Introduction to Statistics (Chapter 8)

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

Population: $X \sim f_\theta(x)$ — pmf / pdf

θ = parameter of this dist. (unknown)

$\downarrow$

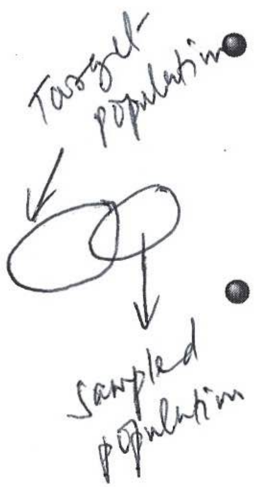
$X_1, X_2, \ldots, X_n$ (random sample)

Goal: Learn about θ using the sample data, e.g., using an estimator $\hat{\theta}$ of θ. — point estimation

a random quantity

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.

Assume: no non-sampling error.

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 (x_1), 36 (x_2), 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 (x_{30}). What is population? X? Sample? Distribution of X?

Population: Collection of all CPU times for this particular type of task.

X = A randomly selected CPU time from this population (A typical "CPU time" for this particular task).

Two ways:

(a) Parametric statistics: Assume a prob. dist. for X (and be sure to check that assumption is reasonable;) simpler, ~~requires~~ doesn't require n to be very large

(b) Nonparametric statistics — Don't assume any particular prob. dist. for X — less flexible, often harder to interpret, requires large n.

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 $\hat{\theta}$

Player 1:

- unbiased player, average throw = target
- high variability

Player 2:

- biased player but has low variability (high precision)

Estimation:

θ

$\hat{\theta}$ is unbiased

Ideal player:

- unbiased + very small variability

θ

$\hat{\theta}$ is biased

θ

unbiased + small variability

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

Some descriptive statistics and what they estimate

Assume: $X_1, X_2, \ldots, X_n$ a random sample from a population; $\mu = E(X)$, $\sigma^2 = var(X)$

Mean:

Population mean: μ

Sample mean: $\overline{X}$

Properties of $\overline{X}$:

consistency of $\overline{X}$ ⇐

- $E[\overline{X}] = \mu \Rightarrow \overline{X}$ is unbiased for μ.
- $var[\overline{X}] = \frac{\sigma^2}{n}$
- LLN: If n is large, $\overline{X} \approx \mu$
- CLT: If n is large, $\overline{X} \dot\sim N[\mu, \frac{\sigma^2}{n}]$ — $E(\overline{X})$, $var(\overline{X})$.
- Greatly affected by outliers

Often a good choice, but in some cases, we can find better estimators than $\overline{X}$.

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

will see later (using R).

Median:

Population median: The smallest value M such that

$$F(M) = P(X \leq M) \geq 0.5. \quad (1)$$

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

M for a Continuous distribution:

X is cont:

$F(M) = 0.5$

$\Rightarrow M = F^{-1}(0.5)$ (2)

(1) and (2) are equivalent.

— M = 0.5 quantile = ~~...~~ 50th percentile.

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

solve:

$$F(M) = 0.5$$

$$\Rightarrow \quad 1 - e^{-\lambda M} = 0.5$$

$$\Rightarrow \quad e^{-\lambda M} = 0.5$$

$$\Rightarrow \quad -\lambda M = \log[0.5]$$

$$\Rightarrow \quad \boxed{M = -\frac{1}{\lambda} \log[0.5]}$$

<u>M for a discrete distribution</u>:

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.

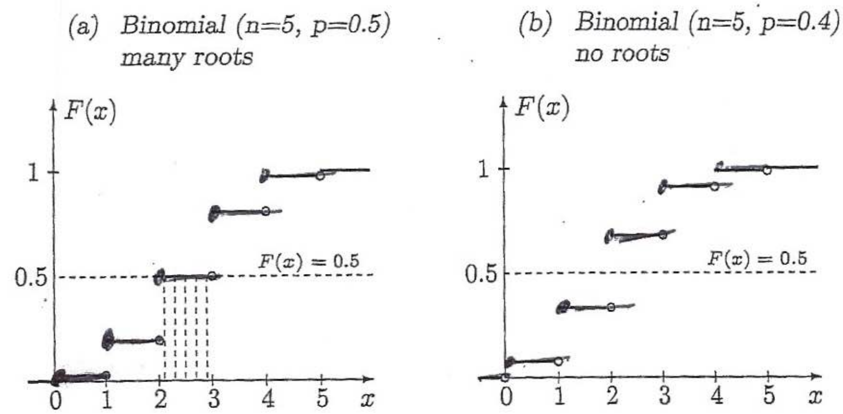
Problem 2: $F(M) = 0.5$ may not have any root.

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.

F(M) = 0.5

⇒ M = 2



No M: F(M) = 0.5

Def: ⇒ M = 2

FIGURE 8.4: *Computing medians of discrete distributions.*

This result agrees with our intuition. With $p = 0.5$, successes and failures are equally likely. Pick, for example, $x = 2.4$ in the interval (2,3). Having fewer than 2.4 successes (i.e., at most two) has the same chance as having fewer than 2.4 failures (i.e., at least 3 successes). Therefore, $X < 2.4$ with the same probability as $X > 2.4$, which makes $x = 2.4$ a central value, a median. We can say that $x = 2.4$ (and any other x between 2 and 3) splits the distribution into two equal parts. Then, it is a median. ◊

Example 8.11 (Asymmetric Binomial, Figure 8.4b). For the Binomial distribution with $n = 5$ and $p = 0.4$,

$$\begin{array}{lll} F(x) < 0.5 & \text{for} & x < 2 \\ F(x) > 0.5 & \text{for} & x \geq 2 \end{array}$$

but there is no value of x where $F(x) = 0.5$. Then, $M = 2$ is the median.

Seeing a value on either side of $x = 2$ has probability less than 0.5, which makes $x = 2$ a center value. ◊

Computing sample medians

A sample is always discrete, it consists of a finite number of observations. Then, computing a sample median is similar to the case of discrete distributions.

In simple random sampling, all observations are equally likely, and thus, equal probabilities on each side of a median translate into an equal number of observations.

Again, there are two cases, depending on the sample size n.

Sample median

If n is odd, the $\left(\frac{n+1}{2}\right)$-th smallest observation is a median.
If n is even, any number between the $\left(\frac{n}{2}\right)$-th smallest and the $\left(\frac{n+2}{2}\right)$-th smallest observations is a median.

Sample median

— Sort the data

$$\hat{M} = \begin{cases} \text{the value in the middle}, & n \text{ is odd} \\ \text{Avg. of the two middle values}, & n \text{ is even} \end{cases}$$

Descriptive statistics and what they estimate (continued)

= ~~100p~~ (100×p)th percentile of X.

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

$$F(\underline{q_p}) = P(X \le q_p) \underline{\ge p}. \quad \text{— (1)}$$

If X is cont.
(1) is equivalent to:
$F(\xi_p) = p$
$\Rightarrow \xi_p = F^{-1}(p)$

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 = median (M).
- **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:** $\hat{Q}_1, \hat{Q}_2, \hat{Q}_3$
- **5-number summary:** $(\min, \hat{Q}_1, \hat{Q}_2, \hat{Q}_3, \max)$ — boxplot is a graphic of this

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

$n = 30$

```
# > 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
# >
```

#15 #16 #8 #23

$\hat{M} = \hat{Q}_2 = \frac{42+43}{2} = 42.5$

$\hat{Q}_1 = (np)$-th largest obs $= \underbrace{(30 \times 0.25)}_{= 7.25}$ th largest obs.

$p = 0.25$

$\underset{\text{rounding up}}{=}$ 8 th largest obs. $= 34$.

$\hat{Q}_3 = (30 \times 0.75)$ th largest obs. $= (22.5)$ th obs.

$\underset{\text{rounding up}}{=}$ 23 rd obs $= 59$.

5-# summary: $[9, 34, 42.5, 59, 139]$

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

Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Properties:

- $E[S^2] = \sigma^2 \Rightarrow S^2$ is unbiased for σ^2.
- S^2 is consistent.
- Measure of spread or variability
- Standard deviation (SD) = $\sqrt{\text{variance}}$
- Estimated standard error (SE) of $\overline{X}$ =

Recall:

$var(\overline{X}) = \frac{\sigma^2}{n}$

$\Rightarrow SE(\overline{X}) = \frac{\sigma}{\sqrt{n}}$

$\Rightarrow \boxed{\widehat{SE}(\overline{X}) = \frac{S}{\sqrt{n}}}$

Ex: (CPU data)

see later using R.

Interquartile range (IQR):

Population: $IQR = Q_3 - Q_1$

Sample: $\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1$

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?