

GOVT 6019: Homework 6 — Solutions

Fall 2026 · Inference

Melissa Sands (Instructor)

Izza Malik (Teaching Assistant)

Question 1: True or false

(a) False. Overlapping intervals can still mean a statistically significant difference. With equal standard errors, the intervals stop overlapping at a gap of $2 \times 1.96 = 3.92$ SE, but the difference is statistically significant once the gap passes $1.96\sqrt{2} \approx 2.77$ SE. Gaps between 2.77 and 3.92 SE give overlapping intervals and a statistically significant difference.

(b) True. If A's interval contains B's mean, the gap is less than 1.96SE_A . That is less than $1.96\sqrt{\text{SE}_A^2 + \text{SE}_B^2}$, the cutoff for the (Welch) test of the difference.

(c) False. The p-value is the probability of a test statistic at least as extreme as t_{obs} , computed assuming H_0 is true. It is not the probability that H_0 is true.

(d) True. The t_9 critical value is 2.262, larger than 1.96, because s estimates σ with error. An interval built with 1.96 is too narrow, so it covers μ less than 95% of the time.

(e) False. Choosing the direction after seeing the data means the researcher would have reported a one-sided result in whichever direction the estimate fell. The test then rejects whenever $|t_{\text{obs}}|$ passes the one-sided 5% cutoff in either tail, so the Type I error rate is 10%, not 5%.

(f) True. A smaller α moves t_{crit} outward, so the test rejects less often. That lowers the Type I error rate. With n fixed, it raises the chance of failing to reject a false H_0 , the Type II error rate.

Question 2: The great British butter shortfall

(a)

- Salted: $H_0: \mu_S = 500$ against $H_1: \mu_S \neq 500$.
- Unsalted: $H_0: \mu_U = 500$ against $H_1: \mu_U \neq 500$.

(b)

With one block of each kind, $\bar{x} = x$ and the standard error is $\sigma/\sqrt{1} = \sigma$:

$$z = \frac{x - 500}{\sigma}.$$

It is a z -test because σ is known, so the statistic has a $\mathcal{N}(0,1)$ distribution under H_0 . A t -test replaces σ with the sample standard deviation s . With $n = 1$ there is no spread to measure: s divides by $n - 1 = 0$ and is undefined. So no t -test is possible.

(c)

```
sigma <- 40
x <- c(salted = 416, unsalted = 423)
z <- (x - 500) / sigma
z
```

```
      salted unsalted
      -2.100  -1.925
```

```
qnorm(.975)
```

```
[1] 1.959964
```

$z_S = -84/40 = -2.10$ and $z_U = -77/40 = -1.925$. The critical value is 1.96.

- Salted: $|z_S| = 2.10 > 1.96$. Reject H_0 .
- Unsalted: $|z_U| = 1.925 < 1.96$. Fail to reject H_0 .

(d)

```
2 * pnorm(-abs(z))
```

```
      salted  unsalted
0.03572884 0.05422936
```

- Salted: $p = .036$. If Asda's salted blocks averaged 500 g, a block at least 84 g away from 500 g would turn up only 3.6% of the time. The data give statistically significant evidence that the mean weight of Asda's salted butter is not 500 g.
- Unsalted: $p = .054$. A block at least 77 g away from 500 g would turn up 5.4% of the time. That is not statistically significant at the 5% level. We cannot conclude that the mean weight of Asda's unsalted butter differs from 500 g. We have not shown that it equals 500 g either.

(e)

- Salted: $H_0: \mu_S = 500$ against $H_1: \mu_S < 500$.
- Unsalted: $H_0: \mu_U = 500$ against $H_1: \mu_U < 500$.

Also accept $H_0: \mu \geq 500$. The test is the same either way.

```
qnorm(.05)
```

```
[1] -1.644854
```

```
pnorm(z)
```

```
      salted  unsalted
0.01786442 0.02711468
```

The test rejects when $z < -1.645$. Both tests now reject: $p_S = .018$ and $p_U = .027$. The unsalted decision changes. The one-sided test puts all of $\alpha = .05$ in the lower tail, so its cutoff, -1.645 , sits closer to 0 than the two-sided cutoff, -1.96 . Its p -value is half the two-sided one, because only the lower tail counts as evidence against H_0 . $z_U = -1.925$ falls between the two cutoffs.

(f)

```
pnorm(485, mean = 500, sd = 40) # (i)
```

```
[1] 0.3538302
```

```
sigma_max <- 15 / qnorm(.975) # (ii)  
sigma_max
```

```
[1] 7.653202
```

```
z_max <- (x - 500) / sigma_max # (iii)  
z_max
```

```
      salted  unsalted  
-10.97580 -10.06115
```

```
2 * pnorm(-abs(z_max))
```

```
      salted      unsalted  
4.996256e-28 8.203571e-24
```

- i. 35.4% of blocks would weigh less than 485 g, against the 2.5% guideline. A packer with $\sigma = 40$ g would fail it even when it fills to 500 g on average.
- ii. The guideline holds when $P(X < 485) \leq .025$, that is, when $500 - 1.96\sigma \geq 485$. So $\sigma \leq 15/1.96 = 7.65$ g.
- iii. $z_S = -10.98$ and $z_U = -10.06$. Both p -values are essentially 0, so both tests reject H_0 . The unsalted decision changes. The statistic divides the gap from 500 g by σ . The same 77 g shortfall counts as weak evidence when blocks vary by 40 g and as overwhelming evidence when they vary by under 8 g. A z -test is only as credible as its assumed σ .

(g)

Both blocks break the rule: 416 g and 423 g are both below 470 g. No hypothesis test is needed. The rule applies to each pack, and each block's weight is observed directly with a calibrated scale. Under the calibration assumption, each block's weight is known.

The tests in (a)–(e) ask a different question. They concern μ , the mean weight of all the blocks Asda packs, which we do not observe. One block is a sample of size 1 from that population. The test asks whether a block this light would be unusual if μ were 500 g.

Question 3: Working 9 to 5?

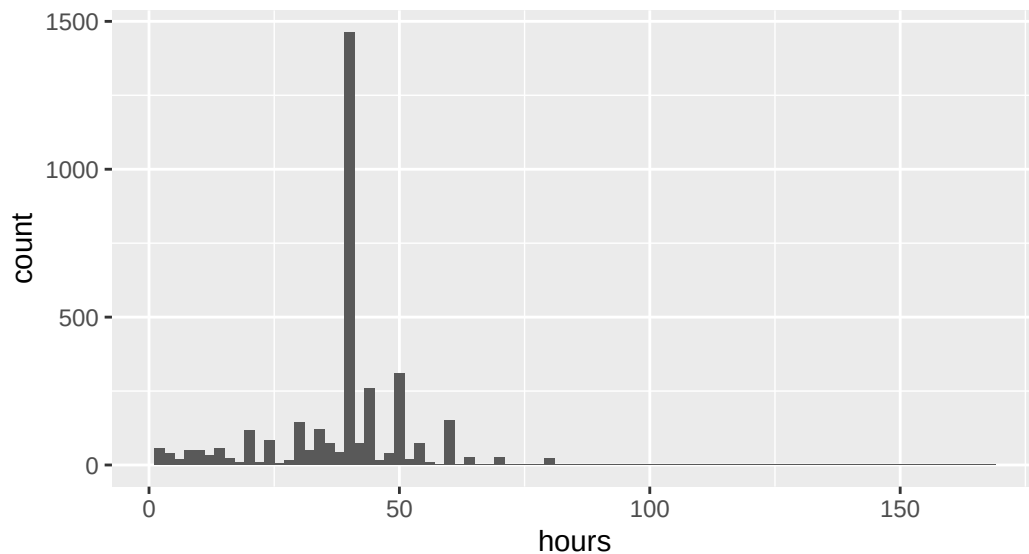
```
library(tidyverse)  
hours <- read_csv("../data/anes2024_hours.csv", show_col_types = FALSE) |>  
  drop_na()  
h <- hours$hours
```

(a)

```
length(h)
```

```
[1] 3505
```

```
ggplot(hours, aes(x = hours)) +
  geom_histogram(binwidth = 2)
```



```
sort(h, decreasing = TRUE)[1:20]
```

```
[1] 168 168 168 160 152 150 120 120 112 110 110 100 100 90 90 90 88 86 85
[20] 84
```

3,505 respondents gave a number. The distribution has a tall spike at 40: 1,459 respondents, 42%, answered exactly 40. Smaller spikes sit at other round numbers: 20, 25, 30, 35, 45, 50, 55 and 60. Apart from the spike, the mass is spread from a few hours to about 60. A long right tail runs out to 168.

A week has $7 \times 24 = 168$ hours, so the 3 answers of 168 are impossible: those respondents would never sleep. Answers of 150, 152 and 160 are just as hard to believe. Anything above about 100 hours, more than 14 hours every day of the week, is implausible for a *usual* week. These answers are probably misunderstandings or typing errors. Some may also be jokes.

(b)

```
n    <- length(h)
xbar <- mean(h)
s    <- sd(h)
se   <- s / sqrt(n)
moe  <- qt(.975, df = n - 1) * se
c(n = n, xbar = xbar, s = s, se = se, moe = moe)
```

n	xbar	s	se	moe
3505.0000000	39.0773181	14.7192529	0.2486233	0.4874610

```
xbar + c(-1, 1) * moe
```

```
[1] 38.58986 39.56478
```

$\bar{x} = 39.08$, $s = 14.72$ and $\widehat{SE} = s/\sqrt{n} = 0.249$. The critical value is $t_{3504,025} = 1.961$, so the margin

of error is 0.49 hours. The 95% confidence interval is [38.59, 39.56].

(c)

We are 95% confident that the mean usual work week of U.S. citizens of voting age who worked in the past 12 months lies between 38.59 and 39.56 hours. The 95% describes the procedure, not this one interval. If we drew many samples and built an interval from each one, about 95% of those intervals would contain μ . This interval either contains μ or it does not.

(d)

```
t_obs <- (xbar - 40) / se
c(t_crit = qt(.975, df = n - 1), t_obs = t_obs,
  p = 2 * pt(-abs(t_obs), df = n - 1))
```

t_crit	t_obs	p
1.9606412325	-3.7111647725	0.0002095127

```
t.test(h, mu = 40)
```

One Sample t-test

```
data: h
t = -3.7112, df = 3504, p-value = 0.0002095
alternative hypothesis: true mean is not equal to 40
95 percent confidence interval:
 38.58986 39.56478
sample estimates:
mean of x
 39.07732
```

- $H_0: \mu = 40$ against $H_1: \mu \neq 40$.
- With $\alpha = 0.05$, the critical value is $t_{3504, 0.025} = 1.961$.
- $t_{\text{obs}} = (39.08 - 40)/0.249 = -3.71$.
- $p = 2 \times F_{t_{3504}}(-|t_{\text{obs}}|) = 0.0002$.
- $|t_{\text{obs}}| > 1.961$ and $p < 0.05$, so we reject H_0 . The mean usual work week is statistically significantly below 40 hours, by about one hour.
- `t.test()` reports the same t , p -value and confidence interval.

(e)

```
t_39 <- (xbar - 39) / se
c(t_obs = t_39, p = 2 * pt(-abs(t_39), df = n - 1))
```

t_obs	p
0.3109851	0.7558305

- $H_0: \mu = 39$ against $H_1: \mu \neq 39$.
- The critical value is again 1.961.
- $t_{\text{obs}} = 0.31$.
- $p = 0.76$.
- $p > 0.05$, so we fail to reject H_0 . The data are consistent with a mean work week of 39 hours.

The decisions match the interval. 40 lies outside $[38.59, 39.56]$ and the test rejects $\mu = 40$. 39 lies inside and the test does not reject $\mu = 39$. The general rule: a two-sided test at level α rejects $H_0: \mu = \mu_0$ exactly when μ_0 lies outside the $1 - \alpha$ confidence interval. Both use the same $\bar{x}$, the same $\widehat{SE}$ and the same critical value.

(f)

```
x <- c(52, 10, 32, 42, 10, 25, 80, 45, 50, 38, 50, 48, 24, 30, 40)
se_x <- sd(x) / sqrt(15)
c(xbar = mean(x), s = sd(x), se = se_x)
```

```
      xbar      s      se
38.40000 17.82374  4.60207
```

```
mean(x) + c(-1, 1) * qt(.975, df = 14) * se_x # t interval
```

```
[1] 28.52954 48.27046
```

```
mean(x) + c(-1, 1) * 1.96 * se_x # 1.96 interval
```

```
[1] 29.37994 47.42006
```

```
t.test(x, mu = 40)
```

One Sample t-test

```
data: x
t = -0.34767, df = 14, p-value = 0.7333
alternative hypothesis: true mean is not equal to 40
95 percent confidence interval:
 28.52954 48.27046
sample estimates:
mean of x
 38.4
```

$\bar{x} = 38.4$, $s = 17.82$ and $\widehat{SE} = 4.60$. The t interval is $[28.53, 48.27]$ and the 1.96 interval is $[29.38, 47.42]$. The t interval is wider. Its critical value, $t_{14,0.025} = 2.145$, is above 1.96. With 15 observations s is a noisy estimate of σ , and the fatter tails of t_{14} account for that extra uncertainty.

$t_{\text{obs}} = (38.4 - 40)/4.60 = -0.35$ and $p = 0.73$. We fail to reject H_0 , and 40 lies inside the interval. The gap from 40 is larger here, 1.6 hours against 0.9, but the standard error is 18.5 times as large. The standard error shrinks with $\sqrt{n}$. With 3,505 respondents, the 0.9-hour gap is 3.7 standard errors. With 15 respondents, the 1.6-hour gap is 0.35 standard errors.

Question 4: Paired or independent?

```
library(tidyverse)
# Path is relative to this file's folder in the course repository;
# adjust to wherever the CSV lives in your own project.
ft <- read_csv("../data/anes2024_ft.csv", show_col_types = FALSE) |>
```

```
drop_na()
```

(a)

```
nrow(ft)
```

```
[1] 5496
```

```
cor(ft$harris, ft$trump)
```

```
[1] -0.8161795
```

5,496 respondents rated both candidates. The correlation is -0.82.

(b)

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n (H_i - T_i) = \frac{1}{n} \sum_{i=1}^n H_i - \frac{1}{n} \sum_{i=1}^n T_i = \bar{H} - \bar{T}.$$

```
mean(ft$harris - ft$trump)
```

```
[1] 7.334607
```

```
mean(ft$harris) - mean(ft$trump)
```

```
[1] 7.334607
```

The mean of the differences equals the difference of the means. Computed on the same respondents, the paired test and the two-sample test have the same point estimate. The two tests can differ only in the standard error.

(c)

i. $\mathbb{V}[\bar{D}] = \mathbb{V}[\bar{H}] + \mathbb{V}[\bar{T}] - 2\text{Cov}(\bar{H}, \bar{T}).$

ii.

$$\mathbb{V}[\bar{D}] = \frac{\sigma_H^2 + \sigma_T^2 - 2\text{Cov}(H_i, T_i)}{n}.$$

iii.

$$\mathbb{V}[\bar{H} - \bar{T}] = \frac{\sigma_H^2}{n_H} + \frac{\sigma_T^2}{n_T}.$$

The two-sample test treats $\bar{H}$ and $\bar{T}$ as independent, so their covariance is 0. Compared with (ii), the formula leaves out the covariance term, $-2\text{Cov}(H_i, T_i)/n$.

(d)

$H_0: \mu_D = 0$ against $H_1: \mu_D \neq 0$, where μ_D is the population mean of $D_i = H_i - T_i$.

```
paired <- t.test(ft$harris, ft$trump, paired = TRUE)
paired
```

Paired t-test

```
data: ft$harris and ft$trump
t = 7.4414, df = 5495, p-value = 1.15e-13
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 5.402353 9.266861
sample estimates:
mean difference
 7.334607
```

```
paired$stderr
```

```
[1] 0.9856446
```

	Two-sample (lecture)	Paired
Estimated difference	7.32	7.33
Standard error	0.73 (= 7.32 / 10.01)	0.99
t	10.01	7.44
95% CI	[5.88, 8.75]	[5.40, 9.27]

The estimated difference barely moves. (b) shows that the two estimates are identical on the same respondents. The small change comes from dropping the few respondents who rated only one candidate. The standard error rises by about a third. The t -statistic falls and the confidence interval widens to match.

The decision does not change. The p -value is far below 0.05, so we reject H_0 : Harris's mean rating is statistically significantly higher than Trump's.

(e)

```
n <- nrow(ft)
sh <- sd(ft$harris)
st <- sd(ft$trump)
r <- cor(ft$harris, ft$trump)
se_paired <- sqrt((sh^2 + st^2 - 2 * r * sh * st) / n)
se_indep <- sqrt(sh^2 / n + st^2 / n)
c(paired = se_paired, indep = se_indep, ratio = se_paired / se_indep)
```

```
      paired      indep      ratio
0.9856446 0.7315362 1.3473628
```

The sample covariance is $r s_H s_T$. The paired standard error, 0.986, equals the `stderr` from (d). The two-sample formula gives 0.732. The paired one is 1.35 times larger. The covariance is negative here, so the term $-2\text{Cov}(H_i, T_i)$ is positive and adds variance. The two-sample formula leaves that term out and understates the variability of the mean difference.

(f)

With a correlation of 0.89 the covariance is positive. The term -2Cov then subtracts variance, so the paired standard error is smaller than the two-sample one. A two-sample test would overstate the standard error.

The general rule: the two-sample formula drops the covariance term.

- Positive correlation: the dropped term would have reduced the variance, so ignoring the pairing overstates the standard error.
- Negative correlation: the dropped term would have added variance, so ignoring the pairing understates the standard error.
- Zero correlation: the two formulas agree.

(g)

The Type I error rate. An understated standard error inflates $|t|$, so the test rejects a true H_0 more often than the nominal α . The analyst believes the test rejects a true null 5% of the time, but it does so more often. The Type II error rate moves the other way: a test that rejects more often fails to reject a false H_0 less often.

The conclusion of the hypothesis test does not change. The paired t is 7.44 and the p -value is far below 0.05, so both tests reject H_0 of equal means. In a smaller sample, or with a smaller gap between the candidates, the two tests could reach different conclusions.