Discrete Univariate and Multivariate Distributions, Intro to Continuous

stat 430
Heike Hofmann
Outline

• Overview: Binomial, Negative Binomial, Poisson, and Geometric
• Compound Discrete Distributions
• Continuous Random Variables: uniform, exponential
Overview

• Binomial: number of success in n independent, identical Bernoulli trials

• Geometric: number of attempts until the first success in independent, identical Bernoulli trials

• Negative Binomial: number of attempts until the r th success in independent, identical Bernoulli trials
Overview

- Poisson: number of ‘successes’ observed in a certain amount of time/space, given the rate at which they occur
Poisson approximation of Binomial

- For large $n$ ($n \geq 30$) and small $p$ ($p \leq 0.05$) the Poisson distribution approximates the Binomial distribution:

$$B_{n,p} \approx Po_\lambda \quad \text{with } \lambda = np$$
Multiple R.Vs

- Let $X$ and $Y$ be two categorical variables, then $p_{X,Y}$ is the joint pmf:

$$p_{X,Y}(x_i, y_j) = P(X = x_i \text{ and } Y = y_j)$$

- Marginal distributions:
  - sum over all $j$ for $X$,
  - sum over all $i$ for $Y$

<table>
<thead>
<tr>
<th></th>
<th>$X=1$</th>
<th>$X=2$</th>
<th>$X=i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y=1$</td>
<td>$\pi_{11}$</td>
<td>$\pi_{12}$</td>
<td>$\pi_{1i}$</td>
</tr>
<tr>
<td>$Y=2$</td>
<td>$\pi_{21}$</td>
<td>$\pi_{22}$</td>
<td>$\pi_{2i}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$Y=J$</td>
<td>$\pi_{J1}$</td>
<td>$\pi_{J2}$</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Joint distributions

• expected value

\[ E[h(X, Y)] := \sum_{x,y} h(x, y)p_{X,Y}(x, y) \]

• variance/covariance

\[ Cov(X, Y) = E[(X - E[X])(Y - E[Y])] \]

• independence

\( X \) and \( Y \) are independent random variables, if \( p_{X,Y} = p_X \ast p_Y \)
Rules for $E$ and $Var$

- Let $X, Y$ be random variables, let $a, b$ real values, then:

$$Var(aX+bY) = a^2 Var(X) + b^2 Var(Y) + 2ab Cov(X,Y)$$

$$Cov(X,Y) = E[XY] - E[X] E[Y]$$

- also:

$$Cov(X,Y) = Cov(Y,X)$$

$$Cov(aX,bY) = ab Cov(X,Y)$$
Correlation

• $\rho := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$

• correlation measures amount of linear association between $X$ and $Y$

• $\rho$ is between -1 and 1

• if $\rho = 1$ or -1, $Y$ is a linear function of $X$

$$
\rho = 1 \quad \rightarrow \quad Y = aX + b \text{ with } a > 0, \\
\rho = -1 \quad \rightarrow \quad Y = aX + b \text{ with } a < 0,
$$
Correlation

• If $X,Y$ are independent, 
  $\text{Cov}(X,Y) = 0$ (then also correlation = 0)

• If $\text{Cov}(X,Y) = 0$, 
  $X$ and $Y$ are not necessarily independent 
  (but there’s no linear relationship between $X$ and $Y$)
Continuous Random Variables
Continuous R.V.s

• $X$ is continuous r.v., if its image is an interval in $\mathbb{R}$

• e.g. “measurements”: temperature, weight of a person, height, time to run one mile,
## Discrete vs Continuous

<table>
<thead>
<tr>
<th>discrete random variable</th>
<th>continuous random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>image $Im(X)$ finite or countable infinite</td>
<td>image $Im(X)$ uncountable</td>
</tr>
<tr>
<td><strong>probability distribution function:</strong></td>
<td><strong>probability density function:</strong></td>
</tr>
<tr>
<td>$F_X(t) = P(X \leq t) = \sum_{k \leq [t]} p_X(k)$</td>
<td>$F_X(t) = P(X \leq t) = \int_{-\infty}^{t} f_X(x)dx$</td>
</tr>
<tr>
<td><strong>probability mass function:</strong></td>
<td><strong>probability density function:</strong></td>
</tr>
<tr>
<td>$p_X(x) = P(X = x)$</td>
<td>$f_X(x) = F_X'(x)$</td>
</tr>
<tr>
<td><strong>expected value:</strong></td>
<td><strong>expected value:</strong></td>
</tr>
<tr>
<td>$E[h(X)] = \sum_x h(x) \cdot p_X(x)$</td>
<td>$E[h(X)] = \int_x h(x) \cdot f_X(x)$</td>
</tr>
<tr>
<td><strong>variance:</strong></td>
<td><strong>variance:</strong></td>
</tr>
<tr>
<td>$Var[X] = E[(X - E[X])^2] = \sum_x (x - E[X])^2 p_X(x)$</td>
<td>$Var[X] = E[(X - E[X])^2] =$</td>
</tr>
<tr>
<td></td>
<td>$= \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)dx$</td>
</tr>
</tbody>
</table>
Density & Distribution

- density function
  \( f \) is density function, if it’s positive and integrates to 1

- distribution function
  \( F \) is distribution, if it’s monotone non-decreasing, has values between 0 and 1, and reaches those limits
Density & Distribution

- density is derivative of distribution
- $F' = f$

- (which makes distribution the antiderivative of the density)
Uniform Distribution

- $X$ has uniform distribution on $[a,b]$ if all values have same probability to be picked
- (idea of a “random” number)

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

- $E[X] =$
- $Var[X] =$
- $U_{a,b}(x) =$
Exponential Distribution

- X is exponentially distributed random variable, if its density is:

\[ f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \]

- lambda is called the rate parameter

- E[X] = 
  Var[X] = 
  Exp_{\lambda}(x) =

This is a very different approach to what we have looked at so far.

How long do we have to wait at most to observe a first hit with a probability of 95%?

What is the probability that we have to wait at most 2/3 min to observe the first hit?

Minutes also: The above probability then becomes:

We know the rate at which hits come to the web page in minutes - so, it's advisable to express the 40 s in hits.

What is the probability that we have to wait at most 4 Min to observe the first hit?

On average there are 2 hits per minute on a specific web page. How long do we have to wait on average to observe a first hit?

What is a sensible value for \( \lambda \)?
Exponential Distribution

• Exponential distribution is memoryless: any event is just as likely to happen within the next minute than it was in the minute ten minutes ago:

Let $X$ be an exponentially distributed r.v., i.e $X \sim \text{Exp}_\mu$ then

$$P( X \leq s) = P( X \leq t+s \mid X \geq t)$$
Exponential Distribution

• When has a variable an Exponential distribution?

• when we observe process, counting #occurrences (which is Poisson), the time or distance until next event is Exponential (continuous equivalent of Geometric)