# DS-100 Practice Midterm Questions

Spring 2017

**Note:** The following questions are intended to be representative of what you'll see on the midterm. The actual exam will have a single page (front and back) answer sheet on which to write each of the answers. The following questions are not guaranteed to cover every topic that's fair game for the exam, and this set is not representative of the length of the exam.

1. True or False

   (1) [1 Pt.] All data science investigations start with an existing dataset.

   > **Solution:** **False.** In many settings a data scientist is tasked with a question or problem and must decide how to collect or obtain data to answer the question or solve the problem.

   (2) [1 Pt.] Because smoking is viewed as a cause for lung cancer, it does not make sense to use lung cancer status to predict smoking status.

   > **Solution:** **False.** Lung cancer is likely a very good predictor of smoking habits. Just because there is no causal relationship (e.g., lung cancer doesn't cause smoking) does not mean that we cannot use one to predict the other.

   (3) It is possible that the null hypothesis will be rejected when it is in fact true.

   > **Solution:** **True.** There is always a small probability that we may erroneously reject the null hypothesis. (This is typically called Type 2 error, by the way.)

   (4) It is possible that the alternative hypothesis is true when the null hypothesis is not rejected.

   > **Solution:** **True.** It could be the case that the alternative hypothesis is true (though our tests can never affirm this) and we incorrectly fail to reject the null hypothesis (typically called a Type 1 error.)

   (5) Practically significant effects will always yield statistically significant results.

   > **Solution:** **False.** Take particle detection in physics for example. Results that are crucial to our understanding of the world are hard to detect and could be missed.

2. Consider two relations with the same schema: `R(A,B)` and `S(A,B)`. Which one of the following relational algebra expressions is not equivalent to the others?

   A. $\pi_{R,A}((R \cup S) - S)$

   **B.** $\pi_{R,A} R - \pi_{R,A}(R \cap S)$

   C. $\pi_{R,A}(R - S) \cap \pi_{R,A} R$

   D. They are all equivalent.

> **Solution:** This one is tricky—without the projection they'd all be equal, but the order of projection and the set operations causes option B to be different. A counter-example for B is $R = \{(1,2),(1,3)\}$, $S = \{(1,2)\}$. A and C will evaluate to $\{(1)\}$, while B will be $\emptyset$.

3. Consider the following real estate schema:

```
Homes(home_id int, city text, bedrooms int, bathrooms int,
      area int)
Transactions(home_id int, buyer_id int, seller_id int,
             transaction_date date, sale_price int)
Buyers(buyer_id int, name text)
Sellers(seller_id int, name text)
```

For the query language questions below, fill in the blanks in the answer to complete the query. For each SQL query and nested subquery, please start a new line when you reach a SQL keyword (SELECT, WHERE, AND, etc.). However, do not start a new line for aggregate functions (COUNT, SUM, etc.), and comparisons (LIKE, AS, IN, NOT IN, EXISTS, NOT EXISTS, ANY, or ALL.)

(1) Fill in the blanks in the SQL query to find the duplicate-free set of id's of all homes in Berkeley with at least 6 bedrooms and at least 2 bathrooms that were bought by "Bobby Tables."

```
SELECT DISTINCT H.home_id
FROM Homes H, Transactions T, Buyers B
WHERE H.home_id=T.home_id
AND T.buyer_id=B.buyer_id
AND H.city="Berkeley"
AND H.bedrooms>=6
AND H.bathrooms>=2
AND B.name='Bobby Tables';
```

(2) Repeat the above query using relational algebra:

$$\pi_{\text{home\_id}}(\sigma_{\text{name = 'Bobby Tables'}}(((\sigma_{\text{city='Berkeley' AND bedrooms >= 6 AND bathrooms >= 2}}\ \text{Homes}) \bowtie \text{Transactions}) \bowtie \text{Buyers}))$$

**Solution:** An alternative solution is to perform the selection on the Buyers table first

$$\pi_{\text{home\_id}}(\sigma_{\text{city='Berkeley' AND bedrooms >= 6 AND bathrooms >= 2}}(((\sigma_{\text{name = 'Bobby Tables'}}\ \text{Buyers}) \bowtie \text{Transactions}) \bowtie \text{Homes}))$$

(3) Fill in the blanks in the SQL query to find the id and selling price for each home in Berkeley. If the home has not ben sold yet, **the price should be NULL**.

```
SELECT H.home_id, T.sale_price
FROM Homes H
LEFT OUTER JOIN Transactions T
ON H.home_id = T.home_id
WHERE H.city = ’Berkeley’;
```

**Solution:** An alternate solution was to use Transactions in the FROM clause and perform a RIGHT OUTER JOIN with Homes.

```
SELECT H.home_id, T.sale_price
FROM Transactions T
RIGHT OUTER JOIN Homes H
ON H.home_id=T.home_id
WHERE H.city = ’Berkeley’
```

4. A biologist wants to figure out how a set of organisms are related to each other, so she runs an algorithm that divides the species into 3 groups based on their traits. This is an example of (choose all that apply):

   A. supervised learning

   **B. unsupervised learning**

   **C. clustering**

   D. classification

**Solution:** It is an example of unsupervised learning because she did not use training data that were labeled with the “right answers” (i.e. the right groups). It is an example of clustering (a particular kind of unsupervised learning) because she divided the unlabeled data into groups based on observed features. (FYI, an example of an unsupervised learning task that isn’t clustering is dimensionality reduction: finding a small number of numbers to encode the traits of each species.) Classification is a supervised learning task, where we are given training data labeled with correct classes (“spam” vs. “ham”, for example), and we use that data to parameterize (“train”) an algorithm to classify subsequent data into those same classes.

5. Chinua decides to investigate the effect of an IQ-enhancing drug, Versivium, on Berkeley students. He magically procures a list of all the students at Berkeley and their contact information. After randomizing the order of the list, he selects the top 1000 to invite to his experiment. Miraculously, all invitees agreed to be part of this clinical trial—anything for that A+. He divides the students into two groups: 800 in control and 200 in treatment. Members of the treatment group were administered a dose of Versivium every day for 20 days. Their counterparts in the control group were given an identical-looking sugar pill in the same dosage. All participants were compliant. At the end of the trial, the participants were asked to take an

IQ test. When Chinua performed a test against the null hypothesis that the two groups had the same mean IQ score, he found that the difference in test scores was significant at the 5% cutoff. Which of the following are valid statements?

A. The results of the test are actually inconclusive since the the control and treatment groups were uneven.

B. We should be doubtful of the results since Chinua took the top 1000 names of his ordered list.

C. The experiment is not well-designed as 1000 is too small a sample to make conclusions about the whole student body

**D. It could be the case that we were unlucky and the difference in IQ scores observed were due to chance**

E. None of the above

6. [3 Pts.] Consider the following plot about how baby boomers describe themselves. Which mistakes does it make? Circle all that apply.

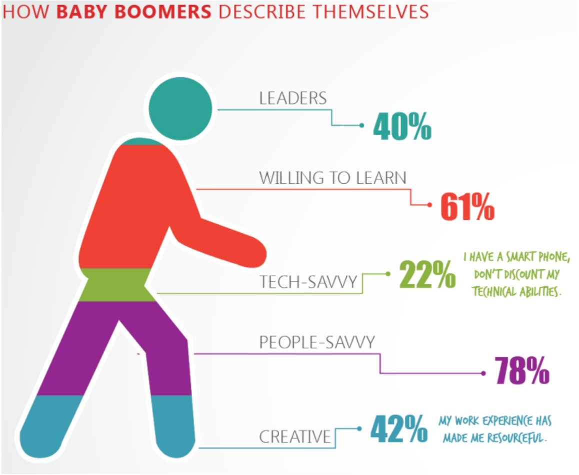
A. sampling bias

B. jiggling base line

**C. stacking**

D. jittering

**E. area perception**



7. [3 Pts.] The FEC data includes contributions to the Clinton and Sanders campaigns. If we want to create a visualization that helps us compare the sizes of donations to their campaigns, which of the following plots should we make? Circle all that apply.

A. scatter plot with donations to Clinton's campaign on one axis and Sanders' on the other.

**B. density curve of Clinton donations over laid on density curve of Sanders donations.**

C. side-by-side bar plot of their donations

**D. Two box plots, one for Clinton donations and one for Sanders.**

E. None of the above

8. [3 Pts.] Maximize the following likelihood with respect to $p$

$$L(p) = (k_1 k_2 k_3) p^6 (1-p)^{(k_1+k_2+k_3)-6}$$

Note, $\bar{k} = (k_1 + k_2 + k_3)/3$.

A. $\bar{k}/2$

**B.** $2/\bar{k}$

C. $-2/\bar{k}$

D. $2/(4+\bar{k})$

E. $2/(4-\bar{k})$

**Solution:** You could solve this very quickly by noticing that it is proportional to the binomial likelihood with $n = k_1 + k_2 + k_3$ when there are $k = 6$ successes, in which case the answer $\frac{6}{k_1+k_2+k_3} = \frac{2}{\bar{k}}$ is just the proportion of successes, which is the MLE for $p$. Or, you can derive the answer. First, take the logarithm:

$$\log L(p) = \log(k_1 k_2 k_3) + 6\log(p) + (k_1 + k_2 + k_3 - 6)\log(1-p)$$

Then find where the derivative with respect to $p$ equals 0 (we'll call the solution to that equation $p^*$):

$$0 = [\frac{d}{dp} \log L](p^*) = \frac{6}{p^*} - \frac{k_1 + k_2 + k_3 - 6}{1 - p^*}$$

$$6(1 - p^*) - (k_1 + k_2 + k_3 - 6)p^* = 0$$

$$p^* = \frac{6}{k_1 + k_2 + k_3}$$

9. Suppose $X$, $Y$, and $Z$ are random variables that are independent and have the same probability distribution. If $\mathrm{Var}(X) = \sigma^2$, then $\mathrm{Var}(X + Y + Z)$ is:

A. $9\sigma^2$

**B.** $3\sigma^2$

**Solution: This is the correct answer because variance is additive for independent random variables.**

C. $\sigma^2$

D. $\frac{1}{3}\sigma^2$

10. A jar contains 3 red, 2 white, and 1 green marble. Aside from color, the marbles are indistinguishable. Two marbles are drawn at random without replacement from the jar. Let $X$ represent the number of red marbles drawn.

(1) [2 Pts.] What is $\mathbb{P}(X = 0)$?

A. $1/9$

**B.** $1/5$

C. $1/4$

D. $2/5$

E. none of the above

**Solution:** The event that $X = 0$ is the same as the event that no red marbles are drawn, which is the same as the event that the first draw isn't red and the second draw isn't red.

$p = P(\text{first draw is not red and second draw is not red})$
$= P(\text{first draw is not red})P(\text{second draw is not red — first draw is not red})$

If the first draw isn't red, there are 5 marbles left, 3 of which are red, so:

$$p = \frac{1}{2}\frac{2}{5} = \frac{1}{5}$$

.

A more brute-force counting argument is as follows. There are $\binom{6}{2} = \frac{6!}{4!2!} = 15$ ways to draw a subset of 2 marbles. Of those, the number of subsets with no red marbles is $\binom{3}{2} = \frac{3!}{2!1!} = 3$, so the proportion of draws without red marbles is $3/15 = 1/5$. However it's probably better to exercise your probabilistic thinking via the previous solution!

(2) [2 Pts.] let $Y$ be the number of green marbles drawn. What is $\mathbb{P}(X = 0, Y = 1)$?

A. $\frac{1}{15}$

**B.** $\frac{2}{15}$

C. $\frac{1}{12}$

D. $\frac{1}{6}$

E. $\frac{7}{15}$

F. $\frac{8}{15}$

**Solution:** For $X$ to be 0 and $Y$ to be 1, means that we drew 1 green and 1 white ball. We can draw green first and then white, which has chance $1/6 \times 2/5$ or white first and green second, which has chance $2/6 \times 1/5$. The combined probability is $4/30$ or $2/15$.

Another approach is to use conditional probability, i.e.,

$$\mathbb{P}(X = 0, Y = 1) = \mathbb{P}(X = 0)\mathbb{P}(Y = 1|X = 0).$$

We found $\mathbb{P}(X = 0)$ above to be $1/5$. For the conditional probability, if we know $X = 0$ then we know that we are drawing from the 2 white and 1 green marbles. There are 3 possible ways to draw 2 marbles from these 3 and 2 of the possibilities give us 1 green and 1 white. Putting these together we have $1/5 \times 2/3 = 2/15$.

Alternatively, we can brute-force count the number of subsets that have 1 green and one white marble, which is 2, and divide by the number of ways to choose 2 marbles out of 6 (which we calculated above to be 15).

11. [3 Pts.] Suppose the random variable $X$ can take on values $-1$, $0$, and $1$ with chance $p^2$, $2p(1-p)$ and $(1-p)^2$, respectively, for $0 \leq p \leq 1$. We observe a sequence of 3 independent observations from this distribution. They are $1, 1, 0$.

    What is the likelihood for $p$?

    A. $2p(1-p)^3$

    B. $p^2(1-p)^2$

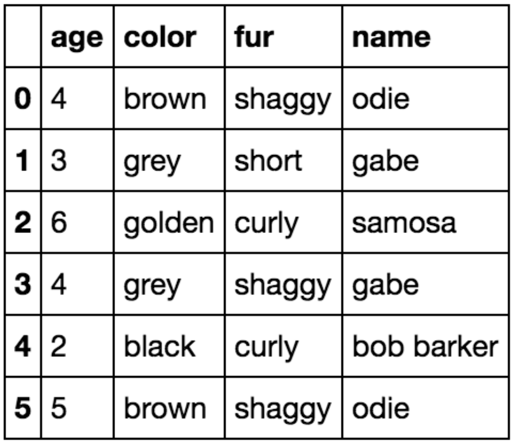
    C. $(1-p)^4$

    **D.** $2p(1-p)^5$

    E. $2p^3(1-p)^3$

**Solution:** Since the 3 observations are independent, we just multiply the probability of each one: $(1-p)^2 \times (1-p)^2 \times 2p(1-p)$.

12. [8 Pts.] The pandas dataframe *dogs* contains information on pets' visits to a veterinarian's office. A portion of the dataframe is shown below.



| | age | color | fur | name |
|---|---|---|---|---|
| 0 | 4 | brown | shaggy | odie |
| 1 | 3 | grey | short | gabe |
| 2 | 6 | golden | curly | samosa |
| 3 | 4 | grey | shaggy | gabe |
| 4 | 2 | black | curly | bob barker |
| 5 | 5 | brown | shaggy | odie |

For each question, provide a snippet of pandas code as your solution. Assume that the table *dogs* has the same column format as the provided table (just more rows).

(1) How many different dogs visited the veterinarian's office? Provide code that outputs the answers as an integer. Assume that no two dogs have the same name.

**A.** `len(dogs.groupby("name").count())`

B. `len(dogs["name"])`

C. `len(dogs)`

> **Solution:** Note that the second and third choices do not account for duplicate appearances by the same name.

(2) What was the name of the oldest dog that visited the veterinarian's office?

A. `dogs.sort_values("age", ascending=False).name[0]`

**B.** `dogs.sort_values("age", ascending=False).name.iloc[0]`

C. `dogs.groupby("name").agg({"age": "max"})`

> **Solution:** The first solution would return the dog which had pandas index 0 (that is, the one that appeared in the first row of the dataframe *before* sorting). The third solution returns the maximum age recorded for each dog, but doesn't choose the oldest among them.

(3) What was the most common fur color among dogs?

A.
```
dogs.groupby("color").count().sort_values("name",
        ascending=False).index[0]
```

B.
```
dogs.groupby("color").count().sort_values("age",
        ascending=False).index[0]
```

C.
```
dogs.groupby("color").count().sort_values("fur",
        ascending=False).index[0]
```

**D. All of the above.**

E. None of the above.

(4) What proportion of dogs had the most common fur type? (For instance, if the most common fur type was curly, what proportion of dogs had curly fur)?

A.
```
dogs.groupby("fur").count().sort_values("age",
        ascending=False).age.iloc[0]/len(dogs)
```

B.
```
dogs.groupby("fur").count().sort_values("age",
        ascending=False).age.iloc[0]/len(dogs["age"])
```

C.
```
dogs.groupby("fur").count().sort_values("age",
        ascending=False).age.iloc[0]/len(dogs["fur"])
```

**D. All of the above.**

E. None of the above.