






Out[139]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	124517	William Van Dijk	M	23.0	185.0	65.0	Belgium	BEL	1984 Summer	1984.0	Summer	Los Angeles	Athletics	Athletics Men's 3,000 metres Steeplechase	0
1	124517	William Van Dijk	M	27.0	185.0	65.0	Belgium	BEL	1988 Summer	1988.0	Summer	Seoul	Athletics	Athletics Men's 3,000 metres Steeplechase	0
2	124517	William Van Dijk	M	31.0	185.0	65.0	Belgium	BEL	1992 Summer	1992.0	Summer	Barcelona	Athletics	Athletics Men's 3,000 metres Steeplechase	0
3	124518	Daniel "Daan" van Dijk	M	21.0	NaN	NaN	Netherlands	NED	1928 Summer	1928.0	Summer	Amsterdam	Cycling	Cycling Men's Tandem Sprint, 2,000 metres	15
4	124519	Everdina "Edin" van Dijk	F	35.0	176.0	66.0	Netherlands	NED	2008 Summer	2008.0	Summer	Beijing	Swimming	Swimming Women's 10 kilometres Open Water	0

Гипотеза H1: Среднее выступлений за 2008 год атлетов из США значимо отличается в большую сторону, чем среднее выступлений атлетов из Германии. Можно использовать z-тест для сравнения двух выборок. Выборка нормально распределена (размер выборки больше 30) и известна дисперсия генеральной совокупности



In [117]:

data.	Medal
0	0
1	0
2	0
3	15
4	0
..	0
22582	0
22583	0
22584	0
22585	0

Out[117]:

Гипотеза H1: Среднее выступлений за 2008 год атлетов из США значительно отличается в большую сторону, чем среднее выступлений атлетов из Германии. Можно использовать z-тест для сравнения двух выборок. Выборка нормально распределена (размер выборки больше 30) и известна дисперсия генеральной совокупности

📄SNOWFALL

```
In [117]: data.Medal
Out[117]:
0      0
1      0
2      0
3     15
4      0
22582  0
22583  0
22584  0
22585  0
22586  0
Name: Medal, Length: 271891, dtype: int64

In [118]: population_usa = data[(data['Sport'] == 'Athletics') & (data['NOC'] == 'USA') & (data['Year'] == 2008)]

In [119]: population_usa = data[(data['Sport'] == 'Athletics') & (data['NOC'] == 'GER') & (data['Year'] == 2008)]

Посчитаем стандартные отклонения в генеральных совокупностях, ведь мы их знаем

In [122]: sd_ger = population_ger["Medal"].std()
sd_usa = population_usa["Medal"].std()

In [123]: import math

Сгенерируем случайные выборки (n = 40) из обоех ген совокупностей

In [140]: np.random.seed(42)

In [124]: idx_ger = np.random.choice(population_ger.index, 40)
idx_usa = np.random.choice(population_usa.index, 40)

In [127]: sample_ger = population_ger.loc[idx_ger, :]
sample_usa = population_usa.loc[idx_usa, :]

In [130]: mean_ger = sample_ger["Medal"].mean()
mean_usa = sample_usa["Medal"].mean()

Считаем z-статистику

In [137]: Z = (mean_usa - mean_ger)/(((sd_usa**2)/40 + (sd_ger**2)/40)**0.5)

In [220]: if Z >= abs(scipy.stats.norm.ppf(0.05)):
print("Различие стат значимо")

Различие стат значимо

для alpha 0.05 -> Z >= 1.645

Можно сказать что среднее выступлений атлетов из Америки стат значимо больше чем среднее выступлений спортсменов из Германии.

Тест №2 Сейчас я хочу рассмотреть доверительный интервал разности средних двух выборок, зная стандартные отклонения их генеральных совокупностей. Каждая выборка имеет нормальное распределение. Выборки независимы и случайно распределены. Найду 99 процентный доверит интервал, разности выступлений двух сборных команд США по плаванию за разные года.
```

📄SNOWFALL

```
In [224]: analyze_ds.ipynb photo_2022-11-24_14-11-55.jpg

In [150]: data.head()
Out[150]:
   ID  Name Sex  Age  Height  Weight  Team  NOC  Games  Year  Season  City  Sport  Event  Medal
0  124517 William Van Dijk  M   23.0   185.0   65.0  Belgium  BEL  1984 Summer  1984.0  Summer  Los Angeles  Athletics  Men's 3,000 metres Steeplechase  0
1  124517 William Van Dijk  M   27.0   185.0   65.0  Belgium  BEL  1988 Summer  1988.0  Summer  Seoul  Athletics  Men's 3,000 metres Steeplechase  0
2  124517 William Van Dijk  M   31.0   185.0   65.0  Belgium  BEL  1992 Summer  1992.0  Summer  Barcelona  Athletics  Men's 3,000 metres Steeplechase  0
3  124518 Daniel "Daan" van Dijk  M   21.0   NaN   NaN  Netherlands  NED  1928 Summer  1928.0  Summer  Amsterdam  Cycling  Cycling Men's Tandem Sprint, 2,000 metres  15
4  124519 Everdina "Edin" van Dijk  F   35.0   176.0   66.0  Netherlands  NED  2008 Summer  2008.0  Summer  Beijing  Swimming  Swimming Women's 10 kilometres Open Water  0

In [160]: idx_2009 = (data["Sport"] == "Swimming") & (data["NOC"] == "USA") & (data["Year"] == 2009)
idx_2004 = (data["Sport"] == "Swimming") & (data["NOC"] == "USA") & (data["Year"] == 2004)

Получим генеральные совокупности

In [161]: population_2009 = data.loc[idx_2009, :]
population_2004 = data.loc[idx_2004, :]

Получим стандартные отклонения из генеральных совокупностей

In [198]: std_2009 = population_2009["Medal"].std()
std_2004 = population_2004["Medal"].std()

Получим random sample

In [199]: idx_2009 = np.random.choice(population_2009.index, 50)
idx_2004 = np.random.choice(population_2004.index, 50)
sample_2009 = population_2009.loc[idx_2009, :]
sample_2004 = population_2004.loc[idx_2004, :]

Вычислим среднее из random sample

In [200]: avg_2009 = sample_2009["Medal"].mean()
avg_2004 = sample_2004["Medal"].mean()

In [201]:

In [202]: alpha = 0.01

In [203]: z_alpha = scipy.stats.norm.ppf(alpha/2)

In [210]: left_bound = avg_2009 - avg_2004 + z_alpha * (std_2009**2/50 + std_2004**2/50)**0.5

In [211]: right_bound = avg_2009 - avg_2004 - z_alpha * (std_2009**2/50 + std_2004**2/50)**0.5

In [215]: print(f"({left_bound} < true_avg_2009 - true_avg_2004 < {right_bound})")
-4.539981215401145 < true_avg_2009 - true_avg_2004 < 1.8839223918717316

In [217]: left_bound / (left_bound - right_bound)
Out[217]: 0.7963218131893735

Можно сделать вывод что в 70 процентах случаев среднее успешности выступлений сборной команды в 2004 году больше чем в 2000

In [ ]:
```