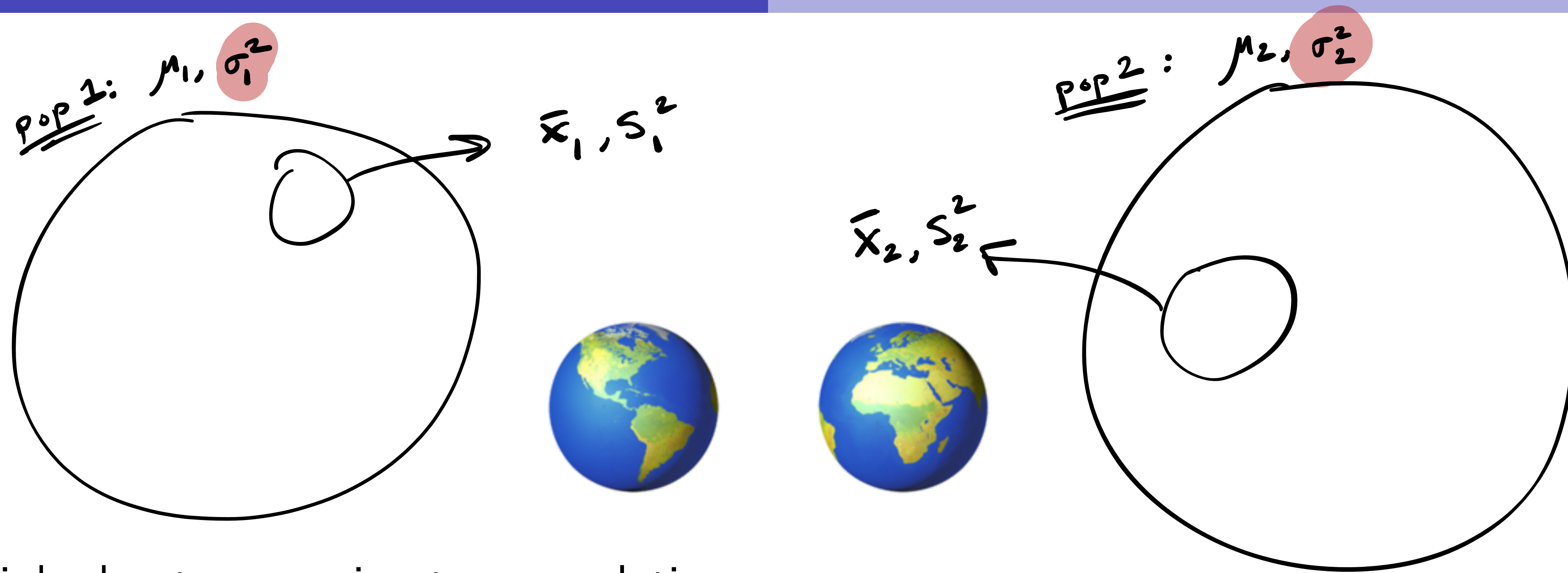
STAT 515 fa 2023 Lec 16 slides

Two-sample testing

Karl Gregory

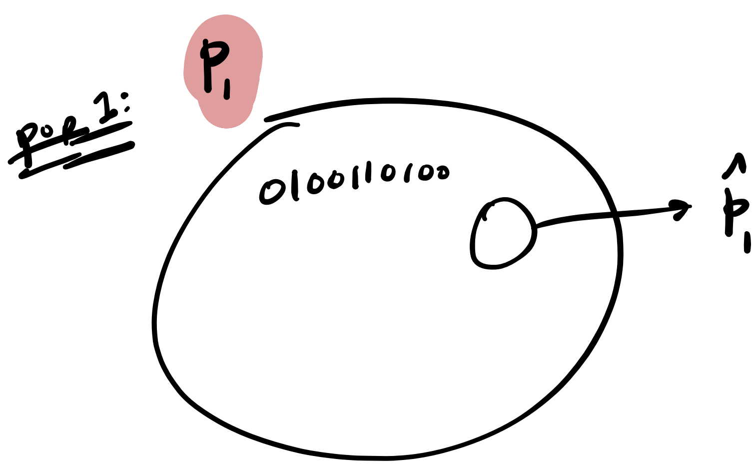
University of South Carolina

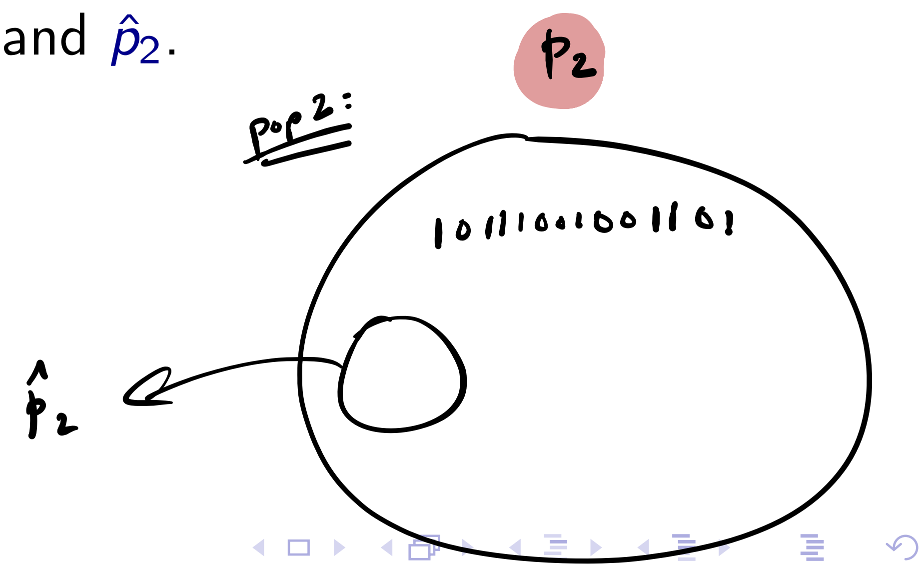
These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.



Think about comparing two populations:

- Compare μ_1 with μ_2 by comparing $\bar{X}_1$ and $\bar{X}_2$.
- Compare p_1 with p_2 by comparing $\hat{p}_1$ and $\hat{p}_2$.





Consider two random samples:

$$X_{11}, \dots, X_{1n_1} \text{ a rs from a pop. with mean } \mu_1 \text{ and variance } \sigma_1^2$$
$$X_{21}, \dots, X_{2n_2} \text{ a rs from a pop. with mean } \mu_2 \text{ and variance } \sigma_2^2$$

Goals: Compare μ_1 and μ_2

1. Build confidence intervals for $\mu_1 - \mu_2$
2. Test null and alternate hypotheses of the form "delta"

$$H_0\colon \mu_1 - \mu_2 \geq \delta_0 \quad \text{or} \quad H_0\colon \mu_1 - \mu_2 = \delta_0 \quad \text{or} \quad H_0\colon \mu_1 - \mu_2 \leq \delta_0$$
$$H_1\colon \mu_1 - \mu_2 < \delta_0 \qquad\quad H_1\colon \mu_1 - \mu_2 \neq \delta_0 \qquad\quad H_1\colon \mu_1 - \mu_2 > \delta_0.$$

In most situations we have $\delta_0 = 0$.

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$
$\rightarrow$
$H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 \neq 0$

Exercise: Write down the null and alternate hypotheses for the following:

1. Do honors grads earn more in first post-grad year than non-honors grads?
2. Do PhD holders earn more than ~~at least~~ twice as much as Bachelor's degree holders?
3. Does a fertilizer increase crop yields?

① μ_1 = honors μ_2 = non-honors

rewrite as

$H_0: \mu_1 - \mu_2 \le 0$ $H_1: \mu_1 > \mu_2 \rightarrow \mu_1 - \mu_2 > 0$

② μ_1 : phD μ_2 : Bachelor

$H_0 : \mu_1 \leq 2\mu_2$ $H_1 : \mu_1 > 2\mu_2$

$\rightarrow \mu_1 - 2\mu_2 > 0$

③ μ_1 : with Fertilizer μ_2 = without Fertilizer

$H_0 : \mu_1 - \mu_2 \leq 0$. $H_1 : \mu_1 > \mu_2$

$\downarrow$

$\mu_1 - \mu_2 > 0$

Sampling distribution of difference in sample means

If both populations are Normal, then $\bar{X}_1 - \bar{X}_2 \sim \text{Normal}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$.

Take $\bar{X}_1 - \bar{X}_2$ into the "t-world" by

replace σ_1^2, σ_2^2 with s_1^2, s_2^2

$$\boxed{\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_{\text{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}} \quad \text{or} \quad \boxed{\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}}$$

not exactly t-distributed.

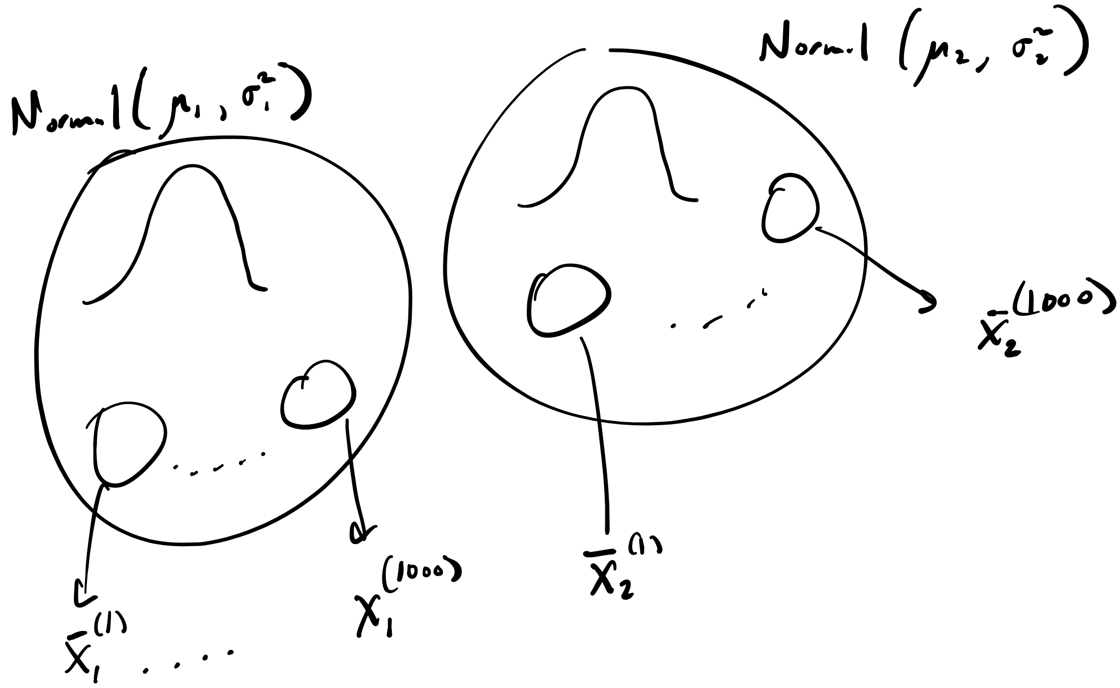
depending on whether $\sigma_1^2 = \sigma_2^2$ or $\sigma_1^2 \neq \sigma_2^2$, respectively.

In the above $S^2_{\text{pooled}} = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$.

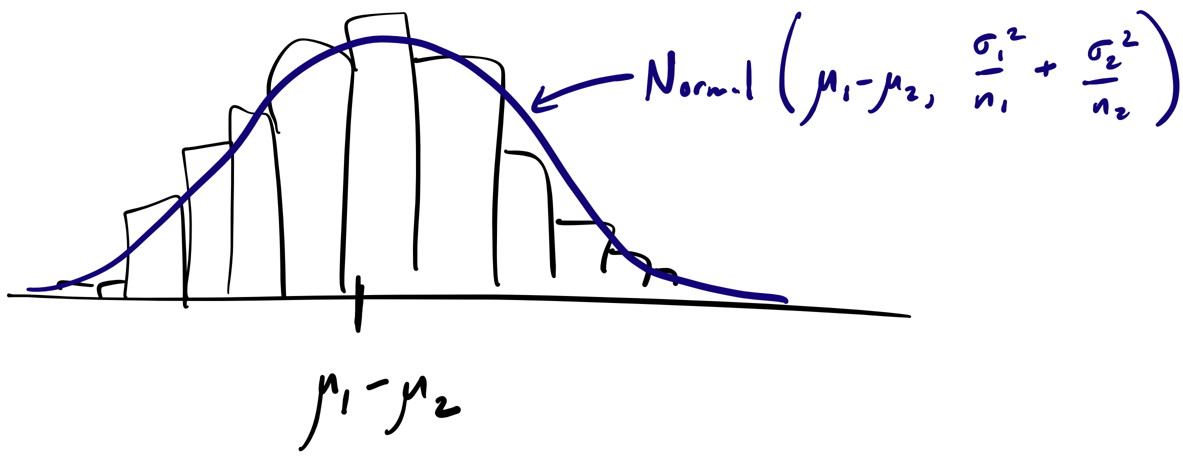
In case $\sigma_1^2 = \sigma_2^2$, this Estimates the common variance.

Think about a C.I. for $\mu_1 - \mu_2$.

Need to know how $\bar{x}_1 - \bar{x}_2$ behaves.

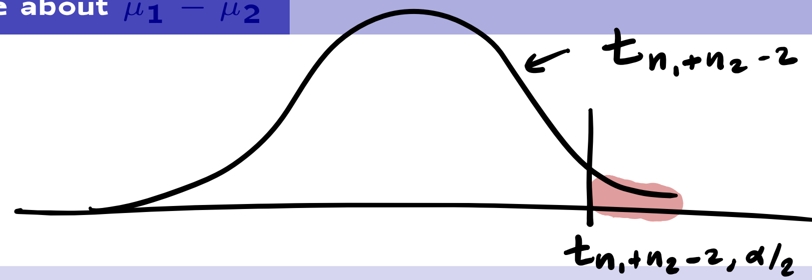


histogram of $\bar{x}_1 - \bar{x}_2$ values looks like



So, send $\bar{x}_1 - \bar{x}_2$ into the Z world with

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$



Confidence intervals for $\mu_1 - \mu_2$ when both populations are Normal

A $(1-\alpha) \times 100\%$ CI for $\mu_1 - \mu_2$ is given by

$$\bar{X}_1 - \bar{X}_2 \pm \boxed{t_{n_1+n_2-2,\alpha/2}} S_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad \text{if } \sigma_1^2 = \sigma_2^2$$

$$\bar{X}_1 - \bar{X}_2 \pm t_{\nu^*,\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \qquad \text{if } \sigma_1^2 \neq \sigma_2^2.$$

In the above $\nu^* = \underbrace{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2 \left[\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}\right]^{-1}}_{\text{t Welch}}.$

Exercise: We wish to know whether drug tablets produced at two different sites have the same average concentration of the drug. (Ex 6.92 in [2]).

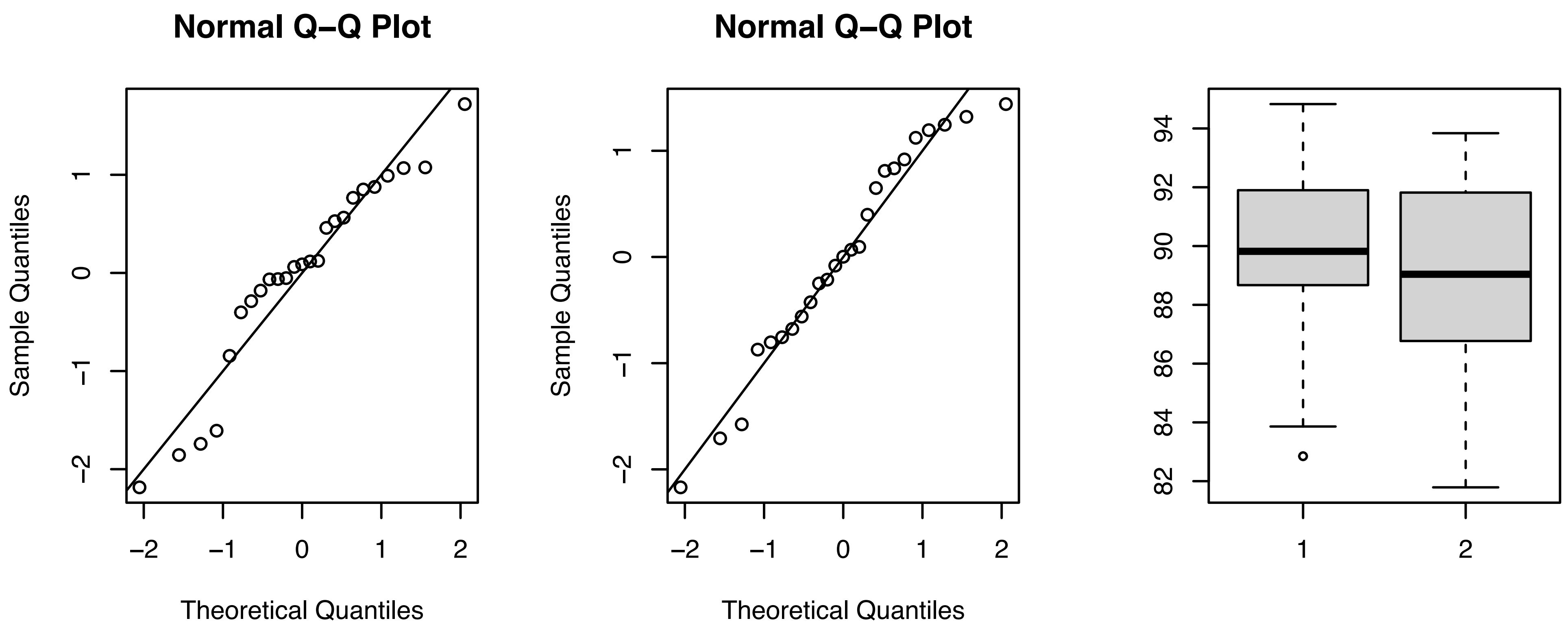
μ_1 = mean for site 1 Build 95% C.I. for $\mu_1 - \mu_2$

μ_2 = mean for site 2

Site 1			Site 2		
91.28	86.96	90.96	89.35	87.16	93.84
92.83	88.32	92.85	86.51	91.74	91.20
89.35	91.17	89.39	89.04	86.12	93.44
91.90	83.86	89.82	91.82	92.10	86.77
82.85	89.74	89.91	93.02	83.33	83.77
94.83	92.24	92.16	88.32	87.61	93.19
89.93	92.59	88.67	88.76	88.20	81.79
89.00	84.21		89.26	92.78	
84.62	89.36		90.36	86.35	

$n_1 = 25$ $n_2 = 25$

Assuming Normality build a 95% confidence interval for the difference in means assuming first $\sigma_1^2 = \sigma_2^2$ and then $\sigma_1^2 \neq \sigma_2^2$.



For drug tablets data:

$H_0: \mu_1 - \mu_2 = 0$ $\qquad$ $H_1: \mu_1 - \mu_2 \neq 0$

```
site1 <- c(91.28,92.83,89.35,91.90,82.85,94.83,89.93,89.00,84.62,
           86.96,88.32,91.17,83.86,89.74,92.24,92.59,84.21,89.36,
           90.96,92.85,89.39,89.82,89.91,92.16,88.67)
site2 <- c(89.35,86.51,89.04,91.82,93.02,88.32,88.76,89.26,90.36,
           87.16,91.74,86.12,92.10,83.33,87.61,88.20,92.78,86.35,
           93.84,91.20,93.44,86.77,83.77,93.19,81.79)

n1 <- length(site1)
n2 <- length(site2)
xbar1 <- mean(site1)
xbar2 <- mean(site2)
s1 <- sd(site1)
s2 <- sd(site2)

sp <- sqrt(((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2))
me <- qt(0.975,n1 + n2 - 2) * sp * sqrt(1/n1 + 1/n2)
d <- xbar1 - xbar2
lo <- d - me
up <- d + me
```

```
nu <- (s1^2/n1 + s2^2/n2)^2 / ((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1))
me2 <- qt(0.975,nu) * sqrt(s1^2/n1 + s2^2/n2)
lo2 <- d - me
up2 <- d + me
```

$H_0: \mu_1 - \mu_2 = 1$ vs $H_1: \mu_1 - \mu_2 \neq 1$

t.test(site1, site2, mu = 1)

$H_0: \mu_1 - \mu_2 \leq 1$ vs $H_1: \mu_1 - \mu_2 > 1$

t.test(site1, site2, mu = 1, alternative = "greater") null difference in means (δ_0)

Build CIs and obtain *p*-values with `t.test()` in R.

1. For $\sigma_1^2 = \sigma_2^2$, we can use

   Also conf.level = 0.99 for example

   ```
   t.test(x1, x2, var.equal=TRUE, mu=0, alternative="two.sided")
   ```

   default

2. For $\sigma_1^2 \neq \sigma_2^2$, we can use

   ```
   t.test(x1, x2, var.equal=FALSE, mu=0, alternative="two.sided")
   ```

Change the `alternative` and `mu` arguments to test other sets of hypotheses.

Run `?t.test` to read the documentation.

"1" "2"

Default: Test $H_0: \mu_1 - \mu_2 = 0$

vs $H_1: \mu_1 - \mu_2 \neq 0.$

```
> t.test(site1,site2,var.equal = TRUE)

Two Sample t-test

data:  site1 and site2
t = 0.57214, df = 48, p-value = 0.5699
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.304376  2.341976
sample estimates:
mean of x mean of y
  89.5520   89.0332
```

$n_1 + n_2 - 2 = 25 + 25 - 2 = 48$

Fail to reject H_0

for testing these hypotheses.

95% CI for $\mu_1 - \mu_2$

$\bar{X}_1$ $\bar{X}_2$

$$T_{test} = \frac{\bar{X}_1 - \bar{X}_2}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

```
> t.test(site1,site2,var.equal = FALSE)

Welch Two Sample t-test

data:  site1 and site2
t = 0.57214, df = 47.659, p-value = 0.5699
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.304712  2.342312
sample estimates:
mean of x mean of y
  89.5520   89.0332
```

$\hat{\nu}^*$

$$T_{test} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Must have both populations Normal *or* $n_1 \geq 30$ and $n_2 \geq 30$ to use these.

HW 8 Q 2 (a) (i)

(i) Enriched leads to heavier?

1 Inference about $\mu_1 - \mu_2$

2 Inference about $p_1 - p_2$

X_i = Enriched$_i$ − Impoverished$_i$

$\Rightarrow X_1, \ldots, X_{11}$

Let μ be the mean difference.

(Enriched − Impoverished).

$H_0: \mu \leq 0$ vs $H_1: \mu > 0$

p_1 and p_2

Exercise: Write down the null and alternate hypotheses for the following:

1. Do same ~~number~~ proportion of honors and non-honors students pursue grad school?
2. Does a vaccine reduce the probability of getting an infection?
3. Do rural and urban voters differ in their preferences for a candidate?

① p_1: Honors p_2: non-Honors

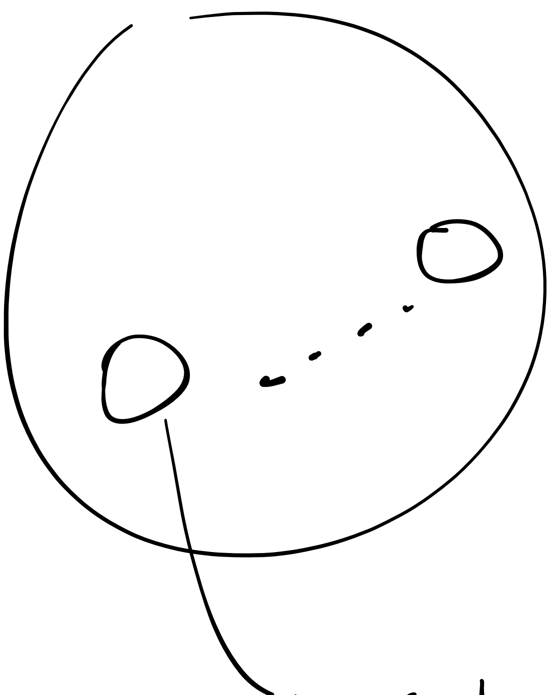
$H_0: p_1 - p_2 = 0$ vs $H_1: p_1 - p_2 \neq 0$.

② p_1: control p_2: Vaccine

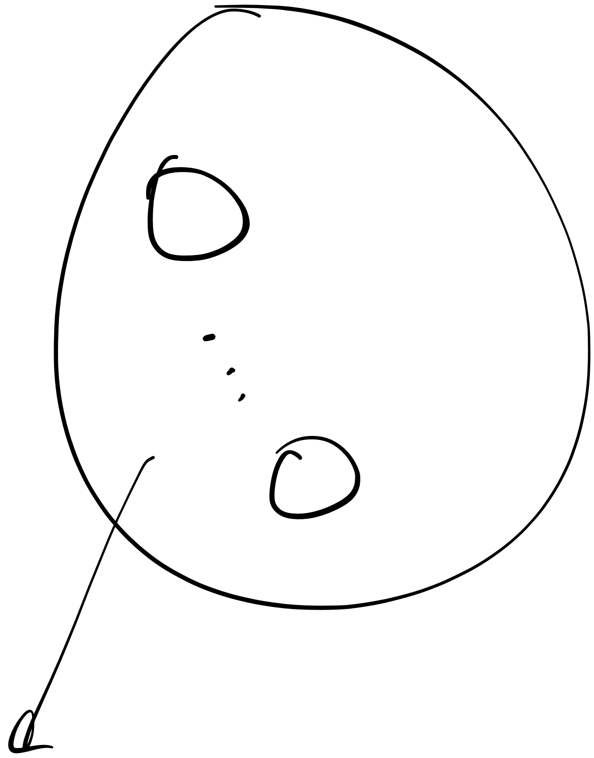
$H_0: p_1 - p_2 \leq 0$ $H_1: p_1 - p_2 > 0$

pop 1: p_1

pop 2: p_2

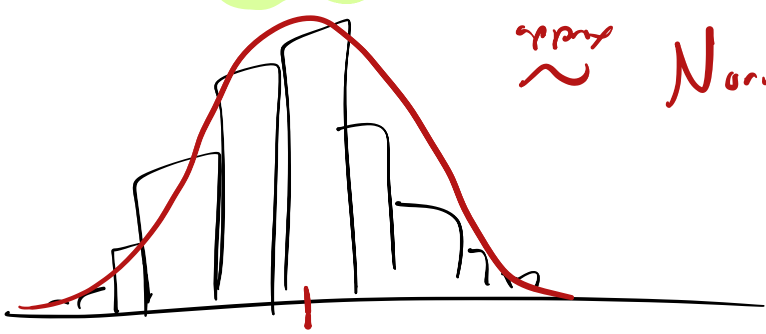


many samples, many $\hat{p}_1$ values



many samples, many $\hat{p}_2$ values

$\hat{p}_1 - \hat{p}_2$



$p_1 - p_2$

$$\hat{p}_1 - \hat{p}_2 \overset{\text{approx}}{\sim} \text{Normal}\left(p_1 - p_2,\ \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

send to Z world.

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Let $X_{k1}, \dots, X_{kn_k} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_k)$, $k = 1, 2$, and let $\hat{p}_1 = \bar{X}_1$, $\hat{p}_2 = \bar{X}_2$.

Sampling distribution of difference in sample proportions

For larger and larger n_1 and n_2, the quantity

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \text{ behaves more and more like } Z \sim \text{Normal}(0,1)$$

Rule of thumb: Need $\min\{n_1\hat{p}_1, n_1(1-\hat{p}_1)\} \geq 15$ and $\min\{n_2\hat{p}_2, n_2(1-\hat{p}_2)\} \geq 15$.

Recall: CI for p: $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (Wald-type)

Confidence interval for difference in proportions

An approximate $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

provided $\min\{n_1\hat{p}_1, n_1(1-\hat{p}_1)\} \geq 15$ and $\min\{n_2\hat{p}_2, n_2(1-\hat{p}_2)\} \geq 15$.

Need at least 15 "successes" & 15 "failures" in each sample.

What is $p_1 - p_2$?

1 : 1^{st} class

2 : 3^{rd} class

Exercise: It is reported that among the 319 adult first class passengers aboard the Titanic, 197 survived, while among the 627 adult third class passengers, 151 survived. The data are taken from [1].

Build a 95% confidence interval for the difference in the "true" proportions as a way of assessing whether the probability of surviving was affected by class.

$$\hat{p}_1 = \frac{197}{319} = 0.618$$

$$\hat{p}_2 = \frac{151}{627} = 0.241$$

$$0.618 - 0.241 \pm \underbrace{1.96}_{= z_{\frac{0.05}{2}}} \sqrt{\frac{.618(1-.618)}{319} + \frac{0.241(1-.241)}{627}}$$

$$= (0.314, .440)$$

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

If $p_1 = p_2$, pool the data to estimate the common proportion, p_0.

Tests about $p_1 - p_2$

Define the test statistic

$$Z_{\text{test}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Then for n_1, n_2 large, the following tests have (approx) $P(\text{Type I error}) \leq \alpha$.

$H_0\colon p_1 - p_2 \geq 0$
$H_1\colon p_1 - p_2 < 0$

Reject H_0 if
$Z_{\text{test}} < -z_\alpha$

$p\text{-val} = P(Z < Z_{\text{test}})$

$H_0\colon p_1 - p_2 = 0$
$H_1\colon p_1 - p_2 \neq 0$

Reject H_0 if
$|Z_{\text{test}}| > z_{\alpha/2}$

$p\text{-val} = 2 \cdot P(Z > |Z_{\text{test}}|)$

$H_0\colon p_1 - p_2 \leq 0$
$H_1\colon p_1 - p_2 > 0$

Reject H_0 if
$T_{\text{test}} > z_\alpha$

$p\text{-val} = P(Z > Z_{\text{test}})$

In the above $\hat{p}_0 = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$. ← estimator of p_0 if $p_1 = p_2 = p_0$.

1: 15-17
2: 25-35

Exercise: Suppose that in random samples of size 1000 of 15-17 yr-olds and 25-35 yr-olds, 6% and 3%, respectively, were found to have used JUUL in the last month. You wish to know if the $\boxed{\text{proportion is higher in the younger age group.}}$ This exercise is based on some summary statistics given in [3].

$p_1 > p_2$

$p_1 - p_2 > 0$

1. Give the hypotheses of interest.
2. What is our conclusion at the $\alpha = 0.01$ significance level?

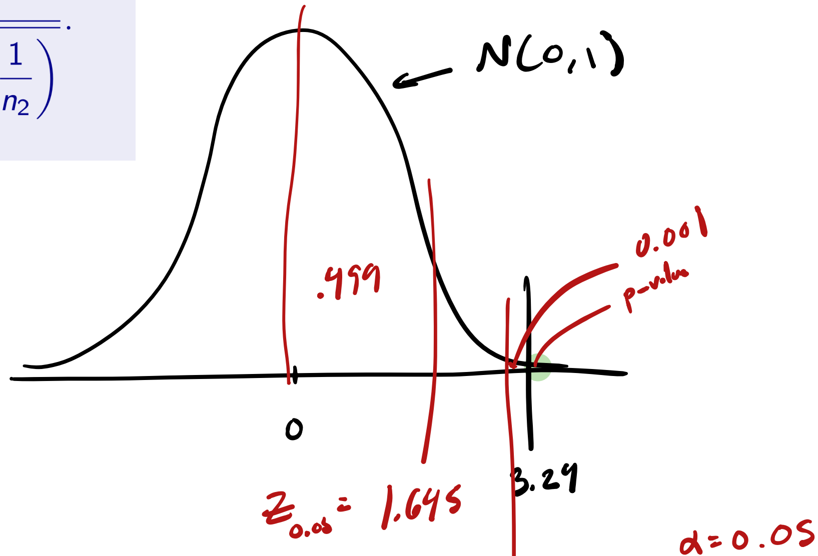
(1) $H_0: p_1 - p_2 \leq 0$ vs $H_1: p_1 - p_2 > 0$

(2) $\hat{p}_1 = 0.06$ $\quad \hat{p}_2 = 0.03$ $\quad \hat{p}_0 = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2} = \dfrac{60 + 30}{2000} = \dfrac{90}{2000} = .045$

$n_1 = 1000$ $\quad n_2 = 1000$

$$Z_{test} = \frac{0.06 - 0.03}{\sqrt{0.045(1-0.045)\left(\frac{1}{1000} + \frac{1}{1000}\right)}} = 3.24$$

$$Z_{\text{test}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$



$H_0: p_1 - p_2 \le 0$ vs $H_1: p_1 - p_2 > 0$

Since $3.24 > Z_{0.05} = 1.645$, Reject H_0 at $\alpha = 0.05$.

OR: p-value is < 0.001. $\Rightarrow$ Reject H_0.

Robert J MacG Dawson.
The "unusual episode" data revisited.
Journal of Statistics Education, 3(3), 1995.

J.T. McClave and T.T. Sincich.
Statistics.
Pearson Education, 2016.

Donna M Vallone, Morgane Bennett, Haijun Xiao, Lindsay Pitzer, and Elizabeth C Hair.
Prevalence and correlates of juul use among a national sample of youth and young adults.
Tobacco Control, 2018.