Homework_03

September 17, 2019

Probability for Data Science

UC Berkeley, Fall 2019

Ani Adhikari and Jim Pitman

1 Homework 3

1.0.1 Instructions

Your homeworks have two components: a written portion and a portion that also involves code. Written work should be completed on paper, and coding questions should be done in the notebook. It is your responsibility to ensure that both components of the homework are submitted completely and properly to Gradescope. Refer to the bottom of the notebook for submission instructions.

1.0.2 How to Do Your Homework

The point of homework is for you to try your hand at using what you've learned in class. The steps to follow:

- Go to lecture and sections, and also go over the relevant text sections before starting on the homework. This will remind you what was covered in class, and the text will typically contain examples not covered in lecture. The weekly Preparation Guide will list what you should read.
- Work on some of the practice problems before starting on the homework.
- Attempt the homework problems by yourself with the text, section work, and practice materials all at hand. Sometimes the week's lab will help as well. The two steps above will help this step go faster and be more fruitful.
- At this point, seek help if you need it. Don't ask how to do the problem — ask how to get started, or for a nudge to get you past where you are stuck.
- For a good measure of your understanding, keep track of the fraction of the homework you can do by yourself or with minimal help. It's a better measure than your homework score, and only you can measure it.

[1]:
```
# Run this cell to set up your notebook
```

```
import numpy as np
from scipy import stats
from datascience import *
from prob140 import *

# These lines do some fancy plotting magic
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')

# These lines make warnings look nicer
import warnings
warnings.simplefilter('ignore', FutureWarning)
```

1.1 1. Binomial and Conditioning

A die is rolled 30 times. Let X be the number of sixes in the first 12 rolls and let Y be the number of sixes in all 30 rolls.

a) [**CODE**] Find $P(X = 3, Y = 7)$.

[2]:
```
# Answer to 1a

# P(X=3, Y=7)
p_x_3 = stats.binom.pmf(3, 12, 1/6)
p_x_4_given = stats.binom.pmf(7-3, 30-12, 1/6)



p_X_3_Y_7 = p_x_3* p_x_4_given
p_X_3_Y_7
```

[2]:
```
0.03630084173970491
```

b) [**CODE**] Construct the joint distribution table of X and Y and name it `joint_dist`. It should be a `prob140` joint distribution object so that you can call `prob140` methods on it in subsequent parts. Use as many lines of code as you need.

It will be a large table, so, not surprisingly, the probabilities in the individual cells will be rather small.

[3]:
```
# Answer to 1b

def joint_prob(x,y):
    if(x<=y):
        p_x = stats.binom.pmf(x,12,1/6)
        p_y_given = stats.binom.pmf(y-x,30-12,1/6)
```

```
        return p_x * p_y_given
    else:
        return 0


k= np.arange(31)

joint_dist = Table().values('X', k, 'Y', k).probability_function(joint_prob)
joint_dist




#def joint_prob(h, s):
#    if (h+s >=0) and (h+s <=5):
#        numerator = special.comb(13, h) * special.comb(13, s) * special.
 ↪comb(26, 5 - h -s)
#        denominator = special.comb(52, 5)
#        return numerator / denominator
#    else:
#        return 0

#k = np.arange(6)
#joint_dist = Table().values('H', k, 'S', k).probability_function(joint_prob)
#joint_dist
```

[3]:

	X=0	X=1	X=2	X=3	X=4	\
Y=30	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
Y=29	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
Y=28	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
Y=27	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
Y=26	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
Y=25	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
Y=24	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
Y=23	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
Y=22	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	8.746368e-16	
Y=21	0.000000e+00	0.000000e+00	0.000000e+00	1.943637e-15	7.871731e-14	
Y=20	0.000000e+00	0.000000e+00	2.915456e-15	1.749274e-13	3.345486e-12	
Y=19	0.000000e+00	2.650414e-15	2.623910e-13	7.434412e-12	8.921295e-11	
Y=18	1.104339e-15	2.385373e-13	1.115162e-11	1.982510e-10	1.672743e-09	
Y=17	9.939054e-14	1.013784e-11	2.973765e-10	3.717206e-09	2.341840e-08	
Y=16	4.224098e-12	2.703423e-10	5.575809e-09	5.204089e-08	2.536993e-07	
Y=15	1.126426e-10	5.068918e-09	7.806133e-08	5.637763e-07	2.174566e-06	
Y=14	2.112049e-09	7.096485e-08	8.456644e-07	4.832368e-06	1.495014e-05	
Y=13	2.956869e-08	7.687858e-07	7.248552e-06	3.322253e-05	8.305633e-05	
Y=12	3.203274e-07	6.589593e-06	4.983380e-05	1.845696e-04	3.737535e-04	
Y=11	2.745664e-06	4.530345e-05	2.768544e-04	8.305633e-04	1.359104e-03	

```
Y=10  1.887644e-05  2.516858e-04  1.245845e-03  3.020230e-03  3.964052e-03
Y=9   1.048691e-04  1.132586e-03  4.530345e-03  8.809004e-03  9.147812e-03
Y=8   4.719109e-04  4.118495e-03  1.321351e-02  2.032847e-02  1.633538e-02
Y=7   1.716040e-03  1.201228e-02  3.049271e-02  3.630084e-02  2.178051e-02
Y=6   5.005116e-03  2.772064e-02  5.445126e-02  4.840112e-02  2.041922e-02
Y=5   1.155027e-02  4.950115e-02  7.260168e-02  4.537605e-02  1.201131e-02
Y=4   2.062548e-02  6.600153e-02  6.806408e-02  2.669180e-02  3.336474e-03
Y=3   2.750064e-02  6.187643e-02  4.003769e-02  7.414388e-03  0.000000e+00
Y=2   2.578185e-02  3.639790e-02  1.112158e-02  0.000000e+00  0.000000e+00
Y=1   1.516579e-02  1.011053e-02  0.000000e+00  0.000000e+00  0.000000e+00
Y=0   4.212720e-03  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00

                X=5           X=6           X=7           X=8           X=9  \
Y=30  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
Y=29  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
Y=28  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
Y=27  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  1.243928e-19
Y=26  0.000000e+00  0.000000e+00  0.000000e+00  1.399419e-18  1.119535e-17
Y=25  0.000000e+00  0.000000e+00  1.119535e-17  1.259477e-16  4.758024e-16
Y=24  0.000000e+00  6.530621e-17  1.007582e-15  5.352777e-15  1.268806e-14
Y=23  2.798838e-16  5.877559e-15  4.282222e-14  1.427407e-13  2.379012e-13
Y=22  2.518954e-14  2.497963e-13  1.141926e-12  2.676388e-12  3.330617e-12
Y=21  1.070555e-12  6.661233e-12  2.141111e-11  3.746944e-11  3.608168e-11
Y=20  2.854814e-11  1.248981e-10  2.997555e-10  4.059189e-10  3.092716e-10
Y=19  5.352777e-10  1.748574e-09  3.247351e-09  3.479305e-09  2.126242e-09
Y=18  7.493888e-09  1.894288e-08  2.783444e-08  2.392022e-08  1.181246e-08
Y=17  8.118378e-08  1.623676e-07  1.913618e-07  1.328901e-07  5.315605e-08
Y=16  6.958610e-07  1.116277e-06  1.063121e-06  5.980055e-07  1.932947e-07
Y=15  4.784044e-06  6.201539e-06  4.784044e-06  2.174566e-06  5.637763e-07
Y=14  2.657802e-05  2.790693e-05  1.739652e-05  6.342483e-06  1.301022e-06
Y=13  1.196011e-04  1.014797e-04  5.073986e-05  1.463650e-05  2.323254e-06
Y=12  4.349131e-04  2.959825e-04  1.170920e-04  2.613661e-05  3.097672e-06
Y=11  1.268497e-03  6.830366e-04  2.090928e-04  3.484881e-05  2.904067e-06
Y=10  2.927300e-03  1.219708e-03  2.787905e-04  3.267076e-05  1.708275e-06
Y=9   5.227321e-03  1.626278e-03  2.613661e-04  1.921809e-05  4.745208e-07
Y=8   6.969762e-03  1.524635e-03  1.537447e-04  5.338359e-06  0.000000e+00
Y=7   6.534152e-03  8.968443e-04  4.270687e-05  0.000000e+00  0.000000e+00
Y=6   3.843619e-03  2.491234e-04  0.000000e+00  0.000000e+00  0.000000e+00
Y=5   1.067672e-03  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
Y=4   0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
Y=3   0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
Y=2   0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
Y=1   0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
Y=0   0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00

      …  X=21  X=22  X=23  X=24  X=25  X=26  X=27  X=28  X=29  X=30
Y=30  …   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
```

```
Y=29  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=28  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=27  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=26  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=25  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=24  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=23  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=22  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=21  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=20  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=19  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=18  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=17  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=16  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=15  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=14  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=13  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=12  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=11  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=10  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=9   …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=8   …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=7   …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=6   …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=5   …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=4   …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=3   …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=2   …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=1   …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Y=0   …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0

[31 rows x 31 columns]
```

c) [CODE] Display the conditional distribution of X given $Y = y$, for all the possible values y of Y.

[4]:
```
# Answer to 1c
joint_dist.conditional_dist('X', 'Y')
```

[4]:
```
                              X=0           X=1       X=2       X=3       X=4  \
Dist. of X | Y=30  0.000000e+00  0.000000e+00  0.000000  0.000000  0.000000
Dist. of X | Y=29  0.000000e+00  0.000000e+00  0.000000  0.000000  0.000000
Dist. of X | Y=28  0.000000e+00  0.000000e+00  0.000000  0.000000  0.000000
Dist. of X | Y=27  0.000000e+00  0.000000e+00  0.000000  0.000000  0.000000
Dist. of X | Y=26  0.000000e+00  0.000000e+00  0.000000  0.000000  0.000000
Dist. of X | Y=25  0.000000e+00  0.000000e+00  0.000000  0.000000  0.000000
Dist. of X | Y=24  0.000000e+00  0.000000e+00  0.000000  0.000000  0.000000
```

Dist. of X	Y=23	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000
Dist. of X	Y=22	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000085
Dist. of X	Y=21	0.000000e+00	0.000000e+00	0.000000	0.000015	0.000623
Dist. of X	Y=20	0.000000e+00	0.000000e+00	0.000002	0.000132	0.002521
Dist. of X	Y=19	0.000000e+00	2.196704e-07	0.000022	0.000616	0.007394
Dist. of X	Y=18	1.156160e-08	2.497305e-06	0.000117	0.002076	0.017512
Dist. of X	Y=17	1.503008e-07	1.533068e-05	0.000450	0.005621	0.035414
Dist. of X	Y=16	1.052106e-06	6.733475e-05	0.001389	0.012962	0.063189
Dist. of X	Y=15	5.260528e-06	2.367237e-04	0.003646	0.026329	0.101554
Dist. of X	Y=14	2.104211e-05	7.070149e-04	0.008425	0.048144	0.148947
Dist. of X	Y=13	7.154318e-05	1.860123e-03	0.017538	0.080384	0.200960
Dist. of X	Y=12	2.146295e-04	4.415236e-03	0.033390	0.123667	0.250427
Dist. of X	Y=11	5.825659e-04	9.612337e-03	0.058742	0.176226	0.288370
Dist. of X	Y=10	1.456415e-03	1.941886e-02	0.096123	0.233026	0.305847
Dist. of X	Y=9	3.398301e-03	3.670165e-02	0.146807	0.285457	0.296436
Dist. of X	Y=8	7.476262e-03	6.524738e-02	0.209335	0.322054	0.258794
Dist. of X	Y=7	1.563218e-02	1.094253e-01	0.277772	0.330681	0.198408
Dist. of X	Y=6	3.126437e-02	1.731565e-01	0.340129	0.302337	0.127548
Dist. of X	Y=5	6.012378e-02	2.576734e-01	0.377921	0.236201	0.062524
Dist. of X	Y=4	1.116585e-01	3.573071e-01	0.368473	0.144499	0.018062
Dist. of X	Y=3	2.009852e-01	4.522167e-01	0.292611	0.054187	0.000000
Dist. of X	Y=2	3.517241e-01	4.965517e-01	0.151724	0.000000	0.000000
Dist. of X	Y=1	6.000000e-01	4.000000e-01	0.000000	0.000000	0.000000
Dist. of X	Y=0	1.000000e+00	0.000000e+00	0.000000	0.000000	0.000000
Marginal of X		1.121567e-01	2.691760e-01	0.296094	0.197396	0.088828

		X=5	X=6	X=7	X=8	X=9	…	\
Dist. of X	Y=30	0.000000	0.000000	0.000000	0.000000	0.000000	…	
Dist. of X	Y=29	0.000000	0.000000	0.000000	0.000000	0.000000	…	
Dist. of X	Y=28	0.000000	0.000000	0.000000	0.000000	0.000000	…	
Dist. of X	Y=27	0.000000	0.000000	0.000000	0.000000	0.054187	…	
Dist. of X	Y=26	0.000000	0.000000	0.000000	0.018062	0.144499	…	
Dist. of X	Y=25	0.000000	0.000000	0.005558	0.062524	0.236201	…	
Dist. of X	Y=24	0.000000	0.001556	0.024009	0.127548	0.302337	…	
Dist. of X	Y=23	0.000389	0.008170	0.059523	0.198408	0.330681	…	
Dist. of X	Y=22	0.002436	0.024154	0.110419	0.258794	0.322054	…	
Dist. of X	Y=21	0.008470	0.052700	0.169392	0.296436	0.285457	…	
Dist. of X	Y=20	0.021510	0.094107	0.225856	0.305847	0.233026	…	
Dist. of X	Y=19	0.044365	0.144924	0.269145	0.288370	0.176226	…	
Dist. of X	Y=18	0.078455	0.198318	0.291406	0.250427	0.123667	…	
Dist. of X	Y=17	0.122768	0.245536	0.289382	0.200960	0.080384	…	
Dist. of X	Y=16	0.173320	0.278034	0.264794	0.148947	0.048144	…	
Dist. of X	Y=15	0.223420	0.289618	0.223420	0.101554	0.026329	…	
Dist. of X	Y=14	0.264794	0.278034	0.173320	0.063189	0.012962	…	
Dist. of X	Y=13	0.289382	0.245536	0.122768	0.035414	0.005621	…	
Dist. of X	Y=12	0.291406	0.198318	0.078455	0.017512	0.002076	…	
Dist. of X	Y=11	0.269145	0.144924	0.044365	0.007394	0.000616	…	

Dist. of X | Y=10	0.225856	0.094107	0.021510	0.002521	0.000132	…
Dist. of X | Y=9	0.169392	0.052700	0.008470	0.000623	0.000015	…
Dist. of X | Y=8	0.110419	0.024154	0.002436	0.000085	0.000000	…
Dist. of X | Y=7	0.059523	0.008170	0.000389	0.000000	0.000000	…
Dist. of X | Y=6	0.024009	0.001556	0.000000	0.000000	0.000000	…
Dist. of X | Y=5	0.005558	0.000000	0.000000	0.000000	0.000000	…
Dist. of X | Y=4	0.000000	0.000000	0.000000	0.000000	0.000000	…
Dist. of X | Y=3	0.000000	0.000000	0.000000	0.000000	0.000000	…
Dist. of X | Y=2	0.000000	0.000000	0.000000	0.000000	0.000000	…
Dist. of X | Y=1	0.000000	0.000000	0.000000	0.000000	0.000000	…
Dist. of X | Y=0	0.000000	0.000000	0.000000	0.000000	0.000000	…
Marginal of X	0.028425	0.006632	0.001137	0.000142	0.000013	…

	X=22	X=23	X=24	X=25	X=26	X=27	X=28	X=29	X=30	Sum
Dist. of X | Y=30	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=29	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=28	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Dist. of X | Y=0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Marginal of X	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

```
[32 rows x 32 columns]
```

d) [WRITTEN & CODE] In the table in Part **c)**, identify the value of $P(X = 3 \mid Y = 7)$. Now complete the line of code to compute it in another way. The expression involves `p_X_3_Y_7` from Part **a)**.

[5]:
```
# Answer to 1d

# P(X = 3 | Y = 7)

p_X_3_Y_7 / stats.binom.pmf(7, 30, 1/6)
```

[5]:
```
0.3306808134394365
```

e) [WRITTEN] Now let $1 \leq m < n$ be integers. Consider n i.i.d. Bernoulli trials with probability of p of success on each trial. Let V be the number of successes in the first m trials and let W be the number of successes in all n trials.

Fix an integer w in the range 0 through n. For $0 \leq v \leq \min(w, m)$, find $P(V = v \mid W = w)$ and show that it doesn't depend on p.

f) [WRITTEN] With no further calculation, fill in the blanks below. The first blank should be filled with the name of a distribution and the second with the parameters in terms of n, m, v, and w.

The conditional distribution of V given $W = w$ is _____ with parameters _____.

g) [CODE] Return to the random variables X and Y in Parts **a-d**. Complete the cell below to calculate the probabilities in the conditional distribution of X given $Y = 7$. Your answer should agree with the relevant portion of the display in Part **c**.

[6]:
```
# Answer to 1g

# Conditional distribution of X given Y=7

x = np.arange(31)      # array of possible values of X given Y=7
stats.hypergeom.pmf(x, 30,12,7)
```

[6]:
```
array([0.01563218, 0.10942529, 0.27777188, 0.33068081, 0.19840849,
       0.05952255, 0.00816976, 0.00038904, 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        ])
```

2 newpage

2.1 2. Phone Calls

In an hour, a student receives X phone calls from people he knows and Y phone calls from people he doesn't know. Assume that X has the Poisson (λ) distribution and Y has the Poisson (μ) distribution. Also assume that X and Y are independent. Finally, assume that each call has chance 0.1 of being missed by the student, independently of all other calls.

In some parts below, you are asked to provide the name of a distribution and its parameters. Keep in mind that a distribution might have only one parameter. We aren't saying "parameter or parameters" every time because that's clunky.

a) [WRITTEN] Fill in the blank with the name of a distribution and its parameters:

The total number of calls that the student receives has the _____ distribution.

b) [WRITTEN] Fill in the blank with the name of a distribution and its parameters:

Given that the student receives a total of k calls, the conditional distribution of the total number of missed calls is _____.

c) [WRITTEN] Fill in the blank with the name of a distribution and its parameters:

The total number of missed calls has the _____ distribution.

d) [WRITTEN] For non-negative integer n, find the chance that the student misses at most n calls from people he knows.

e) [WRITTEN] Let n and m be non-negative integers. Find the chance that the student misses at most n calls from people he knows and also misses at least m calls from people he doesn't know. Don't leave infinite sums in your answer.

2.2 3. Poisson Approximation at Both Ends

Consider n independent Bernoulli (p) trials.

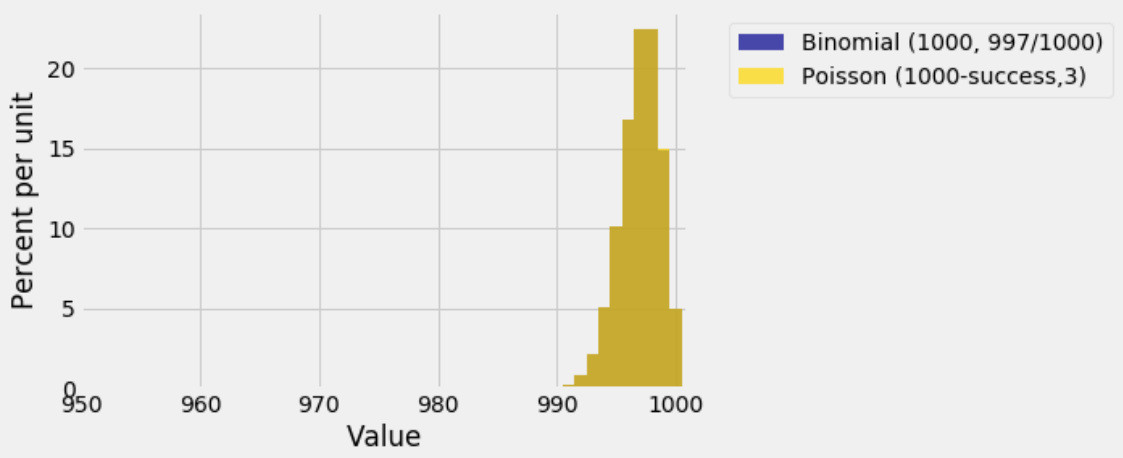
a) [WRITTEN] Fill in the blanks with names of distributions along with parameters in parentheses: If $n = 1000$ and $p = 0.003$, the distribution of the number of successes is exactly _____ (_____) and approximately _____ (_____).

b) [WRITTEN] Let n be large and let p be close to 1. Find a Poisson approximation to p_k (the chance of k successes) by an appropriate use of the Poisson approximation to the binomial derived in class.

Note: You shouldn't have to derive a limit. Just use the limit already derived in class and in the text, but appropriately.

c) [CODE] Plot the probability histogram of the binomial (1000, 0.997) distribution, and overlay your Poisson approximation from Part **b**. Please don't plot the entire range. Choose an informative range of values on the horizontal axis.

[7]:
```
# Answer to 3c

n = 1000
p = 0.997

k = np.arange(1001)                 # array of possible values
binomial_probs = stats.binom.pmf(k, 1000, 0.997)    # array of exact binomial␣
 ↪probabilities
#binomial_probs
#poisson_test = stats.poisson.pmf(1000-k,3)
#poisson_test
#poisson_test

def poisson_approximation_pmf(j):
    """Returns the Poisson approximation to the
    exact binomial probability of j successes"""
    return stats.poisson.pmf(1000-j, 3)

exact_binomial = Table().values(k).probabilities(binomial_probs)
poisson_approximation = Table().values(k).
 ↪probability_function(poisson_approximation_pmf)
#Plots("titles",exact_binomial)

Plots('Binomial (1000, 997/1000)', exact_binomial, 'Poisson (1000-success,3)',␣
 ↪poisson_approximation)
#Plots(...)
plt.xlim(950, 1001);  # left and right endpoints of range of values on␣
 ↪horizontal axis
```



2.3 4. Understanding Total Variation Distance, Part 1

In this exercise and the next, you are going to show that the formula for TVD used in Data 8 and in Lab 2 has an interpretation as a maximum difference of probabilities.

We are going to compare two probability distributions P_{blue} and P_{gold} on a finite set of values S. Suppose the values are labeled $1, 2, \ldots, n$.

The *total variation distance between P_{blue} and P_{gold}* is defined as

$$\|P_{blue} - P_{gold}\|_{TVD} \;=\; \max\{|P_{blue}(A) - P_{gold}(A)| : A \subseteq S\}$$

The definition says: For every event A, compute how far off $P_{blue}(A)$ is from $P_{gold}(A)$. The TVD is the biggest value among all these differences.

That doesn't look at all like what we have been calculating as the TVD starting way back in Data 8. But in fact it's the same thing. It's your job to show how.

Before you get started, confirm your understanding of the definition. Suppose you calculate the TVD between two distributions and get 0.003. That says that if you list all possible events and compare their probabilities under the two distributions, the biggest difference you will get is 3/1000. The two distributions are pretty close.

The goal of these exercises is to show that this new definition of TVD is equivalent to the calculation we have been doing all along. Let's start by setting up some notation. For each i in S, let $P_{blue}(i) = b_i$ and let $P_{gold}(i) = g_i$. If you imagine a bar graph or histogram of each distribution, then b_i is the area of the blue bar at the value i, and g_i is the area of the gold bar at i.

In this notation, our familiar calculation of the TVD is

$$\frac{1}{2}\sum_{i \in S}|b_i - g_i|$$

In this question and the next you will show that

$$\max\{|P_{blue}(A) - P_{gold}(A)| : A \subseteq S\} \;=\; \frac{1}{2}\sum_{i \in S}|b_i - g_i|$$

Three events will be important in the calculations.

The set of values for which the blue bars exceed the gold:

$$B = \{i : b_i > g_i\}$$

The set of values for which the gold bars exceed the blue:

$$G = \{i : g_i > b_i\}$$

The set of values for which the blue bars and gold bars are equal:

$$E = \{i : b_i = g_i\}$$

Keep in mind that for any event A,

$$P_{blue}(A) = \sum_{i \in A} b_i \qquad \text{and} \qquad P_{gold}(A) = \sum_{i \in A} g_i$$

a) [WRITTEN] Find the value of

$$\sum_{i \in B} b_i \; + \; \sum_{i \in G} b_i \; + \; \sum_{i \in E} b_i$$

Repeat the calculation after replacing b_i by g_i in all three sums above.

Hence show that

$$\sum_{i \in B} (b_i - g_i) \; = \; \sum_{i \in G} (g_i - b_i)$$

This proves a statement we have made in Data 8 and Lab 2: "The amount by which the blue bars exceed the gold is the same as the amount by which the gold bars exceed the blue."

b) [WRITTEN] Our usual calculation of TVD is

$$\frac{1}{2} \sum_{i \in S} |b_i - g_i|$$

Partition the sum into two pieces to show that

$$\frac{1}{2} \sum_{i \in S} |b_i - g_i| \; = \; \sum_{i \in B} (b_i - g_i) \; = \; \sum_{i \in G} (g_i - b_i)$$

This proves another statement we made in Data 8 and Lab 2: "The TVD is the amount by which the blue bars exceed the gold."

3 newpage

3.1 5. Understanding Total Variation Distance, Part 2

a) [WRITTEN] Now let A be any event. Show that

$$P_{blue}(A) - P_{gold}(A) \; = \; \sum_{i \in AB} (b_i - g_i) \; - \; \sum_{i \in AG} (g_i - b_i)$$

Hence show that

$$P_{blue}(A) - P_{gold}(A) \; \leq \; \sum_{i \in AB} (b_i - g_i) \qquad \text{and} \qquad P_{gold}(A) - P_{blue}(A) \; \leq \; \sum_{i \in AG} (g_i - b_i)$$

b) [WRITTEN] Use the first of the two inequalities in (a) to show that if $P_{blue}(A) - P_{gold}(A) > 0$ then

$$|P_{blue}(A) - P_{gold}(A)| \ \leq \ \sum_{i \in B}(b_i - g_i)$$

Use the second of the two inequalities in (a) to show that if $P_{blue}(A) - P_{gold}(A) < 0$ then

$$|P_{blue}(A) - P_{gold}(A)| \ \leq \ \sum_{i \in G}(g_i - b_i)$$

c) [WRITTEN] Identify an event for which one of the inequalities in **b** is an equality.

Explain why you now have a complete proof of

$$\max\{|P_{blue}(A) - P_{gold}(A)| : A \subseteq S\} \ = \ \frac{1}{2}\sum_{i \in S}|b_i - g_i|$$

That is, our usual calculation of the TVD is equivalent to finding the biggest difference between probabilities assigned by the two distributions to any event.

3.2 Submission Instructions

Many assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

3.2.1 Written Portion

- Scan all the pages into a PDF. You can use any scanner or a phone using an application. Please **DO NOT** simply take pictures using your phone.
- Please start a new page for each question. If you have already written multiple questions on the same page, you can crop the image or fold your page over (the old-fashioned way). This helps expedite grading.
- It is your responsibility to check that all the work on all the scanned pages is legible.

3.2.2 Code Portion

- Save your notebook using File > Save and Checkpoint.
- Generate a PDF file using File > Download as > PDF via LaTeX. This might take a few seconds and will automatically download a PDF version of this notebook.
  - If you have issues, please make a follow-up post on the general HW 3 Piazza thread.

3.2.3 Submitting

- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so.
- Submit the assignment to Homework 3 on Gradescope.
- **Make sure to assign each page of your pdf to the correct question.**
- **It is your responsibility to verify that all of your work shows up in your final PDF submission.**

Hw 3

1.

e. let $1 \le m < n \Rightarrow$ integers $\Rightarrow n$ (independent & with replacement).

$\Rightarrow p$ success in each trail

$\Rightarrow V \rightarrow$ # success in first m trails

$\Rightarrow W \rightarrow$ # success in all n trails.

$\Rightarrow$ fix integer $w \Rightarrow 0 \le w \le n$.

$\Rightarrow$ for $0 \le v \le \min(w, m)$

$\Rightarrow$ find $P(V=v \mid W=w)$

$$\Rightarrow P(V=v \mid W=w) = \frac{P(VW)}{P(W)}$$

$$= \frac{\binom{m}{v} p^v (1-p)^{m-v} \cdot \binom{n-m}{w-v} p^{w-v} \cdot (1-p)^{n-m-(w-v)}}{\binom{n}{w} \cdot p^w \cdot (1-p)^{n-w}}$$

$$= \frac{\binom{m}{v} \cdot \cancel{p^w} \cdot \cancel{(1-p)^{n-w}} \cdot \binom{n-m}{w-v}}{\binom{n}{w} \cdot \cancel{p^w} \cdot \cancel{(1-p)^{n-w}}}$$

$$\Rightarrow P(V=v \mid W=w) = \frac{\binom{m}{v}\cdot\binom{n-m}{w-v}}{\binom{n}{w}}$$

⇒ Thus, this probability does not depend on p

∴

⇒ binomial distribution

⇒ parameters n and $w \cdot \frac{v}{n-m}$

2.

X phone calls from people know

Y phone calls from people don't know.

X has poisson (λ) distribution)

Y has poisson (μ) distribution.

X & Y are independent.

calls → 0.1 chance of being missed.

a.

possion distribution & Possion ($\mu + \lambda$)

⇒ if $X \to P(\mu)$ & $Y \to P(\lambda) \Rightarrow$ possion

then $S = X + Y \Rightarrow \text{Poisson}(\mu + \lambda)$

b.

⇒ let Total = total number of calls receive

let Y = Number of missed calls

$$P(\text{Total} = k,\ Y = y)$$

$$= e^{-(\mu+\lambda)} \cdot \frac{(\mu+\lambda)^k}{k!} \cdot \frac{k!}{y!(k-y)!} \cdot (0.1)^y (0.9)^{k-y}$$

⇒ Total number of missed call given k is binomial distribution, with parameter $(k, 0.1)$

c.

⇒ let y = number of missed calls

N = Total number of calls.

$$\Rightarrow P(Y = y) = \sum_{n=y}^{\infty} P(N = n,\ Y = y)$$

$$= e^{-(\mu+\lambda)} \cdot \frac{(0.1 \cdot (\mu+\lambda))^y}{y!}$$

⇒ Thus, the distribution is Possion distribution with the parameter $(\mu + \lambda) \cdot 0.1$

d.

$\Rightarrow$ people he known $\rightarrow$ Possion (λ)

$$\Rightarrow P = \sum_{x=0}^{n} e^{\lambda} \cdot \frac{\lambda^x}{x!}$$

e.

$\Rightarrow$ $\nearrow$ at most n knowns $\Rightarrow$ Possion (λ)

$\searrow$ at least m unknowns $\Rightarrow$ possion (μ)

$$\Rightarrow P = \left(\sum_{i=0}^{n} e^{-\lambda} \cdot \frac{\lambda^i}{i!} \right) \cdot \left(\sum_{j=m}^{\infty} e^{-\mu} \cdot \frac{\mu^j}{j!} \right)$$

$$= \left(\sum_{i=0}^{n} e^{-\lambda} \cdot \frac{\lambda^i}{i!} \right) \cdot \left(1 - \sum_{j=0}^{m-1} e^{-\mu} \cdot \frac{\mu^j}{j!} \right).$$

3.

a.

exact $\Rightarrow$ binomial $(n, p) \Rightarrow$ binomial $(1000, 0.003)$

approximate $\Rightarrow$ Possion $(n \cdot p) \Rightarrow$ Possion $(\lambda = 3)$

b.

⇒ let k = number of success

then the total number of failure f = n − k

$$\Rightarrow B(p,n) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \binom{n}{f} (1-p)^f (p)^{n-f}$$

⇒ let a = 1−p → since p is large, then a is small

$$B(p,n) = \binom{n}{f} a^f (1-a)^{n-f}$$

⇒ where f = n−k and a = 1−p

⇒ this fits the drive we have in class

$$\Rightarrow B(p,n) = e^{-\mu} \cdot \frac{\mu^f}{f!} \qquad \mu = n(1-p) \Rightarrow f = n-k$$

$$= e^{-n(1-p)} \cdot \frac{(n(1-p))^{n-k}}{(n-k)!}$$

4.

a.

$P(\Omega) = 1$

$\Rightarrow$ and the outcome space $\Rightarrow \Omega = B + G + E$

$\Rightarrow$ Thuse $P(\Omega) = P(B) + P(G) + P(E)$

$\Rightarrow$ Since $P_{blue}(A) = \sum_{i \in A} b_i$

$\Rightarrow \sum_{i \in B} b_i + \sum_{i \in G} b_i + \sum_{i \in E} b_i$

$= P_{blue}(B) + P_{blue}(G) + P_{blue}(E)$

$= P_{blue}(\Omega)$

$= 1$

$\Rightarrow$ Since P_{blue} & P_{gold} a the whole outcome space for that color

$$\sum_{i \in B} g_i + \sum_{i \in G} g_i + \sum_{i \in E} g_i = P_{gold}(\Omega) = 1.$$

$\Rightarrow$ from above we have prove that

$$\sum_{i \in B} g_i + \sum_{i \in G} g_i + \sum_{i \in E} g_i = \sum_{i \in B} b_i + \sum_{i \in G} b_i + \sum_{i \in E} b_i$$

$\Rightarrow \sum g_i$ & $\sum b_i$ describe the same event

$i \in S$ $\quad$ $i \in B$

thus $\sum_{i \in E} g_i = \sum_{i \in B} b_i$

$\Rightarrow \sum_{i \in B} g_i + \sum_{i \in G} g_i = \sum_{i \in B} b_i + \sum_{i \in G} b_i$

$\sum_{i \in B} g_i - \sum_{i \in B} b_i = \sum_{i \in G} b_i - \sum_{i \in G} g_i$

$\Rightarrow \sum_{i \in B} (b_i - g_i) = \sum_{i \in G} (g_i - b_i)$

b. $\Rightarrow S = G + E + B$

$\frac{1}{2} \sum_{i \in S} |b_i - g_i|$

$= \frac{1}{2} \left[\sum_{i \in B} |b_i - g_i| + \sum_{i \in G} |b_i - g_i| + \sum_{i \in E} \overset{0}{|b_i - g_i|} \right]$

$= \frac{1}{2} \left[\sum_{i \in B} |b_i - g_i| + \sum_{i \in G} |b_i - g_i| \right]$

$\Rightarrow$ since $\sum_{i \in B} (b_i - g_i) = \sum_{i \in G} (g_i - b_i)$

① when $i \in B$

$b_i > g_i \Rightarrow b_i - g_i > 0$

$\Rightarrow \sum_{i \in B} |b_i - g_i| = \sum_{i \in B} (b_i - g_i)$

when $i \in G$

② when $i \in G$

$g_i - b_i \Rightarrow b_i - g_i < 0$.

$$\sum_{i \in G} |b_i - g_i| = \sum_{i \in G} g_i - b_i$$

$$\Rightarrow \text{original} = \frac{1}{2}\left[\sum_{i \in B}(b_i - g_i) + \sum_{i \in G}(g_i - b_i)\right]$$

$$\Rightarrow \text{since } \sum_{i \in B}(b_i - g_i) = \sum_{i \in G}(g_i - b_i)$$

$$\Rightarrow \text{original} = \frac{1}{2} \cdot 2 \sum_{i \in B}(b_i - g_i) = \frac{1}{2} \cdot 2 \sum_{i \in G}(g_i - b_i)$$

$$\Rightarrow \frac{1}{2}\sum_{i \in S}|b_i - g_i| = \sum_{i \in B}(b_i - g_i) = \sum_{i \in G}(g_i - b_i).$$

5.

a.

$$P_{blue}(k) = \sum_{i \in AB} b_i + \sum_{i \in AG} b_i + \sum_{i \in AE} b_i$$

$$P_{gold}(k) = \sum_{i \in AB} g_i + \sum_{i \in AG} g_i + \sum_{i \in AE} g_i$$

$$\Rightarrow P_{blue}(A) - P_{gold}(A)$$

$$= \sum_{i \in AB}(b_i - g_i) + \sum_{i \in AG}(b_i - g_i) + \sum_{i \in AE}(b_i - g_i)$$

$$\Rightarrow \text{since } \sum_{i \in AE} b_i = \sum_{i \in AE} g_i$$

$$= \sum_{i \in AB} (b_i - g_i) + \sum_{i \in AG} (b_i - g_i)$$

$$= \sum_{i \in AB} (b_i - g_i) - \sum_{i \in AG} (g_i - b_i).$$

$\Rightarrow$ in B $b_i > g_i \rightarrow$ thus $b_i - g_i > 0$ in $i \in AB$

in G $b_i < g_i \rightarrow$ thus $g_i - b_i > 0$ in $i \in AB$

$\Rightarrow$ since $\sum_{i \in AG} (g_i - b_i) > 0.$

$$\sum_{i \in AB} (b_i - g_i) - \sum_{i \in AG} (g_i - b_i) \leq \sum_{i \in AG} (g_i - b_i)$$

$$\Rightarrow P_{blue}(A) - P_{gold}(A) \leq \sum_{i \in AB} (b_i - g_i)$$

$$\Rightarrow -(P_{blue}(A) - P_{gold}(A)) = \sum_{i \in AG} (g_i - b_i) - \sum_{i \in AB} (b_i - g_i)$$

since $\sum_{i \in AB} (b_i - g_i) > 0$

$$\Rightarrow P_{gold}(A) - P_{blue}(A) \leq \sum_{i \in AG} (g_i - b_i)$$

b.

① $P_{blue}(A) - P_{gold}(A) > 0 \rightarrow |P_{blue}(A) - P_{gold}(A)| \leq \sum_{i \in B} (b_i - g_i)$

$\Rightarrow$ since $P_{blue}(A) - P_{gold}(A) > 0$

$\rightarrow$ $|P$ (illegible)

$|P_{blue}(A) - P_{gold}(A)| = P_{blue}(A) - P_{gold}(A)$

⇒ for the equation $P_{blue}(A) - P_{gold}(A) \le \sum_{i \in B}(b_i - g_i)$ to hold true

→ since $P_{blue}(A) - P_{gold}(A) \le \sum_{i \in AB}(b_i - g_i)$

⇒ $\sum_{i \in AB}(b_i - g_i) \le \sum_{i \in B}(b_i - g_i)$

⇒ for any event $A \rightarrow P(AB) \le P(B)$
↳ less likely (more condition)

⇒ Thus $\sum_{i \in AB}(b_i - g_i)$ always $\le \sum_{i \in B}(b_i - g_i)$

⇒ so it's true that $|P_{blue}(A) - P_{gold}(A)| \le \sum_{i \in B}(b_i - g_i)$

⊖ if $P_{blue}(A) - P_{gold}(A) < 0$

$|P_{blue}(A) - P_{gold}(A)| = P_{gold}(A) - P_{blue}(A)$

⇒ since $P_{gold}(A) - P_{blue}(A) \le \sum_{i \in AG}(g_i - b_i)$

for the equation to hold true

$\sum_{i \in AG}(g_i - b_i)$ must $\le \sum_{i \in G}(g_i - b_i)$

⇒ for any event $P(AG)$ alway $\le P(G)$

⇒ Thus $\sum_{i \in AG}(g_i - b_i) \le \sum_{i \in G}(g_i - b_i)$ always hold true

⇒ Thus, under the given condition

$$|P_{blue}(A) - P_{gold}(A)| \leq \sum_{i \in G} (g_i - b_i)$$

c.

$$\max \{ |P_{blue}(A) - P_{gold}(A)| : A \subseteq S \}$$

can only be $\Rightarrow$ ① $\sum_{i \in B} (b_i - g_i)$

② $\sum_{i \in G} (g_i - b_i)$

$\Rightarrow$ we have also prove that

$$\frac{1}{2} \sum_{i \in S} |b_i - g_i| = \sum_{i \in B} (b_i - g_i) = \sum_{i \in G} (g_i - b_i)$$

$\Rightarrow$ thus

$$\max \{ |P_{blue}(A) - P_{gold}(A)| : A \subseteq S \} = \frac{1}{2} \sum_{i \in S} |b_i - g_i|$$

$\Rightarrow$ for equality in b.

$P(AB) = P(B)$ or $P(AB) = P(G)$

$\Rightarrow$ Thus for equality $A = B$ or $A = G$