Questions

Information for Questions 1–2
We observe data $y = (12.5, 15.2, 14.8, 13.9, 16.1)$.

1. The sample mean $\bar{y}$ is closest to

   A. 13.9
   B. 14.5
   C. 15.2
   D. 12.5
   E. 16.1

2. The sample standard deviation s is closest to

   A. 0.9
   B. 1.1
   C. 1.4
   D. 2.0
   E. 3.0

3. While browsing past marathon results, we notice a finishing time of 2:07:30 in one race. Using only this information, select the best option:

   A. This must be an unusually fast time.
   B. Since it's one race, this time is typical for marathons.
   C. We can't tell if it's unusual without knowing the distribution of marathon times.
   D. It is normal that all elite marathons finish under 2:05.
   E. Every marathon winner finishes at exactly 2:07:30.

4. Are variation and uncertainty important concepts in statistics?

   A. Yes—models describe variation, and we quantify uncertainty wherever possible.
   B. No—statistics finds exact population parameters without error.
   C. Yes—but only to make analysis look sophisticated.
   D. No—big data eliminates uncertainty.
   E. No—once data are collected, uncertainty disappears.

   Information for Questions 5–7
   A discrete variable Y takes values $\{1, 2, 5\}$ with probabilities 0.20, 0.50, 0.30 respectively.

5. $\Pr(Y \geq 2)$ equals
   A. 0.20
   B. 0.30
   C. 0.50
   D. 0.80
   E. 1.00

6. $E[Y]$ is
   A. 2.1
   B. 2.5
   C. 2.7
   D. 3.0
   E. 3.2

7. Which best describes $E[Y]$?
   A. A randomly selected person will have exactly $E[Y]$.
   B. The expected/average value in repeated sampling from this distribution.
   C. Half the population will be below $E[Y]$.
   D. Having exactly $E[Y]$ implies increased risk.
   E. Y is normally distributed with mean $E[Y]$.

Information for Questions 8–11
Let A be "a customer is a subscriber" and B be "they purchase a premium plan". Suppose $\Pr(A) = 0.40$, $\Pr(B \mid A) = 0.35$, $\Pr(B \mid A^c) = 0.10$.

8. Best interpretation of $\Pr(B \mid A^c)$:
   A. Probability a non-subscriber purchases the premium plan.
   B. Probability someone is not a subscriber given they purchased.
   C. Probability a subscriber purchases the premium plan.
   D. Probability someone is not a subscriber and purchases.
   E. Probability of purchasing regardless of subscription.

9. $\Pr(A \mid B)$ is closest to
   A. 0.30
   B. 0.50
   C. 0.60
   D. 0.70

E. 0.90

10. Pr(A or B) is closest to

    A. 0.20
    B. 0.40
    C. 0.46
    D. 0.55
    E. 1.00

11. Pr(B) is closest to

    A. 0.10
    B. 0.20
    C. 0.30
    D. 0.35
    E. 0.50

12. What is the best description of a random variable?

    A. A summary of both population and sample.
    B. A variable that must be normal.
    C. A random process with a numerical outcome.
    D. A fixed but unknown value.
    E. A value chosen by the analyst.

13. Which is best modelled as continuous?

    A. Number of emails received today.
    B. Count of customers who clicked "Buy".
    C. CO_2 concentration in air (ppm).
    D. Number of red cards in a match.
    E. Count of goals in a game.

Information for Questions 14–15
Let X and Y be independent with $E[X] = 8$, $\text{Var}(X) = 1.2$ and $E[Y] = 5$, $\text{Var}(Y) = 0.8$. Define the score $S = 3X - 2Y$.

14. $E[S]$ is

    A. 6
    B. 10
    C. 12
    D. 14

E. 16

15. The standard deviation of S, $sd(S)$, is closest to

A. 1.10
B. 2.24
C. 3.74
D. 4.50
E. 6.00

Information for Questions 16–19
A test score $Y \sim \mathcal{N}(\mu = 70,\ \sigma = 12)$ for healthy adults.

16. A person has $z = -0.75$. Their score is closest to

A. 52.0
B. 58.0
C. 61.0
D. 70.0
E. 79.0

17. Which R command computes $\Pr(Y > 82)$?

A. 1 - pnorm(82)
B. 1 - pnorm(1.0)
C. pnorm(82)
D. pnorm(1.0)
E. 1 - pnorm(-1.0)

18. If $n = 100$ independent adults are sampled, the sampling distribution of $\bar{Y}$ is

A. Normal, mean 70, sd 12
B. Normal, mean 70, sd 1.2
C. Normal, mean 0, sd 1
D. t with 99 df
E. Unknown

19. If the population is not normal but not extremely skewed (still $\mu = 70, \sigma = 12$), for $n = 100$ the sampling distribution of $\bar{Y}$ is approximately

A. Normal, mean 70, sd 12
B. Normal, mean 70, sd 1.2
C. Normal, mean 0, sd 1

D. t with 99 df

E. Unknown

Information for Questions 20–22
A researcher measures a continuous outcome with unknown σ. Sample size $n = 25$. Data look roughly normal.

20. A 95% CI for μ should be

A. $\hat{p} \pm z_{0.975}\sqrt{\hat{p}(1-\hat{p})/n}$

B. $\bar{y} \pm z_{0.975}\,\sigma/\sqrt{n}$

C. $\mu \pm z_{0.975}\,\sigma/\sqrt{n}$

D. $\bar{y} \pm z_{0.975}\,s/\sqrt{n}$

E. $\bar{y} \pm t_{24,0.975}\,s/\sqrt{n}$

21. If we raise confidence from 95% to 99%, what changes?

A. Estimate changes; unclear effect on width

B. Standard error changes; interval wider

C. Standard error changes; interval narrower

D. Multiplier changes; interval wider

E. Multiplier changes; interval narrower

22. A 99% CI for μ is (3.54, 3.82). Best interpretation:

A. 99% of individual observations lie in $(3.54, 3.82)$.

B. A random future observation has 99% chance to lie in $(3.54, 3.82)$.

C. There is probability 0.99 that the fixed μ is in $(3.54, 3.82)$.

D. 99% of sample means fall in $(3.54, 3.82)$.

E. We are 99% confident the true mean lies in $(3.54, 3.82)$.

23. For $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$, with $\alpha = 0.01$ and $p = 10^{-6}$:

A. Very small p means a very large effect size must exist.

B. We can be certain $\mu \neq \mu_0$.

C. Data would be very unlikely if H_0 were true.

D. $\Pr(\mu = \mu_0) = 10^{-6}$.

E. Pr(our calculations are correct) $= 10^{-6}$.

24. Which best describes estimation?

A. Assume population mean equals sample mean.

B. Compute SE as $s/\sqrt{n}$.

C. Use population data to compute a statistic.

D. Guess data values to improve model fit.
E. Use a statistic to make an informed guess about an unknown parameter.

Information for Questions 25–26
The R object `otter` has variables `weight` (kg) and `length` (cm).

25. What does `mean(otter$weight)` compute?
    A. Sample mean of length
    B. Sample sd of length
    C. Sample mean of weight
    D. Sample sd of weight
    E. Median weight

26. What does `sd(otter$length)` compute?
    A. Sample mean of length
    B. Sample sd of length
    C. Sample mean of weight
    D. Sample sd of weight
    E. Median length

Information for Questions 27–30 (Paired data)
Thirty patients' stress hormone levels are recorded before and after mindfulness training. R output:

```
t = -2.57, df = 29, p-value = 0.012
95 percent CI for mean difference (after - before):  (-1.15, -0.15)
mean difference = -0.65
```

27. With $\alpha = 0.05$, best interpretation:
    A. Since $p < \alpha$, the data would be unusual if there were truly no mean difference.
    B. Since $p < \alpha$, the true after-level is 0.
    C. Since $p < \alpha$, the data would not be unusual under no difference.
    D. Since $p > \alpha$, there's no change.
    E. Therefore mindfulness should be mandatory.

28. Which model is used?
    A. Normal model with paired data
    B. Normal model with two independent groups
    C. Chi-squared test

D. Linear regression
E. Binomial model

29. In this context, power is

A. $\Pr(H_0 \text{ true})$
B. $\Pr(\text{reject } H_0 \mid H_0 \text{ true})$
C. $\Pr(\text{reject } H_0 \mid H_A \text{ true})$
D. Same as the p-value
E. $\Pr(p < 0.05)$ without context

30. Which definitely increases power?

A. Setting `paired = TRUE` in any test
B. Increasing α and decreasing n
C. Decreasing α
D. Repeating identical tests without more data
E. Increasing the sample size

Information for Questions 31–33 (Two independent groups)
Two drugs are compared on blood pressure reduction. Group 1 (Drug A): $\bar{y}_1 = 23.4$, $s_1 = 5.8$, $n_1 = 45$. Group 2 (Drug B): $\bar{y}_2 = 20.1$, $s_2 = 6.2$, $n_2 = 40$.

31. The estimated standard error of $\bar{y}_1 - \bar{y}_2$ is closest to

A. 0.95
B. 1.10
C. 1.31
D. 1.60
E. 2.00

32. A two-sided test of equal means gives $t \approx 2.52$ with $p \approx 0.013$. At $\alpha = 0.05$, we should

A. Fail to reject; no evidence of a difference.
B. Reject; evidence of a difference.
C. Conclude Drug B is superior.
D. Conclude both drugs are identical in effect.
E. Need a paired test instead.

33. A plausible 95% CI for $\mu_1 - \mu_2$ is

A. $(-5.9, \ -0.7)$

B. $(-0.7,\ 5.9)$

C. $(0.7,\ 5.9)$

D. $(1.3,\ 6.5)$

E. $(3.3,\ 6.0)$

Information for Questions 34–36 (Proportions)
In $n = 150$ visitors, $x = 45$ purchased.

34. A 95% CI for the proportion p should be

    A. $\bar{y} \pm t\, s/\sqrt{n}$

    B. $\hat{p} \pm z\, \sqrt{\hat{p}(1-\hat{p})/n}$

    C. $\hat{p} \pm t\, \sqrt{\hat{p}(1-\hat{p})/n}$

    D. $\mu \pm z\, \sigma/\sqrt{n}$

    E. Not possible without σ

35. If we increase n from 150 to 600 (same $\hat{p}$), the CI width

    A. Doubles

    B. Stays the same

    C. Halves

    D. Becomes four times wider

    E. Is unpredictable

36. A 90% CI for p is $(0.22,\ 0.36)$. Best interpretation:

    A. 90% of individuals have outcomes in this range.

    B. There's a 90% chance the random future sample proportion lies in this range.

    C. We are 90% confident the true population proportion lies between 0.22 and 0.36.

    D. 90% of possible samples will produce $\hat{p}$ equal to 0.29.

    E. p varies from study to study.

Normal tools in R

37. Which gives $\Pr(Z > 1.2)$ for $Z \sim \mathcal{N}(0, 1)$?

    A. pnorm(1.2)

    B. pnorm(-1.2)

    C. 1 - pnorm(1.2)

    D. 1 - pnorm(-1.2)

E. pnorm(0)

38. Which gives the 97.5th percentile of $Z \sim \mathcal{N}(0,1)$?

A. qnorm(0.025)
B. qnorm(0.975)
C. pnorm(0.975)
D. 1 - qnorm(0.975)
E. qnorm(1.975)

39. Which best describes a random sample?

A. Every unit is selected with equal probability and independently (or via a valid random design).
B. The most convenient units are selected.
C. Units chosen until results look "about right".
D. Units with the largest values are over-sampled deliberately (without model).
E. Volunteers only.

40. Which action reduces the standard error of $\bar{Y}$?

A. Halving n
B. Doubling n
C. Quadrupling n
D. Doubling s
E. None of the above

41. Which is a parameter, not a statistic?

A. Sample mean sodium = 3.4 mg
B. Sample sd = 1.1 mg
C. Population mean sodium μ
D. $\bar{y}$ from a study of 40 people
E. The sample median

42. Suppose $\sigma = 8$, $n = 64$, $\bar{y} = 102$, $H_0 : \mu = 100$. The two-sided p-value equals

A. pnorm(2)
B. $2 * (1 - \text{pnorm}(2))$
C. $1 - \text{pnorm}(2)$
D. pnorm(-2)

E. $2 * \text{pnorm}(2)$

43. When should a paired t-test be preferred over a two-sample t-test?

A. When two groups are independent.

B. When measurements are "before–after" on the same units.

C. When one group is twice as large.

D. When variances are equal.

E. When sample sizes are both ≥ 30.

44. Even if the population is not normal, the distribution of $\bar{Y}$ is approximately normal when

A. n is large (and the population isn't extremely skewed).

B. $s = 0$.

C. $n = 2$.

D. We take medians instead.

E. Only if the population is uniform.

45. For $Y \sim \mathcal{N}(100, 15^2)$, which R command gives the 90th percentile?

A. qnorm(0.1, 100, 15)

B. pnorm(0.9, 100, 15)

C. qnorm(0.9, mean=100, sd=15)

D. 1 - qnorm(0.9, 100, 15)

E. qnorm(1.9, 100, 15)

46. Which statement about the sampling distribution of $\bar{Y}$ is correct?

A. Mean is μ, standard error is $\sigma\sqrt{n}$.

B. Mean is $\bar{y}$, standard error is $s/\sqrt{n}$.

C. Mean is μ, standard error is $\sigma/\sqrt{n}$.

D. Mean is 0, standard error is 1.

E. It is never approximately normal.

Solutions

1. B. 14.5. $\bar{y} = (12.5 + 15.2 + 14.8 + 13.9 + 16.1)/5 = 72.5/5 = 14.5$.
2. C. 1.4. Compute s (unbiased, $n-1$ in denominator); $s \approx 1.37$, closest is 1.4.
3. C. We need the distribution (range/variability) of times to judge "unusual".
4. A. Models describe variability; we quantify uncertainty (confidence, p-values, etc.).
5. D. 0.50+0.30=0.80.
6. C. $E[Y] = 1(0.20) + 2(0.50) + 5(0.30) = 0.2 + 1.0 + 1.5 = 2.7$.
7. B. It's the long-run average under the distribution.
8. A. Probability a non-subscriber purchases the premium plan.
9. D. $\Pr(A \mid B) = \dfrac{0.35 \cdot 0.40}{0.35 \cdot 0.40 + 0.10 \cdot 0.60} = 0.14/0.20 = 0.70$.
10. C. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = 0.40 + 0.20 - 0.14 = 0.46$.
11. B. $\Pr(B) = 0.35(0.40) + 0.10(0.60) = 0.14 + 0.06 = 0.20$.
12. C. A random process with a numerical outcome.
13. C. CO_2 concentration is continuous (in practice, measured on a continuum).
14. D. $E[3X - 2Y] = 3(8) - 2(5) = 24 - 10 = 14$.
15. C. $\mathrm{Var}(3X - 2Y) = 9(1.2) + 4(0.8) = 10.8 + 3.2 = 14$; $sd = \sqrt{14} \approx 3.74$.
16. C. $y = \mu + z\sigma = 70 + (-0.75) \cdot 12 = 70 - 9 = 61$.
17. B. $z = (82 - 70)/12 = 1.0$; $\Pr(Y > 82) = 1 - \Phi(1.0) = 1 - \mathrm{pnorm}(1.0)$.
18. B. $\bar{Y} \sim \mathcal{N}(70,\ (12/\sqrt{100})^2) \Rightarrow sd = 1.2$.
19. B. By CLT, approximately normal with mean 70, sd $12/\sqrt{100} = 1.2$.
20. E. With unknown σ, smallish $n = 25$, use t: $\bar{y} \pm t_{24,0.975}\, s/\sqrt{n}$.
21. D. Only the multiplier changes (larger), making the interval wider.
22. E. Confidence refers to the true mean lying in the interval.
23. C. A tiny p-value indicates data are very unlikely under H_0.
24. E. Estimation uses a statistic (e.g., $\bar{y}$) to infer a parameter (e.g., μ).

25. C. The sample mean of weight.

26. B. The sample standard deviation of length.

27. A. Since $p = 0.012 < 0.05$, data would be unusual if there were truly no difference.

28. A. Paired t-test (normal model for paired differences).

29. C. Power = Pr(reject $H_0 \mid H_A$ true).

30. E. Increasing sample size increases power (all else equal).

31. C. SE $= \sqrt{s_1^2/n_1 + s_2^2/n_2} = \sqrt{5.8^2/45 + 6.2^2/40} \approx 1.31$.

32. B. With $p \approx 0.013 < 0.05$, reject H_0; there is evidence of a difference.

33. C. 95% CI $\approx 3.3 \pm 1.96 \times 1.31 \approx (0.7, 5.9)$.

34. B. For a single proportion: $\hat{p} \pm z\sqrt{\hat{p}(1-\hat{p})/n}$.

35. C. Width scales with $1/\sqrt{n}$. $n \times 4 \Rightarrow$ width $\times 1/2$.

36. C. Confidence speaks to the population proportion.

37. C. $\Pr(Z > 1.2) = 1 - \Phi(1.2) = 1 - \text{pnorm}(1.2)$.

38. B. 97.5th percentile is qnorm(0.975).

39. A. Random sampling requires a legitimate random mechanism (independence/equal chance).

40. C. Quadrupling n halves $s/\sqrt{n}$.

41. C. μ is a parameter (population quantity).

42. B. $z = (102 - 100)/(8/\sqrt{64}) = 2$; two-sided $p = 2(1 - \Phi(2))$.

43. B. Paired test for before–after on the same units.

44. A. CLT: for large n, $\bar{Y}$ approximately normal (non-extreme skewness).

45. C. qnorm(0.9, mean=100, sd=15).

46. C. Mean μ, standard error $\sigma/\sqrt{n}$.