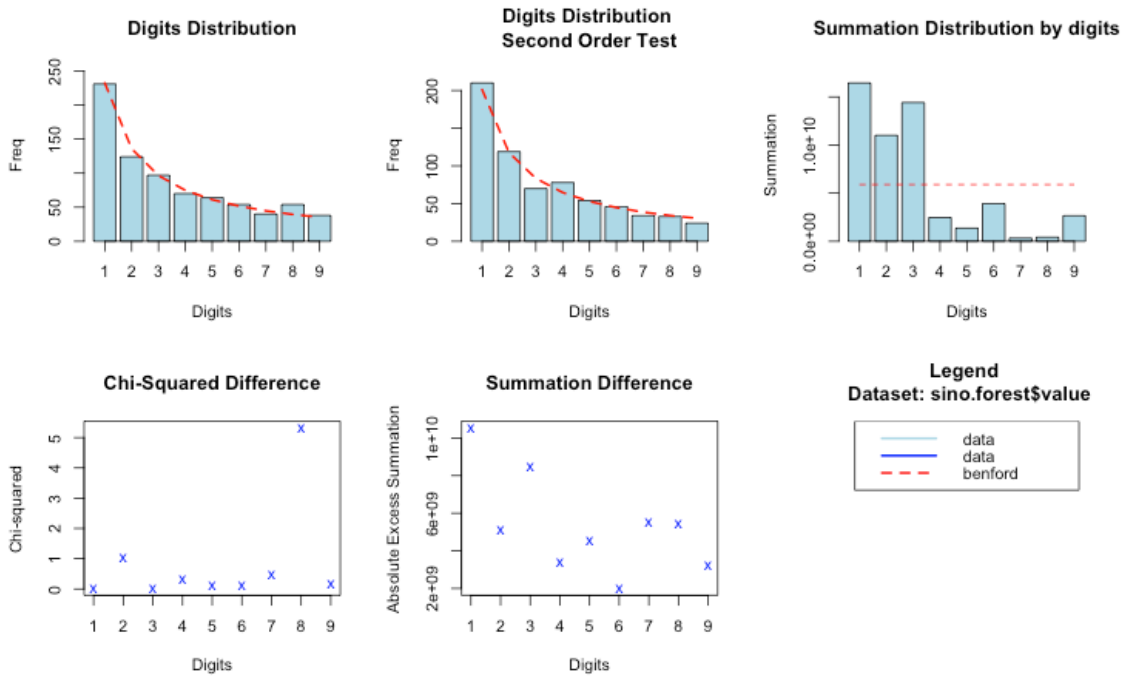
Activity #8: Chi-squared tests for goodness-of-fit and independence

Read the article on Benford's Law: http://www.bradthiessen.com/html5/stats/m301/11a.pdf
Check out some examples at: http://testingbenfordslaw.com/

Scenario: Benford's Law (aka the first digit law) refers to the frequency distribution of digits in many (but not all) real-life sources of data.

The following table displays the frequency of first digits from 2010 financial statements from Sino Forest* (a leading Chinese commercial forest plantation operators).

First digit	1	2	3	4	5	6	7	8	9	
Expected relative frequency under Benford's Law	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	
Observed relative frequency	0.299	0.161	0.126	0.091	0.083	0.070	0.052	0.070	0.049	
Expected frequency	232.37	135.87	96.5	74.88	60.99	51.72	44.78	39.37	35.51	N=772
Observed frequency	231	124	97	70	64	54	40	54	38	N=772



*In June of 2011, the company's stock price plummeted as a research report made allegations that the company had been fraudulently inflating its assets and earnings.

Source: Nigrini, M. J. (2012). Benford's Law: Application for Forensic Accounting, Auditing and Fraud Detection. Wiley and Sons: New Jersey.
Sino Forest info: http://www.theglobeandmail.com/globe-investor/the-empire-sino-forest-built-and-the-farmers-who-paid-the-price/article4183298/?page=all

1. Verify the expected frequency and relative frequency values. Based on the table and subsequent plots, do we have evidence that Sino Forest's financial statements do not follow Benford's Law?

2. We don't expect a sample of first digits (from any source) to **perfectly** follow Benford's Law. But how unlikely were we to get the first digits we observed from Sino Forest if, in fact, the numbers represent a sample of first digits that do come from Benford's Law?

   To derive a method for estimating this p-value, let's first turn to a simple example: Rolling one die. The following table displays the results from rolling one (simulated) die 120 times:

Die roll	1	2	3	4	5	6	Total
Observed frequency	23	15	19	24	21	18	120
Expected frequency							

   If we have a fair die, we assume the results are all equally likely. Fill-in the expected frequency values above.

3. We need to derive a method for determining how "far off" our expectations are from our expectations. Ideally, it would be single value we could calculate from our data.

4. Calculate this value for the following two tables (looking only at equally-likely outcomes 1-3). Are the observed values in each table the same "distance" from their expectations? Should they be?

	Table A				Table B			
	1	2	3	Total	1	2	3	Total
Observed	2	4	6	12	200	400	600	1200
Expected	4	4	4	12	400	400	400	1200

To test whether an observed set of frequencies follows a specific (expected) distribution, we can use the:

Chi-squared Goodness-of-Fit Test: $\chi^2_{r-1} = \sum \frac{\left(\text{observed}_i - \text{expected}_i\right)^2}{\text{expected}_i}$

Note: For the chi-square approximation to be valid, the expected frequency should be at least 5. If some of the counts are less than five, you may need to combine some bins in the tails

Online calculator: http://lock5stat.com/statkey/advanced_goodness/advanced_goodness.html

5. Conduct this chi-squared goodness of fit test for our simulated die. What can we conclude?

Die roll	1	2	3	4	5	6	Total
Observed frequency	23	15	19	24	21	18	120
Expected frequency	20	20	20	20	20	20	120

6. Below, I've pasted output from a chi-square goodness-of-fit test on the Sino Forest data. What can we conclude?

First digit	1	2	3	4	5	6	7	8	9
Expected relative frequency	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046
Expected frequency	232.37	135.87	96.5	74.88	60.99	51.72	44.78	39.37	35.51
Observed frequency	231	124	97	70	64	54	40	54	38

```
Stats:  Pearson's Chi-squared test
X-squared = 7.6517, df = 8, p-value = 0.4682
```

To conduct this test on a computer, you can:

- Copy the data at: http://www.bradthiessen.com/html5/data/sinoforestdigits.csv
- Open http://lock5stat.com/statkey/advanced_goodness/advanced_goodness.html and click EDIT DATA
- Paste data & modify the NULL HYPOTHESIS to match the expected relative frequencies from Benford's Law
- The chi-squared test statistic will be displayed on the top-right
- To get a p-value, use: http://lock5stat.com/statkey/theoretical_distribution/theoretical_distribution.html#chi

7. Since we're already using the applet, let's look at a randomization-based method for estimating the p-value. Generate 10,000 samples and explain what is happening. Then, estimate the p-value.

8. This chi-squared goodness-of-fit test is useful whenever we want to test if a set of data came from a particular distribution. For example, when we conduct an ANOVA, we make a normality assumption. We could use a chi-square test to determine if that assumption is reasonable. (Note that we could also use other methods, such as Q-Q plots and other normality tests.)

   Suppose the following numbers represent the time between breakdowns of a machine. Is the *exponential distribution* an appropriate model for this data?

0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.4	0.5	0.6	0.6
0.6	0.6	0.7	0.7	0.8	0.9	0.9	1.0	1.1	1.2	1.3
1.3	1.6	1.7	1.8	1.9	1.9	2.0	2.0	2.1	2.2	2.2
2.4	2.4	2.4	2.7	2.8	2.8	2.8	3.0	3.1	3.5	3.7
3.7	3.7	4.1	4.1	4.2	4.2	4.3	4.3	4.5	4.9	4.9
4.9	5.0	5.2	5.3	5.3	5.7	5.7	5.9	6.1	6.2	6.2
6.2	6.3	7.4	7.5	7.5	7.7	8.1	8.6	9.2	9.5	10.0
10.3	10.6	10.9	12.3	13.9	13.7	13.9	14.8	14.9	17.3	17.6

   Recall (if you ever learned) the following properties of an exponential function:

   - It can be useful when modeling waiting times
   - The cumulative distribution function (cdf) is: $P(X \le x) = 1 - e^{-\lambda x}$, where $\lambda = \dfrac{1}{E[X]}$

   Step #1: Calculate the parameters of the (assumed) underlying distribution.
   In this case, we need to calculate lambda. To get lambda, we need the expected value (long-run average).
   The best estimate of the population average may be the sample average.
   The sample average of our sample data is 4.69. Therefore, lambda = 1 / 4.69 = 0.213.

   Step #2: The chi-squared test requires categorical (binned) data, so we must categorize our data.

Category:	0-3	3-6	6-9	9-12	12-15	15-18	Total
# of observations:	40	23	11	6	6	2	88

   Notice that last bin only has 2 observations. Let's go ahead and merge that with the preceding bin.

Category:	0-3	3-6	6-9	9-12	12+	Total
# of observations:	40	23	11	6	8	88

Step #3: We can now use the exponential distribution to calculate the expected frequencies.

Category	Probability	Expected frequency
0-3	$P(0 < x < 3) = \left(1 - e^{-.213(3)}\right) - \left(1 - e^{-.213(0)}\right) = 0.472$	0.472 x 88 = 41.55
3-6	$P(3 < x < 6) = \left(1 - e^{-.213(6)}\right) - \left(1 - e^{-.213(3)}\right) = 0.249$	0.249 x 88 = 21.93
6-9	$P(6 < x < 9) = \left(1 - e^{-.213(9)}\right) - \left(1 - e^{-.213(6)}\right) = 0.1315$	0.1315x88 = 11.58
9-12	$P(9 < x < 12) = \left(1 - e^{-.213(12)}\right) - \left(1 - e^{-.213(9)}\right) = 0.069$	0.069 x 88 = 6.11
12+	$P(x > 12) = 1 - \left(1 - e^{-.213(12)}\right) = 0.078$	0.078 x 88 = 6.83

Step #4: We can now calculate the chi-squared statistic.

Category:	0-3	3-6	6-9	9-12	12+	Total
# of observations:	40	23	11	6	8	88
expected:	41.55	21.93	11.58	6.11	6.83	88
(O-E)^2 / E:	**0.058**	**0.052**	**0.029**	**0.002**	**0.200**	**Sum = 0.341**

Step #4: Compare it to a chi-squared distribution with _____ degrees of freedom. Estimated p-value = _____.
Distribution calculator: http://lock5stat.com/statkey/theoretical_distribution/theoretical_distribution.html#chi

9. Thus far, we've used the chi-square distribution to test the distribution of a single variable. We can also use the chi-square statistic to test whether two categorical variables are independent. This is called the *chi-squared test for independence*.

Explain in plain language what it means if two variables are independent. Then, write the definition of independence using probability notation.

Independence means: _____

If A and B are independent, P(A | B) = _____ and P(A and B) = _____

10. As a simple example, suppose we're interested in the relationship between two variables: gender and coin flips. We get 50 men and 50 women to each flip a coin and record the result. Based on an assumption of independence, fill-in the expected values for each cell in the table to the right:

Expected	Male	Female	Total
Heads			
Tails			
Total			

11. Suppose we conduct this experiment and get the results in the table to the right. Compute a chi-square statistic and estimate the p-value. What conclusions could you make?

Observed	Male	Female	Total
Heads	30	10	40
Tails	20	40	60
Total	50	50	100

Scenario: Is acupuncture effective in reducing pain? A 2007 (Haake et al.) involved 1,162 patients with chronic lower back pain. The subjects were randomly assigned to one of three groups:

- Verum acupuncture practiced according to traditional Chinese principles
- Sham acupuncture where needles were inserted into the skin but not at acupuncture points
- Traditional (non-acupuncture) therapy consisting of drugs, physical therapy, and exercise

The following table summarizes the results from this study:

	Verum acupuncture	Sham acupuncture	Traditional	Total
Substantial reduction in pain	184	171	106	461
Not a substantial reduction in pain	203	216	282	701
Total	387	387	388	1162

Source: Tintle, et al.
Applet: http://www.rossmanchance.com/applets/ChiSqShuffle.html?hideExtras=2

12. Calculate the probability that an individual undergoing verum acupuncture experienced a substantial reduction in pain. Then, calculate this same probability for the subjects in the traditional group. Finally, calculate and interpret the ratio of the two values you just calculated.

P(reduction in pain | verum acupuncture) = _____

P(reduction in pain | traditional) = _____

Ratio = _____ = _____

13. Under a null hypothesis that treatment method and reduction in pain are independent, what would be the expected frequencies of each cell?

	Verum acupuncture	Sham acupuncture	Traditional	Total
Substantial reduction in pain				461
Not a substantial reduction in pain				701
Total	387	387	388	1162

14. To speed things up, let's use an applet to conduct the chi-square test for independence.
    - Go to: http://www.rossmanchance.com/applets/ChiSqShuffle.html?hideExtras=2
    - Enter the following data and click USE TABLE

```
       real sham trad
better 184  171  106
not    203  216  282
```

    - Check the boxes for SHOW TABLE and SHOE CHI-SQUARE OUTPUT
    - Select CHI-SQUARED from the STATISTIC MENU

Record the test statistic and p-value here. What can you conclude?

15. The website also allows you to conduct a randomization-based test for independence. Click SHOW SHUFFLE OPTIONS and generate 10,000 shuffles. Explain the process and record the estimated p-value.

Scenario: Researchers selected a sample of 661 heart disease patients and a control group of 771 males not suffering from heart disease. Each subject was classified into 5 baldness categories:

Drawing by Kelly Martelle Comparative risk of heart disease increases from zero with no hair loss to 36 percent for severe crown baldness.

The data from this study are as follows:

	Baldness level					
	None	Little	Some	Much	Extreme	Total
Heart disease	251	165	195	50	2	663
No heart disease	331	221	185	34	1	772
Total	582	386	380	84	3	1435

Source: Tintle, et al.
Applet: http://lock5stat.com/statkey/advanced_association/advanced_association.html

16. Let's use a randomization-based chi-square test.
    - Go to: http://lock5stat.com/statkey/advanced_association/advanced_association.html
    - Click EDIT DATA and paste the following data

```
[blank], none, little, some, much, extreme
disease, 251, 165, 195, 50, 2
not, 331, 221, 185, 34, 1
```

    - Generate 10,000 samples and estimate the p-value.

Record the test statistic and p-value here. What can you conclude?

17. P-values can be useful, but we're typically more interested in estimate the effect size. One measure of the strength of the relationship between two categorical variables is the *phi-coefficient*.

		Variable #1		
Variable #2		a	b	r1
		c	d	r2
		c1	c2	TOTAL

A phi coefficient measures the association between two binary (dichotomous) variables. For the general 2x2 table displayed to the right, the phi coefficient is calculated as: $\phi = \frac{ad - bc}{\sqrt{r_1 r_2 c_1 c_2}}$

Let's take another look at our acupuncture and baldness datasets. This time, let's collapse them into 2x2 tables:

	Verum acupuncture	Traditional or sham acupuncture	Total
Substantial reduction in pain	184	277	461
Not a substantial reduction in pain	203	498	701
Total	387	775	1162

	No baldness	At least some baldness	Total
Heart disease	251	412	663
No heart disease	331	441	772
Total	582	853	1435

Calculate and interpret the phi coefficient for each dataset.

18. Finally, calculate and interpret an odds ratio for each dataset.

Project ideas: G-test, Fisher's exact test, nonparametric analysis of contingency tables