



# Statistics for Health Care

---

## Unit 4: Overview/Teasers



# Overview

---

- Probability distributions; expected value and variance; the binomial and normal distributions



# Teaser 1, Unit 4

---

- A 2012 Mega Millions lottery had a jackpot of \$656 million (\$474 immediate payout).
- Question I received: "If the odds of winning the Mega millions is 1 in 175,000,000 is there a significant statistical advantage in playing 100 picks rather than one?"
- "For a half-billion dollars it almost seems worth it."



# Teaser 2, Unit 4

---

Imagine that you are in a resource-poor area and you want to screen the population for a fairly rare disease. But the antibody test is prohibitively expensive.

A clever cost-saving strategy is to pool the blood from multiple samples (using half of a person's blood sample and saving the other half). If the pooled lot is negative, this saves  $n-1$  tests. If it's positive, then you go back and test each sample individually, requiring  $n+1$  tests total.

If a particular disease has a prevalence of 10% in a population, will the pooling strategy save you tests? If so, what's the optimal number of samples to pool per lot?



## Teaser 3, Unit 4

---

- Ten patients with wrinkles were photographed before and after treatment with a new anti-aging treatment. An independent dermatologist was able to distinguish the pre and post photographs for 9 out of the 10 subjects.
- If the anti-aging treatment is completely ineffective, what's the probability that the dermatologist could have gotten at least 9 right purely by lucky guessing?



# Statistics for Health Care

---

## Module 1: Probability distributions (functions)



# Probability function

---

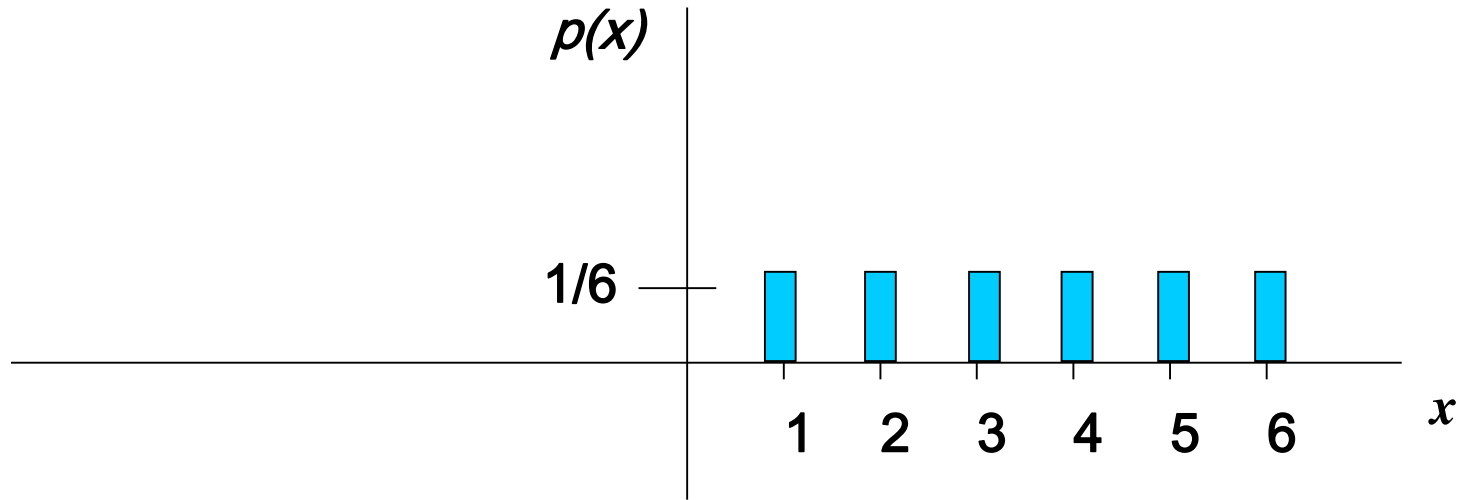
- Gives the probabilities of all possible outcomes.
- A mathematical function that maps each possible outcome  $x$  to its probability  $p(x)$ .
- The probabilities must sum (or integrate) to 1.0.

# Probability functions can be discrete or continuous

- **Discrete:** can only take on certain values
  - Examples: Dead/alive, treatment/placebo, dice, whole numbers, counts, etc.
- **Continuous:** can theoretically take on any value within a given range (has an infinite continuum of possible values).
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6



# Discrete example: roll of a die



$$\sum_{\text{all } x} P(x) = 1$$

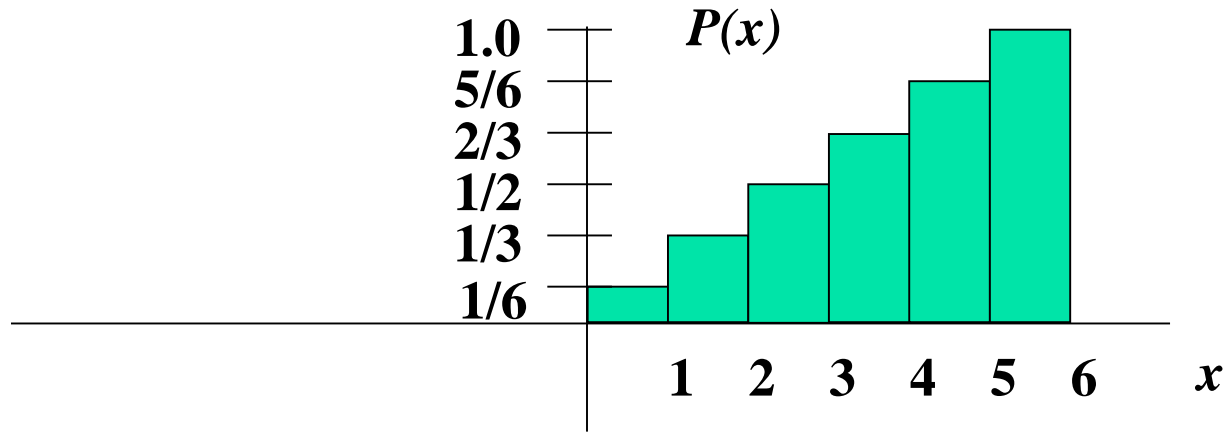


# Probability mass function (pmf)


| $x$ | $p(x)$                         |
|-----|--------------------------------|
| 1   | $p(x=1)=1/6$                   |
| 2   | $p(x=2)=1/6$                   |
| 3   | $p(x=3)=1/6$                   |
| 4   | $p(x=4)=1/6$                   |
| 5   | $p(x=5)=1/6$                   |
| 6   | <u><math>p(x=6)=1/6</math></u> |

1.0

# Cumulative distribution function (CDF)



# Cumulative distribution function for a die



| $x$ | $P(x \leq A)$       |
|-----|---------------------|
| 1   | $P(x \leq 1) = 1/6$ |
| 2   | $P(x \leq 2) = 2/6$ |
| 3   | $P(x \leq 3) = 3/6$ |
| 4   | $P(x \leq 4) = 4/6$ |
| 5   | $P(x \leq 5) = 5/6$ |
| 6   | $P(x \leq 6) = 6/6$ |



# Recall: blood types

***Out of 100 donors . . . . .***

| <b>84 donors are RH+</b> | <b>16 donors are RH-</b> |
|--------------------------|--------------------------|
| <b>38 are O+</b>         | <b>7 are O-</b>          |
| <b>34 are A+</b>         | <b>6 are A-</b>          |
| <b>9 are B+</b>          | <b>2 are B-</b>          |
| <b>3 are AB+</b>         | <b>1 is AB-</b>          |

Source: [AABB.ORG](http://AABB.ORG)

# Probability distribution for blood types (discrete function):

| <u><math>x</math></u> | <u><math>P(x)</math></u> |
|-----------------------|--------------------------|
| <b>0</b>              | <b>45%</b>               |
| <b>A</b>              | <b>40%</b>               |
| <b>B</b>              | <b>11%</b>               |
| <b>AB</b>             | <b>4%</b>                |



# Practice Problem:

- The number of patients seen in the ER in any given hour is a random variable represented by  $x$ . The probability distribution for  $x$  is:

|        |    |    |    |    |    |
|--------|----|----|----|----|----|
| $x$    | 9  | 10 | 11 | 12 | 13 |
| $P(x)$ | .3 | .3 | .2 | .1 | .1 |

Find the probability that in a given hour:

- exactly 13 patients arrive  $p(x=13) = .1$
- At least 12 patients arrive  $p(x \geq 12) = (.1 + .1) = .2$
- At most 11 patients arrive  $p(x \leq 11) = (.3 + .3 + .2) = .80$



# Important discrete distributions in medical research:

---

- Binomial

- Yes/no outcomes (dead/alive, treated/untreated, smoker/non-smoker, sick/well, etc.)

- Poisson

- Counts (e.g., how many cases of disease in a given area)





# Continuous case

---

- Any continuous mathematical function that integrates to 1 is a probability function.
- The probabilities associated with continuous functions are just areas under the curve (integrals!).
- Probabilities are given for a range of values, rather than a particular value.



# Continuous case

---

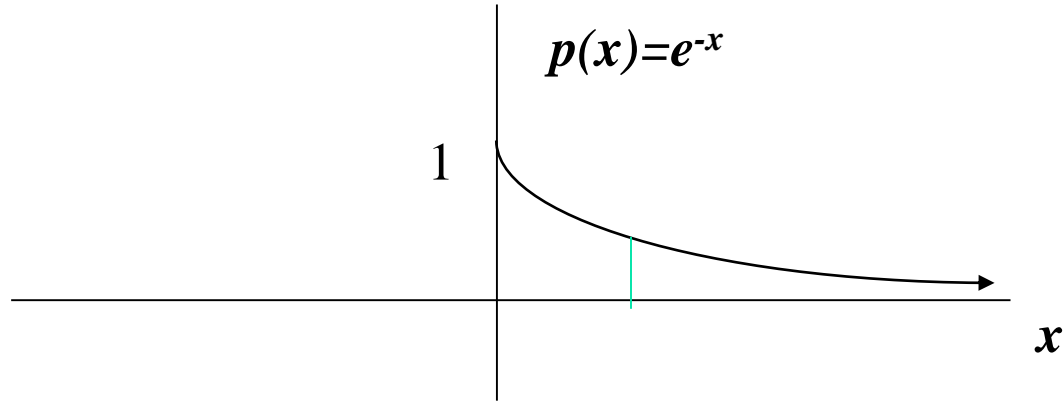
- For example, the exponential distribution is a continuous probability function, because the area under the curve is 1.0.

$$f(x) = e^{-x}$$

- This function integrates to 1 :

$$\int_0^{+\infty} e^{-x} = -e^{-x} \Big|_0^{+\infty} = 0 + 1 = 1$$

# Continuous case: "probability density function" (pdf)

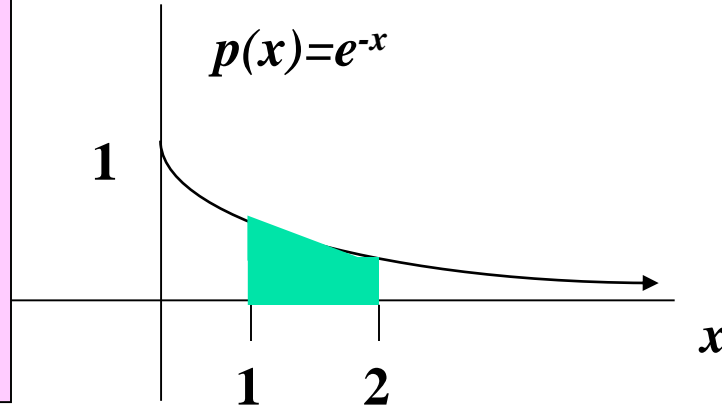


The probability that  $x$  is any exact particular value (such as 1.9976) is 0; we can only assign probabilities to possible ranges of  $x$  (=cumulative distribution function).

# Continuous case: Cumulative distribution function (CDF)

Clinical example: Imagine that survival times after lung transplant roughly follow an exponential function.

Then, the probability that a patient will die in the second year after surgery (between years 1 and 2) is 23%.



**The integral**

$$P(1 \leq x \leq 2) = \int_1^2 e^{-x} = -e^{-x} \Big|_1^2 = -e^{-2} - (-e^{-1}) = -.135 + .368 = .23$$

# Cumulative distribution function

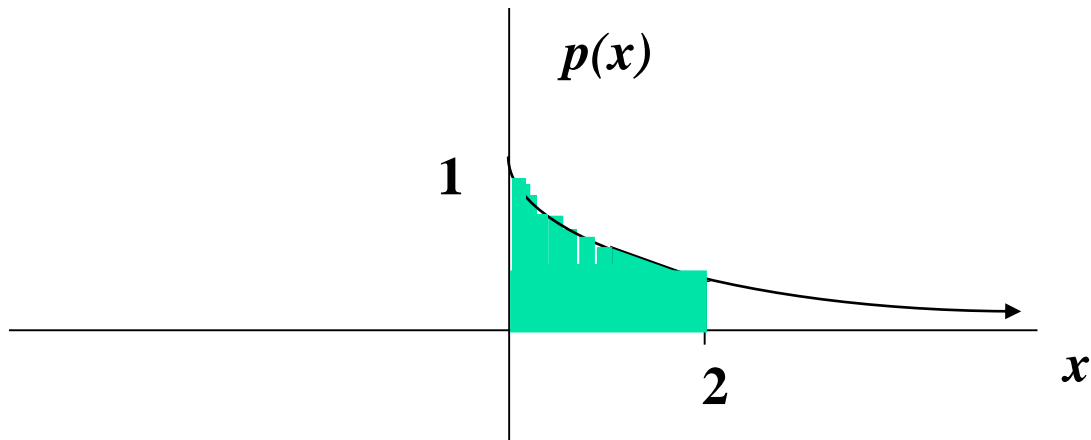


As in the discrete case, we can specify the “cumulative distribution function” (CDF):

The CDF here =  $P(x \leq A) =$

$$\int_0^A e^{-x} = -e^{-x} \Big|_0^A = -e^{-A} - (-e^0) = -e^{-A} + 1 = 1 - e^{-A}$$

# Example



$$P(x \leq 2) = 1 - e^{-2} = 1 - .135 = .865$$



# Practice Problem

---

Suppose that survival drops off rapidly in the year following diagnosis of a certain type of advanced cancer. Suppose that the length of survival (or time-to-death) is a random variable that approximately follows an exponential distribution with parameter 2 (makes it a steeper drop off):

$$\text{probability function : } p(x = T) = 2e^{-2T}$$

$$\text{note : } \int_0^{+\infty} 2e^{-2x} = -e^{-2x} \Big|_0^{+\infty} = 0 + 1 = 1$$

**What's the probability that a person who is diagnosed with this illness survives a year?**



# Answer

---

The probability of dying within 1 year can be calculated using the cumulative distribution function:

Cumulative distribution function is:

$$P(x \leq T) = -e^{-2x} \Big|_0^T = 1 - e^{-2(T)}$$

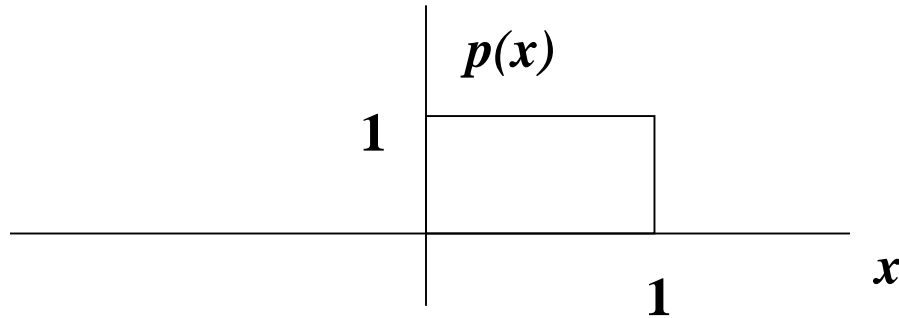
The chance of surviving past 1 year is:  $P(x \geq 1) = 1 - P(x \leq 1)$

$$1 - (1 - e^{-2(1)}) = .135$$



# Example 2: Uniform distribution

The uniform distribution: all values are equally likely.  
 $f(x) = 1$ , for  $1 \geq x \geq 0$

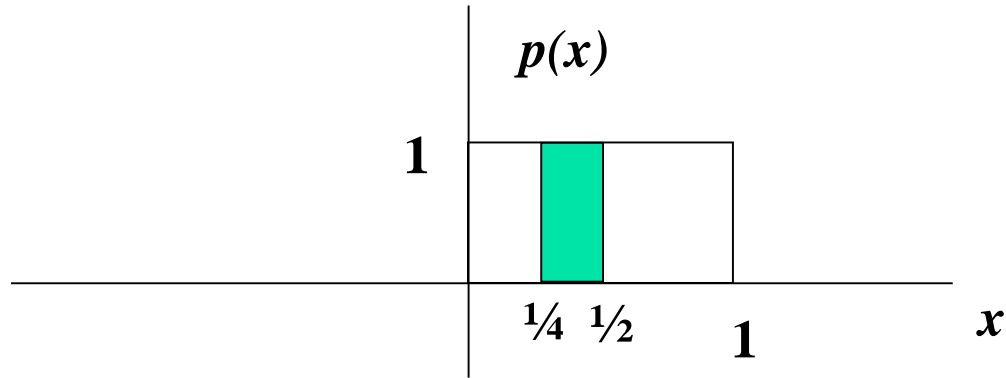


We can see it's a probability distribution because the area under the curve is 1:

$$\int_0^1 1 = x \Big|_0^1 = 1 - 0 = 1$$

# Example: Uniform distribution

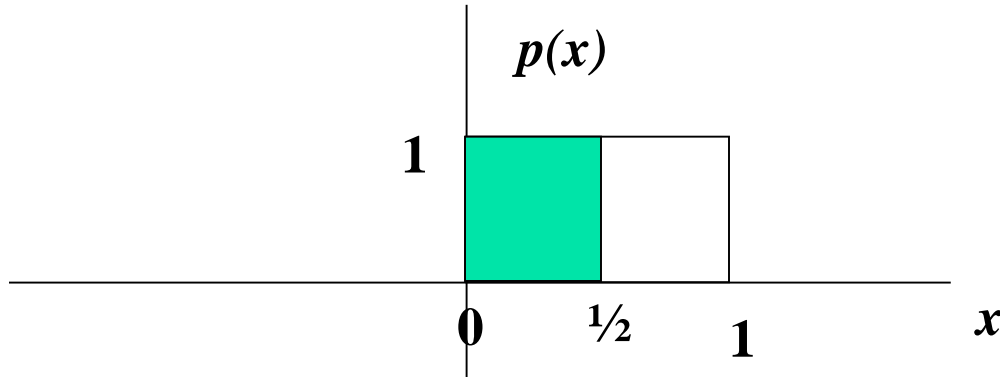
What's the probability that  $x$  is between  $1/4$  and  $1/2$ ?



$$P(1/2 \geq x \geq 1/4) = 1/4$$

# Example: Uniform distribution

What's the probability that  $x$  is between 0 and  $1/2$ ?



$$P(1/2 \geq x \geq 0) = 1/2$$

Clinical Research Example:  
When randomizing patients in an RCT, we often use a random number generator on the computer. These programs work by randomly generating a number between 0 and 1 (with equal probability of every number in between). Then a subject who gets  $X < .5$  is control and a subject who gets  $X > .5$  is treatment.



# Expected value and Variance

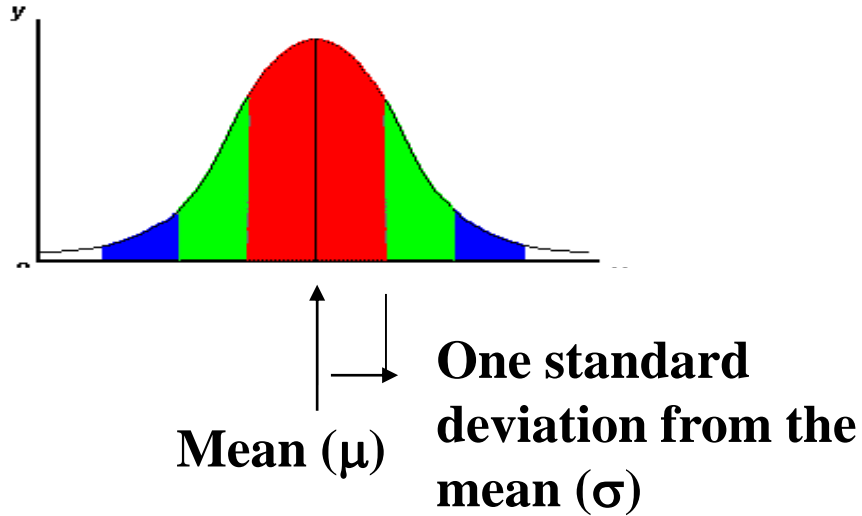
---

- All probability distributions are characterized by an expected value (mean) and a variance (standard deviation squared).



**For example, bell-curve (normal) distribution:**

---





# Statistics for Health Care

---

## Module 2: Expected value



# Expected value

---

- Expected value is just the mean ( $\mu$ ) of a probability distribution.
- It is a weighted average, calculated by weighting the value of each possible outcome by its probability.
- Expected value helps us make informed decisions based on how we expect  $x$  to behave on-average over the long-run.



# Expected value, formally

**Discrete case:**

$$E(X) = \sum_{\text{all } x} x_i p(x_i)$$

**Continuous case:**

$$E(X) = \int_{\text{all } x} x_i p(x_i) dx$$

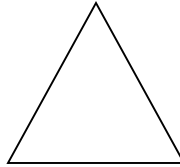




# Example: expected value

- Recall the following probability distribution of ER arrivals:

|        |    |    |    |    |    |
|--------|----|----|----|----|----|
| $x$    | 9  | 10 | 11 | 12 | 13 |
| $P(x)$ | .3 | .3 | .2 | .1 | .1 |



$$\sum_{i=1}^5 x_i p(x) = 9(.3) + 10(.3) + 11(.2) + 12(.1) + 13(.1) = 10.4$$

# A Sample Mean is a special case of Expected Value...

Sample mean, for a sample of  $n$  subjects: =

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \sum_{i=1}^n x_i \left( \frac{1}{n} \right)$$

**The probability (frequency) of each person in the sample is  $1/n$ .**



# Symbol Interlude

---

- $E(X) = \mu$ 
  - these symbols are used interchangeably



# Expected Value

---

- Expected value is an extremely useful concept for good decision-making!



# Example: the lottery

---

- The Lottery (also known as a tax on people who are bad at math...)
- A certain lottery works by picking 6 numbers from 1 to 49. It costs \$1.00 to play the lottery, and if you win, you win \$2 million after taxes.

*If you play the lottery once, what are your expected winnings or losses?*



# Lottery

**Calculate the probability of winning in 1 try:**

$$\frac{1}{\binom{49}{6}} = \frac{1}{\frac{49!}{43!6!}} = \frac{1}{13,983,816} = 7.2 \times 10^{-8}$$

“49 choose 6”

Out of 49 numbers,  
this is the number  
of distinct  
combinations of 6.

**The probability function (note, sums to 1.0):**

| $x$         | $p(x)$               |
|-------------|----------------------|
| -1          | .999999928           |
| + 2 million | $7.2 \times 10^{-8}$ |



# Expected Value

## The probability function

| $x$         | $p(x)$               |
|-------------|----------------------|
| -1          | .999999928           |
| + 2 million | $7.2 \times 10^{-8}$ |

## Expected Value

$$\begin{aligned} E(X) &= P(\text{win}) * \$2,000,000 + P(\text{lose}) * -\$1.00 \\ &= 2.0 \times 10^6 * 7.2 \times 10^{-8} + .999999928 (-1) = .144 - .999999928 = -\$ .86 \end{aligned}$$

Negative expected value is never good!

You shouldn't play if you expect to lose money!



# Expected Value

---

**If you play the lottery every week for 10 years, what are your expected winnings or losses?**

$$520 \times (-.86) = -\$447.20$$



# 2012 record Mega Millions jackpot...



---

- 2012 Mega Millions had a jackpot of \$656 million (\$474 immediate payout).
- Question I received: "If the odds of winning the Mega millions is 1 in 175,000,000 is there a significant statistical advantage in playing 100 quick picks rather than one?"
- "For a half-billion-with-a-B dollars it almost seems worth it."



# Expected value for 1 ticket:

- Chances of losing, 1 ticket:

$$1 - 1/175,000,000 = 99.99999994\%$$

| $x\$$         | $p(x)$                               |
|---------------|--------------------------------------|
| -1            | . 9999999994                         |
| + 500 million | <b><math>6 \times 10^{-9}</math></b> |

## Expected Value

$$\begin{aligned} E(X) &= P(\text{win}) * \$500,000,000 + P(\text{lose}) * -\$1.00 \\ &= 6.0 \times 10^{-9} * 500,000,000 + .9999999994 (-1) = +2 \end{aligned}$$



# Answer, 100 tickets:

- Chances of losing, 100 tickets:  
99.999943%

| $x\$$         | $p(x)$                                 |
|---------------|--|
| -100          | . 99999943                             |
| + 500 million | <b><math>5.7 \times 10^{-7}</math></b> |

## Expected Value

$$\begin{aligned} E(X) &= P(\text{win}) * \$500,000,000 + P(\text{lose}) * -\$100 \\ &= 5.7 \times 10^{-7} * 500,000,000 + .99999943 (-100) = +185 \end{aligned}$$



# So...

---

- One could make a case for playing!
- You can work out that the expected payout only has to be  $> \$176$  million for expected value to be positive (for either 1 ticket or 100 tickets).
- BUT...



# BUT then consider the high chance of multiple winners!

- When the jackpot is huge, lots of people play. The chance of multiple winners (who will share the jackpot) is quite high!
- Assume 600 million tickets are sold, then the probability distribution here is (where  $x$  is the number of winners):

| x    | 0   | 1   | 2   | 3   | 4   | 5   | 6   | 7    | 8    | 9    | 10   |
|------|-----|-----|-----|-----|-----|-----|-----|------|------|------|------|
| P(x) | .03 | .11 | .19 | .22 | .19 | .13 | .07 | .036 | .015 | .005 | .002 |

- $E(X) = 0 + 1 \cdot .11 + 2 \cdot .19 + 3 \cdot .22 + 4 \cdot .19 + 5 \cdot .13 + 6 \cdot .07 + 7 \cdot .036 + 8 \cdot .015 + 9 \cdot .005 + 10 \cdot .002 = 3.4$
- Therefore, the expected winnings if you win are actually  $500 \text{ million} / 3.4 = \mathbf{147 \text{ million}}$



# Not to mention taxes!

---

- You can also assume that about half is going to be lost in taxes.
- And, the fact is, you're still going to lose with almost near certainty!
  - Probability 99.9999...%!



# Gambling (or how casinos can afford to give so many free drinks...)

---

A roulette wheel has the numbers 1 through 36, as well as 0 and 00. If you bet \$1 that an odd number comes up, you win or lose \$1 according to whether or not that event occurs. If random variable  $X$  denotes your net gain,  $X=1$  with probability  $18/38$  and  $X=-1$  with probability  $20/38$ .

$$E(X) = 1(18/38) - 1(20/38) = -\$0.053$$

∴ On average, the casino wins (and the player loses) 5 cents per game.

The casino rakes in even more if the stakes are higher:

$$E(X) = 10(18/38) - 10(20/38) = -\$0.53$$

If the cost is \$10 per game, the casino wins an average of 53 cents per game. If 10,000 games are played in a night, that's a cool \$5300.



# Challenge Problem

---

- Imagine that you are in a resource-poor area and you want to screen the population for a fairly rare disease. But the antibody test is prohibitively expensive.
- A clever cost-saving strategy is to pool the blood from multiple samples (using half of a person's blood sample and saving the other half). If the pooled lot is negative, this saves  $n-1$  tests. If it's positive, then you go back and test each sample individually, requiring  $n+1$  tests total.
- If a particular disease has a prevalence of 10% in a population, will the pooling strategy save you tests? If so, what's the optimal number of samples to pool per lot?
- Solve by "brute force" assuming you want to screen 100 people.





# Try pooling 20...

---

If you pool 20 samples at a time (5 lots), how many tests do you expect to have to run (assuming the test is perfect!)?



# Pooling 20...

---

If you pool 20 samples at a time (5 lots), how many tests do you expect to have to run (assuming the test is perfect!)?

$X$  = the number of tests you have to run per lot:

$$E(X) = P(\text{pooled lot is negative})(1) + P(\text{pooled lot is positive})(21)$$

$$E(X) = (.90)^{20} (1) + [1 - .90^{20}] (21) = 12.2\% (1) + 87.8\% (21) = 18.56$$

$$E(\text{total number of tests}) = 5 * 18.56 = 92.8$$



# Pooling 10...

---

What if you pool only 10 samples at a time?

$$E(X) = (.90)^{10} (1) + [1 - .90^{10}] (11) = 35\% (1) + 65\% (11) = 7.5$$

average per lot

$$10 \text{ lots} * 7.5 = 75$$



# Pooling 5...

---

5 samples at a time?

$$E(X) = (.90)^5 (1) + [1-.90^5] (6) = 59\% (1) + 41\% (6) = 3.05 \text{ average per lot}$$

$$20 \text{ lots} * 3.05 = 61$$



# Pooling 4...

---

4 samples at a time?

$$E(X) = (.90)^4 (1) + [1 - (.90)^4] (5) = 2.38 \text{ average per lot}$$

$$25 \text{ lots} * 2.38 = 59$$



# Pooling 3...

---

3 samples at a time?

$$E(X) = (.90)^3 (1) + [1 - .90^3] (4) = 1.81 \text{ average per lot}$$

$$33 \text{ lots} * 1.81 = 60$$



# Extension to continuous case:

---

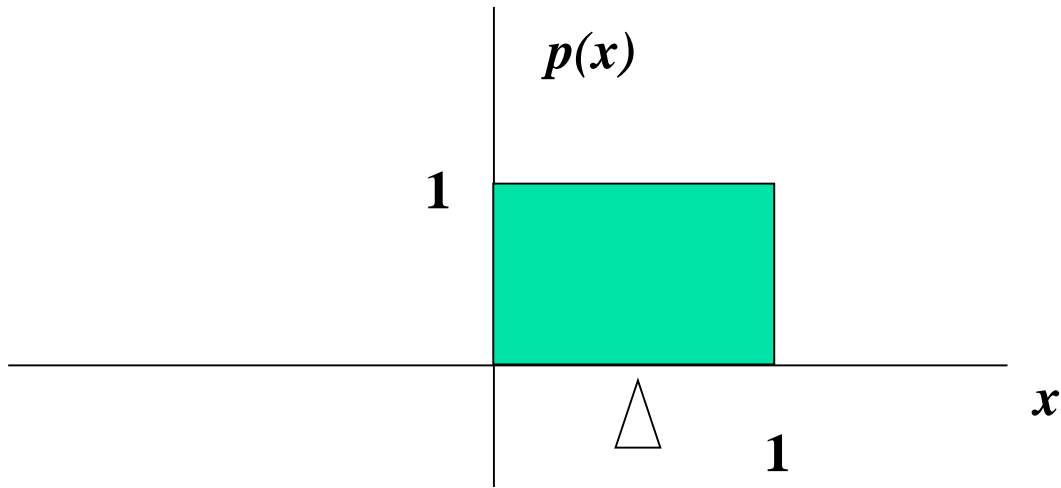
**Discrete case:**

$$E(X) = \sum_{\text{all } x} x_i p(x_i)$$

**Continuous case:**

$$E(X) = \int_{\text{all } x} x_i p(x_i) dx$$

# Extension to continuous case: uniform distribution



**Calculus:**

$$E(X) = \int_0^1 x(1)dx = \frac{x^2}{2} \Big|_0^1 = \frac{1}{2} - 0 = \frac{1}{2}$$





# Statistics for Health Care

---

## Module 3: Variance



# Variance or standard deviation

---

The variance (or standard deviation) quantifies the variability of a probability distribution.

Variance/standard deviation is calculated similarly to how we calculated variance/standard deviation for sample data. However, outcomes are weighted by their probabilities.



# Variance

---

Variance=the average squared distance from the mean

$$\sigma^2 = Var(x) = \sum_{\text{all } x} (x_i - \mu)^2 p(x_i)$$



# Example: variance

Find the variance and standard deviation for the number of patients to arrive in the ER (recall that the mean is 10.4).

|        |    |    |    |    |    |
|--------|----|----|----|----|----|
| $x$    | 9  | 10 | 11 | 12 | 13 |
| $P(x)$ | .3 | .3 | .2 | .1 | .1 |

$$Var(x) = \sum_{i=1}^5 (x_i - 10.4)^2 * p(x_i) = (1.4^2)(.3) + (0.4^2)(.3) + (0.6^2)(.2) + (1.6^2)(.1) + (2.6^2)(.1) = 1.64$$

$$SD(x) = \sqrt{1.64} = 1.28$$

**Interpretation: In an average hour, we expect 10.4 patients to arrive in the ER, plus or minus 1.28. This gives you a feel for what would be considered a typical hour!**



# Variance, formally

---

**Discrete case:**

$$Var(X) = \sum_{\text{all } x} (x_i - \mu)^2 p(x_i)$$


**Continuous case:**

$$Var(X) = \int_{-\infty}^{\infty} (x_i - \mu)^2 p(x_i) dx$$



# Similarity to empirical variance

The variance of a sample:  $S^2 =$

$$\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{n - 1} = \sum_{i=1}^N (x_i - \bar{x})^2 \left( \frac{1}{n - 1} \right)$$


Division by  $n-1$  reflects the fact that we have lost a "degree of freedom" (piece of information) because we had to estimate the sample mean before we could estimate the sample variance.



# Practice Problem

---

A roulette wheel has the numbers 1 through 36, as well as 0 and 00. If you bet \$1.00 that an odd number comes up, you win or lose \$1.00 according to whether or not that event occurs. If  $X$  denotes your net gain,  $X=1$  with probability  $18/38$  and  $X=-1$  with probability  $20/38$ .

We already calculated the mean to be  $= -\$0.053$ .

What are the variance and standard deviation of  $X$ ?



# Answer

---

$$\begin{aligned}\sigma^2 &= \sum_{\text{all } x} (x_i - \mu)^2 p(x_i) \\&= (+1 - -.053)^2 (18/38) + (-1 - -.053)^2 (20/38) \\&= (1.053)^2 (18/38) + (-1 + .053)^2 (20/38) \\&= (1.053)^2 (18/38) + (-.947)^2 (20/38) \\&= .997 \\ \sigma &= \sqrt{.997} = .99\end{aligned}$$

Standard deviation is \$.99. Interpretation: On average, you're either 1 dollar above or 1 dollar below the mean, which is just under zero. Makes sense!






## \*\*A few notes about Variance as a mathematical operator:

---

- If  $c$  = a constant number (i.e., not a variable) and  $X$  and  $Y$  are random variables, then  $\text{Var}(c) = 0$
- $\text{Var}(c+X) = \text{Var}(X)$
- $\text{Var}(cX) = c^2\text{Var}(X)$
- $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$  ***ONLY IF  $X$  and  $Y$  are independent!!!!***


$$\text{Var}(c) = 0$$

---

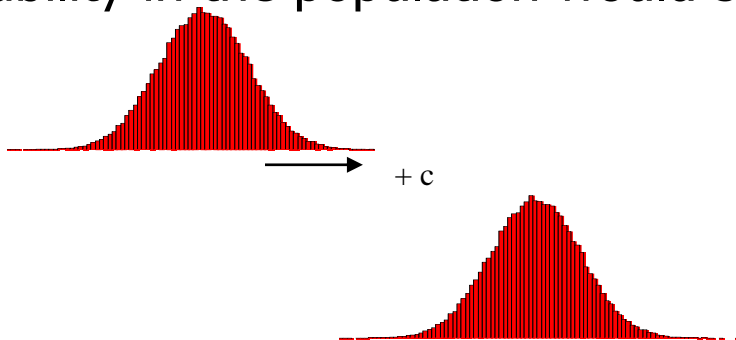
$$\text{Var}(c) = 0$$

Constants don't vary!


$$\text{Var}(c+X) = \text{Var}(X)$$

$$\text{Var}(c+X) = \text{Var}(X)$$

Adding a constant to every instance of a random variable doesn't change the variability. It just shifts the whole distribution by  $c$ . If everybody grew 5 inches suddenly, the variability in the population would still be the same.

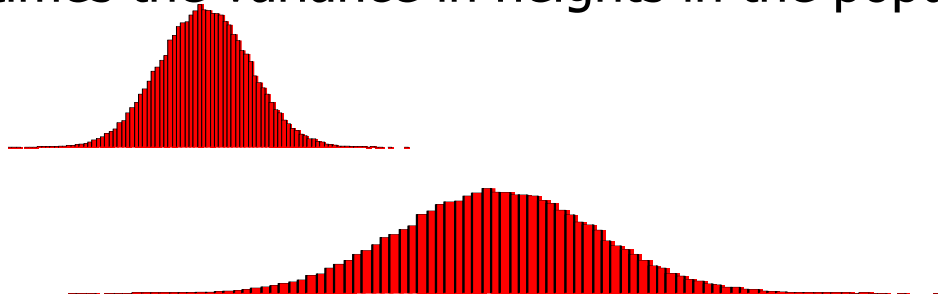



$$\text{Var}(cX) = c^2 \text{Var}(X)$$

---

$$\text{Var}(cX) = c^2 \text{Var}(X)$$

Multiplying each instance of the random variable by  $c$  makes it  $c$ -times as wide of a distribution, which corresponds to  $c^2$  as much variance (deviation squared). For example, if everyone suddenly became twice as tall, there'd be twice the deviation and 4 times the variance in heights in the population.




$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

---

$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$  ***ONLY IF X and Y are independent!!!!!!!***



# Statistics for Health Care

---

## Module 4:

## The binomial distribution



# Binomial Probability Distribution

---

- A fixed number of trials,  $n$ 
  - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- A binary outcome
  - e.g., head or tail in each toss of a coin; disease or no disease
  - Probability of “success” is  $p$ , probability of “failure” is  $1 - p$
- Constant probability for each trial
  - e.g., Probability of getting a tail is the same each time we toss the coin



# Binomial distribution

---

Take the example of 5 coin tosses.  
What's the probability that you flip  
exactly 3 heads in 5 coin tosses?





# Binomial distribution

---

*Solution:*

One way to get exactly 3 heads: HHHTT

What's the probability of this exact arrangement?

$$P(\text{heads}) \times P(\text{heads}) \times P(\text{heads}) \times P(\text{tails}) \times P(\text{tails}) = (1/2)^3 \times (1/2)^2$$

Another way to get exactly 3 heads: THHHT

$$\text{Probability of this exact outcome} = (1/2)^1 \times (1/2)^3 \times (1/2)^1 = (1/2)^3 \times (1/2)^2$$



# Binomial distribution

---

In fact,  $(1/2)^3 \times (1/2)^2$  is the probability of each unique outcome that has exactly 3 heads and 2 tails.

So, the overall probability of 3 heads and 2 tails is:

$(1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 + \dots$   
for as many unique arrangements as there are—but  
how many are there??

$$\binom{5}{3}$$

ways to  
arrange 3  
heads in  
5 trials

$${}_5C_3 = 5!/3!2! = 10$$

| Outcome   | Probability              |
|---|--------------------------|
| THHHT   | $(1/2)^3 \times (1/2)^2$ |
| HHHTT   | $(1/2)^3 \times (1/2)^2$ |
| TTHHH   | $(1/2)^3 \times (1/2)^2$ |
| HTTHH   | $(1/2)^3 \times (1/2)^2$ |
| HHTTH   | $(1/2)^3 \times (1/2)^2$ |
| HTHHT   | $(1/2)^3 \times (1/2)^2$ |
| THTHH   | $(1/2)^3 \times (1/2)^2$ |
| HTHTH   | $(1/2)^3 \times (1/2)^2$ |
| HHTHT   | $(1/2)^3 \times (1/2)^2$ |
| THHTH   | $(1/2)^3 \times (1/2)^2$ |
| 10 arrangements $\times (1/2)^3 \times (1/2)^2$ |                          |

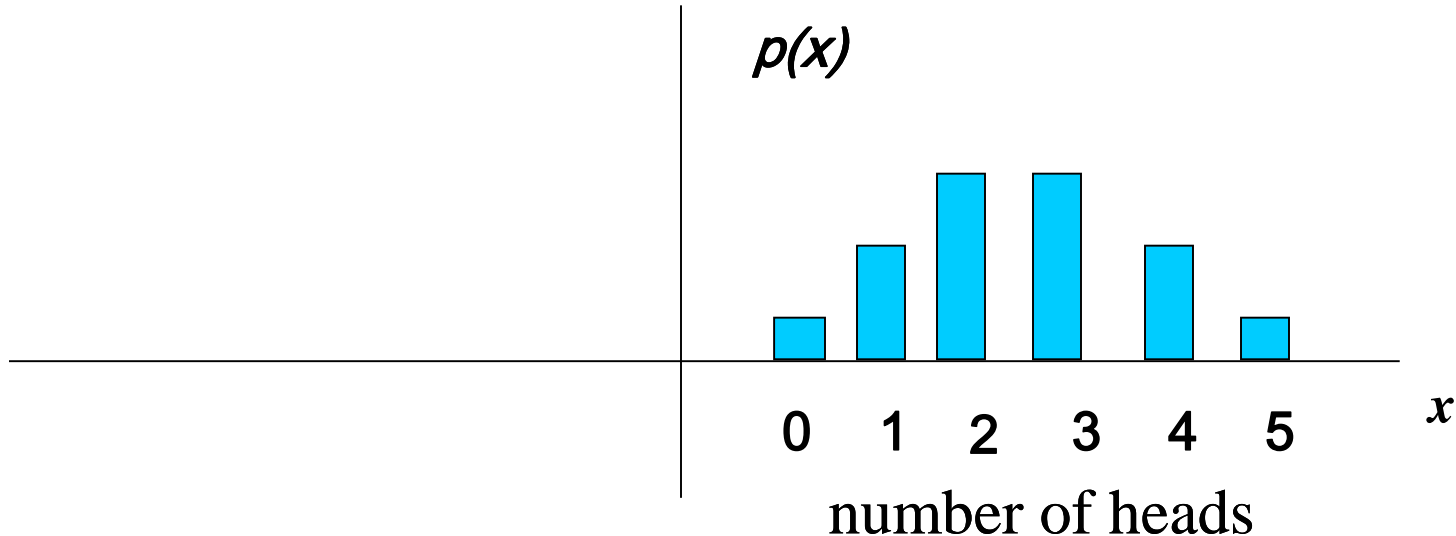
The probability  
of each unique  
outcome (note:  
they are all equal)

$$\therefore P(3 \text{ heads and } 2 \text{ tails}) = \binom{5}{3} \times P(\text{heads})^3 \times P(\text{tails})^2 =$$

$$10 \times (1/2)^5 = 31.25\%$$

# Binomial distribution function

$X$  = the number of heads tossed in 5 coin tosses





## Example 2

---

As voters exit the polls, you ask a representative random sample of 6 voters if they voted for a candidate A. If the true percentage of voters who vote for the candidate A is 55.1%, what is the probability that, *in your sample*, exactly 2 voted for the candidate A and 4 did not?



# *Solution:*

$\binom{6}{2}$   
 ways to  
 arrange 2 Prop  
 100 votes  
 among 6  
 voters

| Outcome | Probability   |
|---------|---|
| YYNNNN  | $= (.551)^2 \times (.449)^4$  |
| NYYNNN  | $(.449)^1 \times (.551)^2 \times (.449)^3 = (.551)^2 \times (.449)^4$ |
| NNYYNN  | $(.449)^2 \times (.551)^2 \times (.449)^2 = (.551)^2 \times (.449)^4$ |
| NNNYYN  | $(.449)^3 \times (.551)^2 \times (.449)^1 = (.551)^2 \times (.449)^4$ |
| NNNNYY  | $(.449)^4 \times (.551)^2 = (.551)^2 \times (.449)^4$                 |

15 arrangements  $\times (.551)^2 \times (.449)^4$

$$\therefore P(2 \text{ yes votes exactly}) = \binom{6}{2} \times (.551)^2 \times (.449)^4 = 18.5\%$$

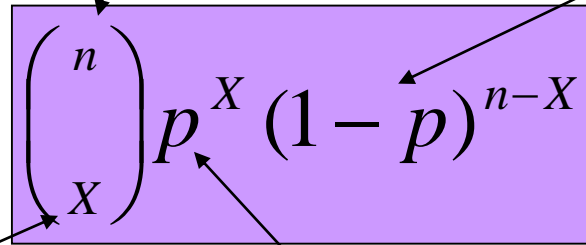


# Binomial distribution, generally

Note the general pattern emerging → if you have only two possible outcomes (call them 1/0 or yes/no or success/failure) in  $n$  independent trials, then the probability of exactly  $X$  “successes”=

$n$  = number of trials

$1-p$  = probability  
of failure


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

$X$  = # successes  
out of  $n$  trials

$p$  = probability  
of success





# Binomial

---

- We write:  $\mathbf{X} \sim \mathbf{Bin}(n, p)$ 
  - *Read as: "X is distributed binomially with parameters n and p"*
- And the probability that there are exactly  $X$  successes is:

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



# Binomial distribution: example

---

- Ten patients with wrinkles were photographed before and after treatment with a new anti-aging treatment. An independent dermatologist was able to distinguish the pre and post photographs for 9 out of the 10 subjects.
- If the anti-aging treatment is completely ineffective, what's the probability that the dermatologist could have gotten at least 9 right purely by lucky guessing?



# Example

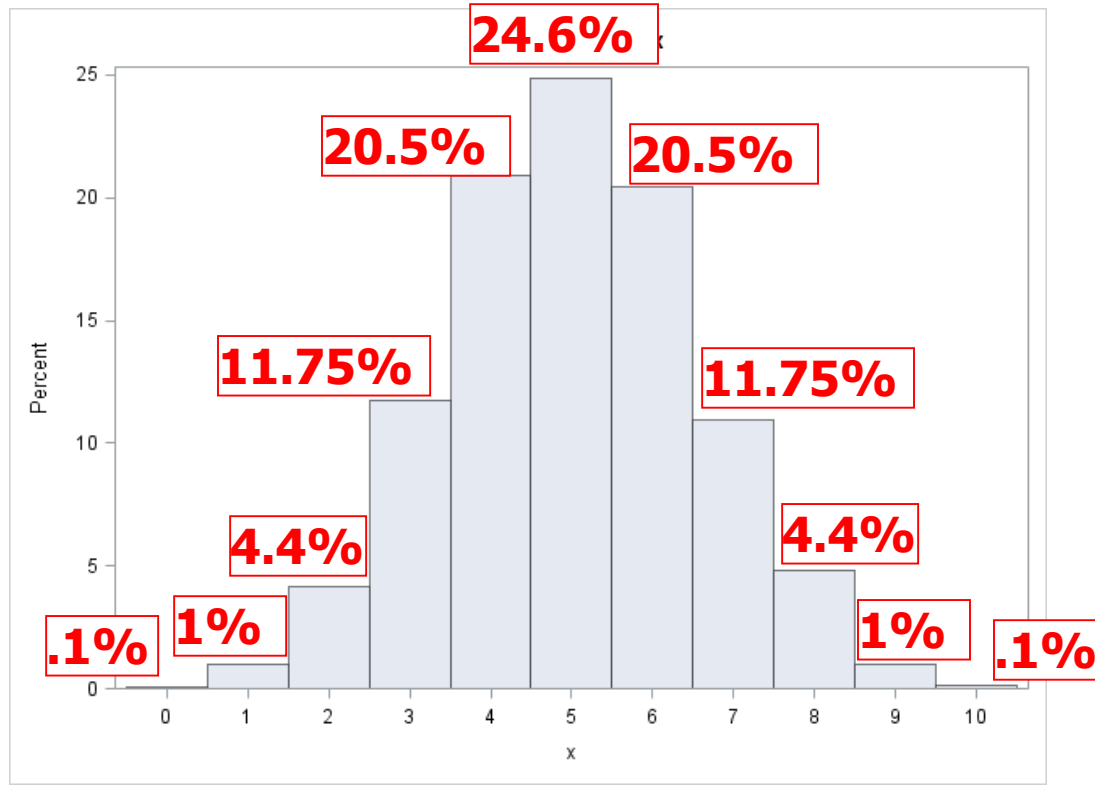
---

$$***X \sim Bin(10, 0.5)***$$

$$***P(X \geq 9) = P(X=9) + P(X=10)***$$

$$\begin{aligned} P(X \geq 9) &= \binom{10}{9} (.5)^9 (1-.5)^1 + \binom{10}{10} (.5)^{10} (1-.5)^0 \\ &= \frac{10!}{9!1!} (.5)^9 (.5)^1 + \frac{10!}{10!0!} (.5)^{10} = 10 * (.5)^9 + (.5)^{10} = 0.01 + .001 = 0.011 \end{aligned}$$

# The full probability distribution:





# Practice Problem:

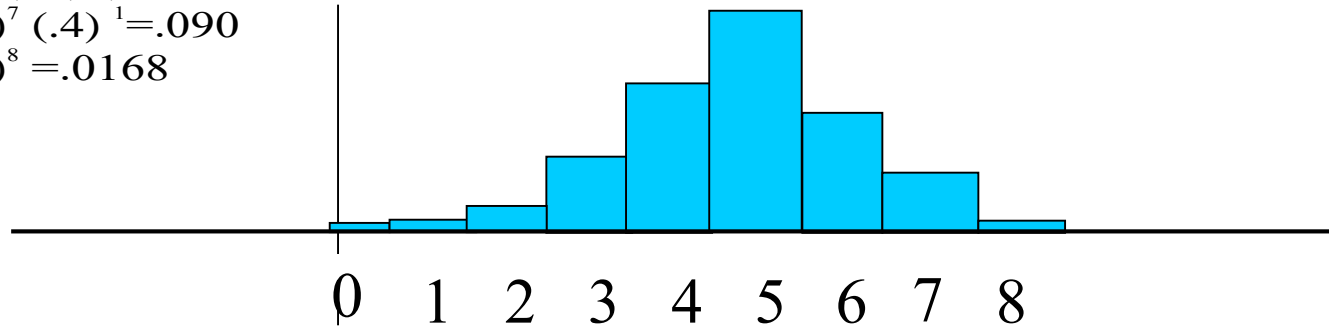
---

You are conducting a case-control study of smoking and lung cancer. If the probability of being a smoker among lung cancer cases is .6, what's the probability that in a group of 8 cases you have:

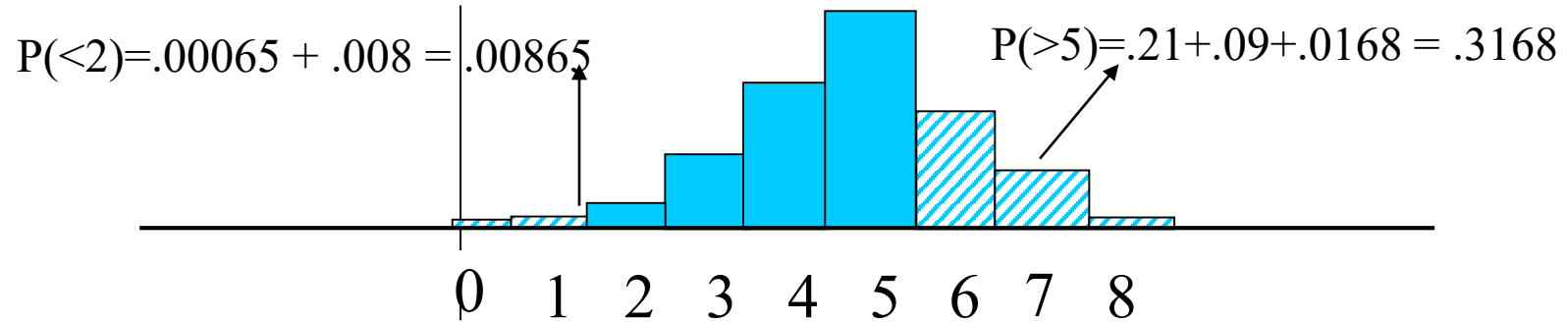
- a. Less than 2 smokers?
- b. More than 5?

# Answer

| X | P(X)                    |
|---|-------------------------|
| 0 | $1(.4)^8 = .00065$      |
| 1 | $8(.6)^1 (.4)^7 = .008$ |
| 2 | $28(.6)^2 (.4)^6 = .04$ |
| 3 | $56(.6)^3 (.4)^5 = .12$ |
| 4 | $70(.6)^4 (.4)^4 = .23$ |
| 5 | $56(.6)^5 (.4)^3 = .28$ |
| 6 | $28(.6)^6 (.4)^2 = .21$ |
| 7 | $8(.6)^7 (.4)^1 = .090$ |
| 8 | $1(.6)^8 = .0168$       |



# Answer, continued





**\*\*All probability distributions are characterized by an expected value and a variance:**

---

If  $X$  follows a binomial distribution with parameters  $n$  and  $p$ :  **$X \sim \text{Bin}(n, p)$**

Then:

$$E(X) = np$$

$$\text{Var}(X) = np(1-p)$$

$$\text{SD}(X) = \sqrt{np(1-p)}$$

**Note: the variance will always lie between**

$$0 * N \sim 0.25 * N$$

**$p(1-p)$  reaches maximum at  $p=.5$**

$$P(1-p) = .25$$





# Practice Problem

---

- You flip a coin 100 times. What are the expected value, variance, and standard deviation for the number of heads?



# Answer

---

$$E(X) = 100 (.5) = 50$$

$$\text{Var}(X) = 100 (.5) (.5) = 25$$

$$\text{SD}(X) = \text{square root } (25) = 5$$

Interpretation: When we toss a coin 100 times, we expect to get 50 heads plus or minus 5.

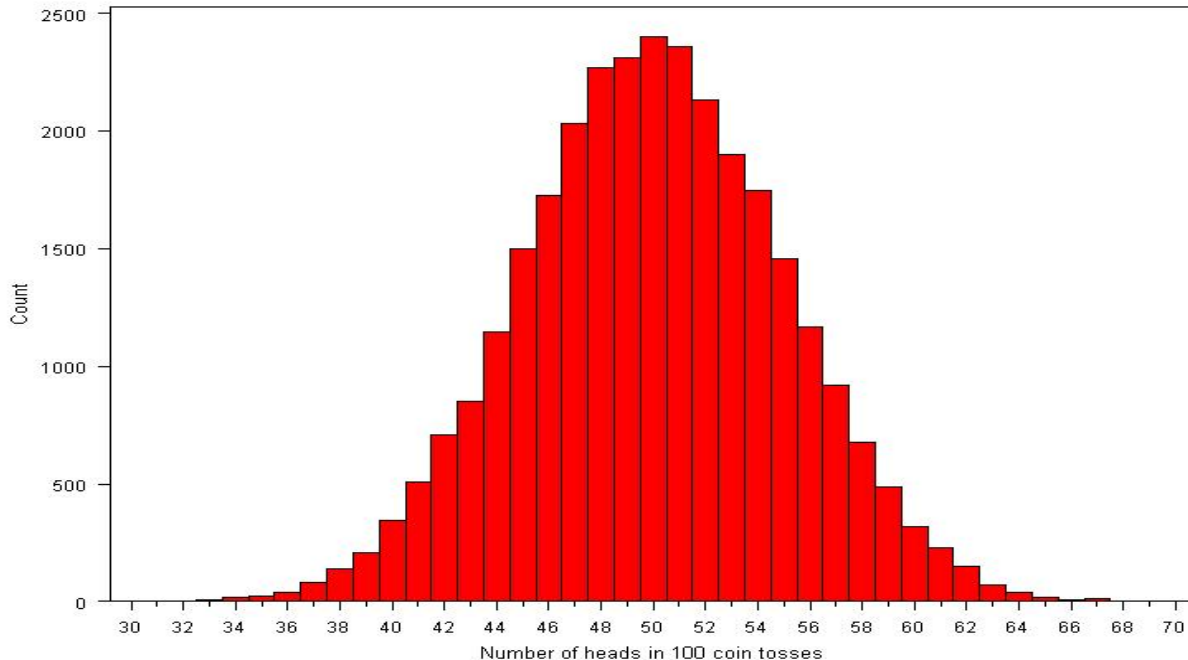


# Or use computer simulation...

---

- Flip coins virtually!
  - Flip a virtual coin 100 times; count the number of heads.
  - Repeat this over and over again a large number of times (we'll try 30,000 repeats!)
  - Plot the 30,000 results.

# Coin tosses...



**Mean = 50**

**Std. dev = 5**

**Follows a normal  
distribution**

**$\therefore$  95% of the time,  
we get between 40  
and 60 heads...**



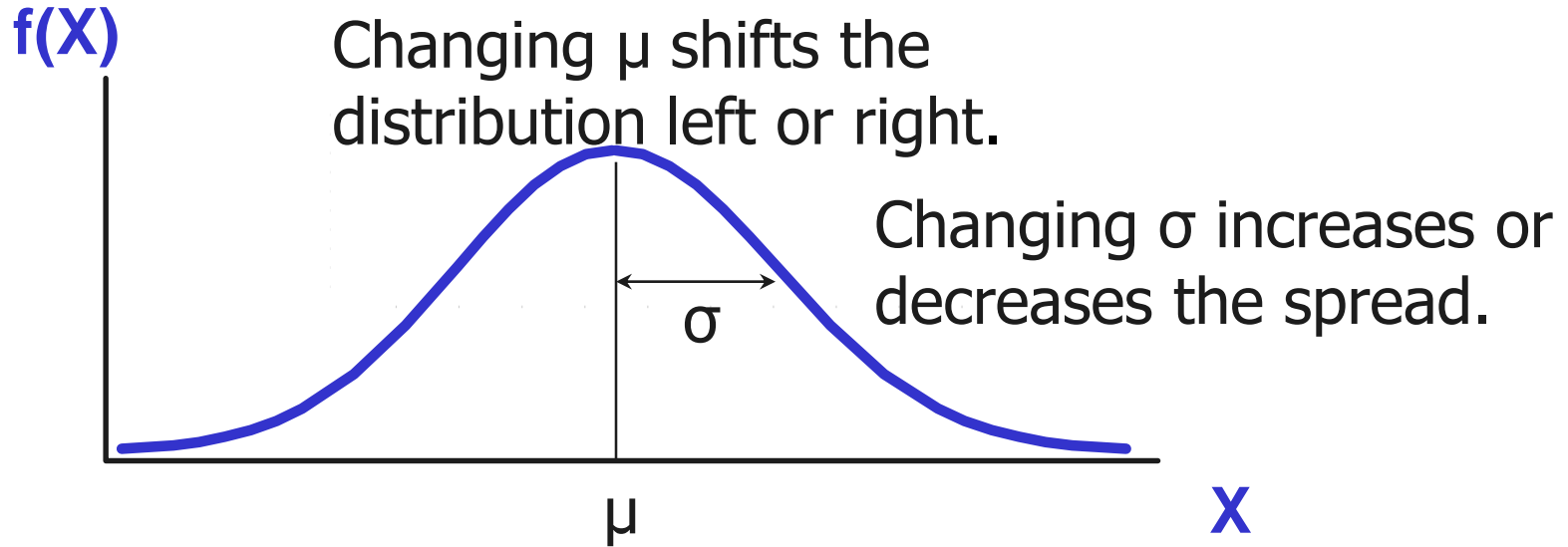
# Statistics for Health Care

---


## Module 5: The normal and standard normal distributions



# The Normal Distribution



# The Normal Distribution: as mathematical function (pdf)


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

Note constants:

$\pi=3.14159$

$e=2.71828$

This is a bell shaped curve with different centers and spreads depending on  $\mu$  and  $\sigma$



# The Normal PDF

---

It's a probability function, so no matter what the values of  $\mu$  and  $\sigma$ , must integrate to 1!

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$





# Normal distribution is defined by its mean and standard dev.

$$E(X)=\mu = \int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$\text{Var}(X)=\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$\text{Standard Deviation}(X)=\sigma$$

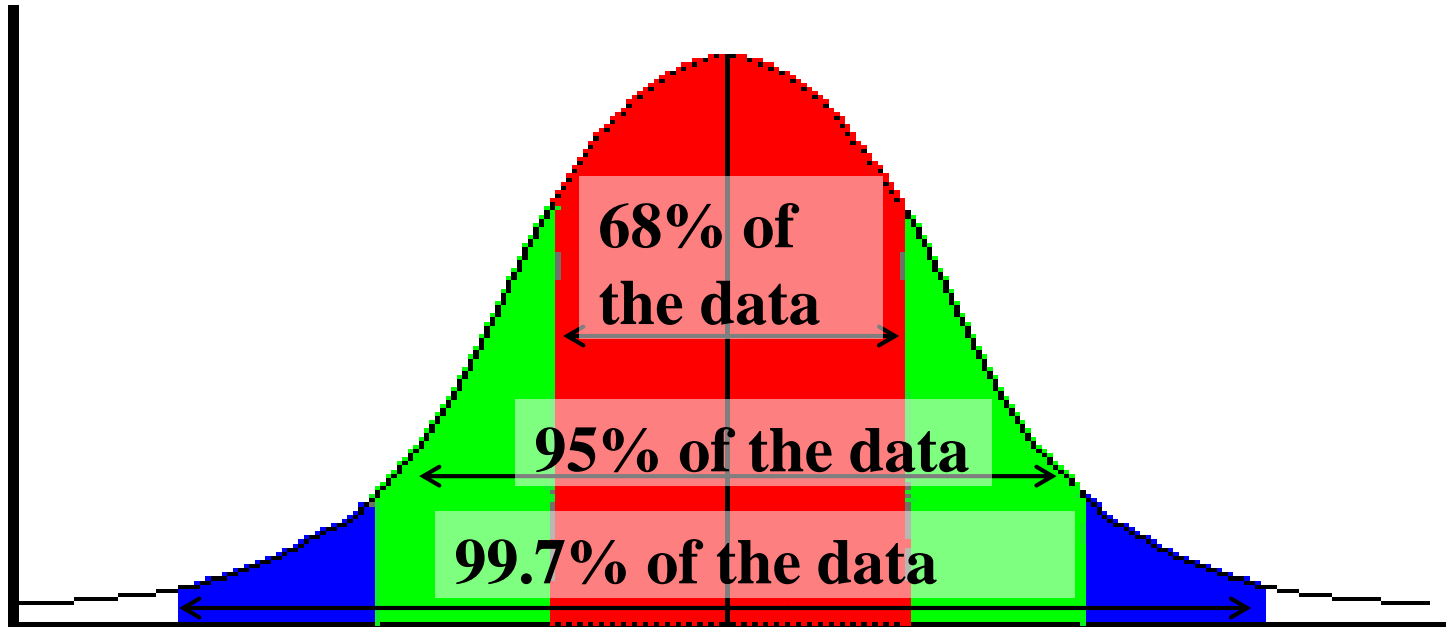


# Recall: 68-95-99.7 Rule

---

No matter what  $\mu$  and  $\sigma$  are, the area between  $\mu - \sigma$  and  $\mu + \sigma$  is about 68%; the area between  $\mu - 2\sigma$  and  $\mu + 2\sigma$  is about 95%; and the area between  $\mu - 3\sigma$  and  $\mu + 3\sigma$  is about 99.7%. Almost all values fall within 3 standard deviations.

# 68-95-99.7 Rule





# 68-95-99.7 Rule in Math!

---

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .68$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .997$$



# Example

---

- Suppose SAT scores roughly follows a normal distribution in the U.S. population of college-bound students (with range restricted to 200-800), and the average math SAT is 500 with a standard deviation of 50, then:
  - 68% of students will have scores between 450 and 550
  - 95% will be between 400 and 600
  - 99.7% will be between 350 and 650



# Example

---

- BUT: What's the probability of getting a math SAT score of 575 or less,  $\mu=500$  and  $\sigma=50$ ?

68-95-99.7 rule doesn't help here!

$$\int_{-\infty}^{575} \frac{1}{50\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-500}{50}\right)^2} dx = ?$$

**Solve this integral?**  
**No thanks!**



# The Standard Normal (Z): “Universal Currency”

The standard normal curve has a mean of 0  
and a standard deviation of 1.

$$p(Z) = \frac{1}{(1)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{Z-0}{1}\right)^2} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(Z)^2}$$



# The Standard Normal Distribution (Z)

---

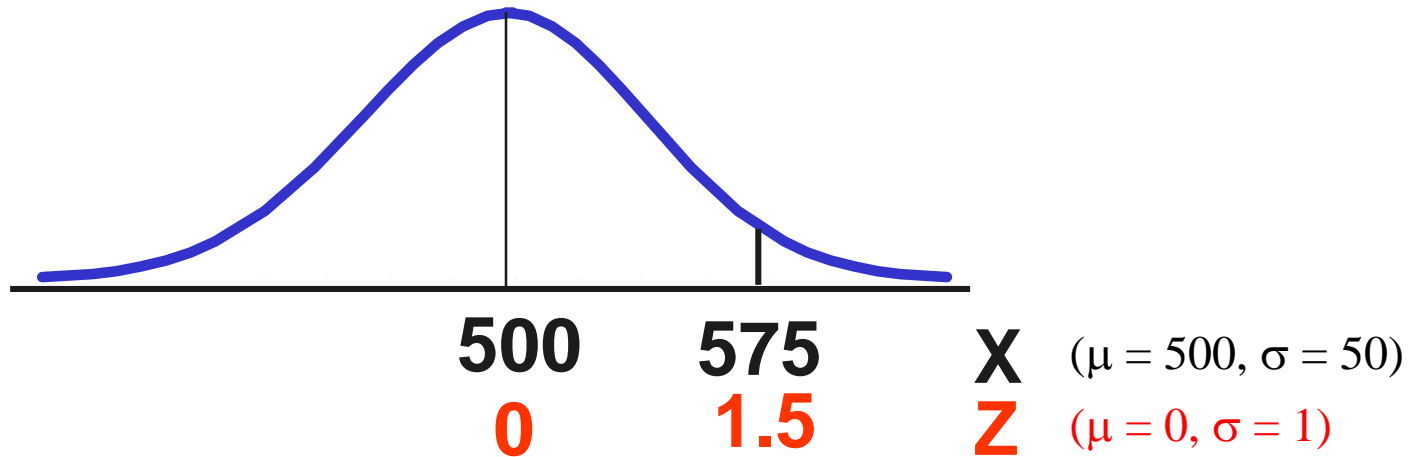
All normal distributions can be converted into the standard deviation units ("Z-scores") by subtracting the mean and dividing by the standard deviation:

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

Standard deviation units! Universal currency!



# Converting to the standard normal...



$$Z = \frac{575 - 500}{50} = 1.5$$



# Example

---

- What's the probability of getting a math SAT score of 575 or less,  $\mu=500$  and  $\sigma=50$ ?

$$Z = \frac{575 - 500}{50} = 1.5$$

- i.e., A score of 575 is 1.5 standard deviations above the mean



# Standard Normal Charts

---

- Someone integrated all the areas under the standard normal curve and put them in a chart.
- Look up  $Z = 1.5$  in standard normal chart  $\rightarrow .9332$

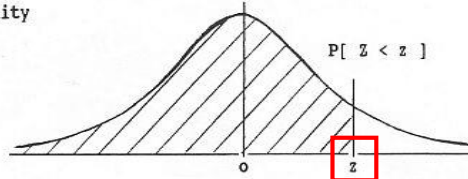
# Looking up probabilities in the standard normal table

## STANDARD STATISTICAL TABLES

### 1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value  $z$  i.e.

$$P[Z < z] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz$$



What is the area to the left of  $Z=1.50$  in a standard normal curve?

| z   | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   | 0.08   | 0.09   |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5159 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7854 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8804 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |

$Z=1.50$

Area is 93.32%

$Z=1.50$

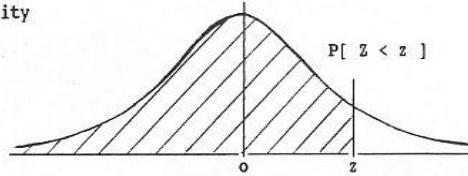
# Looking up probabilities in the standard normal table

## STANDARD STATISTICAL TABLES

### 1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value  $z$  i.e.

$$P[Z < z] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2) dz$$



What is the area to the left of  $Z=1.51$  in a standard normal curve?

| $z$ | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   | 0.08   | 0.09   |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5159 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7854 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8804 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |

$Z=1.51$

Area is 93.45%

$Z=1.51$



# Practice problem

---

If birth weights in a population are normally distributed with a mean of 109 oz and a standard deviation of 13 oz,

- a. What is the chance of obtaining a birth weight of 141 oz *or heavier* when sampling birth records at random?
- b. What is the chance of obtaining a birth weight of 120 *or lighter*?



# Answer

---

- a. What is the chance of obtaining a birth weight of 141 oz *or heavier* when sampling birth records at random?

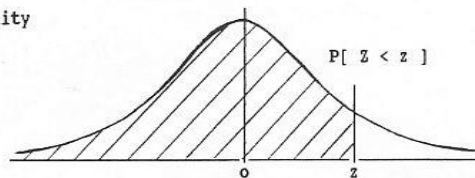
$$Z = \frac{141 - 109}{13} = 2.46$$

# STANDARD STATISTICAL TABLES

## 1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value  $z$  i.e.

$$P[Z < z] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2) dz$$



| z   | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   | 0.08   | 0.09   |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5159 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7854 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8804 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9773 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |
| 2.1 | 0.9821 | 0.9826 | 0.9830 | 0.9834 | 0.9838 | 0.9842 | 0.9846 | 0.9850 | 0.9854 | 0.9857 |
| 2.2 | 0.9861 | 0.9865 | 0.9868 | 0.9871 | 0.9874 | 0.9878 | 0.9881 | 0.9884 | 0.9887 | 0.9890 |
| 2.3 | 0.9893 | 0.9896 | 0.9898 | 0.9901 | 0.9904 | 0.9906 | 0.9909 | 0.9911 | 0.9913 | 0.9916 |
| 2.4 | 0.9918 | 0.9920 | 0.9922 | 0.9924 | 0.9927 | 0.9929 | 0.9931 | 0.9932 | 0.9934 | 0.9936 |
| 2.5 | 0.9938 | 0.9940 | 0.9941 | 0.9943 | 0.9945 | 0.9946 | 0.9948 | 0.9949 | 0.9951 | 0.9952 |
| 2.6 | 0.9953 | 0.9955 | 0.9956 | 0.9957 | 0.9959 | 0.9960 | 0.9961 | 0.9962 | 0.9963 | 0.9964 |
| 2.7 | 0.9965 | 0.9966 | 0.9967 | 0.9968 | 0.9969 | 0.9970 | 0.9971 | 0.9972 | 0.9973 | 0.9974 |
| 2.8 | 0.9974 | 0.9975 | 0.9976 | 0.9977 | 0.9977 | 0.9978 | 0.9979 | 0.9980 | 0.9980 | 0.9981 |
| 2.9 | 0.9981 | 0.9982 | 0.9982 | 0.9983 | 0.9984 | 0.9984 | 0.9985 | 0.9985 | 0.9986 | 0.9986 |
| z   | 3.00   | 3.10   | 3.20   | 3.30   | 3.40   | 3.50   | 3.60   | 3.70   | 3.80   | 3.90   |
| P   | 0.9986 | 0.9990 | 0.9993 | 0.9995 | 0.9997 | 0.9998 | 0.9998 | 0.9999 | 0.9999 | 1.0000 |

Area to the left of  $Z=2.46$  is .9931

Area to the right of 2.46 is:  
 $1 - .9931 = .0069$  or .69%





# Answer

---

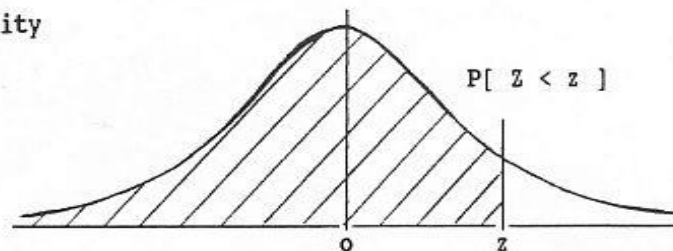
- b. What is the chance of obtaining a birth weight of 120 *or lighter*?

$$Z = \frac{120 - 109}{13} = .85$$

# Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value  $z$  i.e.

$$P[ Z < z ] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2) dz$$



| z   | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   | 0.08   | 0.09   |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5159 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7854 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8804 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9773 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |

Area to the left of  $Z=0.85$  is .8023 or 80.23%.



# Probit function: the inverse of the standard normal

---

$\phi(\text{area}) = Z$ : gives the Z-value that goes with the probability you want

For example, what if you wanted to know the math SAT score that corresponding to the 90<sup>th</sup> percentile (assuming a mean of 50 and a standard deviation of 50)?

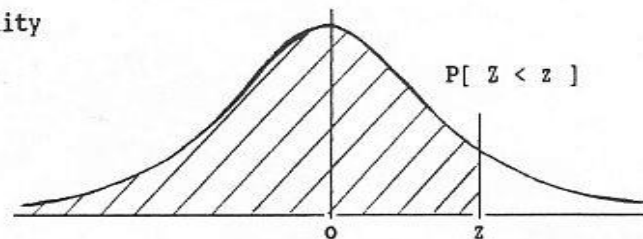
In the Table, find the Z-value that corresponds to an area of 90%...

# STANDARD STATISTICAL TABLES

## Areas under the Normal Distribution

The table gives the cumulative probability  
up to the standardised normal value  $z$   
i.e.

$$P[ Z < z ] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2) dz$$



| z   | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   | 0.08   | 0.09   |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5159 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7854 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8804 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9773 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |
| 2.1 | 0.9821 | 0.9826 | 0.9830 | 0.9834 | 0.9838 | 0.9842 | 0.9846 | 0.9850 | 0.9854 | 0.9857 |

90% area  
corresponds to a Z  
score of about 1.28.



# Probit function: the inverse

---

$Z=1.28$ ; convert back to raw SAT score →



# Statistics for Health Care

---

## Module 6:

The normal approximation to the  
binomial

# Normal approximation to the binomial

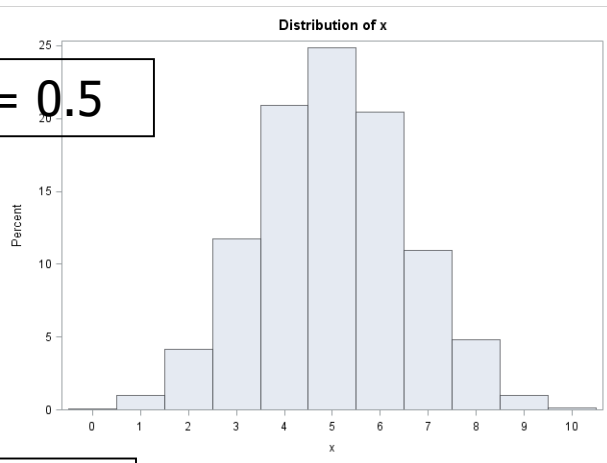


---

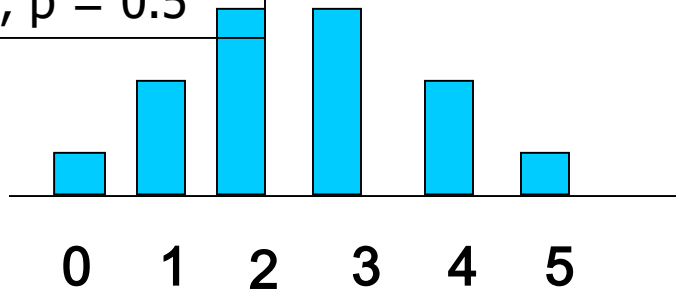
When you have a binomial distribution where the expected value is greater than 5 ( $np > 5$ ), then the binomial starts to look like a normal distribution...

# Binomial looks normal!

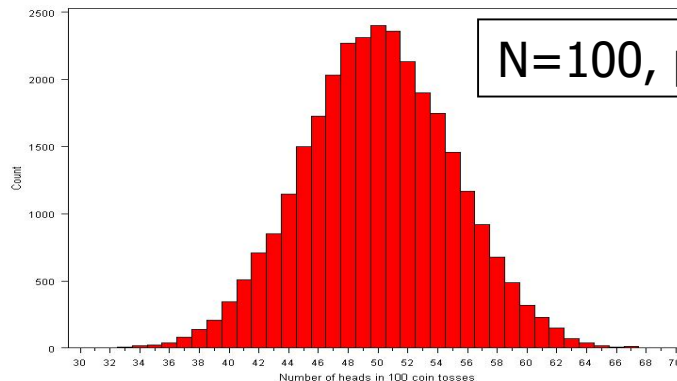
$N=10, p = 0.5$



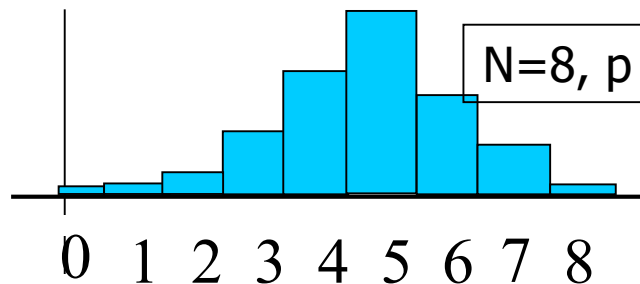
$N=5, p = 0.5$



$N=100, p = 0.5$



$N=8, p = 0.6$







# Normal approximation to the binomial

---

So, we can approximate it as a normal curve with mean= $np$  and variance =  $np(1-p)$ .



# Example

---

You are performing a cohort study. If the probability of developing disease in the exposed group is .25 for the study duration, then if you sample (randomly) 500 exposed people, What's the probability that **at most** 120 people develop the disease?



# Answer

---

**By hand:**

$$P(X \leq 120) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + \dots + P(X=120) =$$
$$\binom{500}{120} (.25)^{120} (.75)^{380} + \binom{500}{2} (.25)^2 (.75)^{498} + \binom{500}{1} (.25)^1 (.75)^{499} + \binom{500}{0} (.25)^0 (.75)^{500} \dots$$

**OR use, normal approximation:**

$$\mu = np = 500(.25) = 125 \text{ and } \sigma^2 = np(1-p) = 93.75; \sigma = 9.68$$

$$Z = \frac{120 - 125}{9.68} = -.52$$

$$P(Z < -.52) = .3015$$



# The binomial forms the basis of statistics on proportions...

---

- A proportion is just a binomial count divided by  $n$ .
  - For example, if we sample 200 cases and find 60 smokers,  $X=60$  but the observed proportion  $=.30$ .
- Statistics for proportions are similar to binomial counts, but differ by a factor of  $n$ .



# Stats for proportions

For binomial:

$$\mu_x = np$$

$$\sigma_x^2 = np(1-p)$$

$$\sigma_x = \sqrt{np(1-p)}$$

Differs by  
a factor of  
n.

For proportion:

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}}^2 = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Differs  
by a  
factor  
of n.

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

P-hat stands for "sample  
proportion."



# It all comes back to normal...

---

- Statistics for proportions are based on a normal distribution, because the binomial can be approximated as normal if  $np > 5$ !
- If  $np < 5$ , we instead use an “exact binomial” approach.



# Statistics for Health Care

---

## Module 7:

### Assessing normality in data




# Are my data “normal”?

---

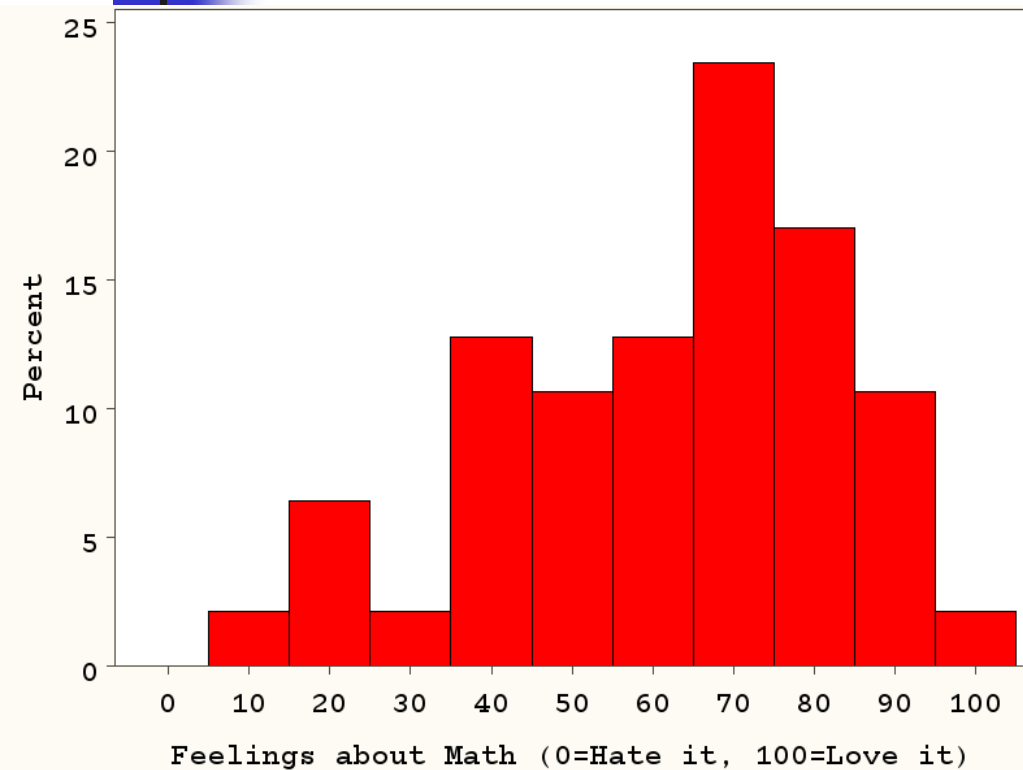
- Some statistical tests assume that the data are normally distributed (especially important for small samples).
- Not all continuous data are normally distributed!
- How do you test for normality?



# Are my data normally distributed?

- 
1. Look at the histogram! Does it appear bell shaped?
  2. Look at a normal probability plot—is it approximately linear?
  3. Look at descriptive statistics. Are the mean and median similar? Do 2/3 of observations lie within 1 std dev of the mean? Do 95% of observations lie within 2 std dev of the mean?
  4. Run tests of normality (such as Kolmogorov-Smirnov). But, be cautious, highly influenced by sample size!

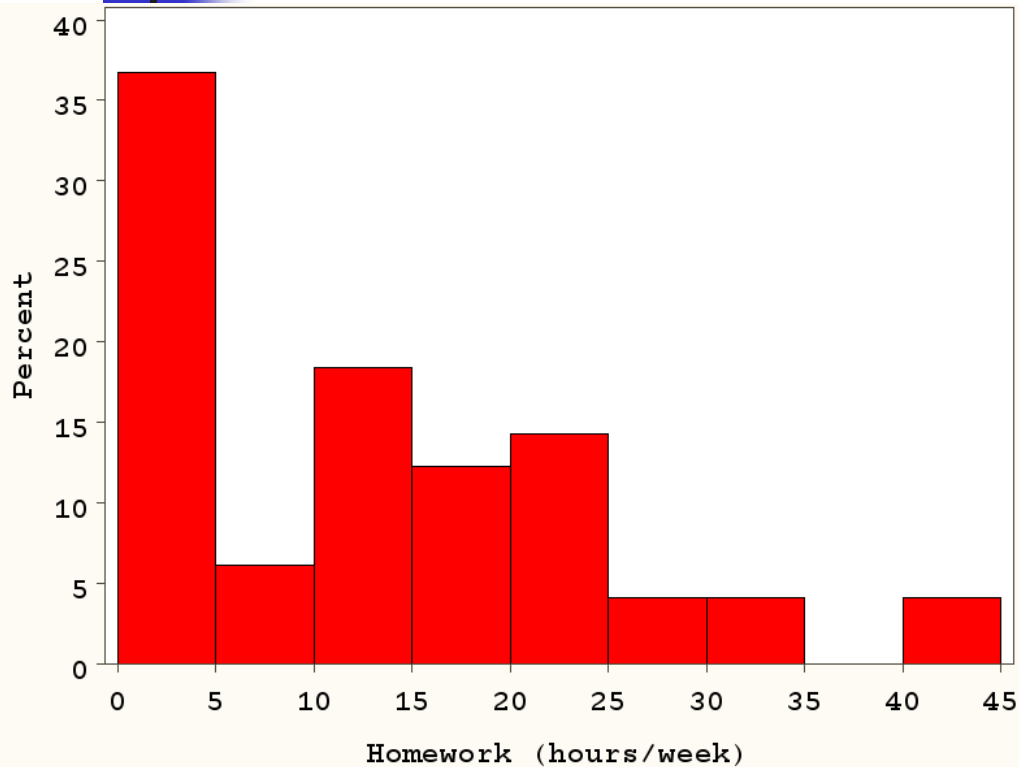
# Feelings about math...



Median = 65

Mean = 61

# Homework...



Median = 10.0

Mean = 11.4