Vocabulary score vs. self identified social class

INFERENCE FOR NUMERICAL DATA IN R

Mine Cetinkaya-Rundel
Associate Professor of the Practice, Duke University

datacamp

Vocabulary score and self identified social class

- `wordsum` : 10 question vocabulary test (scores range from 0 to 10)
- `class` : self identified social class (lower, working, middle, upper)

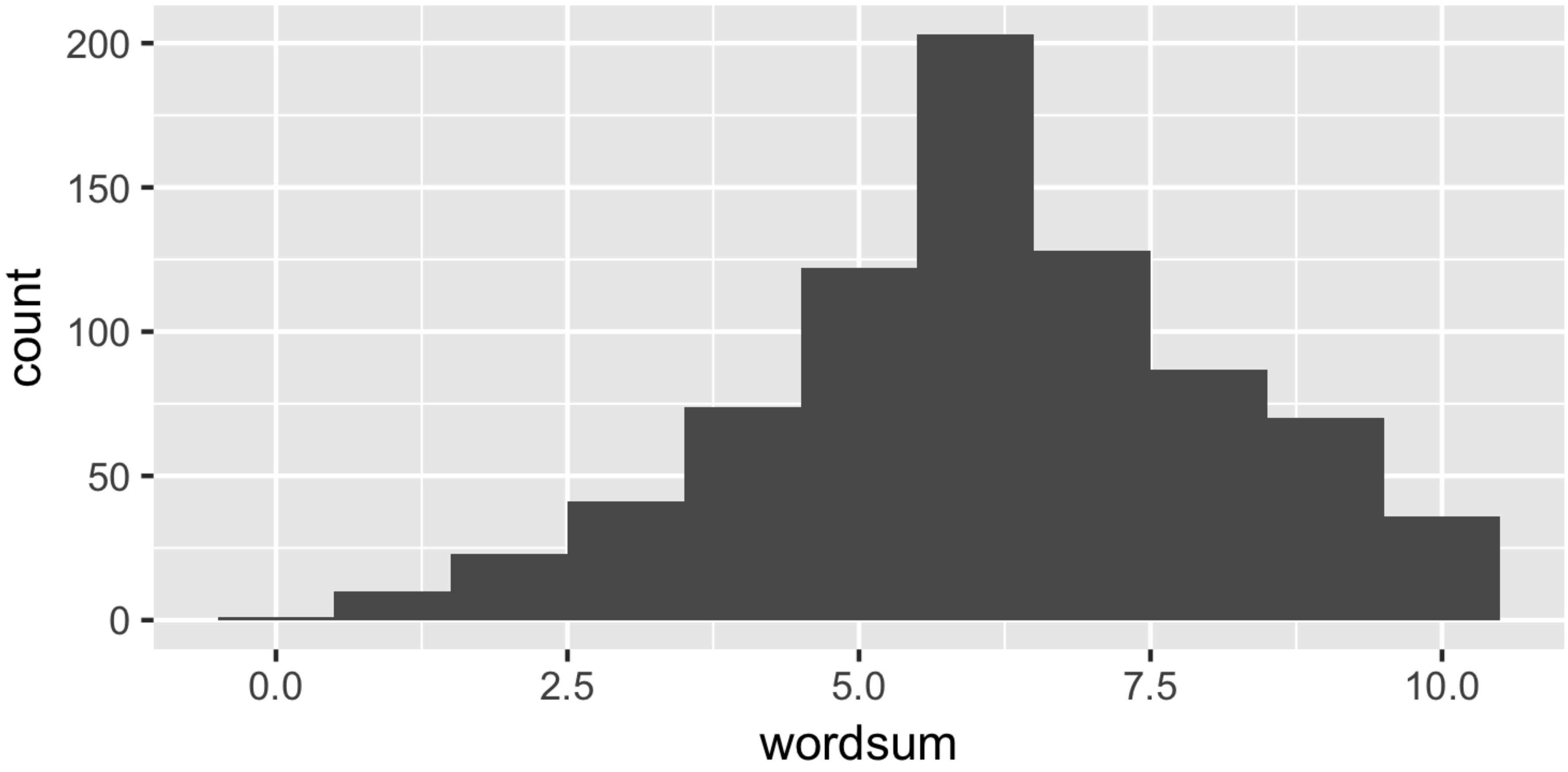
```
    wordsum  class
1         6 MIDDLE
2         9 WORKING
3         6 WORKING
4         5 WORKING
5         6 WORKING
6         6 WORKING
...     ...      ...
795       9 MIDDLE
```

1. SPACE (school, noon, captain, room, board, don't know)
2. BROADEN (efface, make level, elapse, embroider, widen, don't know)
3. EMANATE (populate, free, prominent, rival, come, don't know)
4. EDIBLE (auspicious, eligible, fit to eat, sagacious, able to speak, don't know)
5. ANIMOSITY (hatred, animation, disobedience, diversity, friendship, don't know)
6. PACT (puissance, remonstrance, agreement, skillet, pressure, don't know)
7. **CLOISTERED (miniature, bunched, arched, malady, secluded, don't know)**
8. CAPRICE (value, a star, grimace, whim, inducement, don't know)
9. ACCUSTOM (disappoint, customary, encounter, get used to, business, don't know)

0. ALLUSION (reference, dream, eulogy, illusion, aria, don't know)

Distribution of vocabulary score

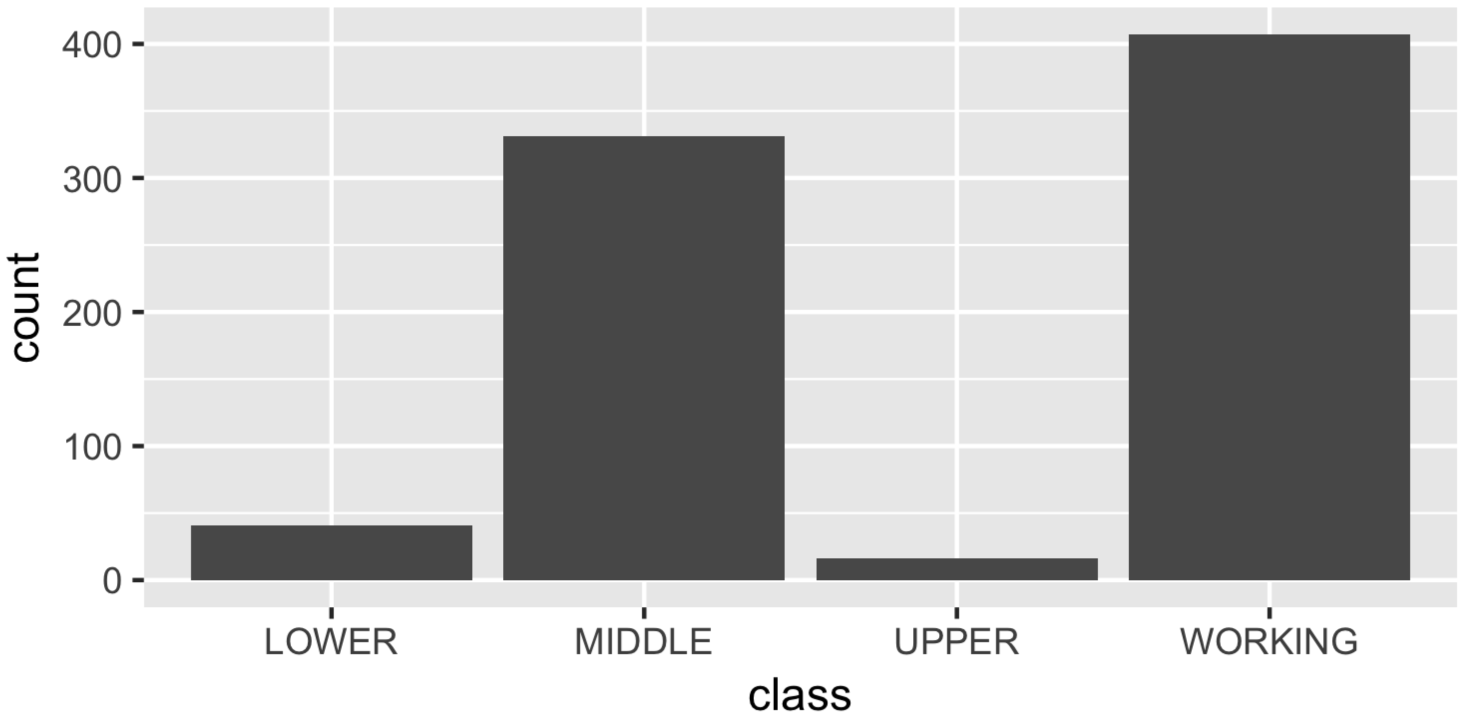
```
ggplot(data = gss, aes(x = wordsum)) +
  geom_histogram(binwidth = 1)
```



Self identified social class: `class`

If you were asked to use one of four names for your social class, which would you say you belong in: the lower class, the working class, the middle class, or the upper class?

```
ggplot(data = gss, aes(x = wordsum)) +
  geom_histogram(binwidth = 1)
```



Let's practice!

INFERENCE FOR NUMERICAL DATA IN R

ANOVA

INFERENCE FOR NUMERICAL DATA IN R

Mine Cetinkaya-Rundel

Associate Professor of the Practice,
Duke University

FINIS

ANOVA for vocabulary scores vs. self identified social class

H_0: The average vocabulary score is the same across all social classes, $\mu_{lower} = \mu_{working} = \mu_{middle} = \mu_{upper}$.

H_A: The average vocabulary scores differ between at least one pair of social classes.

Variability partitioning

Total variability in vocabulary score:

- Variability that can be attributed to differences in social class - **between group** variability
- Variability attributed to all other factor - **within group** variability

ANOVA output

```
library(broom)


aov(wordsum ~ class, gss) %>%
  tidy()
```

```
term         df      sumsq     meansq   statistic p.value
class         3   236.5644  78.854810  21.73467         0
Residuals   791  2869.8003   3.628066        NA        NA
```

Sum of squares

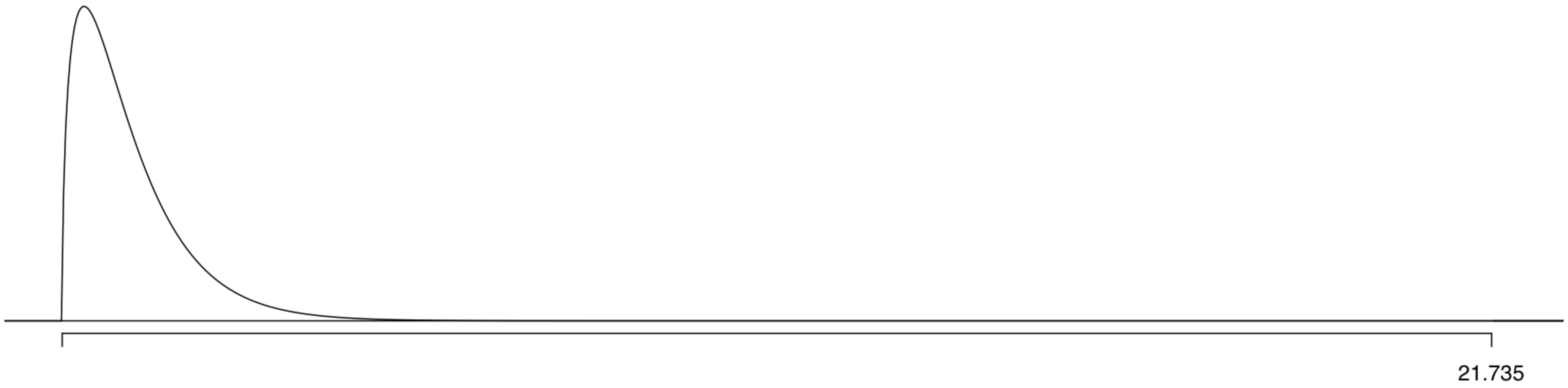
```
term        df      sumsq    meansq   statistic p.value
class        3   236.5644  78.854810   21.73467       0
Residuals  791  2869.8003   3.628066         NA      NA
```

- $SST = 236.5644 + 2869.8003 = 3106.365$ - Measures the total variability in the response variable
- Calculated very similarly to variance (except not scaled by the sample size)
- Percentage of explained variability = $\frac{236.5644}{3106.365} = 7.6\%$

F-statistic

```
term         df     sumsq    meansq   statistic p.value
class         3  236.5644 78.854810  21.73467       0
Residuals   791 2869.8003  3.628066        NA      NA
```

F-statistic = 21.73467 = $\frac{between\ group\ var}{within\ group\ var}$



Let's practice!

INFERENCE FOR NUMERICAL DATA IN R

Conditions for ANOVA

INFERENCE FOR NUMERICAL DATA IN R

Mine Cetinkaya-Rundel

Associate Professor of the Practice,
Duke University

datacamp

Conditions for ANOVA

- **Independence:**
  - within groups: sampled observations must be independent
  - between groups: the groups must be independent of each other (non-paired)
- **Approximate normality:** distribution of the response variable should be nearly normal within each group
- **Equal variance:** groups should have roughly equal variability

Independence

- **Within groups:** Sampled observations must be independent of each other
  - Random sample / assignment
  - Each n_j less than 10% of respective population always important, but sometimes difficult to check
- **Between groups:** Groups must be independent of each other
  - Carefully consider whether the groups may be dependent

Approximately normal

- Distribution of response variable within each group should be approximately normal
- Especially important when sample sizes are small
- Check with visuals

Constant variance

- Variability should be consistent across groups (homoscedasticity)
- Especially important when sample sizes differ between groups

Let's practice!

INFERENCE FOR NUMERICAL DATA IN R

Post-hoc testing

INFERENCE FOR NUMERICAL DATA IN R

Mine Cetinkaya-Rundel

Associate Professor of the Practice,
Duke University

datacamp

Which means differ?

- Two sample t-tests for differences in each possible pair of groups
- Multiple tests $\longrightarrow$ inflated Type 1 error rate
- Solution: use modified significance level

Multiple comparisons

- Testing many pairs of groups is called multiple comparisons
- The Bonferroni correction suggests that a more stringent significance level is more appropriate for these tests
    - Adjust α by the number of comparisons being considered
    - $\alpha^\star = \frac{\alpha}{K}$, where $K = \frac{k(k-1)}{2}$

Pairwise comparisons

- Constant variance → re-think standard error and degrees of freedom: Use consistent standard error and degrees of freedom for all tests
- Compare the p-values from each test to the modified significance level

Let's practice!

INFERENCE FOR NUMERICAL DATA IN R

Congratulations!

INFERENCE FOR NUMERICAL DATA IN R

Mine Cetinkaya-Rundel

Associate Professor of the Practice,
Duke University

Let's practice!

INFERENCE FOR NUMERICAL DATA IN R