

Microquiz 4

1.

1-1. John ran a single trial of 10,000 Monte Carlo simulations of a game with a binary outcome. He won 1,000 times and lost 9,000 times.

- ** The best estimate of the probability of winning is 0.1.
- It is appropriate to compute a confidence interval using SD.
- None of the above.

1-2. John ran a single trial of 10,000 Monte Carlo simulations of a game with a continuous outcome between 0 and 100. The average score was 50.

- ** The best estimate of the expected score is 50.
- It is appropriate to compute a confidence interval using SD.
- ** It is appropriate to compute a confidence interval using SE.
- None of the above.

1-3. D is a normal distribution with a mean of 0 and a standard deviation of 1.

- More than half the values in D are between 0 and 1.
- ** The median value of D is 0.
- ** The probability of drawing the value 0 from D is less than 0.0001
- None of the above.

1-4. Consider the following code:

```
def rSquared(m, p):  
    eErr = ((p-m)**2).sum()  
    mean = m.sum()/len(m)  
    var = ((m - mean)**2).sum()  
    return 1 - eErr/var  
  
def f(X, epsilon):  
    Y = []  
    for x in X:  
        Y.append(x**2 + random.gauss(0, epsilon))  
    return pylab.array(Y)
```

```
X = range(1, 100)  
data1 = (X, f(X, 1000))  
data2 = (X, f(X, 10))  
model1 = pylab.polyfit(data1[0], data1[1], 2)  
model2 = pylab.polyfit(data1[0], data1[1], 3)  
model3 = pylab.polyfit(data2[0], data2[1], 2)
```

- ** R-squared for model2 should be better than for model1.
- ** R-squared for model1 and for model2 should be close to the same.
- ** R-squared for model3 will be larger than R-squared for model2
- None of the above.

1-5. Which of the following is true about k-means clustering?

- ** Once the initial centroids have been chosen, the algorithm is deterministic.
- One problem with k-means clustering is that for small k it often takes a long time to converge.
- ** One problem with k-means clustering is that it can generate an empty cluster.
- As k grows, the average intra-cluster distance tends to grow.
- The clustering found is independent of the distance metric used.
- None of the above.

1-6. Which of the following are true?

- ** Z-scaling ensures that the values for each feature will have a mean of 0 and a standard deviation of 1.

- Linear interpolation ensures that the values for each feature will lie between 0 and 1 with a mean of 0.5
- None of the above.

1-7. Which of the following is true about KNN classification

- The larger k, the more accurate the classification
- The larger k, the longer classification takes.
- When k=1, KNN is the same as linear regression.
- KNN tends to work poorly when classes are reasonably well balanced.
- ** None of the above.

2.

```
def optimize(s):
    """
    s: positive integer, what the sum should add up to
    Solves the following optimization problem:
        x1 + x2 + x3 + x4 is minimized
        subject to the constraint x1*25 + x2*10 + x3*5 + x4 = s
        and that x1, x2, x3, x4 are non-negative integers.
    Returns a list of the coefficients x1, x2, x3, x4 in that order
    """
    denom = [25, 10, 5, 1]
    result = []
    for i in denom:
        div = s//i
        s -= div*i
        result.append(j)
    return result
```

3.

```
def estimate_g(times, velocities, planet):
    model = np.polyfit(times, velocities, 1)
    estVals = np.polyval(model, times)
    r2 = rSquared(velocities, estVals)
    return (model[0], model[1], r2)
```