Chapter 5 sections

Discrete univariate distributions:

- 5.2 Bernoulli and Binomial distributions
- Just skim 5.3 Hypergeometric distributions
- 5.4 Poisson distributions
- Just skim 5.5 Negative Binomial distributions

Continuous univariate distributions:

- 5.6 Normal distributions
- 5.7 Gamma distributions
- Just skim 5.8 Beta distributions

Multivariate distributions

- Just skim 5.9 Multinomial distributions
- 5.10 Bivariate normal distributions

Families of distributions

We will study a few useful families of distributions. That includes identifying some or all of the following

- pf /pdf and cdf - new notation: $f(x|$ parameters $)$
- Mean, variance and the m.g.f. $\psi(t)$
- Parameter space
- Special features and connections to other distributions, approximations
- Reasoning behind a distribution

Some distributions will have a natural justification for a certain kind of experiment

Other distributions are useful as a *model* for the uncertainty in an experiment

All models are wrong, but some are useful – George Box

Bernoulli distributions

Def: Bernoulli distributions – $\text{Bernoulli}(p)$

A r.v. X has the *Bernoulli distribution with parameter p* if $P(X = 1) = p$ and $P(X = 0) = 1 - p$. The pf of X is

$$f(x|p) = \begin{cases} p^x(1-p)^{1-x} & \text{for } x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

Parameter space: $p \in [0, 1]$

- In an experiment with only two possible outcomes, "success" and "failure", let X = number successes. Then the distribution of X is $\text{Bernoulli}(p)$ where p is the probability of success.
- $E(X) = p$, $\text{Var}(X) = p(1-p)$ and $\psi(t) = E(e^{tX}) = pe^t + (1-p)$
- The cdf is $F(x|p) = \begin{cases} 0 & \text{for } x < 0 \\ 1-p & \text{for } 0 \leq x < 1 \\ 1 & \text{for } x \geq 1 \end{cases}$

Binomial distributions

Def: Binomial distributions – $\text{Binomial}(n, p)$

A r.v. X has the *Binomial distribution with parameters n and p* if X has the pf

$$f(x|n,p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Parameter space: n is a positive integer and $p \in [0, 1]$

If X is the number of "successes" in n independent tries where prob. of success is p each time, then $X \sim \text{Binomial}(n, p)$

Theorem 5.2.1

If $X_1, X_2, \ldots, X_n$ form n *Bernoulli trials* with parameter p (i.e. are i.i.d. $\text{Bernoulli}(p)$) then $X = X_1 + \cdots + X_n \sim \text{Binomial}(n, p)$

Binomial distributions

Let $X \sim \text{Binomial}(n, p)$

- $E(X) = np$
- $\text{Var}(X) = np(1-p)$
- To find the m.g.f. of X write $X = X_1 + \cdots + X_n$ where X_i's are i.i.d. $\text{Bernoulli}(p)$. Then $\psi_i(t) = pe^t + 1 - p$ and we get

$$\psi(t) = \prod_{i=1}^{n} \psi_i(t) = \prod_{i=1}^{n} (pe^t + 1 - p) = (pe^t + 1 - p)^n$$

- cdf: $F(x|n,p) = \sum_{t=1}^{x} \binom{n}{t} p^t (1-p)^{n-t} =$ yikes!

Theorem 5.2.2

If $X_i \sim \text{Binomial}(n_i, p)$, $i = 1, \ldots, k$ and the X_i's are independent, then $X = X_1 + \cdots + X_k \sim \text{Binomial}(\sum_{i=1}^{k} n_i, p)$

Example: Blood testing (Example 5.2.7)

The setup:

- 1000 people need to be tested for a disease that affects 0.2% of all people.
- The test is guaranteed to detect the disease if it is present in a blood sample.
- Let $X_j = 1$ if person j has the disease and $X_j = 0$ if not. Then X_j has a Bernoulli distribution.
- *Assume* that X_j's are independent and that

$$P(X_j = 1) = p = 0.002 \quad \text{for all } j$$

- Then $X = \sum_{j=1}^{1000} X_j \sim \text{Binomial}(1000, 0.002)$ and the expected number of people that have the disease is $1000 \times 0.002 = 2$

Testing 1000 samples is expensive, so consider a more effective scenario.

Example: Blood testing (Example 5.2.7) – continued

- Divide the people into 10 groups of 100.
- For each group take a portion of each of the 100 blood samples and combine into one sample.
- Then test the combined blood samples (10 tests).
  - If all of these tests are negative then none of the 1000 people have the disease. Total number of tests needed: 10
  - If one of these tests are positive then we test each of the 100 people in that group. Total number of tests needed: 110
  - If two of these tests are positive then we test each of the 200 people in these groups. Total number of tests needed: 210
  - etc.
  - If all of the 10 tests are positive we end up having to do 1010 tests

What is the expected number of tests needed?

Example: Blood testing (Example 5.2.7) – continued

What is the expected number of tests needed?

- Let Z_i = number of people in group i that have the disease, $i = 1, \ldots, 10$.
  Then $Z_i \sim \text{Binomial}(100, 0.002)$
- Let $Y_i = 1$ if test for group i is positive and $Y_i = 0$ otherwise
- Then Y_i is a $\text{Bernoulli}(p)$ r.v. where

$$p = P(Y_i = 1) = P(Z_i > 0) = 1 - P(Z_i = 0)$$
$$= 1 - \binom{100}{0} 0.002^0 (1 - 0.002)^{100} = 1 - 0.998^{100} = 0.181$$

- Let $Y = Y_1 + \cdots + Y_{10}$ = the number of groups where every individual has to be tested.
  Then $Y \sim \text{Binomial}(10, 0.181)$

Example: Blood testing (Example 5.2.7) – continued

The expected number of tests we need is $10 + 100Y$ and

$$E(10 + 100Y) = 10 + 100E(Y) = 10 + 100(10 \times 0.181) = 191$$

much better than having to test 1000 blood samples!

What is the probability of the worst case scenario, i.e. that we have to do 1010 tests?

$$P(Y = 10) = \binom{10}{10} 0.181^{10} 0.819^{0} \approx 3.8 \times 10^{-8}$$

Hypergeometric distributions

Def: Hypergeometric distributions

A random variable X has the *Hypergeometric distribution with parameters N, M and n* if it has the pf

$$f(x|N,M,n) = \frac{\binom{N}{x}\binom{M}{n-x}}{\binom{N+M}{n}}$$

Parameter space: N, M and n are nonnegative integers with $n \leq N + M$

Reasoning:

- Say we have a finite population with N items of type I and M items of type II.
- Let X be the number of items of type I when we take n samples without replacement from that population
- Then X has the hypergeometric distribution

Hypergeometric distributions

- Binomial: Sampling with replacement (effectively infinite population)
- Hypergeometric: Sample without replacement from a finite population
- You can also think of the Hypergeometric distribution as a sum of dependent Bernoulli trials

In some cases we can use the Binomial distribution instead of the Hypergeometric:

- Theorem 5.3.4: If the samples size n is much smaller than the total population $N + M$ then the Hypergeometric distribution with parameters N, M and n will be nearly the same as the Binomial distribution with parameters

$$n \quad \text{and} \quad p = \frac{N}{N+M}$$

Poisson distributions

Def: Poisson distributions – Poisson(λ)

A random variable X has the *Poisson distribution with mean* λ if it has the pf

$$f(x|\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{for } x = 0, 1, 2 \ldots \\ 0 & \text{otherwise} \end{cases}$$

Parameter space: $\lambda > 0$

Show that

- $f(x|\lambda)$ is a pf
- $E(X) = \lambda$
- $Var(X) = \lambda$
- $\psi(t) = e^{\lambda(e^t - 1)}$

The cdf does not have a particular form:
$F(x|\lambda) = \sum_{k=0}^{x} \frac{e^{-\lambda}\lambda^x}{x!} =$ yikes.

Why Poisson?

- The Poisson distribution is useful for modeling uncertainty in counts / arrivals
- Examples:
  - How many calls arrive at a switch board in one hour?
  - How many busses pass while you wait at the bus stop for 10 min?
  - How many bird nests are there in a certain area?
- Under certain conditions (Poisson postulates) the Poisson distribution can be shown to be the distribution of the number of arrivals (Poisson process). However, the Poisson distribution is often used as a *model* for uncertainty of counts in other types of experiments.
- The Poisson distribution can also be used as an approximation to the Binomial(n, p) distribution when n is large and p is small.

Poisson Postulates

For $t \geq 0$, let X_t be a random variable with possible values in $\mathbb{N}_0$
(Think: $X_t =$ number of arrivals from time 0 to time t)

(i) Start with no arrivals: $X_0 = 0$

(ii) Arrivals in disjoint time periods are ind.: X_s and $X_t - X_s$ ind. if $s < t$

(iii) Number of arrivals depends only on period length:

$$X_s \text{ and } X_{t+s} - X_t \text{ are identically distributed}$$

(iv) Arrival probability is proportional to period length, if length is small:

$$\lim_{t \to 0} \frac{P(X_t = 1)}{t} = \lambda$$

(v) No simultaneous arrivals: $\lim_{t \to 0} \frac{P(X_t > 1)}{t} = 0$

If (i) - (v) hold then for any integer n

$$P(X_t = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \quad \text{that is, } X_t \sim \text{Poisson}(\lambda t)$$

Can be defined in terms of spatial areas too.

Properties of the Poisson Distributions

Useful recursive property: $P(X = x) = \frac{\lambda}{x} P(X = x - 1)$ for $x \geq 1$

Theorem 5.4.4: Sum of Poissons is a Poisson

If $X_1, \ldots, X_k$ are independent r.v. and $X_i \sim \text{Poisson}(\lambda_i)$ for all i, then

$$X_1 + \cdots + X_k \sim \text{Poisson}\left(\sum_{i=1}^{k} \lambda_i\right)$$

Theorem 5.4.5: Approximation to Binomial

Let $X_n \sim \text{Binomial}(n, p_n)$, where $0 < p_n < 1$ for all n and $\{p_n\}_{n=1}^{\infty}$ is a sequence so that $\lim_{n\to\infty} np_n = \lambda$. Then

$$\lim_{n\to\infty} f_{X_n}(x|n, p_n) = e^{-\lambda}\frac{\lambda^x}{x!} = f^{\text{Poisson}}(x|\lambda)$$

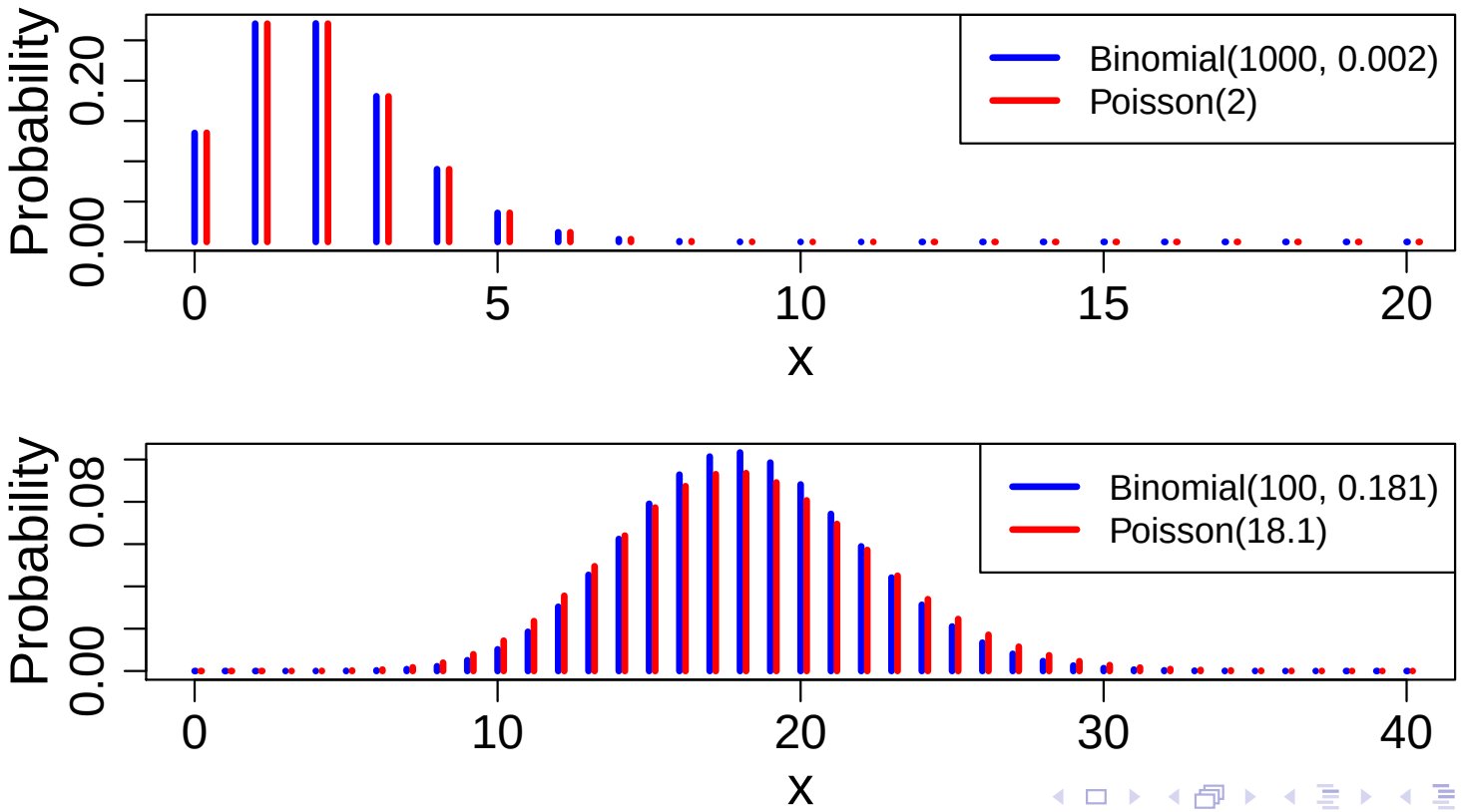
for all $x = 0, 1, 2, \ldots$

Example: Poisson as approximation to Binomial

Recall the disease testing example. We had

$$X = \sum_{i=1}^{1000} X_i \sim \text{Binomial}(1000, 0.002) \quad \text{and}$$

$$Y \sim \text{Binomial}(100, 0.181)$$



Geometric distributions

Def: Geometric distributions $\text{Geometric}(p)$

A random variable X has the *Geometric distribution with parameter p* if it has the pf

$$f(x|r,p) = \begin{cases} p(1-p)^x & \text{for } x = 0, 1, 2 \ldots \\ 0 & \text{otherwise} \end{cases}$$

Parameter space: $0 < p < 1$

- Say we have an infinite sequence of Bernoulli trials with parameter p
- $X =$ number of "failures" before the first "success" . Then $X \sim \text{Geometric}(p)$

Negative Binomial distributions

Def: Negative Binomial distributions – $\text{NegBinomial}(r, p)$

A random variable X has the *Negative Binomial distribution with parameters r and p* if it has the pf

$$f(x|r,p) = \begin{cases} \binom{r+x-1}{x} p^r (1-p)^x & \text{for } x = 0, 1, 2 \ldots \\ 0 & \text{otherwise} \end{cases}$$

Parameter space: $0 < p < 1$ and r positive integer.

- Say we have an infinite sequence of Bernoulli trials with parameter p
- $X =$ number of "failures" before the r^{th} "success". Then $X \sim \text{NegBinomial}(r, p)$
- $\text{Geometric}(p) = \text{NegBinomial}(1, p)$
- Theorem 5.5.2: If $X_1, \ldots, X_r$ are i.i.d. $\text{Geometric}(p)$ then $X = X_1 + \cdots + X_r \sim \text{NegBinomial}(r, p)$