Continuous data – ANOVA, multiple testing

Fundamental Rule of Data Analysis

Different types of data require different statistical analyses.

Number of Groups

- We’ve discussed estimation and testing for a continuous variable in one group
  - Compare the mean of one population to a known value
- We’ve discussed estimation and testing for a continuous variable in two groups
  - Compare the mean of one population to the mean of another population
- What if there are more than two groups?

>2 Groups

- Two-group t-test can be extended to accommodate more than two groups
- This extension is called one-way analysis of variance (ANOVA)

$$H_0\colon \mu_1 = \mu_2 = \mu_3$$

$$H_A\colon \text{at least one of the means differs from the others}$$

- ANOVA procedure returns a p-value. As usual, reject H_0 when the p-value is less than or equal to α. Fail to reject H_0 when the p-value is greater than α.

Analysis of Variance (ANOVA)

- We're comparing the mean in each group, so why is it called analysis of *variance*?
- When working with several different populations, two measures of variance can be calculated:
    1. The variation of the individual values around their own group means
    2. The variation of the group means around the overall combined mean

Variability

- Within-group variability: The variation of the individual values around their own group means
- Between-group variability: The variation of the group means around the overall combined mean

If between-group variability is large relative to within-group variability, there is evidence that the group means are different.

ANOVA F value: $\frac{\text{between−group variability}}{\text{within−group variability}}$

Low Birth Weight Data

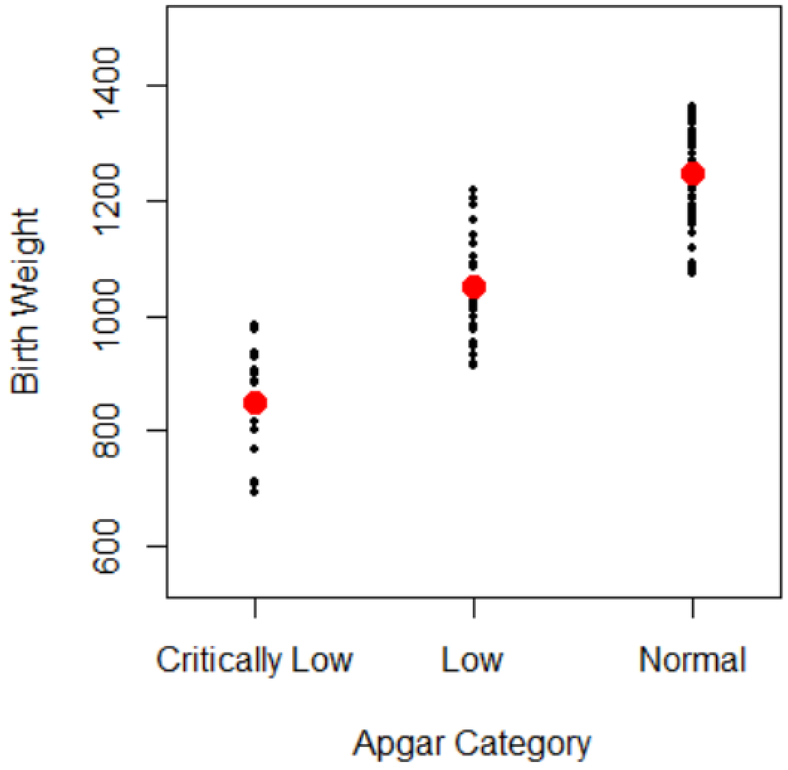
- Information on 100 low birth weight infants born in two teaching hospitals in Boston, Massachusetts

Variable	Description
birthwt	Birth weight of the baby (g)
momage	Mother's age (years)
apgar	Apgar score (integers, min=0, max=10). This is a scoring system used for assessing the clinical status of a newborn. 7 or higher is generally considered normal, 4-6 is low, and 3 or below is critically low.
apgar3	3-category variable indicating either Normal, Low, or Critically Low Apgar score

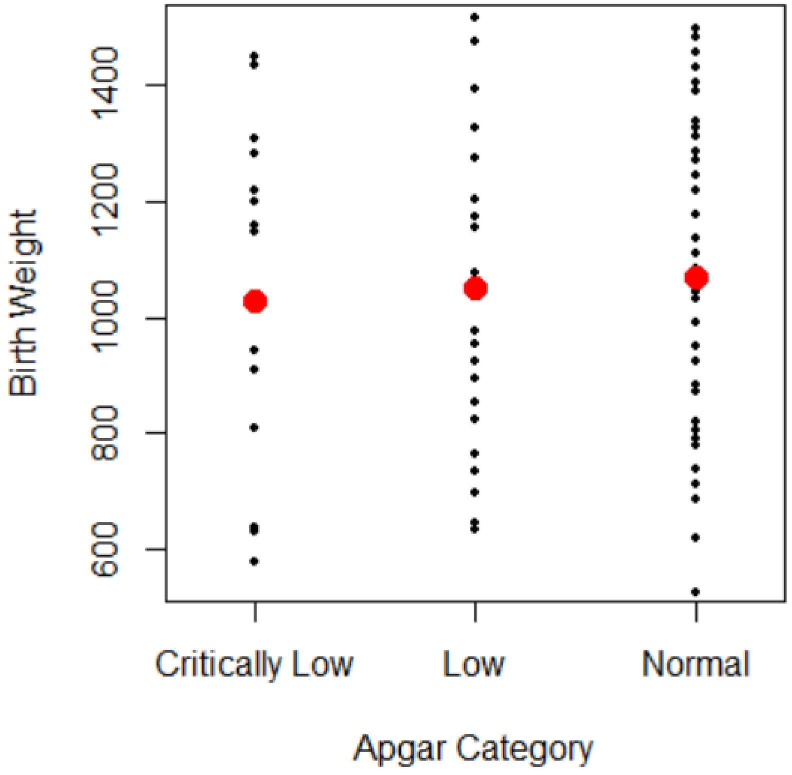
Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Hypothetical Example: Birth Weight by Apgar Category

High between-group variability
Low within-group variability



Low between-group variability
High within-group variability



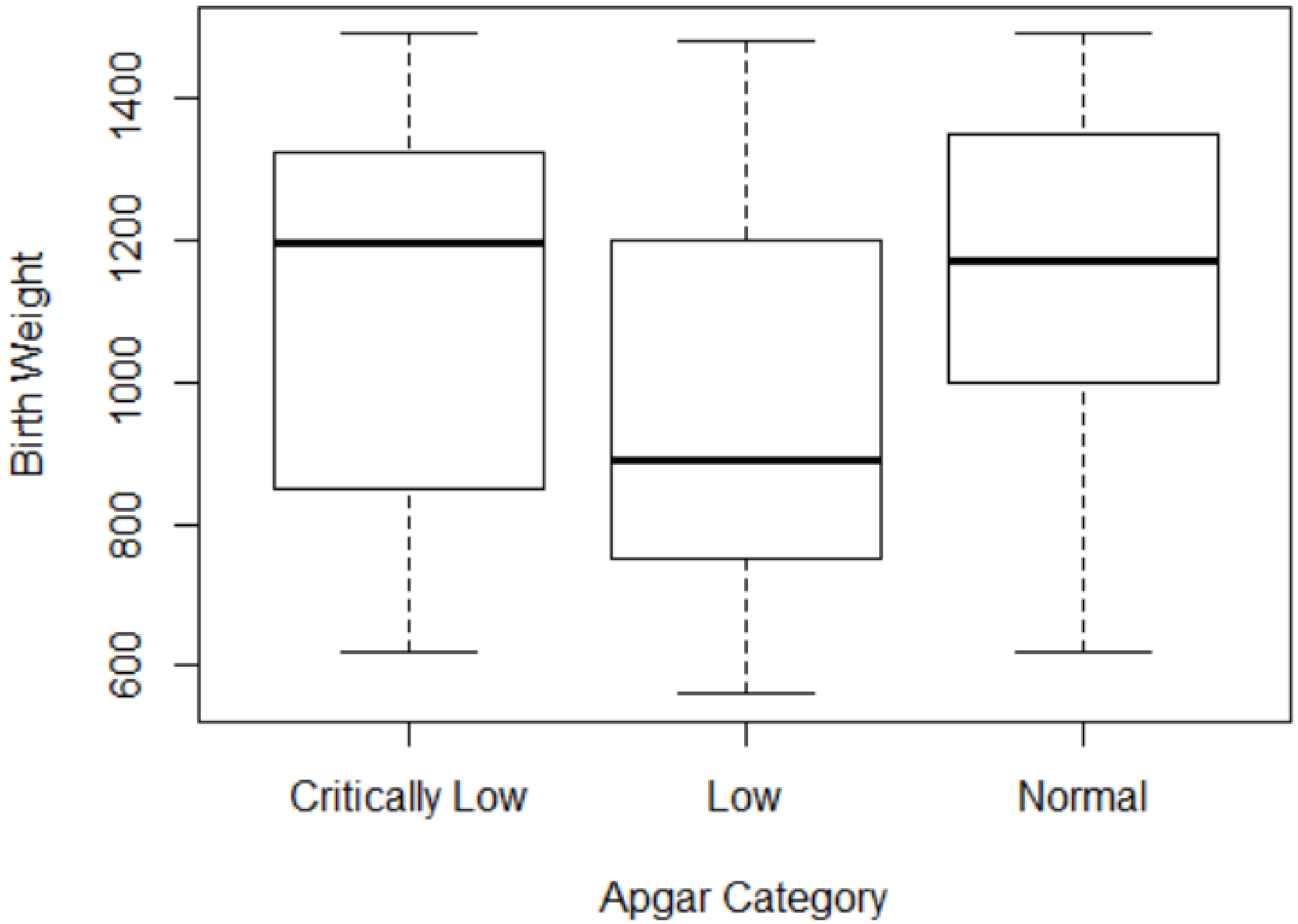
Example: Birth Weight by Apgar Category

- Researchers are interested in investigating whether the average birth weight differs between infants with a Critically Low, Low, or Normal Apgar score.

Sample summary statistics for birth weight in each Apgar score group:

	mean	sd	IQR	0%	25%	50%	75%	100%	birthwt:n
Critically Low	1091.7857	298.4744	433.75	620	875.0	1195	1308.75	1490	14
Low	974.6154	287.2383	425.00	560	757.5	890	1182.50	1480	26
Normal	1154.3333	240.3836	340.00	620	1010.0	1170	1350.00	1490	60

Example: Birth Weight by Apgar Category



Example: Birth Weight by Apgar Category

- Researchers are interested in investigating whether the average birth weight differs between infants with a Critically Low, Low, or Normal Apgar score.

mu1 = average birth weight of infants with critically low apgar score
mu2 = average birth weight of infants with low apgar score
mu3 = average birth weight of infants with normal apgar score

H0: mu1=mu2=mu3
HA: at least one of the means is different

Example: Birth Weight by Apgar Category

H_0: $\mu_1 = \mu_2 = \mu_3$ H_A: mean birth weight is different in at least one Apgar category

variance between groups

ANOVA model comparing average birth weight between Apgar groups:

```
            Df  Sum Sq Mean Sq F value Pr(>F)
apgar3       2  586693  293346   4.292 0.0164 *
Residuals   97 6630050   68351
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

variance within groups

p-value = 0.016

Since the p-value is less than 0.05, we reject H_0 and conclude that there is sufficient evidence to say that the average birth weight of babies is different in at least one of the three Apgar score groups (Normal/Low/Critically Low).

Multiple Testing

- ANOVA test can only tell us that at least one group mean is different… it does not tell us *which* group(s) are different from the others
- We could do two-group t-tests for each combination of groups

Test 1:	$H_0: \mu_1 = \mu_2$	$H_A: \mu_1 \neq \mu_2$
Test 2:	$H_0: \mu_1 = \mu_3$	$H_A: \mu_1 \neq \mu_3$
Test 3:	$H_0: \mu_2 = \mu_3$	$H_A: \mu_2 \neq \mu_3$

- Problem: What happens to our type 1 error rate when we perform 3 tests?

We set signficance level in advance to limit our type 1 error rate (usually 0.05). When we perform a hypothesis test on three different groups, this gives us 3 opportunities to make a type 1 error...so the type 1 error rate is greater than 5%. This is known as multiple testing problem

Tukey's Adjustment

- For performing all pairwise comparisons (Group 1 vs. Group 2, Group 1 vs. Group 3, Group 2 vs. Group 3)
- Do a standard two-group t-test then *adjust* the p-values using Tukey's adjustment
  - Makes the p-values a little larger
  - Harder to reach statistical significance → not as many type 1 errors
  - We say this "controls the family-wise type 1 error rate"
  - Ensures the overall probability that you make a type 1 error in at least one of the tests is α

So significance level in each individual test is, in effect, less than alpha

Example: Birth Weight by Apgar Category

H_0: $\mu_1 = \mu_2 = \mu_3$ H_A: at least one of the means differs from the others

ANOVA model comparing average birth weight between Apgar groups:

```
            Df  Sum Sq Mean Sq F value Pr(>F)
apgar3       2  586693  293346   4.292 0.0164 *
Residuals   97 6630050   68351
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value $= 0.016$

Since the p-value is less than 0.05, we reject H_0 and conclude that there is sufficient evidence to say that the average birth weight of babies is different in at least one of the three Apgar score groups (Normal/Low/Critically Low).

Example: Birth Weight by Apgar Category

family-wise type 1 error rate $\boxed{\alpha = 0.05}$

Pairwise comparisons of means using Tukey's adjustment:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = birthwt ~ apgar3, data = lowbwt)

$apgar3
                             diff        lwr       upr     p adj
Low-Critically Low     -117.17033 -323.45604  89.11538 0.3702135
Normal-Critically Low    62.54762 -122.15202 247.24726 0.7001920
Normal-Low              179.71795   33.60902 325.82688 0.0117441
```

Test 1:

H_0: $\mu_1 = \mu_2$ H_A: $\mu_1 \neq \mu_2$

Adjusted p-value = 0.37

Fail to reject H_0

Test 2:

H_0: $\mu_1 = \mu_3$ H_A: $\mu_1 \neq \mu_3$

Adjusted p-value = 0.70

Fail to reject H_0

Test 3:

H_0: $\mu_2 = \mu_3$ H_A: $\mu_2 \neq \mu_3$

Adjusted p-value = 0.01

Reject H_0

Multiple Testing

- Tukey's adjustment is just one of many types of adjustment for multiple testing
    - Only appropriate when performing all pairwise tests

- Bonferroni adjustment is most popular for performing a set of tests that are not pairwise
    - Large number of unrelated tests being performed

- All adjustments have the goal of maintaining your overall type 1 error rate to be α. This helps prevent false positive findings.

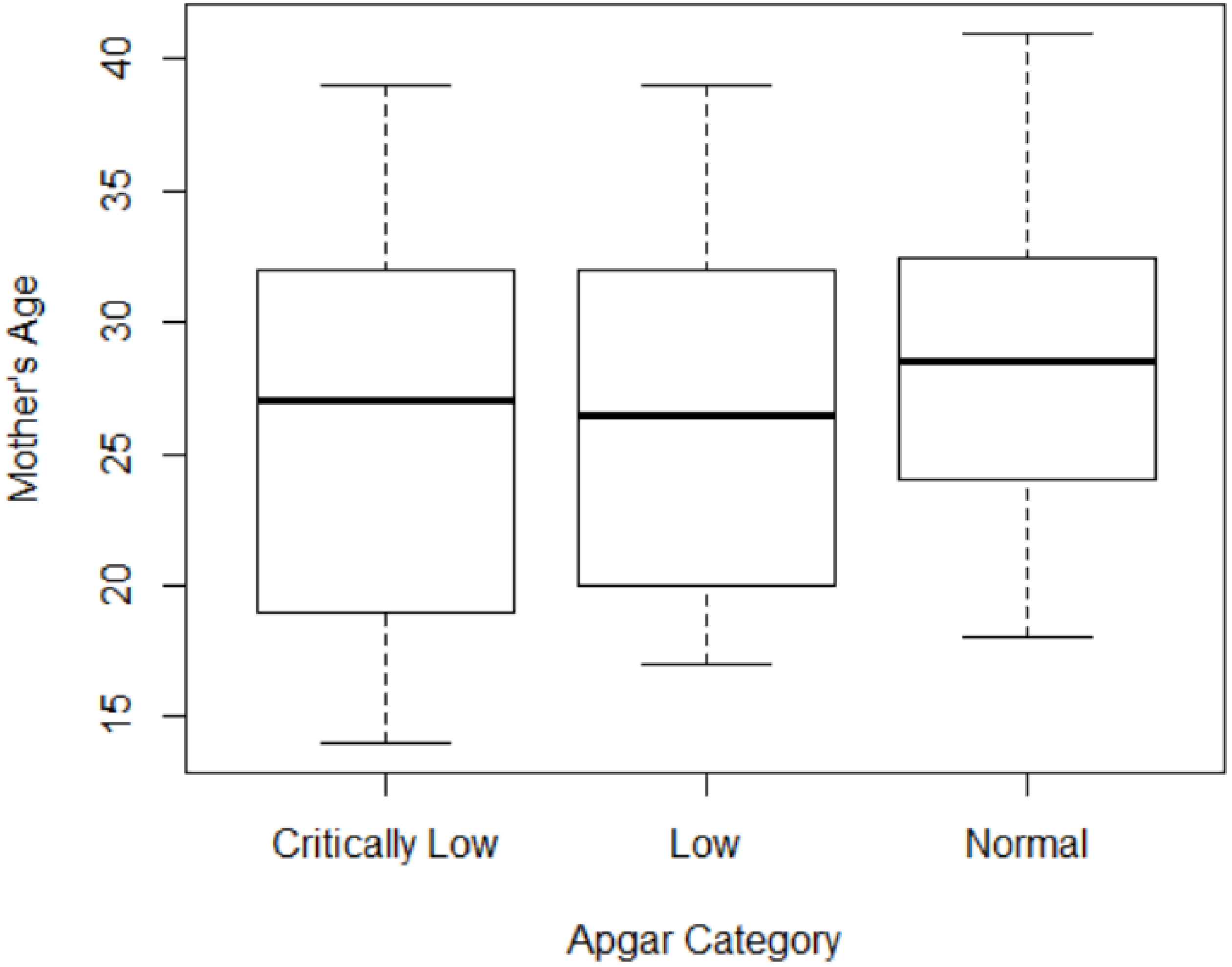
Example: Mothers' Age by Apgar Category

- Researchers are interested in investigating whether the average mothers' age differs between infants with a Critically Low, Low, or Normal Apgar score.

Sample summary statistics for mothers' age in each Apgar score group:

	mean	sd	data:n
Critically Low	26.71429	7.064849	14
Low	26.65385	6.626868	26
Normal	28.43333	5.403598	60

Example: Mothers' Age by Apgar Category



Example: Mothers’ Age by Apgar Category

- Researchers are interested in investigating whether the average age of the mother differs between infants with a Critically Low, Low, or Normal Apgar score.

Example: Mothers' Age by Apgar Category

H_0: $\mu_1 = \mu_2 = \mu_3$ H_A: at least one of the means differs from the others

ANOVA model comparing average mothers' age between Apgar groups:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
apgar3	2	74	37.12	1.038	0.358
Residuals	97	3469	35.77		

p-value = 0.358

Since the p-value is greater than 0.05, we fail to reject H_0 and conclude that there is not sufficient evidence to say that the average age of mothers is different for babies in any of the three Apgar score groups (Normal/Low/Critically Low).

Assumptions of ANOVA

- Groups/populations are independent
  - If not, may be able to use repeated measures ANOVA (beyond the scope of this course)
- Variance in each group is equal
  - If not, use Welch F-test not assuming equal variances (beyond the scope of this course)
- Underlying data are normally distributed
  - Central Limit Theorem helps us as long as the sample size in each group is large enough

Important Points

- Concepts of between-group and within-group variation
- Set up and interpretation of ANOVA
- Multiple testing
  - Interpreting results that use Tukey's adjustment
  - Why adjustment is necessary
  - Consequences of not adjusting