# Data 8 Foundations of Data Science
# Spring 2016

FINAL SOLUTIONS

**INSTRUCTIONS**

- You have 3 hours to complete the exam.
- The exam is closed book, closed notes, closed computer, closed calculator, except one hand-written 8.5" × 11" crib sheet of your own creation and the two official study guides provided with the exam.
- Mark your answers **on the exam itself**. We will *not* grade answers written on scratch paper.

| | |
|---|---|
| Last name | |
| First name | |
| Student ID number | |
| BearFacts email (`_@berkeley.edu`) | |
| GSI | |
| Name of the person to your left | |
| Name of the person to your right | |
| *All the work on this exam is my own.* **(please sign)** | |

**1. (16 points) Tables**

The `cafe` table (left) describes the Yelp reviews for three cafes on Euclid. Every cafe has a count for the number of 3-star, 4-star, and 5-star reviews, in that order. The `price` table (right) describes coffee prices.

```
name    | stars | count
Nefeli  | 3     | 37
Nefeli  | 4     | 75
Nefeli  | 5     | 50
Brewed  | 3     | 56
Brewed  | 4     | 71
... (4 rows omitted)
```

```
name    | $
Nefeli  | 3
Brewed  | 3
Abe     | 2
```

Complete the **Python expressions** below to compute each result. For example, if the result prompt said, "The total number of reviews of all cafes," then you would write: `sum ( cafe.column( 2 ) )`
***** You must fit your solution into the lines and spaces provided to receive full credit. *****
The last line of each answer should evaluate to the result requested; you never need to call `print`.

**(a) (2 pt)** The total number of reviews of the cafe named `Nefeli`.

```
sum(cafe.where('name', 'Nefeli').column('count'))
```

**(b) (2 pt)** The total number of reviews of the cafe with the fewest reviews.

```
min(cafe.group('name', sum).column(2))
```

**(c) (2 pt)** The average number of stars for reviews of the cafe named `Nefeli`.

```
n = cafe.where('name', 'Nefeli')
sum(n.column('stars') * n.column('count')) / sum(n.column('count'))
```

**(d) (3 pt)** The total variation distance between the distributions of stars for `Nefeli` and `Brewed`.

```
a = cafe.where('name', 'Nefeli').column('count')
b = cafe.where('name', 'Brewed').column('count')
0.5 * sum(abs(a/sum(a) - b/sum(b)))
```

**(e) (2 pt)** An array containing the names of all cafes that have above-average coffee prices.

```
price.where(price.column('$') > np.average(price.column('$'))).column('name')
```

**(f) (3 pt)** Among all reviews of cafes with $3 coffee, the proportion that are 3-star reviews.

```
j = cafe.join('name', price).where('$', 3)
sum(j.where('stars', 3).column('count')) / sum(j.column('count'))
```

**(g) (2 pt)** The table below, in which each row describes the number of reviews with a particular star rating for every cafe.

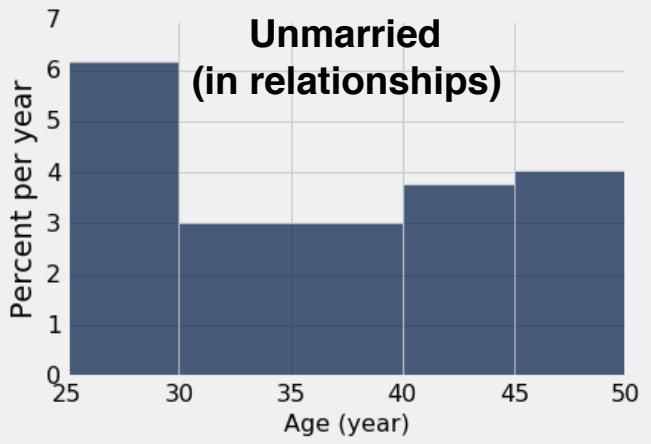
```
stars | Abe  | Brewed | Nefeli
3     | 1    | 56     | 37
4     | 2    | 71     | 75
5     | 17   | 37     | 50
```

```
cafe.pivot('name', 'stars', 'count', sum))
```

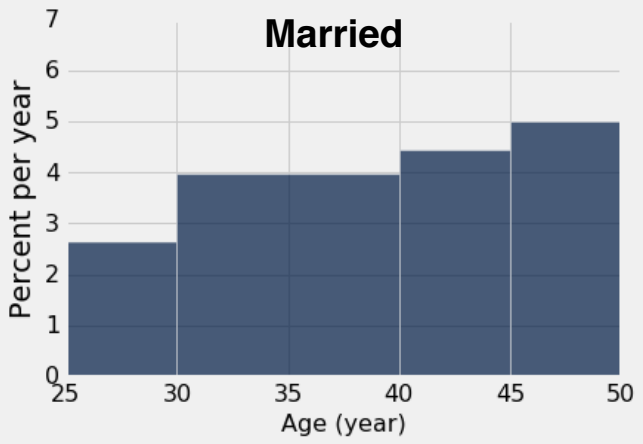
**2. (17 points) Distributions**

**500 women** age 25 to 49 in steady relationships were surveyed. Each woman was asked her age in years and whether she was married to her partner. There were **400 unmarried** and **100 married** women among those surveyed. The histograms below visualize the ages of these two groups of women.

```
unmarried.hist('Age',bins=[25,30,40,45,50])
```




```
married.hist('Age',bins=[25,30,40,45,50])
```




**(a) (10 pt)** For each pair of quantities, compare them using the information above and choose one of following:
(A): (I) is larger.
(B): (II) is larger.
(C): (I) and (II) are about the same.
(D): There is not enough information to compare (I) and (II).
*** **You must briefly justify your answer to receive full credit.** ***

- (I) The **number** of unmarried women age 25-29 vs (II) The **number** of unmarried women age 30-39

  (C): Both are about 30% of 400; also accept (A) with justification 31% ¿ 30%

- (I) Among the unmarried women, the **proportion** who are of age 25-29 vs
  (II) Among the married women, the **proportion** who are of age 45-49

  (A): 30% vs 25%

- (I) The **number** of unmarried women age 30-39 vs (II) The **number** of married women

  (A): 120 vs 100

- (I) The **proportion** of married women age 30-34 vs (II) The **proportion** of married women age 35-39

  (D): You can't tell how ages are distributed within a bin.

- (I) The 20th percentile age of unmarried women vs (II) The 20th percentile age of married women

  (B): Somewhere within 25-29 vs somewhere within 30-39.

**(b) (3 pt)** What proportion of everyone surveyed were in the age range 30-39?

(0.3 * 400 + 0.4 * 100) / 500 = 8/25 or 32%

**(c) (4 pt)** If you select a woman uniformly at random from those surveyed and find out that her age is in the range 30-39, what is the chance that she is married?

(2/5) * (1/5) / (8/25) = 1/4 or 25%

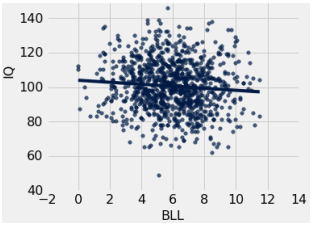
**3. (25 points) Regression**

The `lead` table (left) contains one row per child in a study of 1000 children's Blood Lead Levels (BLL) measured in micrograms per deciliter and their intelligence quotients (IQ). Assume that the data were collected by sampling children at random from a very large population. Summary statistics (middle) and a scatter diagram (right) are shown below. All BLLs are measured to one decimal place, and all IQ scores are integers.

| BLL | IQ |
|---|---|
| 7.9 | 90 |
| 6.2 | 78 |
| 3.2 | 110 |
| 4.1 | 128 |
| 7.3 | 88 |

(995 rows omitted)

| Expression | Value |
|---|---|
| `np.average(lead.column('BLL'))` | 6 |
| `np.std(lead.column('BLL'))` | 2 |
| `np.average(lead.column('IQ'))` | 100 |
| `np.std(lead.column('IQ'))` | 15 |
| `correlation(lead, 'BLL', 'IQ')` | -0.1 |




**(a) (2 pt)** What is the value of `correlation(lead, 'IQ', 'BLL')`?
*Hint*: The `correlation` function appears on your midterm study guide.

-0.1

**(b) (3 pt)** What is the estimated average IQ of a child with a BLL that is 2 standard deviations above the mean BLL? Use the regression line to find this estimate, assuming BLL and IQ are linearly related.

$2 * -0.1 * 15 + 100 = 97$

**(c) (4 pt)** Write the equation of the regression line through this sample for the IQ $y$ in terms of the BLL $x$.

$y = -0.1 \cdot \frac{15}{2} \cdot x + 100 - 6 \cdot (-0.1 \cdot \frac{15}{2}) = -0.75x + 104.5$

**(d) (4 pt)** Complete the code below so that the last line prints out a 95% confidence interval for the IQ value at a BLL of 10.0 on the *regression line of the population* from which this sample was collected.
*Hint*: The `slope` and `intercept` functions appears on your midterm study guide.

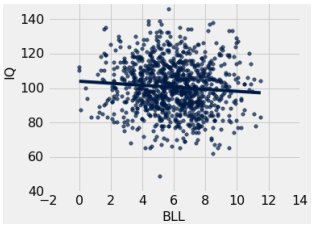
```
estimate_table = Table(['Resample #', 'Estimate'])

for i in np.arange(400):

    r = lead.sample(with_replacement=True)


    e = slope(r, 0, 1) * 10 + intercept(r, 0, 1)

    estimate_table.append([i, e])


estimates = estimate_table.column(1)

print(percentile(2.5, estimates), percentile(97.5, estimates))
```

*The data from the previous page are repeated here for your reference.*

| BLL | IQ |
|---|---|
| 7.9 | 90 |
| 6.2 | 78 |
| 3.2 | 110 |
| 4.1 | 128 |
| 7.3 | 88 |

(995 rows omitted)

| Expression | Value |
|---|---|
| `np.average(lead.column('BLL'))` | 6 |
| `np.std(lead.column('BLL'))` | 2 |
| `np.average(lead.column('IQ'))` | 100 |
| `np.std(lead.column('IQ'))` | 15 |
| `correlation(lead, 'BLL', 'IQ')` | -0.1 |




**(e)** **(6 pt)** Mark each of the following statements about **the confidence interval you computed in part (d)** as *True* or *False* based on the definition of a confidence interval and the details of your implementation. The terms "most" should be interpreted as around 95%.

- Circle *True* or *False*: It contains most IQ scores in the population for children with a BLL of 10.0. False.
- Circle *True* or *False*: It contains most IQ scores in the sample for children with a BLL of 10.0. False.
- Circle *True* or *False*: If `with_replacement=False` were used (line 3), the interval would have 0 width. True.
- Circle *True* or *False*: If the study were repeated many times, most confidence intervals computed in this way would overlap. True.
- Circle *True* or *False*: If the study were repeated many times, the average IQ in the population would fall within most intervals computed in this way. False.
- Circle *True* or *False*: If the study were repeated many times, the average IQ in the population for children with a BLL of 10.0 would fall within most intervals computed in this way. True.

**(f)** **(2 pt)** What null hypothesis would you evaluate in a statistical test to determine whether BLL and IQ are negatively correlated in the population?

The correlation coefficient (or slope) is greater than or equal to 0.

**(g)** **(2 pt)** Based on the summary statistics provided, what is the minimum proportion of IQ scores in the sample that are between 70 and 130 according to Chebyshev's inequality?

$1 - (\frac{1}{2})^2 = 75\%$ because 130 - 100 = 100 - 70 = 30 IQ, which is 2 standard deviations.

**(h)** **(2 pt)** Based on the scatter diagram and summary statistics, what proportion of IQ scores in the sample do you think are between 70 and 130? **Describe your reasoning.** *Hint*: You don't need to count dots.

95% because 130 - 100 = 100 - 70 = 30 IQ, which is 2 standard deviations, and both IQ and BLL appear to be normally distributed in the scatter diagram.

**4. (12 points) Experiments**

To test whether lead exposure affects cognitive function, a researcher conducts the following experiment. 90 goldfish are placed in a tank and left to swim around for 24 hours. Then, a glass plate is inserted, separating the tank into two equal sides that cannot mix. Each fish ends up on one side. The researcher adds a high level of lead to the left side only. After 1 week, the IQ of each fish is measured. The `fish` table (left), counts of fish on each side (middle), and the average IQs on each side (right) describe the results.

| Side | Fish IQ |
|---|---|
| Left | 98 |
| Right | 118 |
| Left | 105 |

(87 rows omitted)

`fish.group('Side')`

| Side | count |
|---|---|
| Left | 44 |
| Right | 46 |

`fish.group('Side', np.average)`

| Side | Fish IQ average |
|---|---|
| Left | 94.6 |
| Right | 103.9 |

**(a) (4 pt)** Complete the null hypothesis in this experiment about lead exposure affecting fish IQ.

Among the fish in the experiment,
the distribution of hypothetical IQ scores for all the fish, had none been exposed to lead

is the same as hypothetical IQ scores for all the fish, had they all been exposed to lead.

**(b) (2 pt)** You choose as your test statistic the absolute difference in average fish IQ scores for the two groups. Complete `absolute_difference_of_averages`. The argument `t` is a two-column table such as `fish` in which column 0 indicates to which group each example belongs, and column 1 contains quantitative data.

```
def absolute_difference_of_averages(t):

    averages = t.group(0, np.average).column(1)

    return abs(np.diff(averages).item(0))
```

**(c) (4 pt)** Complete the `permutation_test` function, which returns an approximate P-value from a permutation test. It takes a two-column table `t` and a test statistic function `f`. An example call would be `permutation_test(fish, absolute_difference_of_averages)`. The body of your function **should not** refer to `fish` or `absolute_difference_of_averages`; only `t` and `f`.

```
def permutation_test(t, f):

    stats = Table(["Test statistic for a shuffled table"])

    for i in np.arange(1000):

        r = Table().with_columns(["X", t.column(0), "Y", t.sample().column(1)])


        stats.append([f(r)])


    more_extreme = stats.column(0) >= f(t)


    return np.count_nonzero(more_extreme) / 1000
```

**(d) (2 pt)** Circle *True* or *False*: Assuming that fish are randomly distributed among the two sides, a P-value of 0.003 indicates that lead exposure causes a change in IQ. **Briefly justify your answer.**

Yes. This is a randomized controlled experiment.

**5. (12 points) Sampling**

**(a) (6 pt)** For each statement below, choose one:
(A): Almost always true for any uniform random sample with replacement.
(B): Almost always true for **large** uniform random samples with replacement (e.g., $n$=1000), but not reliably for small ones (e.g., $n$=5).
(C): Never true.
(D): None of the above.

- For categorical data with three categories, the total variation distance between the population distribution and an observed sample distribution will be small.

  (B): For a large sample, the empirical distribution almost always looks like the population distribution.

- For quantitative data, the sample average is an unbiased estimate of the population average.

  (A): Sample averages are unbiased estimators of population averages.

- For quantitative data, a histogram of the values in a sample will have a bell shape.

  (D):The sampling distribution of a statistic such as the average is normal, but the values in the sample are distributed like the population, which can have any shape.

**(b) (6 pt)** You have a table named `t` with at least 20 rows. All rows are different. Write a Python expression that computes each of the following probabilities. Use `t.num_rows` for the number of rows in `t`.

- Row 10 of `t` appears among the rows in `t.sample(20, with_replacement=False)`.

  ```
  20 / t.num_rows
  ```

- Row 10 of `t` **does not appear** among the rows in `t.sample(20, with_replacement=True)`.

  ```
  ((t.num_rows - 1)/t.num_rows) ** 20
  ```

- All rows of `t.sample(20, with_replacement=True)` are different (no repeats).

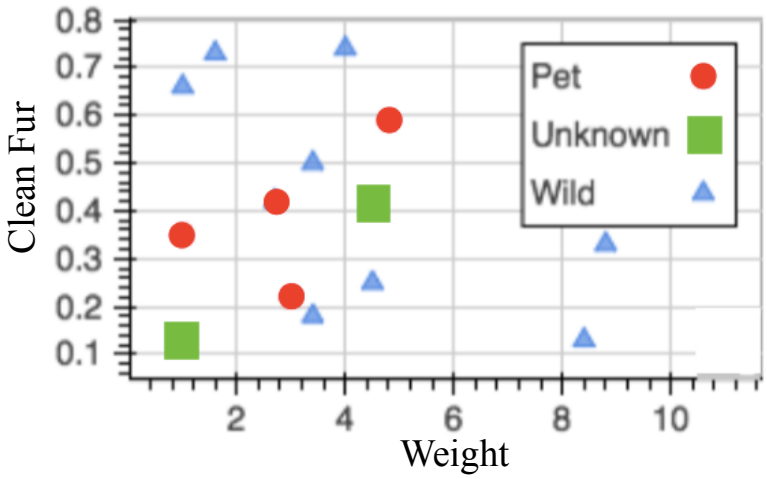
  ```
  np.prod(1-np.arange(20)/t.num_rows)
  ```

**6. (8 points) Classification**

Some squirrels on the Berkeley campus are wild animals and some are pets! Your friend is a Berkeley squirrel expert. She says that two useful features for classifying squirrels are the proportion of their fur that is clean (vertical axis) and their weight in ounces (horizontal axis). She identifies several squirrels for you, but leaves two "Unknown" squirrels for you to classify. No two squirrels have exactly the same features.

You decide to construct a classifier for the "Unknown" squirrels using all of the known "Pet" and "Wild" squirrels as the training set.

***** The horizontal and vertical axes have different scales in this scatter diagram! *****




**(a) (2 pt)** What is the distance in this feature space, roughly, between the bottom-left unknown squirrel and its nearest neighbor in the training set? Use both features to compute the distance. **Show your work!** *Hint*: The `distance` function appears on your final study guide.

(0.34 - 0.12) ** 2 ** 0.5 = 0.22

**(b) (2 pt)** How would each unknown squirrel be classified by a 1-nearest-neighbor classifier that uses only the "clean fur" feature on the vertical axis and ignores the weight?

Bottom-left unknown: Wild

Middle unknown: Pet

**(c) (2 pt)** On the scatter plot at the top of the page, circle the three nearest neighbors in the training set to the bottom-left unknown squirrel, using both features to compute distance.

**(d) (2 pt)** Circle *True* or *False*: Every possible pair of feature values would be classified as Wild for a 9-nearest-neighbor classifier using this training set. **Briefly justify your answer or describe a counterexample.**

True. There are only four Pets, so Wild will always have a majority among 9 points.

**7. (0 points) (Optional) The Art of Data Science**

Draw a picture (or better yet, a data visualization) of life before, during, and/or after taking Data 8.