Inspire...Educate...Transform.

Inspire…Educate…Transform.

Stat Skills

Discrete Distributions, Normal Distribution, Sampling Distribution

Dr. Anand Jayaraman

Dec 24, 2016

Thanks to Dr.Sridhar Pappu for the material

REVIEW

Review

- Data
  - Type of Data
  - Central Tendencies, Spread, Data Presentation - Box Plot
- Probability
  - Probability Rules, Venn Diagram
  - Conditional Probability
  - Bayes Theorem
  - Confusion Matrix & Classification Errors
- Probability Distributions
  - Connection to Histogram
  - Expectation Values (mean) and Spread (Variance/Standard Deviation)

Probability

- A pair of dice are rolled. What is the probability that the sum of the dice is 7?
  - Total number of possibilities = 36
  - Sum 7 events= (1,6),(2,5),(3,4),(4,3),(5,2),(6,1)
  - Probability = 6/36 = 1/6

Election Math:Independent or Mutually Exclusive?

Event BP: BJP wins Punjab (Prob=50%)

Event AP: AAP wins Punjab (Prob=50%)

Event BU: BJP wins UP (Prob=75%)

Event CU: Congress wins UP (Prob=25%)

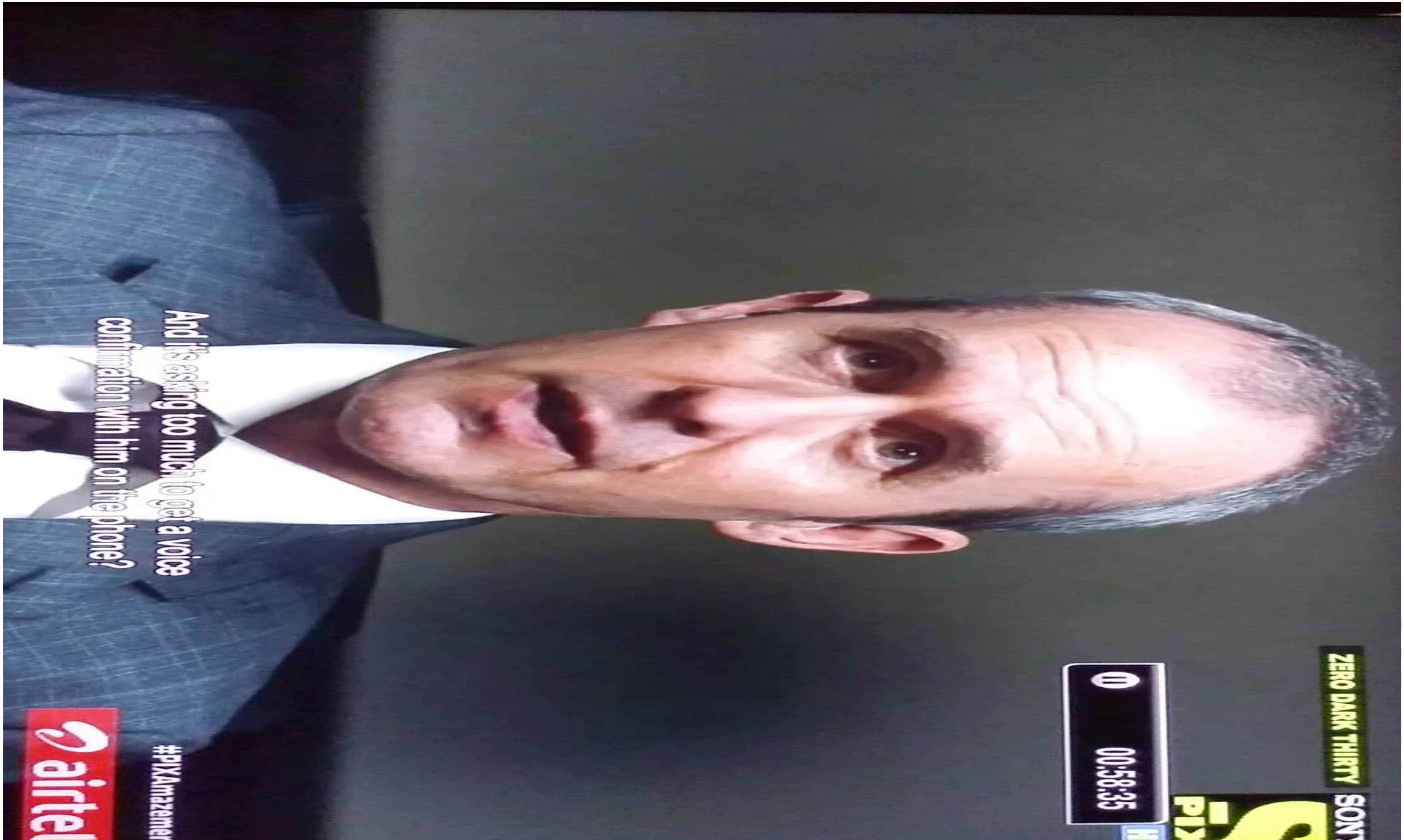
What kinds of events are the below scenarios?

Event BP and Event AP	*Mutually Exclusive*
Event BU and Event CU	*Mutually Exclusive*
Event AP and Event BU	*Independent*

Given the above probability, what is:

P(BP and AP)	*0*	P(BP or AP)	*¾ + ¼ = 1*
P(CU and BP)	*¼ * ½*	P(CU or BP)	*¼ + ½ - (¼* ½) = 0.625*

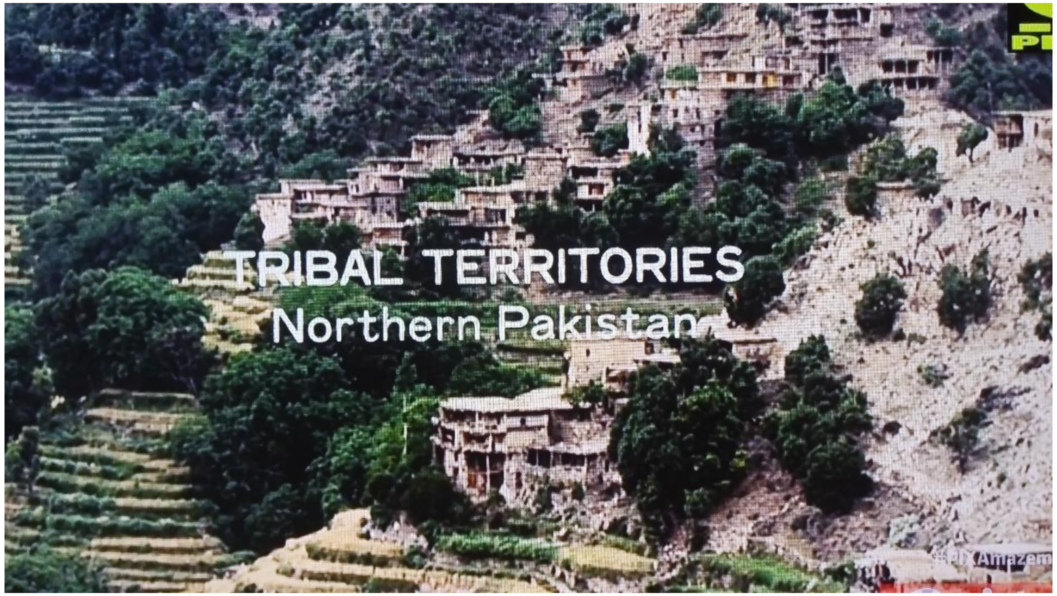
CSE 7315c


And it's asking too much to get a voice
confirmation with him on the phone?
ZERO DARK THIRTY
00:58:35

CSE 7315c

Bayes' Theorem : Finding Osama Bin Laden

Bin Laden was equally likely to be in a compound in a city or in the mountain caves of tribal areas.



Bayes’ Theorem : Finding Osama Bin Laden

The criteria to mark a compound as positive include high walls, inmates excepting a messenger not entering or leaving the compound, inmates carefully avoiding detection including by satellites, etc.

Given a compound meets all the criteria, what is the probability that Bin Laden is in that compound?

Bayes' Theorem : Finding Osama Bin Laden

P(Bin Laden is in Compound) = 0.5

P(Compound is + | Bin Laden is in Compound) = 0.9

P(Compound is + | Bin Laden is NOT in Compound) = 0.6

$$P(Bin\ Laden\ is\ in\ Compound|Compound\ is\ +) = \frac{P(Bin\ Laden\ is\ in\ Compound) * P(Compound\ is\ + |Bin\ Laden\ is\ in\ Compound)}{P(Bin\ Laden\ is\ in\ Compound) * P(Compound\ is\ + |Bin\ Laden\ is\ in\ Compound) + P(Bin\ Laden\ is\ NOT\ in\ Compound) * P(Compound\ is\ + |Bin\ Laden\ is\ NOT\ in\ Compound)}$$

$$P(Bin\ Laden\ is\ in\ Compound|\text{Compound is +}) = \frac{0.5 * 0.9}{0.5 * 0.9 + 0.5 * 0.6} = 60\%$$

Bayes' Theorem : Finding Osama Bin Laden

P(Bin Laden is in Compound) = 0.5

P(Compound is + | Bin Laden is in Compound) = 0.95

P(Compound is + | Bin Laden is NOT in Compound) = 0.05

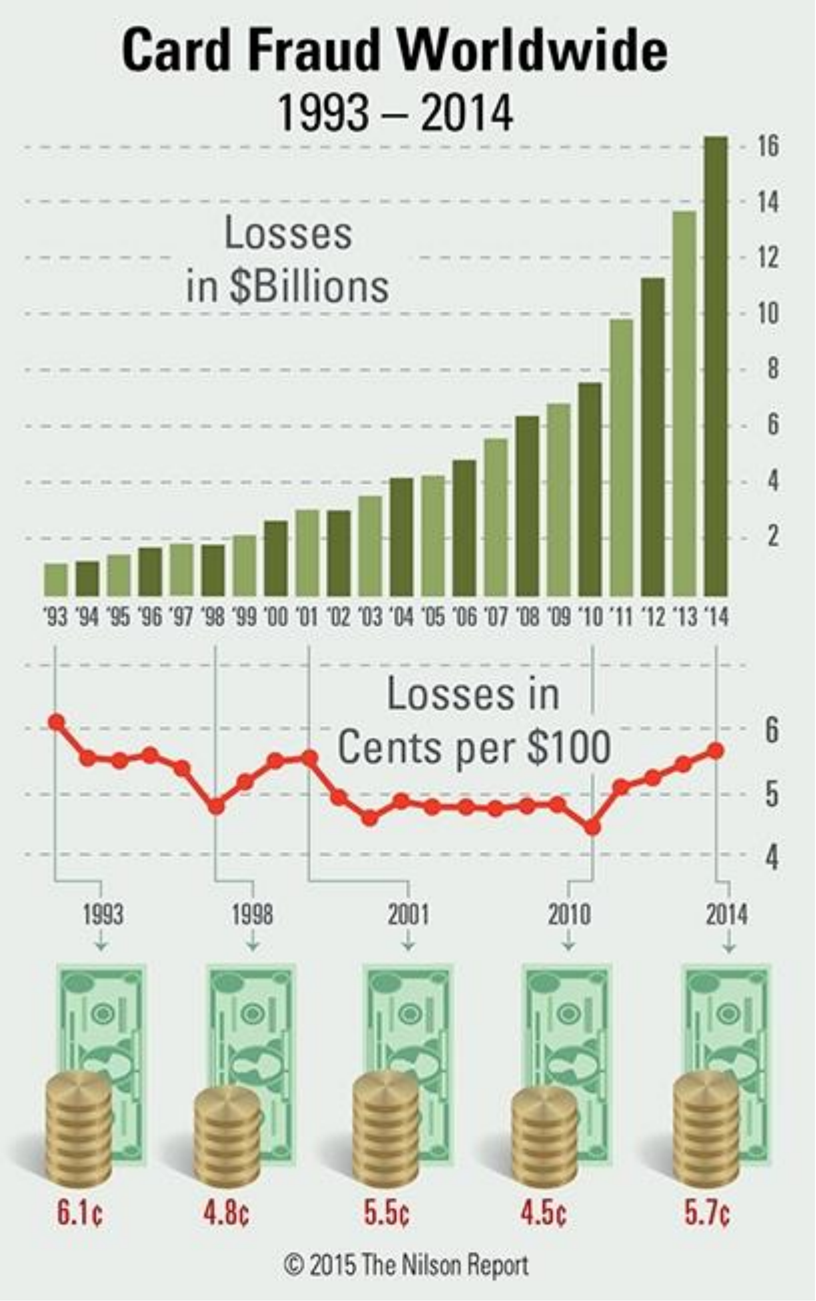
$$P(Bin\ Laden\ is\ in\ Compound|Compound\ is\ +)$$
$$= \frac{0.5 * 0.95}{0.5 * 0.95 + 0.5 * 0.05} = 95\%$$

Probabilities

Credit Card Fraud Detection

An e-commerce site has been offered a fraud detection system promising high accuracy.

It correctly catches 99% of fraudulent transactions and incorrectly flags 1% of honest transactions as fraudulent.



Source: http://www.nilsonreport.com/publication_chart_and_graphs_archive.php?1=1&year=2015
Last accessed: October 16, 2015

CSE 7315c

Probabilities

Credit Card Fraud Detection

1) What will be the implications to the retailer of mistaking honest transactions as fraudulent?

2) What will be the implications to the retailer of mistaking fraudulent transactions as honest?

3) What types of errors are the above two? Which one is more critical to the business?

True Positive	**True Negative**
False Positive	**False Negative**

1

2

Probabilities

Credit Card Fraud Detection

What conditional probabilities are given? What others are important for business?

Given:

P(Flagged fraud | Fraud) = 0.99

P(Flagged fraud | Honest) = 0.01

Other important conditional probabilities:

P(Fraud | Flagged fraud)

P(Honest | Not flagged fraud)

Probabilities

Credit Card Fraud Detection

Suppose 1% of the transactions at the e-commerce site are fraudulent. What are the chances that the transaction is honest given the system has incorrectly flagged it as fraudulent?

To calculate:

P(Honest | Flagged fraud) when P(Fraud) = 0.01

$$P(Honest|Flagged\ fraud) = \frac{P(Honest) * P(Flagged\ fraud|Honest)}{P(Honest) * P(Flagged\ fraud|Honest) + P(Fraud) * P(Flagged\ fraud|Fraud)}$$

$$= \frac{0.99 * 0.01}{0.99 * 0.01 + 0.01 * 0.99} = \frac{1}{2} = 0.50$$

Probabilities

Credit Card Fraud Detection

Suppose 5% of the transactions at the e-commerce site are fraudulent. What are the chances that the transaction is honest given the system has incorrectly flagged it as fraudulent?

To calculate:

P(Honest | Flagged fraud) when P(Fraud) = 0.05

$$P(Honest|Flagged\ fraud) = \frac{P(Honest) * P(Flagged\ fraud|Honest)}{P(Honest) * P(Flagged\ fraud|Honest) + P(Fraud) * P(Flagged\ fraud|Fraud)}$$

$$= \frac{0.95 * 0.01}{0.95 * 0.01 + 0.05 * 0.99} = \frac{0.0095}{0.059} = 0.16$$

Probabilities

Credit Card Fraud Detection

What is your evaluation of the system? Is it adequate for the e-commerce site's needs?

If fraud is rare (1% or less), too many honest transactions are labeled as fraud. This will be disastrous for the company.

At higher fraud rates (5% or more), the system's performance may be acceptable, depending on size of transactions and costs involved in contacting annoyed customers and taking measures to retain them.

Confusion Matrix

Buyer or Non-Buyer

A retail store's marketing team uses analytics to predict who is likely to buy a newly introduced high-end (read "expensive") product. Indicate which measure is more important for the business to track and explain why. Calculate other measures also.

Buyer or Not		**Actual**		**Total**
		Negative	Positive	
Predicted	Negative	725	158	883
	Positive	75	302	377
Total		800	460	1260

Confusion Matrix

Buyer or Not		Actual Negative	Actual Positive	Total
Predicted	Negative	725	158	883
	Positive	75	302	377
Total		800	460	1260

Buyer or Non-Buyer

State/Calculate:

TP = 302 TN = 725 FP = 75 FN = 158

Should the business be more worried about FP or FN or equally worried about both of them? Why?

FN. If the model predicts that the person will not buy, the product will not be marketed to him/her, and the business will lose…er, business. FP is not such a big worry since only the cost of a phone call, SMS or sending a catalog will be lost.

Confusion Matrix

Buyer or Non-Buyer

Buyer or Not		Predicted		Total
		Positive	Negative	
Actual	Positive	302	158	460
	Negative	75	725	800
Total		377	883	1260

What is more important: Recall, Precision or Accuracy?

		Predicted		
		Positive	Negative	
Actual	Positive	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	Negative	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Confusion Matrix

Buyer or Non-Buyer

Buyer or Not		**Predicted** Positive	**Predicted** Negative	**Total**
Actual	Positive	302	158	460
Actual	Negative	75	725	800
Total		377	883	1260

What is more important: Recall, Precision or Accuracy?

$$Recall\ (or\ Sensitivity) = \frac{TP}{TP + FN} = \frac{302}{460} = 65.6\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{302}{377} = 80.1\%$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} = \frac{1027}{1260} = 81.5\%$$

$$Specificity = \frac{TN}{TN + FP} = \frac{725}{800} = 90.6\%$$

$$F_1 = \frac{2 * Recall * Precision}{Recall + Precision} = \frac{2 * 0.656 * 0.801}{0.656 + 0.801} = 72.1\%$$

Sensitivity - Specificity

TEST NAME	TECHNOLOGY	VALUE	UNITS
ANTI CCP (ACCP)	E.L.I.S.A	0.48	OD Ratio

Reference Range :

Negative : < 0.80
Equivocal: 0.80 - 1.20
Positive : > 1.20

Clinical Significance :

Anti-Cyclic-Citrullinated-Peptide (Anti-CCP) Antibodies hold promise for early and more accurate detection of Rheumatoid Arthritis before the disease proceeds into an irreversible damage.

Analytical Specifications :

Anti-Cyclic-Citrullinated-Peptide (Anti-CCP) antibodies are detected using a solid phase enzyme immuno assay having an analytical sensitivity of 1.0 U/ml. No cross reactivity to other auto antigen is found. Sensitivity of the method is 68% and specificity is 92%.

Method : SOLID PHASE CAPTURE ENZYME IMMUNOASSAY

TEST NAME	TECHNOLOGY	VALUE	UNITS
ANTI NUCLEAR ANTIBODIES (ANA)	E.L.I.S.A	0.29	OD Ratio

Reference Range :

Negative < 0.80
Equivocal 0.8 – 1.20
Positive > 1.20

Sensitivity - Specificity

Rheumatoid arthritis is prevalent among 0.75% of adult Indian population. Assuming a test with a certain sensitivity and specificity gives the following results, what are the values for these metrics?

Rheumatoid Arthritis Diagnostic Test		Predicted		Total
		Positive	Negative	
Actual	Positive	20	10	30
	Negative	318	3652	3970
Total		338	3662	4000

$$Recall\ (Sensitivity) = \frac{20}{30} = 0.67$$

$$Specificity = \frac{3652}{3970} = 0.92$$

Distributions - Properties

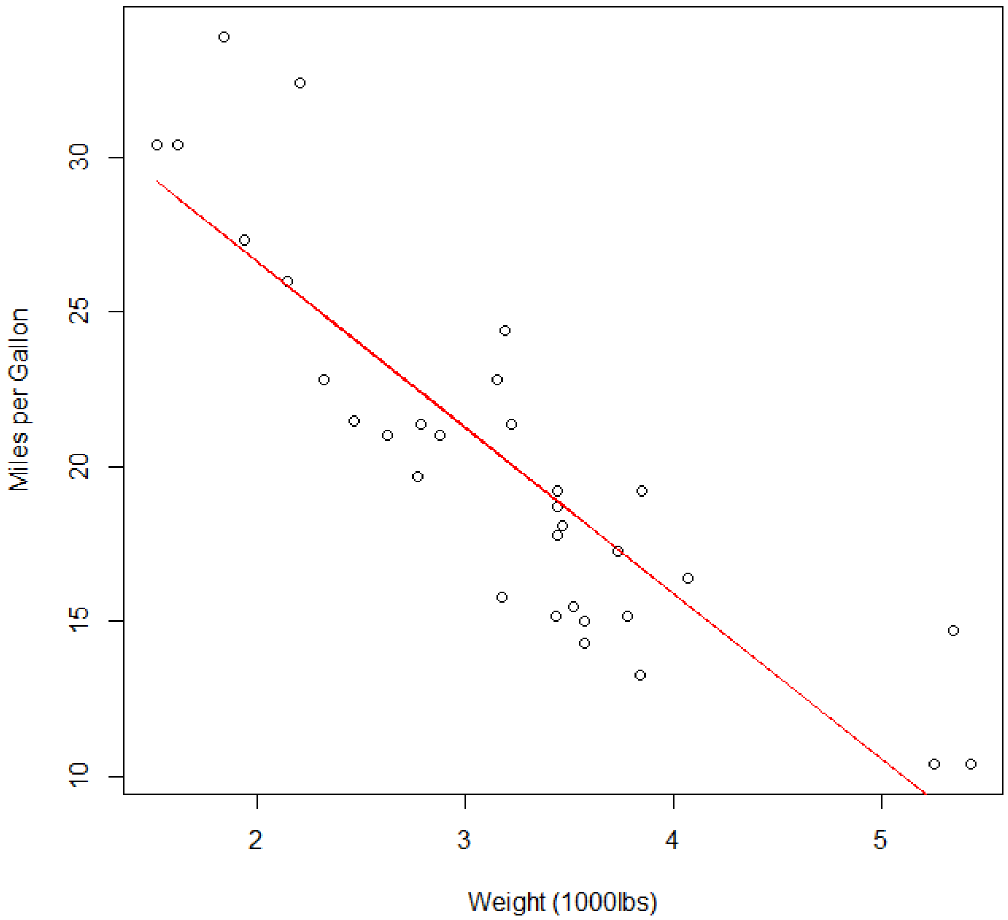
Var(-X) = ?

Options:

- Var(X) ✓
- -Var(X)
- Variance of negative numbers cannot be calculated
- None of the above (Explain)

Present Day

Weight of Cars vs Fuel Efficiency



Source: mtcars database

Sample Software Output

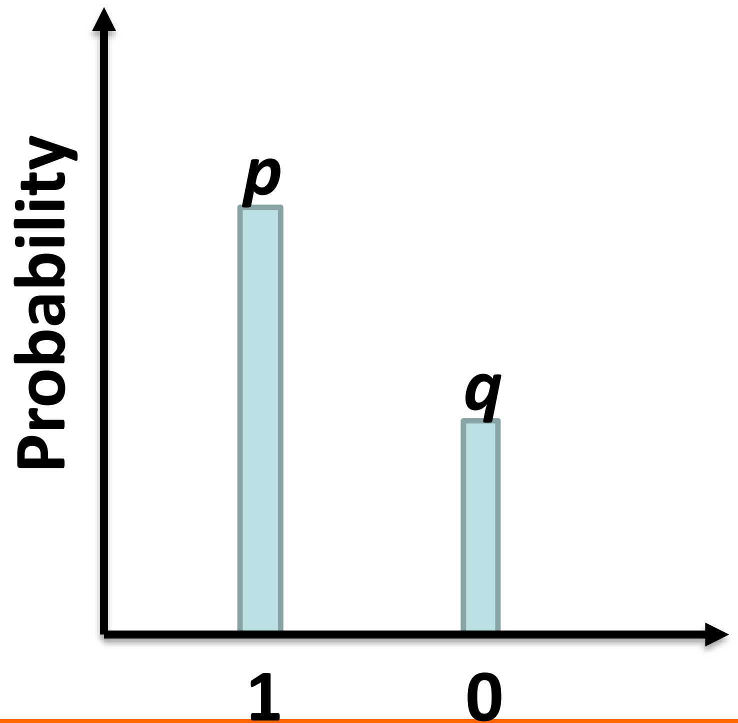
```
Call:
glm(formula = mpg ~ wt, data = mtcars)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-4.5432  -2.3647  -0.1252   1.4096   6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 9.277398)

    Null deviance: 1126.05  on 31  degrees of freedom
Residual deviance:  278.32  on 30  degrees of freedom
AIC: 166.03

Number of Fisher Scoring iterations: 2
```

CSE 7315c

SOME COMMON DISTRIBUTIONS

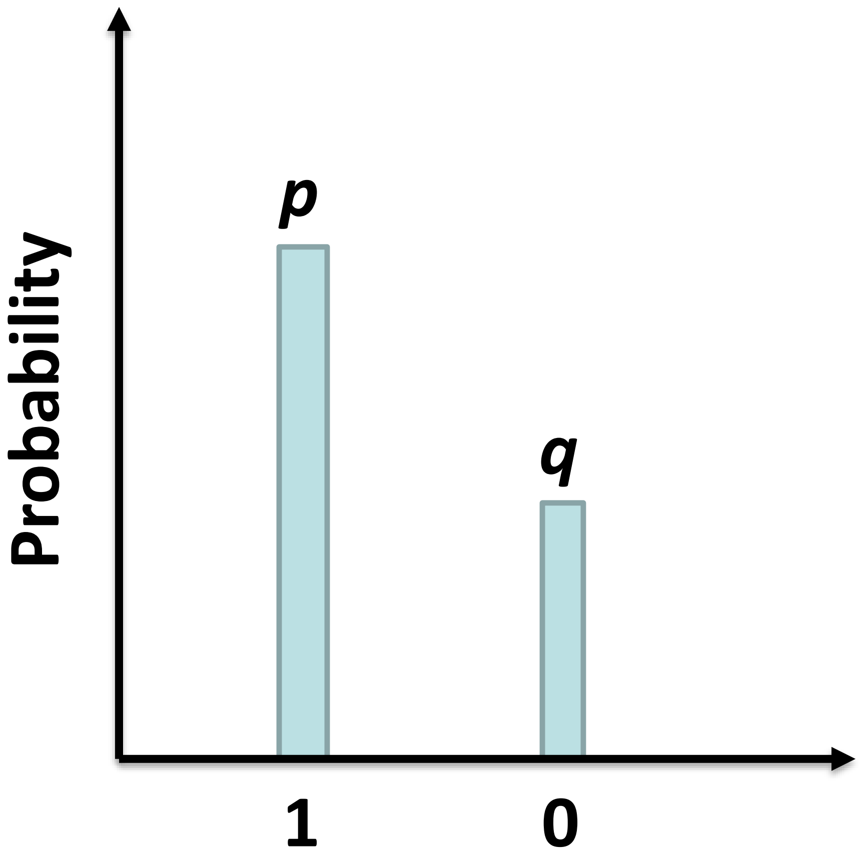
Bernoulli

There are two possibilities (pass an exam or fail) with probability *p* of success and *q=1-p* of failure



Expectation: *p*
Variance : *pq*

Bernoulli



$$Expectation, E(X) = \sum x_i P(x_i)$$

$$= 1 * p + 0 * q = p$$

$$Variance, Var = \sum (x_i - \mu)^2 P(x_i)$$

$$= (1-p)^2 * p + (0-p)^2 * (1-p)$$

$$= p(1-p)$$

$=pq$

Geometric Distribution

Number of independent and identical Bernoulli trials needed to get ONE success, e.g., number of attempts before I pass the exam.

Geometric Distribution

- You run a series of independent trials.
- There can be either a success or a failure for each trial, and the probability of success is the same for each trial.
- The main thing you are interested in is how many trials are needed in order to get the first successful outcome.

Geometric Distribution

PMF*, $P(X = r) = q^{r-1}p$ — *(r-1)* failures followed by ONE success.

$P(X > r) = q^r$ — Probability you will need more than *r* trials to get the first success.

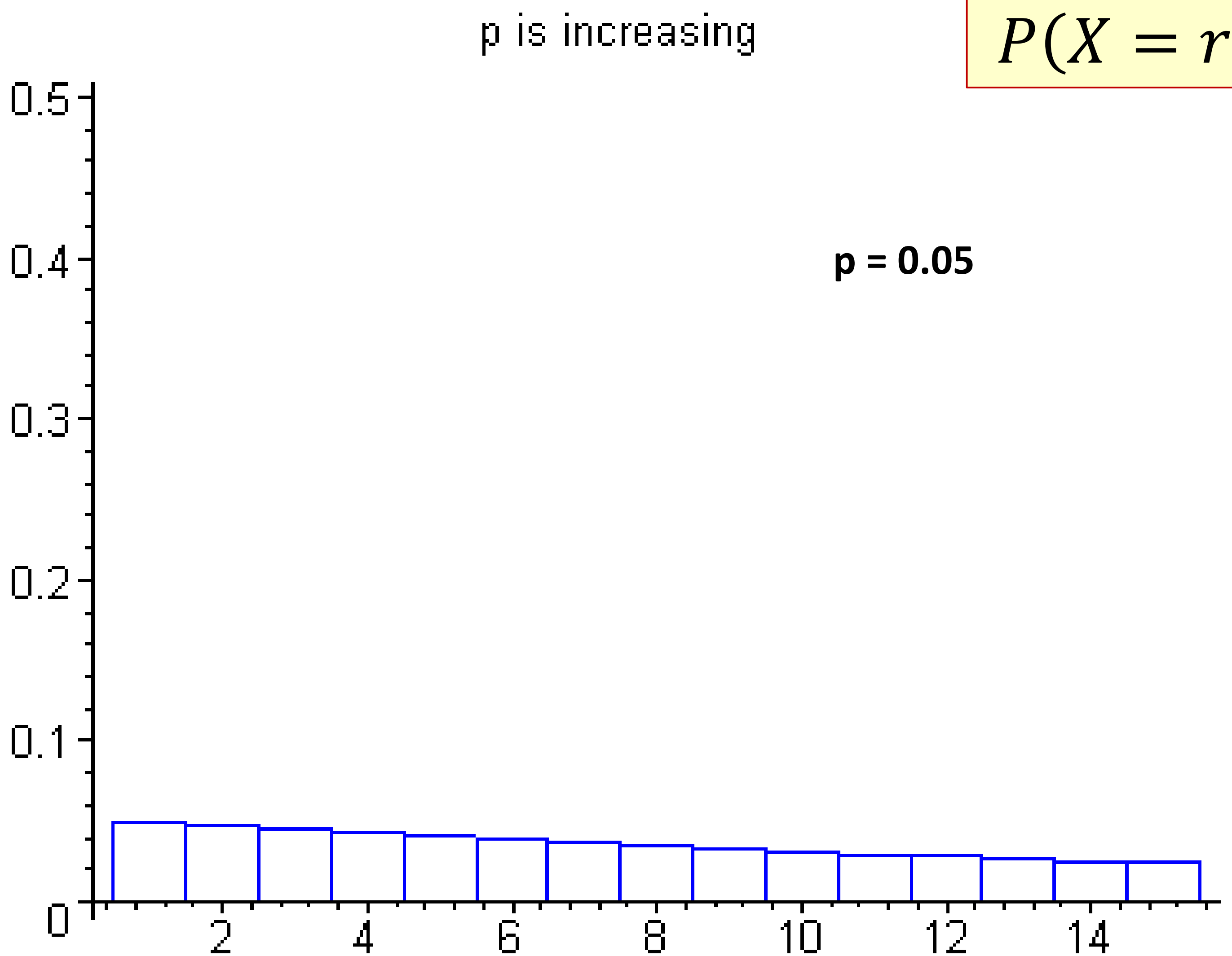
CDF**, $P(X \leq r) = 1 - q^r$ — Probability you will need *r* trials or less to get your first success.

$$E(X) = \frac{1}{p} \qquad Var(X) = \frac{q}{p^2}$$

* Probability Mass Function   ** Cumulative Distribution Function

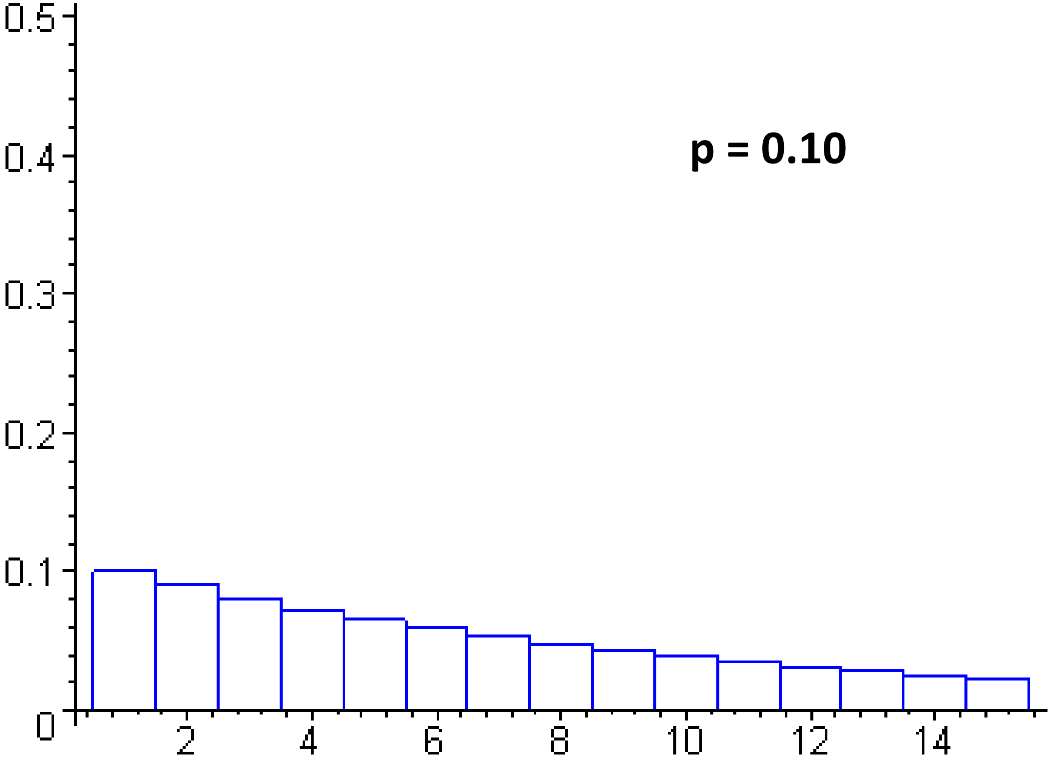
X~Geo(p)

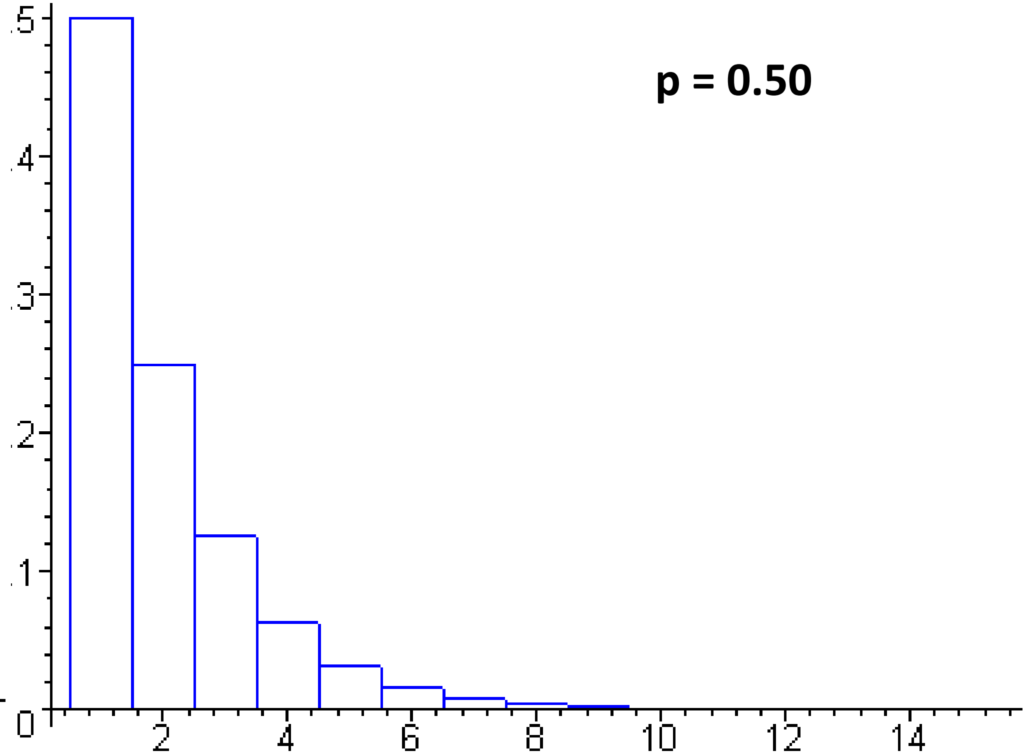
$$P(X = r) = q^{r-1}p$$



Ref: http://personal.kenyon.edu/hartlaub/MellonProject/Geometric2.html
Last accessed: June 12, 2015

X~Geo(p)





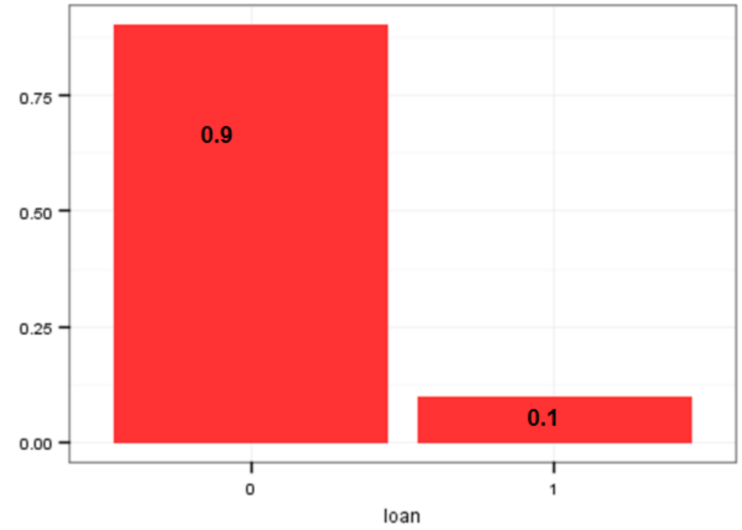
$$P(X = r) = q^{r-1}p$$

Ref: http://personal.kenyon.edu/hartlaub/MellonProject/Geometric2.html
Last accessed: June 12, 2015

Binomial Distribution

If I randomly pick 10 people, what is the probability that I will get exactly

- 0 loan takers = 0.9^{10}
- 1 loan taker = $10 * 0.1^{1} * 0.9^{9}$
- 2 loan takers = $C_2^{10} * 0.1^{2} * 0.9^{8}$



Binomial Distribution

If there are two possibilities with probability *p* for success and *q* for failure, and if we perform *n* trials, the probability that we see *r* successes is

PMF, $\mathrm{P}(\mathrm{X} = \mathrm{r}) = C_r^n p^r q^{n-r}$

CDF, $\mathrm{P}(\mathrm{X} \leq \mathrm{r}) = \sum_{i=0}^{r} C_i^n p^i q^{n-i}$

Binomial Distribution

$E(X) = np$ $Var(X) = npq$

When to use?

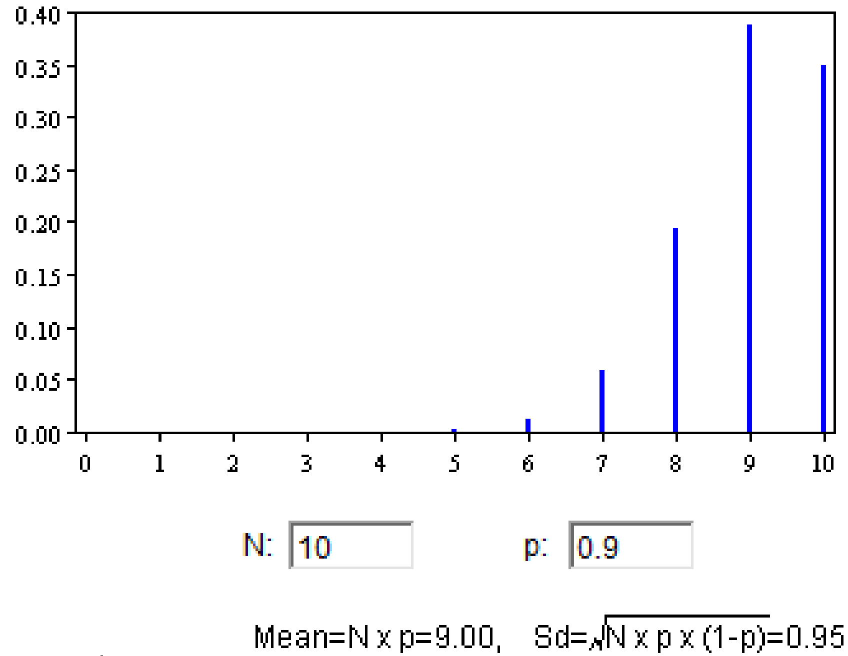
- You run a series of independent trials.
- There can be either a success or a failure for each trial, and the probability of success is the same for each trial.
- There are a finite number of trials, and you are interested in the number of successes or failures.

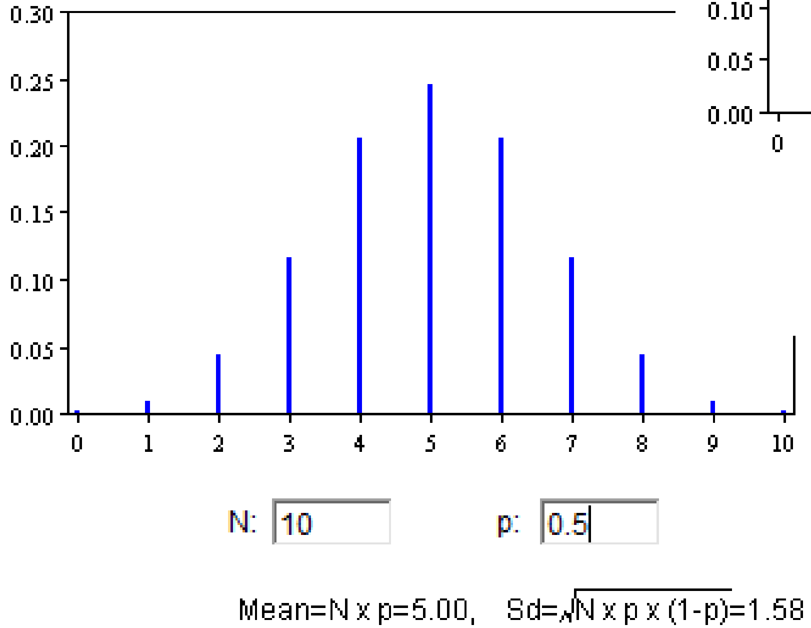
Binomial Distribution

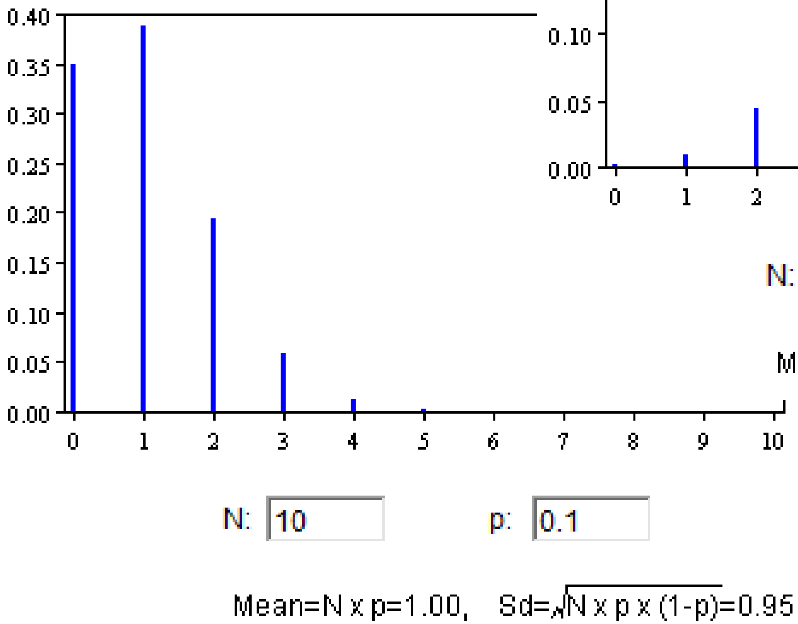
- What is the probability of getting exactly 5 heads in 10 tosses of a fair coin?
- p=0.5 , q=0.5 , r=5, n=10
- $\mathrm{P}(\mathrm{X} = \mathrm{r}) = C_5^{10} p^5 q^5$ = 252*$(0.5)^{10}$

  = 0.24

X~B(n,p)







$$P(X = r) = C_r^n p^r q^{n-r}$$

Ref: http://onlinestatbook.com/2/probability/binomial_demonstration.html
Last accessed: June 12, 2015

Poisson Distribution

- A distribution that expresses the probability of *r* number of events occurring when the average rate of occurrence (λ) is known.

CSE 7315c

Poisson Distribution

Probability of getting 15 customers requesting for loans in a given day given on average we see 10 customers

$\lambda = 10 \; and \; r = 15$

PMF, $\mathrm{P(X = r)} = \frac{e^{-\lambda}\lambda^{r}}{r!}$

CDF, $\mathrm{P(X \leq r)} = e^{-\lambda} \sum_{i=0}^{r} \frac{\lambda^{i}}{i!}$

Poisson Distribution

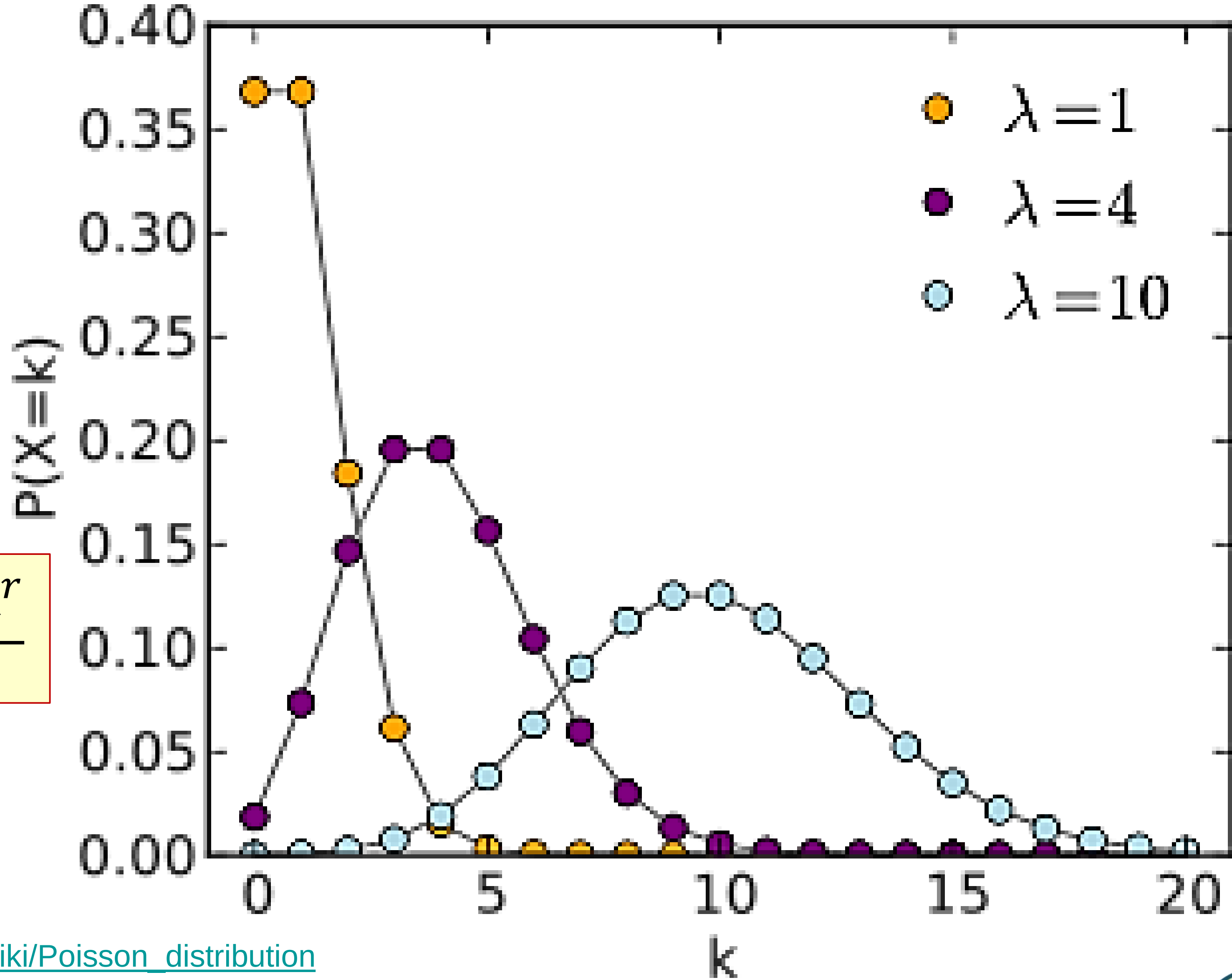
$E(X) = \lambda$ Can be equated to *np* of Binomial if *n* is large (>50) and *p* is small (<0.1)

$Var(X) = \lambda$ Can be equated to *npq* of Binomial in the above situation.

When to use?

- Individual events occur at random and independently in a given interval (time or space).
- You know the mean number of occurrences, λ, in the interval or the rate of occurrences, and it is finite.

X~Po(λ)



$$P(X = r) = \frac{e^{-\lambda}\lambda^{r}}{r!}$$

Ref: https://en.wikipedia.org/wiki/Poisson_distribution
Last accessed: February 12, 2016

Poisson Distribution

The probability that no customer will visit the store in one day

$$P(X=0) = \frac{e^{-\lambda}\lambda^{0}}{0!} = e^{-\lambda}$$

Probability that she will not have a customer for *n* days

$$e^{-n\lambda}$$

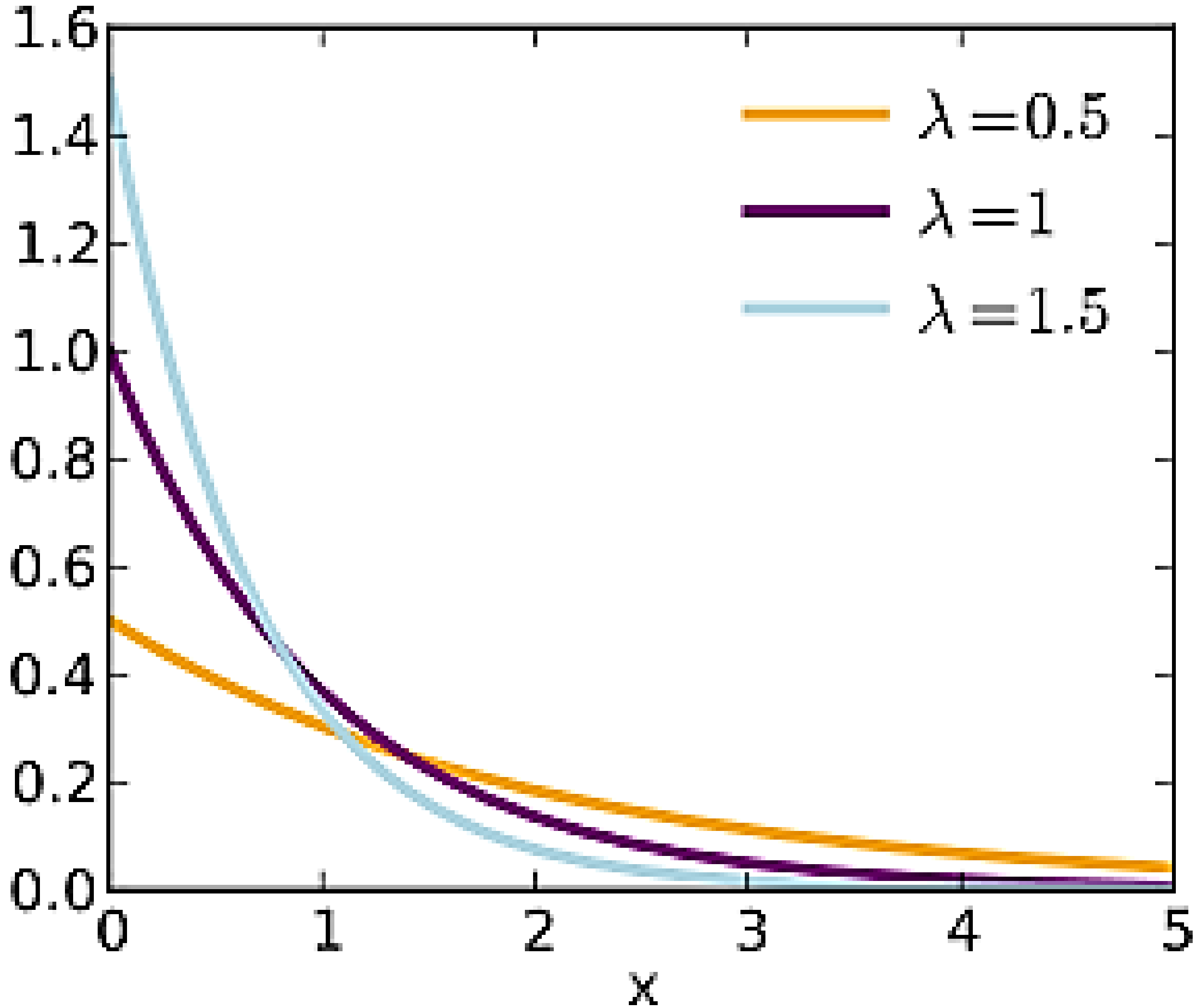
Exponential Distribution

Probability that a customer will visit in *n* days: $1 - e^{-n\lambda}$

$$CDF = 1 - e^{-n\lambda}, n \geq 0$$
$$PDF = \lambda e^{-n\lambda}, n \geq 0$$

X~Exp(λ)

$$PDF = \lambda e^{-n\lambda}, n \geq 0$$



Ref: http://en.wikipedia.org/wiki/Exponential_distribution
Last accessed: June 12, 2015

Exponential Distribution

- Poisson process
- Continuous analog of Geometric distribution

$$E(X) = \frac{1}{\lambda} \qquad Var(X) = \frac{1}{\lambda^2}$$

Distributions

Geometric: For estimating number of attempts before first success

Binomial: For estimating number of successes in *n* attempts

Poisson: For estimating *n* number of events in a given time period when on average we see *m* events

Exponential: Time between events

Probability Distributions

Here are a few scenarios. Identify the distribution and calculate expectation, variance and the required probabilities.

Q1. In an Archery competition a man has 0.3 probability of hitting bulls-eye. If he has 10 tries, what is the probability he will hit bulls-eye less than 3 times?

Q2. On average, 1 bus stops at a certain point every 15 minutes. What is the probability that no buses will turn up in a single 15 minute interval?

Q3. 20% of cereal packets contain a free toy. What is the probability you will need to open fewer than 4 cereal packets before finding your first toy?

Probability Distributions

Solutions

In an Archery competition a man has 0.3 probability of hitting the target. If he has 10 tries, what is the probability he will hit bulls-eye less than 3 times?

X ~ B(10,0.3); n=10, p=0.3, q=1-0.3=0.7, r=0, 1, 2 (< 3)

E(X) = np = 3

Var(X) = npq = 2.1

$$P(X = r) = {}^{n}C_{r} p^{r} q^{n-r}$$

P(X=0) = 0.028; P(X=1) = 0.121; P(X=2) = 0.233

∴ P(X<3) = 0.028 + 0.121 + 0.233 = 0.382

Probability Distributions

Solutions

On average, 1 bus stops at a certain point every 15 minutes. What is the probability that no buses will turn up in a single 15 minute interval?

X ~ Po(1); λ=1, r=0

E(X) = λ = 1

Var(X) = λ = 1

$$P(X = r) = \frac{e^{-\lambda}\lambda^{r}}{r!}$$

P(X=0) = 0.368

Probability Distributions

Solutions

20% of cereal packets contain a free toy. What is the probability you will need to open fewer than 4 cereal packets before finding your first toy?

X ~ Geo(0.2); p=0.2, q=1-0.2=0.8, r<4 or ≤3

E(X) = $\frac{1}{p}$ = 5

Var(X) = $\frac{q}{p^2}$ = 20

$$P(X \leq r) = 1 - q^r$$

P(X ≤ 3) = 0.488

Poisson Distribution Formula Differences?

$$P(X=r)=\frac{e^{-\lambda}\lambda^r}{r!} \; or \; \frac{e^{-\lambda t}(\lambda t)^r}{r!} \; ?$$

Suppose births in a hospital occur randomly at an average rate of 1.8 births per hour. What is the probability of 5 births in a given 2 hour interval?

What is λ?

$$P(X=5)=\frac{e^{-3.6}3.6^5}{5!} \; or \; \frac{e^{-1.8*2}(1.8*2)^5}{5!} \; ?$$

If you use 1.8, use t=2 in the second formula. Alternatively, you could say that since the average is 1.8 per hour, it is 3.6 per 2 hours (the interval of interest).

Poisson Distribution Formula Differences?

$$P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!} \; or \; \frac{e^{-\lambda t}(\lambda t)^r}{r!} \; ?$$

Now suppose head injury patients (due to not wearing helmets) arrive in Hospital A randomly at an average rate of 0.25 patients per hour, and in Hospital B randomly at an average rate of 0.75 per hour. What is the probability of more than 3 such patients arriving in a given 2 hour interval in both hospitals together?

What is the probability distribution?

$$X \sim Po(\lambda_1) \; and \; Y \sim Po(\lambda_2)$$
$$X + Y \sim Po(\lambda_1 + \lambda_2)$$

What are λ_1 and λ_2 if we use first formula?

$$\lambda_1 = 0.5 \text{ and } \lambda_2 = 1.5$$
$$\lambda_1 + \lambda_2 = 2$$

$$P(X + Y > 3) = P(X + Y = 4) + P(X + Y = 5) + P(X + Y = 6) + \cdots$$
$$= 1 - P(X + Y \leq 3) = 1 - (P(X + Y = 0) + P(X + Y = 1) + P(X + Y = 2) + P(X + Y = 3))$$

See http://math.stackexchange.com/questions/221078/poisson-distribution-of-sum-of-two-random-independent-variables-x-y

Poisson or Exponential?

Given a Poisson process:

- The *number* of events in a given time period — Poisson
- The *time* until the first event
- The *time* from now until the next occurrence of the event
- The *time interval* between two successive events

Exponential

Poisson or Exponential?

The tech support centre of a computer retailer receives 5 calls per hour on an average. What is the probability that the centre will receive 8 calls in the next hour? What is the probability that more than 30 minutes will elapse between calls?

$$P(X = 8) = \frac{e^{-5}5^{8}}{8!} = 0.065$$

$$P(Time\ between\ calls > 0.5) = \int_{0.5}^{\infty} \lambda e^{-\lambda T} dT$$

$$= -e^{-\lambda T}\Big]_{0.5}^{\infty} = e^{-5*0.5} = 0.082$$

Distributions

Babyboom Data - Excel

Forty-four babies -- a new record -- were born in one 24-hour period at the Mater Mothers' Hospital in Brisbane, Queensland, Australia, on December 18, 1997. For each of the 44 babies, *The Sunday Mail* recorded the time of birth, the sex of the child, and the birth weight in grams.

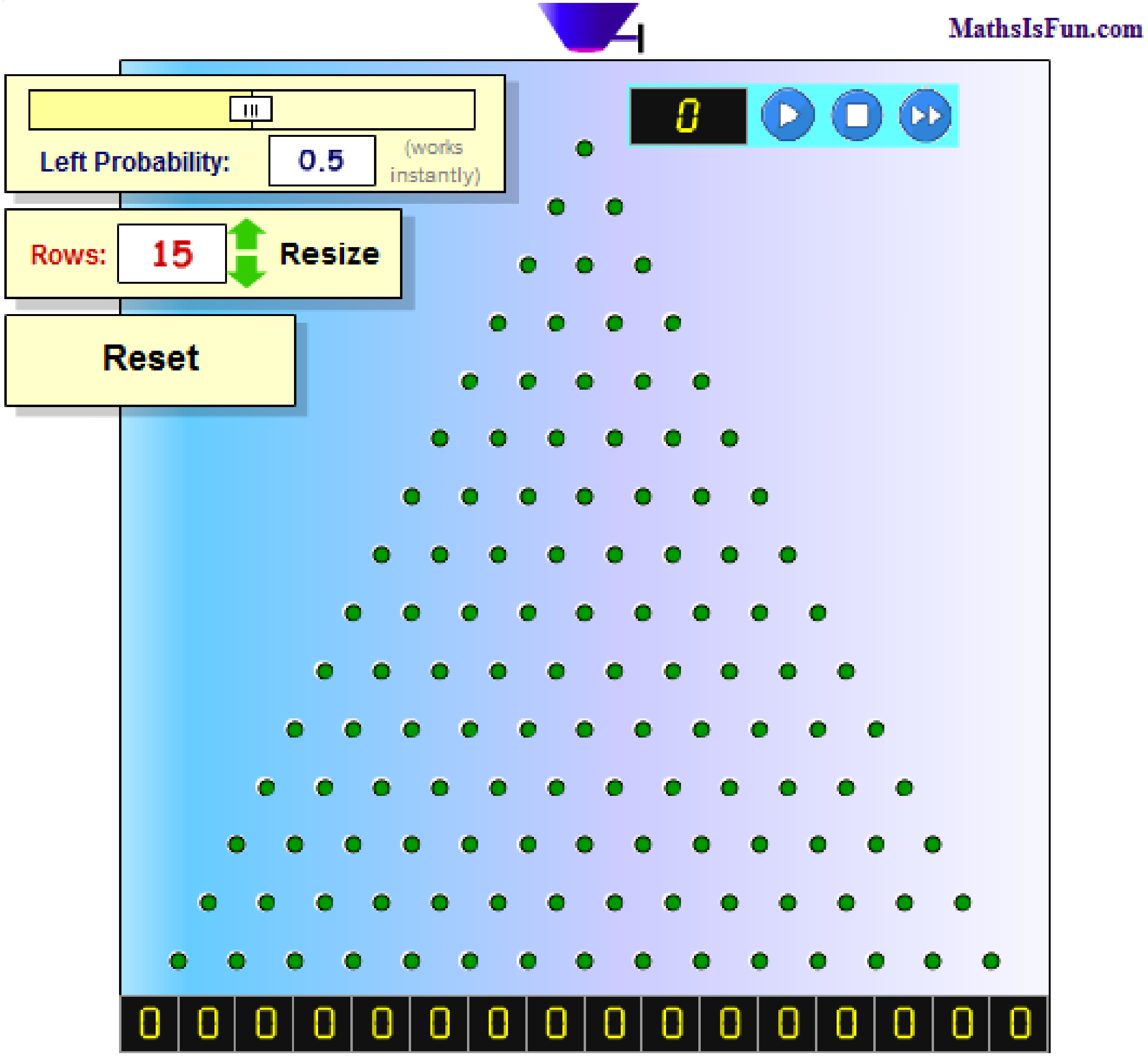
Distributions

Determine the distributions for the following scenarios for this dataset:

1. Probability of observing at least 26 boys in 44 births assuming equal probability of a boy or a girl being born.
2. Probability that 3 births occur before the birth of a girl.
3. Probability of 4 births per hour given 44/24 = 1.83 births per hour on average.
4. Probability that more than 60 minutes will elapse between births.

1. Binomial; 2. Geometric; 3. Poisson; 4. Exponential

Quincunx Demo

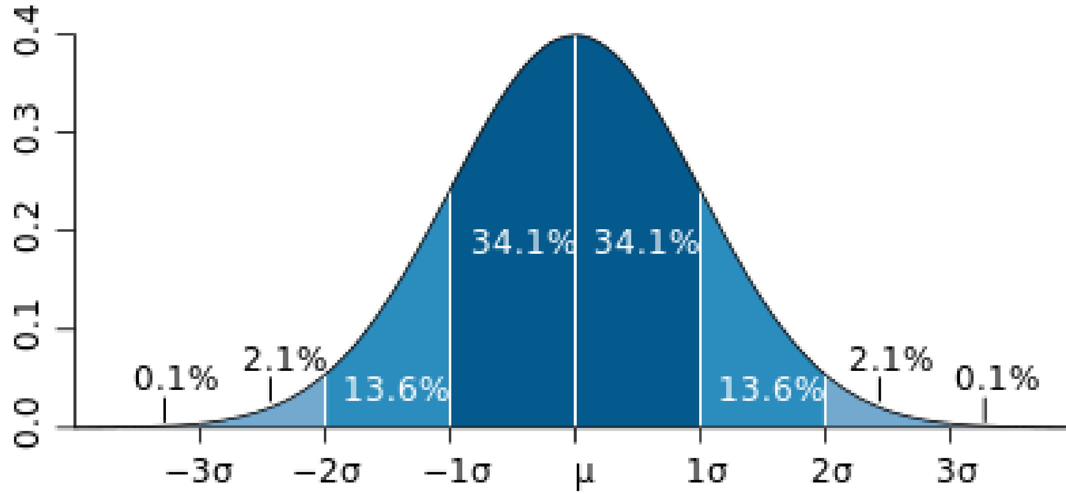


Source: http://www.mathsisfun.com/data/quincunx.html

GAUSSIAN DISTRIBUTION

Normal (Gaussian) Distribution

- Mean = Median = Mode
- Zero Skew and Kurtosis=3
- X ~ N(μ, σ^2)
- 68-95-99.7 empirical rule

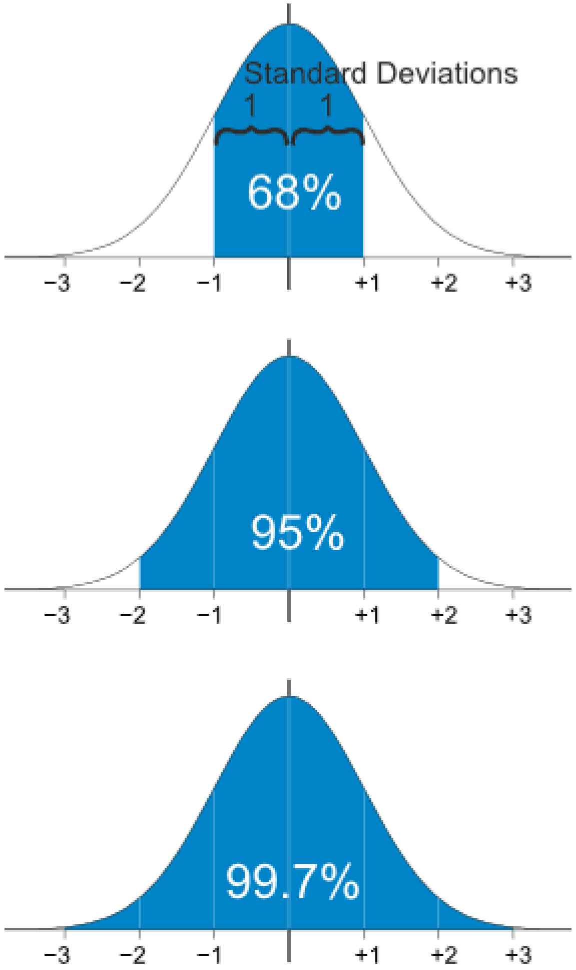


- Shaded area gives the probability that X is between the corresponding values

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Measures of Spread (Dispersion)

You know the 68-95-99.7 rule.



A company produces a lightweight valve that is specified to weigh 1500g, but there are imperfections in the process. While the mean weight is 1500g, the standard deviation is 300g.

Q1. What is the range of weights within which 95% of the valves will fall?

Q2. Approximately 16% of the weights will be more than what value?

Q3. Approximately 0.15% of the weights will be less than what value?

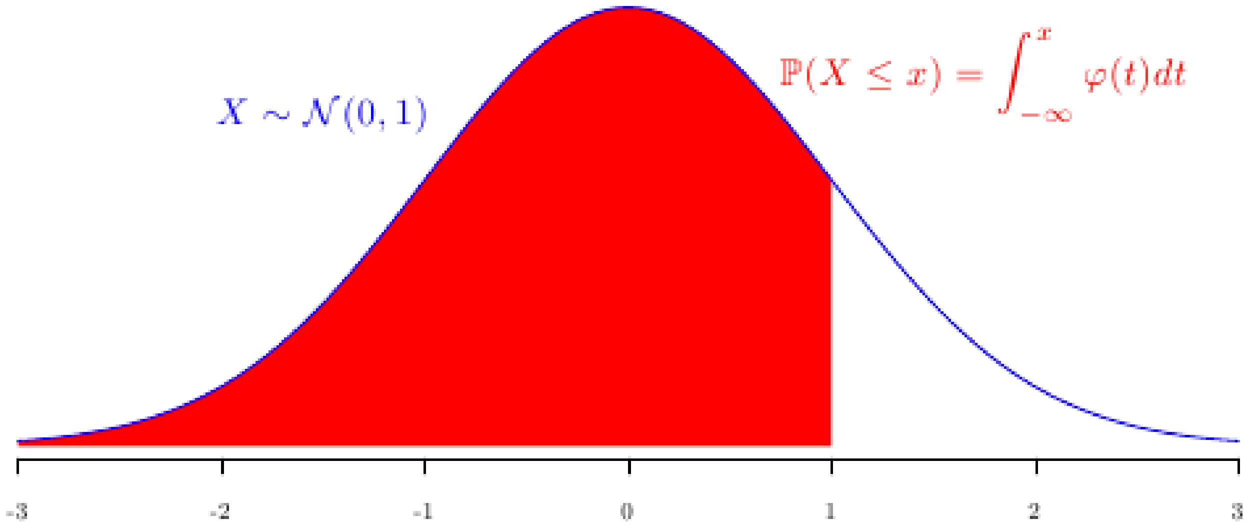
Image source: http://www.mathsisfun.com/data/standard-normal-distribution.html
Last accessed: May 03, 2014

Standardize a Normal Distribution

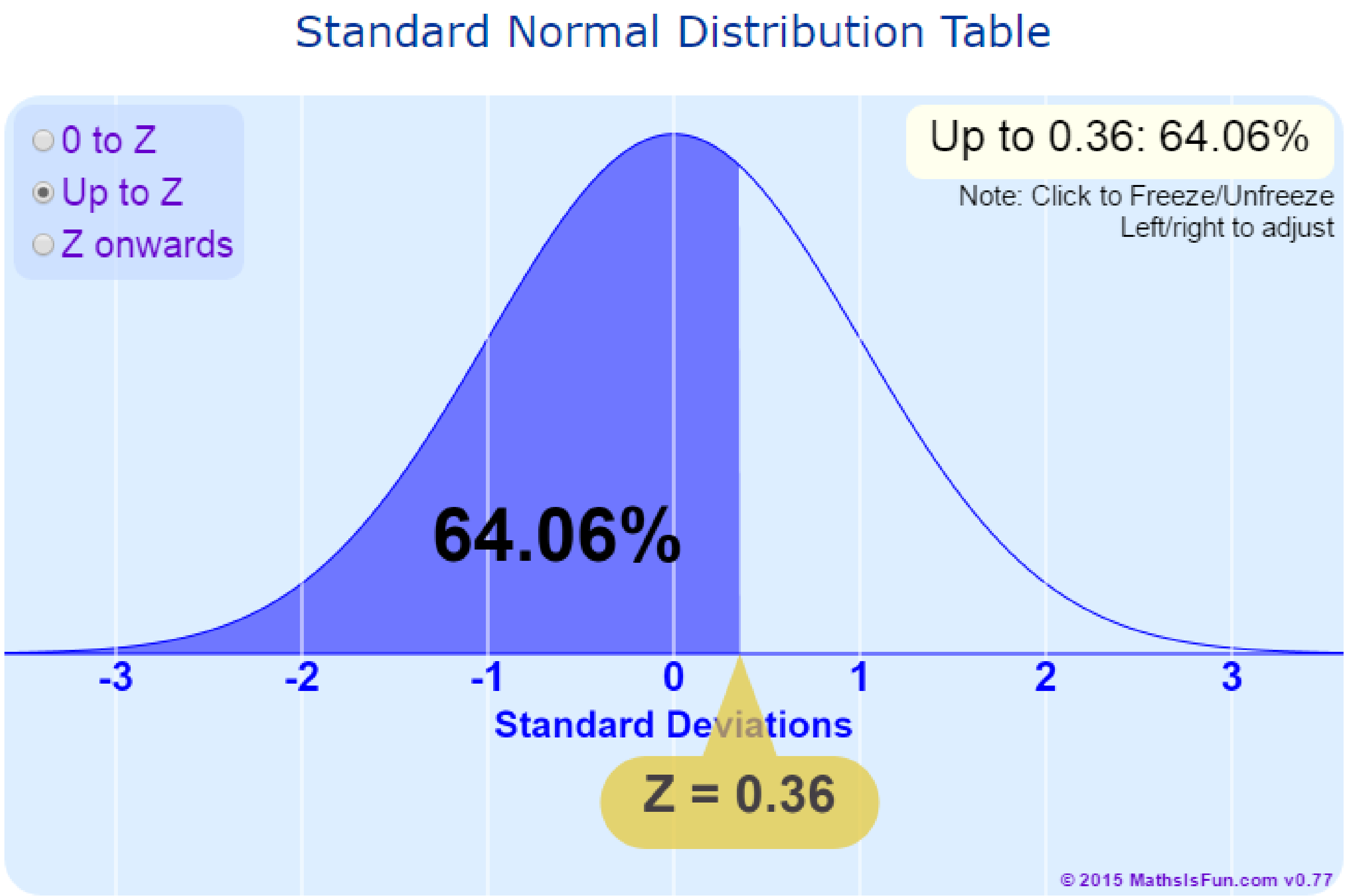
No closed form formula exists for area under curve for the Gaussian.
Hence we need to use pre-computed tables for obtaining area under the curve

We rescale the Gaussian distribution using the following transformation

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



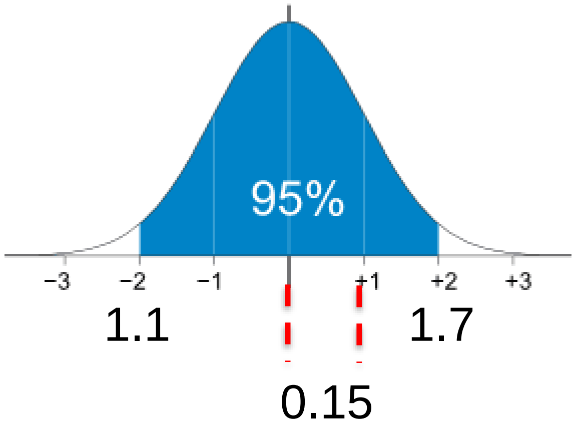
Standardize a Normal Distribution



Source: https://www.mathsisfun.com/data/standard-normal-distribution-table.html

Attention Check

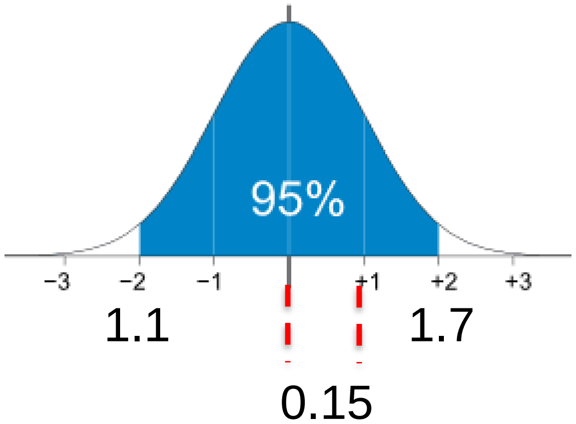
95% of students at a school are between 1.1m and 1.7m tall. Assuming this data is normally distributed, calculate the mean and the standard deviation.



- Mean is halfway between 1.1 and 1.7, and is equal to (1.1+1.7)/2 = 1.4m
- 95% is the area covered between -2σ and +2σ. That is, this spread is equal to 4 standard deviations. Therefore, 1 standard deviation is equal to (1.7-1.1)/4 = 0.15m.

Attention Check

In the same school, one of your friends is 1.85m tall. What is the z-score of your friend's height?



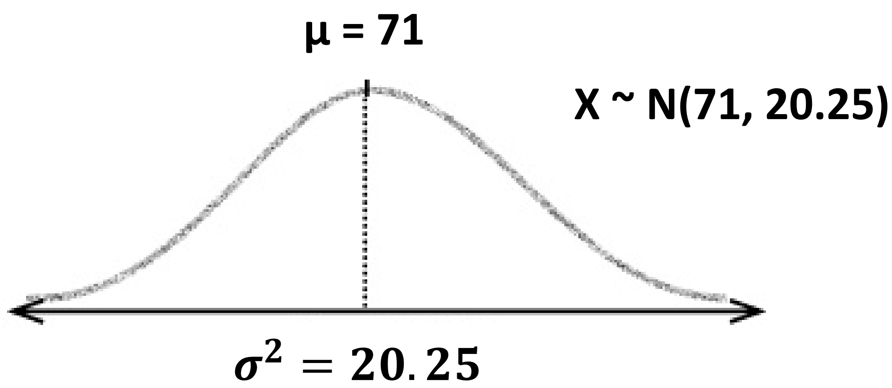
- Step 1: How far is 1.85 from the mean?
  - It is 1.85 – 1.4 = 0.45m from the mean.
- Step 2: How many standard deviations is that?
  - If 0.15m is 1σ, 0.45m is 0.45/0.15 = 3 standard deviations. Therefore, the z-score of your friend's height is 3.

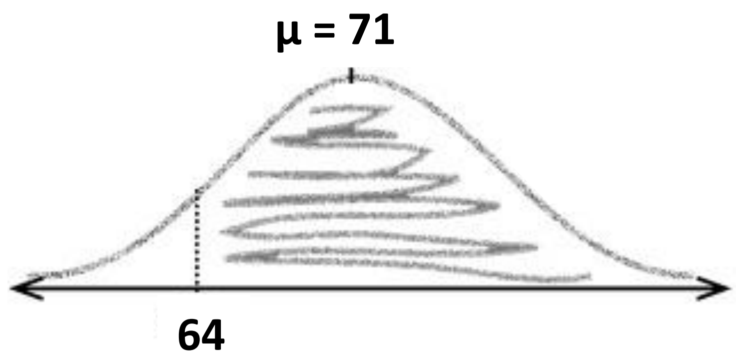
Calculating Normal Probabilities

Step 1: Determine the distribution

Julie wants to marry a person taller than her and is going on blind dates. The mean height of the ‘available’ guys is 71” and the variance is 20.25 inch^2 .

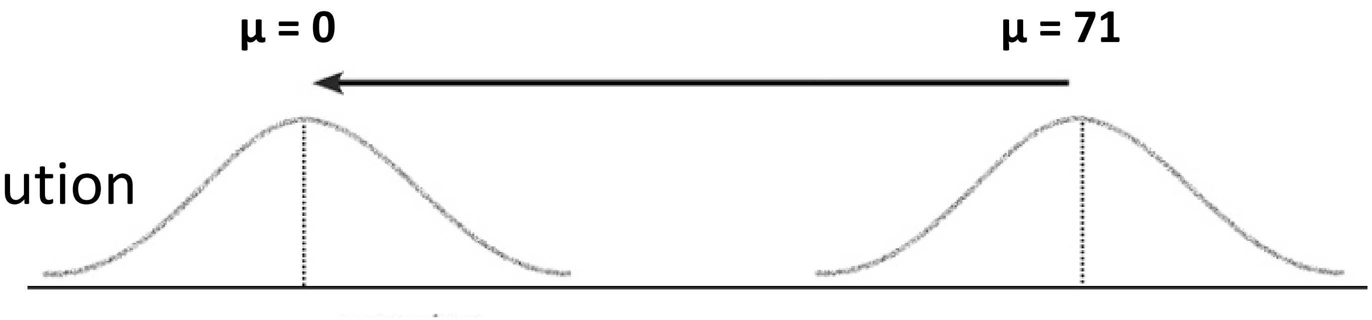
Oh! By the way, Julie is 64” tall.





Calculating Normal Probabilities

Step 2: Standardize to Z ~ N(0,1)



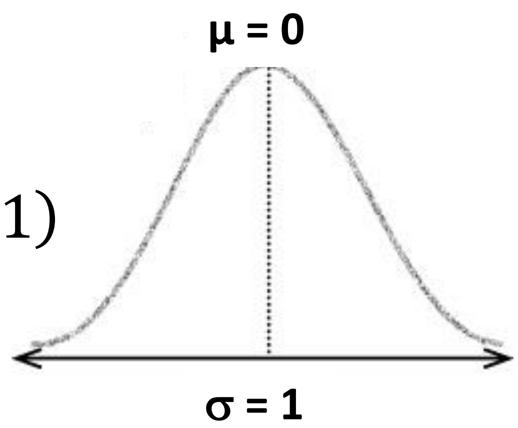
1. Move the mean

   This gives a new distribution

   X-71 ~ N(0,20.25)

2. Rescale the width by dividing by the standard deviation

   This gives us $\frac{X-71}{4.5} \sim N(0,1)$

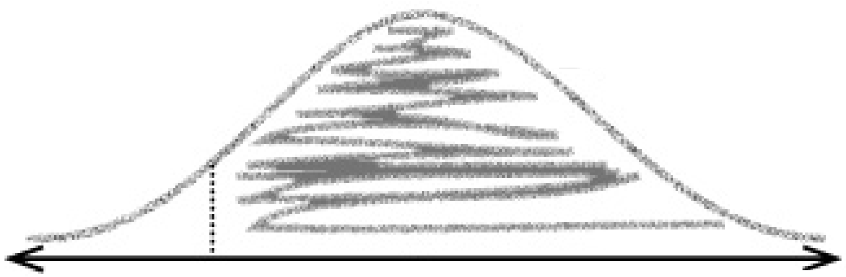


$Z = \frac{X-\mu}{\sigma}$ is called the Standard Score or the z-score.

Calculating Normal Probabilities

Step 2: Standardize to Z ~ N(0,1)

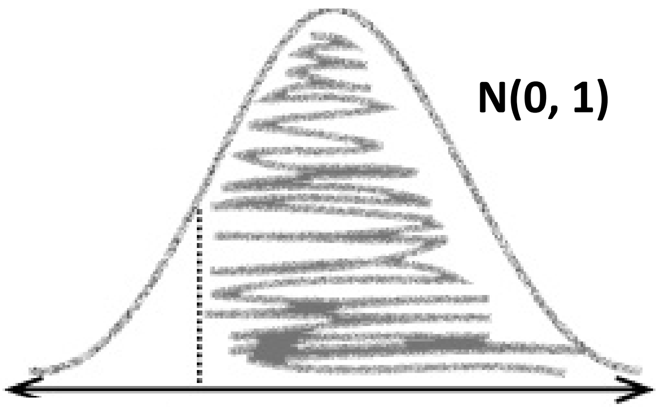
N(71, 20.25)



64

μ = 0

N(0, 1)



z

Note: R does this step internally but you must understand the concept of z-score as this is fundamental to most statistical thinking

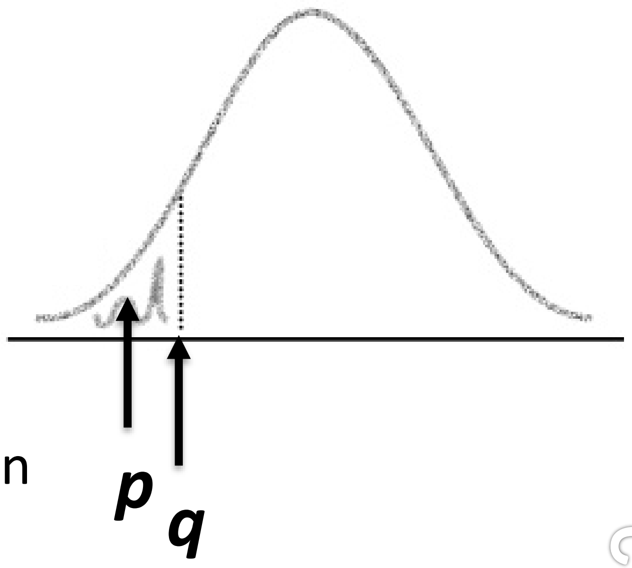
$Z = \frac{64-71}{4.5} = -1.56$ in the case of our problem.

Calculating Normal Probabilities

Step 3: Look up the probability in the tables

Note the tables give P(Z<z).

In R functions, the distribution is abbreviated and prefixed with an alphabet.



pnorm: Probability (Cumulative Distribution Function, CDF) in a *Normal Distribution*

qnorm: Quantile (Inverse CDF) in a *Normal Distribution* – The value corresponding to the desired probability.

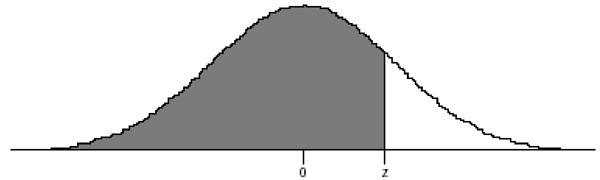
Calculating Normal Probabilities

Step 3: Look up the probability in the tables

Note the tables give P(Z<z).

$z = \frac{64-71}{4.5} = -1.56$ in the case of our problem.

P(Z>-1.56) = 1- P(Z<-1.56) = 1- 0.0594 = 0.9406



Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.7	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-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

Calculating Normal Probabilities

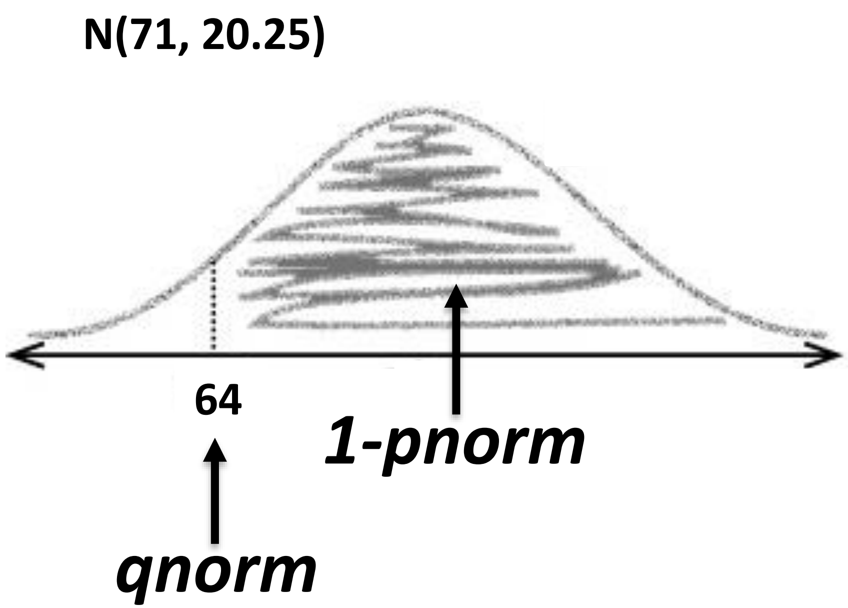
Step 3: Get the probability from R

1-pnorm(64, mean=71, sd=sqrt(20.25))
or
1-pnorm(64, 71, 4.5)

Answer: 1-0.0599 = 94.01%

qnorm(0.0599, 71, 4.5)

Answer: 64



Attention Check

Q. What is the standard score for N(10,4), value 6?

A. $z = \frac{6-10}{2} = -2$

Q. The standard score of value 20 is 2. If the variance is 16, what is the mean?

A. $2 = \frac{20-\mu}{4}.\ \therefore\ \mu = 20 - 8 = 12$

Attention Check

Q. Julie just realized that she wants her date to be taller when she is wearing her heels, which are 5" high. Find the new probability that her date will be taller.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

A. $z = \frac{69-71}{4.5} = -0.44$; P(Z<-0.44) = 0.33, ∴ P(Z>-0.44) = 0.67 or 67%

A. 1-pnorm(69, 71, 4.5). This gives P(X>69) = 67%

Attention Check

Q. Julie wants to have at least 80% probability of finding the right guy. What is the maximum size of heels she can wear?

A. qnorm(0.20, 71, 4.5). This gives a value of 67.2”. As Julie is 64” tall, the maximum heel size she should wear is about 3”.

Outlier Detection - Excel

Hadlum vs Hadlum case

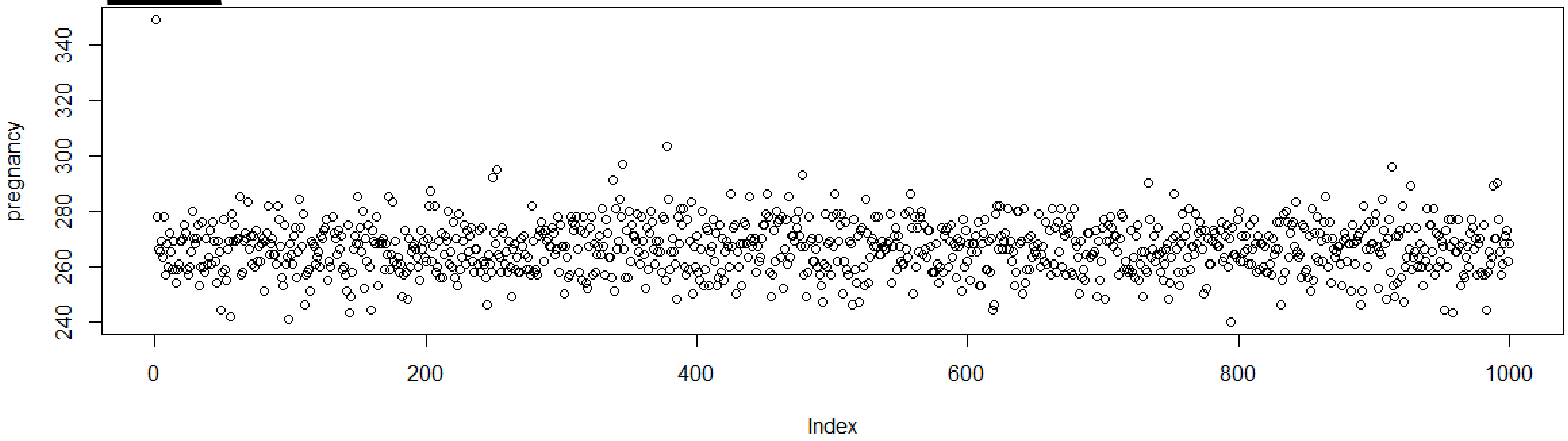
Source: http://www.alphamom.com/legacy/pregnancy-calendar/week36.jpg
Last accessed: November 01, 2014

Source: http://3.bp.blogspot.com/-0YwIRjLMWr0/T4DqOwVClgI/AAAAAAAAAgg/Yjf-ttkQLSg/s1600/fishy.jpg
Last accessed: November 01, 2014

CSE 7315c

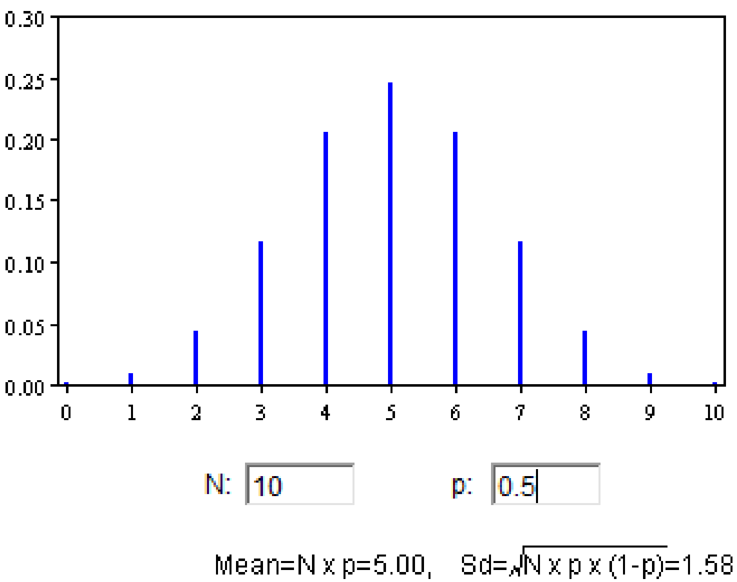
Outlier Detection

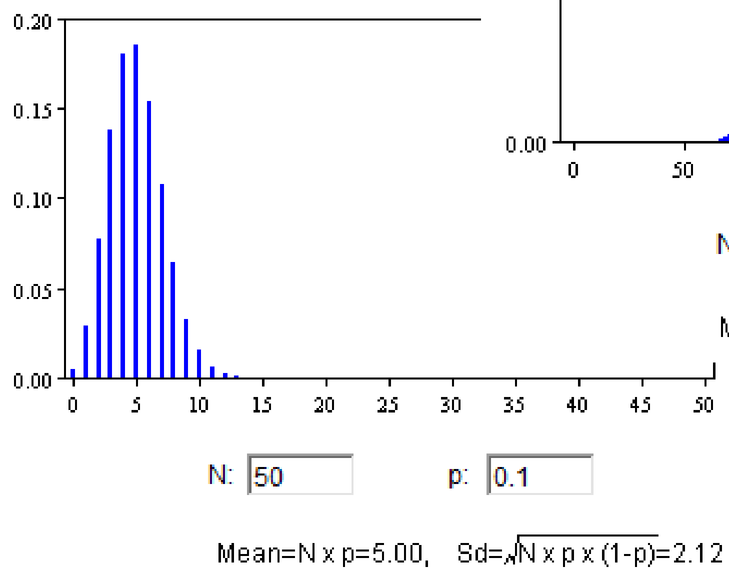
Hadlum vs Hadlum case

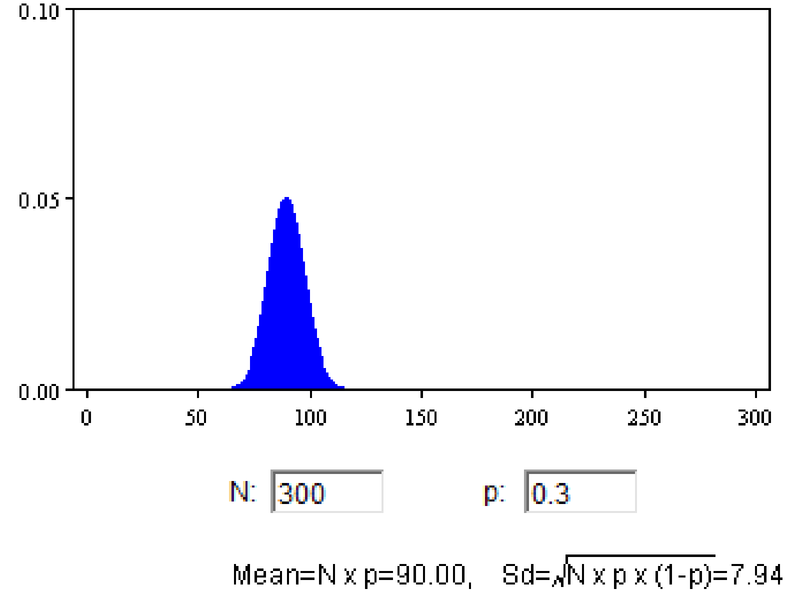


Normal Distribution

Binomial distribution can be approximated to a Normal distribution if np>5 and nq>5 (Continuity Correction required).

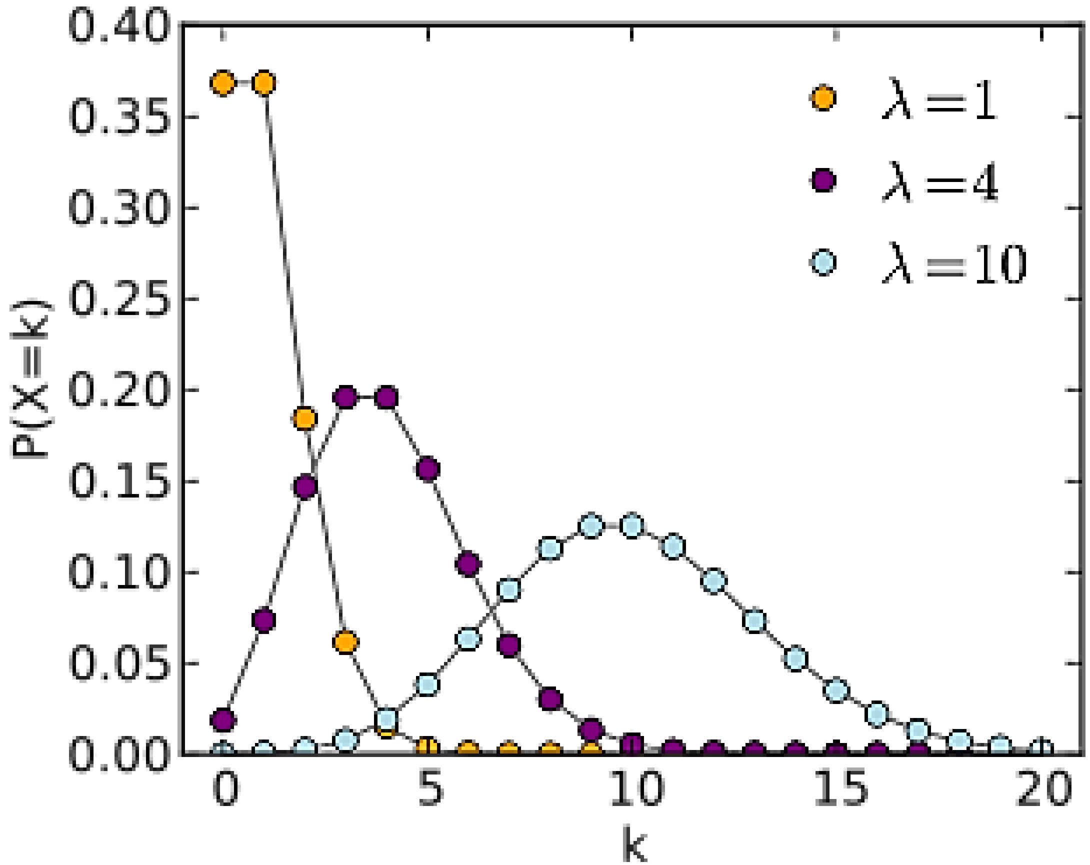






Normal distribution

Poisson distribution can be approximated to a Normal distribution when $\lambda > 15$ (Continuity Correction required).



Continuity Correction?

You are playing Who Wants to Win a 

What is the probability you will get 5 or fewer correct out of 12, given that each question has only 2 possible choices? You have no lifelines.

X ~ B(12, 0.5) and we need to find P(X < 6).

$$P(X=0) = {}^{12}C_0(0.5)^0(0.5)^{12-0} = 0.5^{12}$$
$$P(X=1) = {}^{12}C_1(0.5)^1(0.5)^{12-1} = 12 * 0.5^{12}$$
$$P(X=2) = {}^{12}C_2(0.5)^2(0.5)^{12-2} = 66 * 0.5^{12}$$
$$P(X=3) = {}^{12}C_3(0.5)^3(0.5)^{12-3} = 220 * 0.5^{12}$$
$$P(X=4) = {}^{12}C_4(0.5)^4(0.5)^{12-4} = 495 * 0.5^{12}$$
$$P(X=5) = {}^{12}C_5(0.5)^5(0.5)^{12-5} = 792 * 0.5^{12}$$
$$\therefore P(X<6) = (1+12+66+220+495+792) * 0.5^{12} \cong 0.387$$

Continuity Correction?

X ~ B(12, 0.5) can be approximated to X ~ N(6, 3). How/Why?

n = 12, p = 0.5 and q = 0.5. Since np and nq are both > 5, the Binomial distribution can be approximated to a Normal distribution, i.e., X ~ B(n, p) can be approximated to X ~ N(np, npq).

If we want to get P(X < 6), what is the next step to do in the Normal distribution?

Calculate the z-score (or the standard-score).

$$z = \frac{x - \mu}{\sigma} = \frac{6 - 6}{\sqrt{3}} = 0$$

What do we do with the z-score?

Look it up in the probability tables or R.

What is the probability corresponding to the z-score of 0?

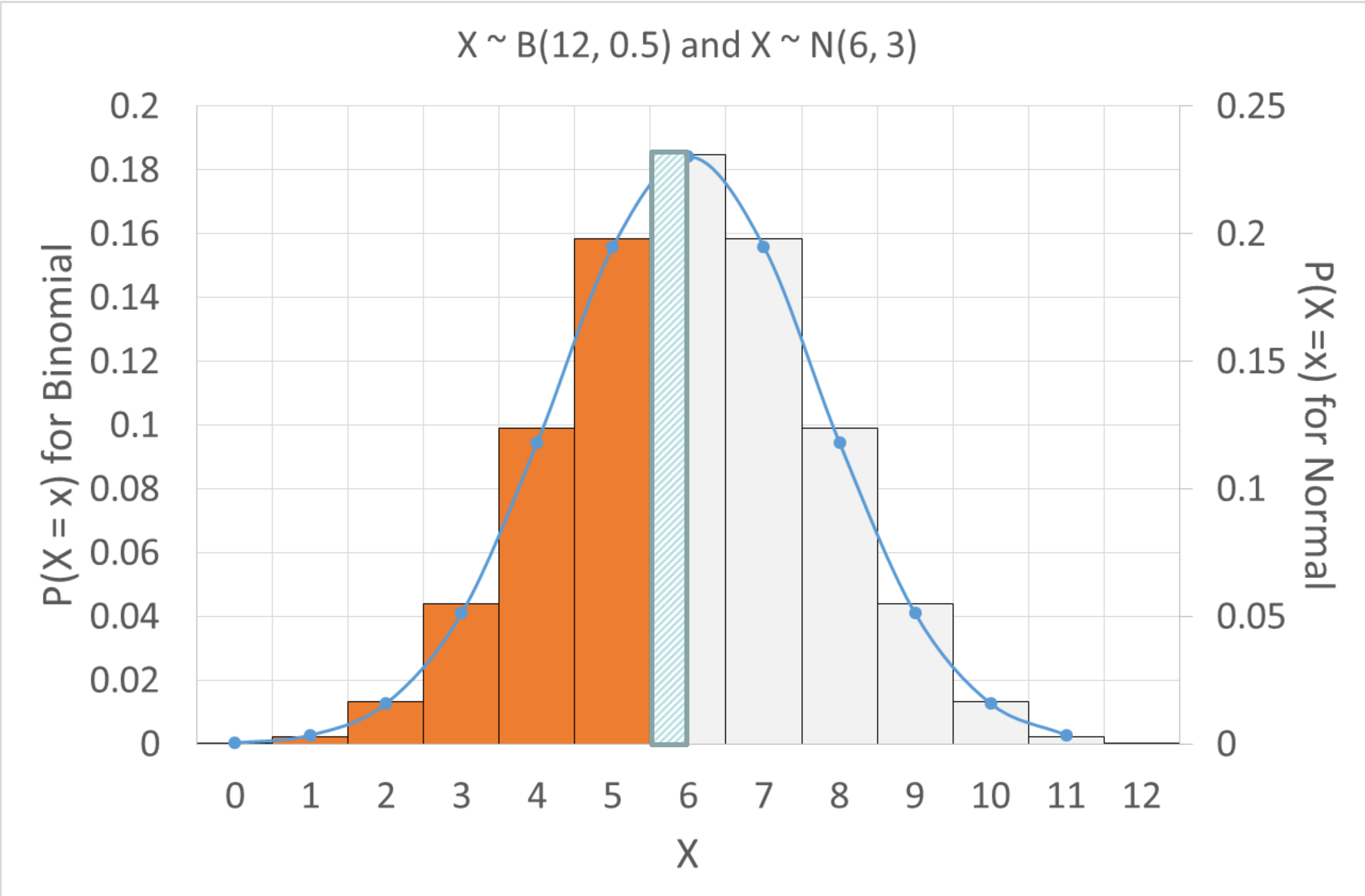
P(X < 6) = 0.5

Z	0.00	0.01	0.02	0.03	0.04	0.05
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647

Continuity Correction?

So, P(X < 6) = 0.387 for X ~ B(12, 0.5)

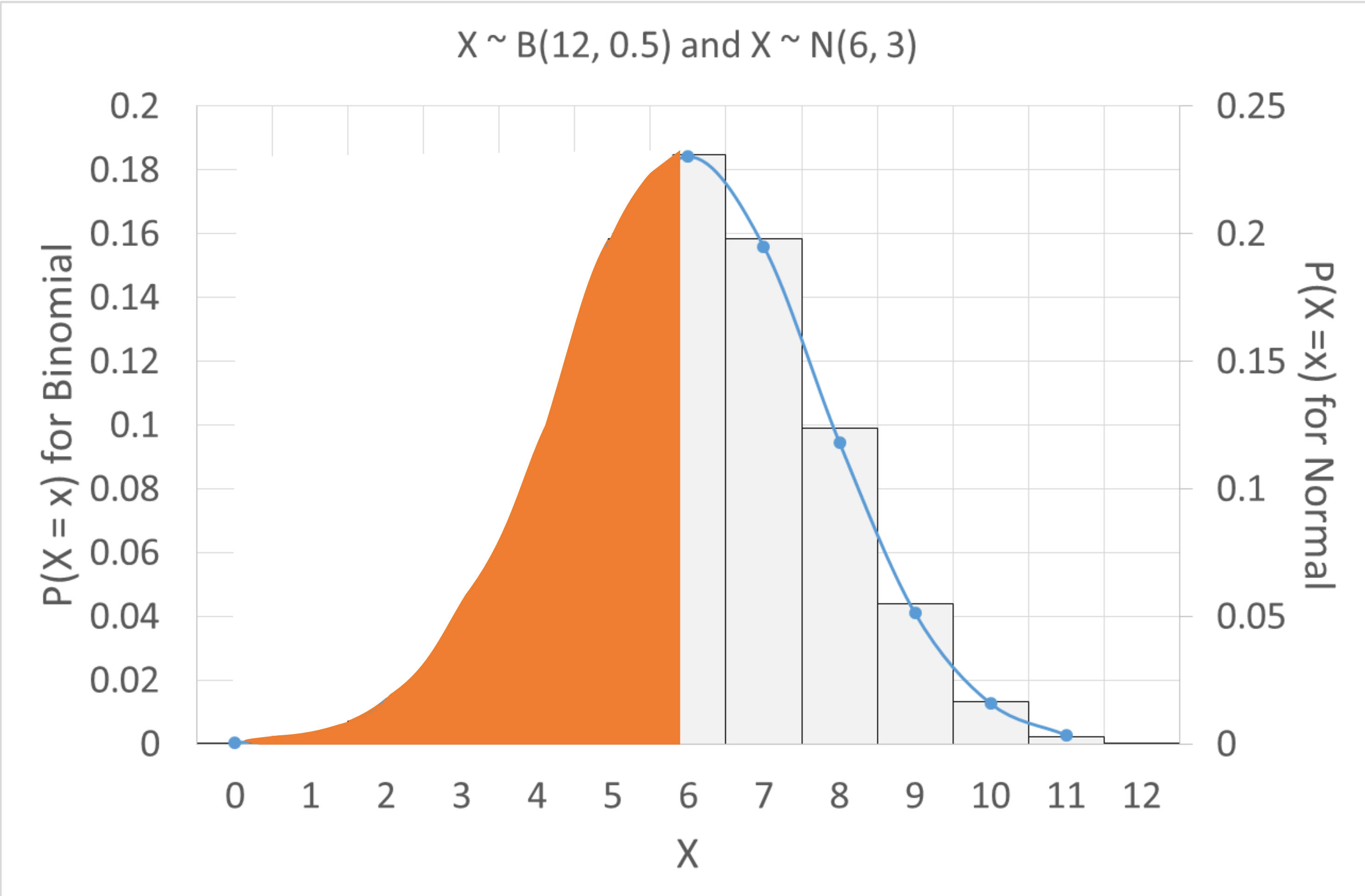
and P(X < 6) = 0.5 for X ~ N(6, 3). Is this a good approximation?



Continuity Correction?

So, $P(X < 6) = 0.387$ for $X \sim B(12, 0.5)$

and $P(X < 6) = 0.5$ for $X \sim N(6, 3)$. Is this a good approximation?



Continuity Correction?

So, P(X < 6) = 0.387 for X ~ B(12, 0.5)

and P(X < 6) = 0.5 for X ~ N(6, 3).

$$z = \frac{5.5 - 6}{\sqrt{3}} = -0.29$$

P(X < 5.5) = 0.3859 for X ~ N(6, 3)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Continuity Correction

Identify the right continuity correction for each discrete probability distribution.

Discrete	Continuous
$X < 3$	$X < 2.5$
$X > 3$	$X > 3.5$
$X \leq 3$	$X < 3.5$
$X \geq 3$	$X > 2.5$
$3 \leq X < 10$	$2.5 < X < 9.5$
$X = 0$	$-0.5<X<0.5$
$3 \leq X \leq 10$	$2.5<X<10.5$
$3 < X \leq 10$	$3.5 < X < 10.5$
$X > 0$	$X>0.5$
$3 < X < 10$	$3.5<X<9.5$

Continuity Correction

You are playing Who Wants to Win a 

What is the probability of getting at least 30 out of 40 questions correct, where each question has 2 possible choices?

$P(X \geq 30)$ where $X \sim B(40, 0.5)$

$P(X > 29.5)$ where $X \sim N(20, 10)$

$$z = \frac{29.5 - 20}{\sqrt{10}} = 3.004$$

P(X>29.5) = 0.003/2

Normal Distribution

You have designed a new game, Angry Buds. The key to success is that it should not be so difficult that people get frustrated, nor should it be so easy that they don't get challenged. Before building the new level, you want to know what the <u>mean</u> and <u>standard deviation</u> are of the number of minutes people take to complete level 1. You know the following:

1. The # of minutes follows a normal distribution.
2. The probability of a player playing for less than 5 minutes is 0.0045.
3. The probability of a player playing for less than 15 minutes is 0.9641.

Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−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

$P(X < 5) = 0.0045$

$z_1 = -2.61$

Normal Distribution

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
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

$P(X < 15) = 0.9641$

$z_2 = 1.8$

Normal Distribution

$-2.61 = \frac{5-\mu}{\sigma}$ and $1.8 = \frac{15-\mu}{\sigma}$

Solving for the above 2 equations, we get

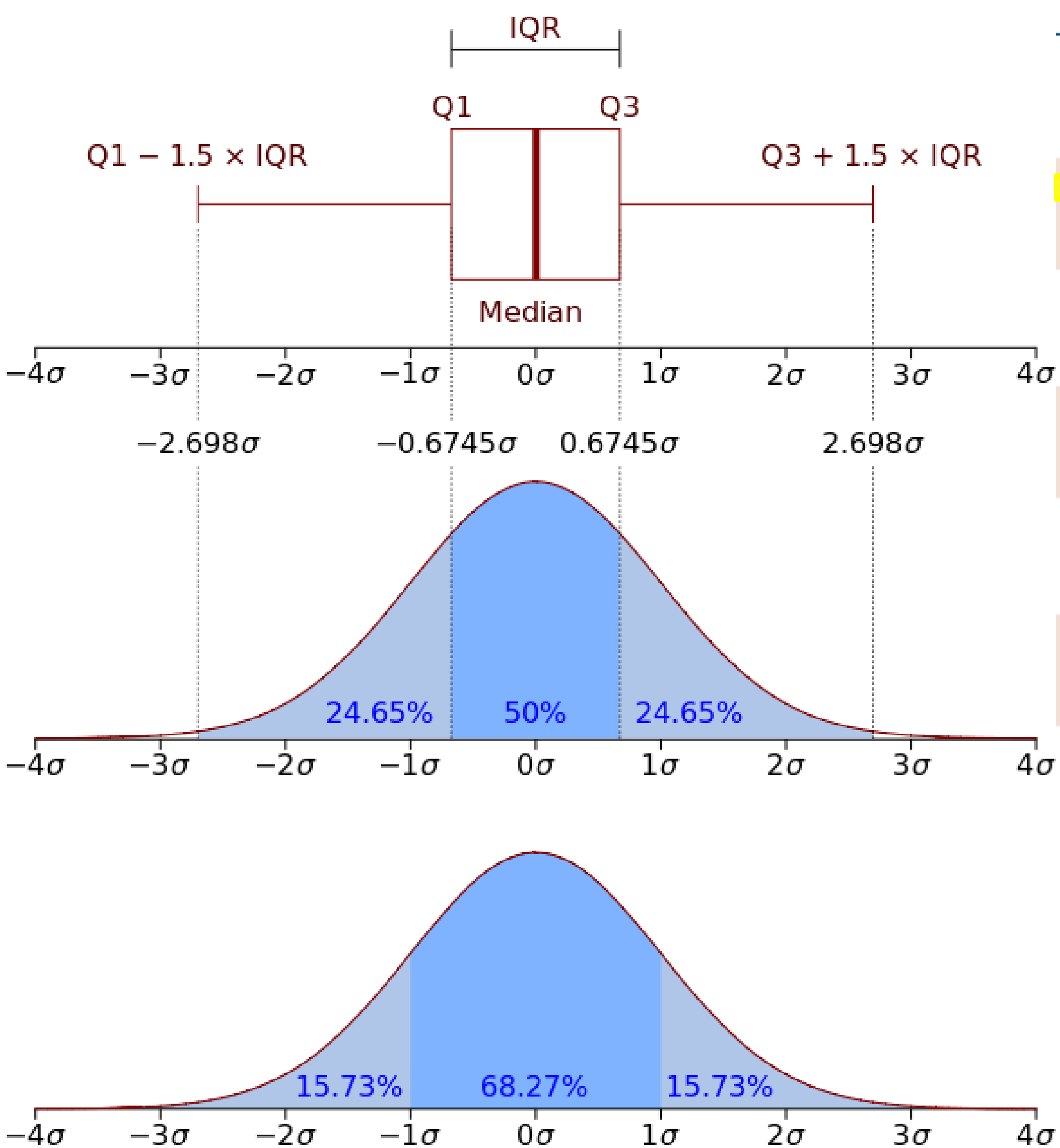
$\mu = 5 + 2.61\sigma$

$\mu = 15 - 1.8\sigma$

Subtracting the two, we get

$0 = -10 + 4.41\sigma \Rightarrow \ \sigma = 10 \div 4.41 = 2.27$

Substituting this value of σ in either of the above 2 equations, we get $\mu = 5 + 2.61 * 2.27 = 10.925$



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
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Source: http://en.wikipedia.org/wiki/Box_plot#mediaviewer/File:Boxplot_vs_PDF.svg; Last accessed: July 01, 2014

Academic Exercise

0.75 is not in the z-table. The closest values are 0.7486 corresponding to a z-score of 0.67 and 0.7517 corresponding to a z-score of 0.68.

Method 1: Take the closest value, 0.7486 in this case.

Method 2: Interpolate.

z-value	P(0<Z<z)	Proportional distance from top
0.67	0.7486	0.0000
c (to be found)	0.75 (desired)	0.4516
0.68	0.7517	1.0000

$$\frac{0.75 - 0.7486}{0.7517 - 0.7486} = \frac{0.0014}{0.0031} = 0.4516$$

$$\Rightarrow \frac{c-0.67}{0.68-0.67} = 0.4516$$

$$\therefore c = 0.67 + 0.4516 * 0.01 = 0.6745$$

SAMPLING DISTRIBUTION OF MEANS

Sampling Distribution

- The core goal of inferential statistics is to be able to make intelligent conclusions about the population parameters by looking at sample statistics.
- For eg: Estimate the mean height of the students in a class, from a small sample.

Sampling Distribution of the Means

- The sampling distribution of means is what you get if you consider all possible samples of size *n* taken from the same population and form a distribution of their means.

- Each randomly selected sample is an independent observation.

Central Limit Theorem

- http://onlinestatbook.com/2/sampling_distributions/clt_demo.html

- As sample size goes large and number of buckets are high, the means will follow a normal distribution with same mean (μ) and $\frac{1}{n}$ of variance (σ^2).

Expectation and Variance for $\bar{X}$

$$E(\bar{X}) = \mu$$

Mean of all sample means of size *n* is the mean of the population.

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

<u>Standard deviation of $\bar{X}$</u> tells how far away from the population mean the sample mean is likely to be and is called the **Standard Error of the Mean**, and is given by

$$Standard\ Error\ of\ the\ Mean = \frac{\sigma}{\sqrt{n}}$$

If X ~ N(μ, σ^2), then $\bar{X}$ ~ N(μ, σ^2/n)

Using the Central Limit Theorem

Let us say the mean number of Gems per packet is 10, and the variance is 1. If you take a sample of 30 packets, what is the probability that the sample mean is 8.5 Gems per packet or fewer?

Using the Central Limit Theorem

We know that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, μ = 10, σ^2 = 1 and n = 30.

We need the value of P($\bar{X}$ < 8.5) when $\bar{X}$ ~ N(10, 0.0333).

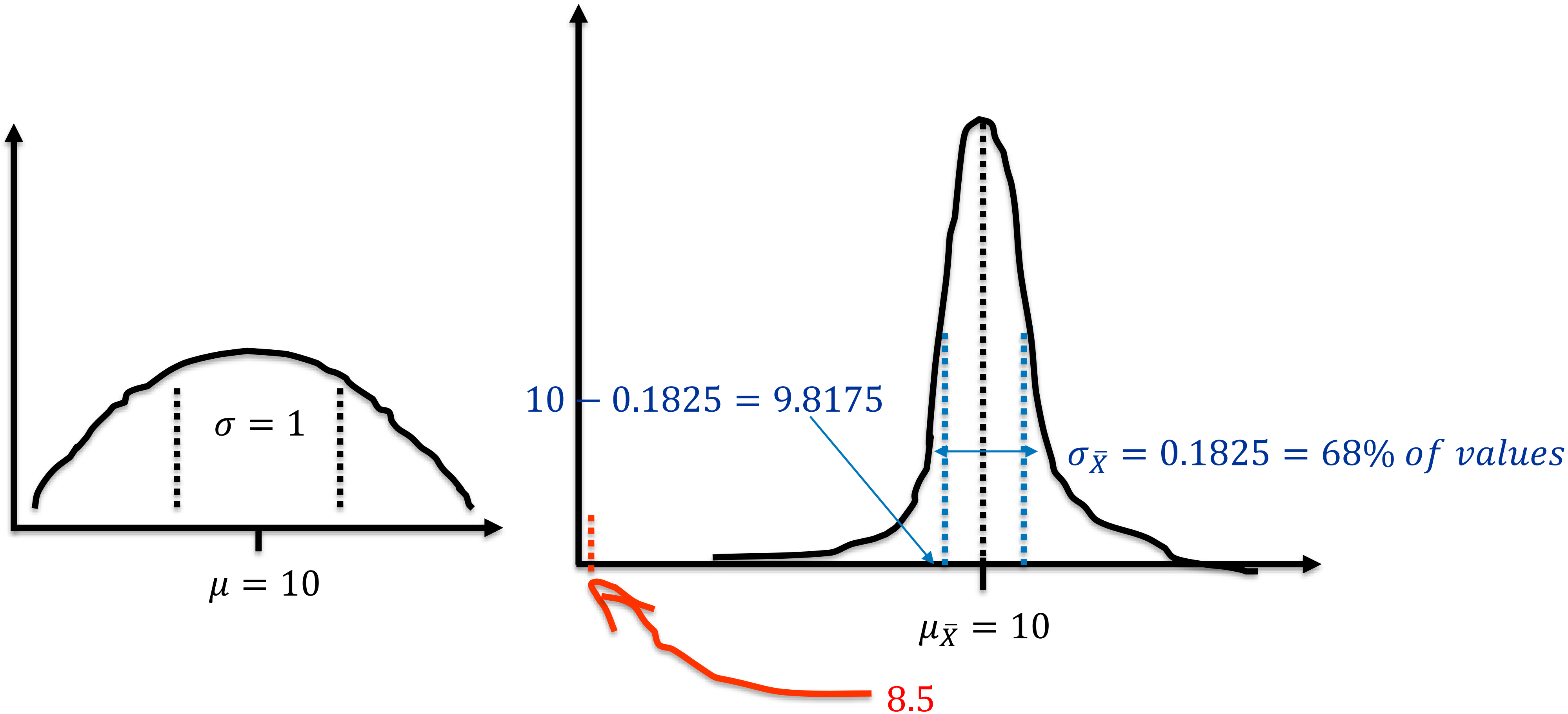
$$z = \frac{8.5 - 10}{\sqrt{0.0333}} = -8.22$$

P(Z < z) = P(Z < -8.22)

This doesn't exist in probability tables. What does it mean?

Using the Central Limit Theorem

How do we visualize it?

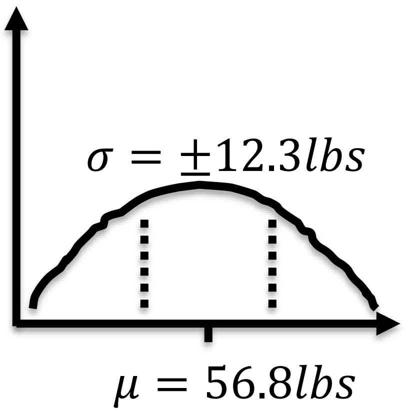


Using the Central Limit Theorem

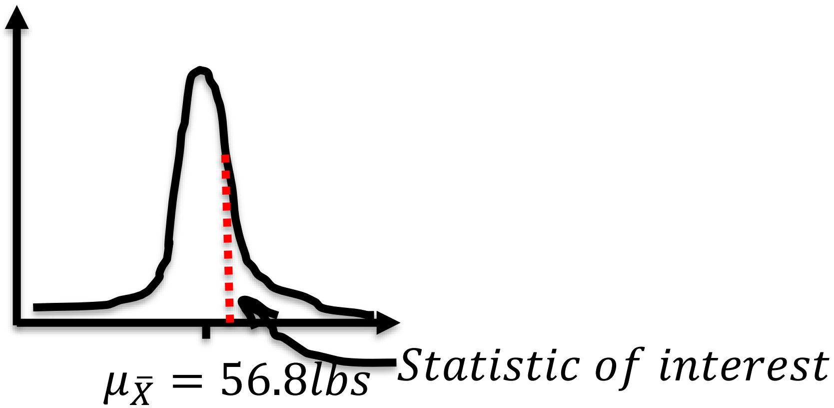
The Aluminum Association of America reports that the average American household uses 56.8 lbs of aluminium in a year. A random sample of 51 households is monitored for one year to determine aluminium usage. If the population standard deviation of annual usage is 12.3 lbs, what is the probability that the sample mean will be $> 60\ lbs$?

Sampling Distribution

Population distribution

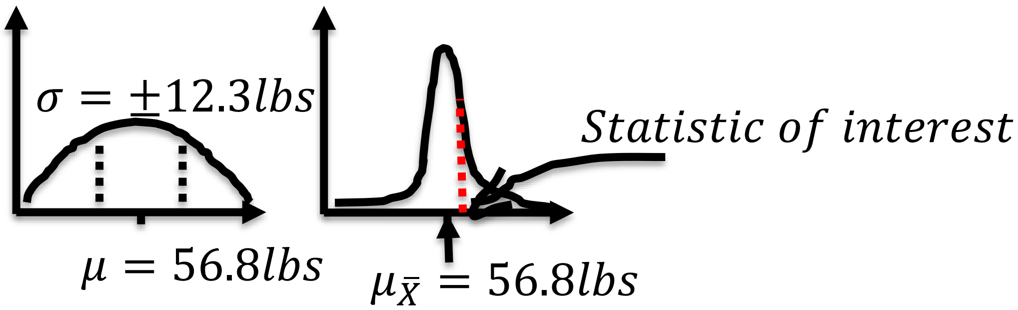


Sampling distribution of sample mean when $n = 51$



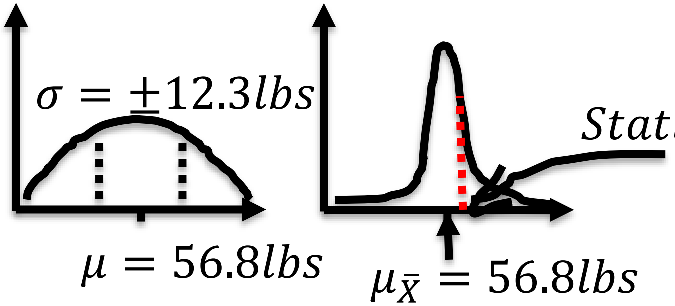
- Step 1: List all known parameters and values
- Step 2: Calculate others, or estimate if cannot be calculated
- Step 3: Find probabilities using tables, Excel or R

Sampling Distribution



- Step 1: List all known parameters and values
  - Population mean, $\mu = 56.8\ lbs$
  - Population standard deviation, $\sigma = 12.3\ lbs$
  - Sample size, $n = 51$
  - Sample mean, $\bar{x} > 60\ lbs$
  - Mean of sample means, $\mu_{\bar{X}} = \mu = 56.8\ lbs$

Sampling Distribution



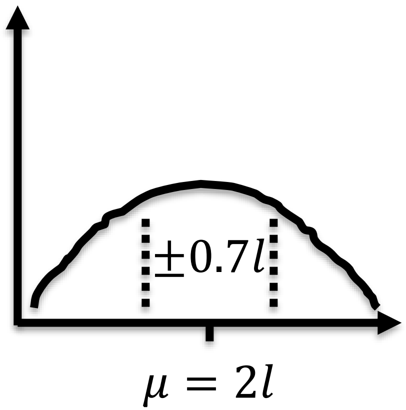
- Step 2: Calculate others or estimate, if cannot be calculated
  - Standard deviation of sample means, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{12.3}{\sqrt{51}} = 1.72$
  - $\therefore z = \frac{60-56.8}{1.72} = 1.86$
- Step 3: Find probabilities using tables, Excel or R
  - Excel: 1-NORM.S.DIST(z,TRUE) = 0.0316
  - Please calculate these for:
    - $> 58\ lbs$
    - $> 56\ lbs < 57\ lbs$
    - $< 50\ lbs$

Sampling Distribution

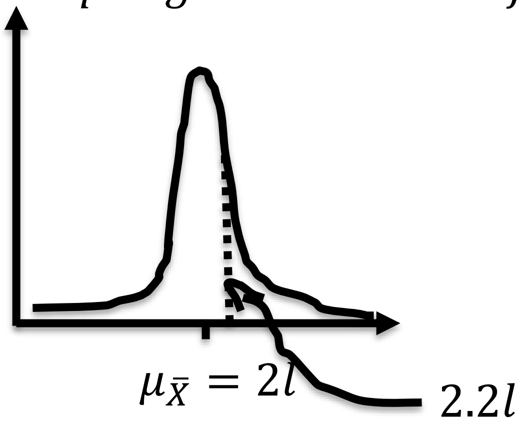
The average male drinks $2l$ of water when active outdoors with a standard deviation of $0.7l$. You are planning a trip for 50 men and bring $110l$ of water. What is the probability that you will run out of water?

$\mu = 2, \sigma = 0.7$

$P(run\ out) \Rightarrow P(use > 110l) \Rightarrow P(average\ water\ use\ per\ male > 2.2l)$



$Sampling\ distribution\ of\ sample\ mean\ when\ n = 50$



$\mu_{\bar{X}} = \mu = 2l, {\sigma_{\bar{X}}}^2 = \frac{\sigma^2}{n} = \frac{0.49}{50}$

$\Rightarrow \sigma_{\bar{X}} = 0.099$

$z = \frac{2.2 - 2}{0.099} = 2.02$

$P(\bar{X} < 2.02) = 0.9783$

The probability of running out is

1-0.9783 = 0.0217 or 2.17%

Source: https://www.khanacademy.org/math/probability/statistics-inferential/sampling_distribution/v/sampling-distribution-example-problem

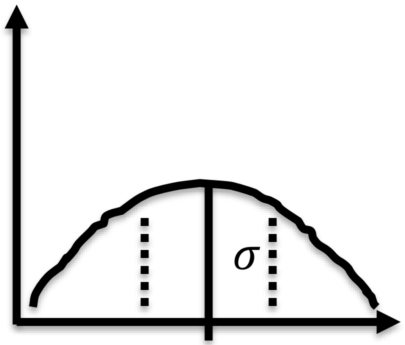
Last accessed: May 9, 2014

Using the Central Limit Theorem

You sample 36 apples from your farm's harvest of 200,000 apples. The mean weight of the sample is 112g with a 40g sample standard deviation. What is the probability that the mean weight of all 200,000 apples is between 100 and 124g?

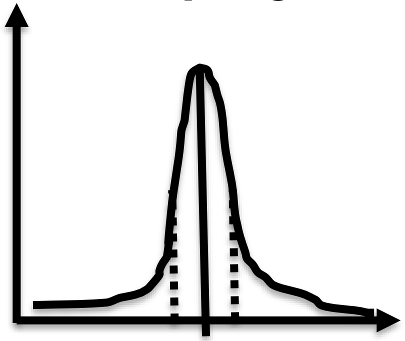
Sampling Distribution

Population distribution



μ

Sampling distribution of sample mean when $n = 36$



$\mu_{\bar{X}} = \mu$

$$\sigma_{\bar{X}}{}^2 = \frac{\sigma^2}{n} = \frac{\sigma^2}{36} \Rightarrow \sigma_{\bar{X}} = \frac{\sigma}{6}$$

What are we trying to find out?

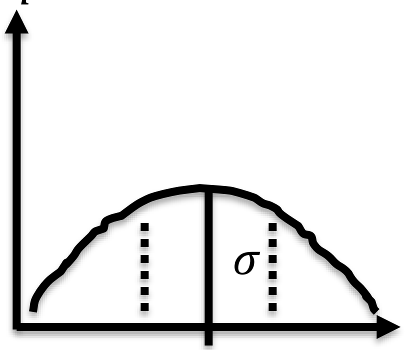
We need to know if population mean, μ, is within $\pm$12g of the sample mean, $\bar{X}$.

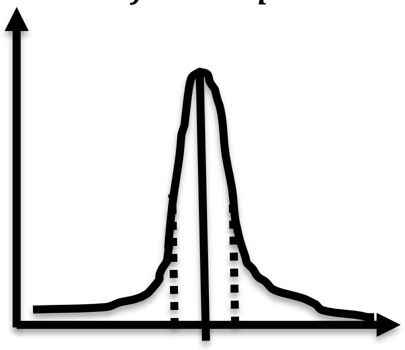
This is the same as saying that we need to know if sample mean, $\bar{X}$, is within $\pm$12g of the population mean, μ. Since $\mu = \mu_{\bar{X}}$, we can now use the sampling distribution of the means.

Source: https://www.khanacademy.org/math/probability/statistics-inferential/confidence-intervals/v/confidence-interval-1
Last accessed: May 9, 2014

Sampling Distribution

$Population\ distribution$ $\quad$ $Sampling\ distribution\ of\ sample\ mean\ when\ n = 36$





We need to find out how many standard deviations away from $\mu_{\bar{X}}$ is 12g. But, we don't know $\sigma_{\bar{X}}$ because we don't know σ. We use the sample standard deviation, *s* (40g), as the best estimate of population standard deviation. $\sigma \approx s = \pm 40g$. $\therefore \sigma_{\bar{X}} = \frac{\sigma}{6} = \frac{40}{6} = 6.67$. So 12g is 12/6.67 = 1.8 standard deviations.

The z-table gives the probability as 0.9641 but that is the entire region below +1.8 z.

Find the region between -1.8 and +1.8 z.

0.9282. How would you get this answer if you did not have the negative z table?

Source: https://www.khanacademy.org/math/probability/statistics-inferential/confidence-intervals/v/confidence-interval-1
Last accessed: May 9, 2014

Normal Distribution

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
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

$(0.9641-0.5000)*2$

$= 0.4641*2$

$= 0.9282$

Resources

- Binomial Distribution: https://www.khanacademy.org/video/binomial-distribution
- Poisson Distribution: https://www.khanacademy.org/video/poisson-process-1
- Normal Distribution: https://www.youtube.com/watch?v=RKdB1d5-OE0
- Central Limit Theorem: https://www.khanacademy.org/video/central-limit-theorem

International School of Engineering

Plot 63/A, Floors 1&2, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals:	+91-9502334561/63 or 040-65743991
For Corporates:	+91-9618483483
Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering