My Data Science Notes

Michael Foley

2020-02-20

Contents

Intro

These notes are pulled from various classes, tutorials, books, etc. and are intended for my own consumption. If you are finding this on the internet, I hope it is useful to you, but you should know that I am just a student and there's a good chance whatever you're reading here is mistaken. In fact, that should probably be your null hypothesis… or your prior. Whatever.

Chapter 1

Probability

1.1 Principles

Here are three rules that come up all the time.

- $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(AB)$. This rule generalizes to $Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C) - Pr(AB) - Pr(AC) - Pr(BC) + Pr(ABC)$.
- $Pr(A|B) = \frac{P(AB)}{P(B)}$
- If A and B are independent, $Pr(A \cap B) = Pr(A)Pr(B)$, and $Pr(A|B) = Pr(A)$.

Uniform distributions on finite sample spaces often reduce to counting the elements of A and the sample space S, a process called combinatorics. Here are three important combinatorial rules.

Multiplication Rule. $|S| = |S_1| \cdots |S_k|$.

How many outcomes are possible from a sequence of 4 coin flips and 2 rolls of a die? $|S| = |S_1| \cdot |S_2| \ldots |S_6| = 2 \cdot 2 \cdot 2 \cdot 2 \cdot 6 \cdot 6 = 288$.

How many subsets are possible from a set of n=10 elements? In each subset, each element is either included or not, so there are $2^n = 1024$ subsets.

How many subsets are possible from a set of n=10 elements taken k at a time with replacement? Each experiment has n possible outcomes and is repeated k times, so there are n^k subsets.

Permutations. The number of *ordered* arrangements (permutations) of a set of $|S| = n$ items taken k at a time *without* replacement has $n(n-1) \ldots (n-k+1)$

subsets because each draw is one of k experiments with decreasing number of possible outcomes.

$$ {}_nP_k = \frac{n!}{(n-k)!} $$

Notice that if $k = 0$ then there is 1 permutation; if $k = 1$ then there are n permutations; if $k = n$ then there are $n!$ permutations.

How many ways can you distribute 4 jackets among 4 people? ${}_nP_k = \frac{4!}{(4-4)!} = 4! = 24$

How many ways can you distribute 4 jackets among 2 people? ${}_nP_k = \frac{4!}{(4-2)!} = 12$

Subsets. The number of *unordered* arrangements (combinations) of a set of $|S| = n$ items taken k at a time *without* replacement has

$$ {}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!} $$

combinations and is called the binomial coefficient. The binomial coefficient is the number of different subsets. Notice that if k=0 then there is 1 subset; if k=1 then there are n subsets; if k=n then there is 1 subset. The connection with the permutation rule is that there are $n!/(n-k)!$ permutations and each permutation has $k!$ permutations.

How many subsets of 7 people can be taken from a set of 12 persons? ${}_{12}C_7 = \binom{12}{7} = \frac{12!}{7!(12-7)!} = 792$

If you are dealt five cards, what is the probability of getting a "full-house" hand containing three kings and two aces (KKKAA)?

$$ P(F) = \frac{\binom{4}{3}\binom{4}{2}}{\binom{52}{5}} $$

Distinguishable permutations. The number of *unordered* arrangements (distinguishable permutations) of a set of $|S| = n$ items in which n_1 are of one type, n_2 are of another type, etc., is

$$ \binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \ldots n_k!} $$

How many ordered arrangements are there of the letters in the word PHILIPPINES? There are n=11 objects. $|P| = n_1 = 3$; $|H| = n_2 = 1$; $|I| = n_3 = 3$; $|L| = n_4 = 1$; $|N| = n_5 = 1$; $|E| = n_6 = 1$; $|S| = n_7 = 1$.

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{11!}{3!1!3!1!1!1!1!} = 1,108,800$$

How many ways can a research pool of 15 subjects be divided into three equally sized test groups?

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{15!}{5!5!5!} = 756,756$$

1.2 Discrete Distributions

These notes rely heavily on PSU STATS 504 course notes.

The most important discrete distributions are Bernoulli, Binomial, Poisson, and Multinomial. Less important, but sometimes useful, are the negative binomial, geometric, and hypergeometric distributions.

A discrete random variable X is described by its probability mass function $f(x) = P(X = x)$. The set of x values for which $f(x) > 0$ is called the *support*. If the distribution depends on unknown parameter(s) θ we write it as $f(x; \theta)$ (frequentists) or $f(x|\theta)$ (Bayesian).

1.2.1 Bernoulli

If X is the result of a trial with two outcomes of probability $P(X = 1) = \pi$ and $P(X = 0) = 1 - \pi$, then X is a random variable with a Bernoulli distribution

$$f(x) = \pi^x (1 - \pi)^{1-x}, \qquad x \in (0, 1)$$

with $E(X) = \pi$ and $V(X) = \pi(1 - \pi)$.

1.2.2 Binomial

If X is the count of successful events in n identical and independent Bernoulli trials of success probability π, then X is a random variable with a binomial distribution $X \sim Bin(n, \pi)$

$$f(x; \pi) = \frac{n!}{x!(n-x)!} \pi^x (1 - \pi)^{n-x} \qquad x \in (0, 1, ..., n), \;\; \pi \in [0, 1]$$

with $E(X) = n\pi$ and $V(X) = n\pi(1 - \pi)$.

The Bernoulli distribution is a special case of the binomial with $n = 1$. As n increases for fixed π, the binomial distribution approaches normal distribution $N(n\pi, n\pi(1-\pi))$. The binomial distribution assumes independent trials - if sampling *without replacement from a finite population*, then the hypergeometric distribution is appropriate.

Example

What is the probability 2 out of 10 coin flips are heads if the probability of heads is 0.3?

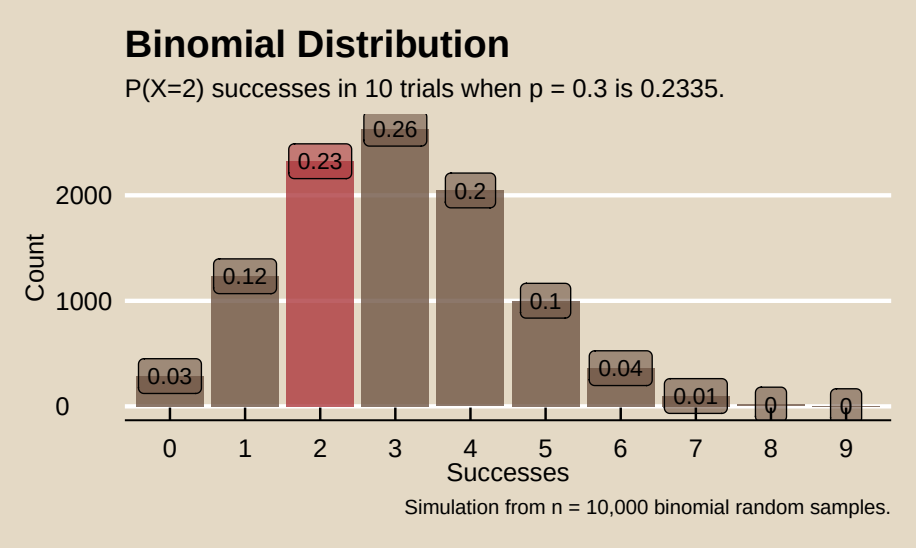
Function `dbinom()` calculates the binomial probability.

```
dbinom(x = 2, size = 10, prob = 0.3)
```

```
## [1] 0.23
```

A simulation of n = 10,000 random samples of size 10 gives a similar result. `rbinom()` generates a random sample of numbers from the binomial distribution.

```
data.frame(cnt = rbinom(n = 10000, size = 10, prob = 0.3)) %>%
  count(cnt) %>%
  ungroup() %>%
  mutate(pct = n / sum(n),
         X_eq_x = cnt == 2) %>%
  ggplot(aes(x = as.factor(cnt), y = n, fill = X_eq_x, label = pct)) +
  geom_col(alpha = 0.8) +
  scale_fill_mf() +
  geom_label(aes(label = round(pct, 2)), size = 3, alpha = .6) +
  theme_mf() +
  theme(legend.position = "none") +
  labs(title = "Binomial Distribution",
       subtitle = paste0(
         "P(X=2) successes in 10 trials when p = 0.3 is ",
         round(dbinom(2, 10, 0.3), 4), "."
       ),
       x = "Successes",
       y = "Count",
       caption = "Simulation from n = 10,000 binomial random samples.")
```



Example

What is the probability of <=2 heads in 10 coin flips where probability of heads is 0.3?

The cumulative probability is the sum of the first three bars in the simulation above. Function `pbinom()` calculates the *cumulative* binomial probability.

```
pbinom(q = 2, size = 10, prob = 0.3, lower.tail = TRUE)
```

```
## [1] 0.38
```

Example

What is the expected number of heads in 25 coin flips if the probability of heads is 0.3?

The expected value, $\mu = np$, is 7.5. Here's an empirical test from 10,000 samples.

```
mean(rbinom(n = 10000, size = 25, prob = .3))
```

```
## [1] 7.5
```

The variance, $\sigma^2 = np(1-p)$, is 5.25. Here's an empirical test.

```
var(rbinom(n = 10000, size = 25, prob = .3))
```

```
## [1] 5.2
```

Example

Suppose X and Y are independent random variables distributed $X \sim Bin(10, .6)$ and $Y \sim Bin(10, .7)$. What is the probability that either variable is <=4?

Let $P(A) = P(X <= 4)$ and $P(B) = P(Y <= 4)$. Then $P(A|B) = P(A) + P(B) - P(AB)$, and because the events are independent, $P(AB) = P(A)P(B)$.

```
p_a <- pbinom(q = 4, size = 10, prob = 0.6, lower.tail = TRUE)
p_b <- pbinom(q = 4, size = 10, prob = 0.7, lower.tail = TRUE)
p_a + p_b - (p_a * p_b)
```

```
## [1] 0.21
```

Here's an empirical test.

```
df <- data.frame(
  x = rbinom(10000, 10, 0.6),
  y = rbinom(10000, 10, 0.7)
  )
mean(if_else(df$x <= 4 | df$y <= 4, 1, 0))
```

```
## [1] 0.21
```

1.2.3 Poission

If X is the number of successes in n (many) trials when the probability of success λ/n is small, then X is a random variable with a Poisson distribution $X \sim Poisson(\lambda)$

$$f(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x \in (0, 1, ...), \quad \lambda > 0$$

with $E(X) = \lambda$ and $V(X) = \lambda$.

$Poison(\lambda) \to Bin(n, \pi)$ when $n\pi = \lambda$ and $n \to \infty$ and $\pi \to 0$. Because the Poisson is limit of the $Bin(n, \pi)$, it is useful as an approximation to the binomial when n is large ($n >= 20$) and π small ($p <= 0.05$).

When the observed variance is greater than λ (overdispersion), the Negative Binomial distribution can be used instead of Poisson.

Example

What is the probability of making 2 to 4 sales in a week if the average sales rate is 3 per week?

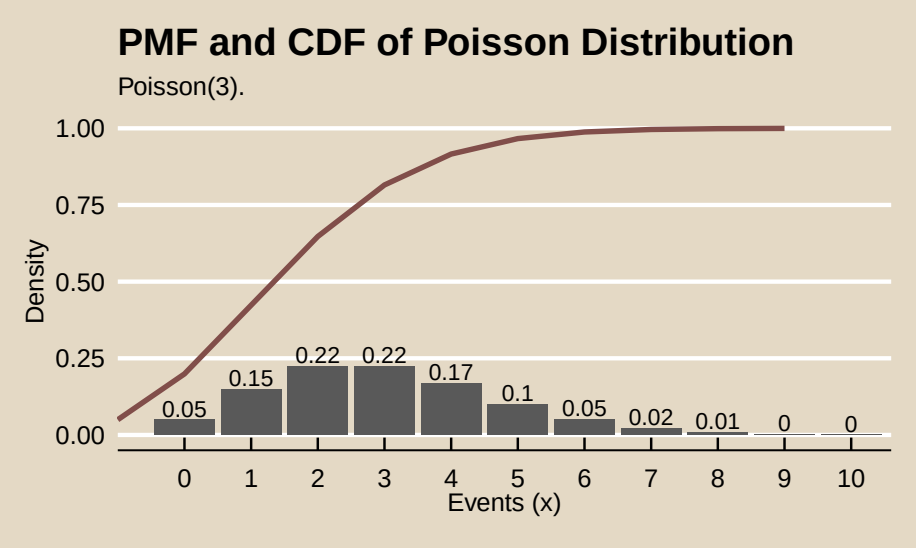
Function `dpois()` calculates the binomial probability.

```
# Using cumulative probability
ppois(q = 4, lambda = 3, lower.tail = TRUE) -
  ppois(q = 1, lambda = 3, lower.tail = TRUE)
```

```
## [1] 0.62
```

```
# Using exact probability
dpois(x = 2, lambda = 3) +
  dpois(x = 3, lambda = 3) +
  dpois(x = 4, lambda = 4)
```

```
## [1] 0.64
```



Example

Suppose a baseball player has a p=.300 batting average. What is the probability of X<=150 hits in n=500 at bats? X=150? X>150?

```
ppois(q = 150, lambda = .300 * 500, lower.tail = TRUE)
```

```
## [1] 0.52
```

```
dpois(x = 150, lambda = .300 * 500)
```

```
## [1] 0.033
```

```
ppois(q = 150, lambda = .300 * 500, lower.tail = FALSE)
```

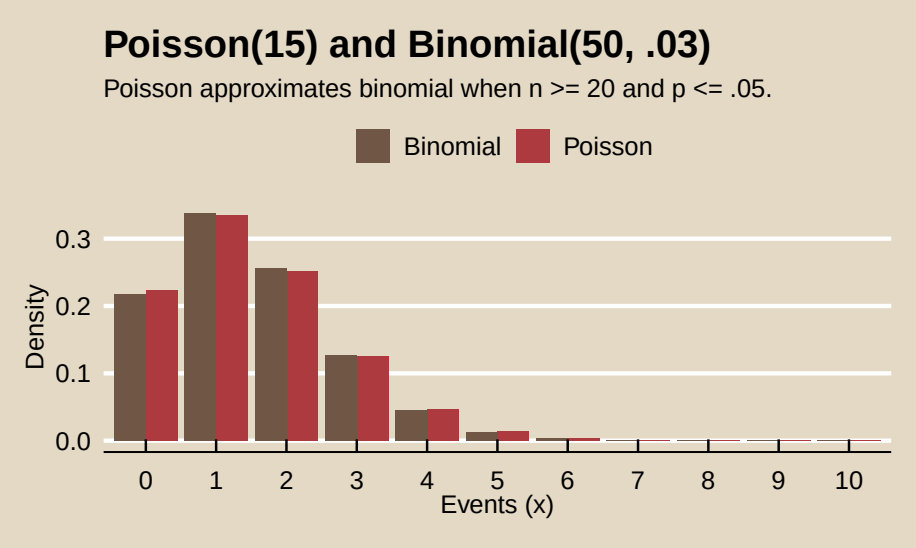
```
## [1] 0.48
```

The Poisson distribution approximates the binomial distribution with $\lambda = np$ if $n >= 20$ and $p <= 0.05$.

Example

What is the distribution of successes from a sample of n = 50 when the probability of success is p = .03?

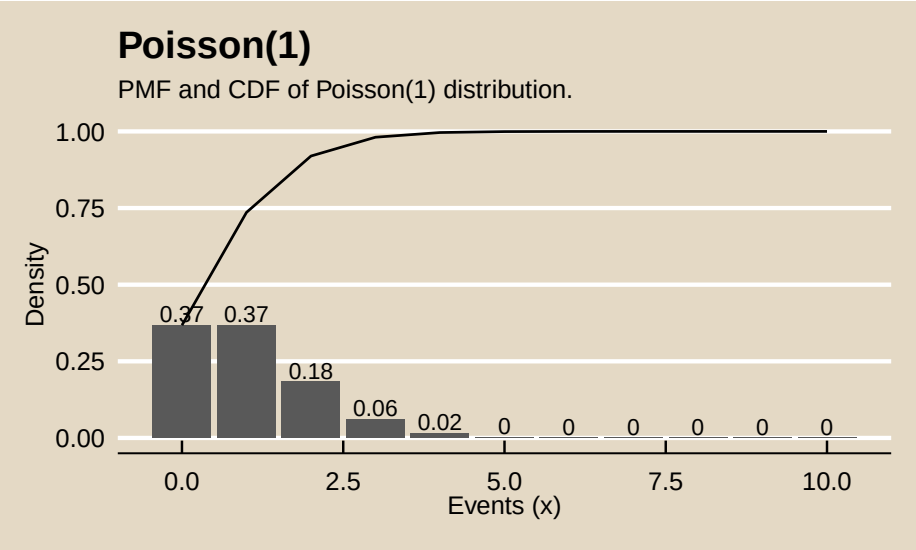
```
options(scipen = 999, digits = 2) # sig digits

n = 0:10
df <- data.frame(events = 0:10,
                      Poisson = dpois(x = n, lambda = .03 * 50),
                      Binomial = dbinom(x = n, size = 50, p = .03))
df_tidy <- gather(df, key = "Distribution", value = "density", -c(events))
ggplot(df_tidy, aes(x = factor(events), y = density, fill = Distribution)) +
  geom_col(position = "dodge") +
  theme_mf() +
  scale_fill_mf() +
  labs(title = "Poisson(15) and Binomial(50, .03)",
       subtitle = "Poisson approximates binomial when n >= 20 and p <= .05.",
       x = "Events (x)",
       y = "Density",
       fill = "")
```



Example

Suppose the probability that a drug produces a certain side effect is p = = 0.1% and n = 1,000 patients in a clinical trial receive the drug. What is the probability 0 people experience the side effect?

The expected value is np, 1. The probability of measuring 0 when the expected value is 1 is `dpois(x = 0, lambda = 1000 * .001)` = 0.37.



1.2.4 Negative-Binomial

If X is the count of trials required to reach a target number r of successful events in identical and independent Bernoulli trials of success probability p, then X is a random variable with a negative-binomial distribution $X \sim nb(r, p)$. The probability of $X = x$ trials prior to r successes is

$$f(x; r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}.$$

with $E(X) = r/p$ and $V(X) = r(1-p)/p^2$.

When the data has overdispersion, model the data with the negative-binomial distribution instead of Poission.

Example

An oil company has a $p = 0.20$ chance of striking oil when drilling a well. What is the probability the company drills $x = 7$ wells to strike oil $r = 3$ times?

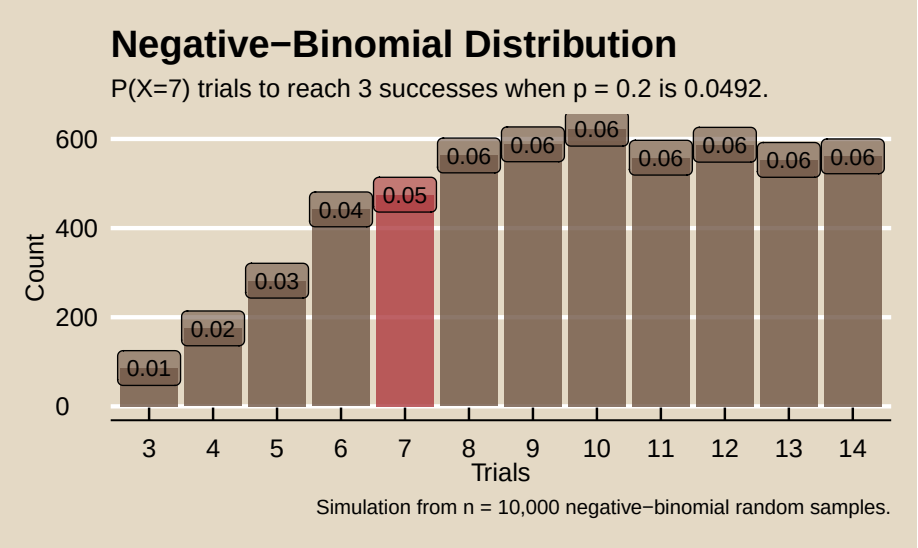
$$P(X = 7) = \binom{7-1}{3-1} (0.2)^3 (1-0.2)^{(7-3)} = 0.049.$$

Function `dnbinom()` calculates the negative-binomial probability. Parameter `x` equals the number of failures, $x - r$.

```
dnbinom(x = 4, size = 3, prob = 0.2)
```

```
## [1] 0.049
```

Here is a simulation of n = 10,000 random samples. `rnbinom()` generates a random sample of numbers from the negative-binomial distribution.



1.2.5 Geometric

If X is the count of independent Bernoulli trials of success probability p required to achieve the first successful trial, then X is a random variable with a geometric distribution $X \sim G(p)$ with mean $\mu = \frac{n}{p}$ and variance $\sigma^2 = \frac{(1-p)}{p^2}$. The probability of $X = n$ trials is

$$f(X = n) = p(1-p)^{n-1}.$$

The probability of $X <= n$ trials is

$$F(X = n) = 1 - (1-p)^n.$$

Example. A sports marketer randomly selects persons on the street until he encounters someone who attended a game last season. What is the probability the marketer encounters $x = 3$ people who did not attend a game before the first success if $p = 0.20$ of the population attended a game?

Function `pgeom()` calculates the geometric distribution probability.

```
dgeom(x = 3, prob = 0.20)
```

```
## [1] 0.1
```