Final Exam Review

The format of the final exam will be similar to the in-class midterm. There will be mostly multiple choices questions with a few short answer prompts.

You will not be expected to write R code from scratch, but you may be asked about R.

Which of the following geometries require TWO variables?

- `geom_bar`
- `geom_col`
- `geom_histogram`
- `geom_density`

Which of the following geometries require TWO variables?

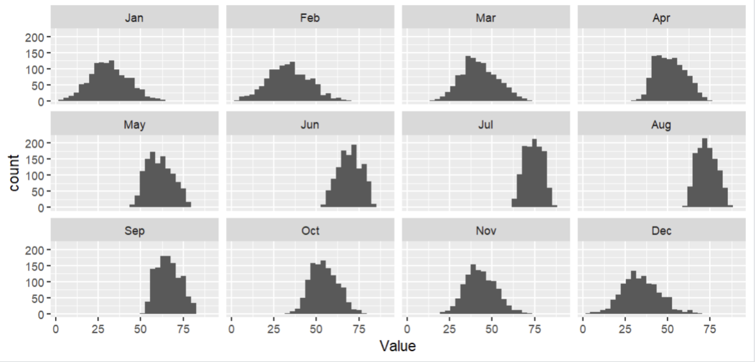
- ○ `geom_bar`
- ● `geom_col`
- ○ `geom_histogram`
- ○ `geom_density`

Other three automatically count occurrences.

You wish to visualize the distribution of Temperature (continuous) within each of 12 months.

Should you facet by Temperature or Month?

`facet_wrap` by month:



Which commands may behave differently when preceded by `group_by`? Select all that apply.

- `select`
- `mutate`
- `slice_max`
- `summarize`

Which commands may behave differently when preceded by `group_by`? Select all that apply.

- ○ `select`
- ● `mutate`
- ● `slice_max`
- ● `summarize`

`select` only changes columns (not affected by group level).

"myDF" has 50 rows. How many rows will the following output?

```
myData %>%
    group_by(group) %>%
    mutate(nRows = n())
```

"`myDF`" has 50 rows. How many rows will the following output?

```
myData %>%
    group_by(group) %>%
    mutate(nRows = n())
```

It will still have 50 rows! `mutate` adds a column.

"`df1`" has 5 rows, and "`df2`" has 4 rows; they have a variable in common, and 3 rows that match.

df1

	Variable A
Row 1	
Row 2	
Row 3	

df2

	Variable A
Row 1	
Row 2	
Row 3	

df1

	Variable A
Row 1	
Row 2	
Row 3	

df2

	Variable A
Row 1	
Row 2	
Row 3	

Rank the following by how many rows the resulting dataframe would have.

- `left_join`
- `right_join`
- `innner_join`
- `full_join`

df1

	Variable A
Row 1	
Row 2	
Row 3	

df2

	Variable A
Row 1	
Row 2	
Row 3	

Rows:

$$\texttt{inner} < \texttt{right} < \texttt{left} < \texttt{full}$$

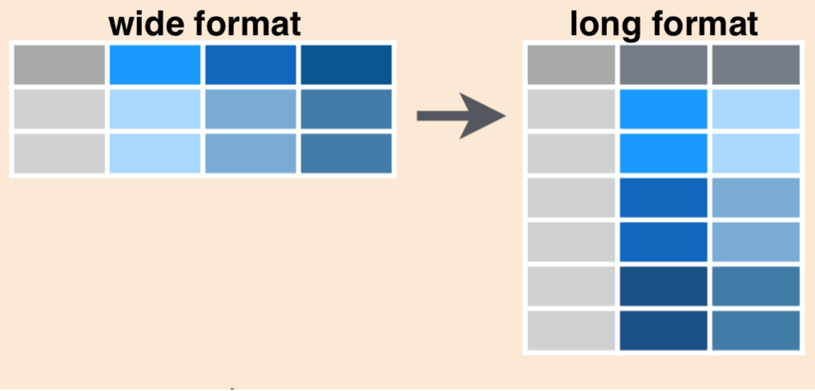
Which statements are true about `pivot_longer`? Select all that apply.

- ◯ It can decrease the number of rows.
- ◯ It can increase the number of rows.
- ◯ It can decrease the number of columns.
- ◯ It can increase the number of columns.

Which statements are true about `pivot_longer`? Select all that apply.

- ○ It can decrease the number of rows.
- ● It can increase the number of rows.
- ● It can decrease the number of columns.
- ○ It can increase the number of columns.

`pivot_wider` is the opposite.


wide format
long format

What value make X into a valid probability distribution?

X	1	2	3	4	5
P(X=x)	0.1	?	0.3	0.3	0.1

What value make X into a valid probability distribution?

X	1	2	3	4	5
P(X=x)	0.1	0.2	0.3	0.3	0.1

What is the cumulative probability $P(X \leq 4)$?

X	1	2	3	4	5
P(X=x)	0.1	0.2	0.3	0.3	0.1

X	1	2	3	4	5
P(X≤x)	0.1	0.3	0.6	?	1

What is the cumulative probability $P(X \leq 4)$?

X	1	2	3	4	5
P(X=x)	0.1	0.2	0.3	0.3	0.1

X	1	2	3	4	5
P(X≤x)	0.1	0.3	0.6	0.9	1

What is the 80th percentile of X?

X	1	2	3	4	5
P(X=x)	0.1	0.2	0.3	0.3	0.1

X	1	2	3	4	5
P(X≤x)	0.1	0.3	0.6	0.9	1

What is the 80th percentile of X?

X	1	2	3	4	5
P(X=x)	0.1	0.2	0.3	0.3	0.1

X	1	2	3	4	5
P(X≤x)	0.1	0.3	0.6	0.9	1

4 is the first x such that $P(X \leq x)$ is at least 0.8.

Consider bowling a ball at ten bowling pins, and let X be the total number of pins knocked down.

Does X follow a binomial distribution? Why or why not?

Consider bowling a ball at ten bowling pins, and let X be the total number of pins knocked down.

X is not binomial, since independence (and perhaps same probability) are not met.

Let $X \sim Binom(n, p)$. Which of the following returns $P(1 < X \leq 4)$?

- ○ `pbinom(4, n, p) - pbinom(1, n, p)`
- ○ `pbinom(5, n, p) - pbinom(0, n, p)`
- ○ `pbinom(4, n, p) - pbinom(2, n, p)`
- ○ `pbinom(3, n, p) - pbinom(1, n, p)`

Let $X \sim Binom(n, p)$. Which of the following returns $P(1 < X \leq 4)$?

- ● `pbinom(4, n, p) - pbinom(1, n, p)`
- ○ `pbinom(5, n, p) - pbinom(0, n, p)`
- ○ `pbinom(4, n, p) - pbinom(2, n, p)`
- ○ `pbinom(3, n, p) - pbinom(1, n, p)`

`pbinom` is $P(X \leq x)$.

True or False: these two commands will return the same value.

```
dbinom(0, size = 5, prob = 0.2)

pbinom(0, size = 5, prob = 0.2)
```

True or False: these two commands will return the same value.

```
dbinom(0, size = 5, prob = 0.2)

pbinom(0, size = 5, prob = 0.2)
```

True - 0 is the minimum.

Let $X \sim N(40, 5)$. Which of the following probabilities are equal? Select all that apply.

- ◯ $P(35 \leq x \leq 50)$
- ◯ $P(35 < x < 50)$
- ◯ $P(35 \leq x < 50)$
- ◯ $P(35 < x \leq 50)$

Let $X \sim N(40, 5)$. Which of the following probabilities are equal? Select all that apply.

- $P(35 \leq x \leq 50)$
- $P(35 < x < 50)$
- $P(35 \leq x < 50)$
- $P(35 < x \leq 50)$

$P(X = x) = 0.$

Let $X \sim N(40, 5)$. Which lines of R code calculate $P(35 \leq X \leq 50)$? Select all that apply.

- ◯ pnorm(35, 50)
- ◯ pnorm(2) - pnorm(-1)
- ◯ pnorm(50) - pnorm(35)
- ◯ pnorm(50, 40, 5) - pnorm(35, 40, 5)

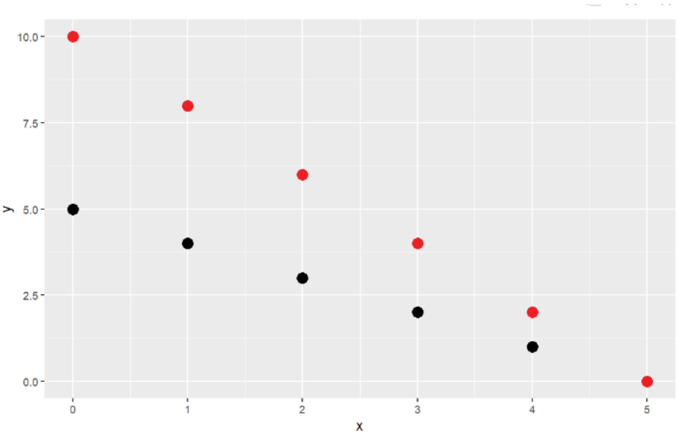
Let $X \sim N(40, 5)$. Which lines of R code calculate $P(35 \leq X \leq 50)$? Select all that apply.

- ○ pnorm(35, 50)
- ● pnorm(2) - pnorm(-1)
- ○ pnorm(50) - pnorm(35)
- ● pnorm(50, 40, 5) - pnorm(35, 40, 5)

No arguments = standard normal.

Sketch out two different scatterplots showing variables with correlation -1.

Is it possible to have two different plots?



Correlation 1 or -1 = perfect line.

Consider a dataset with
$\bar{x} = 0, \bar{y} = 50, s_X = 10, s_Y = 40, r = 0.5.$

Write expressions to find the estimated linear regression intercept and slope $\hat{\beta}_0$ and $\hat{\beta}_1$.

Consider a dataset with
$\bar{x} = 0, \bar{y} = 50, s_X = 10, s_Y = 40, r = 0.5.$

$$\hat{\beta}_1 \;=\; 0.5 \cdot \frac{40}{10} \;=\; 2$$

$$\hat{\beta}_0 \;=\; 50 - 2(0) \;=\; 50$$

Consider a dataset with
$\bar{x} = 0, \bar{y} = 50, s_X = 10, s_Y = 40, r = 0.5.$

$$\hat{y}_i \;=\; 50 + 2x_i$$

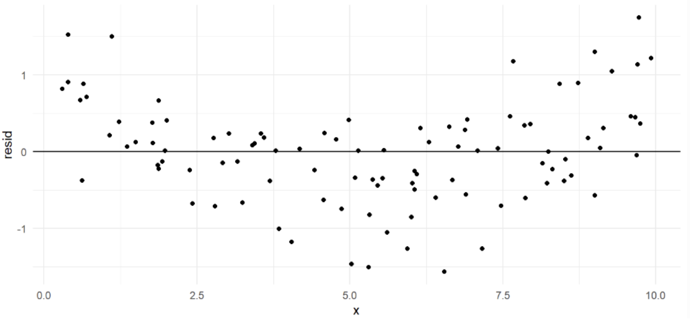
- What is the predicted y at $x = 4$?
- What is the predicted y at $x = 2$?
- Which estimate has more error?

Consider a dataset with
$\bar{x} = 0, \bar{y} = 50, s_X = 10, s_Y = 40, r = 0.5.$

$$\hat{y}_i \;=\; 50 + 2x_i$$

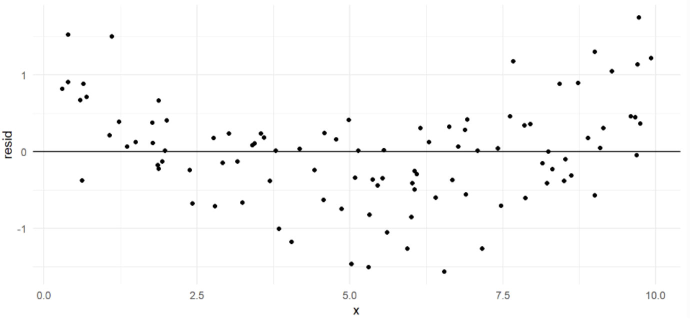
Predicting closer to $\bar{x}$ $(\hat{y} \mid 2)$ has smaller estimation error.

Which linear modeling assumption is not met?



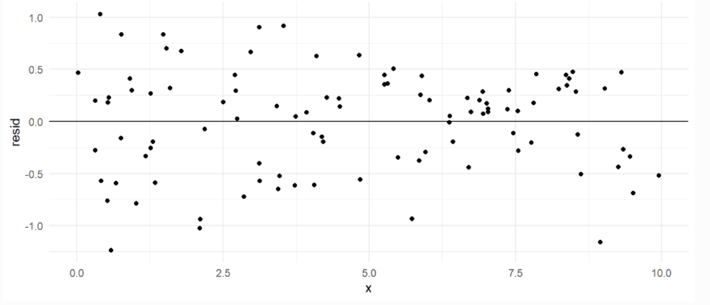
- Linearity
- Constant variance
- Normality
- None

Which linear modeling assumption is not met?



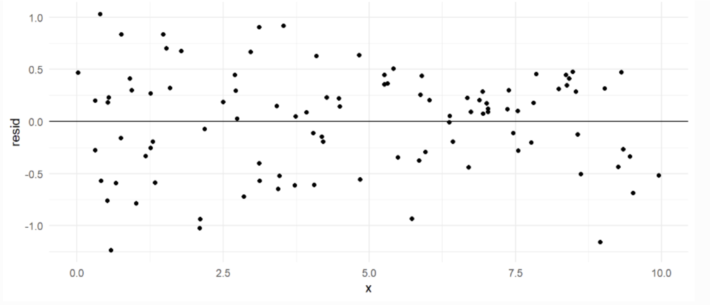
- Linearity
- Normality
- Constant variance
- None

Which linear modeling assumption is not met?



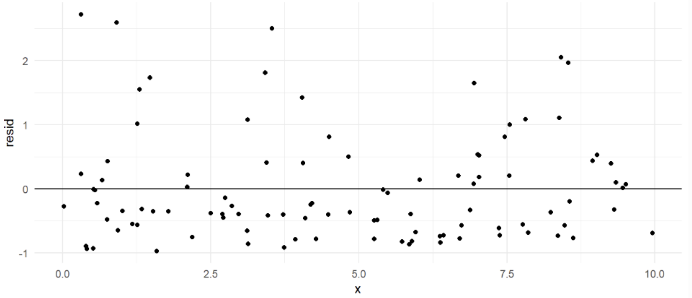
- Linearity
- Normality
- Constant variance
- None

Which linear modeling assumption is not met?



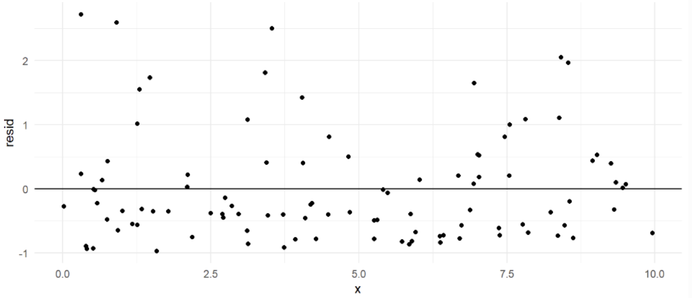
- ○ Linearity
- ○ Normality
- ○ Constant variance
- ● None

Which linear modeling assumption is not met?



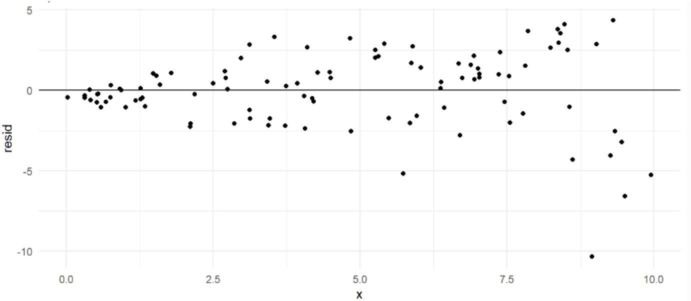
- Linearity
- Normality
- Constant variance
- None

Which linear modeling assumption is not met?



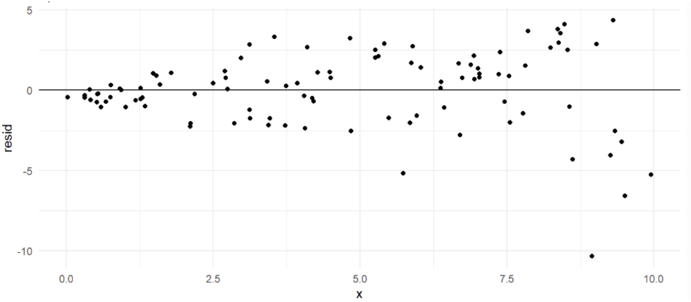
- ○ Linearity
- ● Normality
- ○ Constant variance
- ○ None

Which linear modeling assumption is not met?



- Linearity
- Normality
- Constant variance
- None

Which linear modeling assumption is not met?



- Linearity
- Normality
- Constant variance
- None

The R dataset `women`, and contains heights and weights for a random sample of 15 American women.
We wish to test for the existence of a relationship between height (x) and weight (y).

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height        3.45000    0.09114   37.85 1.09e-14 ***
---
```

How was the `t value` for slope calculated?

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height        3.45000    0.09114   37.85 1.09e-14 ***
---
```

$$t_{obs} = \frac{3.45 - 0}{0.09114} = 37.85$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height        3.45000    0.09114   37.85 1.09e-14 ***
---
```

How is the p-value calculated? $n = 15$

```
2 * (1 - p___(___, df = ___))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height        3.45000    0.09114   37.85 1.09e-14 ***
---
```

```
2 * (1 - pt(37.85, df = 13))
```

New hypotheses:
$H_0 : \beta_1 = 3.3 \quad \text{versus} \quad H_A : \beta_1 \neq 3.3$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height        3.45000    0.09114   37.85 1.09e-14 ***
---
```

Write an expression for the test statistic.

New hypotheses:

$H_0 : \beta_1 = 3.3 \quad \text{versus} \quad H_A : \beta_1 \neq 3.3$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height        3.45000    0.09114   37.85 1.09e-14 ***
---
```

$$t_{obs} = \frac{3.45 - 3.33}{0.09114}$$

Which of the following statements are true? Select all that apply.

- A 95% confidence interval for $E(y \mid x^*)$ is always wider than a 95% prediction interval for $y \mid x^*)$.
- For any type of interval, changing the confidence level from 95% to 99% will widen the interval.
- A 95% prediction interval for $y \mid x^*)$ at $\bar{x}$ is narrower than a 95% prediction interval at a different point.

Which of the following statements are true? Select all that apply.

- A 95% confidence interval for $E(y \mid x^*)$ is always wider than a 95% prediction interval for $y \mid x^*)$.
- For any type of interval, changing the confidence level from 95% to 99% will widen the interval.
- A 95% prediction interval for $y \mid x^*)$ at $\bar{x}$ is narrower than a 95% prediction interval at a different point.

Let's study a difference in proportions, $p_1 - p_2$.

Is there a difference in the proportion of people who are left-handed among basketball players and non-players?

	Left-Handed	Total
Basketball Players	49	538
Non-Players	64	500

Let p_B be the proportion of lefty basketball players, and p_N be the proportion of lefty non-players. State the model for this scenario.

Write hypotheses for a *difference* in proportions.

	Left-Handed	Total
Basketball Players	49	538
Non-Players	64	500

$$X_B \sim Binom(p_B, 538)$$

$$X_N \sim Binom(p_N, 500)$$

$$H_0 : p_B - p_N = 0 \quad \text{versus} \quad H_A : p_B - p_N \neq 0$$

	Left-Handed	Total
Basketball Players	49	538
Non-Players	64	500

What is the test statistic and null distribution for this two-sample proportion Z test? Write (do not evaluate) expressions.

$$z_{obs} = \frac{\frac{49}{538} - \frac{64}{500}}{SE}$$

$$SE = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{538} + \frac{1}{500}\right)}$$

$$\hat{p} = \frac{49+64}{538+500}$$

If the test statistic is -1.91, what R code calculates the two-sided p-value?

- `pnorm(-1.91)`
- `-pnorm(-1.91)`
- `2*pnorm(-1.91)`
- `2*(1-pnorm(-1.91))`

If the test statistic is -1.91, what R code calculates the two-sided p-value?

- ○ `pnorm(-1.91)`
- ○ `-pnorm(-1.91)`
- ● `2*pnorm(-1.91)`
- ○ `2*(1-pnorm(-1.91))`

The p-value is 0.056. What conclusions can we draw at the 5% level?

“Reject/Fail to reject the null” or “Significant/insignificant” result is not in-context.

	Left-Handed	Total
Basketball Players	49	538
Non-Players	64	500

Write expressions for the point estimate and standard error of $p_B - p_N$ with the Agresti-Coffe adjustment.

$$\hat{p}_{B,AC} = \frac{49+1}{538+2}, \qquad \hat{p}_{N,AC} = \frac{64+1}{500+2}$$

$$\text{pt est.} = \hat{p}_{B,AC} - \hat{p}_{N,AC}$$

$$\text{se} = \sqrt{\frac{\hat{p}_{B,AC}(1-\hat{p}_{B,AC})}{n_{B,AC}} + \frac{\hat{p}_{N,AC}(1-\hat{p}_{N,AC})}{n_{N,AC}}}$$

Let's study a difference in means, $p_1 - p_2$.

Is the average height of corn plants with growth treatment higher than the average height of corn plants with just water?

Is the average height of corn plants with growth treatment higher than the average height of corn plants with just water?

State the model and hypotheses for this scenario.

Is the average height of corn plants with growth treatment higher than the average height of corn plants with just water?

$$X_G \sim D_G(\mu_G, \sigma_G)$$
$$X_W \sim D_W(\mu_W, \sigma_W)$$

$$H_0 : \mu_G - \mu_W \leq 0, \quad \text{versus} \quad H_A : \mu_G - \mu_W > 0$$

	Average	Std. Dev	N
Growth Treatment	84.2	4.2	14
Water	80.8	3.4	13

Write an expression for the observed two-sample T test.

	Average	Std. Dev	N
Growth Treatment	84.2	4.2	14
Water	80.8	3.4	13

$$t_{obs} = \frac{84.2 - 80.8}{\sqrt{\frac{4.2^2}{14} + \frac{3.4^2}{13}}}$$

R output of one-sided Welch test:

```
        Welch Two Sample t-test

data:  x and y
t = 2.109, df = 24.61, p-value = 0.02265
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6535047       Inf
sample estimates:
mean of x mean of y
 84.22838  80.78093
```

How was the above p-value calculated? What would the p-value be if we were using a two-sided test?

```
        Welch Two Sample t-test

data:  x and y
t = 2.109, df = 24.61, p-value = 0.02265
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6535047       Inf
sample estimates:
mean of x mean of y 
 84.22838  80.78093 
```

- One-sided: `1 - pt(2.109, 24.61)` = 0.023
- Two-sided: `2*(1 - pt(2.109, 24.61))` = 0.046

	Average	Std. Dev	N
Growth Treatment	84.2	4.2	14
Water	80.8	3.4	13

Write an expression for a 95% CI for $\mu_G - \mu_W$. Include an R expression of the form `qt(__, df = 24.61)` for the critical value.

$$\text{point estimate} \;\; \pm \;\; \text{critical value} \;\; \times \;\; \text{standard error}$$

$$\text{pt est} = 84.2 - 80.8, \quad \text{cv} = \texttt{qt(0.975, df = 24.61)}$$

$$\text{SE} = \sqrt{\frac{4.2^2}{14} + \frac{3.4^2}{13}}$$