Normal Distribution / Z-score

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

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:

$$\begin{aligned}
\mu_{\bar{x}} = & \mu \\
\sigma_{\bar{x}} = & \frac{\sigma}{\sqrt{n}}
\end{aligned}$$

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

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

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

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

In practice n has to relatively much larger like > 100.
Recap:

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

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.

$$E(s^2) = \sigma^2$$

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 a given CI = (Lower, Upper):

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

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

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

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

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

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

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

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

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

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

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:

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

For small n, use t distribution:

$$t_{stat} = \frac{\bar{x} - \mu_0}{s/\sqrt{n-1}}$$

Hypothesis Test - Proportion

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

For small n, use t distribution $(n-1)$.

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.

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

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

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