What to do today (Feb 14, 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)

§2.4 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)

Some Logistics.

- Homework 5 has been assigned. It's due on Monday next week, the reading week.

§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 $Z_{obs} \notin \mathcal{R}$

§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:
  $T_{obs} = 6.18 \in \mathcal{R} \implies$ 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.

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

- **Testing.**
  - *test statistic:* taking 191 and 64 $\gg 1$,

$$Z = \frac{(\hat{p}_X - \hat{p}_Y) - 0}{\sqrt{\hat{p}(1-\hat{p})\big(1/m + 1/n\big)}} \sim N(0,1)$$

    approximately under H_0. Under $H_0 : p_X = p_Y$, which is estimated by $\hat{p} = (101+56)/(m+n)$.
  - *rejection region:* $\mathcal{R} = \{z : |z| > 2.58\}$
  - *making decision:* $Z_{obs} = -4.94 \in \mathcal{R}$, reject H_0.
- **Alternative 1.** p-value
  $= P_{H_0}(|Z| > |Z_{obs}|) = 2(1 - \Phi(4.94)) = 0.0004$
  $\Longrightarrow$ the data indicate strong evidence against H_0.
- **Alternative 2.** Approximate 99% CI of $p_X - p_Y$ is

$$(\hat{p}_X - \hat{p}_Y) \pm (2.58)\sqrt{\hat{p}_X(1-\hat{p}_X)/m + \hat{p}_Y(1-\hat{p}_Y)/n}$$

  $\Longrightarrow (-0.488, -0.205) \not\ni 0.$

§2.4.3 Concerning Population Variances with Normal Populations

Setup.

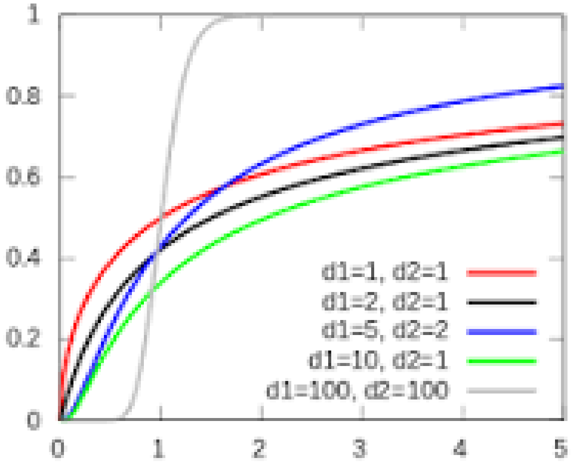
- *Formulation:* $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ and $X \perp Y$
- *Data:* Available a random sample from each of the two populations: $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$
- *Hypotheses:* $H_0 : \sigma_X^2 = \sigma_Y^2$ vs H_1 : otherwise.

thinking ...

- a hunch: to compare S_X^2 and S_Y^2?
- how to realize it?
  - to examine $S_X^2 - S_Y^2$ or S_X^2 / S_Y^2?
  - the two quantities' distributions?

§2.4.3A Preparation: F-distribution

- **Definition.** If $W_1 \sim \chi^2(\nu_1)$, $W_2 \sim \chi^2(\nu_2)$ and $W_1 \perp W_2$, the distribution of $F = \frac{W_1/\nu_1}{W_2/\nu_2}$ is called F-distribution with dfs ν_1 and ν_2, denoted by $F \sim F(\nu_1, \nu_2)$.



§2.4.3A Preparation: F-distribution

- **Definition.** If $W_1 \sim \chi^2(\nu_1)$, $W_2 \sim \chi^2(\nu_2)$ and $W_1 \perp W_2$, the distribution of $F = \frac{W_1/\nu_1}{W_2/\nu_2}$ is called F-distribution with dfs ν_1 and ν_2, denoted by $F \sim F(\nu_1, \nu_2)$.

Properties.

(i) If $\nu_2 \to \infty$, $\nu_1 F \sim \chi^2(\nu_1)$ approximately

(ii) If $T \sim t(\nu)$, $T^2 \sim F(1, \nu)$

(iii) Closely related to (a) Hotelling's T-distn, (b) Beta-distn.

(iv) If $F \sim F(\nu_1, \nu_2)$, $F^{-1} \sim F(\nu_2, \nu_1)$
Thus, if $P(F > f_\alpha(\nu_1, \nu_2)) = \alpha$, the critical value $f_\alpha(\nu_1, \nu_2)$ is $1/f_{1-\alpha}(\nu_2, \nu_1)$.

Proposition. $X \sim N(\mu_X, \sigma_X^2)$ with a random sample $X_1, \ldots, X_m$, and $Y \sim N(\mu_Y, \sigma_Y^2)$ with a random sample $Y_1, \ldots, Y_n$. If $X \perp Y$, then

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(m-1, n-1)$$

§2.4.3B Hypothesis testing

Setup.

- *Formulation:* $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ and $X \perp Y$
- *Data:* Available a random sample from each of the two populations: $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$
- *Hypotheses:* $H_0 : \sigma_X^2 = \sigma_Y^2$ vs H_1 : otherwise. (or $H_1 : \sigma_X^2 < \sigma_Y^2$ or $H_1 : \sigma_X^2 > \sigma_Y^2$)

Test Statistic. Under H_0,

$$F = \frac{S_X^2}{S_Y^2} \sim F(m-1, n-1)$$

Rejection Region. desired level of α

(i) $H_1 : \sigma_X^2 \neq \sigma_Y^2$
$\mathcal{R} = \{f : f < c_1 \text{ or } f > c_2\}$ with $P_{H_0}(F < c_1) = \alpha/2$ and $P_{H_0}(F > c_2) = \alpha/2$.
$c_2 = f_{\alpha/2}(m-1, n-1)$;
$c_1 = f_{1-\alpha/2}(m-1, n-1) = 1/f_{\alpha/2}(n-1, m-1)$

§2.4.3B Hypothesis testing

Rejection Region. desired level of α

- (i) $H_1 : \sigma_X^2 \neq \sigma_Y^2$
  $\mathcal{R} = \{f : f < c_1 \text{ or } f > c_2\}$ with $P_{H_0}(F < c_1) = \alpha/2$ and $P_{H_0}(F > c_2) = \alpha/2$.
  $c_2 = f_{\alpha/2}(m-1, n-1)$;
  $c_1 = f_{1-\alpha/2}(m-1, n-1) = 1/f_{\alpha/2}(n-1, m-1)$

- (ii) $H_1 : \sigma_X^2 < \sigma_Y^2$
  $\mathcal{R} = \{f : f < c\}$ with $P_{H_0}(F < c) = \alpha$.
  $c = f_{1-\alpha}(m-1, n-1) = 1/f_{\alpha}(n-1, m-1)$

- (iii) $H_1 : \sigma_X^2 > \sigma_Y^2$
  $\mathcal{R} = \{f : f > c\}$ with $P_{H_0}(F > c) = \alpha$.
  $c = f_{\alpha}(m-1, n-1)$

Making Decision. If $F_{obs} \in \mathcal{R}$, reject H_0; otherwise, don't reject H_0.

Two alternative approaches: significance test? by CI?

Example 5.4

- **Study.** to compare elderly men and young men in a Serum ferritin.
- **Data.** $m = 28$ obs from elderly with $s_x = 52.6$; $n = 26$ obs from young with $s_y = 84.2$
- **Formulation.** an elderly man's $X \sim N(\mu_X, \sigma_X^2)$ and a young man's $Y \sim N(\mu_Y, \sigma_Y^2)$; suppose $X \perp Y$; to test $H_0 : \sigma_X^2 = \sigma_Y^2$ vs $H_1 : \sigma_X^2 < \sigma_Y^2$ with $\alpha = .01$
- **Testing.**

Test statistic:

$$F = \frac{S_X^2}{S_Y^2} \sim F(m-1, n-1)$$

under H_0.
Rejection region: type (ii) of H_1
$c = f_{1-\alpha}(27, 25) = 1/f_\alpha(25, 27) = .394$; $\mathcal{R} = \{f : f < .394\}$
Making decision:
$F_{obs} = .390 \in \mathcal{R} \Longrightarrow$ reject H_0.

Suggested approach: Since H_1 is one-sided, to consider one-sided CI, i.e. upper bound in the example:

a 99% upper bound of σ_X^2/σ_Y^2 is $\frac{S_X^2}{S_Y^2}/c$: $c = f_{1-\alpha}(m-1, n-1))$
It's $0.989 < 1$. $\Longrightarrow$ reject H_0.

Discussion about the discrepancy

– significance level (type I error)

– test power/efficiency (type II error)

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)

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

§3.1A One-Factor Analysis of Variance (Chp 10)

§3.1B Multi-Factor ANOVA (Chp 11)

§3.2A Simple Linear Regression Analysis (Chp 12)

§3.2B More on Regression (Chp 13)

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