Chapter 9: Essential Probability Rules

Corinne Riddell

September 22, 2021

Learning objectives for today

- Why does probability matter?
- How to estimate a probability from a sample
- How do estimates change as the sample size increases?
- Learn new terminology: sample space (discrete vs. continuous), event, discrete probability model, continuous probability model
- Basic probability rules
- What is a random variable
- More definitions: risk, odds, case fatality rate

Why is probability important

- Statistics can be misleading, knowing the basic rules of probability helps you interpret statistics more clearly.

Three examples on the next slides show that

1) Misleading statistics can be used to make and defend policy decisions
    - You need to know how to interpret statistics for yourself rather than trusting someone else's interpretation
2) Determining probability can be difficult and not intuitive. Our gut instinct about the probability of an event may be way off and lead to poor decision making in a medical setting
3) Using a predictive models to calculate a probability sounds robotic and unbiased. But these models encode bias and discrimination.

Example. 1: Misleading statistics can be used to make and defend policy decisions

- Kirstjen Nielsen, former Homeland Security Secretary, stated the following about individuals crossing the US-Mexico border:

  "Again, let's just pause to think about this statistic: **314 percent** increase in adults showing up with kids that are not a family unit... Those are traffickers, those are smugglers, that is MS-13, those are criminals, those are abusers."

- Nielsen was speaking about a relative increase in the probability of the event of "adults with kids who are not their own at the US-Mexico border". The increase on the relative scale is very large (314%). However, how often did the event happen in the first place?

- In a Washington Post analysis[1], the increase was from 0.19% in 2017 to 0.61% in 2018. Thus the actual chance of the event happening is very small and increased by 0.61%-0.19% = 0.42 percentage points.

- Takeaway: looking at the increase in absolute percentage points provides a different interpretation than the increase on the relative scale.

Reference: https://www.washingtonpost.com/news/politics/wp/2018/06/18/how-to-mislead-with-statistics-dhs-secretary-nielsen-edition/?noredirect=on&utm_term=.9193534ee80c

Example 2: Calculating probabilities in medical settings can be difficult and not intuitive

- Suppose that there is test for a specific type of cancer that has a 90% chance of testing positive for cancer if the individual truly has cancer and a 90% chance of testing negative for cancer when the individual does not have it.
- 1% of patients in the population have the cancer being tested for.
- What is the chance that a patient has cancer given that they test positive?

a) Between 0% - 24.9%
b) Between 25.0% - 49.9%
c) Between 50.0% - 74.9%
d) Between 75.0% - 100%

Example 2: Calculating probabilities in medical settings can be difficult and not intuitive

Many people choose d), but the true answer is a)! Why do we get this so wrong?

Video link (2 mins): click here

Example 3: Algorithms can be biased and discriminatory

- Algorithms are increasingly used to automate decision-making.
- One algorithm aids in decision-making by judges during criminal sentencing. Judges want to determine who is likely to re-offend.
- An investigative report by ProPublica[2] found that:

  "the formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants. White defendants were mislabeled as low risk more often than black defendants"

- Takeaway: algorithms to calculate probability can be biased and discriminatory.

Reference: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Understanding probability

- These three examples illustrate the importance of understanding probability in medicine, public policy, and decision-making
- Often, our understanding of probability is based on estimates from a sample drawn from a probability
- Last class, we talked about how larger samples provide better estimates on average.
- The following slides review this concept (for you to read on your own time)

From B&M page 216: How common is the common cold?

- Suppose there are 100,000 people in your community. You work for your community's public health office and want to estimate the number of people who had a common cold. You are not able to sample everyone but suppose you could randomly call people in your community and ask them "Did you have a cold yesterday?" and then calculate the proportion of the sample who had a cold.

From B&M page 216: How common is the common cold?

- Suppose you somehow had access to data on the whole population in a data frame.
- For each person, the variable `had_cold_yesterday` equals 0 if they did not have a cold yesterday, and 1 if they did have a cold.
- Suppose these data are called `cold_data` with the following outputs:

```r
dim(cold_data)
```

```
## [1] 100000      2
```

```r
head(cold_data, 20)
```

```
##    id had_cold_yesterday
## 1   1                  0
## 2   2                  0
## 3   3                  0
## 4   4                  0
## 5   5                  0
## 6   6                  0
## 7   7                  0
## 8   8                  0
## 9   9                  0
## 10 10                  0
## 11 11                  0
## 12 12                  0
## 13 13                  0
## 14 14                  1
## 15 15                  0
## 16 16                  0
## 17 17                  0
## 18 18                  0
## 19 19                  0
## 20 20                  0
```

```r
cold_data %>% summarize(population_mean = mean(had_cold_yesterday))
```

```
##   population_mean
## 1         0.11214
```

- The average (mean) of a variable containing only 0's and 1's is called a **proportion**. This is because the mean is the number of individuals with a cold (coded as `had_cold_yesterday = 1`) divided by the total number of individuals.

From B&M page 216: How common is the common cold?

Realistically, you would not have access to data on every person in the population. You need to take a sample instead. How many people should be included in the sample?

From B&M page 216: How common is the common cold?

Realistically, you would not have access to data on every person in the population. You need to take a sample instead. How many people should be included in the sample?

- We want to sample enough people such that the proportion of those with colds in the sample is close to the proportion of those with colds in the population
- Let's take samples of size 5, 100, 1000, and so on, using `dplyr`'s `slice_sample(n=)` function:

```r
sample_5 <- cold_data %>% slice_sample(n = 5)
sample_100 <- cold_data %>% slice_sample(n = 100)
sample_1000 <- cold_data %>% slice_sample(n = 1000)
sample_10000 <- cold_data %>% slice_sample(n = 10000)
sample_100000 <- cold_data %>% slice_sample(n = 100000)
```

Proportion with cold as a function of sample size

```
sample_5 %>% summarize(sample_mean_n5 = mean(had_cold_yesterday))
```

```
##   sample_mean_n5
## 1              0
```

```
sample_100 %>% summarize(sample_mean_n100 = mean(had_cold_yesterday))
```

```
##   sample_mean_n100
## 1             0.08
```

```
sample_1000 %>% summarize(sample_mean_n1k = mean(had_cold_yesterday))
```

```
##   sample_mean_n1k
## 1           0.112
```

```
sample_10000 %>% summarize(sample_mean_n10k = mean(had_cold_yesterday))
```

```
##   sample_mean_n10k
## 1           0.1091
```

```
sample_100000 %>% summarize(sample_mean_n100k = mean(had_cold_yesterday))
```

```
##   sample_mean_n100k
## 1           0.11214
```

- What do you notice about the proportion estimates?
- Do they approach the true estimate as the sample size increases?

How many people should we sample?

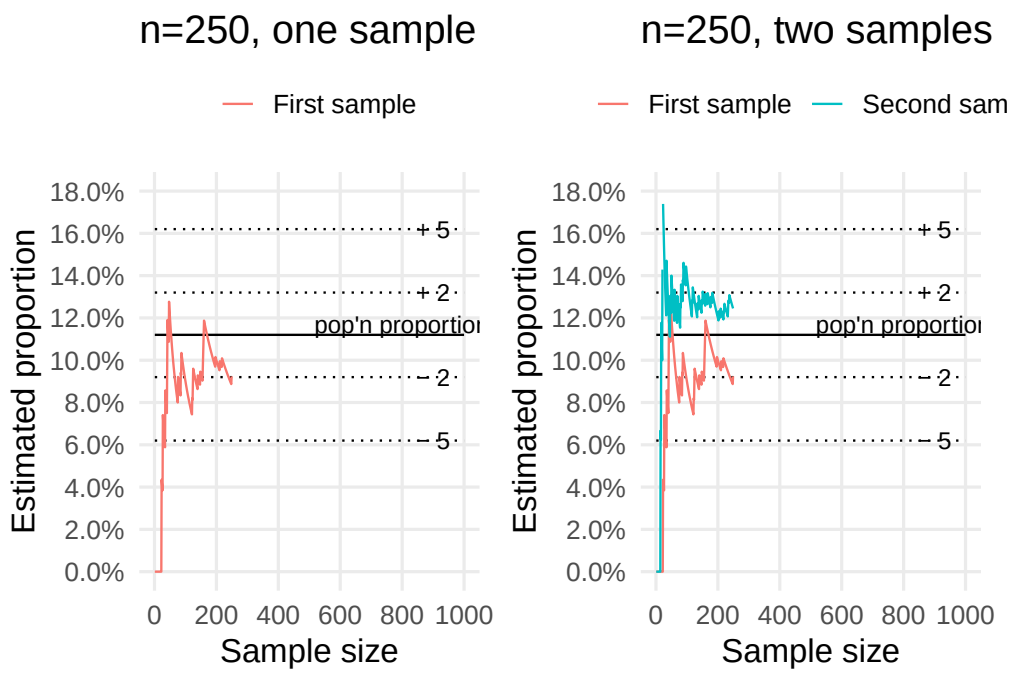
- We know we should sample more than five people, but practically speaking we can't sample everyone.
- We need to sample *enough* people to reasonably estimate the true chance of having a cold. But how many is enough?

Proportion with cold as a function of sample size

- To help us decide I first sampled one person and took the mean of that sample.
- Then I added another person and took the mean of that sample of size 2 (n=2) and so on.
- The plot on the next slide shows the estimated proportion vs. the sample size for samples up to size n=250.
- It also shows variation across samples, pretending I took the first sample and you took a second sample and we both end up with samples of size 250.

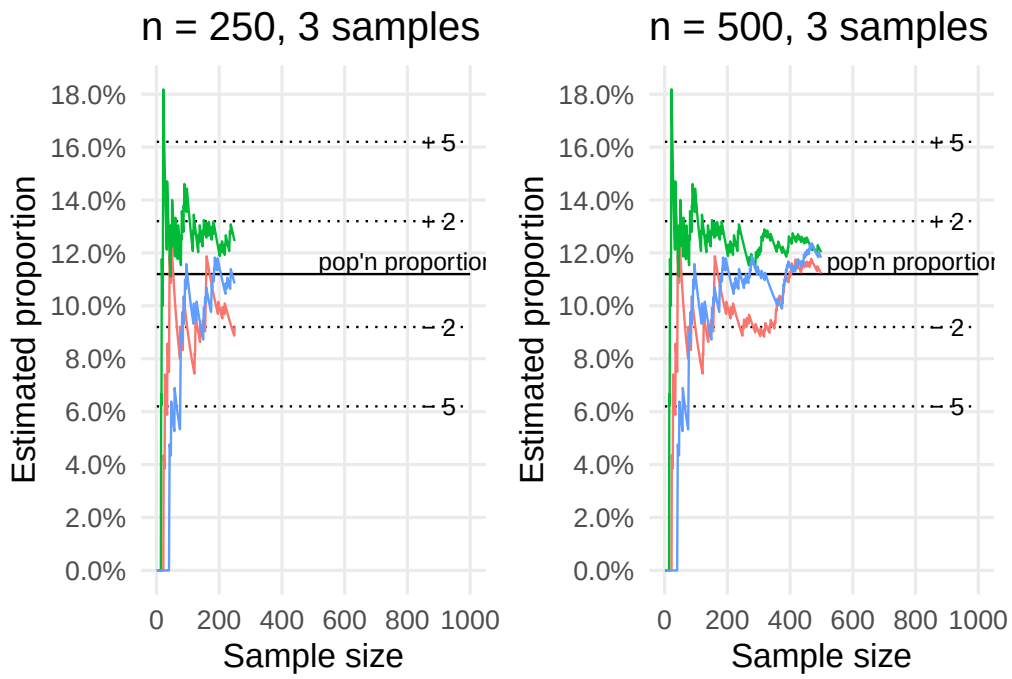
Estimated proportion vs. sample size for n = 250

The true proportion is shown on the graph using the solid horizontal line. How close are we to the true proportion? Do we get closer as sample size increases?



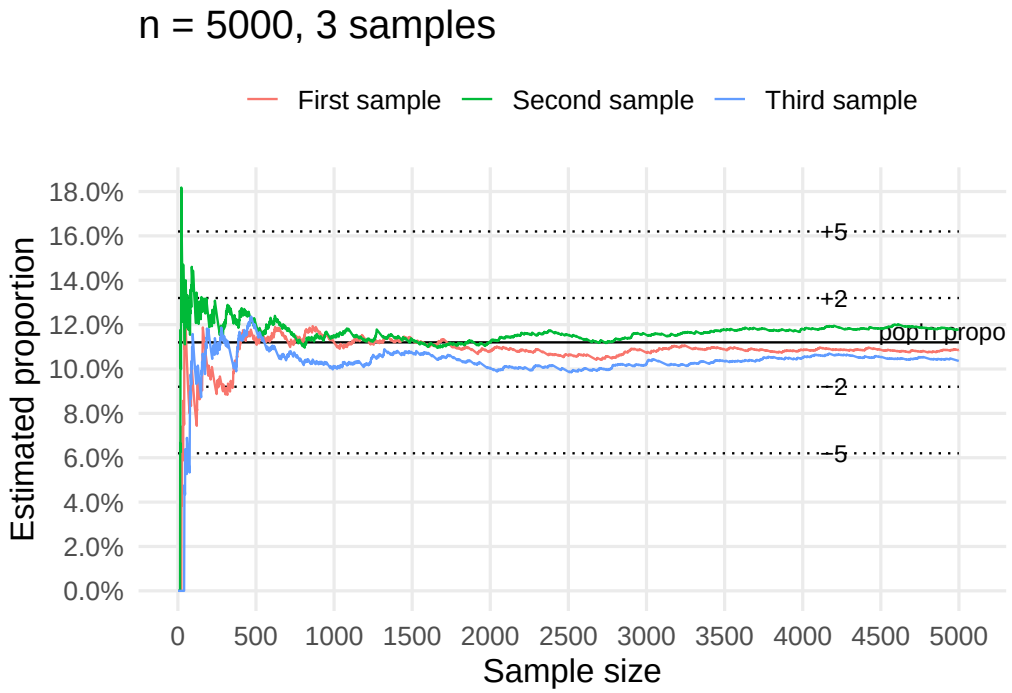
Estimated proportion vs. sample size for n = 250 and n = 500

- Increase the sample size how the estimate becomes closer to the true value
- Add in a third sample to compare how different samples perform in the short vs. the long run



Estimated proportion vs. sample size for n = 5000

The graph shows we all get fairly close when we take a sample of size 5000, but the samples are all slightly off.



Summary of the example on estimating the proportion with a cold

- As the sample size increases, the estimated proportion becomes closer to the true proportion
- Random samples of the same size will provide different estimates of the true proportion, but will be closer to each other (and the true value) if they are "large enough"

Some vocabulary, definitions, and rules

Sample space

A **sample space** is the set of all possible outcomes. It is denoted by the letter S.

Discrete sample space

- A discrete sample space describes all possible events for a particular variable or outcome
- For example, the discrete sample space for marital status can be defined by S, where S = {Single, married, divorced, widowed}
- Discrete spaces remind us of nominal, ordinal, or discrete variables
- Discrete sample spaces can be written for any variable that is countable with "gaps" between the events
- The notation is important: S = {elements in the space} = (element 1, element 2, ... element n)

Continuous sample space

- A continuous sample space describes all possible measures for a particular variable or outcome
- Because continuous variables don't have "gaps" (they can be measured very precisely), we denote their sample spaces using numerical ranges. For example, all possible values between 0 and 1 can be represented by writing "[0, 1]: S = {all numbers between 0 and 1}".
- Continuous sample spaces remind us of continuous variables only
- The events are not countable (i.e., we cannot list the numbers between 0 and 1; they are infinite)

Event

- An event is one possible outcome or a set of outcomes from a sample space.

- For the discrete example, the event could be being divorced. Or it could be being divorced or being widowed.
- For the continuous example, the event could be a value between 0.4 and 0.5.
- We are interested in determining the probability (or chance or risk) that an event will occur. To calculate this probability, wee need to know the probabilities associated with the sample space.

Proability model

- A probability model is a description of a random phenomenon.
- It consists of two parts: i) defining the sample space S, and ii) listing the probability associated with different events.
- In a few slides we will discuss **discrete** probability models and **continuous** probability models.

Rules of probability

1. Probabilities are numbers between 0 and 1. Let A denote an event, and `P(A)` denote the probability of A occurring. Then we can write this sentence in probability notation as:
   - $0 \leq P(A) \leq 1$
2. All possible outcomes that compose a sample space together have a probability of 1. Let S denote the sample space, then:
   - $P(S) = 1$
3. If two events (here denoted by A and B have a joint probability of 0 (i.e., there is no overlap in their event spaces and they cannot both occur at the same time) then they are **disjoint** and the probability of either event occurring is the summation of their individual probabilities.
   - $P(AorB) = P(A) + P(B)$, if A and B are disjoint events.
4. The probability of an event not occurring is 1 minus the probability of the event occurring. This event (i.e., "not occurring") is called the **complement**.
   - P(A does not occur) = 1 - P(A).
   - Another way to write "A not occurring" is using "A"' (pronounced A prime or A not) or "A^c": $P(A') = 1 - P(A)$ or $P(A^c) = 1 - P(A)$

Discrete probability model

- A probability model with a sample space made up of a list of individual outcomes is called discrete
- To assign probabilities in a discrete model, list the probabilities of all the individual outcomes. These probabilities must be numbers between 0 and 1 and must sum to 1. The probability of any event is the sum of the probabilities of the outcomes making up the event.

Discrete probability model example

For example, we could survey a sample of people and ask them their marital status. Based on this survey we can calculate the portion of each event in the sample space:

Single	Married	Divorced	Widowed
47%	30%	18%	5%

This is a discrete probabiltiy model shown in a table.

How else could you display these data?

Continuous probability model

- A continuous probability model assigns probabilities as areas under a **density** curve. The area under the curve and between a range of specified values on the horizontal axis is the probability of an outcome in that range.

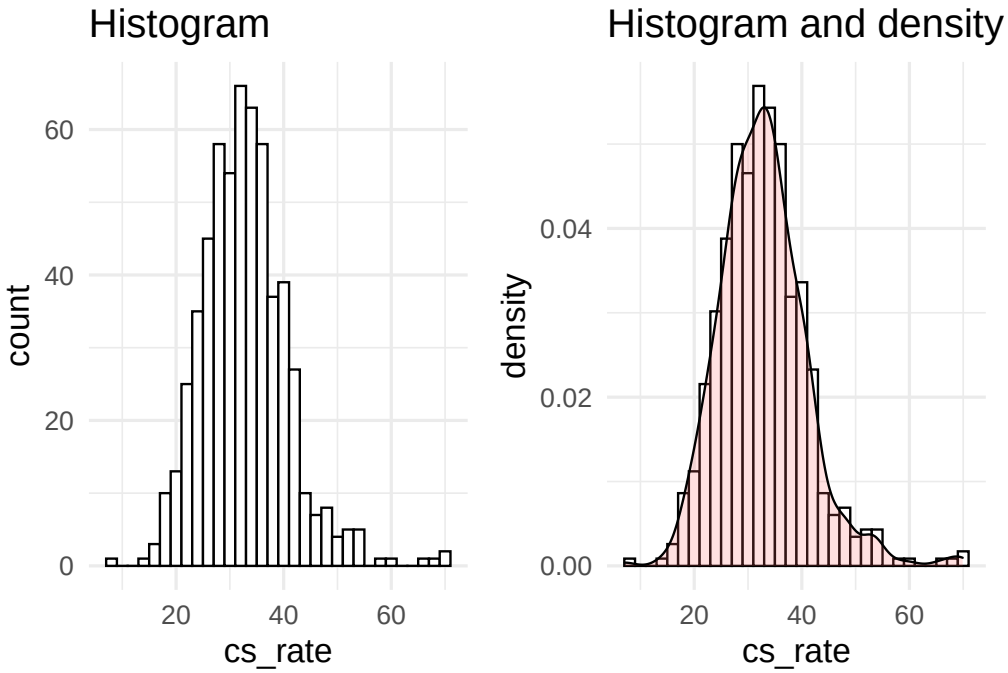
- What is a density curve?

Density curves

- Density curves are also known as **probabilty density functions**
- You can think of density curves as **smoothed histograms**.

Density curves using `geom_density()`

- Recall the data on cesarean delivery rates across hospitals in the US. We can use these data to also make a density plot (also called density curve)



Density curves

- From this plot, we can see that the density curve approximates the shape of the histogram very well.
- Remember because there are infinitely many cesarean delivery rates that could be observed between 0 and 1 it is impossible to assign a finite probability to any specific number.
- If we did, we could do this infinitely and their summed probability would surpass 100%.
- Instead, the density curve is used to determine the probability of an observed event being between a range of values.

Density curves

- Define CS to denote a cesarean delivery rate at a specific hospital. You could use the density curve to calculate:
- The probability that the CS rate is less than 20%, or $P(CS < 0.20)$
- The probability that the CS rate is between 20% and 40%, or $P(0.20 < CS < 0.40)$
- The probability the the CS rate is less than 20% or greater than 40%, or $P(CS < 0.2 \text{ or } CS > 0.4)$. Since these ranges do not overlap (they are disjoint), we can rewrite this as $P(CS < 0.2) + P(CS > 0.4)$
- The probability that the CS rate is larger than 40% is written as $P(CS > 0.4)$. We can also write it as $1 - P(CS \leq 0.4)$ by applying the complement rule.

Interpretations of probability

- The calculations on the previous slide can be interpreted as either:
  - the **probability** that a randomly chosen hospital will have cesarean delivery rate in the specified range.
  - the **proportion** of hospitals with cesarean delivery rates in the specified range
- Thus, the interpretation can both be applied to an individual or a population.

Random variables

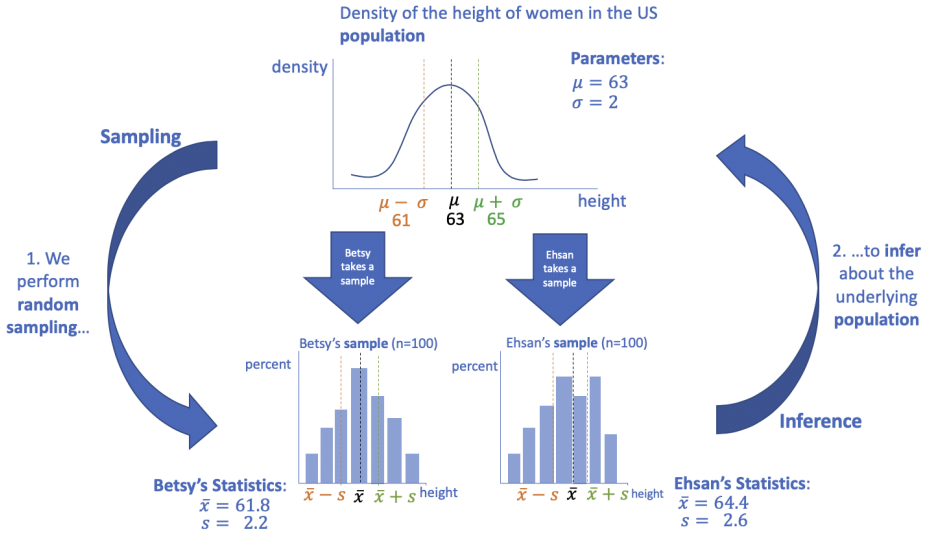
- A random variable is a variable whose value is a numerical outcome of a random phenomenon
- Random variables are represented by capital letters, most popularly the letter X.
- A lower case letter represents a particular value for the random variable has been taken. For example $P(X = x)$ asks, what is the probability that random variable X takes the value x?
- For continuous random variables, we only calculate the probability of X falling in a specified range, e.g., $P(X < x)$, $P(X \geq x)$, or $P(x_1 < X < x_2)$
- Remember: for continuous random variables $P(X = x) = 0$!
- Examples: Let MS denote marital status. Calculate the probability that marital status is equal to divorce, or P(MS = divorce)=?

The mean and standard deviation revisted

- In Chapter 2, we learned about how to calculate the mean ($\bar{x}$) and standard deviation (s) of a sample.
- We can also calculate the mean (μ) and standard deviation (σ) of a population.
- The mean and standard deviation of a population are represented using different notation to remind us that we are describing a population **parameter** vs. the **sample** mean and **sample** standard deviation that are **statistics** used to describe samples.

Putting it all together

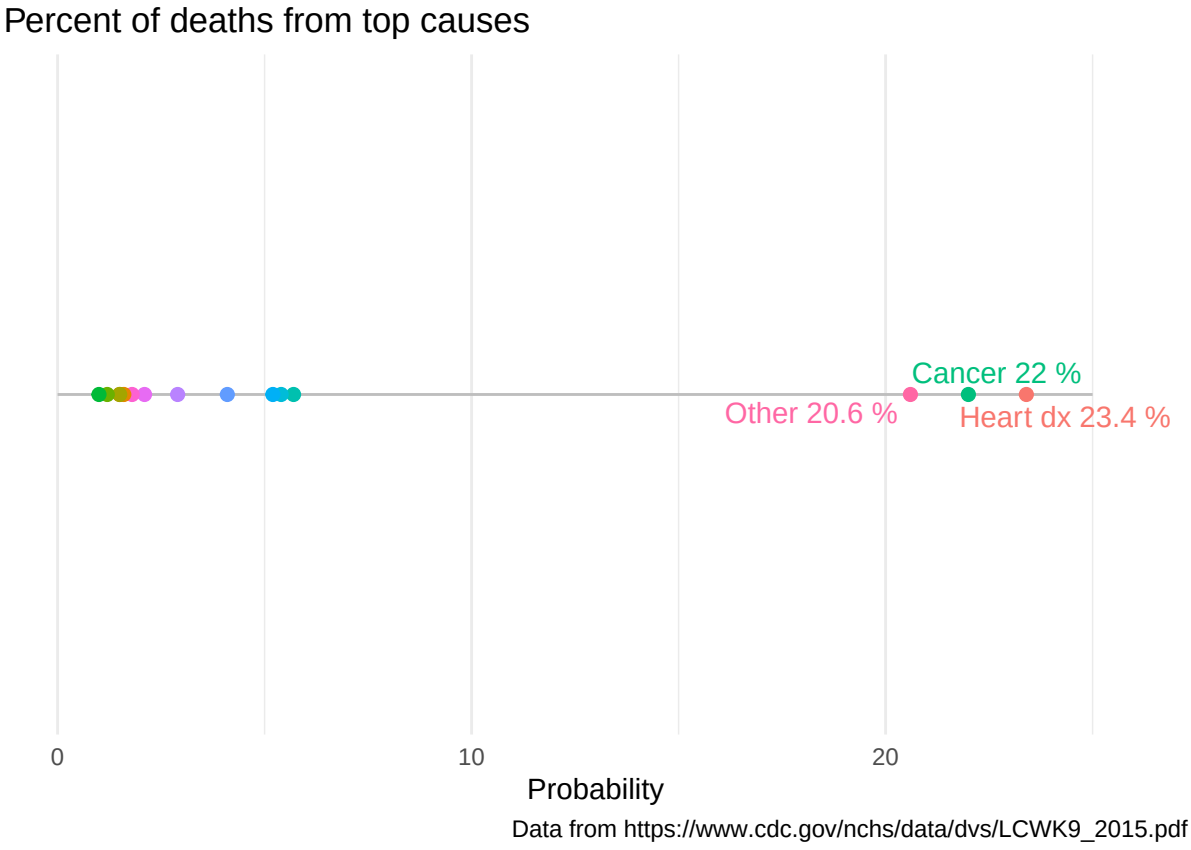
The figure illustrates the difference between a (hypothetical) underlying distribution of heights among women in the US and its mean and standard deviation vs. that of the sampled distribution.



Risk

- Generally speaking, **risk** is another word for the probability or chance of an event occurring. In epidemiology and public health, we often use the word risk to represent the risk of some adverse health outcome among a group of individuals.

- Here is the probability of dying from a specific cause given that a death from one of the top causes has occurred
- Notice that even among the top 10 causes, there is a wide range in the chance of occurrence



Other risk definitions

- **Attack rate** of a virus: the risk (probability) of becoming afflicted during an infectious period, like the flu season. If the attack rate of influenza during a specific flu season was 10% or 10 per 100, this would imply that 10 out of every 100 individuals develop influenza during the epidemic period.
- **Case fatality rate**: the risk (probability) of dying among individuals with a specified condition. For example, the case fatality rate of measles is 1.5 per 1000 cases or 0.15%.
- Note that these definitions use the word "rate" but not in the same way that we define rates in epidemiology. The attack rate and case fatality rate are risks, not rates!

Odds

- The **odds** is another commonly used measure in epidemiology that is a ratio of the probability of the adverse event over the probability of adverse event not occurring. That is, if the risk of event A is equal to p then the odds of the event is equal to $\frac{p}{1-p}$
- If both parents have the gene for sickle cell anemia, the chance that their biologic child will have the disease is 25%. The odds their child will have the disease is $\frac{0.25}{1-0.25} = 0.33$ or 1:3.
- Sometimes, the popular press uses the word "odds" when they actually mean risk or probability. You can usually tell what they mean by the context.

Comparing risks and odds between populations or groups

- In public health and epidemiology, we often contrast the risks and odds of a health outcome between two groups of individuals to ask: Is the risk of the outcome lower (or higher) in individuals who are exposed vs. not exposed (or treated with drug A vs. drug B)?
- We contrast these using ratios (using division) or differences (using subtraction)

Recap: What new concepts did we learn?

- Many definitions: Event, random variable, sample space (discrete and continuous), probability model (discrete and continuous), rules of probability, density curves, risk, odds