Chapter 1

ST 512 - Review

Readings: Chapters 1-8 as needed

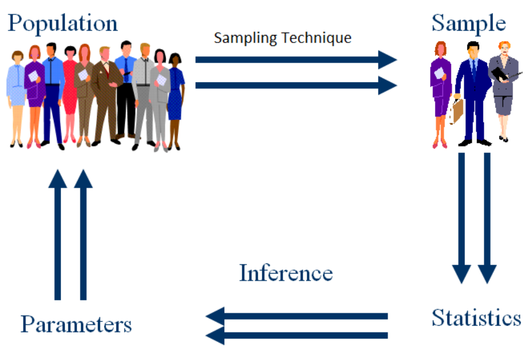
Big ideas in stats:

- Population - all the values, items, or individuals of interest

- Parameter - a (usually) unknown summary value about the population

- Sample - a subset of the population we observe data on

- Statistic - a summary value calculated from the sample observations



Examples of paramters - (true) mean μ, (true) variance σ^2.
Examples of statistics - sample mean $\bar{y}$, sample variance $s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$
Inference - Making mathematically sound claims about the poulation using sample data.

Scales (Types) of Data:

- Qualitative or Categorical - A variable that is described by attributes or labels
  Subscales:
  Nominal - categories have no ordering (Male, Female)
  Ordinal - can order categories (Lickert scale data)

- Quantitative - A variable that is described by numerical measurements where arithmetic can be performed
  Subscales:
  Discrete - finite or countable finite number of values (# of flowers on a plant, 0, 1, 2, ...)
  Continuous - any value in an interval is possibel (Temperature, $(-459.67 \deg F, \infty)$

Random Variables and Things of Interest:

- Random Variable (RV) - Function that takes in outcomes from an experiment and outputs real numbers, or a numeric outcome to a random process

  Things of interest

  - Distribution - pattern and frequency of observable values
    For continuous RVs, visualized with a smooth curve.
  - Mean/Median - measures of center of the distribution

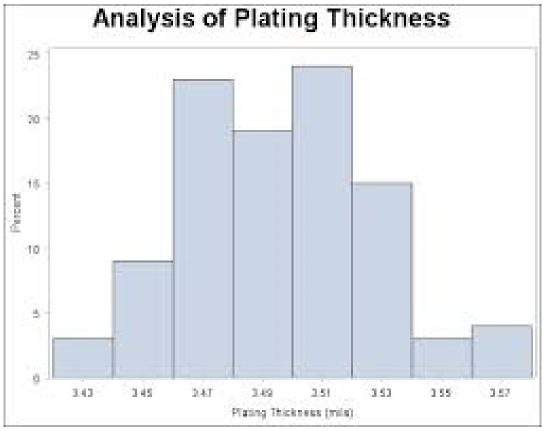
    Focus on mean: true mean μ, RV sample mean $\bar{Y}$, observed sample mean $\bar{y}$
  - Standard Deviation, Variance, IQR, Range - measures of spread for the distribution

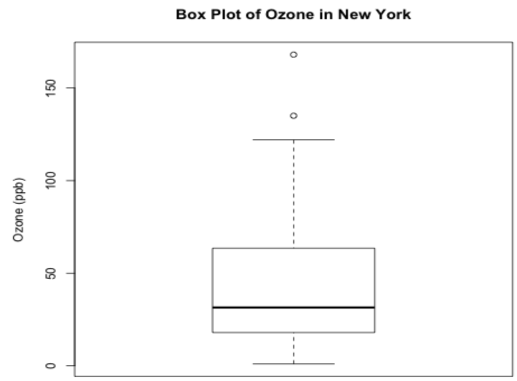
    Focus on SD and Variance: true variance σ^2, true SD σ, observed sample variance s^2, observed SD s

Graphical Descriptions of RV's:

- Histogram - Graphs the frequencies or relative frequencies of realizations of a RV

- Boxplot - Uses the Five Number Summary to display the realizations of a RV
  Five number summary: $min,\ Q_1,\ M,\ Q_3,\ max$





Statistics are also RVs. The distribution of a statistic is called a <u>sampling distribution</u>
Central Limit Theorem (CLT):

If a RV Y has a (true) mean μ and (true) variance σ^2, and a random sample is of size $n \geq 30$ is taken then

$$\bar{Y} \sim N(\mu, \sigma^2/n)$$

Note: If $Y \sim N(\mu, \sigma^2)$ then $\bar{Y} \sim N(\mu, \sigma^2/n)$ for any n.

2 main ways to make inference about a (true) mean, μ:

1. When the true SD, σ, is known we looked at the sampling distribution of the statistic

   $Z = \frac{\bar{Y}-\mu}{\sigma/n} \sim N(0,1)$ valid if $\bar{Y}$ has a normal distribution

   Allows us to form a CI: And a test statistic: Testing $H_0 : \mu = \mu_0$

   $\bar{y} \pm z_{\alpha/2}\sigma/\sqrt{n}$ $z_{obs} = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}$

2. When the true SD, σ, is unknown we looked at the sampling distribution of the statistic

   $T = \frac{\bar{Y}-\mu}{s/n} \sim t_{n-1}$ valid if $\bar{Y}$ has a normal distribution, allow for $n \geq 15$ or so in CLT

   Allows us to form a CI: And a test statistic: Testing $H_0 : \mu = \mu_0$

   $\bar{y} \pm t_{(n-1,\alpha/2)}s/\sqrt{n}$ $t_{obs} = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$

Inference about two (true) means, μ_1 and μ_2:

- From paired samples, $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ where difference is normally distributed

  CI: $(\bar{x} - \bar{y}) \pm t_{(n-1,\alpha/2)} s_{diff}/\sqrt{n}$

  HT: $H_0 : \mu_1 = \mu_2$, i.e. $\mu_1 - \mu_2 = 0$ $\quad t_{obs} = \frac{(\bar{x}-\bar{y})-0}{s_{diff}/\sqrt{n}}$

- Two separate samples from normal populations, $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$

  CI: $(\bar{x} - \bar{y}) \pm t_{(\nu,\alpha/2)} \sqrt{s_X^2/n + s_Y^2/m}$ where ν is an estimate of df

  HT: $H_0 : \mu_1 = \mu_2$, i.e. $\mu_1 - \mu_2 = 0$ $\quad t_{obs} = \frac{(\bar{x}-\bar{y})-0}{\sqrt{s_X^2/n+s_Y^2/m}}$

Extension to inference about t (true) means, $\mu_1, \mu_2, ..., \mu_t$:

Balanced One-way ANOVA table (same number of replicates per group)

Source	DF	SS	MS	F-stat	P-value
Treatment	$t-1$	$n\sum_{i=1}^{t}(\bar{Y}_{i+} - \bar{Y}_{++})^2$	$\frac{SS(Trt)}{t-1}$	$\frac{MS(Trt)}{MS(E)}$	Use $F(t-1, t(n-1))$
Error	$t(n-1)$	$\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i+})^2$	$\frac{SS(E)}{t(n-1)}$		
Total	$nt-1$	$\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{++})^2$			

Analysis used for a completely randomized design.

P-value tests $H_0 : \mu_1 = \mu_2 = ... = \mu_t$ vs H_A : at least 1 mean differs

One Way ANOVA model:

$$Y_{ij} = \mu + \alpha_i + E_{ij}$$

where $i = 1, 2, ...t$ and $j = 1, 2, ..., n$ (total sample size $= nt = N$)

μ = overall mean
α_i = effect from group i
E_{ij} = random error assumed to be *iid* $N(0, \sigma^2)$

For two quantitative variables measured on the same units, the linear relationship can be investigated:

Simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + E_i$ or use correlation.

For a hypothesis test, the p-value means

probability of observign a test statistic as extreme or more extreme than the one observed, assuming the null hypothesis is true.

For a given a null hypothesis, statistical significance implies

the observed value was unlikely to have occurred by random chance alone (assuming the null hypothesis is true).

For an observed confidence interval (cL, cU) we can say

We are ___% confident the true parameter value is contained in the interval. (***Do not say probability or chance!)

The idea of Confidence means

The procedure used to create the interval has a ___% probability of producing an interval that contains the parameter.

i.e. If the experiment were done repeatedly and an interval made for each sample, ___% of the intervals would contain the parameter value.