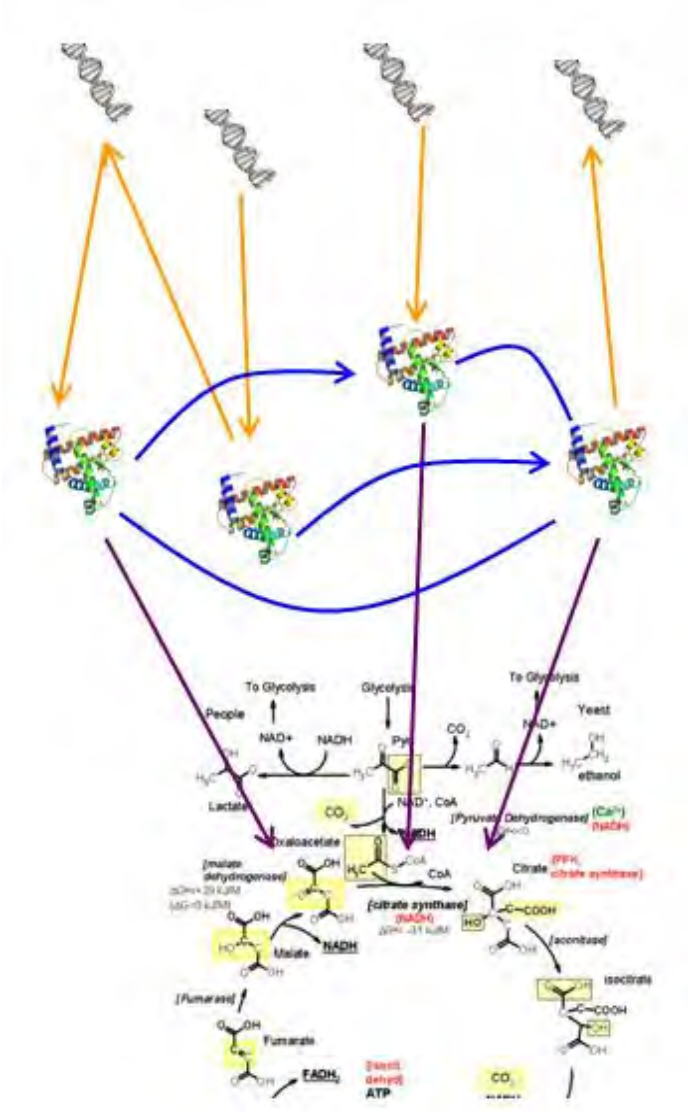
Fitting a Gaussian distribution: a biological example

Molecular binding is used at multiple levels

Each level has its own molecular interaction network



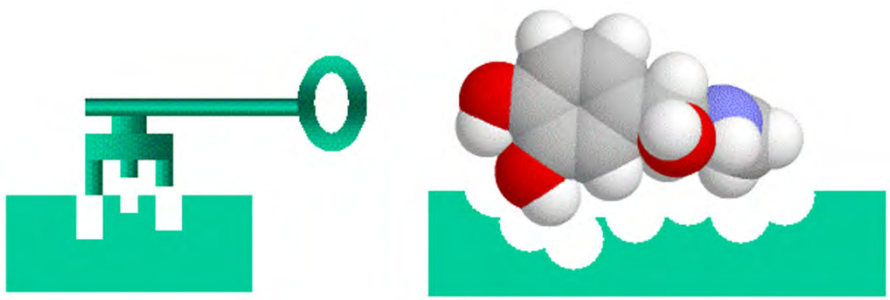
Regulatory network:
RNA-level regulation
By DNA-binding Proteins

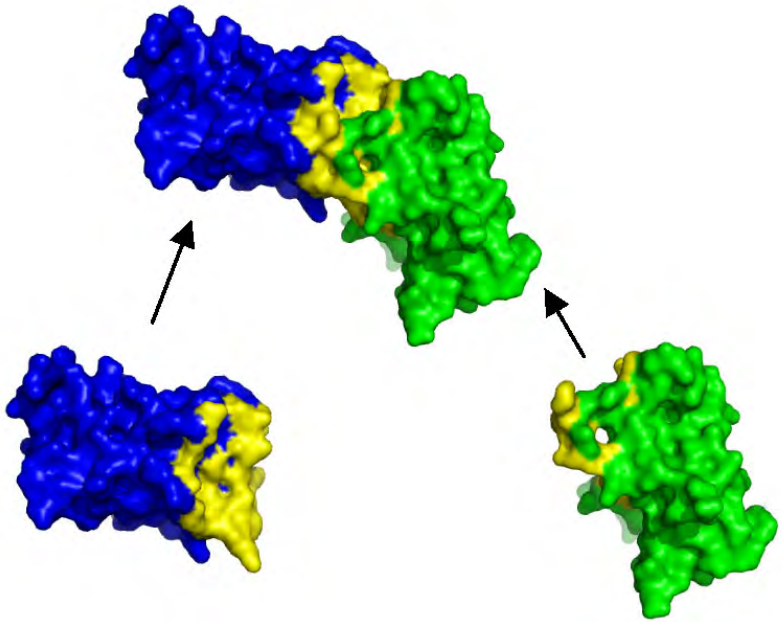
Protein-Protein (binding) Interaction Network

Protein-Metabolite Interactions: Metabolic network

Biological example of a Gaussian: Energy of Protein-Protein Binding Interactions

- Proteins and other biomolecules (metabolites, drugs, DNA) specifically (and non-specifically) bind each other
- For specific bindings: Lock-and-Key theory
- For non-specific bindings: random contacts





A simple physical model for scaling in protein–protein interaction networks

Eric J. Deeds*, Orr Ashenberg†, and Eugene I. Shakhnovich‡§

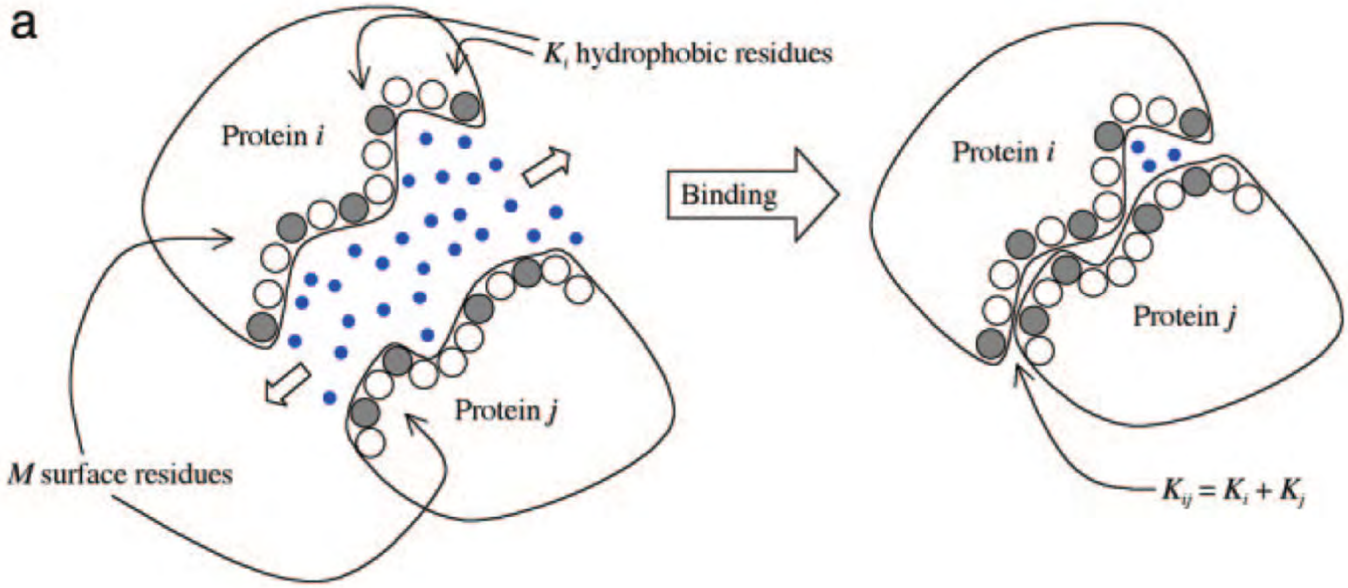
*Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138; †Harvard College, 12 Oxford Street, Cambridge, MA 02138; and ‡Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

Communicated by David Chandler, University of California, Berkeley, CA, November 10, 2005 (received for review September 23, 2005)

It has recently been demonstrated that many biological networks exhibit a "scale-free" topology, for which the probability of observing a node with a certain number of edges (k) follows a power law: i.e., $p(k) \sim k^{-\gamma}$. This observation has been reproduced by

(19–22). Indeed, when the two major *S. cerevisiae* p protein interaction (PPI) experiments are compared w another, one finds that only ≈150 of the thousands of tions identified in each experiment are recovered in th

Most Binding energy is due to hydrophobic amino-acid residues being screened from water

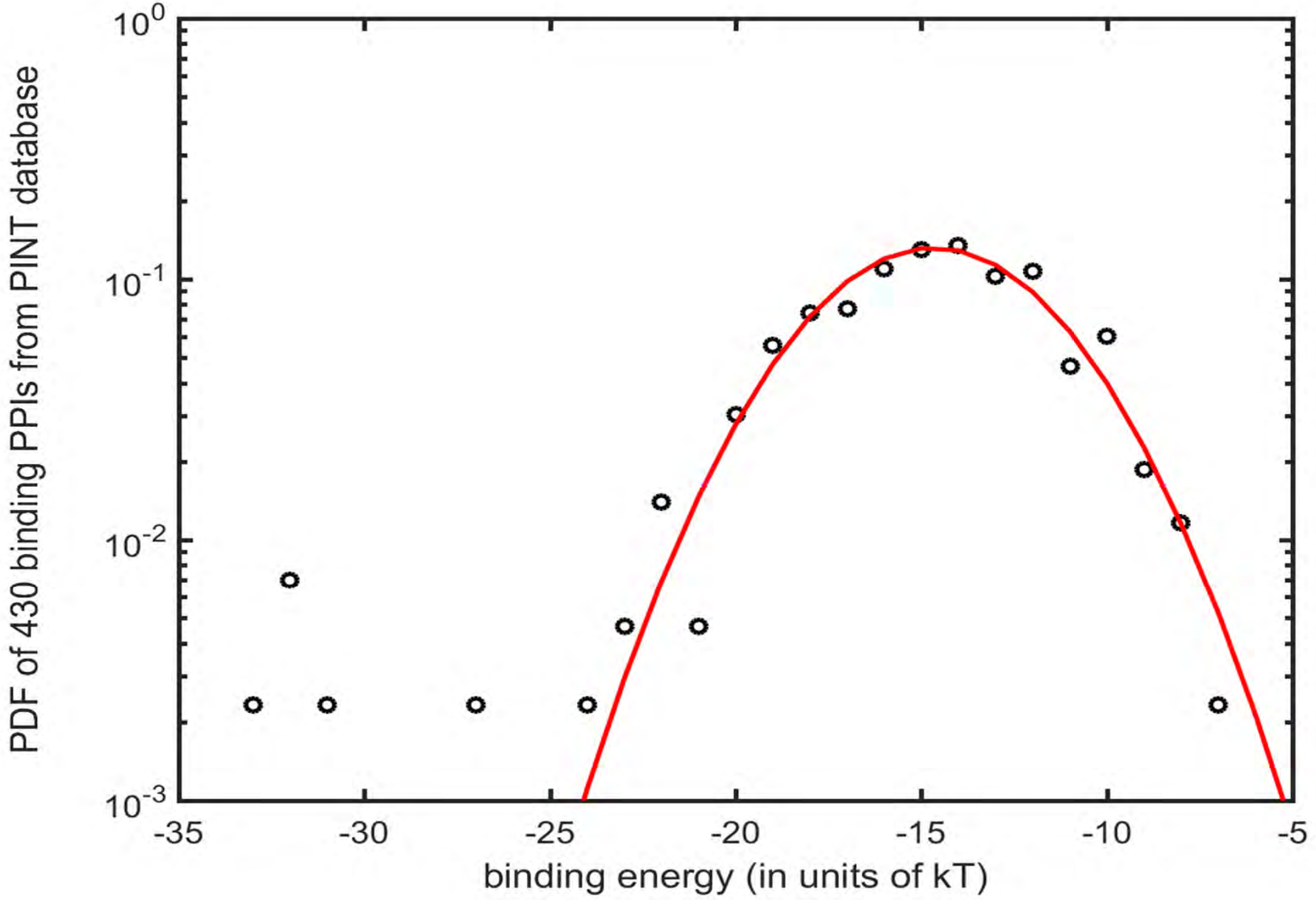


Predicted Gaussian distribution: PDF(E_{ij}=E)– because E_{ij} – sum of hydrophobicities of many independent residues

Matlab exercise

- In Matlab load PINT_binding_energy.mat with binding energy E_{ij} (in units of kT at room temperature) for 430 pairs of interacting proteins from human, yeast, etc.
- Data collected in 2007 from the PINT database http://www.bioinfodatabase.com/pint/
  and analyzed in J. Zhang, **S. Maslov**, E. Shakhnovich, Molecular Systems Biology (2008)
- Fit Gaussian to the distribution of E_{ij} using dfittool
- Use "Exclude" button to generate the new exclusion rule to drop all points with X<-23 from the fit
- Use "New Fit" button to generate the new "Normal" fit with the exclusion rule you just created
- Find mean (mu) and standard deviation (sigma)
- Select "probability plot" from "Display type" dropdown menu to evaluate the quality of the plot. Where does the probability plot deviate from a straight line?

How does it compare with the experimental data ?



J. Zhang, **S. Maslov**, E. Shakhnovich,
Nature/EMBO Molecular Systems Biology (2008)

Data on binding interactions
from PINT database

Dissociation constant

- Interaction between two molecules (say, proteins) is usually described in terms of dissociation constant
  $K_{ij}=1M \exp(-E_{ij}/kT)$
- Law of Mass Action: the concentration D_{ij} of a heterodimer formed out of two proteins with free (monomer) concentrations C_i and C_j : $D_{ij}=C_iC_j/K_{ij}$
- What is the distribution of K_{ij}?
- Answer: it is called log-normal since the logarithm of K_{ij} is the binding energy $-E_{ij}/kT$ which is normally distributed

Lognormal Distribution

- Let *W* denote a normal random variable with mean of θ and variance of ω^2, i.e., $E(W) = \theta$ and $V(W) = \omega^2$
- As a change of variable, let $X = e^W = \exp(W)$ and $W = \ln(X)$
- Now X is a lognormal random variable.

$$F(x) = P[X \le x] = P[\exp(W) \le x] = P[W \le \ln(x)]$$

$$= P\left[Z \le \frac{\ln(x) - \theta}{\omega}\right] = \Phi\left[\frac{\ln(x) - \theta}{\omega}\right] = \quad \text{for} \quad x > 0$$

$$= 0 \quad \text{for} \quad x \le 0$$

$$f(x) = \frac{dF(x)}{dx} = \frac{1}{x\omega\sqrt{2\pi}} e^{-\left[\frac{\ln(x)-\theta}{2\omega}\right]^2} \qquad \text{for } 0 < x < \infty$$

$$E(X) = e^{\theta + \omega^2/2} \qquad \text{and} \qquad V(X) = e^{2\theta+\omega^2}\left(e^{\omega^2} - 1\right) \qquad (4\text{-}22)$$

Lognormal Graphs

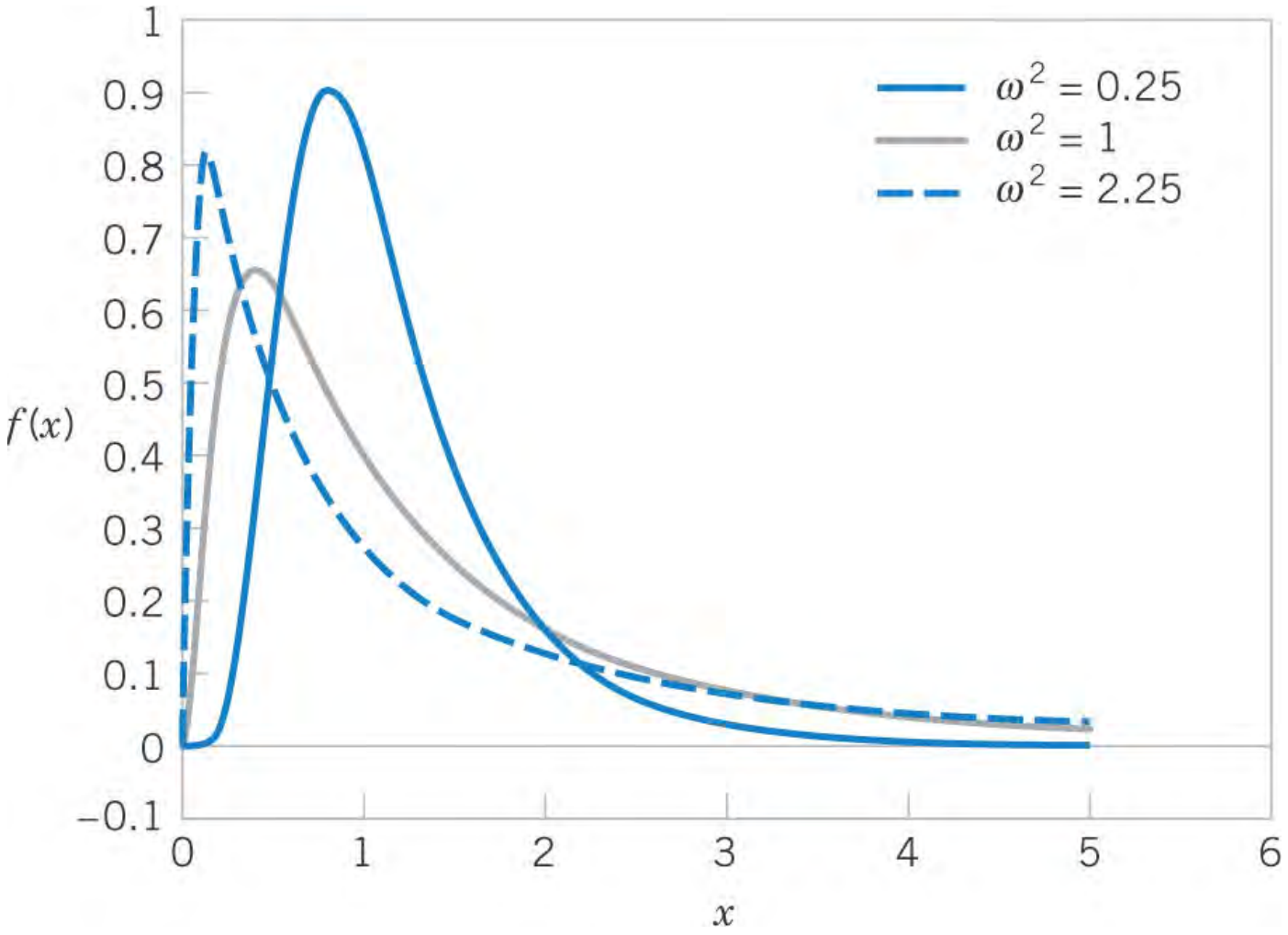


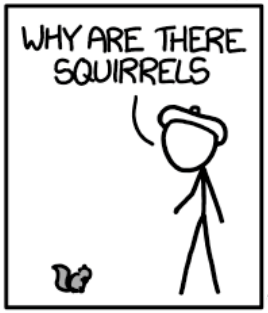
Figure 4-27 Lognormal probability density functions with θ = 0 for selected values of ω^2.

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

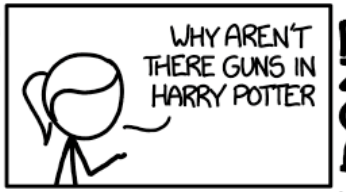
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Multiple random variables, Correlations

What we learned so far…

- Random Events:
  - Working with events as sets: union, intersection, etc.
    - Some events are simple: Head vs Tails, Cancer vs Healthy
    - Some are more complex: 10<Gene expression<100
    - Some are even more complex: Series of dice rolls: 1,3,5,3,2
  - Conditional probability: $P(A|B)=P(A \cap B)/P(B)$
  - Independent events: $P(A|B)=P(A)$ or $P(A \cap B)= P(A)*P(B)$
  - Bayes theorem: relates $P(A|B)$ to $P(B|A)$
- Random variables:
  - Mean, Variance, Standard deviation. How to work with E(g(X))
  - Discrete (Uniform, Bernoulli, Binomial, Poisson, Geometric, Negative binomial, Power law);
    PMF: f(x)=Prob(X=x); CDF: F(x)=Prob(X≤x);
  - Continuous (Uniform, Exponential, Erlang, Gamma, Normal, Log-normal);
    PDF: f(x) such that Prob(X inside A)= $\int_A f(x)dx$; CDF: F(x)=Prob(X≤x)
- Next step: work with **multiple random variables** measured together in the same series of random experiments

Concept of Joint Probabilities

- Biological systems are usually described not by a single random variable but by many random variables
- Example: The expression state of a human cell: 20,000 random variables X_i for each of its genes
- A joint probability distribution describes the behavior of several random variables
- We will start with just two random variables X and Y and generalize when necessary

Joint Probability Mass Function Defined

The joint probability mass function of the
discrete random variables X and Y,
denoted as $f_{XY}(x, y)$, satifies:

(1) $f_{XY}(x, y) = P$

(2) $f_{XY}(x, y) \geq 0$ All probabilities are non−negative

(3) $\sum_x \sum_y f_{XY}(x, y) = 1$ The sum of all probabilities is 1

Montgomery Runger 5th edition Equation (5−1)

Example 5-1: # Repeats vs. Signal Bars

You use your cell phone to check your airline reservation. It asks you to speak the name of your departure city to the voice recognition system.

- Let Y denote the number of times you have to state your departure city.
- Let X denote the number of bars of signal strength on you cell phone.

y = number of times city name is stated	x = number of bars of signal strength		
	1	2	3
1	0.01	0.02	0.25
2	0.02	0.03	0.20
3	0.02	0.10	0.05
4	0.15	0.10	0.05

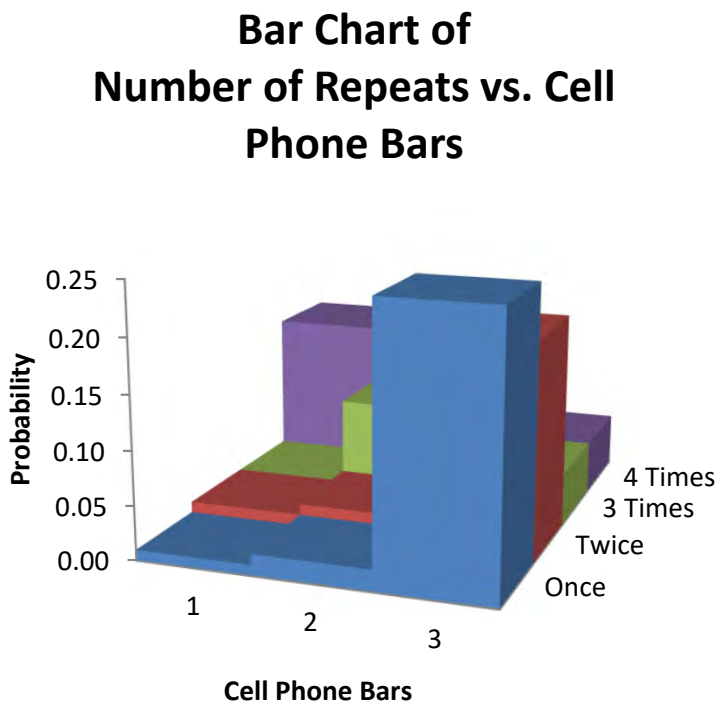


Figure 5-1 Joint probability distribution of X and Y. The table cells are the probabilities. Observe that more bars relate to less repeating.

Marginal Probability Distributions (discrete)

For a discrete joint PDF, there are marginal distributions for each random variable, formed by summing the joint PMF over the other variable.

$$f_X(x) = \sum_y f_{XY}(x,y)$$

$$f_Y(y) = \sum_x f_{XY}(x,y)$$

Called marginal because they are written in the margins

y = number of times city name is stated	x = number of bars of signal strength			
	1	2	3	$f_{Y(}y)$ =
1	0.01	0.02	0.25	0.28
2	0.02	0.03	0.20	0.25
3	0.02	0.10	0.05	0.17
4	0.15	0.10	0.05	0.30
$f_X(x)$ =	0.20	0.25	0.55	1.00

Figure 5-6 From the prior example, the joint PMF is shown in green while the two marginal PMFs are shown in purple.

Mean & Variance of X and Y are calculated using marginal distributions

y = number of times city name is stated	x = number of bars of signal strength					
	1	2	3	$f(y)=$	$y*f(y)=$	$y^2*f(y)=$
1	0.01	0.02	0.25	0.28	0.28	0.28
2	0.02	0.03	0.20	0.25	0.50	1.00
3	0.02	0.10	0.05	0.17	0.51	1.53
4	0.15	0.10	0.05	0.30	1.20	4.80
$f(x)=$	0.20	0.25	0.55	1.00	2.49	7.61
$x*f(x)=$	0.20	0.50	1.65	2.35		
$x^2*f(x)=$	0.20	1.00	4.95	6.15		

$\mu_X = E(X) = 2.35$; $\sigma_X^2 = V(X) = 6.15 - 2.35^2 = 6.15 - 5.52 = 0.6275$

$\mu_Y = E(Y) = 2.49$; $\sigma_Y^2 = V(Y) = 7.61 - 2.49^2 = 7.61 - 16.20 = 1.4099$

Conditional Probability Distributions

Recall that $P(B|A) = \frac{P(A \cap B)}{P(A)}$

$$P(Y=y|X=x)=P(X=x,Y=y)/P(X=x)=$$
$$=f(x,y)/f_X(x)$$

From Example 5-1

$P(Y=1|X=3) = 0.25/0.55 = 0.455$

$P(Y=2|X=3) = 0.20/0.55 = 0.364$

$P(Y=3|X=3) = 0.05/0.55 = 0.091$

$P(Y=4|X=3) = 0.05/0.55 = 0.091$

Sum = *1.00*

y = number of times city name is stated	x = number of bars of signal strength			
	1	2	3	$f_Y(y)$ =
1	0.01	0.02	0.25	0.28
2	0.02	0.03	0.20	0.25
3	0.02	0.10	0.05	0.17
4	0.15	0.10	0.05	0.30
$f_X(x)$ =	0.20	0.25	0.55	1.00

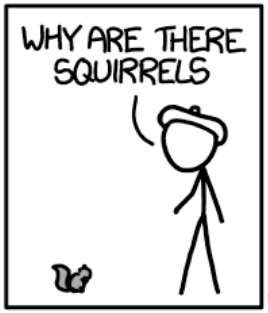
Note that there are 12 probabilities conditional on *X*, and 12 more probabilities conditional upon *Y*.

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

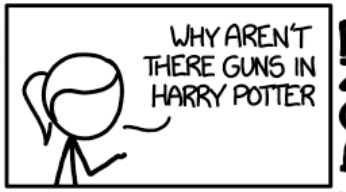
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Joint Probability Density Function Defined

The joint probability density function for the continuous random variables X and Y, denotes as $f_{XY}(x,y)$, satisfies the following properties:

(1) $f_{XY}(x,y) \geq 0$ for all x, y

(2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1$

(3) $P\big((X,Y) \subset R\big) = \iint_R f_{XY}(x,y)dxdy \quad (5-2)$

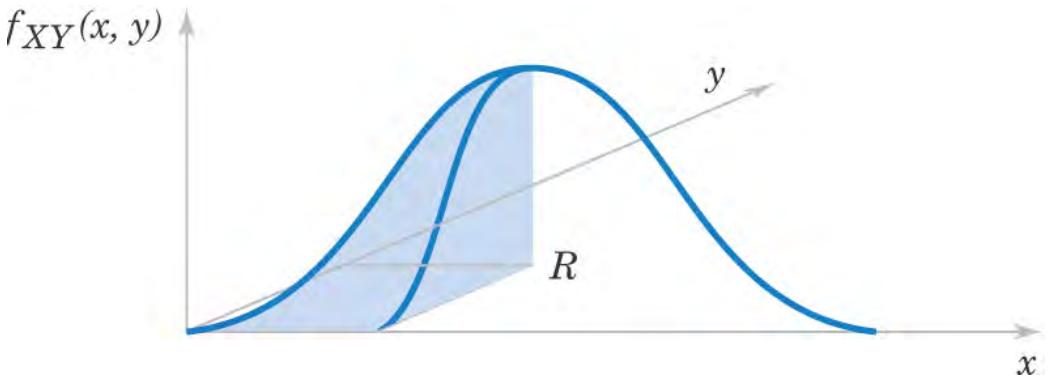


Figure 5-2 Joint probability density function for the random variables X and Y. Probability that (X, Y) is in the region R is determined by the volume of $f_{XY}(x,y)$ over the region R.

Joint Probability Density Function Graph

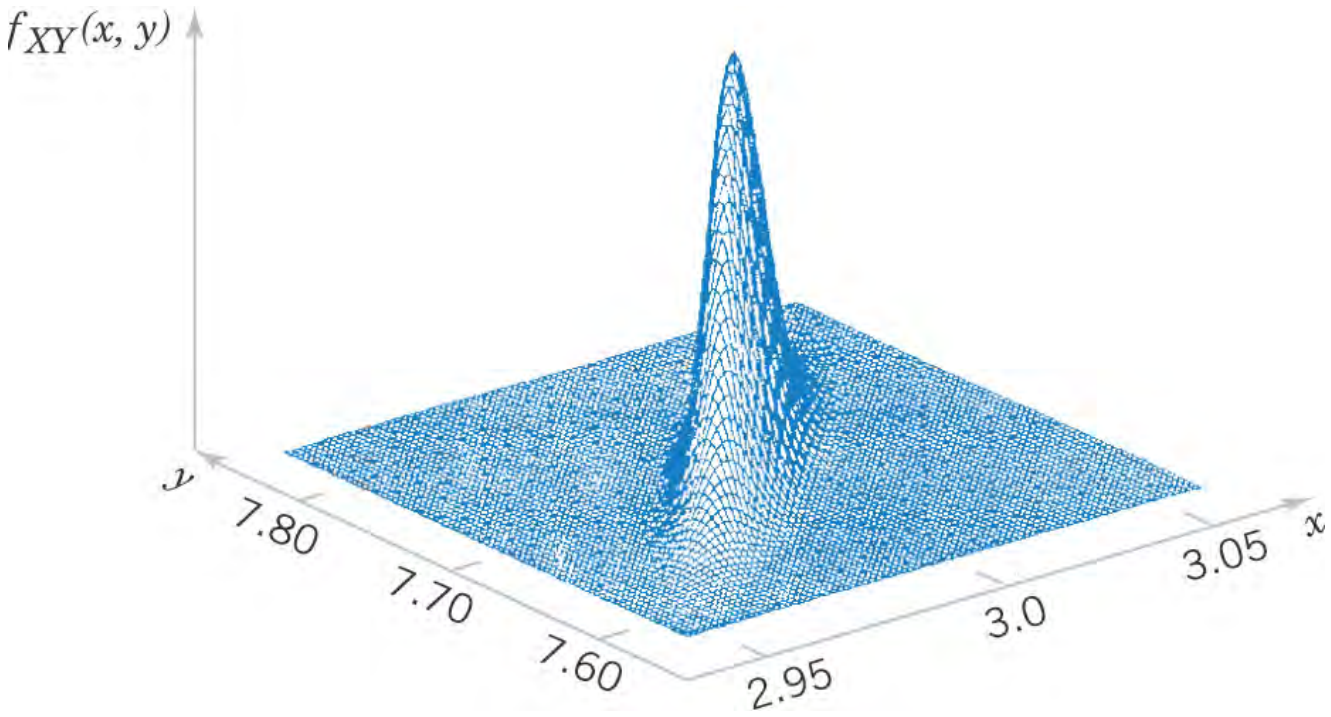


Figure 5-3 Joint probability density function for the continuous random variables *X* and *Y* of expression levels of two different genes. Note the asymmetric, narrow ridge shape of the PDF – indicating that small values in the *X* dimension are more likely to occur when small values in the *Y* dimension occur.

Marginal Probability Distributions (continuous)

- Rather than summing a discrete joint PMF, we integrate a continuous joint PDF.
- The marginal PDFs are used to make probability statements about one variable.
- If the joint probability density function of random variables X and Y is $f_{XY}(x,y)$, the marginal probability density functions of X and Y are:

$$f_X(x) = \int_y f_{XY}(x,y)\,dy$$

$$f_Y(y) = \int_x f_{XY}(x,y)\,dx \qquad (5-3)$$

$$f_X(x) = \sum_y f_{XY}(x,y)$$

$$f_Y(y) = \sum_x f_{XY}(x,y)$$

Conditional Probability Density Function Defined

Given continuous random variables X and Y with
joint probability density function $f_{XY}(x,y)$,
the conditional probability densiy function of Y given X=x is

$$f_{Y|x}(y) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{XY}(x,y)}{\int_y f_{XY}(x,y)\,dy} \text{ if } f_X(x) > 0 \qquad (5-4)$$

which satifies the following properties:

(1) $f_{Y|x}(y) \geq 0$

(2) $\int f_{Y|x}(y)dy = 1$

(3) $P(Y \subset B|X = x) = \int_B f_{Y|x}(y)dy$ for any set B in the range of Y

Compare to discrete: $P(Y=y|X=x)=f_{XY}(x,y)/f_X(x)$

Conditional Probability Distributions

- Conditional probability distributions can be developed for multiple random variables by extension of the ideas used for two random variables.
- Suppose $p = 5$ and we wish to find the distribution of X_1, X_2 and X_3 conditional on $X_4=x_4$ and $X_5=x_5$.

$$f_{X_1X_2X_3|x_4x_5}(x_1, x_2, x_3) = \frac{f_{X_1X_2X_3X_4X_5}(x_1, x_2, x_3, x_4, x_5)}{f_{X_4X_5}(x_4, x_5)}$$

for $f_{X_4X_5}(x_4, x_5) > 0$.

Independence for Continuous Random Variables

For random variables *X* and *Y*, if any one of the following properties is true, the others are also true. Then *X* and *Y* are independent.

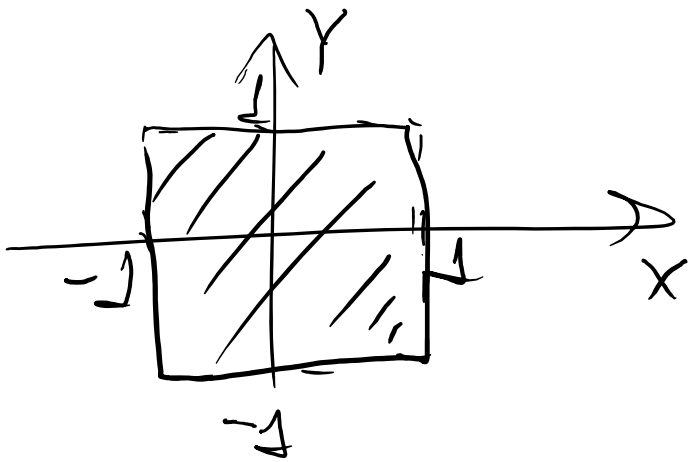
(1) $f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$
(2) $f_{Y|x}(y) = f_Y(y)$ for all x and y with $f_X(x) > 0$
(3) $f_{X|y}(y) = f_X(x)$ for all x and y with $f_Y(y) > 0$
(4) $\mathrm{P}(X \subset A, Y \subset B) = P(X \subset A) \cdot P(Y \subset B)$ for any sets A and B in the range of X and Y, respectively. (5−7)

P(Y=y|X=x)=P(Y=y) **for any *x*** or
P(X=x|Y=y)=P(X=x) **for any *y*** or
P(X=x, Y=y)=P(X=x)·P(Y=y) **for any *x* and *y***

Example 1:

Uniform distribution in the square

$-1 < x < 1, \quad -1 < y < 1$



$$\begin{cases} f_{XY}(x,y) = c \text{ if } -1 < x < 1 \text{ and } -1 < y < 1 \\ 0 - \text{otherwise} \end{cases}$$

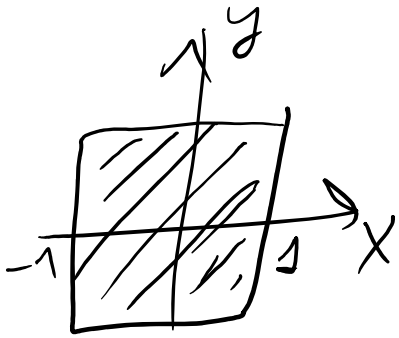
$$1 = \int_{\text{square}} dx\,dy\, f_{XY}(x,y) = c \cdot \text{Area} = c \cdot 4 \Rightarrow c = \frac{1}{4}$$

Are X and Y independent? Yes they are

Let's test if $f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dy =$$

$$= \int_{-1}^{1} \frac{1}{4}\,dy = \frac{1}{2} \quad \text{if } -1 < x < 1$$



Same for $f_Y(y) = \frac{1}{2}$ if $-1 < y < 1$

$$\frac{1}{4} = f_{XY}(x,y) = \frac{1}{2} \cdot \frac{1}{2} = f_X(x) \cdot f_Y(y)$$

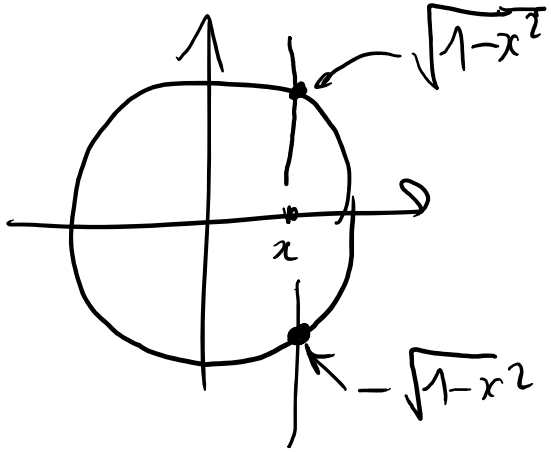
0 otherwise if both x & y are in $[-1,1]$

Joint PDF $f_{XY}(x,y) = \frac{1}{\text{area}} = \frac{1}{\pi}$ if x, y in the disc

0 – otherwise

Marginal distributions:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{dy}{\pi} = \frac{2\sqrt{1-x^2}}{\pi}$$



Same for $f_Y(y) = \frac{2\sqrt{1-y^2}}{\pi}$

$$\frac{1}{\pi} = f_{XY}(x,y) \neq \frac{2}{\pi}\sqrt{1-x^2} \cdot \frac{2}{\pi}\sqrt{1-y^2} = f_X(x) \cdot f_Y(y)$$

Variables are NOT independent

Covariation, Correlations

Quick and dirty check for **linear** (in)dependence between variables

Covariance Defined

Covariance is a number quantifying the average ***linear*** dependence between two random variables.

The covariance between the random variables X and Y, denoted as $\text{cov}(X, Y)$ or σ_{XY} is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$$

Montgomery, Runger 5th edition Eq. (5−14)

The units of σ_{XY} are the units of X times the units of Y.

Unlike the range of the variance, covariance can be negative: $-\infty < \sigma_{XY} < \infty$.

Covariance – 1 number to measure dependance between random variables

$Cov(X,Y)$ or σ_{XY}

$$\sigma_{XY} = E\left[(X-\mu_X)\cdot(Y-\mu_Y)\right] =$$

$$= E(X\cdot Y) - \mu_X \cdot \mu_Y$$

- $Var(X) = Cov(X,X)$
- If X & Y are independent

$$Cov(X,Y) = E[X-\mu_X]\cdot E[Y-\mu_Y] = 0$$

- $-\infty < Cov(X,Y) < +\infty$

Can be negative!

Covariance and PMF tables

y = number of times city name is stated	x = number of bars of signal strength		
	1	2	3
1	0.01	0.02	0.25
2	0.02	0.03	0.20
3	0.02	0.10	0.05
4	0.15	0.10	0.05

The probability distribution of Example 5-1 is shown.

By inspection, note that the larger probabilities occur as *X* and *Y* move in opposite directions. This indicates a negative covariance.

Covariance and Scatter Patterns

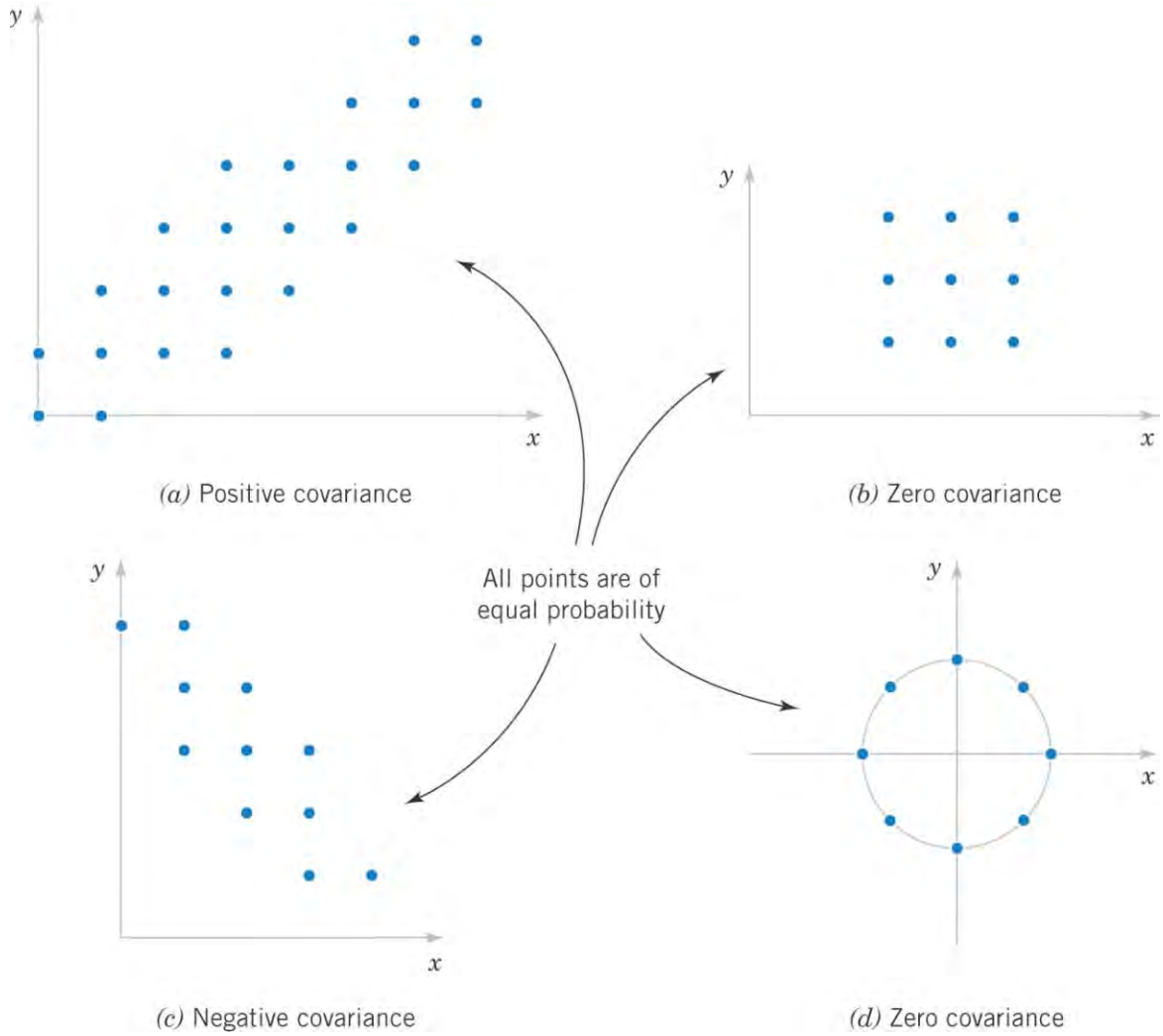


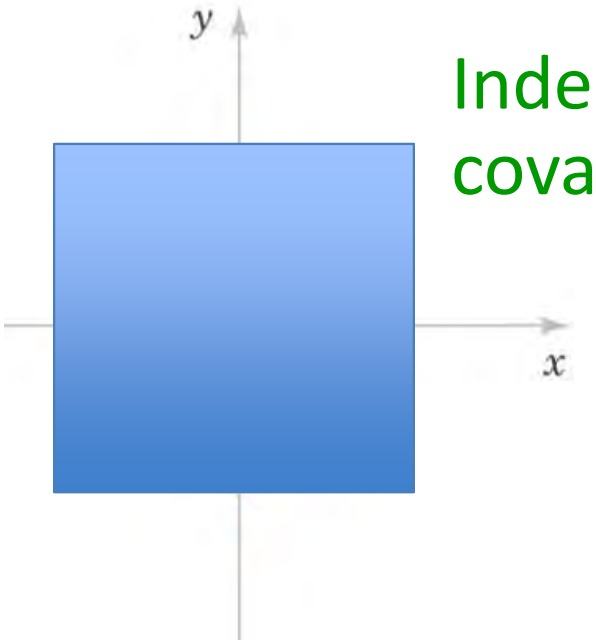
Figure 5-13 Joint probability distributions and the sign of cov(X, Y). Note that covariance is a measure of linear relationship. Variables with non-zero covariance are correlated.

Independence Implies σ=ρ = 0 but not vice versa

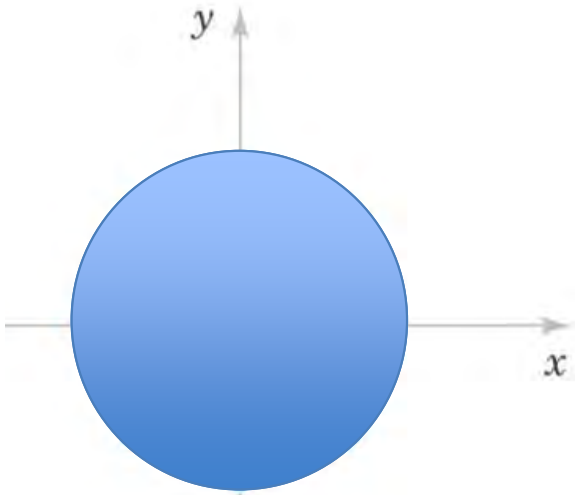
- If *X* and *Y* are independent random variables,

$$\sigma_{XY} = \rho_{XY} = 0 \qquad (5\text{-}17)$$

- $\rho_{XY} = 0$ is necessary, but not a sufficient condition for independence.



Independent
covariance=0



NOT independent
covariance=0

Correlation is “normalized covariance”

- Also called:
  Pearson correlation coefficient

$\rho_{XY}=\sigma_{XY}/\sigma_X\sigma_Y$
is the covariance normalized to be $-1 \leq \rho_{XY} \leq 1$

Karl Pearson (1852– 1936)
English mathematician and biostatistician

Prove that ρ_{XY} is in $[-1, 1]$

$$Z_X = \frac{X - \mu_X}{\sigma_X} \; ; \quad Z_Y = \frac{Y - \mu_Y}{\sigma_Y}$$

$$0 \le E\left((Z_X - Z_Y)^2\right) = E(Z_X^2) + E(Z_Y^2) -$$

$$- 2E(Z_X \cdot Z_Y) = 2 - 2\frac{1}{\sigma_X \sigma_Y} E\left((X - \mu_X)(Y - \mu_Y)\right) =$$

$$2 - 2\rho_{XY} \Longrightarrow \boxed{\rho_{XY} \le 1}$$

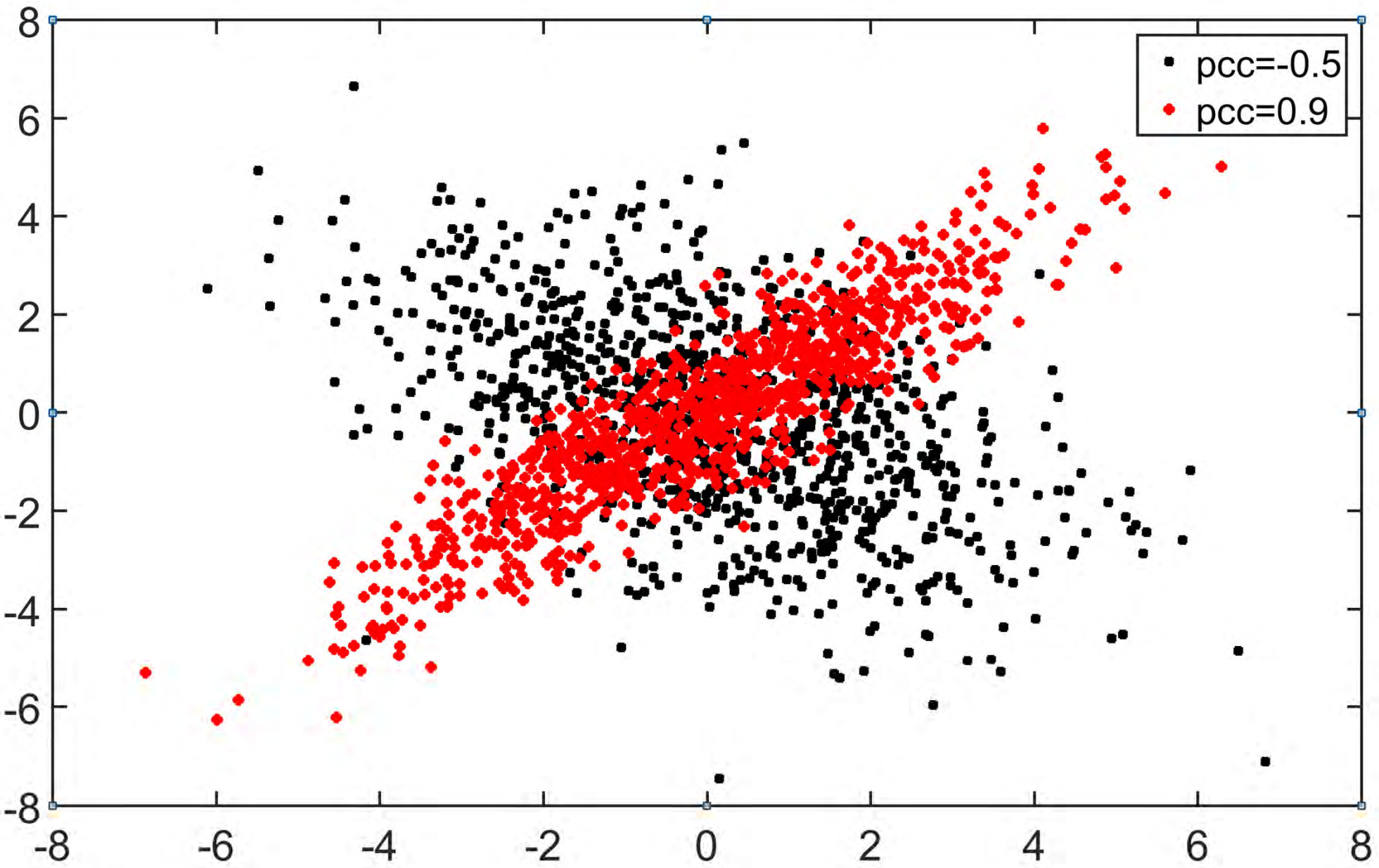
$$0 \le E\left((Z_X + Z_Y)^2\right) = E(Z_X^2) + E(Z_Y^2) +$$

$$+ 2E(Z_X \cdot Z_Y) = 2 + 2\rho_{XY} \Longrightarrow$$

$$\Longrightarrow \boxed{\rho_{XY} \ge -1}$$

Spearman rank correlation

- Pearson correlation tests for linear relationship between X and Y
- Unlikely for variables with broad distributions → non-linear effects dominate
- Spearman correlation tests for any monotonic relationship between X and Y
- Calculate ranks (1 to n), $r_X(i)$ and $r_Y(i)$ of variables in both samples. Calculate Pearson correlation between ranks: Spearman(X,Y) = Pearson(r_X, r_Y)
- Ties: convert to fractions, e.g. tie for 6s and 7s place both get 6.5. This can lead to artefacts.
- If lots of ties: use Kendall rank correlation (Kendall tau)

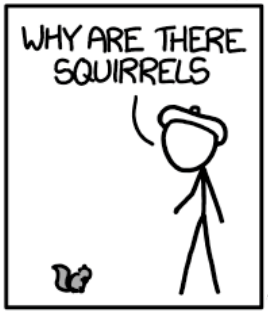

pcc=-0.5
pcc=0.9
8
6
4
2
0
-2
-4
-6
-8
-8
-6
-4
-2
0
2
4
6
8

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

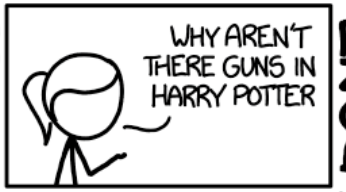
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

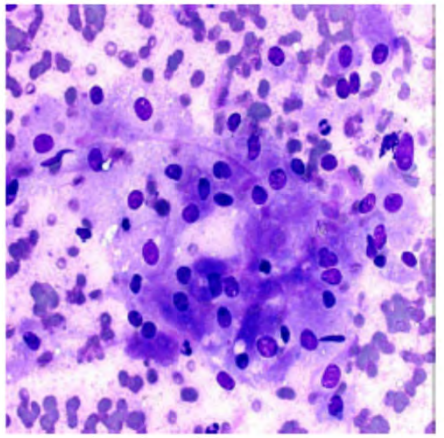
WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Let's work with real cancer data!

- Data from Wolberg, Street, and Mangasarian (1994)
- Fine-needle aspirates = biopsy for breast cancer
- Black dots – cell nuclei. Irregular shapes/sizes may mean cancer
- Statistics of all cells in the image
- 212 cancer patients and 357 healthy individuals (column 1)
- 30 other properties (see table)



Variable	Mean	S.Error	Extreme
Radius (average distance from the center)	Col 2	Col 12	Col 22
Texture (standard deviation of gray-scale values)	Col 3	Col 13	Col 23
Perimeter	Col 4	Col 14	Col 24
Area	Col 5	Col 15	Col 25
Smoothness (local variation in radius lengths)	Col 6	Col 16	Col 26
Compactness ($perimeter^2$ / area - 1.0)	Col 7	Col 17	Col 27
Concavity (severity of concave portions of the contour)	Col 8	Col 18	Col 28
Concave points (number of concave portions of the contour)	Col 9	Col 19	Col 29
Symmetry	Col 10	Col 20	Col 30
Fractal dimension ("coastline approximation" - 1)	Col 11	Col 21	Col 31

Matlab exercise #2

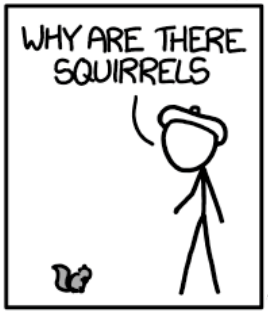
- Download cancer data in cancer_wdbc.mat
- Data in the table cancerwdbc (569x30). First 357 patients are healthy. The remaining 569-357=212 patients have cancer.
- Make scatter plots of area vs perimeter and texture vs radius.
- Calculate Pearson and Spearman correlations
- Calculate the correlation matrix of all-against-all variables: there are 30*29/2=435 correlations. Hint: corr_mat=corr(cancerwdbc);
- Plot the histogram of these 435 correlation coefficients. Hint: use [i,j,v]=find(corr_mat); then find all i>j and analyze v evaluated on this subset of 435 matrix elements

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

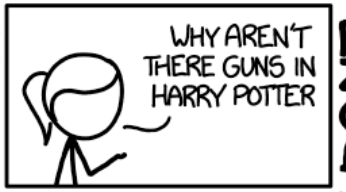
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

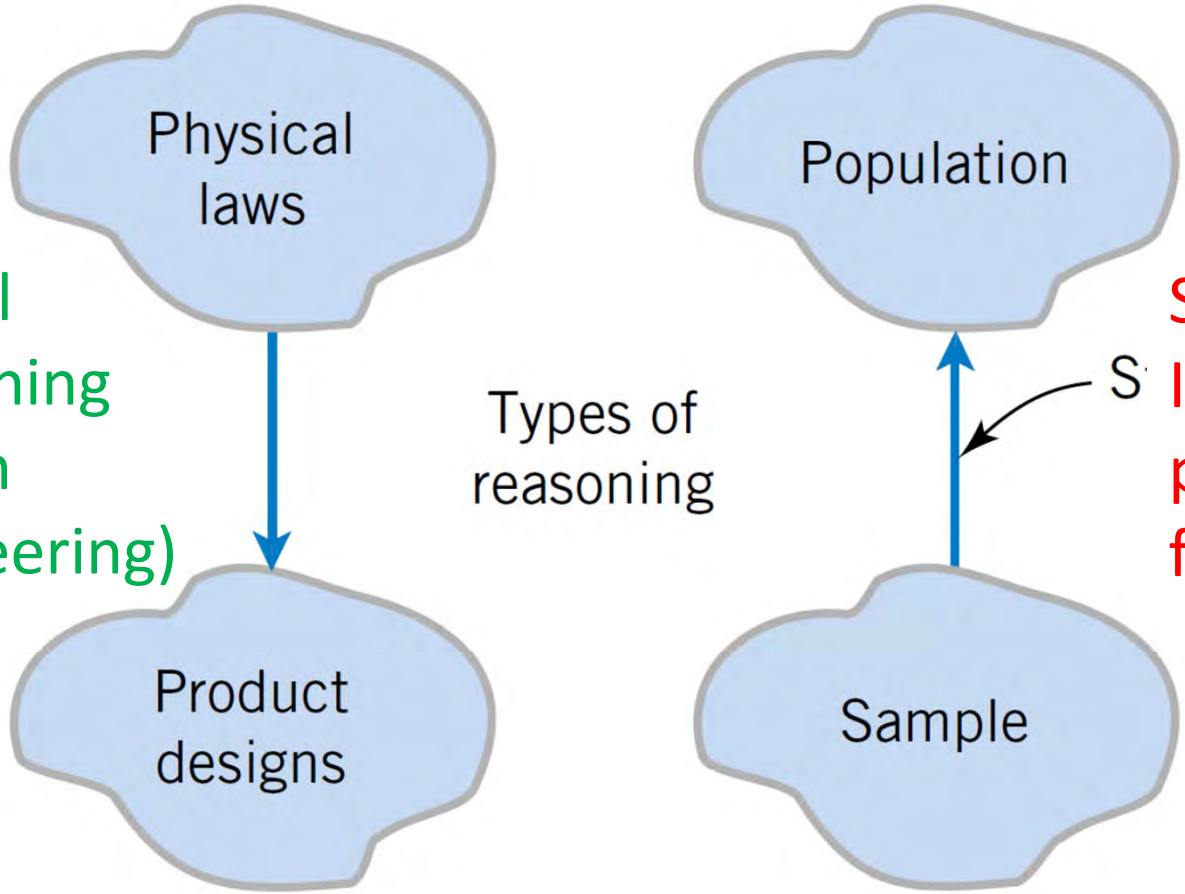
Descriptive statistics:

Populations, Samples

Histograms, Quartiles

Sample mean and variance

Two types of reasoning



Logical Reasoning (e.g. in Engineering)

Statistical reasoning: Inference of population properties from a finite sample

Numerical Summaries of Data

- Data are the numerical observations of a phenomenon of interest.
- The totality of all observations is a population.
  - Population can be infinite (e.g. abstract random variables)
  - It can be very large (e.g. 7 billion humans or all patients who have cancer of a given type)
- A (usually small) portion of the population collected for analysis is a random sample.
- We want to use sample to infer facts about populations
- The inference is not perfect but gets better and better as sample size increases.

Some Definitions

- The random variables $X_1, X_2,\ldots,X_n$ are a random sample of size n if:
  a) The X_i are independent random variables.
  b) Every X_i has the same probability distribution.

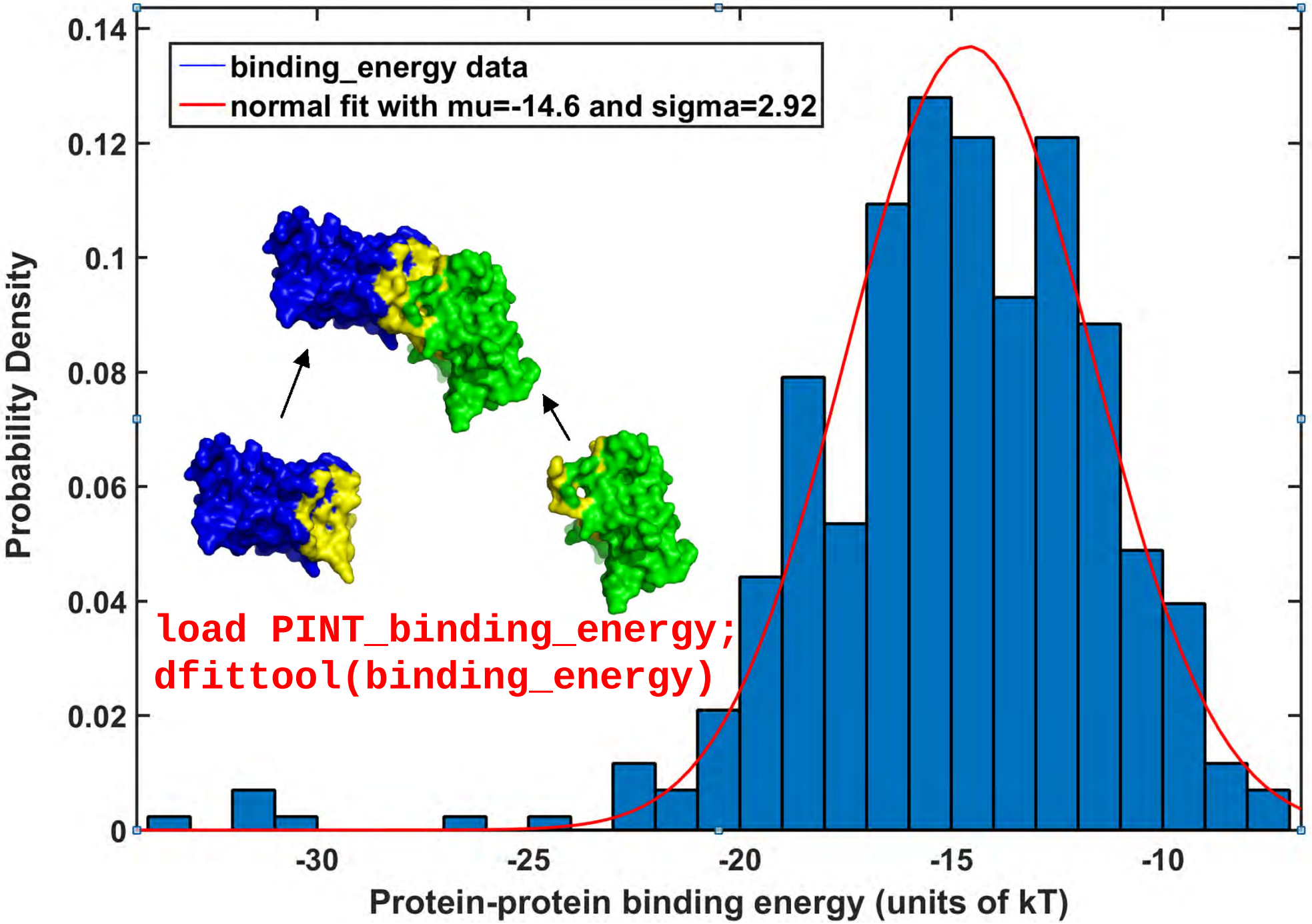
- Such $X_1, X_2,\ldots,X_n$ are also called independent and identically distributed (or i. i. d.) random variables

Ways to describe a sample:

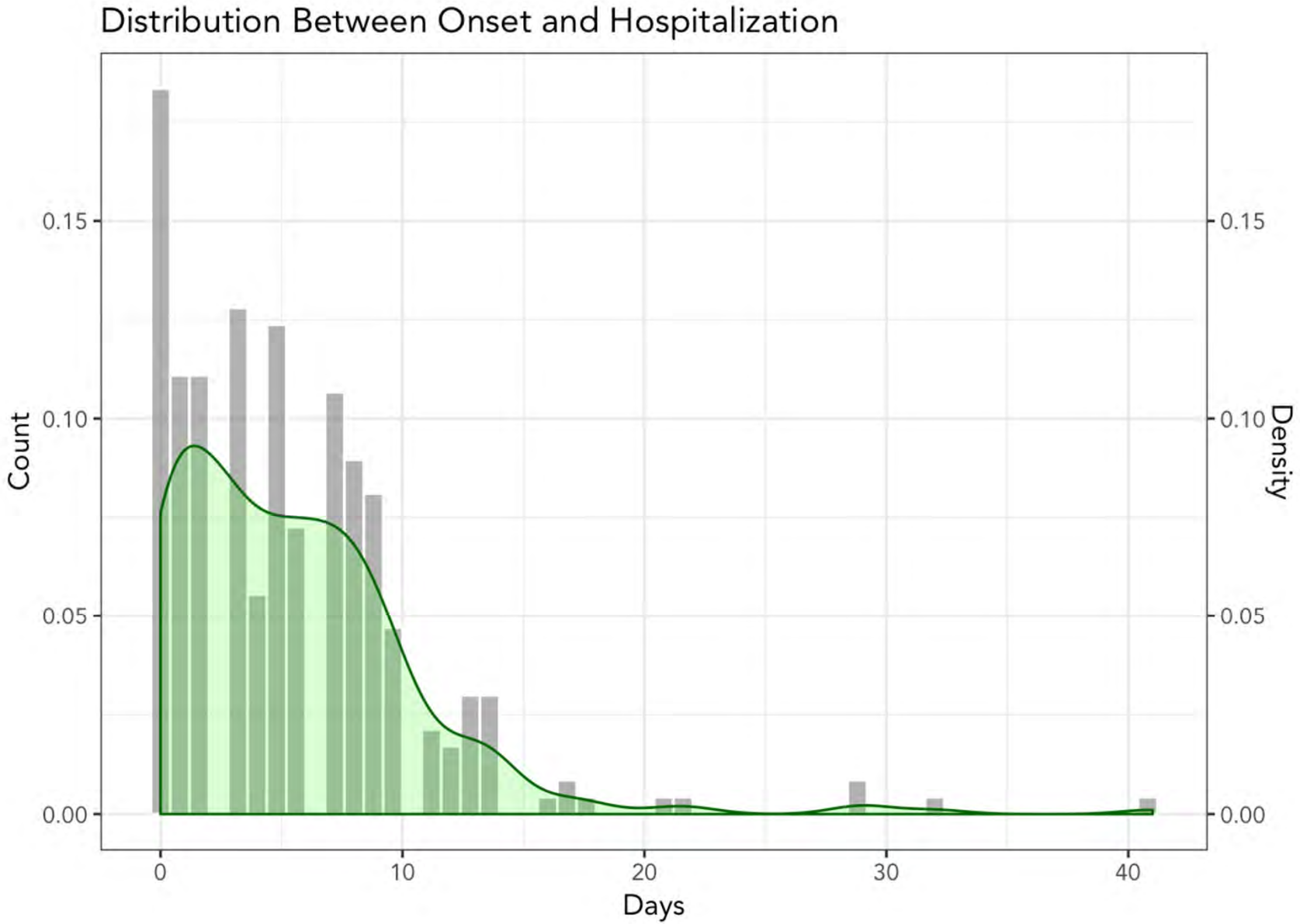
Histogram

approximates PDF

(or PMF)


binding_energy data
normal fit with mu=-14.6 and sigma=2.92
load PINT_binding_energy;
dfittool(binding_energy)
Probability Density
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-30
-25
-20
-15
-10
Protein-protein binding energy (units of kT)

PDF of time between COVID-19 symptoms onset and hospitalization in IL, April 2020

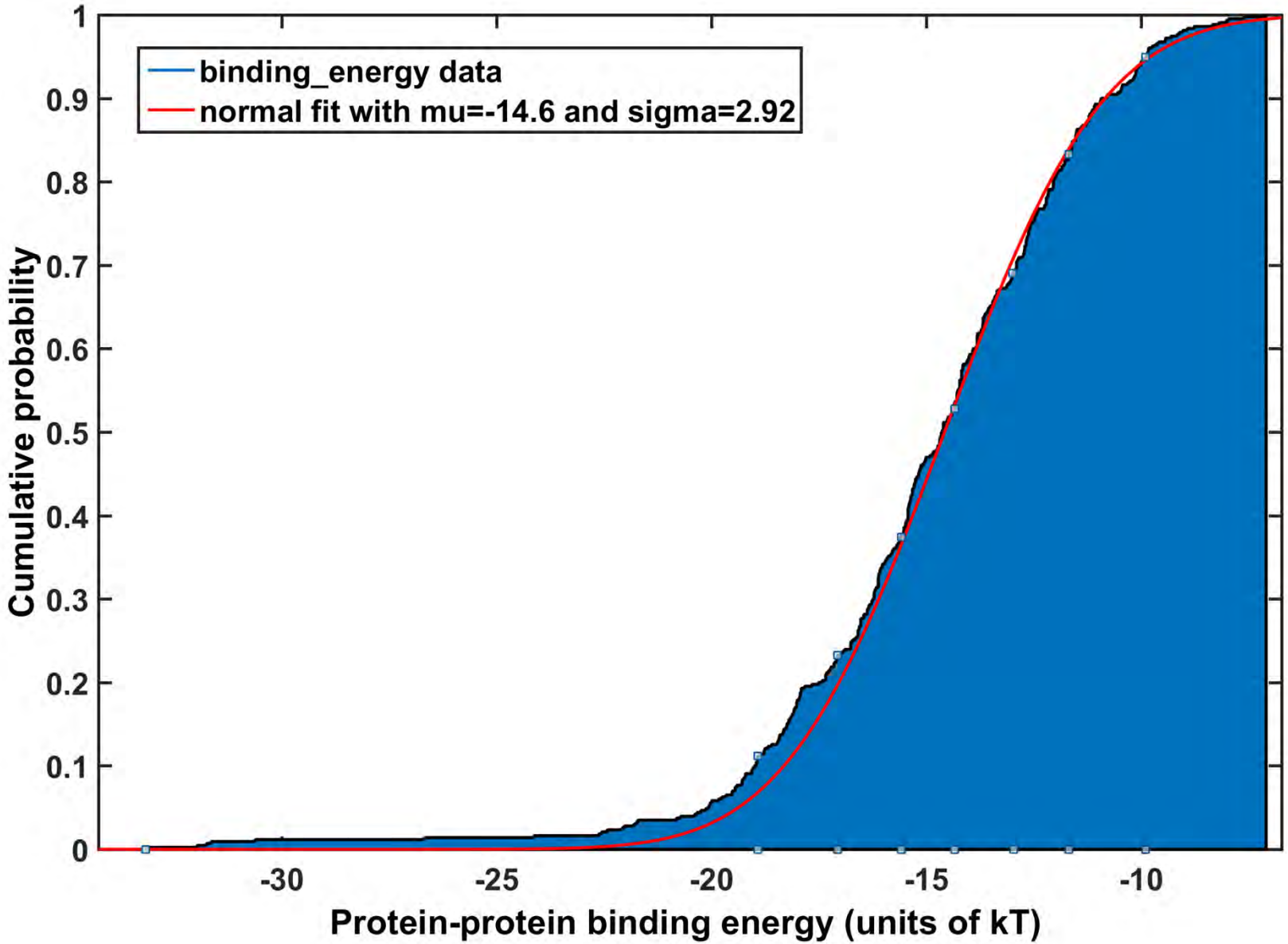


Histograms with Unequal Bin Widths

- If the data is tightly clustered in some regions and scattered in others, it is visually helpful to use narrow bin widths in the clustered region and wide bin widths in the scattered areas.
- To approximate the PDF, the rectangle area, not the height, must be proportional to the bin relative frequency.

$$\text{Rectangle height} = \frac{\text{bin relative frequency}}{\text{bin width}}$$

Cumulative Frequency Plot



Median, Quartiles, Percentiles

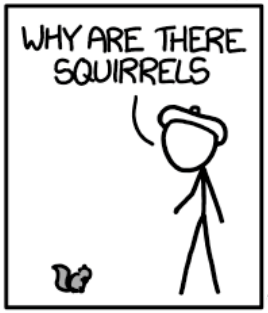
- The median q_2 divides the sample into two equal parts: 50% (n/2) of sample points below q_2 and 50% (n/2) points above q_2
- The three quartiles partition the data into four equally sized counts or segments.
  - 25% of the data is less than q_1.
  - 50% of the data is less than q_2, the median.
  - 75% of the data is less than q_3.
- There are 100 percentiles. n-th percentile p_n is defined so that n% of the data is less than p_n

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

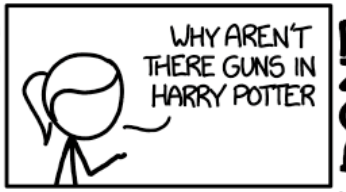
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Box-and-Whisker Plot (or better use Cat-and-Whiskers plots)

- A box plot is a graphical display showing Spread, Outliers, Center, and Shape (SOCS).
- It displays the 5-number summary: *min*, q_1, *median*, q_3, and *max*.

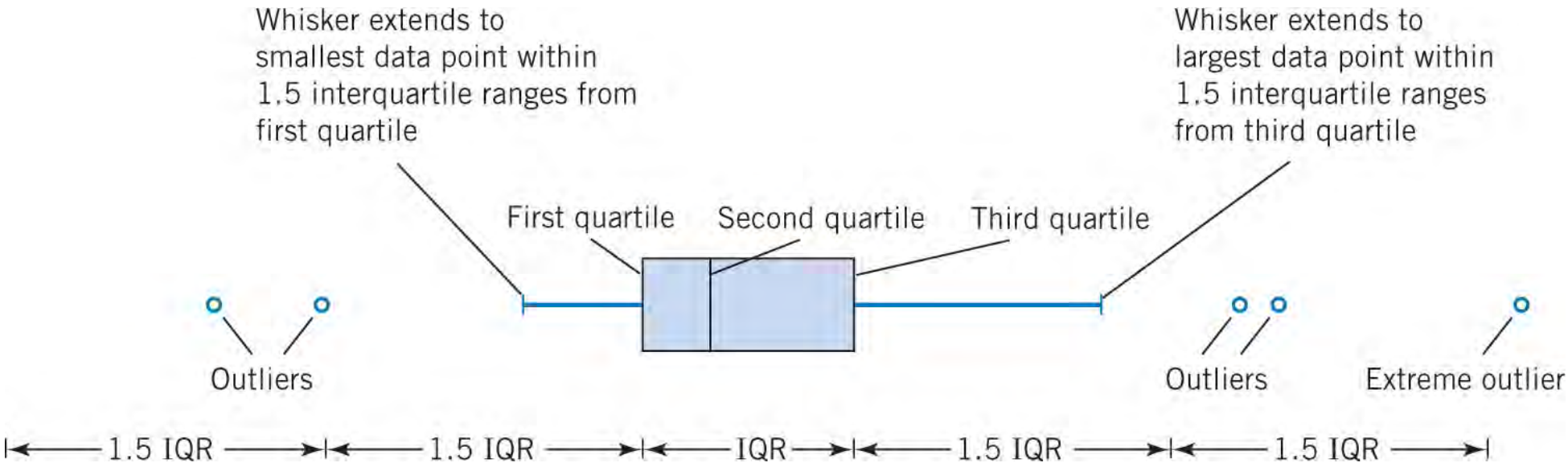
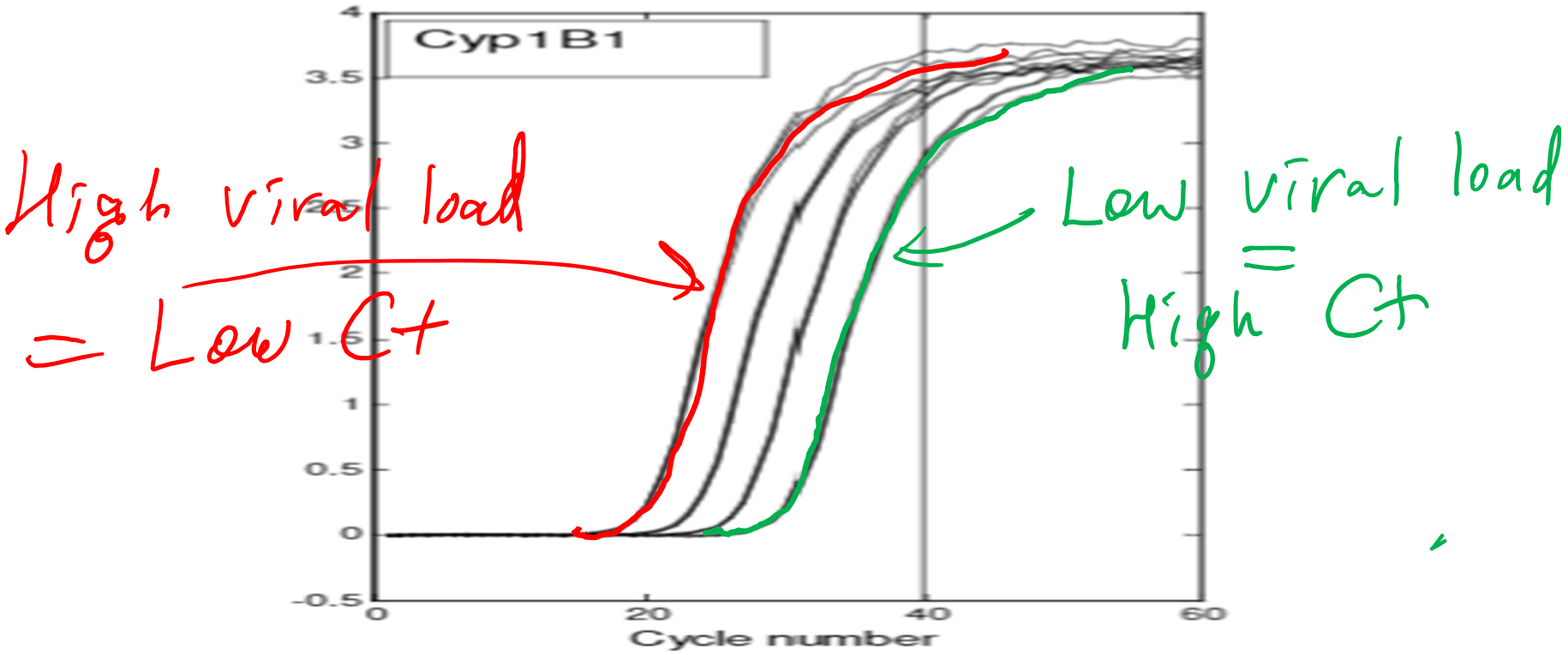


Figure 6-13 Description of a box plot.

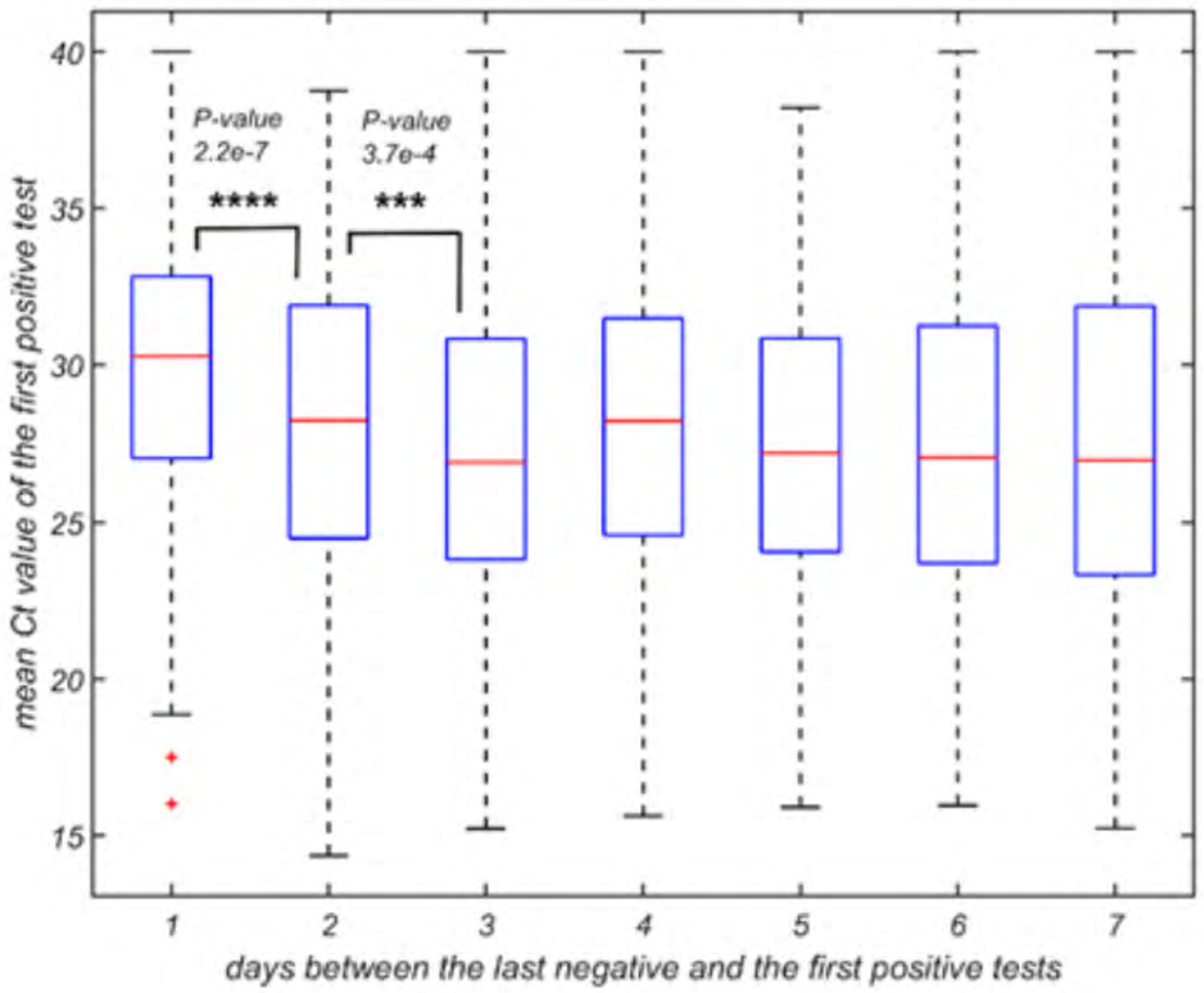
Reminder

What is the Cycle threshold (Ct) value of a PCR test?

Ct = const – log2(viral DNA concentration)



Bar plot based on COVID-19 tests at UIUC



Mitigation of SARS-CoV-2 Transmission at a Large Public University
Diana Rose E. Ranoa, et al. , medRxiv 2021 https://doi.org/10.1101/2021.08.03.21261548

Midterm will be held
here in class
this Tuesday 11/07
during regular class hours
12pm-1:50pm

Midterm Info

- Closed book exam; no books, notes, laptops, phones…
- Calculators (not on smartphones) can be used
- You can prepare one 2-sided cheat sheet
- The following two printouts will be provided

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

Name	Probability Distribution	Mean	Variance	Section in Book
Discrete				
Uniform	$\frac{1}{n}, a \le b$	$\frac{(b+a)}{2}$	$\frac{(b-a+1)^2-1}{12}$	3-5
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$, $x = 0, 1, \ldots, n, 0 \le p \le 1$	np	$np(1-p)$	3-6
Geometric	$(1-p)^{x-1}p$, $x = 1, 2, \ldots, 0 \le p \le 1$	$1/p$	$(1-p)/p^2$	3-7.1
Negative binomial	$\binom{x-1}{r-1}(1-p)^{x-r}p^r$ $x = r, r+1, r+2, \ldots, 0 \le p \le 1$	r/p	$r(1-p)/p^2$	3-7.2
Hypergeometric	$\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$ $x = \max(0, n-N+K), 1, \ldots$ $\min(K, n), K \le N, n \le N$	np, where $p = \frac{K}{N}$	$np(1-p)\left(\frac{N-n}{N-1}\right)$	3-8
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, 0 < \lambda$	λ	λ	3-9
Continuous				
Uniform	$\frac{1}{b-a}, a \le x \le b$	$\frac{(b+a)}{2}$	$\frac{(b-a)^2}{12}$	4-5
Normal	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $-\infty < x < \infty, -\infty < \mu < \infty, 0 < \sigma$	μ	σ^2	4-6
Exponential	$\lambda e^{-\lambda x}, 0 \le x, 0 < \lambda$	$1/\lambda$	$1/\lambda^2$	4-8
Erlang	$\frac{\lambda^r x^{r-1}e^{-\lambda x}}{(r-1)!}, 0 < x, r = 1, 2, \ldots$	r/λ	r/λ^2	4-9.1
Gamma	$\frac{\lambda^r x^{r-1}e^{-\lambda x}}{\Gamma(r)}, 0 < x, 0 < r, 0 < \lambda$	r/λ	r/λ^2	4-9.2

What is included in the midterm?

- Probability of events (set operations), Multiplication rules. Combinatorics
- Bayes Theorem
- Discrete Random Variables
- Continuous Random Variables
- Other topics covered
  (see HW1-HW2 for inspiration)
- No joint probabilities, correlation and covariation
- No Matlab exercises (since no computers)

Probability Multiplication Rules Combinatorics

Mr. Jones has 6 different books that he is going to put on his bookshelf. Of these, 3 are chemistry books, 2 are physics books, and 1 is a mathematics book. Jones wants to arrange his books so that two conditions are met:

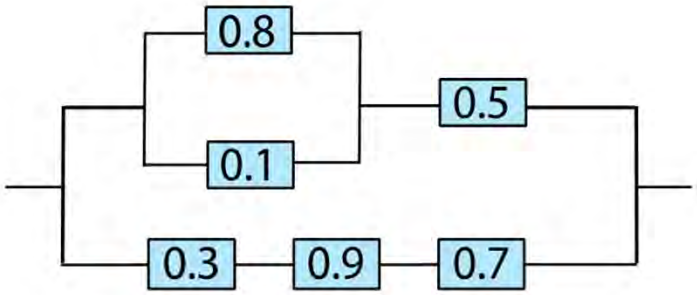
(1) all the books dealing with the same subject are together on the shelf

AND

(2) all chemistry books are on the leftmost side.

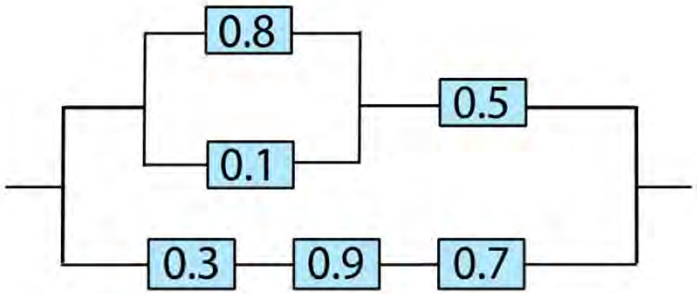
How many such different arrangements are possible?

4. **(4 points)** The following circuit operates if and only if there is a path of functional devices from left to right. The probability that each device functions is as shown. Assume that the probability that a device is functional does not depend on whether or not other devices are functional. What is the probability that the circuit operates?



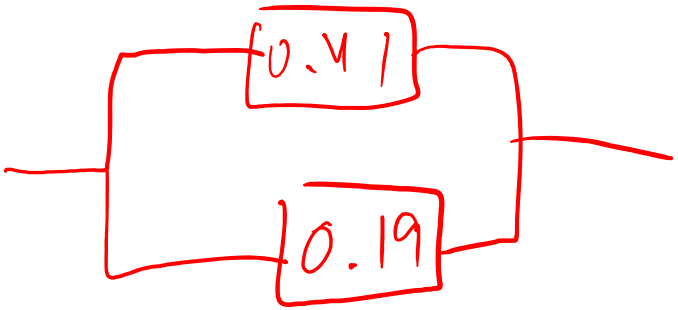
4. **(4 points)** The following circuit operates if and only if there is a path of functional devices from left to right. The probability that each device functions is as shown. Assume that the probability that a device is functional does not depend on whether or not other devices are functional. What is the probability that the circuit operates?

$1-(1-0.8)\cdot(1-0.1)=0.82$



$0.3\cdot 0.9\cdot 0.7=0.19$

$0.82\times 0.5=0.41$



$1-(1-0.41)\cdot(1-0.19)=0.52$

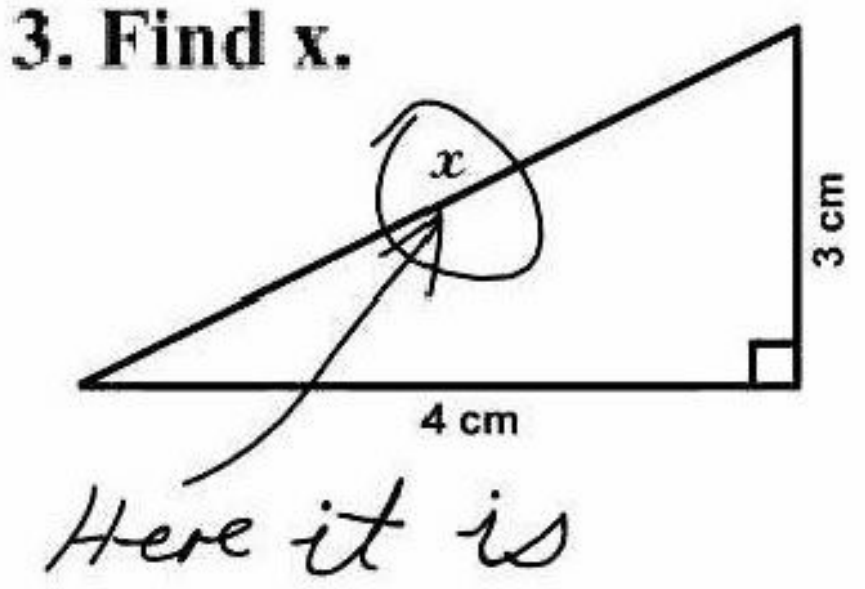
Bayes theorem

(10 points) Suppose that a bag contains ten coins, three of which are fair, while the remaining seven are biased: they have probability of 0.6 of heads when flipped. A coin was taken at random from the bag and flipped five times. All five flips gave heads. What's the probability that this coin is fair?

Discrete Probability Distributions

What is X in this problem?

- What is the random variable: Look for keywords:
  - Find the probability that….
  - What is the mean (or variance) of…
- What are parameters? Look for keywords:
  - Given that…
  - Assuming that…



Guide to probability distributions

- Binomial: # of samples, n, is fixed, # of successes, x, is variable

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

- Geometric: # of samples, x is variable. # of successes 1 is fixed. Success comes in the end

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

- Negative binomial: # of samples, x is variable. # of successes, r, is fixed rth success in ~~the~~ end

$$P(X=x) = \frac{(x-1)!}{(r-1)!(x-r)!} p^r (1-p)^{x-r}$$

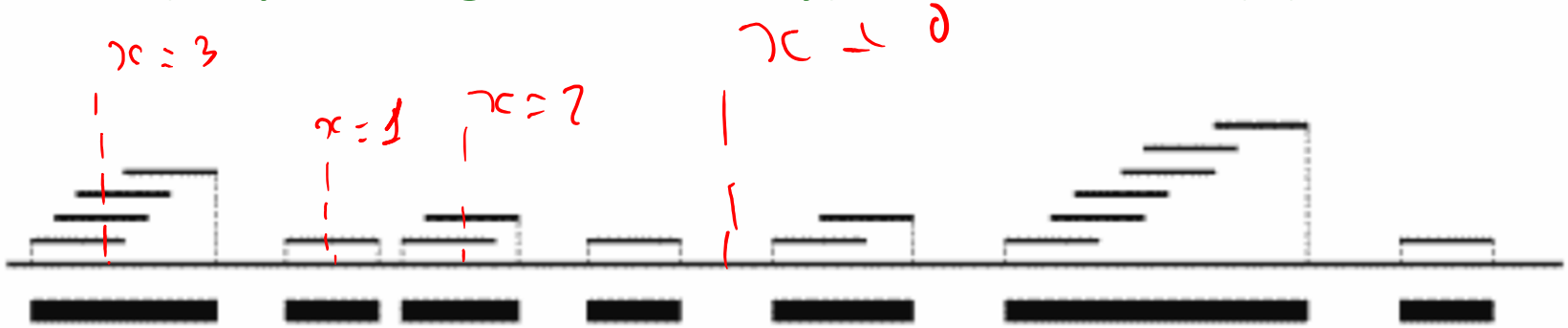
Poisson distribution in genomics

- *G - genome length (in bp)*
- *L - short read average length*
- *N – number of short read sequenced*
- *λ – sequencing redundancy = LN/G*
- *x- number of short reads covering a given site on the genome*

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

Ewens, Grant, Chapter 5.1

Poisson as a limit of Binomial. For a given site on the genome for each short read Prob(site covered): p=L/G is very small. Number of attempts (short reads): N is very large. Their product (sequencing redundancy): λ = NL/G is O(1).



Probability that a base pair in the genome is not covered by any short reads is 0.1

One randomly selects base pairs until exactly 5 uncovered base pairs are found.

Which discrete probability distribution describes the number of attempts?

A. Poisson
B. Binomial
C. Geometric
D. Negative Binomial
E. I have no idea

Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, 0 < \lambda$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$
	$x = 0, 1, \ldots, n, 0 \le p \le 1$
Geometric	$(1-p)^{x-1} p$
	$x = 1, 2, \ldots, 0 \le p \le 1$
Negative binomial	$\binom{x-1}{r-1}(1-p)^{x-r} p^r$
	$x = r, r+1, r+2, \ldots, 0 \le p \le 1$

Get your i-clickers

Probability that a base pair in the genome is not covered by any short reads is 0.1

One randomly selects base pairs until exactly 5 uncovered base pairs are found.

What are the values of p, r?

A. p=0.5, r=5
B. p=0.1, r=0.5
C. p=0.1, r=5
D. p=0.5, r=0.1
E. I have no idea

Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, 0 < \lambda$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$
	$x = 0, 1, \ldots, n, 0 \le p \le 1$
Geometric	$(1-p)^{x-1} p$
	$x = 1, 2, \ldots, 0 \le p \le 1$
Negative binomial	$\binom{x-1}{r-1}(1-p)^{x-r} p^r$
	$x = r, r+1, r+2, \ldots, 0 \le p \le 1$

Get your i-clickers

Cancer happens when the gene p53 mutates.
Probability of p53 to mutate per year is 5%.

How many years before a patient gets disease?

Which discrete probability distribution would you use to answer?

A. Poisson
B. Binomial
C. Geometric
D. Negative Binomial
E. I have no idea

Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, 0 < \lambda$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$
	$x = 0, 1, \ldots, n, 0 \le p \le 1$
Geometric	$(1-p)^{x-1} p$
	$x = 1, 2, \ldots, 0 \le p \le 1$
Negative binomial	$\binom{x-1}{r-1}(1-p)^{x-r} p^r$
	$x = r, r+1, r+2, \ldots, 0 \le p \le 1$

Get your i-clickers

Continuous Probability Distributions

1. **(8 points)** The expression level of a *TP53* tumor suppressor gene in a randomly selected cell is normally distributed with mean $\mu = 20$, and standard deviation $\sigma = 8$.

   (A)(4 points) What is the probability that the expression level in a given cell will be between 24 and 16?

   (B)(4 points) How many cells does one have to sample (on average) until there will be exactly 2 cells with such "close to average" *TP53* expression?

I can show you how to solve any HW1-HW2 problem.

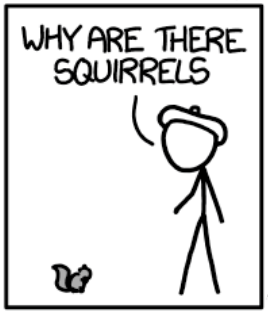
Which one do you choose?

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

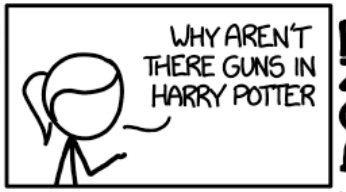
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Box-and-Whisker Plot

- A box plot is a graphical display showing Spread, Outliers, Center, and Shape (SOCS).
- It displays the 5-number summary: *min*, q_1, *median*, q_3, and *max*.

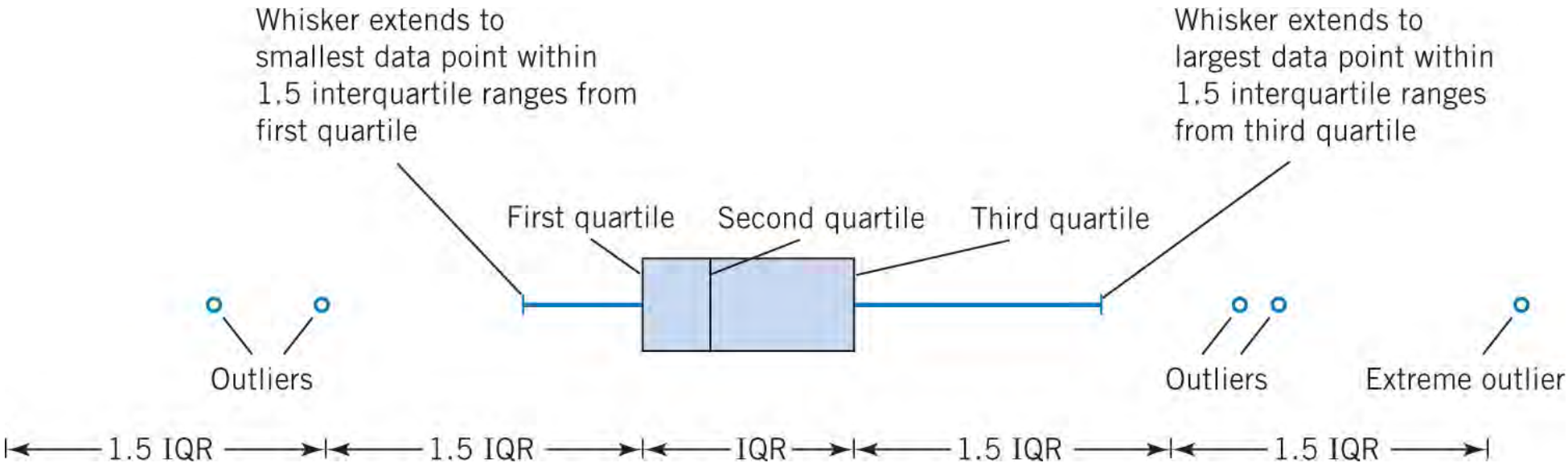
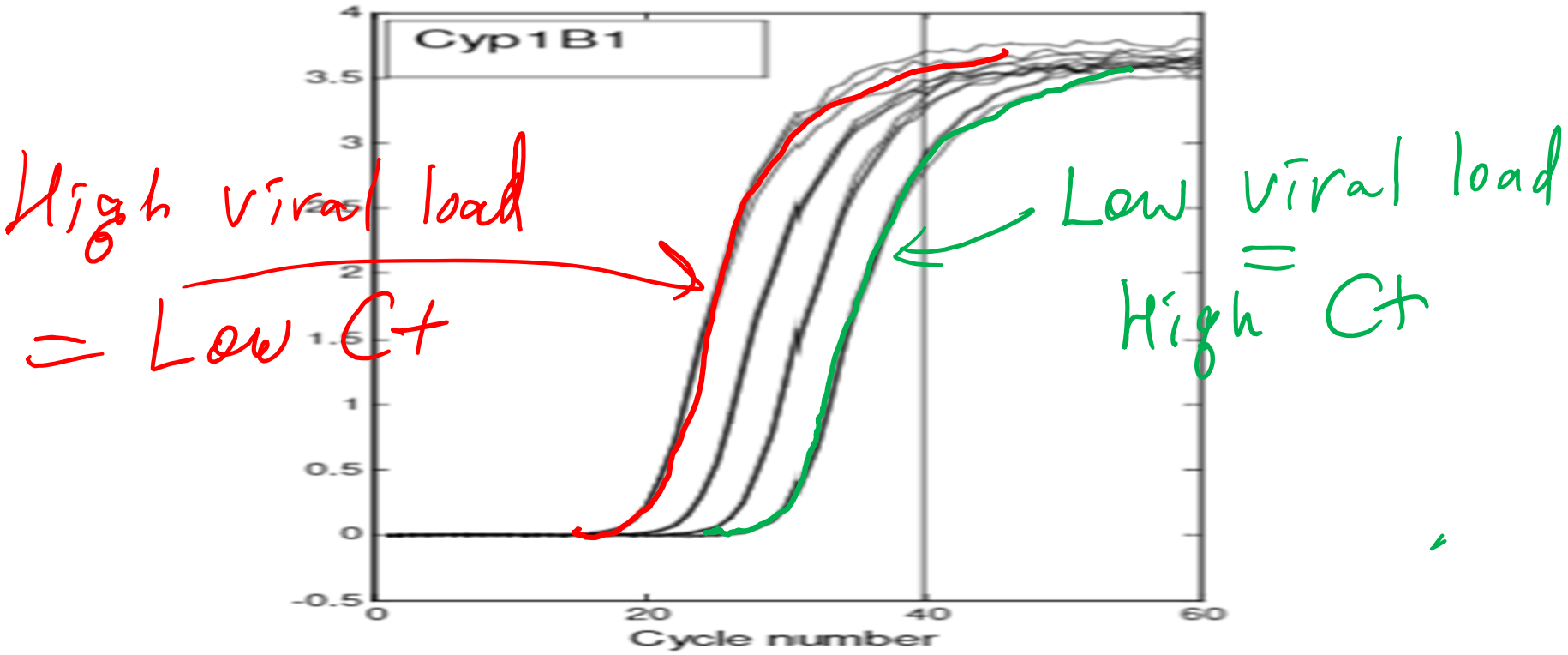


Figure 6-13 Description of a box plot.

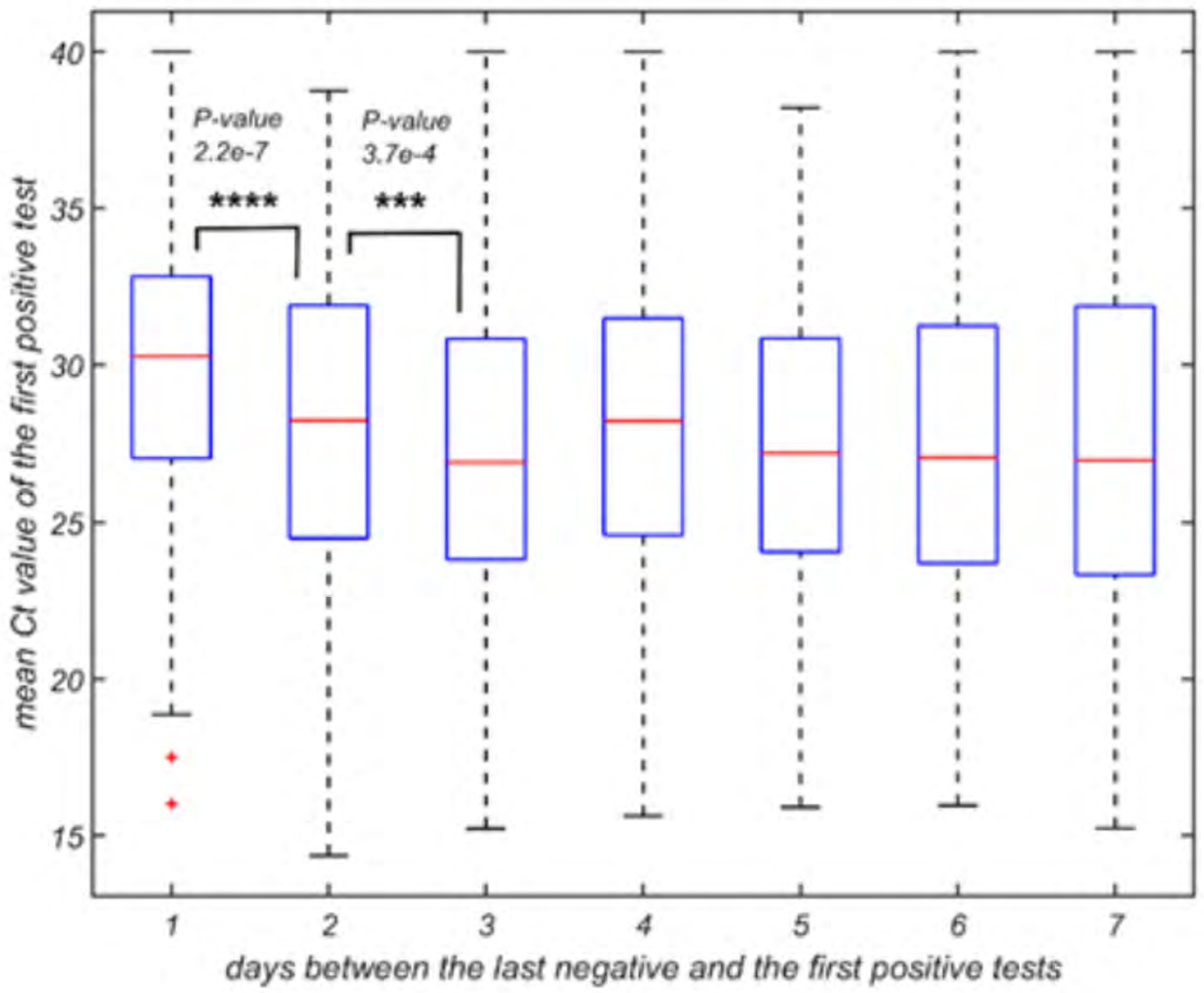
Reminder

What is the Cycle threshold (Ct) value of a PCR test?

Ct = const – log2(viral DNA concentration)



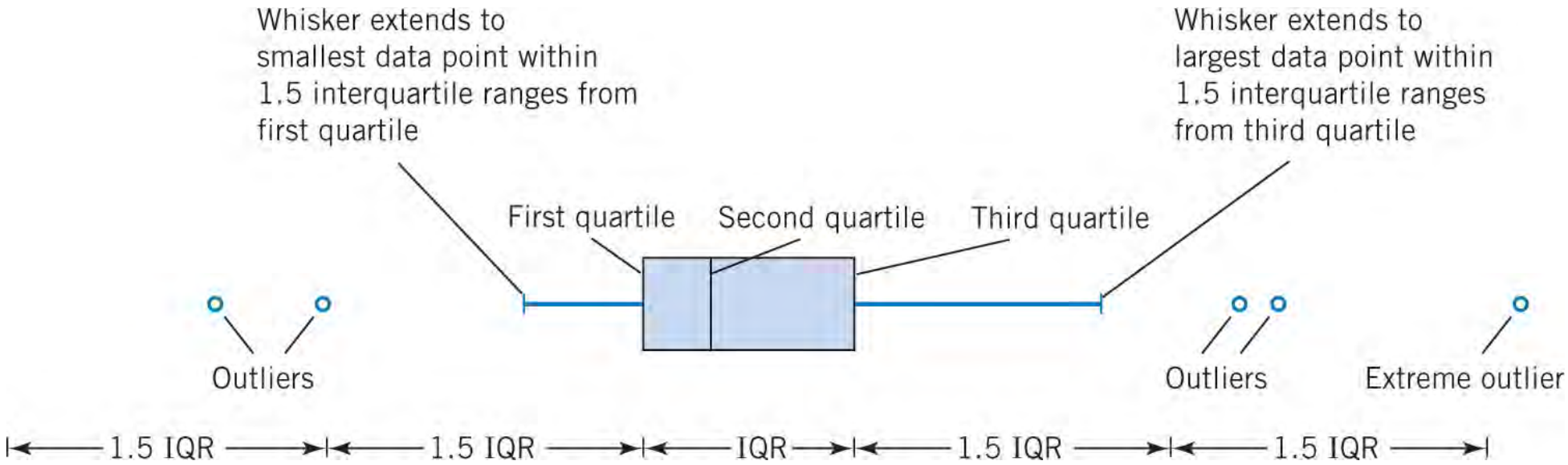
Bar plot based on COVID-19 tests at UIUC



Ranoa, D. R. E. et al. Mitigation of SARS-CoV-2 transmission at a large public university. Nat Commun 13, 3207 (2022)

Matlab exercise #2:

- Generate a sample with n= 1000 following standard normal distribution
- Calculate median, first, and third quartiles
- Calculate IQR and find ranges shown below
- Find and count left and right outliers
- Do not use built-in Matlab functions for this!
- Make box and whisker plot: use boxplot

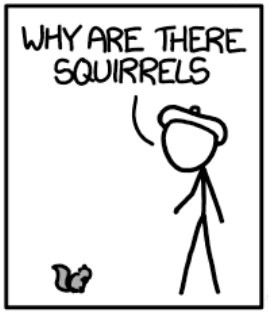


Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

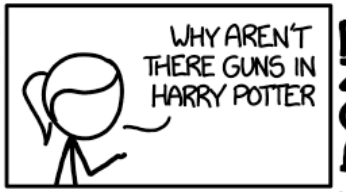
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Descriptive statistics:

Sample mean and its variance

Standard error vs Standard deviation

Some Definitions

- The random variables $X_1, X_2,\ldots,X_n$ are a random sample of size n if:
  a) The X_i are independent random variables.
  b) Every X_i has the same probability distribution.

  Such $X_1, X_2,\ldots,X_n$ are also called independent and identically distributed (or i. i. d.) random variables
- A **statistic** is any function of the observations in a random sample.
- The probability distribution of a statistic is called a **sampling distribution**.

Statistic #1: Sample Mean

If the values of n observations in a random sample are denoted by $x_1, x_2, \ldots, x_n$, the sample mean is

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (6-1)$$

New random variable $\overline{X}$ is a linear combination of *n* independent identically distributed variables $X_1, X_2, \ldots, X_n$

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Mean & Variance of a Linear Function

$$Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p$$

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \ldots + c_pE(X_p) \quad (5\text{-}25)$$

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) + 2\sum_{i<j}\sum c_ic_j \operatorname{cov}(X_iX_j) \quad (5\text{-}26)$$

If $X_1, X_2, \ldots, X_p$ are independent, then $\operatorname{cov}(X_iX_j) = 0$,

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) \quad (5\text{-}27)$$

IMPORTANT:

Sample mean $\bar{x}$ is drawn from a random variable

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$$E(\bar{X}) = \frac{n \cdot E(X_i)}{n} = \frac{n \cdot \mu}{n} = \mu$$

$$V(\bar{X}) = \frac{n \cdot V(X_i)}{n^2} = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\text{Stand. dev.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample of size n is taken from a population with mean μ and finite variance σ^2, and **any distribution**. If $\bar{X}$ is the sample mean, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (7-1)$$

for large n, is the standard normal distribution.

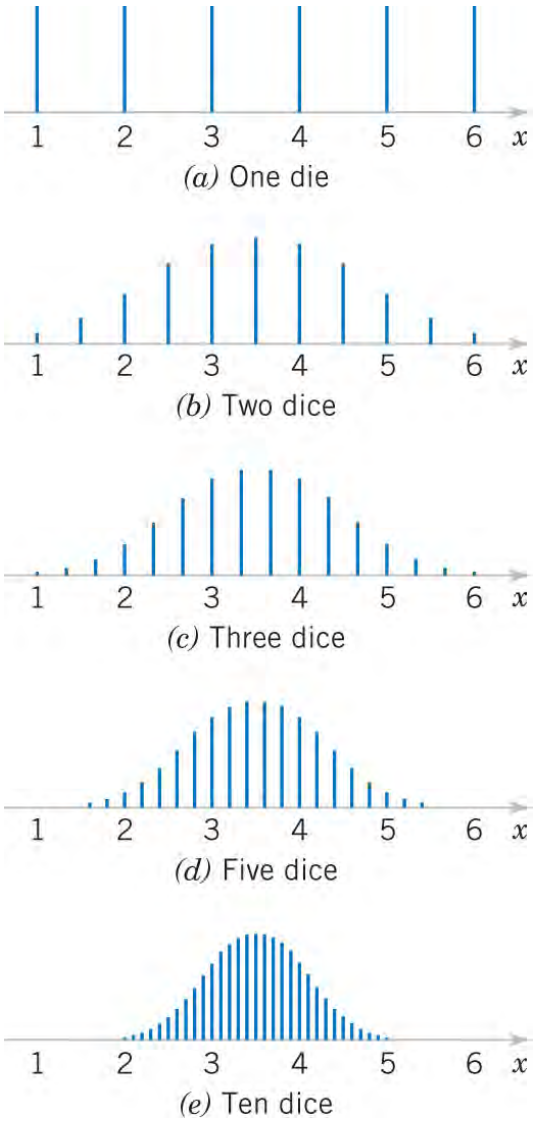
If $X_1, X_2, \ldots, X_n$ are themselves normally distributed – for any n

Test CLT for your own random variable

- Go to: https://onlinestatbook.com/stat_sim/sampling_dist/
- Select “Custom” at the top and use mouse to sketch the PMF of your own random variable
- Select “mean” and n=5 in the third panel
- Choose “Animated” in the second panel and use number_of_experiments=5 to see one sample being generated
- Repeat with number_of _experiments =10,000
- Now select “mean” and n=25 in the fourth panel
- Skewness and Curtosis are measures of how good is the normal (Gaussian) fit (choose “fit normal”)

Sampling Distributions of Sample Means

Figure 7-1 Distributions of average scores from throwing dice.
Mean = (6+1)/2=3.5
Sigma^2 = [(6-1+1)^2-1]/12=2.92
Sigma=1.71



Formulas

$$\mu = \frac{b+a}{2} = 3.5$$

$$\sigma_X^2 = \frac{(b-a+1)^2-1}{12} = 35/12$$

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$$

show Matlab

Matlab exercise

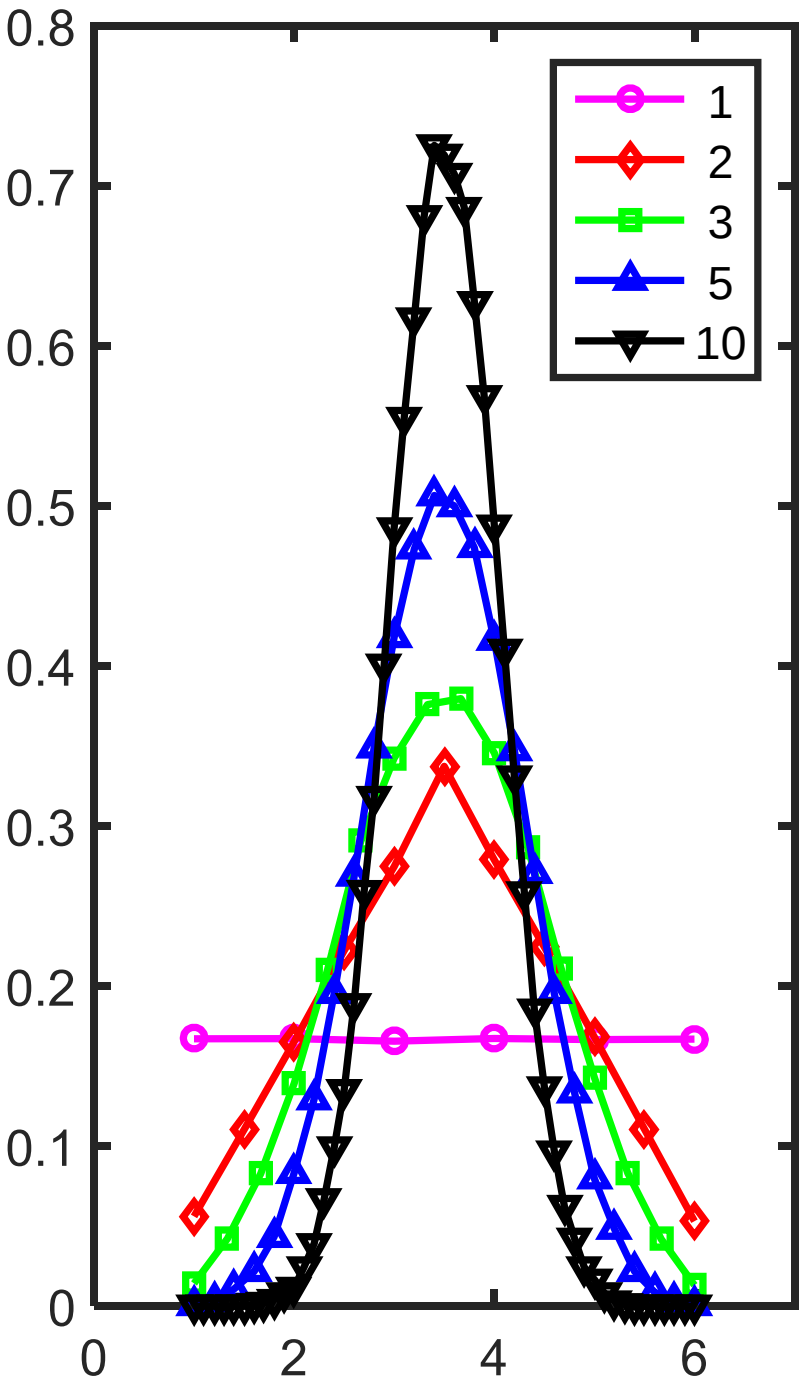
- Do a numerical experiment: generate a sample of size n by rolling n fair dice
- Calculate the sample mean $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
- Repeat Stats=100,000 times
- Generate PDFs of sample means for different samples sizes: n=1, n=2, n=3, n=5, and n=10
- Plot them in the same (semi-logarithmic) figure
- **What do you see?**
- Template is at the website: central_limit_theorem_template.m

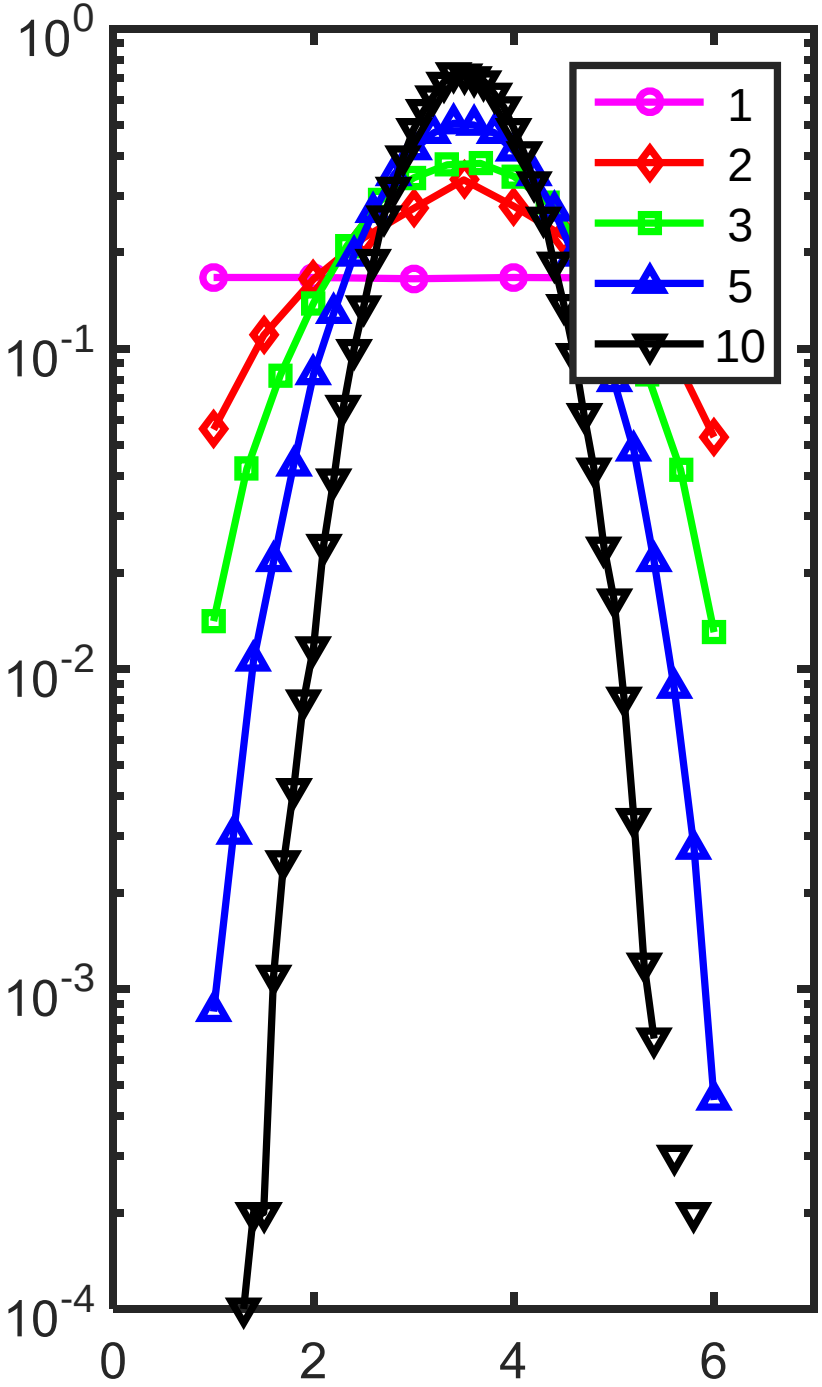
How did I do it?

- **Stats=100000;**
- **figure;**
- **for n=[1,2,3,5,10];**
- **r_sample=floor(6.*rand(Stats,n))+1;**
- **sample_mean=sum(r_sample,2)./n;**
- **step=1./n;**
- **[a,b1]=hist(sample_mean,1:step:6);**
- **pdf_r1=a./sum(a)./step;**
- **semilogy(b1,pdf_r1,'o-'); hold on;**
- **end;**
- **legend('1','2','3','5','10');**

Matlab demonstration

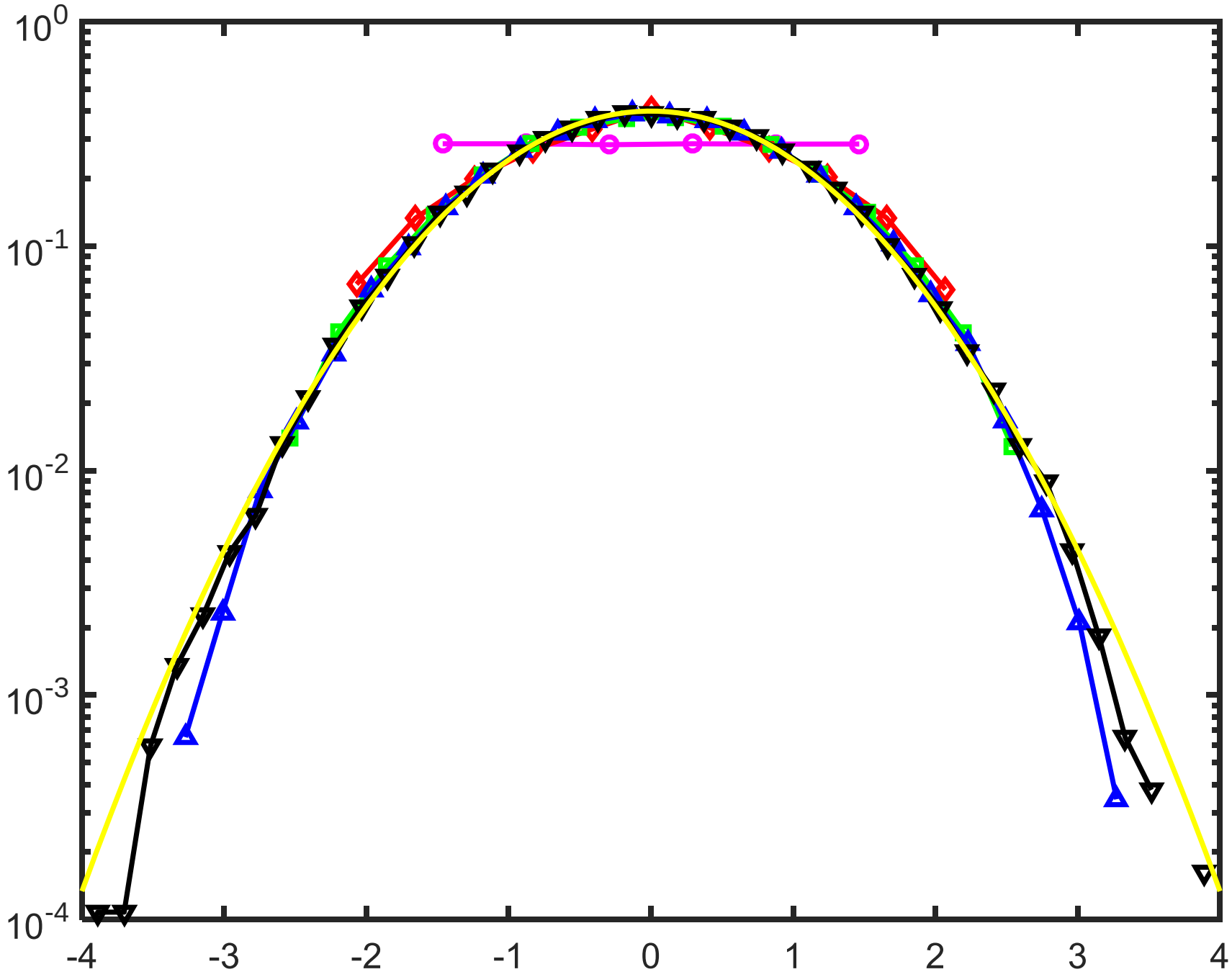
```
Stats=100000; N=10;
r_table=floor(6.*rand(Stats,N))+1;
%%
r1=r_table(:,1);
step=1; [a,b1]=hist(r1,1:step:6);
pdf_r1=a./sum(a)./step;
figure; hold on; subplot(1,2,1); plot(b1,pdf_r1,'mo-'); hold on; axis([0 7 0 0.2]); subplot(1,2,2);
semilogy(b1,pdf_r1,'mo-'); hold on; axis([0 7 1e-3 1]);
%%
r2=(r_table(:,1)+r_table(:,2))./2;
step=0.5; [a,b2]=hist(r2,1:step:6); pdf_r2=a./sum(a)./step;
subplot(1,2,1); plot(b2,pdf_r2,'rd-'); axis([0 7 0 0.4]); subplot(1,2,2); semilogy(b2,pdf_r2,'rd-');
%%
r3=(r_table(:,1)+r_table(:,2)+r_table(:,3))./3;
step=1./3; [a,b3]=hist(r3,1:step:6); pdf_r3=a./sum(a)./step;
subplot(1,2,1); plot(b3,pdf_r3,'gs-');  axis([0 7 0 0.4]); subplot(1,2,2); semilogy(b3,pdf_r3,'gs-');
%%
r5=sum(r_table(:,1:5),2)./5;
step=1./5; [a,b5]=hist(r5,1:step:6); pdf_r5=a./sum(a)./step;
subplot(1,2,1); plot(b5,pdf_r5,'b^-');  axis([0 7 0 0.6]); subplot(1,2,2); semilogy(b5,pdf_r5,'b^-'); axis([0 7 1e-4 1]);
%%
r10=sum(r_table(:,1:10),2)./10;
step=1./10; [a,b10]=hist(r10,1:step:6); pdf_r10=a./sum(a)./step;
subplot(1,2,1); plot(b10,pdf_r10,'kv-'); axis([0 7 0 0.8]); legend(num2str([1,2,3,5,10]'));
subplot(1,2,2); semilogy(b10,pdf_r10,'kv-'); legend(num2str([1,2,3,5,10]'));
```


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
1
2
3
5
10


10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
2
4
6
1
2
3
5
10

Matlab demonstration; part 2

```
%%Now plot all of them normalized to 0 and std 1
sigma=sqrt(35/12);
mu=3.5;
figure;
sigma1=sigma;
semilogy((b1-mu)./sigma1,pdf_r1.*sigma1,'mo-');
axis([-4 4 1e-3 1]);
hold on;
%%
sigma2=sigma./sqrt(2);
semilogy((b2-mu)./sigma2,pdf_r2.*sigma2,'rd-');
%%
sigma3=sigma./sqrt(3);
semilogy((b3-mu)./sigma3,pdf_r3.*sigma3,'gs-');
%%
sigma5=sigma./sqrt(5);
semilogy((b5-mu)./sigma5,pdf_r5.*sigma5,'b^-');
axis([-4 4 1e-4 1]);
%%
sigma10=sigma./sqrt(10);
semilogy((b10-mu)./sigma10,pdf_r10.*sigma10,'kv-');
axis([-4 4 1e-4 1]);
%%
%Let's see how well does the Gaussian fits it
x=-4:0.1:4;
semilogy(x,1./sqrt(2*pi)*exp(-x.^2./2),'y-');
```

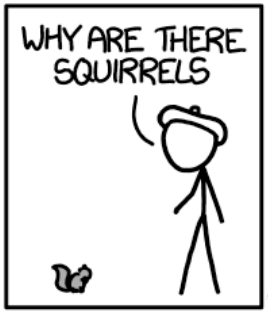

10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
-4
-3
-2
-1
0
1
2
3
4

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

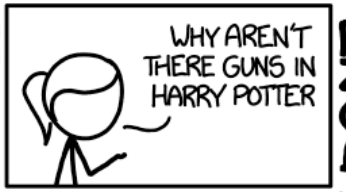
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of $n = 25$ resistors will have an average resistance of less than 95 ohms?

Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of n = 25 resistors will have an average resistance of less than 95 ohms?

$$\mu = 100 \text{ ohms}, \quad \sigma = 10 \text{ ohms}, \quad n = 25$$

$$\mu_{\bar{x}} = \mu; \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2 \text{ ohms}$$

$$Z_{\bar{x}} = \frac{95 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{95 - 100}{2} = -2.5$$

$$\text{Prob}(\bar{X} < 95) = \Phi(Z_{\bar{x}}) = \Phi(-2.5) = 0.0062$$

Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of $n = 25$ resistors will have an average resistance of less than 95 ohms?

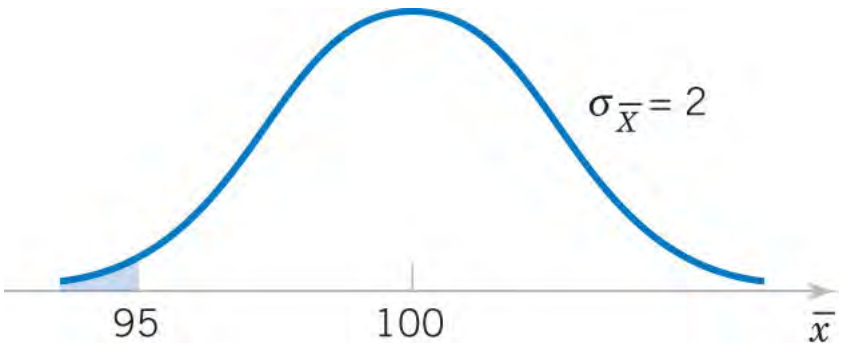


Figure 7-2 Desired probability is shaded

Answer:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2.0$$

$$\Phi\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}}\right) = \Phi\left(\frac{95 - 100}{2}\right)$$

$$= \Phi(-2.5) = 0.0062$$

Two Populations

We have two independent populations. What is the distribution of the difference of their sample means?

The sampling distribution of $\bar{X}_1 - \bar{X}_2$

has the following mean anad variance:

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Sampling Distribution of a Difference in Sample Means

- If we have two independent populations with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2,
- And if X-bar$_1$ and X-bar$_2$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations:
- Then the sampling distribution of:

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (7\text{-}4)$$

  is approximately standard normal, if the conditions of the central limit theorem apply.
- If the two populations are normal, then the sampling distribution is exactly standard normal.

Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a random variable with μ_{old}=5000 hours and σ_{old}=40 hours (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to μ_{new}=5050 hours and σ_{new}=30 hours (new).

Random samples of 16 components manufactured using "old" process and 25 components using "new" process are chosen.

What is the probability new sample mean is at least 25 hours longer than old?

Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a random variable with μ_{old}=5000 hours and σ_{old}=40 hours (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to μ_{new}=5050 hours and σ_{new}=30 hours (new).

Random samples of 16 components manufactured using "old" process and 25 components using "new" process are chosen.

What is the probability new sample mean is at least 25 hours longer than old?

$$\sigma_{\bar{X}_{old}} = \frac{\sigma_{old}}{\sqrt{16}} = 10 \text{ hrs}$$

$$\sigma_{\bar{X}_{new}} = \frac{\sigma_{new}}{\sqrt{25}} = 7 \text{ hrs}$$

$$\sigma_{TOT} = \sqrt{\sigma^2_{\bar{X}_{old}} + \sigma^2_{\bar{X}_{new}}} =$$

$$= \sqrt{100 + 36} \approx 11.7 \text{ hrs}$$

$$\mu_{new} - \mu_{old} = 50 \text{ hrs}$$

$$Z = \frac{25 - (50)}{11.7} = -2.14$$

$$Prob(Z > -2.14) = 0.9840$$

Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a normal-distributed random variable with parameters shown (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters mu and sigma as shown (new).

Random samples are selected from the "old" process and "new" process as shown.

What is the probability new sample mean is at least 25 hours longer than old?

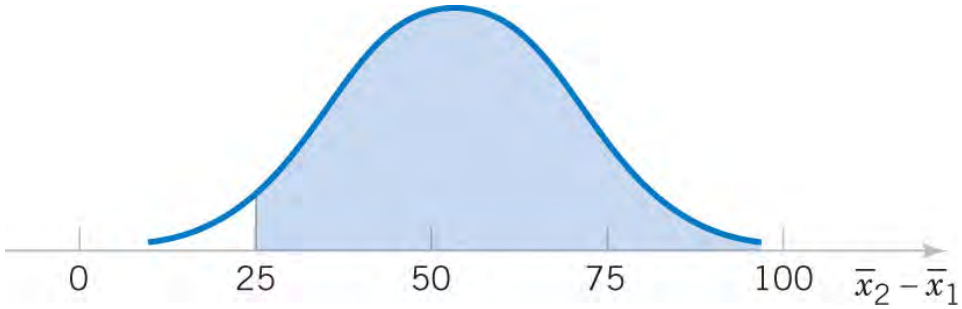


Figure 7-4 Sampling distribution of the sample mean difference.

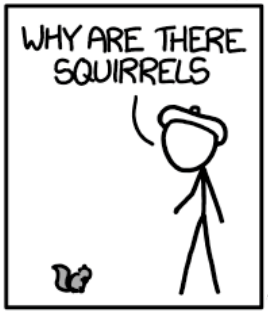
	Process		
	Old (1)	New (2)	Diff (2-1)
mu =	5,000	5,050	50
sigma =	40	30	50
n =	16	25	
Calculations			
$s / \sqrt{n}$ =	10	6	11.7
		z =	-2.14
$P(\bar{x}_2 - \bar{x}_1 > 25) = P(Z > z) =$			0.9840

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

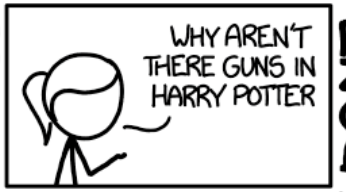
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Descriptive statistics:

Point estimation:

Some Definitions

- The random variables $X_1, X_2,\ldots,X_n$ are a random sample of size n if:
  a) The X_i are independent random variables.
  b) Every X_i has the same probability distribution.

  Such $X_1, X_2,\ldots,X_n$ are also called independent and identically distributed (or i. i. d.) random variables
- A **statistic** is any function of the observations in a random sample.
- The probability distribution of a statistic is called a **sampling distribution**.

Point Estimation

- A sample was collected: $X_1, X_2, \ldots, X_n$
- We suspect that sample was drawn from a random variable distribution $f(x)$
- $f(x)$ has k parameters that we do not know
- Point estimates are estimates of the parameters of the $f(x)$ describing the population based on the sample
  - For exponential PDF: $f(x)=\lambda \exp(-\lambda x)$ one wants to estimate λ
  - For Bernoulli PDF: $p^x(1-p)^{1-x}$ one wants to estimate p
  - For normal PDF one wants to estimates both μ and σ
- Point estimates are uncertain: therefore we can talk of averages and standard deviations of point estimates

Point Estimator

A point estimate of some parameter θ describing population random variable is a single numerical value $\hat{\theta}$ depending on all values $x_1, x_2, \ldots x_n$ in the sample.
The sample statistic (whis a random variable $\widehat{\Theta}$ defined by a function $\widehat{\Theta}(X_1, X_2, \ldots X_n)$) is called the
point estimator.

- There could be multiple choices for the point estimator of a parameter.
- To estimate the mean of a population, we could choose the:
  - Sample mean
  - Sample median
  - Peak of the histogram
  - ½ of (largest + smallest) observations of the sample.
- We need to develop criteria to compare estimates using statistical properties.

Unbiased Estimators Defined

The point estimator $\widehat{\Theta}$ is an unbiased estimator for the parameter θ if:

$$E(\widehat{\Theta}) = \theta \qquad (7-5)$$

If the estimator is not unbiased, then the difference:

$$E(\widehat{\Theta}) - \theta \qquad (7-6)$$

is called the bias of the estimator $\widehat{\Theta}$.

Mean Squared Error

The **mean squared error** of an estimator $\hat{\Theta}$ of the parameter θ is defined as:

$$\mathrm{MSE}(\hat{\Theta}) = E(\hat{\Theta} - \theta)^2 \qquad (7-7)$$

Can be rewritten as

$$= E[\hat{\Theta} - E(\hat{\Theta})]^2 + [\theta - E(\hat{\Theta})]^2$$

$$= V(\hat{\Theta}) + (\mathrm{bias})^2$$

Methods of Point Estimation

- We will cover two popular methodologies to create point estimates of a population parameter.
  - Method of moments
  - Method of maximum likelihood
- Each approach can be used to create estimators with varying degrees of biasedness and relative MSE efficiencies.

Method of moments for point estimation

What are moments?

- A k-th moment of a random variable is the expected value $E(X^k)$
  - First moment: $\mu = \int_{\infty}^{+\infty} xf(x)\,dx$
  - Second moment: $\mu^2 + \sigma^2 = \int_{\infty}^{+\infty} x^2 f(x)\,dx$
- A population moment relates to the entire population
- A sample moment is calculated like its population moments but for a finite sample
  - Sample first moment = sample mean = $\frac{1}{n}\sum_{i=1}^{n} x_i$
  - Sample k-th moment $\frac{1}{n}\sum_{i=1}^{n} x_i^k$

Moment Estimators

Let $X_1, X_2, ..., X_n$ be a random sample from either a probability mass function or a probability density function with m unknown parameters $\theta_1, \theta_2, ..., \theta_m$.

The moment estimators $\hat{\Theta}_1$ $\hat{\Theta}_2$ ..., $\hat{\Theta}_m$ are found by equating the first m population moments to the first m sample moments and solving the resulting simultaneous equations for the unknown parameters.

Exponential Distribution: Moment Estimator-1st moment

- Suppose that $x_1, x_2, \ldots, x_n$ is a random sample from an exponential distribution $f(x)=\lambda\exp(-\lambda x)$ with parameter λ.
- There is only one parameter to estimate, so equating population and sample first moments, we have one equation: $E(X) = \bar{x}$.
- $E(X) = 1/\lambda$ thus
  $\lambda = 1/\bar{x}$ is the 1st moment estimator.

Descriptive statistics:

Point estimation:

Point Estimation

- A sample was collected: $X_1, X_2, \ldots, X_n$
- We suspect that sample was drawn from a random variable distribution $f(x)$
- $f(x)$ has k parameters that we do not know
- Point estimates are estimates of the parameters of the $f(x)$ describing the population based on the sample
  - For exponential PDF: $f(x)=\lambda \exp(-\lambda x)$ one wants to estimate λ
  - For Bernoulli PDF: $p^x(1-p)^{1-x}$ one wants to estimate p
  - For normal PDF one wants to estimates both μ and σ
- Point estimates are uncertain: therefore, we can talk of averages and standard deviations of point estimators

Point Estimator

A point estimate of some parameter θ describing population random variable is a single numerical value $\hat{\theta}$ depending on all values $x_1, x_2, \ldots x_n$ in the sample.
The sample statistic (whis a random variable $\widehat{\Theta}$ defined by a function $\widehat{\Theta}(X_1, X_2, \ldots X_n)$) is called the point estimator.

- There could be multiple choices for the point estimator of a parameter.
- To estimate the mean of a population, we could choose the:
  - Sample mean
  - Sample median
  - Peak of the histogram
  - ½ of (largest + smallest) observations of the sample.
- We need to develop criteria to compare estimates using statistical properties.

Unbiased Estimators Defined

The point estimator $\widehat{\Theta}$ is an unbiased estimator for the parameter θ if:

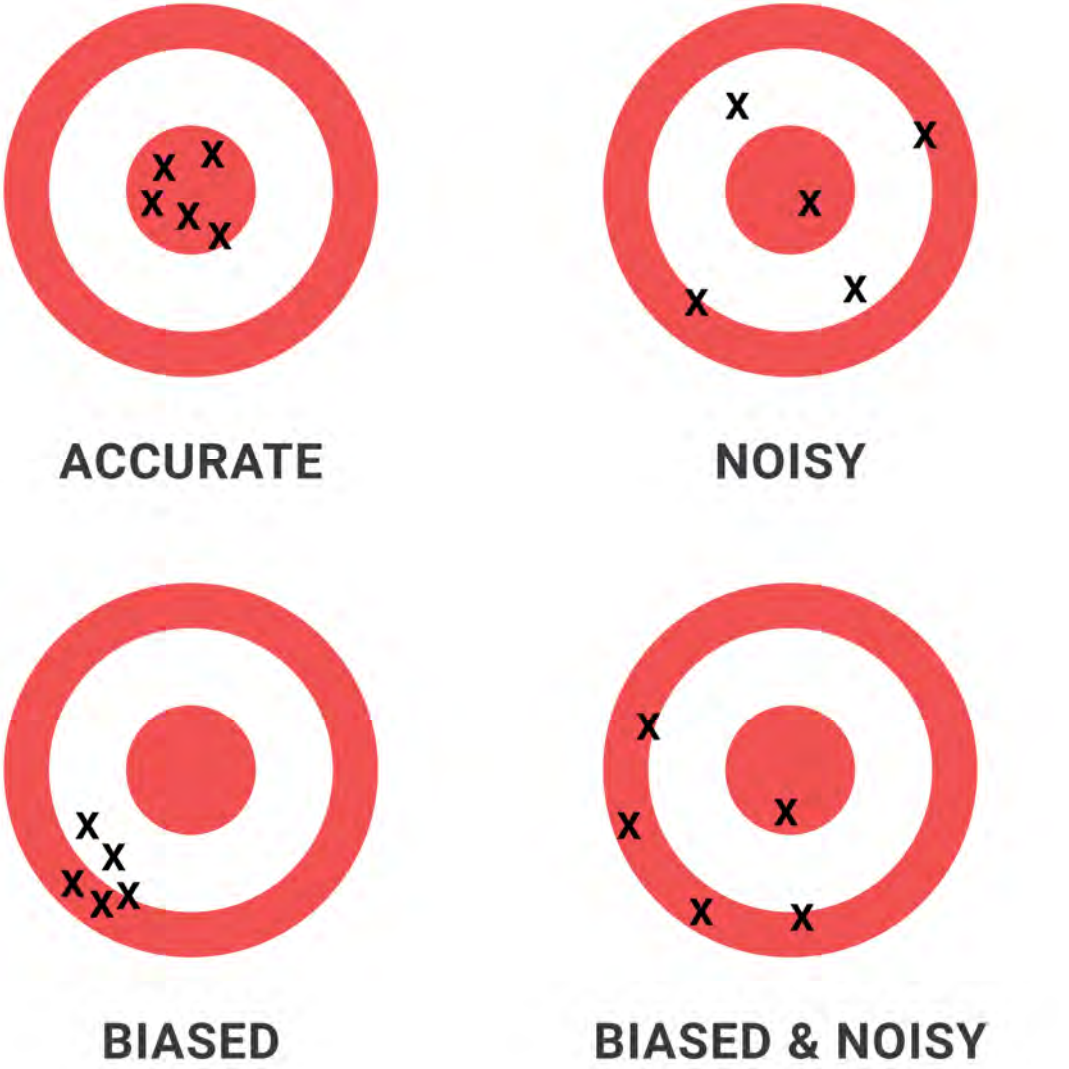
$$E(\widehat{\Theta}) = \theta \qquad (7-5)$$

If the estimator is not unbiased, then the difference:

$$E(\widehat{\Theta}) - \theta \qquad (7-6)$$

is called the bias of the estimator $\widehat{\Theta}$.

Bias vs Noise



NOISE: Daniel Kahneman, Olivier Sabony, and Cass R. Sunstein

Mean Squared Error

The **mean squared error** of an estimator $\hat{\Theta}$ of the parameter θ is defined as:

$$\mathrm{MSE}(\hat{\Theta}) = E(\hat{\Theta} - \theta)^2 \qquad (7-7)$$

Can be rewritten as

$$= E[\hat{\Theta} - E(\hat{\Theta})]^2 + [\theta - E(\hat{\Theta})]^2$$

$$= V(\hat{\Theta}) + (\mathrm{bias})^2$$

Statistic #1: Sample Mean

If the values of n observations in a random sample are denoted by $x_1, x_2, \ldots, x_n$, the sample mean is

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (6-1)$$

New random variable $\overline{X}$ is a linear combination of *n* independent identically distributed variables $X_1, X_2, \ldots, X_n$

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Sample mean $\bar{x}$ is drawn from a random variable

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$$E(\bar{X}) = \frac{n \cdot E(X_i)}{n} = \frac{n \cdot \mu}{n} = \mu$$

Sample mean, $\bar{X}$, is an unbiased estimator of the population mean, μ

Sample variance S^2 –
is an estimator of
the population variance σ^2

Sample Variance

If n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad (6-3)$$

If one knows the population average, μ, one divides by n to estimate the variance

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

Why divide by n-1 instead of n?

- The sample mean $\bar{x}$ is on average closer to points $x_1, x_2, \ldots x_n$ than the true mean μ
  $\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq \sum_{i=1}^{n}(x_i - \mu)^2$
- Consider a sample of size n=1.
  Then $\bar{x} = x_1$ while $\mu \neq x_1$. Dividing by n gives $s^2 = 0$, while dividing by n-1 leaves s^2 undefined (0/0)
- For n=2, $\bar{x}$ is exactly halfway between x_1 and x_2 making its sum of squares smaller than that of μ
- Dividing by n-1 on average corrects for a smaller sum of squares: S^2 is an unbiased estimator of σ^2

Show that S^2 is unbiased estimate of σ^2

$$E(S^2) = E\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right) = \frac{1}{n-1} E\left[\sum_{i=1}^{n}\left(X_i^2 + \bar{X}^2 - 2\bar{X}X_i\right)\right] =$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_i^2 + n\bar{X}^2 - 2\bar{X}\,n\bar{X}\right] =$$

$$= \frac{1}{n-1} E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \frac{1}{n-1}\left(nE(X_i^2) - nE(\bar{X}^2)\right)$$

$$= \frac{1}{n-1}\left(n\left(\mu^2 + \sigma^2\right) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right) = \frac{n-1}{n-1}\sigma^2 = \underline{\underline{\sigma^2}}$$

Example 7-4: Sample Variance S^2 is Unbiased

$$E(S^2) = E\left(\frac{\sum_{i=1}^{n}(X-\bar{X})^2}{n-1}\right)$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i^2 + \bar{X}^2 - 2\bar{X}X_i\right)\right]$$

$$= \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right)\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\mu^2+\sigma^2\right) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right]$$

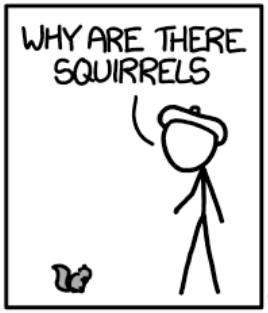
$$= \frac{1}{n-1}\left[n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2\right] = \frac{1}{n-1}\left[(n-1)\sigma^2\right]$$

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

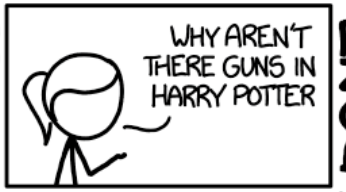
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Methods of Point Estimation

- We will cover two popular methodologies to create point estimates of a population parameter.
  - Method of moments
  - Method of maximum likelihood
- Each approach can be used to create estimators with varying degrees of biasedness and relative MSE efficiencies.

Method of moments for point estimation

What are moments?

- The p-th population moment of a random variable is the expected value of X^p
  - First moment: $\mu = \int_{\infty}^{+\infty} x f(x)\, dx$
  - Second moment: $\mu^2 + \sigma^2 = \int_{\infty}^{+\infty} x^2 f(x)\, dx$
  - p-th moment: $\int_{\infty}^{+\infty} x^p f(x)\, dx$
  - The population moment relates to the entire population

- A sample moment is calculated like its population moments but for a finite sample
  - Sample first moment = sample mean = $\frac{1}{n}\sum_{i=1}^{n} x_i$
  - Sample p-th moment $\frac{1}{n}\sum_{i=1}^{n} x_i^p$

Moment Estimators

Let $X_1, X_2, \ldots, X_n$ be a random sample from either a probability mass function or a probability density function with p unknown parameters $\theta_1, \theta_2, \ldots, \theta_p$.

The moment estimators $\hat{\Theta}_1\ \hat{\Theta}_2 \ldots, \hat{\Theta}_p$ are found by equating the first p population moments to the first p sample moments and solving the resulting simultaneous equations for the unknown parameters.

Exponential Distribution: Moment Estimator-1st moment

- Suppose that $x_1, x_2, \ldots, x_n$ is a random sample from an exponential distribution $f(x)=\lambda\exp(-\lambda x)$ with parameter λ.
- There is only one parameter to estimate, so equating population and sample first moments, we have one equation: $E(X) = \bar{x}$.
- $E(X) = 1/\lambda$ thus
  $\lambda = 1/\bar{x}$ is the 1st moment estimator.

How I solved it

- Stats=100000;
- Y=random('Exponential', 1/3, Stats, 1);

%parametrization in MATLAB is 1/lambda

- 1/mean(Y) %matching the first moment

% ans = 3.0086

- sqrt(2/mean(Y.^2)) %matching the second moment

% ans = 3.0081

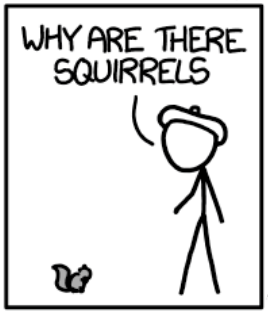
- (factorial(20)/mean(Y.^20))^(1./20) %matching the 20th moment

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

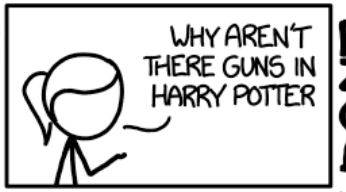
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Method of Maximum Likelihood for point estimation

Maximum Likelihood Estimators

- Suppose that X is a random variable with probability distribution $f(x, \theta)$, where θ is a single unknown parameter. Let $x_1, x_2, \ldots, x_n$ be the observed values in a random sample of size n. Then the likelihood function of the sample is the probability to get it in a random variable with PDF $f(x, \theta)$:

$$L(\theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n, \theta) \qquad (7\text{-}9)$$

- Note that the likelihood function is now a function of only the unknown parameter θ. The maximum likelihood estimator (MLE) of θ is the value of θ that maximizes the likelihood function $L(\theta)$.
- Usually, it is easier to work with logarithms: $l(\theta) = \ln L(\theta)$

Exponential MLE:

$$f(x_i) = \lambda e^{-\lambda x_i}$$

$$L(\lambda) = P(x_1, x_2 \ldots x_n \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} =$$

$$= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$\ln L(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum x_i = 0$$

$$\hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{X}}$$

Same as 1st moment estimator

Example 7-11: Exponential MLE

Let *X* be a exponential random variable with parameter λ. The likelihood function of a random sample of size n is:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$\ln L(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{X}} \quad \text{(same as moment estimator)}$$

Bernoulli MLE

$$f(x, p) = p^x (1-p)^{1-x}$$

$$L(p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} =$$

$$= p^{\sum x_i} (1-p)^{n - \sum x_i}$$

$$\ln L(p) = \left(\sum x_i\right) \ln p + \left(n - \sum x_i\right) \ln(1-p)$$

$$\frac{d \ln L(p)}{dp} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0 \quad \text{at } \hat{p}$$

$$0 = \frac{(1-\hat{p}) \sum x_i - \hat{p}\left(n - \sum x_i\right)}{\hat{p}(1-\hat{p})}$$

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

Example 7-9: Bernoulli MLE

Let X be a Bernoulli random variable. The probability mass function is $f(x;p) = p^x(1-p)^{1-x}$, $x = 0, 1$ where P is the parameter to be estimated. The likelihood function of a random sample of size n is:

$$L(p) = p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \cdot \ldots \cdot p^{x_n}(1-p)^{1-x_n}$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

$$\ln L(p) = \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

$$\frac{d \ln L(p)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{(n - \sum_{i=1}^{n} x_i)}{(1-p)} = 0$$

$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$ (same as moment estimator)

Normal MLE for μ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$L(\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(-\frac{\sum(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\ln L(\mu,\sigma) = -n\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{d\ln L(\mu,\sigma)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \quad \text{at } \hat{\mu}$$

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

Example 7-10: Normal MLE for μ

Let *X* be a normal random variable with unknown mean μ and variance σ^2. The likelihood function of a random sample of size n is:

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{(2\sigma^2)}}$$

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

$$\ln L\,(\mu) = \frac{-n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{d\ln L\,(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X} \text{ (same as moment estimator)}$$

Example 7-11: Normal MLE for σ^2

Let *X* be a normal random variable with the estimate of mean μ determined by MLE (see the previous slide) and an unknown variance σ^2. The likelihood function of a random sample of size n is:

$$L(\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{(2\sigma^2)}}$$

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

$$\ln L(\sigma) = \frac{-n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{d\ln L(\sigma)}{d\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} \text{ (biased estimator)}$$

MLE for Poisson distribution

$$f(x_1, \ldots, x_n|\lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$

$$= \frac{e^{-n\lambda}\lambda^{\Sigma_1^n x_i}}{x_1! \ldots x_n!}$$

$$\log f(x_1, \ldots, x_n|\lambda) = -n\lambda + \sum_1^n x_i \log \lambda - \log c$$

where $c = \prod_{i=1}^n x_i!$ does not depend on λ, and

$$\frac{d}{d\lambda} \log f(x_1, \ldots, x_n|\lambda) = -n + \frac{\sum_1^n x_i}{\lambda}$$

By equating to zero, we obtain that the maximum likelihood estimate $\hat{\lambda}$ equals

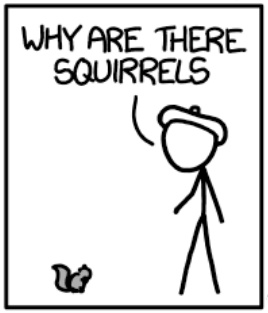
$$\hat{\lambda} = \frac{\sum_1^n x_i}{n}$$

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

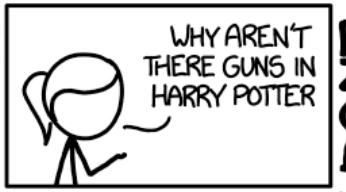
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

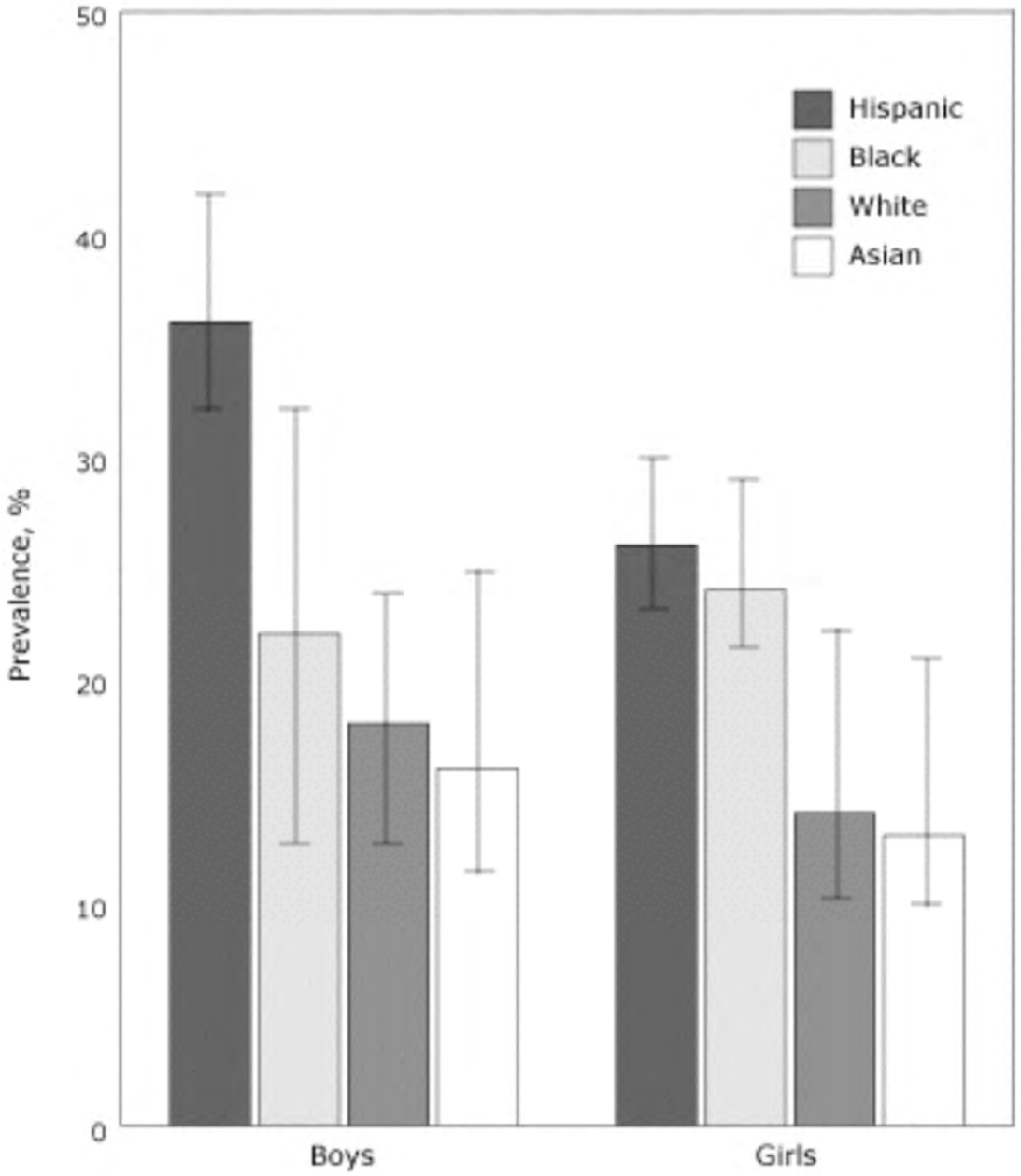
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Confidence Intervals



Prevalence (with 95% CI bars) of obesity among New York City public elementary schoolchildren, by sex and race/ethnicity, 2003.

(source: CDC.GOV)

What do those bars actually mean?

Vol 446|8 March 2007|doi:10.1038/nature05610 nature

ARTICLES

Patterns of somatic mutation in human cancer genomes

What does confidence interval mean?

The numbers of passenger and driver mutations present can be estimated from these results (see Supplementary Methods). Of the 921 base substitutions in the primary screen, 763 (95% confidence interval, 675–858) are estimated to be passenger mutations. Therefore, the large majority of mutations found through sequencing cancer genomes are not implicated in cancer development, even when the search has been targeted to the coding regions of a gene family of high candidature. However, there are an estimated 158 driver mutations (95% confidence interval, 63–246), accounting for the observed positive selection pressure. These are estimated to be distributed in 119 genes (95% confidence interval, 52–149). The number of samples containing a driver mutation is estimated to be 66 (95% confidence interval, 36–77). The results, therefore, provide statistical evidence for a large set of mutated protein kinase genes implicated in the development of about one-third of the cancers studied.

- We have talked about how a parameter can be estimated from sample data. However, it is important to understand how good is the estimate obtained.

- Bounds that represent an interval of plausible values for a parameter are an example of an interval estimate.

Two-sided confidence intervals

- Calculated based on the sample $X_1, X_2,\ldots,X_n$
- Characterized by:
  - lower- and upper- confidence limits L and R
  - the confidence coefficient $1-\alpha$
- Objective: for two-sided confidence interval, find L and R such that
  - Prob$(\mu>R)=\alpha/2$
  - Prob$(\mu<L)=\alpha/2$
  - Therefore, Prob$(L<\mu<R)=1-\alpha$
- For one-sided confidence interval, say, upper bound of μ , find R that
  - Prob$(\mu>R)=\alpha$
- **Assume standard deviation sigma is known**

Consider $1-\alpha = 95\% = 0.95$

$\alpha = 0.0$; $\frac{\alpha}{2} = 0.025$

$z_{\alpha/2} = 1.96 \Rightarrow \text{Prob}(z > z_{\alpha/2}) = \frac{\alpha}{2}$

$$\text{Prob}\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\text{Prob}\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

For one sided lower bound on μ

$$\text{Prob}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha}\right) \rightarrow$$

$$\mu > \bar{X} - z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

$z_{\alpha} = 1.65 <$

$z_{\alpha/2} = 1.96$

Exercise

Ishikawa et al. (Journal of Bioscience and Bioengineering 2012) studied the force with which bacterial biofilms adhere to a solid surface.

Five measurements for a bacterial strain of Acinetobacter gave readings 2.69, 5.76, 2.67, 1.62, and 4.12 dyne-cm2.

Assume that the standard deviation is known to be 0.66 dyne-cm2

(a) Find 95% confidence interval for the mean adhesion force

(b) If scientists want the width of the confidence interval to be below 0.55 dyne-cm2 what number of samples should be?

Ishikawa et al. (Journal of Bioscience and Bioengineering 2012) studied the force with which bacterial biofilms adhere to a solid surface. Five measurements for a bacterial strain of Acinetobacter gave readings 2.69, 5.76, 2.67, 1.62, and 4.12 dyne-cm2. Assume that the standard deviation is known to be 0.66 dyne-cm2

(a) Find 95% confidence interval for the mean adhesion force

(b) If scientists want the width of the confidence interval to be below 0.55 dyne-cm2 what number of samples should be?

a) 95% CI for μ, $n = 5 \quad \sigma = 0.66 \quad \bar{x} = 3.372, z = 1.96$

$$\bar{x} - z\sigma/\sqrt{n} \le \mu \le \bar{x} + z\sigma/\sqrt{n}$$

$$3.372 - 1.96(0.66/\sqrt{5}) \le \mu \le 3.372 + 1.96(0.66/\sqrt{5})$$

$$2.79 \le \mu \le 3.95$$

b) Width is $2z\sigma/\sqrt{n} = 0.55$, therefore $n = [2z\sigma/0.55]2 = [2(1.96)(0.66)/0.55]2 = 22.13$
Round up to $n = 23$.

Confidence Intervals

- We have talked about how a parameter can be estimated from sample data. However, it is important to understand how good is the estimate obtained.

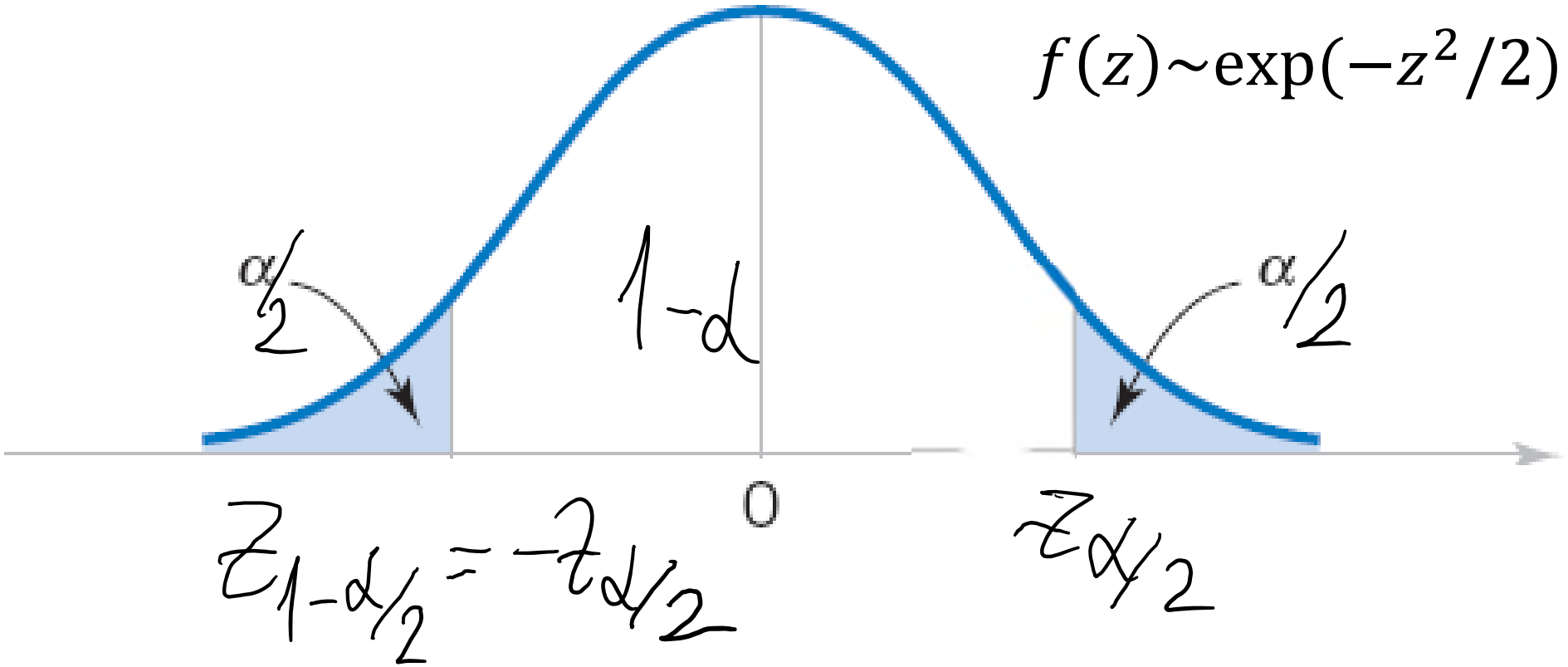
- Bounds that represent an interval of plausible values for a parameter are an example of an interval estimate.

Two-sided confidence intervals

- Calculated based on the sample $X_1, X_2,\ldots,X_n$
- Characterized by:
  - lower- and upper- confidence limits L and U
  - the confidence coefficient $1-\alpha$
- Objective: for two-sided confidence interval, find L and R such that
  - Prob$(\mu>U)=\alpha/2$
  - Prob$(\mu<L)=\alpha/2$
  - Therefore, Prob$(L<\mu<U)=1-\alpha$
- For one-sided confidence interval, say, upper bound of μ , find R that
  - Prob$(\mu>U)=\alpha$
- **Assume standard deviation σ is known**

Confidence Interval on the Population Mean, Variance Known

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



Matlab exercise

- 1000 labs measured average P53 gene expression using n=20 samples drawn from the Gaussian distribution with mu=3; sigma=2;
- Each lab found 95% confidence estimates of the population mean mu **based on its sample only**
- Count the number of labs, where the population mean lies **outside their bounds**
- You should get ~50 labs out of 1000 labs

8-2 Confidence Interval on the Mean of a Normal Distribution, Variance Known

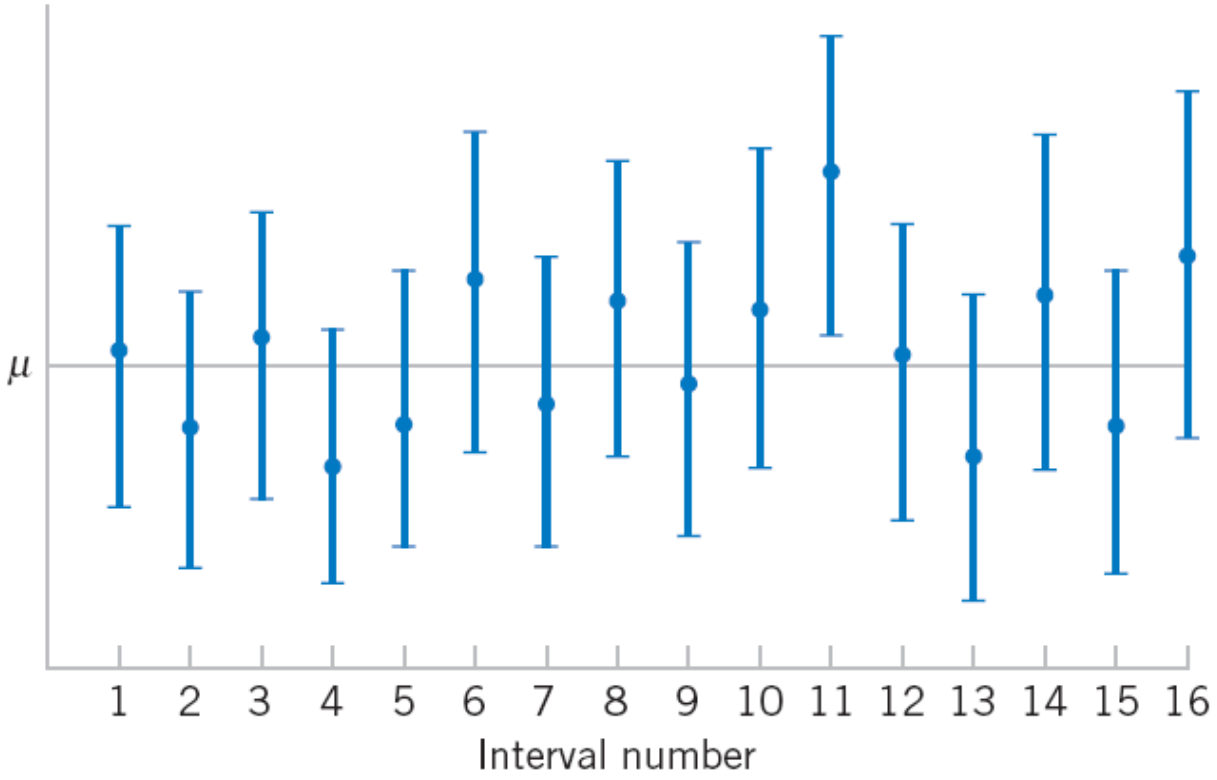


Figure 8-1 Repeated construction of a confidence interval for μ.

Figure 8-1 Repeated construction of a confidence interval for μ.

So far in estimating
confidence intervals for population mean μ
we assumed that the population variance σ^2
is known

Then (or when n>>1, say 20 and above)
one can use the Normal Distribution
to calculate confidence intervals

Q: What to do if the sample is small
& the population variance is **not known**?

A: Use the sample variance

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

but carefully:

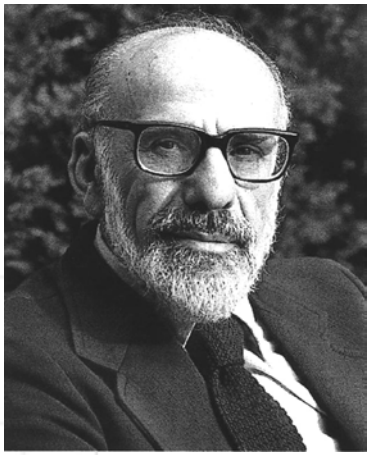
- Variable X has to be **normally distributed**
- **Student t-distribution** has to be used instead of the normal distribution (z-distribution).

Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery. To prevent further disclosure of confidential information, Guinness prohibited its employees from publishing any papers regardless of the contained information. However, after pleading with the brewery and explaining that his mathematical and philosophical conclusions were of no possible practical use to competing brewers, he was allowed to publish them, but under a pseudonym ("Student"), to avoid difficulties with the rest of the staff. Thus, his most noteworthy achievement is now called Student's, rather than Gosset's, t-distribution.

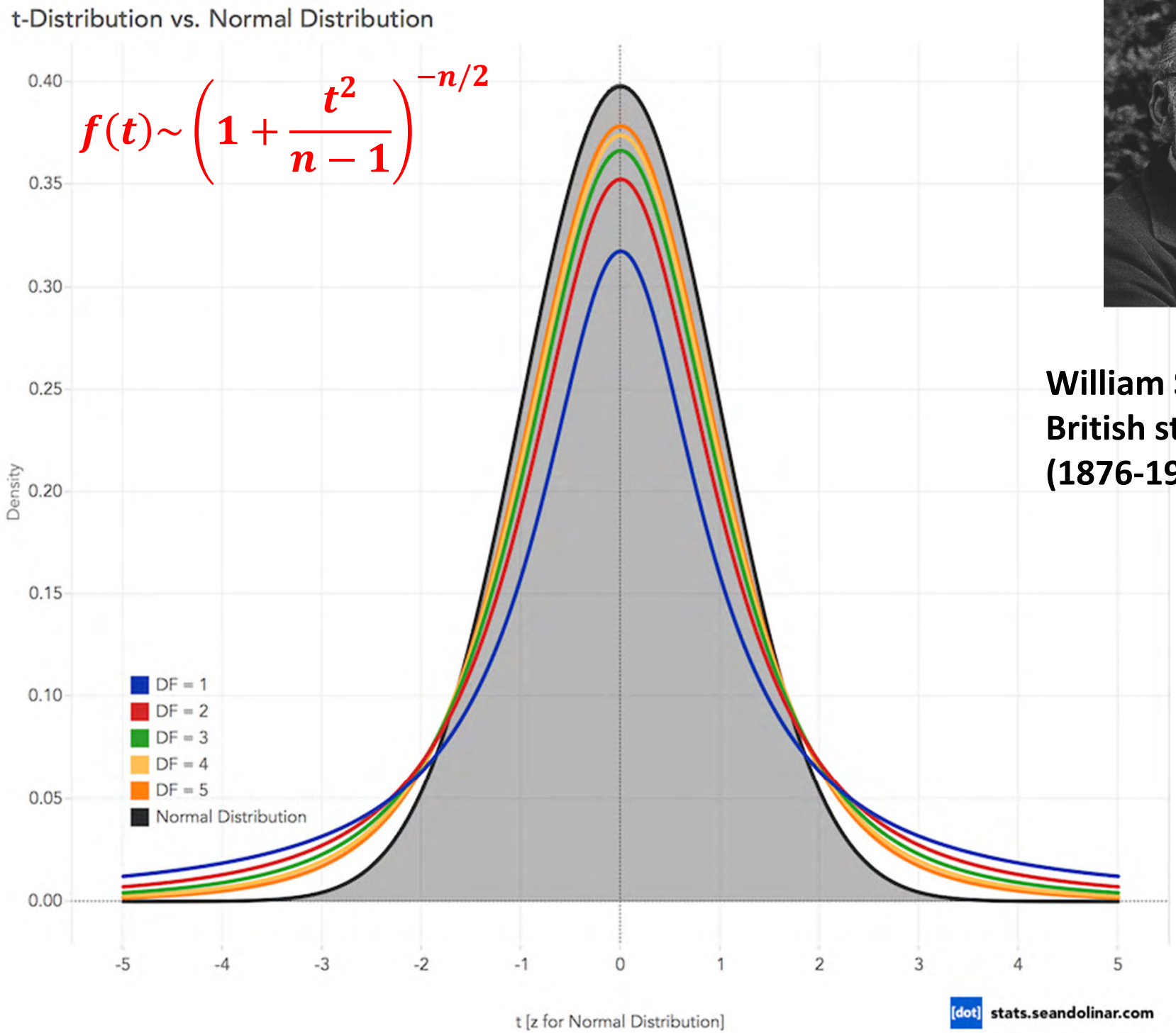
Gosset had almost all his papers including “The probable error of a mean” (1908) published in Pearson's journal Biometrika under the pseudonym Student

Student's t-distribution

$$f(t) \sim \left(1 + \frac{t^2}{n-1}\right)^{-n/2}$$



William Sealy Gosset
British statistician
(1876-1937)



Play with Mathematica notebook

http://demonstrations.wolfram.com/ComparingNormalAndStudentsTDistributions/

By Gary McClelland

8-3 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

Student's *t* distribution

$$f(t) \sim \left(1 + \frac{t^2}{n-1}\right)^{-n/2}$$

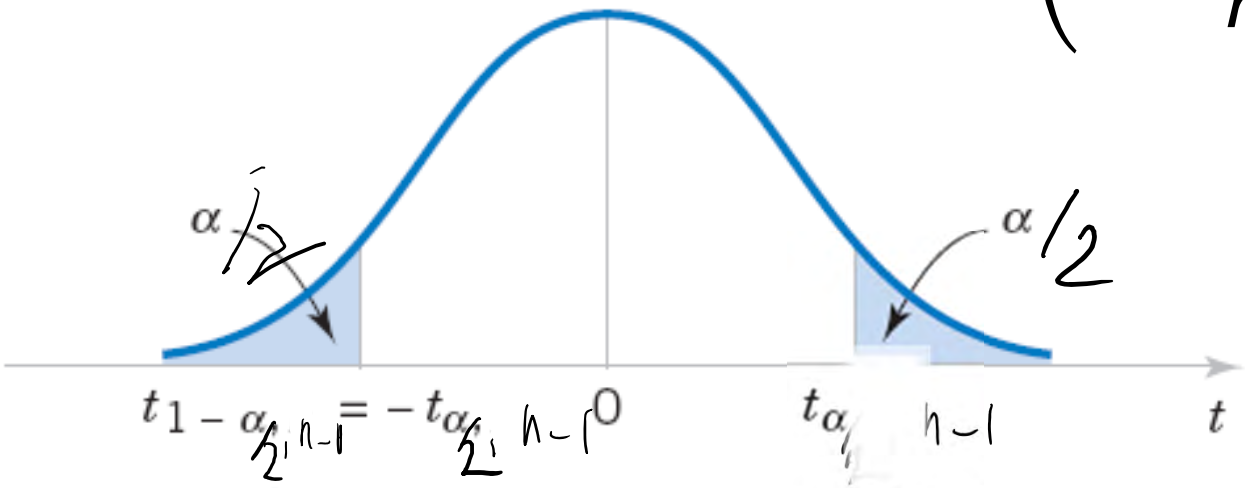


Figure 8-5 Percentage points of the *t* distribution.

8-3 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

8-3.2 The *t* Confidence Interval on μ (Eq. 8-16)

If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a **100(1 − α)% confidence interval on μ** is given by

$$\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n} \qquad (8\text{-}16)$$

where $t_{\alpha/2,n-1}$ is the upper $100\alpha/2$ percentage point of the t distribution with $n - 1$ degrees of freedom.

One-sided confidence bounds on the mean are found by replacing $t_{\alpha/2,n-1}$ in Equation 8-16 with $t_{\alpha,n-1}$.

Confidence intervals for the population variance σ^2 based on the sample variance s^2

Confidence interval for the population variance σ^2

- Up until now we were calculating the confidence interval on the **population average μ**
- What if one wants to put **confidence interval on the population variance $\boldsymbol{\sigma^2}$**?
- We know an unbiased estimator of σ^2:

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

- How to determine the confidence interval?

Assume $\sigma = 1$, $\mu = 0$

$$\vec{X} = (X_1, \dots X_n)$$

$$y = |\vec{X}|^2 = \sum_{i=1}^{n} X_i^2 = (n-1)S^2$$

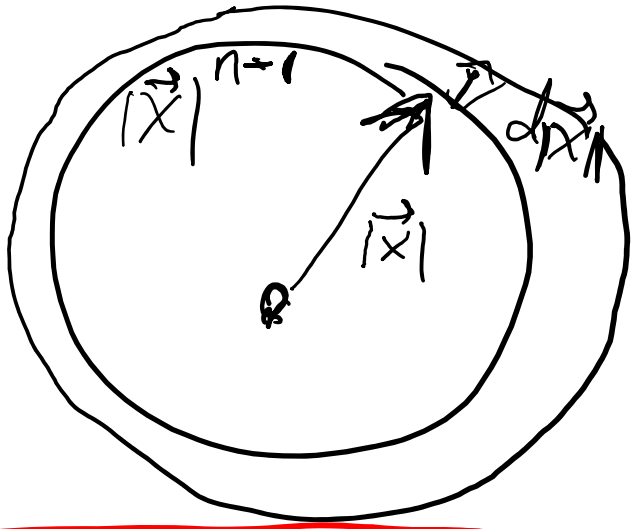
$$P(X_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{X_i^2}{2}\right)$$

$$P(|\vec{X}|) \sim \exp\left(-\frac{|\vec{X}|^2}{2}\right) \cdot \text{Surface of the sphere}$$

$$|\vec{X}| = \sqrt{y}$$

$$d|\vec{X}| = \frac{1}{2}\frac{dy}{\sqrt{y}}$$

$$|\vec{X}|^{n-1} d|\vec{X}| = y^{\frac{n-1}{2}} y^{-\frac{1}{2}} dy =$$



$$= y^{\frac{n}{2}-1} dy$$

$$\boxed{P(y)dy = y^{\frac{n}{2}-1} \exp\left(-\frac{y}{2}\right) dy}$$

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

Definition

(Eq. 8-17)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2, and let S^2 be the sample variance. Then the random variable

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \qquad (8\text{-}17)$$

has a chi-square (χ^2) distribution with $n-1$ degrees of freedom.

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

$X=(n-1)S^2/\sigma^2$

We know n, S^2

want to estimate σ^2

$f(x,n) \sim x^{(n-1)/2-1}\exp(-x/2)$

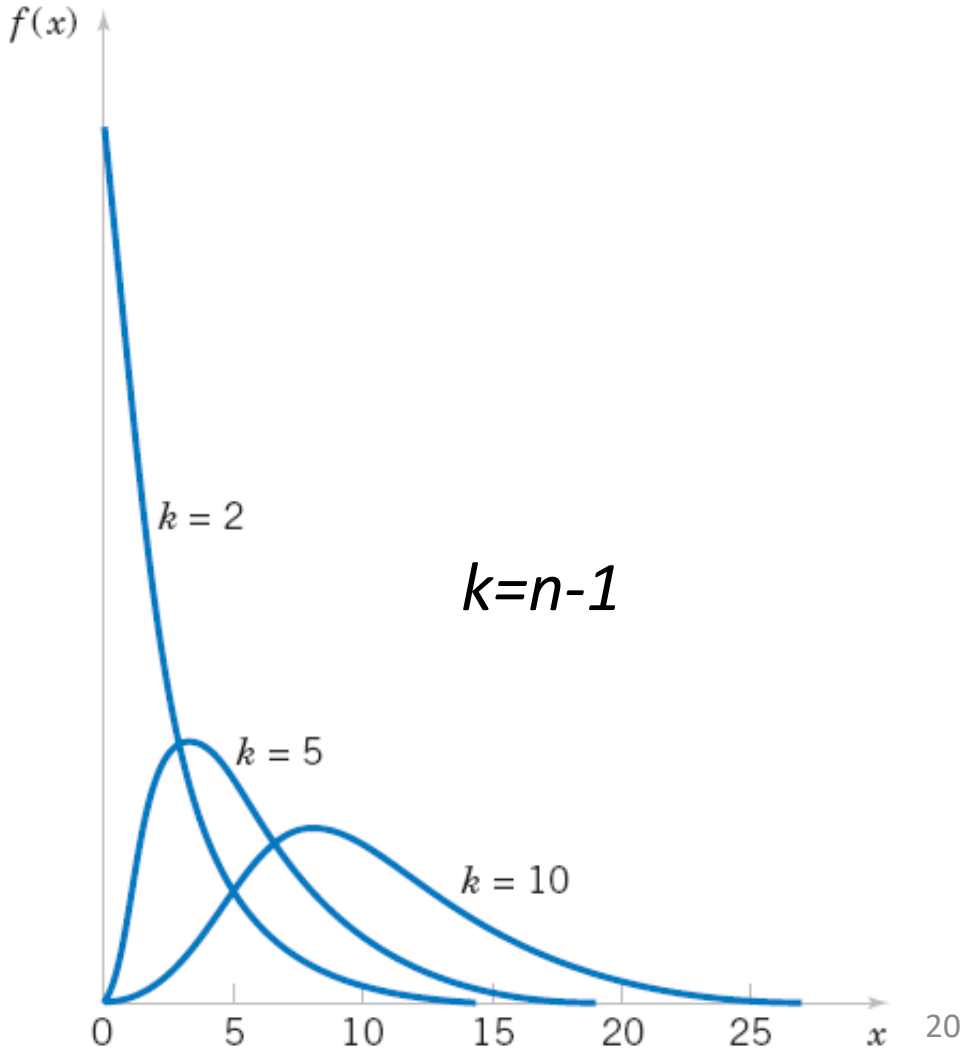
It is just Gamma PDF with $r=(n-1)/2$, and $\lambda=1/2$

Mean value:

$n-1$

Standard deviation:

$$\sqrt{2(n-1)}$$

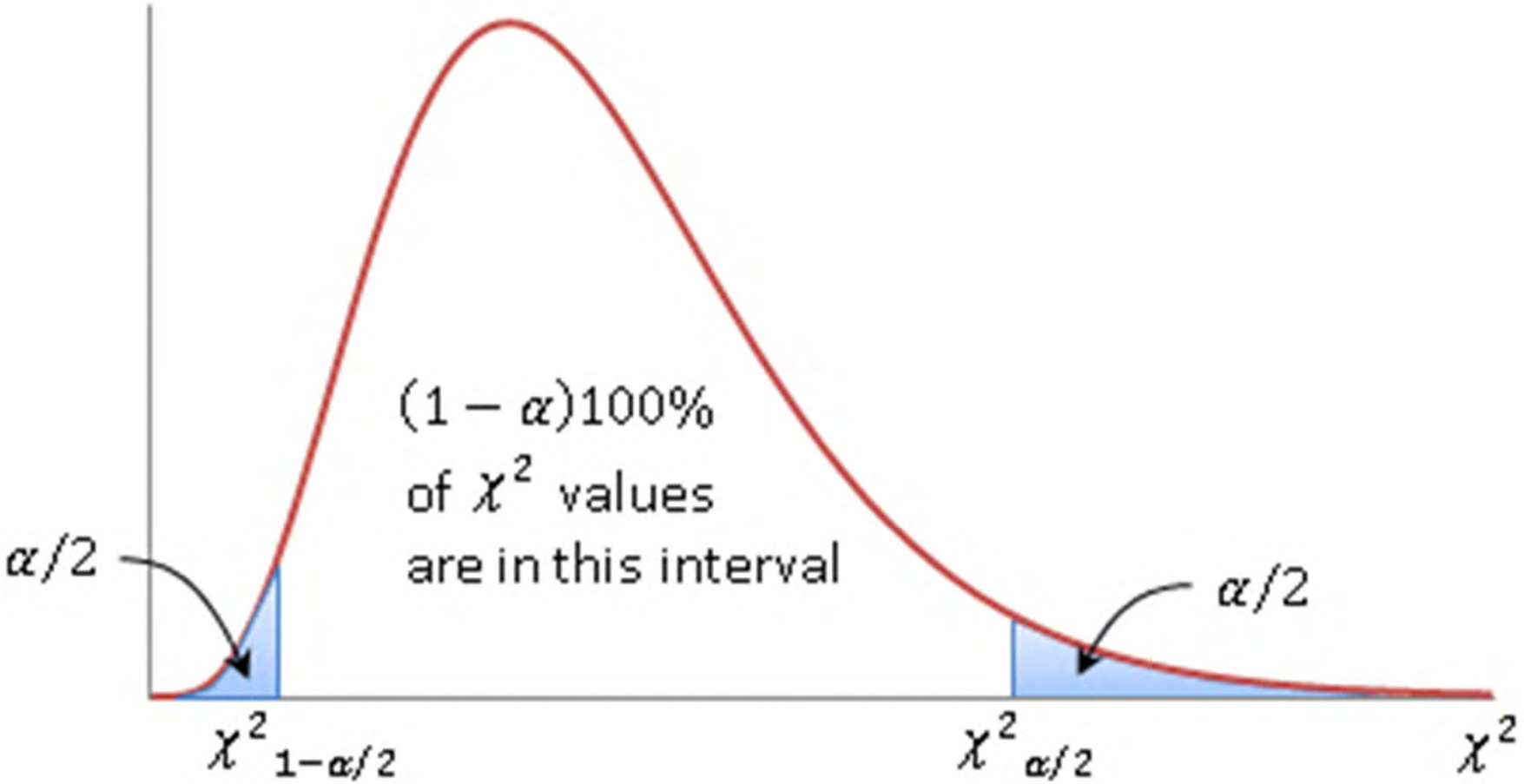


$k=n-1$

Play with Mathematica notebook

http://demonstrations.wolfram.com/ChiSquaredDistributionAndTheCentralLimitTheorem/

By Peter Falloon



$$\chi^2_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

Definition

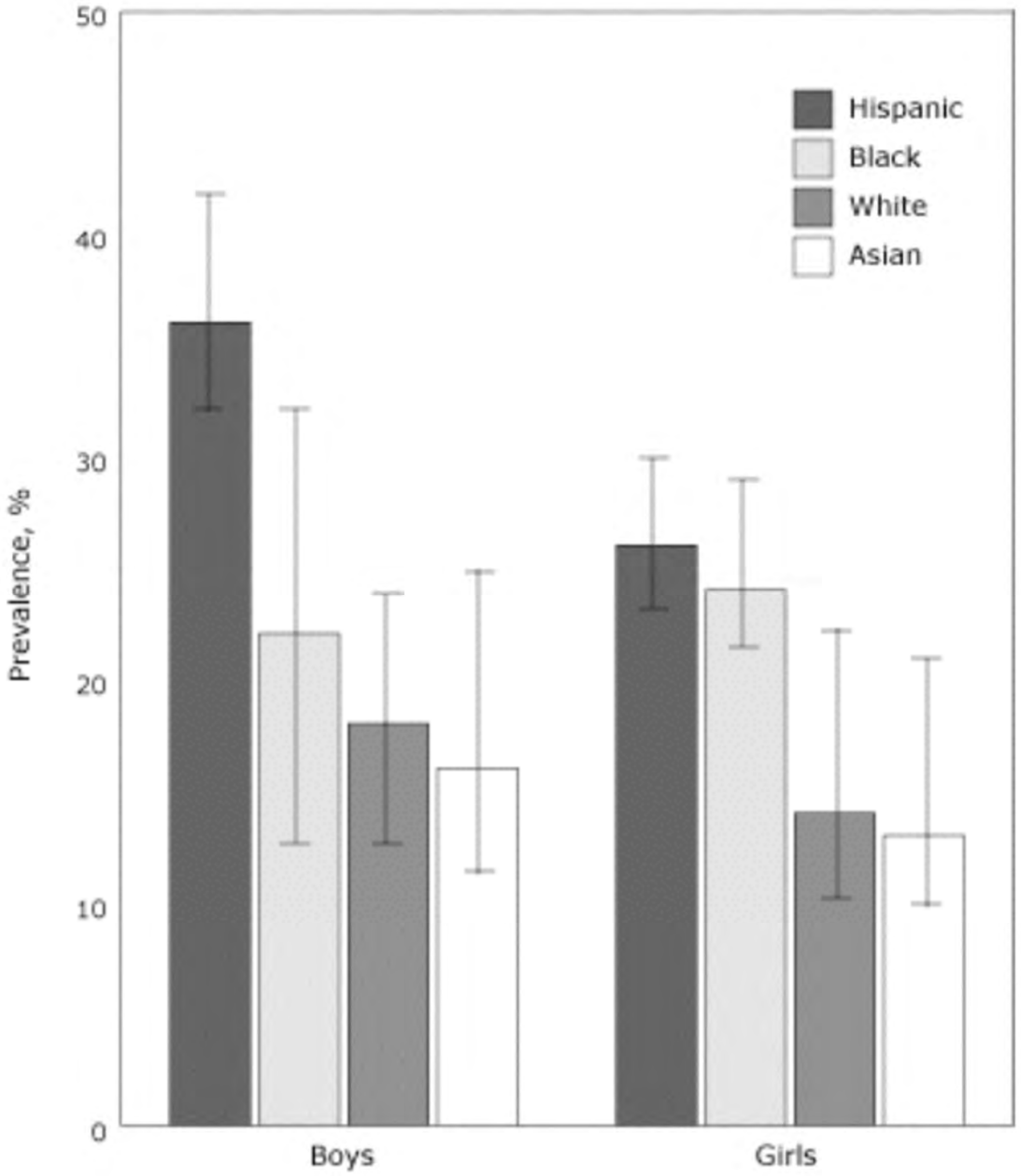
(Eq. 8-19)

If s^2 is the sample variance from a random sample of n observations from a normal distribution with unknown variance σ^2, then **a $100(1-\alpha)\%$ confidence interval on σ^2 is**

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \qquad (8\text{-}19)$$

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the upper and lower $100\alpha/2$ percentage points of the chi-square distribution with $n-1$ degrees of freedom, respectively. A **confidence interval for σ** has lower and upper limits that are the square roots of the corresponding limits in Equation 8-19.

Confidence estimates of the population proportion



Prevalence (with 95% CI bars) of obesity among New York City public elementary schoolchildren, by sex and race/ethnicity, 2003.

(source: CDC.GOV)

Collect a sample of BMI values

Obese means BMI > 30

What do those bars actually mean?

Large sample confidence estimate of population proportion

- Want to know the fraction *p* of the population that belongs to a class, e.g., the class "obese" kids defined by BMI>30.
- Each variable is a Bernoulli trial with one parameter p. We can use moments or MLE estimator to estimate p
- Both give the same estimate: sample fraction $\hat{p}$*=(# of obese kids in the sample)/(sample size n)*
- How to put confidence bounds on *p* based on $\hat{p}$
- # of obese kids in the sample follows the binomial distribution: "success" = sampled kid is obese : -(*p* – probability of success, *1-p* – failure
- Expected # of successes is *np* → Expected fraction of successes is *p*
- Standard deviation of # of successes is $\sqrt{np(1-p)}$→

Standard deviation of fraction of successes is $\sqrt{p(1-p)/n}$

8-5 A Large-Sample Confidence Interval For a Population Proportion

Normal Approximation for Binomial Proportion

If n is large, the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

is approximately standard normal.

The quantity $\sqrt{\hat{p}(1-\hat{p})/n}$ is the standard error of the point estimator $\hat{P}$.

8-5 A Large-Sample Confidence Interval For a Population Proportion (Eq. 8-23)

If $\hat{p}$ is the proportion of observations in a random sample of size n that belongs to a class of interest, an approximate $100(1-\alpha)\%$ confidence interval on the proportion p of the population that belongs to this class is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (8\text{-}23)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

This interval is known as the Wald interval (Wald and Wolfowitz, 1939).

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

http://www.scientificameriken.com/candy5.asp

"To our surprise M&Ms met our demand to review their procedures in determining candy ratios. It is, however, noted that the figures presented in their email differ from the information provided from their website (http://us.mms.com/us/about/products/milkchocolate/). An email was sent back informing them of this fact. To which M&Ms corrected themselves with one last email:

In response to your email regarding M&M'S CHOCOLATE CANDIES

Thank you for your email.
On average, our new mix of colors for M&M'S® Chocolate Candies is:

M&M'S® Milk Chocolate: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown.

M&M'S® Peanut: 23% blue, 23% orange, 15% green, 15% yellow, 12% red, 12% brown.

M&M'S® Kids MINIS®: 25% blue, 25% orange, 12% green, 13% yellow, 12% red, 13% brown.

M&M'S® Crispy: 17% blue, 16% orange, 16% green, 17% yellow, 17% red, 17% brown.

M&M'S® Peanut Butter and Almond: 20% blue, 20% orange, 20% green, 20% yellow, 10% red, 10% brown.

Have a great day!

Your Friends at Masterfoods USA
A Division of Mars, Incorporated

How to estimate these probabilities from a finite sample and how to set confidence interval on these estimates?

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

How large is a sample needed for 95% CI
on the percentage of blue M&Ms to be less than +/- 4%

Same question for red M&Ms?

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

How large is a sample needed for 95% CI
on the percentage of blue M&Ms to be less than +/- 4%

Same question for red M&Ms?

For blue M&Ms $p = 0.24$

$$1.96\sqrt{\frac{0.24(1-0.24)}{n}} < 0.04$$

$$n > \left(\frac{1.96}{0.04}\right)^2 0.24 \times (1-0.24) = 438 \text{ M\&Ms or}$$

$\sim$ 2 x 7oz bags with 210 candies each

For red M&Ms $p = 0.13$

$$n > \left(\frac{1.96}{0.04}\right)^2 \times 0.13 \times (1-0.13) \approx 271 \text{ M\&Ms or}$$

$\sim$ 1 x 7oz bag

Hypothesis testing: one sample

Is P53 gene expressed at a lower level in cancer patients than in healthy people?

- We are interested if a P53 gene expression is **lowered** in population of cancer patients compared to the healthy population.
- We know that mean gene expression in the healthy population is μ_h=50 mRNAs/cell We are interested in deciding whether or not the mean expression in cancer population is **lower than** in healthy population. Let's call hypothesis H_1. Here H_1 is **one-sided**
- If we asked: cancer is not equal to healthy H_1 would be a **two-sided hypothesis**
- Assume we have a sample of 100 cancer patients with sample mean $\bar{x}$ =48 mRNAs/cell and standard deviation σ=10 mRNA/cell
- Can we use our sample to reject the "business as usual" or **null hypothesis H_0**: cancer = healthy and select **one-sided hypothesis H_1**: cancer < healthy

Two types of errors

	decide H_0	decide H_1
true H_0 probability	Correct action $1-\alpha$	Type I error α
true H_1 probability	Type II error β	Correct action power $= 1-\beta$

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

Sometimes the **type I error probability α**
is called the **significance level,** or the **α-error**

Instructions: get α from your boss or PI (e.g., 5% or 1%)

Prob(H_0 is true given the sample data) < α
→reject H_0 and accept H_1

Prob(H_0 is true given the sample data) > α
→ accept H_0 and reject H_1

Type II error is much harder to estimate. Will deal with it later

P-Values of Hypothesis Tests

- P-value: what is the probability to get the observed value of sample mean of $\bar{x}$ =48 mRNAs/cell (or even smaller) and σ=10 mRNAs/cell in a healthy population with μ_h=50 mRNAs/cell

- If P-value is small – the null hypothesis is likely wrong and thus, the probability of making a type I error (incorrectly rejecting the null hypothesis) is small

- P-value answers the question: if I reject the null hypothesis H_0 based on the sample, what is the probability that I am making a type I error?

P-Value vs α in Hypothesis Testing

- Problem with using a predefined α: you don't know by how much you exceeded it
- Another approach is to calculate Prob(H_0 is true given the sample data) referred to as P-value.
  It the smallest α that would lead to rejection of null hypothesis
- You give your boss the P-value and let him/her decide if it is good enough
- Routinely with big datasets in genomics and systems biology P-values can be $10^{-\text{large number}\sim 10\text{-}100}$. This number is used to judge the quality of the hypothesis

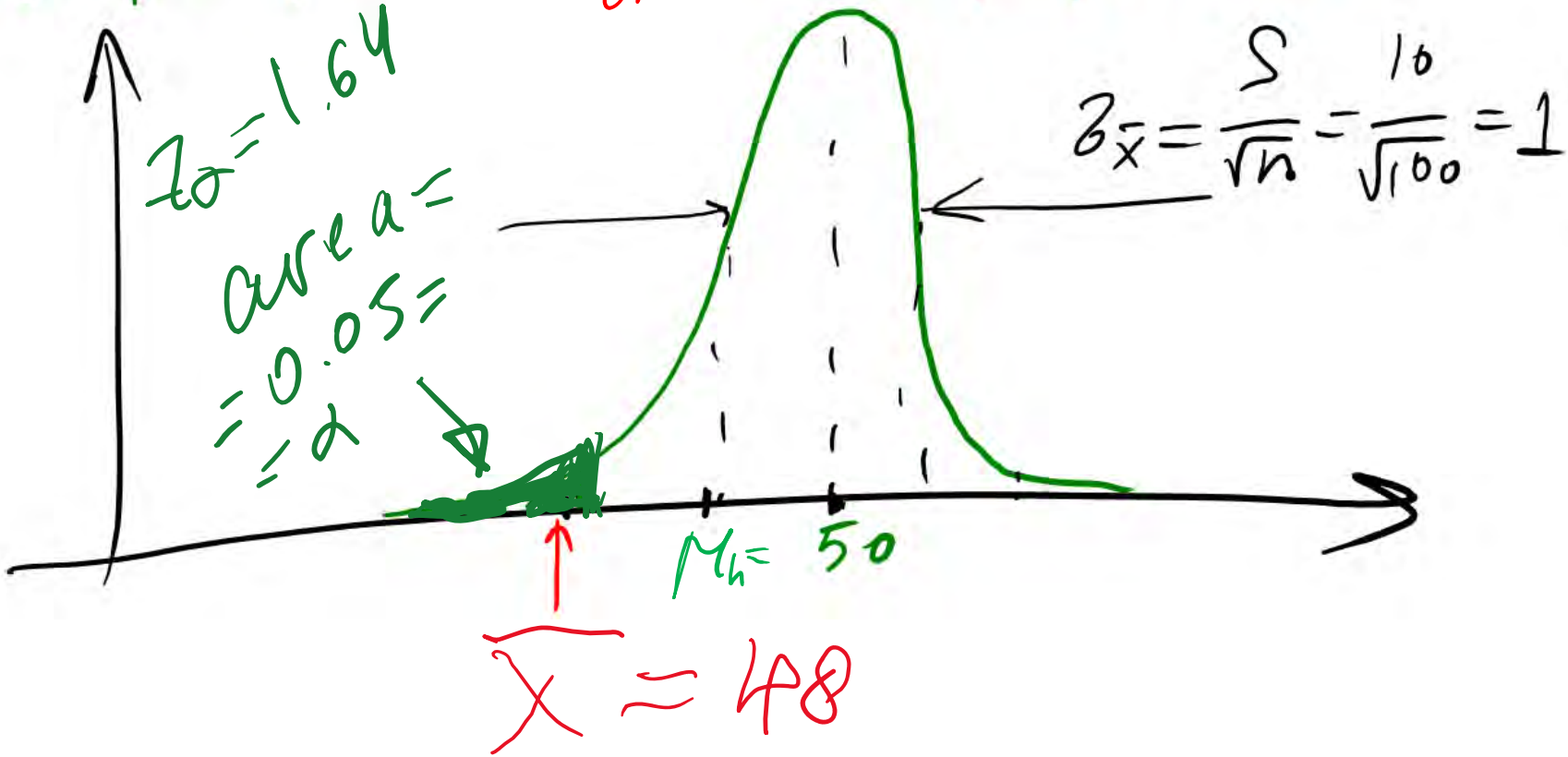
$\mu_h = 50$

$H_0: \mu_c = \mu_h$

$n = 100, \ \bar{X} = 48, \ S = 10$

One-sided hypothesis

$H_1: \mu_c < \mu_h$



$$P\text{-value} = \text{Prob}(\bar{X}_h < 48 \mid H_0) =$$

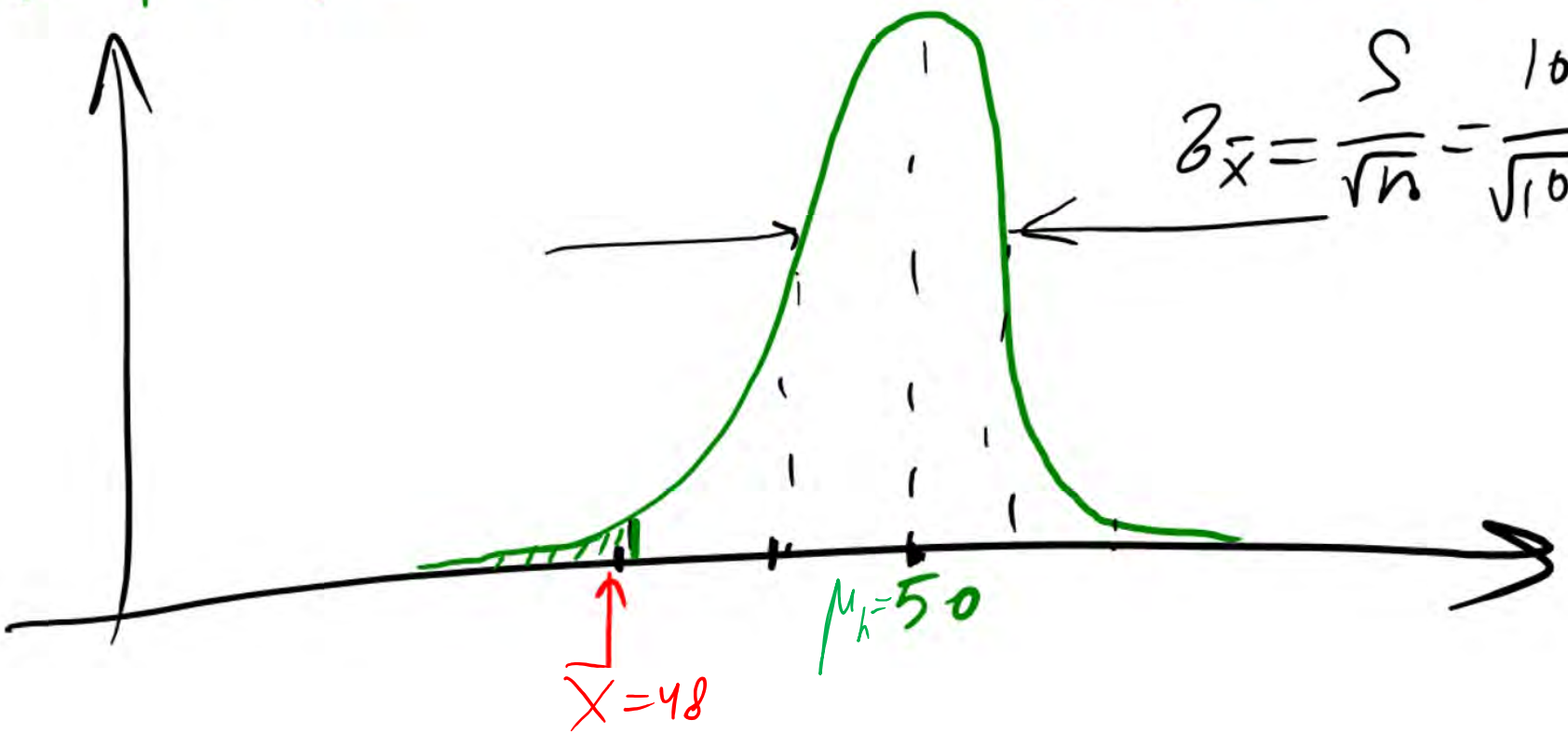
$$\cong 2.5\%$$

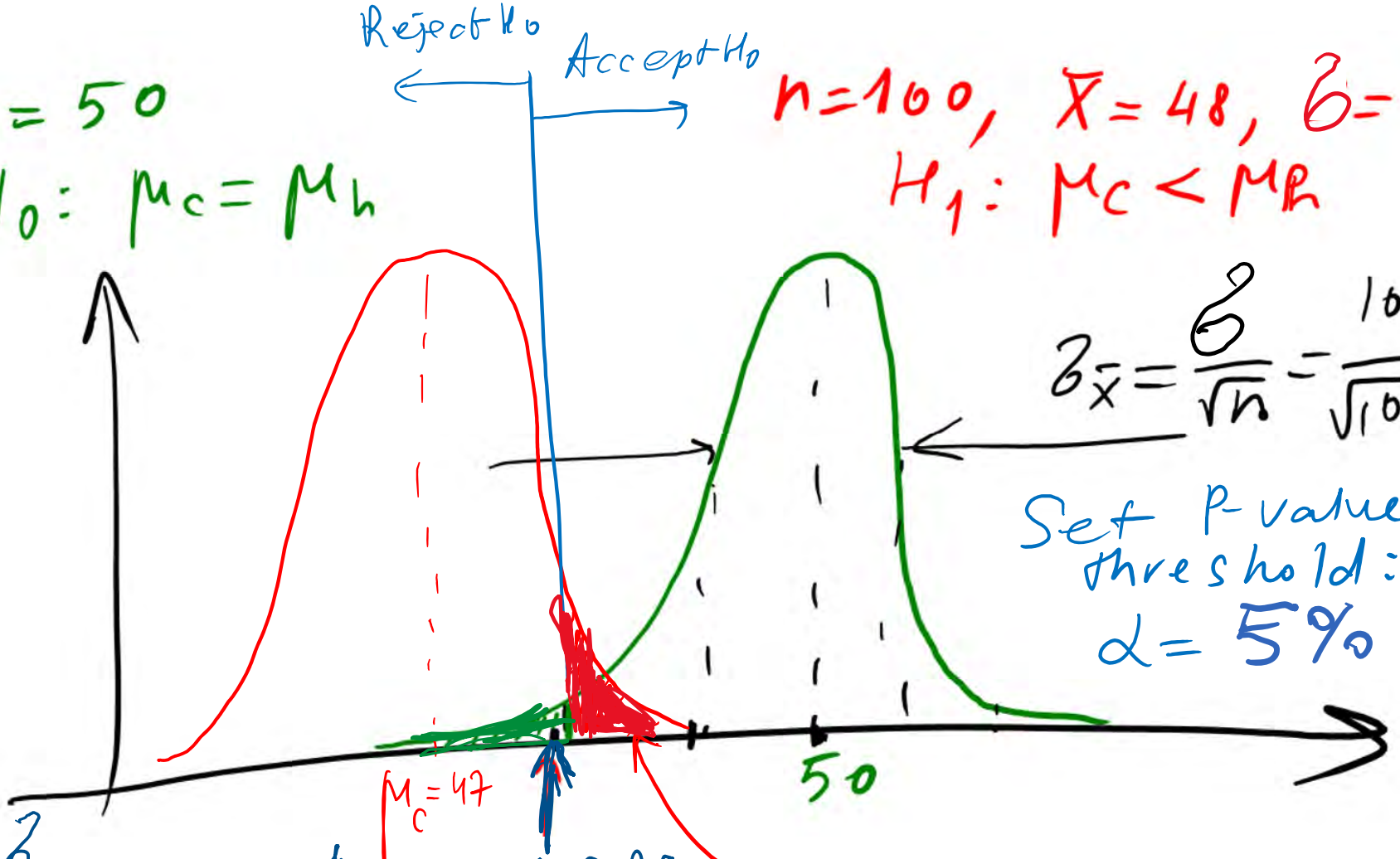
$\mu_h = 50$

$H_0: \mu_c = \mu_h$

$n = 100, \quad \bar{X} = 48, \quad S = 10$

$H_1: \mu_c < \mu_h$





$\mu_R = 50$

$H_0: \mu_c = \mu_R$

$n = 100, \ \bar{X} = 48, \ \sigma = 10$

$H_1: \mu_c < \mu_R$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$

Set P-value threshold: $\alpha = 5\%$

$\mu_R - Z_\alpha \frac{\sigma}{\sqrt{n}} = 50 \rightarrow 1.64 \rightarrow 48.36$

Type II error

$\beta = P(\text{Accept } H_0 \mid H_1 \text{ is true}) = \int_{48.36}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-47)^2}{2}\right) dx =$

$\alpha = 1 - \Phi(1.64) = 5\%$

$= 1 - \Phi(1.36) = 8.8\%$

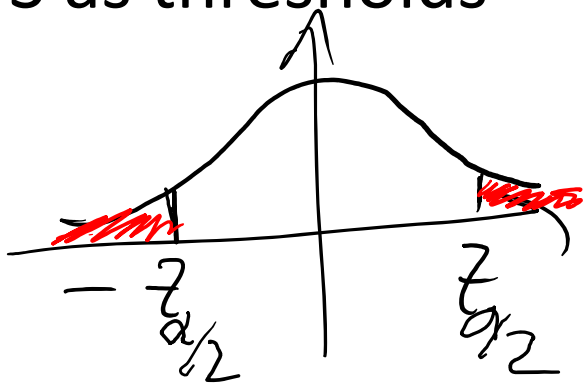
Generalizations

- What if H_1 is a two-sided hypothesis?
- A: P-value is $2(1-\Phi(|Z|))$, where $Z=(\bar{X}-\mu_0)/[S/\sqrt{n}]$

  Compare it to: For one sized $\mu_1>\mu_0$ it is $1-\Phi(Z)$

  For one sized $\mu_1<\mu_0$ it is $\Phi(Z)$
- If α is given, use μ_0 +/- $z_{\alpha/2}$*S as thresholds

to reject the null hypothesis



- What if the sample size n is small (say n<10):
- A: Use t-distribution with n-1 degrees of freedom for 2-sided $P\text{-}value=2(1-CDF\_Tdist(|T|))$ where $T=(\bar{X}-\mu_0)/[S/\sqrt{n}]$.
- For a given α use μ_0 +/- $t_{\alpha/2,n-1}$ T to reject the null hypothesis

Type II Error and Choice of Sample Size

Assume you know the minimum $\delta = |\mu_1 - \mu_0|$ that you care about.
What is the minimal sample you should use to separate H0 and H1 hypotheses if your tolerance to type I and type II errors is α and β ?

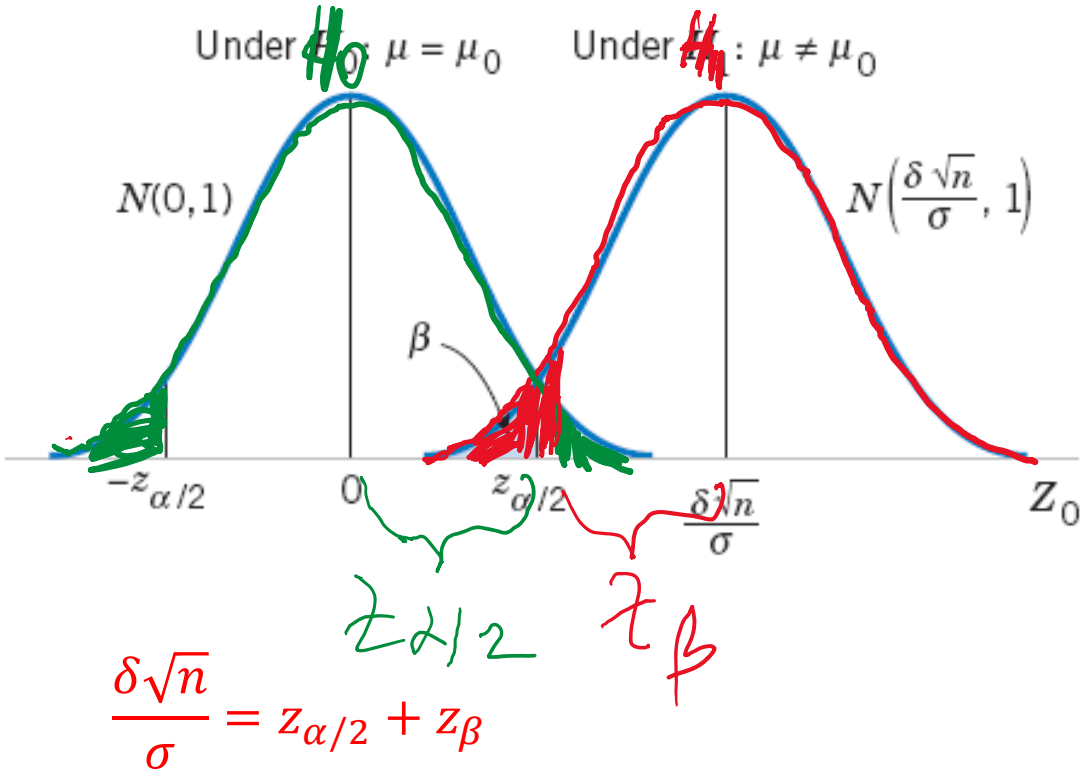


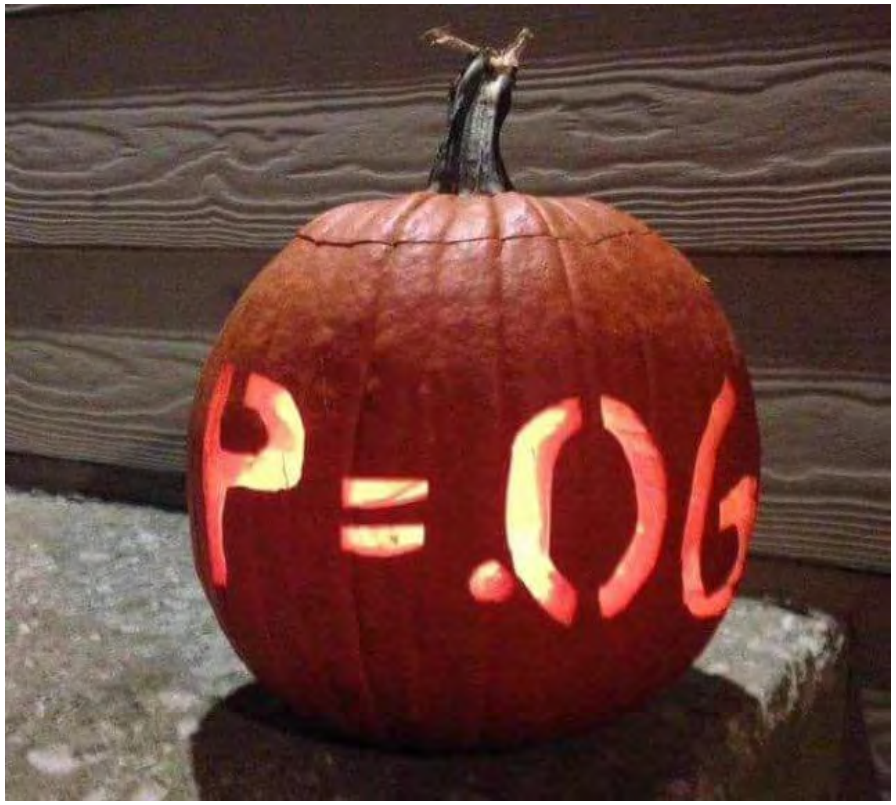
Figure 9-9 The distribution of Z_0 under H_0 and H_1.

$$n \simeq \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2} \quad \text{where} \quad \delta = \mu - \mu_0 \qquad (9\text{-}22)$$

Standard notation to indicate P-value with *, **, ***

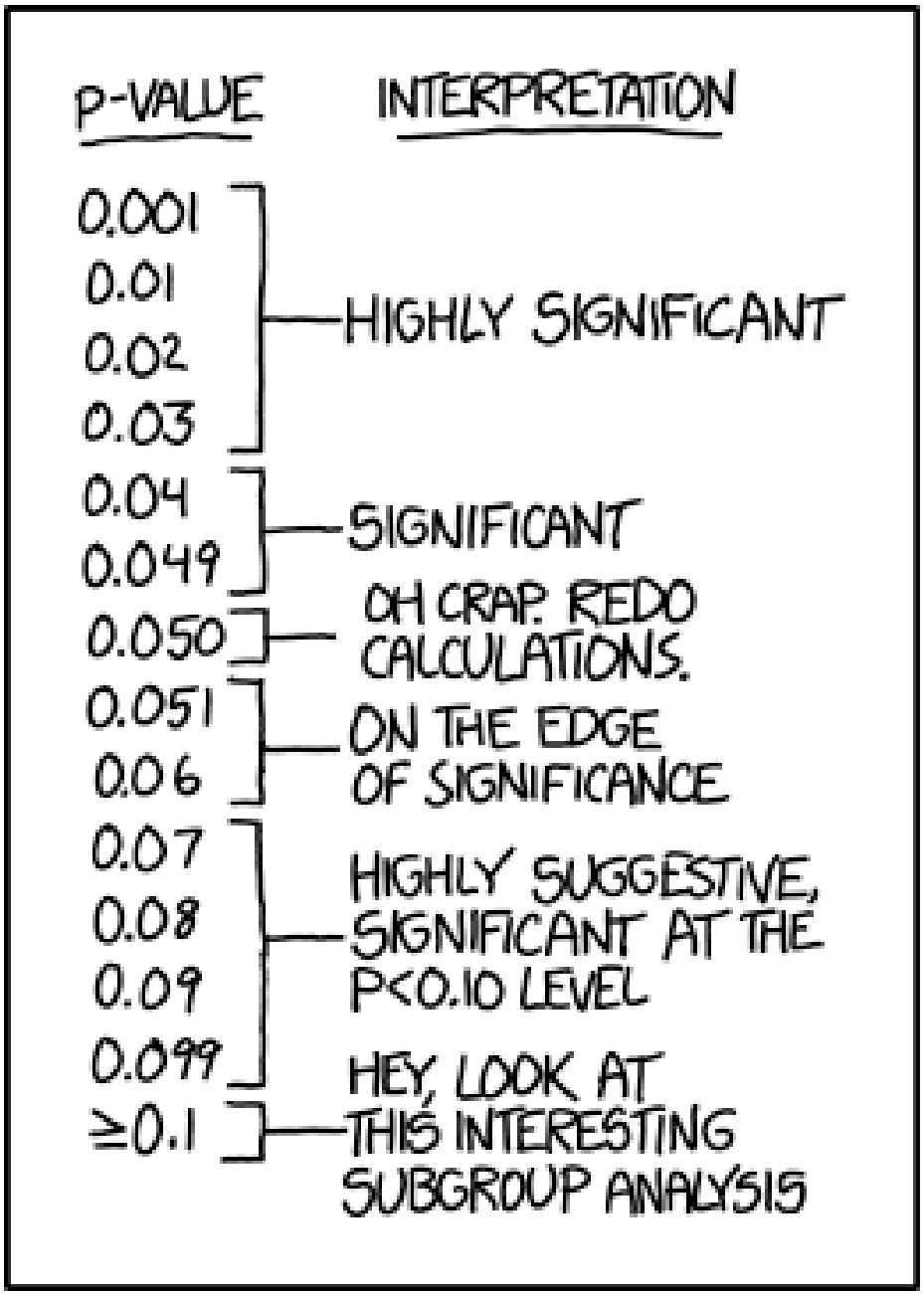
Table 11.1: A commonly adopted convention for reporting p values: in many places it is conventional to report one of four different things (e.g., $p < .05$) as shown below. I've included the "significance stars" notation (i.e., a * indicates $p < .05$) because you sometimes see this notation produced by statistical software. It's also worth noting that some people will write *n.s.* (not significant) rather than $p > .05$.

Usual notation	Signif. stars	English translation	The null is...
$p > .05$		The test wasn't significant	Retained
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.	Rejected
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$.	Rejected
$p < .001$	***	The test was significant at all levels	Rejected



Happy Halloween! (belated)

Credit: Trust me, I'm a "Biologist" Facebook community



Credit: XKCD comics

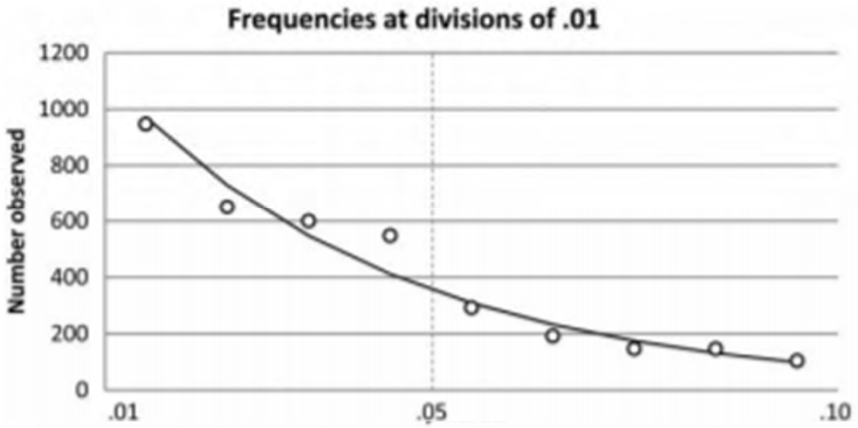
THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
2012, 65 (11), 2271–2279

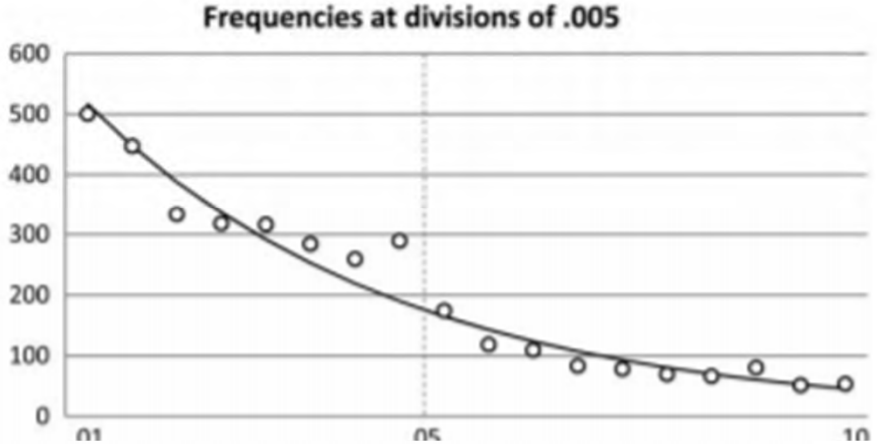
Psychology Press
Taylor & Francis Group

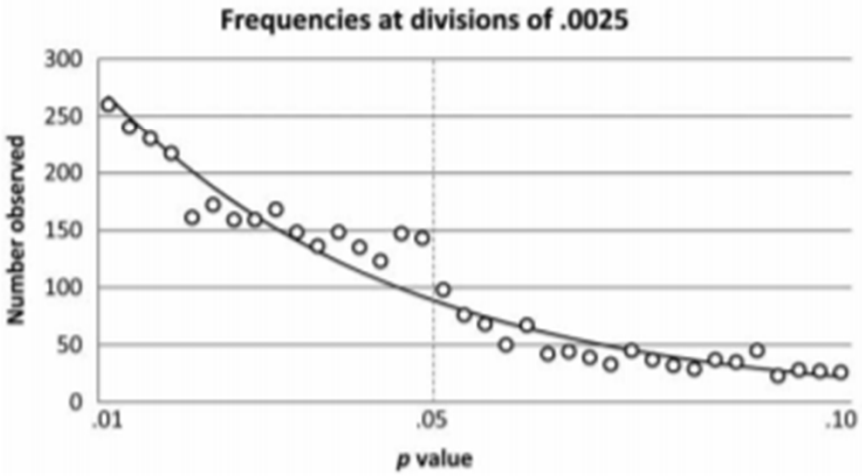
A peculiar prevalence of *p* values just below .05

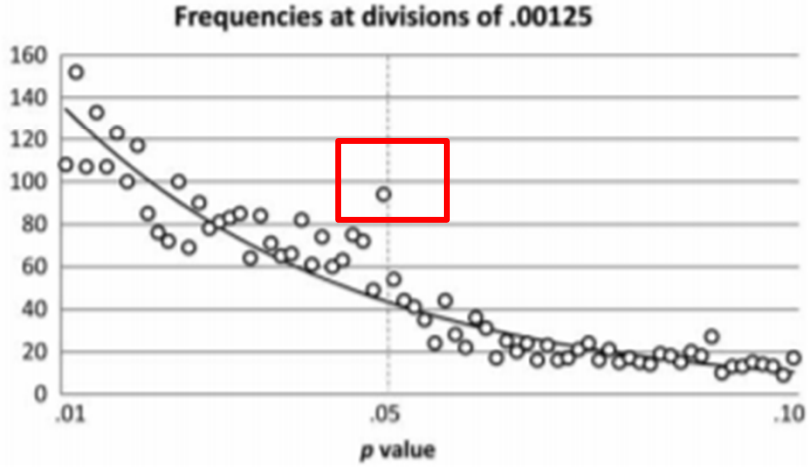
E. J. Masicampo[1], and Daniel R. Lalande[2]

[1]Department of Psychology, Wake Forest University, Winston-Salem, NC, USA
[2]Department of Health Sciences, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada







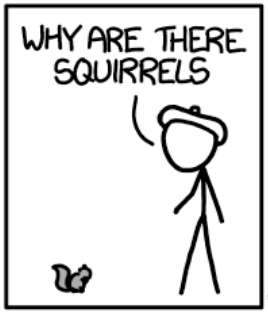


Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

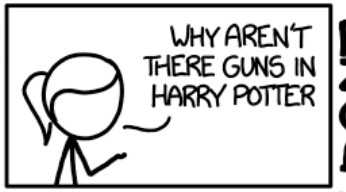
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND