MONASH University

MONASH
BUSINESS
SCHOOL

Statistical Thinking (ETC2420/ETC5242)

Associate Professor Catherine Forbes

Week 4: Resampling techniques for assessing variability in means

Learning Goals for Week 4

- Review the Central Limit Theorem
- Apply one and two sample t-tests and confidence intervals
- Build Bootstrap confidence interval for numerical data
- Distinguish between independent and paired samples

Assigned reading for Week 4:

- Chapter 4 (skip Section 4.4) in ISRS

These slides supplement the Week 4 videos (OpenIntro)

- Involving the application of CLT-based tests and confidence intervals for a mean parameter
  - or for a difference in two mean parameters
  - in paired or independent samples

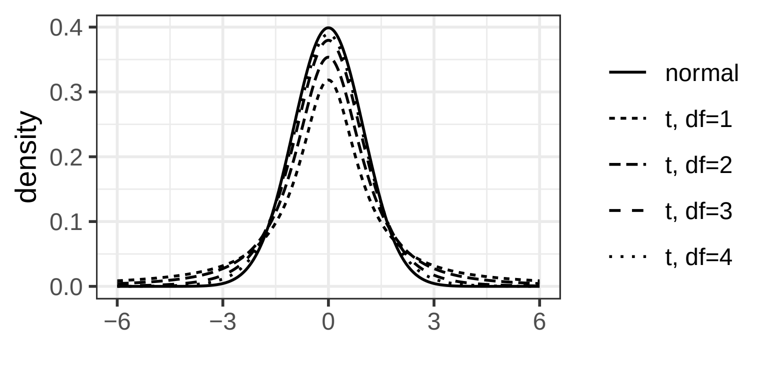
1. The t-distribution

Review: What purpose does a large sample serve?

- As long as observations are independent
- And the population distribution is not “extremely skewed”
- A “large” sample would ensure that…
  - the sampling distribution of the mean is nearly normal
  - the estimate of the standard error ($\frac{s}{\sqrt{n}}$) is reliable
  - (if skewed, even larger sample size is needed)
- $\Rightarrow$ s will be a good estimate of population standard deviation, σ

Video 1: The t-distribution

- Use when σ is unknown (almost always the case) to address the uncertainty of standard error estimate
- Is “bell shaped” but with thicker tails than the normal
  - centered at zero
  - one parameter: degrees of freedom (df) determine thckness of tails
  - compare with $N(\mu, \sigma^2)$, two parameters (mean=μ and SD =σ)



Video 1:: The t-statistic

- for inference on a mean where
  - σ unknown, which is almost always
- calculated the same way

$$T = \frac{obs - null}{SE}$$

Here:

- *obs* refers to the value of an observed statistic
  - $\bar{x}$ from one sample, including $\bar{x}_{DIff}$ for paired samples, where $d_i = x_{1i} - x_{2i}$ is the "DIff" for pair i
  - $\bar{x}_1 - \bar{x}_2$ from two independent samples
- *null* refers to the corresponding value of the population quantity under H_0
- *SE* refers to the *Standard Error*, which is the standard deviation of the statistic

- p-value calculated using *R*
  - one or tail areas, based on H_1

Video 1: Practice

- Find the following probabilities:
  a. $\Pr(|Z| > 2)$ = 0.0455 ($\rightarrow$ reject)
  b. $\Pr(|t_{df=50}| > 2)$ = 0.0509 ($\rightarrow$ fail to reject?)
  c. $\Pr(|t_{df=10}| > 2)$ = 0.0734 ($\rightarrow$ fail to reject)
- (And suppose you have a two sided hypothesis test, and your test statistic is 2. Under which scenario would you be able to reject H_0 and the 5% significance level?)
- Generally degrees of freedom (df) is tied to the sample size

Video 2: Inference for a mean

- Biscuits after lunch study

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

- estimating the mean (single sample): point estimate ± margin of error

$$\bar{x} \pm t^{\star}_{df} SE_{\bar{x}}$$

$$\bar{x} \pm t^{\star}_{df} \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t^{\star}_{n-1} \frac{s}{\sqrt{n}}$$

- Degrees of freedom for t statistic for inference on one sample mean: $df = n - 1$
- Find the critical t score using R

```
[1] -2.07961
```

Video 2: Application: Confidence interval

- Estimate the average after-lunch snack compumption (in grams) of people who eat lunch **distracted** using a 95% confidence interval

$$\begin{aligned}\bar{x} \pm t^{\star} SE &= 52.1 \pm 2.08 \times \frac{45.1}{\sqrt{22}} \\ &= 52.1 \pm 2.08 \times 9.62 \\ &= 52.1 \pm 20 \\ &\Rightarrow (32.1,\ 72.1)\end{aligned}$$

- $\Rightarrow$ we are 95% confident that distracted eaters consume between 32.1 to 72.1 grams of snacks post-meal.

Video 2: Application: Hypothesis test

- Suppose the suggested service size of these biscuits is 30g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size?

$$H_0 : \mu = 30 \quad \text{vs.} \quad H_1 : \mu \neq 30$$

$$T = \frac{52.1 - 30}{9.62} = 2.3$$

- p-value: use 2 × probability greater than T under t distribution with df=21:

```
[1] 0.0318023
```

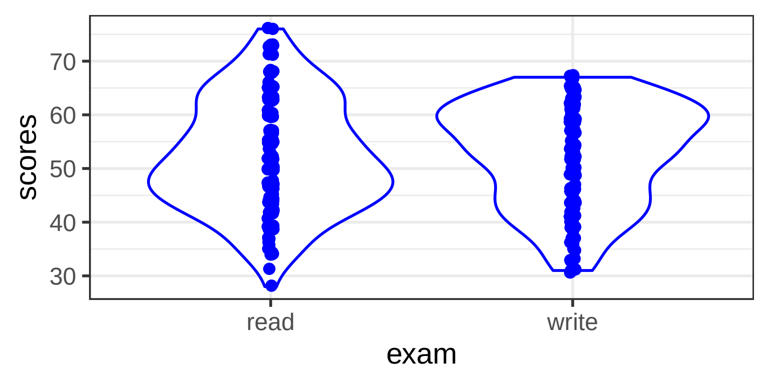
- $\Rightarrow$ Reject H_0 at level $\alpha = 0.05$ since p-value = $0.0318 < 0.05$

Video 2: Conditions

- independent observations
  - random assignment
  - 22 < 10% of all distracted eaters
- sample size/ skew
  - data are likely right-skewed
- We would like to have all of the data, not just the summary statistics

Video 3: Inference for paired data

- We can sumarise paired data to re-use the same method above



Video 3: Analysing paired data

- When two sets of observations have this special correspondence (not independent), they are said to be **paired**
- To analyse paired data, it is often useful to look at the difference in outcomes of each pair of observations:

$$DIff_i = read_i - write_i, \text{ for each } i = 1, 2, \ldots, n = 200$$

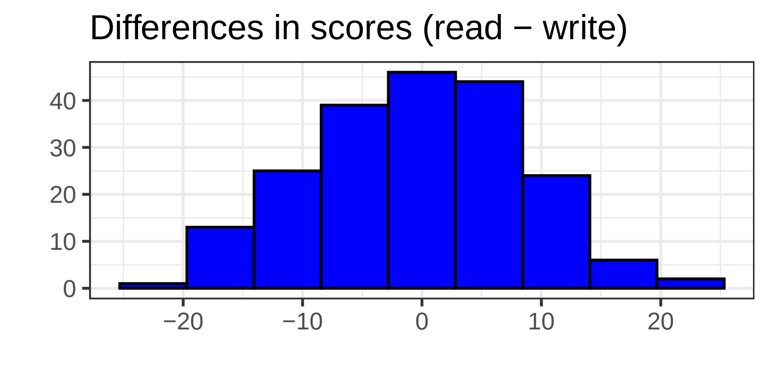
- parameter of interest: Average difference between the reading and writing scores of **all** high school students

xbar_DIff	s_DIff	n_DIff	tstat
-0.545	8.88667	200	-0.867

- point estimate? Average difference between the reading and writing scores of **sampled** high school students: $\bar{x}_{DIff}$

Video 3: Using data to understand the population quantities

- If in fact there was no population difference between the average reading and writing scores, what would you expect the average sample difference to be?



Video 3: Hypothesis for paired means

- $H_0 : \mu_{DIff} = 0 \quad \text{vs.} \quad H_1 : \mu_{DIff} \neq 0$
- Same structure as one-sample mean test
- Test statistic:
$$T = \frac{-0.545 - 0}{8.887 \big/ \sqrt{200}} = -0.867$$
- degrees of freedom: 200 - 1 = 199
- $\Rightarrow$ p-value: 2*pt(-0.867, df=199) = 0.387
- Interpretation?: *p-value of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the scores is zero.*

$$\text{Pr(observed or more extreme outcome} \mid H_0 \text{ is true)}$$

Video 4: Inference for comparing two independent means

- Refer back to earlier **Biscuits after lunch study**
- point estimate ± margin of error

$$(\bar{x}_1 - \bar{x}_2) \pm t^{\star} SE_{(\bar{x}_1 - \bar{x}_2)}$$

- Standard error of difference between two independent means: $SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- We add the two variances (inside the square root) even though we are looking for the Standard error of the difference ($SE_{(\bar{x}_1 - \bar{x}_2)}$)
- Video advocates:
  DF for t statistic for inference on difference of two independent means: $df = min(n_1 - 1, n_2 - 1)$

Video 4: Conditions for inference for comparing two independent means

Video 1: Independence:

- Within groups: sampled observations must be independent
  - random sample/assignment
  - if sampling without replacement, $n < 10\%$ of population
- Between groups: the two groups must be independent of each other (non-paired)

2. Sample size/skew: The more skewed the populations, the large the sample size we need from those distributions

Video 4: Calculations

- $\Rightarrow$ Confidence interval for $\mu_{wd} - \mu_{wod}$: (1.83g, 48.17g)
  - obtained from:

$$\begin{aligned}(\bar{x}_{wd} - \bar{x}_{wod}) \pm t^{\star}_{df}\, SE &= (52.1 - 27.1) \pm 2.08 \times \sqrt{\frac{45.1^2}{22} + \frac{26.4^2}{22}} \\ &= 25 \pm 2.08 \times 11.14 \\ &= 25g \pm 23.17g\end{aligned}$$

- $\Rightarrow$ Hypothesis test:

$$H_0 : \mu_{wd} - \mu_{wod} = 0 \quad \text{vs.} \quad H_1 : \mu_{wd} - \mu_{wod} \neq 0$$

- $T_{21} = \frac{25-0}{11.14} = 2.24 \Rightarrow$ p-value: 2*pt(2.24, df=21, lower.tail=FALSE) = 0.036
  - Reject $H_0 : \mu_{wd} - \mu_{wod} = 0$ at the 5% level

- Now on to Week 5...