Describing data:

·Parameter vs. statistic – numerical summary of the population (denote with Greek letters: $\alpha, \beta, \mu, \sigma$) **vs.** numerical summary of sample (our guess of population) (denote with lower case letters x, s, r, bars $\bar{x}$, hats $\hat{y}, \hat{\beta}$). We use sample statistics to estimate population parameters.
·Data – continuous, categorical, binary, etc.
·Observations – $y_1, y_2, y_3, \ldots y_n$ (or row in the table)
·Population – The entire group
·Sample – A subset of the population
·Variable – X (input) independent, Y (output) dependent.
·Qualitative/quantitative data – categorical (could be discrete), non-numerical, characteristic / Numerical: mean, median, var, sd
·Nominal (unordered)/ordinal(ordered) – (both Binary/Not Binary) colour, sex (mode)/social class: low, middle, high income, polit.ideology (mode, median)
·Granularity (continuous/discrete) – (infinity/finite)
·Skew – describes if such distribution is asymmetrical about its mean; Right Skew (mean>median), Left Skew (mean<median), Symmetric Skew (mean ≈ median)
·Outliers- observations that are far away from most observations, can affect the mean, sd

Measures of Central Tendency:

·Mean $\bar{y} = \frac{y_1+y_2+\cdots+y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$; affected by outliers.
·Median - n odd: middle value of data;
n even: average of two middle value; not affected by outliers
·Mode – value that appears more often in a set of data values.

Measures of dispersion (variability):

·Variance $var = S^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$, difference between observed value and the mean $(y_i - \bar{y})$
·Standard deviation $S = \sqrt{var} = \sqrt{S^2}$, average distance data from mean
·Standard error (SE) $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$, where σ - sd of population; describes how much $\bar{y}$ varies from sample to sample

Distributions and probability:

·Normal Distribution $N(\mu, \sigma^2)$ mean is the center, sd controls the spread of the data. Bell-shaped
·T Distribution – centred at zero, bell-shaped like normal, fatter tails than normal, degrees of freedom define "fatness", as df↑ it gets close to Standard Normal, sd is a bit large than 1.
·Degrees of Freedom is the number of values in the final calculation of a statistic that are free to vary; df = n-1 (size -1)
·Standard deviation(sd) – one standard deviation will be equal to something in original units (10 crimes) s ≥ 0 and s=0 if Y is a constant. The greater the variability about the mean, the larger the value of s.
·Variance – measures the average squared deviation of the observation from the mean.
·Standard Normal Distribution N=(0, 1)
·Sampling distribution – distribution of a statistic given repeated sampling (probability distribution of sample statistics, such as sample mean or sample proportion – tell me how often (or likely) each possible value of the statistic is), C.L.T important. $\bar{y} = \frac{1}{n}\sum y_i$
·Central Limit Theorem (C.L.T.) $for\ n \to \infty,\ \bar{y} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
·Standard error (SE) $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$, as n→ ∞, SE gets smaller
·Sampling distribution of a sample mean $\bar{y}$ is approximately normal distribution; "If we repeatedly took samples, then in the long run, the mean of the sample means $\bar{y}$ would equal the population mean μ"
·Calculate a t-score (if n<30) to find p-value $t^* = \frac{\bar{y}-\mu_0}{\hat{\sigma}_{\bar{y}}}$, df = (n-1)
·Calculate a Z-score (if n≥30) to find p-value $Z = \frac{\bar{y}-\mu_0}{\hat{\sigma}_{\bar{y}}}$, where $\hat{\sigma}_{\bar{y}} = \frac{S}{\sqrt{n}}$

Remember $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ for population, $\hat{\sigma}_{\bar{y}} = \frac{S}{\sqrt{n}}$ for sample

Estimation

·Point estimate – $mean = \hat{\mu}, sd = \hat{\sigma}$ – best estimated by $\bar{y}$ and S - a sample statistic that gives a good guess about population parameter.
·Confidence interval – is a range of numbers within which a population parameter is believed to fall (Point Estimate ± Margin of Error): 1.find $\bar{y}$, 2.find S, 3.find $\hat{\sigma}_{\bar{y}} = \frac{S}{\sqrt{n}}$, 4. find critical value (z or t) = (1-Confidence Coefficient)/2=((1 − α)/2 = (1 − 0.90)/2 = 0.05); 5.calculate $\bar{y} \pm Z \times \hat{\sigma}_{\bar{y}}$ or $\bar{y} \pm T \times \hat{\sigma}_{\bar{y}}$
·Confidence level – the probability that this method produces an interval that contains the parameter (0.90, 0.95, 0.99)
Critical value for big n: 90% $\pm 1.64\sigma_{\bar{y}}$; 95% $\pm 1.96\sigma_{\bar{y}}$; 99% $\pm 2.57\sigma_{\bar{y}}$
·Bias – we want our estimator to be unbiased (accurate) with repeated sampling, $E(\hat{\mu}) = \mu$
·Efficiency – we want our estimators to be precise, $\hat{\sigma}$ is smaller.

Hypothesis testing (5 steps)

1.Assumptions, 2. Hypotheses, 3.Test statistics, 4.P-value, 5.Conclusion("There isn't enough evidence to reject/not reject…")
·Hypothesis – a statement about characteristics of variable
·Null/alternative hypotheses – H_0 and H_a
$H_0: \mu = \mu_0$, and $H_a: \mu > \mu_0$, $H_a: \mu < \mu_0$, $H_a: \mu \neq \mu_0$,
one-sided test (>, <, ≥, ≤) or two-sided (=, ≠) We use both tails because we want to find the probability of error in both directions.
Adopting a two-sided approach effectively splits your α evenly into two tails. A one-sided alternative assigns α entirely into one of two tails.
Test statistic $TS = \frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}}$, (comes from sampling distribution).
·P-value - is the probability that the test statistic equals the observed value or a value even more extreme in the direction predicted by Ha.
$p \leq \alpha - reject\ H_0$ (result is "statistically significant")
$p > \alpha - cannot\ reject\ H_0$

$$p = 2 \times Pr\left(z \geq \left|\frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}}\right|\right) \quad or \quad p = 2 \times Pr\left(t \geq \left|\frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}}\right|\right)$$

·Significance level (α-level) is a number such that we reject H_0 if p-value is less than or equal to it (0.05, 0.01)
·Type I error–The probability of rejecting a true H_0
·Type II error-The probability of failing to reject a false H_0

Chi-square test of independence (H_0: variables are independent)

$F_{observed}$, $f_{expected} = \frac{Row\ total}{Grand\ total} \times Column\ total$
$\chi^2 = \sum \frac{(f_o-f_e)^2}{f_e}$, df =(row-1)(columns-1) ; when H_0 is true, χ^2 is small
Stand. residuals $z = \frac{f_o-f_e}{\sqrt{f_e(1-\frac{Row\ total}{Grand\ Total})(1-\frac{Column\ total}{Grand\ total})}}$

Regression

·Linear model $E(y) = Y = \alpha + \beta X$
·Linear Model $Y_i = \beta_0 + \beta_1 X + \varepsilon_i\ with\ \varepsilon_i \sim N(0, \sigma^2)$
·Interpret y-intercept= α (value of Y when X=0) and slope=β (on average, 1 unit↑ in X is associated with β unit↑ in Y)
·Least squares line (LS)- the prediction line $\hat{y} = a + bx$, because its one with the smallest RSS.
·Prediction equation y = a + bx + e, also (?) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
·Sum of Squared Errors $SSE = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$
OLS estimators: $\hat{\beta} = \frac{\sum((X_i-\bar{X})(Y_i-\bar{Y}))}{\sum(X_i-\bar{X})^2}$, $\hat{\alpha} = \hat{\beta}_0 = \bar{Y} - \hat{\beta}\bar{X}$, SSE - the squared distance from the Y values to the fitted line; want to min. this
·Residual sum of squares RSS=SSE = $\sum(y - \hat{y})^2$
·Regression Sum of Squares RegSS = $\sum(\hat{y} - \bar{y})^2$
·Total Sum of Squares $TSS = RegSS + RSS = \sum(y - \bar{y})^2$
·Interpret a scatterplot - form/pattern, direction, strength, outliers
·Construct a CI around β: $\hat{\beta}_1 \pm t_{\alpha/2} se_{\hat{\beta}_1}$, $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
·Hypothesis test for β, where H_0: β = 0, $t = \frac{\hat{\beta}_0 - 0}{se_{\hat{\beta}_0}}$, df = n-2
·Calculate standard error (SE) for β=SE(β_1)
$se_{\hat{\beta}_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}$ and $se_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum(x_i-\bar{x})^2}}$
·Calculate standard deviation $S_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$ and n $S_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$
·Correlation coefficient r - strength and direction of linear relationship between X and Y; -1 ≤ r ≤ 1, sensitive to outliers, unit-free, only linear, not casual.
$r = R = \frac{S_{xy}}{S_x S_y} = \frac{1}{n-1}\sum\frac{(x-\bar{x})(y-\bar{y})}{S_x S_y}$; where $S_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$
Pearson Correlation (assumed that X and Y normally distributed)
$r_{xy} = \frac{covariance_{xy}}{SD_x SD_y} = \frac{S_{xy}}{S_x S_y} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)}\sqrt{(\sum y_i^2 - n\bar{y}^2)}}$
·Square of Correlation coeff $r^2/R^2 = (\frac{S_{xy}}{S_x S_y})^2 = \frac{explained\ variability}{total\ variability} = \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS}$; interpreted as % of variability in y explained by x; $0 \leq R^2 \leq 1$; evaluate the strength of the fit of linear model; R^2 near 1 suggest a good fit to the data, if R^2=1, all points fall exactly on the line
·Interpret a regression analysis table: Intercept, slope, standard error, p-value. Focus is on the slope.
Assumption:
Type of data (quantitative/continuous, categorical, etc), Sample size (small, large), Population distribution (normally), Sampling method (random)
Assumption for linear regression:
Randomized data generation, independent observation, size (small, large), there is **no** linear relationship, there is error normally distributed $\epsilon_i \sim N(o, \sigma^2)$
Estimation: There is a positive/negative correlation; there is/isn't linear relationship.