1 Introduction

- population vs. sample, parameter vs. statistic
- mumerical 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 $\frac{1}{2}$)
- pth quantile is point with proportion p of data smaller
- $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- range = maximum − minimum
- $IQR = Q_3 - Q_1$; outlier $> 1.5 \times IQR$ from $[Q_1, Q_3]$
- ~~dotplot~~, histogram, ~~boxplot~~, ~~scatterplot~~ (only need to know histogram)

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 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 $p(x) = P(X = x)$ → Pmf
- mean or expected value $\mu_X = E(X) = \sum_x x \cdot p(x)$
  properties: $E(c) = c$, $E(cX) = cE(X)$, $E(X + c) = E(X) + c$, $E(X + Y) = E(X) + E(Y)$
- 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, $VAR(X + Y) = VAR(X) + VAR(Y)$
- 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

pdf

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

5 Estimation ~~[illegible]~~ through MSE

- simple random sample
- $X_1, \ldots, X_n$ are IID from population with μ and $\sigma^2 \implies E(\bar{X}) = \mu$ and $VAR(\bar{X}) = \frac{\sigma^2}{n}$
- standard error of $\bar{X}$ is its estimated standard deviation, $S/\sqrt{n}$
- ~~in normal probability (or QQ) plot, points (≈) lined up leaves normal population plausible~~
- normal population implies normal sample mean: $X_1, \cdots, X_n \sim N(\mu, \sigma^2) \implies \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

1. (20 points) Suppose NBA player weights (in pounds) are $N(221, 15^2)$.

   (a) Find the weight such that 20% of players weigh less than that weight.

Let W = weight of a randomly choosen player

x = the desired 20th percentile weight,

then $P(W < x) = 0.2$

$\Rightarrow P\left(Z < \frac{x-221}{15}\right) = 0.2$

$\Rightarrow \frac{x-221}{15} = -0.845 \Rightarrow x = 208.325$

   (b) A random group of 5 NBA players (still with weights from $N(221, 15^2)$) cross a playground bridge together, even though its breaking strength is only 1000 pounds. What is the probability that it breaks?

Let S = Sum of their weights $= 5\bar{W}$ (Since $\bar{W} = \frac{1}{5}\sum_{i=1}^{5} W_i$). each $W_i \sim N(221, 15^2)$,

So $\bar{W} \sim N(221, \frac{15^2}{5})$ and $S = 5\bar{W} \sim N(5 \times 221, 5^2 \times \frac{15^2}{5}) = N(1105, 33.54^2)$

So $P(S > 1000) = P\left(Z > \frac{1000-1105}{33.54}\right)$

$= P(Z > -3.13) = 1 - P(Z < -3.13)$

$= 1 - 0.0009 = 0.999$

2. (10 points) Here are several questions about summary statistics.

(a) Consider these summary statistics:

- IQR = interquartile range
- M = sample median
- Q_1 = first quartile
- S = sample standard deviation
- $\bar{X}$ = sample mean

Which of them is least affected by an outlier? (Circle one.)

i. IQR, M, and Q_1
ii. IQR and S
iii. M and $\bar{X}$
iv. S and $\bar{X}$
v. None of the above

(a) Not required

(b) R was used to get summary statistics on data on the average commute time (in minutes) for each of 51 states (or, rather, 50 states and the District of Columbia):

```
commute = c(15.2, 15.4, 16.5, 16.9, 17.5, 17.5, 18.1, 18.9, 19.1, 19.4, 19.5, 19.7,
  19.9, 20.3, 20.4, 21, 21.2, 21.6, 21.7, 21.8, 21.8, 22.1, 22.1, 22.5, 22.6,
  22.7, 22.7, 22.9, 23, 23.2, 23.3, 23.3, 23.4, 23.4, 23.6, 23.7, 23.8, 24.5,
  24.6, 24.7, 24.8, 24.8, 25.8, 26, 26.1, 26.5, 27, 28.4, 28.5, 30.2, 30.4)
> summary(commute)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.20   20.10   22.70   22.43   24.55   30.40
>
```

What is the interquartile range (IQR) of the commute times? (Hint: You do not need to use the raw data.)

$IQR = Q_3 - Q_1 = 24.55 - 20.10 = 4.45$

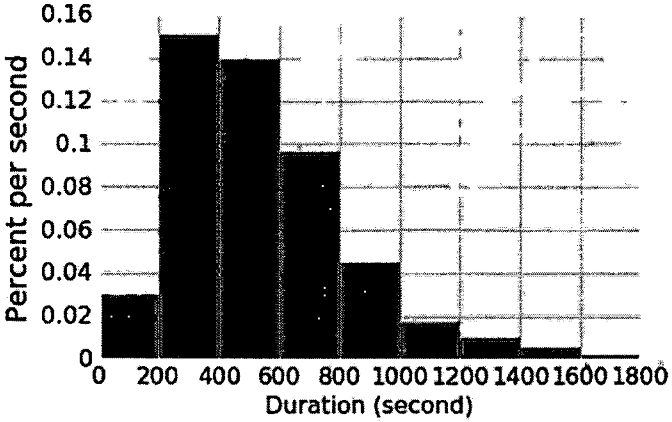
(c) Find the third quartile of this set of spruce log lengths (in feet):
8.7 9.2 8.7 8.0 8.5 10.1 7.5 7.8 8.8 8.0

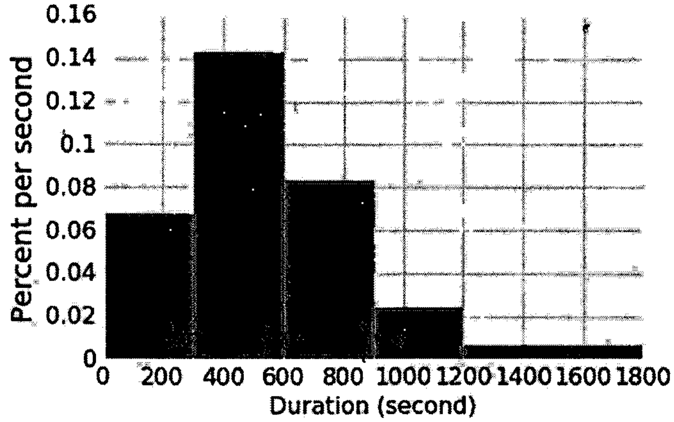
Sort: 7.5, 7.8, 8.0, 8.0, 8.5, 8.7, 8.7, 8.8, 9.2, 10.1

So $Q_3 = 8.8$

3. (11 points) Distributions

The two histograms of bike trip durations below were both generated by `trip.hist(...)` using different bins.





(a) **(8 pt)** Write the proportion of trips that fall into each range of durations below. *Show your work.* If it is not possible to tell from the histograms, instead write **Not enough information**.

- Between 200 (inclusive) and 400 (exclusive) seconds

  $0.0015 \times 200 = 0.3$

- Between 300 (inclusive) and 900 (exclusive) seconds

  $0.0014 \times 300 + 0.0008 \times 300 = 0.66$

- Between 400 (inclusive) and 900 (exclusive) seconds

  $0.0014 \times 200 + 0.0008 \times 300 = 0.52$

- Between 200 (inclusive) and 300 (exclusive) seconds

  $0.0007 \times 300 - 0.0003 \times 200 = 0.15$

(b) **(3 pt)** A study followed 369 people with cardiovascular disease, randomly selected from hospital patients. A year later, those who owned a dog were four times more likely to be alive than those who didn't.

- Circle *True* or *False*: This study is a randomized controlled experiment.

  (b) not required

- Circle *True* or *False*: This study shows that dog owners live longer than cat owners on average.

- Circle *True* or *False*: This study shows that for someone with cardiovascular disease, adopting a dog will probably cause them to live longer.

4. (10 points) Suppose each ticket in a lottery has a $\frac{1}{8}$ chance of being a winner. What is the probability of having exactly 4 winners in a randomly selected group of 10 tickets?

Let X = Number of winners in the 10 tickets, then $X \sim Bin(n=10, \pi=\frac{1}{8})$,

so $P(X=4) = \binom{10}{4}(\frac{1}{8})^4(1-\frac{1}{8})^{10-4} = 0.0230$

5. (20 points) A class of students took a quiz whose score distribution is in the table below.

score	1	2	3	4
proportion of class with score	0	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{2}$

(a) What is the mean score?

$\sum_x x \cdot P(x) = 1 \times 0 + 2 \times \frac{1}{8} + 3 \times \frac{3}{8} + 4 \times \frac{1}{2} = 3.375$

(b) What is the variance of the scores?

$\sum_x (x - \mu_X)^2 P(x) =$

$(1-3.375)^2 \times 0 + (2-3.375)^2 \times \frac{1}{8} + (3-3.375)^2 \times \frac{3}{8} +$

$(4-3.375)^2 \times \frac{1}{2} = 0.484375$

6. (20 points) A grandmother will put a pile of money into two uncertain investments:

- a stock whose return has a mean of 6% and a standard deviation of 1%
- a bond whose return has a mean of 4% and a standard deviation of 0.5%

Suppose she puts half her money in the stock and half in the bond. Let R_s = the return on the stock and R_b = the return on the bond (each as a percentage). Then R = her return $= \frac{1}{2}R_s + \frac{1}{2}R_b = \frac{1}{2}(R_s + R_b)$.

(a) What is the expected value of her return?

$$E(R) = E\left(\frac{1}{2}R_S + \frac{1}{2}R_b\right) =$$

$$\frac{1}{2}E(R_S) + \frac{1}{2}E(R_b) =$$

$$\frac{1}{2} \times 6\% + \frac{1}{2} \times 4\% = 5\%$$

(b) What is the standard deviation of her return?

$$VAR(R) = VaR\left(\frac{1}{2}R_S + \frac{1}{2}R_b\right)$$

$$= \frac{1}{4}VAR(R_S) + \frac{1}{4}VAR(R_b)$$

$$= \frac{1}{4} \times (1\%)^2 + \frac{1}{4} \times (0.5\%)^2$$

$$= 0.3125$$

$$\Rightarrow \quad SD(R) = 0.559$$