## Percentile rank

From Wikipedia, the free encyclopedia
Jump to navigationJump to search


Percentile ranks (PRs or percentiles) compared to Normal curve equivalents (NCEs)
The percentile rank of a score is the percentage of scores in its frequency distribution that are equal to or lower than it. For example, a test score that is greater than $75 \%$ of the scores of people taking the test is said to be at the 75th percentile, where 75 is the percentile rank. In educational measurement, a range of percentile ranks, often appearing on a score report, shows the range within which the test taker's "true" percentile rank probably occurs. The "true" value refers to the rank the test taker would obtain if there were no random errors involved in the testing process. ${ }^{[1]}$

Percentile ranks are commonly used to clarify the interpretation of scores on standardized tests. For the test theory, the percentile rank of a raw score is interpreted as the percentage of examinees in the norm group who scored at or below the score of interest. ${ }^{[2][3]}$

Percentile ranks are not on an equal-interval scale; that is, the difference between any two scores is not the same between any other two scores whose difference in percentile ranks is the same. For example, $50-25=25$ is not the same distance as $60-35=25$ because of the bell-curve shape of the distribution. Some percentile ranks are closer to some than others. Percentile rank 30 is closer on the bell curve to 40 than it is to 20 .

The mathematical formula is
where $\mathrm{c}_{\ell}$ is the count of all scores less than the score of interest, $f_{i}$ is the frequency of the score of interest, and $N$ is the number of examinees in the sample. If the distribution is normally distributed, the percentile rank can be inferred from the standard score.

Frequency distribution
From Wikipedia, the free encyclopedia

Jump to navigationJump to search

In statistics, a frequency distribution is a list, table or graph that displays the frequency of various outcomes in a sample. ${ }^{[1]}$ Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval.
$\ulcorner$

## Contents

- 1Example
- 2Construction
- 3Joint frequency distributions
- 4Applications
- 5See also
- 6Notes
- 7External links

Example[edit]

Here is an example of a univariate (single variable) frequency table. The frequency of each response to a survey question is depicted.

| Rank | Degree of agreement | Number |
| :--- | :--- | :--- |
| 1 | Strongly agree | 20 |
| 2 | Agree somewhat | 30 |
| 3 | Not sure | 20 |
| 4 | Disagree somewhat | 15 |
| 5 | Strongly disagree | 15 |
|  |  |  |

A different tabulation scheme aggregates values into bins such that each bin encompasses a range of values. For example, the heights of the students in a class could be organized into the following frequency table.


Example of a pie chart

A frequency distribution shows us a summarized grouping of data divided into mutually exclusive classes and the number of occurrences in a class. It is a way
of showing unorganized data notably to show results of an election, income of people for a certain region, sales of a product within a certain period, student loan amounts of graduates, etc. Some of the graphs that can be used with frequency distributions are histograms, line charts, bar charts and pie charts. Frequency distributions are used for both qualitative and quantitative data.

## Construction[edit]

1. Decide the number of classes. Too many classes or too few classes might not reveal the basic shape of the data set, also it will be difficult to interpret such frequency distribution. The maximum number of classes may be determined by formula: [base 10 or natural ${ }^{\text {logs?] }}$ or (approximately) where $n$ is the total number of observations in the data. (The latter will be much too large for large data sets such as population statistics.)
2. Calculate the range of the data (Range $=$ Max - Min) by finding the minimum and maximum data values. Range will be used to determine the class interval or class width.
3. Decide the width of the classes, denoted by $h$ and obtained by (assuming the class intervals are the same for all classes).

Generally the class interval or class width is the same for all classes. The classes all taken together must cover at least the distance from the lowest value (minimum) in the data to the highest (maximum) value. Equal class intervals are preferred in frequency distribution, while unequal class intervals (for example logarithmic intervals) may be necessary in certain situations to produce a good spread of observations between the classes and avoid a large number of empty, or almost empty classes. ${ }^{[2]}$

1. Decide the individual class limits and select a suitable starting point of the first class which is arbitrary; it may be less than or equal to the minimum value. Usually it is started before the minimum value in such a way that the midpoint (the average of lower and upper class limits of the first class) is properly ${ }^{[\text {clarification needed }]}$ placed.
2. Take an observation and mark a vertical bar (|) for a class it belongs. A running tally is kept till the last observation.
3. Find the frequencies, relative frequency, cumulative frequency etc. as required.

Joint frequency distributions[edit]
Bivariate joint frequency distributions are often presented as (twoway) contingency tables:

Two-way contingency table with marginal frequencies

|  | Dance | Sports | TV | Total |
| :--- | :--- | :--- | :--- | :--- |
| Men | 2 | 10 | 8 | 20 |
| Women | 16 | 6 | 8 | 30 |
| Total | 18 | 16 | 16 | 50 |

The total row and total column report the marginal frequencies or marginal distribution, while the body of the table reports the joint frequencies. ${ }^{[3]}$

Applications[edit]

Managing and operating on frequency tabulated data is much simpler than operation on raw data. There are simple algorithms to calculate median, mean, standard deviation etc. from these tables.

Statistical hypothesis testing is founded on the assessment of differences and similarities between frequency distributions. This assessment involves measures of central tendency or averages, such as the mean and median, and
measures of variability or statistical dispersion, such as the standard deviation or variance.

A frequency distribution is said to be skewed when its mean and median are significantly different, or more generally when it is asymmetric. The kurtosis of a frequency distribution is a measure of the proportion of extreme values (outliers), which appear at either end of the histogram. If the distribution is more outlier-prone than the Normal distribution it is said to be leptokurtic; if less outlier-prone it is said to be platykurtic.

Letter frequency distributions are also used in frequency analysis to crack ciphers, and are used to compare the relative frequencies of letters in different languages and other languages are often used like Greek,Latin, etc.

See also[edit]

- $\sqrt{\boldsymbol{x}}_{\text {Mathematics portal }}$
- Count data
- Cross tabulation
- Cumulative frequency
- Empirical distribution function


## Percentile

From Wikipedia, the free encyclopedia

Jump to navigationJump to search

## Not to be confused with Percentile function.

A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls. For example, the 20th percentile is the value (or score) below which $20 \%$ of the observations may be found. Similarly, $80 \%$ of the observations are found above the 20th percentile.
The term percentile and the related term percentile rank are often used in the reporting of scores from norm-referenced tests. For example, if a score is at the 86th percentile, where 86 is the percentile rank, it is equal to the value below which $86 \%$ of the observations may be found (carefully contrast with in the 86th percentile, which means the score is at or below the value below which $86 \%$ of the observations may be found-every score is in the 100th percentile). The 25th
percentile is also known as the first quartile $\left(Q_{1}\right)$, the 50th percentile as the median or second quartile $\left(Q_{2}\right)$, and the 75th percentile as the third quartile $\left(Q_{3}\right)$. In general, percentiles and quartiles are specific types of quantiles.

## Contents

- 1Applications
- 2The normal distribution and percentiles
- 3Definitions
- 4The nearest-rank method
- 4.1Worked examples of the nearest-rank method
- 5The linear interpolation between closest ranks method
- 5.1Commonalities between the variants of this method
- 5.2First variant,
- 5.2.1 Worked example of the first variant
- 5.3Second variant,
- 5.3.1 Worked examples of the second variant
- 5.4Third variant,
- 5.4.1 Worked example of the third variant
- 6The weighted percentile method
- 7See also
- 8References


## Applications[edit]

When ISPs bill "burstable" internet bandwidth, the 95th or 98th percentile usually cuts off the top $5 \%$ or $2 \%$ of bandwidth peaks in each month, and then bills at the nearest rate. In this way infrequent peaks are ignored, and the customer is charged in a fairer way. The reason this statistic is so useful in measuring data throughput is that it gives a very accurate picture of the cost of the bandwidth. The 95th percentile says that $95 \%$ of the time, the usage is below this amount: so, the remaining $5 \%$ of the time, the usage is above that amount.
Physicians will often use infant and children's weight and height to assess their growth in comparison to national averages and percentiles which are found in growth charts.
The 85th percentile speed of traffic on a road is often used as a guideline in setting speed limits and assessing whether such a limit is too high or low. ${ }^{[1][2]}$
In finance, Value at Risk is a standard measure to assess (in a model dependent way) the quantity under which the value of the portfolio is not expected to sink within a given period of time and given a confidence value.

The normal distribution and percentiles[edit]


Representation of the three-sigma rule. The dark blue zone represents observations within one standard deviation ( $\sigma$ ) to either side of the mean $(\mu)$, which accounts for
about $68.3 \%$ of the population. Two standard deviations from the mean (dark and medium blue) account for about $95.4 \%$, and three standard deviations (dark, medium, and light blue) for about $99.7 \%$.

The methods given in the Definitions section are approximations for use in smallsample statistics. In general terms, for very large populations following a normal distribution, percentiles may often be represented by reference to a normal curve plot. The normal distribution is plotted along an axis scaled to standard deviations,
or sigma ( ) units. Mathematically, the normal distribution extends to negative infinity on the left and positive infinity on the right. Note, however, that only a very small proportion of individuals in a population will fall outside the -3
to +3 range. For example, with human heights very few people are above the +3 height level.
Percentiles represent the area under the normal curve, increasing from left to right.
Each standard deviation represents a fixed percentile. Thus, rounding to two
decimal places, $-3 \quad$ is the 0.13 th percentile, -2 the 2.28 th percentile, -1
the 15.87th percentile, 0 the 50th percentile (both the mean and median of the distribution), +1 the 84.13th percentile, +2 the 97.72 nd percentile, and +3 the 99.87 th percentile. This is related to the 68-95-99.7 rule or the threesigma rule. Note that in theory the 0th percentile falls at negative infinity and the 100th percentile at positive infinity, although in many practical applications, such as test results, natural lower and/or upper limits are enforced.

## Definitions[edit]

There is no standard definition of percentile, ${ }^{[3][4][5]}$ however all definitions yield similar results when the number of observations is very large and the probability distribution is continuous. ${ }^{[6]}$ In the limit, as the sample size approaches infinity, the $100 p^{\text {th }}$ percentile $(0<p<1)$ approximates the inverse of the cumulative distribution function (CDF) thus formed, evaluated at $p$, as $p$ approximates the CDF. This can be seen as a consequence of the Glivenko-Cantelli theorem. Some methods for calculating the percentiles are given below.

## The nearest-rank method[edit]



The percentile values for the ordered list $\{15,20,35,40,50\}$

One definition of percentile, often given in texts, is that the $P$-th percentile of a list of $N$ ordered values (sorted from least to greatest) is the smallest value in the list such that no more than $P$ percent of the data is strictly less than the value and at least $P$ percent of the data is less than or equal to that value. This is obtained by first calculating the ordinal rank and then taking the value from the ordered list that corresponds to that rank. The ordinal rank $n$ is calculated using this formula

Note the following:

- Using the nearest-rank method on lists with fewer than 100 distinct values can result in the same value being used for more than one percentile.
- A percentile calculated using the nearest-rank method will always be a member of the original ordered list.
- The 100 th percentile is defined to be the largest value in the ordered list.

Worked examples of the nearest-rank method[edit]

## Example 1

Consider the ordered list $\{15,20,35,40,50\}$, which contains 5 data values. What are the 5th, 30th, 40th, 50th and 100th percentiles of this list using the nearest-rank method?

| $\begin{gathered} \text { Percentile } \\ P \end{gathered}$ | Number in list $N$ | Ordinal rank $n$ | Number from the ordered list that has that rank | Percentile value | Notes |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 5th | 5 |  | the first number in the ordered list, which is 15 | 15 | 15 is the smallest element of the list; $0 \%$ of the data is strictly less than 15 , and $20 \%$ of the data is less than or equal to 15 . |
| 30th | 5 |  | the 2 nd number in the ordered list, which is 20 | 20 | 20 is an element of the list |
| 40th | 5 |  | the 2 nd number in the ordered list, which is 20 | 20 | In this example it is the same as the 30th percentile. |
| 50th | 5 |  | the 3 rd number in the ordered list, which | 35 | 35 is an element of the ordered list. |


|  |  |  | is 35 |  |  |
| :--- | :--- | :--- | :---: | :---: | :---: |
| 100 th | 5 |  | 50, which <br> is the last <br> number in <br> the ordered <br> list | 50 | The 100th <br> percentile is <br> defined to be the <br> largest value in the <br> list, which is 50. |

So the 5 th, 30 th, 40 th, 50 th and 100 th percentiles of the ordered list $\{15,20,35$, $40,50\}$ using the nearest-rank method are $\{15,20,20,35,50\}$.

## Example 2

Consider an ordered population of 10 data values $\{3,6,7,8,8,10,13,15,16$, $20\}$. What are the 25th, 50th, 75th and 100th percentiles of this list using the nearest-rank method?

| Percentile <br> $\boldsymbol{P}$ | Number <br> in list <br> $\boldsymbol{N}$ | Ordinal <br> rank <br> $\boldsymbol{n}$ | Number <br> from the <br> ordered list <br> that has <br> that rank | Percentile <br> value | Notes |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 25th | 10 |  | the 3rd <br> number in <br> the ordered <br> list, which is <br> 7 | 7 | 7is an element of <br> the list |
| 50th | 10 |  | (he 5th <br> number in <br> the ordered <br> list, which is | 8 | 8 is an element of <br> the list. |


|  |  |  | 8 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 75th | 10 |  | the 8th <br> number in <br> the ordered <br> list, which is <br> 15 | 15 | 15 is an element <br> of the list. |
| 100th 10 | Last | 20, which is <br> the last <br> number in <br> the ordered <br> list | 20 | The 100th <br> percentile is <br> defined to be the <br> largest value in <br> the list, which is <br> 20. |  |

So the 25th, 50th, 75th and 100th percentiles of the ordered list $\{3,6,7,8,8$, $10,13,15,16,20\}$ using the nearest-rank method are $\{7,8,15,20\}$.

## Example 3

Consider an ordered population of 11 data values $\{3,6,7,8,8,9,10,13,15$, $16,20\}$. What are the 25 th, 50 th, 75 th and 100th percentiles of this list using the nearest-rank method?

| Percentile | Number <br> in list | Ordinal <br> rank <br> $\boldsymbol{n}$ | Number <br> from the <br> ordered list <br> that has <br> that rank | Percentile <br> value | Notes |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 25th | 11 |  | the 3rd <br> number in <br> the ordered <br> list, which is | 7 | 7 is an element of <br> the list |


|  |  |  | 7 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 50th | 11 |  | the 6th number in the ordered list, which is 9 | 9 | 9 is an element of the list. |
| 75th | 11 |  | the 9th number in the ordered list, which is 15 | 15 | 15 is an element of the list. |
| 100th | 11 | Last | 20 , which is <br> the last number in the ordered list | 20 | The 100th percentile is defined to be the largest value in the list, which is 20. |

So the 25th, 50th, 75 th and 100 th percentiles of the ordered list $\{3,6,7,8,8,9$, $10,13,15,16,20\}$ using the nearest-rank method are $\{7,9,15,20\}$.

The linear interpolation between closest ranks method[edit]
An alternative to rounding used in many applications is to use linear interpolation between adjacent ranks.
Commonalities between the variants of this method [edit]
All of the following variants have the following in common. Given the order statistics
we seek a linear interpolation function that passes through the points This is simply accomplished by
where uses the floor function to represent the integral part of positive , whereas uses the mod function to represent its fractional part (the remainder after division by 1). (Note that, though at the endpoint , is undefined, it does not need to be because it is multiplied by .) As we can see, is the continuous version of the subscript , linearly interpolating between adjacent nodes. There are two ways in which the variant approaches differ. The first is in the linear relationship between the rank , the percent rank , and a constant that is a function of the sample size :

There is the additional requirement that the midpoint of the range , corresponding to the median, occur at :
and our revised function now has just one degree of freedom, looking like this:

The second way in which the variants differ is in the definition of the function near the margins of the range of : should produce, or be forced to produce, a result in the range , which may mean the absence of a one-to-one correspondence in the wider region.

First variant, [edit]


The result of using each of the three variants on the ordered list $\{15,20,35,40,50\}$
(Sources: Matlab "prctile" function, ${ }^{[7][8]}$ )
where

Furthermore, let

The inverse relationship is restricted to a narrower region:

## Worked example of the first variant[edit]

Consider the ordered list $\{15,20,35,40,50\}$, which contains five data values. What are the 5th, 30th, 40th and 95th percentiles of this
list using the Linear Interpolation Between Closest Ranks method? First, we calculate the percent rank for each list value.

| List <br> value <br> Position <br> of that <br> value <br> in the <br> ordered <br> list | Number <br> of values | Calculation <br> of <br> percent rank | Percent <br> rank, | Notes |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 15 | 1 | 5 |  | 10 |  |
| 20 | 2 | 5 |  | 30 |  |
| 35 | 3 | 5 |  | 50 |  |
| 40 | 4 | 5 |  | 70 |  |
| 50 | 5 | 5 |  | 90 |  |
| 5 |  |  |  |  |  |
| 4 |  | 5 |  |  |  |

Then we take those percent ranks and calculate the percentile values as follows:

| Percent rank | $\begin{gathered} \text { Number } \\ \text { of } \\ \text { values } \end{gathered}$ | Is <br> ? | Is ? | Is there a percent rank equal to ? | What do we use for percentile value? | Percentile value | Notes |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 5 | Yes | No | No | We see that $\mathrm{P}=5$, which is less than the first percent rank $\mathrm{p} 1=10$, so use the first list value v 1 , which is 15 | 15 | 15 is a member of the ordered list |
| 30 | 5 | No | No | Yes | We see that $\mathrm{P}=30$ is the same as the second percent rank p2 $=30$, so use the second list value v2, which is 20 | 20 | 20 is a member of the ordered list |


| 40 | 5 | No | No | No | We see that $\mathrm{P}=40$ is between percent rank p2=30 and p3 $=50$, so we take $\mathrm{k}=2$, $k+1=3$, $\mathrm{P}=40$, $\mathrm{pk}=\mathrm{p} 2=30$, $\mathrm{vk}=\mathrm{v} 2=20$, $v k+1=v 3=35$, $\mathrm{N}=5$. <br> Given those values we can then calculate v as follows: | 27.5 | 27.5 is <br> not a member of the ordered list |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 95 | 5 | No | Yes | No | We see that $\mathrm{P}=95$, which is greater than the last percent rank $\mathrm{pN}=90$, so use the last list value, which is 50 | 50 | 50 is a <br> member of the ordered list |

So the 5th, 30th, 40th and 95th percentiles of the ordered list \{15, 20, 35, 40, 50\} using the Linear Interpolation Between Closest Ranks method are \{15, 20, 27.5, 50\}

## Second variant, [edit]

(Source: Some software packages, including NumPy ${ }^{[9]}$ and Microsoft Excel ${ }^{[5]}$ (up to and including version 2013 by means of the PERCENTILE.INC function). Noted as an alternative by NIST ${ }^{[10]}$ )

Note that the relationship is one-to-one for , the only one of the three variants with this property; hence the "INC" suffix, for inclusive, on the Excel function.

## Worked examples of the second variant[edit]

## Example 1:

Consider the ordered list $\{15,20,35,40,50\}$, which contains five data values. What is the 40th percentile of this list using this variant method?
First we calculate the rank of the 40th percentile:
So, $x=2.6$, which gives us and . So, the value of the 40th percentile is

## Example 2:

Consider the ordered list $\{1,2,3,4\}$ which CONTAINS four data values. What is the 75th percentile of this list using the Microsoft Excel method?
First we calculate the rank of the 75th percentile as follows:

So, $x=3.25$, which gives us an integral part of 3 and a fractional part of 0.25 . So, the value of the 75th percentile is

## Third variant, [edit]

(The primary variant recommended by NIST. ${ }^{[10]}$ Adopted by Microsoft Excel since 2010 by means of PERCENTIL.EXC function. However, as the "EXC" suffix
indicates, the Excel version excludes both endpoints of the range of $p$, i.e., , whereas the "INC" version, the second variant, does not; in fact, any number smaller than $1 /(\mathrm{N}+1)$ is also excluded and would cause an error.)

The inverse is restricted to a narrower region:

Worked example of the third variant [edit]Consider the ordered list $\{15,20$, $35,40,50\}$, which contains five data values. What is the 40th percentile of this list using the NIST method?
First we calculate the rank of the 40th percentile as follows:
So $x=2.4$, which gives us and . So the value of the 40th percentile is calculated as:

So the value of the 40 th percentile of the ordered list $\{15,20,35,40,50\}$ using this variant method

## 

Missing in \%న్నూడ

- Automatic translation
- Contribute

 ซైళడాఱారు)





 బळుఖః $\dot{\operatorname{N}} \mathrm{భ}$












 గణణితద ヴ๐




ఇదేనె $\Omega_{2}$ నిలలడి



## లుల్లియవగేళు








## 

నేమ్ళుల్లి లుళిదేవెరిగి అంళిఅంఠగేళు!


- صునె



- అంశిఅంత బอగో





##  

 ఱరపిడి

1. రైడాలారు


2. $\quad$.

## 












 లుత్తెఱుఱాగిది.





## 









 ळుుడుచు త్తిద్దినఱె.


 లుదలळరరణ్యయ




30 1

33
2

## 




छرలణణ $=25 / 100 *(8+1)=0.25 * 9=2.25$.



ळంత 2: ๗ల్య2ల్యన 1 అథఱల 2 అన్నె ఆరిసి:
 లో(యృంం).



 யjయి్ని $\vec{\sim} J త ్ త ద ి: ~$



 $(0.25)(43-33)=2.5$




 ఎరాడు బలర ఱరిరిలిసుఎపుదు జలణతన.

## 
















అల2నేయేన్ను ळృడ్రుః
ळుడృృช Kannada

 ిిఖారేసుు మోడడేలాదే గి,లడో-ळొభు

Install our FREE Android app on Google Play !




