7- descriptive statistics

Statistics

focus on
- data collection
- data analysis
  - visualization
  - estimation of distribution parameter
  - finding correlation
  - testing statements about parameter
  - assessing the reliability of the parameters

population = set of all possible sources of a random variable

parameter = any numerical characteristic of a population

sample = a set of observed sources from the population

statistic = any function of a sample

θ = population parameter

$\hat{\theta}$ = its estimator, obtained from the sample

population → data collection (sampling) → sample

parameters: μ, σ, σ^2..

statistics: $\bar{x}, s, s^2$..

Sampling

errors
1. sampling errors:
   - caused by only a portion of a population is observed
   - sampling errors decrease as the size increase
2. non-sampling errors: caused by wrong statistical techniques

simple random sampling:
- data points are independent from each other
- all data points are equally likely to be sampled

descriptive statistics (tanımlayıcı)

mean = average value of a sample, estimates population mean ($\mu = E(x)$)

$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$ (we expect $\bar{x}$ to converge to μ as sample gets large)

- unbiased: an estimator $\hat{\theta}$ is unbiased for a parameter θ if its expectation equals the parameter, $E(\hat{\theta}) = \theta$ for all possible values of θ → $Bias(\hat{\theta}) = E(\hat{\theta} - \theta)$

  means collecting a large number of samples and computing $\hat{\theta}$ from each of them, on the average we hit the unknown parameter θ exactly
- consistent: if the sampling error converges to 0 as the size increases

  $P(|\hat{\theta} - \theta| > \varepsilon) \to 0$ as $n \to \infty$

- asymptotically normal: by the central limit theorem, the sum of observations, the sample mean have approximately normal distribution if they are computed from a large sample

μ: population mean $\longrightarrow$ $\bar{X}$: sample mean — estimator of μ

σ: population standard deviation $\longrightarrow$ s: sample standard deviation — estimator of σ

σ^2: population variance $\longrightarrow$ s^2: sample variance — estimator of σ^2

sample mean's disadvantage is, its sensitivity to extreme observations (outliers)

median = the central value, sample median ($\hat{M}$) is exceeded by at most a half of observations and it is preceed by at most a half of observations.

population median (M) → $P(X > M) \leq 0.5$
→ $P(X < M) \geq 0.5$

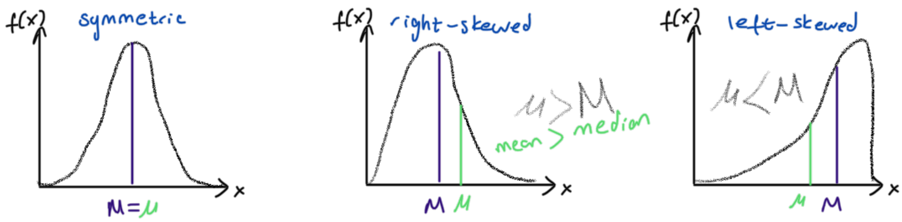
cdf of M is → 0.5

ex; exponential dist:
$F(x) = 1 - e^{-\lambda x}$ (cdf)
$F(M) = 1 - e^{-\lambda M} = 0.5$
$M = \frac{0.6931}{\lambda}$

for discrete distributions ⇒ $F(x) \geq 0.5$, smallest x
in this case median is not unique, often the middle of interval

sample median = sort the samples
- if n is odd, median is the unique middle element
- if n is even, median is any point between the two middle elements

mean vs median =



example: exponential distribution $M \rightarrow F(M) = 1 - e^{-\lambda M} = 0.5 = \frac{0.6931}{\lambda}$

$\mu \rightarrow \frac{1}{\lambda}$ → $M < \mu$ (right skewed)

quantiles = generalizing the notion of a median, we replace 0.5 in its definition by some $0 < p < 1$
(quantity)

p-quantile of a population = $P(X < x) \leq p$
$P(X > x) \leq 1-p$
$\}$ cdf $(x) = p$
$F(x) = p$

Sample p-quantile = any number that exceeds at most $100p\%$ of the sample, and is exceed by at most $100(1-p)\%$ of the sample

percentiles = γ-percentile is 0.01γ quantile

quartiles = first, second, and third quartiles are the 25^{th}, 50^{th}, 75^{th} percentiles
(çeyrek)
they split a population or a sample into four equal parts

median = 0.5 quantile = 50^{th} percentile = 2^{nd} quartile
(M) ($q_{\frac{1}{2}}$) (π_{50}) (Q_2)

q_p: population p-quantile $\longrightarrow$ $\hat{q}_p$: sample p quantile — estimator of q_p

π_γ: population γ-percentile $\longrightarrow$ $\hat{\pi}_\gamma$: sample γ-percentile — estimator of π_γ

Q_1, Q_2, Q_3: population quartiles $\longrightarrow$ $\hat{Q}_1, \hat{Q}_2, \hat{Q}_3$: sample quartiles — estimator of Q_1, Q_2, Q_3

M : population median $\longrightarrow$ $\hat{M}$: sample median — estimator of M

variance $= s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{\sum_{i=1}^{n}(x_i^2 - n\bar{x}^2)}{n-1}$ ($\frac{1}{n-1}$ ensures that s^2 is unbiased)

standard deviation $= s = \sqrt{\text{sample variance}}$

margin of errors $= \frac{s}{\sqrt{N}}$ (confidence interval)

interquartile range $= IQR = Q_3 - Q_1$ (difference between the first and the third quartiles)

- sample mean, variance, and standard deviation are sensitive to outliers. if an extreme observation (an outlier) erroneously appears in data set, it can significantly affect the values of $\bar{x}$ and s^2.
- to detect and identify outliers, we use measures of variability that are not very sensitive to them (which is IQR)

outliers = data that lie below 1.5 IQR below Q_1 and above 1.5 IQR above Q_3

↳ outside of $[\hat{Q}_1 - 1.5(\widehat{IQR}), Q_3 + 1.5(\widehat{IQR})]$