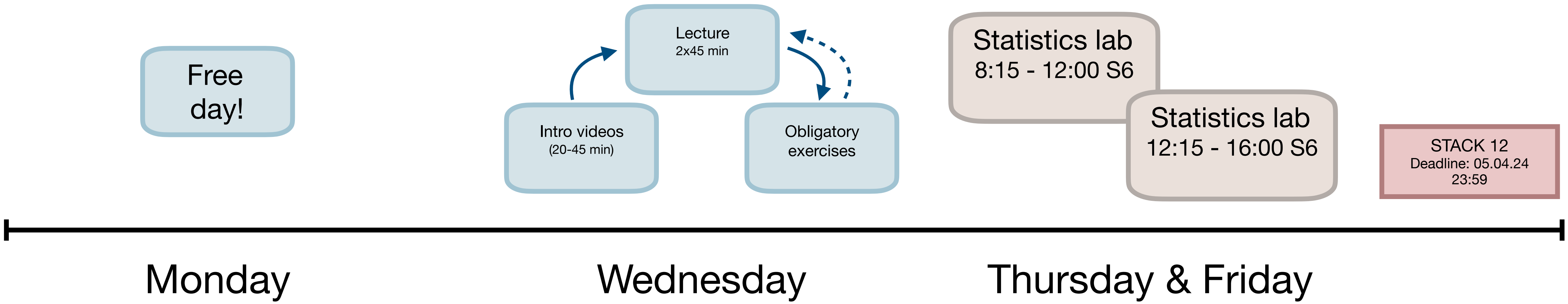
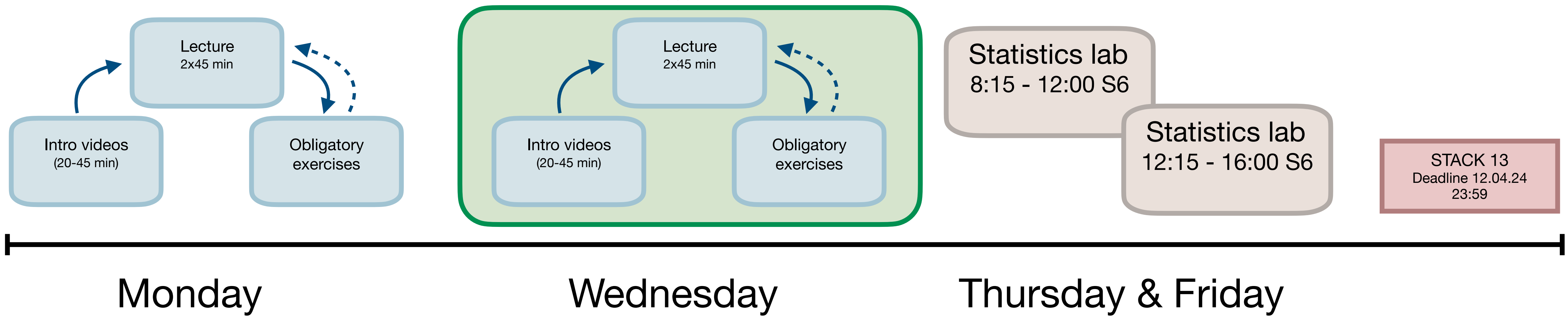
TMA4245 Statistikk

Week 15 - Wednesday

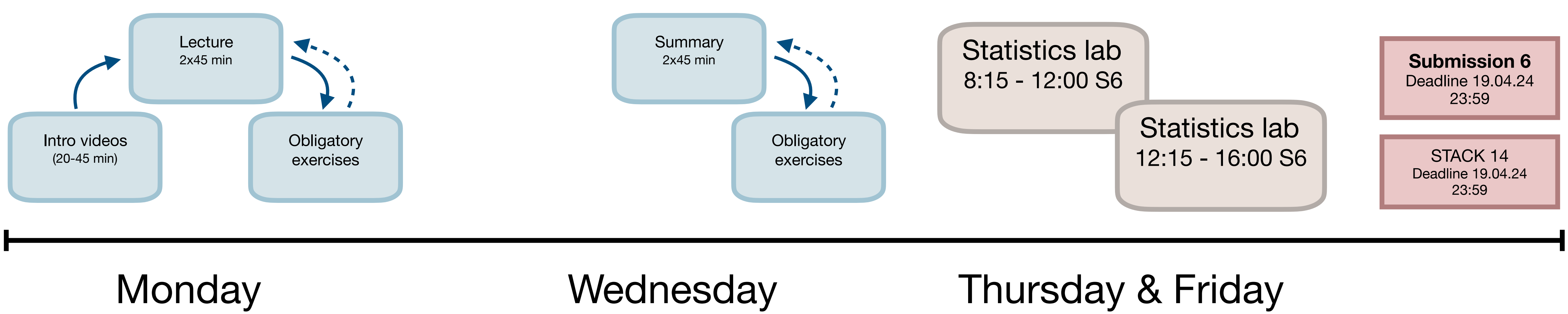
Week 14 (Mon. April 1 - Friday April 5)



Week 15 (Mon. April 8 - Friday April 12)



Week 16 (Mon. April 15 - Friday April 19)



Brief review: Effect of ε assumptions

If x has has linear relationship with y
and x_i are fixed numbers:

1. $E[\varepsilon_1] = \ldots = E[\varepsilon_n] = 0 \;\Rightarrow$

$$E[\hat{\alpha}] = \alpha$$
$$E[\hat{\beta}] = \beta$$

Brief review: Effect of ε assumptions

If x has has linear relationship with y and x_i are fixed numbers:

1. $E[\varepsilon_1] = \ldots = E[\varepsilon_n] = 0 \;\Rightarrow$

$$E[\hat{\alpha}] = \alpha$$
$$E[\hat{\beta}] = \beta$$

2. $\varepsilon_1, \ldots, \varepsilon_n$ are iid with mean 0 and variance $\sigma^2 \;\Rightarrow$

$$Var(\hat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$Var(\hat{\beta}) = \frac{\sigma^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$

Brief review: Effect of ε assumptions

If x has has linear relationship with y and x_i are fixed numbers:

1. $E[\varepsilon_1] = \ldots = E[\varepsilon_n] = 0 \;\Rightarrow$

$$E[\hat{\alpha}] = \alpha$$
$$E[\hat{\beta}] = \beta$$

2. $\varepsilon_1, \ldots, \varepsilon_n$ are iid with mean 0 and variance $\sigma^2 \;\Rightarrow$

$$Var(\hat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$Var(\hat{\beta}) = \frac{\sigma^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$

3. $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2) \Rightarrow$

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$
$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

Brief review: Effect of ε assumptions

If x has has linear relationship with y and x_i are fixed numbers:

1. $E[\varepsilon_1] = \ldots = E[\varepsilon_n] = 0 \ \Rightarrow$

$$E[\hat{\alpha}] = \alpha$$
$$E[\hat{\beta}] = \beta$$

2. $\varepsilon_1, \ldots, \varepsilon_n$ are iid with mean 0 and variance $\sigma^2 \ \Rightarrow$

$$Var(\hat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$Var(\hat{\beta}) = \frac{\sigma^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$

3. $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2) \Rightarrow$

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$
$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

Notes:

- $\hat{\alpha}$ and $\hat{\beta}$ are still approximately normal under **2.** provided n is large enough.
- In this class we assume **3.** is necessary for simple linear regression unless otherwise stated

Brief review

Assuming $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$:

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Brief review

Assuming $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$:

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

But σ^2 is unknown! Must instead be estimated. MLE (SME) is:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta} x_i))^2,$$

Brief review

Assuming $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$:

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

But σ^2 is unknown! Must instead be estimated. MLE (SME) is:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta} x_i))^2,$$

What can we say about $\hat{\alpha}$, $\hat{\beta}$ if σ^2 is unknown?

Recall:

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2} \quad \text{and} \quad \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

for $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2$.

Also:

$$\frac{Z}{\sqrt{\frac{V}{\nu}}} \sim t_\nu$$

for $Z \sim N(0,1)$ and $V \sim \chi^2_\nu$ independent.

Also, S^2 is independent of $\hat{\alpha}$, $\hat{\beta}$!

Brief review

Hence, we get the following pivotal quantities under H_0:

$$\frac{\hat{\alpha} - \alpha_0}{\sqrt{S^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}$$

$$\frac{\hat{\beta} - \beta_0}{\sqrt{S^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}.$$

Brief review

Similar logic shows, for prediction at x_0 given by $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta} x_0$:

$$E[\hat{\mu}_0] = \alpha + \beta x_0 = \mu_0$$
$$Var(\hat{\mu}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right),$$

and:

$$\frac{\hat{\mu}_0 - \mu_0}{\sqrt{S^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}} \sim t_{n-2}.$$

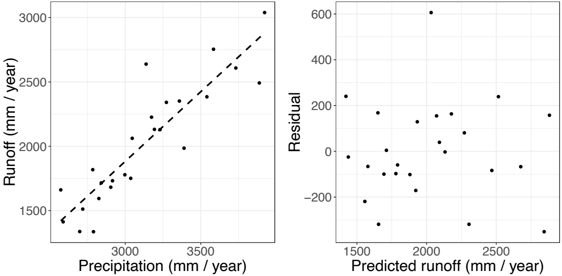
Constructing confidence intervals

We can then use the pivotal quantities to construct $(1-\alpha) \times 100\%$ confidence intervals:

$$
\begin{aligned}
\frac{\hat{\theta}-\theta_0}{\sqrt{S^2 \cdot C}} &\sim t_{n-2} \\
-t_{\alpha/2,n-2} <& \frac{\hat{\theta}-\theta_0}{\sqrt{S^2 \cdot C}} < t_{\alpha/2,n-2} \\
-t_{\alpha/2,n-2}\sqrt{S^2 \cdot C} <& \hat{\theta}-\theta_0 < t_{\alpha/2,n-2}\sqrt{S^2 \cdot C} \\
-\hat{\theta}-t_{\alpha/2,n-2}\sqrt{S^2 \cdot C} <& -\theta_0 < -\hat{\theta}+t_{\alpha/2,n-2}\sqrt{S^2 \cdot C} \\
\hat{\theta}-t_{\alpha/2,n-2}\sqrt{S^2 \cdot C} <& \theta_0 < \hat{\theta}+t_{\alpha/2,n-2}\sqrt{S^2 \cdot C},
\end{aligned}
$$

for constant C. This gives us the tools we need for hypothesis testing and confidence intervals!

Runoff (based on problem 3, Spring 2019)

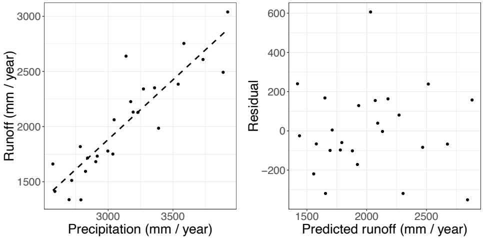


Assume the following linear relationship between annual runoff Y and annual precipitation x within a drainage basin,

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{1}$$

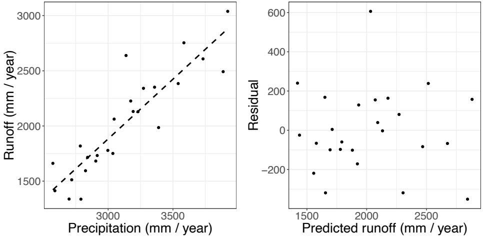
where β_0 and β_1 are unknown constants and ε is normally distributed with expected value (mean) 0 and unknown variance σ^2.

Runoff (based on problem 3, Spring 2019)



Assume that we now are interested in predicting future runoff for a new year μ_0 given annual precipitation $x = x_0$, from the model defined in (1), where (x_0, Y_0) is independent of $(x_1, Y_1), (x_2, Y_2), \ldots, (x_{25}, Y_{25})$. Assume that $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is a reasonable point estimator for the expected runoff $\mu_{Y|x_0} = \beta_0 + \beta_1 x_0$ when annual precipitation is x_0. Assume $\hat{\beta}_0 = -1364$, $\hat{\beta}_1 = 1.08$, and $S^2 = 156^2$.

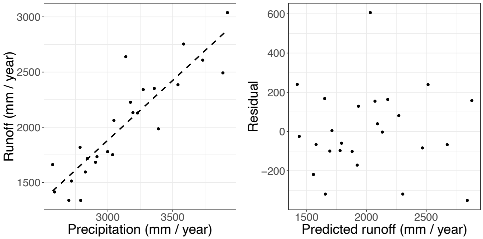
Runoff (based on problem 3, Spring 2019)



Assume $\hat{\beta}_0 = -1364$, $\hat{\beta}_1 = 1.08$, and $S^2 = 156^2$. Questions:

1. Show that $\hat{\mu}_0$ is unbiased
2. Show that $Var(\hat{\mu}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$

Runoff (based on problem 3, Spring 2019)



Assume $\hat{\beta}_0 = -1364$, $\hat{\beta}_1 = 1.08$, and $S^2 = 156^2$. Questions:

1. Show that $\hat{\mu}_0$ is unbiased
2. Show that $Var(\hat{\mu}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$
3. Assume the average yearly precipitation in a year, $\bar{x}$, is 3200 mm/yr. Predict runoff for an average year.
4. Give a 95% confidence interval for the average runoff in a year with average precipitation.

Confidence intervals for the trend line

Recall:

$$E[\hat{\mu}_0] = \alpha + \beta x_0 = \mu_0$$
$$Var(\hat{\mu}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right),$$

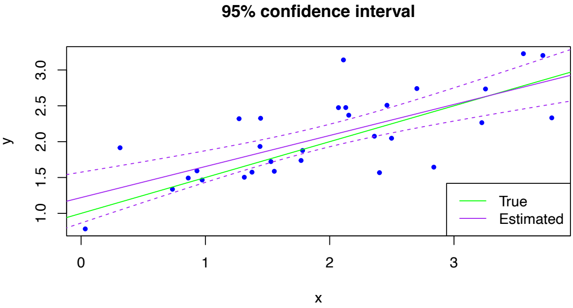
and:

$$\frac{\hat{\mu}_0 - \mu_0}{\sqrt{S^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}} \sim t_{n-2},$$

so:

$$\hat{\mu}_0 - t_{\alpha/2,n-2}\sqrt{S^2 \cdot C} < \mu_0 < \hat{\mu}_0 + t_{\alpha/2,n-2}\sqrt{S^2 \cdot C}.$$

Confidence intervals for the trend line

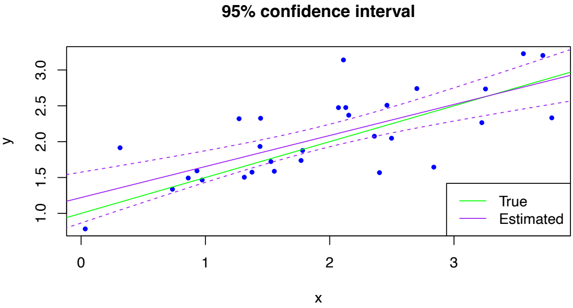


The confidence interval (CI) margin of error (MOE) is then:

$$t_{\alpha/2,n-2}\sqrt{S^2 \cdot C} = t_{\alpha/2,n-2}\sqrt{S^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

.

Confidence intervals for the trend line



The confidence interval (CI) margin of error (MOE) is then:

$$t_{\alpha/2,n-2}\sqrt{S^2 \cdot C} = t_{\alpha/2,n-2}\sqrt{S^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

$$= t_{\alpha/2,n-2}S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Constructing approximate confidence intervals

What if the ϵ_i are non-Gaussian?

If $\varepsilon_1, \ldots, \varepsilon_n$ iid, $E[\varepsilon_i] = 0$, and $Var(\varepsilon_i) = \sigma^2$, then for large enough n:

Constructing approximate confidence intervals

What if the ϵ_i are non-Gaussian?

If $\varepsilon_1, \ldots, \varepsilon_n$ iid, $E[\varepsilon_i] = 0$, and $Var(\varepsilon_i) = \sigma^2$, then for large enough n:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are still approximately Gaussian,

Constructing approximate confidence intervals

What if the ϵ_i are non-Gaussian?

If $\varepsilon_1, \ldots, \varepsilon_n$ iid, $E[\varepsilon_i] = 0$, and $Var(\varepsilon_i) = \sigma^2$, then for large enough n:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are still approximately Gaussian,
- $\hat{\mu}_0$ is still approximately Gaussian,

Constructing approximate confidence intervals

What if the ϵ_i are non-Gaussian?

If $\varepsilon_1, \ldots, \varepsilon_n$ iid, $E[\varepsilon_i] = 0$, and $Var(\varepsilon_i) = \sigma^2$, then for large enough n:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are still approximately Gaussian,
- $\hat{\mu}_0$ is still approximately Gaussian,
- S^2 still converges to $Var(\varepsilon)$ in some sense,

Constructing approximate confidence intervals

What if the ϵ_i are non-Gaussian?

If $\varepsilon_1, \ldots, \varepsilon_n$ iid, $E[\varepsilon_i] = 0$, and $Var(\varepsilon_i) = \sigma^2$, then for large enough n:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are still approximately Gaussian,
- $\hat{\mu}_0$ is still approximately Gaussian,
- S^2 still converges to $Var(\varepsilon)$ in some sense,
- **approximate** $(1 - \alpha) \times 100\%$ confidence intervals can still be created via:

$$\hat{\theta} - Z_{\alpha/2}\sqrt{S^2 \cdot C} < \theta_0 < \hat{\theta} + Z_{\alpha/2}\sqrt{S^2 \cdot C}$$

  only using $Z_{\alpha/2}$ instead of $t_{\alpha/2,n-2}$, and

Constructing approximate confidence intervals

What if the ϵ_i are non-Gaussian?

If $\varepsilon_1, \ldots, \varepsilon_n$ iid, $E[\varepsilon_i] = 0$, and $Var(\varepsilon_i) = \sigma^2$, then for large enough n:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are still approximately Gaussian,
- $\hat{\mu}_0$ is still approximately Gaussian,
- S^2 still converges to $Var(\varepsilon)$ in some sense,
- **approximate** $(1-\alpha) \times 100\%$ confidence intervals can still be created via:

$$\hat{\theta} - Z_{\alpha/2}\sqrt{S^2 \cdot C} < \theta_0 < \hat{\theta} + Z_{\alpha/2}\sqrt{S^2 \cdot C}$$

  only using $Z_{\alpha/2}$ instead of $t_{\alpha/2,n-2}$, and
- **approximate** hypothesis tests can be conducted via the pivot:

$$\frac{\hat{\theta} - \theta_0}{\sqrt{S^2 \cdot C}} \dot{\sim} Z \quad (\text{‘}\dot{\sim}\text{’ meaning ‘approximately distributed as’})$$

Hot chocolate sales (based on problem 4, Fall 2015)

Last winter, on Sundays, Alexander sold cups of hot chocolate by the ski tracks near his house. This winter he plans to have a similar business. Alexander experienced that the sales changed dramatically with the weather and skiing conditions. He made a condition index, x, where $x = 1$ means “bad conditions”, $x = 2$ means “good conditions”, $x = 3$ means “very good conditions” and $x = 4$ means “excellent conditions”.

Hot chocolate sales (based on problem 4, Fall 2015)

For 20 Sundays, $i = 1, \ldots, 20$, he registered both the number of cups sold, denoted y_i, and the associated condition index, x_i. We will phrase the sales as a regression model taking condition as an explanatory variable:

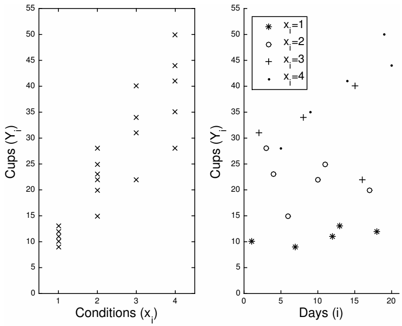
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, 20,$$

where $\epsilon_1, \ldots, \epsilon_{20}$ are independent variables with expected value 0 and variance σ^2, and β_0 and β_1 are fixed but unknown regression parameters.

Hot chocolate sales (based on problem 4, Fall 2015)

$\sum_{i=1}^{20}(x_i - \bar{x})^2 = 24.95$
$\frac{1}{18}\sum_{i=1}^{20}(y_i - \hat{\beta}_0 - \hat{\beta}_1)^2 = 5.65^2$
$\hat{\beta}_1 \approx 9.51.$

Questions:

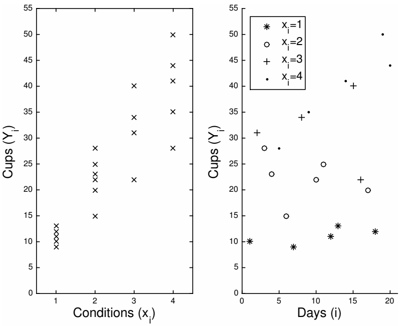


Hot chocolate sales (based on problem 4, Fall 2015)

$\sum_{i=1}^{20}(x_i - \bar{x})^2 = 24.95$
$\frac{1}{18}\sum_{i=1}^{20}(y_i - \hat{\beta}_0 - \hat{\beta}_1)^2 = 5.65^2$
$\hat{\beta}_1 \approx 9.51.$

Questions:

1. Would it be reasonable to construct a 95% confidence interval for β_0 or β_1?

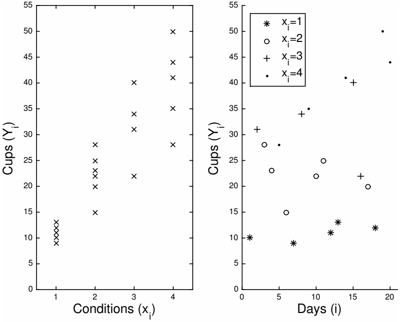


Hot chocolate sales (based on problem 4, Fall 2015)

$\sum_{i=1}^{20}(x_i - \bar{x})^2 = 24.95$
$\frac{1}{18}\sum_{i=1}^{20}(y_i - \hat{\beta}_0 - \hat{\beta}_1)^2 = 5.65^2$
$\hat{\beta}_1 \approx 9.51.$

Questions:

1. Would it be reasonable to construct a 95% confidence interval for β_0 or β_1?
2. Alexander thinks he can improve skiing conditions by 1 unit by grooming the snow, but for it to be worth it he would need to be 95% confident he would sell at least 8 more cups of hot chocolate on average. Use a 95% approximate confidence test to decide on your recommendation assuming $\varepsilon_1, \dots, \varepsilon_n$ are iid with mean 0 and variance σ^2.



Reminders

- Team based learning on Monday April 15