Normal Distribution / Z-score

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

Central Limit Theorem

Requires $n \ge 30$ 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 $n \cdot p \ge 5$ and $n(1-p) \ge 5$
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"):

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

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:

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

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
$H_0 : \mu = \mu_0$ $H_a : \mu \ne \mu_0$	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 $(p_1 - p_2)$
$H_0 : p_1 = p_2$ $H_a : p_1 \ne p_2$	If $\lvert T \rvert > 1.96$, reject H_0	H_a : diff $\ne 0$
$H_0 : p_1 = p_2$ $H_a : p_1 < p_2$	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*.

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