10

Statistical Inference for Two Samples

LO

- On difference of *two population mean/proportions:*
    - *confidence intervals*
    - *test of hypotheses*
- Discuss on sample sizes

Inference on the Difference in Means of Two Normal Distributions, Variances Known

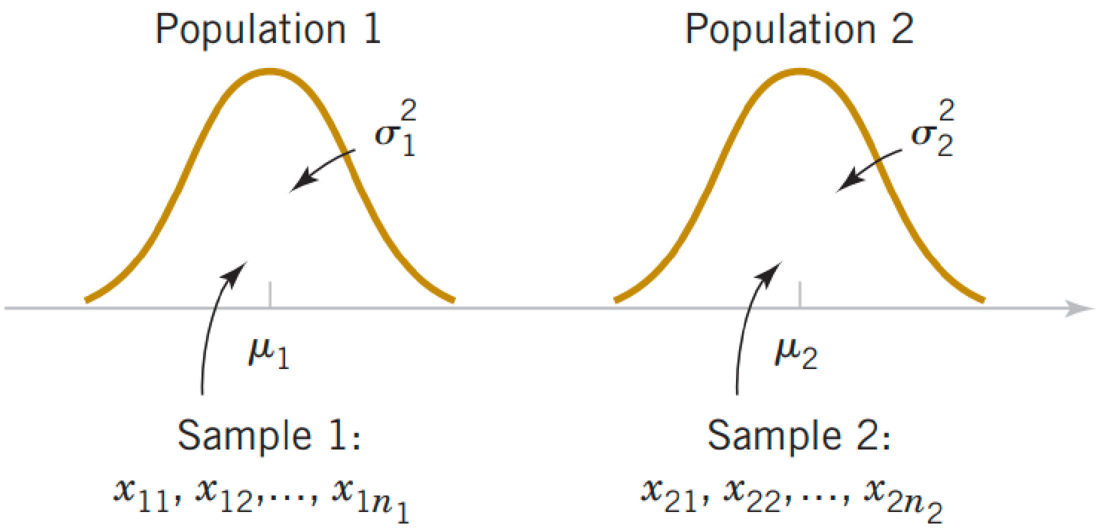
Assumptions for Two-Sample Inference

(1) X_{11}, X_{12}, …, X_{1n1} is a random sample from population

(2) X_{21}, X_{22}, …, X_{2n2} is a random sample from population

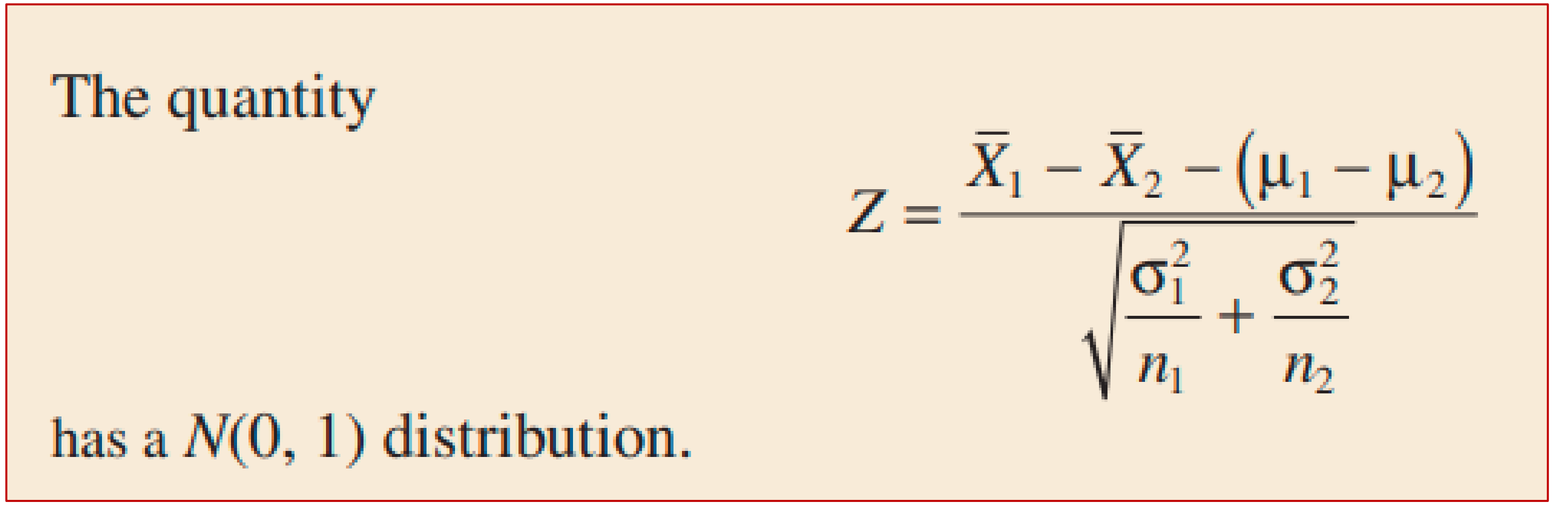
(3) The two populations represented by X_1 and X_2 are *independent*.

(4) Both populations are *normal*.



Two independent populations

Tests on the Difference in Means, Variances Known



The quantity

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

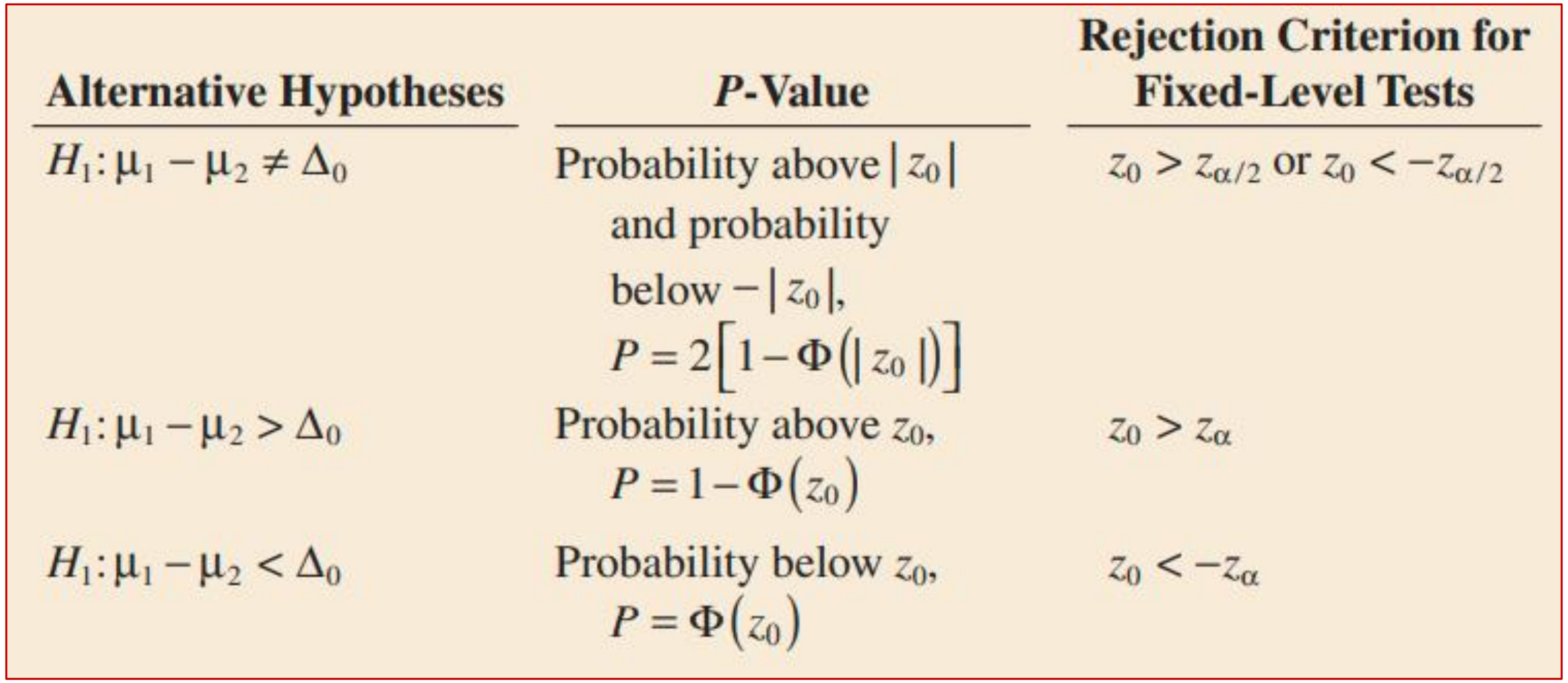
has a $N(0, 1)$ distribution.

Tests on the Difference in Means, Variances Known

Null hypothesis and test statistic:

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

$$Z_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$



Alternative Hypotheses	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert z_0 \rvert$ and probability below $-\lvert z_0 \rvert$, $P = 2\left[1 - \Phi\left(\lvert z_0 \rvert\right)\right]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

Ex. A manager evaluates effectiveness of a major hardware upgrade by running a certain process 50 times before the upgrade and 50 times after it. Based on these data, the average running time is 8.5 minutes before the upgrade, 7.2 minutes after it. Historically, the standard deviation has been 1.8 minutes, and presumably it has not changed. Test on the difference between the means of running time. Use α = 0.05.

Confidence Interval

$$P\left(\bar{X}_1-\bar{X}_2-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\leq\mu_1-\mu_2\leq\bar{X}_1-\bar{X}_2+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)=1-\alpha$$

Ex. A manager evaluates effectiveness of a major hardware upgrade by running a certain process 50 times before the upgrade and 50 times after it. Based on these data, the average running time is 8.5 minutes before the upgrade, 7.2 minutes after it. Historically, the standard deviation has been 1.8 minutes, and presumably it has not changed. Construct a 95% confidence interval showing how much the mean running time reduced due to the hardware upgrade.

Choice of Sample Size

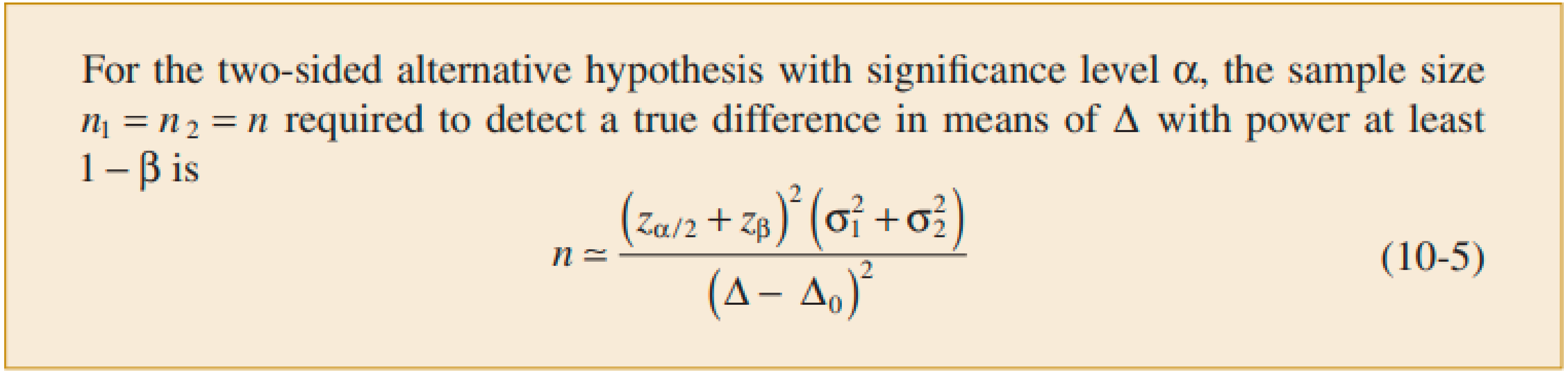
The sample size required so that the error in estimating $\mu_1 - \mu_2$ by x_1 - x_2 will be less than E at $100(1 - \alpha)\%$ confidence.

$$n_1 = n_2 = n = \left[\left(\frac{z_{\alpha/2}}{E} \right)^2 \left(\sigma_1^2 + \sigma_2^2 \right) \right]$$

Ex. A manager evaluates effectiveness of a major hardware upgrade by running a certain process 50 times before the upgrade and 50 times after it. Based on these data, the average running time is 8.5 minutes before the upgrade, 7.2 minutes after it. Historically, the standard deviation has been 1.8 minutes, and presumably it has not changed. What sample size would be required in each population if you wanted to be 95% confident that the error in estimating the difference in mean running time is less than 0.5 minute?

Sample Size Formulas

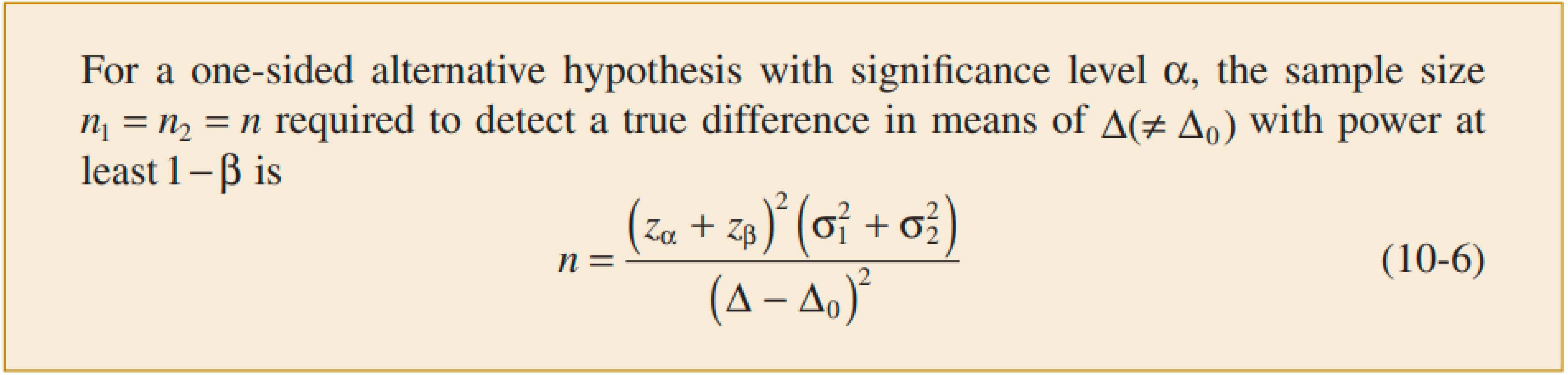
Sample Size for a Two-Sided Test on the Difference in Means with $n_1 = n_2$, Variances Known



For the two-sided alternative hypothesis with significance level α, the sample size $n_1 = n_2 = n$ required to detect a true difference in means of Δ with power at least $1 - \beta$ is

$$n \simeq \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 \left(\sigma_1^2 + \sigma_2^2\right)}{\left(\Delta - \Delta_0\right)^2} \tag{10-5}$$

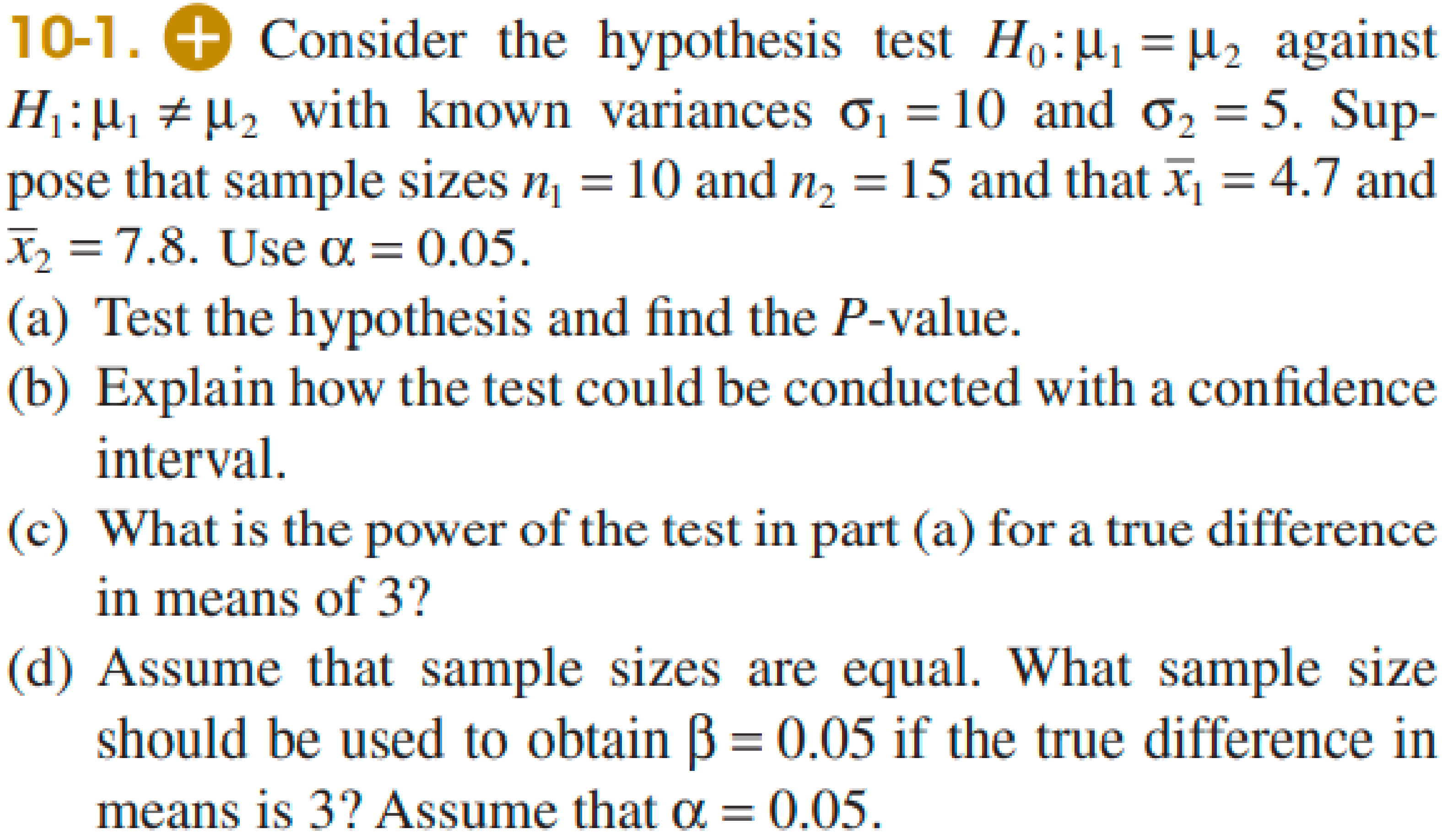
Sample Size for a One-Sided Test on the Difference in Means with $n_1 = n_2$, Variances Known



For a one-sided alternative hypothesis with significance level α, the sample size $n_1 = n_2 = n$ required to detect a true difference in means of $\Delta(\neq \Delta_0)$ with power at least $1 - \beta$ is

$$n = \frac{\left(z_{\alpha} + z_{\beta}\right)^2 \left(\sigma_1^2 + \sigma_2^2\right)}{\left(\Delta - \Delta_0\right)^2} \tag{10-6}$$

Exercise



10-1. Consider the hypothesis test $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$ with known variances $\sigma_1 = 10$ and $\sigma_2 = 5$. Suppose that sample sizes $n_1 = 10$ and $n_2 = 15$ and that $\overline{x}_1 = 4.7$ and $\overline{x}_2 = 7.8$. Use $\alpha = 0.05$.

(a) Test the hypothesis and find the *P*-value.

(b) Explain how the test could be conducted with a confidence interval.

(c) What is the power of the test in part (a) for a true difference in means of 3?

(d) Assume that sample sizes are equal. What sample size should be used to obtain $\beta = 0.05$ if the true difference in means is 3? Assume that $\alpha = 0.05$.

Unknown σ^2

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

$$Z_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- What if variance σ^2 is unknown?

➔ Use large sample size (n_1 and $n_2 \geqslant 30$) and sample variances s_1^2 s_2^2 instead of population variances σ_1^2, σ_2^2 which are unknown.

Ex. A test is conducted to compare breaking strength of cell phones manufactured by two companies. Summary data are given below.

Company A: $n_1 = 65$, $\overline{x}_1 = 107$ pounds, $s_1 = 10$ pounds

Company B: $n_2 = 60$, $\overline{x}_2 = 113$ pounds, $s_2 = 13$ pounds

Use $\alpha = 0.05$.

Example

- Parameter of interest: $\mu_1 - \mu_2$ and $\Delta_0 = 0$
- H_0: $\mu_1 - \mu_2 = \Delta_0 = 0$
- H_1: $\mu_1 < \mu_2$
- Test statistic: $z_0 = \dfrac{\overline{x}_1 - \overline{x}_2 - 0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{107 - 113}{\sqrt{\dfrac{10^2}{65} + \dfrac{13^2}{60}}} = -2.875$

- P-value = $P(Z < -2.875) = 0.002$
- Conclusion: Reject H_0 at the $\alpha = 0.05$ level

Unknown variances, large samples

Ex. Internet connections are often slowed by delays at nodes. Let us determine if the delay time increases during heavy-volume times. Five hundred packets are sent through the same network between 5 pm and 6 pm (sample X_1), and three hundred packets are sent between 10 pm and 11 pm (sample X_2). The early sample has a mean delay time of 0.8 sec with a standard deviation of 0.1 sec whereas the second sample has a mean delay time of 0.5 sec with a standard deviation of 0.08 sec. Construct a 95% confidence interval for the difference between the mean delay times.

Unknown variances, large samples

Ex. Internet connections are often slowed by delays at nodes. Let us determine if the delay time increases during heavy-volume times. Five hundred packets are sent through the same network between 5 pm and 6 pm (sample X_1), and three hundred packets are sent between 10 pm and 11 pm (sample X_2). The early sample has a mean delay time of 0.8 sec with a standard deviation of 0.1 sec whereas the second sample has a mean delay time of 0.5 sec with a standard deviation of 0.08 sec. Test on the difference between the mean delay times at 0.05 level of significance.

Exercise

Grades of two classes:

Class A: $n_1 = 30$, $\bar{x}_A = 6.2$, $s_1 = 0.8$

Class B: $n_2 = 31$, $\bar{x}_B = 5.8$, $s_2 = 0.6$

a/ Test H_0: $\mu_1 = \mu_2$

vs H_1: H_0: $\mu_1 > \mu_2$

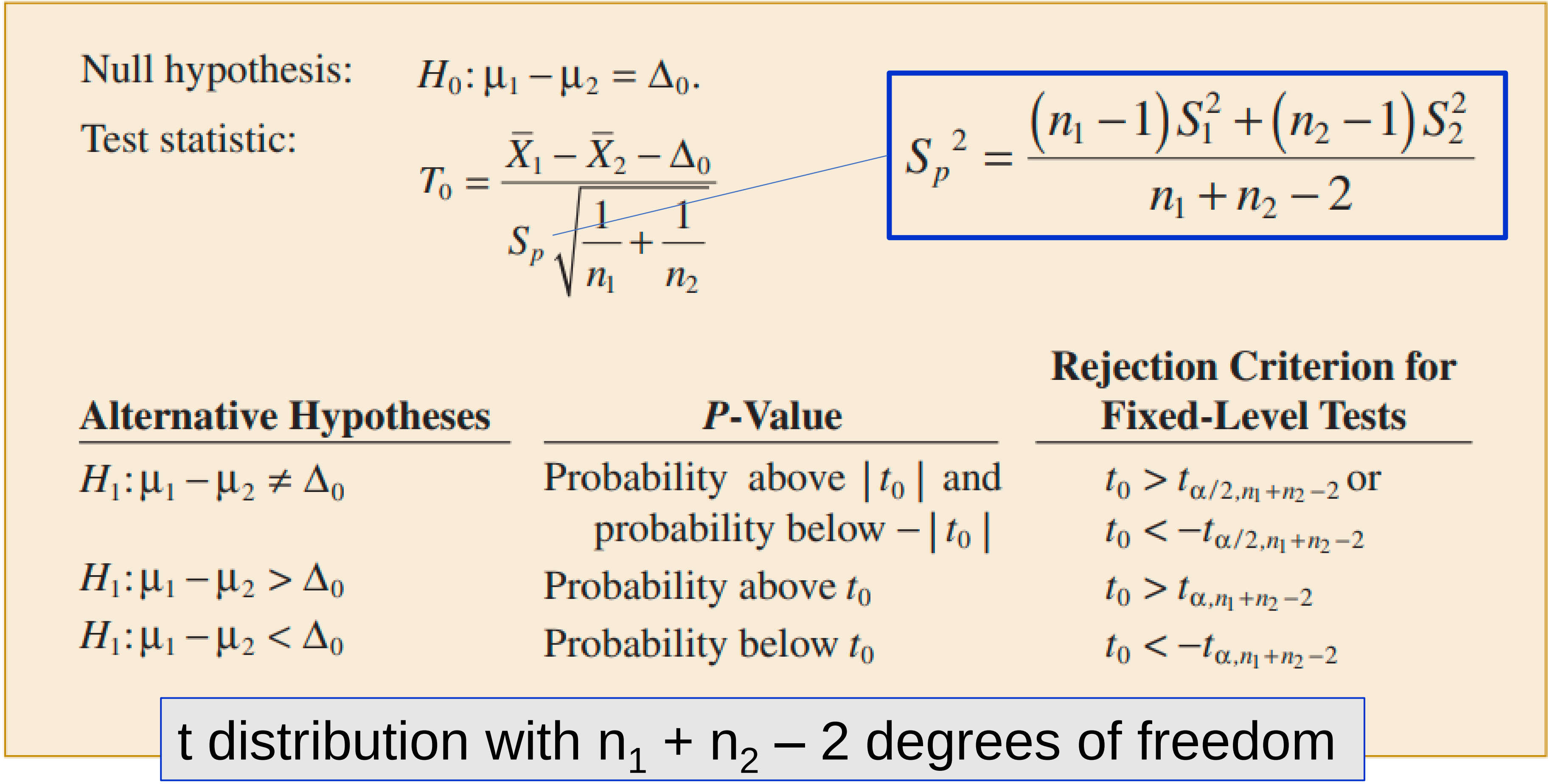
at the $\alpha = 0.05$ level of significance

b/ Construct 95% confidence interval for $\mu_1 - \mu_2$

Unknown variance, small samples

Tests on the Difference in Means of Two Normal Distributions, Variances Unknown and Equal*

$\sigma_1 = \sigma_2 = \sigma$



Null hypothesis: $H_0: \mu_1 - \mu_2 = \Delta_0$.

Test statistic:

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\boxed{S_p{}^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}}$$

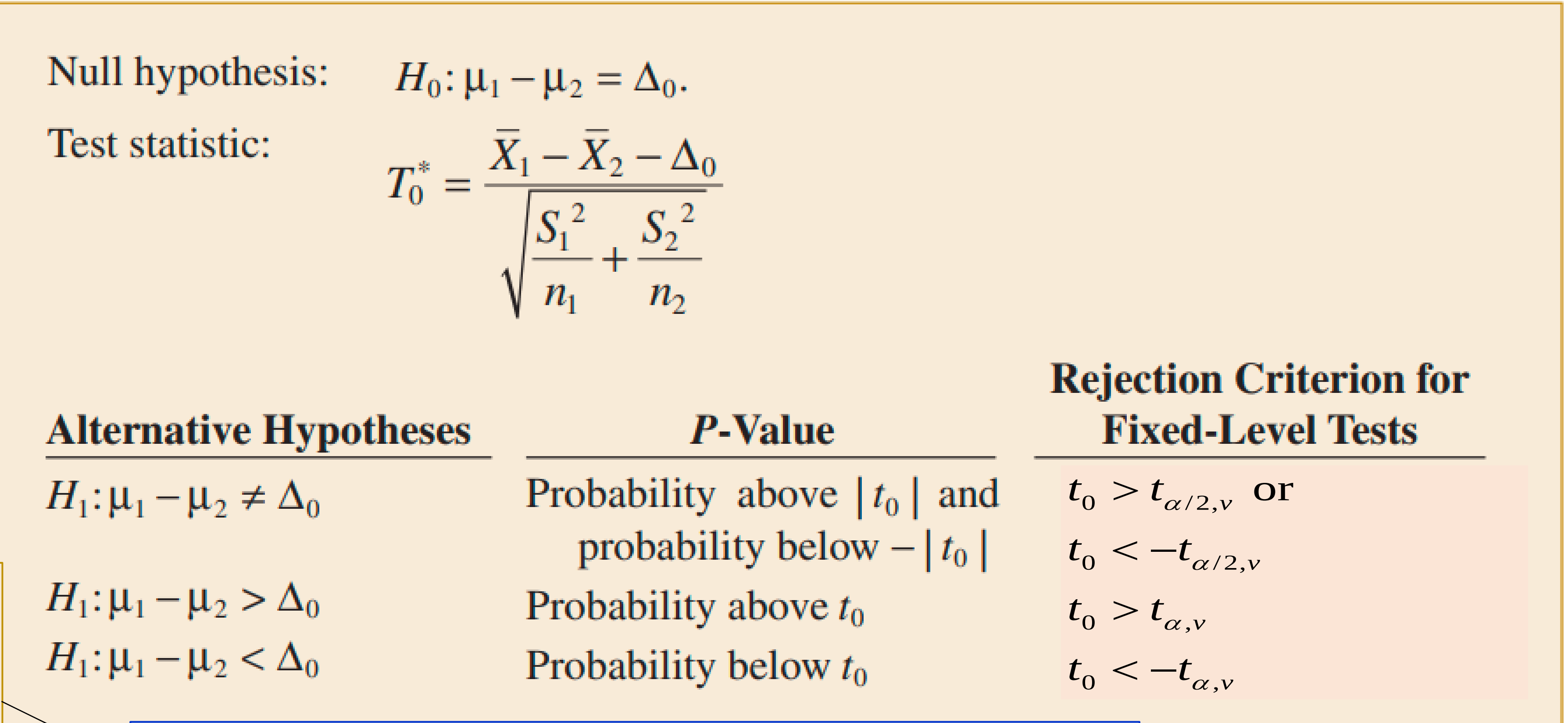
Alternative Hypotheses	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert t_0 \rvert$ and probability below $-\lvert t_0 \rvert$	$t_0 > t_{\alpha/2, n_1+n_2-2}$ or $t_0 < -t_{\alpha/2, n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probability above t_0	$t_0 > t_{\alpha, n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probability below t_0	$t_0 < -t_{\alpha, n_1+n_2-2}$

t distribution with $n_1 + n_2 - 2$ degrees of freedom

Unknown variance, small samples

Case 2: Test Statistic for the Difference in Means, Variances Unknown and Not Assumed Equal

$\sigma_1 \neq \sigma_2$

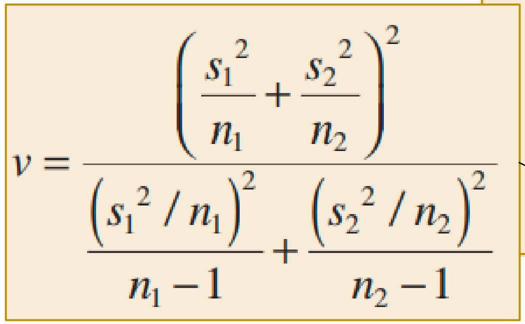


Null hypothesis: $H_0\colon \mu_1 - \mu_2 = \Delta_0.$

Test statistic:

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{S_1^{\,2}}{n_1} + \dfrac{S_2^{\,2}}{n_2}}}$$

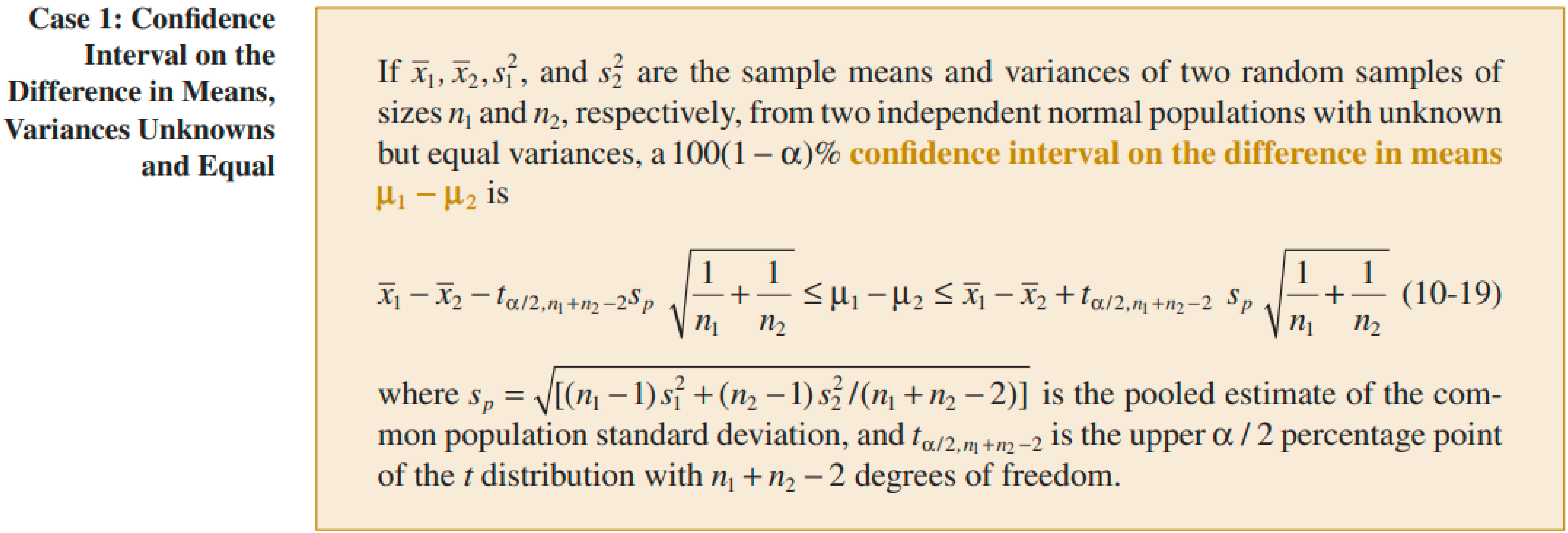
Alternative Hypotheses	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1\colon \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert t_0 \rvert$ and probability below $-\lvert t_0 \rvert$	$t_0 > t_{\alpha/2,v}$ or $t_0 < -t_{\alpha/2,v}$
$H_1\colon \mu_1 - \mu_2 > \Delta_0$	Probability above t_0	$t_0 > t_{\alpha,v}$
$H_1\colon \mu_1 - \mu_2 < \Delta_0$	Probability below t_0	$t_0 < -t_{\alpha,v}$



$$v = \frac{\left(\dfrac{s_1^{\,2}}{n_1} + \dfrac{s_2^{\,2}}{n_2}\right)^2}{\dfrac{\left(s_1^{\,2}/n_1\right)^2}{n_1 - 1} + \dfrac{\left(s_2^{\,2}/n_2\right)^2}{n_2 - 1}}$$

t distribution with ν degrees of freedom

Confidence Interval



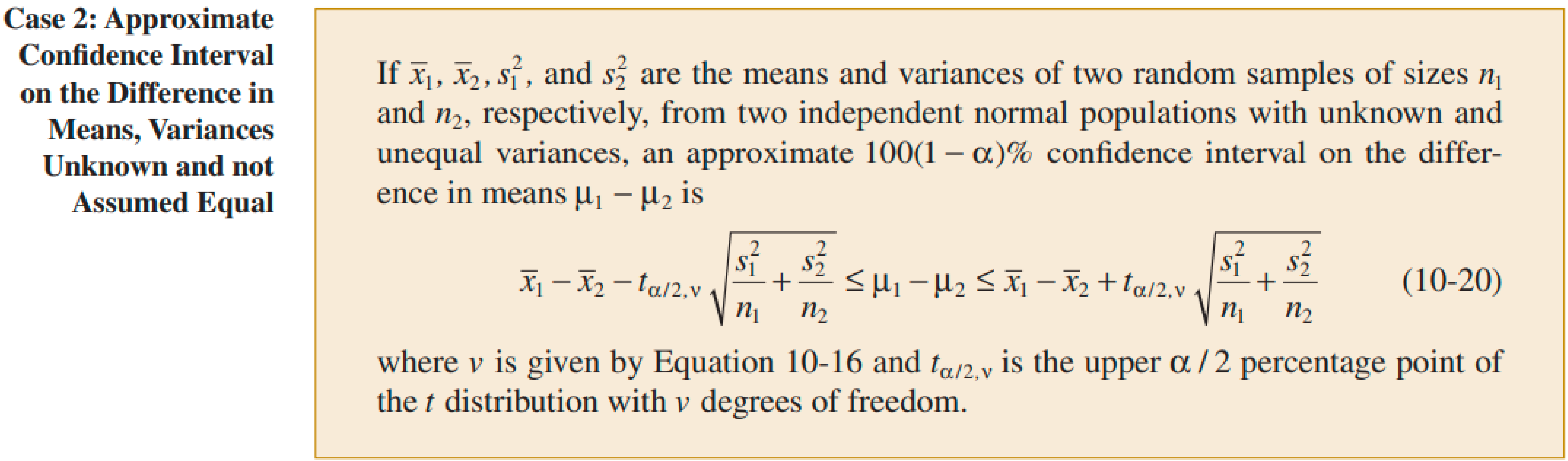
Case 1: Confidence Interval on the Difference in Means, Variances Unknowns and Equal

If $\bar{x}_1, \bar{x}_2, s_1^2$, and s_2^2 are the sample means and variances of two random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown but equal variances, a 100(1 – α)% **confidence interval on the difference in means $\mu_1 - \mu_2$** is

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2, n_1+n_2-2}\ s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (10\text{-}19)$$

where $s_p = \sqrt{[(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2 / (n_1 + n_2 - 2)]}$ is the pooled estimate of the common population standard deviation, and $t_{\alpha/2, n_1+n_2-2}$ is the upper $\alpha / 2$ percentage point of the *t* distribution with $n_1 + n_2 - 2$ degrees of freedom.

Confidence Interval

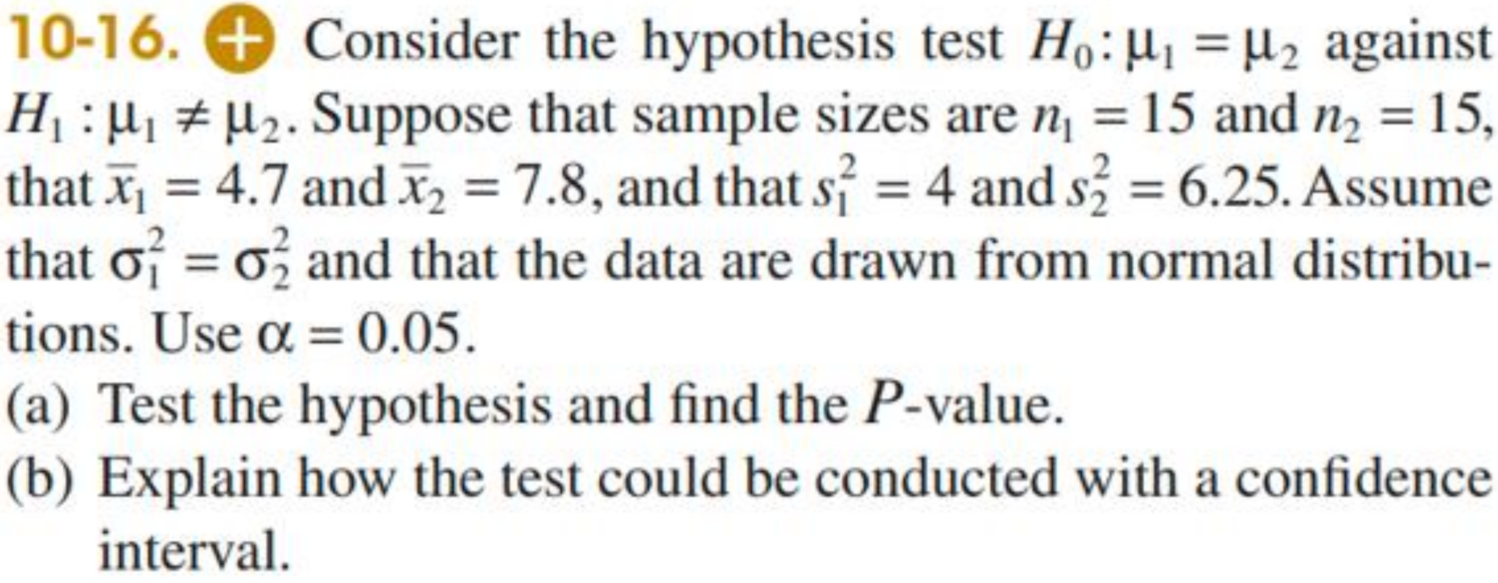


Case 2: Approximate Confidence Interval on the Difference in Means, Variances Unknown and not Assumed Equal

If $\bar{x}_1$, $\bar{x}_2$, s_1^2, and s_2^2 are the means and variances of two random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown and unequal variances, an approximate $100(1-\alpha)\%$ confidence interval on the difference in means $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \qquad (10\text{-}20)$$

where v is given by Equation 10-16 and $t_{\alpha/2,\nu}$ is the upper $\alpha/2$ percentage point of the t distribution with v degrees of freedom.



10-16. ⊕ Consider the hypothesis test $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$. Suppose that sample sizes are $n_1 = 15$ and $n_2 = 15$, that $\overline{x}_1 = 4.7$ and $\overline{x}_2 = 7.8$, and that $s_1^2 = 4$ and $s_2^2 = 6.25$. Assume that $\sigma_1^2 = \sigma_2^2$ and that the data are drawn from normal distributions. Use $\alpha = 0.05$.

(a) Test the hypothesis and find the P-value.

(b) Explain how the test could be conducted with a confidence interval.

On the difference of proportions

Parameter of interest: $\theta = p_1 - p_2$

Estimated by: $\hat{\theta} = \hat{p}_1 - \hat{p}_2$

Its standard error: $$\sigma(\hat{\theta}) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Estimated by: $$s(\hat{\theta}) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Confidence interval for the difference of proportions

$$\boxed{\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

Approximate Tests on the Difference of Two Population Proportions

Null hypothesis: $H_0\colon p_1 = p_2$

Test statistic: $$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}\left(1-\hat{P}\right)\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} \qquad (10\text{-}35)$$

$$\boxed{P = \frac{X_1 + X_2}{n_1 + n_2}}$$

Alternative Hypothesis	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1\colon p_1 \neq p_2$	Probability above $\lvert z_0 \rvert$ and probability below $-\lvert z_0 \rvert$. $P = 2\left[1-\Phi\left(\lvert z_0 \rvert\right)\right]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1\colon p_1 > p_2$	Probability above z_0. $P = 1-\Phi\left(z_0\right)$	$z_0 > z_\alpha$
$H_1\colon p_1 < p_2$	Probability below z_0. $P = \Phi\left(z_0\right)$	$z_0 < -z_\alpha$

10-82. Consider the following computer output.

Test and CI for Two Proportions

```
Sample     X      N       Sample p
   1       54     250      0.216000
   2       60     290      0.206897
Difference = p(1) - p(2)
Estimate for difference: 0.00910345
95% CI for difference: (-0.0600031,
0.0782100)
Test for difference
 =0(vs not = 0): Z = ? P-Value = ?
```

(a) Is this a one-sided or a two-sided test?

(b) Fill in the missing values.

(c) Can the null hypothesis be rejected?

(d) Construct an approximate 90% CI for the difference in the two proportions.

Summary

- On difference of *two population mean/proportions:*
  - *confidence intervals*
  - *test of hypotheses*
- Discuss on sample sizes