

# **Statistics 431:**

# **Statistical Inference**

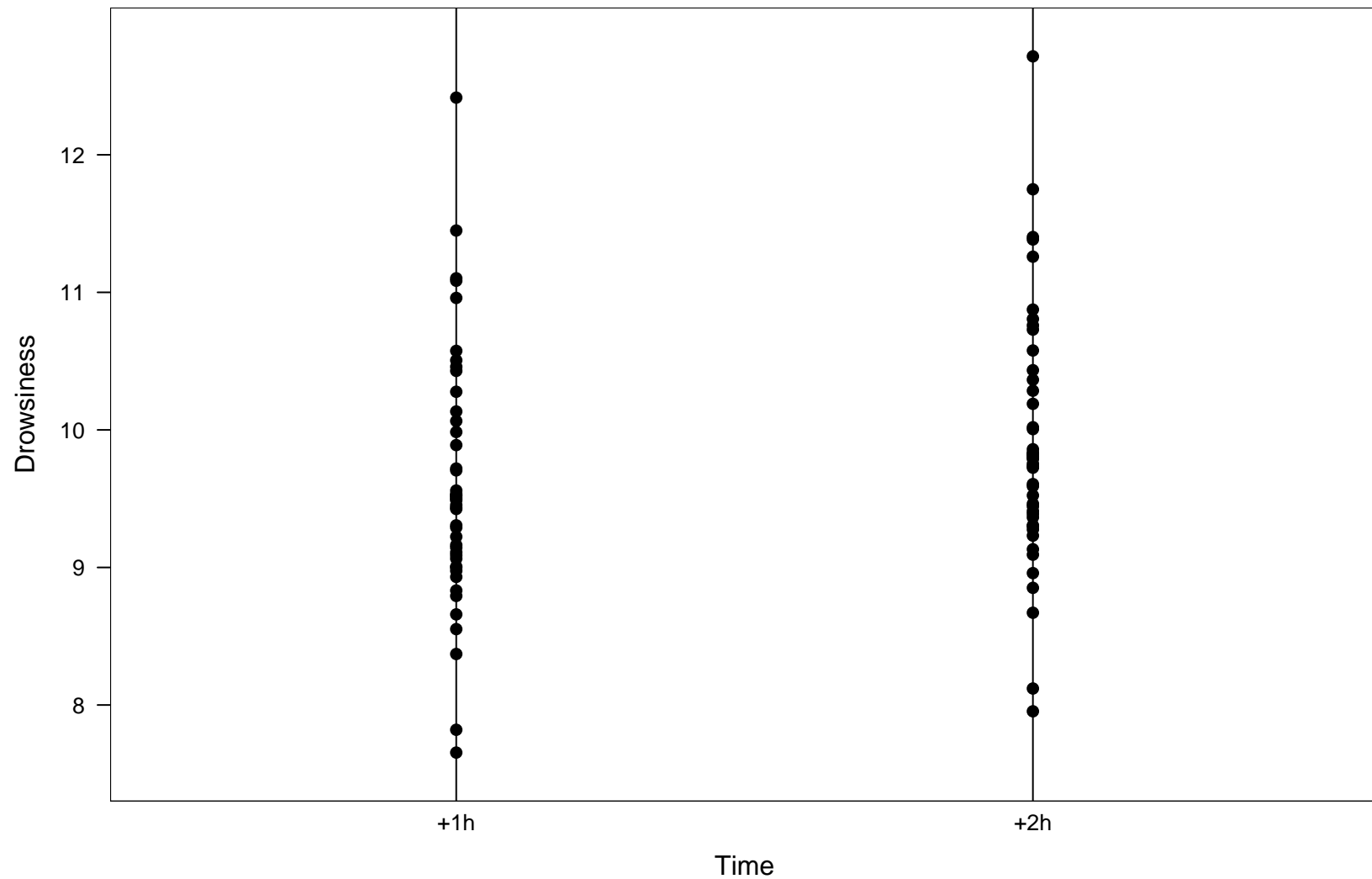
**Lecture 8: More on two-sample inference**

# Paired samples: basics

- Our two samples  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$  up to now have been *unpaired*: there's no correspondence between observations in the first sample and observations in the second. (The number of  $X_i$ 's is not even the same as the number of  $Y_i$ 's, in general.)
- In some situations where  $m = n$ , there is a natural matching of each  $X_i$  with one  $Y_i$ , to form the  $n$  pairs  $(X_i, Y_i)$ .
- Examples:
  - $n$  patients in a clinical trial are given a sedative. The drowsiness of each patient  $i$  is measured one hour ( $X_i$ ) and two hours ( $Y_i$ ) after treatment.
  - $n$  houses are selected at random in Philadelphia. At each house  $i$ , the radon level is measured in the basement ( $X_i$ ) and on the highest floor ( $Y_i$ ).
  - For each of  $n$  sets of identical twins, we measure the body mass index of the older ( $X_i$ ) and younger ( $Y_i$ ) twin.
- Instead of comparing the sampled  $X_i$ 's to the sampled  $Y_i$ 's, as for unpaired data, we should focus on comparing the pairs  $(X_i, Y_i)$  to each other.

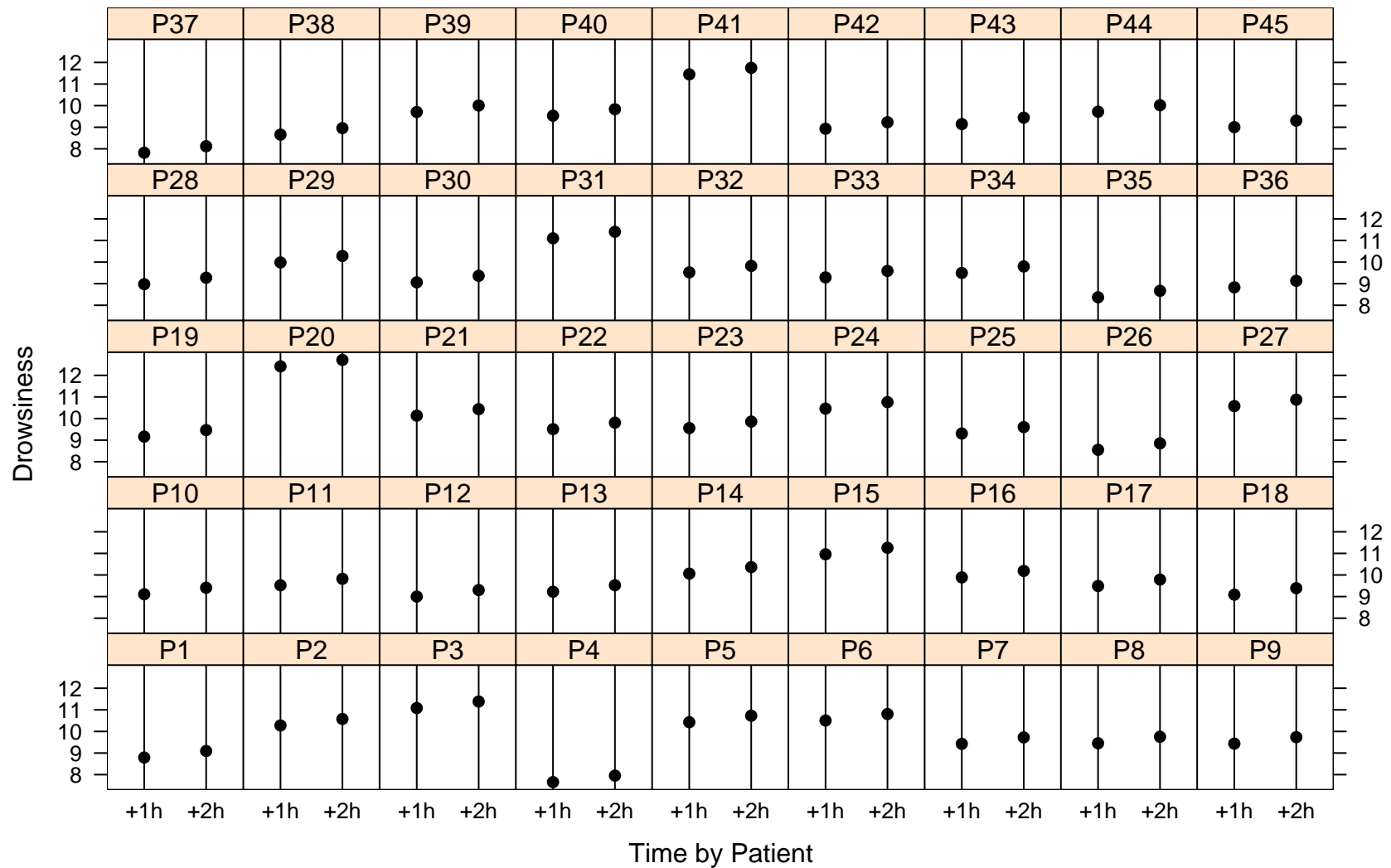
# Why look at pairs?

- Ignore pairing



# Why look at pairs?

- Include pairing



# Difference between means: paired test

- The data:  $(X_1, Y_1), \dots, (X_n, Y_n) \sim p(x, y)$ , IID. We are not assuming the  $X_i$ 's are independent of the  $Y_i$ 's as we did with unpaired data.
- $EX_i = \mu_1, EY_i = \mu_2$ .
- Define the within-pair differences  $D_1 = X_1 - Y_1, \dots, D_n = X_n - Y_n$ .
- Let  $\mu_D = ED_i$  (we know  $\mu_D = \mu_1 - \mu_2$ ), and let  $\sigma_D^2 = \text{Var}(D_i)$ .
- Under  $H_0: \mu_D = \Delta_0$  (the same null as with unpaired data), and substituting in the sample variance, we get

$$T = \frac{\bar{D} - \Delta_0}{S_D/\sqrt{n}} \approx N(0, 1)$$

for large samples. For small samples,  $T \sim t_{n-1}$ , once we add the assumption that each  $D_i$  is normal.

- As before:  $H_A: \mu_D \neq \Delta_0 \Rightarrow$  reject when  $|T| > c_\alpha$   
 $H_A: \mu_D > \Delta_0 \Rightarrow$  reject when  $T > c_\alpha$   
 $H_A: \mu_D < \Delta_0 \Rightarrow$  reject when  $T < -c_\alpha$ .
- As before, set  $c_\alpha$  using  $N(0, 1)$  in large samples and  $t_{n-1}$  in small samples.

# Difference between means: paired CI

- This should start to look familiar:
- $T = (\bar{D} - \mu_D)/(S/\sqrt{n})$  is a pivot. In large samples,  $T \approx N(0, 1)$ , and in small samples under a normal assumption,  $T \sim t_{n-1}$ .
- Using the same pivot statement as always,

$$P\left(-z_{\alpha/2} < \frac{\bar{D} - \mu_D}{S/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha,$$

we get the  $100(1 - \alpha)\%$  large-sample paired CI

$$\bar{D} \pm z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

- The same goes for a small normal sample, with  $z_{\alpha/2}$  replaced by  $t_{\alpha/2;n-1}$ .

# Collect paired or unpaired data?

- We can be more precise about the benefits of paired data.
- Just like  $\text{Var}(\bar{X}) = \sigma_X^2/m$  and  $\text{Var}(\bar{Y}) = \sigma_Y^2/n$ , we know that  $\text{Var}(\bar{D}) = \sigma_D^2/n$ .
- But  $\sigma_D^2 = \text{Var}(D_i) = \text{Var}(X_i - Y_i) = \sigma_X^2 + \sigma_Y^2 - 2 \cdot \text{Cov}(X_i, Y_i)$ .
- So: when  $X_i$  and  $Y_i$  are positively correlated,  $\text{Var}(\bar{D})$  is smaller than when they are uncorrelated.
- Very often, the within-pair correlation is positive, rather than negative. In a paired analysis, the positive correlation is reflected in  $S_D^2$ , which usually turns out *smaller* than  $S_X^2 + S_Y^2$ .
- However: with  $n$   $X_i$ 's and  $n$   $Y_i$ 's, we have  $2n$  unpaired obsvns, but only  $n$  pairs. So the reference  $t$  distrn in a small sample will have  $2n - 1$  obsvns for unpaired data, which is better than  $n - 1$  for paired data.
- Despite this, for moderate  $n$ , the (usually) smaller SE of  $\bar{D}$  leads to increased testing power via a larger  $T$  statistic, and narrower CIs.
- If  $X_i$  and  $Y_i$  are negatively correlated in a paired dataset, you still need to do a paired analysis, but you would have been better off without pairing!

# Inference about two population proportions

- Recall the one-sample setup:  $p$  is the proportion of “successes” in a population, i.e. the fraction of the population possessing some characteristic.
- Now we have two populations: the success fraction in population “A” is  $p_1$ , and in “B” it is  $p_2$ .
- We draw  $m$  times independently from population “A”, and call  $X$  the number of successes in the sample. Similarly, we make  $n$  independent draws from population “B”, and call  $Y$  the number of successes.
- Then  $X \sim \text{Bin}(m, p_1)$  and  $Y \sim \text{Bin}(n, p_2)$ . We assume the two samples are drawn independently, so  $X$  is independent of  $Y$ .
- Example: in 1954, a large-scale randomized controlled experiment was conducted to study Salk’s polio vaccine. Randomizing a child to the control (placebo) group is like drawing from a population having probability  $p_1$  of “success” (contracting polio). Randomizing a child to the treatment (vaccination) group is like drawing from a population with probability  $p_2$  of contracting polio.
- In 1954, people were *very* interested in  $p_1 - p_2$ .

# Two population proportions: large-sample tests

- The natural estimator of  $p_1 - p_2$  is  $\hat{p}_1 - \hat{p}_2$ , where  $\hat{p}_1 = X/m$  and  $\hat{p}_2 = Y/n$ .
- Facts:  $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$ ,  $\text{Var}(\hat{p}_1 - \hat{p}_2) = p_1(1 - p_1)/m + p_2(1 - p_2)/n$ .
- So, when both  $m$  and  $n$  are large,

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{p_1(1 - p_1)/m + p_2(1 - p_2)/n}} \approx N(0, 1) .$$

We have a (large-sample) pivot: all we need for testing and CIs.

- Based on  $Z$ , you should be able to derive a test procedure for  $H_0 : p_1 - p_2 = \Delta_0$  vs.  $H_A : p_1 - p_2 \neq \Delta_0$ , for any  $\Delta_0$ .
- However, there is a wrinkle when  $\Delta_0 = 0$ . In this case,  $p_1 = p_2 = p$ , and the denominator of  $Z$  is just  $\sqrt{p(1 - p)(1/m + 1/n)}$ . We should estimate this directly, rather than plugging  $\hat{p}_1$  and  $\hat{p}_2$  into the original denominator.

- If  $p_1 = p_2 = p$ , then  $X \sim \text{Bin}(m, p)$ , and, independently,  $Y \sim \text{Bin}(n, p)$ . But then  $X + Y \sim \text{Bin}(m + n, p)$ .
- So a natural estimator of  $p$  under  $H_0 : p_1 = p_2 = p$  is just  $\hat{p} = (X + Y)/(m + n)$ .
- This results in the large-sample test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(1/m + 1/n)}}$$

and the procedures

$H_A : p_1 - p_2 \neq 0 \Rightarrow$  reject when  $|Z| > z_{\alpha/2}$

$H_A : p_1 - p_2 > 0 \Rightarrow$  reject when  $Z > z_{\alpha}$

$H_A : p_1 - p_2 < 0 \Rightarrow$  reject when  $Z < -z_{\alpha}$ .

# Two population proportions: large-sample CIs

- With both  $m$  and  $n$  large, and substituting sample proportions in the denominator,

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}_1(1 - \hat{p}_1)/m + \hat{p}_2(1 - \hat{p}_2)/n}} \approx N(0, 1) .$$

- Try this at home: write down the  $100(1 - \alpha)\%$  pivoting confidence statement for  $p_1 - p_2$ , using  $Z$ . Rearrange it to obtain the confidence interval

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{m} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n}} .$$

- A suggested correction when  $m$  and  $n$  are not huge: replace  $\hat{p}_1 = X/m$  with  $\tilde{p}_1 = (X + 1)/(m + 2)$  and  $\hat{p}_2$  with the analogous  $\tilde{p}_2$ . This can improve the quality of the normal approximation, hence the correctness of the CI.
- When  $m$  or  $n$  is quite small? A different approach is needed, which we won't cover. (Take more statistics courses.)