

# Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

R Demo for Chapter 11 : Inference for Two Means

Jinko Graham

2018-09-13

## Example Data: Low Birthweight Infants

- ▶ Data on 100 infants born with birth weight less than 1500g.
  - ▶ Variables are: head circumference (cm), birth length (cm), gestational age (wks), birth weight (g), mother's age (yrs), and toxemia (1=high blood pressure during pregnancy, 0=not)

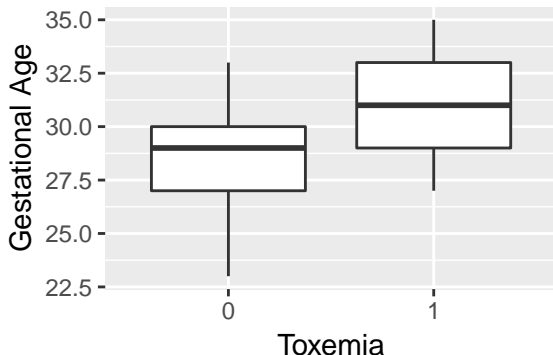
```
uu <- url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_17/Data/lbwt.csv")
lbwt <- read.csv(uu)
head(lbwt)
```

##	headcirc	length	gestage	birthwt	momage	toxemia
## 1	27	41	29	1360	37	0
## 2	29	40	31	1490	34	0
## 3	30	38	33	1490	32	0
## 4	28	38	31	1180	37	0
## 5	29	38	30	1200	29	1
## 6	23	32	25	680	19	0

# Gestational Age by Toxemia: Boxplots

- Explore differences graphically with boxplots:

```
library(dplyr)
lbwt <- mutate(lbwt,toxCateg = factor(toxemia))
library(ggplot2)
ggplot(lbwt,aes(x=toxCateg,y=gestage)) +
  labs(x="Toxemia",y="Gestational Age") + geom_boxplot()
```

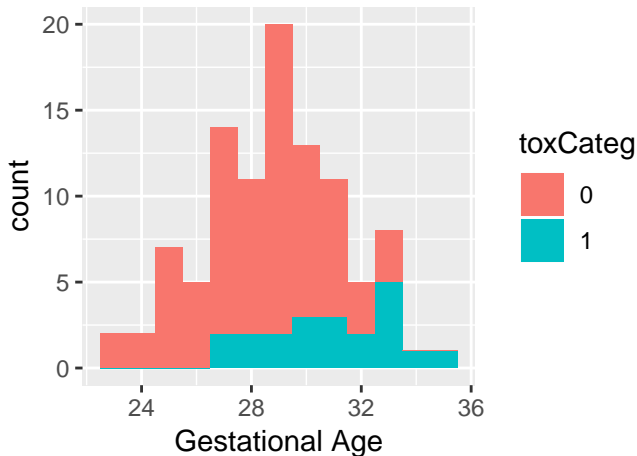


# Software Notes

- ▶ Use `mutate()` to make changes to a dataset.
  - ▶ In the example, we added a new variable called `toxCateg` to the `lbwt` dataset.
  - ▶ Whereas `toxemia` is a numeric variable with values 0 and 1, `toxCateg` is an explicitly categorical (or factor) variable, still having values 0 and 1.
  - ▶ In R, categorical variables are known as “factors”.
- ▶ We have used `ggplot()` to do the boxplots of `gestage` by `toxemia` categories (`toxCateg`).
  - ▶ For a boxplot, the call to aesthetic argument, `aes`, must specify an x-variable that is a factor.
  - ▶ `labs()` specifies the x- and y-axis labels.
  - ▶ `geom_boxplot()` adds the boxplots.
- ▶ Note: Data “wrangling”, processing and graphics take up 90% of an analyst’s time (that’s why data scientists get paid big bucks). Not to worry though because we will have templates to work from in this class.

# Gestational Age by Toxemia: Histograms

```
ggplot(lbwt, aes(x=gestage, fill=toxCateg)) +  
  labs(x="Gestational Age") + geom_histogram(binwidth=1)
```



## Software Note

- ▶ For a histogram, setting the aesthetic argument `fill` to `fill=toxCateg` specifies that the bars of the histogram are to be filled with different colors for the different categories of `toxCateg`.
  - ▶ Gives the impression of histograms stacked one upon the other.

## Gestational Age by Toxemia: Summary Statistics

- ▶ The sample means and SDs of gestational age for each toxemia category are summarized below.

```
library(dplyr)
lbwt %>%
  group_by(toxCateg) %>%
  summarize(mean=mean(gestage), sd=sd(gestage))
```

```
## # A tibble: 2 x 3
##   toxCateg mean    sd
##   <fct>    <dbl> <dbl>
## 1 0        28.4  2.32
## 2 1        30.9  2.32
```

# Software Notes

```
lbwt %>%  
  group_by(toxCateg) %>%  
  summarize(mean=mean(gestage), sd=sd(gestage))
```

- ▶ The code that produced the summaries should be read as:
  - ▶ Start with the `lbwt` dataset,
  - ▶ Group observations in this dataset by the variable `toxCateg`; i.e., partition the observations into groups defined by the categories of `toxCateg`, and
  - ▶ Summarize the sample mean and SD of the `gestage` variable within each group.
  - ▶ The “forward pipe” `%>%` is the “glue” that connects these steps together.
- ▶ This short video (time 5:22) gives a nice explainer on the forward pipe.



# Gestational Age Differences by Toxemia

Can use a 2-sample t-test:

```
t.test(gestage ~ toxCateg, data=lbwt, conf.level=0.90)
```

```
##
## Welch Two Sample t-test
##
## data:  gestage by toxCateg
## t = -4.4745, df = 31.465, p-value = 9.365e-05
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -3.516280 -1.584383
## sample estimates:
## mean in group 0 mean in group 1
##      28.35443      30.90476
```

- ▶ In the call to `t.test()`, the “formula” `gestage ~ toxCateg` tells R to think of `gestage` as a function of `toxCateg`.
  - ▶ We’ll use R formulas again when we study regression.
- ▶ The argument `conf.level` sets the level, or coverage probability,  $C$ , of the CI.

## Reading the output

```
##  
## Welch Two Sample t-test  
##  
## data: gestage by toxCateg  
## t = -4.4745, df = 31.465, p-value = 9.365e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 90 percent confidence interval:  
## -3.516280 -1.584383  
## sample estimates:  
## mean in group 0 mean in group 1  
## 28.35443 30.90476
```

- ▶ Under  $H_0$ , the test statistic  $T$ 's distribution is approximately a t distribution on  $\nu = 31.465$  degrees of freedom.
- ▶ The software compares the observed value -4.4745 of the test statistic to this reference distribution, to get a  $p$ -value of .00009366115.
- ▶ It gives the requested 90% CI for the difference in group means.
- ▶ It also gives us the sample means in the 2 groups.