Probability II

9.07

9/17/2004

Today

- The gambler's fallacy
- Expected value
- Bayes' theorem
- The problem of false positives
- Making a decision
- Making a decision the Bayesian way
- Probability with continuous random variables

The gambler's fallacy

- If you get 8 heads in a row, on a fair coin, what is the probability that the next one will be a tail?
- 0.5
  - Remember, coin tosses are independent
  - $P(T \mid 8\ H) = P(T) = 0.5$
- But, most people feel like, having tossed 8 heads, it must be more likely you get a tail next, to "even out" the number of heads. This is the *gambler's fallacy*.
- Basically, most human beings don't really, in their gut, believe in probability. But we can learn to rationally get the right answer.

Some paradoxes & puzzles, if the gambler's fallacy were true

- Anna gets 10 heads in a row on her coin. Lee gets 10 tails in a row on his coin. They then switch coins, and flip.
  - Is Anna more likely to get a head, to balance out all the tails on Lee's coin (which she's now flipping), or
  - Is Anna more likely to get a tail, to balance out all the heads she just flipped on her own coin?

Some paradoxes & puzzles, if the gambler's fallacy were true

- If you manage to get 10 heads in a row on a coin, have you "stored up" heads? Would you carry that coin to the nearest bar, and try to bet someone a lot of money the coin would come up tails?

Some paradoxes & puzzles, if the gambler's fallacy were true

- If I get 10 heads in a row on my coin, how do I know which situation I have:
  - The coin was "balanced" before the 10 heads. I expect to see some tails, soon, to balance out the recent heads. $(p(T)>0.5)$
  - Before, the coin had a surplus of tails. The 10 heads I saw evened it up. Now the coin is balanced, and I expect $p(T)=0.5$.
  - We still have a surplus of tails. I should expect to see more heads, to balance. $(p(T)<0.5)$

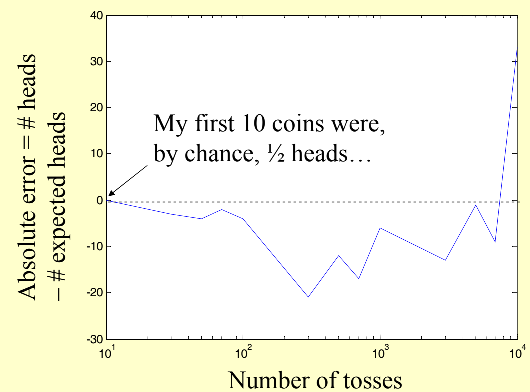
In some sense, the fact that you don't know which situation you're in means you should consider it equally likely that we now have a surplus of heads, vs. a surplus of tails. Your best guess would be to assume that the coin is now balanced.

So, even with the gambler's fallacy, your best guess should be that $p(T) = 0.5$

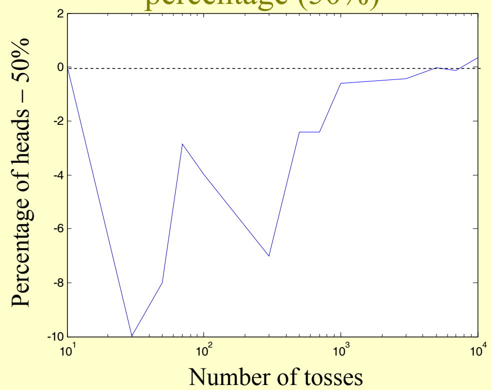
The law of averages

- In the *long run*, the relative frequency of an event approximates the probability of that event.
- So, why doesn't this mean that if you get 10 heads, you should expect to see some tails to balance them out?
  - Toss 550 heads in 1000 tosses. "Error" of 50 more heads than expected. Don't we need to toss some tails to get this error down to zero?

No, the absolute error actually goes *up* with more coin tosses. An example:



What does go down is the difference between the percentage of heads, and the expected percentage (50%)



A problem

- Which would you rather:
  - I give you $50, or
  - You flip a (fair) coin, and I give you $100 if you get a head, $0 if you get a tail.
- To answer this (rationally), we need the concept of *expected value*.

Expected value

- There are three balls in a bag. Two are black, one is white.
- If you draw a white ball, I'll pay you $1.
- If you draw a black ball, you pay me $0.50.
- Should you take the bet? What is the expected value, i.e. how much money do you expect to win or lose?

Expected value

- "Well, there's a certain probability that I'll pick a white ball, in which case I'll make $x, and a certain probability that I'll pick a black ball, in which case I'll have to pay $y, so on average…"

Expected value

- Expected value =

$$\sum p_i \cdot \text{value}(i)$$

sum over all possible outcomes ($\sum$); of the probability of that outcome (in the long run, this is the fraction of times you expect this event to occur) (p_i); times the value, or cost, of that outcome ($\text{value}(i)$)

Back to our problem

- There are three balls in a bag. Two are black, one is white.
- If you draw a white ball, I'll pay you $1.
- If you draw a black ball, you pay me $0.50.
- Your expected value = E =
  (1/3) $1 + (2/3) (-$0.50) = $0
- In the long run, you expect to break even with this bet.

Another problem

- I'll give you $0.90 each time you roll a six. You pay $0.30 to play (i.e. to roll the die). Should you take the bet?
- Your E = (1/6) ($0.60) + (5/6) (-$0.30) = -$0.15
  - Negative, so you expect to *lose* money
  - You should not take the bet.
  - If you play 10 times, you expect to lose $1.50 (this is just the *expected* loss, however)
- Note that my E = (1/6)(-$.60)+(5/6)($.30)=$0.15

Back to our original problem

- Which would you rather:
  - I give you $50.
  - You flip a (fair) coin, and I give you $100 if you get a head, $0 if you get a tail.
- E1 = $50.
- E2 = (0.5) $100 + (0.5) $0 = $50
- The two are equivalent. (People tend to prefer the certain case where I give you $50, because somehow that's "certain"…)

Another example: what does it mean if someone gives you 3-to-1 odds?

- You bet \$x on an outcome.
- If it occurs, you win \$3x.
- If it does not occur, you pay \$x.
- If the bookie has correctly set the odds, what is the probability that the outcome occurs, i.e. what is the probability, p, for which you expect to break even?
- What about for n-to-m odds?

Bayes' Theorem

- Recall from last time, a version of the multiplication rule:

  P(E and F) = P(F|E) P(E) = P(E|F) P(F)
- Rearranging, we get *Bayes' Theorem* (a.k.a. Bayes' Rule):

  P(F|E) = P(E|F) P(F) / P(E)
- This is a powerful rule for many practical applications.

Example: deciding if you played the good chess program or the easier one

- There are two chess programs on the computer. The good one (G) beats you 75% of the time. The mediocre (M) one wins 50% of the time.
- You randomly pick one of the programs (you don't know which), and play two games. It wins both times. (It wins = W, it loses = L)
- What is the probability that you played the mediocre program, M?

Which program did you play?

- P(M | WW) = ?
  - This is difficult to directly compute
  - But, switching it around, it's easy to figure out P(WW | M)
  - Bayes' rule lets you use the easy P(WW | M) to compute the more difficult P(M | WW)

Which program did you play?

- Bayes' rule, again:
  $P(F|E) = P(E|F) P(F) / P(E)$
- $P(M \mid WW) = P(WW \mid M) P(M) / P(WW)$
- $P(WW \mid M) = P(W \mid M)^2 = (½)^2 = ¼$
- $P(M) = ½$
- P(WW) is a bit tricky

Which program did you play?

- There are two ways you could have seen the computer program win the 2 games: 1) You could be playing the mediocre game, and it could win twice; 2) You could be playing the good game, and it could win twice.
- P(WW) = P(M and WW) + P(G and WW)
  - Mutually exclusive because M and G can't both be true.
- Using the multiplication rule,
  - $P(WW) = P(WW \mid M) P(M) + P(WW \mid G) P(G)$
    $= ½ * ½ * ½ + ¾ * ¾ * ½$
    $= 0.40625$

Which program did you play?

- Bayes' rule, again:
  $P(F|E) = P(E|F) P(F) / P(E)$
- $P(M \mid WW) = P(WW \mid M) P(M) / P(WW)$
- $P(WW \mid M) = P(W \mid M)^2 = (½)^2 = ¼$
- $P(M) = ½$
- $P(WW) = 0.40625$
- So, $P(M \mid WW) = ¼ * ½ / 0.40625 \approx 0.31$
- Note that this is < 0.5. The probability of playing the mediocre program has gone down a small amount, since you lost twice in a row (which makes it more likely you played the good program).

Bayes theorem and the problem of false positives

- Suppose a rare disease infects 1 out of 1000 people in the population.
- There exists a good, but imperfect, test for the disease.
  - If you have the disease, the test is positive 99% of the time.
  - However, there are *false positives*. About 2% of uninfected patients who are tested also test positive.
- You just tested positive. What are your chances of having the disease?

Bayes & false positives

- Two events. Let
  D = patient has the disease (not D = they don't)
  + = patient tests positive ("–" = negative test)
- We want to know P(D | +)
  = P(+ | D) P(D) / P(+)
- We know:
  P(+ | D) = 0.99
  P(D) = 0.001
  P(+ | not D) = 0.02 what is P(+)?

Bayes & false positives

- P(+) = P(+ and not D) + P(+ and D)
  = P(+ | not D) P(not D) + P(+ | D) P(D)
  = 0.02 * (1-P(D)) + 0.99 * 0.001
  = 0.02 * 0.999 + 0.99 * 0.001
  = 0.02097
- So, P(D | +) = P(+ | D) P(D) / P(+)
  = 0.99 * 0.001 / 0.02097
  ≈ 0.0472
- If you test positive for the disease, you have less than a 5% chance of actually having the disease!
  - Phew, only about 1 in 20! Although, on the other hand, before the test you thought your chances were 1 in 1000…

Some implications

- Probably you want to have some more tests done.
- Because of this problem of false positives, ironically some tests actually *reduce* your life span, and thus are not recommended.
  - Cancer A occurs in 1 in 1000 men < 50 years old.
  - Probability of having cancer A if test is positive: 1 in 20, as in the previous example.
  - 90% of patients testing positive have surgery to remove the cancer. (They know they may not have it, but they're worried.)
    2% of them die due to complications, the rest are cured.
  - Death rate (in 10 years) from cancer A = 25%
  - P(death due to cancer A) = P(death | cancer A) P(cancer A)
    = 0.25 * 0.001 = 0.00025
  - P(death due to test) ≈
    P(die | surgery) P(surgery | positive) P(test positive)
    = 0.02 * 0.90 * 0.02 = 0.00036 > 0.00025

Do you have a fair coin?

- You flip a coin 5 times, and see 5 heads in a row. Is this a fair coin?
- P(5 heads in a row | fair coin) =
  $(\frac{1}{2})^5 \approx 0.03$
- This is pretty unlikely to happen – you might want to bet it's not a fair coin.

  That's not a bad way to make the decision, but if you have more information you can do better.

Bayes, and deciding if the coin is fair

- P(fair coin | 5 heads) =
  P(5 heads | fair coin) P(fair coin)/P(5 heads)
- What is P(fair coin)? P(5 heads)?
  - Often you don't know.
  - But, for this example, pretend we know…

Bayes, and deciding if the coin is fair

- Suppose you have a jar of 100 coins. 99 are fair, one is unfair. The unfair one has two heads.
- P(fair coin) = 0.99
- P(5 H) =
  P(5 H|fair)P(fair) + P(5 H|not fair)P(not fair)
  $= (\frac{1}{2})^5 (0.99) + 1\ (0.01)$
  = 0.0409
- P(fair coin | 5 H) =
  P(5 H | fair coin) P(fair coin) / P(5 H)
  $= (\frac{1}{2})^5(0.99)/0.0409 \approx 0.76$
- Thus, despite the fact that you got 5 heads on 5 flips of the coin, it's pretty likely it *was* a fair coin…

With a different prior on P(fair coin)

- P(fair coin) = 0.5
- P(5 H) =
  P(5 H|fair)P(fair) + P(6 H|not fair)P(not fair)
  $= (\frac{1}{2})^5 (0.5) + 1\ (0.5)$
  ≈ 0.52
- P(fair coin | 5 H) =
  P(5 H | fair coin) P(fair coin) / P(5 H)
  $= (\frac{1}{2})^5(0.5)/0.52 \approx 0.03$
- Thus, if we don't know how likely it is that the coin is fair, and thus assume the fair coin is just as likely to happen as not, then we get about the same result as before we used Bayes': this isn't likely to be a fair coin.

Decisions, decisions…

- The statistical tests we'll talk about in this class, and 99.99% of tests done in BCS research, essentially amount to saying, "gosh, 5 heads in a row is pretty unlikely if you've got a fair coin – maybe it's not fair"
- If you've got more information about the prior probabilities, you can make better decisions with Bayes'.
- But often we don't, so the first method is not bad.

Continuous random variables

- So far, a lot of this probability stuff has involved discrete random variables
  - There were a discrete, finite set of possible outcomes, e.g. get a tail on one coin flip, roll a double 6
- Much of what we do in statistics, however, involves continuous random variables, with a (theoretically) infinite number of possible outcomes.
  - E.G. height of a college student, reaction time on a test
  - (Note that infinite does not mean unbounded.)

What's different, with continuous random variables?

- $P(O_i) = 0$
  - There are an infinite number of possible outcomes, and the sum (integral, really) of their probabilities must equal 1. So each one has probability 0.
- As a result, we need to look at probability that the outcome falls in some range, instead, e.g.
  $P(x>4.5)$
  $P(2.0<x<6.3)$
- Discrete case: $P(O_1 + O_2 \ldots) = \text{sum}(P(O_i))$ for elementary outcomes.
- Continuous case: $P(x>4.5)$ an integral of the distribution function from x=4.5 to x=infinity

The "normal approximation"

- A lot of times in statistics, we'll compute probabilities like $P(x>4.5)$ by assuming that the data is distributed according to a familiar distribution, such as a normal distribution.
  - This is the "normal approximation" previously mentioned.

The normal curve approximation for data

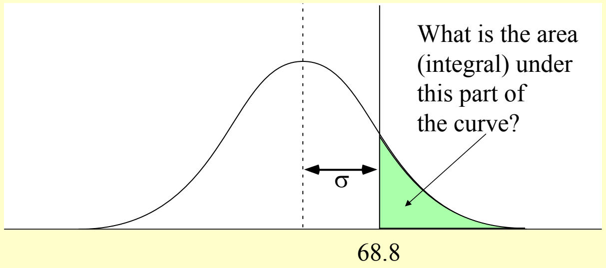
- For many types of data, the normal curve is a good approximation to the distribution of the data.
- Use this approximation to generalize from the sample to the larger population.
- First:
  - Estimate the population mean, μ, from the sample mean $\bar{x}$.
  - Estimate the population std. deviation, σ, from the sample standard deviation, s.

Example question you can ask

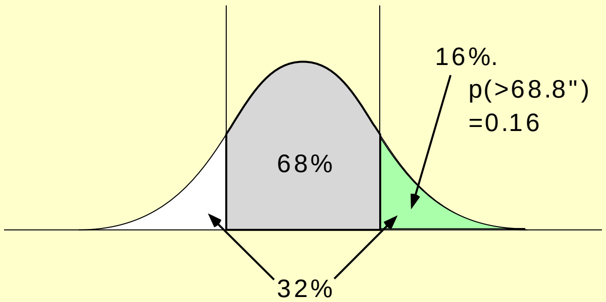
- Mean height for this class is 65.7 inches.
- Std. dev. = 3.1 inches
- Assume the broader population we're interested in is MIT students in general.
- What's the probability that a student will be taller than 68.8 inches (P(h>68.8"))?

Using what we know about normal distributions

- Well, 68.8" is one std. dev. above the mean height.



Recall that 68% of a normal distribution lies between +/- 1 s.d.



For less simple examples, use a table

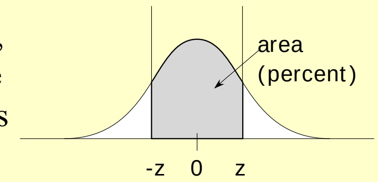
- We know the area for a normal distribution for +/- 1 s.d. (68%), +/- 2 s.d. (95%), and +/- 3 s.d. (99.7%), but what about for 1.65 s.d.?
- For this, we use a table, e.g. p. A-105 of your book. (Statistical packages will have these tables built in.)

Using a table

- The table is for a z-distribution
- This is a “standardized” normal distribution, i.e. its mean is 0, and its standard deviation is 1.
- This means you don’t look up P(x>4.5), x~N(3, 0.5), you instead look up P(x more than 3 s.d. bigger than the mean), i.e. P(z>3).
  - This is just like what we were doing in the last example There’s nothing new but to tell you this means we’re using a z-distribution.

What the table tells you

- For each value of z, what fraction of the total probability lies between +/- z s.d.?

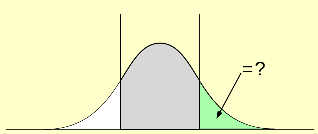


z	Height	Area
0.00	39.89	0
0.05	39.84	3.99
0.15	39.69	7.97

(ignore the “Height” column…)

An example of using the table

- Same as our earlier example. Mean height = 65.7”, s.d. = 3.1”, what is the probability that a student is taller than 68.8”?
- I.E., what is P(z>1)?



An example of using the table

- Want P(z>1). Look up P(-1<z<1)
- = 68.27.
  (Some tables will give you $P(z>z_o)$ directly.)
- P(z>1) = (1-0.6827)/2 ≈ 0.16, as before.

z	Height	Area
…	…	…
0.95	25.41	65.79
1.00	24.20	68.27
1.05	22.99	70.68
…	…	…

Another example of using the table

- Mean(x) = 5,
  s.d.(x) = 2,
  what is
  P(x>6.5 or x<3.5)?
- We need the z-scores for 6.5 and 3.5 – how many standard deviations are 6.5 and 3.5 from the mean, 5?
- $z_1 = (6.5 – 5)/2 = 0.75$
- $z_2 = (3.5 – 5)/2 = -0.75$
- So we want P(z>0.75 or z<-0.75), and we can get this from looking up P(-0.75 < z < 0.75).

Another example of using the table

- P(-0.75<z<0.75) = 0.5467
- P(z<-0.75 or z>0.75) = 1-0.5467 ≈ 0.45
- That's our answer.

z	Height	Area
…	…	…
0.70	31.23	51.61
0.75	30.11	54.67
0.80	28.97	57.63
…	…	…

What do I do if I need to look up z=0.752?

- For this class, don't worry about it. Round to the nearest z in the table.
- In your research life, the computer will look it up for you, interpolating if necessary.

Another example

- Mean = 4, s.d. = 1, what is p(x<1.5 or x>4.5)?
- z(0.5) = (1.5 – 4)/1 = -2.5
- z(4.5) = (4.5 – 4)/1 = 0.5
- What do we look up? Well, all we really can look up is z=2.5 and z=0.5, and figure it out from there.

Another example

- P(-2.5 < z < 2.5) = 98.76/100
- So, as in an earlier example, P(-2.5<z) = P(z<2.5) = (1-0.9876)/2 ≈ 0.006
- P(-0.5 < z < 0.5) = 38.29/100
- P(z>0.5) = (1-0.3829)/2 ≈ 0.31
- So P(x<1.5 or x>4.5) ≈ 0.316

More differences with continuous random variables

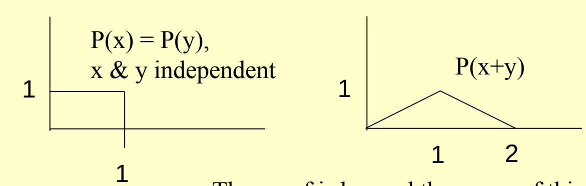
- More complicated events are trickier to do in your head.
  - P(sum of 2 dice = 3) = P({1,2},{2,1}) easy
  - x~*N*(3, 2), y~*N*(4,1), P(x+y > 5) = ??

Normal distribution

- Distribution of x+y, given x and y both normal distributions?
- You can do it empirically, in MATLAB, to get an estimate.
- This is a homework problem.

Some examples

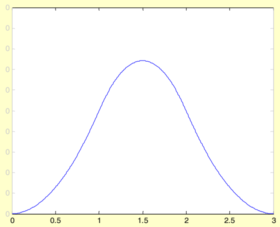
- If x and y are both distributed according to a uniform distribution on (0,1), what is the distribution of x+y?



The proof is beyond the scope of this course, but if you know about convolution, P(x+y) = P(x) convolved with P(y).

x+y+z?

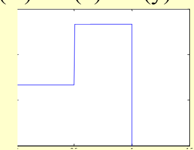
- P(x+y+z), x, y, z all uniform on (0,1)?, x, y, & z independent.



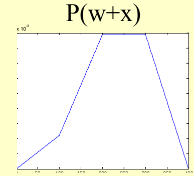
- Hmm, this is starting to look familiar…

Another example (again, assume independence)

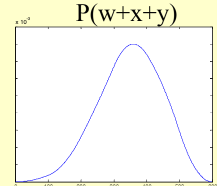
P(w) = P(x) = P(y) = P(z)



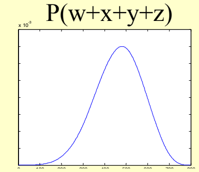
P(w+x)



P(w+x+y)



P(w+x+y+z)



Hmm…

The central limit theorem

- As you add N independent random variables, their sum tends to look more and more like it is normally distributed as N gets larger, regardless of the distributions of the random variables.
- This a fuzzy version of what is known as the "central limit theorem".
- What it means: data that are influenced by (i.e. the "sum" of) many small and unrelated random effects tend to be approximately normally distributed.
- This is why so many kinds of data are approximately normally distributed.
  - E.G. weight is a result of genetics, diet, health, what you ate last night, … And thus it tends to be approximately normally distributed.