Statistics as a Process for Inference

Name: ____________________	Date: ______________	Score: _____ / 33

Quick Review

Statistics is a way of *reasoning under uncertainty*. You can't measure everyone, so you measure some – a **sample** – and use what you find to say something about the whole group, the **population**. That step from sample to population is called **statistical inference**.

Population vs. sample. The *population* is the entire collection of objects, people, or measurements you care about (all U.S. voters, every light bulb a factory makes this year, every tree in a forest). The *sample* is the subset you actually measure.

Parameter vs. statistic. A *parameter* is a number describing the population: μ (population mean), σ (population std dev), p (population proportion). A *statistic* is a number computed from the sample: $\bar{x}$, s, $\hat{p}$. Statistics are what you observe and compute; parameters are what you want to know but usually can't measure directly.

Sampling variability. Different samples give different statistics, even from the same population. This variability is *not* a flaw – it's the heart of inference. Knowing *how much* statistics typically vary from the parameter lets us build confidence intervals and run hypothesis tests.

The inference workflow. (1) Define the population and the question. (2) Plan how to collect data (random sampling, random assignment, sample size). (3) Collect the data. (4) Summarize with statistics. (5) Use probability to quantify the uncertainty and draw a conclusion. (6) Communicate honestly, including the limitations.

Common slips. Mixing up parameter and statistic ("$\bar{x} = 72$" is a statistic, "$\mu = 72$" is a parameter). Treating the sample as if it were the population ("everyone in our class said yes" $\neq$ "everyone in the school says yes"). Forgetting that good inference requires a *random* sample – a convenience sample (your friends, the people who happened to be at the mall) brings in bias that no formula can fix.

PRACTICE

Match each term to its definition; identify parameter vs. statistic.

1. What is a population? ________

2. What is a sample? ________

3. In the table, the value $\bar{x} = 72$ comes from a sample of 30. Is it a parameter or a statistic? ________

Symbol	Value	Source
$\bar{x}$	72	sample of 30
μ	unknown	whole population

4. In the same table, μ describes the whole population. Is it a parameter or a statistic? ________

Symbol	Value	Source
$\bar{x}$	72	sample of 30
μ	unknown	whole population

5. What is the goal of statistical inference? ________

6. Does inference require eliminating all variability? ________

7. Why is a random sample preferred over a convenience sample? ________

8. Is $\hat{p} = 0.62$ from a poll of 1000 voters a parameter or a statistic? ________

9. True or false: two random samples from the same population always give the same statistic. ________

Scan Me

10. Using the table, explain the difference between σ and s. ________

Symbol	Describes	Type
σ	population	parameter
s	sample	statistic

11. A descriptive summary of sample data is called ________
12. True or false: a population must always be a group of people. ________
13. What does the law of large numbers say about $\bar{x}$ as $n \to \infty$? ________
14. Sampling variability shrinks as n ________
15. Is "a sample of 1 person" useful for inference about a population? ________
16. True or false: a statistic is exactly equal to the parameter. ________
17. In inference, the sample size affects the ________
18. True or false: if everyone in your random sample says "yes," you can conclude everyone in the population says "yes." ________
19. What's a *census*? ________
20. Why use a sample at all instead of a census? ________

♦ Word Problems

21. A teacher gives a math test to all $50{,}000$ high-schoolers in her district. She computes the mean to be 74 and the standard deviation to be 11. Are these numbers parameters or statistics? Explain. ____

22. A pharmaceutical company tests a new drug on 400 patients sampled at random from the eligible population. The sample proportion who improved is $\hat{p} = 0.72$. Identify the parameter the company wants to learn about, and explain why $\hat{p}$ is only an estimate. ____

23. A small-town blogger surveys her 200 Twitter followers about a national policy and reports the results as "representative of America." What's the major flaw in her inference, and what would fix it? ____

24. A quality engineer tests 50 randomly selected widgets and finds 3 defective. She reports a sample proportion of $\hat{p} = 0.06$. Explain what the next steps of inference would do with this number. ____

Additional Practice

25. Find the mean of $4, 6, 8, 10$. ________
26. Find the median of $3, 9, 4, 10, 7$. ________
27. Find the range of $12, 5, 9, 20$. ________
28. Find the mode of $2, 3, 3, 5, 8$. ________
29. Find z for $x = 72$, mean 60, standard deviation 6. ________
30. Interpret $z = -1.5$. ________
31. Predicted y for $\hat{y} = 2x + 5$ at $x = 6$. ________
32. Residual if actual $= 20$ and predicted $= 17$. ________
33. Positive association: slope sign? ________

Scan Me

Answer Keys

1. all individuals/items of interest
2. a subset of the population
3. statistic
4. parameter
5. use sample to say something about population
6. No – it quantifies variability with probability
7. random samples avoid systematic bias
8. statistic
9. False
10. σ is population SD; s is sample SD
11. descriptive statistics
12. False
13. $\bar{x} \to \mu$
14. increases
15. Generally no
16. False
17. precision (standard error)
18. False
19. measuring the whole population
20. cheaper, faster, sometimes the only option
21. parameters (she measured the whole population)
22. parameter p; $\hat{p}$ varies
23. biased convenience sample; use random sample
24. build a CI for p (defect rate)

Additional Practice Answers

25. 7
26. 7
27. 15
28. 3
29. 2
30. 1.5 SD below mean
31. 17
32. 3
33. positive

Additional Practice: Answers for all numbered items, including the added practice, are shown in the grid above.

Step-by-Step Explanations

1. The full group you want to study – could be people, bulbs, trees, anything. Not just the ones you measure.
2. The piece of the population you actually measure. Inference uses the sample to talk about the population.
3. It's computed from the *sample*, so it's a statistic. The unknown population mean μ in the next row is the parameter.
4. It describes the population, so it's a parameter. Usually its exact value is unknown – which is why we take samples and compute statistics like $\bar{x}$.
5. Inference is the bridge from what you measured (the sample) to what you want to know (the population).
6. Variability is built into sampling. Inference uses probability to *describe* that variability, not eliminate it.
7. Random selection gives every member of the population a fair shot, which lets the math of inference work. Convenience samples bake in bias that no formula can correct.
8. Start with the key idea: $\hat{p}$ is the sample proportion – a statistic. The parameter p is the (unknown) true proportion in the whole voting population. That gives a quick check on the answer.
9. Different samples give different statistics. That variability is what standard error measures.
10. Greek letters (σ) name population parameters; Roman letters (s) name sample statistics. Same idea – spread – but computed from different groups.
11. Means, medians, ranges, histograms – numbers and pictures that describe what you have in hand. Inference comes *after* descriptive statistics.
12. Populations can be bulbs, trees, transactions, atoms – any well-defined group of items.
13. As the sample grows, the sample mean converges to the true population mean. Bigger samples are more reliable – the LLN is the formal version of that intuition.
14. Standard error $= \sigma/\sqrt{n}$ shrinks as n grows. Larger samples mean less sample-to-sample variability in $\bar{x}$.
15. One data point gives no sense of variability, and one person's value can be wildly different from the population mean. The math of inference assumes the sample reflects the population's distribution.
16. Statistics estimate parameters but generally differ from them due to sampling variability. The hope is they're close.
17. A careful way to see it: Bigger $n \Rightarrow$ smaller standard error $\Rightarrow$ tighter intervals $\Rightarrow$ more precise estimates. That gives a quick check on the answer.
18. Even with a unanimous sample, the population almost certainly has dissent. Inference gives confidence intervals – never absolute certainty.
19. A census measures every member of the population. No sampling, no inference – you just have the parameter directly. Usually too expensive or impossible.
20. Measuring everyone is usually impractical (cost, time, physical impossibility). A well-designed sample gives reliable inference at a fraction of the effort.
21. Because the teacher measured every high-schooler in her district – the entire population – the mean and standard deviation she computed are *parameters*, denoted $\mu = 74$ and $\sigma = 11$. If she had instead sampled, say, 500 students, the same numbers would be *statistics* (denoted $\bar{x}$ and s) used to estimate unknown parameters.
22. The parameter of interest is p, the true proportion of eligible patients who would improve on the drug if every one of them received it. Since the company can only test 400 at a time, they use the sample proportion $\hat{p} = 0.72$ as an estimate of p. Another random sample of 400 would give a slightly different $\hat{p}$. The *sampling distribution* of $\hat{p}$ tells us how much $\hat{p}$ typically varies around the true p, which is the basis for the confidence interval the company will report.
23. Her followers are a *convenience sample* – not randomly drawn from the American population. They likely share her geography, interests, and political leanings, so their opinions don't generalize. No clever formula can fix selection bias – the only repair is to use a properly designed *random sample* of the actual population she wants to describe (which, for "America," is a big undertaking).
24. Start with the key idea: $\hat{p} = 0.06$ is the engineer's point estimate of the true defect rate p. The next steps of inference would: (1) compute the standard error $\sqrt{\hat{p}(1-\hat{p})/n}$, (2) use that to build a confidence interval (e.g., a 95% CI of roughly 0.06 ± 0.066), and (3) decide whether the true defect rate is plausibly above some company threshold. The point estimate alone ("6%") hides the uncertainty; the interval makes it visible. That gives a quick check on the answer.