

Sampling, Regression, Experimental Design and
Analysis for Environmental Scientists,
Biologists, and Resource Managers

C. J. Schwarz

Department of Statistics and Actuarial Science, Simon Fraser University
cschwarz@stat.sfu.ca

January 16, 2011

Contents

101Medians and Percentiles; Prediction and Tolerance Intervals	2
101.1Median	7
101.1.1 Estimation	7
101.1.2 Confidence intervals	7
Non-parametric interval	8
Parametric interval	11
101.2Percentiles	11
101.2.1 Estimation	11
101.2.2 Confidence intervals for percentiles	13
Non-parametric interval	13
Parametric interval	17

Chapter 101

Medians and Percentiles; Prediction and Tolerance Intervals

An analyst computes that the mean daily averaged phosphorus reading in a river is 12.5 mg/L. How reliable is this estimate? As you saw earlier, the precision of this estimate can be quantified though the standard error (*se*) of the estimate and the computation of a 95% confidence interval.

The sample mean is perhaps the most popular measure of central tendency that is applied to data. However, the mean and standard deviation are both sensitive to outliers, and a few outlying points can severely distort the estimate, standard error, and 95% confidence intervals. Is there another measure of central tendency that is more robust?

As you will see in this chapter, the *median* is another measure of central tendency that is more robust to outliers and odd distributions than the simple mean. You will also see that the median is a special case of a *percentile*, a general way to describe locations of data.

Based on the actual data observed, can you make predictions about the next value? For example, given the daily NO values observed in January, what can you say is a likely range of values for future days in January?

Given a range of data, how likely are extreme values to occur? For example, an estimate of the 100-year flood, the 99th percentile of annual flood peaks, was determined to be 10,000 cubic feet per second (cfs). Assuming that the

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

choice of a particular distribution to model these floods (Log Pearson Type III) is correct, what is the reliability of this estimate? The Canada Wide Standard on $PM_{2.5}$ is based on the 98th percentiles of daily averaged values. How are these computed and interpreted?

The results in this chapter will be illustrated using the data on annual peak discharges for the Saddle R. at Lodi, NJ from 1925 to 1967.¹

Here is the raw data. It is also available in the *flow* tab in the *ALLOf-DATA.xls* workbook in the Sample Program Library at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>.

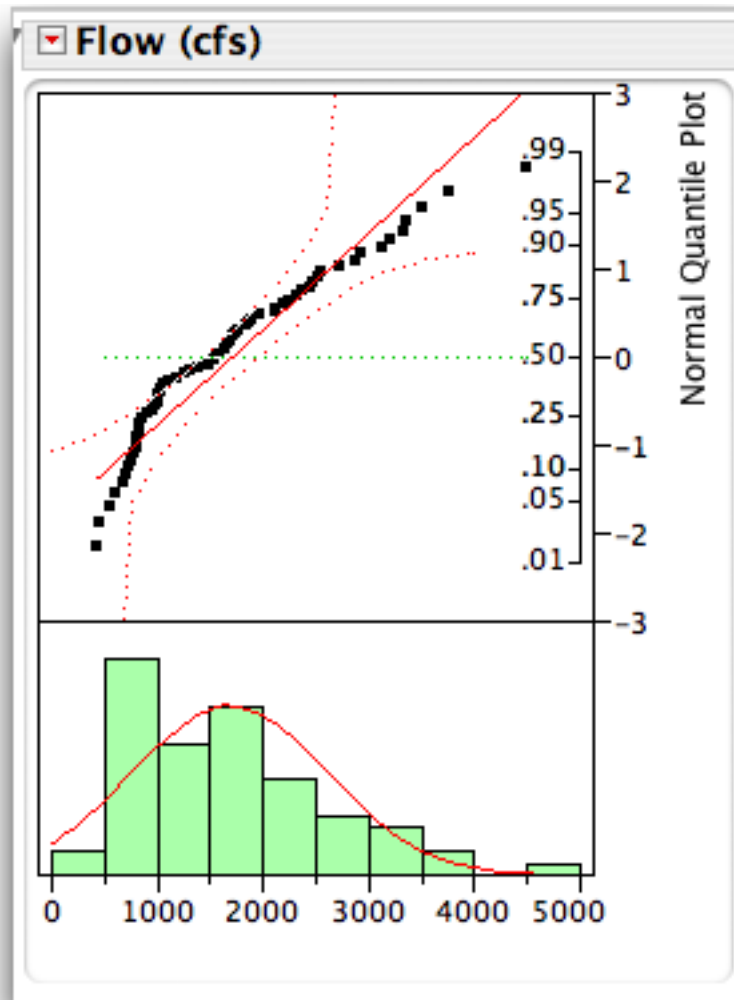
Annual Peak Discharges for Saddle River, NJ	
Year	Flow (cfs)
1925	980
1926	741
1927	1630
1928	829
1929	903
1930	418
1931	549
1932	686
1933	1320
1934	850
1935	614
1936	1720
1937	1060
1938	1680
1939	760
1940	1380
1941	1030
1942	820
1943	1020
1944	998
1945	3500
1946	1100
1947	1010
1948	830
1949	1030
1950	452
1951	2530
1952	1740
1953	1860

¹Data available as dataset C1 from Helsel and Hirsch (2002) Statistical Methods in Water Resources available at <http://pubs.usgs.gov/twri/twri4a3/>

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND
TOLERANCE INTERVALS

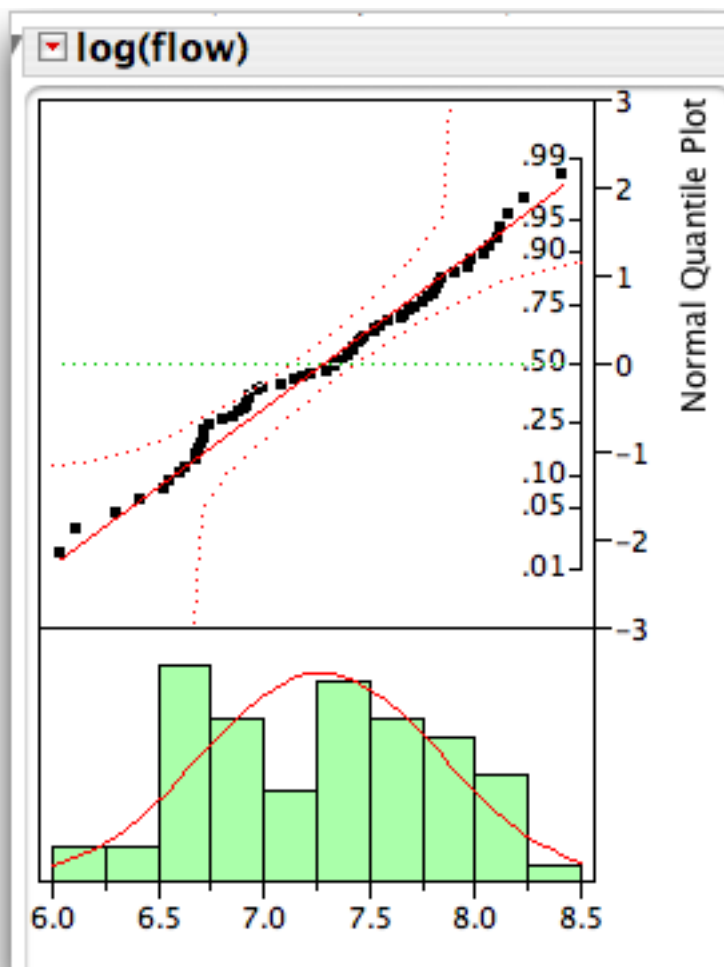
1954	1270
1955	2200
1956	1530
1957	795
1958	1760
1959	806
1960	1190
1961	952
1962	1670
1963	824
1964	702
1965	1490
1966	1600
1967	800
1968	3330
1969	1540
1970	2130
1971	3770
1972	2240
1973	3210
1974	2940
1975	2720
1976	2440
1977	3130
1978	4500
1979	2890
1980	2470
1981	1900
1982	1980
1983	2550
1984	3350
1985	2120
1986	1850
1987	2320
1988	1630
1989	2380

A preliminary plot of the data shows evident skewness and the normal probability plot (see earlier sections of the notes) shows good evidence of a poor fit to a normal distribution:



Data such as annual maximums are often analyzed using a type of Extreme Value distribution, but in many cases there is little difference from fitting a log-normal distribution, i.e. the logarithms of the values are fit to a normal distribution. This seems to fit the data well except perhaps in the extreme upper tails: ²

²In many cases, these are the exact quantities of interest and so a good fit to the upper tail is crucial!



A formal test for normality (the Shapiro-Wilk test, not shown) shows no evidence of non-normality on the log-scale. Consequently, the $\log(\text{flow})$ will be used in the remainder of this section.

101.1 Median

101.1.1 Estimation

The sample median is simply the data point that lies in the middle of the data (after sorting). These can be easily computed by hand, but more often are computed using a computer.

The median is computed by:

- sorting the data from smallest to largest
- if n is odd then median is $Y[(n+1)/2]$ if n is even, median is $(Y[(n/2)] + Y[n/2 + 1])/2$,³ i.e. the middle value if n is odd, or the average of the two middle values if n is even

The median is valid for ordinal, interval, or ratio data. It is relatively insensitive to outliers and can also be computed with censored data.

For example:

- The median of 13, 11, 17 is 13.
- The median of 13, 11, 17, 15 is $(13 + 15)/2 = 14$.
- The median of 13, 11, 20234234 is still 13.
- The median of 13, 11, <5 , <5 , 9⁴ is 9.

101.1.2 Confidence intervals

As with confidence intervals for means, it is possible to compute confidence intervals for medians. Both a non-parametric and a parametric confidence interval can be computed for the median - however, unless you have a fairly large sample size, you may find that the non-parametric confidence intervals are not very narrow.

³Note that $Y[n]$ is a shorthand notation for the n^{th} value from smallest to largest after sorting.

⁴The notation <5 means that the value is less than 5, but the exact value is unknown. If you are using a computer package, you may have to specify an actual value for measurements below the detection limit so that computer know how to sort them. A “rule of thumb” is to use half of the detection limit in these cases.

Non-parametric interval

The non-parametric confidence interval for the median is based on using the Binomial distribution. For small sample sizes (say $n \leq 20$), an exact Binomial table should be used.⁵ For larger sample sizes (say $n > 20$), the following approximation will usually be sufficient.

- Compute $R_{lower} = n(.5) - z\sqrt{n(.5)(1-.5)}$ rounded to the nearest integer.
- Compute $R_{upper} = n(.5) + z\sqrt{n(.5)(1-.5)}$ rounded to the nearest integer.
- The confidence interval is then $Y[R_{lower}]$ to $Y[R_{upper}]$.

The value of z depends upon the confidence level required. For a 95% confidence interval use $z = 1.96$ or $z = 2$. Note that the .5 values occur in the above formulae because you are interested in the median which turns out to be the 50th percentile or the 0.5 quantile. These formulae will reappear later in this chapter with suitable modifications.

For example, if you have 50 observations, the median is the average of the 25th and 26th observations, and the computations for a 95% confidence interval above reduce to:

- $R_{lower} = 50(.5) - 1.95\sqrt{50(.5)(.5)} = 25 - 6.92 = 18$
- $R_{upper} = 50(.5) + 1.95\sqrt{50(.5)(.5)} = 25 + 6.92 = 32$
- The 95% confidence interval will range from the 18th to the 32nd observation.

While most computer packages will compute medians (and other percentiles) only a few will compute confidence intervals for medians.

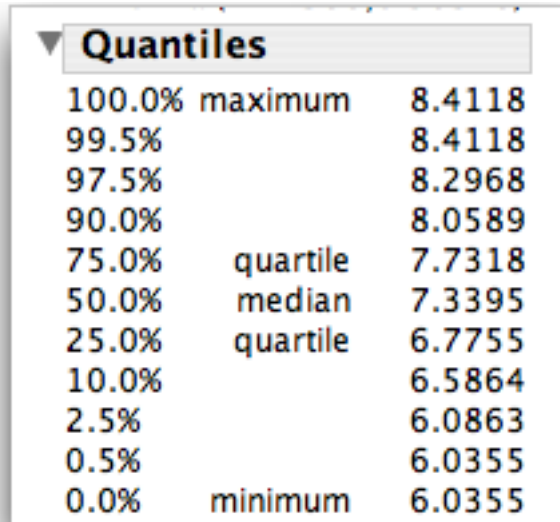
JMP analysis

The data and JMP scripts are available in the *flow.jmp* file in the Sample Program Library at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>. The median can be computed using *Analyze->Distribution* platform, but confidence intervals for the median (and other percentiles) are not available.

⁵However for small sample sizes, the interval will be so wide as typically not to be very useful, and so is not demonstrated here.

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

For example, the *Analyze->Distribution* platform applied to the *log_flow* data gives:

A screenshot of a SAS window titled 'Quantiles'. It displays a list of percentiles from 0.0% to 100.0% along with their corresponding numerical values. The 50.0% percentile is labeled as the 'median', the 75.0% as the 'quartile', and the 0.0% as the 'minimum'.

Percentile	Label	Value
100.0%	maximum	8.4118
99.5%		8.4118
97.5%		8.2968
90.0%		8.0589
75.0%	quartile	7.7318
50.0%	median	7.3395
25.0%	quartile	6.7755
10.0%		6.5864
2.5%		6.0863
0.5%		6.0355
0.0%	minimum	6.0355

The median $\log(\text{flow})$ is 7.3395.

SAS analysis

The median can be computed using many of the SAS procedures such as *Proc Means*, *Proc Univariate*, but confidence intervals for the median (and other percentiles) are available through *Proc Capability*. The SAS program is available in the *flow.sas* file in the Sample Program Library available at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>. The following code fragment:

```
/* test for normality and other attributes of original and logged data */
/* The cipctlDF      requests confidence intervals for percentiles that
   are distribution free
   cipctlNORMAL requests confidence intervals based on assumption
   of normality. Both of these can be requested for
   single sided, two sided, asymmetric etc.
   See the SAS documentation for more details.
   normaltest      requests test for normality */

proc capability data=flow normaltest gout=graph cipctldf cipctlnormal;
  title2 'Use log(flow) as the response variable';
  var logflow;
```

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

```
qqplot logflow / normal; /* create qq plot */
interval logflow / k=1 methods=(1,3) ;
```

was used to compute various statistics including confidence intervals for the median *log_flow*. This gives:

Quantiles (Definition 5)					
		95% Confidence Limits		----Order Statistics-----	
Quantile	Distribution Free	LCL Rank	UCL Rank	Coverage	
50% Median	6.96602 7.47307	25	41	95.36	

The *Definition 5* for the computation of quantiles is based the definition used previously in the notes. We find that the distribution-free 95% confidence interval for the median *log_flow* ranges from 6.967 to 7.473 which correspond to discharges of $e^{6.967} = 1060$ to $e^{7.473} = 1760$ cfs.

The actual coverage of this interval is estimated to be just over 95% (as expected).⁶

SYStat analysis

The data are available in the *flow* tab of the *ALLofDATA.xls* workbook in the Sample Program Library at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>. The median can be computed using Analysis→DescriptiveStatistics platform, but confidence intervals for the median (and other percentiles) are not available.

For example, the platform applied to the *log_flow* data gives:

⁶As you will see later in the notes, it is sometimes impossible to get a distribution free confidence interval for an extreme percentile (near 0 or 1) to have the proper coverage.

Method = CLEVELAND	
1 %	6.04721
5 %	6.39202
10 %	6.60800
20 %	6.71720
25 %	6.79060
30 %	6.90575
40 %	7.04239
50 %	7.33954
60 %	7.45586
70 %	7.65917
75 %	7.72300
80 %	7.80586
90 %	8.04879
95 %	8.12767
99 %	8.38528

The median $\log(\text{flow})$ is 7.3395.

Parametric interval

In some cases, a parametric interval for the median can be computed.

For example, if you are willing to assume that data comes from a log-normal distribution, then you can compute a confidence interval for the mean on the log-scale and back-transform the intervals.

It is also possible to compute a confidence interval assuming that the original data comes from a normal distribution. See the section on confidence intervals for percentiles later in these notes for details.

101.2 Percentiles

101.2.1 Estimation

A percentile is a measure of relative standing against all other people. The p^{th} percentile has at least $p\%$ of the values are less than or equal to that point and

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

at least $(100 - p)\%$ of the data values are greater than or equal to that point. For example, if the 98th percentile for $PM_{2.5}$ is 10, then 98% of all daily values are less than or equal to 10.

There are many ways of computing percentiles, and computer packages differ in which method they use. The two most common methods are the

- The $Y[np+]$ rule which means to take the next observation above np if np is not an integer, and the average of this and the next observation if np is an integer. Percentiles are normally computed only for interval or ratio data.
- The extrapolation rule which means that rather than averaging two neighboring observations, you extrapolate between the two observations depending on the value of the fractional part of np . For example, if $np = 4.75$, the percentile is found as 75% of the way between the 4th and 5th observation.

Percentiles are normally computed only for interval or ratio data.

For example, if a dataset had 75 observations, then the 25th percentile would be the $Y[(.25)(75)+] = Y[19]$, i.e., the 19th smallest observation after sorting. The 40th percentile would be $Y[(.40)(75)+] = (Y[30] + Y[31])/2$, i.e. the average of the 30th and 31st smallest observations after sorting.

Some percentiles are given special names. The median is the 50th percentile; the first quartile is the 25th percentile, the third quartile is the 75th percentile.

Important Don't confuse percentiles with percentages. A percentile is only related to the relative standing of an observation compared to other data values. So, a person who scores 75% on a test but whose score is the 40th percentile, means that only 40% of students did worse than this student, and 60% of students did better.

Percentiles are only valid for ordinal, interval, or ratio scaled data. If using percentiles for ordinal data, you have to be a little careful if you need to interpolate between two values because this is not well defined.

Most computer packages display standard percentiles and many also allow you to request special percentiles.

101.2.2 Confidence intervals for percentiles

Quantiles or percentiles have often been used to describe rare events. For example, a 100-year flood is the 99th percentile (0.99 quantile) of the distribution of annual flood peaks. It is the magnitude of flood which is expected to be exceeded only once in 100 years. The 20-year flood is the value that is expected to be exceeded only once in 20 years, or is the 95th percentile of annual peaks. Often parametric forms are used to estimate these percentiles. The log Pearson Type III is often used in the United States while European countries have used the Gumbel (extreme value) distribution, though the generalized extreme value distribution is now more common.

It is also becoming more common to express regulatory rules in terms of percentiles. For example, if the 98th percentile is exceeded, some action must be taken.

But, these percentiles are computed from raw data, and usually the uncertainty in these values has not been quantified.

It is possible to compute confidence intervals for percentiles.⁷ These should not be confused with prediction intervals which are ranges for future values of individual events.

Non-parametric interval

A non-parametric confidence interval for a percentile was already seen previously – the confidence interval for the median was a confidence interval for the the same basic results can be used. Let p be the percentile of interest. Then for large samples⁸ compute

- $R_{lower} = n(p) - z\sqrt{n(p)(1-p)}$
- $R_{upper} = n(p) + z\sqrt{n(p)(1-p)}$
- The confidence interval is then $Y[R_{lower}]$ to $Y[R_{upper}]$.

The value of z depends upon the confidence level required. For a 95% confidence interval use $z = 1.96$ or $z = 2$.

⁷Some article erroneously call these tolerance intervals. This is not quite correct as you will see in a later section.

⁸Use Binomial tables for small samples as shown in Conover (1999), *Non-parametric statistics*.

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

For example, suppose you have 500 values and wish to find a 95% confidence interval for the 99th percentile. The estimated percentile is the $500(.99) = 495^{th}$ value. Then $z = 1.96$ and:

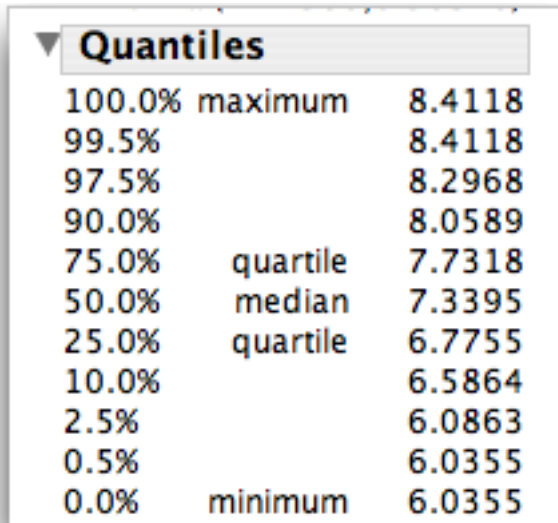
- $R_{lower} = 500(.99) - 1.96\sqrt{500(.99)(1 - .99)} = 495 - 4.36 = 490.64$
- $R_{upper} = 500(.99) + 1.96\sqrt{500(.99)(1 - .99)} = 495 + 4.36 = 499.36$
- The confidence interval is then $Y[490.64]$ to $Y[499.36]$.

Some interpolation may be required to find $Y[490.64]$ and $Y[499.36]$.

Unless sample sizes are large, you may find that the non-parametric confidence intervals are so large as to be not useful. It is also possible that for extreme percentiles (i.e. close to 0 or to 1) that it is impossible to find a non-parametric confidence interval that has a 95% coverage.

JMP analysis

The data and JMP scripts are available in the *flow.jmp* file in the Sample Program Library at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>. Various percentiles are automatically generated using the *Analyze* > *Distribution* platform:



▼ Quantiles		
100.0%	maximum	8.4118
99.5%		8.4118
97.5%		8.2968
90.0%		8.0589
75.0%	quartile	7.7318
50.0%	median	7.3395
25.0%	quartile	6.7755
10.0%		6.5864
2.5%		6.0863
0.5%		6.0355
0.0%	minimum	6.0355

The estimated 10-year return period flood is estimated to be 8.0589 (on the *log* scale) which corresponds to 3160 cfs. No confidence intervals are available for

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

percentiles.

SAS analysis

Proc Capability can be used to compute confidence intervals for percentiles. The SAS program is available in the *flow.sas* file in the Sample Program Library available at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>. The same basic code fragment is used:

```
/* The cipctlDF      requests confidence intervals for percentiles that are distribution free
   cipctlNORMAL      requests confidence intervals based on assumption of normality
   Both of these can be requested for single sided,
   two sided, asymetric etc - see the SAS documentation */
proc capability data=flow normaltest gout=graph cipctldf cipctlnormal;
  title2 'Use log(flow) as the response variable';
  var logflow;
```

This produces output:

Quantiles (Definition 5)					
95% Confidence Limit ---Order Statistics---					
Quantile	Distribution Free		LCL Rank	UCL Rank	Coverage
99%	8.16052	8.41183	63	65	45.21
95%	8.04879	8.41183	59	65	92.11
90%	7.83597	8.23483	54	64	96.49
75% Q3	7.52833	7.96901	43	57	95.03
50% Median	6.96602	7.47307	25	41	95.36
25% Q1	6.67834	6.93731	9	23	95.03
10%	6.11368	6.70930	2	12	96.49
5%	6.03548	6.60800	1	7	92.11
1%	6.03548	6.30810	1	3	45.21

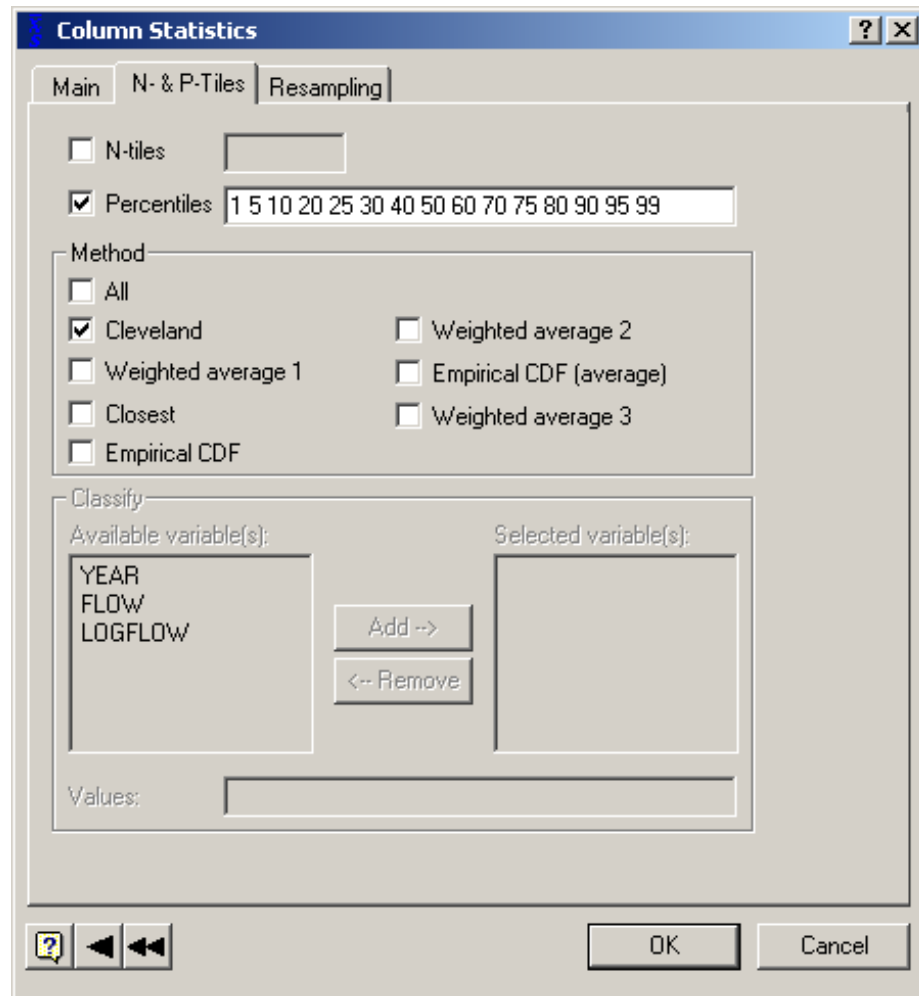
These are interpreted in exactly the same way as for the median. For example, we are 95% confident that the 90th percentile (which corresponds to the 10 year return period) is between 7.836 and 8.235 (on the *log* scale) which corresponds to between 2530 and 3770 cfs on the original scale. Note that coverage for the non-parametric confidence interval for the 99th and 1st percentiles is far below the nominal 95% coverage and should be used cautiously.

SYStat analysis

The data are available in the *flow* tab of the *ALLOfDATA.xls* workbook in the Sample Program Library at <http://www.stat.sfu.ca/~cschwarz/Stat-650/>

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

Notes/MyPrograms. Various percentiles can be computed using the *N- and P-tiles* tab of the *Analysis→DescriptiveStatistics* platform



There are number of ways to compute the percentiles - look at the SYStat documentation for more details. This gives:

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

Method = CLEVELAND	
1 %	6.04721
5 %	6.39202
10 %	6.60800
20 %	6.71720
25 %	6.79060
30 %	6.90575
40 %	7.04239
50 %	7.33954
60 %	7.45586
70 %	7.65917
75 %	7.72300
80 %	7.80586
90 %	8.04879
95 %	8.12767
99 %	8.38528

The estimated 10-year return period flood is estimated to be 8.0589 (on the *log* scale) which corresponds to 3160 cfs. No confidence intervals are available for percentiles in SYStat.

Parametric interval

In cases with moderate sample sizes, the non-parametric confidence intervals are typically very wide. It is common practice then to compute confidence intervals for the percentiles assuming a distribution (e.g. normal, or log-normal) for the data. These parametric confidence intervals can be considerably more precise, but of course the danger is that if your assumption about the distribution is wrong, then the confidence intervals are also wrong.

These notes will find confidence intervals for percentiles based on a normal distribution.⁹

A review article on computing confidence intervals for percentiles from a normal distribution is found at:

⁹If data are skewed, they may follow a log-normal distribution. In this case, the logarithms of the data now follow a normal distribution. Use the methods of this section on the logarithm of the values, and then simply back transform the endpoints of the intervals using the anti-logarithm

CHAPTER 101. MEDIANs AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

Chakraborti S. and Li J. (2007).
 Confidence Interval Estimation of a Normal Percentile.
 American Statistician 61, 331-336.
<http://dx.doi.org/10.1198/000313007X244457>

The estimated percentile is found by working backwards from a normal distribution. In large samples, this is easily found as:

$$\widehat{Y}_p = \bar{Y} + zs$$

where \widehat{Y}_p is the estimated percentile, \bar{Y} is the sample mean, s is the sample standard deviation, and z is the appropriate percentage point from a standard normal curve, with a TOTAL probability p of being less than z . For example,

p	z
.50	0.00
.75	0.67
.80	0.84
.90	1.28
.95	1.65
.98	2.05
.99	2.33

These values should NOT be confused with the multipliers used for confidence intervals for means of a normal distribution obtained from tables. For small samples, the sample standard deviation (s) is a biased estimator of σ ¹⁰ For small samples, a small adjustment is made for the bias. The adjusted estimate of the percentile is:

$$\bar{Y}_p = \bar{Y} + z \frac{s}{M(df)}$$

where $M(df)$ is the mean of a standardized chi-distribution¹¹ with df degrees of freedom, and

$$M(df) = \frac{\sqrt{2}\Gamma[(df+1)/2]}{\sqrt{df}\Gamma(df/2)}$$

The confidence interval for the percentile must take into account the uncertainty in \bar{Y} and s . The interval takes the form of

$$(\bar{Y} - ks, \bar{Y} + ks)$$

¹⁰Note that s^2 is unbiased for σ^2 , but because the standard deviation is a non-linear function of s^2 , it is not unbiased.

¹¹A standardized chi-distribution random variable is generated by taking a χ_{df}^2 random variable, dividing by the df and taking the square root.

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

where k is taken from a non-central t -distribution.¹² Stedinger (1983)¹³ has tables of the k values.

Computation of these intervals by hand is tedious at best, and can rarely be done without a computer. Gerow and Bielen (1999)¹⁴ have a nice article on the application of these intervals as well in a fisheries context showing how to use modern software to extract the necessary information using simulation to estimate the various quantities.

JMP analysis

The data and JMP scripts are available in the *flow.jmp* file in the Sample Program Library at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>.

The parametric percentiles are found by first using the *Fit Distribution* of the *Analyze->Distribution* platform:

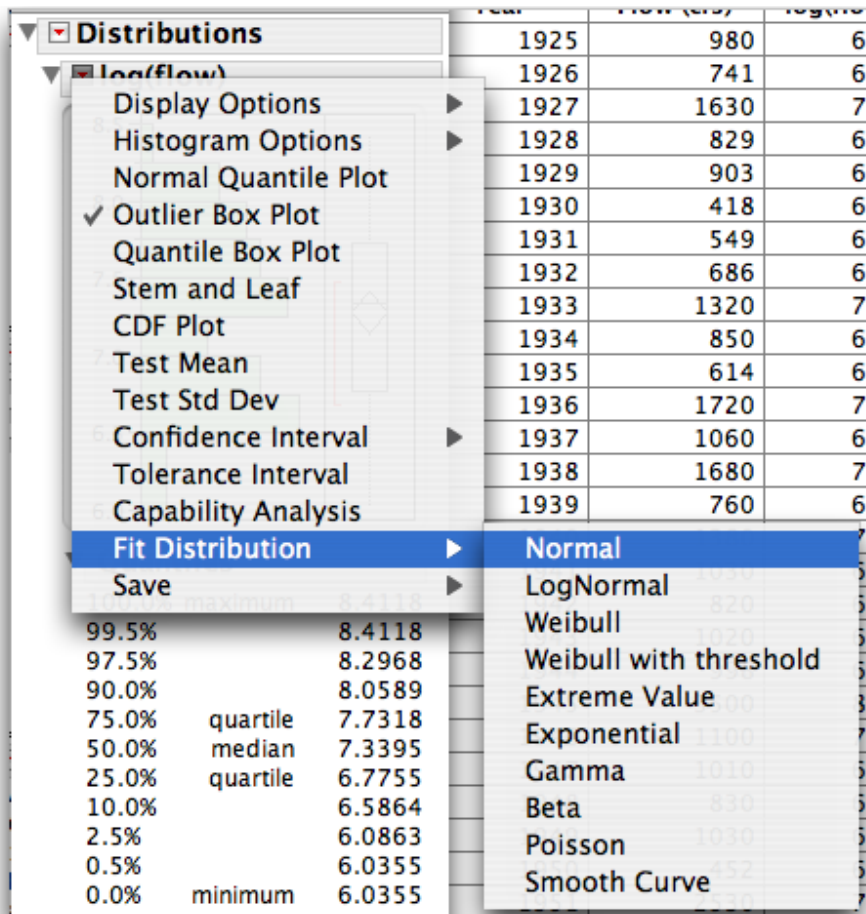
¹²In particular, k is the $(1 - \alpha/2)$ percentile of a non-central t distribution with $z_p \times \sqrt{n}$ noncentrality parameter and $n - 1$ degrees of freedom, divided by \sqrt{n} , *i.e.*

$$k = \frac{t(z_p \times \sqrt{n}, n - 1)_{1-\alpha/2}}{\sqrt{n}}$$

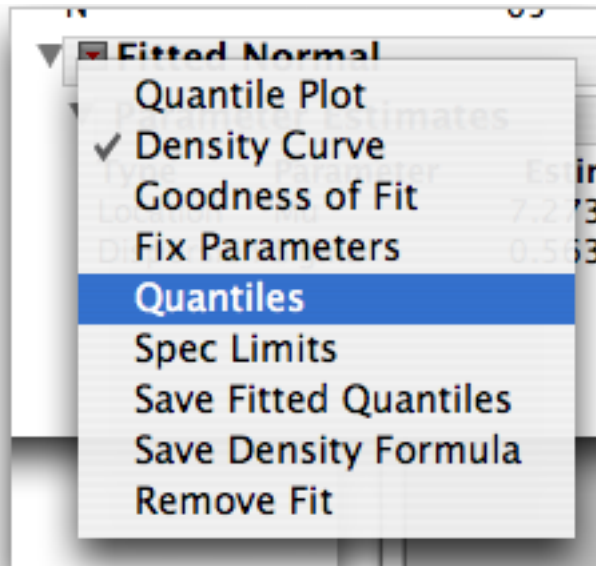
¹³Stedinger, J. R., 1983, Confidence intervals for design events: Journal of Hydraulic Eng., ASCE 109, 13-27. [http://dx.doi.org/10.1061/\(ASCE\)0733-9429\(1983\)109:1\(13\)](http://dx.doi.org/10.1061/(ASCE)0733-9429(1983)109:1(13))

¹⁴Gerow, K. and Bielen, Confidence Intervals for Percentiles: An Application to Estimation of Potential Maximum Biomass of Trout in Wyoming Streams. North American Journal of Fisheries Management 19, 149-151. [http://dx.doi.org/10.1577/1548-8675\(1999\)019<0149:CIFPAA>2.0.CO;2](http://dx.doi.org/10.1577/1548-8675(1999)019<0149:CIFPAA>2.0.CO;2)

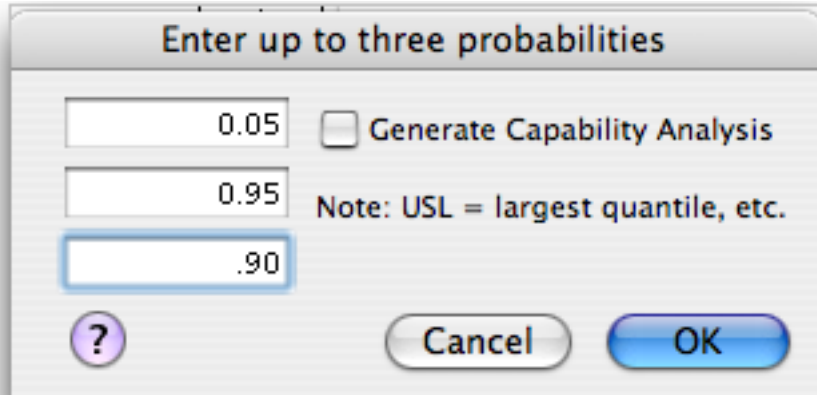
CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS



followed by requesting various percentiles:



finally followed by requesting the quantiles of interest:



which gives:

Quantiles: Uncentered and Unscaled	
Percentile	Quantile
0.0500000	6.347067
0.9000000	7.995993
0.9500000	8.200701

The 90th percentile (corresponding to a 10-year return period) is estimated to be 7.996 (on the *log* scale) which corresponds to 2970 cfs on the original scale.

Confidence intervals for percentiles are not available in JMP.

SAS analysis

The SAS program is available in the *flow.sas* file in the Sample Program Library available at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>. The code fragment seen earlier will compute the parametric confidence intervals for the percentiles assuming a normal distribution (the *cipctlnormal* keyword):

```
/* The cipctlDF      requests confidence intervals for percentiles that
                        are distribution free.
      cipctlNORMAL requests confidence intervals based on assumption
                        of normality. Both of these can be requested for
                        single sided, two sided, asymmetric etc.
                        See the SAS documentation */
proc capability data=flow normaltest gout=graph cipctldf cipctlnormal;
  title2 'Use log(flow) as the response variable';
  var logflow;
```

gives the following output:

Quantiles (Definition 5)			
		95% Confidence Limits	
Quantile	Estimate	Assuming Normality	
99%	8.41183268	8.35463683	8.89855321
95%	8.11671562	8.01476793	8.44696503
90%	8.04878828	7.83011137	8.20956044
75% Q3	7.71423114	7.51176755	7.82197067
50% Median	7.33953770	7.13426422	7.41350375
25% Q1	6.80572255	6.72579730	7.03600042

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

10%	6.60800063	6.33820753	6.71765660
5%	6.41999493	6.10080294	6.53300004
1%	6.03548143	5.64921476	6.19313114

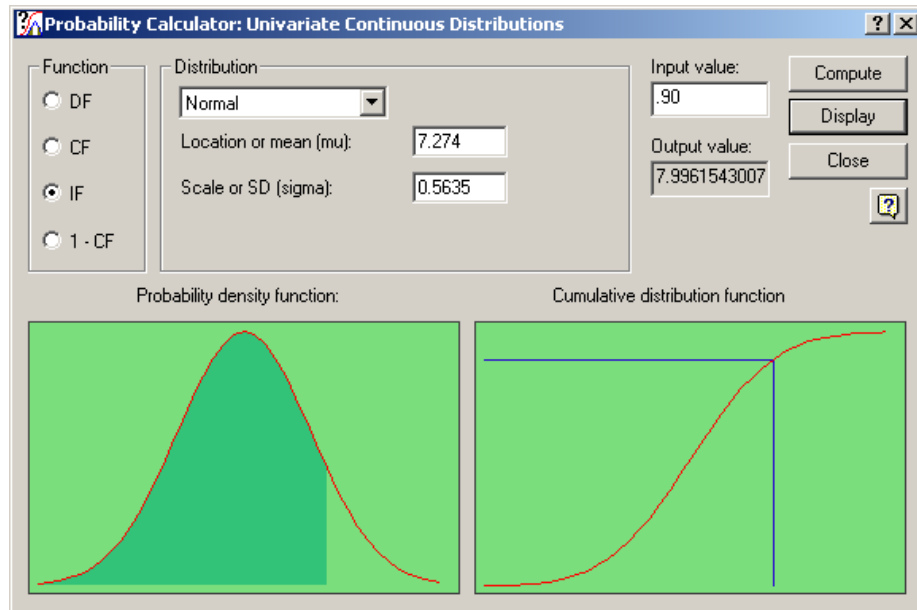
Based on a normal distribution, we are 95% confident that the 90th percentile (corresponding to a 10-year return period) lies between 7.830 and 8.210 on the *log* scale which corresponds to 2514 to 3680 cfs on the original scale.

These intervals are very sensitive to the assumption of normality, particularly for the more extreme percentiles!

SYStat analysis

In order to estimate the quantiles, you must first estimate the mean and standard deviation from the raw data. Based on 65 data values, the sample mean on *log* scale is 7.27388 and the standard deviation is .56346.

Use the *Utilities*→*ProbCalculator*→*UnivariateCont* platform to get the dialogue box where you enter the mean and standard deviation. Then select the *IF=Inverse Function* button and specify the desired quantile (.90). After pressing the *Compute* button the estimated percentile is displayed:



Based on a the fitted normal distribution, the 90th percentile is estimated to be 799 on the *log* scale which corresponds to 2970 cfs on the original scale.

SYStat is unable to compute confidence intervals for these percentiles.

101.3 Prediction Intervals

101.3.1 Introduction

A major source of confusion to many people is that confidence intervals do NOT say anything about future observations.¹⁵ Confidence intervals are ranges of reasonable values for the true population MEAN; there is no relationship to individual values.

Prediction intervals give ranges for future individual observations. Prediction intervals are computed for a different purpose than confidence intervals – they deal with individual data values as opposed to a summary statistic such as the mean. A prediction interval is wider than the corresponding confidence interval, because an individual observation is more variable than is a summary statistic computed from several observations. Unlike a confidence interval, a prediction interval takes into account the variability of single data points around the median or mean, in addition to the error in estimating the center of the distribution.

A “quick and dirty” prediction interval is often computed as the mean \pm two standard DEVIATIONS. However, because both the mean and the standard deviation are based on data and are subject to uncertainty. Consequently, despite being a nominal 95% prediction interval, the interval actually has lower coverage than expected.

101.3.2 Non-parametric interval

This is simply found by finding the middle 95% of the data and using the endpoints of this interval. This is done more formally as:

- Let α represent the 1-confidence level. For example, for a 95% confidence interval, $\alpha = .05$.
- Compute $R_{lower} = (n + 1)\alpha/2$ rounded to the nearest integer.
- Compute $R_{upper} = (n + 1)(1 - \alpha/2)$ rounded to the nearest integer.

¹⁵This same confusion arises in regression settings where there are confidence intervals for the mean response at a new X and prediction intervals for individual values at a new X .

- The prediction interval ranges from $Y[R_{lower}]$ to $Y[R_{upper}]$.

For example, if you have 500 observations, then a 95% prediction interval is found as

- Let α represent the 1-confidence level. For example, for a 95% confidence interval, $\alpha = .05$.
- Compute $R_{lower} = (501)(.05)/2 = 13$
- Compute $R_{upper} = (501)(1 - (.05)/2) = 489$
- The prediction interval ranges from $Y[13]$ to $Y[489]$.

Alternatively, given a set of data, how confident am I that the next observation will fall between the minimum and maximum values? ¹⁶ Rather surprisingly, this question has a simple result that are true regardless of the underlying distribution.

You will be $\frac{n-1}{n+1} \times 100\%$ confident that the next future observation will fall between the minimum and maximum of the observed data.

For example, suppose that 50 observations are taken with a minimum of 42.017 and a maximum of 46.050. Then you can be $\frac{50-1}{50+1} \times 100\% = 96.1\%$ confident that the next independent observation will lie between 42.017 and 46.050.

For the flow data with 65 years of data, you will be $\frac{65-1}{65+1} \times 100\% = 97\%$ confident that the next future observation will lie between the minimum *log_flow* of 6.0354 and the maximum *log_flow* of 8.4118. Or after back transforming, you will be 97% confident that the next future observation will be between 425 and 4500 cfs.

I am unaware of any computer package that does these computations automatically.

101.3.3 Parametric interval

Parametric prediction intervals can also be computed to give a range of future observations. However, you must now make strong assumptions about the particular distribution which need to be verified before these procedures are used.

¹⁶These assume that the observations are independent of each other. If autocorrelation exists, the results presented here are not valid.

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

Unlike confidence intervals for the mean which will be valid in large samples regardless of the distribution from which the data are collected because of the Central Limit Theorem in Statistics, there is no equivalent large sample result and the distribution is crucial. In other parts of the notes, you way how to verify different distributions though the use of quantile plots.

The most common assumption is that the data follow a normal distribution. Prediction intervals are then constructed to be symmetric around the sample mean, and account for uncertainty in both the estimated mean and standard deviation. The interval is computed as:

$$\bar{Y} \pm t \sqrt{s^2 \left(1 + \frac{1}{n}\right)}$$

where t represents the appropriate value from a t -distribution (for small sample sizes) or a normal distribution for large sample sizes, s is the sample standard DEVIATION, and n is the sample size. The term $(1 + \frac{1}{n})$ is what accounts for uncertainty in the mean because you took a sample, and the t value accounts for uncertainty in the standard deviation. For large samples, a reasonable approximation for a 95% prediction interval is to use $t = 2$.

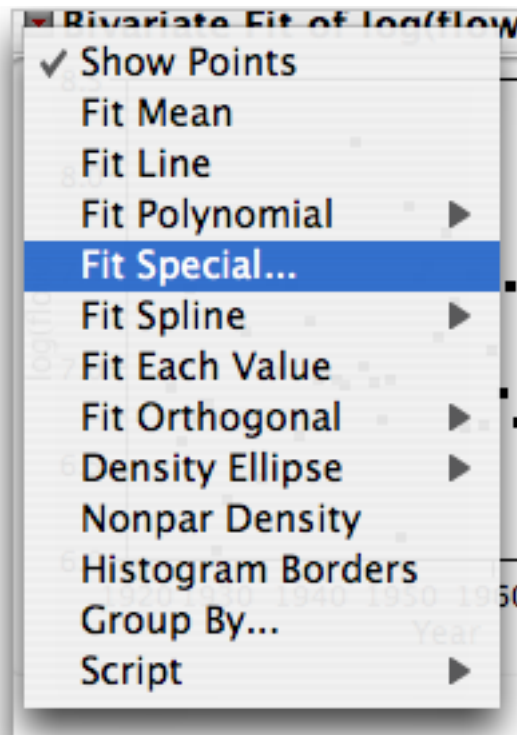
It is also possible to make prediction intervals based on other distributions (e.g. extreme values, etc). This is beyond the scope of these notes.

JMP analysis

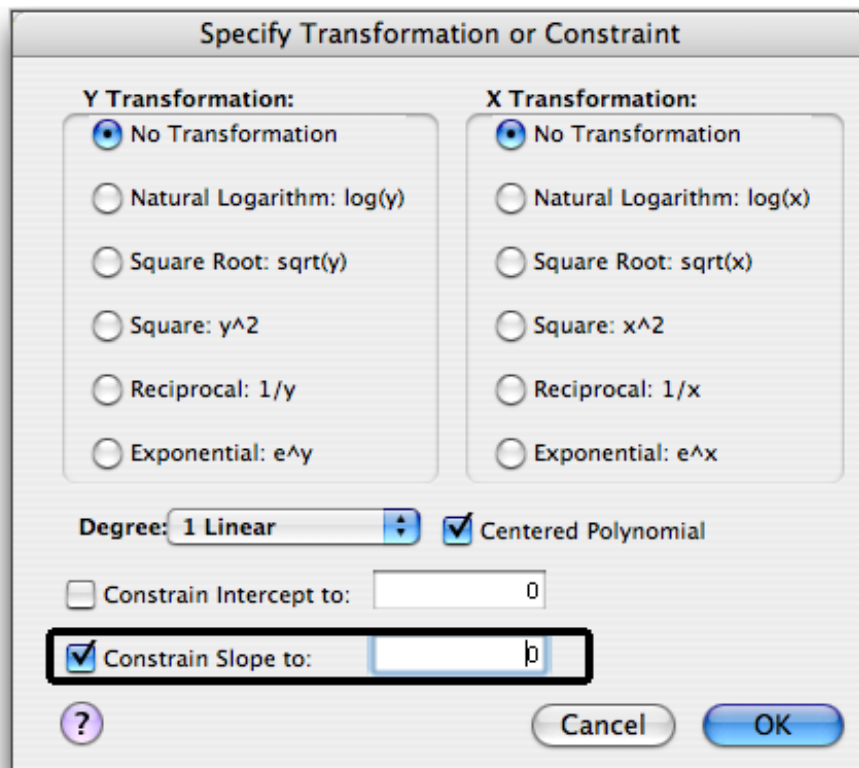
The data and JMP scripts are available in the *flow.jmp* file in the Sample Program Library at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>.

While a simple prediction interval is not directly available from the *Analyze->Distribution* platform, JMP can be “tricked” into obtaining a prediction interval.

Use the *Analyze->Fit Y-by-X* platform with Y specified a *log_flow* and another variable specified as the X variable. Then use the *Fit Special* option:



to fit a line with a 0-constrained slope:



The dialog box is titled "Specify Transformation or Constraint". It contains two main sections: "Y Transformation:" and "X Transformation:". Each section has a "No Transformation" option selected by default, along with other options like "Natural Logarithm: log(y)", "Square Root: sqrt(y)", "Square: y^2", "Reciprocal: 1/y", and "Exponential: e^y". Below these sections, there is a "Degree:" dropdown menu set to "1 Linear", a checked "Centered Polynomial" checkbox, and two checkboxes for constraints: "Constrain Intercept to:" (unchecked) and "Constrain Slope to:" (checked). The "Constrain Slope to:" checkbox is highlighted with a black border, and its corresponding input field contains the letter "b". At the bottom, there are buttons for "?", "Cancel", and "OK".

Specify Transformation or Constraint

Y Transformation:

- ☒ No Transformation
- ☐ Natural Logarithm: $\log(y)$
- ☐ Square Root: \sqrt{y}
- ☐ Square: y^2
- ☐ Reciprocal: $1/y$
- ☐ Exponential: e^y

X Transformation:

- ☒ No Transformation
- ☐ Natural Logarithm: $\log(x)$
- ☐ Square Root: \sqrt{x}
- ☐ Square: x^2
- ☐ Reciprocal: $1/x$
- ☐ Exponential: e^x

Degree: 1 Linear

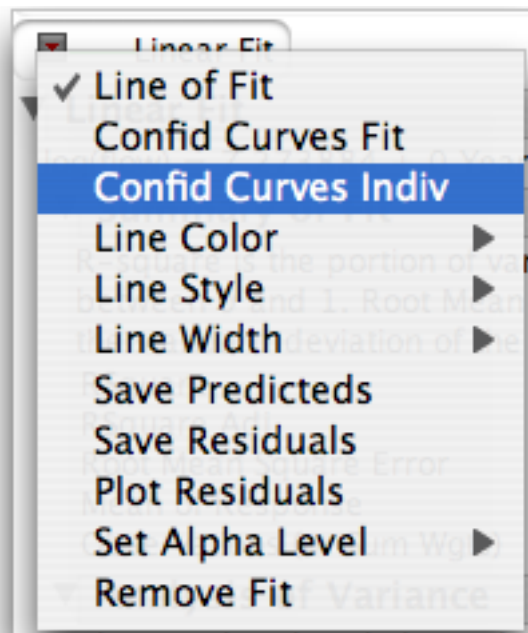
☒ Centered Polynomial

☐ Constrain Intercept to: 0

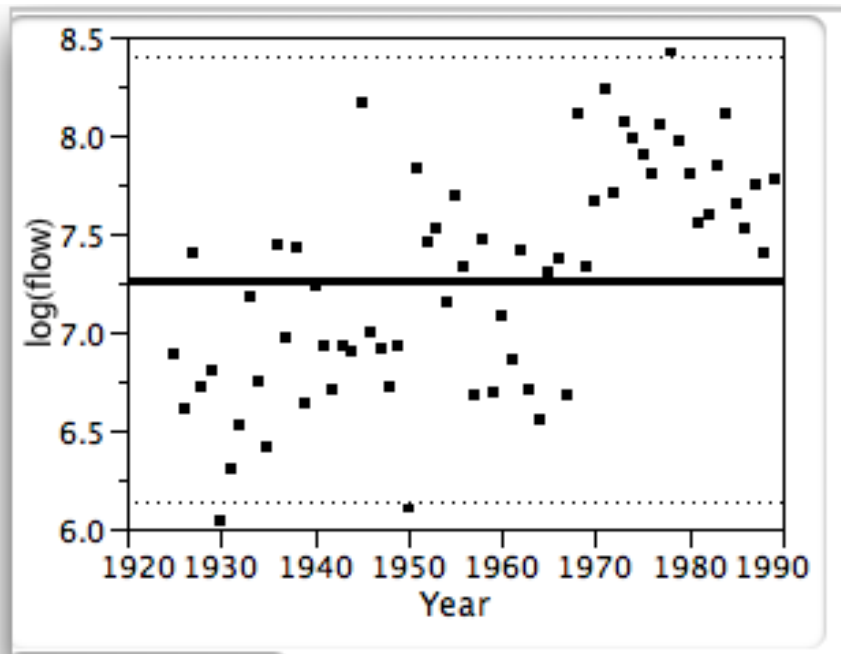
☒ Constrain Slope to: b

? Cancel OK

This will fit the mean to the data, and then use the *Confidence Indiv* fitting option to get a prediction interval for the next observation:



which creates a plot with the individual prediction limits:



Reading off the graph, the 95% prediction interval for the next observation would be from (6.14 to 8.40) on the *log* scale.

When these notes were being put together, I did not first check to see if the series was stationary over time until I tried this trick in JMP. The graph clearly shows that the maximum yearly discharge has been increasing over time and the regression is statistically significant. Sigh....That is what happens when you fail to practice what you preach, i.e. always graph the data before doing any formal analyses. I'm afraid that all the neat computations on the stream discharge data don't much sense in this chapter but it is too late in the day to go back and change everything! Sigh!!!!

SAS analysis

Proc Capability can provide prediction intervals for the next future observation assuming normality. We earlier saw that there was no evidence of non-normality for the *log_flow* based on the Shapiro-Wilk statistic.

The SAS program is available in the *flow.sas* file in the Sample Program Library available at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>. The *interval* statement in following code fragment request a 95%

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

¹⁷ prediction interval (method=1) for a single ($k = 1$) observation:

```
proc capability data=flow normaltest gout=graph cipctldf cipctlnormal;
  title2 'Use log(flow) as the response variable';
  var logflow;
  qqplot logflow / normal; /* create qq plot */
  /* the following interval statement asks for prediction intervals
     (method=1) for k=1 future observations */
  interval logflow / k=1 methods=(1,3) ;
```

This gives:

```
Two-Sided Statistical Intervals for logflow Assuming Normality
Approximate Prediction Interval
Containing All of k
Future Observations
Confidence      k  Prediction Limits
99.00%          1   5.767    8.781
95.00%          1   6.140    8.408
90.00%          1   6.326    8.222
```

We are 95% confident that the next future observation will have a *log_flow* between 6.14 and 8.40 which correspond to 464 to 4482 cfs on the untransformed scale. **These prediction intervals are VERY SENSITIVE to the assumption of normality!**

SYStat analysis. SYStat is unable to compute prediction intervals.

101.4 Tolerance Intervals

Finally, the prediction interval for a single future observation can be extended. These are known as tolerance intervals, and are typically expressed as being (say) 95% confident that 80% of future observations are between two values. ¹⁸

A simple rule of thumb can be established based on the minimum and maximum of a set of data:

¹⁷Other confidence levels are possible as well. Consult the SAS documentation

¹⁸One sided tolerance intervals can also be found, but this is beyond the scope of these notes.

CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

You will be $(1 - p^n - n(1 - p)p^{n-1}) \times 100\%$ confident that at least the fraction p of the future observations will lie between the minimum and maximum of the observed data.

If you are interested in knowing the tolerance level for $p = .95$, then you are $(1 - .95^{50} - 50(1 - .95).95^{50-1}) \times 100\% = 72\%$ confident that at least 95% of future observations will lie between 42.017 and 46.050.

For the stream discharge data, there were 65 observation with minimum and maximum values of 6.0354 and 8.4118 respectively on the *log* scale. Suppose we wish a tolerance interval for 90% of future values. Then we are $(1 - .90^{65} - 65(1 - .90)(.90)^{65-1}) \times 100\% = 99\%$ confident that 90% of future observations are between the minimum and maximum values.

Not surprisingly, much work has been done when data come from a normal distribution but comparable intervals are available for many other distributions as well. In the case of a normal distribution, these intervals will take the form

$$\bar{Y} \pm ks$$

where k is obtained from tables, s is the sample standard deviation, and \bar{Y} is the sample mean.¹⁹ A nice web-based java script for computing tolerance intervals from a normal distribution is available at <http://statpages.org/tolintvl.html> **The tolerance intervals are very sensitive to the distribution chosen!**

For example, suppose that we wish to be 95% confident about the next 90% of future observations from the stream discharge data. There were 65 observations with a sample mean of 7.2738 and the sample standard deviation was .5625 (on the *log* scale). The tolerance interval is found as:

If I measure a sample consisting of items,
 and get a mean value of
 and a standard deviation of
 then I can be % certain
 that % of the population
 lies within the interval to (a Two-sided Tolerance Interval)
 from:

¹⁹See for example, <http://www.itl.nist.gov/div898/handbook/prc/section2/prc263.htm> for the actual formulae.

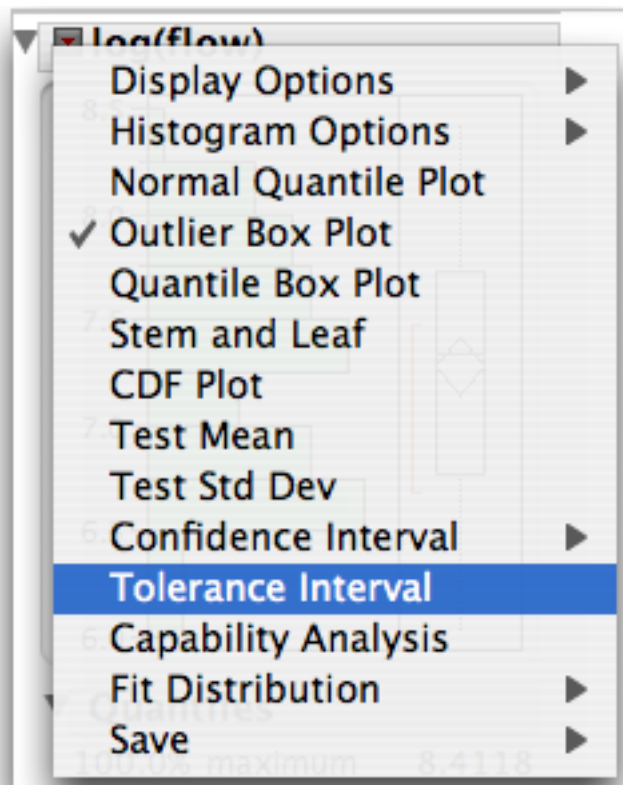
CHAPTER 101. MEDIAN AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS

So we are 95% confident that the next 90% of observations will lie between 6.1794 and 8.3684 (on the *log* scale).

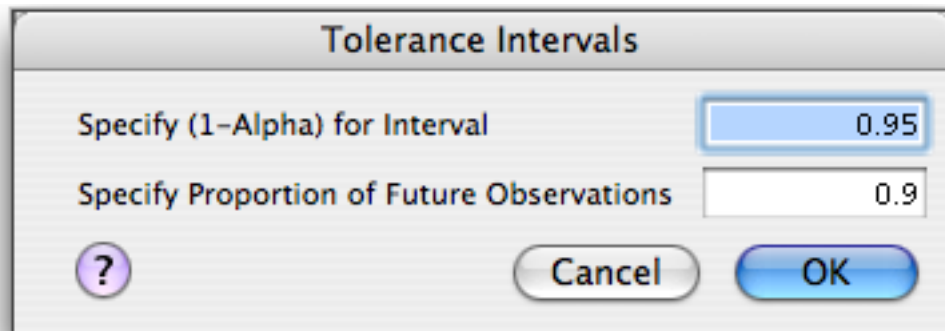
JMP analysis

The data and JMP scripts are available in the *flow.jmp* file in the Sample Program Library at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>.

A tolerance interval is available in the *Analyze->Distribution* platform:



which request the confidence level and the proportion of future observations of interest:



This gives:

Tolerance Intervals					
Parameter	Estimate	Lower TI	Upper TI	1-Alpha	Proportion
Mean	7.273884	6.178312	8.369456	0.950	0.900

We are 95% confident that 90% of future observations will lie between 6.178 and 8.37 on the *log* scale.

SAS analysis

The SAS program is available in the *flow.sas* file in the Sample Program Library available at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>. The same code fragment as use earlier also computes tolerance intervals (the *method=3* option on the *interval* statement:

```
proc capability data=flow normaltest gout=graph cipctldf cipctlnormal;
  title2 'Use log(flow) as the response variable';
  var logflow;
  /* the following interval statement asks for prediction intervals
     (method=1) for k=1 future observations.
     method=3 requests a tolerance interval */
  interval logflow / k=1 methods=(1,3) ;
```

This gives:

Approximate Tolerance Interval
Containing At Least Proportion

CHAPTER 101. MEDIANs AND PERCENTILES; PREDICTION AND TOLERANCE INTERVALS ---

p of the Population			
Confidence	p	Tolerance Limits	
99.00%	0.900	6.102	8.446
99.00%	0.950	5.877	8.670
99.00%	0.990	5.439	9.109
95.00%	0.900	6.179	8.368
95.00%	0.950	5.970	8.578
95.00%	0.990	5.560	8.988
90.00%	0.900	6.217	8.331
90.00%	0.950	6.015	8.533
90.00%	0.990	5.619	8.929

So we are 95% confident that the next 90% of observations will lie between 6.1794 and 8.3684 (on the *log* scale).

SYStat analysis. SYStat is unable to compute tolerance intervals.