Quantitative Sociological Analysis

Sample Data & Uncertainty
A Foundation in Probability

Part 5

February 27-?, 2025

Science: a **process** of organizing, and acquiring new, knowledge

Steps in the process

1. Start with a perspective
2. Select a theory
3. Derive a research proposition
4. Derive a research question
5. Derive a hypothesis
6. Find or collect data
7. Analyze data
8. Report results & Answer question
9. Interpret results in terms of theory
10. Draw implications for theory

Without knowing this, what follows is worse than lies or even damn lies

Typically, a sample

There is uncertainty in sample data. Probability can help us understand why and how statistics are used to address this uncertainty

Part 5: preface

- I’m going to throw a lot out there to see what we, on average, pick up on
  - if you feel like you’re getting overwhelmed, refocus on the key takeaways expressed throughout
    - Stop me and ask questions or write down questions and follow up with me.

- There will be many different equations, which you will not be expected to memorize
  - if you find equations overwhelming, focus on understanding the concepts that they express

- There will be some math, which you will not be tested on nor expected to recompute
  - if you find the math overwhelming, just know it generally shows how and why concepts can be applied

- Consider throughout how this material forms a foundation for extending descriptive statistics so they can be applied to the broader population
  - Let’s see how this goes, and I’ll design Exam 2 accordingly to meet us where we’re at
    - This is another reason why stopping me and asking questions and/or following up with me is critical

Part 5

Learning objective: begin to understand that sample data have uncertainty due to chance, which must be addressed to make generalizable statements that can be applied to the broader population

recognize how:

probability theory underlies sampling

the Central Limit Theorem (CLT) connects probability and sampling

differences between a population and sample due to chance can be addressed

Takeaway: descriptive statistics help summarize sample data, but they cannot produce generalizable conclusions because they do not account for sampling variability

Data

- Population: entire group of individuals or entities of interest
    - e.g., if you're interested in college student's mental health, then the population might be all college students…in the world, in a continent, in a specific country, in a region or state within a country, in a certain group of universities, in a specific university, etc.
        - it's often impractical to study an entire population, even at relatively small levels

- Sample: subset of a population selected for data collection
    - Sampling: process of selecting a subset of individuals or entities from a population
        - goal is to draw a sample from which generalizable conclusions can be made

Sample data and quantitative methods

- Descriptive statistics are used to summarize data, but they are
  - unable to produce generalizable conclusions, inferences, because this
    - requires us to know how likely it is that findings can be applied to the broader population

- Inferential statistics are used to quantify the likelihood that a sample statistic (e.g., $\bar{x}$) approximates its true population parameter (e.g., μ), because they
  - enable us to address a key source of uncertainty
    - sampling variability: difference in sample vs population characteristics due to chance

- Thus, sampling strategies must ensure that observations are selected in a way that allows for the *probability* of obtaining a representative sample
  - subset of observations that is characteristically comparable to its population

Probability

- mathematical measurement of the chance that an event will occur
  - essential for addressing sampling error
    - difference between a sample statistic and the true population parameter
      - e.g., the likelihood that a sample mean (e.g., $\bar{X}$) contains the population mean (e.g., μ)

- Simple random sampling (SRS): every individual or entity in the population has an equal probability of being selected
  - Many different sampling strategies, but beyond scope of this course.
    - Maybe learned in methods or other courses?

- Consider the following example based on SRS, which
  - lays groundwork for beginning to understand inferential statistics
    - Note: SRS is conceptually simple and effective, but it is often practically unreasonable

Example preface

- I'm going to use some math and equations
  - to demonstrate how probability is foundational for inferential statistics

- You will not be tested on the math nor equations
  - although some might find this useful

Learning objective

- Begin to understand why uncertainty in sample data must be accounted for
  - if we wish to produce generalizable conclusions
    - make statements about a broader population with some degree of certainty

SRS and probability: example

- Population of interest: a box of ten marbles
- Characteristics of interest: marble color
  - 5 black marbles and 5 white marbles
    - Note: often rely on Census data to inform sampling strategies of population characteristic

- Due to resource constraints, we can only draw 4 samples
  - What is the probability that the sample is representative?
    - proportional to the population in terms of characteristics of interest
      - color: 0.50 black and 0.50 white

SRS and probability: example continued

- First: How many possible samples of 4 are there in a population of 10?
  - since order doesn't matter in the present example, we will use the combination formula

$$C(n,x) = \binom{n}{x} = \frac{n!}{x!\,(n \ - x)!}$$

Side Note that may be of interest to folks who plan to move on to more advanced quantitative methods courses:

Recall, in Exercise 3, how we computed the number of variables necessary to account for all possible rankings of 5 genres.

Since order mattered in that example, we used the permutation formula to determine this would require 120 variables.

$$P(n,n) = \frac{n!}{(n \ - n)!} = n!$$

SRS and probability: example continued

$$C(10,4) = \binom{10}{4} = \frac{10!}{4!\,(10\ -4)!} = \frac{3628800}{17280} = 210$$

- Second: adjust for matching proportional population characteristics
  - 0.50 black and 0.50 white. n = 4, so choose 2 black and 2 white
    - ways to choose 2 black: $\binom{5}{2} = \frac{5!}{2!(5\ -2)!} = 10$
    - ways to choose 2 white: $\binom{5}{2} = \frac{5!}{2!(5\ -2)!} = 10$

- Third: calculate the total ways: $\binom{5}{2} \times \binom{5}{2} = 10\ \times 10 = 100$

SRS and probability: example continued

- Finally: adjust for all other possible samples of 4, and
  - compute probability of each respective possibility

# black	# white	possible samples	probability of selecting
4	0	$\binom{5}{4} \times \binom{5}{0} = \quad 5$	$\frac{5}{210} = 0.024$
3	1	$\binom{5}{3} \times \binom{5}{1} = \quad 50$	$\frac{50}{210} = 0.238$
2	2	$\binom{5}{2} \times \binom{5}{2} = 100$	$\frac{100}{210} = 0.476$
1	3	$\binom{5}{1} \times \binom{5}{3} = \quad 50$	$\frac{50}{210} = 0.238$
0	4	$\binom{5}{0} \times \binom{5}{4} = \quad 5$	$\frac{5}{210} = 0.024$

Note: $0.024 + 0.238 + 0.476 + 0.238 + 0.024 = 1.000$

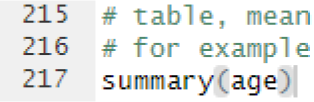
All probabilities naturally
range from 0 to 1

$0 \leq P(A) \leq 1$

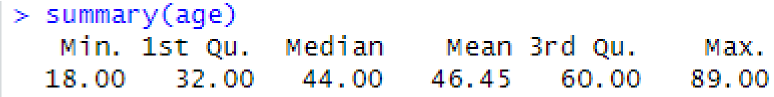
Building a foundation for inferential statistics

- The GSS data we’re using is a nationally representative sample
    - actually, each wave (i.e., 1972-2022) composed of a nationally representative sample
        - respective to year-specific population characteristics

- Recall descriptive statistics unable to quantify uncertainty, which is required to
    - make statements about a broader population with some degree of certainty



```
# table, mean, summary, and sd commands unable to quantify uncertainty
# for example...
summary(age)
```



```
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   32.00   44.00   46.45   60.00   89.00 
```

- To produce generalizable conclusions about age, must account for differences
    - between its sample statistic (e.g., $\bar{x}$) and population parameter (e.g., μ)
        - e.g., to quantify the likelihood that the mean age of adults in the US is 46.45
            - it will take some additional work to build up to this understanding…

Sample space

- set of all possible outcomes of a random trial
    - e.g., selecting a respondent with certain characteristics, like a specific age, to be included in the GSS
        - 18, 19, 20, 21, … 89

- Probability distribution: assigns probabilities to the outcomes in a sample space
    - defined by either a probability density or mass function, whether random variable is
        - continuous: takes on an infinite number of values within a range (e.g., interval-ratio variables)
        - discrete: takes on a countable number of values (e.g., nominal/ordinal variables)

Beyond the scope of this course: there are many different types of probability distributions
Continuous: e.g., normal, exponential, uniform, gamma, beta. Discrete: e.g., binomial, Poisson, geometric, negative binomial.

Key takeaway: because probability distributions have known properties they can be converted into a normal distribution, which has special properties that enable us to address uncertainty in sampling due to chance by quantifying likelihood

*this explanation overlooks some details and caveats that are not necessary to address at this entry level

Let’s work through a hypothetical example to begin to understand this key takeaway…

Probability and sample space: an example

- The nature of the sample space determines which probability function to use
  - Let’s see how a function is used to assign probabilities to a random variable within a sample space, and
    - establish the properties that define its probability distribution

Again, the details and equations in this example are simply to help us build up to an understanding of how we can use the normal distribution to address uncertainty in sampling due to chance by quantifying likelihood

- Let’s consider a coin flip (i.e., random variable)
  - sample space: $S = \{Heads, Tails\}$
- with 20 flips (i.e., random trials)
  - $S = 220$, so there are 1,048,576 possible sequences of 20-coin flips

Binomial probability distribution: example cont.

- we will use a binomial mass function to compute the probabilities assigned to this probability distribution, since
  - only two possible outcomes per trial (e.g., H, T), a fixed number of trials (e.g., 20), independence of trials (e.g., outcome of one flip does not affect outcome of any other flip), and constant probability of success (e.g., fair coin)

Note: above are just details that show one example for how the nature of a sample space determines which probability function to use

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$P(X = k)$ is the probability of getting exactly k success (e.g., heads)

n is the number of trails (e.g., how many coin flips)

k is the number of success (e.g., how many flips were heads)

p is the probability of success on a single trial (e.g., with a fair coin this is 0.5)

$\binom{n}{k}$ is the binomial coefficient: the number of ways to choose k success from n trials

$(1 - p)$ is the probability of failure (e.g., tails)

Binomial probability distribution: example cont.

- What is probability of observing 10 heads (successes) in 20 flips of a fair coin?

Anyone know what this probability is?

Binomial probability distribution: example cont.

- What is probability of observing 10 heads (successes) in 20 flips of a fair coin?

$$P(X = 10) = \binom{20}{10} 0.5^{10}(1 - 0.5)^{20-10}$$

$$= \left(\frac{20!}{10!\,(20 - 10)!}\right) 0.000977(0.5)^{10}$$

$$= 184{,}756 \times \frac{1}{1{,}048{,}576}$$

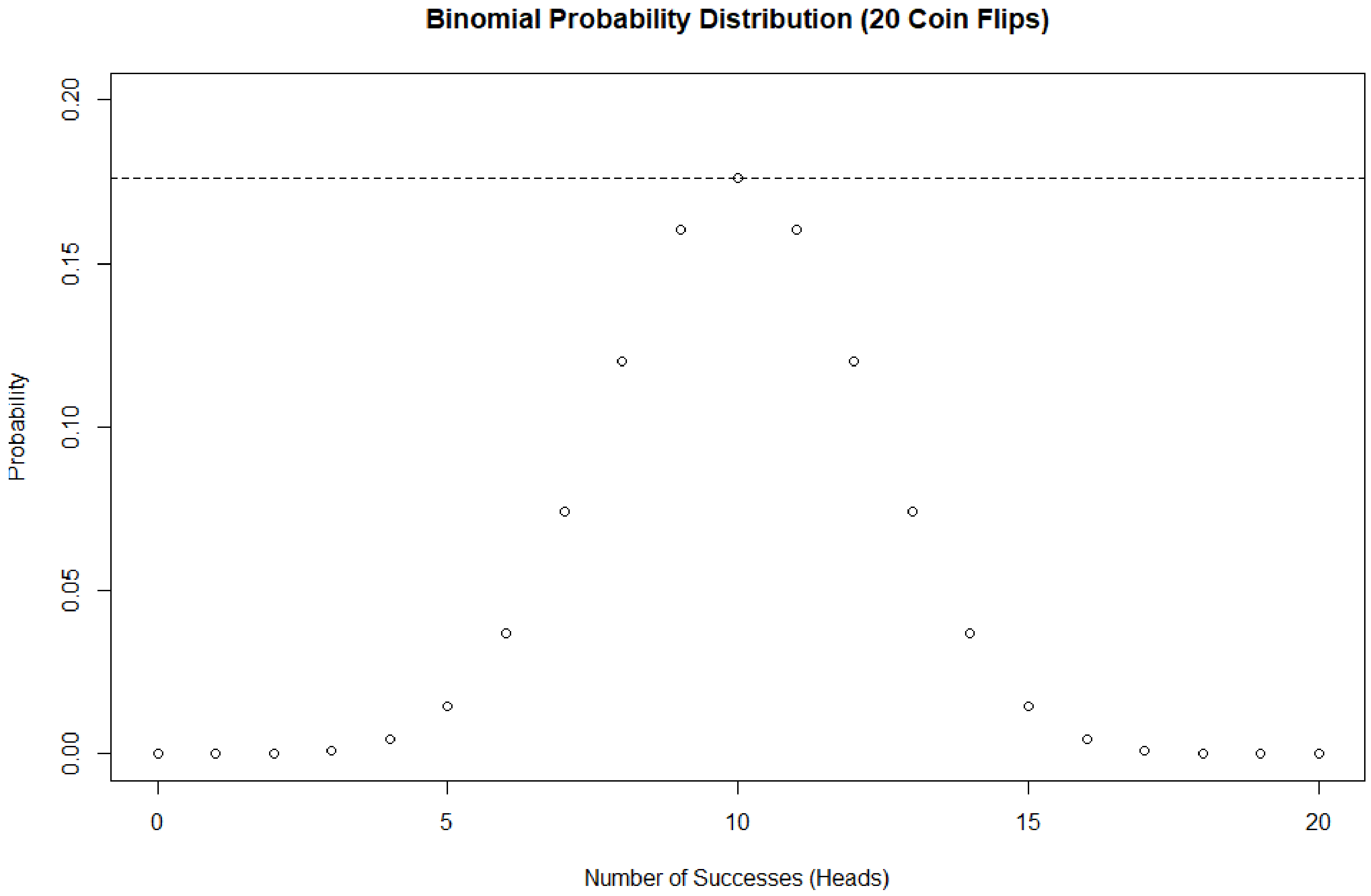
Does this denominator look familiar?

$$= 0.176$$

Recall: $S = 220$, so there are 1,048,576 possible sequences of 20-coin flips

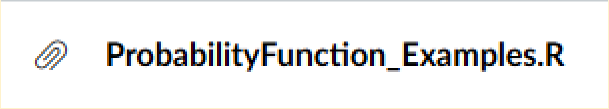
Let’s check out the distribution of probabilities for all possible success (number of heads) in 20 flips…

Binomial probability distribution: example cont.



If we summed each probability for all possible success, what would this equal and how do you know this?

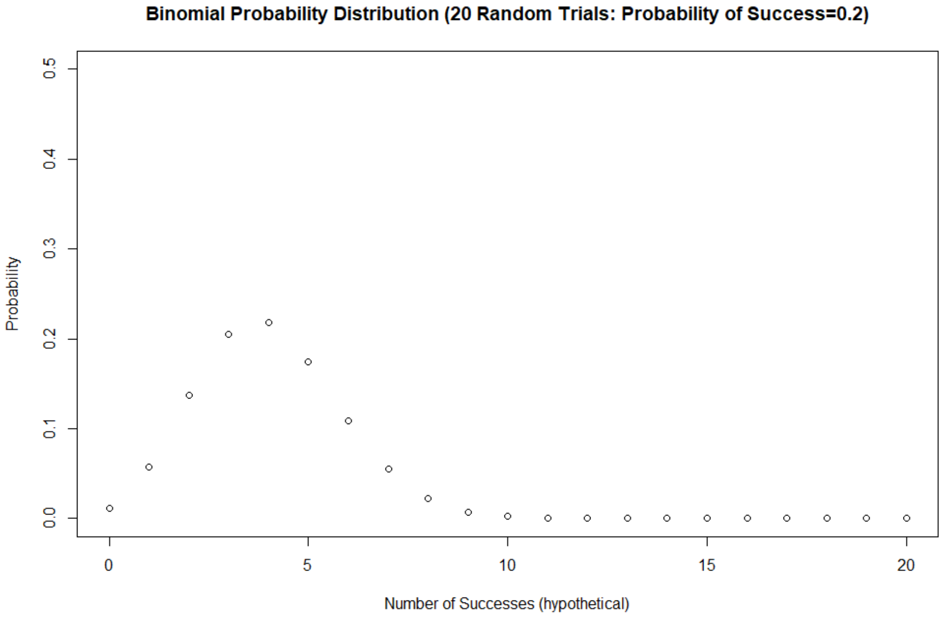
RScript for this example, including plot, available on Canvas if interested

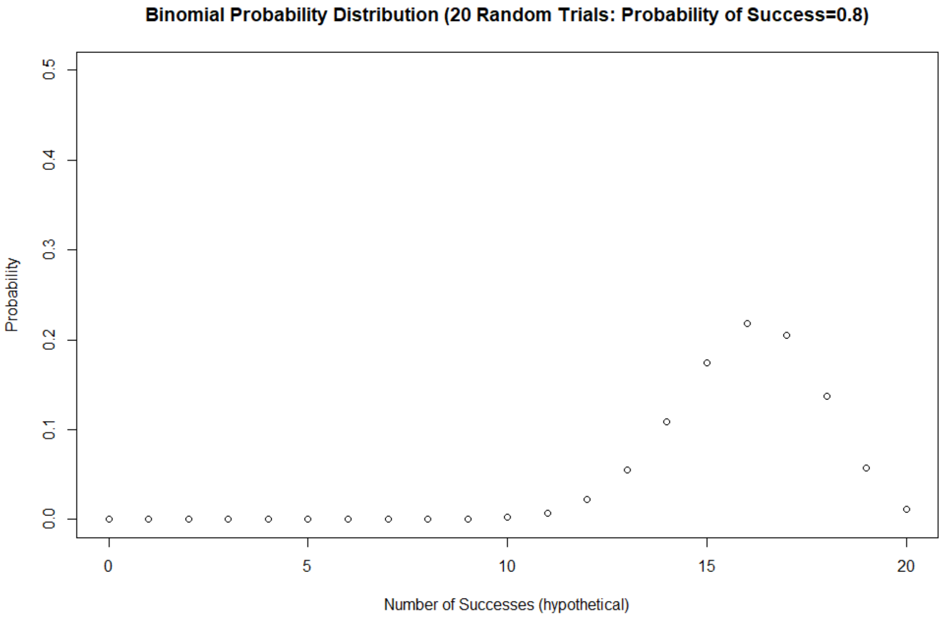


ProbabilityFunction_Examples.R

Probability distributions, some more insights

- The binomial probability distribution we computed was bell shaped
  - we obtained a bell curve because the probability of success in a single trial was 0.5
- Why do binomial distributions with a probability of success in a single trial below vs above 0.5 look different?
  - Recall with descriptive statistics of dispersion for interval-ratio variables that when
    - median less than mean then negative/left skewness
    - median greater than mean then positive/right skewness





The properties of a probability distribution are determined by the nature of its sample space, and not by its shape

Let’s consider the normal probability distribution and its density function, determined by continuous sample spaces…

Normal probability distribution: example

- Let’s consider height of adult males in the US (i.e., random variable)
    - sample space: $S = \{shortest\ possible\ inches, tallest\ possible\ inches\}$
- and selecting a male of a certain height from the population (i.e., random trial)
    - theoretical sample space: $S = \{x \in R \mid x > 0\}$
        - unknown because not entirely governed by natural rules of chance

Given that continuous variables (e.g., height) involve an infinite number of values within a range, the probability of any certain value is typically zero. Whereas discrete random variables (e.g., coin tosses) follow defined probability rules, making it possible to compute the probability of any specific value. Thus, we generally must either know or make assumptions about the population characteristics when working with continuous variables.

- Let’s assume population’s mean height ($\mu = 70$ inches) and standard deviation ($\sigma = 3$ inches)
    - Note how these parameters for a continuous variable differ from those used for a discrete variable

Normal probability distribution

- normal density function to compute the probabilities assigned to this probability distribution

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

$f(x)$ is the value of the probability density function at x

µ is mean of the population distribution (location at peak)

σ is the standard deviation of the population distribution

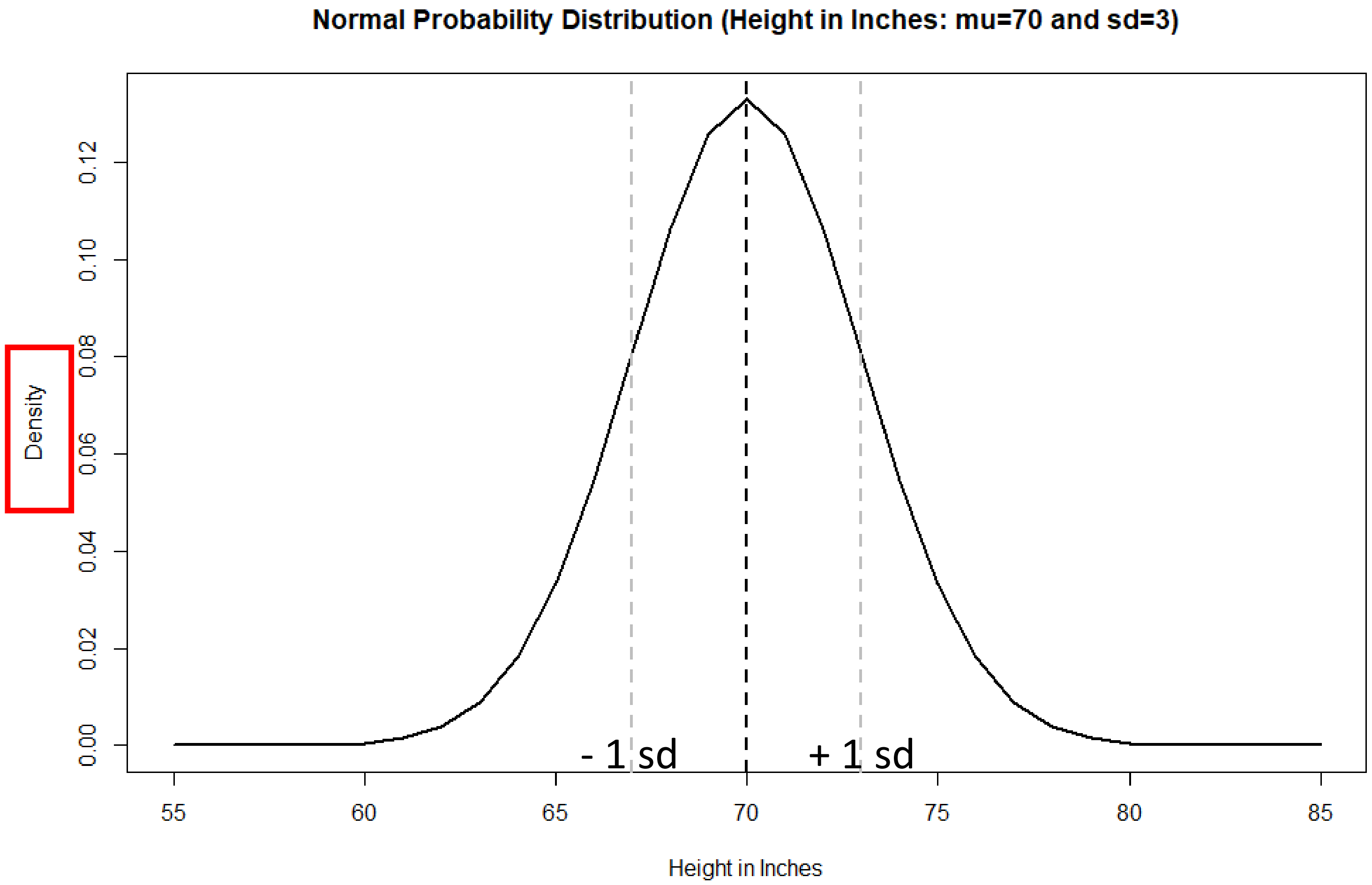
σ^2 is the variance of the population distribution (the square of the standard deviation)

exp is the exponential function (ex)

Let’s apply this to our height ($\mu = 70$ inches, $\sigma = 3$ inches) example…

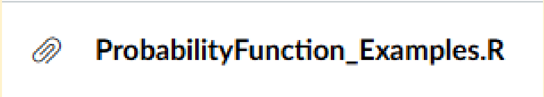
Normal probability distribution: example cont.

- What is probability of selecting an adult male of a certain height from the population?



RScript for this example

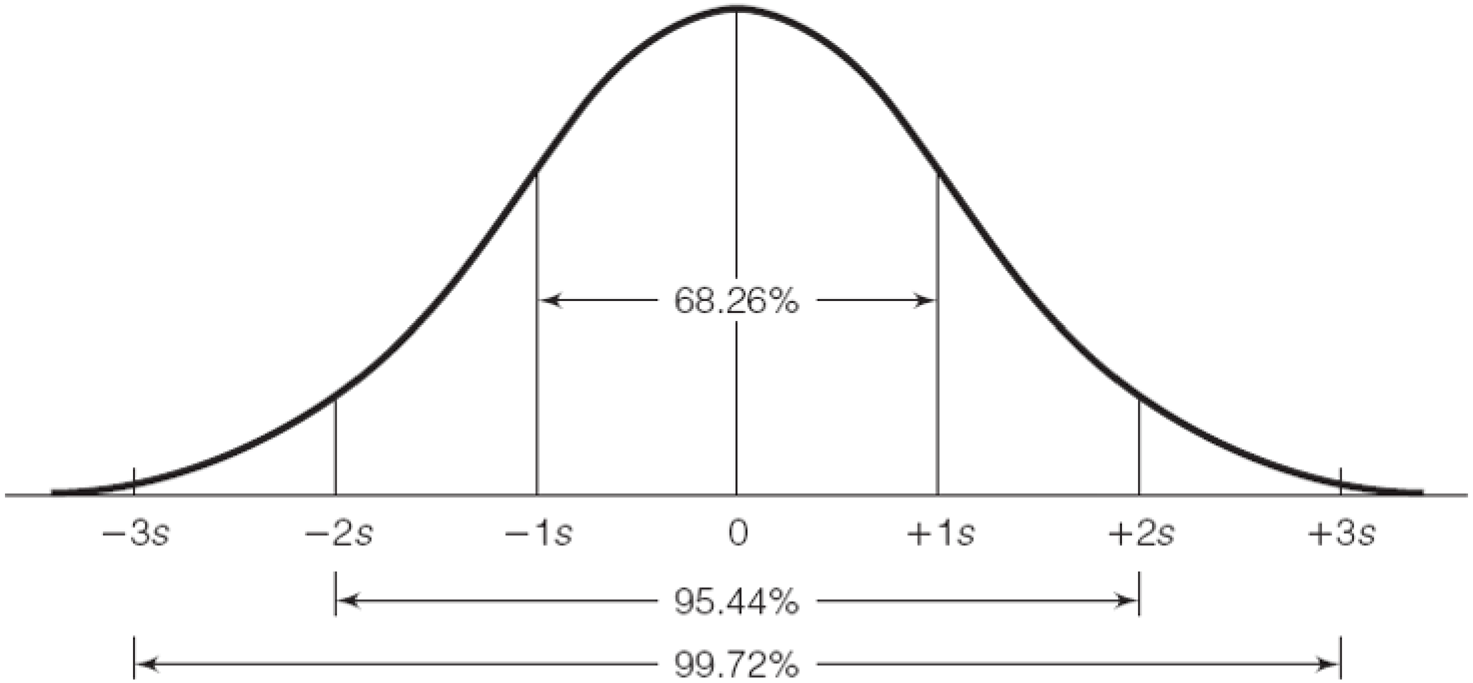
available on Canvas if interested


ProbabilityFunction_Examples.R

The probability of any certain value in a continuous sample space is typically zero, so we are generally more concerned with a range of probabilities that a certain value may fall in. How do we do that?

Theoretical normal distribution

- We can use this distribution's special properties to describe a range of probabilities
  - see how standard deviation (sd) plays an important role in the normal distribution's special properties



This involves some more math, because values will need to be standardized…

Normal probability distribution: Z scores

- values standardized to match the properties of the theoretical normal curve, which
  - enable us to find probabilities inherent in this distribution’s special properties

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

X is the individual data point (e.g., certain height)

μ is the population mean

σ is the population standard deviation

Let’s use our height example to see how

68.26% of the population within ± 1 sd

95.44% of the population within ± 2 sd

99.72% of the population within ± 3 sd

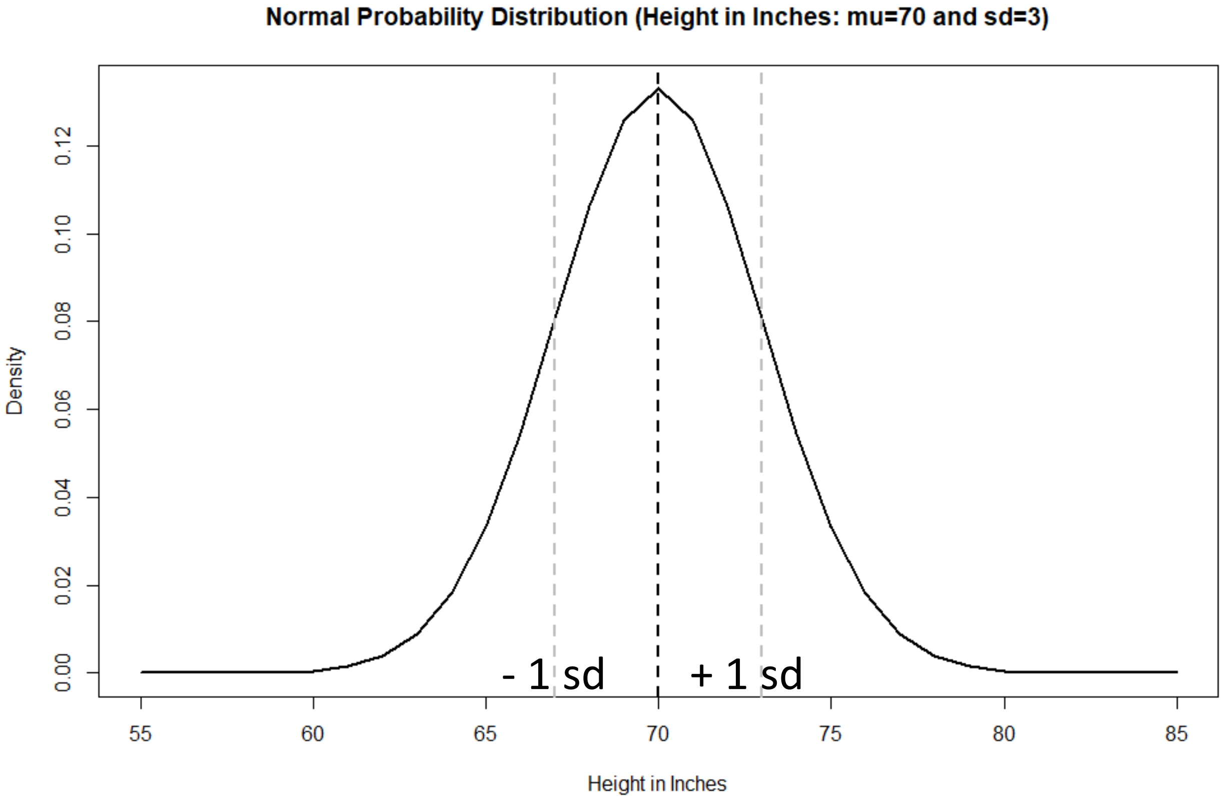
Normal probability distribution: Z scores example

- an American adult male's height of 73 inches falls within 1 sd of the mean

$$Z = \frac{73\ - 70}{3} = \ +1.00$$

- an American adult male's height of 67 inches falls within 1 sd of the mean

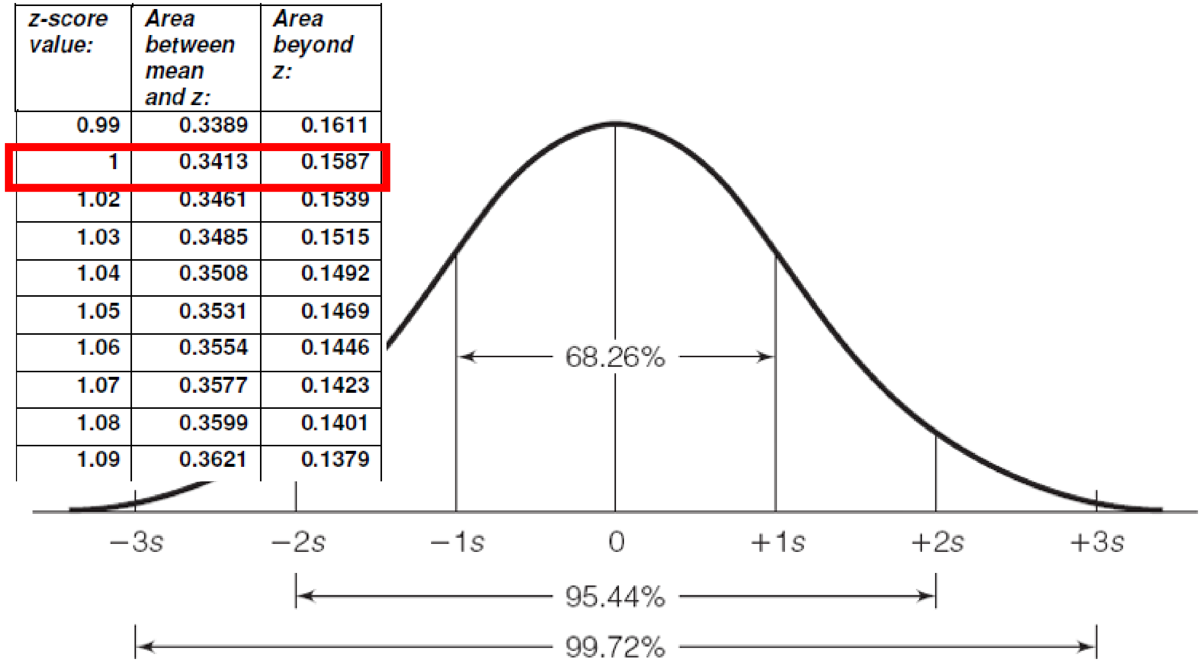
$$Z = \frac{67\ - 70}{3} = -1.00$$



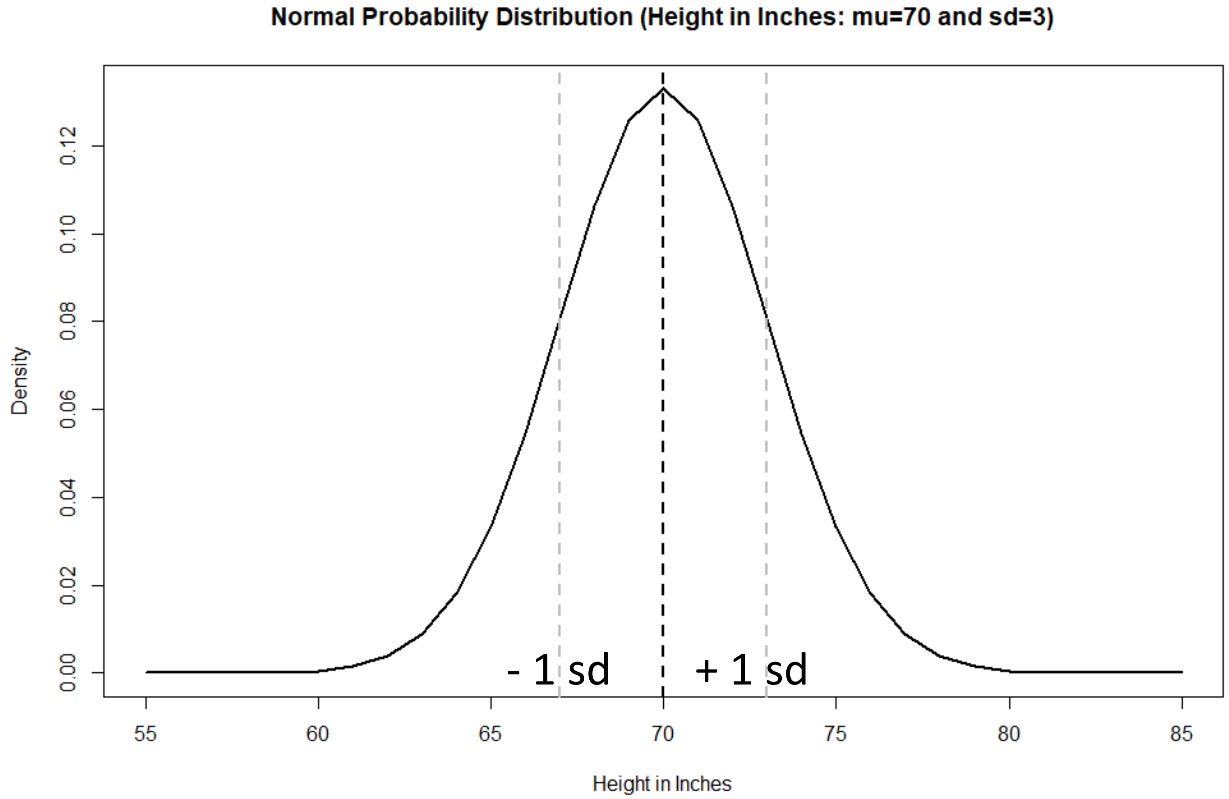
After a value is standardized to match the properties of the theoretical normal curve, the Z-table can be used to find the probability that this value would be selected from a population with the given parameters (i.e., μ and σ).

Normal probability distribution: Z scores example cont.

Z-table available on Canvas if interested



z-score value:	Area between mean and z:	Area beyond z:
0.99	0.3389	0.1611
1	0.3413	0.1587
1.02	0.3461	0.1539
1.03	0.3485	0.1515
1.04	0.3508	0.1492
1.05	0.3531	0.1469
1.06	0.3554	0.1446
1.07	0.3577	0.1423
1.08	0.3599	0.1401
1.09	0.3621	0.1379



The probability of a male being taller than 73-inches is 0.1587

The probability of a male being shorter than 67-inches is 0.1587

The probability of a male being between 67-73-inches is 0.6826

1.0000

Takeaway: begin to understand that the normal distribution has special properties that can be broadly used to address uncertainty in sampling due to chance by quantifying likelihood

Approximating the normal probability distribution

- Many probability distributions can be approximated to the normal probability distribution
  - For example, the binomial probability distribution

$$Z = \frac{(x \pm 0.5) - np}{\sqrt{np(1-p)}}$$

x is number of success in the binomial distribution

n is sample size

p is probability of success

np is expected count of success given a certain sample size

$np(1-p)$ is the variance of a count. Note: $p(1-p)$ is the variance of a proportion

± 0.5 is continuity correction factor

Let’s use our coin flip example to see how this works…

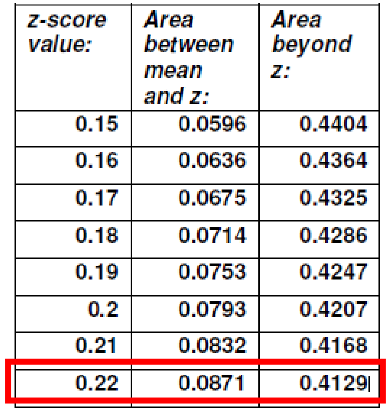
Approximating the normal probability distribution: example

- What is probability of 10 heads (successes) in 20 flips of a fair coin?

$$Z = \frac{(10-0.5)-10}{\sqrt{10(1-0.5)}} = \frac{-0.5}{2.236} = -0.224 \text{ for } P(Z \leq -0.224) \approx 0.4129$$

$$Z = \frac{(10+0.5)-10}{\sqrt{10(1-0.5)}} = \frac{0.5}{2.236} = 0.224 \text{ for } P(Z \geq 0.224) \approx 1 - 0.4129 \approx 0.5871$$

$$P(9.5 \leq X \leq 10.5) = 0.5871 - 0.4129 = 0.1742$$



z-score value:	*Area between mean and z:*	*Area beyond z:*
0.15	0.0596	0.4404
0.16	0.0636	0.4364
0.17	0.0675	0.4325
0.18	0.0714	0.4286
0.19	0.0753	0.4247
0.2	0.0793	0.4207
0.21	0.0832	0.4168
0.22	0.0871	0.4129

Would have been closer to 0.176 if used exact Z score, but Z = 0.224 not available in this table

Overall takeaway: the probability of sample characteristics matching its population is fundamental to understanding uncertainty due to chance. The normal probability distribution has special properties that can be used to help quantify and make sense of this uncertainty.

Thus far, we've only considered

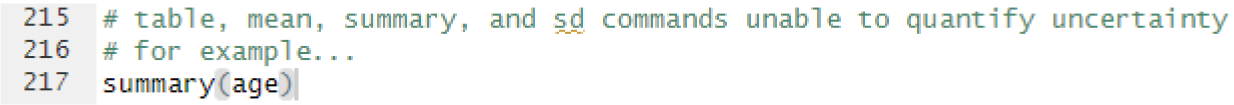
- the probability of sampling specific values of a variable
  - While useful, we are also interested in summaries of those values
- Let's build from here to see how we can produce generalizable conclusions
  - For example, quantifying the likelihood that a sample's descriptive statistics can be applied to the broader population from which they were collected

Recall the following slide…

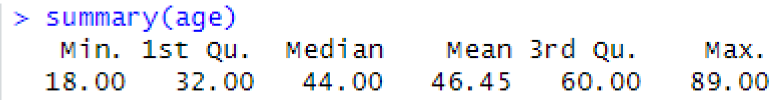
Building a foundation for inferential statistics

- The GSS data we’re using is a nationally representative sample
  - actually, each wave (i.e., 1972-2022) composed of a nationally representative sample
    - respective to year-specific population characteristics

- Recall descriptive statistics unable to quantify uncertainty, which is required to
  - make statements about a broader population with some degree of certainty



```
# table, mean, summary, and sd commands unable to quantify uncertainty
# for example...
summary(age)
```



```
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   32.00   44.00   46.45   60.00   89.00 
```

- To produce generalizable conclusions about age, must account for differences
  - between its sample statistic (e.g., $\bar{x}$) and population parameter (e.g., μ)
    - e.g., to quantify the likelihood that the mean age of adults in the US is 46.45
      - it will take some additional work to build up to this understanding…

Sampling distribution

- The theoretical probability distribution of a statistic based on all possible samples of a given size from a population
  - a probability distribution that assigns probabilities to sample statistics based on repeated random sampling

- Imagine randomly selecting a sample of (n) people from a population (N)
  - computing a sample mean $(\bar{x})$,
  - replacing the sample back into the population, and
  - repeating this until every possible combination of samples is complete $(\bar{x}_1 \dots \bar{x}_n)$

Properties of the Sampling Distribution

The distribution of sample means $(\bar{x}_1 \dots \bar{x}_n)$, would have a

1. mean $(\bar{X})$ equal to the population mean (μ)

2. standard error (σ_x) equal to the population standard deviation divided by square root of N $\left(\frac{\sigma}{\sqrt{N}}\right)$,

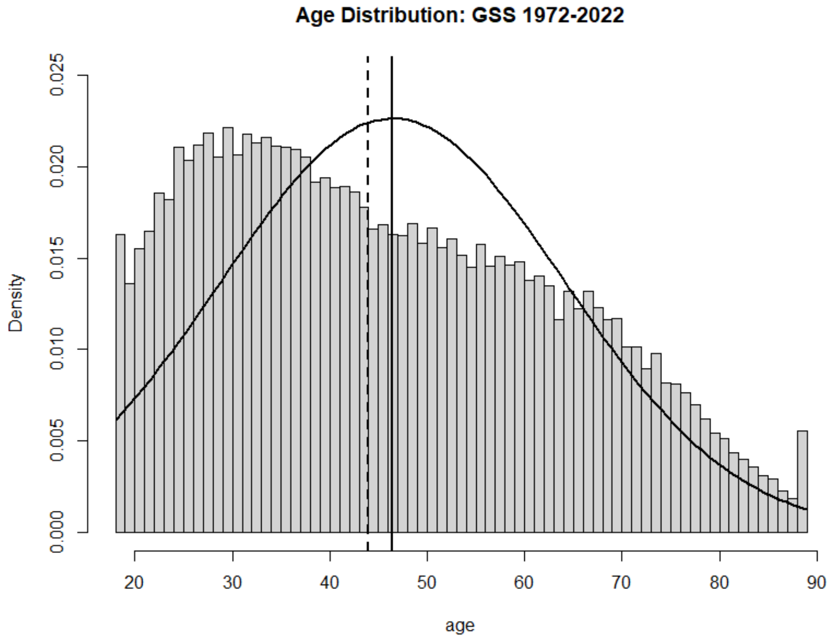
3. normal shape given $n \geq 100$

Central Limit Theorem (CLT)

- The sampling distribution of the sample mean approaches a normal distribution as the sample size increases, when samples are randomly selected and independent
  - If random samples of size N ($\geq$ 100) are drawn from any population, then the distribution of sample means $(\bar{x}_1 \ldots \bar{x}_n)$ will converge on a bell shape with a mean $(\bar{X})$ of μ and a standard error (σ_x) of $\left(\frac{\sigma}{\sqrt{N}}\right)$
    - Note how probability theory and the CLT together help establish key properties of the sampling distribution

This holds for any variable, even variables that are not normally distributed in the population.

Let's see how this works with our age example from the GSS, which we already know is a variable that is not normally distributed – it's positively (right) skewed.

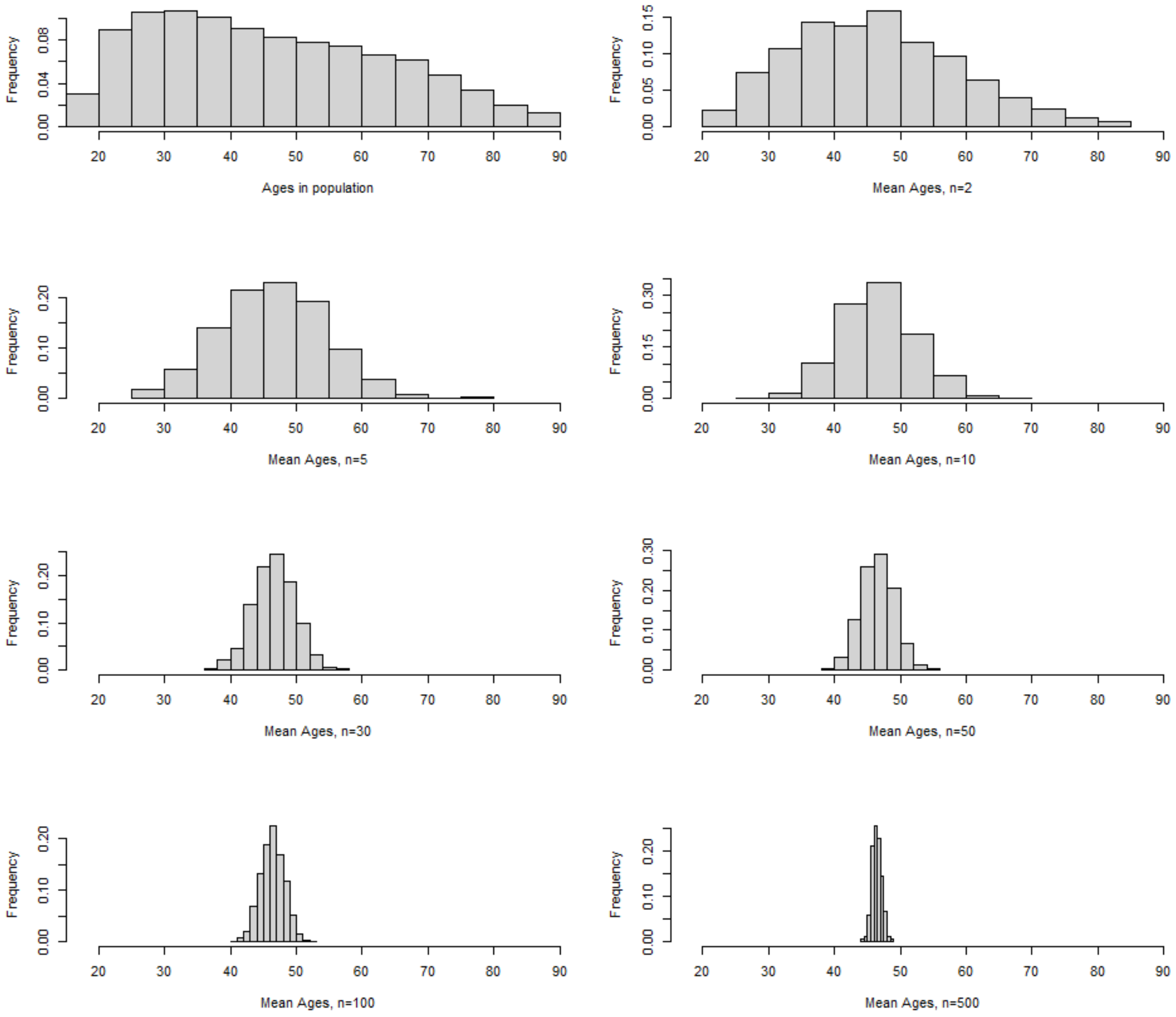


Sampling Distribution and CLT: example

- Let's consider the GSS as if it were the population
  - from which random samples will be repeatedly selected
    - same data as previous histogram, just binned into age groups
- Randomly select 1,000 samples of size n_i
  - impractical to select all possible combinations of sample size n_i,
    - so arbitrarily set number of samples to a reasonably large amount
- Compute the mean for all 1,000 samples of size n_i
  - $\bar{X}_{n_i} = \frac{(\bar{x}_{ni1} + \bar{x}_{ni2} + \cdots \bar{x}_{ni1000})}{1000}$
- Repeat this for different sample sizes n_i
  - $n_2, n_5, n_{10}, n_{30}, n_{50}, n_{100}, n_{500}$

See how approaches normal and converges around 46.45 (μ) as n_i increases

*since based on sample data, this is not an effective example for demonstrating how σ_x converges around $\left(\frac{\sigma}{\sqrt{N}}\right)$ as n_i increases

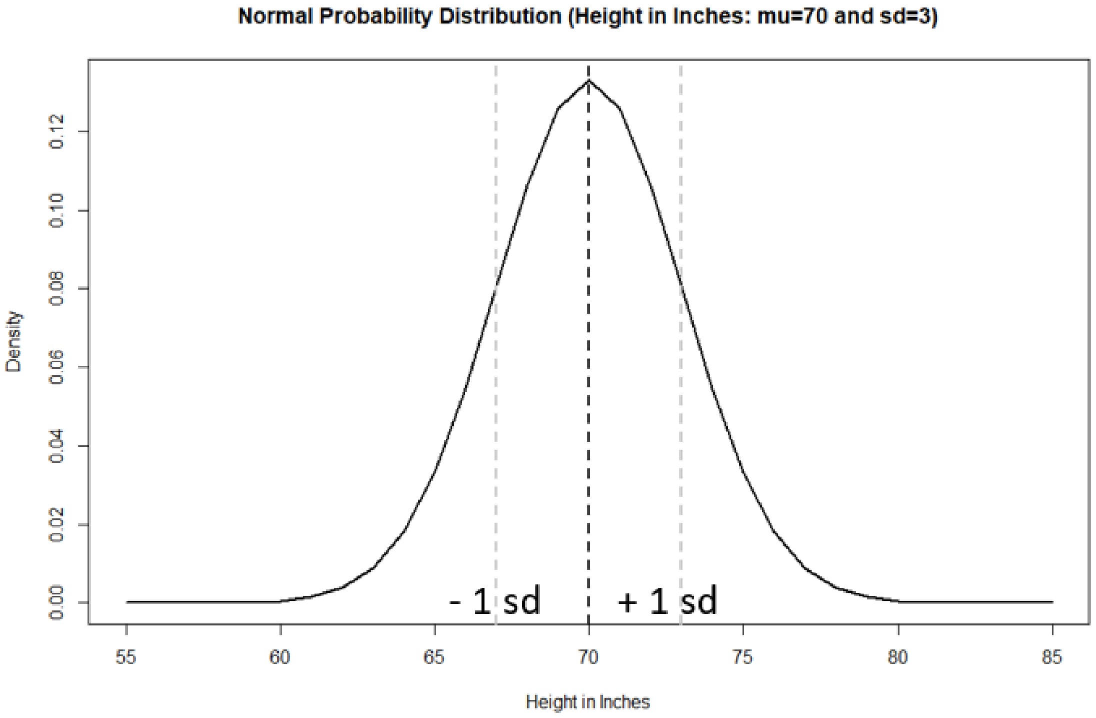


RScript for this example available on Canvas if interested

SamplingDistributionCLT_Example.R

Sampling Distribution and CLT: Z scores

- since the sampling distribution approximates the standard normal distribution
  - we can compute the probability of a sample mean using a modification of the Z formula
    - $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$
- Let’s go back to our height example to see how this can also work for another variable
  - $\mu = 70$ inches, $\sigma = 3$ inches



Sampling Distribution and CLT: Z scores

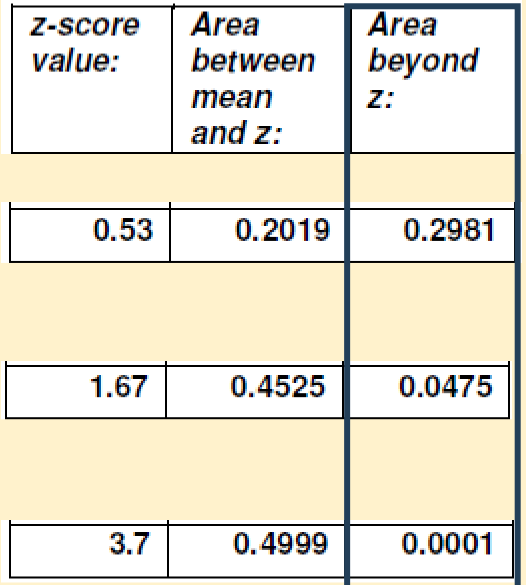
- Let’s consider theoretically randomly selecting 1,000 samples of size n_i from the US adult male population
  - $\boldsymbol{n_{10}}$, $\boldsymbol{n_{100}}$, $\boldsymbol{n_{500}}$
- What is the probability of selecting a sample with a mean $\geq$70.5 inches, and how does this differ by n_i ?

See how narrows and converges around 70 (μ) as n_i increases

$$z_{n10} = \frac{70.5 - 70}{3/\sqrt{1000}} = 0.53$$

$$z_{n100} = \frac{70.5 - 70}{3/\sqrt{1000}} = 1.67$$

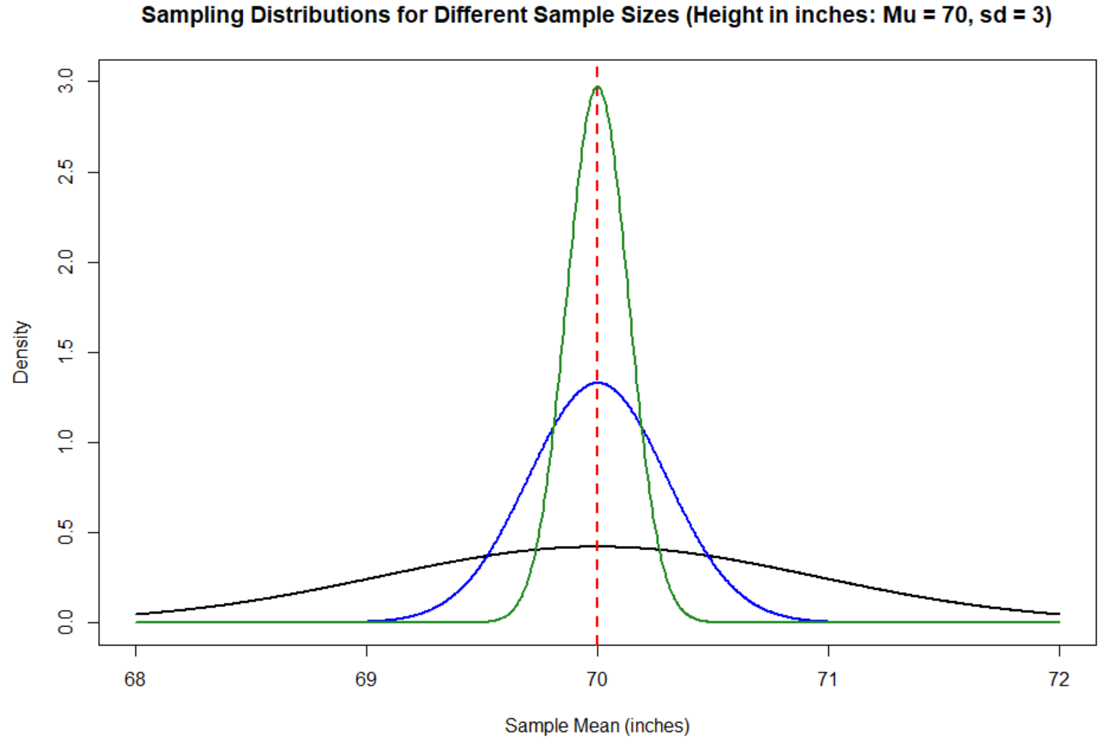
$$z_{n500} = \frac{70.5 - 70}{3/\sqrt{1000}} = 3.73$$



z-score value:	*Area between mean and z:*	*Area beyond z:*
0.53	0.2019	0.2981
1.67	0.4525	0.0475
3.7	0.4999	0.0001

The probability of obtaining a sample mean $\geq$70.5 with n_i □

or, 1 - □ to compute probability of $\leq$ 70.5



Sampling Distributions for Different Sample Sizes (Height in inches: Mu = 70, sd = 3)

Sampling Distribution and CLT: standard error (SE)

- measures the variability of a sample statistic (e.g., $\bar{X}$) across repeated samples
  - helps determine the likelihood that sample results reflect the true population parameter
    - e.g., for the mean ($\bar{X}$)

when parameters are known $SE = \frac{\sigma}{\sqrt{N}}$ or $SE = \frac{s}{\sqrt{N}}$ when parameters are unknown

σ is the population standard deviation

s is the sample standard deviation

N is the sample size

* equation for standard error specific to sample statistic type, not just whether un/known parameters

In general, as sample size increases SE deceases → more precise estimates

Useful for extending the normal distribution’s properties to further understand the likelihood of sample statistics

Z scores and SE: confidence interval (CI)

- range of values within which the estimated parameter is expected to fall
  - with a level of certainty, based on repeated sampling from the population

Places upper and lower bounds around a statistic in a way that expresses likelihood of where true parameter falls

Useful for making generalizable statements about estimates of true parameters, inferences about a population

$$CI = \hat{\theta} \pm Z\left(\frac{\theta}{\sqrt{N}}\right)$$

when parameters are known

or

$$CI = \hat{\theta} \pm Z \times SE$$

when parameters are unknown

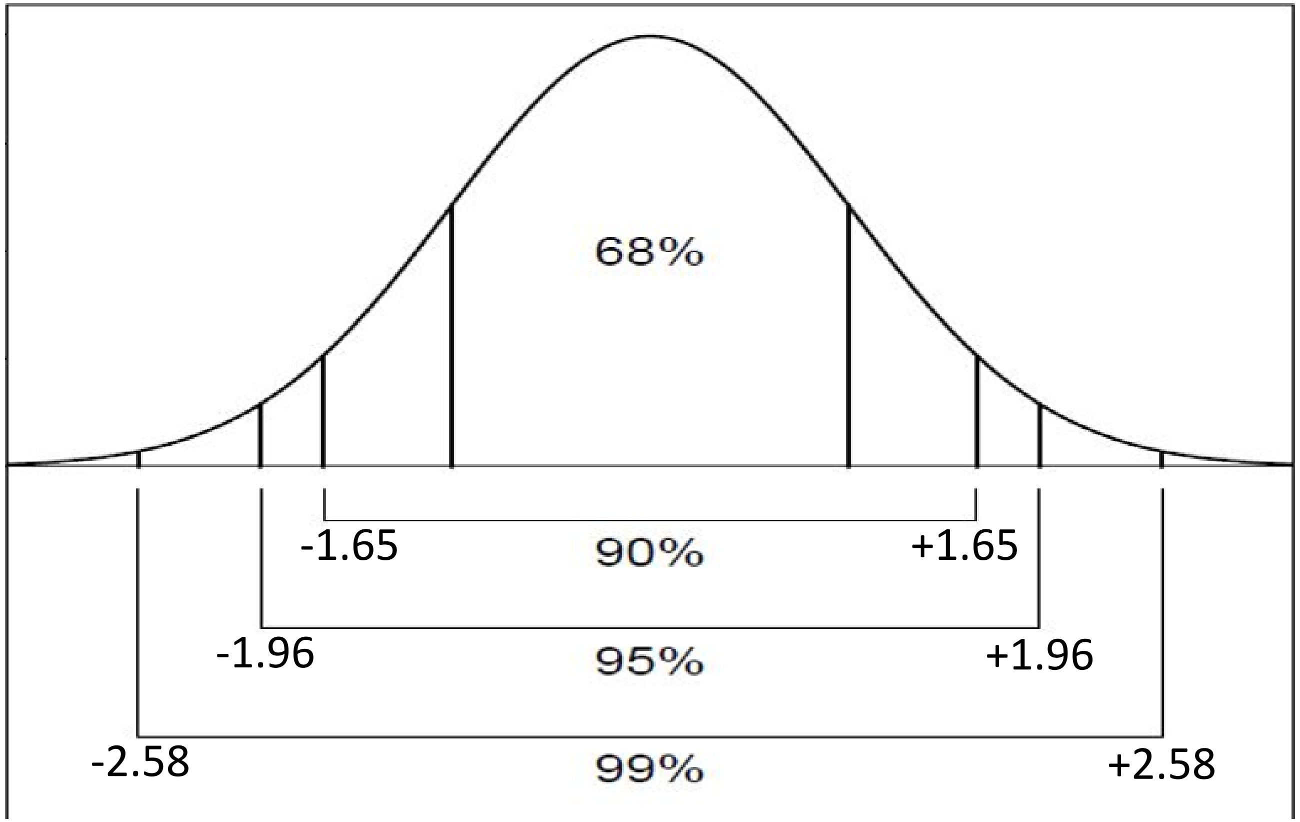
$\hat{\theta}$ is a sample statistic, e.g., mean, proportion, variance, sd

θ is a population parameter, e.g., mean, proportion, variance, sd

Recall how the standard error (σ_x) for the mean ($\bar{X}$) of a theoretical sampling distribution is equal to the population standard deviation divided by square root of N

$$\left(\frac{\sigma}{\sqrt{N}}\right)$$

* parameters do not have error, there is no uncertainty



lower bound Z score

confidence level

upper bound Z score