Biostatistics and Informatics

BTM6000

Session 3: Probability distributions

Properties of probability mutually exclusive events

- If events A and B are mutually exclusive (e.g. C^+ = having cancer, C^- =not having cancer):

(Venn Diagram)

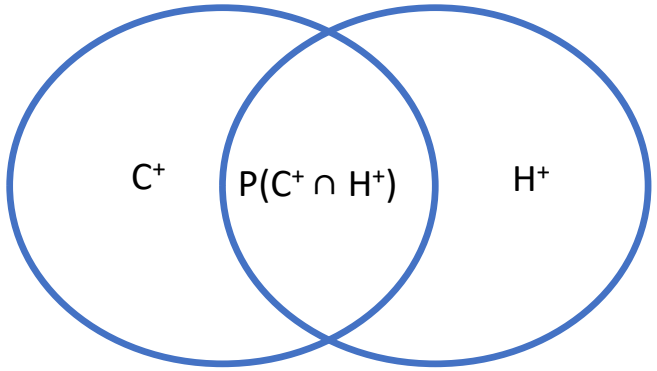
- $P(C^+ \cup C^-) = P(C^+) + P(C^-)$

  = 1 since options are also exhaustive

$P(C+ \cap C-) = 0$ since they are exclusive

Properties of probability non-exclusive events

e.g. C^+ = having cancer, H^+= having heart disease



U = union, “or”
∩ = intersection, “and”

- $P(C^+ \cup H^+) = P(C^+) + P(H^+) - P(C^+ \cap H^+)$
- $P(C^+ \cap H^+) = P(C^+) * P(H^+)$ *if independent*
- $P(C^+ \cap H^+) = P(C^+ \mid H^+) * P(H^+)$ *if non-independent*

Probability notation cheat-sheet

- A U B = *union* of A and B, “A or B”
- A ∩ B = *intersection* of A and B, “A and B”
- A^c = A *complement*, “not A”.
- P(A | B) = *conditional probability*. “Probability of A conditionally on B being true,” or “probability of A given B.”
- P(A ∩ B) = *joint probability*. Probability of both A and B being true.
- P(A)=*marginal probability* of A independently of B

Random variables - definitions

- A ***random variable*** – any characteristic that can be measured or categorized, and where any particular outcome is determined at least partially by chance.
  - Examples:
    - # of new diabetes cases in a population during a given year
    - The weight of a randomly selected individual in a population
  - Types:
    - Categorical random variable
    - Discrete random variable
    - Continuous random variable

Random variables - notation

- Random variables will be written in upper case roman letters: X, Y, etc
- Particular realizations of a random variable will be written in corresponding lower case letters: x, y, etc.
  - For example, x_1, x_2, …, x_n could be several observations of the random variable X.

Examples

- Landsteiner blood groups (data for the US)

GROUP	%
O+	37.40%
A+	35.70%
B+	8.50%
AB+	3.40%
O-	6.60%
A-	6.30%
B-	1.50%
AB-	.60%
TOTAL	*100%*

What kind of random variable is this?

Categorical

Examples

- Number of visits to the obstetrician's office during an uncomplicated pregnancy

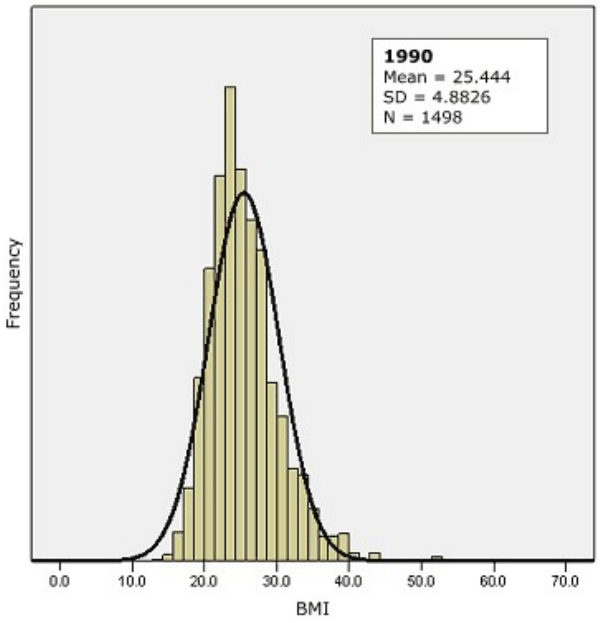
Visit	Fraction of women
0	.2
1	.2
2	.5
3	.1

What kind of Random Variable is this?

Discrete Quantitative

Examples

- The distribution of BMI in a population

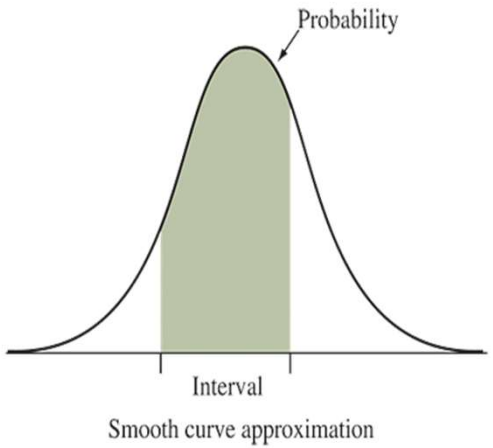


What kind of random variable is this?

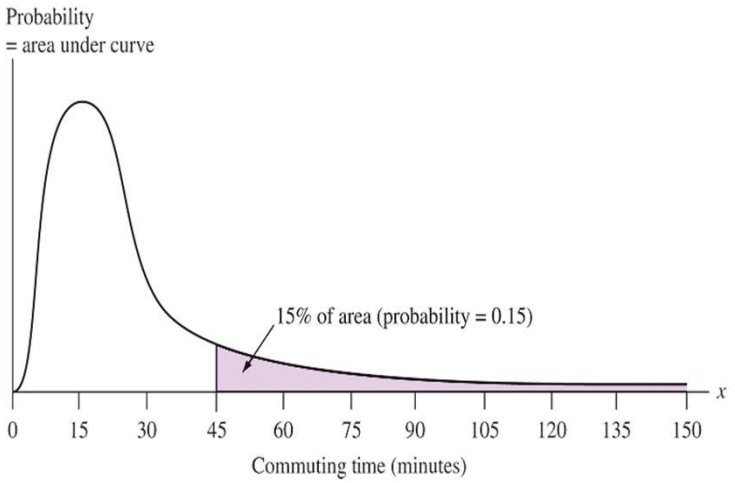
Continuous

Continuous Random Variables

- The probability distribution of a continuous random variable is a curve with allowed values on the x-axis, and probabilities on the y-axis
  - We will be interested in *intervals* along this curve
  - Each interval has cumulative probability between 0 and 1
  - The interval containing all possible values has probability equal to 1



Smooth curve approximation



Commuting time (minutes)

Two most important probability distributions

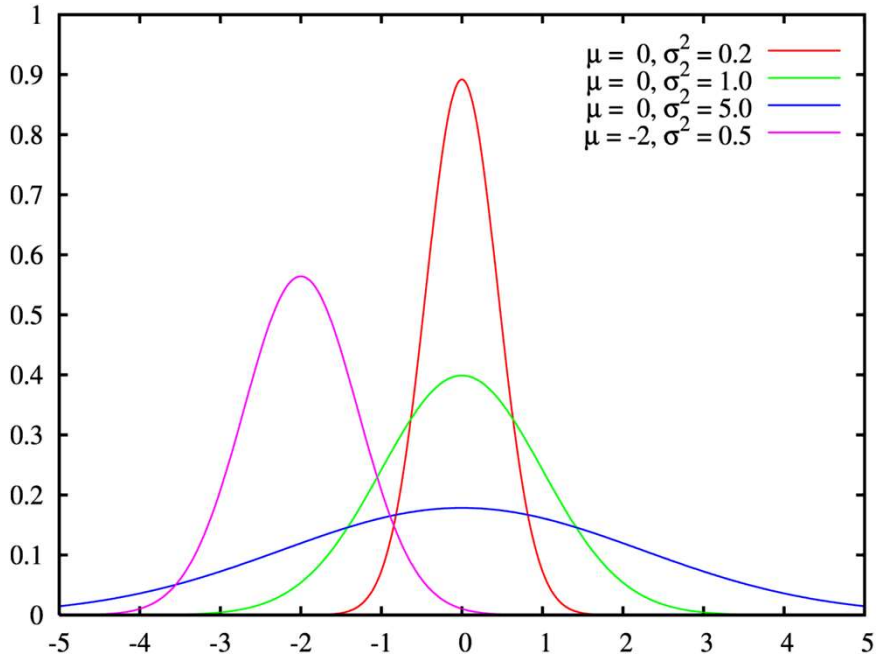
- For continuous outcome: **Normal Distribution**
  - Used when the distribution of a continuous outcome is described by a bell-shaped curve
- For binary outcome: **Binomial distribution**
  - A discrete probability distribution for the outcome of a number of independent yes/no experiments, each of which has a fixed probability of yes / no (or heads/tails, health/sick, etc).

Normal Distribution

The **normal distribution** is

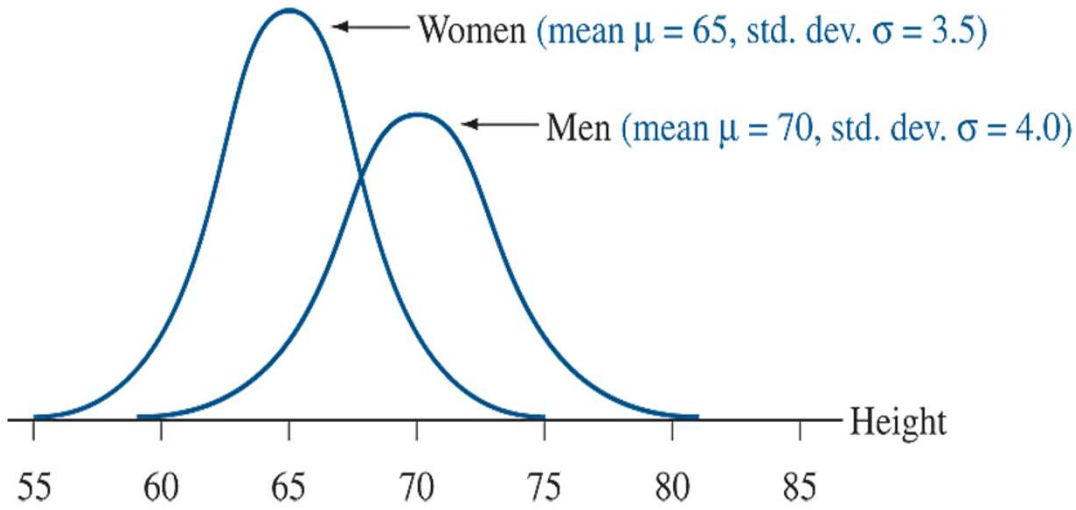
- symmetric, bell-shaped
- characterized by mean μ and standard deviation σ

Examples:



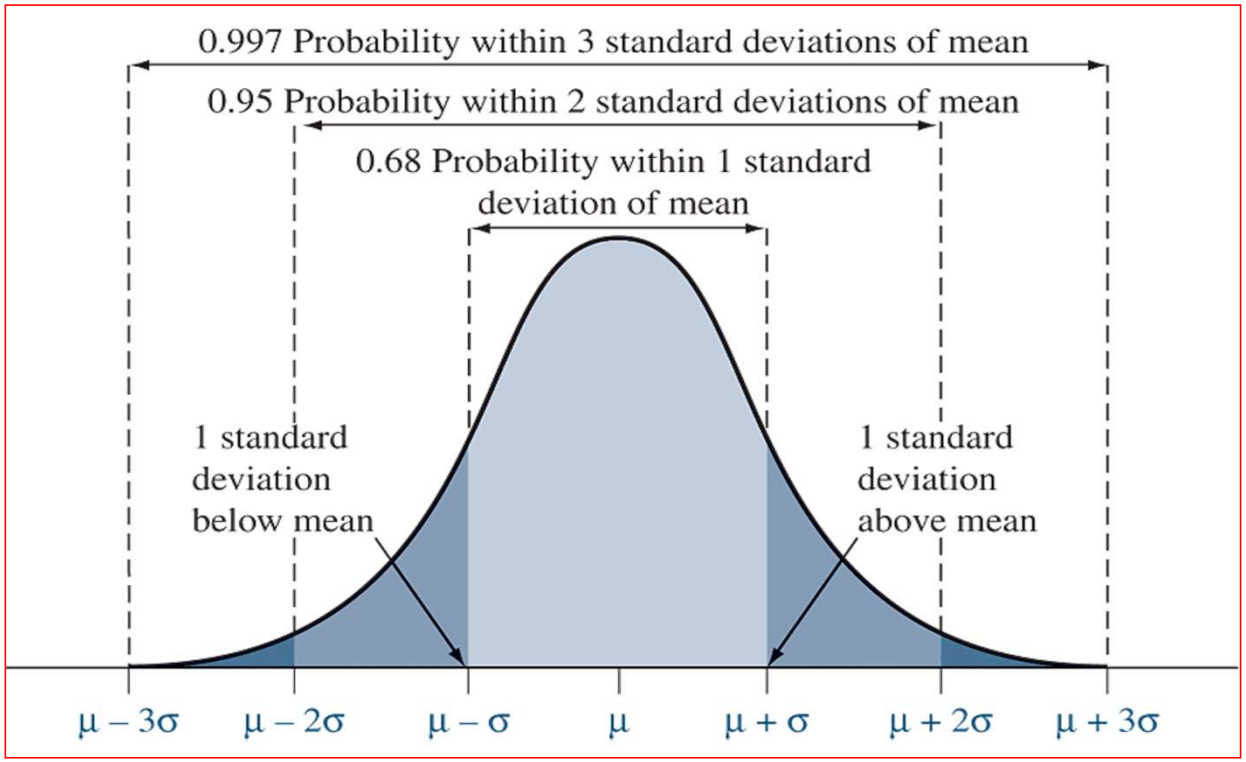
Normal Distribution

Examples:



- Within what interval do almost all of the men's heights fall? Women's height?

Empirical Rule
68-95-99.7 Rule for Any Normal Curve



Example: Empirical (68-95-99.7%) Rule

- Heights of adult women is approximately normal with μ= 65 inches and σ=3.5 in.

  - **68-95-99.7 Rule for women's heights**
    - 68% are between 61.5 and 68.5 inches

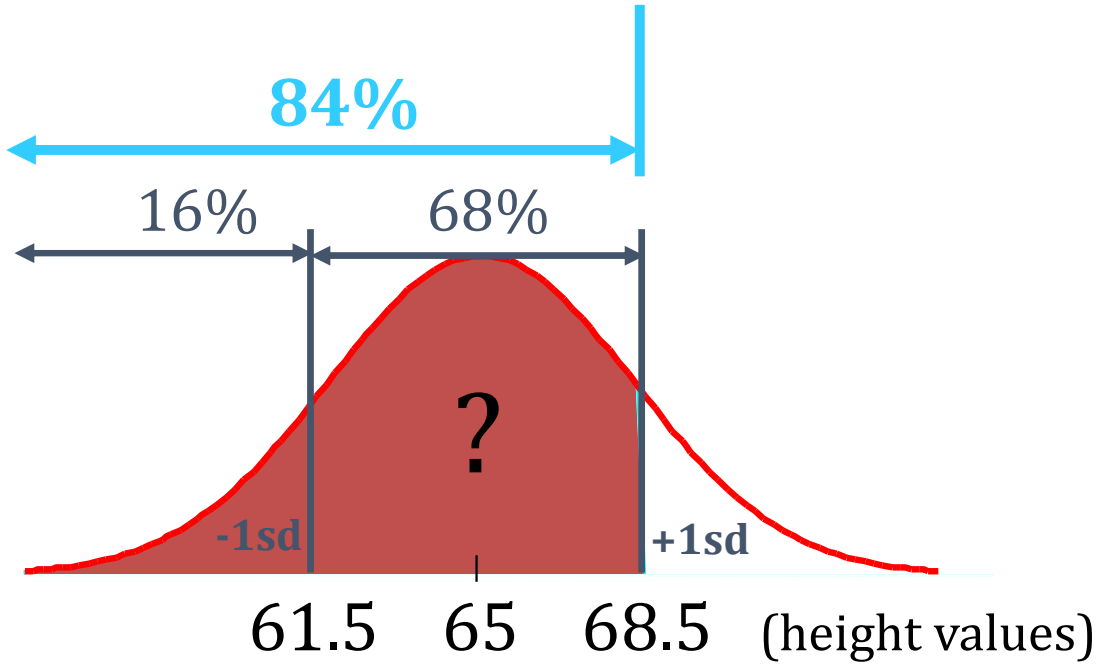
      $[\ \mu \pm \sigma = 65 \pm 3.5\ ]$
    - 95% are between 58 and 72 inches

      $[\ \mu \pm 2\sigma = 65 \pm 2(3.5) = 65 \pm 7\ ]$
    - 99.7% are between 54.5 and 75.5 inches

      $[\ \mu \pm 3\sigma = 65 \pm 3(3.5) = 65 \pm 10.5\ ]$

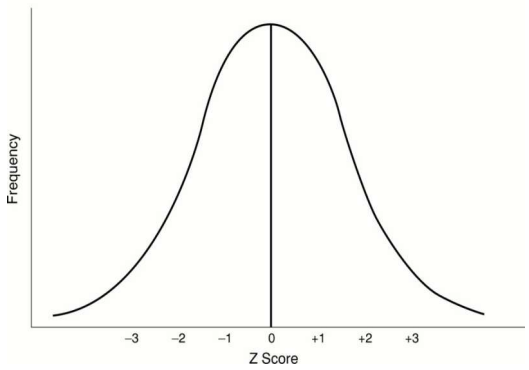
Example: Empirical (68-95-99.7%) Rule

- What proportion of women are less than 68.5 inches tall?



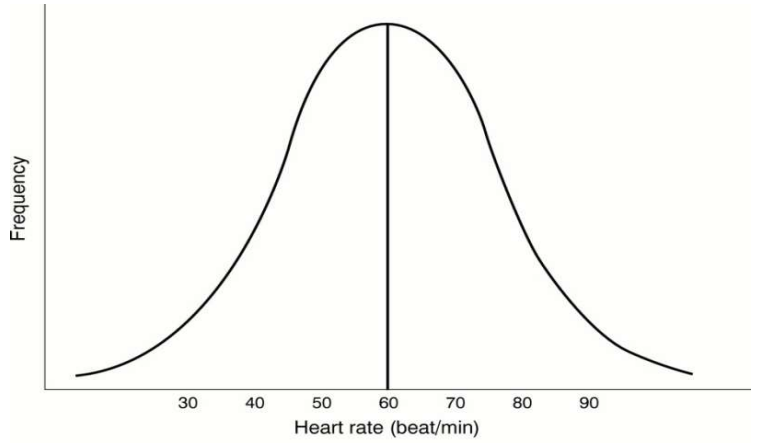
Z-Scores and the Standard Normal Distribution

- Every normal distribution can be standardized to mean zero (μ=0) and standard deviation one (σ=1). This is the **Standard Normal Distribution.**
  - The equation is **$Z=(X-\mu)/\sigma$**
  - Observations transformed using this equation are called *Z scores*

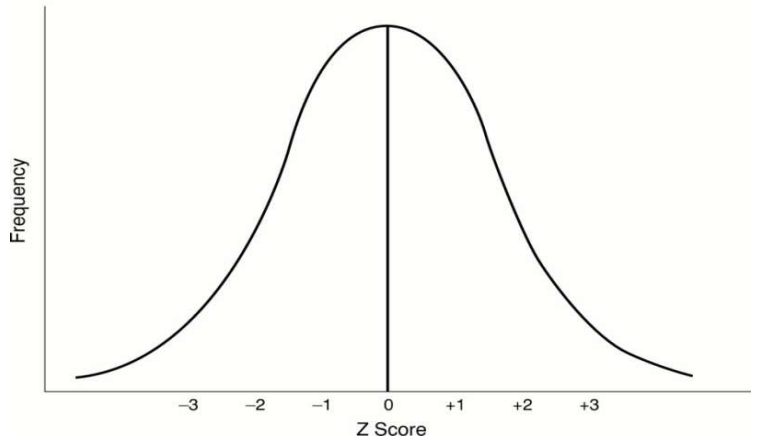


Example of standardization

Normal distribution

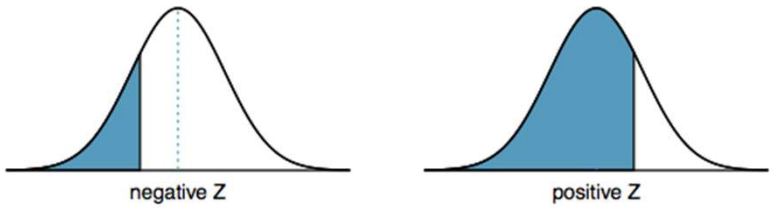


The same distribution but standardized

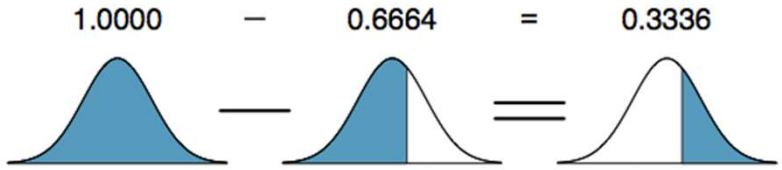


Z-Scores and the Standard Normal Distribution

- Makes it easier to compute probabilities using a table of probabilities for the standard normal distribution, for the *lower tail* or *upper tail.*
  - OpenIntro text, Appendix B.1 “Normal Probability Table” (this is area in the left or lower tail)

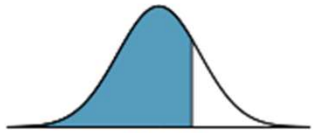


To find the area to the right, calculate 1 minus the area to the left.



Finding probabilities
Table B.1 “Normal Probability Table”

- Find the probability that a normal random variable takes a value **less than** 1.43 standard deviations above μ



	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

P(z<1.43)= 0.9236

Probability that the variable takes a value **greater** than μ+1.43*σ:

P(z>1.43) = 1 – P(z<1.43) = 0.0764

Some important Z values

- $P(Z > 1.28) = 0.10$
- $P(Z > 1.645) = 0.05$
- $P(Z > 1.96) = 0.025$
- $P(Z > 2.32) = 0.01$
- $P(Z > 2.57) = 0.005$

- NORM.DIST function in Excel

Finding probabilities using the Z distribution

- Find the probability that a normal random variable assumes a value within 1.43 standard deviations of μ

$P(-1.43 < Z < 1.43) = P(Z < 1.43) - P(Z < -1.43)$

or, $= P(Z > -1.43) - P(Z > 1.43)$

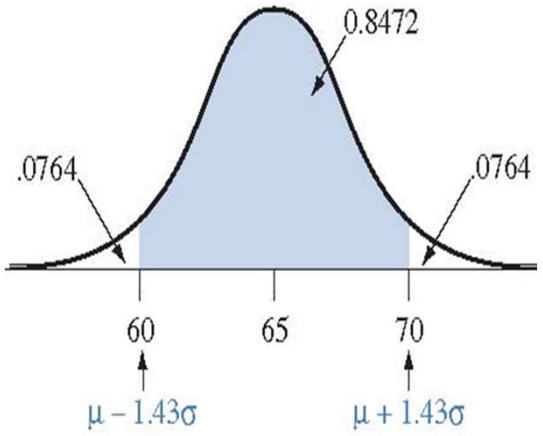
$= 0.9236 - 0.0764 = 0.8472$

OR $= 1 - 2 * P(Z < -1.43)$

$= 1 - 2 * P(Z > 1.43)$

$= 1 - 2*0.0764$

$= 0.8472$

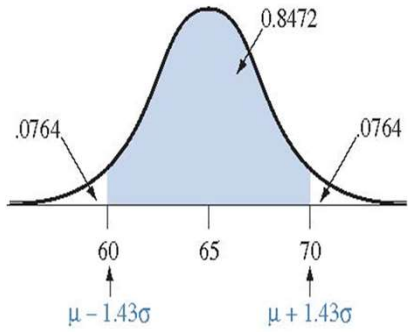


Finding probabilities using the Z distribution

- **SUGGESTION**: When solving problems draw normal curve and shade area corresponding to desired probability
- Find probability to the left of z = -1.64
  - P(Z < -1.64)=.0505
- Find probability to the right of z = 1.56
  - P(Z > 1.56)= 1 - 0.9406 = 0.0594
- Lab will cover how to get these probabilities from software using probability functions.

How Can We Find the Value of z for a Certain Cumulative Probability?

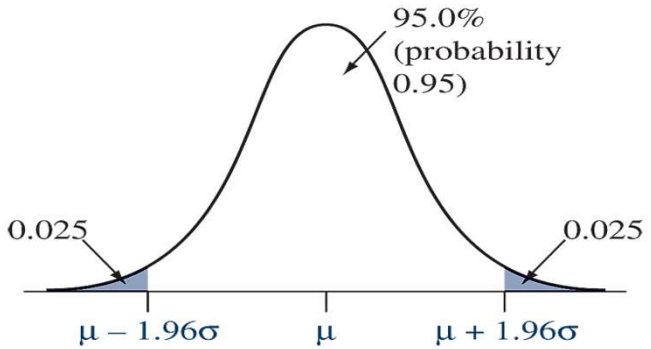
- To solve some of our problems, we will need to find the value of *Z* that corresponds to a certain normal cumulative probability
- To do so, we use table in reverse
  - Find the probability in the body of the table
  - Read the corresponding z-score
  - In software, use quantile function



How Can We Find the Value of *z* for a Certain Cumulative Probability?

E.g. find the value of *z* for a cumulative probability of 0.025.

- Look up the cumulative probability of 0.025 in the body of Table A.
- A cumulative probability of 0.025 corresponds to z =-1.96.



Find the value of *z* for a cumulative probability of 0.975.

- Look up the cumulative probability of 0.975 in the body of the table
- A cumulative probability of 0.975 corresponds to z =1.96

Finding Probabilities for Normally Distributed Random Variables

1. **State the problem in terms of X and x** (random variable and one possible realization), i.e., P(X<x)
2. **Standardize X** to restate the problem in terms of a standard normal variable Z

$$P(X < x) = P\left(Z < \frac{x-\mu}{\sigma}\right)$$

3. **Draw a picture** to show the desired probability under the standard normal curve
4. **Find the area** under the standard normal curve using a software

The Binomial Distribution

- Can be used for **repeated, independent** events with only two possible values
  - E.g., binary categorical variables
- Example of one type of event (also called trial):
  - Has, or does not have cancer
- Sample question:
  - What is the probability that among 100 randomly selected individuals in NYC, 60 have private health insurance?

The Binomial Distribution

- We can use the binomial distribution to compute probabilities if:
  - Each trial has two possible outcomes: "success" & "failure"
  - Each trial has the same probability of success (*p)*
  - The *n* trials are independent

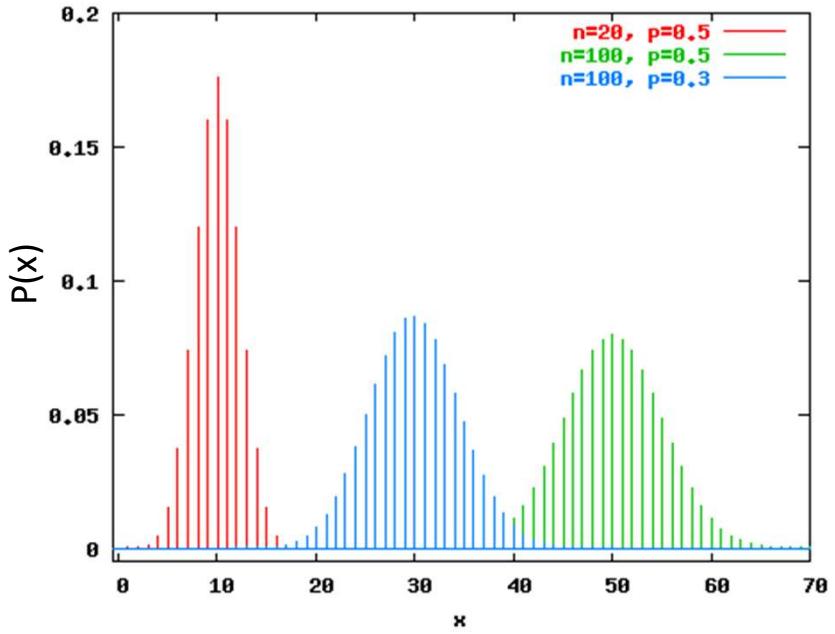
- The probability of x successes in *n* trials, where each has a probability of success p:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

(! is factorial)

The Binomial Distribution

- Binomial Distribution looks like a probability for each number of "successes" (x) out of n "trials" each with a probability of success p:



Example

- The prevalence of diabetes in an elderly population is 10% (p=0.10)

- What is the probability that if you randomly select **4** individuals from this population,

  <u>exactly</u> 3 of them will be diabetic?

$$P(x=3) = \frac{4!}{3!(4-3)!} 0.1^3 (1-0.1)^{4-3} = 0.0036$$

Example

What is the probability that if you randomly select **4** individuals from this population, **at most one** individual will be diabetic?

$$P(x=0) + P(x=1) = 0.6561 + 0.2916 = 0.9477$$

Example

- What is the probability that if you randomly select **4** individuals from this population that <u>at **least one**</u> of them will be diabetic?

- Note: P(x is at least one) = 1 – P(none)
- BINOM.DIST function in Excel

Normal approximation to binomial distribution

- The mean is $\mu = n * p$
- The standard deviation is $\sigma = \sqrt{n * p * (1 - p)}$

- Example of application:
  - If we sample 1000 individuals from a population with prevalence of diabetes equal to .1, we expect that about $n * p \rightarrow 1000 * .01 = 10$ individuals in the sample will have diabetes

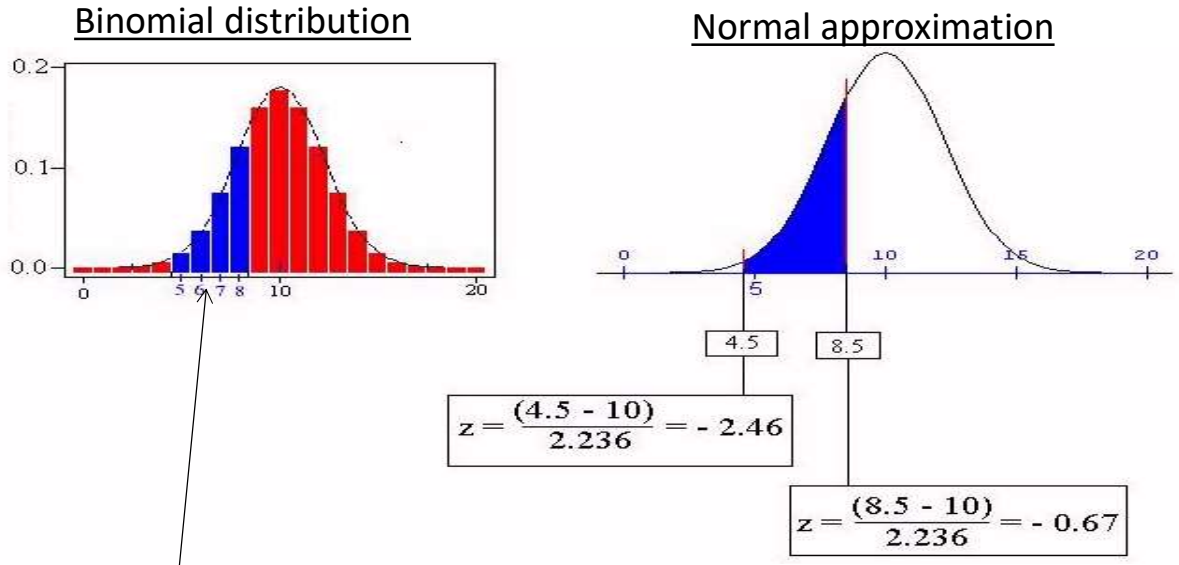
Normal approximation to binomial distribution

- The mean is $\mu=n*p$
- The standard deviation is $\sigma = sqrt(n*p*(1-p))$

- Example of application:
  - If we sample 1000 individuals from a population with prevalence of diabetes equal to .1, we expect that about $n*p \rightarrow 1000*.01= 10$ individuals in the sample will have diabetes

Approximating the Binomial Distribution with the Normal Distribution

- When the number of trials is large, binomial distribution is well approximated by the normal distribution
  - Works well if np & $n(1-p)$ are both at least 15



P(x=5)+P(x=6)+P(x=7)+P(x=8)

P(4.5<x<8.5)

Summary

- Probability notation and calculations are key
- We will use the normal distribution, binomial distribution, and normal approximation to the binomial distribution throughout this course
- New notation: X, x, n, p, μ, σ, Z
- probability calculations and look-ups for binomial and normal distributions
- Upper, lower tails of the normal distribution