Biostatistics Hands-on Workshop:

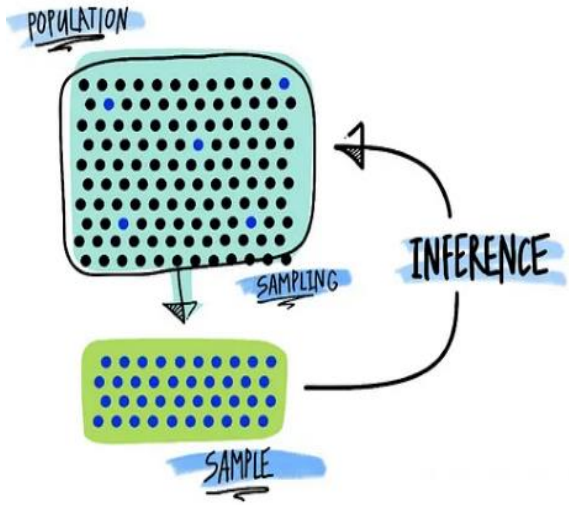
Common Statistical Methods with R/Rstudio

May 3, 2024

Wonsuk Yoo, PhD

Biostatistics Core
Barrow Neurological Institute
wonsuk.yoo@barrowneuro.org

Components of Data Analysis:



Data Investigation

Data Description

Data Inference

R: Set up “working directory”

- **`getwd()`**: Shows current directory
- **`setwd()`**: Create new working directory.
- Two types of directory expression:

```
> setwd("C:/training/biostat2024“)
> setwd("C:\\training\\biostat2024“)
```

- Or you can do from the top menu of RStudio:

```
Session – Set working directory – Choose Directory…
```

“dplyr”: data manipulation package

- **`dplyr`** is a R-package that help you manipulate the data for describing and analyzing the data.
- **`dplyr`** functions : The packages includes several key commands to allow data manipulation steps.
- The easiest way to get **`dplyr`** is to install the whole tidyverse package:

```
> install.packages("tidyverse")
```

- For package loading,

```
> library(dplyr)
```

▶ `library`: Loading of R-packages for use

- More information...

```
> vignette("dplyr")
```

“dplyr”: useful functions

- The pipe (`%>%`) operator is a very useful command to pipe the results from one step into the next step.
- You can use the pipe to rewrite multiple operations that you can read left-to-right, top-to-bottom (reading the pipe operator as “then”).
- Thus, “`x %>% f(y)`” turns into “`f(x,y)`”.

```
> 1 %>% exp()
```
▶ Same as “`exp(1)`”

```
> 1 %>% exp() %>% log()
```
▶ Same as “`log(exp(1))`”

- The group-by operation (**`group_by()`**) allows you to perform any operation “by group”.

“dplyr”: Selected commands related to rows

- Commands related to rows: `filter(), slice(), arrange()`

1. **`filter()`**: choose rows based on column values. It allows you to select a subset of rows in a data frame.

```
> dim(starwars)
```
▶ data (startwars): 87-by-14

```
> starwars2 <- starwars %>% filter(sex =='male', height > 170)
> dim(starwars2)
```
▶ data (starwars2): 45-by-14

2. **`summarise()`**: it collapses a group into a single rows.

```
> starwars %>% summarise(height = mean(height, na.rm = TRUE))
> starwars %>% group_by(sex) %>% summarise(height = mean(height, na.rm
  = TRUE))
```

“dplyr”: Selected commands related to columns

- Commands related to columns: `select(), rename(), mutate(), relocate()`

1. **`select()`**: decide whether or not a column is included. When a few of columns are only of interest to you, you can designate the columns using select operator.

```
> starwars3 <- starwars %>% select(hair_color, skin_color, eye_color)
> dim(starwars3)
```

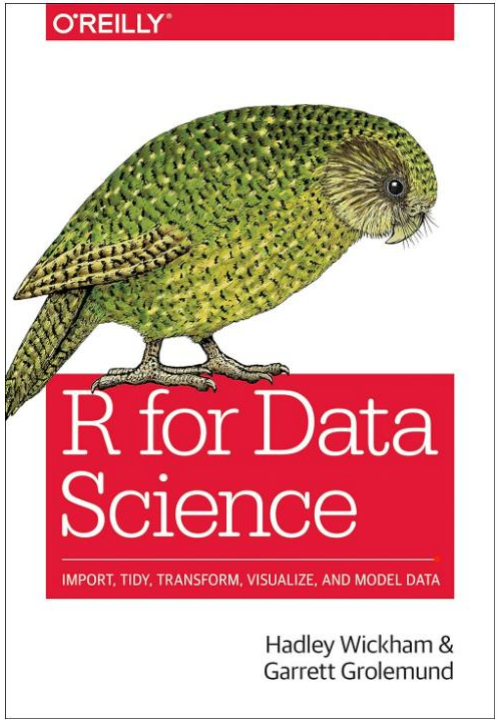
2. **`mutate()`**: When you want to add new columns that are functions of existing columns, this is the job of mutate()

```
> starwars %>%
+   mutate(height_m = height/100, BMI = mass / (height_m^2))
```

Reference for Data Science

- Tons of online information on google
- Books:

https://www.rstudio.com/resources/books/



Common Statistical Methods:

- [1] Statistical Methods in **Mean Difference** among Independent Groups ➡ Cross-sectional studies
- [2] Statistical Methods in **Mean Change** between two time points for Paired Groups ➡ Longitudinal studies
- [3] Statistical Methods in **Proportion Differences** from Frequency Table

Statistical Methods in Mean Difference among Independent Groups

- For these methods, dependent (or testing) variable should be in continuous scale and approximately normally distributed.
- These methods are based on the cross-sectional time points.
- Independent **t-tests** for comparing group means from two independent group.
- Analysis of Variance (**ANOVA**) for comparing group means from two or more than two independent groups
- Analysis of Covariance (**ANCOVA**) for comparing group means from two or more than two independent groups after adjusting for the covariates

Datasets: NHEFS

- We will be primarily relying on the data of NHANES Epidemiologic Follow-Up Study (NHEFS) to demonstrate the applications on common statistical methods.
- This is a subset data (n=350) of NHEFS data for illustrative purpose in this workshop. Thus, the analyses do not show scientific meaningful results.
- This subset data will be used to estimate the effect of quitting smoking on weight (1982) and weight change (1971 to 1982).
- Throughout this course, when using this dataset, our exposure of interest will be the indicator of whether the individual quit smoking (**qsmk**), our outcome will be **weight change** between 1971 (wt71) and 1982 (wt82), and our confounders of this relation will be all remaining variables.
- URL: https://wwwn.cdc.gov/nchs/nhanes/nhefs/#ndl

Data Loading:

- Import CSV data (nhefs.csv) using "`read_csv`" from "`tidyverse`" R-package.

```
> library(tidyverse)
> dat <- read_csv('data/nhefs.csv')
> dim(datdel)
[1] 350  13
> names(datdel)
 [1] "seqn"    "qsmk"    "sex"     "age"     "income"    "race"    "sbp"
 [8] "exercise" "wt71"   "wt82"    "wt82_71"  "wt_delta"  "wt_med"
> head(dat)
# A tibble: 6 × 13
   seqn  qsmk   sex   age income  race   sbp exercise  wt71  wt82 wt82_71
  <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl>   <dbl>
1  6960     0     0    38     11     1   120        2  52.6  56.7    4.08
2 20947     1     0    41     20     0   139        1  68.4  82.1   13.7
3 21369     0     1    25     18     0   118        1  48.4  51.7    3.29
4 21490     1     0    53     22     0   125        1  66.4  77.1   10.7
5 22946     1     1    35     16     0   133        1  71.8  76.2    4.42
6 24140     0     1    51     21     0   120        2  57.0  60.8    3.74
```

T-test:

- The t-test is one of the most popular statistical methods used to test whether mean difference between (independent) two groups is statistically significant.
- The null hypothesis is that both means are statistically equal, while alternative hypothesis is that both means are not statistically equal.
- When we do not know population variance, we apply t-test.
- One-sample t-test:
- Two sample independent t-test:

T-test: One-sample

From the data, BMI was given by mean of 24.5 and standard deviation of 2.19, whereas population mean was assumed to be 25.5. A researcher would like to know whether the sample data showed less BMI compared to that of national data.

- Hypothesis:
  Null: The sample group has the same mean BMI level as that of the national group.
  Alternative: Both groups do not have same mean BMI level.
- R-code:

```
> t.test(dat$BMI, mu=25.5)
```

- Result: p-value=0.04498

T-test: One-sample

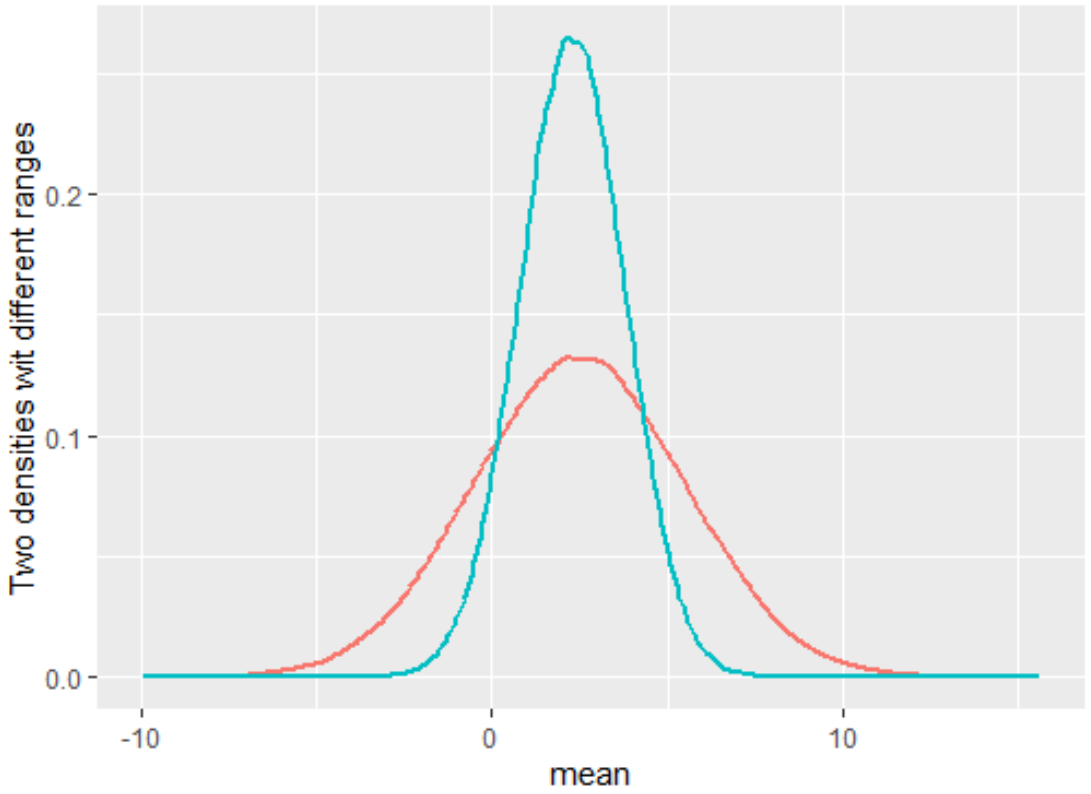
```
> t.test(dat$BMI, mu=25.5)

One Sample t-test
data:  dat$BMIt = -2.1462, df = 19, p-value = 0.04498
alternative hypothesis: true mean is not equal to 25.5
95 percent confidence interval:
  23.42604 25.47396
sample estimates:
mean of x
    24.45
```

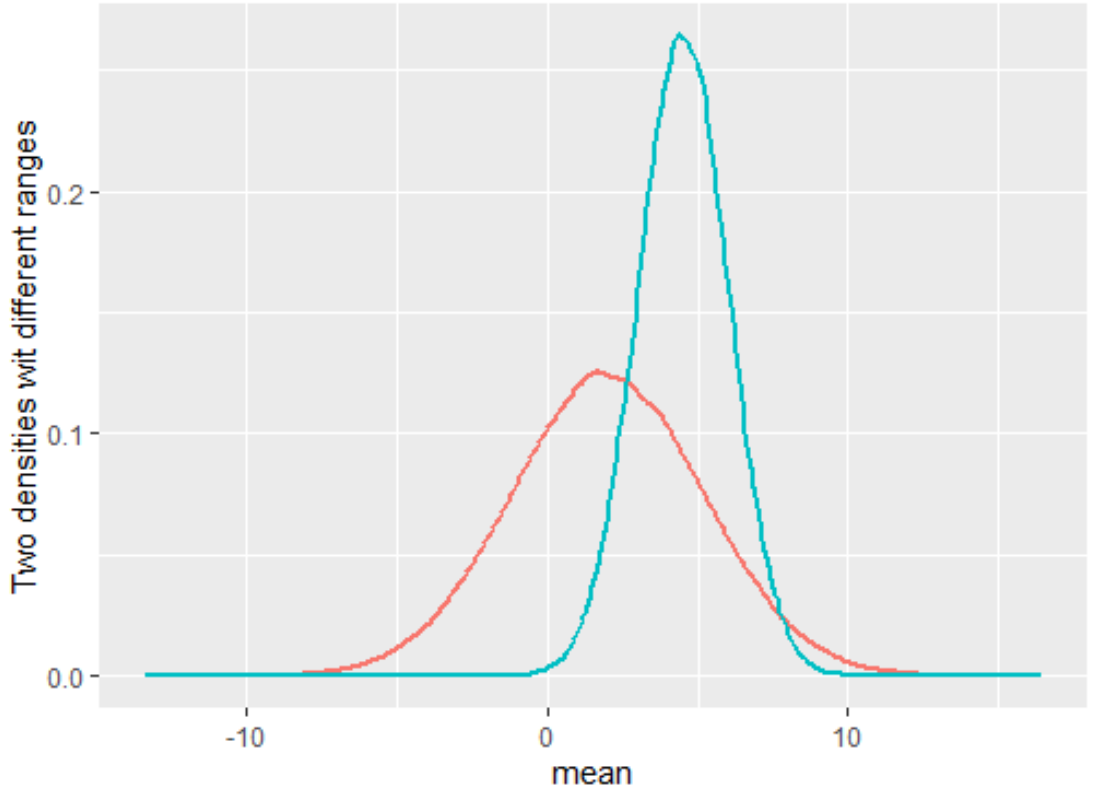
Two-sample independent t-test:

From the data, mean “wt82” of the “qsmk (=1)” and “wt82” of the “qsmk (=0)” were 79.7 (SD=17.2) and 74.3 (SD=16.0), respectively. We want to know whether the mean wt82 with quit-smoking (qsmk=1) group is significantly different from that of smoking (qsmk=0) group.

- Two sample t-test was used because we don’t know population variance.
- Set hypothesis:
  - Null: The mean weight on 1982 are the same for both groups on QSMK.
  - Alternative: Both groups do not have same mean weight (1982) levels.
- We should check the homogeneity of variance (HOV) assumption
- Interpretation: Compare the p-value with significance level (5%).

Two distributions on means and variances





Why Homogeneity of Variance be important?

- Homogeneity of variance (HOV) means that the variance of the dependent variable should be the same for all groups.
- This assumption is critical because if it is not met, the results may produce a biased estimate of the mean differences between groups, leading to incorrect conclusions.
- This bias can occur because the group with the largest variance will dominate the ANOVA results, leading to an overestimation of the group differences.
- The test statistics are generally robust to violations of HOV as long as group sizes are equal.
- HOV test is important to obtain non-biased type 1 error rate for maintaining the power of the test.

Overcoming Violation of HOV Assumption:

- There exist a couple of options to overcome when HOV assumption is violated.
- A parametric Welch's t-test/ANOVA does not assume homogeneity of variance and the test statistic is calculated with unequal variances.
  - The formula is a bit more complex than the traditional t-test formula.
  - Welch's t-test that is a more robust test that can be used in a wider range of situation.
- A non-parametric Wilcoxon test is a robust test that are less sensitive to violations of assumptions… Since the tests use the ranks rather than raw data.

Test of Homogeneity of Variance (HOV test)

- There are several methods for testing HOV including graphical methods and statistical tests.
- The most common graphical method is the **Q-Q plot**.
- Statistical tests include Levene's test and Bartlett's test.
  - **Levene's test** is the most widely used test and is recommended when the sample sizes are equal or nearly equal.
  - **Bartlett's test** is more appropriate when the sample sizes are unequal.
- Null hypothesis: No violation of HOV assumption (meet HOV assumption)
- D/M: Perform the standard t-test for high p-value. Or we need to perform other options to overcome the results from the violation of HOV

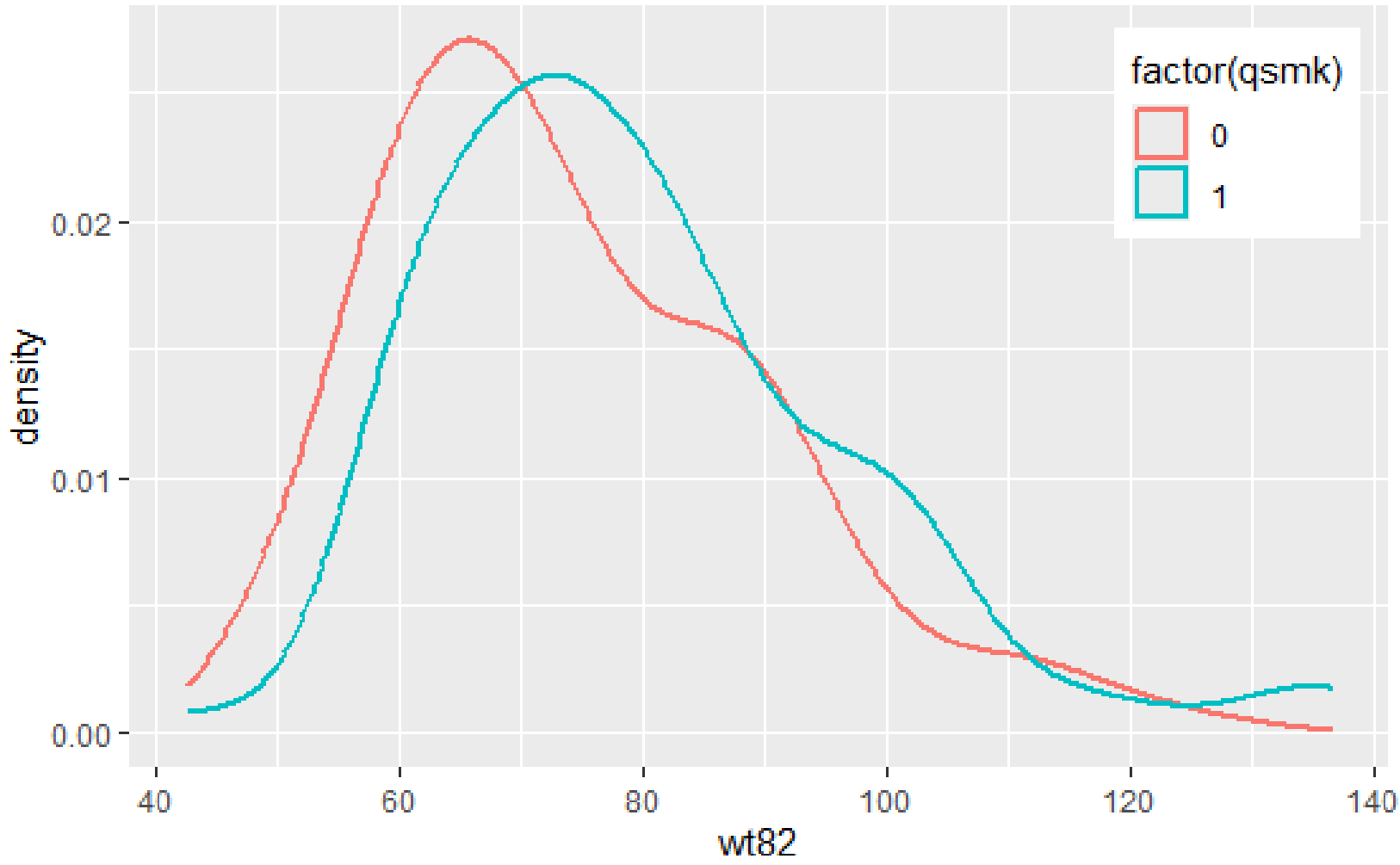
Two-sample independent t-test: HOV Test

```
> leveneTest(y=wt82, group=qsmk, data=datdel)

Levene's Test for Homogeneity of Variance (center = median: datdel)

      Df F value Pr(>F)
group  1  0.0619 0.8037
      348
```

```
> bartlett.test(wt82~qsmk, data = datdel)

Bartlett test of homogeneity of variances

data:  wt82 by qsmk
Bartlett's K-squared = 0.7237, df = 1,
p-value = 0.3949
```

➡ What's your conclusion?

Distributions of wt82 by qsmk:



Two-sample independent t-test:

- Aim: We want to know whether the mean wt82 with quit-smoking (qsmk=1) group is significantly different from that of smoking (qsmk=0) group.

```
> t.test(wt82 ~ qsmk, var.equal=TRUE, data=dat)

        Two Sample t-test

data:  wt82 by qsmk
t = -2.8323, df = 348, p-value = 0.00489
alternative hypothesis: true difference in means between group 0
and group 1 is not equal to 0
95 percent confidence interval:
 -9.140162 -1.648406
sample estimates:
mean in group 0 mean in group 1
       74.29248        79.68677
```

➡ What's your conclusion?

Analysis of Variance (ANOVA):

Let’s assume we are interested in the effect of mean weight in 1982 based on the levels of exercise (0: no exercise, 1: mild exercise, 2: strong exercise

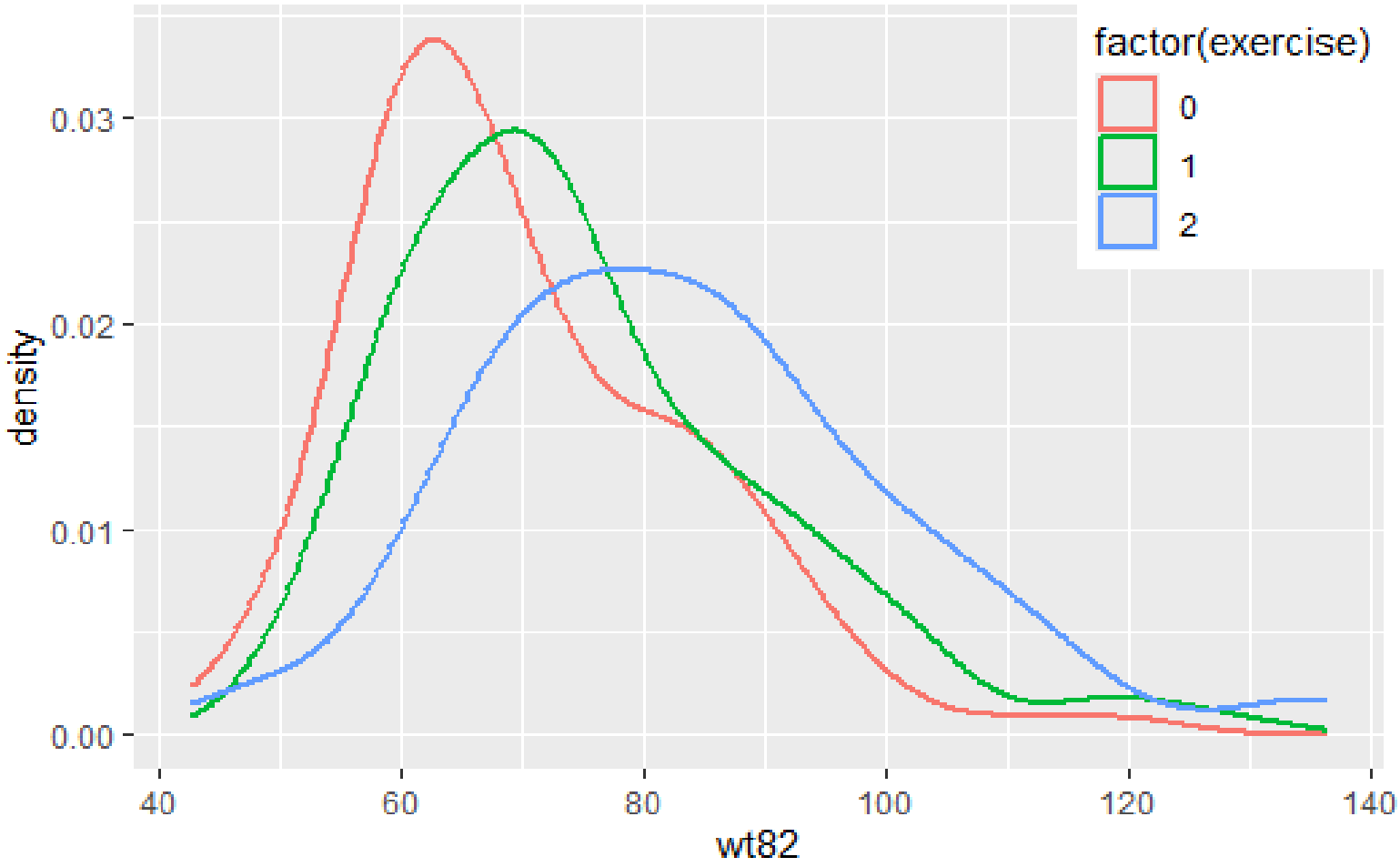
- Analysis purpose: To compare the means in wt82 among three age groups.
- Set **hypothesis**:
  - null: No difference of means in weight among three age groups.
  - alternative: At least one mean in weight is different from other groups.
- We need to test the homogeneity of variance (HOV) assumption.
- When the ANOVA test is performed, the two results occur:
  - ➡ For $p>0.05$, no means are different among the groups.
  - ➡ For $p<0.05$, at least one mean is different from those from other groups.

ANOVA: HOV Test

```
> library(car) #"car" package
> leveneTest(y=wt82, group=exercise, data=datdel)
Levene's Test for Homogeneity of Variance (center = median: dat)
       Df F value Pr(>F)
group   2  1.1501 0.3178
      347
```

- Conclusion?

Distributions of weight (1982) by Exercise



ANOVA: Hypothesis Test

```
> aov <- aov(wt82 ~ exercise, data = dat)
> summary(aov)
             Df Sum Sq Mean Sq F value   Pr(>F)
exercise      2  10569    5284   21.61 1.43e-09 ***
Residuals   347  84862     245
---
> summary.lm(aov)

Call:
aov(formula = wt82 ~ exercise, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   69.946      1.446  48.380  < 2e-16 ***
exercise1      4.893      2.024   2.418   0.0161 *
exercise2     13.478      2.072   6.505 2.72e-10 ***
```

- ➡ Conclusion
- ➡ What's next?

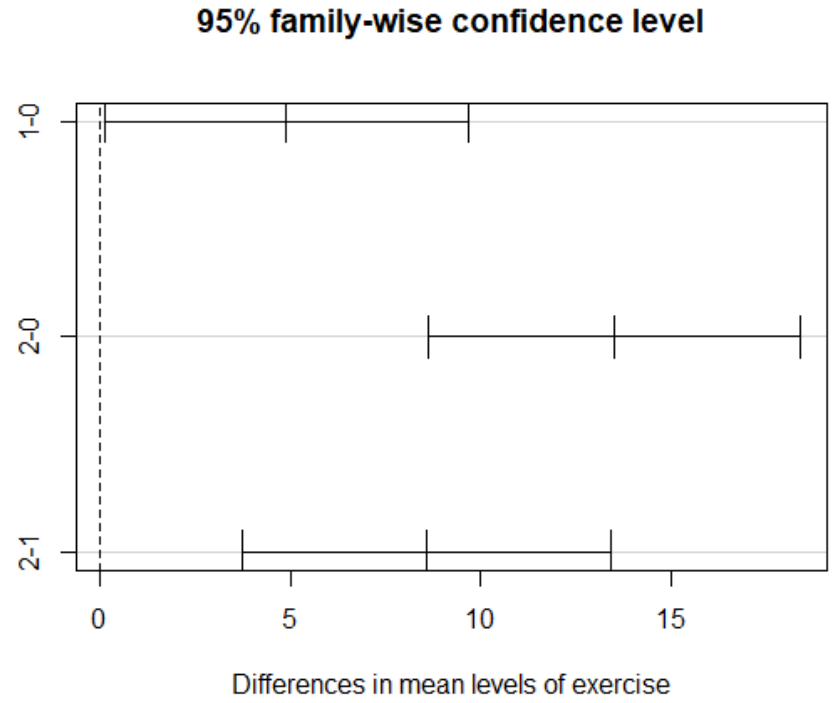
ANOVA: Multiple Comparisons (Post Hoc)

- Since p-value is less than 5%, we can conclude that at least one group mean on wt82 is different among the groups of exercise status.
- However, we do not know which pairwise groups are significant different.
- We need to explore which pairwise groups are statistically different

  ➡ This is why the multiple comparison procedure is called a post hoc analysis
- What is the **fundamental theory** on the study design with the multiple comparison?
- Post Hoc tests should control the type 1 error rate with the view of good study design.
- Several methods… Tukey, Bonferroni, etc

Post Hoc Test with ANOVA: Multiple Comparisons

```
> library(multcomp)
> TukeyHSD(aov_del)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = wt82 ~ exercise,
data = datdel)

$exercise
         diff       lwr       upr     p adj
1-0  4.892754 0.1296331  9.655874 0.0425182
2-0 13.477949 8.6006601 18.355237 0.0000000
2-1  8.585195 3.7568089 13.413582 0.0001069
```



Statistical Methods in Mean Change between two time points for Paired Groups

- For these methods, dependent (or testing) variable should be in continuous scale and approximately normally distributed.
- These methods are based on the repeated measures on multiple time points.
- Paired t-tests is to test whether there exist a meaningful change in paired data data between two time points (before and after, pre- and post-)
- ANCOVA approach is appropriate to compare the mean changes after adjusting for the baseline measurements.

One-sample Paired t-test:

From the data, means “wt82” and “wt71” were 75.9 (SD=16.5) and 69.0 (SD=14.4), respectively. We want to know whether the mean of weight in 1982 (wt82) is significantly different from that in 1971 (wt71).

- Set hypothesis:
  - Null: The mean weights between 1982 and 1971 are the same.
  - Alternative: The means in weight between 1982 and 1971 are different.
- Even though we have two timepoint data in 1982 and 1971, we handle one data (difference between two time points).
- Frequently used in pilot studies to estimate the effect size of the intervention or treatment effect.

One-sample Paired t-test:

```
> pairtd <- t.test(datdel$wt71, datdel$wt82, paired = TRUE)
> pairtd

        Paired t-test

data:  datdel$wt71 and datdel$wt82
t = -20.133, df = 349, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -7.596297 -6.244232
sample estimates:
mean difference
      -6.920265
```

- Conclusion?

Analysis of Variance (ANCOVA): Introduction

- Let's assume we are interested in whether the changes in mean weights between 1982 and 1971 are different among the quit-smoking status (qsmk)?
- Analysis of covariance (ANCOVA) can be a most appropriate method to compare the treatment effect with two time points.
- ANCOVA combines ANOVA method and Regression method.
  - ➡ ANOVA: to compare the exposure group.
  - ➡ Regression: Pre and Post data change.

ANCOVA Layout

GROUP 1	
X	y
X_{11}	y_{11}
X_{12}	y_{12}
X_{13}	y_{13}
...	...
X_{1n_1}	y_{1n_1}

GROUP 2	
x	y
x_{21}	y_{21}
x_{22}	y_{22}
x_{23}	y_{23}
...	...
x_{2n_2}	y_{2n_2}

+ COVs

$$Post_i = \beta_0 + \beta_1 Pre_i + \beta_2 Trt_i + COVs_i + \varepsilon_i$$

ANCOVA: “lm” vs “aov”

```
> fit1 <- aov(wt82 ~ wt71+qsmk+sex+exercise, data = datdel)
> summary(fit1)
             Df Sum Sq Mean Sq  F value   Pr(>F)
wt71          1  81273   81273 2856.047  < 2e-16 ***
qsmk          1    586     586   20.602 7.83e-06 ***
sex           1    125     125    4.383    0.037 *
exercise      2   3658    1829   64.280  < 2e-16 ***
Residuals   344   9789      28
---
```

```
> mod1 <- lm(wt82 ~ wt71+qsmk+sex+exercise, data = datdel)
> summary(mod1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6950     1.7572  -0.396   0.6927
wt71          1.0413     0.0223  46.704  < 2e-16 ***
qsmk1         1.4703     0.6418   2.291   0.0226 *
sex1          0.9791     0.6430   1.523   0.1287
exercise1     3.3767     0.6968   4.846 1.91e-06 ***
exercise2     8.2388     0.7298  11.288  < 2e-16 ***
---
Residual standard error: 5.334 on 344 degrees of freedom
Multiple R-squared:  0.8974,  Adjusted R-squared:  0.8959
F-statistic: 601.9 on 5 and 344 DF,  p-value: < 2.2e-16
```

Nonparametric Statistical Methods for t-tests and ANOVA

- For small sample size and/or far from normality assumption, we should use a non-parametric method or even a permutation-based test.
- Nonparametric methods use ranked data rather than actual data.
- Wilcoxon rank sum test for independent t-tests
- Wilcoxon signed rank test for paired t-test
- Kruskal-Wallis test for Oneway ANOVA

Wilcoxon rank sum test:

- Let’s assume our data showed small sample size and were violated from normality assumption. Wilcoxon rank sum test is a non-parametric version of independent t-tests,

```
> wilcx2a <- wilcox.test(datdel$wt71, datdel$wt82)
> wilcx2a

        Wilcoxon rank sum test with continuity
        correction

data:  datdel$wt71 and datdel$wt82
W = 45940, p-value = 1.046e-08
alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon signed rank test:

- Let’s assume our data showed small sample size and were violated from normality assumption. Wilcoxon rank sum test is a non-parametric version of paired t-tests,

```
> wilcx = wilcox.test(datdel$wt71, datdel$wt82, paired = TRUE,
exact=FALSE)
> print(wilcx)

        Wilcoxon signed rank test with continuity
        correction

data:  datdel$wt71 and datdel$wt82
V = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Statistical Methods in Proportion Differences from Frequency Table

- For these methods, dependent (or testing) variable and independent variable should be in discrete scale.
- The chi-square tests are based on approximate chi-square distribution.
- Fisher Exact test for exact calculation using hypergeometric distribution.
- Chi-square and Fisher’s test.. Does it really matter?
- Logistic regression analysis for binary outcome with covariate adjustments

Chi-square t-test:

- From the data, let's assume we want to examine the relationship between two discrete variable of "qsmk"(quit smoking status) and "exercise"(exercise levels 0,1,2).
- This is a problem of independency between two categorical variables. Thus,
  - Null: Two discrete variables are independent (no association).
  - Alternative: Two discrete variables are not independent.
- Pearson's chi-square tests are usually used to test the independency.
- Frequently used in pilot studies to estimate the effect size of the intervention or treatment effect.

Chi-square t-test:

```
> qsmkexer <- table(dat$exercise, dat$qsmk)
> qsmkexer

    0  1
  0 96 21
  1 84 38
  2 64 47
> chisq.test(qsmkexer)

        Pearson's Chi-squared test
data:  qsmkexer
X-squared = 16.119, df = 2, p-value = 0.000316
```

➡ Conclusion: Once the null hypothesis from the chi-square tests is rejected, we need to move to the measure the association between two categorical variable.

Chi-square t-test: Association

```
> library(grid)
> library(vcd)
> assocstats(qsmkexer)
                    X^2 df   P(> X^2)
Likelihood Ratio 16.550  2 0.00025481
Pearson          16.119  2 0.00031604

Phi-Coefficient   : NA
Contingency Coeff.: 0.21
Cramer's V        : 0.215
```

- Phi-coefficient
- Cramer’s coefficient
- Contingency coefficient

They ranges zero (independence) to one (completed dependence).

Fisher Exact Test:

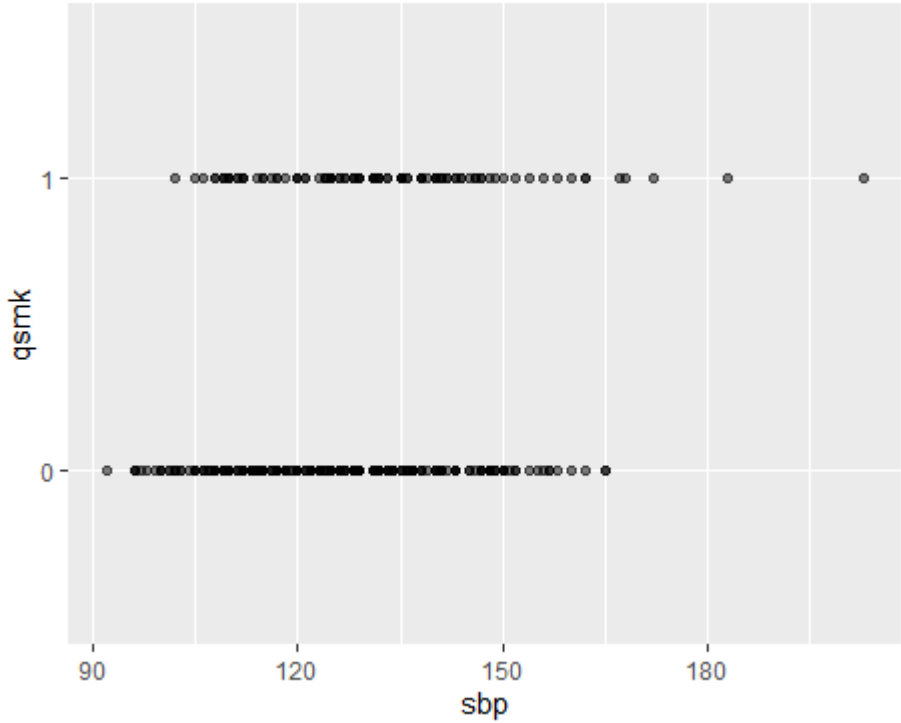
- It estimates the test statistic based on the hypergeometric distribution from a finite population size (exact number of a success or failure). The formula is based on the permutations to factorials without replacement.
- Example: if only 2 successes from 5 trial for a total of 15 population. P(X=2) = [5C2*10C3]/(15C5)=0.399.
- The odds mean a probability an event happens over a probability an event not happens. The odds of an event A: P(A)/(1-P(A)). From 2-by-2 table, OR=ad/bc.
- This is a problem of independency between two categorical variables. Thus,
  - Null: Two relative frequencies are independent (no association).
  - Alternative: Two relative frequencies are not independent.

Fisher Exact Test:

```
> fisher.test(qsmkexer)

        Fisher's Exact Test for Count Data

data:  qsmkexer
p-value = 0.0002706
alternative hypothesis: two.sided
```

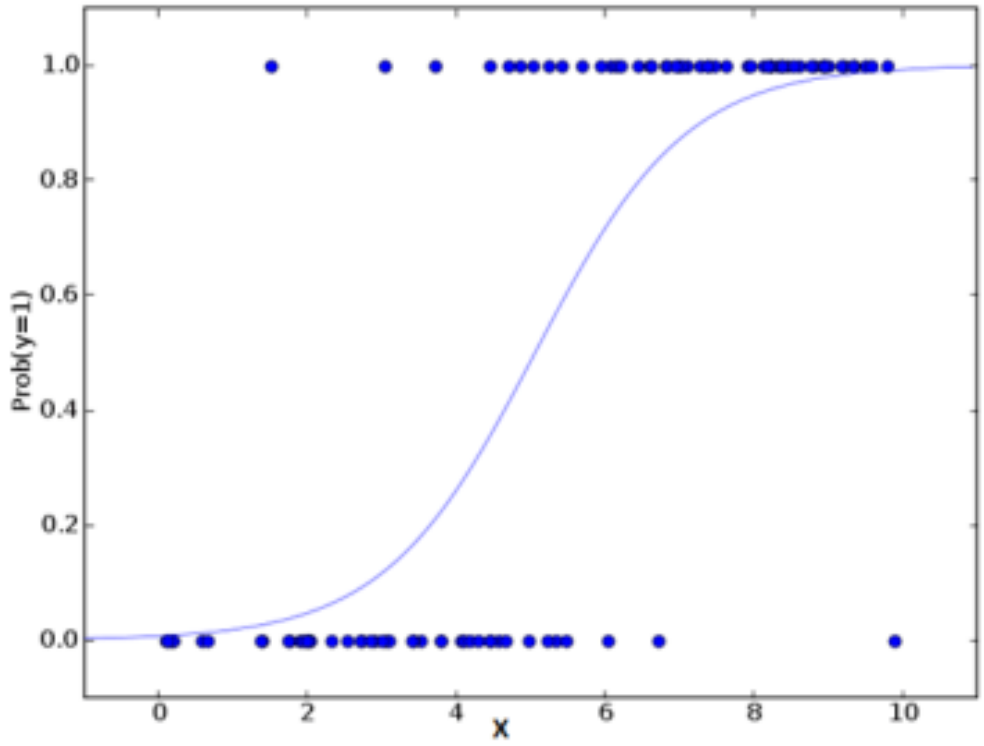
Logistic Regression Analyses:

- A logistic regression model is a type of regression model with binary outcome (disease status: yes/no).
- Let’s assume that we are interested in identifying the factors affecting the “qsmk” (quit smoking status). Thus, we have a binary variable as an outcome (qsmk) and a continuous predictor (sbp).
- Look at the scatter plot of sbp and qsmk. The plot looks far away from linear.



Logistic Regression Analyses (2):

- Since the outcome is dichotomous, the linear assumptions are not satisfied.
- For that, the logistic regression uses the logit (log-odds) function on the probability of a success, which shows a linear relation of logit function between 0 and 1.
- Thus, a logistic regression is the logit transformed linear regression model for binary outcomes.



Logistic Regression Analyses: Model

- The logistic regression is

$$logit(CHD) = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 AGE + \varepsilon$$

where p is the probability that CHD=1,

$$p = \frac{\exp(\beta_0 + \beta_1 AGE)}{1 + \exp(\beta_0 + \beta_1 AGE)}$$

- The model is fitted using "glm" function in R.

```
> glm(formula, family = binomial, data=dataname)
```

where formula: y ~ x1 + x2
family=binomial

Logistic Regression Analyses: Fit

```
> fit1 <- glm(qsmk ~ age+income+sbp+exercise, data=dat, family="binomial")
> summary(fit1)

Call:
glm(formula = qsmk ~ age + income + sbp + exercise, family = "binomial",
    data = dat)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.571206   1.630619  -5.256 1.47e-07 ***
age          0.026210   0.012630   2.075 0.037964 *
income       0.170324   0.057848   2.944 0.003237 **
sbp          0.021719   0.008103   2.680 0.007352 **
exercise1    0.769854   0.321235   2.397 0.016550 *
exercise2    1.220309   0.322093   3.789 0.000151 ***
```

Key Elements of Common Statistical Methods:

- The research type (analysis purpose): exploratory or confirmatory studies.
  - ➡ Exploratory studies without formal inferential tests
  - ➡ Confirmatory studies with formal hypothesis tests
- The study type: Randomized studies or Observational studies
  - ➡ Randomized studies like clinical trials or intervention trials
  - ➡ Observational studies needs to solve pre-existing condition in data
- The outcome data type: Continuous, categorical or Survival data
- The size of study data: more than 30 or less than 30 (small)?

The Types of Outcomes and Predictors:

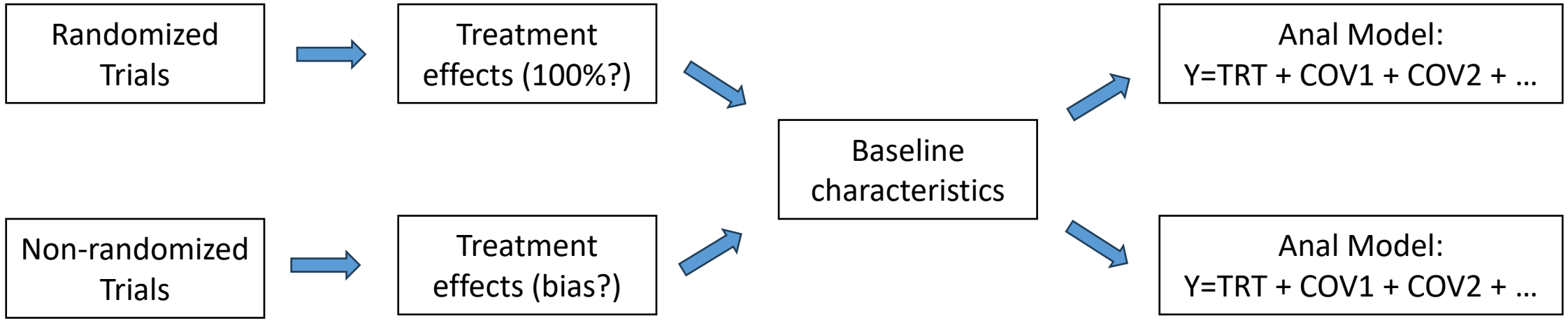
		OUTCOME		
		Categorical	**Continuous**	**Survival**
INPUT	**Categorical** (N=2)	Chi-square test Logistic regression	Student's t-test Wilcoxon rank sum test	Log-rank test Cox regression
	Categorical (N>2)	Chi-square test Logistic regression	ANOVA Kruskal-Wallis test	Log-rank test Cox regression
	Continuous	Logistic regression	Correlation analysis Linear regression	Cox regression

Evaluation link. . .

- Please go to the below survey link for your evaluation:

https://docs.google.com/forms/d/e/1FAIpQLScDfQQKoo3Vbl6VrRM71F75B4xvBVnrj-q4ULMVqdD9c9IDLg/viewform

Analysis Flow:



Final comments…

- **R/Rstudio …**
  - Tons of online materials on R/RStudio… Do coding for yourself.
- **Beyond Common Statistical Methods . . .**
  - Model building and Prediction Model (regression methodology)
  - Survival data analysis: Kaplan-Meier curve method, Cox regression
  - Repeated measures analysis
  - Longitudinal data analysis
  - Statistical diagnosis analysis
  - Bayesian data analysis

Barrow
Neurological Institute