Chapter 10

Pooled (combined) sample proportion The overall proportion of successes in the two samples is

$$\hat{p}_C = \frac{\text{count of successes in both samples combined}}{\text{count of individuals in both samples combined}} = \frac{X_1 + X_2}{n_1 + n_2}$$

Pooled two-sample *t* statistic A special version of the two-sample *t* statistic that assumes that the two populations have the same variance.

Randomization distribution The distribution of a statistic (like $\hat{p}_1 - \hat{p}_2$ or $\bar{x}_1 - \bar{x}_2$) in repeated random assignments of experimental units to treatment groups assuming that the specific treatment received doesn't affect individual responses. When the Random, Normal, and Independent conditions are met, our usual inference procedures based on the sampling distribution of the statistic will be approximately correct.

Sampling distribution of $\hat{p}_1 - \hat{p}_2$ Choose an SRS of size n_1 from population 1 with proportion of successes p_1 and an independent SRS of size n_2 from population 2 with proportion of successes p_2.

- **Shape:** When $n_1p_1, n_1(1-p_1), n_2p_2,$ and $n_2(1-p_2)$ are all at least 10, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal.
- **Center:** The mean of the sampling distribution is $p_1 - p_2$. That is, the difference in sample proportions is an unbiased estimator of the difference in population proportions.
- **Spread:** The standard deviation of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is
$$\sqrt{\frac{\hat{p}_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$
as long as each sample is no more than 10% of its population (the 10% condition).

Sampling distribution of $\bar{x}_1 - \bar{x}_2$ Choose an SRS of size n_1 from population 1 with mean μ_1 and standard deviation σ_1 and an independent SRS of size n_2 from population 2 with mean μ_2 and standard deviation σ_2.

- **Shape:** When the population distributions are Normal, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is Normal. In other cases, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will be approximately Normal if the sample sizes are large enough ($n_1 \geq 30$ and $n_2 \geq 30$).
- **Center:** The mean of the sampling distribution is $\mu_1 - \mu_2$. That is, the difference in sample means is an unbiased estimator of the difference in population means.
- **Spread:** The standard deviation of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is
$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
as long as each sample is no more than 10% of its population (the 10% condition).

Standard error of $\hat{p}_1 - \hat{p}_2$ The estimated standard deviation of the statistic $\hat{p}_1 - \hat{p}_2$, given by

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

Standard error of $\bar{x}_1 - \bar{x}_2$ The estimated standard deviation of the statistic $\bar{x}_1 - \bar{x}_2$, given by

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Two-sample *t* interval for a difference between two means When the Random, Normal, and Independent conditions are met, an approximate level *C* confidence interval for $\bar{x}_1 - \bar{x}_2$ is

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

Here *t** is the critical value for confidence level *C* for the *t* distribution with degrees of freedom from either Option 1 (technology) or Option 2 (the smaller of $n_1 - 1$ and $n_2 - 1$).

Two-sample *t* statistic When we standardize the estimate $\bar{x}_1 - \bar{x}_2$, the result is the two-sample *t* statistic

$$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The statistic *t* has the same interpretation as any *z* or *t* statistic: it says how far $\bar{x}_1 - \bar{x}_2$ is from its mean in standard deviation units.

Two-sample *t* test for the difference between two means Suppose the Random, Normal, and Independent conditions are met. To test the hypothesis $H_0 : \mu_1 - \mu_2 = \text{hypothesized value}$, compute the two-sample *t* statistic

$$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\mathrm{s}_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Find the *P*-value by calculating the probability of getting a *t* statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the *t* distribution with degrees of freedom approximated by technology or the smaller of $n_1 - 1$ and $n_2 - 1$.

Two-sample *z* interval for a difference between two proportions When the Random, Normal, and Independent conditions are met, an approximate level *C* confidence interval for $\hat{p}_1 - \hat{p}_2$ is $\left(\hat{p}_1 - \hat{p}_2\right) \pm z^* \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$ where *z** is the critical value for the standard Normal curve with area *C* between –*z** and *z**.

Two-sample *z* test for the difference between two proportions Suppose the Random, Normal, and Independent conditions are met. To test the hypothesis $H_0 : p_1 - p_2 = 0$, first find the pooled proportion $\hat{p}_C$ of successes in both samples combined. Then compute the *z* statistic

$$z = \frac{\left(\hat{p}_1 - \hat{p}_2\right) - 0}{\sqrt{\frac{\hat{p}_C\left(1-\hat{p}_C\right)}{n_1} + \frac{\hat{p}_C\left(1-\hat{p}_C\right)}{n_2}}}$$

Find the *P*-value by calculating the probability of getting a *z* statistic this large or larger in the direction specified by the alternative hypothesis H_a.