AP Statistics Unit 2 Quick Notes

1. Two Categorical Variables

- **Two-way table**: joint + marginal distributions.
- **Conditional probability:** $P(A|B) = \frac{\#(A \cap B)}{\#B}$.
- **Relative bar chart**: compares proportions across groups.
- **Mosaic plot**: Width = proportion in x category; Height = proportion in y category.
- **Independence**: Conditional distributions are the same across groups.

2. Two Quantitative Variables

Scatterplots: Describe *form* (linear/nonlinear), *direction* (positive/negative), *strength* (strong/weak).

- **Outlier**: Far from trend in y.
- **High-leverage**: Extreme x value.
- **Influential**: Greatly changes slope/intercept if removed.

Simpson's Paradox: Trend reverses when groups combined due to lurking variable.

3. Covariance & Correlation

$$\text{Cov}(X,Y) = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y}), \quad r = \frac{\text{Cov}(X,Y)}{s_x s_y}$$

- $-1 \leq r \leq 1$, $|r| \approx 1$: strong, ≈ 0: weak.
- No units, not affected by scaling, but affected by outliers.
- **Caution:** Correlation $\neq$ causation.

4. Linear Regression Model

$$\hat{y} = a + bx, \quad b = r\frac{s_y}{s_x}, \quad a = \bar{y} - b\bar{x}$$

- **Slope** b: Average change in $\hat{y}$ per unit x.
- **Intercept** a: Predicted $\hat{y}$ when $x = 0$ (may be meaningless).

Residuals: $e_i = y_i - \hat{y}_i$
Mean residual = 0; Positive $\Rightarrow$ underestimation; Negative $\Rightarrow$ overestimation. LSRL minimizes $\sum e_i^2$.

5. Coefficient of Determination & Standard Error

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}, \quad s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

r^2: Proportion of variation in y explained by x. s: Typical prediction error in y-units.

6. Regression Conditions (LINER)

- L: Linearity — Scatter/residual plots show no curve.
- I: Independence — From study design.
- N: Normal residuals — Histogram/Normal plot.
- E: Equal variance — Residual spread constant.
- R: Randomness — From sampling/assignment.

7. Warnings

- **Extrapolation:** Avoid beyond observed x-range.
- **Transformations:** log, $\sqrt{\ }$, reciprocal for curvature or spread issues.

8. AP-Style Reminders

- Correlation and slope have same sign.
- $r^2 \geq 0$ always.
- High r^2 does not prove causation.
- Residual plot: randomness = good; patterns = bad.
- Categorical data relationships: segmented bar, relative bar, mosaic plots.