Probability II

Lula Chen for UIUC Political Science Math Camp
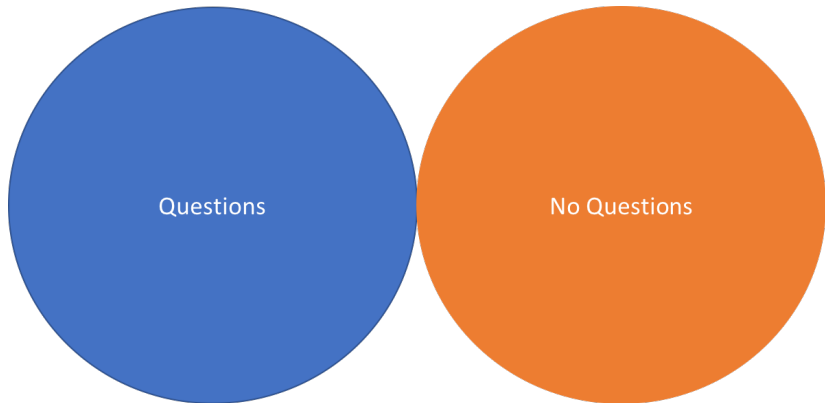
8/16/2018

Plan for the Morning[1]

Questions from yesterday's session?

- ▶ Types of Data
- ▶ Random Variable
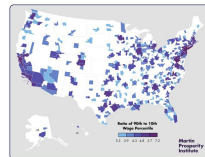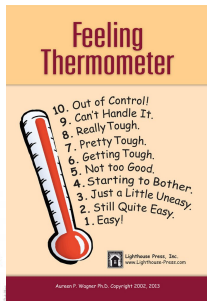- ▶ Distributions
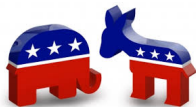- ▶ Expected Values
- ▶ Variance

---

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

Questions from yesterday?
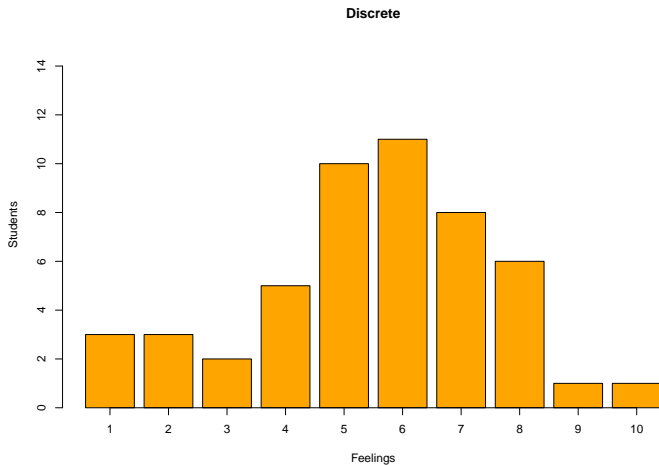
Questions

No Questions

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
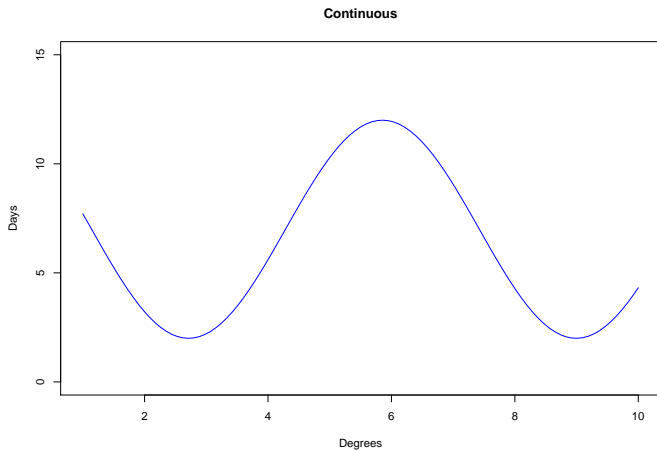
Continuous

Distribution

What percentage of students are feeling like the class is "starting to bother" (x=4)?
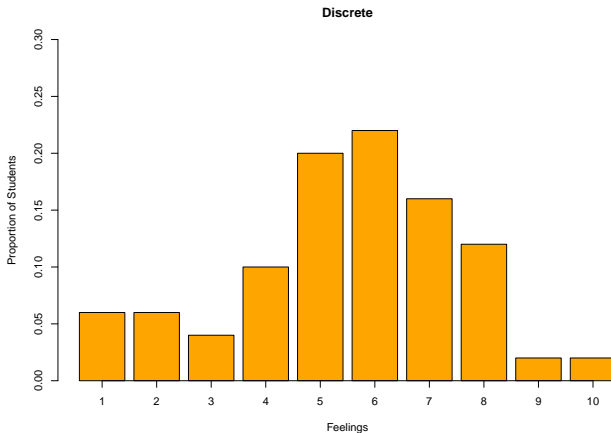
Distribution

What percentage of students are feeling like the class is "starting to bother" (x=4)?
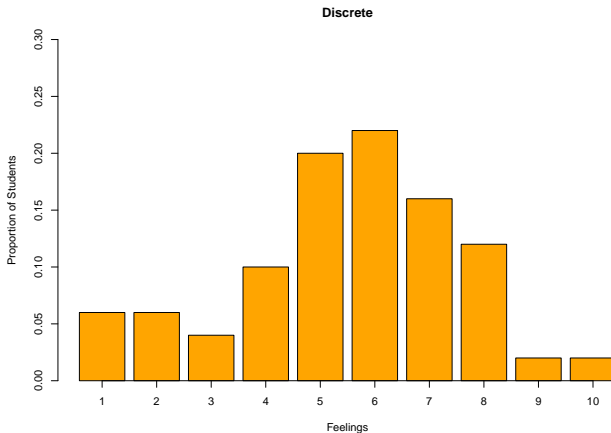
We can tell because we plotted the *distribution* of the variable.

Distribution

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

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.

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*.

Random Variable

A random variable, X, is a function, X, that assigns one and only one real number to all possible outcomes, given a random experiment.

Random Variable

A random variable, X, is a function, X, that assigns one and only one real number to all possible outcomes, given a random experiment.

$$X = \begin{cases} 1 & \text{if heads} \\ 2 & \text{if tails} \end{cases}$$

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?

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

Discrete pmf

f(x) = Proportion of Students

x = Feelings

Discrete CDF

F(x) = Cumulative Proportion of Students

x = Feelings

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):

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

$f(x) \geq 0$

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

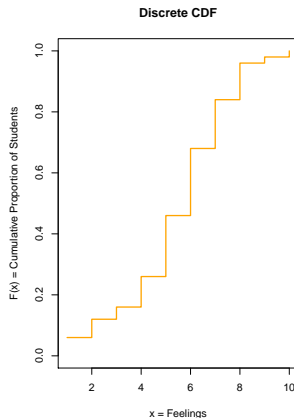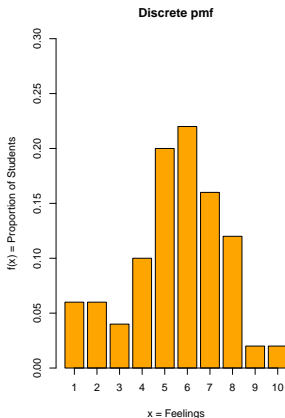$P(X = x) = 0$ (because the integral at a single point is 0)

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.

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

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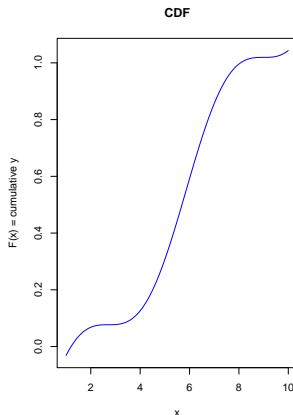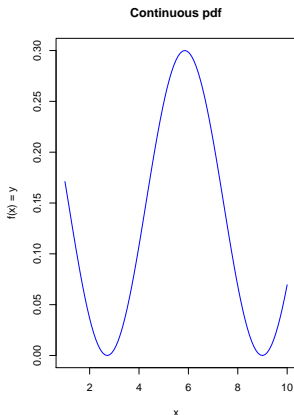
Distributions
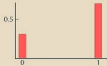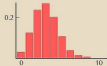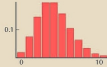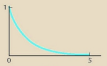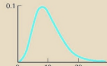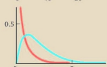
Distributions
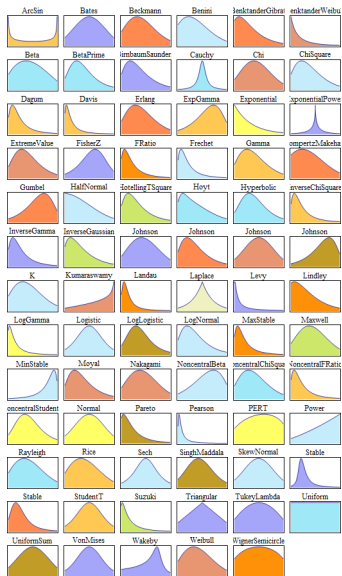
Distributions

TABLE 28.1. Common probability distributions

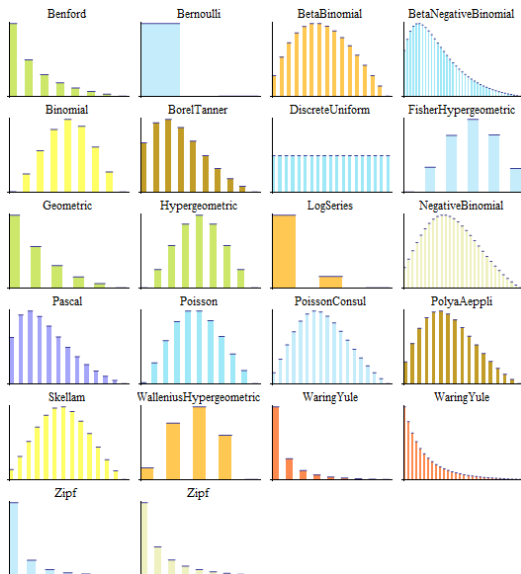| Distribution		Mean	Variance	
Discrete				
Two-valued	$f_0 = q, f_1 = p$	p	pq	
Binomial	$\dfrac{n!}{j!(n-j)!}\, q^{n-j} p^j$	np	npq	
Poisson	$e^{-\lambda}\dfrac{\lambda^j}{j!}$	λ	λ	
Continuous				
Exponential	$\lambda e^{-\lambda x}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$	
Gaussian (or normal)	$\dfrac{1}{2\pi\sigma^2}\exp\left(-\dfrac{(x-\bar{x}/2}{2\sigma^2}\right)$	$\bar{x}$	σ^2	
Chi-square	$\dfrac{1}{2\Gamma(n/2)}\left(\dfrac{x}{2}\right)^{n/2-1} e^{-x/2}$	n	$2n$	
Gamma	$\dfrac{1}{\beta\Gamma(\alpha)}\left(\dfrac{x}{\beta}\right)^{\alpha-1} e^{-x/\alpha\beta}$	$\alpha\beta$	$\alpha\beta^2$	

Application
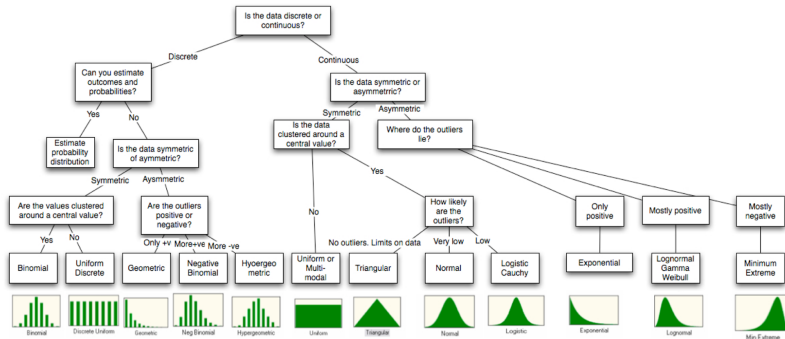
Describing Distributions

Expected Value or Expectation

Describing Distributions

Expected Value or Expectation

- The *expected value* of X 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

Expected Value or Expectation

- The *expected value* of X is the weighted average that X will take on after many trials.
- The *expected value* of X is the *mean* of X: $E[X] = \mu$

- Discrete: $E[X] = \sum x f(x)$
- Continuous: $E[X] = \int x f(x) dx$
- This is the sum or integral of all the possible values of X times the probability that outcome x occurs.

Describing Distributions

Expected Value or Expectation

Normal Distribution
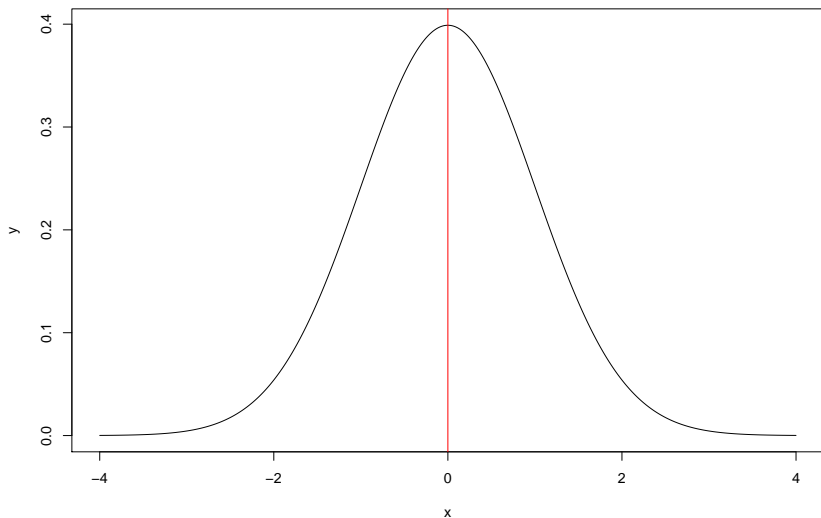
Exercise

Calculate the expected value of the feelings of students:

Feelings	Students	Feelings x Students
1	3.00	3.00
2	3.00	6.00
3	2.00	6.00
4	5.00	20.00
5	10.00	50.00
6	11.00	66.00
7	8.00	56.00
8	6.00	48.00
9	1.00	9.00
10	1.00	10.00

Exercise

Calculate the expected value of the feelings of students:

Feelings	Students	Feelings x Students
1	3.00	3.00
2	3.00	6.00
3	2.00	6.00
4	5.00	20.00
5	10.00	50.00
6	11.00	66.00
7	8.00	56.00
8	6.00	48.00
9	1.00	9.00
10	1.00	10.00

$\sum(FeelingsxStudents) = 3 + 6 + 6 + 20 + 50 + ... + 10 = 274$

Exercise

Calculate the expected value of the feelings of students:

Feelings	Students	Feelings x Students
1	3.00	3.00
2	3.00	6.00
3	2.00	6.00
4	5.00	20.00
5	10.00	50.00
6	11.00	66.00
7	8.00	56.00
8	6.00	48.00
9	1.00	9.00
10	1.00	10.00

$\sum(FeelingsxStudents) = 3 + 6 + 6 + 20 + 50 + ... + 10 = 274$

$E[FeelingsofStudents] = 274/50 = 5.48$

Exercise

Calculate the expected value of the feelings of students:

Feelings (x)	Students	Feelings x Students	f(x)	xf(x)
1	3.00	3.00	0.06	0.06
2	3.00	6.00	0.06	0.12
3	2.00	6.00	0.04	0.12
4	5.00	20.00	0.10	0.40
5	10.00	50.00	0.20	1.00
6	11.00	66.00	0.22	1.32
7	8.00	56.00	0.16	1.12
8	6.00	48.00	0.12	0.96
9	1.00	9.00	0.02	0.18
10	1.00	10.00	0.02	0.20

Exercise

Calculate the expected value of the feelings of students:

Feelings (x)	Students	Feelings x Students	f(x)	xf(x)
1	3.00	3.00	0.06	0.06
2	3.00	6.00	0.06	0.12
3	2.00	6.00	0.04	0.12
4	5.00	20.00	0.10	0.40
5	10.00	50.00	0.20	1.00
6	11.00	66.00	0.22	1.32
7	8.00	56.00	0.16	1.12
8	6.00	48.00	0.12	0.96
9	1.00	9.00	0.02	0.18
10	1.00	10.00	0.02	0.20

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

Exercise

Calculate the expected value of the feelings of students:

Feelings (x)	Students	Feelings x Students	f(x)	xf(x)
1	3.00	3.00	0.06	0.06
2	3.00	6.00	0.06	0.12
3	2.00	6.00	0.04	0.12
4	5.00	20.00	0.10	0.40
5	10.00	50.00	0.20	1.00
6	11.00	66.00	0.22	1.32
7	8.00	56.00	0.16	1.12
8	6.00	48.00	0.12	0.96
9	1.00	9.00	0.02	0.18
10	1.00	10.00	0.02	0.20

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

$E[X] = 0.06 + 0.12 + 0.12 + 0.40 + ... + 0.20 = 5.48$

Describing Distributions

Variance

Describing Distributions

Variance

- Variance tells us the spread of the data, or how far apart the data are from the mean.

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 $E[X]$.

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

- ▶ The standard deviation is the square root of the variance.

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

Application

How would you calculate the variance?

$$VAR[X] = E[X - E[X]]^2 = E[X^2] - (E[X])^2$$

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

$$VAR[X] = 34.32 - 5.48^2 = 4.2896$$

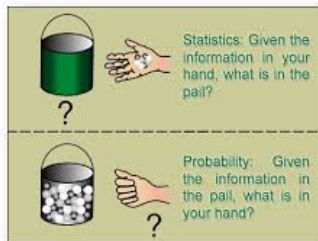| x	f(x)	F(x)	$x^2f(x)$	xf(x)
1	0.06	0.06	0.06	0.06
2	0.06	0.12	0.24	0.12
3	0.04	0.16	0.36	0.12
4	0.1	0.26	1.6	0.4
5	0.2	0.46	5	1
6	0.22	0.68	7.92	1.32
7	0.16	0.84	7.84	1.12
8	0.12	0.96	7.68	0.96
9	0.02	0.98	1.62	0.18
10	0.02	1	2	0.2
		Total	34.32	5.48

From Probability to Statistics

From Probability to Statistics

- ▶ By building from different mathematical theories, probability tells us how our outcomes will be distributed.

From Probability to Statistics

Statistics: Given the information in your hand, what is in the pail?

Probability: Given the information in the pail, what is in your hand?

- ▶ 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, $X_n = X_1, X_2, X_3, ..., X_n$, 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.

Statistics

Central Limit Theorem

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

Central Limit Theorem

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

As n approaches ∞, the distribution of $\bar{X}_n$ (mean of X) 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.

Statistics

Law of Large Numbers

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

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.

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 work with probabilities, expected values, variances, etc. in R this afternoon. You'll learn to determine if what you're seeing in the data is due to randomness or chance, or due to an actual effect tomorrow!

Questions?

Thanks!

Lula - nchen3@illinois.edu