به نام خدا

تمرین ٤

درس مقدمه ای بر علم داده

استاد : دکتر رضاپور

کسری صمدی<۹۹۳۶۲۳۰۳۰>

بهمن ۱٤۰۲

۱- تئوری Central Limit را روی توزیع‌های مختلف داده شامل چوله به چپ، راست، یونیفورم و نرمال بررسی کنید. برای این کار ابتدا به ازای هر یک از شکل های توزیع ذکر شده لیستی از صد هزار عدد را به وسیله زبان پایتون ایجاد کنید. سپس به ازای تغییر دو فاکتور مهم در این تئوری شامل تعداد سمپل‌ها (k) و تعداد نمونه های هر سمپل (n) بررسی کنید که نمودار توزیع میانگین این سمپل ها چه شکلی می‌گیرد. حاصل بررسی خود را در قالب یک فایل ورد شامل هیستوگرام توزیع مجموعه داده ایجاد شده، به علاوه هیستوگرام‌های حاصل از تغییرات اعمال شده روی n و k و توضیحات مکفی به ازای هر کدام و در نهایت شرحی از نتیجه‌گیریتان از این تئوری در پاسخ به این تمرین ارسال نمایید.

ابتدا کتابخانه‌های موردنیاز را ایمپورت می‌کنیم و تعداد اعداد (nums) را ۱۰۰۰۰۰ در نظر می‌گیریم.

Import Libraries

In [1]:
```
import matplotlib.pyplot as plt
import numpy as np
```

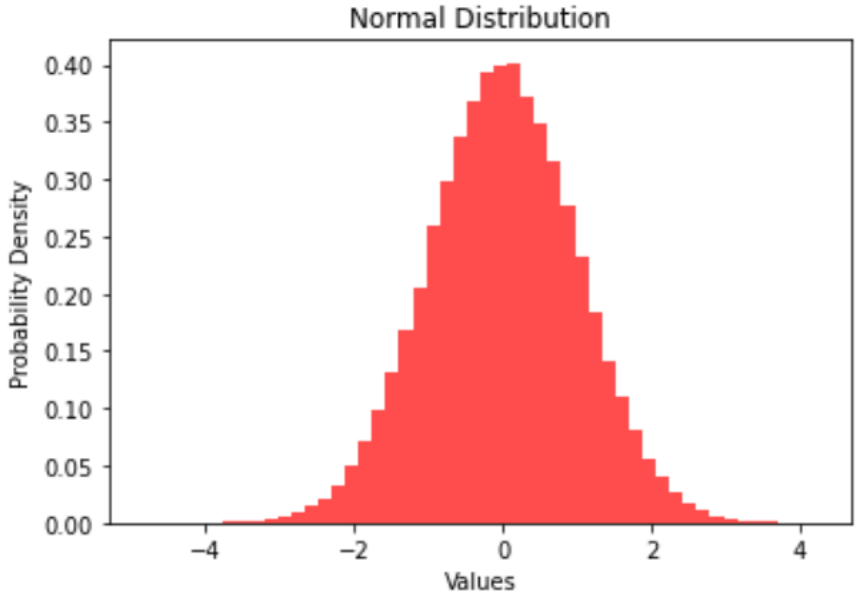
In [2]:
```
nums = 100000
```

حال توزیع‌های مختلف داده گفته شده را ایجاد می‌کنیم :

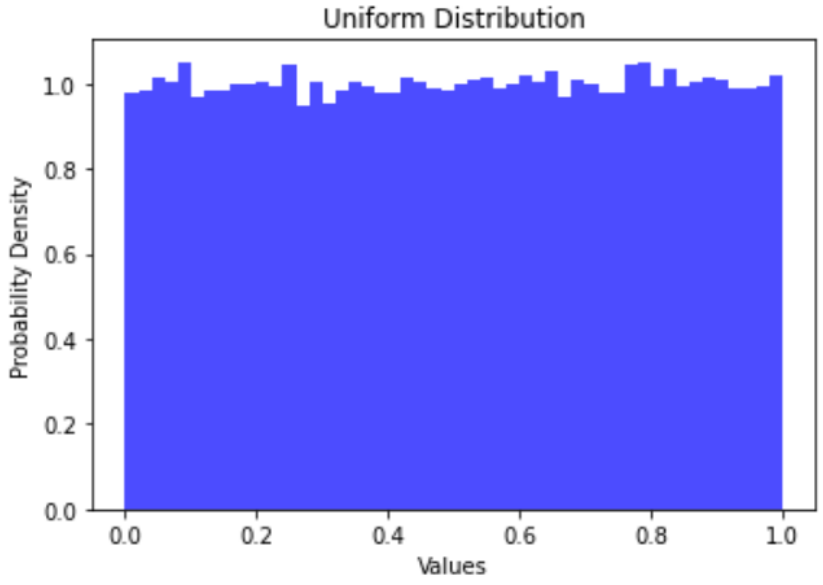
In [3]:
```
np.random.seed(0)
normal_data = np.random.normal(loc=0, scale=1, size=nums) # نرمال
uniform_data = np.random.uniform(low=0, high=1, size=nums) # یونیفرم
right_skewed_data = np.random.exponential(scale=2, size=nums) # چوله به راست
left_skewed_data = -np.random.exponential(scale=2, size=nums) # چوله به چپ
```

نمودار هیستوگرام برای توزیع‌های گفته شده به شکل زیر است:

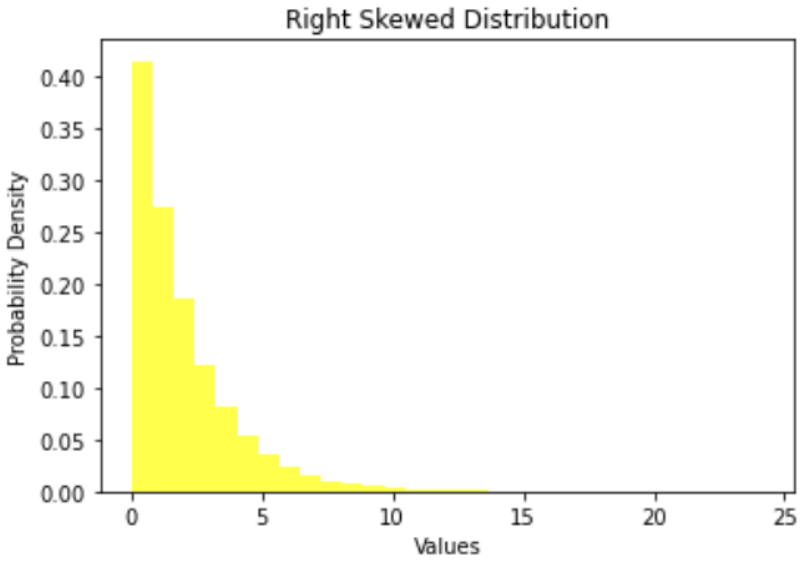
In [15]:
```
plt.hist(normal_data, bins=50, density=True, alpha=0.7, color='red')
plt.title('Normal Distribution')
plt.xlabel('Values')
plt.ylabel('Probability Density')
plt.show()
```



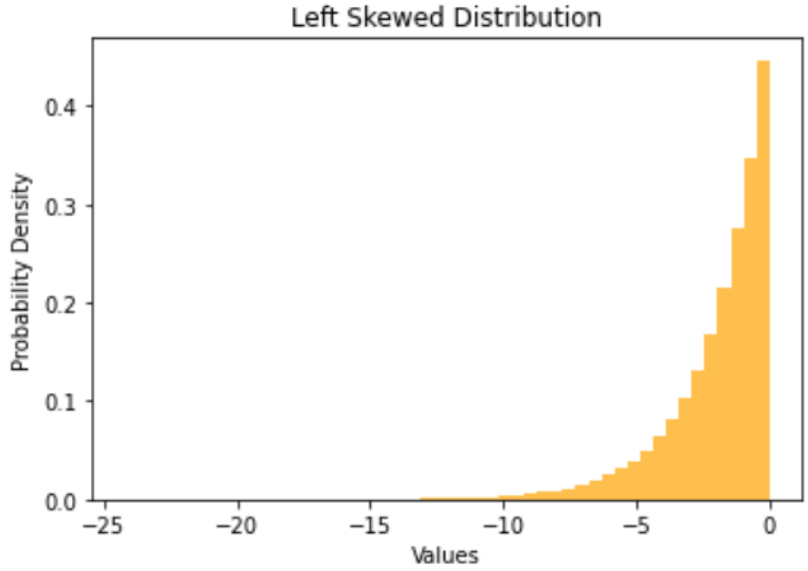
In [16]:
```
plt.hist(uniform_data, bins=50, density=True, alpha=0.7, color='blue')
plt.title('Uniform Distribution')
plt.xlabel('Values')
plt.ylabel('Probability Density')
plt.show()
```



In [17]:
```
plt.hist(right_skewed_data, bins=30, density=True, alpha=0.7, color='yellow')
plt.title('Right Skewed Distribution')
plt.xlabel('Values')
plt.ylabel('Probability Density')
plt.show()
```



In [18]:
```
plt.hist(left_skewed_data, bins=50, density=True, alpha=0.7, color='orange')
plt.title('Left Skewed Distribution')
plt.xlabel('Values')
plt.ylabel('Probability Density')
plt.show()
```



حال لیستی از k ها (تعداد سمپل‌ها) و لیستی از n ها (تعداد نمونه‌های هر سمپل) تعریف می‌کنیم :

```
In [24]: Ks = [20, 50, 100, 500] # k تعداد سمپل ها
         Ns = [10 , 30 , 50, 500] # n تعداد نمونه های هر سمپل
```

با توجه به تئوری Central Limit ، به اندازه n بار و به تعداد k سمپل از داده‌های اصلی هر توزیع، نمونه‌برداری می‌شود و میانگین این نمونه‌ها در یک آرایه به نام sample_means ذخیره می‌شود.

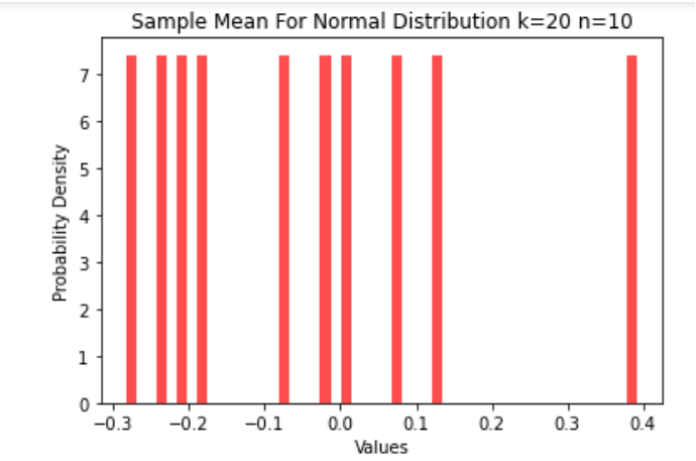
تئوری Central Limit میگوید که وقتی از توزیع‌های مختلفی سمپل‌هایی (n) با حجم‌های کافی (k) برداشت کنیم و میانگین آن‌ها را محاسبه کنیم، توزیع میانگین‌ها به شکلی نرمال ترند می‌گیرد حتی اگر توزیع اصلی نرمال نباشد.

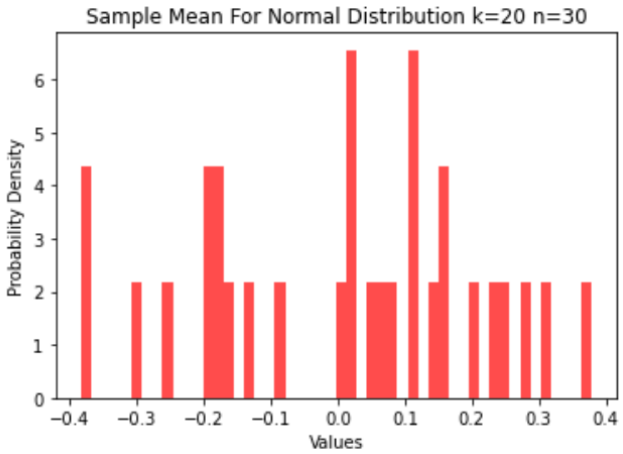
کد این قسمت به همراه خروجی‌های آن برای هر توزیع به صورت زیر است:

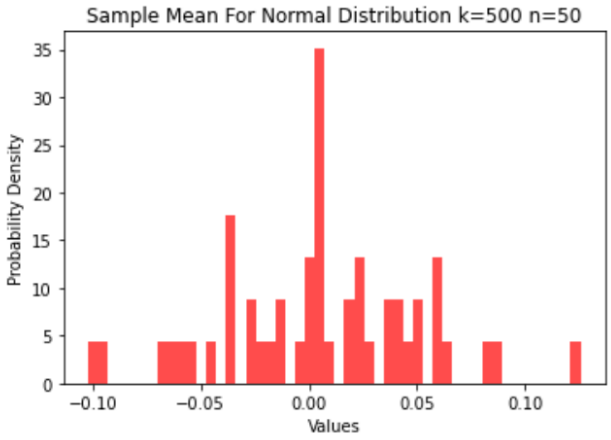
Central Limit برای توزیع نرمال :

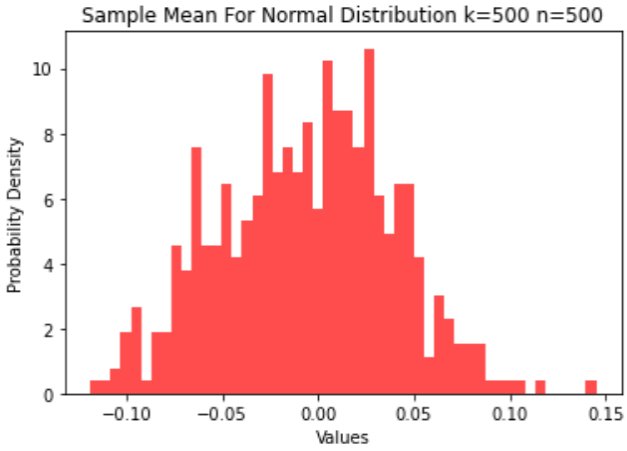
```
In [9]: for k in Ks:
            for n in Ns:
                sample_means = []
                for i in range(n):
                    sample = np.random.choice(normal_data, size=k)
                    sample_means.append(np.mean(sample))

                plt.hist(sample_means, bins=50, density=True, alpha=0.7, color='red')
                plt.title(f'Sample Mean For Normal Distribution k={k} n={n}')
                plt.xlabel('Values')
                plt.ylabel('Probability Density')
                plt.show()
```





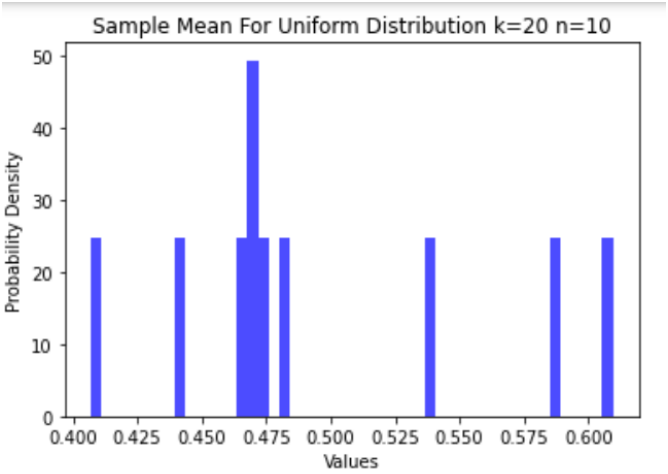


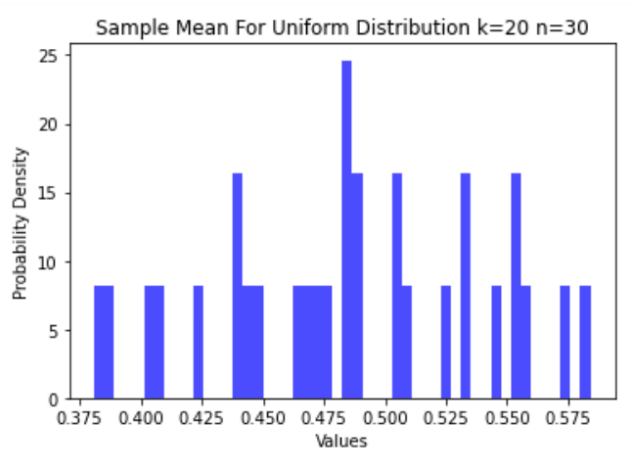


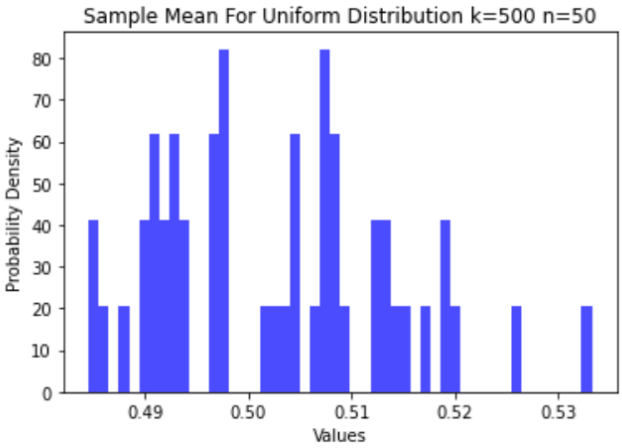
Central Limit برای توزیع یونیفورم :

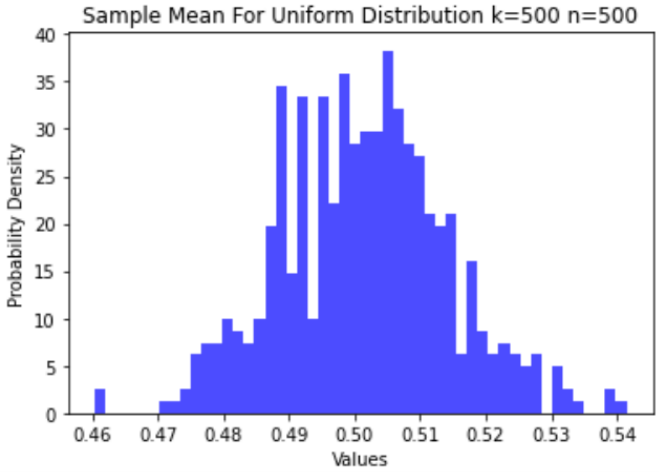
In [10]:
```
for k in Ks:
    for n in Ns:
        sample_means = []
        for i in range(n):
            sample = np.random.choice(uniform_data, size=k)
            sample_means.append(np.mean(sample))

        plt.hist(sample_means, bins=50, density=True, alpha=0.7, color='blue')
        plt.title(f'Sample Mean For Uniform Distribution k={k} n={n}')
        plt.xlabel('Values')
        plt.ylabel('Probability Density')
        plt.show()
```





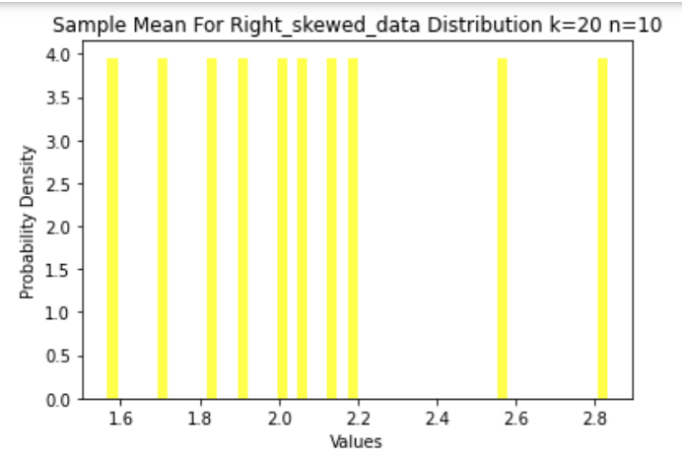


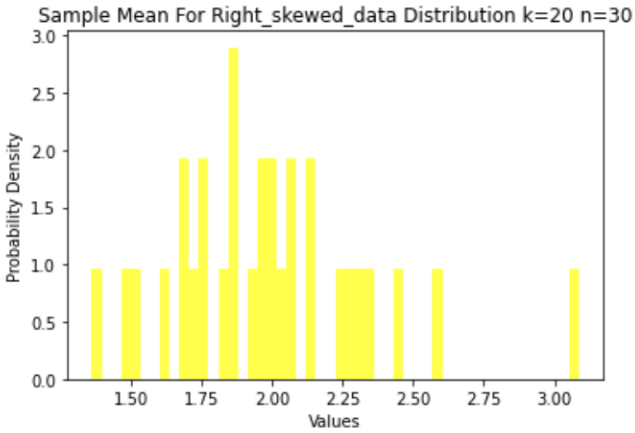


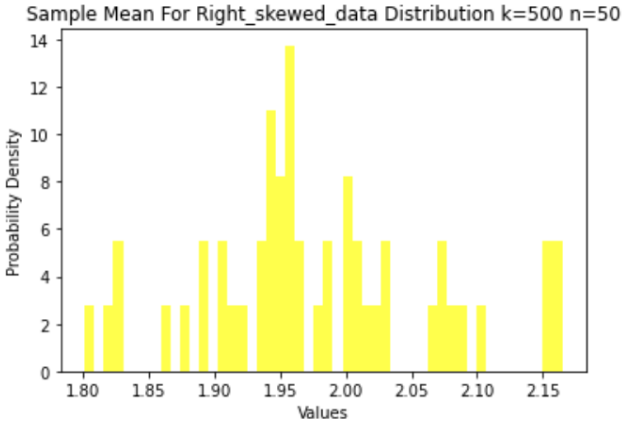
Central Limit برای توزیع چوله به راست :

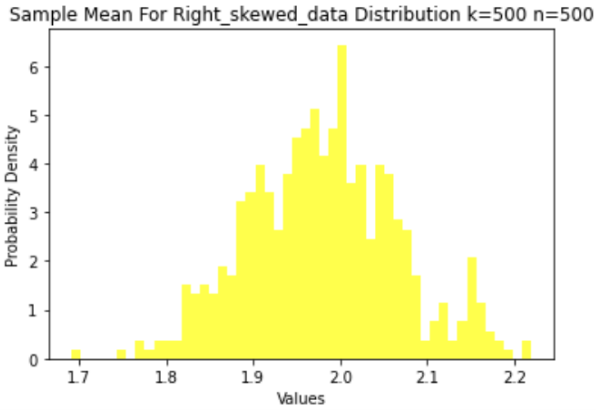
In [13]:
```
for k in Ks:
    for n in Ns:
        sample_means = []
        for i in range(n):
            sample = np.random.choice(right_skewed_data, size=k)
            sample_means.append(np.mean(sample))

        plt.hist(sample_means, bins=50, density=True, alpha=0.7, color='yellow')
        plt.title(f'Sample Mean For Right_skewed_data Distribution k={k} n={n}')
        plt.xlabel('Values')
        plt.ylabel('Probability Density')
        plt.show()
```





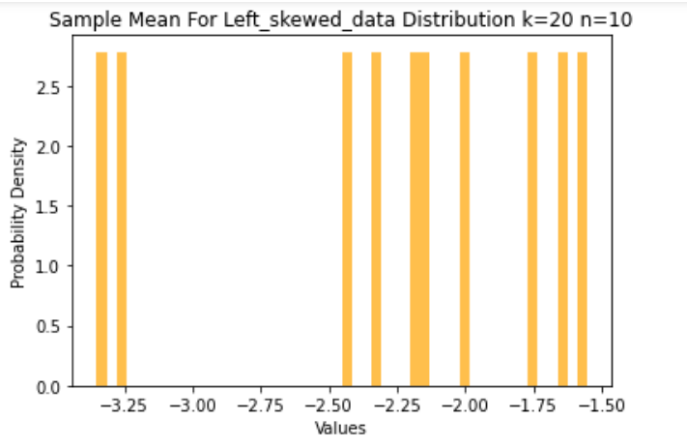


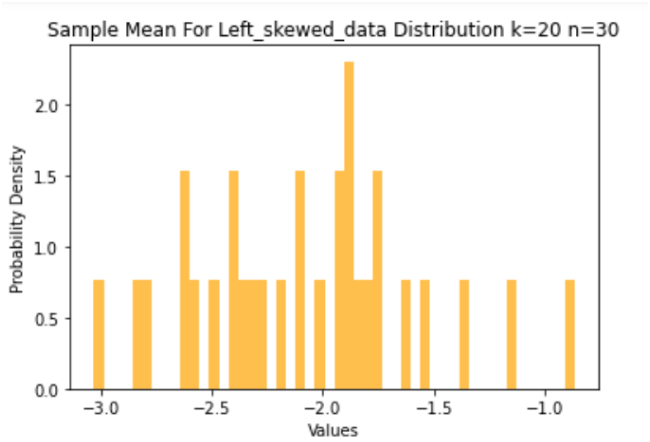


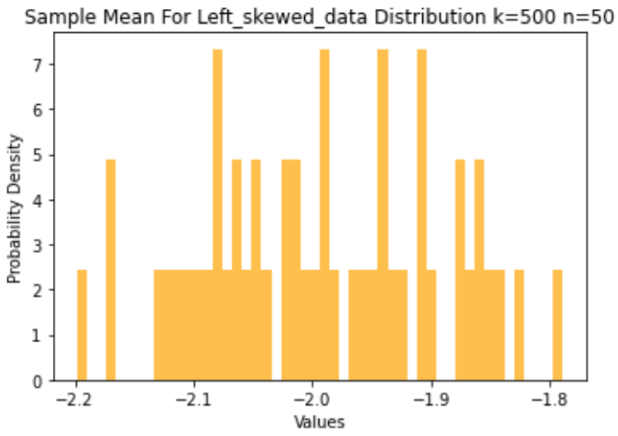
Central Limit برای توزیع چوله به چپ :

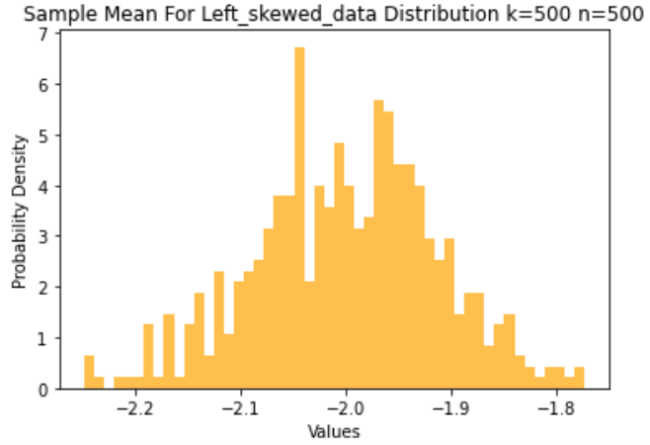
In [14]:
```
for k in Ks:
    for n in Ns:
        sample_means = []
        for i in range(n):
            sample = np.random.choice(left_skewed_data, size=k)
            sample_means.append(np.mean(sample))

        plt.hist(sample_means, bins=50, density=True, alpha=0.7, color='orange')
        plt.title(f'Sample Mean For Left_skewed_data Distribution k={k} n={n}')
        plt.xlabel('Values')
        plt.ylabel('Probability Density')
        plt.show()
```









این تئوری بر روی توزیع‌های نرمال، یونیفورم، چوله به راست و چپ انجام شد. همانطور که در نمودارهای توزیع مجموعه داده‌ها مشاهده می‌شود، توزیع‌های چوله به چپ و راست به ترتیب دارای دمای کوچکتر و بزرگتر هستند و توزیع یونیفورم نیز دمای یکسانی در تمام محدوده دارد و توزیع نرمال به صورت متقارن و مرکزی است.

نکته : در مفهوم آماری، دما (Skewness) به معیاری اشاره دارد که نشان می‌دهد که توزیع داده‌ها به چه اندازه از توزیع نرمال تقریباً به چپ یا راست منحرف است. دما یک اندازه‌گیری از تقارن توزیع است و با اندازه‌گیری اختلاف میانگین توزیع و مد آن نسبت به پراکندگی داده‌ها محاسبه می‌شود. اگر دما برابر با صفر باشد، توزیع داده‌ها به صورت کاملاً تقارنی است و شبیه به توزیع نرمال است. اگر دما بزرگتر از صفر باشد، توزیع داده‌ها به چپ منحرف (چوله به چپ) است و اگر دما کوچکتر از صفر باشد، توزیع داده‌ها به راست منحرف (چوله به راست) است.

نتیجه‌گیری :

با تغییر تعداد سمپل‌ها (k) و تعداد نمونه‌های هر سمپل(n) ، توزیع میانگین سمپل‌ها به شکلی نرمال ترند می‌گیرد. با افزایش تعداد سمپل‌ها و تعداد نمونه‌ها، توزیع میانگین به شکلی نرمال‌تر و متقارن‌تر نزدیک می‌شود.