(b)

Hypothesis: California(CA) and Texas(TX) have the most significant difference in their charges in DRG Charges 190. And CA's average charges is greater than TX's average charges.

H0: Average charges in DRG Charges 190 in CA and TX are the same.
H1： CA's average charges in DRG Charges 190 is greater than TX's average charges.

In [17]:

```
import scipy.stats as st

df1_hypo = df_6states[['DRG Charges 190','Provider State']]
df1_CA = df1_hypo[df1_hypo['Provider State']=='CA']['DRG Charges 190']
df1_TX = df1_hypo[df1_hypo['Provider State']=='TX']['DRG Charges 190']
df1_CA = df1_CA.dropna()
df1_TX = df1_TX.dropna()

t,p = st.ttest_ind(df1_CA, df1_TX)
print("t statistic: " + str(t))
print("p-value: " + str(p))

# Proform one-sided test and use significant value 0.05
if p < 0.05/2:
    print("Reject H0.")
else:
    print("Accept H0.")
```

```
t statistic: 9.804105998394789
p-value: 1.3330879399800308e-20
Reject H0.
```

Proform **one-sided** test and use significant value **0.05**.
Since p-value < 0.05/2, null hypothesis H0 is rejected. Therefore, we tentatively conclude H1 to be the case, which support the claim.

(c)

Hypothesis: Pennsylvania(PA) and Georgia(GA) have the significant difference in their charges in DRG Charges 190.

H0: Average charges in DRG Charges 190 in PA and GA are the same.
H1： Average charges in DRG Charges 190 in PA and GA are the different.

In [25]:

```
df2_hypo = df_6states[['Provider State','DRG Charges 190','DRG Charges 392','DRG Charges 871']]

df2_PA = df2_hypo[df2_hypo['Provider State']=='PA']
df2_PA = df2_PA.dropna()
df2_PA = pd.concat([df2_PA['DRG Charges 190'], df2_PA['DRG Charges 392'], df2_PA['DRG Charges 871']], ignore_index=True)

df2_GA = df2_hypo[df2_hypo['Provider State']=='GA']
df2_GA = df2_GA.dropna()
df2_GA = pd.concat([df2_GA['DRG Charges 190'], df2_GA['DRG Charges 392'], df2_GA['DRG Charges 871']], ignore_index=True)

length = min(len(df2_PA), len(df2_GA))
df2_GA = df2_GA.sample(n = length, random_state=3)
df2_PA = df2_PA.sample(n = length, random_state=3)

t_rel, p_rel = st.ttest_rel(df2_GA, df2_PA)
print("Two sample paired Student's t-test.")
print("t statistic(paired): " + str(t_rel))
print("p-value(paired): " + str(p_rel))
if p_rel < 0.05:
    print("Reject H0.")
else:
    print("Accept H0.")
```

```
Two sample paired Student's t-test.
t statistic(paired): -3.2213719626126287
p-value(paired): 0.0014578800424925055
Reject H0.
```

Proform **two sample paired Student's t-test** and use significant value **0.05**.
Since p-value < 0.05, null hypothesis H0 is rejected. Therefore, we tentatively conclude H1 to be the case, which support the claim.

In [24]:

```
t_ind, p_ind = st.ttest_ind(df2_GA, df2_PA)
print("\nTwo sample unpaired t-test(two sided).")
print("t statistic(unpaired): " + str(t_ind))
print("p-value(unpaired): " + str(p_ind))
if p_ind < 0.05:
    print("Reject H0.")
else:
    print("Accept H0.")
```

```
Two sample unpaired t-test(two sided).
t statistic(unpaired): -3.133704313100077
p-value(unpaired): 0.0018350152260808122
Reject H0.
```

Proform **two sample unpaired t-test(two sided)** and use significant value **0.05**.
Since p-value < 0.05, null hypothesis H0 is rejected. Therefore, we tentatively conclude H1 to be the case, which support the claim.

As is shown above, paired t-test gets p-value slightly less then unpaired t-test, which means it's more likely to reject H0.