

## Solutions 5

### Jumping Rivers

#### Z-score

1. Write a function that processes data by applying the slide and squish rule to implement the z-score transformation

```
import numpy as np
```

```
def zscore(x):  
    m = np.mean(x)  
    s = np.std(x, ddof=1)  
    z = (x - m)/s  
    return z
```

2. Load the random data from the first practical using `jrpyml.get_numeric_list()` and apply your transformation. Ensure that the mean and variance are indeed as expected.

```
import jrpyml
```

```
x1 = jrpyml.get_numeric_list()  
xz = zscore(x1)
```

```
xz.mean()
```

```
## -5.110992987009403e-17
```

```
xz.var(ddof=1)
```

```
## 1.0
```

#### Means of distributions

The `numpy.random.choice` function can be used to generate samples from an array e.g

```
x = np.array([3, 6, 2, 1, 10, 2])  
np.random.choice(x, 3)
```

```
## array([1, 2, 2])
```

1. Write a function that takes 3 arguments
  - Some array of data, `x`
  - A sample size, `n`

- A replicate count,  $r$

This function should find  $r$  means of sample size  $n$  from the array  $x$  and return them

```
def replicate_means(x, n, r):
    output = []
    for i in range(r):
        sample = np.random.choice(x, n)
        output.append(sample.mean())
    return output
```

2. Generate a histogram of the number of votes on the movies data

```
import matplotlib.pyplot as plt
movies = jrpyml.datasets.load_movies()

movies['votes'].plot.hist(bins=100)
plt.show()
```

3. Does this distribution look normal?

```
answer = """
No, Very skewed towards zero
"""
```

4. Using your newly written function for calculating means, calculate 1000 averages on samples of size 5000 from the collection of all votes

```
means = replicate_means(movies['votes'], 100, 100)
```

5. Draw a histogram of these means

```
plt.hist(means)
plt.show()
```

6. Does the distribution now look normal?

```
answer = """
Definitely looks a lot closer
"""
```

7. Experiment with different sample sizes, what do you observe

```
answer = """
The larger the sample size, the closer to a normal distribution it looks
"""
```

What we are doing here is exploring a piece of mathematics theory known as the central limit theorem which we will explore in the next chapter.