Unit 1: Probability Theory

Probability

Introducing Probability

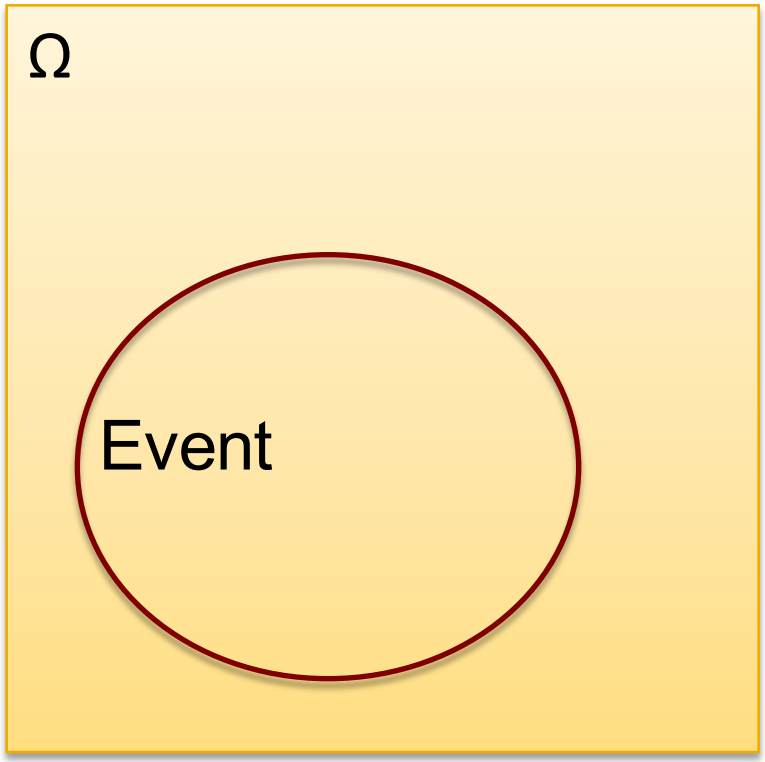
- Statistics is the study of data that is influenced by chance
  - The data we collect are almost never deterministic
  - Variation is implicit in the processes we study, introduced by sampling, or introduced by our measurement technique
  - We always have uncertainty, and our conclusions are never definite.
- Because of this, we're going to use probabilities over and over in this course.
  - Researchers have different views of what probability means – and we'll discuss those soon
  - For now, let's review the basic mechanics of probability – the rules that govern how it works.

Probability Spaces

Mathematicians define a probability space as having 3 components:

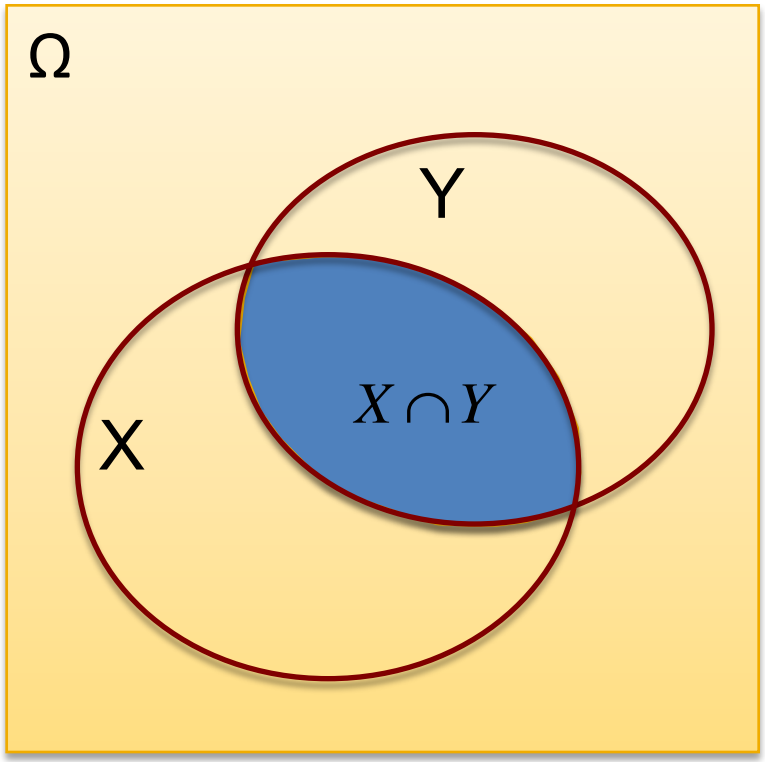
1. A set of possible outcomes, Ω.
   - These could represent the possible results of an experiment or study, or states of nature.
   - Every instance of a real world situation produces exactly one outcome.
     - If two hypothetical runs of a study differ in any way that we care about, they must be different outcomes.
     - If we roll a regular die, the possible outcomes might be Ω={0,1,2,3,4,5,6}
2. A set of events, F.
   - These are subsets of Ω.
   - An event could be a single outcome, in which case we call it an elementary event, or it could include multiple outcomes.
   - In the dice example, events include rolling a 6, rolling an odd number.
3. A probability function, *P*. That assigns probabilities to each event, and it has to satisfy axioms of probability.

The probability space is the triple, (Ω, F, P)



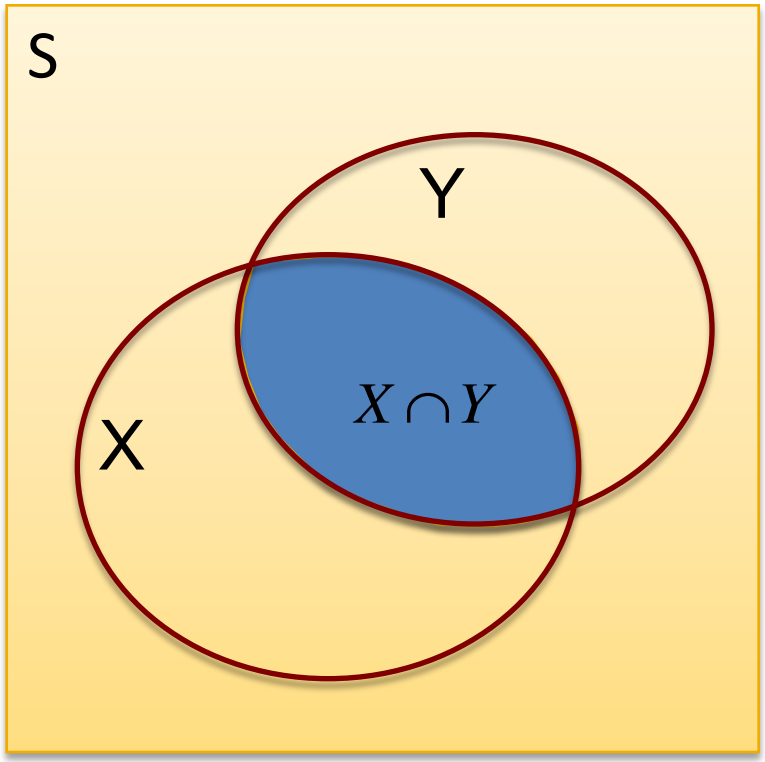
Area Diagrams

- To make probability spaces more intuitive, we often depict them using area diagrams.
  - Imagine that Ω is a dart board, and each bit of the surface is just as likely to be hit as any other bit.
- Then a probability can be expressed as a ratio.
  - The probability of event X is the area of X over the area of Ω, since all we know is that the dart hit the board somewhere. But we usually assume that the area of Ω is 1, to keep things simple.
    - Then the probability of X is just the area of X.
- The axioms of probability follow the same mathematical structure as areas
  - For example, the addition rule: If we want the probability that X or Y occurs, the probabilities add up just like the areas would.
  - $Pr(X \cup Y) = P(X) + Pr(Y) - Pr(X \cap Y)$



Conditional Probability

- What if we learn some information? Let's say we observe that event Y occurs, and then we want to know the probability of X. This is the conditional probability, Pr(X|Y).
  - Looking at the board, we know the dart hit somewhere in Y, and we want to know the probability that it also landed in X. This is the shaded region.
  - Again, this is the ratio of areas. The area of X and Y over the area of Y.
    - And that's the same as the probability of X and Y, over the probability of Y
  - We write this as Pr(X|Y) = Pr(X and Y) / Pr(Y)
  - Notice that we didn't include the part of X that doesn't fall in Y. We already know the dart isn't in this region!

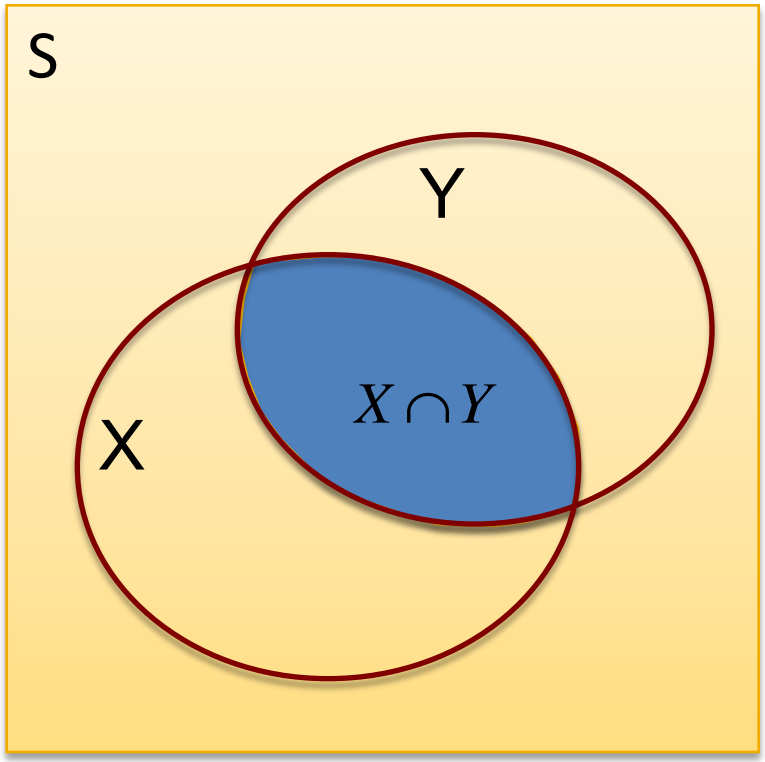


Bayes Rule (to be added to Week 2, right before Embracing Subjectivity)

Bayes' Rule

- Before we take a closer look at Bayesian statistics, we need to review Baye's rule
- This is the famous property of probability that underlies Bayesian statistics.
- To derive Baye's rule, we can take the definition of conditional probability and rearrange it to get:

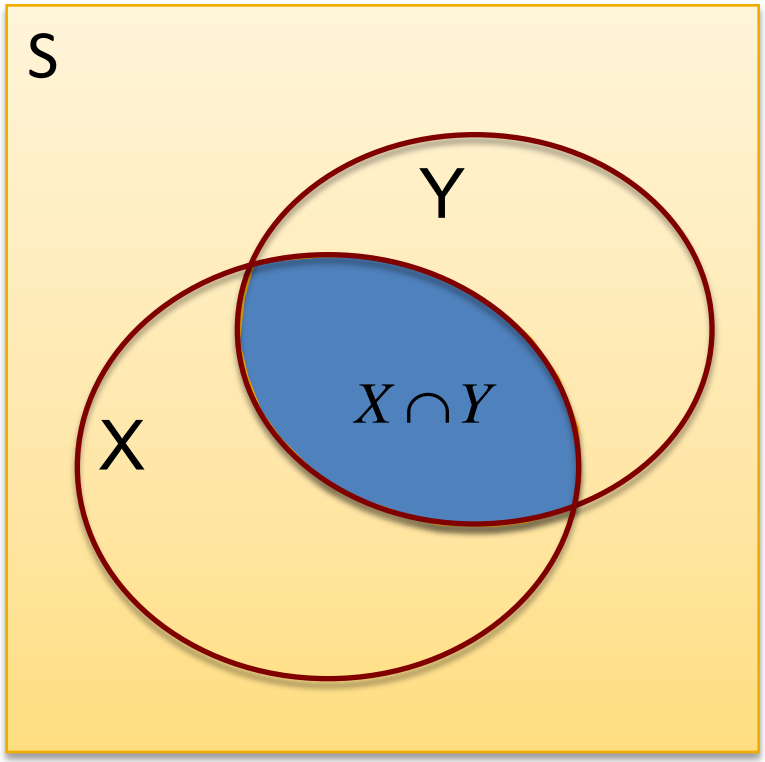
  Pr(X and Y) = P(Y) Pr(X | Y)
  - In other words, the probability that our dart lands in the intersection is the probability that it lands in Y, times the probability that it lands in X, given that it already lands in Y.
  - This is called the multiplication rule.
- Notice that we could have done this the other way around, first assuming that X occurs, then adding the fact that Y occurs.
- Pr(X ∩ Y) = P(X) Pr(Y | X)
- You can combine these two equations to get
- P(X) Pr(Y | X) = P(Y) Pr(X | Y)
- Or P(Y|X) = Pr(X | Y) Pr(Y)/ Pr(X)
  - This is a famous relationship known as Bayes' Rule
  - It relates P(X|Y) to its inverse, P(Y|X)
  - If you ever forget it, it's easy to derive.



Bayes' Rule Quiz

- $P(Y|X) = Pr(X \mid Y)\, Pr(Y) / Pr(X)$

- Suppose that 0.2% of the US population suffers from coca colic.
- There's a test for the disease, but it gives a false positive 10% of the time for healthy individuals, and a false negative 10% of the time for sick individuals.
- Suppose you take the test and the result comes back positive.
- What is the probability you have coca colic?



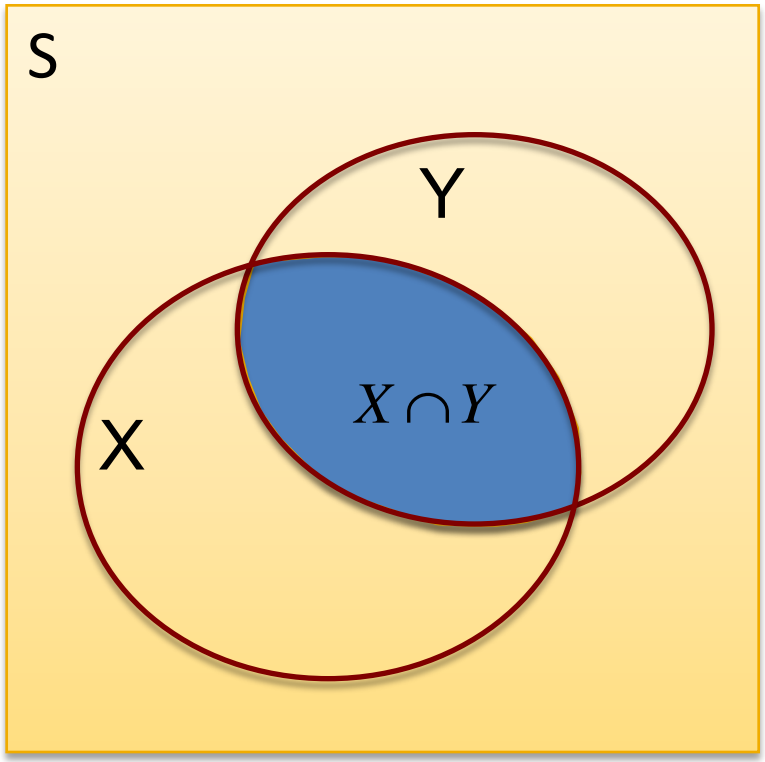
- Hint (should be hidden until students click something) Let X be the event that the test comes back positive. Let Y be the event that you have coca colic.
- Ans: 0.01771653543 (or .018, .0177, 1.77%, etc)

Quiz Solution (to be added to Week 2, right before Embracing Subjectivity)

Static text card (no video)

Bayes' Rule Quiz Solution

- Let X be the event that the test comes back positive. Let Y be the event that you have coca colic.
- The quantity we want is P(Y|X)
- We get this by writing down Bayes' rule:
- P(Y|X) = Pr(X | Y) Pr(Y)/ Pr(X)
- To get the denominator, we note that X has two parts, the part in Y and the part in !Y. Here, the ! means NOT – the event not Y.
- Pr(X) = Pr(X ∩ Y) + Pr(X ∩ !Y)
- Now, each of these terms can be expanded using the mutiplication rule:
- Pr(X) = Pr(X | Y) Pr(Y) + Pr(X| !Y) Pr(!Y)
- This formula has a special name: the law of total probability.
- Plugging this back into Bayes' rule, we get the answer
- P(Y|X) = Pr(X | Y) Pr(Y)/ [Pr(X | Y) Pr(Y) + Pr(X| !Y) Pr(!Y)]
- = .9 * 0.002 / (.9 * 0.002 + .1 * .998)
- Ans: 0.01771653543 (or .018, .0177, 1.77%, etc)



Random Variables

Reading Appendix B1 – B2

Introducing Random Variables

- The concept of a random variable is very important in statistics.
- The correct mathematical definition of a random variable is tricky to understand, so we won't define it here.
- Intuitively, we're talking about something that doesn't have a fixed numerical value – it varies probabilistically.
  - Every time you measure the heart rate of a patient, you'll get a different number.
  - Every day you get a different number of visitors to your website
  - If you collect a sample of people and measure their mean height – you only get one number. But if you imagine running the study over and over, you'd get a different mean each time.

- Instead of defining random variables in general, which is tough, we'll talk about the two most important types of random variables:
- Discrete random variables
- Continuous random variables

Discrete Random Variables

A ***discrete random variable*** X takes on values that form a discrete set in the real numbers.

- It can be defined as a discrete set of possible values, O, and a probability function, P_X:O -> [0, 1], that takes each value in O to its probability.

A discrete set in the real number line

- What's a discrete set? Roughly speaking, it's a set of disconnected points, no continuous intervals.
  - Any finite set of numbers is a discrete set. So we could represent a perfect die as a random variable that takes on the values O = {1,2,3,4,5,6}, each with probability 1/6.
  - We could represent a coin toss as a random variable with values O={0,1}, where 1 represents Heads or success, and it has probability p. This is a very common tool and it's called a Bernoulli distribution
  - Some infinite sets are discrete too. For example, the integers. We could represent a respondent's number of children as a random variable. Here, O = {0,1,2,…}, with some fixed probability for each number of children.

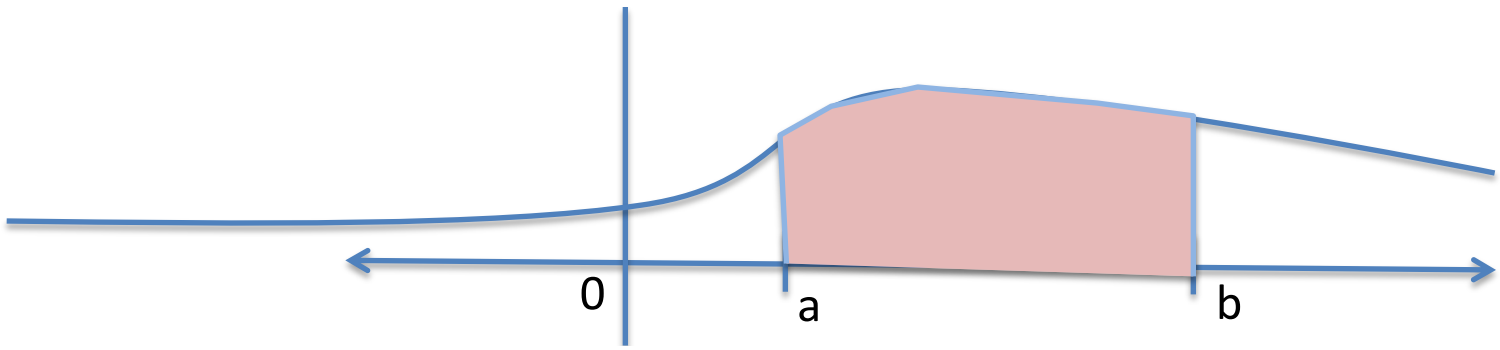
Continuous Random Variables

- A ***continuous random variable*** X takes on values in a continuous interval or set of intervals, but the probability of any one value is zero.
- Consider the length of a new squid species in meters. This could take on any positive value, but it's impossible to find a specimen that's exactly 2.5 meters long – the probability is zero.
- Since individual values have zero probability, we can't write down a probability function for X. Instead, we can define X in terms of a probability density function, *f*.
- *f* defines a differential probability, probability per unit on the real number line.
- To find the probability that X falls in some interval, [a,b], we integrate *f* from a to b.

$$P(a \le X \le b) = \int_{x=a}^{b} f(x)\,dx$$

- We can also define the cumulative probability distribution, F(x), as the area under *f* from -∞ to x.

$$F(x) = \int_{x=-\infty}^{b} f(x)\,dx$$



Manipulating Random Variables

- We often have to standardize distributions, which requires two operations:
  - Addition of a constant
  - Multiplication by a constant
- Addition of a constant can be easily defined for a random variable:
  - Given random variable X with probability (or probability density) function *f,* and constant c, X + c is the random variable with probability (density) function

$$f'(x) = f(x - c)$$

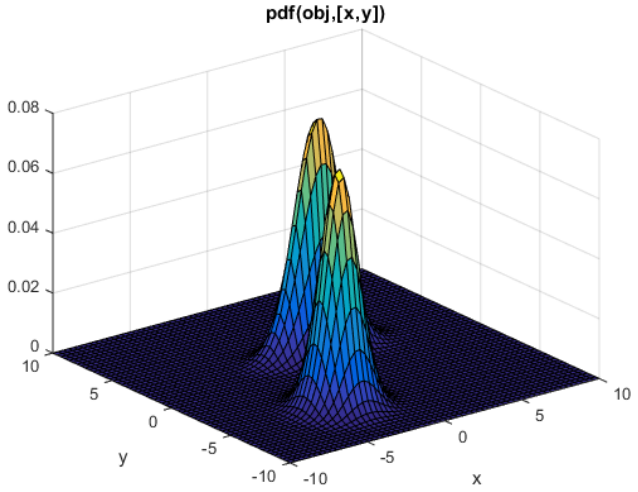
  - We just shift the values, but the probabilities are the same.
- Multiplication by a constant works the same for a discrete random variable, just multiply each value and leave the probabilities unchanged.
- If X is continuous, multiplication is a bit tricky, because it stretches the probability density function
  - we have to rescale it so that the total probability is still one.
  - If X is a continuous random variable with probability density function *f*, cX has the probability density function

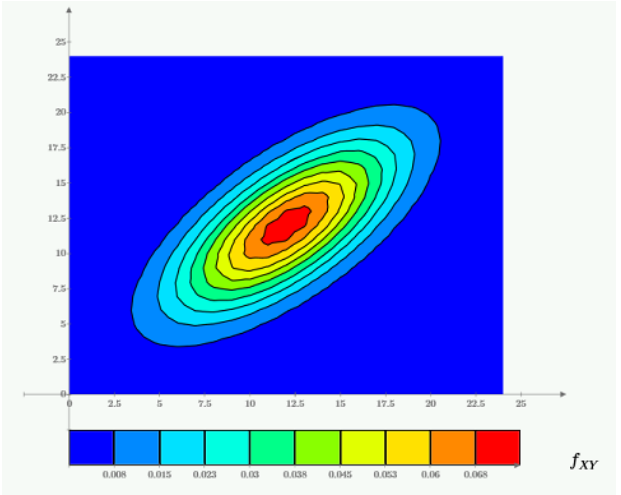
$$f'(x) = \frac{1}{c} f(x / c)$$

Joint Distributions

Joint Distribution

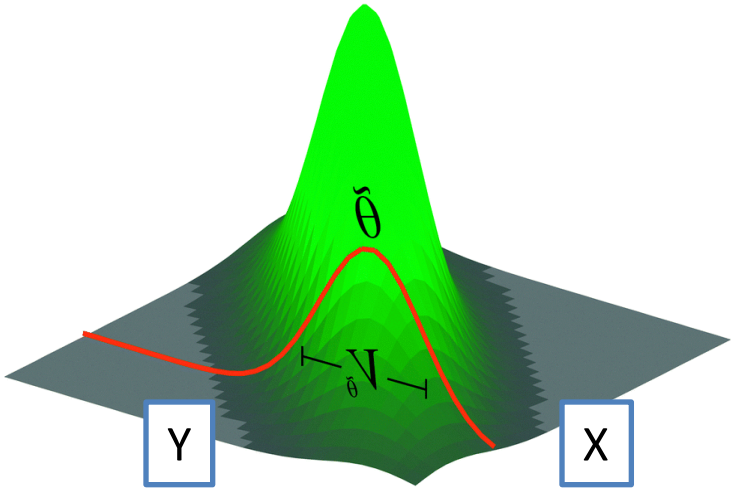
- When we have two random variables, X and Y, things get more complicated. We have to specify how probable each *combination* of values is.
- For continuous random variables, this is done with a ***joint probability density function***, $f_{X,Y}(x,y)$, which gives the probability density at X=x and Y=y.
  - We would integrate this function within a region of the x-y plane to find the probability that our variables are located there.
- We can do something similar for two discrete random variables, but we'd define a joint probability function, which gives the actual probability, P(X=x, Y=y).
- We could visualize a joint distribution using a 3D plot, or using a heatmap, as shown.
- A joint distribution can have some intricate features:
  - Dependency between variables. Perhaps x is likely to be high when y is high, and low when y is low.
  - Elliptical peaks, elliptical peaks in a diagonal direction, multiple peaks, etc.





Conditional Distribution

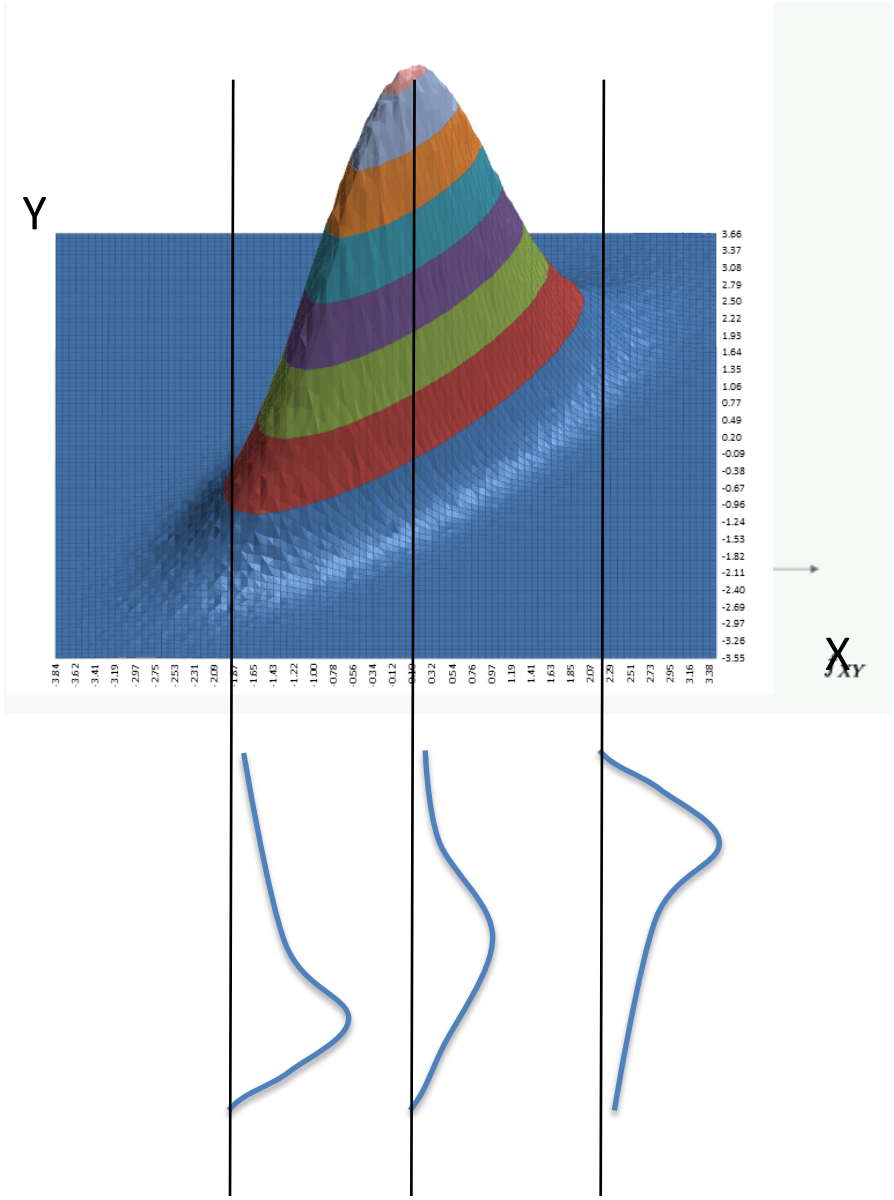
- Let's say we have two random variables, X and Y.
  - You might want to think about Y as our outcome variable.
- If we know the value of X, say X=x, we get more information about Y.
  - Looking at the joint distribution, we know that we're located on a particular line– the line X=x.
  - This is like taking a slice, or cross section, through the joint probability distribution.
- The result that we get is a probability density function (we have to rescale it so that the total probability is 1) for Y.
- We call this the conditional distribution of Y, given X. $f_{Y|X}(Y|X)$
  - You can define a specific X, or view the conditional distribution as a function of X.



Artist: need a plot that looks like this, but none of the labels. Just a 3D plot with a slice through it and X and Y labeled.

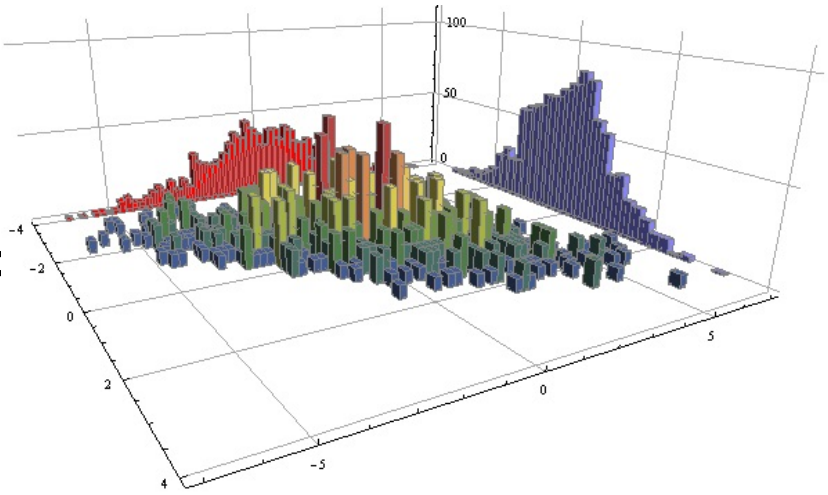
Application of Conditional Distribution

- Conditional distributions are a very general way of understanding the relationship between variables.
- In this figure, suppose X represents snowfall and Y represents number of visitors to a ski resort
- For each value of X, the conditional distribution of Y looks different
  - Notice that for smaller X, the distribution of Y is more concentrated at lower values
  - As we move to higher X, the distribution of Y moves upwards.
  - The conditional distribution of Y is a function of X.
  - It tells us in a very precise way what values of Y we can expect for each X.
  - There's a lot of information here, and we can usually only present graphs of the conditional distribution for a few values of X.
    - So we'll usually summarize the conditional distribution, often with its mean
  - But the conditional distribution is an important concept that we'll use many times.



Marginal Distribution

- Suppose you're running the ski resort and trying to plan capacity for the entire season.
- You don't know how much snowfall there will be.
- You want to know how probable each number of visitors is, without any other information
- This is just $f_Y(Y = y)$, the regular probability distribution
  - Not $f_{Y|X}(Y = y \mid X = x)$, the conditional distribution.
- We get this distribution by integrating the joint probability over all possible values of x:
- When we do this, the distribution we get is called a marginal distribution of the joint distribution
  - The word marginal comes from the idea of writing totals in the margins of a table.



$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_{y=-\infty}^{\infty} f_{Y|X}(y \mid x) f_X(x)\,dx$$

Expectation (lightboard)

Reading: Appendix B3-

Expectation

- The expectation of a random variable is its mean, averaged over probabilities.
  - I'm sure you've taken hundreds of means, so you have an intuitive understanding of expectation already.
- For a discrete random variable X, with a finite set of outcomes O={x_1, x_2,…,x_k} and probability function *f*, we write,

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_k f(x_k) = \sum_{j=1}^{k} x_j f(x_j)$$

  - Notice that if all *k* values have equal probability, we can rewrite this as the more familiar

$$E(X) = \sum_{j=1}^{k} x_j / k$$

- For a continuous random variable X with probability density function *f*, the expectation is similar, but we need an integral instead of a sum:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

- You could use the word mean instead of expectation, but the word expectation emphasizes that we're thinking of expectation as a function.
  - Let's look at some if its properties…

Properties of Expectations

- The expectation of a constant c, is the constant, $E(c) = c$
- Expectation is a linear function. That means:
  1. For a constant c, $E(cX) = cE(X)$
  2. For random variables X and Y, $E(X + Y) = E(X) + E(Y)$
     - You can move the expectation inside scalar multiplication and distributed it over sums.
- We can use these two to derive the expectation of a linear combination of two random variables:
- $E(aX + bY) = E(aX) + E(bY) = aE(X) + bE(Y)$
  - Not that a simlar rule does hold for multiplication. In general, $E(XY)$ is not the same as $E(X)E(Y)$.
    - One exception: if X and Y are independent, you can check that $E(XY) = E(X)E(Y)$

Variance (lightboard)

Reading: appendix B3-B4

Variance

Let's look at some special expectations.

- Let $\mu=E(X)$. We know this is the mean of X.
- We can take the mean away from X to get random variable $X - \mu$, and then square this to get $(X - \mu)^2$.
  - This should be familiar, its the squared deviation from the mean.
- The expectation of this random variable is the variance,
  - $var(X) = E[(X - \mu)^2]$
- We can rewrite this by expanding the square:
- $var(X) = E[X^2 - 2X\mu + \mu^2]$
- Then use the linearity of expectation:
- $var(X) = E(X^2) - E(2X\mu) + E(\mu^2)$
- $= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2$
- $= E(X^2) - \mu^2 = E(X^2) - [E(X)]^2$

Properties of Variance

For random variable X and constant c,

- $\text{var}(X+c) = E[((X+c)-E(X+c))^2]$
- $= E[(X+c-E(X) - c)^2] = E[(X-E(X))^2] = \text{var}(X)$

For random variable X and constant c,

- $\text{var}(cX) = E[(cX-E(cX))^2] = E[(cX-cE(X))^2]$
- $= E[c^2(X - E(X))^2] = c^2E[(X - E(X))^2] = c^2\text{Var}(X)$

- The square root of variance is the standard deviation, $\sigma_X = \sqrt{\text{var}(X)}$

Covariance (lightboard)

Covariance

- Consider random variables X and Y, with means $\mu_X=E(X)$ and $\mu_Y=E(Y)$
- You should remember covariance from your beginner statistics course. The definition looks similar to variance:

  $cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$
- This is a measure of how much two variables covary.
  - If $(X-\mu_X)$ tends to be positive when $(Y-\mu_Y)$ is positive (and vice-versa), $cov(X,Y)$ will be positive.
  - If $(X-\mu_X)$ tends to be positive when $(Y-\mu_Y)$ is negative (and vice-versa), $cov(X,Y)$ will be negative.
- We can do some math to rewrite covariance in another way:
- $cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)] = E[XY - \mu_X Y - \mu_y X + \mu_X\mu_Y]$
- $=E(XY) - E(\mu_X Y) - E(\mu_y X) + E(\mu_X\mu_Y) = E(XY) - \mu_X E(Y) - \mu_y E(X) + \mu_X\mu_Y$
- $=E(XY) - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y = E(XY) - \mu_X\mu_Y$

Properties of Covariance

- It is easy to check that for constants a and b cov(aX,bY) = ab cov(X,Y)
- We can also check that cov(X, Y + Z) = $E[(X-\mu_X)(Y + Z -\mu_Y -\mu_z)]$
- = $E[(X-\mu_X)(Y -\mu_Y) + (X-\mu_X)(Z -\mu_z)] = E[(X-\mu_X)(Y -\mu_Y)] + E[(X-\mu_X)(Z -\mu_z)]$
- = cov(X,Y) + cov(X,Z)

- If X and Y are independent, cov(X,Y) = 0.
- You may remember that you can scale covariance to get the correlation coefficient:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- This quantity is always between -1 and 1, so it gives us a standard way of measuring the strength of a linear relationship.

Relating Variance to Covariance

- We can also go back and find a nice formula for the variance of a sum of random variables
- $\mathrm{Var}(X + Y) = E[(X-\mu_X+Y-\mu_Y)^2]$
- $= E[(X-\mu_X)^2+(Y-\mu_Y)^2+2(X-\mu_X)(Y-\mu_Y)]$
- $= E[(X-\mu_X)^2]+E[(Y-\mu_Y)^2]+2E[(X-\mu_X)(Y-\mu_Y)]$
- $= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$

- This also tells us that if X and Y are independent, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Variance of Sum of 3 Random Variables

- What happens if we have more variables to add together?
- $Var(X + Y + Z) = Var(X + (Y + Z)) =$
- $Var(X) + Var(Y + Z) + 2Cov(X,Y+Z)$
- $=Var(X) + Var(Y) + Var(Z) + 2Cov(Y,Z)+ 2Cov(X,Y) + 2Cov(X,Z)$
- You can see that we get one term for every combination of variables in the sum. The pattern continues for even more variables.

Covariance Matrix

Covariance Matrix

- There will be times when we have an entire vector of random variables,
- For example, we use a sample of data to estimate the coefficients of a linear regression.
  - We have an entire vector of coefficients, and they will all vary in some way – but maybe not independently
- We need a way to summarize the types of variation in **X**.
- This can be done with a covariance matrix.
- In position i, j, we have the covariance of X_i and X_j.
- On the diagonal, we get the covariance of one X_i with itself, which is just variance.
- Notice also that the matrix is symmetric.
- These matrices are commonly used in regression analysis.

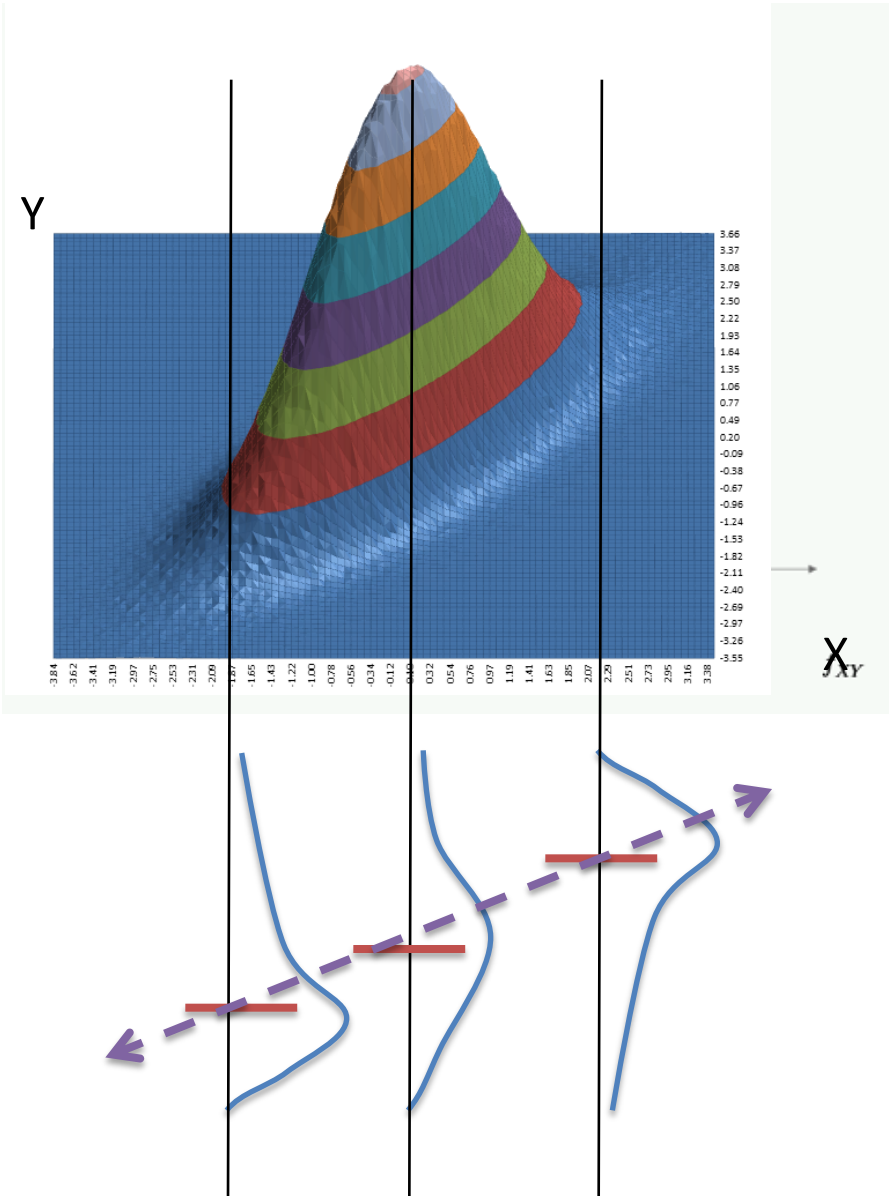
$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

$$\mathrm{var}(\mathbf{X}) = \begin{bmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & & \mathrm{cov}(X_1, X_n) \\ \vdots & & \ddots & \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & & \mathrm{var}(X_n) \end{bmatrix}$$

Conditional Expectation

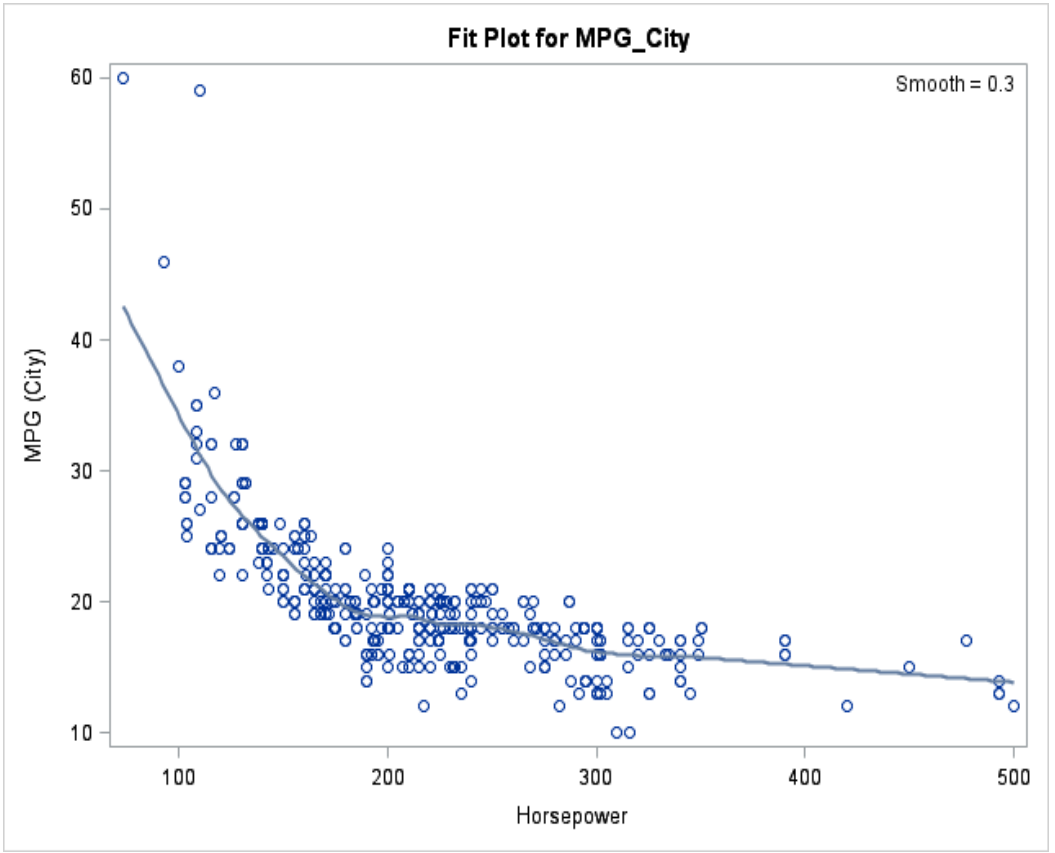
Conditional Expectation

- Remember that we can use a conditional distribution, $f_{Y|X}$, to understand how Y varies with X
  - But this is a lot of information – more than we can draw in a single figure.
- In the picture, we've drawn the conditional expectations of Y for three values of X.
  - We can't do that for every value of X.
  - Instead, let's summarize each one of these by recording just its expectation.
  - Now we have a single value for every X. This is the conditional expectation, E(Y|X)
  - This is much less information, but we can plot the conditional expectation for every value of X
    - In this case, E(Y|X) looks linear, but it could be much more complicated.



Using Conditional Expectation

- Here's an example, from a dataset on cars.
- The curve estimates the conditional expectation of MPG, given Horsepower.
- Notice that we lose some information compared to the conditional distribution function
  - For example, there seems to be more variation to the left of the graph, and the conditional distribution might be skewed upwards
- On the other hand, there's a lot more information here than there would be if we fitted a line to the data.
  - We would miss the upward curve

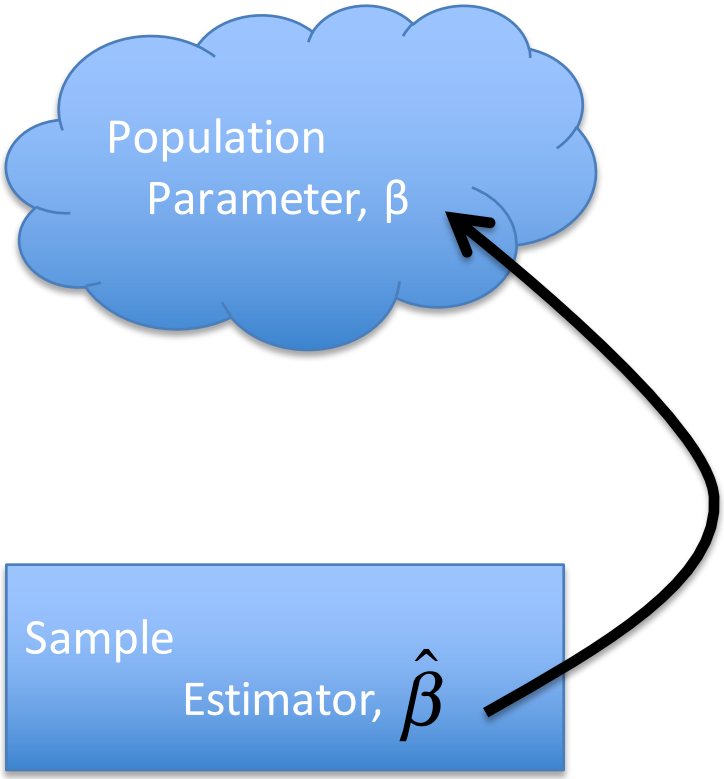


Parameters and Estimators

Reading: Appendix C1-C4

Parameters vs. Estimators

- In statistics, we have to be careful about whether we're talking about population parameters or estimators.
  - A parameter describes a population.
    - It's the true value
    - it has a fixed value.
    - It's something we're interested in.
    - It's usually unknown.
  - An estimator is calculated from our sample
    - We use the estimator to approximate the population parameter.
      - It's our guess for the parameter's value
    - It changes from sample to sample – it's a random variable
    - We usually write the name of the parameter with a hat over it to represent the estimator.
- Let's take a closer look at each one..



Population Parameters

- Even to say that a parameter describes a population, we need to make some assumptions.
  - This may seem counterintuitive: you may not always realize when you've made assumptions about a population.
  - The population has to belong to a mathematical class of populations that can be described by that parameter
    - This is what we call a ***population model.***
  - One common parameter is the mean, μ
    - The population model is $y = \mu + \varepsilon$, where $E(\varepsilon) = 0$.
    - Even here, there are some mathematical distributions for which we can't define a mean – it's infinite or undefined. We have to assume that this doesn't happen
  - In linear regression, our parameters are the slope coefficients and intercepts
    - For simple regression, our population model is usually $y = \beta_0 + \beta_1 x + \varepsilon$ (with some assumptions about the error term)
    - This (usually) means that we're assume there's a linear relationship in the population.
      - This is a strong assumption, and you should examine your scatterplots and diagnostics to see how realistic it is.

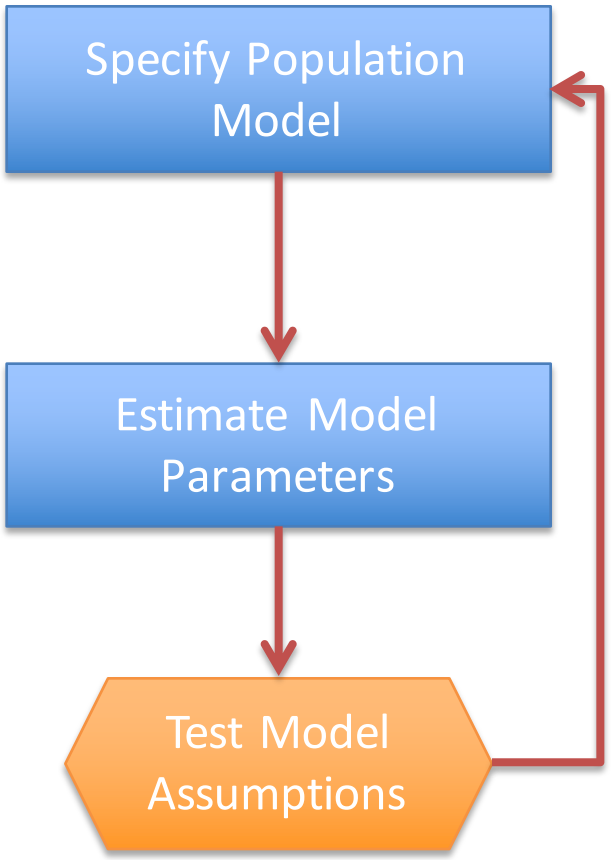
Estimators

- Intuitively, an estimator is our approximation of a population parameter.
  - It changes from sample to sample, so it has some probability distribution over possible values
    - Since we never know the parameter's value, we never know the sampling distribution, but we can still say some things about it.
- There are often many estimators that we can choose for our parameter
  - Ordinary Least Squares (OLS) Estimators
    - An estimator which minimizes the sum squared residuals (i.e., error terms) between each Yi and the model's estimate of Yi
    - In essence you're defining a loss function and trying to minimize it across all observations
  - Method of Moments Estimators (MMEs)
    - We know to expect certain relationships between parameters we care about (like the mean) and some expected values, like $E(Y)$ or $E(Y^2)$, so we take the sample analogs of those expected values and rearrange them to get sample estimates of our parameters
    - This is like taking the sample average and using it to infer the population average
    - The "moments" here are higher powers of the variable being taken in expectation, e.g., $E(Y^2)$ or $E(Y^3)$
  - Maximum Likelihood Estimators (MLEs)
    - For every estimate of a population parameter θ, we can calculate the likelihood of our data – the probability of observing values Y1 … YN conditional on that estimate. MLEs choose the estimate that maximizes that probability.
    - A very general estimator that works in many situations.
    - MLEs have some very nice mathematical properties that often make them good choices.

The Statistical Analysis Workflow

Let's see how parameters and estimators fit together in a statistical analysis:

- First, we specify a population model. This includes the parameters that describe our population and the assumptions needed for our estimation procedure to work.
- We then get some data, and we use it to compute estimates for our population parameters.
- Finally – and this is very important – we test the assumptions of our population model.
  - We test them against our dataset, so the question is whether our data looks like a sample that could come from our model
  - Note that we can never know for sure if our assumptions are correct.
    - We perform specific tests that measure particular features of the data
      - If those measurements deviate enough from what we expect under our model, we conclude that our model is unrealistic.
      - But even if our test looks reasonable, we still don't know that our model is true – we can only conclude that it's plausible.
  - If we decide that our model is unrealistic, we return to the first step and adjust our model specification.
- This is one example of a statistical analysis. In reality, things could be more complicated.
  - We may begin testing before we estimate parameters. We may also make some assumptions, then test, then add more assumptions…
  - But it may help you to keep this picture in mind as we go on.



Properties of Estimators

Estimator Properties

- We introduced estimators, but we were a bit vague about their behavior.
  - Intuitively, we want an estimator to be a good guess for a population parameter.
- But we need to explain what makes an estimator a good guess.
  - In other words, what are the properties that we'd like an estimator to have?

Bias

- One property that we might want for our estimator is that it is unbiased.
- This means that in expectation it actually equals the parameter
  - If you took a huge number of samples, the average of your estimator from each one would equal your parameter.
- We write this as $E(\hat{\theta}) = \theta$
- The bias of the estimator is defined as $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$
- Examples:
  - The mean of a random sample is an unbiased estimator for the population mean, $E(\bar{y}) = \mu$
  - As we'll soon discuss, Ordinary Least Squares is a way of estimating the parameters of a linear population model. We'll give a set of assumptions under which OLS regression is unbiased.

Consistency

- As we'll learn later in the course, there are situations in which we cannot get an unbiased estimator
  (at least not without sacrificing other attractive properties)
- We may instead use a biased estimator if we know the bias will be small, as long as we have a lot of data.
- A consistent estimator is one for which bias approaches zero for large sample sizes.
  - Given sample size n, $\hat{\theta}$ is a consistent estimator for θ if for every ε >0, we know $\hat{\theta}$ has a higher and higher chance of being within ε of θ
    - Prob(| $\hat{\theta}$ – θ| > ε) -> 0 as n -> ∞
    - Imagine setting up a window of size ε around θ, you can crank up n until you're 99% sure you're in the window. Then you can choose an even smaller window if you like, keep cranking up n...
    - This is called a probability limit. It is often written $\text{plim}_{n\to\infty}\hat{\theta} = \theta$

Examples of Consistent Estimators

Some important estimators that are biased, but consistent:

- The standard deviation of a random sample is a biased estimator for the population standard deviation
  - The sample variance is an unbiased estimator, but taking the square root introduces bias
  - But we still use the sample standard deviation as an estimator because its consistent.

Efficiency

- Often we have two estimators that are both unbiased (or both consistent). How do we choose between them?
- We want some measure of how precise they are.
- Assume some population parameter θ.
  - (this is unknown, but just imagine that it has a fixed value.
  - Different samples will yield different values of our estimator $\hat{\theta}$
  - A natural measurement of precision is how much variance there is in these values, var($\hat{\theta}$)
- Definition: Estimator $\theta\hat{}_1$ is relatively efficient to estimator $\theta\hat{}_2$ if $Var(\theta\hat{}_1) \leq Var(\theta\hat{}_2)$ for all possible values of θ with at least one θ such that $Var(\theta\hat{}_1) < Var(\theta\hat{}_2)$
- Efficiency means we're more confident of being closer to the true population parameter.
- A parameter is asymptotically efficient if for all sample sizes, it has variance equal to the Cramer-Rao bound – a lower bound on how much variance any estimator can have.

Artist: the ^s should go over the thetas, not after them.

The Two Big Laws of Asymptotics

You already know these laws, but let's state them in terms of the properties we just defined.

- **The Law of Large Numbers:** If we randomly sample y_1, $y_2,\ldots,y_n$ from a population with mean μ, the sample average,

$$\bar{Y}_n = \frac{1}{n} \sum Y_i$$

  is an unbiased and consistent estimator of μ
  - Consistency says we get closer to μ with higher and higher probability
- What's the variance of $\bar{Y}_n$?
  - Remember that Var(X + Y) = Var(X) + Var(Y) when X and Y are independent. The y_i are assumed to be independent, so

$$\text{var}(\frac{1}{n}\sum y_i) = \frac{1}{n^2}\text{var}(\sum y_i) = \frac{1}{n^2} n\,\text{var}(y_i) = \text{var}(y_i)/n$$

  - Note that this approaches zero, which is what we need for consistency.
- What about the standard deviation of $\bar{Y}_n$?

$$\sigma_{\bar{Y}_n} = \sqrt{\text{var}(\bar{Y}_n)} = \sqrt{\text{var}(Y_n)/n} = \sigma/\sqrt{n}$$

The Two Big Laws of Asymptotics

- **The Central Limit Theorem:** Let Y1 … Yn be a random sample of random variable Y_i from a population with mean μ and variance σ^2
  - In such circumstance, the Z-score (i.e., demeaned and renormalized version) of Y

$$Z_n = \frac{Y_n - \mu}{\sigma / \sqrt{n}}$$

  - Has an asymptotically normal distribution
  - That means that the cumulative distribution, approaches the standard normal one.
  - For any z, $P(Z_n<z) \rightarrow \Phi(z)$, where $\Phi(z)$ is the cumulative distribution of the standard normal curve at z.
  - Basically, the shape of the sampling distribution approaches the normal curve.
    - This is what allows us to test hypotheses about parameters in our models.
    - The normal curve gives us a sense of how much a parameter might vary due to chance.