(§5.1-5.3 Large-Sample Confidence Interval for Mean; Proportion; Small-Sample Interval for Mean)

§5.4 (Large-Sample) Confidence Inferences for the Difference of Two Means

§5.5 Confidence Inferences for the Difference of Two Proportions

(§5.6 Small-Sample Confidence Inferences for the Difference of Two Means)

5.4 Confidence Intervals for the Difference of Two Means

We compare two population means, μ_X and μ_Y, by studying their difference, $\mu_X - \mu_Y$. Notation:

	Population 1	Population 2
Variable	X	Y
Mean	μ_X	
Standard deviation		σ_Y
Sample size	n_X	
Sample mean	$\bar{X}$	
Sample standard deviation		s_Y

For inference about $\mu_X - \mu_Y$, use the statistic _____________.

To find a confidence interval for $\mu_X - \mu_Y$, we need the distribution of _____________. Recall for independent X and Y:

- (§2.5) $\mu_{X-Y} =$
- (§2.5) $\sigma^2_{X-Y} =$
- (§2.5) $\sigma^2_{\bar{X}} =$
- (§4.5) If $X \sim N(\mu_X, \sigma^2_X)$ and $Y \sim N(\mu_Y, \sigma^2_Y)$, then $X - Y \sim$
- (§4.11) For large n, the CLT says $\bar{X} \sim$ (≈)

It follows that, for large n_X and n_Y, $\bar{X} - \bar{Y} \sim$ ____.

Confidence Intervals on the Difference of Two Means

Recall that many confidence intervals have the form

(point estimate) ± (margin of error)
=(point estimate) ± (______ value for confidence) × [(estimated or true) _________________ of point estimate]
$=\hat{\theta} \pm$ (table value for confidence) $\times\, \sigma_{\hat{\theta}}$

Derive a Confidence Interval

Here's our previous derivation of a confidence interval for a normally distributed statistic:

- Consider a statistic $\hat{\theta}$ as an estimator for a parameter θ, where $\hat{\theta} \sim N(\theta, \sigma^2)$

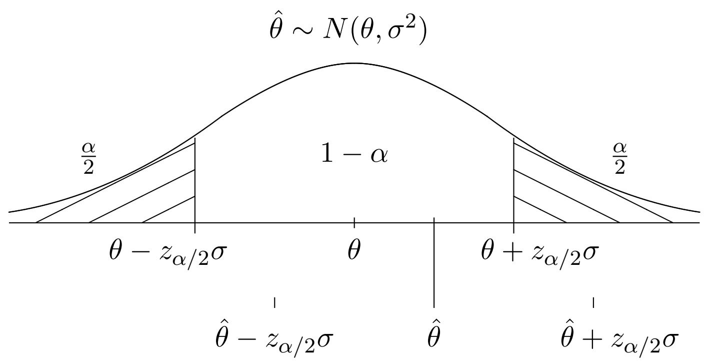
  (Generalize because it's ____________ to write θ than ____________, $\hat{\theta}$ than ____________, and ______ than $\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$.)

- Let $z_{\alpha/2}$ = the z-score cutting off a right tail area ______ from $N(0,1)$ (as before), so $P(-z_{\alpha/2} < Z < z_{\alpha/2}) =$ ____________ (draw)

- Unstandardize using $Z = \frac{}{}$ to get $P(-z_{\alpha/2} < \frac{\hat{\theta}-\theta}{\sigma} < z_{\alpha/2}) = 1 - \alpha$; solve in two ways:

  - for $\hat{\theta}$ in the middle: $P(\theta - z_{\alpha/2}\sigma < \hat{\theta} < \theta + z_{\alpha/2}\sigma) = 1 - \alpha$ (pictured __________________)
  - for θ in the middle: $P(\hat{\theta} - z_{\alpha/2}\sigma < \theta < \hat{\theta} + z_{\alpha/2}\sigma) = 1 - \alpha$ (draw)

That is, $\hat{\theta} \pm z_{\alpha/2}\sigma$ contains ______ for a proportion ____________ of random samples (see picture, below). It's the $100\%(1-\alpha)$ *confidence interval* for θ.



The Case of a Difference of Two Means

Letting $\theta =$ ____________ and $\hat{\theta} =$ ____________, gives the confidence interval we need:

Let $X_1, \cdots, X_{n_X}$ and $Y_1, \cdots, Y_{n_Y}$ be independent large random samples from populations with means μ_X and μ_Y and standard deviations σ_X and σ_Y. A $100\%(1-\alpha)$ confidence interval for $\mu_X - \mu_Y$ is

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

(We usually need to use $\sigma_X \approx$ ______ and $\sigma_Y \approx$ ______.)

e.g. A crayon maker is comparing the effects of two yellow dyes on crayon brittleness. Dye B is more expensive than dye A, but might produce a stronger crayon. 40 crayons are tested with each dye, and the impact strength (in joules) is measured for each. The A strength averaged 2.6, with standard deviation 1.4. The B strength averaged 3.8, with standard deviation 1.2. Find a 99% confidence interval for the difference, B − A, in population strengths.

5.5 Confidence Intervals for the Difference of Two Proportions

We compare two population proportions, p_X and p_Y, by studying their difference, $p_X - p_Y$.

Notation:

	Population 1	Population 2
Success probability	p_X	p_Y
#Trials	n_X	n_Y
#Successes	X	Y
Sample proportion of successes	$\hat{p}_X = \dfrac{X}{n_X}$	$\hat{p}_Y = \dfrac{Y}{n_Y}$

For inference about $p_X - p_Y$, use the statistic _____________.

To find a confidence interval for $p_X - p_Y$, we need the distribution of $\hat{p}_X - \hat{p}_Y$. Recall for independent X and Y:

- If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then $X - Y \sim$ (§4.5)
- If $X \sim \text{Bin}(n, p)$, and $np > 10$ and $n(1-p) > 10$, then $X \sim N($_______, ______________$)$ ($\approx$; because CLT applies to $X = \sum_{i=1}^{n} B_i$, where $B_i \sim \text{Bernoulli}(p)$) (§4.11)

  $\implies \hat{p} = \dfrac{X}{n} \sim$

It follows that, for $n_X p_X > 10, n_X(1 - p_X) > 10, n_Y p_Y > 10$, and $n_Y(1 - p_Y) > 10$,

$$\hat{p}_X - \hat{p}_Y \sim$$

We need the standard deviation for inference about the unknown $p_X - p_Y$, but we don't know _______ or _______. If the #successes and #failures are more than _______in each sample, we can approximate them with _______ and _______.

Confidence Intervals on the Difference of Two Proportions

Recall, again, that many confidence intervals have the form

(point estimate) $\pm$ (margin of error)
=(point estimate) $\pm$ (table value for confidence) $\times$ [(estimated or true) standard deviation of point estimate]
=$\hat{\theta} \pm$ (table value for confidence) $\times\ \sigma_{\hat{\theta}}$

The Old Confidence Interval

If the #successes and #failures are more than 10 in each sample, then the old $100\%(1-\alpha)$ confidence interval for $p_X - p_Y$ is

$$(\hat{p}_X - \hat{p}_Y) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}$$

For small samples, this interval ____________________ $p_X - p_Y$ for a proportion $1-\alpha$ of samples.

The New Plus-Four Confidence Interval

Recent research (2000) describes an improvement: add four fake observations, two successes and two failures, ____________________ to each sample. (The §5.2 plus-four interval for a single proportion added _______ successes and _______ failures to the single sample.)

Let independent $X \sim \text{Bin}(n_X, p_X)$ and $Y \sim \text{Bin}(n_Y, p_Y)$. Define

$$\tilde{n}_X = \_\_\_\_\_\_\_\_\_, \tilde{n}_Y = \_\_\_\_\_\_\_\_\_, \tilde{p}_X = \text{―――――}, \text{ and } \tilde{p}_Y = \text{―――――}$$

Then the $(100\%)(1-\alpha)$ plus-four confidence interval for $p_X - p_Y$ is

$$(\tilde{p}_X - \tilde{p}_Y) \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}_X(1-\tilde{p}_X)}{\tilde{n}_X} + \frac{\tilde{p}_Y(1-\tilde{p}_Y)}{\tilde{n}_Y}}$$

This interval can be used if $n_X > 4$ and $n_Y > 4$, without regard for the #successes and #failures. (Since $(p_X - p_Y) \in$ [_______, _______], trim the interval if it extends outside [_______, _______].)

e.g. A randomized double-blind experiment assigned 244 smokers who wanted to quit to receive nicotine patches and another 245 to receive patches and an antidepressant. After a year, 40 in the first group and 87 in the second had quit. Give a 99% plus-four confidence interval for the difference (treatment $-$ control) in proportions of smokers who quit.