Lesson 2

Probability

What is Probability?

- The study of uncertainty and randomness in the world.

Context

- So far
  - Summarizing data (based on variable type, numerical vs graphical)
- Now
  - Probability (accounting for uncertainty)
- Next
  - Statistical inference (generalize from sample to population)

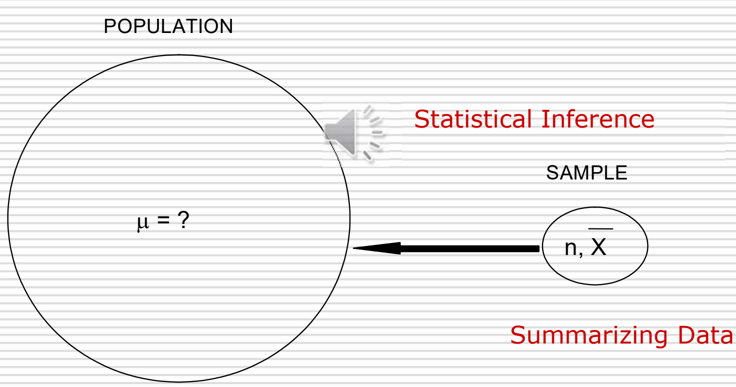
Probability

- What is the probability you will develop cardiovascular disease in the next 20 years?
- What is the likelihood that you have a hip replacement?
- What is the chance that you will develop cancer in your lifetime?

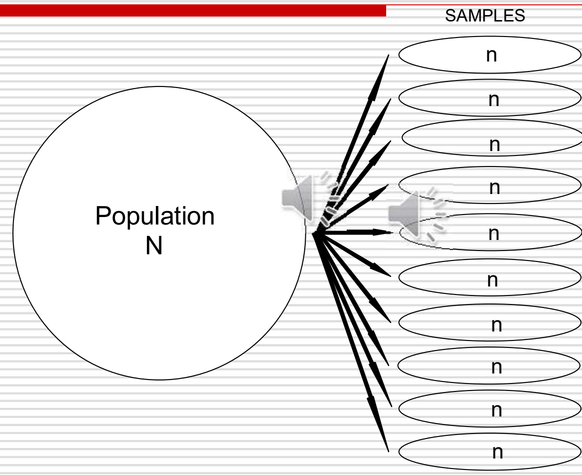
Objectives

- Understand probability as it pertains to statistical inference
- Understand the attributes and applications of popular probability models
- Understand and apply the results of the Central Limit Theorem

Two Areas of Biostatistics



Sampling from a Population



Basics

- Probability reflects the likelihood that outcome will occur.
- $0 \leq \text{Probability} \leq 1$.

$$\text{Probability} = \frac{\text{Number with outcome}}{\text{N}}$$

New York City - Cancer Registry 2008*

Type	White Male	White Female	Black Male	Black Female	Total
Colorectal	1236	1251	449	584	3520
Liver	330	134	149	57	670
Lung	1449	1332	537	497	3815
Thyroid	175	537	29	135	876
Non-Hodgkins Lymphoma	582	523	170	159	1434
Leukemia	348	285	87	85	805
Total	4120	4062	1421	1517	11120

*Selected data from
http://www.health.state.ny.us/statistics/cancer/registry/about.htm

Probability
A case is selected at random:

P(White Male)= 4120/11120 = 0.37

P(Black Male)= 1421/11120 = 0.13

P(Thyroid Cancer)= 876/11120 = 0.08

P(White Female with Liver Cancer)= 134/11120 = 0.01

P(Black Patient with Lung Cancer)= (537+497)/11120 = 0.09

What is the probability of selecting a male?

Blood Pressure Category

	Optimal	Normal	Pre-Htn	Htn	Total
Male	20	15	15	30	80
Female	5	15	25	25	70
Total	25	30	40	55	150

What is the probability of selecting a male with optimal blood pressure?

Blood Pressure Category

	Optimal	Normal	Pre-Htn	Htn	Total
Male	20	15	15	30	80
Female	5	15	25	25	70
Total	25	30	40	55	150

What is the probability of selecting a patient with Pre-Htn or Htn?

Blood Pressure Category

	Optimal	Normal	Pre-Htn	Htn	Total
Male	20	15	15	30	80
Female	5	15	25	25	70
Total	25	30	40	55	150

What proportion of men have prevalent CVD?

	CVD	Free of CVD
Men	35	265
Women	45	355

What proportion of patients with CVD are men ?

	CVD	Free of CVD
Men	35	265
Women	45	355

What proportion of white females have thyroid cancer?

Type	White Male	White Female	Black Male	Black Female	Total
Colorectal	1236	1251	449	584	3520
Liver	330	134	149	57	670
Lung	1449	1332	537	497	3815
Thyroid	175	537	29	135	876
Non-Hodgkins Lymphoma	582	523	170	159	1434
Leukemia	348	285	87	85	805
Total	4120	4062	1421	1517	11120

P(thyroid cancer|white female) = 537/4062 = 0.13

Probability

Definition of Independent Events:

A and B are independent if:

P(A) = P(A|B) or

P(B)= P(B|A) or

P(A and B) = P(A) * P(B)

For data with N=11120:

Type of cancer and race/ethnicity are not independent.

(P(Thyroid cancer) = 0.08 ≠ P(Thyroid cancer | White Female) = 0.13)

Probability

Example. Consider the following table which cross classifies subjects by their family history of CVD and current (prevalent) CVD status.

	Current CVD	
Family History	No	Yes
No	215	25
Yes	90	15

Are family history and current status independent?

Probability

P(Current CVD)= 40/345 = 0.116

P(Current CVD| Family Hx)

= 15/105 = 0.143

P(Current CVD| No Family Hx)

= 25/240 = 0.104

Independent?

Performance Characteristics of Screening Tests

	Disease +	**Disease -**
Test +	a	b
Test -	c	d

Performance Characteristics of Screening Tests

Disease + means you have disease.
Disease - means you don't have disease.

Test + means a positive test result
Test - means a negative test result.

Performance Characteristics

- Sensitivity = True Positive Fraction = P(Test + | Disease)
- Specificity = True Negative Fraction = P(Test - | No Disease)

Performance Characteristics

- False Positive Fraction = P(Test + | No Disease)
- False Negative Fraction = P(Test - | Disease)

Which is worse?

1. A higher false positive fraction
2. A higher false negative fraction
3. Equally bad
4. It depends

Which of the following affect your decision to take the test?

1. The chance you test positive if you have disease.
2. The chance you test negative if you do not have disease.
3. The chance you have disease if you test positive.
4. The chance you don't have disease if you test negative.

Performance Characteristics

- Positive Predictive Value = P(Disease | Test +)
- Negative Predictive Value = P(No Disease | Test -)

Sensitivity and Specificity

	Affected Unborn Baby	Unaffected Unborn Baby	Total
Positive Screen	9	351	360
Negative Screen	1	4449	4450
Total	10	4800	4810

Sensitivity and Specificity

Sensitivity = P(test +|disease)
=9/10=0.90
Specificity = P(test -|disease free)
= 4449/4800 = 0.927

False negative fraction= P(test -|disease)
= 1/10 = 0.10
False positive fraction=P(test +|disease free)
= 351/4800 = 0.073

Should you have this test?

I had the test. Now what?

My test was positive…
Positive Predictive Value = P(Disease|Test +)
= 9/360 = 0.025

My test was negative…

Negative Predictive Value = P(No Disease|Test -)
= 4449/4450 = 0.9998

I was afraid to have the test.
What is the chance I have an affected fetus?
P(Disease | No Test) = 10/4810 = 0.002

What is the sensitivity of the test summarized below?

	Positive	Negative	Total
Disease	12	5	17
No Disease	8	22	30
Total	20	27	47

What is the false positive fraction of the test?

Practice problem

- 1% of women at age forty who participate in routine screening have breast cancer.
- 80% of women with breast cancer will get positive mammographies.
- 9.6% of women without breast cancer will also get positive mammographies.

A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

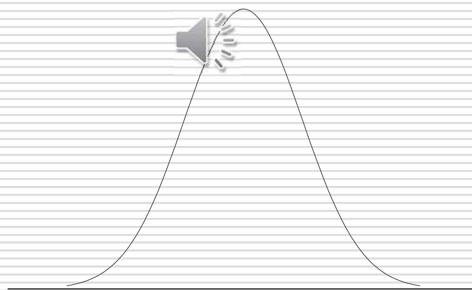
Hint

Create a hypothetical population of women and complete the following table.

	BC	No BC
S+		
S-		

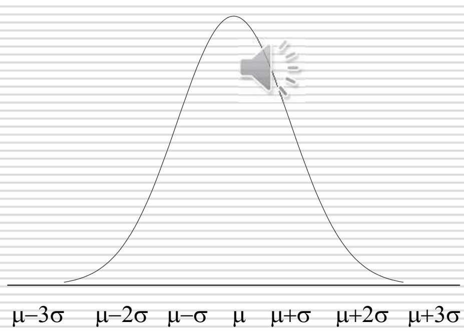
Normal Distribution

- Model for continuous outcome
- Mean=median=mode



Normal Distribution

Notation: μ=mean and σ=standard deviation



Properties of the Normal Distribution

i) The normal distribution is symmetric about the mean

(i.e., $P(X > \mu) = P(X < \mu) = 0.5$).

ii) The mean and variance, μ and σ^2, completely characterize the normal distribution.

iii) The mean = the median = the mode.

$P(\mu - \sigma < X < \mu + \sigma) = 0.68$,

$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$,

$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.99$

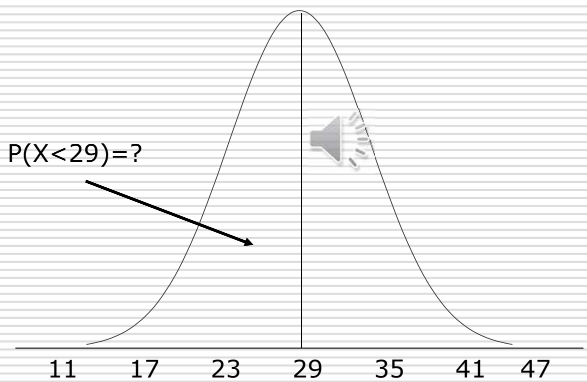
iv) $P(a < X < b)$ = the area under the normal curve from a to b.

Normal Distribution

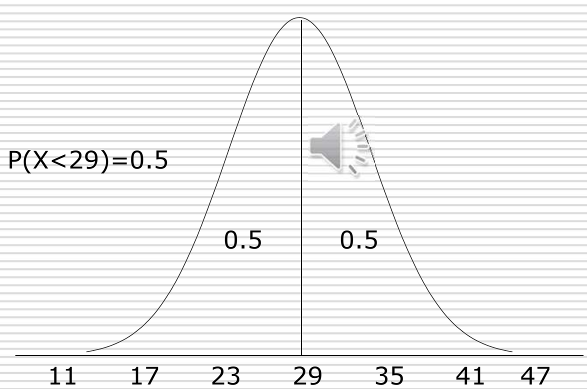
Body mass index (BMI) for men age 60 is normally distributed with a mean of 29 and standard deviation of 6

What is the probability that a male has BMI less than 29?

Normal Distribution



Normal Distribution

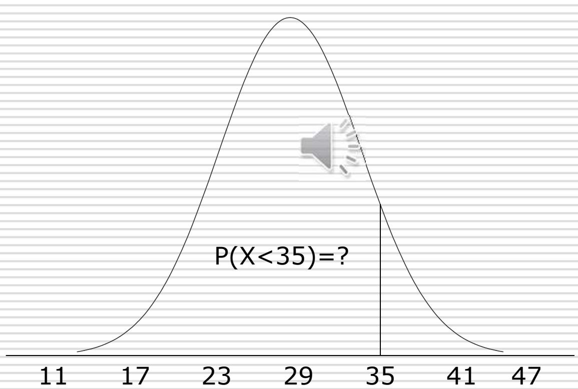


Normal Distribution

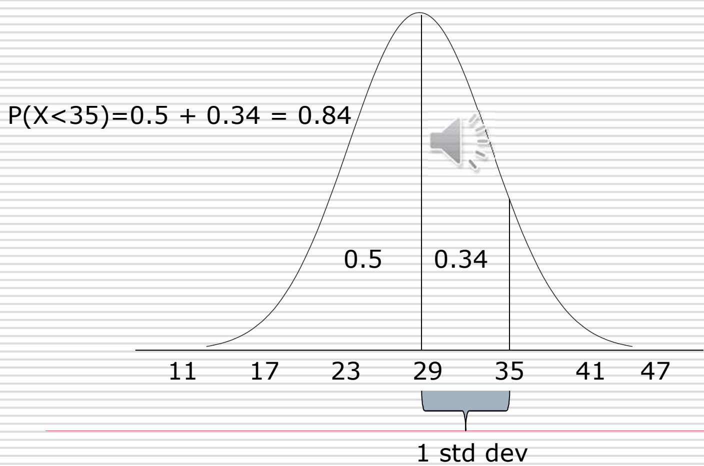
Body mass index (BMI) for men age 60 is normally distributed with a mean of 29 and standard deviation of 6

What is the probability that a male has BMI less than 35?

Normal Distribution

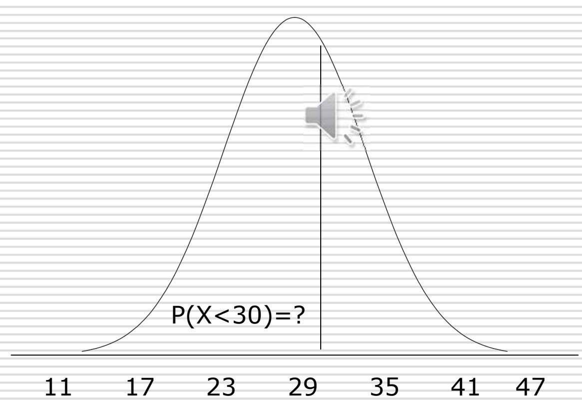


Normal Distribution



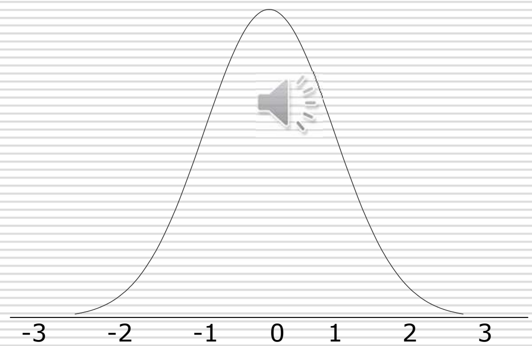
Normal Distribution

What is the probability that a male has BMI less than 30?



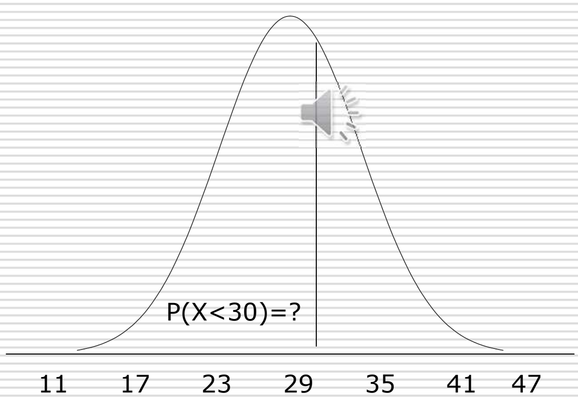
Standard Normal Distribution Z

Normal distribution with μ=0 and σ=1

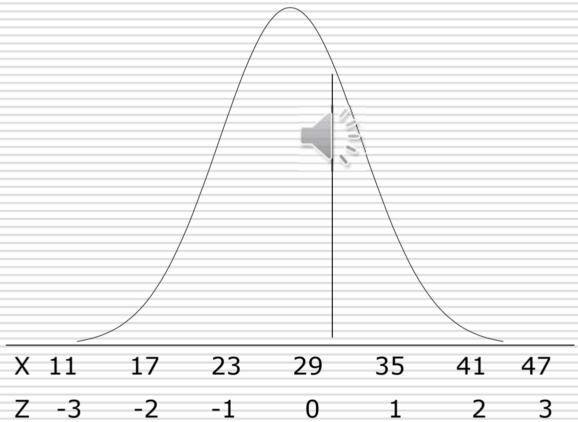


Normal Distribution

What is the probability that a male has BMI less than 30?



Normal Distribution



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

Standardize

$$Z = \frac{x - \mu}{\sigma} = \frac{30 - 29}{6} = 0.17$$

P(X<30)= P(Z<0.17) = 0.5675

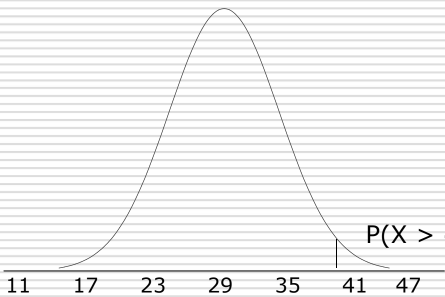
Table 1. Probabilities of the Standard Normal Distribution Z (continued)

Table entries represent P(Z < Z_i)

e.g., P(Z < -1.96) = 0.0250, P(Z < 1.96) =0.9750

Z_i	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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

What is the probability that a male has BMI greater than 40?



$$Z = \frac{x - \mu}{\sigma} = \frac{40 - 29}{6} = 1.83$$

P(X > 40)= P(Z>1.83) = 1 – 0.9664 = 0.0336

Comparing Systolic Blood Pressure (SBP)

Comparing systolic blood pressure (SBP): Suppose

- Males Age 50, SBP is approximately normally distributed with a mean of 108 and a standard deviation of 14.
- Females Age 50, SBP is approximately normally distributed with a mean of 100 and a standard deviation of 8.

If a Male Age 50 has a SBP = 140 and a Female Age 50 has a SBP = 120, who has the "relatively" higher SBP ?

Normal Distribution

Z_M = (140 - 108) / 14 = 2.29

Z_F = (120 - 100) / 8 = 2.50

JMP example

- Open arrhythmia dataset in JMP PRO
- Examine the distributions of the following variables: QRS duration, P-R interval, Q-T interval, T interval and P interval
- Check each for normality using Normal Quantile Plot option

Percentiles of the Normal Distribution

The k^{th} *percentile* is defined as the score that holds k percent of the scores below it.

For example: 90^{th} percentile is the score that holds 90% of the scores below it.

Q1 = Lower Quartile = 25^{th} percentile,

median = 50^{th} percentile

Q3 = Upper Quartile = 75^{th} percentile

Percentiles

For the normal distribution, the following is used to compute percentiles:

$$X = \mu + Z\,\sigma$$

where

μ = mean of the random variable X,

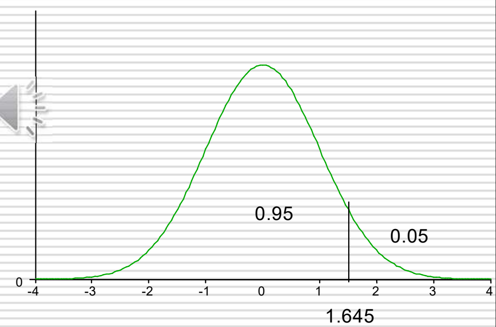
σ = standard deviation, and

Z = value from the standard normal distribution for the desired percentile (See next slide).

Percentiles

Percentiles of the Standard Normal Distribution

Percentile	Z
1^{st}	-2.326
2.5^{th}	-1.960
5^{th}	-1.645
10^{th}	-1.282
50^{th}	0
90^{th}	1.282
95^{th}	1.645
97.5^{th}	1.960
99^{th}	2.326



Percentiles of the Normal Distribution

BMI in men follows a normal distribution with $\mu=29$, $\sigma=6$. BMI in women follows a normal distribution with $\mu=28$, $\sigma=7$.

The 90^{th} percentile of BMI for men:

$X = 29 + 1.282\ (6) = 36.69.$

The 90^{th} percentile of BMI for women:

$X = 28 + 1.282\ (7) = 36.97.$

JMP example

- Open arrhythmia dataset in JMP PRO
- Interpret the 5th and 95th percentiles of QRS duration, P-R interval, Q-T interval, T interval and P interval?
- Check out the Quantile Box Plot option
- Compute the 20th and 80th percentiles using the Display Options – Custom Quantiles

Central Limit Theorem

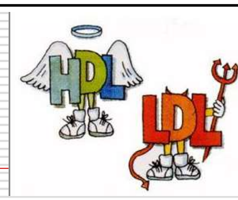
Suppose we have a population with known mean μ and standard deviation σ. If we take simple random samples of size n with replacement, then for large n, the sampling distribution of the sample means is approximately normal with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

Application

- Non-normal population
- Take samples of size n – as long as n is sufficiently large (usually n $\geq$ 30 suffices)
- The distribution of the sample mean is approximately normal, therefore can use Z to compute probabilities

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

Central Limit Theorem



HDL cholesterol has a mean of 54 and standard deviation of 17 in patients over 50. A physician has 40 patients over age 50 and wants to know the probability that their mean cholesterol is above 60.

$$P(\bar{X} > 60) = ?$$

Central Limit Theorem

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{60 - 54}{17 / \sqrt{40}} = 2.22$$

$$P(\bar{X} > 60) = P(Z > 2.22) = 1 - 0.9868 = 0.0132$$

Practice Problem- Comparing Systolic Blood Pressure (SBP)

- Total cholesterol in children aged 10-15 is assumed to follow a normal distribution with a mean of 191 and a standard deviation of 22.4.
  1. What proportion of children 10-15 years of age have total cholesterol between 180 and 190?
  2. What proportion of children 10-15 years of age would be classified as hyperlipidemic (defined as a total cholesterol level over 200)?
  3. If a sample of 20 children are selected, what is the probability that the mean cholesterol level in the sample will exceed 200?

1. What proportion of children 10-15 years of age has total cholesterol between 180 and 190?

$P(180 < X < 190) = P(\frac{180-191}{22.4} < Z < \frac{190-191}{22.4}) = P(-0.49 < Z < -0.04) = 0.4840 - 0.3121 = 0.1719.$

2. What proportion of children 10-15 years of age would be classified as hyperlipidemic (Assume that hyperlipidemia is defined as a total cholesterol level over 200)?

$P(X > 200) = P(Z > \frac{200-191}{22.4}) = P(Z > 0.40) = 1-0.6554 = 0.3446.$

3. If a sample of 20 children are selected, what is the probability that the mean cholesterol level in the sample will exceed 200?

$P(\bar{X} > 200) = P(Z > \frac{200-191}{22.4/\sqrt{20}}) = P(Z > 1.80) = 1-0.9641=0.0359.$

JMP Project 2