ST112
Notes for Week VIII
Spring 2011

Probability and Statistics

Goals: (i) A bare bones understanding of probability and conditional probability; (ii) An understanding of margin of errors in polls/sampling.

Probability basics

Given an experiment or observation, the *sample space* is the set S of all possible outcomes.

Examples: (1) Flip a coin four times. $S = \{HHHH, HHHT, ..., TTTT\}$.
(2) Measure the temperature in Boston. $S = \{t : -25 \leq t \leq 105\}$.

Let $\mathcal{S}$ be the set of all subsets of S. A *probability measure* is a function $P : \mathcal{S} \to [0, 1]$ with

$$P(A) = 1, P(A^c) = 1-P(A), A \subset B \Rightarrow P(A) \leq P(B), A\cap B = \emptyset \Rightarrow P(A\cup B) = P(A)+P(B).$$

Think of $P(A)$ as the probability that the outcome of your experiment lies in A.

Examples: (1) For a fair coin, $P(A) = 1/16$ if A has one element, e.g. $A = \{HTHH\}$. This is a *uniform distribution.*
(2) Maybe $P(\{t : 50 \leq t \leq 70\}) = .65$. We usually just write $P(50 \leq t \leq 70) = .65$

Conditional probabilities

This keeps track of prior information. We want to compute $P(B|A)$, which informally is the probability of B occurring knowing that A has occurred.

Defintion: The conditional probability of B given A is

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

For a uniform distribution, this computes the fraction of A taken up by $B \cap A$, those outcomes lying in both B and A.

Examples: (!) (1) If you've flipped a fair coin three times and gotten HHH, what are the chances you'll get H on the fourth flip? Many people will say, "Very small." So let B be the subset of S consisting of those outcomes with an H on the last flip, and let $A = \{HHHH, HHHT\}$ be the subset consisting of all outcomes with three heads in the beginning. Then

$$\begin{aligned} & P(\text{heads on the 4th flip }|\text{heads on the first 3 flips}) \\ &= P(B|A) = \frac{P(B \cap A)}{P(A)} \\ &= \frac{P(\text{all heads})}{P(\text{heads on first 3 flips})} = \frac{1/16}{1/8} = \frac{1}{2}. \end{aligned}$$

Note that $P(B|A) = P(B)$, so the outcome on the 4th flip is independent of what happened on the first three flips – there is always a 50-50 chance of getting H on any given flip. Since you can compute $P(A|B) = P(A)$, A and B just don't care if the other happens or not. So if $P(B|A) = P(B) = P(A|B)$, we call A and B *independent* sets/events.

(2) I estimated I've commuted 3000 days to BU. Since I don't always commute at rush hour, I think that 60% of the time there's a traffic jam (TJ) during my commute. If there is a traffic jam, the chances that I get to work on time (WOT) is 35%. If there isn't a traffic jam (NTJ), chances of WOT jump to 90%. What is $P(WOT)$?

Answer: By the definition of conditional probability,

$$\begin{aligned} P(WOT) &= P((WOT \cap TJ) \cup (WOT \cap NTJ)) \\ &= P(WOT \cap TJ) + P(WOT \cap NTJ) \\ &= P(WOT|TJ)P(TJ) + P(WOT|NTJ)P(TJ) \\ &= (.35)(.6) + (.9)(.4) \\ &= .576 \end{aligned}$$

Note that TJ and NTJ break the sample set of 3000 commutes into two distinct subsets. We can turn this process around to predict the chances that there was a traffic jam from observing if I get to work on time.

Here's the abstract but not difficult theorem:

Bayes' Theorem: If S is partitioned into disjoint sets $B_1, ...B_k$, then for any r,

$$P(B_r|A) = \frac{P(B_r)P(A|B_r)}{P(B_1)P(A|B_1) + ... + P(B_k)P(A|B_k)}.$$

Example: If I get to work on time, the chances that there was a traffic jam is

$$P(TJ|WOT) = \frac{P(TJ)P(WOT|TJ)}{P(TJ)P(WOT|TJ) + P(NTJ)P(WOT|NTJ)} = \frac{(.6)(.35)}{(.6)(.35) + (.4)(.9)} \approx .365$$

Random Variables

A *random variable* is just a function $X : S \to \mathbb{R}$ on the sample space. We think of X as a measurement applied to our observed set of elements S.

Examples: (1) For the coin flip, X of any sequence of 4 H's and T's could be the number of H's.

(2) If S is the set of people in the US, for $p \in S$, set $X(p)$ to be the height of person p in feet and inches.

The *probability distribution* of X is a function f : (range of X) $\to [0, 1]$ given by

$$f(x_0) = P(\{p \in S : X(p) = x_0\}) = P(X = x_0).$$

(This is for the case where the range of X is a discrete set of real numbers – see below).

Examples: (1) Draw the probability distribution for the 4 coin flip, where X counts the number of heads.

(2) Do the same for the US population, where X measures the height.

(3) If S is the set of all times between 1850 and now, and $X(t_0)$ is the temperature at time t_0 in Boston, then the range of X is the whole interval $[-25, 105]$. So this is the opposite of the discrete case. Almost surely the probability distribution has $f(x_0) = 0$, so we'll define the *cumulative distribution* to be $F(x_0) = P(X \leq x_0)$, so now $F : [-25, 105] \to [0, 1]$ is an increasing function. For example $F(55) = 1/2$ (maybe), and $F(105) = 1$.

In the continuous case (3), for nice X, we can find a function f with

$$F(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} f(x)dx.$$

f is called the *probability density* of F. In the discrete case, this f coincides with the previous f.

Expectation values

The *expectation value* or *mean* of a random variable X with probability distribution/density f is

$$\mu = \sum xf(x) \left(= \int_{-\infty}^{\infty} xf(x)dx \right)$$

depending on whether we're in the discrete or continuous case. The expectation is the expected/average output from the measurement X.

Example: For the 4 times coin flip, let's say I get one dollar for every H. The expected pay off, or average payoff in the long run (doing a 4 coin flip experiment a thousand times) is

$$0(1/16) + 1(1/4) + 2(3/8) + 3(1/4) + 4(1/16) = 2.$$

So if the person running this game charges me more than \$2 per 4 coin flip, I should decline.

The *variance* of a random variable is

$$\sigma^2 = \sum x^2 f(x) \left(= \int_{-\infty}^{\infty} x^2 f(x)dx \right),$$

and $\sigma > 0$ is the *standard deviation.* A small σ means that f is concentrated near its mean μ, and a small σ means f is smeared out.

The normal distribution

The most important probability distribution is the *normal distribution with mean μ and variance σ^2*, which has probability density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

The *standard normal distribution* ($\mu = 0, \sigma^2 = 1$) is the *bell curve* $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

So for a random variable X with standard normal distribution,

$$P(a \leq X \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}x^2} dx,$$

which we can only compute via tables.

Note: If X has a normal distribution with mean μ, variance σ^2, then $Z = \frac{X-\mu}{\sigma}$ has a standard normal distribution.

Random Sampling

If $S = \{\text{people in the US}\}$, and $X : S \to \mathbb{R}$ is the height function, in practice we can't compute X on all of S. We can take sample sets e.g. $\{p_1, p_2, ..., p_{100}\} \subset S$ and compute X on this subset. To guard against a poor choice of this sample set, let's pick 20 such subsets $S_1, ...S_{20}$, and 20 random variables $X_1 : S_1 \to \mathbb{R}, ..., X_{20} : S_{20} \to \mathbb{R}$. To make sure we're sampling and measuring wisely, we assume

(i) The probability distributions of each X_i are the same. (This wouldn't be the case if we stupidly picked S_{17} to consist of only basketball players and measured their heights.)

(ii) The X_i are independent, i.e. $P(X_i = a, X_j = b) = P(X_i = a)P(X_j = b)$ for $i \neq j$. (This wouldn't be the case if we stupidly picked S_i to be the same as S_j.

Given these conditions, $X_1, ..., X_{20}$ are called a *random sample* of X.

We are then interested in

$$\bar{X} = \frac{X_1 + ... + X_{20}}{20},$$

which is the sample average. So we can think of $\bar{X} : \{1, 2, ..., 100\} \to \mathbb{R}$, with e.g.

$$\bar{X}(7) = \frac{1}{20} \left(X_1(\text{seventh person in } S_1) + ... + X_{20}(\text{seventh person in } S_{20}) \right).$$

The Central Limit Theorem: Let $X_1, ..., X_n$ have the same probability distribution functions with mean μ and variance σ^2. Assume that the X_i are all independent. Then in the limit as $n \to \infty$, the random variable

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

goes towards the standard normal distribution.

This means that as we sample more and more often, the average of the samples looks more and more like a normal distribution with mean μ, and which is concentrated more and more near μ. So our sampled average is looking more and more like a only slightly smeared version of the true average.

Confidence Intervals

We feel that by sampling techniques, we should be able to measure e.g. the mean μ of a random variable X on a sample space S which is too big for practical purposes. We start by taking e.g. 20 samples, and produce an average sample function $\bar{X}$. We would like to ensure that $|\bar{X} - \mu|$ is small, but we can't. There's always the possibility that our samples produce an $\bar{X}$ that is very far from μ. (It could happen that although we tried to pick

$S_1, ..., S_{20}$ “randomly,” they all consisted of NBA players.) On the other hand, coming up with such a bad result seems highly unlikely.

So we compromise, and ask for some interval $[a, b]$ such that $P(\mu \in [a, b]) \geq .95$. Then $[a, b]$ is called the 95% confidence interval. Of course, we can change .95 to any number we want in $[0, 1]$. In concrete terms, this says that the odds are greater than 95% that the true mean lies in $[a, b]$.

Warmup example: Let X have probability distribution the standard normal distribution. Using tables, we see that

$$\frac{1}{\sqrt{2\pi}} \int_{-1.96}^{1.96} e^{-\frac{1}{2}x^2} dx = .95.$$

Thus $P(X \in [-1.96, 1.96]) = .95$, so $[-1.96, 1.96]$ is the 95% confidence interval for X. In summary, if a measurement X on a sample space S has $P(X \in [a, b]) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}x^2} dx$, then 95% of the time $X(s) \in [-1.96, 1.96]$ for $s \in S$.

Back to our sampling situation, the Central Limit Theorem says that for $n \gg 0$ (usually $n > 30$ is good enough)

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \in [-1.96, 1.96]\right) = .95,$$

up to some small error, so a little algebra and the (nontrivial) theory of estimators give

Theorem: If a random variable X has a normal distribution with mean μ and variance σ^2, then

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95.$$

(This is unrealistic, as we’re trying to estimate μ, so why should we know σ? There are more refined versions when σ is unknown.)

Example: We want to sample the age of the senior population in the US. Somehow we know that $\sigma^2 = 225$. We take 20 samples and observe an average sampled age of $\bar{X} = 64.3$. Then

$$64.3 - 1.96\frac{15}{\sqrt{20}} = 57.7, \quad 64.3 + 1.96\frac{15}{\sqrt{20}} = 70.9,$$

so

$$P(57.7 < \mu < 70.9) = .95.$$

In other words, with 95% confidence, the average age of the senior population lies in $[57.7, 70.9]$.

References

1. J. Freund and R. Walpole. *Mathematical Statistics.* Prentice-Hall, New Jersey, 1980.

2. R. Walpole and R. Myers. *Probability and Statistics for Engineers and Scientists.* Macmillan, New York, 1985.