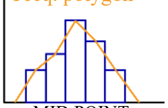
Tabulating and Graphing Numerical Data

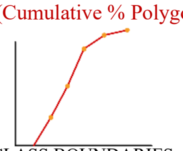
- method
  1) $Range = max - min$
  2) Define $number\ of\ class\ (NOC)$
  3) $Class\ interval\ (I) = \frac{Range}{NOC}$, round up.
     a. Round up.
     b. No remain, +1 instead.
  4) Class boundaries → $[min, min + I), \ldots$
  5) Class midpoint → $\frac{min+(min+I)}{2}, \ldots$
  6) Assign value to the table.
- Table example

Class	Freq.	Percent.
[32, 45)	4	$\frac{4}{16} = 25\%$
⋮	⋮	⋮

Histogram
Freq. polygon

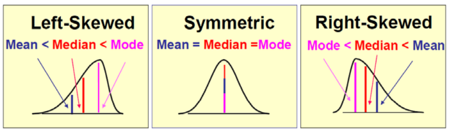

MID POINT

Ogive (Cumulative % Polygon)


CLASS BOUNDARIES

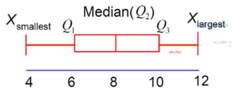
Numerical Descriptive Measure (sample, population)

- Sample size (n).
  Population size (N).
- Central Tendency
  - Mean ($\bar{x}, \mu$)
  - Mode: most common value
  - Median → Index of median: $((n+1)/2)$

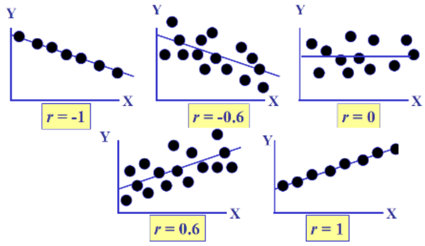


- Quartiles
  - Index of Q_i: $\frac{i}{4}(n+1)$
  - Interquartile range (IR): $Q_3 - Q_1$
- Variance & Standard deviation (S^2, σ^2)
  - Variance = SD^2
  - $S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
    $\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$
  - Coefficient of Variation (CV) $= \frac{S}{\bar{x}} \times 100\%$

- Box n' Whisker plot

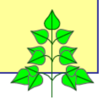


  - $x_{largest} \le Q_3 + IR$
    $x_{smallest} \ge Q_1 - IR$
  - If not, considered as outlier.
- Coefficient of correlation (r)
  - Measure the strength of linear relationship.
  - $r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$



Stem-Leaf Display

2 144677
3 028
4 1



Discrete Probability Distribution (DPD)

- Random variable (X): possible value of a random event
- Probability distribution ($P(X = x)$)
  - A set of probability of a scenario when value of random variable is x
  - Add up to 1.0
- Summary Measures
  - Mean: $\mu = \sum x_j \cdot P(X = x_j)$
    SD: $\sigma^2 = \sum(x_j - \mu)^2 \cdot P(X = x_j) \to \sum x_j^2 P(X = x_j) - \mu^2$
- Considering population

Binomial Distribution

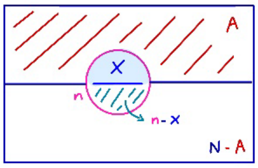
- Condition
  - Have only 2 outcomes: yes/no
  - All trials are INDEPENCENT & IDENTICAL
- Probability of x success.
  - $P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$
  - $n \to number\ of\ trial$
    $x \to number\ of\ success$
    $p \to propability\ of\ success\ (q = 1 - p)$
  - $P(X < 5) = P(X = 4) + P(X = 3) + \cdots + P(X = 0)$
- Mean: $\mu = np$
  SD: $\sigma = \sqrt{np(1-p)}$
- Excel: BIONOM.DIST(x, n, p, cumulative) → cumulative: TRUE/FALSE

Poisson Distribution

- Describe how likely an event is to occur over a time interval.
- Probability of x success.
  - $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$
  - $\lambda \to average\ NOS\ in\ a\ unit\ time$
    if $\lambda = 5\ u/hr \to when\ 2\ hr,\ \lambda = 5 \times 2 = 10\ u/hr$
- Mean: $\mu = \lambda$
  SD: $\sigma = \sqrt{\lambda}$
- Excel: POISSON.DIST(x, λ, cumulative)

Hypergeometric Distribution

- Describe the success of sample when know the success of population.



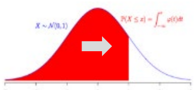
$P(X = x) = \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}}$

$< prop\ of\ x\ in\ n\ when\ A\ in\ N >$

- Mean: $\mu = \frac{nA}{N}$
  SD: $\sigma = \sqrt{\frac{NA(N-A)}{N^2}} \cdot \sqrt{\frac{N-n}{N-1}}$ → Finite population correction factor
- Excel: HYPGEOM.DIST(x, λ, cumulative)

Continuous Probability Distribution

- Calculate Z and use the area under the normal distribution (ND).
  - Density func. of ND: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$
  - $P(X < x) = P(Z_i < Z) \to Z = \frac{x-\mu}{\sigma}$
    Z: Standard normal rando variable
- Make sure the distribution is normal before calculating. (method)
  - Construct chart:
    - Small: Stem-leaf, box & whisker → symetric
    - Large: histogram, polygon → bell-shaped
  - Find summary measure.
    - $mean \approx med \approx mode \mid IR \approx 1.33\sigma \mid Range\ \approx 6\sigma$
  - Observe dataset: 2/3 of data → lie between $\pm 1\sigma$
    4/5 → $\pm 1.28\sigma$, 19/20 → $\pm 2\sigma$
  - Plot (z, x) must be a straight line.
- Excel: NORM.DIST(x, μ, σ, cumulative)
  - Give value count from the right.



- Normal approximation (Binomial Dist.)
  - When $n \to \infty, p \to 0.5, np\ \&\ nq > 5$
  - Mean: $\mu = np$
    Variance: $\sigma^2 = nqp$
    $Z = \frac{x-np}{\sqrt{nqp}}$
- Normal approximation (Poisson Dist.)
  - When $\lambda \to \infty$
  - Mean: $\mu = \lambda$
    Variance: $\sigma^2 = \lambda$
    $Z = \frac{x-\lambda}{\sqrt{\lambda}}$

$*cal\ using\ definition$
$\mu = \int x f(x) dx$
$\sigma^2 = \int (x-\mu)^2 f(x) dx$

- Uniform distribution
  - Each value is equally likely to occur anywhere.
    - Range: between [a, b]
    - graph: horizontal line
  - Density function $f(x) = \begin{cases} \frac{1}{b-a}; a \le x \le b \\ 0; \quad elsewhere \end{cases}$
    → $P(a \le x \le b) = \int_a^b f(x)\,dx$
  - Mean: $\mu = \frac{a+b}{2}$
    Variance: $\sigma^2 = \frac{(b-a)^2}{12}$
    $< cal\ using\ definition >$

Sampling Distribution

- Statistical inference: estimatre charactor of population from info of sample.
- Terminology
  - Parameter: charactor of population (μ, σ, p)
    Sample statistics: charactor of sample ($\bar{x}, s, \bar{p}/p_s/\hat{p}$)
  - Point estimate: numerical value of ($\bar{x}, s, \bar{p}$)
  - Sampling error: absolute difference $|\mu - x|$
- Large sample gives more accurate estimation.

Sampling distribution of mean: Prop-dist. of $\bar{x}$

- Take multiple groups of n data from N and find $\bar{x}$, and plot a distribution of that multiple $\bar{x}$
  - Graphs will look like normal distribution.
  - When $N \to \infty$, graph look like population distribution (PD).
  - When $N \to \infty$ & large n, graph approach the normal distribution. Even when PD is not normal.
- Mean: $\mu_{\bar{x}} = \mu \to \bar{x}$ is unbiased
  SD: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \to \sigma_{\bar{x}} < \sigma$ always
  $Z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$
- Size of n to get normal sample mean distribution.
  - Most dist: $n > 30$ | Symmetric dist: $n > 15$ | Normal dist: n can be any value.

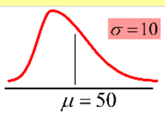
Sampling distribution of sample proportions (p_S)

- population proportions (p): A measure of how common a characteristic is in a population.
  - $p = \frac{x}{N} \to x$ is the number of successes | $p_s = \frac{x}{n} \to x$ success in sample group.
- Normally focus on x with two outcomes (have/not have) → binomial distribution.
- Mean: $\mu_{ps} = p$
  SD: $\sigma_{ps} = \sqrt{\frac{p(1-p)}{n}}$
  $Z = \frac{p_s - \mu_{ps}}{\sigma_{ps}}$

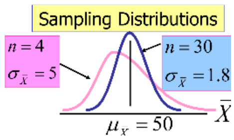
Finite population correction factor (FPC)

- Used when n is very large: $n/N > 0.05$
- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$
  $\sigma_{ps} = \sqrt{\frac{p(1-p)}{n}} \sigma_{ps} \cdot \sqrt{\frac{N-n}{N-1}}$
  FPC

Population Distribution

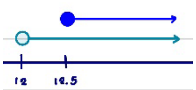


Sampling Distributions



- Continuity correction (Binomial & Poisson)
  - Use when convert discrete dist. to continuous dist.
  - ± 0.5 depends on the situation.

$P(X > 12) \to P(X \ge 12.5)$



$P(X \le 25) \to P(X < 25.5)$