BMR 617

Exploring data and relationships between variables:

Review and Summary

Types of data

- Categorical data
  - Nominal
  - Ordinal
- Quantitative data
  - Interval
  - Ratio

Roles of variable

- Response variable
  - The particular focus of a question in the study or experiment
- Explanatory variables
  - Variables that predict or explain changes in the response variable
- Role of variable defined by experimental design
- The type and role of variables determine much of the analysis

C → C

- Example: vaccine trial
  - Explanatory variable is the treatment (factor with levels "placebo" and "vaccine")
  - Response variable is whether or not the subject became infected (factor with levels "infected" and "not infected"
- When the explanatory and response variables are both categorical, typically present data using a contingency table
- We typically compute the risk for each group and the relative risk
  - Other measures include the attributable risk and number needed to treat

C → Q

- Examples include our mouse data
  - Response variables are quantitative: cholesterol level, fat mass, body weight, triglycerides level, glucose level
  - Explanatory variables are categorical: mouse strain (C57BL/6 or TALLYHO), and diet (Chow, High-fat, Low-fat high calorie)
- Present data using box plots, column scatter plots, or bar charts
- Compute mean or median, along with measures of spread, such as standard deviation, standard error of the mean, or interquartile range

Q → Q

- Examples include the insulin resistance data
- Present data using a scatter plot
- Compute correlation coefficient
  - Later we'll also talk about linear regression

Confidence intervals

- Framework:
  - Our data is a sample from a population
  - We compute statistics (such as a proportion, relative risk, mean, etc.) from the sample
  - These are estimates of the same statistic in the population
- Given a chosen level of confidence (e.g. 95%), we calculate a range of values which we are 95% confident will include the "true" value of the population statistic
- Computed confidence intervals for proportions and for means of quantitative variables

Distributions

- A distribution describes the probability a variable will take on a value, or a range of values
- The "Normal" or "Gaussian" distribution is a particular distribution with "nice" mathematical properties
  - Symmetric about the mean
  - Unimodal
  - Determined entirely by the mean and standard deviation
  - Software (or statistical tables) can be used to find the probability a normally-distributed value takes on any given range of values

The central limit theorem

- The central limit theorem says:
- Given any population distribution, if we take a sample of size n and take the mean of the sample, then:
  - The collection of all possible sample means is approximately normally distributed
    - The approximation improves as n gets larger
  - The mean of all sample means is the same as the mean of the population
  - The standard deviation of all sample means is the standard deviation of the population divided by $\sqrt{n}$
    - The last quantity is called the "standard error of the mean"

R

- Using libraries:
  - Install a package (once only per installation of R):

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

  - Load a library (once each session):

```
library(tidyverse)
```

Types of variable in R

- Can see the type using class(x) or looking in the "Environment" tab
- Categorical
  - character or factor
- Quantitative
  - numeric
- Special NA value represents missing data

Loading data

- Use the tidyverse function

`read_csv(…)`

- There are similar functions for loading other data formats (tab-delimited, etc.)
- Results in a "tibble" (data table)

Confidence intervals of proportions

- Use the binom library
- Choose from a variety of methods
  - Typically either "asymptotic" or "exact"
  - Avoid "asymptotic" if the proportion is close to 0 or 1

Confidence interval of a mean

- Computed this from a formula, using the qt(…) function
- If C is the desired confidence level, s is the sample standard deviation, and n is the sample size, the margin of error is

```
qt((1+C)/2, df=n-1)*s/sqrt(n)
```

- and then the confidence interval is the mean plus or minus the margin of error

Plots

- Use the ggplot2 library, which is part of tidyverse
- Start a plot with

```
ggplot(dataTable, aes(x=…, y=…))
```

- Then add layers using various geom_ functions
  - `geom_boxplot`
  - `geom_point`
  - `geom_bar`
- Configure axis with `xlab`, `ylab`
- Create title with `ggtitle`

Colors

- Change color of bar chart with `scale_fill_xxx` functions
- `scale_fill_manual` will let you specify colors manually:

```
ggplot(met_summary, aes(x=Diet, y=MeanCholesterol, fill=Strain)) +
geom_bar(stat="identity", position=position_dodge()) +
geom_errorbar(aes(ymin=MeanCholesterol-sem,
ymax=MeanCholesterol+sem),
position=position_dodge(0.9), width=0.2) +
scale_fill_manual(values=c("red","blue"))
```