MPO1: Quantitative Research Methods
Session 2: Random Variables and Probability Distributions

Thilo Klein

University of Cambridge
Judge Business School

- Probability – Bayes' Theorem
- Probability distributions, random variables, moments of r.v.s
- Specific probability distributions: e.g., Normal
- Properties of estimators
- Sampling distributions of the mean estimator and confidence intervals

Probability Rules

Gambling consult from last week

- Chevalier de Mere to Blaise Pascal : What is more likely?
  - Rolling at least one 6 in four throws of a single die
  - Rolling at least one double 6 in 24 throws of a pair of dice

Simulation Exercise (de Mere's problem)

```
set.seed(123)

dice4 = sample(6,4,T)
dice4 == 6
6 %in% dice4
e = sapply(1:100, function(x) 6 %in% sample(6,4,T))
mean(e)

dice24.1 = sample(6,24,T); dice24.2 = sample(6,24,T)
(dice24.1 + dice24.2) == 12
12 %in% (dice24.1 + dice24.2)
f = sapply(1:100, function(x) 12 %in% (sample(6,24,T) +
sample(6,24,T)))
mean(f)
```

Bayes' Theorem example

Joint distribution

	A: patient has disease	$\overline{A}$: patient does not have disease	
B: patient tests positive	0.00099	0.01998	$P(B) = 0.02097$
$\overline{B}$: patient does not test positive	0.00001	0.97902	$P(\overline{B}) = 0.97903$
	$P(A) = 0.001$	$P(\overline{A}) = 0.999$	1

A finance example: Let's say we want to know how a change in interest rates would affect the value of a stock market index.
Historical data available:

	A: Interest declines	Interest increases	
B: Stock decline	200	950	1150
$\overline{B}$: Stock increase	800	50	850
	1000	1000	2000

- P(SI) = the probability of the **S**tock index **I**ncreasing, etc
- Thus with our example plugging in our number we get:

$$P(SD|II) = \frac{P(SD)\cdot P(II|SD)}{P(II)} = \frac{(\frac{1150}{2000})(\frac{950}{1150})}{\frac{1000}{2000}} = 0.9499$$

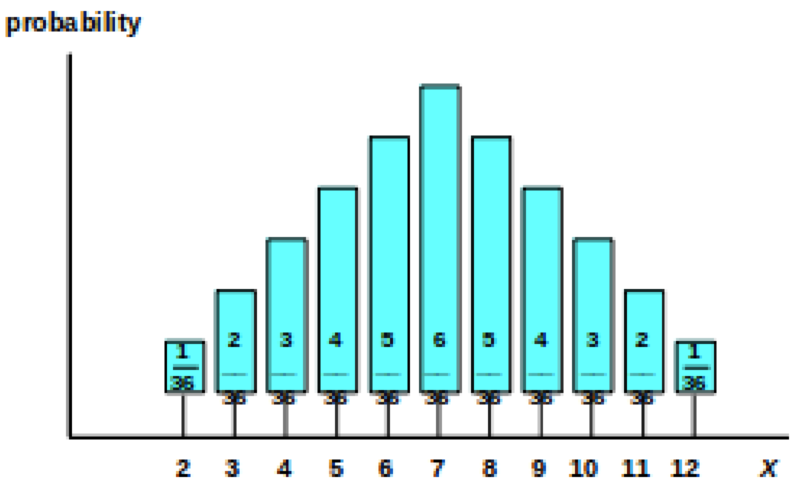
- Table: out of 2000 observations, 1150 instances showed the stock index decreased. the prior probability based on historical data, 57.5% (1150/2000). This doesn't take into account any information about interest rates, we wish to update. After updating this prior probability with information that interest rates have risen leads us to update the probability of the stock market decreasing from 57.5% to 95%. 95% is the posterior probability.

Quiz

Discrete random variable, probability and cumulative distribution functions

- Experiment: Throw a pair of fair dice
- X is a random variable defined as the *sum* of two die faces
- What does the probability distribution look like?
- What does the CDF look like?

The distribution is symmetrical, highest for $X = 7$, and declining on either side.



What does the CDF look like?

$\rightarrow$ sigmoid.

Linear transformed Random variable

Quiz:

$$X_{new} = aX + b$$

- Mean of X_{new} = ?
- Median of X_{new} = ?
- Variance of X_{new}= ?

CAMBRIDGE Judge Business School

Derive.

$$\begin{aligned}
\mu_{X_{new}} = E[aX+b] &= \sum_x (ax+b)P(X=x) && (1)\\
&= \sum_x axP(X=x) + bP(X=x) && (2)\\
&= \sum_x axP(X=x) + \sum_x bP(X=x) && (3)\\
&= a\sum_x xP(X=x) + b\sum_x P(X=x) && (4)\\
&= aE[x] + b = a\mu_X + b && (5)
\end{aligned}$$

$$Med[X_{new}] = aMed[X] + b \quad (6)$$

$$\begin{aligned}
\sigma^2_{X_{new}} = V[aX+b] &= E[((aX+b) - (aE[x]+b))^2] && (7)\\
&= E[((aX+b) - aE[x] - b)^2] && (8)\\
&= E[a^2(X - E[x])^2] && (9)\\
&= a^2E[(X-E[x])^2] = a^2V[X] = a^2\sigma^2_X && (10)
\end{aligned}$$

Adding independent random variables

Quiz:

- Mean and Variance of a sum of independent random variables
- Many useful *statistics* are linear combinations of data, i.e., of random variables
- If X and Y are *independent*:
  - $E(X+Y) = ?$
  - $Var(X+Y) = ?$

CAMBRIDGE Judge Business School

- X and Y independent – generated by independent mechanisms.
- Say the die example: Quiz:
- triangular between 2 and 12
- Can be generalised to any number of random variables
- derive

$$E[X+Y] = E[X] + E[Y] = \mu_X + \mu_Y \tag{11}$$

$$\begin{aligned}
V[X+Y] &= E[((X+Y) - (\mu_X - \mu_Y))^2] \\
&= E[((X - \mu_X) + (Y - \mu_Y))^2] \\
&= E[(X - \mu_X)^2] + E[(Y - \mu_Y))^2] + 2E[(X - \mu_X)(Y - \mu_Y)] \\
&= V(X) + V(Y) + \underbrace{2Cov(X,Y)}_{= 0 \text{ if } X,\, Y \text{ independent}}
\end{aligned}$$

Moments of a Random variable

Quiz:

- Experiment: Throw a pair of fair dice, independently.
- Outcome = X = sum of the faces
- Mean?
- Variance?
- Skewness?

CAMBRIDGE Judge Business School

- Mean: $E[X+Y] = \mu_{X+Y} = \mu_X + \mu_Y = 3.5 + 3.5 = 7$ (= mode = median)
- Variance: $V(X+Y) \overset{ind.}{=} V(X) + V(Y) = 2 \cdot V(X)$
  $V(X) = (1-3.5)^2 \cdot \frac{1}{6} + ...(6-3.5)^2 \cdot \frac{1}{6}$

  2 * sum((1:6 - 3.5)^2 * 1/6)
- or:

$$\sum_i (x_i - \bar{x})^2 \cdot p_i = (1-7)^2 \cdot 1/36 + ... + (12-7)^2 \cdot 1/36 = 5.8333$$

- Skewness = 0 (because of symmetry)

Covariance of functions of r.v.s

Covariance between linear functions of r.v.s:

- $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
- $Cov(X,Y)$:

$$\begin{aligned}
\sigma_{XY} &= E\left((X-\mu_X)(Y-\mu_Y)\right) \\
&= E\left(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y\right) \\
&= E(XY) - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y \\
&= E(XY) - \mu_X\mu_Y
\end{aligned}$$

- Likewise, can show: $Cov(a+bX+cV,Y) = b\sigma_{XY} + c\sigma_{VY}$

CAMBRIDGE Judge Business School

Derive.

$$\begin{aligned}
Cov(a+bX+cV,Y) &= E[(a+bX+cV-(a+b\mu_X+c\mu_V))(Y-\mu_Y)] \\
&= E[((a-a)+b(X-\mu_X)+c(V-\mu_V))(Y-\mu_Y)] \\
&= E[b(X-\mu_X)(Y-\mu_Y)+c(V-\mu_V)(Y-\mu_Y)] \\
&= bE[(X-\mu_X)(Y-\mu_Y)]+cE[(V-\mu_V)(Y-\mu_Y)] \\
&= b\sigma_{XY}+c\sigma_{VY}
\end{aligned}$$