Sta 111 - Summer II 2017
Probability and Statistical Inference

5. Normal distribution

Lu Wang

Duke University, Department of Statistical Science

July 6, 2017

Outline

1. Continuous random variable

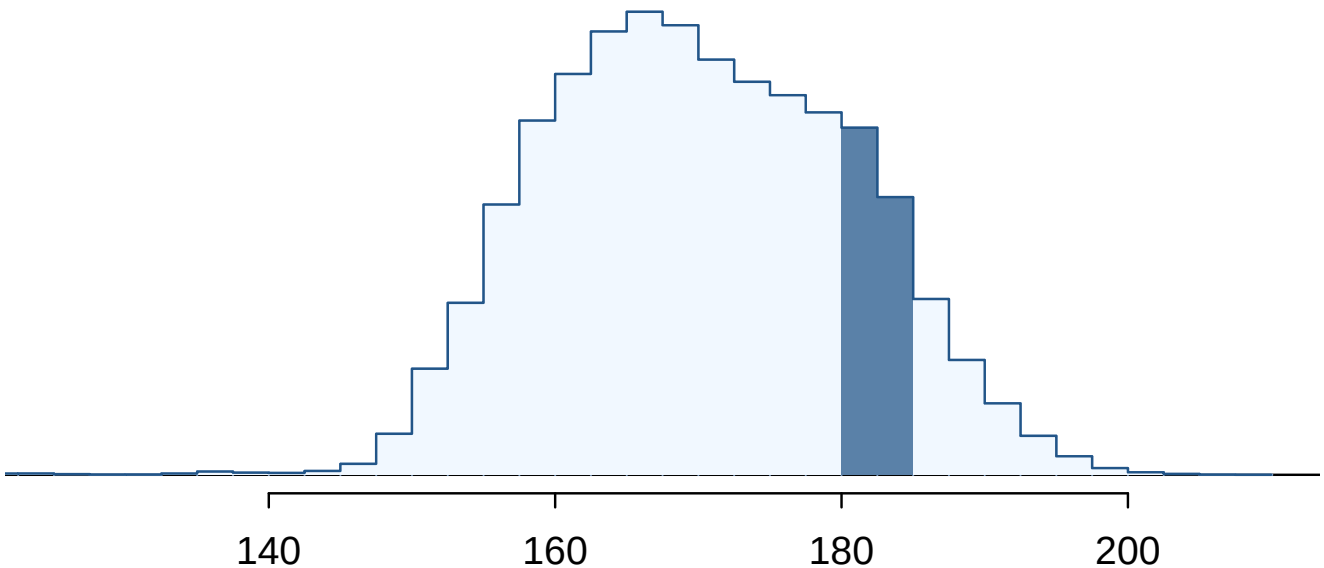
1. Probability density function
2. Cumulative distribution function

2. Normal distribution

1. Standardizing with Z scores
2. Percentiles
3. 68-95-99.7 rule
4. Normal approximation to the binomial

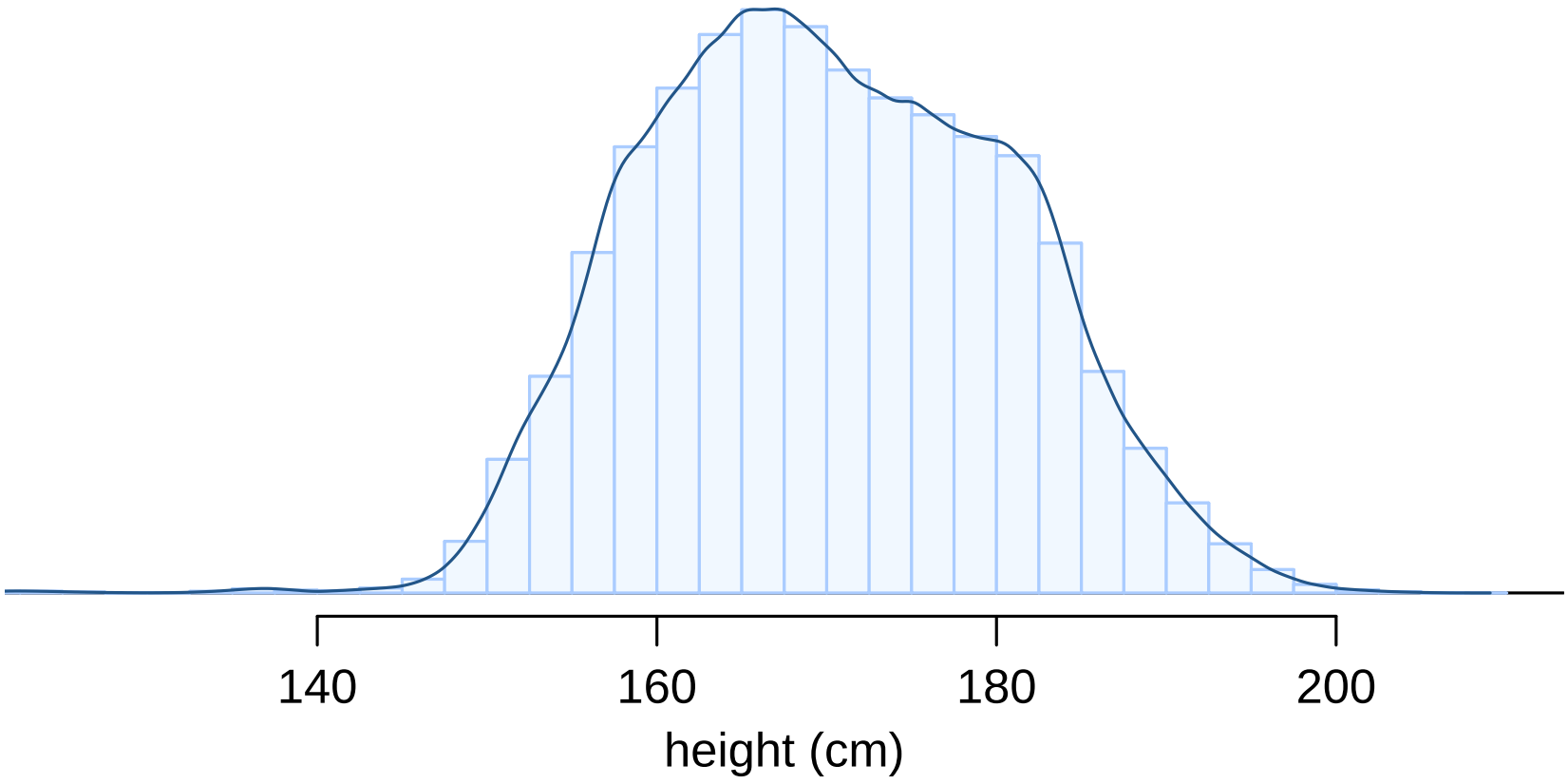
Continuous random variable

- In the previous lecture, we described discrete distributions, such as the binomial, geometric. A discrete distribution has positive probability on a discrete set.
- For continuous random variables, as we shall soon see, the probability that X takes on any particular value x is 0. One can only assign probabilities to intervals.
- Below is a histogram of heights of US adults (height is a continuous numerical variable). The proportion of data that falls in the shaded bins gives the probability that a randomly sampled US adult is between 180 cm and 185 cm (about 5'11" to 6'1").



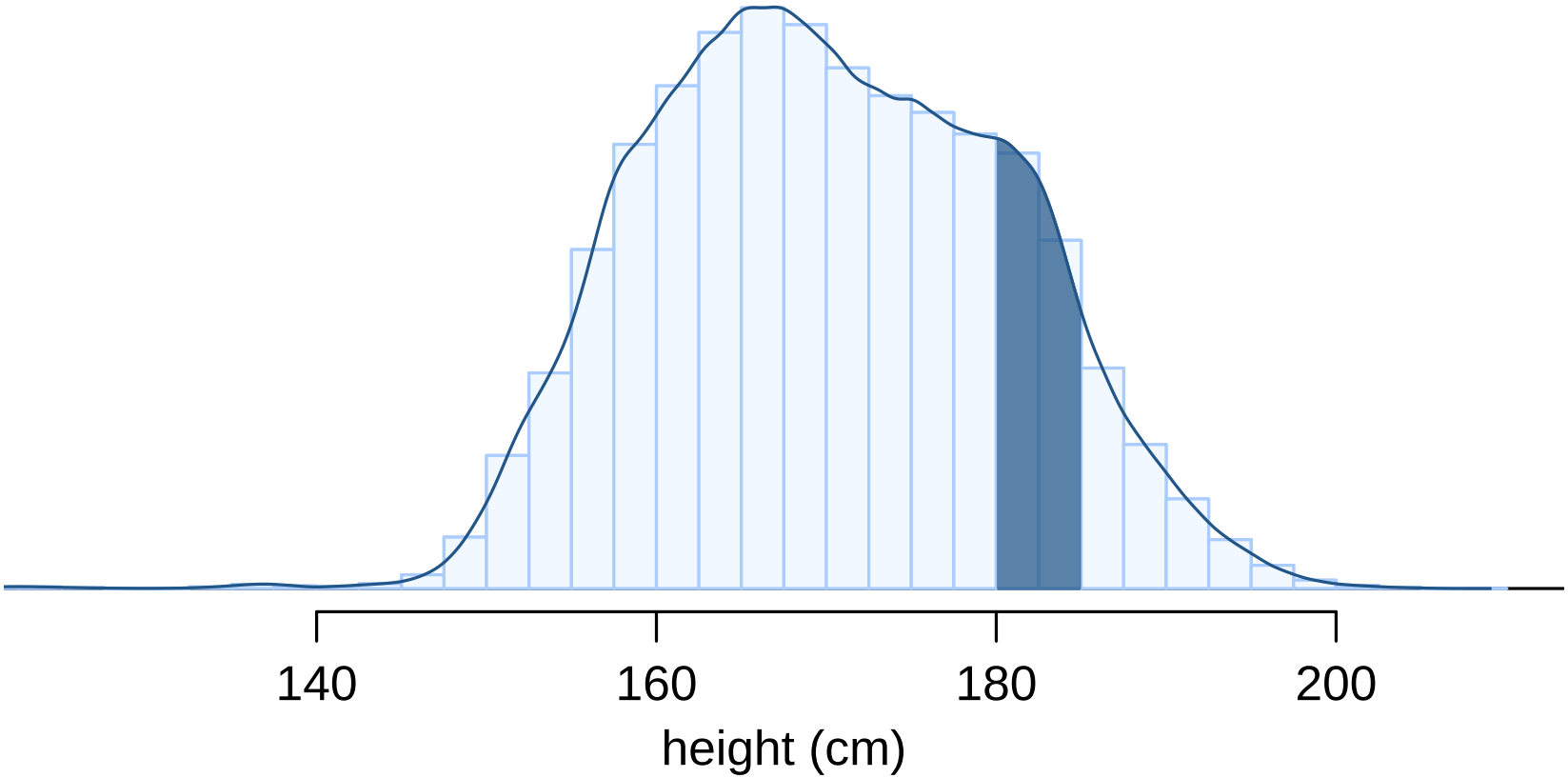
From histograms to continuous distributions

If we keep decreasing the length of each interval, the intervals would eventually get so small that we could represent the probability distribution of height, not as a histogram, but rather as a smooth curve, we call this curve *probability density function*.



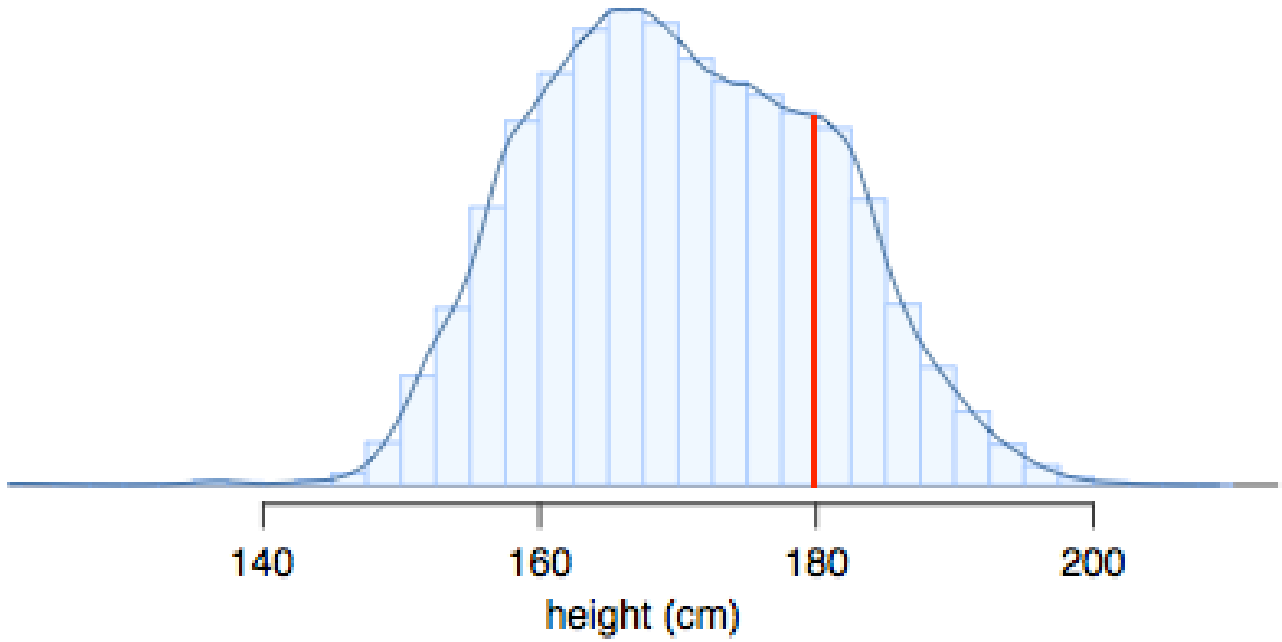
Probabilities from continuous distributions

Therefore, the probability that a randomly sampled US adult is between 180 cm and 185 cm can also be estimated as the shaded area under the curve.



Probability of a particular value

Since continuous probabilities are estimated as “the area under the curve”, the probability of a person being exactly 180 cm (or any exact value) is defined as 0.



Probability density function

A probability density function is any function $f(x)$ such that

- $f(x)$ is non-negative, $f(x) \geq 0$
- $f(x)$ integrates to 1, $\int_{-\infty}^{+\infty} f(x)\,dx = 1$.

This definition of the density function ensures

(1) all probabilities are between 0 and 1:

$$P[a \leq X \leq b] = \int_a^b f(x)\,dx \in [0, 1].$$

(2) $P[-\infty < X < \infty] = 1$

(3) if A and B are disjoint intervals, then

$$P[X \in A \text{ or } X \in B] = P[X \in A] + P[X \in B].$$

Cumulative distribution function

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(y)\, dy.$$

Then $P[a \leq X \leq b] = \int_a^b f(x)\, dx = F(b) - F(a)$.

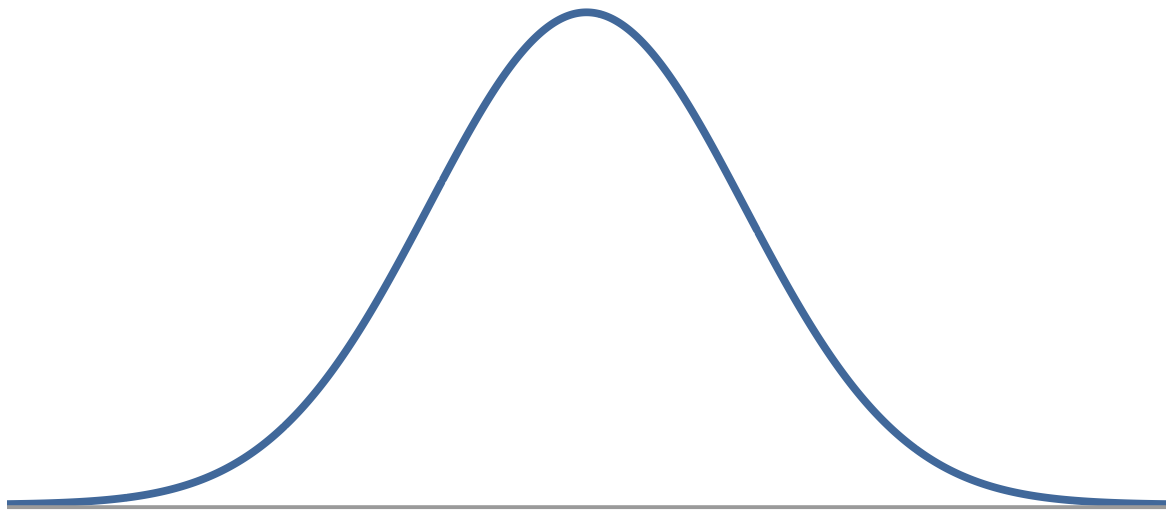
- *Expectation*: $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)\, dx$.
  - The expectation of a function $h(X)$ of X is $E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)\, dx$.
- *Variance*

$$\begin{aligned} Var(X) &= E[(X-\mu)^2] \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

- As before, $E(aX+bY) = aE(X) + bE(Y)$;
  $V(aX+bY) = a^2 V(X) + b^2 V(Y)$ if X and Y are **independent**.

Normal distribution

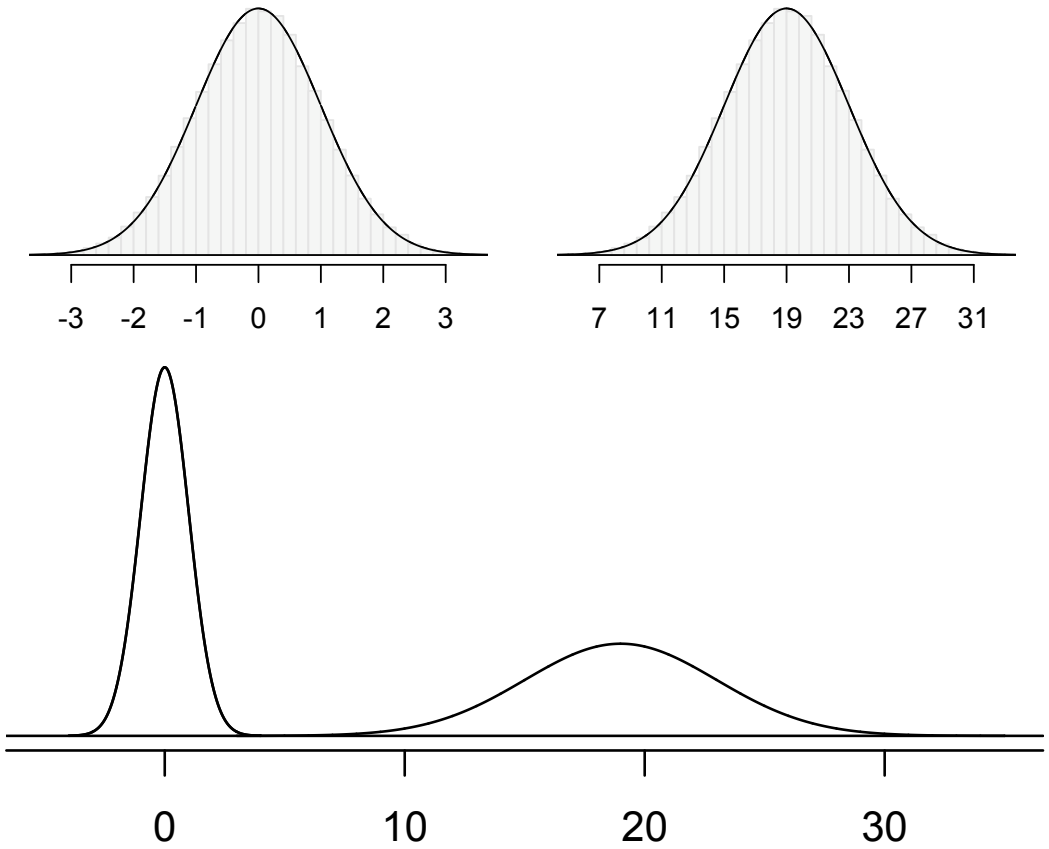
- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as $N(\mu, \sigma) \rightarrow$ Normal with mean μ and standard deviation σ



Normal distributions with different parameters

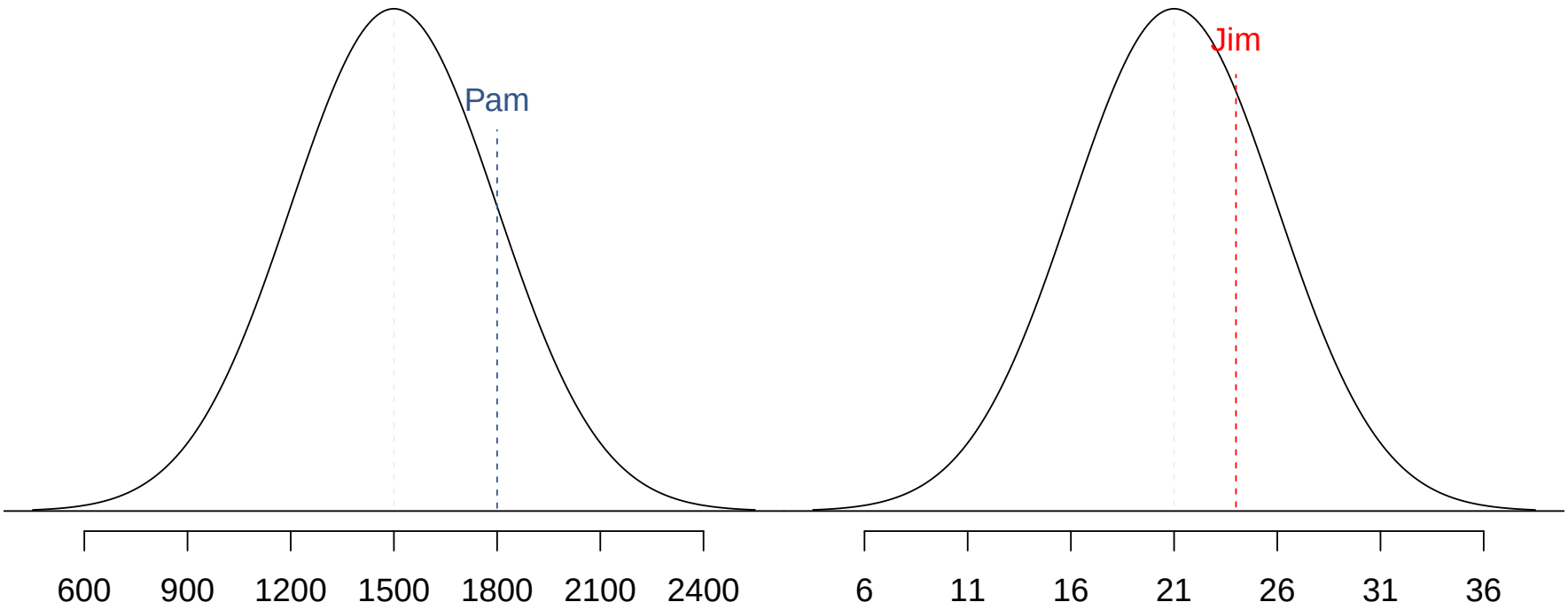
μ: mean, σ: standard deviation

$N(\mu = 0, \sigma = 1)$ $N(\mu = 19, \sigma = 4)$



SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.
ACT scores are distributed nearly normally with mean 21 and standard deviation 5.

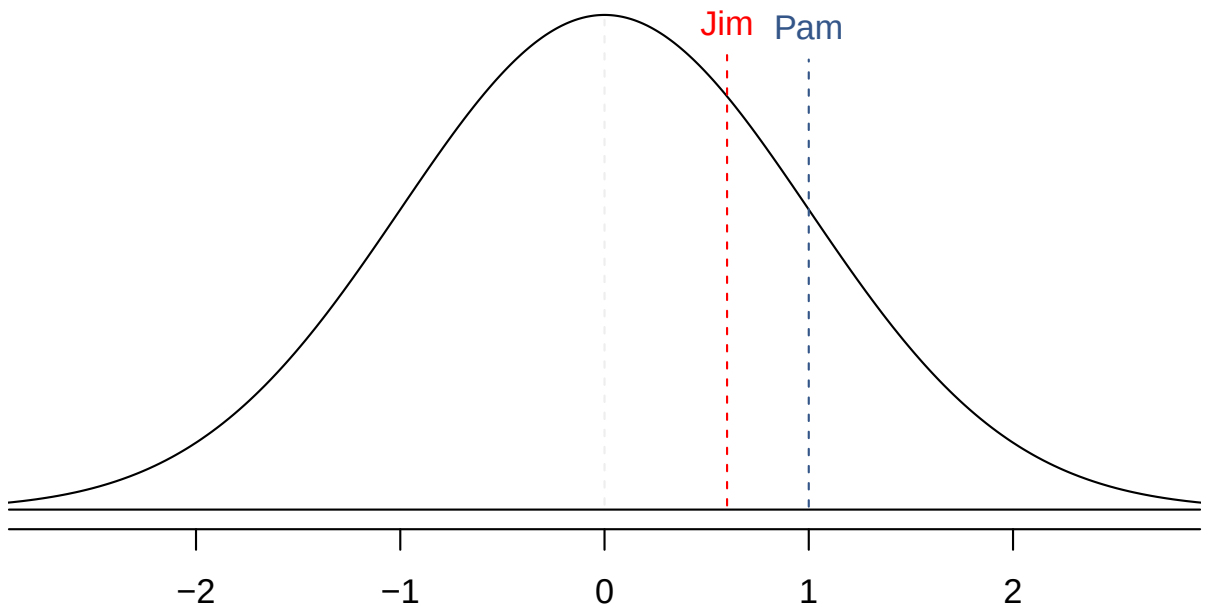
A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?



Standardizing with Z scores

Since we cannot just compare these two raw scores, we instead compare how many standard deviations beyond the mean each observation is.

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



Standardizing with Z scores (cont.)

Z scores

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

- Z scores are linearly transformed data values having mean of zero and standard deviation of 1.
- Z score creates a common scale so you can assess data without worrying about the specific units in which it was measured.
- Z score of an observation is the number of standard deviations it falls above or below the mean.
- Observations that are more than 2 SD away from the mean ($|Z| > 2$) are usually considered unusual.
- Z scores are defined for distributions of any shape and *not necessarily normally distributed*. Only when the distribution of data is normal can we say Z scores have a *standardized* normal distribution $N(0, 1)$.

Practice

Scores on a standardized test are normally distributed with a mean of 100 and a standard deviation of 20. If these scores are converted to standard normal Z scores, which of the following statements will be correct?

(a) The mean will equal 0, but the median cannot be determined.

(b) The mean of the standardized Z-scores will equal 100.

(c) The mean of the standardized Z-scores will equal 5.

(d) Both the mean and median score will equal 0.

(e) A score of 70 is considered unusually low on this test.

Percentiles

If X is a continuous random variable, then the $(100p)^{th}$ percentile is a number π_p such that the area under $f(x)$ and to the left of π_p is p.

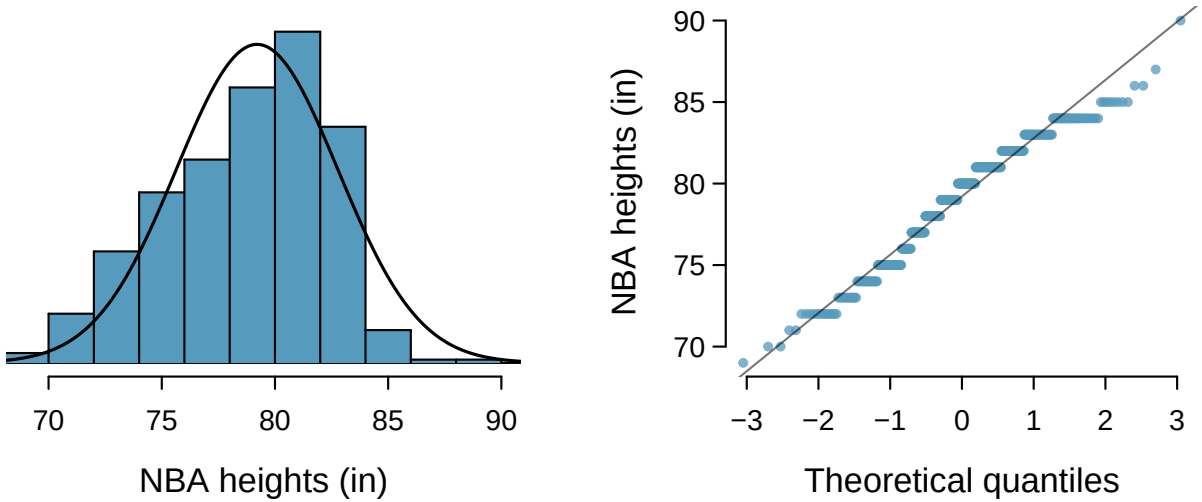
$$\int_{-\infty}^{\pi_p} f(x)\,dx = p$$

Some percentiles are given special names:

- The 25th percentile, $\pi_{0.25}$, is called the first quartile (denoted $q1$).
- The 50th percentile, $\pi_{0.50}$, is called the median (denoted m) or the second quartile (denoted $q2$).
- The 75th percentile, $\pi_{0.75}$, is called the third quartile (denoted $q3$).

Evaluating the normal approximation

Below is a histogram and a quantile-quantile (Q-Q) plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?

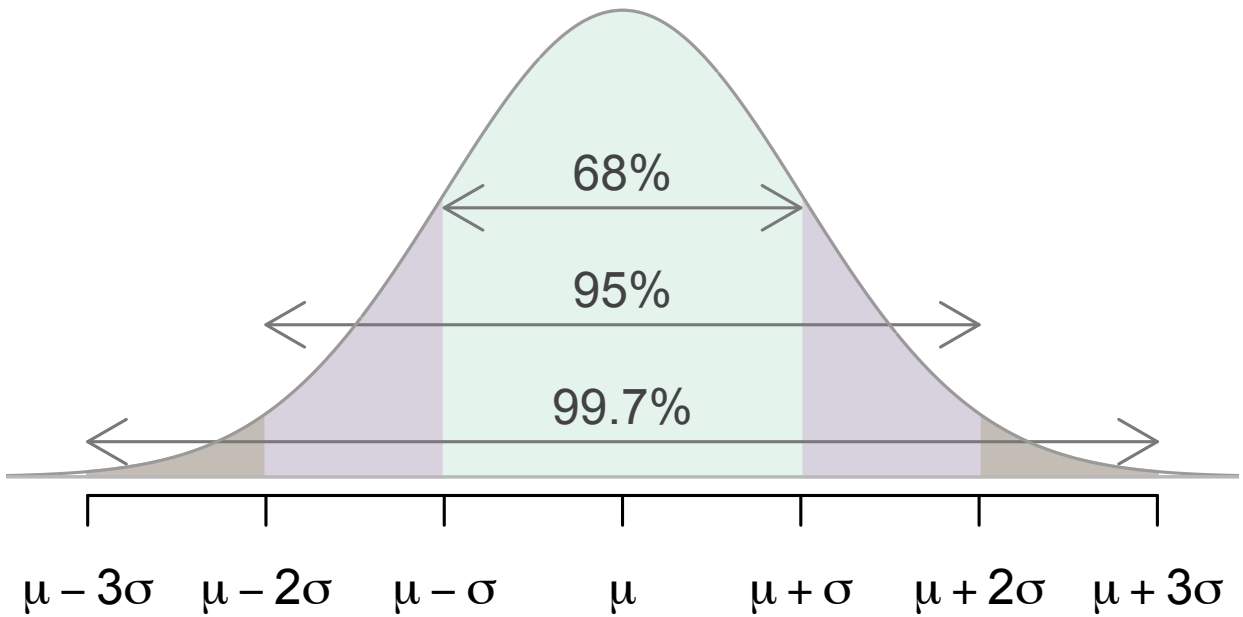


Q-Q Plot

- Z scores are plotted on the y-axis, and theoretical quantiles (from $N(0, 1)$) on the x-axis. (`qqnorm` in R)
- If there is a linear relationship (with slope 1) in the plot, then the data follow a nearly normal distribution.

68-95-99.7 Rule

- For nearly normally distributed data,
  - about 68% falls within 1 SD of the mean,
  - about 95% falls within 2 SD of the mean,
  - about 99.7% falls within 3 SD of the mean.
- It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



Practice

Which of the following is false?

(a) Majority of Z scores in a right skewed distribution are negative.

(b) In skewed distributions the Z score of the mean might be different than 0.

(c) For a normal distribution, IQR is less than $2 \times SD$.

(d) Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

Practice

The temperature in June in LA closely follow a normal distribution with mean $77°$F and a standard deviation of $5°$F. Which of the following is the most reasonable guess for the temperature in June in LA with 95% probability?

(a) 77 ± 5

(b) $77 \pm 2 \times 5$

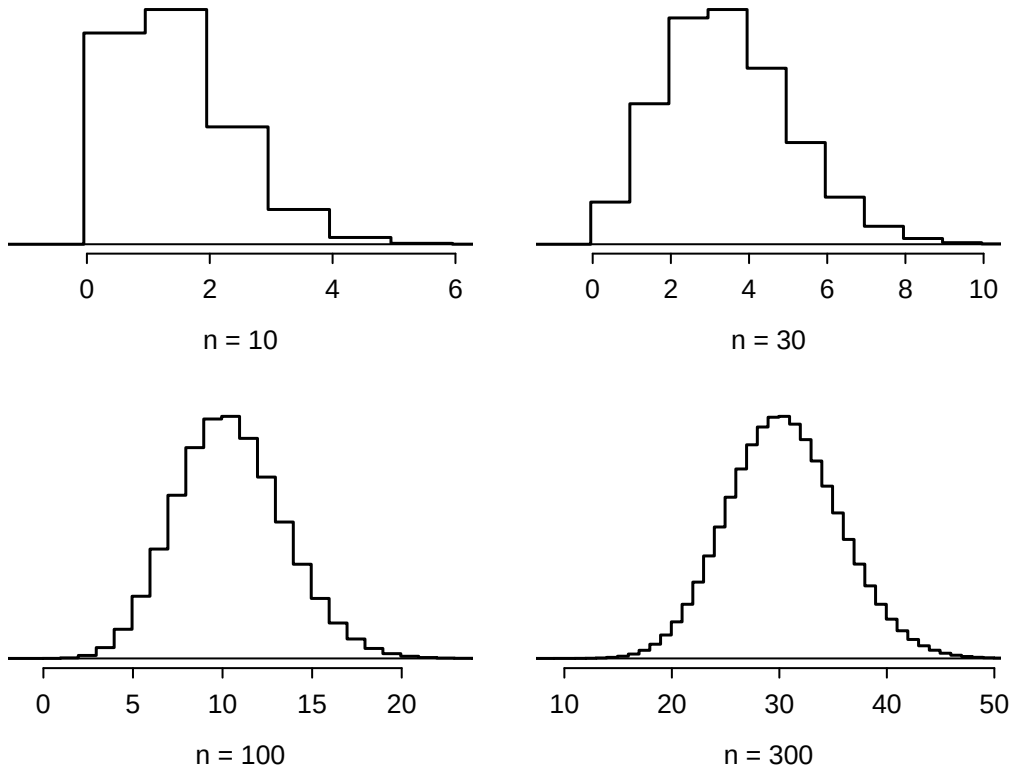
(c) $77 \pm 3 \times 5$

(d) cannot tell from the information given

Shapes of binomial distributions

Go to *https://gallery.shinyapps.io/dist_calc/* and choose Binomial coin experiment in the drop down menu on the left.

Follow histograms of samples from the binomial model where $p = 0.10$ and $n = 10$, 30, 100, and 300. What happens as n increases?



Normal approximation to the binomial

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ *and* $\sigma = \sqrt{np(1-p)}$.

A study found that approximately 25% of Facebook users are power users.
The same study found that the average Facebook user has 245 friends.
What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?

We are given that $n = 245, p = 0.25$, and we are asked for the probability $P(X \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

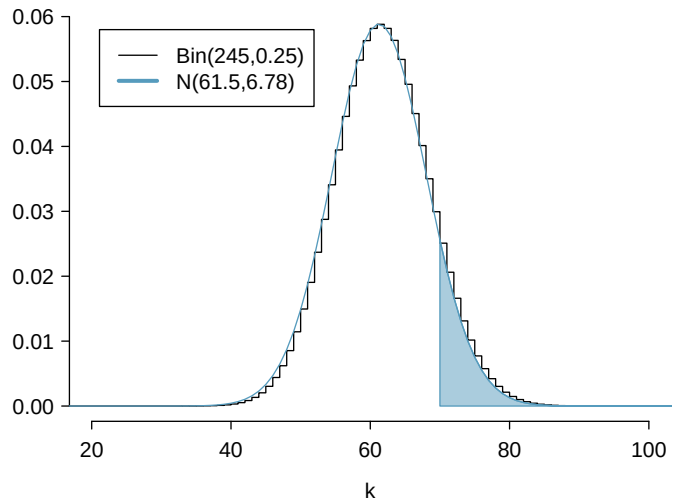
$$\begin{aligned} P(X \geq 70) &= P(X = 70 \text{ or } X = 71 \text{ or } X = 72 \text{ or } \cdots \text{ or } X = 245) \\ &= P(X = 70) + P(X = 71) + P(X = 72) + \cdots + P(X = 245) \end{aligned}$$

This seems like an awful lot of work...

- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.



- R:

```
> 1 - pnorm(70, mean=61.25, sd = 6.78)
[1] 0.09842807
```