ASSIGNMENT 1 ON STATISTICAL INFERENCE WITH APPLICATION TO HEALTH

Student's Code	AIMS African Institute for Mathematical Sciences CAMEROON	Deadline
o03wY2R1W1		**16.11.25, 11:59 pm**
November 16, 2025		Ac. Year: 2025 - 2026

Lecturer(s): Mia Holley

Dataset Description

This report explores the `birthwt` dataset from the `MASS` package, which records 189 births and maternal characteristics from Baystate Medical Center (1986).[1].The birthwt data frame has 189 rows and 10 column

Dataset variables

Table below describes the dataset variables used.

Data dictionary for `birthwt`

Variable	Type	Description
low	Categorical (0/1)	Indicator of birth weight < 2.5 kg
age	Numeric	Mother's age (years)
lwt	Numeric	Mother's weight (pounds) at last menstrual period
race	Categorical (1/2/3)	1 = white, 2 = black, 3 = other
smoke	Categorical (0/1)	Smoking during pregnancy
ptl	Numeric	Number of previous premature labours
ht	Categorical (0/1)	History of hypertension
ui	Categorical (0/1)	Uterine irritability present
ftv	Numeric	First-trimester physician visits
bwt	Numeric	Infant birth weight (grams)

Distributions and Frequency Tables

Frequency tables

The two categorical variables chosen are race and smoking status (smoke). In the dataset, both variables are represented by number i.e, mother's race (1 = white, 2 = black, 3 = other) and smoking status (0 = non smoker, 1= smoker).

Smoking Status: Frequency, Relative, and Cumulative

Smoking Status	Count	Relative (%)	Cumulative (%)
Non-smoker	115	60.8	60.8
Smoker	74	39.2	100.0

[1] https://rdrr.io/cran/MASS/man/birthwt.html

Race: Frequency, Relative, and Cumulative

Race	Count	Relative (%)	Cumulative (%)
White	96	50.8	50.8
Black	26	13.8	64.6
Other	67	35.5	100.0

Frequency of race by smoking status

Race	Non-smoker	Smoker
White	44	52
Black	16	10
Other	55	12

Bar chat

My two chosen variables are **Race** and **Smoking status**.

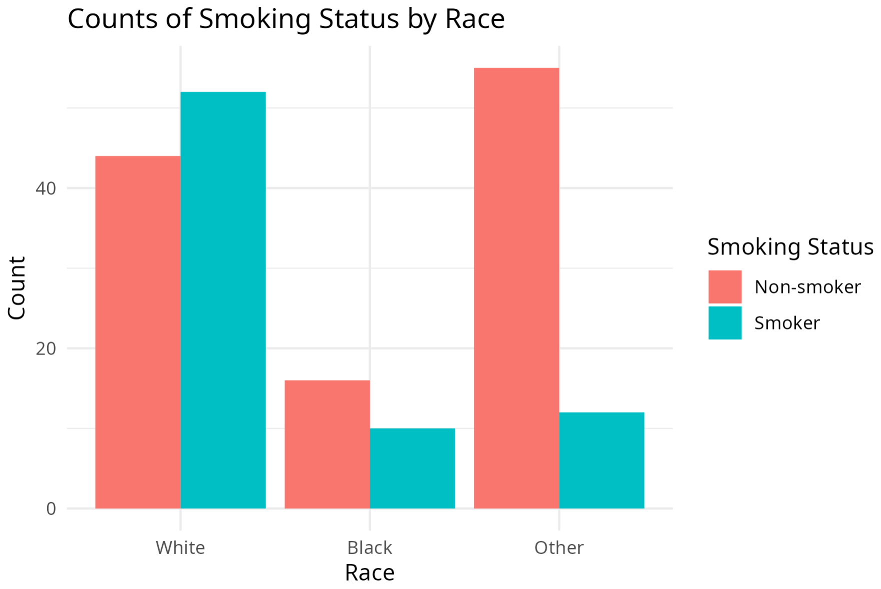


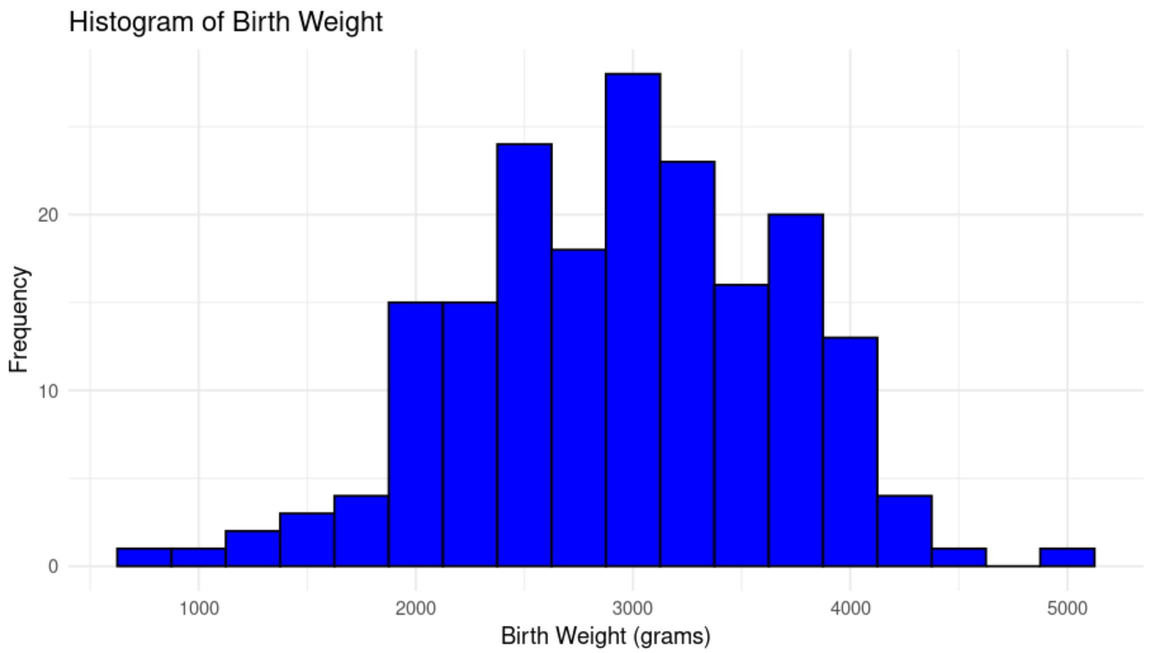
Figure 1

Interpretation:

As we can see, the race of mothers who smoke a lot during pregnancy are white women while other race category has the highest number of non-smokers who are pregnant.

Histogram

The created summary is that which describes the distribution of **Infant birth weight**(in grams).



Interpretation:

The histogram appears roughly bell-shaped, which suggests a normal distribution of birth weights. The highest bar is centered around $3000g$, indicating that most babies were born with a weight close to $3000g$. Birth weights range widely, but extreme low $< 2000g$ and high $> 4000g$ are less common. Quite frankly, the distribution is fairly symmetric.

Comparing Groups

Maternal Age by Smoking Status

Smoking Status	Median Age	IQR Age	SD Age	Min Age	Max Age
Non-smoker	23	6	5.47	14	45
Smoker	22	7	5.05	14	35

Maternal Weight (lbs) by Smoking Status

Smoking Status	Median LWT	IQR LWT	SD LWT	Min LWT	Max LWT
Non-smoker	124	29.5	28.4	85	241
Smoker	120	30	33.8	80	250

Interpretation

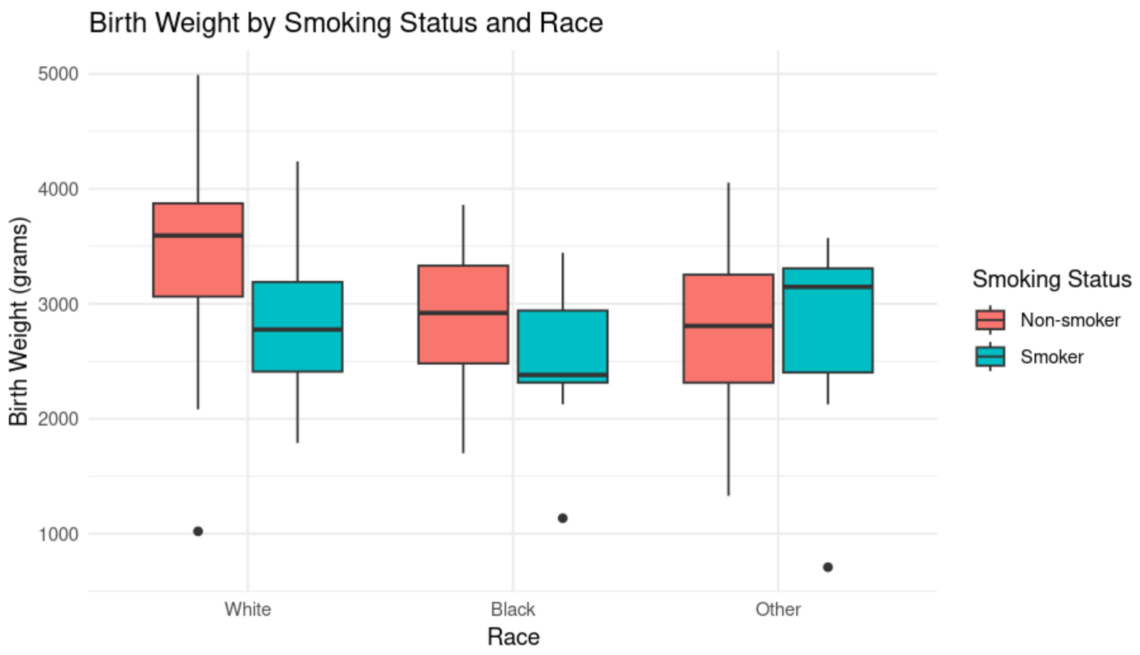
- Age: Non-smokers have a slightly higher median age (23 vs. 22). The variability (IQR and SD) is similar across groups.

- Maternal weight (LWT): Non-smokers show a slightly higher median maternal weight (124 lbs vs. 120 lbs). Smokers have greater variability (SD = 33.8 vs. 28.4).

- Range: Both groups include very young mothers (min age 14), but non-smokers extend to older ages (max 45 vs. 35).

Outcome - Birth Weight

A boxplot was created to compare birth weight of an infant by smoking status and race.

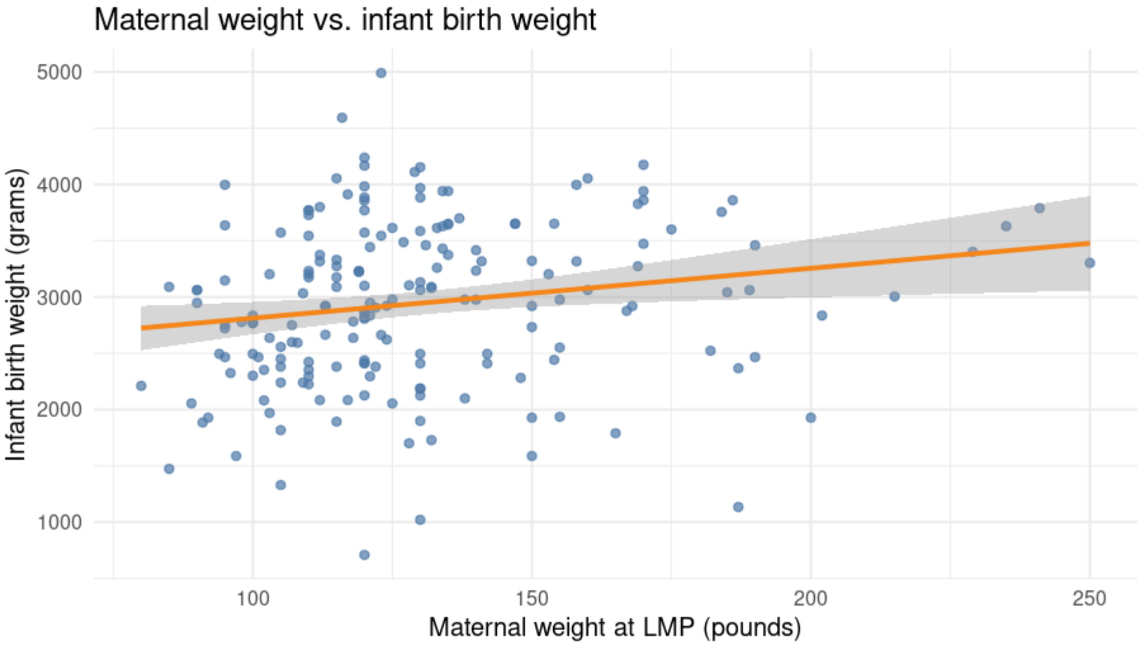


Intepretation

- **White:** Non smokers have the highest meadian birth weight among all race which drops in smokers. We can conclude that smoking reduces weight in white infants
- **Black:** Median birth weight is lower than white non-smokers, but still higher than black smokers. The black smokers have the lowest median birth rate overall with possible outliers. Smoking impactly impacts the birth weight in this group.
- **Other:** The impact of smoking in this group is less severe because the median in smoking women is higher than non smokers.

Relationships

A scatterplot was created to plot maternal weight vs. infant birth weight.



Interpretation

Our best fit line (regression) has a positive slope, which indicates a positive correlation between our 2 variables. Since the data points are spread out around the line, our correlation is not that strong. This is due to the fact that the martenal weight is not the only factor influencing birth weight.

Outliers: There are a few points that seem unusually low, say infant weight around $1000g$ with a martenal weight of $130 - 180lbs$ or high , infant weight close to $5000g$.