Recap

Statistical inference:

Population: $X \sim f_\theta(x)$
θ is unknown

$\downarrow$

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

- use data to learn about θ
- (Point) estimator: $\hat{\theta}$ ← a statistic used to estimate θ

Desirable properties of $\hat{\theta}$:

- Unbiased
- small variance
- consistent
- Asymptotically normal

Descriptive statistics:

- sample mean $\bar{X}$ estimates population mean μ.

Some descriptive statistics and what they estimate

Have a random sample $X_1, \ldots, X_n$ from the population of X.

Mean: → measure of center of the dist.

Population mean: $\mu = E[X]$

Sample mean: $\bar{X} = \sum_{i=1}^{n} x_i / n$

Properties of $\overline{X}$: natural estimator of μ [may be possible to find better estimator]

- $\bar{X}$ is a
- LLN: $\bar{X}$ is consistent.
- $E[\bar{X}] = \mu$ for all $\mu \Rightarrow \bar{X}$ is unbiased. $\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n}$ ($\mathrm{Var}(X)$)
- CLT: $\bar{X} \dot{\sim} N[\mu, \frac{\sigma^2}{n}]$ if n is large (Asymptotic normality)
- Greatly affected by outliers — Have outliers in the data if the parent distn. is not symmetric.

Ex: (CPU data): $\overline{X}$=? $\bar{X} = \frac{70 + 36 + \cdots + 19}{30} = 48.22$ (?) using R

SE[$\bar{X}$] = ? — will see this later.

Median:

Another measure of center of a dist.

Population median: A value M such that

soln. may not be unique if dist. is discrete.

$$\left\{ P(X > M) \leq 0.5, \text{ and } P(X < M) \leq 0.5. \right\} \quad (*)$$

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

M for a Continuous distribution:

- $F(M) = p$
- CDF $F(x) = P[X \leq x] = P[X < x]$ is cont. and ↑ fn. of x.

$(*)$ $F(M) \leq 0.5$,

$P[X > M] = 1 - F(M) \leq 0.5$

$\Rightarrow F(M) \geq 0.5$

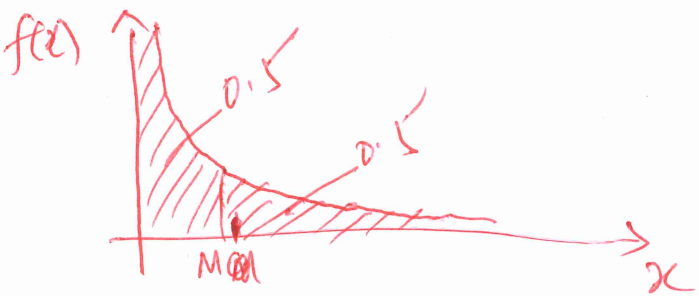
$\Rightarrow$ $F(M) = 0.5$

Solve this equation to get M — unique.

To force uniqueness, revise (K) as:

M is the smallest x such that $F(x) \geq 0.5$. — (K*)

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



Solve for M:

$$F(M) = \boxed{1 - e^{-\lambda M} = 0.5}$$

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

$$\Rightarrow \quad -\lambda M = \log(0.5)$$

$$\Rightarrow \quad \boxed{M = -\frac{1}{\lambda}\log(0.5)}$$

M 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. Take the smallest x with $F(x) \geq 0.5$ as the median.

Ex: Look at Figure 8.4 and find the median.

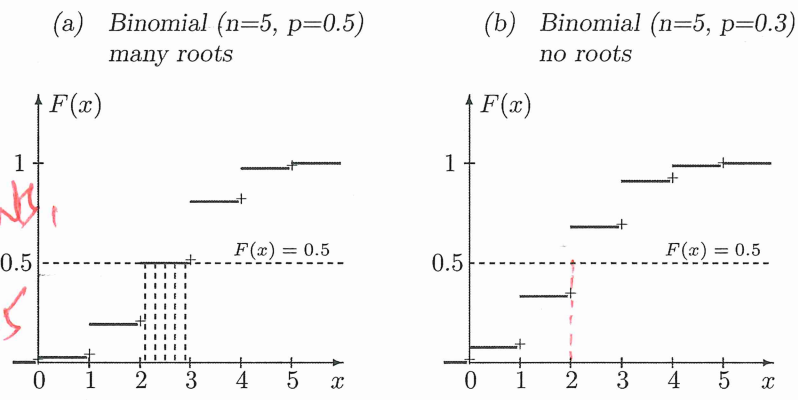


Figure 8.4 *Computing medians of discrete distributions.*

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 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.

Example 8.11 (MEDIAN CPU TIME). Let's compute the median of $n = 30$ CPU times from the data on p. 225.

Sample median

– midpoint (or middle value) in the sample.

– Book:

$$\hat{M} = \begin{cases} \text{middle obs.} & , n \text{ is odd} \\ \text{average of the two middle obs.} & , n \text{ is even} \end{cases}$$

– can also use new def. of median.

- $\hat{M}$ estimate M.
- Difficult to estimate $\text{Var}[\hat{M}]$ – will use 'bootstrap' for this purpose.

Ex: CPU data: $n = 30$.

Sorted data: 9, 15, 19, ——, 42, 43, 46, ——, 139

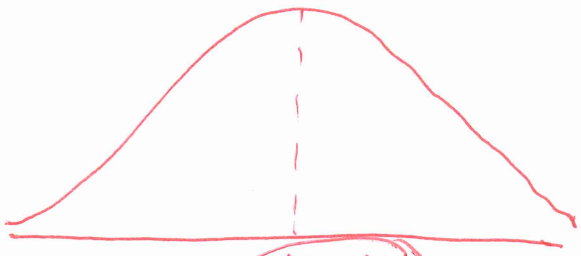
#15 #16

14 obs. 14 obs.

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

X = income of a randomly selected Dallas resident → typically right-skewed.

Symmetric dist:

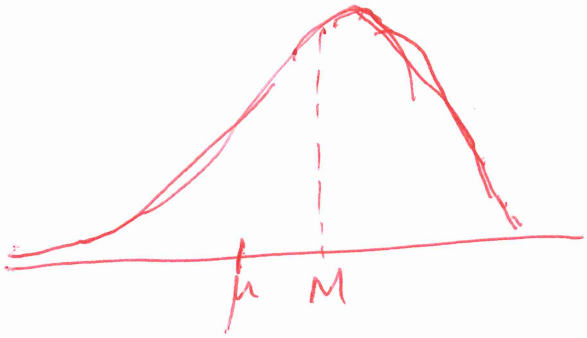


Two tails are equal

- Makes sense to use mean for center of the dist.
- use $\bar{X}$ to estimate μ.

$M = \mu$ → provided μ exists.

Left-skewed dist:



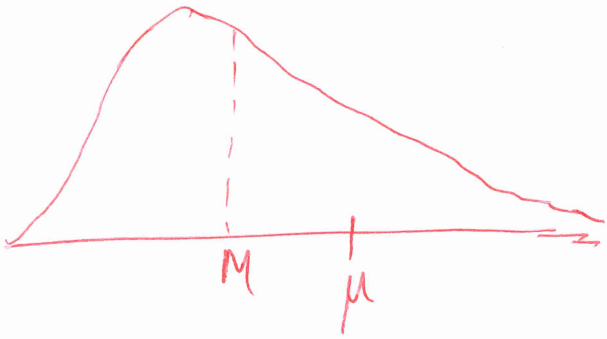
Left tail is longer

- Mean (μ) is not a good measure of center of dist.
- Don't use $\bar{X}$.

↑ same situation with right-skewed dist.

Right-skewed dist.

Long right tail



Characteristic shapes of a distribution

Symmetric:

Right-skewed:

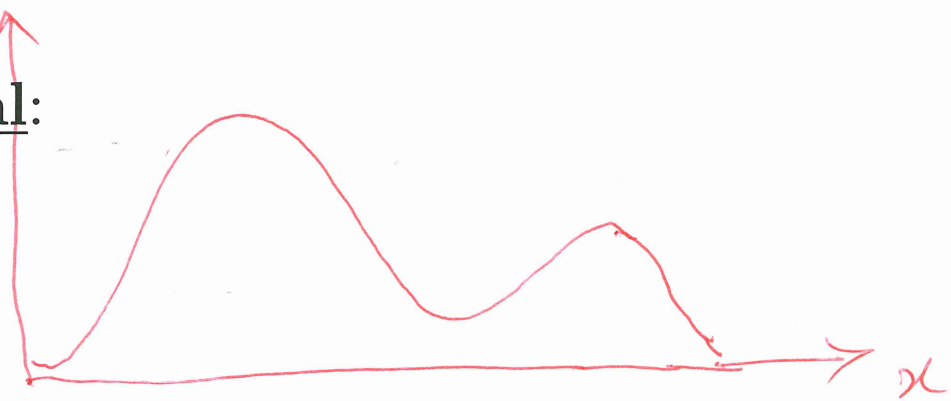
Saw already.

Left-skewed:

Bimodal or multi-modal:

Several bumps in the dist.

(mixture distn.)



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

when the dist. is symmetric

when the distn. is skewed.

Descriptive statistics and what they estimate (continued)

Same issues (consequences) arise here also.

p-quantile of a population: A value q_p such that

$$P(X < q_p) \leq p, \text{ and } P(X > q_p) \leq 1 - p.$$

(*)

If $p = 0.5$, then $q_p = M$.

Essentially X has p probability on the left of q_p.

p-quantile of a sample: A number $\hat{q}_p$ that exceeds *at most* $100p\%$ of the sample (i.e., np observations) and is exceeded by *at most* $100(1-p)\%$ of the sample (i.e., $n(1-p)$ observations).

- $\hat{q}_p$ estimates q_p
- 0.5-quantile $= M$
- **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. $Q_3 = q_{0.75}$
- **Sample quartiles:** Simply divide the data in 4 equal parts using sample counterpart of (*).
- **5-number summary:** of a sample dataset: $(\min, \hat{Q}_1, \hat{Q}_2, \hat{Q}_3, \max)$.

$Q_1, \hat{Q}_2, \hat{Q}_3$