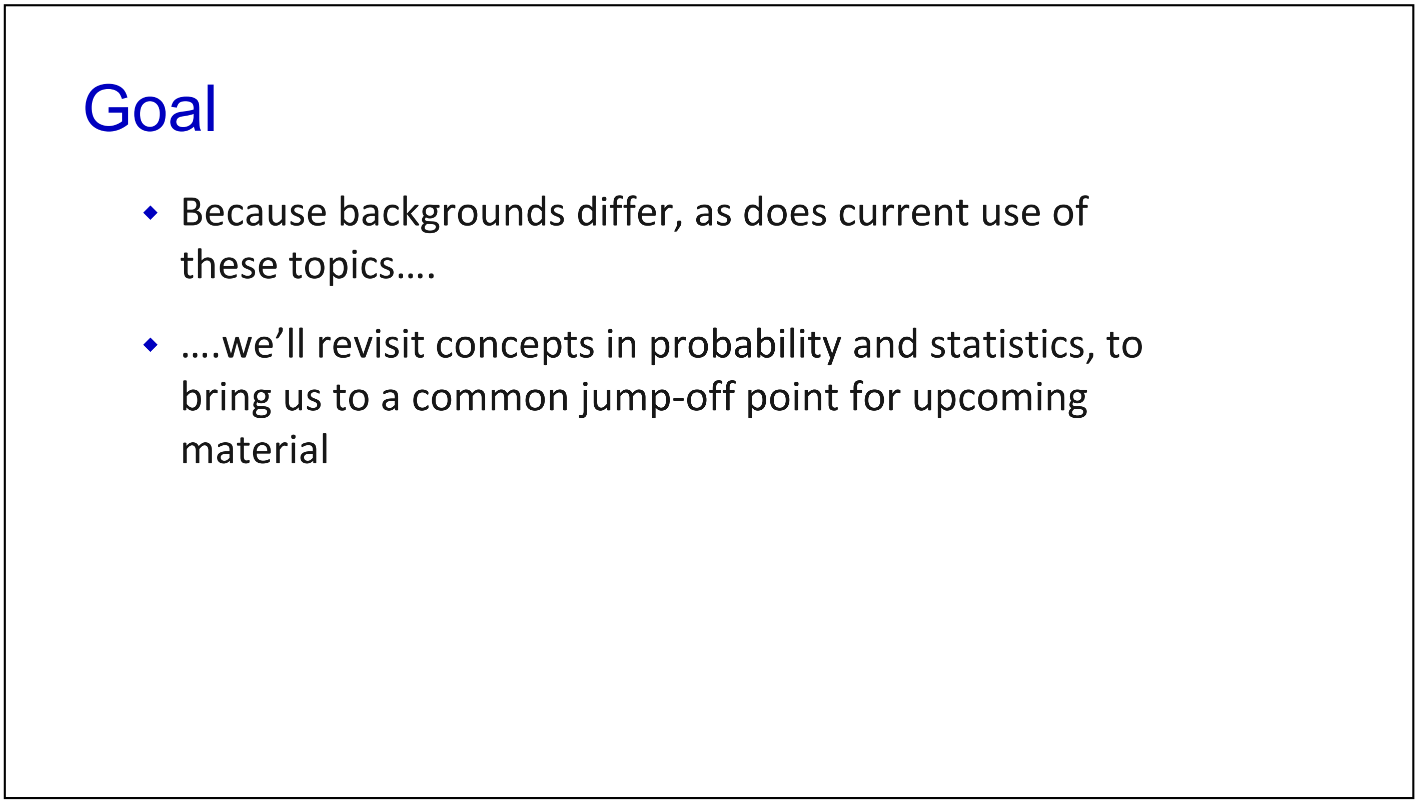
Basic Probability and Statistics

Flood Frequency Analysis

Beth Faber, PhD, PE

Greg Karlovits, PE, PH, CFM

Hydrologic Engineering Center, May 2022

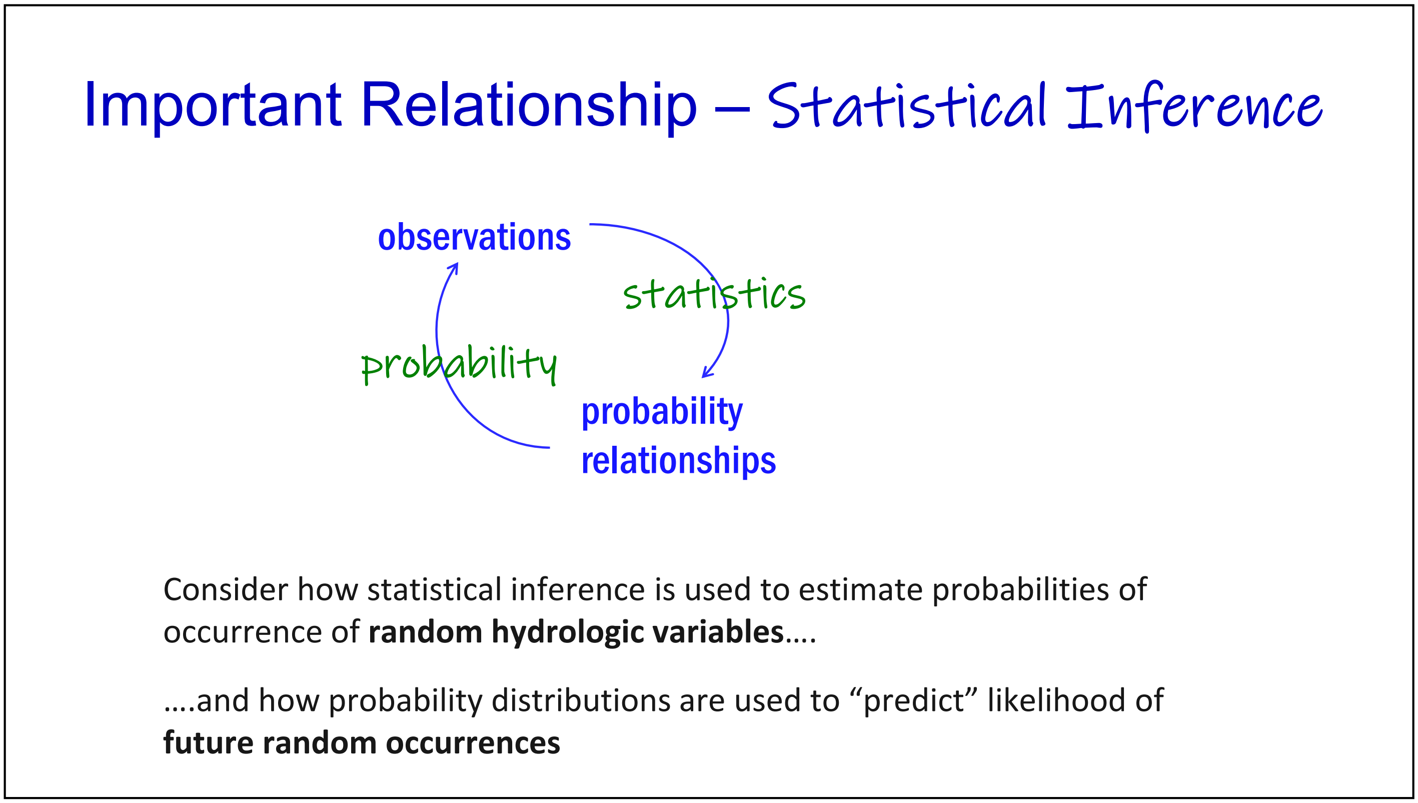


Goal

- Because backgrounds differ, as does current use of these topics….
- ….we’ll revisit concepts in probability and statistics, to bring us to a common jump-off point for upcoming material

We all have different backgrounds and experience with probability ideas, so this review of the college course will hopefully bring us to a similar place.

This lecture is kind of a slow walk through the basic topics, to give a chance to dwell on the ideas a bit. Much or most of it will be summarized more briefly in later lectures.

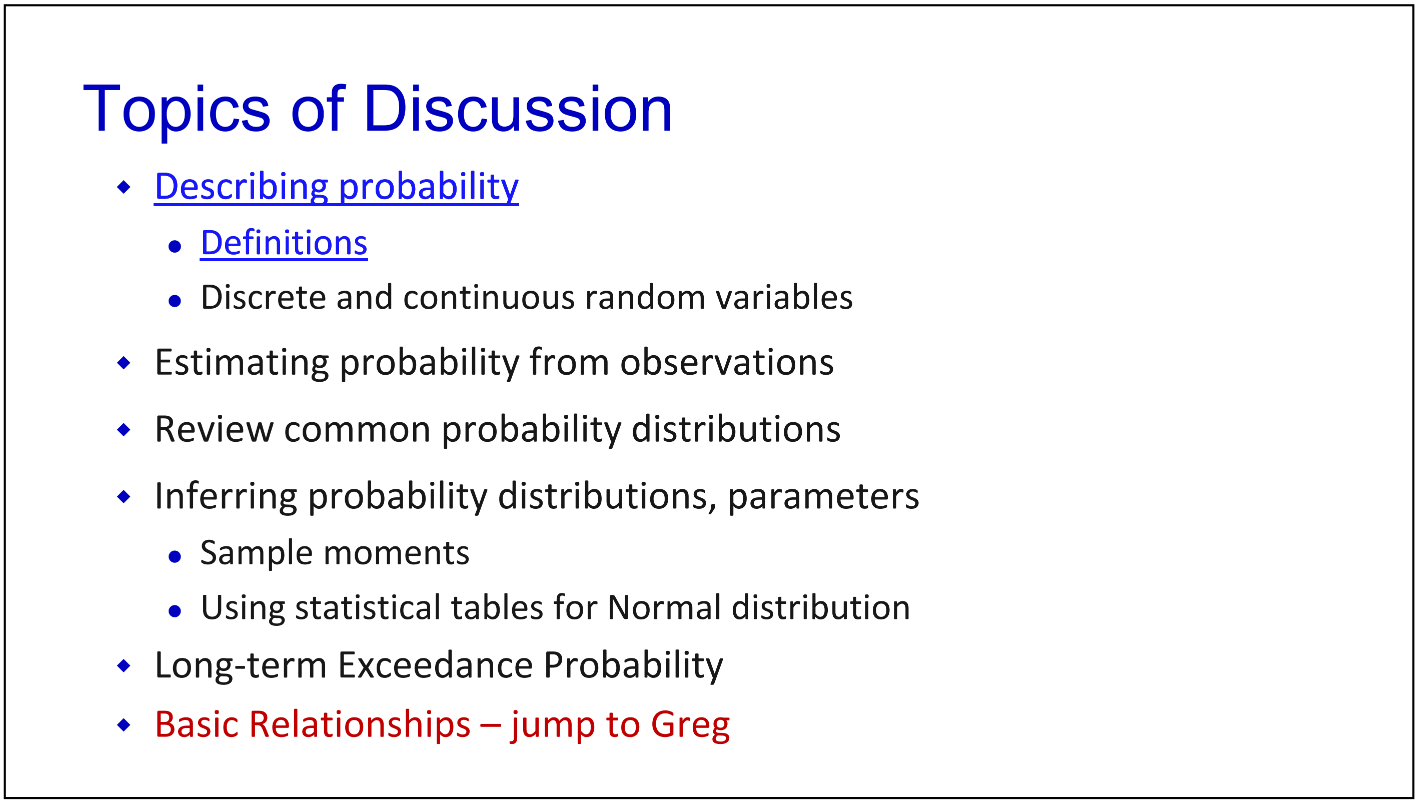


Important Relationship – Statistical Inference

Consider how statistical inference is used to estimate probabilities of occurrence of **random hydrologic variables**….

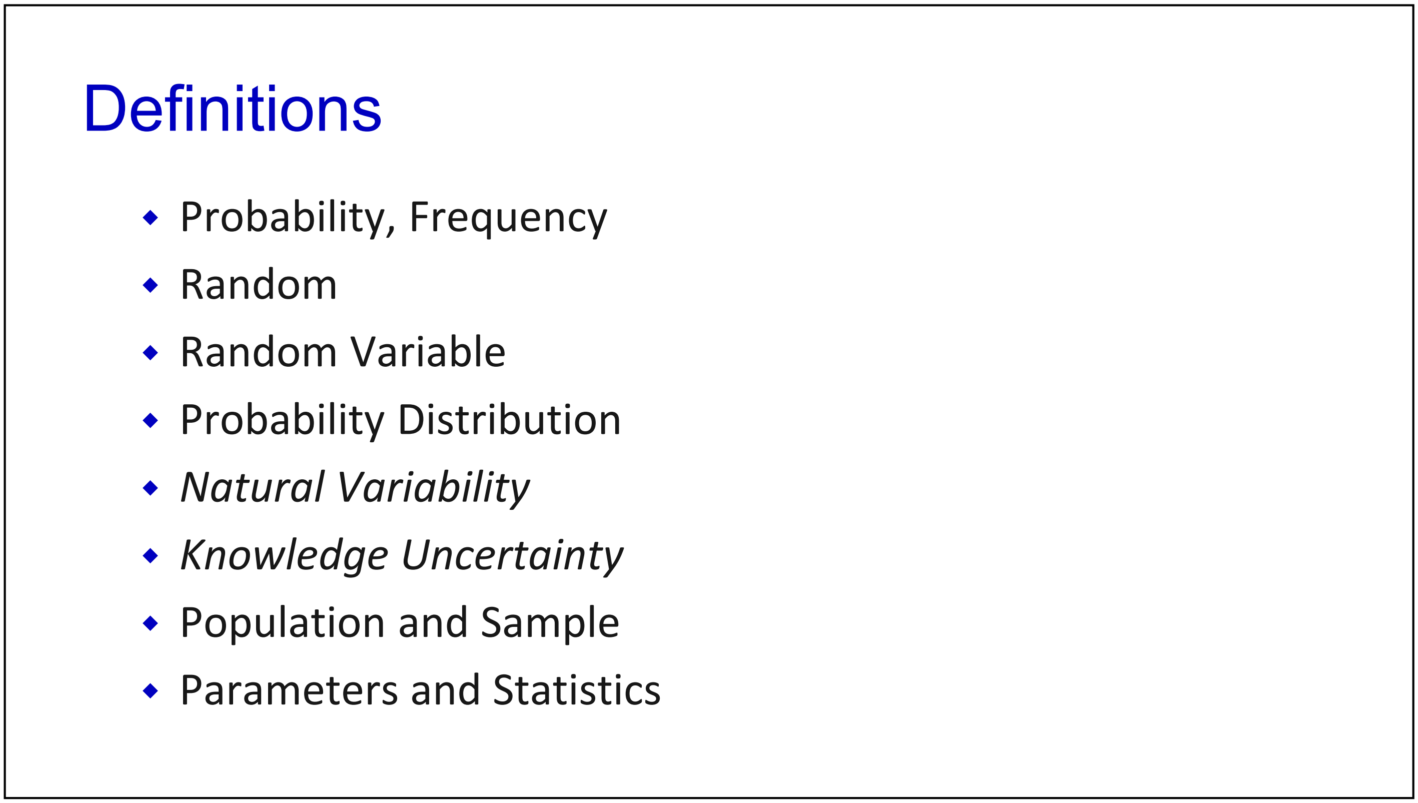
….and how probability distributions are used to “predict” likelihood of **future random occurrences**

This is an idea that summarizes the relationship between probability and basic statistics. We use statistical analysis to estimate probability from data or observations, and then use that probability description to determine the likelihood or frequency of future observations. Once those future observations occurred, we can use them to improve probability estimates, and so on…



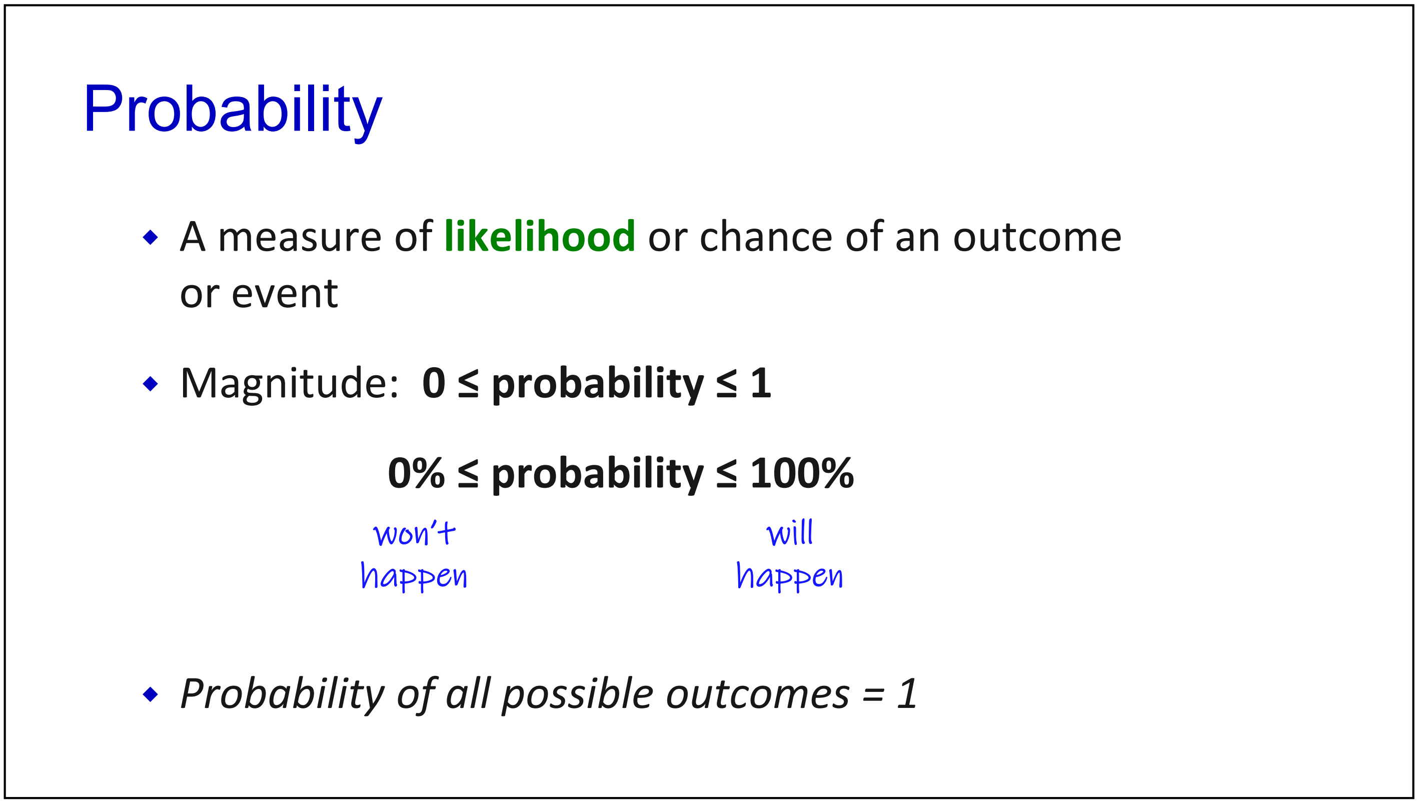
Topics of Discussion

- Describing probability
  - Definitions
  - Discrete and continuous random variables
- Estimating probability from observations
- Review common probability distributions
- Inferring probability distributions, parameters
  - Sample moments
  - Using statistical tables for Normal distribution
- Long-term Exceedance Probability
- Basic Relationships – jump to Greg



Definitions

- Probability, Frequency
- Random
- Random Variable
- Probability Distribution
- *Natural Variability*
- *Knowledge Uncertainty*
- Population and Sample
- Parameters and Statistics



Probability

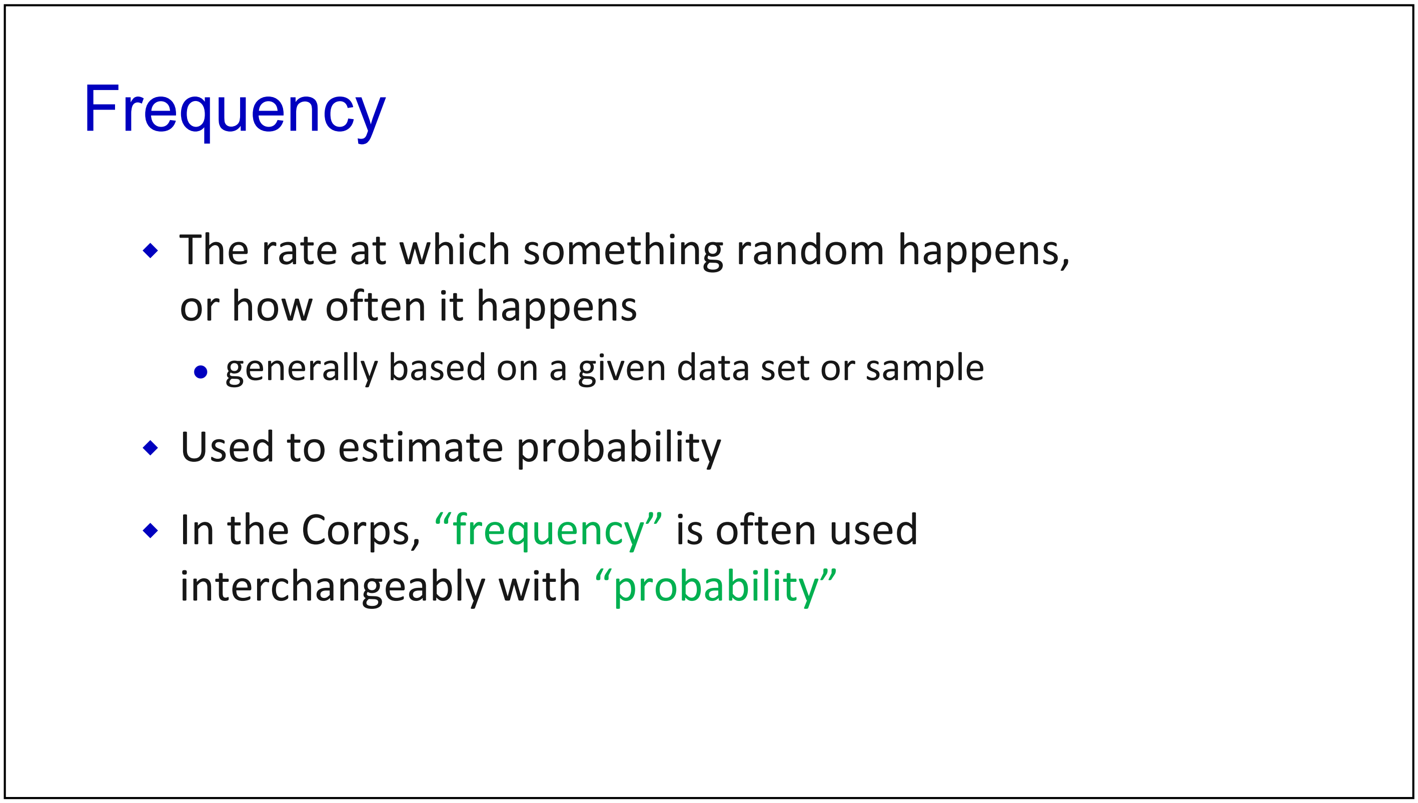
- A measure of **likelihood** or chance of an outcome or event
- Magnitude: **0 ≤ probability ≤ 1**

  0% ≤ probability ≤ 100%

  won't happen   will happen

- *Probability of all possible outcomes = 1*

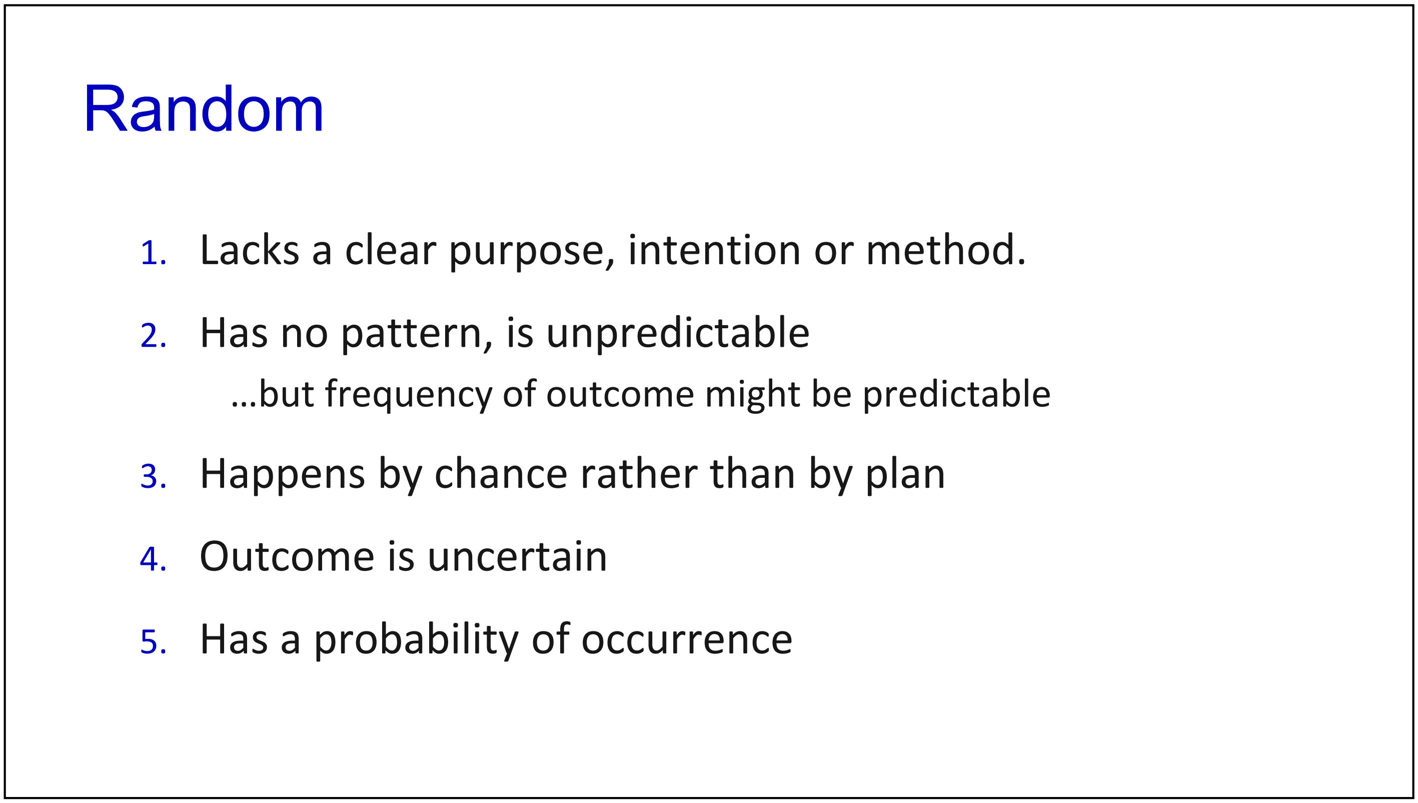
… what's really important in our dealings with probability is where the estimates come from.



Frequency

- The rate at which something random happens, or how often it happens
  - generally based on a given data set or sample
- Used to estimate probability
- In the Corps, "frequency" is often used interchangeably with "probability"

How likely something is to happen is closely tied with how often it happens.

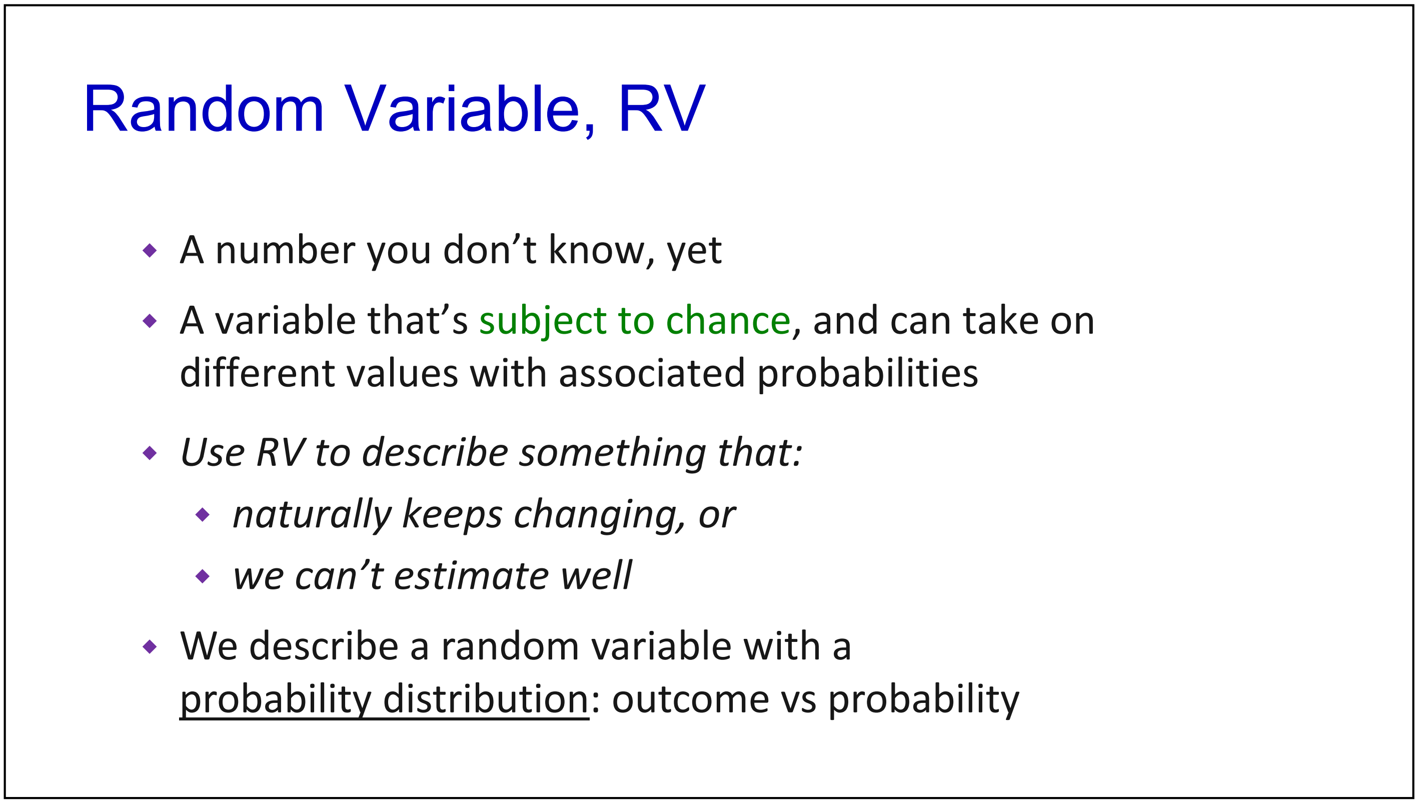


Random

1. Lacks a clear purpose, intention or method.
2. Has no pattern, is unpredictable

   …but frequency of outcome might be predictable
3. Happens by chance rather than by plan
4. Outcome is uncertain
5. Has a probability of occurrence

There are many definitions of “Random.” Some outcomes are naturally random. But the word is sometimes used to describe a process so complex we can’t understand it completely enough to predict.



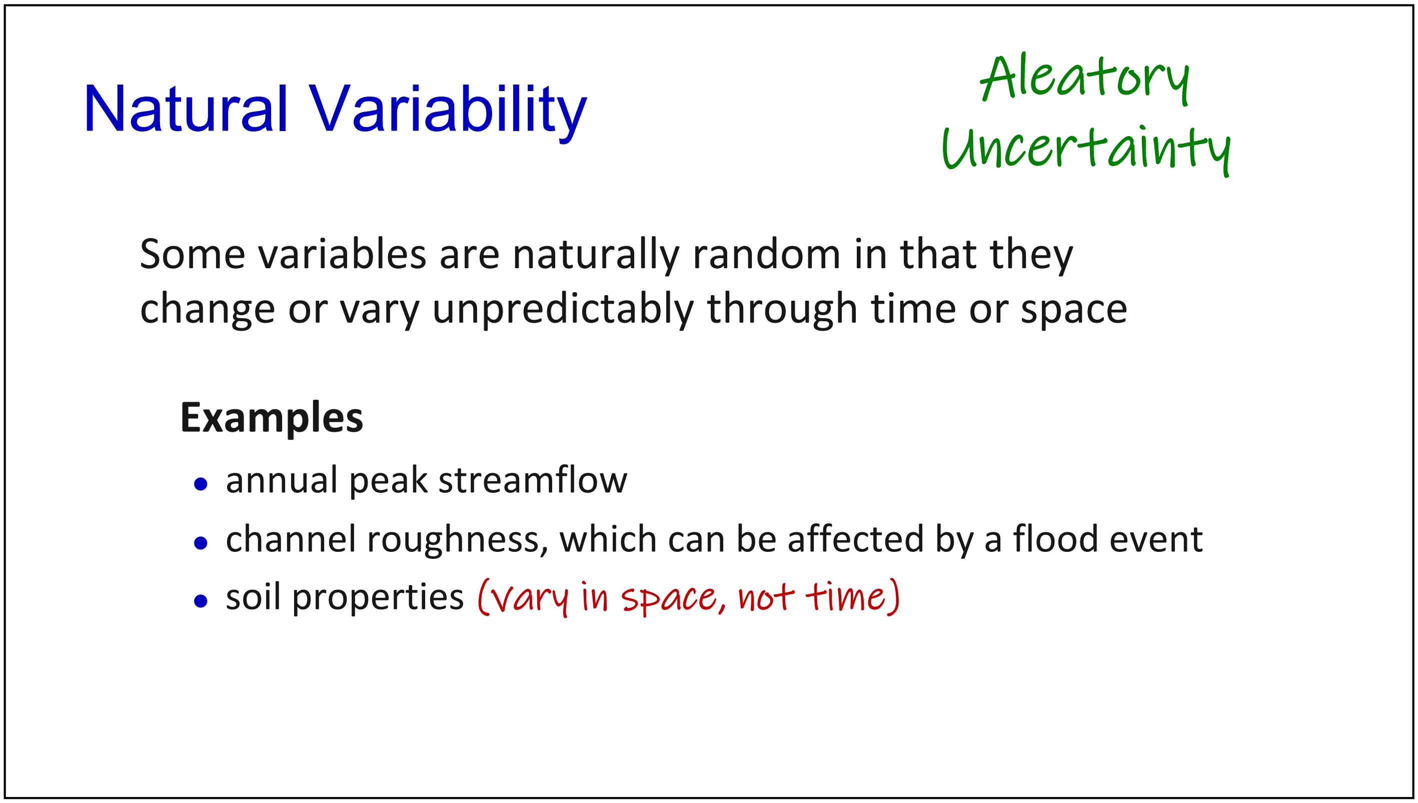
Random Variable, RV

- A number you don't know, yet
- A variable that's subject to chance, and can take on different values with associated probabilities
- *Use RV to describe something that:*
  - *naturally keeps changing, or*
  - *we can't estimate well*
- We describe a random variable with a probability distribution: outcome vs probability

Any value that's relevant to our analysis, that either hasn't happened yet or can't be measured accurately, is a random variable.

Probability Distribution

- A relationship between a random variable’s value and probability
- A function that describes all possible values of a random variable, and their likelihoods
- We use probability distributions to describe what is random, and what is unknown



Natural Variability

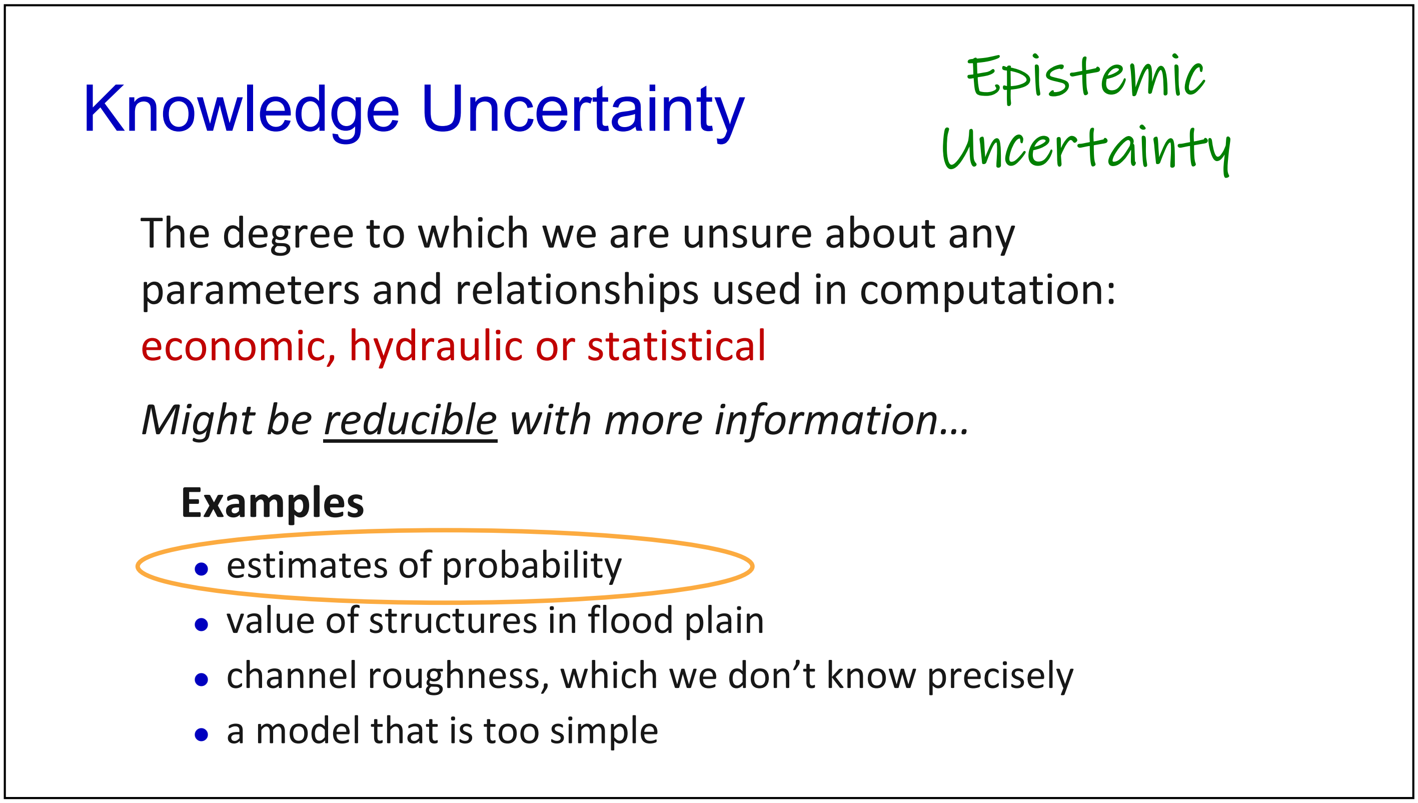
Aleatory Uncertainty

Some variables are naturally random in that they change or vary unpredictably through time or space

Examples

- annual peak streamflow
- channel roughness, which can be affected by a flood event
- soil properties (vary in space, not time)

When dealing with hydrologic phenomena such as flood events, natural variability tends to be things that vary from flood event to flood event.



Knowledge Uncertainty

Epistemic Uncertainty

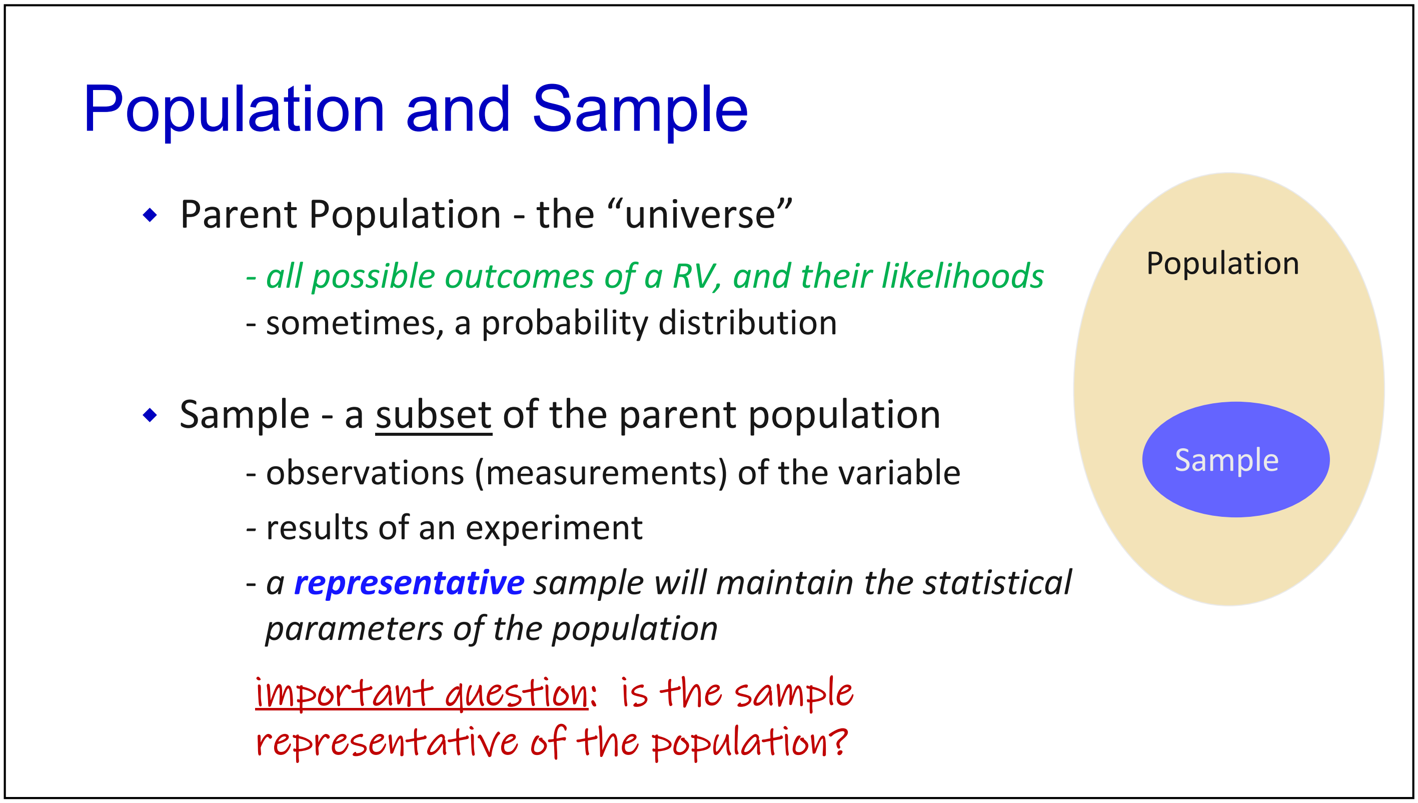
The degree to which we are unsure about any parameters and relationships used in computation: economic, hydraulic or statistical

Might be reducible with more information…

Examples

- estimates of probability
- value of structures in flood plain
- channel roughness, which we don't know precisely
- a model that is too simple

When dealing with hydrologic phenomena such as flood events, knowledge uncertainties tend to be things that are true across all flood events, but just aren't known or able to be estimated well.



Population and Sample

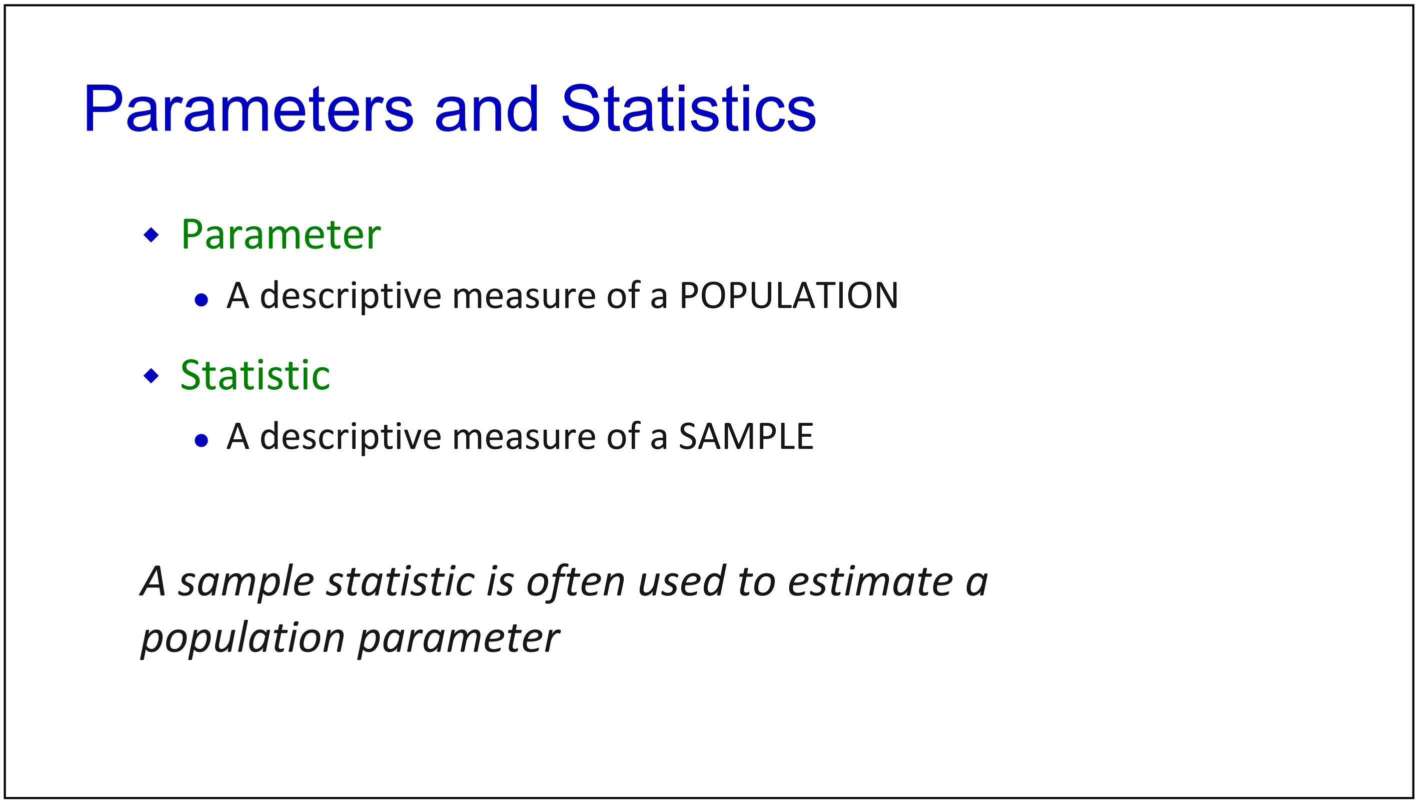
- Parent Population - the “universe”
  - *- all possible outcomes of a RV, and their likelihoods*
  - - sometimes, a probability distribution
- Sample - a <u>subset</u> of the parent population
  - - observations (measurements) of the variable
  - - results of an experiment
  - - *a **representative** sample will maintain the statistical parameters of the population*

<u>important question</u>: is the sample representative of the population?

Population

Sample

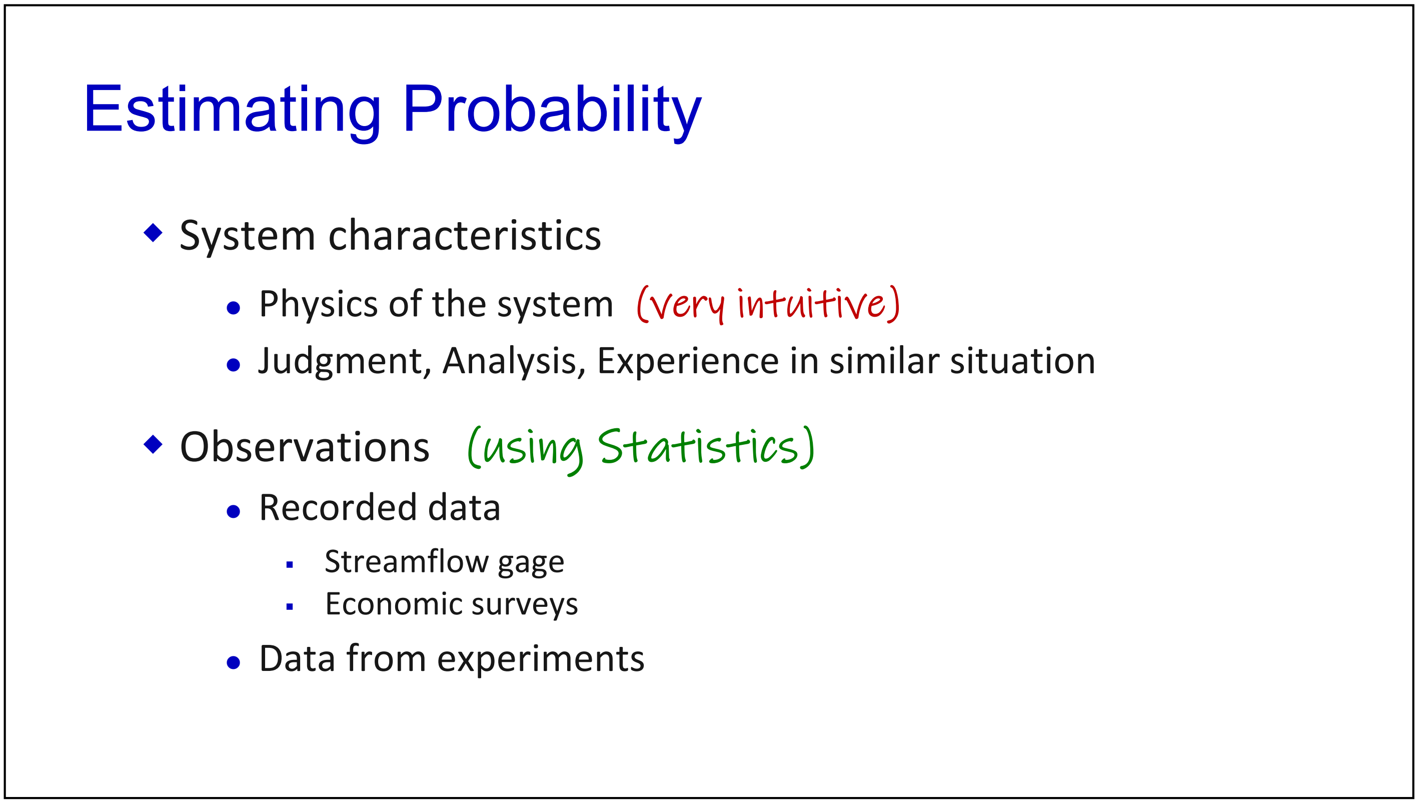
As an example, for flood frequency the population is any flood that could occur in the watershed, and how likely each magnitude is. The sample is what’s been observed.



Parameters and Statistics

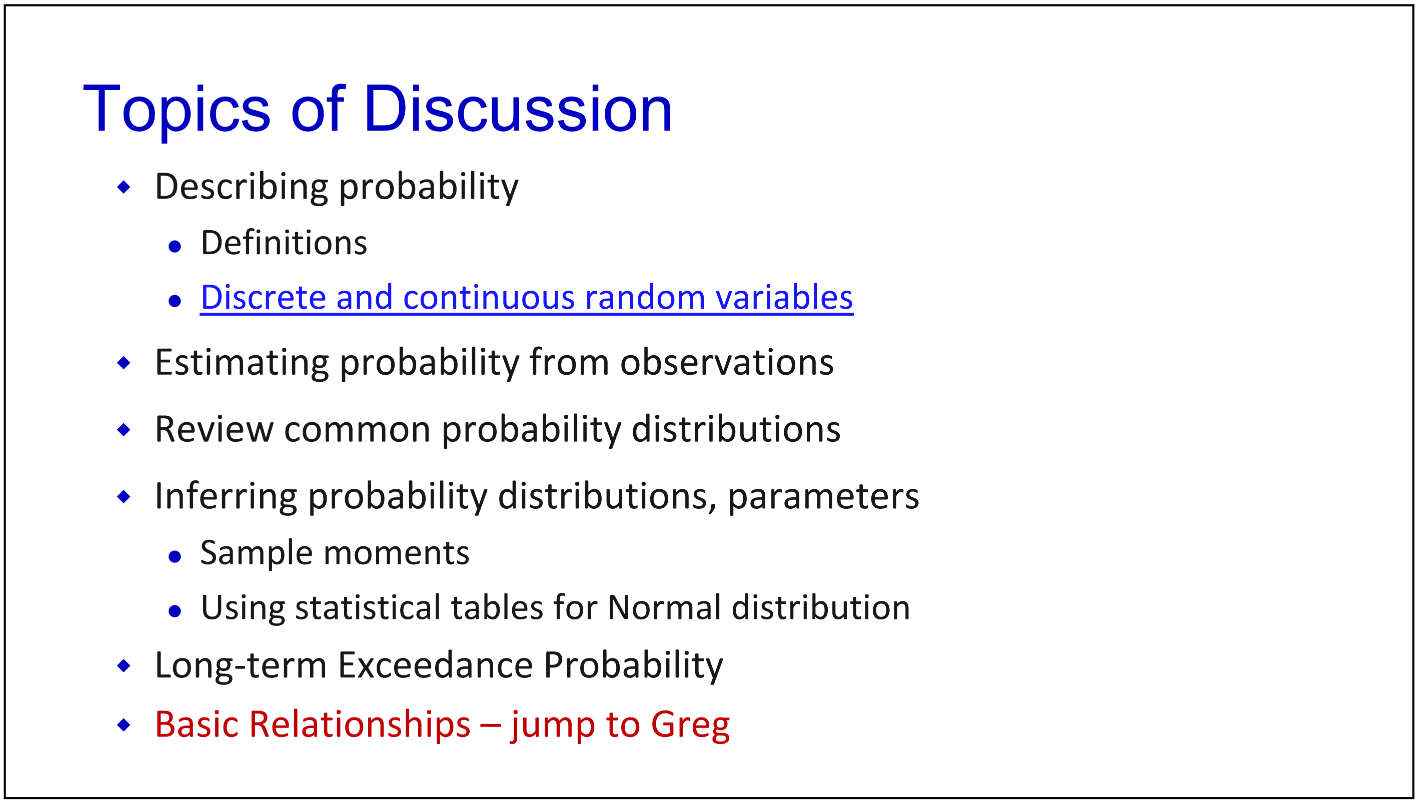
- Parameter
  - A descriptive measure of a POPULATION
- Statistic
  - A descriptive measure of a SAMPLE

A sample statistic is often used to estimate a population parameter



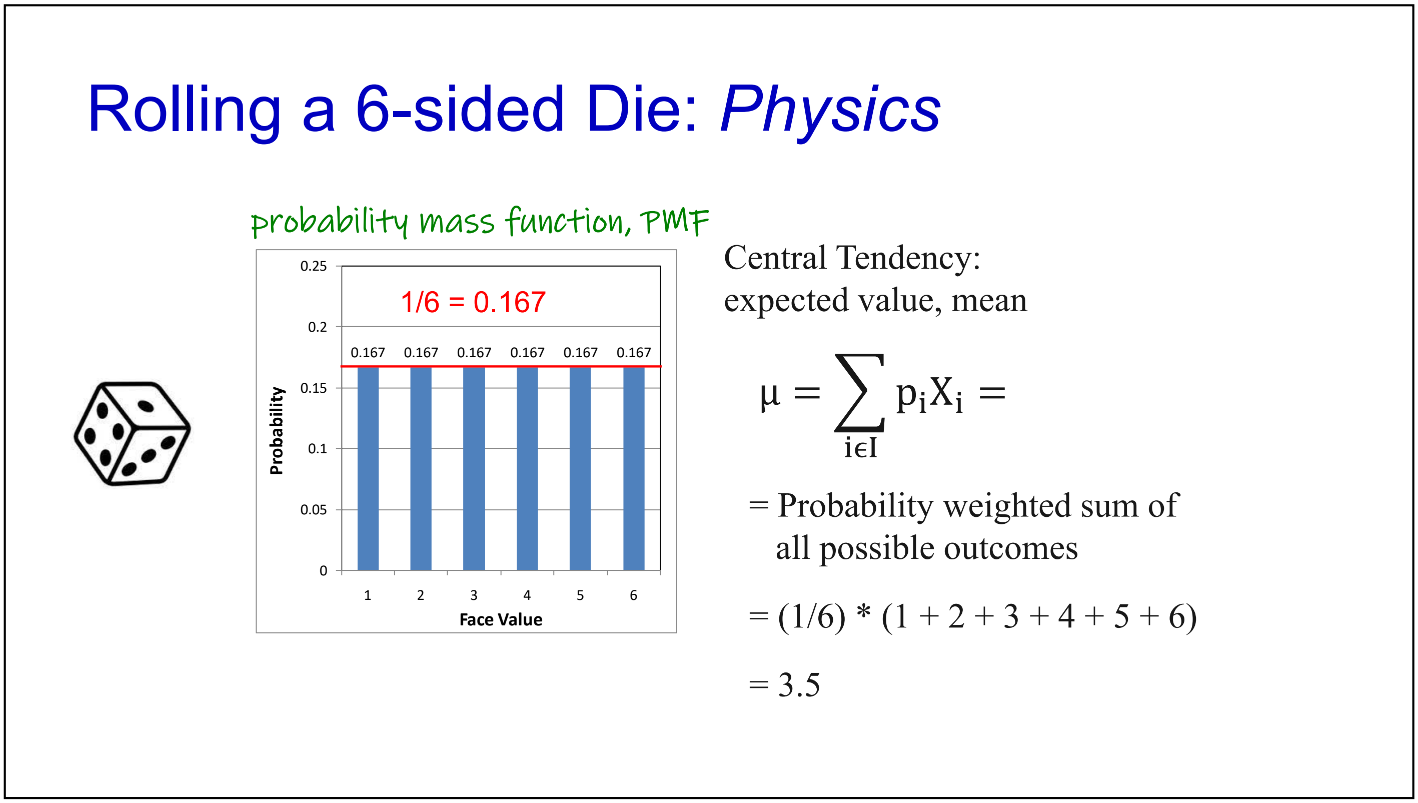
Estimating Probability

- System characteristics
  - Physics of the system (very intuitive)
  - Judgment, Analysis, Experience in similar situation
- Observations (using Statistics)
  - Recorded data
    - Streamflow gage
    - Economic surveys
  - Data from experiments



Topics of Discussion

- Describing probability
  - Definitions
  - Discrete and continuous random variables
- Estimating probability from observations
- Review common probability distributions
- Inferring probability distributions, parameters
  - Sample moments
  - Using statistical tables for Normal distribution
- Long-term Exceedance Probability
- Basic Relationships – jump to Greg



Rolling a 6-sided Die: *Physics*

probability mass function, PMF

Central Tendency:
expected value, mean

$$\mu = \sum_{i \epsilon I} p_i X_i =$$

= Probability weighted sum of all possible outcomes

= (1/6) * (1 + 2 + 3 + 4 + 5 + 6)

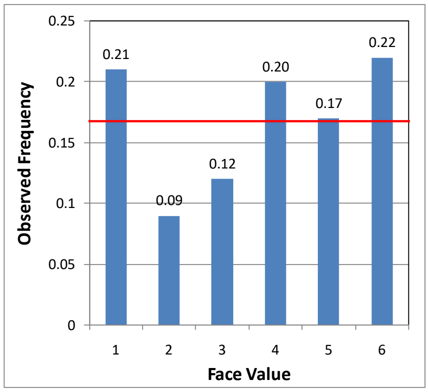
= 3.5

This is the estimate of probability that results from recognizing the simple system of a 6-sided die suggests that each side is equally likely. Therefore, the total probability of all outcomes of 1.0 divided by the 6 possible outcomes provides probability 1/6 for each outcome.

The mean of the discrete distribution is the probability weighted sum of all possible outcomes, and is not necessarily a possible outcome.

Rolling a 6-sided Die: *Experiment*

histogram - estimate of PMF



Sample size = 100

Central Tendency:
expected value, mean, average

$$\hat{\mu} = \sum_{i\epsilon I} \hat{p}_i X_i = \overline{X} =$$

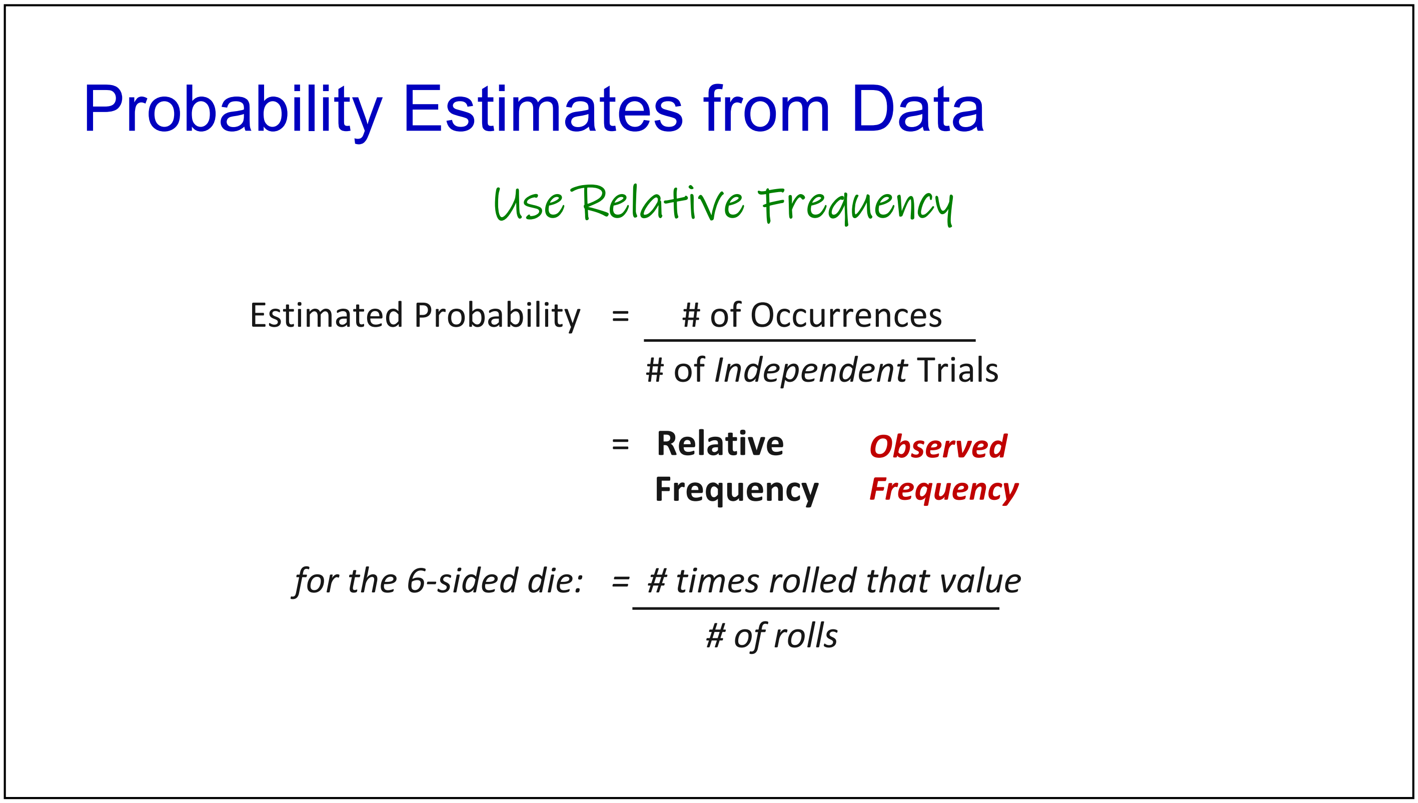
sample estimate of mean

= Probability weighted sum of all outcomes

= (0.21*1) + (0.09*2) + (0.12*3) + (0.20*4) + (0.17*5) + (0.22*6)

= 3.67

This is the same example of the 6-sided die, with probability estimated by experimentation. This method is only possible with systems that can be repeated replicated at will, to produce a sample. The relative frequency of each possible outcome is used to estimate the probability of each outcome, with the resulting histogram becoming the estimate of the PMF. The sample estimate of the mean is either the probability weighted sum computed with estimated probabilities, or simply the average of all outcomes.



Probability Estimates from Data

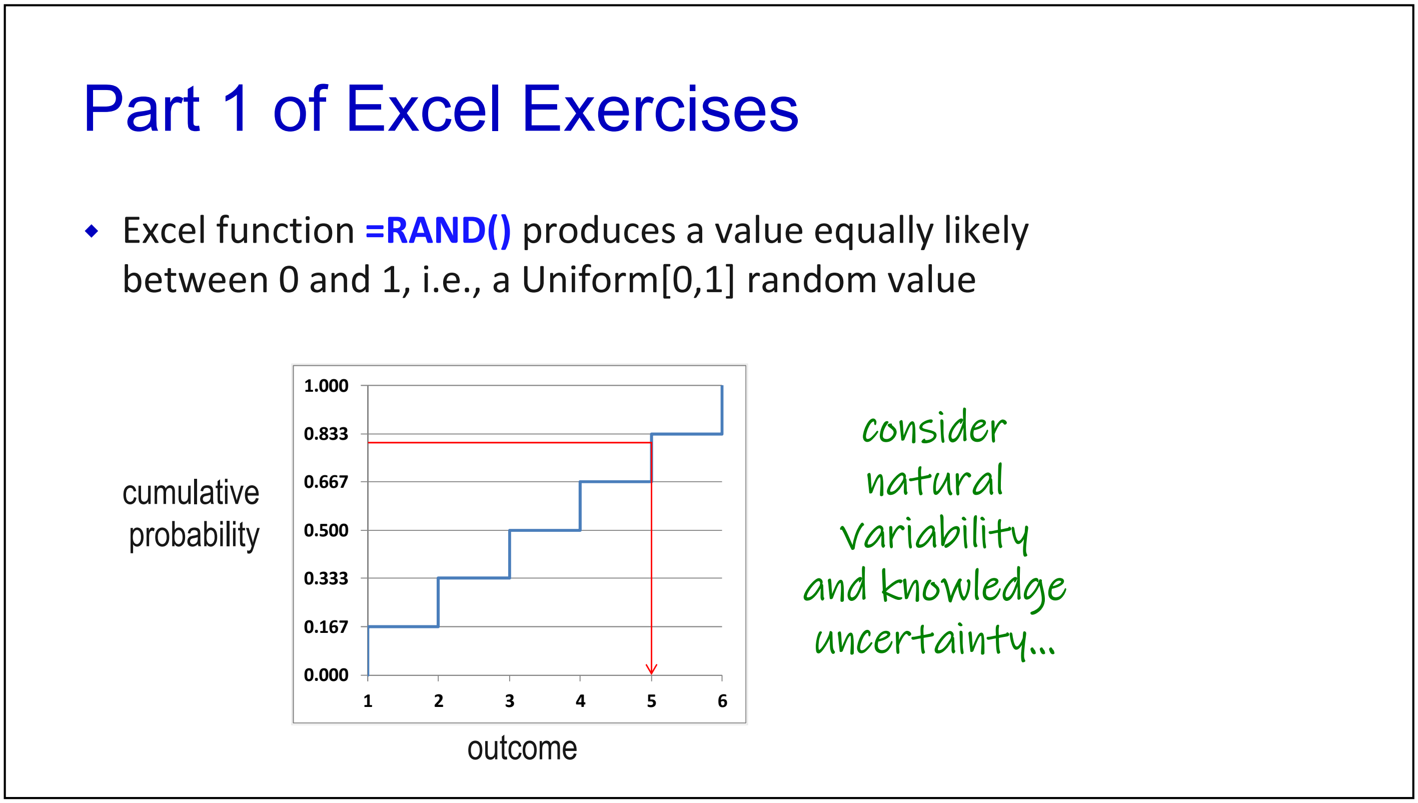
Use Relative Frequency

$$\text{Estimated Probability} = \frac{\text{\# of Occurrences}}{\text{\# of } \textit{Independent} \text{ Trials}}$$

= **Relative Frequency** ***Observed Frequency***

$$\textit{for the 6-sided die:} = \frac{\textit{\# times rolled that value}}{\textit{\# of rolls}}$$

Relative frequency is a very simple but very effective means of estimating probability from a random sample.



Part 1 of Excel Exercises

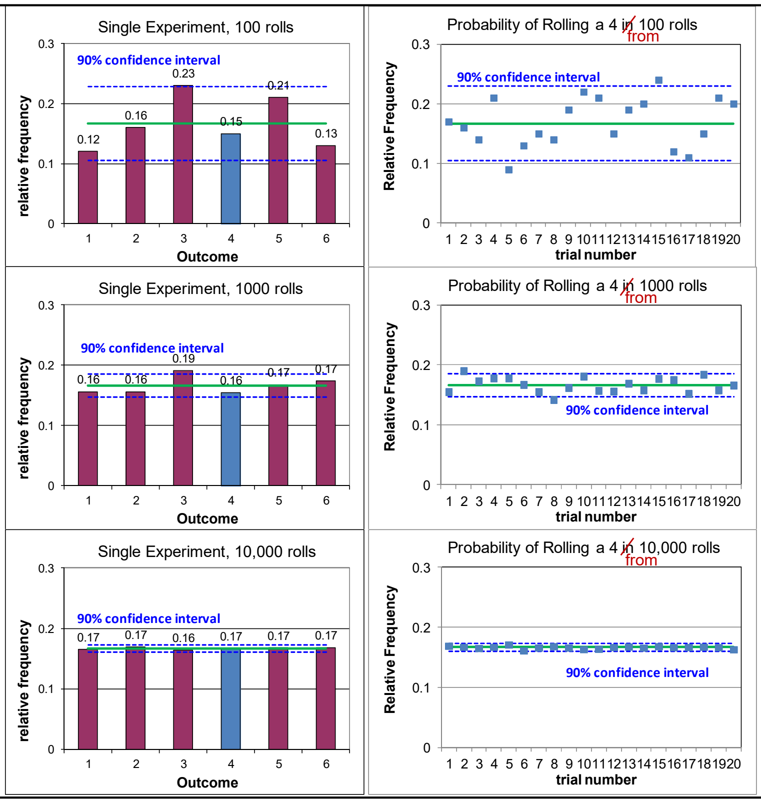
- Excel function **=RAND()** produces a value equally likely between 0 and 1, i.e., a Uniform[0,1] random value

consider natural variability and knowledge uncertainty…

After repeatedly rolling a real 6-sided die to estimate probability, we’ll be using Excel to simulate the rolling of a 6-sided die using pseudorandom numbers.

Dice Experiment Confidence Intervals

can compute by experiment, or approximate with a Normal distribution



These are images from the Excel exercises. A reason for repeating an experiment in which we estimate probability from a limited sample, rather than just pooling all the samples for a better estimate, is to explore the uncertainty in an estimate from a limited sample.

These figures show single-sample histograms on the left for increasing sample size, and 20 replicates of the experiment on the right. From repeating the experiment, we see how much error is in our estimates from limited samples and can draw 90% confidence intervals. The interval shows the range in which 90% of random sample estimates of probability will fall, but can also be drawn around a single estimate as an interval for the "population" probability.

Rolling 2 Dice: Physics & Experiment

36 possible rolls resulting from **2** random numbers, but here we're interested in the **SUM** of both dice.

only 11 possible outcomes

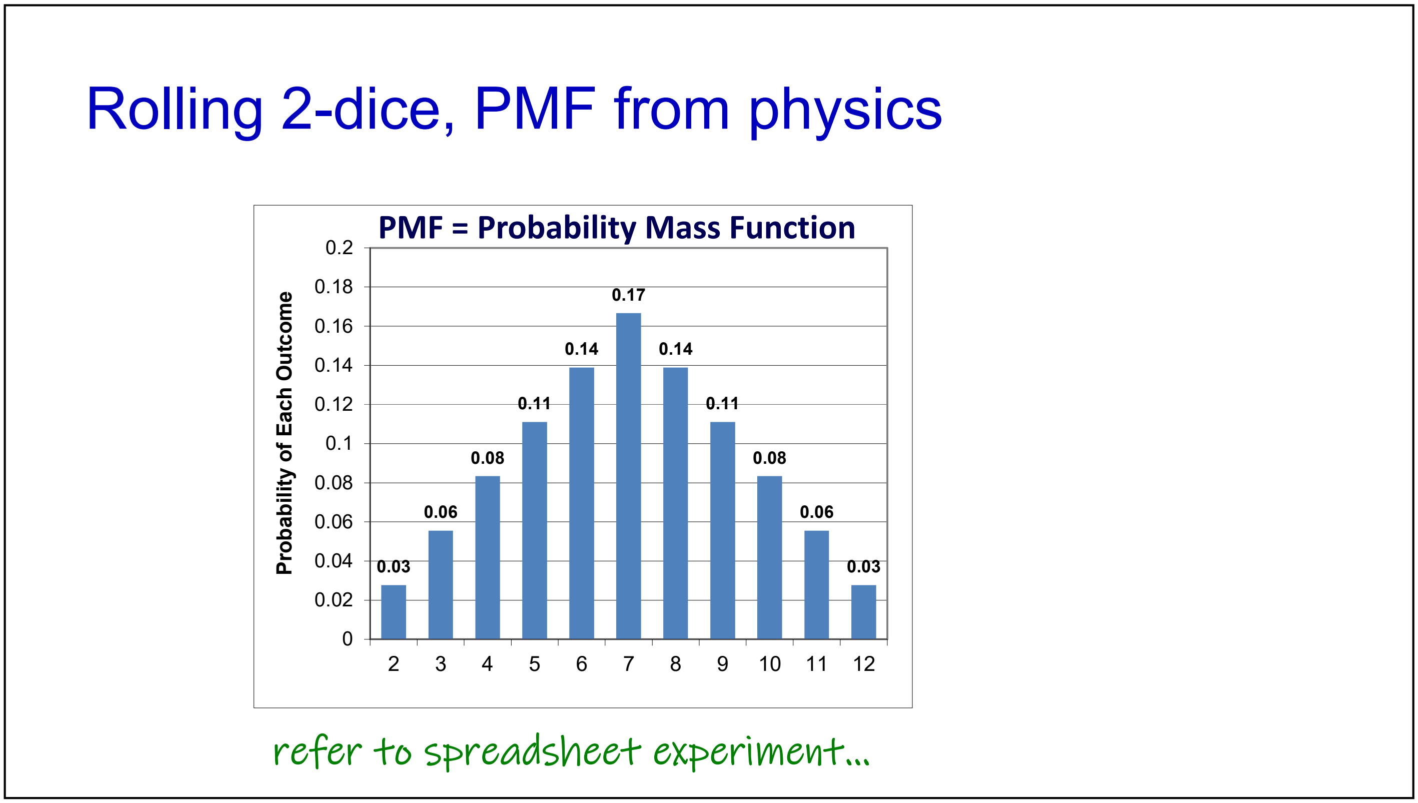
		Die #2					
		1	2	3	4	5	6
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
Die #1	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

these outcomes are
mutually exclusive and exhaustive

Expected Value of the Sum =

$$2*{}^{1}/_{36} + 3*{}^{2}/_{36} + 4*{}^{3}/_{36} +$$
$$5*{}^{4}/_{36} + 6*{}^{5}/_{36} + 7*{}^{6}/_{36} +$$
$$8*{}^{5}/_{36} + 9*{}^{4}/_{36} + 10*{}^{3}/_{36} +$$
$$11*{}^{2}/_{36} + 12*{}^{1}/_{36} = 7$$

This is a similar system using two dice, and computing the sum. The system is simple enough to estimate probability from the basic understanding. There are 36 possible rolls, as each value on die 1 can be paired with each value on die 2. There are 11 possible sums, between 2 and 12. The likelihood of each possible outcome is simply the number of ways to produce that outcome, divided by the 36 possible outcomes, with 7 the most likely as 6 out of 36.

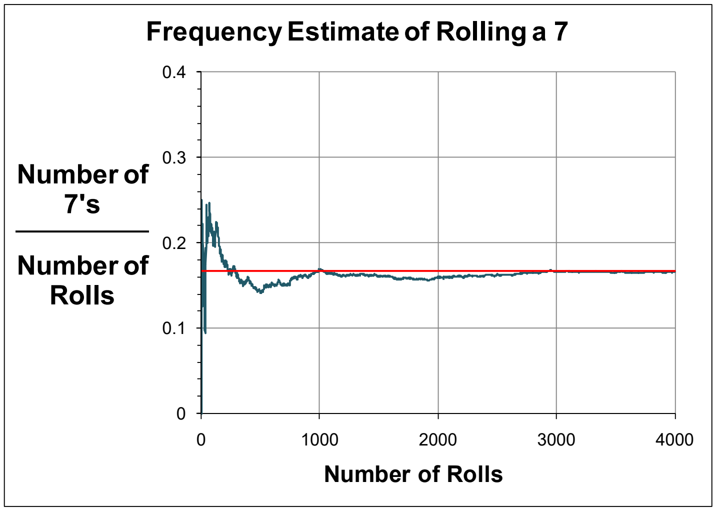


Rolling 2-dice, PMF from physics

refer to spreadsheet experiment...

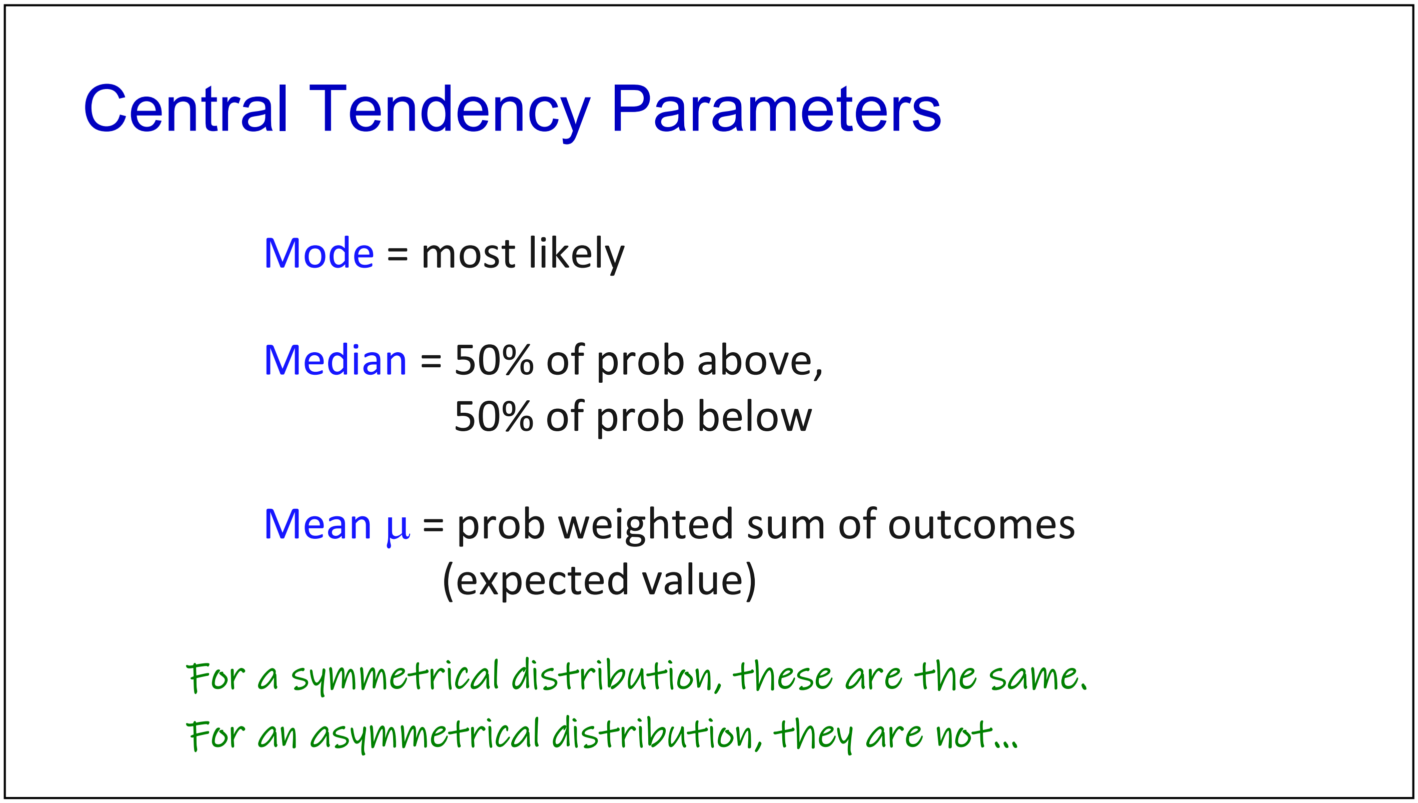
This is the PMF of the sum of two dice, estimated by understanding the simple system. It can also be estimated by repeated random sampling, like the just-completed workshop, as shown in a spreadsheet.

Estimated Probability of Rolling a 7



The ERROR in our estimate of probability from a limited sample decreases as sample size increases

This figure shows how, in using repeated random sampling to estimate probability, the estimate gets closer to the true value as the number of rolls increases.



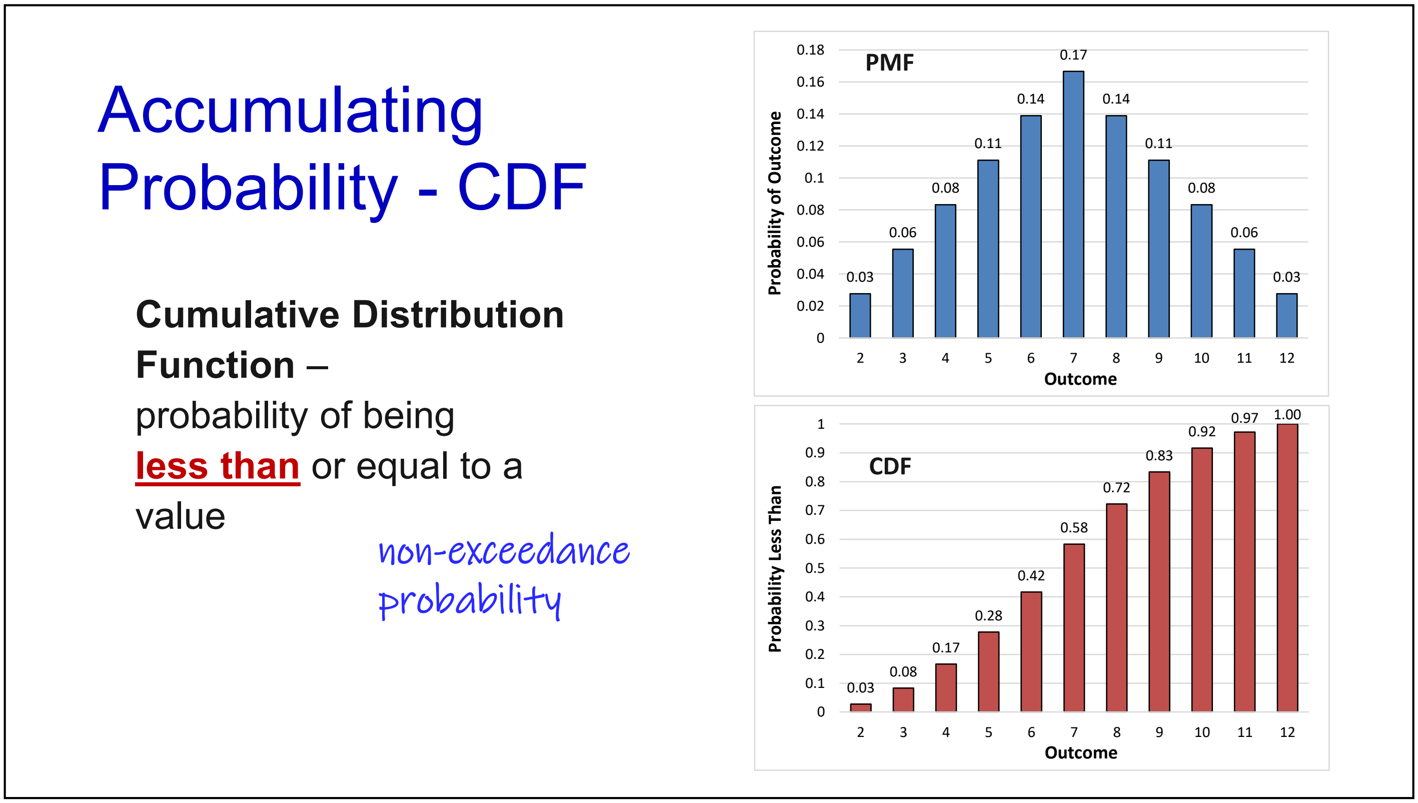
Central Tendency Parameters

Mode = most likely

Median = 50% of prob above,
50% of prob below

Mean μ = prob weighted sum of outcomes
(expected value)

For a symmetrical distribution, these are the same.
For an asymmetrical distribution, they are not...



Accumulating Probability - CDF

Cumulative Distribution Function –
probability of being
less than or equal to a
value

non-exceedance
probability

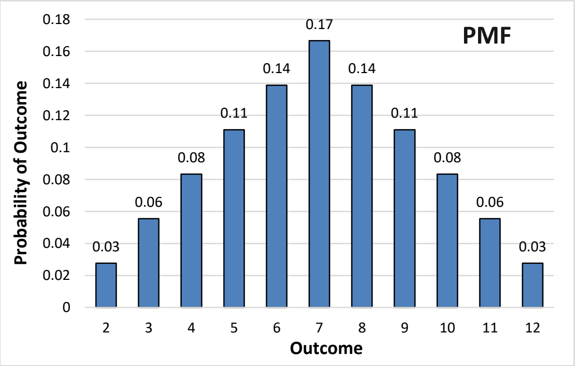
It is sometimes more useful to work with cumulative probabilities. The Probability Mass Function (PMF) displays the probability of a discrete outcome occurring. The PMF can be accumulated from the bottom into a Cumulative Distribution Function (CDF) that displays the probability that an occurrence is less than or equal to than each discrete outcome. The probability of an occurrence less than or each to 5 is the sum of the probabilities of 2, 3, 4 and 5. The sum of all possible outcomes is 1.0, and so the CDF spans the range between 0 and 1.

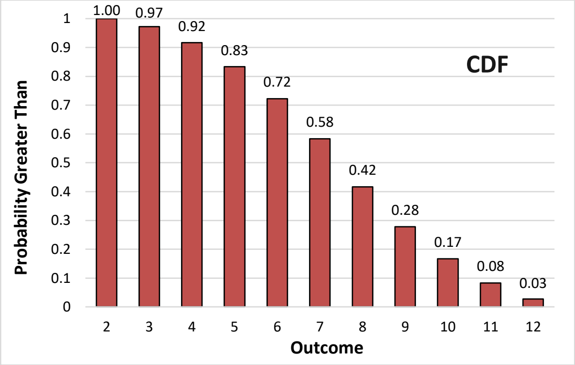
Accumulating Probability - CDF

Cumulative Distribution Function –
probability of being **greater than** or equal to a value

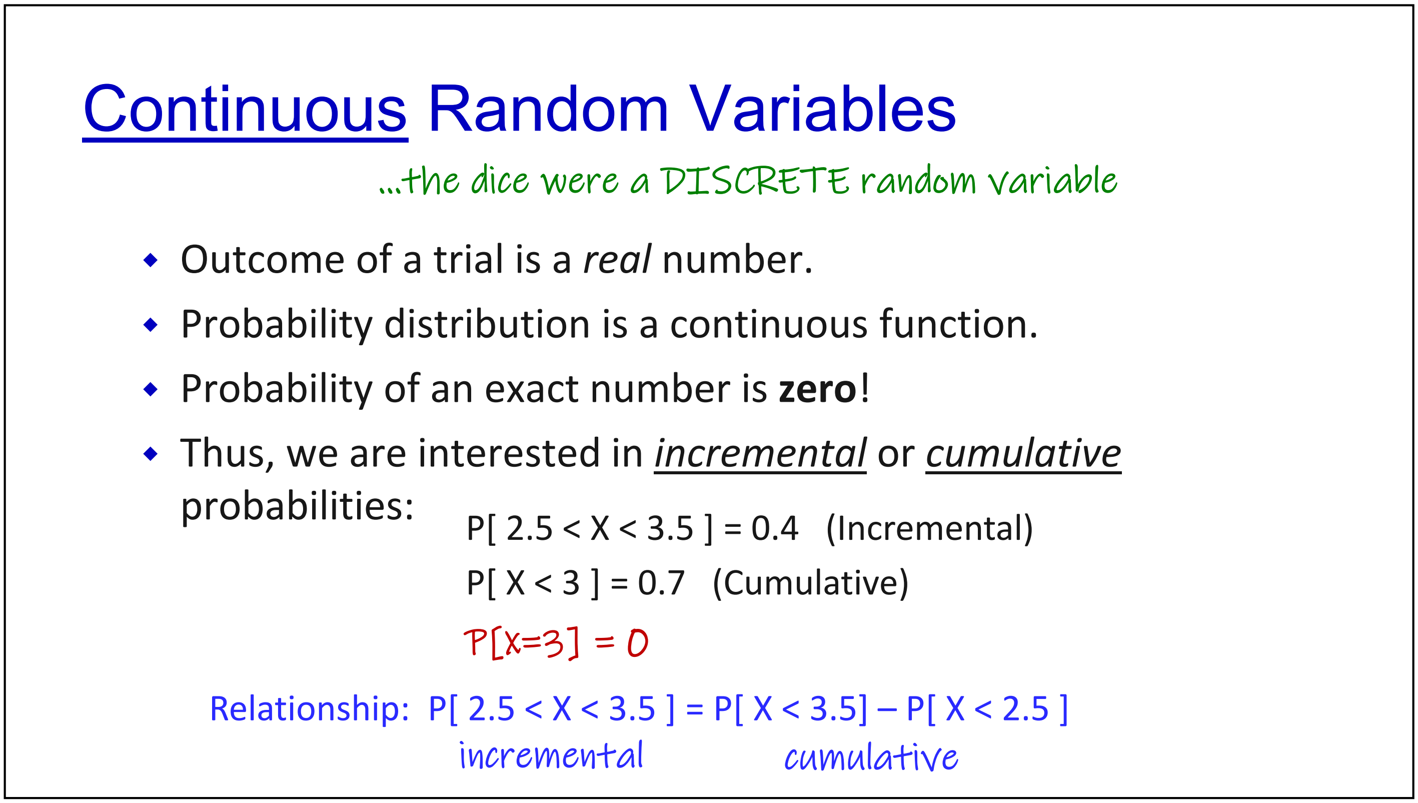
exceedance probability

Can be called Complementary CDF





Sometimes the probability of being greater than or equal to some value is more interesting than less than. So we can accumulate the PMF from the top to compute the probability of an occurrence greater than or equal to each possible outcome. The produce is still a Cumulative Distribution Function, but is sometimes referred to as a complementary CDF.



Continuous Random Variables

...the dice were a DISCRETE random variable

- Outcome of a trial is a *real* number.
- Probability distribution is a continuous function.
- Probability of an exact number is **zero**!
- Thus, we are interested in *incremental* or *cumulative* probabilities:

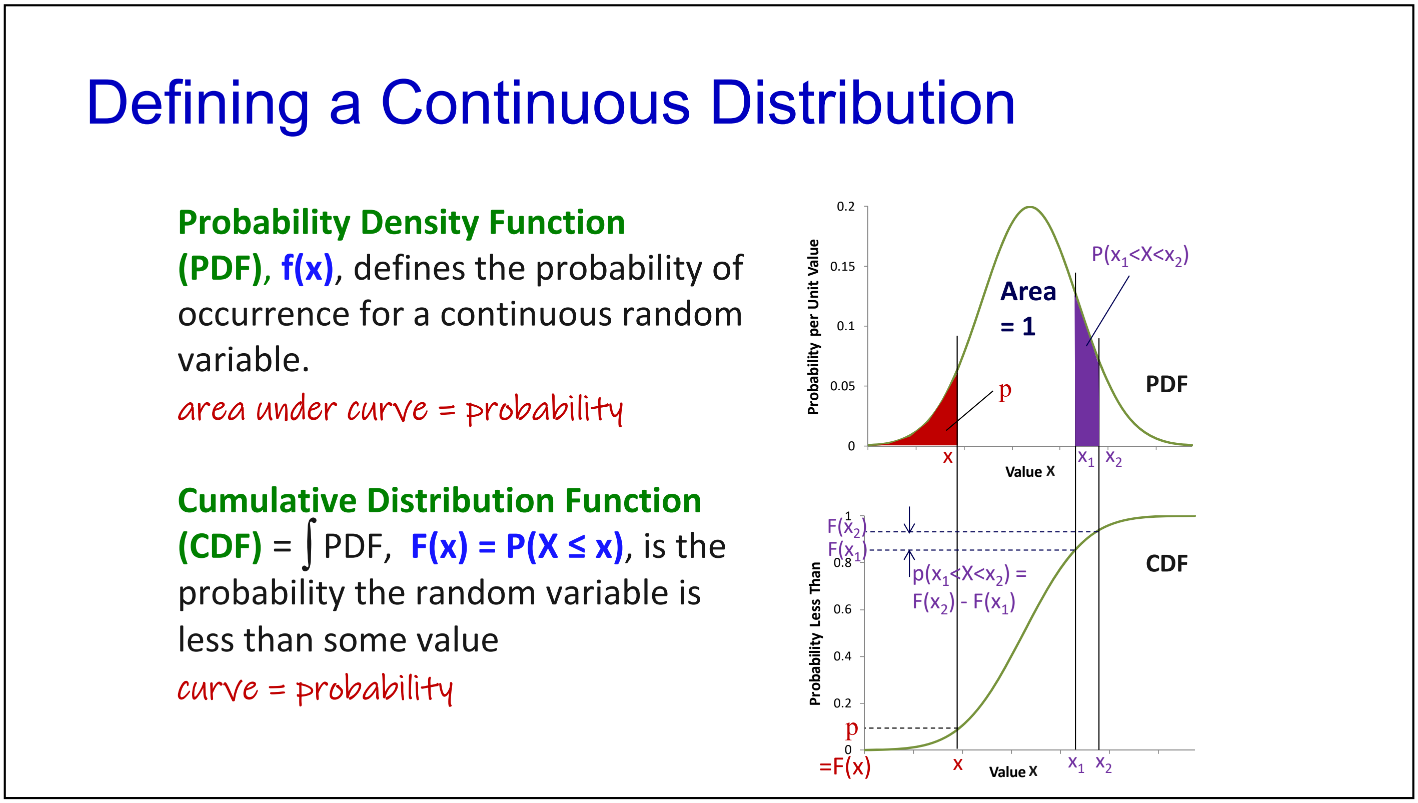
$P[\ 2.5 < X < 3.5\ ] = 0.4$ (Incremental)

$P[\ X < 3\ ] = 0.7$ (Cumulative)

$P[X=3] = 0$

Relationship: $\underbrace{P[\ 2.5 < X < 3.5\ ]}_{\text{incremental}} = \underbrace{P[\ X < 3.5] - P[\ X < 2.5\ ]}_{\text{cumulative}}$

Continuous random variables have an infinite number of possible outcomes, making the probability of any exact outcome equal to zero. Therefore, we work with either cumulative or incremental probabilities with continuous random variables.



Defining a Continuous Distribution

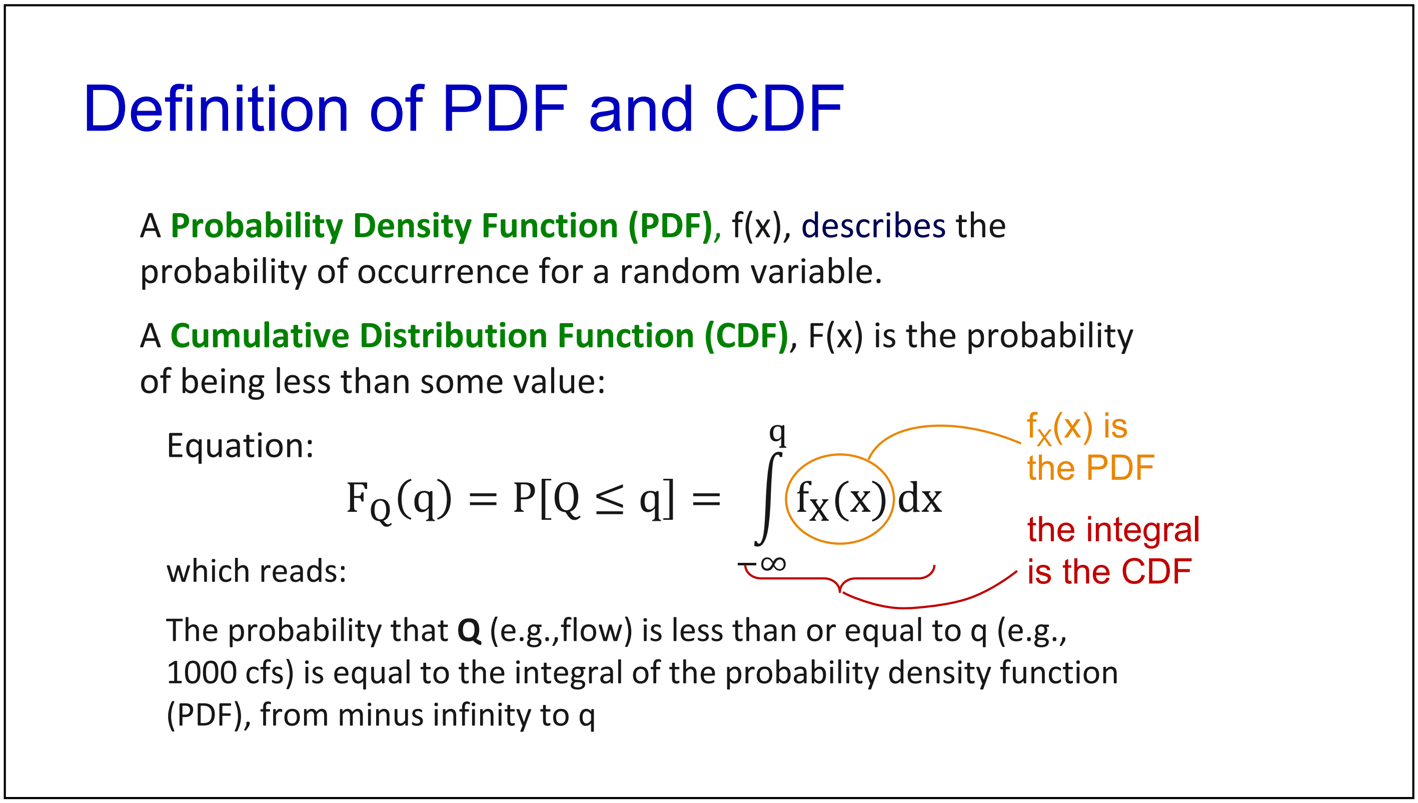
Probability Density Function (PDF), **f(x)**, defines the probability of occurrence for a continuous random variable.

area under curve = probability

Cumulative Distribution Function (CDF) = ∫ PDF, **$F(x) = P(X \le x)$**, is the probability the random variable is less than some value

curve = probability

These are the basic images and definitions of the Probability Density Function and the Cumulative Distribution function for a continuous random variable.



Definition of PDF and CDF

A **Probability Density Function (PDF),** f(x), describes the probability of occurrence for a random variable.

A **Cumulative Distribution Function (CDF)**, F(x) is the probability of being less than some value:

Equation:

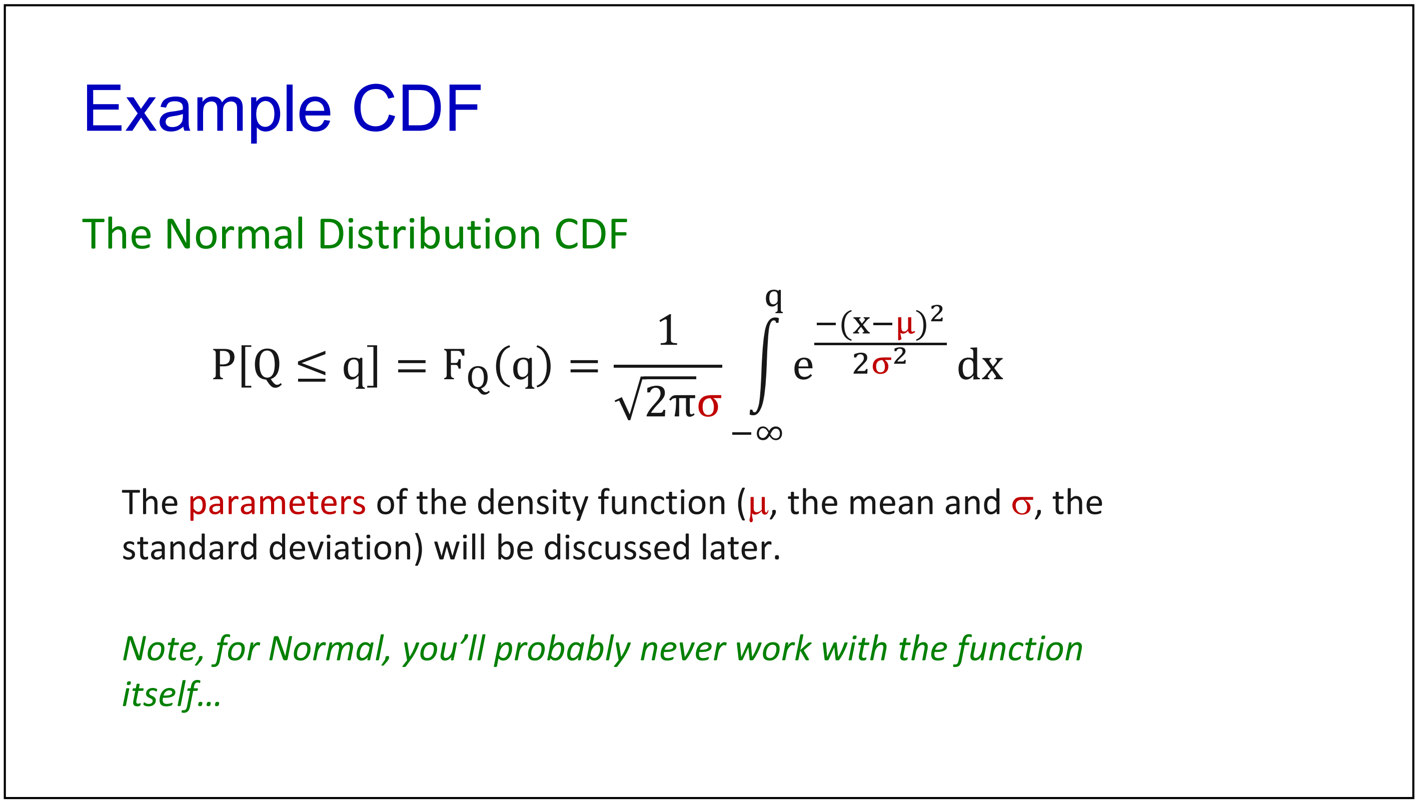
$$F_Q(q) = P[Q \le q] = \underbrace{\int_{-\infty}^{q} \boxed{f_X(x)}\, dx}_{\text{the integral is the CDF}}$$

$f_X(x)$ is the PDF

which reads:

The probability that **Q** (e.g.,flow) is less than or equal to q (e.g., 1000 cfs) is equal to the integral of the probability density function (PDF), from minus infinity to q

The CDF is the integral of the PDF, and a probability for a given value is the integral up to that value.



Example CDF

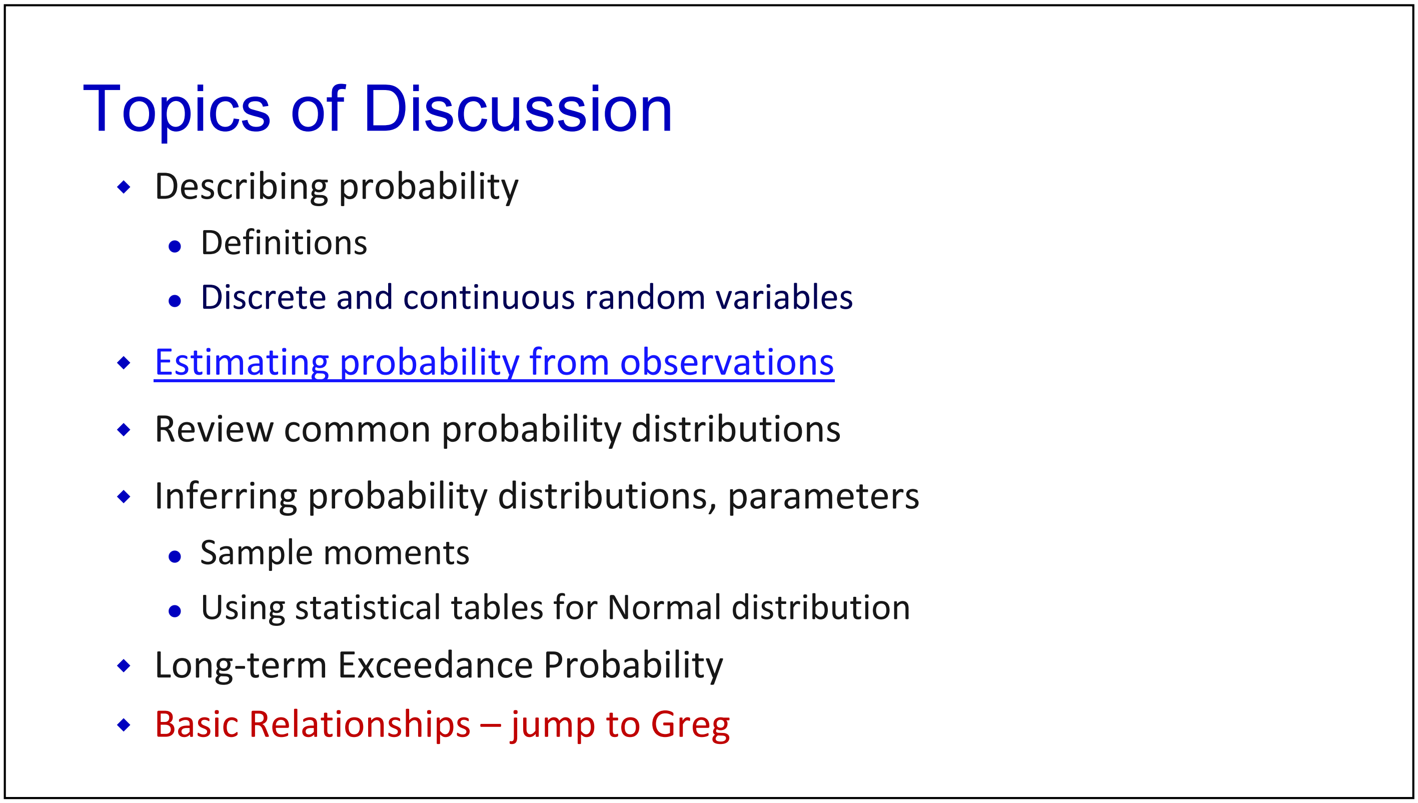
The Normal Distribution CDF

$$P[Q \le q] = F_Q(q) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{q} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \, dx$$

The parameters of the density function (μ, the mean and σ, the standard deviation) will be discussed later.

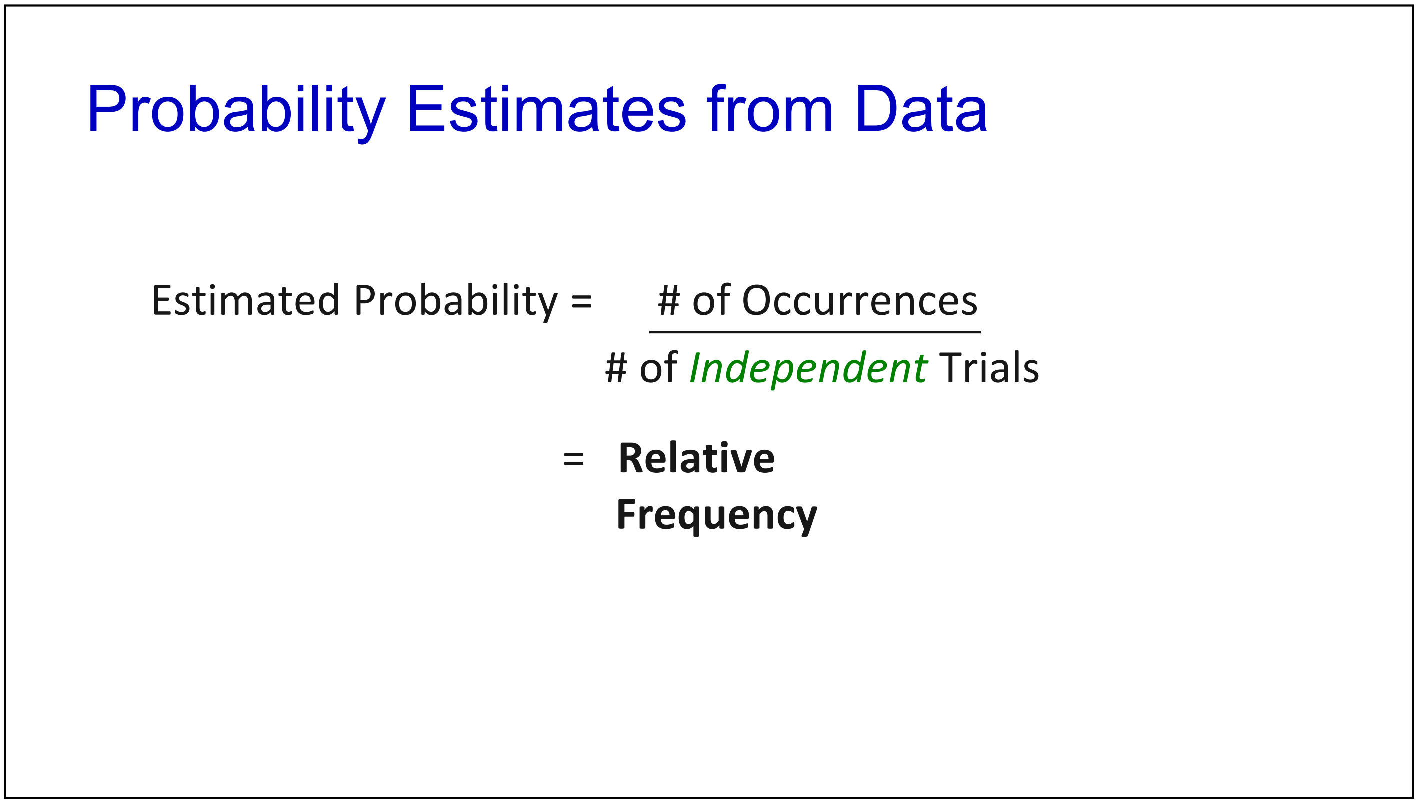
Note, for Normal, you'll probably never work with the function itself…

The equations for the Normal distribution are awkward. We typically use a tabulation, or an approximation.



Topics of Discussion

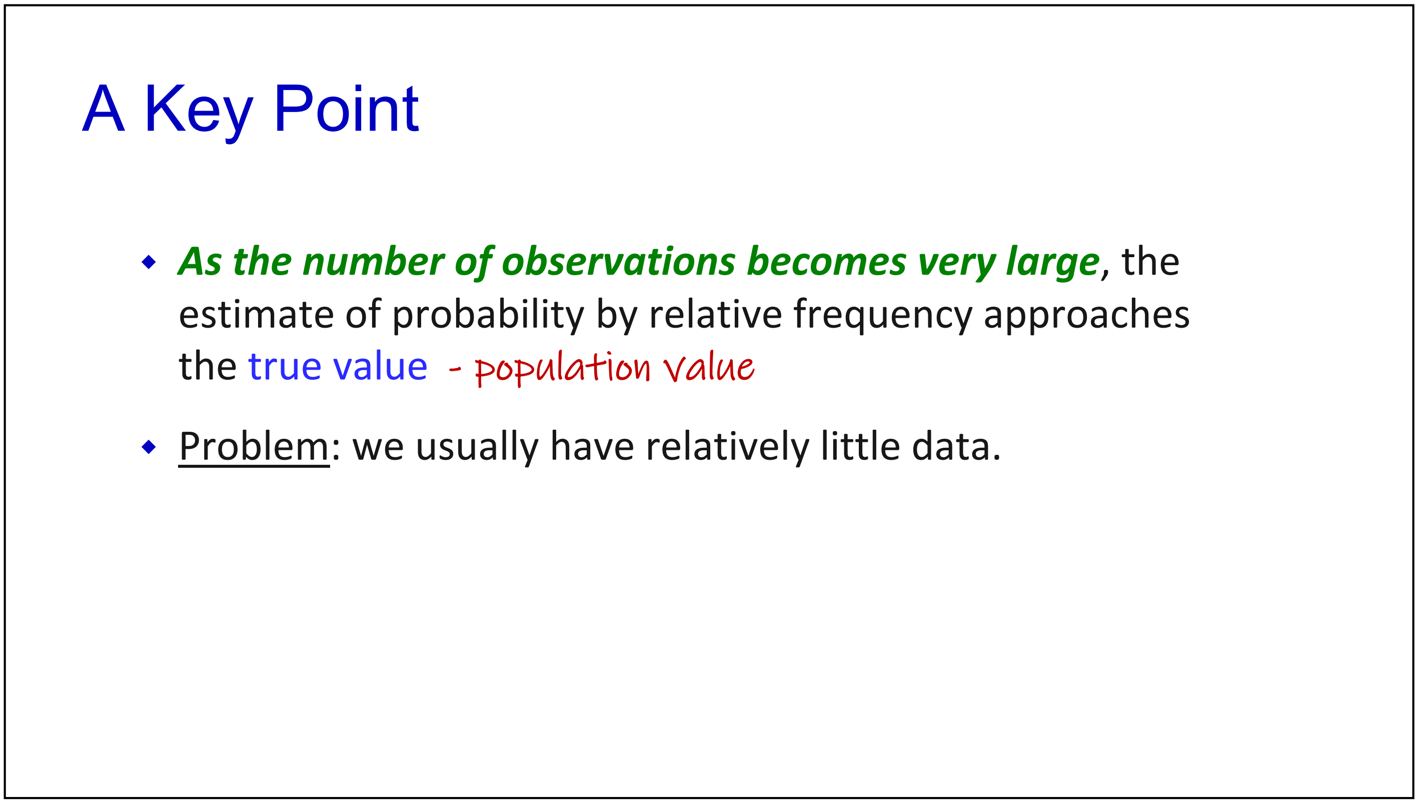
- Describing probability
  - Definitions
  - Discrete and continuous random variables
- Estimating probability from observations
- Review common probability distributions
- Inferring probability distributions, parameters
  - Sample moments
  - Using statistical tables for Normal distribution
- Long-term Exceedance Probability
- Basic Relationships – jump to Greg



Probability Estimates from Data

$$\text{Estimated Probability} = \frac{\text{\# of Occurrences}}{\text{\# of \textit{Independent} Trials}} = \textbf{Relative Frequency}$$

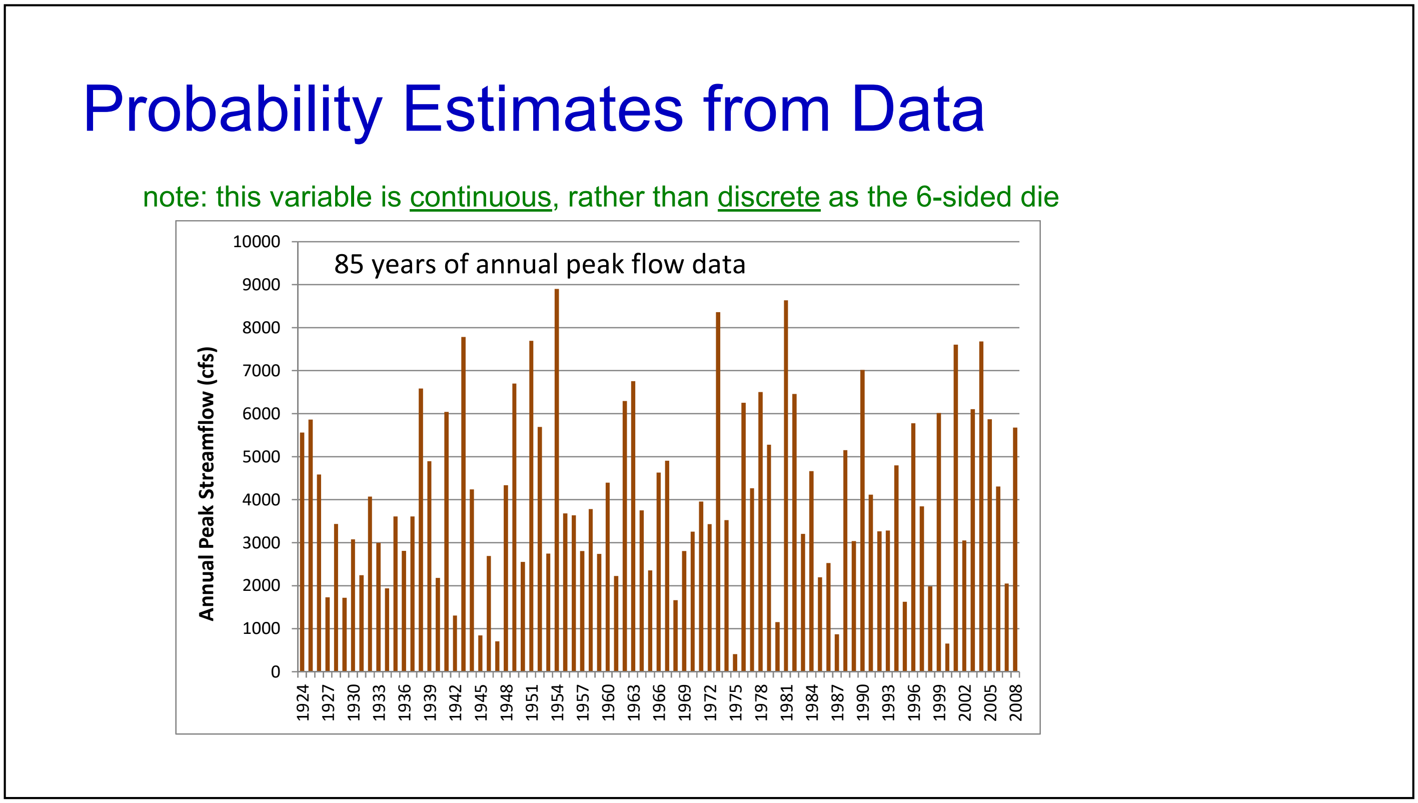
Relative Frequency is a very simple but very effective method of estimating probability from observations. Outcomes that are more probable to occur more frequently.



A Key Point

- ***As the number of observations becomes very large***, the estimate of probability by relative frequency approaches the true value - population value
- <u>Problem</u>: we usually have relatively little data.

As an estimator, relative frequency improves as the number of observations increases. This is true of most of the estimators we will use this week. Unfortunately, except with systems that can be made to produce an outcome at will, natural systems provide us a limited number of observations that only increases with time.

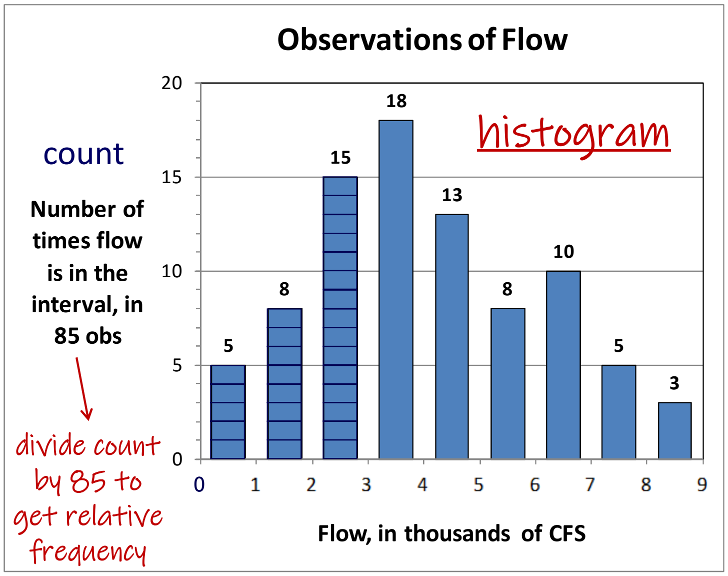


Probability Estimates from Data

note: this variable is continuous, rather than discrete as the 6-sided die

This is an example data set of the annual maximum flow value of each year for an 85 year period of record. Note this is a continuous random variable, because flow can be any value in the range of possibility.

Probability Estimates from Data



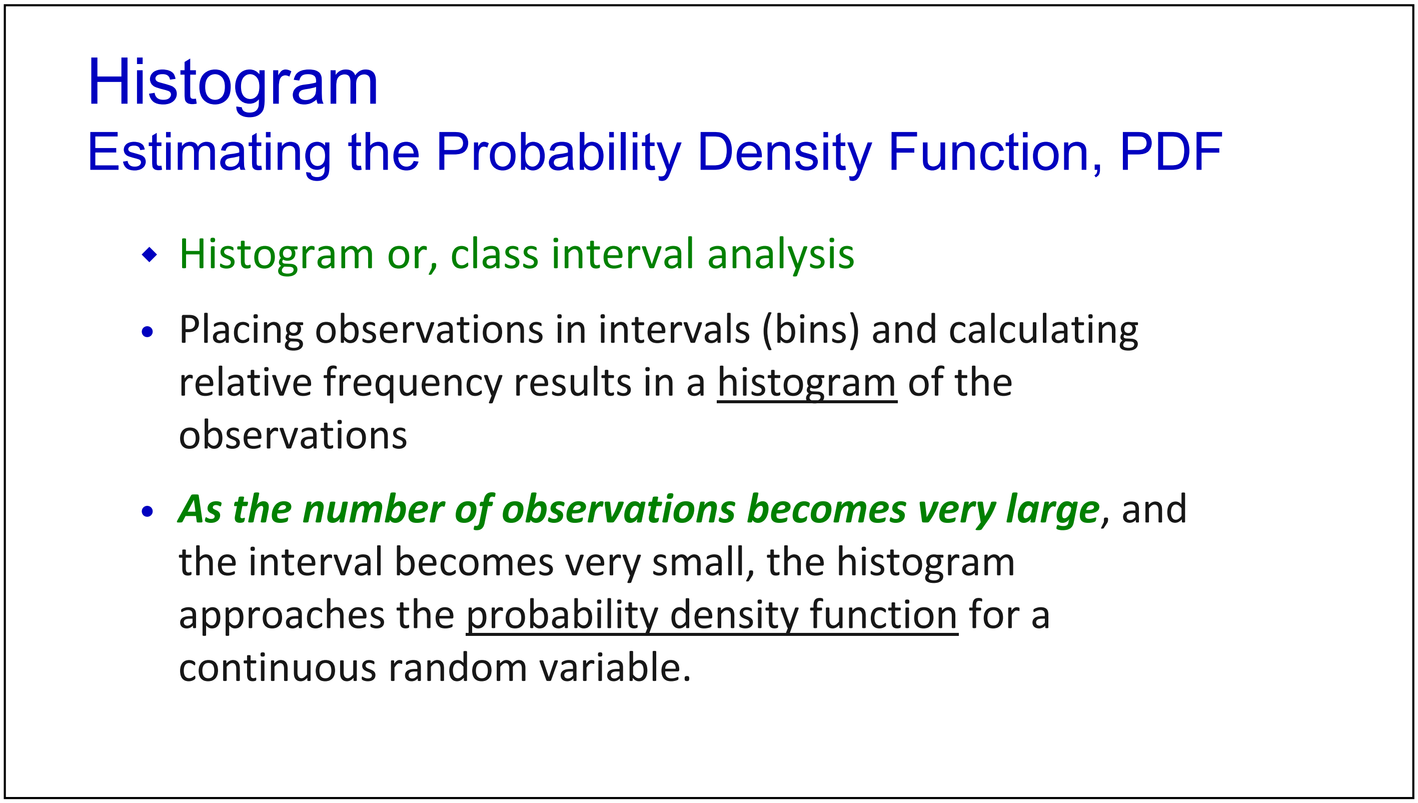
count

histogram

divide count by 85 to get relative frequency

note: this variable is continuous, rather than discrete as the 6-sided die

One way to view the variable probabilistically is to discretize the range and produce a histogram. We see higher bars where there are more occurrences, implying higher probability of those values. The histogram is an estimate of the PDF.



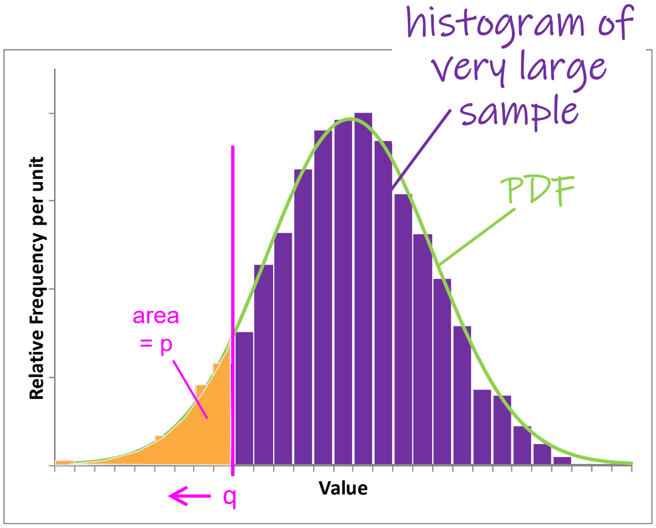
Histogram

Estimating the Probability Density Function, PDF

- Histogram or, class interval analysis
- Placing observations in intervals (bins) and calculating relative frequency results in a histogram of the observations
- ***As the number of observations becomes very large***, and the interval becomes very small, the histogram approaches the probability density function for a continuous random variable.

We'll see this phrase "as the number of observations becomes very large" a few times in the next few slides. Estimators that get better with more data are called "consistent" estimators.

Density function (PDF) vs. histogram

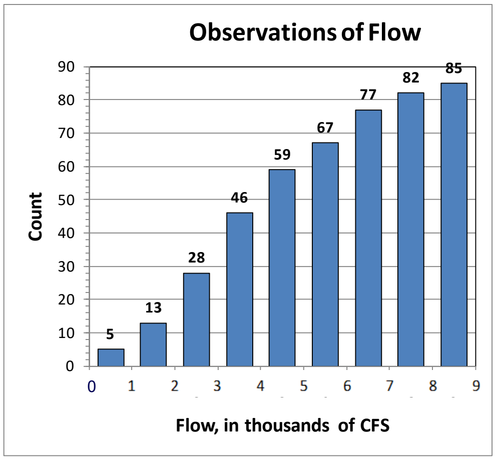


area under the probability density function is <u>probability</u>, and so the area under the *entire* curve is equal to 1.0

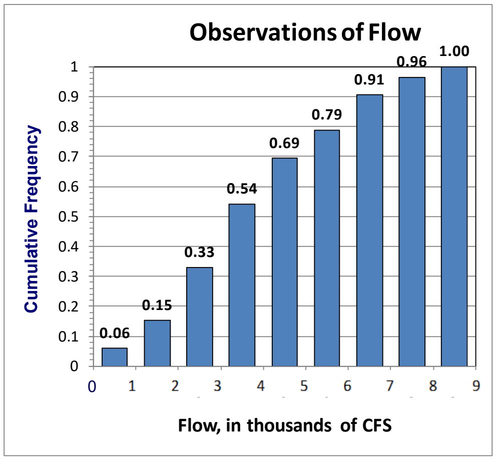
The green PDF is a parent population, and the purple histogram represents a very large sample from that population. This is intended to show that for a very large sample, the histogram approaches the PDF.

Cumulative Histogram

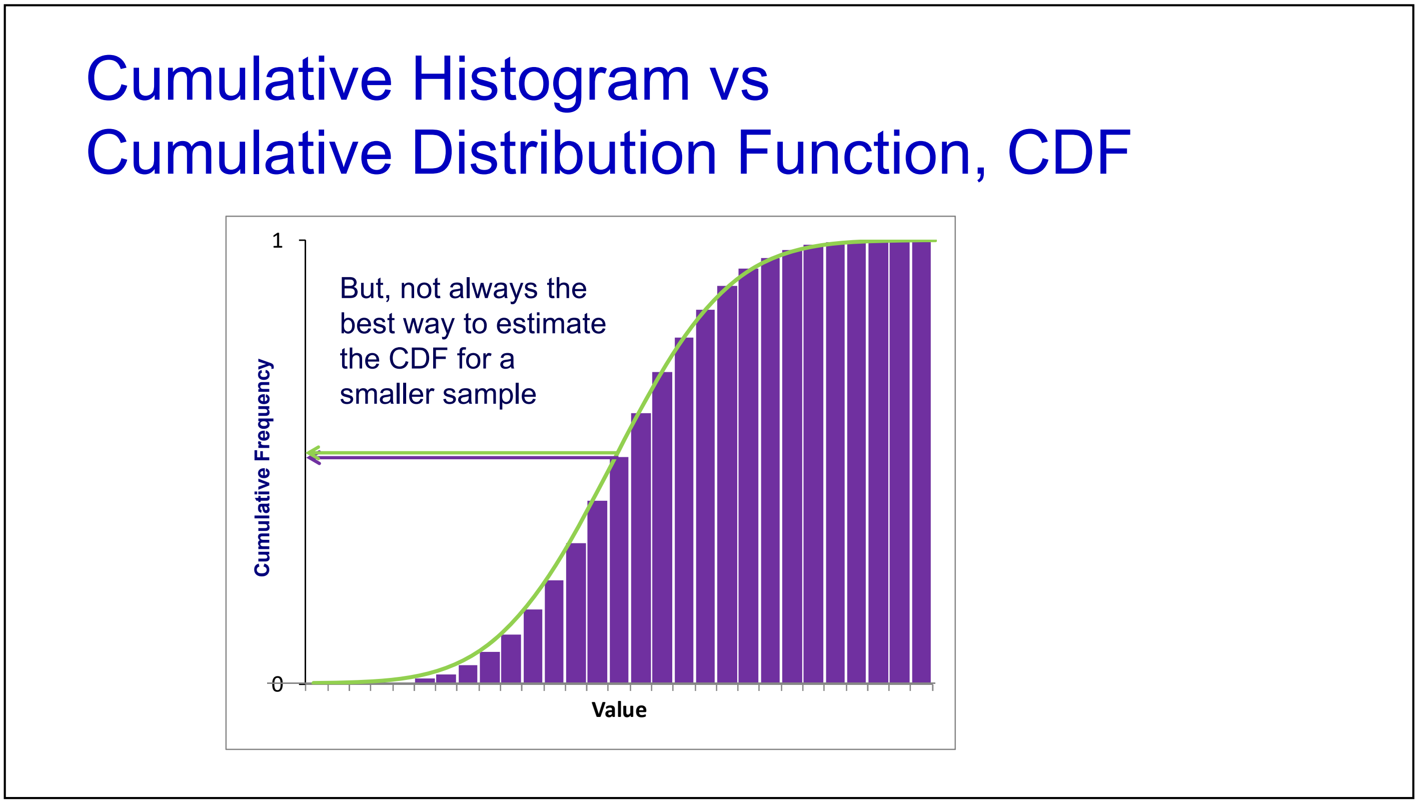
Cumulative Count



Cumulative Frequency



Histograms can be accumulated for continuous random variables in the same way they are for discrete random variables, and are an estimate of the CDF. This is the cumulative histogram of the sample of 85 values, as count on the left and as relative frequency on the right.



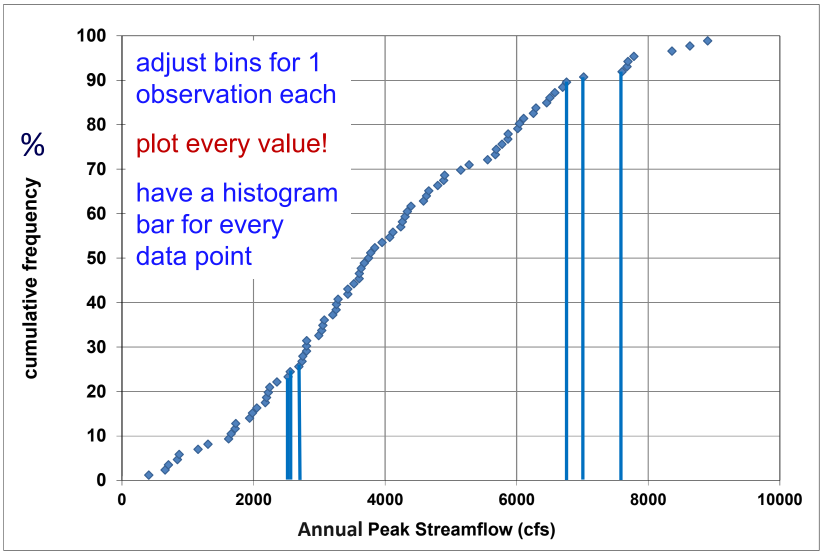
Cumulative Histogram vs Cumulative Distribution Function, CDF

But, not always the best way to estimate the CDF for a smaller sample

Similarly, the green CDF is a parent population, and the purple cumulative histogram represents a very large sample from that population. This is intended to show that for a very large sample, the cumulative histogram approaches the CDF.

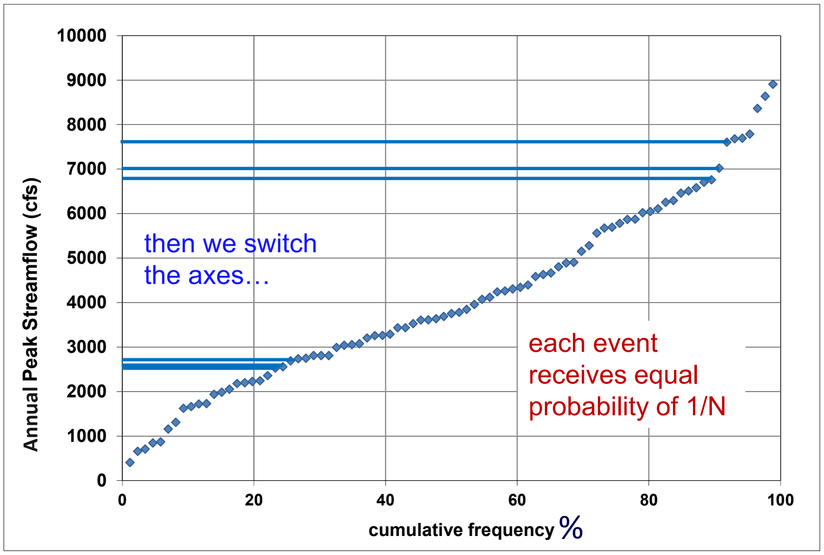
However, the cumulative histogram is not the ideal representation of the CDF…

Another Estimate of CDF



A better estimate of the CDF involves plotting all the data points, rather than condensing them into a histogram. It generalizes the histogram idea by assuming that bins are sized such that there is exactly one value per bin. This is the sample of 85 values. The vertical lines are an approximation of showing this as a form of cumulative histogram.

Another Estimate of CDF

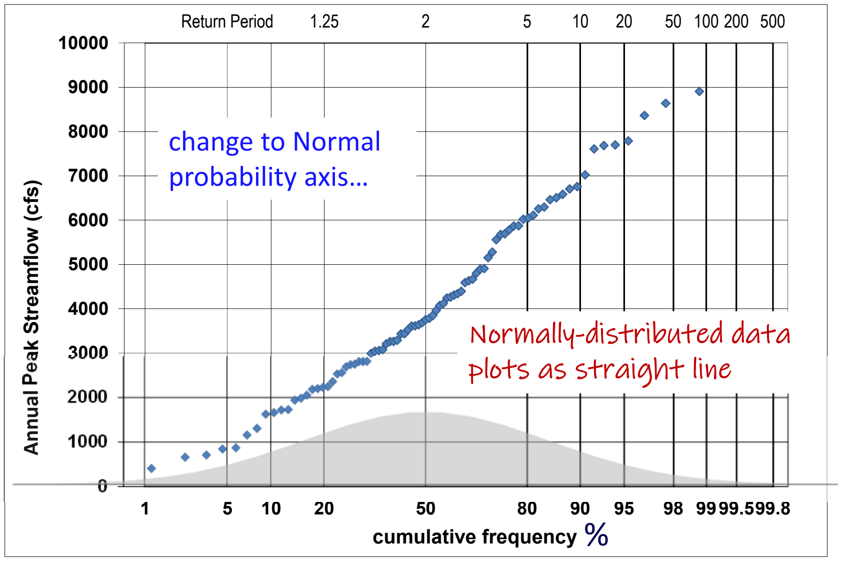


then we switch the axes…

each event receives equal probability of 1/N

In this method of estimating the CDF, the axes are switched to put the variable on the vertical and frequency on the horizontal. Note that cumulative frequency increases by the same amount with each event, because the count of events equal to or below that value is increasing by 1 each time.

Another Estimate of CDF



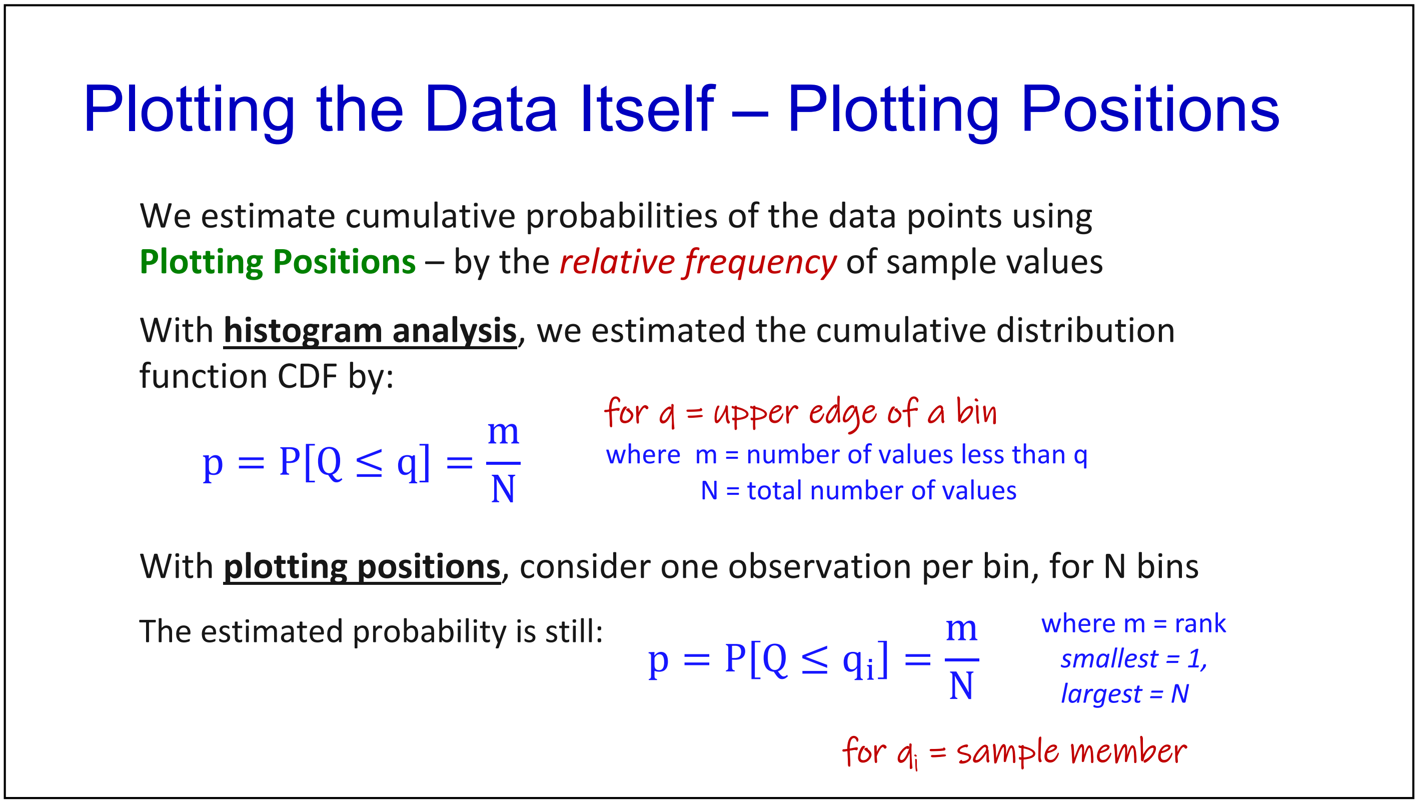
change to Normal probability axis…

Normally-distributed data plots as straight line

Next, the formerly linear probability/frequency axis has been turned into a Normal Probability axis. Note, this axis scaling brings values in the middle closer together than the ones at the tails farther apart. Normally distributed data will plot as a straight line on a Normal probability axis.

The Normal probability axis is scaled based on the Normal PDF that can be seen sitting on the horizontal axis. Starting on the left, the value on the axis represents the area under the PDF that has been accumulated by that point. 5% of the area has been accumulated to the left of where the axis shows 5. 50% has been accumulated to the left of where the axis shows 50. So, plot a probability where that PDF has accumulated that probabilty.

The Normal probability axis is actually linear in the standard Normal deviate, Z, which will be discussed in later slides.



Plotting the Data Itself – Plotting Positions

We estimate cumulative probabilities of the data points using **Plotting Positions** – by the *relative frequency* of sample values

With **histogram analysis**, we estimated the cumulative distribution function CDF by:

$$p = P[Q \leq q] = \frac{m}{N}$$

for q = upper edge of a bin

where m = number of values less than q
N = total number of values

With **plotting positions**, consider one observation per bin, for N bins

The estimated probability is still:

$$p = P[Q \leq q_i] = \frac{m}{N}$$

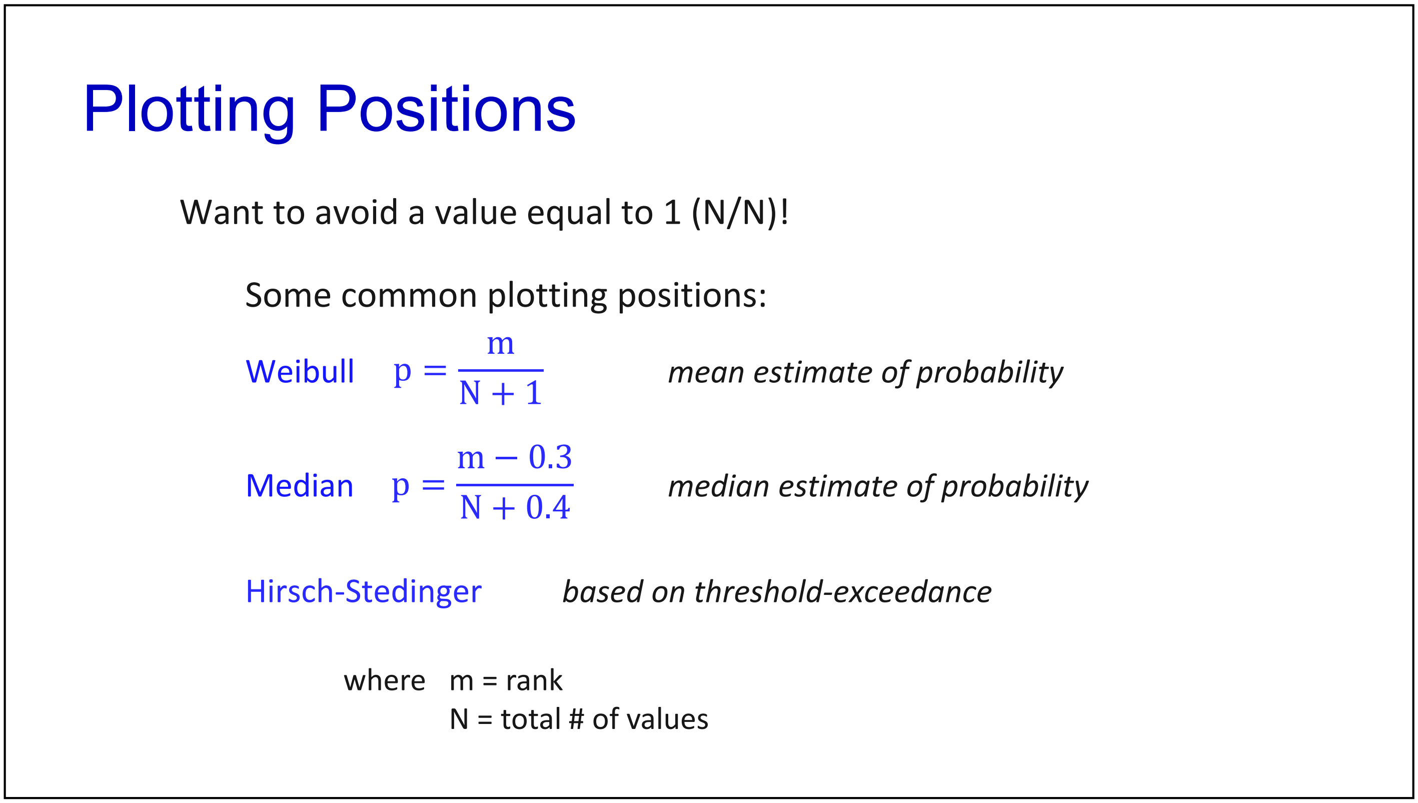
where m = rank
smallest = 1,
largest = N

for q_i = sample member

In general, we're using relative frequency to estimate the probability of being less that a certain value by how often the observations WERE less than that value, given our available data. That probability estimate by relative frequency is called a PLOTTING POSITION, because is defines where to plot the observation on the probability axis.

For a histogram, we compute that relative frequency at the edge of every bin. The probability that variable Q is less than value q is the number of values that were less than q. m is the number of values less q, and is the sum of all histogram bins below q. N is the total number of values. m/N is thus the relative frequency of values less than q.

For a plotting positions, we compute that relative frequency for every member of the sample. The probability that variable Q is less than value q is still the number of values that were less than q. We sort the observed values and rank them from 1 to N. The rank, m, is the number of values less than or each to that observation. N is the total number of values. m/N is thus the relative frequency of values less than q, which is now computed for every sample member, rather than only for every bin.



Plotting Positions

Want to avoid a value equal to 1 (N/N)!

Some common plotting positions:

Weibull $p = \dfrac{m}{N+1}$ *mean estimate of probability*

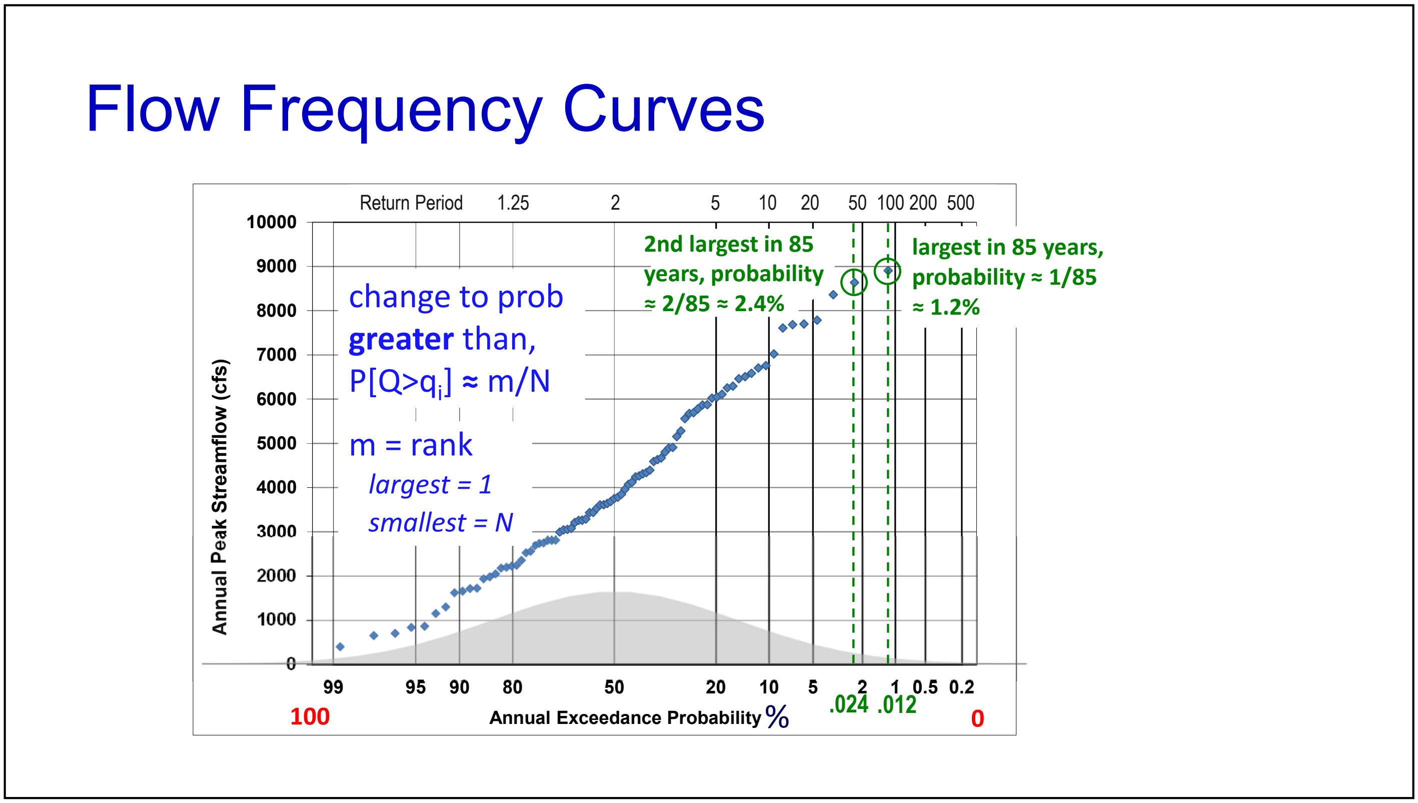
Median $p = \dfrac{m - 0.3}{N + 0.4}$ *median estimate of probability*

Hirsch-Stedinger *based on threshold-exceedance*

where m = rank
N = total # of values

Since either the largest or smallest event will have rank m = N, the simple m/N plotting position can equal 1.0. For non-exceedance, this implies that it's impossible to have an event larger than the largest in the record, which is a poor assumption. Even for exceedance, the assumption that we can't have an event smaller than the smallest is problematic, and either way, this outcome implies a bias in all of the plotting positions.

There are many other plotting positions derived using m and N that have good properties. Weibull and Median, shown on this slide, are commonly used. Weibull estimates the mean of the PDF describing the uncertainty in the estimate of probability, which means it is unbiased. Median provides the median of the PDF of uncertainty, which is a good comparison to fitted analytical probability distributions.



Flow Frequency Curves

change to prob **greater** than, $P[Q>q_i] \approx m/N$

m = rank

largest = 1

smallest = N

2nd largest in 85 years, probability ≈ 2/85 ≈ 2.4%

largest in 85 years, probability ≈ 1/85 ≈ 1.2%

The next adjustment is switching from P[Q>q], the probability of being less than, to P[Q>q], the probability of being greater than. We do this by sorting the data from high to low, so m represents the number of values greater than or equal to.

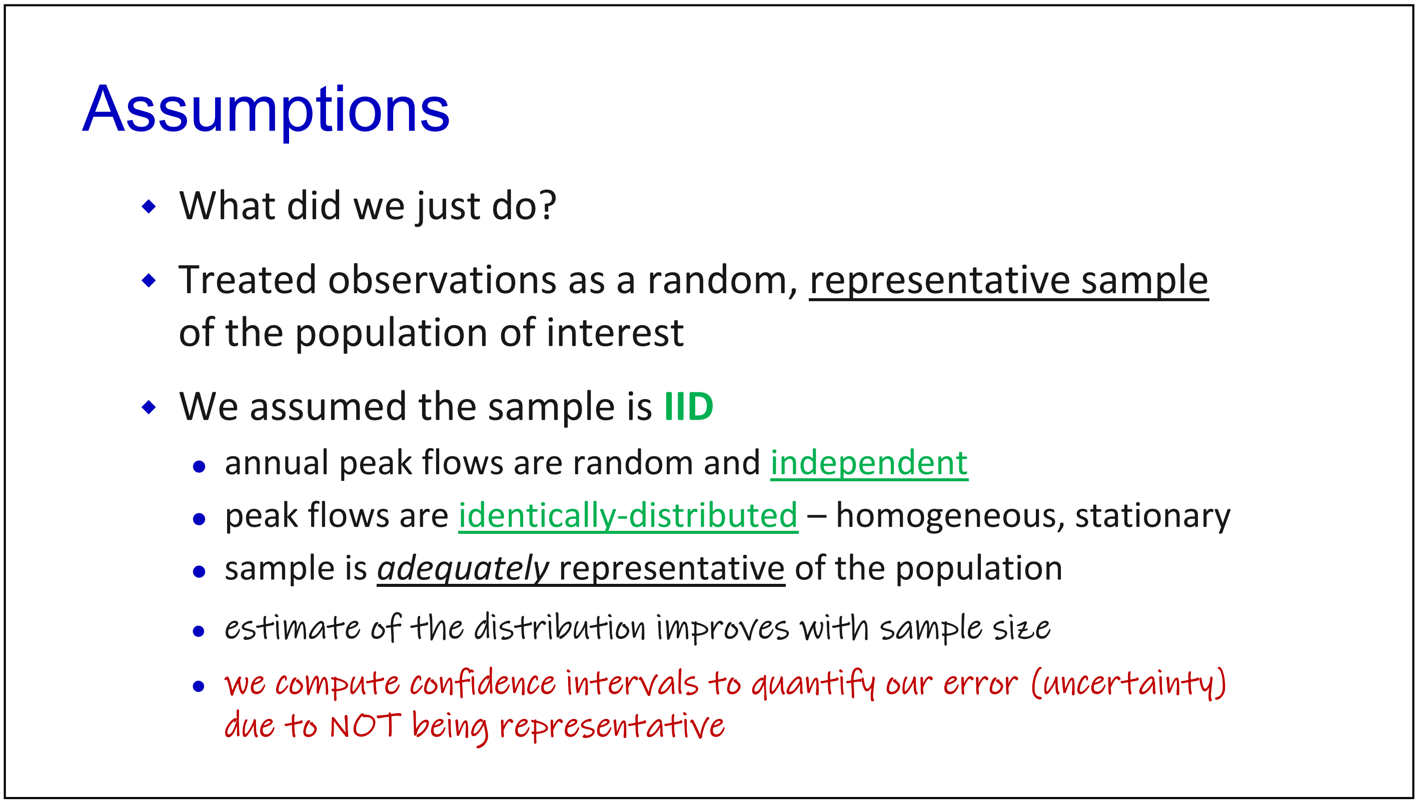
The axis is now Exceedance Probability (the probability of being greater), but has been reversed so that zero is on the right. There is now also an upper horizontal axis showing return period. Return period is the average number of years between exceedances of the flow value, and is equal to 1 / exceedance probability. Note that the largest sample member plots as the largest in 85 years, and plots at approximately 1/85 = 1.2 % chance of exceedance. The second largest sample member is equaled or exceeded twice in 85 years, and so plots about 2/85 = 2.4 %. Each sample member plots at the relative frequency of exceedance of its value as defined by its rank in the sample.

Cumulative Distribution Function

We can estimate a CDF from ranked observations versus plotting positions.

As the number of observations becomes very large, the *estimated* cumulative distribution function approaches the *true* cumulative distribution function for a continuous random variable

- population CDF

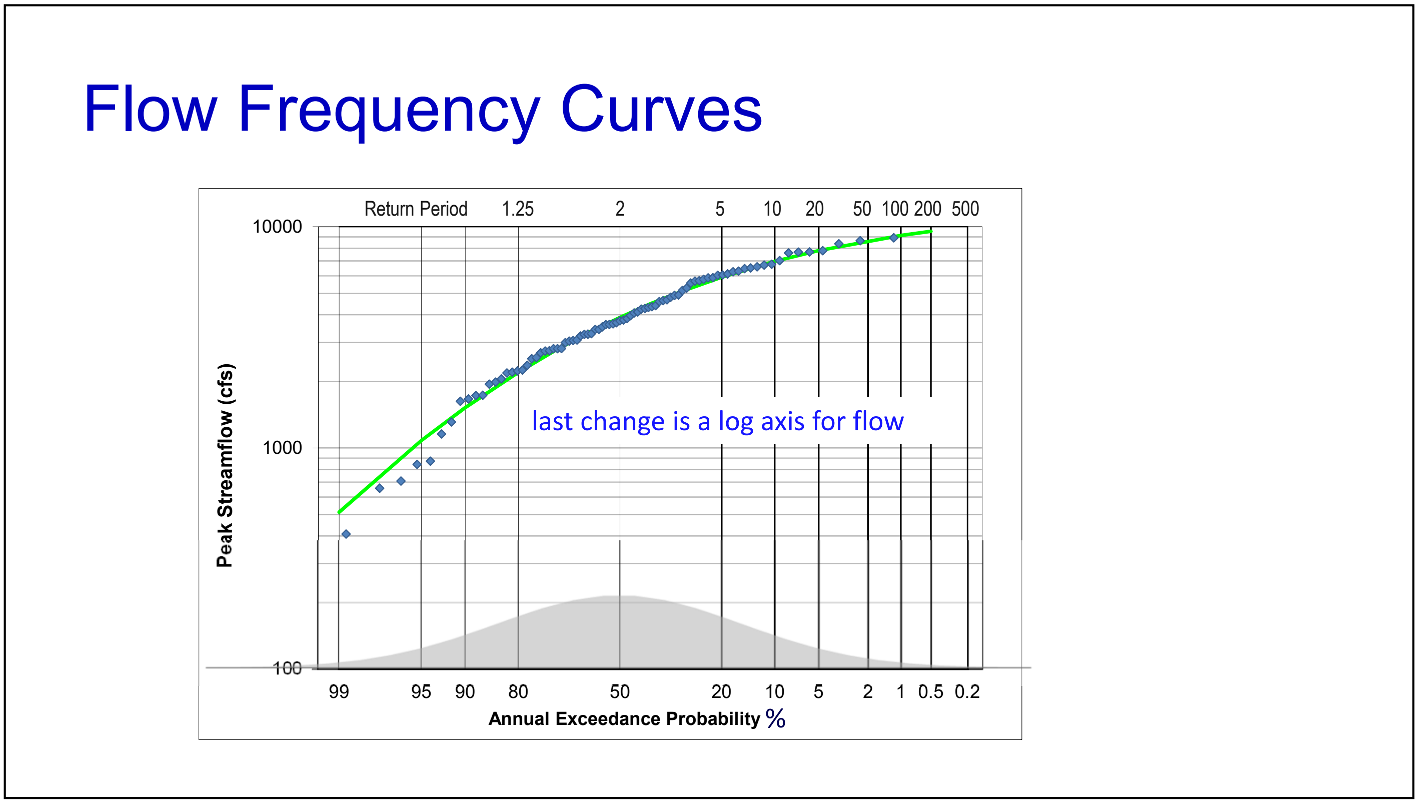


Assumptions

- What did we just do?
- Treated observations as a random, representative sample of the population of interest
- We assumed the sample is **IID**
  - annual peak flows are random and independent
  - peak flows are identically-distributed – homogeneous, stationary
  - sample is *adequately* representative of the population
  - estimate of the distribution improves with sample size
  - we compute confidence intervals to quantify our error (uncertainty) due to NOT being representative

Not so much that they ARE identically distributed as that it is reasonable to fit a single distribution to them.

ID: All value from the same prob distr. OR, all values can be effectively represented by the same probability distribution



Flow Frequency Curves

last change is a log axis for flow

The final axis adjustment is to put streamflow on a log axis. Often this brings the plotted data closer to a straight line, which would imply a log Normal distribution or something close to it. In the case of this data, it is more curved on a log axis.

So far, we've done an empirical analysis in which we've looked only at the data set itself and made no further assumptions. The next step is estimating an analytical probability distribution for this data, which is the green line on the plot.

Alternate Labels for Frequency Axis

Annual

Exceedance Probability = Prob. of being greater than value

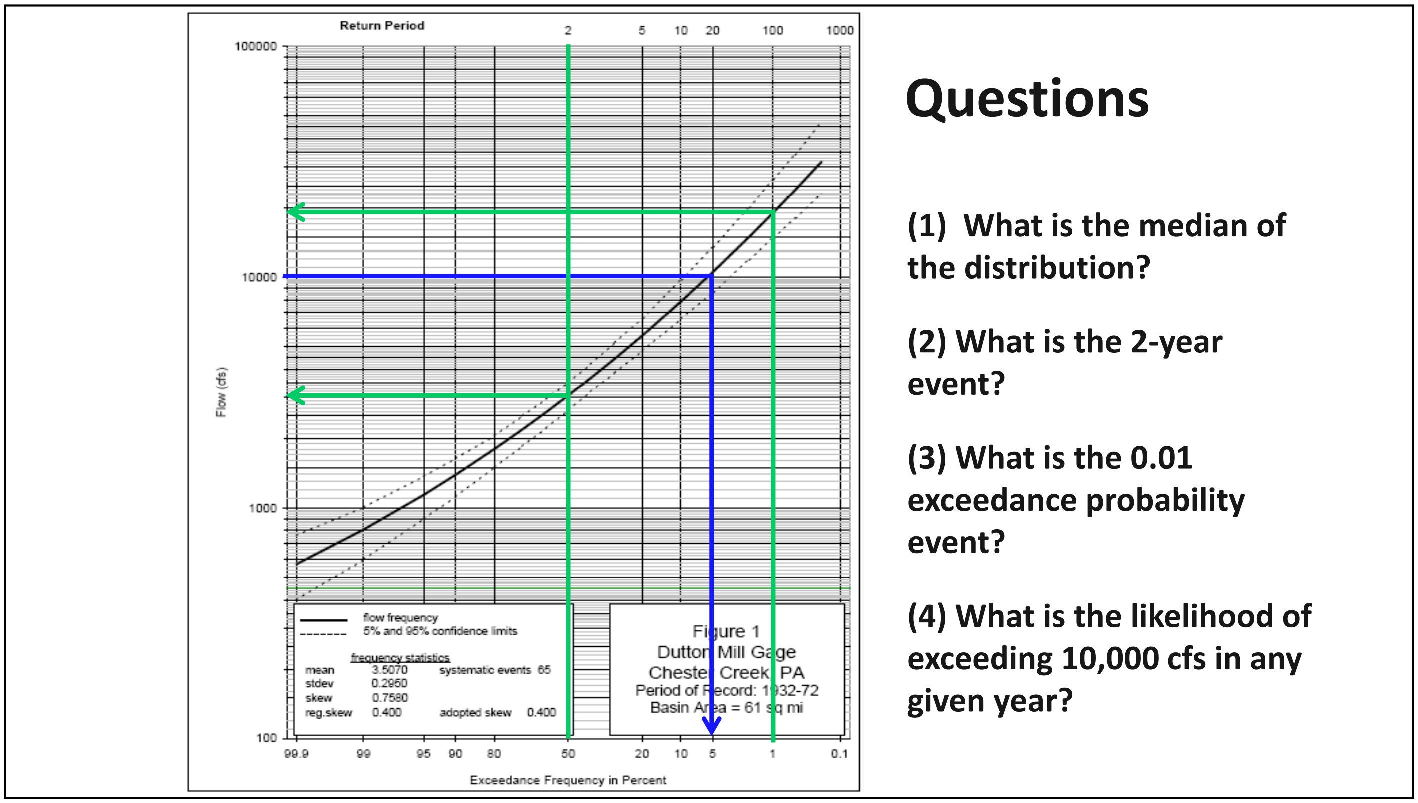
Exceedance Frequency = Exc. Probability * 100

%-Chance Exceedance = Exc. Probability * 100

Return Period = 1 / Exceedance Probability

Recurrence Interval = 1 / Exceedance Probability

Exceedance Interval = 1 / Exceedance Probability



Questions

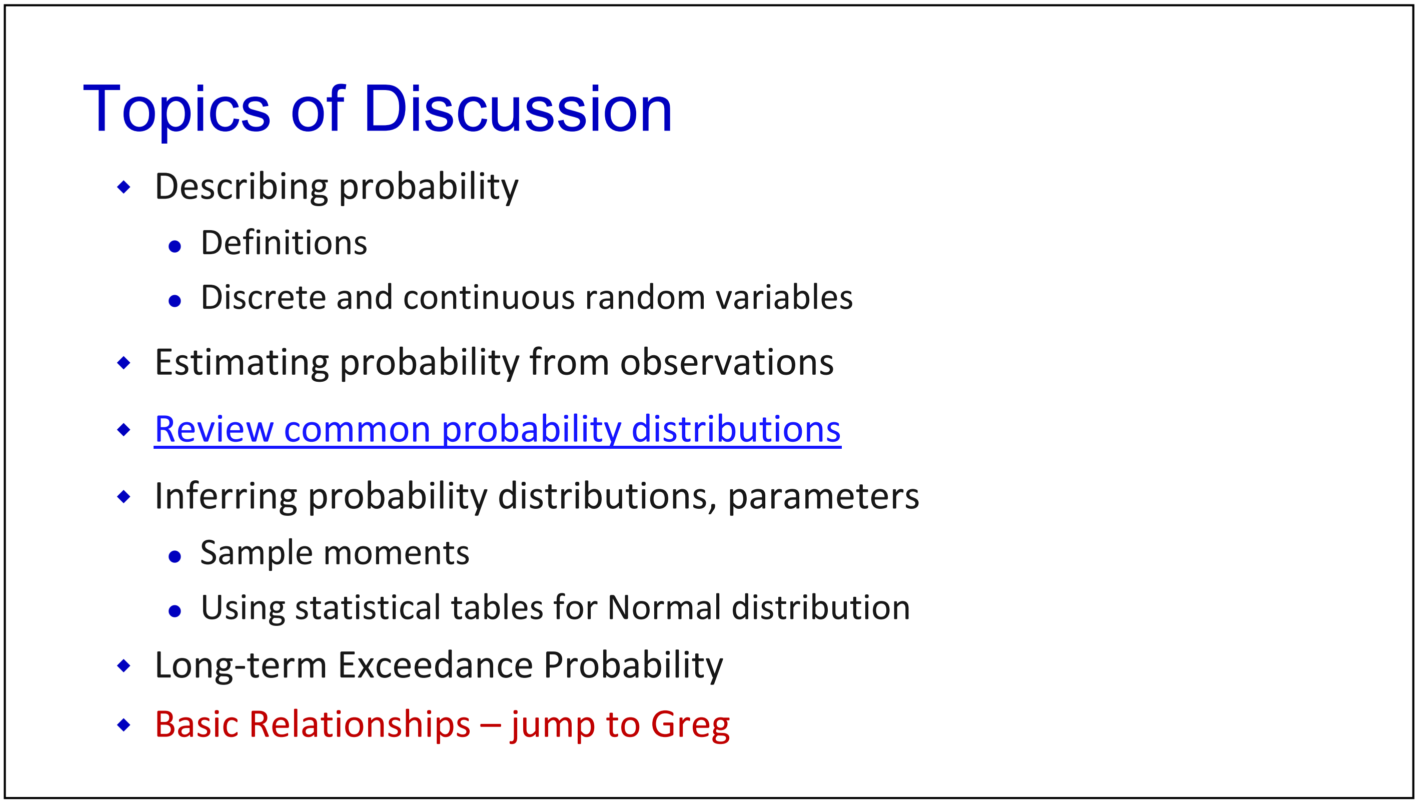
(1) What is the median of the distribution?

(2) What is the 2-year event?

(3) What is the 0.01 exceedance probability event?

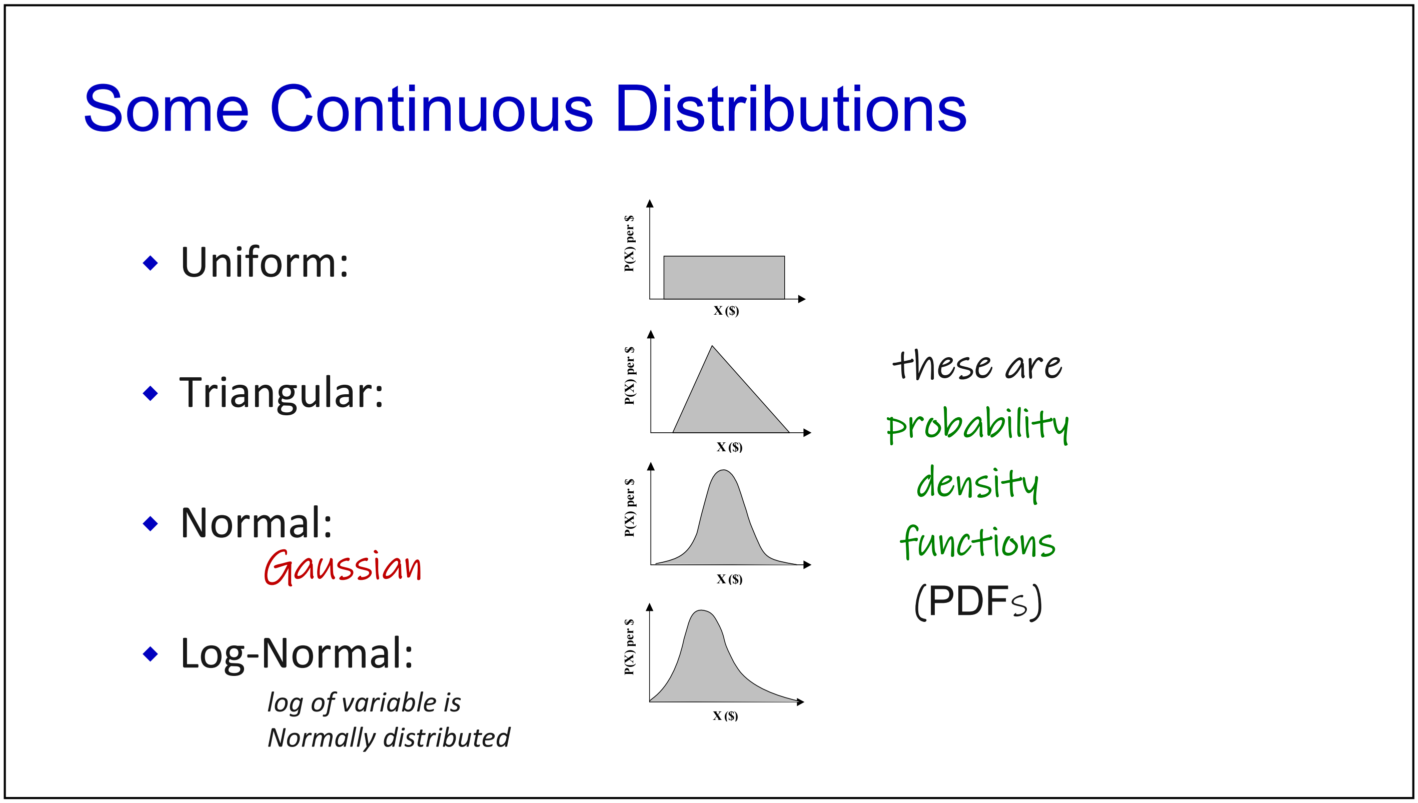
(4) What is the likelihood of exceeding 10,000 cfs in any given year?

This figure shows a fitted analytical distribution and a 90% confidence interval, which we'll discuss further in later slides. Once we have an estimated probability distribution, it is a relationship between our variable (annual maximum flow) and probability that can answer questions in either direction. We can specify flow and determine its exceedance probability, and we can specify exceedance probability and determine the flow that has that probability, called a flow quantile.



Topics of Discussion

- Describing probability
  - Definitions
  - Discrete and continuous random variables
- Estimating probability from observations
- Review common probability distributions
- Inferring probability distributions, parameters
  - Sample moments
  - Using statistical tables for Normal distribution
- Long-term Exceedance Probability
- Basic Relationships – jump to Greg



Some Continuous Distributions

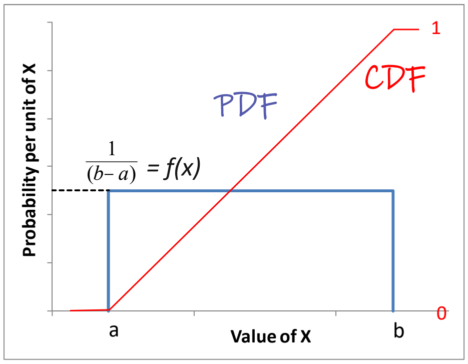
- Uniform:
- Triangular:
- Normal:
  Gaussian
- Log-Normal:
  log of variable is Normally distributed

these are probability density functions (PDFs)

P(X) per $
X ($)
P(X) per $
X ($)
P(X) per $
X ($)
P(X) per $
X ($)

In this lecture, we'll be looking at these four common probability distributions. These are useful for different purposes. Some can be used for describing hydrologic variables, and others are more useful to describe uncertainty distributions.

Uniform Distribution

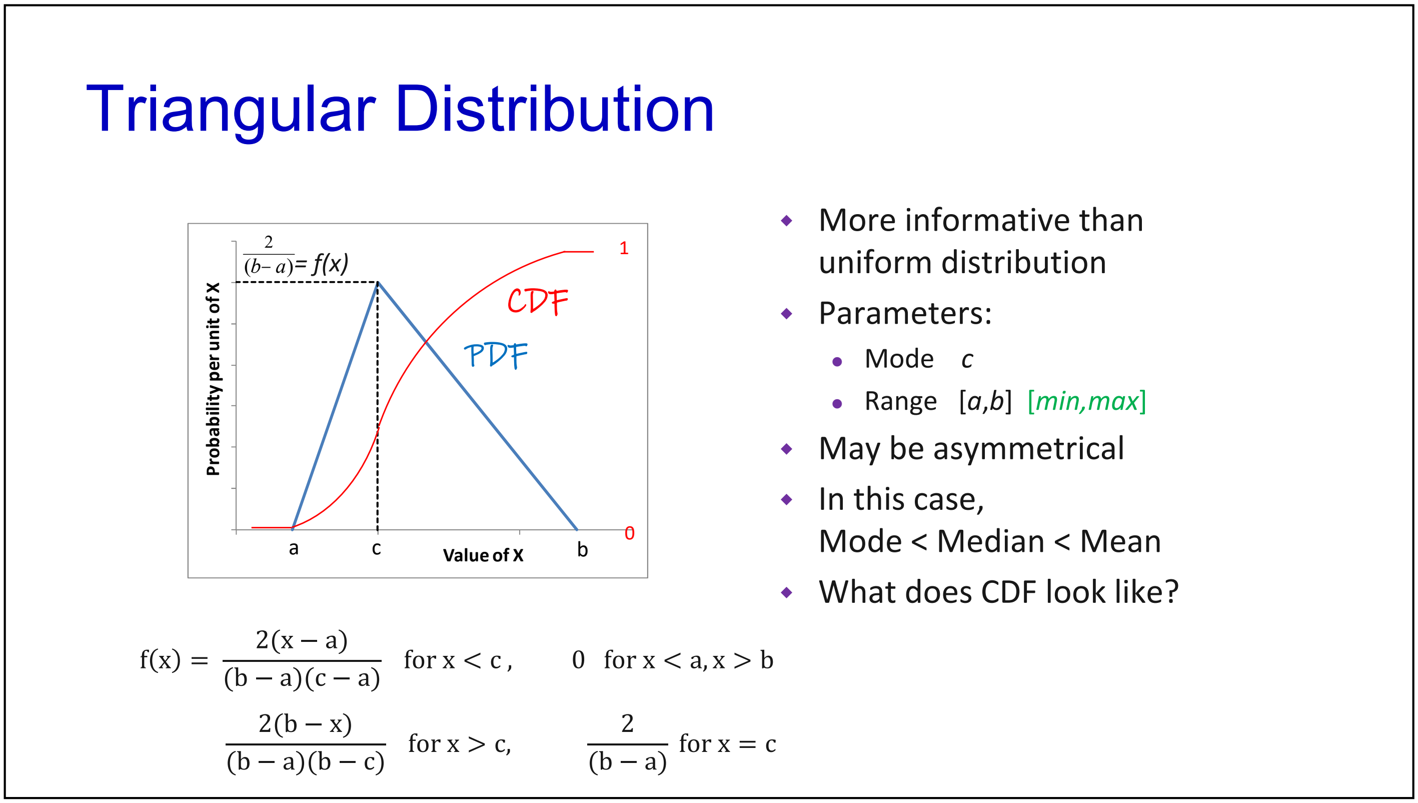


$$f(x) = \frac{1}{b-a} \quad \text{for a } \le X \le b$$

$$0 \quad \text{otherwise}$$

- Interpretation:
  - All values in range are equally likely
- Parameters:
  - Range [*a*,*b*] [*min*,*max*]
- What is the mean, median, mode?
- What does the CDF look like?

A common use of the Uniform distribution with min and max of 0 and 1, designated as U[0,1], is for generating values to act as cumulative probabilities to randomly sample from any probability distribution for Monte Carlo simulation, as we did with the dice-rolling spreadsheet.



Triangular Distribution

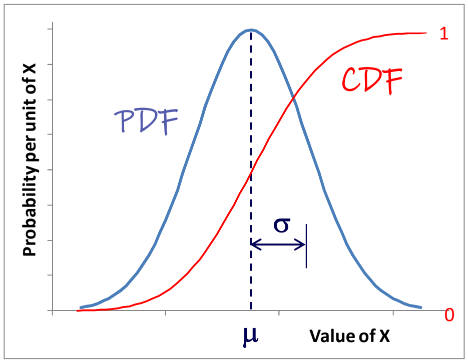
- More informative than uniform distribution
- Parameters:
  - Mode *c*
  - Range [*a*,*b*] [*min*,*max*]
- May be asymmetrical
- In this case,
  Mode < Median < Mean
- What does CDF look like?

$$f(x) = \frac{2(x-a)}{(b-a)(c-a)} \text{ for } x < c, \quad 0 \text{ for } x < a, x > b$$

$$\frac{2(b-x)}{(b-a)(b-c)} \text{ for } x > c, \quad \frac{2}{(b-a)} \text{ for } x = c$$

The triangular distribution can be useful in describing uncertainty around an estimate because it can be either symmetrical or asymmetrical.

Normal (Gaussian) Distribution

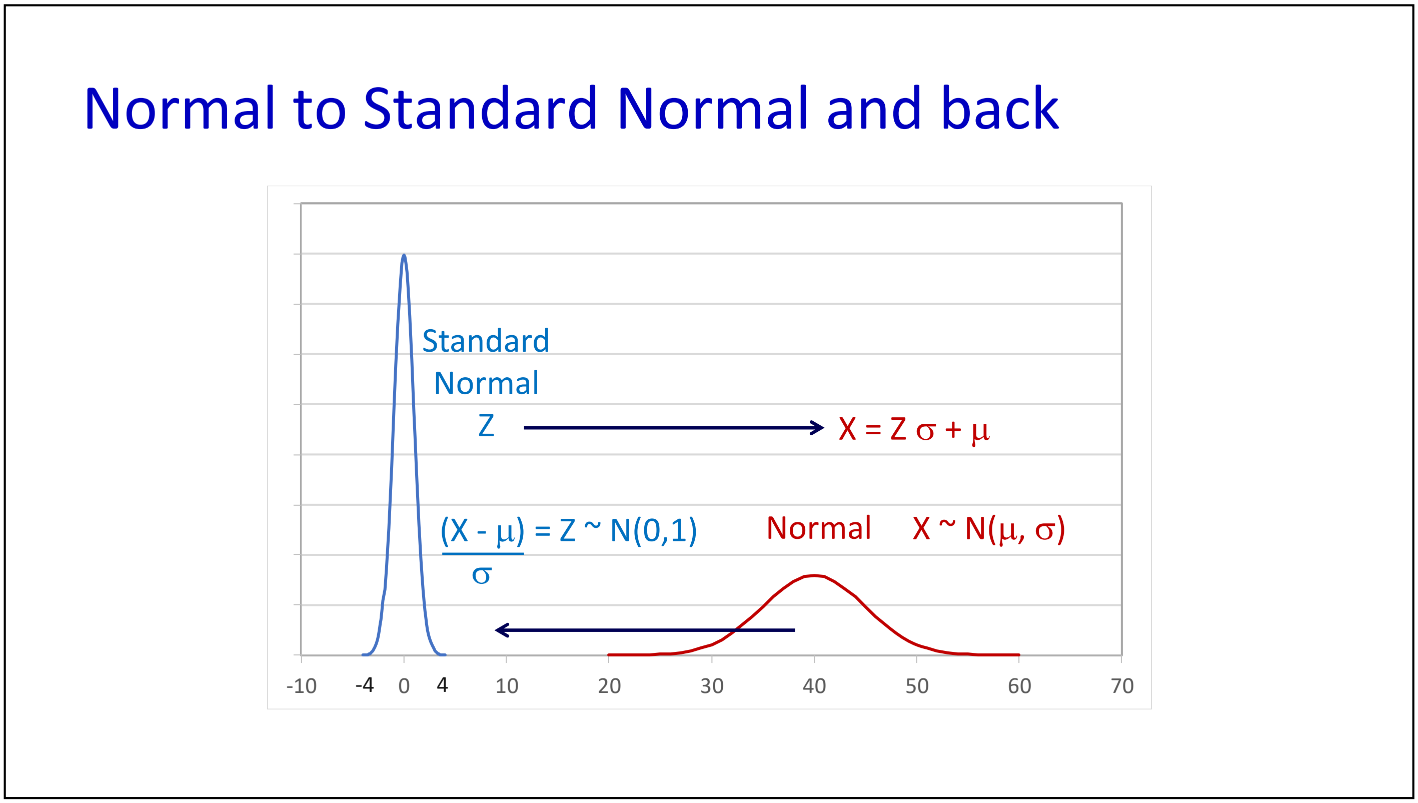


$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

- Most common distribution found in nature, very useful
- Has a defined shape, eqn
- Parameters:
  - Mean μ
  - Variance σ^2 (standard deviation σ)
- Symmetrical
  - Mean = Median = Mode
  - Skew coefficient = 0
- Scalable from Standard Normal

  Mean = 0, St.Dev = 1
- What does CDF look like?

The Normal distribution is useful both to describe variables and to describe uncertainty.

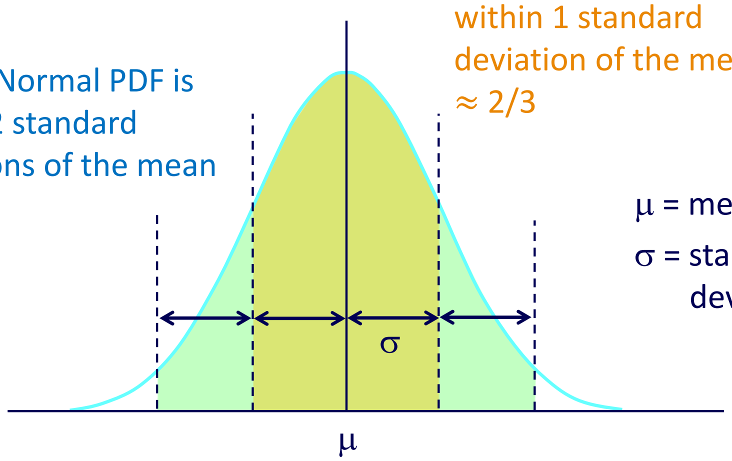


Normal to Standard Normal and back

This slides shows the transformation between Normal and Standard Normal distributions. Since Standard Normal has mean of zero and standard deviation of one, we produce a Standard Normal variate by subtracting the mean and the dividing by the standard deviation. We transform back to the Normal distribution by multiplying by the standard deviation and adding back the mean.

NOTE, the area under a PDF is 1.0, and so the two PDF in this image SHOULD have the same area. They do not, to make an easier image to read, but note that the red PDF should be much lower.

Normal Distribution



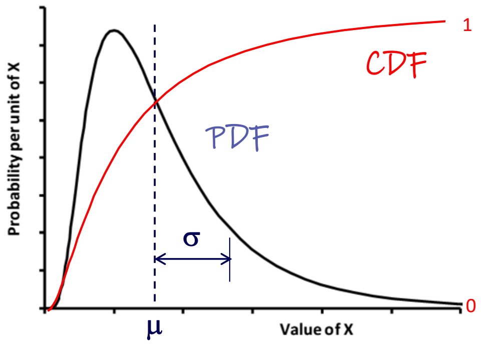
95% of Normal PDF is within 2 standard deviations of the mean

68% of Normal PDF is within 1 standard deviation of the mean ≈ 2/3

μ = mean

σ = standard deviation

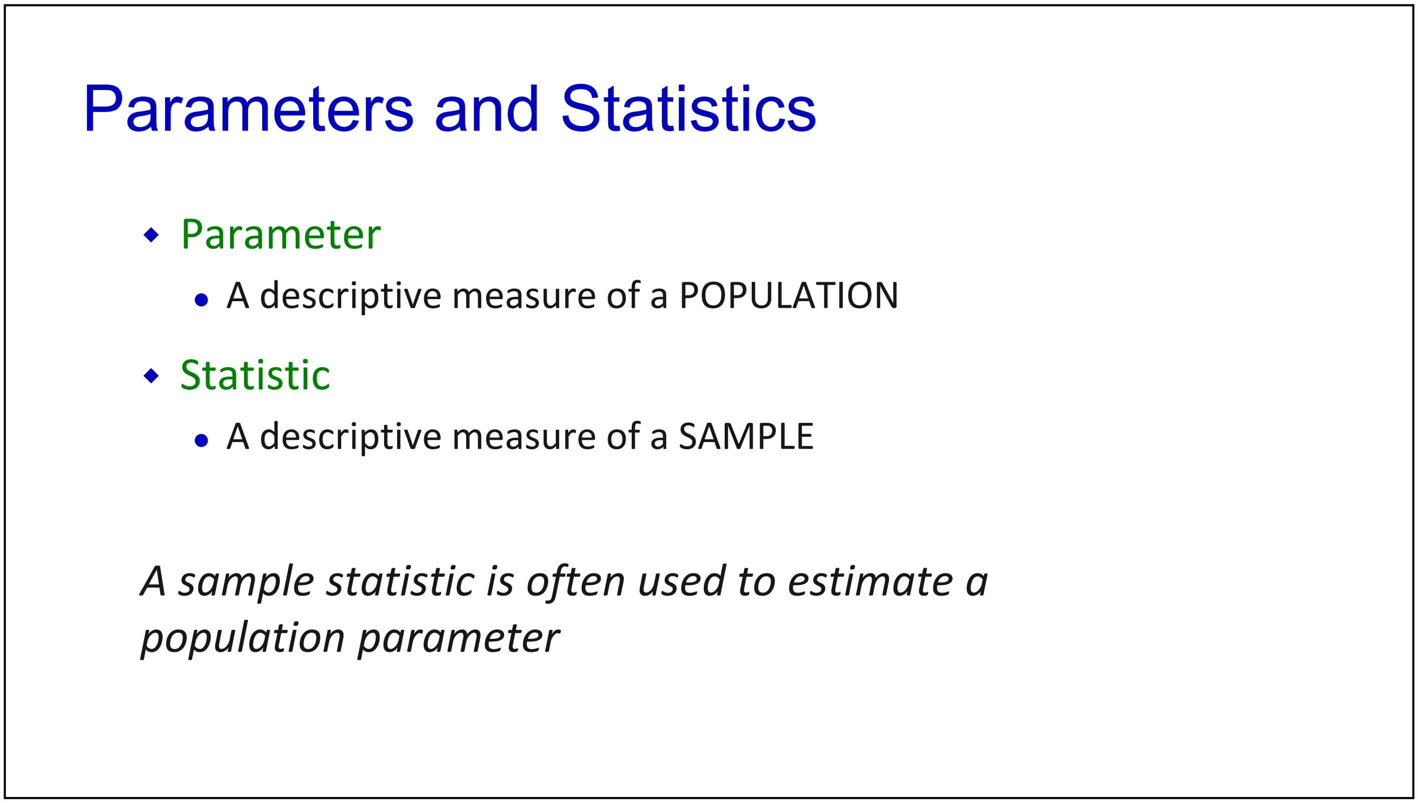
LogNormal Distribution



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

- The **LOG** of the variable X has a Normal distribution
- natural or base-10 logarithm
- Has a defined shape, eqn
- Parameters:
  - Mean of logX = μ
  - Variance of logX = σ^2 (st.dev σ)
- Has "fixed" skew defined by μ and σ
- **Can't be less than zero!**
- What does CDF look like?

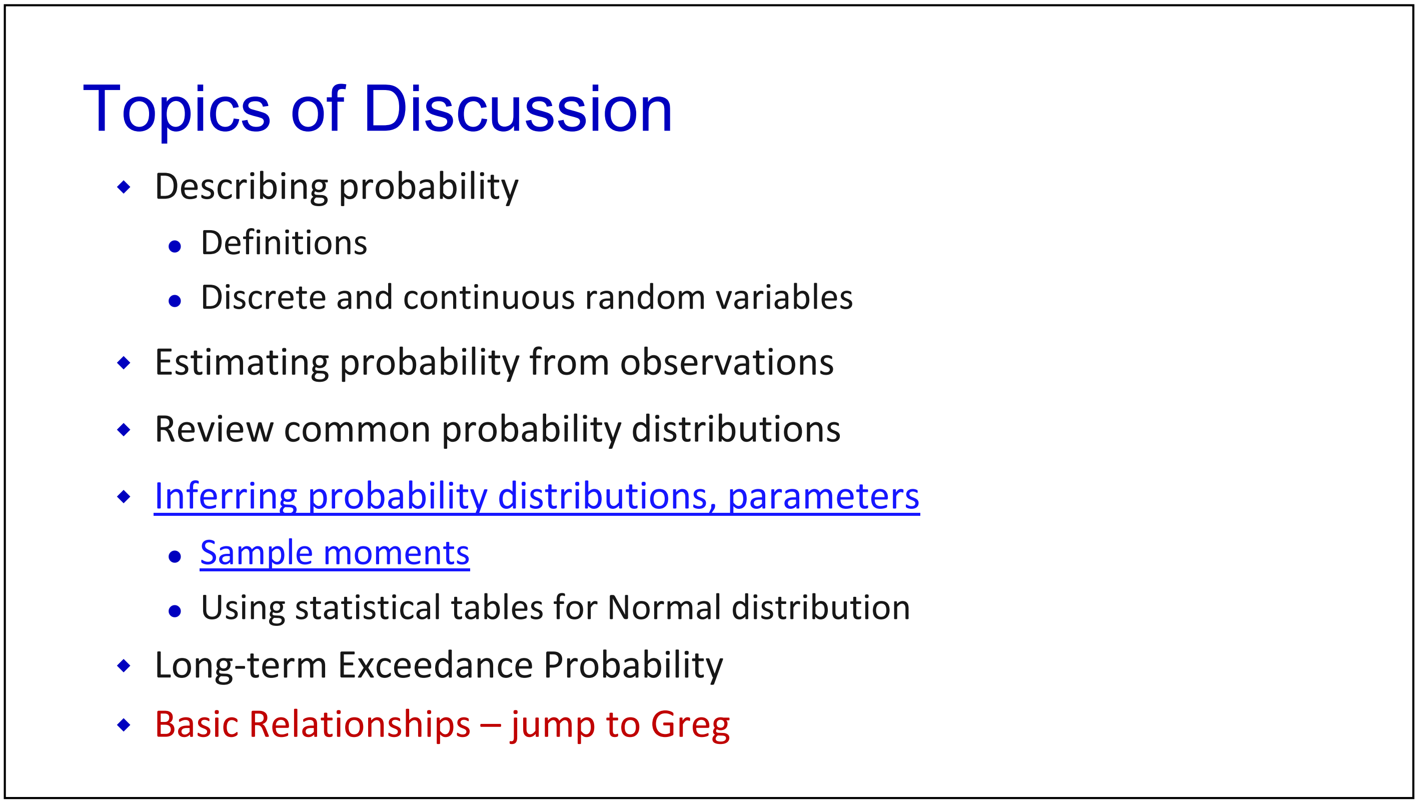
A log transform of the Normal distribution is useful when a variable has a positive skew and cannot have values less than zero. The parameters are based on the log transform of the variable.



Parameters and Statistics

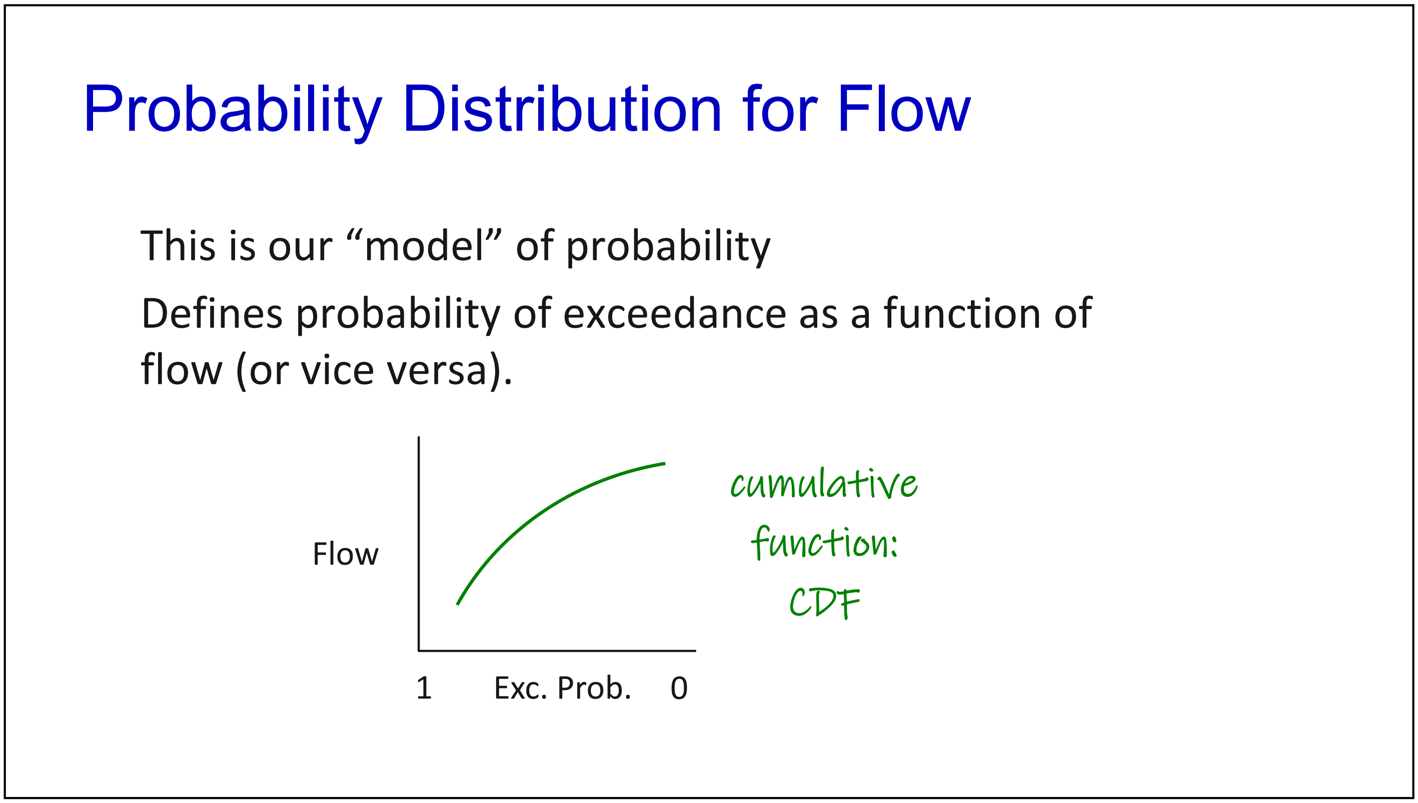
- Parameter
  - A descriptive measure of a POPULATION
- Statistic
  - A descriptive measure of a SAMPLE

A sample statistic is often used to estimate a population parameter



Topics of Discussion

- Describing probability
  - Definitions
  - Discrete and continuous random variables
- Estimating probability from observations
- Review common probability distributions
- Inferring probability distributions, parameters
  - Sample moments
  - Using statistical tables for Normal distribution
- Long-term Exceedance Probability
- Basic Relationships – jump to Greg



Probability Distribution for Flow

This is our “model” of probability

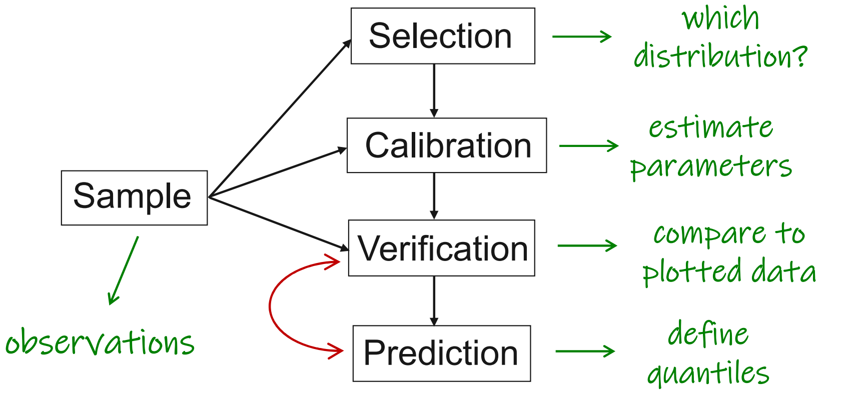
Defines probability of exceedance as a function of flow (or vice versa).

Flow
1
Exc. Prob.
0

cumulative function: CDF

A probability distribution is just another kind of model that represents the relationship between the variable of interest and its probability of occurrence.

Distribution Fitting Procedure



Quantile = value exceeded with a certain probability, p

We use the same steps in fitting a probability model as we do in developing other models. All steps are based on the sample of data that we assume is representative of parent population probability distribution.

We select a probability distribution, estimate its parameters from the sample, compute the CDF of that distribution, and compare that CDF to the sample, either as histogram or plotted points.

Because of how they're defined, the "prediction" step comes before the "verification" step because the CDF (the quantiles) must be computed before it can be compared to the data.

Elements of a Distribution

- Equation

  Defines the relationship between the variable and cumulative (or exceedance) probability

- Parameters

  A parameter is a coefficient in the equation

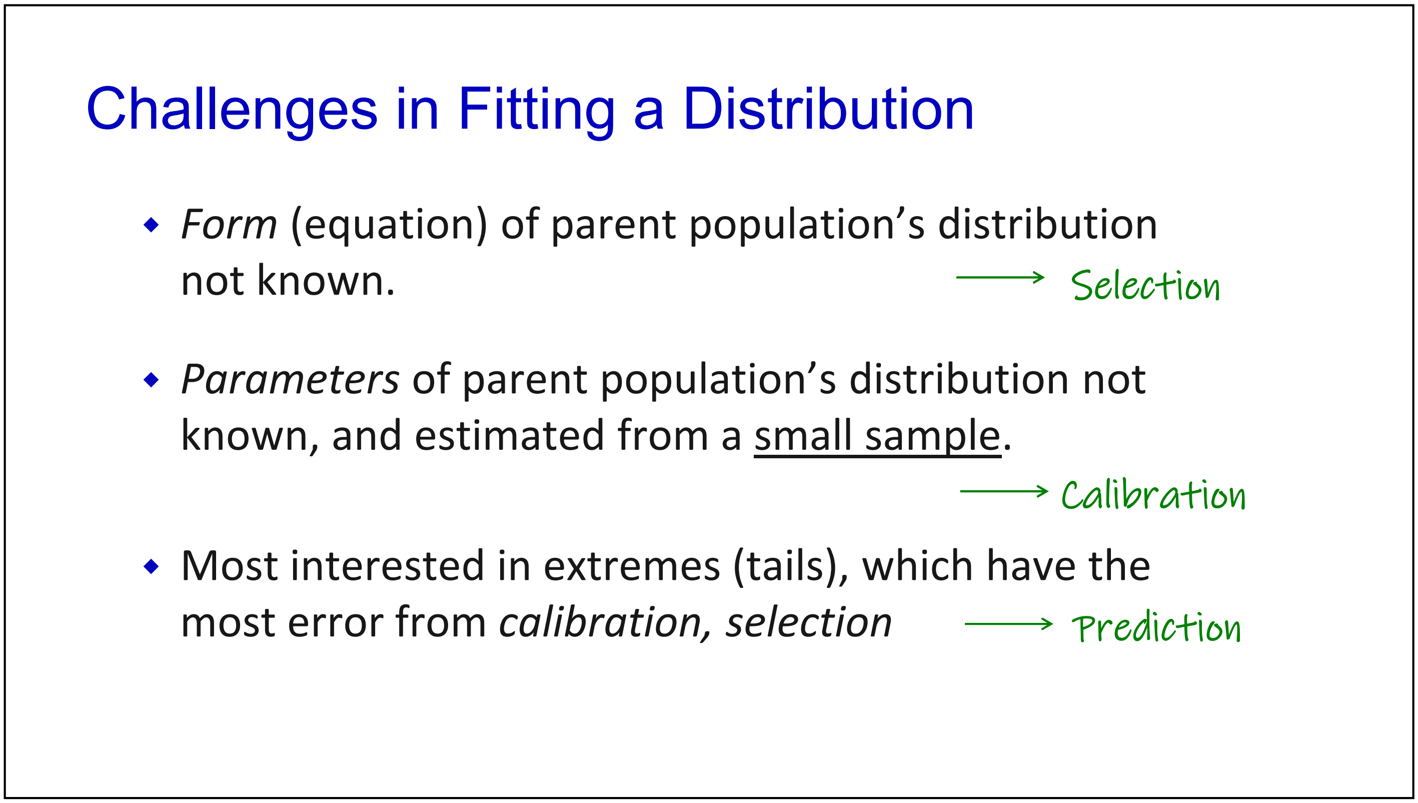
Example: Exponential Distribution

$$\text{Probability}< = 1 - e^{-\lambda \text{Flow}}$$

cumulative distribution function CDF

$$\text{Flow} = \frac{-\log(1-\text{Prob}<)}{\lambda}$$

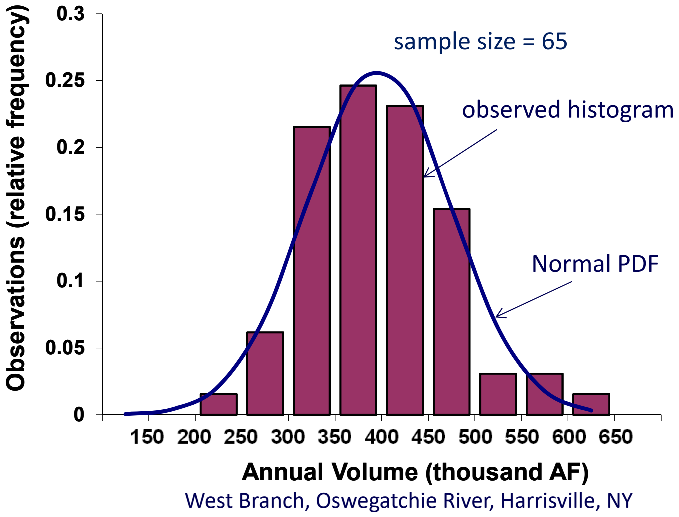
quantile function



Challenges in Fitting a Distribution

- *Form* (equation) of parent population's distribution not known. → Selection
- *Parameters* of parent population's distribution not known, and estimated from a small sample. → Calibration
- Most interested in extremes (tails), which have the most error from *calibration, selection* → Prediction

Selection of Distribution

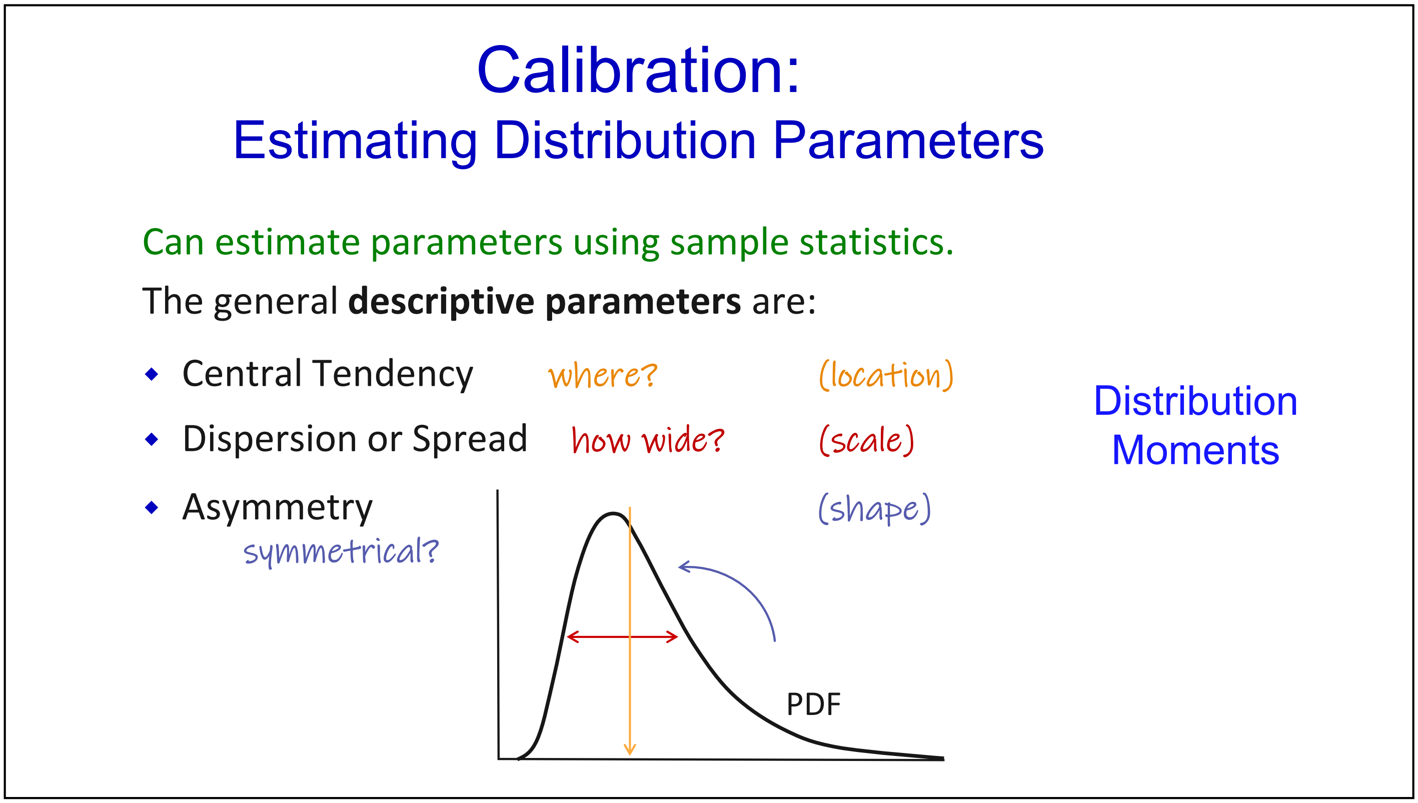


West Branch, Oswegatchie River, Harrisville, NY

This is the histogram of a 65 year record of annual flow volumes for a river in New York. By examining the histogram as a first step, we can decide what of the distributions we've looked at might be reasonable.

Uniform is clearly incorrect. Triangular is possible, but there are more values cluster in the middle, which is closer to Normal. It's not completely symmetrical, with a longer upper tail than lower, but perhaps close enough that the sample could have been generated by a symmetrical PDF.

The example will proceed with fitting a Normal distribution to the 65-member sample of flow volume.



Calibration:
Estimating Distribution Parameters

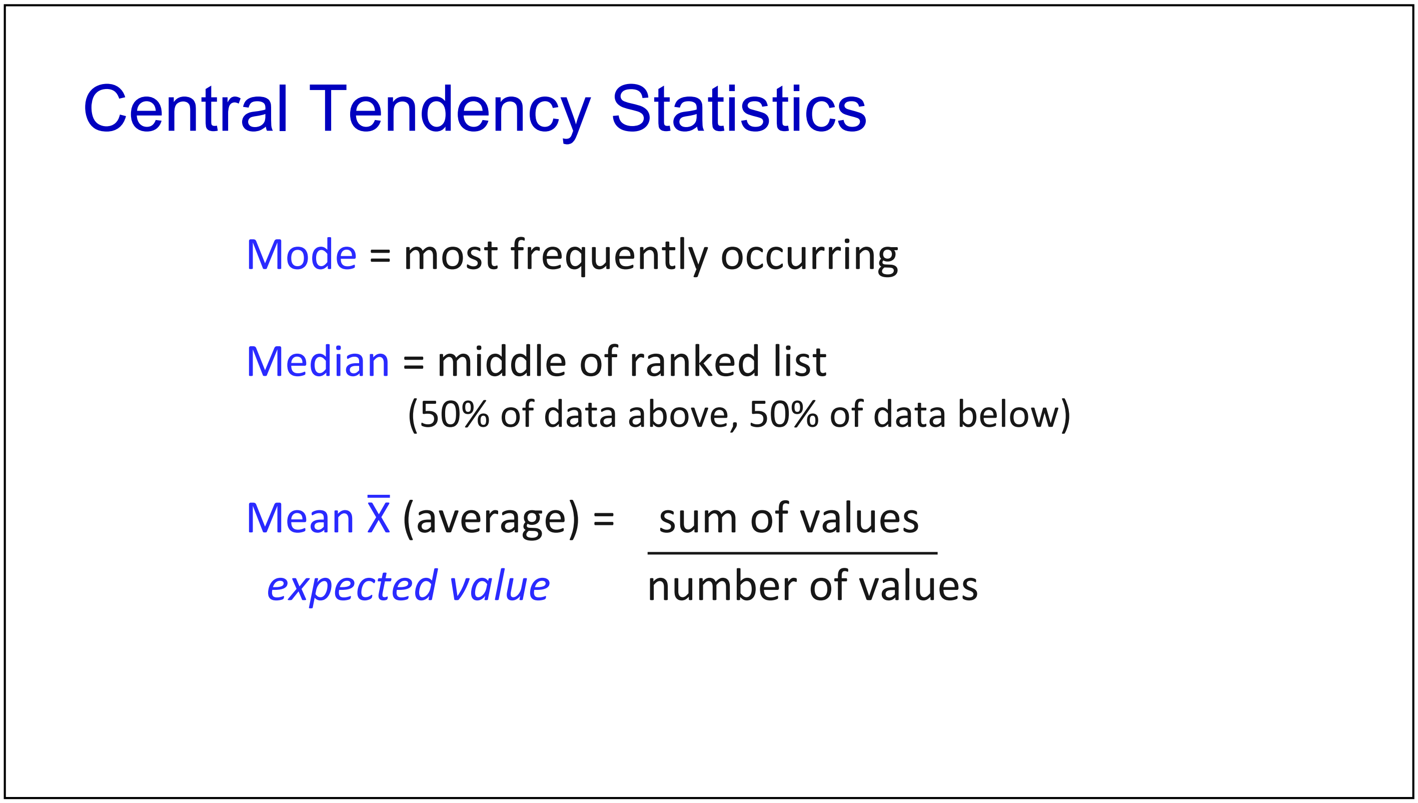
Can estimate parameters using sample statistics.

The general **descriptive parameters** are:

- Central Tendency where? (location)
- Dispersion or Spread how wide? (scale)
- Asymmetry (shape)
  symmetrical?

Distribution Moments

These description parameters are also known as the MOMENTS of the probability distribution.



Central Tendency Statistics

Mode = most frequently occurring

Median = middle of ranked list
(50% of data above, 50% of data below)

Mean $\bar{X}$ (average) = $\dfrac{\text{sum of values}}{\text{number of values}}$

expected value

These are the statistics of central tendency as estimated from a sample. The earlier mention of the these measures described their definitions for a probability distribution. These sample statistics can be used to estimate the distribution values.

Sample Mean, $\overline{X}$

expected value

Central value statistic

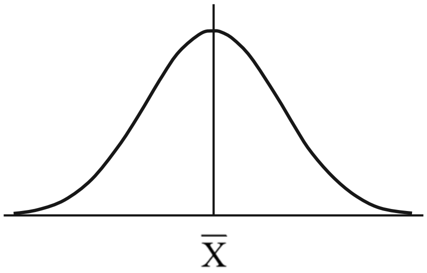
Sample mean is an estimate of population mean, μ

1st moment
about 0

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\overline{X} = \hat{\mu}$$

X_{bar} mu_{hat}



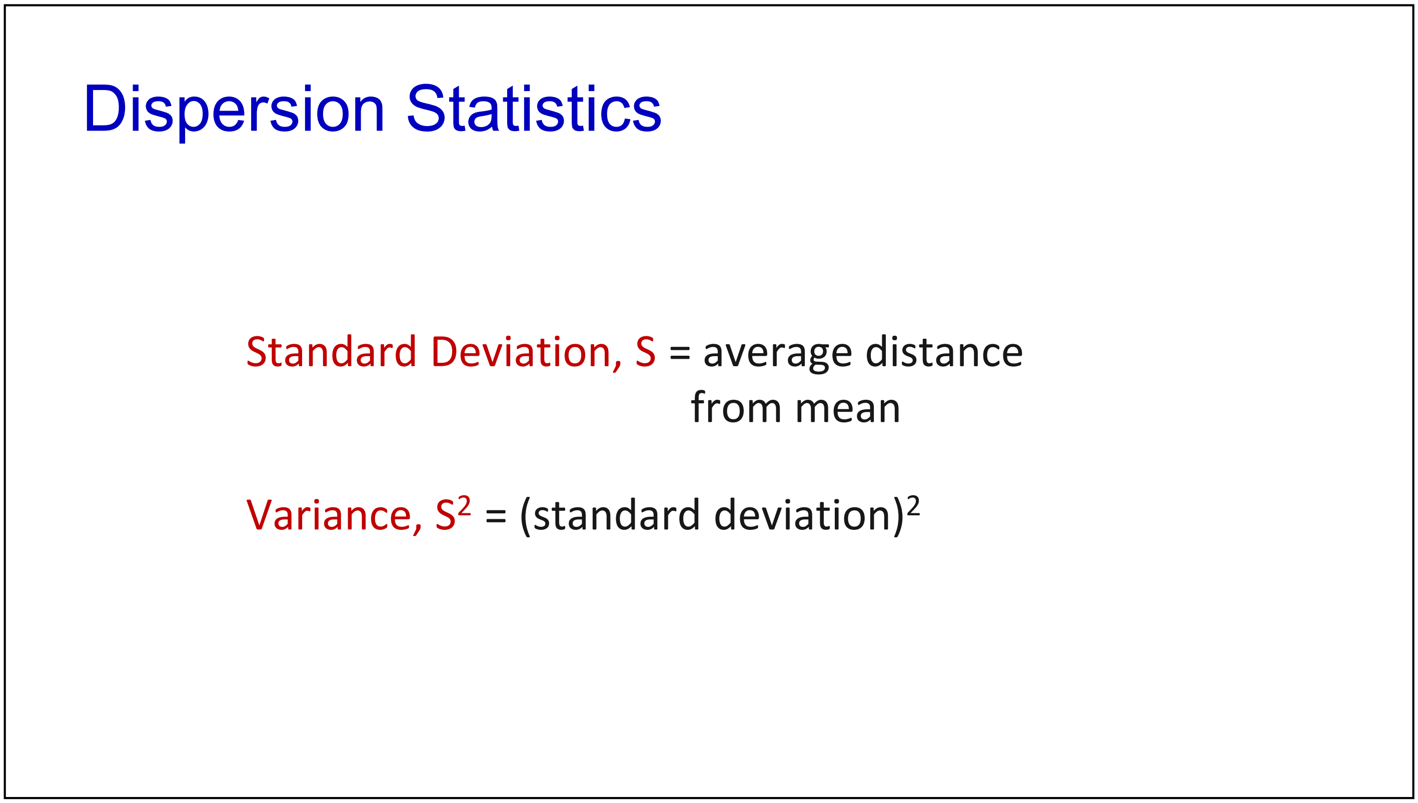
where: X_i = sample member, i

N = sample size

has same unit as variable

Sample mean is also known as expected value, and is noted as X-bar, or whatever variable designation with a bar over it.

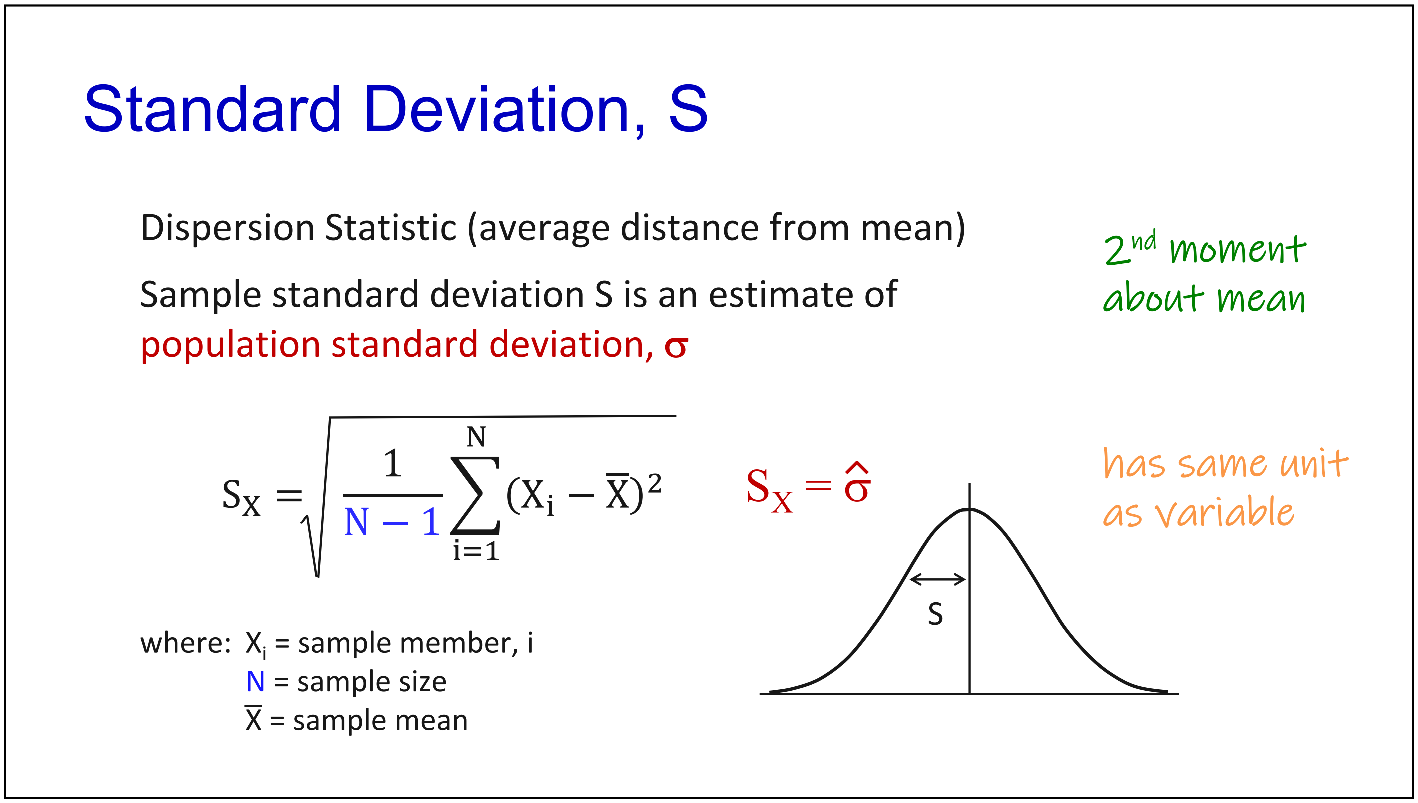
A change in the mean moves the PDF left (for lower) or right (for higher).



Dispersion Statistics

Standard Deviation, S = average distance
from mean

Variance, S^2 = (standard deviation)^2



Standard Deviation, S

Dispersion Statistic (average distance from mean)

Sample standard deviation S is an estimate of population standard deviation, σ

2nd moment about mean

$$S_X = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

$$S_X = \hat{\sigma}$$

has same unit as variable

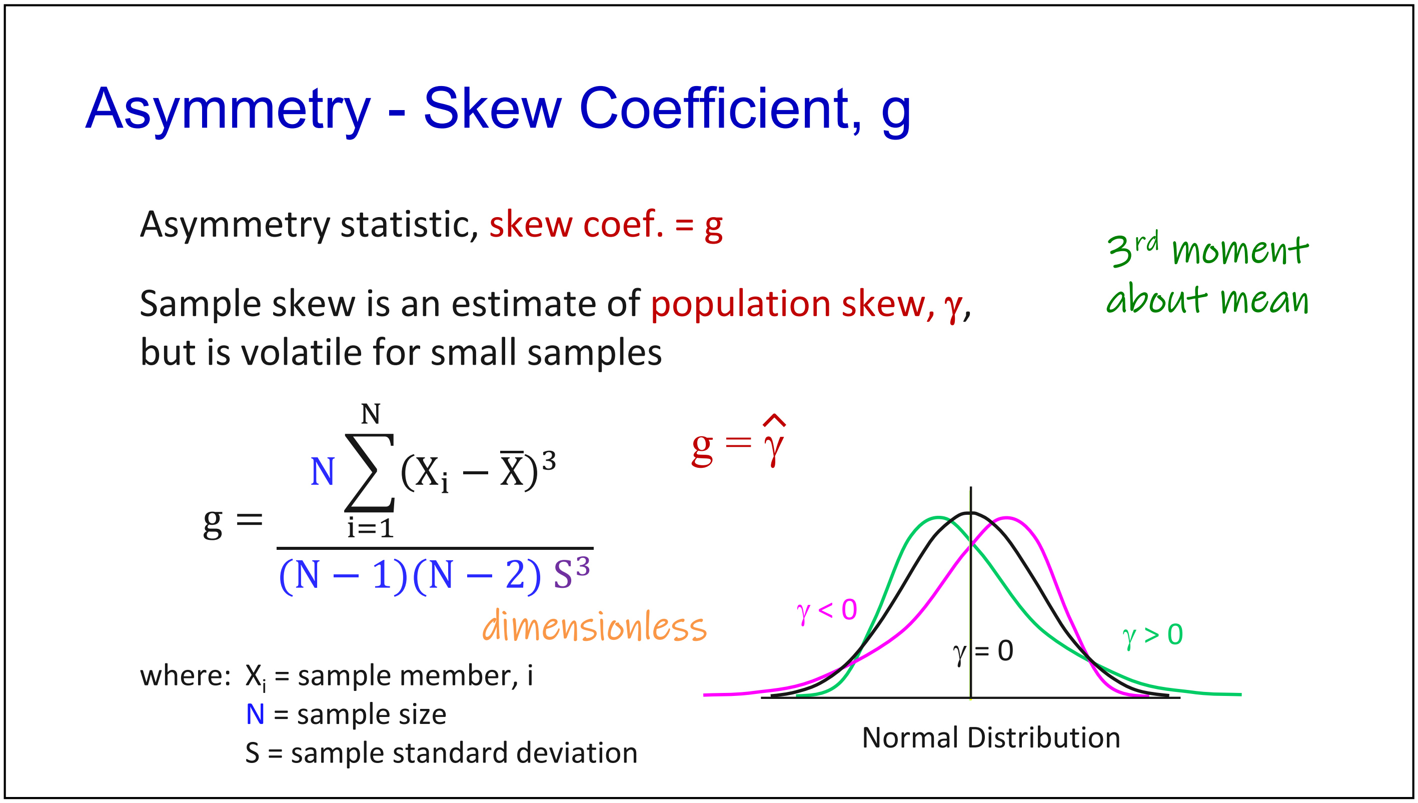
where: X_i = sample member, i

N = sample size

$\overline{X}$ = sample mean

Standard deviation, noted S, is the average distance from the sample mean. Because the sum of distances from the sample mean is equal to zero, this estimator squares the distances Xi – Xbar to remove the sign and make all values positive. After averaging, the square root it taken to return the metric to the original unit of the variable. The equation divides by N-1 rather than N as a correction for bias, to produce an unbiased estimator.

A change in the standard deviation makes the PDF narrower (for smaller) or wider (for larger).



Asymmetry - Skew Coefficient, g

Asymmetry statistic, skew coef. = g

3rd moment about mean

Sample skew is an estimate of population skew, γ, but is volatile for small samples

$$g = \frac{N\sum_{i=1}^{N}(X_i - \bar{X})^3}{(N-1)(N-2)\,S^3}$$

$$g = \hat{\gamma}$$

dimensionless

where: X_i = sample member, i
N = sample size
S = sample standard deviation

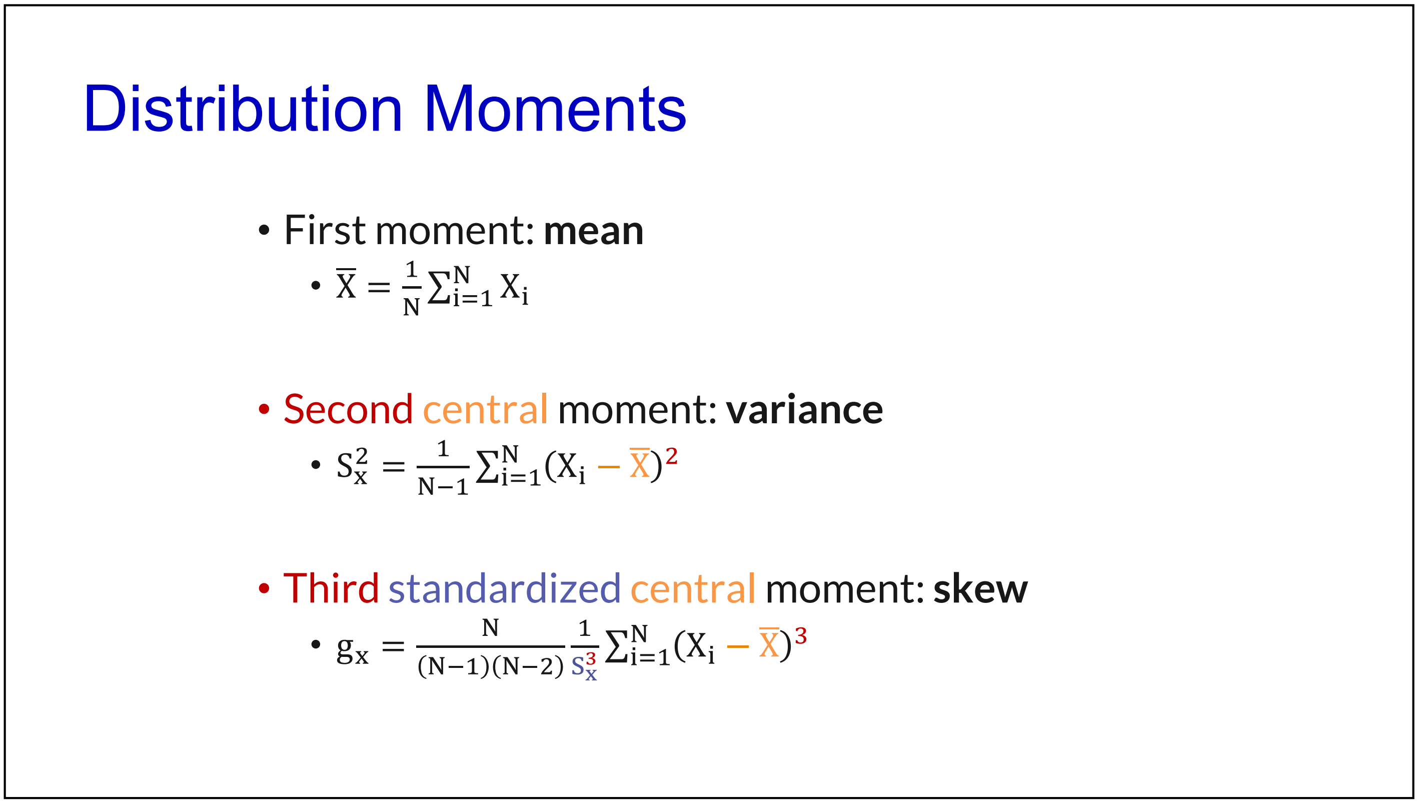
$\gamma < 0$
$\gamma = 0$
$\gamma > 0$

Normal Distribution

The skew coefficient again uses distances from the sample mean, but cubes them, which reestablishes the sign (positive or negative) and exaggerates the value. Values far from the mean, above or below, have more influence on the statistic than values close to the mean. More values far above the mean than far below the mean will produce a positive sum, and more values far below will produce a negative sum.

N-1 and N-2 are again corrections for bias. But with a single N on top and two Ns on the bottom, this is still an average cubed distance from the mean. Finally, it is divided by the standard deviation cubed, both normalizing the value and making it dimensionless (unitless).

A symmetrical sample will have a skew of zero, because the distances above and below the mean are balanced. Values much larger than the mean, producing a positive skew, show as a long upper tail and short lower tail. Values much smaller than the mean, producing a negative skew, show as a long lower tail and short upper tail.

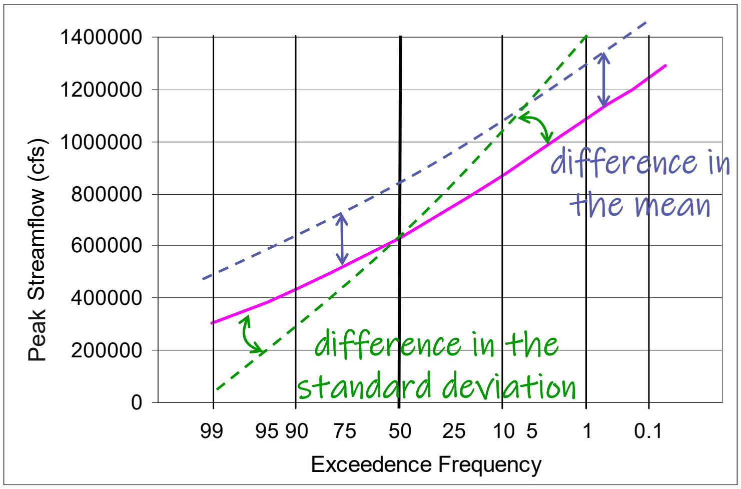


Distribution Moments

- First moment: **mean**
  - $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$
- Second central moment: **variance**
  - $S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$
- Third standardized central moment: **skew**
  - $g_x = \frac{N}{(N-1)(N-2)}\frac{1}{S_x^3}\sum_{i=1}^{N}(X_i - \overline{X})^3$

Distribution Parameters

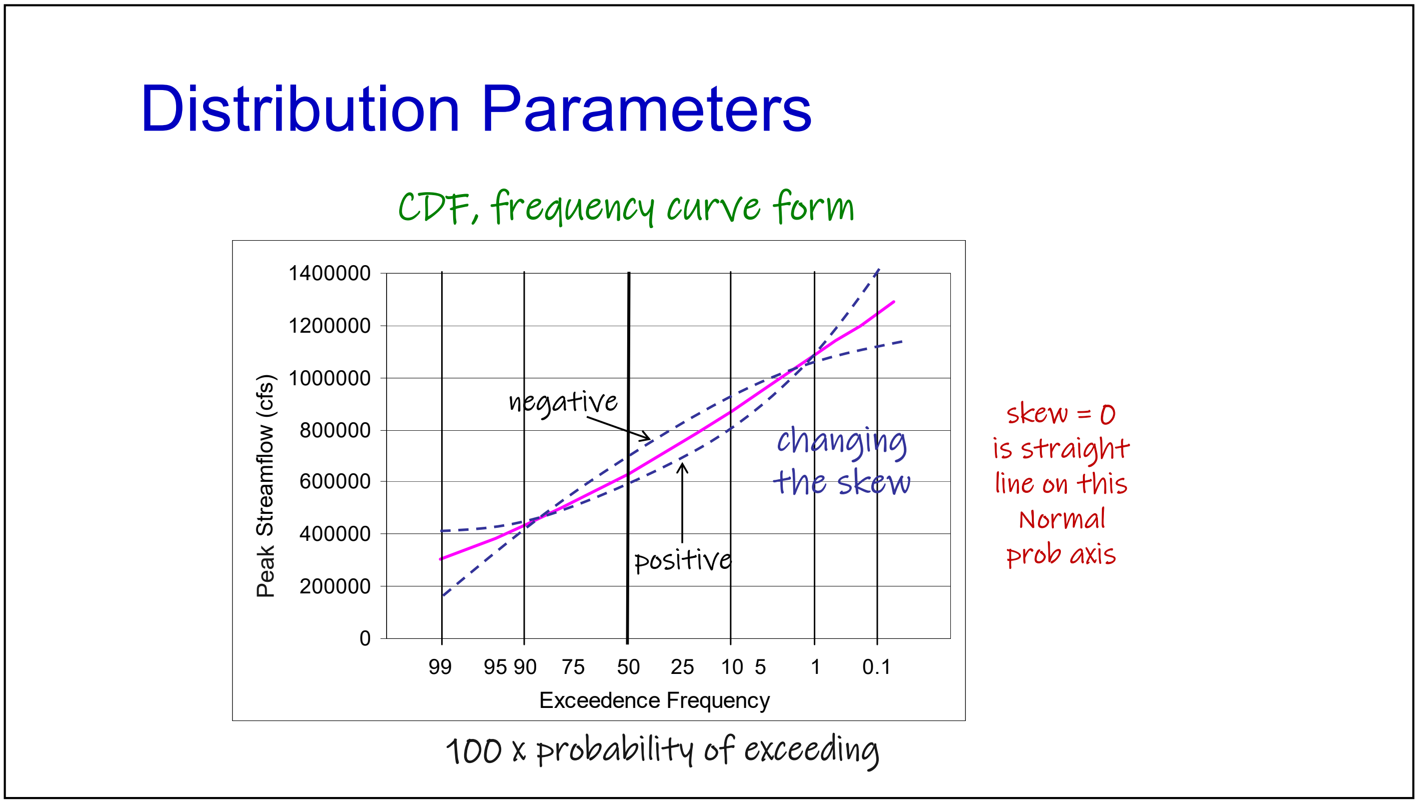
CDF, frequency curve form



100 x probability of exceeding

Earlier slides showed the sample statistics, and the parameters they estimate, on the PDF form of the distribution or sample. This figure shows them on the frequency curve form, which is the CDF with the axes switched to put the variable on the vertical and probability or frequency on the horizontal.

A change in the mean moves the frequency up or down. A change in the standard deviation changes the slope of the frequency curve. A larger standard deviation produces a steeper slope that spans a wider range of the variable on the vertical axis. A smaller standard deviation thus produces a smaller slope.



Distribution Parameters

CDF, frequency curve form

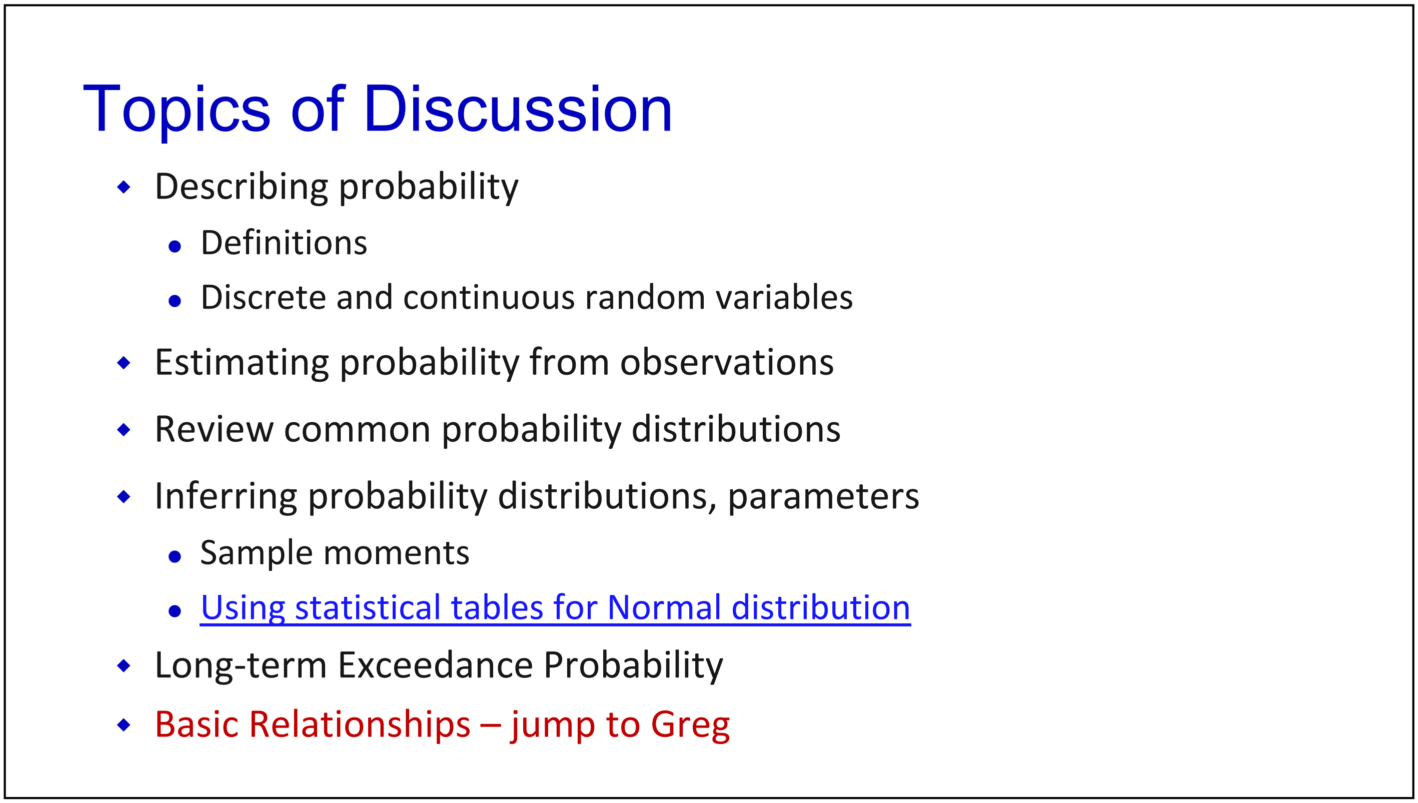
100 x probability of exceeding

skew = 0
is straight
line on this
Normal
prob axis

Earlier slides showed the sample statistics, and the parameters they estimate, on the PDF form of the distribution or sample. This figure shows them on the frequency curve form, which is the CDF with the axes switched to put the variable on the vertical and probability or frequency on the horizontal.

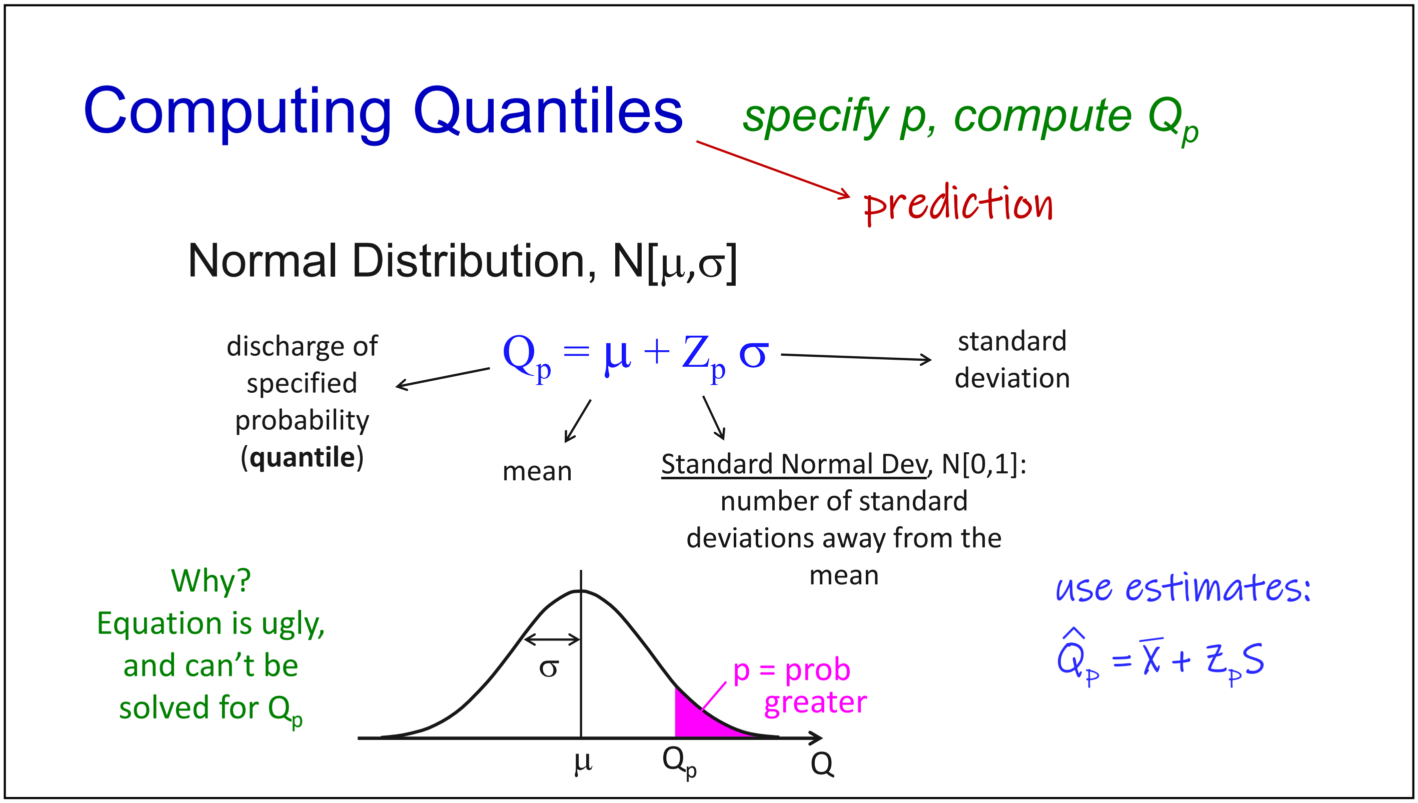
A Normal distribution has a zero skew, and plots as a straight line on a Normal probability axis. A positive skew produces an upward curvature, as the long upper tail reaches higher vertically on the right. A negative skew produces a downward curvature, as the long lower tail reaches downward on the left, and the short upper tail pulls downward on the right.

As a mnemonic, positive skew is happy and produces a smiling upper curvature, and negative skew is sad and produces a frowning downward curvature.



Topics of Discussion

- Describing probability
  - Definitions
  - Discrete and continuous random variables
- Estimating probability from observations
- Review common probability distributions
- Inferring probability distributions, parameters
  - Sample moments
  - Using statistical tables for Normal distribution
- Long-term Exceedance Probability
- Basic Relationships – jump to Greg



Computing Quantiles

specify p, compute Q_p → prediction

Normal Distribution, N[μ,σ]

$$Q_p = \mu + Z_p\,\sigma$$

discharge of specified probability (**quantile**)

mean

Standard Normal Dev, N[0,1]: number of standard deviations away from the mean

standard deviation

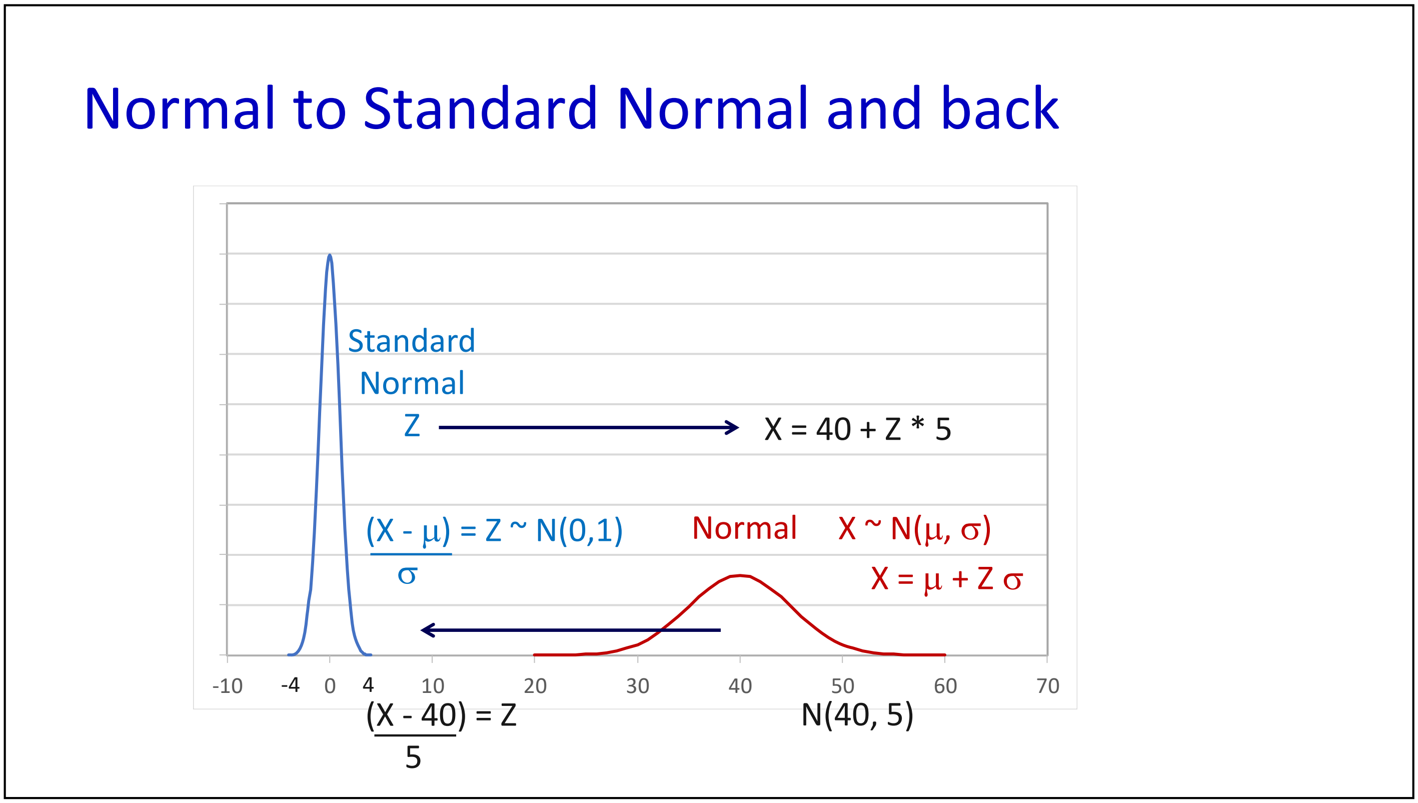
Why? Equation is ugly, and can't be solved for Q_p

use estimates:

$$\hat{Q}_p = \bar{X} + z_p S$$

The Normal distribution has an equation that is hard to work with, and so we generally work with the Standard Normal distribution that is fully tabulated in all statistics text books. The Standard Normal distribution has mean of 0 and standard deviation of 1, and is referred to as Z. Zp is the value for a given probability p.

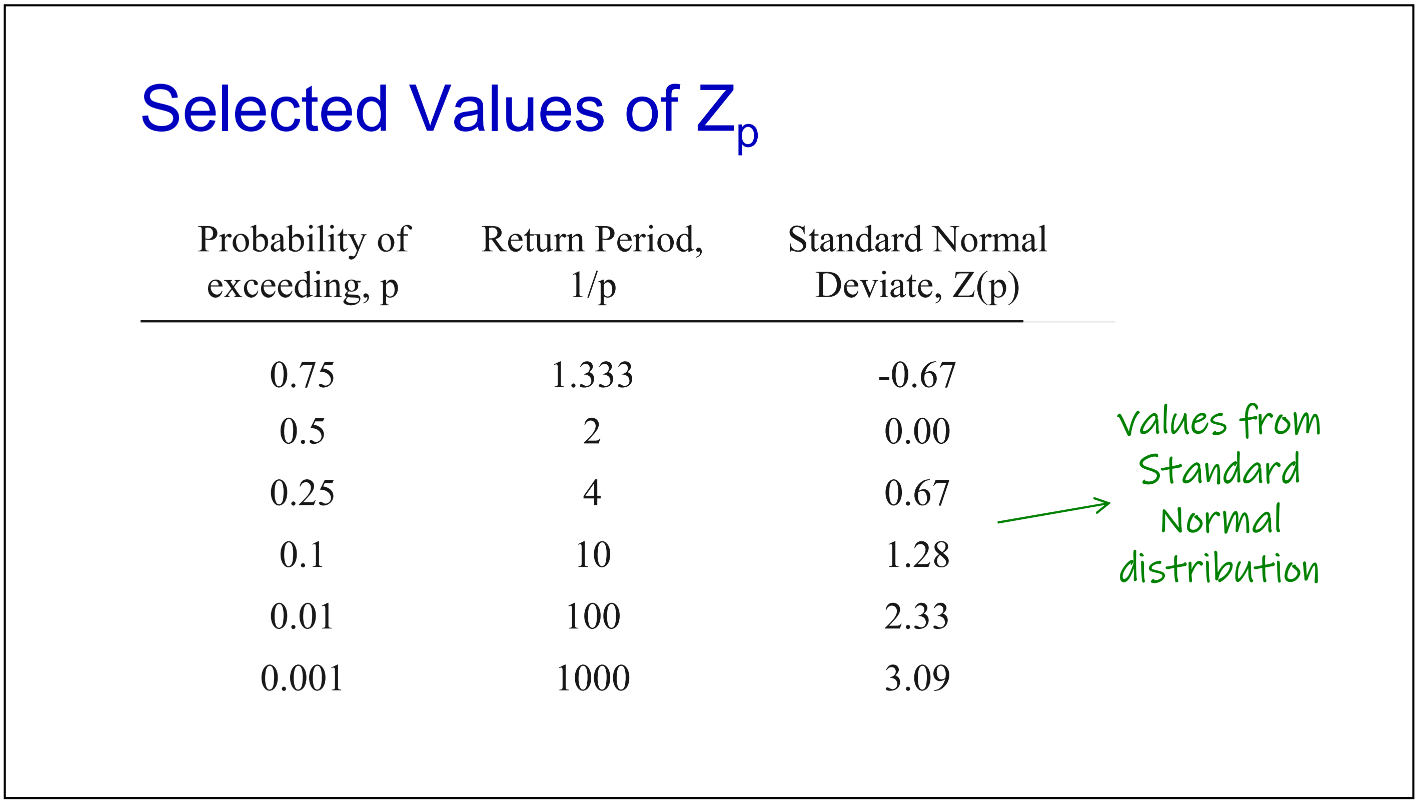
To produce a quantile from the Normal distribution for probability p, we first look up the value Zp from the Standard Normal, then multiply it by our standard deviation and add our mean. The Z value can be interpreted as the number of standard deviations above or below the mean.



Normal to Standard Normal and back

This slides shows the transformation between Normal and Standard Normal distributions. Since Standard Normal has mean of zero and standard deviation of one, we produce a Standard Normal variate by subtracting the mean and the dividing by the standard deviation. We transform back to the Normal distribution by multiplying by the standard deviation and adding back the mean.

NOTE, the area under a PDF is 1.0, and so the two PDF in this image SHOULD have the same area. They do not, to make an easier image to read, but note that the red PDF should be much lower.

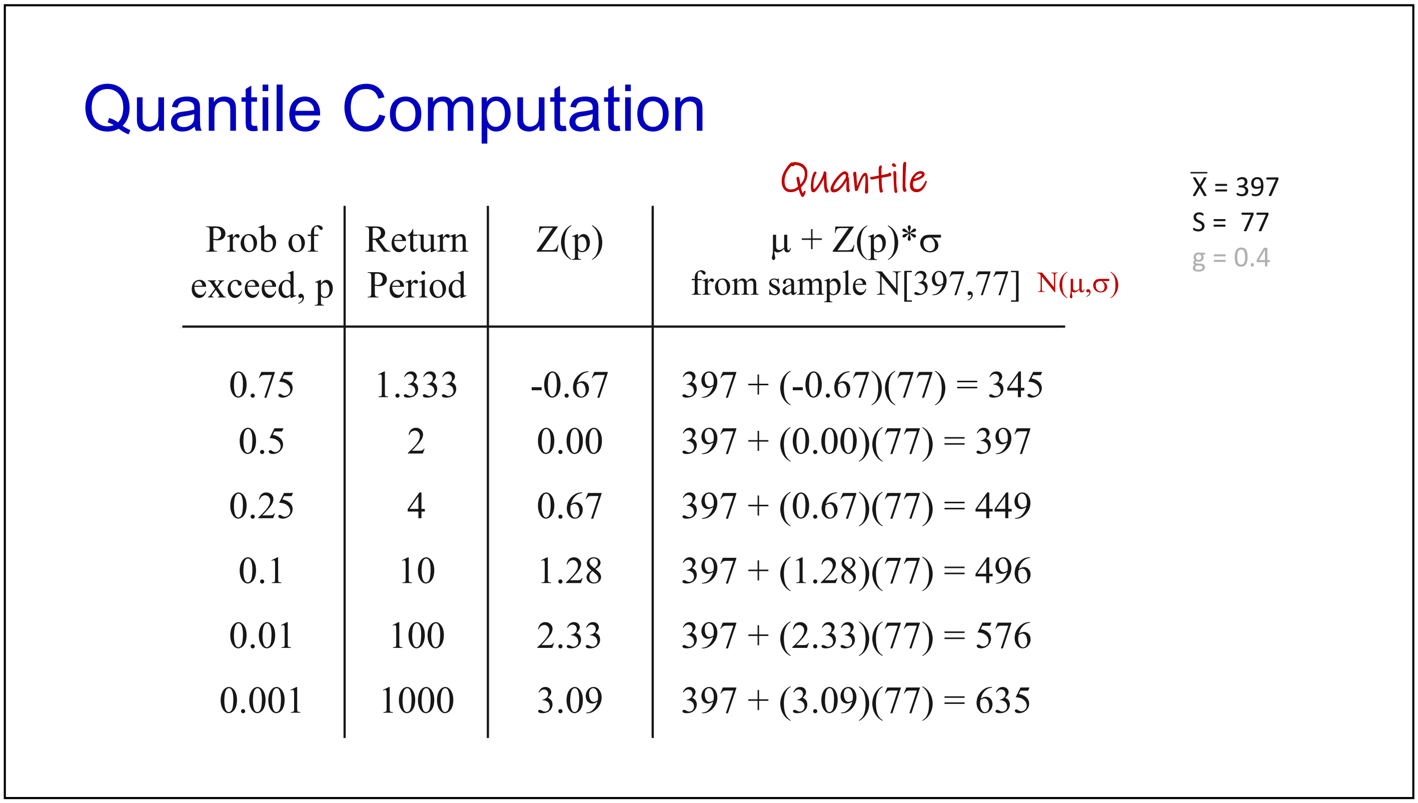


Selected Values of Z_p

Probability of exceeding, p	Return Period, 1/p	Standard Normal Deviate, Z(p)
0.75	1.333	-0.67
0.5	2	0.00
0.25	4	0.67
0.1	10	1.28
0.01	100	2.33
0.001	1000	3.09

values from Standard Normal distribution

Here we have several exceedance probabilities, the associated return periods 1/p, and the Zp value.



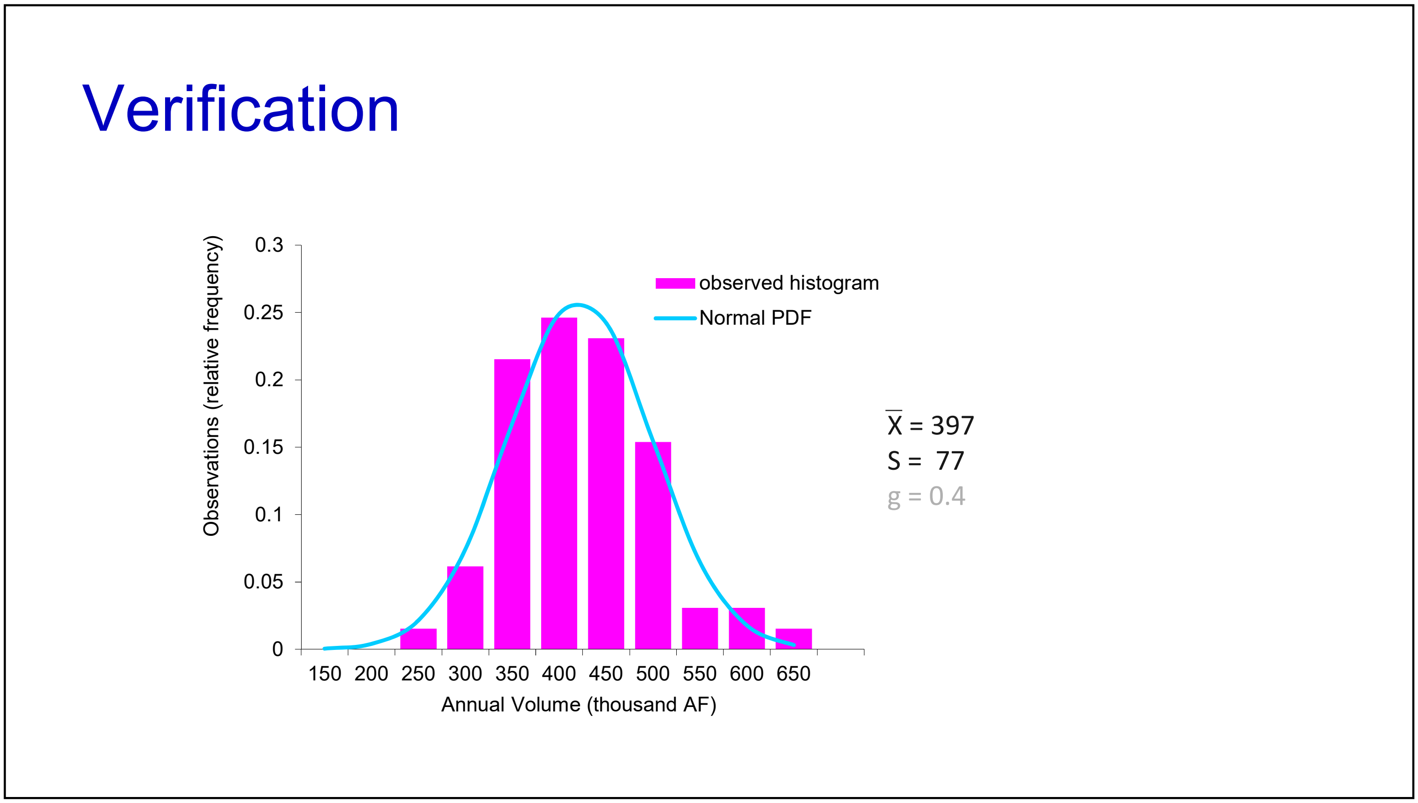
Quantile Computation

Prob of exceed, p	Return Period	Z(p)	Quantile $\mu + Z(p)*\sigma$ from sample N[397,77] N(μ,σ)
0.75	1.333	-0.67	397 + (-0.67)(77) = 345
0.5	2	0.00	397 + (0.00)(77) = 397
0.25	4	0.67	397 + (0.67)(77) = 449
0.1	10	1.28	397 + (1.28)(77) = 496
0.01	100	2.33	397 + (2.33)(77) = 576
0.001	1000	3.09	397 + (3.09)(77) = 635

$\overline{X}$ = 397
S = 77
g = 0.4

A fourth column is added showing the transformation from the Standard Normal back to the Normal distribution using the computed sample mean and standard deviation. Mean + SD * Zp produces the quantiles of the distribution, given the sample statistics shown.

Sample statistics computed as sample mean Xbar = 397, sample standard deviation S = 77 and sample skew coefficient g = 0.4. Note, the skew coefficient is not a parameter of the Normal distribution, which has a skew coefficient of zero, and so was not used here.



Verification

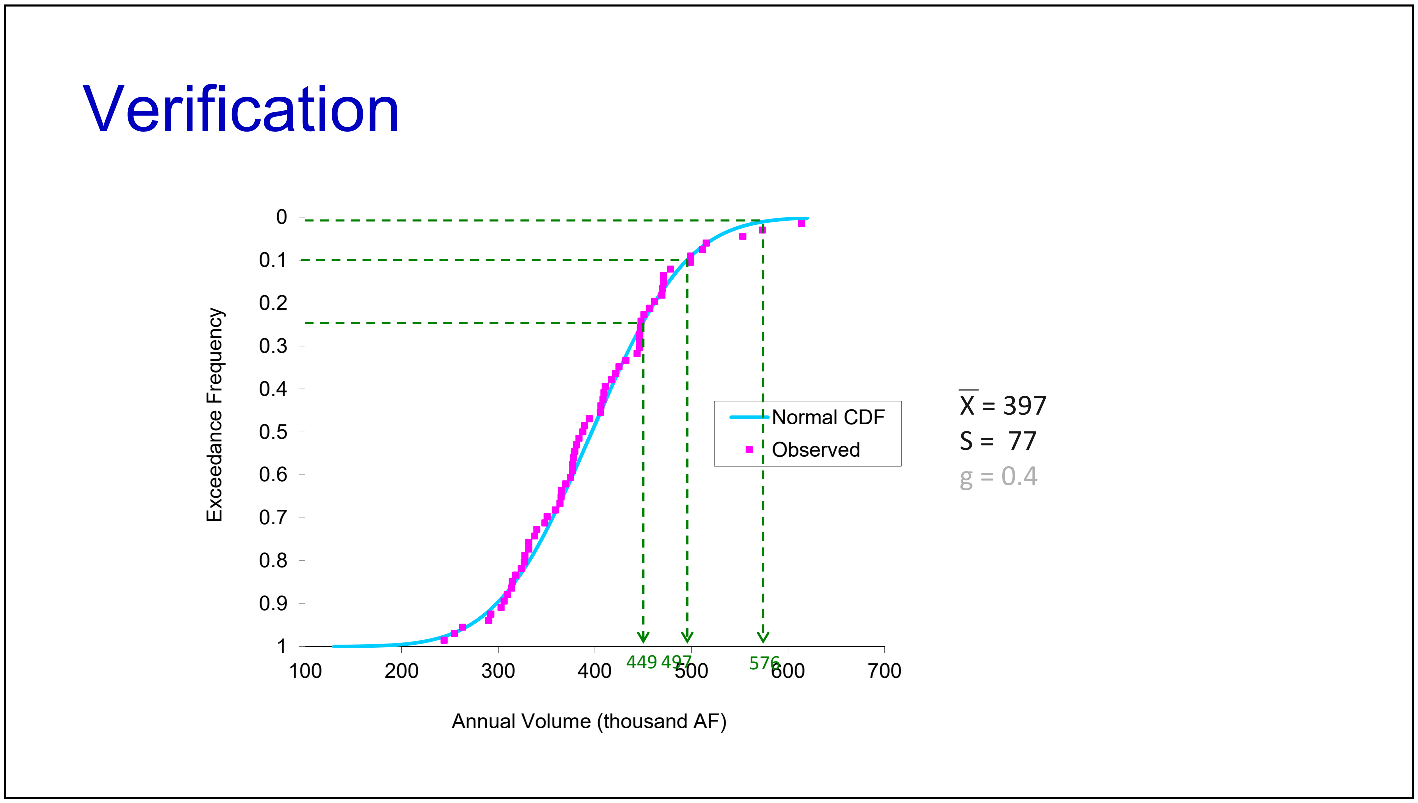
$\overline{X} = 397$

$S = 77$

$g = 0.4$

The blue PDF is the computed Normal Distribution, compared to the sample histogram. How well does the distribution agree with the data?

Sample statistics computed as sample mean Xbar = 397, sample standard deviation S = 77 and sample skew coefficient g = 0.4. Note, the skew coefficient is not a parameter of the Normal distribution, which has a skew coefficient of zero, and so was not used here.



Verification

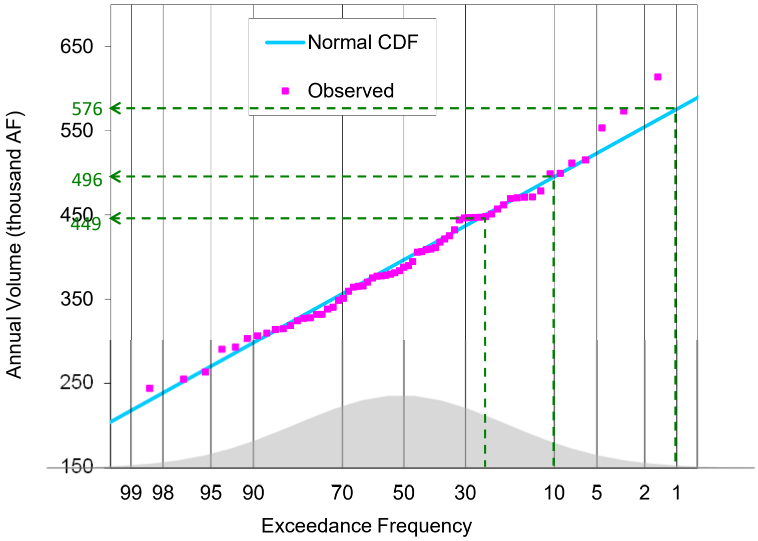
$\overline{X} = 397$

$S = 77$

$g = 0.4$

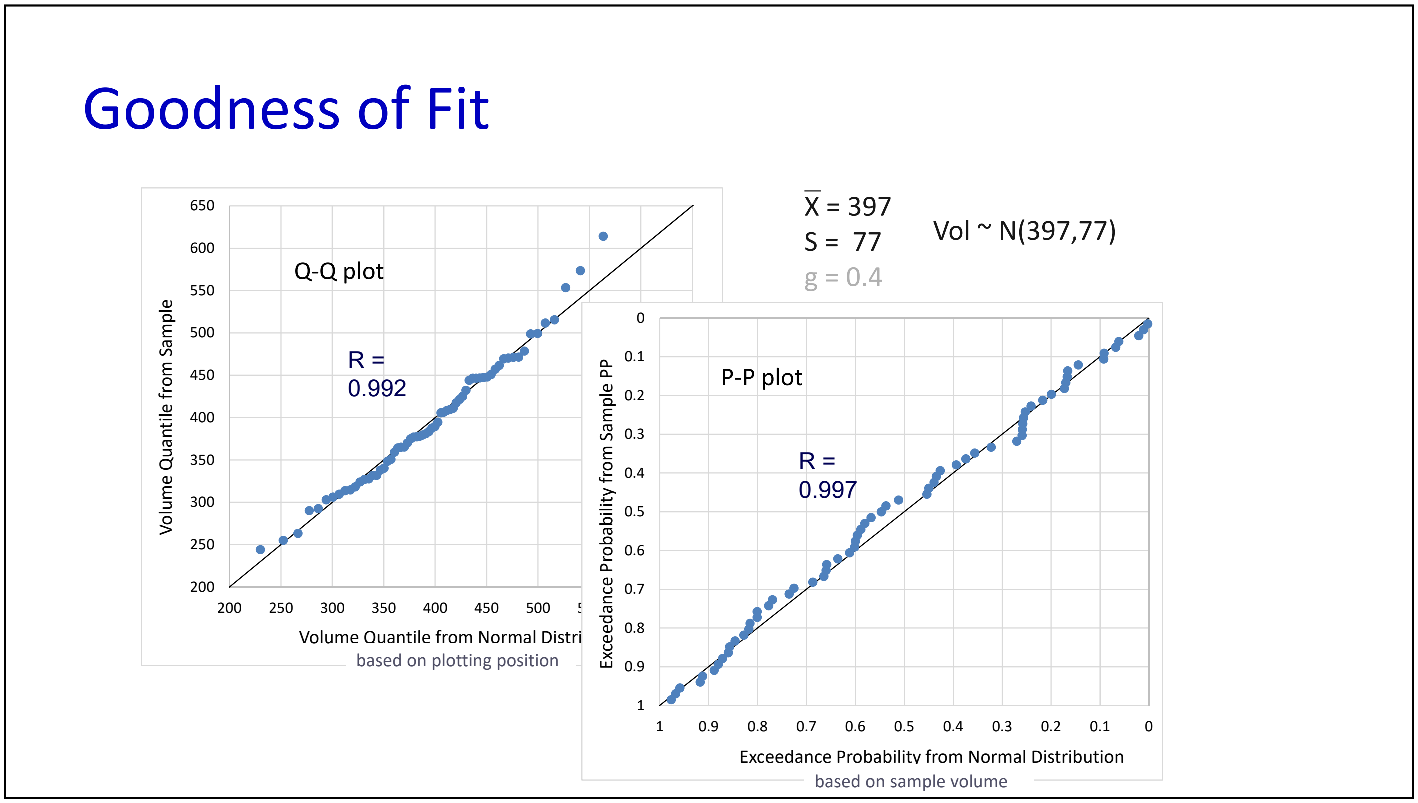
This is the CDF form, plotted with the sample values against median plotting positions. How well does the distribution agree with the data?

Verification



$\bar{X} = 397$
$S = 77$
$g = 0.4$

This is the frequency curve form, plotted with sample values versus plotting positions. How well does the distribution agree with the data? This seems a good fit except the highest three points, as expected when the original histogram showed a longer upper tail than lower.



Goodness of Fit

$\overline{X}$ = 397

S = 77

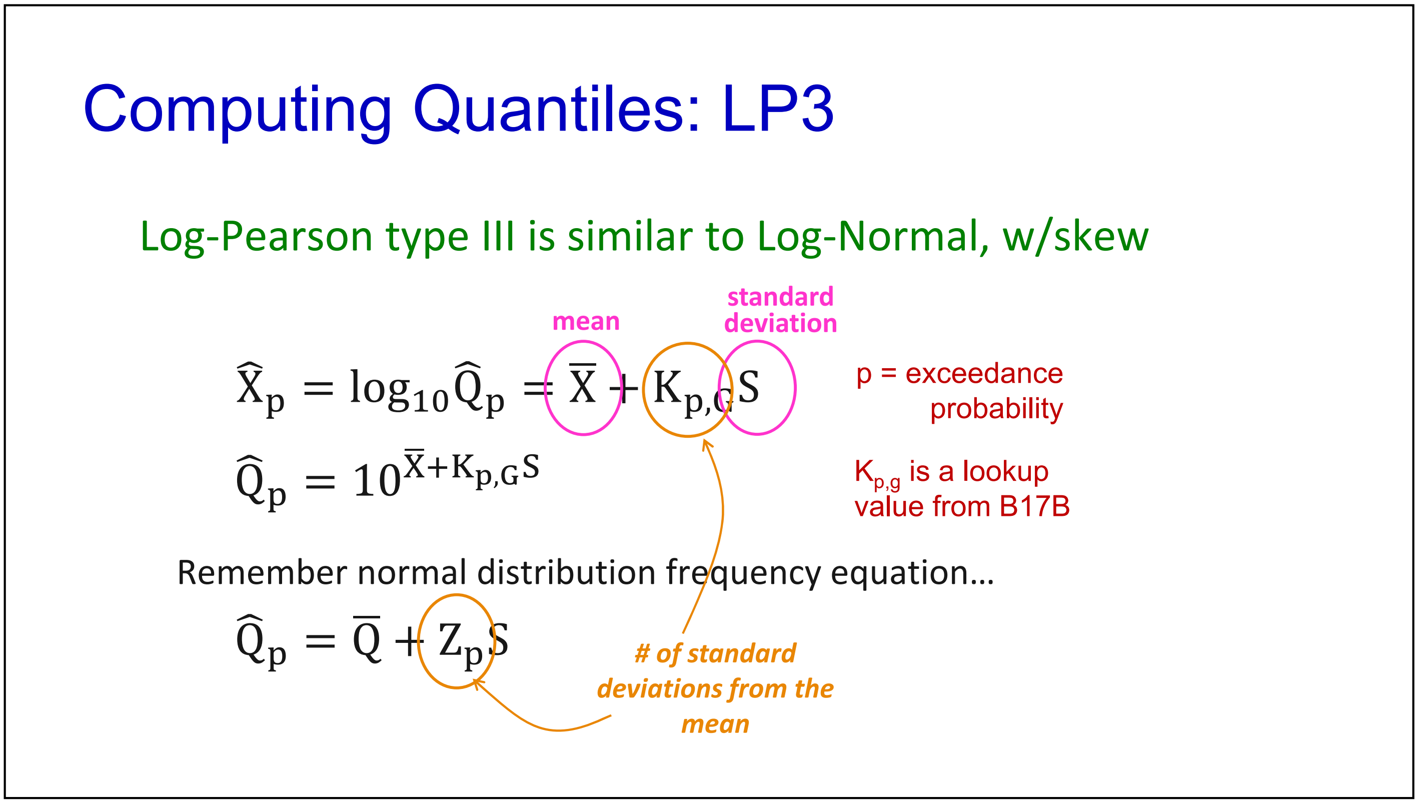
g = 0.4

Vol ~ N(397,77)

Another way to assess how well the distribution fits is with Goodness of Fit tests. These two plots compare the sample values to what a Normal distribution would have expected the sample values to be.

The QQ plot shows the sample data on the vertical axis, paired with the value of the Normal distribution that goes with its plotting position on the horizontal axis. The closer to the straight 1:1 line, the better the fit. We often compute a correlation coefficient for the pairs, with a value closer to 1.0 showing better fit, to have a numerical value of fit to work with.

The PP plot is similar, with the plotting positions for any 65-member sample on the vertical, and the cumulative probabilities of the fitted Normal distribution for each of the sample members on the horizontal. Again, a correlation coefficient can be used to produce a numerical value of the fit.



Computing Quantiles: LP3

Log-Pearson type III is similar to Log-Normal, w/skew

mean

standard deviation

$$\hat{X}_p = \log_{10}\hat{Q}_p = \bar{X} + K_{p,G}S$$

$$\hat{Q}_p = 10^{\bar{X}+K_{p,G}S}$$

p = exceedance probability

$K_{p,g}$ is a lookup value from B17B

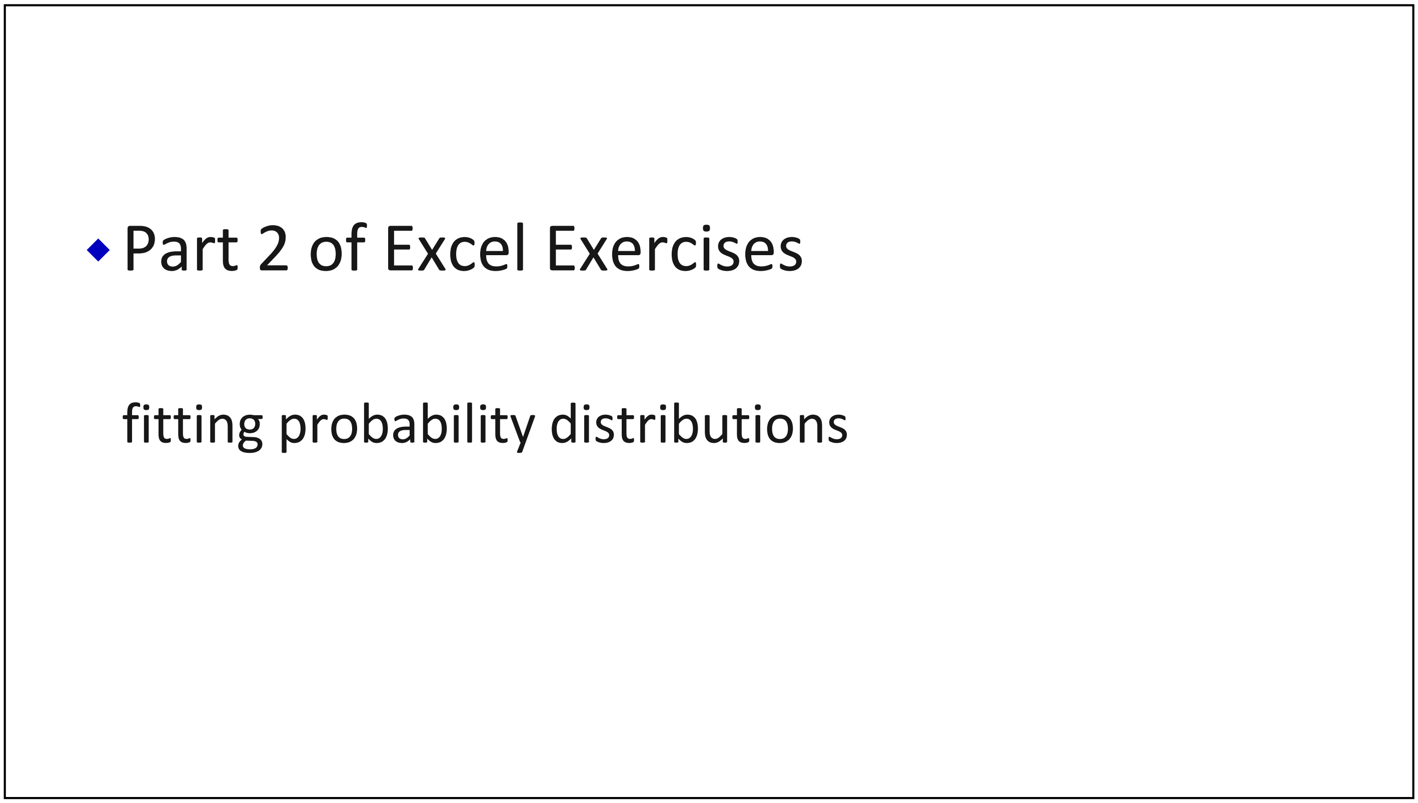
Remember normal distribution frequency equation…

$$\hat{Q}_p = \bar{Q} + Z_p S$$

# of standard deviations from the mean

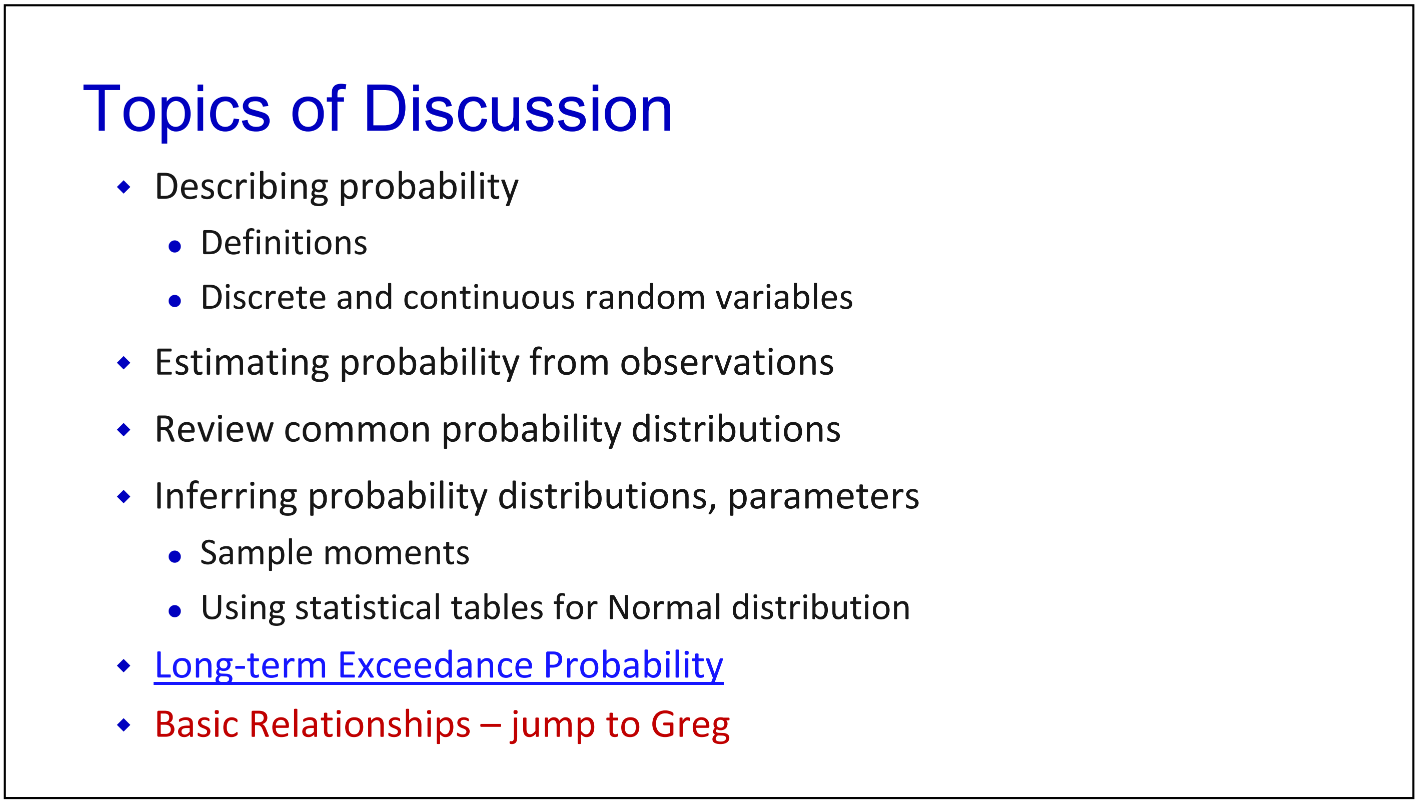
The LogPearson III curve is estimated by computing its quantiles. Recall the estimation of the Normal distribution curve, defined as the mean plus Zp standard deviations, where Zp is the “standard normal deviate” for probability p.

The LogPearson III curve is estimated in the same way, except that the deviate K is based on both probability p and skew g.



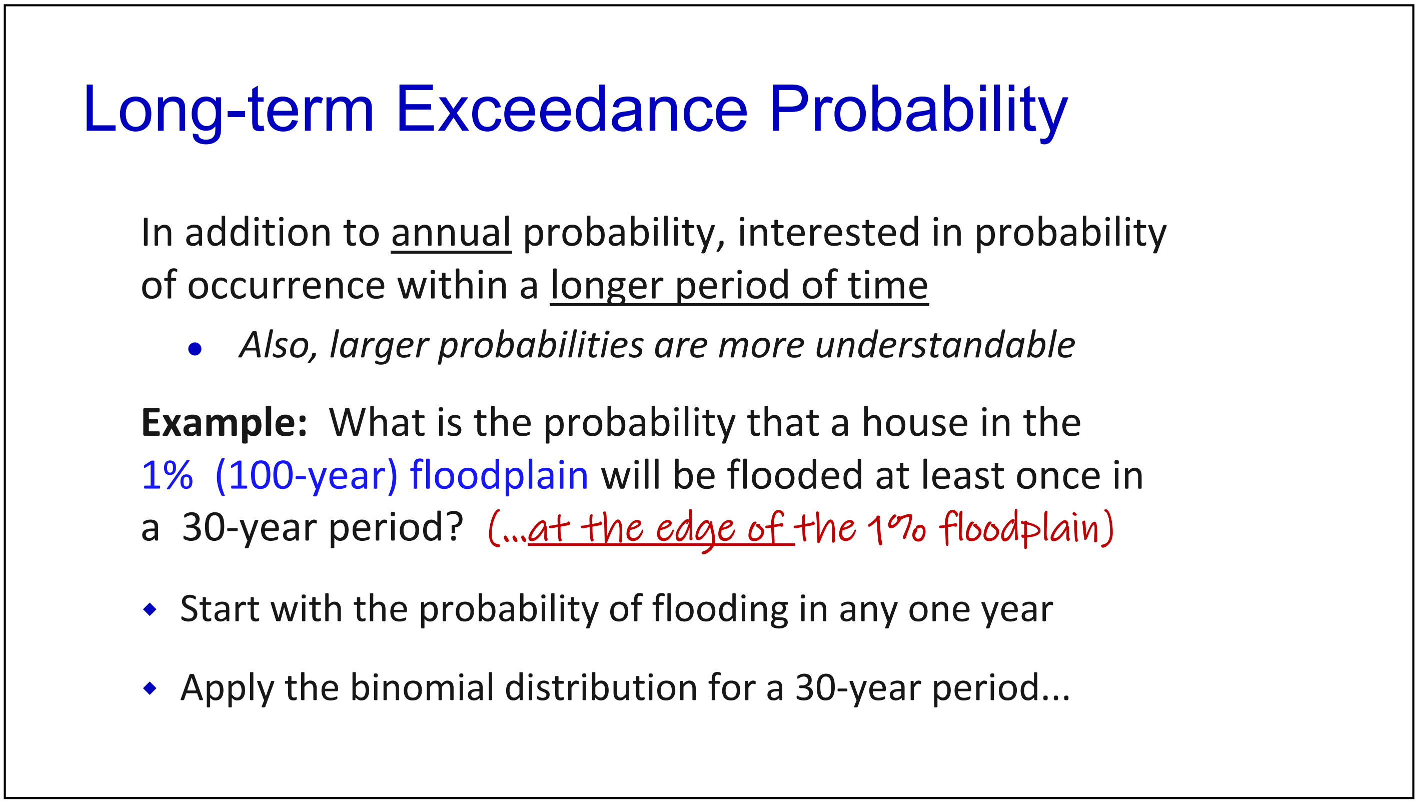
- Part 2 of Excel Exercises

fitting probability distributions



Topics of Discussion

- Describing probability
  - Definitions
  - Discrete and continuous random variables
- Estimating probability from observations
- Review common probability distributions
- Inferring probability distributions, parameters
  - Sample moments
  - Using statistical tables for Normal distribution
- Long-term Exceedance Probability
- Basic Relationships – jump to Greg



Long-term Exceedance Probability

In addition to annual probability, interested in probability of occurrence within a longer period of time

- *Also, larger probabilities are more understandable*

Example: What is the probability that a house in the 1% (100-year) floodplain will be flooded at least once in a 30-year period? (...at the edge of the 1% floodplain)

- Start with the probability of flooding in any one year
- Apply the binomial distribution for a 30-year period...

Binomial Distribution

assumes independence of trials

$$P[\text{x successes in N trials}] = \begin{bmatrix} N \\ x \end{bmatrix} p^x (1-p)^{N-x}$$

where: $\begin{bmatrix} N \\ x \end{bmatrix} = \dfrac{N!}{x!\,(N-x)!}$ = number of ways to get x successes in N trials

p = probability of success in a single trial

Example: $P[\text{2 successes in 3 trials} \mid p = 0.4]$

$$= \frac{3!}{2!\,(3-2)!}\,(0.4)^2\,(1-0.4)^{3-2}$$

$$= 3\ (0.4)^2 (0.6)^1$$

$$= 0.288$$

1. 1-1-0
2. 1-0-1
3. 0-1-1

probability of 2 successes

probability of 1 failure

Long-term Exceedance Probability

x=0

- Prob of **no** exceedances in *N* years is
  $(1 - p)^N$ *(x=0, where p = probability of exceedance)*

- Prob of one or more exceedances in *N* years is
  $1 - (1 - p)^N$ *(complement of none is at least 1)* x=1, ..., x=30

➔ Prob of one or more "1%" floods in 30 years is
$1 - (1 - 0.01)^{30} = 0.26$ *because* p = 0.01 *and* N = 30

Answer is:
26%

The risk of 1 or more "10-yr" floods in 30 years is $1 - (1-{}^1/_{10})^{30} = 0.96$ 96%

The risk of 1 or more "100-yr" floods in 30 years is $1 - (1-{}^1/_{100})^{30} = 0.26$ 26%

The risk of 1 or more "500-yr" floods in 30 years is $1 - (1-{}^1/_{500})^{30} = 0.06$ 6%

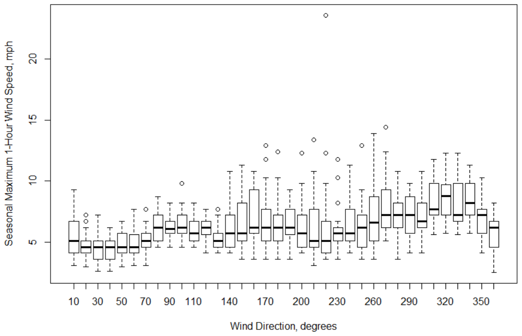
Goals addressed so far….

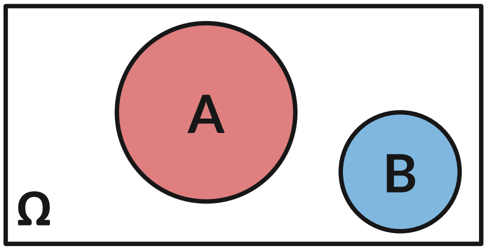
- revisit probability concepts
- estimating probability
  - from understanding the system or from data
- using probability distributions to describe what is random, and what is unknown
- reviewing common probability distributions
- fitting probability distributions to data
- computing subsequent probabilities

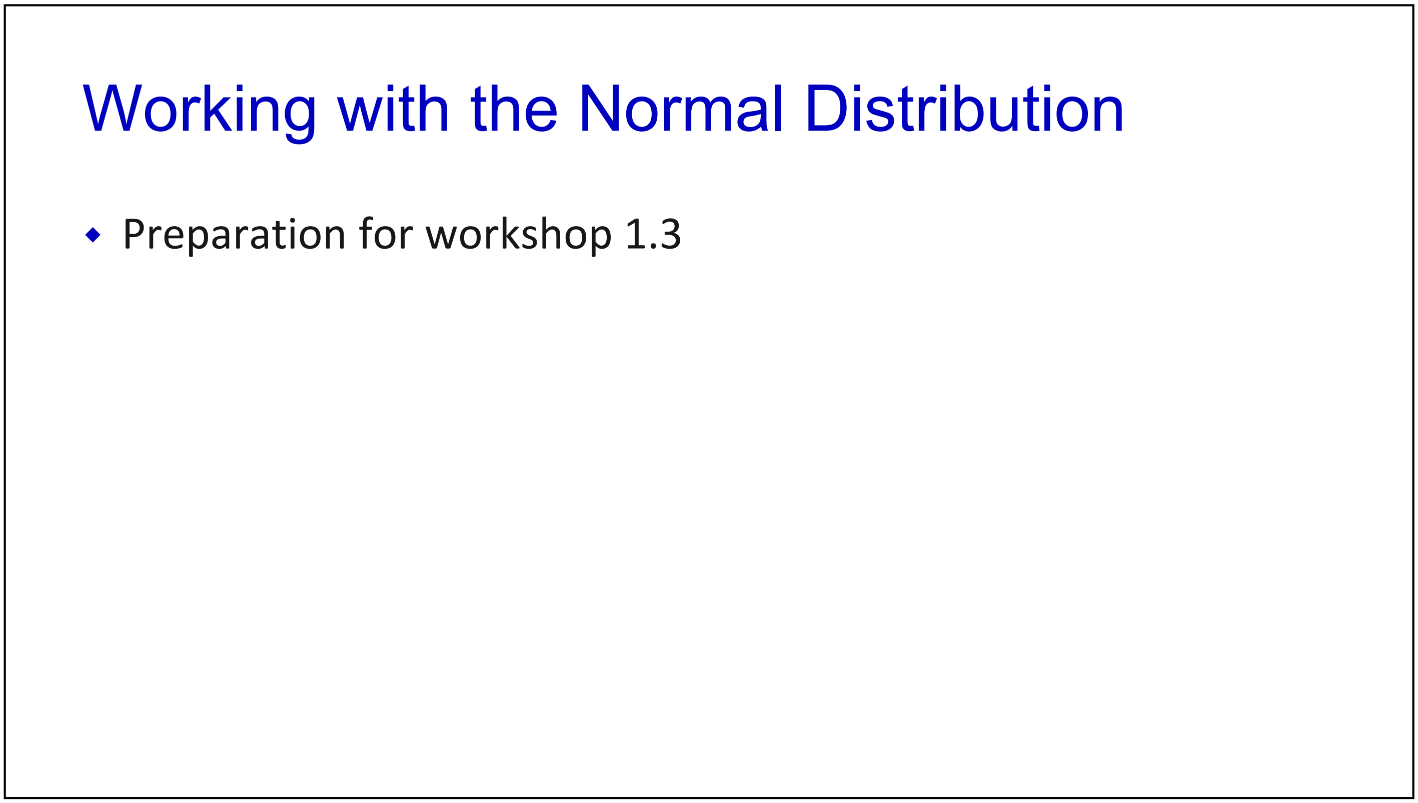
More topics!

Greg Karlovits, HEC

- Evaluating Data
  - Types
  - Numerical Summaries
  - Diagnostics
- Venn Diagrams
  - Events / Axioms of Probability

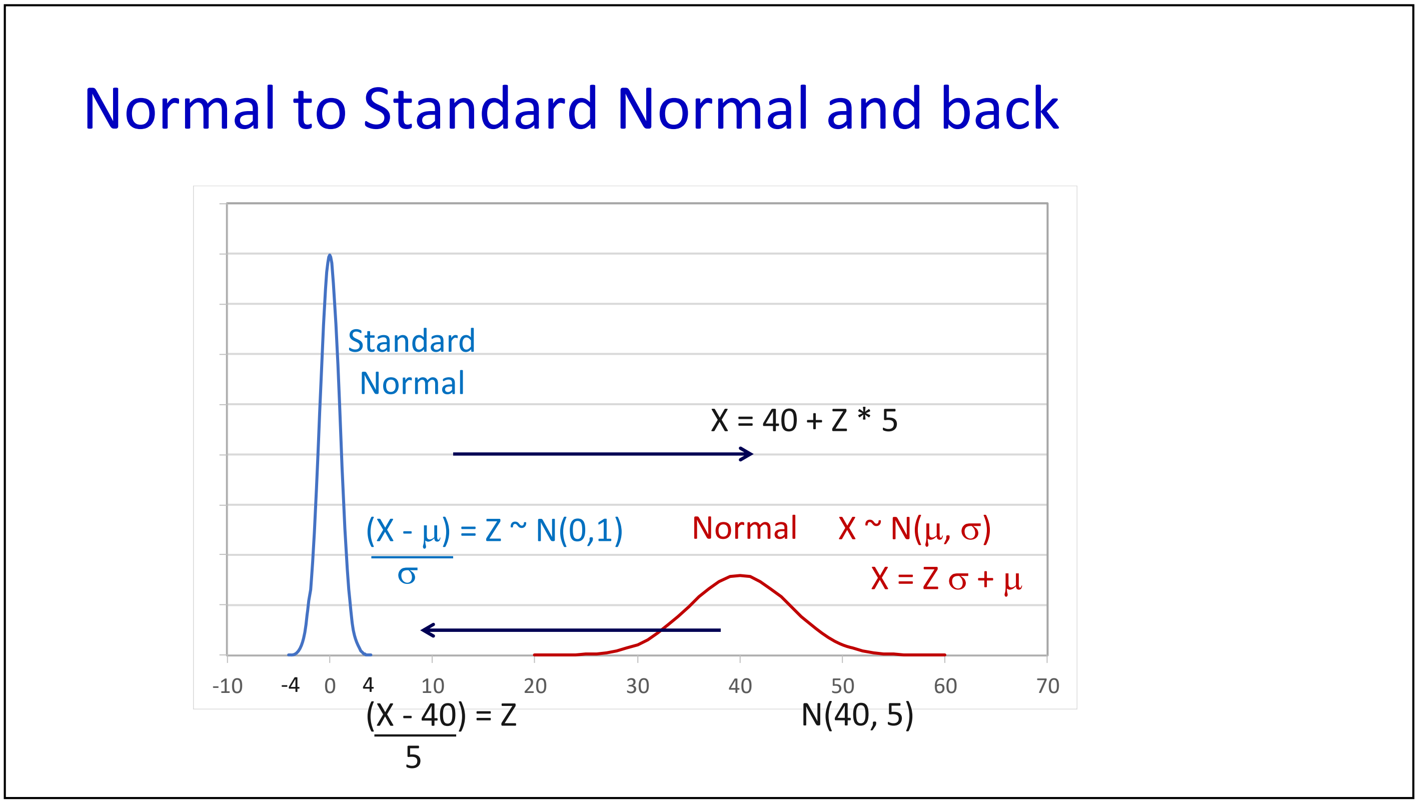






Working with the Normal Distribution

- Preparation for workshop 1.3



Normal to Standard Normal and back

$X = 40 + Z * 5$

$\frac{(X - \mu)}{\sigma} = Z \sim N(0,1)$

$X \sim N(\mu, \sigma)$

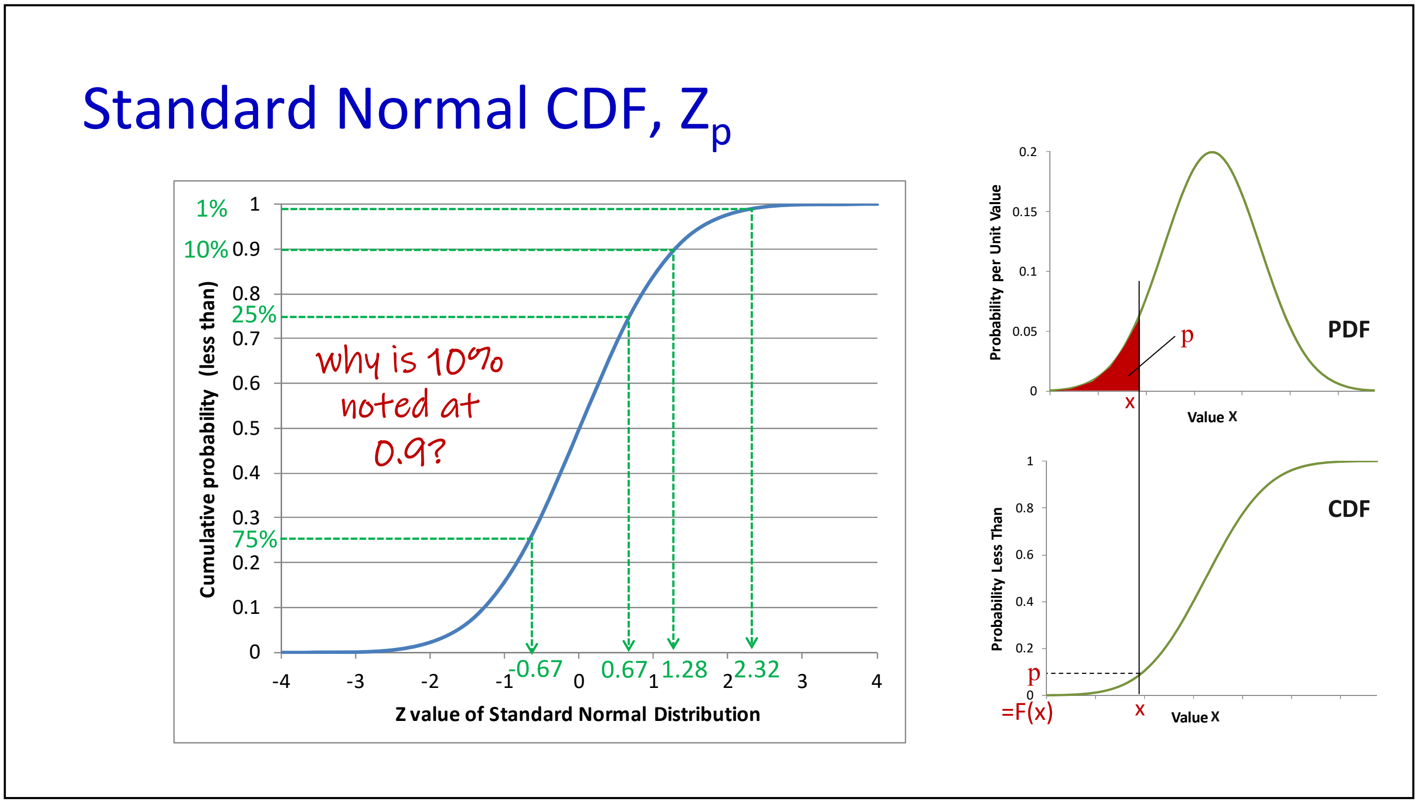
$X = Z\,\sigma + \mu$

$\frac{(X - 40)}{5} = Z$

N(40, 5)

This slides shows the transformation between Normal and Standard Normal distributions. Since Standard Normal has mean of zero and standard deviation of one, we produce a Standard Normal variate by subtracting the mean and the dividing by the standard deviation. We transform back to the Normal distribution by multiplying by the standard deviation and adding back the mean.

NOTE, the area under a PDF is 1.0, and so the two PDF in this image SHOULD have the same area. They do not, to make an easier image to read, but note that the red PDF should be much lower.

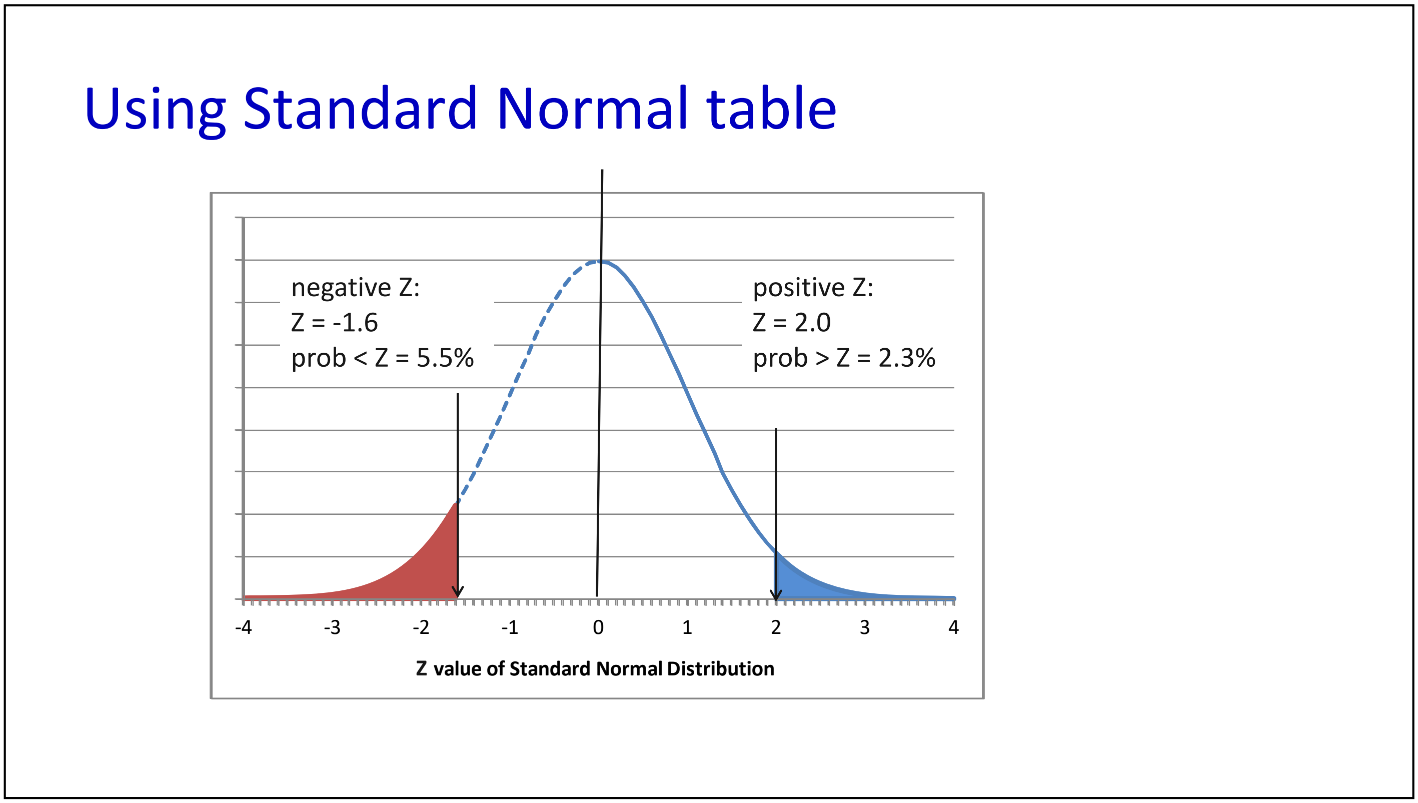


Standard Normal CDF, Z_p

why is 10% noted at 0.9?

This is the CDF of the Standard Normal distribution, showing values that are also tabulated in text books.

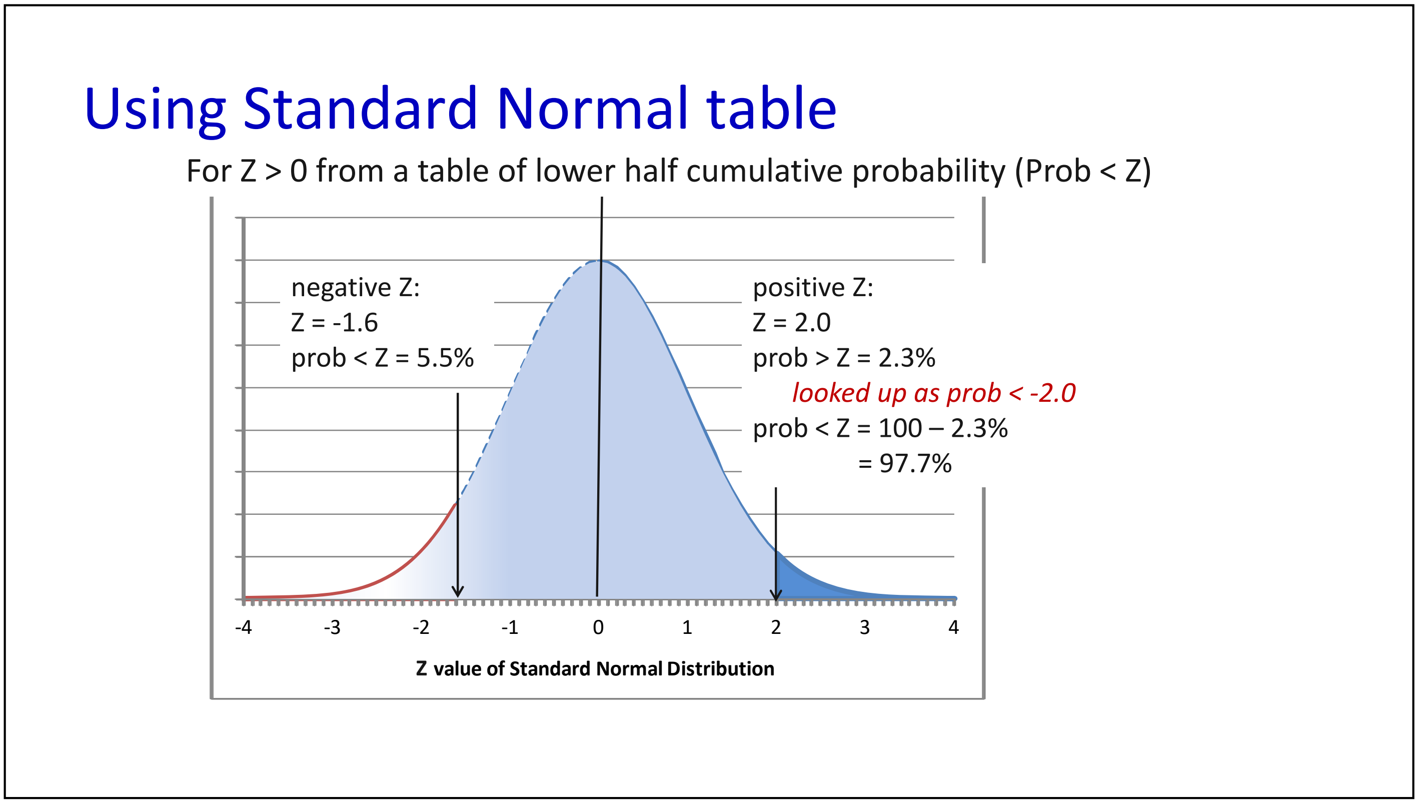
Cumulative probability (probability less than) is on the vertical and the Z variate is on the horizontal. Note that most of the distribution is between -3 and 3, and nearly all is between -4 and 4.



Using Standard Normal table

Because the distribution is symmetrical, most Standard Normal tabulations only show half, and leave it to the user to reverse to result for values on the other half of the distribution.

In the table coming up in this lecture, the negative side of the distribution is tabulated. So the pairing of Z = -1.6 having cumulative probability (P<Z) of 5.5% can be read directly. But to determine the cumulative probability (P<Z) of Z = 2.0, we must transform a value assumed from the upper half of the distribution. The table will tell us that Z = -2.0 has cumulative probability (P<Z) of 2.3%, and we therefore know by symmetry that that Z = 2.0 has exceedance probability (P>Z) of 2.3%.



Using Standard Normal table

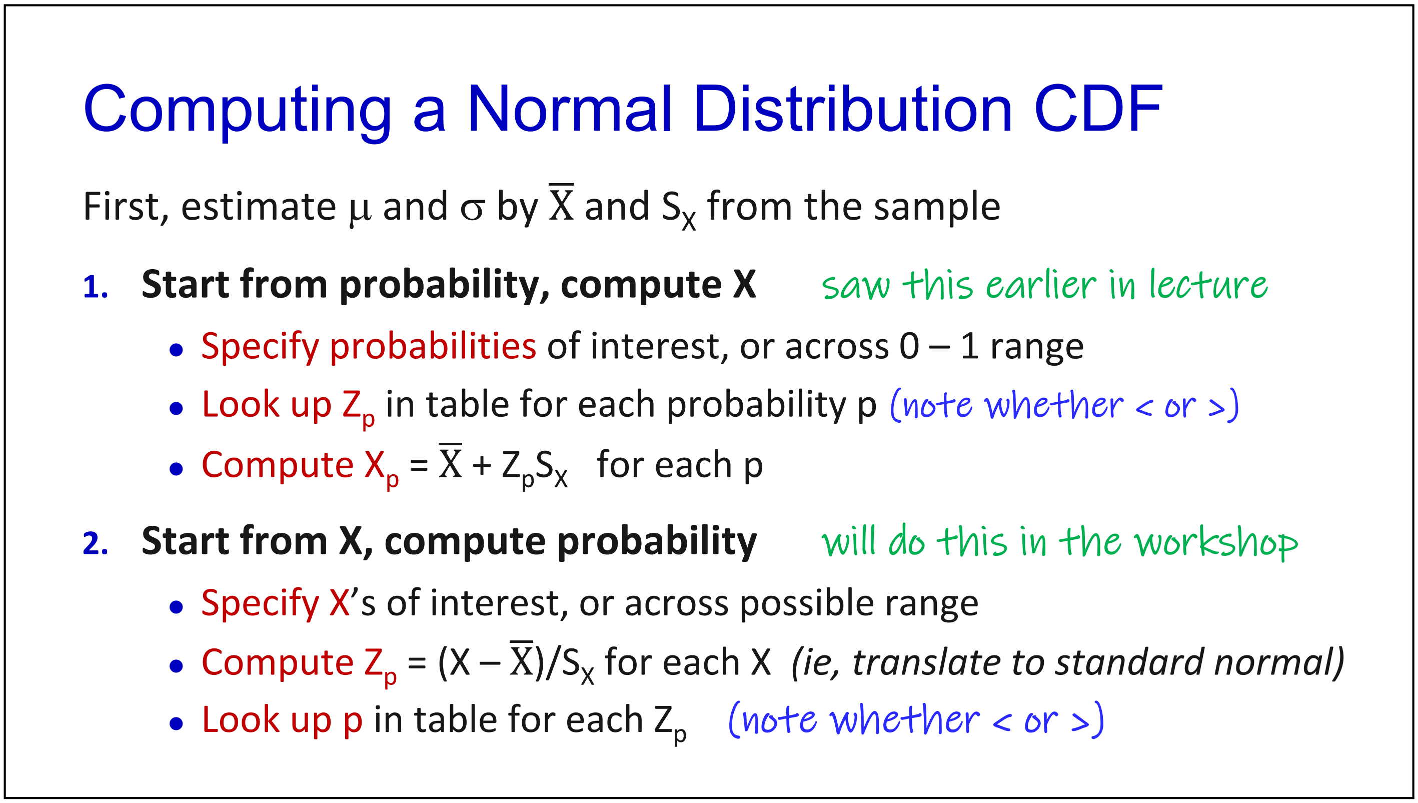
For Z > 0 from a table of lower half cumulative probability (Prob < Z)

negative Z:
Z = -1.6
prob < Z = 5.5%

positive Z:
Z = 2.0
prob > Z = 2.3%
looked up as prob < -2.0
prob < Z = 100 – 2.3%
= 97.7%

Because the distribution is symmetrical, most Standard Normal tabulations only show half, and leave it to the user to reverse to result for values on the other half of the distribution.

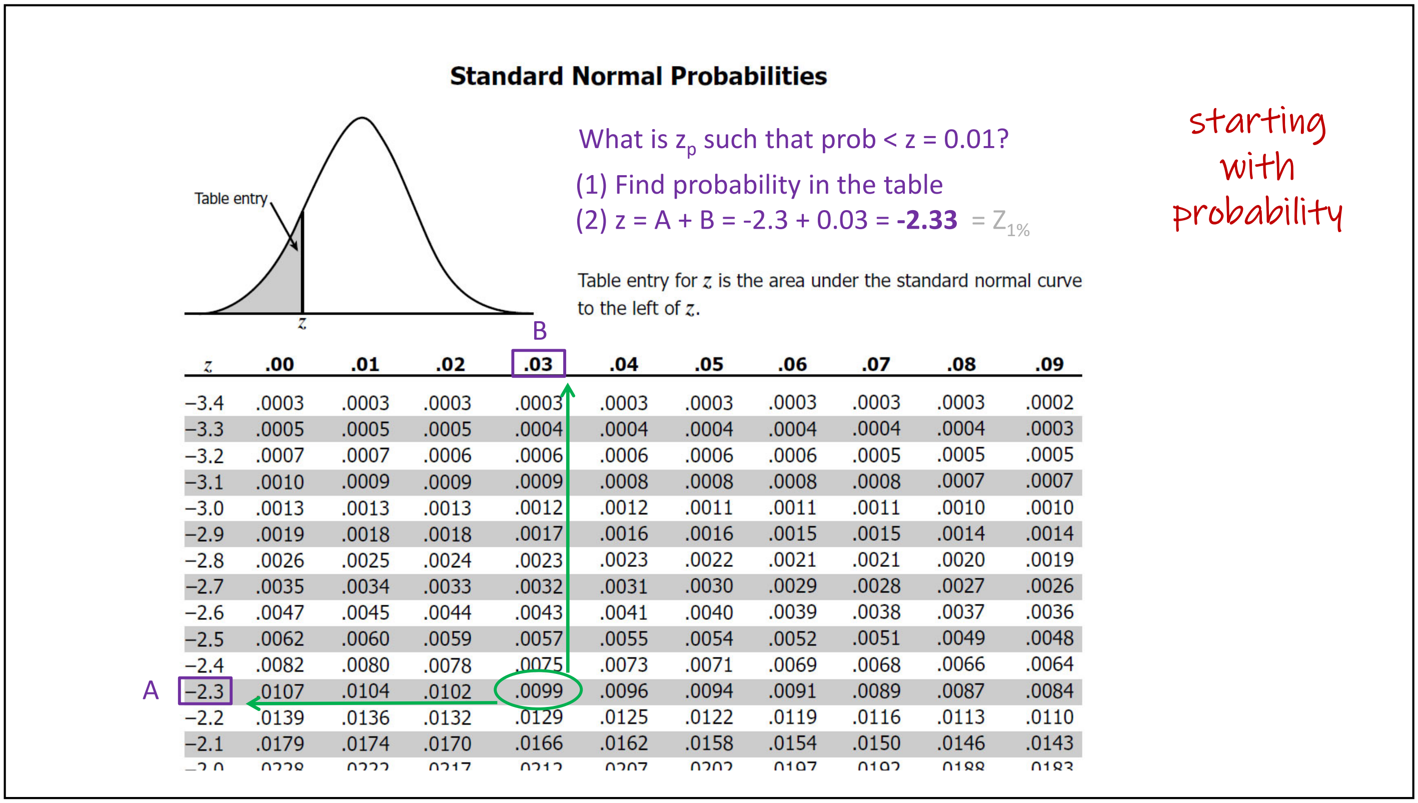
In the table coming up in this lecture, the negative side of the distribution is tabulated. So the pairing of Z = -1.6 having cumulative probability (P<Z) of 5.5% can be read directly. But to determine the cumulative probability (P<Z) of Z = 2.0, we must transform a value assumed from the upper half of the distribution. The table will tell us that Z = -2.0 has cumulative probability (P<Z) of 2.3%, and we therefore know by symmetry that that Z = 2.0 has exceedance probability (P>Z) of 2.3%. We can then compute the probability of Z < 2.0 as 100 – 2.3% = 97.7%.



Computing a Normal Distribution CDF

First, estimate μ and σ by $\bar{X}$ and S_X from the sample

1. **Start from probability, compute X** saw this earlier in lecture
   - Specify probabilities of interest, or across 0 – 1 range
   - Look up Z_p in table for each probability p (note whether < or >)
   - Compute $X_p = \bar{X} + Z_p S_X$ for each p
2. **Start from X, compute probability** will do this in the workshop
   - Specify X's of interest, or across possible range
   - Compute $Z_p = (X - \bar{X})/S_X$ for each X *(ie, translate to standard normal)*
   - Look up p in table for each Z_p (note whether < or >)



Standard Normal Probabilities

What is z_p such that prob < z = 0.01?

(1) Find probability in the table

(2) z = A + B = -2.3 + 0.03 = **-2.33** = $Z_{1\%}$

starting with probability

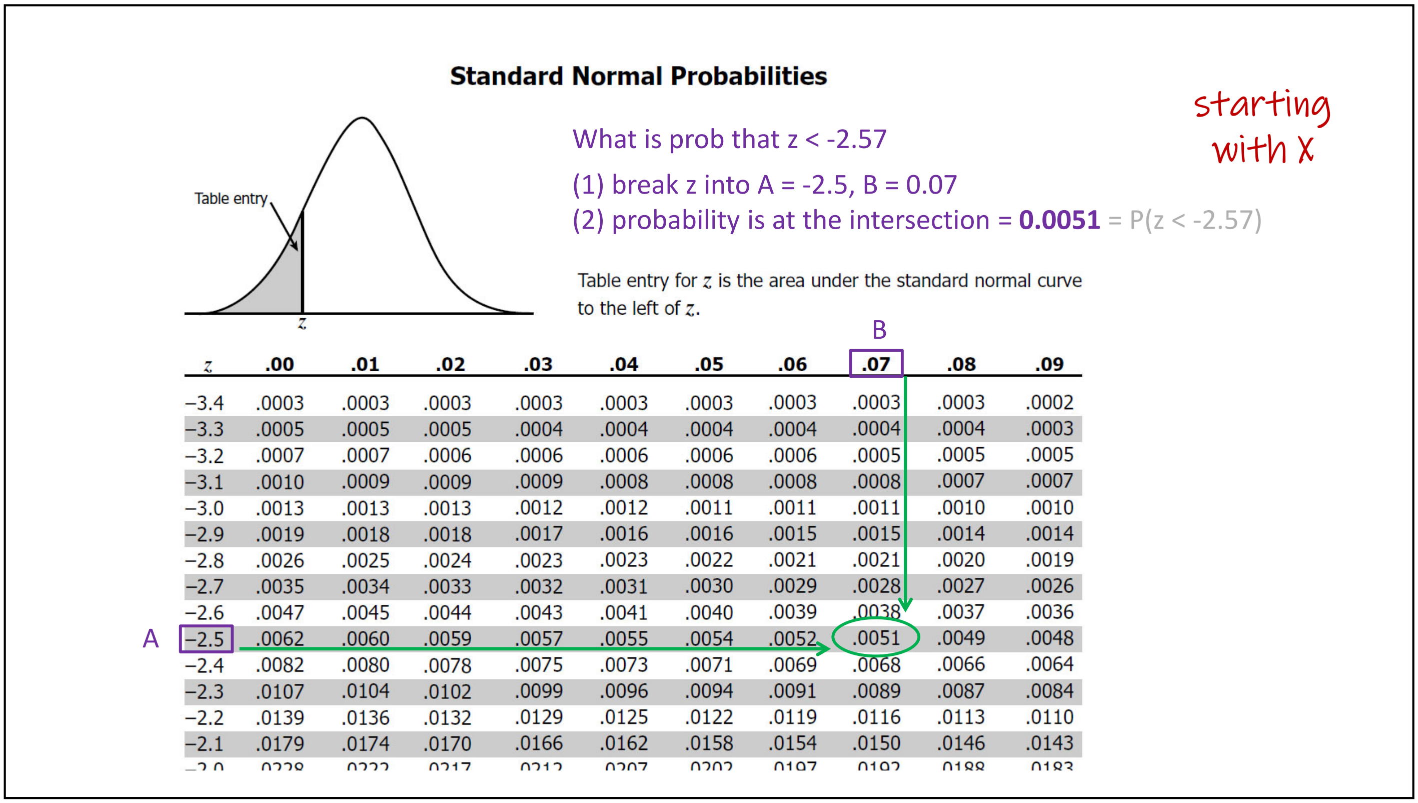
Table entry for *z* is the area under the standard normal curve to the left of *z*.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183

This is a tabulation of the lower half of the Standard Normal distribution, providing cumulative probability P<Z. The value of Z is the headers of the rows and columns, with Z to the first decimal as the row, and the second decimal as the column, and the cumulative probability (P<Z) is the value in the table.

To start with probability and look up Z, find the probability of interest in the table, and read the row and column headers. For (P<Z) of 1% or 0.01, the closest value is .0099, which has Z of -2.3 from the row and .03 from the column for a total of -2.33.

Note symmetry of the distribution and we know Z = 2.33 has exceedance probability (P>Z) of 1%.



Standard Normal Probabilities

starting with X

What is prob that z < -2.57

(1) break z into A = -2.5, B = 0.07

(2) probability is at the intersection = **0.0051** = P(z < -2.57)

Table entry for *z* is the area under the standard normal curve to the left of *z*.

B

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
A −2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183

This is a tabulation of the lower half of the Standard Normal distribution, providing cumulative probability P<Z. The value of Z is the headers of the rows and columns, with Z to the first decimal as the row, and the second decimal as the column, and the cumulative probability (P<Z) is the value in the table.

Thus, for Z = -2.57, we go to the row for -2.5, the column for .07 and read the P<Z = 0.51%

If we actually have Z = 2.57, we look up the probability of Z = -2.57 and assume symmetry. Thus cumulative probability P<Z of Z = 2.57 is 1 – 0.0051 = 0.9949 of 99.49%.