STAT 5200 Handout #4.a Tests and Distributions (Ch. 2)

Recall general framework of a test:

1. State null ("empty") and alternative hypothesis

   H_0: "no effect" vs. H_a: "some nonzero effect"

2. Calculate test statistic

   (standardized difference)

3. Obtain p-value from <u>Sampling distribution</u>
   (If had lots of samples, and H_0 were true,
   what is the [theoretical] distribution of the test statistic?) (Not: distribution of this sample of data).

4. Make conclusion (in context)

Handout #4: A vs. B: $\underbrace{Y_{A1}, \ldots, Y_{An_A}}_{\text{symbolic rep. of data}}$ "iid" $N(\mu_A, \sigma_A^2)$, $Y_{B1}, \ldots, Y_{Bn_B}$ iid $N(\mu_B, \sigma_B^2)$

iid → independent and identically distributed

subscript: group & rep. #

n_A = total sample size in group A

μ_B ← mean, σ_B^2 ← variance

1. H_0: $\mu_A = \mu_B$ vs. H_a: $\mu_A \neq \mu_B$

   ↳ true pop. mean # lice same for both chemicals

2. With s_p^2 = pooled variance estimate, (assume $\sigma_A^2 = \sigma_B^2$)

   pooled ↓ like a weighted ave. of s_A^2 & s_B^2

   $s_A^2 = \hat{\sigma}_A^2$

$$t = \frac{\overbrace{\bar{Y}_B - \bar{Y}_A}^{\text{diff. in means}}}{\underbrace{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}_{\text{standardizing value}}}$$

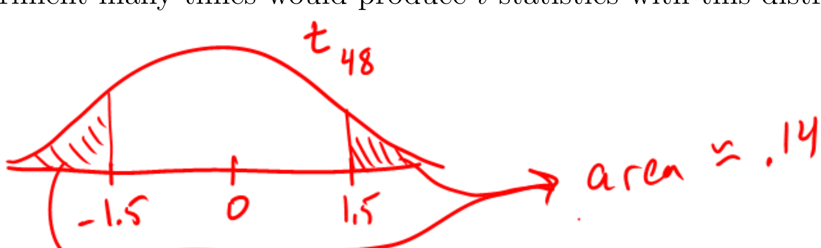
   t ↓ -1.50 here

3. Under H_0 (i.e., if <u>H_0 is true</u>, and model assumptions are met),

$$t \sim t_{n_A+n_B-2},$$

   so repeating this experiment many times would produce t statistics with this distribution:

total area under curve is 1



   P-value = probability of observing a result [test statistic] at least as extreme as this, just by chance when H_0 is true.

   extreme: H_a tells direction

4. Fail to reject H_0 because: .14 > .05 (α-level)
   Conclude:

   No signif. evidence of difference between chemicals A & B

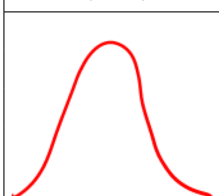
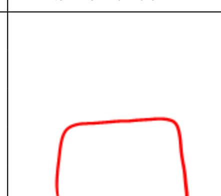
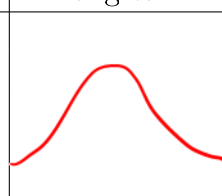
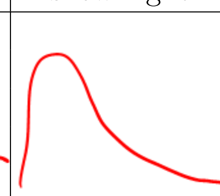
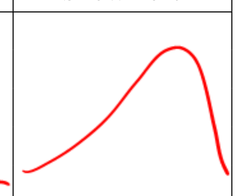
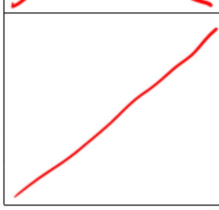
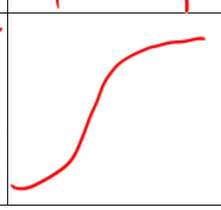
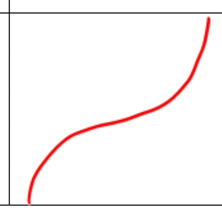
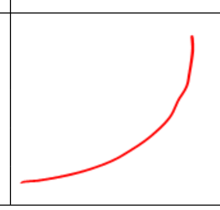
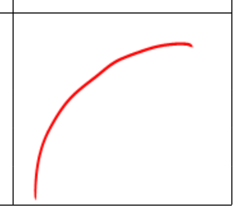
Important aside:

- "p-values, confidence intervals, and other statistical measures are all uncertain. Treating them otherwise is immodest overconfidence."[1]
- ATOM: Accept uncertainty. Be thoughtful, open, and modest.[1]
- We will use small p-values as suggestive of further interest, not as definitive proof

But wait – the t distribution is only appropriate when: model assumptions are met
(need to check distribution of # lice)

Graphical checks: (for normality; more later)

- boxplot – kind of clunky, but useful for summarizing symmetry and "outliers" (don't over-interpret)
- histogram – hard to gauge deviations from normal curve
- normal probability plot – compare observed values (vertical axis) to expected [normal] values (horizontal axis); examples:

distribution:	normal	short-tail	long-tail	skew right *	skew left
histogram:					
normal prob: (QQ plot)					

What to do if assumptions are violated?

- We'll come back to this more in Ch. 6, but in general:

  transform response variable
  (log, or √, or square, or ...) -or- assume a different distribution (later in semester)

  (drop "outliers" only as a last resort)
- Can also consider a non-parametric test → makes no assumption about distribution

[1]Wasserstein et al. 2019. Moving to a World Beyond "p < 0.05"

Example: if t distribution with $n_A + n_B - 2$ d.f. isn't appropriate distribution, what is?

- One way: Generate sampling distribution by permutations.
  - H_0 says treatment labels don't matter, so randomly rearrange them and re-calculate t; so this many [all possible] times and use the distribution of all possible t values

  - Example of "resampling" (book refers to as randomization testing; related to bootstrap sampling)

    ↳ generate sampling dist'n w/ repeat samples of data

- Another way: Rather than using actual data values, look at permutation of ranks

  (Wilcoxon Rank-Sum or Mann-Whitney or Kruskal-Wallis)

  → see PROC NPAR1WAY

A numerical test for normality? (NOTE: try "energy" test, too)

- Several exist, but easy to over-interpret. If distribution is roughly bell-shaped, any test requiring normality usually okay
- H_0: data normal vs. H_a: data not normal
  - For $10 < n < 2000$: Shapiro-Wilks Test
  - For $n > 2000$: Kolmogorov-Smirnov Test

  But beware, because tests are hyper-sensitive (end of HO #4)

  - even slight departures from normality will be exaggerated

- Best to rely on: Graphical checks (esp. QQ or normal prob. plot)