The 80th percentile is the value in a set that is at least as large as 80% of the elements in the set

For \( s = [1, 7, 3, 9, 5] \), \( \text{percentile}(80, s) \) is 7

The 80th percentile is ordered element 4: \((80/100) \times 5\)

For a percentile that does not exactly correspond to an element, take the next greater element instead

\[
\begin{align*}
\text{percentile}(10, s) &= 1 \\
\text{percentile}(20, s) &= 3 \\
\text{percentile}(21, s) &= 3 \\
\text{percentile}(40, s) &= 3
\end{align*}
\]

**Inference:** Making conclusions from random samples

**Population:** The entire set that is the subject of interest

**Parameter:** A quantity computed for the entire population

**Sample:** A subset of the population

In a Random Sample, we know the chance that any subset of the population will enter the sample, in advance

**Statistic:** A quantity computed for a particular sample

Estimation is a process with a random outcome

Population (fixed) → Sample (random) → Statistic (random)

A 95% **Confidence Interval** is an interval constructed so that it will contain the parameter for 95% of samples

For a particular sample, the interval either contains the parameter or it doesn't; the process works 95% of the time

**Resampling:** When we wish we could sample again from the population, instead sample from the sample

Using a confidence interval to test a hypothesis:

- Null hypothesis: Population mean = \( x \)
- Alternative hypothesis: Population mean ≠ \( x \)
- Cutoff for P-value: \( p\%

Method:

- Construct a (100-\( p\% \)) confidence interval for the population average
- If \( x \) is not in the interval, reject the null
- If \( x \) is in the interval, can't reject the null

**Permutation test** for comparing two samples

- **E.g.:** Among babies born at some hospital, is there an association between birth weight and whether the mother smokes?
- **Null hypothesis:** The distribution of birth weights is the same for babies with smoking mothers and non-smoking mothers.
- **Inferential Idea:** If maternal smoking and birth weight were not associated, then we could simulate new samples by replacing each baby's birth weight by a randomly picked value from among all babies.
  - Permute (shuffle) the outcome column \( K \) times. Each time:
    - Create a shuffled table that pairs each individual with a random outcome.
    - Compute a sampled test statistic that compares the two groups, such as the difference in mean birth weights.
  - Compare the observed test statistic to these sampled test statistics to see whether it is typical under the null.

Computing a confidence interval for an estimate from a sample:

- Collect a random sample
- Resample \( K \) times from the sample, with replacement
  - Compute the same statistic for each resampled sample
  - Take percentiles of the resampled estimates
  - 95% confidence interval: [2.5 percentile, 97.5 percentile]

The Central Limit Theorem (CLT)

If the sample is large, and drawn at random with replacement, then, **regardless of the distribution of the population**, the probability distribution of the sample average (or sample sum) is roughly bell-shaped

- Fix a large sample size
- Draw all possible random samples of that size
- Compute the mean of each sample
- You'll end up with a lot of means
- The distribution of those is the probability distribution of the sample mean
- It's roughly normal, centered at the population mean
- The SD of this distribution is the (population SD) / \( \sqrt{\text{sample size}} \)

Choosing sample size so that the 95% confidence interval is small

- CLT says the distribution of a sample proportion is roughly normal, centered at population mean
- 95% confidence interval:
  - **Center** ± 2 SDs of the sample mean
  - **Width** is 4 SDs of the sample mean
    
    \[
    = 4 \times (\text{SD of population})/\sqrt{\text{sample size}}
    \]

To find the \( k \) nearest neighbors of an example:

- Find the distance between the example and each example in the training set
- Augment the training data table with a column containing all the distances
- Sort the augmented table in increasing order of the distances
- Take the top \( k \) rows of the sorted table

To classify a point:

- Find its \( k \) nearest neighbors
- Take a majority vote of the \( k \) nearest neighbors to see which of the two classes appears more often
- Assign the point the class that wins the majority vote

<table>
<thead>
<tr>
<th>Expression</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimize(fn)</td>
<td>Return an array of argument values that minimize a function.</td>
</tr>
<tr>
<td>table.row(i)</td>
<td>Return the row of a table at index ( i ).</td>
</tr>
<tr>
<td>table.rows</td>
<td>All rows of a table; Used in for row in table.rows:</td>
</tr>
</tbody>
</table>

```python
def bootstrap_mean(original_sample, label, replications):
    means = make_array()
    for i in np.arange(replications):
        bootstrap_sample = original_sample.sample()
        resampled_mean = np.mean(bootstrap_sample.column(label))
        means = np.append(means, resampled_mean)
    return means
resampled_means = bootstrap_mean(some_table, some_label, 5000)
right = percentile(97.5, resampled_means)
left = percentile(2.5, resampled_means)
confidence_interval = [left, right]
```
Measures how clustered the scatter is around a straight line
Describing the deviation of $x$ from $0$ (the average of $x$'s)
Measured in $r: -1 \leq r \leq 1$; $r = 1$ (or -1) if the scatter is a perfect straight line
Chebyshev: At most $1/z^2$ are more than $z$ SDs from the mean
Almost all standard unit values are in the range (-5, 5)

Correlation Coefficient ($r$) =
$$\text{average of product of } x \text{ in standard units and } y \text{ in standard units}$$

Measures how clustered the scatter is around a straight line
- $-1 \leq r \leq 1$; $r = 1$ (or -1) if the scatter is a perfect straight line
- $r$ is a pure number, with no units
- $r$ is not affected by changing units of measurement
- $r$ is not affected by switching the horizontal and vertical axes

Regression to the mean: a statement about $x$ and $y$ pairs
- Measured in standard units
- Describing the deviation of $x$ from 0 (the average of $x$'s)
- And the deviation of $y$ from 0 (the average of $y$'s)

Regression line: $y = \text{average of } y$ + slope * $x$ + intercept

Correlation
$$y(\text{su}) = r \times \bar{x}(\text{su})$$

In original units, the regression line has this equation:
$$\frac{\text{estimate of } y - \text{average of } y}{\text{SD of } y} = r \times \frac{\text{the given } x - \text{average of } x}{\text{SD of } x}$$

Properties of fitted values, residuals, and the correlation $r$:
- $r^2 = \frac{(\text{Variance of fitted values})}{(\text{Variance of y})}$
- $1 - r^2 = \frac{(\text{Variance of residuals})}{(\text{Variance of y})}$
- $r = \sqrt{1 - r^2}$

In original units, the regression line has this equation:
$$y = \text{average of } y + \text{slope} \times x$$

If $r > 0$: A positive slope
If $r < 0$: A negative slope
If $r = 0$: A horizontal line

The slope and intercept are unique for linear regression
$$\text{slope of the regression line} = \frac{\text{SD of } y}{\text{SD of } x}$$
$$\text{intercept of the regression line} = \text{average of } y - \text{slope} \times \text{average of } x$$

Bootstrap the scatter plot & find the slope of the regression line through the bootstrapped plot many times.
- Draw the empirical histogram of all the resampled slopes.
- Get the “middle 95%” interval: that’s an approximate 95% confidence interval for the slope of the true line.

Null hypothesis: The slope of the true line is 0.
- Construct a bootstrap confidence interval for the true slope.
- If the interval doesn’t contain 0, reject the null hypothesis.
- If the interval does contain 0, there isn’t enough evidence to reject the null hypothesis.

Fitted value: Height of the regression line at some $x$: $a \times x + b$.
Residual: Difference between $y$ and regression line height at $x$.
Regression model: $y$ is a linear function of $x$ + normal "noise"
Residual plot looks like a formless "noise" cloud under this model
Average of residuals is always 0 for any scatter diagram
Properties of fitted values, residuals, and the correlation $r$:
$$\text{Variance of residuals} / \text{Variance of fitted values} = 1 - r^2$$
$$\text{Variance of y} = \text{Variance of residuals} + \text{Variance of fitted values}$$

Properties of residuals:
- Residuals are uncorrelated with the fitted values
- Mean of residuals is 0
- Variance of residuals: $\text{var}(r) = r^2$ (in standard units)
- SD of residuals: $\text{SD}(r) = \sqrt{r^2}$
- SD of residuals: $\text{SD}(r) = r \times \text{SD of } y$
- Relationship w/ fitted values: $\text{SD}(\text{residuals}) = \text{SD}(\text{fitted values}) / \sqrt{\text{SD}(y)}$
- Relationship w/ fitted values: $\text{SD}(\text{residuals}) / \text{SD}(y) = \sqrt{1 - r^2}$

Data 8 Final Study Guide — Page 2
Mean (or average): Balance point of the histogram
- Not the “half-way point” of the data; the mean is not the median unless the histogram is symmetric
- If the histogram is skewed, then the mean is pulled away from the median in the direction of the tail

Bootstrap the scatter plot & find the slope of the regression line through the bootstrapped plot many times.
- Draw the empirical histogram of all the resampled slopes.
- Get the “middle 95%” interval: that’s an approximate 95% confidence interval for the slope of the true line.
- Null hypothesis: The slope of the true line is 0.
- Construct a bootstrap confidence interval for the true slope.
- If the interval doesn’t contain 0, reject the null hypothesis.
- If the interval does contain 0, there isn’t enough evidence to reject the null hypothesis.

Fitted value: Height of the regression line at some $x$: $a \times x + b$.
Residual: Difference between $y$ and regression line height at $x$.
Regression model: $y$ is a linear function of $x$ + normal "noise"
Residual plot looks like a formless "noise" cloud under this model
Average of residuals is always 0 for any scatter diagram
Properties of fitted values, residuals, and the correlation $r$:
$$\text{Variance of residuals} / \text{Variance of fitted values} = 1 - r^2$$
$$\text{Variance of y} = \text{Variance of residuals} + \text{Variance of fitted values}$$

Properties of residuals:
- Residuals are uncorrelated with the fitted values
- Mean of residuals is 0
- Variance of residuals: $\text{var}(r) = r^2$ (in standard units)
- SD of residuals: $\text{SD}(r) = \sqrt{r^2}$
- SD of residuals: $\text{SD}(r) = r \times \text{SD of } y$
- Relationship w/ fitted values: $\text{SD}(\text{residuals}) = \text{SD}(\text{fitted values}) / \sqrt{\text{SD}(y)}$
- Relationship w/ fitted values: $\text{SD}(\text{residuals}) / \text{SD}(y) = \sqrt{1 - r^2}$

Data 8 Final Study Guide — Page 2
Mean (or average): Balance point of the histogram
- Not the “half-way point” of the data; the mean is not the median unless the histogram is symmetric
- If the histogram is skewed, then the mean is pulled away from the median in the direction of the tail

Bootstrap the scatter plot & find the slope of the regression line through the bootstrapped plot many times.
- Draw the empirical histogram of all the resampled slopes.
- Get the “middle 95%” interval: that’s an approximate 95% confidence interval for the slope of the true line.
- Null hypothesis: The slope of the true line is 0.
- Construct a bootstrap confidence interval for the true slope.
- If the interval doesn’t contain 0, reject the null hypothesis.
- If the interval does contain 0, there isn’t enough evidence to reject the null hypothesis.

Fitted value: Height of the regression line at some $x$: $a \times x + b$.
Residual: Difference between $y$ and regression line height at $x$.
Regression model: $y$ is a linear function of $x$ + normal "noise"
Residual plot looks like a formless "noise" cloud under this model
Average of residuals is always 0 for any scatter diagram
Properties of fitted values, residuals, and the correlation $r$:
$$\text{Variance of residuals} / \text{Variance of fitted values} = 1 - r^2$$
$$\text{Variance of y} = \text{Variance of residuals} + \text{Variance of fitted values}$$

Properties of residuals:
- Residuals are uncorrelated with the fitted values
- Mean of residuals is 0
- Variance of residuals: $\text{var}(r) = r^2$ (in standard units)
- SD of residuals: $\text{SD}(r) = \sqrt{r^2}$
- SD of residuals: $\text{SD}(r) = r \times \text{SD of } y$
- Relationship w/ fitted values: $\text{SD}(\text{residuals}) = \text{SD}(\text{fitted values}) / \sqrt{\text{SD}(y)}$
- Relationship w/ fitted values: $\text{SD}(\text{residuals}) / \text{SD}(y) = \sqrt{1 - r^2}$