ISE 315: Engineering Statistics

Lecture 16: Software Tools for Statistical Inference

Instructor: Mansur M. Arief, PhD
Industrial and Systems Engineering, KFUPM

Office: 22-219 — Email: mansur.arief@kfupm.edu.sa

Based on Montgomery & Runger, Applied Statistics and Probability for Engineers, 6th Ed.

Lecture 16

Software Tools for Automating Statistical Inference & Regression
(Chapter 10)

Lecture 16 Outline

- Why use software for statistical inference? (§10.1)

Lecture 16 Outline

- Why use software for statistical inference? (§10.1)
- The tool landscape: Minitab, Excel, Python, R (Table 10.1)

Lecture 16 Outline

- Why use software for statistical inference? (§10.1)
- The tool landscape: Minitab, Excel, Python, R (Table 10.1)
- Minitab walkthrough: CIs, tests, regression, ANOVA (Procedure 10.1, Procedure 10.2, Procedure 10.3, Procedure 10.4)

Lecture 16 Outline

- Why use software for statistical inference? (§10.1)
- The tool landscape: Minitab, Excel, Python, R (Table 10.1)
- Minitab walkthrough: CIs, tests, regression, ANOVA
  (Procedure 10.1, Procedure 10.2, Procedure 10.3, Procedure 10.4)
- What the software automates vs. what **you** still must do (Table 10.2)

Lecture 16 Outline

- Why use software for statistical inference? (§10.1)
- The tool landscape: Minitab, Excel, Python, R (Table 10.1)
- Minitab walkthrough: CIs, tests, regression, ANOVA (Procedure 10.1, Procedure 10.2, Procedure 10.3, Procedure 10.4)
- What the software automates vs. what **you** still must do (Table 10.2)
- Python `scipy.stats` and the course Jupyter notebook

Why Use Software?

From Tables to Tools

What we have been doing:

- Computing test statistics by hand: Z_0, T_0, χ^2_0, F_0
- Looking up critical values from appendix tables
- Bounding p-values between table entries

Why Use Software?

From Tables to Tools

What we have been doing:

- Computing test statistics by hand: Z_0, T_0, χ_0^2, F_0
- Looking up critical values from appendix tables
- Bounding p-values between table entries

What software gives us:

- Exact p-values (e.g., $p = 0.0447$ instead of $0.02 < p < 0.05$)
- Graphical diagnostics (residual plots, distribution checks)
- Handles large datasets and complex designs effortlessly

Why Use Software?

From Tables to Tools

What we have been doing:

- Computing test statistics by hand: Z_0, T_0, χ^2_0, F_0
- Looking up critical values from appendix tables
- Bounding p-values between table entries

What software gives us:

- Exact p-values (e.g., $p = 0.0447$ instead of $0.02 < p < 0.05$)
- Graphical diagnostics (residual plots, distribution checks)
- Handles large datasets and complex designs effortlessly

Software speeds up *calculation*, but you still need to understand the results.

Textbook: Chapter 10, §10.1 ((10.1) onward)

Statistical Software Landscape

Table 10.1 in the Textbook

Tool	Strengths	Limitations	Typical Use
Minitab	Menu-driven, built for quality/industrial stats	Commercial license, less flexible	Six Sigma, QC, DoE in industry

Statistical Software Landscape

Table 10.1 in the Textbook

Tool	Strengths	Limitations	Typical Use
Minitab	Menu-driven, built for quality/industrial stats	Commercial license, less flexible	Six Sigma, QC, DoE in industry
Excel	Widely available, familiar interface	Limited stat functions, rounding issues	Quick calculations, data entry

Statistical Software Landscape

Table 10.1 in the Textbook

Tool	Strengths	Limitations	Typical Use
Minitab	Menu-driven, built for quality/industrial stats	Commercial license, less flexible	Six Sigma, QC, DoE in industry
Excel	Widely available, familiar interface	Limited stat functions, rounding issues	Quick calculations, data entry
Python (`scipy`)	Free, fully programmable, reproducible	Requires coding knowledge	Research, automation, ML

Statistical Software Landscape

Table 10.1 in the Textbook

Tool	Strengths	Limitations	Typical Use
Minitab	Menu-driven, built for quality/industrial stats	Commercial license, less flexible	Six Sigma, QC, DoE in industry
Excel	Widely available, familiar interface	Limited stat functions, rounding issues	Quick calculations, data entry
Python (`scipy`)	Free, fully programmable, reproducible	Requires coding knowledge	Research, automation, ML
R	Purpose-built for statistics, rich packages	Steeper learning curve	Academic research, biostatistics

Statistical Software Landscape

Table 10.1 in the Textbook

Tool	Strengths	Limitations	Typical Use
Minitab	Menu-driven, built for quality/industrial stats	Commercial license, less flexible	Six Sigma, QC, DoE in industry
Excel	Widely available, familiar interface	Limited stat functions, rounding issues	Quick calculations, data entry
Python (`scipy`)	Free, fully programmable, reproducible	Requires coding knowledge	Research, automation, ML
R	Purpose-built for statistics, rich packages	Steeper learning curve	Academic research, biostatistics

In ISE 315: We focus on understanding the methods. Any of these tools can execute them. Today we look at Minitab (industry standard for ISE) and Python

Minitab

Walkthrough for Chapters 3–5 and a Regression/ANOVA Preview

Minitab: Confidence Interval on the Mean

Procedure 10.1 — Reading a CI Output

Scenario (AI safety testing): An eval suite scores an LLM guardrail on $n = 25$ benchmark runs. Historical variation is known: $\sigma = 2.0$ points. Sample mean $\bar{x} = 78.40$.

Minitab: Confidence Interval on the Mean

Procedure 10.1 — Reading a CI Output

Scenario (AI safety testing): An eval suite scores an LLM guardrail on $n = 25$ benchmark runs. Historical variation is known: $\sigma = 2.0$ points. Sample mean $\bar{x} = 78.40$.

Minitab — 1-Sample Z: EvalScore

```
Descriptive Statistics
 N     Mean   StDev   SE Mean   95% CI for μ
25   78.400   2.000     0.400   (77.616, 79.184)

μ:  population mean of EvalScore

Known standard deviation = 2
```

Minitab: Confidence Interval on the Mean

Procedure 10.1 — Reading a CI Output

Scenario (AI safety testing): An eval suite scores an LLM guardrail on $n = 25$ benchmark runs. Historical variation is known: $\sigma = 2.0$ points. Sample mean $\bar{x} = 78.40$.

Minitab — 1-Sample Z: EvalScore

```
Descriptive Statistics
 N     Mean   StDev   SE Mean   95% CI for μ
25   78.400   2.000     0.400   (77.616, 79.184)

μ:  population mean of EvalScore

Known standard deviation = 2
```

This output implements (3.1). Verify by hand: $\bar{x} \pm z_{0.025}\dfrac{\sigma}{\sqrt{n}} = 78.40 \pm 1.96 \times 0.40 = (77.62,\ 79.18)$ ✓

Minitab: CI When σ Is Unknown

Same Menu, t Instead of Z

Scenario (supply chains): $n = 16$ container dwell times (hours) at a port terminal. We do not know σ. Sample: $\bar{x} = 62.75$, $s = 8.40$.

Minitab: CI When σ Is Unknown

Same Menu, t Instead of Z

Scenario (supply chains): $n = 16$ container dwell times (hours) at a port terminal. We do not know σ. Sample: $\bar{x} = 62.75$, $s = 8.40$.

Minitab — 1-Sample t: DwellTime

```
Descriptive Statistics
 N     Mean   StDev   SE Mean   95% CI for μ
16   62.750   8.400     2.100   (58.274, 67.226)

μ:  population mean of DwellTime
```

Minitab: CI When σ Is Unknown

Same Menu, t Instead of Z

Scenario (supply chains): $n = 16$ container dwell times (hours) at a port terminal. We do not know σ. Sample: $\bar{x} = 62.75$, $s = 8.40$.

Minitab — 1-Sample t: DwellTime

```
Descriptive Statistics
 N     Mean   StDev   SE Mean   95% CI for μ
16   62.750   8.400     2.100   (58.274, 67.226)

μ:  population mean of DwellTime
```

This output implements (3.5). Note the differences from the Z-interval:

- Uses $s = 8.40$ (sample StDev) instead of a known σ
- Uses $t_{0.025,\,15} = 2.131$ instead of $z_{0.025} = 1.96$
- Resulting CI is **wider** due to extra uncertainty in estimating σ

Minitab: Hypothesis Test on the Mean

Procedure 10.2 — Reading a Test Output

Scenario (startup): After a redesign, your app's mean session length should exceed the old baseline of 12 minutes. $n = 10$, $\bar{x} = 13.40$, $s = 1.90$. Test $H_0 : \mu = 12$ vs. $H_1 : \mu \neq 12$ at $\alpha = 0.05$.

Minitab: Hypothesis Test on the Mean

Procedure 10.2 — Reading a Test Output

Scenario (startup): After a redesign, your app's mean session length should exceed the old baseline of 12 minutes. $n = 10$, $\bar{x} = 13.40$, $s = 1.90$. Test $H_0 : \mu = 12$ vs. $H_1 : \mu \neq 12$ at $\alpha = 0.05$.

Minitab — 1-Sample t: SessionMin

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
10	13.400	1.900	0.601	(12.041, 14.759)

Test: H_0: μ = 12 vs. H_1: $\mu \neq 12$

T-Value	P-Value
2.33	0.045

Minitab: Hypothesis Test on the Mean

Procedure 10.2 — Reading a Test Output

Scenario (startup): After a redesign, your app's mean session length should exceed the old baseline of 12 minutes. $n = 10$, $\bar{x} = 13.40$, $s = 1.90$. Test $H_0 : \mu = 12$ vs. $H_1 : \mu \neq 12$ at $\alpha = 0.05$.

Minitab — 1-Sample t: SessionMin

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
10	13.400	1.900	0.601	(12.041, 14.759)

Test: H_0: μ = 12 vs. H_1: $\mu \neq 12$

T-Value	P-Value
2.33	0.045

The T-Value implements (4.6): $T_0 = \frac{13.40-12}{1.90/\sqrt{10}} = 2.33$. From the t-table, $0.02 < p < 0.05$; Minitab gives the **exact** p-value $p = 0.045$. **Reject H_0.**

Minitab: Hypothesis Test on the Mean

Procedure 10.2 — Reading a Test Output

Scenario (startup): After a redesign, your app's mean session length should exceed the old baseline of 12 minutes. $n = 10$, $\bar{x} = 13.40$, $s = 1.90$. Test $H_0 : \mu = 12$ vs. $H_1 : \mu \neq 12$ at $\alpha = 0.05$.

Minitab — 1-Sample t: SessionMin

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
10	13.400	1.900	0.601	(12.041, 14.759)

Test: H_0: μ = 12 vs. H_1: $\mu \neq 12$

T-Value	P-Value
2.33	0.045

The T-Value implements (4.6): $T_0 = \frac{13.40-12}{1.90/\sqrt{10}} = 2.33$. From the t-table, $0.02 < p < 0.05$; Minitab gives the **exact** p-value $p = 0.045$. **Reject H_0.**

Minitab: Test on a Variance

Scenario (AI safety testing): A perception model's inference latency must stay consistent: target $\sigma^2 = 0.25$ ms^2. From $n = 20$ runs, $s = 0.44$ ms. Test $H_0 : \sigma^2 = 0.25$ at $\alpha = 0.05$.

Minitab: Test on a Variance

Scenario (AI safety testing): A perception model's inference latency must stay consistent: target $\sigma^2 = 0.25$ ms^2. From $n = 20$ runs, $s = 0.44$ ms. Test $H_0 : \sigma^2 = 0.25$ at $\alpha = 0.05$.

Minitab — 1 Variance: LatencyMs

Test: H_0: $\sigma^2 = 0.25$ vs. H_1: $\sigma^2 \neq 0.25$

Method: Chi-Square

Statistic	DF	P-Value
14.71	19	0.519

95% CI for σ^2: (0.112, 0.413)

Minitab: Test on a Variance

Scenario (AI safety testing): A perception model's inference latency must stay consistent: target $\sigma^2 = 0.25$ ms^2. From $n = 20$ runs, $s = 0.44$ ms. Test $H_0 : \sigma^2 = 0.25$ at $\alpha = 0.05$.

Minitab — 1 Variance: LatencyMs

Test: H_0: $\sigma^2 = 0.25$ vs. H_1: $\sigma^2 \neq 0.25$

Method: Chi-Square

Statistic	DF	P-Value
14.71	19	0.519

95% CI for σ^2: (0.112, 0.413)

The Statistic implements (4.7): $\chi_0^2 = \frac{(20-1)(0.44)^2}{0.25} = \frac{19 \times 0.1936}{0.25} = 14.71$.
From the χ^2-table, $p > 0.10$. Minitab gives $p = 0.519$: **fail to reject** H_0.

Minitab: Test on a Variance

Scenario (AI safety testing): A perception model's inference latency must stay consistent: target $\sigma^2 = 0.25$ ms^2. From $n = 20$ runs, $s = 0.44$ ms. Test $H_0 : \sigma^2 = 0.25$ at $\alpha = 0.05$.

Minitab — 1 Variance: LatencyMs

Test: H_0: $\sigma^2 = 0.25$ vs. H_1: $\sigma^2 \neq 0.25$

Method: Chi-Square

Statistic	DF	P-Value
14.71	19	0.519

95% CI for σ^2: (0.112, 0.413)

The Statistic implements (4.7): $\chi_0^2 = \frac{(20-1)(0.44)^2}{0.25} = \frac{19 \times 0.1936}{0.25} = 14.71$.
From the χ^2-table, $p > 0.10$. Minitab gives $p = 0.519$: **fail to reject** H_0.

Minitab also reports the CI for σ^2 from (3.6), plus a Bonett method for non-

Minitab: Two-Sample t-Test

Scenario (transportation): Delivery lateness (minutes) on two routes. Test H_0 : $\mu_1 = \mu_2$ at $\alpha = 0.05$, assuming equal variances.

Minitab: Two-Sample t-Test

Scenario (transportation): Delivery lateness (minutes) on two routes. Test H_0 : $\mu_1 = \mu_2$ at $\alpha = 0.05$, assuming equal variances.

Minitab — 2-Sample t: RouteA vs RouteB (pooled)

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
RouteA	10	41.20	5.10	1.61
RouteB	12	37.80	4.60	1.33

Estimate for difference: 3.400 95% CI: $(-0.915, 7.715)$

Test: H_0: $\mu_1 - \mu_2 = 0$ vs. H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
1.64	20	0.116

Both use Pooled StDev = 4.8314

Minitab: Two-Sample t-Test

Scenario (transportation): Delivery lateness (minutes) on two routes. Test H_0 : $\mu_1 = \mu_2$ at $\alpha = 0.05$, assuming equal variances.

Minitab — 2-Sample t: RouteA vs RouteB (pooled)

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
RouteA	10	41.20	5.10	1.61
RouteB	12	37.80	4.60	1.33

Estimate for difference: 3.400 95% CI: $(-0.915, 7.715)$

Test: H_0: $\mu_1 - \mu_2 = 0$ vs. H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
1.64	20	0.116

Both use Pooled StDev = 4.8314

This implements (5.4) with s_p from (5.3). $p = 0.116 > 0.05$: **fail to reject** H_0. The CI includes 0 — consistent.

Minitab: Complete Two-Sample Menu Map

Table 10.5 in the Textbook

Test	**Minitab Menu Path**	**Key Options**
Two-sample t (pooled)	`Stat` → `Basic Stat` → `2-Sample t`	Check "Assume equal variances"
Two-sample t (Welch)	`Stat` → `Basic Stat` → `2-Sample t`	Uncheck equal variances
Paired t	`Stat` → `Basic Stat` → `Paired t`	Two matched columns
Two variances (F-test)	`Stat` → `Basic Stat` → `2 Variances`	Levene's test also shown
Two proportions	`Stat` → `Basic Stat` → `2 Proportions`	Event counts and trials

Minitab: Complete Two-Sample Menu Map

Table 10.5 in the Textbook

Test	Minitab Menu Path	Key Options
Two-sample t (pooled)	`Stat → Basic Stat → 2-Sample t`	Check "Assume equal variances"
Two-sample t (Welch)	`Stat → Basic Stat → 2-Sample t`	Uncheck equal variances
Paired t	`Stat → Basic Stat → Paired t`	Two matched columns
Two variances (F-test)	`Stat → Basic Stat → 2 Variances`	Levene's test also shown
Two proportions	`Stat → Basic Stat → 2 Proportions`	Event counts and trials

Minitab automatically provides:

- The test statistic, the exact p-value, and the confidence interval
- Degrees of freedom (including the Welch df from (5.7))

Minitab: Complete Two-Sample Menu Map

Table 10.5 in the Textbook

Test	Minitab Menu Path	Key Options
Two-sample t (pooled)	`Stat → Basic Stat → 2-Sample t`	Check "Assume equal variances"
Two-sample t (Welch)	`Stat → Basic Stat → 2-Sample t`	Uncheck equal variances
Paired t	`Stat → Basic Stat → Paired t`	Two matched columns
Two variances (F-test)	`Stat → Basic Stat → 2 Variances`	Levene's test also shown
Two proportions	`Stat → Basic Stat → 2 Proportions`	Event counts and trials

Minitab automatically provides:

- The test statistic, the exact p-value, and the confidence interval
- Degrees of freedom (including the Welch df from (5.7))

Tip: Run `2 Variances` first. If the F-test (5.13) rejects equal variances, uncheck "Assume equal variances" in the 2-Sample t dialog.

Minitab: Regression Analysis

Procedure 10.3 — `Stat > Regression > Fit Regression Model`

Scenario (startup): Monthly ad spend x (SAR thousands) vs. new signups y for $n = 12$ months.

Minitab: Regression Analysis

Procedure 10.3 — Stat > Regression > Fit Regression Model

Scenario (startup): Monthly ad spend x (SAR thousands) vs. new signups y for $n = 12$ months.

Minitab — Regression: Signups vs AdSpend

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	30.50	3.81	8.01	0.000
AdSpend	8.930	0.377	23.70	0.000

Model Summary

S	R-sq	R-sq(adj)
4.505	98.25%	98.08%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	11403.7	11403.7	561.85	0.000
Error	10	203.0	20.3		
Total	11	11606.7			

Minitab: Regression Analysis

Procedure 10.3 — `Stat > Regression > Fit Regression Model`

Scenario (startup): Monthly ad spend x (SAR thousands) vs. new signups y for $n = 12$ months.

Minitab — Regression: Signups vs AdSpend

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	30.50	3.81	8.01	0.000
AdSpend	8.930	0.377	23.70	0.000

Model Summary

S	R-sq	R-sq(adj)
4.505	98.25%	98.08%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	11403.7	11403.7	561.85	0.000
Error	10	203.0	20.3		
Total	11	11606.7			

Map to the book: Coef column — $\hat{\beta}_0$, $\hat{\beta}_1$ from (6.10), (6.9); T-Value — (6.18);

Minitab: One-Way ANOVA

Procedure 10.4 — A Chapter 8 Preview

Scenario (AI safety testing): Three guardrail filter configurations (A, B, C), 6 red-team batches each. Response: false positives per 1000 prompts. Are the means equal?

Minitab: One-Way ANOVA

Procedure 10.4 — A Chapter 8 Preview

Scenario (AI safety testing): Three guardrail filter configurations (A, B, C), 6 red-team batches each. Response: false positives per 1000 prompts. Are the means equal?

Minitab — One-Way ANOVA: FalsePos vs Filter

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Filter	2	108.00	54.00	27.00	0.000
Error	15	30.00	2.00		
Total	17	138.00			

Means

Filter	N	Mean
A	6	15.000
B	6	12.000
C	6	9.000

Minitab: One-Way ANOVA

Procedure 10.4 — A Chapter 8 Preview

Scenario (AI safety testing): Three guardrail filter configurations (A, B, C), 6 red-team batches each. Response: false positives per 1000 prompts. Are the means equal?

Minitab — One-Way ANOVA: FalsePos vs Filter

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Filter	2	108.00	54.00	27.00	0.000
Error	15	30.00	2.00		
Total	17	138.00			

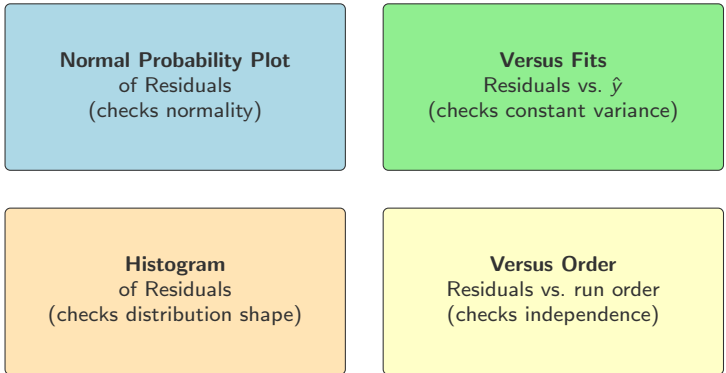
Means

Filter	N	Mean
A	6	15.000
B	6	12.000
C	6	9.000

Minitab: Residual Diagnostics

The "Four-in-One" Plot (Figure 10.4)

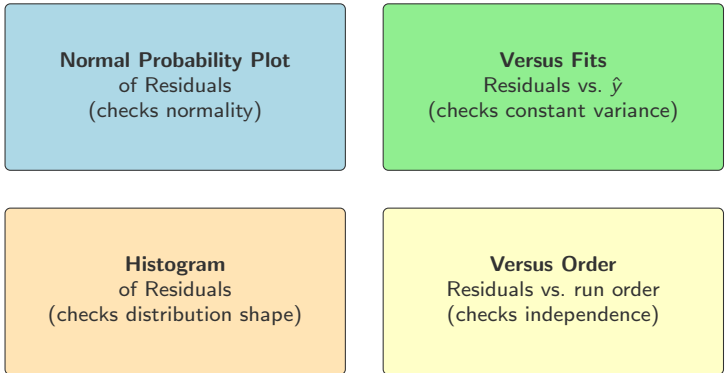
Minitab's `Graphs` button in the regression menu gives the **four-in-one residual plot**:



Minitab: Residual Diagnostics

The "Four-in-One" Plot (Figure 10.4)

Minitab's `Graphs` button in the regression menu gives the **four-in-one residual plot**:



Two Facts for Reading Software Output

§10.9 in the Textbook

Fact 1: One-sided p-values from two-sided output. Most software reports a two-sided p-value by default. If your H_1 is one-sided and the estimate lies on the H_1 side:

$$p_{\text{one-sided}} = \frac{1}{2}\, p_{\text{two-sided}} \tag{10.1}$$

Two Facts for Reading Software Output

§10.9 in the Textbook

Fact 1: One-sided p-values from two-sided output. Most software reports a two-sided p-value by default. If your H_1 is one-sided and the estimate lies on the H_1 side:

$$p_{\text{one-sided}} = \frac{1}{2}\, p_{\text{two-sided}} \tag{10.1}$$

Fact 2: The t and F columns agree. In simple linear regression output, the slope t-statistic (6.18) and the ANOVA F-statistic (6.20) test the same hypothesis, and:

$$T_0^2 = F_0 \tag{10.2}$$

Two Facts for Reading Software Output

§10.9 in the Textbook

Fact 1: One-sided p-values from two-sided output. Most software reports a two-sided p-value by default. If your H_1 is one-sided and the estimate lies on the H_1 side:

$$p_{\text{one-sided}} = \tfrac{1}{2}\, p_{\text{two-sided}} \tag{10.1}$$

Fact 2: The t and F columns agree. In simple linear regression output, the slope t-statistic (6.18) and the ANOVA F-statistic (6.20) test the same hypothesis, and:

$$T_0^2 = F_0 \tag{10.2}$$

Check it on our ad-spend output: $T_0 = 23.70$ for the slope, and $23.70^2 = 561.7 \approx 561.85 = F_0$ (rounding). Use this as a built-in sanity check when reading

What Software Automates

The Division of Labor (Table 10.2)

Software handles this	**You still must do this**
Compute test statistics (Z_0, T_0, χ^2_0, F_0)	Choose the correct test for the problem
Look up exact p-values	Interpret the p-value in context
Calculate CI bounds	Select the right confidence level
Degrees of freedom (incl. Welch)	Verify assumptions (normality, independence)
Residual plots	Judge whether assumptions are met
Regression coefficients	Decide which variables to include
ANOVA decomposition	Explain what the results mean to a stakeholder

What Software Automates

The Division of Labor (Table 10.2)

Software handles this	**You still must do this**
Compute test statistics (Z_0, T_0, χ^2_0, F_0)	Choose the correct test for the problem
Look up exact p-values	Interpret the p-value in context
Calculate CI bounds	Select the right confidence level
Degrees of freedom (incl. Welch)	Verify assumptions (normality, independence)
Residual plots	Judge whether assumptions are met
Regression coefficients	Decide which variables to include
ANOVA decomposition	Explain what the results mean to a stakeholder

Warning: Software will happily run a t-test on non-normal data, or fit a regression with meaningless

What We Still Need From You

1. **Problem Formulation**
   - Writing H_0 and H_1 correctly (one-sided vs. two-sided), as in Procedure 4.2
   - Choosing α based on the engineering context and consequences

What We Still Need From You

1. **Problem Formulation**
   - Writing H_0 and H_1 correctly (one-sided vs. two-sided), as in Procedure 4.2
   - Choosing α based on the engineering context and consequences
2. **Test Selection**
   - σ known $\rightarrow Z$, unknown $\rightarrow t$, variance $\rightarrow \chi^2$, two variances $\rightarrow F$

What We Still Need From You

1. **Problem Formulation**
   - Writing H_0 and H_1 correctly (one-sided vs. two-sided), as in Procedure 4.2
   - Choosing α based on the engineering context and consequences
2. **Test Selection**
   - σ known $\rightarrow Z$, unknown $\rightarrow t$, variance $\rightarrow \chi^2$, two variances $\rightarrow F$
3. **Assumption Checking**
   - Is the population approximately normal? Is n large enough for the CLT?
   - Are the two samples truly independent, or are they paired?
   - For proportions: are $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$?

What We Still Need From You

1. **Problem Formulation**
   - Writing H_0 and H_1 correctly (one-sided vs. two-sided), as in Procedure 4.2
   - Choosing α based on the engineering context and consequences
2. **Test Selection**
   - σ known $\rightarrow Z$, unknown $\rightarrow t$, variance $\rightarrow \chi^2$, two variances $\rightarrow F$
3. **Assumption Checking**
   - Is the population approximately normal? Is n large enough for the CLT?
   - Are the two samples truly independent, or are they paired?
   - For proportions: are $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$?
4. **Interpretation and Communication**

Suggested Workflow

You Decide, Software Computes, You Conclude

Problem (AI safety testing): A red team measures jailbreak success counts on $n = 15$ prompt suites, before and after a safety patch. Did the patch reduce the mean success count?

Suggested Workflow

You Decide, Software Computes, You Conclude

Problem (AI safety testing): A red team measures jailbreak success counts on $n = 15$ prompt suites, before and after a safety patch. Did the patch reduce the mean success count?

You (the engineer) decide:

- This is **paired data** (same suites, before vs. after)
- $H_0 : \mu_D = 0$ vs. $H_1 : \mu_D > 0$, where $D_i = \text{Before}_i - \text{After}_i$
- Use $\alpha = 0.05$ — the paired statistic is (5.8)

Suggested Workflow

You Decide, Software Computes, You Conclude

Problem (AI safety testing): A red team measures jailbreak success counts on $n = 15$ prompt suites, before and after a safety patch. Did the patch reduce the mean success count?

You (the engineer) decide:

- This is **paired data** (same suites, before vs. after)
- $H_0 : \mu_D = 0$ vs. $H_1 : \mu_D > 0$, where $D_i = \text{Before}_i - \text{After}_i$
- Use $\alpha = 0.05$ — the paired statistic is (5.8)

Software computes: $\bar{d} = 2.60$, $s_D = 3.90$, $T_0 = 2.58$, one-sided $p = 0.011$

Suggested Workflow

You Decide, Software Computes, You Conclude

Problem (AI safety testing): A red team measures jailbreak success counts on $n = 15$ prompt suites, before and after a safety patch. Did the patch reduce the mean success count?

You (the engineer) decide:

- This is **paired data** (same suites, before vs. after)
- $H_0 : \mu_D = 0$ vs. $H_1 : \mu_D > 0$, where $D_i = \text{Before}_i - \text{After}_i$
- Use $\alpha = 0.05$ — the paired statistic is (5.8)

Software computes: $\bar{d} = 2.60$, $s_D = 3.90$, $T_0 = 2.58$, one-sided $p = 0.011$

You (the engineer) conclude: "At the 5% significance level, there is sufficient evidence that the patch reduces the mean jailbreak success count."

Python Alternative

Free, Programmable, Reproducible

Python for Statistical Inference

Table 10.4 in the Textbook

Key functions in `scipy.stats`:

Our Test	Python Function	Returns
1-sample Z (CI/HT)	`norm.cdf()`, `norm.ppf()`	CDF values, critical values
1-sample t-test	`ttest_1samp(data, `μ_0`)`	T_0, p-value
2-sample t-test	`ttest_ind(data1, data2)`	T_0, p-value
Paired t-test	`ttest_rel(before, after)`	T_0, p-value
χ^2 on variance	`chi2.cdf()`, `chi2.ppf()`	CDF values, critical values
F-test on variances	`f.cdf()`, `f.ppf()`	CDF values, critical values
Regression	`linregress(x, y)`	$\hat{\beta}_0, \hat{\beta}_1$, r, p-value
One-way ANOVA	`f_oneway(g1, g2, g3)`	F_0, p-value

Python for Statistical Inference

Table 10.4 in the Textbook

Key functions in `scipy.stats`:

Our Test	Python Function	Returns
1-sample Z (CI/HT)	`norm.cdf()`, `norm.ppf()`	CDF values, critical values
1-sample t-test	`ttest_1samp(data, `μ_0`)`	T_0, p-value
2-sample t-test	`ttest_ind(data1, data2)`	T_0, p-value
Paired t-test	`ttest_rel(before, after)`	T_0, p-value
χ^2 on variance	`chi2.cdf()`, `chi2.ppf()`	CDF values, critical values
F-test on variances	`f.cdf()`, `f.ppf()`	CDF values, critical values
Regression	`linregress(x, y)`	$\hat{\beta}_0, \hat{\beta}_1$, r, p-value
One-way ANOVA	`f_oneway(g1, g2, g3)`	F_0, p-value

Advantage: You can write a script once and rerun it on any new dataset. Fully

Python Example: One-Sample t-Test

The Session-Length Test, Reproduced in Code

```
from scipy import stats
import numpy as np

# minutes per session after the redesign (n = 10)
data = [13.1, 15.2, 11.8, 14.6, 12.9, 16.0, 12.2, 13.7, 14.4, 10.1]

# H0: mu = 12 vs H1: mu != 12, alpha = 0.05
mu_0 = 12
t_stat, p_value = stats.ttest_1samp(data, mu_0)

print(f"T-statistic: {t_stat:.4f}")
print(f"P-value:     {p_value:.4f}")

# 95% confidence interval (book eq. 3.5)
n = len(data)
xbar, s = np.mean(data), np.std(data, ddof=1)
t_crit = stats.t.ppf(0.975, df=n-1)
ci = (xbar - t_crit*s/np.sqrt(n),
      xbar + t_crit*s/np.sqrt(n))
```

Common Pitfalls When Using Software

- **Blindly trusting the output.**
  Software does not check your assumptions. A t-test on highly skewed data with $n = 5$ will still produce a p-value, but it may not be valid.

Common Pitfalls When Using Software

- **Blindly trusting the output.**
  Software does not check your assumptions. A t-test on highly skewed data with $n = 5$ will still produce a p-value, but it may not be valid.
- **Choosing the wrong test.**
  Running a two-sample t-test on paired data inflates the standard error and loses power. The software will not warn you.

Common Pitfalls When Using Software

- **Blindly trusting the output.**
  Software does not check your assumptions. A t-test on highly skewed data with $n = 5$ will still produce a p-value, but it may not be valid.
- **Choosing the wrong test.**
  Running a two-sample t-test on paired data inflates the standard error and loses power. The software will not warn you.
- **Ignoring practical significance.**
  With $n = 10{,}000$, even a trivial difference can be "statistically significant" ($p < 0.001$). Always report effect size alongside the p-value.

Common Pitfalls When Using Software

- **Blindly trusting the output.**
  Software does not check your assumptions. A t-test on highly skewed data with $n = 5$ will still produce a p-value, but it may not be valid.
- **Choosing the wrong test.**
  Running a two-sample t-test on paired data inflates the standard error and loses power. The software will not warn you.
- **Ignoring practical significance.**
  With $n = 10{,}000$, even a trivial difference can be "statistically significant" ($p < 0.001$). Always report effect size alongside the p-value.
- **Forgetting one-sided vs. two-sided.**
  Most software defaults to a two-sided test. If your H_1 is one-sided, apply (10.1) or select the correct option in the dialog.

Demo: ISE 315 Statistical Inference Notebook

We have prepared a **Jupyter Notebook** that covers all tests from Chapters 3–5:

Demo: ISE 315 Statistical Inference Notebook

We have prepared a **Jupyter Notebook** that covers all tests from Chapters 3–5:

Features:

- Upload your raw data (CSV) or enter summary statistics
- Select the test type from a menu
- The notebook recommends settings based on your inputs
- Outputs: test statistic, exact p-value, CI, and a plain-English conclusion

Demo: ISE 315 Statistical Inference Notebook

We have prepared a **Jupyter Notebook** that covers all tests from Chapters 3–5:

Features:

- Upload your raw data (CSV) or enter summary statistics
- Select the test type from a menu
- The notebook recommends settings based on your inputs
- Outputs: test statistic, exact p-value, CI, and a plain-English conclusion

Supported tests:

- Chapter 3: CI for μ (σ known/unknown), CI for σ^2, CI for p
- Chapter 4: Z-test, t-test, χ^2-test on variance, proportion test
- Chapter 5: two-sample Z, pooled t, Welch t, paired t, F-test, two-proportion Z

Demo: ISE 315 Statistical Inference Notebook

We have prepared a **Jupyter Notebook** that covers all tests from Chapters 3–5:

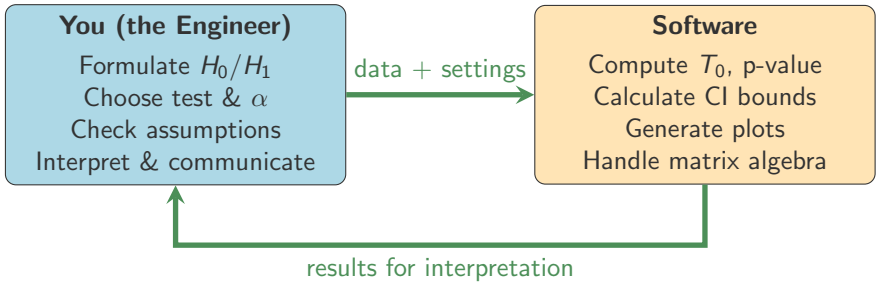
Features:

- Upload your raw data (CSV) or enter summary statistics
- Select the test type from a menu
- The notebook recommends settings based on your inputs
- Outputs: test statistic, exact p-value, CI, and a plain-English conclusion

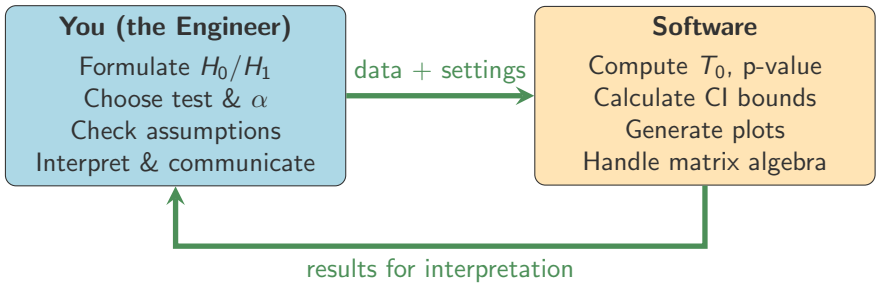
Supported tests:

- Chapter 3: CI for μ (σ known/unknown), CI for σ^2, CI for p
- Chapter 4: Z-test, t-test, χ^2-test on variance, proportion test
- Chapter 5: two-sample Z, pooled t, Welch t, paired t, F-test, two-proportion Z

Summary: The Right Balance



Summary: The Right Balance



Bottom line: Master the concepts first (that is what exams test, by hand with tables). Then use software — with the master menu map in Table 10.7 — to be efficient and accurate in practice.

Remarks

- The Jupyter notebook is posted on Blackboard (Supplementary Materials)

Remarks

- The Jupyter notebook is posted on Blackboard (Supplementary Materials)
- Software is useful for homework verification and your own projects

Remarks

- The Jupyter notebook is posted on Blackboard (Supplementary Materials)
- Software is useful for homework verification and your own projects
- Next lecture: back to Chapter 6 for a full worked regression fit

Remarks

- The Jupyter notebook is posted on Blackboard (Supplementary Materials)
- Software is useful for homework verification and your own projects
- Next lecture: back to Chapter 6 for a full worked regression fit
- Skim §10.8 (the master menu map) before you try the notebook