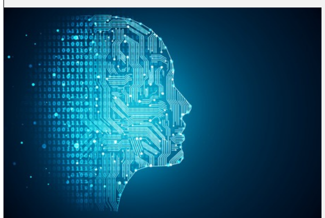


# Data Science for Researchers and Scholars

**Vasant G. Honavar**
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

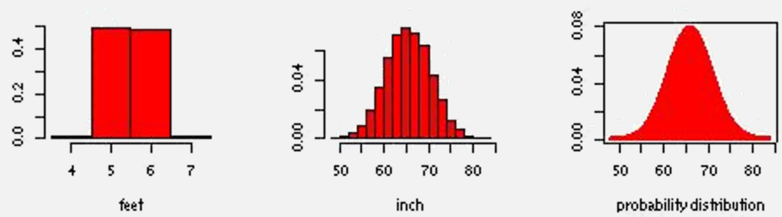
vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

# Continuous Random Variables

- A random variable is continuous if it can assume the infinitely many values corresponding to points on a line interval.
- Examples
  - Height, weight
  - Scores on a test
  - Measurement error

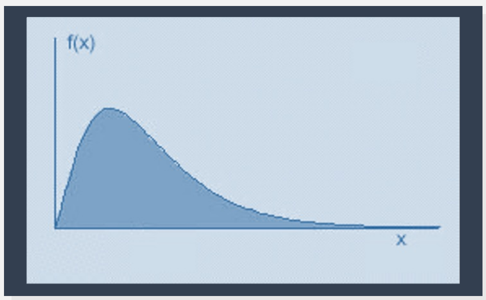
# Continuous Probability Distribution

- Suppose we measure height of students in this class.
- If we “discretize” by rounding to the nearest feet, the discrete probability histogram is shown on the left.
- Now if height is measured to the nearest inch, a possible probability histogram is shown in the middle.
- We get more bins and much smoother appearance. Imagine we continue in this way to measure height more and more finely, the resulting probability histograms approach a smooth curve shown on the right.



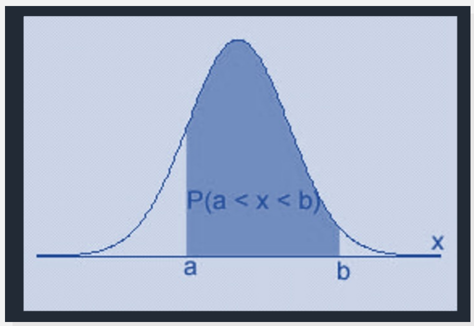
# Probability Distribution of a Continuous Random Variable

- Probability distribution describes how the probabilities are distributed over all possible values.
- A probability distribution for a continuous random variable $x$ is specified by a mathematical function denoted by f(x) which is called the density function.
- The graph of a density function is a smooth curve.

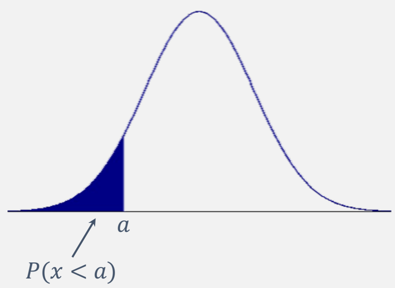


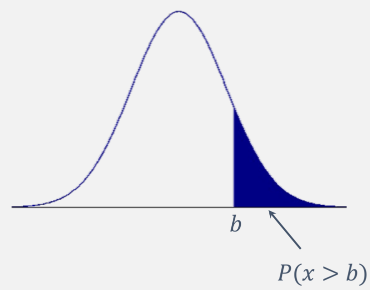
# Properties of Continuous Probability Distributions

- $f(x) \geq 0$
- The area under the curve is equal to **1.**
- $P(a \leq x \leq b)$ $=$ area under the curve between $a$ and $b$.



# Notes




For a continuous random variable $x$,

$$P(x = a) = 0$$

Specifically this means

$$P(x < a) = P(x \leq a)$$

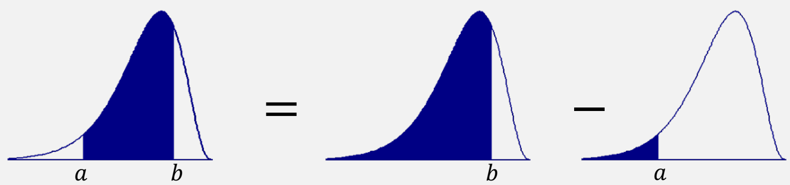
$$P(a < x < b) = P(a \leq x < b) = P(a < x \leq b) = P(a \leq x \leq b)$$

# Method of Probability Calculation

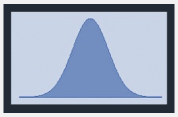
The probability that a continuous random variable $x$ lies between a lower limit $a$ and an upper limit $b$ is

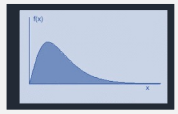
$P(a < x < b)$ = (cumulative area to the left of $b$) – (cumulative area to the left of $a$)

$= P(x < b) - P(x < a)$



# Continuous Probability Distributions

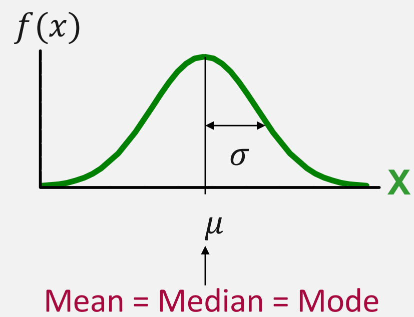




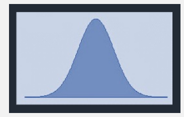
- There are many different types of continuous random variables
- Goal is to pick a model that
  - Fits the data well
  - Allows us to make the best possible inferences using the data.
- Machine learning can be used to fit complex models to data
- One important continuous random variable is the normal random variable.

# The Normal Distribution

- Bell Shaped
- Symmetrical
- Mean, Median and Mode are Equal
- Central tendency is determined by the mean, $\mu$
- Spread is determined by the standard deviation, $\sigma$
- The random variable has an infinite theoretical range: $+\infty$ to $-\infty$



# The Normal Distribution



Normal distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for} -\infty < x < \infty$$

$e = 2.7183 \qquad \pi = 3.1416$

$\mu$ and $\sigma$ are the population mean and standard deviation.

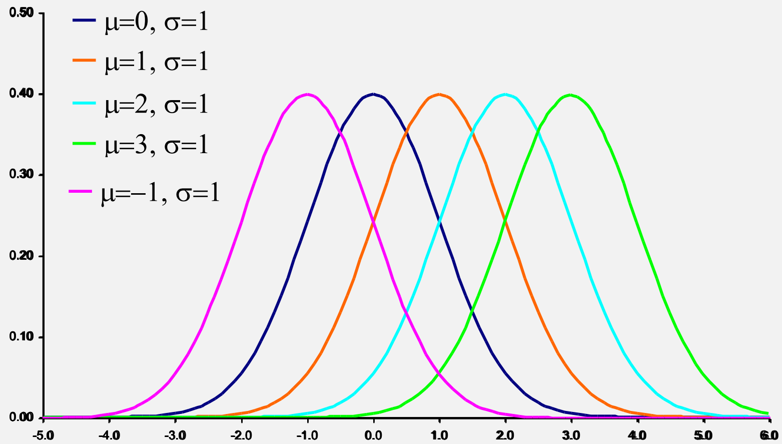
- Two parameters, mean and standard deviation, completely specify the Normal distribution.
- The shape and location of the normal curve changes as the mean and standard deviation change.

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory

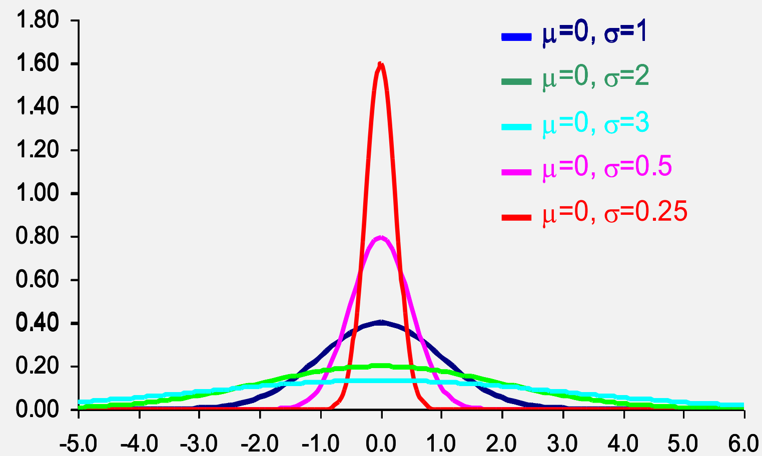
PennState Institute for Computational and Data Sciences

PennState Clinical and Translational Science Institute

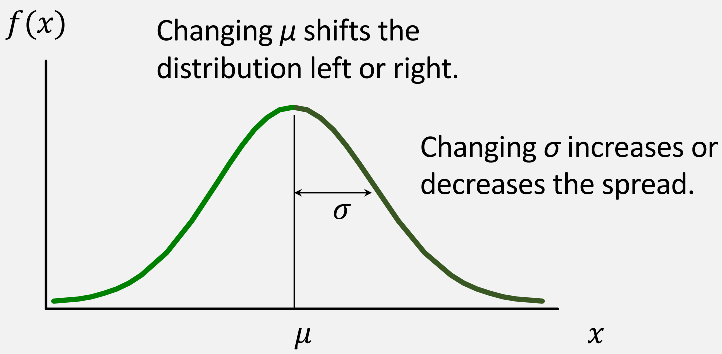
# Normal Distributions: σ=1



# Normal Distributions: $\mu$=0



## The Normal Distribution Shape



# From Normal to the Standardized Normal Distribution

- Translate from $x$ to the standardized normal (the "$z$" distribution) by subtracting the mean of X and dividing by its standard deviation:

$$z = \frac{x - \mu}{\sigma}$$

- The $z$ distribution always has mean = 0 and standard deviation = 1
- $z$, also called $z$-score, the number of standard deviations $\sigma$ it lies from the mean $\mu$.

## The Standardized Normal Probability Density Function

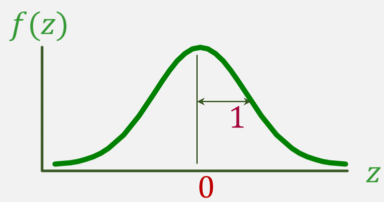
- The formula for the standardized normal probability density function is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{1}{2}\right)z^2}$$

- where
  - $e$ = the base of the natural logarithm ≈ 2.71828
  - $\pi$ = the mathematical constant ≈ 3.14159
  - $z$ = any value of the standardized normal distribution
  - Mean of $f(z)$ is $0$ and its standard deviation is $1$

# The Standardized Normal Distribution

- The standard normal distribution is also known as the "$z$" distribution
- Mean is 0
- Standard Deviation is 1




- Values above the mean have positive $z$-values.
- Values below the mean have negative $z$-values.

# Example

- If $x$, say, cost of a pair of running shoes is distributed normally with mean of \$100 and standard deviation of \$50, the $z$ value for $x = \$200$ is

$$z = \frac{x-\mu}{\sigma} = \frac{200-100}{50} = 2$$

- What does this tell us?
- That $x = \$200$ is two standard deviations (2 increments of \$50 units) above the mean of \$100.

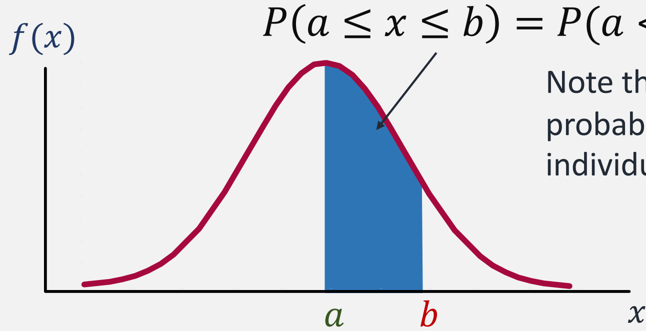
# Comparing $x$ and $z$ units


- Normalizing $x$ to get $z$ preserves the shape of the distribution but changes the scale.
- We can express the problem in the original units ($x$ in dollars) or in standardized units ($z$)

# Finding Normal Probabilities

Probability is measured by the area under the curve

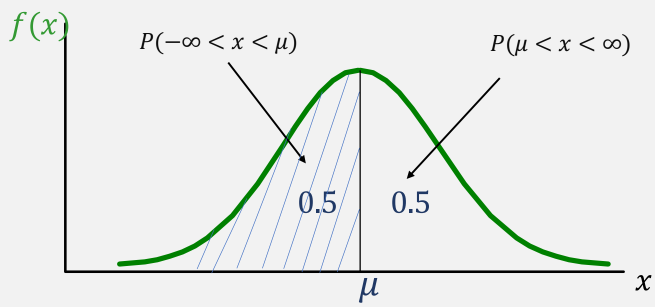
$$P(a \leq x \leq b) = P(a < x < b)$$




Note that the probability of any individual value is 0

## Probability as Area Under the Curve

- The total area under the curve is 1.0, and
- The curve is symmetric,
- So half is above the mean, half is below the mean

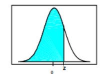



$$P(- < \infty < x < \infty) = 1$$

# Standard cumulative probability table


Cumulative probabilities for NEGATIVE z-values are shown in the following table:

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| -3.0 | .0013 | .0013 | .0013 | .0012 | .0012 | .0011 | .0011 | .0011 | .0010 | .0010 |
| -2.9 | .0019 | .0018 | .0018 | .0017 | .0016 | .0016 | .0015 | .0015 | .0014 | .0014 |
| -2.8 | .0026 | .0025 | .0024 | .0023 | .0023 | .0022 | .0021 | .0021 | .0020 | .0019 |
| -2.7 | .0035 | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |
| -2.6 | .0047 | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |
| -2.5 | .0062 | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |
| -2.4 | .0082 | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |
| -2.3 | .0107 | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |
| -2.2 | .0139 | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |
| -2.1 | .0179 | .0174 | .0170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |
| -2.0 | .0228 | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |
| -1.9 | .0287 | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |
| -1.8 | .0359 | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |
| -1.7 | .0446 | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |
| -1.6 | .0548 | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |
| -1.5 | .0668 | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |
| -1.4 | .0808 | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |
| -1.3 | .0968 | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |
| -1.2 | .1151 | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |
| -1.1 | .1357 | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |
| -1.0 | .1587 | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |
| -0.9 | .1841 | .1814 | .1788 | .1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |
| -0.8 | .2119 | .2090 | .2061 | .2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |
| -0.7 | .2420 | .2389 | .2358 | .2327 | .2296 | .2266 | .2236 | .2206 | .2177 | .2148 |
| -0.6 | .2743 | .2709 | .2676 | .2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |
| -0.5 | .3085 | .3050 | .3015 | .2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |
| -0.4 | .3446 | .3409 | .3372 | .3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |
| -0.3 | .3821 | .3783 | .3745 | .3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |
| -0.2 | .4207 | .4168 | .4129 | .4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |
| -0.1 | .4602 | .4562 | .4522 | .4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |
| 0.0 | .5000 | .4960 | .4920 | .4880 | .4840 | .4801 | .4761 | .4721 | .4681 | .4641 |




Cumulative probabilities for POSITIVE z-values are in the following table:

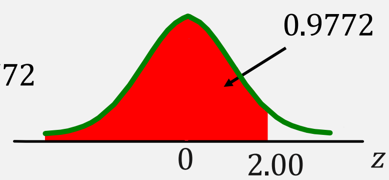
| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| 0.1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| 0.2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| 0.3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| 0.4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |
| 0.5 | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 | .7123 | .7157 | .7190 | .7224 |
| 0.6 | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| 0.7 | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| 0.8 | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| 0.9 | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| 1.0 | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| 1.1 | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| 1.2 | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | .9015 |
| 1.3 | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9279 | .9292 | .9306 | .9319 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 | .9418 | .9429 | .9441 |
| 1.6 | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 | .9525 | .9535 | .9545 |
| 1.7 | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 | .9616 | .9625 | .9633 |
| 1.8 | .9641 | .9649 | .9656 | .9664 | .9671 | .9678 | .9686 | .9693 | .9699 | .9706 |
| 1.9 | .9713 | .9719 | .9726 | .9732 | .9738 | .9744 | .9750 | .9756 | .9761 | .9767 |
| 2.0 | .9772 | .9778 | .9783 | .9788 | .9793 | .9798 | .9803 | .9808 | .9812 | .9817 |
| 2.1 | .9821 | .9826 | .9830 | .9834 | .9838 | .9842 | .9846 | .9850 | .9854 | .9857 |
| 2.2 | .9861 | .9864 | .9868 | .9871 | .9875 | .9878 | .9881 | .9884 | .9887 | .9890 |
| 2.3 | .9893 | .9896 | .9898 | .9901 | .9904 | .9906 | .9909 | .9911 | .9913 | .9916 |
| 2.4 | .9918 | .9920 | .9922 | .9925 | .9927 | .9929 | .9931 | .9932 | .9934 | .9936 |
| 2.5 | .9938 | .9940 | .9941 | .9943 | .9945 | .9946 | .9948 | .9949 | .9951 | .9952 |
| 2.6 | .9953 | .9955 | .9956 | .9957 | .9959 | .9960 | .9961 | .9962 | .9963 | .9964 |
| 2.7 | .9965 | .9966 | .9967 | .9968 | .9969 | .9970 | .9971 | .9972 | .9973 | .9974 |
| 2.8 | .9974 | .9975 | .9976 | .9977 | .9977 | .9978 | .9979 | .9979 | .9980 | .9981 |
| 2.9 | .9981 | .9982 | .9982 | .9983 | .9984 | .9984 | .9985 | .9985 | .9986 | .9986 |
| 3.0 | .9987 | .9987 | .9987 | .9988 | .9988 | .9989 | .9989 | .9989 | .9990 | .9990 |

# The Standardized Normal Table

- The Cumulative Standardized Normal table gives the probability less than a desired value of $z$ (i.e., from negative infinity to $z$)

Example:

$P(z < 2.00) = 0.9772$



# The Standardized Normal Table

The column gives the value of $z$ to the second decimal point

| $z$ | 0.00 | 0.01 | 0.02 … |
|---|---|---|---|
| 0.0 | | | |
| 0.1 | | | |
| ⋮ | | | |
| **2.0** | .9772 | | |

The row shows the value of $z$ to the first decimal point

The value within the table gives the probability from $z = -\infty$ up to the desired $z$ value

$$P(z < 2.00) = 0.9772$$

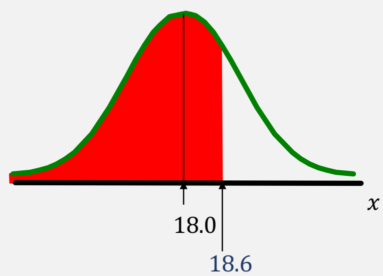
# General Procedure for Finding Normal Probabilities

To find $P(a < x < b)$ when $x$ is distributed normally:

- Draw the normal curve for the problem in terms of $x$
- Translate $x$-values to $z$-values by subtracting the mean $\mu$ and dividing by the standard deviation $\sigma$
- Use the Standardized Normal Table to read off the relevant probabilities $P(x < b)$ and $P(x < a)$
- $P(a < x < b) = P(x < b) - P(x < a)$

# Finding Normal Probabilities

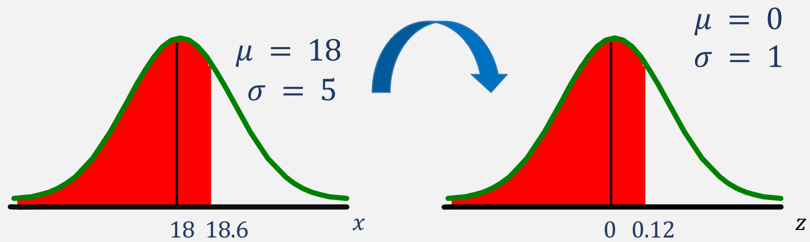
- Let $x$ represent the time it takes (in seconds) to download an a data set from the internet.
- Suppose $x$ is normal with a mean of 18.0 seconds and a standard deviation of 5.0 seconds. Find $P(X < 18.6)$



# Finding Normal Probabilities

- Let $x$ represent the time it takes, in seconds to download a data set from the internet.
- Suppose $x$ is normal with a mean of 18.0 seconds and a standard deviation of 5.0 seconds. Find $P(X < 18.6)$

$$z = \frac{x - \mu}{\sigma} = \frac{18.6 - 18.0}{5.0} = 0.12$$



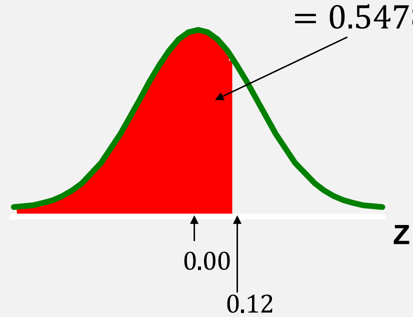

$P(x < 18.6)$   $P(z < 0.12)$

# Finding $P(z < 0.12)$

Standardized Normal Probability Table

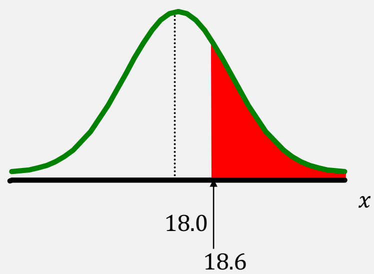
| $z$ | .00 | .01 | **.02** |
|---|---|---|---|
| 0.0 | .5000 | .5040 | .5080 |
| **0.1** | .5398 | .5438 | **.5478** |
| 0.2 | .5793 | .5832 | .5871 |
| 0.3 | .6179 | .6217 | .6255 |

$$P(x < 18.6) = P(z < 0.12) = 0.5478$$



# Finding Normal Upper Tail Probabilities

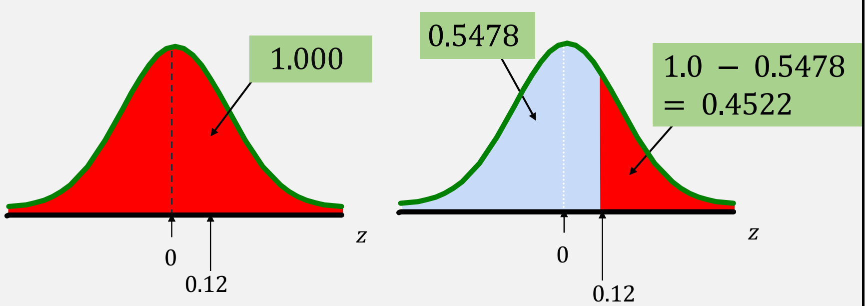
- Suppose $x$ is normal with mean 18.0 and standard deviation 5.0.
- Now find $P(X > 18.6)$



# Finding Normal Upper Tail Probabilities

- Find $P(x > 18.6)$

$$P(x > 18.6) = P(z > 0.12) = 1.0 - P(z \leq 0.12)$$
$$= 1.0 - 0.5478 = 0.4522$$



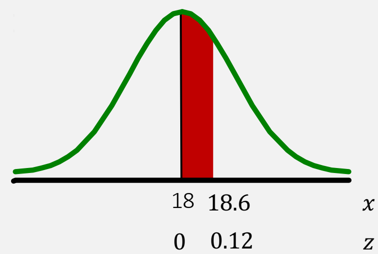
# Finding a Normal Probability Between Two Values

- Suppose $x$ is normal with mean 18.0 and standard deviation 5.0.
- Find $P(18 < x < 18.6)$

Calculate the $z$-values:

$$z_1 = \frac{x_1 - 18}{5} = \frac{18 - 18}{5} = 0$$

$$z_2 = \frac{x_2 - 18}{5} = \frac{18.6 - 18}{5} = 0.12$$



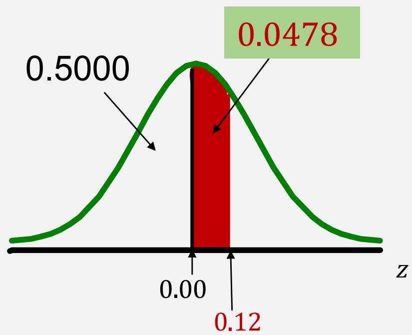

$$P(18 < x < 18.6) = P(0 < z < 0.12)$$

# Finding $P(0 < z < 0.12)$

Standardized Normal Probability Table

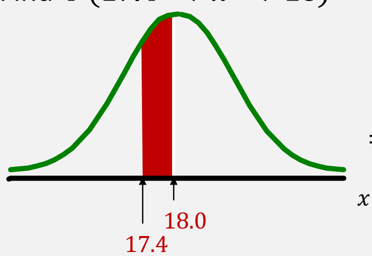
| $z$ | .00 | .01 | .02 |
|---|---|---|---|
| 0.0 | .5000 | .5040 | .5080 |
| 0.1 | .5398 | .5438 | .5478 |
| 0.2 | .5793 | .5832 | .5871 |
| 0.3 | .6179 | .6217 | .6255 |

$$P(18 < x < 18.6)$$
$$= P(0 < z < 0.12)$$
$$= P(z < 0.12) - P(z \leq 0)$$
$$= 0.5478 - 0.5 = 0.0478$$



# Probabilities in the Lower Tail

- Suppose $x$ is normal with mean 18.0 and standard deviation 5.0.
- Find $P(17.4 < x < 18)$

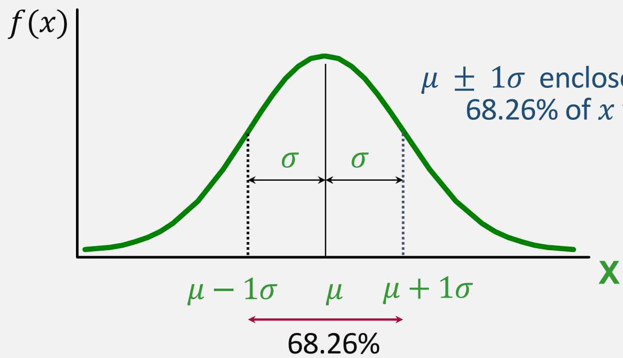



$$P(17.4 < x < 18)$$
$$= P(-0.12 < z < 0)$$
$$= P(z < 0) - P(z \leq -0.12)$$
$$= 0.5000 - 0.4522 = 0.0478$$

- Note that because of symmetry of the normal distribution, $P(-0.12 < z < 0) = P(0 < z < 0.12$

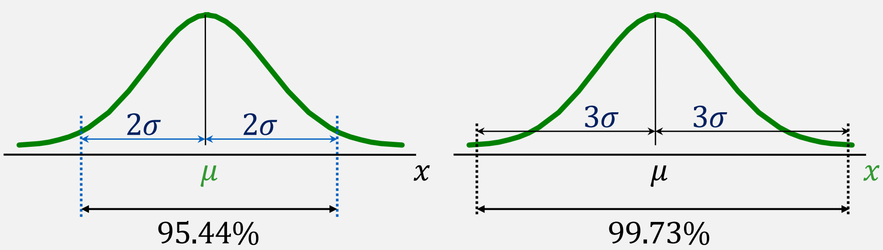
# Empirical Rule

- What can we say about the distribution of values around the mean?
- For any normal distribution:



# Empirical Rule

- $\mu \pm 2\sigma$ covers about 95.44% of $x$ values
- $\mu \pm 3\sigma$ covers about 99.73% of $x$ values



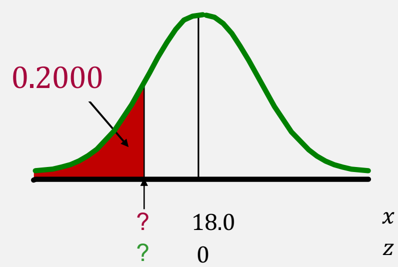
# Given a Normal Probability Find the $x$ Value

- Steps to find the $x$ value for a known probability:
- Find the $z$ value for the known probability
- Convert to $x$ units using the formula: $x = \mu + z\sigma$

# Finding the $x$ value for a Known Probability

Example:

- Let $x$ represent the time it takes (in seconds) to download a data set from the internet.
- Suppose $x$ is normal with mean 18.0 and standard deviation 5.0
- Find $x$ such that 20% of download times are less than $x$.



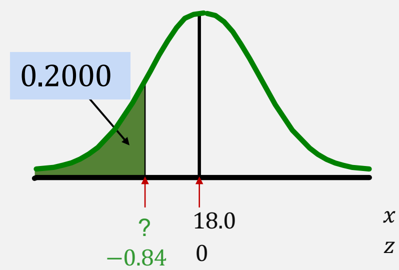
# Find the $z$ value for 20% in the Lower Tail

- Find the $z$ value for the known probability

Standardized Normal Probability Table

| $z$ | … | .03 | **.04** | .05 |
|---|---|---|---|---|
| -0.9 | … | .1762 | .1736 | .1711 |
| **-0.8** | … | .2033 | **.2005** | .1977 |
| -0.7 | … | .2327 | .2296 | .2266 |

- 20% area in the lower tail is consistent with a z value of $-0.84$



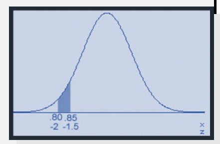
# Finding the $x$ value

Convert to $x$ units using the formula:

$$x = \mu + z\sigma = 18.0 \ + (-0.84)(5.0) = 13.8$$

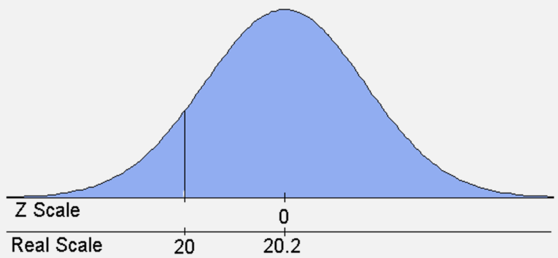
# Exercise

- The weights of packages of salad are normally distributed with mean 1 pound and standard deviation .10.
- What is the probability that a randomly selected package weighs between 0.80 and 0.85 pounds?



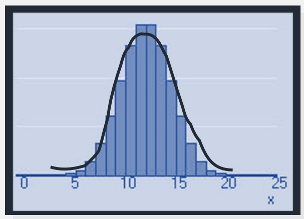
# Exercise

- A Company produces “20 ounce” jars of a picante sauce.
- The true amounts of sauce in the jars of this brand sauce follow a normal distribution.
- Suppose the companies “20 ounce” jars follow a normally distribution with a mean $\mu$=20.2 ounces with a standard deviation $\sigma$=0.125 ounces.
- What proportion of jars are under-filled?



# The Normal Approximation to the Binomial

- We can calculate binomial probabilities using
  - The binomial formula
  - The cumulative binomial tables
- When $n$ is large, and $p$ is not too close to 0 or 1, areas under the normal curve with mean $np$ and variance $npq$ can be used to approximate binomial probabilities.



# Sampling distributions

- Parameters are numerical descriptive measures for populations.
  - Two parameters for a normal distribution: mean $\mu$ and standard deviation $\sigma$.
  - One parameter for a binomial distribution: the success probability of each trial $p$.
- Often the values of parameters that specify the exact form of a distribution are unknown.
- You must rely on the sample to learn about these parameters.

# Examples of Sampling

- A pollster is sure that the responses to his “agree/disagree” question will follow a binomial distribution, but $p$, the proportion of those who “agree” in the population, is unknown.
- An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean $\mu$ and the standard deviation $\sigma$ of the yields are unknown.
- If you want the sample to provide reliable information about the population, you must select your sample such that it is representative of the population!

# Simple Random Sampling

- The sampling plan or experimental design determines
  - The amount of information you can extract, and
  - Often allows you to measure the reliability of your inference.
- Simple random sampling ensures that each possible sample of size $n$ has an equal probability of being selected.

# Sampling Distributions

- Any numerical descriptive measures calculated from the sample are called statistics.
- Statistics vary from sample to sample and hence are random variables. This variability is called sampling variability.
- The probability distributions of the statistics are called sampling distributions.
- In repeated sampling, they tell us what values of the statistics can occur and how often each value occurs.

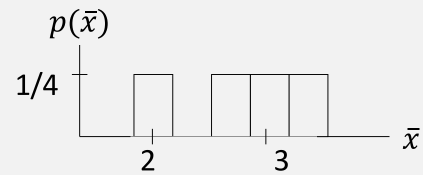
# Example

<u>Population:</u> 3, 5, 2, 1

Draw samples of size *n* = 3 without replacement

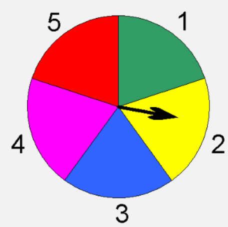
| Possible samples | $\overline{x}$ |
|---|---|
| 3, 5, 2 | $10/3 = 3.33$ |
| 3, 5, 1 | $9/3 = 3$ |
| 3, 2, 1 | $6/3 = 2$ |
| 5, 2, 1 | $8/3 = 2.67$ |

Each value of $\bar{x}$ is equally likely, with probability 1/4



# Example

- Consider a population that consists of the numbers 1, 2, 3, 4, 5 generated such that the probability of each of the values is 0.2 regardless of previous selections.
- This population could be described as the outcome associated with a roulette wheel shown with the distribution.




| x | p(x) |
|---|---|
| 1 | 0.2 |
| 2 | 0.2 |
| 3 | 0.2 |
| 4 | 0.2 |
| 5 | 0.2 |

# Example

- If the sampling distribution for the means of samples of size two is analyzed, it looks like

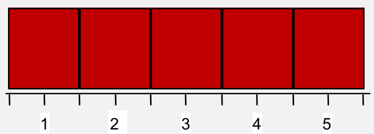
| Sample | |
|---|---|
| 1, 1 | 1 |
| 1, 2 | 1.5 |
| 1, 3 | 2 |
| 1, 4 | 2.5 |
| 1, 5 | 3 |
| 2, 1 | 1.5 |
| 2, 2 | 2 |
| 2, 3 | 2.5 |
| 2, 4 | 3 |
| 2, 5 | 3.5 |
| 3, 1 | 2 |
| 3, 2 | 2.5 |
| 3, 3 | 3 |

| Sample | |
|---|---|
| 3, 4 | 3.5 |
| 3, 5 | 4 |
| 4, 1 | 2.5 |
| 4, 2 | 3 |
| 4, 3 | 3.5 |
| 4, 4 | 4 |
| 4, 5 | 4.5 |
| 5, 1 | 3 |
| 5, 2 | 3.5 |
| 5, 3 | 4 |
| 5, 4 | 4.5 |
| 5, 5 | 5 |

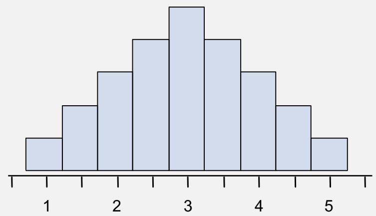
| | frequency | p($\bar{x}$) |
|---|---|---|
| 1 | 1 | 0.04 |
| 1.5 | 2 | 0.08 |
| 2 | 3 | 0.12 |
| 2.5 | 4 | 0.16 |
| 3 | 5 | 0.20 |
| 3.5 | 4 | 0.16 |
| 4 | 3 | 0.12 |
| 4.5 | 2 | 0.08 |
| 5 | 1 | 0.04 |
| | 25 | |

# Example

The original distribution and the sampling distribution of means of samples with $n = 2$ are given below.



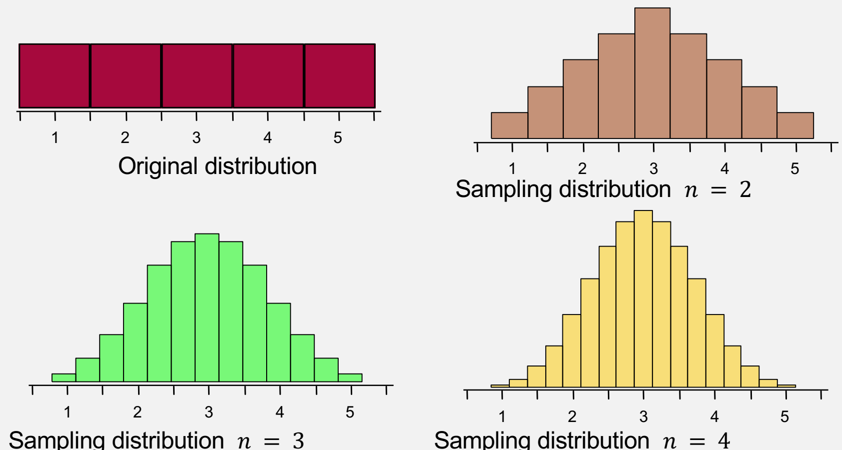

Original distribution




Sampling distribution for $n = 2$

# Example

- Sampling distributions for n=3 and n=4 are shown below.
- What do you notice as n gets larger?
- The sampling distribution approaches normal distribution



## Sampling Distribution of $\overline{x}$

If a random sample of *n* measurements is selected from a population with mean $\mu$ and standard deviation $\sigma$, the sampling distribution of the sample mean $\overline{x}$ will have a mean

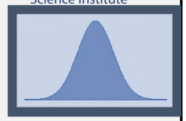
$$\mu_{\overline{x}} = \mu$$

and a standard deviation

$$\sigma_{\overline{x}} = \sigma / \sqrt{n}$$

Central Limit Theorem: If random samples of *n* observations are drawn from a nonnormal population with finite $\mu$ and standard deviation $\sigma$ , then, when *n* is large, the sampling distribution of the sample mean $\overline{x}$ is approximately normally distributed, with mean $\mu$ and standard deviation $\sigma / \sqrt{n}$ .
The approximation becomes more accurate as *n* becomes large.

## Why is this Important?

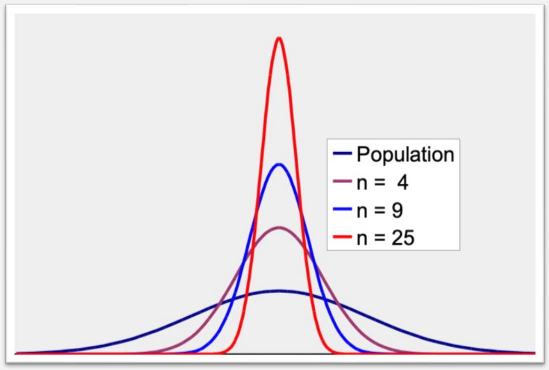


- The Central Limit Theorem also implies that the sum of $n$ measurements is approximately normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$ .
- Many statistics that are used for statistical inference are sums or averages of sample measurements.
- When $n$ is large, these statistics will have approximately normal distributions.
- This will allow us to describe their behavior and evaluate the reliability of our inferences.

## How Large is Large?

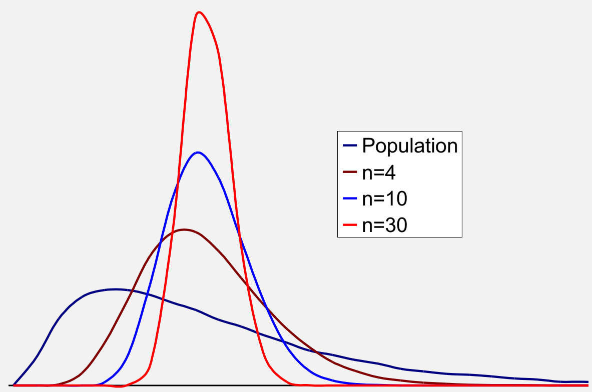
- If the sample is normal, then the sampling distribution of $\bar{x}$ will also be normal, no matter what the sample size.
- When the sample population is approximately symmetric, the distribution becomes approximately normal for relatively small values of $n$.
- When the sample population is skewed, the sample size must be at least 30 before the sampling distribution of $\bar{x}$ becomes approximately normal.

# Sampling Distributions Illustrated




Symmetric distributions that resemble the normal distribution

# Sampling distributions illustrated




Skewed population

# Finding Probabilities for the Sample Mean

- If the sampling distribution of $\bar{x}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- Find the appropriate area using the z distribution.

Example: A random sample of size *n* = 16 from a normal distribution with $\mu = 10$ and $\sigma = 8$.

$$P(\bar{x} > 12) = P(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < \frac{12 - 10}{8 / \sqrt{16}})$$

$$= P(z > 1) = .5 - .3413 = .1587$$

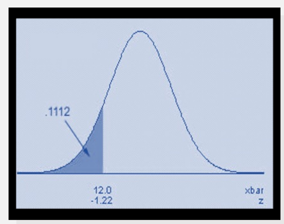
## Example

- A soda filling machine is supposed to fill cans of soda with 12 fluid ounces.
- Suppose that the fills are actually normally distributed with a mean of 12.1 oz and a standard deviation of .2 oz.
- The probability of one can less than 12 is

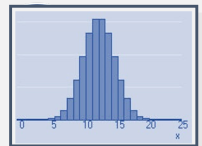
$$P(x<12) = P(\frac{x-\mu}{\sigma} < \frac{12-12.1}{.2}) = P(z<-.5) = .5-.1915 = .3085$$

What is the probability that the average fill for a 6-pack is less than 12 oz?

$$P(\bar{x}<12) =$$
$$P(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < \frac{12-12.1}{.2/\sqrt{6}}) =$$
$$P(z<-1.22) = .1112$$

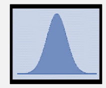


# The Sampling Distribution of the Sample Proportion




- The Central Limit Theorem can be used to conclude that the binomial random variable *x* is approximately normal when *n* is large, with mean *np* and variance $npq$.
- The sample proportion, $\hat{p} = \frac{x}{n}$ is simply a *rescaling* of the binomial random variable *x*, dividing it by $n$.
- From the Central Limit Theorem, the sampling distribution of $\hat{p}$ will also be approximately normal, with a *rescaled* mean and standard deviation.

## The Sampling Distribution of the Sample Proportion



✓A random sample of size *n* is selected from a binomial population with parameter *p*.

✓The sampling distribution of the sample proportion, $\hat{p} = \frac{x}{n}$

will have mean *p* and standard deviation $\sqrt{\frac{pq}{n}}$

✓If *n* is large, and *p* is not too close to zero or one, the sampling distribution of $\hat{p}$ will be approximately normal.

- The standard deviation of *p*-hat is sometimes called the STANDARD ERROR (SE) of *p*-hat.

## Finding Probabilities for the Sample Proportion

✓If the sampling distribution of $\hat{p}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

✓Find the appropriate area using the normal table.

## Example

- The soda bottler in the previous example claims that only 5% of the soda cans are underfilled.
- A quality control technician randomly samples 200 cans of soda.
- What is the probability that more than 10% of the cans are underfilled?

$n$ = 200

S: underfilled can

$p$ = P(S) = .05

$q$ = .95

$np$ = 10  $nq$ = 190

OK to use the normal approximation

$$P(\hat{p} > .10)$$

$$= P\left(z > \frac{.10 - .05}{\sqrt{\frac{.05(.95)}{200}}}\right) = P(z > 3.24)$$

$$< .5 - .4990 = .001$$

This would be very unusual, if indeed $p$ = .05!

## Example

- Suppose 3% of the people contacted by phone are receptive to a certain sales pitch and buy your product. If your sales staff contacts 2000 people, what is the probability that more than 100 of the people contacted will purchase your product?

$$n = 2000, \qquad p = 0.03, np = 60, nq = 1940,$$

OK to use the normal approximation

$$P(\hat{p} > 100/2000) = P(z > \frac{.05 - .03}{\sqrt{\frac{.03(.97)}{2000}}}) = P(z > 5.24) \approx 0$$

# Sampling - Summary

- Simplest sampling technique – random sampling
  - Each possible sample is equally likely
- Sampling distributions describe the possible values of a statistic and how often they occur in repeated sampling.
- If the underlying data are normally distributed, sampling distribution is a normal distribution
- The **Central Limit Theorem** states that sums and averages of measurements from an arbitrarily distributed population with finite mean $\mu$ and standard deviation $\sigma$ have approximately normal distributions for large samples of size $n$.

# Sampling - Sampling Distribution of the Sample Mean

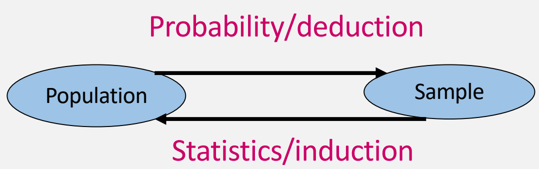
- When samples of size $n$ are drawn from a normal population with mean $\mu$ and variance $\sigma^2$, the sample mean $\bar{x}$ has a normal distribution with mean $\mu$ and variance $\sigma^2/n$.
- When samples of size $n$ are drawn from a nonnormal population with mean $\mu$ and variance $\sigma^2$, the Central Limit Theorem ensures that the sample mean $\bar{x}$ will have an approximately normal distribution with mean $\mu$ and variance $\sigma^2/n$ when $n$ is large ($n$ 3 30).
- Probabilities involving the sample mean $\mu$ can be calculated by standardizing the value of $\bar{x}$ using $z = \frac{x-\mu}{\sigma/\sqrt{n}}$

## Summary – Sampling Distribution of the Sample Proportion

- When samples of size $n$ are drawn from a binomial population with parameter $p$, the sample proportion $\hat{p}$ will have an approximately normal distribution with mean *p* and variance $pq\ /n$ as long as $np > 5$ and $nq > 5$.
- Probabilities involving the sample proportion $\hat{p}$ can be calculated by standardizing the value $\hat{p}$ using $z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$

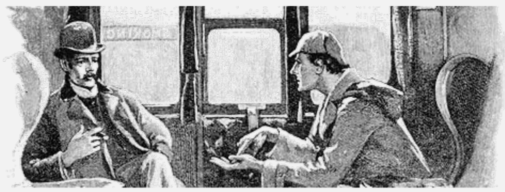
# Probabilistic  vs Statistical Reasoning

- In the last lecture, we looked at how the properties of a population govern what we see in sample(s) of the population
- Now we turn to going in the other direction: Given a sample, we try to understand the data generating process that could have generated the observed data
- This shift our mode of thinking from deductive reasoning to induction



# Probabilistic vs Statistical Reasoning

- In many ways, science, or scholarly inquiry, is like detective work.
- We begin with a set of observations, we ask what can be said about the data generating process



- "Data! Data! Data!.. I can't make bricks without clay"

— Sherlock Holmes, 1892

— " The Adventure of the Copper Beeches"

# Shadows: Shadow Puppetry :: Data : Data Generating Process

Image source: Annie Katsura Rollins, Ballard Institute and Museum of Puppetry, photo by Kenneth Best

# Parameters

- Populations are described by their probability distributions
  - If we assume a parametric form for the distribution, e.g., Normal, binomial, etc., then populations are described by the parameters of the respective distributions
    - Binomial populations are determined by a single parameter, $p$.
    - Normal distributions are described by the mean $\mu$ and the standard deviation $\sigma$.
- If the values of parameters are unknown, we have to make inferences about them using information provided by a sample from the underlying distribution
- Sample or data : distribution :: shadows : shadow puppetry
- The puppeteer whose machinations generate the shadows you see is hidden from you. Your goal is to learn his or her modus operandi.

# Two types of statistical inference

- Estimation
  - Estimating or inferring the value of the parameter(s)
    - Maximum likelihood: What is the mean height of individuals of Asian descent given the sample of individuals of Asian descent you have observed?
    - Bayesian: What is the likely height of the next person of Asian descent you may encounter, given your prior belief about the heights of individuals of Asian descent, the heights of individuals of Asian descent that you have observed?
- Hypothesis testing
  - Deciding if the data support a preconceived idea or theory one has about a population
    - “Did the sample of individuals you have come from a population with mean height of 5.6” ?
    - “Was the newly discovered manuscript of unknown authorship written by Shakespeare?

## Specify the type of statistical inference

- A consumer wants to estimate the average price of similar homes in her city before putting her home on the market.
- Estimation: Estimate the average price of similar homes in the city

- A manufacturer wants to know if a new type of steel is more resistant to high temperatures than an old type was.
- Hypothesis testing: Is the average efficacy of the new Covid vaccine $\mu_{New}$ greater than that of the old Covid vaccine $\mu_{Old}$?

# Methods of Statistical Inference

- Whether you are estimating parameters or testing hypotheses, statistical methods
  - Offer a sound basis for inference
  - A measure of the goodness or reliability of the inference

# What is an estimator?

An estimator is a formula, that tells you how to calculate the estimate of a parameter of interest from the given sample.

- Point estimation yields a single value for the parameter
  - Example: The estimated probability of a coin coming up heads is 0.4
  - Underlying assumption:
    - The coin has a fixed parameter $p$
    - Our job is to estimate it.
    - How realistic is this assumption?
- Confidence interval is an interval such that for a chosen degree of confidence, expressed as a probability, the true value of the parameter is likely to fall inside the interval.
  - Example: 95% confidence interval for $p$ is [0.3, 0.5]

# Point Estimator of Population Mean

- Given a sample $S = \{x_1 \cdots x_n\}$, the point estimate of the population mean, $\mu$, is the sample mean

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  - Example: Suppose $S = \{4,5,6,3,4,6,3,5,8,1\}$ were the ratings given by viewers of the movie "Back to the future".
  - Suppose you believe that the viewers are a random sample of the viewers of "Back to the future".
  - What is the sample estimate of the mean rating of "Back to the future?
  - 4.5
- But why?

# Point Estimation of Population Proportion

- A point estimate of $p$, population success rate of a binary experiment (e.g., coin tosses with outcomes $H$ and $T$) is sample proportion of successes observed in the sample:

$$\hat{p} = \frac{n_H}{n_H + n_T}$$

- Example: Out of 100 people tested for Covid, 10 were positive.
  - What is the point estimate of $p$, the Covid positive rate in the population?

$$\hat{p} = \frac{10}{100} = 0.1$$

- But why?

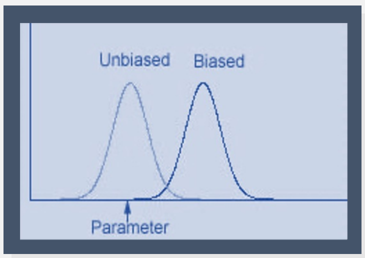
# Properties of Point Estimators

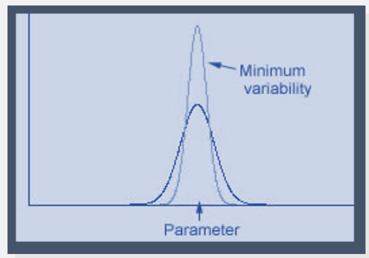
- Since an estimator is calculated from sample values, it varies from sample to sample according to its sampling distribution.
- An estimator is unbiased if
  - The mean of its sampling distribution equals the parameter of interest.
  - It does not systematically overestimate or underestimate the target parameter.
- Both sample mean and sample proportion are unbiased estimators of population mean and proportion.
- Given $n$ samples, the following sample variance is an unbiased estimator of population variance $\sigma^2$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n-1}$$

# Properties of Point Estimators

- Of all the unbiased estimators, we prefer the estimator whose sampling distribution has the smallest spread or variability.



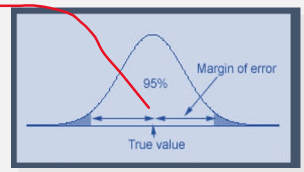



# Confidence Intervals

- Confidence intervals depend on sampling distributions
- The shape of sampling distributions depend on sample sizes
- For large sample sizes, central limit theorem applies which allow us to use normal distributions
- For small sample sizes, we need to choose the right sampling distribution

# Quantifying the error of Point Estimates

- Assumption: The sample sizes are large
- From the Central Limit Theorem, the sampling distributions of $\hat{\mu}$ and $\hat{p}$ will be approximately normal
- For unbiased estimators with normal sampling distributions, 95% of all point estimates will lie within 1.96 standard deviations of the parameter of interest.
- Margin of error: an upper bound on the difference between a particular estimate and the parameter that it estimates.
- Margin of error = 1.96 $\times$ standard deviation of the estimate



# Estimating Means and Proportions

Point estimator of population mean $\mu : \bar{x}$

Margin of error $(n \geq 30) : \pm 1.96 \frac{s}{\sqrt{n}}$

For a binomial population,

Point estimator of population proportion $\mathrm{p} : \hat{\mathrm{p}} = \mathrm{x/n}$

Margin of error $: \pm 1.96\sqrt{\frac{\hat{\mathrm{p}}\hat{\mathrm{q}}}{\mathrm{n}}}$

*Assumption* : $\mathrm{np} > 5$ and $\mathrm{nq} > 5$; or $0 < \mathrm{p} \pm 2\sqrt{\frac{\mathrm{pq}}{\mathrm{n}}} < 1$

# Example

- A homeowner randomly samples 64 homes similar to her own and finds that the average selling price is $250,000 with a standard deviation of $15,000.
- Estimate the average selling price for all similar homes in the city.
- What is the margin of error?

$$\text{Point estimator of } \mu : \bar{x} = 250{,}000$$

$$\text{Margin of error} : \pm 1.96 \frac{\hat{\sigma}\ s}{\sqrt{n}} = \pm 1.96 \frac{15{,}000}{\sqrt{64}} = \pm 3675$$

## Example

- A quality control technician wants to estimate the proportion of soda cans that are underfilled.
- He randomly samples 200 cans of soda and finds 10 underfilled cans.

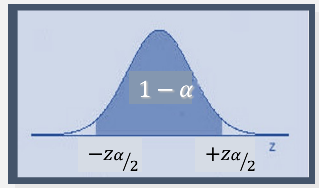
$n = 200 \qquad p = \text{proportion of underfilled cans}$

$\text{Point estimator of } p : \hat{p} = x/n = 10/200 = .05$

$$\text{Margin of error} : \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = \pm 1.96\sqrt{\frac{(.05)(.95)}{200}} = \pm .03$$

# Confidence Interval

- Create an interval so that you are fairly sure that the parameter lies between these two values.
- "Fairly sure" means "with high probability", measured using the confidence coefficient, $1-\alpha$.
- Usually, $1-\alpha = 0.9, 0.95, 0.99$ ...



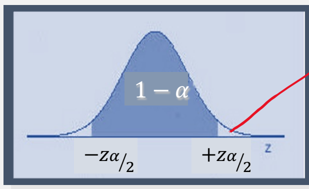

- For large-Sample size,

$100(1-\alpha)\%$ confidence Interval:

Point Estimate $\pm z_{\alpha/2}$

# Confidence Level

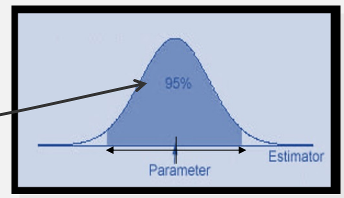
- To change to a general confidence level, $1-\alpha$, pick a value of $z$ that puts area $1-\alpha$ in the center of the $z$ distribution.




| Tail area $\alpha/2$ | $z_{\alpha/2}$ |
|---|---|
| .05 | 1.645 |
| .025 | 1.96 |
| .005 | 2.58 |

- Suppose 1-$\alpha$ = .95

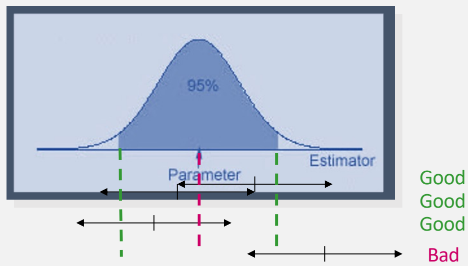
There is 95% probability that the interval constructed in this manner will contain the population mean



# Confidence Interval

- Since we don't know the value of the parameter, consider which has a variable center.

Point Estimator $\pm$ 1.96 std error




- Only if the estimator falls in the tail areas will the interval fail to enclose the parameter. This happens only 5% of the time.

## Interpretation of a Confidence Interval

- A confidence interval is calculated from one given sample.
- The interval either covers or misses the true parameter.
- Since the true parameter is unknown, you'll never know with certainty
- If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter.
- The confidence level associated with a confidence interval is the success rate of the confidence interval.

# Confidence Intervals for Means and Proportions

Confidence interval for a population mean $\mu$ :

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

For a binomial population:

Confidence interval for a population proportion $p$ :

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$


## Example

A random sample of $n = 50$ males showed a mean average daily intake of dairy products equal to 756 grams with a standard deviation of 35 grams. Find a 95% confidence interval for the population average $\mu$.

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} \Rightarrow 756 \pm 1.96 \frac{35}{\sqrt{50}} \Rightarrow 756 \pm 9.70$$

$$\text{or } 746.30 < \mu < 765.70 \text{ grams.}$$

## Example

Find a 99% confidence interval for $\mu$, the population average daily intake of dairy products for men.

$$\bar{x} \pm 2.58\frac{s}{\sqrt{n}} \Rightarrow 756 \pm 2.58\frac{35}{\sqrt{50}} \Rightarrow 756 \pm 12.77$$

$$\text{or } 743.23 < \mu < 768.77 \text{ grams.}$$

## Example

- Of a random sample of $n$ = 150 college students, 104 of the students said that they had played on a soccer team during their K-12 years.
- Estimate the proportion of college students who played soccer in their youth with a 90% confidence interval.

$$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}\hat{q}}{n}} \Rightarrow \frac{104}{150} \pm 1.645\sqrt{\frac{.69(.31)}{150}}$$

$$\Rightarrow \quad .69 \pm .06 \quad \text{or} \;\; .63 < p < .75.$$

## Estimating the Difference between Two Means

- Sometimes we are interested in comparing the means of two populations.
  - The average growth of plants fed using two different nutrients.
  - The average scores for students taught with two different teaching methods.
- To make this comparison

A random sample of size $n_1$ drawn from population 1 with mean $\mu_1$ and variance $\sigma_1^2$.

A random sample of size $n_2$ drawn from population 2 with mean $\mu_2$ and variance $\sigma_2^2$.

# Comparing Two Means

| | Mean | Variance | Standard Deviation |
|---|---|---|---|
| Population 1 | $\mu_1$ | $\sigma_1^2$ | $\sigma_1$ |
| Population 2 | $\mu_2$ | $\sigma_2^2$ | $\sigma_2$ |

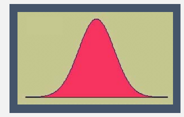
| | Sample size | Mean | Variance | Standard Deviation |
|---|---|---|---|---|
| Sample from Population 1 | $n_1$ | $\bar{x}_1$ | $s_1^2$ | $s_1$ |
| Sample from Population 2 | $n_2$ | $\bar{x}_2$ | $s_2^2$ | $s_2$ |

# Estimating the Difference between Two Means

- We compare the two averages by making inferences about $\mu_1$-$\mu_2$, the difference in the two population averages.
  - If the two population averages are the same, then $\mu_1$-$\mu_2$ = 0.
  - The best estimate of $\mu_1$-$\mu_2$ is the difference in the two sample means

$$\bar{x}_1 - \bar{x}_2$$

# The Sampling Distribution of $\bar{x}_1 - \bar{x}_2$



1. The mean of $\bar{x}_1 - \bar{x}_2$ is $\mu_1 - \mu_2$, the difference in the population means.

2. The standard deviation of $\bar{x}_1 - \bar{x}_2$ is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

3. If the sample sizes (both $n_1$ and $n_2$) are large, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal, and standard deviation can be estimated as $\text{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

# Estimating $\mu_1$-$\mu_2$

- For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal ($z$) distribution.

$$\text{Point estimate for } \mu_1 - \mu_2 : \bar{x}_1 - \bar{x}_2$$

$$\text{Margin of Error} : \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

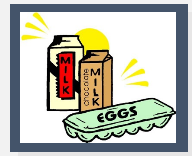
$$\text{Assumption}:$$
$$\text{Both } n_1 \geq 30 \text{ and } n_2 \geq 30$$

$$\text{Confidence interval for } \mu_1 - \mu_2 :$$

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

# Example

| Avg Daily Intakes | Men | Women |
|---|---|---|
| Sample size | 50 | 50 |
| Sample mean | 756 | 762 |
| Sample Std Dev | 35 | 30 |




- Compare the average daily intake of dairy products of men and women using a 95% confidence interval.

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\Rightarrow (756 - 762) \pm 1.96\sqrt{\frac{35}{50} + \frac{30}{50}} \quad \Rightarrow \quad -6 \pm 12.78$$

or $-18.78 < \mu_1 - \mu_2 < 6.78.$

# Example, continued


$$-18.78 < \mu_1 - \mu_2 < 6.78$$

- Could you conclude, based on this confidence interval, that there is a difference in the average daily intake of dairy products for men and women?
- The confidence interval contains the value $\mu_1 - \mu_2 = 0$.
- Therefore, it is possible that $\mu_1 = \mu_{2.}$
- You would not want to conclude that there is a difference in average daily intake of dairy products for men and women.

# Estimating the Difference between Two Proportions

- Sometimes we are interested in comparing the proportion of “successes” in two binomial populations.
  - The germination rates of untreated seeds and seeds treated with a fungicide.
  - The proportion of male and female voters who favor a particular candidate for governor.
- To make this comparison

A random sample of size $n_1$ drawn from binomial population 1 with parameter $p_1$.

A random sample of size $n_2$ drawn from binomial population 2 with parameter $p_2$.

# Comparing Two Proportions

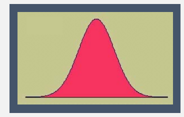
| | Sample size | Sample Proportion | Sample Variance | Standard Deviation |
|---|---|---|---|---|
| Sample from Population 1 | $n_1$ | $\hat{p}_1 = \frac{x_1}{n_1}$ | $\frac{\hat{p}_1\hat{q}_1}{n}$ | $\sqrt{\frac{\hat{p}_1\hat{q}_1}{n}}$ |
| Sample from Population 2 | $n_2$ | $\hat{p}_2 = \frac{x_2}{n_2}$ | $\frac{\hat{p}_2\hat{q}_2}{n}$ | $\sqrt{\frac{\hat{p}_2\hat{q}_2}{n}}$ |

# Estimating the Difference between Two Means

- We compare the two proportions by making inferences about $p_1$-$p_2$, the difference in the two population proportions.
  - If the two population proportions are the same, then $p_1$-$p_2$ = 0.
  - The best estimate of $p_1$-$p_2$ is the difference in the two sample proportions,

$$\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$$

## The Sampling Distribution of $\hat{p}_1 - \hat{p}_2$



1. The mean of $\hat{p}_1 - \hat{p}_2$ is $p_1 - p_2$, the difference in the population proportions.

2. The standard deviation of $\hat{p}_1 - \hat{p}_2$ is $\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}$.

3. If the sample sizes (both $n_1$ and $n_2$) are large, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal, and stanard deviation can be estimated as

$$SE = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}.$$

# Estimating $p_1$-$p_2$

For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal ($z$) distribution.

$$\text{Point estimate for } p_1\text{-}p_2 : \hat{p}_1 - \hat{p}_2$$

$$\text{Margin of Error} : \pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\text{Confidence interval for } p_1 - p_2 :$$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Assumption : both $n_1$ and $n_2$ are sufficiently large so that $-1 \le \hat{p}_1 - \hat{p}_2 \pm 2SE \le 1$

# Example

| Youth Soccer | Male | Female |
|---|---|---|
| Sample size | 80 | 70 |
| Played soccer | 65 | 39 |

- Compare the proportion of male and female college students who said that they had played on a soccer team during their K-12 years using a 99% confidence interval.

$$(\hat{p}_1 - \hat{p}_2) \pm 2.58\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\Rightarrow (\frac{65}{80} - \frac{39}{70}) \pm 2.58\sqrt{\frac{.81(.19)}{80} + \frac{.56(.44)}{70}} \quad \Rightarrow \quad .25 \pm .19$$

$$\text{or } .06 < p_1 - p_2 < .44.$$

## Example, continued

$$.06 < p_1 - p_2 < .44$$

- Could you conclude, based on this confidence interval, that there is a difference in the proportion of male and female college students who said that they had played on a soccer team during their K-12 years?
- The confidence interval does not contains the value $p_1$-$p_2$ = 0. Therefore, it is not likely that $p_1$= $p_2$. You would conclude that there is a difference in the proportions for males and females.

A higher proportion of males than females played soccer in their youth.

# Summary – Large Sample Point Estimators

To estimate one of four population parameters when the sample sizes are large, use the following point estimators with the appropriate margins of error.

| Parameter | Point Estimator | Margin of Error |
|---|---|---|
| $\mu$ | $\bar{x}$ | $\pm 1.96\left(\frac{s}{\sqrt{n}}\right)$ |
| $p$ | $\hat{p} = \frac{x}{n}$ | $\pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$ |
| $\mu_1 - \mu_2$ | $\bar{x}_1 - \bar{x}_2$ | $\pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ |
| $p_1 - p_2$ | $(\hat{p}_1 - \hat{p}_2) = \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)$ | $\pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$ |

# Summary – Large Sample Confidence Intervals

To estimate one of four population parameters when the sample sizes are large, use the following interval estimators.

| Parameter | $(1-\alpha)100\%$ Confidence Interval |
|---|---|
| $\mu$ | $\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$ |
| $p$ | $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ |
| $\mu_1 - \mu_2$ | $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ |
| $p_1 - p_2$ | $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$ |

## Summary: Large Sample Confidence intervals

- All values in the interval are possible values for the unknown population parameter.
- Any values outside the interval are unlikely to be the value of the unknown parameter.
  - To compare two population means or proportions, look for the value 0 in the confidence interval.
  - If 0 is in the interval, it is possible that the two population means or proportions are equal, and you should not declare a difference.
  - If 0 is not in the interval, it is unlikely that the two means or proportions are equal, and you can confidently declare a difference.

# Estimation – A little deeper dive

- Recall that the point estimate assumes that the parameter to be estimated is constant. Who can say that it truly is?
- Estimating or inferring the value of the parameter(s)
  - Maximum likelihood: What is the mean height of individuals of Asian descent given the sample of individuals of Asian descent you have observed?
  - Maximum a posteriori: What is the mean height of individuals of Asian descent given your prior belief about the heights of individuals of Asian descent and the heights of individuals of Asian descent that you have encountered?
  - Bayesian: What is the likely height of the next person of Asian descent you may encounter, given your prior belief about the heights of individuals of Asian descent, the heights of individuals of Asian descent that you have observed?

# Estimating probabilities from data (discrete case)

- Maximum likelihood estimation
- Bayesian estimation
- Maximum a posteriori estimation

## Example: Binomial Experiment

Head

Tail

- When tossed, the thumbtack can land in one of two positions: ***Head*** or ***Tail***
- We denote by θ the (unknown) probability $P(H)$.
- Estimation task
  - Given a sequence of toss samples $x_1 \cdots x_N$, we want to estimate the probabilities $P(H) = \theta$ and $P(T) = 1 - \theta$

# Population Parameter Estimation from Data

Consider samples $x_1 \cdots x_N$ such that

- The values that *the random variable* $x$ can take is known
- Each is sampled from the same distribution
- Each is sampled independently of the rest

$\Bigg\}$ i.i.d. samples

The task is to find a parameter $\theta$ so that our belief about the data can be summarized by a probability $P(x|\theta)$.

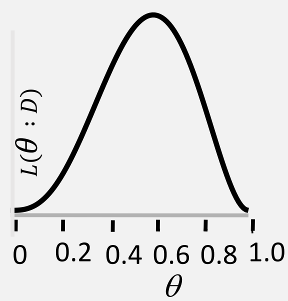
- The parameters depend on the given family of probability distributions: multinomial, Gaussian, Poisson, etc.
- We will focus first on binomial and then on multinomial distributions
- The main ideas generalize to other distribution families

## The Likelihood Function

- How good is a particular estimate of $\theta$?
- It depends on how likely it is to generate the observed data as specified by the likelihood function

$$L(\theta : D) = P(D|\theta) = \prod_{i=1}^{N} P(x_i|\theta$$

The likelihood for the sequence $H, T, T, H, H$ is




$$L(\theta : D) \propto \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta$$

# Likelihood function

- The likelihood function $L(\square : D)$ provides a measure of relative preferences for various values of the parameter $\mu$ given a collection of observations $D$ drawn from a distribution that is parameterized by fixed but unknown $\theta$.
- $L(\theta: D)$ is proportional to the probability of the observed data D viewed as a function of θ.
- Suppose data $D$ is 5 heads out of 8 tosses.
- What is the likelihood function assuming that the observations were generated by a binomial distribution with an unknown but fixed parameter $\theta$?

$$\binom{8}{5}\theta^5(1-\theta)^3$$

# Sufficient Statistics

- To compute the likelihood in the thumbtack example we only require $N_H$ and $N_T$ (the number of heads and the number of tails)

$$L(\theta : D) \propto \theta^{N_H} \cdot (1-\theta)^{N_T}$$

- $N_H$ and $N_T$ are sufficient statistics for the parameter $\theta$ that specifies the binomial distribution
- A statistic is simply a function of the data
- A sufficient statistic $s$ for a parameter $\theta$ is a function that summarizes from the data $D$, the relevant information $s(D)$ needed to compute the likelihood $L(\theta : D)$.
- If $s$ is a sufficient statistic for $\theta$, and $s(D) = s(D')$, then $L(\theta : D) = L(\theta : D')$

# Maximum Likelihood Estimation

- Main Idea: Estimate from the given data, parameters that maximize the likelihood function
- Maximum likelihood estimator is
  - Intuitively appealing
  - One of the most commonly used estimators in statistics
  - Assumes that the parameters to be estimated are fixed, but unknown

## Example: MLE for Binomial Data

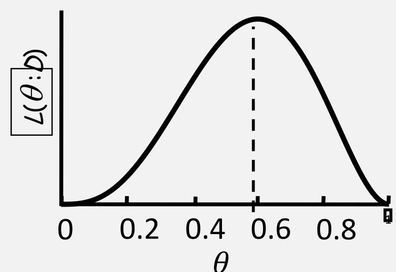
- Applying the MLE principle we get
- (Why?)

$$\hat{\theta} = \frac{N_H}{N_H + N_T}$$

Example:

$(NH, NT) = (3,2)$

ML estimate is $\theta$



# MLE for Binomial data

$$L(\theta : D) = \binom{N}{N_H} \theta^{N_H} \cdot (1-\theta)^{N_T}$$

$$\log L(\theta : D) = N_H \log \theta + N_T \log(1-\theta)$$

The likelihood is positive for all legitimate values of $\theta$

So maximizing the likelihood is equivalent to maximizing its logarithm i.e. log likelihood

$$\frac{\partial}{\partial\theta} \log L(\theta{:}D) = 0 \text{ at extrema of } L(\theta{:}D)$$

$$\frac{\partial}{\partial\theta} \log L(\theta{:}D) = \frac{N_H}{\theta} + \frac{N_T(-1)}{(1-\theta)} = 0$$

$$(N_H + N_T)\theta = N_H$$

$$\theta_{ML} = \frac{N_H}{(N_H + N_T)}$$

Note that the likelihood is indeed maximized at $\theta = \theta_{ML}$ because in the neighborhood of $\theta_{ML}$, the value of the likelihood is smaller than it is at $\theta = \theta_{ML}$

## Behavior of the likelihood around the maximum

- At the maximum, the derivative of the log likelihood is zero
- At the maximum, the second derivative is negative
- The curvature of the log likelihood is defined as

$$I(\theta) = -\frac{\partial}{\partial \theta^2} \log L(\theta : D)$$

- Large observed curvature $I\ (\theta_{ML})$ at $\theta = \theta_{ML}$ is associated with a sharp peak, intuitively indicating less uncertainty about the maximum likelihood estimate
- $I\ (\theta_{ML})$ is called the Fisher information

# Maximum Likelihood Estimate

ML estimate can be shown to be

- Asymptotically unbiased

$$\lim_{N\to\infty} E(\theta_{ML}) = \theta_{True}$$

- Asymptotically consistent - converges to the true value as the number of examples approaches infinity

$$\lim_{N\to\infty} \Pr\{\|\theta_{ML} - \theta_{True}\| \le \varepsilon\} = 1$$

$$\lim_{N\to\infty} E\left(\|\theta_{ML} - \theta_{True}\|^2\right) = 0$$

- Asymptotically efficient – achieves the lowest variance that any estimate can achieve for a training set of a certain size (satisfies the Cramer-Rao bound)


# Maximum Likelihood Estimate

- ML estimate can be shown to be representationally invariant – If $\theta_{ML}$ is an ML estimate of $\theta$, and $g(\theta)$ is a function of $\theta$, then $g(\theta_{ML})$ is an ML estimate of $g(\theta)$
- When the number of samples is large, the probability distribution of $\theta_{ML}$ has Normal distribution with mean $\theta_{True}$ (the actual value of the parameter) – a consequence of the central limit theorem
  - A random variable which is a sum of a large number of random variables has Normal distribution – ML estimate is related to the sum of random variables
- We can use the likelihood ratio to reject the null hypothesis corresponding to $\theta = \theta_0$ as unsupported by data if the ratio of the likelihoods evaluated at $\theta_0$ and at $\theta_{ML}$ is small.

## From Binomial to Multinomial

- Suppose a random variable $x$ can take the values $1,2, \ldots, K$
- We want to learn the parameters $\theta_1, \theta_2. \ldots, \theta_K$
- Sufficient statistics: $N_1, N_2, \ldots, N_K$ - the number of times each outcome is observed
- Likelihood function

$$L(\theta : D) \propto \prod_{k=1}^{K} \theta_k^{N_k}$$

- ML estimate

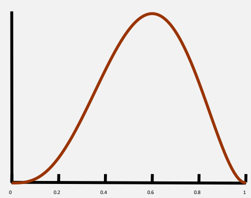
$$\hat{\theta}_k = \frac{N_k}{\sum_{\ell} N_\ell}$$

# Summary of Maximum Likelihood Estimation

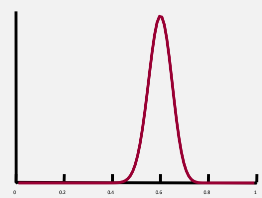
- Define a likelihood function which is a measure of how likely it is that the observed data were generated from a probability distribution with a particular choice of parameters
- Select the parameters that maximize the likelihood
  - In simple cases, ML estimate has a closed form solution
  - In complex cases, ML estimation may require numerical optimization – as in the case of distributions parameterized by Neural networks, e.g., large language models
- Problem with ML estimate – assigns zero probability to unobserved values – can lead to difficulties when estimating from small samples

# Bayesian Estimation

- MLE commits to a specific value of the unknown parameter (s)
- MLE is the same in both cases shown




vs.




- Of course, in general, one cannot summarize a function by a single number!
- Intuitively, the confidence in the estimates should be different

# Bayesian Estimation

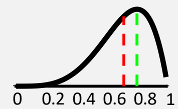
Maximum Likelihood approach is Frequentist at its core

- Assumes there is an unknown but fixed parameter $\theta$
- Estimates $\theta$ with some confidence
- Prediction of probabilities using the estimated parameter value

Bayesian Approach

- Represents uncertainty about the unknown parameter
- Uses probability to quantify this uncertainty:
  - Unknown parameters are treated as random variables
- Prediction follows from the rules of probability:
  - Expectation over the unknown parameters

# Binomial Data Revisited




- Suppose $D$ is such that $(NH, NT) = (4,1)$
- Suppose that we choose a uniform prior $p(\theta) = 1 \quad \forall \theta \in [0,1]$
- $P(\theta|D)$ is proportional to the likelihood $L(\theta : D)$

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{\int_0^1 p(D \mid \theta)p(\theta)d\theta}$$

In this case, $p(D \mid \theta) = \begin{pmatrix} 5 \\ 1 \end{pmatrix} \theta^4 (1-\theta)^1$ and $\forall \theta \in [0,1], p(\theta) = \frac{1}{1-0} = 1$

$$\int_0^1 p(D \mid \theta)p(\theta) = \begin{pmatrix} 5 \\ 1 \end{pmatrix} \int_0^1 \left(\theta^4 - \theta^5\right) d\theta = \begin{pmatrix} 5 \\ 1 \end{pmatrix} \left[\frac{\theta^5}{5} - \frac{\theta^6}{6}\right]_0^1 = \begin{pmatrix} 5 \\ 1 \end{pmatrix} \frac{1}{30}$$

$$p(\theta \mid D) = 30\theta^4(1-\theta)$$

$$P(x_6 = H|D) = \int_0^1 p(\theta \mid D)\theta d\theta = 30\int_0^1 \theta^4(1-\theta)\theta d\theta = 30\left[\frac{\theta^6}{6} - \frac{\theta^7}{7}\right]_0^1 = \frac{5}{7} = 0.7142$$

# Example: Binomial Data Revisited

- Suppose $D$ has $M = N_H + N_T$ samples where $(NH, NT\,) = (4{,}1)$
- MLE for $\theta = P(x_6 = \ H\,)$ is $4/5\ =\ 0.8$
- Bayesian estimate is $P(x_6 =\ H\,)\ = 0.7142\cdots$

In this example, MLE and Bayesian predictions differ

It can be proved that

- If the prior is well-behaved – i.e. does not assign 0 density to any feasible parameter value
  - Then both MLE and Bayesian estimate converge to the same value in the limit as the number of samples approach ∞
- Both almost surely converge to the underlying distribution of $X$
- The ML and Bayesian approaches behave differently when the number of samples is small

## All prior beliefs are not created equal

- In practice we may have reason to believe that the prior distribution of the parameter of interest is not uniform
- We might want to assert priors that allow us to express our beliefs regarding the parameter to be estimated
- Example: We might want a prior that assigns a higher probability to parameter values that describe a fair coin than it does to an unfair coin
- The beta distribution allows us to capture such prior beliefs

## Beta distribution

- Let $x$ be an integer that is greater than $0$
- Let $\Gamma(x) = (x-1)!$
- This implies $\frac{\Gamma(x+1)}{\Gamma(x)} = x$
- The Beta density function $Beta(\theta: a, b)$, with parameters $a, b$, and $N = a + b$ where $a > 0$ and $b > 0$ are positive integers, is given by:

$$p(\theta) = \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} \theta^{a-1} \theta^{b-1} \text{ where } 0 \leq \theta \leq 1$$

# Beta distribution

- Suppose $D \ = \{x_1, \cdots x_N\}$ be random samples from Binomial distribution where $N_H + N_T = N$
- Then it can be shown that if $p(\theta) = Beta(\theta; a, b)$,

$$p(\theta|D) = Beta(\theta; a + N_H, b + N_T)$$

Update of the parameter with a beta prior based on data yields a beta posterior

## Conjugate Families

- When the posterior distribution follows the same parametric form as the prior distribution we say that we have a conjugate prior
- Conjugate priors are useful because:
  - For many distributions we can represent them with hyper parameters
  - They permit sequential update of the posterior based on data
  - In many cases we have closed-form solution for prediction
- Beta prior is a conjugate prior for the binomial likelihood

## Bayesian prediction

- Beta prior implies Beta posterior

$$P(x_M = H|D) = \frac{a + N_H}{N + M} = \frac{a + N_H}{(a + b) + (N_H + N_T)}$$

- Thus, we can update the posterior by simply replacing
  - $a$ by $(a + N_H)$ and
  - $b$ by $(b + N_T)$
- That is, we are doing relative frequency estimates, where $a$ and $b$ are counts that represent prior beliefs about $\theta$, the probability of heads
- Choosing $a = b$, implies we assume that the random mechanism is fair unless the data tells us otherwise

# Dirichlet Priors

- Recall that the multinomial likelihood function is $L(\Theta : D) = \prod_{k=1}^{K} \theta_k^{N_k}$
- A Dirichlet prior with hyperparameters $\alpha_1, \cdots, \alpha_K$ is defined as

$$P(\Theta) = \frac{\Gamma(N)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}; \qquad 0 \leq \theta_k \leq 1; \qquad \sum_{k=1}^{K} \theta_k = 1$$

where $\Theta = (\theta_1 \ldots \theta_K)$

- Under the Dirichlet prior, $P(x_1 = k) = \frac{\alpha_k}{\sum_{j=1}^{K} \alpha_j}$
- Then given the samples $D$ with observed counts $N_1, \cdots, N_K$ for the K different outcomes, the posterior has the same form, with hyperparameters $\alpha_1 + N_1, \cdots, \alpha_K + NK$
- Dirichlet posterior is $P(x_{M+1} = k|D) = \frac{\alpha_k + N_k}{\sum_{j=1}^{K} (\alpha_j + N_j)}$
- Dirichlet priors are conjugate priors for the multinomial distribution

# Intuition behind Beta and Dirichlet priors

- The hyperparameters $\alpha_1, \cdots, \alpha_K$ can be thought of as imaginary counts from our prior experience
- Equivalent sample size $= \alpha_1 + \cdots + \alpha_K$
- The larger the equivalent sample size the more confident we are in our prior

# Conjugate Families

- The property that the posterior distribution follows the same parametric form as the prior distribution is called conjugacy
  - Dirichlet prior is a conjugate family for the multinomial likelihood
- Conjugate families are useful because:
  - For many distributions we can represent them with hyperparameters
  - They allow for sequential update within the same representation
  - In many cases we have closed-form solution for prediction

# Summary of Bayesian estimation

- Treat the unknown parameters as random variables
- Assume a prior distribution for the unknown parameters
- Update the distribution of the parameters based on data – easy if we have conjugate priors
- Use Bayes rule to make prediction