1 Introduction

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

2 Graphical and Numerical Summaries

- $\bar{X} = \frac{1}{n} \sum X_i$ (sample mean)
- $M =$ 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 $\frac{1}{2}$)
- pth quantile is point with proportion p of data smaller
- $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$ (sample standard deviation)
- 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 processs, 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 $\iff P(A|B) = P(A)$ and $P(B|A) = P(B) \iff P(A \text{ and } B) = P(A)P(B)$
- $P(A|B) = \dfrac{P(A \text{ and } B)}{P(B)}$

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 $p(x) = P(X = x)$
- (population) mean or expected value $\mu_X = E(X) = \sum_x x \cdot p(x)$
  properties: $E(c) = c$, $E(cX) = cE(X)$, $E(X + c) = E(X) + c$, $E(X + Y) = E(X) + E(Y)$
- (population) variance $\sigma_X^2 = E\left([X - \mu_X]^2\right) = \sum_x (x - \mu_X)^2 \cdot p(x)$
  properties: $VAR(c) = 0$, $VAR(cX) = c^2 VAR(X)$, $VAR(X + c) = VAR(X)$, and,
  for independent X and Y, $VAR(X + Y) = VAR(X) + VAR(Y)$
- (population) standard deviation $\sigma_X = \sqrt{\sigma_X^2}$

Bernoulli trials

$$Y = \begin{cases} 1, \text{ for success} \\ 0, \text{ for failure} \end{cases};\ P(Y=1)=\pi, P(Y=0)=1-\pi \implies \mu_Y=\pi, \sigma_Y^2=\pi(1-\pi)$$

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 $n! = 1\times 2\times 3\times ... \times n$
- $P(X=x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$ for $x=0,1,\ldots,n$
- $\mu_X = n\pi, \sigma_X^2 = n\pi(1-\pi), \sigma_X = \sqrt{n\pi(1-\pi)}$

continuous X

- values fill interval
- $P(a \le X \le b) =$ area under $f(x)$ between a and b (area between $-\infty$ and ∞ is 1)
- cumulative distribution function $F(x) = P(X \le x)$

normal distributions

- in curve $f(x)$ for $N(\mu,\sigma^2)$, μ is at center and σ is distance from center to curvature change
- $X \sim N(\mu,\sigma^2) \implies Z = \frac{X-\mu}{\sigma} \sim N(0,1^2)$
- $Z \sim N(0,1^2) \implies X = Z\sigma + \mu \sim N(\mu,\sigma^2)$
- $P(X < x) = P\left(\left[Z = \frac{X-\mu}{\sigma}\right] < \frac{x-\mu}{\sigma}\right)$
- $P\left(Z < [z = a.bc]\right)$ is in row $a.b$ and column $.0c$ of $N(0,1^2)$ table
- $X \sim N(\mu,\sigma^2) \implies P(|X-\mu| < \begin{array}{c}1\\2\\3\end{array}\ \sigma) \approx \begin{array}{l}68\\95\\99.7\end{array}\ \%$

5 Estimation

- simple random sample
- $X_1,\ldots,X_n$ are IID from population with μ and $\sigma^2 \implies E(\bar{X}) = \mu$ and $VAR(\bar{X}) = \frac{\sigma^2}{n}$
- in normal probability (or QQ) plot, points ($\approx$) lined up leaves normal population plausible
- normal population implies normal sample mean: $X_1,\cdots,X_n \sim N(\mu,\sigma^2) \implies \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- CLT: large enough sample (rule of thumb: $n > 30$) implies ($\approx$) normal sample mean: $X_1,\cdots,X_n$ a large SRS from (almost) any population with μ 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 $N(0,1^2)$
- $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ contains μ for a proportion $1-\alpha$ of SRSs $X_1,\ldots,X_n$ from population with unknown μ and known σ, provided n is large enough or population is normal

  sample size $n = \left(\frac{z_{\alpha/2}\sigma}{m}\right)^2$ suffices to give error margin m (if σ unknown, use $\sigma \approx s$)