R Series Workshop: Using R for Statistical Analysis

Dr. Charlie Keown-Stoneman

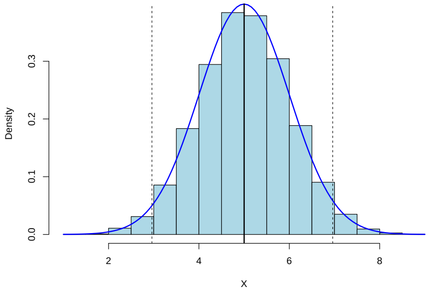
Applied Health Research Centre (AHRC), St. Michael's Hospital
Dalla Lana School of Public Health (DLSPH), University of Toronto

April 23, 2019

Some basics

Basic descriptive statistics, e.g.

- mean
- standard deviation (sd)
- quantiles
- median



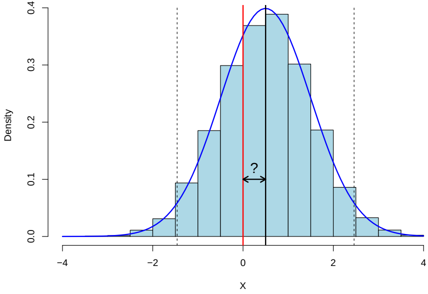
t-tests

- One of the most basic statistical tests is the *t-test*.
- Invented as a way of monitoring the quality of Guinness Stout.
- Various versions of the *t-test*:
  - Single sample t-test
  - Two-sample t-test
    - Independent samples
    - Paired samples
- Assumptions:
  - Independent observations (if paired, sets of pairs are independent)
  - Normal distribution in each group (least important)
  - If using pooled SD, true variance/SD in the populations are equal

Two-sided hypothesis test:

- Ho (null hypothesis) : The population mean is equal to μ_o
- Ha (alternative hypothesis) : The population mean is not equal to μ_o

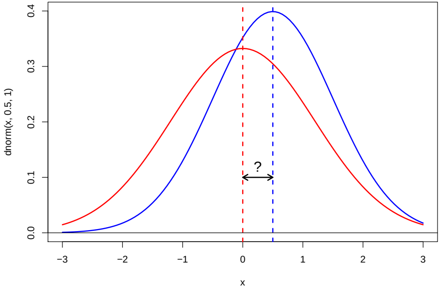
μ_0 is predetermined by researchers (e.g. $\mu_o = 0$)



Two sample t-test

Two-sided hypothesis test:

- Ho : No difference in the true population means of the two groups
- Ha : There is a difference between the true population means of the two groups



Pearson's χ^2 test

χ^2 (chi-square) tests can be used to test if two categorical variables are associated

- Based on the expected value in each cell of the contingency table if we assume independence
- Requires expected values of all cells be greater than 5
- Fairly straightforward in R

ANOVA

Analysis **o**f **Va**riance (ANOVA) is one method for testing the association between covariates and a continuous outcome.

- Order of covariates in the model is important
- Useful if more than 2 groups want to be compared
- Assumptions:
  - independence – observations are independent of other observations in the same group and observations in other groups
  - constant variance – all groups have the same variance
  - normality (least important assumption)

Simple Linear Regression (SLR)

$$E(y) = \beta_0 + \beta_1 x$$

- continuous, random Y variable, we cannot predict Y exactly due to measurement error and/or biological variability
- X can be either fixed (e.g. experiment) or random (e.g. observational study)
- in observational studies often an underlying causal hypothesis determines which variable is the Y variable
- single (often continuous) X variable
- linear relationship between X and the average of Y

SLR Assumptions

Assumptions for Simple Linear Regression:

- **Linear relationship holds on the average in population:** For every value of x there is a theoretical subpopulation of y's. The mean of this subpopulation falls on the straight line $\beta_0 + \beta_1 x$
- **Normality:** Each subpopulation has a normal distribution (least important assumption)
- **Constant Variance:** All subpopulations have the same variance/standard deviation
- **Independence:** Observations are independent

Multiple Linear Regression

If there is more than one x variable in a linear regression model, we have "multiple linear regression":

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

- Associations/effects are now adjusted for the other covariates in the model
- X variables can be continuous or binary; categorical variables can be made into binary "dummy" variables

Confidence Intervals

General Form of Wald-type confidence intervals (the most common type):

[Estimate] $\pm$ [Theoretical Test Statistic] $\times$ [Standard Error of Estimate]

For example, the 95% confidence interval for the mean is,

$$\bar{x} \pm t_{0.975,df} \times SE_{\bar{x}}$$

Logistic Regression

Logistic regression is a form of regression used when the outcome is binary (e.g. Yes/No, On/Off, Dead/Alive, etc)
In general:

$$\log\left(\frac{Prob(Y=1)}{1-Prob(Y=1)}\right) = \log(\text{Odds}) = \beta_0 + \beta_1 x$$

We could also write the logistic regression model as

$$\left(\frac{Prob(Y=1)}{1-Prob(Y=1)}\right) = \text{Odds} = e^{\beta_0+\beta_1 x}$$