Basics of Probability

Brian Healy, PhD

Harvard Catalyst Applied Biostatistics Certificate Program

Lecture outline

- Probability
- Conditional probability
- Binomial distribution
- Normal distribution
- Central limit theorem

Note on practicum

- The practicum for this week will focus on creating new variables
  - Continuous --> Categorical
- We will not show how to calculate specific probabilities in STATA since this is rarely done in practice
  - *display binomial(n,x,p)*
  - *display normal(z)*

Goals of lecture

- At the end of this lecture, you will:
  - Understand different types of probabilities
  - Understand the concept of conditional probability
  - Understand the binomial distribution
  - Be able to calculate the probability of specific events based on the binomial and normal distributions

Big picture

- A basic understanding of probability is critical for anyone using biostatistics
- In this lecture, we will describe the basics to set up future lectures
  - We move a little quickly in this lecture, but please ask questions if you are confused
- More than any other lecture in this course, you may need to review some of the concepts on your own after the lecture

Examples of probability

- Weather forecasts
  - "There is a 80% chance of rain on Thurs"
- Gambling odds
  - "The Patriots are favored by 6 points in Sunday’s game"
- Medical research
  - p-value
  - Relative risk
  - Sensitivity/specificity

Definition of probability

- From Rosner: "Probability of an event is the relative frequency of this outcome over an infinitely large number of trials"
- How likely is an event?
- Must be between 0 and 1
  - 0 implies there is no chance that the event occurs
  - 1 implies the event must occur

Single events

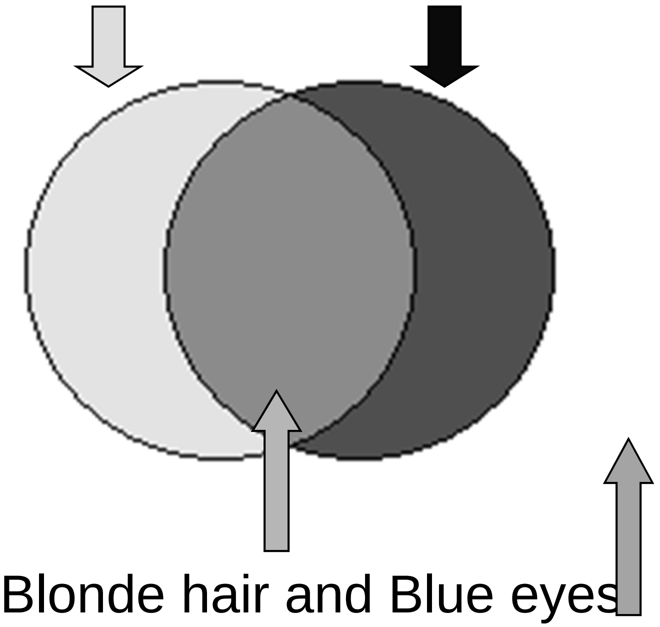
- P(A): probability of event A
  - P(blonde hair): probability that a person has blond hair
- P(not A): probability of event A not occurring
  - Complement
  - P(not blonde hair): probability that a person does not have blonde hair

Multiple events

- P(A and B)=P(A,B): probability that BOTH event A and event B occur
  - Intersection
  - P(blonde hair and blue eyes): probability that a person has BOTH blonde hair and blue eyes
- P(A and/or B): probability of either event A or B or both have occurred
  - Union
  - P(blonde hair and/or blue eyes): probability that a person has either blonde hair or blue eyes or both

Venn diagram

- A **Venn diagram** is a simple graphical way to represent 2 (or more events)
- The box represents all possible events
- What color or colors represent:
  - – P(blonde hair and blue eyes)
  - – P(blonde hair and/or blue eyes)



Blonde Hair

Blue eyes

Blonde hair and Blue eyes

Neither blonde hair nor blue eyes

Conditional probability

- Definition: Given that an event has occurred, what is the probability of a second event?
- Conditional probability restricts the sample space by requiring that a specific event occurs
- Examples:
  - Given that a person has blonde hair, what is the probability that the person has blue eyes?
  - Given that it is raining today, what is the probability that it will rain tomorrow?
  - Given that I want the Red Sox to win, what is the probability that they will win tonight?

- The probability of A given B (conditional probability) is

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

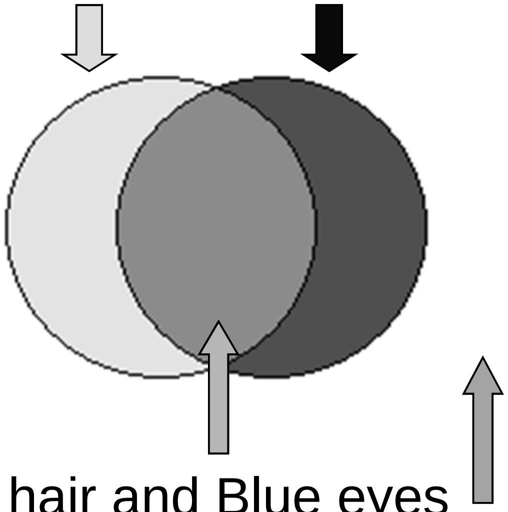
  - The denominator is the probability of event B which we know has happened
  - The numerator of this expression is the probability that both events occur together
- As we will see in couple of slides, sensitivity is an example of conditional probability
- As we will see next lecture, p-values are examples of conditional probabilities

Example

- Suppose we are interested in the probability of having blue eyes given that you have blonde hair. In terms of the Venn diagram, conditioning on having blonde hair means that we restrict our attention to people with blonde hair

Blonde Hair

Blue eyes



Blonde hair and Blue eyes

Neither blonde
hair or blue eyes

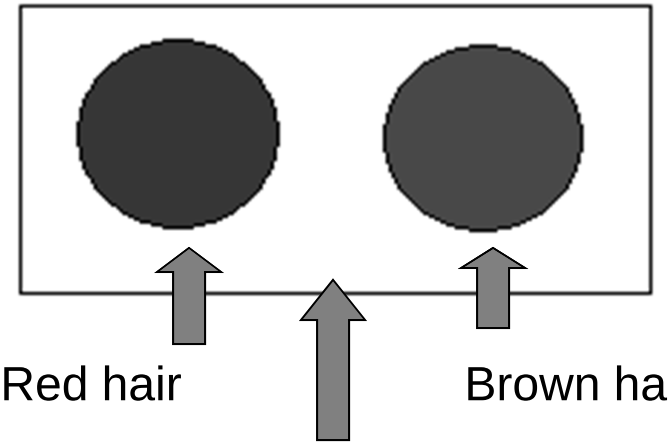
$$P(BE \mid BH) = \frac{P(BE, BH)}{P(BH)} = \frac{green}{yellow + green}$$

Other types of events

- Mutually exclusive: Two events (A and B) that cannot occur at the same time
  - P(A|B) = 0
  - Ex. P(smoker | nonsmoker) = 0
  - **When two events are mutually exclusive, P(A and/or B) = P(A) + P(B)**
- Exhaustive: A set of events that covers all possible events
- Mutually exclusive and exhaustive events: The set of events that cover all possibilities and the probability sums to 1

Venn diagram

- The diagram shows three mutually exclusive events, red hair, brown hair, and other color hair
- The three events are also exhaustive because they cover all possible events
- P(Red hair)+P(Brown hair)+P(Any other color)=1
- Does P(Red hair) + P(Not red hair) = 1? Why?



Red hair

Brown hair

Any other color hair

Does P(Red and/or Brown hair)= P(Red hair)+P(Brown hair)?

Independent events

- Independent: The occurrence of event B does not affect the probability of A occurring
  - P(A|B) = P(A)
  - Ex. P(coin toss 2 is heads | coin toss 1 is tails) = P(coin toss 2 is heads).
  - **When two events are independent, P(A and B)=P(A|B)P(B)=P(A)P(B)**
- Are two mutually exclusive events independent?

Examples

- Is having eye color independent of hair color?
  - P(Blue eyes)=1/6=0.1667
  - P(Blue eyes|Blonde hair)=4/5=0.8
- Is the outcome of the Red Sox game independent of me?
  - P(Red Sox win)=0.5
  - P(Red Sox win|I want them to win)=0.5

Summary

- Groups of events can be
  - Mutually exclusive
  - Exhaustive
  - Mutually exclusive and exhaustive
  - Independent
  - None of the above
- Classification of the events impacts the how to compute the probability of the events

Diagnostic tests

- Sensitivity: the probability of a positive test given that the patient has the disease, $P(T^+|D^+)$
- Specificity: the probability of a negative test given that the patient does not have the disease, $P(T^-|D^-)$
  - These two are when the test was correct
- False positive: the probability of a positive test given that the patient does not have the disease, $P(T^+|D^-)$
- False negative: the probability of a negative test given that the patient has the disease, $P(T^-|D^+)$
- Dr. Feldman will discuss this in great depth

Question

- Which of these sum to 1?
  - P(T+|D+)+P(T-|D+)=1
    - Sensitivity + False negative=1
    - If you have the disease, either you are a true positive or a false negative.
    - These are mutually exclusive and exhaustive
  - P(T+|D-)+P(T-|D-)=1

Probability distributions

Probability distributions

- In our initial discussions of probability, we assumed that the probability of an event was known
- In medical research, we often assume that the probability of an event follows a specific distribution
- These distributions describe the probability of a set of events with a small number of parameters

Discrete probability distribution

Give the possible outcomes of a future event and the probability associated with each (the sum of the probabilities must be 1)

Examples: Coin Toss

Event	Probability
Heads	0.5
Tails	0.5

Side Effects from a Medication

Event	Probability
Severe	0.08
Mild	0.12
None	0.8

Example

- In the classroom for this course, we have 30 people
- We know that the probability that a person in the US is left-handed equals 0.1
- What is the probability that there are
  - 0 left-handed people?
  - 3 left-handed people?
  - 0, 1, 2 or 3 left handed people?

Binomial distribution

- Discrete distribution for the number of successes (X) in a specific number of trials
- Defined by two values:
  - n=number of trials (n=30 in our example)
  - p=probability of success on each trial (p=0.1 in our example)
- Assumes only two options for each trial and each trial is independent
- Total number of successes is between 0 and n

Intuition

- A classic example of the binomial distribution is the number of tails on a set of coin tosses
- If we toss the coin once, what is the probability that we get tails?
  - P(tails)=0.5
  - P(heads)=1-P(tails)=0.5
- If we flipped a coin 2 times, what is the probability of 1 tails?

Possible outcomes

- In the coin example, there are four scenarios

Scenarios	Probability
T,T	p*p=0.5*0.5=0.25
T,H	p*(1-p)=0.5*0.5=0.25
H,T	(1-p)*p=0.5*0.5=0.25
H,H	(1-p)*(1-p)=0.5*0.5=0.25

- In terms of number of tails, this leads to

Number of tails	Probability
2	p*p=0.25
1	2*p*(1-p)=0.5
0	(1-p)*(1-p)=0.25

Class example

- If we returned to the left-handed example and focused on two people, there are four scenarios

Scenarios	Probability
L,L	p*p=0.1*0.1=0.01
L,R	p*(1-p)=0.1*0.9=0.09
R,L	(1-p)*p=0.9*0.1=0.09
R,R	(1-p)*(1-p)=0.9*0.9=0.81

- In terms of number of left-handed people:

Number of lefties	Probability
0	(1-p)*(1-p)=0.81
1	2*p*(1-p)=0.09+0.09=0.18
2	p*p=0.01

Class example

- Three people

	Probability		Probability
L,L,L	p*p*p=0.1*0.1*0.1=0.001	L,R,R	p*(1-p)*(1-p)=0.1*0.9*0.9=0.081
L,L,R	p*p*(1-p)=0.1*0.1*0.9=0.009	R,L,R	(1-p)*p*(1-p)=0.9*0.1*0.9=0.081
L,R,L	p*(1-p)*p=0.1*0.9*0.1=0.009	R,R,L	(1-p)*(1-p)*p=0.9*0.9*0.1=0.081
R,L,L	(1-p)*p*p=0.1*0.9*0.1=0.009	R,R,R	(1-p)*(1-p)*(1-p)=0.9*0.9*0.9=0.729

- In terms of number of left-handed people:

Number of lefties	Probability
0	(1-p)*(1-p)*(1-p)=0.729
1	3*p*(1-p)*(1-p)=0.081+0.081+0.081=0.243
2	3*p*p*(1-p)=0.009+0.009+0.009=0.027
3	p*p*p=0.001

What does this mean?

- In these two scenarios, we can calculate the probability of a specific event by just listing all possible events
- As the number of events increases, it becomes difficult to list all of the possibilities
- Is there a formula for the probability for a specific number of successes assuming a binomial distribution?

Formula

- Binomial formula

We have n-x "failures"

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

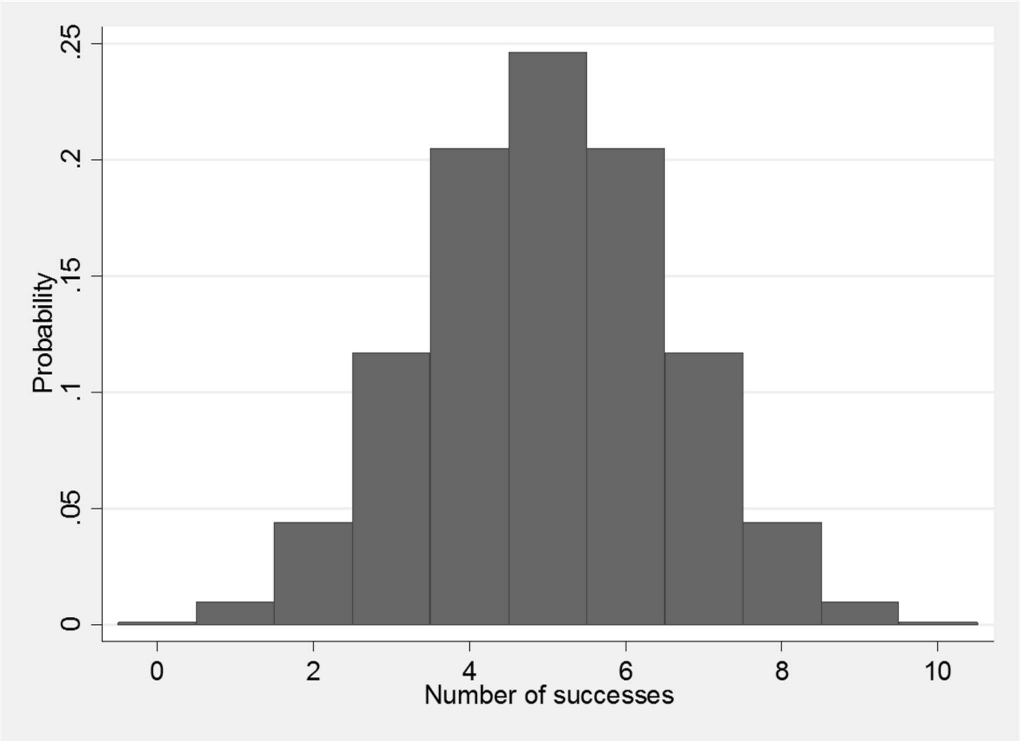
This is called "n choose x"

This corresponds to the number of ways you can get x successes out of n trials

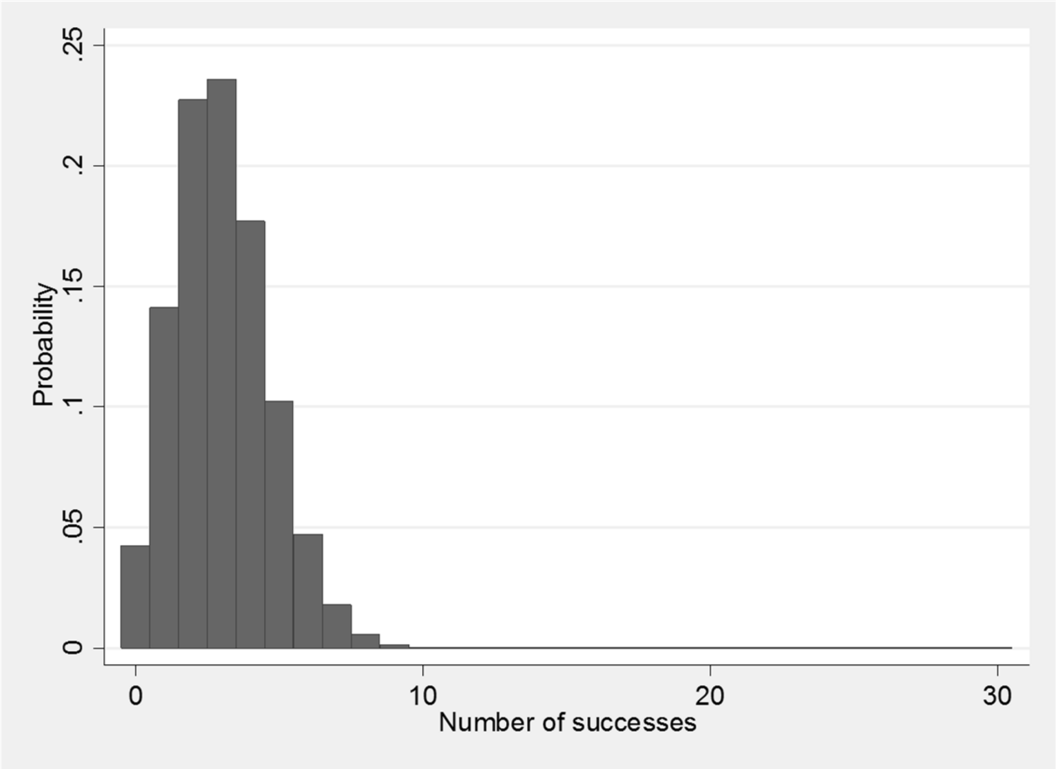
We have x "successes"

$$P(X = 1) = \binom{3}{1} 0.1^1 (1-0.1)^{3-1} = 3 * 0.1 * 0.9 * 0.9 = 0.243$$

Binomial with n=10 and p=0.5



Binomial with n=30 and p=0.1



Example

- Returning to our original example, we can calculate the probability of
  - 0 left-handed people
    - P(X=0)=0.042 – based on formula or STATA
  - 3 left-handed people
    - P(X=3)=0.236 – based on formula or STATA
  - 0, 1, 2 or 3
    - P(X<=3)=P(X=0)+ P(X=1)+ P(X=2)+ P(X=3)=.042+.141+.228+.236=0.647
    - We can add these since they are mutually exclusive

Mean and variance

- If we flip a fair coin 10 times, how many tails would we expect?
- In our class of 30 people, if the probability of being left-handed is 0.1, how many lefties would we expect?
- Expected value=Mean=np
- Variance=np(1-p), $SD = \sqrt{np(1-p)}$

Summary

- Binomial distribution appropriate for independent trials of a dichotomous outcome
- Using the binomial distribution, we can calculate the probability of a specific number of successes out of a specific number of trials
  - Very useful for hypothesis testing and confidence intervals
  - Lecture 6

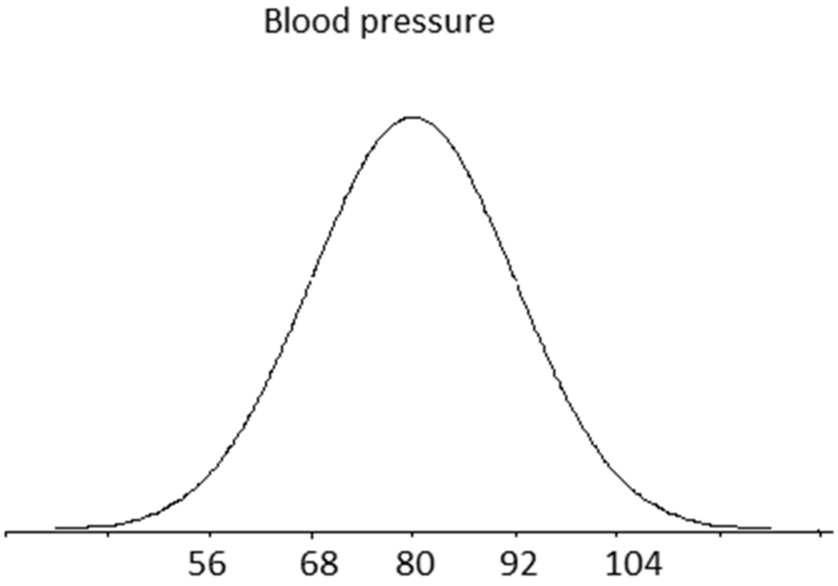
Normal distribution

Example

- One of the most important probability distributions in biostatistics is the normal distribution
- Several measurements follow an approximately normal distribution
  - Blood pressure
  - Birth weight

Picture

- Assume blood pressure in middle aged men has a normal distribution mean 80 and standard deviation 12 (from Rosner)



Properties of normal distribution

- Continuous random variable
- Range (-inf,inf)
- Two parameters
  - Mean: μ
  - Variance: σ^2
- Symmetric

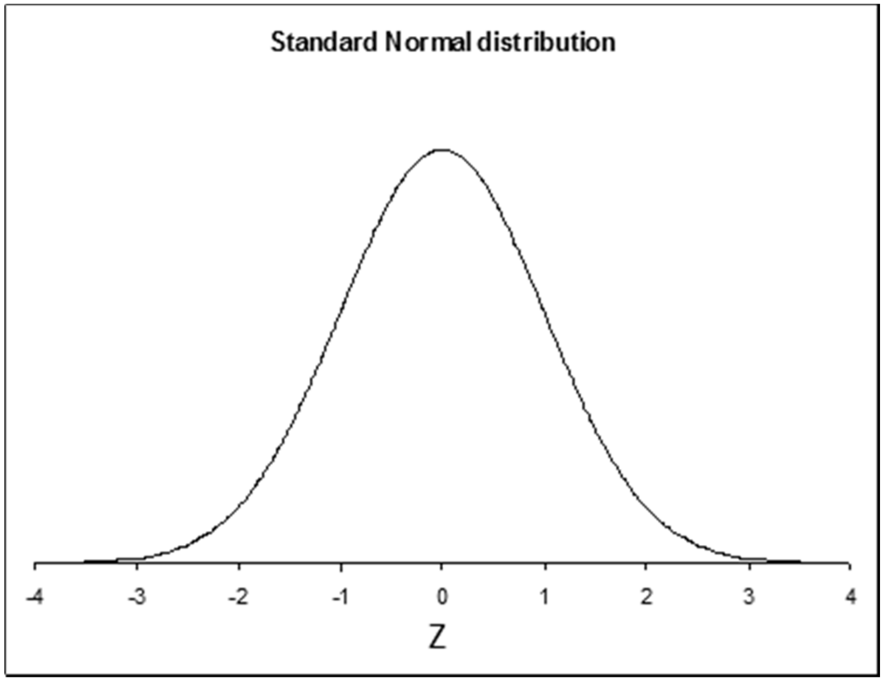
Probability

- Our focus in biostatistics will often be calculating the probability of a specific event, like with the binomial distribution
- For the normal, our focus will be the probability of being between two values or less than/greater than a value
  - For example, in the blood pressure example, we might want to know the probability of having a blood pressure less than 70

Normal distributions

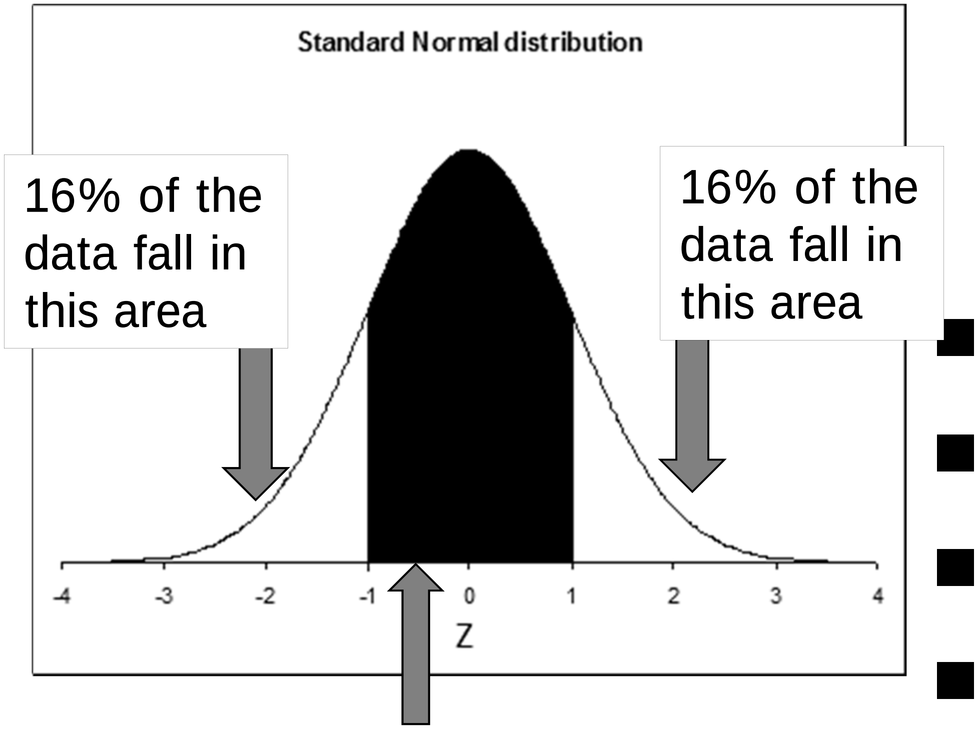
- A great feature of the normal distribution is that any normal distribution can be converted into a standard normal distribution by subtracting the mean and dividing by the standard deviation
- $Z = \frac{X-\mu}{\sigma}$

Standard normal distribution



- Mean=0
- SD=1
- "z-score" or "z-statistic"
- P(Z<=0)=0.5
- To calculate probabilities, we use the area under curve

Standard normal distribution



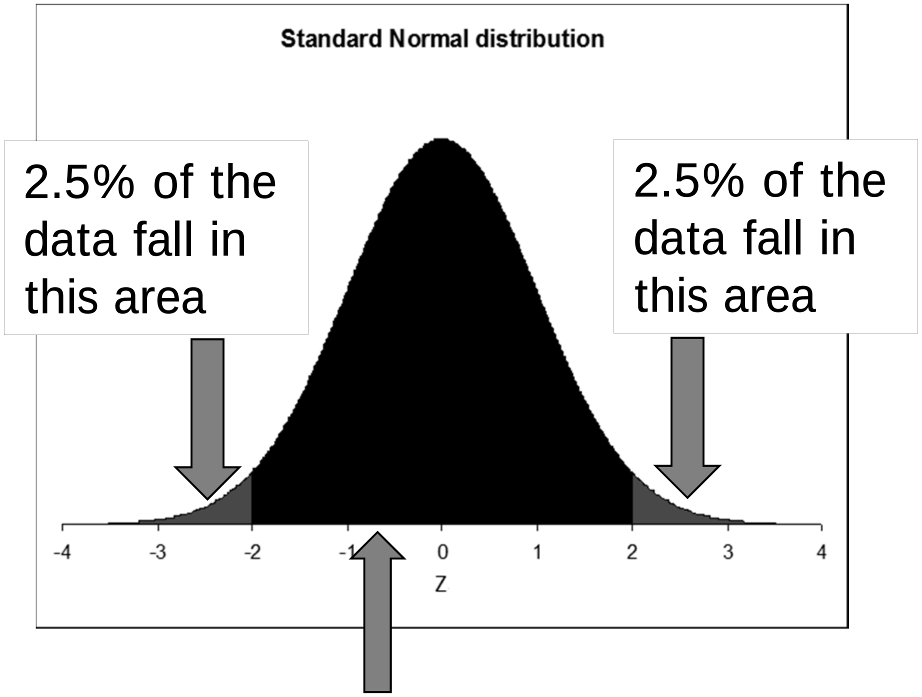
16% of the data fall in this area

16% of the data fall in this area

68% of the data fall in the shaded area

- In a normal distribution 68% of the data lie within one standard deviation of the mean
- $P(-1 \leq Z \leq 1) = 0.68$
- $P(Z \leq -1) = 0.16$
- $P(Z \geq 1) = 0.16$
- $P(Z \leq 1) = P(-1 \leq Z \leq 1) + P(Z \leq -1) = 0.84$

Standard normal distribution



2.5% of the data fall in this area

2.5% of the data fall in this area

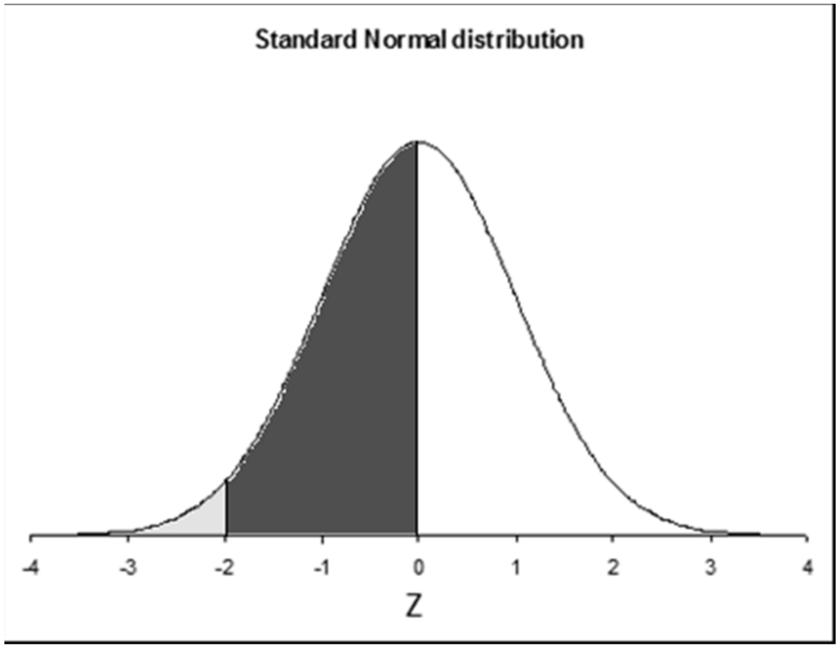
95% of the data fall in the black area

- In a normal distribution about 95% of the data lie within two standard deviations of the mean
- $P(-1.96 \leq Z \leq 1.96) = 0.95$
- $P(Z \leq -1.96) = 0.025$
- $P(Z \geq 1.96) = 0.025$
- $P(Z \leq 1.96) = P(-1.96 \leq Z \leq 1.96) + P(Z \leq 0.975$

Tail probabilities

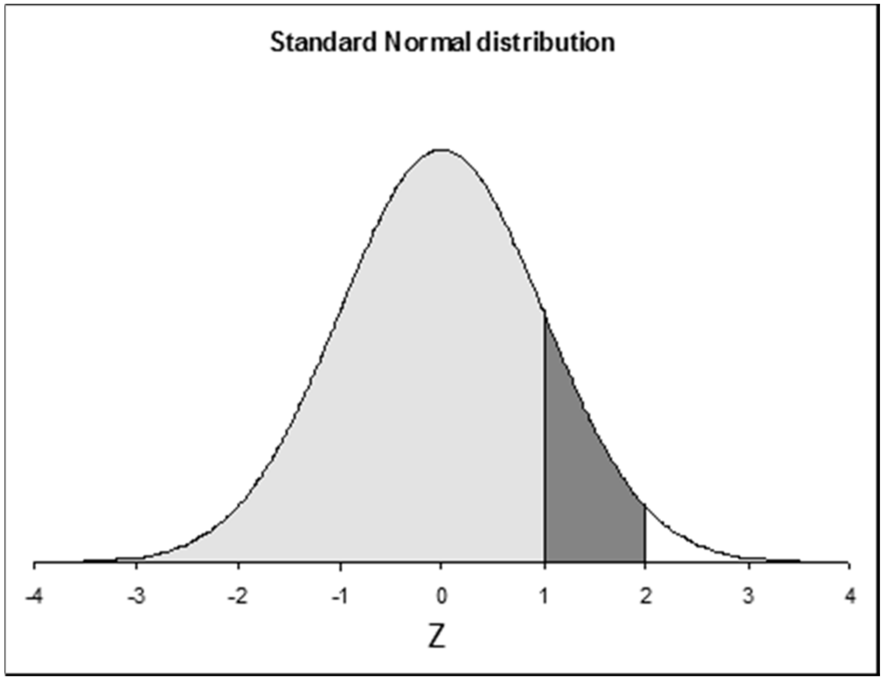
- The previous examples have described the probabilities associated with specific events
- The lower tail probability for all values of a standard normal, $P(Z \leq z)$, have been calculated
- This result is useful in calculating the probability of specific events and will help in calculating p-values

Examples



- The probability of any other event can be calculated as the appropriate area (blue)
- $P(-1.96 \leq Z \leq 0) = P(Z \leq 0) - P(Z \leq -1.96) = 0.5 - 0.025 = 0.475$

Examples



- The probability of any other event can be calculated as the appropriate area (blue)
- $P(1 \leq Z \leq 1.96) = P(Z \leq 1.96) - P(Z \leq 1) = 0.975 - 0.84 = 0.135$
- $P(1 \leq Z \leq 1.96) = P(Z \geq 1) - P(Z \geq 1.96) = 0.16 - 0.025 = 0.135$

Summary

- The probability of events from a normal distribution can be calculated
- STATA can be used to calculate the tail probabilities using this command:
  - *display normal(0)*
  - *display normal(-1.96)*
  - *display normal(0) -normal(-1.96)*

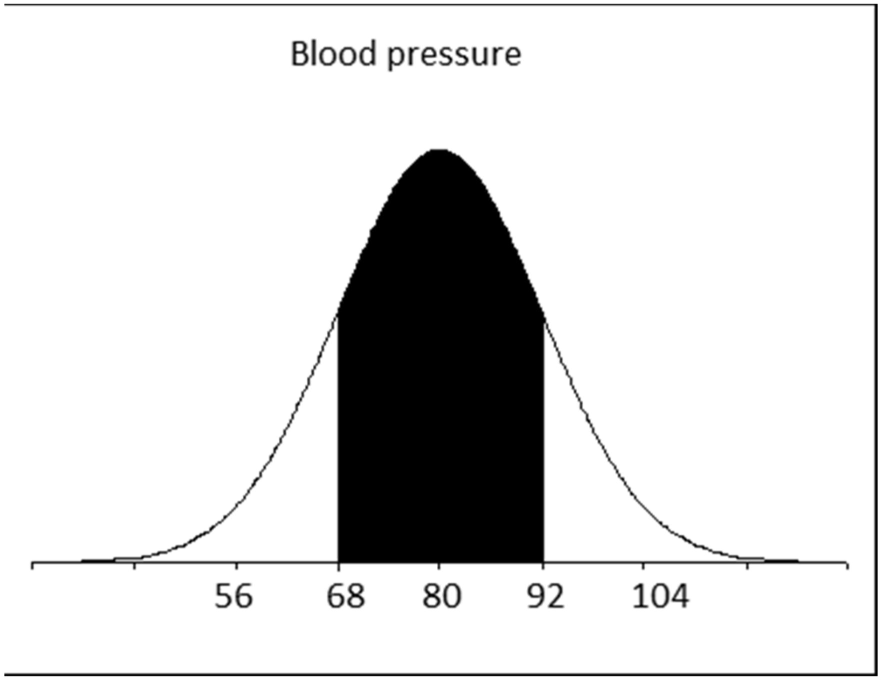
Other normal distributions

- Since any normal distribution can be converted into a standard normal distribution by subtracting the mean and dividing by the standard deviation, we can use this relationship to calculate the probability of events using the same procedure described above
- $Z = \frac{X-\mu}{\sigma}$

Example

- Assume the distribution of blood pressure in middle aged men is normal with $\mu = 80$ and $\sigma = 12$
- How could we calculate the probability that a man has a blood pressure between 68 and 92?
  - $P(68 \leq X \leq 92) = P\left(\frac{68-80}{12} \leq \frac{X-80}{12} \leq \frac{92-80}{12}\right) =$
  - $P\left(\frac{68-80}{12} \leq Z \leq \frac{92-80}{12}\right) =$
  - $P(-1 \leq Z \leq 1) = 0.68$

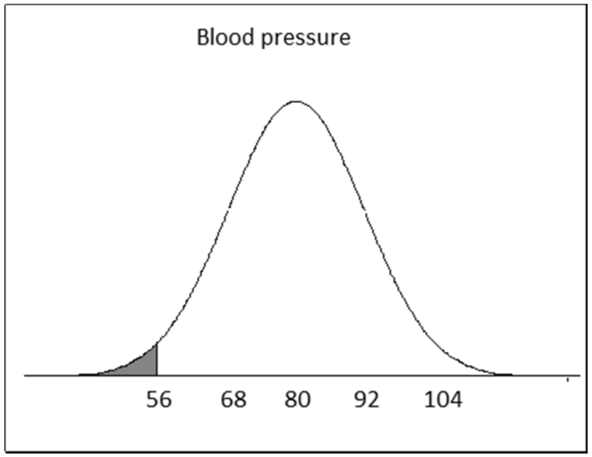
Picture



- The amount of area between 68 and 92 in this graph is exactly the same as between -1 and 1 on the standard normal

Additional examples

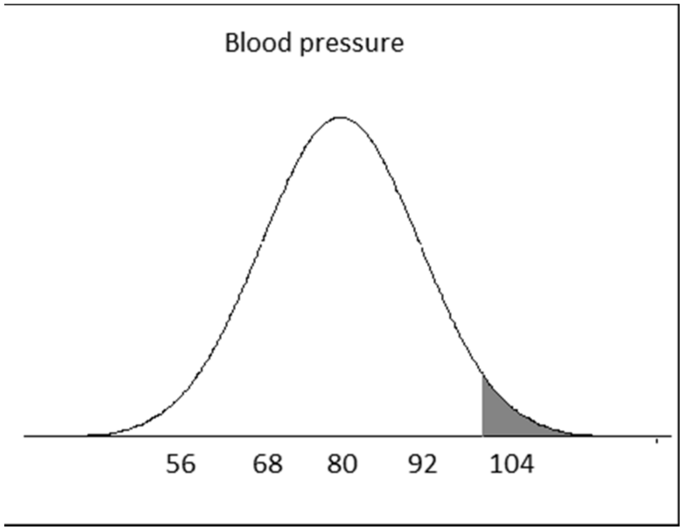
- To calculate the probability of other events we use the same idea
  - What is the probability that a male has blood pressure less than 56?
    - $P(X \leq 56) = P\left(Z \leq \frac{56-80}{12}\right) = P(Z \leq -2) = 0.023$



Calculating the tail probability like this is going to be very similar to our approach for calculating the p-value

Additional examples

- What is the probability that a male has blood pressure more than 100?
  - $P(X \geq 100) =$
  - $P\left(Z = \frac{X-80}{12} \geq \frac{100-80}{12}\right) =$
  - $P(Z \geq 1.67) =$
  - $1 - P(Z \leq 1.67) = 0.048$

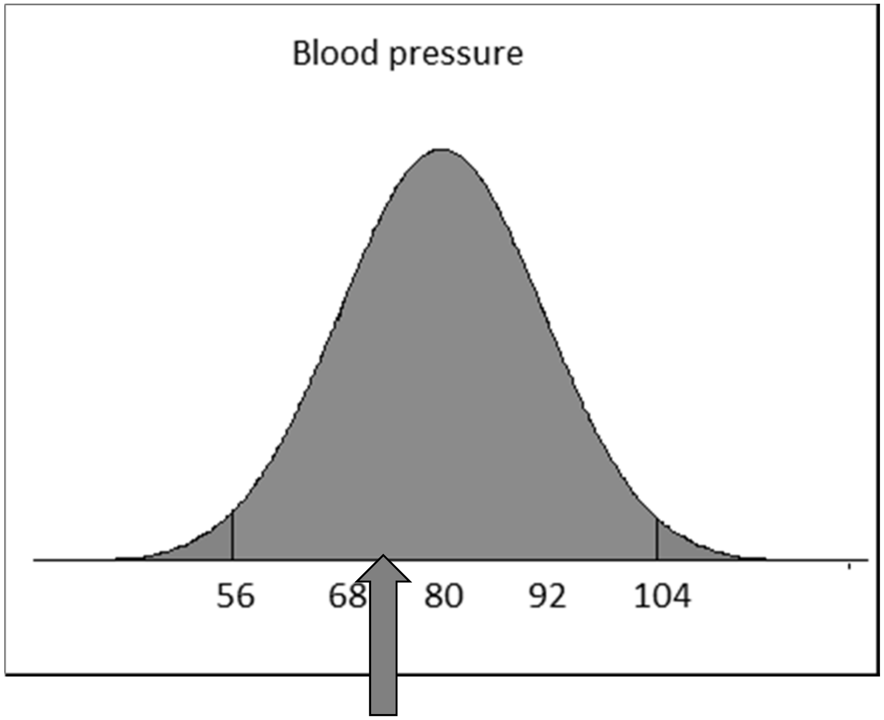


Bounds

- What if we wanted to know the normal range for blood pressure?
  - Let's define normal range to be the middle 95% of the data
  - We know $P(-1.96 \leq Z \leq 1.96) = 0.95$
  - $P\left(-1.96 \leq \frac{X-80}{12} \leq 1.96\right) =$
  - $P(80 - 1.96 * 12 \leq X \leq 80 + 1.96 * 12) =$
  - $P(56.5 \leq X \leq 103.5) = 0.95$

Bounds

- 95% of the data fall between (56.5, 103.5)



95% of the data fall
in the green area

There is a relationship between these bounds and the 95% confidence interval that we will see next week

Summary

- The probability of events from any normal distribution can be calculated using the relationship to a standard normal
- We can also use properties of the normal distribution to calculate upper and lower bounds

Distribution of the mean

Mean of multiple observations

- In the previous examples, we have looked at the distribution of a single observation
  - This is relevant for establishing normal ranges for clinical care
- In medical research, we are often interested in the distribution of the mean of a group of observations
  - What is the distribution of a the mean of a group of observations?

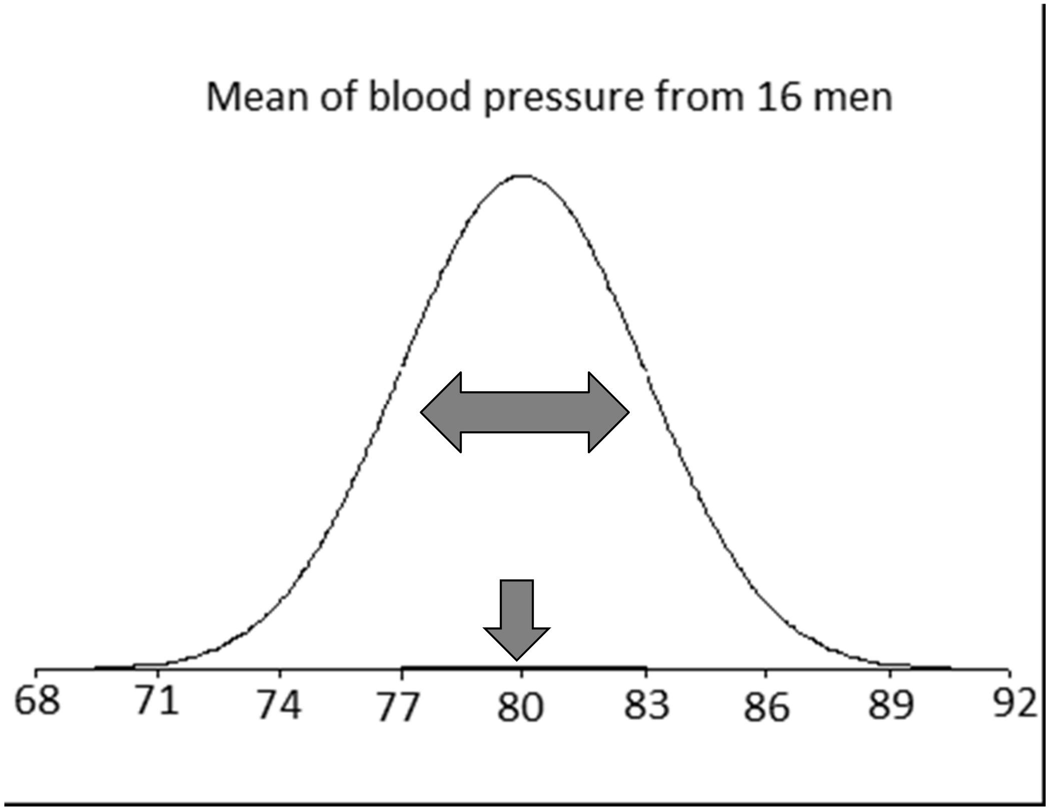
Example

- We have collected a set of 16 patients from MGH and calculated their mean blood pressure
  - Normal distribution
  - Population mean=80
  - Population standard deviation=12
- What is the distribution of the MEAN of 16 patients?
- http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Results

- Based on the simulation of data from a normal distribution, we see
  - The mean of the distribution of the sample means is equal to the population mean
  - The standard deviation of the distribution of the sample means is equal to the standard deviation of population divided by the square root of the sample size
  - The distribution of the sample means is normal

Picture



Distribution of sample means in normal

Mean is equal to the population mean=80

Standard deviation is equal to population standard deviation divided by square root of sample size=3

Standard error

- The standard deviation of the sample means $\left(\frac{\sigma}{\sqrt{n}}\right)$ is also called the standard error
- This tells you about the variability in the distribution of the sample means
  - Goes down as the sample size goes up regardless of the standard deviation
  - Some journals prefer SD rather than SE
  - Why?

Technical

- Mean of sample mean:

$$E\left(\overline{x}\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(x_i)=\frac{1}{n}\left(n\mu\right)=\mu$$

- Variance of sample mean:

$$Var\left(\overline{x}\right)=Var\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)=\frac{1}{n^2}Var\left(\sum_{i=1}^{n}x_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}Var\left(x_i\right)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n}$$

- Variance of sample mean decreases as n increases
- This holds regardless of the distribution of the data

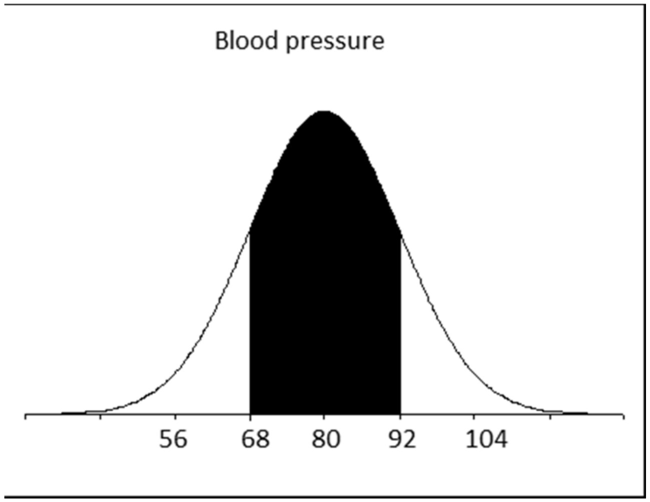
Example

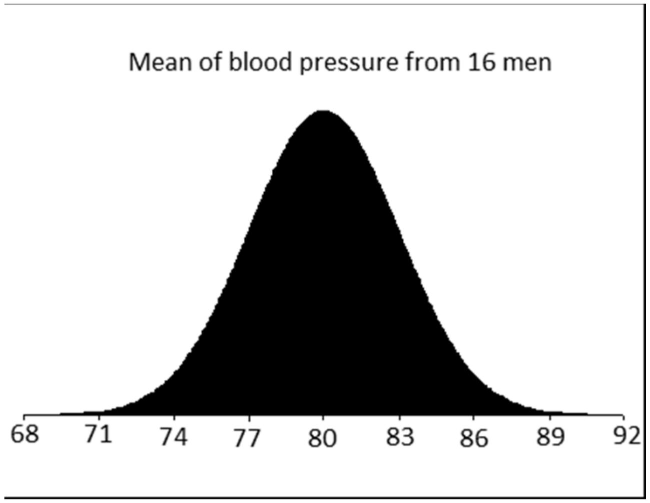
- Before, we calculated the probability that a man had a blood pressure between 68 and 92 using the following?
  - $P(68 \leq X \leq 92) = P\left(\frac{68-80}{12} \leq \frac{X-80}{12} \leq \frac{92-80}{12}\right) =$
  - $P\left(\frac{68-80}{12} \leq Z \leq \frac{92-80}{12}\right) =$
  - $P(-1 \leq Z \leq 1) = 0.68$
- What changes if we want to calculate the probability that the MEAN of 16 men is between 68 and 92?

Example

- Since we are dealing with the MEAN of 16 men rather than a single men, we must use the variance associated with the mean
  - $P(68 \leq \bar{X} \leq 92) = P\left(\frac{68-80}{12/\sqrt{16}} \leq \frac{\bar{X}-80}{12/\sqrt{16}} \leq \frac{92-80}{12/\sqrt{16}}\right)$
  - $P\left(\frac{68-80}{3} \leq Z \leq \frac{92-80}{3}\right) =$
  - $P(-4 \leq Z \leq 4) = 0.999$
- What happened?

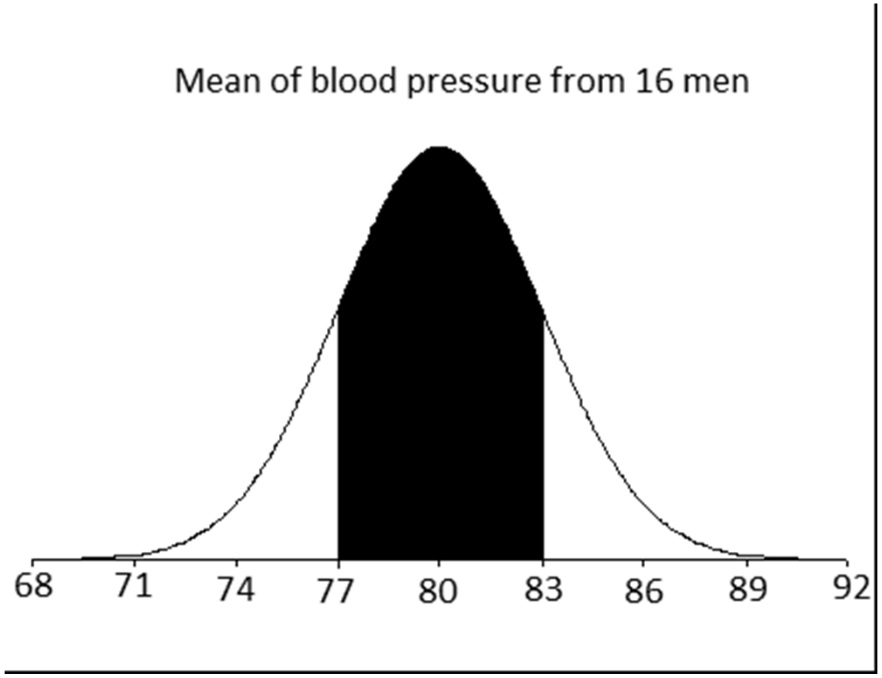
Picture





- Since the spread in the distribution of the mean is much less than the distribution of the individual observations, much more of the distribution is between 68 and 92

Picture



- What percentage of sample means would we expect to be between 77 and 83?

$$P(77 \le \bar{X} \le 83) = P\left(\frac{77-80}{3} \le \frac{\bar{X}-80}{3} \le \frac{83-80}{3}\right) =$$

$$P\left(\frac{77-80}{3} \le Z \le \frac{83-80}{3}\right) =$$

$$P(-1 \le Z \le 1) = 0.68$$

Amazing!!

- Using the same approach to calculate the probability of a single observation, we can calculate the probability of observing means equal to specific values
- Since the sample mean has a normal distribution, we can use the same approach as before
- This concept will be the basis of calculating p-values next class

Conclusions

- The mean of a set of observations from a normal distribution also has a normal distribution
  - Mean of this distribution equals the population distribution
  - Standard deviation of this distribution is reduced based on the sample size
- The approach to calculate probabilities is the same

Central limit theorem

Experiment

- Let’s try another experiment...
- http://onlinestatbook.com/stat_sim/sampling_dist/index.html
- What happens as we increased the sample size?
  - What changes?
- Why do we want a large sample size?
- What is the distribution of the sample mean?

Central limit theorem

- If you take a sample of size n and n is large, the distribution of the sample means $(\bar{x})$ is
  - NORMAL!!!
  - Mean=μ
  - Standard deviation= $\frac{\sigma}{\sqrt{n}}$
- So when we take a sample, we know the distribution of the sample mean
- We also know $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

Conclusions

- The central limit theorem tells us that the distribution of the sample mean is actually normal even if the underlying data are not normal if we have a large sample size
- Since we know the distribution, we can calculate probabilities of specific events
- This result is important for calculation of p-values and confidence intervals

What we learned

- In this lecture, we described
  - Different types of probabilities
  - Conditional probability
  - Binomial distribution
  - Probability of specific events based on the binomial and normal distributions