```python
from scipy.special import comb
from scipy.stats import norm
import scipy.stats as stats
import matplotlib.pyplot as plt
```

MCB 111 week 4 Section

p-values and p-hacking

There is an infectious disease that if left untreated, affected people have a probability of recovery of $p_0 = 0.4$. Let's call this the null hypothesis H_0.

Your lab is testing two new treatments, treatment A (*TrA*) and treatment B (*TrB*). You run two different experiments (*expA* and *expB*) where *TrA* or *TrB* is given to two different groups of affected people.

Your analysis tells you that the p-values for the outcomes of *expA* and *expB* respect to the null hypothesis H_0 are

$$\begin{aligned}\text{pval}(expA) &= 0.2131,\\ \text{pval}(expB) &= 2.09 \times 10^{-6}.\end{aligned}$$

Can we conclude that *TrB* is more effective than *TrA*? The p-value is lower!

No!

Let's see a simple confounding variable that can underly this observation. Say that

- *expA* involves 15 infected people, 8 of which recovered using *TrA*
- *expB* involes 300 infected people, 160 of which recovered using *TrB*

What are the p-values under the null hypothesis of no treatment?

$$\begin{aligned}pval(expA) &= P(\text{observing result as or more extreme than 8 infected} \mid N = 15, p_0 = 0.4)\\ &= \sum_{n=8}^{15} P(n \mid N = 15, p_0 = 0.4)\\ &= \sum_{n=8}^{15} \frac{15!}{n!(15-n)!} 0.4^n 0.6^{15-n}\\ &= 0.2131\end{aligned}$$

$$\begin{aligned}pval(expB) &= P(\text{observing result as or more extreme than 160 infected} \mid N = 300, p_0 = 0.4)\\ &= \sum_{n=160}^{300} \frac{300!}{n!(300-n)!} 0.4^n 0.6^{300-n}\\ &= 2.09 \times 10^{-6}\end{aligned}$$

But, let's compare the effectiveness for each treatment:

$$\hat{p}_A = \frac{8}{15} = 0.5\bar{3}$$

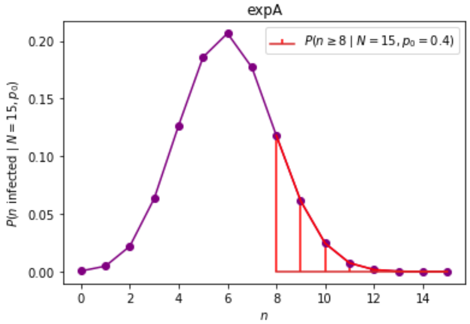
$$\hat{p}_B = \frac{160}{300} = 0.5\bar{3}$$

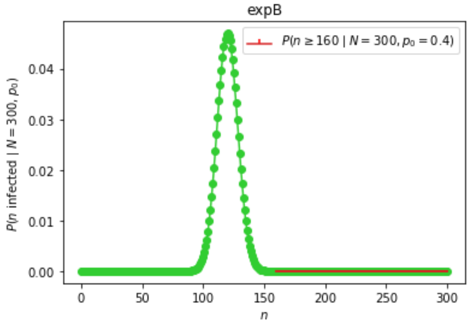
These correspond to the modes of the respective posterior distributions obtained for each probability assuming a uniform prior (i.e. maximum likelihood estimators for p_A and p_B)

```python
def binom_pmf(N, n, p):
 return comb(N, n) * p ** n * (1-p) ** (N-n)
```

```python
p0 = 0.4
nA = 8
NA = 15
nB = 160
NB = 300

ns_A = np.arange(NA+1)
ns_B = np.arange(NB+1)

more_extreme_ns_A = np.arange(nA, NA+1)
```

```
more_extreme_ns_B = np.arange(nB, NB+1)

plt.figure()
plt.plot(ns_A,
         binom_pmf(NA, ns_A, p0),
         'o-', c='purple')
plt.stem(more_extreme_ns_A,
         binom_pmf(NA, more_extreme_ns_A, p0),
         markerfmt='red',
         linefmt='red',
         use_line_collection=True,
         label='$P(n \geq 8 \mid N = 15, p_0 =0.4)$')
plt.xlabel('$n$')
plt.ylabel('$P(n$ infected$ \mid N = 15, p_0)$')
plt.title('expA')
plt.legend()
plt.show()


plt.figure()
plt.plot(ns_B,
         binom_pmf(NB, ns_B, p0),
         'o-', c='limegreen')
plt.stem(more_extreme_ns_B,
         binom_pmf(NB, more_extreme_ns_B, p0),
         markerfmt='red',
         linefmt='red',
         use_line_collection=True,
         label='$P(n \geq 160 \mid N = 300, p_0 =0.4)$')
#plt.semilogy()
plt.xlabel('$n$')
plt.ylabel('$P(n$ infected$ \mid N = 300, p_0)$')
plt.title('expB')
plt.legend()
plt.show()
```





```
np.sum(binom_pmf(N=NA, n=more_extreme_ns_A, p=p0))
```

```
0.21310318261043207
```

```
np.sum(binom_pmf(N=NB, n=more_extreme_ns_B, p=p0))
```

```
2.0919886135867843e-06
```

So what have we seen here? These two treatments have the same effectiveness, but because the sample sizes of the experiments are so different, the null distribution for expB is much narrower than that of expA, so the p-value for expB is smaller as a result.

You can never compare two hypotheses by looking to their p-values relative to a third null hypothesis. There are many hidden and possibly confounding variables in that calculation, one of them the sample size.

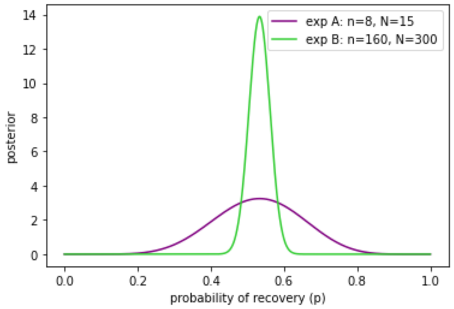
Let's instead look at the posterior distributions for the two experiments, as done before in week 2:

$$
\begin{aligned}
P(p_A \mid expA) &= \frac{16!}{8!7!}\, p_A^8 (1-p_A)^7 \\
P(p_B \mid expB) &= \frac{301!}{160!140!}\, p_B^{160} (1-p_B)^{140}
\end{aligned}
$$

```
def binom_post(p, N, n):
  return (N + 1) * comb(N, n) * p**n * (1-p)**(N-n)
```

```
ps = np.linspace(0, 1, 10000)

post_expA = binom_post(ps, N=NA, n=nA)
post_expB = binom_post(ps, N=NB, n=nB)
```

```
plt.figure()
plt.plot(ps, post_expA, c='purple', label='exp A: n=8, N=15')
plt.plot(ps, post_expB, c='limegreen', label='exp B: n=160, N=300')
plt.legend()
plt.xlabel('probability of recovery (p)')
plt.ylabel('posterior')
plt.show()
```



These posteriors are compatible with both treatments having the same effectiveness, as they are both centered around the same value, but the estimate of p_B is more precise than that of p_A because there *expB* had much more data.

As we did in the homework of w02, you remember that when the data follows a binomial distribution, the best estimate of the Bernoulli parameter p and its confidence value are given by

$$
\hat{p} = \tilde{\mu} = \frac{n}{N} \qquad \tilde{\sigma} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}.
$$

The best estimates and confidence values for the effectiveness of the two treatments are

$$
\begin{aligned}
p_A &\approx 0.533 \pm 0.129 \\
p_B &\approx 0.523 \pm 0.029.
\end{aligned}
$$

In fact, if the two experiments had been run with using the same treatment, you would have obtained one “significant” p-value and one “non significant” p-value just because of the different sample size.

$$
P(p_A < p_B) = 1 - \mathrm{cdf}(0 | \mu = p_A - p_B, \sigma^2 = \sigma_A^2 + \sigma_B^2)
$$

$$
P(p_A < p_B) = 1 - \Phi\left(\frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}} \right)
$$

$$P(p_A < p_B) = P(p_A - p_B < 0)$$

$$P(p_A < p_B) = 1 - \text{cdf}(0|\mu = p_A - p_B, \sigma^2 = \sigma_A^2 + \sigma_B^2)$$

$$P(p_A < p_B) = 1 - \Phi\left(\frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)$$

$$= 1 - \Phi\left(\frac{0.5\bar{3} - 0.5\bar{3}}{\sqrt{0.129^2 + 0.029^2}}\right)$$

$$= 1 - \Phi\,(0)$$

$$P(p_A < p_B) \approx 50\%$$

Do not compare p-values of different experiments!