From probability to statistics, and back

Data to decisions

The premise

- Data (i.e. samples) are draws from a *data-generating-process* (DGP) or *population*

- Probability is the formal language we use for defining/describing DGPs

- Statistical inference is the process of using samples to learn about the DGP

The language of probability

- Let S be the set of possible states of the world (the "*universe*")
- Roll of a fair dice: $S = \{1,2,3,4,5,6\}$
- An *event* is a subset of S
- Ex: $A = \{2,4,6\}$ is the event that the roll is even
- A *probability distribution* is a function that assigns probabilities to each possible state of the word
- Ex: If dice is fair, $P(s) = 1/6$ for all $s \in \{1,2,3,4,5,6\}$, and, for any event A:

$$P(A) = \frac{\#A}{\#S}$$

Random variables

- A random variable X attaches a value to each possible state of the world

- Called *discrete* or *categorical* if it can assume only a finite (or at least *countable*) set of values…
- … *continuous* if it can take any value on an interval or collection of intervals

- Ex: X pays $\$1$ of roll of dice is even, nothing otherwise:
$$P(X = 1) = P(s \in \{2,4,6\}) = 0.5$$

- $P(X)$ is *the probability distribution* of X

Expectations

- The *expected value* of a random variable X is defined as:

$$E(X) = \sum_{s \in S} P(s)\, X(s)$$

- X pays \$1 if roll of dice is even, nothing otherwise:

$$\begin{aligned} E(X) = {} & P(s=1) \times 0 + P(s=2) \times 1 \\ & + P(s=3) \times 0 + P(s=4) \times 1 \\ & + P(s=5) \times 0 + P(s=6) \times 1 \\ = {} & 0.5 \end{aligned}$$

Variances and standard deviations

- $VAR(X) \quad = \sum_{s \in S} P(s)\,(X(s) - E(X))^2$
  $\qquad\qquad = \; E[X - E(X)]^2$

- X pays \$1 of roll of dice is even, nothing otherwise:
  $VAR(X) =$

  $P(s = 1)\,(0 - 0.5)^2 \; + \; P(s = 2)\,(1 - 0.5)^2$
  $\; + P(s = 3)(0 - 0.5)^2 + \; P(s = 4)\,(1 - 0.5)^2$
  $\; + P(s = 5)(0 - 0.5)^2 \; + \; P(s = 6)(1 - 0.5)^2$

  $\quad = \; 0.25$

- The standard deviation of X is the square root of its variance

Risk

- A random variable X is *risk-free* if:

$$VAR(X) = 0 \;\Leftrightarrow\; X(s) = x \;\text{ for all }\; s \in S$$

- It is *risky* if $VAR(X) > 0$

Entropy

- Like variance, it is a measure of uncertainty
- But it measures not so much how far apart realizations of X can be
- Rather how "complex" the distribution is
- Very elegantly, it measures how many words/messages you'd have to send on average to describe the draw
- Formally,

$$Entropy(p) = -\sum_x p(x)\log[p(x)]$$

where log base 2 is often used for elegance's sake

- Notes:
  1. If there is only one possible value, entropy is zero
  2. Without constraints, entropy is maximized when all possible values of X are *equiprobable*
  3. If there are two possible values entropy is maximized at 50-50, where entropy is 1 (in base 2)

Multiple random variables

- Data are joint observations of multiple random variables: age, income, spending…
- One of the main game we play is to try and use some of these variables to predict others
- Ex: given someone's age and income, what is the best possible forecast of their spending over the next year?
- This amounts to learning about the *joint probability distribution* of these variables
- DGPs are joint probability distributions
- Given data (a few joint observations of X, Y, Z), what can we say about the DGP? And with what confidence?

Covariance

- We need a notion of how two random variables X and Y are related:

$$\begin{aligned} COV(X,Y) &= \sum_{s\in S} P(s)\,(X(s)-E(X))(Y(s)-E(Y)) \\ &= E[(X-E(X))(Y-E(Y))] \end{aligned}$$

- $COV(X,Y) > 0$ means that X tends to be high when Y tends to be high, and vice-versa

- Note 1: if X is risk-free, then $COV(X,Y) = 0$
- Note 2: $COV(X,X) = VAR(X)$
- Note 3: $COV(X,Y) = COV(Y,X)$

Example

- X pays \$1 if roll of dice is even, Y pays \$1 if roll of dice is 4 or more

- Then $E(X) = E(Y) = 0.5$, and:

$$COV(X,Y) =$$
$$P(s=1)(0-.5)(0-.5) + P(s=2)(1-.5)(0-.5) +$$
$$P(s=3)(0-.5)(0-.5) + P(s=4)(1-.5)(1-.5) +$$
$$P(s=5)(0-.5)(1-.5) + P(s=6)(1-.5)(1-.5)$$
$$= 1/12$$

Coefficient of correlation

- $\rho_{X,Y} = COV(X, Y)\,/(\sigma_X \sigma_Y)$
- Varies from -1 to 1
- $\rho_{X,Y} = 1$ means that $Y = a\,X\ + b$, where $a > 0$
- $\rho_{X,Y} = -1$ means that $Y = a\,X\ + b$, where $a < 0$

Example

- X pays $1 of roll of dice is even, Y pays $1 if roll of dice is 4 or more:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \frac{\frac{1}{12}}{\sqrt{0.25 \times 0.25}} = \frac{1}{3}$$

Independence

- X is *independent* of Y if knowing something about X does not change the probability distribution of Y
- If X and Y are independent then $COV(X, Y) = 0$
- If X and Y are *dependent* then knowing X is useful for forecasting Y
- We just need to understand or *model* that dependence in order to exploit it

A very useful expression

- $VAR(aX + bY) =$

$$a^2 VAR(X) + b^2 VAR(Y) + 2abCOV(X, Y)$$

- In words, when you add/combine two random variables, the variance of the combination depends on how risky each variable is but also on how they co-vary with one-another

Statistics

- Data are draws from the DGP
- Given data, what can learn about the DGP?
- In particular, can we find systematic patterns that will be useful for forecasting purposes?

Sample description: univariate

- Sample means, standard deviations and other such statistics are all "estimates" of the corresponding features of the DGP…
- … as long as the sample is representative
- i.e. as long as it was drawn without bias

Map from samples to DGP

- Sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- An estimate of the DGP's expectation or population mean
- Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

- An estimate of the DGP's variance
- …

Multivariate inference

- Sample covariance:

$$s(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

- An estimate of the DGP's covariance

- And more generally, any regression on a sample is an estimate of what the same regression would produce on whole population

A quick aside on the Bessel Correction

- Why are sample variances and covariances divided by $n - 1$ rather than n?
- This is called the *Bessel correction*
- We have already used the sample to estimate the mean, which *biases* estimates of dispersion down
- Dividing by $n - 1$ removes that bias
- Immaterial for n large, obviously

Law of large numbers

As long as draws are not overly correlated, sample estimates converge to their population counterparts

Building confidence

- Samples enable us to make statements about the population/DGP
- But how confident should we be about those statements?
- Statistics are random variables (different samples give different answers) so they have a distribution
- The dispersion in those distributions is telling us how confident we should be about our sample-based generalizations

Key law 1: the normal distribution

- The sacrosanct bell curve, ubiquitous in nature
- Describes a continuous random variable whose distribution is completely described by its expectation and variance
- For any two numbers a and b, gives the probability that the variable will fall in $[a, b]$
- **Ex1: with 95% probability, a draw from a normal distribution is within $\mathbf{1.96}\boldsymbol{\sigma}$ of its expectation, where $\boldsymbol{\sigma}$ is the standard deviation**
- **Ex2: with 95% probability, a draw from a normal distribution is less than $\boldsymbol{\mu} + \mathbf{1.645}\boldsymbol{\sigma}$**
- Many statistics (the mean of a suitably large and representative sample, e.g) are approximately normally distributed
- The standard normal distribution is the normal distribution with mean 0 and standard deviation 1

A very useful fact

- If X is normally distributed with mean μ and standard deviation $\sigma > 0$ then

$$Z = \frac{X - \mu}{\sigma}$$

follows a standard normal distribution

- If X is a sample statistic with known expectation μ and standard error σ then Z is called a z-score

- If statistic X is roughly normally distributed Z should be within -1.96 and +1.96 in 95% of samples

Key law 2: the t-distribution

- To the naked eye, looks a lot like the standard normal distribution
- But it has father tails, it attaches more likelihood to draws far away from the mean
- Characterized by its number n of degrees of freedom.
- Useful for hypothesis testing:
  1. The mean of n independent draws from a normal distribution (properly scaled, see next chapter) follows a t-distribution with $n-1$ degrees of freedom
  2. The ratio of coefficients to standard errors in a regression is a statistic that is also t-distributed
  3. Z-scores when σ is unknown
- When n is large, the t-distribution becomes the standard normal distribution

Key law 3: the chi-squared distribution

- Distribution of the sum of the square of n independent draws from a standard normal distribution
- Characterized by its number n of degrees of freedom
- Expectation is n, variance is $2n$
- Useful for hypothesis testing:
  1. Do two samples come the same population?
  2. Are two random variables independent?

The central limit theorem

- Assume we draw a random sample of size n from a population/DGP with mean μ and standard deviation σ
- For n "large", the sample mean $\hat{\mu}$ is roughly normally distributed with mean μ and standard deviation $\sqrt{\sigma^2/_n}$

Back to our quality control example

- $H0$: Machine true defect rate is $\pi \leq 1\%$
- The DGP/Population's standard deviation is $\sqrt{\pi(1-\pi)}$
- So, for a sample of size n, the standard deviation (or *standard error*) of the mean is $\sqrt{\frac{\pi(1-\pi)}{n}}$