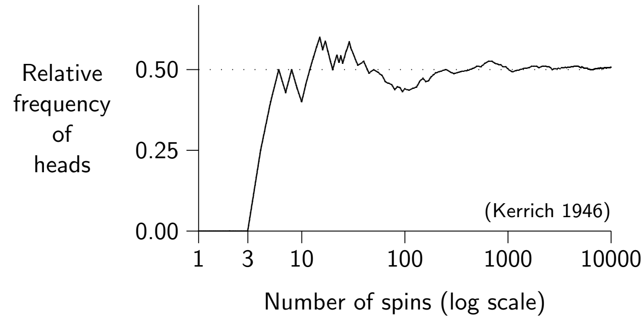
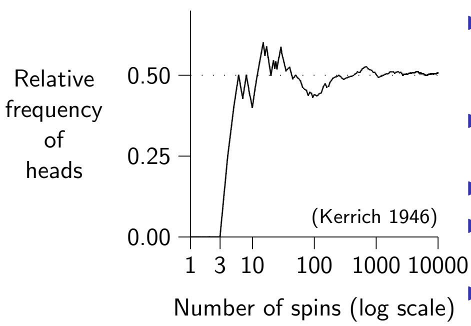
Probability

Alan R. Rogers

January 16, 2024



Probability and relative frequency in repeated trials



- rel. freq. of heads gradually approaches limiting value.
- Limiting value is the *probability* of heads
- Need not equal 1/2.
- We estimate probabilities from relative frequencies.
- We never know them exactly.

Kerrich's "urn" experiment

(○ ○ ● ●)

- Urn contains 4 balls: 2 black and 2 white
- Mix them up.
- Draw one at random
- Draw a second *without* replacing first.
- Repeat 5000 times.

Results from Kerrich's urn experiment

First ball	Second ball: Black	Second ball: White	sum
Black	756	1689	2445
White	1688	867	2555
sum	2444	2556	5000

- If 1st ball is B, 2nd is likely to be W
- And vice versa

Model of Kerrich's urn experiment

Assumption: we are equally likely to draw any ball in urn.

1st Ball

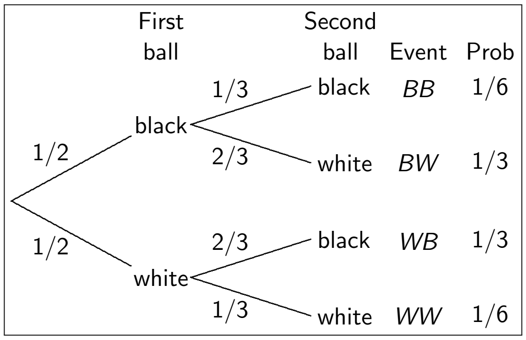
(○ ○ ● ●)

We are equally likely to draw black or white

2nd Ball

First ball	Remaining balls	Prob. of black
●	(○ ○ ●)	1/3
○	(○ ● ●)	2/3

2nd ball usually black if 1st was white, and vice versa.



Tree diagram for urn model

Kerrich's urn experiment: model versus data

Event	Theoretical probability	Observed relative frequency
BB	0.167	0.151
BW	0.333	0.338
WB	0.333	0.338
WW	0.167	0.173

Theory and observation are not identical, but they are close.

Why do we multiply along branches?

Conditional probability

- What is the conditional probability that the 2nd ball is white given that the first was black?
- 2/3.
- Called a *conditional probability* and written

$$\Pr[\text{2nd ball white}|\text{1st one black}].$$

- "|" is pronounced "given."

Conditional relative frequencies

First ball	Second ball Black	White	sum
Black	**756**	$\boxed{\textbf{1689}}$	**2445**
White	1688	867	2555
sum	2444	2556	5000

- On trials where the 1st ball was black, how often was the 2nd white?
- A fraction 1689/2445 of the time, or ≈ 0.69.

This is a conditional relative frequency. If the number of trials is large, this approximates a conditional probability.

The results of 20,000 throws with two dice (Wolf 1850, cited in Bulmer 1967)

Black	White 1	2	3	4	5	6	$\sum$	f
1	547	587	500	462	621	690	3407	.170
2	609	655	497	535	651	**684**	**3631**	.182
3	514	540	468	438	587	629	3176	.159
4	462	507	414	413	509	611	2916	.146
5	551	562	499	506	658	672	3448	.172
6	563	598	519	487	609	646	3422	.171
$\sum$:	3246	3449	2897	2841	3635	3932	20000	1.000
f:	.162	.172	.145	.142	.182	.197	1.000	

- What is the conditional frequency of $W6$ given $B2$?
- $684/3631 \approx 0.188$

Product rule for relative frequencies

How often did Kerrich get $B1$ and $W2$?

First ball	Second ball: Black	Second ball: White	sum
Black	756	$\boxed{\mathbf{1689}}$	2445
White	1688	867	2555
sum	2444	2556	**5000**

A fraction 1689/5000 of the time.

$$\frac{1689}{5000} = \frac{1689}{2445} \times \frac{2445}{5000}$$

$$\overbrace{\frac{1689}{5000}}^{f(B1\,\&\,W2)} = \overbrace{\frac{1689}{2545}}^{f(W2|B1)} \times \overbrace{\frac{2445}{5000}}^{f(B1)}$$

As N increases, relative frequencies (f) become probabilities.

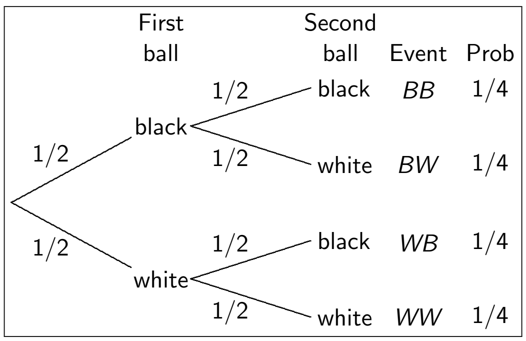
Product rule

The probability of A and B is

$$\Pr[A\,\&\,B] = \Pr[B|A]\Pr[A]$$

This is why we multiply along the branches of a tree diagram.

Statistical independence: sampling w/ replacement



$$\Pr[W_2|B_1] = \Pr[W_2|W_1] = \Pr[W_2] = 1/2$$

Sampling with replacement: model versus data

Event	Theoretical probability	Observed relative frequency
BB	0.25	0.254
BW	0.25	0.255
WB	0.25	0.252
WW	0.25	0.239

Data from computer simulation

Theory and observation are not identical, but they are very close.

Sum rule: Pr[black 4 or white 5 (or both)]

Black \ White	1	2	3	4	5	6	$\sum$
1	547	587	500	462	**621**	690	3407
2	609	655	497	535	**651**	684	3631
3	514	540	468	438	**587**	629	3176
4	**462**	**507**	**414**	**413**	$\boxed{\mathbf{509}}$	**611**	2916
5	551	562	499	506	**658**	672	3448
6	563	598	519	487	**609**	646	3422
$\sum$:	3246	3449	2897	2841	3635	3932	*20000*

Relative frequency is the sum of the bold-face values divided by 20,000.

$$f[b4 \text{ or } w5] = \overbrace{\frac{2916}{20000}}^{f[b4]} + \overbrace{\frac{3635}{20000}}^{f[w5]} - \overbrace{\frac{509}{20000}}^{f[b4\,\&\,w5]}$$

Sum rule for probabilities

$$\Pr[A \text{ or } B] = \Pr[A] + \Pr[B] - \Pr[A\,\&\,B]$$

Sum rule again: Pr[white 3 or white 5]

For mutually exclusive events, there is nothing to subtract.

	White						
Black	1	2	3	4	5	6	$\sum$
1	547	587	**500**	462	**621**	690	3407
2	609	655	**497**	535	**651**	684	3631
3	514	540	**468**	438	**587**	629	3176
4	462	507	**414**	413	**509**	611	2916
5	551	562	**499**	506	**658**	672	3448
6	563	598	**519**	487	**609**	646	3422
$\sum$:	3246	3449	2897	2841	3635	3932	*20000*

What is rel. frq. of white 3 or white 5?

$$f[w4 \text{ or } w5] = \overbrace{\frac{2897}{20000}}^{f[w4]} + \overbrace{\frac{3635}{20000}}^{f[w5]}$$

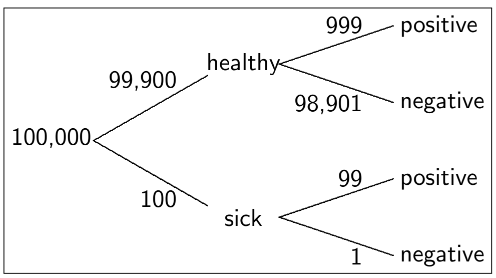
Bayes's rule

Problem: Our emphasis has been on the probability of an outcome given a hypothesis. But we often want to know the probability of the hypothesis, given the outcome.

Example: The probability the patient is sick given a positive result on some test.

Suppose that 0.1% of people have some disease. When tested for the disease 99% of sick people test positive, but so do 1% of well people. What fraction of those with positive results are really sick?

Bayes's rule in terms of counts



What fraction of those who test positive are really sick?

$$\frac{99}{99+999} \approx 0.09 \qquad \text{Fewer than 1 in 10!}$$

Bayes's rule in terms of probabilities

Recall the multiplication law:

$$\Pr[A\&B] = \Pr[B]\Pr[A|B] = \Pr[A]\Pr[B|A]$$

Divide through by $\Pr[B]$:

$$\Pr[A|B] = \frac{\Pr[A]\Pr[B|A]}{\Pr[B]} \quad \text{(Bayes's rule)}$$

Allows us to calculate $\Pr[A|B]$ from $\Pr[B|A]$.

Back to example

$$\Pr[A|B] = \frac{\Pr[A]\Pr[B|A]}{\Pr[B]} \quad \text{(Bayes's rule)}$$

A: patient is sick. $\Pr[A] = 1/1000$.

B: patient tested positive.
$\Pr[B] = (999+99)/100000 = 1098/100000$.

Pr[testing positive if sick] is $\Pr[B|A] = 99/100$.

Using Bayes's rule,

$$\Pr[A|B] = \frac{1/1000 \times 99/100}{1098/100000} = \frac{99}{1098} \approx 0.09$$

This is the same answer we got using counts.

Summary

Sum rule

$$\Pr[A \text{ or } B] = \Pr[A] + \Pr[B] - \Pr[A \,\&\, B]$$

Product rule

$$\Pr[A \,\&\, B] = \Pr[A]\Pr[B|A]$$

Bayes's rule

$$\Pr[A|B] = \frac{\Pr[A]\Pr[B|A]}{\Pr[B]}$$

Problems

1. You toss a fair coin twice. What is the probability that the number of heads is one?
2. You toss two fair dice, one red and one black. What is the probability that you observe either a red 4 or a black 6 (or both)?
3. Imagine a modified version of Kerrich's urn experiment in which each trial begins with 3 balls of each color (red and black). What is the probability that, in a single trial, both of the balls drawn are red?

Random Variables, Expectations, Variance, and Covariance

Alan R. Rogers

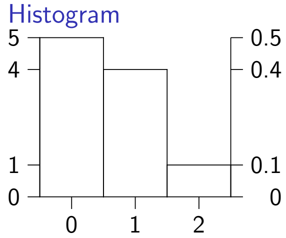
January 16, 2024

Distributions of counts and of relative frequencies

$$\text{data} = [0, 1, 0, 0, 1, 1, 1, 0, 0, 2]$$

Value	Count	Relative frequency
0	5	0.5
1	4	0.4
2	1	0.1
	10	1.0

Histogram



The mean (m)

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where N is sample size and x_i is ith data value.
Example: If $x = [3, 2, 2]$, then

$$m = \frac{1}{3} \times (3 + 2 + 2) = 7/3$$

Using relative frequencies to calculate the mean

$$m = \sum_x x f_x$$

where x is a sample value and f_x is the relative frequency of that value.
Example: If $x = [3, 2, 2]$, then

$$\begin{aligned} f_2 &= 2/3 \\ f_3 &= 1/3 \\ m &= (2 \times 2/3) + (3 \times 1/3) = 7/3 \end{aligned}$$

Larger example

$$\text{data} = [0, 1, 0, 0, 1, 1, 1, 0, 0, 2]$$

Value	Relative frequency
0	0.5
1	0.4
2	0.1

What is the mean?

$$\begin{aligned} m &= 0 \times 0.5 + 1 \times 0.4 + 2 \times 0.1 \\ &= 0.4 + 0.2 = 0.6 \end{aligned}$$

Measures of variation

- range of data
- interquartile range: range of middle half of data
- variance: average of $(x - m)^2$, where m is the mean
- square root of variance: the standard deviation

In population genetics, the variance is most useful.

Calculating the variance (v)

$$v = \sum_x (x - m)^2 f_x$$

where m is the mean, x is a sample value, and f_x is the relative frequency of that value.

What are the mean and variance of this data set: $[3, 2, 2]$?

Calculations

Frequency distribution:

Value	Relative frequency
2	2/3
3	1/3

Mean:

$$\begin{aligned} m &= 2 \times \frac{2}{3} + 3 \times \frac{1}{3} \\ &= 7/3 \approx 2.33 \end{aligned}$$

Variance:

$$\begin{aligned} V &= (2 - 2.33)^2 \times \frac{2}{3} \\ &\quad + (3 - 2.33)^2 \times \frac{1}{3} \\ &= 0.22 \end{aligned}$$

These ideas work not only for relative frequencies but also for probabilities.

- Frequency distributions become probability distributions.
- Means become expected values.
- Nothing else changes.

Probability distribution

- Assigns a probability to every event.
- When events have numeric values, the probability distribution translates one number (the event) into another (the probability).
- A set of events with associated probabilities is a *random variable* (r.v.).
- Distributions of numerical r.v.s are often described using mathematical functions.

A **random variable** is a variable whose values occur with particular probabilities.
(We would need to modify this slightly for variables that vary along a continuum, such as height or weight. But I'm going to ignore that distinction here.)

Example 1: a fair coin

Suppose that X (a random variable) is the number of heads in one toss of a fair coin. The *probability distribution* of X is

X	Probability (p_X)
0	1/2
1	1/2

Probabilities

- lie between 0 and 1,
- sum to 1.

Problem

In the previous slide, X was the number (either 0 or 1) of heads in one toss of a fair coin. What is the probability distribution of $Y = X^2$?

Example 2: a loaded die

Let X be the number obtained on a roll of the die. This die is "loaded," so that 1s and 2s are twice as probable as other values.

X	(p_X)
1	0.250
2	0.250
3	0.125
4	0.125
5	0.125
6	0.125
	1.0000

The mean (or expectation) of a random variable

The mean of X is written $E(X)$ and equals

$$E(X) = \sum_i p_i x_i$$

where x_i is the ith value that X can take, and p_i is its probability. If X is the number obtained on a roll of our loaded die, then

$$\begin{aligned} E[X] &= 1 \times 0.25 + 2 \times 0.25 + 3 \times 0.125 \\ &\quad + 4 \times 0.125 + 5 \times 0.125 + 6 \times 0.125 \\ &= 3 \end{aligned}$$

The same as an average, except that p_i is a probability rather than a relative frequency.

Allele frequency as expectation

G'type	G'type freq	Cond. allele freq
A_1A_1	P_{11}	1
A_1A_2	P_{12}	0.5
A_2A_2	P_{22}	0

Allele frequency

$$\begin{aligned} p_1 &= 1 \times P_{11} \\ &\quad + 0.5 \times P_{12} \\ &\quad + 0 \times P_{22} \end{aligned}$$

The variance

If μ is the mean of X, then its variance is

$$V[X] = E[(X - \mu)^2] \tag{1}$$

For our loaded die, the mean was $\mu = 3$. The variance is

$$\begin{aligned} V[X] &= (1-3)^2 \times 0.25 \\ &\quad + (2-3)^2 \times 0.25 \\ &\quad + (3-3)^2 \times 0.125 \\ &\quad + (4-3)^2 \times 0.125 \\ &\quad + (5-3)^2 \times 0.125 \\ &\quad + (6-3)^2 \times 0.125 \\ &= 3 \end{aligned}$$

A single toss of an unfair coin

The probability of "heads" is an unknown value p.
Your winnings: $X = 1$ for heads and $X = 0$ for tails. What's the probability distribution of X? The mean? The variance?

Properties of expectations

If X and Y are random variables and a is a constant,

$$E[a] = a \quad (2)$$
$$E[aX] = aE[X] \quad (3)$$
$$E[X+Y] = E[X] + E[Y] \quad (4)$$

See JEPr for details.

Using rules of expectations to re-express the variance

Let $\mu = E[X]$. The variance of X is

$$\begin{aligned}
V &= E[(X-\mu)^2] && \text{(by Eqn. 1)}\\
&= E[X^2 - 2\mu X + \mu^2] \\
&= E[X^2] - E[2\mu X] + E[\mu^2] && \text{(by Eqn. 4)}\\
&= E[X^2] - E[2\mu X] + \mu^2 && \text{(by Eqn. 2)}\\
&= E[X^2] - 2\mu E[X] + \mu^2 && \text{(by Eqn. 3)}\\
&= E[X^2] - 2\mu^2 + \mu^2 && \text{(by definition of } \mu)\\
&= E[X^2] - \mu^2 && (5)
\end{aligned}$$

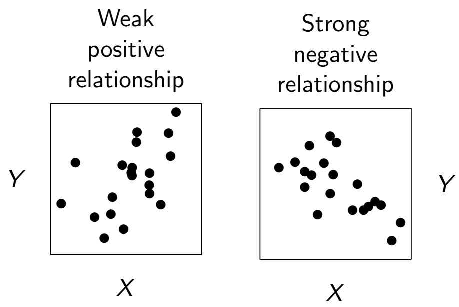
Variance of our loaded die

A moment ago, we found for our loaded die that $E[X] = V[X] = 3$. Let us recalculate this using Eqn. 5. We need

$$\begin{aligned}
E[X^2] &= 1^2 \times 0.25 + 2^2 \times 0.25 \\
&\quad + 3^2 \times 0.125 + 4^2 \times 0.125 \\
&\quad + 5^2 \times 0.125 + 6^2 \times 0.125 \\
&= 12
\end{aligned}$$

$$V = E[X^2] - \mu^2 = 12 - 3^2 = 3$$

Association between variables



Positive and negative relationships between variables.

Covariance: a measure of association

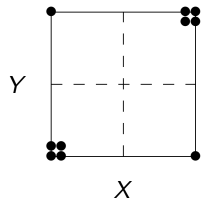
$$\begin{aligned}
C(X,Y) &= \sum_{x,y} (x - E[X])(y - E[Y]) P_{x,y} \\
&= E\Big[(X - E[X])(Y - E[Y])\Big] \\
&= E[XY] - E[X]E[Y]
\end{aligned}$$

When X and Y are independent, $C(X,Y) = 0$.

A bivariate probability distribution

X	Y	$P_{X,Y}$	$(X - E[X])(Y - E[Y])$
0	0	0.4	+0.25
0	1	0.1	–0.25
1	0	0.1	–0.25
1	1	0.4	+0.25

Note: the $P_{X,Y}$ column lists the probabilities of the (X, Y) pairs. In column 4, $E[X] = E[Y] = 0.5$.



Numerical value of covariance in previous slide

$$\begin{aligned} C(X,Y) &= 0.4 \times 0.25 \\ &\quad - 0.1 \times 0.25 \\ &\quad - 0.1 \times 0.25 \\ &\quad + 0.4 \times 0.25 \\ &= 0.15 \end{aligned}$$

Problem

In Kerrich's urn experiment, suppose you get $1 for each red ball and $0 for each green one, and let X and Y represent the dollars you receive on the two draws within a single trial of the experiment. Write down the probability distribution of X and Y in tabular form. Your table should have columns for X, for Y, and for the joint probability of X and Y, i.e. $\Pr[X, Y]$.

This is exactly like Fig. 2 of JEPr, which presents the following probability distribution:

Event	Prob
RR	1/6
RG	1/3
GR	1/3
GG	1/6

where "R" and "G" stand for "red" and "green". Now, "R" becomes "1," "G" becomes "0," and the probability distribution becomes

X	Y	$\Pr(X, Y)$
1	1	1/6
1	0	1/3
0	1	1/3
0	0	1/6

Probability Distributions

Alan R. Rogers

January 16, 2024

Probability distributions

A probability distribution is a function.

Input event

Output probability of event

- So far we have described probability distributions using tables.
- When events are numbers, distributions can be expressed as mathematical functions.

The Urn Metaphor

Imagine two urns: metaphors for a population in two successive generations. Urn 1 has 50 balls, some red, some white, representing parental gene copies. Urn 2 is empty until urn 1 has "reproduced" as follows:

1. Examine a random ball from urn 1.
2. Put a ball of the same color into urn 2.
3. Replace the ball from urn 1.
4. Repeat until there are 50 balls in urn 2.

The number of red balls in urn 2 is likely to differ from that in urn 1, because of random sampling. This metaphor is used as a model of genetic drift.

Binomial random variable

In probability theory, the number of red balls in urn 2 is a *binomial random variable*.

1. Balls drawn from the urn are statistically independent.
2. Each ball is red with probability p, the fraction of red balls in urn 1.

This distribution has two parameters: N, the number of balls put into urn 2, and p, the probability of "red" each time a ball is drawn.

Probability of HT

Consider tosses of an unfair coin, for which the probability of "heads" is p and that of "tails" is $q = 1 - p$. Assume that the tosses are statistically independent.

Experiment	Toss a coin 2 times.
Result	HT
Probability	pq

This is an event of form $\Pr[A\&B]$, where A is the event that the first toss is H and B is the event that the 2nd is T. By assumption, $\Pr[A] = p$ and $\Pr[B] = q$. The tosses are statistically independent, so

$$\Pr[A\&B] = pq$$

by the multiplication law of probability.

Probability of HHT

Experiment	Toss a coin 3 times.
Result	HHT
Probability	p^2q

Probability of 2 heads in 3 tosses

There are 3 ways to get 2 heads in 3 tosses:

Event	Probability
THH	p^2q
HTH	p^2q
HHT	p^2q

The probability of 2 heads in 3 tosses is

$$\begin{aligned} P_2 &= 3p^2q \\ &= \binom{3}{2} p^2 q \end{aligned}$$

where $\binom{3}{2}$ is pronounced "3 choose 2" and means the number of ways to choose 2 items out of a collection of 3.

Binomial distribution

The probability of x heads in K tosses is

$$P_X = \binom{K}{x} p^x q^{K-x}$$

$$E[X] = pK \qquad \text{mean}$$
$$V[X] = Kpq \qquad \text{variance}$$

Poisson distribution

Consider the lineage that connects me to an ancestor who lived t generations ago. The expected number of mutations along that lineage is $\lambda = ut$, where u is the mutation rate per generation. The number of mutations is a random variable (r.v.). If the mutation rate is constant, then the distribution of this r.v. is *Poisson*.

Poisson distribution function

If X is a Poisson-distributed r.v. with mean λ, then X takes value x with probability

$$P_x = \frac{\lambda^x e^{-\lambda}}{x!}$$

where e is the base of natural logarithms and $x!$ is "x factorial," or $x \cdot (x-1) \cdot (x-2) \cdots 1$.

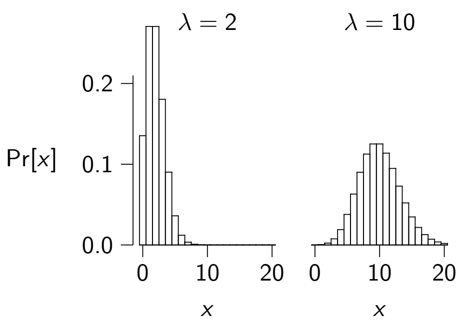
Mean equals variance.

$$E[X] = V[X] = \lambda$$

What is P_0? (Hint: $0! = 1$ and $\lambda^0 = 1$.)

$$P_0 = e^{-\lambda}$$

Poisson distribution



Poisson distribution functions

Mutation rates at autosomal nucleotide sites are roughly 10^{-9} per year. Consider a nucleotide in you. If you could trace its ancestry back across the last 10^9 years, what is the probability that you would find no mutations?

The expected number of mutations is $\lambda = ut$, where $u = 10^{-9}$ and $t = 10^9$. Thus, $\lambda = 1$. The probability of no mutations is

$$e^{-1} \approx 0.37$$

Raisin data

```
Date: Aug 28, 2009
N=41, Mean=21.756098, Var=26.239024, Max=33
 1- 3: *
 4- 6: *
 7- 9:  *
10-12: -*
13-15: --*
16-18: -----*-
19-21: ----------  *
22-24: -------*--
25-27: ------- *
28-30: --- *
31-33: *

Key: ---- Poisson distribution w/ mean 21.756098
        * Data
```

Raisin data

```
Date: Sept 6, 2013
N=32, Mean=20.375000, Var=15.080645, Max=34
 1- 3: *
 4- 6: *
 7- 9: *
10-12: *
13-15: --*
16-18: -------  *
19-21: --------*
22-24: -------   *
25-27: --*-
28-30: *
31-33: *

Key: ---- Poisson distribution w/ mean 20.375000
        * Data
```

Raisin data

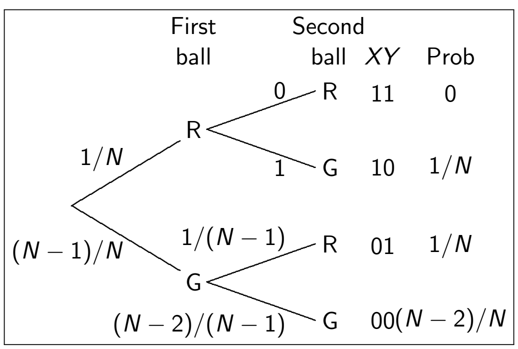
```
Date: Sept 6, 2017
N=36, Mean=14.111111, Var=13.473016, Max=21
 1- 3: *
 4- 6: *
 7- 9: ---  *
10-12: --------*
13-15: --------*--
16-18: --------   *
19-21: --- *

Key: ---- Poisson distribution w/ mean 14.111111
        * Data
```

Homework problem 1.41

Imagine an urn with N balls, of which 1 is red and the rest are black. You draw 2 balls from the urn at random *without* replacement. Let $X = 1$ if the first ball is red and $X = 0$ otherwise. Define Y similarly for the second ball.

What is the covariance of X and Y?

Probability tree & distribution of (X, Y)



Goal: calculate $C(X, Y) = E[XY] - E[X]E[Y]$

$E[X]$

X	Y	Pr
1	1	0
1	0	$1/N$
0	1	$1/N$
0	0	$(N-2)/N$

$$\begin{aligned} E[X] &= 1 \times 0 \\ &\quad + 1 \times 1/N \\ &\quad + 0 \times 1/N \\ &\quad + 0 \times (N-2)/N \\ &= 1/N \end{aligned}$$

$E[Y]$

X	Y	Pr
1	1	0
1	0	$1/N$
0	1	$1/N$
0	0	$(N-2)/N$

$$\begin{aligned} E[Y] &= 1 \times 0 \\ &\quad + 0 \times 1/N \\ &\quad + 1 \times 1/N \\ &\quad + 0 \times (N-2)/N \\ &= 1/N \end{aligned}$$

$E[XY]$

X	Y	XY	Pr
1	1	1	0
1	0	0	$1/N$
0	1	0	$1/N$
0	0	0	$(N-2)/N$

$$\begin{aligned} E[XY] &= 1 \times 0 \\ &\quad + 0 \times (\text{the rest}) \\ &= 0 \end{aligned}$$

Covariance of X and Y:

$$\overbrace{E[XY]}^{0} - \overbrace{E[X]}^{1/N}\overbrace{E[Y]}^{1/N} = -1/N^2$$