Statistics Refresher for Faculty

Su-Miao Lai (sml10@uw.edu) and Haley Skipper (hskipper@uw.edu)

October 2, 2018

Acknowledgements

The part of material presented in these slides were developed and/or inspired by Dr. Timothy Thornton and Dr. Scott Emerson in the Department of Biostatistics at UW Seattle.

Outline

- How to conduct a research
- How to approach a data analysis
- Basic concepts review in Statistics
- Common statistical methods

How to conduct a research

- Ask a research question
- Consider a suitable study design
- Collect the data properly
- Pre-specify the statistcial methods
- Analyze the data
- Discuss the results

Research Question

- A research question is a scientific question, not every scientific question can be answered by a statistical question
- Identify overall goal of the study
- Identify the specific aims and how they relate to the overall goal
- Refine scientific hypotheses into the statistical hypotheses

Study Design

- Observational vs. Experimental
- Study subjects
- Sampling scheme/method
- Sample size

Data Collection

- Source of data
- Potential bias
- What type of variables you need for your statistical method
- Measurement of variables

Data Collection

- Levels of Measurement of Variables

Measurement	Description	Examples
Ratio	There is a natural zero starting point and ratios are meaningful.	Heights, lengths, distances, volumes, weights
Interval	Differences are meaningful, but there is no natural zero starting point and ratios are meaningless.	Body's temperature in degrees Fahrenheit or Celsius, IQ scores
Ordinal	Data can be arranged in order, but differences either can't be found or are meaningless.	Ranks of colleges in U.S. News & World Report
Nominal	Categories only. Data cannot be arranged in order.	Eye colors, blood types, gender, Religion

R Output

```
#>    No gender indays dx age education marriage residence religion caregiver
#> 1   1      1      5  2  89         4        2         1        1         1
#> 2   2      2      3  1  80         1        4         1        3         1
#> 3   3      2      5  1  83         1        4         4        3         3
#> 4   4      2      7  2  94         2        4         1        3         1
#> 5   5      1      6  1  92         6        2         1        2         2
#> 6   6      2     10  1  84         4        4         1        4         1
#> 7   7      1      4  1  90         6        2         1        1         1
#> 8   8      2      6  1  67         6        1         1        2         1
#> 9   9      1      7  1  94         3        2         2        1         1
#> 10 10      2      4  1  81         4        4         1        2         1
```

Exploratory Data Analysis

- Numerically
  - Measures of center (mean, median, mode)
  - Measures of spread (sd, variance, IQR, range)

Why summarize the data

- Identify errors in the data
- Verify the measurement of variables
- Identify pattern of missing data
- Summary statistics table

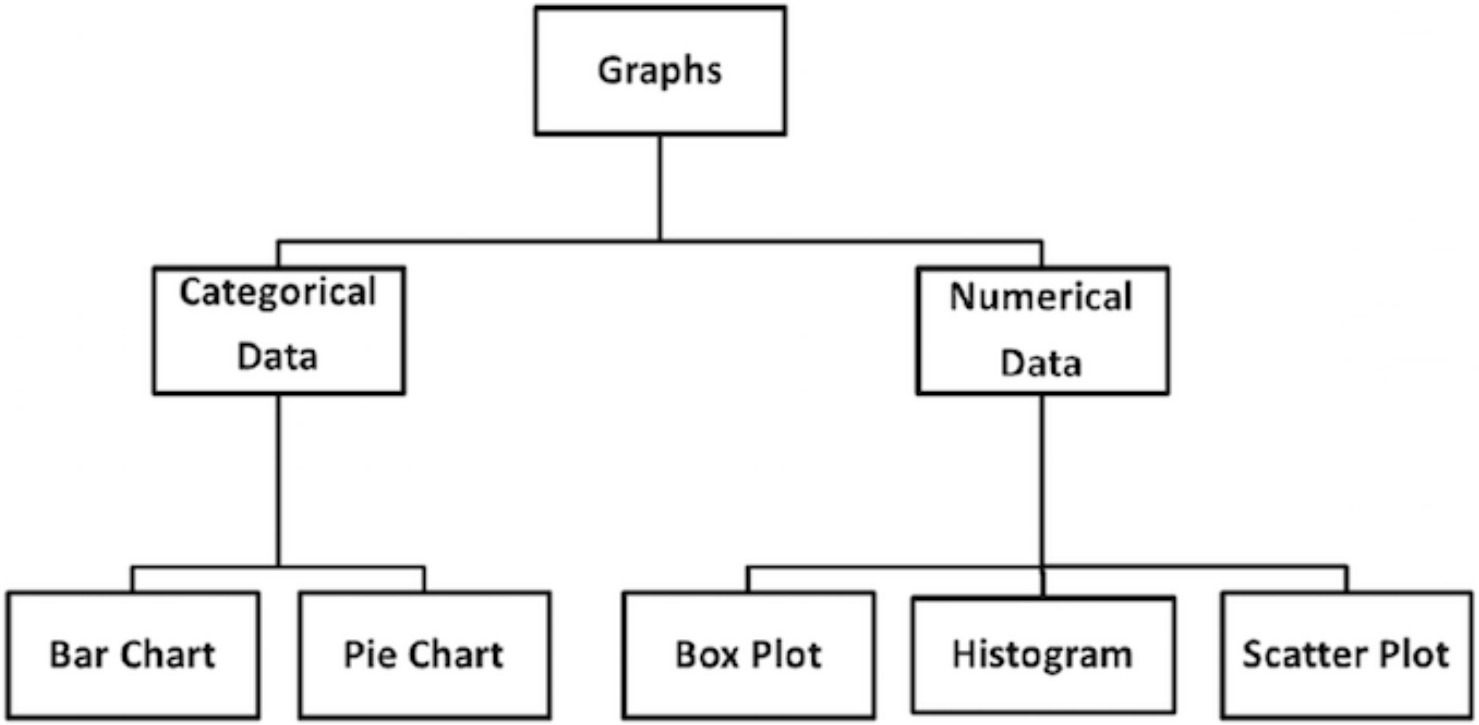
R Output

```
#>        No            gender        indays            dx
#>  Min.   : 1.00   Min.   :1.0   Min.   : 3.00   Min.   :1.0
#>  1st Qu.: 3.25   1st Qu.:1.0   1st Qu.: 4.25   1st Qu.:1.0
#>  Median : 5.50   Median :2.0   Median : 5.50   Median :1.0
#>  Mean   : 5.50   Mean   :1.6   Mean   : 5.70   Mean   :1.2
#>  3rd Qu.: 7.75   3rd Qu.:2.0   3rd Qu.: 6.75   3rd Qu.:1.0
#>  Max.   :10.00   Max.   :2.0   Max.   :10.00   Max.   :2.0
#>       age          education       marriage     residence      religion
#>  Min.   :67.0   Min.   :1.00   Min.   :1.0   Min.   :1.0   Min.   :1.00
#>  1st Qu.:81.5   1st Qu.:2.25   1st Qu.:2.0   1st Qu.:1.0   1st Qu.:1.25
#>  Median :86.5   Median :4.00   Median :3.0   Median :1.0   Median :2.00
#>  Mean   :85.4   Mean   :3.70   Mean   :2.9   Mean   :1.4   Mean   :2.20
#>  3rd Qu.:91.5   3rd Qu.:5.50   3rd Qu.:4.0   3rd Qu.:1.0   3rd Qu.:3.00
#>  Max.   :94.0   Max.   :6.00   Max.   :4.0   Max.   :4.0   Max.   :4.00
#>    caregiver
#>  Min.   :1.0
#>  1st Qu.:1.0
#>  Median :1.0
#>  Mean   :1.3
#>  3rd Qu.:1.0
#>  Max.   :3.0
```

Exploratory Data Analysis

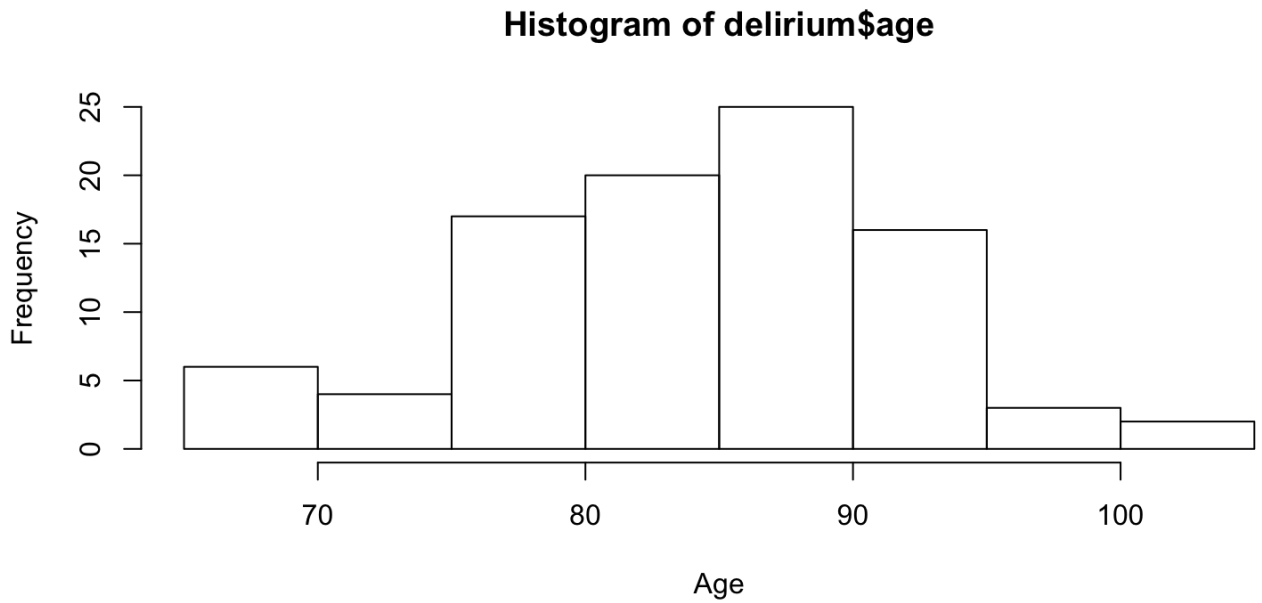
- Graphically
  - Histogram
  - Bar graph
  - Boxplot
  - Scatter plot

How to choose a graph



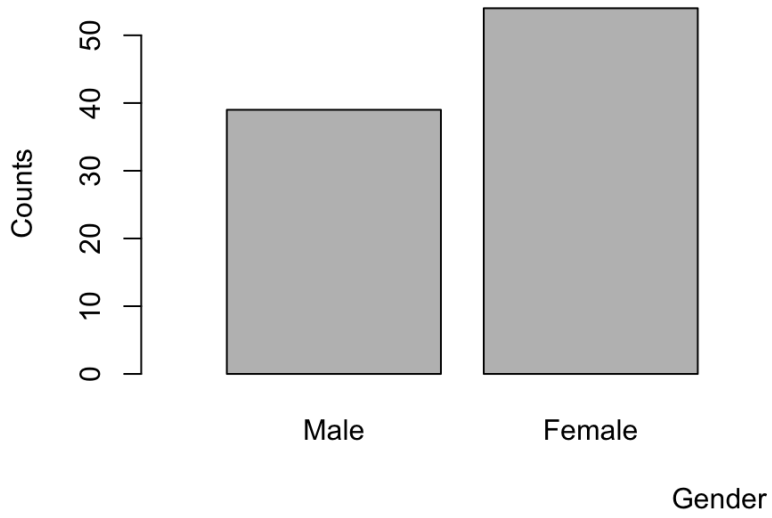
Histogram

```
hist(delirium$age, xlab="Age")
```



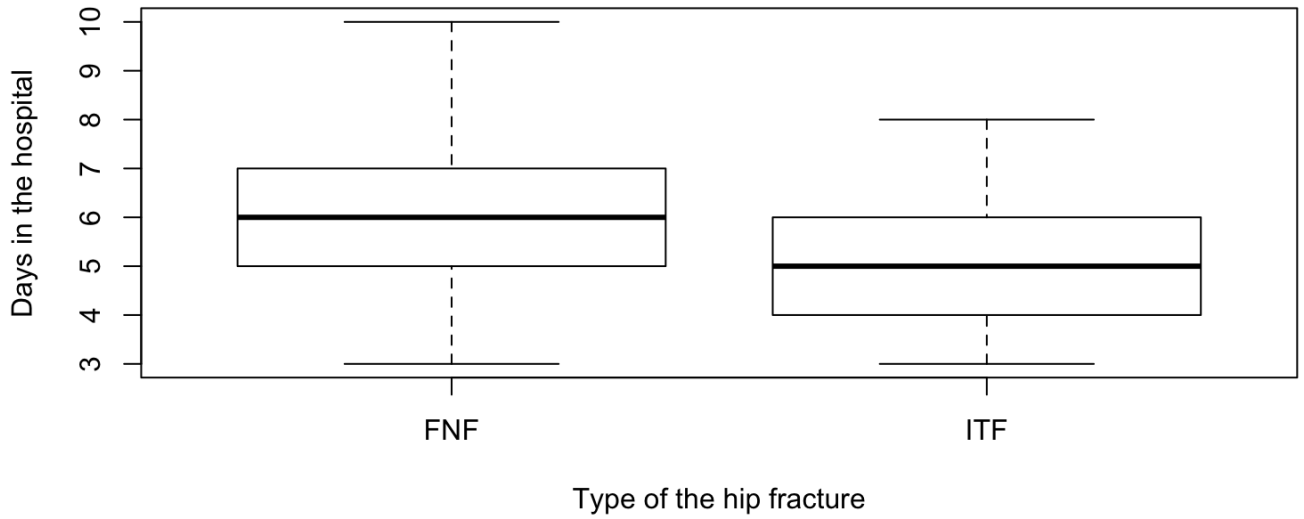
Bar Plot

```
counts <- table(delirium$gender)
barplot(counts, xlab="Gender", ylab="Counts",names.arg=c("Male", "Female"),
        xlim = c(0, 5), width = 1)
```



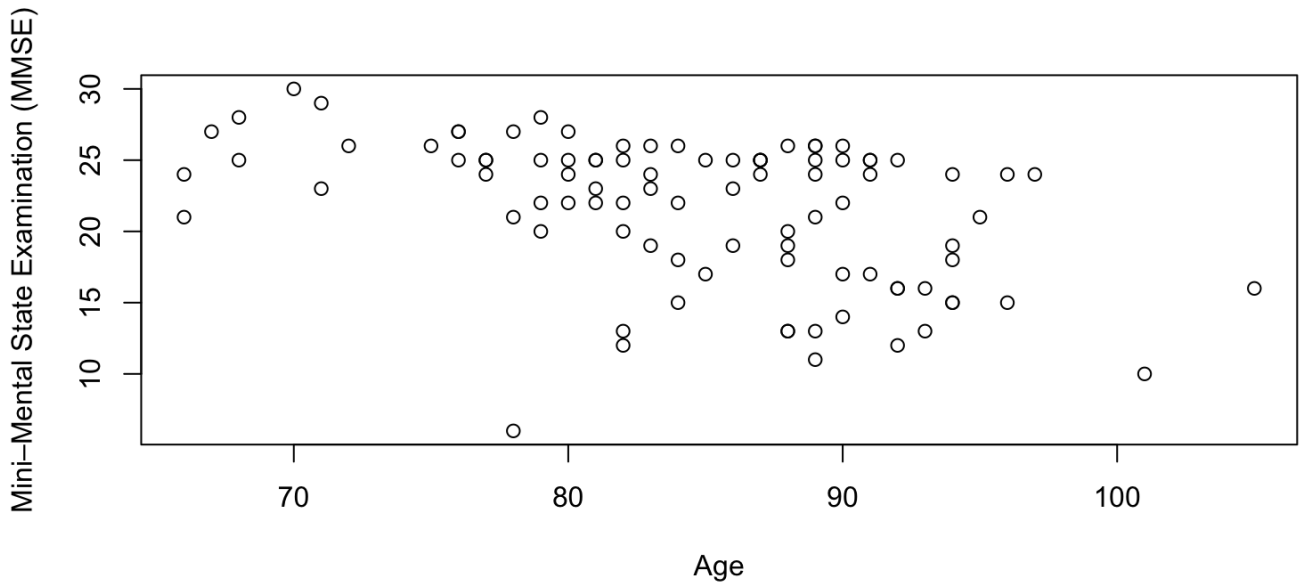
Boxplot

```
boxplot(indays~ as.factor(dx),data=delirium,
    xlab="Type of the hip fracture", ylab="Days in the hospital", names=c("FNF","ITF"))
```



Scatter plot

```
plot(delirium$age, delirium$MMSE,
     xlab="Age", ylab="Mini-Mental State Examination (MMSE)")
```



Statistical Hypotheses

- Hypothesis Testing Example
  - *H0: The defendant is innocent*
  - *Ha: The defendant is quilty*

	H0 is True (Truth is that defendant is innocent)	Ha is True (Truth is that defendant is guilty)
Do not reject H0 (I think the defendant is innocent)	Correct decision (Freed)	Type II error (β) (Freed)
Reject H0 (I think the defendant is guilty)	Type I error (α) (Convicted)	Correct decision (Convicted)

Caveat

- When we **reject** the null hypothesis, we conclude that the data provide sufficient evidence to support the alternative hypothesis.
- When we **fail to reject** the null hypothesis, we do not conclude that the data provide sufficient evidence to support the null hypothesis.

Common statistical methods

- Two independent samples t test
- One Way ANOVA (Analysis of Variance)
- Correlation Analysis
- Regression Modeling
- Chi-Squared test

Post-operative Delirium

- Delirium is a syndrome with acute but reversible changes in mental status
- Disorientation and inattention
- Imbalance
- Hallucinations

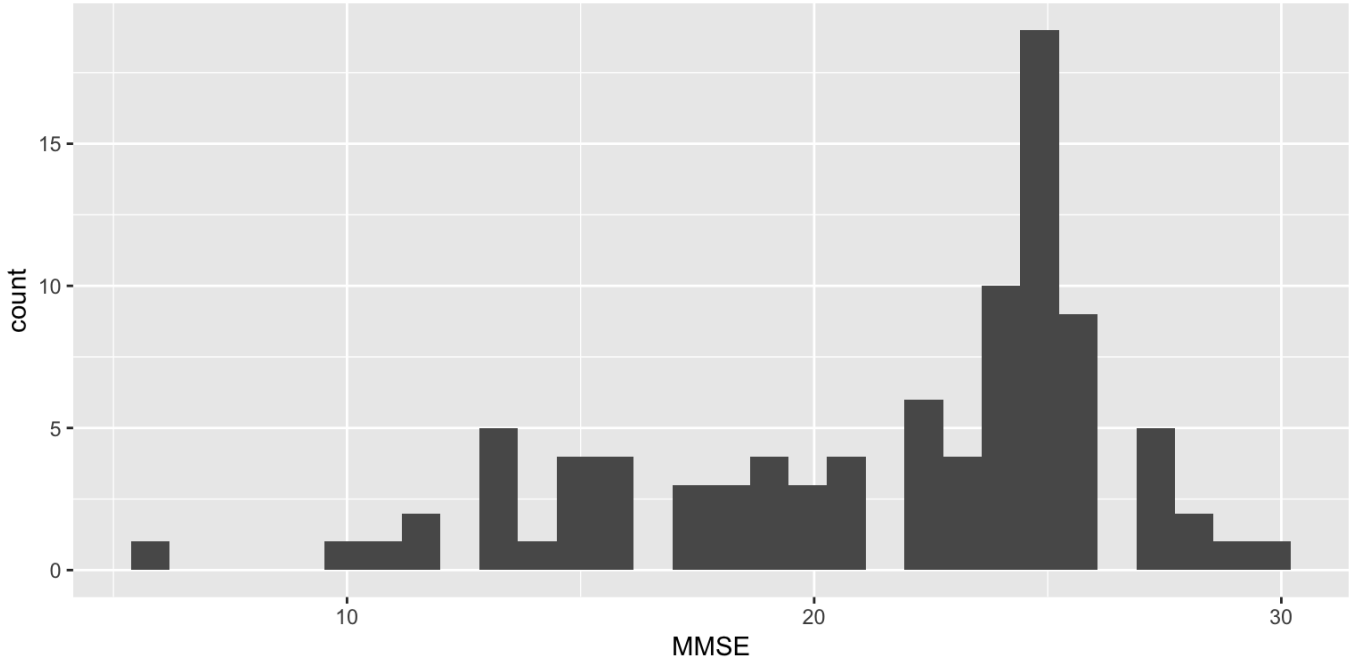
Study Subjects

- Convenience sample
- Orthopedic ward of medical center
- 93 patients
- All participants 65+ years old
- All articipants had hip fractures and underwent surgery
- 26 variables
    - Demographic Characteristics
    - Pre-Existing Health Conditions
    - Surgical Risk Factors

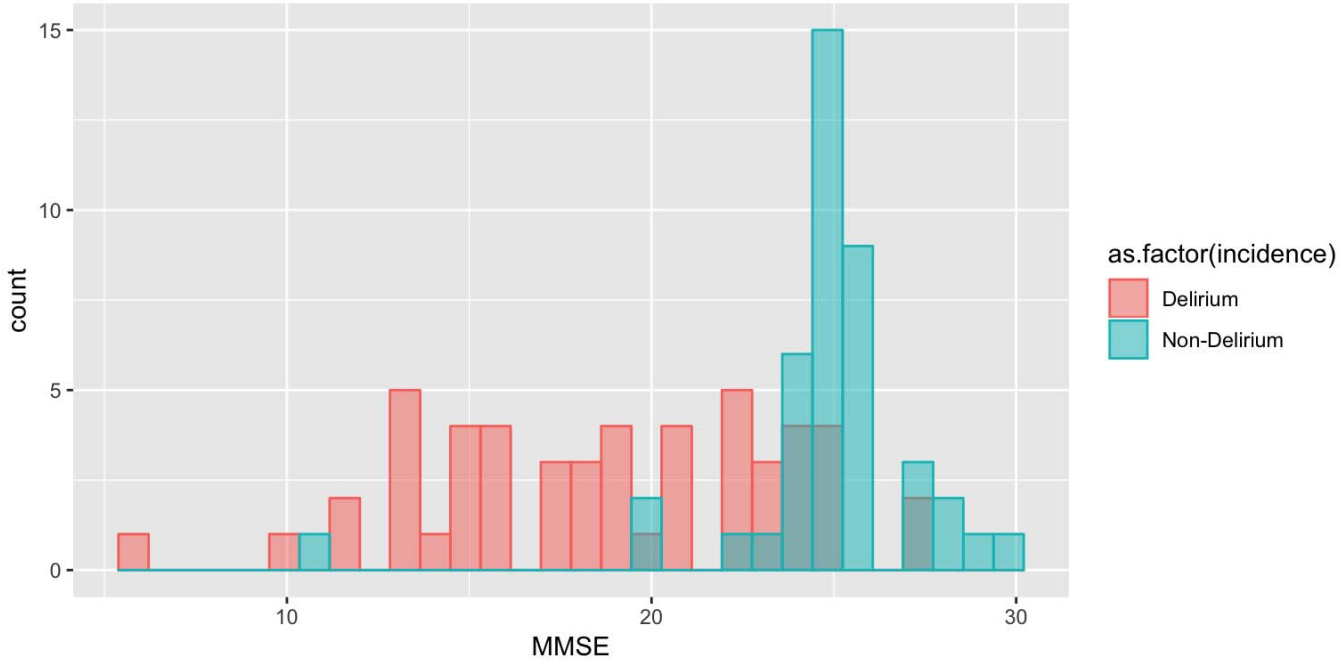
Two independent samples t test

- Compare the means of two independent samples
- Example of Hypotheses
  - *H0: There is no difference in means of MMSE between delirium group and non-delirium group*
  - *Ha: There is a difference in means of MMSE between delirium group and non-delirium group*
- Variables:
  - MMSE (Mini–Mental State Examination)
  - incidence (Delirium incidence)

R Output - Histogram



R Output - Histogram



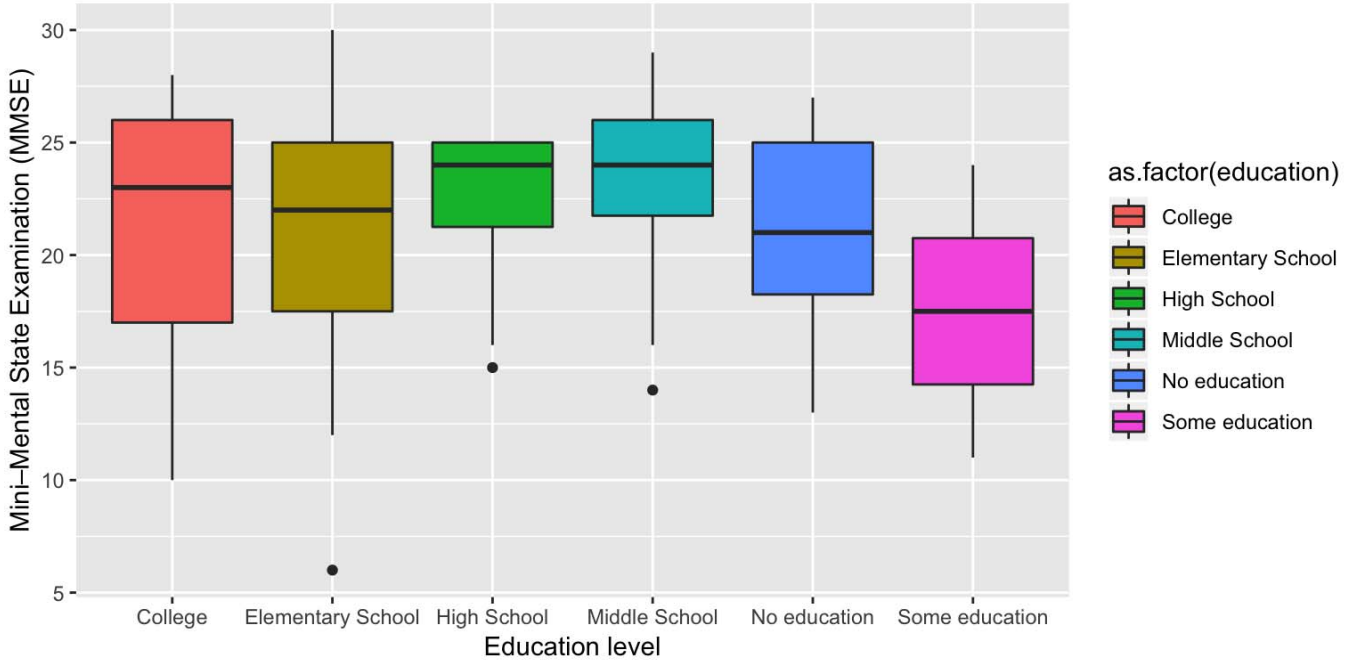
R Output

```
#>
#>  Welch Two Sample t-test
#>
#> data:  delirium$MMSE by delirium$incidence
#> t = -7.587, df = 84.075, p-value = 4.025e-11
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  -7.719341 -4.513152
#> sample estimates:
#>     mean in group Delirium mean in group Non-Delirium
#>                   18.76471                   24.88095
```

One Way ANOVA

- Compare the means of independent samples (more than two groups)
- Example of Hypotheses
  - *H0: There is no difference in means of MMSE among education level*
  - *Ha: There is a difference in means of MMSE among education level*
- Variables:
  - MMSE (Mini–Mental State Examination)
  - education (education level)

R Output - Box Plot



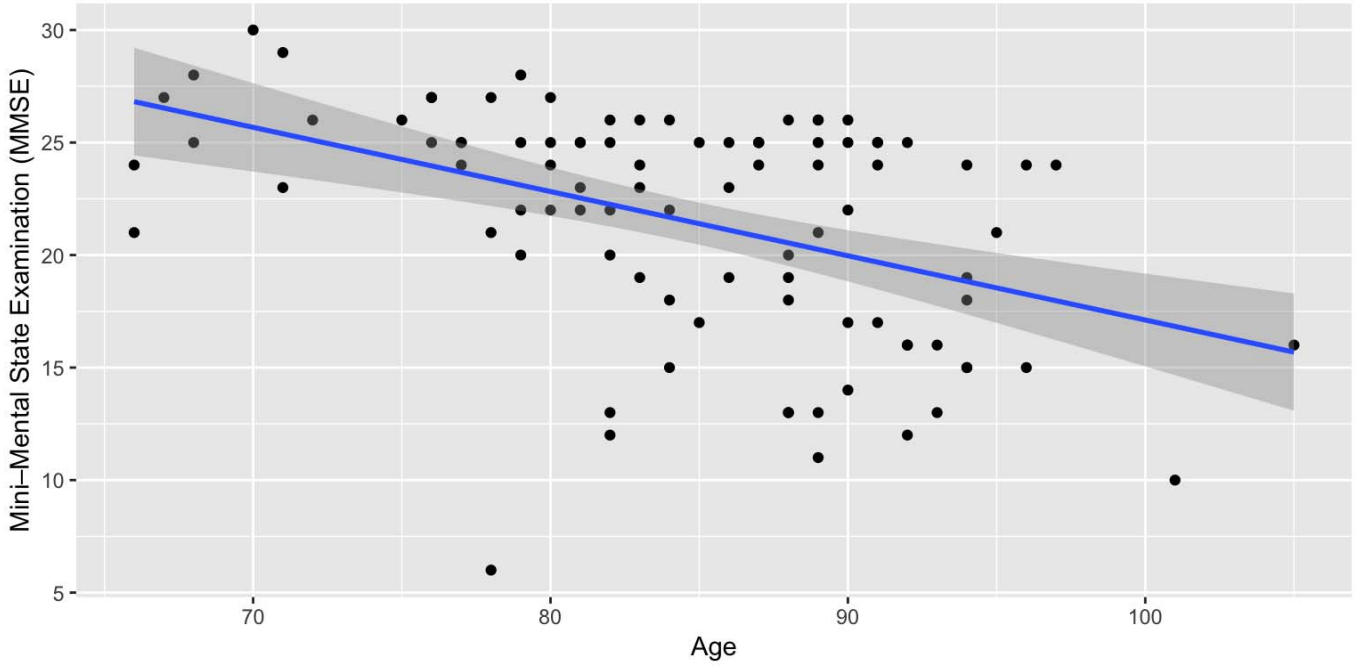
R Output

```
#>                     Df Sum Sq Mean Sq F value Pr(>F)
#> as.factor(education)  5   87.5   17.50   0.672  0.646
#> Residuals            87 2265.7   26.04
```

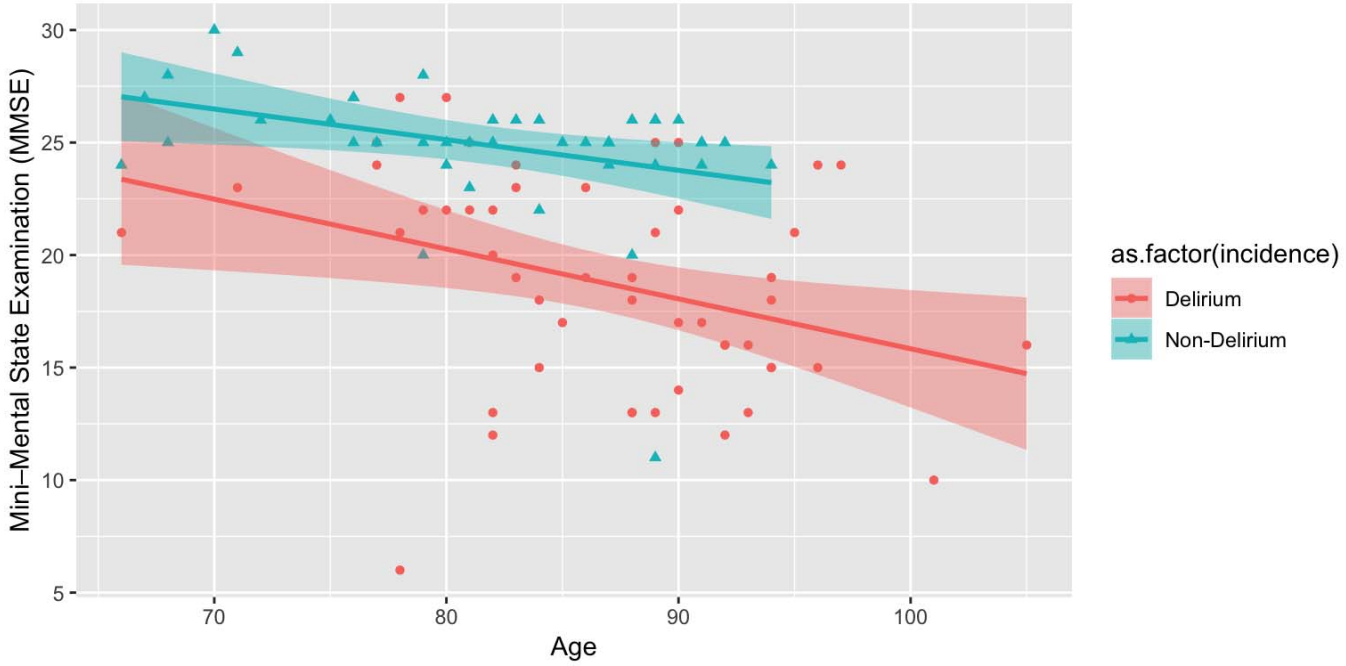
Correlation Analysis

- Correlation between two numerical variables
- Example of Hypotheses
  - *H0: Age is not correlated to MMSE*
  - *Ha: Age is correlated to MMSE*
- Variables:
  - Age
  - MMSE (Mini–Mental State Examination)

R Output - Scatter Plot



R Output - Scatter Plot



R Output

```
#>
#>  Pearson's product-moment correlation
#>
#> data:  delirium$age and delirium$MMSE
#> t = -4.7702, df = 91, p-value = 6.972e-06
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#>  -0.5966046 -0.2679574
#> sample estimates:
#>       cor
#> -0.4472524
```

Regression Modeling

- Association between an outcome of interest and a predictor/treatment of interest
- Example of Hypotheses
  - *H0: There will be no significant prediction of OpToAmb by MMSE*
  - *Ha: There will be a significant prediction of OpToAmb by MMSE*
- Variables:
  - MMSE (Mini–Mental State Examination)
  - OpToAmb (How many hours after surgery for patient to be able to leave bed)

R Output

```
#>
#> Call:
#> glm(formula = OpToAmb ~ MMSE, family = (gaussian), data = delirium)
#>
#> Deviance Residuals:
#>     Min       1Q   Median       3Q      Max
#> -16.390   -4.998   -1.998    2.502   26.281
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  25.7504     3.8144   6.751 1.34e-09 ***
#> MMSE         -0.3101     0.1725  -1.797   0.0756 .
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for gaussian family taken to be 70.05722)
#>
#>     Null deviance: 6601.5  on 92  degrees of freedom
#> Residual deviance: 6375.2  on 91  degrees of freedom
#> AIC: 663.09
#>
#> Number of Fisher Scoring iterations: 2
```

Chi-Squared test

- Association between two categorical variables
- Example of Hypotheses
  - *H0: Religion is not associated with delirium incidence*
  - *Ha: Religion is associated with delirium incidence*
- Variables:
  - religion (religion)
  - delirium (delirium incidence)

R Output

```
#>
#>              Delirium Non-Delirium
#>   Buddhism          15           13
#>   Catholicism        0            5
#>   Christian          6            6
#>   None              26           10
#>   Taoism             4            8
#>
#>  Pearson's Chi-squared test
#>
#> data:  tbl
#> X-squared = 12.837, df = 4, p-value = 0.0121
```

What to report

- Point estimate
- Confidence interval (lower, upper bound)
- P value
  - p-value is the probability of obtaining a result that is as extreme as or more extreme than the one observed, assuming the null hypothesis is true.
  - Small p-values provide evidence against the null hypothesis; larger p-values don't.
  - A small p-value is the same as a result being statistically significant.

Something to keep in mind

- Determine what your research question is and refine the question of interest into the specific aim (statistical question).
- Spend more time to study your variables and to collect data in a proper way in order to maintain the quality of data.
- Always double check what kind of variables you are going to use (numerical or categorical) before running the statistical test.
- Properly defining the measurement of variables will save you a lot of time while running the statistical tests on RStudio.
- Be sure of what you are testing– detecting the association or comparing the means.

Questions?

- UW Tacoma Statistics Consulting Group http://www.tacoma.uw.edu/office-research/statistical-consulting-analytical-support