Chapter 7: Inference for Distributions

MATH 560-01
Statistical Data Analysis

February 22, 2021
(Last corrected: 2/22)

These slides are based on material from *Introduction to the Practice of Statistics* by David S. Moore, George P. McCabe, and Bruce A. Craig, 9th edition.

Sections

7.1 Inference for the Mean of a Population

7.2 Comparing Two Means

7.1 Inference for the Mean of a Population

After completing this section, students should be able to:

- use a t-table to find *percentile*-related values for t-distributions

7.1 Inference for the Mean of a Population

After completing this section, students should be able to:

- use a t-table to find *percentile*-related values for t-distributions
- compute *confidence intervals for the population mean* μ when the population standard deviation σ is unknown

7.1 Inference for the Mean of a Population

After completing this section, students should be able to:

- use a t-table to find *percentile*-related values for t-distributions
- compute *confidence intervals for the population mean* μ when the population standard deviation σ is unknown
- use a t-table to obtain bounds on probabilities for random variables following t-distributions

7.1 Inference for the Mean of a Population

After completing this section, students should be able to:

- use a t-table to find *percentile*-related values for t-distributions
- compute *confidence intervals for the population mean* μ when the population standard deviation σ is unknown
- use a t-table to obtain bounds on probabilities for random variables following t-distributions
- perform *one-sided* or *two-sided* significance tests involving the population mean μ when the population standard deviation σ is unknown

7.1 Inference for the Mean of a Population

After completing this section, students should be able to:

- use a t-table to find *percentile*-related values for t-distributions
- compute *confidence intervals for the population mean* μ when the population standard deviation σ is unknown
- use a t-table to obtain bounds on probabilities for random variables following t-distributions
- perform *one-sided* or *two-sided* significance tests involving the population mean μ when the population standard deviation σ is unknown
- perform *paired t-tests* for equality of the means of two populations based on a matched pairs design

7.1 Inference for the Mean of a Population

- In Chapter 6, we considered inference procedures for the unknown population mean μ under the unrealistic assumption that the population standard deviation σ is known.

7.1 Inference for the Mean of a Population

- In Chapter 6, we considered inference procedures for the unknown population mean μ under the unrealistic assumption that the population standard deviation σ is known.
- In this chapter, we consider how to make inferences about μ in the more realistic setting where σ is unknown so that we need to use the standard deviation of the sample

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

to estimate it.

7.1 Inference for the Mean of a Population

- If a population has a $N(\mu, \sigma)$ distribution, then the sampling distribution of $\bar{x}$ follows a $N(\mu, \frac{\sigma}{\sqrt{n}})$ distribution.

7.1 Inference for the Mean of a Population

- If a population has a $N(\mu, \sigma)$ distribution, then the sampling distribution of $\bar{x}$ follows a $N(\mu, \frac{\sigma}{\sqrt{n}})$ distribution.
- The methods in Chapter 6 are based on the fact that $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ then follows a standard normal distribution.

7.1 Inference for the Mean of a Population

- If a population has a $N(\mu, \sigma)$ distribution, then the sampling distribution of $\bar{x}$ follows a $N(\mu, \frac{\sigma}{\sqrt{n}})$ distribution.
- The methods in Chapter 6 are based on the fact that $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ then follows a standard normal distribution.
- When σ is unknown, we cannot use these methods and we must replace σ with s.

7.1 Inference for the Mean of a Population

- If a population has a $N(\mu, \sigma)$ distribution, then the sampling distribution of $\bar{x}$ follows a $N(\mu, \frac{\sigma}{\sqrt{n}})$ distribution.
- The methods in Chapter 6 are based on the fact that $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ then follows a standard normal distribution.
- When σ is unknown, we cannot use these methods and we must replace σ with s.
- The random variable

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

follows a different distribution called a t **distribution with** $n - 1$ **degrees of freedom**.

7.1 Inference for the Mean of a Population

- If a population has a $N(\mu, \sigma)$ distribution, then the sampling distribution of $\bar{x}$ follows a $N(\mu, \frac{\sigma}{\sqrt{n}})$ distribution.
- The methods in Chapter 6 are based on the fact that $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ then follows a standard normal distribution.
- When σ is unknown, we cannot use these methods and we must replace σ with s.
- The random variable

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

  follows a different distribution called a t **distribution with** $n-1$ **degrees of freedom**.
- The denominator $\frac{s}{\sqrt{n}}$ is sometimes called the **standard error** of t.

7.1 Inference for the Mean of a Population

t distribution

- The density curve of the t distribution with df degrees of freedom has the form

$$f(t) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{\pi df}\,\Gamma\left(\frac{df}{2}\right)\left(1+\frac{t^2}{df}\right)^{\frac{df+1}{2}}}, t \in \mathbb{R}.$$

7.1 Inference for the Mean of a Population

<u>t distribution</u>

- The density curve of the t distribution with df degrees of freedom has the form

$$f(t) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{\pi df}\,\Gamma\left(\frac{df}{2}\right)\left(1+\frac{t^2}{df}\right)^{\frac{df+1}{2}}}, t \in \mathbb{R}.$$

- This curve is symmetric about 0 and unimodal. When $df > 1$, the mean of this distribution is 0. (For $df = 1$, the expected value $\int_{-\infty}^{\infty} t f(t)\, dt$ does not exist.)

7.1 Inference for the Mean of a Population

t distribution

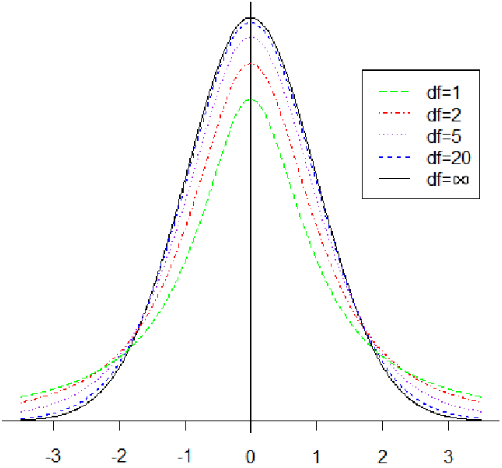
- The density curve of the t distribution with df degrees of freedom has the form

$$f(t) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{\pi df}\,\Gamma\left(\frac{df}{2}\right)\left(1+\frac{t^2}{df}\right)^{\frac{df+1}{2}}}, t \in \mathbb{R}.$$

- This curve is symmetric about 0 and unimodal. When $df > 1$, the mean of this distribution is 0. (For $df = 1$, the expected value $\int_{-\infty}^{\infty} t f(t)\, dt$ does not exist.)
- Sometimes, we abbreviate the t distribution with df degrees of freedom by $t(df)$.

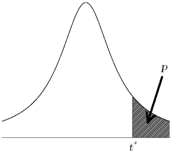
7.1 Inference for the Mean of a Population

<u>t distribution</u>



7.1 Inference for the Mean of a Population

Table for t distribution



t distribution critical values

	Upper-tail probability p								
df	.25	.20	.15	.10	.05	.025	.02	.01	.005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.90	31.82	63.66
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841
					⋮				
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581
∞	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576
	50%	60%	70%	80%	90%	95%	96%	98%	99%
	Confidence level C								

7.1 Inference for the Mean of a Population

Confidence interval for μ when σ is unknown and the population is normal

A level C confidence interval for μ is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the value for the $t(n-1)$ density with area C between $-t^*$ and t^*.

The confidence interval is approximately correct when n is large even when the population is not normal.

7.1 Inference for the Mean of a Population

Example 7A: Suppose we want information on the number of hours that U.S. college students who have mobile phones watch videos on their phone. Here is a SRS with the number of hours for $n = 8$ students:

$$7 \quad 9 \quad 1 \quad 6 \quad 13 \quad 10 \quad 3 \quad 5$$

Find a 95% two-sided confidence interval for the average number of hours that all U.S. college students who have mobile phones watch videos.

7.1 Inference for the Mean of a Population

- *Answer*:

7.1 Inference for the Mean of a Population

Two-sided one sample t-test for a population mean when σ is unknown and the population is normal

1. Test $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$

7.1 Inference for the Mean of a Population

Two-sided one sample t-test for a population mean when σ is unknown and the population is normal

1. Test $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$
2. Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

7.1 Inference for the Mean of a Population

Two-sided one sample t-test for a population mean when σ is unknown and the population is normal

1. Test $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$
2. Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
3. Use the t-table to find the value t^* such that $P(T > t^*) = \frac{\alpha}{2}$ where T follows a t distribution with $n - 1$ degrees of freedom.

7.1 Inference for the Mean of a Population

Two-sided one sample t-test for a population mean when σ is unknown and the population is normal

1. Test $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$
2. Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
3. Use the t-table to find the value t^* such that $P(T > t^*) = \frac{\alpha}{2}$ where T follows a t distribution with $n - 1$ degrees of freedom.
4. Reject H_0 if $|t| > t^*$

7.1 Inference for the Mean of a Population

One-sided (left-sided) one sample t-test for a population mean when σ is unknown and the population is normal

1. Test $H_0 : \mu = \mu_0$ vs. $H_a : \mu < \mu_0$
2. Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
3. Use the t-table to find the value t^* such that $P(T < -t^*) = P(T > t^*) = \alpha$ where T follows a t distribution with $n - 1$ degrees of freedom.
4. Reject H_0 if $t < -t^*$

7.1 Inference for the Mean of a Population

One-sided (right-sided) one sample t-test for a population mean when σ is unknown and the population is normal

1. Test $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$
2. Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
3. Use the t-table to find the value t^* such that $P(T > t^*) = \alpha$ where T follows a t distribution with $n - 1$ degrees of freedom.
4. Reject H_0 if $t > t^*$

7.1 Inference for the Mean of a Population

Example 7B: A population follows a Normal distribution with unknown mean μ and unknown standard deviation σ. A SRS of size $n = 25$ is taken from this population to perform a significance test of $H_0 : \mu = 10$ versus $H_a : \mu < 10$. The observed sample statistics are $\bar{x} = 8$ and $s = 3$.
(a) Calculate an appropriate test statistic.
(b) Locate the two critical values t^* from the t-table that bracket t.
(c) Between what two values does the P-value fall?
(d) Should H_0 be rejected at level $\alpha = .01$?

7.1 Inference for the Mean of a Population

- *Answer*:

7.1 Inference for the Mean of a Population

Paired t-test

- When designing experiments, comparative designs are often more useful than single sample studies.

7.1 Inference for the Mean of a Population

Paired t-test

- When designing experiments, comparative designs are often more useful than single sample studies.
- However, the one sample t-test procedure is useful when subjects are matched in pairs and their outcomes are compared within each matched pair.

7.1 Inference for the Mean of a Population

Paired t-test

- When designing experiments, comparative designs are often more useful than single sample studies.
- However, the one sample t-test procedure is useful when subjects are matched in pairs and their outcomes are compared within each matched pair.
- For example, if we take measurements on the same subjects under two different experimental conditions, then this is a **matched pairs design**.

7.1 Inference for the Mean of a Population

Paired t-test

- When designing experiments, comparative designs are often more useful than single sample studies.
- However, the one sample t-test procedure is useful when subjects are matched in pairs and their outcomes are compared within each matched pair.
- For example, if we take measurements on the same subjects under two different experimental conditions, then this is a **matched pairs design**.
- In analyzing data from a matched pairs design, x_i is the difference between the two measurements for the ith subject, and we use one-sample procedures (confidence intervals or t-tests to make inferences).

Example 7C: Consider the following study to compare two popular energy drinks. For each subject, a coin was flipped to determine which drink to rate first. Each drink was rated on a 0 to 100 scale, with 100 being the highest rating.

	Subject				
Drink	1	2	3	4	5
A	43	83	66	89	78
B	45	78	64	79	71

(a) Is there a difference in preferences at level $\alpha = .05$? State appropriate hypotheses, calculate an appropriate test statistic, compute the P-value or critical value, and state your conclusion.
(b) Compute a 95% two-sided confidence interval for the difference in preference between Drink A and Drink B.

7.1 Inference for the Mean of a Population

- *Answer*:

7.1 Inference for the Mean of a Population

Robustness of t procedures

- A procedure is **robust** if the probability calculations are insensitive to violations of the assumptions made.

7.1 Inference for the Mean of a Population

Robustness of t procedures

- A procedure is **robust** if the probability calculations are insensitive to violations of the assumptions made.
- The t procedures are usually robust against non-Normality of the population; outliers and strong skewness do affect the accuracy of the procedures, though.

7.1 Inference for the Mean of a Population

Robustness of t procedures

- A procedure is **robust** if the probability calculations are insensitive to violations of the assumptions made.
- The t procedures are usually robust against non-Normality of the population; outliers and strong skewness do affect the accuracy of the procedures, though.
- As the sample size grows, the CLT implies the distribution of $\bar{x}$ is close to normal and s will be an accurate estimate of σ.

7.1 Inference for the Mean of a Population

Robustness of t procedures

- A procedure is **robust** if the probability calculations are insensitive to violations of the assumptions made.
- The t procedures are usually robust against non-Normality of the population; outliers and strong skewness do affect the accuracy of the procedures, though.
- As the sample size grows, the CLT implies the distribution of $\bar{x}$ is close to normal and s will be an accurate estimate of σ.
- Except when the sample size is small, the assumption that the data are a SRS is more crucial than that the population is Normal.

7.2 Comparing Two Means

After completing this section, students should be able to:

- compute *confidence intervals for the difference* $\mu_1 - \mu_2$ *between population means* when the population standard deviations σ_1 and σ_2 are unknown but equal

7.2 Comparing Two Means

After completing this section, students should be able to:

- compute *confidence intervals for the difference* $\mu_1 - \mu_2$ *between population means* when the population standard deviations σ_1 and σ_2 are unknown but equal
- perform *one-sided* or *two-sided tests* for equality of the means of two populations when the population standard deviations σ_1 and σ_2 are unknown but equal based on random samples from each population

7.2 Comparing Two Means

After completing this section, students should be able to:

- compute *confidence intervals for the difference* $\mu_1 - \mu_2$ *between population means* when the population standard deviations σ_1 and σ_2 are unknown but equal
- perform *one-sided* or *two-sided tests* for equality of the means of two populations when the population standard deviations σ_1 and σ_2 are unknown but equal based on random samples from each population
- compute confidence intervals for the difference $\mu_1 - \mu_2$ between population means when the population standard deviations σ_1 and σ_2 are unknown and not necessarily equal based on *Satterthwaite's Approximation*

7.2 Comparing Two Means

After completing this section, students should be able to:

- compute *confidence intervals for the difference* $\mu_1 - \mu_2$ *between population means* when the population standard deviations σ_1 and σ_2 are unknown but equal
- perform *one-sided* or *two-sided tests* for equality of the means of two populations when the population standard deviations σ_1 and σ_2 are unknown but equal based on random samples from each population
- compute confidence intervals for the difference $\mu_1 - \mu_2$ between population means when the population standard deviations σ_1 and σ_2 are unknown and not necessarily equal based on *Satterthwaite's Approximation*
- perform one-sided or two-sided tests for equality of the means of two populations when the population standard deviations σ_1 and σ_2 are unknown and not necessarily equal based on *Satterthwaite's Approximation*

7.2 Comparing Two Means

- In two sample problems, the goal of inference is to compare the responses in the two groups.

7.2 Comparing Two Means

- In two sample problems, the goal of inference is to compare the responses in the two groups.
- Each group is considered to be a sample from a distinct population.

7.2 Comparing Two Means

- In two sample problems, the goal of inference is to compare the responses in the two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group.

7.2 Comparing Two Means

- Suppose we have two populations. The first population has a $N(\mu_1, \sigma_1)$ distribution and the second population has a $N(\mu_2, \sigma_2)$ population.

7.2 Comparing Two Means

- Suppose we have two populations. The first population has a $N(\mu_1, \sigma_1)$ distribution and the second population has a $N(\mu_2, \sigma_2)$ population.
- We take a random sample of size n_1 from population 1 and compute the sample mean $\bar{x}_1$, and we take a random sample of size n_2 from population 2 and compute the sample mean $\bar{x}_2$.

7.2 Comparing Two Means

- Suppose we have two populations. The first population has a $N(\mu_1, \sigma_1)$ distribution and the second population has a $N(\mu_2, \sigma_2)$ population.
- We take a random sample of size n_1 from population 1 and compute the sample mean $\bar{x}_1$, and we take a random sample of size n_2 from population 2 and compute the sample mean $\bar{x}_2$.
- The samples from the two populations are taken independently.

7.2 Comparing Two Means

- Suppose we have two populations. The first population has a $N(\mu_1, \sigma_1)$ distribution and the second population has a $N(\mu_2, \sigma_2)$ population.
- We take a random sample of size n_1 from population 1 and compute the sample mean $\bar{x}_1$, and we take a random sample of size n_2 from population 2 and compute the sample mean $\bar{x}_2$.
- The samples from the two populations are taken independently.
- We are interested in making inferences about the difference between the population means, $\mu_1 - \mu_2$.

7.2 Comparing Two Means

- We use $\bar{x}_1 - \bar{x}_2$ to estimate $\mu_1 - \mu_2$.

7.2 Comparing Two Means

- We use $\bar{x}_1 - \bar{x}_2$ to estimate $\mu_1 - \mu_2$.
- The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is $N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$ since

7.2 Comparing Two Means

- We use $\bar{x}_1 - \bar{x}_2$ to estimate $\mu_1 - \mu_2$.
- The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is $N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$ since $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$

7.2 Comparing Two Means

- We use $\bar{x}_1 - \bar{x}_2$ to estimate $\mu_1 - \mu_2$.
- The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is $N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$ since
  $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$
  $\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

7.2 Comparing Two Means

- We use $\bar{x}_1 - \bar{x}_2$ to estimate $\mu_1 - \mu_2$.
- The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is $N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$ since
  $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$
  $\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.
- When σ_1 and σ_2 are unknown, we must replace them by estimates from the sample and use

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

7.2 Comparing Two Means

If we make the additional assumption that $\sigma_1 = \sigma_2$ and use the pooled estimate $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$, then

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

follows a $t(n_1 + n_2 - 2)$ distribution.

7.2 Comparing Two Means

Confidence interval for $\mu_1 - \mu_2$ when $\sigma_1 = \sigma_2$ are unknown and the populations are normal

A level C confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where t^* is the value for the $t(n_1 + n_2 - 2)$ density with area C between $-t^*$ and t^*.

The confidence interval is approximately correct when n_1 and n_2 are large even when the population is not normal.

7.2 Comparing Two Means

Two-sided two sample t-test for equality of population means when $\sigma_1 = \sigma_2$ is unknown and the population is normal

1. Test $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$
2. Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where
$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$
3. Use the t-table to find the value t^* such that $P(T > t^*) = \frac{\alpha}{2}$ where T follows a t distribution with $n_1 + n_2 - 2$ degress of freedom.
4. Reject H_0 if $|t| > t^*$

7.2 Comparing Two Means

One-sided (left-sided) two sample t-test for equality of population means when $\sigma_1 = \sigma_2$ is unknown and the population is normal

1. Test $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 < \mu_2$
2. Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where
$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$
3. Use the t-table to find the value t^* such that $P(T < -t^*) = P(T > t^*) = \alpha$ where T follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom.
4. Reject H_0 if $t < -t^*$

7.2 Comparing Two Means

<u>One-sided (right-sided) two sample t-test for equality of population means when $\sigma_1 = \sigma_2$ is unknown and the population is normal</u>

1. Test $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 > \mu_2$
2. Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where
$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$
3. Use the t-table to find the value t^* such that $P(T > t^*) = \alpha$ where T follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom.
4. Reject H_0 if $t > t^*$

7.2 Comparing Two Means

Example 7D (Book p.464): To assess the effect of calcium intake on blood pressure, an experiment was conducted on a Calcium and Placebo group, and the change in blood pressure was recorded from the beginning and end of the study for each group. Here are the summary statistics for the decrease in blood pressure:

Group	Treatment	n	$\bar{x}$	s
1	Calcium	10	5.000	8.743
2	Placebo	11	-0.273	5.901

Assume that the standard deviations of both populations are the same. Should $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 > \mu_2$ be rejected at level $\alpha = .05$? Carefully show all steps of your test (i.e., calculate an appropriate test statistic, compute the P-value or critical value, and state the conclusion).

7.2 Comparing Two Means

- *Answer*:

7.2 Comparing Two Means

- Unfortunately, the exact distribution of the random variable t is not as simple when $\sigma_1 \neq \sigma_2$.

7.2 Comparing Two Means

- Unfortunately, the exact distribution of the random variable t is not as simple when $\sigma_1 \neq \sigma_2$.
- (Satterthwaite's Approximation) It approximately follows a $t(k)$ distribution where

$$k = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}.$$

7.2 Comparing Two Means

- Unfortunately, the exact distribution of the random variable t is not as simple when $\sigma_1 \neq \sigma_2$.
- (Satterthwaite's Approximation) It approximately follows a $t(k)$ distribution where

$$k = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}.$$

- A more conservative approach uses $k = \min\{n_1 - 1, n_2 - 1\}$.

7.2 Comparing Two Means

Confidence interval for $\mu_1 - \mu_2$ when $\sigma_1 \neq \sigma_2$ are unknown and the populations are normal

A level C confidence interval for $\mu_1 - \mu_2$ is

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

where t^* is approximated by the $t\left(\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \right)$ density with area C between $-t^*$ and t^*.

The confidence interval is approximately correct when n_1 and n_2 are large even when the population is not normal.

7.2 Comparing Two Means

Two-sided two sample t-test for equality of population means when $\sigma_1 \neq \sigma_2$ is unknown and the population is normal

1. Test $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$
2. Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
3. Use the t-table to find the value t^* such that $P(T > t^*) = \frac{\alpha}{2}$ where T approximately follows a t distribution with
$$k = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$ degrees of freedom.
4. Reject H_0 if $|t| > t^*$

7.2 Comparing Two Means

One-sided (left-sided) two sample t-test for equality of population means when $\sigma_1 \neq \sigma_2$ is unknown and the population is normal

1. Test $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 < \mu_2$
2. Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
3. Use the t-table to find the value t^* such that $P(T < -t^*) = P(T > t^*) = \alpha$ where T approximately follows a t distribution with $k = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$ degrees of freedom.
4. Reject H_0 if $t < -t^*$

7.2 Comparing Two Means

<u>One-sided (right-sided) two sample t-test for equality of population means when $\sigma_1 \neq \sigma_2$ is unknown and the population is normal</u>

1. Test $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 > \mu_2$
2. Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
3. Use the t-table to find the value t^* such that $P(T > t^*) = \alpha$ where T approximately follows a t distribution with
$k = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$ degrees of freedom.
4. Reject H_0 if $t > t^*$

7.2 Comparing Two Means

Example 7E: Population A follows a Normal distribution with unknown mean μ_1 and unknown standard deviation σ_1, and Population B follows a Normal distribution with unknown mean μ_2 and unknown standard deviation σ_2. Here are the summary statistics for two independent SRSs

Population	n	$\bar{x}$	s
1	20	60	1
2	5	43	5

(a) Find an approximate 99% two-sided confidence interval for $\mu_1 - \mu_2$ without assuming equality of population standard deviations based on Satterthwaite's approximation.
(b) Assume that the population standard deviations are the same. Find a 99% two-sided confidence interval for $\mu_1 - \mu_2$.

7.2 Comparing Two Means

- *Answer*: