

Probability and Statistics - Module 2

Lula Chen for UIUC Political Science Math Camp

8/25/2017

Plan for the Afternoon¹

Questions from morning session?

Types of Solutions

- ▶ Application

Distributions

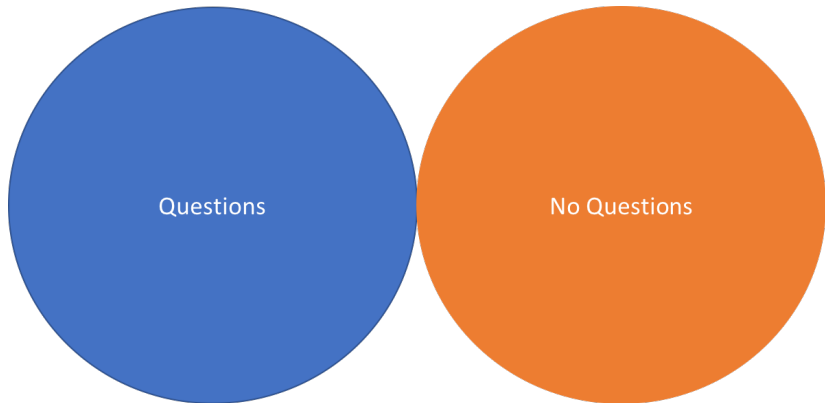
Variance and Covariance

Expectation

- ▶ Application

¹I used Hogg, Tanis, and Zimmerman (2015), Harvard's Political Science Math Camp, and Wolfram Alpha as sources for these slides.

Questions from the morning?



Types of Solutions

You have a favorite restaurant, open Monday-Saturday. The restaurant rotates through a menu of **six** items and it serves **one** item only every day. The rotation is random, so it may serve the same dish two or more days in a row.

You are determined to try every item on the menu. On average, how many consecutive days do you have to go to the restaurant to try every dish?

How would you solve this problem?

Definitions

Analytic solution

A solution to a problem that can be written in “closed-form” formula in terms of well-known functions, constants, etc.

A closed-form solution solves a given problem in terms of functions and mathematical operations from a given, generally-accepted set.

Imagine that you find your solution by writing an equation and solving it.

Example: Solve for x

$$x + 2 = 5$$

$$x + 2 - 2 = 5 - 2$$

$$x = 3$$

Defintions

Computational solution (approximation)

A solution (approximation) to a problem that comes from using numerical methods.

Numerical methods are methods that use algorithms (series of steps) that give approximations to study mathematical problems.

Imagine that you find your solution by plugging in many different numbers and finding an approximation.

Example: Solve for x , but you don't know subtraction

$$x + 2 = 5$$

Plug in $x = 2$. $2 + 2 = 4$, it's too small!

Plug in $x = 4$. $4 + 2 = 6$, it's too big!

Keep plugging in numbers until you arrive at (or close enough) to 5.

Why is this distinction important?

- ▶ Sometimes, we can write out and solve a formula for our problem. But sometimes we can't or we don't know the math to help us solve it.
- ▶ When we face situations where the analytic solution seems difficult, programs such as R can be helpful to us to find an approximation through computation.

Application: Analytic or computational?

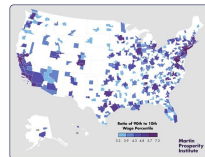
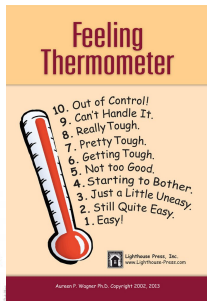
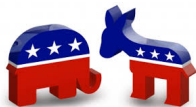
You have a favorite restaurant, open Monday-Saturday. The restaurant rotates through a menu of six items and it serves one item only every day. It rotates through its items, but the rotation is random, so it may serve the same dish two or more days in a row.

You are determined to try every dish. How many consecutive days do you have to go to the restaurant, on average, to try every dish?

In what way did you try to solve this problem?

Types of data:

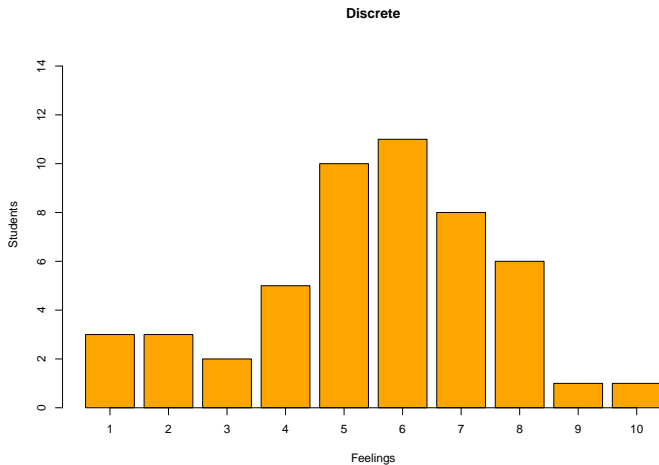
Categorical, Ordinal, Interval, Ratio



Types of data

Discrete

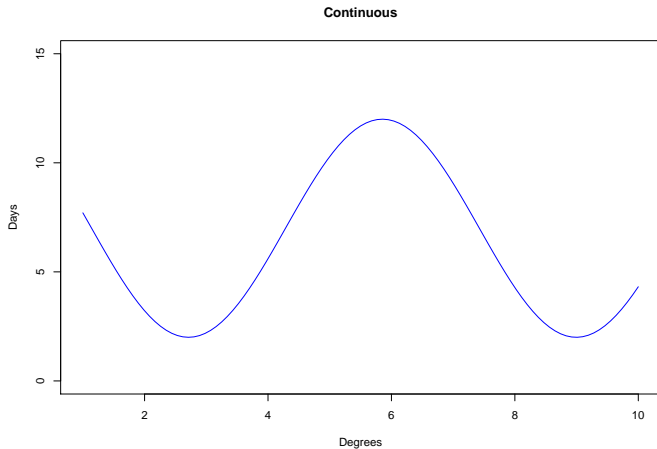
- ▶ Finite or countably infinite
- ▶ Categorical, Ordinal, (some) Interval



Types of data

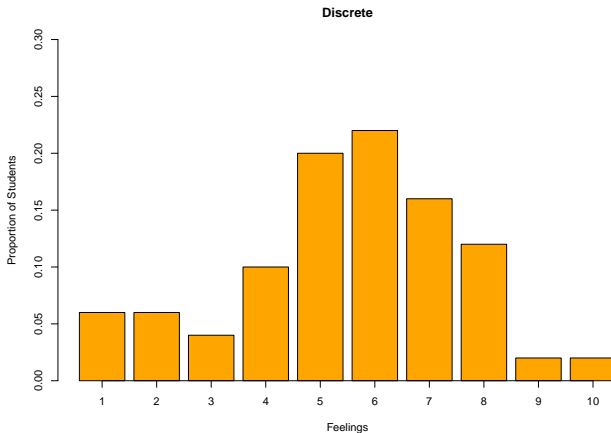
Continuous

- ▶ Uncountably infinite
- ▶ Interval, Ratio



Distribution

What percentage of students are feeling like the class is “starting to bother” ($x=4$)?



We can tell because we plotted the *distribution* of the variable.

Distribution

Distribution

The distribution of a variable is a description of the relative numbers of times each possible outcome will occur in a number of trials.

Importantly, distributions tell us the probability that certain values will occur.

We are interested in distributions of *random variables*.

Distribution: Discrete

probability mass function (pmf)

For a discrete random variable, the pmf $f(x)$ is the function that tells the probability that a given value will occur $P(X = x)$. It has the following properties (among others):

$$0 \leq f(x) \leq 1$$

$$\sum f(x) = 1$$

Distribution: Discrete

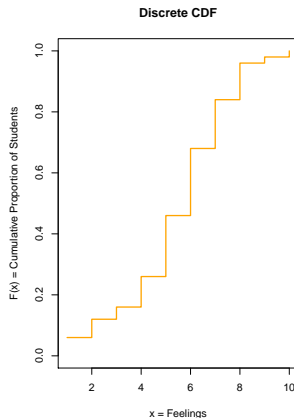
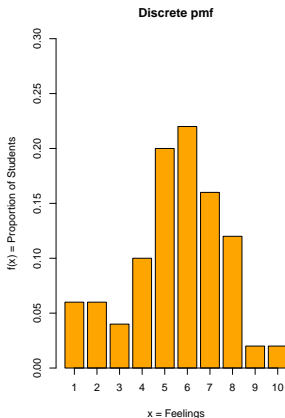
cumulative distribution function (CDF)

For a discrete random variable, the CDF $F(x)$ is the function that tells us the cumulative probability that a given value or any value smaller than it will occur $P(X \leq x)$, $-\infty < x < \infty$.

Distribution: Discrete

How is CDF related to pmf?

$$F(x) = P(X \leq x) = \sum_i^x f(x)$$



Distribution: Continuous

probability density function (pdf)

For a continuous random variable, the pdf $f(x)$ is the function that tells us the probability that a random variable will fall within a particular range of values. It has the following properties (among others):

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$P(X = x) = 0 \text{ (because the integral at a single point is 0)}$$

Distribution: Continuous

cumulative distribution function (CDF)

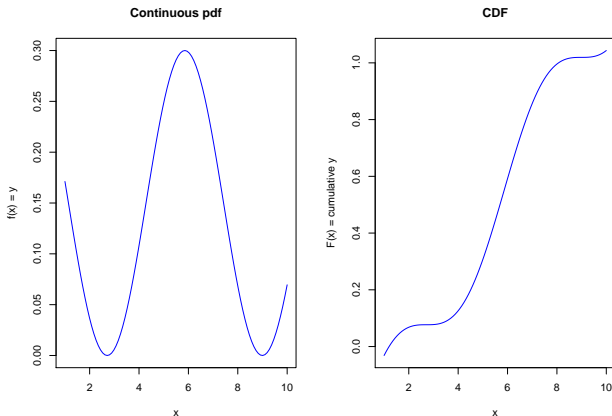
For a continuous random variable, the CDF $F(x)$ is the function that tells us the cumulative probability that a given value or any value smaller than it will occur $P(X \leq x)$, $-\infty < x < \infty$.

Distribution: Continuous

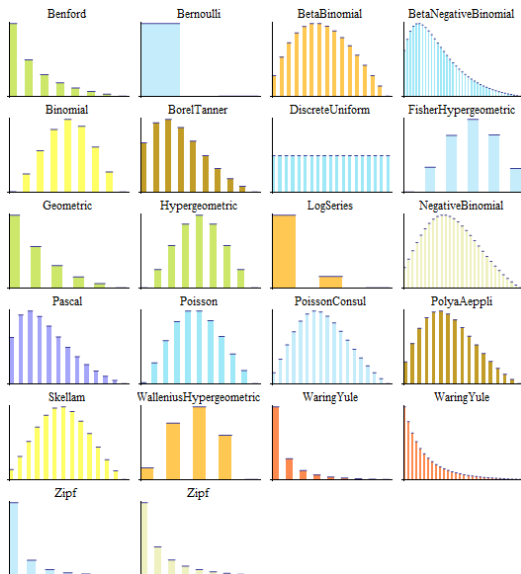
The pdf $f(x)$ and CDF $F(x)$ are related in the same way that the pmf and CDF are related. Since the pdf is continuous, the pdf is the derivative of the CDF and the CDF is the integral of the pdf.

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt, -\infty < x < \infty$$

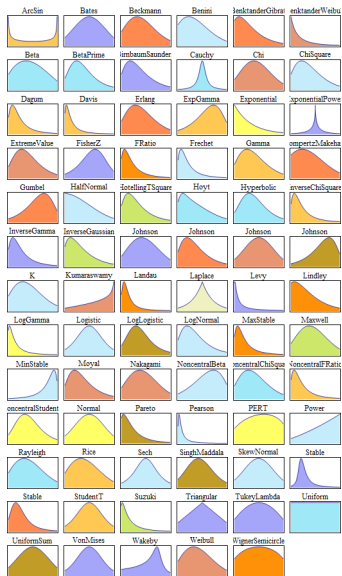
$$F'(x) = f(x)$$



Distributions

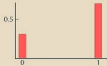
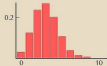
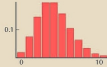
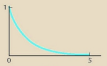

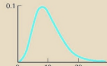
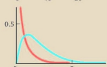


Distributions

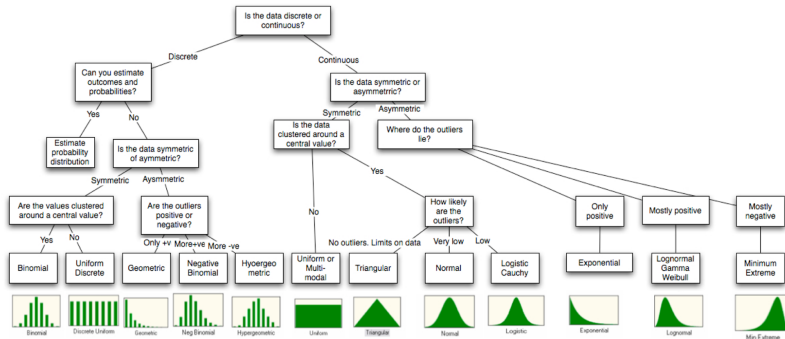


Distributions

TABLE 28.1. Common probability distributions

Distribution	Mean	Variance	
Discrete			
Two-valued $t_0 = q, t_1 = p$	p	pq	
Binomial $\frac{n!}{j!(n-j)!} q^j p^{n-j}$	np	npq	
Poisson $E^{-\lambda} \frac{\lambda^j}{j!}$	λ	λ	
Continuous			
Exponential $\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	
Gaussian (or normal) $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right)$	\bar{x}	σ^2	
Chi-square $\frac{1}{2^{n/2} \Gamma(n/2)} \left(\frac{x}{2}\right)^{n/2-1} e^{-x/2}$	n	$2n$	
Gamma $\frac{1}{\Gamma(\alpha)\beta^\alpha} \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta}$	$\alpha\beta$	$\alpha\beta^2$	

Application



Describing Distributions

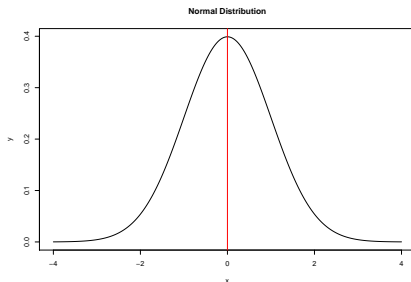
Expectation

Discrete: $E[X] = \sum xf(x)$

Continuous: $E[X] = \int xf(x)dx$

This is the sum or integral of all the possible values of X times the probability that x occurs. It is the weighted average that X will take on after many trials.

The *expected value* of X is the *mean* of x : $E[X] = \mu$



Describing Distributions

Variance

Variance tells us the spread of the data, or how far apart the data are from the mean. Variance is also an expectation; it is the weighted average of the squares of the distances between X and $E[X]$.

Discrete: $VAR[X] = \sigma^2 = E[(X - E[X])^2] = E[X^2] - (E[X])^2$

Continuous: $VAR[X] = \sigma^2 = E[(X - \mu)^2] = \int (x - \mu)^2 f(x) dx$

Describing Distributions

Standard Deviation

$$SD[X] = \sigma = \sqrt{VAR[X]}$$

Covariance

Covariance tells us how much two random variables covary together.

$$COV[X, Y] = E[(X - E[X])^2(Y - E[Y])^2] = E[XY] - E[X]E[Y]$$

Application

How would you calculate the expected value?

$$E[X] = \sum xf(x)$$

$$E[X] = 5.48$$

x	f(x)	F(x)	xf(x)
1	0.06	0.06	0.06
2	0.06	0.12	0.12
3	0.04	0.16	0.12
4	0.1	0.26	0.4
5	0.2	0.46	1
6	0.22	0.68	1.32
7	0.16	0.84	1.12
8	0.12	0.96	0.96
9	0.02	0.98	0.18
10	0.02	1	0.2
Total			5.48

Application

How would you calculate the variance?

$$\text{VAR}[X] = E[X - E[X]]^2 = E[X^2] - (E[X])^2$$

$$E[X^2] - E[X]^2 = \sum (x^2)f_x - (\sum xf(x))^2$$

$$\text{VAR}[X] = 34.32 - 5.48^2 = 4.2896$$

x	$f(x)$	$F(x)$	$x^2f(x)$	$xf(x)$
1	0.06	0.06	0.06	0.06
2	0.06	0.12	0.24	0.12
3	0.04	0.16	0.36	0.12
4	0.1	0.26	1.6	0.4
5	0.2	0.46	5	1
6	0.22	0.68	7.92	1.32
7	0.16	0.84	7.84	1.12
8	0.12	0.96	7.68	0.96
9	0.02	0.98	1.62	0.18
10	0.02	1	2	0.2
Total			34.32	5.48

From Probability to Statistics



- ▶ By building from different mathematical theories, probability tells us how our outcomes will be distributed.
- ▶ By measuring our outcomes, statistics tells us which mathematical models fit our data.

Statistics

Before, we knew about the entire population, or all of the outcomes.

Now, imagine that we only know about some of the outcomes, a sample. Imagine that these outcomes, $X_n = X_1, X_2, X_3, \dots, X_n$, are independent, and come from a distribution with finite mean (μ) and finite positive variance (σ^2). Note that we don't know *all* outcomes, only *some* of them. Therefore, we don't know what numbers μ or σ^2 equal for the distribution, just that they exist.

Central Limit Theorem

$$\bar{X}_n = \frac{x_1 + x_2 + \dots + x_n}{n} \rightarrow \mathcal{N}(\mu, \frac{\sigma^2}{n}), n \rightarrow \infty$$

As n approaches ∞ , the distribution of \bar{X}_n approaches a normal distribution, with the mean at μ .

If our observations come from independent processes, we don't need too many outcomes before we start to feel some level of comfort (maybe) for claiming that they fall within a normal distribution.

Law of Large Numbers

$$\bar{X}_n = \frac{x_1 + x_2 + \dots + x_n}{n} \rightarrow \mu, n \rightarrow \infty$$

As we have more and more outcomes in our study, the average of the outcomes will converge on μ .

As a result of the Central Limit Theorem and the Law of Large Numbers, notice that sample size matters and be cautious when you have a small sample.

Review questions

- ▶ When might you use an analytic solution vs. computational approximation?
- ▶ What is a distribution? What is the difference between the pmf, pdf, and CDF?
- ▶ How do mean and variance describe a distribution?
- ▶ Why are we comfortable using normal distributions in statistics?

Takeaways

- ▶ When we do research, we collect a lot of data and have theories about why the data look the way they do, or why they have the relationships that they do.
- ▶ The tools from probability lay the groundwork to help us determine the chances of seeing the data that we see.
- ▶ You'll learn more about how to determine if what you're seeing in the data is due to randomness or chance, or due to an actual effect in the next session!

Questions?

Thanks!

Lula - nchen3@illinois.edu