Biostatistics Preparatory Course: Methods and Computing

Lecture 3

Probability Distributions

Yay Rmarkdown!

- Install LaTeX on your computer
- Download the '2018_Lecture3_Exercises.Rmd' from the course website
- Open file in R
- Choose 'Knit to pdf' and cross your fingers

Probability Distributions

- In statistics, we try to draw conclusions about a larger population from a sample of observations
- We use mathematical models to capture probabilistic behavior of a population
- This behavior is modeled using probability distributions

Density/Distribution Functions

Definition (Cumulative Distribution Function)

$$F_X(x) = P(X \leq x) \; \forall x \in \mathbb{R}$$

- A CDF is associated with every RV X
- A RV is continuous if $F_X(x)$ is continuous in x, and discrete if $F_X(x)$ is a step function in x
  - A discrete RV takes on a finite/countable number of values, e.g. subset of natural numbers
  - A continuous RV takes on value from some uncountable subset of the reals

Density/Distribution Functions cont.

Definition (Probability Mass Function)

For a discrete RV, the **probability mass function (PMF)** is:

$$f_X(x) = P(X = x) \, \forall x \in \mathbb{R}$$

Density/Distribution Functions cont.

Definition (Probability Mass Function)

For a discrete RV, the **probability mass function (PMF)** is:

$$f_X(x) = P(X = x) \forall x \in \mathbb{R}$$

Definition (Probability Density Function)

For a continuous RV, the **probability density function (PDF)** is:

$$f_X(x) = \frac{\partial}{\partial t} F(t)\Big|_{t=x}$$

So $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt \forall x \in \mathbb{R}$.

Note that $f_X \geq 0$ for $\forall x$, and thus F_X is an increasing function

Expectation and Variance

Definition (Expectation)

A measure of central tendancy (a weighted average of the values of X)

$$E[X] = \sum_{x \in S} x \; P(X = x) \text{ for discrete RV taking values from } S$$

$$E[X] = \int_{-\infty}^{\infty} x \; f_X(x) dx \text{ for continuous RV}$$

Expectation and Variance

Definition (Expectation)

A measure of central tendancy (a weighted average of the values of X)

$$E[X] = \sum_{x \in S} x \ P(X = x) \text{ for discrete RV taking values from } S$$

$$E[X] = \int_{-\infty}^{\infty} x \ f_X(x) dx \text{ for continuous RV}$$

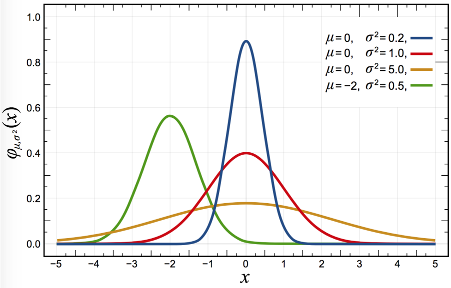
Definition (Variance)

A measure of the spread of a distribution

$$Var(X) = \textstyle\sum_{x \in S} (x - E[X])^2 \ \ P(X = x) \text{ for discrete RV}$$
$$Var(X) = \textstyle\int_{-\infty}^{\infty} (x - E[X])^2 f_X(x) dx \text{ for continuous RV}$$

Example of Continuous Distribution (Normal)

- The normal distribution is a very important distribution because:
  - A lot of things look normal
  - Analytically tractable
  - Central limit theorem
- $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for $\forall x \in \mathbb{R}$
- Characterized by mean, μ, and variance, σ^2



How to Generate Samples from Normal Distribution

The following commands are for a normal random variable with mean μ and variance σ^2, that is, $X \sim N(\mu, \sigma^2)$,

- To calculate the probability density function at a value x, i.e. $f_X(x)$

```
dnorm(x,mu,sigma)
```

- To calculate the cumulative distribution function at a value x, i.e. $P(X \leq x)$

```
pnorm(x,mu,sigma)
```

- To generate a size m sample from the normal distribution, i.e. $X_1, ..., X_m$ where $X_i \sim N(\mu, \sigma^2)$.

```
rnorm(m,mu,sigma)
```

- Note that the third argument is the **square root of the variance**, this is because the R function for normal distribution asks for the standard deviation, which is defined as the square root of the variance

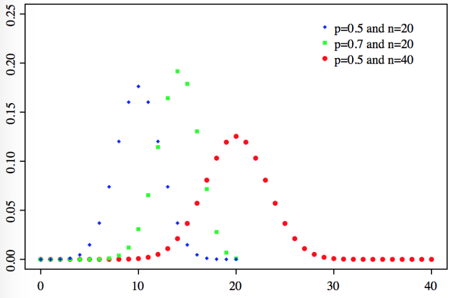
Normal Distribution Exercises

In general: Probability Distributions in R

- Many probability distributions are defined in R
- All common distributions (and most others) have four functions associated with them:
  - **Density**: the probability mass function for discrete or the probability density function for continuous random variables. Prefixed by `d` (eg., `dnorm`).
  - **Distribution function**: the cumulative distribution function, $P(X \leq x)$. Prefixed by `p` (eg., `pnorm`).
  - **Quantile function**: The inverse cdf. Prefixed by `q` (eg., `qnorm`).
  - **Random generation**: Generate n random values from the distribution. Prefixed by `r` (eg., `rnorm`).
- Using `?` with any of the four functions brings the help for all of them (eg., `?rnorm`).

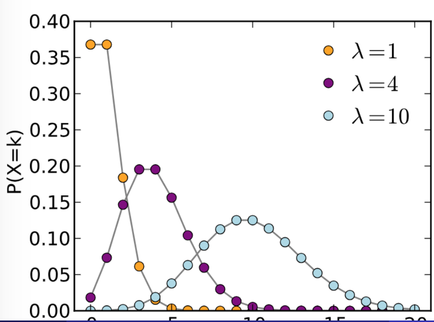
Example of Discrete Distribution (Binomial)

- Bernoulli (p) RV, is 1 with probability p and 0 with probability $1-p$
- Binomial (n,p) RV, sum of n independent Bernoulli (p) RV
  - Fixed number, n, of Bernoulli trials
  - Each trial has the same probability of success
- $f_X(k) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0 \ldots n$
- Characterized by p and n



Example of Discrete Distribution (Poisson)

- Poisson (λ) RV has an event rate $\lambda > 0$
- $f_X(k) = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$
  - k can be thought of as the number of events that occur in a given time period or space
  - λ can also be though of as the mean number of events that occur in a given time period or space as $E[X] = \lambda$
- Characterized by λ



Some Probability Distributions in R: The Most Common

Distribution	R Abbrev	Description
Normal	`norm`	everyone's favorite bell curve
t	`rt`	standard normal distribution with wider tails
Uniform	`unif`	equal probability on the chosen interval
Binomial	`binom`	probability for a given number of successes in a fixed number of experiments
Poisson	`pois`	probability of a given number of events occurring in a fixed interval of time or space
Geometric	`geom`	probability that it takes a given number of failures until one success

Some Probability Distributions in R: Other Useful Ones

Continuous

- Beta (`?rbeta`)
- Chi-sq (`?rchisq`)
- Exponential (`?rexp`)
- F (`?rf`)
- Logistic (`?rlogis`)
- Lognormal (`?rlnorm`)

Discrete

- Geometric (`?rgeom`)
- Negative Binomial (`?rnbinom`)
- Multinomial (`?rmultinom`)

Empirical vs. Theoretical CDF

In statistics, an empirical distribution function is the distribution function associated with the empirical measure of a **sample**.
We can denote the theoretical CDF as (this is what dnorm gives you!):

$$F_X(k) = Pr(X \leq k)$$

and the empirical as:

$$\hat{F}_n(k) = \frac{\text{number of elements in the sample} \leq k}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq k}$$

where $X_1, ..., X_n$ make up some random sample from the underlying distribution.

Poisson Distribution Exercises

Group Exercises