Bihan Qian(bq2150)

"Three ways to test differences between two groups, along with their assumptions, validities, and remedial solutions."

A) Consider the crabs data frame in R library MASS which has 200 rows and 8 columns, describing 5 morphological measurements on 50 crabs each of two color forms ("B" or "O" for blue or orange) and both sexes, of the species Leptograpsus variegatus collected at Fremantle, W. Australia. data(crabs, package="MASS")

(1). Determine whether there is a significant difference between blue and orange crabs in mean carapace length (mm) [CL] using each of the following procedures:

a. A parametric procedure

b. A non-parametric procedure

c. A re-sampling procedure

(2) Discuss the assumptions underlying the analyses in (1) above, their validity, and any remedial measures to be taken.

In [1]:
```
import statsmodels.api as sm
crabs = sm.datasets.get_rdataset('crabs','MASS')
```

In [2]:
```
crabs.data
```

Out[2]:

	sp	sex	index	FL	RW	CL	CW	BD
0	B	M	1	8.1	6.7	16.1	19.0	7.0
1	B	M	2	8.8	7.7	18.1	20.8	7.4
2	B	M	3	9.2	7.8	19.0	22.4	7.7
3	B	M	4	9.6	7.9	20.1	23.1	8.2
4	B	M	5	9.8	8.0	20.3	23.0	8.2
...	...	...	...	...	...	...	...	...
195	O	F	46	21.4	18.0	41.2	46.2	18.7
196	O	F	47	21.7	17.1	41.7	47.2	19.6
197	O	F	48	21.9	17.2	42.6	47.4	19.5
198	O	F	49	22.5	17.2	43.0	48.7	19.8
199	O	F	50	23.1	20.2	46.2	52.5	21.1

200 rows × 8 columns

In [3]:
```
import pandas as pd
crabs_data = pd.DataFrame(crabs.data)
```

In [4]:
```
crabs_data.shape
```

Out[4]:
```
(200, 8)
```

A(1)a parametric procedure

Determine whether there is a significant difference between blue and orange crabs in mean carapace length (mm) [CL] using each of the following procedures:

In [5]:
```
import pandas as pd
from scipy import stats

blue_crabs = crabs_data[crabs_data['sp'] == 'B']['CL']
orange_crabs = crabs_data[crabs_data['sp'] == 'O']['CL']
t_statistic, p_value = stats.ttest_ind(blue_crabs, orange_crabs)

print("T-statistic:", t_statistic)
print("P-value:", p_value)
```

```
T-statistic: -4.237156779280929
P-value: 3.4675924330865825e-05
```

In [6]:
```
import numpy as np
confidence_level = 0.95
conf_interval = stats.t.interval(confidence_level, df=len(blue_crabs) + len(orange_cra

mean_blue_crabs = blue_crabs.mean()
mean_orange_crabs = orange_crabs.mean()
```

```
print(f"{confidence_level * 100}% Confidence Interval for the Difference in Means:", c
print("Sample Mean for Blue Crabs:", mean_blue_crabs)
print("Sample Mean for Orange Crabs:", mean_orange_crabs)
```

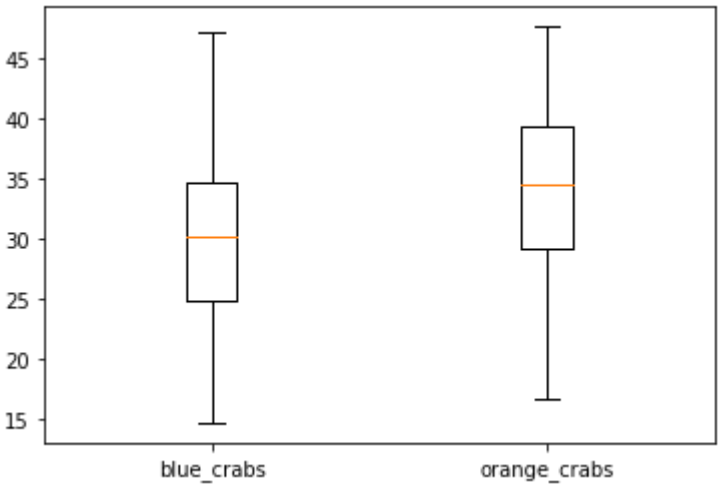
```
95.0% Confidence Interval for the Difference in Means: (nan, nan)
Sample Mean for Blue Crabs: 30.058000000000003
Sample Mean for Orange Crabs: 34.153
```

In [7]:
```
#variance test check for variance assumption
statistic, p_value = stats.levene(blue_crabs, orange_crabs)
p_value
```

Out[7]:
```
0.7484592344995541
```

In [8]:
```
#check for normality
import matplotlib.pyplot as plt
plt.boxplot([blue_crabs, orange_crabs], labels=["blue_crabs", "orange_crabs"])
```

Out[8]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x2068dbcbd30>,
  <matplotlib.lines.Line2D at 0x2068dbea040>,
  <matplotlib.lines.Line2D at 0x2068e296190>,
  <matplotlib.lines.Line2D at 0x2068e296460>],
 'caps': [<matplotlib.lines.Line2D at 0x2068dbea3d0>,
  <matplotlib.lines.Line2D at 0x2068dbea5e0>,
  <matplotlib.lines.Line2D at 0x2068e296730>,
  <matplotlib.lines.Line2D at 0x2068e296a00>],
 'boxes': [<matplotlib.lines.Line2D at 0x2068dbcba60>,
  <matplotlib.lines.Line2D at 0x2068dbeae80>],
 'medians': [<matplotlib.lines.Line2D at 0x2068dbea8b0>,
  <matplotlib.lines.Line2D at 0x2068e296cd0>],
 'fliers': [<matplotlib.lines.Line2D at 0x2068dbeab80>,
  <matplotlib.lines.Line2D at 0x2068e296fa0>],
 'means': []}
```



In [9]:
```
blue_crabs.shape, orange_crabs.shape
```

Out[9]:
```
((100,), (100,))
```

Answer:

We use two-sample t-test here.

The null hypothesis: true difference in means of blue and orange crabs' carapace length is equal to 0.

The alternative hypothesis: true difference in means of blue and orange crabs' carapace length is not equal to 0.

This two-sample t-test provides strong evidence that blue and orange crabs in mean carapace length (mm) are different (two-sided p-value=3.4675924330865825e-05).

Advancedly, we make variance test and get p_value at 0.7484592344995541, we fail to reject the null that true ratio of variance is equal to 1. The "equal variances" assumption for the t-test is considered valid. Similarly, from the boxplot, data of both groups are not skewed.

A(1)b non-parametric procedure

Determine whether there is a significant difference between blue and orange crabs in mean carapace length (mm) [CL] using each of the following procedures:

In [10]:
```
#Wilcoxon Rank-Sum Test
u_statistic, p_value = stats.mannwhitneyu(blue_crabs, orange_crabs, alternative='two-s
print("Wilcoxon Rank-Sum Test Statistic:", u_statistic)
print("P-value:", p_value)
```

```
Wilcoxon Rank-Sum Test Statistic: 3378.5
P-value: 7.468646222532377e-05
```

Answer

The null hypothesis: there is no difference in the distribution (location, typically median) of the two independent samples being compared.

This Wilcoxon Rank-Sum Test provides strong evidence that the distribution of blue and orange crabs in mean carapace length (mm) are different (two-sided p-value=7.468646222532377e-05).

A(1)C re-sampling procedure

In [19]:
```
n_resamples = 10000
resampled_diffs = np.zeros(n_resamples)
combined_data = np.concatenate([blue_crabs, orange_crabs])

# Bootstrap
for i in range(n_resamples):
    resample = np.random.choice(combined_data, size=len(combined_data), replace=True)
    resampled_diffs[i] = np.mean(resample[:len(blue_crabs)]) - np.mean(resample[len(bl

p_value = np.mean(np.abs(resampled_diffs) >= np.abs(observed_diff))
```

```
print("Observed Difference in Means:", observed_diff)
print("Bootstrap p-value:", p_value)
```

```
Observed Difference in Means: -4.094999999999995
Bootstrap p-value: 0.0001
```

Answer

Null hypothesis: there is no significant difference between blue and orange crabs in mean carapace length (mm).

This bootstrap resampling test provides strong evidence that there is significant difference between blue and orange crabs in mean carapace length (mm) (two-sided p-value=0.0001).

A (2) Discuss the assumptions underlying the analyses in (1) above, their validity, and any remedial measures to be taken.

Two-Sample T-Test

Assumptions:

- Normality for each group. Skewed from box plot.
- Homogenity of variance. Satisfied by variance test.

Validity:

- T-test is robust to deviations from normality and homogeneity of variance, especially with large sample sizes where larger than 30. Normality assumption will not be major the major concern even for the small sample size, as long as the skewness is same in two populations and sample size is roughly equal.
- Besides large sample size and same skewness and group size for both two groups, if normality or homogenity is invalid, the result of two-sample t-test is going to be unreliable.

Remedial Measures:

- By variance test and box plot we have above, two assumptions for two-sample t-test are not violated. We can trust the result of our two-sample t-test. No further actions needed currently.
- If normality is violated, consider using non-parametric tests like the Wilcoxon Rank-Sum Test or bootstrapping.
- If variance is violated, consider using Welch's t-test, which is robust to unequal variances.
- If there are outliers or numbers in different scales, transforming the data like log.

Wilcoxon Rank-Sum Test

Assumptions:

- The values are independent and identically distributed.
- Two groups are independent of each other.
- Ordinal Scale.(rankable)
- The populations from which the two samples were taken differ only in location. That is, the populations may differ in their means or medians, but not in their dispersions or distributional shape (such as skewness)

Validity:

- The Wilcoxon Rank-Sum Test is non-parametric without assumptions about the shape of the distribution or homogeneity of variances.

Remedial Measures:

- In our example above, the collection of crabs in two colors and their carapace length are seems independent.
- If dependence like pair-wise data, using paired t-test or Wilcoxon signed-rank test.

Bootstrap

Assumptions:

- No specific assumptions about data distribution.
- the validity of bootstrap results largely depends on the quality and representativeness of the original data.

Validity:

- Bootstrap is robust when the associated distributions are not tractable.

Remedial Measures:

- In our example, result of bootstrap shows the data do not provide support for a significant difference of carapace length between two colored crab goups. The fair different result may contributed by substantial overlap of data points for bootstrap incresing variability.
- If result not ideal, change data of original sample or try to increase the number of bootstrap resamples for better accuracy.
- Try out permutation test or Jackknife resampling.

B) Consider the ToothGrowth data in R, concerning the Effect of Vitamin C on Tooth Growth in Guinea Pigs.

(3) Assume that if “len” is above 20, it is classified as “HIGH”; and ”LOW”, otherwise. Ignore ‘dose’, and determine whether there is a significant difference in the proportions of the two groups classified as “HIGH” using a suitable test and a 95% confidence interval.

In [22]:
```
import statsmodels.api as sm
Tooth = sm.datasets.get_rdataset('ToothGrowth','datasets')
Tooth.data
```

Out[22]:

	len	supp	dose
0	4.2	VC	0.5
1	11.5	VC	0.5
2	7.3	VC	0.5
3	5.8	VC	0.5
4	6.4	VC	0.5
5	10.0	VC	0.5
6	11.2	VC	0.5
7	11.2	VC	0.5
8	5.2	VC	0.5
9	7.0	VC	0.5
10	16.5	VC	1.0
11	16.5	VC	1.0
12	15.2	VC	1.0
13	17.3	VC	1.0
14	22.5	VC	1.0
15	17.3	VC	1.0
16	13.6	VC	1.0
17	14.5	VC	1.0
18	18.8	VC	1.0
19	15.5	VC	1.0
20	23.6	VC	2.0
21	18.5	VC	2.0
22	33.9	VC	2.0
23	25.5	VC	2.0
24	26.4	VC	2.0
25	32.5	VC	2.0
26	26.7	VC	2.0
27	21.5	VC	2.0
28	23.3	VC	2.0
29	29.5	VC	2.0
30	15.2	OJ	0.5
31	21.5	OJ	0.5
32	17.6	OJ	0.5
33	9.7	OJ	0.5

	len	supp	dose
34	14.5	OJ	0.5
35	10.0	OJ	0.5
36	8.2	OJ	0.5
37	9.4	OJ	0.5
38	16.5	OJ	0.5
39	9.7	OJ	0.5
40	19.7	OJ	1.0
41	23.3	OJ	1.0
42	23.6	OJ	1.0
43	26.4	OJ	1.0
44	20.0	OJ	1.0
45	25.2	OJ	1.0
46	25.8	OJ	1.0
47	21.2	OJ	1.0
48	14.5	OJ	1.0
49	27.3	OJ	1.0
50	25.5	OJ	2.0
51	26.4	OJ	2.0
52	22.4	OJ	2.0
53	24.5	OJ	2.0
54	24.8	OJ	2.0
55	30.9	OJ	2.0
56	26.4	OJ	2.0
57	27.3	OJ	2.0
58	29.4	OJ	2.0
59	23.0	OJ	2.0

In [23]:
```
tooth_data = pd.DataFrame(Tooth.data)
tooth_data.shape
```

Out[23]:
```
(60, 3)
```

In [26]:
```
tooth_data['degree'] = np.where(tooth_data['len'] > 20, 'HIGH', 'LOW')
contingency_table = pd.crosstab(tooth_data['supp'], tooth_data['degree'])
contingency_table
```

Out[26]:

degree	HIGH	LOW
supp		
OJ	18	12
VC	10	20

In [25]:
```
tooth_data.head(3)
```

Out[25]:

	len	supp	dose	degree
0	4.2	VC	0.5	LOW
1	11.5	VC	0.5	LOW
2	7.3	VC	0.5	LOW

In [27]:
```
#chi-squared test
# we have all frequncy in the table larger than 5, thus no need for Yates' correction
chi2, p, dof, expected = stats.chi2_contingency(contingency_table,correction=False)
print("Chi-squared statistic:", chi2)
print("P-value:", p)
print("Degrees of Freedom:", dof)
print("Expected frequency table:\n", expected)
```

```
Chi-squared statistic: 4.285714285714286
P-value: 0.03843393023678176
Degrees of Freedom: 1
Expected frequency table:
 [[14. 16.]
 [14. 16.]]
```

In [29]:
```
#difference in proportions
proportion_VC_HIGH = contingency_table.loc['VC', 'HIGH'] / contingency_table.loc['VC']
proportion_OJ_HIGH = contingency_table.loc['OJ', 'HIGH'] / contingency_table.loc['OJ']
diff_proportions = proportion_VC_HIGH - proportion_OJ_HIGH

#SE
se_diff_proportions = np.sqrt(
    (proportion_VC_HIGH * (1 - proportion_VC_HIGH) / contingency_table.loc['VC'].sum()
    (proportion_OJ_HIGH * (1 - proportion_OJ_HIGH) / contingency_table.loc['OJ'].sum()
)

#Margin of error for a 95%CI
z_stats = stats.norm.ppf(0.975)  # 1.96 for a 95% confidence interval
margin_of_error = z_stats * se_diff_proportions

# Confidence interval
ci_lower = diff_proportions - margin_of_error
ci_upper = diff_proportions + margin_of_error

print("95% Confidence Interval for Difference in Proportions:", (ci_lower, ci_upper))
```

```
95% Confidence Interval for Difference in Proportions: (-0.5099502893978591, -0.02338
304393547419)
```

Assumptions: independence are satisfied because we use treatment of VC or OJ for every individual Guinea Pig. Moreover, from the expected frequencies table, all of them are greater

than 5. Frequency assumption is satisfied and no need for Yates' correction.

Null of chi-squared test: there is no significant difference in the proportions of "HIGH" and "LOW" groups based on the "len".

We have suggestive evidence to conclude that there is a significant difference in proportions between "HIGH" and "LOW" groups based on the "len" column(two-sided p-value 0.03843393023678176).

There is a noticeable significant difference in proportions between the groups that we can detect in 95% cases, where the difference in proportions is expected to fall in (-0.5099502893978591, -0.02338304393547419).