Continuous Distributions

Lecture 9

Reading: Sections 9.8 – 9.11, "Normal Table: Read It, Use It" (Optional: 9.7, 9.12, "Normal Probability Plots" pp. 280-2)

1

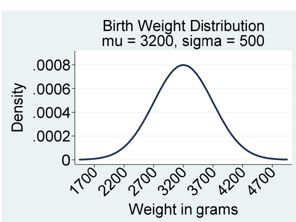
Discrete versus Continuous

- Discrete random variable: can find $P(X = x_1), \ldots, P(X = x_m)$
- Continuous random variable: probability that X equals a specific number is 0
  - E.g.: If measure weight *perfectly* at birth, $P(W = 3302g) = ?$
- Typically round weight to nearest gram
  - Technically a finite # of possible values
  - But treat as continuous: a good approximation
- Discrete vs. continuous:
  - # cars through a toll booth in a month?
  - # UoT snow days in a year?

2

Probability Density Function: $f(x)$

- Density function requirements:
  - $f(x) \geq 0$ for all x
  - Total area under $f(x)$ equals 1
- $f(x)$ is height, not a probability
- Area under a density function is probability of a range of values



$P(2700 < W < 3700) = 0.6827$

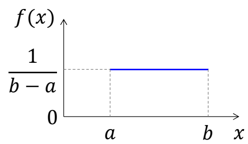
$P(W = 3302) = 0$

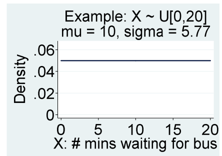
$P(3301.5 < W < 3302.5) = 0.0008$

3

Uniform Distribution: $X \sim U[a, b]$

- Uniform: $f(x) = \frac{1}{b-a}$ for $a < x < b$ and 0 otherwise
  - Two parameters: a and b
  - Bounded support
  - $E[X] = \mu = \frac{a+b}{2}$
  - $V[X] = \sigma^2 = \frac{(b-a)^2}{12}$
  - Can $f(x)$ be greater than one?



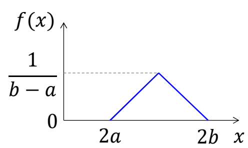


$P(12 < X \leq 14) = ?$

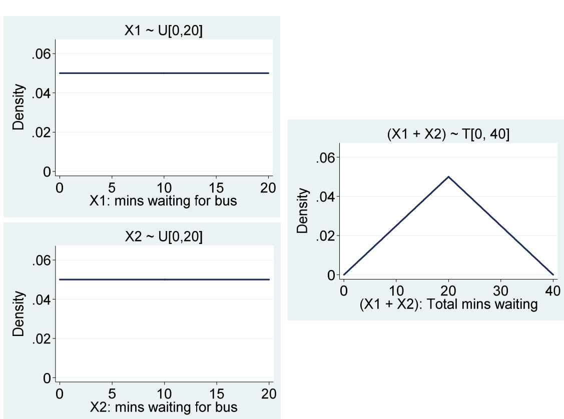
4

Triangle Distribution: $X \sim T[2a, 2b]$

- Triangle distribution created as sum of two independent and identically distributed Uniform random variables
  - X_1 and X_2 independent
  - $X_1 \sim U[a, b]$ and $X_2 \sim U[a, b]$ (identically distributed)
  - $(X_1 + X_2) \sim T[2a, 2b]$
  - E.g. If wait time is Uniformly distributed from 0 to 20 minutes, $U[0, 20]$, then total wait time for two buses $(X_1 + X_2)$ follows the Triangle distribution: $T[0, 40]$

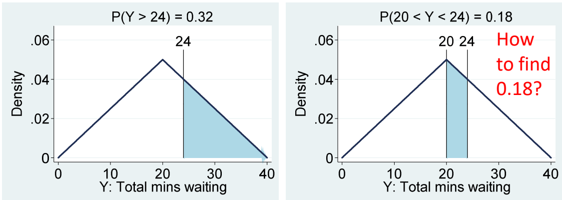


5



6

Finding Areas: $A = \frac{1}{2} base * height$



$y - y_1 = m(x - x_1)$

$y - 0 = -\frac{0.05}{20}(x - 40)$

$y = 0.1 - 0.0025x$

$A = 0.5 * (40 - 24) * (0.1 - 0.0025 * 24) = 0.32$

7

Summary: Uniform $X \sim U[a, b]$

- Symmetric, rectangle-shaped, even density throughout
- Parameters: a and b
  - Bounded support: $[a, b]$
- Find probabilities with $A = bh$ and symmetry
- $E[X] = \mu = \frac{a+b}{2}$
- $V[X] = \sigma^2 = \frac{(b-a)^2}{12}$

Summary: Triangle $X \sim T[2a, 2b]$

- Symmetric, triangle-shaped, more density around mean
- Parameters: $2a$ and $2b$
  - Bounded sup.: $[2a, 2b]$
- Find probabilities with $A = \frac{1}{2}bh$ and symmetry
- $E[X] =$
- $V[X] =$

8

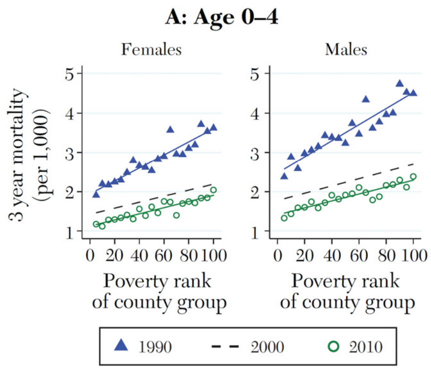
Percentiles & Uniform Distribution

- Percentiles are often interesting
  - Standardized tests: for example, a score of 590 on the math SAT is at the 73rd percentile (good)
  - Birth weights: for example, a 10-month old girl who weighs 6.8 kg is at the 3rd percentile (small)
- No matter how a variable's distribution is shaped – positively skewed, negatively skewed, bi-modal, Normal (Bell) – percentiles will be Uniformly distributed from 0 to 100

The WHO Child Growth Standards, Weight-for-age, http://www.who.int/childgrowth/standards/weight_for_age/en/

9

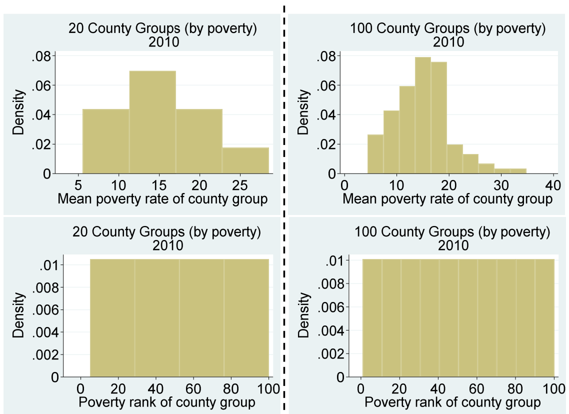
Recall: Currie & Schwandt (2016)



What is the meaning of the x-axis variable?

Excerpt: Figure 3 "Three-year mortality rates across groups of counties ranked by their poverty rate" (p. 41), which appears on page 46 of the DACM Handbook (Module B.1).

10

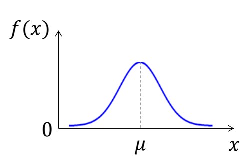


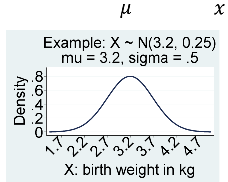
Dataset (right): mort_in_figure_a2.xlsx, Datasets page, DACM Quercus site

11

Normal Distribution: $X \sim N[\mu, \sigma^2]$

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ for $-\infty < x < \infty$
  - Parameters: μ and σ^2
  - Unbounded support
  - For area under $f(x)$, no formula (calculus cannot solve): use tables/software





12

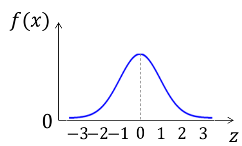
Recall: Empirical Rule

- In sample from Normal population, then <u>about</u>:
  - 68.3% of obs lie w/in 1 s.d. of mean (i.e. btwn $\bar{X} - s$ and $\bar{X} + s$)
  - 95.4% of obs lie w/in 2 s.d. of mean (i.e. btwn $\bar{X} - 2s$ and $\bar{X} + 2s$)
  - 99.7% of obs lie w/in 3 s.d. of mean (i.e. btwn $\bar{X} - 3s$ and $\bar{X} + 3s$)
- For the Normal model, <u>exactly</u> (with rounding):
  - 68.3% of obs lie w/in 1 s.d. of mean (i.e. btwn $\mu - \sigma$ and $\mu + \sigma$)
  - 95.4% of obs lie w/in 2 s.d. of mean (i.e. btwn $\mu - 2\sigma$ and $\mu + 2\sigma$)
  - 99.7% of obs lie w/in 3 s.d. of mean (i.e. btwn $\mu - 3\sigma$ and $\mu + 3\sigma$)

13

Standard Normal: $Z \sim N(0, 1)$

- If $X \sim N(\mu, \sigma^2)$ then $Z \sim N(0, 1)$ where $Z = \frac{X - \mu}{\sigma} = -\frac{\mu}{\sigma} + \frac{1}{\sigma}X$
  - Recall that linear transformations do NOT change shape & Laws of Expectation & Variance:



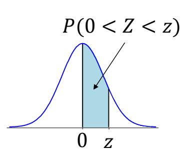
$$E\left[-\frac{\mu}{\sigma} + \frac{1}{\sigma}X\right] = -\frac{\mu}{\sigma} + \frac{1}{\sigma}E[\mathrm{X}] = -\frac{\mu}{\sigma} + \frac{\mu}{\sigma} = 0$$

$$V\left[-\frac{\mu}{\sigma} + \frac{1}{\sigma}X\right] = \frac{1}{\sigma^2}V[\mathrm{X}] = \frac{1}{\sigma^2}\sigma^2 = 1$$

14

Using Standard Normal Table

$$P(x_1 < X < x_2) = P\left(\frac{x_1 - \mu}{\sigma} < Z < \frac{x_2 - \mu}{\sigma}\right)$$



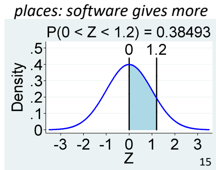
For example, if birth weights are Normally distributed with a mean of 3,200 grams and a standard deviation of 500 grams, what is the probability that a randomly selected newborn weighs less than 2,600 grams?

$$P(X < 2{,}600) = P\left(Z < \frac{2{,}600 - 3{,}200}{500}\right)$$

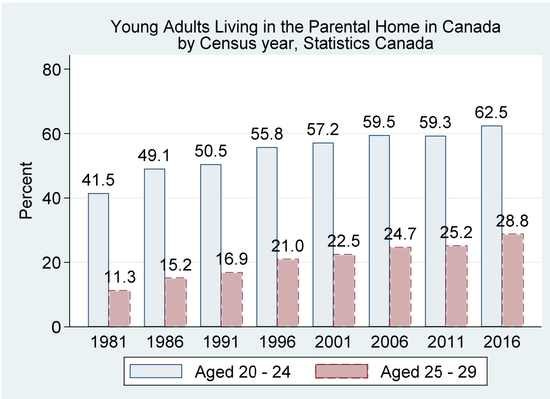
$$P(Z < -1.2) = 0.5 - 0.3849$$
$$= 0.1151$$

"Normal Table: Read it, Use it": Use one-page table for HW

Table rounds to 4 decimal places: software gives more



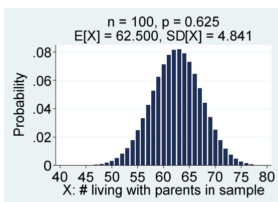
15



Statistics Canada, https://www150.statcan.gc.ca/n1/pub/75-006-x/2016001/article/14639-eng.htm; 2016 from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/fam/Table.cfm?Lang=E&T=32&Geo=00 16

Sampling (Sampling Error) Example

- In random sample of 100 young adults aged 20-24 in 2016, what is the chance that *56 or fewer* live with parents (recalling that Census says 62.5%)?
  - Define X as count of those living with parents
  - How is X distributed?



$P(X \leq 56) = P(X < 57) = ?$

Reasonable to compute with handheld calculator?

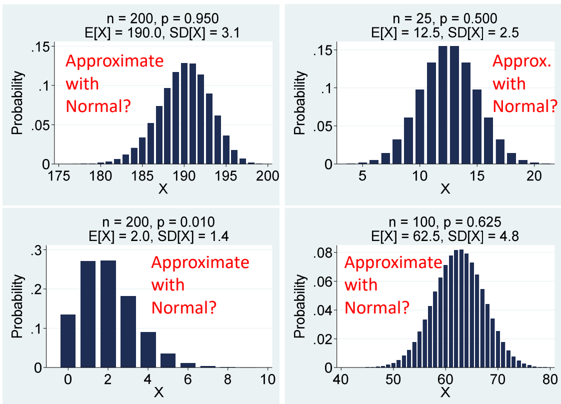
$$P(X = 56) = \frac{100!}{56!(100-56)!} 0.625^{56}(1 - 0.625)^{100-56} = 0.0331$$

17

Normal Approximation to Binomial

- $X \sim B(n, p)$ is *sometimes* well approximated by $X \sim N(\mu, \sigma^2)$ w/ $\mu = np$ and $\sigma^2 = np(1-p)$
  - Two common rules of thumb (you may use either) to check if it's a reasonable approximation:
    1) $np \geq 10$ and $n(1-p) \geq 10$
    2) $np - 3\sqrt{np(1-p)} \geq 0$ and $np + 3\sqrt{np(1-p)} \leq n$
- As n rises, Normal approximation improves
  - There is *nothing* magic about marginally passing or marginally failing either rule of thumb

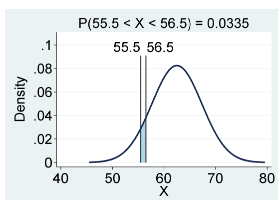
18



19

Continuity Correction

- Binomial is *discrete*
  - $P(X = 56) \neq 0$ but $P(56 < X < 57) = 0$
- Normal is *continuous*
  - $P(X = 56) = 0$ but $P(56 < X < 57) \neq 0$
- "Continuity correction" addresses this disparity
  - Page 286 textbook (box)

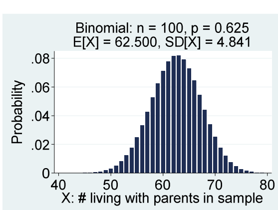


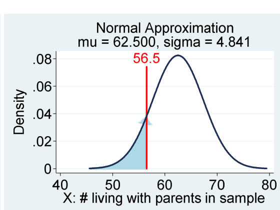
$P(X = 56) \approx 0.0335$

To approximate $P(X \leq 56)$, use Normal to find $P(X < 56.5)$

Note: Correction (slightly) improves the approximation. With a large n it makes little difference: researchers usually ignore it.

20





$P(X \leq 56) \approx P(X < 56.5)$

$$P(X < 56.5) = P\left(Z < \frac{56.5 - 62.5}{4.841}\right) = P(Z < -1.24) = 0.5 - 0.3925 = 0.1075$$

Is having so few (56) living at home in a sample of 100 surprising?

21

Normal + Normal = Normal?

- Linear combination: $a_0 + a_1X_1 + \cdots + a_JX_J$ where $a_0, a_1, \ldots, a_J$ are constants
- A linear combination of independent Normal random variables yields a Normal random variable

In 2016, chance a random sample of 50 people aged 25-29 has *more* living w/ parents than sample of 50 aged 20-24?

$P(X_{25-29} > X_{20-24}) = ?$

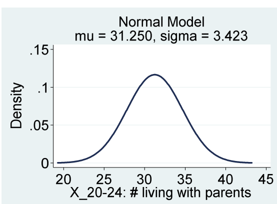
$P\big((X_{25-29} - X_{20-24}) > 0\big) = ?$

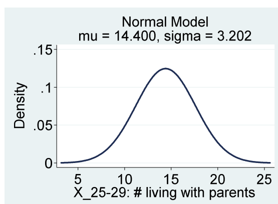
$X_{20-24} \sim B(n = 50, p = 0.625)$

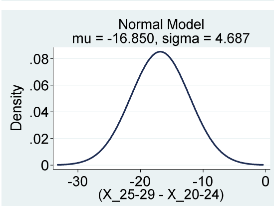
$X_{25-29} \sim B(n = 50, p = 0.288)$

Use Normal approx. for each?

If so, is difference Normal?







$P\big((X_{25-29} - X_{20-24}) > 0\big) = ?$

$P\left(Z > \frac{0 - -16.850}{4.687}\right)$

$= P(Z > 3.60) = 0.0002$

Would it be surprising if X_{25-29} were larger than X_{20-24}?