# Describing Uncertainty 02

EN5423 | Spring 2024

w04_uncertainty_02.pdf
(Week 4)

# Contents


**1 NONPARAMETRIC PREDICTION INTERVALS ..... 2**

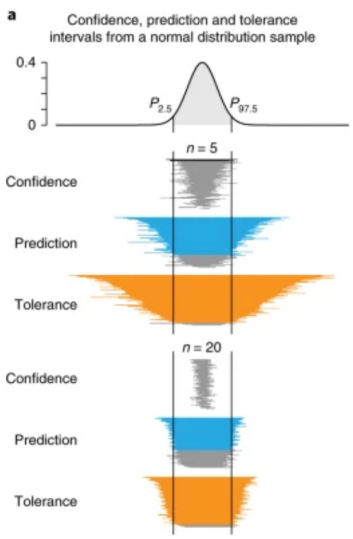
**1.1 TWO-SIDED NONPARAMETRIC PREDICTION INTERVAL (FIGURE 1) ..... 2**
**1.2 ONE-SIDED NONPARAMETRIC PREDICTION INTERVAL (FIGURE 3) ..... 5**
**1.3 PARAMETRIC PREDICTION INTERVALS ..... 6**
*1.3.1 Symmetric Prediction Interval ..... 6*
*1.3.2 Asymmetric Prediction Intervals ..... 9*
*When to Consider Non-Parametric Methods: ..... 10*
*When Parametric Methods Might Be Better: ..... 10*

**2 CONFIDENCE INTERVALS FOR QUANTILES AND TOLERANCE LIMITS ..... 11**

**2.1 CONFIDENCE INTERVALS FOR PERCENTILES VERSUS TOLERANCE INTERVALS ..... 11**
**2.2 TWO-SIDED CONFIDENCE INTERVALS FOR PERCENTILES ..... 12**
**2.3 LOWER ONE-SIDED TOLERANCE LIMITS (FIGURE 6) ..... 18**
**2.4 UPPER ONE-SIDED TOLERANCE LIMITS (FIGURE 7) ..... 19**

**3 OTHER USES FOR CONFIDENCE INTERVALS ..... 21**

**3.1 IMPLICATIONS OF NON-NORMALITY FOR DETECTION OF OUTLIERS ..... 21**
**3.2 IMPLICATIONS OF NON-NORMALITY FOR QUALITY CONTROL ..... 22**
**3.3 IMPLICATIONS OF NON-NORMALITY FOR SAMPLING DESIGN ..... 22**


| | Confidence Interval (CI) | Prediction Interval (PI) | Tolerance Interval (TI) |
|---|---|---|---|

| | | | |
|---|---|---|---|
| **What it represents** | A confidence interval quantifies the uncertainty in estimating a population parameter (e.g., mean, proportion). It provides a range within which we are confident (to a certain degree, e.g., 95%) that the true parameter value lies. | A prediction interval quantifies the uncertainty in predicting a single new observation. It provides a range within which we expect a new observation (with a certain level of confidence) to fall, taking into account the variability of the data. | A tolerance interval quantifies the range within which a specified proportion (e.g., 95%) of the population falls with a certain level of confidence. It accounts for both the mean and variability of the data and is useful for making statements about the distribution of the data. |
| **Key point** | It is about the uncertainty of the parameter estimate, not the variability of individual observations. | It is about the range for individual future observations, considering both the estimate's uncertainty and the variability in the data. | It is about covering a specific proportion of the population with a certain level of confidence. |
| **Example** | If the 95% CI for the mean height of a population is 160 to 170 cm, we are 95% confident that the true mean height of the entire population lies within this range. | If the 95% PI for the height of a new person from the population is 150 to 180 cm, we are 95% confident that this new person's height will fall within this range. | A 95% tolerance interval with 90% coverage for the population's height might be 145 to 185 cm, meaning we are 95% confident that 90% of the population's heights are within this range. |

• **Confidence intervals** are used to estimate the uncertainty of a population parameter.

• **Prediction intervals** are used to predict the range for a new observation.

• **Tolerance intervals** are used to cover a specific portion of the population with a certain confidence level.

# 1 Nonparametric Prediction Intervals

• **Prediction Interval Purpose**: It is used to determine if a new observation might belong to a different distribution than existing data by checking if it falls outside a specifically computed prediction interval from the existing data.

• **Selection of α Level**: The $\alpha$ level represents the likelihood that a new observation falls outside the prediction interval. A $100\cdot(1-\alpha)\%$ prediction interval suggests that there is a $100\cdot\alpha\%$ chance for a new observation to be outside this interval.

• **Interpretation of Being Outside the Interval**: If a new observation lies outside the interval, it does not conclusively indicate it is from a different distribution, but it raises the likelihood under certain assumptions about the distribution. The significance of this depends on the $\alpha$ level chosen.

• **Difference Between Prediction and Confidence Intervals**: Prediction intervals cater to ***individual data points***, making them ***wider than confidence intervals***, which relate to summary statistics like the mean (because an individual observation is more variable than a summary statistic computed from several observations). This width accounts for the variability around the median or mean and the error in estimating the distribution's center.

• **Nonparametric Prediction Intervals**: These intervals do not assume any specific distributional shape for the data, making them ***versatile for various data types***.

• **Application to New Data Sets**: Beyond individual observations, prediction intervals can help assess if the median or mean of a new dataset significantly differs from an existing dataset, a concept further explored in hypothesis testing in later chapters.

## 1.1 Two-sided Nonparametric Prediction Interval (Figure 1)

• To understand and apply two-sided nonparametric prediction intervals to environmental data analysis, it is crucial to grasp the underlying concepts and practical implementation steps in programming language. The nonparametric approach is particularly useful when dealing with datasets where the distribution is unknown or non-normal, common in environmental sciences.

• **Prediction Intervals** differ from confidence intervals in that they aim to capture where future observations will likely fall, as opposed to estimating a population parameter like the mean or median. A 90% prediction interval, for example, suggests that 90% of future data points are expected to lie within this range, given the data already observed.

• **Nonparametric Methods** are robust techniques that do not assume your data follow a specific distribution (e.g., normal distribution). This flexibility makes them valuable for environmental data, which can exhibit various distributional shapes due to natural variability.

• **Significance Level ($\alpha$)**: The choice of $\alpha$ affects the breadth of the prediction interval. A lower $\alpha\alpha$ (implying a higher confidence level, such as 95% vs. 90%) necessitates a wider interval to increase the probability that it encompasses the true parameter value. This approach reduces the risk of excluding the true parameter but results in less precise (wider) intervals. A balance

must be struck based on the data analysis context, where one weighs the need for confidence against the desire for interval precision.

• For instance, using the arsenic data from example 3.1 with $\alpha = 0.1$ and the Weibull plotting position, the interval can be computed as

```
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])
percentiles = [5, 95]  # Example: 5th and 95th percentiles

# Calculate percentiles using Weibull plotting positions
weibull_percentile_values = weibull_percentiles(arsenic_concentrations,
percentiles)
print(f"Weibull 5th percentile: {weibull_percentile_values[0]}")
print(f"Weibull 95th percentile: {weibull_percentile_values[1]}")
>> Weibull 5th percentile: 1.36
Weibull 95th percentile: 507.99999999999955
```

• With the arsenic dataset, the lower limit of the prediction interval is 1.36 ppb and the upper limit of the prediction interval is 508 ppb. If we conclude that any value that is less than 1.36 ppb or greater than 508 ppb comes ***from a different distribution than our original data***, then there is a 10 percent chance of drawing that conclusion if, in fact, all of the values did come from the same population.

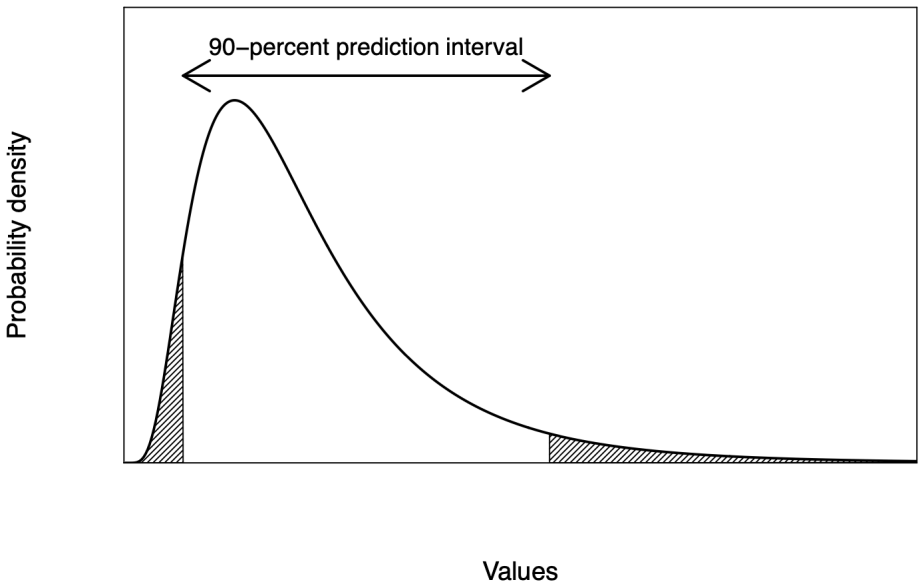



**Figure 1**. Example of a probability distribution showing the 90-percent prediction interval, with $\alpha = 0.10$. Each of the two shaded areas in the tails has a probability of $\alpha/2 = 0.05$.

Figure 3.8 uses the James River discharge data introduced in chapter 2 to show the histogram of the dataset as well as the 98-percent prediction interval for annual mean discharge. The sample median is 195 cubic meters per second ($m^3/s$) (shown with the solid line in **Fig. 2**). The prediction interval includes the interval from 64 to 350 $m^3/s$. This implies that values less than

64 $m^3/s$ have a probability of about 1 percent of occurring in any given year and values greater than 350 $m^3/s$ also have a probability of about 1 percent of occurring in any given year.

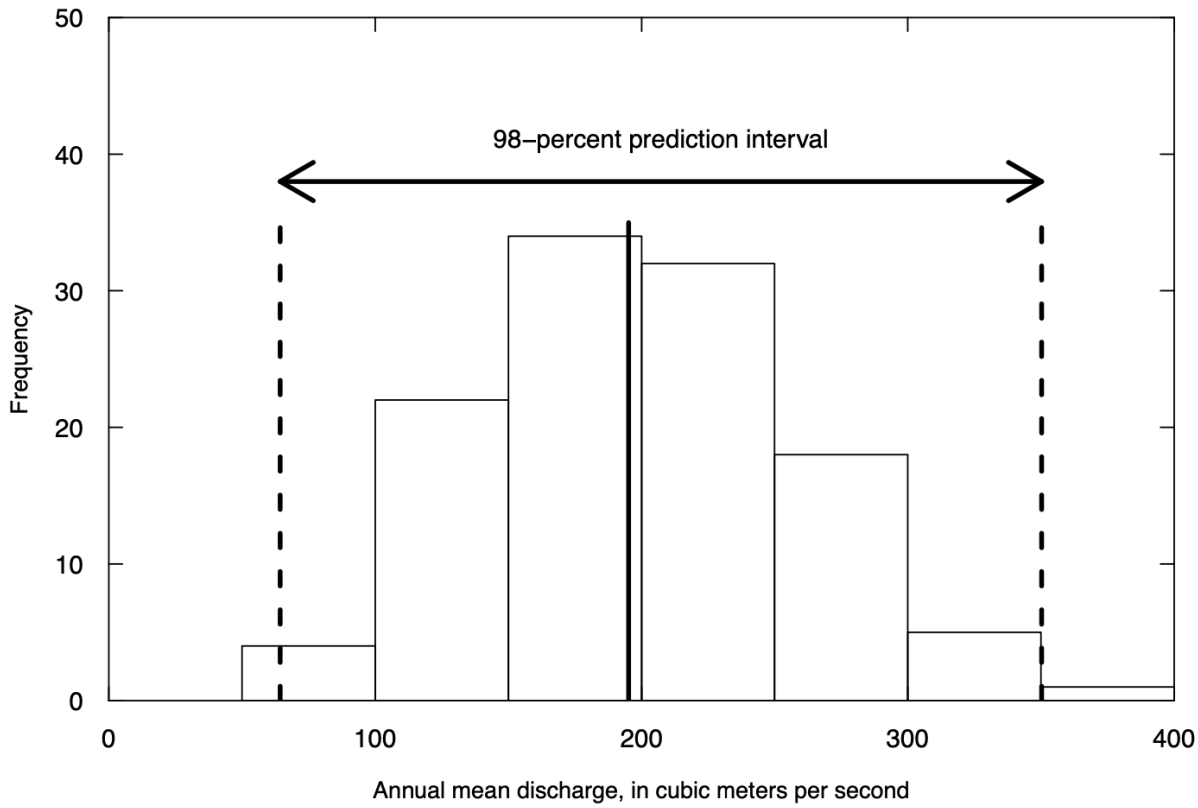



**Figure 2**. Histogram of the James River annual mean discharge dataset. The solid vertical line is the sample median and the two vertical dashed lines are the lower and upper bound of the 98-percent prediction interval.

| **Note** |
| --- |
| The concept of using the Weibull plotting position for calculating percentiles, particularly in nonparametric approaches such as prediction intervals, does seem to introduce a form of assumption about the data distribution. However, it's important to clarify that the Weibull plotting position itself is not about assuming a specific distribution for the data but rather about a way to estimate empirical cumulative distribution function (ECDF) positions more accurately, especially for the extreme values in a dataset.<br><br>When we talk about nonparametric methods, we're referring to techniques that do not rely on assumptions about the form or parameters of the distribution from which the data are drawn. The Weibull plotting position is used in nonparametric statistics as a way to adjust the rank positions used in percentile calculations to provide a more accurate reflection of the distribution's tails. This adjustment is particularly useful in hydrology and environmental sciences, where the distribution of data (like flood events or pollutant concentrations) may not follow common parametric distributions neatly, and the extremes are of particular interest. |

• Statistical Theory and Uncertainty Quantification

- Understanding Variability: The core reason behind calculating prediction intervals is to quantify the ***uncertainty associated with future observations***. Unlike a ***confidence interval, which estimates the uncertainty around a population parameter*** (e.g., the mean), a prediction interval accounts for the ***expected variability in individual future observations***.

- Capturing Distribution Extremes: The use of methods like the Weibull plotting position is particularly relevant when the data distribution is skewed or has heavy tails. These conditions are common in environmental and geophysical data, where extreme values (like flood peaks or maximum pollutant concentrations) have critical practical importance. The Weibull plotting position helps in estimating these extremes more accurately by providing a way to assign empirical probabilities to ranked observations, thus offering a better grasp of the distribution's tails.

## 1.2 One-sided Nonparametric Prediction Interval (Figure 3)

• One-sided prediction intervals are appropriate if the scientist is interested in whether a new observation is larger than existing data or smaller than existing data, ***but not both***. The decision to use a ***one-sided interval must be based entirely on the question of interest***. It should not be determined after looking at the data and deciding that the new observation is likely to be only larger, or only smaller, than existing information. One-sided intervals use $\alpha$ rather than $\alpha/2$ as the error risk, placing all the risk on one side of the interval (fig. 3.9).

If the 90-percent prediction interval is on the right tail, it would be the interval from PI to $\infty$, where PI is determined as:

```
> PI = weibull_percentiles(arsenic_concentrations, 0.9)
```

If the 90-percent prediction interval is on the left tail, it would be the interval from 0 to PI, where PI is determined as:

```
> PI = weibull_percentiles(arsenic_concentrations, 0.1)
```

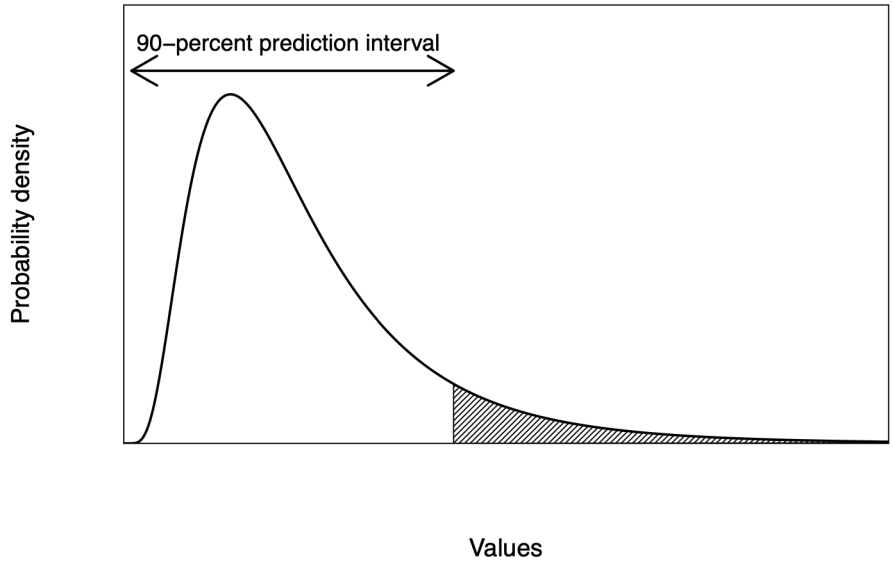



**Figure 3**. Example of a probability distribution showing the one-sided 90-percent prediction interval (α = 0.10). The shaded area in the right tail has a probability of 0.10.

**Example 1**

An arsenic concentration of 350 ppb is found in a southeastern New Hampshire groundwater sample. Does this indicate a shift in the distribution to larger values as compared to the distribution of concentrations for this arsenic_concentrations dataset ? Use $\alpha = 0.10$.

As only large concentrations are of interest, the new data point will be considered larger than the original dataset if it exceeds the $\alpha = 0.10$ one-sided prediction interval, or the upper 90th percentile of the existing data. Using the quantile function I provided above, we compute this upper 90th percentile as ***PI = weibull_percentiles(arsenic_concentrations, 0.9)***, which has a value of 316 ppb.

A concentration of 350 ppb is considered to come from a different distribution than the existing data at an $\alpha$ level of 0.10. However, 316 ppb or greater will occur approximately 10 percent of the time if the distribution of data has not changed; therefore, a concentration of 350 ppb is considered larger than the existing data at an $\alpha$ level of 0.10.

## 1.3 Parametric Prediction Intervals

• Parametric prediction intervals help us figure out if a new piece of data is probably similar to or different from data we've already collected.

• To do this, we make a guess about the pattern or “shape” of the data collection. This guess helps us build a more informed range of values.

• But, if our guess about the data’s pattern is wrong, then our range can end up being way off.

### 1.3.1 Symmetric Prediction Interval

• If the assumption is that the data follow a normal distribution, prediction intervals are then constructed to be symmetric around the sample mean and wider than the confidence intervals on the mean.

• The equation for this interval (1) differs from that for a confidence interval around the mean by adding a term $\sqrt{s^2} = s$, the standard deviation of individual observations around their mean:

$$\text{PI} = \bar{X} + t_{(\alpha/2,n-1)} \cdot \sqrt{s^2 + (s^2/n)} \text{ to } \bar{X} + t_{\left(1-\frac{\alpha}{2},n-1\right)} \cdot \sqrt{s^2 + (s^2/n)} \qquad \text{Eq. (1)}$$

• One-sided intervals are computed as before, using $\alpha$ rather than α/2 and comparing new data to only one end of the prediction interval.

**Example 2**

Using the arsenic data from ***Example 1***, we will proceed as if the data were symmetric (which we know is a ***poor assumption***). Using that assumption and α = 0.10, how would we answer the question:
Is a concentration of 370 ppb different (not just larger) than what would be expected from the previous distribution of arsenic concentrations?

The parametric two-sided $\alpha$ = 0.10 prediction interval is:

$$98.4 + t_{(0.05,24)} \cdot \sqrt{144.7^2 + \frac{144.7^2}{25}} \text{ to } 98.4 + t_{(0.95,24)} \cdot \sqrt{144.7^2 + \frac{144.7^2}{25}}$$

$$98.4 - 1.711 \cdot 147.6 \text{ to } 98.4 + 1.711 \cdot 147.6, \text{ and}$$

-154.1 to 350.9

In simple terms, if we find a value like 370 parts per billion (ppb) outside of our 90% prediction interval, we would think it is not likely from the same data set, assuming a 10% chance of being wrong ($\alpha$ = 0.10). But, there is a big problem if our interval suggests that negative values are possible because, in reality, you cannot have negative amounts of something like a concentration.

This shows that using a prediction interval that assumes data is symmetric (the same on both sides) is not right here, as we cannot have negative concentrations. Instead, we should use an interval that does not assume symmetry or try transforming the data (like taking logarithms) to make a symmetric interval work better.

On a different note, when looking at data like the water flow in the James River where the data is ***pretty symmetric***, calculating a 98% prediction interval makes sense and follows a straightforward method.

$$198.8 - 2.36\sqrt{60.6^2 + \frac{60.6^2}{116}} \text{ to } 198.8 + 2.36\sqrt{60.6^2 + \frac{60.6^2}{116}}$$

$$198.8 - 2.36 \cdot 60.86 \text{ to } 198.8 + 2.36 \cdot 60.86, \text{ and}$$

55.2 to 342.4

These results are rather similar to the nonparametric prediction interval presented above, which was 64 to 350.

### 1.3.2 Asymmetric Prediction Intervals

• When dealing with data that is not evenly spread out (i.e., skewed data), like the arsenic example we have, it is better to use prediction intervals that are not the same on both sides, known ***as asymmetric intervals***.

• This approach fits well with most environmental data, including water resources, which often lean more towards higher values than lower ones (this is called positive skewness).

• There are two main ways to create these uneven ranges: one way is by using methods from section 1 that **1) do not rely on assuming a certain data pattern**, and **2) another is by assuming the data, when transformed with logarithms**, has a symmetric distribution, and then applying standard methods on the transformed data.

• Symmetric prediction intervals, which assume data is evenly spread, should really only be used when the data clearly matches a normal (bell-curve) pattern. This is because these intervals are looking at the possible range of individual future data points, and the general rule that averages tend to become more normal with larger samples (the Central Limit Theorem) does not apply here.

• Proving data follows a normal distribution is tough, especially with ***smaller sample sizes*** common in environmental studies, and just looking at graphs might not always reveal the true shape of the data. Despite this, symmetric intervals are still frequently used, even though most water data is not symmetric and samples are often small.

• For datasets that follow a pattern where values are ***multiplicative*** (think of how times, not plus or minus, changes impact the data), using the logarithms of the data to calculate an asymmetric interval is a good approach. This method is considered ***parametric*** because it works under the assumption that the transformed data fits a lognormal distribution, which means after taking logarithms, the data behaves as if it is normally distributed.

An asymmetric (but parametric) prediction interval can be computed using logarithms. This interval is parametric because percentiles are computed assuming that the data, x, follow a lognormal distribution. Thus from equation 1:

$$PI = \exp\left[\bar{y} + t_{(\alpha/2,n-1)}\sqrt{s_y^2 + \left(s_y^2/n\right)}\right] \text{ to } \exp\left[\bar{y} + t_{\left(1-\frac{\alpha}{2},n-1\right)}\sqrt{s_y^2 + \left(s_y^2/n\right)}\right] \quad \text{Eq. (2)}$$

where $y = \ln(x)$, $\bar{y}$ is the mean, and, $s_y^2$ is the variane of $y$.

**Example 3**

An asymmetric parametric prediction interval is computed using the logs of the arsenic data from ***Example 1***. A 90-percent prediction interval becomes:

$$PI = \exp\left[3.172 - 1.71\sqrt{1.96^2 + \frac{1.96^2}{25}}\right] \text{ to } \exp\left[3.172 + 1.71\sqrt{1.96^2 + \frac{1.96^2}{25}}\right]$$

$$PI = \exp\left[3.172 - 1.71 \cdot 2.00\right] \text{ to } \exp\left[3.172 + 1.71 \cdot 2.00\right]$$

$$PI = \exp\left(-0.248\right) \text{ to } \exp\left(6.592\right), \text{ and}$$

$$PI = 0.78 \text{ to } 729.2$$

When we use percentiles, we can easily change them from one type of measurement to another. So, if we have a prediction interval in log (logarithmic) units, we can turn it back into the original measurement scale just by exponentiating (raising e to the power of) those log units.

This method, called the ***parametric prediction interval***, is different from the one using sample percentiles because it assumes that the data follows a lognormal distribution.

If we are confident that our data really does follow this lognormal pattern, then this parametric method is the way to go. However, if we are unsure about this assumption or we prefer not to assume any specific pattern for how our data is distributed, then we would lean towards the nonparametric method.

This nonparametric approach is better when we need a more flexible interval that does not rely on the data fitting a particular model.

**Note**

**When to Consider Non-Parametric Methods:**

- **Uncertainty About Distribution:** If you are not sure whether your data follows a specific distribution (like the normal or lognormal distribution), non-parametric methods are safer because they do not require such assumptions.
- **Small Sample Sizes:** Non-parametric methods can be more reliable when you are working with a small number of data points, where the distribution is unclear.
- **Robustness Desired:** These methods are less sensitive to outliers or deviations from assumed distributions, making them more robust in many real-world scenarios.

**When Parametric Methods Might Be Better:**

- **Well-Understood Distribution:** If you have strong reasons to believe your data follows a specific distribution, parametric methods can provide more precise and efficient estimates and predictions.
- **Large Samples:** With larger datasets, the Central Limit Theorem suggests that sample means will tend to follow a normal distribution regardless of the original data distribution, making parametric methods more applicable.
- **Specific Statistical Tests:** Some advanced statistical analyses and tests are designed to work with parametric assumptions, providing insights that non-parametric methods can't.

# 2 Confidence Intervals for Quantiles and Tolerance Limits

• Quantiles are a key tool used in water resources to understand how often floods happen and to analyze flow rates over time. They are similar to percentiles but are scaled from 0 to 1 instead of 0 to 100. For example, a 100-year flood, which is a very big flood that has a 1% chance of happening in any given year, is called the 99th percentile or 0.99 quantile. Similarly, a 20-year flood, with a 5% chance each year, is the 95th percentile, and a 2-year flood is right in the middle, the 50th percentile or 0.50 quantile.

• To figure out these flood quantiles, scientists assume that the amount of water in floods follows a certain pattern. In the U.S., they often use the log-Pearson Type III distribution to do this. European countries used to prefer the Gumbel distribution, but now the generalized extreme value (GEV) distribution is more popular.

• Quantiles are not just for floods. They are also used to understand low water levels, like the lowest flow in a river over a 7-day period every 10 years, known as the 7Q10, which is the 10th percentile. These measures help us define the characteristics of river flow at different times.

• Percentiles are also becoming crucial for managing water quality. They are used more and more to set and check water quality standards. Because of this, it is important to understand how variable these percentiles can be, especially when they are compared to health or legal standards.

• Here, we will also talk about ***confidence intervals*** and ***tolerance intervals*** related to percentiles, explaining how they are used in different situations to ensure water quality meets certain criteria.

• There is a distinction between one-sided lower and upper confidence limits, also known as ***tolerance limits***, and their importance in comparing percentiles to standards is highlighted. Both nonparametric methods (that do not assume a specific distribution) and distributional methods (that do) are important for understanding and applying these concepts in water resource management.

## 2.1 Confidence Intervals for Percentiles Versus Tolerance Intervals

• A ***two-sided tolerance interval*** is a range calculated ***to include a specific part***, say P% (like 90%), of all the data with a certain level of confidence (like 95%). This interval is about covering a part of the population and being sure about it.

• For example, if we are looking at the amount of stuff dissolved in water from the Cuyahoga River, Ohio, a 90% tolerance interval might capture the middle 90% of data with 95% confidence. This means we are pretty sure that this interval has 90% of all possible values.

• Because we are sampling, not measuring every single thing, this interval is wider than just looking at the 5th to 95th percentiles directly (e.g., sampling error). Two-sided tolerance intervals are not common in water studies but are used in quality control.

• There is also something called a ***two-sided confidence interval around a percentile***, which shows how accurately we have estimated that percentile (similar to a confidence interval around a mean or median), like how close our observed 10-year flood level is to what we would expect. This is different from a tolerance interval.

• Often, though, in water studies, ***we use one-sided limits more***. These are like only caring about either the high end or the low end, not both. They help us understand the extremes - like how high floods can get for designing structures, or if pollution exceeds safe levels. One-sided limits are similar to taking half of the confidence interval around a percentile.

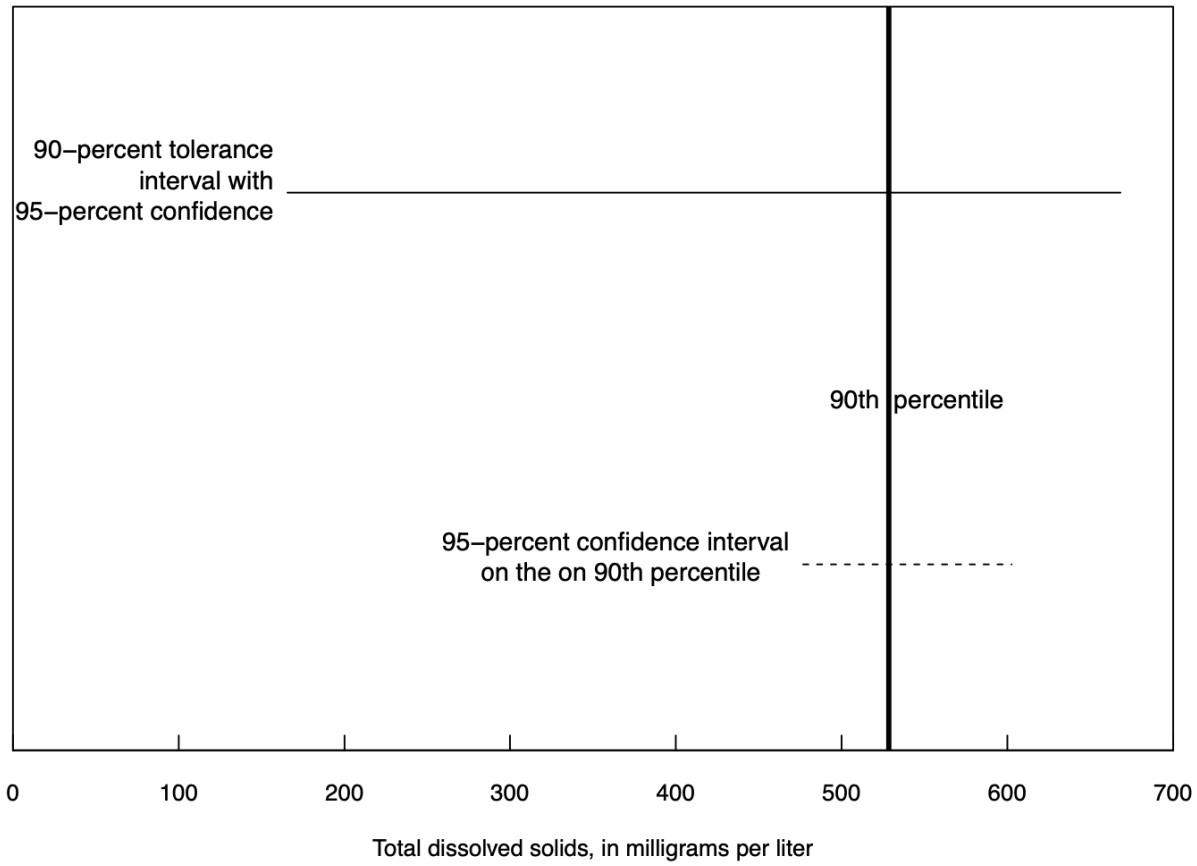



**Figure 4**. A two-sided tolerance interval with 90-percent coverage (solid line), and two-sided confidence interval on the 90th percentile (dashed line). Both were computed using a 95-percent confidence coefficient. The solid vertical line is the sample 90th percentile. The data are total dissolved solids concentrations for the Cuyahoga River, at Independence, Ohio (70 observations from 1969–73).

## 2.2 Two-sided Confidence Intervals for Percentiles

• A two-sided confidence interval around a percentile is a way to show how accurately we've guessed where that percentile falls within our data, similar to how we would show our confidence in the average or middle value. This can be done in two main ways:

**1) Non-Distributional Method:**

- Just like we might find the middle value (median) by looking directly at the data, we can find the confidence interval around a percentile by simply counting observations from the edges of

our data set towards the middle. This approach does not assume the data follows any specific pattern (distribution), but it does require having a lot of data, especially for percentiles that are at the extreme ends (like very high or very low values).

**2) Using a Distribution Model:**

- Alternatively, we can assume our data fits a certain shape or distribution, which allows us to estimate the upper limit of the interval even beyond the largest observation we have. This method's accuracy depends on ***how well the assumed distribution matches the real data distribution***. For smaller sets of data, or when we need to estimate around a high percentile, this might be our only option.

- For any percentile, we can create nonparametric (not based on a data distribution model) confidence intervals similar to how we would for the median. This involves identifying the lower and upper positions in our data that mark the ends of the confidence interval. We use a statistical tool (like the `scipy.stats.binom.ppf` function in Python) to find these positions based on the percentile we are interested in (for example, p = 0.75 for the 75th percentile).

- To ensure our confidence level (like 95%) accurately reflects the chance of being above the lower end, we add one to the lower rank. This step makes sure the interval starts just beyond the lower data point, not at it.

**Example 4**

For the arsenic concentrations data used in ***Example 1***, we aim to determine a 95-percent confidence interval for the 20th percentile of concentration ($C_{0.20}$), which stands at p = 0.2.

This specific percentile is selected for illustration purposes due to its moderate extremity. For a small sample size of $n = 25$, computing a 95-percent nonparametric interval around a high percentile, like the 90th, would not be feasible.

```
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])
C_0_20 = weibull_percentiles(arsenic_concentrations, 20)

print(f"20th percentile (C_0.20) = {C_0_20} ppb")
>> 20th percentile (C_0.20) = 2.94 ppb
```

The sample 20th percentile $\hat{C}_{0.20} = 2.94$ ppb, the $0.20 \cdot (25+1) = 5.2$th smallest observation, or two-tenths of the distance between the 5th and 6th smallest observations. The order statistics corresponding to $\frac{\alpha}{2} = 0.0025$ are at ranks 1 and 9 in the dataset:

```
n = len(arsenic_concentrations)
p = 0.2
lower_rank, upper_rank = binom.ppf([0.025, 0.975], n, p).astype(int)
print(f"Lower rank: {lower_rank}, Upper rank: {upper_rank}")
>> Lower rank: 1, Upper rank: 9
```

Adding 1 to the lower rank produced by the binom.cdf function and summing the probabilities of inclusion for the 2nd through 9th ranked observations yields:

```
sum_probabilities = binom.cdf(upper_rank, n, p) -
binom.cdf(lower_rank, n, p)
print(f"Sum of probabilities of inclusion: {sum_probabilities}")
>> Sum of probabilities of inclusion: 0.9552784048575103
```

The interval between and including the 2nd and 9th ranked observations (1.5 ppb to 8 ppb) contains the true population 20th percentile of arsenic with confidence 95.5 percent. Note that the asymmetry around $\hat{C}0.20 = 2.94$ reflects the asymmetry of the data.

**Example 5**

An alternate method to compute a nonparametric interval is bootstrapping.

```
# Number of bootstrap samples
n_bootstrap_samples = 10000

# Function to calculate the 20th percentile
def percentile_20(data):
    return weibull_percentiles(data, 20)

# Store bootstrapped 20th percentiles
bootstrapped_20th_percentiles = []

for _ in range(n_bootstrap_samples):
    # Sample with replacement
    bootstrap_sample = np.random.choice(arsenic_concentrations,
size=len(arsenic_concentrations), replace=True)
    # Calculate and store the 20th percentile of this bootstrap
sample

bootstrapped_20th_percentiles.append(percentile_20(bootstrap_sample))

# Calculate the 95% confidence interval from the bootstrapped 20th
percentiles
lower_bound = np.percentile(bootstrapped_20th_percentiles, 2.5)
upper_bound = np.percentile(bootstrapped_20th_percentiles, 97.5)

print(f"Bootstrap 95% Confidence Interval for the 20th percentile:
({lower_bound}, {upper_bound})")
>> Bootstrap 95% Confidence Interval for the 20th percentile:(1.56, 9.5)
```

• For sample sizes larger than 20, a large-sample (normal) approximation to the ***binomial distribution*** is a third method to obtain nonparametric interval estimates for percentiles. Ranks corresponding to the upper and lower confidence limits are determined by equations 3.7 and 3.8 using quantiles of the standard normal distribution, $z_{\frac{\alpha}{2}}$ and $z_{\left[1-\frac{\alpha}{2}\right]}$. Those ranks are:

$$R_L = np + z_{\alpha/2}\sqrt{np(1-p)} + 0.5 \qquad \text{Eq. (3)}$$

$$R_U = np + z_{[1-\alpha/2]}\sqrt{np(1-p)} + 0.5 \qquad \text{Eq. (4)}$$

whereas in example 3.9, $n$ is the sample size and $p$ is the quantile value for the percentile around which the interval is computed. The 0.5 terms added to equations 3 and 4 reflect a continuity correction (you will learn about this later) of 0.5 for the lower bound and −0.5 for the upper bound (which otherwise would be a value of +1). The computed ranks $R_U$ and $R_L$ are rounded to the nearest integer.

• After rounding, the 2nd and 9th ranked observations are found to be the approximate $\alpha = 0.05$ confidence limit on $C_{0.2}$, agreeing with the exact confidence limit computed above.

• For a test of whether a percentile significantly differs (either larger or smaller) from a prespecified value $X_0$, simply compute a $(1 - \alpha)$-percent two-sided confidence interval for the percentile. If $X_0$ falls within this interval, the percentile does not significantly differ from $X_0$ at a significance level $\alpha$ (Fig. 5*A*). If $X_0$ is not within the interval, the percentile significantly differs from $X_0$ at the significance level of $\alpha$ (Fig. 5*B*).

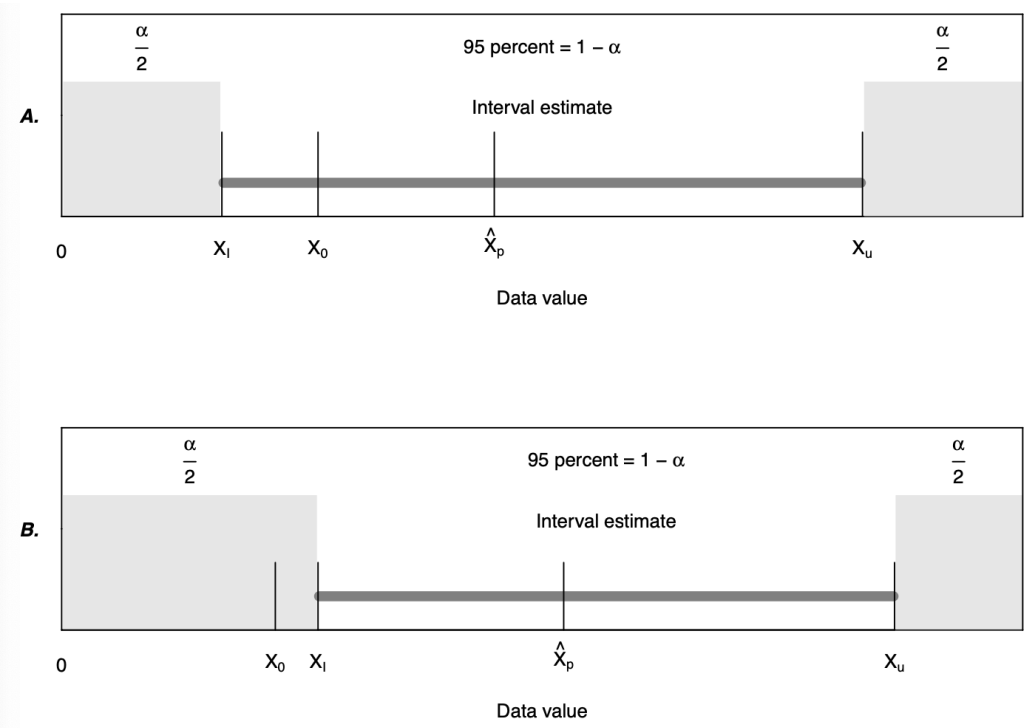



**Figure 5**. Confidence interval on the pth percentile Xp as a test for H0: Xp = X0. **A**. $X_0$ inside the interval estimate, Xp not significantly different from $X_0$. B. $X_0$ outside the interval estimate, Xp significantly different from $X_0$.

**Example 6**

James.Q.csv include annual peak discharges in cubic meters per second. Of interest is the 5-year flood, the flood that is likely to be equaled or exceeded once every 5 years (20 times in 100 years), and so is the 80th percentile of annual peaks.

Let's assume that using some method of calculation (such as a regional statistical model or a deterministic rainfall-runoff model) it has been determined that the 5-year design flood is 220 m3/s. We would like to consider if the actual data are consistent with this estimate, at $\alpha = 0.05$.

The 5-year flood is equivalent to the 80th percentile of the population of floods because 0.8 = 1 − (1/5). The 80th percentile is estimated from the data sets (n = 116) as follows:

```
# Calculate the length of Q
n = len(Q)

# Calculate the 80th percentile using numpy, which by default uses a
method similar to type = 6 in R
quantile_80 = np.percentile(Q, 80)

print(f"80th percentile: {quantile_80}")
>> 80th percentile: 252.4
```

Therefore, $\hat{Q}_{0.8} = 252.4\ \mathrm{m}^3/\mathrm{s}$. Following equations 3 and 4, a two-sided confidence interval on this percentile is:

$$R_L = np + z_{\alpha/2}\sqrt{np(1-p)} + 0.5$$
$$R_U = np + z_{1-\alpha/2}\sqrt{np(1-p)} + 0.5$$

```
# Calculate the ranks (RL and RU) based on the formula provided in R
RL = n * 0.8 + norm.ppf(0.025) * np.sqrt(n * 0.8 * 0.2) + 0.5
RU = n * 0.8 + norm.ppf(0.975) * np.sqrt(n * 0.8 * 0.2) + 0.5

# Calculate the quantiles at RL/n and RU/n
quantiles = np.percentile(Q, [RL/n * 100, RU/n * 100])

print(f"RL (converted to percentile): {RL/n * 100}%")
print(f"RU (converted to percentile): {RU/n * 100}%")
print(f"Quantile at RL/n: {quantiles[0]}")
print(f"Quantile at RU/n: {quantiles[1]}")
>>
RL (converted to percentile): 73.15191098427847%
RU (converted to percentile): 87.71015798123878%
Quantile at RL/n: 236.9496371583043
Quantile at RU/n: 266.23336335684917
```

Thus, the 95-percent confidence interval for the 5-year flood lies between the 73.2th and 87.7th ranked peak flows, or 236.95 < $\hat{Q}_{0.8}$<266.23. The interval does not include the design value $X_0 = 220\ \mathrm{m}^3/\mathrm{s}$. Therefore the 20-year flood does differ from the design value at a significance level of $\alpha = 0.05$.

• When working with small datasets, estimating ranges around very high or very low percentiles can be tricky without assuming the data follows a certain pattern (distribution).

• If we try to use methods that do not rely on a distribution (nonparametric methods) with too few data points, we might end up with the smallest or largest values as our range limits, which might not be accurate.

• To avoid this, sometimes we assume the data fits a certain distribution, which can make our estimates more precise as long as our assumption matches the real data pattern.

• However, if our assumed distribution does not really fit the data, our estimates could be off. The tricky part is, when we really need to assume a distribution due to having only a few observations, it is also harder to check if our assumption about the data pattern is correct.

• Like the non-distribution based methods, when we use a distribution to estimate ranges, we can check if a specific value ($X_0$) falls within our estimated range to see if our observed data significantly differs from $X_0$.

• For example, assuming our data fits a log-normal distribution (meaning the logarithms of our data follow a normal distribution), calculating these estimates becomes more straightforward. (***Homework04 #4***)

• We take the logarithm of our data points, calculate the average (mean) and the spread (standard deviation) of these logged values, and from there, we can estimate the percentile values we are interested in.

• Let $y = \ln(x)$ where the $x$ values are the original units. The sample mean of the $y$ values is denoted $y$ and sample standard deviation of the $y$ values is $sy$. The point estimate of any percentile is then:

$$\hat{X}_p = \exp\left(\bar{y} + z_p \cdot s_y\right) \qquad \text{Eq. (5)}$$

where $zp$ is the $p$th quantile of the standard normal distribution.

For percentiles other than the median, confidence intervals are computed using the noncentral $t$-distribution (Stedinger, 1983). The confidence interval on $Xp$ is:

$$CI(X_p) = \left[\exp\left(\bar{y} - n^{1/2} \cdot \zeta_{\left[1-\frac{a}{2}\right]} \cdot s_y\right), \exp\left(\bar{y} - n^{1/2} \cdot \zeta_{[a/2]} \cdot s_y\right)\right] \qquad \text{Eq. (6)}$$

where $\zeta_{[a/2]}$ is the $\alpha/2$ quantile of the noncentral $t$-distribution with $n - 1$ degrees of freedom and noncentrality parameter $-n^{1/2}$ for the desired percentile with sample size of $n$. Using Python, the $z_p$ values are computed with the function `scipy.stats.norm.ppf` and the $\zeta_{[a/2]}$ values are computed with the function `scipy.stats.nct.ppf`.

## 2.3 Lower One-sided Tolerance Limits (Figure 6)

• A lower tolerance limit (LTL) for a certain percentile, with a confidence level of $(1-\alpha)$ and covering a proportion $p$ of the data, is basically the same as a one-sided lower confidence limit for that percentile.

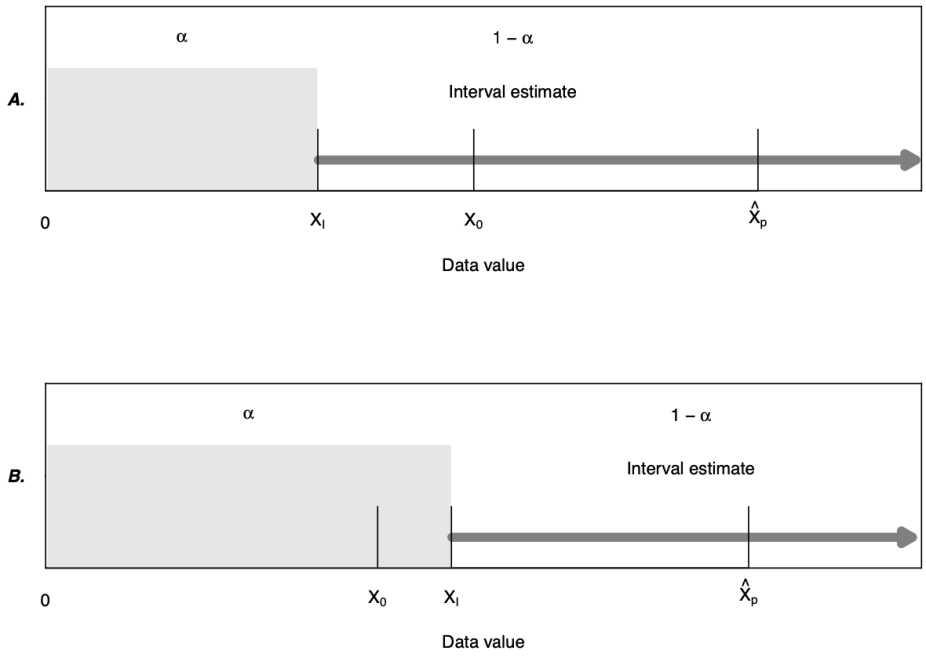



**Figure 6**. Lower tolerance limit as a test for whether the percentile $X_p > X_0$. **A**. $X_0$ above lower limit $X_l$: $X_p$ not significantly greater than $X_0$. **B**. $X_0$ below lower limit $X_l$: $X_p$ significantly greater than $X_0$.

• When we calculate an LTL, we focus only on the lower end of the data range, treating the upper end as if it does not matter (sometimes, this is shown as infinity in software programs).

• Lower tolerance limits are useful in environmental studies, like checking water quality, ***to see if the amount of a substance exceeds safe levels***. If the LTL ($X_l$) is above the safe level, it means we are $(1-\alpha)$ confident that more than $(1-p)\times 100\%$ of the water samples have higher concentrations of the substance than allowed.

• To calculate a nonparametric LTL for any percentile, we use a method similar to that for two-sided confidence intervals, but we assign all the error probability $\alpha$ to the lower side. This method involves a large-sample approximation to find the rank of the observation that marks the lower tolerance limit.

• The rank of the observation corresponding to the lower confidence limit on the percentile (lower tolerance limit) is:

$$R_L = np + z_a\sqrt{np(1-p)} + 0.5 \qquad \text{Eq. (7)}$$

• If you want to check if a certain percentile (let's say the amount of a pollutant) is significantly higher than what is considered safe (a criterion), you compute the LTL for that percentile.

• If the LTL is higher than the criterion, it indicates that the pollutant level is significantly higher than what is safe, according to our statistical confidence and coverage proportion.

**Example 7: Nonparametric lower tolerance limit as a test for whether the 90th percentile exceeds a standard.**

A water-quality standard states that the 90th percentile of arsenic concentrations in drinking water shall not exceed 10 ppb. Has this standard been violated at the $\alpha = 0.05$ confidence level by the New Hampshire arsenic data in ***Example 1***?

PI = weibull_percentiles(arsenic_concentrations, 0.9), which has a value of 316 ppb.

or by hand

$\hat{C}_{0.90} = (25 + 1) \cdot 0.9 = 23.4^{\text{th}}$ data point = 300 + 0.4(340-300) = 316 ppb.

Following equation 7, the rank of the observation corresponding to a one-sided 95-percent lower confidence bound on $C_{0.90}$ is

$$R_L = np + z_\alpha \sqrt{np(1-p)} + 0.5$$
$$= 25 \cdot 0.9 + z_{0.05} \cdot \sqrt{25 \cdot 0.9(0.1)} + 0.5$$
$$= 22.5 + (-1.64)\sqrt{2.25} + 0.5 = 20.5^{\text{th}} \text{ lowest observation}$$

thus, the lower confidence limit is the 20.5th lowest observation—or 215 ppb—halfway between the 20th and 21st observations. This confidence limit is greater than $X_0 = 10$ and therefore the standard has been proven to be exceeded at the 95-percent confidence level.

## 2.4 Upper One-sided Tolerance Limits (Figure 7)

• An upper tolerance limit (UTL) with a certain confidence level (1 - $\alpha$) and covering a proportion $p$ of the data works like a one-sided upper confidence limit for a specific percentile.

• When calculating a UTL, the focus is only on the upper end, essentially ignoring the lower end (which might be represented as negative infinity in software).

• UTLs are particularly useful in environmental and natural resource studies for setting a threshold. For instance, it helps in establishing a limit that we expect only (1 - $p$) × 100% of future observations to exceed, with (1 - $\alpha$)% confidence.

• If more new observations than expected exceed the UTL, it suggests that the environmental conditions or the characteristics of the natural resource have changed from the original assumptions. For example, UTLs can be used ***to set a baseline for acceptable sediment loads or chemical concentrations in studies aimed at controlling pollution***.

• To calculate a nonparametric UTL for a percentile, you use a similar approach to that for calculating upper confidence limits, but here, all the uncertainty ($\alpha$) is considered only for the upper range. This calculation often relies on a large-sample approximation to identify the observation rank that serves as the UTL.

• If you need to check whether a particular percentile ($X_p$) is significantly lower than a given standard ($X_0$), you would compute the UTL for $X_p$. If this UTL falls below $X_0$, it indicates that the actual percentile is significantly lower than the standard, based on the established confidence level and coverage. ***This approach is particularly relevant for environmental monitoring and management, where exceeding certain thresholds can indicate a need for intervention or reassessment of conditions.***

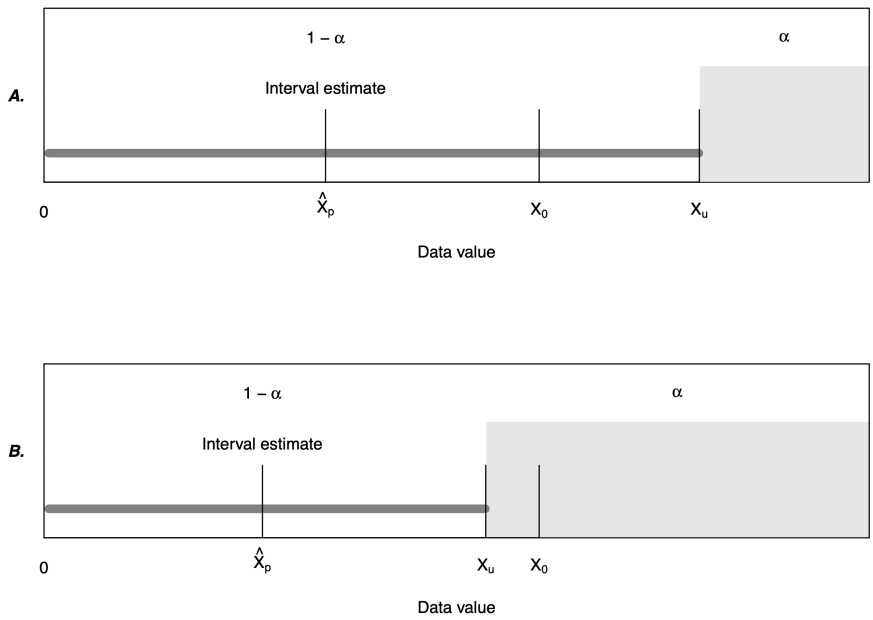



**Figure 7.** Upper tolerance limit as a test for whether percentile $X_p < X_0$. **A**. $X_0$ below the upper tolerance limit $X_u$; $X_p$ not significantly less than $X_0$. **B**. $X_0$ above the upper tolerance limit $X_u$; $X_p$ significantly less than $X_0$.

**Example 8: Nonparametric upper tolerance limit as a test for whether the 90th percentile is below the regional $_7Q_{10}$**

We have 68 values of the annual 7-day minimum flows for climate years 1942–2009 (a climate year runs from April 1 to March 31) on the Little Mahoning Creek at McCormick, Pennsylvania, expressed in cubic meters per second. We can load and view the data and then estimate the 7-day, 10-year, low flow ($_7Q_{10}$) which has a value of 0.04911 m3/s:

```
n = len(Q_7min)  # Number of observations in your dataset
RU = 0.1  # Define the rank of the upper quantile you are interested in
# Calculate the Weibull plotting position for the given rank
probs = (np.arange(1, n+1)) / (n + 1)
RU = 0.1
xUpper = np.interp(RU, probs, Q_7min)
xUpper
>> 0.04911
```

The estimate of the $_7Q_{10}$ based on the data (with no distributional assumptions) is 0.04911 m3/s. However, based on a regional regression model of $_7Q_{10}$ values versus watershed characteristics the $_7Q_{10}$ was expected to equal 0.06 m3/s (call that $X_0$ = 0.06 m3/s). The question is: Should we reject at $\alpha = 0.05$ the null hypothesis that the $_7Q_{10}$ is 0.06 m3/s versus the alternate hypothesis that it is below 0.06 m3/s? To answer this question we need to compute the 95-percent upper confidence limit for the true 0.1 percentile of the distribution of annual minimum 7-day low flows. We can compute that as follows:

$$R_U = np + z_{1-\alpha}\sqrt{np(1-p)} + 0.5$$
$$R_U = 68 \cdot 0.1 + 1.644\sqrt{68 \cdot 0.1 \cdot (0.9)} + 0.5$$
$$R_U = 11.369$$

The rank of the upper 95-percent confidence bound on the 10th percentile of the distribution of annual 7-day minimum flows is the 11.369th rank out of the 68 observed values.

We calculate the discharge associated with this fractional rank as xUpper, 16.71933%, which is 0.076 m3/s.

```
x0 = np.interp(11.369/n, probs, Q_7min)
x0
>> 0.0757
```

Our best estimate based only on the quantiles of the data at the site is 0.04911 m3/s. We cannot reject the hypothesis that the true value of the $_7Q_{10}$ is 0.06 m3/s because the upper 95-percent confidence bound for the $_7Q_{10}$ is above 0.06, at 0.076 m3/s.

# 3 Other Uses for Confidence Intervals

Confidence intervals are used for purposes other than interval estimates. Three common uses are **(1) to detect outliers**, **(2) for quality control charts**, and **(3) for determining sample sizes necessary to achieve a stated level of precision**. However, the implications of data non-normality for the three applications are often overlooked; these issues are discussed in the following sections.

## 3.1 Implications of Non-normality for Detection of Outliers

Outliers in a dataset can dramatically influence the interpretation of data, especially when constructing confidence intervals under the assumption of normality. Confidence intervals are used to determine the range within which we expect a certain percentage of the data to fall. When data are non-normal, especially with outliers, the traditional confidence intervals (which usually assume normality) might not accurately reflect the variability and distribution of the data. We need for caution when automatically labeling data points as outliers based on standard confidence intervals derived from normal distribution assumptions. Instead, we need a more nuanced approach to evaluating outliers by considering whether they represent true extreme

values or errors, recognizing that confidence intervals based on normality may not always be appropriate.

## 3.2 Implications of Non-normality for Quality Control

In quality control processes, confidence intervals are crucial for monitoring the consistency of product quality or laboratory measurements. The boxplot, for instance, uses confidence intervals around sample means to signal when the process mean shifts away from the control mean. However, when underlying data are skewed or contain outliers, the confidence intervals calculated under normality assumptions might not accurately detect shifts in the process mean. So there is a limitation of traditional control charts in the presence of non-normal data and suggests alternative approaches, such as transforming data or using charts that do not rely on the assumption of normality, to better capture the true variability and ensure the reliability of quality control measures.

## 3.3 Implications of Non-normality for Sampling Design

Sampling design is significantly impacted by the assumptions about data distribution. Confidence intervals are often used to estimate the sample size needed to achieve a desired level of precision in estimating parameters like the mean. However, these calculations typically assume that the data are normally distributed. There is challenges of designing samples based on normality assumptions, especially for skewed data. The non-normality can lead to overestimated sample size requirements or inappropriate confidence intervals that do not accurately reflect the data's variability. Alternative methods, including transforming data to approximate normality or employing nonparametric approaches, are suggested to derive more realistic confidence intervals and sample sizes that account for the actual distribution of the data.