In [1]:
```python
from IPython.core.display import HTML
def css_styling():
    styles = open("cs109.css", "r").read();
    return HTML(styles)
css_styling()
```

Out[1]:

S-109A Introduction to Data Science

Homework 0

Harvard University
Summer 2018
Instructors: Pavlos Protopapas and Kevin Rader

This is a homework which you must turn in but it won't be graded.

This homework has the following intentions:

1. To get you familiar with the jupyter/python environment.
2. You should easily understand these questions and what is being asked. If you struggle, this may not be the right class for you.
3. You should be able to understand the intent (if not the exact syntax) of the code and be able to look up on Google and provide code that is asked of you. If you cannot, this may not be the right class for you.

Basic Math and Probability/Statistics Calculations

We'll start you off with some basic math and statistics problems questions to make sure you have the appropriate background to be comfortable with concepts that will come up in CS 109a.

Question 1: Mathiness is What Brings Us Together Today

Matrix Operations

Complete the following matrix operations (show your work as a markdown/latex notebook cell)

1.1. Let $A = \begin{pmatrix} 3 & 4 & 2 \\ 5 & 6 & 4 \\ 4 & 3 & 4 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 4 & 2 \\ 1 & 9 & 3 \\ 2 & 3 & 3 \end{pmatrix}$.

Compute $A \cdot B$.

1.2. Let $A = \begin{pmatrix} 0 & 12 & 8 \\ 1 & 15 & 0 \\ 0 & 6 & 3 \end{pmatrix}$.

Compute A^{-1}.

Answers

1.1 $A \cdot B = \begin{pmatrix} 3+4+4 & 12+36+6 & 6+12+6 \\ 5+6+8 & 20+54+12 & 10+18+12 \\ 4+3+8 & 16+27+12 & 8+9+12 \end{pmatrix} = \begin{pmatrix} 11 & 54 & 24 \\ 19 & 86 & 40 \\ 15 & 55 & 29 \end{pmatrix}$

1.2

$$\left(\begin{array}{ccc|ccc} 0 & 12 & 8 & 1 & 0 & 0 \\ 1 & 15 & 0 & 0 & 1 & 0 \\ 0 & 6 & 3 & 0 & 0 & 1 \end{array}\right) \xrightarrow[\text{\& R2}]{\text{switch R1}} \left(\begin{array}{ccc|ccc} 1 & 15 & 0 & 0 & 1 & 0 \\ 0 & 12 & 8 & 1 & 0 & 0 \\ 0 & 6 & 3 & 0 & 0 & 1 \end{array}\right) \xrightarrow{\text{R2/12}}$$

$$\left(\begin{array}{ccc|ccc} 1 & 15 & 0 & 0 & 1 & 0 \\ 0 & 1 & \frac{2}{3} & \frac{1}{12} & 0 & 0 \\ 0 & 6 & 3 & 0 & 0 & 1 \end{array}\right) \xrightarrow[\text{R3 - 6*R2}]{\text{R1 - 15*R2}} \left(\begin{array}{ccc|ccc} 1 & 0 & -10 & \frac{-5}{4} & 1 & 0 \\ 0 & 1 & \frac{2}{3} & \frac{1}{12} & 0 & 0 \\ 0 & 0 & -1 & \frac{-1}{2} & 0 & 1 \end{array}\right) \xrightarrow{\text{R3/-1}}$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & -10 & \frac{-5}{4} & 1 & 0 \\ 0 & 1 & \frac{2}{3} & \frac{1}{12} & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{2} & 0 & -1 \end{array}\right) \xrightarrow[\text{R2 - }\frac{2}{3}\text{*R2}]{\text{R1 + 10*R3}} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{15}{4} & 1 & -10 \\ 0 & 1 & 0 & \frac{-1}{4} & 0 & \frac{2}{3} \\ 0 & 0 & 1 & \frac{1}{2} & 0 & -1 \end{array}\right)$$

$$A^{-1} = \begin{pmatrix} \frac{15}{4} & 1 & -10 \\ \frac{-1}{4} & 0 & \frac{2}{3} \\ \frac{1}{2} & 0 & -1 \end{pmatrix}$$

Calculus and Probability

Complete the following (show your work as a markdown/latex notebook cell)

1.3. From Wikipedia:

> In mathematical optimization, statistics, econometrics, decision theory, machine learning and computational neuroscience, a loss function or cost function is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event. An optimization problem seeks to minimize a loss function.

We've generated a cost function on parameters $x, y \in \mathcal{R}$ $L(x, y) = 3x^2y - y^3 - 3x^2 - 3y^2 + 2$. Find the critical points (optima) of $L(x, y)$.

1.4. Two central aspects of call center operations are the busy-ness of the customer service reps and the caller demographics. Because of historical data, the distribution of these measures can often take on well-known distributions. In the CS109 Homework Helpdesk, X and Y are discrete random variables with Y measuring the total number of help lines that are busy with callers (there are three lines available), and X measuring the number of help lines that are busy with female callers. We've determined historically the joint pmf of (X, Y) and found it to be:

$$p_{X,Y}(x, y) = \binom{y}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{y-x}\frac{(1+y)}{10}$$

where $y \in 0, 1, 2, 3, x \in [0, y]$ (That is to say the total number of callers in a minute is a non-negative integer and the number of female callers naturally assumes a value between 0 and the total number of callers inclusive). Note: by definition, $\binom{0}{0} = 1$.

(i) What is the conditional distribution of $X|Y$ (it is a well-known family)?

(ii) Find the mean and variance of the marginal distribution of X.

Answers

1.3

$\frac{dL}{dx} = 6xy - 6x = 0$ $\qquad$ $\frac{dL}{dy} = 3x^2 - 3y^2 - 6y = 0$

$6x(y-1) = 0$

$x = 0$ or $y = 1$

$\frac{dL}{dy}(x = 0) = -3y^2 - 6y = -3y(y+2) = 0 \implies y = 0$ or $-2,\ x = 0$

$\frac{dL}{dy}(y = 1) = 3x^2 - 3 - 6 = 3x^2 - 9 = 0 \implies x = \pm\sqrt{3},\ y = 1$

Critical points are: $(0, 0),\ (0, -2),\ (\sqrt{3}, 1),\ (-\sqrt{3}, 1)$

In [2]:
```python
from ipy_table import *
print('**1.4**')
print('Table of Probabilities')
joint_pmf = [
    [' ', 'y=1', 'y=2', 'y=3', 'y=4', 'p(X)'],
    ['x0', 1/10, 2/15, 2/15, 16/135, 131/270],
    ['x1', 0, 1/15, 2/15, 8/45, 17/45],
    ['x2', 0, 0, 1/30, 4/45, 11/90],
    ['x3', 0, 0, 0, 2/135, 2/135]];
make_table(joint_pmf)
apply_theme('basic_both')
```

```
**1.4**
Table of Probabilities
```

Out[2]:

	y=1	y=2	y=3	y=4	p(X)
x0	0.1000	0.1333	0.1333	0.1185	0.4852
x1	0	0.0667	0.1333	0.1778	0.3778
x2	0	0	0.0333	0.0889	0.1222
x3	0	0	0	0.0148	0.0148

In [3]:
```python
from ipy_table import *
print('Conditional Distribution of X|Y')
print('top axis: y-values; left axis: x-values')
x_y = [
    [' ', '0', '1', '2', '3'],
    ['0', 1, 2/3, 4/9, 8/27],
    ['1', 0, 1/3, 4/9, 4/9],
    ['2', 0, 0, 1/9, 2/9],
    ['3', 0, 0, 0, 1/27]];
make_table(x_y)
apply_theme('basic_both')
```

```
Conditional Distribution of X|Y
top axis: y-values; left axis: x-values
```

Out[3]:

	0	1	2	3
0	1	0.6667	0.4444	0.2963
1	0	0.3333	0.4444	0.4444
2	0	0	0.1111	0.2222
3	0	0	0	0.0370

The conditional distribution of X|Y is a binomial distribution, with p = 1/3.
Note - wasn't able to get Markdown or Latex tables to work for this problem, so used a different format which didn't provide as extensive options.

In [4]:
```
print('expectation of marginal distribution of X:')
exp_X = 0*131/270 + 1*17/45 + 2*11/90 + 3*2/135
print(round(exp_X, 2))
print('variance of marginal distribution of X:')
exp_X2 = 0*131/270 + 1*17/45 + (2**2)*11/90 + (3**2)*2/135
var_X = exp_X2 - (exp_X)**2
print(round(var_X,2))
```

```
expectation of marginal distribution of X:
0.67
variance of marginal distribution of X:
0.56
```

In [5]:
```
# The line %... is a jupyter "magic" command, and is not part of the Python
# In this case we're just telling the plotting library to draw things on
# the notebook, instead of on a separate window.
%matplotlib inline
# See the "import ... as ..." contructs below? They're just aliasing the pac
# That way we can call methods like plt.plot() instead of matplotlib.pyplot.
import numpy as np
import scipy as sp
import scipy.stats as stats
import matplotlib.pyplot as plt
```

Basic Statistics

Complete the following: you can perform the calculations by hand (show your work) or using software (include the code and output, screenshots are fine if it is from another platform).

1.5. 37 of the 76 female CS concentrators have taken Data Science 1 (DS1) while 50 of the 133 male concentrators haven taken DS1. Perform a statistical test to determine if interest in Data Science (by taking DS1) is related to sex. Be sure to state your conclusion.

Answer: With a p-value > 0.05 (0.078), there is not a significant difference between the sexes wrt interest in data science.

```r
library(dplyr)

fem_ds <- c(rep(1, 37), rep(0, 76-37))
fem_sex <- c(rep('Female', 76))
male_ds <- c(rep(1, 50), rep(0, 133-50))
male_sex <- c(rep('Male', 133))
fem_tab <- data.frame(fem_ds, fem_sex)
fem_tab <- rename(fem_tab, sex = fem_sex)
fem_tab <- rename(fem_tab, DS1 = fem_ds)
male_tab <- data.frame(male_ds, male_sex)
male_tab <- rename(male_tab, sex = male_sex)
male_tab <- rename(male_tab, DS1 = male_ds)
full_ds <- bind_rows(fem_tab, male_tab)
id_m <- which(full_ds$sex == "Male")
ds_m <- mean(male_ds)
id_f <- which(full_ds$sex == "Female")
ds_f <- mean(fem_ds)
obs <- ds_f - ds_m; obs
ds_col <- full_ds$DS1
# permutation test
N <- 10^5 - 1
diffs <- numeric(N)
for (i in 1:N) {
  # sample size of 76, from 1 to 209, without replacement
  index <- sample(209, size = 76, replace = FALSE)
  diffs[i] <- mean(ds_col[index]) - mean(ds_col[-index])
}
# calculate p-value
pvalue <- (sum(diffs >= obs)+1)/(N+1); pvalue
```

```
> # calculate p-value
> pvalue <- (sum(diffs >= obs)+1)/(N+1); pvalue
[1] 0.07769
```

r code

```
library(dplyr)


fem_ds <- c(rep(1, 37), rep(0, 76-37))
fem_sex <- c(rep('Female', 76))
male_ds <- c(rep(1, 50), rep(0, 133-50))
male_sex <- c(rep('Male', 133))
fem_tab <- data.frame(fem_ds, fem_sex)
fem_tab <- rename(fem_tab, sex = fem_sex)
fem_tab <- rename(fem_tab, DS1 = fem_ds)
male_tab <- data.frame(male_ds, male_sex)
male_tab <- rename(male_tab, sex = male_sex)
male_tab <- rename(male_tab, DS1 = male_ds)
full_ds <- bind_rows(fem_tab, male_tab)
id_m <- which(full_ds$sex == 'Male')
ds_m <- mean(male_ds)
id_f <- which(full_ds$sex == 'Female')
ds_f <- mean(fem_ds)
obs <- ds_f - ds_m; obs
ds_col <- full_ds$DS1
N <- 10^5 - 1
diffs <- numeric(N)
for (i in 1:N) {
  index <- sample(209, size = 76, replace = FALSE)
  diffs[i] <- mean(ds_col[index]) - mean(ds_col[-index])
}
pvalue = obs)+1)/(N+1); pvalue
```

Simulation of a Coin Throw

We'd like to do some experiments with coin flips, but we don't have a physical coin at the moment. So let's **simulate** the process of flipping a coin on a computer. To do this we will use a form of the **random number generator** built into `numpy`. In particular, we will use the function `np.random.choice` which picks items with uniform probability from a list. If we provide it a list ['H', 'T'], it will pick one of the two items in the list. We can also ask it to do this multiple times by specifying the parameter `size`.

In [6]:
```python
def throw_a_coin(n_trials):
    return np.random.choice(['H','T'], size=n_trials)
```

`np.sum` is a function that returns the sum of items in an iterable (i.e. a list or an array). Because python coerces `True` to 1 and `False` to 0, the effect of calling `np.sum` on the array of `True` s and `False` s will be to return the number of of `True` s in the array (which can then effectively count the number of heads).

Question 2: The 12 Labors of Bernoullis

Now that we know how to run our coin flip experiment, we're interested in knowing what happens as we choose larger and larger number of coin flips.

2.1. Run one experiment of flipping a coin 40 times storing the resulting sample in the variable `throws1` . What's the total proportion of heads?

2.2. **Replicate** the experiment in 2.1 storing the resulting sample in the variable `throws2` . What's the proportion of heads? How does this result compare to that you obtained in question 2.1?

2.3. Write a function called `run_trials` that takes as input a list, called `n_flips` , of integers representing different values for the number of coin flips in a trial. For each element in the input list, `run_trials` should run the coin flip experiment with that number of flips and calculate the proportion of heads. The output of `run_trials` should be the list of calculated proportions. Store the output of calling `run_trials` in a list called `proportions` .

2.4. Using the results in 2.3, reproduce the plot below.

2.5. What's the appropriate observation about the result of running the coin flip experiment with larger and larger numbers of coin flips? Choose the appropriate one from the choices below.

> A. Regardless of sample size the probability of in our experiment of observing heads is 0.5 so the proportion of heads observed in the coin-flip experiments will always be 0.5.
>
> B. The proportions **fluctuate** about their long-run value of 0.5 (what you might expect if you tossed the coin an infinite amount of times), in accordance with the notion of a fair coin (which we encoded in our simulation by having `np.random.choice` choose between two possibilities with equal probability), with the fluctuations seeming to become much smaller as the number of trials increases.
>
> C. The proportions **fluctuate** about their long-run value of 0.5 (what you might expect if you tossed the coin an infinite amount of times), in accordance with the notion of a fair coin (which we encoded in our simulation by having `np.random.choice` choose between two possibilities with equal probability), with the fluctuations constant regardless of the number of trials.

Answers

2.1

In [7]:
```
throws1 = throw_a_coin(40)
print(throws1)
print(np.sum(throws1 == "H")/40)
```

```
['H' 'H' 'T' 'T' 'T' 'H' 'T' 'T' 'T' 'T' 'H' 'H' 'T' 'T' 'H' 'H' 'T' 'T'
 'H' 'H' 'H' 'T' 'T' 'H' 'T' 'H' 'T' 'T' 'T' 'H' 'T' 'H' 'T' 'H' 'H' 'T'
 'T' 'H' 'H' 'T']
0.45
```

2.2

In [8]:
```
throws2 = throw_a_coin(40)
print(throws2)
print(np.sum(throws2 == "H")/40)
```

```
['T' 'H' 'H' 'T' 'H' 'H' 'H' 'H' 'H' 'T' 'T' 'H' 'H' 'H' 'H' 'T' 'T' 'H'
 'H' 'T' 'H' 'T' 'T' 'T' 'H' 'T' 'H' 'H' 'H' 'H' 'H' 'T' 'T' 'T' 'H' 'H'
 'T' 'H' 'H' 'T']
0.6
```

2.3

In [9]:
```
flips_list = [10, 30, 50, 70, 100, 130, 170, 200, 500, 1000, 2000, 5000, 10(
```

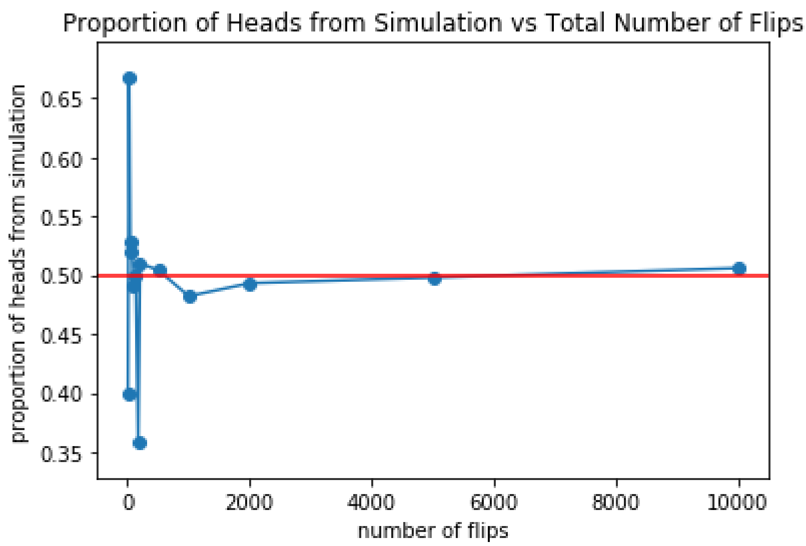
In [10]:
```
def run_trials(n_flips):
    result = list()
    for i in range(0, len(n_flips)):
        throw = throw_a_coin(n_flips[i])
        heads = (np.sum(throw == "H")) / n_flips[i]
        result.append(heads)
    return result
```

In [11]:
```
proportions = run_trials(flips_list)
print(proportions)
```

```
[0.4, 0.6666666666666666, 0.52, 0.5285714285714286, 0.49, 0.5, 0.35882352
94117647, 0.51, 0.504, 0.482, 0.493, 0.4978, 0.5058]
```

2.4

```
In [12]: plt.scatter(flips_list, proportions)
plt.plot(flips_list, proportions)
plt.axhline(y=0.5, color='r', linestyle='-')
plt.xlabel('number of flips'); plt.ylabel('proportion of heads from simulati
plt.title('Proportion of Heads from Simulation vs Total Number of Flips');
```



2.5

What's the appropriate observation about the result of applying the coin flip experiment to larger and larger numbers of coin flips? Choose the appropriate one.

Because of the Central Limit Theorem, option B is correct. As the n increases, the proportion of heads gets closer to 50%. Fluctuations are getting much smaller.
"The proportions fluctuate about their long-run value of 0.5 (what you might expect if you tossed the coin an infinite amount of times), in accordance with the notion of a fair coin (which we encoded in our simulation by having np.random.choice choose between two possibilities with equal probability), with the fluctuations seeming to become much smaller as the number of trials increases."

Multiple Replications of the Coin Flip Experiment

The coin flip experiment that we did above gave us some insight, but we don't have a good notion of how robust our results are under repetition as we've only run one experiment for each number of coin flips. Lets redo the coin flip experiment, but let's incorporate multiple repetitions of each number of coin flips. For each choice of the number of flips, n, in an experiment, we'll do M replications of the coin tossing experiment.

Answers

Question 3. So Many Replications

3.1. Write a function `make_throws` which takes as arguments the `n_replications` (M) and the `n_flips` (n), and returns a list (of size M) of proportions, with each proportion calculated by taking the ratio of heads to to total number of coin flips in each replication of n coin tosses. `n_flips` should be a python parameter whose value should default to 20 if unspecified when `make_throws` is called.

3.2. Create the variables `proportions_at_n_flips_100` and `proportions_at_n_flips_1000`. Store in these variables the result of `make_throws` for `n_flips` equal to 100 and 1000 respectively while keeping `n_replications` at 200. Create a plot with the histograms of `proportions_at_n_flips_100` and `proportions_at_n_flips_1000`. Make sure to title your plot, label the x-axis and provide a legend.(See below for an example of what the plot may look like)

3.3. Calculate the mean and variance of the results in the each of the variables `proportions_at_n_flips_100` and `proportions_at_n_flips_1000` generated in 3.2.

3.4. Based upon the plots what would be your guess of what type of distribution is represented by histograms in 3.2? Explain the factors that influenced your choice.

> A. Gamma Distribution
>
> B. Beta Distribution
>
> C. Gaussian

3.5. Let's just assume for arguments sake that the answer to 3.4 is **C. Gaussian**. Plot a **normed histogram** of your results `proportions_at_n_flips_1000` overlayed with your selection for the appropriate gaussian distribution to represent the experiment of flipping a coin 1000 times. (**Hint: What parameters should you use for your Gaussian?**)

3.1

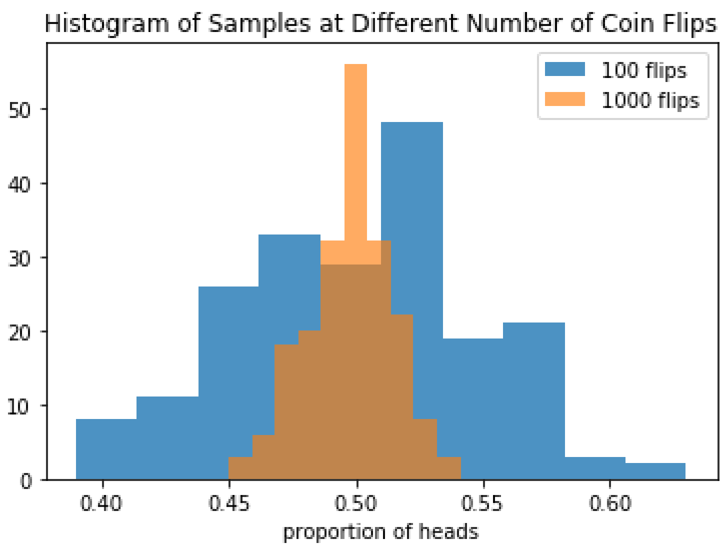
In [13]:
```
def make_throws(n_replications, n_flips=20):
    flips = np.repeat(n_flips, n_replications)
    return run_trials(flips)
```

3.2

In [14]:
```python
proportions_at_n_flips_100 = make_throws(200, 100)
proportions_at_n_flips_1000 = make_throws(200, 1000)
print(proportions_at_n_flips_100)
print(len(proportions_at_n_flips_100))
print(proportions_at_n_flips_1000)
print(len(proportions_at_n_flips_1000))
```

```
[0.56, 0.55, 0.48, 0.49, 0.5, 0.51, 0.52, 0.51, 0.43, 0.54, 0.52, 0.56,
0.46, 0.53, 0.55, 0.46, 0.48, 0.47, 0.57, 0.46, 0.5, 0.51, 0.53, 0.54, 0.
56, 0.54, 0.52, 0.42, 0.47, 0.46, 0.45, 0.54, 0.52, 0.5, 0.47, 0.4, 0.58,
0.51, 0.49, 0.5, 0.48, 0.47, 0.56, 0.56, 0.51, 0.47, 0.58, 0.51, 0.51, 0.
46, 0.55, 0.41, 0.51, 0.48, 0.6, 0.39, 0.55, 0.51, 0.48, 0.48, 0.43, 0.4
2, 0.5, 0.49, 0.49, 0.47, 0.56, 0.52, 0.49, 0.53, 0.61, 0.47, 0.55, 0.46,
0.6, 0.53, 0.49, 0.53, 0.51, 0.43, 0.54, 0.5, 0.48, 0.41, 0.54, 0.42, 0.5
4, 0.52, 0.56, 0.43, 0.58, 0.44, 0.51, 0.5, 0.47, 0.51, 0.48, 0.46, 0.41,
0.56, 0.48, 0.54, 0.44, 0.52, 0.46, 0.46, 0.57, 0.47, 0.44, 0.53, 0.4, 0.
47, 0.56, 0.46, 0.55, 0.45, 0.47, 0.53, 0.51, 0.52, 0.5, 0.49, 0.49, 0.4
6, 0.5, 0.53, 0.42, 0.5, 0.52, 0.53, 0.53, 0.45, 0.48, 0.49, 0.6, 0.47,
0.48, 0.51, 0.43, 0.52, 0.58, 0.52, 0.56, 0.51, 0.55, 0.49, 0.47, 0.47,
0.44, 0.41, 0.47, 0.44, 0.44, 0.47, 0.42, 0.51, 0.51, 0.47, 0.55, 0.58,
0.47, 0.53, 0.52, 0.5, 0.51, 0.53, 0.51, 0.58, 0.56, 0.58, 0.63, 0.49, 0.
55, 0.49, 0.57, 0.47, 0.52, 0.52, 0.54, 0.41, 0.46, 0.49, 0.46, 0.5, 0.4
8, 0.48, 0.43, 0.5, 0.53, 0.45, 0.5, 0.49, 0.46, 0.44, 0.53, 0.52, 0.5,
0.54, 0.47, 0.46]
200
[0.505, 0.502, 0.488, 0.474, 0.518, 0.487, 0.498, 0.48, 0.497, 0.517, 0.5
25, 0.485, 0.503, 0.489, 0.492, 0.501, 0.499, 0.512, 0.504, 0.498, 0.504,
0.476, 0.498, 0.503, 0.481, 0.516, 0.502, 0.509, 0.509, 0.488, 0.516, 0.
5, 0.499, 0.524, 0.498, 0.499, 0.485, 0.477, 0.5, 0.507, 0.492, 0.489, 0.
494, 0.49, 0.467, 0.45, 0.52, 0.463, 0.513, 0.489, 0.521, 0.502, 0.47, 0.
505, 0.509, 0.476, 0.494, 0.523, 0.487, 0.514, 0.498, 0.503, 0.474, 0.49
9, 0.488, 0.488, 0.493, 0.535, 0.475, 0.504, 0.489, 0.497, 0.514, 0.496,
0.487, 0.53, 0.525, 0.504, 0.506, 0.526, 0.479, 0.484, 0.503, 0.501, 0.50
4, 0.476, 0.501, 0.491, 0.456, 0.504, 0.501, 0.515, 0.52, 0.469, 0.509,
0.487, 0.482, 0.532, 0.502, 0.469, 0.476, 0.462, 0.484, 0.507, 0.484, 0.5
16, 0.506, 0.506, 0.482, 0.495, 0.488, 0.507, 0.502, 0.527, 0.476, 0.489,
0.463, 0.508, 0.496, 0.504, 0.514, 0.512, 0.506, 0.479, 0.51, 0.478, 0.50
1, 0.504, 0.502, 0.516, 0.491, 0.507, 0.501, 0.511, 0.507, 0.515, 0.5, 0.
475, 0.505, 0.453, 0.508, 0.497, 0.505, 0.502, 0.477, 0.52, 0.541, 0.511,
0.502, 0.503, 0.499, 0.491, 0.504, 0.467, 0.488, 0.502, 0.486, 0.484, 0.4
99, 0.475, 0.493, 0.522, 0.518, 0.513, 0.517, 0.467, 0.509, 0.501, 0.495,
0.514, 0.518, 0.5, 0.508, 0.479, 0.477, 0.484, 0.48, 0.475, 0.521, 0.501,
0.499, 0.506, 0.48, 0.491, 0.489, 0.527, 0.49, 0.481, 0.499, 0.473, 0.49
9, 0.497, 0.489, 0.507, 0.489, 0.515, 0.5, 0.507, 0.511, 0.484]
200
```

In [15]:
```
plt.hist(proportions_at_n_flips_100, alpha=0.8, label='100 flips')
plt.hist(proportions_at_n_flips_1000, alpha=0.65, label='1000 flips')
plt.legend()
plt.xlabel('proportion of heads')
plt.title('Histogram of Samples at Different Number of Coin Flips')
plt.show()
```



3.3

In [16]:
```
print('mean and variance for 100 flips:')
print(np.mean(proportions_at_n_flips_100))
print(np.var(proportions_at_n_flips_100))
print('mean and variance for 1000 flips:')
mean_1000 = np.mean(proportions_at_n_flips_1000)
var_1000 = np.var(proportions_at_n_flips_1000)
print(mean_1000)
print(var_1000)
```

```
mean and variance for 100 flips:
0.49965000000000004
0.0021743775000000005
mean and variance for 1000 flips:
0.4975
0.0002660500000000003
```

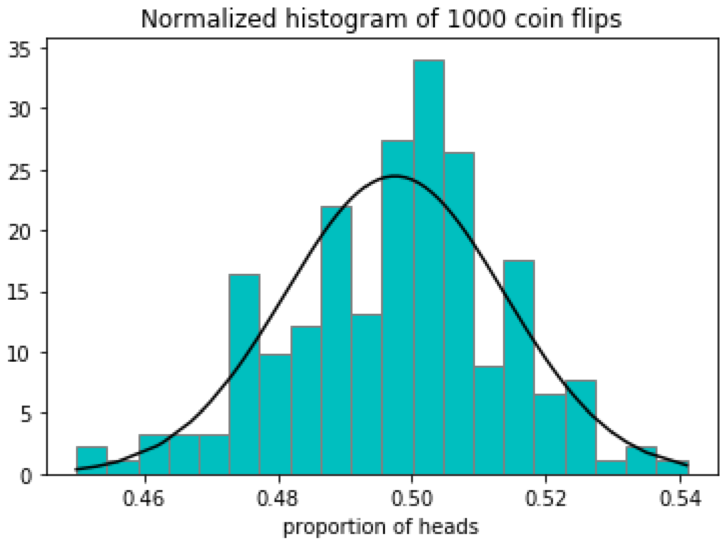
3.4

Gaussian distribution

As the *n* of our sample increases, the proportion of heads tends towards a Gaussian distribution. Follows Central Limit Theorem.

3.5

In [30]:
```
x = sorted(proportions_at_n_flips_1000)
fit = stats.norm.pdf(x, mean_1000, np.sqrt(var_1000))
plt.plot(x, fit, color = 'k')
plt.hist(x, density=True, bins=20, color='c', edgecolor="grey")
plt.xlabel('proportion of heads')
plt.title('Normalized histogram of 1000 coin flips')
plt.show()
```



Working With Distributions in Numpy/Scipy

Earlier in this problem set we've been introduced to the Bernoulli "aka coin-flip" distribution and worked with it indirectly by using np.random.choice to make a random selection between two elements 'H' and 'T'. Let's see if we can create comparable results by taking advantage of the machinery for working with other probability distributions in python using numpy and scipy.

Question 4: My Normal Binomial

Let's use our coin-flipping machinery to do some experimentation with the binomial distribution. The binomial distribution, often represented by $k \sim Binomial(n, p)$ is often discribed the number of successes in `n` Bernoulli trials with each trial having a probability of success `p`. In other words, if you flip a coin `n` times, and each coin-flip has a probability `p` of landing heads, then the number of heads you observe is a sample from a binomial distribution.

4.1. Sample the binomial distribution with $p = 0.5$ using coin flips by writing a function `sample_binomial1` which takes in integer parameters `n` and `size`. The output of `sample_binomial1` should be a list of length `size` observations with each observation being the outcome of flipping a coin `n` times and counting the number of heads. By default `size` should be 1. Your code should take advantage of the `throw_a_coin` function we defined above.

4.2. Sample the binomial distribution directly using scipy.stats.binom.rvs by writing another function `sample_binomial2` that takes in integer parameters `n` and `size` as well as a float `p` parameter `p` where $p \in [0 \ldots 1]$. The output of `sample_binomial2` should be a list of length `size` observations with each observation a sample of $Binomial(n, p)$ (taking advantage of scipy.stats.binom). By default `size` should be 1 and `p` should be 0.5.

4.3. Run sample_binomial1 with 25 and 200 as values of the `n` and `size` parameters respectively and store the result in `binomial_trials1`. Run sample_binomial2 with 25, 200 and 0.5 as values of the `n`, `size` and `p` parameters respectively and store the results in `binomial_trials2`. Plot normed histograms of `binomial_trials1` and `binomial_trials2`. On both histograms, overlay a plot of the pdf of $Binomial(n = 25, p = 0.5)$

4.4. How do the plots in 4.3 compare?

4.5. Find the mean and variance of `binomial_trials1`. How do they compare to the true mean and varaince of a $Binomial(n = 25, p = 0.5)$ distribution?

Answers

4.1

In [18]:
```
def sample_binomial1(n, size=1):
    flips = np.repeat(n, size)
    result = list()
    for i in range(0, len(flips)):
        throw = throw_a_coin(flips[i])
        heads = np.sum(throw == "H")
        result.append(heads)
    return result
```

4.2

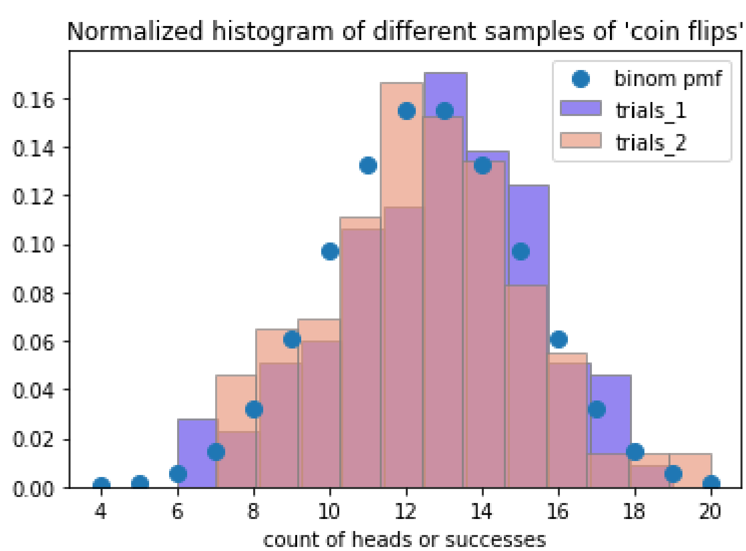
In [19]:
```
def sample_binomial2(n, size=1, p=0.5):
    return stats.binom.rvs(n=n, size=size, p=p)
```

4.3

In [20]:

```
binomial_trials1 = sample_binomial1(25, 200)
binomial_trials2 = sample_binomial2(25, 200, 0.5)
# check that these both worked appropriately:
print(binomial_trials1)
print(len(binomial_trials1))
print(binomial_trials2)
print(len(binomial_trials2))
# fit = stats.binom.pdf(x, mean_1000, np.sqrt(var_1000))
# plt.plot(x, fit, color = 'k')
plt.hist(binomial_trials1, density=True, bins=12, color='mediumslateblue', e
plt.hist(binomial_trials2, density=True, bins=12, color='darksalmon', edgec
plt.xlabel('count of heads or successes')
plt.title("Normalized histogram of different samples of 'coin flips'")
x = np.arange(stats.binom.ppf(0.0001, 25, 0.5), stats.binom.ppf(0.9999, 25,
plt.plot(x, stats.binom.pmf(x, 25, 0.5), 'o', ms=8, label='binom pmf')
plt.legend()
plt.show()
```

```
[16, 13, 12, 14, 13, 9, 11, 13, 11, 12, 13, 10, 13, 12, 17, 10, 11, 14, 1
4, 16, 11, 6, 13, 14, 15, 15, 17, 10, 10, 14, 15, 15, 16, 11, 9, 12, 15,
17, 11, 17, 11, 9, 13, 14, 13, 14, 9, 14, 10, 12, 15, 16, 14, 14, 16, 13,
15, 12, 13, 14, 7, 16, 13, 10, 12, 12, 13, 12, 12, 15, 13, 12, 8, 12, 11,
14, 15, 13, 11, 13, 14, 13, 10, 13, 13, 15, 8, 12, 12, 7, 13, 15, 11, 9,
11, 14, 14, 14, 14, 11, 15, 14, 17, 14, 13, 16, 17, 13, 14, 14, 15, 14, 1
0, 14, 8, 17, 12, 10, 12, 17, 14, 13, 17, 9, 12, 11, 9, 12, 12, 19, 13, 1
0, 17, 14, 14, 11, 13, 16, 13, 13, 16, 15, 11, 13, 6, 11, 15, 9, 14, 16,
13, 11, 15, 13, 15, 12, 11, 9, 13, 12, 11, 15, 7, 12, 15, 15, 13, 13, 13,
14, 15, 8, 15, 13, 18, 11, 11, 12, 15, 11, 14, 15, 7, 10, 13, 11, 12, 8,
15, 15, 15, 14, 13, 12, 10, 10, 9, 16, 9, 13]
200
[10 14 15 17  8 18 14 13 13 10 12  7 12 13 12 15 10 12 16 14 14  9 10 10
 12 13 10 14 15 14  8 10 18 11 11 11  7 14 12 16 14 12 13 17 14  9  9  9
 12 11 12 11 10  9 14 11 12 12 14 12 13 12 13 12 11 13 11  9 10  8 15 15
 14 19 14 13  9 11 11 13 14 16 13 11 15 13 12 11 14 12 15 16 13 13 13 12
 12 15 10 13 14 14 14 11 15 15 12 11 13 17 15 12 15 12 12 13 11 13 14 14
 15 10 10 19 13 10 16 15 14 13 14 13 15 11 12 14 16 16 15 15  8 12  9 14
 13 10 14 13  7 16  8  9 11 13 11 13  7  9 13 20  8 11 12 14  9 12 10 11
 16 12 16 13 12 16 12 11 11 13  9 13 16 12 14 14 13 12 14 12 13  9 11 15
 18 12  9 13 12 12 11 12]
200
```



4.4

The two histograms are quite similar. The coin-flip scenario can be generated by sampling the binomial distribution.

4.5

In [21]:
```
print('mean and variance of binomial_trials1:')
mean_trials1 = np.mean(binomial_trials1)
var_trials1 = np.var(binomial_trials1)
print(mean_trials1, var_trials1)
print('true mean and variance:')
mean, var = stats.binom.stats(25, 0.5, moments='mv')
print(mean, var)
```

```
mean and variance of binomial_trials1:
12.755 6.114975000000002
true mean and variance:
12.5 6.25
```

The mean of trials1 is similar to the true mean. The variance of trials1 is larger than the true variance.

Testing Your Python Code

In the following section we're going to do a brief introduction to unit testing. We do so not only because unit testing has become an increasingly important part of of the methodology of good software practices, but also because we plan on using unit tests as part of our own CS109 grading practices as a way of increasing rigor and repeatability decreasing complexity and manual workload in our evaluations of your code. We'll provide an example unit test at the end of this section.

Introduction to unit testing

In [22]:
```
import ipytest
```

Unit testing is one of the most important software testing methodologies. Wikipedia describes unit testing as "a software testing method by which individual units of source code, sets of one or more computer program modules together with associated control data, usage procedures, and operating procedures, are tested to determine whether they are fit for use."

There are many different python libraries that support software testing in general and unit testing in particular. PyTest is one of the most widely used and well-liked libraries for this purpose. We've chosen to adopt PyTest (and ipytest which allows pytest to be used in ipython notebooks) for our testing needs and we'll do a very brief introduction to Pytest here so that you can become familiar with it too.

If you recall the function that we provided you above `throw_a_coin`, which we'll reproduce here for convenience, it took a number and returned that many "coin tosses". We'll start by seeing what happens when we give it different sizes of N. If we give $N = 0$, we should get an empty array of "experiments".

```
In [23]: def throw_a_coin(N):
             return np.random.choice(['H','T'], size=N)
```

```
In [24]: throw_a_coin(0)
```

```
Out[24]: array([], dtype='<U1')
```

Great! If we give it positive values of N we should get that number of 'H's and 'T's.

```
In [25]: throw_a_coin(5)
```

```
Out[25]: array(['H', 'T', 'T', 'H', 'H'], dtype='<U1')
```

```
In [26]: throw_a_coin(8)
```

```
Out[26]: array(['H', 'T', 'H', 'H', 'H', 'T', 'T', 'T'], dtype='<U1')
```

Exactly what we expected!

What happens if the input isn't a positive integer though?

```
In [27]: throw_a_coin(4.5)
```

```
---------------------------------------------------------------------------
--
TypeError                                 Traceback (most recent call las
t)
<ipython-input-27-7a98054470df> in <module>()
----> 1 throw_a_coin(4.5)

<ipython-input-23-9b62022d816e> in throw_a_coin(N)
      1 def throw_a_coin(N):
----> 2     return np.random.choice(['H','T'], size=N)

mtrand.pyx in mtrand.RandomState.choice()

mtrand.pyx in mtrand.RandomState.randint()

mtrand.pyx in mtrand.RandomState.randint()

randint_helpers.pxi in mtrand._rand_int64()

TypeError: 'float' object cannot be interpreted as an integer
```

or

```
In [31]: throw_a_coin(-4)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-31-8560c28a4e91> in <module>()
----> 1 throw_a_coin(-4)

<ipython-input-23-9b62022d816e> in throw_a_coin(N)
      1 def throw_a_coin(N):
----> 2     return np.random.choice(['H','T'], size=N)

mtrand.pyx in mtrand.RandomState.choice()

mtrand.pyx in mtrand.RandomState.randint()

mtrand.pyx in mtrand.RandomState.randint()

randint_helpers.pxi in mtrand._rand_int64()

ValueError: negative dimensions are not allowed
```

It looks like for both real numbers and negative numbers, we get two kinds of errors a `TypeError` and a `ValueError` . We just engaged in one of the most rudimentary forms of testing, trial and error. We can use pytest to automate this process by writing some functions that will automatically (and potentially repeatedly) test individual units of our code methodology. These are called ***unit tests***.

Before we write our tests, let's consider what we would think of as the approrpriate behavior for `throw_a_coin` under the conditions we considered above. If `throw_a_coin` receives positive integer input, we want it to behave exactly as it currently does -- returning an output consisting of a list of characters 'H' or 'T' with the length of the list equal to the positive integer input. For a positive floating point input, we want `throw_a_coin_properly` to treat the input as if it were rounded down to the nearest integer thus returning a list of 'H' or 'T' integers whose length is the same as the input rounded down to the next highest integer. For a any negative number input or an input of 0, we want `throw_a_coin_properly` to return an empty list.

We create pytest tests by writing functions that start or end with "test". We'll use the **convention** that our tests will start with "test".

We begin the code cell with ipytest's clean_tests function as a way to clear out the results of previous tests starting with "test_throw_a_coin" (the * is the standard wild card charater here).

In [32]:
```python
## the * after test_throw_a_coin tells this code cell to clean out the resul
## of all tests starting with test_throw_a_coin
ipytest.clean_tests("test_throw_a_coin*")

## run throw_a_coin with a variety of positive integer inputs (all numbers l
## verify that the length of the output list (e.g ['H', 'H', 'T', 'H', 'T'],
def test_throw_a_coin_length_positive():
    for n in range(1,20):
        assert len(throw_a_coin(n)) == n

## verify that throw_a_coin produces an empty list (i.e. a list of length 0)
## of 0
def test_throw_a_coin_length_zero():
    ## should be the empty array
    assert len(throw_a_coin(0)) == 0


## verify that given a positive floating point input (i.e. 4.34344298547201,
## coin flips of length equal to highest integer less than the input
def test_throw_a_coin_float():
    for n in np.random.exponential(7, size=5):
        assert len(throw_a_coin(n)) == np.floor(n)


## verify that given any negative input (e.g. -323.4), throw_a_coin produces
def test_throw_a_coin_negative():
    for n in range(-7, 0):
        assert len(throw_a_coin(n)) == 0


ipytest.run_tests()
```

```
unittest.case.FunctionTestCase (test_throw_a_coin_float) ... ERROR
unittest.case.FunctionTestCase (test_throw_a_coin_length_positive) ... ok
unittest.case.FunctionTestCase (test_throw_a_coin_length_zero) ... ok
unittest.case.FunctionTestCase (test_throw_a_coin_negative) ... ERROR

======================================================================
ERROR: unittest.case.FunctionTestCase (test_throw_a_coin_float)
----------------------------------------------------------------------
Traceback (most recent call last):
  File "<ipython-input-32-78a86d656b91>", line 22, in test_throw_a_coin_f
loat
    assert len(throw_a_coin(n)) == np.floor(n)
  File "<ipython-input-23-9b62022d816e>", line 2, in throw_a_coin
    return np.random.choice(['H','T'], size=N)
  File "mtrand.pyx", line 1163, in mtrand.RandomState.choice
  File "mtrand.pyx", line 995, in mtrand.RandomState.randint
  File "mtrand.pyx", line 996, in mtrand.RandomState.randint
  File "randint_helpers.pxi", line 253, in mtrand._rand_int64
TypeError: 'numpy.float64' object cannot be interpreted as an integer

======================================================================
ERROR: unittest.case.FunctionTestCase (test_throw_a_coin_negative)
----------------------------------------------------------------------
Traceback (most recent call last):
  File "<ipython-input-32-78a86d656b91>", line 28, in test_throw_a_coin_n
```

```
egative
    assert len(throw_a_coin(n)) == 0
  File "<ipython-input-23-9b62022d816e>", line 2, in throw_a_coin
    return np.random.choice(['H','T'], size=N)
  File "mtrand.pyx", line 1163, in mtrand.RandomState.choice
  File "mtrand.pyx", line 995, in mtrand.RandomState.randint
  File "mtrand.pyx", line 996, in mtrand.RandomState.randint
  File "randint_helpers.pxi", line 253, in mtrand._rand_int64
ValueError: negative dimensions are not allowed

----------------------------------------------------------------------
Ran 4 tests in 0.006s

FAILED (errors=2)
```

As you see, we were able to use pytest (and ipytest which allows us to run pytest tests in our ipython notebooks) to automate the tests that we constructed manually before and get the same errors and successes. Now time to fix our code and write our own test!

Question 5: You Better Test Yourself before You Wreck Yourself!

Now it's time to fix `throw_a_coin` so that it passes the tests we've written above as well as add our own test to the mix!

5.1. Write a new function called `throw_a_coin_properly` that will pass the tests that we saw above. For your convenience we'll provide a new jupyter notebook cell with the tests rewritten for the new function. All the tests should pass. For a positive floating point input, we want `throw_a_coin_properly` to treat the input as if it were rounded down to the nearest integer. For a any negative number input, we want `throw_a_coin_properly` to treat the input as if it were 0.

5.2. Write a new test for `throw_a_coin_properly` that verifies that all the elements of the resultant arrays are 'H' or 'T'.

Answers

5.1

In [33]:
```
def throw_a_coin_properly(n_trials):
    if n_trials < 0:
        return np.random.choice(['H','T'], size=0)
    elif isinstance(n_trials, int):
        return np.random.choice(['H','T'], size=n_trials)
    else:
        return np.random.choice(['H','T'], size=int(np.floor(n_trials)))
```

In [34]:
```python
ipytest.clean_tests("test_throw_a_coin*")

def test_throw_a_coin_properly_length_positive():
    for n in range(1,20):
        assert len(throw_a_coin_properly(n)) == n


def test_throw_a_coin_properly_length_zero():
    ## should be the empty array
    assert len(throw_a_coin_properly(0)) == 0


def test_throw_a_coin_properly_float():

    for n in np.random.exponential(7, size=5):
        assert len(throw_a_coin_properly(n)) == np.floor(n)


def test_throw_a_coin_properly_negative():

    for n in range(-7, 0):
        assert len(throw_a_coin_properly(n)) == 0


ipytest.run_tests()
```

```
unittest.case.FunctionTestCase (test_throw_a_coin_properly_float) ... ok
unittest.case.FunctionTestCase (test_throw_a_coin_properly_length_positiv
e) ... ok
unittest.case.FunctionTestCase (test_throw_a_coin_properly_length_zero)
... ok
unittest.case.FunctionTestCase (test_throw_a_coin_properly_negative) ...
ok

----------------------------------------------------------------------
Ran 4 tests in 0.004s

OK
```

5.2

In [35]:

```
ipytest.clean_tests("test_throw_a_coin*")

## write a test that verifies you don't have any other elements except H's a
def test_throw_a_coin_properly_verify_H_T():
    assert all(y in ('T', 'H') for y in throw_a_coin_properly(100))

ipytest.run_tests()
```

```
unittest.case.FunctionTestCase (test_throw_a_coin_properly_verify_H_T)
... ok

----------------------------------------------------------------------
Ran 1 test in 0.001s

OK
```