Continuous Probability Distributions

Isaac D. Mehlhaff

September 11, 2024

Logistics

Problem set 1 due this Friday at 1:50 PM

- Submit two things: typed answers (Word doc, PDF, etc.) and R script

Capstone project component 2 due next Friday

Review

Probability distribution:

Review

Probability distribution: statistical function describing probability of observing each possible value of RV

Three rules of probability distributions:

Review

Probability distribution: statistical function describing probability of observing each possible value of RV

Three rules of probability distributions:

- Outcomes must be independent
- $0 \leq Pr(x) \leq 1$
- $1 = \sum_x f(x)$ for discrete, $1 = \int_{min(x)}^{max(x)} f(x)$ for continuous

Bernoulli distribution:

Review

Probability distribution: statistical function describing probability of observing each possible value of RV

Three rules of probability distributions:

- Outcomes must be independent
- $0 \leq Pr(x) \leq 1$
- $1 = \sum_x f(x)$ for discrete, $1 = \int_{min(x)}^{max(x)} f(x)$ for continuous

Bernoulli distribution: probability distribution of a binary variable $(n = 1)$

Binomial distribution:

Review

Probability distribution: statistical function describing probability of observing each possible value of RV

Three rules of probability distributions:

- Outcomes must be independent
- $0 \leq Pr(x) \leq 1$
- $1 = \sum_x f(x)$ for discrete, $1 = \int_{min(x)}^{max(x)} f(x)$ for continuous

Bernoulli distribution: probability distribution of a binary variable $(n = 1)$

Binomial distribution: probability distribution of number of successes in n independent trials

Continuous Probability Distributions

Normal distribution: symmetric, bell-shaped, unimodal probability distribution of a continuous variable

- Takes two parameters: $X \sim \mathrm{N}(\mu, \sigma)$

Continuous Probability Distributions

Normal distribution: symmetric, bell-shaped, unimodal probability distribution of a continuous variable

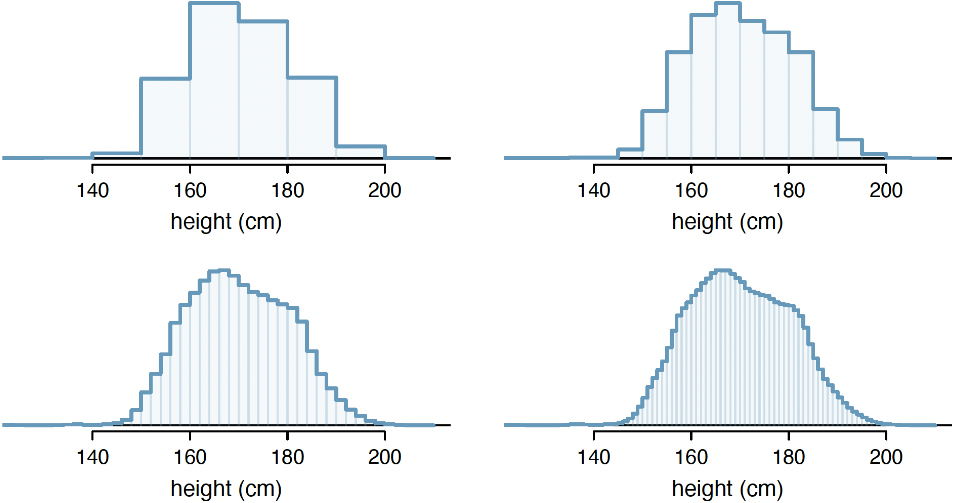
- Takes two parameters: $X \sim \mathrm{N}(\mu, \sigma)$

Lots of things are (approximately) normally distributed:

- Height
- Birth weight
- ACT/SAT scores
- Retirement age of NFL players

Many statistical methods require assumption of normality

Continuous Probability Distributions



Probability Density Function

TEXAS A&M UNIVERSITY

For a continuous RV, the **probability density function** (PDF) assigns probabilities to ranges of outcomes:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Density $\neq$ probability; $Pr(X = x) = 0$

Intuition: pick any real number X between $-\infty$ and ∞.

Probability Density Function

For a continuous RV, the **probability density function** (PDF) assigns probabilities to ranges of outcomes:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Density $\neq$ probability; $Pr(X = x) = 0$

Intuition: pick any real number X between $-\infty$ and ∞.
How many chose $X = 1048.23328456$?

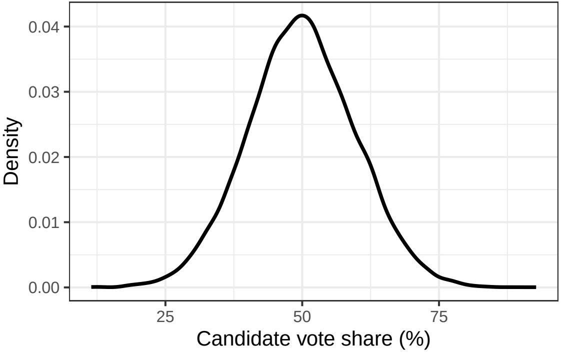
Probability Density Function

Suppose this illustrates the random process generating vote share

Outcomes around 50% most likely, but probability of 50% exactly is zero

Density gives likelihood of being in the neighborhood of X

Need to know area under the curve: **integral**



Probability Density Function

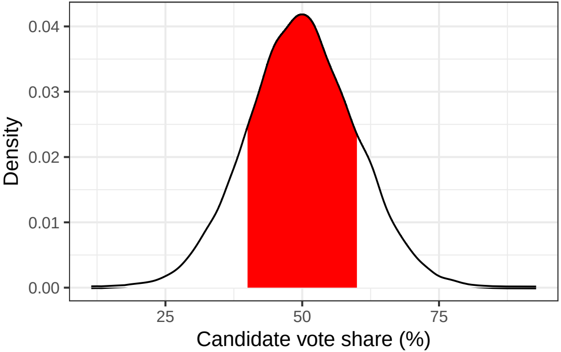
What is the probability of a candidate getting 40-60% of the vote?

Looking for area of red region

This is given by $\int_{40}^{60} f(x)$—
not reasonable to do by hand

In R: `pnorm()` gives area under normal curve **to the left** of a specified value

```
pnorm(x = 60, mean = 50, sd = 10) -
pnorm(x = 40, mean = 50, sd = 10)
```
$= 0.683$

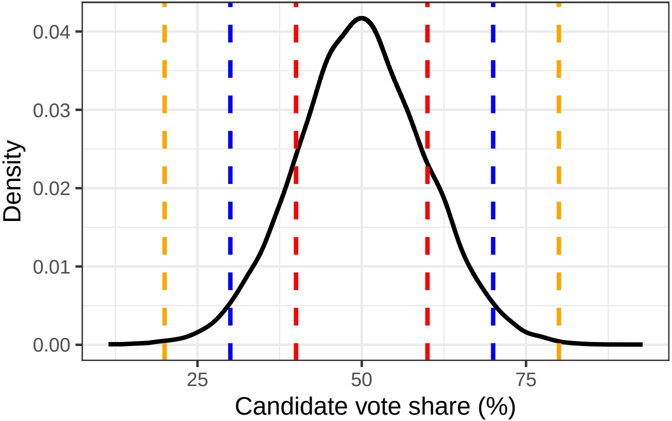


68-95-99 Rule

Not a coincidence we got 68% on the last slide

Property of normal distributions:

- Approx. 68% of obs. fall within 1 SD of mean
- Approx. 95% fall within 2 SD
- Approx. 99.7% fall within 3 SD
- Any real number possible, but extremely unlikely beyond 3 SD



Standard Normal Distribution

Special case of the normal distribution with $\mu = 0$ and $\sigma = 1$

Useful for characterizing uncertainty and comparing across variables with different distributions

Standard Normal Distribution

Special case of the normal distribution with $\mu = 0$ and $\sigma = 1$

Useful for characterizing uncertainty and comparing across variables with different distributions

Typically denote standard normal RV as Z such that $Z \sim \mathrm{N}(0, 1)$

Also has special notation for its PDF and CDF

- Probability density function (PDF): $\phi(z)$
- Cumulative density function (CDF): $\Phi(z)$ (gives the percentile)

Standard Normal Distribution

Special case of the normal distribution with $\mu = 0$ and $\sigma = 1$

Useful for characterizing uncertainty and comparing across variables with different distributions

Typically denote standard normal RV as Z such that $Z \sim \mathrm{N}(0, 1)$

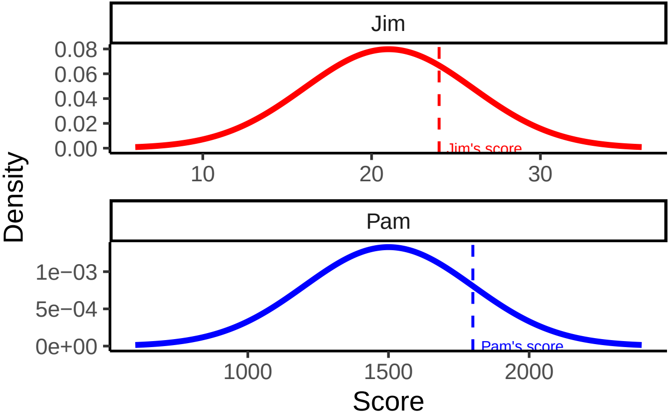
Also has special notation for its PDF and CDF

- Probability density function (PDF): $\phi(z)$
- Cumulative density function (CDF): $\Phi(z)$ (gives the percentile)

For any RV $X \sim \mathrm{N}(\mu, \sigma)$:

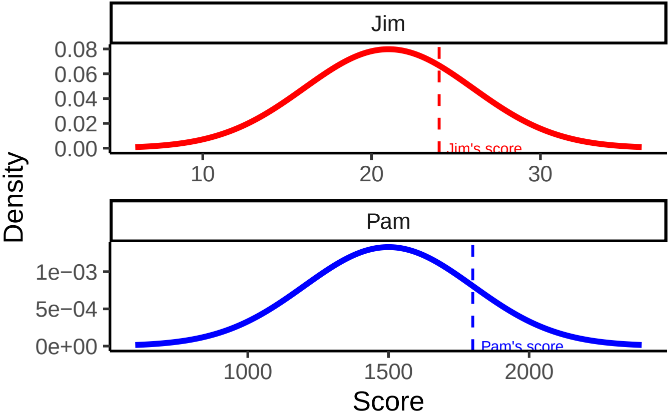
$$\frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1)$$

Standardization



Pam and Jim are applying to Texas A&M. Pam got 1800 on the SAT and Jim got 24 on the ACT. Who has a more impressive test score?

Standardization



Pam and Jim are applying to Texas A&M. Pam got 1800 on the SAT and Jim got 24 on the ACT. Who has a more impressive test score?

Can't compare the two raw scores. Need info on distribution of each test:

- SAT: $\mu = 1500$, $\sigma = 300$
- ACT: $\mu = 21$, $\sigma = 5$

Lecture Exercise

TEXAS A&M UNIVERSITY

Distribution of each test:

- SAT: $\mu = 1500$, $\sigma = 300$
- ACT: $\mu = 21$, $\sigma = 5$

Convert Pam and Jim's scores into a standardized score ("Z-score") using the formula:

$$\frac{X - \mu}{\sigma}$$

How do you interpret their standardized scores? Who has a more impressive score?

Lecture Exercise

Distribution of each test:

- SAT: $\mu = 1500$, $\sigma = 300$
- ACT: $\mu = 21$, $\sigma = 5$

Pam's standardized score:

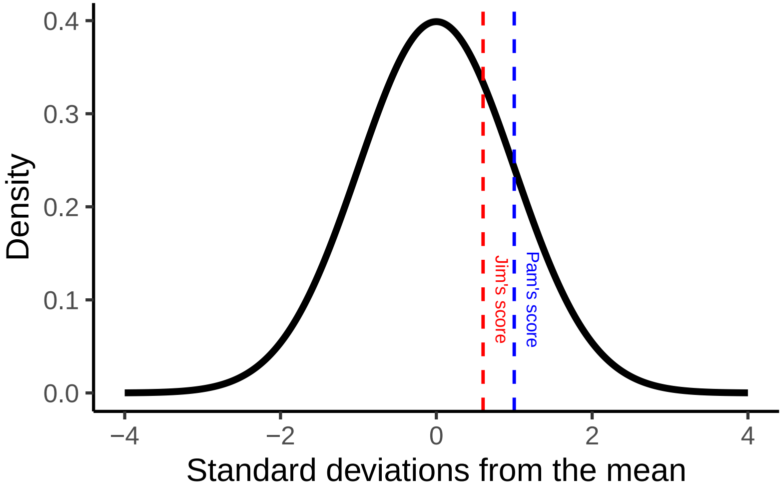
$$\frac{X - \mu}{\sigma} = \frac{1800 - 1500}{300} = 1$$

Jim's standardized score:

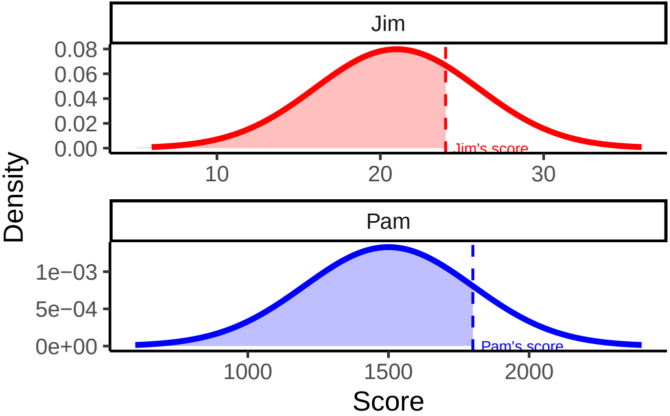
$$\frac{X - \mu}{\sigma} = \frac{24 - 21}{5} = 0.6$$

Pam scored 1 SD above the mean. Jim scored 0.6 SD above the mean $\rightarrow$ Pam has a more impressive test score

Standardization



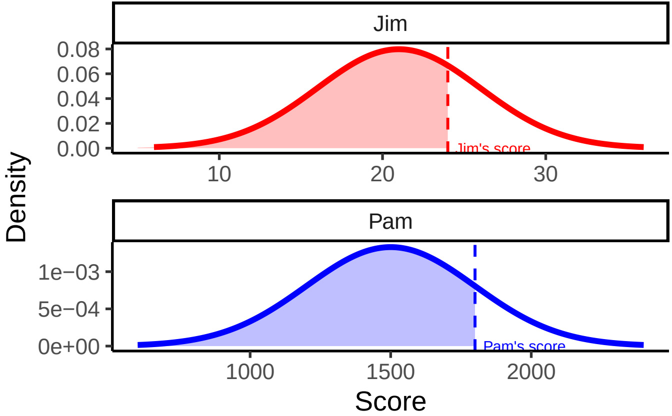
Percentiles



How many test-takers got scores **worse** than Pam and Jim? i.e. In what **percentile** are Pam and Jim among all test-takers?

Straightforward with standard normal CDF: $\Phi(z) = Pr(Z < z)$

- Note the $<$ instead of $\leq$ in the continuous CDF—why?

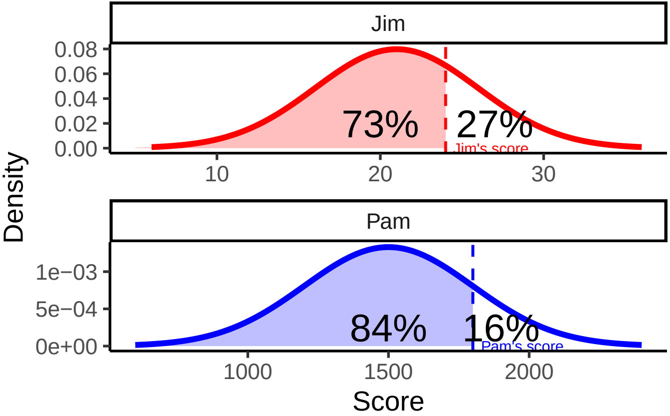


How many test-takers got scores **worse** than Pam and Jim? i.e. In what **percentile** are Pam and Jim among all test-takers?

Straightforward with standard normal CDF: $\Phi(z) = Pr(Z < z)$

- Note the $<$ instead of $\leq$ in the continuous CDF—why?
- In a continuous distribution, $Pr(Z = z) = 0$

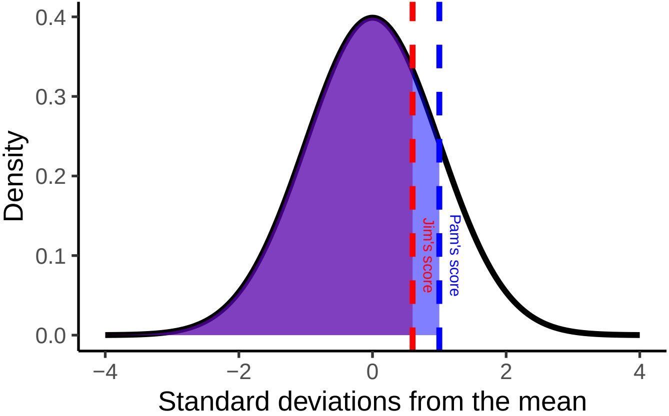
Percentiles



In R:

- `pnorm(q = 1800, mean = 1500, sd = 300)` = 0.841
- `pnorm(q = 24, mean = 21, sd = 5)` = 0.726

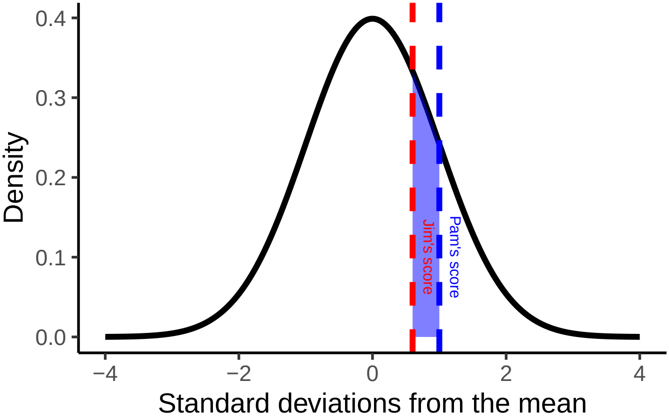
Percentiles



Can also do this with **standardized** scores:

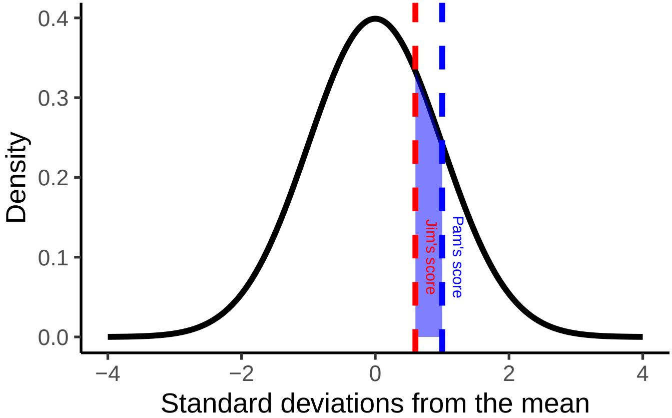
- `pnorm(q = 1)` $= 0.841$
- `pnorm(q = 0.6)` $= 0.726$
- `pnorm()` defaults to standard normal

Percentiles



How would we find the percentage of test-takers with scores **between** Pam and Jim?

Percentiles



How would we find the percentage of test-takers with scores **between** Pam and Jim?

Still using standard normal CDF:

$$\Phi(1) - \Phi(0.6)$$

In R:

- `pnorm(q = 1) - pnorm(q = 0.6)` $= 0.116$

Inverse CDF

We just found the **percentage** below Pam and Jim, given their scores. What if we wanted to know the **score** z, given the percentage of scores below it?

- What is the 95th percentile of scores?
- What is z such that $Pr(Z < z) = 0.95$?

Inverse CDF

We just found the **percentage** below Pam and Jim, given their scores. What if we wanted to know the **score** z, given the percentage of scores below it?

- What is the 95th percentile of scores?
- What is z such that $Pr(Z < z) = 0.95$?

Use inverse CDF: $\Phi^{-1}(p) = z$ for some percentile p (no formula for this)

In R:

- For SAT score: `qnorm(p = 0.95, mean = 1500, sd = 300)` $=$ 1993.456
- For ACT score: `qnorm(p = 0.95, mean = 21, sd = 5)` $=$ 29.224
- For Z-score: `qnorm(p = 0.95)` $=$ 1.645

Inverse CDF

We just found the **percentage** below Pam and Jim, given their scores. What if we wanted to know the **score** z, given the percentage of scores below it?

- What is the 95th percentile of scores?
- What is z such that $Pr(Z < z) = 0.95$?

Use inverse CDF: $\Phi^{-1}(p) = z$ for some percentile p (no formula for this)

In R:

- For SAT score: `qnorm(p = 0.95, mean = 1500, sd = 300)` $= 1993.456$
- For ACT score: `qnorm(p = 0.95, mean = 21, sd = 5)` $= 29.224$
- For Z-score: `qnorm(p = 0.95)` $= 1.645$
- What will `pnorm(qnorm(p = 0.95))` return?

Inverse CDF

We just found the **percentage** below Pam and Jim, given their scores. What if we wanted to know the **score** z, given the percentage of scores below it?

- What is the 95th percentile of scores?
- What is z such that $Pr(Z < z) = 0.95$?

Use inverse CDF: $\Phi^{-1}(p) = z$ for some percentile p (no formula for this)

In R:

- For SAT score: `qnorm(p = 0.95, mean = 1500, sd = 300)` $= 1993.456$
- For ACT score: `qnorm(p = 0.95, mean = 21, sd = 5)` $= 29.224$
- For Z-score: `qnorm(p = 0.95)` $= 1.645$
- What will `pnorm(qnorm(p = 0.95))` return? **0.95**