بسم الله الرحمن الرحیم

پروژه یک درس مبانی علوم داده

دکتر بهرک و دکتر یعقوب زاده

محمد امین توانایی - ۸۱۰۱۰۱۳۹۶

سید علی تهامی - ۸۱۰۱۰۱۳۹۷

مهدی وجهی - ۸۱۰۱۰۱۵۵۸

فهرست

مقدمه

اولین بخش این پروژه، به شبیه‌سازی مونت‌کارلو اختصاص دارد. شبیه‌سازی مونت‌کارلو یک روش عددی برای حل مسائلی است که در آن‌ها استفاده از روش‌های تحلیلی یا رسیدن به یک پاسخ انتگرالی بسیار دشوار یا غیرممکن است. در این بخش، ما از این روش برای محاسبه تقریبی عدد پی و همچنین تحلیل احتمال برد در بازی منچ استفاده می‌کنیم.

در ادامه، به بررسی قضیه حد مرکزی (Central Limit Theorem) که یکی از نتایج بنیادین در آمار و احتمالات است، می‌پردازیم. این قضیه، پایه‌ای برای بسیاری از روش‌های آماری مانند آزمون فرض است. ما با استفاده از توزیع‌های احتمال مختلف، اثر قضیه حد مرکزی را مشاهده و نتایج آن را تحلیل می‌کنیم.

در نهایت، به آزمون‌های فرض می‌پردازیم که ابزارهای مهمی برای تحلیل داده‌ها و اتخاذ تصمیمات آگاهانه در پژوهش‌های علمی هستند. در این بخش، با استفاده از آزمون فاصله اطمینان و p-value، فرضیه عادلانه بودن یک سکه را آزمایش می‌کنیم. سپس از آزمون t برای بررسی این فرض که داشتن شغل در دوران دانشجویی بر نمرات تاثیر منفی می‌گذارد، استفاده می‌کنیم.

شبیه ساز مونته کارلو

تخمین عدد پی

شرح مساله

در این قسمت از ما خواسته شده تا با دادن نقاط رندوم در یک مربع ۲×۲ و محاط کردن یک دایره در آن ببینیم نسبت نقاطی که در دایره هستند به کل نقاط چقدر است.
در واقع با این کار به تخمین مساحت دایره ای به شعاع یک که طبق فرمول S=r*(pi)^2 برابر با عدد پی است پرداخته شده است.اگر تعداد کل نقاط تخمینی از مساحت مربع باشند که برابر با ۴ است تعداد نقاط درون دایره به همان نسبت تخمینی از عدد پی هستند که با قرار دادن دو کسر با هم مقدار عدد پی تخمین زده خواهد شد.
(برای چک کردن اینکه نقطه ای درون دایره است باید چک کرد فاصله آن از مرکز کمتر از شعاع باشد)

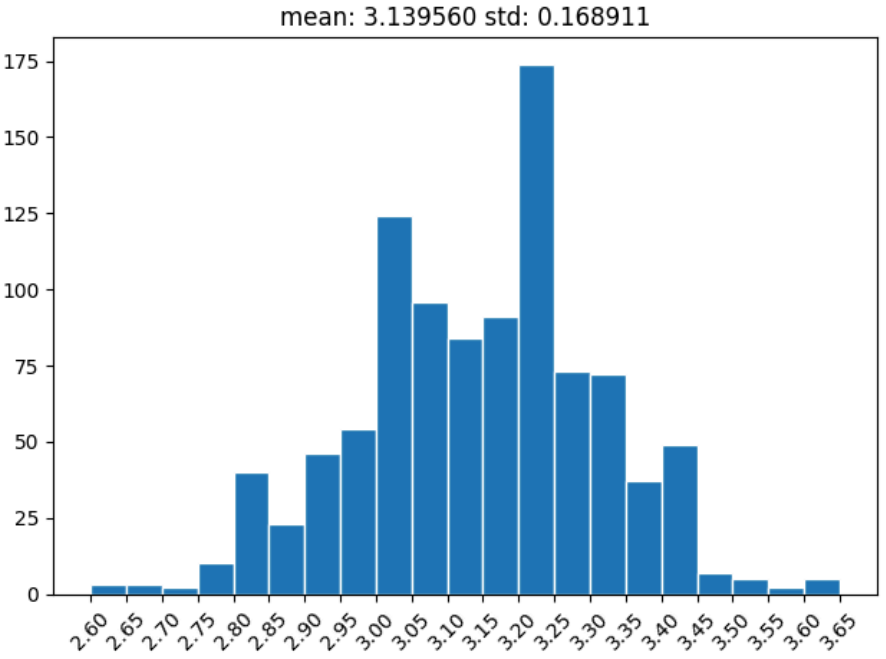
```
def estimate_Pi(x_lim,y_lim,num_of_gen):
    inside_circle_num = 0
    radius = (x_lim[1] - x_lim[0])/2
    points = generate_random_point(x_lim,y_lim,num_of_gen)
    for i in range(len(points)):
        if(distance(points[i],(0,0)) <= radius):
            inside_circle_num += 1
    return (inside_circle_num/len(points))*4
```

این تابع پی را با تولید num_of_gen نقطه تخمین میزند.

تحلیل داده

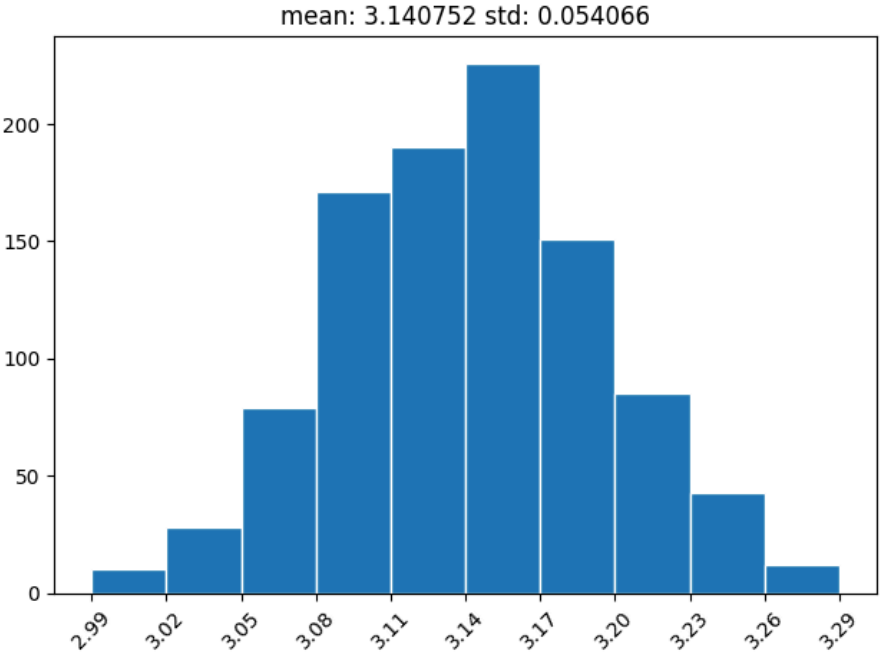
در این بخش من ۱۰۰۰ بار الگوریتم را با تولید تعداد نقاط مشخص اجرا و نمودار های آنها را نشان داده ام

۱.در این نمودار پی با تولید ۱۰۰ نقطه و ۱۰۰۰ بار اجرای الگوریتم نشان داده شده است.



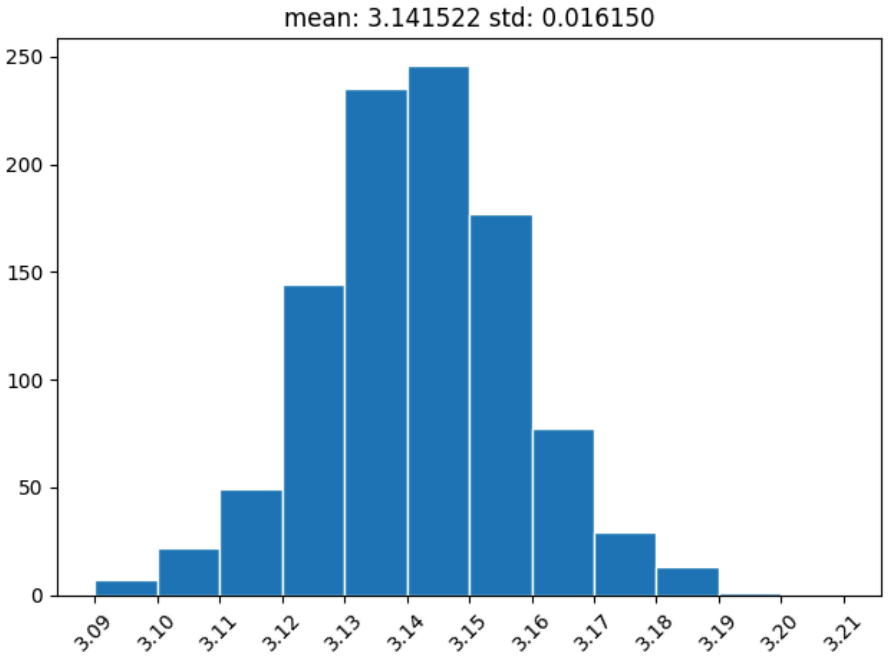
دیده می شود علاوه بر اینکه تخمین تخمین درست نیست پراکندگی نیز زیاد است.

۲.برای همین برای همین الگوریتم را با تولید ۱۰۰۰ نقطه انجام دادم نتایج بصورت زیر است

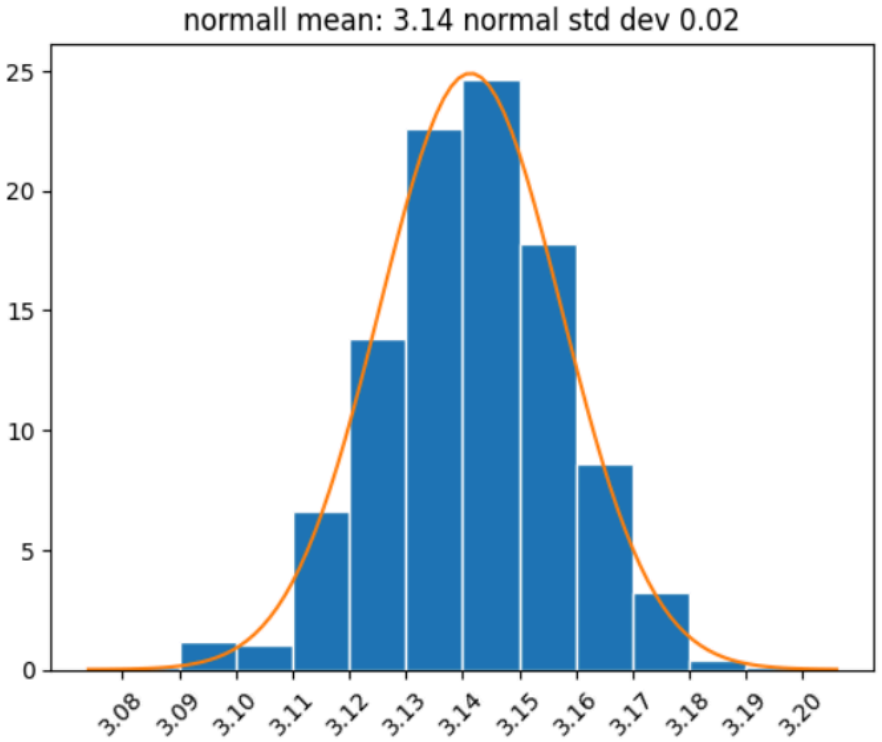


دیده میشود که پراکندگی و دقت بهتر شده است.

۳.تولید ۱۰۰۰۰ نقطه



پراکندگی به شدت کاهش یافته و دقت هم بهتر شده است.نکته دیگری که در این نمودارها دیده شده است اینست که تخمین پی بر اساس این روش گویا از توزیع نرمال پیروی می کند زیرا نمودار ها به نمودار نرمال نزدیک میشوند. برای همین رویی این نمودار یک توزیع نرمال نیز فیت کردم.



تخمین برنده در بازی منچ

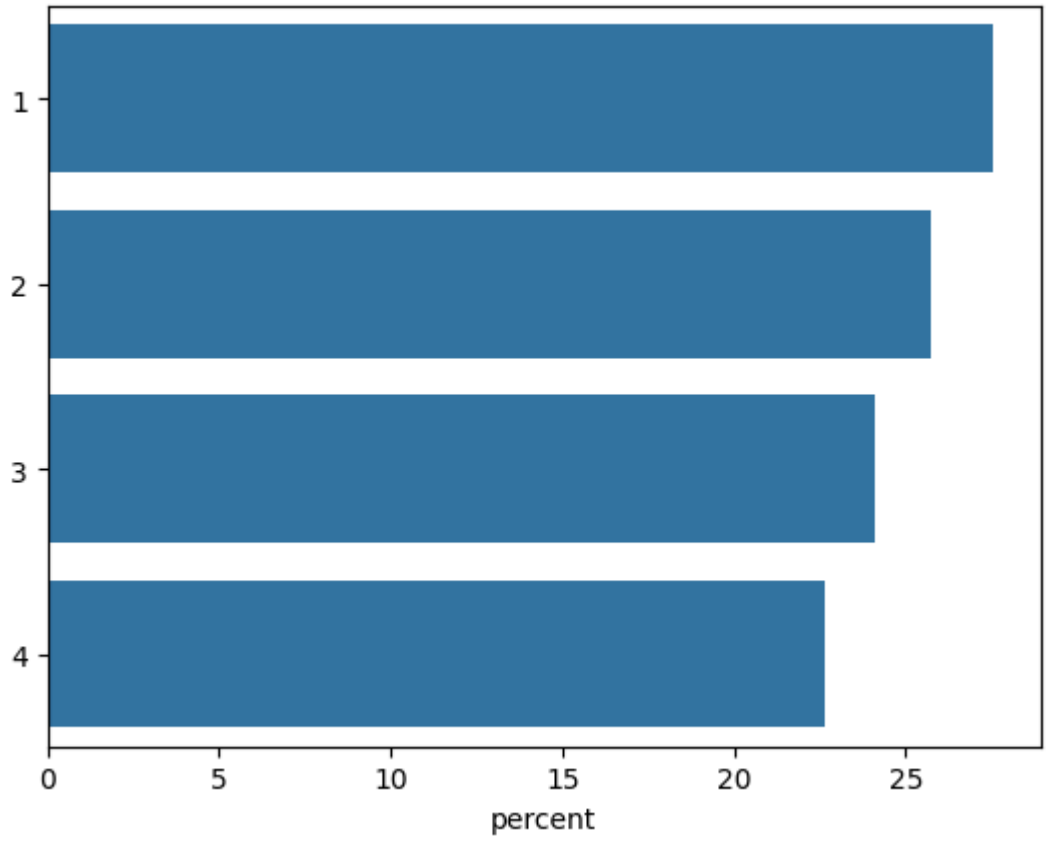
در این قسمت خواسته شده است که با اجرای بازی منچ با یک مهره برای هر کس ببینیم درصد برد هر نفر چقدر است.

توضیح بازی

بازی به اینصورت است که هر نفر در نوبتش تاس می اندازد اگر مهره اش درون صفحه بازی نباشد باید برای ورود به بازی شش بیاورد و اگر مهره درون بازی باشد به اندازه مقدار تاس جلو میرود تا به خانه امن خودش برسد و ببرد و اگر مهره ای در خانه ای باشد و فرد دیگری روی آن خانه بیاید مهره زیرین حذف میشود و باید دوباره فرد مورد نظر شش بیاورد تا بتواند مهره اش را حرکت بدهد.
هر کس که زودتر مهره اش به خانه امن برسد برنده است.

تحلیل نتایج

نتایج پس از یک میلیون بار اجرای بازی منچ به صورت زیر است




```
the win probability of first player is:  0.275151
the win probability of second player is 0.256867
the win probability of third player is 0.241548
the win probability of fourth player is 0.226434
```

این نمودار نشان دهنده اینست که با اینکه همه تقریبا شانس یکسانی برای برد دارند اما هر کس بازی را زودتر شروع کند احتمال برنده شدنش بیشتر است.نتایج نشان دهنده اینست که بازی کمی ناعادلانه است .این ناعادلانگی از قانون بازی نشات میگیرد که هر کس مهره اش به خانه امن برسد در جا بازی تمام میشود در صورتی که باید صبر کنیم تا نفر آخر هم تاسش را بیاندازد و بعد بازی تمام شود در واقع چون نفر اول ، بازی را زودتر شروع احتمالا زودتر هم به خانه امن برسد.

قضیه حد مرکزی - Central Limit Theorem

شبیه سازی قضیه حد مرکزی

شرح مساله

در این بخش، می خواهیم با استفاده از سه توزیع متفاوت، قضیه حد مرکزی را مورد بررسی قرار دهیم و نتیجه آن را مشاهده کنیم.

توضیح مساله و کد

برای شبیه سازی و دیدن نتایج، ما از سه توزیع نرمال، برنولی و یونیفرم استفاده خواهیم کرد. برای هر یک از این سه توزیع، پارامتر های مورد نیاز را مقدار دهی خواهیم کرد.

```
exponential_params = {"scale": 1}
bernoulli_params = {"p": 0.5}
uniform_params = {"loc": 0, "scale": 1}


distributions = {
    "Exponential": expon(**exponential_params),
    "Bernoulli": bernoulli(**bernoulli_params),
    "Uniform": uniform(**uniform_params)
}
```

برای اینکه بتوانیم به شبیه سازی و نتیجه درست برای این قضیه برسیم، سایز نمونه های مان را اعدادی با اختلاف مناسب انتخاب میکنیم؛ در اینجا ما از نمونه های 1،10،100،500،1000 استفاده خواهیم کرد و نمودار مربوط به هر شکل را رسم میکنیم.

```
 for sample_size in sample_sizes:
        sample_means = []

        # Generate the samples and calculate their means
        for _ in range(num_samples):
            sample = dist.rvs(size=sample_size)
            sample_means.append(np.mean(sample))


        fig, axs = plt.subplots(1, 2, figsize=(10, 5))
```

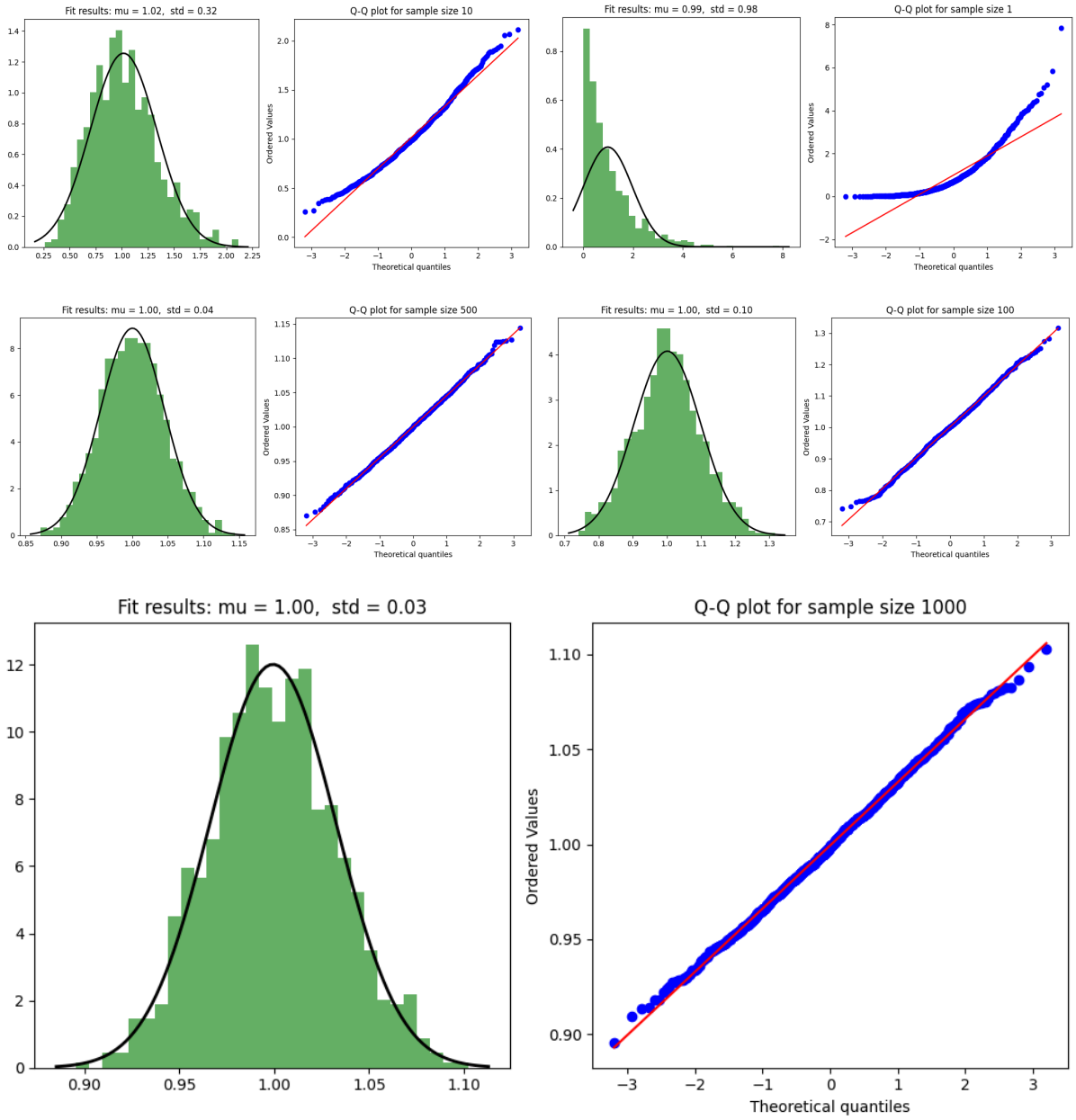
```
        axs[0].hist(sample_means, bins=30, density=True, alpha=0.6,
color='g')


        mu, std = norm.fit(sample_means)
        xmin, xmax = axs[0].get_xlim()
        x = np.linspace(xmin, xmax, 100)
        p = norm.pdf(x, mu, std)
        axs[0].plot(x, p, 'k', linewidth=2)
        title = f"Fit results: mu = {mu:.2f},  std = {std:.2f}"
        axs[0].set_title(title)


        probplot(sample_means, dist="norm", plot=axs[1])
        axs[1].set_title(f"Q-Q plot for sample size {sample_size}")


        plt.tight_layout()
        plt.show()
```

نتایج و نمودار ها

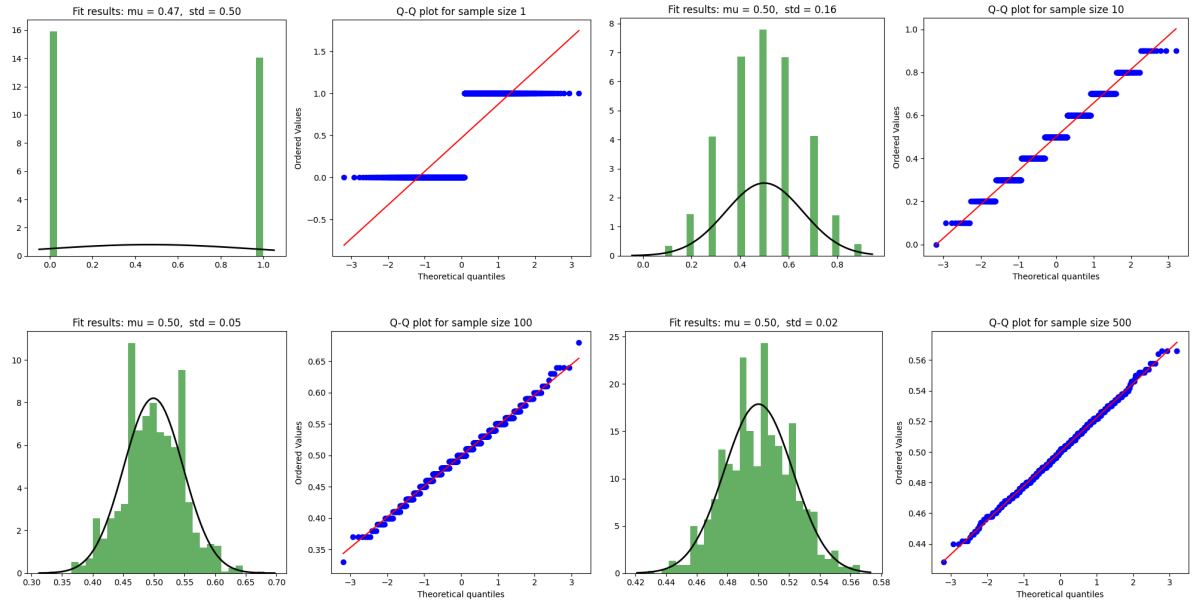
برای اینکه ببینیم توزیع های ما از توزیع نرمال پیروی میکنند یا خیر، از نمودار Q-Q استفاده خواهیم کرد.
اگر نقاط در نمودار Q-Q روی یک خط راست قرار بگیرند، این نشان می‌دهد که داده‌ها از توزیع مورد نظر پیروی می‌کنند. در این مورد، اگر نقاط روی خط y=x قرار بگیرند، این نشان می‌دهد که میانگین نمونه‌ها از یک توزیع نرمال پیروی می‌کند.

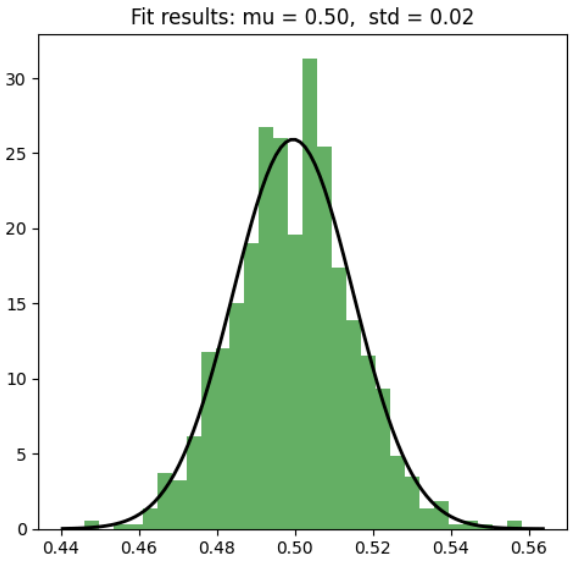
با افزایش اندازه نمونه، نقاط باید به خط y=x نزدیک‌تر شوند. این نشان می‌دهد که با افزایش اندازه نمونه، توزیع میانگین نمونه‌ها به یک توزیع نرمال نزدیک می‌شود، که این دقیقاً همان چیزی است که قانون حد مرکزی پیش‌بینی می‌کند. اگر نقاط از خط y=x فاصله داشته باشند، این نشان می‌دهد که توزیع میانگین نمونه‌ها از توزیع نرمال فاصله دارد. این ممکن است به دلیل اندازه نمونه کوچک یا توزیع غیر نرمال اولیه باشد.

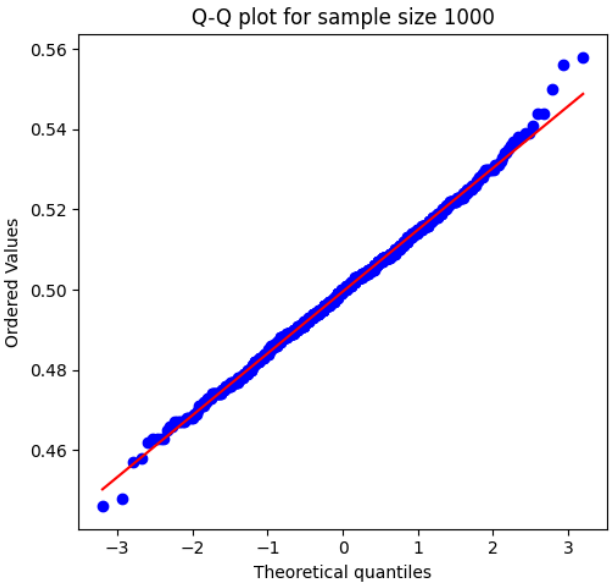
نمودار های توزیع نمایی



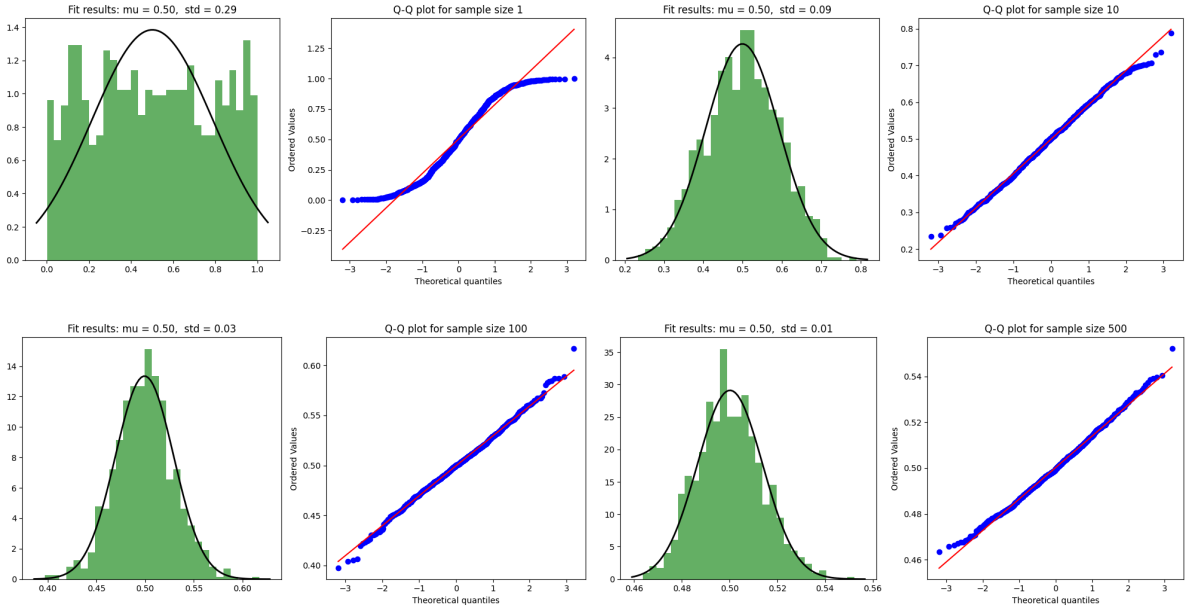
نمودار های توزیع برنولی

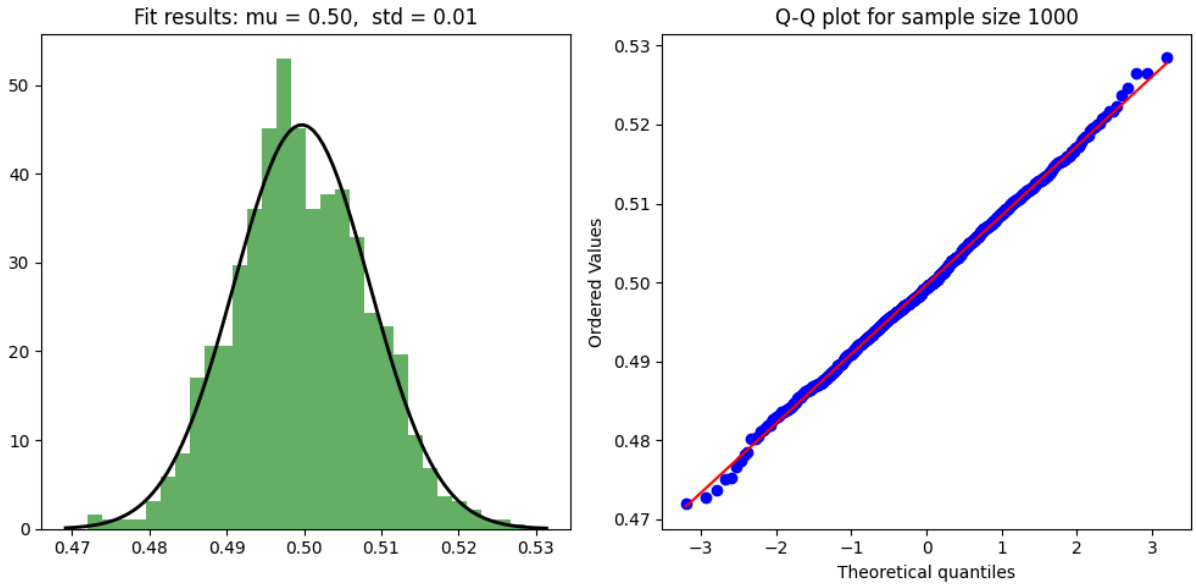






نمودار های توزیع یونیفرم





تحلیل نتایج

در این قسمت ما با استفاده از نمودار های هیستوگرام و Q-Q قضیه حد مرکزی را بررسی خواهیم کرد.
همانگونه که در نمودار ها مشاهده میشود، هر چقدر سایز نمونه ما افزایش می‌یابد، نمودار Q-Q ما به خط x=y نزدیک تر می‌شود و این نشانه از این دارد که توزیع ما به مقدار بیشتری به توزیع نرمال میل کرده است.

با توجه به نتایج اجرای کد، چند نکته مهم قابل مشاهده است:

1. توزیع میانگین نمونه به توزیع نرمال نزدیک می شود: با افزایش حجم نمونه، توزیع میانگین نمونه به توزیع نرمال نزدیک می شود. این یکی از نتایج اصلی قانون حد مرکزی است.
2. افزایش دقت با افزایش حجم نمونه: با افزایش حجم نمونه، توزیع میانگین نمونه ها به میانگین جامعه نزدیکتر می شود. این نشان می دهد که با افزایش حجم نمونه، برآورد ما از میانگین جامعه دقیق تر می شود.
3. کاهش انحراف معیار با افزایش حجم نمونه: با افزایش حجم نمونه، انحراف معیار میانگین توزیع نمونه ها کاهش می یابد. این نشان می دهد که با افزایش حجم نمونه، توزیع میانگین نمونه متمرکزتر می شود، به این معنی که برآورد ما از میانگین جامعه قابل اعتمادتر است.
4. استقلال از توزیع اولیه: قانون حد مرکزی برای هر توزیع احتمال برقرار است، خواه این توزیع نرمال باشد یا نباشد. در این آزمایش، ما این را برای سه توزیع مختلف مشاهده کردیم: نرمال، برنولی و یکنواخت.

این نتایج نشان می دهد که چگونه قانون حد مرکزی می تواند در عمل به ما کمک کند تا برآوردهای دقیق تری از پارامترهای جمعیت بر اساس آمار نمونه بدست آوریم.

آزمون فرض

سکه ناعادلانه

در این قسمت باید سکه ای ناعادلانه (۵۵٪, ۴۵٪) طراحی کرد و سپس با آزمون فرض بررسی کرد که سکه سالم است یا خیر.

شبیه سازی سکه

در این بخش تنها کافیست که با دادن احتمالات به تابع انتخاب شانسی شبیه سازی خودمان را انجام دهیم.

```
ONE_CHANCE = 0.55
def flip_coin(sample_size:int)->list[int]:
  return np.random.choice([1, 0], sample_size, p=[ONE_CHANCE, 1-ONE_CHANCE])
```

خروجی نمونه

```
array([0, 1, 1, 1, 0, 0, 1, 0, 1, 0])
```

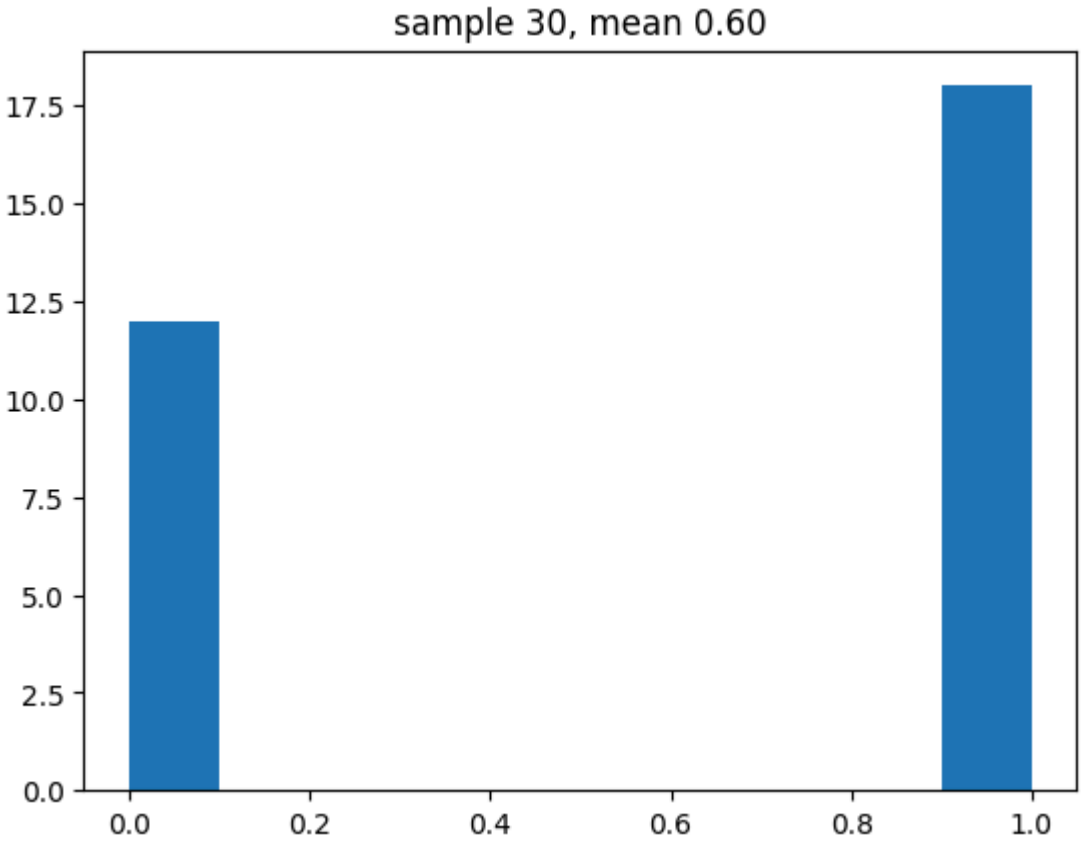
تنظیم آزمون فرض

آزمون فرض را به شکل زیر تشکیل می دهیم.

$$H_0: p_{tail} = \frac{1}{2}$$
$$H_A: p_{tail} \neq \frac{1}{2}$$

نمونه گیری

با استفاده از تابع شبیه سازی که نوشته بودیم نمونه گیری می کنیم. نتیجه به شکل زیر است:



بررسی شرایط آزمون

به بررسی شرایط آزمون فرض می پردازیم.

- **استقلال نمونه ها**: برقرار بودن این شرط در پرتاب سکه بدیهی می باشد و واضح است که پرتاب ها از یکدیگر مستقل هستند.
- **اندازه و چولگی نمونه**: به نظر میاد طبق نمونه گیری اولیه ای که کردیم یک مقداری چولگی داریم و اندازه نمونه باید از ۳۰ تا بیشتر باشد و ۳۰ کافی نیست.

محاسبه p-value

با فرمول های زیر p value را حساب می کنیم.

$$\bar{x} \sim N(\mu = 0.5, SE)$$

$$p\,value = p(Z > z) + p(Z <- z) = 2p(Z > z)$$

$$z = \frac{\bar{x} - \mu}{SE}$$

$$SE = \frac{s}{\sqrt{n}}$$

تابع مربوط به آن به صورت زیر است:

```
def calculate_p_value(sample:list[bool]):
  n = len(sample)
  x_bar = np.mean(sample)
  s = np.std(sample)
  se = s / (n ** 0.5)
  z = abs(H0_VALUE - x_bar) / se
  p_value = 2 * st.norm.cdf(-z)
  confidence_interval = (x_bar - 1.96 * se, x_bar + 1.96 * se)
  return p_value, z, confidence_interval
```

اجرای آزمون روی مقادیر گفته شده و بررسی نتایج

sample size	p_value	z score	confidence interval	result z test	result confidence interval
30	0.714393	0.365963	(0.355, 0.712)	True	True
100	0.024117	2.255254	(0.514, 0.706)	False	False
1000	0.005212	2.793642	(0.513, 0.575)	False	False

همانطور که از نتایج مشخص است (False معادل رد شدن است) ما با ۳۰ ورودی نتوانستیم متوجه ایراد سکه بشویم دلیل این موضوع همان چولگی بود که در بالا گفته شد.در ۲ مورد بعد می بینیم که هم با آزمون فرض و هم با بررسی بازه ۹۵ درصد متوجه ناعادلانه بودن سکه می شویم.

تاثیر شغل بر تحصیل

به عقیده بعضی از افراد کار کردن همزمان با تحصیل تاثیر منفی روی تحصیل دارد. در این بخش با داده هایی که از دانشجویان چند دانشگاه آمریکا داریم می خواهیم صحت این موضوع را بررسی کنیم.

تنظیم آزمون فرض

آزمون فرض را به شکل زیر تشکیل می دهیم.

$$H_0: \mu_{Not\ placed} - \mu_{placed} = 0$$
$$H_A: \mu_{Not\ placed} - \mu_{placed} < 0$$

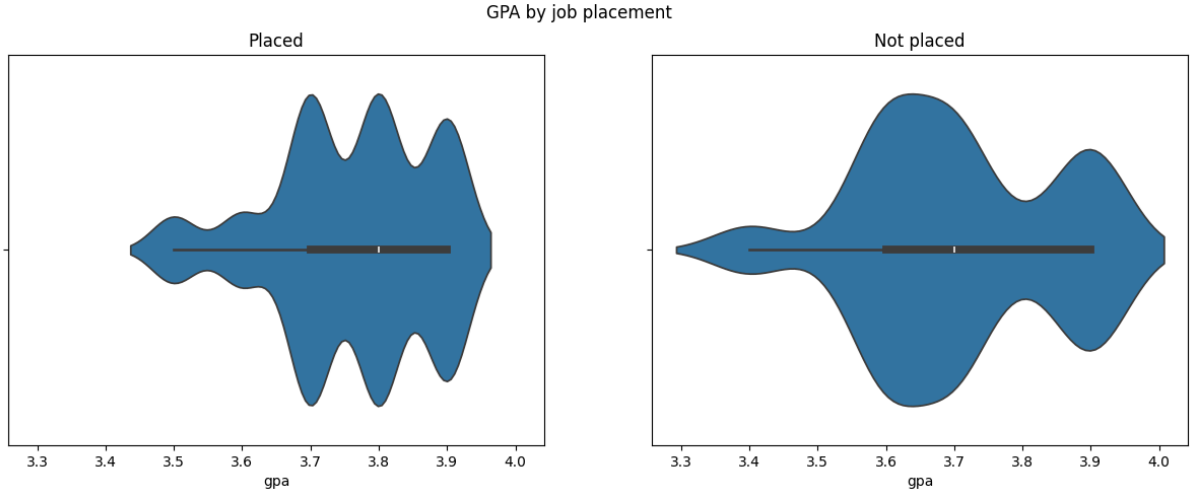
خواندن داده ها

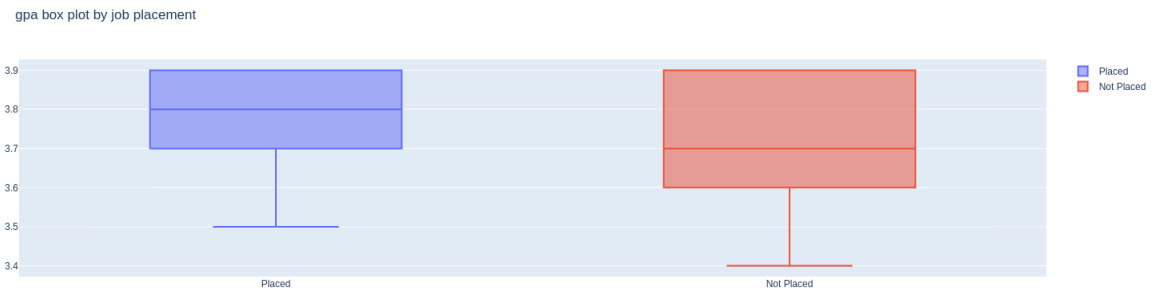
در این بخش با مجموعه داده همراه با صورت پروژه را بارگذاری می کنیم که شامل اطلاعات ۷۰۰ دانشجوی آمریکایی چند دانشگاه مختلف است. داده ها در قالب زیر هستند.

id	name	gender	age	degree	stream	college_name	placement_status	salary	gpa	years_of_experience
1	John Doe	Male	25	Bachelor's	Computer Science	Harvard University	Placed	60000	3.7	2.0
2	Jane Smith	Female	24	Bachelor's	Electrical Engineering	Massachusetts Institute of Technology	Placed	65000	3.6	1.0
3	Michael Johnson	Male	26	Bachelor's	Mechanical Engineering	Stanford University	Placed	58000	3.8	3.0

با بررسی داده ها به یک داده ی از دست رفته می رسیم. ستون مدت تجربه کاری برای ردیف ۵۴۵ خالی است. با توجه به این که ما فقط با دو ستون معدل و وضعیت کاری تحلیل خود را انجام می دهیم لازم نیست برای آن داده خالی فکری بکنیم.

حال دو گروه شاغل و غیر شاغل را از هم جدا می کنیم.





با بررسی نمودار متوجه می شویم که احتمالا فرض تاثیر منفی شغل احتمالا درست نیست.

بررسی شرایط آزمون

به بررسی شرایط آزمون فرض می پردازیم.

- **استقلال نمونه ها**: شرط کمتر از ۱۰ درصد برای استقلال برقرار است ولی باید جزئیات بیشتری از داده ها بدانیم تا بتوانیم درباره استقلال درون و برون گروهی اظهار نظر کنیم.
- **اندازه و چولگی نمونه**: با توجه به اندازه داده ها و چولگی آن به نظر مشکلی وجود ندارد.

محاسبه p-value

برای دو گروه آماره ها را حساب می کنیم.

	count	mean	std	min	25%	50%	75%	max
placement_status								
Not Placed	130.0	3.702308	0.141676	3.4	3.6	3.7	3.9	3.9
Placed	570.0	3.761404	0.113352	3.5	3.7	3.8	3.9	3.9

سپس موارد زیر را حساب می کنیم.

$$SE_{(\bar{x}_1-\bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

$$df = \frac{(n_1-1)(n_2-1)}{(n_2-1)C^2+(1-C)^2(n_1-1)}$$

$$C = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

البته درجه آزادی را می توان با روشی تقریبی حساب کرد که در آن این مقدار برابر کمترین اندازه در دو گروه منهای یک است.

حال با موارد بالا t-score را حساب می کنیم.

$$t = \frac{\bar{x}_{diff}-\mu_{diff}}{SE}$$

سپس از روی توزیع مقدار مربوط به آن را پیدا می کنیم که همان p-value ما است.

نتایج به صورت زیر است:

SE	0.013302
mu_diff	0.059096
df	683.052055
score	4.442633
p_value	0.999995
result	No evidence of negative impact (do not reject)

اجرای تست با کتابخانه scipy

تست را به صورت زیر اجرا می کنیم.

```
st.ttest_ind(df[df['placement_status'] == 'Placed']['gpa'],
        df[df['placement_status'] == 'Not Placed']['gpa'],
        alternative=HA)
```

```
TtestResult(statistic=5.105318956468754, pvalue=0.9999997868177752, df=698.0)
```

بررسی نتایج

همانطور هم که از نمودار ها متوجه شدیم نتوانستیم شواهدی مبنی بر تاثیر منفی کار کردن روی معدل تحصیل افراد پیدا کنیم.

دلیل تفاوت statistic تابع آماده تفاوت در انحراف معیار است اگر ما انحراف معیار ها را از میانگین وزن دار ۲ واریانس حساب کنیم دقیقا به نتایج تابع آماده می رسیم.

	count	mean	std	min	25%	50%	75%	max
placement_status								
Not Placed	130.0	3.702308	0.119095	3.4	3.6	3.7	3.9	3.9
Placed	570.0	3.761404	0.119095	3.5	3.7	3.8	3.9	3.9

SE	0.011575
mu_diff	0.059096
df	697.996243
score	5.105319
p_value	1.0
result	No evidence of negative impact (do not reject)

با توجه به صورت سوال که گفته شده می دانیم واریانس ها یکسان هستند استفاده از این روش معقول به نظر می رسد و راه خود را اصلاح می می کنیم. بنابراین دیگر تفاوتی بین جواب ها وجود ندارد.

پاسخ به سوالات

سوال ۱: کاربرد های شبیه سازی مونت کارلو در زندگی واقعی

شبیه ساز مونته کارلو در بسیاری از صنایع کاربرد دارد:

1. برای ارزیابی ریسک های مالی استفاده میشود
2. در پروژه های ساختمانی برای تخمین طول مدت ساخت
3. در هوانوردی برای پیش بینی خرابی در سیستم ها
4. بهینه سازی در طراحی یک سازه
5. تخمین اندازه مخازن نفتی و همینطور ترکیب شیمیایی آنها
6. پیش بینی ساختار پیچیده پروتئین ها

سوال ۲: تاثیر اندازه نمونه در نمودار ها در بخش ۲

در بخش نتایج توضیح داده شد.

سوال ۳: تاثیر اندازه نمونه در آزمایش سکه

همانطور که مشاهده کردیم با افزایش نمونه جواب ما قطعی تر می شد و هم p-value به صفر نزدیک تر می شد و هم بازه اطمینان محدود تر و به مقدار حقیقی نزدیک می شد. همانطور هم که دیدیم مقداری چولگی در داده با نمونه اولیه مشاهده شد باید برای استفاده از حد مرکزی و آزمون فرض نمونه از ۳۰ بیشتر می بود و مقدار ۳۰ کافی نبود که به همین علت تست نتوانست با نمونه ۳۰ تایی ما را به جواب درست برساند و بازه اطمینان بسیار بزرگ بود.

سوال ۴: t-test و مقایسه دو مجموعه

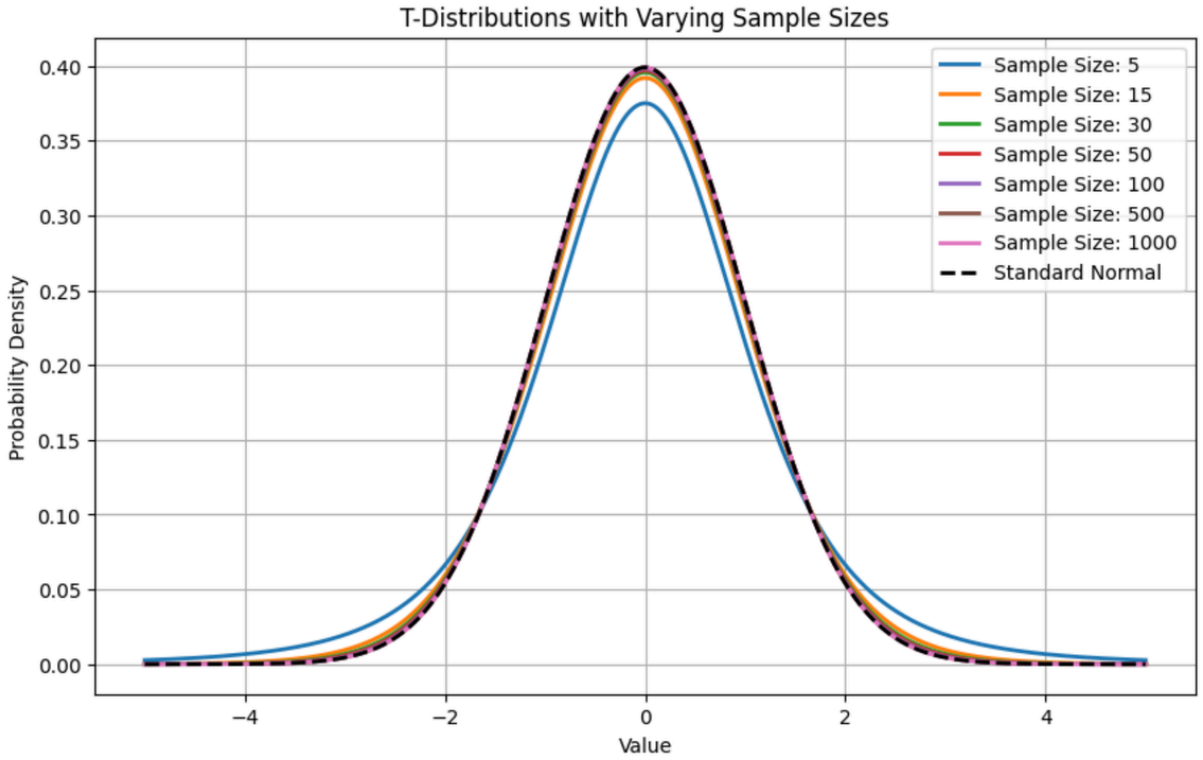
می دانیم که قضیه حد مرکزی برای اختلاف میانگین نیز برقرار است پس از تست هایی که ما در میانگین از آنها استفاده می کردیم می توانیم در تفریق میانگین نیز استفاده کنیم. در نتیجه می توان برای این موضوع از t-test استفاده کرد. نحوه استفاده از آن در گزارش توضیح داده شد و در ادامه نیز توضیحات تکمیلی ارائه می شود.

ابتدا به تعریف چند مفهوم و متغیر آماری می پردازیم که به آنها نیاز داریم:

- **درجه آزادی**: همانطور که می دانید از توزیع t می توان برای نمونه های کمتر از ۳۰ نیز استفاده کرد دلیل این موضوع و این برتری نسبت به توزیع نرمال این است که این توزیع اگر درجه آزادی پایینی داشته باشد محتاط در عمل می کند و توزیع بعد از ۳ برابر مقدار انحراف معیار نیز مقدار دارد. مقدار درجه آزادی وابسته به تعداد نمونه و واریانس آن دارد و اگر تعداد نمونه ها زیاد باشد توزیع t به

توزیع نرمال میل می کند. این پارامتر در زمانی که یک میانگین داریم برابر با تعداد نمونه منهای ۱ است و در وقتی که ۲ نمونه داریم به صورتی که در پروژه توضیح داده شد حساب می شود.

- **توزیع t**: این توزیع مانند توزیع نرمال استاندارد است با این تفاوت که یک پارامتر درجه آزادی دارد که بالا توضیح داده شد و این پارامتر باعث می شود که توزیع به توزیع نرمال نزدیک شود یا کمی محتاط تر و باز تر از آن باشد در تصویر زیر شکل توزیع را به ازای درجه آزادی های مختلف مشاهده می کنید.



- **آماره t**: این آماره اختلاف نتایج به دست آمده را با ادعای اولیه نشان می دهد سپس با استفاده از آن و توزیع t می توان احتمال رخ دادن این اختلاف را محاسبه کرد و سپس با سطح اطمینان (که معمولا ۰.۰۵) است مقایسه کرد و اگر احتمال کمتری داشت آن ادعا را رد کرد.

سوال ۵: شرایط استفاده از t test

شرایط استفاده از این تست به دارا بودن شرایط زیر است:

- استقلال
  - کمتر از ۱۰ درصد جامعه
  - نمونه گیری تصادفی
  - در صورتی که تست مقایسه ای
    - استقلال درون گروهی
    - استقلال برون گروهی
- اندازه نمونه و چولگی
  - در صورتی که چولگی وجود دارد باید به میزان لازم اندازه نمونه را افزایش داد

- بر خلاف z test نیازی نیست تعداد نمونه بیش از ۳۰ باشد

سوال ۶: معرفی چند آزمون فرض

- **آزمون ANOVA (آنالیز واریانس)**: این آزمون برای مقایسه میانگین ۲ جامعه یا بیشتر است و تعمیمی از آزمون t است و برای مقایسه های بیش از ۳ گروه استفاده می شود.
- **آزمون F**: این آزمون برای مقایسه واریانس ۲ جامعه است. مثلا بین دو شرکت تولید پیچ کدام یک پیچ هایی با واریانس کمتر در طول تولید می کند.
- **آزمون کولموگروف اسمیرنف**: این آزمون برای سنجش این است که یک نمونه آماری بر یک توزیع آماری تطابق دارد یا خیر.
- **آزمون یو من ویتنی**: این آزمون مانند آزمون t برای مقایسه ۲ متغیر عمل می کند با این تفاوت که به جای میانگین از رتبه ها برای مقایسه استفاده می کند و برای داده هایی با چولگی بالا مناسب است.

منابع

- ایردا
- آکادمی همراه
- فرادرس
- ویکی پدیا
- اسلاید های آمار و احتمال دکتر توسلی پور
- فیلم های استنباط آماری دکتر بهرک
- اسلاید های درس