Review

- population
- sample
- parameters (e.g. $\mu, \sigma, p$)
- statistics (e.g. $\bar{x}, s, \hat{p}$)

Types of data

Qualitative
- Nominal
- Ordinal

Quantitative
- Discrete
- Continuous

Sampling: We want representative samples (not biased).

→ Ideally: simple random sample.
Stem-and-leaf plot:

0  | 1 1 2 7
1  | 0 1 1 2 4 7 9 ← shows the distribution
2  | 0 0 2 2 3 4 4 5 of the data
3  | 1 3 7
4  | 0 6

Histogram

→ For qualitative data:
  • Pareto chart
  • Pie chart
Measures of center

- Mean $\bar{x} = \frac{\sum x_i}{n}$: usually best estimator of $\mu$
- Median: 50% of data is below, 50% of data is above
- Mode: most popular observation

Measures of variation

- Range: max-min
- Standard deviation
  - $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$ (population)
  - $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$ (sample)
Empirical Rule:
- 68% of data is within 1 sd from mean
- 95% of data is within 2 sd from mean

Standardized score: $z = \frac{x - \mu}{\sigma} \\
\rightarrow$ If $|z| > 2$, then the observation is unusual.

Quartiles:
$Q_1$: 25% of the data is $< Q_1$
$Q_3$: 75% of the data is $< Q_3$

Boxplot
Probability

* Simple events: Ambidextrous (A), Hyperactive (H)
* Sample space: set of all possible simple events
  \[ \{AH, A\bar{H}, \bar{A}H, \bar{A}\bar{H}\} \]
* Event: set of simple events
  \[ \rightarrow \text{ If each simple event is equally likely, each has probability } \frac{1}{n}. \]

Rules of Probability

- \[ 0 \leq P(A) \leq 1 \]
- \[ P(\text{not } A) = 1 - P(A) \]
★ Addition Rule:

\[
P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)
\]

or both!

\[
P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0
\]

if A and B are mutually exclusive

e.g. \( P(\text{Ambidextrous or Hyperactive}) = P(A) + P(H) - P(A \text{ and } H) \)

★ Multiplication Rule:

\[
P(A \text{ and } B) = P(A|B)P(B)
\]

\[
P(A \text{ and } B) = P(A)P(B) \text{ if } A \text{ & } B \text{ are independent}
\]

e.g. \( P(\text{Ambidextrous and Hyperactive}) = P(A|H)P(H) \)

★ Conditional probability:

\[
P(A|B) = \frac{P(A \text{ and } B)}{P(B)}
\]

e.g. \( P(\text{Ambidextrous} \mid \text{Hyperactive}) = \frac{P(A \text{ and } H)}{P(H)} \)
**Question:** A friend who works in a big city owns two cars, one small and one large. Three-quarters of the time he drives the small car to work, and one-quarter of the time he drives the large car. If he takes the small car, he usually has little trouble parking, and so is at work on time with probability 0.8. If he takes the large car, he is at work on time with probability 0.6. Given that he was on time on a particular morning, what is the probability that he drove the small car?
Random variable

→ Probability distribution

★ Binomial: $n$ trials, $k$ successes, $p$ probability of success

$$P(k \text{ out of } n) = \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k}$$

→ Probability distributions have theoretical mean and sd:

★ For Binomial: $\mu = np$ and $\sigma = \sqrt{np(1 - p)}$
★ For Poisson: $P(k) = \frac{\mu^k \cdot e^{-\mu}}{k!}$, $k$ is the number of occurrences.
  mean is $\mu$ and sd is $\sigma = \sqrt{\mu}$. 
**Question** Suppose you play a daily lottery where you have a 1 in 10 chance of winning something.

a) If you play for 10 days in a row, what is the probability you win at least once?

b) What is the expected number of times you win?

**Question** Suppose the number of fleas on a stray dog follows a Poisson distribution with mean 28. Would it be unusual to find a dog with 38 fleas?
Normal: bell-shaped distribution with $y = \frac{1}{2\pi} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Standard normal: $\mu = 0$ and $\sigma = 1$

We convert to standard normal: $z = \frac{x-\mu}{\sigma}$

If we know the shaped area we can find this $z$-value for the table.
**Central Limit Theorem (CLT):** Sample mean $\bar{x}$ is normally distributed with

- mean $\mu_{\bar{x}} = \mu$ and
- sd $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

→ Good if population is normally distributed or $n$ is "large", i.e. $n \geq 30$.

**Normal approximation to Binomial**

- Requires at least 5 successes ($np \geq 5$) and 5 failures ($np(1 - p) \geq 5$)
- $\mu = np$ and $\sigma = \sqrt{np(1 - p)}$
- Continuity correction:
  Binomial: $P(X > 10) \rightarrow$ Normal approx: $P(X > 10.5)$
* * Question: SAT math scores are normally distributed with a mean of 512 points and a standard deviation of 112 points.

a) If a random sample of 12 people take the test, what is the probability their average score will be under 500?

b) Suppose a college wants to admit only those applicants in the top 10%. What cut-off should they use?
Confidence Intervals (CIs)

- A \((1 - \alpha)\) CI for the \textit{population mean} \(\mu\) is given
  - when \(\sigma\) is known by: \(\bar{x} \pm E = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\)
  - when \(\sigma\) is unknown by: \(\bar{x} \pm E = \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}\)

→ Interpretation!!!

→ Sample size for fixed \(E\): \(n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2\)

- A \((1 - \alpha)\) CI for the \textit{population proportion} \(p\) is given by
  \(\hat{p} \pm E = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\)

→ Sample size for fixed \(E\): \(n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1 - \hat{p})\)
  
  If \(\hat{p}\) is \textbf{not known} in advance: \(n = \left(\frac{z_{\alpha/2}}{E}\right)^2 (0.5)^2\)
Question: A study of perception tested 80 randomly selected men for color blindness and found 7 with red/green color blindness.

a) Construct a 95% confidence interval for the population proportion of men with red/green color blindness.

b) Women have a 0.25% rate of red/green color blindness. Can we safely conclude that women have a lower rate of red/green color blindness than men?