1 Introduction

- population vs. sample, parameter vs. statistic
- numerical data, discrete vs. continuous
- categorical data, ordinal vs. nominal

2 Graphical and Numerical Summaries

- $\bar{X} = \frac{1}{n} \sum X_i$
- $M =$ sorted sample midpoint: $n$ odd $\implies$ at position $\frac{n+1}{2}$, $n$ even $\implies$ average of points $\frac{n}{2}$ and $\frac{n}{2} + 1$
- $Q_1 =$ median of first $\frac{1}{2}$ of data, $Q_3 =$ median of second $\frac{1}{2}$ ($n$ odd $\implies$ include median in each $\frac{1}{2}$)
- $p$th quantile is point with proportion $p$ of data smaller
- $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$
- range = maximum - minimum
- $IQR = Q_3 - Q_1$; outlier $> 1.5 \times IQR$ from $[Q_1, Q_3]$
- dotplot, histogram, boxplot, scatterplot

3 Probability

- probability uses population information to describe samples in long run
- statistics uses sample information to make uncertain claims about population
- random process, outcome, sample space, event, probability
- $P(E) =$ sum of probabilities of outcomes in $E$
- $0 \leq P(E) \leq 1$
- $P(\text{not } E) = 1 - P(E)$
- $A$ and $B$ are independent if occurrence of one doesn’t change $P()$ of other; then $P(A \text{ and } B) = P(A)P(B)$

4 Random Variables and Distributions

- random variable, distribution
- RV represents population, while collection of realizations of RV represents sample

discrete $X$

- values can be put in sequence
- probability mass function $p(x) = P(X = x)$
- mean or expected value $\mu_X = E(X) = \sum x \cdot p(x)$
  - properties: $E(c) = c$, $E(cX) = cE(X)$, $E(X + c) = E(X) + c$, $E(X + Y) = E(X) + E(Y)$
- variance $\sigma_X^2 = E([X - \mu_X]^2) = \sum (x - \mu_X)^2 \cdot p(x)$
  - properties: $VAR(c) = 0$, $VAR(cX) = c^2VAR(X)$, $VAR(X + c) = VAR(X)$, and, for independent $X$ and $Y$, $VAR(X + Y) = VAR(X) + VAR(Y)$
- standard deviation $\sigma_X = \sqrt{\sigma_X^2}$
Bernoulli trials

\[ Y = \begin{cases} 
1, & \text{for success} \\
0, & \text{for failure} 
\end{cases} ; P(Y = 1) = \pi, P(Y = 0) = 1 - \pi \implies \mu_Y = \pi, \sigma^2_Y = \pi(1 - \pi) \]

binomial distribution

- \( X \sim \text{Bin}(n, \pi) \) is \#successes in \( n \) independent Bernoulli trials, each with \( P(\text{success}) = \pi \)
- \( \binom{n}{x} = \frac{n!}{x!(n-x)!} \), where \( 0! = 1 \) and \( n! = 1 \times 2 \times 3 \times ... \times n \)
- \( P(X = x) = \left( \binom{n}{x} \right) \pi^x (1 - \pi)^{n-x} \) for \( x = 0, 1, \ldots, n \)
- \( \mu_X = n\pi, \sigma^2_X = n\pi(1 - \pi), \sigma_X = \sqrt{n\pi(1 - \pi)} \)

continuous \( X \)

- values fill interval
- \( P(a \leq X \leq b) = \text{area under } f(x) \) between \( a \) and \( b \) \( (\text{area between } -\infty \text{ and } \infty \) is 1 \)
- cumulative distribution function \( F(x) = P(X \leq x) \)

normal distributions

- in curve \( f(x) \) for \( N(\mu, \sigma^2) \), \( \mu \) is at center and \( \sigma \) is distance from center to curvature change
- \( X \sim N(\mu, \sigma^2) \implies Z = \frac{X - \mu}{\sigma} \sim N(0, 1^2) \)
- \( Z \sim N(0, 1^2) \implies X = Z\sigma + \mu \sim N(\mu, \sigma^2) \)
- \( P(X < x) = P \left( \left| Z = \frac{X - \mu}{\sigma} \right| < \frac{x - \mu}{\sigma} \right) \)
- \( P(Z < [z = a.b]) \) is in row \( \text{a.b} \) and column \( .0c \) of \( N(0, 1) \) table
- \( X \sim N(\mu, \sigma^2) \implies P(|X - \mu| < 2\sigma) \approx 95\% \)
- \( X \sim N(\mu, \sigma^2) \implies P(|X - \mu| < 3\sigma) \approx 99.7\% \)

5 Estimation (through Student’s \( t \) confidence interval)

- simple random sample
- \( X_1, \ldots, X_n \) are IID from population with \( \mu \) and \( \sigma^2 \implies E(\bar{X}) = \mu \) and \( \text{VAR}(\bar{X}) = \frac{\sigma^2}{n} \)
- standard error of \( \bar{X} \) is its estimated standard deviation, \( S/\sqrt{n} \)
- in normal probability (or QQ) plot, points \((\approx)\) lined up leaves normal population plausible
- normal population implies normal sample mean: \( X_1, \ldots, X_n \sim N(\mu, \sigma^2) \implies \bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \)
- CLT: large enough sample (rule of thumb: \( n > 30 \)) implies \((\approx)\) normal sample mean: \( X_1, \ldots, X_n \) a large SRS from (almost) any population with \( \mu \) and \( \sigma^2 \implies \bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \) \((\approx)\)
- \( z_{\alpha/2} \) cuts off right tail area \( \alpha/2 \) from \( N(0, 1^2) \)
- \( \bar{X} \pm \frac{z_{\alpha/2}}{\sqrt{n}} \frac{\sigma}{\sqrt{n}} \) contains \( \mu \) for a proportion \( 1 - \alpha \) of SRSs \( X_1, \ldots, X_n \) from population with unknown \( \mu \) and known \( \sigma \), provided \( n \) is large enough or population is normal
  
  sample size \( n = \left( \frac{z_{\alpha/2}^2 \sigma}{m} \right)^2 \) suffices to give error margin \( m \) (if \( \sigma \) unknown, use \( \sigma \approx s \))
- for normal \( X_1, \ldots, X_n, T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \)
- \( t_{n-1, \alpha} \) cuts off right tail area \( \alpha \) from \( t_{n-1} \)
- \( \bar{X} \pm \frac{t_{n-1, \alpha/2}}{\sqrt{n}} S \) contains \( \mu \) for a proportion \( 1 - \alpha \) of SRSs \( X_1, \ldots, X_n \) from population with unknown \( \mu \) and unknown \( \sigma \), provided \( n \) is large enough or population is normal