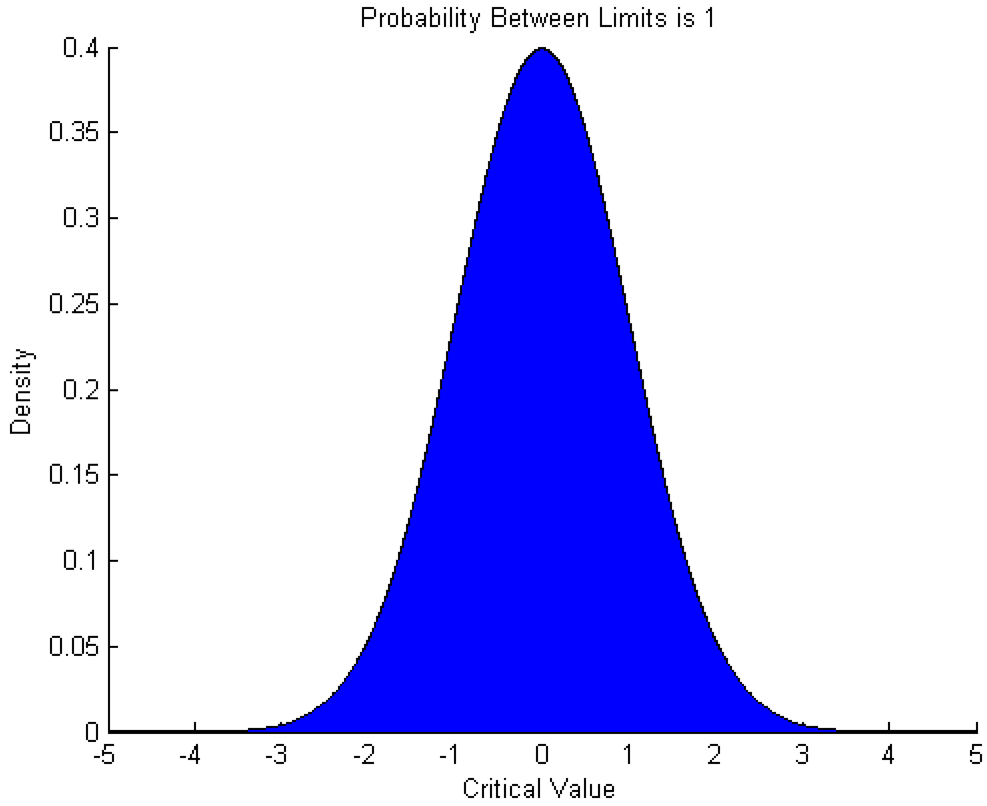
Assignments

- Undergrads: Dept and CMES scholarship deadline this Friday
- Turn in in-class Assignment
- Wednesday in MLIB again
- Time to Work on Probability Assignment (#8) today, due Wednesday
- Read Chapter 3 and 4 notes in Canvas (Chapter 5 also available)
- Read text: Text Chapter 3.1-3.6, Chapter 4
- Will discuss Assignment #10 on Wednesday

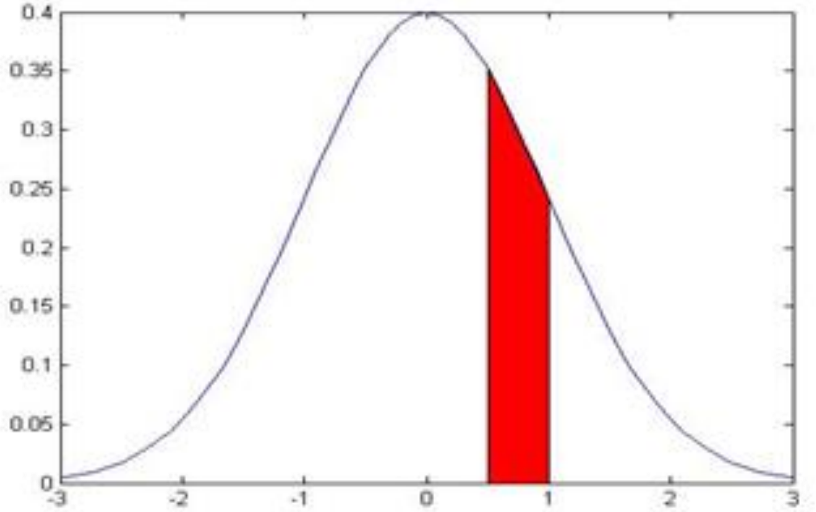
Empirical vs. Parametric Distributions

- Parameteric distributions:
  - Theoretical approach to define populations with known properties
  - Can be defined by a function with couple parameters and assumption that population composed of random events



Random Continuous Variable x

- f(x) probability density function (PDF) for a random continuous variable x
- f(x)dx incremental contribution to total probability

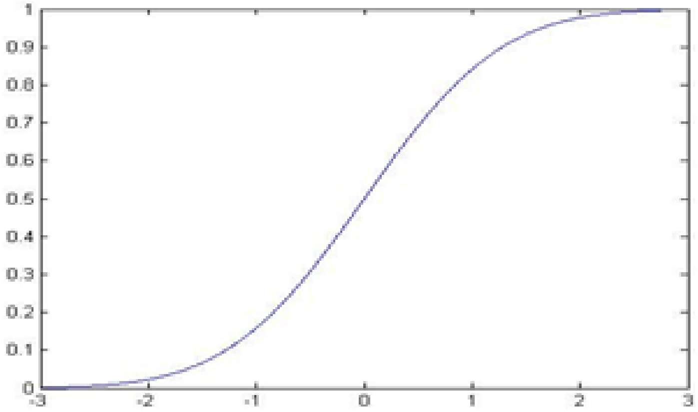


$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Cumulative Density Function of Continuous Variable

- F(X)- total probability below a threshhold
- F(0) = 50%
- F(.66) = 75%
- X(F) – quantile function- value of random variable corresponding to particular cumulative probability
- X(75%) = 0.66

$$F(X) = \Pr\{x \le X\} = \int_{-\infty}^{X} f(x)dx$$



Gaussian Parametric Distribution

- PDF $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$
- CDF $$F(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{X} \exp(-\frac{(x-\mu)^2}{2\sigma^2})dx$$

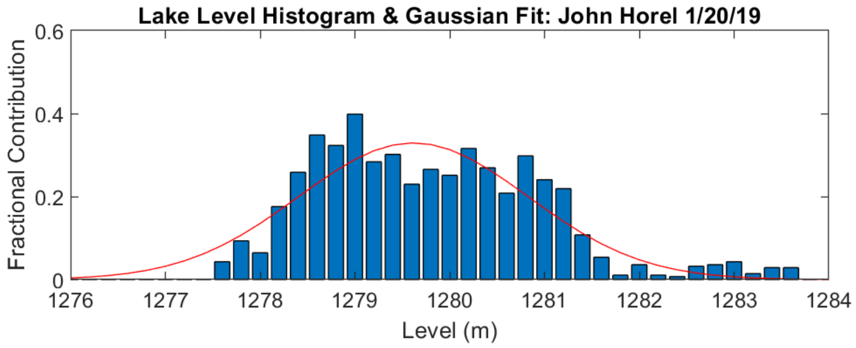
- Two parameters define Gaussian distribution: μ and σ
- Nothing magic or “normal” about the Gaussian distribution- it is a mathematical construct

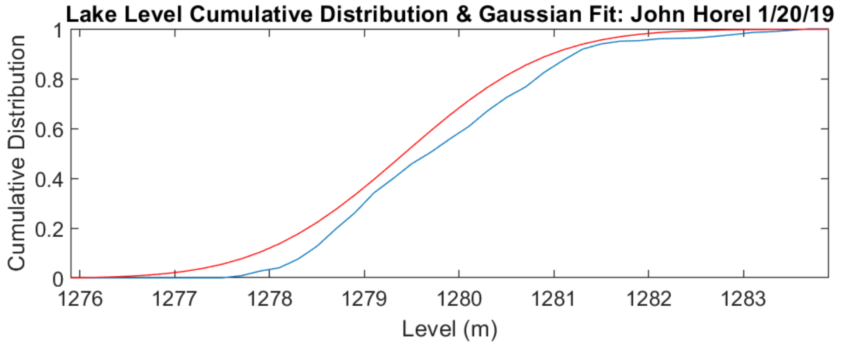
Using parametric distributions

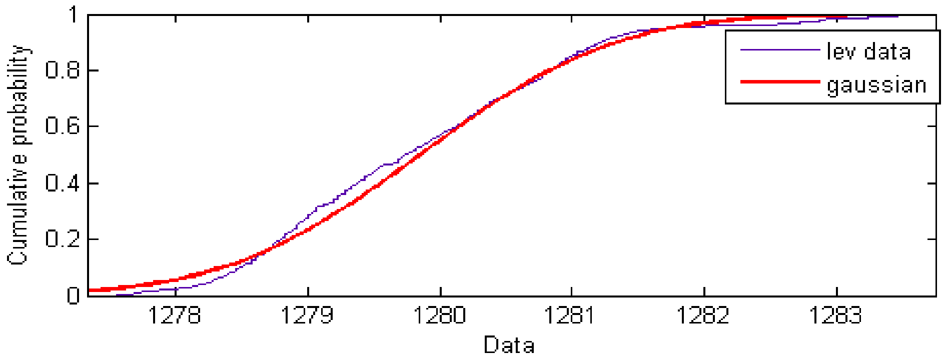
- Generate an empirical cumulative probability (CDF)
- Use dfittool to see if there is a good match between the empirical CDF and a particular parametric distribution
- Use the parameters from that parametric distribution to estimate the probabilities of values above below a threshhold or extreme events

Lake Level

- Fit using sample mean and sample estimate of population std dev



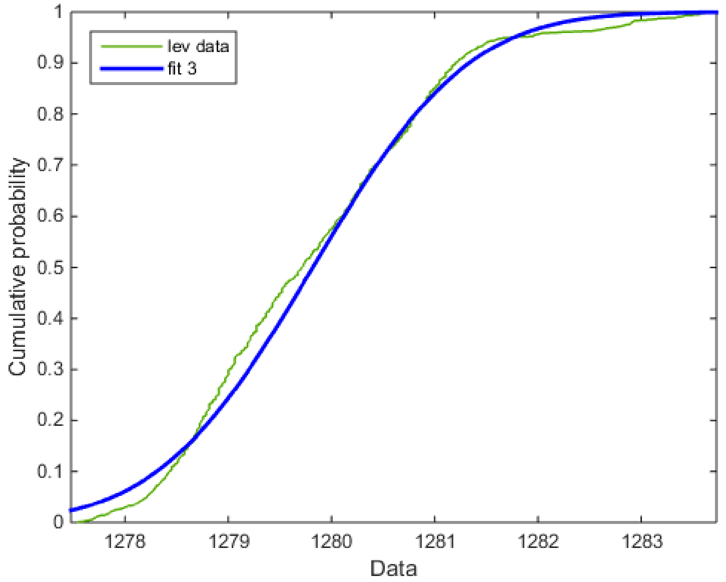




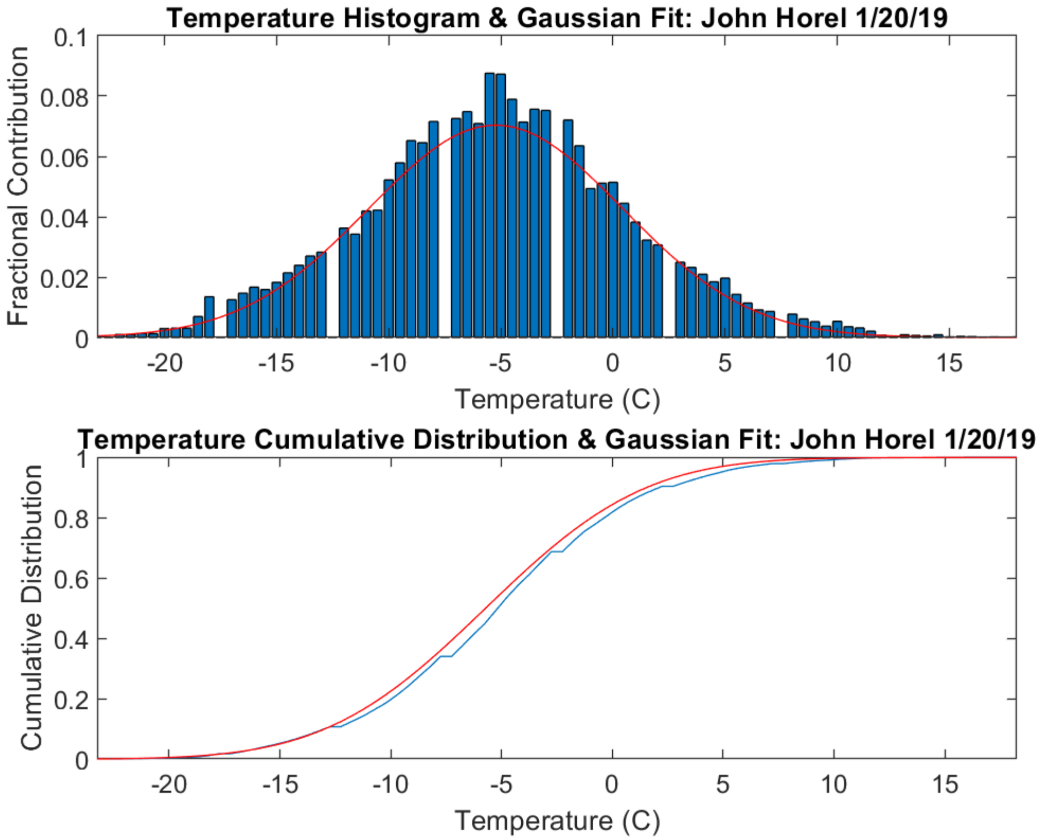
Gaussian fit to the annual level of the Great Salt Lake in terms of a histogram (top) and cumulative distribution (middle) and cumulative distribution using the dfittool.

Run dfittool

- Select the “lev” data for a cumulative prob distribution
- Select the “normal” fit and apply



Alta-Collins Temperature

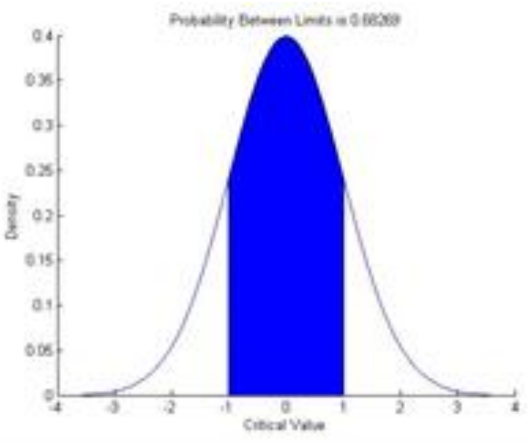


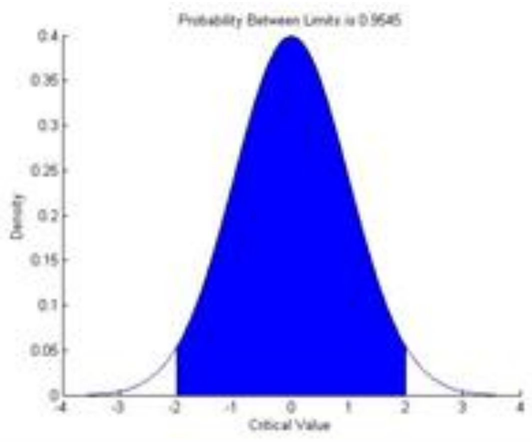
PDF and CDF of Alta-Collins hourly temperature data with Gaussian fit.

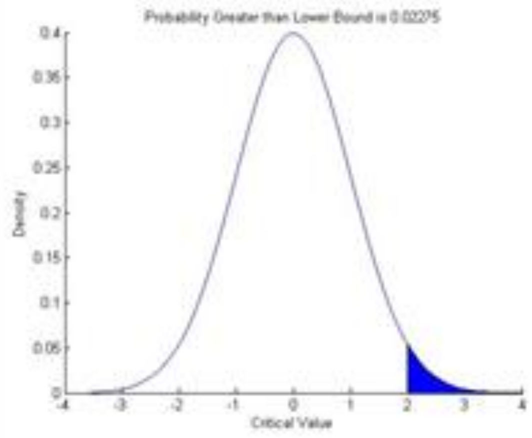
- Mean: -5.1
- Variance: 34.8
- In dfittool: Clear out lev and bring in “cln” for density plot

Using normspec: normal density plot

- 68.3% between -1 and 1
- 95.% between -2 and 2
- 2.3% of time variable explained by Gaussian distribution > 2 std dev of mean

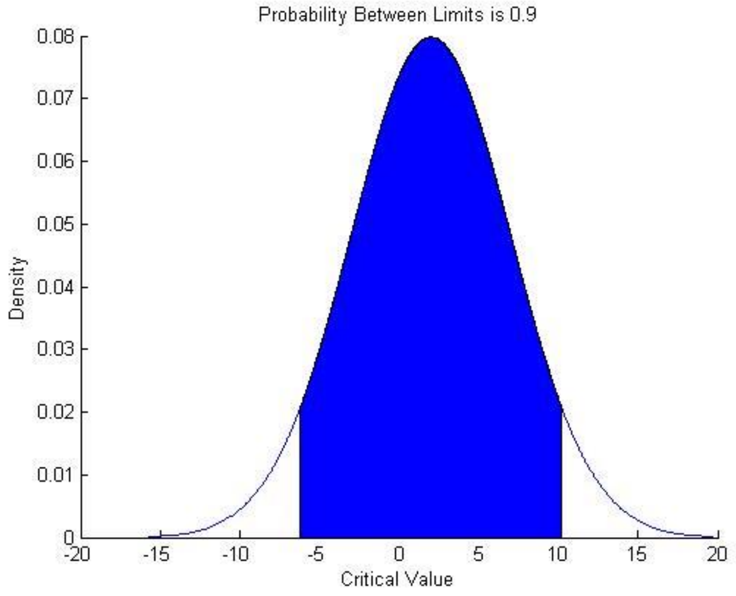






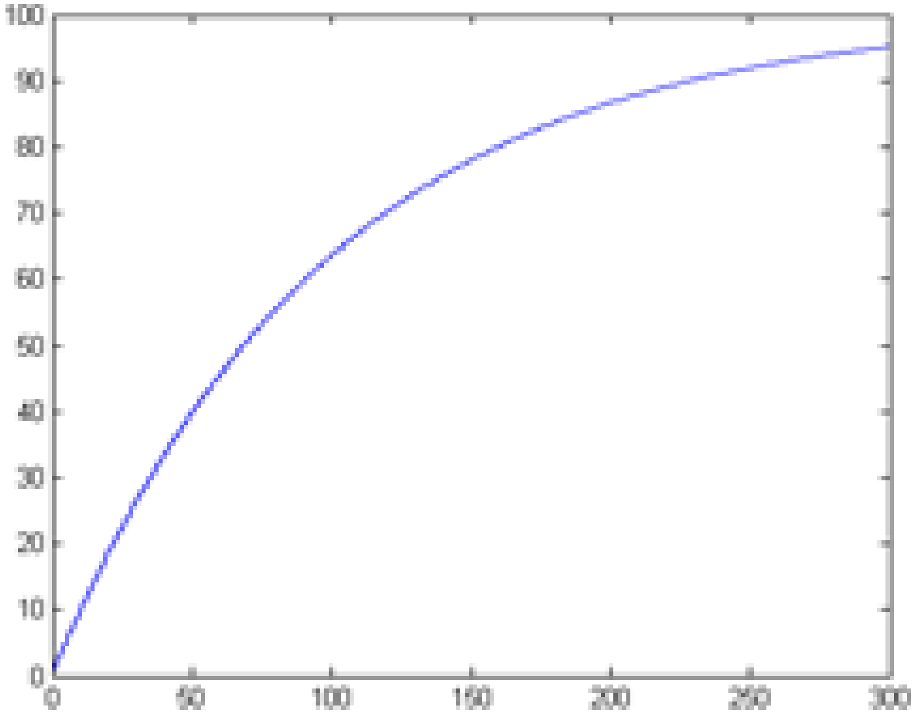
Using norminv: quantile function

- norminv([0.05,0.95],2,5)
- 90% of total variance between -6.2243   10.2243
- normspec([-6.2243,10.2243],2,5)

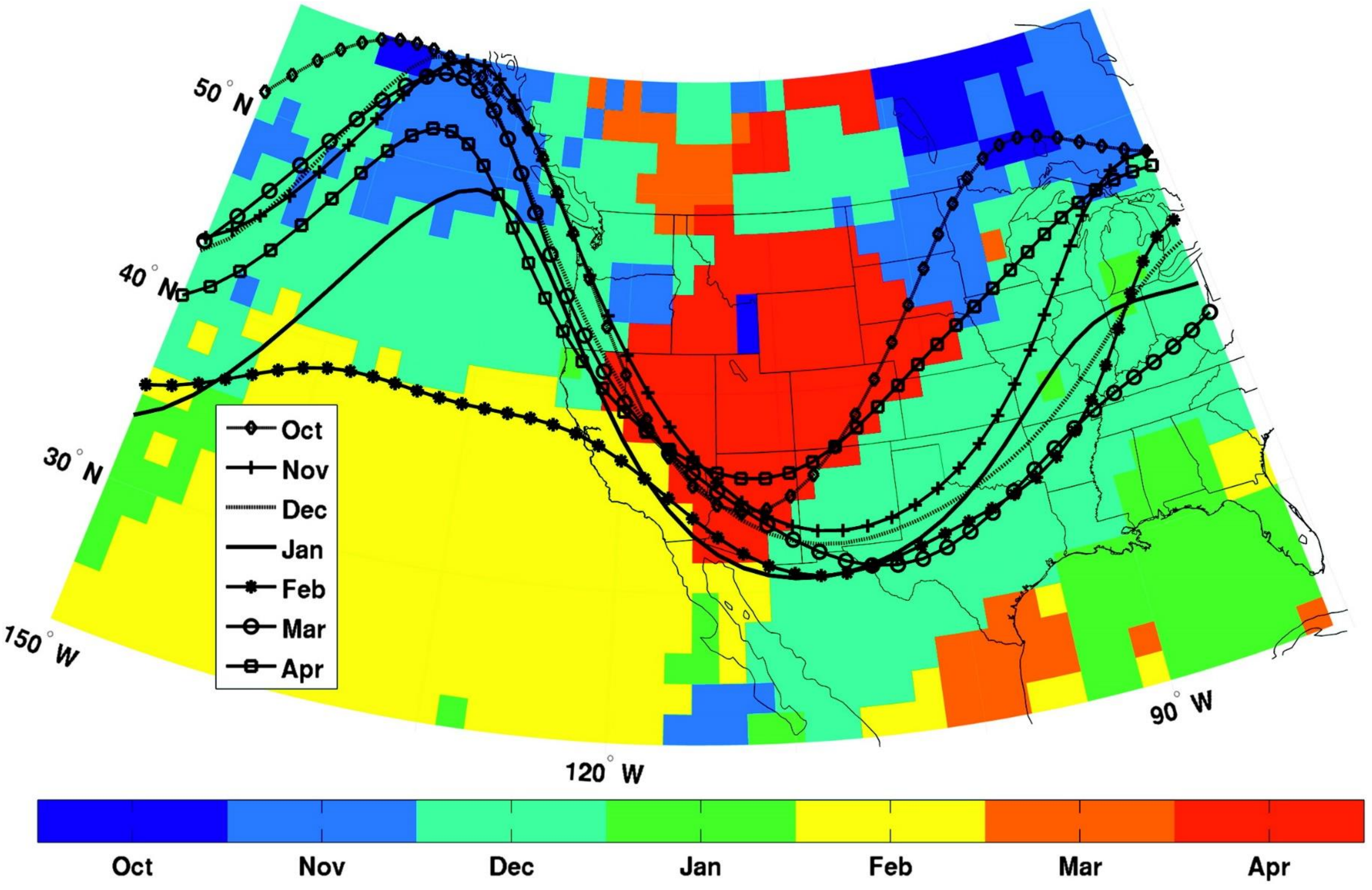


Geometric Distribution

- Estimating how likely rare events can happen by chance
- Pr{0.01}- probability of a 1 in 100 year event
- geocdf(x,0.01)- probability for the next event to happen in 1, 10, 30, 100, 200, 300 years
- 63% chance in next 100 years
- 12% chance not until 200 years
- This is nothing “rea”, just one of many assumptions



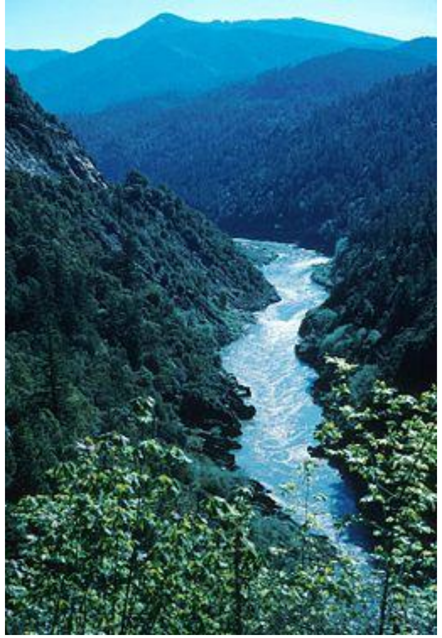
Lareau and Horel (2012)

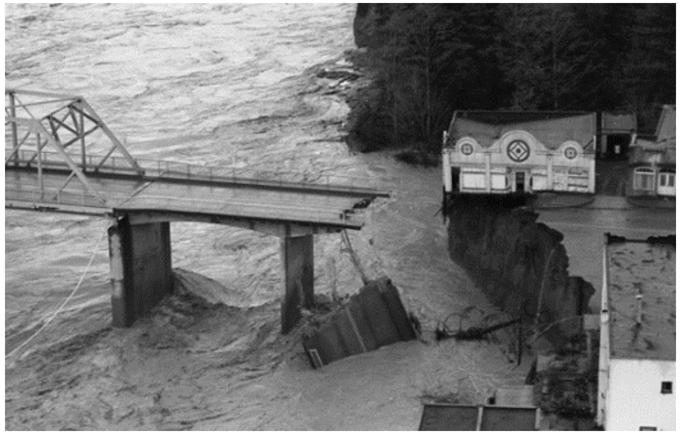


Calendar month of largest mean synoptic-scale ascent (color shading). The monthly mean storm-track centers are shown for reference.

Klamath River, northern CA

http://water.weather.gov/ahps2/hydrograph.php?wfo=eka&gage=klmc1





Flood Categories (in feet)

Major Flood Stage:	46
Moderate Flood Stage:	42
Flood Stage:	38
Action Stage:	30

Historic Crests

(1) 61.29 ft on 12/23/1964
(2) 47.12 ft on 12/31/2005
(3) 43.80 ft on 01/01/1997
(4) 41.64 ft on 02/10/2017 **(P)**
(5) 40.52 ft on 12/29/2005
Show More Historic Crests

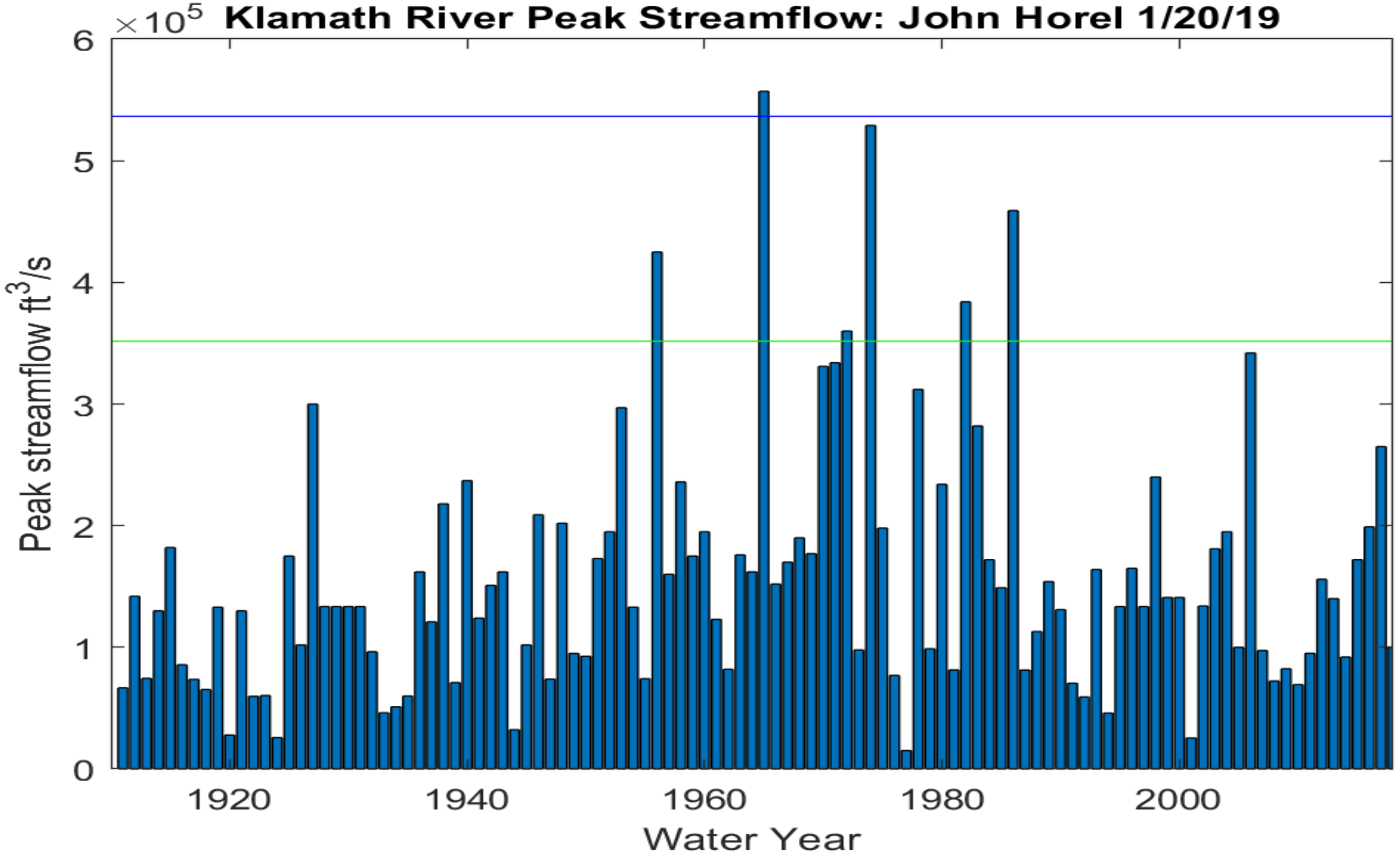
(P): Preliminary values subject to further review.

Recent Crests

(1) 41.64 ft on 02/10/2017 **(P)**
(2) 25.51 ft on 03/10/2014
(3) 30.83 ft on 12/02/2012
(4) 32.33 ft on 03/31/2012
(5) 25.82 ft on 12/29/2010

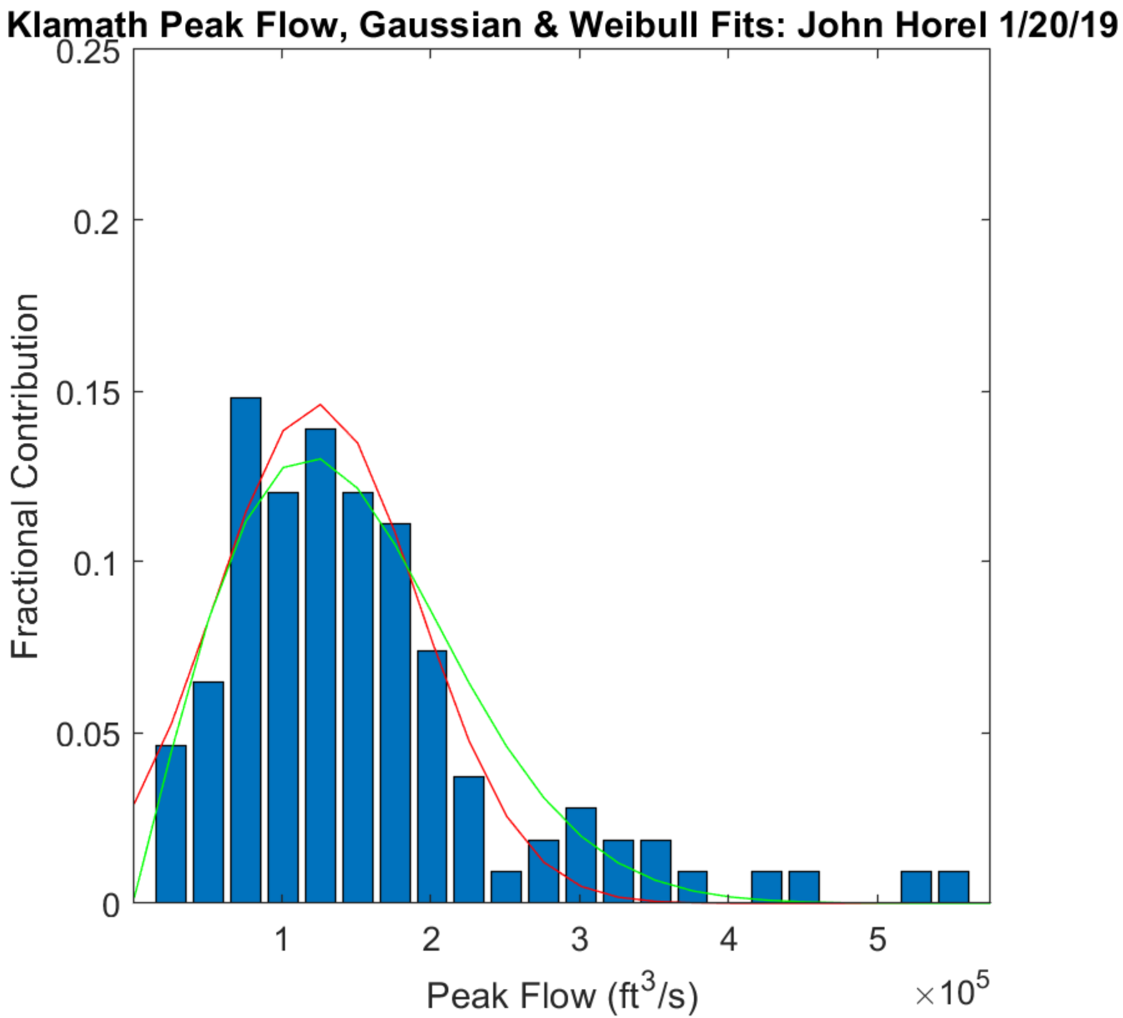
Klamath River Streamflow

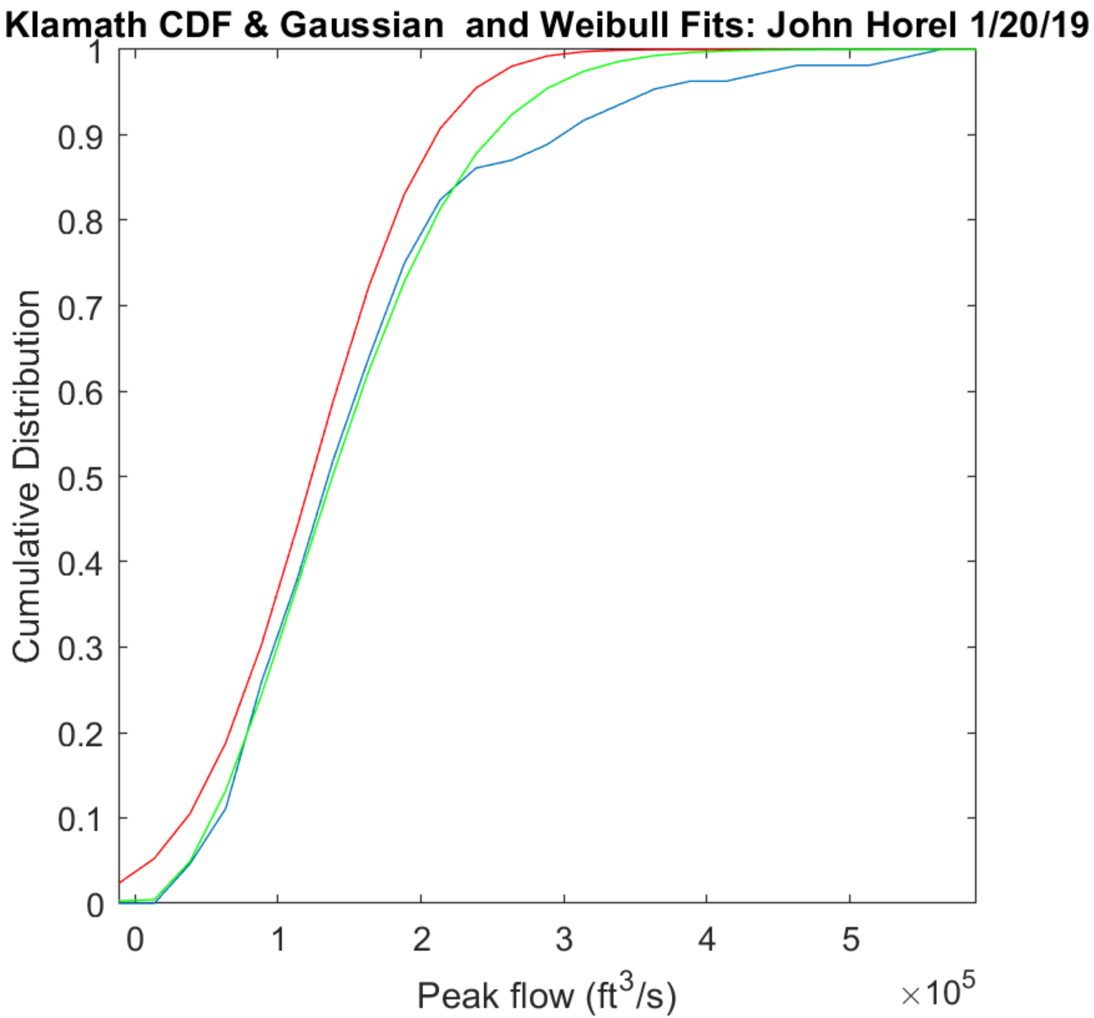
- Analyzing “flow” variable



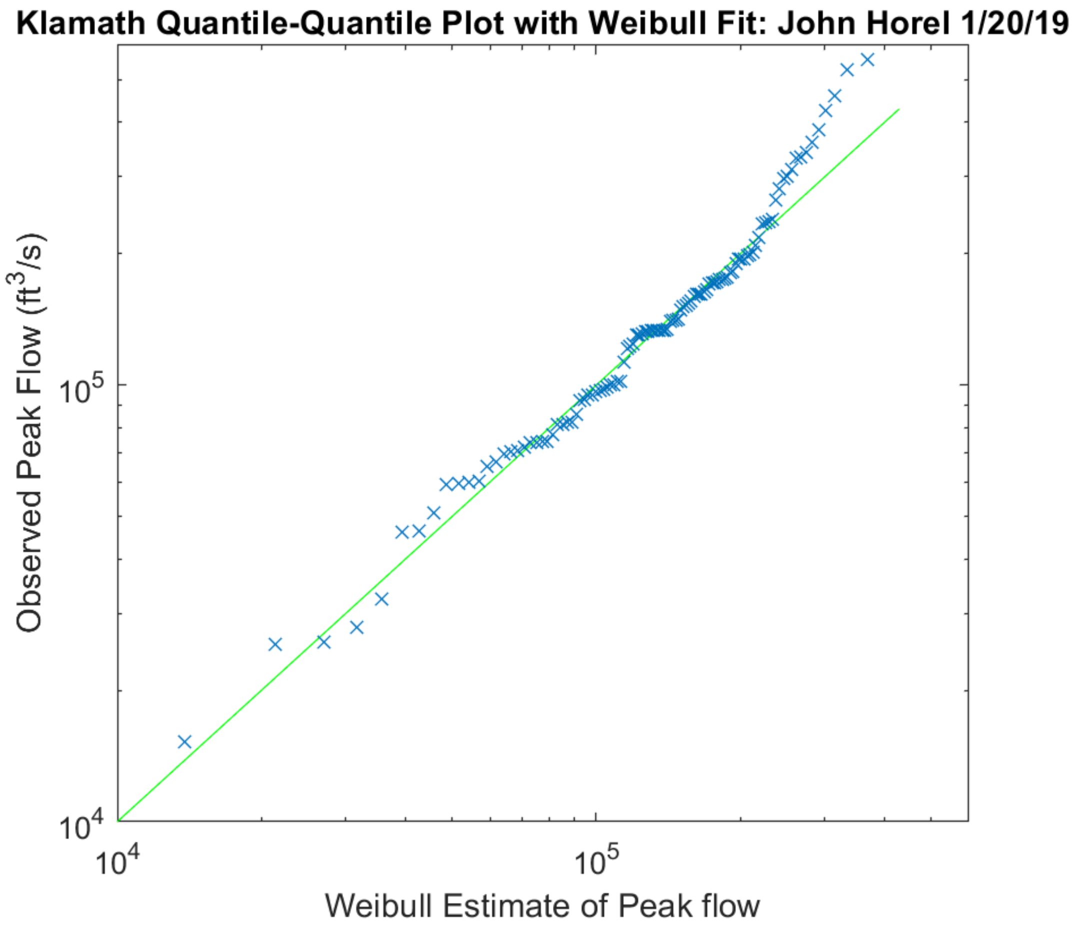
Klamath River Streamflow

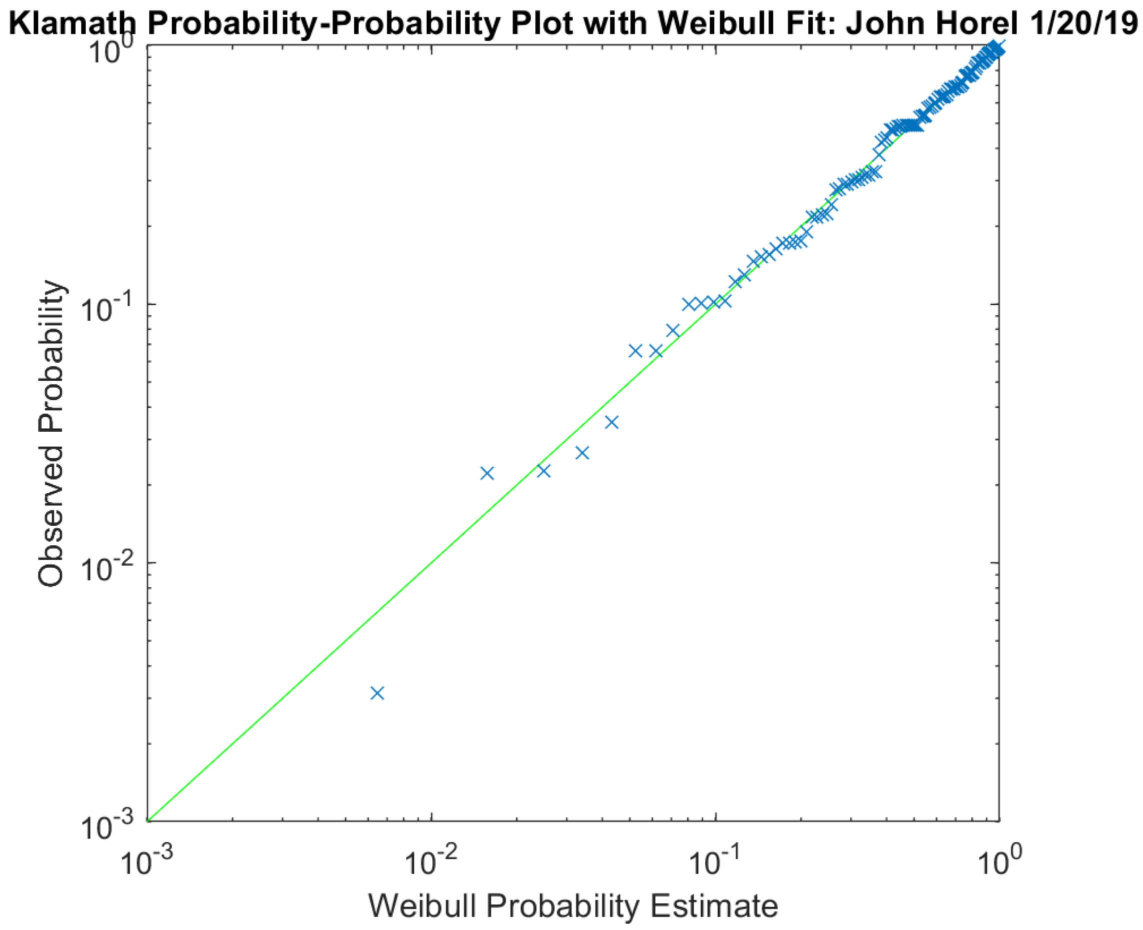
- Weibull parametric fit





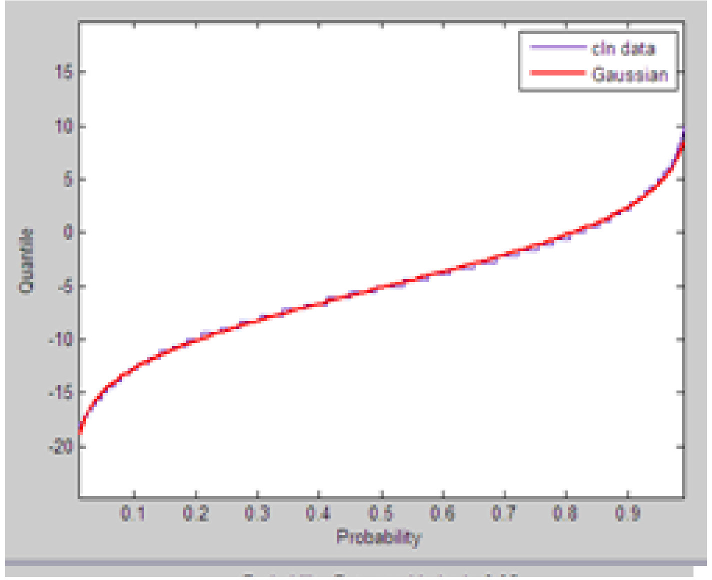
Klamath streamflow





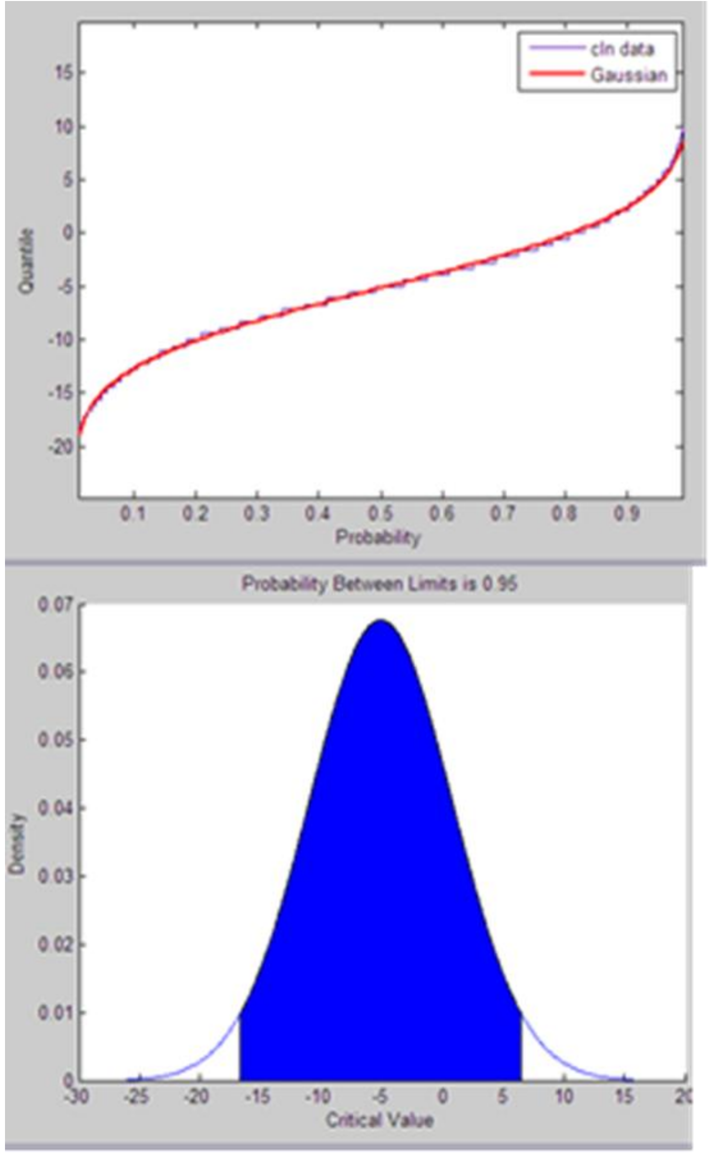
Hypothesis Testing

- Alta temperature:
- Empirically: probability of temperature less than -15C is low
- Empirical estimates:
  - Mean= -5.1C
  - Std dev = 5.9C
- What are chances of getting temp of -20C IF this was a population of random numbers with that mean and std dev?



Null hypothesis

- Null hypothesis: Temp of -20C does not differ significantly from mean of -5.1C
- 95% of time, random value would be within -16 and 6C
- So 5% of time, random value would be outside this range
- REJECT the null hypothesis accepting a 5% risk that we are rejecting the null hypothesis incorrectly
- If null hypothesis: Temp of -15C does not differ significantly from mean of -5.1C
- CANNOT reject the null hypothesis since 95% of the time the value could be within -16 and 6C



Warning

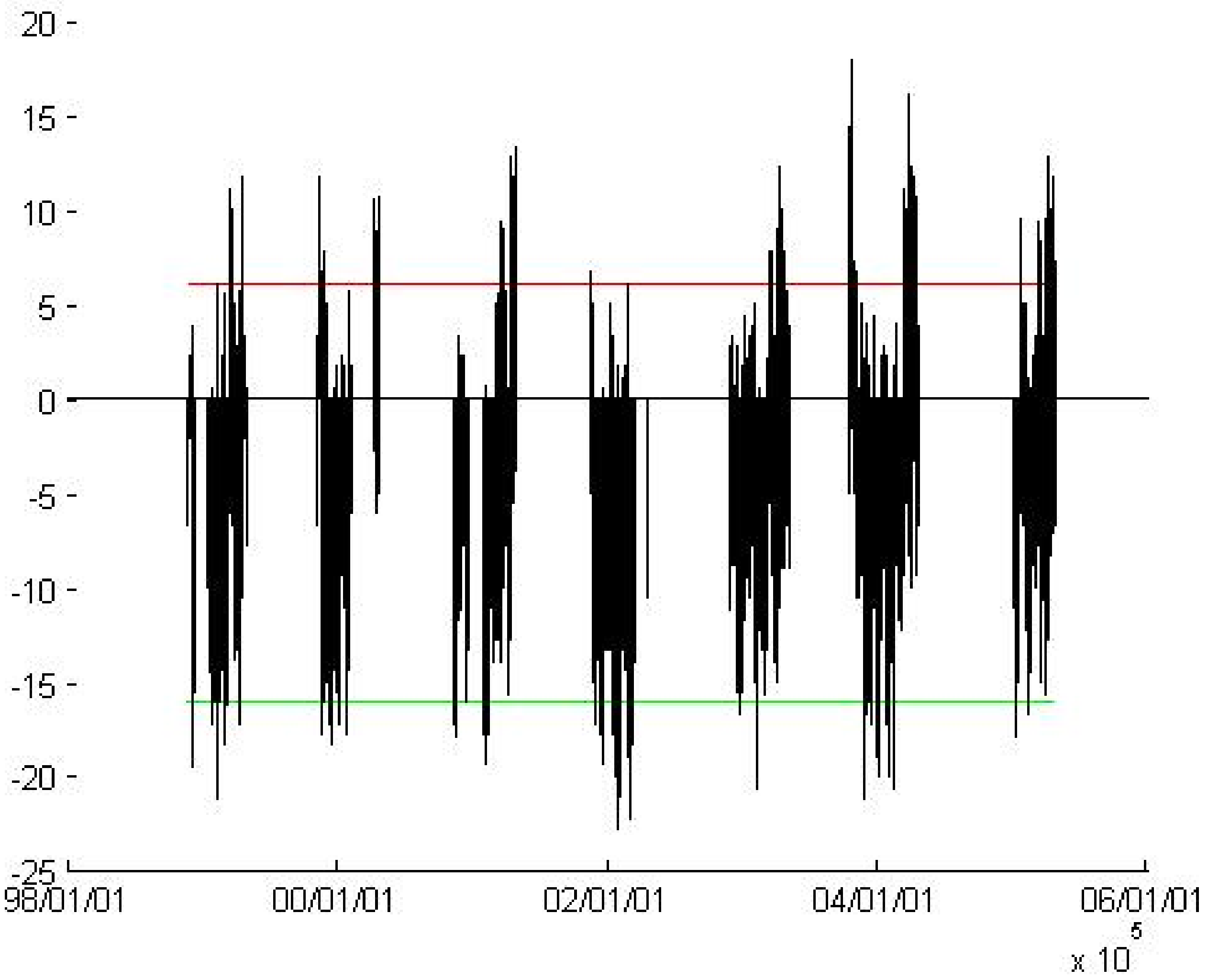
Don't use language such as:

- the results are significant at the 5% level

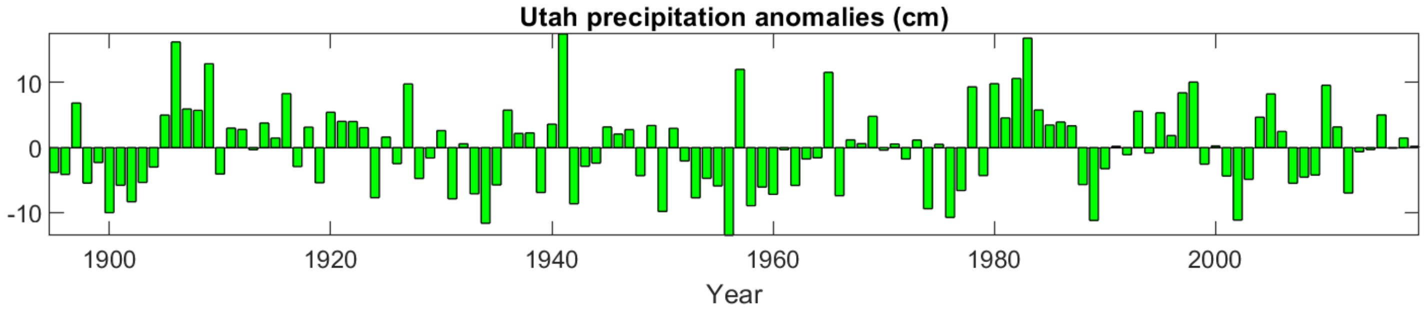
You can only state:

- Null hypothesis is rejected as too unlikely to be true with a risk of 5%

Collins: Confidence Intervals



Annual Precip in Utah



What 3 yr periods have experienced droughts?

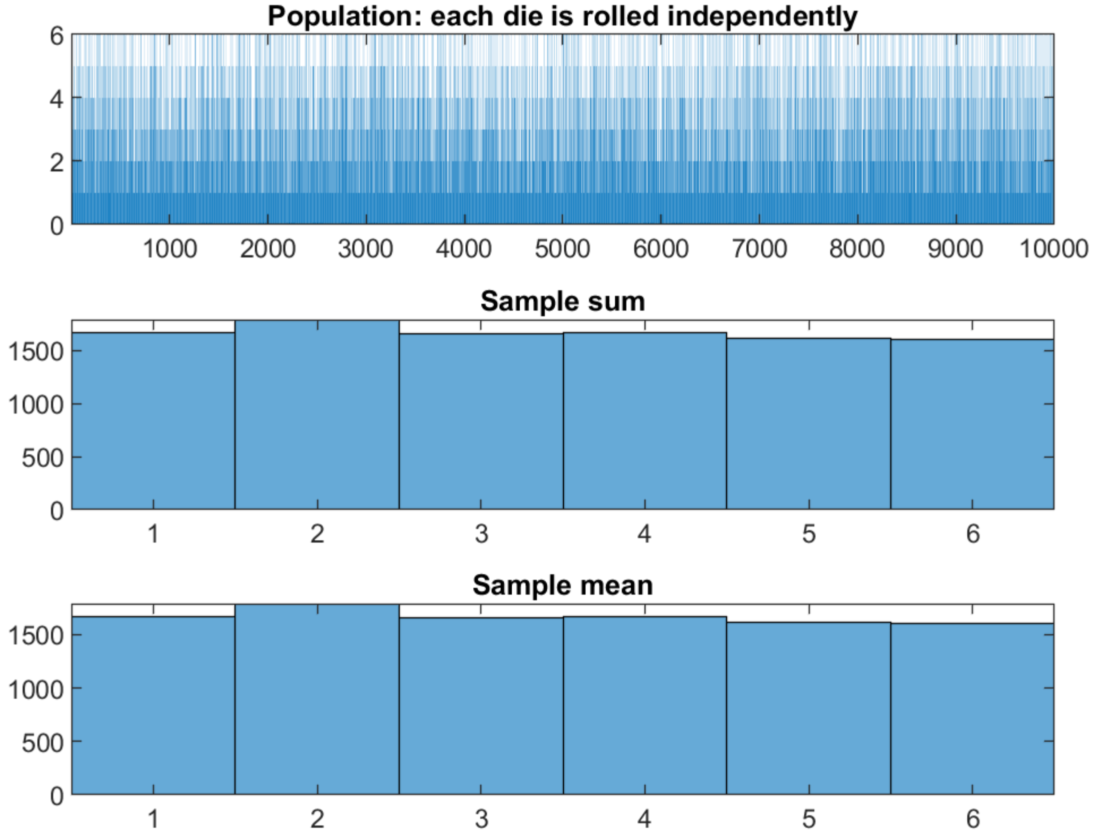
Steps of Hypothesis Testing

- Identify a test statistic that is appropriate to the data and question at hand
  - Computed from sample data values. 3 yr sample means
- Define a null hypothesis, H_0 to be rejected
  - 3 yr sample mean 0
- Define an alternative hypothesis, H_A
  - 3 yr sample mean < 0
- Estimate the null distribution
  - Sampling distribution of the test statistic IF the null hypothesis were true
  - Making assumptions about which parametric distribution to use (Gaussian, Weibull, etc.)
  - Use sample mean of 0 and 124 yr sd of 6.4
- Compare the observed test statistics (3-yr means) to the null distribution. Either
  - Null hypothesis is rejected as too unlikely to have been true IF the test statistic fall in an improbable region of the null distribution
    - Possibility that the test statistics has that particular value in the null distribution is small
    - normspec([-1.96*6.4,1.96*6.4],0,6.4)
    - OR
    - The null hypothesis is not rejected since the test statistic falls within the values that are relatively common to the null distribution

Caution!

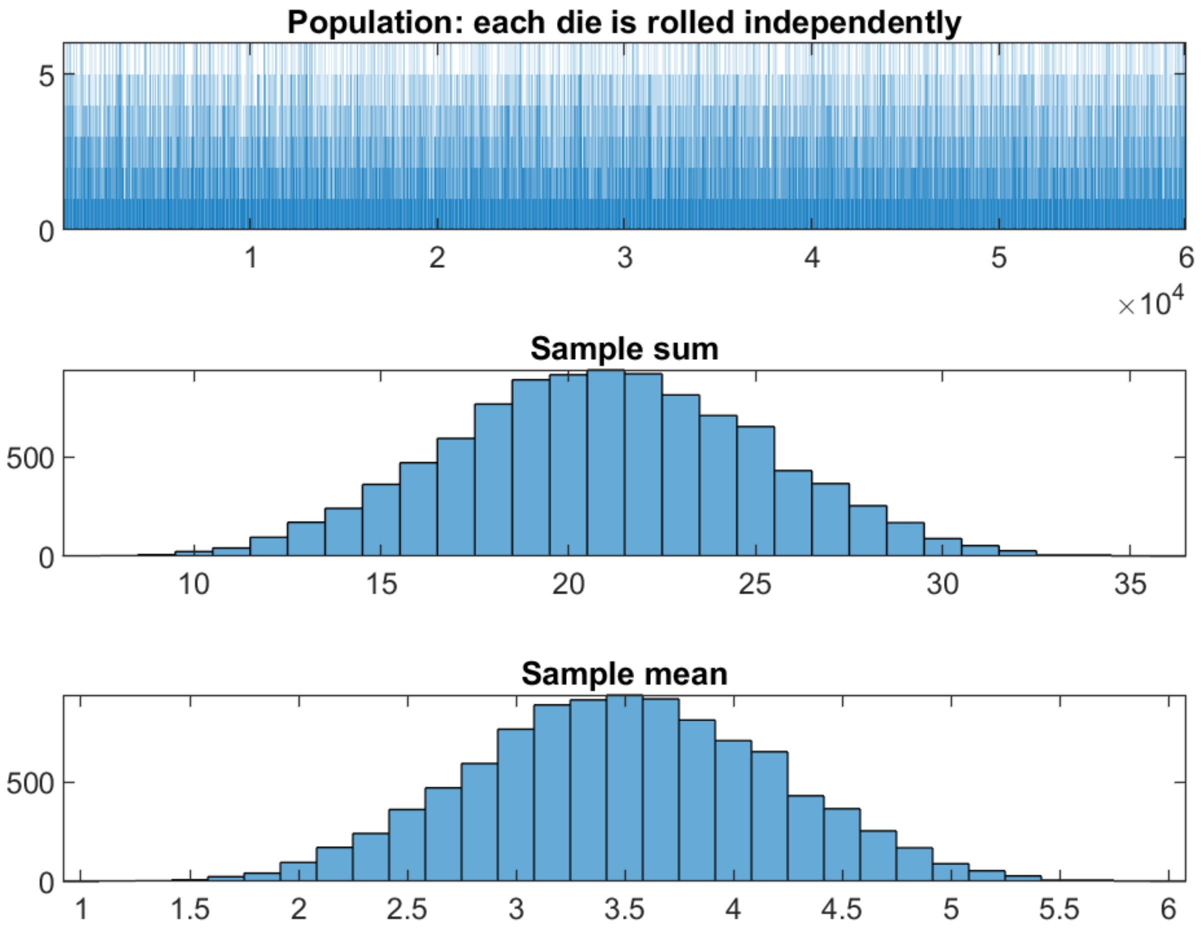
- NOT rejecting the null hypothesis is not the same as saying the null hypothesis is true
  - There is insufficient evidence to reject H_0
- H_0 is rejected if the probability p of the observed test statistic is ≤ α significance or rejection level
- If odds of test statistic occurring in the null distribution less than 1 or 5%, then we may choose to reject the null hypothesis
- Rejecting the null hypothesis MAY be same as accepting alternative hypothesis BUT there may be many other possible alternative hypotheses
- You must define ahead of time the α significance or rejection level
  - 1% or 5%, 1 in 100 or 5 in 100 chance that you accept the risk of rejecting the null hypothesis incorrectly
  - Type 1 category error of a false rejection of the null hypothesis

Roll 1 die 10000 times



Roll 6 die 10000 times

- Getting the sum or mean of 6 numbers



Central Limit Theorem

sum (or mean) of a sample (6 dice) will have a Gaussian distribution even if the original distribution (one die) does not have a Gaussian distribution, especially as the sample size increases.

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

$\sigma_{\bar{x}}$ standard deviation of the sample means

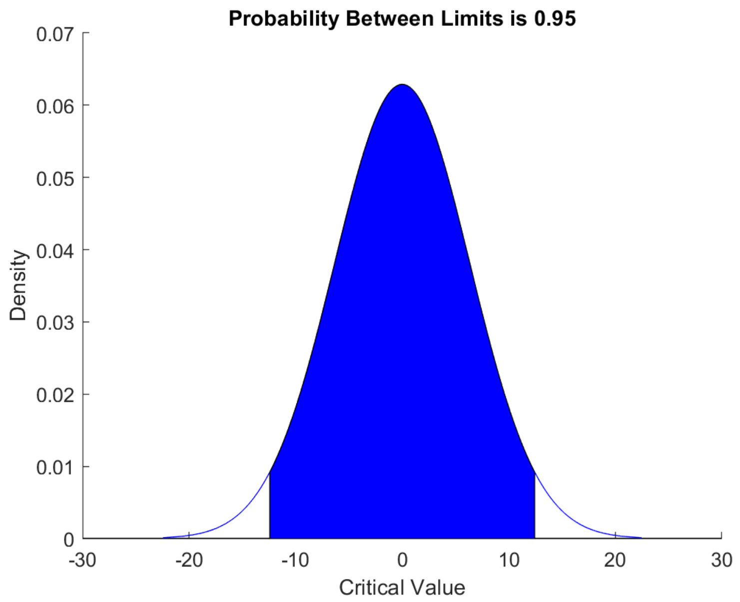
σ standard deviation of the original population

n sample size

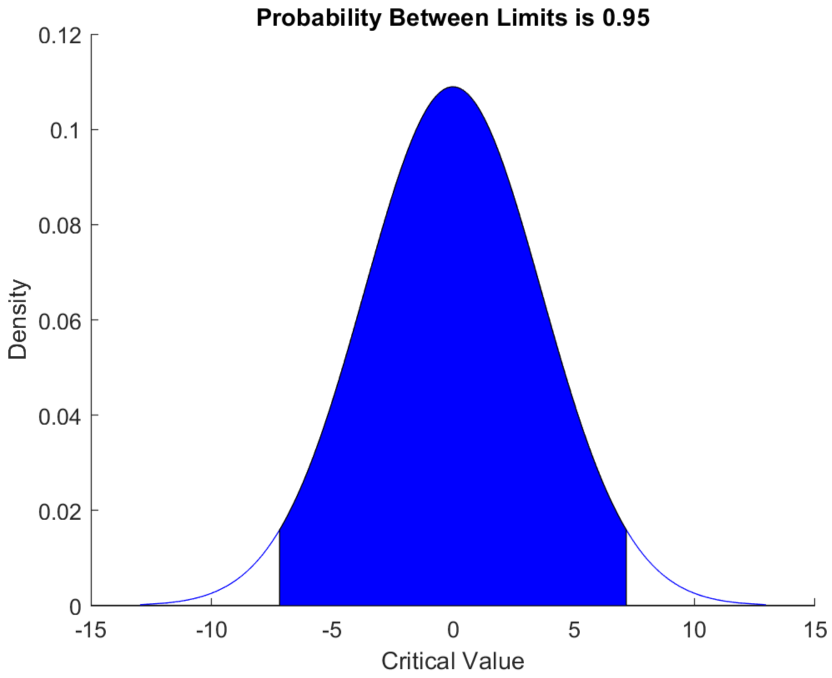
$$s_x = \sqrt{\frac{n-1}{n}}\sigma \qquad \sigma_{\bar{x}} = s_x / \sqrt{n-1}$$

degrees of freedom: n-1, since sample can be described by the mean (1 value) plus n-1 others

Left: normspec([-1.96*6.3,1.96*6.3],0,6.3)
right: normspec([-1.96*3.7,1.96*3.7],0,3.7)



95% chance that individual year within 12.4 cm



95% chance that 3-yr mean anomaly within 7.3 cm

Less likely to have a 3-yr drought (really large 3-yr mean) than to have a single really dry year

Students' t test

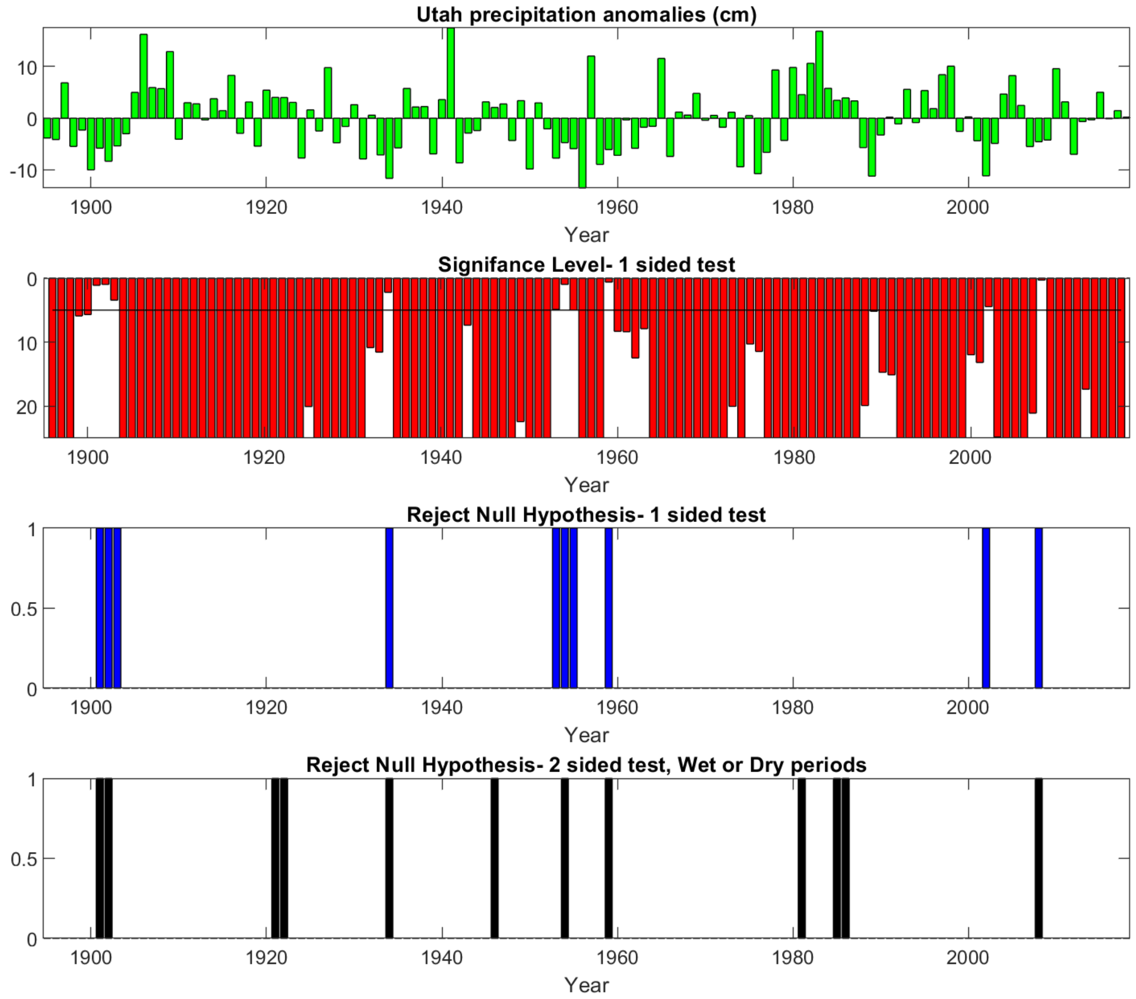
- $\sigma_{\overline{X}} = \frac{s_x}{\sqrt{n-1}}$
- Estimate of population variance from sample
- T value:
- Numerator: signal
- Denominator: noise
- At t gets larger, confidence in rejecting the null hypothesis (sample mean differs from population mean) gets higher
- T large IF:
  - Spread between sample and population means large
  - Degrees of freedom is large
  - Variability in sample is small

$$t = (\bar{x} - \mu)\sqrt{n-1}/s_x$$

Using t test

- **[h,p,ci,stat]= ttest(valy,0,.05,'left');**
- where on input valy is the vector of values in each 3-year sample
- 0 is the mean value for the null hypothesis
- .05 is the significance level chosen (5%)
- ‘left’ indicates that we are assuming that we have ruled out that large positive anomalies are relevant (the other options are ‘both’ a two-tailed test and ‘right’ where we rule out large negative anomalies, i.e., look for wet periods)
- Output:
  - h is a flag, 0 means the null hypothesis can not be rejected, 1 means it can be rejected
  - p is the significance level corresponding to the t value, the smaller the number the better
  - ci is the confidence interval (low and high values for the sig level chosen)
  - stat- is an array that returns the value of the t statistic, the number of degrees of freedom, and the estimated population standard deviation

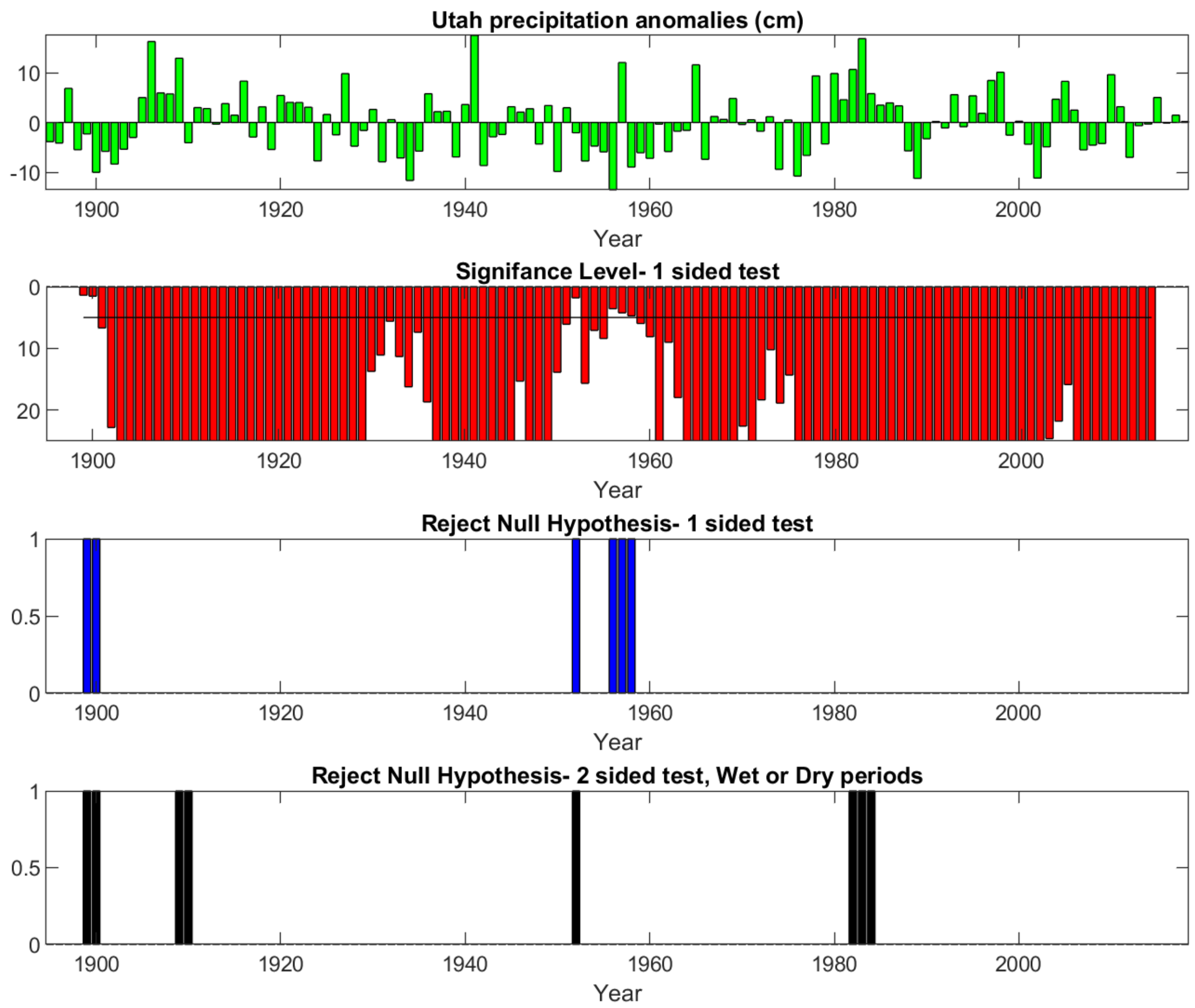
Which 3-yr samples would be considered a drought?



Left, Right, 2 Sided (both)

- Left- alternative hypothesis is drought
- Right- alternative hypothesis is flood
- 2 Sided- either drought or flood
- 2 sided tests are weaker and should be avoided
- [h,p,ci,stat]= ttest(valy,0,.05,'both')
- Sample value must be further from 0 (smaller p value) since α is smaller by 2 (2.5% in each tail)

Two Sided Test: Flood or Drought



Summary

- Research involves defining a testable hypothesis and demonstrating that any statistical test of that hypothesis meets basic standards
- Typical failings of many studies include:
  - (1) ignoring serial correlation in environmental time series that reduces the estimates of the number of degrees of freedom and
  - (2) ignoring spatial correlation in environmental fields that increases the number of trials that are being determined simultaneously.
    - Inflates the opportunities for the null hypothesis to be rejected falsely.
- Use common sense
- Be very conservative in estimating the degrees of freedom temporally and spatially
- Avoid attributing confidence to a desired result when similar relationships are showing up far removed from your area of interest for no obvious reason
- The best methods for testing a hypothesis rely heavily on independent evaluation using additional data not used in the original statitiscal analysis

Assignments

- Undergrads: Dept and CMES scholarship deadline this Friday
- Turn in in-class Assignment
- Wednesday in MLIB again
- Time to Work on Probability Assignment (#8) today, due Wednesday
- Read Chapter 3 and 4 notes in Canvas (Chapter 5 also available)
- Read text: Text Chapter 3.1-3.6, Chapter 4
- Will discuss Assignment #10 on Wednesday