6. Random Variables calculations and wrap up

Transfer exploration seminar: Statistics and Data Science

Dr. Uma Ravat
PSTAT 194TR

Summary: Random Variables

- Random Variables: Discrete , Continuous
  - distribution function: pmf $P(X = x)$, pdf $f(x)$
  - probabilities : sums, series and integrals
  - Expected value: μ, $\bar{X}$, $E(X)$
  - Variance: σ^2, $Var(X)$

Become familiar with the notation and concepts, the algebra will follow much more easily

Next we will see...

- Random variables
  - Using calculus (integrals and series) in probability calculations
- Wrap up with a look at connection of Intro to R and Intro to Probability.
- Feedback about these modules.
- Discuss anything else you want.

Become familiar with the notation and concepts, the algebra will follow much more easily

Pre-reading: Math Review:

- Sum and series
- Integrals

Comparison of Discrete and Continuous Random Variables

Random Variable	Discrete	Continuous
Takes on	finite or a countable number of distinct values	any value in a given range/interval
Example	# of correct answers on a 100 question test	Time taken to hike around the lagoon
Prob distribution function	probability mass function (PMF) $P(X = x)$	probability density function (PDF) $f(x)$
Properties	$0 \leq P(X = x) \leq 1$ $\sum_x P(X = x) = 1$	$0 \leq f(x)$ $\int_x f(x) = 1$ (total area under PDF = 1)

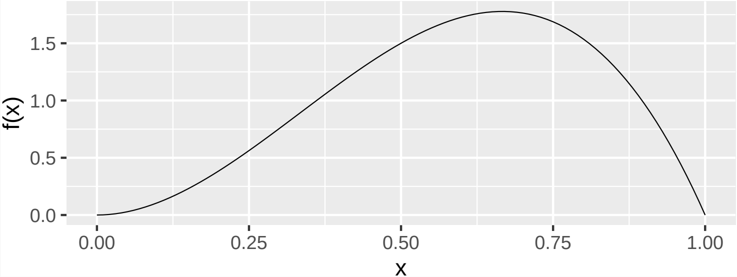
Comparison of Discrete and Continuous Random Variables

Random Variable	Discrete	Continuous
Probability Calculation	$P(X = x)$ gives the probability of a specific value	$P(a \leq X \leq b) = \int_a^b f(x)dx$ gives probability over an interval
Mean (Expected Value)	$E(X) = \sum_x x \cdot P(X = x)$	$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$
Variance	$Var(X) = \sum_x (x - E(X))^2 \cdot P(X = x)$	$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x)\, dx$

Revisit: Example of continuous random variable

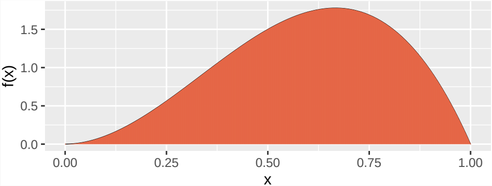
A statistical quality control example.

$$f(x) = \begin{cases} 12\ x^2\ (1-x) & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$



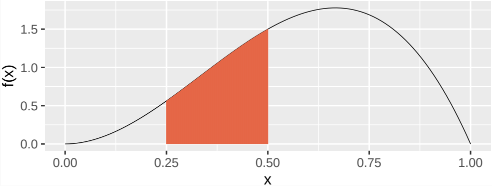
Area under the curve = 1

$f(x) = 12\ (x^2)\ (1 - x),\ 0 \leq x \leq 1$



$$\int_{x\ \epsilon\ S} f(x)dx = \int_0^1 12(x^2)(1 - x)dx$$
$$= 12\int_0^1 (x^2 - x^3)dx = 12\Big[\frac{x^3}{3} - \frac{x^4}{4}\Big|_0^1\Big] = 1$$

Probability is Area Under the Curve



$$P(0.25 < X < 0.50) = \int_{0.25}^{0.50} 12(x^2)(1-x)dx = 12\int_{0.25}^{0.50}(x^2 - x^3)dx$$

$$= 12\Big[\frac{x^3}{3} - \frac{x^4}{4}\Big|_{0.25}^{0.50}\Big] = 0.2617188$$

Expected value

$$f(x) = \begin{cases} 12\ x^2\ (1-x) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} E(X) &= \int_{x\ \epsilon\ S} x f(x) dx \\ &= \int_0^1 x 12 (x^2)(1-x) dx \\ &= 12 \int_0^1 (x^3 - x^4) dx \\ &= 12 [\frac{x^4}{4} - \frac{x^5}{5} \Big|_0^1] \\ &= \boxed{0.6} \end{aligned}$$

Variance

$$f(x) = 12\,x^2\,(1-x) \text{ if } 0 \le x \le 1$$

$$\begin{aligned} Var(X) &= \int_0^1 (x - E(X))^2 \cdot f(x)\,dx \\ &= \int_0^1 (x-0.6)^2 \cdot 12 \cdot \; x^2 \; \cdot (1-x)\,dx \\ &= \dots \\ &= \dots \\ &= \dots \\ &= \dots \\ &= \dots \\ &= \boxed{0.04} \end{aligned}$$

Variance

$$f(x) = 12\ x^2\ (1 - x) \text{ if } 0 \leq x \leq 1$$

$$\begin{aligned}
Var(X) &= \int_0^1 (x - E(X))^2 \cdot f(x)\, dx \\
&= \int_0^1 (x - 0.6)^2 12\ x^2\ (1 - x)\, dx = \int_0^1 (x^2 - 1.2x + 0.36) 12x^2 (1 - x)\, dx \\
&= 12 \int_0^1 (x^4 - 1.2x^3 + 0.36x^2)(1 - x)\, dx = 12 \int_0^1 (x^4 - 1.2x^3 + 0.36x^2) - (x^5 - 1.2x^4 + 0.36x^3)\, dx \\
&= 12 \int_0^1 (x^4 - 1.2x^3 + 0.36x^2) - x^5 + 1.2x^4 - 0.36x^3\, dx = 12 \int_0^1 -x^5 + 2.2x^4 - 1.56x^3 + 0.36x^2\, dx \\
&= 12 \left[\frac{-x^6}{6} + \frac{2.2x^5}{5} - \frac{1.56x^4}{4} + \frac{0.36x^3}{3} \bigg|_0^1 \right] \\
&= 12 \left[-16 + \frac{2.2}{5} - \frac{1.56}{4} + \frac{0.36}{3} \right] \\
&= \boxed{0.04}
\end{aligned}$$

Variance: Equivalent definitions

$$
\begin{aligned}
Var(X) &= \int_x (x - E(X))^2 \cdot f(x)\, dx \\
&= E[(X - E(X))^2] \\
&= E(X^2) - [E(X)]^2
\end{aligned}
$$

$E(X^2) = ?$

$$
\begin{aligned}
E(X^2) &= \int_x x^2 f(x) dx = \int_0^1 x^2 12(x^2)(1-x) dx \\
&= 12 \int_0^1 (x^4 - x^5) dx \\
&= 12 \Big[\frac{x^5}{5} - \frac{x^6}{6} \Big|_0^1 \Big] \\
&= \boxed{0.4} \\
Var(X) &= 0.4 - 0.6^2 = \boxed{0.04}
\end{aligned}
$$

Example: Flipping two coins

$X =$ number of heads in two independent coin flips

X is a __________________ random variable.

Sample space is S $= \{$ ______________ $\}$

PMF of X is is ______________

Example: Flipping two coins

$X =$ number of heads in two independent coin flips

X is a *discrete* random variable.

Sample space is $S = \{HH, HT, TH, TT\} = \{0, 1, 2\}$

PMF is

Values: $\text{X} = \text{x}$	0	1	2
Probability: $P(X = x) = P(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Expected value and variance

X = number of heads in two independent coin flips, $S_X = \{0, 1, 2\}$

$E(X) = \sum_x xP(x)$

$\boldsymbol{x}$	0	1	2	
$\boldsymbol{P(x)}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	
$\boldsymbol{xP(x)}$	$0 = (0)\frac{1}{4}$	$\frac{1}{2} = (1)\frac{1}{2}$	$\frac{1}{2} = (2)\frac{1}{4}$	$1 = E(X)$

$$\text{So}, \quad E(X) = \sum_x xP(x) = 1$$

On average, you expect to see 1 head if you flip two coins.

Variance

X = number of heads in two independent coin flips, S_X=$\{0, 1, 2\}$

$$VarX = E(X^2) - [E(X)]^2 = \sum_{\text{all } x} x^2 P(x) - [E(X)]^2$$

$E(X) = 1, E(X^2) =?$

$\boldsymbol{x}$	0	1	2	
$\boldsymbol{P(x)}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	
$\boldsymbol{xP(x)}$	$0 = (0)\frac{1}{4}$	$\frac{1}{2} = (1)\frac{1}{2}$	$\frac{1}{2} = (2)\frac{1}{4}$	$1 = E(X)$
$\boldsymbol{x^2P(x)}$	$0 = (0^2)\frac{1}{4}$	$\frac{1}{2} = (1^2)\frac{1}{2}$	$1 = (2^2)\frac{1}{4}$	$1.5 = E(X^2)$

$Var(X) = E(X^2) - [E(X)]^2 = 1.5 - 1^2 = 0.5$

Recall: Series and it's sum

If b is any number, the geometric series and it's sum is given by

$$\sum_{k=0}^{\infty} b^k = \left(1 + b^1 + b^2 + \cdots + b^k + \cdots\right) = \frac{1}{1-b} = \frac{\text{first term}}{1 - \text{base}}, \textit{if}\, |b| < 1$$

$k! = 1 \cdot 2 \cdot 3 \cdots (k-1) \cdot k$ eg. $5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$

If a is any number, the exponential series and it's sum is given by

$$\begin{aligned} \sum_{k=0}^{\infty} \frac{a^k}{k!} &= \frac{a^0}{0!} + \frac{a^1}{1!} + \frac{a^2}{2!} + \cdots + \frac{a^k}{k!} + \cdots \\ &= 1 + a + \frac{a^2}{2!} + \cdots + \frac{a^k}{k!} + \cdots \\ &= e^a \end{aligned}$$

See also, series review and math review sheet.

Example: uncountable sample space

Find $E(X)$ of a discrete random variable X that has the following probability distribution:

$$P(X=0) = P(0) = 2-\sqrt{e}; P(X=k) = P(k) = \frac{1}{2^k \cdot k!},\ k = 1,2,3,\dots$$

$$\begin{aligned}
E(X) &= \sum_x xP(x) = 0(2-\sqrt{e}) + \sum_{k=1}^{\infty} k \cdot \frac{1}{2^k k!} \\
&= \sum_{k=1}^{\infty} \frac{\left(\frac{1}{2}\right)^k}{(k-1)!} \\
&= \frac{1}{2} \sum_{k-1=0}^{\infty} \frac{\left(\frac{1}{2}\right)^{(k-1)}}{(k-1)!} \\
&= \frac{1}{2} \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)^{(n)}}{(n)!} \quad (k-1=n) \\
&= \frac{1}{2} e^{1/2} = \frac{1}{2}\sqrt{e}
\end{aligned}$$

Summary:

- Probability, E(X), V(X) calculations using
  - integrals for continuous random variables
  - series for discrete random variables with uncountable sample space.

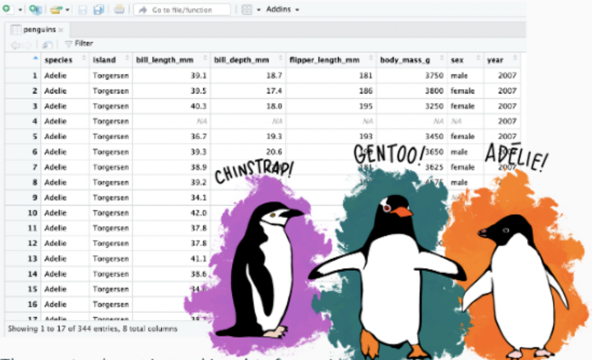
Overall summary:

- Study skills for success in PSTAT courses
- Introduction to R (in preparation for PSTAT 10)
- Introduction to Probability (in preparation for PSTAT 120a)
  - Also, review Math review slides provided.
  - Review Double Integrals

Overall Connection

Module 1(Intro to R) and Module 2(Introduction to Probability)

Sample



The penguins dataset is stored in a **data frame** with

- **344 observations/samples/cases/subjects** (rows)
  - each case represents a penguin
- **8 variables** (columns)
  - species, island, bill_length_mm, bill_depth_mm etc
  - each corresponds to some measurement of the penguin

Module 2(Introduction to Probability) -> PStat 120A -> Pstat 120B

Population

Random variables

Population parameters

- Population mean
- Population variance

Sampling distributions

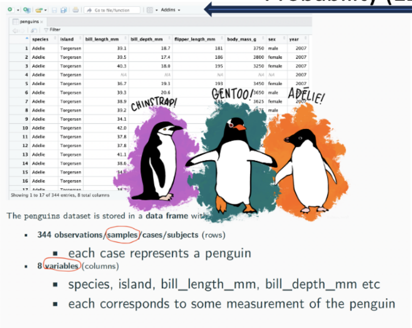
Central Limit Theorem

Connection

Statistics (120B)

Sample ⇄ Population

Probability (120A)



Variables

Summary statistics

- sample mean
- sample variance

Visualizations

Random variables

Population parameters

- Population mean
- Population variance

Sampling distributions

Central Limit Theorem

Overall summary:

- Study skills for success in PSTAT courses
- Introduction to R (in preparation for PSTAT 10)
- Introduction to Probability (in preparation for PSTAT 120a)
    - Also, review Math review slides provided.
    - Review Double Integrals

Next, Intro to Python