# Assignments

- Assignment 6: Due at beginning of class next Tuesday
- Assignment 8: Due  Friday*
- Read Chapter 3 notes
- Read text: Text Chapter 3.1-3.6

# Superbowl

- https://projects.fivethirtyeight.com/2018-nfl-predictions/

UPDATED JAN. 20, 2019, AT 10:16 PM

## 2018 NFL Predictions

For the regular season and playoffs, updated after every game.

**More NFL:** Every team's Elo history

Can you outsmart our forecasts?

**Standings** | Games

| | | | | PLAYOFF CHANCES | | | |
|---|---|---|---|---|---|---|---|
| ELO RATING | 1-WEEK CHANGE | TEAM | DIVISION | MAKE DIV. ROUND | MAKE CONF. CHAMP | MAKE SUPER BOWL | WIN SUPER BOWL |
| 1686 | +25 | **New England** 13-5 | AFC East | ✓ | ✓ | ✓ | 53% |
| 1667 | +19 | **L.A. Rams** 15-3 | NFC West | ✓ | ✓ | ✓ | 47% |
| 1664 | | **New Orleans** 14-4 | NFC South | ✓ | ✓ | – | – |
| 1651 | | **Kansas City** 13-5 | AFC West | ✓ | ✓ | – | – |
| 1626 | | **L.A. Chargers** 13-5 | AFC West | ✓ | – | – | – |
| 1620 | | **Philadelphia** 10-8 | NFC East | ✓ | – | – | – |

# SuperBowl Results

# Two Statistical Frameworks: Frequency vs. Bayesian

- Frequency- probability of an event is its relative frequency after many trials
- a- number of occurrences of E
- n- number of opportunities for E to take place
- a/n- relative frequency of E occurring
- $Pr\{E\} \rightarrow a/n$ as $n \rightarrow \infty$

# Two Statistical Frameworks: Frequency vs. Bayesian

- Bayesian- probability represents the degree of belief of an individual about an outcome of an uncertain event
- Some events occur so rarely that there is no long-term relevant probability
- Two individuals can have different probabilities for same outcome
- Bookies are Bayesian
  - **Super Bowl odds 2019: Heavy Patriots money has sportsbooks rooting for  Rams**

# Conditional Probability

- Conditional probability: probability of $\{E_2\}$ given that $\{E_1\}$ has occurred
- $Pr\{E_2 \mid E_1\} = Pr\{E_1 \cap E_2\} / Pr\{E_1\}$
- $E_1$ is the conditioning event
- If $E_1$ and $E_2$ are independent of each other, then $Pr\{E_2 \mid E_1\} = Pr\{E_2\}$ and $Pr\{E_1 \mid E_2\} = Pr\{E_1\}$
- Fair coin- Pr{heads} = 0.5
  - chance of getting heads on second toss is independent of the first

    Pr{heads | heads} = 0.5

    Pr{heads} twice = 0.5 * 0.5 = .25

# Bayes Theorem

- $Pr\{E_2 \mid E_1\} = Pr\{E_1 \cap E_2\} / Pr\{E_1\}$
- $E_1$ is the conditioning event
- What is the advantage? Probability of conditioning event $E_1$ only computed once

- $Pr\{E_1 \mid E_2\} = Pr\{E_2 \mid E_1\} * Pr\{E_1\} / Pr\{E_2\}$

- $Pr\{E_1 \cap E_2\} = Pr\{E_2 \mid E_1\} * Pr\{E_1\}$
- $Pr\{E_1 \cap E_2\} = Pr\{E_1 \mid E_2\} * Pr\{E_2\}$ then

# Application of Bayes theorem: how to be rational responding to probabilities

| | Pos Test | Neg Test | TOTAL |
|---|---|---|---|
| DRUG USER | 0.495% | 0.005% | 0.5% |
| NOT DRUG USER | .995% | 98.505% | 99.5% |
| TOTAL | 1.49% | 98.51% | 1 |

What are odds of a drug user skating?

$E_4$ – drug user

$E_3$ - negative test

$Pr\{E_3\}$ – 98.51%

$Pr\{E_4\}$ – 0.5%

$Pr\{E_3 \cap E_4\}$ – .005%

$Pr\{E_4 \mid E_3\} = Pr\{E_4 \cap E_3\} / Pr\{E_3\} =$

$= 0.005 / 98.51 = .0051\%$

Out of 10000 people, maybe 1 drug user will test negative

Conclusion: people who give drug tests are more interested in making sure drug users are caught than worrying about innocent people being falsely accussed

# Forecast Verification

- What is your reason for doing it?
- (Brier and Allen 1951; Compendium of Meteorology)
  - Administrative: who's blowing the forecasts?
  - Scientific: why do errors happen?
  - Economic: what's the impact of forecast errors?

# Measures oriented: “give me a number!”

- Distill set of forecasts and observations into small # of metrics

| | | Observed | Observed | Forecast marginal totals |
|---|---|---|---|---|
| | | Yes | No | |
| Forecast | Yes | a | b | a+b |
| Forecast | No | c | d | c+d |
| | Observed marginal totals | a+c | b+d | n=a+b+c+d sample size |

- PC = percent correct = $\frac{a+d}{n}$
- FAR = false alarm ratio = $\frac{b}{a+b}$
- TS = CSI = $\frac{a}{a+b+c}$
- POD = HR = $\frac{a}{a+c}$

# What if it just happened by chance?

| | | Observed | Observed | Forecast marginal totals |
|---|---|---|---|---|
| | | Yes | No | |
| Forecast | Yes | a | b | a+b |
| Forecast | No | c | d | c+d |
| | Observed marginal totals | a+c | b+d | n=a+b+c+d sample size |

- Random correct yes forecast by chance = $\frac{(a+b)}{n}\frac{(a+c)}{n}$
- Random correct no forecast by chance = $\frac{(b+d)}{n}\frac{(c+d)}{n}$
- $SS = \frac{(correct\ forecasts - random\ correct\ forecasts)}{(total\ forecasts - random\ correct\ forecasts)}$
- $HSS = \frac{2(ad-bc)}{(a+c)(b+d)+(a+b)(b+d)}$

# Verifying wind forecasts

| | | Observed | Observed | Forecast Marginal totals |
|---|---|---|---|---|
| | | ≥ 5m/s | <5 m/s | |
| Forecast | ≥ 5m/s | 11 | 6 | 17 |
| Forecast | <5 m/s | 16 | 44 | 60 |
| | Observed Marginal totals | 27 | 50 | 77 |

PC= 71.4%; FAR= 35.3%; TS= 33.3%; and POD = 40.7%
randomly correct yes forecast: 7.7%
randomly correct no forecast: 50.1%
HSS= 31.4%

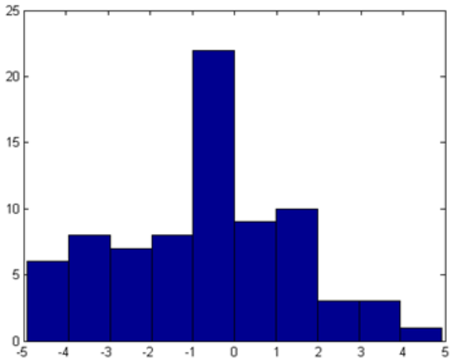
## Distributions oriented: “ how close am I?”

- Assessing the characteristics of joint distribution of errors
- Categorize errors: which errors are smallest, which are biggest as a function of the range of values?
- Relies heavily on conditional probabilities
- http://meso1.chpc.utah.edu/jfsp/

# Forecast Verification

- http://meso1.chpc.utah.edu/jfsp/
- Select Wildfires by WFO
- Select SLC
- Look at over all years, then focus on wildfires in Utah in 2016
- Then follow along in class

# Assessing Forecast Accuracy



| | | Observed | Observed | Observed | Error Marginal totals |
|---|---|---|---|---|---|
| | | ≤3 m/s | 3-6 m/s | ≥6 m/s | |
| Error | ≤ -2 m/s | 0 | 10 | 11 | 21 |
| Error | ± 2 m/s | 22 | 20 | 7 | 49 |
| Error | > 2 m/s | 0 | 7 | 0 | 7 |
| | Observed Marginal totals | 22 | 37 | 18 | 77 |

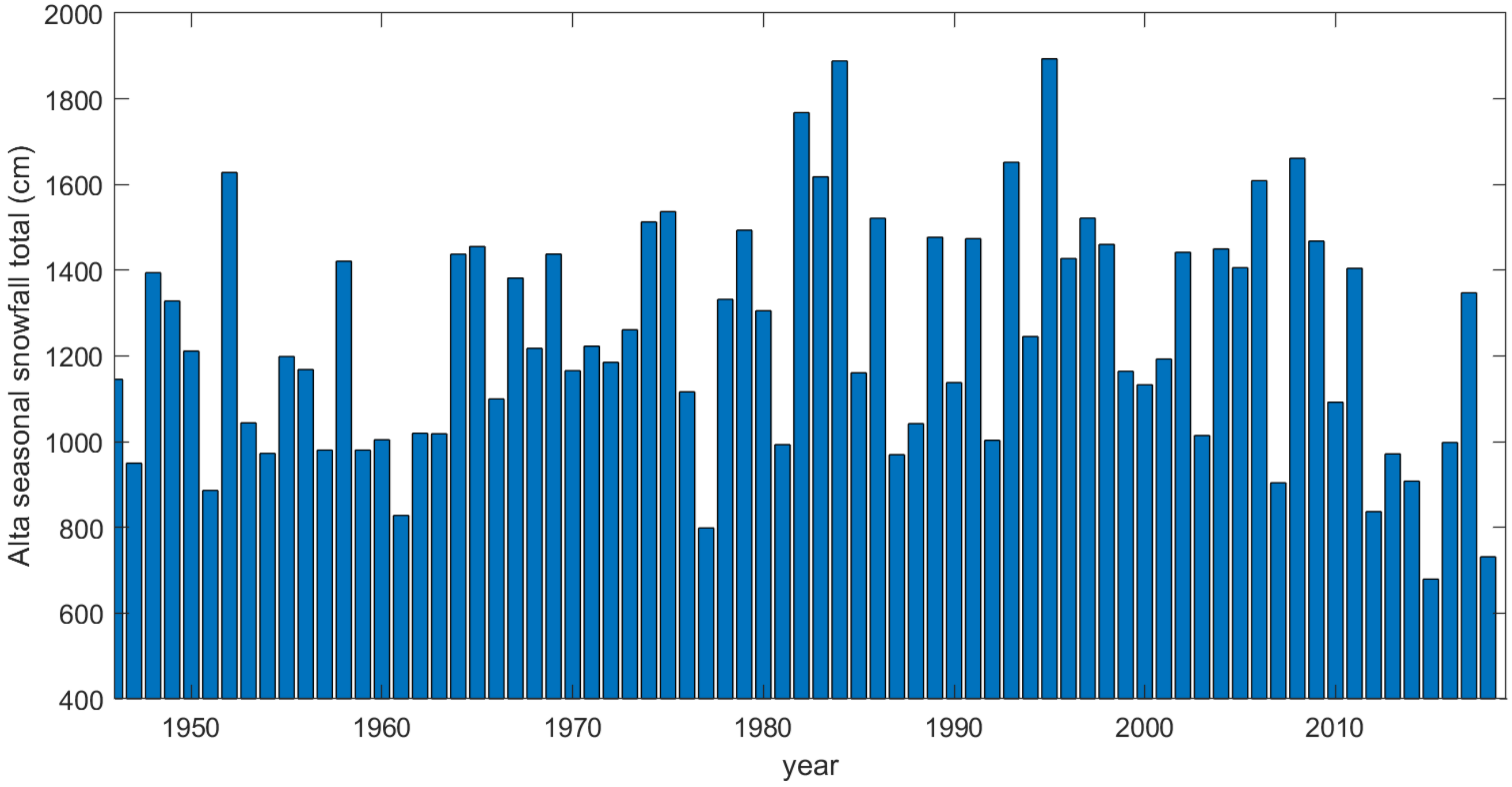
- 26% of the forecasts were within 2 m/s when the wind speeds were between 3 and 6 m/s (20/77)
- Given that the observed wind speed is greater than 6 m/s: ($Pr\{E_1\}$ = 18/77= 23.4%)
- Probability that the forecasters predict strong winds to be too light $Pr\{E_2 \mid E_1\}$:

$Pr\{E_2 \mid E_1\} = Pr\{E_1 \cap E_2\} / Pr\{E_1\} = ((11/77)/(18/77)) = 64.7\%$

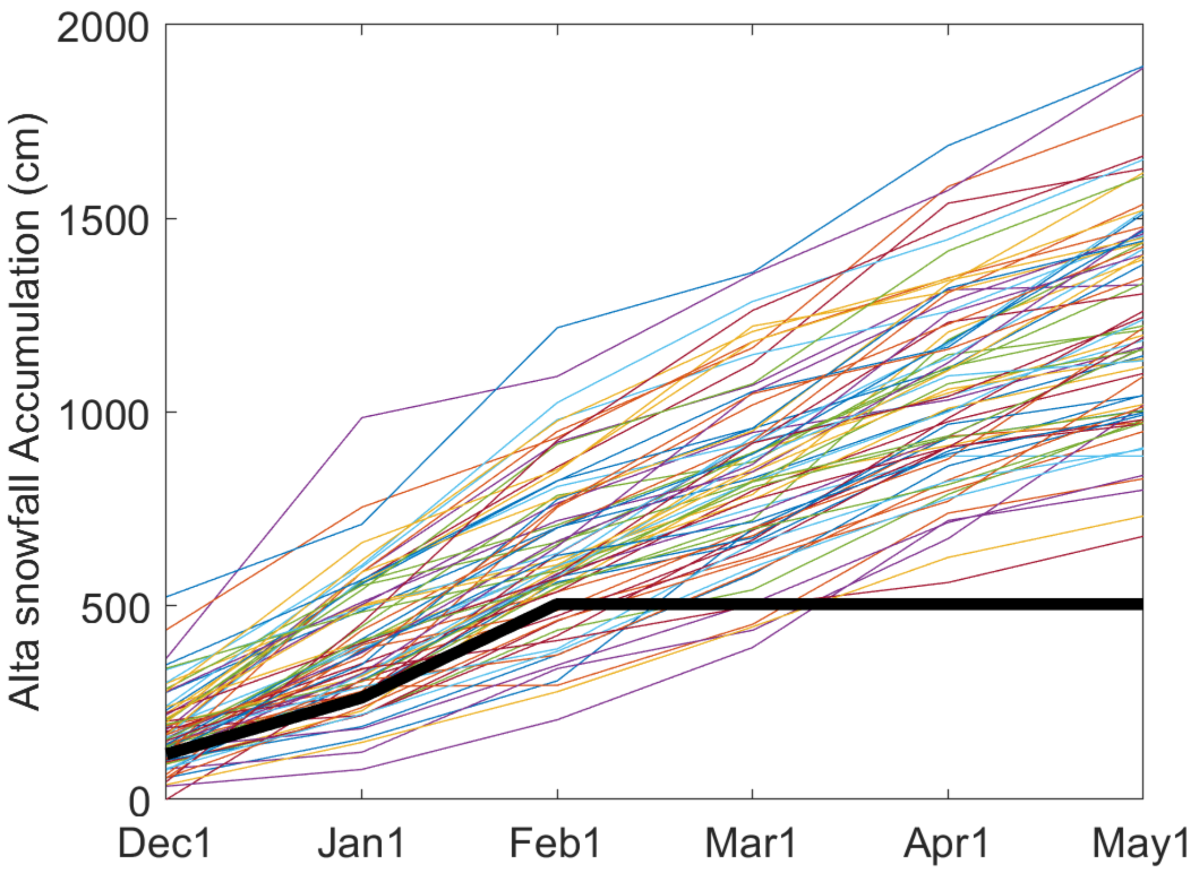
## What can we say about estimating this winter’s snow total will be?

- What physically is happening?
- Could we use last winter’s snow total to predict this winter’s?
  - Persistence from one year to next
- What about the amount of snow earlier this winter or right now?
  - Persistence from one month to the next…

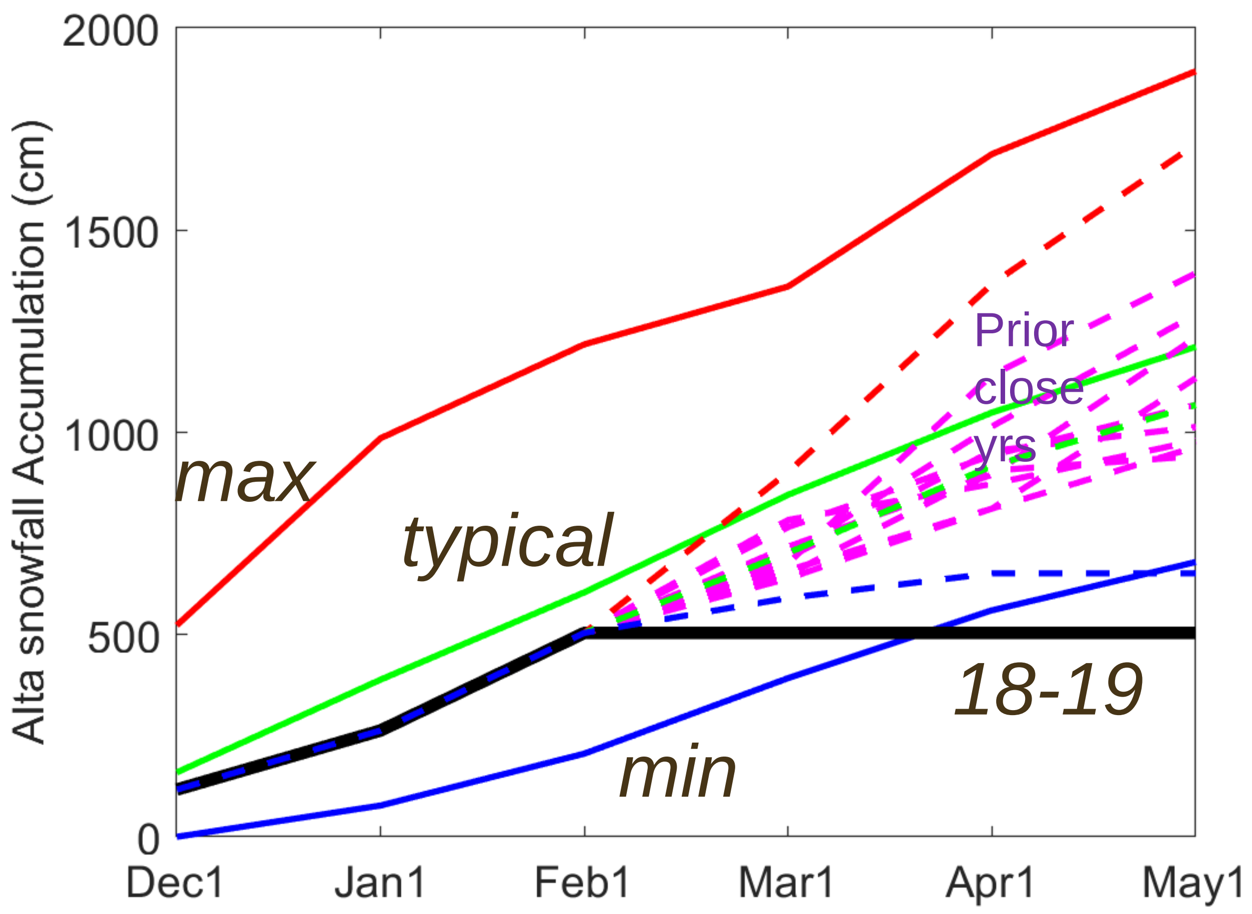
# Alta Snowfall Seasonal Totals



# Atla Snowfall Accumulation Each Winter



# How Much Snow Might Accumulate During the Season at Alta?



# Predict May 1 Snowfall from Dec 1 Snowfall

| Case 1. Predictor: Dec1 total snowfall (cm) | | | | | |
|---|---|---|---|---|---|
| **Predictand: May 1 Total snowfall at Alta (cm)** | | Below | Near | Above | Marginal Totals |
| | Below | 14 | 9 | 1 | 24 |
| | Near | 6 | 9 | 10 | 25 |
| | Above | 4 | 7 | 13 | 24 |
| | Marginal Totals | 24 | 25 | 24 | 73 |

# Predict May 1 Snowfall from Dec 1 Snowfall

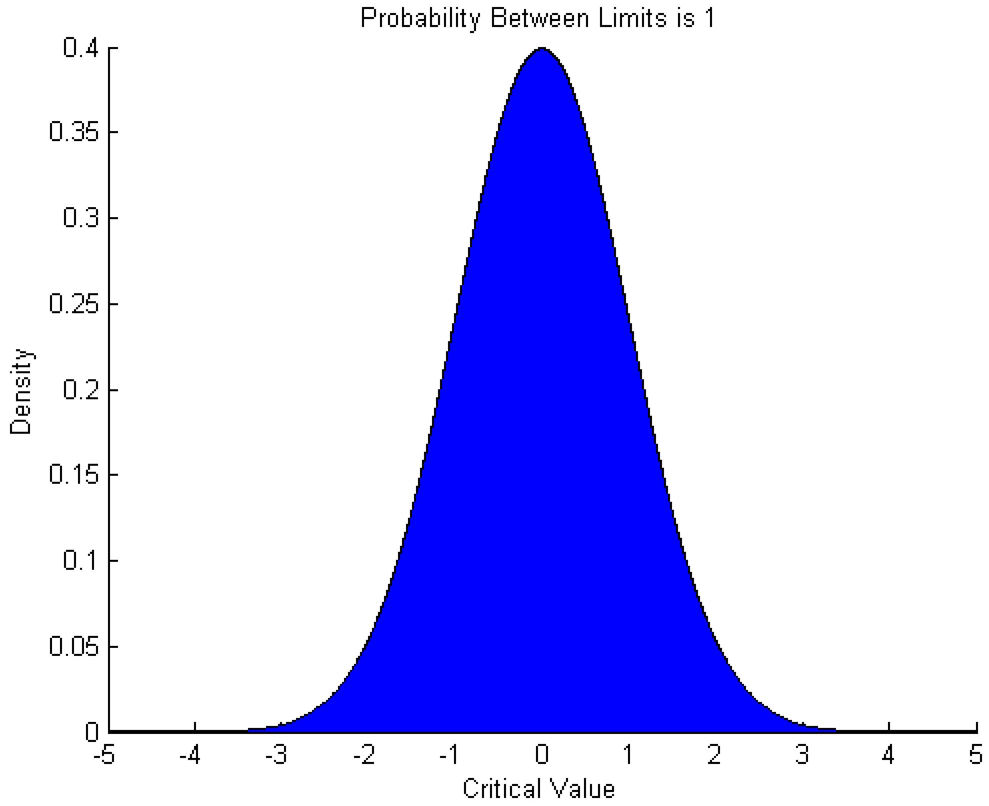
| | | | |
|---|---|---|---|
| **$Pr\{E_1\}$** | 24/73=33% | $Pr\{M_3 \mid E_3\}$ | 13/24=54% |
| **$Pr\{E_1 \cap E_2\}$** | 0 | $Pr\{E_1 \mid M_1\}$ | 14/24=58% |
| **$Pr\{E_1 \cap M_1\}$** | 14/73=19% | $Pr\{E_2 \mid M_1\}$ | 9/24=38% |
| **$Pr\{E_1 \cap M_3\}$** | 4/73= 5% | $Pr\{E_3 \mid M_1\}$ | 1/24=4% |
| **$Pr\{M_1 \mid E_1\}$** | 14/24=58% | $Pr\{E_3 \mid M_3\}$ | 13/24=54% |
| **$Pr\{M_3 \mid E_1\}$** | 4/24=17% | $Pr\{E_1 \cap M_1\}$ IF random | 9/72=11% |
| **$Pr\{M_3 \mid E_2\}$** | 7/24=29% | % May 1 total same as Predictor | 36/73= 49% |

# Predict May 1 Snowfall from Last Year's May 1 Snowfall

| Case 6. Predictor: May1 Prior Year total snowfall (cm) | | | | | |
|---|---|---|---|---|---|
| **Predictand: May 1 Total snowfall at Alta (cm)** | | Below | Near | Above | Marginal Totals |
| | Below | 10 | 7 | 7 | 24 |
| | Near | 5 | 12 | 7 | 24 |
| | Above | 8 | 6 | 10 | 24 |
| | Marginal Totals | 23 | 25 | 24 | 72 |

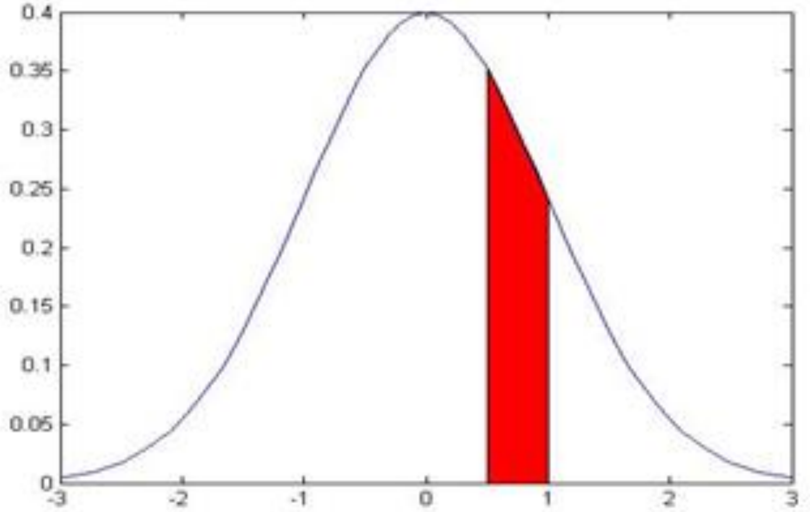
# Empirical vs. Parametric Distributions

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



# Random Continuous Variable x

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

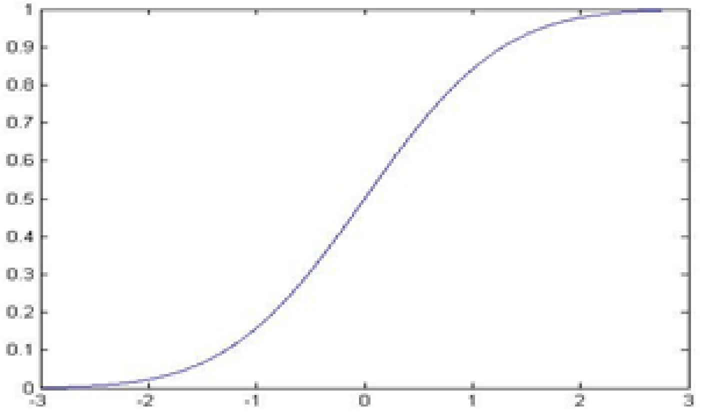



$$\int_{-\infty}^{\infty} f(x)dx = 1$$

# Cumulative Density Function of Continuous Variable

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

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



# Gaussian Parametric Distribution

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

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

# Using parametric distributions

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