CS109: Introduction to Probability for Computer Scientists

Lecture 19: Bootstrapping

Olivia Beyer Bruvik

Winter 2022

Lecture 19

Sample

A sample of sample size 8: $(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$

A realization of sample size 8: $(59, 87, 94, 99, 87, 78, 69, 91)$

Sample mean, $\bar{x}$

Sample mean is an RV with known Var.

By central limit theorem, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ $Var(\bar{X}) = \frac{\sigma^2}{n}$

$E[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} X_i$

Unbiased estimate

The expected value of the sample mean is the true mean.

Your estimate of the true mean is the average of your samples.

Estimating population variance

Problem: Estimate is σ^2, the variance of happiness of Bhutanese people.

1. Population variance: $\sigma^2 = E[(X-\mu)^2] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
2. Sample variance: $E[S^2] = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - E[\bar{X}])^2$

Sample variance is an estimate using an estimate so it needs additional scaling. We also systematically underestimate distance between datapoints and the true mean when using the estimated mean.

Unbiased estimate

S^2 is an unbiased estimator of the population variance, σ^2. $E[S^2] = \sigma^2$

Error bars

$Std(\bar{X}) = \sqrt{\frac{E[S^2]}{n}}$

Example: p-set time

```
def analyse(data):
  for question_key, timings_list in data.items():

    # calculate n
    n = len(timings_list)

    # estimate the mean
    sample_mean = np.mean(timings_list)

    # estimate the variance
    sample_var = np.var(timings_list)

    # estimate the standard error of the mean
    standard_err = math.sqrt(sample_var / n)

    # sample std
    sample_std = math.sqrt(sample_var)

    # print them out
    display_name = question_key[:12]
    print(f'{display_name},\tmean: {sample_mean:.1f} ± {standard_err:.1f},\tstd: {sample_std:.1f}')
```

Bootstrap

- Uses:
  - know the distribution of statistics
  - calculate p-values

Hypothetical

- What is the probability that the mean of a sample of 200 people is within the range of 81 to 85?
- What is the std of the sample variance, calculated from 200?
- What is the std of the sample variance, calculated from 200, if you know the true distribution?
  - 10,000 times take a mock sample of 200, calculate the sample variance.

Bootstrapping method

1. Bootstrapping assumption: your sample is the best guess you have as to the full distribution.

$$F \approx \hat{F}, \text{ where F is the undelying distribution, and } \hat{F} \text{ is the sample distribution}$$

2. Normalize a histogram of your data to estimate the PMF of the underlying distribution, using your sample.
3. Bootstrapping assumption:

Algorithm

1. Estimate the PMF using sample
2. Repeat 10,000 times

a. Resample len(sample) from PMF
b. Recalculate the stat (mean/var) on the resample

3. You now have a distribution of your stat (mean/var)