STAT 113
Sampling Distributions and Confidence Intervals

Colin Reimer Dawson

Oberlin College

September 28, 2017

Outline

Inference Goals

Sampling Distributions

Two Main Goals of Inference

1. Assessing strength of evidence about "yes/no" questions (hypothesis testing)
2. Estimating unknown quantities in a population using a sample (confidence intervals)

Statistics vs. Parameters

- Summary values (like mean, median, standard deviation) can be computed for populations or for samples.
- In a population, such a summary value is called a **parameter**
- In a sample, these values are called **statistics**, and are used to *estimate* the corresponding parameter

Value	Population Parameter	Sample Statistic
Mean	μ	$\bar{X}$
Proportion	p	$\hat{p}$
Correlation	ρ	r
Slope of a Line	β_1	$\hat{\beta}_1$
Difference in Means	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$
...	...	...

Outline

Inference Goals

Sampling Distributions

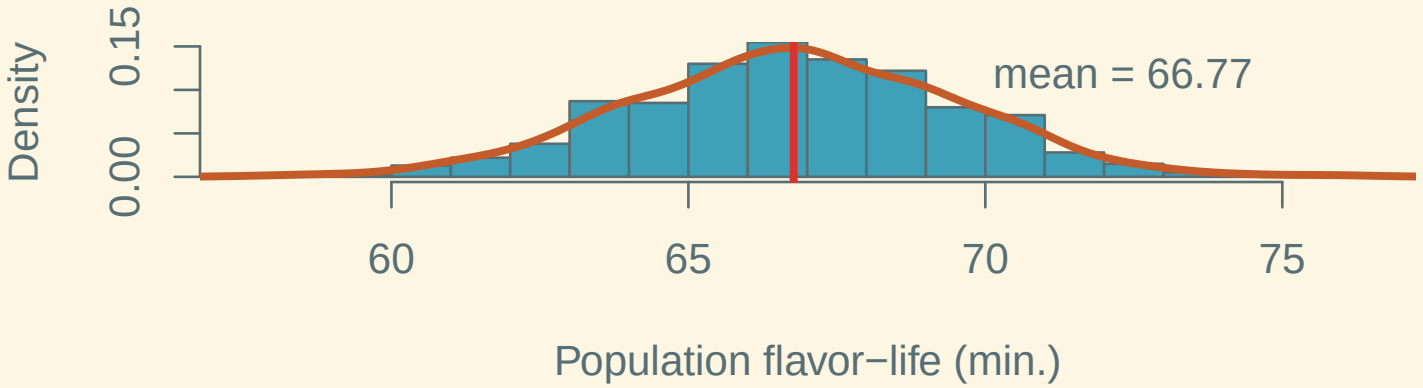
Using Samples to Make Estimates About Populations

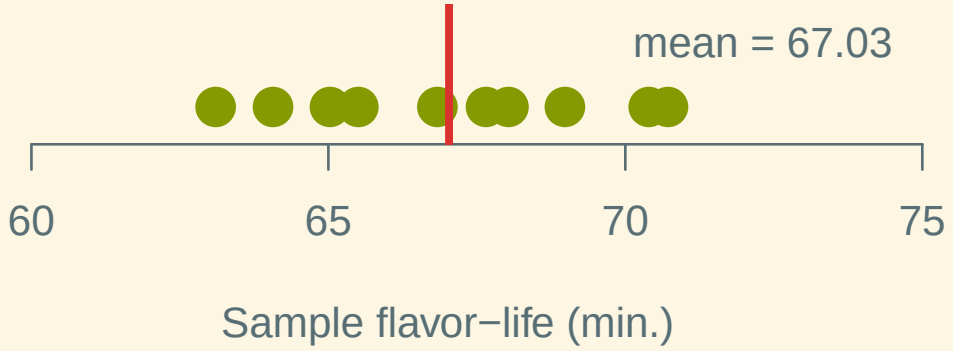
- I want to know the mean flavor-life (in minutes) of gumballs from my gumball factory.
- The set of all gumballs is my **population**.
- The mean flavor-life of all the gumballs produced from the factory is a **population parameter** (write μ for the pop. mean)
- I can only test a **sample** — ideally, a random one.
- The mean flavor-life in the sample is a **sample statistic** (write $\bar{x}$ for the sample mean).

Statistic : Sample :: Parameter : Population

Variability due to Sampling

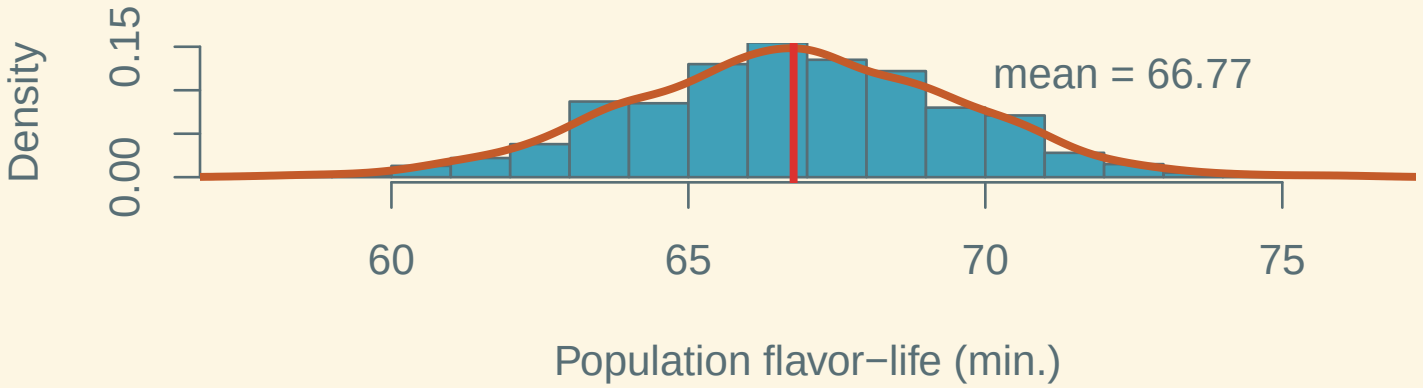
If we take a random sample from this population, it might look like this

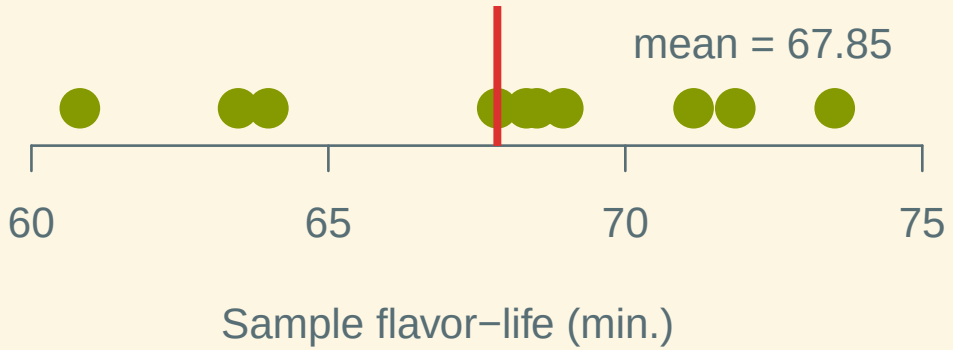




Variability due to Sampling

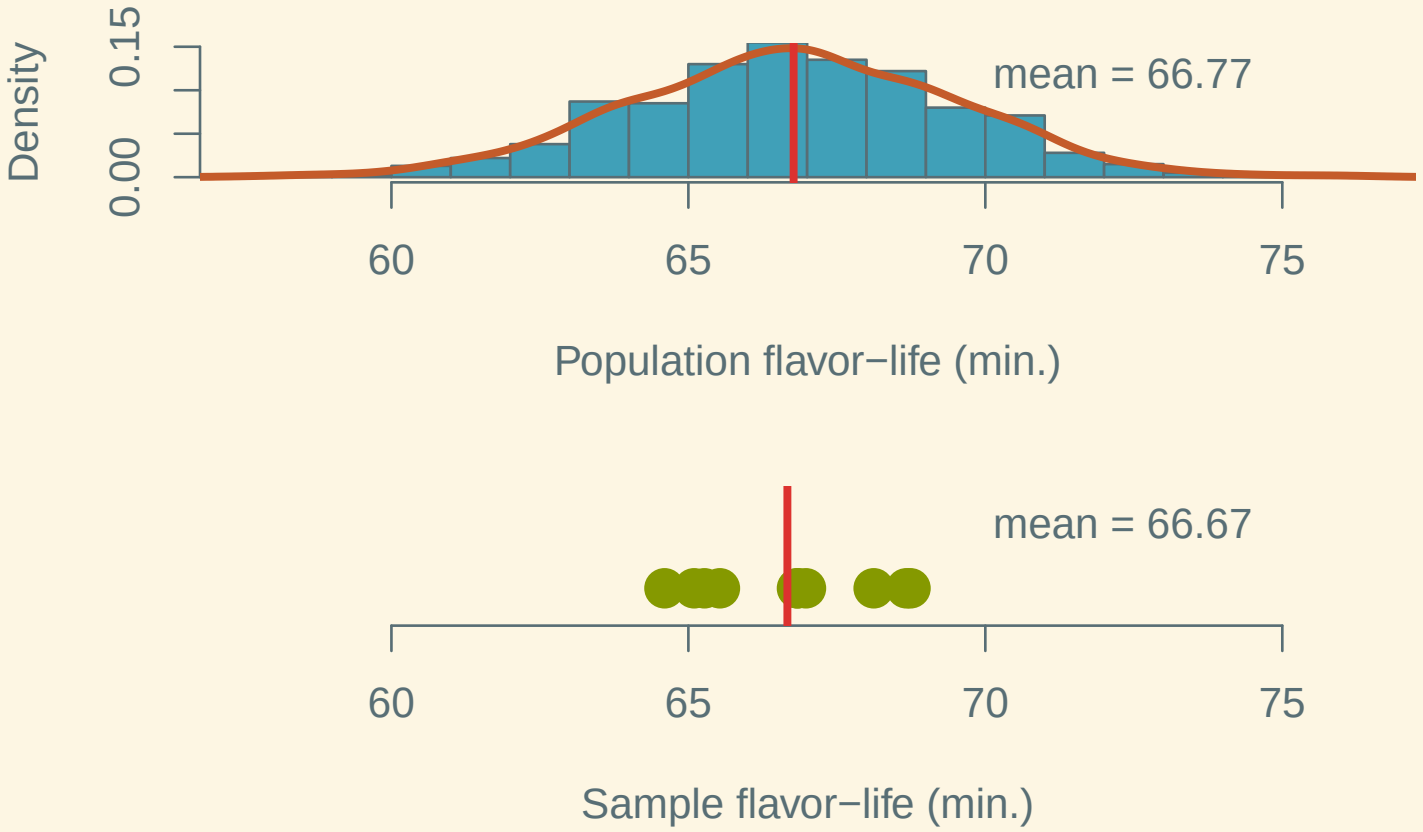
Or this





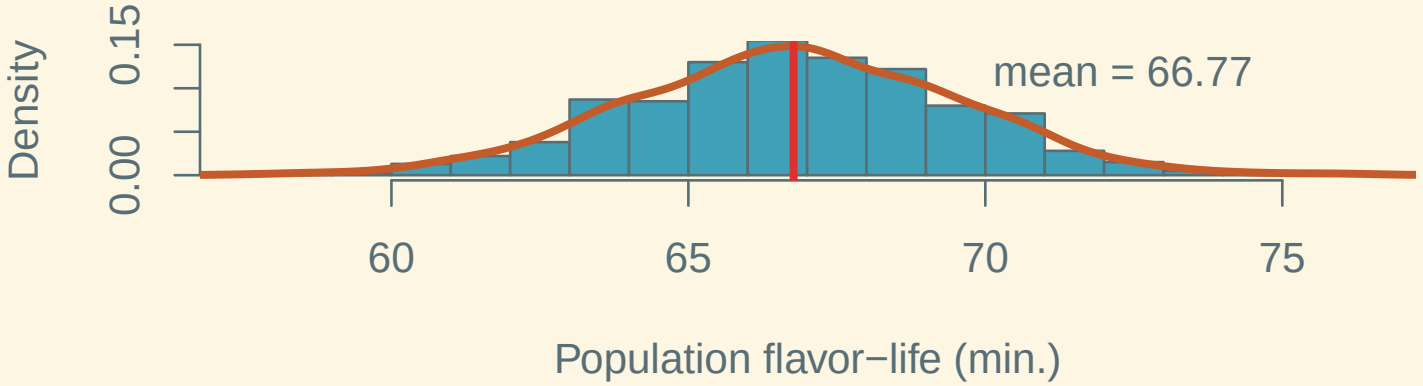
Variability due to Sampling

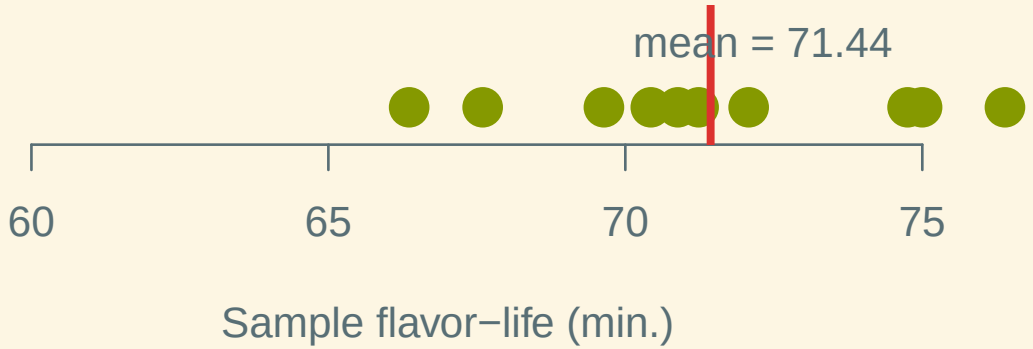
Or this



Variability due to Sampling

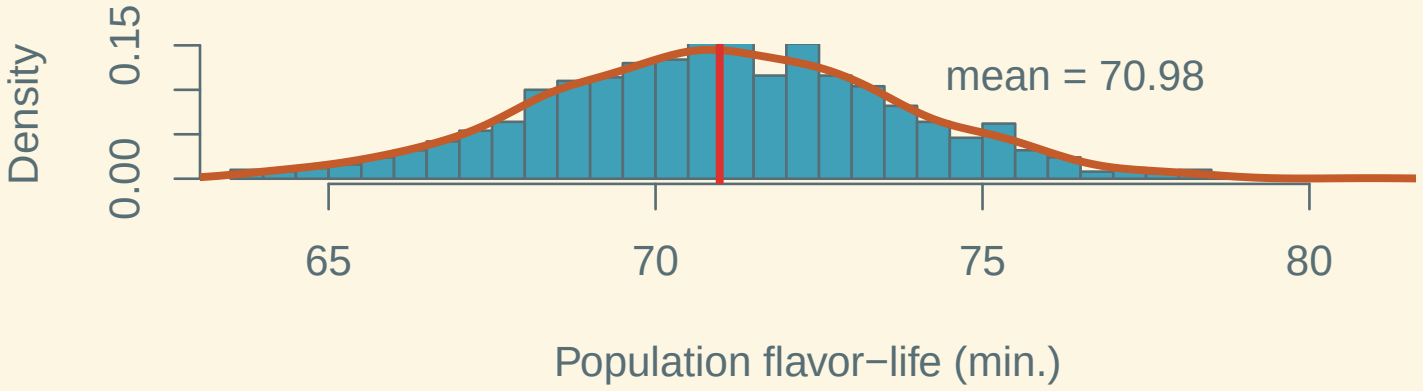
This one could happen, but it's not very likely.

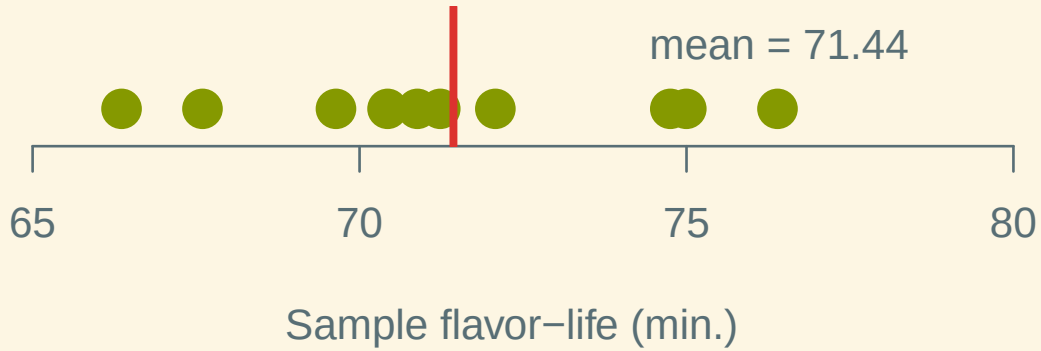




Variability due to Sampling

But if the population looked like this instead...





then the first three samples are unlikely, whereas the last one is more likely.

Variability due to Sampling

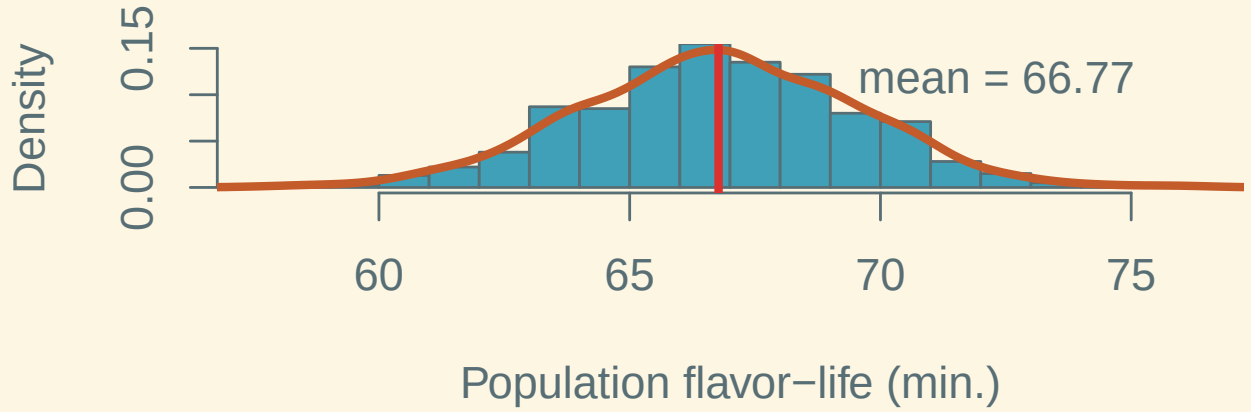
- Each sample differs from the population, so sample information is an imperfect reflection.
- However, there is information about the population, since some populations are more likely than others to produce the given sample.
- If we imagine a continuum of populations (or just population means), some are more plausible than others *because they make the data more likely*.

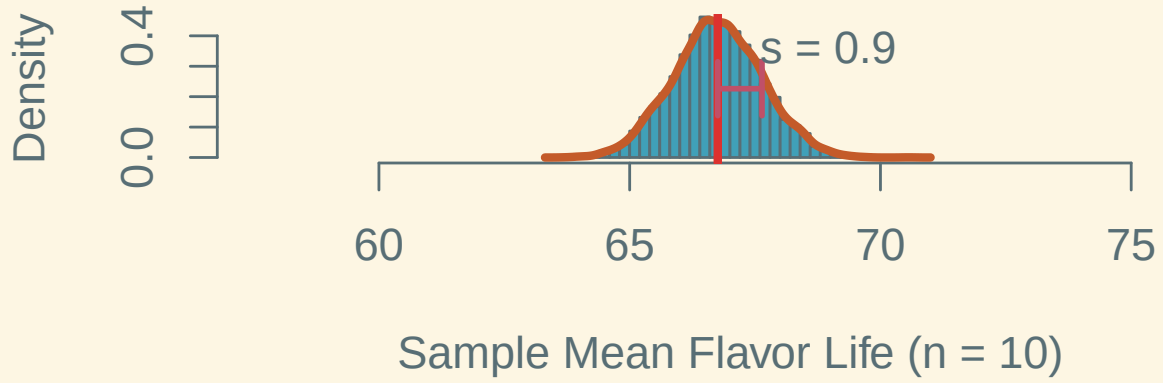
Sampling Distributions

Sampling Distribution Definition

Consider all possible random samples of a fixed size, n from a population. Each one has its own value for a particular **statistic** (like $\bar{x}$). A **sampling distribution** is the collection of all of of those $\bar{x}$ values (or whatever the statistic is)

Sampling Distribution of Gumball Means





Demo: *StatKey*

`http://lock5stat.com/statkey`

Self-Check

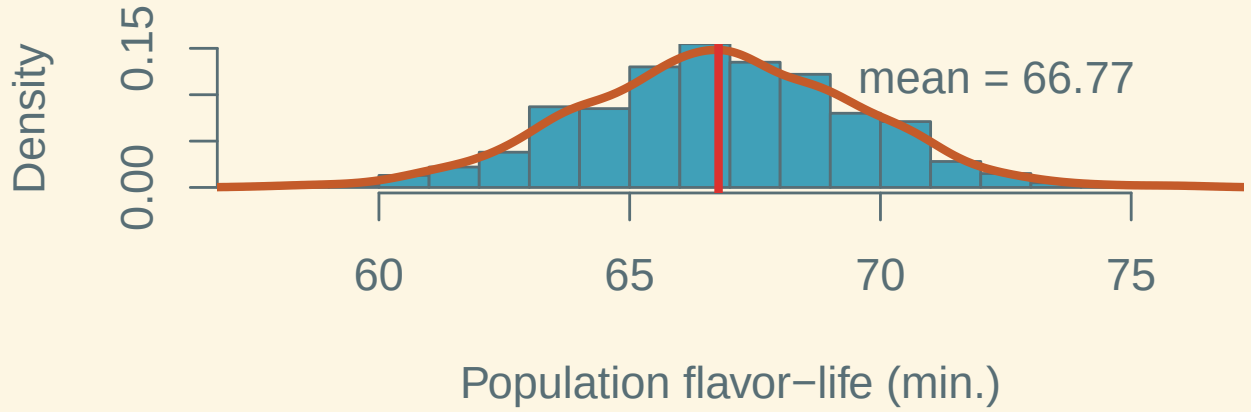
1. What are the cases in the context of a sampling distribution?
   Possible samples of a fixed size n
2. What is the variable in the relevant sampling distribution for the gumball life example? Each case has its own **sample mean**

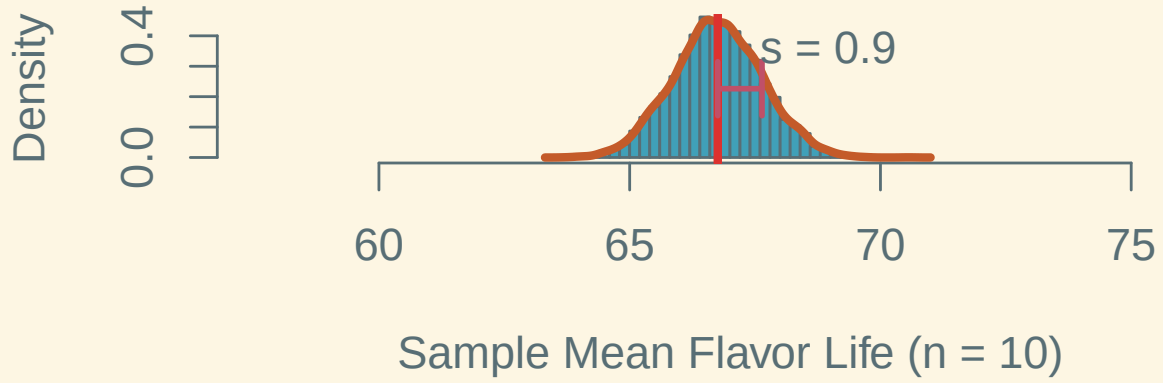
Standard Error

Standard Error Definition

The distribution of a quantitative variable has a standard deviation. The **sampling distribution** of a quantitative *sample statistic* (like a mean) has a standard deviation too. This has a special name: the **standard error** (e.g., "of the mean").

Sampling Distribution of Gumball Means





Properties of Sampling Distributions

Most (about 95%) of *simple random* samples have a sample mean ($\bar{x}$) which is within 2 Standard Errors of the population mean (μ).

So, if I have a sample mean, $\bar{x}$, there is a good chance the population mean, μ is within 2 Standard Errors in either direction.

So I can *estimate* that the population mean is between $\bar{x} - 2SE$ and $\bar{x} + 2SE$. This statement should be correct about 95% of the time.