Probability vs. Statistics

- Probability and statistics are often blurred together by students. They are related but accomplish two completely different tasks.
- Statistics
  - A field of applied mathematics. Used in other applied fields to draw conclusions about data gathered.
  - Used for collection, analysis, and interpretation of data.
  - Thus, the analysis part of our studies in this class.
- Probability
  - A field of theoretical mathematics used in applied fields as a modeling tool.
  - Predicts the likelihood of events given a set of assumptions (e.g. a distribution)
  - Thus, the modeling part of our studies in this class.
Probability Theory

• Probability is a formal framework to model likelihoods mainly used to make models from which we can come up with predictions

• There are two main (interchangeable) views of probability
  - Subjective / uncertainty view: Probabilities summarize the effects of uncertainty on the state of knowledge
    In Bayesian probability all types of uncertainty are combined in one number
  - Frequency view: Probabilities represent relative frequencies of events
    \[ P(e) = \frac{\text{(# of times of event } e)}{\text{(total number of events)}} \]
Random Variables

- Random variables are the main objects of interest in probability theory.
- In general they can be:
  - Propositional (i.e., only true or false values)
    - e.g., ishot, isblue, israining
  - Multivalued (i.e., textual values)
    - e.g., temp_feels_like, color, weather
  - Discrete Real-valued
    - e.g., temperature_degrees, hue_##, rain_volume_inch
  - Continuous Real-valued
    - E.g., temp_frac, wavelength, rain_volume_frac
- We use different notations to describe them.

Axioms of Probability (1)

- Probability follows a fixed set of rules
- For **propositional** random variables
  - \( P(A) \in [0,1] \) (notation!!)
  - \( P(\text{true}) = 1, \ P(\text{false})=0 \)
  - \( P(A \lor B) = P(A) + P(B) - P(A \land B) \)
  - \( P(A \land B) = P(A) \cdot P(B|A) \) (conditional probability!)
  - \( P(A) + P(\neg A) = 1 \)
Axioms of Probability (2)

• Similar axioms are for real-valued and multivalued variables. Mass function.
  • For multivalued random variables (e.g., \( X \in \{\text{blue,red,yellow,purple}\} \) or \( Y,Z \in S = \{y_1,y_2,\ldots,y_N\} \))
    • \( P(Y=y_i) \in [0,1], \ \forall 1 \leq i \leq N \) (notation!!)
    • \( P(S) = 1, \ P(\emptyset) = 0 \)
    • \( P(Y \cup Z) = P(Y) + P(Z) - P(Y \cap Z) \) (notation!!)
    • \( P(Y \cap Z) = P(Y) \ P(Z|Y) \) (conditional probability!)
    • \( \sum_{y_i \in S} P(Y = y_i) = 1 \)

Continuous Random Variables

• To describe continuous random variables we need additional tools
  • If \( X \in [0,10] \), uniformly distributed, how much is \( P(X=5) \)?
  • For most distributions and most values the answer is zero!
  • However the value \( P(4.5 \leq X \leq 5.5) \) is likely not zero. We need to define intervals on which we calculate probabilities.
  • The concept of probability densities is needed to describe the behavior of continuous random variables.
Cumulative and Probability Density

• Cumulative distribution:
  - $P(-\infty<X<x)$ is indeed a probability (behaves as such) and the function with the factor $x$ is the cdf: $F(X=x)= P(-\infty<X<x)$ or sloppily: $F(x)= P(X<x)$

• Probability density:
  - The intuitive “what’s the distribution like?” function
  - $p(X = x) = f(x) = \frac{dF(x)}{dx}$, i.e., the derivative of the cdf
  - Thus, $P(a \leq X \leq b) = \int_a^b p(X = x)dx = \int_a^b f(x)dx$

Other Probability Terms

• Unconditional or prior probabilities represent the state of knowledge before new observations or evidence
  - $P(H)$

• A probability distribution (pmf, or pdf) gives values for all possible assignments to a random variable
  - $F(x), f(x)$

• A joint probability distribution gives values for all possible assignments to all random variables in a system with multiple variables.
  - $X, Y, Z$, we should know $P(X=x_i, Y=y_j, Z=z_k)$ for all $i, j, k$

• Independent variables do not influence each other
BAYES THEOREM AND ITS IMPLICATIONS

Conditional Probability

• Conditional probabilities represent the probability after certain observations or facts have been considered
  • $P(H|E)$ is the posterior probability of $H$ after evidence $E$ is taken into account
• If A, and B are statistically independent, then $P(A|B)=P(A)$ and $P(B|A)=P(B)$
• Bayes rule allows to derive posterior probabilities from prior probabilities
  • $P(H|E) = P(E|H)*P(H)/P(E)$
Conditioning and Marginalization

- Probability calculations can be conditioned by conditioning all terms
  - Often it is easier to obtain conditional probabilities
  - E.g., if $P(A)=P(B)P(C)$, then $P(A|Z)=P(B|Z)P(C|Z)$
- Conditions can be removed by marginalization (rule of total probability aka. rule of elimination)
  - $P(H) = \sum_E P(H|E)P(E)$

More on Bayes

- Bayes theorem helps us to move any terms from the left to the right and vice versa:
  - $P(A \land B|C)=P(A \land C|B)*P(B)/P(C)$
  - $P(A \land B|C,D)=P(A \land C|B,D)*P(B|D)/P(C|D)$
  - $P(A|B)=P(A \land B)/P(B)$
  - $P(A|B,C)=P(A \land B|C)/P(B|C)$
  - $P(A\land B)=P(A|B)*P(B)$
Joint Distributions

• In real life decision processes we usually need one of the joint probabilities or one of the conditional probabilities.

• To calculate any of those it would be great to have the full joint distribution, as the joint distribution defines the probability values for all possible assignments to all random variables.

• Then conditional probabilities could be easily determined from the joint distributions (multiplication rule)

• However the size of the joint distribution increases exponentially with the number of random variables (a new variable needs to be combined with all other variables and variable combinations).

Inference

• Inference in probabilistic representation involves the computation of (conditional) probabilities from the available information.

• Most frequently the computation of a posterior probability \( P(H|E) \) form a prior probability \( P(H) \) and new evidence \( E. \) s

• This can involve many calculations and is cumbersome; it would be great to have simple algorithmic representations… (more - later)
PROBABILITY DISTRIBUTIONS

Functions of Distributions

- We know that deeper investigation of random variables requires us to assign real values to their outcomes (e.g., zero to head, one to tails)
- We already know:
  - Probability mass function (pmf) $P(X=x_i)=p_i$
  - Cumulative mass function (cmf) $P(X\leq x_i)=c_i$
  - Probability density function (pdf) $f(X)$
  - Cumulative distribution function (cdf) $F(X)$
The Expected Value

• The expected value $E[X]$ is a function which is defined:
  • For a finite discrete variable over the pmf:
    $$E[X] = \frac{\sum_{i=1}^{N} p_i \cdot x_i}{\sum_{i=1}^{N} p_i} = \sum_{i=1}^{N} p_i \cdot x_i$$
  • For a countable discrete variable:
    $$E[X] = \frac{\sum_{i} p_i \cdot x_i}{\sum_{i} p_i} = \sum_{i} p_i \cdot x_i$$
  • For a continuous random variable:
    $$E[X] = \frac{\int_{-\infty}^{\infty} xf(x)dx}{\int_{-\infty}^{\infty} f(x)dx} = \int_{-\infty}^{\infty} xf(x)dx$$

Moments of Distributions

• Moments of distributions are their very important properties.
• The $r^{th}$ moment of a distribution (sloppily denoted by) $X$ around $a$ is:
  $$E[(X - a)^r]$$
• Some special moments:
  • Mean is the first moment around zero:
    $$\mu = E[(X - 0)^1] = E[X]$$
  • Variance is the second moment around the mean:
    $$\sigma^2 = Var[X] = E[(X - \mu)^2] = E[X^2] - E[X]^2$$
  • Skewness tells us how "unsymmetrical" a distribution is:
    $$\gamma = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] = \frac{E[(X - \mu)^3]}{(E[(X - \mu)^2])^{3/2}}$$
SIMPLE MULTIVARIATE DISTRIBUTIONS

Multivariate Distributions

- Multivariate distributions sometimes arise when combining the outcomes of multiple random variables
- Sometimes we are interested in the joint effect of multiple random variables
  - Distribution of the product of two random variables
  - Distribution of the joint additive effect of multiple variables
Multivariate Distributions (2)

• For some operations combining multiple variables we can determine the moments of the distribution relatively easily
  • Usually assumptions made about random variables
    • Independently distributed
    • Moments of the distributions of the individual variables are known
  • If variables are not independent we have to use conditional distributions and the laws of probability

---

Significant Moments of Random Variable Products

• The mean and variance of the distribution of the product of two independent random variables $X$ and $Y$ are:
  • $\mu_{XY} = \mu_X \mu_Y$
  • $\sigma_{XY}^2 = \sigma_X^2 \sigma_Y^2 + \sigma_X^2 \mu_Y^2 + \sigma_Y^2 \mu_X^2$
Significant Moments of Random Variable Sums

- The mean and variance of the distribution of the sum of two independent random variables $X$ and $Y$ are:
  - $\mu_{X+Y} = \mu_X + \mu_Y$
  - $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$

DISTRIBUTION FAMILIES (i.e., what’s the question?)
Families of Distributions

• There are families of important distributions that are useful to model or analyze events
  • Families of distributions are parameterized
  • Different distributions are used to answer different questions about events
    • What is the probability of an individual event
    • How many times would an event happen in a repeated experiment
    • How long will it take until an event happens

DISCRETE DISTRIBUTIONS
Discrete Uniform Distribution

What is the probability of an individual event if all events are equally likely?

- Parameterized by the number of discrete events $x_i$, $N$
- Probability function:
  $$p(x; N) = P(X = x) = \frac{1}{N} \quad \forall x \in \{X\}$$
- Mean: $E[X] = \mu = \frac{1}{N} \sum_{i=1}^{N} x_i$
- Variance: $Var[X] = \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$

Discrete Uniform Distribution (2)

What is the probability of an individual event if all events are equally likely?

- Parameterized by the lowest possible value $a$ and the highest possible value $b$. Thus $N = b - a + 1$
- Probability function:
  $$p(x; N) = P(X = x) = \begin{cases} \frac{1}{N} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$
- Mean: $E[X] = \mu = \frac{a + b}{2}$
- Variance: $Var[X] = \sigma^2 = \frac{N^2 - 1}{12}$
**Bernoulli Distribution**

- **If an event happens with a given probability, what is the probability that it happens?**
  - Parameterized on the probability $p$ of the event happening.
  - Probability function:
    \[
    P(x; p) = P(X = x) = \begin{cases} 
      p & \text{if } x = 1 \\
      1 - p & \text{otherwise}
    \end{cases}
    \]
  - Mean: $\mu = p$
  - Variance: $\sigma^2 = p \times (1 - p)$

---

**Binomial Distribution**

- **What is the likelihood that an event will occur a certain number of times $k$ in $n$ Bernoulli experiments?**
  - Parameterized by the likelihood, $p$, of the event in the Bernoulli experiment, the number of positive outcomes $k$ in the experiment, and the number of experiments, $n$
  - Probability function:
    \[
    P(k; n; p) = \binom{n}{k} p^k (1 - p)^{n-k}
    \]
  - Mean: $\mu = np$
  - Variance: $\sigma = np(1 - p)$
Hypergeometric Distribution

- **What is the probability of **k** successes in **n** draws from a finite population without replacements?**
  - Parameterized by the initial population **N**, the initial number of positive events **M**, the number of experiments **n**, and the number of positive outcomes **x**.
  - The probability mass function is:
    \[ P(x, M, N, n) = \binom{M}{x} \binom{N-M}{n-x} \binom{N}{n} \]
  - The mean is: \( \mu = n \frac{M}{N} \)
  - The variance is: \( \sigma^2 = n \frac{M}{N} \frac{(N-M)}{N} \frac{(N-n)}{N-1} \)

Multinomial Distribution

- **What is the likelihood that if there are **K** independent events, each will occur a certain number of times \( (x_i) \) given that there are **n** independent experiments?**
  - Parameterized by the likelihoods, \( p_i \), of the **K** events in the experiment and the number of experiments, **n**
  - Probability function:
    \[ P(x_1, ..., x_K, n, p_1, ..., p_K) = \frac{n!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} p_k^{x_k} \]
  - Mean: \( E[X_i] = \mu_i = np_i \)
  - Variance: \( \text{Var}[X_i] = np_i(1-p_i) \)
Poisson Distribution

- **What is the likelihood that an event will occur a certain number of times in a continuous experiment?**
  - Parameterized by the expected number of occurrences, \( \lambda \), of the event within one time period (rate)
  - Probability function: \( P(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \)
  - Mean: \( \mu = \lambda \)
  - Variance \( \sigma^2 = \lambda \)

Geometric Distribution

- **What is the likelihood of the event happening first at exactly the \( x^{th} \) trial in a Bernoulli experiment?**
  - Parameterized by the probability, \( p \), of the event in each Bernoulli experiment
  - Probability mass function: \( P(x; p) = (1-p)^{x-1}p \)
  - Mean: \( \mu = \frac{1}{p} \)
  - Variance: \( \sigma^2 = \frac{1-p}{p^2} \)
CONTINUOUS DISTRIBUTIONS

37

Uniform Distribution

- **What is the probability of a particular outcome if all outcomes are equally likely?**
  - Parameterized by the lowest possible value $a$ and the highest possible value $b$. $[a,b]$
  - Probability density function:
    $$f(x; N) = \begin{cases} 
    \frac{1}{b - a} & \forall x, a \leq x \leq b \\
    0 & \text{otherwise}
    \end{cases}$$
  - Cumulative distribution function:
    $$F(x; N) = \begin{cases} 
    0 & x < a \\
    \frac{x - a}{b - a} & a \leq x \leq b \\
    1 & b < x
    \end{cases}$$
  - Mean: $\mu = \frac{a + b}{2}$
  - Variance: $\sigma^2 = \frac{(b-a)^2}{12}$
Normal Distribution

- **Models the likelihood of results if the results are either distributed with a “Bell curve” or, alternatively, the result of the summation of a large number of random effects. This is a good approximation for a wide range of natural processes or noises.**
  - Parameterized by a mean, \( \mu \), and the standard deviation \( \sigma \), frequently denoted as \( N(\mu, \sigma^2) \)
  - Probability density function:
    \[
    f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
    \]
  - Cumulative distribution function: no closed form
  - Mean and variance are trivial

Exponential Distribution

- **Models the likelihood of an event happening for the first time at time \( x \) in a Poisson process (i.e., a process where events occur with the same likelihood at any point in time, independent of the time since the last occurrence).**
  - Parameterized by event rate: \( \lambda \)
  - Probability density function: \( f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \)
  - Cumulative distribution function: \( F(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \)
  - Mean \( \mu = \frac{1}{\lambda} \)
  - Variance \( \sigma^2 = \frac{1}{\lambda^2} \)
Moment Generating Function

- Moments are sometimes hard to derive using their definition. The moment generating function can come in handy in these cases.
- The moment generating function for random variable $X$ is defined as:
  \[ m_X(t) = E[e^{tx}] \]
- The $r^{th}$ moment around 0 can be calculated as:
  \[ \lim_{t \to 0} \frac{\partial^r m_X(t)}{\partial t^r} \]
- The $r^{th}$ moment around $a$ can be calculated as:
  \[ \lim_{t \to 0} \frac{\partial^r m_{X-a}(t)}{\partial t^r} \]
- These cannot always be computed.
Mean and Variance

• The mean and variance thus can be calculated as (for continuous variables):
  
  - Mean: \( \mu = \lim_{t \to 0} \frac{\partial}{\partial t} \int e^{xt} f(x) dx \)
  
  - Variance: \( \sigma^2 = \lim_{t \to 0} \frac{\partial^2}{\partial t^2} \int e^{(x-\mu)t} f(x) dx = \lim_{t \to 0} \frac{\partial^2}{\partial t^2} \int e^{xt} f(x) dx - \mu^2 \)

Example: Poisson Distribution

• Probability mass function:
  \( P(x, \lambda) = p_x = \frac{\lambda^x e^{-\lambda}}{x!} \)

• Moment generating function:
  \( m_x(t) = E[e^{xt}] = e^{\lambda(e^t-1)} \)

• Mean:
  \( \mu = \lim_{t \to 0} \frac{\delta}{\delta t} e^{\lambda(e^t-1)} = \lambda \)

• Variance
  \( \sigma^2 = \lim_{t \to 0} \frac{\delta^2}{\delta t^2} e^{\lambda(e^t-\mu-1)} = \lambda \)
DISTRIBUTION RELATIONSHIPS

Hypergeometric to Binomial

- If the population is large and the number of samples drawn is small, then the Hypergeometric distribution can be approximated by the Binomial distribution.
  - \( p = \frac{M}{N} \)
Normal to Standard Normal

- We usually denote Normal as: \( N(\mu, \sigma^2) \)
- The standard normal as: \( N(0, 1) = Z \)
- If random variable \( X \) is normally distributed, i.e., \( X = N(\mu, \sigma^2) \) then \( Z = (X - \mu) / \sigma \)

Binomial to Poisson

- Binomial’s pdf is: \( P(x; n; p) = \binom{n}{x} p^x (1 - p)^{n-x} \)
- The problem is in the binomial coefficient; if \( n \) is large it is very hard to calculate.
- Poisson asks a similar question but on a continuous time scale. \( P(x, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \)
- If \( n \) is large and \( p \) is small, then the binomial can be approximated by the Poisson with rate \( \lambda = np \)
Binomial to Normal

• We cannot use Poisson to approximate binomial if $p$ is not very small (as $np$ goes towards infinity).

• However, we can use the Normal distribution: $N(np, np(1-p))$

• Thus we can also approximate the Poisson as $N(\lambda, \lambda)$ for large $\lambda$-s
(not) All that Glitters is Gold

- If a function $f(x)$ is always positive and converges to 0, $\lim_{x \to \infty} f(x) = 0$, can we make a pdf out of it?
- Let’s say $f(x) = \frac{a}{x}$ between 1 and $\infty$. Can we find a value for $a$ to make this into a pdf?
  - No we cannot.

Maybe Silver?

- How about some “quicker” convergence. Let’s look at the function $f(x) = \frac{a}{x^2}$. (Still on $[1, \infty]$ and zero otherwise.)
- Can this be made into a pdf? What is $a$? What is the cdf?
- What is the mean?
  - Infinite. The tail of the function is too heavy.
- What is the variance?
  - Infinite
Lower Polynomial Powers?

- Can \( f(x) = \frac{a}{x^3} \) be made into a pdf on \([1, \infty]\)?
  - Does it have a mean? Does it have a variance? Is the tail still heavy?
- How about \( f(x) = \frac{a}{x^4} \)?
- The above functions can be seen as pdfs but they do not model any real or intuitive probability densities or do they?
- What does all of this imply? Recall the normal function? Recall the sum of variances?
- Heavy tailed pdfs do not exist in real life though, or do they?

Pareto Distribution

- The Pareto distribution has two parameters, one relates to its shape (\( \alpha \)) the other to the minimum value of the domain (\( x_m \)).
  - The pdf is: \( f(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{(\alpha+1)}} & x \geq x_m \\ 0 & x < x_m \end{cases} \)
  - The cdf is: \( F(x) = \begin{cases} 1 - \left(\frac{x}{x_m}\right)^\alpha & x \geq x_m \\ 0 & x < x_m \end{cases} \)
- This is the 80-20 rule in economics (where \( \alpha \) is 1.161)
- The \( \alpha \) value can make this heavy-tailed.