Introduction to Bayes Theorem

Lesson 5

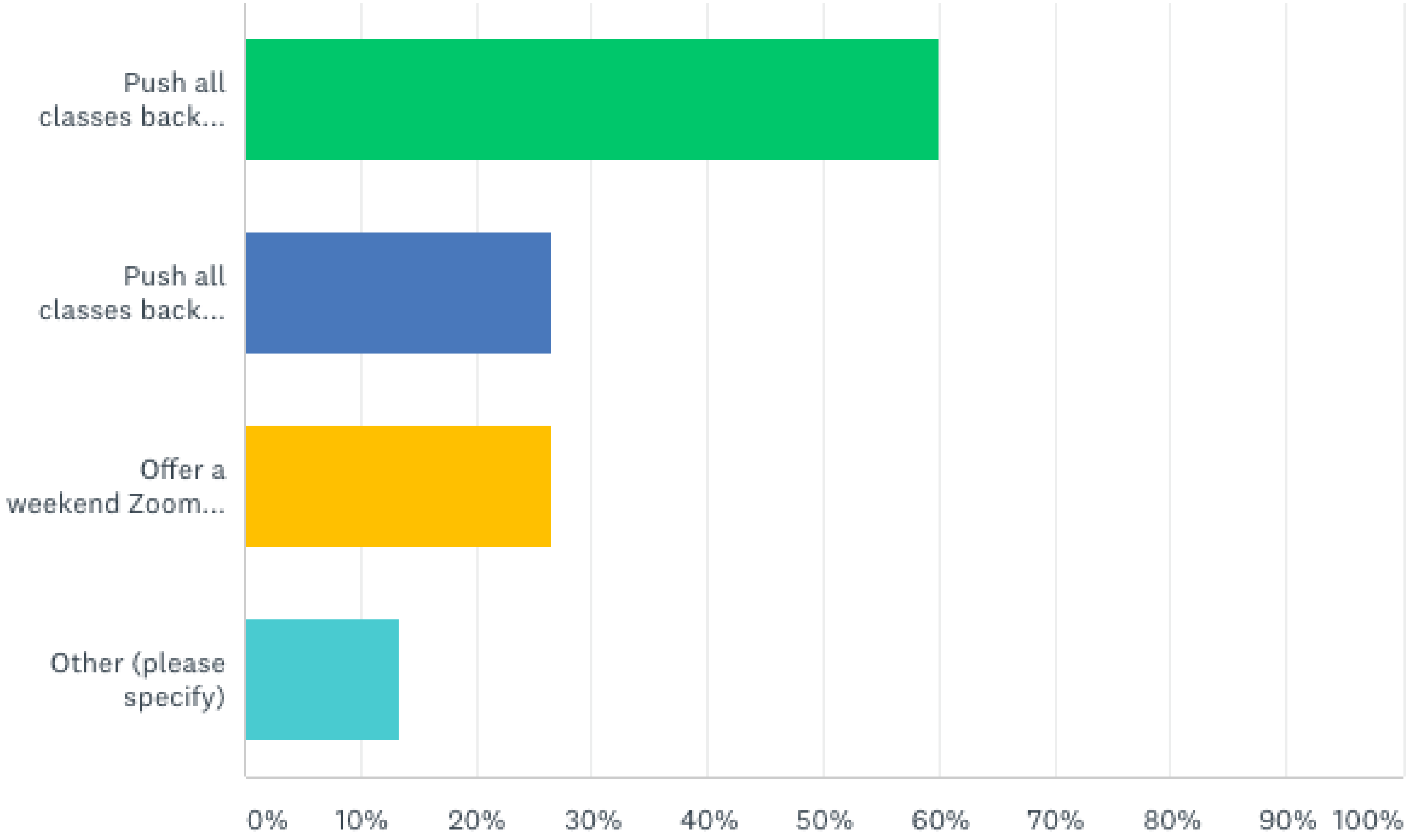
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Topics

- **Survey Results**
- **** Milestone 2 Due Feb 17th ****
- **Review HW4**
- **Central Limit Theorem**
- **Confidence Intervals**
- **Bayes Statistics**

Snow Day Survey Results

Class extended by one week, all due dates pushed back



Central Limit Theorem

C**entral Limit Theorem** (CLT) is a statistical theory states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.

Central Limit Theorem

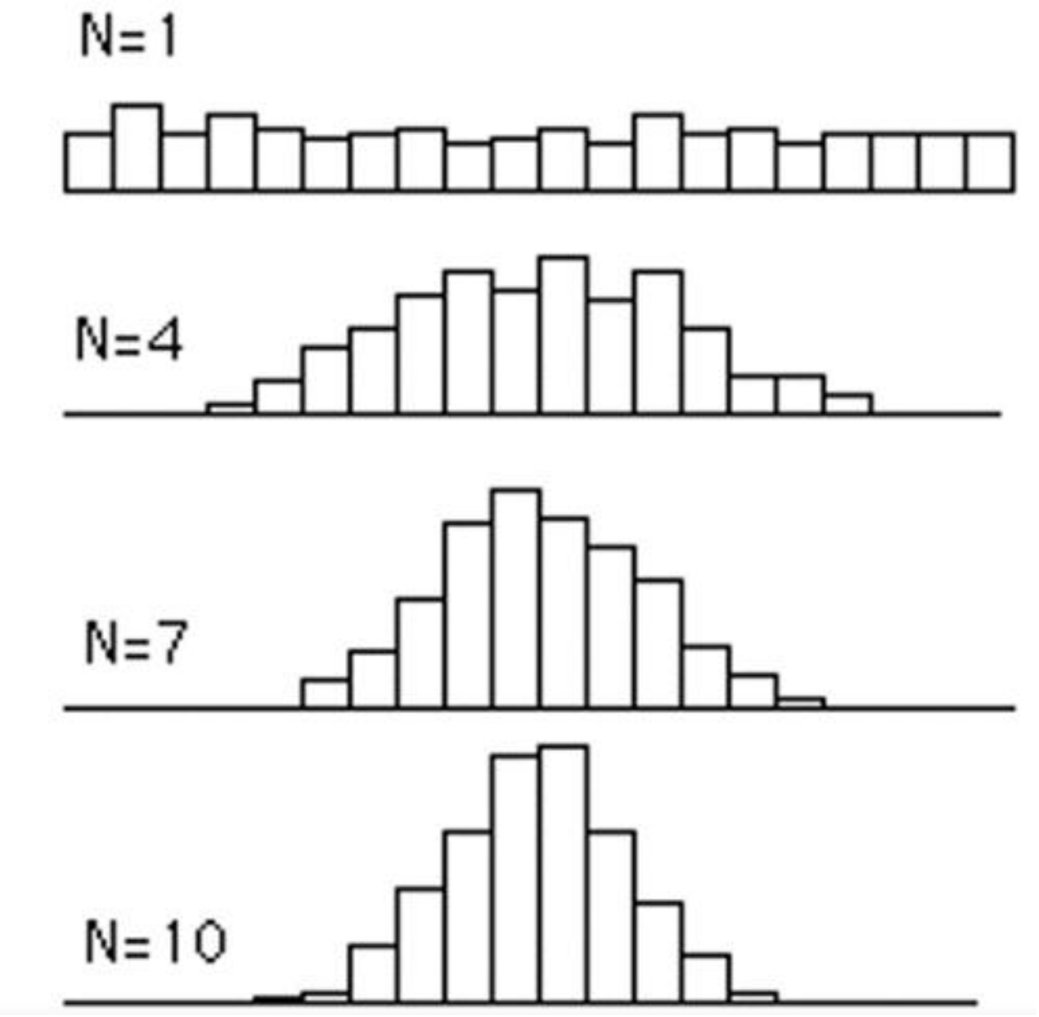
- **If we sample a population over and over, the set of means of all samples are normally distributed, regardless of the population distribution.**
- **The more samples, the closer to normal.**

$\bar{X}$=sample mean. $\bar{X} \sim N(mean, \frac{st.dev}{\sqrt{n}})$ $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

- μ is the population mean
- σ is the population standard deviation
- *n* is the sample size

Central Limit Theorem

1. We have a distribution that contains the means from 500 samples of our total population
2. For n = 4, 4 scores were randomly sampled, and the means computed. Same for n=7 and n=10.
3. We can see as n increases, distribution gets more normal



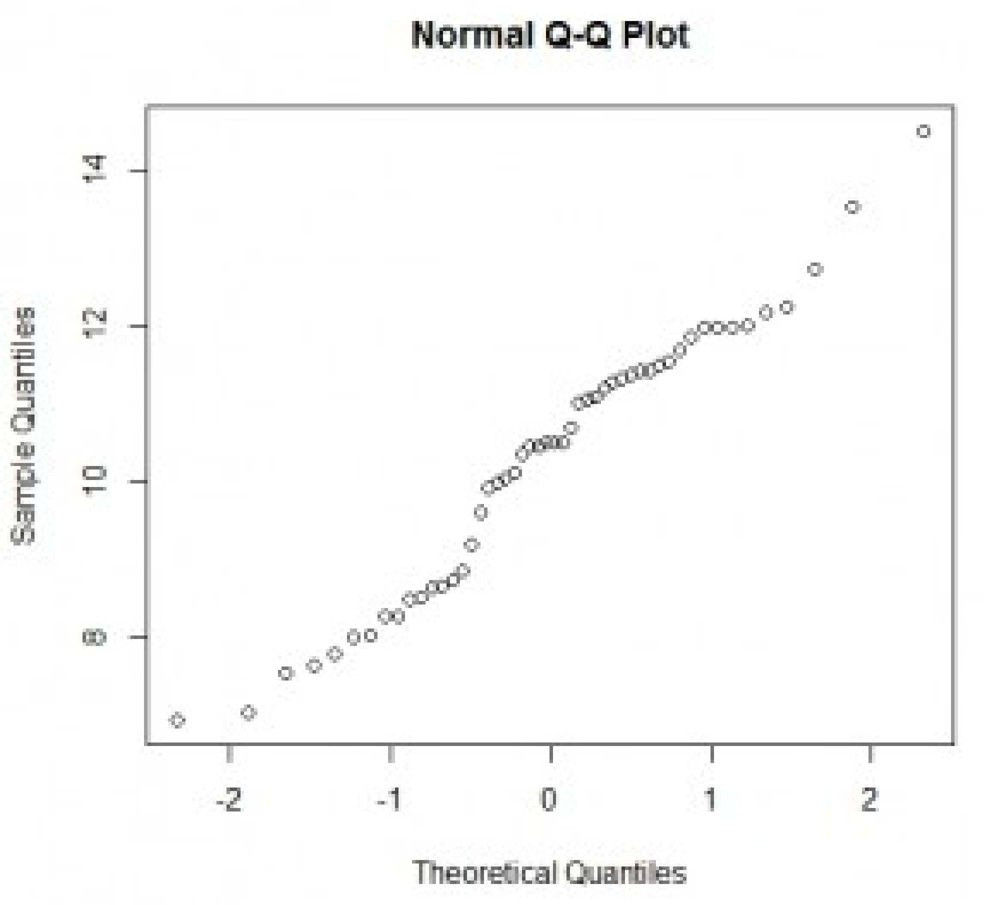
Central Limit Theorem

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- **We can use this central limit theorem to generate confidence intervals on expressing the population mean.**
- **We know the sample mean, sample variance, and number of samples.**
- **Then we know how our estimate of the population mean is distributed (from above formula).**
- **We can then generate 90%, 95%, … confidence intervals around our sample mean.**

QQ Plots

- Help assess if a set of data plausibly came from some theoretical distribution (normal)
- If both sets came from same distribution, you would see a straight line
- Quantiles = points in your data below which a certain proportion of your data fall.
- X axis = theoretical (normal) dist



Confidence Interval

C**entral Limit Theorem** ((**CI**) is a type of interval estimate, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter. The interval has an associated **confidence level** that, loosely speaking, quantifies the level of confidence that the parameter lies in the interval.

Confidence Intervals

Confidence intervals are a way to express uncertainty in *population* parameters, as estimated by the sample.

E.g. If we create a 95% confidence interval for the population mean, say

– Then we can say that the true population mean, μ, has a 95% chance of being between 5 and 15. $\hat{\mu} = \bar{X} = 10 \pm 5$

It is **not** correct to say:

- ~~"95% of the sample values are in this range."~~
- ~~"There is a 95% chance that the mean of another sample will be in this range."~~

Confidence Intervals

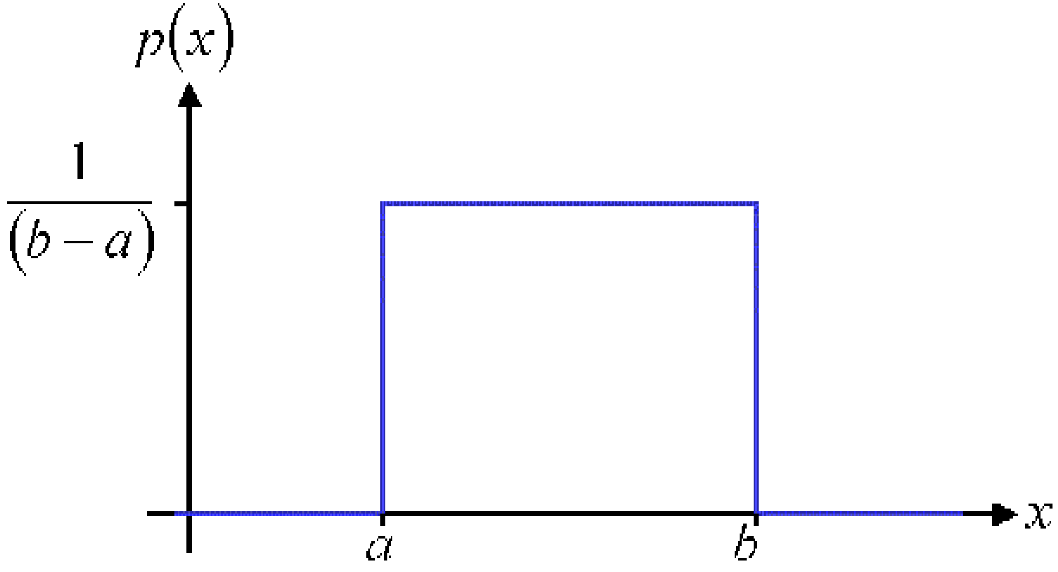
To create confidence intervals for population means, we use the central limit theorem and create confidence intervals based on the normal distribution.

- –Repeatedly sample from the population.
- –Calculate the mean for each sample.
- –Use the average of the sample means as the population estimate and create a C.I. based on the Std. Dev of the sample means.
- –R demo

Review: Uniform Distribution

- Uniform (flat, bounded)

$$P(x) = \begin{cases} \dfrac{1}{(b-a)} \; if \; a \le x \le b \\ 0 \; if \; x < a \; or \; x > b \end{cases}$$



- Constant probability
- Length of the base of the rectangle is ($b-a$), while the length of the height of the rectangle is $1/(b-a)$
- Area under rectangle = 1

Resampling Methods

Why use resampling?

- Allow computation of statistics from limited data
- Compute statistics from multiple subsamples of dataset
- Minimal distribution assumptions
- But it can be computationally expensive

Bootstrapping

- If we have a sample of 100 values (x) and we'd like to get an estimate of the mean of the sample:

```
mean(x) =  sum(x)/count
```

- Since our sample is small, the mean is not robust

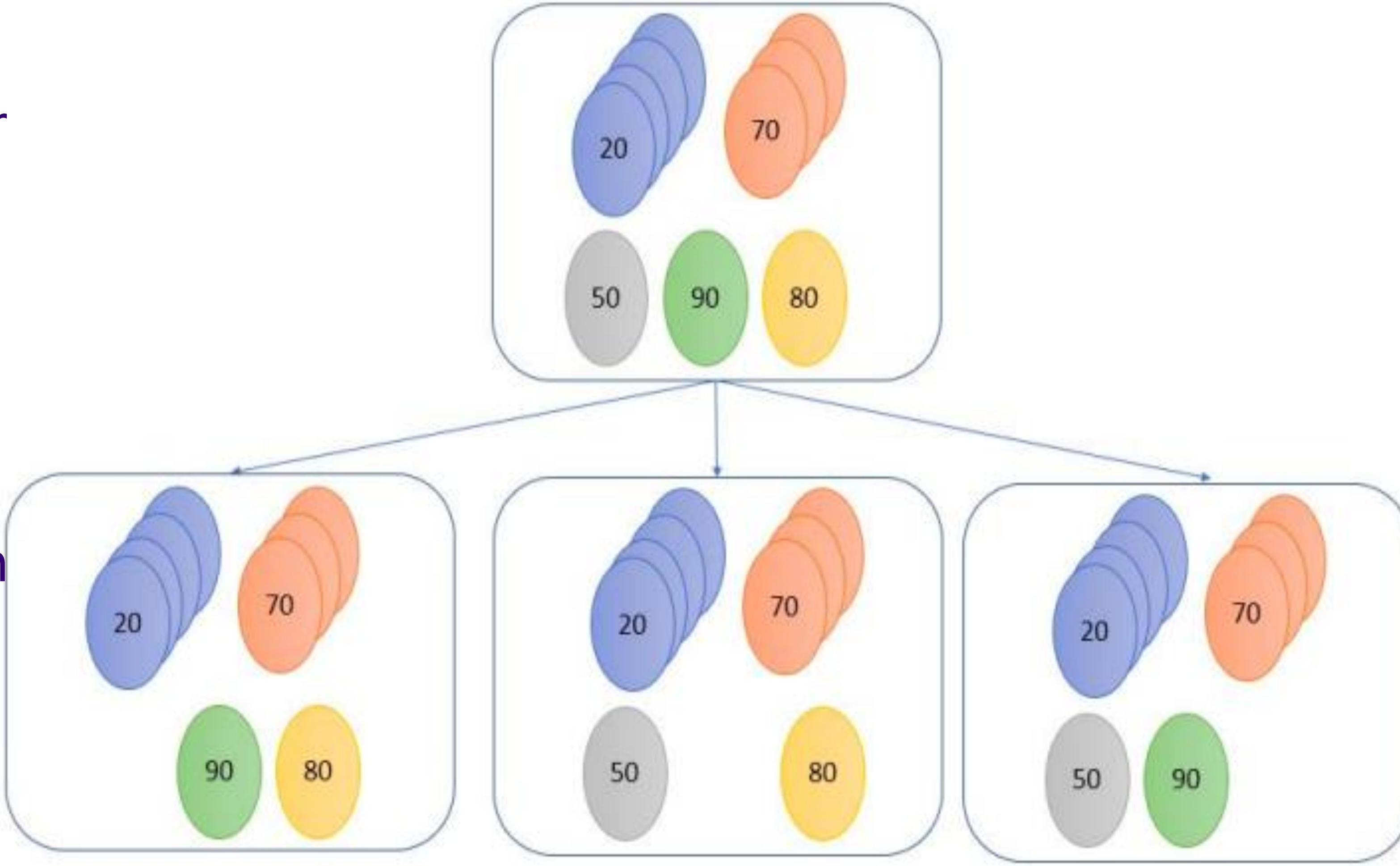
Bootstrapping

1. Create 1000 subsamples of our dataset with replacement
2. Calculate the mean of each subsample
3. Calculate the average of all the means we collected

- Can also use other measurements (SD, coeff, etc)

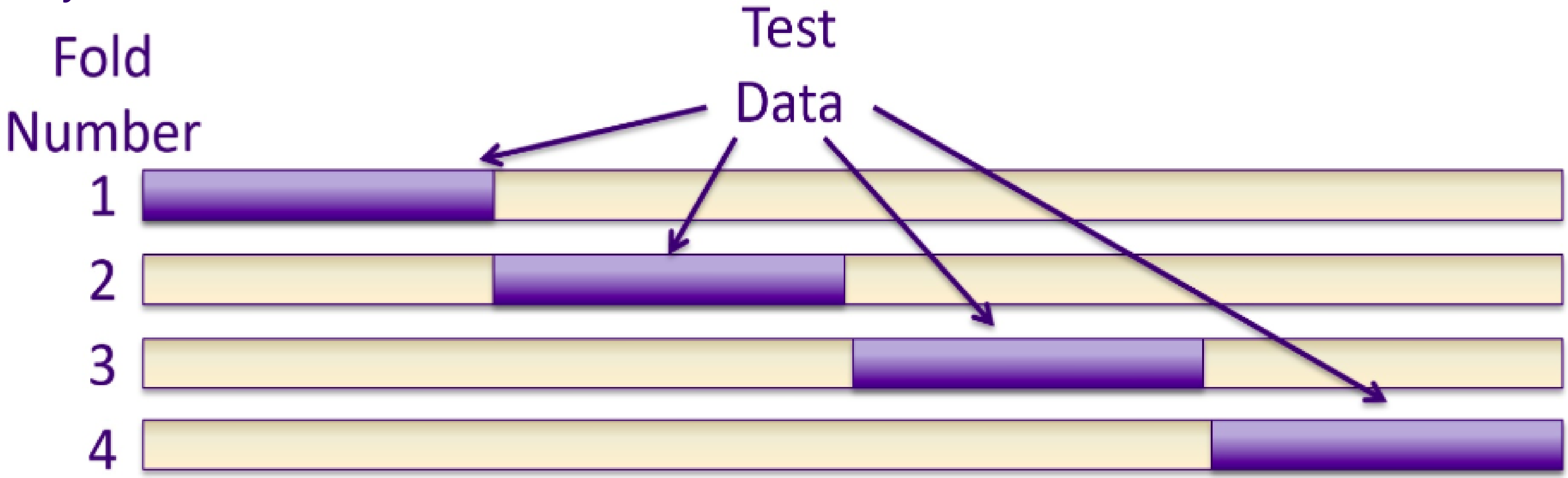
Leave some-out resampling: Jackknife

- Uses resampling to estimate the bias of a sample statistic
- Resamples systematically rather than randomly (like bootstrapping)
- Less computationally intensive than resampling
- Deletes each observation and calculates an estimate based on the remaining n-1 of them
  - Repeat for each observation in set



Leave out many resampling: K folds Cross Validation

- Basic idea is to split the training data into “k” independent pieces (called folds)
  - Train on (k-1) folds and test on the remaining fold
  - Repeat this “k” times, testing once on each fold
  - Average the model and performance metrics from each of these “k” runs
- Typically, k ~ 10



Bayes Theorem

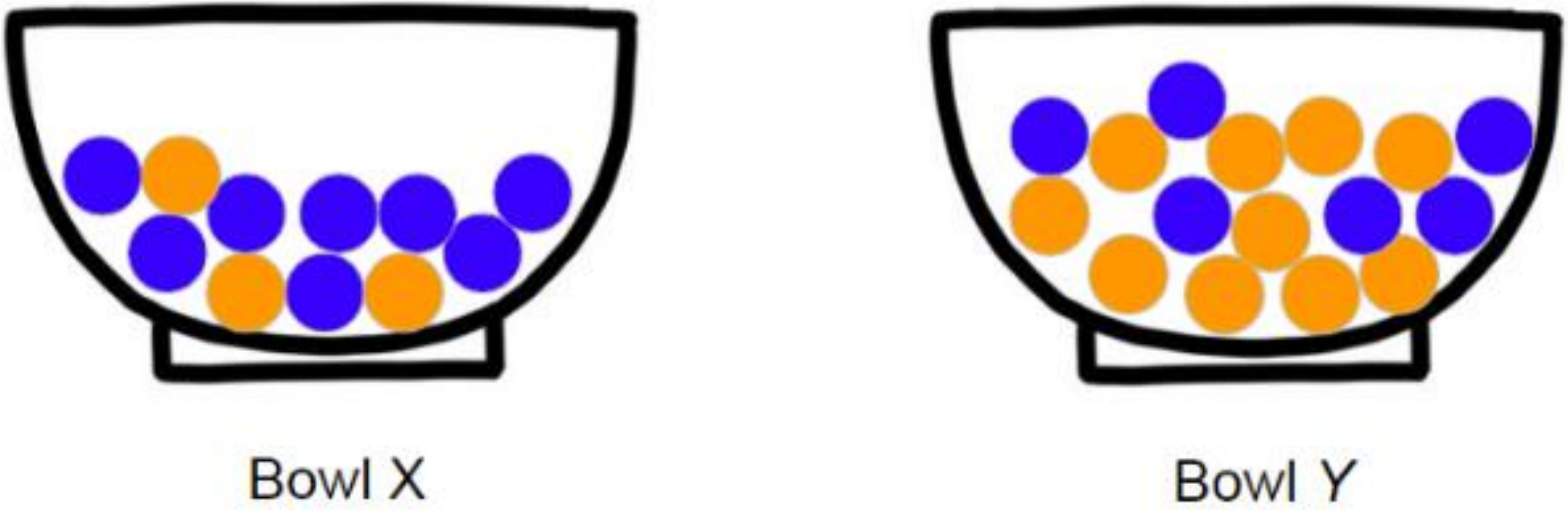
Describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have cancer.

Introduction to Bayesian Statistics

- Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available
- Widely used in science fields
- Subjective depending on how you calculate your prior knowledge

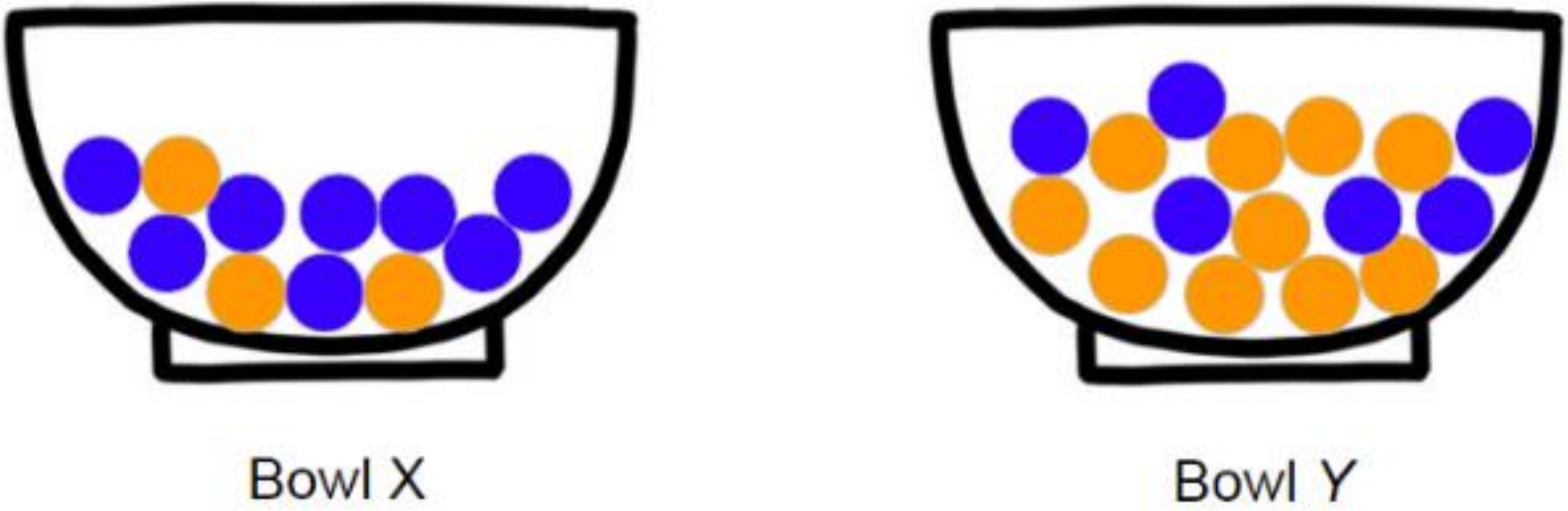
Introduction to Bayesian Statistics

Bowl X and Y are filled with orange and blue marbles



Introduction to Bayesian Statistics

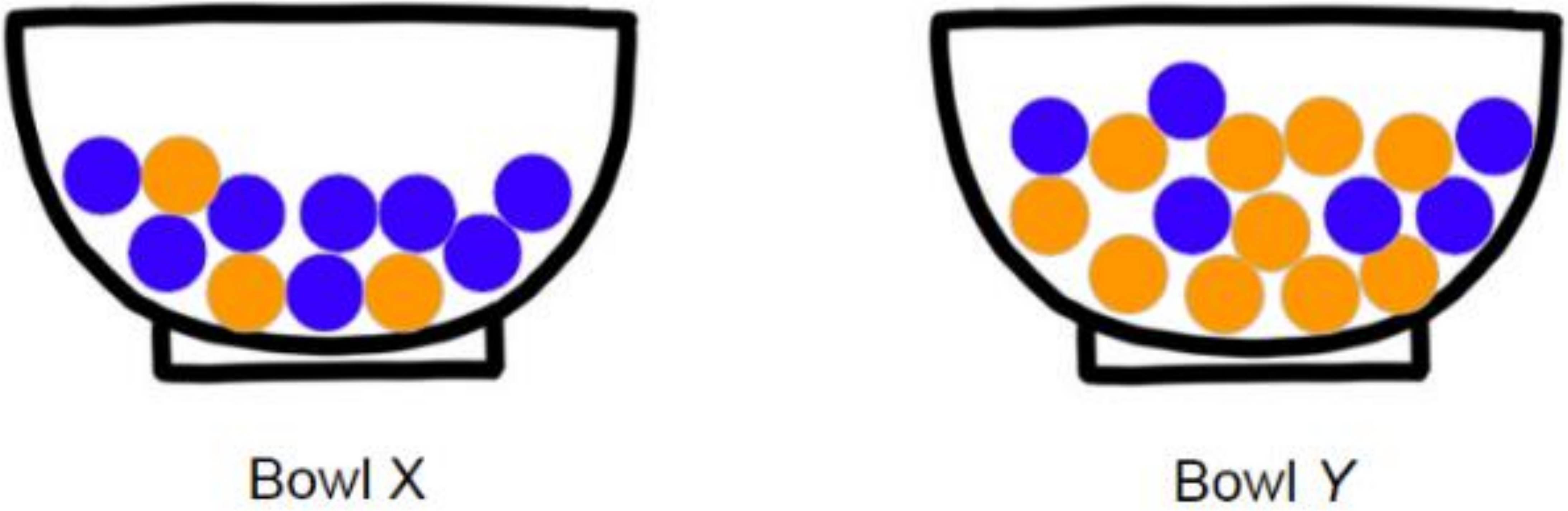
How likely is it to pick an organge marble from bowl X?



Introduction to Bayesian Statistics

11 items in bowl X, 3 of those are orange =

p(orange)= 3/11



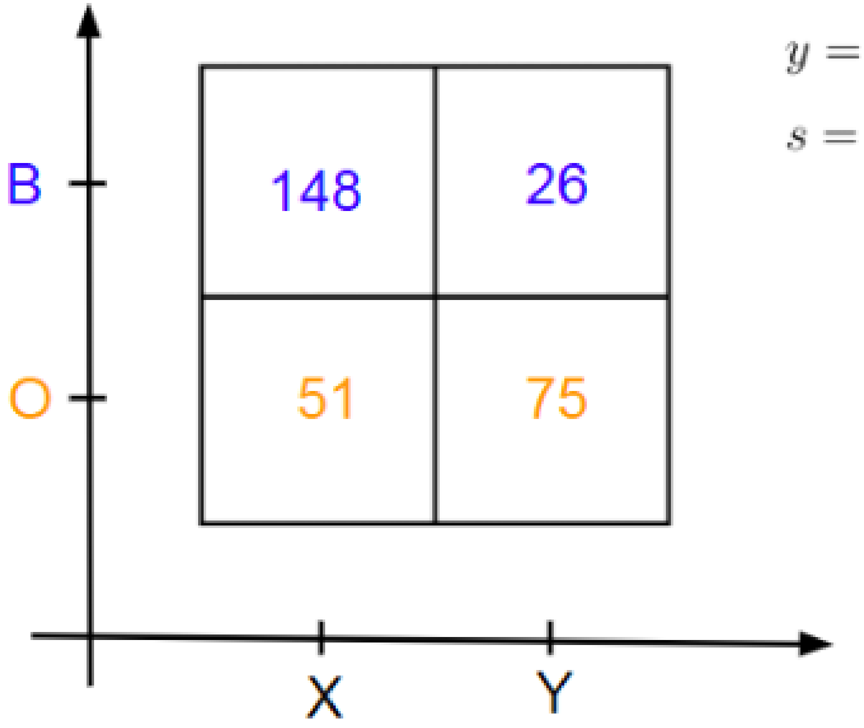
Introduction to Bayesian Statistics

Now suppose I asked you to predict which bowl the blue marble was taken from?

Introduction to Bayesian Statistics

If we pick a marble from each bowl randonly many times and write down which one gave us a blue marble, we start to gain information we can use

S = source (bowl)
Y = observable variable



$y = \{B, O\}$
$s = \{X, Y\}$

Bayes Theorem

The formula is:

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

Which tells us: how often A happens *given that B happens*, written **P(A|B)**,

When we know: how often B happens *given that A happens*, written **P(B|A)**

and how likely A is on its own, written **P(A)**

and how likely B is on its own, written **P(B)**

A Simpler Way to Write Bayes Law:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|not\ A)P(not\ A)}$$

$$P(A|B) \propto P(B|A)P(A)$$

Posterior Distribution

The Likelihood

Prior Distribution

Statistical Drama

What is the controversy?

- Bayesian methods use priors to quantify what we know about parameters.
- Frequentists do not quantify anything about the parameters, using p-values and confidence intervals to express the unknowns about parameters.

Remember Bayes Law:

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

Important points to make:

- Tests are not the event. We have a disease test, which is different than the event of actually having the disease.
- Tests are flawed. Tests have false positives and false negatives.
- Tests return test probabilities, not the event probabilities.
- False positives skew results.
  - E.g. If fraud is rare, then the likelihood of a positive result of fraud is probably due to a false positive

Interpretation with Modeling

Steps:

- Identify data relevant to the research question. E.g.: what are the measurement scales of the data? (Helps set uninformative priors)
- Define a descriptive model for the data. E.g.: pick a linear model formula.
- Specify a prior distribution of the parameters. E.g. We think the error in the linear model is Normally distributed as N$(0, \sigma^2)$.
- Use the Bayesian inference formula (above) to re-assess parameter probabilities.
- Optionally, iterate if more data is observed.

$$P(parameters|data) \propto P(data|parameters)P(parameters)$$

Bayes Theorem Example

Let us say P(Fire) means how often there is fire, and P(Smoke) means how often we see smoke, then:

P(Fire|Smoke) means how often there is fire when we can see smoke
P(Smoke|Fire) means how often we can see smoke when there is fire

So the formula kind of tells us "forwards" P(Fire|Smoke) when we know "backwards" P(Smoke|Fire)

Example: If dangerous fires are rare (1%) but smoke is fairly common (10%) due to barbecues, and 90% of dangerous fires make smoke then:

$$P(\text{Fire}|\text{Smoke}) = \frac{P(\text{Fire})\ P(\text{Smoke}|\text{Fire})}{P(\text{Smoke})} = \frac{1\% \times 90\%}{10\%} = 9\%$$

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

So the "Probability of dangerous Fire when there is Smoke" is 9%

Choosing Priors

- Uniform (flat)
  - consistent, flat prior value
  - Use when we have no expectation for the outcome
  - Flat prior p(θS, θN): every hypothesis (square) has equal probability
- Informed Prior
  - Experiment has already been run with priors
  - We saw what happened with our previous priors (aka our treatment has a stronger effect on Thyroid Cancer than on Pancreatic Cancer) and we use that information to weight the priors

Conjugate Priors

- You must make distributional assumptions about your data
- The supporting evidence you find is contained within the parameters of your distribution
- Integrating the actual prior for every distribution, especially for multi-dimensional models, is computationally impossible
- We can calculate priors and update our models with priors that are already know to be contained within the parameters of our model

Likelihood		Conjugate
Binomial	β	Beta
Bernoulli		Beta
Poisson	γ	Gamma
Categorical	$B(\alpha)$:	Dirichlet
Normal		Normal, Inverse Gamma

Credible Intervals

Frequentist Concept

- Data has one unknown true value
- Confidence Interval- range of values designed to include the true value

Bayesian Concept

- Parameter's value is fixed but has been chosen from some (prior) probability distribution
- Confidence Interval for an unknown (fixed) parameter θ is an interval of numbers that we believe is likely to contain the true value of θ
- If our confidence level is 95% and our interval is (L, U). Then we are 95% confident that the true value of θ is contained in (L, U) in the long run

Metropolis Hastings

In statistics and statistical physics, the **Metropolis–Hastings algorithm** is a Markov chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution from which direct sampling is difficult.