Statistical Methods in Hydrology

Prof. Belize Lane

belize.lane@usu.edu

CEE 6400 Fall 2020

COLLEGE of ENGINEERING
UtahStateUniversity

Why statistics??

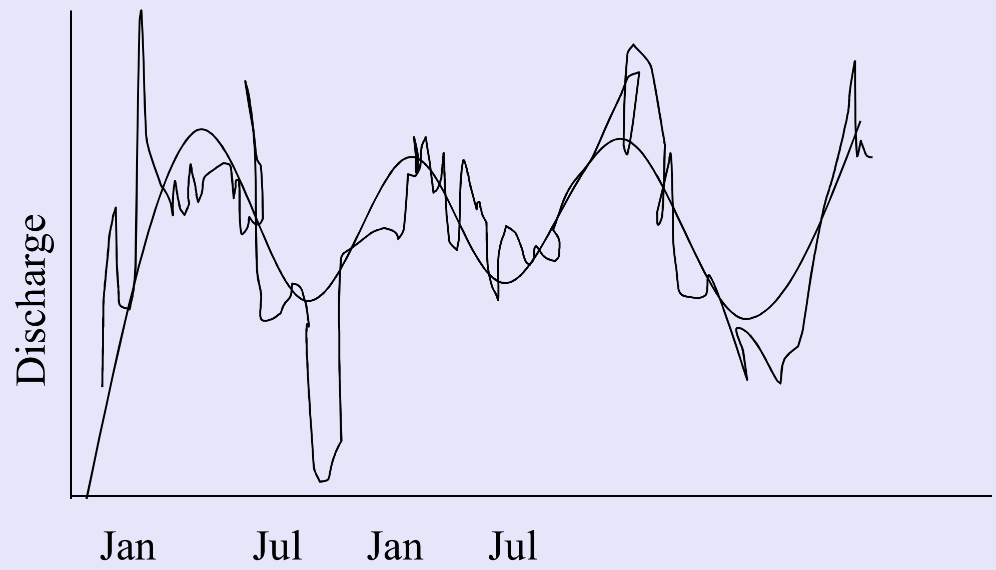
- Reduce & summarize observed data
- Present information in precise and meaningful form
- Determine underlying characteristics of observed phenomena
- Make predictions concerning future behavior

Why statistics??

Hydrologic processes:

Predictable (*deterministic*) + Random (*stochastic*)

Probability theory & statistics



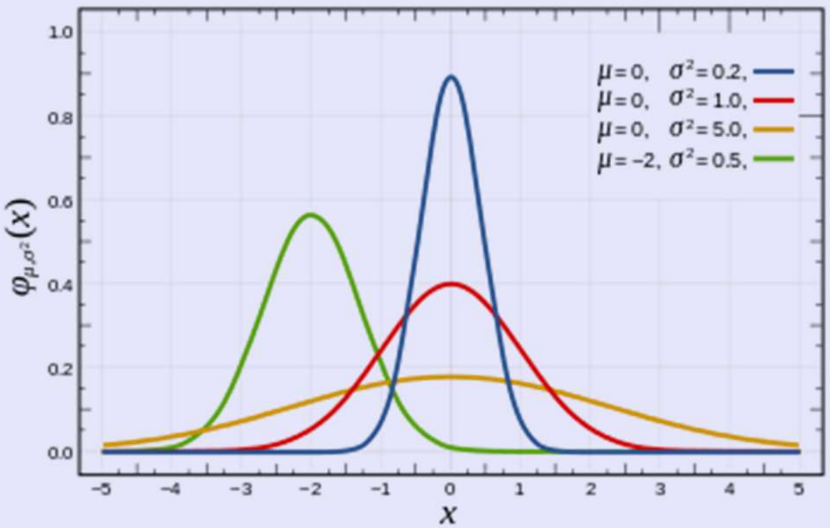
COLLEGE of ENGINEERING
UtahStateUniversity

Lane 2020

Hydrologic data often exhibit…

1. A **lower bound of zero**
2. Presence of **'outliers'**
3. Positive **skewness**. Skewness can be expected when outlying values occur in only one direction. (eg log-normal distribution)
4. **Non-normal distribution** of data. Data may be reported only as below or above some threshold (eg annual flood stage records)
5. **Seasonal patterns**. Values tend to be higher or lower in certain seasons.
6. **Autocorrelation**. Consecutive observations are highly correlated (high follow high, or low follow low values)
7. **Dependence** on other uncontrolled variables (eg precipitation, hydraulic conductivity)

Concepts to Understand



- Random variable
- PDF and CDF
- Expected value
- Parametric v. non-parametric
- Quantiles
- Method of Moments
- Flow exceedance
- Frequency/ return period
- Confidence intervals

COLLEGE of ENGINEERING UtahStateUniversity

Lane 2020

Summarizing time-series data

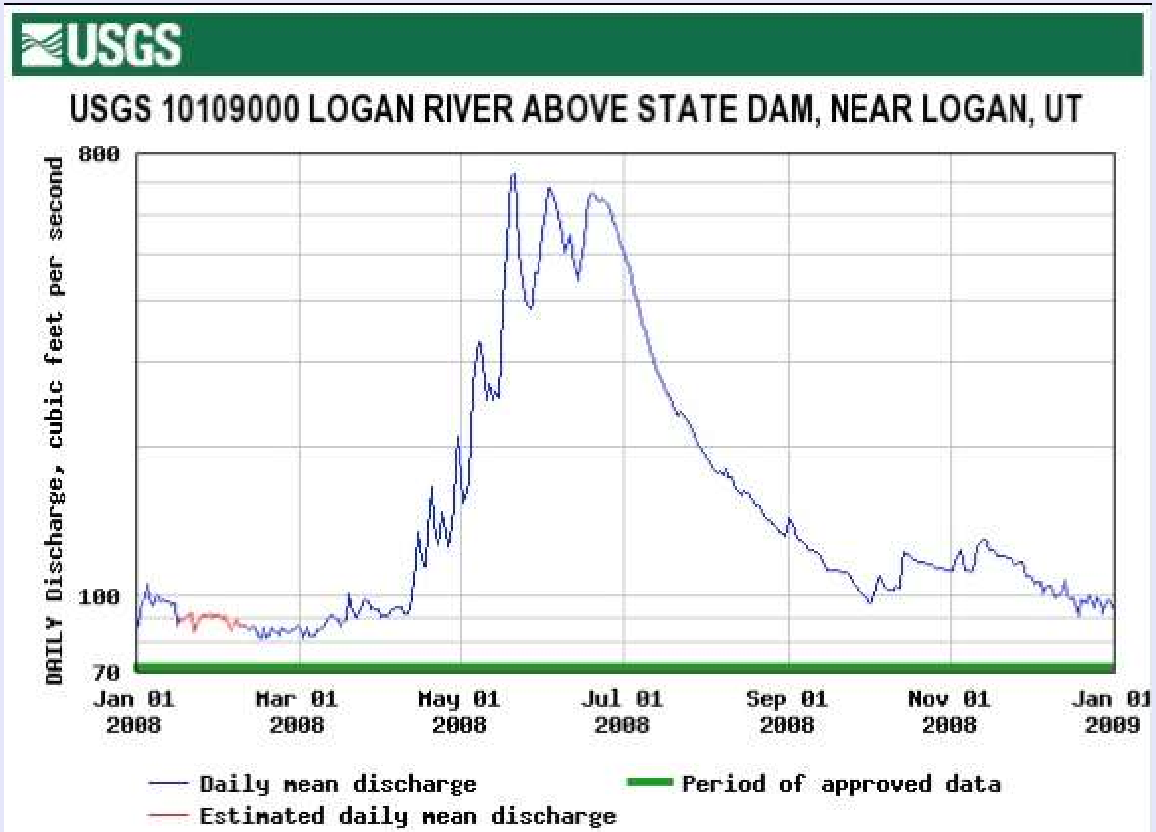
- Time series plots
- Histogram/ frequency distribution
- Box plots
- Flow duration curves (FDC)

Summarizing time-series data

Time series plot

- Plot variable versus time (bar/line/points)

  Example: Daily discharge, monthly streamflow

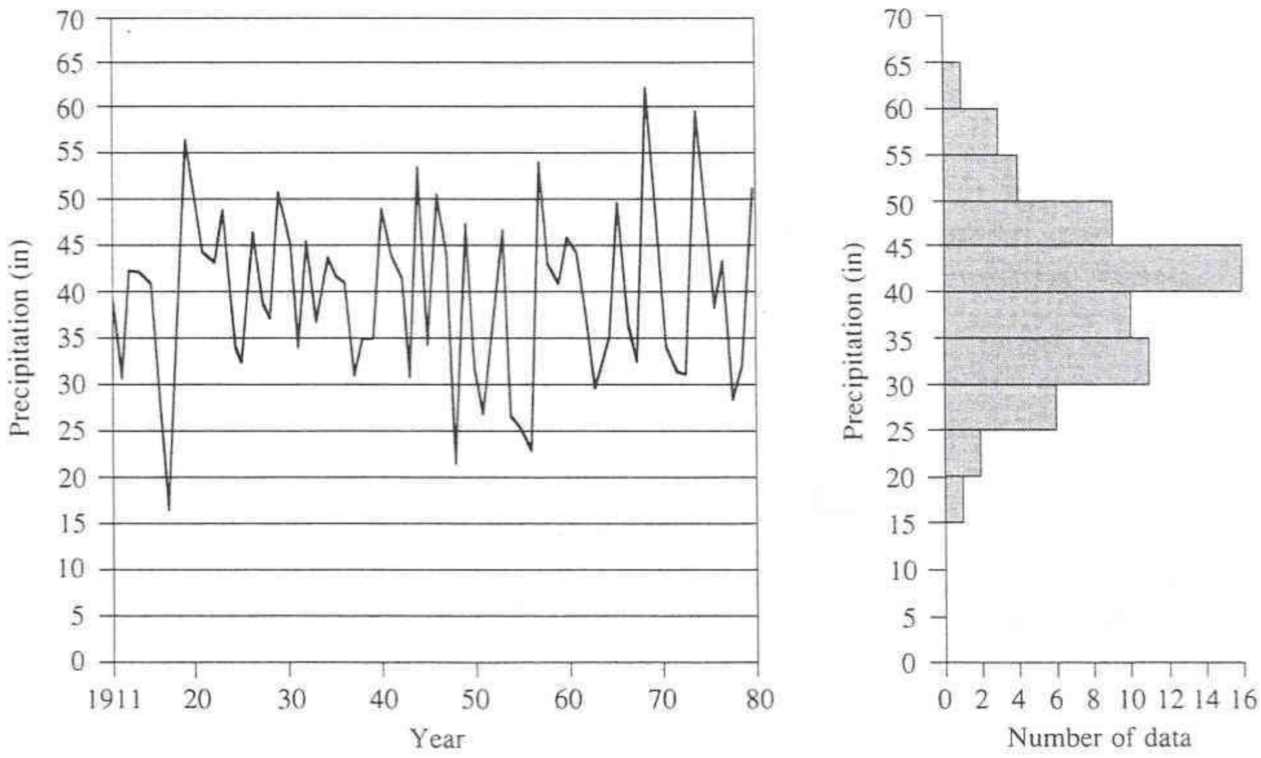


Lane 2020

Summarizing time-series data

Histogram

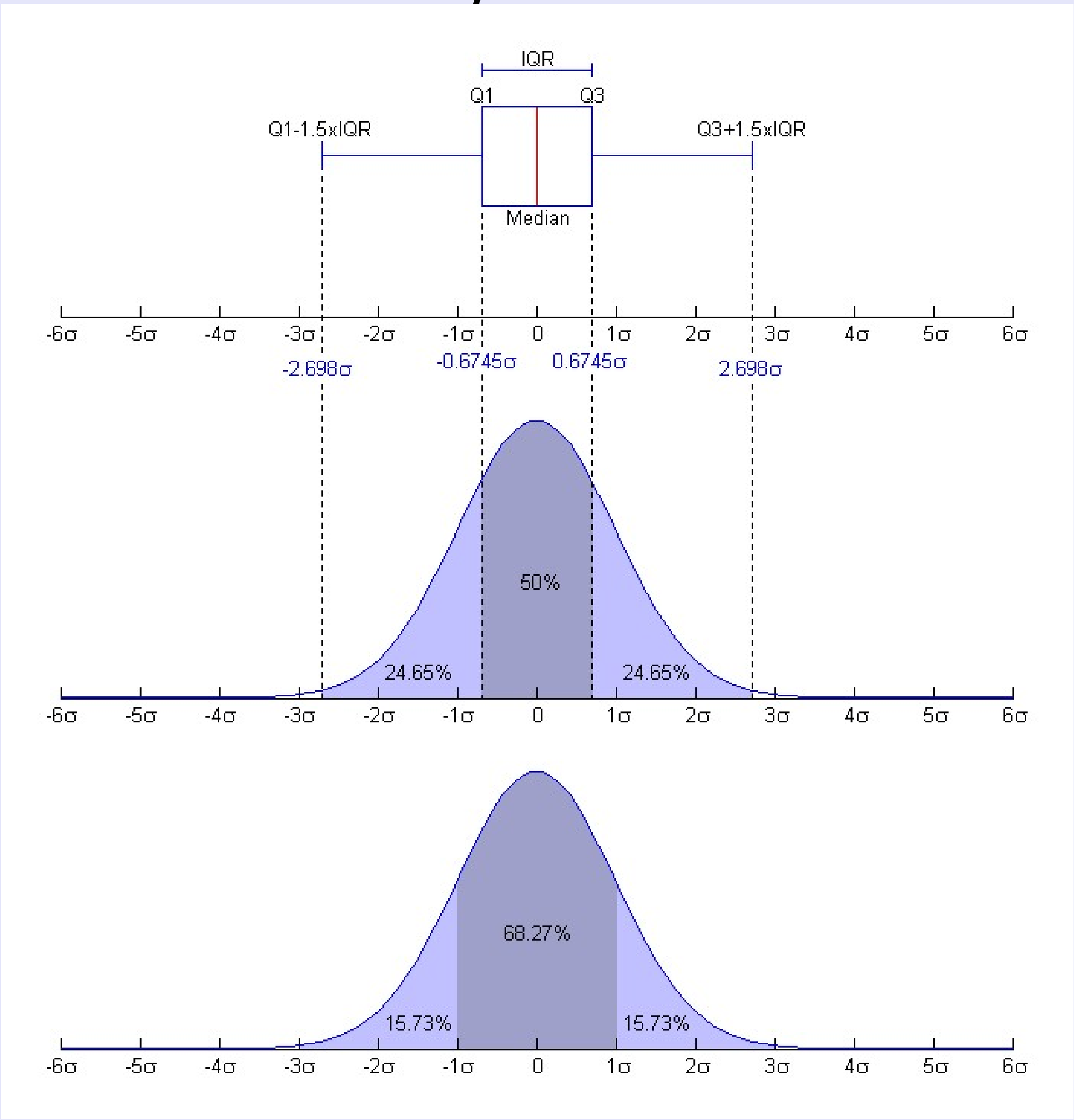
- Bar plots of the number n_i or fraction (n_i/N) of data falling into equal width intervals of data values ("bins")



(*a*) Annual precipitation. (*b*) Frequency histogram.

Summarizing time-series data

Boxplots



Summarizing time-series data

Flow Duration Curve (FDC)

Plot of the percent of time that flow exceeds some specified value.

Step 1: Sort (rank) average daily discharges for period of record from largest to smallest for a total of *n* values.

Step 2: Assign each discharge value a rank (*i*), starting with 1 for the largest daily discharge value.

Step 3: Calculate the exceedence probability (*P*) as follows:

$$P = i / (n + 1)$$

P = the probability that a given flow will be equaled or exceeded (% of time)
i = ranked position
n = number of events in period of record

Summarizing time-series data

Flow Duration Curve (FDC)

Date	Q (cfs)	Rank (i)	Exc. Probability (P)	Return period (T)
7/2/1905	20100	1	0.0001	7306
7/2/1905	18700	2	0.0003	3653
7/2/1905	17300	3	0.0004	2435
6/20/1905	15100	4	0.0005	1827
7/2/1905	15100	5	0.0007	1461
6/20/1905	15000	6	0.0008	1218
6/15/1905	11700	7	0.0010	1044
7/2/1905	11400	8	0.0011	913
6/23/1905	10800	9	0.0012	812
6/23/1905	10700	10	0.0014	731
6/15/1905	10500	11	0.0015	664
6/23/1905	10400	12	0.0016	609
6/15/1905	10100	13	0.0018	562
7/3/1905	10100	14	0.0019	522
7/3/1905	9970	15	0.0021	487
6/26/1905	9940	16	0.0022	457
6/23/1905	9770	17	0.0023	430
6/15/1905	9650	18	0.0025	406
6/15/1905	9600	19	0.0026	385
6/23/1905	9600	20	0.0027	365
6/26/1905	9480	21	0.0029	348
7/2/1905	9380	22	0.0030	332
6/15/1905	9300	23	0.0031	318
6/26/1905	9130	24	0.0033	304

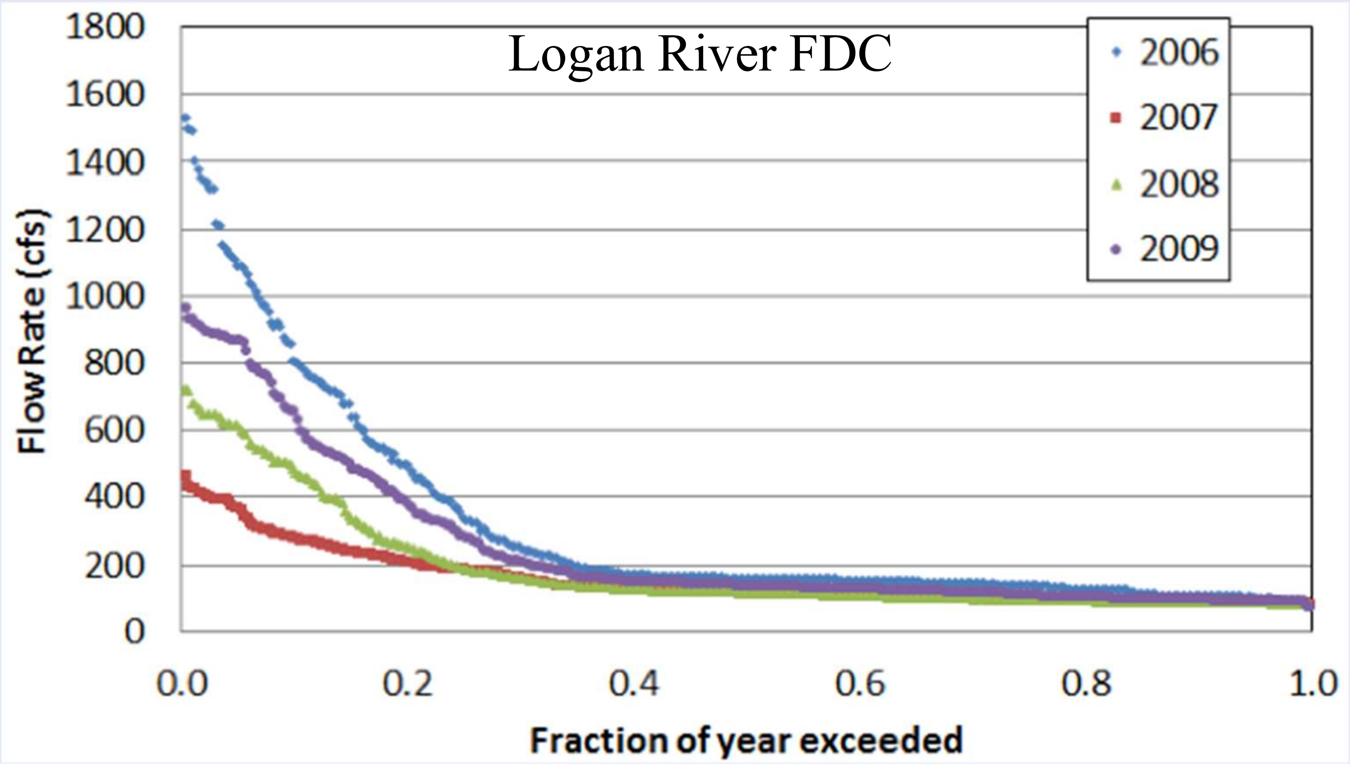
$$P = 100 * [\, i / (n + 1) \,]$$

$$T = 1/P$$

Summarizing time-series data

Flow Duration Curve (FDC)

The relationship between the magnitude and frequency of a hydrologic variable for a particular basin / year



What percentage of time does daily flow exceed a given value?

COLLEGE of ENGINEERING
UtahStateUniversity

Lane 2020

Parametric vs Non-Parametric

Nonparametric statistics (NP) are **based on the ranking of the data rather than the data values themselves.** This fact has many desirable properties in hydrologic data analysis:

- Fewer assumptions about the data distribution
- Easier to apply
- Robust to the presence of outliers

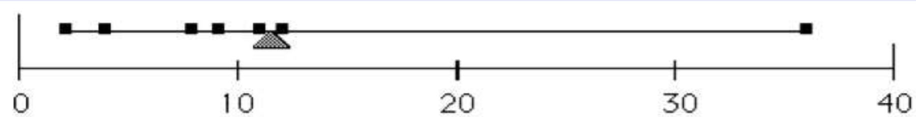


Figure 1.3a The mean (triangle) as balance point of a data set.

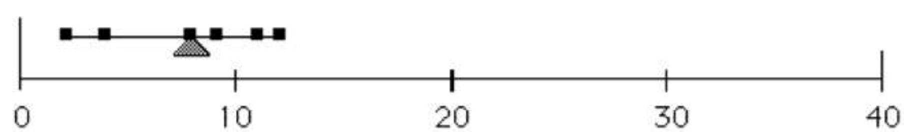


Figure 1.3b Shift of the mean downward after removal of outlier.

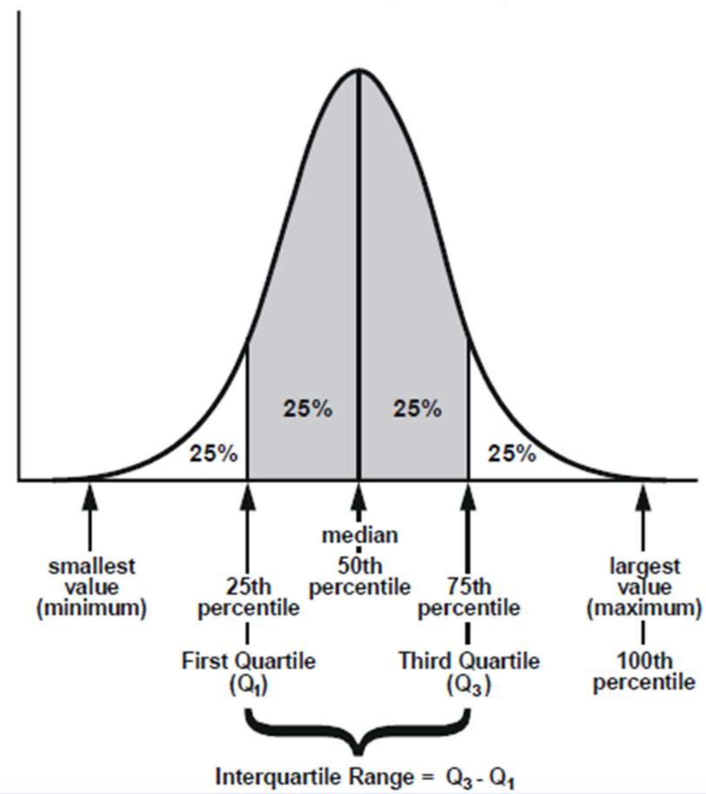
Helsel and Hirsch 2002

Lane 2020

Quantiles

The pth quantile of a random variable X divides the PDF so that p% of the values lie below and (100-p)% of the values lie above.

Figure 3.8
The middle half of the observations in a frequency distribution lie within the interquartile range



COLLEGE of ENGINEERING
UtahStateUniversity

Moments of a Distribution

Expected Value $E(X) = \int_{-\infty}^{\infty} xf(x)dx$

	Population	Sample
Mean	$\mu = \int_{-\infty}^{\infty} xf(x)dx$	$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$
Variance	$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = E([X - E(X)]^2)$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$ $S = \sqrt{S^2}$
Skewness	$\gamma = \frac{1}{\sigma^3}\int_{-\infty}^{\infty} (x-\mu)^3 f(x)dx = E([X - E(X)]^3)/\sigma^3$	$\hat{\gamma} = \frac{\frac{1}{N}\sum_{i=1}^{N} (X_i - \overline{X})^3}{S^3}$

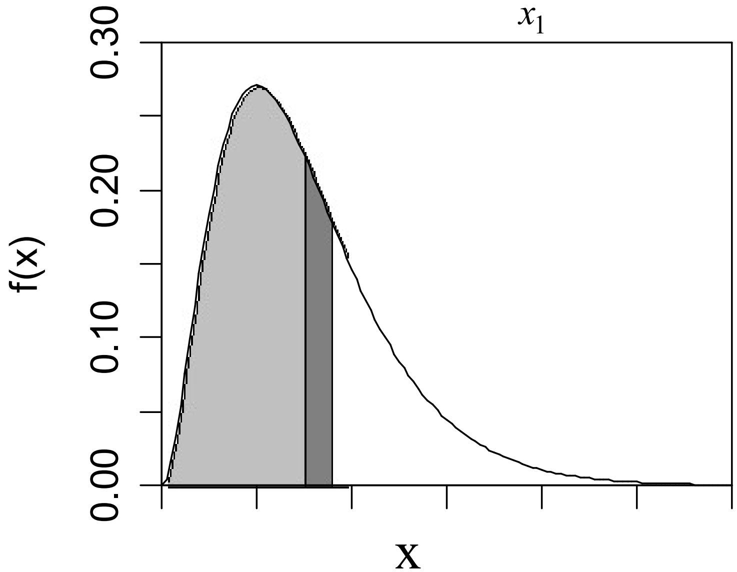
COLLEGE of ENGINEERING UtahStateUniversity

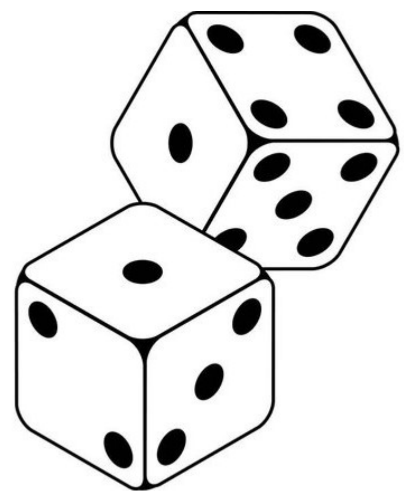
Lane 2020

Expected Value

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x)dx \qquad E[X] = \sum_i x_i p_X(x_i)$$

$$P(x_1 < X \le x_2) = \int_{x_1}^{x_2} f(x)dx$$





Lane 2020

Measures of location

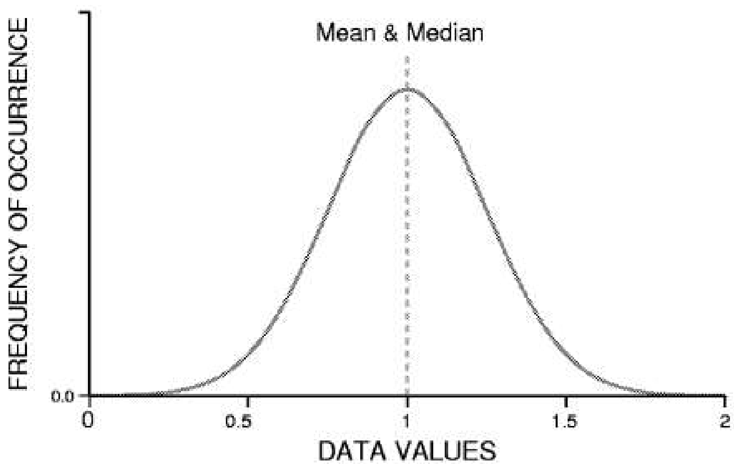


Figure 1.2 Density Function for a Normal Distribution

- Mean (P)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Median (NP)

median ($P_{0.50}$) = $X_{(n+1)/2}$ when n is odd, and

median ($P_{0.50}$) = $\frac{1}{2}(X_{(n/2)} + X_{(n/2)+1})$ when n is even.

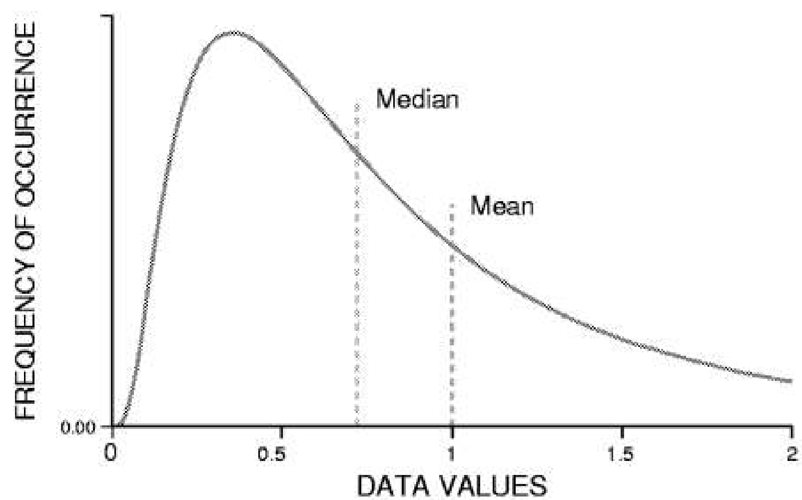


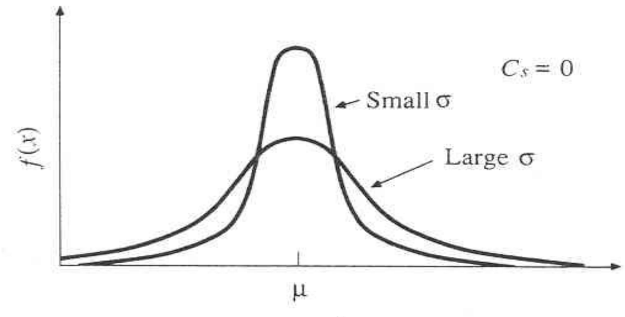
Figure 1.1 Density Function for a Lognormal Distribution

Measures of spread

- Standard deviation (P)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$$

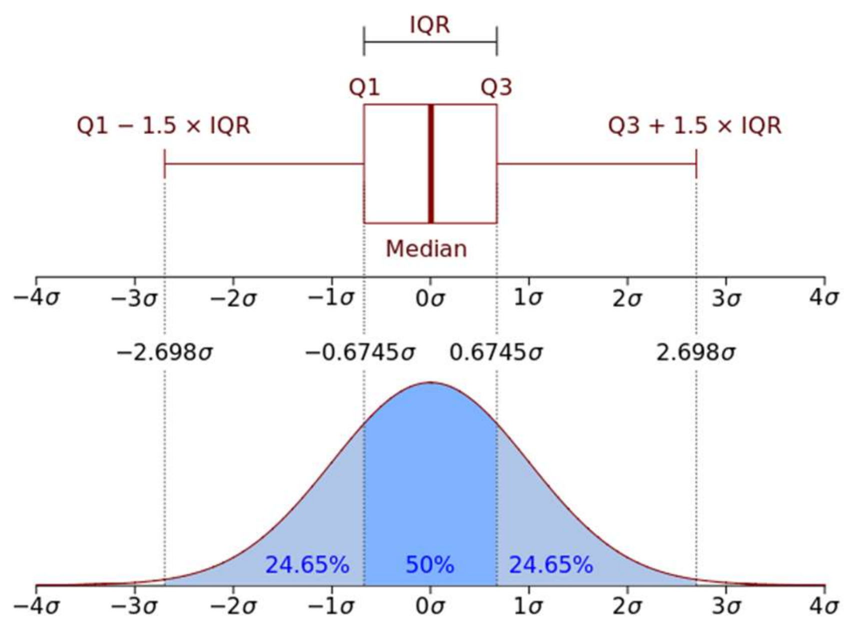
$$S = \sqrt{S^2}$$



- CV (P)

$$\mathrm{CV} = S/\bar{X}$$

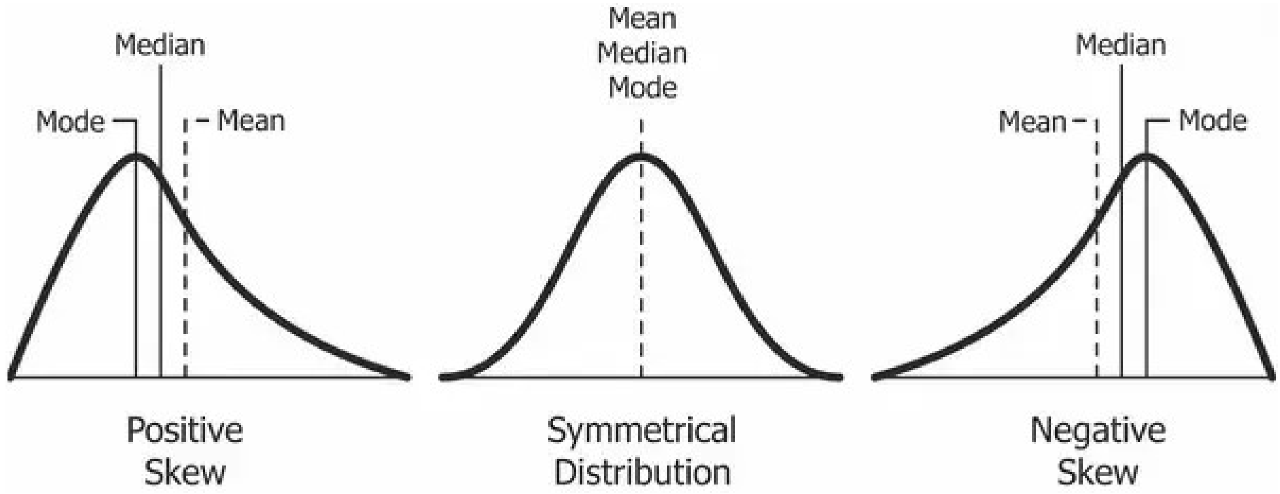
- IQR (NP)



COLLEGE of ENGINEERING
UtahStateUniversity

Lane 2020

Measures of skewness



- Skewness (P)

$$g = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (x_i - \overline{X})^3$$

Frequency Analysis

- The probability that X exceeds a given event discharge x_p is:

$$F_x(x) = P(X \geq x_p) = p$$

- The **return period (T)** corresponding to this exceedance probability is:

$$T = 1/p$$

- So, the 100-year return period is an event with an exceedance probability $p =$ **0.01** or a non-exceedance probability $1 - p =$ **0.99**

Frequency Analysis

- Find the probability that $X \geq x_T$ ***at least once*** in N years

$$p = P(X \geq x_T)$$

$$P(X < x_T) = (1 - p)$$

$$P(X \geq x_T \ at\ least\ once\ in\ N\ years) = 1 - P(X < x_T\ all\ N\ years)$$

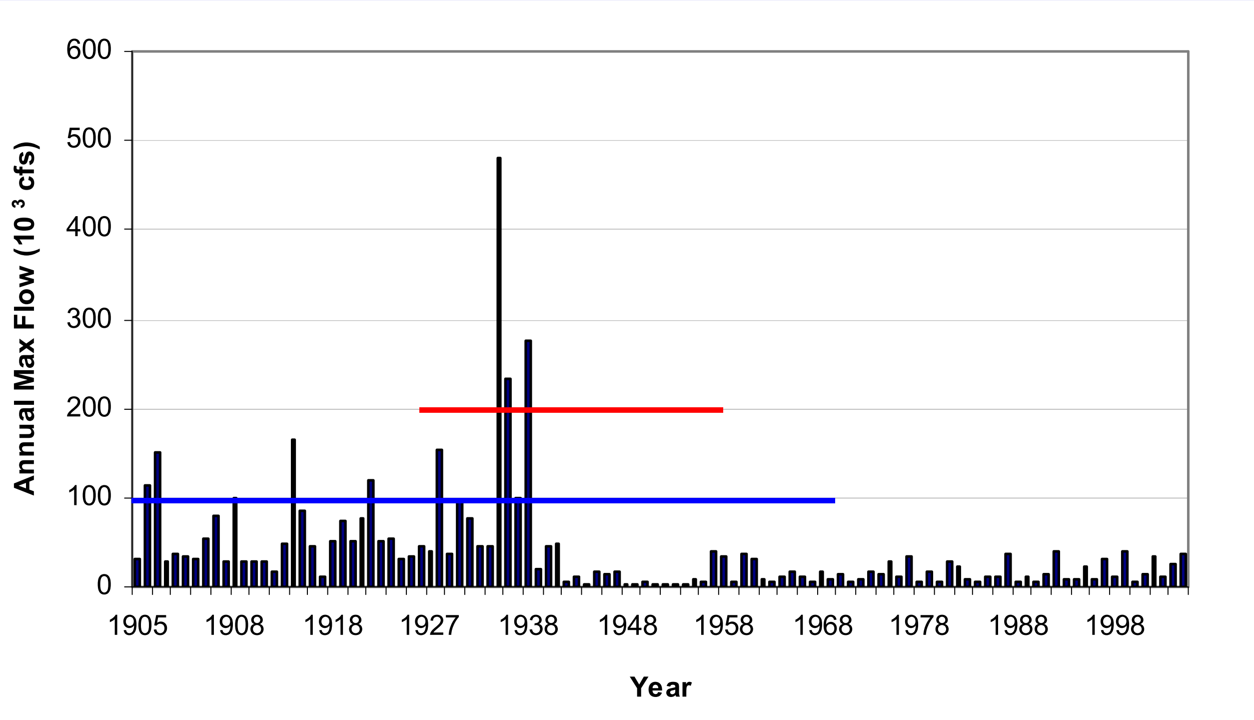
$$= 1 - (1-p)^N = 1 - \left(1 - \frac{1}{T}\right)^N$$

COLLEGE of ENGINEERING
Utah State University

Lane 2020

Frequency Analysis

- Annual maximum discharge for 106 years on the Colorado River



x_T = 200,000 cfs

No. of occurrences = 3

P=

T=

If x_T = 100,000 cfs

No. of occurrences =8

P = 8/106=7.5%

T = 106/8 = 13.5 yrs

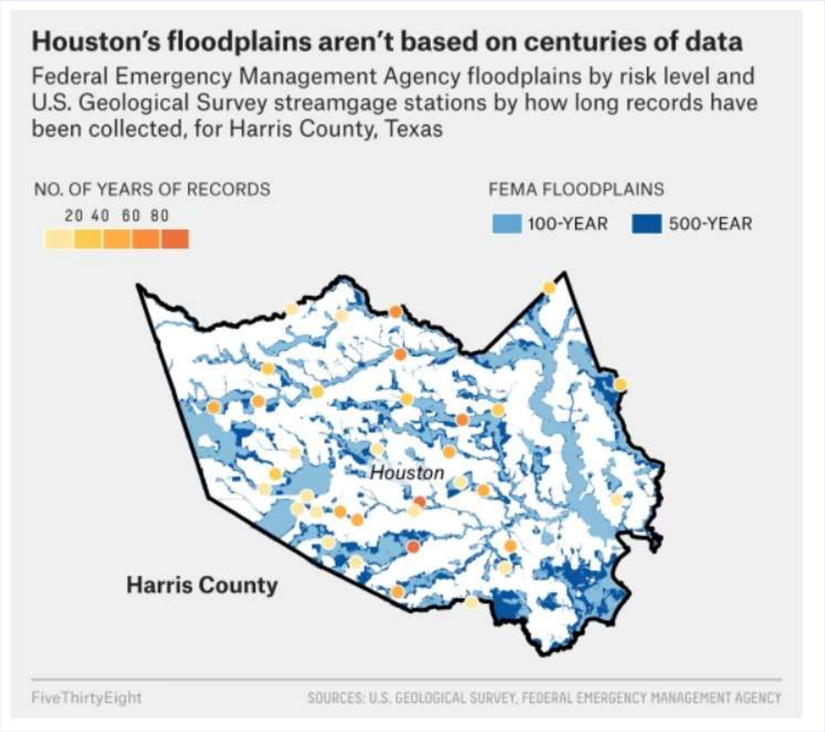
P(X ≥ 100,000 cfs *at least once* in the next 5 years) = 1- $(1-.075)^5$ = 32%

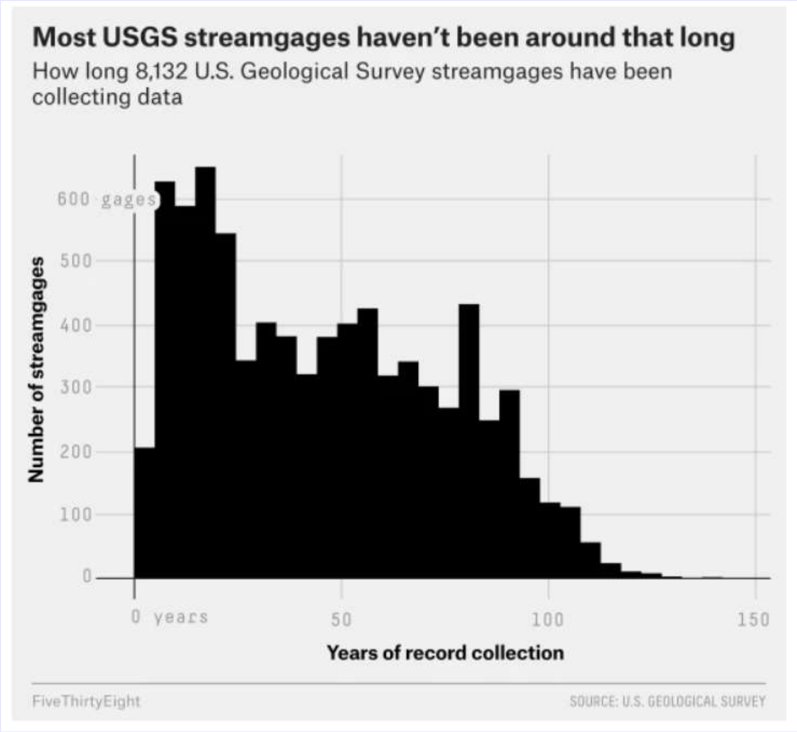
COLLEGE *of* ENGINEERING
UtahStateUniversity

Lane 2020

The ‘100-year flood’

Probability that $X \geq x_T$ *at least* once in 100 years $=1-(1-1/100)^{100}$ = **63.4%**





FiveThirtyEight, USGS

Assumptions:

- Independent observations
- From same PDF
- Stationarity

Lane 2020

Random Variables

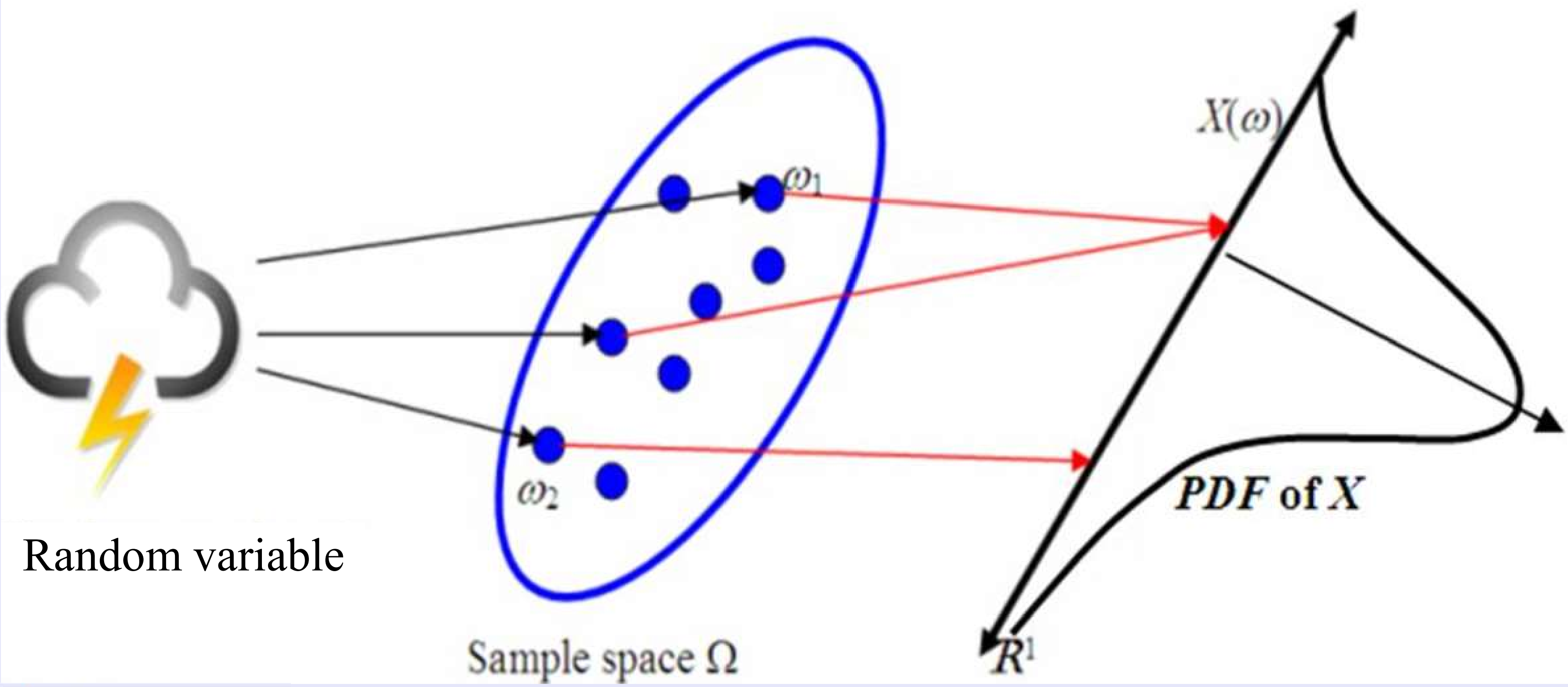
- Variables that demonstrate variability that is *not sufficiently explained* by analytical measures of a physical process
- Hydrologic processes are often random variables
  (e.g. precipitation, runoff)
- Random variable X is described by a **probability distribution,** a set of probabilities associated with the values in the random variable's sample space
- Probability statistics provide **models** to deal with uncertainty of random variables so we can still **quantify processes**

XKCD

COLLEGE of ENGINEERING
UtahStateUniversity

Lane 2020

Random Variables



Random variable

Probability Distributions

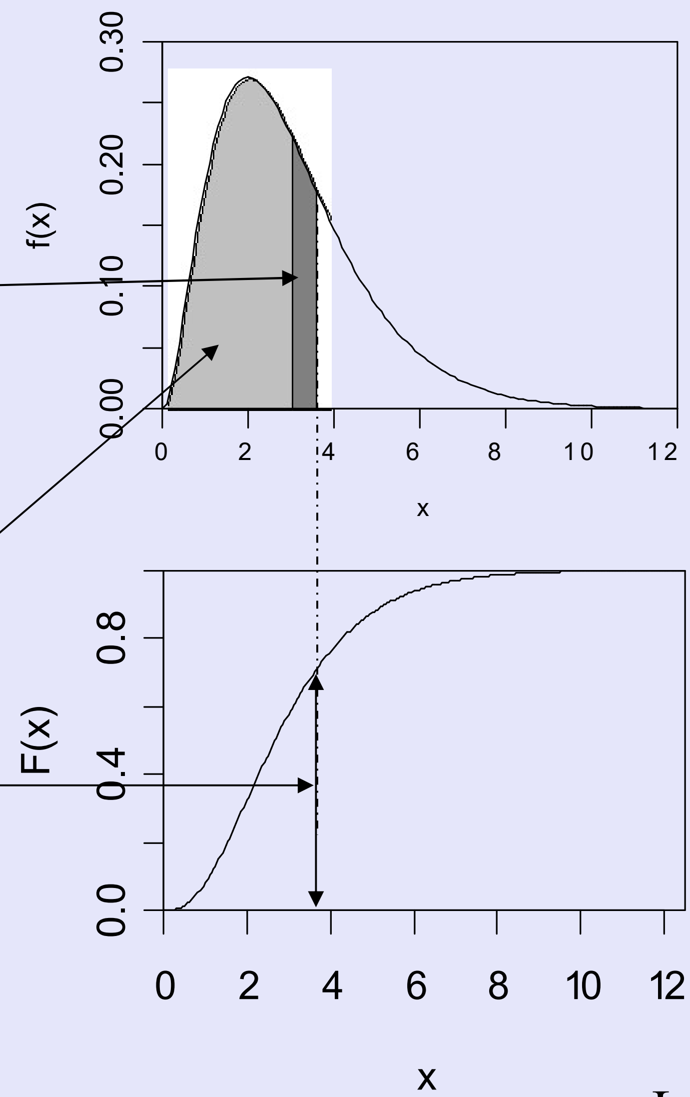
Probability density function (PDF)

$$P(x_1 < X \le x_2) = \int_{x_1}^{x_2} f(x)dx$$

Cumulative distribution function (CDF)

$$F(x) = P(X \le x) = \int_{-\infty}^{x_2} f(x)dx$$

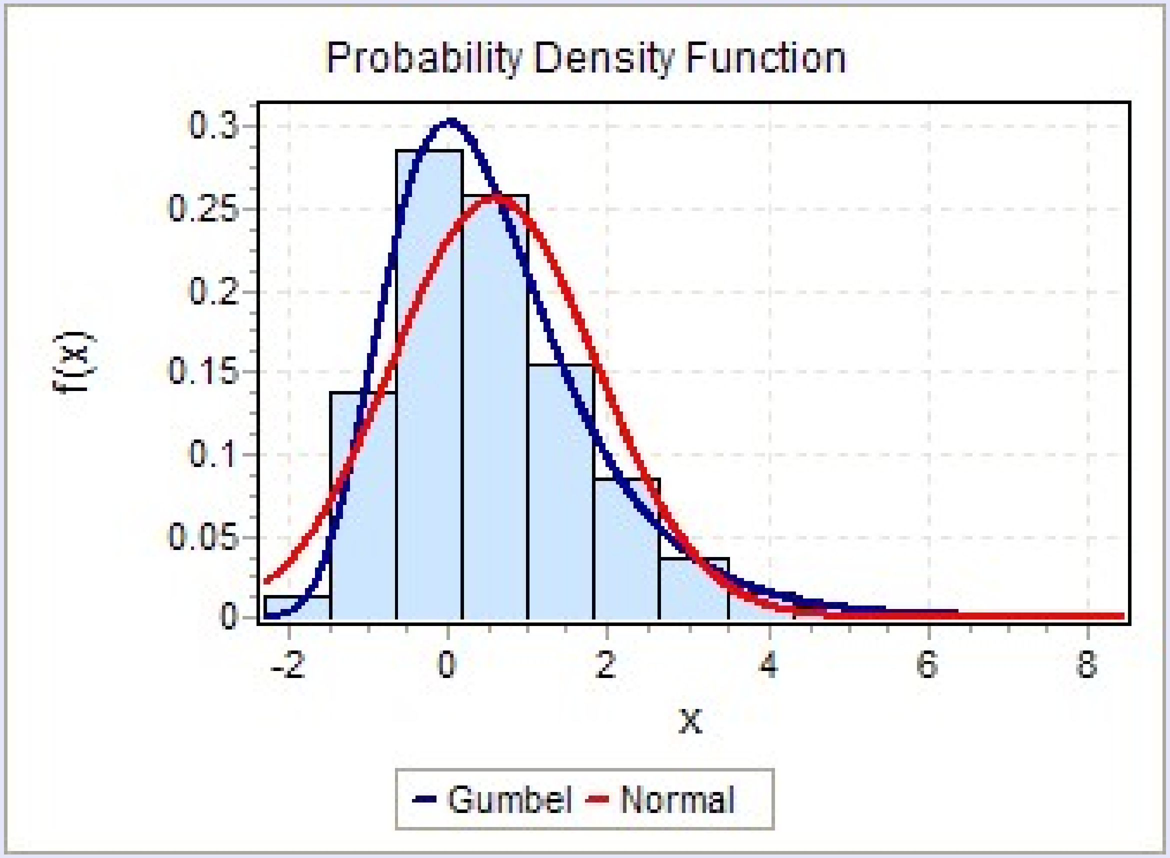
$$f(x) = \frac{dF}{dx}$$



Lane 2020

Probability Distributions

- Many different distributions and analytical expressions



COLLEGE of ENGINEERING UtahStateUniversity

Lane 2020

Probability Distributions

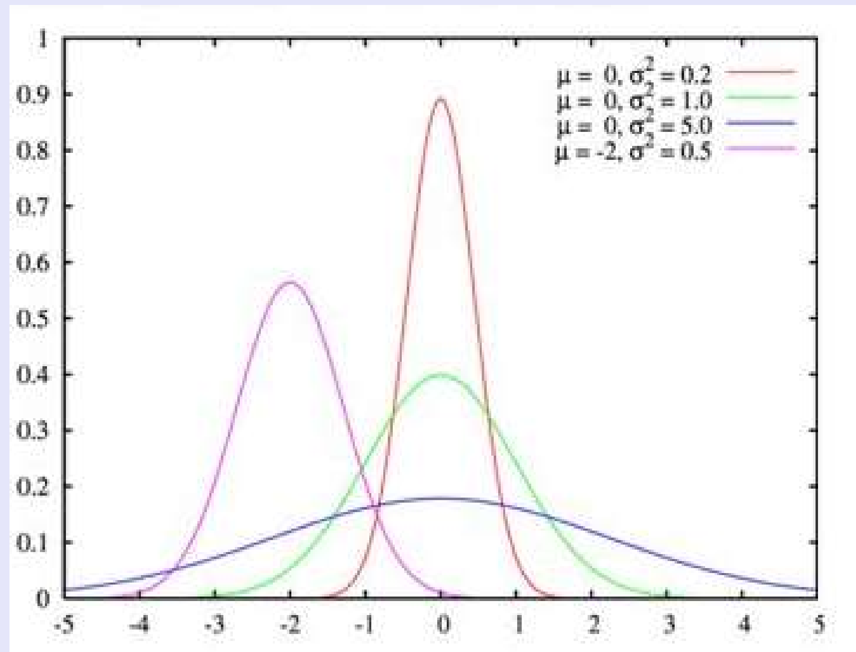
Normal Distribution

Central limit theorem – if X is the **sum** of *n* independent and identically distributed random variables, with increasing *n* the distribution of X trends towards normal regardless of the distribution of random variables.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ is the mean and σ is the standard deviation

- Most *average* variables
- Many error distributions



COLLEGE of ENGINEERING
UtahStateUniversity

Lane 2020

Probability Distributions

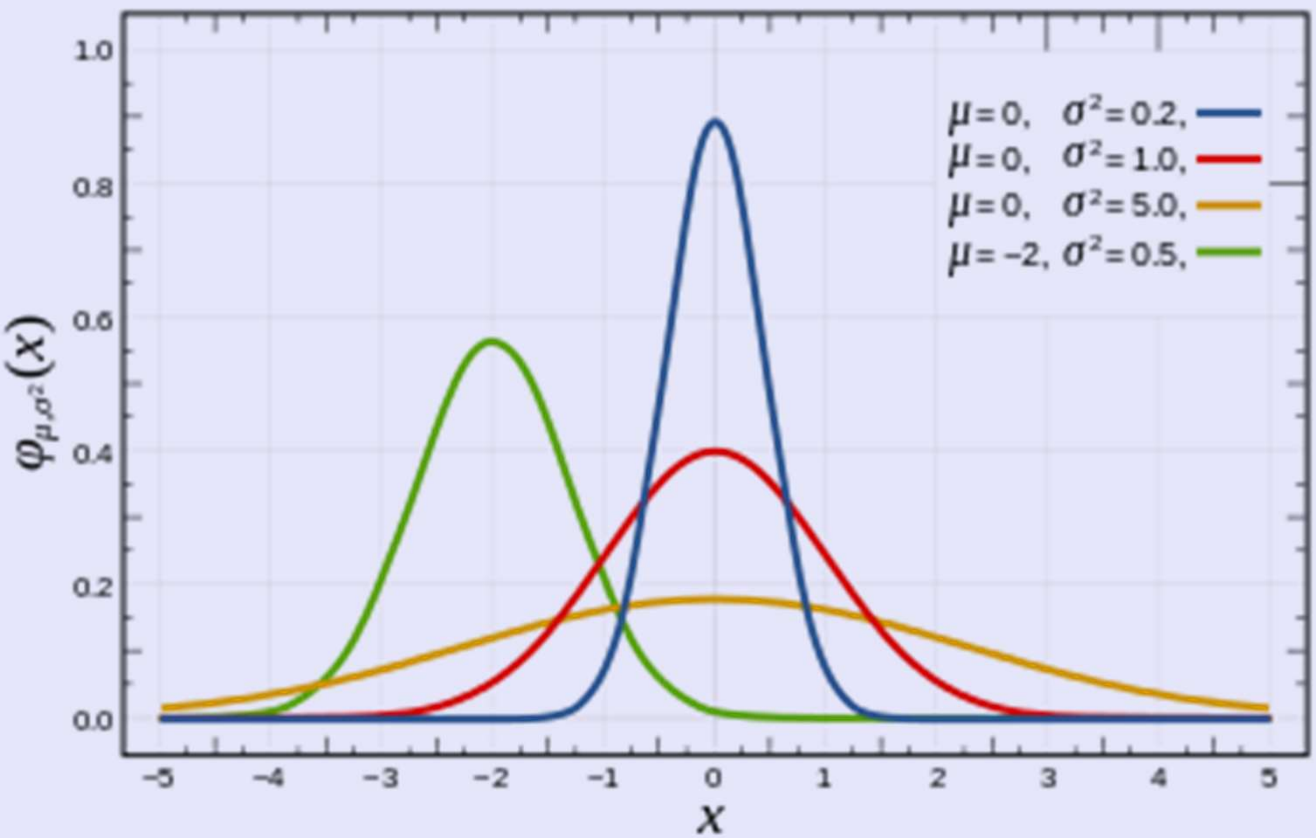
- Normal family
  - Normal (average annual P and Q),
  - Log-normal (hydraulic conductivity)

- Generalized extreme value (GEV) family
  - Gumbel (annual max streamflow), GEV, and Weibull (7-day min flow)

- Pearson family
  - Exponential, Log-Pearson type III (annual max flows)

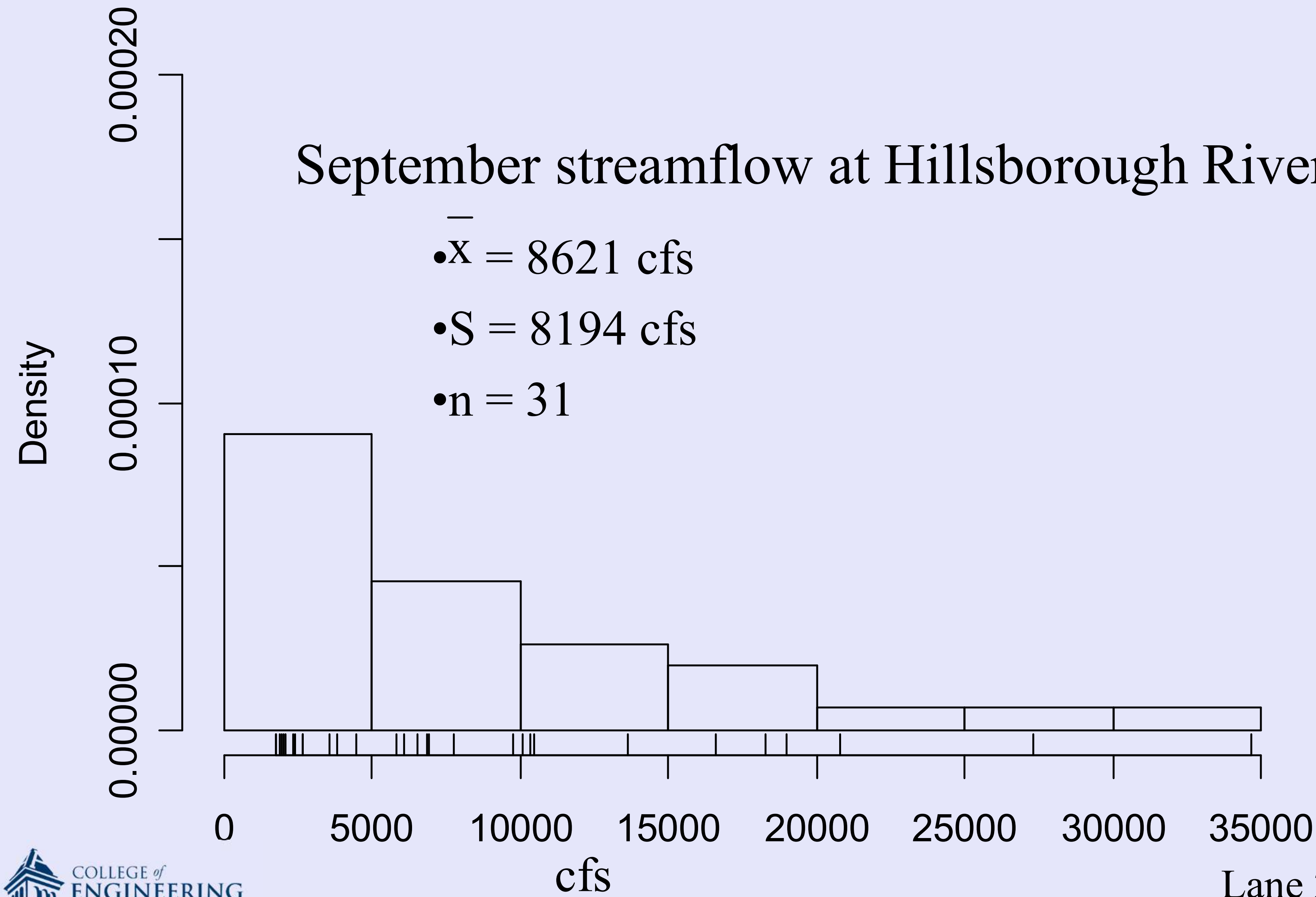
COLLEGE of ENGINEERING UtahStateUniversity

Characterizing Probability Distributions



Lane 2020

Fitting a probability distribution to data



September streamflow at Hillsborough River

- $\bar{x}$ = 8621 cfs
- S = 8194 cfs
- n = 31

Lane 2020

COLLEGE of ENGINEERING
UtahStateUniversity

Fitting a probability distribution to data

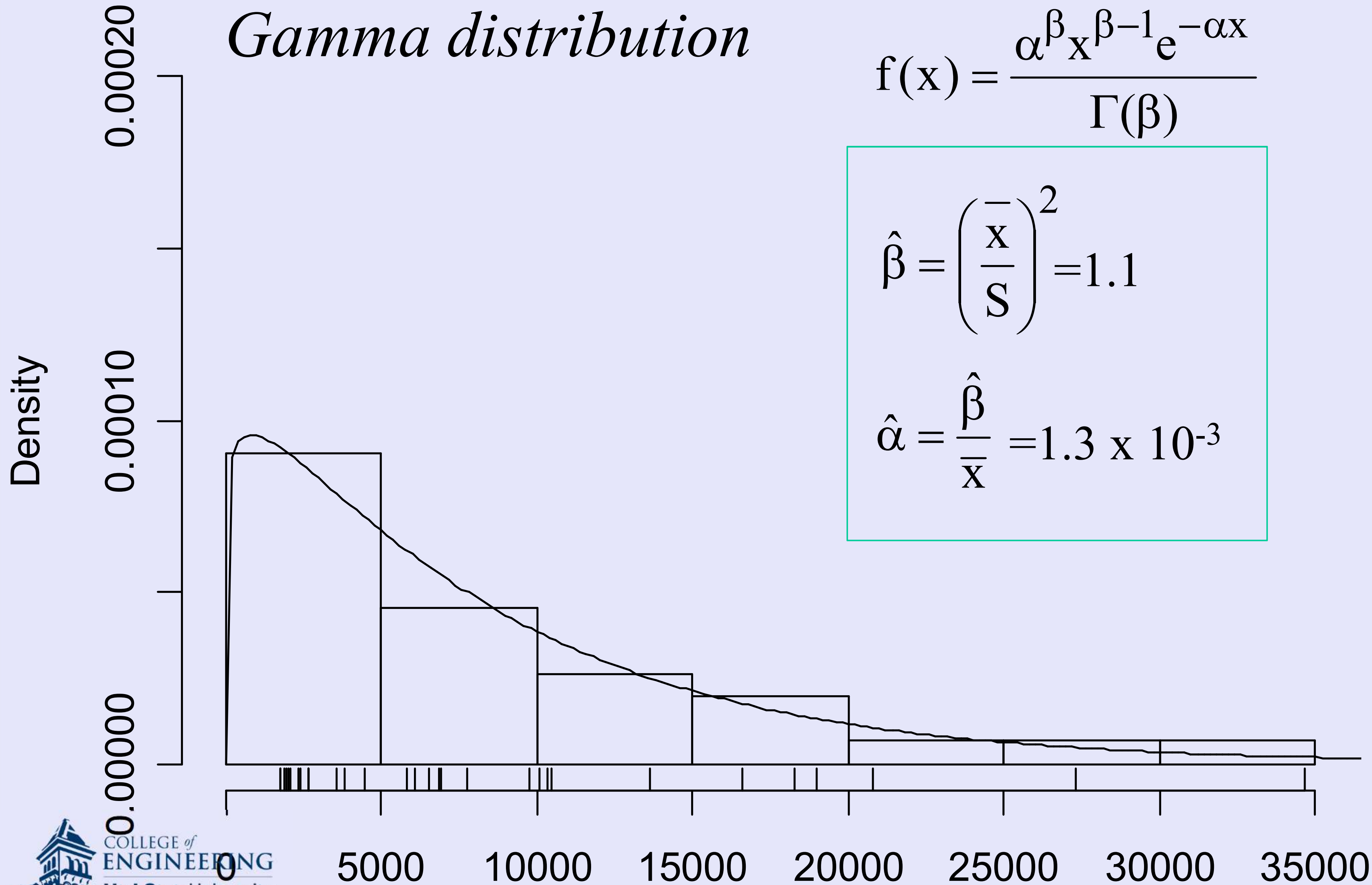
Set the **sample** moments as the estimate for the **population** parameters

$$\hat{E}(X) = \overline{x}\,; \hat{V}ar(X) = \sigma^2$$

	Population		Sample
Mean	$E(X) = \int_{-\infty}^{\infty} x f(x) dx$	=	$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

COLLEGE of ENGINEERING
UtahStateUniversity

Fitting a probability distribution to data



$$f(x) = \frac{\alpha^{\beta} x^{\beta-1} e^{-\alpha x}}{\Gamma(\beta)}$$

$$\hat{\beta} = \left(\frac{\bar{x}}{S}\right)^2 = 1.1$$

$$\hat{\alpha} = \frac{\hat{\beta}}{\bar{x}} = 1.3 \text{ x } 10^{-3}$$

Fitting a probability distribution to data

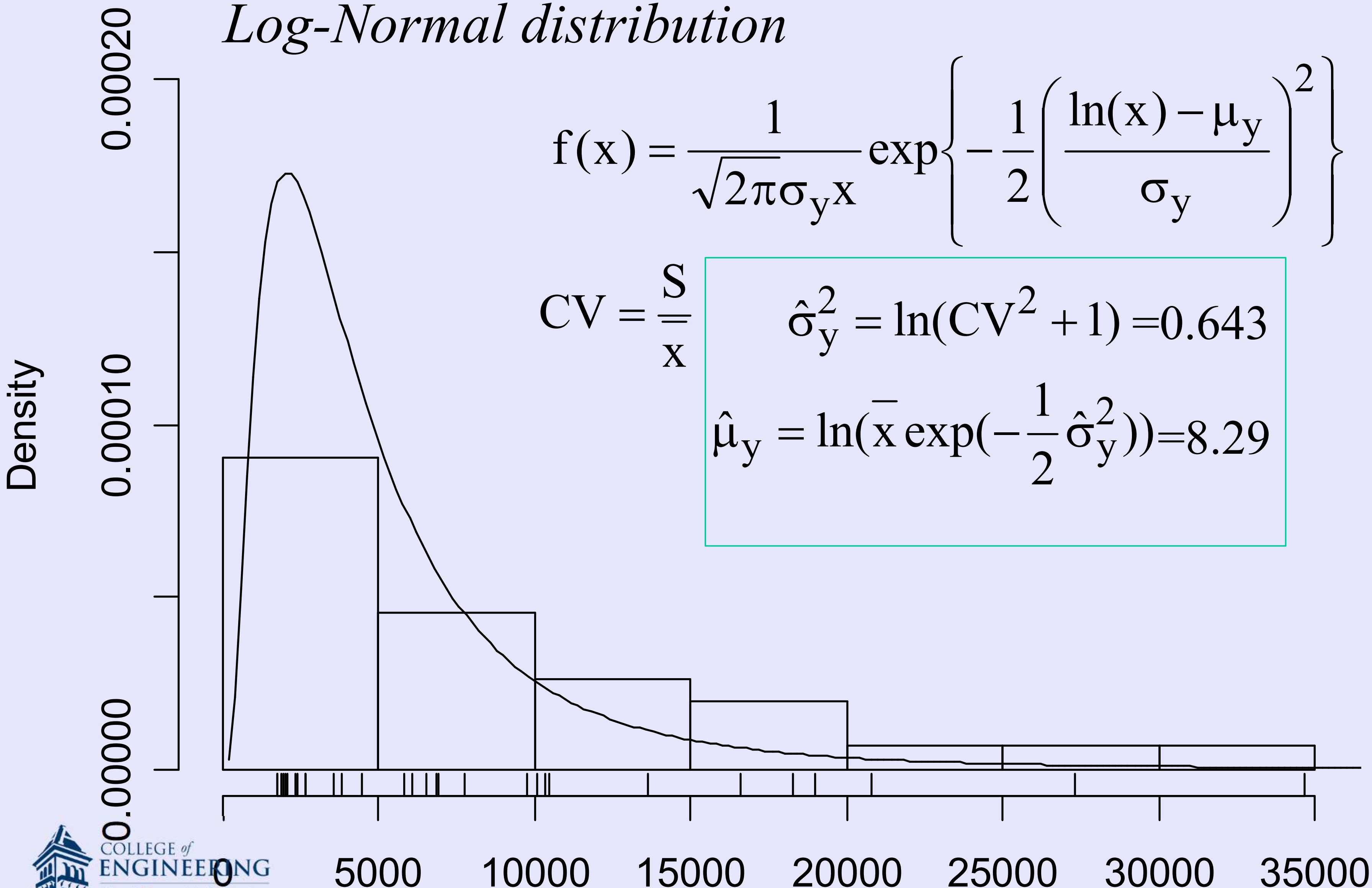
Log-Normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_y x} \exp\left\{ -\frac{1}{2}\left(\frac{\ln(x) - \mu_y}{\sigma_y} \right)^2 \right\}$$

$$CV = \frac{S}{\bar{x}}$$

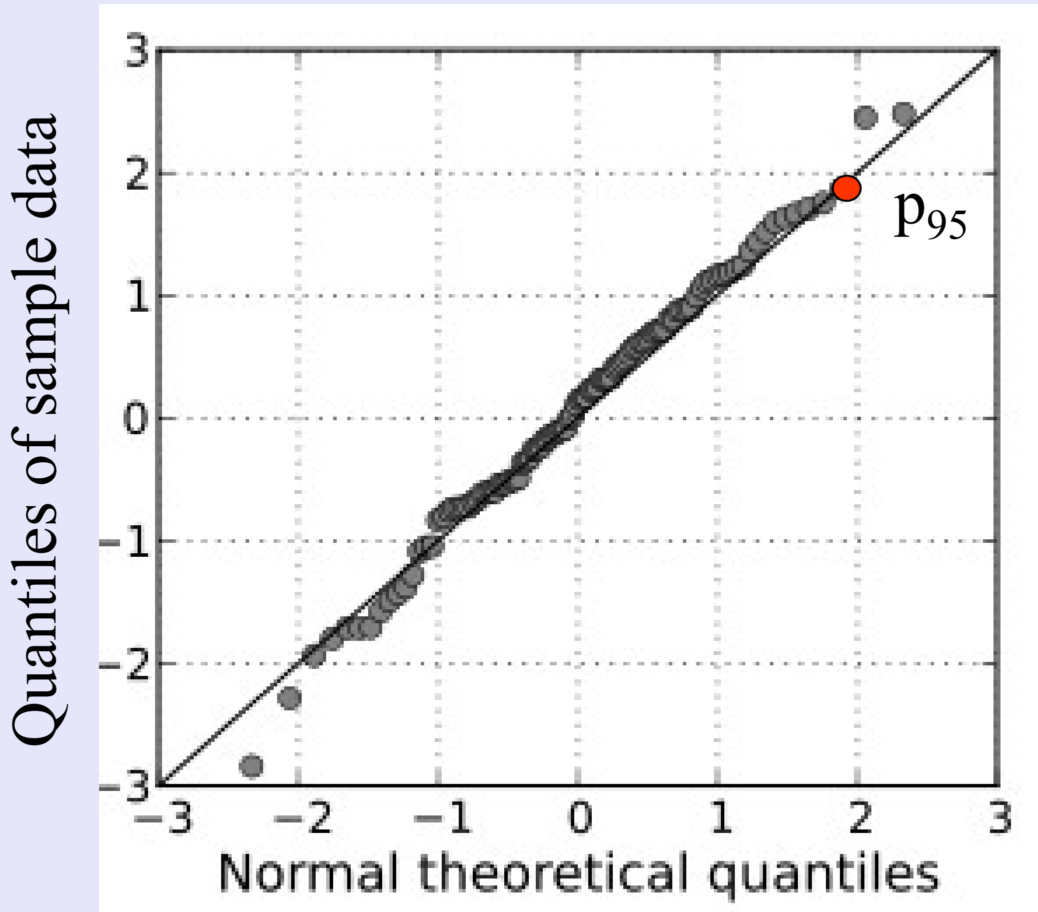
$$\hat{\sigma}_y^2 = \ln(CV^2 + 1) = 0.643$$

$$\hat{\mu}_y = \ln(\bar{x}\exp(-\frac{1}{2}\hat{\sigma}_y^2)) = 8.29$$



Quantifying Uncertainty

Quantile - Quantile plots



A graphical "goodness of fit" test

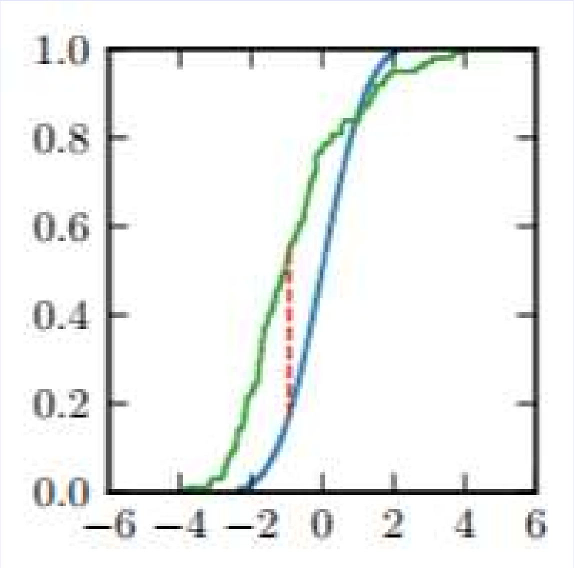
COLLEGE of ENGINEERING UtahStateUniversity

Lane 2020

Quantifying Uncertainty

Kolmogorov-Smirnov Test

- Computes the largest difference between the target CDF $F_X(x)$ and the observed CDF, $F^*(X)$.
- The test statistic D_2 is:



$$D_2 = \max_{i=1}^{n}\left[\left|F^*(X^{(i)}) - F_X(X^{(i)})\right|\right]$$

$$= \max_{i=1}^{n}\left[\left|\frac{i}{n} - F_X(X^{(i)})\right|\right]$$

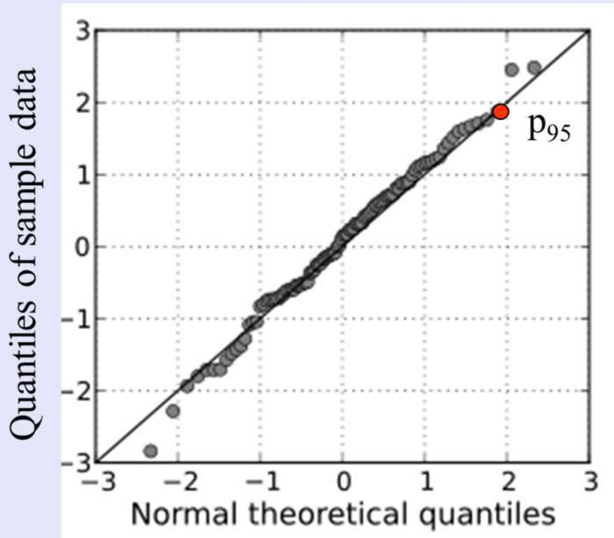
where $X^{(i)}$ is the ith largest observed value in the random sample of size n.

Lane 2020

Quantifying Uncertainty

Probability Plot Correlation Coefficient

$$r = \frac{\sum (x_{(i)} - \overline{x})(w_i - \overline{w})}{\left[\left(\sum (x_{(i)} - \overline{x})^2 \sum (w_i - \overline{w})^2\right)\right]^{0.5}} \quad (7.74)$$



Probability Plot Correlation Coefficient test employs the correlation r between the ordered observations $x_{(i)}$ and the corresponding fitted quantiles $w_i = G^{-1}(p_i)$, determined by plotting positions p_i for each $x_{(i)}$. Values of r near 1.0 suggest that the observations could have been drawn from the fitted distribution: r measures the linearity of the probability plot providing a quantitative assessment of fit. If $\overline{x}$ denotes the average value of the observations and $\overline{w}$ denotes the average value of the fitted quantiles, then

Lane 2020

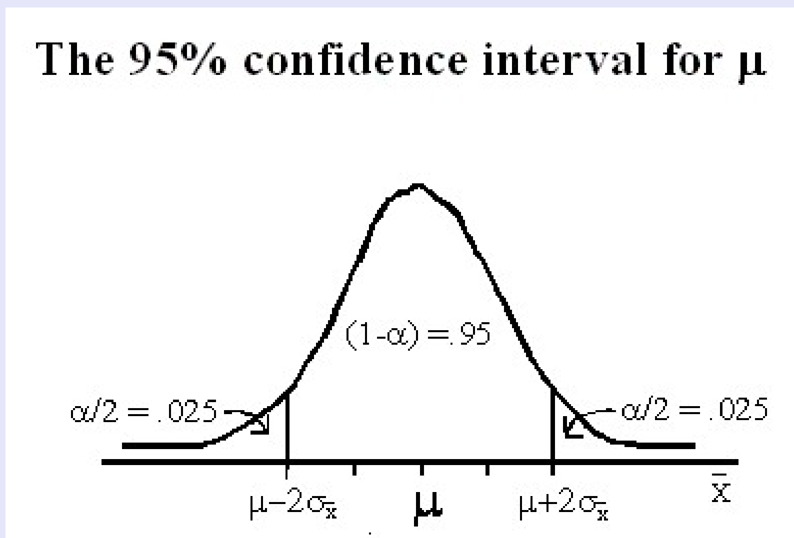
COLLEGE of ENGINEERING
UtahStateUniversity

Quantifying Uncertainty

Normal Distribution

"*95% confidence interval*": the true population mean will be contained in these intervals an average of 95% of the time

For a Normal distribution, P[$\mu - 1.96\sigma \le$ *true mean* $\le \mu + 1.96\sigma$] = 0.95

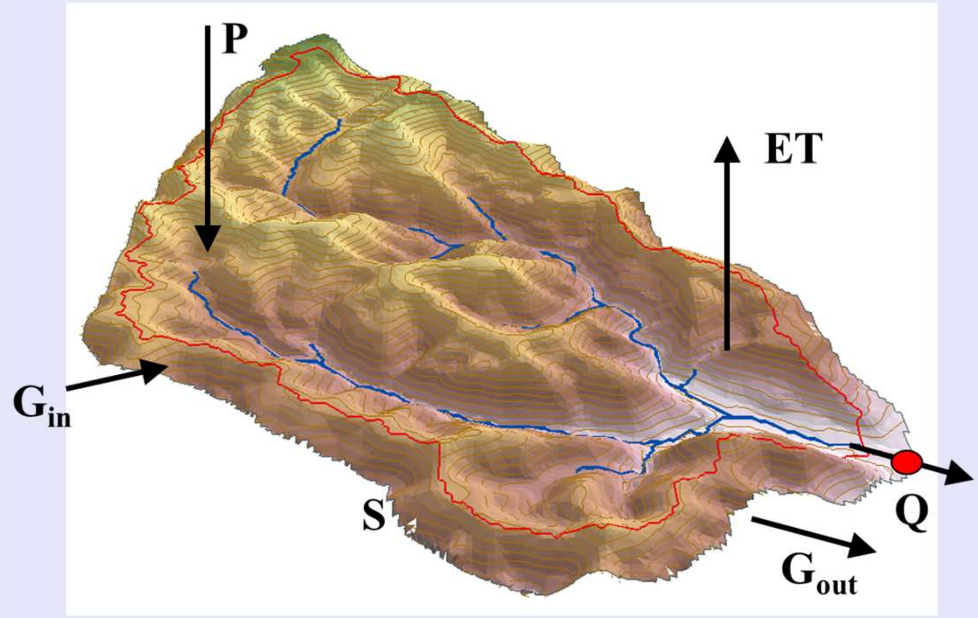


Z-scores

α	$(1-\alpha)$	z
.10	.90	1.645
.05	.95	1.96
.01	.99	2.575

Uncertainty in catchment water balance

Estimate average annual ET and error



Tarboton, 2003

A very useful resource, updated 2020!

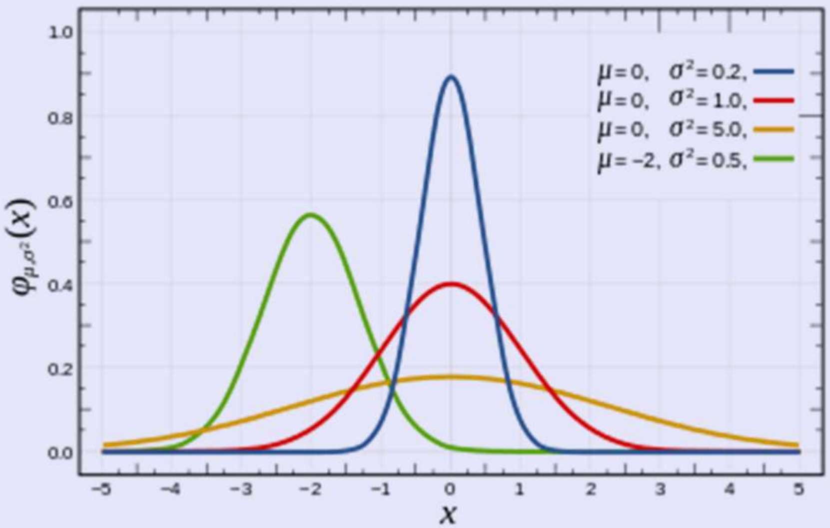
USGS
science for a changing world

- Chapter 1 Summarizing Univariate Data
- Chapter 2 Graphical Data Analysis
- Chapter 3 Describing Uncertainty
- Chapter 4 Hypothesis Tests
- Chapter 5 Testing Differences Between Two Independent Groups
- Chapter 6 Paired Difference Tests of the Center
- Chapter 7 Comparing Centers of Several Independent Groups
- Chapter 8 Correlation
- Chapter 9 Simple Linear Regression
- Chapter 10 Alternative Methods for Regression
- Chapter 11 Multiple Linear Regression
- Chapter 12 Trend Analysis
- Chapter 13 How Many Observations Do I Need?
- Chapter 14 Discrete Relations
- Chapter 15 Regression for Discrete Responses
- Chapter 16 Presentation Graphics
- References Cited
- Index

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., and Gilroy, E.J., 2020, Statistical methods in water resources: U.S. Geological Survey Techniques and Methods, book 4, chapter A3, 458 p.,

https://doi.org/10.3133/tm4a3

Concepts to Understand



- Random variable
- PDF and CDF
- Expected value
- Parametric v. non-parametric
- Quantiles
- Method of Moments
- Flow exceedance
- Frequency/ return period
- Confidence intervals

COLLEGE of ENGINEERING
UtahStateUniversity

Lane 2020