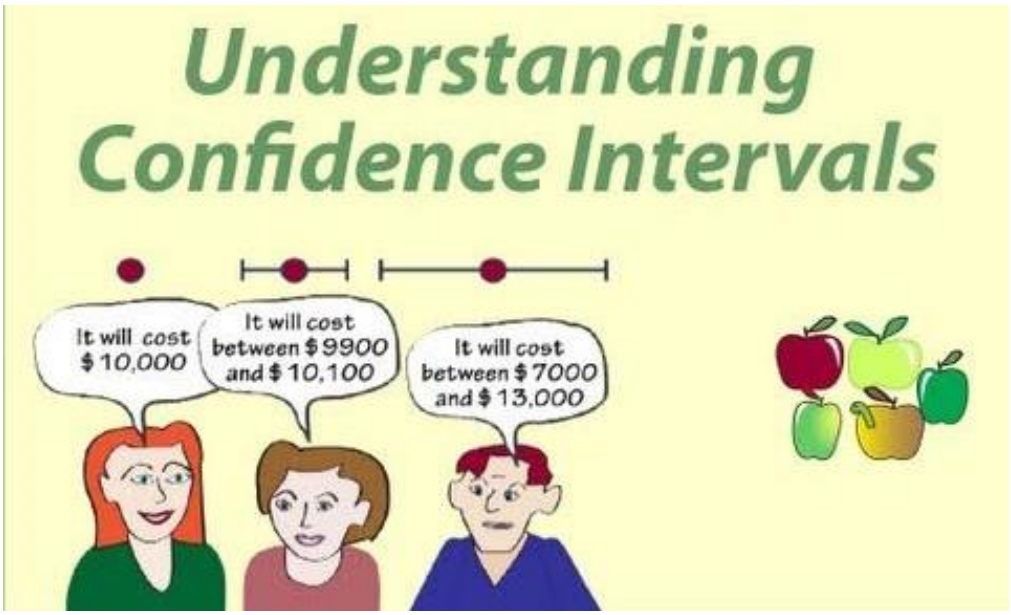
Describing Uncertainty 01

EN5423 | Spring 2024

w04_uncertainty_01.pdf
(Week 4)

Contents



Introduction to Uncertainty

• **Uncertainty and Reliability of Estimates**

- Here, I will introduce the concepts of uncertainty and reliability in sample estimates, emphasizing the use of interval estimates (both parametric and nonparametric) over single point estimates to assess and report on the variability and confidence in data analysis.

• **Interval Estimates for Population Parameters**:

- You will learn the utility of interval estimates for comparing sample statistics to population parameters and for conducting hypothesis tests on the significance of differences from specific values.

• Examples in Environmental Engineering:

(1) Evaluating Mean Nitrate Concentration:
The mean nitrate concentration in a shallow aquifer under agricultural land was calculated as 5.1 milligrams per liter (mg/L). How reliable is this estimate? Is 5.1 mg/L in violation of a health advisory limit of 5 mg/L? Should it be treated differently than another aquifer having a mean concentration of 4.8 mg/L?

(2) Specific Capacity of Wells:
Thirty wells over a five-county area were found to have a mean specific capacity of 1 gallon per minute per foot, and a standard deviation of 7 gallons per minute per foot. A new well was drilled and developed with an acid treatment. The well produced a specific capacity of 15 gallons per minute per foot. To determine whether this increase might be a result of the acid treatment, we wonder how unusual is it to have a well with a specific capacity of 15 gallons per minute per foot given our observations about the distribution of specific capacity values we see in the wells we have sampled?

(3) 100-Year Flood Estimate Reliability:
An estimate of the 100-year flood, the 99th percentile of annual flood peaks, was determined to be 1,000 cubic meters per second (m3/s). Assuming that the choice of a particular distribution to model these floods (log-Pearson Type III) is correct, what is the reliability of this estimate?

1 Definition of Interval Estimates

• Here, we will discuss the concept of estimating central tendency in statistical analysis, contrasting point estimates with interval estimates, and elaborating on their significance, especially in terms of reliability and variability.

• This section sets the groundwork for understanding how interval estimates enhance the interpretation of data, particularly in expressing the certainty of statistical estimates and accommodating data variability.

1.1 Point vs. Interval Estimates

The sample median and mean provide point estimates of a population's ***central tendency***, but they ***do not reflect the estimate's reliability or variability***. This limitation is illustrated by comparing two datasets, X and Y, both with a mean of 5 but differing in variability below:

Example 1

1.2 Reliability of Point Estimates

A point estimate like a mean of 5 does not reveal the underlying variability of the data. For example (**Example 1** above), dataset X, with high variability, offers less certainty in its point estimate compared to dataset Y, which clusters closely around the mean.

1.3 Introduction of Interval Estimates

Interval estimates address this by offering a range with a stated probability (e.g., 95%) of containing the true population value, providing a clearer picture of estimate reliability. Greater data variability necessitates wider intervals to maintain the same level of confidence.

1.4 Types of Interval Estimates

Two main types will be discussed: 1) ***confidence intervals***, which assess the ***likelihood*** of an interval containing the true population value, and 2) ***prediction intervals***, which evaluate the probability of a single new data point belonging to the population.

1.5 Application of Interval Estimates

Confidence and prediction intervals serve distinct purposes and ***cannot be used interchangeably***. The next section outlines plans to explore confidence intervals for medians and means, along with ***parametric*** and ***nonparametric*** prediction intervals, and tolerance intervals for percentiles other than the median.

2 Interpretation of Interval Estimates

This section introduces the concept and application of interval estimates in statistical analysis, emphasizing their utility in providing a more nuanced understanding of data reliability and variability.

2.1 Understanding Interval vs. Point Estimates

• Unlike a point estimate that gives a single value for a parameter (e.g., mean), an interval estimate provides a range within which we expect the true parameter value to lie, along with a confidence level (e.g., 90% or 95%) indicating the reliability of this range. This distinction is crucial as ***it introduces the concept of statistical confidence and acknowledges data variability***.

• Interval estimates capture the uncertainty inherent in sample data, especially important when comparing datasets with different variabilities.

• The probability that the interval ***does include*** the true value is called the ***confidence level***.

• The probability that this interval ***will not cover*** the true value is called the ***significance level***, α, which is computed as:

$$\alpha = 1 - \text{ confidence level} \quad (1)$$

2.2 Simulating Interval Estimates (Table 1; Figure 1)

• A simulation, based on known population parameters (mean $\mu = 5$ mg/L, variance $\sigma^2 = 1$), generates interval estimates for the mean from 12 samples, repeated across 10 independent trials (Although in reality only one set of 12 samples would be taken each year).

• **Figure 1 and Table 1** Demonstrates how the intervals can vary across samples and that while most will contain the true mean, not all will, illustrating the concept of a 90-percent confidence level in practical terms.

Table 1. Ten replicate datasets of 12 samples each of chloride concentrations, each with mean = 5 and standard deviation = 1. All units are in milligrams per liter.
[Xbar, sample mean; sd, sample standard deviation; lcl, lower confidence limit for the mean; ucl, upper confidence limit for the mean, where the confidence interval is a 90-percent two-sided interval]

Replicate	xbar	sd	lcl	ucl
1	5.21	1.072	4.66	5.77
2	5.23	0.754	4.84	5.62
3	5.30	1.314	4.62	5.98
4	5.19	1.182	4.58	5.80
5	5.26	0.914	4.79	5.73
6	5.60	0.713	5.23	5.97
7	4.93	0.917	4.45	5.40
8	5.34	0.871	4.89	5.79
9	4.91	1.132	4.32	5.50
10	5.32	1.098	4.75	5.89

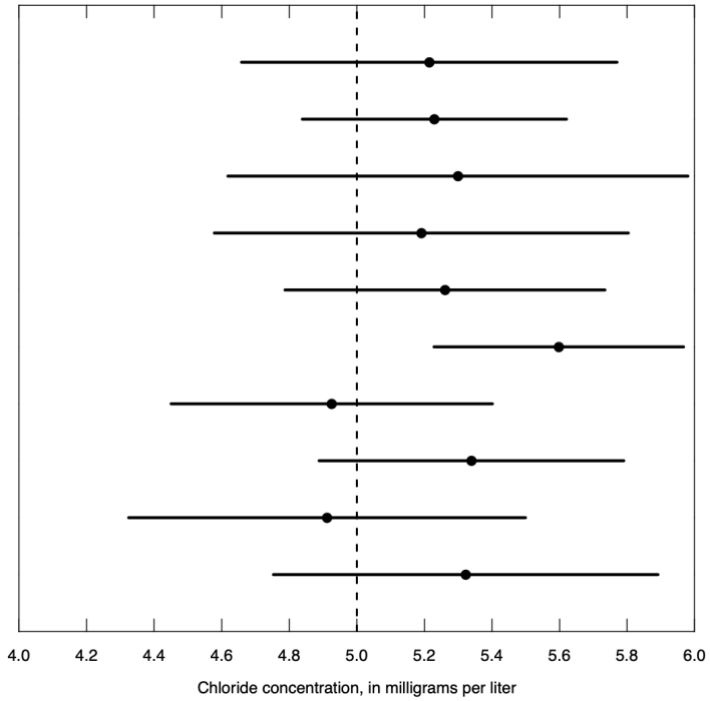


Figure 1. Ten 90-percent confidence intervals for normally distributed data with true mean = 5 and standard deviation = 1, in milligrams per liter. Dots indicate the sample mean from each sample.

2.3 Role of Confidence Levels in Interval Width

• The width of an interval estimate depends on the desired confidence level, sample size, and data variability. Higher confidence levels result in wider intervals, as they increase the probability of encompassing the true population parameter.

• Explores the balance between confidence level and interval precision, showing how adjustments to the confidence level affect the interval's range and our certainty about the parameter estimate.

2.4 Challenges with Skewed Data (Figures 2 & 3)

• The width of an interval estimate depends on the desired confidence level, sample size, and data variability. Higher confidence levels result in wider intervals, as they increase the probability of encompassing the true population parameter.

• Explores the balance between confidence level and interval precision, showing how adjustments to the confidence level affect the interval's range and our certainty about the parameter estimate.

2.5 Challenges with Skewed Data (Figures 2 & 3)

• For skewed datasets or those with small sample sizes, symmetric confidence intervals based on the normal distribution may not accurately reflect the true parameter's location.

• Introduces asymmetric confidence intervals as an alternative for skewed data, highlighting their importance in ensuring accurate and reliable interval estimates.

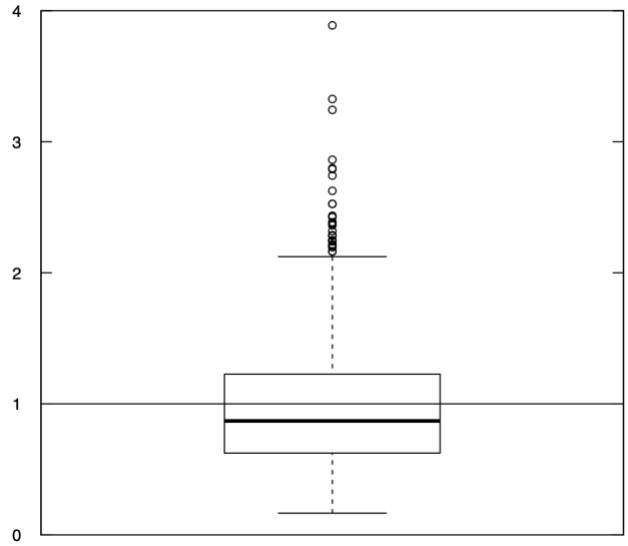


Figure 2. Boxplot of a random sample of 1,000 observations from a lognormal distribution. Population mean = 1, population coefficient of variation = 1. The horizontal line that crosses the entire plot is the true population mean value.

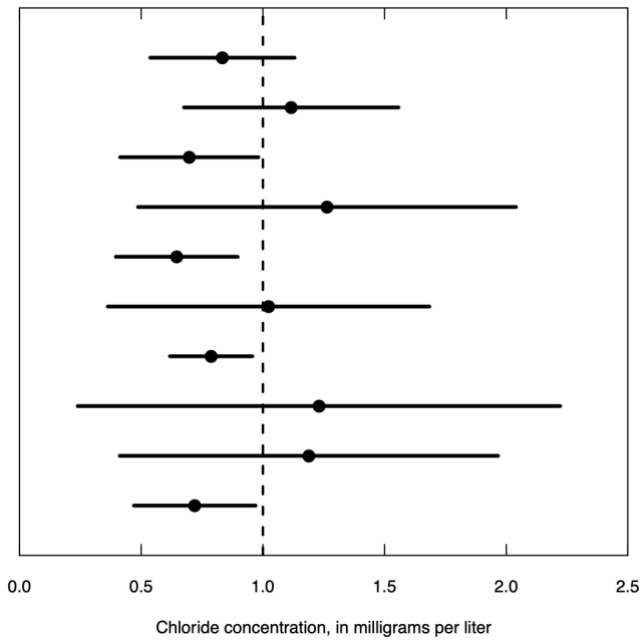


Figure 3. Ten 90-percent confidence intervals around a true mean of 1, each one based on a sample size of 12. Data are from a log normal distribution of mean = 1.0 and coefficient of variation = 1.0. Dots indicate the sample mean values. Four out of the 10 intervals do not include the true value.

2.6 Importance of the Student's t-Distribution (*Example 1*)

• The t-distribution, also known as Student's t-distribution, is a family of probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and the population standard deviation is unknown.

• It resembles the normal distribution but has heavier tails. This means that it is more prone to producing values that fall far from its mean, accounting for the additional uncertainty when estimating a population parameter from a small sample.

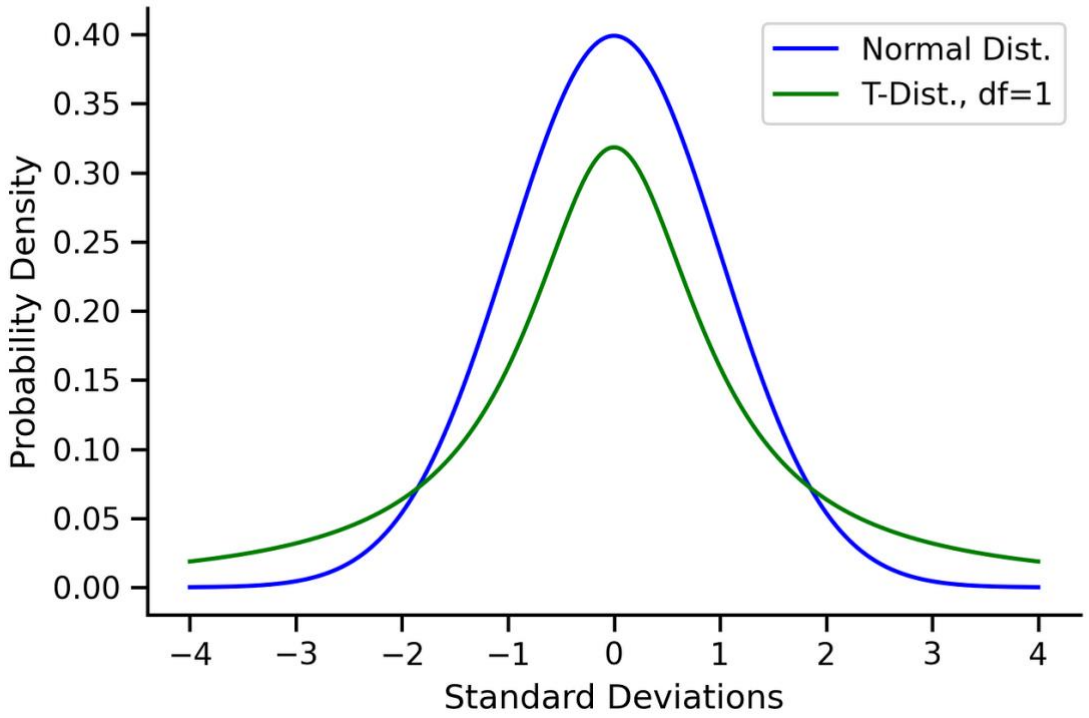


Figure 4. The Normal Distribution vs. Student's t-Distribution

• Calculating Intervals with the t-distribution:

(1) Confidence Intervals: The formula for a confidence interval using the t-distribution takes into account the sample mean, the sample standard deviation, and the critical t-value associated with the desired confidence level and degrees of freedom:

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1} \times \left(\frac{s}{\sqrt{n}}\right) \tag{2}$$

- $\bar{x}$ is the sample mean.
- s is the sample standard deviation.
- n is the sample size.

- $t_{\frac{\alpha}{2},n-1}$ is the critical value from the t-distribution for $n-1$ degrees of freedom and the desired confidence level (e.g., 95%).

• This formulation acknowledges that the precision of the sample mean as an estimator of the population mean depends not only on the sample's variability (as measured by s) but also on the sample size. The critical t-value widens or narrows the confidence interval based on the confidence level and the degrees of freedom, reflecting the estimator's reliability.

• Impact on Estimator Reliability

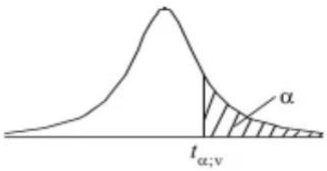
- **Reliability and Sample Size**: As the sample size increases, the degrees of freedom increase, causing the t-distribution to more closely resemble the normal distribution. This reflects reduced uncertainty in the estimator as more data points provide a clearer picture of the population parameter.
- **Variability Reflection**: The use of *n-1* in calculating the sample variance (a component of the confidence interval formula) ensures that the sample's variability is not underestimated, which is crucial for small samples where each data point significantly influences the overall variability.

In summary, the t-distribution, through its dependency on degrees of freedom, provides a nuanced framework for assessing the reliability of statistical estimators derived from ***small samples***. This framework adjusts for sample size and variability, enabling the calculation of confidence intervals that more accurately reflect the underlying uncertainty and estimator's reliability.

Example 2
n = 16, 5% significance level, two-tailed test what is $t_{\frac{\alpha}{2},n-1}$?

Table of the Student's *t*-distribution

The table gives the values of $t_{\alpha;\nu}$ where
$\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

3 Confidence Intervals for the Median

A confidence interval for the true population median may be computed in two ways: (1) without assuming the data follow any specific distribution (nonparametric; section 3.1.), or (2) assuming they follow a distribution such as the lognormal (parametric; section 3.2.).

3.1 Nonparametric Interval Estimate for the Median

• We often estimate the median of a population without making assumptions about the population's distribution. Two common nonparametric methods for this are the binomial distribution method and the bootstrap method.

(1) Binomial Distribution Method (*Example 2*) (*HW04 #1*) (link)

• We use the binomial distribution to answer the following question: How likely is it that the true population median, $c_{0.5}$, would be such that k of the n observed data would be above $c_{0.5}$ and $n-k$ below $c_{0.5}$, where for example, k could be 0, 1, 2, 3,…, 25 out of n = 25?

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x} \quad (3)$$

Where, n = the number of trials (or the number being sampled); x = the number of successes desired; p = probability of getting a success in one trial; q = 1 - p = the probability of getting a failure in one trial.

• **Significance Level (α)**: Before calculating the interval estimate, decide on a significance level. This level represents the risk you are willing to accept that the true median is not captured in your interval estimate. The risk is evenly divided at both ends of the interval ($\frac{\alpha}{2}$).

• **Calculating the Interval**:

- Use the cumulative distribution function (CDF) of the binomial distribution to calculate the confidence interval for the median. In practical applications, tools like the SciPy library in Python to find the critical values corresponding to $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$.

- These critical values help identify the data points that define the upper and lower bounds of the median's confidence interval.

• **Interpreting the Binomial Approach**:

- The binomial approach ***assumes a 50% chance of an observation being above or below*** the true population median. The resulting confidence interval will reflect the shape (skewed or symmetric) of the original data.

- For small sample sizes, achieving the exact desired confidence level might not be possible due to the discrete nature of data points. However, you can still obtain intervals close to the desired confidence level.

Example 2: Nonparametric Interval estimate of the median (Binomial Dist. approach)

• To compute a nonparametric confidence interval for the median in Python, particularly for a dataset with arsenic concentrations, we can use Python's `scipy.stats` for binomial distribution calculations and managing the dataset with Python's built-in functionalities or packages like NumPy.

• Given a dataset of 25 arsenic concentrations from southeastern New Hampshire (**Table 2**; **Figure 5**), our goal is to calculate the 95% confidence interval for the median concentration, using a significance level of $\alpha = 0.05$.

Table 2. Arsenic concentrations (in parts per billion) for groundwaters of southeastern New Hampshire (from Boudette and others, 1985), ranked in ascending order.

Rank	Value	Rank	Value	Rank	Value
1	1.3	10	9.5	19	120
2	1.5	11	12	20	190
3	1.8	12	14	21	240
4	2.6	13	19	22	250
5	2.8	14	23	23	300
6	3.5	15	41	24	340
7	4.0	16	80	25	580
8	4.8	17	100		
9	8.0	18	110		

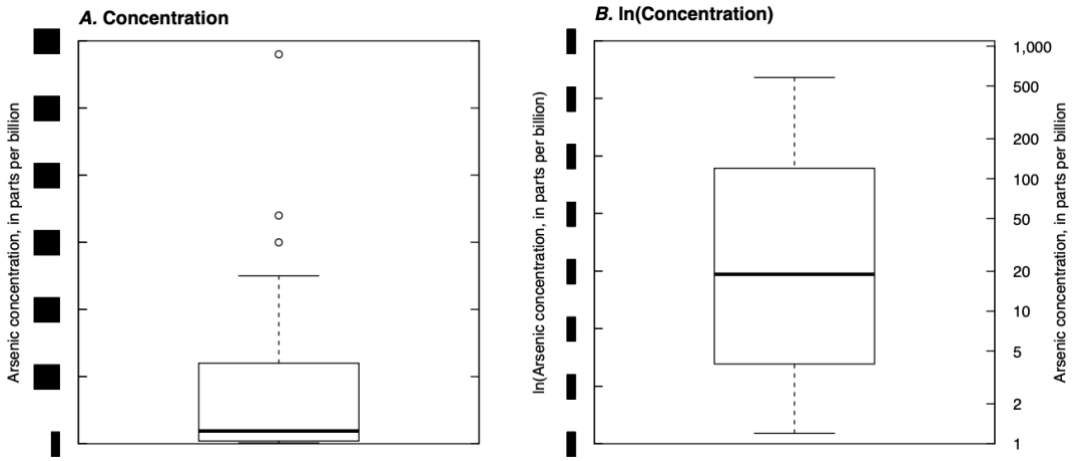


Figure 5. Boxplots of the (A) original and (B) log-transformed arsenic data from Boudette and others (1985)

Step 1:
The sample median $\hat{c}_{0.5} = 19$, which is the 13th observation ranked from smallest to largest in this sample size of 25. The binomial distribution is used to determine the 95-percent confidence interval for the true median concentration, $c_{0.5}$. We obtain the critical values from the `binom.ppf([alpha / 2, 1 - alpha / 2], n, 0.5.` Because we are focused on the median here, the prob value is always 0.5 for this calculation. For the population median, half the population values are above the median and half below. The α value here is 0.05 and thus the end points of the confidence intervals are at $\frac{\alpha}{2}$ and at$1 - \frac{\alpha}{2}$, which are 0.025 and 0.975, respectively. The values returned by this function are 8 and 17,

which are the ranks of the two end points. We can then compute the concentration values that are associated with these two ranks as follows:

```
# Define arsenic concentrations data
arsenic_concentrations = np.array([1.3, 1.5, 1.8, 2.6, 2.8, 3.5, 4.0,
4.8, 8.0, 9.5, 12, 14, 19, 23, 41, 80, 100, 110, 120, 190, 240, 250,
300, 340, 580])

arsenic_concentrations = np.sort(arsenic_concentrations)
# Calculate critical ranks using the binomial distribution
n = len(arsenic_concentrations)
alpha = 0.05  # Significance level

lower_rank, upper_rank = binom.ppf([alpha / 2, 1 - alpha / 2], n,
0.5).astype(int)
print([lower_rank, upper_rank])
>> [8, 17]

arsenic_concentrations[[lower_rank-1,upper_rank-1]]
>> array([  4.8, 100. ])
```

This code indicates that the lower and upper confidence intervals are at ranks 8 and 17, and that these translate to concentration values of 4.8 and 100 (the 8th and 17th values on the sorted list of concentration values in **Table 2**). Because the sample size is relatively small (n =25) we know that the ***interval will not be an exact 95-percent confidence interval***.

Step 2:

We can compute the probability that the interval will contain the true value using the `binom.ppf` function. This function returns the probability density of the binomial distribution, and we can sum the density values from 8 through 17 to determine the prob ability of the true median being in the range of the 8th through 17th values in the sorted vector of values.

```
# Parameters for the binomial distribution
n = 25  # number of samples
p = 0.5  # probability of success

# Calculate the PMF for each value from 8 to 17 and sum them
probability_sum = np.sum(binom.pmf(np.arange(8, 18), n, p))

print("Sum of probabilities from 8 to 17:", probability_sum)
>> Sum of probabilities from 8 to 17: 0.9567147493362426
```

• The result tells us that the true probability for this range is 0.9567, which is very close to the desired probability of 0.95. Thus, one could say the closed interval [4.8, 100] is the best approximation to a 95-percent confidence interval for the median. This means that the probability that $c_{0.5}$ will be less than the 8th ranked sample value is ≤0.025 and similarly the probability that $c_{0.5}$ will be greater than the 17th ranked sample value is also ≤0.025.

• Thus, we can state that a 95-percent confidence interval for the median is [4.8, 100] because 4.8 and 100 are the 8th and 17th ranked values in the sample. The results of these computations

are shown in **Figure 6**; note the substantial amount of asymmetry in the confidence interval, which is what we would expect given the asymmetry of the full sample.

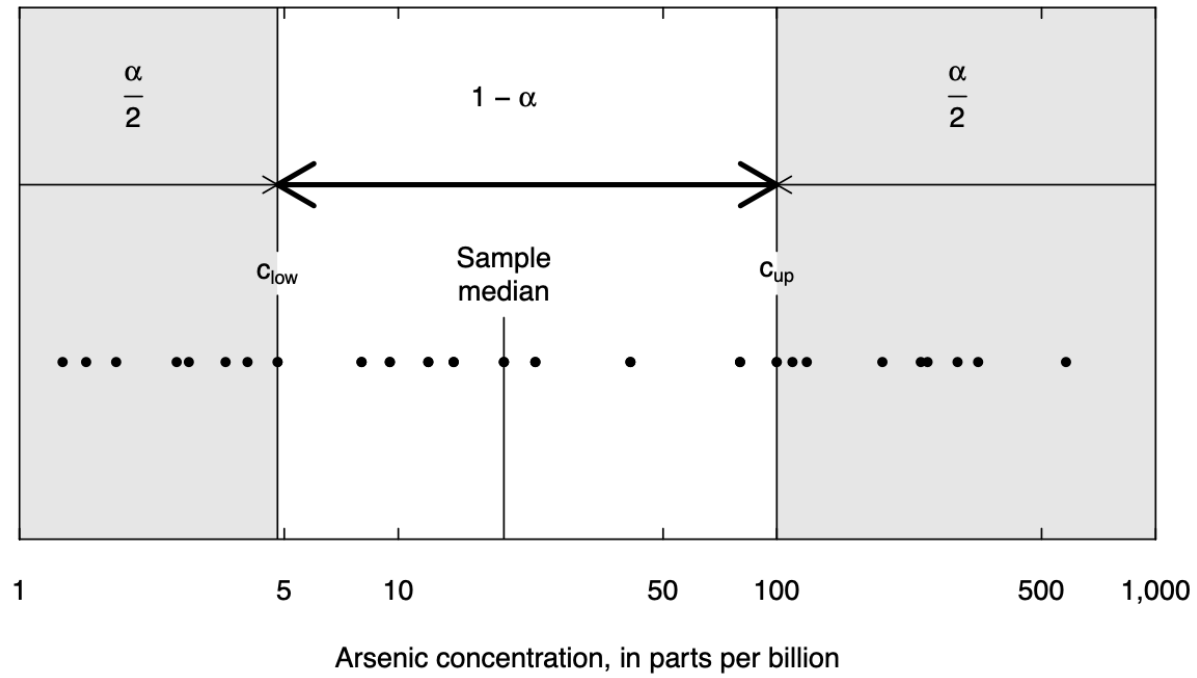


Figure 6. Plot of the 95-percent confidence interval for the true median in example 3.1. The dots represent the 25 observations in the arsenic dataset on a logarithmic scale. The sample median (19 parts per billion [ppb]) and upper (c_{up}) and lower (c_{low}) bounds on the confidence interval for the true median (4.8 and 100 ppb) are shown. The confidence interval is computed using $\alpha = 0.05$. Thus, the probability that the population median will be greater than c_{up} is 0.025 and the probability that the population median is less than c_{low} is also 0.025.

(2) Bootstrapping Method (*Example 3*) (*HW04 #2*) (link)

• For complex issues, bootstrapping might take a lot of time, and deciding on the number of samples involves balancing computation time against precision. In bootstrapping, each resample randomly picks, with replacement, the same number of observations as in the original dataset. This means each data point has the same chance of being included in a resample.

• So, if you start with 25 data points, a bootstrap sample will also have 25 data points, possibly with some repeats or omissions. The variance among the original and resampled datasets illustrates the data's randomness.

• For each bootstrap sample, we calculate and record the statistic of interest, like the median or mean, leading to thousands of such calculated statistics. These form the basis of a confidence interval for the statistic.

Example 3: Nonparametric Interval estimate of the median (Bootstrapping approach).

• The percentile method is one type of bootstrapping; it names itself from using a specific percentile from these thousands of calculations as the final estimate. For a 95% confidence interval of the median, with 2,000 bootstrap samples, we would sort the medians and pick the 50th and 1,950th as the lower and upper limits, reflecting the 2.5th and 97.5th percentiles. This process does not assume any specific distribution, relying instead on the data to mimic the actual distribution, with larger datasets typically giving more reliable results.

• However, when data shows strong positive skewness, the upper confidence limit from percentile bootstrapping might be underestimated. An adjustment known as the bca (bias-corrected and accelerated) bootstrap can correct this by accounting for skewness and bias.

• In summary, bootstrapping, particularly the percentile method, is a powerful tool that generates confidence intervals from the data itself, without needing to assume a particular distribution shape.

(3) ***HW04 #3*** **(link for Bootstrap Sample):** Watch this video and submit a half-page summary of what the Bootstrap sample is.

Note
Use the **binomial method** for simpler, smaller datasets, or when dealing with proportions or medians, and the assumptions of the binomial distribution are met. Opt for **bootstrapping** when dealing with complex data, larger datasets, or when you need flexibility in the type of statistics for which you're calculating confidence intervals, or the distribution of the data is unknown or non-standard.

3.2 Parametric Interval Estimate for the Median

• When we have data that, once converted into logarithms, look evenly spread out, the geometric mean gives us a good guess of the midpoint value in the original data format.

• By calculating the average of these log-transformed numbers and then applying certain formulas, we can get back to our original data scale, along with an estimate of how confident we are about this midpoint.

$$\text{Let } GM_x = \exp(\bar{y}) \tag{4}$$

Where

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$y_i = \ln(x_i)$$

then the lower and upper confidence intervals for median are

$$\exp\left(\bar{y} - t_{(a/2,n-1)}\sqrt{s_y^2/n}\right) \leq GM_x \leq \exp\left(\bar{y} + t_{(a/2,n-1)}\sqrt{s_y^2/n}\right) \quad (7)$$

where s_y^2 is the sample variance of *y*, and $t_{(a/2,n-1)}$ is the critical value of the *t*-distribution with *n*−1 degrees of freedom and a cumulative probability value of $\frac{\alpha}{2}$ (see section 4.1. for more discussion of the use of the *t*-distribution). Using the same arsenic concentration dataset as in **table 2**, we now take the natural logarithms of each value (**table 3.3**).

• The mean of the log-transformed data is 3.17, with a standard deviation of 1.96. Box plots of the original and log-transformed data are shown in **Figure 5**. Clearly, the log-transformed data are much closer to being symmetric and are well approximated by a normal distribution.

From equations 4 and 5, the geometric mean and its 95-percent confidence interval are

$$GM_C = \exp(3.17) = 23.8$$

$$\exp\left(3.17 - 2.064 \cdot \sqrt{1.96^2/25}\right) \leq \mathrm{GM_C} \leq \exp\left(3.17 + 2.064 \cdot \sqrt{1.96^2/25}\right)$$

$$\exp(2.36) \leq \mathrm{GM_C} \leq \exp(3.98), \text{ and}$$

$$10.6 \leq \mathrm{GM_C} \leq 53.5$$

• These steps assume our data follows a certain pattern (specifically, a lognormal distribution), making this method based on a specific assumption about our data's shape.

• If this assumption holds true, the geometric mean and its confidence range can more accurately pinpoint the median and its range than simpler methods that do not rely on any assumptions. However, if the log-transformed data are still skewed or have outliers (data points that are much higher or lower than the rest), it is better to use a nonparametric without assuming any specific data distribution.

Table 3. Log-transformed arsenic concentrations (in parts per billion) for groundwaters of southeastern New Hampshire (from Boudette and others, 1985), ranked in ascending order.

Rank	Value
1	0.262
2	0.405
3	0.588
4	0.956
5	1.030
6	1.253
7	1.387
8	1.569
9	2.079

Rank	Value
10	2.251
11	2.485
12	2.639
13	2.944
14	3.135
15	3.714
16	4.382
17	4.605
18	4.700

Rank	Value
19	4.787
20	5.247
21	5.481
22	5.521
23	5.704
24	5.829
25	6.363

4 Confidence Intervals for the Mean

• Interval estimates can also be made for the average value (μ) of a whole population. This approach is useful when you are interested in the overall average rather than just the midpoint of your data.

• Typically, we calculate these intervals to symmetrically encompass the average value we calculate from our sample (***X***). Thanks to a principle known as the ***Central Limit Theorem***, if we have a lot of data points, the average will tend to follow a predictable pattern (a normal distribution) even if the original data do not.

• This principle applies as long as our data do not behave too wildly, which is generally the case for water resource studies. However, if the data set is small or if the data are heavily skewed (not balanced) or have extreme values (outliers), the average of these data would not fit this predictable pattern unless the data itself is normally distributed.

• In cases where the data is skewed, we might need up to 100 data points before we can safely say that the distribution of the average value is balanced and unaffected by any extreme values in the data set.

4.1 Symmetric Confidence Interval for the Mean

Symmetric confidence intervals for the mean are computed using equation:

$$\bar{x} + t_{\left(\frac{\alpha}{2},n-1\right)} \cdot \sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{x} + t_{\left(1-\frac{\alpha}{2},n-1\right)} \cdot \sqrt{\frac{s^2}{n}} \tag{8}$$

• For a chosen confidence level of $1 - \alpha$ (e.g., 95% confidence level, $\alpha = 0.05$), the critical values from the *t*-distribution are needed to calculate the confidence interval. These critical *t*-values, which depend on the sample size (*n*), are identified using statistical software or *t*-distribution tables.

• For instance, with a sample size of 25, the critical *t*-values for a 95% confidence interval are found to be approximately -2.064 and +2.064.

• The width of the confidence interval is influenced by these critical *t*-values, the sample's standard deviation, and the size of the sample.

However, when ***the sample is small (less than 70) and the data is highly skewed or has outliers***, the standard *t*-interval approach might not be suitable. In such situations, the confidence interval can be unusually broad, sometimes even suggesting impossible negative values for data that should only be positive, indicating the need for a different approach for skewed data.

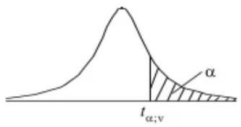
• For such data, assuming a lognormal distribution as described in **section 4.2**. will probably result in more realistic confidence intervals.

Quiz

Assume that the sample mean arsenic concentration, x = 98.4 parts per billion (ppb), is the point estimate for the true unknown population mean, μ. The standard deviation of the arsenic concentrations, s, is 144.7 ppb. Using equation 8, a 95-percent confidence interval (α = 0.01) for the true mean, μ, is:

Table of the Student's *t*-distribution

The table gives the values of $t_{\alpha;\nu}$ where
$\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

4.2 Asymmetric Confidence Interval for the Mean (for Skewed Data)

• When logarithmically transforming data, if the transformed data approximates a normal distribution, this method can offer a more precise estimate of the mean compared to computing the mean directly from untransformed data.

• This is particularly useful when other transformations do not normalize the data. The mean of the original data can be estimated using the formula $\hat{\mu}_x = \exp\left(\bar{y} + \frac{s_y^2}{2}\right)$, where $\bar{y}$ is the mean of the logarithms of x, and s_y^2 is their variance. This formula assumes the logarithms follow a normal distribution and may be biased for small samples or high variances. For larger sample sizes or smaller variances, the bias in this estimate becomes minimal.

• However, the confidence interval for the original mean, $\hat{\mu}_x$, is not directly obtained by exponentiating the confidence interval of $\bar{y}$. Calculating an exact confidence interval for the mean from log-normal data involves complex equations not covered in this course. For skewed data, a more accurate confidence interval estimation can be achieved through bootstrapping, offering a versatile solution regardless of data skewness or sample size.

(1) Bootstrap Confidence Interval for the Mean for Cases with Small Sample Sizes or Highly Skewed Data (*Example 4*)

• Just as we used the bootstrap to develop confidence intervals for the median, in section 3.1., we can also use the bootstrap to develop confidence intervals for the mean. The following example illustrates how that is done.

• Using the values from the 25 observations in the arsenic dataset used in ***Example 3*** are selected and their mean is computed (sampling with replacement). This is repeated 2,000 times and a two-sided 95-percent confidence interval for the mean is the 0.025·10,000th and 0.975·10,000th ordered resample estimates for the mean.

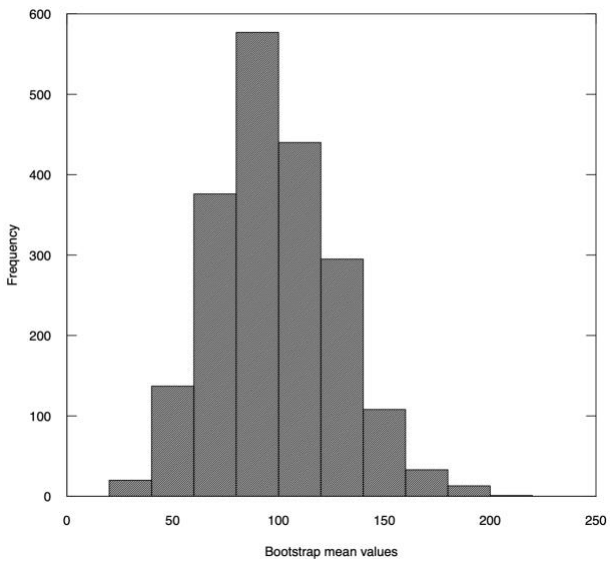


Figure 7. Histogram of bootstrapped estimates of the mean of arsenic concentrations used in ***Example 3***.

Table 4. Comparison of 95-percent confidence interval estimators for various measures of central tendency for the arsenic data, in parts per billion.
[*clow*, lower confidence interval; *cup*, upper confidence interval]

Parameter and estimation method	Estimate	c_{low}	c_{up}
Mean using *t*-interval	98.4	38.7	158.1
Median using binomial confidence interval	19	4.8	100
Geometric mean based on retransformation of *t*-interval estimates on the logs	23.8	10.6	53.5
Mean using percentile bootstrap	98.4	47.78	159.70

• In this example, the bootstrap method gives a confidence interval ranging from 47.78 ppb to 159.70 ppb for arsenic concentration. However, changing the random seed or altering the number of bootstrap samples, such as using 10,000 instead of 2,000, can lead to slightly different confidence intervals.

• This variability is a notable characteristic of the bootstrap approach, highlighting that results can vary with each execution.

• Despite this, bootstrap is highly regarded for its flexibility, especially when working with complex statistical models or when higher precision is not critical.

• For simpler statistics like the mean, running more replicates, like 10,000, is manageable and can enhance result stability. Yet, for more intricate analyses, reducing the number of replicates may be necessary, accepting a minor decrease in precision for practicality. The variability and outcomes of bootstrap methods, including confidence intervals for central tendency measures, can be effectively illustrated through histograms of bootstrap estimates.