

STAT 401-F; EXAM 1 - SOLUTIONS

Q2: FALSE: Once the CI is computed, μ is either in the CI or not, we don't know which one is true.
We do know that 95% of the times we sample and compute the CI, it will contain μ (5% of the sample CI will not contain μ).

FALSE: \bar{y} is always in the CI: $\bar{y} \pm t \times \text{se.}(\bar{y})$

FALSE: The CI is to estimate μ , not to predict individual sample values.

TRUE: 95% of the CI should contain μ (95% of 300 is 285).

Q3: a) Precision of the mean is measured by its standard error.
In this case, the s.e. for the 3 years are: 5, 4 and 6.
Therefore, year 2 had the most precise estimator of prevalence.

b) iii: Width of CI: $2 \times t \times \frac{s}{\sqrt{n}}$. If $\uparrow \Rightarrow$ width \uparrow

c) reduces bias; allows for inference to the larger population.

Q4: a) Independent samples, equal sample sizes,
some indication of skewness but not enough to prevent
a t-test.

$$H_0: \mu_1 = \mu_2$$

$$s_p = 20.915$$

$$H_a: \mu_1 \neq \mu_2$$

$$t = \frac{382 - 59}{20.915 \sqrt{\frac{1}{5} + \frac{1}{6}}} = -1.64$$

$$d.f. = 9$$

2 sided p-value lies between (0.1 and 0.2)

No evidence from the data that one drink caused a higher mean
score than the other.

b) Paired data, some large values but not enough to prevent
a paired t-test: 3 -13 -3 -6 5 -15 -19 -4 -12 → diffs.

$$H_0: \mu_D = 0$$

$$H_a: \mu_D \neq 0$$

$$\text{mean} = -7$$

$$s.d. = 8.28$$

$$d.f. = 8$$

$$\Rightarrow t = \frac{-7}{\frac{8.28}{\sqrt{9}}} = -2.53$$

2 sided p-value between (0.025 and 0.05)

⇒ moderate evidence that one drink caused a higher mean
score than the other.

c) Analysis in part (b) is better because it accounts for
variation among subjects (which is evidently large in
this experiment. See subject 16 compared to the others).

Q5: a) Sign test. (paired data & no numerical values)

Example:

Parent	1	2	3	4	5
Large	y	y	y	Ny	y
Small	y	Ny	y	y	N	
Sign	0	+	0	-	+	

count # of non-zero pairs $\Rightarrow n$

count # of positive diffs $\Rightarrow k$...

b) t-test (paired diffs) or signed-rank

c) signed-rank (outlier in differences, small sample)

d)

differences:	3	3	-3	-14	2
abs. diffs:	2	3	3	3	14
rank	1	2	3	4	5
rank	1	3	3	3	5

negative diffs.

$$S = \text{sum of positive (diff.) ranks} = 1 + 3 + 3 = 7$$

e) No, too many ties.

f) p-value = proportion of obs. S values as large as 7
there are 18 such values, 32 total # of values
p-value is: $\frac{18}{32} = 0.5625 \rightarrow$ VERY large

\Rightarrow NO evidence that the seedling height is different for the large and small seeds.

Q6: a) Independent samples

Right skewed (median > mean)

Unequal standard deviations

Group with larger mean has larger s.d.

Log data doesn't have any of these problems

→ log transform data and perform a 2-indep. sample t-test on it.

b) To compute a CI for the medians, first compute a CI for the means of the log-transformed data and use the back-transform (and use the apparent symmetry of the log-transformed data)

$$\bar{Z}_{\text{grassy}} - \bar{Z}_{\text{forest}} = 4.462 - 3.874 = 0.588$$

$$\text{Thus } \frac{\text{Med grassy}}{\text{Med forest}} = \exp(0.588) = 1.8$$

$$\text{and } 90\% \text{ CI: } 0.588 \pm \underbrace{0.599}_{\substack{\text{Sp} \\ \text{for log-data}}} \sqrt{\frac{1}{11} + \frac{1}{21}} \times \underbrace{1.697}_{\substack{\text{t with} \\ 30 \text{ d.f.}}} = (0.21, 0.97)$$

$\downarrow \exp$
(1.23, 2.63)

It appears that the median ^{height} in the grassy area is about 1.8 times larger than the median height in the forest area, with a 90% CI of (1.23 to 2.63). Thus 2 seems plausible.