Computer Science 477/577 - Homework 1

Matt Jensen

Very Late

The simple mean (exemplified on page 88).

```
data = pd.read_csv(paths('data') / 'hw' / 'q1.csv', sep="|")
data = data.sort_values('salary').reset_index(drop=True)
mean = sum(data.salary) / len(data.salary)
print(f"mean: {mean:.1f}")
```

mean: 3458.9

The weighted mean (exemplified on page 89).

```
count = data.groupby('salary')['salary'].count()
weighted_mean = sum([a * b for a, b in zip(list(count.index), list(count))]) / len(data)
print(f"weighted: {weighted_mean:.1f}")
```

weighted: 3458.9

The median.

```
median = data.iloc[len(data) // 2]['salary']
print(f"median: {median}")
```

median: 2100

The mode.

```
counts = dict(zip(list(count.index), list(count)))
mode = max(counts, key=counts.get)
print(f"mode: {mode}")
```

mode: 1400

The geometric mean

```
total = 1
for i in data.salary:
    total *= i
geometric = total ** (1 / len(data))
print(f"geometric: {geometric:.1f}")
```

geometric: 2421.0

The variance

```
variance = sum(((data - mean) ** 2)['salary']) / len(data)
print(f"variance: {variance:.1f}")
```

variance: 21835940.5

The standard deviation.

```
std = math.sqrt(variance)
print(f"std: {std:.2f}")
```

std: 4672.89

The Z score.

```
z_scores = round((data - mean) / std, 2)
z_scores = list(zip(data.salary, z_scores.salary))
print(f"z_scores: {z_scores}")
```

z_scores: [(1400, -0.44), (1400, -0.44), (1400, -0.44), (1400, -0.44), (1400, -0.44), (1400, -0.44), (1400, -0.44), (1400, -0.44), (1400, -0.44), (1400, -0.44), (1400, -0.44), (2100, -0.29), (2500, -0.21), (2500, -0.21), (2500, -0.21), (2500, -0.21), (3472, 0.0), (3500, 0.01), (3500, 0.01), (3500, 0.01), (5500, 0.44), (5500, 0.44), (7600, 0.89), (25000, 4.61)]

The coefficient of variation.

```
coeff_v = std / mean * 100
print(f"coeff. of var.: {coeff_v:.2f}%")
```

coeff. of var.: 135.10%

The first quartile.

```
q_1 = data.iloc[len(data) // 4]['salary']
print(f"first quartile: {q_1}")
```

first quartile: 1400

The third quartile.

```
q_3 = data.iloc[(len(data) // 4) * 3]['salary']
print(f"third quartile: {q_3}")
```

```
third quartile: 3500
```

In a colophon to table 4.1 on page 93 of History by Numbers, it remarks that "The mode is interpolated from the mean and the median and not derived directly from the data." How exactly was this calculated?

```
data = pd.read_csv(paths('data') / 'hw' / 'a1_q2.csv', sep="|")
mode = (3 *data['median']) - (2 * data['mean'])
print(f"mode: {mode.values}")
```

```
mode: [23.  23.8 23.9 24.4 23.7 23.3 23.5 21.8]
```