

STAT 22000 Lecture Slides

Continuous Distributions and Normal Distributions

Yibi Huang
Department of Statistics
University of Chicago

Coverage: Section 2.5 and 3.1 of OpenIntro Statistics (3ed)

- Continuous distribution (2.5)
- Normal distribution (3.1)

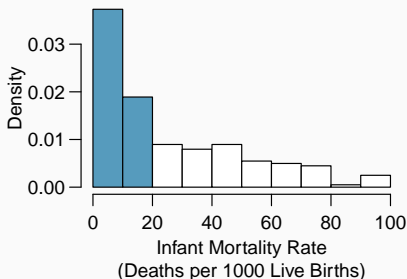
Please skip section 3.2

Continuous distributions

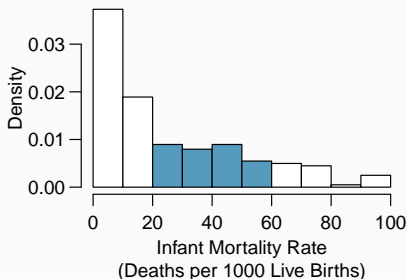
Recall: Area Under a Histogram = Proportion

Recall in L01, we introduced the density scale of a histogram,

Blue *area* = *proportion* of
countries/regions with
 $0 \leq \text{IMR} \leq 20$

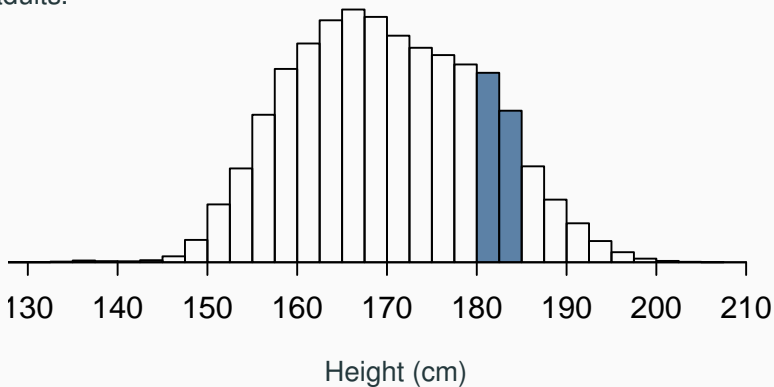


Blue *area* = *proportion* of
countries/regions with
 $20 \leq \text{IMR} \leq 60$



In the density scale, the total area under a histogram is 1 (why?).

Below is a histogram showing the distribution of the heights of US adults.



The percentage of US adults that are 180-185 cm tall is closest to which of the following?

10%

25%

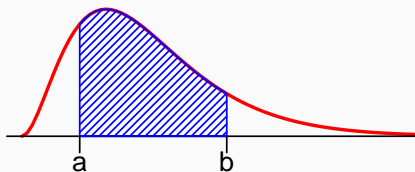
50%

75%

Continuous Random Variables & Density Curves

The probability distribution of a continuous random variable is described by a **density curve**, which can be regarded as a highly smoothed histogram.

If Y is a continuous random variable, $P(a < Y < b)$ is the area under the density curve of Y above the interval between a and b

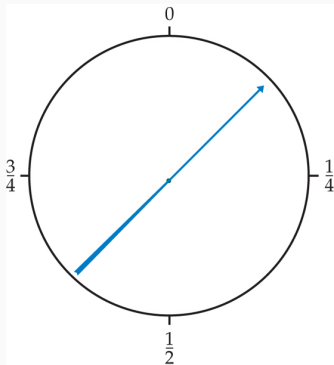


- Note: all continuous probability distributions assign zero probability to every individual outcome: $P(Y = y) = 0$

Example — Spinner

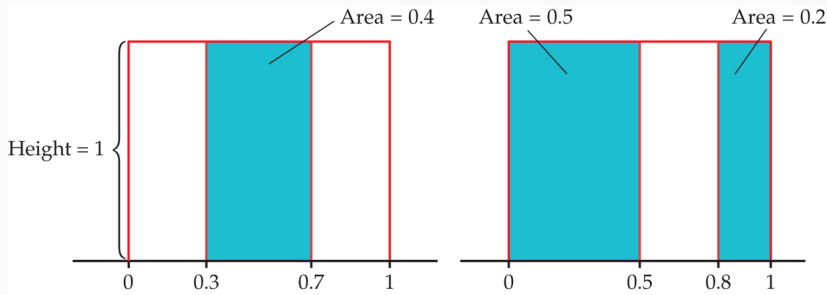
A spinner turns freely on its axis and slowly comes to a stop.

- Define a random variable X as the location of the pointer when the spinner stops. It can be anywhere on a circle that is marked from 0 to 1.
- Sample space $S = \{ \text{all numbers } x \text{ such that } 0 \leq x < 1 \}$
- $P(0.3 < X < 0.7) = ?$
- $P(X < 0.5 \text{ or } X > 0.8) = ?$
- $P(X = 0.75) = ?$



Density Curve for the Spinner Example

For the spinner example, the density curve for X is constant at 1 on the interval $[0, 1]$, and 0 elsewhere.



$$P(0.3 < X < 0.7) = 0.4$$

$$P(X < 0.5 \text{ or } X > 0.8) = 0.7$$

Expected Value (=Mean) and Variance for a Continuous Random Variable

If X is a **continuous** random variable with density curve $f(x)$, the *expected value* or the *mean* of X is defined as the integral

$$\mu_X = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

The variance of X is defined as the integral

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)dx$$

in which μ_X is the mean of X .

The SD of X is the square root of the variance:

$$\sigma_X = SD(X) = \sqrt{V(X)}.$$

Example — Spinner

The density of X is a constant 1 on $[0,1]$ and 0 elsewhere

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x < 0 \text{ or } x > 1 \end{cases}$$

The mean of X is

$$\mu_X = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 x \cdot 1 dx = \frac{1}{2}x^2 \Big|_0^1 = \frac{1}{2},$$

the variance is

$$V(X) = \int_{-\infty}^{\infty} (x - \frac{1}{2})^2 f(x) dx = \int_0^1 (x - \frac{1}{2})^2 \cdot 1 dx = \frac{1}{3}(x - \frac{1}{2})^3 \Big|_0^1 = \frac{1}{12}.$$

The SD is

$$SD(X) = \sqrt{V(X)} = \sqrt{1/12} \approx 0.289.$$

In STAT 220, you will NEVER have to do integration to find probabilities or expected values or variances.

Normal distribution

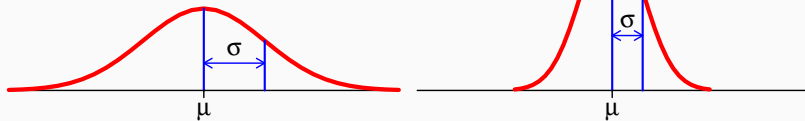
Normal Distributions

Normal distributions (aka. Gaussian distributions) are a family of *symmetric*, *bell-shaped* density curves defined by

a mean μ , and an SD σ

denoted as $N(\mu, \sigma)$. The formula for the $N(\mu, \sigma)$ curve is

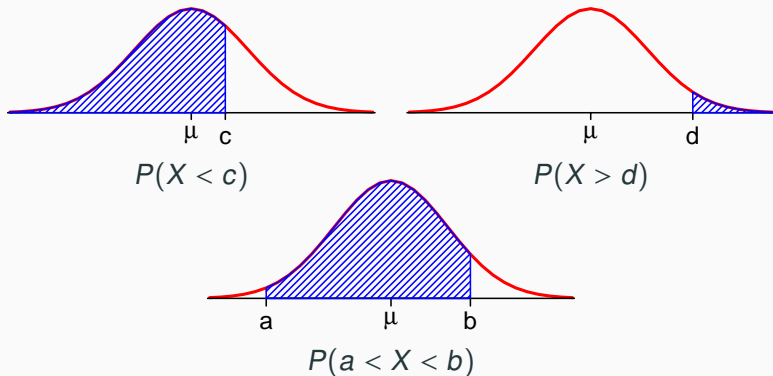
$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}.$$



A normal distribution with $\mu = 0$, and $\sigma = 1$ is called the *standard normal distribution*, denoted as $N(0, 1)$.

Normal Probabilities

If X has a normal distribution, then to find probabilities about X is to find **areas** under a normal curve $N(\mu, \sigma)$.



But,... there is no simple formula for areas under a Normal curve.
Must use **softwares (e.g. R)** or the **normal probability table**.

Find Normal Probabilities in R

The R command `pnorm()` can find the lower-tail area under the standard normal $N(0, 1)$ curve.

$P(Z < z) = \text{pnorm}(z)$ = area of shaded region in



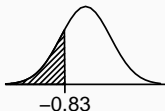
```
> pnorm(-0.83)
```

```
[1] 0.2032694
```

```
> pnorm(1.573)
```

```
[1] 0.9421406
```

$P(Z < -0.83) = \text{pnorm}(-0.83) =$



$= 0.2032694$

$P(Z < 1.573) = \text{pnorm}(1.573) =$



$= 0.9421406$

Finding Upper Tail Probabilities

$$\begin{aligned} P(Z > -0.83) &= \text{[Diagram 1]} = \text{[Diagram 2]} - \text{[Diagram 3]} \\ &= 1 - \text{pnorm}(-0.83) \\ &= 1 - 0.2032694 = 0.7967306 \end{aligned}$$

The diagrams illustrate the calculation of the upper tail probability $P(Z > -0.83)$. Diagram 1 shows a standard normal distribution curve with the area to the right of -0.83 shaded. Diagram 2 shows the entire area under the curve shaded. Diagram 3 shows the area to the left of -0.83 shaded. The equation shows that the area to the right of -0.83 is equal to the total area (1) minus the area to the left of -0.83 ($\text{pnorm}(-0.83)$).

```
> 1 - pnorm(-0.83)
[1] 0.7967306
```

Or one find the upper tail probability directly by specifying `lower.tail=FALSE`

```
> pnorm(-0.83, lower.tail=FALSE)
[1] 0.7967306
```


$$\begin{aligned}
 P(-0.83 < Z < 2) &= \text{[Graph of } Z \text{ between } -0.83 \text{ and } 2\text{]} = \text{[Graph of } Z < 2\text{]} - \text{[Graph of } Z < -0.83\text{]} \\
 &= P(Z < 2) - P(Z < -0.83) \\
 &= \text{pnorm}(2) - \text{pnorm}(-0.83) \\
 &= 0.9772499 - 0.2032694 \\
 &= 0.7739805
 \end{aligned}$$

```

> pnorm(2)
[1] 0.9772499
> pnorm(-0.83)
[1] 0.2032694
> pnorm(2) - pnorm(-0.83)
[1] 0.7739805

```

Finding the z for a Given (Lower-Tail) Probability

E.g, we want to find the first quartile (Q_1) of the standard normal, i.e., the z such that

$$P(Z < z) = \text{shaded area in } \img alt="A standard normal distribution curve with the area to the left of a vertical line at z=? shaded with diagonal lines." data-bbox="531 301 716 463"/> = 0.25?$$

The R command `qnorm()` is for finding the z such that $P(Z < z)$ equals a specific probability

```
> qnorm(0.25)
[1] -0.6744898
```

So the first quartile of the standard normal is about -0.6744898 .

Quartiles of the Standard Normal Distribution

The quartiles of the standard normal distributions are:

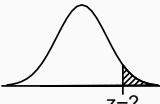
$$Q_1 \approx -0.6745 \quad \dots\dots \text{(from the previous slide)}$$

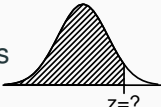
$$Q_2 = 0 \quad \dots\dots \text{(why?)}$$

$$Q_3 = -Q_1 \approx 0.6745 \quad \dots\dots \text{(why?)}$$

The interquartile range (IQR) for the standard normal curve is

$$\text{IQR} = Q_3 - Q_1 \approx 0.6745 - (-0.6745) \approx 1.349 \approx 1.35$$

If $P(Z > z) =$  $= 0.05$, then $z = ?$

This implies  $= 0.95$.

So $z = \text{qnorm}(1-0.05) \approx 1.644854$.

```
> qnorm(1-0.05)
```

```
[1] 1.644854
```

```
> qnorm(0.05, lower.tail=F) # alternative way
```

```
[1] 1.644854
```

Finding Probabilities for General Normal Distributions

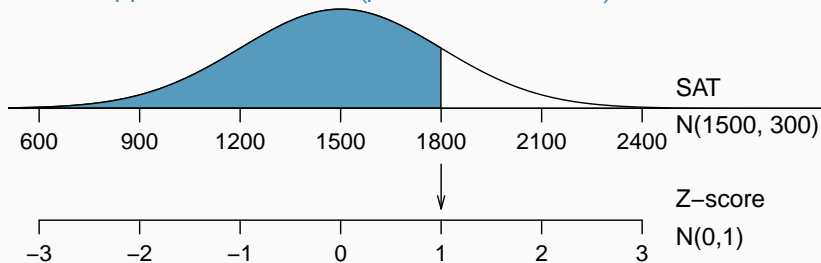
Now we've learned how to find probabilities about the standard normal $N(0, 1)$. To compute probability about general normal distribution $N(\mu, \sigma)$, we need to use the property

$$\text{if } X \sim N(\mu, \sigma), \text{ then } Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

The $Z = \frac{X - \mu}{\sigma}$ is called the *Z-score*, the *standardized value*, or the *standardized score*.

Example: Calculating Normal Probabilities

Approximately what percent of students score below 1800 on the SAT? Suppose that $\text{SAT} \sim N(\mu = 1500, \sigma = 300)$



$$\begin{aligned} P(X < 1800) &= P\left(\underbrace{\frac{X - 1500}{300}}_{=Z} < \underbrace{\frac{1800 - 1500}{300}}_{=1}\right) = P(Z < 1) \\ &= \text{pnorm}(1) \\ &\approx 0.8413447. \end{aligned}$$

So about 84% of students score below 1800 on the SAT.

The R command `pnorm()` can calculate probabilities for general normal distributions without converting to z-scores.

```
> pnorm(1)
[1] 0.8413447
```

```
> pnorm(1800, mean = 1500, sd = 300) # Alternatively
[1] 0.8413447
```

```
> pnorm(1800, m = 1500, s = 300)
[1] 0.8413447
```

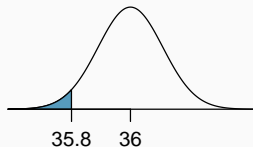
```
> pnorm(1800, 1500, 300)
[1] 0.8413447
```

Ketchup Quality Control

At a ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and SD 0.11 oz. What percent of bottles have less than 35.8 ounces of ketchup?

Let X = amount of ketchup in a bottle

$$X \sim N(\mu = 36, \sigma = 0.11)$$



$$\begin{aligned} P(X < 35.8) &= P\left(\underbrace{\frac{X - 36}{0.11}}_{=Z} < \frac{35.8 - 36}{0.11}\right) = P(Z < \frac{35.8 - 36}{0.11}) \\ &= \text{pnorm}((35.8 - 36)/0.11) \\ &\approx 0.03451817 \end{aligned}$$

```
> pnorm((35.8-36)/0.11)
```

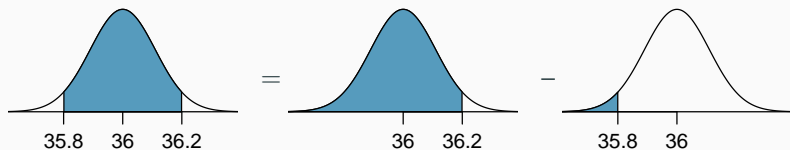
```
[1] 0.03451817
```

```
> pnorm(35.8, mean = 36, sd = 0.11)
```

```
[1] 0.03451817
```

Answer: About 3.45%

What percent of bottles have between 35.8 oz. and 36.2 oz. of ketchup, which meet quality control requirement?



$$\begin{aligned}P(35.8 < X < 36.2) &= P(X < 36.2) - P(X < 35.8) \\&= P\left(Z < \frac{36.2 - 36}{0.11}\right) - P\left(Z < \frac{35.8 - 36}{0.11}\right) \\&= \text{pnorm}((36.2-36)/0.11) - \text{pnorm}((35.8-36)/0.11) \\&= 0.9309637\end{aligned}$$

Answer: About 93.1%.

```
> pnorm((36.2-36)/0.11)-pnorm((35.8-36)/0.11)
[1] 0.9309637
> pnorm(36.2, m = 36, s = 0.11)-pnorm(35.8, m = 36, s = 0.11)
[1] 0.9309637
```

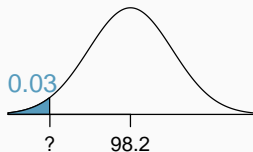
Finding the Cutoff Point For A Percentile

Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and SD 0.73°F . What is the cutoff for the lowest 3% of human body temperatures?

Let $X = \text{Body Temp. in degree F}$

$$X \sim N(\mu = 98.2, \sigma = 0.73)$$

$$\Rightarrow Z = \frac{X - 98.2}{0.73} \sim N(0, 1)$$



The z such that $P(Z < z) = 0.03$ is `qnorm(0.03)` = -1.880794 .

$$z = \frac{x - 98.2}{0.73} \approx -1.88 \quad \Rightarrow \quad x \approx (-1.88 \times 0.73) + 98.2 \approx 96.8$$

```
> qnorm(0.03)*0.73+98.2
```

```
[1] 96.82702
```

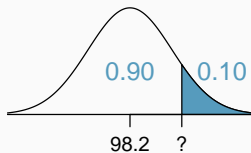
```
> qnorm(0.03, m = 98.2, s = 0.73)
```

```
[1] 96.82702
```

Answer: 96.8°F

Exercise

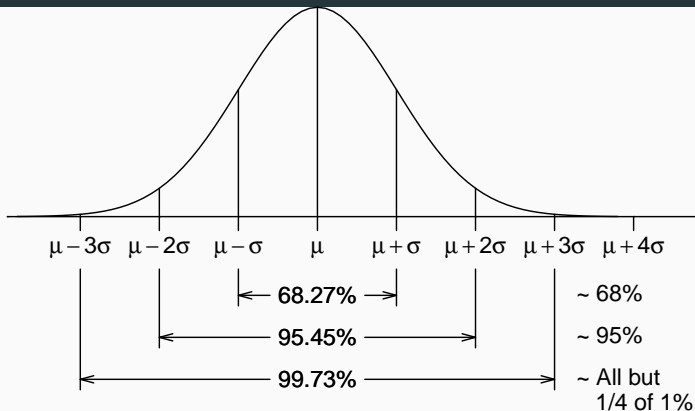
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F . What is the cutoff for the highest 10% of human body temperatures?



```
> qnorm(0.9, m = 98.2, s = 0.73)  
[1] 99.13553
```

Answer: 99.1°F

68-95-99.7% Rule for Normal Distributions



```
> pnorm(1) - pnorm(-1)
```

```
[1] 0.6826895
```

```
> pnorm(2) - pnorm(-2)
```

```
[1] 0.9544997
```

```
> pnorm(3) - pnorm(-3)
```

```
[1] 0.9973002
```

Other Application of Z-scores

- Z scores = $\frac{X - \mu}{\sigma}$ can also be defined if X has some other distribution, not normal, but only when X is normal can we use Z scores to calculate normal probabilities.
- Z score of an observation is *the number of SDs it falls above or below the mean.*

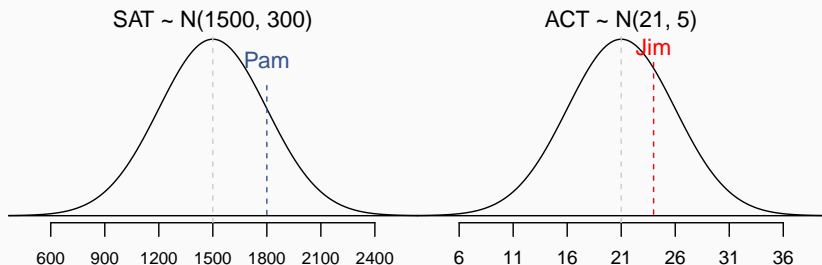
$$Z = \frac{\text{observation} - \text{mean}}{SD}$$

for whatever distributions

- Hence the Z score can be used as a measure of how extreme an observation is. Observations that are more than 3 SD away from the mean ($|Z| > 3$) are usually considered unusual.

Other Applications of the Z-score (ACT v.s. SAT)

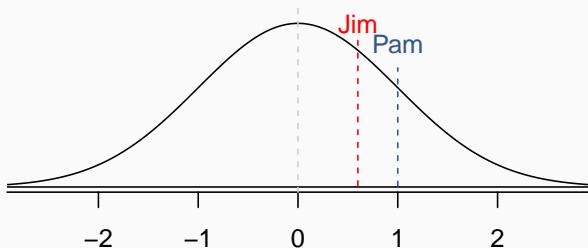
SAT scores are distributed nearly normally with mean 1500 and SD 300. ACT scores are distributed nearly normally with mean 21 and SD 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?



Other Applications of the Z-score (ACT v.s. SAT)

Since we cannot just compare the two raw scores, we instead compare how many SDs beyond the mean each observation is, i.e., we compare the Z scores.

- Pam's score is $\frac{1800 - 1500}{300} = 1$ SD above the mean.
- Jim's score is $\frac{24 - 21}{5} = 0.6$ SDs above the mean.



Recap: Ways to Detect Outliers

- 1.5 IQR rule
- Observations with $|Z\text{-scores}| > 3$ (or sometimes > 2)
- Histograms
- Scatterplots