What to do today (Feb 10, 2023)?

Part 1. Introduction and Review (Chp 1-5)

Part 2. Basic Statistical Inference (Chp 6-9)

§2.1 Point Estimation (Chp 6)

§2.2 Interval Estimation (Chp 7)

§2.3 One-Sample Tests of Hypotheses (Chp 8)

§2.3.1 Introduction and Basic Concepts

§2.3.2 Tests about Population Mean

§2.3.3 Large Sample Tests*

§2.3.4 Discussion

§2.4 Inference Based on Two-Samples (Chp 9)

§2.4.1 Population Means with Normal Populations

§2.4.2 Concerning Population Means Based on Large Sample

§2.4.3 Inferences on Two Population Variances

Part 3. Important Topics in Statistics (Chp 10-13)

Part 4. Further Topics (Selected from Chp 14-16)

Example 4.6 (cont'd)

$X \sim N(\mu, 1)$ with $X_1, \ldots, X_{100}$.
To test on $H_0 : \mu = 0$ vs $H_1 : \mu < 0$ with $\alpha = 0.05$ (rate of type I error or significance level), when $\bar{x} = -0.2$.

- Recall that $Z = (\bar{X} - 0)/\frac{1}{\sqrt{n}} \sim N(0, 1)$ under H_0.
- hypothesis testing:
  $\mathcal{R} = \{z : z < -z_{0.05}\} = \{z : z < -1.65\}$;
  $Z_{obs} = -0.2/(1/\sqrt{100}) = -2 \in \mathcal{R}$; reject H_0
- significance test:
  p-value$= P_{H_0}(Z < Z_{obs}) = 0.023 < 0.05$; the data show strong evidence against H_0.

§2.4. Two-Sample Tests of Hypotheses (Chp 9)

(2-Sample Problems)

Setup.

- *populations.* r.v. $X \sim F(\cdot;\theta)$ and r.v. $Y \sim F(\cdot;\phi)$
- *data.* a random sample from each population: $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$
- *goal.* to test on $H_0 : \theta - \phi = \Delta_0$ vs H_1 with desired α

§2.4.1 Population Means with Normal Populations

$$X \sim N(\mu_X, \sigma_X^2); \quad Y \sim N(\mu_Y, \sigma_Y^2)$$

(i) $H_0 : \mu_X - \mu_Y = \Delta_0$ vs (i) $H_1 : \mu_X - \mu_Y \neq \Delta_0$

(ii) $H_0 : \mu_X - \mu_Y = \Delta_0$ vs (ii) $H_1 : \mu_X - \mu_Y < \Delta_0$

(iii) $H_0 : \mu_X - \mu_Y = \Delta_0$ vs (iii) $H_1 : \mu_X - \mu_Y > \Delta_0$

§2.4.1A Two independent populations

If the variances are known ... ...

- *Test statistic.* Under H_0, consider $Z = \frac{(\bar{X}-\bar{Y})-\Delta_0}{\sqrt{\sigma_X^2/m+\sigma_Y^2/n}} \sim N(0,1)$
- *Rejection region.*

  (i) $H_1 : \mu_X - \mu_Y \neq \Delta_0$ to choose c such that
  $P_{H_0}(|Z| > c) = \alpha \Longrightarrow \mathcal{R} = \{z : |z| > z_{\alpha/2}\}$
  (ii) $H_1 : \mu_X - \mu_Y < \Delta_0$ to choose c such that $P_{H_0}(Z < c) = \alpha$
  $\Longrightarrow \mathcal{R} = \{z : z < -z_\alpha\}$
  (iii) $H_1 : \mu_X - \mu_Y > \Delta_0$ to choose c such that
  $P_{H_0}(Z > c) = \alpha \Longrightarrow \mathcal{R} = \{z : z > z_\alpha\}$
- *Making decision.*
  to obtain Z_{obs} and check if $Z_{obs} \in \mathcal{R}$:
  - reject H_0 if $Z_{obs} \in \mathcal{R}$
  - don't reject H_0 if $Z_{obs} \notin \mathcal{R}$

Example 5.1 (p364)

- **Study.** to compare yield strengths of cold-rolled steel and two-sided galvanized steel
- **Data.** 1st sample: $m = 20$ with $\bar{x} = 29.8$ksi; 2nd sample: $n = 25$ with $\bar{y} = 34.7$ksi.
- **Formulation.** $X \sim N(\mu_X, 4.0^2)$ and $Y \sim N(\mu_Y, 5.0^2)$; to test $H_0 : \mu_X = \mu_Y$ vs $H_1 : \mu_X \neq \mu_Y$ with $\alpha = .01$
- **Testing.** *Test statistic:* under H_0: $\Delta_0 = 0$,

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/20 + \sigma_Y^2/25}} \sim N(0,1)$$

  Rejection region: type (i) of H_1
  $c = z_{\alpha/2} = 2.58$; $\mathcal{R} = \{z : |z| > 2.58\}$
  Making decision:
  $Z_{obs} = -3.66 \in \mathcal{R} \Longrightarrow$ reject H_0.

§2.4.1A Two independent populations

If the variances are unknown and $\sigma_X^2 = \sigma_Y^2$ ... ...

- *Test statistic.* to consider
  $T = \frac{(\bar{X}-\bar{Y})-\Delta_0}{\sqrt{\hat{\sigma}_X^2/m+\hat{\sigma}_Y^2/n}} \sim t(m+n-2)$ under H_0 with
  $\hat{\sigma}_X^2 = \hat{\sigma}_Y^2 = s_{pooled}^2 = \frac{s_X^2(m-1)+s_Y^2(n-1)}{m+n-2}$.
- *Rejection region.*

  (i) $H_1 : \mu_X - \mu_Y \neq \Delta_0$ to choose c such that $P_{H_0}(|T| > c) = \alpha$
  $\Longrightarrow \mathcal{R} = \{t : |t| > t_{\alpha/2}(m+n-2)\}$

  (ii) $H_1 : \mu_X - \mu_Y < \Delta_0$ to choose c such that $P_{H_0}(T < c) = \alpha$
  $\Longrightarrow \mathcal{R} = \{t : t < -t_\alpha(m+n-2)\}$

  (iii) $H_1 : \mu_X - \mu_Y > \Delta_0$ to choose c such that $P_{H_0}(T > c) = \alpha$
  $\Longrightarrow \mathcal{R} = \{t : t > t_\alpha(m+n-2)\}$

- *Making decision.*
  to obtain T_{obs} and check if $T_{obs} \in \mathcal{R}$:
  - reject H_0 if $T_{obs} \in \mathcal{R}$
  - don't reject H_0 if $Z_{obs} \notin \mathcal{R}$

2.4.1A Two independent populations

If the variances are unknown ... ...

- *Test statistic.* with $\hat{\sigma}_X^2 = s_X^2$ and $\hat{\sigma}_Y^2 = s_Y^2$, to consider
  $T = \frac{(\bar{X}-\bar{Y})-\Delta_0}{\sqrt{\hat{\sigma}_X^2/m+\hat{\sigma}_Y^2/n}}$
  The distribution of T under H_0 is approximately $t(\nu)$:
  ν can be obtained using the formula in (9.2) of the textbook.
- *Rejection region.*

  (i) $H_1 : \mu_X - \mu_Y \neq \Delta_0$ to choose c such that $P_{H_0}(|T| > c) = \alpha$
  $\Longrightarrow \mathcal{R} = \{t : |t| > t_{\alpha/2}(\nu)\}$

  (ii) $H_1 : \mu_X - \mu_Y < \Delta_0$ to choose c such that $P_{H_0}(T < c) = \alpha$
  $\Longrightarrow \mathcal{R} = \{t : t < -t_{\alpha}(\nu)\}$

  (iii) $H_1 : \mu_X - \mu_Y > \Delta_0$ to choose c such that $P_{H_0}(T > c) = \alpha$
  $\Longrightarrow \mathcal{R} = \{t : t > t_{\alpha}(\nu)\}$

- *Making decision.*
  to obtain T_{obs} and check if $T_{obs} \in \mathcal{R}$:
  - reject H_0 if $T_{obs} \in \mathcal{R}$
  - don't reject H_0 if $Z_{obs} \notin \mathcal{R}$

What if X and Y are not independent?

$2.4.1B When data are paired

Data. $m = n$, $(X_1, Y_1), \ldots, (X_n, Y_n)$

Reformulating. $D = X - Y \sim (\mu_X - \mu_Y, \sigma_D^2)$;
$D_i = X_i - Y_i$, $i = 1, \ldots, n$; $H_0 : \mu_D = \Delta_0$

$\Longrightarrow$ one-sample problem on population mean with normal population: *known* σ_D^2; *unknown* σ_D^2

Remarks:

- the type of data are common
- no need to assume X and Y are independent
- no need to specify the dependence of X and Y

Example 5.2

- **Study.** to compare slide retrieval time and gigital retrieval time
- **Data.** in pair $m = n = 13$, $\bar{d} = 20.5$ and $s_D = 11.96$
- **Formulation.** $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$; to test $H_0 : \mu_X = \mu_Y$ vs $H_1 : \mu_X \neq \mu_Y$ with $\alpha = .05$
- **Testing.**
  Test statistic: $\Delta_0 = 0$, under H_0

$$T = \frac{\bar{D}}{\sqrt{\sigma_D^2/13}} \sim t(13-1)$$

Rejection region: type (i) of H_1
$c = t_{\alpha/2}(12) = 2.18$; $\mathcal{R} = \{t : |t| > 2.18\}$
Making decision:
$T_{obs} = 6.18 \in \mathcal{R} \Longrightarrow$ reject H_0.

§2.4.2 Concerning Population Means Based on Large Sample

Setup.

- *Formulation:* $X \sim F(\cdot)$ with population mean μ_X and $Y \sim G(\cdot)$ with population mean μ_Y
- *Data:* Available a random sample from each of the two populations: $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ with $m \gg 1$ and $n \gg 1$.
- *Hypotheses:* $H_0 : \mu_X - \mu_Y = \Delta_0$ vs $H_1 : \mu_X - \mu_Y \neq \Delta_0$ (or $\mu_X - \mu_Y < \Delta_0$ or $\mu_X - \mu_Y > \Delta_0$)

§2.4.2 Concerning Population Means Based on Large Sample

§2.4.2A With independent populations ($X \perp Y$)

Test statistic.

$$Z = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{S_X^2/m + S_Y^2/n}} \sim N(0, 1) \text{ approximately under } H_0.$$

§2.4.2B With paired data (not necessarily $X \perp Y$)

Re-formulation. $n = m$; $D_i = X_i - Y_i$ for $i = 1, \ldots, n$ iid from population $D = X - Y$ with population mean $\mu_D = \mu_X - \mu_Y$.

Test statistic.

$$Z = \frac{\bar{D} - \Delta_0}{\sqrt{S_D^2/n}} \sim N(0, 1)$$

approximately under H_0.

HWQ: what if data aren't paired and $X \not\perp Y$?

Example 5.3

- **Study.** to find out whether the proportion of all defendants who plead guilty and are sent to prison differs from the proportion who are sent to prison after pleading innocent and being found guilty?
- **Data.**

	Plea Guilty	Plea Innocent
Judged Guilty	m=191	n=64
Sentenced to Prison	101	56

- **Formulation.** to test $H_0 : p_X = p_Y$ vs $H_1 : p_X \neq p_Y$ at $\alpha = 0.01$
  - Initially pleaing for guilty: $X \sim B(1, p_X)$ with $X = 1$ or 0 for being sentenced or not
  - Initially pleaing for innocent: $Y \sim B(1, p_Y)$ with $Y = 1$ or 0 for being sentenced or not

What will we study next?

Part 1. Introduction and Review (Chp 1-5)

Part 2. Basic Statistical Inference (Chp 6-9)

§2.1. Point Estimation (Chp 6)

§2.2. Interval Estimation (Chp 7)

§2.3. One-Sample Tests of Hypotheses (Chp 8)

§2.4. Two-Sample Tests of Hypotheses (Chp 9)

- *2.4.1 Population Means with Normal Populations*
- *2.4.2 Population Means Based on Large Sample*
- **2.4.3 Inferences on Two Population Variances**

Part 3. Important Topics in Statistics (Chp 10-13)

Part 4. Further Topics (Selected from Chp 14-16)