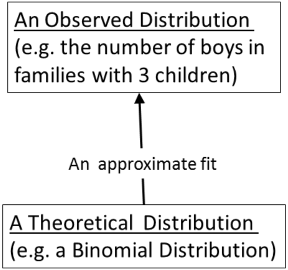
Basic Probability

Machine Learning

Models for Data I

- Sometimes a known functional form (theoretical distribution) can be used to summarise the observed data (like a Normal distribution)

Models for Data II



An Observed Distribution
(e.g. the number of boys in families with 3 children)

An approximate fit

A Theoretical Distribution
(e.g. a Binomial Distribution)

Models for Data III

- Probability calculus gives us the deductive machinery for inference with *probability* functions,

Random Variables I

- A unifying idea underlying all of probability theory is that of a *random variable*
- A random variable can also be seen as the main link connecting statistical observations to the world of probability theory
- Mathematically, a random variable is a function
- Suppose you conduct an experiment which has some outcomes. If we cannot predict the outcome of an experiment, it is called a *random experiment*. The set of outcomes of a random experiment is called the *sample space* (this is the mathematician's term for what we were calling "the population")

Random Variables II

- A random variable is a function from the sample space to some set denoting the range of the random variable. Often, though not always, the range is the set of real numbers. That is, it is set of pairs $\{(s_1, p_1), (s_2, p_2), \ldots\}$ where the s_i are elements of the sample space
  - When you toss 3 coins, the experiment can have different outcomes based on each one landing H or T
  - The sample space has 8 outcomes. One random variable h is a function that takes each outcome and maps it to the number of H's in the outcome. That is $h(HHH) = 3$, $h(HHT) = 2$ and so on
- *Events* are subsets of outcomes. Sometimes, we are interested in events that are related to the value(s) taken by some random variable. For example, the subset of outcomes in tosses of 3 coins, s.t. the number of *Heads* is 2
  - $TwoHeads = \{s : s \in S, h(s) = 2\}$ $(TwoHeads = \{HHT, HTH, THH\})$

Random Variables III

  - If all these outcomes are equally likely, then we can answer the question $Prob(TwoHeads) = 3/8$
- Textbooks usually use a capital letter (like “X”) to denote the function representing the random variable, and often even just say $X = 2$ instead of of $X(s) = 2$
  - So, you may end up confusing the experimental outcome (like HHT) with some function of that outcome (like $X(HHT)$)
- Even if this is done, do not forget that a random variable is, in the end a function over outcomes
- Now, if X is a random variable that maps outcomes (H or T) from n Bernoulli trials to the number of successes, then we know that the distribution of values of X will be distributed according to the Binomial distribution

$$p(X = k) \ = \ p(X(HHT \cdots H) = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Random Variables IV

- Similarly, if Y is the number of failures before the first success, then the distribution of $Y = k$ values is a geometric distribution
- If the range of a random variable X (remember, X is a function) is finite, or a countably infinite set (that is, it can be put in a 1-1 corrspondence to the natural numbers), then X is said to be *discrete* and the corresponding probability function over X values is called a *probability mass function* (or p.m.f). Otherwise X is *continuous* and the probability over X values is calculated using a *probability density function* (or p.d.f)

Random Variables V

- This dichotomy is unpleasant, but we will have to live with it in this course. There can be probability functions that are *mixed* in the sense of being both continuous and discrete, depending on the outcomes. This causes difficulty to the p.m.f/p.d.f based formulation, and requires a reformulation of probability as a measure over the set of events.
- For now, we will simply go with random variables being discrete or continuous, and the corresponding probabilities being obtained using a p.m.f. or a p.d.f.

Aside: Mass Density and Total Mass I

- Suppose you are told that the mass density of a rod of metal is 1 gm/cm. This means that the distribution of metal is the same across the entire rod (the mass density is constant). Then it is easy to see that the total mass of a 10cm rod is 10gm
- Now, suppose instead you were told that the mass density varied from point to point in the rod, given by some function $f(x)$ gm/cm where x is the point on the rod. Then how would you calculate the total mass of a rod 10cm long?
  - One way is to imagine the rod to be made up of small pieces of size Δ each, within which the mass density is approximately the same. Suppose these pieces were centred at $x = x_1, x_2, \ldots x_n$. Then, the total mass is $f(x_1)\Delta + f(x_2)\Delta + \cdots + f(x_n)\Delta = \sum_{i=1}^{n} f(x_n)\Delta$

Aside: Mass Density and Total Mass II

- In general, instead of Δ if we use a small interval dx that contains the point x, and increase n, the total mass between $x = a$ and $x = b$ becomes:

$$\int_a^b f(x)dx$$

- With discrete random variables we are able to compute the probability ("mass") at each point. With continuous random variables we are able to calculate the probability between points (or less than, or greater than some point x), using the density function. With a small enough interval dx,
  $P(X \approx x_1) = f(x_1)dx$
  But you cannot use it to obtain the probability at an exact point $x = x_1$

Probability Spaces I

- Let us look at a random experiment like a coin-toss. Any one toss of the coin has two outcomes, H or T. The sample space therefore has 2 elements in it.
- In general a sample space S could contain a finite number of elements or be an infinitely large set. An subset of S is called an *event*
- Associated with each event $A \subseteq S$, is a real number probability $P(A)$ such that:
  - $0 \leq P(\{s\}) \leq 1$ for $s \in S$
  - $\sum_i P(\{s_i\}) = 1$
  - If A and B are disjoint events, then $P(A \cup B) = P(A) + P(B)$

  (The last condition has to be made more precise to account for the case when S is an infinite set.)

Probability Spaces II

- It follows from this that if we want to find the probability of a set say $\{s_1, s_2, s_3\}$ for $s_{1,2,3} \in S$, then this is $P(\{s_1\}) + P(\{s_2\}) + P(\{s_3\})$. It is also common to write $P(s)$ rather than $P(\{s\})$ for $s \in S$.
- "Equally likely" events are those that have the same probability value. If all events are equally likely, then for finite sets of size n, their probability value must be $1/n$
- The sample space S along with an assignment of probability to every subset of S, is called a *probability space* (correctly, a probability space is the pair (S, P) where P is a probability function).

Probability Distributions and Random Variables I

- Recall that a random variable X is a function from the set of outcomes (often to a real number)
- This function can be a many-to-one mapping, in the sense that several outcomes may map to the same number. We will use $X = a$ to denote the set of outcomes that a random variable X maps to the value a
  - if X is the function that counts the number of heads in 3 coin tosses, then $X = 2$ refers to the set $\{HHT, THH, HTH\}$
- That is, $X = a$ denotes the event: $\{s : s \in S, X(s) = a\}$
- Given an probability space (S, P), we will call the values of $P(X = x)$ for all values x of X, the probability distribution of the random variable X, and write this as $P(X)$.
- Sometimes we will write $P(x)$ to stand for $P(X = x)$. Remember x is a (real) value, and X is a function, and $X = x$ is a set of outcomes (that is, an event)

Probability Distributions and Random Variables II

- If we have more than 1 random variable — say the functions X and Y, then for all values x of X and y of Y, $P(X = x, Y = y)$ will be called the joint distribution of X and Y. This will refer to the probability of the event $\{s : s \in X = x \text{ and } s \in Y = y\}$
  - Suppose the set of outcomes is the pair of entries from throwing a pair of dice. Then $S = \{(1,1), (1,2), \ldots, (6,6)\}$
  - Let:

$$X((a,b)) = a + b$$

$$Y((a,b)) = \begin{cases} odd & \text{if } odd(a) \text{ and } odd(b) \\ even & \text{otherwise} \end{cases}$$

Then $P(X = 4, y = odd) = P(4, odd)$ is the probability of $\{(1,3),(3,1)\} = P((1,3)) + P((3,1)) = 1/18$

Probability Distributions and Random Variables III

- Just like with one r.v., when we want to refer to the joint distribution of all values of X and Y as $P(X, Y)$. For more than 2 variables, the same principle applies.

Unconditional and Conditional Probability I

- Given a probability space (S, P), the unconditional probability of a event $E \subset S$ is the sum of the probabilities assigned by the probability function P to individual outcomes in E
- Given events (subsets of the sample space) E and F, such that $P(F) \neq 0$, the conditional probability of E given F, is

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \quad (P(F) \neq 0)$$

- There is a difference between $P(Y|X)$ and $P(X, Y)$. For example, suppose you were given 4 (x, y) data points $D = (1,0),(1,0),(2,0),(2,1)$. Then:

X	Y	$P(X,Y)$
1	0	0.5
1	1	0.0
2	0	0.25
2	1	0.25

Unconditional and Conditional Probability II

- But:

X	Y	$P(Y\|X)$
1	0	1.0
1	1	0.0
2	0	0.5
2	1	0.5

Independence I

- Two events E and F are independent if either $P(E) = 0$ or $P(F) = 0$ or:

$$P(E|F) \;=\; P(E) \quad (P(E) \neq 0, \;\; P(F) \neq 0)$$

(It follows from this that $P(F|E) = P(F)$

- A variant of this is

$$P(E \cap F) \;=\; P(E)P(F)$$

- Events E and F are conditionally independent given G if either $P(E|G) = 0$ or $P(F|G) = 0$ or

$$P(E|F \cap G) \;=\; P(E|G)$$

This is similar, but not the same as unconditional independence. Once we know the outcomes in G, then the outcomes in F are irrelevant for calculating the probability of the outcomes in E

Total Probability

- if we have the events $E_1, E_2, \ldots, E_n$ that are partitions of the space of outcomes S. That is, $E_i \cap E_j = \emptyset$ $(i \neq j)$ and $S = \cup E_i$
- Then, for some event F:

$$P(F) \;=\; P(F \cap E_1) + P(F \cap E_2) + \cdots + P(F \cap E_n)$$

- The conditional form of this rule is used quite often:

$$P(F) \;=\; P(F|E_1)P(E_1)+P(F|E_2)P(E_2)+\cdots+P(F|E_n)P(E_n)$$

Marginals I

- Recall the two forms of the law of total probability for events F and $E_1, \dots, E_n$

$$P(F) \;=\; P(F \cap E_1) + P(F \cap E_2) + \cdots + P(F \cap E_n)$$

$$P(F) \;=\; P(F|E_1)P(E_1) + P(F|E_2)P(E_2) + \cdots + P(F|E_n)P(E_n)$$

- If F denotes the events $X = x$ for some r.v. X and $E_1, \dots, E_n$ the events $Y = y_1, \dots Y = y_n$ for some r.v. Y, then it follows that:

$$P(X = x) \;=\; \sum_i P(X = x, Y = y_i)$$

- $P(X = x)$ is the *marginal* probability distribution of X given the joint distribution $P(X, Y)$. Marginals with more variables can be obtained in the same way. For example $P(X = x, Y = y) = \sum_z P(X = x, Y = y|Z = z)P(Z = z)$ and so on

Marginals II

- A variant of marginalisation is called conditioning:

$$P(X = x) \; = \; \sum_i P(X = x | Y = y_i) P(Y = y_i)$$

  This follows from the conditional version of the law of total probability

Product Rule

Given two events A and B

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ P(B \cap A) &= P(B|A)P(A) \end{aligned}$$

A generalisation of the product rule is the *chain rule*:

$$\begin{aligned} \mathrm{P}(x_1, \ldots, x_n) &= \mathrm{P}(x_n|x_{n-1}, \ldots, x_1)\mathrm{P}(x_{n-1}, \ldots, x_1) \\ &= \mathrm{P}(x_n|x_{n-1}, \ldots, x_1)\mathrm{P}(x_{n-1}|x_{n-2}, \ldots, x_1) \cdots \mathrm{P}(x1) \\ &= \textstyle\prod_{i=1}^{n} \mathrm{P}(x_i|x_{i-1}, \ldots, x_1) \end{aligned}$$

Independence (again)

- Recall events A and B are independent if and only if

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

- Or, by the product rule

$$\mathrm{P}(A|B) = \mathrm{P}(A) \quad \text{and} \quad \mathrm{P}(A|B) = \mathrm{P}(B)$$

Bayes' Rule I

- Bayes' rule or Bayes Theorem:

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

- Here is some terminology:
  - $P(A)$ is called the *prior probability* of A
  - $P(A|B)$ is called the *posterior probability* of A given B
  - $P(B|A)$ is called the *likelihood* of B given A
  - $P(B)$ is called the *marginal likelihood* of B

$$P(B) = \sum_a P(B|A=a)P(A=a)$$

(This is just the rule of total probability.)

Bayes' Rule II

- Conditional form of Bayes' rule:

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

(Check this!)

- Side-stepping the normalisation step:

$$\begin{aligned} P(A|B) &= \frac{P(B|A) \times P(A)}{P(B)} \\ P(\neg A|B) &= \frac{P(B|\neg A) \times P(\neg A)}{P(B)} \end{aligned}$$

So, $P(A|B) = \alpha P(B|A)P(A)$ and
$P(\neg A|B) = \alpha P(B|\neg A)P(\neg A)$.
Since $P(A|B) + P(\neg A|B) = 1$, we get α as:

Bayes' Rule III

$$\alpha = \frac{1}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

Although we knew this anyway, it still gives us a slightly different way of calculating with Bayes' rule, by postponing the calculation of the normalising factor α

Bayes' Rule with Several Variables

$$\mathrm{P}(Y_1, \cdots, Y_m | X_1, \cdots, X_n) = \alpha \mathrm{P}(X_1, \cdots, X_n | Y_1, \cdots, Y_m) \mathrm{P}(Y_1, \cdots, Y_m)$$

- We need to know the conditional probabilities for all possible combinations of the X_i, Y_js
  - With just boolean variables, this is $2^n 2^m$
  - Just like using the full joint distribution
- Can we use the idea of independence to reduce these requirements?

Conditional Independence

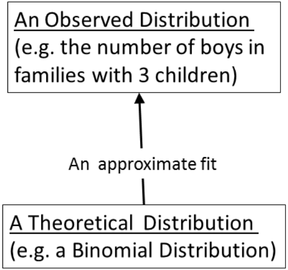
- If X and Y are conditionally independent given Z then
  $\mathrm{P}(X, Y|Z) = \mathrm{P}(X|Z)\mathrm{P}(Y|Z) \quad \text{and}$
  $\mathrm{P}(X|Y, Z) = \mathrm{P}(X|Z) \quad \text{and} \quad \mathrm{P}(Y|X, Z)\mathrm{P}(Y|Z)$
- $X_1, \cdots, X_n$ are conditionally independent, given $Y_1, \cdots, Y_m$ if and only if

  $$\mathrm{P}(X_1, \cdots, X_n|Y_1, \cdots, Y_m) = \\ \alpha\mathrm{P}(X_1|Y_1, \cdots, Y_m)\cdots\mathrm{P}(X_n|Y_1, \cdots, Y_m)$$
- For boolean variables, this reduces the representation to $2n2^m$ combinations

Probability Models I

- Sometimes a known functional form (theoretical distribution) can be used to summarise the observed data (like a Normal distribution)

Probability Models II



An Observed Distribution
(e.g. the number of boys in families with 3 children)

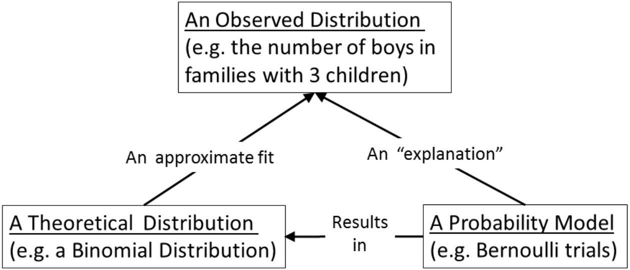
An approximate fit

A Theoretical Distribution
(e.g. a Binomial Distribution)

Probability Models III

- Sometimes a *stochastic* model can be used to justify the theoretical distribution for the data

Probability Models IV



An Observed Distribution

(e.g. the number of boys in families with 3 children)

An approximate fit

An "explanation"

A Theoretical Distribution

(e.g. a Binomial Distribution)

Results in

A Probability Model

(e.g. Bernoulli trials)

The Binomial Probability Model I

- If the probability of success in a Bernoulli trial is p and the probability of failure is q, then the probability of i successes in a sequence of n trials is given by $B(n,i)p^i q^{(n-i)}$, where $B(n,i)$ is some number that we have so far obtained using Pascal's triangle
- It should be easy to see that $B(n,i)$ is simply $\binom{n}{i}$
- Thus, the probability of i successes and $(n-i)$ failures is:

$$p_i = \binom{n}{i} p^i q^{(n-i)}$$

- It can be shown that $\binom{n}{i} = \frac{n!}{i!(n-i)!}$. So:

$$p_i = \frac{n!}{i!(n-i)!} p^i q^{(n-i)}$$

Mean and Spread of the Probability Model I

- The “theoretical mean” is called the *expected value* (usually denoted μ) is simply a weighted average that multiplies each outcome by its (theoretical) probability. That is, the mean for the probability model over a set of outcomes $x_1, x_2, \ldots$ with probabilities $p(x_1), p(x_2), \ldots$ is:

$$\mu \; = \; E(X) \; = \sum_k x_k p(x_k)$$

- The spread of the probability distribution

$$\sigma^2 = \sum_k (x_k - \mu)^2 p(x_k)$$

- It is not hard to show that

$$\sigma^2 \; = \; E(X^2) - (E(X))^2$$

Continuous Probability Distributions I

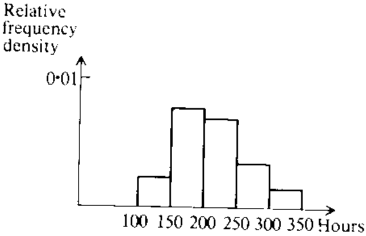
- So far, we have looked mostly at outcomes that have been *discrete*, like the number of heads (this is the same as saying we have been looking at discrete random variables)
- As a result, we have been able to tabulate these values and their relative frequencies. We have been able to think of constructing functions that approximate the relative frequency of an outcome

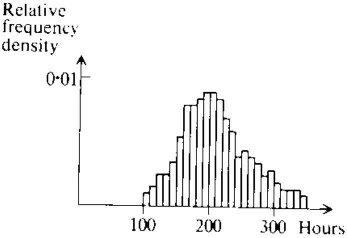
Continuous Probability Distributions II

- To model such probabilities we will have to use a continuous function However, the function will not be a probability mass function (i.e. one that computed the probability of a particular value—like 240 hours), but a probability *density* function (p.d.f.)
- We will have to go back to histograms

Probability Density Functions I

- Recall that histograms were frequency density diagrams. Here are two histograms for battery-life, using different interval widths



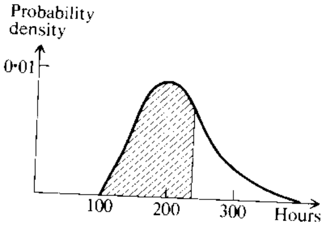


Probability Density Functions II

  - Now the proportion of instances (= relative frequency) with battery life less than 240 can be calculated by summing up the areas of rectangles to the left of 240 in any one of the two histograms
- The probability density function (p.d.f.) is a curve that approximates this relative frequency density. For any interval, it the area under the curve gives the probability of a data

Probability Density Functions III

instance falling in that interval



- As in the discrete case, we will have two requirements on on any p.d.f. $\phi(x)$:
  1. $\phi(x) \geq 0$ for all values of x;
  2. $\int_{-\infty}^{+\infty} \phi(x) = 1$

Mean and Spread (Continuous Models) I

- We saw earlier how we can obtain the mean and spread of a discrete probability function:

$$\mu = \sum_i x_i p(x_i) \qquad \sigma^2 = \sum_i (x_i - \mu)^2 p(x_i)$$

- The natural extension to continuous models is:

$$\mu = \int x\phi(x)dx \qquad \sigma^2 = \int (x - \mu)^2 \phi(x)dx$$

(where the integration is over the domain of the p.d.f.)

Modelling Evenly Scattered Frequency Distributions I

- The (continuous) uniform distribution is a family of curves such that is used to model relative frequencies that are approximately the same in any interval.
- The distribution has two parameters, a and b, which are its minimum and maximum values. The distribution is often abbreviated $U(a, b)$

$$\phi(x) = \begin{cases} \frac{1}{b-a} & \mathrm{if}(x \in [a, b]) \\ 0 & \mathrm{otherwise} \end{cases}$$

Modelling Evenly Scattered Frequency Distributions II

- The mean and spread of this deviation can be easily shown to be:

$$\text{Mean} = E(X) = \int_{-\infty}^{+\infty} x\phi(x)dx$$

That is,

$$\begin{aligned}\text{Mean} &= \int_a^b \frac{x}{b-a}dx \\ &= \frac{1}{2}\frac{(b^2-a^2)}{b-a} \\ &= \frac{b+a}{2}\end{aligned}$$

Modelling Evenly Scattered Frequency Distributions III

- Similarly, for the variance:

  $$\text{Variance} \;=\; E(X^2) - (E(X))^2$$

  If you do the calculations correctly, you should find:

  $$\text{Variance} \;=\; \frac{(b-a)^2}{12}$$

- The uniform distribution has another important role: it can be used to *simulate* many other distributions. We will see how to do this later.

Modelling Symmetric, Bell-shaped Frequency Distributions

- The Gaussian or Normal distribution has 2 parameters: the mean μ and standard deviation σ. In this distribution the frequency with which a value x occurs is:

Modelling Symmetric, Bell-shaped Frequency Distributions

- The Gaussian or Normal distribution has 2 parameters: the mean μ and standard deviation σ. In this distribution the frequency with which a value x occurs is:

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

Modelling Symmetric, Bell-shaped Frequency Distributions

- The Gaussian or Normal distribution has 2 parameters: the mean μ and standard deviation σ. In this distribution the frequency with which a value x occurs is:

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

  This formula is hardly ever needed in practice. What is more useful is to know that:

  - The frequency distribution is symmetric with a bell-shape;

Modelling Symmetric, Bell-shaped Frequency Distributions

- The Gaussian or Normal distribution has 2 parameters: the mean μ and standard deviation σ. In this distribution the frequency with which a value x occurs is:

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

  This formula is hardly ever needed in practice. What is more useful is to know that:

  - The frequency distribution is symmetric with a bell-shape;
  - About 70% of the x values lie within 1 s.d. of the mean μ;

Modelling Symmetric, Bell-shaped Frequency Distributions

- The Gaussian or Normal distribution has 2 parameters: the mean μ and standard deviation σ. In this distribution the frequency with which a value x occurs is:

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

This formula is hardly ever needed in practice. What is more useful is to know that:

  - The frequency distribution is symmetric with a bell-shape;
  - About 70% of the x values lie within 1 s.d. of the mean μ;
  - 95% of the observations lie within 2σ of μ; and

Modelling Symmetric, Bell-shaped Frequency Distributions

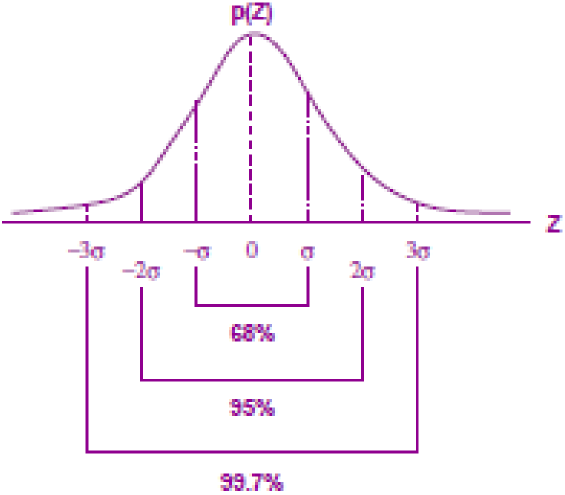
- The Gaussian or Normal distribution has 2 parameters: the mean μ and standard deviation σ. In this distribution the frequency with which a value x occurs is:

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

  This formula is hardly ever needed in practice. What is more useful is to know that:

  - The frequency distribution is symmetric with a bell-shape;
  - About 70% of the x values lie within 1 s.d. of the mean μ;
  - 95% of the observations lie within 2σ of μ; and
  - Nearly all (99.7%) of the observations lie within 3σ of μ
- If this theoretical distribution is used to fit observed data, then these properties must hold (at least approximately: a perfect fit will rarely happen)

The Standard Normal Distribution



Mean and Spread of the Standard Normal Distribution I

- The calculation of the mean uses the expression for expected values:
$$\text{Mean} = E(X) = \int_{-\infty}^{+\infty} x\phi(x)dx$$
Ignoring the constant for the moment,

$$\begin{aligned}
\text{Mean} &= \int_{-\infty}^{+\infty} xe^{-x^2/2}dx \\
&= \int_{-\infty}^{0} xe^{-x^2/2}dx + \int_{0}^{+\infty} xe^{-x^2/2}dx \qquad (\text{let } t = x^2/2) \\
&= \int_{\infty}^{0} e^{-t}dt + \int_{0}^{\infty} e^{-t}dt \\
&= -\int_{0}^{\infty} e^{-t}dt + \int_{0}^{\infty} e^{-t}dt \\
&= 0
\end{aligned}$$

Mean and Spread of the Standard Normal Distribution II

- Similarly, for the variance:

$$\text{Variance} = E(X^2) - (E(X))^2$$

- Now

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx \\ &= \int_{-\infty}^{0} x^2 e^{-x^2/2} dx + \int_{0}^{\infty} x^2 e^{-x^2/2} dx \end{aligned}$$

  With $u = x$ and $v = e^{-x^2/2}$, we get $dv = -xe^{-x^2/2}dx$. So, each of the integrals above is of the form $\int udv$, which can be evaluated by by parts as $\int udv = uv - \int vdu$.

- If you get all the steps right, you will find Variance = 1

Bounds using Expectations: Markov's Inequality I

- Suppose you are observing values X, and want to know how likely is it that X will be very large
  - If you knew the relative frequencies of the X's was well modelled by, say, a normal distribution with some μ and σ^2, you could provide a very precise answer to this
  - But what if you did not have a good theoretical distribution to model the data; or you did not know what the distribution was?
- Under some circumstances, Markov's inequality gives an upper bound on the probability of an unusually large value:
  - Suppose you did not know a theoretical distribution for the values, but you knew the mean μ of the distribution
  - Suppose all X values were known to be non-negative
- Then Markov's inequality states:

$$P(X \geq k) \leq \frac{\mu}{k} \qquad\qquad (k > 0)$$

Bounds using Expectations: Markov's Inequality II

- Or, in terms of expectations:

$$P(X \geq k) \leq \frac{E(X)}{k} \qquad (k > 0)$$

- Although a proof of Markov's inequality should really be done after you have had more experience with random variables, we can use what you know already:

$$\begin{aligned} E(X) &= \int_0^\infty xf(x)dx \\ &\geq \int_k^\infty xf(x)dx \\ &\geq \int_k^\infty kf(x)dx \\ &\geq kP(X \geq k) \end{aligned}$$

Bounds using Expectations: Markov's Inequality III

from which the result follows

Bounds: Chebyshev's Inequality I

- Suppoise you want to know how likely X will be from the mean. Again, suppose we know very little about X other than it is well-modelled by a theoretical distribution with mean μ variance σ^2. We want to know the value of

$$P(|X - \mu| \geq k) \qquad (k > 0)$$

- We cannot find this value when we do not know the precise functional form of the theoretical distribution. But we can find an upper bound on the value, using Markov's inequality:

$$\begin{aligned} P(|X - \mu| \geq k) &= P((X - \mu)^2 \geq k^2) \\ &\leq \frac{E[(X - \mu)^2]}{k^2} \qquad \text{Markov} \\ &\leq \frac{Var(X)}{k^2} \end{aligned}$$

Bounds: Chebyshev's Inequality II

- A variant of this is:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

  You can show that this follows in the same way as the other version

- Clearly, the bound is only useful if $k > 1$, since for $0 < k \leq 1$, the bound is trivially equal to 1

Chernoff Bounds I

- Chebyshev's inequality follows from a more general feature of Markov's inequality, namely: we can substitute any positive function f, such that:

  $$P(f(X) \geq f(k)) \leq \frac{E(f(X))}{f(k)}$$

  In order to obtain Chebyshev's inequality, we use a function $f(X) = X^2$.
- In general, it is possible to obtain tighter bounds using other kinds of functions that grow even faster than X^2. One example is the use of an exponential, which results in Chernoff bounds

Chernoff Bounds II

- Suppose X is the sum of n independent observations, each modelled by a theoretical distribution with mean $p_1, p_2, \ldots, p_n$. Then, the expected value of the sum μ is clearly $\sum_i p_i$. Using an exponential function in Markov's inequality, it is possible to show that

$$P(X \geq (1+\delta)\mu) \leq e^{\frac{-\delta^2\mu}{3}} \qquad (0 < \delta < 1)$$

$$P(X \leq (1+\delta)\mu) \leq e^{\frac{-\delta^2\mu}{2}} \qquad (0 < \delta < 1)$$

- More general Chernoff bounds that follow from these are:

$$P(X \geq (1+\delta)\mu) \leq e^{\frac{-\delta^2\mu}{2+\delta}} \qquad (0 \leq \delta)$$

$$P(X \leq (1+\delta)\mu) \leq e^{\frac{-\delta^2\mu}{2+\delta}} \qquad (0 \leq \delta)$$