Lecture 2
Basic Concepts and Simple Comparative Experiments
Montgomery: Chapter 2
Random Variable and probability distribution

Discrete random variable $Y$:
- Finite possible values $\{y_1, y_2, y_3, \ldots, y_k\}$
- Probability mass function $\{p(y_1), p(y_2), \ldots p(y_k)\}$ satisfying

$$p(y_i) \geq 0 \text{ and } \sum_{i=1}^{k} p(y_i) = 1.$$ 

Continuous random variable $Y$:
- Possible values form an interval
- Probability density function $f(y)$ satisfying

$$f(y) \geq 0 \text{ and } \int f(y) dy = 1.$$
Mean and Variance

Mean $\mu = E(Y)$: center, location, etc.

Variance $\sigma^2 = \text{Var}(Y)$: spread, dispersion, etc.

Discrete $Y$:

$$\mu = \sum_{i=1}^{k} y_i p(y_i); \quad \sigma^2 = \sum_{i=1}^{k} (y_i - \mu)^2 p(y_i)$$

Continuous $Y$:

$$\mu = \int y f(y) dy; \quad \sigma^2 = \int (y - \mu)^2 f(y) dy$$

Formulas for calculating mean and variance

If $Y_1$ and $Y_2$ are independent, then

$$E(Y_1 Y_2) = E(Y_1) E(Y_2)$$

$$\text{Var}(aY_1 \pm bY_2) = a^2 \text{Var}(Y_1) + b^2 \text{Var}(Y_2)$$

Other formulas refer to Page 28 (Montgomery, 6th Edition)
Statistical Analysis and Inference:
Learn about population from (randomly) drawn data/sample

Model and parameter:
Assume population ($Y$) follows a certain model (distribution) that depends on a set of unknown constants (parameters): $Y \sim f(y, \theta)$.

Example 1: $Y \sim N(\mu, \sigma^2)$

\[
Y \sim \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}; \text{ where } \theta = (\mu, \sigma^2)
\]

Example 2: $Y_1$ and $Y_2$ are yields of tomato plants fed with fertilizer mixtures $A$ and $B$ respectively:

$Y_1 = \mu_1 + \epsilon_1; \ \epsilon_1 \sim N(0, \sigma_1^2)$

$Y_2 = \mu_2 + \epsilon_2; \ \epsilon_2 \sim N(0, \sigma_2^2)$

$\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$
Random sample or observations

Random Sample (conceptual)

\[ X_1, X_2, \ldots, X_n \sim f(x, \theta) \]

Random Sample (realized)

\[ x_1, x_2, \ldots, x_n \sim f(x, \theta) \]

Example 1:

\[ 0.0 \ 4.9 \ -0.5 \ -1.2 \ 2.1 \ 2.8 \ 1.2 \ 0.8 \ 0.9 \ -0.9 \]

Example 2:

A: 19.6 17.9 18.0 20.3 19.3 17.1 16.7 19.2 19.9 19.3

B: 19.6 19.9 21.8 18.4 19.4 21.4 20.5 20.0 18.2 19.9
Statistical Inference: Estimating Parameter $\theta$

- **Statistics**: a statistic is a function of the sample.
  
  $Y_1, \ldots, Y_n$: $\hat{\theta} = g(Y_1, Y_2, \ldots, Y_n)$ called estimator
  
  $y_1, \ldots, y_n$: $\hat{\theta} = g(y_1, y_2, \ldots, y_n)$ called estimate

- **Example 1**:

  Estimators for $\mu$ and $\sigma^2$

  $\hat{\mu} = \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$; $\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n - 1}$

  Estimates

  $\hat{\mu} = \bar{y} = 1.01; \hat{\sigma}^2 = s^2 = 3.49$
Example 2:

Estimators:

\[ \hat{\mu}_i = \bar{Y}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \]

\[ \hat{\sigma}_i^2 = S_i^2 = \frac{\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{n_i - 1} \]

for \( i = 1, 2 \).

Estimates:

\[ \bar{y}_1 = 18.73; s_1^2 = 1.50; \bar{y}_2 = 19.91; s_2^2 = 1.30; \]

Assume \( \sigma_1^2 = \sigma_2^2 \):

\[ S_{pool}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}; s_{pool}^2 = 1.40 \]
Statistical Inference: Testing Hypotheses

Use test statistics and their distributions to judge hypotheses regarding parameters.

- \( H_0: \text{null hypothesis vs } H_1: \text{alternative hypothesis} \)
  
  Example 1: \( H_0: \mu = 0 \) vs \( H_1: \mu \neq 0 \)
  
  Example 2.1: \( H_0: \mu_2 = \mu_1 \) vs \( H_1: \mu_2 > \mu_1 \)
  
  Example 2.2: \( H_0: \sigma_1^2 = \sigma_2^2 \) vs \( H_1: \sigma_1^2 \neq \sigma_2^2 \)

Details refer to Table 2-3 on Page 47 and Table 2-7 on Page 53

- **Test statistics:**
  
  Measures the amount of deviation of estimates from \( H_0 \)
  
  Example 1:

  \[
  T = \frac{\bar{Y} - 0}{S/\sqrt{n}} \sim^{H_0} t(n-1); \quad T_{obs} = 1.71
  \]
Example 2:

\[
T = \frac{(\bar{Y}_2 - \bar{Y}_1) - 0}{S_{pool} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim H_0 \quad t(n_1 + n_2 - 2); \quad T_{obs} = 2.22
\]

**Decision Rules**

- Given significance level \( \alpha \), there are two approaches:
  - Compare observed test statistic with critical value
  - Compute the \( P \)-value of observed test statistic
    * Reject \( H_0 \), if the \( P \)-value \( \leq \alpha \).
Statistical Inference: Testing Hypotheses

- $P$-value is the probability that test statistic takes on a value that is **at least as extreme as** the observed value of the statistic when $H_0$ is true.

  “Extreme” in the sense of the alternative hypothesis $H_1$.

Example 1:

$$P - \text{value} = P(T \leq -1.71 \text{ or } T \geq 1.71 \mid t(9)) = .12$$

Conclusion: fail to reject $H_0$ because $12\% \geq 5\%$.

Example 2:

$$P - \text{value} = P(T \geq 2.22 \mid t(18)) = 0.02$$

Conclusion: reject $H_0$ because $2\% \leq 5\%$. 
Type I Error, Type II Error and Power of a Decision Rule

Type I error: when $H_0$ is true, reject $H_0$.

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

Type II error: when $H_0$ is false, not reject $H_0$.

$$\beta = P(\text{type II error}) = P(\text{not reject } H_0 \mid H_0 \text{ is false})$$

Power

$$\text{Power} = 1 - \beta = P(\text{reject } H_0 \mid H_0 \text{ is false})$$

Details refer to Chapter 2, Stat511, etc.

In testing hypotheses, we usually control $\alpha$ (the significance level) and prefer decision rules with small $\beta$ (or high power). Requirements on $\beta$ (or power) are usually used to calculate necessary sample size.
Statistical Inference: Confidence Intervals:
Interval statements regarding parameter $\theta$

100(1-$\alpha$) percent confidence interval for $\theta$: $(L, U)$
Both $L$ and $U$ are statistics (calculated from a sample), such that

$$P(L < \theta < U) = 1 - \alpha$$

Given a real sample $x_1, x_2, \ldots, x_n$, $l = L(x_1, \ldots, x_n)$ and $u = U(x_1, \ldots, x_n)$ lead to a confidence interval $(l, u)$.

Question:

$$P(l < \theta < u) = ?$$
Example 1.
A 95% Confidence Interval for $\mu$:

$$(L, U) = (\bar{Y} - t_{0.025}(9) \frac{S}{\sqrt{n}}, \bar{Y} + t_{0.025}(9) \frac{S}{\sqrt{n}})$$

For the given sample;

$$(l, u) = (1.01 - 2.26 \times \frac{1.87}{\sqrt{10}}, 1.01 + 2.26 \times \frac{1.87}{\sqrt{10}}) = (-.33, 2.35)$$

Example 2.
A 95% Confidence interval for $\mu_2 - \mu_1$:

$$(L, U) = \bar{Y}_2 - \bar{Y}_1 \pm t_{0.025}(18) S_{pool} \sqrt{1/n_1 + 1/n_2}$$

$$(l, u) = (19.91 - 18.83) \pm 2.10 \times 1.18 \times \sqrt{1/10 + 1/10} = (.07, 2.29)$$

Connection between two-sided hypothesis testing and C.I.
If the C.I. contains zero, fail to reject $H_0$; otherwise, reject $H_0$. 
Sampling Distributions

Distributions of statistics used in estimation, testing and C.I. construction

Random sample: \( Y_1, Y_2, \ldots, Y_n \sim N(\mu, \sigma^2) \)

**Sample mean** \( \bar{Y} = (Y_1 + Y_2 + \cdots + Y_n)/n \)

\[
E(\bar{Y}) = E\left( \frac{1}{n} \sum Y_i \right) = \frac{1}{n} \sum E(Y_i) = \frac{1}{n} n \mu = \mu
\]

\[
\text{Var}(\bar{Y}) = \text{Var}\left( \frac{1}{n} \sum Y_i \right) = \frac{1}{n^2} \sum \text{Var}(Y_i) = \frac{1}{n^2} n \sigma^2 = \sigma^2/n
\]

\( \bar{Y} \) follows \( N(\mu, \sigma^2/n) \)
The Central Limit Theorem

$Y_1, Y_2, \ldots, Y_n$ are $n$ independent and identically distributed random variables with $E(Y_i) = \mu$ and $\text{Var}(Y_i) = \sigma^2$. Then

$$Z_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

has an approximate $N(0, 1)$ distribution.

Remark

1. Do not need to assume that the population distribution is normal
2. When the population distribution is normal, then $Z_n$ exactly follows $N(0, 1)$. 
Sampling Distributions: Sample Variance

\[ S^2 = \frac{(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2}{n - 1} \]

\[ E(S^2) = \sigma^2 \]

\[ \frac{(n - 1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{n-1} \]

Chi-squared distribution

If \( Z_1, Z_2, \ldots, Z_k \) are i.i.d as \( N(0, 1) \), then

\[ W = Z_1^2 + Z_2^2 + \cdots + Z_k^2 \]

follows a Chi-squared distribution with degree of freedom \( k \), denoted by \( \chi^2_k \)
Density functions of $\chi^2_k$
Sampling Distributions

- **t-distribution: \( t(k) \)**

  If \( Z \sim N(0, 1), W \sim \chi^2_k \) and \( Z \) and \( W \) independent, then

  \[
  T_k = \frac{Z}{\sqrt{W/k}}
  \]

  follows a \( t \)-distribution with d.f. \( k \), i.e., \( t(k) \).

**For example, in \( t \)-test:**

\[
T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} = \frac{\sqrt{n}(\bar{Y} - \mu_0)/\sigma}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{W/(n-1)}} \sim t(n-1)
\]

**Remark:**

As \( n \) goes to infinity, \( t(n - 1) \) converges to \( N(0, 1) \).
Density functions of $t(k)$ distributions

Figure 2-7 Several $t$ distributions.
Sampling Distribution

- **$F$-distributions: $F_{k_1,k_2}$**
  Suppose random variables $W_1 \sim \chi^2_{k_1}$, $W_2 \sim \chi^2_{k_2}$, and $W_1$ and $W_2$ are independent, then

  $$F = \frac{W_1/k_1}{W_2/k_2}$$

  follows $F_{k_1,k_2}$ with numerator d.f. $k_1$ and denominator d.f. $k_2$.

- **Example:** $H_0 : \sigma_1^2 = \sigma_2^2$, the test statistic is

  $$F = \frac{S_1^2}{S_2^2} = \frac{S_1^2/\sigma^2}{S_2^2/\sigma^2} = \frac{W_1/(n_1 - 1)}{W_2/(n_2 - 1)} \sim F_{n_1-1,n_2-1}$$

  Refer to Section 2.6 for details.
Density functions of $F$-distributions

Figure 2.8 Several $F$ distributions.
Normal Probability Plot

used to check if a sample is from a normal distribution

\[ Y_1, Y_2, \ldots, Y_n \] is a random sample from a population with mean \( \mu \) and variance \( \sigma^2 \).

Order Statistics: \( Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)} \) where \( Y^{(i)} \) is the \( i \)th smallest value.

if the population is normal, i.e., \( N(\mu, \sigma^2) \), then

\[ E(Y^{(i)}) \approx \mu + \sigma r_{\alpha_i} \text{ with } \alpha_i = \frac{i-3/8}{n+1/4} \]

where \( r_{\alpha_i} \) is the 100\( \alpha_i \)th percentile of \( N(0, 1) \) for \( 1 \leq i \leq n \).

Given a sample \( y_1, y_2, \ldots, y_n \), the plot of \( (r_{\alpha_i}, y^{(i)}) \) is called the normal probability plot or QQ plot.

the points falling around a straight line indicate normality of the population; Deviation from a straight line pattern indicates non-normality (the pen rule)
Example 1

\[ y(i) \begin{array}{cccccccccc}
-1.2 & -0.9 & -0.5 & 0.0 & 0.8 & 0.9 & 1.2 & 2.1 & 2.8 & 4.9 \\
\end{array} \]

\[ \alpha_i \begin{array}{cccccccccc}
.06 & .16 & .26 & .35 & .45 & .55 & .65 & .74 & .84 & .94 \\
\end{array} \]

\[ r_{\alpha_i} \begin{array}{cccccccccc}
-1.6 & -1.0 & -0.7 & -0.4 & -0.1 & 0.1 & 0.4 & 0.7 & 1.0 & 1.6 \\
\end{array} \]

Note: \( r_{\alpha_i} \) were obtained from the \( Z \)-chart (table)
QQ Plot 1
QQ Plot 1 (continued): True Population Distribution

Concave-upward shape indicates right-skewed distn
QQ plot 2.
QQ plot 2 (continued)

Concave-downward shape indicates left-skewed distn
QQ plot 3.

![QQ plot image]

- Quantiles of Standard Normal
- x-axis: Quantiles of Standard Normal
- y-axis: xt
flipped $S$ shape indicates a distribution with two heavier tails
SAS Code for QQ plot

data one;
input observation @@;
datalines;
0.89  2.79  2.27  2.58  1.72  2.93  -0.82  -1.40  0.08  1.97
0.84  -2.74  2.62  3.48  1.95  2.23  1.02  -0.76  0.20  -1.69
-1.69  0.89  1.98  1.61  0.22  2.60  -0.52  0.40  2.71  2.19;

proc univariate data=one;
var observation;
histogram observation / normal;
qqplot observation / normal (L=1 mu=est sigma=est);
run;
quit;
Output
Choice of Sample Size

- **Type II error**: \( \beta = P(\text{fail to reject } H_0 \mid H_1 \text{ is correct}) \)
  
  In testing hypotheses, one first wants to control type I error. If type II error is too large, the conclusion would be too conservative.

- **Example 2**  
  \( H_0 : \mu_2 - \mu_1 = 0 \) vs \( H_1 : \mu_2 - \mu_1 \neq 0 \)
  
  - Significance level: \( \alpha = 5\% \)
  
  - For convenience, we assume two samples have the same size \( n \)
  
  - Decision Rule based on two-sample \( t \)-test:
    
    reject \( H_0 \), if \( \frac{\bar{Y}_2 - \bar{Y}_1}{S_{pool} \sqrt{1/n + 1/n}} > t_{0.025}(2n-2) \) or \( < -t_{0.025}(2n-2) \)
    
    Equivalently
    
    fail to reject \( H_0 \) if \( -t_{0.025}(2n-2) \leq \frac{\bar{Y}_2 - \bar{Y}_1}{S_{pool} \sqrt{1/n + 1/n}} \leq t_{0.025}(2n-2) \)

  The type I error of the decision rule is 5%, we want to know how large \( n \) should be so that the decision rule has type II error less than a threshold, say, 5%.
Recall

\[ \beta = P(\text{type II}) = P(\text{accept } H_0 | H_1 \text{ holds}) \]

Hence

\[ \beta = P(-t_{0.025}(2n - 2) \leq \frac{\bar{Y}_2 - \bar{Y}_1}{S_{\text{pool}} \sqrt{1/n + 1/n}} \leq t_{0.025}(2n - 2) | H_1) \]

Under \( H_1 \), the test statistic does not follow \( t(2n - 2) \), in fact, it follows a noncentral \( t \)-distribution with df \( 2n - 2 \) and noncentral parameter

\[ \delta = \frac{|\mu_2 - \mu_1|}{\sigma \sqrt{2/n}} \].

Hence \( \beta \) is a function of \( |\mu_2 - \mu_1|/2\sigma \), and \( n \),

\[ \beta = \beta(|\mu_2 - \mu_1|/2\sigma, n) \]
Choice of Sample Size (continued)

- Let $d = \frac{|\mu_2 - \mu_1|}{2\sigma}$. So $\beta = \beta(d, n)$, which is the probability of type II error when $\mu_1$ and $\mu_2$ are apart by $d$. Intuitively, the smaller $d$ is, the larger $n$ needs to be such that $\beta \leq 5\%$.

- In terms of power $(1 - \beta(d, n))$. The smaller $d$ is, the larger $n$ needs to be in order to detect $\mu_1$ and $\mu_2$ are different from each other.

- Suppose we are interested in making the correct decision when $\mu_1$ and $\mu_2$ are apart by at least $d = 1$ with high probability (power), that is, we want to guarantee the type II error at $d = 1$, $\beta(1, n)$ to be small enough, say $< 5\%$. How many data points we need to collect?:

  **Find the smallest $n$ such that $\beta(1, n) < 5\%$**

- Calculate $\beta(d, n)$ for $d > 0$ and fixed $n$ and plot $\beta(d, n)$ against $d$, until the smallest $n$ is found.
• Case 1: n=4

operating characteristic curve

prob of accepting H₀
Case 2: n=7

operating characteristic curve
Case 3: n=9

operating characteristic curve
Operating characteristic Curves

- Curves of $\beta(d, n)$ versus $d$ for various given $n$ are called operating characteristic curves, O.C. Curves, which can be used to determine sample size

- O.C. Curves for two-sided $t$ test (next slide)

- $n = n_1 + n_2 - 1$. From the curves,

\[ n_1 + n_2 - 1 \approx 16 \]

If equal sample size is required, then $n_1 = n_2 \approx 9$. 
O.C. Curves for two-sided $t$ test

Figure 2-12  Operating characteristic curves for the two-sided $t$-test with $\alpha = 0.05$. (Reproduced with permission from “Operating Characteristics for the Common Statistical Tests of Significance,” C. L. Ferris, F. E. Grubbs, and C. L. Weaver, *Annals of Mathematical Statistics*, June 1946.)
SAS code for plotting O.C. Curves

```sas
data one;
  n=9; df=2*(n-1); alpha=0.05;
  do d=0 to 1 by 0.10;
    nc=d*sqrt(2*n);
    rlow=tinv(alpha/2,df); rhight=tinv(1-alpha/2,df);
    p=probt(rhigh,df,nc)-probt(rlow,df,nc);
    output;
  end;

proc print data=one;
symbol1 v=circle i=sm5;
title1 'operating characteristic curve';
axis1 label=('prob of accepting H_0'); axis2 label=('d');

proc gplot;
plot p*d/haxis=axis2 vaxis=axis1;
run;
quit;
```