Definitions

A **population** (big N) is the entire collection of objects or individuals about which information is desired. A **sample** (little n) is a subset that is being studied. We assume in this class that the sample is $< 10\%$ of population, and that the population is large.
Descriptive statistics consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs. A **statistic** is a numerical summary based on a sample.
Inferential statistics uses methods that take results from a sample, extends them to the population, and measures the reliability of the result. This is how predictions are made.
Bias in sampling: Voluntary response, Convenience, Selection, Nonresponse, Response, Wording-Deliberate, Wording-unintentional

Descriptive Statistics

	Sample statistic	Population Parameter
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Correlation	r	ρ
Noise	e_i	ϵ_i
	Guess	True, but unknown

The Mean, Variance and StdDev are subject to outliers.

$$\begin{aligned}
\bar{x} =& \frac{1}{n}\sum_{i=1}^{n} x_i \\
\mu = E(W) =& E(a+bX) = a + bE(X) \\
\sigma^2 = Var(X) =& E[X^2] - \mu_x^2 \\
s^2 =& \sum_{all\ i} \frac{(x_i - \bar{x})^2}{n-1} \\
\sigma = StdDev(X) =& \sqrt{\sigma^2} \\
Var(X+c) =& Var(X) \\
Var(cX) =& c^2 Var(X) \\
Var(X+Y) \neq& Var(X) + Var(Y) \\
Var(X+Y) =& Var(X-Y) \\
Var(X) =& E[(X-\mu)^2] \\
Var(X) =& E(X^2) - E(X)^2 (= E(X - E(X))^2)
\end{aligned}$$

Quartiles split the data into 4 equal groups by number of values, or 25% percentiles. Q2 is the median.

```
25% | 25% | 25% | 25%
    Q1    Q2    Q3
```

Chebyshev's Rule:

$$1 - (\frac{1}{k^2})$$

yields the % of data that falls with k std deviations.

Normal Distribution / Z-score / Standardization

$$\begin{aligned}
X \sim& \mathcal{N}(\mu, \sigma^2) \\
Z =& \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1) \\
X =& \sigma Z + \mu \\
P(a \leq X \leq b) =& P[(a-\mu) \leq (X-\mu) \leq (b-\mu)] \\
=& P[\frac{(a-\mu)}{\sigma} \leq \frac{(X-\mu)}{\sigma} \leq \frac{(b-\mu)}{\sigma}] \\
=& P[\frac{(a-\mu)}{\sigma} \leq Z \leq \frac{(b-\mu)}{\sigma}]
\end{aligned}$$

Covariance and Correlation

Covariance gives direction. Correlation gives direction and strength.
Both are for *linear* relationships only.

$$\begin{aligned}
Cov(X,Y) =& \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \\
Cor =& \frac{Cov}{\sigma_x \cdot \sigma_y} \\
r_{xy} =& \frac{s_{xy}}{s_x s_y}
\end{aligned}$$

If $Z = aX + bY$, then

$$\begin{aligned}
\overline{Z} =& a\overline{X} + b\overline{Y} \\
Var(Z) = s_z^2 =& a^2 Var(X) + b^2 Var(Y) + 2abCov(X,Y)
\end{aligned}$$

Return on Portfolio

$$\begin{aligned}
Avgreturn =& w_1\bar{r}_1 + (1-w_1)\bar{r}_2 \\
\sigma^2 =& w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_1\sigma_2 \cdot Correlation \\
Stockreturn =& \alpha + \beta \cdot Indexreturn
\end{aligned}$$

Probabilities

$0 \leq$ All probabilities ≤ 1
Mutually exclusive and Exhaustive

$$\begin{aligned}
P(\overline{A}) =& 1 - P(A) \\
P(A \text{ or } B) =& P(A) + P(B) - P(A \text{ and } B) \\
P(A|B) =& \frac{P(A \text{ and } B)}{P(B)}
\end{aligned}$$

Joint vs. Marginal probabilities

Independent if $P(A|B) = P(A)$
For Independent only: $P(E \text{ and } F) = P(E) \cdot P(F)$

	B	$\overline{B}$
A	$P(AandB) = P(B)P(A\|B)$	$P(Aand\overline{B}) = P(\overline{B})P(A\|\overline{B})$
$\overline{A}$	$P(\overline{A}andB) = P(\overline{A})P(B\|\overline{A})$	$P(\overline{A}and\overline{B}) = P(\overline{A})P(\overline{B}\|\overline{A})$

Random Variables

$$\begin{aligned}
P(X \leq x) \rightarrow& CDF \\
E(cX) =& c \cdot E(X) \\
E(X+c) =& E(X) + c
\end{aligned}$$

Discrete Probability Distributions

$$\mu_x = E(X) = \sum_{all\ x_i} x_i P(X = x_i)$$

$$\begin{aligned}
\sigma_x^2 = Var(X) =& \sum_{all\ x_i} (x_i - \mu_x)^2 P(X = x_i) \\
=& E[X^2] - (\mu_x)^2
\end{aligned}$$

	Chebyshev Any Prob	Empircal Normal
$P(\mu - \sigma < x < \mu + \sigma)$	≥ 0	68%
$P(\mu - 2\sigma < x < \mu + 2\sigma)$	$\geq 75\%$	95%
$P(\mu - 3\sigma < x < \mu + 3\sigma)$	$\geq 89\%$	100%

$$\begin{aligned}
W =& a + bX \\
\mu_W = E(W) =& E(a+bX) \\
=& a + bE(X) \\
=& a + b\mu_x \\
Var(W) =& Var(a+bX) \\
=& b^2\sigma_x^2
\end{aligned}$$

Binomial Distribution

- n independent trials
- binary result
- same probability of success
- total number of successes

$$\begin{aligned}
X \sim& B(n,p) \\
\binom{n}{x} =& \frac{n!}{x!(n-x)!} \\
P(X = x) =& \frac{n!}{x!(n-x)!} p^x q^{(n-x)} \\
\mu_x = E(X) =& n \cdot p \\
\sigma^2 = Var(X) =& n \cdot p \cdot q = n \cdot p \cdot (1-p)
\end{aligned}$$

Shape of distribution depends on p, n.
Small p, left skewed. Large p, right skewed

all success	p^n
all failure	$(1-p)^n$
at least one failure	$1-p^n$
at least one success	$1-(1-p)^n$

Joint Distribution

$$P_{x,y} = P(X = x \text{ and } Y = y)$$

Independent if **for all values**:
$P_{x,y}(X = x \text{ and } Y = y) = P_X(x)P_Y(y)$
Conditional Distribution:
$P_{X|Y}(X = x|Y = y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}$
Conditional Expectation:
$E(X|Y = y) = \sum_{all x} x \cdot P(X = x|Y = y)$
Covariance:

$$\sigma_{XY} = \sum_{i=1}^{N} [x_i - E(X)][(y_i - E(Y)]P(x_i, y_i)$$

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

Correlation:

$$\rho = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

If X and Y are **independent**:

$$E(X + Y) = E(X) + E(Y)$$
$$Var(X + Y) = Var(X) + Var(Y)$$
$$Cov(X, Y) = 0 \text{because} E(XY) = E(X)E(Y)$$

If X and Y are **not independent**:

$$E(X + Y) = E(X) + E(Y)$$
$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Most general combination of random variables:

$$E((a + bX) + (c_d Y)) = a + bE(X) + c + dE(Y)$$
$$Var((a + bX) + (c + dY) = b^2 Var(X) + d^2 Var(Y) + 2bdCov(X, Y)$$

Continuous Probability Distribution

$P(X = x) = 0$ because it is always possible to be more precise.
Probability of an interval: $P(a \leq X \leq b) = F_x(b) - F_x(a)$

Uniform Distribution

$$E(X) = \frac{(a+b)}{2}$$
$$Var(X) = \frac{(b-a)^2}{12}$$

y axis should be a fraction to make area = 1

Decision Analysis

Maximax takes the maximum of each row, and then the maximum of the resulting set. "What is the best that can happen?" Most aggressive.
Maximin takes the minimum of each row, and then the maximum of the resulting set. "What is the worst that can happen?" Most conservative.
Decision Trees show the problem with all possible outcomes and payoffs. Squares are decision nodes. Circles are uncertain external events (probabilistic node, like a coin). Walk the tree to find which gives the best expected value. "Fold back the tree" walking from right to left. (In CompSci, this is called a Depth First Search). Multiply the end states times the probabilities and then aggregate to one level up in the tree. At any given branch, the best path can be determined by the aggregated value of the path.
Expected Monetary Value does not include utility and risk. Since it involves a predicted average over repetition, it may not be appropriate for one-off decisions. It also factors in only Monetary value so it does not take into account other objectives (e.g. environment, aesthetic, social)

Central Limit Theorem

Requires $n \geq 30$ or population be normally distributed.
For a mean, a sample distribution will have:

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$P(\bar{x} < A) = P(Z < \frac{A - \mu_{\bar{x}}}{\sigma_{\bar{x}}/\sqrt{n}})$$

Discrete data requires $n \cdot p \geq 5$ and $n(1-p) \geq 5$
For proportions, a sample distribution will have:

$$p_{\hat{p}} = p$$
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

To find if a population totals less a value ("the Swan Problem"):

$$P\left(\sum_{i=1}^{n} x_i < \max\right)$$

Divide both sides by n:

$$P(\overline{X} < \text{avg})$$

From CLT:

$$\overline{X} \sim \mathcal{N}(\mu, \frac{s^2}{n})$$

Then Z-score:

$$P(Z < \frac{\text{avg} - \mu}{s/\sqrt{n}})$$

Bias of an Estimator

In practice n has to relatively much larger like > 100.
Guesses should be *unbiased* and have *minimum variance*.
MVUE (Minimum Variance, Unbiased Estimates).

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Unbiased if $bias = 0$ (expected value equals true, not a particular value of $\bar{x}$).
For samples, we divide by $n-1$ instead of n to make an unbiased estimator. The guess would otherwise be too low.

$$E(\bar{x}) = \sum_{i=1}^{\infty} x_i p_i = \mu$$
$$E(s^2) = \sigma^2$$
$$E[X + c] = E[X] + c$$
$$E[X + Y] = E[X] + E[Y]$$
$$E[aX] = aE[X]$$
$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

Example: Roulette has a \$1 bet with a \$35 payoff for $\frac{1}{38}$ odds.

$$E[\text{gain from a \$1 bet}] = -\$1 \cdot \frac{37}{38} + \$35 \cdot \frac{1}{38} = -\$0.0526$$

Confidence Interval - Margin of Error

$$\text{Margin of error(MoE)} = Z_{\alpha/2} \times \text{standard error}$$

For a 95% confidence, $Z = 1.96$. (For 99%, $Z = 2.58$, For one sided 90%, $Z = 1.28$)
For a given CI = (Lower, Upper):

$$\bar{x} = (U + L)/2$$
or
$$\hat{p} = (U + L)/2$$
$$\text{MoE} = (U - L)/2$$

The CI interval is the confidence percentage that the true population mean is within the interval. It does not imply what percentage of the population is outside the interval.
The larger the CI Percentage, the wider it is and the smaller the risk of being incorrect.
Factors affecting the margin of error:

- data variation σ. Direct relation
- sample size n. Inverse relation
- level of confidence, $1 - \alpha$. Direct relation

Confidence Interval - Mean

$$Var(\bar{x}) = \frac{s^2}{n}$$

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$$

$$n = \left(\frac{1.96s}{MoE}\right)^2$$

μ, the population mean, cannot be determined from the CI. We are only 95% certain that μ is in the CI range.

Confidence Interval - Proportion

$$\hat{p} = \frac{x}{n}$$

$$Var(\hat{p}) = \frac{p(1-p)}{n} \text{ or } \frac{\hat{p}(1-\hat{p})}{n}$$

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$n = (\frac{1.96}{0.05})^2 \hat{p}(1-\hat{p})$$
$$= 1536.64 * \hat{p}(1-\hat{p})$$

Worst case, use $\hat{p} = 0.50$. e.g. 95% confident, 2% accuracy, find n:

$$n = (\frac{1.96}{0.02})^2 (0.5)^2$$

For small n, use Agresti with a different $\hat{p}$:

$$\hat{p} = \frac{x+2}{n+4}$$

Hypothesis Test - General

The purpose of hypothesis testing is to help the researcher reach a conclusion about a population by examining the data contained in a sample.
H_0 is default position, the status quo. It requires significant evidence to be disproven.

Hypotheses	Decision Rule
$H_0 : \mu = \mu_0$	
$H_a : \mu \neq \mu_0$	If $\lvert t_{stat}\rvert > 1.96$, reject H_0
$H_0 : \mu = \mu_0$	
$H_a : \mu < \mu_0$	If $t_{stat} < -1.64$, reject H_0
$H_0 : \mu = \mu_0$	
$H_a : \mu > \mu_0$	If $t_{stat} > 1.64$, reject H_0

We use 1.96 because it is 2.5% on either side. We use 1.64 because it is 5% on a single side.

Hypothesis Test - Mean

Calculation by hand using a t test:

$$t_{stat} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Hypothesis Test - Proportion

$$t_{stat} = \frac{(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)/n}}$$

Types of Errors

Type I the null hypothesis is rejected when it is true

Type II the null hypothesis is accepted when it is false

α is *level of significance* - probability of making a Type I error. The greater the cost of an error, the smaller α should be. β is the probability of making a Type II error. There is an inverse relation between Type I and II errors. Reducing one increases the other. The only way to reduce both is to increase n the sample size.

Comparing Two Sets - General

The null hypothesis is always $H_0 : p_1 = p_2$.

Hypotheses	Decision Rule	Stata Diff $(p_1 - p_2)$
$H_0 : p_1 = p_2$		
$H_a : p_1 \neq p_2$	If $\lvert T\rvert > 1.96$, reject H_0	H_a : diff $\neq 0$
$H_0 : p_1 = p_2$		
$H_a : p_1 < p_2$	If $T < -1.64$, reject H_0	H_a : diff < 0
$H_0 : p_1 = p_2$		
$H_a : p_1 > p_2$	If $T > 1.64$, reject H_0	H_a : diff > 0

If the interval is all positive then $\hat{p}_1 > \hat{p}_2$. If the interval is all negative then $\hat{p}_1 < \hat{p}_2$. If the interval spans 0, then one is not significantly bigger than the other (or cannot be determined). As long as $n > 30$, it doesn't matter if the sample size is different between the random variables.

Comparing Two Proportions

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$$

The 95% confidence interval for $p_1 - p_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Decision Rules for Testing Two Proportions:

$$T = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \text{ where } \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

$\hat{p}$ is called the *pooled proportion*.

Comparing Two Means

Requirements:

1. σ_1 and σ_2 are unknown. No assumption made about their equality.
2. The two samples are independent.
3. Both samples are simple random samples.
4. The two samples size are both large (ie. > 30) or both populations have normal distributions.

A confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Comparing Two Normal Distributions

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$
$$B \sim \mathcal{N}(\mu_B, \sigma_B^2)$$

Find $P(A < B + c)$:

$$P(A - B - c < 0)$$
$$(A - B - c) \sim \mathcal{N}(\mu_A - \mu_B - c, \sigma_A^2 + \sigma_B^2)$$
$$P(X = A - B - c > 0)$$

then Z-score

Matched Pairs

This when there are two samples that are **not** independent, e.g. Weight Watchers, Before / After or matched, shared characteristics.

Is the data matched or independent?

If we don't take into account the match, the results are wrong.

To account for this, take the difference between $\overline{X}_1 - \overline{X}_2$ and then do a hypothesis test on the *difference*.

$$H_0 : \mu_D = 0$$
$$H_a : \mu_D > 0$$

Chi-Square Test - Goodness of Fit

A class of two tests: *goodness of fit* and *statistical independence*.
Tests several proportions at the same time, aka the multinomial setting.
k categories of interest with $p_1, p_2, ..., p_k$ probabilities that a value is in a particular cell. All p's add up to 1, as usual.

$$H_0 : p_1 = a_1, p_2 = a_2, ..., p_k = a_k$$

where $a_1, a_2, ..., a_k$ are the values to be tested.
H_a : at least one p_i is not equal to the specified value.

O observed frequency of an outcome, given

E expected frequency of an outcome, calculated

k number of different categories

n number of trials

s_i sample standard deviation

Calculate Observed and Expected to see if they are consistent. Known as Chi-Squared Goodness of Fit (GOF) Test.

$$e_i = n \cdot p_i$$

$$\chi^2 = \sum_{\text{all } i} \frac{(o_i - e_i)^2}{e_i}$$

Smallest possible value is zero. Smaller χ^2 means H_0 is plausible. Larger χ^2 means reject the null.
Use table to determine cut-off values (determined by degrees of freedom $k-1$). As before, we typically use $\alpha = 5\%$ level of significance.
If $\chi^2 > \chi^2_{\alpha,k-1}$, then reject the null in favor of H_a.
Something has changed (but we don't know what or which direction).
Requirements:

1. Data is random
2. Data has frequency counts per category
3. $e_i \geq 5, o_i$ can be anything. Might need to group smaller categories.

Chi-Squared Test of Independence

aka Two-way Chi-Squared Test.
Tests if r rows and c columns are independent or not. H_0 is independent, H_a is dependent.
Look for P value. Again, if P is low, H_0 must go.
Need to figure out the probabilities in order to determine e_i.
Recall, for independent variables:

$$P(A \text{ and } B) = P(B)P(A)$$

If $e_{ij} = P(r_i)P(c_j)$ then independent

Regression - Errors

$$\begin{aligned} Residual =& Actual - Predicted \\ =& Observed - Fitted \\ =& y - \hat{y} \end{aligned}$$

Mean Absolute Error:

$$e = \frac{1}{n} \sum_{all\ i} |y_i - \hat{y}_i|$$

Least squares error:

$$\sum_{i=1}^{n} (y_i - b_0 - b_i x_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Least Squares has two interesting properties:
1) The mean of the residuals $\frac{1}{n}\sum e_i = 0$. This implies that $\sum e_i = 0$.
2) The mean of the fitted values equals the mean of the original values ie. $\bar{y} = \bar{\hat{y}}$.
Since $\hat{y}$ is defined as a linear relationship with x, there is a perfect correlation, ie. $corr(\hat{y}, x) = 1$.
$corr(E, X) = 0$ as there is no linear relationship between the errors and x values.
The sum of squares total (SST) = regression (SSR) + error (SSE)

$$Var(Y) = Var(\hat{Y}) + Var(e)$$

SST is the total information in y. SSR is total information explained by x. SSE is the information in y not explained by x. We want to maximize SSR and minimize SST. If SST = SSR, then SSE = 0 and we have a perfect fit.

Regression - Single Variable Diagnostics

R^2: what percentage of the variation of the predicted y is explained by the variation in x. The rest is unexplained by the model.
This is the *coefficient of determination*:

$$R^2 = \frac{SSR}{SST}$$

R^2 is in the range $[0, 1]$. The closer R^2 is to 1, the better the fit. If it is 0, the model explains none of the variability of the response data around its mean.
The most accurate guess is around $\bar{x}$. Further away from this point, the errors become quadratically wrong.
R^2 does not indicate whether a regression model is adequate. You can have a low R^2 for a good model, or a high R^2 value for a model that does not fit the data!
Adding "junk" x variables will increase R^2.

Estimating Error Variance s_e and Noise

To estimate variance and standard deviation of the error:

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-2}$$

s_e^2 is our estimate of σ^2. $s_e = \sqrt{s_e^2}$ is our estimate of σ. The standard deviation of the noise, aka s_e or **Root MSE**, is important as it is good estimate and unbiased. It is more important than R^2 as a measure of quality of regression.
Use s_e to form bands around the regression line:

Percent of y Values	Band
68%	$b_0 + b_1 X \pm 1 s_e$
95%	$b_0 + b_1 X \pm 1.96 s_e$
99%	$b_0 + b_1 X \pm 3 s_e$

Assuming a 95% confidence interval, use:

$$\hat{y} = b_0 + b_1 \times x \pm 1.96 s_e$$

For 68%, use 1 instead of 1.96. For 99% use 3 instead of 1.96.
A confidence interval for β_1 is:

$$b_1 \pm 1.96(s_{b_1})$$

where:

$$Var(b_1) = s_{b_1}^2 = \frac{s_e^2}{(n-1)s_x^2}$$

Badness occurs with small n, big s_e, small s_x^2. We actually want more variance in the x's so that less of the overall variance is due to noise.
A confidence interval for β_0 is:

$$b_0 \pm 1.96(s_{b_0})$$

where:

$$Var(b_0) = s_{b_0}^2 = s_e^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)$$

Regression - Hypothesis Testing

Recap: Assumptions: linear model, noise is modeled by a standard distribution.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

These tests work identically for β_1 and β_0.
We want to test whether β_1 equals a proposed value. We always use a two-sided test.

$$\begin{aligned} H_0 :& \beta_1 = \beta_1^* \\ H_a :& \beta_1 \neq \beta_1^* \end{aligned}$$

To know if x affects y, test if $\beta_1 = 0$. Use the following test statistic:

$$T = \frac{b_1 - \beta_1^*}{s_{b_1}}$$

Test Statistic	Decision Rule
$H_0 : \beta_1 = \beta_1^*$	
$H_a : \beta_1 \neq \beta_1^*$	If $\lvert T\rvert > 1.96$, reject H_0
$H_0 : \beta_1 \geq \beta_1^*$	
$H_a : \beta_1 < \beta_1$	If $T < -1.64$, reject H_0
$H_0 : \beta_1 \leq \beta_1^*$	
$H_a : \beta_1 > \beta_1^*$	If $T > 1.64$, reject H_0

As before if $n < 30$, use t distribution.

Multiple Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

Comparison with Simple Linear Regression:

- intercept is the same
- Slope b_i is the change in y given a unit change in x_i, while holding all other variables constant
- SST, SSE, SSR and R^2 are the same. Instead of variance of y explain by a single x, it is explained by a set of x's
- s_e has a new formula: $s_e = \sqrt{SSE/(n-k-1)}$
- Slope coefficient confidence intervals are the same
- p-values (one for each x_i) are the same
- *Interpretation* is different due to multiple x_i's

Confidence Intervals are the same:

$$b_j \pm 1.96 s_{b_j}$$

The hypothesis test:

$$H_0 : \beta_j = \beta_j^*$$

when

$$t = \left| \frac{b_j - \beta_j^*}{s_{b_j}} \right| \geq 1.96$$

or p-value < 0.05.

By adding dimensions (x's), the error sum of squares (SSE) will decrease so R^2 will always increase. R^2 becomes even less useful in multiple regression. To counteract, the **adjusted R-squared** is available:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Adjusted R-squared imposes an "artificial" penalty for adding dimensions.

The Overall F Test null hypothesis:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

ie. you need nothing. The alternative hypothesis is that at least one x_n is required.

If there are no x's in the model, then SSR=0, and SST=SSR+SSE = SSE.

However, it at least one x is useful, then SSR $\neq$ 0, and ideally if some x's are useful, then SSR>SSE. So we compare SSR to SSE in some fashion.

The test statistic is:

$$f = \frac{(SSR)/k}{SSE/(n-k-1)}$$

and reject for large values of f (the x's explain a significant portion of the model).

The f distribution is a series of tables (like χ^2). It tells us when to reject the null by

$$f \sim F_{k,n-k-1}$$

The decision rule is to reject H_0 if $f \geq f_{k,n-k-1,\alpha}$.

Variable Selection

Forward Stepwise Regression starts with no variables and then adds one at a time.

Backward Stepwise Regression starts with all variables and then delete the least important one based on largest p-value > 0.05. Refit and repeat. Stop when all variables are significant.

Use the p-value or the Confidence Interval rather than the t stat for small ($n < 30$) datasets. The p-value is automatically adjusted by Stata, whereas the correct t stat must be looked up in a table.

The smallest p-value determines the most important variable (do not use the coefficients!).

Predictions on the full model might still be better than the simplified model. Backward regression may not result in the same results as forward regression.

Dummy Variables

Binary - takes on a value of 0 or 1.

Convert x variables with discrete values into a set of binary / dummy variables.

This effectively allows two regressions at once.

β_1 describes the difference the two discrete values, then use hypothesis testing of β_1. Could be used for discrimination testing. The intercept or baseline is when $\beta_1 = 0$. When $\beta_1 \neq 0$, it explains the difference in value of the discrete variable.

β_0, the intercept, represents when all dummy variables have a value of zero.

Remember to use **"while holding everything else the same."**

Example Recoding problem: what if the baseline group was 'other' instead of 'white'?

Original regression:

$$\hat{y} = 30 - 4f + 5b - 2o + 0.3e$$

Recoding the baseline w for o:

$$\hat{y} = 28 - 4f + 7b + 2w + 0.3e$$

Just to be clear, each of the non-baseline values already have the baseline included, i.e. the coefficient of x_1 is $\beta_0 + \beta_1$.

We could use a t-test to compare just one variable:

```
test salary, by (males) unequal
```

The regression allows more variables to be tested than the t-test.

Interaction Term Model

Models two variables that interact, meaning that the coefficients can combine. Multiply a dummy variable by another variable in the model to create a new variable called the *interaction variable*.

This allows the effect of a continuous variable to differ depending on the value of discrete variable.

It is possible to create an interaction variable with two continuous variables, but then partial derivatives are required for interpretation.

When a discrete variable has multiple values, leave one out to create a baseline and then have dummy variables for each additional category. The baseline represents when all other categories are zeroed out and equals β_0.

Regression - Multivariable Diagnostics

Residuals vs. Fitted plot should have random pattern.

The standardized residuals should be in the range $[-2, +2]$ 95% of the time. `sres = res /` s_e

The following are all in units of y: yi,ei,se. ri is unitless since division by the standard deviation removes the units. If y = b0 + b1x, then b0 must be in units of y and b1 must be in units of y/x. (This may appear on an exam question as a hypothetical ?if the units of x changed, how would b1 or b0 change).

Omnibus plot: If all assumptions are met, (meaning $y = \hat{y} + e$, all the information about the x's is stored in $\hat{y}$ and everything else is in e), then the correlation$(\hat{y}, e)$ should be = 0. Plotting e vs. $\hat{y}$ should be a random blob.

For multiple variables, plots should be done for each e vs x_i.

Normality

a huge assumption that errors in our regression model are normally distributed. This enables constructing confidence intervals and doing hypothesis tests. This assumptions *must* be validated.

Three different normality tests in Stata. H_0 : errors are normally distributed. H_a : errors are not normally distributed.

```
predict res, r
swilk res
sfrancia res
sktest res
```

We want to fail to reject the null hypothesis, looking for high p-values. If a transformation is done, this cycle must be restarted.
Look at standardized residuals.
If normality is not satisfied:

- try transforming the dependent variable
- log transform either an x_i or less preferably, the y
- remove outliers

Heteroskedasticity

non-constant variance. We assume that the variance is consistent across all values of all x's. The may not be the case.
Homoskedastic noise:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

A plot would have a straight band around the data, ie. $\text{Var}(\epsilon_i) = \sigma^2$. Irrespective of x_i, the variance is the same.
If we assume homoskedasticity, we have verify our assumption as always.
Heteroskedastic noise:

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

Note the subscript $\sigma_\mathbf{i}^2$ which means the variance changes for a given x_i. This implies that there are a lot of different variances to estimate, complicating the model. A plot of either y vs x or residuals vs. fitted values would have a spread of data points in a fan or tunnel shape. A frequently occurring situation is that $\text{Var}(\epsilon_i) = x^2\sigma^2$ meaning that the variance increases with larger values of x.
With heteroskedasaticity, the estimates are ok, but the standard errors are incorrect, which is a problem in the model. The coefficients are useful, but the p-values are wrong.
An easy way (at this level) to resolve this is by taking a log(y). This makes comparison of models more difficult. x transforms can be compared against prior models because the units of y have not changed.
Test for homoskedasticity in Stata: `hettest`. H_0 is constant variance, and H_a is heteroskedasaticity. Once again, we want a large p-value, if possible.

Multicollinearity

some or all x's are related to each other, when they should be related to y. Two highly related x's confuse the regression algorithm. The standard deviation of the regression coefficients (s_{b_i}) will be disproportionately large, resulting in t ratios that are too small because, recall:

$$t_{stat} = \frac{b_i}{s_{b_i}}$$

When the t stat is too small, it will seem that some variables are not needed when in fact they *are* needed.

"The regression coefficient estimates will be unstable. Because of the high standard errors, reliable estimates are hard to obtain. Signs of the coefficients may be opposite of what is intuitive or reasonable. Dropping one variable from the regression will cause large changes in the estimates of the other variables."
The **overall F test** is another way of quickly finding that at least one variable is necessary. If the individual p-values say all the variables aren't necessary but the F test disagrees, then there is multicollinearity.
Early on, do a correlation of all x's. If two or more x's are highly related (corr > 0.8), then some should be thrown out. Keep the variables that have a better correlation with y.
Dealing with multicollinearity:

- Throw out redundant explanatory variables
- Get more data
- Redefine variables, such as creating an index, e.g. $\frac{x_1+x_2}{2}$
- Step-wise regression

Variance Inflation Factors (VIF) is a good automated way of discovering this.
Take regression of all permutations of x, ie. Take regress of x_1 on x_2, x_3. Take regress of x_2 on x_1, x_3. Take regress of x_3 on x_1, x_2.
Note R^2 on each regression, and then calculate:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Interpretation: If there is no relationship, $R_j^2 = 0$, so $VIF_j = 1$. As R_2 increases (due to a better fit on the regression), VIF_j also increases. If $R_j^2 = 0.90$, then $VIF_j = 10$.
As a rule of thumb, if $VIF_j > 10$, then multicollinearity of x_j may be a problem.

Nonlinearities

By default, regresses y on $\hat{y}^2, \hat{y}^3, \hat{y}^4$. A significant value means that polynomial terms should be added. H_0: no transformations of x needed, H_a: polynomial or other transformations are needed. A small p-value means one of the x's (but not which) needs to be transformed.
In Stata, use `ovtest`. Re-run again. Might need a x^2 and x^3 and maybe other terms. (Adding a cubic term might need to drop the linear term). As usual, be wary of overfitting especially with polynomials.

Finding Outliers

Recall: influential observations are the worst (extreme in x and y).
Z scores and standard deviations are a simple and quick way to find outliers.
Cook's distance is (approx) $|e_i| * |x_i - \bar{x}|$ (high residual and far away from mean(x)). Looking for extreme Cook values to know which to drop.