Introduction to Statistics (Chapter 8)

Statistics: Learning about a population based on a sample from it. Recall a general *statistical inference* framework:

Statistic: Any feature of the sample data. They are used construct *estimators* of features of the population.

Sampling and non-sampling errors: Discrepancy between a sample and the whole population.

- *Sampling error* is caused by the fact that only a portion of the population is sampled. In most cases, this error reduces as n increases.
- *Non-sampling error* occurs if the sample is biased, i.e., it is not representative of the population of interest. Avoid well-know problems, such as selection bias, non-response bias, investigator bias, etc., while collecting data.

Random sample: $X_1, \ldots, X_n$ are independent and have the same distribution as X

- IID (independently and identically distributed) data
- Sample is representative of population.

Ex: To evaluate effectiveness of a processor for a certain type of tasks, we recorded the CPU time for $n = 30$ random chosen jobs (in seconds): 70, 36, 43, 69, 82, 48, 34, 62, 35, 15, 59, 139, 46, 37, 42, 30, 55, 56, 36, 82, 38, 89, 54, 25, 35, 24, 22, 9, 56, 19. What is population? X? Sample? Distribution of X?

Desirable properties of an estimator $\hat{\theta}$ of θ

$\hat{\theta}$ will have a *probability distribution* — induced by randomness in the sampling process. It is called *sampling distribution* of $\hat{\theta}$.

Unbiasedness:

- $\hat{\theta}$ is unbiased for θ if $E(\hat{\theta}) = \theta$ for all θ.
- Estimator is correct on average.

Small variance

- Variance = uncertainty.
- Larger variance = less precise.
- We would like to have small variance or high precision.
- Standard error (se) of $\hat{\theta}$ = standard deviation of $\hat{\theta}$

Consistency:

- $\hat{\theta}$ is consistent for θ if it converges to θ as $n \to \infty$.
- Necessary for a reasonable estimator.
- Why use an estimator that does not become more accurate as n increases?

Asymptotic normality:

- For large n, $\hat{\theta}$ approximately follows $N(\theta, \text{var}(\hat{\theta}))$.
- Consequence of CLT and related results.
- Useful for designing inference procedures that are valid for large n

Some descriptive statistics and what they estimate

Mean:

Population mean:

Sample mean:

Properties of $\overline{X}$:

-
-
-
-
- Greatly affected by outliers

Ex: (CPU data): $\overline{X}$=?

Median:

Population median: The smallest value M such that

$$F(M) = P(X \leq M) \geq 0.5.$$

Essentially M is a *middle* value — it divides the probability distribution in two halves.

M for a Continuous distribution:

Ex: Suppose $X \sim \text{Exponential}(\lambda)$. Recall its cdf, $F(x) = 1 - e^{-\lambda x}$ for $x > 0$. What is M?

$\underline{M \text{ for a discrete distribution}}$:

Problem 1: $F(M) = 0.5$ may have a whole interval of roots.

- Median not unique
- Take the mid-point of the interval as the median.

Problem 2: $F(M) = 0.5$ may not have any root.

This is why we take M to be the smallest value for which $F(M) \geq 0.5$. We now have a unique value for median.

Ex: Look at Figure 8.4 and find the median.

Sample median

Characteristic shapes of a distribution

Symmetric:

Right-skewed:

Left-skewed:

Bimodal or multi-modal:

Which measure of center to use — mean or median?:

Descriptive statistics and what they estimate (continued)

p-quantile of a population: The smallest value q_p such that

$$F(q_p) = P(X \leq q_p) \geq p.$$

Essentially X has p probability on the left of q_p.

p-quantile of a sample: Take $\hat{q}_p$ to be the (np)-th largest value in the sample. If np is not an integer, round it up to the next integer (i.e., apply the ceiling function). Alternatively, $\hat{q}_p$ is the smallest value in the sample that has at least p proportion of observations on its left (including itself).

- $\hat{q}_p$ estimates q_p
- 0.5-quantile =
- **Population quartiles**: $(Q_1, Q_2, Q_3) = (q_{0.25}, q_{0.50}, q_{0.75})$ — they divide the distribution in four equal parts.
- **Sample quartiles:**
- **5-number summary:**

Ex: (CPU data) Sample quartiles of the CPU data.

```
# > sort(cpu)
# [1]    9  15  19  22  24  25  30  34  35  35  36  36
37  38  42  43  46  48
# [19]  54  55  56  56  59  62  69  70  82  82  89 139
# >
```

Population variance: $\sigma^2 = E((X-\mu)^2) = E(X^2) - \mu^2$

Sample variance:

Properties:

-
-
- Measure of spread or variability
- Standard deviation (SD) $= \sqrt{\text{variance}}$
- Estimated standard error (SE) of $\overline{X} =$

Ex: (CPU data)

Interquartile range (IQR):

Population:

Sample:

Properties:

-
-
-

Rule of thumb for "outlier" detection: An observation may be considered an "oulier" if it falls outside the interval from $\hat{Q}_1 - 1.5 * \widehat{IQR}$ to $\hat{Q}_3 + 1.5 * \widehat{IQR}$.

Ex: (CPU data): Estimated (or sample) IQR=? Could the observation 139 be an outlier?

Which measure of spread to use — SD or IQR?: