Unit 9 Notebook: Introduction to Inference – The Central Limit Theorem and Confidence Intervals for μ and p

Case Study Pew Survey Analysis 1

What is a plausible range of values for the average age of ALL adults living in the U.S.?

Case Study Pew Survey Analysis 2

What is a plausible range of values for the proportion of ALL adults living in the U.S. that are satisfied with the way things are going in the country at the time of the survey (2017)?

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()
```

1. Two Main Types of Inference for Unknown Population Parameters

See Unit 9 slides (section 1).

2. Proving Properties of Sampling Distributions of Sample Means

See Unit 9 slides (section 2).

2.1 Proving that the Mean of Sampling Distributions $\approx E[\bar{X}] = \mu$

See Unit 9 slides (section 2.1).

2.2 Proving that the Standard Deviation of Sampling Distributions (ie. the Standard Error) $\approx SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}$

See Unit 9 slides (section 2.2).

2.3 Proving that the sampling distribution of sample means is approximately normal if either:

1.) the sample size n>30 OR

2.) the population distribution (or sample distribution) is approximately normal.

See Unit 9 slides (section 2.3).

3. Confidence Intervals

See Unit 9 slides (section 3).

4. Confidence Interval for a Population Mean μ

4.1. Confidence Interval for a Population Mean μ - General Framework

See Unit 9 slides (section 4.1).

<u>Pew Survey Analysis 1</u> What is a plausible range of values for the average age of ALL adults living in the U.S.?

Ex: Suppose we wanted to calculate a 95% confidence interval (ie. range of plausible values) for μ (the average age of ALL adults living in the U.S.). We have a random sample of size n=1489 that has a mean age of 50.49 years and a standard deviation of 17.84 years. **Suppose we also know that the standard deviation of ALL**

4.1.1 Dataset Cleaning and Inspection

The February 2017 Pew Research Center random phone number dialing survey had 1,503 respondents in total.

First, let's learn a little more about this dataset.

In [2]:
```
# Read in the data
missing_values = ["NaN", "nan", "Don't know/Refused (VOL.)"]
df_pew = pd.read_csv('Feb17public.csv',
                     na_values=missing_values)
df_pew.head()
```

Out[2]:

	psraid	sample	int_date	fcall	version	attempts	refusal	ilang	cregion	state
0	100008	Landline	21017	170207	Client changes	4	No	English	Midwest	Illinois
1	100019	Landline	21217	170207	Client changes	4	Yes	English	South	North Carolina
2	100020	Landline	21217	170207	Client changes	4	Yes	English	Northeast	New York
3	100021	Landline	20717	170207	Initial version	1	No	English	Midwest	Minnesota
4	100024	Landline	20717	170207	Initial version	1	No	English	Midwest	Illinois

5 rows × 130 columns

What is the shape of the dataset?

In [3]:
```
df_pew.shape
```

Out[3]:
```
(1503, 130)
```

What columns are contained in this dataset?

In [4]:
```
df_pew.columns
```

Out[4]:
```
Index(['psraid', 'sample', 'int_date', 'fcall', 'version', 'attempts',
       'refusal', 'ilang', 'cregion', 'state',
       ...
       'ql1a', 'qc1', 'money2', 'money3', 'iphoneuse', 'hphoneuse', 'll', 'c
p',
       'cellweight', 'weight'],
      dtype='object', length=130)
```

Let's use a for-loop so we can inspect all 130 column names.

In [5]:
```
for col in df_pew.columns:
    print(col)
```

psraid
sample
int_date
fcall
version
attempts
refusal
ilang
cregion
state
density
sstate
form
stimes
igender
irace
llitext0
susr
usr
scregion
qs1
q1
q1a
q2
q5af1
q5bf1
q5cf1
q5df1
q6af2
q6bf2
q6cf2
q6df2
q10a
q10b
q15af1
q15b
q15cf2
q15df1
q15ef1
q15ff1
q15gf2
q15hf2
q15if2
q16
q19
q35
q36
q37
q39
q43
q44
q45
q45vb
Q45VB0
Q45VB1
Q45VB2
q45oem1
q45oem2
q45oem3
q52
q53
q54
q55
q61a
q61b
q61c
q61d
q61e
q62f1
q63f1
q64f2
q65
q66
q68f1

```
q69f2
q70f1
q71f2
q74
q75
q81
q82
q84a
q84bf1
q84cf1
q84df1
q84ef2
q84ff2
q84gf2
q88
q90f1
q91f2
sex
age
gen5
educ2
hisp
adults
racethn
racethn2
birth_hisp
citizen
child
relig
chr
born
attend
q92
q92a
income
reg
party
partyln
partysum
partyideo
q93
q94
ideo
hh1
hh3
ql1
ql1a
qc1
money2
money3
iphoneuse
hphoneuse
ll
cp
cellweight
weight
```

How many missing values are in each column?

In [6]:
```
df_pew.isna().sum()
```

Out[6]:
```
psraid          0
sample          0
int_date        0
fcall           0
version         0
              ...
hphoneuse       0
ll              0
cp              0
cellweight    377
weight          0
Length: 130, dtype: int64
```

How many missing values are in the 'age' column?

In [7]:
```
df_pew.isna().sum().loc['age']
```

Out[7]:
```
14
```

Let's create a pandas series that is just the the age column of this dataframe and drop the missing values from this series.

In [8]:
```
df_pew_age=df_pew['age'].dropna()
df_pew_age
```

Out[8]:
```
0       80.0
1       70.0
2       69.0
3       50.0
4       70.0
        ...
1498    37.0
1499    30.0
1500    72.0
1501    67.0
1502    35.0
Name: age, Length: 1489, dtype: float64
```

The code below confirms that we dropped 14 (=1503-1489) entries from this series that had missing values.

In [9]:
```
df_pew_age.shape
```

Out[9]:
```
(1489,)
```

4.1.2. Collecting Sample Information

If we consider our **population** to be ALL adults living in the U.S. then we can think of this Pew dataset as a **random sample** of size $n = 1489$ from this population. Because the sample is **random** we can use this dataset to **make inferences** about our population.

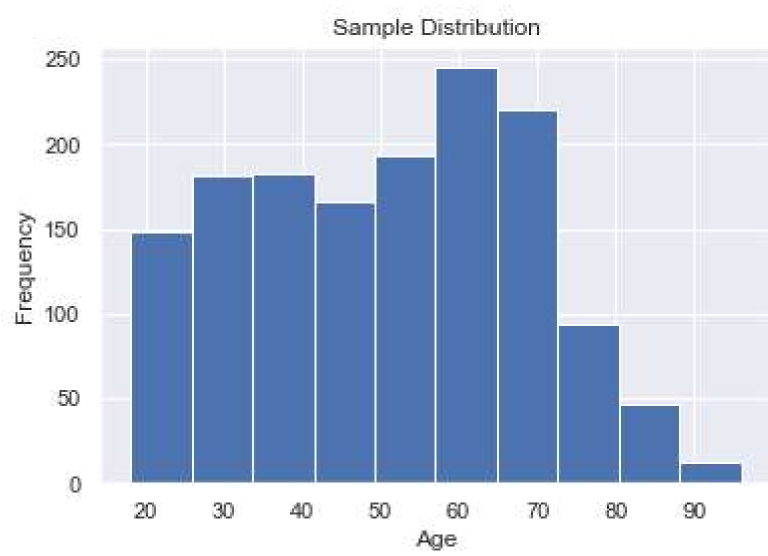
In [10]:
```
# compute and display sample statistics
sample_mean_age = df_pew_age.mean()
sample_std_age = df_pew_age.std()
n_age = df_pew_age.shape[0]
print('sample mean age=', round(sample_mean_age, 2),
      'sample std age=', round(sample_std_age,2),
      'sample size n=', n_age)

pop_std_age=18

print('popuolation standard deviation age=',pop_std_age)
```

```
sample mean age= 50.49 sample std age= 17.84 sample size n= 1489
popuolation standard deviation age= 18
```

In [11]:
```
df_pew_age.hist()
plt.title('Sample Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



4.1.3 Are we allowed to calculate a confidence interval for μ using this sample that was collected and the equations we just learned? Why or why not?

Yes, the Central Limit Theorem Conditions (for Sample Means) below are met by this sample.

1. Condition: The observations are independent.

   - Because the sample is collected via random sampling and $n < 10\%$ of the population of ALL adults living in the U.S.

2. Condition: Either n>30 OR the population distribution is normal..

   - It looks like the population distribution of ages is NOT normal. How do we know?
     - The sample distribution and the population distribution tend to mirror each other.
     - The sample distribution above is NOT symmetric and unimodal, therefore, it is not a good assumption to say that it is normal.
   - However, because our sample size $n = 1489 > 30$,this condition is met.

4.1.4 What if one of the conditions above was not met and we calculated our confidence interval anyway using the given equations?

Then our interpretations about the confidence interval may not be valid. For instance, we are calculating a 95% confidence interval. However, if our assumptions are not met, it may (for instance) be the case that we are only 90% confident that our population mean is in the range we produced.

4.1.5 What is the critical value for this 95% confidence interval

Goal: Find the POSITIVE z-score z^* in the standard normal distribution in which:

- an area of 0.95 is in between $-z^*$ and z^*.

Put another way: We want to find the POSITIVE z-score z^* in the standard normal distribution in which:

- an area of 0.975=0.025+0.95 is to the left of z^* and
- an area of 0.025 is to the right of z^*.

We can find the x-axis value (ie. the z-score) that has a left tail area of 0.975 by using the **norm.ppf()** function.

In [12]:
```
from scipy.stats import norm
critical_value=norm.ppf(0.975)
critical_value
```

Out[12]:
```
1.959963984540054
```

The us the **critical value** for this 95% confidence interval is $z^* = 1.96$.

4.1.6 Calculate the 95% confidence interval.

Thus, using our confidence interval equation we get:

$$(\bar{x} - z^*\frac{\sigma}{\sqrt{n}}, \bar{x} + z^*\frac{\sigma}{\sqrt{n}})$$

$$(50.49 - (1.96)\frac{18}{\sqrt{1489}}, 50.49 + (1.96)\frac{18}{\sqrt{1489}})$$

$$(49.57, 51.40).$$

In [13]:
```
lower_bound=sample_mean_age-critical_value*(pop_std_age/np.sqrt(n_age))
upper_bound=sample_mean_age+critical_value*(pop_std_age/np.sqrt(n_age))

print(lower_bound,',',upper_bound)
```

```
49.57397971873396 , 51.40251457273683
```

4.1.7 Interpret this 95% confidence interval.

We are 95% confident that the average age of all adults living in the U.S. (ie. μ) is between 49.57 and 51.40.

4.2 What to do with you don't know σ?

See Unit 9 slides (section 4.2).

Ex: Suppose we wanted to calculate a 90% confidence interval (ie. range of plausible values) for μ (the average age of ALL adults living in the U.S.). We have a random sample of size n=1503 that has a mean age of 50.49 years and a standard deviation of 17.84 years. Suppose we didn't know what the population standard deviation was.

Because $n > 30$ (for now) we can plug in s for σ and still get a relatively valid confidence interval:

$(\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}})$

$(\bar{x} - z^* \frac{s}{\sqrt{n}}, \bar{x} + z^* \frac{s}{\sqrt{n}})$

$(50.49 - (1.96)\frac{17.84}{\sqrt{1489}}, 50.49 + (1.96)\frac{17.84}{\sqrt{1489}})$

$(49.58, 51.39)$.

In [14]:
```
lower_bound=sample_mean_age-critical_value*(sample_std_age/np.sqrt(n_age))
upper_bound=sample_mean_age+critical_value*(sample_std_age/np.sqrt(n_age))

print(lower_bound,',',upper_bound)
```

```
49.581904861535484 , 51.3945894299353
```

We are 95% confident that the average age of all adults living in the U.S. (ie. μ) is between 49.58 and 51.39.

4.3 What does "95% confident" mean?

See Unit 9 slides (section 4.3).

5. Binomial Random Variables

See Unit 9 slides (section 4.3).

Ex: About 35% of American adults use Instagram. We decide to collect a random sample of 5 American adults and ask if they use Instagram or not.

7. What is the probability that we select a random sample with 2 people that use Instagram (using Python)?

If we let Let X = # of randomly selected American adults (out of 5) that are Instagram users, then we know that

$X \sim Bin(n = 5, p = 0.35)$.

$P(X = 2) = \binom{n}{2} p^k (1-p)^{n-k} = \binom{5}{2} (0.35)^2 (1-0.35)^{5-2} = (10)(0.35)^2(1-0.35)^{5-2} = 0.336$

In [15]:
```
from scipy.stats import binom
binom.pmf(2, n=5, p=0.35)
```

Out[15]:
```
0.33641562499999983
```

8. What is the probability that we select a random sample with more than 2 people that use Instagram (using Python)?

$$P(X > 2) = P(X \geq 3) = 1 - P(X \leq 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$

$$= 1 - [\binom{n}{0} p^0 (1-p)^{n-0} + \binom{n}{1} p^1 (1-p)^{n-1} + \binom{n}{2} p^2 (1-p)^{n-2}]$$

$$= 1 - [\binom{5}{0} p^0 (1-0.35)^{5-0} + \binom{5}{1} (0.35)^1 (1-0.35)^{5-1} + \binom{5}{2} (0.35)^2 (1-0.35)^{5-2}]$$

$$= 0.235.$$

In [16]:
```
1-binom.cdf(2, n=5, p=0.35)
```

Out[16]:
```
0.2351693749999998
```

7. Confidence Interval for a Population Proportion *p*

7.1. Confidence Interval for a Population Proportion *p* - General Framework

See Unit 9 slides (section 7.1).

7.2. What do you do when you need to plug in a 'p' in the conditions and confidence interval equation, butyou don't know p?

See Unit 9 slides (section 7.2).

Pew Survey Analysis 2 What is a plausible range of values for the proportion of ALL adults living in the U.S. that are satisfied with the way things are going in the coutnry at the time of the survey (2017)?

Ex: Suppose we wanted to calculate a 99% confidence interval (ie. range of plausible values) for p, the proportion of all adults living in the U.S. that are satisfied with the way things are going in the country at the time of the survey (2017). We collect a sample of size n=1503 that has a sample proportion of __.

7.2.1 Dataset Cleaning and Inspection

We wil be using the same 2017 Pew dataset as in the previous case study. The 'q2' column contains responses to the following question: 'All in all, are you satisfied or dissatisfied with the way things are going in this country today?'

Let's create a pandas series that is just the the q2 column of this dataframe and drop the missing values from this series.

In [17]:
```
df_pew_q2=df_pew['q2'].dropna()
df_pew_q2
```

Out[17]:
```
0       Dissatisfied
1       Dissatisfied
2       Dissatisfied
3          Satisfied
4       Dissatisfied
            ...
1498       Satisfied
1499    Dissatisfied
1500    Dissatisfied
1501    Dissatisfied
1502       Satisfied
Name: q2, Length: 1435, dtype: object
```

It looks like we ended up dropping 68 = (1503 - 1435) entries in this column with missing values.

In [18]:
```
df_pew_q2.shape
```

Out[18]:
```
(1435,)
```

7.2.2 Collect information from the problem.

How many of each type of response is there in this column?

In [19]:
```
q2sum=df_pew_q2.value_counts()
q2sum
```

Out[19]:
```
Dissatisfied    1003
Satisfied        432
Name: q2, dtype: int64
```

Now we can compute the sample proportion that are satisfied as $\hat{p} = 0.301$.

In [20]:
```
prop = q2sum['Satisfied']/q2sum.sum()
round(prop, 4)
```

Out[20]:
```
0.301
```

Sample size n=1435

In [21]:
```
n_prop=df_pew_q2.shape[0]
n_prop
```

Out[21]:
```
1435
```

7.2.3. Are we allowed to calculate a confidence interval for p using this sample that was collected and the equations we just learned? Why or why not?

Yes, the Central Limit Theorem Conditions (for Sample Proportions) below are met by this sample.

1. Condition: The observations are independent.
   - Because the sample is collected via random sampling and $n < 10\%$ of the population of ALL adults living in the U.S.
2. Condition: $np \geq 10$ and $n(1-p) \geq 10$.
   - Because we don't know p, we plug in $\hat{p} = 0.301$ in for p in the conditions above.
   - $n\hat{p} = 1435 \cdot 0.301 \geq 10$
   - $n(1-\hat{p}) = 1435 \cdot (1-0.301) \geq 10$.

In [22]:
```
n_prop*prop
```

Out[22]:
```
432.0
```

In [23]:
```
n_prop*(1-prop)
```

Out[23]:
```
1003.0
```

7.2.4. What is the critical value for this 99% confidence interval?

<u>Goal:</u> Find the POSITIVE z-score z^* in the standard normal distribution in which:

- an area of 0.99 is in between $-z^*$ and z^*.

<u>Put another way:</u> We want to find the POSITIVE z-score z^* in the standard normal distribution in which:

- an area of 0.995=0.001+0.99 is to the left of z^* and
- an area of 0.001 is to the right of z^*.

We can find the x-axis value (ie. the z-score) that has a left tail area of 0.995 by using the **norm.ppf()** function.

In [24]:
```
from scipy.stats import norm
critical_value=norm.ppf(0.995)
critical_value
```

Out[24]:
```
2.5758293035489004
```

The critical value for this 99% confidence interval is $z^* = 2.576$.

7.2.5 Calculate the 99% confidence interval.

$$(\hat{p} - z^*\sqrt{\frac{p(1-p)}{n}}, \hat{p} - z^*\sqrt{\frac{p(1-p)}{n}})$$

$$(\hat{p} - z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} - z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

$$(0.301 - 2.576\sqrt{\frac{0.301}{(1-0.301)}1435}, 0.301 + 2.576\sqrt{\frac{0.301}{(1-0.301)}1435})$$

$(0.27, 0.33)$.

In [25]:
```
lower_bound=prop-critical_value*np.sqrt(prop*(1-prop)/n_prop)
upper_bound=prop+critical_value*np.sqrt(prop*(1-prop)/n_prop)

print(lower_bound,',',upper_bound)
```

```
0.26985413319300744 , 0.33223645914148736
```

7.2.6. Interpret this 99% confidence interval.

We are 99% confident that the proportion of ALL adults living in the U.S. that approve of the way things are going in the country (in 2017) is between 0.27 and 0.33.

STAT 207, Victoria Ellison and Douglas Simpson, University of Illinois at Urbana-Champaign

```
In [ ]:
```