Probability & Sampling

http://www.pelagicos.net/classes_biometry_fa16.htm

High-rise Syndrome in Cats

Observation: Cats land unharmed on their feet, no matter how far they fall

132 cats falling more than 2 stories (mean = 5.5 +/- 0.3 S.E., max = 32 stories), have a survival rate of about 90%, assuming they are treated for the injuries that occur because of the ground impact.

Data:

17 / 132 euthanized by owners

11 / 115 died due to injuries

104 / 115 survived

(Diamond, 1988)

High-rise Syndrome in Cats

Remove the 17 euthanized cats from the sample

p = P (surviving) = 90% (104 / 115)
q = P (dying) = 10% (11 / 115)

$$p + q = 1$$

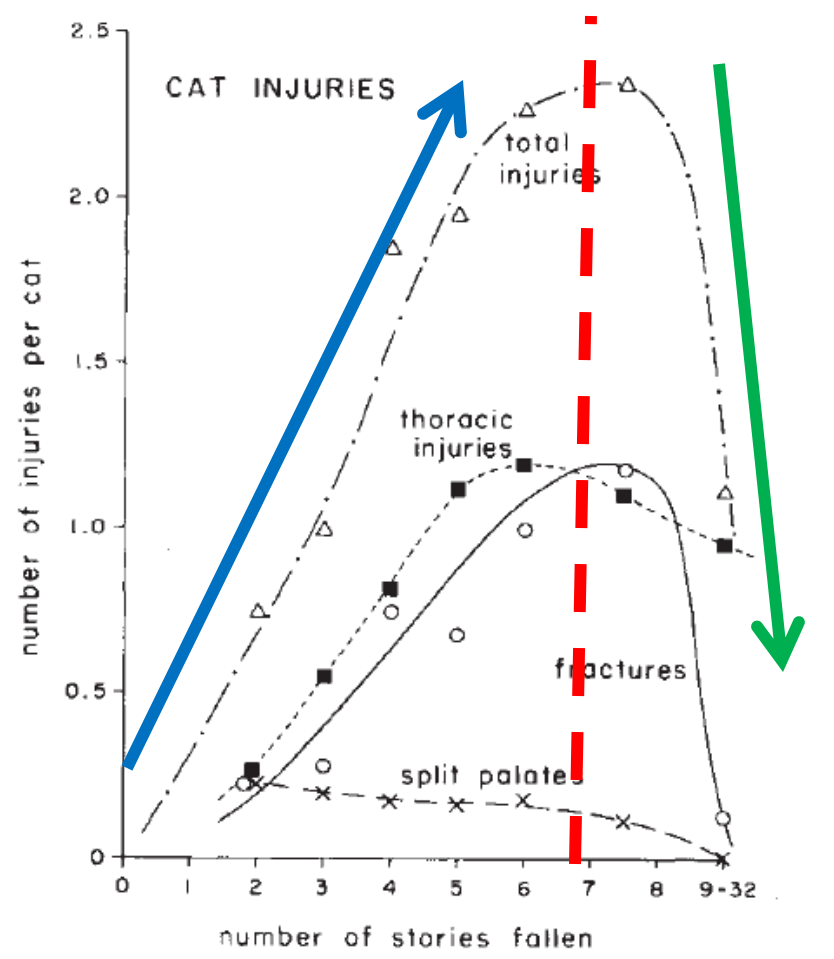
Proposed Explanation:

Cats are able to right themselves, and reach terminal velocity (60 mph) after about 5 storeis

At this point, they relax and glide… and land safely on their feet (NOTE: Force = mass * acceleration)

High-rise Syndrome in Cats

Closer look at the Data :

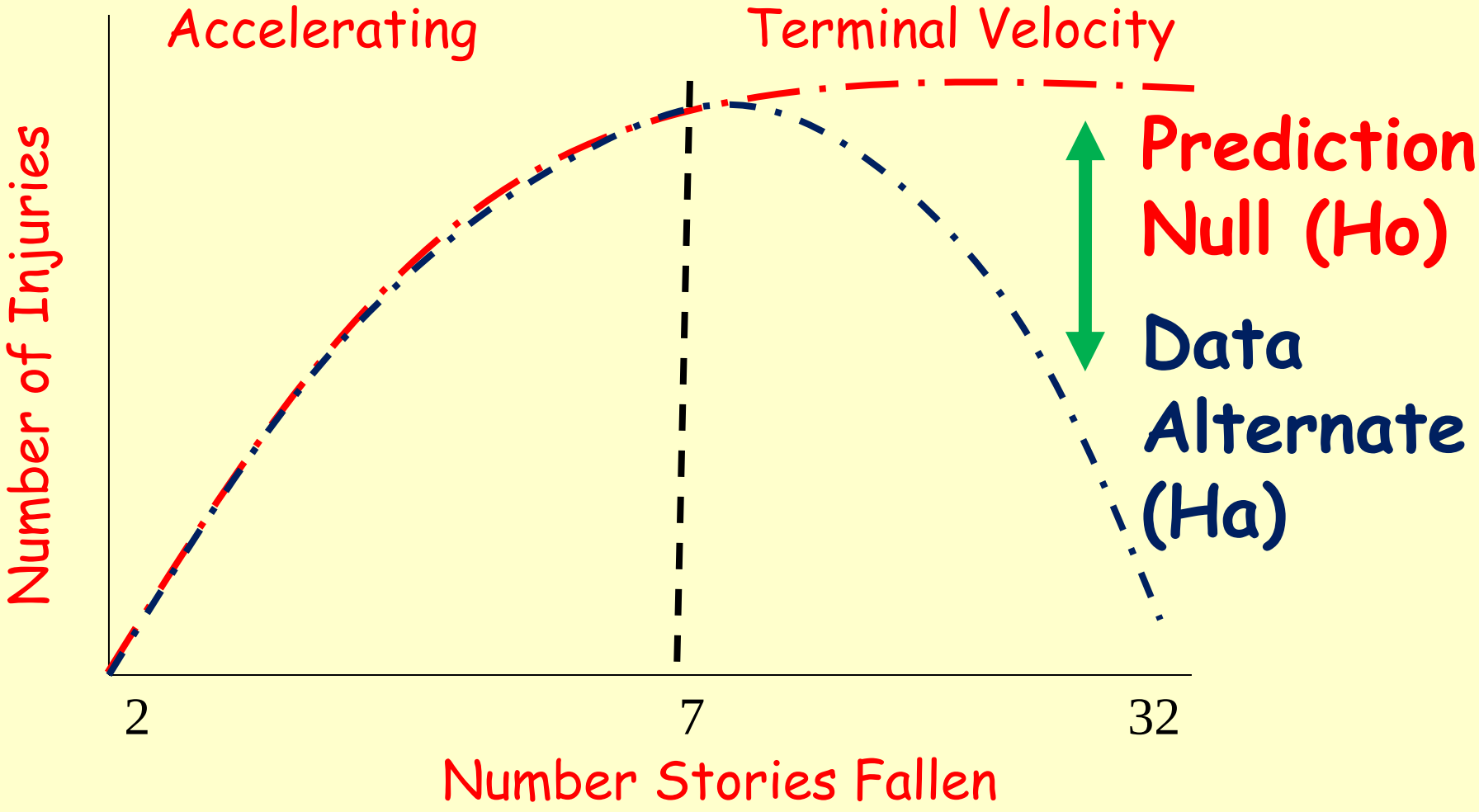


From 1 – 6 storeis:
Injuries increase with height

From 7 – 32 storeis:
Injuries decrease with height

High-rise Syndrome in Cats

Reconciling the Observations with the Predictions:



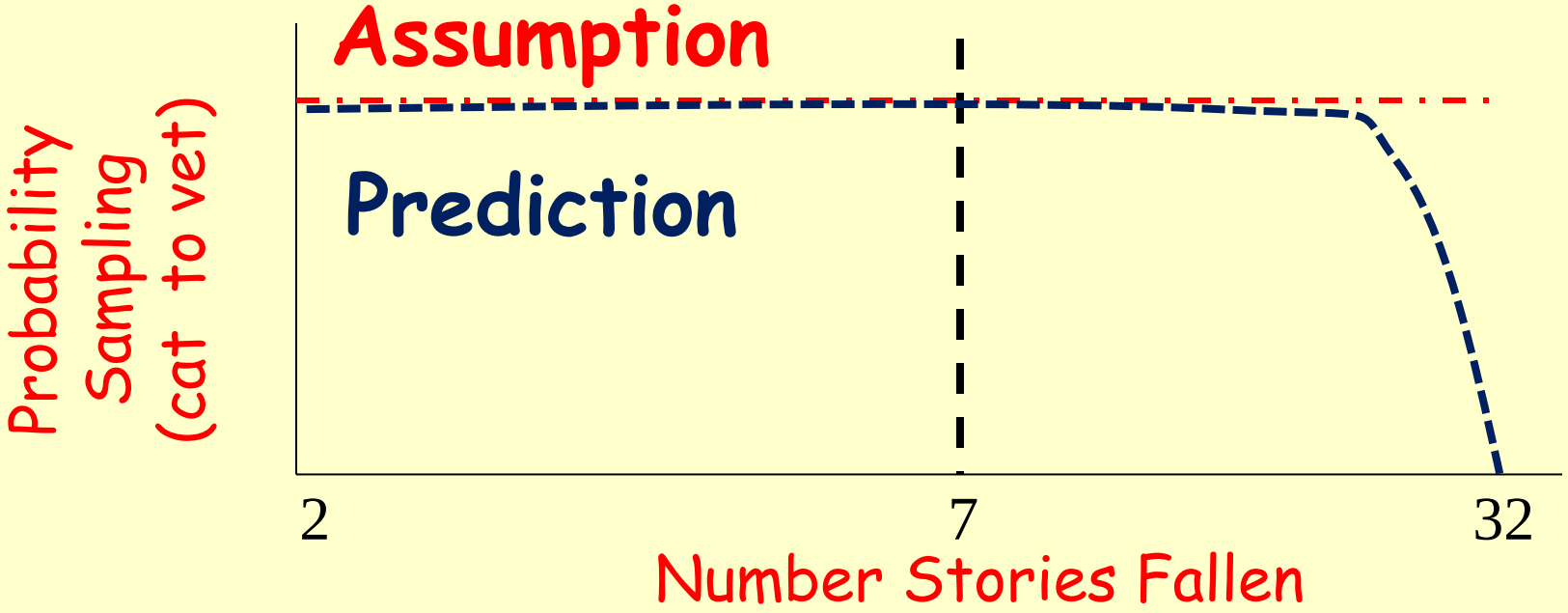
High-rise Syndrome in Cats

Reconciling the Observations with the Predictions:

Another Explanation: Two processes involved in this study

- Cat falls from the high-rise and is brought to vet
- Vet quantifies the injuries sustained by the cat

Question: Are these events independent?



High-rise Syndrome in Cats

Take Home Lessons:

1. Carefully consider the matches / mis-matches between the "biological population" and the "statistical sample"

2. Sampling and probability estimation require determining the sample space and the independence of events

How to Quantify Probability ?

Probability

Relative occurrence of a particular result, from a given number of trials – in a finite sample

Relative occurrence of a particular result, from an infinite number of trials – in an infinite sample

Probability Estimation

1. Determine biological population and define sample space
2. Estimate probability from finite sample (observation)
3. Use estimate to assess the actual rate of occurrence of the phenomenon of interest (expectation)

Defining the Sample Space

Sample Space

The universe of all possible events

e.g., flip of a coin: heads OR tails

Axiom 1: The sum of all the probabilities of outcomes within a single sample space = 1.0

$$\sum_{i=1}^{n} P(Oi) = 1.0$$

NOTE: In a properly defined "sample space", the outcomes are mutually exclusive

Complex and Shared Events

Complex Events

Composites of simple events in the same space

e.g., probability (heads) OR probability (tails)

Shared Events

Simultaneous occurrence of multiple simple events in the same space

e.g., probability (heads) AND probability (tails)

Axiom 2: Probability of a complex event equals the sum of outcomes that make up that event

Shared Events & Independence

e.g., Imagine we are flipping two coins sequentially

Probability of Heads & Heads = (1/2) * (1/2)

Event 1

Event 2

What is the Critical Assumption? Independence

Axiom 3: If two events are independent, the probability of a shared event (both events occur) equals the product of the individual probabilities.

Independence – Card Example

Sampling with Replacement: Choosing an ace from a deck of cards, replacing it, AND then choosing an ace as the second card.

P (ace in first draw) = (4 / 52) = 0.077

P (ace in second draw) = (4 / 52) = 0.077

Sampling without Replacement: Choosing an ace from a deck of cards, AND then choosing an ace as the second card.

P (ace in first draw) = (4 / 52) = 0.077

P (ace in second draw) = (3 / 51) = 0.058

Conditional Probability

When calculating the probability of a complex event, information on known outcomes can be used to revise the probability calculations.

These updated estimates are termed "conditional probabilities" $P(A \mid B)$

Probability of A, given than B occurred

e.g., How would knowing that the first coin was a head, modify the calculation?

Event 1

(1/2)

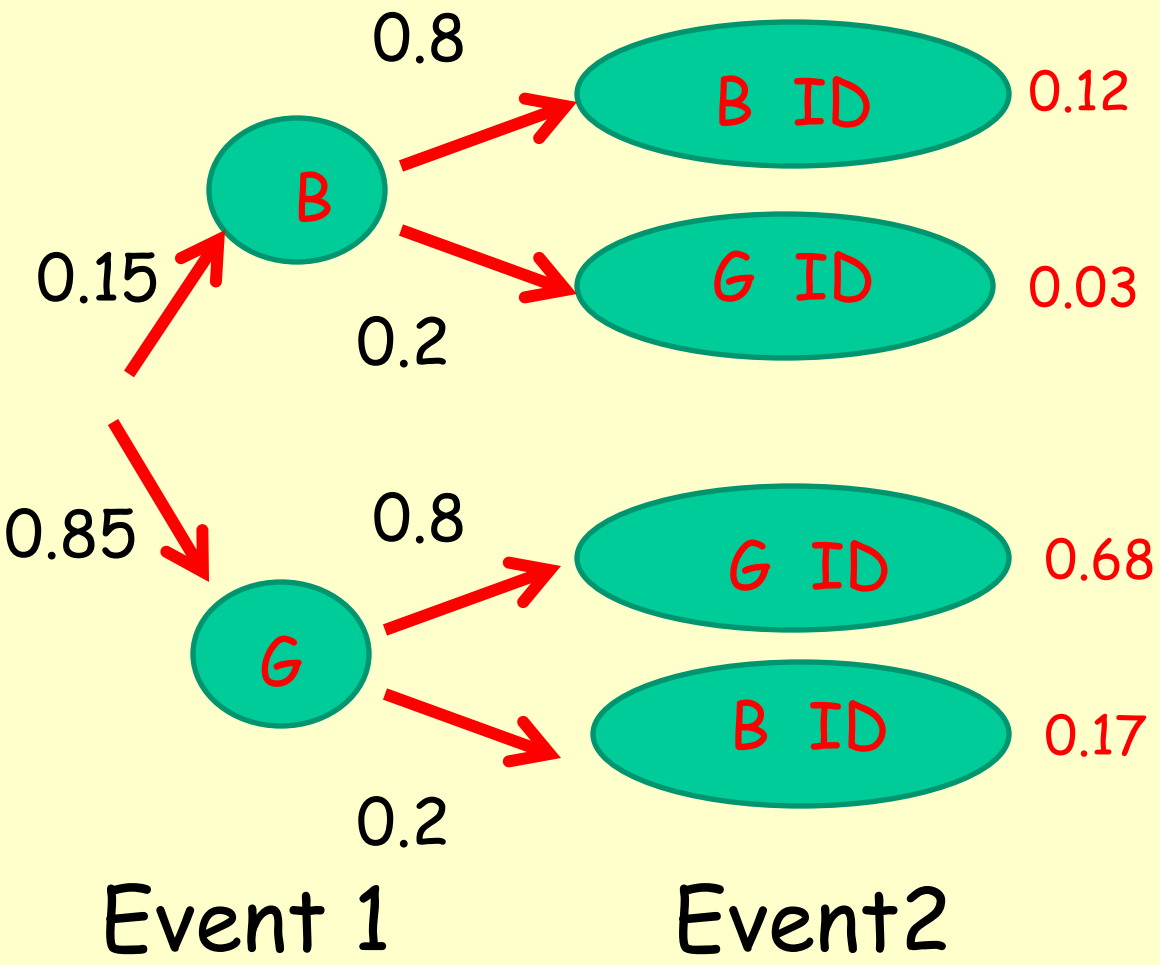
Event 2

Conditional Probability

Question: What is probability that car was blue if identified as a blue car?

Car colors:
- 85% green
- 15% blue

ID:
- 80% correct
- 20% incorrect



Solution:

$$\frac{\text{Blue ID \& Blue Car}}{\text{Blue ID}}$$

0.12 / (0.12 + 0.17)

Probabilistic Independence

Definition: Two events, A and B, are independent if the fact that A occurs does not affect the probability of B occurring.

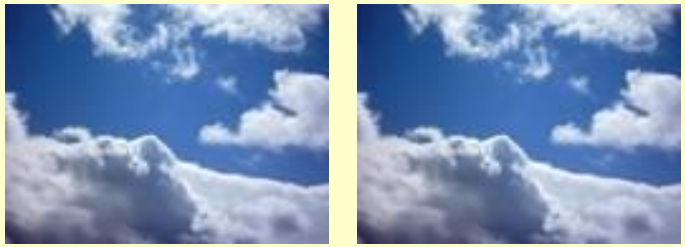
Some examples of independent events are:

Landing on heads after tossing a coin AND rolling a 5 on a single 6-sided die.

Choosing a marble from a jar AND landing on heads after tossing a coin.

Choosing a 3 from a deck of cards, replacing it, AND then choosing an ace as the second card.

Conceptually, how can you show whether two events are independent ?




P (sun)
= 2 / 5 = 0.4

P (rain)
= 3 / 5 = 0.6

	EXPECTED	OBSERVED
P (sun - sun)	= 0.4 * 0.4 = 0.16	1 / 4 = 0.25
P (sun - rain)	= 0.6 * 0.4 = 0.24	1 / 4 = 0.25
P (rain - sun)	= 0.6 * 0.4 = 0.24	0 / 4 = 0.00
P (rain - rain)	= 0.6 * 0.6 = 0.36	2 / 4 = 0.50

RESULT: Not independent

Another Example of Sample Independence

Bycatch in longlines: How many sea turtles are caught?

Estimate Parameter from the Sample:

Calculate Bycatch Rate:

$$\frac{\text{\# turtles}}{\text{\# fishing events}}$$

Use parameter to Extrapolate:

(Bycatch Rate) * (# fishing events)

Assessing Sample Independence

Bycatch in longlines: Catches of many species in clumps

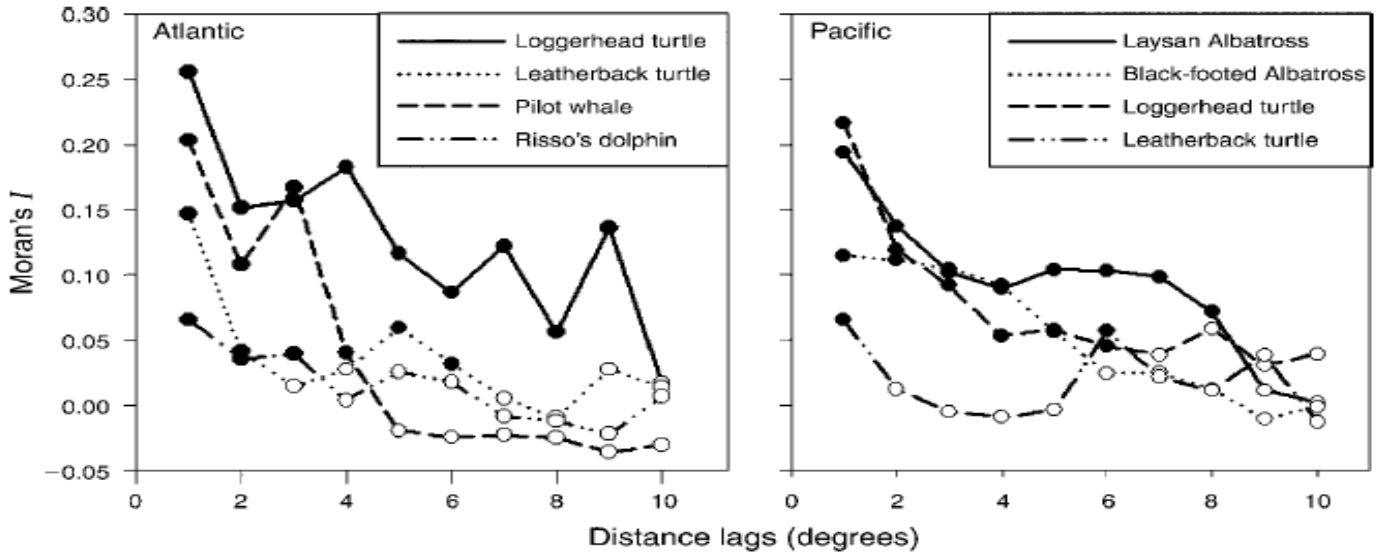


FIG. 3. Autocorrelation of bycatch for individual species in each of the ocean basins. Moran's *I* values were calculated at 1° increments for distance lags ranging from 1° to 10°. Solid circles indicate pseudo-95% confidence intervals that did not overlap zero; open cirlces indicate pseudo-95% confidence intervals that overlapped zero. Values significantly greater than zero indicate positive spatial autocorrelation of bycatch (i.e., clustering), whereas values significantly less than zero indicate negative spatial autocorrelation of bycatch (i.e., overdispersion).

(Lewison et al. 2009)

Implications:

Difficult to assess overall fisheries catch

Possible mitigation (if bycatch, then move)

Assessing Sample Independence

Sea turtle bycatch in Hawaiian longline fishery:

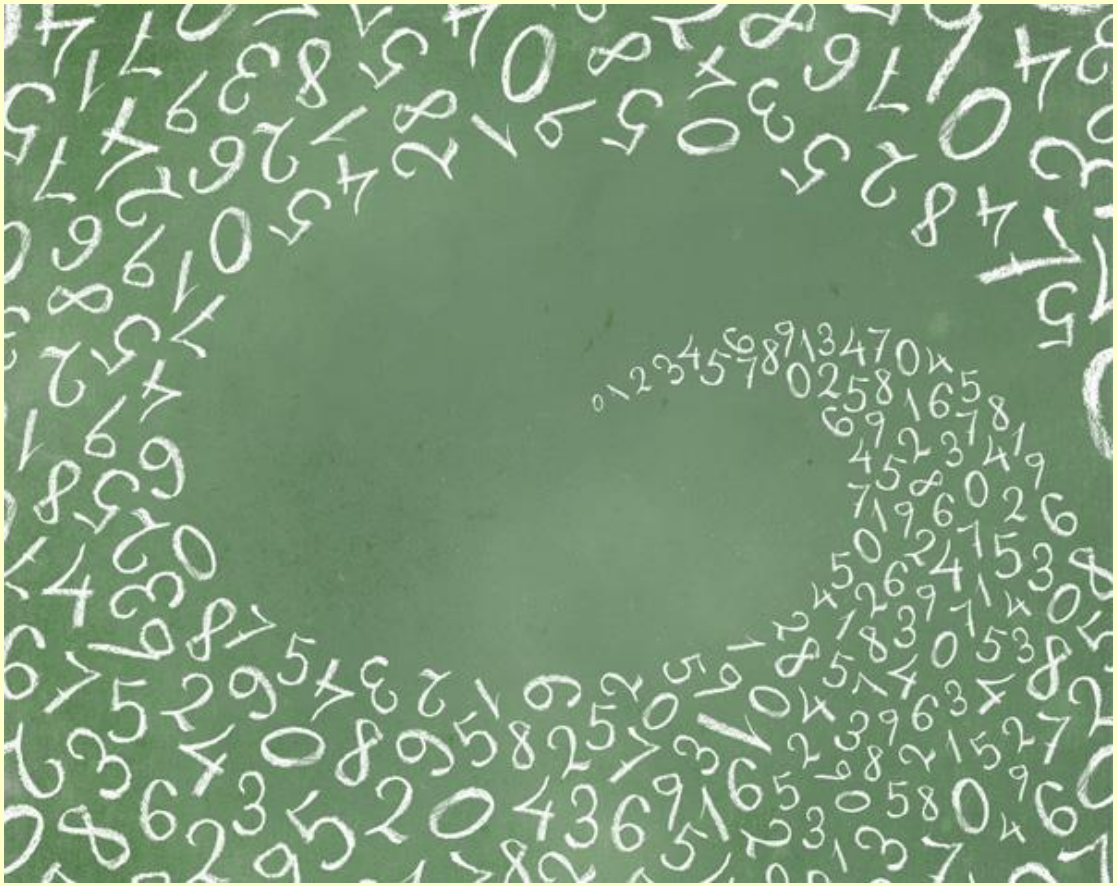
- Data: 25% of sets with turtle catches in clusters.
  Is this likely… if the catches are independent?

- Probability that 25% of 231 sets with turtle bycatch would be consecutive, if the events were truly independent (given null hypothesis is true) = 0.005

- Conclusion: Reject null hypothesis.
  Turtle Catches are not independent

Di Nardo, G.T. 1993. Statistical guidelines for a pilot observer programme to estimate turtle takes in the Hawaii longline fishery. NOAA-TM-NMFS-SWFSC-190.

Variables



Definition: Anything that can be measured and (potentially) can differ across entities or over time

Variables and Hypothesis Testing

Testing hypotheses requires making predictions and taking measurements of different variables

Often, the goal is to measure the response of one variable to an experimental change in other variables

Independent	Dependent
Variable denotes the cause (the driver of the pattern)	Variable denotes the effect (responds to the driver)
Termed: predictor variable	Termed: outcome variable

For the ant example:

Dependent: Ant Nest Density Independent: Habitat

Classes of Variables

Numerical Variables:

Variable takes on numerical values (e.g., ant nests, length)

Can be ordered and ranked

Subclasses:
-Discrete: few possible values (integers)
- Continuous: Measurements take any value within range

Categorical Variables:

Variable takes on different entities or categories (e.g., color)

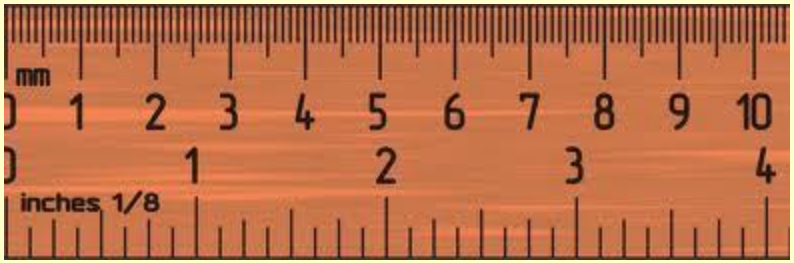
Can be ordered and ranked (ordinal variables)

Subclasses:
-Binary: Two Options
-Nominal: > 2 Options

Continuous Variables

Interval Variables:

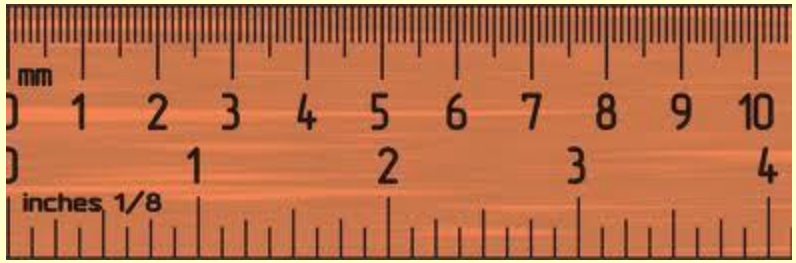
Differences in one unit of measurement equal along entire measurement scale



e.g., Temperature

Ratio Variables:

Differences in one unit of measurement equal along entire measurement scale



Rations are meaningful:
(There is a real zero value)

e.g., Length

Sampling – Summarizing Measurements

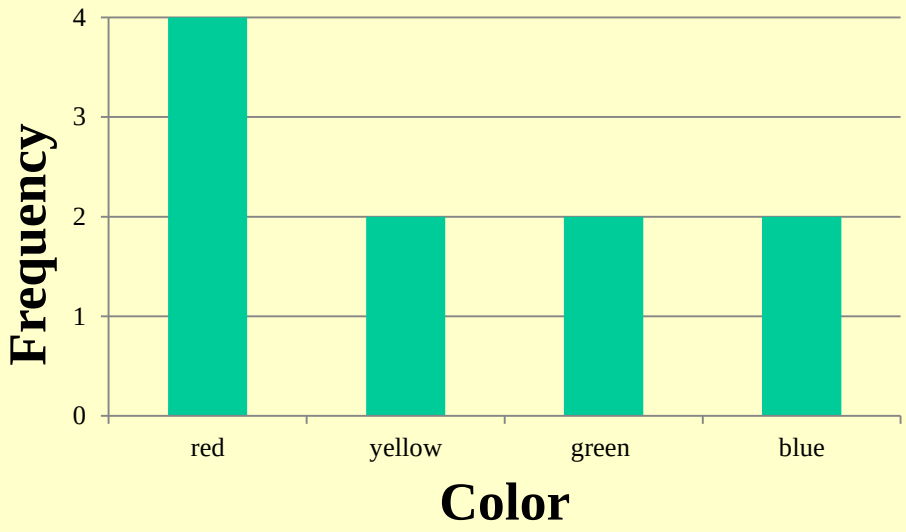
Simplest way to summarize observations: Make a list

red, red, blue, green, red, blue, yellow, yellow, red, red
1.3, 1.5, 2.2, 1.1, 1.0. 1.4. 1.7

Frequency Table

Color	Frequency	Rel. Freq.
Red	4	0.4
Yellow	2	0.2
Green	2	0.2
Blue	2	0.2

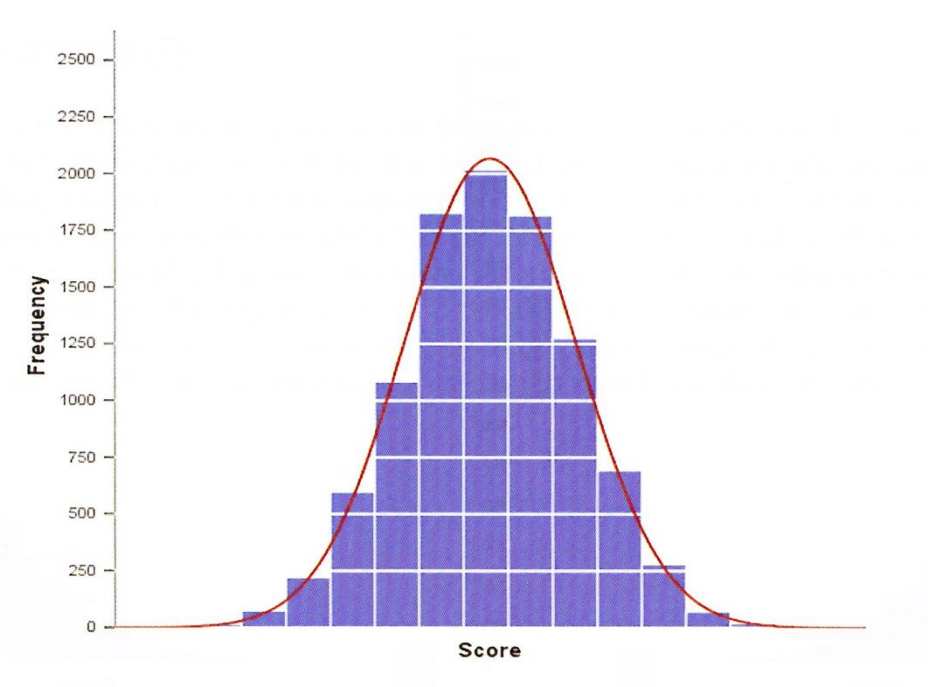
Frequency Distribution



Summarizing Data Using Figures

Frequency Distributions (Histograms) facilitate the summarization of categorical and numerical data

Important features of frequency distributions:



Range:

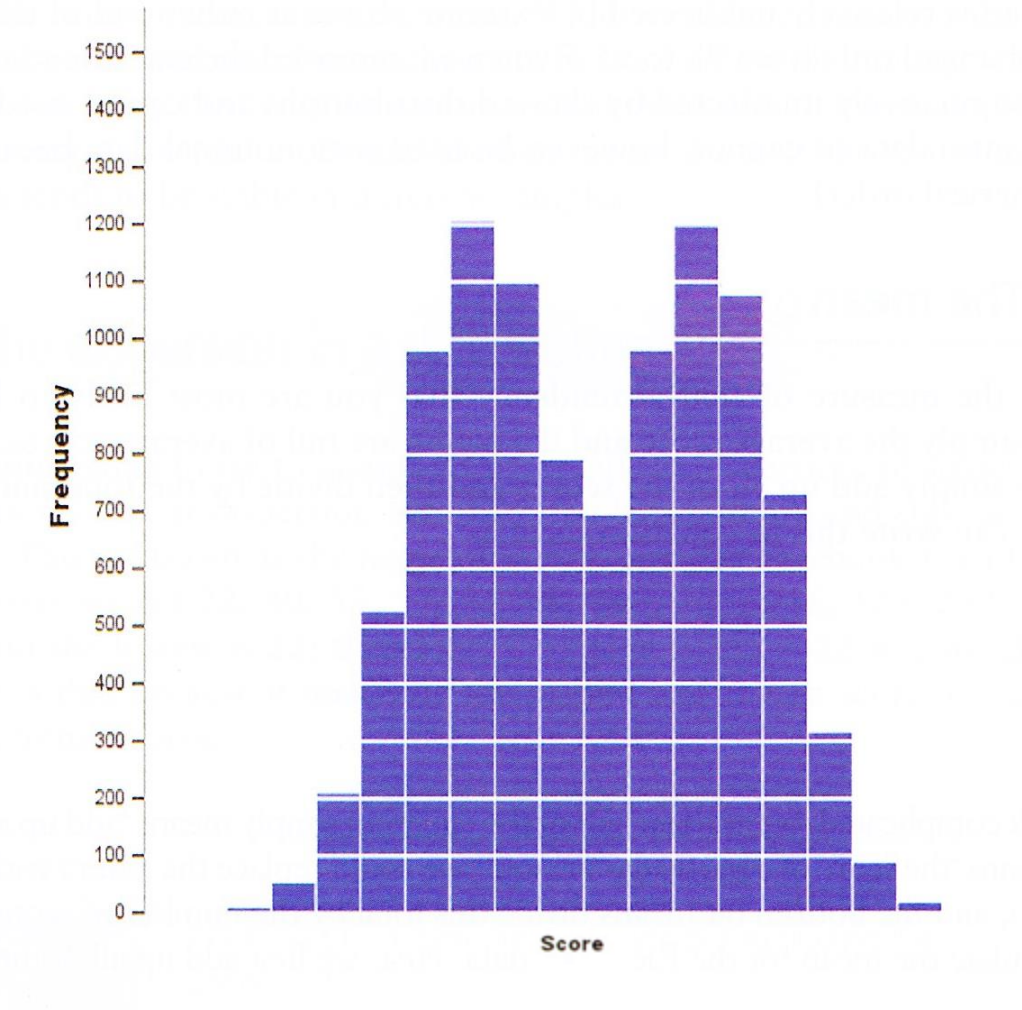
Maximum – Minimum

Beware of Bins:

X axis Categories

Quantifying Distributions

Several metrics characterize shape of distributions



Mode: Most numerous observation(s)

Beware of Bins:

X axis Categories

Note:

Distributions can have multiple modes, or none

Uni-modal: one mode

Bi-modal: two modes

Multi-modal: > 2 modes

Quantifying Distributions

Several metrics characterize shape of distributions

Median: Mid-point of the distribution (50%)

Note:

All distributions have only one median

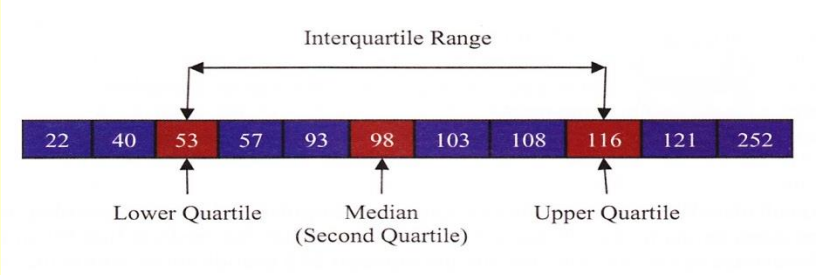
What are the medians of dist1 and dist2, below?

dist1: 1, 2, **3**, 4, 5

dist2: 1, 2, **3**, 4, 50

Percentiles: Distributions are characterized using certain percentages of observations

(e.g., 25%, 50%, 75%)



Quantifying Distributions

Several metrics characterize shape of distributions

Mean: The average of the scores in the distribution

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

What are the means of dist1 and dist2, below?

dist1: 1, 2, 3, 4, 5

mean1: 15 / 5 = 3

dist2: 1, 2, 3, 4, 50

mean2: 60 / 5 = 12

Quantifying Distributions

Distribution shapes categorized by symmetry (skew)

Skew: Measure of the symmetry of a distribution.

Symmetric distributions have a skew = 0.

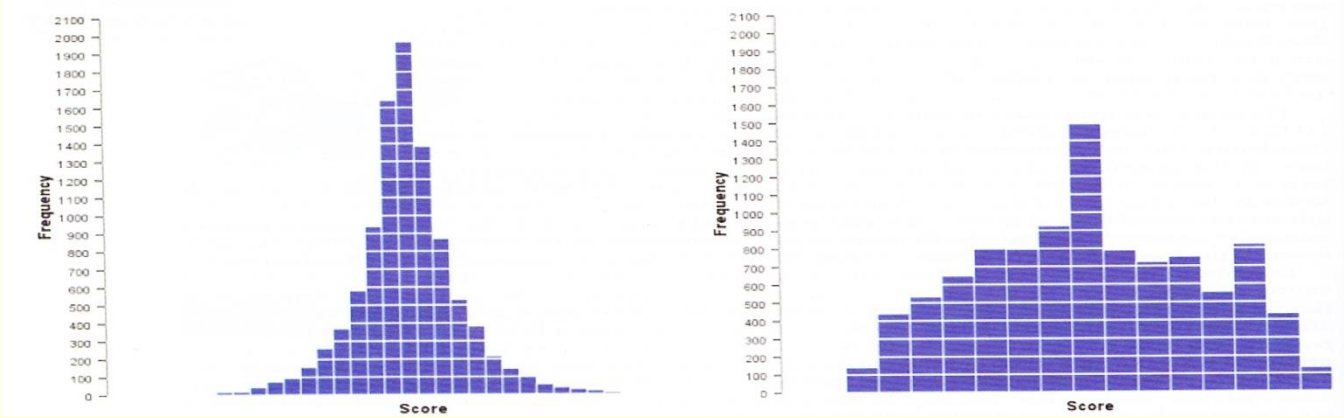
Positive skew:
the mean is larger than the median,
skewness > 0

Negative skew:
the mean is smaller than the median,
skewness < 0

Quantifying Distributions

Distribution shapes categorized by kurtosis

Kurtosis: Measure of the degree to which observations cluster in the tails or the center of the distribution.



Positive kurtosis:
Less values in tails and more values close to mean.
Leptokurtic.

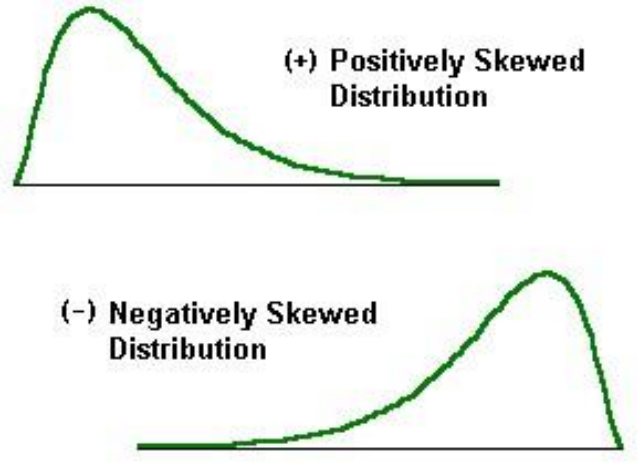
Negative kurtosis:
More values in tails and less values close to mean.
Platykurtic.

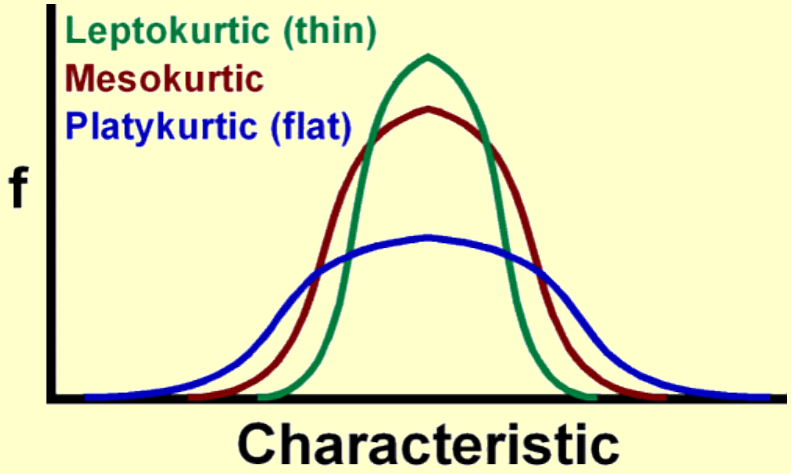
Quantifying Distributions

Formulas for skewness and kurtosis:

$$\text{Skew} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

$$\text{Kurtosis} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$





Summary

Quantifying Probability

Consider sample space and conditional probability

Be ready to perform calculations like in examples

Defining Variables

Influence how we measure / quantify observations

Memorize important definitions

Characterizing Variables

Use frequency tables and distributions

Distributions characterized with certain metrics:

range, median, mean, mode, skew, kurtosis

Memorize important definitions

Readings

Diamond, J. (1988).
Why cats have nine lives.
Nature 332 (6165): 586–587.

Lewison, R.L., Soykan, C.U., Franklin J. (2009).
Mapping the bycatch seascape: multispecies and multi-scale spatial patterns of fisheries bycatch.
Ecological Applications 19(4): 920–930