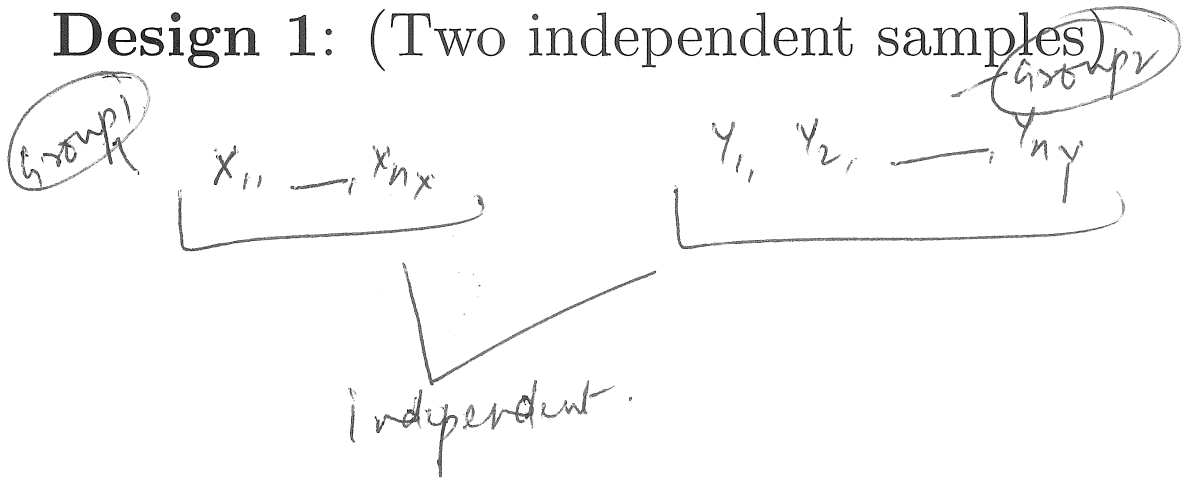
# Independent or paired samples?

**Design 1**: (Two independent samples)




Given a choice, if feasible, a paired design is better than the two-sample (indep) design. This is because in case of the former, any diff. in X and Y is almost due to the effect of the intervention in study. In case of the latter, the effect of the intervention gets confounded with the difference in the groups.

**Design 2**: (Paired samples)

| ID | $(X, Y)$ | $D = X - Y$ |
|---|---|---|
| 1 | $(X_1, Y_1)$ | $D_1$ |
| 2 | $(X_2, Y_2)$ | $D_2$ |
| ⋮ | | ⋮ |
| n | $(X_n, Y_n)$ | $D_n$ |

Generally: Assume different subjects are indep.

# CI for $\mu_X - \mu_Y$ with paired design

- Data: $(X_i, Y_i)$, $i = 1, 2, \ldots, n$ ($i$ → subject-index)
- Parameter of interest: $\mu_X - \mu_Y = E[D] = \mu_D$.
- Define: $D = X - Y$, $D_i = X_i - Y_i$, $i = 1, \ldots, n$
- $\Rightarrow E[D] = E[X] - E[Y] = \mu_X - \mu_Y$
- Apply one-sample procedure to the differences.

**$100(1-\alpha)\%$ CI for $\mu_D$ assuming $D_1, \ldots, D_n \sim N(\mu_D, \sigma_D^2)$:**

Pivot: $\frac{\bar{D} - \mu_D}{\sigma_D/\sqrt{n}} \sim N(0,1)$

$\bar{D} \pm z_{\alpha/2} \frac{\sigma_D}{\sqrt{n}}$, $\sigma_D$ is known

$\frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$

$\bar{D} \pm t_{n-1,\alpha/2} \frac{S_D}{\sqrt{n}}$ // $\sigma_D$ is unknown,

$S_D^2$ = sample variance of the differences

Approximate $100(1-\alpha)\%$ CI for $\mu_D$ if $n$ is large:

$$\bar{D} \pm z_{\alpha/2} \frac{\sigma_D}{\sqrt{n}} \approx \bar{D} \pm z_{\alpha/2} \frac{s_D}{\sqrt{n}}$$

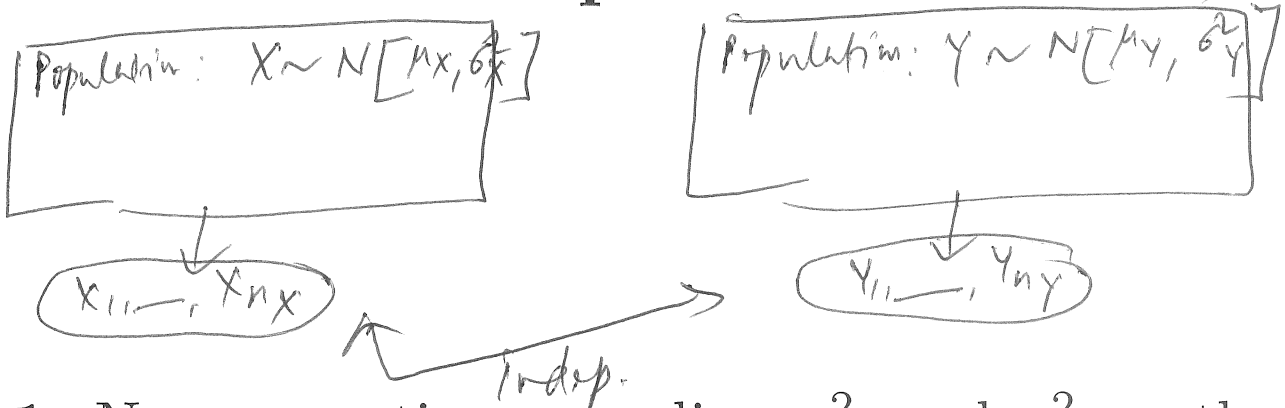
Pivot: $\dfrac{\bar{D}-\mu_D}{\sigma_D/\sqrt{n}} \approx \dfrac{\bar{D}-\mu_D}{s_D/\sqrt{n}} \overset{\cdot}{\sim} N(0,1)$, provided $n$ is large.

**Q:** What is the pivot here?

# CI for $\mu_X - \mu_Y$ with two independent samples (First: Normality)

goal:

**Setup:**




**Scenario 1:** No assumption regarding $\sigma_X^2$ and $\sigma_Y^2$ — they may be equal or unequal.

**Estimators of parameters:**

$$S_X^2 = \frac{1}{n_X - 1} \sum_{i=1}^{n_X} (X_i - \bar{X})^2$$

- $\mu_X$: $\bar{X}$
- $\mu_Y$: $\bar{Y}$
- $\sigma_X^2$: $S_X^2$ — sample var.
- $\sigma_Y^2$: $S_Y^2$ — sample var.
- $\mu_X - \mu_Y$: $\bar{X} - \bar{Y}$

**Pivot:** $\dfrac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\widehat{SE}(\bar{X}-\bar{Y})}$

$\sim t_\nu$

$\text{Var}(\bar{X}-\bar{Y}) \overset{\bar{X} \text{ and } \bar{Y} \text{ are indep.}}{=} \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}$

$\Rightarrow \widehat{SE}[\bar{X}-\bar{Y}] = \sqrt{\dfrac{S_X^2}{n_X} + \dfrac{S_Y^2}{n_Y}}$

- **Satterthwaite's approximation:** The distribution of the pivot can be approximated by $t_\nu$ distribution where

$$\nu = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{S_X^4}{n_X^2(n_X-1)} + \frac{S_Y^4}{n_Y^2(n_Y-1)}}$$

- Approximate $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$:

  $(\bar{X}-\bar{Y}) \pm t_{\nu,\alpha/2}\ \widehat{SE}(\bar{X}-\bar{Y})$.

- When $n_X, n_Y$ are large, the CI is: $(\bar{X}-\bar{Y}) \pm z_{\alpha/2}\ \widehat{SE}(\bar{X}-\bar{Y})$
- Is the normality assumption needed when $n_X, n_Y$ are large? NO.

Normality + equal variance. — boxplot can be used to check this assumption.

**Scenario 2:** Assume common variance, i.e., $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. ↑ unknown.

- Estimators of $(\mu_X, \mu_Y, \mu_X - \mu_Y)$: $\overline{X}$, $\overline{Y}$, $\overline{X} - \overline{Y}$
- Estimate $\sigma^2$ using the *pooled* sample variance $S_p^2$:

$$S_p^2 = \frac{(n_X - 1)\, S_X^2 + (n_Y - 1)\, S_Y^2}{n_X + n_Y - 2}$$

- Is $S_p^2$ unbiased for $\sigma^2$?

$$E[S_p^2] = \frac{(n_X - 1)\, \overbrace{E[S_X^2]}^{=\sigma^2} + (n_Y - 1)\, \overbrace{E[S_Y^2]}^{=\sigma^2}}{n_X + n_Y - 2} = \frac{\sigma^2 (n_X - 1 + n_Y - 1)}{n_X + n_Y - 2} = \sigma^2.$$

- ~~SE~~ $\operatorname{var}(\overline{X} - \overline{Y}) = \sigma^2 \left( \frac{1}{n_X} + \frac{1}{n_Y} \right)$
- Estimated SE: $\sqrt{S_p^2 \left( \frac{1}{n_X} + \frac{1}{n_Y} \right)} = S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$

**Pivot:** $\dfrac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\widehat{SE}(\bar{X}-\bar{Y})} = \dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}}$

$\sim t_{n_X+n_Y-2}$ (Result).

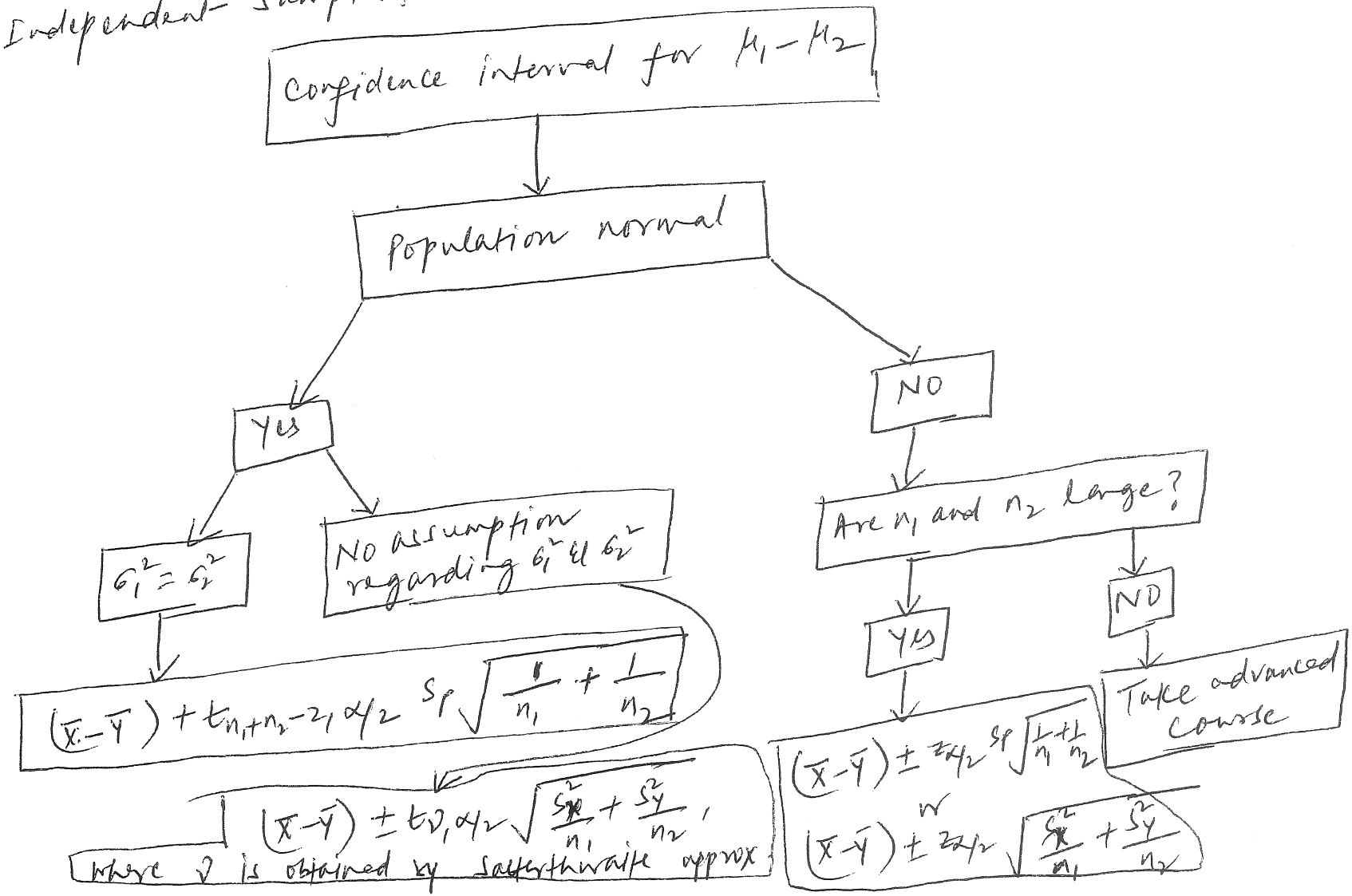
- $100(1-\alpha)\%$ CI for $\mu_X-\mu_Y$: $(\bar{X}-\bar{Y}) \pm t_{n_X+n_Y-2,\alpha/2}\,\widehat{SE}(\bar{X}-\bar{Y})$.
- When $n_X, n_Y$ are large, the CI is: $(\bar{X}-\bar{Y}) \pm z_{\alpha/2}\,\widehat{SE}(\bar{X}-\bar{Y})$.
- Is the normality assumption needed when $n_X, n_Y$ are large? No.

# Recap

## Comparing two population means:

(1) Paired samples: Apply one-sample procedures to the differences.

(2) Independent samples:



# Large-sample CI for $p_X - p_Y$ with two independent samples

**Setup:**

Pop 1: $X \sim \text{Bernoulli}(p_X)$ — $X_1, \ldots, X_{n_X}$ (indep.) $Y_1, \ldots, Y_{n_Y}$ — Pop 2: $Y \sim \text{Bernoulli}(p_Y)$

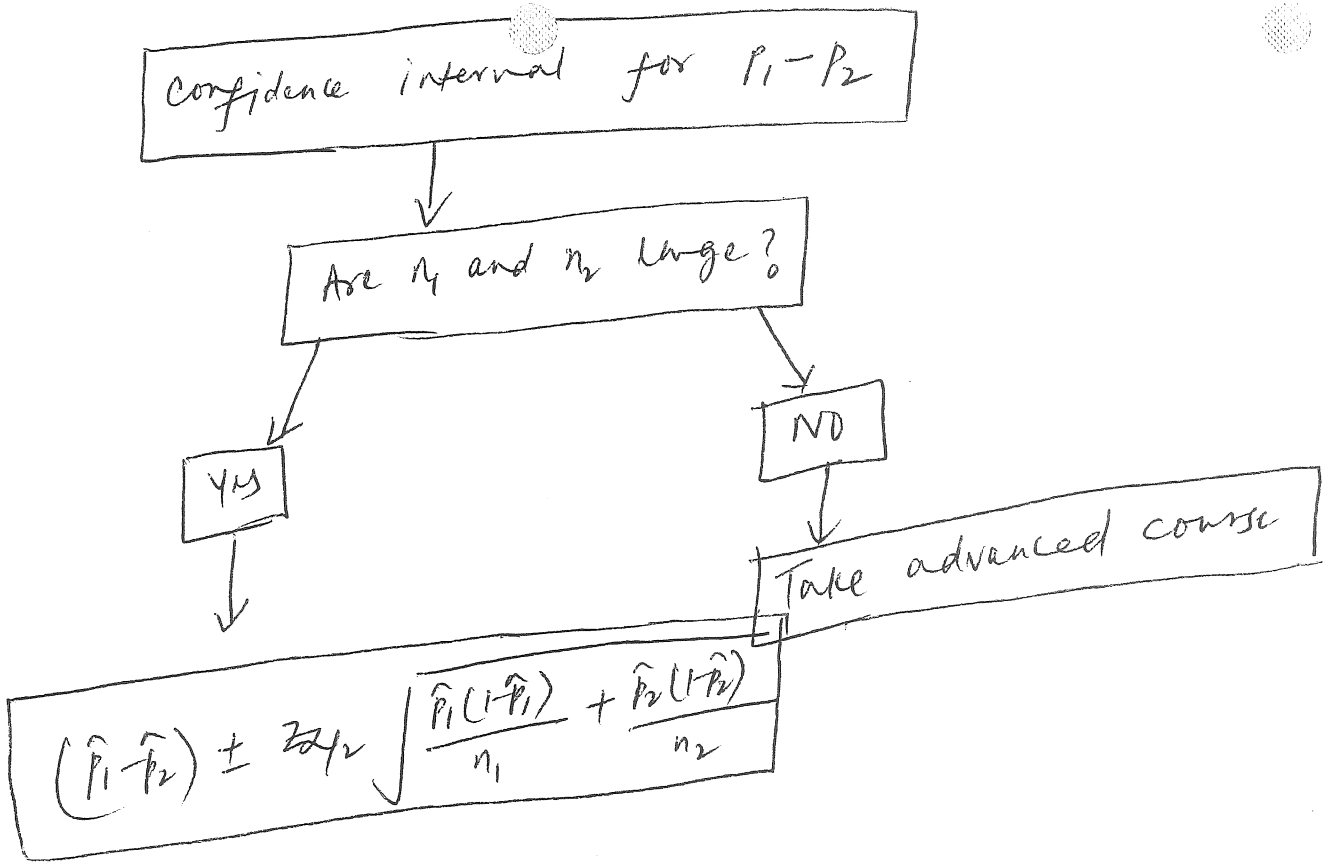
$p_X - p_Y = E[X] - E[Y] = \mu_X - \mu_Y$

- Estimators of $(p_X, p_Y, p_X - p_Y)$: $(\hat{p}_X, \hat{p}_Y, \hat{p}_X - \hat{p}_Y)$ ($\hat{p}_X = \bar{X}$, $\hat{p}_Y = \bar{Y}$)
- $SE(\hat{p}_X - \hat{p}_Y) = \sqrt{\text{Var}(\hat{p}_X - \hat{p}_Y)} = \sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}$
- Estimated SE: $\widehat{SE}(\hat{p}_X - \hat{p}_Y) = \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}$
- **Pivot:**

$$\frac{(\hat{p}_X - \hat{p}_Y) - (p_X - p_Y)}{\widehat{SE}(\hat{p}_X - \hat{p}_Y)} \approx N(0,1)$$

- Approximate $100(1-\alpha)\%$ CI: $(\hat{p}_X - \hat{p}_Y) \pm z_{\alpha/2}\, \widehat{SE}(\hat{p}_X - \hat{p}_Y)$,

provided both $n_X$ and $n_Y$ are large.


Confidence interval for $p_1 - p_2$
Are $n_1$ and $n_2$ large?
YES
NO
Take advanced course
$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

# CI for a function of $\theta$

**Issue:** Suppose $(L, U)$ is a $100(1-\alpha)\%$ CI for $\theta$. How to get a $100(1-\alpha)\%$ CI for $g(\theta)$, where $g$ is a monotonically increasing function of $\theta$?

Have: $P[L \le \theta \le U] = 1-\alpha$ (for all $\theta$).

$\Updownarrow$

$P[g(L) \le g(\theta) \le g(U)] = 1-\alpha$

$\Rightarrow$ $[g(L), g(U)]$ is a $100(1-\alpha)\%$ CI for $g(\theta)$.

# Example 2

A study shows that 61 of 414 adults who grew up in a single-parent household report that they suffered at least one incident of abuse during childhood. By contrast, 74 of 501 adults who grew up in two-parent households report abuse.

(a) Is there a difference in single-parent and two-parent households when it comes to reporting abuse?

$X$: $\begin{cases} 1, & \text{if abuse is reported in a single-parent household} \\ 0, & \text{o/w.} \end{cases}$

Bernoulli $(p_X)$

$Y = \begin{cases} 1, & \text{if} \quad " \quad " \quad " \quad \text{two-parent} \quad " \\ 0, & \text{o/w.} \end{cases}$

Bernoulli $(p_Y)$

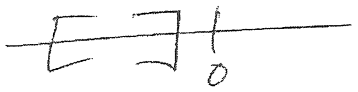
$p_X - p_Y$ = diff. of proportions in the two populations.

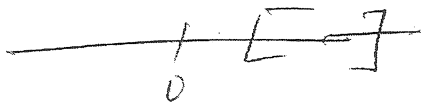
No difference: $p_X - p_Y = 0$.

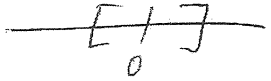
Construct 95% CI for $p_X - p_Y$: large sample Z interval.

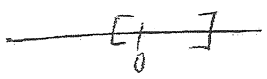
$\hat{p}_X = \frac{61}{414}$, $\hat{p}_Y = \frac{74}{501}$

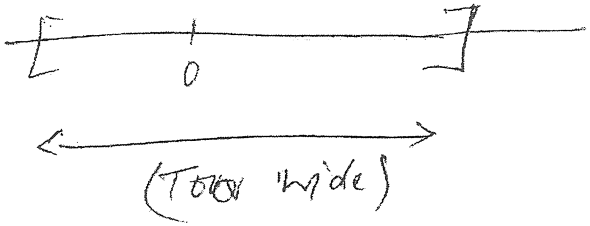
Consider the following scenarios. CI for $p_1 - p_2$

Case 1:  clearly: $p_1 < p_2$

Case 2:  clearly: $p_1 > p_2$

Case 3:  clearly: $p_1 = p_2$

Case 4:  conclude: $p_1 = p_2$

Case 5: 

Need more data (so that CI is not that wide)

Case 6: 
— may still conclude that $p_1 = p_2$, but this case is a borderline (need more data for a more exact conclusion)

From R: 95% CI for $p_x - p_y$ : $[-0.05, 0.05]$

(b) What assumptions, if any, did you make to compute the interval in (a)? Do the assumptions seem reasonable?

→ since 0 is right in the middle of the CI, we can conclude that $p_x = p_y$.

both $n_x$ and $n_y$ are large.

# Example 1

95% CI for $\mu_X - \mu_Y$:

The data below show the sugar content (as a % of weight) of several national brands of children's and adults' cereals.

$X$ = sugar content in a typical children's cereal — $\mu_X = E[X]$

**Children's cereals**: 40.3, 55, 45.7, 43.3, 50.3, 45.9, 53.5, 43, 44.2, 44, 47.4, 44, 33.6, 55.1, 48.8, 50.4, 37.8, 60.3, 46.5

$Y$ = sugar content in a typical adults' cereal — $\mu_Y = E[Y]$

**Adults' cereals**: 20, 30.2, 2.2, 7.5, 4.4, 22.2, 16.6, 14.5, 21.4, 3.3, 6.6, 7.8, 10.6, 16.2, 14.5, 4.1, 15.8, 4.1, 2.4, 3.5, 8.5, 10, 1, 4.4, 1.3, 8.1, 4.7, 18.4

(a) Is it reasonable to assume that each sample comes from a normal distribution?

using R — look at boxplot — NO, may be OK for children's cereals but not for adults' cereals

(b) Can the variances of the two distributions be assumed to be equal? Justify your answer.