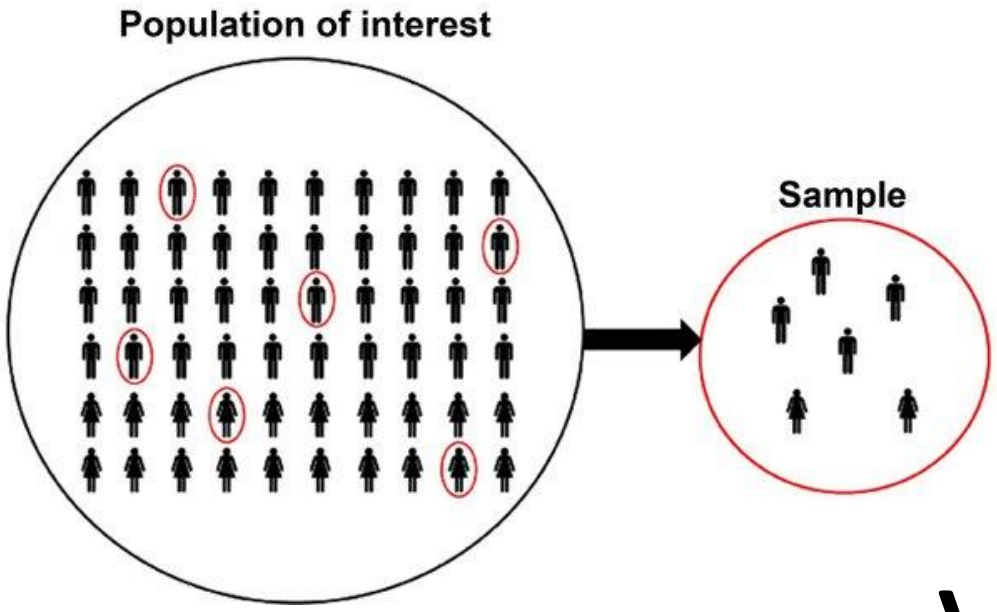
Lecture Outline

- Quiz 1 – 15 minutes
- Introduce the concept of sampling and sampling distributions
- Distribution of the Sample Mean
- Central Limit Theorem
- Python code to demonstrate the proposed behavior
- Additional CLT example

What do we mean by “Sampling”?

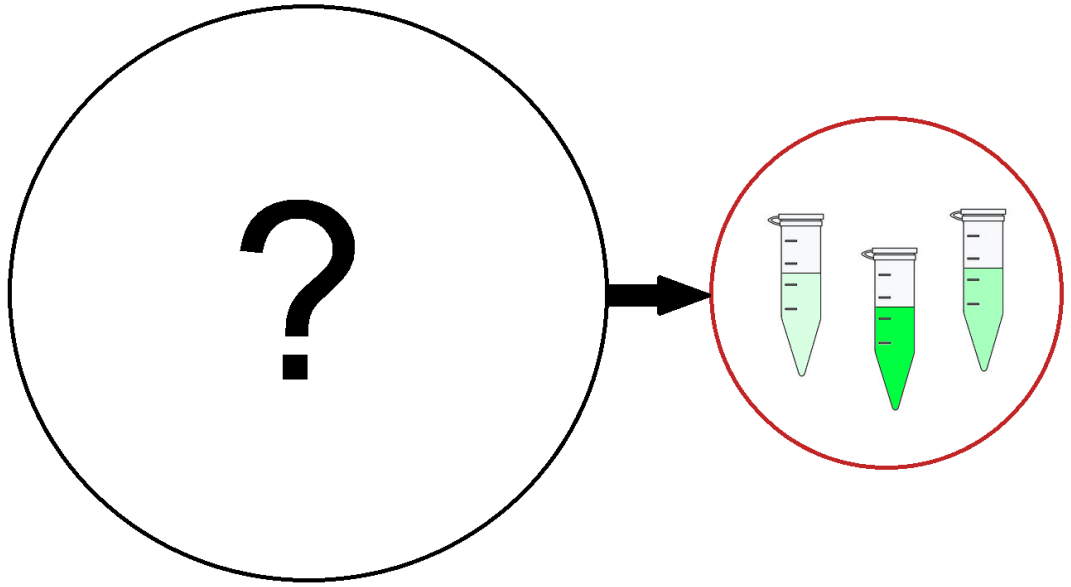
- A simple random sample (SRS) can be thought of as set of n **random** representative members picked from a population
- What does this mean?
  - The term population can be confusing.
  - Random people picked from a larger group could be considered a sample of that larger group or population, which may be of a known size (e.g., U.S. citizenry)
  - Experiments performed in the lab can also be thought of as samples of a theoretically infinite population (the protocol of the experiment could potentially be repeated any number of times)

Thinking of Experiments As Samples



VS

Theoretical Population of All
Possible Experimental Values
For Given Protocol



Sampling Distributions – How do they fit in?

- So far, we have looked at some concepts that are fundamental to understanding statistical inference:
  - Descriptive statistics
  - Basic Probability
  - Probability Distributions
- Understanding sampling distributions is the key to understanding inferential statistics
- Sampling distributions serve two core purposes:
  - Answer probability questions about sample statistics
  - Provide the necessary theory that validates statistical inference procedures

Sampling Distribution - Definition

The distribution of all possible values that can be assigned to some statistic, computed from samples of the same size (n) randomly drawn from the same population.

Sampling Distribution Construction

- From a finite population of size *N*, randomly draw all possible samples of size *n (with replacement)*
- Compute the statistic for each sample
- Create a frequency table (and/or histogram) that shows number of occurrences for each distinct value

- Sampling distributions of infinite populations may be approximated by taking a large number of samples of a given size (n)

Sampling Distribution Characteristics

- We are typically interested in three things about a given sampling distribution:
  1. Mean
  2. Variance
  3. Functional form (shape of the plotted distribution)

Distribution of the Sample Mean

Example of All Possible Samples n=2 from Population N=5

Population members = [6, 8, 10, 12, 14]

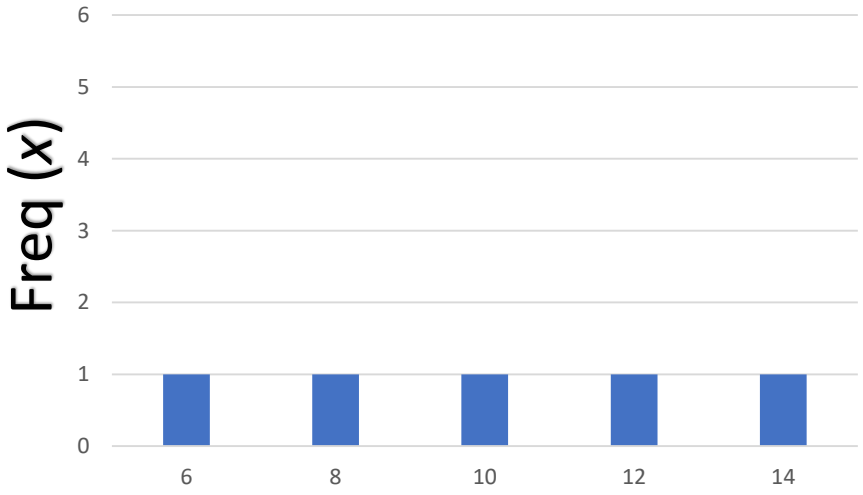
Table shows all possible combinations of 2 samples… **with replacement!**

The average of the two sample values is shown in parentheses.

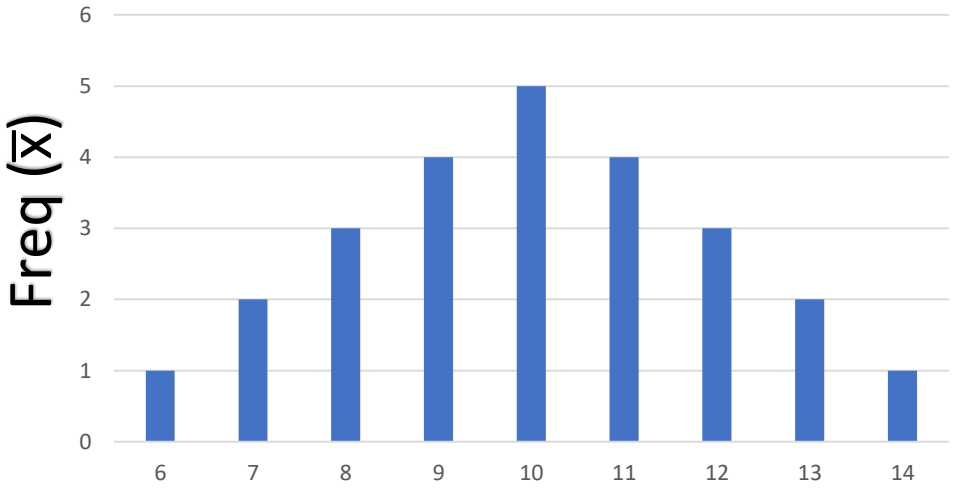
		Second Draw				
		6	**8**	**10**	**12**	**14**
	6	6, 6 (6)	6, 8 (7)	6, 10 (8)	6, 12 (9)	6, 14 (10)
	8	8, 6 (7)	8, 8 (8)	8, 10 (9)	8, 12 (10)	8, 14 (11)
First Draw	**10**	10, 6 (8)	10, 8 (9)	10, 10 (10)	10, 12 (11)	10, 14 (12)
	12	12, 6 (9)	12, 8 (10)	12, 10 (11)	12, 12 (12)	12, 14 (13)
	14	14, 6 (10)	14, 8 (11)	14, 10 (12)	14, 12 (13)	14, 14 (14)

Note! – This is the set of **ALL** possible samples of size n=2. We are using it to demonstrate a complete sample distribution. This is the parent population that random samples of size n=2 could be taken from.

Population Distribution vs Sampling Distribution



Distribution of population



Sampling distribution of $\bar{x}$

(n = 2)

$$\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{N^n} = \frac{6+7+7+8+\cdots+14}{25} = \frac{250}{25} = 10$$

Sampling Distribution of $\bar{x}$: Variance

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \frac{\sum (\bar{x}_i - \mu_{\bar{x}})^2}{N^n} \\ &= \frac{(6-10)^2 + (7-10)^2 + (7-10)^2 + \cdots + (14-10)^2}{25} \\ &= \frac{100}{25} = 4\end{aligned}$$

Note: the variance of the sampling distribution is not equal to the population variance. However, the variance of the sampling distribution IS equal to the population variance divided by the size of the sample used to obtain the distribution:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{8}{2} = 4$$

Sampling Distribution of $\bar{x}$: Sampling from a Normally Distributed Population

When sampling is from a normally distributed population, the distribution of the sample mean will possess the following properties:

1. The distribution of $\bar{x}$ will be normal.
2. The mean, $\mu_{\bar{x}}$, of the distribution of $\bar{x}$ will be equal to the mean of the population from which the samples were drawn.
3. The variance, $\sigma^2_{\bar{x}}$ of the distribution of $\bar{x}$ will be equal to the variance of the population divided by the sample size.

Sampling From Nonnormally Distributed Populations

The Central Limit Theorem:

Given a population of any nonnormal functional form with a mean μ and finite variance σ^2, the sampling distribution of $\bar{x}$, computed from samples of size *n* from this population, will have mean μ and variance σ^2/n and will be approximately normally distributed when the sample size is large.

Central Limit Theorem - Python Example

Population of integers from 1 to 99.

```
import random
#This method will return the average of n integer samples pick at random between
#1 and 99
def get_sample_average(sample_size):
    sum = 0
    for x in range(sample_size):
        sum = sum + random.randint(1,99)
    avg = sum/sample_size
    return avg
```

```
import pandas as pd
df = pd.DataFrame()
for n in range(1,31):
    name = "n = "+str(n)
    samp_list = []
    for sample in range(1000):
        samp_list.append(get_sample_average(n))
    df[name] = samp_list
```

Dataframe of samples for size n = 1 to 30

```
df.describe()
```

	n = 1	n = 2	n = 3	n = 4	n = 5	n = 6	n = 7	n = 8	n = 9
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	50.447000	50.006500	50.326333	50.453250	49.481400	49.645333	50.076714	50.192250	49.858111
std	29.018602	20.260039	17.076661	14.343334	13.474014	12.061535	10.978519	10.219945	10.142869
min	0.000000	2.500000	5.000000	2.500000	11.600000	18.166667	19.428571	15.375000	20.444444
25%	26.000000	36.000000	37.333333	40.437500	40.600000	41.333333	42.535714	43.125000	42.888889
50%	51.000000	49.750000	50.666667	51.000000	49.200000	49.833333	50.285714	50.500000	49.666667
75%	76.000000	64.500000	62.333333	60.750000	59.400000	58.166667	57.714286	57.375000	56.583333
max	100.000000	99.000000	98.000000	92.500000	91.800000	86.000000	80.000000	80.750000	81.666667

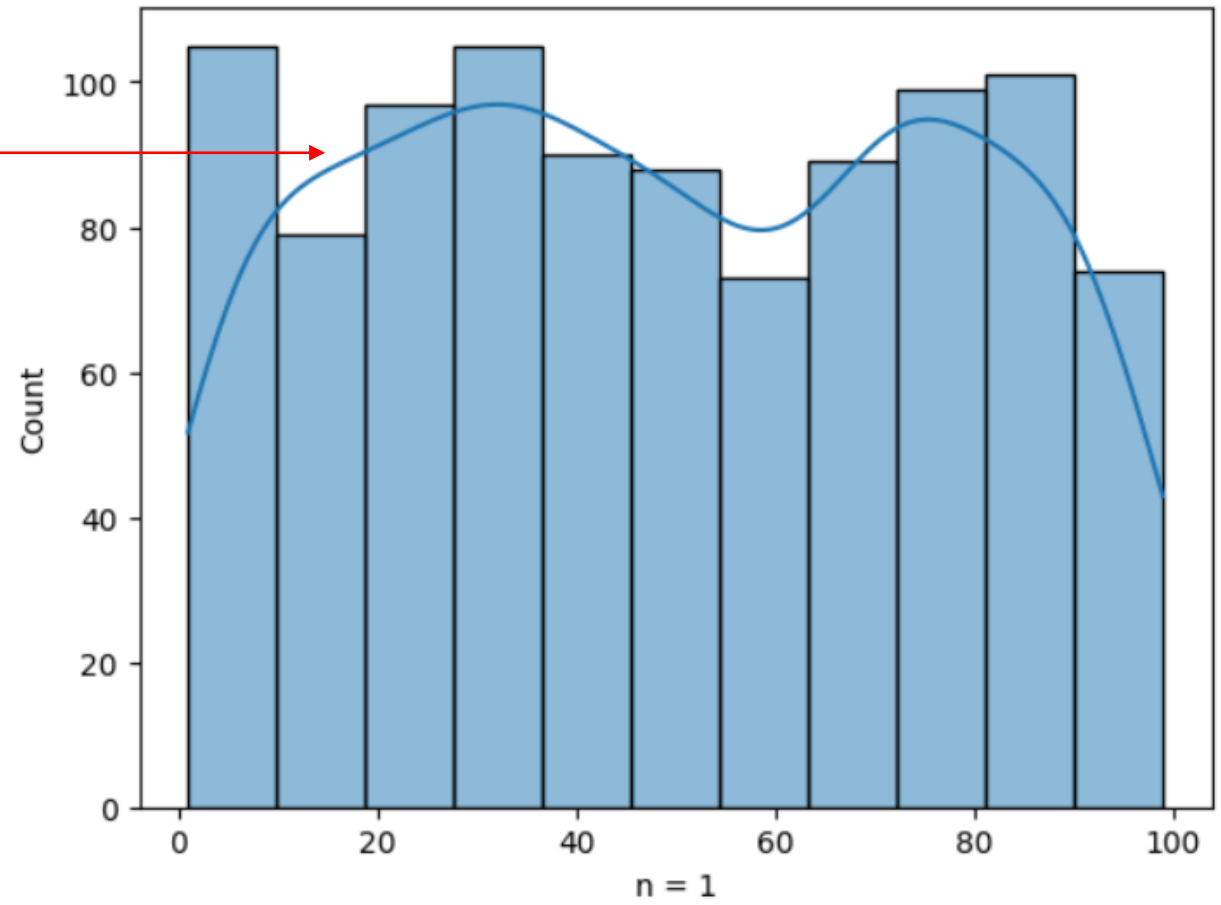
8 rows × 30 columns

Sample size(n) = 1

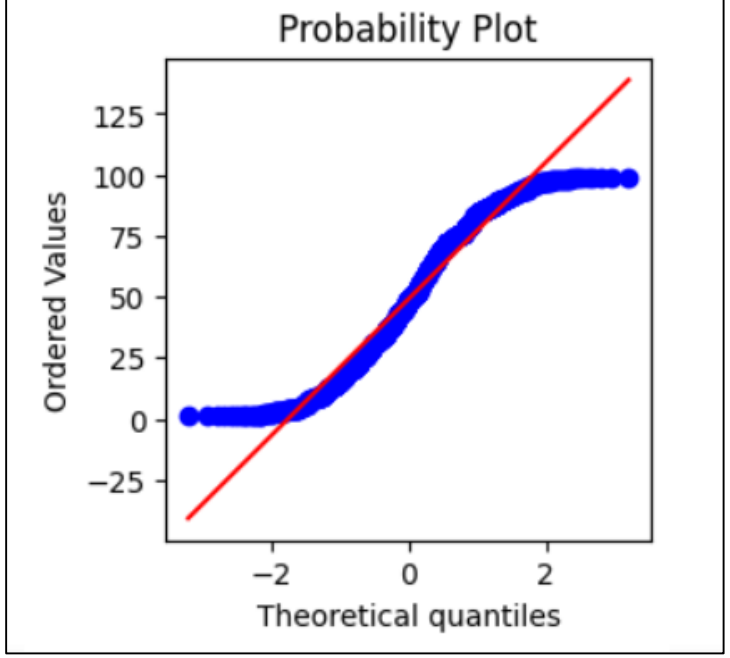
Approximately uniform

```
import seaborn as sns
sns.histplot(df["n = 1"], kde=True)
```

```
<AxesSubplot:xlabel='n = 1', ylabel='Count'>
```




```
##probability plot to help assess normality
import matplotlib.pyplot as plt
import scipy
fig, ax = plt.subplots(figsize=(3, 3))
scipy.stats.probplot(df['n = 1'], plot=ax)
```



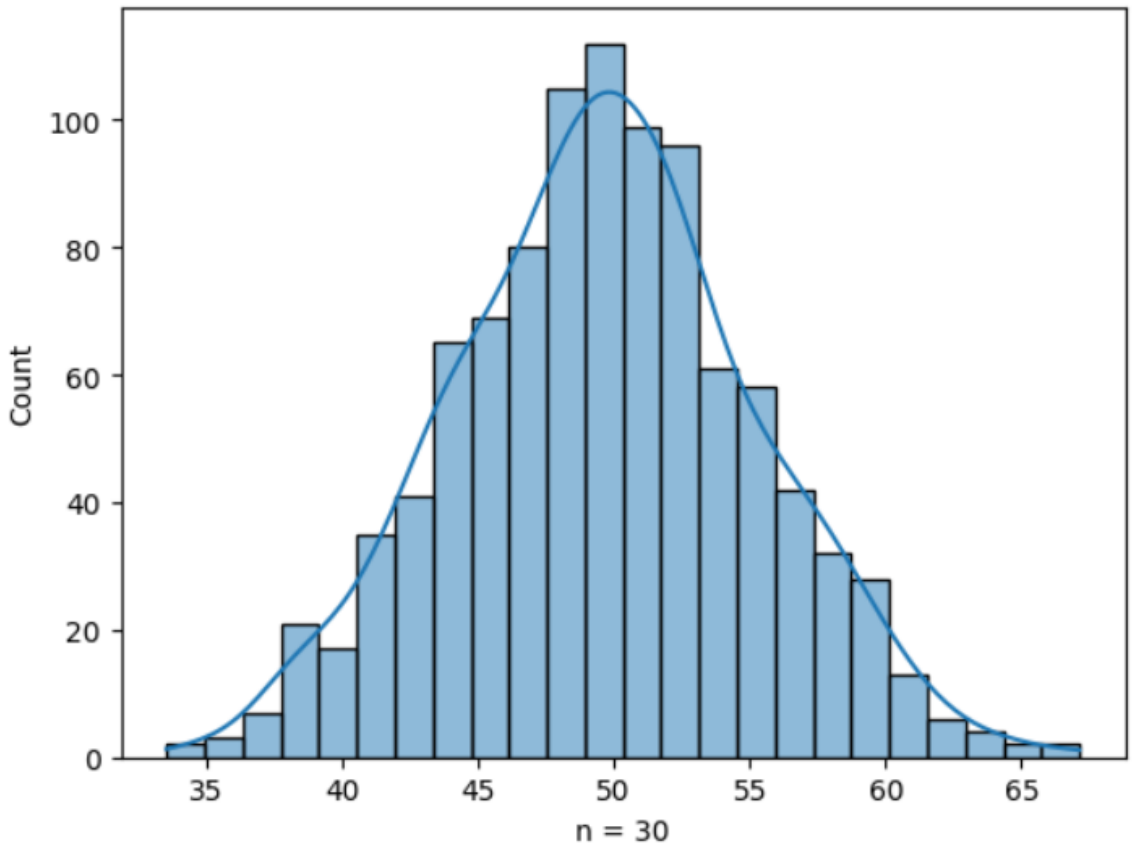

```
scipy.stats.normaltest(df['n = 1'])
```

```
NormaltestResult(statistic=519.6753331150117, pvalue=1.4253949750971322e-113)
```

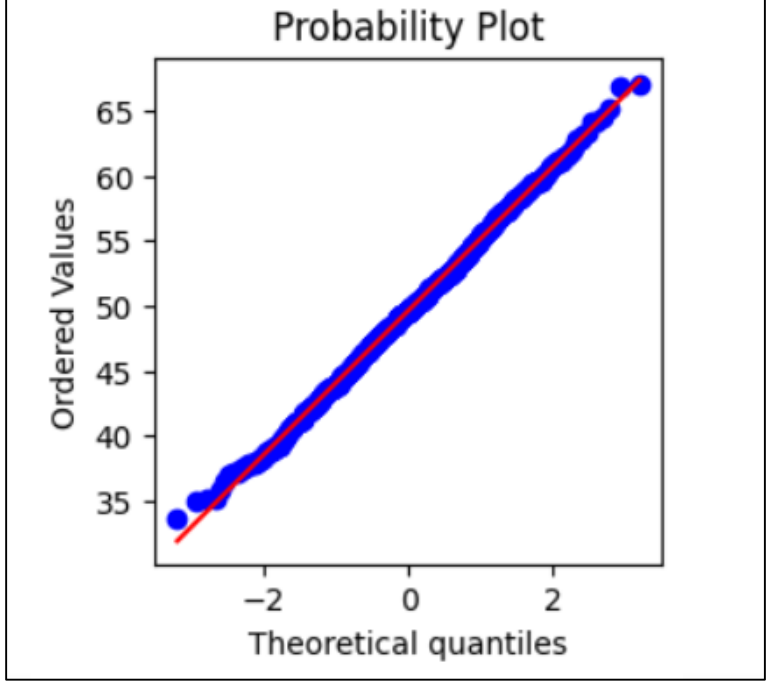
Sample size(n) = 30

```
sns.histplot(df["n = 30"], kde=True)
```

```
<AxesSubplot:xlabel='n = 30', ylabel='Count'>
```

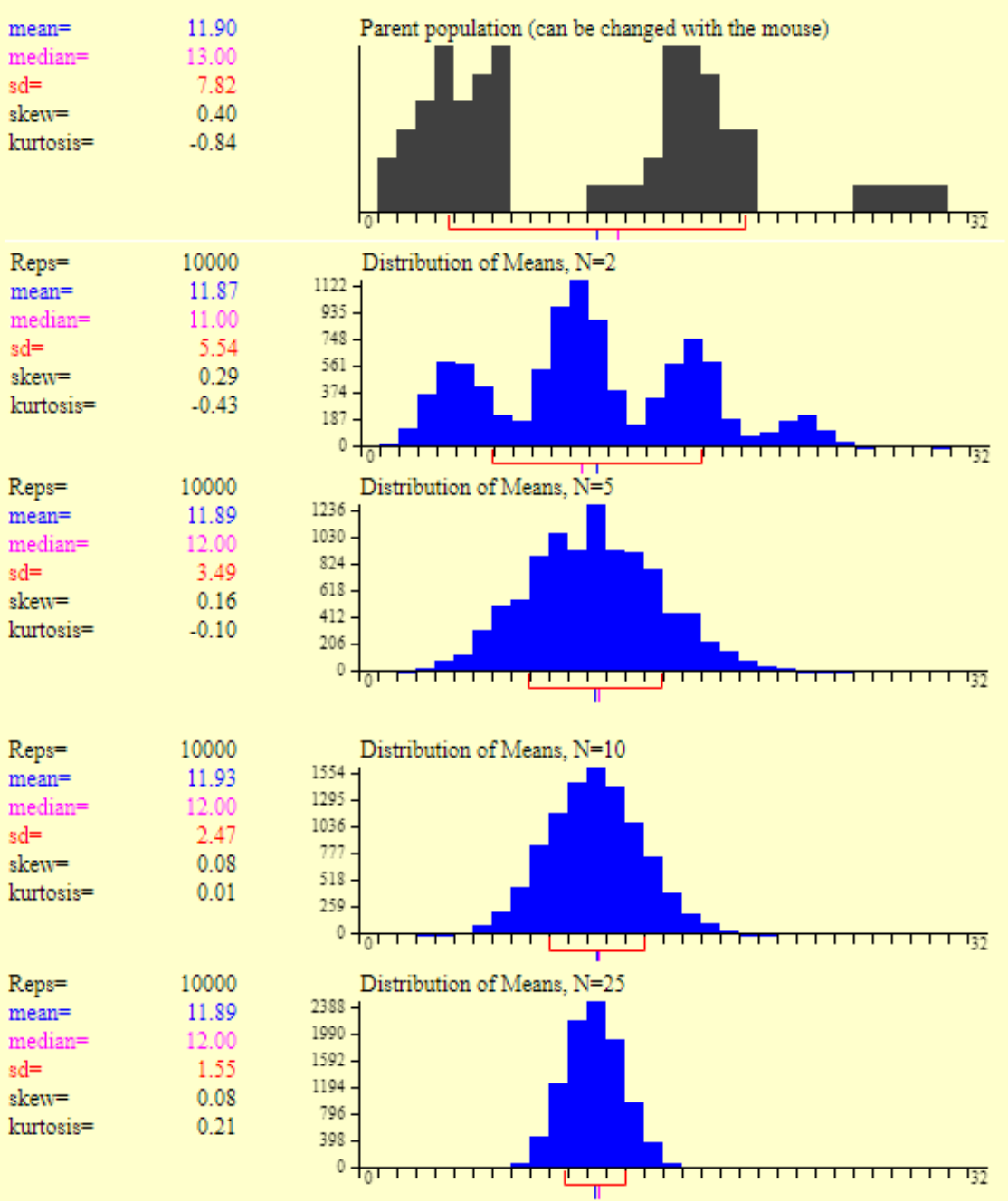



```
##probability plot to help assess normality
import matplotlib.pyplot as plt
import scipy
fig, ax = plt.subplots(figsize=(3, 3))
scipy.stats.probplot(df['n = 30'],plot=ax)
```




```
scipy.stats.normaltest(df['n = 30'])
```

```
NormaltestResult(statistic=0.7648871040223206, pvalue=0.6821923982784968)
```



Central Limit Theorem - Recap

Not only is the parent population not normal, it is composed of discontinuous blocks!

However, note the shape of the sampling distributions as *n* increases.

If the population distribution is Normal, then so is the sampling distribution of $\bar{x}$. This is true no matter what the sample size n is.

If the population distribution is not Normal, the central limit theorem tells us that the sampling distribution of $\bar{x}$ will be approximately Normal in most cases if $n \geq 30$.

Graphs generated at site: https://onlinestatbook.com/stat_sim/sampling_dist/index.html