STATISTICAL INFERENCE

The process of generating conclusions about a population from a noisy sample.

The only formal system of inference to make predictions about the larger sample of data.

We will be working with frequency style of thinking. Repeat an experiment over and over again to get a probabilistic style study of the population.

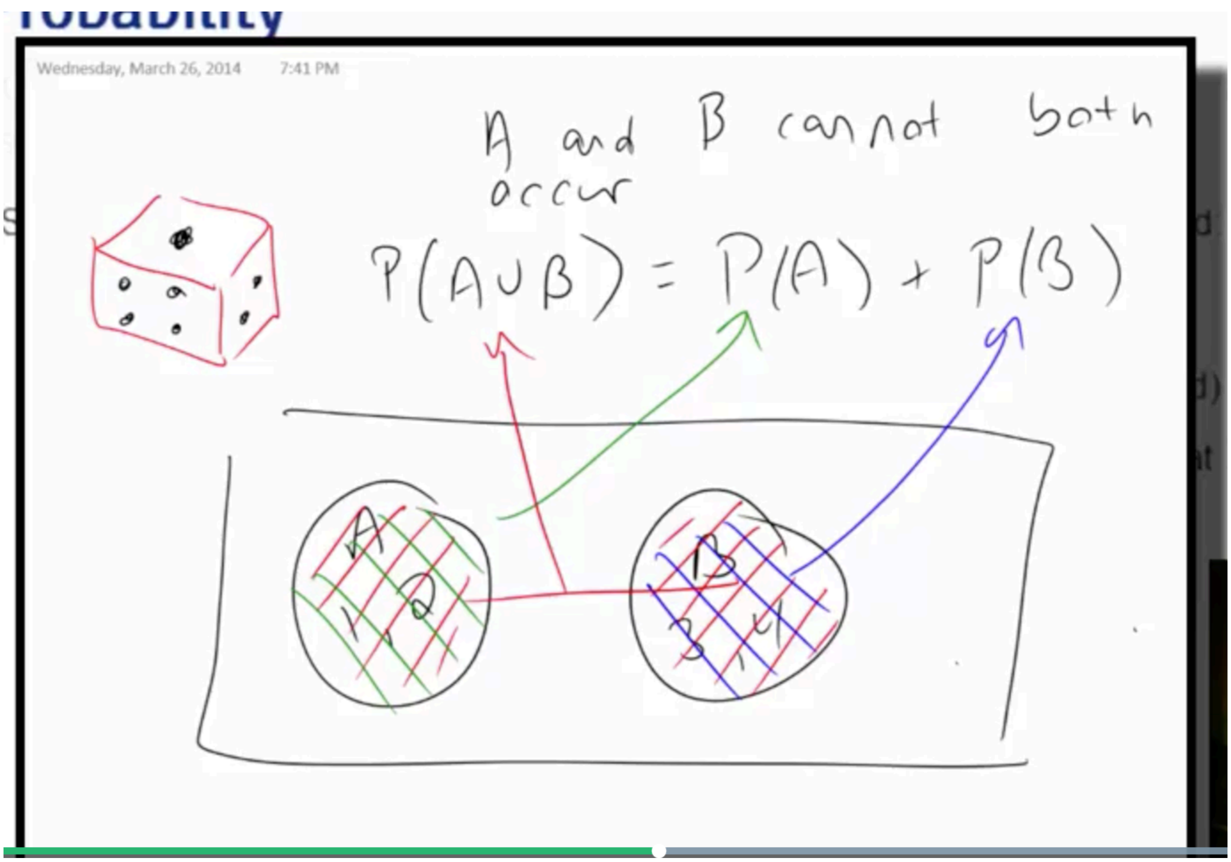
Questions such as studying causality rather than just looking at causation

Probability assigns a number between 0 and 1 to events to give a sense of the "chance" of the event.

MODULE 2 : Probability

A probability measure is a population quality that summarizes the randomness.

It operates on the outcomes from the experiment. Take in a set of outcomes and assign a number between 0 and 1. The union of the sets should be the sum of the probabilities of the sets.



Probability must follow some rules: P that nothing occurs is zero, p of something occurs is 0 minus 1. prob of something occurring is also 1 minus the p of the opposite occurring. P of mutually exclusive events is the sum of the events. for any two, P is the sum of A plus B - intersection.

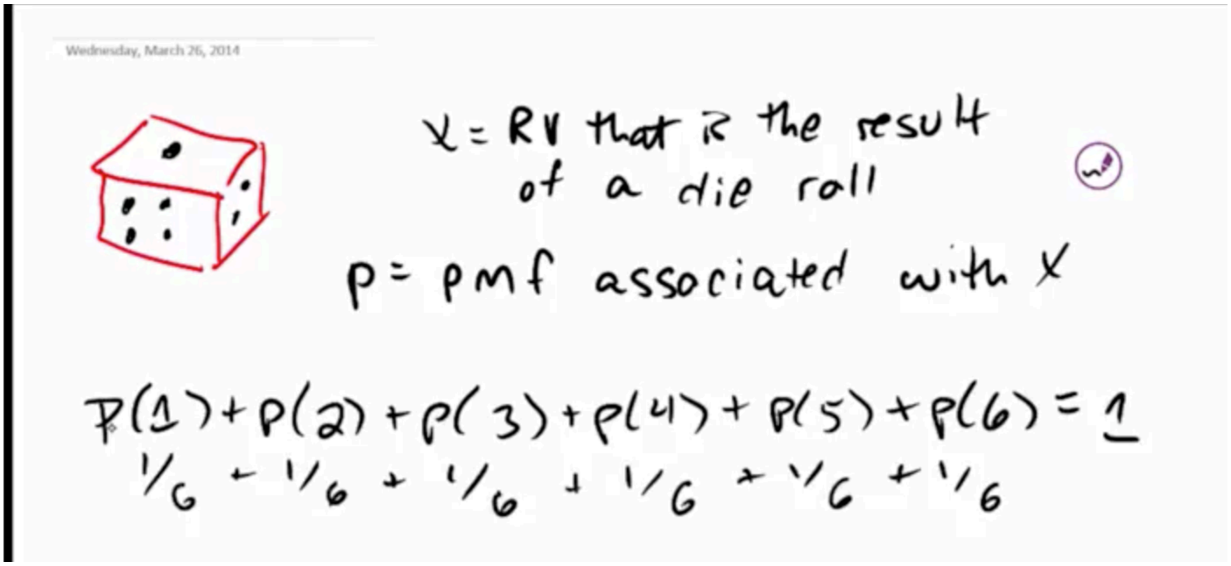
Probability Calculus

Foundation of probability rules. Densities and mass functions for random variables are the best starting point

A random variable, discrete or continuous , is the numerical outcome of an experiment.

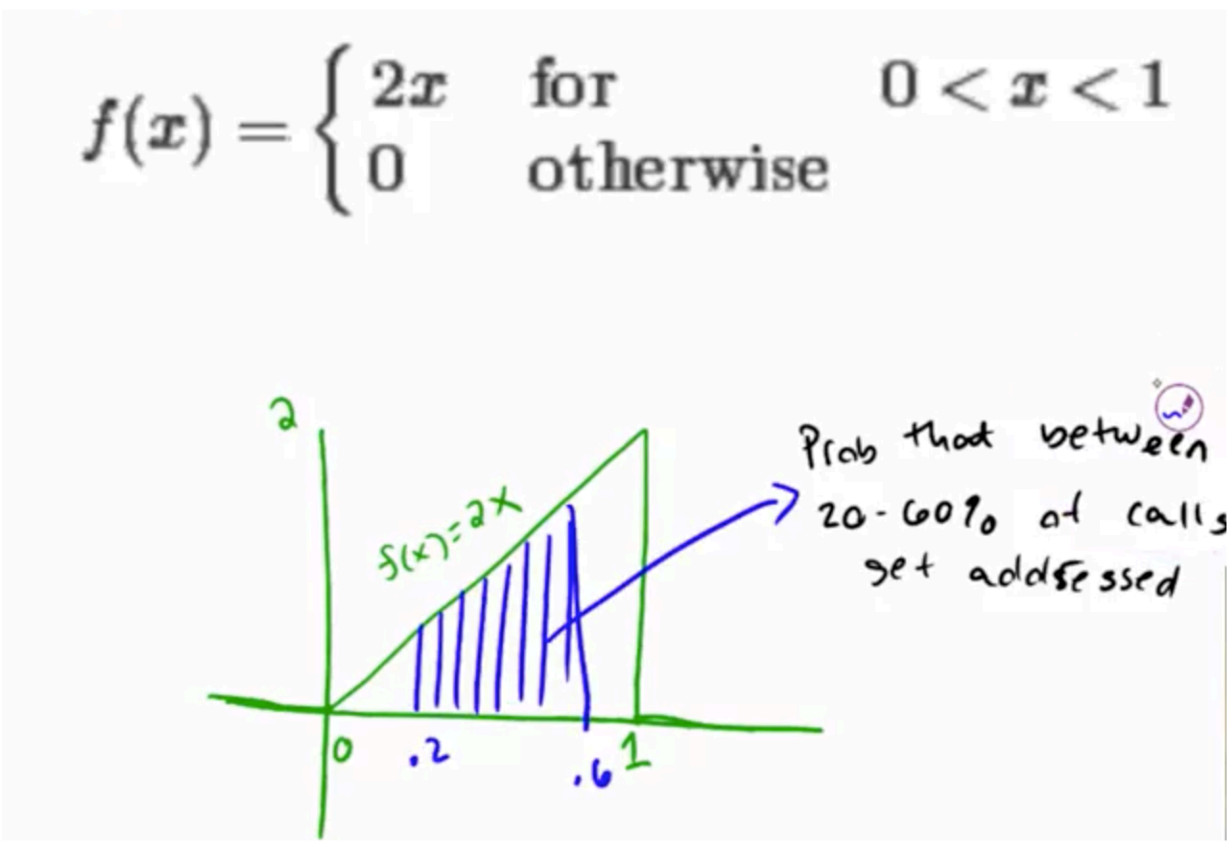
Discrete if its value is not in a range. Eg, flip of a coin or a roll off a dice.

PMF - Probability Mass function that takes in a function that a discrete random variable can take and assigns a value to it. Must always be greater than zero and the sum of all the random variables it can take must add up to one.



represent probability mass functions in terms of P(x) = O^x (1-O)^1-x   x, 0,1

Probability Density Function (PDF) , function associated with continuous function. Total area must be under one. The area must represent the prob of the event happening. Used for continuous data. for example bell shaped density curve .



For something to be a mathematically valid probability density function, the area must equal 1. the areas of the right triangles gives you an estimate of probabilities- beta distributions.

$ pbeta(0.75, 2, 1) - R function to get probability in a beta distribution. cumulative distribution function.

Quantiles - sample quantiles. Lie the observations from least to greatest and line the distribution

95th quantile, what is the value such that the the probability that a random variable drawn is lower than that point is 95% and greater is 5%.

$ qbeta(0.5, 2,1) - get the quantile value for a data.

sample median is estimating the population median- estimand, sample median is the estimator.

MODULE 3 - CONDITIONAL PROBABILITY

Conditional probability - death rate for lightning strikes for people standing near a tree when a lightning strike is 1 in 6 even though the overall is 1 in 7mil

Let B be an

P(A|B) = P(A intersection B) / P (B) ..
if independent then P(A).P(B) / P(B) — if in related simply P(A).

$A = \{1\} \qquad B = \{1, 3, 5\}$

$$P(\text{one given that roll is odd}) = P(A \mid B)$$

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

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

$$= \frac{1/6}{3/6} = \frac{1}{3}$$

BAYES Rule ::
Allows us to reverse the roles . You can relate B given A, instead of A/B

Diagnostic Tests : Sensitivity is the P that the test is positivity if the subject has the disease.
Specificity is when they don’t have the diseases and the test is negative.

Positive predictive value - necessary when you have the disease.
Negative predictive value -useful when you don’t have the diseases.

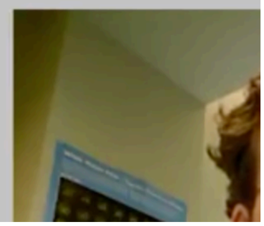
Using Bayes' formula

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)}$$

$$= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + \{1 - P(- \mid D^c)\}\{1 - P(D)\}}$$

$$= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999}$$

$$= .062$$



$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}$$

You can get the p(B|A) if you know the P(A|B) and other diagnostic values .

Independence : P(A) and P(B) are independent if P A intersection B is P(A) X P(B) . meaning not all probabilities can be multiplied. Events that are independent should not be multiplied.
iid sampling - used as a conceptual mode to analyze data .

MODULE 4 : EXPECTED VALUE

Expected values are the process of making conclusions about population from noisy data.
Assume that population and randomness is given my densities and mass functions.

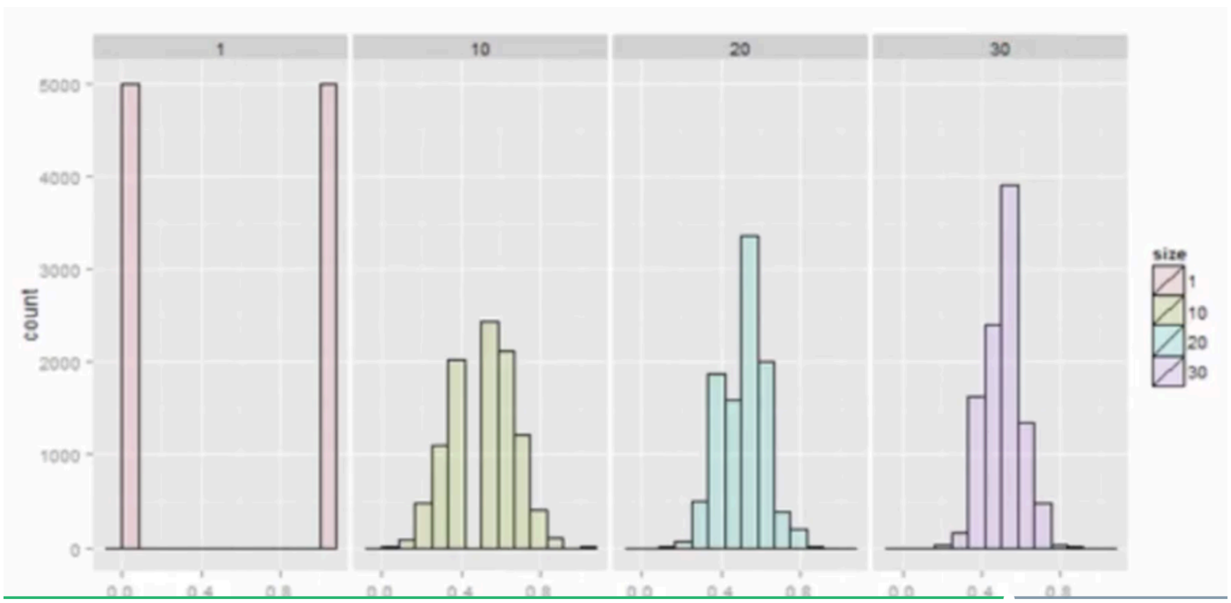
Can be seen as the mean of a random variable at the centre of its distribution.
Variance is how spread out the data is.

Height is P(X). Idea of centre of mass. Population mean founding using the centre of mass of for the sample mean.

$ manipulate(myHist(mu) , mu = slider(62, 74, step = 0.5)

expected value is the properties of the population, centre of mass of the distribution.

For continuous random variable: thinking about how you are going to divide a dense object (cutting a piece of wood) to find the mean.
The population distribution of the sample mean is centered in the same place as the population mean, when this happens this is estimator is called an unbiased estimator.



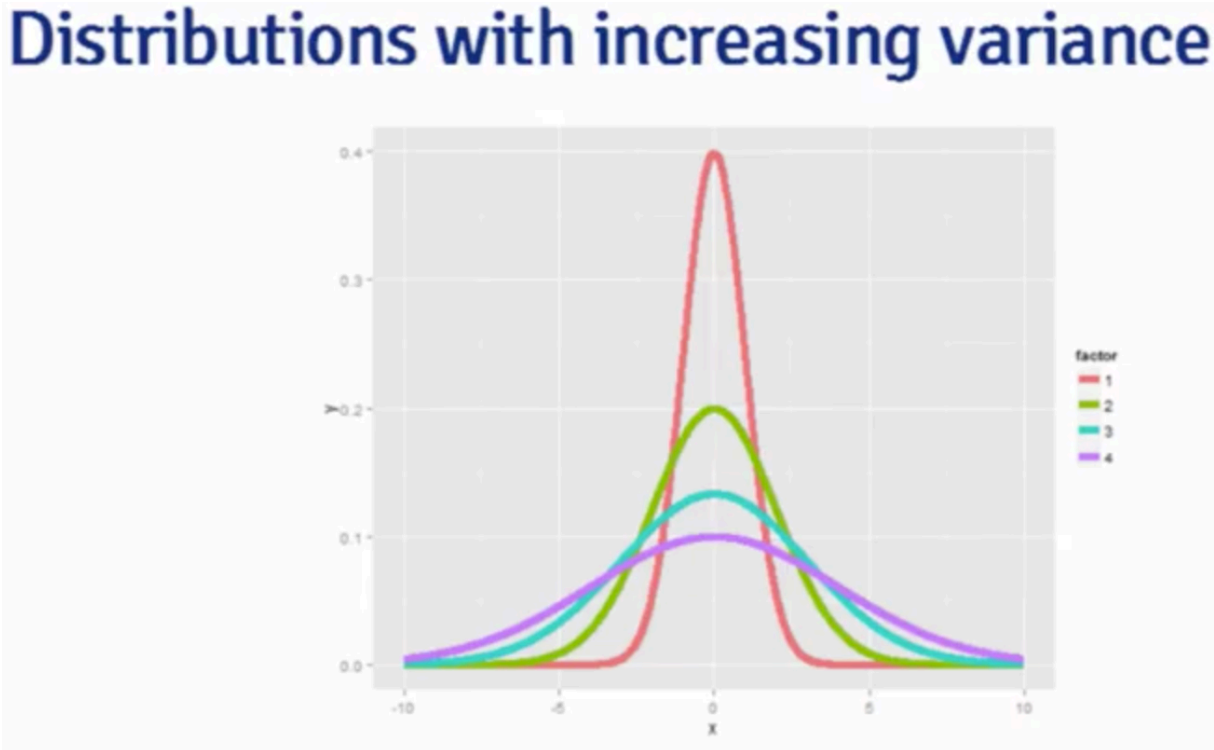
As more and more random sampling is done, the centre remains the same and the distribution gets smoother. more concentrated densities, more gaussian-ish.

MODULE 5 VARIABILITY

If mean is the where the distribution is centered at variability is how fat or how thin or how spread out the density is around the mean.

$$Var(X) = E[(X-\mu)^2] = E[X^2] - E[X]^2$$

Sq root of variance is the standard deviation. Higher variance is more spread out.



Sq of distance from the mean.

Population variance and sample variance are analogous. divide by n-1 rather than 1.

$$Var(X) = E[(X-\mu)^2] = E[X^2] - E[X]^2$$

Sample variance :

$$S^2 = \frac{\sum_{i-1}(X_i - \bar{X})^2}{n-1}$$

Expected value of the population variance is what the sample variance is trying to predict.
if enough of the samples are taken, its variance will be the same as the population , same as the concept of mean. will be centered around the same but it

get concentered at the centered.

Standard Error

Variance of the sample mean becomes zero as we get closer to the real data , we can estimate sigma2.
Variance of the distribution of sample mean:

$$Var(\bar{X}) = \sigma^2/n$$

$$S/\sqrt{n}$$

- standard error of the mean . S is the standard deviation of how variable your population is.

Poission - discrete random variable with a distribution of 4.
Simulations are run look at how variations work.
for coin flip distribution :

$$1/(2\sqrt{n})$$

MODULE 6 DISTRIBUTIONS

Bernoulli Distribution: Rises out of coin flips, only takes in random values 0 and 1. mean is p and variance is (1-p) .

Binomial Distribution: Bunch of sums of bernoulli values. for biased coin.
example

If each gender has an independent 50% probability for each birth, what's the probability of getting 7 or more girls out of 8 births?

$$\binom{8}{7}.5^7(1-.5)^1 + \binom{8}{8}.5^8(1-.5)^0 \approx 0.04$$

$ pbinom - get probabilities for binomial probabilities in R.

Normal Distribution:

Gaussian distribution with mean μ and variance σ^2

$$(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$$

mean is zero and std dev and variance are one. total area within 1 sd = 68%, 2 sd (-2 to +2) 95%, leaves 2.5% on each tail . 3 sd = 99% of the mass.

Normal can be taken to standard and normal .

If $X \sim \mathrm{N}(\mu, \sigma^2)$ then

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

If Z is standard normal

$$X = \mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$$

R formula for normal distirubution.

```
qnorm(0.75, mean = 1020, sd = 50)
```

Poisson Distribution: {{ REVIEW More }}
mean is lambda
variance is also lambda

Used for modeling unbounded count, event-time survival data.
standard mass distribution function :

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

MODULE 7 : Asymptotics - check CI again

Behavior of the statistics as sample size limits to infinity. Frequency interpretation of probabilities.

Limits of Random Variables

Law of large numbers - the average limits to what its estimating , the population mean. An estimator is consistent if it converges to what you want it to estimate. Law of large numbers, sample mean of iid (identically distributed random variables) samples is consistent with the population mean

The central limit theorem, *** very imp
Sates that the distribution of iid samples becomes standard normal as sample size increases.

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}$$

Replacing the population std deviation with the sample std deviation does not change the results.

MODULE 8 : Confidence Intervals

T-Distribution: The t distribution has thicker tails, centred around 0 with one parameter, the degrees of freedom. if the formula for gaussian distribution is divided by something, the no longer becomes gaussian , so when

Interval is $\bar{X} \pm t_{n-1} S/\sqrt{n}$ where t_{n-1} is the relevant quantile

Different from a normal distribution but difference is seen in lower degrees of freedom.